PROGRAM:

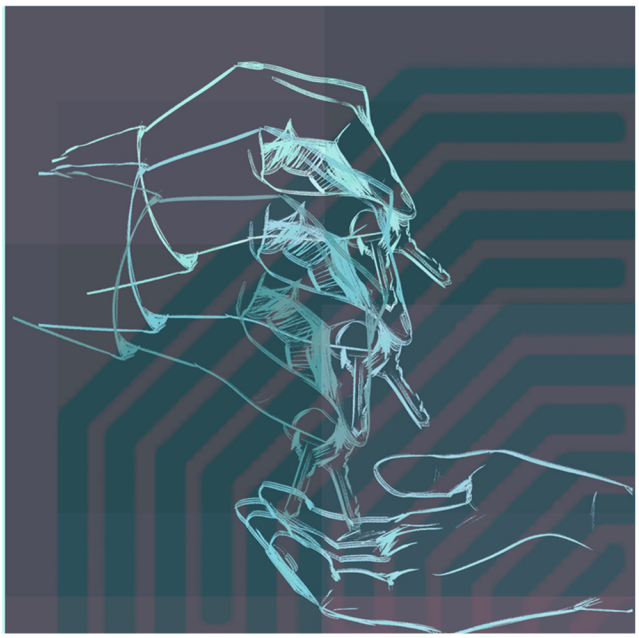
COGSCI 2016

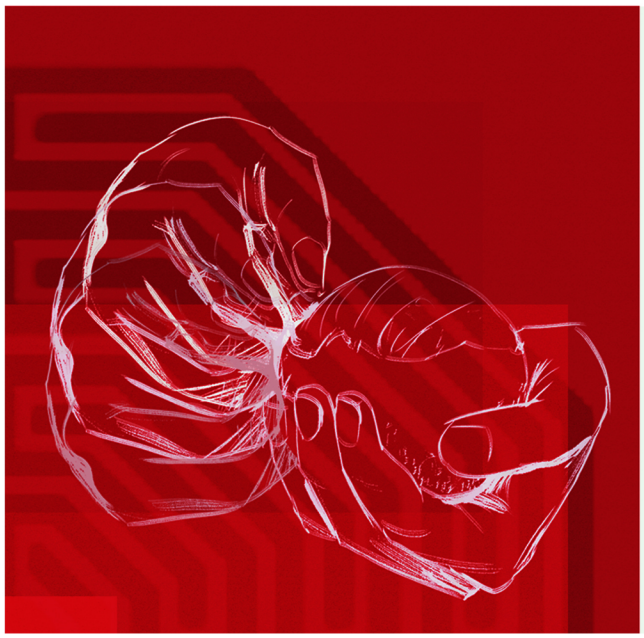
38th Annual Meeting of the Cognitive Science Society



Recognizing and Representing Events

Philadelphia, Pennsylvania August 10 - 13, 2016

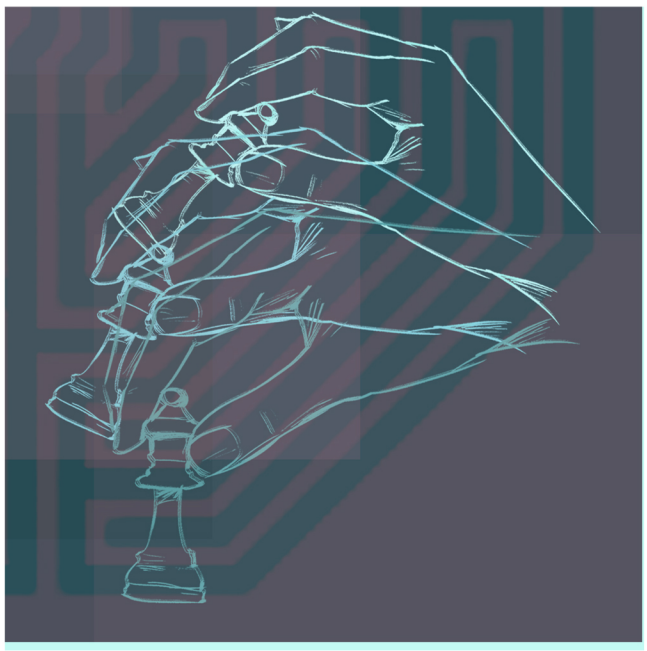






Edited by:

Anna Papafragou
Daniel Grodner
Daniel Mirman
John Trueswell



ISBN: 978-0-9911967-3-9

Table of Contents

Introduction 1

Committees 4

Conference Awards 8

Invited Presentations 11

Workshops

Active learning: Cognitive development, education, and computational models 13
Elizabeth Bonawitz, Fei Xu

Learning to Talk about Events: Grounding Language Acquisition in Intuitive Theories and Event Cognition 15
Eva Wittenberg, Melissa Kline, Joshua Hartshorne

Proposal of the Second Workshop on Physical and Social Scene Understanding 17
Tao Gao, Chenfanfu Jiang, Yixin Zhu, Yibiao Zhao, Lap-Fai (craig) Yu

Tutorial Workshop on Contemporary Deep Neural Network Models 19
James L. McClelland, Steven S. Hansen, Andrew Saxe

Workshop on Corpus Collection, (Semi)-Automated Analysis, and Modeling of Large-Scale Naturalistic Language Acquisition Data 21
Elika Bergelson

Interactive spatiotemporal cognition: Data, theories, architectures, and autonomy 23
Sangeet Khemlani, Greg Trafton

Tutorials

Full Day Tutorial on Quantum Models of Cognition and Decision 25
Jennifer Trueblood, James Yearsley, Peter Kvam, Zheng Wang, Jerome Busemeyer

Wallace: Automating Cultural Evolution Experiments Through Crowdsourcing 27
Jordan W. Suchow, Thomas J. H. Morgan, Jessica B. Hamrick, Michael D. Pacer, Stephan C. Meylan, Thomas L. Griffiths

Analysing discourse relations in natural language: The case of space and time 29
Thora Tenbrink

Cognitive models of transfer of cognitive skills 31
Niels Taatgen

Tutorial: Meta-Analytic Methods for Cognitive Science 33
Sho Tsuji, Molly Lewis, Christina Bergmann, Mike Frank, Alejandrina Cristia

Symposia

Empirical and Computational Approaches to Metaphor and Figurative Meaning 35
Justine Kao, Noah Goodman

Comics and cognitive systems: The processing of visual narratives 37
Neil Cohn, Emily Coderre, L.n. Kendall, Joe Magliano

Effecting re-representation: revising false beliefs and fostering creativity 39
Micah Goldwater, Zachary Horne, Jeffrey Loewenstein, Robert Goldstone

The Emergence of Conventions 41
Robert Hawkins, Noah Goodman, Olga Feher, Kenny Smith, Robert Goldstone, Tom Griffiths

The cultural evolution of cognition 43
Sieghard Beller, Andrea Bender, Christine Caldwell, Morten Christiansen, Karenleigh Overmann, Kenny Smith

Brain Science and Education: Is it Still a Bridge Too Far? 45
Ray Perez, Danielle McNamara, Gregg Solomon, Wayne Gray

Research at the Interface of Cognition, Education, and Disciplinary Science 47
Laura Novick, Mary Hegarty, Richard Catrambone, John Pani, Thomas Shipley

Beyond the language explosion: What gradual word learning tells us about conceptual development 49
Katharine Tillman, Katie Wagner, Junyi Chu, Mutsumi Imai, Barbara Malt, Sherri Widen, Marilyn Shatz

Aligning implicit learning and statistical learning: Two approaches, one phenomenon 51
Patrick Rebuschat, Padraic Monaghan, Rebecca Gomez, Gary Dell, Nathaniel Anderson, Christopher Conway

Higher-Level Goals in the Processing of Human Action Events 53
Michelle Eisenberg, Jeffrey Zacks, Shaney Flores, Lauren Howard, Amanda Woodward, Jeff Loucks, Andrew Meltzoff, Rick Cooper

Concepts from Event Semantics in Cognition 55
Alexis Wellwood, Jeremy Kuhn, Philippe Schlenker, Carlo Geraci, Brent Strickland, Susan Hespos, Lance Rips, E. Matthew Husband, Alexander Williams

Papers

Investigating the Effects of Transparency and Ambiguity on Idiom Learning 57
Mehrgol Tiv, Evelyn Milburn, Tessa Warren

Explaining December 4, 2015: Cognitive Science Ripped from the Headlines 63
Samuel Johnson

Modeling category learning using a dual-system approach: A simulation of Shepard, Hovalnd and Jenkins (1961) by COVIS 69
Charlotte Edmunds, Andy Wills

Using Prior Data to Inform Model Parameters in the Predictive Performance Equation 75
Michael Collins, Kevin Gluck, Matthew Walsh, Michael Krusmark, Glenn Gunzelmann

Using Statistics to Learn Words and Grammatical Categories: How High Frequency Words Assist Language Acquisition 81
Rebecca Frost, Padraic Monaghan, Morten Christiansen

Modeling Impairments in Lexical Development 87
Michael Vinos, Angeliki Andrikopoulou, Christina F. Papaeliou, Athanassios Protopapas

Monolinguals' and Bilinguals' Use of Language in Forming Novel Object Categories 93
Sarah Fairchild, Anna Papafragou

Learning How To Throw Darts: The Effect Of Modeling Type And Reflection On Dart-Throwing Skills 99
Janneke van der Loo, Eefje Frissen, Emiel Krahmer

On the Tragedy of Personnel Evaluation 105
Momme von Sydow, Niels Braus

Are we ON the same page? Monolingual and bilingual acquisition of familiar and novel relational language 111
Nathan George, Junko Kanero, Dorothee Chwilla, Daniel Weiss

Dissociable effects of cue validity on bias formation and reversal 117
Angelo Pirrone, Qi Zhang, Sheng Li

Effects of experience in a developmental model of reading 123
Ya-Ning Chang, Padraic Monaghan, Stephen Welbourne

Differentiating between Encoding and Processing during Sequential Diagnostic Reasoning: An Eye tracking study. 129
Anja Klichowicz, Agnes Scholz, Sascha Strehlau, Josef F. Krems

Young children and adults integrate past expectations and current outcomes to reason about others' emotions 135
Desmond Ong, Mika Asaba, Hyowon Gweon

Language does not explain the wine-specific memory advantage of wine experts 141
Ilja Croijmans, Asifa Majid

Sub-Categorical Properties of Stimuli Determine the Category-Order Effect 147
Jordan Schoenherr, Robert Thomson

Causal Learning With Continuous Variables Over Time 153
Kevin Soo, Benjamin Rottman

Viewing time affects overspecification: Evidence for two strategies of attribute selection during reference production 159
Ruud Koolen, Albert Gatt, Roger Van Gompel, Emiel Krahmer, Kees Van Deemter

Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions 165
Jens Koed Madsen

Expressive faces are remembered with less pictorial fidelity than neutral faces 171
Martina Lorenzino, Giorgio Gronchi, Corrado Caudek

Three barriers to effective thought experiments, as revealed by a system that externalizes students' thinking 176
Miki Matsumuro, Kazuhisa Miwa

Mind reading: Discovering individual preferences from eye movements using switching hidden Markov models 182
Tim Chuk, Antoni B. Chan, Shinsuke (shin) Shimojo, Janet H. Hsiao

Semantic Contamination of Visual Similarity Judgments 188
Layla Unger, Anna Fisher

Human Reinforcement Learning of Sequential Action 193
George Kachergis, Floris Berends, Roy de Kleijn, Bernhard Hommel

The Role of the Prefrontal Cortex in Inductive Reasoning: An fNIRS Study 199
Layla Unger, Jaeah Kim, Theodore Huppert, Julia Badger, Anna Fisher

The Emergence of Linguistic Consciousness and the "hard problem" 205
Christina Behme

An Empirical Evaluation of Models for How People Learn Cue Search Orders 211
Percy Mistry, Michael Lee, Ben Newell

A Deep Siamese Neural Network Learns the Human-Perceived Similarity Structure of Facial Expressions Without Explicit Categories 217
Sanjeev Jagannatha Rao, Yufei Wang, Garrison Cottrell

On the adaptive nature of memory-based false belief 223
Hidehito Honda, Toshihiko Matsuka, Kazuhiro Ueda

Causal Action: A Fundamental Constraint on Perception of Bodily Movements 229
Yujia Peng, Steven Thurman, Hongjing Lu

A Computational Exploration of Problem-Solving Strategies and Gaze Behaviors on the Block Design Task 235
Maithilee Kunda, Mohamed El Banani, James Rehg

Preferring the Mighty to the Meek: Toddlers prefer Superior to Subordinate Individuals. 241
Ashley J. Thomas, Meline Abramyan, Angela Lukowski, Lotte Thomsen, Barbara W. Sarnecka

Unifying Conflicting Perspectives in Group Activities: Roles of Minority Individuals 247
Kazuhisa Miwa, Yugo Hayashi, Hitoshi Terai

Blink durations reflect mind wandering during reading 253
Stephanie Huette, Ariel Mathis, Art Graesser

The Influence of Reputation Concerns and Social Biases on Children's Sharing Behavior 259
Haleh Yazdi, David Barner, Gail Heyman

Are There Hidden Costs to Teaching Mathematics with Incorrect Examples? 265
Min Kyung Hong, Darren J. Yeo, Bethany Rittle-Johnson, Lisa K. Fazio

Interaction of Instructional Material Order and Subgoal Labels on Learning in Programming 271
Laura Schaeffer, Lauren Margulieux, Richard Catrambone

No Effect of Verbal Labels for the Shapes on Type II Categorization Tasks 277
Fotis A. Fotiadis, Athanassios Protopapas

Context-dependent Processes and Engagement in Reading Literature 283
Miho Fuyama, Shohei Hidaka

Effects of Working Memory Training on L2 Proficiency and Working Memory Capacity 289
Gregory Colflesh, Valerie Karuzis, Polly O'Rourke

Representation: Problems and Solutions 295
Nancy Salay

Lasting Political Attitude Change Induced by False Feedback About Own Survey Responses 301
David Sivén, Thomas Strandberg, Lars Hall, Petter Johansson, Philip Pärnamets

How should autonomous vehicles behave in moral dilemmas? Human judgments reflect abstract moral principles 307
Derek Powell, Patricia Cheng, Michael Waldmann

Examining Search Processes in Low and High Creative Individuals with Random Walks 313
Yoed Kenett, Joseph Austerweil

Determining the alternatives for scalar implicature 319
Benjamin Peloquin, Mike Frank

The Role of Similarity in Constructive Memory: Evidence from Tasks with Children and Adults . 325
Georgi Petkov, Margarita Pavlova

Dual process theory of reasoning and recognition memory errors: Individual differences in a memory prose task 331
Giorgio Gronchi, Stefania Righi, Giacomo Parrini, Lapo Pierguidi, Maria Pia Viggiano

A neurocomputational model of the effect of learned labels on infants' object representations 336
Arthur Capelier-Mourguy, Katherine Twomey, Gert Westermann

A neural network model of hierarchical category development 342
Chris Gorman, Alistair Knott

Improving Visual Memory with Auditory Input 348
Scott R. Schroeder, Viorica Marian

Measuring lay theories of parenting and child development 354
Emily Hembacher, Michael C Frank

Emotions in lay explanations of behavior 360
Desmond Ong, Jamil Zaki, Noah Goodman

But vs. Although under the microscope 366
Fatemeh Torabi Asr, Vera Demberg

Tangible models and haptic representations aid learning of molecular biology concepts 372
Kristen Johannes, Jacklyn Powers, Lisa Couper, Matt Silberglitt, Jodi Davenport

Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors 378
Till Poppels, Roger Levy

No stereotype threat effect in international chess 384
Tom Stafford

'Unlikely' Outcomes Might Never Occur, But What About 'Unlikely (20% Chance)' Outcomes? . 390
Sarah Jenkins, Adam Harris, Murray Lark

Visual Statistical Learning Deficits in Children with Developmental Dyslexia: an Event Related Potential Study 396
Sonia Singh, Anne Walk, Christopher Conway

The St. Petersburg Paradox: A Subjective Probability Solution 402
Hongbin Wang, Yanlong Sun, Jack Smith

A test of two models of probability judgment: quantum versus noisy probability 408
Fintan Costello, Paul Watts

Influence of 3D images and 3D-printed objects on spatial reasoning 414
Akihiro Maehigashi, Kazuhisa Miwa, Masahiro Oda, Yoshihiko Nakamura, Kensaku Mori, Tsuyoshi Igami

Making it Right: Can the Right-Hemisphere Compensate for Language Function in Patients with Left-Frontal Brain Tumors? 420
Ethan Jost, Nicole Brennan, Kyung Peck, Andrei Holodny, Morten Christiansen

Spatial Meaning is Retained in Emotion Metaphors: Some Evidence from Spanish 426
Cesar Riano, Florencia Reali

Explanatory Value, Probability, and Abductive Inference 432
Matteo Colombo, Marie Nilsenova, Jan Sprenger

Item Learning vs. High-Level Categorization in Consistent-Mapping Memory Search 438
Rui Cao, Robert Nosofsky, Richard Shiffrin

The impact of biased hypothesis generation on self-directed learning 444
Doug Markant

Representing Sequence: The Influence of Timeline Axis and Direction on Causal Reasoning in Litigation Law 450
Amy Fox, Martin van den Berg, Erica de Vries

Attractivity Weighting: Take-the-Best's Foolproof Sibling 456
Paul Thorn, Gerhard Schurz

A Perception-Based Threshold for Bidirectional Texture Functions 462
Banafsheh Azari, Sven Bertel, Charles Wuethrich

Influencing Categorical Choices through Physical Object Interaction 468
Nicholas Shipp, Frederic Vallee-Tourangeau, Susan Anthony

Dynamics of Strategy Adaptation in a Temporally Extended Monty Hall Dilemma 474
Stephanie Petrusz, Theo Rhodes, Joshua Shields, Abisha Munroe

Causal Contrasts Promote Algebra Problem Solving 480
Jian-Ping Ye, Jessica Walker, Patricia Cheng

A connectionist model for automatic generation of child-adult interaction patterns 486
Moinuddin M. Haque, Paul Vogt, Afra Alishahi, Emiel Krahmer

Connections between ACT-R's declarative memory system and Minerva2 492
Cvetomir Dimov

Distributed Cognition in the Past Progressive: Narratives as Representational Tools for Clinical Reasoning 496
Katherine Lippa, Valerie Shalin

Predicting Overprecision in Range Estimation 502
Matthew Kaesler, Matthew Welsh, Carolyn Semmler

A Dynamic Neural Field Model of Speech Cue Compensation 508
Gavin Jenkins, Paul Tupper

The face-space duality hypothesis: a computational model 514
Jonathan Vitale, Mary-Anne Williams, Benjamin Johnston

Fuse to be used: A weak cue's guide to attracting attention 520
Zara Harmon, Vsevolod Kapatsinski

Stereotype-Based Intuitions: A Psycholinguistic Approach to Experimental Philosophy's 'Sources Project' 526
Eugen Fischer, Paul Engelhardt

Mechanisms for storing and accessing event representations in episodic memory, and their expression in language: a neural network model 532
Martin Takac, Alistair Knott

What Makes You Feel You Are Learning: Cues to Self-Regulated Learning 538
Jessie Chin, Elizabeth Stine-Morrow

A Tale of Two Disasters: Biases in Risk Communication 544
Matthew Welsh, Sandy Steacy, Steve Begg, Daniel Navarro

Individual Differences in Pupil Dilation during Naming Task 550
Kaidi Lõo, Jacolien van Rij, Juhani Järvikivi, Harald Baayen

Modeling Commonsense Reasoning via Analogical Chaining: A Preliminary Report 556
Joseph Blass, Kenneth Forbus

Social Cues Modulate Cognitive Status of Discourse Referents 562
Kara Hawthorne, Anja Arnhold, Emily Sullivan, Juhani Järvikivi

Hysteresis in Processing of Perceptual Ambiguity on Three Different Timescales 568
Marieke van Rooij, Harald Atmanspacher, Jürgen Kornmeier

Critical Features of Joint Actions that Signal Human Interaction 574
Tianmin Shu, Steven Thurman, Dawn Chen, Song-Chun Zhu, Hongjing Lu

Visual constraints modulate stereotypical predictability of agents during situated language comprehension 580
Alba Rodríguez, Michele Burigo, Pia Knoeferle

Biases and Benefits of Number Lines and Pie Charts in Proportion Representation 586
Michelle Hurst, Charlotta Relander, Sara Cordes

Exploring the Relationship between Adolescents' Interest in Algebra and Procedural Declines 592
Natalie Corbett, Julie Booth, Christina Barbieri, Laura Young

Factors Influencing Categorization Strategy in Visual Category Learning 596
Sujith Thomas, Harish Karnick

Simulating Developmental Changes in Noun Richness through Performance-limited Distributional Analysis 602
Daniel Freudenthal, Julian Pine, Gary Jones, Fernand Gobet

The comprehension of English Garden-path sentences by Mandarin and Korean learners of English as a second language 608
Zhiying Qian

Choice adaptation to increasing and decreasing event probabilities 614
Samuel Cheyette, Emmanouil Konstantinidis, Jason Harman, Cleotilde Gonzalez

A Robust Implementation of Episodic Memory for a Cognitive Architecture 620
David Menager, Dongkyu Choi

Consistency and credibility in legal reasoning: A Bayesian network approach 626
Saoirse Connor Desai, Stian Reimers, David Lagnado

Attention and the Development of Inductive Generalization: Evidence from Recognition Memory 632
Tracey Miser, Vladimir Sloutsky

Predicting Decision in Human-Agent Negotiation using functional MRI 638
Eunkyung Kim, Sarah Gimbel, Aleksandra Litvinova, Jonas Kaplan, Morteza Dehghani

Searching large hypothesis spaces by asking questions 644
Alexander Cohen, Brenden Lake

Using a smartphone game to promote transfer of skills in a real world environment 650
Inge Doesburg, Niels Taatgen

Solution of division by access to multiplication: Evidence from eye tracking 656
Shawn Tan, Kasia Muldner, Jo-Anne Lefevre

The role of word-word co-occurrence in word learning 662
Abdellah Fourtassi, Emmanuel Dupoux

Where Should Researchers Look for Strategy Discoveries during the Acquisition of Complex Task Performance? The Case of Space Fortress 668
Marc Destefano, Wayne Gray

Developmentally plausible learning of word categories from distributional statistics 674
Daniel Freudenthal, Julian Pine, Gary Jones, Fernand Gobet

A Learned Label Modulates Object Representations in 10-Month-Old Infants 680
Katherine Twomey, Gert Westermann

An experimental study on the observation of facts in explanation reconstruction 686
Hitoshi Terai, Kazuhisa Miwa, Naohiro Toyama

Gesture reveals spatial analogies during complex relational reasoning 692
Kensy Cooperrider, Dedre Gentner, Susan Goldin-Meadow

What Determines Human Certainty? 698
Louis Marti, Steven Piantadosi, Francis Mollica, Celeste Kidd

Moral Judgments: Studying People with Helping Professions 704
Maurice Grinberg, Evgeniya Hristova, Veselina Kadreva

Children consider others' expected costs and rewards when deciding what to teach 710
Sophie Bridgers, Julian Jara-Ettinger, Hyowon Gweon

Disfluency production in speech and gesture 716
Niloofar Akhavan, Tilbe Goksun, Nazbanou Nozari

Predictions with Uncertain Categorization: A Rational Model 722
Elizaveta Konovalova, Gael Le Mens

Numeric Competencies and Anchoring Biases 728
Sangsuk Yoon, Nathan Fong

A perspective on all cognition? A study of everyday environments from the perspective of distributed cognition 734
Nils Dahlbäck, Mattias Kristiansson

Temporal Structure Modulates ERP Correlates of Visual Sequential Learning 740
Kimberly Ross, Christopher Conway

Knowledge and use of price distributions by populations and individuals 746
Timothy Lew, Edward Vul

A Unified Framework for Bounded and Unbounded Numerical Estimation 752
Dan Kim, John Opfer

Temporal Causal Strength Learning with Multiple Causes 758
Cory Derringer, Benjamin Rottman

Document Cohesion Flow: Striving towards Coherence 764
Scott Crossley, Mihai Dascalu, Stefan Trausan-Matu, Laura Allen, Danielle McNamara

The Relationship Between Inhibitory Control and Free Will Beliefs in 4-to 6-Year-Old-Children . 770
Adrienne Wente, Titus Ting, Rosie Aboody, Tamar Kushnir, Alison Gopnik

Explanatory Biases in Social Categorization 776
Samuel Johnson, Haylie Kim, Frank Keil

Are children flexible speakers? Effects of typicality and listener needs in children's event descriptions 782
Myrto Grigoroglou, Anna Papafragou

From embodied metaphors to metaphoric gestures 788
Margot Lhommet, Stacy Marsella

Relation between bimanual coordination and whole-body balancing on a slackline 794
Kentaro Kodama, Yusuke Kikuchi, Hideo Yamagiwa

Causality, Normality, and Sampling Propensity 800
Thomas Icard, Joshua Knobe

Selecting Explanations from Causal Chains: Transitivity Intuitions Require Exportable Mechanisms 806
Jonas Nagel, Simon Stephan

The Variable Relationship Between On-Task Behavior and Learning 812
Karrie Godwin, Howard Seltman, Ma. Victoria Almeda, Shimin Kai, Ryan Baker, Anna Fisher

The effect of "mood": Group-based collaborative problem solving by taking different perspectives .. 818
Yugo Hayashi

The Impact of Interactivity on Simulation-Based Science Inquiry with Variable-Setting Controls 824
Jung Aa Moon, Michelle Lamar, Carolyn Forsyth, Madeleine Keehner

Causal Reasoning in Infants and Adults: Revisiting backwards-blocking 830
Deon Benton, David Rakison

An Analysis of Frame Semantics of Continuous Processes 836
Clifton McFate, Ken Forbus

Inferring Generic Meaning From Pragmatic Reference Failure 842
Philip Crone, Mike Frank

Effect of Aging on Inhibitory Attentional Mechanisms 848
Maegen Walker, Margeaux Ciraolo, Andrew Dewald, Scott Sinnett

Analytic Eye Movement Patterns in Face Recognition are Associated with Better Performance and more Top-down Control of Visual Attention: an fMRI Study 854
Cynthia Y. H. Chan, J. J. Wong, Antoni B. Chan, Tatia M. C. Lee, Janet H. Hsiao

Establish Trust and Express Attitude for a Non-Humanoid Robot 860
Mei Si, J. Dean McDaniel

Similarity-Based Reasoning is Shaped by Recent Learning Experience 866
Paul Thibodeau, David Myers, Stephen Flusberg

Curiosity and Its Influence on Children's Memory 872
Haley Walin, Shaun O'Grady, Fei Xu

The Effects of Discourse Cues on Garden Path Processing 877
Ana Besserman, Elsi Kaiser

Children's Use of Orthographic Cues in Language Processing 883
Takayo Sugimoto

Modeling language discrimination in infants using i-vector representations 889
M. Julia Carbajal, Radek Fér, Emmanuel Dupoux

Are Financial Advisors Money Doctors or Charlatans? Evidence on Trust, Advice, and Risk Taking in Delegated Asset Management 895
Qizhang Sun, Michael Gibbert, Thomas Hills, Eric Nowak

A Recurrent Network Approach to Modeling Linguistic Interaction 901
Rick Dale, Riccardo Fusaroli, Kristian Tylén, Joanna Rączaszek-Leonardi, Morten Christiansen

The Aging Lexicon: Differences in the Semantic Networks of Younger and Older Adults 907
Dirk U. Wulff, Thomas Hills, Margie Lachman, Rui Mata

Increasing preschoolers' awareness of lexical ignorance to encourage word-learning 913
Sofia Jimenez, Kaitlin Ryan, Megan Saylor

Collective search on rugged landscapes: A cross-environmental analysis 918
Daniel Barkoczi, Pantelis Pipergias Analytis, Charley Wu

Is it a nine, or a six? Prosocial and selective perspective taking in four-year-olds 924
Xuan Zhao, Bertram F. Malle, Hyowon Gweon

Can Monaural Auditory Displays Convey Directional Information to Users? 930
Takanori Komatsu, Seiji Yamada

Spatial Attention to Social Cues is not a Monolithic Process 936
Samuel Harding, Ty Boyer, Bennett Bertenthal

When High WMC Promotes Mental Set: A Model of the Water Jar Task 942
Erin Sovansky, Stellan Ohlsson

Thermodynamics and Cognition: Towards a Lawful Explanation of the Mind 948
Larry Moralez, Luis Favela

Benefits for Grounded Feedback over Correctness in a Fraction Addition Tutor 954
Eliane Wiese, Rony Patel, Kenneth Koedinger

Our morals really depends on our language: The foreign language effect within participants 960
Kuninori Nakamura

Learning and making novel predictions about others' preferences 966
Natalia Vélez, Yuan Chang Leong, Chelsey Pan, Jamil Zaki, Hyowon Gweon

Monitoring the Level of Attention by Posture Measurement and EEG 972
Ryohei Furutani, Yuki Seino, Taro Tezuka, Tetsuji Satoh

Referential choice in identification and route directions 978
Adriana Alexandra Baltaretu, Emiel Krahmer, Alfons Maes

Statistical learning creates novel object associations via transitive relations 984
Yu Luo, Jiaying Zhao

Information Search with Depleting and Non-Depleting Resources 990
Amber Bloomfield, J. Isaiah Harbison, Susan Campbell, Petra Bradley, Lelyn Saner

Conceptual Expansion During Divergent Thinking 996
Richard Hass

Evaluating Causal Hypotheses: The Curious Case of Correlated Cues 1002
Bob Rehder, Zachary Davis

Do Simple Probability Judgments Rely on Integer Approximation? 1008
Shaun O'Grady, Tom Griffiths, Fei Xu

Social Affordance Tracking over Time - A Sensorimotor Account of False-Belief Tasks 1014
Judith Bütepage, Hedvig Kjellström, Danica Kragic

The Illusion of Explanatory Depth in a Misunderstood Field: The IOED in Mental Disorders .. 1020
Andrew Zeveney, Jessecae Marsh

Influences of Speaker-Listener Similarity on Shadowing and Comprehension 1026
Lynn Perry, Emily Mech, Maryellen MacDonald, Mark Seidenberg

Hidden Markov Modeling of eye movements with image information leads to better discovery of regions of interest 1032
Stephan Brueggemann, Antoni B. Chan, Janet Hsiao

Structural Alignment in Incidental Word Learning 1038
Ruxue Shao, Dedre Gentner

Reflexive Spatial Attention to Goal Directed Reaching 1044
Alexis Barton, Bennett Bertenthal, Samuel Harding

Longitudinal L2 Development of the English Article in Individual Learners 1050
Akira Murakami, Theodora Alexopoulou

Spatializing emotion: A mapping of valence or magnitude? 1056
Benjamin Pitt, Daniel Casasanto

Essays about service-learning events can be mined for program assessment 1062
Anne Gilman, Deborah Roney, Victoria Rehr, Helen Hu, Mark Peterson

Language Informativity: Is starfish more of a fish in English than in Dutch? 1068
Farah M. Djalal, Wouter Voorspoels, Tom Heyman, Gert Storms

A Developmental Shift in the Relationship Between Sequential Learning, Executive Function, and Language Ability as Revealed by Event-Related Potentials 1074
Joanne Deocampo, Christopher Conway

Ambiguity and Representational Stability: What is the role of embodied experiences? 1080
Ramon D. Castillo, Talia Waltzer, Heidi Kloos

Questions in informal teaching: A study of mother-child conversations 1086
Yue Yu, Elizabeth Bonawitz, Patrick Shafto

Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech 1092
Fatemeh Torabi Asr, Jon Willits, Michael Jones

Geometric representations of evidence in models of decision-making 1098
Peter Kvam

Centering and the meaning of conditionals 1104
Nicole Cruz, David Over, Mike Oaksford, Jean Baratgin

A rational speech-act model of projective content 1110
Ciyang Qing, Noah Goodman, Daniel Lassiter

Infants' speech and gesture production in Mozambique and the Netherlands 1116
Chiara De Jong, Paul Vogt

A Neural Model of Context Dependent Decision Making in the Prefrontal Cortex 1122
Sugandha Sharma, Brent Komer, Terrence Stewart, Chris Eliasmith

Discourse Analysis as a Solution to Interpretive Problems in Cognitive Development Research . 1128
Patrick Byers

Concept Membership vs Typicality in Sentence Verification Tasks 1134
Francesca Zarl, Danilo Fum

Better safe than sorry: Risky function exploitation through safe optimization 1140
Eric Schulz, Quentin Huys, Dominik Bach, Maarten Speekenbrink, Andreas Krause

Intentionality and the Role of Labels in Categorization 1146
Felix Gervits, Megan Johanson, Anna Papafragou

Constraining the Search Space in Cross-Situational Word Learning: Different Models Make Different Predictions 1152
Giovanni Cassani, Robert Grimm, Steven Gillis, Walter Daelemans

Feature-based Joint Planning and Norm Learning in Collaborative Games 1158
Mark Ho, James MacGlashan, Amy Greenwald, Michael Littman, Elizabeth Hilliard, Carl Trimbach, Stephen Brawner, Josh Tenenbaum, Max Kleiman-Weiner, Joseph Austerweil

Metaphor & Emotion: Metaphorical Frames for Coping with Hardship 1164
Rose Hendricks, Lera Boroditsky

Trust, Communication, and Inequality 1170
Joanna Bryson, Paul Rauwolf

What Were They Thinking? Diagnostic Coding of Conceptual Errors in a Mathematics Learning Software Data Archive 1176
Christine Massey, Jennifer Kregor, Laura Cosgrove, Himchan Lee

Generalisable patterns of gesture distinguish semantic categories in communication without language 1182
Gerardo Ortega, Asli Özyürek

Singular Interpretations Linger During the Processing of Plural Noun Phrases 1188
Nikole Patson

Using determiners as contextual cues in sentence comprehension: A comparison between younger and older adults 1193
Nazbanou Nozari, Daniel Mirman

Scarcity captures attention and induces neglect: Eyetracking and behavioral evidence 1199
Brandon Tomm, Jiaying Zhao

The Influence of Religious Beliefs on False Memory of Fabricated Events 1205
Ellen Searle, Jennifer Vonk, Brock Brothers

Surprising blindness to conversational incoherence in both instant-messaging and face-to-face speech 1211
Gareth Roberts, Benjamin Langstein, Bruno Galantucci

Analytical Thinking Predicts Less Teleological Reasoning and Religious Belief 1217
Jeffrey Zemla, Samantha Steiner, Steven Sloman

Gender Differences in the Effect of Impatience on Men and Women's Timing Decisions 1223
Moojan Ghafurian, David Reitter

Why Sense-Making through Magnitude May Be Harder for Fractions than for Whole Numbers . 1229
Eliane Wiese, Rony Patel, Kenneth Koedinger

The Primary and Convergent Retrieval Model of Recall 1235
William J. Hopper, David E. Huber

Modelling the co-development of word learning and perspective-taking 1241
Marieke Woensdregt, Simon Kirby, Chris Cummins, Kenny Smith

Salience versus prior knowledge - how do children learn rules? 1247
Samuel Rivera, Vladimir Sloutsky

Creative Interaction with Blocks and Robots 1253
Daniel Smithwick, Larry Sass, David Kirsh

A Hierarchical Probabilistic Language-of-Thought Model of Human Visual Concept Learning ... 1259
Matthew Overlan, Robert Jacobs, Steven Piantadosi

Integrating physical reasoning and visual object recognition for fully occluded scene interpretation 1265
Ilker Yildirim, Max Siegel, Josh Tenenbaum

"This problem has no solution": when closing one of two doors results in failure to access any. 1271
Hippolyte Gros, Emmanuel Sander, Jean-Pierre Thibaut

Analogical Generalization and Retrieval for Denominal Verb Interpretation 1277
Clifton McFate, Ken Forbus

Mindfulness meditation as attention control training: A dual-blind investigation 1283
Alexa Romberg, Hank Haarmann

Which Learning Algorithms Can Generalize Identity-Based Rules to Novel Inputs? 1289
Paul Tupper, Bobak Shahriari

Pragmatic relativity: Gender and context affect the use of personal pronouns in discourse differentially across languages 1295
Zeynep Azar, Ad Backus, Asli Ozyurek

Investigating Rational Analogy in the Spirit of John Stuart Mill: Bayesian Analysis of Confidence about Inferences across Aligned Simple Systems 1301
Bradley Rogers, David Landy

Translating testimonial claims into evidence for category-based induction 1307
Tamar Kushnir, Susan Gelman

Simple Search Algorithms on Semantic Networks Learned from Language Use 1313
Aida Nematzadeh, Filip Miscevic, Suzanne Stevenson

Racial Essentialism is Associated With Prejudice Towards Blacks in 5- and 6-Year-Old White Children 1319
Tara Mandalaywala, Marjorie Rhodes

The Combinatorial Power of Experience 1325
Brendan Johns, Randall Jamieson, Matthew Crump, Michael Jones, Douglas Mewhort

Discriminability of sound contrasts in the face of speaker variation quantified 1331
Christina Bergmann, Alejandrina Cristia, Emmanuel Dupoux

A Twist On Event Processing: Reorganizing Attention To Cope With Novelty In Dynamic Activity Sequences 1337
Jessica Kosie, Dare Baldwin

Recursion in Nicaraguan Sign Language 1343
Annemarie Kocab, Ann Senghas, Jesse Snedeker

Comparing competing views of analogy making using eye-tracking technology 1349
Jean-Pierre Thibaut, Bob French, Yannick Glady

Memory for the Random: A Simulation of Computer Program Recall 1355
Fernand Gobet, Oliver Iain

Feature Overlap in Action Sequence 1361
Alexandra Stubblefield, Lisa Fournier

Modeling adaptation to a novel accent 1367
Kasia Hitczenko, Naomi Feldman

Gibson's Reasons for Realism and Gibsonian Reasons for Anti-Realism: An Ecological Approach to Model-Based Reasoning in Science 1373
Guilherme Sanches de Oliveira

Measuring the Causal Dynamics of Facial Interaction with Convergent Cross Mapping 1379
Eric Postma, Marie Postma-Nilsenová

Linguistic niches emerge from pressures at multiple timescales 1385
Molly Lewis, Michael C. Frank

Know Your Adversary: Insights for a Better Adversarial Behavioral Model 1391
Yasaman Abbasi, Debarun kar, Nicole Sintov, Milind Tambe, Noam Ben-Asher, Don Morrison, Cleotilde Gonzalez

Semantic, Lexical, and Geographic Cues in Recall Processes 1397
Janelle Szary, Michael N. Jones

Sentire Decision Making in a Mixed-Motive Game 1403
Joana Campos, Ana Paiva

Modeling the Contribution of Central Versus Peripheral Vision in Scene, Object, and Face Recognition 1409
Panqu Wang, Garrison Cottrell

The Charon Model of Moral Judgment 1415
Deirdre Kelly, Jim Davies

Investigating Semantic Conflict between General Knowledge and Novel Information in Real-Time Sentence Processing 1421
Angele Yazbec, Michael Kaschak, Arielle Borovsky

Extracting Human Face Similarity Judgments: Pairs or Triplets? 1427
Linjie Li, Vicente Malave, Amanda Song, Angela Yu

Decision contamination in the wild: Sequential dependencies in Yelp review ratings 1433
David Vinson, Rick Dale, Michael Jones

Using Motor Dynamics to Explore Real-time Competition in Cross-situational Word Learning: Evidence From Two Novel Paradigms 1439
John Bunce, Drew Abney, Chelsea Gordon, Michael Spivey, Rose Scott

Modeling N400 amplitude using vector space models of word representation 1445
Allyson Ettinger, Naomi Feldman, Philip Resnik, Colin Phillips

Grammatical gender affects odor cognition 1451
Laura Speed, Asifa Majid

How Different Frames of Reference Interact: A Neural Network Model 1457
Weizhi Nan, Yanlong Sun, Xun Liu, Hongbin Wang

Syntax Accommodation in Social Media Conversations 1463
Reihane Boghrati, Joe Hoover, Kate M. Johnson, Justin Garten, Morteza Dehghani

Developmental Differences in Children's Statistical Learning Abilities 1469
Limor Raviv, Inbal Arnon

Under Pressure: How Time-Limited Cognition Explains Statistical Learning by 8-Month Old Infants 1475
Martyn Lloyd-Kelly, Fernand Gobet, Peter Lane

Exploring the Neural Mechanisms Supporting Structured Sequence Processing and Language Using Event-Related Potentials: Some Preliminary Findings 1481
Gretchen Smith, Gerardo Valdez, Anne Walk, John Purdy, Christopher Conway

Bifurcation analysis of a Gradient Symbolic Computation model of incremental processing 1487
Pyeong Whan Cho, Paul Smolensky

Coalescing the Vapors of Human Experience into a Viable and Meaningful Comprehension 1493
Tomer Ullman, Max Siegel, Josh Tenenbaum, Samuel Gershman

Music Reading Expertise Modulates Visual Spans in both Music Note and English Letter Reading 1499
Tze Kwan Li, Susana T. L. Chung, Janet H. Hsiao

Design from Zeroth Principles 1505
Jordan W. Suchow, Michael D. Pacer, Thomas L. Griffiths

Should moral decisions be different for human and artificial cognitive agents? 1511
Evgeniya Hristova, Maurice Grinberg

Attentive and Pre-Attentive Processes in Multiple Object Tracking: A Computational Investigation 1517
Paul Bello, Will Bridewell, Christina Wasylyshyn

How do Distributions of Item Sizes Affect the Precision and Bias in Representing Summary Statistics? 1523
Midori Tokita, Akira Ishiguchi

Who should I tell? Young children correct and maintain others' beliefs about the self 1529
Mika Asaba, Hyowon Gweon

Extended Metaphors are Very Persuasive 1535
Paul Thibodeau, Peace Iyiewuare, Matias Berretta

How Does Generic Language Elicit Essentialist Beliefs? 1541
Emily Foster-Hanson, Sarah-Jane Leslie, Marjorie Rhodes

Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures 1547
Sebastian Musslick, Biswadip Dey, Kayhan Ozcimder, Mostofa Patwary, Theodore L. Willke, Jonathan D. Cohen

Fractal Scaling and Implicit Bias: A Conceptual Replication of Correll (2008) 1553
Mary Jean Amon, John G. Holden

The Relationship Between the Numerical Distance Effect and Approximate Number System Acuity is Non-Linear 1559
Dana Chesney

Strategic search in semantic memory 1565
Evgenii Nikitin, Thomas Hills

Context, but not proficiency, moderates the effects of metaphor framing: A case study in India . 1571
Paul Thibodeau, Daye Lee, Stephen Flusberg

The Determinants of Knowability 1577
Samuel Johnson, Kristen Kim, Frank Keil

A scaleable spiking neural model of action planning 1583
Peter Blouw, Chris Eliasmith, Bryan Tripp

Systems Factorial Analysis of Item and Associative Retrieval 1589
Gregory Cox, Amy Criss

Children learn non-exact number word meanings first 1595
Junyi Chu, Katie Wagner, David Barner

A Dream Model: Reactivation and Re-encoding Mechanisms for Sleep-dependent Memory Consolidation 1601
George Kachergis, Roy de Kleijn, Bernhard Hommel

Solving the knowledge-behavior gap: Numerical cognition explains age-related changes in fairness 1607
Nadia Chernyak, Beth Sandham, Paul Harris, Sara Cordes

The Pragmatics of Spatial Language 1613
Tomer Ullman, Yang Xu, Noah Goodman

Measuring Interest in Science: The Science Curiosity Scale 1619
Asheley Landrum, Joseph Hilgard, Heather Akin, Nan Li, Dan Kahan

Conflict-based regulation of control in language production 1625
Michael Freund, Barry Gordon, Nazbanou Nozari

Active Viewing in Toddlers Facilitates Visual Object Learning: An Egocentric Vision Approach 1631
Sven Bambach, David Crandall, Linda Smith, Chen Yu

Processing Consequences of Onomatopoeic Iconicity in Spoken Language Comprehension 1637
David Peeters

Language Evolution in the Lab: The Case of Child Learners 1643
Limor Raviv, Inbal Arnon

Outcome or Strategy? A Bayesian Model of Intelligence Attribution 1649
Marta Kryven, Tomer Ullman, William Cowan, Josh Tenenbaum

Communicating generalizations about events 1655
Michael Henry Tessler, Noah Goodman

Learning Non-Adjacent Dependencies in Continuous Presentation of an Artificial Language 1661
Hao Wang, Jason Zevin, Toben Mintz

Process Modeling of Qualitative Decision Under Uncertainty 1667
David Broniatowski, Valerie Reyna

A Nonlinear Dynamical Systems Theory Perspective on Dual-Processing Accounts of Decision-Making under Uncertainty 1673
Marieke M. J. W. van Rooij, Luis H. Favela

Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction ... 1679
Max Kleiman-Weiner, Mark Ho, Joseph Austerweil, Michael Littman, Josh Tenenbaum

Effects of Gesture on Analogical Problem Solving: When the Hands Lead you Astray 1685
Autumn Hostetter, Mareike Wieth, Katlyn Foster, Keith Moreno, Jeffery Washington

From uh-oh to tomorrow: Predicting age of acquisition for early words across languages 1691
Mika Braginsky, Daniel Yurovsky, Virginia Marchman, Mike Frank

Lexical Complexity of Child-Directed and Overheard Speech: Implications for Learning 1697
Ruthe Foushee, Tom Griffiths, Mahesh Srinivasan

Degeneracy results in canalisation of language structure: A computational model of word learning 1703
Padraic Monaghan

General Mechanisms Underlying Language and Spatial Cognitive Development 1709
Hilary Miller, Vanessa Simmering

When the Words Don't Matter: Arbitrary labels improve categorical alignment through the anchoring of categories 1715
Ellise Suffill, Martin Pickering, Holly Branigan

Can the High-Level Semantics of a Scene be Preserved in the Low-Level Visual Features of that Scene? A Study of Disorder and Naturalness 1721
Hiroki Kotabe, Omid Kardan, Marc Berman

Is the Self-Concept like Other Concepts? The Causal Structure of Identity 1727
Stephanie Chen, Daniel Bartels, Oleg Urminsky

The Effect of Emotion and Induced Arousal on Numerical Processing 1733
Karina Hamamouche, Michelle Hurst, Sara Cordes

Putting the "th" in Tenths: The Role of Labeling Decimals in Revealing Place Value Structure . 1739
Abbey Loehr, Bethany Rittle-Johnson

Implicit measurement of motivated causal attribution 1745
Laura Niemi, Joshua Hartshorne, Tobias Gerstenberg, Liane Young

Boredom, Information-Seeking and Exploration 1751
Andra Geana, Robert Wilson, Nathaniel Daw, Jonathan Cohen

Analyzing distributional learning of phonemic categories in unsupervised deep neural networks . 1757
Okko Räsänen, Tasha Nagamine, Nima Mesgarani

The Structure of Names in Memory: Deviations from Uniform Entropy Impair Memory for Linguistic Sequences 1763
Melody Dye, Brendan Johns, Michael Jones, Michael Ramscar

What do we learn from rating metaphors? 1769
Paul Thibodeau, Les Sikos, Frank Durgin

Listener sensitivity to foreign-accented speech with grammatical errors 1775
Yuki Asano, Andrea Weber

A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding 1781
Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T. Freeman, Joshua B. Tenenbaum

Learning Behavior-Grounded Event Segmentations 1787
Christian Gumbsch, Jan Kneissler, Martin Butz

Information-Seeking, Learning and the Marginal Value Theorem: A Normative Approach to Adaptive Exploration 1793
Andra Geana, Robert Wilson, Nathaniel Daw, Jonathan Cohen

Learning that numbers are the same, while learning that they are different 1799
Jon Willits, Michael Jones, David Landy

Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem 1805
James Kubricht, Chenfanfu Jiang, Yixin Zhu, Song-Chun Zhu, Demetri Terzopoulos, Hongjing Lu

College Students' Understanding of Linear Functions: Slope is Slippery 1811
Marta Mielicki, Jennifer Wiley

L2 Idiom Processing: Figurative Attunement in Highly Idiomatic Contexts 1817
Sara Beck, Andrea Weber

Synthesized size-sound sound-symbolism 1823
Gwilym Lockwood, Peter Hagoort, Mark Dingemanse

Examining the specificity of the seductive allure effect 1829
Emily Hopkins, Deena Weisberg, Jordan Taylor

More than Words: The Many Ways Extended Discourse Facilitates Word Learning 1835
Sumarga Suanda, Linda Smith, Chen Yu

Two Potential Mechanisms Underlying the Link between Approximate Number Representations and Symbolic Math in Preschool Children 1841
Ariel Starr, Rachel Roberts, Elizabeth Brannon

Construal level affects intuitive moral responses to narrative content 1847
Nicholas Lester, René Weber

Beyond Markov: Accounting for Independence Violations in Causal Reasoning 1853
Bob Rehder

Curiosity-Driven Development of Tool Use Precursors: a Computational Model 1859
Sébastien Forestier, Pierre-Yves Oudeyer

Spatial Memory and Foraging: How Perfect Spatial Memory Improves Foraging Performance .. 1865
Bryan Kerster, Chris Kello

Temporal Expressions in Speech and Gesture 1871
İdil Bostan, Ahmet Börütecene, Oğuzhan Özcan, Tilbe Goksun

Children's Awareness of Authority to Change Rules in Various Social Contexts 1877
Xin Zhao, Tamar Kushnir

An Ecological Model of Memory and Inferences 1883
Daniela Link, Julian N. Marewski, Lael J. Schooler

Conversational expectations account for apparent limits on theory of mind use 1889
Robert X.d. Hawkins, Noah Goodman

Simpler structure for more informative words: a longitudinal study 1895
Uriel Cohen Priva, Emily Gleason

The Influence of Language-specific Auditory Cues on the Learnability of Center-embedded Recursion 1901
Jun Lai, Chiara de Jong, Dingguo Gao, Ren Huang, Emiel Krahmer, Jan Sprenger

U-INVITE: Estimating Individual Semantic Networks from Fluency Data 1907
Jeffrey Zemla, Yoed Kenett, Kwang-Sung Jun, Joseph Austerweil

Contextual Events and Their Role in a Two-Choice Joint Simon Task 1913
Steve Croker, J. Scott Jordan, Daniel Schloesser, Vincent Cialdella, Alex Dayer

A Framework for Evaluating Speech Representations 1919
Caitlin Richter, Naomi Feldman, Harini Salgado, Aren Jansen

Facilitating Spatial Task Learning in Interactive Multimedia Environments While Accounting for Individual Differences and Task Difficulty 1925
Dar-Wei Chen, Richard Catrambone

A Neural Dynamic Model Parses Object-Oriented Actions 1931
Mathis Richter, Jonas Lins, Gregor Schoener

Vector Space Semantic Models Predict Subjective Probability Judgments for Real-World Events 1937
Sudeep Bhatia

Noisy Parameters in Risky Choice 1943
Sudeep Bhatia, Graham Loomes

Do Do Do, The The The: Interactivity and Articulatory Suppression in Mental Arithmetic 1949
Frederic Vallee-Tourangeau, Miroslav Sirota, Gaelle Vallee-Tourangeau

Between versus Within-Language Differences in Linguistic Categorization 1955
Anne White, Gert Storms, Barbara C. Malt, Steven Verheyen

Event participants and linguistic arguments 1961
Roxana-Maria Barbu, Ida Toivonen

Decision-Making and Biases in Causal-Explanatory Reasoning 1967
Samuel Johnson, Marianna Zhang, Frank Keil

Does Chess Instruction Enhance Mathematical Ability in Children? A Three-Group Design to Control for Placebo Effects 1973
Giovanni Sala, Fernand Gobet, Roberto Trinchero, Salvatore Ventura

Are Symptom Clusters Explanatory? A Study in Mental Disorders and Non-Causal Explanation 1979
Daniel Wilkenfeld, Jennifer Asselin, Tania Lombrozo

Not all overlaps are equal: Social affiliation and rare overlaps of preferences 1985
Natalia Vélez, Sophie Bridgers, Hyowon Gweon

What Causal Illusions Might Tell us About the Identification of Causes 1991
Robert Thorstad, Phillip Wolff

Quantifying Joint Activities using Cross-Recurrence Block Representation 1997
Tian Xu, Chen Yu

Temporal Horizons and Decision-Making: A Big Data Approach 2003
Robert Thorstad, Phillip Wolff

Using Subgoal Learning and Self-Explanation to Improve Programming Education 2009
Lauren Margulieux, Richard Catrambone

The Effects of Gender Stereotypes for Structure Mapping in Mathematics 2015
Kreshnik Begolli, Brooke Herd, Hannah Sayonno, Susanne Jaeggi, Lindsey Richland

Improving with Practice: A Neural Model of Mathematical Development 2021
Sean Aubin, Aaron Voelker, Chris Eliasmith

Examining Referential Uncertainty in Naturalistic Contexts from the Child's View: Evidence from an Eye-Tracking Study with Infants 2027
Yayun Zhang, Chen Yu

The Naïve Utility Calculus unifies spatial and statistical routes to preference 2033
Julian Jara-Ettinger, Felix Sun, Laura Schulz, Josh Tenenbaum

Task-set Selection in Probabilistic Environments: a Model of Task-set Inference 2039
Ian W. Eisenberg, Russell A. Poldrack

N400 amplitudes reflect change in a probabilistic representation of meaning: Evidence from a connectionist model 2045
Milena Rabovsky, Steven S. Hansen, James L. McClelland

Asking and evaluating natural language questions 2051
Anselm Rothe, Brenden Lake, Todd Gureckis

Sequential images are not universal, or Caveats for using visual narratives in experimental tasks 2057
Neil Cohn

Deciding to Remember: Memory Maintenance as a Markov Decision Process 2063
Jordan W. Suchow, Tom Griffiths

When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication 2069
Rony Patel, Ran Liu, Kenneth Koedinger

Helping people make better decisions using optimal gamification 2075
Falk Lieder, Tom Griffiths

Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (Bayesian) learning and functional selection 2081
Thomas Brochhagen, Michael Franke, Robert van Rooij

A Computational Model of Perceptual Deficits in Medial Temporal Lobe Amnesia 2087
Patrick Sadil, Rosemary A. Cowell

Linguistic input is tuned to children's developmental level 2093
Daniel Yurovsky, Gabriel Doyle, Michael Frank

The Plausible Impossible: Causal Constraints on Magical Reasoning 2099
Andrew Shtulman, Caitlin Morgan

Statistical learning bias predicts second-language reading efficiency 2105
Luca Onnis, Stefan Frank, Hongoak Yun, Matthew Lou-Magnuson

The Permeability of Fictional Worlds 2111
Meghan Salomon, Lance Rips

The development of heuristics in children: Base-rate neglect and representativeness 2117
Samantha Gualtieri, Stephanie Denison

Combining Multiple Perspectives in Language Production: A Probabilistic Model 2123
Mindaugas Mozuraitis, Suzanne Stevenson, Daphna Heller

Statistical Learning Ability Can Overcome the Negative Impact of Low Socioeconomic Status on Language Development 2129
Leyla Eghbalzad, Joanne Deocampo, Christopher Conway

The Influence of Group Interaction on Creativity in Engineering Design 2135
Trina Kershaw, Rebecca Peterson, Sankha Bhowmick

There is more to gesture than meets the eye: Visual attention to gesture's referents cannot account for its facilitative effects during math instruction 2141
Miriam Novack, Elizabeth Wakefield, Eliza Congdon, Steve Franconeri, Susan Goldin-Meadow

Measuring and modeling distraction by self-referential processing in a complex working memory span task 2147
Jeroen Daamen, Marieke van Vugt, Niels Taatgen

Shakers and Maracas: Action-Based Categorisation Choices in Triads Are Influenced by Task Instructions 2153
Nicholas Shipp, Frederic Vallee-Tourangeau, Susan Anthony

Training prospective abilities through conversation about the extended self 2159
Nadia Chernyak, Kathryn Leech, Meredith Rowe

How people differ in syllogistic reasoning 2165
Sangeet Khemlani, P.n. Johnson-Laird

Grounded Distributional Semantics for Abstract Words 2171
Katsumi Takano, Akira Utsumi

Active Overhearing: Development in Preschoolers' Skill at 'Listening in' to Naturalistic Overheard Speech 2177
Ruthe Foushee, Fei Xu

Towards a Cognitively Realistic Representation of Word Associations 2183
Ivana Kajic, Jan Gosmann, Terrence Stewart, Thomas Wennekers, Chris Eliasmith

Don't Blink! Evaluating Training Paradigms for Overcoming the Attentional Blink 2189
Trudy Buwalda, Jelmer Borst, Marieke van Vugt, Niels Taatgen

Allocation of attention during auditory word learning 2195
Keith Apfelbaum, Vladimir Sloutsky

Examining Cardiac and Behavioral Responses in a Modality Dominance Task 2201
Chris Robinson, Krysten Chadwick, Jessica Parker, Scott Sinnett

From Words to Behaviour via Semantic Networks 2207
Armand S. Rotaru, Gabriella Vigliocco, Stefan L. Frank

Evolution of polysemous word senses from metaphorical mappings 2213
Yang Xu, Barbara Malt, Mahesh Srinivasan

Using Violations of Fitts' Law to Communicate during Joint Action 2219
Cordula Vesper, Laura Schmitz, Günther Knoblich

Perceived Momentum Influences Responsibility Judgments 2225
Jeffrey Parker, Iman Paul, Nicholas Reinholtz

A computational investigation of the Sapir-Whorf hypothesis: The case of spatial relations 2231
Christine Tseng, Alexandra Carstensen, Terry Regier, Yang Xu

Effects of Auditory Input on a Spatial Serial Response Time Task 2237
Chris Robinson, Jessica Parker

Memory for exemplars in category learning 2243
Charlotte Edmunds, Andy Wills, Fraser Milton

Chinese and English speakers' neural representations of word meaning offer a different picture of cross-language semantics than corpus and behavioral measures 2249
Benjamin Zinszer, Andrew Anderson, Rajeev Raizada

Modeling developmental and linguistic relativity effects in color term acquisition 2255
Barend Beekhuizen, Suzanne Stevenson

Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions 2261
Caroline Graf, Judith Degen, Robert Hawkins, Noah Goodman

Do classifier categories affect or reflect object concepts? 2267
Laura Speed, Jidong Chen, Falk Huettig, Asifa Majid

Exploring the Cost Function in Color Perception and Memory: An Information-Theoretic Model of Categorical Effects in Color Matching 2273
Chris R. Sims, Zheng Ma, Sarah R. Allred, Rachel A. Lerch, Jonathan I. Flombaum

Abstraction in time: Finding hierarchical linguistic structure in a model of relational processing 2279
Leonidas Doumas, Andrea Martin

Modeling sampling duration in decisions from experience 2285
Nisheeth Srivastava, Johannes Mueller-Trede, Paul Schrater, Edward Vul

Experience as a Free Parameter in the Cognitive Modeling of Language 2291
Brendan Johns, Michael Jones, Douglas Mewhort

An Information-Processing Account of Representation Change: International Mathematical Olympiad Problems are Hard not only for Humans 2297
Takuya Matsuzaki, Munehiro Kobayashi, Noriko H. Arai

Inductive Ethics: A Bottom-Up Taxonomy of the Moral Domain 2303
Justin Landy, Daniel Bartels

The Interaction of Memory and Attention in Novel Word Generalization: A Computational Investigation 2309
Erin Grant, Aida Nematzadeh, Suzanne Stevenson

Environmental Orientation Affects Emotional Expression Identification 2315
Stephen Flusberg, Derek Shapiro, Kevin Collister, Paul Thibodeau

Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so 2321
Kathryn D. Schuler, Charles Yang, Elissa L. Newport

Variability in category learning: The Effect of Context Change and Item Variation on Knowledge Generalization 2327
Dustin Finch, Paulo Carvalho, Robert Goldstone

Performance Pressure and Comparison in Relational Category Learning 2333
John Patterson, Kenneth Kurtz

Intermediate Judgments Inhibit Belief Updating: Zeno's Paradox in Decision Making 2339
James Yearsley, Emmanuel Pothos

Working memory encoding of events and their participants: a neural network model with applications in sensorimotor processing and sentence generation 2345
Martin Takac, Alistair Knott

What do you expect from an unfamiliar talker? 2351
Dave Kleinschmidt, T. Florian Jaeger

Working Memory Affects Attention to Loss Value and Loss Frequency in Decision-Making under Uncertainty 2357
Bo Pang, Kaileigh Byrne, Darrell Worthy

Adapting Deep Network Features to Capture Psychological Representations 2363
Joshua Peterson, Joshua Abbott, Thomas Griffiths

Neurophysiological Effects of Negotiation Framing 2369
Peter Khooshabeh, Rebecca Lin, Celso de Melo, Jonathan Gratch, Brett Ouimette, Jim Blascovich

Generalization of within-category feature correlations 2375
Nolan Conaway, Kenneth Kurtz

Deconstructing "tomorrow": How children learn the semantics of time 2381
Katharine Tillman, Tyler Marghetis, David Barner, Mahesh Srinivasan

Spatial Interference and Individual Differences in Looking at Nothing for Verbal Memory 2387
Alper Kumcu, Robin L. Thompson

Bayesian Pronoun Interpretation in Mandarin Chinese 2393
Meilin Zhan, Roger P. Levy, Roger P. Levy, Andrew Kehler

Leveling the Field: Talking Levels in Cognitive Science 2399
Luke Kersten, Robert West, Andrew Brook

Know Your Enemy: Applying Cognitive Modeling in Security Domain 2405
Vladislav Daniel Veksler, Norbou Buchler

Probability Prediction in Children with ASD 2411
Zi L. Sim, Fei Xu

Left-right mental timeline is robust to visuospatial and verbal interference 2417
Rose Hendricks, Esther Walker, Benjamin Bergen, Lera Boroditsky, Rafael Nunez

A speed-accuracy trade-off in children's processing of scalar implicatures 2423
Rose Schneider, Michael Frank

Systematic feature variation underlies adults' and children's use of in and on 2429
Kristen Johannes, Colin Wilson, Barbara Landau

Modeling Triage Decision Making 2435
J. Isaiah Harbison, Alan Mishler, Thomas Wallsten

The mismeasurement of mind: How neuropsychological testing creates a false picture of cognitive aging 2441
Michael Ramscar, Ching Chu Sun, Peter Hendrix, Harald Baayen

When are representations of causal events quantum versus classical? 2447
James Yearsley, Jennifer Trueblood, Emmanuel Pothos

A model of conditional probability judgment 2453
Fintan Costello, Paul Watts

When does passive learning improve the effectiveness of active learning? 2459
Kyle MacDonald, Mike Frank

Balancing Structural and Temporal Constraints in Multitasking Contexts 2465
Dario Salvucci, Tuomo Kujala

Do additional features help or harm during category learning? An exploration of the curse of dimensionality in human learners 2471
Wai Keen Vong, Andrew Hendrickson, Amy Perfors, Daniel Navarro

Desirable difficulties in the development of active inquiry skills 2477
George Kachergis, Marjorie Rhodes, Todd Gureckis

Linguistic Signatures of Cognitive Processes during Writing 2483
Laura Allen, Cecile Perret, Danielle McNamara

Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for Sentence Prominence 2489
Sofoklis Kakouros, Okko Räsänen

Integrating identification and perception: A case study of familiar and unfamiliar face processing 2495
Kelsey Allen, Ilker Yildirim, Josh Tenenbaum

On the Link between Fact Learning and General Cognitive Ability 2501
Florian Sense, Rob R. Meijer, Hedderik van Rijn

The paradox of relational development: Could language learning be (temporarily) harmful? 2507
Christian Hoyos, Ruxue Shao, Dedre Gentner

Linguistic Priming and Learning Adjacent and Non-Adjacent Dependencies in Serial Reaction Time Tasks 2513
Hao Wang, Elsi Kaiser

Modeling the Visual Word Form Area Using a Deep Convolutional Neural Network 2519
Sandy Wiraatmadja, Garrison Cottrell

Think Fast! Mental-state Language is Related to the Speed of False-belief Reasoning in Adulthood 2525
Erin Roby, Rose Scott

Simple trees in complex forests: Growing Take The Best by Approximate Bayesian Computation 2531
Eric Schulz, Maarten Speekenbrink, Björn Meder

Extraction of Event Roles From Visual Scenes is Rapid, Automatic, and Interacts with Higher-Level Visual Processing 2537
Alon Hafri, John Trueswell, Brent Strickland

A 3D shape inference model matches human visual object similarity judgments better than deep convolutional neural networks 2543
Goker Erdogan, Robert A. Jacobs

Cognitive Strategies in HCI and Their Implications on User Error 2549
Marc Halbrügge, Michael Quade, Klaus-Peter Engelbrecht

Connectionist Semantic Systematicity in Language Production 2555
Jesus Calvillo, Harm Brouwer, Matthew Crocker

Grammatical Bracketing Determines Learning of Non-adjacent Dependencies 2561
Hao Wang, Jason Zevin, Toben Mintz

Natural science: Active learning in dynamic physical microworlds 2567
Neil Bramley, Tobias Gerstenberg, Joshua Tenenbaum

Inattentional Blindness in a Coupled Perceptual–Cognitive System 2573
Will Bridewell, Paul Bello

Recursive belief manipulation and second-order false-beliefs 2579
Torben Braüner, Patrick Blackburn, Irina Polyanskaya

Syntactic Flexibility in the Noun: Evidence from Picture Naming 2585
Nicholas Lester, Fermin Moscoso del Prado Martin

Alien species and alienable traits: An artificial language game investigating the spread of cultural variants between antagonistic groups 2591
Betsy Sneller, Gareth Roberts

Inferring priors in compositional cognitive models 2597
Eric Bigelow, Steven Piantadosi

Which is in front of Chinese people: Past or Future? A study on Chinese people's space-time mapping 2603
Yan Gu, Yeqiu Zheng, Marc Swerts

A performance model for early word learning 2609
Michael Frank, Molly Lewis, Kyle MacDonald

Cognitive biases and social coordination in the emergence of temporal language 2615
Tessa Verhoef, Esther Walker, Tyler Marghetis

Transfer of Cognitive Skills in Developmental Tasks 2621
Sarah Rupp, Niels Taatgen

The Description-Experience Gap in Risky Choice Framing 2627
Gaëlle Vallée-Tourangeau, Frédéric Vallée-Tourangeau, Madhuri Ramasubramanian

First things first? Top-down influences on event apprehension 2633
Johannes Gerwien, Monique Flecken

Definitely maybe and possibly even probably: efficient communication of higher-order uncertainty 2639
Michele Herbstritt, Michael Franke

Dynamical systems modeling of the child–mother dyad: Causality between child-directed language complexity and language development 2645
Jeremy Irvin, Daniel Spokoyny, Fermin Moscoso del Prado Martin

The Relational SNARC: Spatial Representation of Nonsymbolic Ratios? 2651
Percival Matthews, Rui Meng, Elizabeth Toomarian, Edward Hubbard

Emotional influences on time perception 2657
Anna K. Trapp, Manfred Thüring

Stable Causal Relationships are Better Causal Relationships 2663
Nadya Vasilyeva, Thomas Blanchard, Tania Lombrozo

What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data 2669
Michael Franke, Fabian Dablander, Anthea Schoeller, Erin Bennett, Judith Degen, Michael Tessler, Justine Kao, Noah Goodman

Active control of study leads to improved episodic memory in children 2675
Azzurra Ruggeri, Douglas B. Markant, Todd M. Gureckis, Fei Xu

Cohesive Features of Deep Text Comprehension Processes 2681
Laura Allen, Matthew Jacovina, Danielle McNamara

Specificity at the basic level in event taxonomies: The case of Maniq verbs of ingestion 2687
Ewelina Wnuk

Switch it up: Learning Categories via Feature Switching 2693
Garrett Honke, Nolan Conaway, Kenneth Kurtz

Episodic memory as a prerequisite for online updates of model structure 2699
David Gergely Nagy, Gergo Orban

Stop paying attention: the need for explicit stopping in inhibitory control 2705
Ning Ma, Angela Yu

A cross-linguistic investigation on the acquisition of complex numerals 2711
Pierina Cheung, Meghan Dale, Mathieu Le Corre

tDCS to Premotor Cortex Changes Action Verb Understanding: Complementary Effects of Inhibitory and Excitatory Stimulation 2717
Tom Gijssels, Daniel Casasanto

From low to high cognition: A multi-level model of behavioral control in the primate brain 2723
Jerald Kralik, Dongqing Shi, Omar El-Shroa, Laura Ray

Learning in the wild - how labels influence what we learn 2729
Samuel Rivera, Chris Robinson

Cross-linguistic similarities aid third language learning in bilinguals 2735
James Bartolotti, Viorica Marian

Varieties of experience: A new look at folk philosophy of mind 2741
Kara Weisman, Carol Dweck, Ellen Markman

Benefiting from Being Alike: Interindividual Skill Differences Predict Collective Benefit in Joint Object Control 2747
Basil Wahn, Laura Schmitz, Peter König, Günther Knoblich

Reading experience shapes the mental timeline but not the mental number line 2753
Benjamin Pitt, Daniel Casasanto

Cultural Evolution Across Domains: Language, Technology and Art 2759
Monica Tamariz, Simon Kirby, Jon W. Carr

A Neural Field Model of Word Repetition Effects in Early Time-Course ERPs in Spoken Word Perception 2765
Andrew Valenti, Michael Brady, Matthias Scheutz, Phillip Holcomb, He Pu

Talking with tact: Polite language as a balance between informativity and kindness 2771
Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, Michael C. Frank

Understanding "almost": Empirical and computational studies of near misses 2777
Tobias Gerstenberg, Josh Tenenbaum

Tit-for-Tat: Effects of Feedback and Speaker Reliability on Listener Comprehension Effort 2783
Nicole Craycraft, Zoe Kriegel, Michael Tanenhaus, Jennifer Roche

Distinguishing processing difficulties in inhibition, implicature, and negation 2789
Ann Nordmeyer, Erica Yoon, Mike Frank

Answering Causal Queries about Singular Cases 2795
Simon Stephan, Michael R. Waldmann

The Impact of Granularity on the Effectiveness of Students' Pedagogical Decisions 2801
Guojing Zhou, Collin Lynch, Thomas Price, Tiffany Barnes, Min Chi

Can Musical Engagement Alleviate Age-Related Decline in Inhibitory Control? 2807
Ruben Vromans, Marie Postma-Nilsenova

A 6-month longitudinal study on numerical estimation in preschoolers 2813
Pierina Cheung, Emily Slusser, Anna Shusterman

The distorting effect of deciding to stop sampling 2819
Anna Coenen, Todd Gureckis

Inferring Individual Differences Between and Within Exemplar and Decision-Bound Models of Categorization 2825
Irina Danileiko, Michael Lee

Adults' guesses on probabilistic tasks reveal incremental representativeness biases 2831
Habiba Azab, David Ruskin, Celeste Kidd

Publication-Based Presentations

A Bayesian Metric for Network Similarity 2837
Richard Shiffrin

Cognition without behaviour: Cognitive functions in behaviourally non-responsive individuals ... 2839
Boris Kotchoubey

Visualizing Events: Simulating Meaning in Language 2841
James Pustejovsky, Nikhil Krishnaswamy

An Architectural Account of Variation in Problem Solving and Execution 2843
Pat Langley

Verbalizing navigation: Explicit and implicit concepts 2845
Thora Tenbrink

The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color .. 2847
Emily Cibelli, Yang Xu, Joseph Austerweil, Thomas Griffiths, Terry Regier

Member Abstracts

Do gestures serve an interpersonal function? 2848
Yiwei Chen, Robyn Enns, Elena Nicoladis

Emergence of Euclidean geometrical intuitions in hierarchical generative models 2849
Arianna Yuan, Te-Lin Wu, James L. McClelland

Implicit Emotional Priming of Traumatic Events: The Effects of Semantic Level and Emotional Activation 2850
Jae-Ho Lee, Yun-Kyeung Choi

Current Research in Cognitive Science at Educational Testing Service (ETS) 2851
Irvin Katz

Free-form response vs. yes/no-question methodologies in the study of human reasoning 2852
Salvador Mascarenhas, Philipp Koralus

Noticing causal properties of objects from sequence statistics 2853
Anna Leshinskaya, Sharon L Thompson-Schill

Filling in the gaps: Event segmentation is robust to missing information 2854
Jessica Kosie, Dare Baldwin

Effects of the Plausibility of the Retracting Information on Memory Updating 2855
Kyung Soo Do, Jin Ju Gu, Kwanghyeon Yoo

Which Statistic Matters? Effects of Category Size and Distribution on Statistical Category Learning 2856
Chi-hsin Chen, Paulo Carvalho, Chen Yu

Reciprocal altruism in preschool-aged children 2857
Nadia Chernyak, Yarrow Dunham, Peter Blake

Heuristics in exploration: Distributional information is selectively used for active learning 2858
Elizabeth Lapidow, Elizabeth Bonawitz

On the psychological reality of linguistic event structures 2859
Jayden Ziegler, Jesse Snedeker

Effects of Analogical Processing: Evidence for Re-representation 2860
Daniel Silliman, Kenneth Kurtz

A computational theory of temporal inference 2861
Sangeet Khemlani, Anthony Harrison, Greg Trafton

An interactive model accounts for both ultra-rapid superordinate classification and basic-level advantage in object recognition 2862
Qihong Lu, Timothy Rogers

Handedness and Mathematics: Toward a More Comprehensive Model 2863
Giovanni Sala, Martina Bolognese, Giulia Barsuola, Michela Signorelli, Fernand Gobet

Shifting meanings: The fluidity of signal-meaning mappings in a minimal communicative task . 2864
Jennifer Misyak, Takao Noguchi, Nick Chater

Process of visual input does not decide the accuracy imitation performance 2865
Takashi Mizuguchi

Reduced benefit from regularities in language among Dyslexics 2866
Eva Kimel, Merav Ahissar

How the Physicality of Space Affects How We Think about Time 2867
Jennifer Kolesari, Laura Carlson

Embodiment Effects in Evolutionary Robotics 2868
Nicholas Livingston, Anton Bernatskiy, Kenneth Livingston, Marc Smith, Jodi Schwarz, Joshua Bongard, David Wallach, Evan Altiero, John Long

Perception of math and non-math content in children's storybooks 2869
Colleen Uscianowski, Ma. Victoria Almeda, Cassiopeia Freeman, Herbert Ginsburg

Your Obstacle on My Mind: Task Co-representation in Coordination is Modulated by External Timing Cues 2870
Laura Schmitz, Cordula Vesper, Natalie Sebanz, Guenther Knoblich

Prediction of Single-Trial Behavior using a Layered Dynamic Systems Model with Evolutionary Algorithm Updating 2871
Richard Prather, Sara Heverly-Fitt

The role of viewpoint in event segmentation 2872
Khena Swallow, Ayse Candan, Jovan Kemp

Individual differences in verbalization predict change detection performance: A new perspective on the language-thought debate 2873
Junko Kanero, Eileen Kitrick, Kathy Hirsh-Pasek, Kevin Holmes

Do Subliminal Hints Facilitate Sequential Planning When Solving a Spatial Insight Problem? .. 2874
Hiromitsu Miyata, Hiroaki Suzuki

Electrophysiological markers indicate disturbance of involuntary attention, but not voluntary attention, in adult ADHD patients 2875
Frank Nonnenmacher, Boris Kotchoubey

Biased Attention to Spatial Dimensions Predicts Children's Spatial Word Acquisition 2876
Hilary Miller, Haley Vlach

On Measuring the Difficulty of Scrabble-like Problems 2877
Morgan Magnus Fleming, Paul P. Maglio

Metaphorical Color Representations of Emotional Concepts in English and Chinese Speakers ... 2878
Junqing Chen, Natalie Kacinik, Yingjun Chen, Nianyang Wu

Feature distinctiveness in verbs: links between verb distinctiveness, child directed speech and age of acquisition 2879
Tomas Engelthaler, Thomas Hills

The Effects of Grammatical Aspect and Visual Perspective on Imaginging Accomplishments 2880
Deanna Hall, Todd Ferretti, Jeffrey Hong

A Cognitive Model of Fraction Arithmetic 2881
David Braithwaite, Robert Siegler

The role of regional topography in route planning 2882
Andrew Deslauriers, Kevin Holmes

Sign languages reveal spatial mappings of valence and magnitude 2883
Eileen Kitrick, Kevin Holmes

Informant effort expenditure impacts young children's learning, eye gaze, and trust 2884
Molly Schlesinger, Rebekah Richert, John Franchak

Audio-Visual Task Switching in Multisensory Environments 2885
Christina Wasylyshyn

Causal Representation in Foresight: Can We Improve Memory and Novel Understanding? 2886
Edward Munnich, Emma Weinberger, Ana Maria Hoffmann

From computation to automization: How practice alters initial neural response to familiar arithmetic problems 2887
Caitlin Tenison, John Anderson

Influence of Need for Cognition and Cognitive Closure on Magic Perceptions 2888
Lisa Grimm, Nicholas Spanola

Asymmetric derivational priming in recognition of Greek nouns and verbs 2889
Sofia Loui, Athanassios Protopapas

Linear separability and human category learning: Revisiting a classic study 2890
Kimery Levering, Nolan Conaway, Kenneth Kurtz

Concept of a Deity: Structure and Properties 2891
Nicholas Shaman, Anondah Saide, Rebekah Richert

Do peer interactions influence infant communication development? 2892
Anne Warlaumont, Gina Pretzer

How event endstates are conceptualized in adults and infants 2893
Angela Xiaoxue He, Sudha Arunachalam

ADHD modulates link between event processing and recall 2894
Robbie Ross, Leah Child, Dare Baldwin

The sequence of study changes what is encoded during category learning. 2895
Paulo Carvalho, Robert Goldstone

Finding Clarity Amidst the Clutter: How Parents Name Objects 2896
Charlene Tay, Linda B. Smith, Chen Yu

Relational discovery in category learning: interactions of learning strategy and task structure ... 2897
Micah Goldwater, Hilary Don, Evan Livesey

The effect of loss and gain expression in the riddle on insight problem solving 2898
Reiko Yakushijin, Katsutochi Endo

Beyond the 64 Squares: Does Chess Instruction Enhance Children's Academic and Cognitive Skills? A Meta-Analysis 2899
Giovanni Sala, Fernand Gobet

Memory biases in matching and recall: Evidence from initial consonant clusters 2900
Mackenzie Young, Colin Wilson

Assessing Science Inquiry using MDP Goal Detectors 2901
Michelle Lamar, Janet Koster van Groos

Metaphor and Memory: How Metaphors Instantiate Schemas in and Influence Memory of Narrative 2902
Sarah Cox, Paul Thibodeau

Is human learning driven by Prediction Error? 2903
Jiří Čevora, Máté Lengyel, Richard Henson

Modeling the Influence of Knowledge on Recognition: Connecting visual recognition behavior across development to PDP computational models of semantic knowledge 2904
Clint Jensen, Vanessa Simmering, Timothy Rogers

Explanation-based discourse inferences support early word learning 2905
Jessica Sullivan, Juliana Boucher, Rachel Goodkind, Reina Kiefer, Raymond Skyer, Katie Williams

Moral language in the Basel Accords: A quantitative analysis 2906
Eyal Sagi, Hadar Jabotinsky

Measuring Cognitive Skills through Conversation-Based Assessment 2907
G. Tanner Jackson, Katherine Castellano, Diego Zapata-Rivera

Single-kernel models of single-voxel visual selectivities in convolution neural networks 2908
Daniel Leeds, Ivan Iotzov

Attentional Resource Allocation in Multisensory Processing is Task-dependent 2909
Basil Wahn, Peter König

Cognitive Predictors of Timed and Untimed Early Arithmetic Performance 2910
Olivia Wassing, Marcie Penner-Wilger

On the Evaluability of Effort: Influences of Single and Joint Evaluation on Judgments of Subjective Effort in Memorial, Motor, and Perceptual Domains 2911
Timothy Dunn, Evan Risko

Towards a Pre-Newtonian Intuitive Physics of Object Collisions 2912
Ralf Mayrhofer, Michael Waldmann

Parsing Selective Attention and Executive Function in Children 2913
Staci Weiss, Peter Marshall

The effect of disfluency on mind wandering during text comprehension 2914
Myrthe Faber, Caitlin Mills, Kristopher Kopp, Sidney D'mello

Choosing Poorly: Reward-Induced Strategy Shifts in Estimating the Probabilities of Conjunctions and Disjunctions 2915
James Tripp, Adam Sanborn, Neil Stewart, Takao Noguchi

The presence of meaning constrains productive language processes: A test of the language game hypothesis in type writing. 2916
Sebastian Wallot, Joachim Grabowski

Modular versus Integrated Causal Learning 2917
Bob Rehder, Kelly Goedert, Ciara Willett, Raymond Blattner

A normative theory of visual working memory limitations 2918
Ronald van den Berg, Wei Ji Ma

Verbs of Explanandum seem Crucial in Evaluating Explanations 2919
Kwanghyeon Yoo, Kyung Soo Do

Network Analysis of Characters' Relationship in "Chronicle of Death foretold" using Graph Theory 2920
Jongin Kim, Yohan Moon, Hwiyeol Jo, Jeong Ryu

The construction of function representations 2921
Brian Montambault, Christopher Lucas, Joseph Austerweil

Infants' Developing Coordinated Visual-Manual Object Exploration and Links with Vocabulary Development 2922
Lauren Slone, Chen Yu, Linda Smith

Measuring individual and developmental differences in children's sense of confidence 2923
Carolyn Baer, Darko Odic

Assessing children's reading comprehension by the component processes tasks 2924
Chi-Shun Lien, Yuhtsuen Tzneg, Hsin-Yin Chien, Chin-Chih Chen

Linguistic alignment with artificial entities in the context of second language acquisition 2925
Astrid Marieke Rosenthal-von der Pütten, Carolin Straßmann, Nicole Krämer

Temporal event clustering in speech versus music 2926
Butovens Médé, Ramesh Balasubramaniam, Chris Kello

Developmental deficit in autobiographical episodic memory: Evidence from Williams syndrome . 2927
Katrina Ferrara, Alexandra Cohen, Barbara Landau

Building bilingual semantic representations based on a corpus-based statistical learning algorithm 2928
Xiaowei Zhao

Deconstructing the multi-dimensional Aha! experience 2929
Amory H. Danek, Shannon Menard, Jennifer Wiley

On Constancy in Spatial Perception 2930
Louise Daoust

Analogies and Graphics can lead to Illusions of Understanding 2931
Jennifer Wiley, Thomas Griffin, Allison Jaeger, David Sarmento, Marta Mielicki

Auditory N1 Amplitude Varies Across Multiple Acoustic and Phonological Dimensions in Speech 2932
Olivia Pereira, Joe Toscano

Does Contrast or Comparison Help More? The Role of Learning Mode and Category Type 2933
Jan Andrews, Ken Livingston, Calais Larson, Ken Kurtz

Speech Perception Across the Lifespan: Using a Gaussian Mixture Model to Understand Changes in Cue Weighting Between Younger and Older Adults 2934
Sarah Vrabic, Elke Nordeen, Joe Toscano

Individual Differences in the Acquisition of Strategies in a Complex Task 2935
Edward Cranford, Aaron Wong, Jaymes Durriseau, Kevin Barnes, Winston Jones, Gary Bradshaw, Jarrod Moss

Multimodal Dynamics of Explaining the Mechanisms of Global Warming 2936
Alexandra Paxton, Drew Abney, Kristina Castellanos, Gilbert Sepulveda

Imagining activities: The role of perspective and grammatical aspect 2937
Jeffrey Hong, Todd Ferretti, James Siklos-Whillans, Deanna Hall

The role of higher order relational structure in relational category label extension 2938
Sean Snoddy, Kenneth J. Kurtz

Memory for the Meaningless: Experts' Advantage at Recalling Unstructured Material 2939
Giovanni Sala, Fernand Gobet

"He will try to learn it because he doesn't know it." Young children's understanding of learning based on their knowledge states 2940
Jeein Jeong, Douglas Frye

The Relationship Between Mental-state Language and False-belief Understanding in Adulthood . 2941
Erin Roby, Alexandra Paxton, Rick Dale, Rose Scott

The use of dispositional cues to causality in judgements of mechanical and living interactions .. 2942
Julia Schneider, Sufi Abbaspour, Trix Cacchione

Learning Hierarchical Labels through Cross-situational Learning 2943
Chen Yu, Chi-hsin Chen

Intensional Probability Judgments and Inclusion Fallacies With Generics 2944
Momme von Sydow

Motor cortex excitability during processing of handwritten and typed non-action-related text 2945
Chelsea Gordon, Ramesh Balasubramaniam, Michael Spivey

In the nick of time: Using temporal cues to examine ongoing event representations 2946
Victoria Scharp, Connie Tompkins, Michael Dickey

Toward a Simulation Platform for Comparing Computational Cognitive Neuroscience Models .. 2947
David Noelle

Consequences of bilingualism for perceptions of categories and similarity 2948
Kevin Holmes, Jacob Brodsky

Mandarin-English Bilinguals Match Lexical-Tone Processing to the Language Context 2949
Carolyn Quam, Sarah Creel

Modifying Cognitive Load Component Survey for K-12 STEM Testing 2950
Jingyuan Huang, Janet Koster van Groos, Michelle Lamar

The Development of Intuitions about the Controllability of Thoughts, Emotions, and Behavior . 2951
Brittany Klimek, Amanda Brandone

A shape-heavy vocabulary does not a shape bias make: A comparison of the content of English-learning children's and Spanish-learning children's typical vocabularies 2952
Emily Russell, Christina Schonberg, Shawntel Barreiro

Motion Capture of Phase Change Transitions During Insight Problem Solving 2953
John Hart, Chelsea Johnson, Nicholas Duran

Exploring the Cognitive and Social Profile of Science Rejection 2954
Emilio Lobato, Corinne Zimmerman, Steve Croker

Technological Shaping of Verbal Working Memory: A Difference between Chinese Phonology-Based and Orthography-Based Typing 2955
Jenn-Yeu Chen, Adam Li

Adults' drawing and recognition of familiar objects and substances: Nonsolids are hard to identify 2956
Lynn Perry

How Spatial Ability and Stress Impact Escape Path 2957
Benjamin Nelligan, Alexandra Amorati, Deanne Adams, Christopher Galeucia, Laura Carlson

Rapid emotion discrimination in the infant brain 2958
Alexandra Marquis, Bruno Rossion, Margaret Moulson

Mental representations and processing of radical expressions 2959
Purav Patel, Sashank Varma

Data Shows Human Behavior is Not Random, period. 2960
William G. Kennedy

Feedback Markers in Mandarin: Tracking Cognitive Status in Conversation 2961
Li-chiung Yang

The Semantic Stroop Effect: An Ex-Gaussian Analysis 2962
Darcy White, Evan Risko, Derek Besner

Investigating the Effect of Experience on Concrete and Abstract Word Processing 2963
Selgun Yuceil, Didem Gokcay

A cognitive model of online event segmentation 2964
Anthony Harrison, Sangeet Khemlani, Greg Trafton

Human-object interaction understanding without objects 2965
Anna Belardinelli, Johannes Lohmann, Martin V. Butz

The Effect of Book Design on Beginning Readers' Attention Allocation 2966
Karrie Godwin, Anna Fisher

Generating Predictions Non-consciously: Evidence from Invisible Motion with and without Obstacles 2967
Ariel Goldstein, Ido Rivlin, Ran Hassin

Using analogical comparison to help children learn the day-night cycle 2968
Benjamin Jee, Florencia Anggoro, Natalie Evans, Caitlin Murphy, Jessica Tran, Caroline Morano, Amanda McCarthy, Victoria Jackson

A Preliminary Model of Situation Awareness in a Cognitive Architecture. 2969
Ehsan Khosroshahi, Dario Salvucci

Verifying who "jumped more" or "higher" in simple events 2970
Haley Farkas, Alexis Wellwood

Language influences attention to Japanese event components in native English-speaking 21- to 24-month-olds 2971
Natalie Brezack, Haruka Konishi, Roberta Golinkoff, Kathryn Hirsh-Pasek

Comparing predictions of lexical norm data obtained using word associations and word collocation 2972
Hendrik Vankrunkelsven, Steven Verheyen, Simon De Deyne, Gert Storms

Does GIScience Training Enhance Spatial Navigation Ability? 2973
Annisa Ahmed, Alina Nazareth, Danielle Burnetta, Magda Abozeed, Nora Newcombe

Plasticity of Categorization: Developmental Differences in Category Learning and Transfer between Children and Adults 2974
Sophia Deng, Vladimir Sloutsky

Unsupervised learning of VerbNet argument structure 2975
Jesse Mu, Timothy O'Donnell, Joshua Hartshorne

Moral Social Media: Heavy Facebook Users Accept Harsher Moral Criticism for Microaggressions 2976
Boyoung Kim, Joanna Korman, Bertram F. Malle

Memory-based decision making: Examining the relative influence of experimental and pre-experimental exposure. 2977
Justin Olds, Julian Marewski

The role of emotional mediation in musical and vocal sound-color correspondence 2978
Erin Isbilen

Collaborative Story Construction and Telling for Second Language Learning 2979
Mei Si

Partnering with Autonomous Vehicles 2980
Umesh Krishnamurthy, Paul Maglio

Mind Wandering during Film Comprehension: The Role of Prior Knowledge and Situational Interest 2981
Sidney Dmello, Kris Kopp, Caitlin Mills

A concurrent task facilitates, not impedes, the heel-to-toe standing balance in children: The case of a dual-task benefit 2982
Rong-Ju Cherng, Hsien-Ying Chen, Chiu-Yu Cho, Jenn-Yeu Chen

The effect of language impairment on non-symbolic exact quantity representation 2983
John Verbos, Sarah Wallace, Alexander Kranjec

Do Open Specifier Positions at Clause Edges Alleviate Working Memory Load? 2984
Ryan Rhodes

Attentional Enhancement at Event Boundaries 2985
Dare Baldwin, Eric Pederson

Historical Semantics of Risk 2986
Ying Li, Thomas Hills, Ralph Hertwig

An explicit theory of sortal representation and some evidence for it. 2987
Sandeep Prasada

The influence of temporal order on the recognition of causal relations 2988
Jane Neal, Katja Wiemer

Can Distributional Fitting of Short Semantic Fluency Results Predict ADHD? 2989
Janelle Szary, Michael N. Jones

Representations of Entropy and of the Relations Same and Different Early in Human Development 2990
Rebecca Zhu, Jean-Remy Hochmann, Sophia Sanborn, Susan Carey

Predictors of lexical stability in an artificially learnt language 2991
James Brand, Padraic Monaghan

Predictable stimulus onsets improve memory 2992
George Kachergis, Shannon Tubridy, Todd Gureckis

Can a Bayes' Net approach capture intuitive use of sequential testimonies in a legal reasoning paradigm? 2993
Jens Koed Madsen, Saoirse Connor Desai, Adam Harris, David Lagnado

Exploring Individual Differences in Preschooler's Causal Reasoning Skills in the Physical and Digital Domains 2994
Jessie-Raye Bauer, Amy Booth, Cristine Legare

A subject-object asymmetry in the online processing of 'only': evidence from eye-tracking 2995
Pooja Paul, Tanya Levari, Dylan Hardenbergh, Jesse Snedeker

Taking the Easy Way Out: Effects of Feedback on Children's Metacognitive Control 2996
Allison O'Leary, Vladimir Sloutsky

Instance-based Learning in Multi-cue Diagnosis 2997
Christopher Myers, Kevin Gluck, Andrew Halsey, Jack Harris

Rapid acquisition of novel information: Is disjunctive syllogism necessary for fast mapping? ... 2998
Hillary Abel, Anna Drummey, Stephanie Vicari, Irene Kan

Information Processing during Intertemporal Choice: An Investigation using Eye Movement Data 2999
Kelli L. Johnson, Michael T. Bixter, Christian C. Luhmann

Neural Resonance Theory: Entrainment to Missing Pulse Rhythms 3000
Charles Wasserman, Jung Nyo Kim, Edward Large, Erika Skoe

Young children's estimation of difficulty and time 3001
Hyowon Gweon, Mika Asaba

Exploring the Use of Conversations with Virtual Agents in Assessment Contexts 3002
Diego Zapata-Rivera, G. Tanner Jackson, Irvin R. Katz

Mothers' use of emotion words to their 15- to 18- month old children. 3003
Mai Hamana

Effects of categorization on item memory and forgetting 3004
Kevin Darby, Vladimir Sloutsky

Optimal Predictions in Illness Cognition 3005
Talia Robbins, Pernille Hemmer

Language acquisition as sparse foraging: mapping path-dependence in word acquisition 3006
Riccardo Fusaroli, Malte Lau Petersen

Multiple Systems for Modal Cognition 3007
Jonathan Phillips, Fiery Cushman

The Effect of Varying Problem Contexts on Learning Probability Rules 3008
Tiantian Jin, Marianna Lamnina

Analyzing experimental paradigms under modification on web-based experiment platforms 3009
Michael Pacer

Sex Differences in Mental Rotation Performance: The Self-fulfilling Prophecy of Gender Stereotypes 3010
Alina Nazareth, Shannon Pruden

Relative influence of anchoring and centering biases in reconstructive memory 3011
Sean Duffy, L. Elizabeth Crawford, John Smith

Making invisible "trouble" visible: Self-repair increases abstraction of referring expressions 3012
Gregory Mills, Gisela Redeker

Inferring actions, intentions, and causal relations in a neural network 3013
Andrew Saxe

Forming spatial habits in the wild: Examining stabilization in classroom seating arrangements . 3014
Mona Zhu, Evan Risko

An action dynamics study of the onset of prediction 3015
Maryam Tabatabaeian, Rick Dale

The ON/IN scale of semantic extensions of adpositions: testing through artificial language learning 3016
Matthew Stave, Eric Pederson

Time Course of Fidelity and Contributing Factors to Long-Term Memory 3017
Kimele Persaud, Pernille Hemmer

The Neurorobotics Platform of the Human Brain Project 3018
Florian Roehrbein, Marc-Oliver Gewaltig, Cecilia Laschi, Gudrun Klinker, Paul Levi, Alois Knoll

Understanding human facial attractiveness from multiple views 3019
Amanda Song, Linjie Li, Vicente Malave, Gary Cottrell, Angela Yu

Causal perception is constrained by principles of Newtonian mechanics 3020
Jonathan Kominsky, Brent Strickland, Annie Wertz, Claudai Elsner, Karen Wynn, Frank Keil

Finger Gnosis Predicts Children's Numeracy, Despite Controlling for Visuo-Spatial Memory ... 3021
Marcie Penner-Wilger, Sheri-Lynn Skwarchuk, Carla Sowinski, Jo-Anne Lefevre

Does Experience with Physics Concepts Improve Mental Rotation Performance? 3022
Rosalie Odean, Alina Nazareth, Jessica Bartley, Eric Brewe, Angela Laird, Shannon M. Pruden

Pragmatic inference in definite and indefinite contexts 3023
Kelsey Moty, Mahesh Srinivasan

Using eye tracking data to compare models of numerical estimation 3024
Bradley Morris, Christopher Was, Amy Masnick, Bushra Aldosari

A Cognitively Realistic Model of Decision Making in Ocean Ecology 3025
Philipp Koralus, Jens Madsen, Ernesto Carella, Richard Bailey

Neural bases of semantic-memory deficits for events 3026
Chia-Ming Lei, Haley C. Dresang, Michelle B. Holcomb, Tessa C. Warren, Michael Walsh Dickey

Seeing the Bees Buzz and Hearing the Diamonds Glisten: The Effect of the Mode of Presentation of Stimuli on the Modality-Switch Effect 3032
Elisa Scerrati, Luisa Lugli, Anna M. Borghi, Roberto Nicoletti

Bilingual Proficiency Affects Inhibitory Control: A study of Stroop Performance in 8-year-old English-Chinese Singaporean Children 3033
Carissa Kang

Linguistic recursion and Autism Spectrum Disorder 3034
Irina Polyanskaya, Torben Braüner, Patrick Blackburn

Computational explanation of "fiction text effectivity" for vocabulary improvement: Corpus analyses using latent semantic analysis 3035
Keisuke Inohara, Akira Utsumi

Interaction, abstraction and complexity 3036
Kristian Tylén, Riccardo Fusaroli, Pernille Smith, Jakob Arnoldi

Using Nature Language Processing to Improve Optical Character Recognition 3037
Lining Xu, Yongxu Wu

Artificial Language Learning: The Context of Negation 3038
Ariel Mathis, Stephanie Huette

Source Expertise and Question Type Effects in Conversation-Based Assessment 3039
Jesse R. Sparks, Jessica J. Andrews, Diego Zapata-Rivera, Blair Lehman, Kofi James

Implicit updating of object representation via predictive associations 3040
Ru Qi Yu, Jiaying Zhao

Roles of Metacognitive Suggestions in Hypothesis Revision 3041
Sachiko Kiyokawa, Kazuhiro Ueda, Yoshimasa Ohmoto

Tactile protocol analysis: Observations of novices reading data tables by touch 3042
Peter Cheng, Panayiota Polycarpou, Grecia Garcia Garcia, Frances Aldrich

In the event of a turn exchange: Visual information the perception of turn-taking behavior in natural conversation 3043
Nida Latif, Agnès Alsius, K.g. Munhall

An Eye For Figurative Meaning: The Effects of Familiarity on Metaphor Comprehension 3044
Chris Genovesi, Michael Vertolli

The Concept of Transcendent God and the Emergence of Cognitive Level D in the Four-Level-Cognitive-Development Theory 3045
Vladimir Glebkin

Implicit versus explicit language learning: Differential effects of working memory and learning styles 3046
Sarah Grey, Christine Schoetensack, Kimberly Bell, Padraic Monaghan, Patrick Rebuschat

Enhancing Creativity in Children by Imparting Chess Training 3047
Ebenezer Joseph, Veena Easvaradoss, Suneera Abraham, Michael Brazil, David Chandran

The role of grammatical form in generalizing principled and statistical properties 3048
Kim Kristina Fuellenbach, E. Matthew Husband

Priming Spatial Reference Frames for Memory 3049
Elena Andonova

Deep Learning and Attentional Bias in Human Category Learning 3050
Leyla Roksan Caglar, Stephen Jose Hanson

Re-presenting a Story by Emotional Factors using Sentiment Analysis Method 3051
Hwiyeol Jo, Yohan Moon, Jongin Kim, Jeong Ryu

Reexamining the Unaccusative Hypothesis: a Visual World Paradigm study 3052
Yujing Huang, Laine Stranahan, Jesse Snedeker

Negation affects processing of correct and incorrect information: A visual-world paradigm for misinformation 3053
Jeffrey Viaud, Stephanie Huette

Lexicalization Typology of Realization Events in Mandarin Chinese 3054
Yu Deng, Fuyin Li

An Analogical Model of Pretense 3055
Irina Rabkina, Ken Forbus

A Model of Language-Guided Concept Formation using a Common Framework for Unsupervised and Supervised Learning 3056
Paul Munro

Learning Spiking Neural Controllers for In-Silico Navigation Experiments 3057
Mahmoud Akl, Florian Walter, Florian Röhrbein

Why are we (un)systematic? the (empirical) costs and benefits of learning universal constructions 3058
Steven Phillips, Yuji Takeda, Fumie Sugimoto

Reasoners are influenced by conversational pragmatics in abstract conditional reasoning tasks .. 3059
Paul Thibodeau, Daniel Grodner

A Brain-Based Feature Model of Adjective-Noun Composition 3060
Elizabeth A. Shay, Rajeev D. S. Raizada

Experience Representation of Artificial Cognitive System in Interaction with Real World 3061
Liwei Zhang, Zhen Deng, Jianwei Zhang, Zhixian Chen, Ying Hu

Is it Living? Insights from Modeling Event-Oriented, Self-Motivated, Acting, Learning and Conversing Game Agents 3062
Martin Butz, Mihael Simonic, Marcel Binz, Jonas Einig, Stephan Ehrenfeld, Fabian Schrodt

Cognitive Sciences Strategies for Futures Studies (Foresight) 3063
Ahmad Mahdeyan

Extending the Talmyan Typology: A Macro-event as Event Integration and Grammaticalization 3064
Fuyin Li

Perceptual Decision Making of Humans and Deep Learning Machines: a Behavioral Study 3065
Ge Chen, Yuzhong Liu, Qiyang Zhao

Age Related Differences in Episodic Memory Recollections: Applying Latent Dirichlet Allocation to Free-Writings on Driving Incidents by Older and Young Drivers 3066
Ritsuko Iwai, Ritsuko Iwai, Takatsune Kumada, Takatsune Kumada, Daisuke Kawahara, Sadao Kurohashi

The Temporal Cheerleader Effect: Attractiveness Judgments Depend on Surrounding Faces Through Time 3067
Alexia Toskos Dils, Scott Niedopytalski, Jeffrey Arroyo, Stephen Flusberg

Information Acquisition: Stopping Rules For Varying Levels of Probabilities and Consequences . 3068
Gustavo Gauer, Roberto Nonohay, Richard Gonzalez, Guilherme Lannig

Categorization of Probability Word Problem: Effects of Prior Statistical Training and Semantic Schema 3069
Chenmu Xing, James Corter, Doris Zahner

Differential Processing for Actively Ignored Pictures and Words 3070
Scott Sinnett, Maegen Walker, Margeaux Ciraolo, Andrew Dewald

Kindergarteners and adults learn fraction-rules in a categorization task 3071
Tasha Posid, Vladimir Sloutsky

Promoting prosocial behavior and emotional awareness in preschoolers 3072
Qiuwen Dou, Sarah-Jane Vick, Michael C. W. Yip

A Cognitive Approach to Modeling Sentence Level Prominence Based on Stimulus Unpredictability 3073
Sofoklis Kakouros, Okko Räsänen

Simulating the cost of cooperation: A recipe for collaborative problem solving 3074
Alessandro Lazzeri, Andrea Guazzini, Daniele Vilone, Giorgio Gronchi

Choice magnitude and decision time: investigating magnitude sensitivity in value-based decision making 3075
Angelo Pirrone, Philip Pärnamets, James Marshall, Tom Stafford

Multimodal Object Recognition and Categorisation by Interactive Behaviours 3076
Haojun Guan, Jianwei Zhang

Preschoolers evaluate risk and reward in exploration-exploitation tasks 3077
Elizabeth Lapidow, Elizabeth Baraff Bonawitz

Transfer at the Level of Human-Computer System: Problem Solving using Procedure-Automation Software 3078
Dorrit Billman, Debra Schereckenghost, Zachary Caddick

What Makes Campaign Messages Popular On Twitter ? 3079
M. Afzal Upal, Pritham Marupaka

Investigating the Explore/Exploit Trade-off in Adult Causal Inferences 3080
Erik Herbst, Christopher Lucas, Daphna Buchsbaum

Author Index 3081

Reviewers List 3101

Dear Cognitive Scientists,

Welcome to Philadelphia for the 38th Annual Conference of the Cognitive Science Society! Our meeting brings together some of the most innovative and exciting research in cognitive science. The program features exceptional plenary talks by three international figures: Brian Scholl, Amanda Woodward, and Jeffrey Zacks. It also includes two invited symposia aimed at showcasing the theme for this year: **Recognizing and Representing Events**.

CogSci 2016 received 877 submissions, including 656 full papers, 193 member abstracts, 6 publication-based presentations, 11 symposium proposals, and 11 workshop and tutorial proposals. After a rigorous review process, we selected 222 papers for oral presentation (34%), 258 papers for poster presentation (39%), 193 member abstracts for poster presentation, 6 publication-based talks, and 11 symposia. All 11 tutorials and workshop submissions were deemed promising and of wide interest, and were accepted.

We hope that you enjoy the program this year and the city of Philadelphia, one of the most vibrant and historically significant cities in the United States. There are thousands of shops and restaurants near the convention center and throughout the city, and it is a convenient access point to the greater Philadelphia area and the east coast. We encourage you to set aside the time to enjoy some of the many activities the region has to offer.

Your Hosts,

Dan Grodner, Dan Mirman, Anna Papafragou and John Trueswell

How to Cite Your Paper

APA formatted citation for a 6-page paper:

Author A. & Author B. (2016). This is the title of the paper. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. PAGES). Austin, TX: Cognitive Science Society.

APA formatted citation for a published abstract:

Author A. & Author B. (2016). This is the title of the abstract [Abstract]. In A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (p. NUMBER). Austin TX: Cognitive Science Society.

APA formatted citation for a talk (or poster) presentation:

Author A. & Author B. (2016, August). This is the title of the talk or poster. Paper (or Poster) presented at the 38th Annual Conference of the Cognitive Science Society. Philadelphia, Pennsylvania USA.

Acknowledgments

We are grateful to everyone who contributed to the planning and organization of this year's meeting and to all reviewers who generously donated time to evaluate submissions. We thank the 190 Program Committee members and the 1300 reviewers who were essential to the review process. We thank the awards committee who helped assess the prize-winning papers. All Program Committee members and reviewers are listed below in the proceedings.

We are especially grateful for the assistance of a number of individuals and groups for many organizational aspects of the meeting. We thank Jessica Wong, the Cognitive Science Society's Conference Officer, for managing pretty much everything; Deborah Gruber, the Cognitive Science Society's Business Manager, for handling the business details; Richard Catrambone, Nora Newcombe, Susan Gelman, Susan Goldin-Meadow, and the Cognitive Science Society's Executive Committee and Governing Board for their constant support and advice; the Scarritt Group, Inc., for logistics and support; James Stewart of Precision Conference Solutions for maintaining the conference reviewing system; Sarah Fairchild for all her help, especially for maintaining the schedule on

the web; Rick Dale, Paul Maglio, Teenie Matlock and the other members of last year's Cogsci conference committee for their helpful advice and insights; Martin Butz and Anna Belardinelli, University of Tübingen, who wrangled the student volunteers; the student volunteers, listed below, who keep the conference running smoothly on site; and importantly our conference sponsors, listed below, whose generous support enabled us to focus more on the program and less on the finances.

We are also thankful for the support of the Cognitive Science community. We are sure that the suggestions and contributions we received from this community helped us to make it a more lively, engaging and fun meeting for all. Enjoy!

- Dan Grodner, Dan Mirman, Anna Papafragou and John Trueswell

Program Co-Chairs, Cognitive Science 2016

Sponsors

The Robert J. Glushko and Pamela Samuelson Foundation

Educational Testing Service (ETS)

John Wiley & Sons Ltd.

Nature Human Behaviour – Launching 2017

Swarthmore College, Department of Psychology

Temple University, Department of Psychology

University of Delaware, Department of Psychological and Brain Sciences

University of Pennsylvania, Institute for Research in Cognitive Science

University of Pennsylvania, School of Arts and Sciences

Cognitive Science Society

Thank you again for your support!

Committees

Organizing Committee:

Sudha Arunachalam (Boston University)

Stella Christie (Swarthmore College)

Dan Grodner (Swarthmore College)

Dan Mirman (University of Alabama at Birmingham)

Cat Norris (Swarthmore College)

Jared Novick (University of Maryland)

Anna Papafragou (University of Delaware)

John Trueswell (University of Pennsylvania)

Organizing Committee Chair: Anna Papafragou

Program Co-Chairs: Dan Grodner, Dan Mirman, Anna Papafragou and John Trueswell

Workshop, Tutorial and Symposium Co-Chairs: Stella Christie and Cat Norris

Member Abstract Co-Chairs: Sudha Arunachalam and Jared Novick

Invited Symposium Organizers: Anna Papafragou, John Trueswell and Dan Grodner

Awards Committee: Dan Mirman, Glenn Gunzelmann, and Chris Sims

Website Chair: Jared Medina

Student Volunteer Chairs: Anna Belardinelli and Martin Butz (University of Tübingen)

Sponsorship Chair: John Trueswell

Conference Officer: Jessica Wong (University of Chicago)

Program Committee Members

Abrahamson, Dor
Abramson, Darren
Adrian, Staub
Alan, Garnham
Altmann, Erik
Andrews, Janet
Arik, Engin
Armstrong, Blair
Arunachalam, Sudha
Ashley, Richard
Barley, Mike
Barner, David
Bartels, Daniel
Beer, Randall
Bender, Andrea
Benjamin, Kuipers
Bertel, Sven
Best, Brad
Bhatia, Sudeep
Billman, Dorrit
Blessing, Stephen
Bonatti, Luca
Bonawitz, Elizabeth
Bonnefon, Jean-Francois
Borovsky, Arielle
Bosse, Solveig
Bott, Lewis
Branigan, Holly
Breheny, Richard
Brighton, Henry
Brown, Joshua
Bucciarelli, Monica
Burns, Bruce
Carruthers, Peter
Casasanto, Daniel
Catrambone, Richard
Chang, Franklin
Chen, Lang
Chernyak, Nadia
Christiansen, Morten
Christie, Stella
Clausner, Timothy
Clement, Catherine
Clifton, Charles
Cooper, Rick
Cottrell, Gary
Craig, Scotty
Cravalho, Patrick
Croker, Steve
Culbertson, Jennifer
Cummins, Fred
Cushman, Fiery
Dale, Rick
Danks, David
Danovitch, Judith
Davidoff, Jules
De Brigard, Felipe
de Sa, Virginia
Deák, Gedeon
Dehghani, Morteza
Di Eugenio, Barbara
Dixon, James
Doumas, Alex
Echols, Catharine
Elizabeth, Brannon
Epstein, Susan
Farmer, Thomas
Fausey, Caitlin
Feist, Michele
Fiser, Jozsef
Fisher, Anna
Fisher, Cynthia
Forbus, Ken
Frank, Mike
Frank, Stefan
Freksa, Christian
Fu, Wai-Tat
Fum, Danilo
Galati, Alexia
Gelman, Susan
Gentner, Dedre
Gerry, Altmann

Gershman, Samuel
Gildea, Daniel
Gluck, Kevin
Goel, Ashok
Gomez, Rebecca
Gonnerman, Laura
Gonzalez, Cleotilde
Goodman, Noah
Griffiths, Oren
Griffiths, Tom
Grinberg, Maurice
Gunderson, Elizabeth
Gunzelmann, Glenn
Gureckis, Todd
Gweon, Hyowon
Hale, John
Hare, Mary
Hartshorne, Joshua
Hayhoe, Mary
Houpt, Joseph
Hsiao, Janet
Hubbard, Edward
Husband, E. Matthew
Jaeger, T. Florian
Johnson, Samuel
Kaiser, Elsi
Kan, Irene
Katherine, White
Katz, Irvin
Keller, Frank
Kello, Christopher
Kidd, Celeste
Kloos, Heidi
Knobe, Joshua
Knoeferle, Pia
Kominsky, Jonathan
Kurumada, Chigusa
Kushnir, Tamar
Levin, Beth
Levy, Roger
Lew-Williams, Casey
Li, Peggy
Little, Hannah
Luhmann, Christian
Lupyan, Gary
M., Nieuwland
MacDonald, Maryellen
Machery, Edouard
Magnani, Lorenzo
Marina, Bedny
Markman, Art
Markson, Lori
Masnick, Amy
Matthews, Danielle
Mazuka, Reiko
Medina, Jared
Morett, Laura
Morris, Bradley
Munro, Paul
Norris, Catherine
Noveck, Ira
Novick, Jared
O'Seaghdha, Padraig
Odic, Darko
Olivetti, Marta
Omaki, Akira
Patrick, Sturt
Paul, Muentener
Pearl, Lisa
Perfors, Amy
Rabagliati, Hugh
Rawlins, Kyle
Redeker, Gisela
Rips, Lance
Sarah, Brown-Schmidt
Scott, Rose
Sloutsky, Vladimir
Snyder, William
Soojin, Park
Sprouse, Jon
Srinivasan, Mahesh
Stein, Nancy L
Stenning, Keith
Steven, Most
Strickland, Brent
Swingley, Daniel
Syrett, Kristen
Tenenbaum, Josh

Trafton, Greg
Tversky, Barbara
Varakin, Donald Alexander
Vickery, Timothy
Victor, Ferreira
Warren, Tessa
Was, Christopher
Watson, Duane
Wellwood, Alexis
Werker, Janet
Williams, Joseph Jay
Wills, Andy
Wolff, Phillip
Xu, Fei
Yang, Charles
Zimmerman, Corinne

Student Volunteers

Elbanani, Mohamed (University of Michigan)
Finch, Dustin (Indiana University)
Gordon, Chelsea (University of California)
Hart, John (Arizona State University)
Hopman, Elise (University of Wisconsin, Madison)
Jensen, Clint (University of Wisconsin)
Khosroshahi, Ehsan Baradaran (Drexel University)
Kimel, Eva (The Hebrew University of Jerusalem)
Kluth, Thomas (University Bielefeld)
Médé, Butovens (University of California)
Miller, Hilary (University of Wisconsin-Madison)
Patel, Purav (University of Minnesota Twin Cities)
Raviv, Limor (Max Planck Institute for Psycholinguistics, Netherlands)
Ross, Robbie (University of Oregon)
Sanches de Oliveira, Guilherme (University of Cincinnati)
Ye, Jian-Ping (Teachers College Columbia University)

Conference Awards

Robert J. Glushko Dissertation Prizes

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

The 2016 recipients of the Robert J. Glushko Prizes for Outstanding Doctoral Dissertations / Theses in Cognitive Science are listed below.

Lang Chen, *White matter connectivity explains category-specific brain activation and impairment: A neurocomputational model of semantic cognition*, PhD 2014, University of Wisconsin – Madison.

Isabelle Dautriche, *Weaving an ambiguous lexicon*, PhD 2015, University of Paris.

Andrew Saxe, *Deep linear neural networks: A theory of learning in the brain and mind*, PhD 2015, Stanford University.

Philip Pärnamets, *Observing and influencing preferences in time*, PhD 2015, Lund University.

Jan Engelmann, *An empirical investigation of the evolutionary and ontogenetic roots of trust*, PhD 2014, Max Planck Institute for Evolutionary Anthropology.

For more information about awardees and their dissertations, see

http://www.cognitivesciencesociety.org/about_awards_glushko_recipients.html

The *Glushko Dissertation Prize Symposium* showcases the award winning PhD research projects, moderated by Linda B. Smith (Indiana University). Friday, 12 August, 2:30pm-4:10pm.

Paper Awards

Marr Prize

The Marr Prize, named in honor of the late David Marr, is awarded to the best student paper at the conference. All student first authors were eligible for the Marr Prize for the best student paper. The Marr Prize includes an honorarium of $1000 and is sponsored by The Cognitive Science Society. The winner of the 2016 Marr Prize for the Best Student Paper is

Wai Keen Vong, Andrew Hendrickson, Amy Perfors, and Daniel Navarro: *Do additional features help or harm during category learning? An exploration of the curse of dimensionality in human learners.* Friday, 12 August, 2:55-3:20pm.

Computational Modeling Prizes

Four prizes worth $1000 each are awarded for the best full paper submissions to CogSci 2016 that involve computational cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are sponsored by The Cognitive Science Society. The winners of the 2016 Computational Modeling Prizes are listed below.

Perception and Action: **Joshua Peterson, Joshua Abbott, and Thomas Griffiths:** *Adapting deep network features to capture psychological representations.* Thursday, 11 August, 5:30-5:55pm.

Language: **Padraic Monaghan:** *Degeneracy results in canalisation of language structure: A computational model of word learning.* Friday, 12 August, 10:30-10:55am.

Higher-Level Cognition: **Mathis Richter, Jonas Lins, and Gregor Schoener:** *A neural dynamics parses object-oriented actions.* Thursday, 11 August, 10:30-10:55am.

Applied Cognition: **Sean Aubin, Aaron Voelker, and Chris Eliasmith:** *Improving with practice: A neural model of mathematical development* Friday, 12 August, 3:45-4:10pm.

Student Travel Awards

The Robert J. Glushko and Pamela Samuelson Foundation generously sponsored $10,000 for student travel awards. Travel awards have been provided to students whose submissions were accepted as full papers, received high rankings, and who indicated a need for travel funding. This year's travel awards went to:

Yuki Asano, Sean Aubin, Nolan Conaway, Tom Gijssels, Caroline Graf, Christian Gumbsch, Robert X.D. Hawkins, Rose Hendricks, Michele Herbstritt, Max Kleiman-Weiner, Dave Kleinschmidt, Miriam Novack, Joshua Peterson, Benjamin Pitt, Mathis Richter, Kathryn Schuler, Daniel Spokoyny, Robert Thorstad, Wai Keen Vong, and Tian Xu.

Invited Presentations

Rumelhart Prize Lecture

Dedre Gentner, Northwestern University

Why we're so smart: Analogical processing and relational representation

Friday, August 12, 4:45pm

Terrace Ballroom 4

Rumelhart Prize Symposium

The Symposium in Honor of Dedre Gentner

Panelists:

Kenneth Forbus, Northwestern University

Micah Goldwater, University of Sydney

Kenneth Kurtz, Binghamton University

Stella Christie, Swarthmore College

Jeffrey Loewenstein, University of Illinois, Urbana-Champaign

Arthur Markman, University of Texas, Austin

Friday, 12 August, 10:30am

Room 126A

Thursday's Keynote

Brian Scholl, Yale University

Let's see what happens: Dynamic events as foundational units of perception and cognition

Thursday, August 11, 9:00am

Terrace Ballroom 4

Friday's Keynote

Jeffrey Zacks, Washington University, St. Louis

Event comprehension and memory in the predictive brain

Friday, August 12, 9:00am

Terrace Ballroom 4

Saturday's Keynote

Amanda Woodward, University of Chicago

Putting the action back into infant cognition

Saturday, August 13, 9:00am

Terrace Ballroom 4

Invited Symposia

Invited Symposium 1: Representing events in vision, language and action
Organizer and Moderator: John Trueswell, University of Pennsylvania

Panelists:

John Aloimonos, University of Maryland
Gerry Altmann, University of Connecticut
Dare Baldwin, University of Oregon
Marina Bedny, Johns Hopkins University

Thursday, August 11, 10:30am

Invited Symposium 2: Causal Reasoning: Origins and Development
Organizer and Moderator: Anna Papafragou, University of Delaware

Panelists:

Elizabeth Bonawitz, Rutgers University, Newark
Gergely Csibra, Central European University
Frank Keil, Yale University
L.A. Paul, University of North Carolina, Chapel Hill

Saturday, August 13, 10:30am

Active learning: Cognitive development, education, and computational models

Organizers:
Elizabeth Bonawitz (elizabeth.bonawitz@rutgers.edu)
Rutgers University – Newark, Department of Psychology
101 Warren Street; Newark, NJ 07102 USA

Fei Xu (fei_xu@berkeley.edu)
University of California – Berkeley, Department of Psychology
Tolman Hall; Berkeley, CA 94702 USA

Keywords: active learning, exploration, information gain, computational models, education

Generating data for learning

Behold! Human beings living in an underground den, which has a mouth open towards the light and reaching all along the den; here they have been from their childhood, and have their legs and necks chained so that they cannot move, and can only see before them, being prevented by the chains from turning round their heads. …they see only their own shadows, which the fire throws on the opposite wall of the cave… To them truth would be literally nothing but the shadows of the images.

-Plato's Republic

Socrates' point was to show that, like his prisoners, none of us can ever be sure about the truth of the world. However, unlike his prisoners who passively watch the shadows on the wall, we are equipped with the ability to drive our own learning. From infancy on, our lives are filled with self-directed opportunities for acquiring information about the world. Active learning encompasses many of these self-directed opportunities. It includes attending to a particular event in our environment over another, mentally searching for explanations, asking questions and knowing who to ask, or taking actions. Although there are numerous contexts for which *active learning* applies, they are united by the broader goal of generating data for learning. The factors that influence active learning may be crucial to human intelligence – they afford purposeful data gathering by the learner.

Active learning is not only an important topic for understanding human behavior, but also for developing intelligent machine algorithms. There have been numerous computational approaches towards capturing aspects of how and when a learner might benefit by generating further data. These include algorithm that favor novelty, surprising events (Shannon information), and ambiguity (e.g. as in when Bayesian posteriors equally support multiple hypotheses).

One well-studied computational approach to optimally generating data for learning is known as information gain, or the Kullback-Leibler divergence (KL Divergence) Simply, the KL Divergence measures the degree to which one's beliefs after having seen the evidence differ from the beliefs they held just prior to observing the evidence. An optimal learner takes actions so as to maximize this information gain.

Although computational approaches have begun to lay the foundation for optimal approaches to active learning, less is known about how children and adults approach the problem of data generation. What are the cognitive mechanisms that influence active learning? Are learners systematic? Rational? Are their decisions captured by information gain? How does active learning interact with teaching and instruction?

In this workshop, we invite speakers from a variety of approaches to broadly inform our understanding of active learning, including cognitive development, education, and computational modeling. We examine what "active" means in active learning, and present talks on the cognitive mechanisms that might support active learning, including attention, hypothesis-generation, explanation, pretend play, and question asking. We also explore how efficient learners are when planning and executing actions in the service of learning, and whether there are developmental or socio-economic differences in active learning. We integrate the problem of active learning with teaching to investigate the similarities and differences involved in selecting evidence for oneself and others. Throughout we ask how we can capture these processes with computational models that spell out the underlying assumptions and potential algorithms.

Identifying factors that influence active learning is important because it could lead to understanding broader individual differences in drive for learning, with direct consequences for the development of informal and formal educational practices. Interdisciplinary research leveraging tools from computational and developmental science toward educational goals has the potential for generating critical insights for each of the fields involved. The proposed workshop will bring together these different communities, to encourage interdisciplinary dialog in this important topic.

In active learning we are broken free from the chains of passive observation, affording self-generated discovery of the true nature of the shadows on the wall.

Topics and Speakers

The workshop is divided into three main themes, with speakers from education, modeling, and developmental backgrounds in each. After each set of talks, we schedule ample time for discussion lead by a panel moderator who is an expert in the field, encouraging participants to fully engage with the speakers. The workshop concludes with a final discussion broad discussion on open questions and the future of active learning research.

Opening remarks

Cognitive mechanisms in active learning

1. *Switches in attention reflect a strategy for optimizing learning.* Madeline Pelz & Celeste Kidd. (University of Rochester)
2. *Hypothesis generation processes and how they guide active exploration.* Doug Markant. (Center for Adaptive Rationality, Max Planck Institute for Human Development).
3. *Mechanisms in children's active learning: self-generated explanations.* Caren Walker & Tania Lombrozo (University of California, San Diego; University of California, Berkeley).
4. *Progress in building a machine that can ask interesting and informative questions.* Anselm Rothe, Brenden Lake & Todd Gureckis* (New York University) *presenter.
5. Panel Discussion

Development of active learning (Part 1)

1. *5- and 7-year-olds benefit from selection learning in a category-learning task.* Zi L. Sim, & Fei Xu. (UC Berkeley)
2. *Five-year-old children identify the most informative questions.* Azzurra Ruggeri. Azzurra Ruggeri*, Zi Lin Sim*, & Fei Xu. (Max Planck Institute for Human Development, Berlin, Germany; University of California, Berkeley) *contributed equally.
3. *Children's selective social referencing during word learning.* Emily Hembacher and Michael C. Frank (Stanford University)
4. Invited Panel Moderator Laura Schulz

Development of active learning (Part 2)

1. *Can preschoolers direct their learning based on difficulty? Evidence from word learning*. Stephanie Denison (University of Waterloo).
2. *Cognitive heuristics for computing information gain in children and adults.* Elizabeth Lapidow & Elizabeth Bonawitz. (Rutgers University – Newark)
3. *Socioeconomic status and exploratory play in early childhood.* Julia Leonard, John D.E. Gabrieli, & Laura Schulz (Massachusetts Institute of Technology)
4. Invited Panel Moderator Fei Xu

Active learning and teaching

1. *Cost-and-benefit analysis in planning and helping others learn.* Hyowon Gweon (Stanford University)
2. *Teaching versus active exploration: A computational analysis of conditions that affect learning.* Scott Cheng-Hsin Yang & Patrick Shafto (Rutgers University – Newark)
3. *Why guided play is a form of active learning.* Deena Weisberg, Roberta M. Golinkoff, Kathy Hirsh-Pasek, and Marcia Shirilla (University of Pennsylvania, University of Delaware, Temple University)
4. *The grandparent method: A computational model of experience-based guided learning.* Sophia Ray Searcy, Yue, Yu, Scott Cheng-Hsin Yang, & Patrick Shafto (Rutgers University – Newark)
5. Invited Panel Moderator Kathy Hirsch-Pasek

Closing remarks

Learning to Talk about Events
Grounding Language Acquisition in Intuitive Theories and Event Cognition

Eva Wittenberg (ewittenberg@ucsd.edu)
Department of Linguistics, University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093-0108 USA

Melissa Kline (mekline@mit.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02138 USA

Joshua K. Hartshorne (joshua.hartshorne@bc.edu)
Department of Psychology, Boston College
140 Commonwealth Ave, Chestnut Hill, MA 02467

Keywords: language acquisition; concepts; event cognition; cognitive development; intuitive theories; argument structure

Introduction

Language is a powerful tool for moving thoughts from the mind of one person to another. Thus, any theory of language must make contact with theories of thought and conceptual representation. Any theory of language acquisition must explain how children link words to concepts, whether those concepts were pre-existing or created during the process of language acquisition. Conversely, theories of conceptual representation are constrained by the need to support language and language acquisition. Indeed, there is a rich tradition of cognitive science research that explicitly tackles these issues (Clark, 2004; Bowerman, 1989).

This is an opportune time to assess our current understanding of how language acquisition might be grounded in thought – particularly focusing beyond the better-explored domains of objects and kinds to include concepts whose nature may differ drastically from these. Events, which (at a minimum) require specifying how an activity and one or more participants are related to each another, are a central instance of such concepts. Recent years have seen the emergence of a robust psychological literature on event representations (Tversky & Zacks, 2013). There has likewise been explosive growth in "intuitive theory" or "Theory Theory" approaches to conceptual representation (Gopnik & Wellman, in press; Hartshorne, O'Donnell, & Tenenbaum, 2015; Goodman, Ullman, & Tenenbaum, 2011; Battaglia, Hamrick, & Tenenbaum, 2013). On such accounts, particular concepts are embedded in robust theories and derive their meaning from their roles in those theories. Importantly, our growing understanding of concepts and conceptual relations is not limited to mature representations: Work on infant and child conceptual representations has been particularly productive (Gopnik & Wellman, in press; Jara-Ettinger, Gweon, Tenenbaum, & Schulz, in press; Hernik & Csibra, 2015).

Language research has also seen a flourishing of theories of concepts. Linguists now have well-specified, articulated theories of the semantics of verbs that draw on structured representations of events and other abstract concepts(Levin & Rappaport Hovav, 2011, 2005). There is now increasingly strong evidence for the psychological reality of these structures and their role in language acquisition (Ambridge, Pine, Rowland, Chang, & Bidgood, 2013; Hartshorne et al., in press). While these theories bear certain similarities to the infant cognition work (e.g., representations of agency, intentionality, and causation play key roles in both), in other ways they diverge (e.g., the latter do not have a clear analog to an intuitive theory).

This workshop brings linguists and language acquisition researchers together with experts in event cognition and cognitive development in order to understand how these independent but linked fields inform one another. Because language makes direct contact with thought (and vice versa) it is likely that the achievements of one field will inform the others, and discrepancies between fields must ultimately be resolved to adequately explain phenomena in both areas. Given the interdisciplinary nature of this discussion, the Annual Meeting of the Cognitive Science Society is an ideal venue.

Goals and Scope

This workshop brings together leading researchers in the language and cognition of concepts and events – both as speakers and as audience members – in order to share knowledge, discuss open research questions of mutual concern, and shape the path forward. Precisely because questions about the representation of concepts and events are so broadly applicable across the cognitive sciences, they tend to be studied fairly independently in multiple different (sub)disciplines. Thus, gatherings like this one are crucial for ensuring efficient dissemination of ideas and findings.

The workshop is organized around language acquisition, particularly the acquisition of verbs. This will help focus discussion without necessarily sacrificing breadth: Verb acquisition presents a particularly rich set of phenomena touching upon issues of central concern to the disparate concepts and events literatures. To these ends, the workshop speakers represent diverse research traditions and topics, and many have contributed to multiple of these literatures.

Beth Levin, Joshua Hartshorne, and Eva Wittenberg will

discuss insights into event representations that stem from investigation of **linguistic structure**. Cynthia Fisher, Joshua Hartshorne, and Melissa Kline will discuss the grounding of **verb acquisition** in event representations, and Noah Goodman will discuss his work grounding language processing more generally in conceptual representations. Beyond this, many of the talks will focus squarely on conceptual and event representations. Many speakers will discuss recent work on **infant and child cognition**, especially Dare Baldwin and Gergely Csibra. Barbara Tversky and Jeffrey Zacks will discuss **event perception and segmentation**, and the potential role of **intuitive theories** will be discussed by Noah Goodman and Joshua Tenenbaum.

All speakers will endeavor to draw out connections between the different lines of research. Question periods and discussion during coffee breaks will allow participants and attendees to synthesize the different literatures, generating future directions for research.

Workshop Organization

The workshop will be organized around short presentations by speakers ranging from theoretical overviews to detailed discussion of specific phenomena. A final panel featuring all speakers will finish the day, with the goal of spurring discussion about promising avenues for future research and building a common vocabulary and agenda. Presentations will be geared towards an interdisciplinary audience and should be approachable by a broad cognitive science audience.

Workshop Organizers

Eva Wittenberg is a postdoctoral researcher in the Departments of Linguistics and Psychology at UCSD. Her research focuses on how the mind assembles meaning and how that capacity came to be. **Melissa Kline** is a postdoctoral researcher in the Harvard Psychology department studying how babies' and young children's cognitive representations of events relate to verb meaning and argument structure. **Joshua K. Hartshorne** is an assistant professor of Psychology at Boston College. His work focuses on the interaction between conceptual and linguistic representations.

Target Audience

The target audience for this workshop overlaps significantly with the target audience of CogSci. The workshop's central themes (language acquisition and conceptual representation) are central concerns of the Society and are typically well-represented at its meetings. The multidisciplinary nature of the work is particularly appropriate for a multidisciplinary conference like CogSci. Finally, our specific focus dovetails this year's overall conference theme: "Recognizing and representing events: Integrating psychological, linguistic, computational and neural perspectives."

Confirmed Speakers

Dare Baldwin, University of Oregon
Gergely Csibra, Central European University
Cynthia Fisher, University of Illinois, Urbana-Champaign
Noah Goodman, Stanford University
Joshua Hartshorne, Boston College
Melissa Kline, Harvard University
Beth Levin, Stanford University
Josh Tenenbaum, Massachusetts Institute of Technology
Barbara Tversky, Columbia University
Eva Wittenberg, University of California, San Diego
Jeff Zacks, Washington University in St. Louis

References

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews Cognitive Science*, *4*, 47-62.

Battaglia, P., Hamrick, J., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327-18332.

Bowerman, M. (1989). Learning a semantic system: What role do cognitive predispositions play? In M. L. Rice & R. L. Schiefelbusch (Eds.), *The teachability of language.* Baltimore: Brooks.

Clark, E. V. (2004). How language acquisition builds on cognitive development. *Trends in Cognitive Sciences*, *8*, 472-478.

Goodman, N. D., Ullman, T., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*, 110-110.

Gopnik, A., & Wellman, H. M. (in press). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and theory theory. *Psychological Bulletin*.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., Lee, M., & Snedeker, J. (in press). Psych verbs, the linking problem, and the acquisition of language. *Cognition*.

Hartshorne, J. K., O'Donnell, T. J., & Tenenbaum, J. B. (2015). The causes and consequences explicit in verbs. *Language, Cognition, & Neuroscience*, *30*, 716-734.

Hernik, M., & Csibra, G. (2015). Infants learn enduring functions of novel tools from action demonstrations. *Journal of Experimental Child Psychology*, *130*, 176-192.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (in press). Children's understanding of the costs and rewards underlying rational action. *Cognition*.

Levin, B., & Rappaport Hovav, M. (2005). *Argument realization*. Cambridge, UK: Cambridge University Press.

Levin, B., & Rappaport Hovav, M. (2011). Lexical conceptual structure. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning I.* Berlin: Mouton de Gruyter.

Tversky, B., & Zacks, J. M. (2013). Event perception. In D. Riesberg (Ed.), *Oxford handbook of cognitive psychology.* Oxford, UK: Oxford University Press.

Proposal of the Second Workshop on Physical and Social Scene Understanding

Tao Gao[1] taogao@mit.edu
Chenfanfu Jiang[2] cffjiang@cs.ucla.edu
Yixin Zhu[3] yixin.zhu@ucla.edu
Yibiao Zhao[1] ybz@mit.edu
Lap-Fai (Craig) Yu[4] craigyu@cs.umb.edu

[1]Department of Brain and Cognitive Sciences, MIT
[2]Computer Graphics and Vision Lab, UCLA
[3]Center for Vision, Cognition, Learning, and Autonomy, UCLA
[4]Graphics and Virtual Environments Lab, University of Massachusetts Boston

Keywords: Functionality; Physics; Intentionality; Causality.

Theme

Computer vision has made significant progress in locating and recognizing objects in real images. However, beyond the scope of this "what is where" challenge, it lacks the abilities to understand scenes characterizing human visual experience. The mission of this workshop is to (a) identify the key domains in which human visual perception and cognition outperform computer vision; (b) formalize the computational challenges in these domains; and (c) provide promising frameworks for solving these challenges by conducting cognitive science and computer vision studies.

We propose FPIC as four key domains for exploration beyond "what is where":

- **Functionality** (e.g., what can be done with this slotted spoon?)
- **Physics** (e.g., will the spoon be able to pick up the meatball?)
- **Intentionality** (e.g., is the person trying to scoop up the cheese or point toward it?)
- **Causality** (e.g., why does the gravy pass through the spoon?)

The combination of these largely orthogonal dimensions can span a large space for scene understanding. Despite their apparent differences, these domains do connect with each other in ways that are theoretically important: (a) events happening in these domains usually do not project onto explicit visual features; (b) existing computer vision algorithms are neither competent in these domains nor (in most cases) applicable at all; and (c) human cognition is nevertheless highly efficient in these domains. Therefore, studying FPIC should significantly fill the gap between computer vision and human vision. On the one hand, human studies on FPIC-related topics can inspire the invention of novel, cognitively-motivated computer vision systems. On the other hand, state-of-the-art computer vision systems can expand the scope of cognitive sciences to address challenges in real scenes.

The introduction of FPIC will advance cognitive models in three aspects: (a) *transfer learning*. As higher-level representation, FPIC tends to be globally invariant across the entire human living space. Therefore, learning in one type of scenes can be transferred to reason about novel situations; (b) *small sample learning*. Leaning of FPIC-related knowledge, which is consistent and noise-free, is possible even without a wealth of previous experience or "big data"; and (c) *bidirectional inference*. Inference with FPIC requires the combination of top-down abstract knowledge and bottom-up visual patterns. The bidirectional processes should boost the performance of each other as a result.

Several key themes of our proposed workshop are:

- Physically grounded scene interpretation
- Causal model of vision and cognition
- Reasoning about goals and intents of agents in scenes
- Human-object-scene interaction
- Top-down and Bottom-up inference algorithms

In conjunction with CogSci 2015, our "Physical and Social Scene Understanding" workshop will bring together researchers from cognitive science, computer vision and robotics, to illuminate cognitively-motivated vision systems going beyond labeling "what is where" in an image. These systems coordinate closely to achieve a sophisticated and coherent understanding of scenes with respect to Functionality, Physics, Intentionality and Causality (FPIC). Ultimately, these systems are expected to answer an almost limitless range of questions about an image using a finite and general-purpose model. We also want to note that FPIC is never meant to be an exclusive set of scene understanding problems. We welcome the insights of scholars who share the same perspective but are working on different problems.

Speakers

We will invite speakers working in cognitive science, computer vision, computer graphics and robotics, who have fundamental insights of visual understanding. We plan to choose eight speakers from the list below, but are not limited to.

Andrew Bagnell (Professor, Robotics, CMU)
Robotics

Noah Goodman (Professor, Cognitive Science, Stanford)
Causality and Theory of Mind

Keith Holyoak (Professor, *Cognitive Science*, UCLA)
Causal Reasoning

Josh Tenenbaum (Professor, *Cognitive Science*, MIT)
Intuitive Physics and Theory of Mind

Demetri Terzopoulos (Professor, Computer Graphics, UCLA)
AI-based computer graphics

Emo Todorov (Professor, *Robotics,* Univ. of Washington)
Physics-based planning

Felix Warneken (Professor, *Infant Cognition,* Harvard)
Social Development

Jianxiong Xiao (Professor, *Computer Vision*, Princeton)
Visual Scene Understanding

Song-Chun Zhu (Professor, *Computer Vision*, UCLA)
Causal Parsing with Commonsense Reasoning

Workshop Program

We plan to host a full day workshop consisting of talks given by eight invited speakers who are leading researchers in their research fields.

- Each speaker will have 35 minutes to present.
- One-hour panel discussion at the end.
- All talks and discussions will be video recorded and posted on our workshop website.

Tentative Schedule:

9:00am - 9:10am Welcome speech
9:15am - 9:45am Invited talk 1
9:55am - 10:25am Invited talk 2
10:30pm - 11:00 am Invited talk 3
11:05pm - 11:35 am Invited talk 4
11:40am - 1:00pm Lunch Break
1:00pm - 1:30pm Invited talk 5
1:35pm - 2:05pm Invited talk 6
2:10pm - 2:40pm Invited talk 7
2:45pm - 3:15pm Invited talk 8
3:20pm - 4:20pm Panel Discussion

Potential Financial Support

We are currently looking for sponsorship from the Center of Brian, Mind and Machine (CBMM) at MIT. We are planning to get support from the sponsors (e.g., Microsoft Research, A9 and Huawei) of our previous workshops again. The sponsorship will be used for covering the travel fees of some of our invited speakers; making souvenirs for our contributors and workshop participants; recording a video for all of our talks. With this support, we aim at organizing a workshop that every participant enjoys, further boosting the impact of our workshop and the CogSci conference.

Success of Our Previous Workshops

We held the first workshop on Physical and Social Scene Understanding at CogSci 2015. We also held related workshops on "Vision meets Cognition" at CVPR 2014 and 2015 (a premiere computer vision conference). Our workshops were highly successful and very well-received. This reflects the strong enthusiasm towards recent cognitive science studies in the computer vision community. The following is a brief summary of our previous workshops. We are dedicated to continue the success at CogSci 2016.

Talks: We successfully held 24 keynote talks in our three workshops, given by top experts in cognitive science, computer vision and computer graphics. These talks provided diverse insights from different perspectives which are all highly relevant to the theme of our workshop: Vision meets Cognition. The slides and videos of these talks were posted on our workshop websites:
http://www.visionmeetscognition.org/fpic2014/
http://www.visionmeetscognition.org/fpic2015/
http://www.visionmeetscognition.org/cogsci2015/

Audience: Our workshops were very well-received. For example, in CVPR 2014, there were 367 conference participants who signed up for our workshop during registration. Our keynote talks were very popular and most of the time our room was fully seated. At peak time, our workshop attained attendance of over 200 participants.

Accepted papers: There were 38 carefully peer-reviewed papers accepted by our workshop held in CVPR 2014, which involved more than 200 paper authors. Each paper was reviewed by 2-4 experts in the field chosen among our 30 program committee. We broadcasted a trailer video composed of spotlight slides to promote all of our accepted papers.

Sponsors: Our workshops have also aroused significant industry interests and were generously supported by industrial sponsors: Microsoft Research, Amazon A9, Vicarious, Google DeepMind, Huawei and the Office for Naval Research.

Souvenirs: All of our invited speakers received our custom-designed souvenirs in recognition of their contribution. In CVPR 2014, we also designed and manufactured 200 magic mugs for all our workshop attendants and guests.

Based on our rich organizing experience, we are highly confident that the workshop of "Physical and Social Scene Understanding" at CogSci 2016 can achieve an even bigger success.

Tutorial Workshop on Contemporary Deep Neural Network Models

James L. McClelland (mcclelland@stanford.edu)
Steven Hansen (sshansen@stanford.edu)
Andrew Saxe (asaxe@fas.harvard.edu)
Department of Psychology, Stanford University
Stanford, CA, 95305, USA

Keywords: deep learning; neural networks; unsupervised learning; computational modeling; reinforcement learning, simulation tools

Contemporary Deep Neural Networks

Neural network models were exploited in the late 80's and 90's to model human cognition, based on developments such as the back propagation learning algorithm. Over the last 10 years, massive data sets, enhanced computational resources, and new developments in algorithms have led to explosive growth in use of such models in machine learning and artificial intelligence. We propose a workshop to bring the tools, methods, and insights from this research into contemporary cognitive science and cognitive neuroscience.

Here we mention three relevant recent architectures: Deep Neural Networks (DNNs), Deep Reinforcement Learning Networks (DRNs), and Recurrent Neural Networks with Long-Short-Term Memory Units (RNN-LSTMs). Each is a rich architectural framework rather than a single fixed model and each has been used on a different class of problem. DNNs (including convolutional neural networks, CNNs) are trained using a combination of supervised and unsupervised learning methods, and are used in vision, object recognition and other tasks including numerosity judgment; DRNs learn optimal action policies in video games and other action selection settings; RNN-LSTMs have been used for cross-domain mapping, either between different spoken languages or (combined with DNNs) between video and spoken language and for creating provocative new cognitive models such as the Neural Turing Machine (NTM). Together the architectures incorporate many key tools and concepts, including convolution, unsupervised learning and unsupervised pre-training, enhanced methods for reinforcement learning, contemporary regularization methods (to prevent overfitting), and the LSTM mechanism—a mechanism that learns to gate information into and out of buffers to maintain context over long sequences. These and related architectures, applications, tools, and concepts will be the focus of our workshop.

Presenters and their Contributions

Literature citations highlighting the work of the presenters in the areas described above are contained in the bibliography below. Several of the authors have released preprints on *arXiv* of papers first presented at top Machine Learning conferences (NIPS, CVCPR) prior to publication in journals due to the fast-moving nature of research in this field. Here we mention briefly each presenter's expertise and background.

Marco Zorzi, Professor, University of Padova (PhD in Cognitive Science, University of Trieste, 1999). Has applied deep networks to modeling human numerosity judgment and reading and has developed tools for efficient implementation of these models.
Nikolaus Kriegeskorte, Program Leader, Memory and Perception Group, MRC-CBU. Has applied a deep convolutional neural network architecture introduced in 2012 to model human voxel-level activity patterns in different layers of visual cortex.
Timothy Lillicrap, Senior Research Scientist, Google DeepMind, London UK (PhD in Systems Neuroscience, Queen's University, 2012). Leading the effort to extend Deep Reinforcement Learning approaches that allow simulated agents to learn sophisticated continuous motor control policies.
Greg Wayne, Research Scientist, Google DeepMind, London, UK (PhD in Neuroscience, Columbia University, 2013). One of the creators of the Neural Turing Machine, a Deep Learning Model that relies on the LSTM mechanism for the storage and retrieval of information in memory.
Alberto Testolin, Post-Doctoral Fellow, University of Padova. Testolin is a co-author with Zorzi on several deep learning papers including tutorial articles.

Presenter Contributions

Organizer McClelland will provide historical context and an overall perspective on the role of deep neural networks in contemporary cognitive science, and Zorzi, Kriegeskorte, Lillicrap, and Wayne will each lead tutorials on unsupervised DNN's; CNN's; DQN's and other reinforcement learning methods; and RNN's including LSTM's and other methods that allow networks to span very long distance dependencies. These tutorials will cover the basic network paradigm and architectures, algorithms, applications, and best practices.

Testolin, Kriegeskorte, Lillicrap, and Wayne will each host breakouts in the afternoon in which participants will have to opportunity to run example neural networks of each type. During these sessions participants will try out existing networks and begin to develop their own plans guided by the tutorial leaders. Advanced preparation for this will be coordinated by co-Organizers Hansen, a PhD student in

McClelland's lab who has spent 4 months as an intern at DeepMind, and Saxe, formerly of McClelland's lab and now a post-doc in computational neuroscience at Harvard. Hansen and Saxe will ensure that the relevant software tools are available for remote use and/or download to Workshop Participant's laptops. Their expertise will also allow them to contribute to helping students formulate specific proposals for their own models and to configure software for their implementation.

Workshop Structure

The morning session will begin with a brief introduction by McClelland (10 min), followed by 40 minute presentations by Zorzi, Kriegeskorte, Lillicrap, and Wayne, with a 10 minute break after Kriegeskorte's presentation. The afternoon session will begin with a brief overview of available software tools by Lillicrap with input from others. This will be followed by a description of the specific software tools and demonstration simulations prepared for the workshop, and short descriptions of the focus of each of the four breakout sessions. During the breakouts, workshop participants will discuss with presenters/breakout leaders possible applications they might develop themselves and will begin to work on implementation. Starting at about 3pm, a subset of participants from each group will describe their ideas to all the workshop participants, allowing for all presenters to offer feedback and for broad sharing of ideas. A final Q&A session with the presenters as a panel will round out the session, which will end at 4 pm.

Breakouts will be organized in different corners of the allocated conference room. If there is sufficient interest and available space, part of afternoon session might use a second room; in that case, one room might be allocated to DNN's and CNN's, and the other to reinforcement learning and recurrent networks, depending on the interests of participants. We hope to coordinate on this with the Cognitive Science meeting program committee as participant interest becomes better defined.

Budget

We request $1,200 toward support for registration fees (including Workshop registration). Although registration cost information is not currently available, we expect that this amount would cover approximate 4 participants who can document inability to cover costs of registration.

Other Funding and Coordination Information

Some funding for this tutorial workshop will be provided by the Rumelhart Emergent Cognitive Functions fund created by the proceeds from McClelland's Rumelhart Prize. This will be used to defray presenter's registration for those presenters who will not otherwise be participating in the Cognitive Science Society Meeting, as well as one hotel night for presenters.

We also note that McClelland, with Stefan Frank of Radboud University, Netherlands, and CogSci 2016 program committee member Daniel Mirman will be hosting a related longer co-located workshop (Neural Computation and Psychology – NCPW 2016) at Drexel University in Philadelpia on the 8th and 9th of August, just prior to this Cogsci Tutorial workshop. The NCPW workshop, supported by the W.K. and K. W. Estes Fund, the Rumelhart Fund, and a gift from Google DeepMind, will cover the travel and accommodation for all of the listed presenters, who will also be participating in the earlier workshop. Their presentations in the NCPW workshop will focus on concepts and research results; their presentations in this CogSci tutorial workshop will focus on algorithms, best practices, and implementation.

References

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv preprint arXiv:1410.5401*.

Heess, N., Wayne, G., Silver, D., Lillicrap, T., Erez, T., & Tassa, Y. (2015). Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems* (pp. 2926-2934).

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Computational Biology, 10(11). E1003915.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Silver, D, Huang, A., Maddison, C. J., … Lillicrap, T., … & Haassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484-489.

Stoianov, I., & Zorzi, M. (2012). Emergence of a'visual number sense'in hierarchical generative models. *Nature Neuroscience, 15(2),* 194-196.

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in psychology*, *4*.

Workshop on Corpus Collection, (Semi)-Automated Analysis, and Modeling of Large-Scale Naturalistic Language Acquisition Data

Elika Bergelson (elika.bergelson@gmail.com)
Brain and Cognitive Sciences, University of Rochester, Rochester, NY, 14627 USA
Department of Psychology & Neuroscience, Duke University, Durham, NC, 27708 USA

Keywords: language acquisition; automatic speech recognition; computational models; speech corpora

Introduction

The main goal of this full-day workshop is to bring together researchers from several distinct fields: behavioral psychologists studying language acquisition, speech technology researchers, linguists, and computational modelers of cognitive development. These groups are broadly interested in the same questions, i.e. what is the nature of speech and language, and how might a system learn to process it in supervised or unsupervised ways? Since the groups interested in these questions work on different analysis levels, cross-pollination has been sparse.

Recent technological innovations have made collecting long naturalistic recordings of children's home environment far simpler than in the past. However, the raw output of such recordings is not immediately usable for most analyses. Simultaneously, speech technology (ST) and machine learning tools have improved immensely over the past decade, making it feasible to use such tools with increasingly diverse and noise-laden data. Relatedly, cognitively viable computational models have made recent strides in explaining learning and development, but few such models can be applied to novel data-sets without encountering many hurdles about translatability across frameworks. This workshop brings together experts from all of these areas, and seeks to build bridges across them, with insight from other similar interdisciplinary efforts in other areas of cognitive science. Talks will discuss the match between the theory-driven questions researchers would like to ask, and the answers the current state of the art allows.

The program committee is part of a newly formed group called DARCLE (Daylong Audio Recordings of Children's Language Environment); with the help of an NSF grant, DARCLE has created a repository called HomeBank for raw data, metadata, and analysis/processing tools for long-form recordings of child language. This workshop is an opportunity to network with related efforts in Europe, and for a talk and demo of a related effort, the NSF-funded Speech Recognition Virtual Kitchen.

Workshop Organization

Target Audience

The target audience of this workshop is researchers (from students to PIs) with an interest in collecting, analyzing, and modeling language data. It will be especially useful for participants with background in speech and hearing, computer science, linguistics, or psychology, but will be geared to a non-specialist audience.

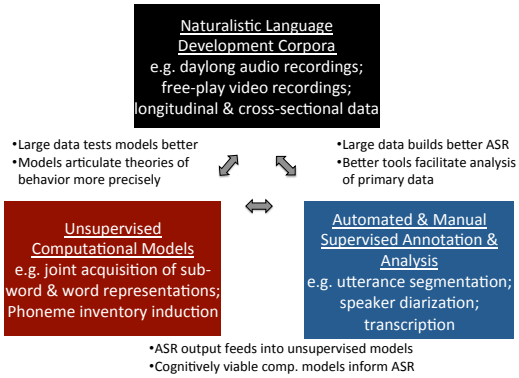


Figure 1: Schematic of Relevant Areas and Questions.

Workshop Format

The workshop comprises ten confirmed speakers in four sets of talks, with interspersed question periods and breaks. Ten institutions from four countries are represented. Recordings will be available on the workshop website. The workshop features experts across a range of disciplines, and provides an opportunity for junior researchers to solicit feedback.

Workshop Organizer and Program Committee

Elika Bergelson is a Research Professor in the University of Rochester's Brain & Cognitive Sciences Department, and an incoming Assistant Professor in Duke's Psychology & Neuroscience Department.

The program committee includes PIs of the HomeBank grant, and the DARCLE board: Elika Bergelson (Rochester/Duke), Alex Cristia (LSCP), Emmanuel Dupoux (LSCP), Brian MacWhinney (CMU), Melanie Soderstrom (Manitoba), Mark VanDam (Washington State), and Anne Warlaumont (UC-Merced).

Summary of Presentation Topics

Speech and Language Development Corpora

Anne Warlaumont After introducing HomeBank, Warlaumont will present a case example of a computational model that uses statistics obtained from daylong home audio recordings of children to provide an account for how growth in speech-related vocalizations varies across groups with

differing clinical and socioeconomic status. She will also discuss how more detailed computational modeling examining specific speech sound production could benefit from daylong home recording data. Finally, she will propose ST challenges that need to be overcome in order to maximize synergy between computational modeling and daylong home recording.

Elika Bergelson Bergelson will discuss aspects of data collection and analysis of daylong audio and hour long video recordings from her longitudinal corpus, SEEDLingS. Specifically focusing on object word acquisition from 6-18 months, Bergelson will discuss links between eyetracking data and home environment, and discuss which aspects of the variability in the input map onto variance in word comprehension and production in the first two years of life.

Kim Oller Oller will articulate his approach to large scale analysis of vocal data from infants. He will highlight cautions about coding infant vocalizations in terms of adult sound categories, emphasizing the fact that in the vast majority of cases in the first year infant utterances are neither phonemic nor lexical. Oller will thus underscore the challenge for computational analysis of infant vocal data: finding ways to monitor infant categories as they are progressively transformed from primitive or discoordinated phonatory/articulatory sequences into mature speech.

Annotation and Automatic Speech Recognition

Reiko Mazuka Mazuka will introduce a corpus of Japanese infant-directed speech (Riken-Japanese Mother-Infant Conversation Corpus), which is fully annotated, and time-aligned both for segmental and intonational details. She will discuss the challenges inherent in attaining expert phonetic annotation for a sizable naturalistic speech corpus, and the necessity of doing so for certain research questions.

Florian Metze Metze will introduce the Speech Recognition Virtual Kitchen, which is dedicated to improving community research and education infrastructure in speech technology. The goal of this work is to allow non-experts to access state of the art ST and Automatic Speech Recognition tools. The preconfigured virtual-machine-based "kitchen" provides the infrastructure for "appliances" (e.g., speech recognition toolkits), "recipes" (scripts for creating state-of-the art systems), and "ingredients" (language data).

Metze will introduce these resources, including demonstrations on users' machines. He will highlight how the "kitchen" can facilitate the development of audio related algorithms in research on child language acquisition using large corpora.

Computational Modelers of Cognitive Development

Okko Räsänen Räsänen will discuss challenges and opportunities in using large-scale audio/audio-visual recordings for computational models of speech perception and early language acquisition (LA). He will focus on how such data could enable a transition from the study of isolated aspects of LA to the development of integrated models (e.g., joint acquisition of sub-word and word representations, including semantic grounding of words to referential contexts). Räsänen will also discuss computational model evaluation, the related requirements this imposes on the metadata/annotation of recordings, and the challenges associated with the use of linguistic representations, such as phonemes or words, as proxies for learning targets in early stages of development.

Tove Gerholm Gerholm will discuss her group's longitudinal study of parent-child interaction from 3 months to 3 years of life. This video-corpus based work seeks to look at modalities across language acquisition, build a reusable corpus, and model interaction during development. Her talk will focus on an expansion of models that map sound strings to objects by adding in tactile and gestural information, and caregiver feedback. Gerholm will discuss model assumptions and highlight her groups modeling efforts that have met both failure and success, helping others with similar research goals.

Synthesis

Early Stage Work There will be a group discussion allowing early-career participants to describe ongoing or planned work, and solicit audience feedback.

Caitlin Fausey Fausey will briefly review findings from her work about the distributions of activities, contexts, and visual instances over the first two years of everyday life at home, addressing relevant inferential issues for learners building multimodal links among language, vision, and action. Fausey will raise discussion points about how coarse- and fine-grained daily rhythms interact in language learning, and how these activity-based rhythms change over the first two years. Finally, she will discuss potential links between distributions of daily events and ST performance.

Mark Liberman Liberman will discuss the present state and future prospects of corpus-based methods in speech and language science. Key questions include infrastructure for sharing data (including annotations of various types), the inventory of tools (and what's missing), and the skills researchers will need in order to participate fully in this area's trans-disciplinary future. Finally, he will discuss some examples, obvious and not so obvious, of the research opportunities created by new social and technological developments.

Acknowledgments

This work was supported by an NIH grant to Bergelson (DP5-OD019812-01), a collaborative NSF grant awarded to Anne Warlaumont (1539129), Mark VanDam (1539133) and Brian MacWhinney (1539010), and an NSF grant to Florian Metze (1305365).

Interactive spatiotemporal cognition:
Data, theories, architectures, and autonomy

Sangeet Khemlani and J. Gregory Trafton
{sangeet.khemlani, greg.trafton}@nrl.navy.mil
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Washington, DC 20375

Keywords: spatiotemporal cognition; interactive inference; context; pragmatics; gesture; architectures

Introduction

Everyday interactions often depend on thinking about space and time: collaborators need to know where events take place – and in what order – to, e.g., communicate driving directions, build pieces of furniture, or carry out strategic operations in military and sports settings (Núñez & Cooperrider, 2013). A simple set of driving directions may require a listener to interpret and reason about the spatial relations – such as *next to* and *behind* – and the temporal relations – such as *after* and *during* – that a speaker describes. The speaker may also use gestures to substitute, supplement, or disambiguate linguistic descriptions (Holle & Gunter, 2007; Perzanowski, Schultz, & Williams, 1998). Such rapid, rich, and productive interactions are transient and difficult to analyze behaviorally, and so they pose a challenge for experimenters. They are grounded in the physical world, and accordingly challenge computational models that cannot digest rich perceptual and environmental input in real time. Robotic systems are geared towards processing and acting upon the physical world – and they increasingly support human-robotic interaction (e.g., Fong et al., 2006; Kawamura et al., 2003; Kortenkamp et al., 1999). But they, too, are uniquely challenged in maintaining productive interactive exchanges with human teammates, because they must be tolerant of human idiosyncrasies, preferences, limitations, and errors (Trafton et al., 2013).

Because these challenges cut across broad interests in cognitive science – such as linguistics, artificial intelligence, robotics, and psychology – progress is unlikely without the engagement of multiple approaches, from psychological experimentation to the construction of autonomous, embodied systems. In recent years, progress towards understanding interactive spatiotemporal cognition has accelerated along parallel paths: there exist new behavioral and imaging methodologies to study event segmentation (e.g., Radvansky & Zacks, 2014), spatial inference (e.g., Knauff & Ragni, 2013), and gestural cognition (e.g., Novack et al., 2016); novel computational theories of understanding physical reasoning (e.g., Battaglia et al., 2013) and mental simulation (e.g., Khemlani & Johnson-Laird, 2013); cognitive architectures that support rich interactivity (Huffman & Laird, 2014; Trafton et al., 2013); and a wide variety of technological platforms on which to transform theory into embodied interaction.

The goal of the workshop is to allow these parallel approaches to converge. Discussants will share recent data and theory, consider novel architectural approaches, and demonstrate burgeoning technological advances that advance the science of spatiotemporal inference. The workshop will promote interdisciplinary collaboration by focusing on three unifying themes.

Spatiotemporal cognition and interaction

The workshop will focus on three general themes that draw from and synthesize varying approaches and methodologies in the cognitive sciences. We summarize each theme below:

The representation of time and space

People represent space and time through a variety of cognitive and neural structures (e.g., Bonato et al., 2012; Gibbon & Church, 1990; Schaeken, 1996); and computer scientists and roboticists have proposed efficient methods of extrapolating spatial and temporal structure from the external environment (e.g., Tang, Fei-Fei, Koller, 2012). This set of talks will leverage recent advances in cognitive neuroscience, computational modeling, and artificial intelligence to make progress towards efficient, domain-general representations of time and space.

Embodied communication

Gestures are often communicative (e.g., Hostetter, 2011) and can improve thinking across many tasks (e.g., Alibali et al., 2011; Bucciarelli et al., 2016; Chu & Kita, 2011; Ehrlich et al., 2006). In particular, gestures appear to be outward signs of mental simulations (Hegarty et al., 2005; Hostetter & Alibali, 2008, 2010). How do gestures help build, refine, or disambiguate mental representations? This set of talks builds upon the previous theme to understand how physical interactions depend on – and facilitate – the construction of representations of space and time.

Architecture and inference

Inferences about space and time occur in the context of building representations from environmental input and using them to act upon the world, e.g., by issuing motor commands to carry out particular gestures. Cognitive and robotic architectures allow researchers to build robust systems that connect psychological theories of

spatiotemporal representation and reasoning with interactions in the physical world. This set of talks brings together researchers whose goal is to build integrated architectures for spatiotemporal reasoning systems.

Workshop structure

The workshop will include three sets of twenty-five minute presentations (including Q&A) organized around three central themes: a) representing space and time; b) gestural and embodied communication; and c) architectures for inference and interaction. Each presentation will conclude with a panel discussion. Presentations will include overviews of theoretical and computational models, discussions of recent experimental phenomena, descriptions of novel technologies, and discussions of setting an agenda for future research.

Organizers

Sangeet Khemlani is a cognitive scientist (Ph.D., Princeton University) at the Naval Research Laboratory in Washington, DC. His research focuses on the development of a unified cognitive model of deductive, probabilistic, kinematic, and temporal reasoning. His recent work focuses on integrating theories of event segmentation of temporal inference (Khemlani et al., 2015).

Greg Trafton is a cognitive scientist (Ph.D., Princeton University) at the Naval Research Laboratory in Washington, DC, who studies cognitive robotics and human robot interaction. His expertise is in building high fidelity computational cognitive models that highlight an aspect of human cognition and then putting those models on embodied platforms so that they take a person's fallacies or weaknesses into account and use that information to either compensate or facilitate the person's goals and actions.

Target audience

The workshop is targeted to a broad spectrum of researchers and practitioners in the cognitive sciences. Because it incorporates several different themes, from computational modeling, to gesture and embodied cognition, to high-level inference, it is appropriate to many disciplines within cognitive science, including psychologists, computer scientists, linguists, and roboticists. Furthermore, the workshop proposes a multidisciplinary approach towards understanding the nexus of interaction and spatiotemporal cognition.

References

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science, 22,* 1138–1144.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327-18332.

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: Evidence for a mental time line. *Neuroscience and Biobehavioral Reviews, 36,* 2257-2273.

Bucciarelli, M., Mackiewicz, R., Khemlani, S., & Johnson-Laird, P.N. (2016). Children's creation of algorithms: simulations and gestures. *Journal of Cognitive Psychology.*

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. Journal of Experimental Psychology: General, 140, 102–116.

Ehrlich, S. B., Levine, S., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology, 42,* 1259–1268.

Fong, T., Kunz, C., Hiatt, L. M., & Bugajska, M. (2006). The human-robot interaction operating system. In *Proceedings of Human-Robot Interaction.*

Gibbon, J., & Church, R. M. (1990). Representation of time. *Cognition*, *37*, 23-54.

Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation, 5,* 333–356.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of cognitive neuroscience*, *19*, 1175-1192.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137,* 297–315.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15,* 495–514.

Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the gesture as simulated action framework. *Journal of Memory & Language, 63,* 245–257.

Huffman, S. B., & Laird, J. E. (2014). Learning procedures from interactive natural language instructions. In *Proc. 10th International Conference on Machine Learning. Amherst, MA* (pp. 143-150).

Kawamura, K., Nilas, P., Muguruma, K., Adams, J. A., & Zhou, C. (2003). An agent-based architecture for an adaptive human-robot interface. In *Proceedings of the 36th Hawaii International Conference on System Sciences.*

Khemlani, S., Harrison, A. M., & Trafton, J. G. (2015). Episodes, events, and models. *Frontiers in Human Neuroscience*, 9, 590, 1-13.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, 4, 1-20.

Kortenkamp, D., Burridge, R., Bonasso, R. P., Schreckenghost, D., & Hudson, M. B. (1999). An intelligent software architecture for semiautonomous robot control. In *Autonomy Control Software Workshop, Autonomous Agents 99* (pp. 36–43).

Novack, M., Wakefield, E. & Goldin-Meadow, S. (2016). What makes a movement a gesture? *Cognition, 146,* 339-348.

Núñez, R., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in cognitive sciences*, *17*, 220-229.

Perzanowski, D., Schultz, A. C., & Adams, W. (1998). Integrating natural language and gesture in a robotics domain. In *Intelligent Control (ISIC) 1998* (pp. 247-252). IEEE.

Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition.* Oxford University Press.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review, 120,* 561-588.

Schaeken, W., Johnson-Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, *60*, 205-234.

Tang, K., Fei-Fei, L., & Koller, D. (2012, June). Learning latent temporal structure for complex event detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 1250-1257). IEEE.

Trafton, J. G., Hiatt, L.M., Harrison, A. M., Tamborello, F.P., II, Khemlani, S., & Schultz, A.C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human Robot Interaction*, 2, 30-55.

Full Day Tutorial on Quantum Models of Cognition and Decision

Jennifer S. Trueblood (jennifer.s.trueblood@vanderbilt.edu)
Department of Psychology, Vanderbilt University, Nashville, TN 37240 USA

James M. Yearsley (james.m.yearsley@vanderbilt.edu)
Department of Psychology, Vanderbilt University, Nashville, TN 37240 USA

Peter Kvam (kvam.peter@gmail.com)
Department of Psychology, Michigan State University, East Lansing, MI 48824 USA

Zheng (Joyce) Wang (wang.1243@osu.edu)
School of Communications, Center for Cognitive and Brain Sciences
The Ohio State University, Columbus, OH 43210 USA

Jerome R. Busemeyer (jbusemey@indiana.edu)
Cognitive Science, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405 USA

Keywords: classical information processing; quantum information processing; logic and mathematical foundation; Bayesian probability; quantum probability; Markov and quantum processes; decision making; quantum entanglement; memory.

General Purpose

This ***full day*** tutorial is an exposition of a rapidly growing new alternative approach to building computational models of cognition and decision based on quantum theory. The cognitive revolution that occurred in the 1960's was based on classical computational logic, and the connectionist/neural network movements of the 1970's were based on classical dynamical systems. These classical assumptions remain at the heart of both cognitive architecture and neural network theories, and they are so commonly and widely applied that we take them for granted and presume them to be true. What are these critical but hidden assumptions upon which all traditional theories rely? Quantum theory provides a fundamentally different approach to logic, reasoning, probabilistic inference, and dynamical systems. For example, quantum logic does not follow the distributive axiom of Boolean logic; quantum probabilities do not obey the disjunctive axiom of Kolmogorov probability; quantum reasoning does not obey the principle of monotonic reasoning. It turns out that humans do not obey these restrictions either, which is why we consider a quantum approach.

This tutorial will provide an exposition of the basic assumptions of classical versus quantum theories. These basic assumptions will be examined, side-by-side, in a parallel and elementary manner. We will show that quantum theory provides a unified and powerful explanation for a wide variety of paradoxes found in human cognition and decision ranging from attitude, inference, causal reasoning, judgment and decision, and memory. This tutorial introduces and trains cognitive scientists on this promising new theoretical and modeling approach.

Presenters

Jennifer Trueblood is an assistant professor at Vanderbilt University. She has published many articles on the topic of quantum cognition, and her work has been funded by NSF. James Yearsley is a postdoctoral fellow at Vanderbilt University. He has a PhD in the foundations of quantum theory from Imperial College, London and worked in the Centre for Quantum Information and Foundations at the University of Cambridge. Peter Kvam is a graduate student at Michigan State who has published many articles on quantum cognition including in top journals such as PNAS. Zheng (Joyce) Wang is an associate professor at The Ohio State University. She was Co-Editor for a special issue on quantum cognition that appeared in *Topics in Cognitive Science* (2013), Vol. 5 (4)). Her work on quantum cognition has been funded by NSF and AFOSR. Jerome Busemeyer is Provost Professor of Psychological and Brain Sciences at Indiana University. He is Editor of *Decision* and Associate Editor of *Psychological Review*, and was Editor of *Journal of Mathematical Psychology*. He is also author with Peter Bruza of the book *Quantum models of Cognition and Decision*.

Previous Tutorials and Symposia

The tutorial has been presented at the Cognitive Science meetings in Nashville (2007), Washington DC (2008), Amsterdam (2009), Sopporo (2012), Berlin (2013), Quebec City (2014), and Pasadena (2015) with about 30 to 50 participants each time. The ratings from participants after the tutorial were all very positive. Also, this tutorial follows a symposium on quantum cognition at the Cognitive Science meeting 2011 whose papers appeared as a special issue in *Topics in Cognitive Science* (2013).

Participants Background

This tutorial will introduce participants to an entirely new area and no previous experience or background with quantum theory will be assumed. ***No background in physics is required.*** In fact, except for a few simple examples to motivate the idea, little or no reference to physics will be made during main part of the tutorial. What is required is an elementary background in classical logic and probability.

Material to be Covered

1. Introduction and Background (1.5 hours). First, we will examine major differences between classical versus quantum theories of probability. The concept of superposition is introduced and distinguished from classical probability mixtures. The important issue of measurement in classical and quantum systems will be compared and examined. We will include several dramatic empirical examples illustrating empirical violations of the classical laws of probability (e.g., conjunction, disjunction, and total probability) and the parsimonious explanation of all these violations by quantum theory.

Then we will examine the differences between classical and quantum dynamical systems. The basic idea of a Markov processes will be introduced and compared with quantum processes. A parallel development of Markov and quantum processes will be shown using a concrete empirical example. The concept of a state will be distinguished for Markov and quantum systems. The effects of measurement on the state of the system are compared for Markov and quantum systems. A key goal is to show when and how quantum processes depart from Markov processes, and how we can empirically test whether a system is Markov or quantum.

2. Quantum logic and heuristics (1 hour). Cues indicating the state of the world play a critical role in decision-making in both inferential and preferential tasks, and are the focus of many heuristic models of cognitive processes. In this section, we present the formal logical structure that is used by most classical information processing models, including fast and frugal heuristics. We review the structure of these heuristics and show that they make a number of implicit assumptions arising from their reliance on a classical binary logic of bits and logic gates. However, many of these assumptions are inconsistent with empirical data from decision tasks, suggesting that there is much to be gained by revising the structure of heuristic models.

As an alternative, we introduce quantum logic and show that it addresses many of the issues arising in classical logic models. We then demonstrate how several fast and frugal heuristics can be reconstructed by integrating them with a quantum logic structure, introducing qubits, U-gates, and state evaluation to model how information is processed when these strategies are executed. This approach opens up a number of new questions and predictions, which we address by reviewing existing literature on expertise, game theory, recognition memory, decision making under uncertainty, and the hindsight bias. The results suggest that integrating heuristics with a quantum logic structure can enhance the empirical accuracy of heuristics as well as ground quantum logic in psychological theory by giving it specific processing rules.

3. Implementing Quantum Models using Bayesian Statistics (2 hours). In this section, we will provide hands-on experience with an easy to use computer program (JAGS) that will allow you to implement quantum models in a Bayesian framework. We will present the details of classical and quantum models of causal reasoning and illustrate how Bayesian modeling can be used to fit the models. At the end of this section you will have gained the technical skills to implement quantum models.

4. Future Directions (30 minutes). Finally, we will review progress in quantum cognition research and propose future directions.

See the references and the website below for some of the material to be covered and relevant background material: http://mypage.iu.edu/~jbusemey/quantum/Quantum Cognition Notes.htm

References

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge, UK: Cambridge University Press.

Busemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment 'errors.' *Psychological Review*, *118*, 193-218.

Kvam, P. D., Pleskac, T. J., Yu, S., & Busemeyer, J. R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *PNAS*, 112(34), 10645–10650.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral & Brain Sciences*, *36*, 255-274.

Trueblood, J. S. & Busemeyer, J. R. (2011). A Quantum probability account of order effects in inference. *Cognitive Science*, *35*, 1518-1552.

Wang, Z., Busemeyer, J. R., Atmanspacher, H., & b Pothos, E. M. (2013). The potential of using quantum theory to build models of cognition. *Topics in Cognitive Science*, *5*, 672-688.

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *PNAS*, *111*(26), 9431-9436.

Yearsley, J.M. and Pothos E.M. (2014). Challenging the classical notion of time in cognition: a quantum perspective. *Proc. R. Soc. B* 281, 1781, 20133056.

Acknowledgments

This tutorial is supported by U.S. National Science Foundation (SES-1153726, SES-1153846, SES-1556415, DGE-142487) and AFOSR (FA9550-15-1-0343).

Wallace: Automating Cultural Evolution Experiments Through Crowdsourcing

Jordan W. Suchow (suchow@berkeley.edu)
Thomas J. H. Morgan (thomas.morgan@berkeley.edu)
Jessica B. Hamrick (jhamrick@berkeley.edu)
Michael D. Pacer (mpacer@berkeley.edu)
Stephan C. Meylan (smeylan@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)

Department of Psychology,
University of California, Berkeley

Keywords: cultural transmission, crowdsourcing, iterated learning, research software

Significance of the topic

The experimental study of cultural evolution, social learning, cooperation, and collective decision-making asks fundamental questions about our capacities to learn, decide, and communicate in a world that is shared with other people. Experimental studies of cultural evolution have revealed a wealth of findings, including how structured forms of communication emerge from individual learning and decision-making (Verhoef, Kirby, & Padden, 2011; Claidière, Smith, Kirby, & Fagot, 2014), the inductive biases underlying human decision making (Griffiths et al., 2008), how innovations accumulate in populations to produce technologies that go beyond what any one individual could create (Caldwell, & Millen, 2008; Derex & Boyd, 2015), and how the mode of communication affects transmission and acquisition of new skills (Morgan et al., 2015).

However, in-laboratory experiments of this kind are resource intensive and logistically complex, requiring recruitment and coordination of participants to perform tasks sequentially and in concert, with enough space and time to isolate and control their interactions. These requirements drive experimental designs towards simple network structures (such as the transmission chain), small groups, and limited interaction between participants. Even where such experiments can be carried out using computers, the complexity of each experiment often means that existing software is unsuitable, leading each researcher to build bespoke software for their particular experiment. In addition to slowing the rate at which such experiments can be carried out, this also makes it hard to share code, to replicate other's experiments, and to build off the work of others.

To address these issues, we created a software-based tool for orchestrating cultural transmission using online crowdsourcing. Our tool, named *Wallace*, builds on psiTurk (Gureckis et al., 2015) to provide efficient high-throughput automation for running behavioral experiments involving cultural transmission. Wallace recruits participants, obtains their informed consent, arranges them into a network, coordinates their communication, records the data they produce, pays them, and validates and manages the resulting data. Wallace runs on commodity hardware and cloud platforms, uses a custom API, and uses widely supported languages and markup languages such as Python, HTML5, JavaScript, and CSS. It is released as open-source software under the permissive MIT license.

Wallace is modular and includes a library of components that can be used to quickly create new experiments. Prepackaged network structures include linear chains (e.g., Bartlett, 1932), scale-free networks (e.g., Bednarik et al., 2014), star and burst formations, micro-society (e.g. McElreath et al., 2005), and the discrete generational structure of the Wright–Fisher model from population genetics (Wright, 1931; Fisher, 1930), among others. Prepackaged behavioral tasks include story recall, category learning, function learning, magnitude estimation, a public goods game, stimulus–response mapping, and numerosity judgment. Nonetheless, experiments can also use custom network structures, processes, and tasks, which can be built by modifying the provided templates, allowing experimental designs of arbitrary complexity.

Structure of tutorial and activities

The tutorial will include a mix of presentations, demonstrations, and hand-on activities with the goal of giving attendees enough knowledge to be able to use Wallace to run their own studies and to understand the new scale of experimentation that Wallace makes feasible.

The tutorial will begin with a 60 minute presentation on cultural evolution experiments, describing the intellectual history of the approach, common experimental designs, and the most notable results produced in these paradigms. This will introduce attendees to the sorts of questions Wallace is designed to help answer.

After this, we will introduce attendees to Wallace itself. We will provide a demonstration replication of Bartlett's early experiment using serial reproduction, illustrating the process of running an experiment with Wallace. We will then guide attendees through the installation of Wallace on their machines such that they can run the same replication on their own systems.

Next, we will have a short presentation describing the software's architecture at a conceptual level. This will start with the key object classes around which Wallace is based as well as the key methods involved in experimental design.

Building on this knowledge, we will then return to the example experiment to see how these classes and methods are used to create experiments. We will discuss potential changes to the experimental design and how to choose which experiment to modify. A demonstration of how to go from an existing example experiment to a new experiment will follow. Participants will be given time to implement some simple extensions, such as changing the network structure or altering the kind of information transmitted.

The final tutorial section will teach attendees how to deal with unexpected events, including details of the inner workings of the code base, how to pause Wallace mid-experiment, how to access and repair the database and how to get support from Wallace's development team. We will also show attendees how they can contribute to the development of Wallace.

The day will conclude with an open discussion section where we will cover issues that came up during the day, experimenter best practices, and how Wallace compares with other existing experimental platforms.

Credentials of organizers

This tutorial will be led by the five members of Wallace's core development team, all of whom work with Tom Griffiths' in the Computational Cognitive Science Lab at UC Berkeley. Collectively, their expertise includes iterated learning, cultural transmission, web-based experimentation, and software engineering:

- Jordan Suchow (Postdoc at UC Berkeley, Ph.D. in psychology from Harvard) studies vision, memory, and learning. He is a lead developer of Wallace and has experience in releasing research software packages (memtoolbox.org).
- Tom Morgan (Postdoc at UC Berkeley, Ph.D. in biology from St. Andrews) studies cultural evolution, human evolution, and social learning. He is a lead developer of Wallace and is an expert in cultural evolutionary studies.
- Jessica Hamrick (Ph.D. student at UC Berkeley) studies mental simulation and software-based tools for research. She is a core contributor to the Jupyter and IPython projects and has given lectures, courses, and workshops on various topics related to programming.
- Michael Pacer (Ph.D. student at UC Berkeley) studies how different aspects of communication shape the information that is being transmitted. He has extensive experience designing and carrying out cultural transmission experiments online.
- Stephan Meylan (Ph.D. student at UC Berkeley) has extensive experience with iterated learning and experiments with multiple microtask types. He previously worked as an analyst at Crowdflower, a major crowdsourcing startup.
- Thomas Griffiths (Professor at UC Berkeley) studies cultural transmission and inductive biases. He has extensive experience of bringing computational methods to bear on psychological questions.

References

Austin, E. et al. Guide to Social Science Data Preparation and Archiving Best Practice Throughout the Data Life Cycle. (2010).

Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* 286, 509-512 (1999).

Bartlett, F.C. *Remembering: An experimental and social study*. Cambridge: Cambridge University (1932).

Bednarik, P., Fehl, K., & Semmann, D. (2014). Costs for switching partners reduce network dynamics but not cooperative behaviour. *Proceedings of the Royal Society B: Biological Sciences*, *281*.

Caldwell, C.A. & Millen, A.E. Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior* 29, 165-171 (2008).

Claidière, N., Smith, K., Kirby, S. & Fagot, J. Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society of London B: Biological Sciences* 281, 20141541 (2014).

Derex, M., & Boyd, R. (2015). The foundations of the human cultural niche. *Nature Communications*, *6*(May), 1–7. doi:10.1038/ncomms9398

Fisher, R. (1930). *The genetical theory of natural selection*. Oxford: Clarendon.

Flynn, E. & Whiten, A. Cultural transmission of tool use in young children: A diffusion chain study. *Social Development* 17, 699-718 (2008).

Galton, F. Vox populi. *Nature* 75, 450-451 (1907).

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., Hamrick, J. B., & Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107.

Horner, V., Whiten, A., Flynn, E. & de Waal, F.B. Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences* 103, 13878-13883 (2006).

Jacobs, R.C. & Campbell, D.T. The perpetuation of an arbitrary tradition through several generations of a laboratory microculture. *The Journal of Abnormal and Social Psychology* 62, 649 (1961).

Kalish, M.L., Griffiths, T.L. & Lewandowsky, S. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review* 14, 288-294 (2007).

Mcelreath, R., Lubell, M., Richerson, P. J., Waring, T. M., Baum, W., Edsten, E., … Paciotti, B. (2005). Applying evolutionary models to the laboratory study of social learning. *Evolution and Human Behavior*, *26*(6), 483–508.

Morgan, T. et al. Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nature Communications* 6 (2015).

Verhoef, T., Kirby, S. & Padden, C. in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* 483-488 (2011).

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

Analysing discourse relations in natural language: The case of space and time

Thora Tenbrink (t.tenbrink@bangor.ac.uk)
School of Linguistics and English Language, College Road, Bangor University (Wales)
Bangor LL57 2DG, UK

Keywords: discourse relations; reference frames; cognitive processes; verbal data; cognitive discourse analysis; spatiotemporal language.

Objectives and Scope

Discourse relations are links between utterances and sentences or sentence parts. These can be implicit or explicit through markers such as *because*, *before*, *but*, etc. Thus, the causal discourse relation between the two clauses that is explicit in 'He fell because he stumbled over a stone' can be inferred in the sentence 'He stumbled over a stone and fell'. Much research in this field focuses on the range of ways in which these connections can be established and inferred by the recipient (or analyst). However, the *use* of such markers in discourse also shows how a speaker conceptualizes the relationships between various aspects of what they are saying. Whenever they occur in discourse, they are therefore reflections of the speakers' mind in this respect.

In this tutorial, we will explore these effects by focusing on how speakers conceptualise and express relationships in time and space in discourse (Tenbrink, 2007). In the first half, we will discuss the linguistic options in this regard, and work out some principles for (and differences between) temporal and spatial relationships. In the second half, we will look at how these effects can be investigated by a systematic analysis of speakers' discourse produced under controlled circumstances. This includes a brief introduction to the methodology of Cognitive Discourse Analysis, with a focus on relationships between spatial and temporal elements in discourse.

Discourse Relations in space and time

Discourse relations are structures in discourse that connect various linguistically expressed elements meaningfully to each other in the language user's mind (Sanders et al., 1992). These effects have been investigated in multiple ways and come under various names, such as Conjunction in Hallidayan theory (Halliday & Matthiessen, 2014), and rhetorical relations in RST (Mann & Thompson, 1988). Explicit markers can connect different kinds of structural elements, such as clauses, sentences, phrases, and more, depending on the type of connector and other linguistic or conceptual choices.

The domains of space and time are particularly interesting in this regard. As fundamental domains of human experience, they are conceptually connected (Piaget, 1946) and can sometimes be used interchangeably, as in 'follow the road for five minutes / for about 300 yards' Moreover, temporal terms are, to some extent, semantically and historically based on spatial terms (Clark, 1973; Haspelmath, 1997). Thus, explicit markers of temporal and spatial relationships (e.g., prepositions like *after* and *behind*) have much in common. Both represent schematic rather than precise metric relationships, and both rely on some kind of conceptual reference frame (Tenbrink, 2011), albeit in different ways for different kinds of structures.

Nevertheless, in natural discourse the expression of temporal relations between *events* differs systematically from the expression of spatial relations between *objects* in rather fundamental ways, related to the ontological differences between objects and events (Tenbrink, 2007). For instance, in the spatial domain objects are typically conceptually related through their *functions*, which affects language use systematically (Coventry et al., 1994). In the temporal domain, language use is crucially affected by various kinds of *causal* concepts: if events are reported as following after one another (as in 'He stumbled over a stone and fell') a causal relation will typically be assumed, as long as this is in accord with world knowledge. These basic concepts affect patterns of language use in different ways in the two domains, depending on the communicative context.

Cognitive Discourse Analysis

To examine patterns and effects of expressing temporal and spatial relations in natural discourse, it is useful to adopt a systematic approach that relates linguistic choices to features of the situation in which language is produced. Cognitive Discourse Analysis (CODA; Tenbrink, 2015) was designed to address features of language use that reflect concepts and cognitive processes systematically, including the conceptualisation of relationships between objects and events. Linguistic features sometimes convey more than what the speaker is aware of, and more than what linguistic content explicitly expresses. For instance, an utterance like *to the left of the cupboard* presupposes an underlying perspective that defines the *left* side without stating it explicitly. Conceivably, the same spatial relationship could have been expressed in other ways (e.g., *in front of the cupboard* using a different perspective, or *next to the cupboard* using a less specific lateral term). Speakers are typically not consciously aware of the *network of options* (Tenbrink & Freksa, 2009) that allows for other linguistic choices beside their own.

The significance of such linguistic and conceptual choices emerges more clearly by considering a larger data set collected under controlled circumstances. By systematically

analysing speakers' choices of relating objects and events in space and time to each other, the analyst can identify features of the cognitive and communicative circumstances that lead to the preference of particular types of expressions in natural discourse. Insights in cognitive linguistics and beyond support the interpretation of these patterns of language use, and behavioural performance data contribute further insights through triangulation (e.g., Hölscher et al., 2011). This kind of empirical approach is suitable for investigating spatial and temporal cognition through the analysis of language use, addressing how spatial environments and temporal structures are conceptualised and which features of the spatial and temporal situation affect these concepts in the human mind.

Format and organization

This tutorial is designed to cover a half day (three hours) and will be interactive, connecting where feasible to participants' prior experience with any of the elements covered in the tutorial: discourse relations, spatial and temporal language including reference frames, discourse analysis, and empirical studies in cognitive science. Contents can be adapted to some degree according to participants' interest, prioritising theoretical insights or practical aspects of data collection and analysis as desired. Participants will be encouraged to consider relevant research studies in their own current or future work, in light of the theoretical and practical aspects discussed in the tutorial.

Target audience information

There is no prerequisite for taking this tutorial. It is open for researchers in cognitive science at any point in their career, ranging from graduate students to established experts. It is particularly relevant for those who consider collecting (or have already collected) natural language data to address questions about the human mind, and in particular conceptualisations of space and time.

Tutor Information

Thora Tenbrink is Reader in Cognitive Linguistics at Bangor University (Wales, UK), and uses linguistic analysis to understand the human mind. For this purpose she has developed Cognitive Discourse Analysis (CODA, Tenbrink, 2015). She is author of "Space, Time, and the Use of Language" (Mouton de Gruyter, 2007), has co-edited three books on spatial language, representation, and dialogue, and is co-organising UK-Cognitive Linguistics Conference 2016 and CogSci 2017. Current research includes cognitive strategies in problem solving tasks such as Origami paper folding, route planning, and wayfinding under conditions of uncertainty, and spatial communication and dialogue. See http://knirb.net for further information.

Previous instantiations

This tutorial has previously been offered as an invited workshop at Fribourg University, as part of the CRUS Doctoral Programme Language & Cognition, November 24, 2015. The methodology Cognitive Discourse Analysis (CODA), which will be introduced here as a method to analyse discourse relations systematically in elicited discourse, was presented on many previous occasions (see http://knirb.net for details, and Tenbrink et al., 2012, for a report). This includes a theme session at ICLC 2015, workshops at UK-CLC 2014 in Lancaster (UK), at ETH Zurich, and at Bremen University, tutorials at CogSci 2013, ICCM 2012 and 2013, Spatial Cognition 2012, and COSIT 2011, and a summer school course at Bangor University. The language of space and time was a topic at a panel at IPrA 2007, and the language of space was further presented as a summer school course at UCSB in 2013, as a workshop in 2005, and in many presentations over the past decade.

References

Clark, H.H. (1973). Space, time, semantics, and the child. In: Moore, Timothy E. (Ed), *Cognitive Development and the Acquisition of Language*. N.Y. Academic Press, pp. 27-63.

Coventry, K. R., Carmichael, R. & Garrod, S. C. (1994). Spatial prepositions, object-specific function and task requirements. *Journal of Semantics* 11, 289-309.

Halliday, M.A.K. & Matthiessen, C.M.I.M. (2014). *Halliday's Introduction to Functional Grammar (4th edition)*. London: Routledge.

Haspelmath, M. (1997). *From Space to Time. Temporal Adverbials in the World's Languages*. Munich: Lincom.

Hölscher, C., Tenbrink, T., & Wiener, J. (2011). Would you follow your own route description? *Cognition 121,* 228-247.

Mann, W.C. & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text 8:*3, 243-281.

Piaget, J. (1946). *Le développement de la notion de temps chez l'enfant*. Paris: PUF.

Sanders, T.J.M., Spooren. W.P.M., & Noordman, L.G.M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes 15*, 1-35.

Tenbrink, T. (2007). *Space, time, and the use of language: An investigation of relationships*. Berlin: Mouton de Gruyter.

Tenbrink, T. (2011). Reference frames of space and time in language. *Journal of Pragmatics 43*:3, 704-722.

Tenbrink, T. (2015). Cognitive Discourse Analysis: Accessing cognitive representations and processes through language data. *Language and Cognition* 7:1, 98 – 137.

Tenbrink, T., D'Odorico, T., Hertzberg, C., Mazman, G., Meneghetti, C., Reshoeft, N., & Yang, J. (2012). Tutorial report: Understanding spatial thought through language use. *Journal of Spatial Information Science 5*:107-114. http://josis.org/index.php/josis/article/viewFile/124/90

Tenbrink, T., & Freksa, C. (2009). Contrast sets in spatial and temporal language. *Cognitive Processing 10 Supplement 2*, S322-S324.

Cognitive Models of Transfer of Cognitive Skill (full day)

Niels A. Taatgen (n.a.taatgen@rug.nl)
Institute of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, Netherlands

Keywords: Cognitive Architectures; Cognitive Modeling; Skill Acquisition; Cognitive Transfer.

Transfer in Cognitive Skills

If you learn mathematics in school, but then proceed to become a lawyer, will you have any benefit from mathematics, for example in terms of abstract reasoning? Very little is know about the cognitive mechanisms behind such transfer of knowledge, and most psychological theories assume that very little transfer exists (Singley & Anderson, 1989, Thorndike & Woodworth, 1901). Nevertheless it is undisputable that many of our cognitive skills build upon each other, and have benefits beyond just the skill itself. Moreover, the new controversial industry of Brain Training games has emerged where spectacular improvements on various cognitive measures is promised, even though empirical evidence for such improvements is rather weak (Chein & Morrison, 2010, Jaeggi et al, 2008, Karbach & Kray, 2009, Owen et al., 2010).

A problem with most models and experiments in cognitive science is their focus on the experiment, and therefore take the particular task for granted. As a consequence, there is very little theory on how cognitive skills are interconnected, and how one skill can build on another. For this purpose, the PRIMs cognitive architecture (Taatgen, 2013) was developed, which will be the central topic of this tutorial. PRIMs is derived from the well-known ACT-R theory (Anderson, 2007), and inherits most of its principles (and therefore benefits) from that architecture. However, whereas one of the basic units of knowledge in ACT-R is the production rule, a fairly complex knowledge representation, the PRIMs theory starts with much smaller primitive information processing units (the PRIMs). PRIMs can be clustered into mental operators that perform tasks. However, many of these clusters are task-general, and can be used for different tasks.

The novel aspect of PRIMs is that cognitive modelers

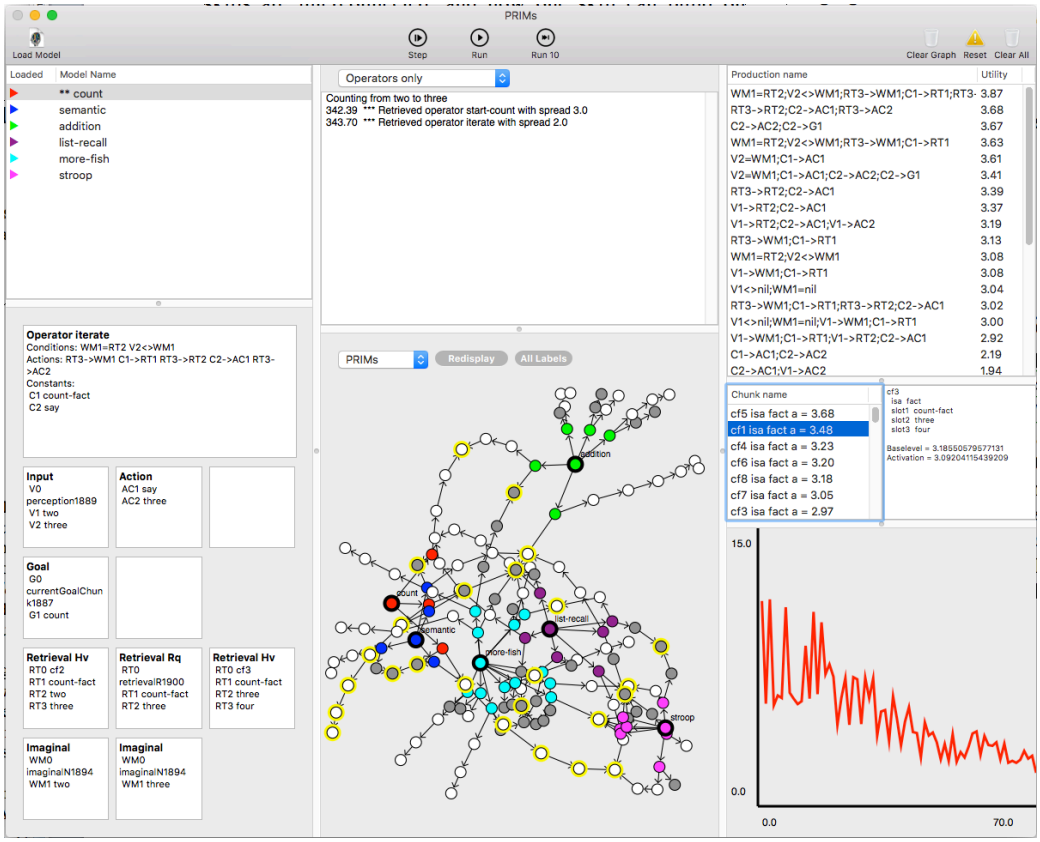


Figure 1. Screenshot of the PRIMs application. Six tasks (top-left corner), have been loaded into the system. The bottom-center panels shows how the knowledge for these tasks is interconnected.

can now move beyond the focus on individual tasks, and start building large-scale models that incorporate knowledge of many different tasks. This means that the modeler typically only has to supply a limited amount of new knowledge for a new task, because the system can use knowledge from other tasks.

Tutorial Structure

The chief goal of the tutorial is to give participants some hands-on experience with the architecture. The PRIMs application is easy to use, and gives a good overview of what happens in the model, how knowledge is interconnected and what the results of running models are. It will therefore start with a relatively brief theoretical overview, followed by a small exercise in which participants have to construct a small model that performs addition by counting, in order to investigate the transfer between regular counting and counting as part of doing addition.

The afternoon session will focus at an example of how PRIMs can model a case of *far transfer*, in which training on task switching improves the model's performance on the Stroop task (we will look at data from Karbach & Kray, 2009, for this). In the subsequent exercise, participants will extend that model with a working memory task.

We will end the session with some discussion about how PRIMs can discover its own knowledge, and future directions of the architecture.

Materials

Participants will be provided with a tutorial reader that has the instructions on how to use the software and build models. They can download and run the software themselves from https://github.com/ntaatgen/ACTransfer (it has both the source code and a precompiled application). Participants will also be provided with the Powerpoint presentations and other relevant files for the tutorial. The current limitation is that it only runs on a Macintosh computer, so participants may have to share machines for the exercises.

The tutorial assumes no prior knowledge apart from a general cognitive science audience. It may be helpful for participants to read the article that outlines the theory (Taatgen, 2013).

Credentials of the Organizer

Taatgen has given tutorials on ACT-R during CogSci 2004, and Taatgen and van Rijn during CogSci 2005-2008, and in 2011 with Jelmer Borst. In addition, Taatgen has organized the ACT-R summer school at CMU in the past, and presently co-hosts the yearly European cognitive modeling Spring School in Groningen that has run from 2009 until present. He has given the present tutorial to a smaller audience in 2015 at Rensselaer Polytechnic University, and will give the tutorial during the upcoming Spring School in Groningen.

Acknowledgements

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen.

References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford university press.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17*(2), 193-199.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829-6833

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science, 12*(6), 978-990.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., . . . Ballard, C. G. (2010). Putting brain training to the test. *Nature, 465*(7299), 775-778.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Taatgen, N.A. (2013). The nature and transfer of cognitive skills. *Psychological Review, 120*(3), 439-471.

Taatgen, N.A. (2013). Diminishing return in transfer: A PRIM model of the Frensch (1991) arithmetic experiment. *Proceedings of the 12th international conference on cognitive modeling* (pp. 29-34). Ottawa: Cartelon University.

Taatgen, N.A. (2014). Between architecture and model: Strategies for cognitive control. *Biologically Inspired Cognitive Architectures, 8*, 132-139.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review, 8, 247-261.

Tutorial: Meta-Analytic Methods for Cognitive Science

Sho Tsuji
tsujish@gmail.com
Département d'Etudes Cognitives
Ecole Normale Supérieure

Molly Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Christina Bergmann
chbergma@gmail.com
Département d'Etudes Cognitives
Ecole Normale Supérieure

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Alejandrina Cristia
alecristia@gmail.com
Département d'Etudes Cognitives
Ecole Normale Supérieure

Meta-analysis is a powerful yet underused tool in cognitive science. It allows researchers to leverage entire bodies of literature to get a broad and quantitative overview of a particular phenomenon, thereby promoting theory development, and to make more precise estimates of effect sizes, which enables robust planning of prospective studies (e.g. through power-analyses). In this tutorial, we will introduce meta-analysis as a tool with which to inform everyday research, and provide participants with hands-on experience conducting their own meta-analysis. We will also present an online platform we have developed for conducting meta-analyses in the field of language development: MetaLab (http://metalab.stanford.edu).

Keywords: meta-analysis; reproducibility

Significance

The empirical social sciences are in crisis: Many subfields are plagued by issues of low reliability and validity of their findings (Ioannidis, 2005; Open Science Collaboration, 2013, 2015). It has become evidently clear that any single study is limited both in interpretability and in scope, as it is a noisy estimate of the underlying effect size, often measured unreliably due to low power, and it only measures an effect in one setting. Meta-analysis is a powerful tool that allows quantitative aggregation of effect sizes across studies in a particular field. Meta-analyses can provide three key pieces of information. First, they allow an estimation of the presence of bias in a field of work. Second, they yield more realistic measures of the size of main effects and their variability, allowing researchers to better inform their power analyses and obtain more accurate estimates of desirable sample sizes. Third, by providing a framework in which different studies can be compared in a quantitative way, meta-analyses allow the exploration of relationships between variables previously not compared in a single study, thereby further developing theories based on a broad overview of a particular phenomenon. In sum, meta-analyses allow both consumers and producers of a given field of work to gain a better appreciation of that research, and they enable researchers to make both practical decisions (such as sample size) and theoretical decisions (such as predictions for particular variables) that are grounded in empirical data.

Despite the salient benefits of using meta-analysis, cognitive scientists use meta-analyses relatively rarely. The most likely reason is that we lack the training to carry out and use meta-analyses effectively. Indeed, meta-analyses traditionally rely on a very few people painfully entering large bodies of research, with little ready-to-use support tools and educational materials available. In addition, the general benefits of meta-analyses, for instance the possibility of conducting power analyses, are often neither evident nor accessible to individual researchers who lack training on this simple tool. Moreover, if a meta-analysis already exists, potential re-users of this valuable data may feel that its value diminishes as time goes on, since traditional meta-analyses remain static after publication, aging quickly as new results emerge. Finally, even if researchers may be keen on utilizing an extant meta-analysis, they may be uncertain about how to deal with "mixed apples and oranges" or the presence of a publication bias.

The goal of this tutorial is to empower participants to harness the power of meta-analyses. We will start with a broad introduction to meta-analysis as an analytical tool. Participants will then get hands-on experience conducting and reusing a meta-analysis in an interactive session. By the end of the tutorial, participants will have a better understanding of the practical and theoretical utility of meta-analysis, as well as working knowledge about how to go about conducting their own meta-analysis, or reusing an extant meta-analysis.

Additionally, we will introduce participants to a novel tool called MetaLab (http://metalab.stanford.edu; Bergmann et al., 2015; Lewis et al., 2015). MetaLab is an online platform that aggregates meta-analyses on topics related to language development (e.g., phoneme discrimination and word segmentation; Tsuji & Cristia, 2014; Bergmann & Cristia, 2015). MetaLab facilitates the learners' and users' task in three ways. First, it supplies templates and analysis scripts, streamlining the process of learning about and conducting a meta-analysis. Second, it supports community-augmented meta-analyses (CAMA; Tsuji, Bergmann, & Cristia, 2014), allowing a meta-analysis to be conducted and extended by multiple researchers, both reducing the workload of the individual researcher as well as allowing for dynamic extensions to always include the newest results. Third, for each meta-analysis conducted in the MetaLab framework, we provide free and easy-to-use tools for power analysis and data exploration.

A broad range of researchers will benefit from this introduction to meta-analysis: Novices to a particular research subfield who are in need of a robust overview can turn their

literature review into a meta-analysis with a few additional steps, providing themselves and the whole research community with a valuable resource. Any researcher faced with new requirements for publishing in top-tier journals, such as providing a reason for sample size decisions, will profit from being familiar with the concept and uses of meta-analyses, including thinking in terms of effect sizes rather than significance and being able to carry out a prospective power analysis.

Structure

This one-day tutorial will introduce participants to the method of meta-analysis, providing a hands-on step-by-step guide to use the MetaLab infrastructure for conducting a meta-analysis, working on it collaboratively, and sharing it with the research community.

We will lead participants through the steps of a meta-analysis based on a pre-selected topic. The topics of literature search and study selection, which precede the actual meta-analysis, will be covered briefly. Participants will be walked through the steps of a meta-analysis with a theoretical and practical component to each step of the process.

1. Coding of variables (2h)
   (a) Theory: How to decide on independent and dependent variables to be included; which pieces of information are mandatory and optional
   (b) Practical: Set-up of a spreadsheet in standardized format, deciding on variables to be included, explain in what format to code variables in order for them to be included in quantitative analyses, coding of one preselected article (different article for each participant)
2. Effect size calculation (1h)
   (a) Theory: Introduction to different types of effect sizes, their calculation, and how to transform between them
   (b) Practical: Effect size calculation for paper coded
3. Meta-analysis (2h)
   (a) Theory: Introduction to meta-analytic regression, choice of model, choice of moderator variables, correction for publication bias, and interpretation of analysis output
   (b) Practical: Putting together the papers coded by each participant and conducting a meta-analysis
4. Integration with MetaLab and use of extant meta-analyses (1h)
   (a) Theory: Advantages of making a meta-analysis publicly available, and how to use extant meta-analyses for informing new study design
   (b) Practical: Examples of power analysis, study design decisions, including, but not restricted to use of MetaLab infrastructure.

Each participant will need a laptop, but no additional materials are required for the tutorial.

Organizer Credentials

All authors have conducted meta-analyses in their field: ST, AC: (Tsuji & Cristia, 2014); ML, MF: (Lewis & Frank, 2015), CB, AC: (Bergmann & Cristia, 2015). ST, CB and AC have proposed the concept of CAMAs (Tsuji et al., 2014) and provided online tutorials for facilitating meta-analysis for researchers in the cognitive sciences. ST, ML, and CB have taught the basics of meta-analysis, including theoretical and hands-on parts, in graduate level university classes and lead meta-analysis workshops. They have also presented the CAMA and MetaLab frameworks on international conferences (Bergmann et al., 2015; Lewis et al., 2015). All authors have collaborated to develop the MetaLab infrastructure and tutorial materials since 2/2015 (together with Mika Braginsky and Page Piccinini).

References

Bergmann, C., & Cristia, A. (2015). Development of infants' segmentation of words from native speech: a meta-analytic approach. *Developmental Science*, Advance online publication.

Bergmann, C., Tsuji, S., Lewis, M., Braginsky, M., Piccinini, P., Cristia, A., & Frank, M. C. (2015). Metalab: Power analysis and experimental design in developmental research made easy. Poster presented at the 6th Annual Budapest CEU Conference on Cognitive Development.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Lewis, M., Braginsky, M., Bergmann, C., Tsuji, S., Cristia, A., & Frank, M. C. (2015). Metalab: tool for power analysis and experimental planning in developmental research. In *Proceedings of 40th annual meeting of the Boston University Child Language Development.*

Lewis, M., & Frank, M. C. (2015). A meta-analytic approach to understanding the disambiguation effect. Poster presented at the 2015 Biennial Meeting of the Society for Research in Child Development, Philadelphia, PA.

Open Science Collaboration. (2013). The reproducibility project: A model of large-scale collaboration for empirical research on reproducibility. *Implementing reproducible computational research. New York, NY: Taylor & Francis*, *2*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.

Empirical and Computational Approaches to Metaphor and Figurative Meaning

Justine T. Kao (justinek@stanford.edu) and Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, Stanford University, Stanford CA, USA

Keywords: Figurative language; Metaphor; Computational modeling; Computational linguistics; Psycholinguistics

Motivation

One of the hallmarks of human intelligence is the ability to go beyond literal meanings of utterances to infer speakers' intended meanings, a feat that remains elusive to the most advanced artificial systems. Figurative language such as metaphor, in particular, provides a striking case where complex meanings arise that cannot be derived from the literal semantics alone. For example, *My lawyer is a shark* is a literally false sentence, yet communicates relevant features of the lawyer in question (e.g. *ruthless*, but not *swims*).

Metaphor and other types of figurative language raise a range of questions core to the study of language and cognition; as a result, how people derive and produce figurative meanings has been approached from various angles and with a diverse set of methods. Some psychologists focus on the cognitive mechanisms that underly interpretations of specific types of figurative use, such as how people align shared properties and analogous relations across domains in order to understand metaphor (Gentner & Wolff, 1997). Other researchers apply theories of communication to explain how people arrive at contextually appropriate interpretations of non-literal utterances (Steen, 2015; Kao, Wu, Bergen, & Goodman, 2014). Finally, natural language processing researchers seek to identify features and principles that can help artificial agents process and produce figurative language (Veale & Hao, 2007).

In this symposium, we will discuss the methods that our speakers have employed to examine how people interpret and produce figurative meaning, as well as ways to combine complementary approaches. By bringing together experts with different theoretical perspectives and from a range of disciplines, we aim to discuss outstanding questions that may require a synthesis of tools to resolve. We will open the symposium with a brief overview of the landscape of metaphor and figurative meaning, provided by **N. Goodman**. We will then present four 20-minute talks, starting with **F. Maravilla** presenting joint work with **D. Gentner** on how conventionality, relationality, and aptness affect figurative language processing. **G. Steen** will then introduce the Deliberate Metaphor Theory and highlight the importance of communicative intent. Continuing this thread, **J. Kao** will present joint work with **Goodman** on applying Bayesian models of communication to interpret figurative uses. Finally, **T. Veale** will describe a computational system that employs cognitive principles to automatically generate figurative language. The symposium will end with a 20-30 minute discussion with the speakers and audience (facilitated by **Goodman** and **Kao**).

Conventionality, Relationality and Aptness in Figurative Language Processing

Francisco Maravilla & Dedre Gentner (Northwestern University)

Figurative statements can be expressed either in comparison syntax, as similes (An X is like a Y), or in categorization syntax, as metaphors (An X is a Y). What determines preference for one form over the other? The Career of Metaphor account (Bowdle & Gentner, 2005; Gentner & Wolff, 1997) links this difference to language evolution. Novel figuratives are processed as comparisons, so they are preferred in simile form. In contrast, conventional figuratives can be expressed as metaphors (categorization form), because they have acquired a standard metaphoric abstraction. A second account (Aisenman, 1999) proposes that people prefer to express relational meanings as metaphors and attributional meanings as similes. A third account, Glucksberg and Keysar (1990) class inclusion theory of metaphor, claims that the preference for metaphor form depends on aptness, not conventionality (Glucksberg, 2003). We tested these accounts in three studies. In Experiment 1, conventionality and relationality significantly predicted preference for metaphor. In Experiment 2, we found that people provided relational interpretations for both forms. In Experiment 3 (underway) we test the claim that apt figuratives are preferred as metaphors, and ask how aptness relates to relationality. Our findings so far support the Career of Metaphor theory and Aisenman?s relationality hypothesis.

Introducing Deliberate Metaphor Theory

Gerard Steen (University of Amsterdam)

99% of all metaphorically used words have meanings that are so conventionalized that they can be found straight in a user's dictionary (Steen et al., 2010). This corpus-linguistic finding raises fundamental questions about how people produce and interpret metaphor in discourse. For instance, if conventionalized metaphorical meanings are so readily available, can a processing model that includes lexical disambiguation between metaphorical and non-metaphorical senses account for most metaphor use? This would go against Conceptual Metaphor Theory (Lakoff & Johnson, 2008) and support part of Sperber and Wilson (2008)'s deflationary account of metaphor and would agree with Giora (2008)'s

position that (most) metaphor is nothing special. However, not all metaphor is conventional, and even conventional metaphor can require processing via on-line cross-domain comparison (Steen, 2008, 2015). This is when metaphor is used deliberately as a metaphor, the referential meaning of the utterance in context involving an intentional comparison between two concepts, categories, spaces or domains. Examples include all novel metaphors (where processing can only go via the meaning of the source domain to arrive at an interpretation of the utterance as a whole) and all explicit comparisons (such as similes and analogies where the target is explicitly compared to some alien source). These deliberate metaphors point to a third dimension of metaphor, next to language and thought, i.e., communication, and can explain some of the current findings and theories, including Bowdle and Gentner (2005)'s Career of Metaphor Theory, in alternative ways. The paper will present a brief account of Deliberate Metaphor Theory.

Modeling figurative communication

Justine T. Kao & Noah D. Goodman (Stanford)

Figurative language often requires listeners to access background knowledge and discourse context in order to arrive at appropriate interpretations. How do listeners incorporate multiple sources of information to derive true and relevant information from figurative utterances? We describe an extended version of the Rational Speech Act (RSA) model, a Bayesian computational model that formalizes core principles of communication. Our model highlights the idea that speakers are assumed to be informative with respect to specific communicative goals, which may be to convey subjective attitudes instead of objective information about the world. By reasoning about the range of communicative goals a speaker may have, listeners can go beyond the literal meanings of utterances to arrive at figurative interpretations. Through a series of experiments, we show that our model predicts people's interpretations of hyperbole, irony, and metaphor with high accuracy. We show that despite apparent differences among these subtypes of figurative language, the same computational framework flexibly produces fine-grained interpretations for diverse figurative uses. We use this as evidence suggesting that the rich and often affectively-laden meanings expressed by figurative language can be explained by basic principles of communication.

Metaphor and the human creative potential

Tony Veale (University College Dublin)

Metaphors in the wild rarely occur in isolation. Rather, metaphors in many texts are supported by, or give support to, other metaphors, similes and blends. These figurative devices present different affordances to an author, and serve complementary roles in the communication of complex ideas. A tentative simile or analogy may lay the groundwork for bolder metaphors to come, which can in turn set the stage for an immersive conceptual blend. Clearly, these are not disjoint phenomena arising from distinct mechanisms, but the manifestations of different settings–such as degree of integration–of a common creative mechanism. Computational modeling offers researchers a generative approach to demonstrating this claim, by supporting the construction of generative systems that can produce their own similes, metaphors, analogies and blends from a small set of core principles. In this talk I explore the workings of such a computational system, called Metaphor Magnet, that is realized in a number of public forms, from a Twitterbot called @MetaphorMagnet to a Web service that provides figurative competence to 3rd-party software systems. In particular, I show how a separation of concerns–chiefly, of conceptual content and linguistic framing–allows the system to achieve a wide variety of human-like figurative outputs from a small number of cognitive principles and AI mechanisms.

References

Aisenman, R. A. (1999). Structure-mapping and the simile–metaphor preference. *Metaphor and Symbol*, *14*(1), 45–51.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*(1), 193.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *J of Memory and Language*, *37*(3), 331–355.

Giora, R. (2008). Is metaphor unique. *The Cambridge handbook of metaphor and thought*, 143–160.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in cognitive sciences*, *7*(2), 92–96.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological review*, *97*(1), 3.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge handbook of metaphor and thought*, 84–105.

Steen, G. (2008). The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, *23*(4), 213–241.

Steen, G. (2015). Developing, testing and interpreting deliberate metaphor theory. *Journal of Pragmatics*.

Steen, G., Dorst, A., Herrmann, B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From mip to mipvu* (Vol. 14). John Benjamins Publishing.

Veale, T., & Hao, Y. (2007). Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *Aaai* (Vol. 2007, pp. 1471–1476).

Comics and cognitive systems: The processing of visual narratives

Neil Cohn (neilcohn@visuallanguagelab.com)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Emily L. Coderre (ecoderr1@jhmi.edu)
Department of Neurology, Johns Hopkins University School of Medicine

L.N. Kendall (lia.kendall@psych.ubc.ca)
Department of Psychology, University of British Columbia

Joseph P. Magliano (jmagliano@niu.edu)
Department of Psychology, Northern Illinois Unviersity

Keywords: comics; narrative; visual narrative; autism; perception; event-related potentials

Introduction

Humans have drawn sequential images as a means of expression throughout history, from cave paintings and frescoes to wall-carvings and tapestries (McCloud, 1993). In contemporary society, we find them most prevalently in comics of the world, and over the past two decades, increasing attention has turned to this communicative system in the cognitive sciences.

Earlier work often focused on theory alone, drawing from paradigms in linguistics or semiotics (for review, see Cohn, 2012; Wildfeuer & Bateman, 2016) or from theorists outside academia (e.g., McCloud, 1993). However, newer studies test theoretical predictions with empirical corpus analyses and both behavioral and neurocognitive experimentation. As in language research, combining these methodologies provides converging evidence on the structure of visual narratives, their diversity across the world, and their comprehension by minds and brains.

Recent research has especially focused on the overlapping cognition between the processing of the "visual languages" constituting drawn visual narratives and the linguistic systems of verbal and signed languages (Cohn, 2013; Magliano, Larson, Higgs, & Loschky, 2015). These presentations further such analyses, and explore questions related to the degree to which these visual languages share mechanisms with linguistic and other cognitive systems.

Visual narratives and converging evidence

The first presentation by **Neil Cohn** will provide an overview of this multipronged approach to visual narratives by discussing recent projects in collaboration with Marta Kutas. Many visual narratives use a construction where successive images depict different characters, thereby "conjoining" them into a larger spatial environment containing all the characters (Cohn, 2013). Corpus analysis has suggested that this "narrative grammatical pattern" appears more in Japanese *manga* than in American comics (Cohn, 2013). We used event-related brain potentials (ERPs) to determine whether comprehension of such "visual narrative conjunctions" invokes "grammatical" combinatorial processing in addition to incremental mental model updating in wordless visual narrative sequences. We crossed (non)conjunction sequences with (in)congruities. Conjunctions elicited a larger anterior negativity (300-500ms) than non-conjunctions, regardless of congruity, implicating only "grammatical" processes because of its insensitivity to semantic incongruence. Both non-conjunction and incongruity elicited larger P600s (500-700ms), indexing a process of mental updating or reanalysis (Cohn & Kutas, 2015; Kuperberg, 2013). Finally, regression analyses revealed that both neural conjunction effects were modulated by participants' frequency of reading manga while growing up: A greater anterior negativity in frequent manga readers suggested more reliance on combinatoric processing, while larger P600 effects in infrequent manga readers suggested more mental updating. Thus, through converging evidence across theoretical, corpus, and experimental methods, we can see that visual narrative comprehension involves multiple processes modulated by fluency in specific "visual languages."

Understanding the moment-to-moment processing of sequential narratives

It is the aspects of mental model updating where we find the most overlap between research on the moment-to-moment processing of visual and text-based narratives. This is particularly true for investigating the role of inferences in mental model construction. In this presentation, **Joseph P. Magliano** (along with co-authors John P. Hutson, Adam M. Larson, Karyn Higgs, & Lester C. Loschky) present a mix of recently published and newly analyzed data that show how processing effort (global picture viewing times) and eye movements are sensitive to mental model construction.

To test this relationship, participants were presented with a visual narrative. The construction of participant mental models was manipulated through the presence or absence of a panel that depicted the bridging event of an action sequence. We predicted that participants who did not see the bridging event would need to generate an inference about

the bridging event in order to construct a coherent mental model of the visual narrative. The data shows that processing pictures in a visual narrative is sensitive to the need to construct bridging inferences and breaks in situational continuity. These results again demonstrate the benefits of converging evidence combining the utility of viewing time data for studying mental model construction in visual narratives, and the role of top-down comprehension processes in generating eye-movements. More importantly, these results speak to the need for a theory of how eye movements and visual narrative comprehension processes are coordinated.

Visual vs. Linguistic Narratives in individuals with Autism

Work presented by **Emily L. Coderre** (along with co-authors Neil Cohn, Sally K. Slipher, Mariya Chernenok, Kerry Ledoux, & Barry Gordon), will further explore the overlaps in processing of language and visual narratives by comparing the cognition of healthy adults with those of individuals with autism spectrum disorders (ASD). Language difficulties are one of the hallmarks of ASD. Previous research has documented particular difficulty with narrative comprehension among individuals with ASD, indicating impairment in understanding and interpreting narratives compared to typically developing controls. However, the majority of these studies have used linguistic stimuli (i.e. written or spoken narratives), making it difficult to determine whether the impairment in narrative comprehension stems from the use of language-based stimuli, which individuals with ASD have difficulty with overall, or from a more global impairment in narrative comprehension. The current study aimed to address this question by analyzing the N400 ERP component, which is shown to be elicited by incongruous stimuli in both linguistic stimuli (sentences) and sequential images (visual narratives), to directly compare semantic comprehension of linguistic narratives (a series of short sentences) and visual narratives (a series of comic panels) in high-functioning adults with autism (HFAs) and normal controls (NCs). Each narrative ended with either a semantically congruent or incongruent word or image. HFAs showed reduced N400 effects compared to NCs for both linguistic and visual narratives, suggesting impaired semantic comprehension of narratives in both modalities. These results, which suggest that difficulties with narratives in individuals with ASD stem from a more global impairment of narrative comprehension regardless of the stimulus modality, have important implications for the use of visual and linguistic stimuli in intervention and treatment strategies in ASD.

Could any real person look as despondent as Charlie Brown?

The higher-level information in visual narratives must draw from basic aspects of object recognition and scene perception. Our final presentation by **L.N. Kendall** (with co-authors Quentin Raffaelli, Alan Kingstone, Rebecca M. Todd) targets one facet of this lower-level processing: the differences in processing the styles of visual representations.

Iconic representations are ubiquitous in visual narratives like comics, graphic novels, and children's cartoons. While there is some evidence that iconic images are processed differently from photorealistic images, the qualities that serve to distinguish the two are largely unknown. We tested whether iconic representations are uniquely advantageous for communicating emotion quickly and efficiently. In Experiment 1 displayed face stimuli at various presentation times to determine the time necessary to detect an expression on a face. As images became more iconic, their information grew easier to discriminate. Additionally, analysis of ERP components P1 and N170 suggested that as images grew more iconic, they were processed faster. Cartoon images also uniquely showed more negativity in both components. In Experiment 2, we tested if the features of cartoon faces hold any processing advantage when they match real life expressions compared to when they do not, e.g., :) versus :+. Participants categorized faces either with known symbols or arbitrary symbols for mouths. Halfway through the task, they relearned the arbitrary faces as representing emotions. We found that arbitrary symbols as facial features can be learned to represent emotions as well as meaningful symbols, as measured by N170 similarity. These experiments provide converging evidence that iconic representations are unique in their ability to better represent and communicate emotional information.

References

Cohn, N. (2012). Comics, Linguistics, and Visual Language: The Past and Future of a Field. In F. Bramlett (Ed.), *Linguistics and the Study of Comics* (pp. 92-118). New York: Palgrave MacMillan.

Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London, UK: Bloomsbury.

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia, 77*, 267-278. doi:10.1016/j.neuropsychologia.2015.08.026

Kuperberg, G. R. (2013). The pro-active comprehender. In *Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension*. Baltimore: Paul Brookes Publishing.

Magliano, J. P., Larson, A. M., Higgs, K., & Loschky, L. C. (2015). The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. *Memory & Cognition*, 1-13. doi:10.3758/s13421-015-0558-7

McCloud, S. (1993). *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.

Wildfeuer, J., & Bateman, J. (2016). Linguistically Oriented Comics Research in Germany. In N. Cohn (Ed.), *The Visual Narrative Reader* (pp. 19-66). London, UK: Bloomsbury.

Effecting re-representation: revising false beliefs and fostering creativity

Micah B. Goldwater (micah.goldwater@sydney.edu.au)
School of Psychology, University of Sydney, New South Wales, Australia

Zachary Horne (horne2@illinois.edu), Derek Powell (derekpowell@ucla.edu), & John Hummel (jehummel@illinois.edu)
Department of Psychology, University of Illinois Urbana-Champaign, IL, USA

Jeffrey Loewenstein (jloew@illinois.edu)
Department of Business Administration, University of Illinois Urbana-Champaign, IL, USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IA, USA

Keywords: false beliefs; belief revision; causal learning; re-representation; negotiation; creativity

Introduction

Perhaps the greatest challenge to successful learning and problem solving is when learners initially represent the novel problem or concept in a way that conflicts with the proper solution or content to be learned. For example, naïve conceptions of *force* often lead to failure to show any learning gains from lessons on Newton's laws of motion (e.g., Muller, Sharma, & Reiman, 2008), and knowledge of typical functions prevent using familiar artifacts in novel ways to solve otherwise simple problems (Adamson, 1952). Further, rampant society-wide misconceptions, such as the belief that vaccines cause autism, or deny the causal role of humans in global warming, are having disastrous effects on public health and environmental policy. On the flip side, overcoming the constraints of an initial representation can spark creative innovation. For example, the Dyson vacuum cleaner is famously based on the re-conception of the vacuum mechanism via an analogy to saw mills.

The goal of this symposium is to present a variety of methods to effect successful re-representation across several domains. We bring together empirical work addressing processes of re-representation in the domains of moral reasoning, causal learning, and negotiation in business. We hope communicating these methods will help grow the group of researchers both trying to improve our basic knowledge of re-representational processes, and solve these real-world problems.

Z. Horne, D. Powell & J. Hummel will present on the effects of revising so called "toxic beliefs"— beliefs that constrain the understanding of entire domains, preventing the learning of new information. M. Goldwater presents a novel approach to overcoming stubborn causal misattributions by teaching learners to categorize disparate phenomena via their shared causal structure. Categorizing phenomena aids learners to discard misattributed causal relations inconsistent with the category's causal model. J. Loewenstein presents a line of studies on when and why negotiators redefine what is under discussion to form creative agreements. Our likelihood of redefining problems appears to be partly a matter of how tightly we are clinging to our existing definitions. R. Goldstone, an expert in the reciprocity between perceptual learning and higher-level cognition, will serve as a discussant, integrating the presentations, making connections between the processes of reinterpreting situations and perceptually re-parsing objects in new ways, and presenting future challenges.

Toxic beliefs

Zachary Horne, Derek Powell, and John Hummel

In the medical, moral, and political domains, people are often deeply entrenched in their beliefs, resistant to any new information. Educational efforts—for instance, pro-vaccine messaging which emphasizes the safety of vaccines—can sometimes fail (e.g., Nyhan et al., 2014; also see Horne, Powell, Hummel, & Holyoak, 2015 for a recent success). Here, we investigate a potential cause of such failures, a class of beliefs we call *toxic beliefs*— which are a kind of over-hypothesis (e.g., Jern et al., 2014) that inform how reasoners interpret evidence in the belief's domain.

To consider an intuitive though controversial example, certain religious beliefs may be considered toxic (in the functional sense just defined). Imagine that some event, say a new archaeological find, would normatively constitute evidence against the hypothesis that God exists (henceforth, evidence E1 and hypothesis H1). However, now imagine that the reasoner also holds the belief that all people of faith will often have their religious convictions tested by secular society (call this H2). This belief affects how reasoners interpret evidence like E1. Rather than viewing E1 as direct evidence against H1, E1 might instead provide further support for H2 in turn *strengthening* H1. Thus, we call H2 a toxic belief because the belief that the faithful will be challenged allows any evidence *against* the existence of God be reinterpreted as evidence *for* the existence of God.

Here we present a line of research applying these ideas to medical denialism – crudely, the idea that modern medicine

provides ineffective treatments and simply supports corporate interests. We argue that medical denialism is a kind of toxic belief that can lead reasoners to be negatively affected by positive medical information and health education. We then discuss avenues to encouraging rational belief revision by altering or undermining toxic beliefs.

Discarding causal misattributions with abstract causal knowledge

Micah B. Goldwater

Once a causal relation is attributed between two variables, it stubbornly persists in learners' causal models despite novel evidence to the contrary— evidence that would have prevented the misattribution in the first place. Taylor and Ahn (2012) refer to this as "causal imprinting." Their work concerned learners first attributing a causal relation between two correlated variables, and then being presented with evidence that the correlation can be explained with a third variable that is the common cause of both. That is, the first two co-occur *because* this third variable causes them to. Taylor and Ahn presented several experiments that show that this stubborn causal imprinting is not due simply to memory limitations, or total failure to update their representations. Subjects did learn that the third variable caused the other two, but these causal relations were simply added to their causal model; they did not replace the original causal (mis-)attribution. On the other hand, when subjects learned about all three variables to start with, they only represented the two correct causal relations.

In a separate line of research, Goldwater and Gentner (2015) taught subjects to classify descriptions of natural and social phenomena via their underlying causal structure, such as a common-cause system (wherein multiple effects have a single common-cause). The current work presents evidence that the learning intervention of Goldwater and Gentner can overcome causal imprinting. That is, subjects with an abstract representation of a "common-cause system" were capable of recognizing how that structure applied to the co-occurrence data from Taylor and Ahn, encouraging learners to discard the causal misattribution. This research is novel both because it shows how causal category learning is applied to the interpretation of co-occurrence data, and argues for learning at the category-level as a method to revise false beliefs about individual exemplars.

Redefining negotiation problems to form creative agreements

Jeffrey Loewenstein

Negotiations, disputes, and other social conflicts sometimes result in creative solutions that make everyone better off for having had the run-in. The pioneering Mary Parker Follet (1940) championed it as "constructive conflict." The question is when and why multiple people coming together to generate a decision end up redefining their understandings and producing creative agreements rather than settling for meager divisions and souring their relationships.

My colleagues and I are finding that one's stance towards the problem and the other parties appear to matter. Unsurprisingly, when parties are unable to reach their goals and so are confronted by an impasse they are more likely to redefine the problem and generate creative agreements. But it is not just flexibility but goal commitment and flexibility together that foster redefining the problem and generating creative agreements. Further, focusing on higher-order goals rather than lower-order goals fosters generating creative agreements.

In addition, considering not just one's own ambiguous information but also the ambiguous information offered by others is important. This is challenging because those others might just be trying to use their influence to gain acceptance of lopsided proposals. Nonetheless, seeking to learn from the current experience rather than maximize current performance fosters generating creative agreements. Seeking to ask questions of others rather than seeking to talk about one's own information fosters generating creative agreements. Not trust in others but respect for others fosters attending to and making use of others' information and then appreciating and accepting their creative proposals.

Taken together, these findings indicate that redefining negotiation problems to produce creative agreements is strongly influenced by the broader goals and approach people adopt towards the negotiation problem and their counterparts.

References

Adamson, R. E. (1952). Functional fixedness as related to problem solving: a repetition of three experiments. *Journal of experimental psychology*, *44*(4), 288.

Follett, M. (1940). *Constructive conflict*. New York: Harper.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: structural alignment and explication in learning causal system categories. *Cognition, 137*, 137-153.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Procee dings of the National Academy of Sciences*,*112*(33), 10321-10324.

Jern, A., Chang, K. M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological review, 121*(2), 206.

Muller, D. A., Sharma, M. D., & Reimann, P. (2008). Raising cognitive load with linear multimedia to promote conceptual change. *Science Education*, *92*(2), 278-296.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. *Pediatrics*, *133*(4), e835-e842.

Taylor, E. G., & Ahn, W. (2012). Causal imprinting in causal structure learning. *Cognitive Psychology, 65*(3), 381-413.

The Emergence of Conventions

Robert X. D. Hawkins (rxdh@stanford.edu)
Noah D. Goodman (ngoodman@stanford.edu)
Stanford University
Kenny Smith (kenny.smith@ed.ac.uk)
Olga Fehér (olga.feher@ed.ac.uk)
University of Edinburgh
Robert L. Goldstone (rgoldsto@indiana.edu)
Indiana University
Tom Griffiths (tom_griffiths@berkeley.edu)
University of California, Berkeley

Overview

Much of our everyday behavior is governed by conventions. The shape of the line we form at the café, the language we use to order our coffee, and the money we use to pay for it are all partly arbitrary but self-sustaining solutions to recurring coordination problems. This definition, first formalized by David Lewis (1969), has provided a potent means of characterizing conventions. For cognitive scientists, however, the outstanding question is how these solutions emerge and adapt in populations of learning, reasoning agents. The aim of this symposium is to gather and integrate several distinct empirical and theoretical perspectives on this question, bridging different domains of application.

There are two primary experimental paradigms used to study the emergence of conventions, each particularly useful for answering certain theoretical questions. The first is an iterated learning or diffusion chain design, where each successive participant produces the data used to train the next participant. Iterated learning experiments have been used, for example, to probe the inductive biases driving the emergence of efficient, compressible structures in language (Kirby, Cornish, & Smith, 2008; Griffiths & Kalish, 2007). The second is a closed group design, drawing from tasks in collective behavior and behavioral game theory, where a single group of two or more participants repeatedly play a coordination game with one another. Because participants directly interact over time, closed group tasks have been useful for capturing the establishment of shared social expectations and the behavioral dynamics of self-organization (Clark & Wilkes-Gibbs, 1986; Goldstone, Roberts, & Gureckis, 2008; Galantucci, 2005).

Recently, these two paradigms have begun to merge, especially within the domain of language evolution, spurring novel models of the pressures and constraints contributed by each mechanism and how they interact to shape stable conventions (Kirby, Tamariz, Cornish, & Smith, 2015; Centola & Baronchelli, 2015). However, many insights from collective behavior and social cognition have not yet been integrated. Moving forward, theoretical questions include sources of variability in convention-formation, the role of active social reasoning vs. emergent inductive biases, and the multiple scales on which conventions are formed; empirical issues include scaling up experiments to encompass larger populations and developing hybrid paradigms to test domain-generality of formation processes.

This set of issues cuts across the interests of many different areas of cognitive science: from linguists concerned with the origins of language, to developmental psychologists interested in how children co-create their culture, to computer scientists attempting to build self-sustaining robot collectives. Because of this breadth and because recent developments have opened up exciting new sources of data and theory, we are confident this symposium will be of broad interest at Cognitive Science.

The symposium will consist of four talks, described below, by the leaders of the field. We will close with a panel discussion on some of these issues, led by Noah Goodman.

Compression and communication in the cultural evolution of linguistic structure

Kenny Smith, Olga Fehér, & Simon Kirby (Co-author)

Language exhibits striking systematic structure. Words are composed of combinations of reusable sounds, and those words in turn are combined to form complex sentences. These properties make language unique among natural communication systems and enable our species to convey an open-ended set of messages. We provide a cultural evolutionary account of the origins of this structure. We show, using simulations of rational learners and laboratory experiments, that structure arises from a trade-off between pressures for compressibility (imposed during learning) and expressivity (imposed during communication). We further demonstrate that the relative strength of these two pressures can be varied in different social contexts, leading to novel predictions about the emergence of structured behaviour in the wild.

Adaptive group coordination and role differentiation

Robert L. Goldstone & Michael E. Roberts (Co-author)

Many real world situations (potluck dinners, academic departments, sports teams, corporate divisions, committees, seminar classes, etc.) involve actors adjusting their contributions in order to achieve a mutually satisfactory group goal, a win-win result. However, the majority of human group research has involved situations where groups perform poorly

because task constraints promote either individual maximization behavior or diffusion of responsibility, and even successful tasks generally involve the propagation of one correct solution through a group.

Here we introduce a group task that requires complementary actions among participants in order to reach a shared goal. Without communication, group members submit numbers in an attempt to collectively sum to a randomly selected target number. After receiving group feedback, members adjust their submitted numbers until the target number is reached. For all groups, performance improves with task experience, and group reactivity decreases over rounds.

Our empirical results provide evidence for adaptive coordination in human groups, and as the coordination costs increase with group size, large groups adapt through spontaneous role differentiation and self-consistency among members. We suggest several agent-based models with different rules for agent reactions, and we show that the empirical results are best fit by a flexible, adaptive agent strategy in which agents decrease their reactions when the group feedback changes. The task offers a simple experimental platform for studying the general problem of group coordination while maximizing group returns, and we distinguish the task from several games in behavioral game theory.

The emergence of conventions in real-time environments

Robert Hawkins

Why are some behaviors governed by strong social conventions while others are not? I will present data from real-time, multi-player experiments that explore factors involved in synchronic convention-formation across two domains: behavioral game theory and language understanding.

In the first domain, we manipulated two factors in a game of impure coordination: the continuity of interaction within each round of play and the stakes of the interaction. To maximize efficiency and fairness in this game, players must coordinate on one of two equally advantageous equilibria. We found that when stakes are low in a real-time environment, players can satisfactorily coordinate 'on the fly,' but when stakes are high, or information is scarce, there is increased pressure to establish and adhere to shared expectations that persist across rounds.

Second, we ran a series of closed group, natural language communication games where the speaker produces a referring expression to help the listener select a target stimulus from a set of three. The stimuli we used – colors, random splines, and tangrams – varied in the extent to which prior conventions exist, context is salient, and specific stimuli persist across the discourse environment. Exploratory analyses of these data begin to characterize the domain-general space of features that govern the emergence of local conventions.

Effects of inductive biases in iterated learning and coordination

Tom Griffiths

There are two dominant approaches to modeling language creation: iterated learning, in which the structure of languages emerge as they are transmitted from learner to learner, and coordination games, where the structure of languages emerge as the result of a joint communication task. In both cases, a natural question to ask is how the inductive biases of learners – the factors that make some languages easier or harder to learn – affect which structures emerge. One way to explore this question is to model learning as Bayesian inference, and ask how the prior distribution of the learners influences the outcome of iterated learning or coordination. For iterated learning, the answer is that the languages that emerge directly reflect this prior distribution, or exaggerate the biases expressed in the prior. I will present results that show that an analogous property holds for the coordination setting: simply engaging in a collaborative learning task results in languages that reflect the prior distribution, and actively seeking to create languages that are mutually comprehensible exaggerates the prior.

References

Centola, D., & Baronchelli, A. (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences*, *112*(7), 1989–1994.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive science*, *29*(5), 737–767.

Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science*, *17*(1), 10–15.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

The cultural evolution of cognition

Sieghard Beller (Sieghard.Beller@uib.no) **& Andrea Bender** (Andrea.Bender@uib.no)
Department of Psychosocial Science, University of Bergen, Norway

Christine A. Caldwell (c.a.caldwell@stir.ac.uk)
Behaviour and Evolution Research Group, Department of Psychology, University of Stirling, UK

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology & Cornell Cognitive Science Program, Cornell University, Ithaca, NY 14853, USA

Karenleigh A. Overmann (karenleigh.overmann@keble.ox.ac.uk)
School of Archaeology, University of Oxford, UK

Kenny Smith (Kenny.Smith@ed.ac.uk)
School of Philosophy, Psychology and Language Sciences, University of Edinburgh, UK

Keywords: cognitive evolution; culture; language; materiality; conceptual tools.

As humans, we share most of our biological make-up with our closest primate relatives, yet we stand out not just from them but from all other species with respect to our cultural diversity, our capacity for language, and particularly our cognitive skills. While of prime interest for the field of cognitive science, the still-open question of which factors gave rise to human uniqueness has rarely surfaced at its conferences or in its journals. A hallmark of humankind, our propensity for engaging in social interactions and cultural transmission, is likely one of the most essential preconditions of cognitive evolution, accompanied by both language (Christiansen & Kirby, 2003) and material culture (Malafouris, 2013). Accounting for evolutionary change in human cognition thus requires new conceptual frameworks that view our psychological, behavioral, and material capabilities as interacting in complex manners in a continuous process of co-evolution (Richerson & Christiansen, 2013).

Our symposium highlights the relevance of culture as a driving force in the evolution of human cognition, with a special focus on material culture, language, and conceptual tools. Researchers from various sub-fields of cognitive science—including archaeology, anthropology, linguistics, cognitive psychology, and comparative psychology—will discuss cultural factors that may have triggered, constrained, or shaped the course of cognitive evolution:

Caldwell uses experimental methods to study cultural evolution in human participants, and carries out comparative research on social learning with nonhuman primates. Cognitive scientist ***Christiansen*** employs a variety of computational, behavioral, and neuroimaging experiments to unravel the interaction between biological and environmental constraints in the evolution, acquisition and processing of language. ***Smith*** develops experimental and computational models of cultural evolution and the co-evolution of language and language learning biases and strategies. Cognitive archaeologist ***Overmann*** draws on insights from various subfields of cognitive science to investigate how mind, behavior, and material artifacts interact to change human cognition. ***Bender & Beller***, finally, take an anthropological perspective on cultural diversity to account for how conceptual tools for cognition may have emerged and evolved.

Together, these participants have published several dozen papers related to the topic, including in high-quality journals such as *Science*, *PNAS*, *BBS*, *Proceedings of the Royal Society*, *TiCS*, *Current Anthropology*, or *Cognition*.

Prerequisites of Cumulative Culture in Laboratory Studies of Cultural Evolution

Christine A. Caldwell

In humans, cultural traditions often change in ways which increase efficiency and functionality (widely referred to as cumulative cultural evolution). However, directional change of this kind appears to distinguish human cultural traditions from behavioral traditions that have been documented in other animals. We (Caldwell et al., 2016) have reviewed experimental studies of cultural evolution to consider how the task is framed for the participants, and what kind of change is being measured by the researchers. Contrasting the broad approaches in the literature provides some enlightening insights into the conditions necessary for eliciting the ratcheting pattern of change regarded as the hallmark of cumulative culture. These insights may shed light on the apparent rarity of cumulative culture in nonhumans.

Language Evolution as Cultural Evolution

Morten H. Christiansen

Culture in its many manifestations—social organization, technology, science, language, religion—is responsible for the striking difference between humans and other organisms. Over the past few decades, a growing body of research has emerged from a variety of disciplines highlighting the importance of cultural evolution in understanding human

behavior (Richerson & Christiansen, 2013). Within the language sciences, this has resulted in a shift toward explaining language evolution in terms of cultural evolution rather than biological adaptation. This work has demonstrated how various nonlinguistic biases amplified by cultural transmission across generations and pressures from interactions within generations may help explain many facets of linguistic structure observable in today's languages (e.g., Christiansen & Chater, 2008). A key challenge for future research is to identify specific constraints on learning and use that shape the various aspects of the cultural evolution of language. Ongoing work will be discussed to shed light on specific constraints on the cultural evolution of linguistic structure (Christiansen & Chater, 2016).

The Cultural Evolution of Linguistic Structure

Kenny Smith

The expressive power of human language comes from its structure: Language provides a generative system for building complex utterances which convey complex meanings. How have humans developed this unique and powerful system of communication? One classic explanation is that language structure represents a biological adaptation to a uniquely socially-interdependent human ecology. Recent evidence (from computational and experimental models of cultural transmission) suggests an alternative: linguistic structure develops gradually through cumulative cultural evolution, as a consequence of language learning and language use (e.g., Kirby et al., 2015). This work predicts that structure should arise from cultural transmission under fairly general conditions, a prediction borne out by recent work on cultural evolution of structure in non-humans (Claidière et al., 2014). These various sources of evidence suggest that we have structured language because we are social, but rather than language being a biological adaptation to this social ecology, it is primarily a cultural adaptation arising from our propensity to learn socially.

Literacy as Cognitive Change Emerging from Material Engagement

Karenleigh A. Overmann

Seen through *Material Engagement Theory* (Malafouris, 2013), literacy is a cognitive change emerging from specific interactions with material forms (Overmann, 2016). Over multiple generations, the manuovisual stimulation inherent in handwriting influences brain functionality and form, and change in psychological processing facilitates the manipulation of writing and scripts into new, stimulating forms with subsequent impact on neurological functioning. Long-term sustainment of the behaviors required to develop the necessary psychological and material changes is culturally motivated. While each individual must acquire the cognitive change needed to participate in the literate system (to whatever degree it has been developed), multi-generational involvement ensures the cumulative change remains synchronized to average cognitive capabilities. What is ultimately realized is beyond what any one individual could invent but something that most individuals can join. As a collective, cumulative process of cognitive change, literacy provides insight into the cultural evolution of human cognition through psychological–behavioral–material interactivity.

Puzzles in the Evolution of Numerical Cognition

Andrea Bender & Sieghard Beller

Counting is central to human cognition and cultural diversification, as it is one of the key activities that fostered abstraction and symbolic thinking at the dawn of humanity (Coolidge & Overmann, 2012). Then and now, however, the ability to count, and numeracy more generally, seems to presuppose a counting sequence, and this raises the question of how such a conceptual tool could be invented in the first place. And how would strikingly different conceptual tools or embodied experiences (Beller & Bender, 2008; Bender & Beller, 2012, 2014) stimulate the same decisive conceptual change believed to be required for eventually grasping the cardinality principle? Puzzles like these are presented, together with an overview of the attempts to tackle them.

References

Beller, S., & Bender, A. (2008). The limits of counting: Numerical cognition between evolution and culture. *Science, 319,* 213-215.

Bender, A., & Beller, S. (2012). Nature and culture of finger counting: Diversity and representational effects of an embodied cognitive tool. *Cognition, 124,* 156-182.

Bender, A., & Beller, S. (2014). Mangarevan invention of binary steps for easier calculation. *PNAS, 111,* 1322-1327.

Caldwell, C. A., Cornish, H., & Kandler, A. (2016, forthcoming). Identifying innovation in laboratory studies of cultural evolution: Rates of retention and measures of adaptation. *Philosophical Transactions of the Royal Society of London B.*

Caldwell, C. A., & Millen, A. (2009). Social learning mechanisms and cumulative cultural evolution: Is imitation necessary? *Psychological Science, 20*, 1478-1483.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral & Brain Sciences, 31*, 489-558.

Christiansen, M. H., & Chater, N. (2016). *Creating language*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences, 7,* 300-307.

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of a systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B, 281,* 20141541.

Coolidge, F. L., & Overmann, K. A. (2012). Numerosity, abstraction, and the emergence of symbolic thinking. *Current Anthropology, 53*, 204-225.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141,* 87-102.

Malafouris, L. (2013). *How things shape the mind*. Cambridge, MA: The MIT Press.

Overmann, K. A. (2016). Beyond writing: The development of literacy in the Ancient Near East. *Cambridge Archaeological Journal, 26*, 285-303.

Richerson, P. J., & Christiansen, M. H. (Eds.) (2013). *Cultural evolution.* Cambridge, MA: MIT Press.

Brain Science and Education: Is it Still a Bridge Too Far?

Organizers

Ray S. Perez (ray.perez@navy.mil)
Office of Naval Research
Danielle S. McNamara (dsmcnama@asu.edu)
Arizona State University
Gregg Solomon (gesolomo@nsf.gov)
National Science Foundation
Wayne D. Gray (grayw@rpi.edu)
Rensselaer Polytechnic Institute

Abstract

Does neuroscience have the potential to inform education? In this debate, three participants will describe emerging results in neuroscience with translatable links to learning and memory. Experts in learning sciences will discuss connections as well as limits in the ways that neuroscience can inform education at the individual and classroom levels. Although significant progress has been made in our understanding of how the brain learns and remembers, the question for this symposium is whether this progress does or can provide a *direct* bridge from brain science to education.

Keywords: neuroscience; brain science; translational research; education; brain-based education; educational neuroscience

Can our knowledge of brain structure and function be used to improve education? Twenty years ago, in "Education and the Brain: A Bridge Too Far" (1997), John Bruer warned that "we should be wary of the claim that neuroscience has much to tell us about education." His was a common position (e.g., Chipman, 1986). However, even at that time, many educational researchers (e.g., Mayer, 1986, Mayer et al., 1998; Brynes & Fox, 1998) disagreed with this view and argued that ignoring how the brain is structured and functions is a mistake and short sighted. Ten years later, the battle still raged with some researchers (e.g., Giedd & Rapport, 2010) arguing in favor of the potential impact of brain science on education, and others (e.g., Hirsh-Pasek & Bruer, 2007) maintaining the argument against, and further, that the perceived benefit of brain science to early childhood education was, in part, based on the misconception of critical periods for learning.

We have three goals for this symposium. First: to provide an assessment of the advances in cognitive neuroscience that are most relevant to the question of educational application. Second: to provide a forum for discussing translational research (i.e., the nature of the bridge) from the neuroscience to education. Third: to provide specific examples of promising brain science to education (e.g., reading, STEM).

The symposium features an introduction by John Bruer, four talks by eminent researchers in brain and cognitive science, a brief discussion by Ray Perez, and a 20 min Q&A among panel members and the audience. *John Bruer* will summarize the major issues in the debate. Presentations, by *Gary Lynch*, *Marcel Just*, and *Terrence Sejnowski*, will summarize the latest results from brain science at three different levels of analysis. Gary Lynch will describe current understandings of the neurobiology and neurophysiology of memory. Marcel Just and Robert Mason will describe their research at the cognitive neuroscience level, in particular the use of machine learning on fMRI data to chart neural signatures of concept formation. Terrence Sejnowski will provide a description of cortical models (modeling of network communication) of spatial and unsupervised learning of brain function and structure that could serve as a framework to integrate results of brain science across various levels of analysis.

Danielle S. McNamara raises questions on the potential impact of neuroscience on reading and literacy and how this research can influence neuroscience. *Richard Mayer*, more broadly, raises questions of how brain research can help educational psychologists better understand the nature of cognitive processing during learning (including markers of heavy or light cognitive load, or deep or shallow processing).

Discussant, *Ray Perez*, will briefly summarize both sides of the arguments followed by recommendations as to what could be the nature of translation research.

Symposium Presentations

Introduction: Revisiting The Bridge Too Far

John T. Bruer

John T. Bruer served as President of the James S. McDonnell Foundation from 1986 through 2015, initiating both *Cognitive Studies for Educational Practice* and the Foundation's program to support the development of cognitive neuroscience. More recently, Bruer helped to organize and fund the *Latin American School in the Neural, Cognitive and Educational Sciences*. His 1996 article, A Bridge Too Far, continues to be a lightning rod for debate about the promise and pitfalls of educational neuroscience. The paper's negative argument was that developmental synaptogenesis, critical periods, and enriched environments did not inform educational practice. Its positive argument, that links between neuroscience and education are most likely to be mediated by psychology, is still being debated. John Bruer will open up this discussion with a summary of the major issues in the debate and principle considerations, including the importance of identifying and prioritizing educational problems and ascertaining the levels of scientific analysis most likely to generate solutions in the near to intermediate future (see e.g., Bruer, 2016).

Brain Substrates for Complex Learning

Gary Lynch, University of California Irvine

Decades of work have led neuroscientists to a detailed, behaviorally predictive model describing the synaptic plasticity required for encoding memory. Substantial progress has also been made in defining forebrain networks

related to fundamental aspects of cognition; combined, these developments describe realistic bridges between brain science and education. For example, recent studies on the synaptic model uncovered a physiological/biochemical analogue of *distributed practice effects* in learning: neurobiological timing rules successfully predict optimal training regimens. Other work has linked episodic memory to specific nodes and types of plasticity in hippocampal networks; individuals need to engage this form of memory when learning how to deal with complexities of real world circumstances. Advances in this area are expected to suggest mechanism-based strategies for optimization and enhancement of teaching.

Using Brain Reading to Understand Concept Learning in Educational Contexts

Marcel Adam Just and Robert Mason
Carnegie Mellon University

Advances in identifying neural signatures of a concept renders it possible to determine the neural basis of the learning of scientific and technical knowledge. Just and Mason present findings observing (a) a high degree of commonality across people in representations of concepts, including abstract scientific concepts; (b) the neural emergence of individual newly learned concepts can be observed, with the ability to predict each concept's brain locations and its neural similarity to other concepts; and (c) a set of concepts in a specific domain can be characterized in terms of a few key underlying neurosemantic dimensions of representation. Such findings (including classroom research) enable the beginning of a neurally informed understanding of how individual scientific concepts are learned, and can potentially lead to enhanced instruction.

Cortical Models of Brain Function

Terrence Sejnowski, UC San Diego and The Salk Institute

From a computational neuroscience perspective, Sejnowski describes research towards understanding the principles that link brain to behavior. Drawing upon studies that use multi-imaging techniques, such as, EEG, TMS, PET, MEG, and behavioral measures, his laboratory uses both experimental and modeling techniques to study the biophysical properties of synapses and neurons and the population dynamics of large networks of neurons. He has developed new computational models and new analytical tools to better understand how the brain represents the world and how new representations are formed through learning algorithms for changing the synaptic strengths of connections between neurons.

Bridges from Brain Science to Reading: Some Key Ingredients of a Direct Translation

Danielle S. McNamara, Arizona State University

Dr. McNamara, an expert in reading comprehension and reading interventions, will discuss potential links between brain science and reading, including neural correlates of strategic reading. She will discuss potentially promising results as well as inherent limitations of neural correlates in informing educational practice, and in particular, the necessity to inform interventions for individuals at various stages of development.

How Can Brain Research Inform Academic Learning and Instruction?

Richard E. Mayer, UC Santa Barbara

Renowned for his work in applying the science of learning to education, Mayer will discuss ways that brain research may help educational psychologists understand the nature of cognitive processing during learning. From the standpoint of educational psychology, he will describe work connecting brain signatures (including markers of heavy or light cognitive load, or deep or shallow processing) and (including markers for meaningful or rote learning) and what brain research can tell us about individual differences in learning strategies and the effectiveness of instructional manipulations.

Discussant: Ray S. Perez, senior scientist and Program Officer at the Office of Naval Research, is responsible for managing ONR's Cognitive Science of Learning Program, with three major multidisciplinary and highly intertwined thrusts. Specifically, (1) training/education research and their core technologies, (2) individual differences research, and (3) neuro-biology of learning research.

References

Bruer, J. (1997). Education and the brain: A bridge too far. *Educational Researcher, 26*(8), 4-16.

Bruer, J. (2016). Where is educational neuroscience? *Educational Neuroscience, 1*, 1-12.

Chipman, S. (1986). Integrating three perspectives on Learning. In S.L. Friedman, K.A. Klivington, &R.W. Peterson (Eds.) *The Brain Cognition and Education* (pp. 203-229). San Diego, CA: Academic Press.

Fiorella, L. & Mayer, R. E. (2015). *Learning as a generative activity: Eight learning strategies that promote understanding.* Cambridge Press, New York.

Laughlin, S. B., & Sejnowski, T. J., (2003). Communication in neuronal networks, *Science, 301*, 1870-1874.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., … Ecker, J. R., (2013). Global epigenomic reconfiguration during mammalian brain development, *Science, 341*, 629.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*, 1191–1195.

Moss, J., Schunn, C. D., Schneider, W., McNamara, D. S., & VanLehn, K. (2011). The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension. *NeuroImage, 58*, 675-686.

Trieu, B. H., Kramar, E. A., Wang, W. Cox, C. D., Gall, C. M., & Lynch, G. (2015). Differences in adenosine processing create marked regional variations in post-induction shaping of memory related synaptic changes. *J. Physiology, 593*, 2889-2907.

Research at the Interface of Cognition, Education, and Disciplinary Science

Laura R. Novick (Laura.Novick@vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University
Nashville, TN 37212 USA

Mary Hegarty (mary.hegarty@psych.ucsb.edu)
Department of Psychological and Brain Sciences, University of California, Santa Barbara
Santa Barbara, CA 93106 USA

Richard Catrambone (richard.catrambone@psych.gatech.edu)
School of Psychology, Georgia Institute of Technology
Atlanta, GA 30332 USA

John R. Pani (jrpani@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292 USA

Thomas F. Shipley (Tshipley@temple.edu)
Department of Psychology, Temple University
Philadelphia, PA 19122 USA

Keywords: STEM education; science literacy; representations; processes; learning

Overview

Meeting the challenges of the 21st century will require both workers who are highly competent in fields of science and engineering and citizens who have a high degree of science literacy. Accomplishing these goals will, in turn, require changes to undergraduate science and engineering education (National Research Council, 2012). With their expertise in learning, memory, problem solving, and diagrammatic and spatial reasoning, cognitive scientists are ideally situated to make important contributions to this endeavor.

To gain competence in STEM fields, students must learn to engage in two distinct types of reasoning: (a) how to interpret and reason with domain-specific representations such as diagrams of evolutionary trees, digital images showing brain anatomy, and 3-D models of chemistry molecules; (b) how to reason about processes that operate on those structures. For example, a variety of transformational processes (e.g., compression, tension, shear) operate to produce geologic events (e.g., folds, faults). The distinction between structures and processes, as well as the interplay between them, is a core issue in cognitive science.

This symposium, moderated by Novick and Hegarty, includes five talks, by leading experts in their fields, on current research at the interface of cognitive science, education, and other science disciplines. Three talks will consider how cognitive and perceptual principles affect comprehension and learning of science concepts and disciplinary representations. Two talks will examine how investigating representations, reasoning strategies, and science practices broadens our understanding of spatial thinking and reveals previously unrecognized spatial reasoning processes.

Physics

Richard Catrambone will talk about research comparing concept-based and traditional curricula for teaching introductory physics (e.g., Caballero et al., 2012). Physics curricula fail to prepare students for solving basic physics problems. The Matter and Interactions (M&I) curriculum was designed to improve learning by organizing concepts around fundamental principles. "Traditional" course students often learn to use special case formulas. Catrambone and colleagues expected that M&I students would learn to approach problems from first principles, but they have found little evidence of this. M&I students' think-aloud protocols during problem solving involving the Force Concept Inventory (FCI) indicate that they failed to employ the fundamental principles around which their curriculum was designed. When presented with probes, students answered using the language from their respective (M&I or traditional) courses. While this suggests the courses have some impact, it does not appear to extend to solution procedures. This has implications for claims that courses created around core concepts can successfully lead learners to reason in domain-appropriate ways.

Neuroanatomy

John Pani will talk about research on developing and testing a computer-assisted system for teaching neuroanatomy (e.g., Pani et al., 2014). Neuroanatomy is a core discipline that is perceived to have a high level of difficulty for obtaining even basic expertise. Students must learn to describe a large

domain that is complex both in structure and function. Modern digital illustration and computational systems have great promise for reducing the barriers to learning in such a field, and yet progress to date has been disappointing. Pani and colleagues' research in learning neuroanatomy suggests that cognitive science can help to design, test, and develop digital systems that will dramatically improve the educational experience in this area. Pani will review how a variety of design principles, including intuitive exploratory graphics, efficient mapping of organizational frameworks, frequent testing with immediate feedback, and continuous adaptive learning have generated successful instructional systems.

Biology

Laura Novick will talk about research on students' ability to interpret and reason with the information depicted in evolutionary trees (e.g., Novick & Catley, 2013). Diagrams are important cultural tools that are central to conceptual understanding and theoretical development in many scientific disciplines. Contrary to what many people believe, however, diagrams do not transparently convey the intended meaning about the relationships among the concepts being depicted. The elements of which diagrams are composed (e.g., lines, arrows, circles) are meaningful symbols in their own right, and they may suggest interpretations that either support or conflict with the scientifically appropriate interpretation of the diagram. How the human perceptual system works also affects students' interpretations of diagrams. In particular, Gestalt principles of grouping (e.g., good continuation, connectedness) play an important role. Finally, prior experience with a superficially similar task—reading text—affects the meaning that students extract from diagrams. Novick will describe research on how these factors affect undergraduate students' interpretations of tree-of-life diagrams from biology, which depict evolutionary relationships among taxa.

Chemistry

Mary Hegarty will talk about research on how to train spatial intelligence in the domain of organic chemistry (e.g., Stieff et al., 2014). In recent years there has been new recognition of the importance of spatial thinking in science. Spatial visualization ability is related to success in science disciplines and performance on spatial ability tests can be enhanced by training. As a result, current approaches aim to increase science achievement by recruiting individuals with high spatial ability, or by training of general spatial abilities. Hegarty and colleagues' alternative approach is to characterize spatially intelligent activities as they occur in science disciplines and focus on how to train these intelligent activities. Focusing on the domain of organic chemistry, they have found that spatial visualization is just one component of spatial thinking in science. Other components include adaptive strategy choice between imagistic and more analytic thinking, and proficiency in using multiple spatial representations. Based on these insights, and recent intervention studies, Hegarty will discuss how to best nurture spatial thinking in science.

Geoscience

Tim Shipley will talk about research on spatial reasoning in the geosciences (e.g., Resnick & Shipley, 2013). Understanding Earth's physical structure and history requires characterization of 3D solid structures from surface observations, and using the current state of the Earth to reason about the processes that have acted on it in the past. Shipley and colleagues have identified a previously unrecognized visual completion process that allows visualizing interior structures of objects but results in systematic errors inferring 3D forms from single cross-sections. Reasoning about Earth processes requires simulation of both rigid events (e.g., rotation—the traditional subject of mental event simulation research) and non-rigid events (e.g., bending and breaking). Expert geologists excel at visualizing both types of events, whereas expert chemists excel only on rigid rotations. Shipley will discuss these findings and their implications for supporting students as they struggle with spatial challenges in STEM practice and training.

References

Caballero, M.D., Greco, E.F., Murray, E.R., Bujak, K.R, Marr, M.J., Catrambone, R., Kohlmyer, M.A., & Schatz, M.F. (2012). Comparing large lecture mechanics curricula using the Force Concept Inventory: A five thousand student study. *American Journal of Physics, 80* (7), 638-644.

National Research Council. (2012). *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering.* S. R. Singer, N. R. Nielsen, & H. A. Schweingruber (Eds.), Committee on the Status, Contributions, and Future Directions of Discipline-Based Education Research. Board on Science Education, Division of Behavioral and Social Sciences and Education, Washington, DC: National Academies Press.

Novick, L. R., & Catley, K. M. (2013). Reasoning about evolution's grand patterns: College students' understanding of the tree of life. *American Educational Research Journal, 50,* 138-177.

Pani, J. R., Chariker, J. H., & Naaz, F., Roberts, J., Sephton, S. E. (2014). Learning with interactive computer graphics in the undergraduate neuroscience classroom. *Advances in Health Sciences Education: Theory and Practice*, *19*, 507-528.

Resnick, I., & Shipley, T.F. (2013). Breaking new ground in the mind: An initial study of mental brittle transformation and mental rigid rotation in science experts. *Cognitive Processing*, 14 (2, May), 143-152.

Stieff, M., Dixon, B. L., Ryu, M., Kumi, B. C., & Hegarty, M. (2014). Strategy training eliminates sex differences in spatial problem solving in a stem domain. *Journal of Educational Psychology*, *106*, 390-402.

Beyond the language explosion: What gradual word learning tells us about conceptual development

Katharine Tillman[1] **(katillman@ucsd.edu)**
Katie Wagner[1] **(kgwagner@ucsd.edu)**
Symposium Organizers

Junyi Chu (junyichu@ucsd.edu)
[1]Department of Psychology
University of California, San Diego, USA

Mutsumi Imai (imai@sfc.keio.ac.jp)
Environment and Information Studies
Keio University, Japan

Barbara Malt (bcm0@lehigh.edu)
Department of Psychology
Lehigh University, USA

Sherri Widen (swiden@stanford.edu)
Graduate School of Education
Stanford University, USA

Marilyn Shatz (mshatz@umich.edu)
Departments of Psychology and Linguistics
University of Michigan, USA
Symposium Moderator

Keywords: word learning, language acquisition, development, concepts, number, color, emotion, locomotion, time

Introduction

For the past 50 years, language acquisition researchers have explored the fascinating question of "the language explosion": how children are able to rapidly acquire hundreds to thousands of new words in a brief period of time. Focusing primarily on how basic-level nouns, such as "cup", are quickly mapped to visible objects, researchers have proposed several constraints guiding early word learning. Their insights include the principle of *mutual exclusivity*, the *whole object bias*, the principle of *contrast*, *fast mapping*, and the notion of "comprehension before production" (e.g., Carey & Bartlett, 1978; Clark, 1988; Markman, 1994). Such work has greatly enhanced the field's understanding of how rapid early word learning can be achieved.

Much less is known about how children acquire words that do not refer to objects, how children's word meanings may differ from those of adults, and the inferential processes that take children from their initial mappings to adult-like meanings. In striking contrast to the notion of rapid word learning, we here explore the claim that in some cases, despite early production, children require additional months or even years to achieve adult-like comprehension (e.g., Ameel et al., 2008; Saji et al., 2011; Tillman & Barner, 2015; Wagner, Dobkins, & Barner, 2013; Widen & Russell, 2008; Wynn, 1990). This symposium brings together researchers studying five conceptual domains in which children's word meanings evolve gradually — color, number, emotion, locomotion, and time — to explore the inferential processes children use to overcome their early assumptions and acquire adult meanings. In each case, we discuss how linguistic and perceptual cues come together to aid the construction of an abstract, domain-specific lexical category, how that category grows and gains internal structure, and how this structure in turn constrains and refines the meanings of individual words within it.

Imai discusses children's acquisition of *color words*, like "red", which are initially mapped quickly and coarsely to perceptual referents, but whose meanings must then be gradually refined to language-specific boundaries. In a similar vein, **Widen** considers how children's use of *emotion labels*, like "sadness", narrows as children acquire new words, like "embarrassment". **Chu** argues that children initially have inexact meanings for *number words*, like "four", long before they acquire their exact meanings and learn how counting works. **Malt** examines children's early meanings for *locomotion words*, like "running" and "jumping", and finds that although children overextend their early meanings, they rely on cues from the biomechanics of gait to help restrict the set of meanings they consider. **Tillman** explores how children learn time words, like "hour," which lack visible referents, increasing children's reliance on linguistic structure to constrain early meanings *prior* to mapping the words to experiences. Finally, **Shatz,** who has examined abstract word learning processes in several of these domains (e.g., Shatz et al., 1996; 2010), will serve as our moderator and discussant.

How young children construct the lexicon as a connected system: The case of color names

(Imai *with N. Saji, M. Asano, Y. Ujihara, K. Yasufuku, M. Ebe & M. Ohba***)**

To acquire adult-like word meanings, children need to construct a system in which the clusters of words in each semantic domain are mutually related, while the boundaries among these words are delineated. This is particularly true for the color lexicon. To investigate the full developmental trajectory through which children construct the color lexicon, we conducted two studies. Study 1 examined how children who knew few color words became able to map eight basic color names (red, blue, yellow, green, purple, pink, brown, orange) to their typical referents, by testing children every month, starting at 24-months of age and ending when the child was able to map all eight words correctly.. In Study 2, three-, 4-, and 5-year-olds and adults were tested on their naming of 93 samples of both chromatic

and achromatic colors. The two studies together showed that children go through a process of continuous reorganization to construct the connected system of color words, but the nature of the reorganization is very different for the initial and later stages of development. In the initial stages, children build up the frame of the system quickly with coarse word-referent mappings; they then refine the system by modifying the boundaries of each word through a slow learning process.

Early inexact meanings of number words

(**Chu** *with K. Wagner and D. Barner)*

Children acquire exact meanings for number words in a series of distinct stages. First, they learn the exact meaning for *one,* then *two,* and then *three.* Sometime after learning *three* or *four,* children learn to apply the counting procedure to their entire count list. Although the stages of acquisition of number word meanings are ubiquitous and well documented, the foundation of these meanings remains highly contested. Here we ask whether or not children assign preliminary meanings to number words prior to learning their exact meanings, by examining their responses on the Give-a-Number task to numbers for which they do not yet have exact meanings. While several research groups have approached this question before, we argue that because these data do not usually conform to a normal distribution, typical methods of analysis likely underestimate their knowledge. Using non-parametric analyses, we show that children acquire non-exact meanings for small number words like *one, two, three, four* and possibly for higher numbers well before they acquire the exact meanings.

The gradual development of children's understanding of emotion labels (Widen)

Toddlers use a variety of emotion labels in spontaneous conversation and children's emotion vocabulary increases with age. But do children understand these labels in the same way adults do – as fairly discrete categories? Across several studies in which children (2-18 years) labeled facial expressions and emotion stories, children's use of emotion labels suggests that their initial understanding of emotions words such as *happy*, *sad*, and *angry* is initially broad and valence based and narrows only gradually with age. The breadth of children's use of emotion labels reflects the nature of their underlying emotion concepts. In addition, later-emerging emotion concepts are initially incorporated in the early, broad concepts (e.g., embarrassment is a part of sadness). Thus, even though children use emotion labels from an early age, their understanding of these labels is initially different from adults' and only gradually do they acquire the adult taxonomy of emotions.

Do children use biomechanical structure to constrain hypotheses about locomotion word meaning? (**Malt** *with E. Ameel, A. White, and G. Storms***)**

Children refine word meanings and patterns of word use into middle childhood and beyond, even for common words appearing in early vocabulary. To address where children past toddlerhood diverge and where they correspond to adults in word use and why, we examined naming of instances of locomotion (walking, running, hopping, etc.) by children aged four to nine. Their performance was compared to that of adults. We evaluated whether the children are sensitive to discontinuities in the biomechanical properties of these actions that constrain adult word use. We also evaluated whether they apply this constraint by age four or only later in childhood. Results of two studies revealed that the children from four onward exploit biomechanical discontinuities to constrain their choice of label for actions. They do so despite having fewer words than adults to cover the domain, resulting in overextension of some terms relative to adults. Perceived domain structure plays a role in explaining later lexical development.

Conceptual structure and the semantics of time words (**Tillman** *with D. Barner*)

How do children learn to label a dimension of experience that cannot be directly perceived? Children produce duration words, like "minute" and "hour," as young as age 2. However, evidence suggests that they may take up to an additional *five years* to acquire adult-like meanings for such words. Here, I will discuss what children think these words mean during the long delay between initial production and mature comprehension, and the linguistic cues they use to build these meanings. Particular attention will be paid to children's acquisition of word ordering within a common semantic domain (e.g., *second* < *minute* < *hour*). While in many cases, mapping a new word to a perceptual experience is the first step in the acquisition process, in the case of time, I will argue that it is the one of the last.

References

Ameel, E., Malt, B., & Storms, G. (2008). Object naming and later lexical development: From baby bottle to beer bottle. *Journal of Memory and Language*, *58*(2), 262-285.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, 15, 317-336.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199-227.

Saji, N., Imai, M., Saalbach, H., Zhang, Y., Shu, H., & Okada, H. (2011). Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain. *Cognition*,*118*, 45-61

Shatz, M., Tare, M., Nguyen, S. P., & Young, T. (2010). Acquiring non-object terms: The case for time words. *Journal of Cognition and Development*,*11*(1), 16-36.

Shatz, M., Behrend, D., Gelman, S., & Ebeling, K. S. (1996). Colour term knowledge in two-year-olds: Evidence for early competence. *Journal of Child Language*, *23*, 177-200.

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. *Cognitive Psychology*, *78*, 57-77.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition*, *127*(3), 307-317.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, *23*(2), 291-312.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

Aligning implicit learning and statistical learning: Two approaches, one phenomenon

Patrick Rebuschat (p.rebuschat@lancaster.ac.uk)
Department of Linguistics and English Language
Lancaster University, Lancaster LA1 4YL, UK

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology
Lancaster University, Lancaster LA1 4YL, UK

Nathaniel D. Anderson (nandrsn3@illinois.edu)
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

Christopher M. Conway (cconway@gsu.edu)
Department of Psychology, Georgia State University
Atlanta, GA 30303-3083, USA

Gary S. Dell (gdell@illinois.edu)
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

Rebecca Gomez (rgomez@u.arizona.edu)
Department of Psychology
University of Arizona, Tucson AZ 85721, USA

Keywords: Implicit learning; statistical learning

Goals and Scope

The past 15-20 years have witnessed a particularly strong interest in our ability to rapidly extract structured information from the environment. This fundamental process of human cognition is widely believed to underpin many complex behaviors – from language development and social interaction to intuitive decision making and music cognition – so this interest spans practically all branches of cognitive science. Research on this topic can be found in two related, yet traditionally distinct research strands, namely "implicit learning" (Reber, 1967) and "statistical learning" (Saffran, Aslin, & Newport, 1996).

Both lines of research focus on how we acquire information from complex stimulus domains and both rely heavily on the use of artificial systems (e.g., finite-state grammars, pseudoword lexicons). In typical experiments, participants are initially exposed to stimuli generated by an artificial system and then tested to determine what they have learned. Given these and other significant similarities, Perruchet and Pacton (2006) argue that these distinct lines of research actually represent two approaches to a single phenomenon, and Conway and Christiansen (2006) propose combining the two in name: "implicit-statistical learning". Yet, despite frequent acknowledgements that researchers in implicit learning and statistical learning might essentially be looking at the same phenomenon, there is surprisingly little alignment between the two strands.

This symposium seeks to remedy this situation by bringing together leading researchers from both areas in order to promote a shared understanding of research questions and methodologies, to discuss similarities and differences between the two approaches, and to work towards a joint research agenda. The symposium comprises four presentations, followed by a thematic discussion, which provide coverage of these phenomena in terms of development (children and adults), different language learning tasks (sublexical phonotactics, word acquisition, grammar learning), and their role in both production and comprehension, each integrating multidisciplinary perspectives. Gomez focuses on implicit-statistical learning in early development, identifying words and grammatical sequences and the memory systems that underlie this learning. Monaghan and Rebuschat measure word learning and grammar learning in adults, while varying the knowledge that participants have of the structure they are acquiring. Dell and Anderson demonstrate how their work on acquisition of phonotactic constraints is exhibited in speakers' productions, and discuss the inter-relation in speech between implicit and statistical learning. Finally, Conway provides an overview of the two fields, and proposes a novel framework that unifies implicit learning and statistical learning.

The Contributions of Emerging Learning Systems to Implicit and Explicit Statistical Learning in Early Development (Gomez)

Here we propose that statistical learning changes qualitatively as a function of the different learning systems emerging across early development. Early in development, infants do not appear to retain veridical details of statistical learning. We argue that this stems from underdevelopment of the hippocampus, a structure well known for supporting the explicit episodic details of our experiences. Instead, very young infants rely on cortical learning which shows a gradual retention function requiring many exposures over time for consolidation in long-term memory. With increasing maturity in the second and third years of life the hippocampus begins to support rapid encoding and retention of details from learning, ensuring a stable initial memory trace and increased fidelity of memory. We argue that implicit, incremental cortical learning is advantageous in allowing infants to tune to fine details in their input while preventing retention of idiosyncratic patterns hippocampal learning would permit. By this view, the infant brain supports a very different profile of learning than is found in older learners with implications for theories of statistical and implicit learning.

A Single Paradigm for Implicit and Statistical Learning (Monaghan & Rebuschat)

There has been a traditional distinction in methods used to study implicit learning and statistical learning in language research. Implicit learning tended to address the unconscious compared to conscious knowledge of grammatical structures of sequences, whereas statistical learning manipulated the probabilities of sequences to measure the role of statistics in acquisition of words or simple local constraints in grammatical sequences. This distinction between grammatical structures and learning of words, or simpler sequences, has in turn been related to the operation of procedural and declarative memory systems. We have developed an experimental paradigm that enables simultaneous testing of both classic implicit learning paradigms and statistical learning of sequence structure. The paradigm involves artificial grammar sentences accompanying two dynamic scenes, one of which is described by the sentence. Through cross-situational learning, participants are able to determine the target scene. In a series of studies, we show that acquisition of both vocabulary and grammar is promoted by explicit information about the language structure, that awareness of structure affects acquisition during learning, but is not distinctive at the endpoint of learning, and furthermore that effective statistical learning facilitates explicit knowledge about the structure. We show that two traditions of learning – statistical and implicit – can be conjoined in a single paradigm to explore both the phenomenological and learning consequences of statistical structural knowledge.

Implicit Learning of Phonotactic Distributions in Language Production (Dell & Anderson)

The statistical patterning of linguistic elements is implicitly learned from experience. This is as much true when this experience involves production as when it involves the processing of linguistic input. We review studies showing that speakers quickly and implicitly learn restrictions on the syllable positions of consonants as they recite lists of syllables. The learning is revealed in slips of the tongue. For example, if /f/ is artificially restricted to onset position throughout the course of an experiment, when an /f/ is erroneously produced instead of some other consonant, it nearly always occurs in onset position. For this symposium, we present experiments and simulations that use reversal shift to determine the nature of the learning. After implicitly learning, say, that /f/ must be an onset and /s/ must be a coda, the rule is then reversed, e.g. /s/ must be an onset and /f/ must be a coda. When a reversed rule is learned more slowly than the original learning (or a comparable new set of restrictions), this suggests that the original rule must be un-learned before the new rule can be learned, which is classically taken as evidence for purely associative learning. We propose that implicit sequence learning in production is purely associative in just this way.

Toward a Unifying Framework for Implicit Learning and Statistical Learning (Conway)

Despite attempts to unify implicit learning and statistical learning research (e.g., Perruchet & Pacton, 2006), these two areas continue to proceed relatively independently of the other. Perhaps one reason for this compartmentalization is that each research realm traditionally has relied on different types of tasks: the serial-reaction time and artificial grammar learning tasks for implicit learning research and the word segmentation task for statistical learning. In addition, there are currently very few theoretical frameworks that adequately encompass findings from both areas. In this paper, we argue that while implicit learning and statistical learning tasks have important methodological differences, all three tap into a common ability of humans and other organisms to become sensitive to underlying environmental regularities under incidental conditions and to make predictions (either implicitly or explicitly) about what will be experienced next. We then evaluate a number of current theories related to implicit and statistical learning (e.g., Daltrozzo & Conway, 2014; Frost et al., 2015). Based both upon this analysis and recent findings from our research group, we propose an integrative framework that attempts to provide a unifying account of learning while simultaneously recognizing the heterogeneity of the processes and mechanisms involved. Specifically, we propose that implicit-statistical learning relies upon a distributed network of brain areas and cognitive processes that involve both lower-level implicit-perceptual learning mechanisms and higher-level abstraction, integration, and prediction processes. We close by presenting testable predictions of this framework and areas for future research.

References

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities pitting abstract against stimulus-specific representations. *Psychological Science*, *17*(10), 905-912.

Daltrozzo, J., & Conway, C. M. (2014). Neurocognitive mechanisms of statistical-sequential learning: what do event-related potentials tell us? *Frontiers in Human Neuroscience*, *8*, 437.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117-125.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233-238.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Higher-Level Goals in the Processing of Human Action Events

Michelle L. Eisenberg (mleisenb@wustl.edu)
Jeffrey M. Zacks (jzacks@wustl.edu)
Shaney Flores (sflores@wustl.edu)
Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

Lauren H. Howard[1] **(lauren.howard@fandm.edu)**
Amanda L. Woodward[2] **(woodward@uchicago.edu)**
[1]Psychology Department, Franklin & Marshall College, Lancaster, PA, USA
[2]Department of Psychology, University of Chicago, Chicago, IL, USA

Jeff Loucks[1] **(jeff.loucks@uregina.ca)**
Andrew N. Meltzoff[2] **(meltzoff@uw.edu)**
[1]Psychology Department, University of Regina, Regina, SK, Canada
[2]Institute for Learning & Brain Sciences, University of Washington, Seattle, WA, USA

Richard P. Cooper (R.Cooper@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck, University of London, UK

Keywords: human action; goals; goal hierarchies; memory; cognitive development; prediction

The concept of a goal critically separates dynamic events involving humans from other events. Human behaviours are motivated by goals, which are known to the actor but typically inferred on the part of the observer. Goals can be hierarchical in nature, such that a collection of sub-goals (e.g., getting a mug, boiling water) can be nested under a higher-level goal (e.g., making tea), which can be further nested under an even higher-level goal (e.g., making breakfast).

The diverse set of talks in this symposia all highlight the foundational role that goals play in action processing and representation. Eisenberg et al. detail how online prediction of others' goals shapes observers' sampling of information during action observation. Howard and Woodward provide evidence that children's memory for non-human events can be facilitated by priming children with their own goal-directed actions. Loucks and Meltzoff highlight the importance of goal structure in children's memory for complex action sequences. Finally, Cooper presents a computational model to explain the emergence of goal-directed action hierarchies.

Eisenberg, Zacks, & Flores: Prediction Around Event Boundaries During Continuous Viewing of Naturalistic Activity

The ability to break down ongoing activity into meaningful events is integral for comprehension and memory. We propose that this ability is driven by monitoring prediction error during perception. As people view ongoing activity, they maintain a model of the current event and use it to make predictions about what is going to happen next. When these predictions fail, viewers must update their event model to better capture the ongoing activity; this is experienced as a boundary between events. This proposed mechanism entails that prediction performance is worse at event boundaries than during the middles of events.

Previous research has supported this hypothesis using an explicit prediction judgment task. Here, we tested the hypothesis that perceptual prediction fails at event boundaries using an implicit eye tracking measure. In a series of two studies, participants passively watched movies of actors completing everyday activities while their eyes were tracked using an eye-tracker. Participants then watched the movies again and identified the boundaries between meaningful units of activity. We hypothesized that participants would predictively look at the objects that the actor was about to touch during the few seconds before the actor contacted the objects. In addition, we hypothesized that participants would begin making predictive eye movements later in time when the object contact occurred close to an event boundary compared to when the object contact occurred within an event. The results provided support for both hypotheses. These results provide further evidence that prediction failures drive the subjective experience of an event boundary and suggest a novel and naturalistic method for studying prediction during comprehension of ongoing activity.

Howard & Woodward: The Effect of Action Priming on Children's Event Memory

Research has demonstrated that the mere inclusion of a person in an event boosts children's memory (Howard & Woodward, under review). However, the mechanism

underlying this effect has yet to be explored. One possibility is that the presence of a person primes children to think of discrete sequential steps as part of a higher-level goal, providing a structure that increases later recall. In fact, we know that even infants spontaneously segment events according to goals, and in adults segmentation according to goals relates to better event memory.

We explored whether priming children to think of goal-directed actions before encoding an event can improve sequential memory. Three-year-old children played either an action game (Action Priming) or sticker game (Control Priming) with a series of novel toys. Participants then watched two picture sequences on a Tobii eye-tracking monitor, each culminating in the creation of an object. Half of the children saw object pieces assembled by a person (Person condition) while half saw the pieces assembled without a person present (No-person condition). After a delay, children were given the pieces and asked to reconstruct the sequence from memory. Our hypotheses were that only action priming would facilitate recall in the No-person condition, and that there would be no benefit of either priming in the Person condition.

Results demonstrated that children in the Person condition remembered the same number of event steps regardless of priming. In contrast, children in the No-Person/Action group remembered significantly more steps than those in the No-Person/Control group, suggesting that action priming can increase memory for non-social events. These results suggest that early learning is strengthened by the goal structures inherently provided by social partners, and that framing even non-social events in terms of higher-level goals can increase later recall.

Loucks & Meltzoff: Higher-Level Goals Structure Children's Action Memory

Multitasking is a common occurrence in action. For example, one might be preparing breakfast at the same time as packing one's briefcase, flitting between these higher-level goals with the completion of individual sub-goals. Such events pose an unexplored representational problem for the observer: should the representation prioritize veridical sequential structure, at the cost of representing higher-level goals, or should it prioritize goals, at a similar cost to sequential structure? This representational problem may be significant for young observers, who are rapidly acquiring action goals. We previously used a deferred imitation task to demonstrate that three-year-old children's memory for familiar interleaved goals prioritizes goal organization over sequential organization (Loucks & Meltzoff, 2013).

In the current research, we more directly observed goal organization by manipulating children's exposure to higher-level goal information in two studies. In Study 1, children in the experimental group learned about a novel goal A on day 1, while children in the control group learned an unrelated novel goal (B). On day 2, all children were shown an interleaved event involving two goals – goal A and goal X – and then given a chance to imitate. Only experimental children organized their imitation according to higher-level goals, and also remembered significantly more actions overall. Study 2 demonstrated that prior goal knowledge also structures children's memory when it is gained in the form of a picture book on day 1.

We believe these results highlight both the early emergence and the benefit of goal prioritization in children's memory for others' action, and underscore the importance of this representation problem for theories of action representation.

Cooper: On the Acquisition of Goal Hierarchies: A Computational Model

Methods based on reinforcement learning, and in particular temporal difference learning, have proven to be highly effective in developing agents that can learn to perform goal-directed sequential behaviors. The direct application of such methods, however, can learn to achieve only one goal, and that goal must be known in advance (so that reinforcement can be provided when it is achieved). Thus, the classical approach does not address the questions of where goals come from or how goals inter-relate. I will consider these questions from the perspective of the contention scheduling model of routine action selection (Cooper & Shallice, 2000), which assumes that action sequences reflect the selection and execution of hierarchically organized goal-directed action schemas.

I will extend the existing model to action acquisition, showing how a set of action primitives may become associated with feature-based object representations through exploration, exploitation and habituation. Representations of action primitives and of objects form separate but connected interactive activation networks. Through experience, objects with similar action-related features come to trigger actions appropriate to those objects, and vice versa. It is further assumed that actions with "interesting" consequences (i.e., that lead to reliable changes to the world) are reified into schemas, which may themselves be invoked to achieve specific ends. Successive application of this approach results, after sufficient toying within a domain, in the emergence of hierarchically-structured goal-directed action schemas.

References

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4), 297-338.

Howard, L. H., Riggins, T., & Woodward, A. L. (under review). Learning from others: The effects of agency of event memory in young children.

Loucks, J., & Meltzoff, A. N. (2013). Goals influence memory and imitation for dynamic human action in 36-month-old children. *Scandinavian Journal of Psychology*, 54, 41-50.

Concepts from Event Semantics in Cognition

organizer and speaker
Alexis Wellwood (wellwood@northwestern.edu)
Department of Linguistics, Northwestern University

invited speakers
Jeremy Kuhn (jeremy.d.kuhn@gmail.com), Philippe Schlenker (philippe.schlenker@gmail.com), Carlo Geraci (carlo.geraci76@gmail.com), & Brent Strickland (stricklandbrent@gmail.com)
Département d'Etudes Cognitives, École Normale Supérieure, PSL Research University Institut Jean Nicod (ENS, EHESS, CNRS)

Susan Hespos (hespos@northwestern.edu) & Lance Rips (rips@northwestern.edu)
Department of Psychology, Northwestern University

E. Matthew Husband (matthew.husband@ling-phil.ox.ac.uk)
Faculty of Linguistics, Philology and Phonetics, University of Oxford

Alexander Williams (alxndrw@umd.edu)
Departments of Linguistics and Philosophy, University of Maryland

Keywords: events; objects; states; scales; mass/count; telicity; thematic roles; event participants; linguistic ontology

Concepts of Events

Common sense intuition distinguishes between events and regular objects; events *happen*, after all, and objects don't. This distinction is deployed in linguistics, psychology and philosophy. Linguists say that sentences describe events, while nouns describe objects. Psychologists describe the principles of event perception, and philosophers debate the metaphysics of event identity. But how do these various discussions relate to each other?

Casati and Varzi (2008) emphasize the importance of delimiting 'common sense' understandings from theoretical usages; it could turn out that the representations posited to explain object or event perception share important common features, and that neither obviously resemble what is actually 'out there'. Worse, what determines whether a given portion of experience falls into a particular event category can seem intimately bound up with the language that we use to describe it (e.g. *chase* versus *flee*).

A prominent strand of theorizing in natural language semantics holds that the basic logic of sentences involves a hidden 'event variable', posited to explain the intuitive validity of a wide variety of productive inferences in natural language (e.g. from *A kicked B in the shin* to *A kicked B*; Davidson 1967). It has sometimes also been taken to burden the theorist with substantial ontological commitments. In contrast, some views suggest that understanding the 'event variable' is a fundamentally cognitive question (see Pietroski 2015).

This symposium aims to illuminate the psychological notion of 'event' from the perspective of event semantics, connecting posits in formal semantics to shared aspects of our perception and cognition. For the semanticist, such attempts encourage a different understanding of the entities that populate our models. For the psychologist, they suggest a rich arena in which to derive empirically-testable predictions about the mind. And, for the philosopher, they could provide new insight into why the common sense notion of 'event' has the structure that it does (Casati & Varzi 2008).

Two questions are broadly relevant here: what is the relationship between formal semantics and cognitive science? And, how can results in formal semantics be used to generate predictions about how the mind works? More specifically, what structure does natural language semantics imply for event perception and cognition? Investigating these questions will enhance interdisciplinary research on events from a lesser-explored perspective in cognitive science.

Intuitive iconicity for events and objects

J. Kuhn, P. Schlenker, C. Geraci, B. Strickland

Telic verbs refer to events bound in time (*decide*), while atelic verbs (*think*) refer to events that are not. Analogously, count nouns typically refer to objects bound in space (*coin*) while mass nouns typically refer to entities that are not (*rain*). Semantic theory (Jackendoff 1991, Bach 1986) draws formal parallels between the logical properties of two domains. Here, we ask whether the notion of boundedness (in space or time) is present in the fundamental cognitive representation of the objects the words denote. A first experiment showed unfamiliar Italian Sign Language signs to non-signers. Participants readily assigned telic and count meanings to signs containing "gestural stops" while they assigned atelic and mass meanings to signs lacking such stops. A second experiment asked if similar results may extend in spoken language. Here, written non-words either contained or lacked a phonological stop. "Stop" words were again more readily assigned telic and count meanings while non-stop words were more readily assigned atelic and mass meanings. Collectively, these results suggest that the telic/atelic and count/mass distinctions play

similar roles in an abstract iconic system (mapping meaning to symbols), thus suggesting that these categories share deep similarities in how their typical referents are represented.

Object : substance :: event : process

A. Wellwood, S. Hespos, L. Rips

Beginning with Bach (1986), semanticists have suggested that the objects/events and substances/processes picked out by nouns and verbs are strongly parallel. We investigate whether these parallels can be understood to reflect a shared representational format in cognition. We hypothesized that a criterion for counting is necessary for 'object' and 'event' representations, unlike 'substance' or 'process' representations (cf. Barner et al 2008). This criterion is strongly implied by plural (e.g. *some gorps*, for novel *gorp*) but not mass language (e.g. *some gorp*). We tested the salience of numerical differences between pairs of minimally different images and animations, designed with 'natural' spatial gaps (images) or temporal gaps (animations), suggesting countability, versus 'unnatural'. We tested preference for matching these stimuli with mass or count language (Expt.1), then compared similarity ratings between pairs presented without linguistic information (Expt.2), with 'matching' linguistic information (Expt.3), and 'mismatching' linguistic information (Expt.4). Our results support Bach's analogy in perception, and highlight the role of countability in object and event represention.

Countability in eventualities and beyond

E. M. Husband

Researchers interested in how our linguistic system structures ontological semantic domains have observed striking parallels between the domain of objects in their count/mass interpretation and the domain of events in their aspectual classification. These parallels have guided research to formalize an underlying property of countability that is shared between the domains of objects and events (e.g. Bach 1986). In this talk, I provide evidence that the parallels found between the domains of objects and events can be extended to the domains of degrees and states with considerations of scale structure on the one side and existential interpretation on the other. Such evidence suggests a role for countability in our linguistic system than is broader and deeper than typically considered. More speculatively, I observe that countability may extend beyond the linguistic system into other cognitive domains. The visual system, for instance, shows certain signature behaviors that distinguish between objects and visual substances. This suggests countability may be a core property reflecting the organization of our minds more broadly.

Thematic relations in different views of meaning

A. Williams

I discuss how different views of semantics affect our understanding of so-called event variables. After Davidson (1967), the semantics of complex predicates has been used to support the view that subjects, objects, adverbs and verbs contribute separate predicates of the same variable, call it an E. The predicates that result from this analysis, such as 'Agent' and 'Patient', have forced either rejection of the analysis (Dowty 1991), or one of three accommodations. The choice depends on the background understanding of sentence meanings and their Es. Taking Es to range over events leads to a surprising metaphysics (Parsons 1990). Taking them to range over events-under-a-perspective leads to a surprising psychology (Schein 2002). The third, less familiar response has Es range over the structured thoughts that sentence-tokens express (Pietroski 2015). 'Agent' and 'Patient' then indicate, not properties satisfied by events(-as-we-view-them), but different positions in a relational thought. This avoids the other problems, but implies a very different view of the generalizations that initially motivate the analysis. I map this landscape to facilitate discussion among linguists and psychologists on the varying relevance of terms like 'Agent' and 'Patient' in our understanding of inference, event perception and language acquisition.

Acknowledgments

Kuhn et al's research was supported by the ERC under the EU's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement 324115-FRONTSEM and ERC Grant 313610-SEMEXP, and conducted at the Institut d'Etudes Cognitives, École Normale Supérieure–PSL, supported by grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*, as well as by a Fyssen Foundation Postdoctoral grant.

References

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, *9*(1), 5-16.

Barner, D., Wagner, L., & Snedeker, J. (2008). Events and the ontology of individuals: Verbs as a source of individuating mass and count nouns. *Cognition*, *106*, 805-832.

Casati, R., & Varzi, A. C. (2008). Event concepts. In T. Shipley & J. Zacks (Eds.), *Understanding events* (p. 31-53). New York: Oxford University Press.

Davidson, D. (1967). The logical form of action sentences. In N. Rescher (Ed.), *The logic of decision and action* (pp. 81–95). Pittsburgh: Pittsburgh University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*, 547-619.

Jackendoff, R. (1991). Parts and Boundaries. *Cognition*, *41*(9-45).

Parsons, T. (1990). Events in the semantics of English: A study in subatomic semantics. In *Current studies in linguistics series no. 19* (p. 334). Cambridge, Massachusetts: MIT Press.

Pietroski, P. (2015). Framing event variables. *Erkenntnis*, *80*(1), 31-60.

Schein, B. (2002). Events and the semantic content of thematic relations. In G. Preyer & G. Peter (Eds.), *Logical Form & Language* (p. 263-344). Oxford University Press.

Investigating the Effects of Transparency and Ambiguity on Idiom Learning

Mehrgol Tiv (met81@pitt.edu)
University of Pittsburgh
4200 Fifth Ave.
Pittsburgh, PA 15260

Evelyn Milburn (eam115@pitt.edu)
University of Pittsburgh
4200 Fifth Ave.
Pittsburgh, PA 15260

Tessa Warren (tessa@pitt.edu)
University of Pittsburgh
4200 Fifth Ave.
Pittsburgh, PA 15260

Abstract

The purpose of this study was to learn how transparency and ambiguity affect idiom learning. To start, 157 French idioms were translated to English and normed for familiarity, transparency, and ambiguity. Experiment 1 was a training study in which 32 of these idioms were taught to 25 native English speakers over two days of training. A cued recall test during a third session showed a reliable effect of transparency, but performance was close to ceiling. In Experiment 2, the amount of training was reduced to one session and a semantic relatedness test was included after the cued recall test. The results of Experiment 2 suggest that high transparency idioms are recalled with greater accuracy in a cued-recall test but low transparency idioms are recalled with greater accuracy in a semantic relatedness test. No significant effect of or interaction with ambiguity was found.

Keywords: Idioms; Language Learning; Figurative Language; Individual Differences

Background

Figurative language, which includes phenomena such as metaphor, metonymy, irony, and idioms, is generally characterized as "going beyond the literal" (Gibbs, 1994). Of these, idioms have been of particular interest to cognitive linguists and psycholinguists. Idioms are traditionally defined by their conventionality; as Nunberg, Sag, and Wasow stated: "their meaning or use can't be predicted, or at least entirely predicted, on the basis of a knowledge of the independent conventions that determine the use of their constituents when they appear in isolation from one another" (1994). The tension that arises between overall expression meaning and literal word meaning gives rise to several unique characteristics, two of which, transparency and ambiguity, will be considered here.

Most research on idiomatic language has focused on how idioms are mentally represented and processed. There has been less investigation of how idioms are learned, despite the practical ramifications it bears for second language learning applications (e.g. Howarth, 1998; Yorio, 1989). The current study investigates the way that the properties of transparency and ambiguity affect idiom learning.

Transparency refers to the ease with which the comprehender can make a connection between the idiom's literal and figurative meanings (Nunberg et al., 1994). The idiom *spill the beans* is highly transparent because the connection between "divulging a secret" and "releasing beans from a container" is very clear. In contrast, the idiom *wet behind the ears* is less transparent because the relationship between the literal and figurative meaning is less obvious. Several factors, including properties of individual words (compositionality) and the capacity to evoke a mental image (imageability), can influence the ease of drawing a connection between literal and figurative meanings and thus contribute to an idiom's level of transparency (Nunberg et al., 1994).

Ambiguity refers to whether an idiom has both literal and figurative meanings, or only a figurative meaning. The idiom *kick the bucket* is ambiguous because it has both a literal meaning (striking a pail with a foot) and a figurative meaning (dying), but the idiom *under the weather* is unambiguous because it only has a figurative meaning (Titone & Connine, 1994).

We were interested in whether transparency and ambiguity influence how successfully individuals will learn the meanings of idioms. There is evidence that transparency and ambiguity both influence the comprehension of idioms, such that greater transparency tends to facilitate idiom comprehension, whereas greater ambiguity usually impedes comprehension (Foss, 1970), but there is only a small literature on how these factors influence learning. In one of the most relevant studies, Steinhel et al. (2007) investigated the effects of transparency on L2 idiom learning and found minimal effects of transparency but significant effects of imageability. However, their study included very few items

in each condition, so its findings must be interpreted with caution.

Most previous idiom learning studies have focused on learning idioms in a second language (e.g. Steinhel et al., 2007) and on the way that context affects learning (e.g. Zyzik, 2011). The current study is novel because it investigates idiom learning in the participants' native language and uses completely unfamiliar idioms, thereby avoiding any learning advantage from previous exposure. This is critical because there is evidence that comprehenders find more familiar idioms more transparent (Keysar & Bly, 1995). We presented participants with unfamiliar phrases that had no English idiomatic meaning. This process, in which the learner must assign idiomatic meaning to unfamiliar strings of words, mirrored natural idiom acquisition and allowed us to examine how unfamiliar idiomatic meanings are learned.

We base our predictions on findings from the growing literature on ambiguous word learning, under the assumption that assigning a new conventionalized meaning for a phrase that already has a compositional meaning is like learning a new meaning for a word that already has a meaning. Given evidence that ambiguous words are harder to learn than unambiguous words (Degani & Tokowicz, 2010), we predict that figurative meanings of less ambiguous idioms will be learned better than figurative meanings of more ambiguous idioms. This is because literal meanings of ambiguous idioms are highly accessible and may interfere with learning of figurative meanings. Drawing from work showing that when new meanings are assigned to old words, increased semantic similarity between the meanings leads to increased recall accuracy (Rodd, Berriman, Landau, Lee, Ho, Gaskell, & Davis, 2012), we also predict that figurative meanings of highly transparent idioms will be learned better than figurative meanings of less transparent idioms. This is because participants will be able to more easily draw connections between the idioms' literal and figurative meanings, thereby boosting recall.

Finally, we investigated three individual difference factors that could interact with the learning process and influence performance. The first was performance on the Operation-Span (O-Span) task (Turner & Engle, 1989), a measure of working memory. Because working memory is important for learning, we predict that individuals who score higher on an O-Span task will learn better than individuals that score lower on the O-Span task (Unsworth, Heitz, Schrock, & Engle, 2005). Additionally, we predict that more creative individuals will be able to create more or stronger connections between the literal and figurative meanings of the idioms, making the low transparency idioms seem more transparent to them. Thus, we expect that participants who score higher on the Abbreviated Torrance Test for Adults (ATTA) (Goff, 2002) will learn better and show less of a difference between performance on high and low transparency idioms. Finally, we expected that better performance on an Author Recognition Test (Stanovich & West, 1989), a proxy for language experience and thus figurative language proficiency, would also be predictive of idiomatic learning.

Norming Idiom Properties

To inform the choice of stimuli for the training study, we conducted multiple norming questionnaires.

Methods

Participants 85 undergraduate students enrolled in Introduction to Psychology at the University of Pittsburgh participated in one of various norming questionnaires for course credit. All participants were over 18 years old, native speakers of American English, and had no prior experience with French.

Materials 305 French idioms were translated word-for-word (literally) and as an idiomatic phrase (nonliterally) using online and print resources (Bonin, Meot, & Bugaiska, 2013; "Idioms," 2006; Lamiroy & Klein, 2010; "Proverbes en français," 2006; "Traduction en contexte," 2006; "Word Reference") For example, the French idiom *avoir un cheveu sur la langue* was translated literally as *to have a hair on your tongue* and nonliterally as *to have a lisp.* French idioms that were extremely similar to English idioms were excluded, leaving 157 translated idioms in the set that was normed.

Procedures We designed online Qualtrics questionnaires to evaluate familiarity, transparency, and ambiguity. Each participant was randomly assigned to one questionnaire. Each questionnaire contained several catch trials to confirm that participants were on task and attentive. All participants completed a brief language history questionnaire.

To verify that the translated idioms were not familiar, 15 participants were shown each idiom and asked to "determine if there is an English phrase that is similar to and means the same thing as the following expression" by answering "Yes" or "No." Any idiom that received more than one "Yes" response was excluded from the set of stimuli.

To measure the level of transparency for each idiom, we presented the literal translation of each idiom (e.g. *to have a hair on your* tongue) with its figurative translation (e.g. *to have a lisp*) and asked participants to rate how likely from 1 (*Very unlikely*) to 5 (*Very likely*) they were to have the same meaning. Each opportunity for rating was followed by a free-response question: "why did you select that rating?" Piloting trials indicated that the questionnaire could take over an hour to complete, so, in order to avoid fatigue, each participant received half of the questionnaire. Fourteen participants completed the first half, and 13 completed the second half.

To index ambiguity, we evaluated the salience of the literal interpretation of the idiom by asking participants to "decide which [interpretation] you think is the more likely interpretation by assigning each option with a percentage (0-100). A higher percentage indicates the more likely

interpretation." For example, for the idiom *to have a hair on your tongue* participants would have to assign a percentage to each of the two options: (1) *to have a hair resting on your* tongue and (2) *to have a lisp*. Increasing the likelihood of the literal meaning increases the chance that it will compete with the figurative meaning. Therefore, we used the rated likelihood of the literal meaning (option 1) as our index of the ambiguity of the idiom. Participants were additionally asked to "indicate how likely (from *very unlikely* to *very likely*) you would be to encounter these expressions when reading a novel or talking with someone." This questionnaire was also time-consuming so it was similarly split in half. Fifteen participants completed each half.

Results

Average ambiguity and transparency scores for each idiom were calculated. We selected 32 idioms that varied across the full range of transparency and ambiguity scores, yet also satisfied a 2x2 design crossing transparency and ambiguity. Table 1 gives examples from each condition.

Table 1: Example Idioms Used in Training Study

Idiom	Translation	Transparency	Ambiguity
To have a hair on your tongue	To have a lisp	High	High
To fall like apples	To faint	High	Low
To do your hair	To worry yourself a lot	Low	High
To not miss the air	To be bold	Low	Low

Experiment 1: Learning of Unfamiliar Idioms

Methods

Participants 26 undergraduate students from the same population as the norming participated for course credit. None had previously participated in the norming.

Materials The 32 items described above.

Procedures The training study and Operation Span test were presented electronically using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA); the ATTA, Author Recognition test, and language history questionnaire were administered on paper.

Participants signed up for three separate sessions about two to three days apart. For each session, presentation of all items was randomized. During the first session (approximately 30-45 minutes), participants viewed an idiom and typed what they thought it meant. After they typed their answer, the correct meaning of the idiom appeared, and participants were asked to indicate whether their answer was correct or not. Then participants were asked to generate a novel sentence that included the idiom. This procedure was repeated for each idiom.

The second session (approximately 20-30 minutes) consisted of three parts. First, each idiom and its definition were presented on the screen for 12 seconds. Then, participants saw each idiom again, and rated their confidence in their memory of the idiom's definition. Then, the idiom and its definition reappeared on the screen for another twelve seconds. The metacognitive confidence rating was included to help ensure that participants were paying attention and to aid learning (Kang, 2007). During the second part, participants verbalized each idiom and its meaning to add a physical and auditory memory trace. Participants concluded the session by generating a novel sentence that included the idiom. This time, the participants saw the idiom and had to type the sentence without a reminder of the definition.

In the third session, participants were tested via a cued-recall test. Participants were instructed to "type the learned definition of the expression to the best of your ability." Emphasis was put on recalling the correct meaning, instead of verbatim recall. On each screen, the participant saw an idiom and typed their answer below. Next, participants completed the Operation Span task, the Abbreviated Torrance Test for Adults, and Author Recognition Test. Finally all participants completed a brief language history questionnaire.

In the Operation Span task, participants viewed variably sized sets of mathematical computations followed by single words. The task was to judge whether the answer to the math problem was correct or incorrect and then remember the subsequent word. At the end of each set, the participants were asked to recall as many words as they could remember from that set. Set size ranged from 3-6 words.

In the Abbreviated Torrance Test, participants had three minutes to complete each of three activities. They had to answer thought questions and create images or pictures with incomplete figures. Participants only worked on one activity at a time.

In the Author Recognition test, participants were given a piece of paper that contained eighty names. They were instructed to "read the names and put a check mark next to the names of the individuals you know to be writers."

Results

Responses from the cued-recall test were coded for accuracy. A verbatim or slightly reworded response of the learned definition was counted as correct. For example, for the idiom *to jump from the rooster to the donkey,* which means "to switch from one subject to another", the response "to switch sides" was marked as incorrect. These criteria are strict enough to measure learning over guessing, yet flexible enough to measure competence over memorization. Table 2 shows means and standard deviations for accuracy by condition. It is important to note that even with this relatively strict coding, there was limited variability in the data. Participants performed with perfect accuracy on

approximately half the items, and approximately half of participants made no errors across all items.

Accuracy was analyzed in R (R Development Core Team, 2015; ver 3.2.3) with the lme4 package. We used linear mixed effect logit models with participants and items as crossed random factors (Baayen, 2008; Bates, Maechler, Bolker, & Walker, 2015). Fixed factors (e.g. ambiguity, transparency) were treated as continuous, and were centered in all analyses. When maximal models failed to converge, the random slopes that captured the least variance were dropped until the model converged (Barr, Levy, Scheepers, & Tily, 2013).

Analyses showed a significant effect of transparency (β= 0.59, SE= 0.28, p= 0.04), such that more transparent idioms were learned better. There was no effect of ambiguity (β= 0.01, SE= 0.01, p= 0.63) and no interaction between ambiguity and transparency (β= 0.01, SE= 0.01, p= 0.32).

Individual difference measures were evaluated according to standardized procedures. For the Operation Span Task, set size was used instead of total words recalled because it is more representative of an individual's maximal working memory potential (Unsworth et al., 2005). The range of scores across subjects for each test was as follows: Operation Span Task Set Size (3-6), Abbreviated Torrance Test for Adults Creativity Index (26-90), and Author Recognition Test hits (4-26). Before inclusion in mixed models, all measures were centered and the ATTA Creativity Index was scaled down by a factor of 10 to make its scale more similar to the others.

Table 2: Cued Recall Accuracy in Experiment 1

Transparency	Ambiguity	Mean	SD
High	High	0.94	0.24
High	Low	0.93	0.26
Low	High	0.85	0.36
Low	Low	0.85	0.36

We found a significant positive interaction between transparency and O-Span (β= 0.42, SE= 0.18, p= 0.02). However, the limited variability in the data made us hesitant to draw any firm conclusions from the current results. Therefore we decided to run a follow-up study with less intensive training.

Experiment 2: Less Training of Idioms

Methods

Participants 17 undergraduate students from the same population participated for course credit. None had participated in the norming or Experiment 1.

Materials The same 32 idioms as Experiment 1.

Procedures Experiment 2 consisted of two sessions: a learning session and, two days later, a testing session. Because Experiment 1 hinted at an interaction between transparency and the O-Span test, this task was also administered in Experiment 2. We did not administer the other individual differences tasks for lack of time.

Experiment 2 only included a subset of Experiment's training procedures. In Session 1, participants viewed an idiom and typed what they thought it meant. The correct meaning of the idiom then appeared on the next screen, and participants indicated whether their answer was correct or not. They did this for all idioms. Next, participants went through the same flashcard activity as Experiment 1, however the duration of each flashcard was reduced from twelve seconds to six seconds. Finally, participants generated a novel sentence that included the idiom. This procedure was also repeated for each idiom.

In Session 2, participants were tested via the same cued-recall test as Experiment 1. After the participant typed the definitions to all idioms, they completed a semantic relatedness test. In this test, participants saw one of the idioms they had learned paired with one of two possible words: one related to the idiom's learned figurative meaning and one related to the idiom's literal meaning. Two versions of the semantic relatedness test were created for counterbalancing purposes; this ensured that across the entire experiment, idioms were paired with both related and unrelated words, but individual participants saw each idiom only once, and had an equal number of related and unrelated trials. Participants indicated whether the idiom and the presented word were or were not related in meaning by pressing keys on the keyboard. Finally, participants completed the O-Span test and language history questionnaire.

Results

All analyses were done in R with the same procedures as Experiment 1.

Cued Recall Accuracy Recall accuracy was coded according to the same criteria as Experiment 1; Table 3 shows means and standard deviations for accuracy by condition. Note that accuracy was lower, and there was more variability in the data than in Experiment 1. Our model included random slopes by subject for transparency and ambiguity, and replicating Experiment 1, it showed a significant positive effect of transparency (β= 0.61, SE= 0.26, p= 0.02), but no effect of ambiguity (β< 0.01, SE= 0.01, p= 0.95) and no interaction (β< 0.01, SE= 0.01, p= 0.79).

Table 3: Cued Recall Accuracy in Experiment 2

Transparency	Ambiguity	Mean	SD
High	High	0.84	0.37
High	Low	0.88	0.33
Low	High	0.68	0.47
Low	Low	0.65	0.48

Semantic Relatedness Accuracy Means and standard deviations for accuracy by condition on the semantic relatedness test are shown in Table 4. A model including random slopes by subject for transparency and ambiguity showed a significant negative effect of transparency (β= -0.52, SE= 0.22, p= 0.02), but no effect of ambiguity (β= -0.01, SE= 0.01, p= 0.25) and no interaction (β< 0.01, SE= 0.01, p= 0.64).

Table 4: Semantic Relatedness Accuracy in Experiment 2

Transparency	Ambiguity	Mean	SD
High	High	0.79	0.41
High	Low	0.79	0.41
Low	High	0.90	0.30
Low	Low	0.92	0.27

Transparency x Test Type Interaction We conducted a repeated-measures ANOVA to investigate the interaction of transparency and test type. It indicated a significant main effect of test mode ($F(1, 16) = 18.54$; $p < 0.05$) and a significant interaction between test mode and transparency ($F(1, 16) = 41.78$; $p < 0.05$). The main effect of transparency approached significance ($F(1, 16) = 3.54$; $p = 0.08$). Pairwise comparisons indicated that recall was more accurate for high transparency idioms than low transparency idioms. However, this was reversed for semantic relatedness judgment accuracy: low transparency idioms were more accurate than high transparency idioms.

Operation Span Test We found no significant effect of O-Span (β= 0.80, SE= 0.60, p= 0.18) and no interaction of O-Span with transparency (β= -0.04, SE= 0.12, p= 0.70).

Discussion

Transparency

The current results suggest that transparency plays an important role in idiom learning. In both Experiments 1 and 2, higher transparency idioms were learned better than lower transparency idioms when learning was indexed by performance on a cued-recall test. One plausible explanation for this is that transparency facilitates a useful memorization strategy: finding connections and links between items.

Transparency may aid learning at both the word and phrase levels. For cases in which compositionality or decomposability contributes to higher transparency, spreading activation from a single word in the phrase could aid in recall. For example, in the experimental item *there is an eel under the rock,* the word *eel* could prompt recall of the adjective *fishy*, thus bringing the learner to the meaning *something fishy is going on.* At the phrasal level, the relationship may rely less on lexical associations, and instead be be more illustrative or story-like. Participants may elaborate on the pairing between the literal and idiomatic meaning through mental imagery or storytelling.

However, our findings suggest that transparency's effect on learning may depend on the test. On the semantic relatedness task in Experiment 2, accuracy was higher for less transparent idioms. This result is surprising, and could potentially be caused by some inadvertently uncontrolled property of the semantic relatedness test itself. But if we take it seriously, a potential mechanism for explaining this data pattern could be borrowed from the Construction-Integration model (Kintsch, 1988), or other accounts according to which comprehenders activate wide networks of inferred relationships when trying to relate less-related concepts (e.g. Mason & Just, 2004).

We speculate that a similar process could come into play during our training sessions, such that participants might generate a larger set of weak possible relations when trying to relate low transparency idioms to their meanings than when trying to relate higher transparency idioms to their meanings. Although activating this wider network of possible relations will not necessarily improve performance on cued-recall tests, it could lead to improved performance on a semantic relatedness test.. This is because having activated a wider network of related concepts during training may make it more likely that the concepts corresponding to the related words at test had already received some activation.

Ambiguity

The current experiments found no effect of ambiguity on recall of idiomatic meaning. It is possible that the way ambiguity was operationalized and the way the current experiments were structured could have contributed to this lack of effect. Unlike in previous word-learning studies in which ambiguity was manipulated by teaching either one or more meanings (e.g. Degani & Tokowicz, 2010), in the current study, ambiguity served as a measure for the potential of the literal meaning of an expression to compete with a newly taught figurative interpretation; this kind of ambiguity has also been referred to as *literal plausibility* (Titone, Columbus, Whitford, Mercier, & Libben, 2015). Critically, although this kind of ambiguity would be expected to have a strong effect during comprehension when multiple meanings compete for activation, it may have a less important role in the context of a training study. Participants in the current study were focused on learning idiomatic meanings for all of the items, and at test they knew that only the idiomatic meanings were relevant. This strong experiment-internal emphasis on the idiomatic meaning could have overwhelmed any potential effects of competition from a literal meaning.

Individual Differences

The current results are inconclusive regarding the influence of individual difference measures on idiom learning. Although there was some indication that working memory might be related to the influence of transparency on idiom learning in Experiment 1, this did not hold in Experiment 2, which had more variability in accuracy and a more even

distribution of O-Span scores. Therefore we are hesitant to draw any strong conclusions. Future research with larger samples will be important for addressing the roles that individual differences play in idiom learning.

Acknowledgments

We would like to thank the University of Pittsburgh Honors College for providing the funding for this project and Dr. Scott Fraundorf for his help with analyses in R.

References

. Word Reference.

(2006). Idioms. The Free Dictionary, Fairlex.

(2006). "Proverbes en francais." from http://parolesfr.kiev.ua/idiomsfr.php.

(2006). Traduction en Contexte. Reverso.

Baayen, R. H. (2008). "Analyzing linguistic data: A practical introduction to statistics using R." Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software 67 (1): 1-48.

Bonin, P., et al. (2013). "Norms and comprehension times for 305 French idiomatic expressions." Behav Res Methods 45(4): 1259-1271.

Degani, T. and N. Tokowicz (2010). "Ambiguous words are harder to learn." Bilingualism: Language and Cognition 13(03): 299-314.

Foss, D. J. (1970). "Some effects of ambiguity upon sentence comprehension." Journal of Verbal Learning and Verbal Behavior 9(6): 699-706.

Gibbs, R. W. (1994). The poetics of mind: Figurative thought, language, and understanding. Cambridge University Press.

Goff, K. (2002). "The abbreviated Torrance test for adults (ATTA)." IL Scholastic Testing Service.

Howarth, P. (1998). "Phraseology and second language proficiency." Applied Linguistics 19(1): 24-44.

Kang, S. H., McDermott, K.B., Roediger III, H.L. (2007). "Test format and corrective feedback modify the effect of testing on long-term retention " European Journal of Cognitive Psychology 19(4-5): 528-558.

Keysar, B., Bly, B. (1995). "Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans?" Journal of Memory and Language 34(1): 89-109.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95, 163–182.

Lamiroy, B., Klein, J.R. (2010). "Les expressions verbales figées de la francophonie: Belgique, France, Québec, et Suisse." OPHRYS.

Mason, R. A., & Just, M. A. (2004). How the brain processes causal inferences in text a theoretical account of generation and integration component processes utilizing both cerebral hemispheres. Psychological Science, 15(1), 1-7.

Nunberg, G., Sag, I.A., Wasow, T. (1994). "Idioms." Language 70(3): 491-538.

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M.G., Davis, M.H. (2012). "Learning new meanings for old words: effects of semantic relatedness." Memory and Cognition 40(7): 1095-1108.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-prime computer software and manual. *Pittsburgh, PA: Psychology Software Tools Inc*.

Stanovich, K. E., West, R.F. (1989). "Exposure to print and orthographic processing." Reading Research Quarterly: 402-433.

Steinel, M. P., Hulstijn, J.H., Steinel, W. (2007). "Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency." Studies in Second Language Acquisition 29(3): 449-484.

Titone, D. A., Connine, C.M. (1994). "Descriptive Norms for 171 Idiomatic Expressions: Familiarity, Compositionality, Predictability, and Literality." Metaphor and Symbolic Activity 9(4): 247-270.

Titone, D., Columbus, G., Whitford, V., Mercier, J., & Libben, M. (2015). Contrasting bilingual and monolingual idiom processing. Bilingual figurative language processing, 171-207.

Turner, M. L., Engle, R.W. (1989). "Is working memory capacity task dependent?" Journal of Memory and Language 28(2): 127-154.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. Behavior research methods, 37(3), 498-505.

Yorio, C. (1989). Idiomaticity as an Indicator of second language proficiency. Bilingualism Across the Lifespan. L. K. O. K. Hyltenstam. Cambridge, Cambridge University Press: 55-72.

Zyzik, E. (2011). "Second language idiom learning: The effects of lexical knowledge and peagogical sequencing." Language Teaching Research 15(4): 413-433.

Explaining December 4, 2015: Cognitive Science Ripped from the Headlines

Samuel G. B. Johnson
(samuel.johnson@yale.edu)
Department of Psychology, Yale University, New Haven, CT 06520 USA

Abstract

Do the discoveries of cognitive science generalize beyond artificial lab experiments? Or do they have little hope of helping us to understand real-world events? Fretting on this question, I bought a copy of the *Wall Street Journal* and found that the three front page headlines each connect to my own research on explanatory reasoning. I report tests of the phenomena of *inferred evidence*, *belief digitization*, and *revealed truth* in real-world contexts derived from the headlines. If my own corner of cognitive science has such explanatory relevance to the real world, then cognitive science as a whole must be in far better shape yet.

Keywords: Explanatory reasoning; ecological validity; everyday thinking; causal reasoning; theory of mind.

The Fretful Voice

Lately, I've been losing sleep.

Cognitive science excites us in part because it helps to explain broad swathes of human experience. Categories guide our stereotypes about social groups and our choices about which toothpaste to buy. Analogies help politicians to learn from history when deciding foreign policy and children to learn from examples when first encountering scientific ideas. Probability judgments determine our willingness to risk our lives and to play the lottery.

Yet, I suspect I am not the only cognitive scientist with a certain existential fear—a fretful inner voice that wonders whether our discoveries really have the generality I confidently boast to my undergrads. When a student asks about the significance of some principle of naïve physics, we can easily point to implications for science education—an important domain to be sure, but one almost custom-tailored to the scientific findings. When justifying the importance of attention research, the applications for traffic safety stand out as critical—but to what extent are various discoveries about attention generalizable across everyday experience, beyond cherry-picked case studies? It is not difficult to find real-world examples explained by cognitive theories, yet one wonders at the degrees of freedom.

I do not believe that scientific research must have direct practical implications, nor do I deny that theory-driven research can reveal genuine scientific truths. But the world is filled with truths: Isn't it our job to find the important ones—the ones that are both deep and general? To use Dennett's (2006) example, there is something undeniably elegant about the game of chess and the results in mathematics and computer science that it has inspired. But what about the (made-up) game of *chmess*, where the king moves two squares instead of one? There are just as many facts to discover about chmess as there are about chess—and they are just as true—yet chmess problems have an air of triviality that chess problems do not suffer. The insecure voice asks: Is my research more like *chess*, or more like *chmess*? Dennett quotes Donald Hebb: "If it isn't worth doing, it isn't worth doing well."

One Friday morning, I listened to the voice. I walked down Whalley Ave. to a supermarket, where I bought a copy of the only available national newspaper—the *Wall Street Journal*. The date was December 4, 2015. I was to give a talk the following week on three lines of research, each on a phenomenon of explanatory reasoning, aiming to use real-life examples from the paper to illustrate each part of the talk—to convince my audience (and myself) that my research resembles chess rather than chmess. The front page featured three principal headlines (one on a shooting, one on a central bank decision, one on military policy). Hence, there were no degrees of freedom in choosing headlines. This paper reports tests of these three phenomena in the context of these real-world events.

Explanatory Logic in Everyday Thinking

Our mental experiences consist largely of understanding observed data in terms of unobserved explanations. We make sense of events in terms of causes, features in terms of categories, behavior in terms of mental states, and retinal data in terms of 3-D organizations of the world.

To what extent do these explanatory inferences qualify as a natural kind? Do they merely share a common informational structure, or does the mind use similar mechanisms for solving these inference problems across very different types of psychological processes? I have argued that the very same mechanisms apply across these processes, via a set of heuristics I refer to as *explanatory logic*. For instance, people use an explanation's simplicity to estimate its probability, in a manner that is similar across causal reasoning (Lombrozo, 2007), categorization (Johnson, Kim, & Keil, 2016), and some visual tasks (Johnson, Jin, & Keil, 2014). Similar empirical cases have been made for several other explanatory strategies (e.g., Johnson, Merchant, & Keil, 2015; Johnson, Rajeev-Kumar, & Keil, 2014, 2015a, 2015b; Murphy & Ross, 1994; Sussman, Khemlani, & Oppenheimer, 2014).

However, if these strategies are really so general across cognition, they should also show up in everyday behavior. Is explanation not a dominant theme in our mental lives? The current studies used newspaper headlines to generate stimuli to demonstrate the wide applicability and ecological validity of three of these explanatory strategies.

General Method

Participants (N = 299) were recruited and compensated via Amazon Mechanical Turk. Participants completed three experiments in a random order, and were randomly assigned to one of three between-subjects conditions for each experiment (see Methods below). Afterwards, participants answered 12 true/false check questions. Participants incorrectly answering 33% or more of these questions were excluded from analysis (N = 9).

"California Shooters Leave Clues, but No Clear Motive"

The banner headline referred to the shooting in San Bernardino—an event that had occurred two days earlier. It was unclear at the time whether the motive was terrorism (as ultimately proved true) or an interpersonal feud. The available "clues" were stockpiles of weapons, which would be equally consistent with either motive. It would be more helpful to know whether a terrorist organization would claim responsibility (likely under the terrorism explanation, but unlikely under the interpersonal explanation); however, at the time of printing, it was too early to know. How do people think about such potentially diagnostic information when it is unavailable?

It turns out that people try to 'fill in' such information, using erroneous strategies to do so—a tendency known as the *inferred evidence heuristic* (Johnson, Rajeev-Kumar, & Keil, 2014, 2015a). People use the base rate of the *evidence* to infer whether the evidence would likely be observed, if available, even if the prior probabilities of the *hypotheses* are known, leading people to make illusory inferences (Khemlani, Sussman, & Oppenheimer, 2011). This is essentially the opposite of base rate neglect (Kahneman & Tversky, 1973)—people use *irrelevant* base rates that should be ignored.

For example, in one experiment with artificial stimuli (Khemlani et al., 2011), participants were told that magic spell *A* led to symptoms such as lumps, spots, and bumps, whereas spell *B* led only to lumps and spots. Given that Daryl has lumps and spots, but that it is unknown whether Daryl has bumps, participants believed that spell *A* was likelier. Subsequent work revealed that this bias occurs because people know that *most* people do not have bumps, and reason erroneously that Daryl must not have bumps either. This strategy explains the bias toward explanations making fewer predictions, over-and-above mechanisms such as biased prior probabilities, beliefs about the non-independence of evidence, and pragmatic inference (Johnson, Rajeev-Kumar, & Keil, 2015a).

In the shooting case, let's suppose that investigators have narrowed down the motive to two possibilities (terrorism or interpersonal feud), which have equal prior probabilities. People may nonetheless try to guess what percentage of shootings have responsibility claimed by a terrorist organization (an irrelevant piece of information once the prior of each hypothesis is known). This number is small (say, 10%), so participants might reason that there is a small chance that responsibility would be claimed in the San Bernardino case. If people then hold that *inferred* negative evidence against the terrorism motive, people would infer that the interpersonal motive is more probable than the terrorism motive—incorrectly, because this inference contradicts the prior probabilities without any new information.

To test this prediction and mechanism, participants were oriented to an anonymized version of the case:

> Imagine that a shooting occurred in the United States. Investigators have narrowed the suspect's motivation down to two possible motivations. Suppose that each motivation accounts for about 2% of shootings in the United States:
>
> The motivation could have been interpersonal problems between the suspect and one of the victims. In such cases, weapons stockpiles are typical.
>
> The motivation could have been terrorist intentions. In such cases, weapons stockpiles are typical, and a terrorist organization usually claims responsibility.
>
> The suspect had stockpiled weapons, but it is too early to tell whether any terrorist organization will claim responsibility.

In the *Neutral* condition, participants were not given a base rate for terrorist organizations claiming responsibility. We would expect these participants to use their tacit base rate, which would be low, and therefore to think the interpersonal motive is more likely. In the *Low Base Rate* condition, participants were explicitly given a low base rate:

> Suppose that for the vast majority of shootings, no terrorist organization claims responsibility.

Conversely, in the *High Base Rate* condition, participants were explicitly given a high base rate:

> Suppose that for the vast majority of shootings, a terrorist organization claims responsibility (regardless of whether or not they are actually responsible).

This parenthetical remark was included only in the latter condition, so that the effect base rate did not contradict the cause base rates given earlier in the problem (in the High condition), but also did not introduce a pragmatic violation (in the Low condition). If the mechanism at work here is inferred evidence, then we would expect participants to favor the interpersonal motive *more strongly* in the Low condition, and *less strongly* (or even favor the terrorism motive) in the High condition.

Method

Participants read either the text of the Neutral, Low Base Rate, or High Base Rate condition (see above). After reading this information, participants were asked "Which explanation do you think is most probable in this case?" Responses could range from -5 ("Very likely interpersonal") to 5 ("Very likely terrorism"). The order of the two explanations was randomized, and the

orientation of the scale was adjusted to match this order.

On a separate page, participants in the Neutral condition were asked to report their tacit base rate: "Of all the shootings in the United States, for what percent do you think a terrorist organization claims responsibility?"

Results and Discussion

Scales were coded so that negative scores show a preference for the explanation making fewer predictions (interpersonal feud) and positive scores favor the explanation making more predictions (terrorism).

In the Neutral condition, participants preferred the interpersonal explanation [M = -0.20, SD = 0.72; $t(95)$ = 2.73, p = .008]. This prediction was predicated on participants having tacit base rates of less than 50% for terrorists claiming responsibility for shootings: Indeed, participants reported a mean 11.1% (SD = 17.1%) base rate. This base rate is normatively irrelevant, because the prior probabilities of the motives were set as equal (2%).

Further, participants strongly preferred the interpersonal motive in the Low Base Rate condition [M = -0.37, SD = 0.77; $t(96)$ = 4.74, p < .001], but not in the High Base Rate condition [M = 0.02, SD = 0.74; $t(86)$ = 0.32, p = .75]—a significant difference [$t(182)$ = 3.64, p < .001].

These results show that inferred evidence mechanisms apply not only to artificial stimuli, but also to realistic stimuli "ripped from the headlines." In addition, insofar as participants were inferring the mental states of the San Bernardino shooters, this finding suggests that people may use explanatory heuristics, such as inferred evidence, in mentalizing. Future research should address this question more fully (but see Johnson & Rips, 2014 for other explanatory heuristics used in mentalizing).

One initially surprising aspect of these results is that participants did not prefer the terrorism explanation in the High Base Rate condition. However, this is consistent with other findings in the literature (Johnson, Rajeev-Kumar, & Keil, 2015a). People's dislike of explanations making unverified predictions is multiply determined, and several other mechanisms make it difficult (though not impossible) to find a preference favoring explanations that do make such predictions.

"ECB Move Crushes Hopeful Markets"

The previous day, there had been a downturn in European markets because the European Central Bank (ECB) had not increased quantitative easing (QE), an inflationary monetary policy, as much as markets had anticipated.

Although seemingly of a very different flavor from the San Bernardino headline, the ECB story also involves an explanatory inference. Investors made inferences about the ECB's intentions based on statements from the ECB chairman. Such explanatory inferences must necessarily be uncertain (interpreting central bank statements relies on many of the same skills as tea-leaf reading). Normatively, then, this uncertainty about the correct interpretation of ECB statements should also propagate to any predictions made on the basis of such inferences.

It turns out, however, that people often *digitize* their beliefs reached through diagnostic reasoning (Johnson, Merchant, & Keil, 2015; Murphy & Ross, 1994). That is, even though people are happy to say that (for example) there is a 60% chance that an object is a skunk or that there is an 80% chance of rain, people do not treat these propositions as having graded truth; instead, they treat them as though they are certainly true or certainly false, when making inferences based on these propositions. Thus, when judging the implications of uncertain evidence (e.g., a very skunk-like and somewhat rabbit-like object), people treat the evidence as pointing to an explanation with certainty (e.g., treating it as though it is certainly a skunk) when thinking about the explanation's implications (e.g., judging whether it is likely to smell).

This tendency could partly explain why markets often react strongly to disconfirmed expectations—if the expectations are formed based on uncertain information treated as certain, the market would be overconfident. For example, suppose the bank's cryptic statement indicates a 70% chance of an aggressive monetary policy and a 30% chance of a modest monetary policy. Suppose further than there is an 80% chance of a major QE expansion, conditional on aggressive intentions, but a 20% chance of major QE expansion, conditional on modest intentions. Then, the probability of a major QE expansion is 0.8*0.7 + 0.2*0.3 = 0.62. But suppose that instead of treating the central bank's intention as uncertain, investors instead treated it as definite—then the probability of a major QE expansion would be 0.8*1 + 0.2*0 = 0.80. Hence, a failure of QE expansion would be more surprising given the 'digital' computation, leading to a bigger adjustment.

Although an experimental study cannot determine what was going through the minds of European investors this past December, the current study tested whether belief digitization occurs in stimuli relevant to such situations.

Method

The method was based on Johnson, Merchant, and Keil (2015, Exp. 2). Participants were assigned to either the *high/low*, the *low/low*, or the *low/high* condition. In the *high/low* condition, the good explanation (aggressive monetary policy) led to an event with high probability and the bad explanation (modest monetary policy) led to an event (introducing a ZT initiative) with low probability:

> Imagine that the central bank of the United States is deciding what policies to adopt.
>
> If they intend to adopt an aggressive monetary policy, they are *likely* to introduce a ZT initiative.
>
> If they intend to adopt a modest monetary policy, they are *unlikely* to introduce a ZT initiative.
>
> Suppose that the central bank chair says that the bank is concerned about the economy and considering a more aggressive monetary policy.

This last statement was intended to lead participants to

think that an aggressive policy was more likely than a modest policy—an aggressive intention would be a better explanation for such a statement than a modest intention.

The *low/low* and *low/high* conditions differed only in the conditional probability of a ZT initiative given each explanation. In the *low/low* condition, the bank was *unlikely* to introduce a ZT initiative under either explanation, and in the *low/high* condition, the bank was *unlikely* to introduce a ZT initiative under an aggressive monetary policy but *likely* to do so under a modest monetary policy. The unfamiliar term "ZT initiative" was used in place of QE in order to make the three conditions equally plausible. The order of listing the good and bad explanations was randomized for each participant.

Participants were then asked a *diagnosis* question and a *prediction* question (in that order, on separate pages).

First, the *diagnosis* question asked "What do you think are the central bank's intentions?" Ratings were made independently for the options "Bank intends to adopt an aggressive monetary policy" and "Bank intends to adopt a modest monetary policy" as percentages. This question was intended to encourage participants to use graded beliefs (working against our hypothesis).

Second, the key dependent measure—the *prediction* question—asked "What do you think is the probability that the bank will introduce a ZT initiative?"

Results and Discussion

First, the results of the diagnosis question indicated that participants thought that an aggressive policy was most probable [M = 73.9%, SD = 18.9%]. However, a modest policy was nonetheless assigned a reasonably high probability [M = 28.4%, SD = 20.3%]. Thus, a failure to account for the low probability explanation could not be due to the explanation having extremely low probability.

As predicted, participants 'digitized', ignoring the low probability explanation when making predictions. There was a large difference between the high/low and low/low conditions [M = 75.4%, SD = 13.8% vs. M = 32.5%, SD = 30.7%; $t(191) = 12.89$, $p < .001$]; that is, participants changed their predictions based on the conditional probability of a ZT initiative, given the high-probability explanation. However, there was no difference at all between the low/high and the low/low conditions [M = 34.3%, SD = 26.8% vs. M = 32.5%, SD = 30.7%; $t(181) = 0.42$, $p < .001$]. Thus, participants did not change their predictions based on the conditional probability of a ZT initiative, given the low-probability explanation. This shows that participants were making inferences as though the high-probability explanation were certainly true.

As in the case of inferred evidence, these results affirm the digitization effect previously found using more artificial stimuli. And also like the inferred evidence case, the context (reading the intention of a central banker) involved mental-state inference. Future work might explore digitization effects more fully in mentalizing.

"U.S. Opening All Military Combat Roles to Women"

The final story concerned a new development in the U.S. military. The military ended a longtime policy of barring women from some combat roles, due to new evidence that women and men were equally capable in these roles.

Once again, this situation involves explanatory inference, and potentially relies on a heuristic studied in previous research (Johnson, Rajeev-Kumar, & Keil, 2015b). Our decisions depend on both the utilities of potential outcomes and our beliefs about those outcomes, which are often reached through inference (Jeffrey, 1965; Johnson, Zhang, & Keil, 2016). Sometimes situations are ambiguous, but it is nonetheless prudent to act as though a "high-stakes" hypothesis were true even if it is uncertain. In such situations, people are subject to a *revealed truth* bias—they not only *act* as though such high-stakes hypotheses are true, but they come to *believe* that they are true. When the evidence is neutral, but one action is more prudent than another, people tend to believe the corresponding hypothesis is likelier to be true. Similarly, evidence favoring a more prudent action is seen as more diagnostic than evidence favoring a less prudent action.

In the current case, the military no doubt believes that it is more problematic to make a Type II error (allowing women to serve when in fact women are less able than men) than a Type I error (forbidding women to serve when in fact women are equally able). In the former "high-stakes" case, there is a potential risk of fatalities, whereas in the latter "low-stakes" case, the risks are more intangible (e.g., discrimination, inefficiency). We would thus expect that if the military waited such a long time to open these roles up to women, it is because they made this trade-off and required overwhelming evidence that they were not making a Type II error in order to allow women to serve in these roles. Would this tendency toward conservative action—acting *as if* the hypothesis were true, that women were less capable in these roles—also make people think that women really *were* less capable?

Method

Participants were assigned to the *Neutral*, the *High-Stakes*, or the *Low-Stakes* condition. All participants were told about a disagreement between two think tanks about the abilities of a particular social group to serve in combat roles. One favors the high-stakes explanation (an error would involve fatalities) and one favors the low-stakes explanation (an error would involve inconvenience):

> One think tank argues that the members of this group are less able to engage in combat and should not be allowed to serve in combat roles. They argue that there will be a serious risk of combat fatalities if they are allowed to serve in combat.
>
> One think tank argues that the members of this group are equally able to engage in combat and should be allowed to serve in combat roles. They argue that there will be a minor inconvenience to this group if

they are not allowed to serve in combat.

In past cases where the think tanks have disagreed on similar issues, the two think tanks have each been proven right about half the time by objective measures.

This last statement was included to equate the prior probabilities of each explanation. The order of listing the explanations was randomized for each participant, and the orientation of each scale was adjusted to match this order.

Next, participants were given evidence concerning this group's combat abilities. In the *Neutral* condition, the evidence was ambiguous between the two explanations:

In this particular case, the evidence is unclear as to which think tank's view is right.

In the *High-Stakes* condition, the evidence favored the high-stakes explanation:

In this particular case, the evidence favors the view that members of this group are less able to engage in combat.

In the *Low-Stakes* condition, the evidence favored the low-stakes explanation:

In this particular case, the evidence favors the view that members of this group are equally able to engage in combat.

Participants then completed measures of *action* and *belief*. For the *action* question, participants were asked "Would you allow members of this social group to engage in combat?" on a scale from -5 ("Definitely no") to 5 ("Definitely yes"). For the *belief* question, participants were asked "Do you think members of this group are equally able to engage in combat or less able to engage in combat?" on a scale from -5 ("Definitely less able") to 5 ("Definitely equally able"). Thus, negative scores correspond to the high-stakes explanation and positive scores correspond to the low-stakes explanation. The order of these two questions was counterbalanced.

Results and Discussion

In the Neutral condition, participants should favor the high-stakes option in choice (i.e., not allowing women to serve), even though the evidence is ambiguous and favors neither hypothesis. Indeed, our predictions about the belief question are predicated on this assumption about the choice question. Unfortunately, this manipulation check failed: Participants were *more* likely to allow members of the group to serve, even when the evidence was ambiguous [M = 0.56, SD = 1.98; $t(97) = 2.80$, p = .006]. In retrospect, it makes sense that many participants would not share the military's priorities, and would view the (certain) social costs of forbidding a social group from participating in the military as potentially more serious than the (potential) risk of combat fatalities. However, the large variance reveals that there were considerable individual differences in their action choices. Thus, although the High Stakes and Low Stakes conditions are uninterpretable because we cannot determine which participants viewed the Type I or Type II error risk as greater, we can analyze the Neutral condition by splitting the sample into those who chose to intervene as though the high-stakes or low-stakes explanation were true.

We first looked at participants whose choices matched our assumptions (N = 32), favoring the high-stakes over the low-stakes explanation in their actions [M = -1.78, SD = 1.23; $t(31) = -8.19$, $p < .001$]. Even though the evidence was ambiguous between the two hypotheses, these participants nonetheless believed the high-stakes explanation was (marginally) more likely to be *true* [M = -0.63, SD = 1.90; $t(31) = -1.96$, p = .059]. Thus, these participants seem to have used their decisions to infer the truth, even though their decisions were the result of prudential, rather than probabilistic, thinking.

The story is similar for those participants whose choices were opposite to our assumptions (N = 55), favoring the low-stakes explanation in their actions [M = 2.23, SD = 1.32; $t(54) = 12.52$, $p < .001$]. These participants also believed the low-stakes explanation more likely to be true [M = 1.39, SD = 1.77; $t(54) = 5.82$, $p < .001$].

These results support the idea that revealed truth is at work in everyday situations such as those covered by the newspaper. A shortcoming of this study was the failure of the manipulation to induce participants to consistently favor one course of action due to prudential concerns—participants appear to differ in which kind of error they deem more problematic. Hence, future research should look at naturalistic cases where the prudential concerns are more clear-cut. (Of course, the original revealed truth effect was found using artificial stimuli where prudential concerns were clear, in order to avoid this problem.)

The individual difference analysis helps to buttress our account, but has two limitations. First, since participants were selected based on their responses to the action question, it is possible that some participants ignored our insistence that the prior probabilities of the hypotheses were equal, and then based their action choices on their own antecedent beliefs. This seems unlikely given the magnitude of the effects (much more extreme for the action question than for the belief question, consistent with previous findings; Johnson, Rajeev-Kumar, & Keil, 2015b), but cannot be ruled out entirely. Second, it was not possible to test the asymmetry in evidence diagnosticity (intended to be tested with the Low-Stakes and High-Stakes conditions). These limitations should be addressed in future work with other naturalistic stimuli.

Despite these limitations, this study is encouraging for the generality of the revealed truth hypothesis. This is so not merely because the results as consistent with that hypothesis as they could be (given the failed manipulation check), but because the sort of *situation* in which the revealed truth phenomenon occurs was highlighted on the front page of the Wall Street Journal—on an arbitrary day. If situations are ecologically frequent where beliefs can be informed by choices, then the laboratory findings are likely to generalize to many real-world situations.

General Discussion

In the trenches, we forget how ubiquitous the principles of cognitive science really are. For me, this project has been an encouraging confirmation of the ecological validity of at least one corner of cognitive science research.

In previous work, I've looked at whether there are a common set of cognitive mechanisms—*explanatory logic*—at play in various diagnostic reasoning tasks, such as causal inference and categorization. To draw theoretically strong inferences requires high internal validity, so that these studies often rely on stimuli that are isolated from participants' background knowledge, such as fake diseases and magical transformations.

The current results show that people also use the same mechanisms to contemplate issues found in front page headlines. Inferences about criminal cases depend on both observed and *inferred evidence*—using irrelevant base rates to fill gaps in knowledge. Predictions in economic contexts involve *belief digitization*—treating uncertain propositions as being a sure thing. And beliefs about the capabilities of social groups may turn on *revealed truth*—choosing based on the riskiness of the options, and using that choice to infer what the truth must have been.

I do not claim that these results tell us how often these patterns of inference arise naturally in day-to-day life. Instead, this exercise demonstrates that (1) diagnostic reasoning problems are common in one naturalistic corpus; and (2) the same reasoning mechanisms found in artificial contexts apply to these types of natural problems. Future research might measure spontaneous explanatory behavior directly, to better estimate the frequency of such fallacious thinking (see Weiner, 1985, for a related effort in the domain of attribution theory).

Empirical studies often involve a trade-off between internal and external validity. Whereas cognitive science approaches (including investigations of explanatory reasoning) typically aim to maximize internal validity at virtually any cost, the current work plots a new point on the trade-off curve, increasing external validity at the expense of some experimental control. Nonetheless, I did draw some lines in the sand: I insisted on an experimental approach, where very similar stimuli could be tested across all conditions. This necessarily meant some artificiality in isolating real-world knowledge from these effects, in order to be sure that causal conclusions could be drawn from the results. Future investigations might swing even further toward external validity, perhaps using a larger variety of items drawn from real corpuses (such as newspaper articles), where the theoretically relevant dimensions (such as effect base rates) naturally vary.

If one research program has this degree of real-world relevance, I am far more hopeful for our science as a whole. This conference features hundreds of talks and posters, each reporting a discovery. This project has reinvigorated my hope that many of these discoveries can contribute toward our understanding of cognition in a broad sense. It's a relief—now I can sleep again.

Acknowledgments

I thank Frank Keil and an audience at the University of Chicago, Booth School of Business for helpful discussion. I dedicate this work to the students who keep us honest.

References

Dennett, D.C. (2006). Higher-order truths about chmess. *Topoi*, *25*, 39–41.

Jeffrey, R.C. (1965). *The logic of decision*. New York, NY: McGraw-Hill.

Johnson, S.G.B., Jin, A., & Keil, F.C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Kim, H.S., & Keil, F.C. (2016). Explanatory biases in social categorization. *Proceedings of the 38th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015). Predictions from uncertain beliefs. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015). Inferred evidence in latent scope explanations. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015a). *Sense-making under ignorance*. Under review.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015b). Belief utility as an explanatory virtue. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., & Rips, L.J. (2014). Predicting behavior from the world: Naïve behaviorism in lay decision theory. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Zhang, M., & Keil, F.C. (2016). Decision-making and biases in causal-explanatory reasoning. *Proceedings of the 38th Conference of the Cognitive Science Society*.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–51.

Khemlani, S.S., Sussman, A.B., & Oppenheimer, D.M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–35.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–57.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–93.

Sussman, A.B., Khemlani, S.S., & Oppenheimer, D.M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin*, *97*, 74–84.

Modeling category learning using a dual-system approach: A simulation of Shepard, Hovland and Jenkins (1961) by COVIS

C. E. R. Edmunds (ceredmunds@gmail.com)
School of Psychology, Plymouth University, UK

Andy J. Wills (andy.wills@plymouth.ac.uk)
School of Psychology, Plymouth University, UK

Abstract

This paper examines the ability of a dual-system, formal model of categorization COVIS (Ashby, Paul & Maddox, 2011) to predict the learning performance of participants on the six category structures described in Shepard, Hovland and Jenkin's (1961) seminal study. COVIS assumes that category learning is mediated by two dissociable neural systems that compete to control responding. The verbal system explicitly tests verbalizable rules, whereas the implicit system gradually associates each stimulus with the appropriate response. Although COVIS is highly influential, there are no published evaluations of the formal model against classic category learning data (COVIS is most typically applied heuristically to the design of new experiments). In the current paper, we begin to address this gap in the literature. Specifically, we demonstrate that COVIS is able to accommodate the ordinal pattern found by Shepard et al., provided that adjustments consistent with the model's theoretical framework are made.

Keywords: category learning; computational modelling; dual-system; implicit; explicit;

The field of category learning is currently inundated with formal models competing to explain categorization behavior (Wills & Pothos, 2012). These theoretical models vary on features such as how stimuli are initially represented, how category membership is determined and how they explain observed phenomena. One approach, the COVIS model of category learning (COmpetition between Verbal and Implicit Systems; Ashby, Alfonso-Reese, Turken, U, & Waldron, 1998; Ashby, Paul, & Maddox, 2011), hypothesizes that category learning is mediated by two competing systems: one explicit, one implicit. The explicit, verbal system learns using working memory to test hypotheses about category membership. In contrast, the procedural, implicit system learns by gradually associating areas of stimulus space with a response. At the beginning of learning, COVIS predicts that responding is controlled by the verbal system. People only switch to the implicit system if the verbal system cannot learn the category structure.

One of the strengths of COVIS is that it is capable of making predictions that affect three strands of the category learning literature: behavioral studies, cognitive neuroscience and cognitive modeling (Lewandowsky, Palmeri, & Waldmann, 2012). This is because the model contains, as well as the broad theoretical conceptualization above, both neurological and computational implementations. However, research with the COVIS model has not taken full advantage of these strengths. Instead, researchers have mainly focused on the behavioral predictions of COVIS. Crucially, these predictions are stated heuristically without any supporting formal modeling (Ashby & Maddox, 2005, 2010). The few papers that have presented formal modeling of COVIS have focused on its ability to account for specific novel behavioral results, such as impaired learning in Parkinson's patients, rather than a range of already-established seminal experiments in the category learning literature (Hélie, Paul, & Ashby, 2012a, 2012b). This neglect of standard results stands in contrast to the approach taken by proponents of other well-known models of category learning (e.g. Love, Medin, & Gureckis, 2004; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994).

We note that at least some of COVIS's predictions do not seem to hold under under closer empirical scrutiny (e.g. Dunn, Newell, & Kalish, 2012; Edmunds, Milton, & Wills, 2015; Newell, Moore, Wills, & Milton, 2013; Stanton & Nosofsky, 2007, 2013). However, a different–arguably more fundamental–question is whether the COVIS model is capable of accommodating results already well-known in the category learning literature. As a first step to beginning to answer this question, we describe an investigation of COVIS's ability to account for the ordinal performance of participants learning the six category structures described in (Shepard, Hovland, & Jenkins, 1961). It is important to establish whether COVIS can accommodate these results for two reasons. First, not only have Shepard et al.'s results been found to be robust, they are also known to be problematic for some category learning models (Nosofsky et al., 1994). Second, Shepard et al.'s results have become a standard dataset against which most other formal models have been compared (e.g. Love et al., 2004; Nosofsky et al., 1994). Therefore, testing COVIS against this data set would enable detailed comparisons between COVIS and other leading models of category learning in respect of these particular data.

COVIS

There are slight variations in the model description between papers. The version described here was reported in Hélie et al. (2012a, 2012b).

The verbal system

The verbal system generates categorization responses to stimuli by selecting explicit, verbalizable rules from the set $\mathbf{R} = \{R_1, R_2, \ldots, R_m\}$, where each R_i represents a different rule. In previous implementations of COVIS, the set $\mathbf{R}$ has exclusively included all relevant one-dimensional rules, although it is also hypothesized to include some conjunction or disjunction rules. Each rule R_i is described mathematically in

terms of a discriminant function $h_V(\underline{x})$, whose form depends on the type of rule being implemented. For a stimulus $\underline{x}$ that varies on r dimensions, denoted $\underline{x} = (x_1, x_2, \ldots, x_r)$, a one-dimensional rule based on dimension i calculates the value of the discriminant function as follows

$$h_V(\underline{x}) = x_i - C_i \tag{1}$$

where C_i is a constant that defines the decision boundary. The value of the discriminant function then determines the response given on that trial as follows

$$\begin{aligned} &\text{If on trial } n \text{ if } h_V(\underline{x}) < \varepsilon_V \text{ respond A} \\ &\quad \text{whereas if } h_V(\underline{x}) > \varepsilon_V \text{ respond B} \end{aligned} \tag{2}$$

where ε_V is a normally distributed random variable with a mean of 0 and standard deviation σ_V.

The rule used on a particular trial depends on the success of the rule used on the previous trial. If the categorization response was correct on the previous trial, the same rule will always be used on the next trial. However, if the response is incorrect the next rule is randomly selected from the set **R**, weighted by each rule's current weight, $Y_i(n)$. The weight of each rule is dependent on the participant's past experience of the rule, the reward history of the rule and the participant's tendency to perseverate with incorrect rules. $Y_i(n)$ is calculated from the salience of the rule on the current trial, denoted by $Z_i(n)$. Initial saliences are pre-defined and in typical applications of COVIS one-dimensional rules are assigned equal, relatively high, saliences.

The salience of rule R_i used on trial n is adjusted after every trial depending on whether it resulted in a correct response or not. If the response was correct on trial n then the salience of the rule is adjusted by

$$Z_i(n+1) = Z_i(n) + \Delta_C \tag{3}$$

where Δ_C is a positive constant that represents the perceived reward associated with the correct answer. However, if rule R_i resulted in an incorrect response on trial n then the salience of that rule decreases by the rule

$$Z_i(n+1) = Z_i(n) - \Delta_E \tag{4}$$

where Δ_E is a positive constant that represents the perceived cost of an error on any trial. All the remaining rules in **R** keep their saliences from the previous trial.

The salience of each rule is then transformed to produce the weight, $Y_k(n)$, according to the following

1. For the rule R_i used on trial n

$$Y_i(n) = Z_i(n) + \gamma \tag{5}$$

   where γ is a constant that represents a participant's tendency to perseverate with a rule in light of receiving disconfirming feedback; as γ increases the participant is less likely to switch rules.

2. For a rule chosen randomly from **R**, R_j, its weight is adjusted by

$$Y_j(n) = Z_j(n) + \mathbf{X} \tag{6}$$

   where **X** is a randomly distributed variable that has a Poisson distribution with mean λ. **X** represents the participant's tendency to select novel rules on each trial; the larger λ is, the more likely they are to switch rules. COVIS assumes that rule selection is mediated by the frontal cortex and that λ is related to dopamine levels in the cortex.

3. For the remaining rules

$$Y_k(n) = Z_k(n) \tag{7}$$

Finally, the rule to be used on the next trial is selected with probability

$$P_{n+1}(R_k) = \frac{Y_k^a(n)}{\sum_{s=1}^{m} Y_s^a(n)} \tag{8}$$

where a determines the decision stochasticity. If $a = 1$ then the rule to be used on the next trial is chosen probabilistically, with rules with higher weights being more likely to be picked. As a becomes larger than one, the rule with the largest weight becomes more and more likely to be chosen on each trial, so rule choice behaves more deterministically. Whereas the closer a is to 0, the smaller the differences between probabilities for each rule and so the decisions are more noisy and much less deterministic. Hélie et al. (2012a, 2012b) interpret a as a gain parameter and state a increases with cortical dopamine levels.

The implicit system

The implicit system of COVIS is based on a procedural learning mechanism that associates a response with each stimulus, whose two or more stimulus dimensions are integrated predecisionally. Broadly, this system consists of a representation of sensory information that leads to a hidden layer representing the striatum, which in turn leads to a decision making process in the prefrontal cortex. The sensory cortex is modeled by COVIS as an ordered array of up to 10,000 units. Each unit responds maximally to one particular stimulus and responds to a lesser extent to stimuli resembling it. The activation of each unit is calculated mathematically by a Gaussian function of the distance between the unit's preferred stimulus and the stimulus currently displayed, $d(K, \text{stimulus})$. So, the activation in sensory cortical unit K on trial n is given by

$$I_K(n) = e^{-\frac{d(K,\text{stimulus})^2}{\alpha}} \tag{9}$$

where α is a positive constant that scales the unit of measurement in stimulus space. The larger α is, the more similar the stimuli are to each other.

The activation of striatal unit $\mathcal{J}$ on trial n is determined by the weighted sum of the activations of all sensory units that

project to it, which is formalized as

$$S_{\mathcal{J}}(n) = \sum_K w_{K,\mathcal{J}}(n) I_K(n) + \varepsilon_I \tag{10}$$

where $w_{K,\mathcal{J}}(n)$ is the strength of the synapse between cortical unit K and striatal cell $\mathcal{J}$ on trial n, $I_K(n)$ is the activation of sensory unit K on trial n and ε_I is normally distributed noise (with mean 0 and variance σ_I^2).

Then, the decision rule is

Respond A on trial n if $S_A(n) > S_B(n)$; otherwise respond B

The synapse strengths, $w_{K,\mathcal{J}}(n)$, are adjusted on each trial via reinforcement learning. The initial value, however, must be predetermined. In Ashby et al. (2011), $w_{K,\mathcal{J}}(0)$ was given by

$$w_{K,\mathcal{J}}(0) = 0.001 + 0.0025 \text{ U} \tag{11}$$

where U is a constant sampled randomly from a uniform $[0,1]$ distribution. This means that all initial synaptic strengths will be between 0.001 and 0.0035 and will be randomly assigned. More recent applications of COVIS to experimental data have not defined this parameter.

Then, $w_{K,\mathcal{J}}(n)$ is adjusted on each trial as follows

$$\begin{aligned} w_{K,\mathcal{J}}(n+1) &= w_{K,\mathcal{J}}(n) \\ &+ \alpha_w I_K(n)[S_{\mathcal{J}}(n) - \theta_{\text{NMDA}}]^+ [D(n) - D_{\text{base}}]^+ [1 - w_{K,\mathcal{J}}(n)] \\ &- \beta_w I_K(n)[S_{\mathcal{J}}(n) - \theta_{\text{NMDA}}]^+ [D_{\text{base}} - D(n)]^+ w_{K,\mathcal{J}}(n) \\ &- \gamma_w I_K(n)[\theta_{\text{NMDA}} - S_{\mathcal{J}}(n)]^+ [S_{\mathcal{J}}(n) - \theta_{\text{AMPA}}]^+ w_{K,\mathcal{J}}(n) \end{aligned} \tag{12}$$

where if $g(n) > 0, [g(n)]^+ = g(n)$, otherwise $[g(n)]^+ = 0$.

As Equation 12 is somewhat complex, further explanation is merited. Broadly, the first line describes the conditions under which synapses would be strengthened whereas lines two and three describe the conditions under which the connections would be weakened.

α_w, β_w, γ_w, θ_{NMDA} and θ_{AMPA} are constants. The first three are learning rates whereas the constants θ_{NMDA} and θ_{AMPA} represent the activation thresholds for post-synaptic NMDA and AMPA glutamate receptors respectively, where $\theta_{\text{NMDA}} > \theta_{\text{AMPA}}$ because NMDA receptors have a higher threshold for activation than AMPA receptors.

Equation 12 requires that we specify the amount of dopamine, $D(n)$, released on every trial in response to feedback. The amount of dopamine released is in turn dependent on the reward prediction error (RPE), which is determined by the following

$$\text{RPE} = \text{Obtained Reward} - \text{Predicted Reward} \tag{13}$$

The reward obtained on each trial is dependent on the feedback given. For example, for applications where all stimuli are rewarded or punished equally, the obtained reward R_n on trial n is defined as $+1$ if correct feedback is given, 0 if no feedback is given or -1 if incorrect feedback is given.

The predicted reward on each trial is calculated using a simplified version of the Rescorla-Wagner model (Rescorla & Wagner, 1972). Assuming that the participant has just responded for the nth time to some stimulus then the reward they should expect to receive is given by COVIS as

$$P_n = P_{n-1} + 0.025(R_{n-1} - P_{n-1}) \tag{14}$$

Then, the dopamine release on each trial, $D(n)$, is calculated from the RPE using the following model

$$D(n) = \begin{cases} D_{\max} & \text{if RPE} > \frac{D_{\max} - D_{\text{base}}}{D_{\text{slope}}} \\ D_{\text{slope}}\text{RPE} + D_{\text{base}} & \text{if } -\frac{D_{\text{base}}}{D_{\text{slope}}} \leq \text{RPE} \leq \frac{D_{\max} - D_{\text{base}}}{D_{\text{slope}}} \\ 0 & \text{if RPE } \leq -\frac{D_{\text{base}}}{D_{\text{slope}}} \end{cases} \tag{15}$$

where $D_{\max}$, D_{s} and D_{b} are constants.

Competition mechanism

COVIS uses a combination of the "confidence" each system has in its response, and the level of "trust" the competition system has for each system, to decide which system will guide responding. The confidence that each system has in its response is related to the degree of activation of each stimulus representation. The confidence in the verbal system equals the absolute value of the discriminant function, i.e. $|h_V(n)|$. If $|h_V(n)|$ is large then the stimulus is a long way from the decision bound and so the verbal system is more confident in its response, whereas if $|h_V(n)|$ is 0 then the stimulus is exactly on the boundary between two categories and the verbal system has no confidence on its categorisation. The confidence in the implicit system, $|h_I(n)|$, is equal to

$$|h_I(n)| = |S_A(n) - S_B(n)| \tag{16}$$

and follows a similar logic. If $|h_I(n)|$ is large then the implicit system favours one response much more than the other. However, if $|h_I(n)|$ is close to zero then both striatal units are activated equally and so the system has little confidence in its decision.

The degree of trust in each system is based on the past successes and failures of the system. The amount of trust in the verbal system is given by

$$\theta_V(n+1) = \theta_V(n) + \Delta_{OC}[1 - \theta_V(n)] \tag{17}$$

if the verbal system suggests a correct response on trial n. However, if the verbal system results in an incorrect response, the amount of trust in the verbal system on the next trial is given by

$$\theta_V(n+1) = \theta_V(n) - \Delta_{OE}\ \theta_V(n) \tag{18}$$

where Δ_{OE} is a parameter. The trust in the implicit system is given by

$$\theta_I(n+1) = 1 - \theta_V(n+1) \tag{19}$$

.

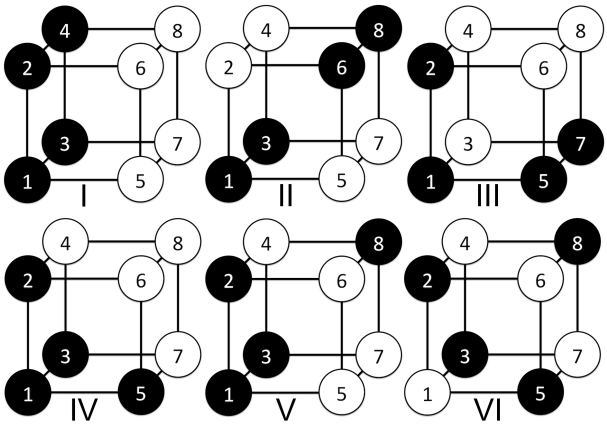


Figure 1: The six category types (I-VI) in Shepard, Hovland & Jenkins (1961). Each number represents a stimulus created from combining the three binary dimensions, which are represented by the sides of the cube. The colour of the circles indicate category membership.

Then, COVIS emits the response suggested by the verbal system if

$$\theta_V(n)\ |h_V(n)| > \theta_I(n)\ |h_I(n)| \tag{20}$$

else it emits the response suggested by the implicit system.

Simulation

The experiment

The category types in Shepard et al. (1961) represent the six qualitatively different ways that eight stimuli, generated from varying three binary dimensions, can be sorted into two, equally sized, categories. These six category structures used are detailed in Figure 1. These category types have been found to systematically vary in difficulty (Shepard et al., 1961; Nosofsky et al., 1994; Smith, Minda, & Washburn, 2004).

Category Type I is a one-dimensional rule and is learned the most easily. Category Type II is an XOR rule and is the next easiest (although see Kurtz, Levering, Stanton, Romero, & Morris, 2013, for a discussion on the relative differences between Types II and IV). Category Types III, IV and V require participants to attend to all three stimulus dimensions and are the next most difficult. Finally, the Type VI category structure has no simplifying rule, is the hardest to learn, and it is generally assumed that the participants learn the stimulus-category assignments separately for each stimulus.

Summary

COVIS is able to predict the pattern of category learning difficulty found in Shepard et al. (1961). However, in order to capture this pattern of data, several adaptations had to be made to the model.

Verbal system adaptations Two changes were made to the verbal system. Although these modifications to the formal model are unique to the current paper, they seem broadly consistent with previously published heuristic descriptions of the COVIS verbal system.

As mentioned above, previous simulations of experimental results using COVIS have limited possible rule selection to only the relevant one-dimensional rules (Ashby et al., 2011; Hélie et al., 2012a, 2012b). However, when using only the 6 available one-dimensional rules (3 dimensions x 2 category mappings) COVIS was unable to capture the ordinal results of Shepard et al. (1961). Specifically, under these conditions COVIS misplaces the difficulty of the Type II problem, placing it as difficult as the Type VI problem.

To rectify this, the model was extended to also include conjunction and disjunction of conjunction rule types, with varying initial saliences. The one-dimensional rules were the most salient, followed by the conjunctions and finally by the disjunction of conjunction rules. We do not include disjunctions as they are behaviorally identical to conjunction models.

The inclusion of rules with different initial saliences implied a knock-on change to part of the rule selection mechanism. It seemed inconsistent with the general operation of the model that rules with different initial saliences would be equally likely to receive the boost produced by Equation 6. Therefore, in this implementation, random selection was weighted by current rule salience.

Implicit system adaptations Two modifications were made to the implicit system. The first was that this simulation included the radial basis function (as stated in Equation 9). This has always been stated as part of the implicit system of COVIS and we state our use explicitly here only because none of the previously simulations have actually used it! (Ashby et al., 2011; Hélie et al., 2012a, 2012b). These previous studies assumed that the stimuli were not confusable and therefore that each sensory unit would respond only to its preferred stimulus. A similar argument could be made here. However, as this simulation aimed to test whether COVIS was capable of learning categories, it seemed important to include the generalizing function as generalization is key to the definition of what it is to have learned a category.

Our second adaptation was to provide separate feedback was given to the implicit system, i.e. feedback as to whether or not the implicit system was correct, not whether the overall response, as determined by the competition system, was correct or not. Providing separate feedback has been included in previous implementations of COVIS, with little explanation as to why (Hélie et al., 2012a, 2012b).

Simulation

1000 simulations were run for each category type using the model and adaptations described above. As in previous reports of the empirical data Nosofsky et al. (1994), blocks 1 and 2 contained 8 trials whilst the remainder contained 16 trials. The stimuli were presented to the verbal system as

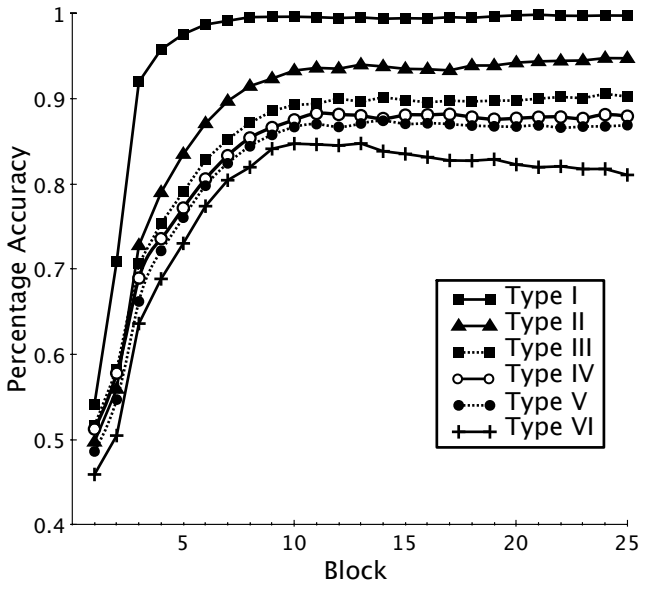


Figure 2: COVIS's predictions of learning the six structures in Shepard, Hovland and Jenkins (1961).

vectors that varied on three binary dimensions as per previous simulations of COVIS (Ashby et al., 2011; Hélie et al., 2012a, 2012b). However, unlike previous simulations and as discussed above, stimuli were presented to the implicit system as vectors that varied on 8 dimensions, corresponding to the eight possible stimuli, with values determined according to Equation 9. This was so that the verbal system received a feature based representation of the stimulus and the implicit system received an object-based representation which included some generalisation across stimuli. The parameters used are displayed in Table 1. No optimisation procedures were used, rather parameters estimates were taken from previous COVIS simulations and adjusted until the predictions shown in Figure 2 were generated.

The results indicate that COVIS can capture the ordinal pattern of difficulty of the six category structures found in Shepard et al. (1961).

Table 1: Parameter values for all learning task simulations.

Parameter	Value	Parameter	Value
Verbal System		**Implicit System**	
σ_V	0	α	0.14
δ_C	0.1	σ_I	0.0125
δ_E	0.2	α_w	0.65
γ	1	β_w	0.1
λ	5	γ_w	0.02
a	10	θ_{NMDA}	0.0022
		θ_{AMPA}	0.01
Competition System		D_{base}	0.2
Δ_{OC}	0.2	D_{slope}	0.8
Δ_{OE}	0.9	D_{max}	1

Discussion

This simulation demonstrates that COVIS is capable of capturing the ordinal pattern of difficulty found by participants in learning the six distinct category structures described by Shepard, Hovland and Jenkins (1961). However, it is interesting to note that although COVIS captured the ordinal pattern, there were still some quantitative discrepancies between this simulation and behavioral replications of the original study. For example, Nosofsky et al. (1994) demonstrated that well before the end of the experiment category types II to V had reached maximal accuracy, whereas COVIS predicts that accuracy does not exceed approximately 85 and 95%. One might argue that this is due to slower learning on the part of COVIS relative to the participants. However, as learning appears to have reached a plateau, this seems unlikely. Another possibility is that there are other parameters values that could allow COVIS to more precisely capture learning of these categories. This seems more likely due to the lack of formal optimization in the current demonstration; an exhaustive search of the parameter space could conceivably discover improved parameter estimates that would enable COVIS to predict the quantitative aspects of Nosofsky et al. (1994) more closely. However, we would also argue that capturing the ordinal aspects of a data set are more important than the precise level of quantitative fit, at least in the early stages of model testing and comparison (Wills & Pothos, 2012).

It should also be emphasised that this and the simulations conducted by Ashby and colleagues (Ashby et al., 2011; Hélie et al., 2012a, 2012b) are first steps in determining whether the formal aspects of COVIS provide the best description of the mechanisms of category learning. Reasons to be cautious come from both simulations and behavioral work conducted to test the COVIS theoretical framework. First, it is possible that the predictions from this simulation and the others mentioned above are parameter dependent. Although COVIS has been shown to be sufficient to capture the ordinal predictions of several experiments, this only determines the model's behaviour at one point in parameter space for each experiment (Pitt, Kim, Navarro, & Myung, 2006). A harder, and more important, question is whether there is a set of parameters that would permit COVIS to capture the ordinal patterns of several experiments simultaneously (as recommended by Wills and Pothos, 2012). One way future work could address this issue would be to employing analysis techniques that look at the model's behaviour across all parameter values, such as parameter space partitioning (Pitt et al., 2006) or landscaping (Navarro, Pitt, & Myung, 2004).

Conclusion

In conclusion, this simulation demonstrates that COVIS is able to capture the order of difficulty of the six category types found by Shepard et al. (1961). Although encouraging, further research is needed to establish whether COVIS can ac-

commodate the broad range of well-established phenomena already present in the category learning literature. Establishing this seems likely to be of similar importance to testing novel predictions of the model.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, U, & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Maddox, W. T. (2005, February). Human Category Learning. *Annual Review of Psychology*, *56*(1), 149–178.

Ashby, F. G., & Maddox, W. T. (2010, December). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011, February). COVIS. In *Formal approaches in categorisation* (pp. 1–13). New York: Cambridge University Press.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures . *The Quarterly Journal of Experimental Psychology*, *68*(2), 1203–1222.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012a, July). A neurocomputational account of cognitive deficits in Parkinson's disease. *Neuropsychologia*, *50*(9), 2290–2302.

Hélie, S., Paul, E. J., & Ashby, F. G. (2012b). Simulating the effects of dopamine imbalance on cognition: From positive affect to Parkinson's disease. *Neural Networks*.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552–572.

Lewandowsky, S., Palmeri, T. J., & Waldmann, M. R. (2012). Introduction to the Special Section on theory and data in categorization: Integrating computational, behavioral, and cognitive neuroscience approaches. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 803–806.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004, August). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*(1), 47–84.

Newell, B. R., Moore, C. P., Wills, A. J., & Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves implicit perceptual categorization: A comment on Filoteo, Lauritzen, and Maddox (2010). *Psychological Science*, *24*(3), 386–389.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994, October). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*(3), 1–20.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57–83.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 64–99.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category Learning in Rhesus Monkeys: A Study of the Shepard, Hovland, and Jenkins (1961) Tasks. *Journal of Experimental Psychology: General*, *133*(3), 398–414.

Stanton, R. D., & Nosofsky, R. M. (2007, January). Feedback interference and dissociations of classification: Evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*(7), 1747–1758.

Stanton, R. D., & Nosofsky, R. M. (2013). Category number impacts rule-based and information-integration category learning: A reassessment of evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1174–1191.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102–125.

Using Prior Data to Inform Model Parameters in the Predictive Performance Equation

Michael G. Collins[1] (michael.collins.74.ctr@us.af.mil)
Kevin A. Gluck[1] (kevin.gluck@us.af.mil)
Mathew Walsh[2] (mmw188@gmail.com)
Michael Krusmark[1] (michael.krusmark.ctr@us.af.mil)
Glenn Gunzelmann[1] (glenn.gunzelmann@us.af.mil)

[1]Air Force Research Laboratory, 2620 Q Street, Building 852
Wright-Patterson Air Force Base, OH 45433
[2]TiER1 Performance Solutions100 E Rivercenter
Blvd #100, Covington, KY 41011

Abstract

The predictive performance equation (PPE) is a mathematical model of learning and retention that attempts to capitalize on the regularities seen in human learning to predict future performance. To generate predictions, PPE's free parameters must be calibrated to a minimum amount of historical performance data, leaving PPE unable to generate valid predictions for initial learning events. We examined the feasibility of using the data from other individuals, who performed the same task in the past, to inform PPE's free parameters for new individuals (prior-informed predictions). This approach could enable earlier and more accurate performance predictions. To assess the predictive validity of this methodology, the accuracy of PPE's individualized and prior-informed predictions before the point in time where PPE can be fully calibrated using an individual's unique performance history. Our results show that the prior data can be used to inform PPE's free parameters, allowing earlier performance predictions to be made.

Keywords: Mathematical model; Performance predictions; Skill learning; Parameter generalization; Educational data mining

Introduction

A common characteristic of training and education programs is that instructors have little or no information about the ability of specific students arriving to a particular class. Without information, instructors must wait until a certain amount of the curriculum has been completed before they can identify who possesses adequate or inadequate knowledge about a given topic, and before they can administer informed training interventions (e.g., removing or adding requirements for additional practice). To make more effective decisions about adaptive education or training interventions, instructors must anticipate the likely effects of specific actions. This would be enabled by models that can predict future performance if a particular training intervention were to be implemented.

In the field of cognitive science, mathematical models of learning and retention have been developed to predict individuals' future performance in a variety of different domains (Anderson & Schunn, 2000; Jastrzembski, Gluck, & Gunzelmann, 2006; Pavlik, & Anderson, 2008). These models can potentially be applied to education and training situations to support more accurate predictions of students' future performance.

One such model is the Predictive Performance Equation (PPE). PPE has been used to predict aggregate group performance and the performance of individuals on declarative (know-what) and procedural (know-how) tasks (Jastrzembski et al., 2006). Prior research has validated PPE with performance data collected in laboratory settings (Jastrzembski et al. 2006), training settings such as Air Force F–16 pilot testbeds (Jastrzembski et al. 2010), and educational settings involving classroom learning and tutoring systems (Collins, Gluck, & Jastrzembski, 2015).

The Predictive Performance Equation

PPE is a model of learning and retention that predicts future performance on the basis of three factors: (1) total amount of practice; (2) elapsed time since practice occurred; and (3) how practice was distributed across time. In general, performance increases with amount of practice (Factor 1), and decreases with elapsed time since practice occurred (Factor 2)(Anderson, 1995). The third factor, distribution of practice over time, is central to research on the spacing effect (for a review, see Cepeda, Vul, Wixted, & Rohrer, 2006). This research has shown that separating practice repetitions by a delay (i.e. *spacing*) slows acquisition but enhances retention. The spacing effect is one of the most widely replicated results in psychology research, and its potential implications for education and training are substantial.

PPE has three free parameters that are calibrated based on historical performance data (Equation 1).

$$\text{Performance} = S * St * N^{c} * T^{-d} \qquad (1)$$

The three free model parameters are S (scalar), used to accommodate the performance measure of interest (e.g.,

error rate, percent correct, response time, etc.), *c* (learning rate), and *d* (decay rate). The model's fixed parameters are determined by the timing and frequency of events in the protocol, such as *T*, the amount of time passed since the onset of training, and *N*, the number of training events that occurred in the training period. *St* (Equation 2) is the stability term that "captures the effects of spacing, by calibrating experience amassed as a function of temporal training distribution and true time passed" (Jastrzembski, Addis, Krusmark, Gluck, & Rodgers, 2010, p. 110).

$$ST = \frac{T_{diff_i}}{PT_i \cdot T_i} + 1 \qquad (2)$$

St is 1 for the first event and is calculated for all other events based on the elapsed time between the current and previous events (T_{diff_i}), the total amassed practice time (PT_i) and the elapsed time between the current event and the first event (T_i).

The individualized approach to using PPE involves gathering data from an individual during a series of calibration sessions, finding the values of PPE's free parameters that maximize the correspondence between the model's output and the individual's observed performance during the calibration sessions, and using the parameterized model to predict the individual's future performance. The basic idea is that although the structure of the model is invariant across individuals, cognitive processes and the psychological parameters that control them, such as rate of learning and rate of forgetting, may vary. Once PPE has been calibrated based on an individual's training history, it can be used to make personalized performance predictions and training prescriptions (Jastrzembski et al., 2010).

Motivation for Analyses

One limitation of PPE is that is has shown to be necessary to calibrate to a minimum of three instances of prior performance before generating a prediction. A minimum of three data points are needed to estimate values for PPE's three free parameters (*S*, *c*, *d*). With fewer than three points, parameter estimates for PPE are unlikely to be accurate because they are under constrained. Multiple different combinations of parameter values may account equally well for the existing data. Consequently, out-of-sample predictions will be highly uncertain at best, and highly inaccurate at worst.

PPE is moderately complex. When such a model is fit to a small number of data points, it is likely to capture the structure of data in addition to noise. This causes poor out-of-sample prediction (i.e., generalization, Geman, Bienenstock, & Doursat, 1992). When PPE is fit using too few data points, its parameter estimates may reflect unrealistic assumptions about psychological processes (e.g., complete learning or forgetting), subsequently causing it to fail to account for the future performance of a sample.

The model's inability to make valid out-of-sample performance predictions after calibrating to fewer than three events reduces its utility early in a training regimen. This creates a lag period during which students start to complete part of the curriculum and personalized performance predictions are not available. One source of information that can help inform predictions of initial performance is data collected from others who performed the same task in the past. Indeed, educational data mining (EDM) attempts to make use of previously collected student data for exactly this purpose. EDM researchers use various data mining techniques (e.g., clustering analysis, rule mining, Bayesian models) to leverage existing student data to predict the future performance of students and to inform educational decisions (Romero & Ventura, 2010). Despite the fact that EDM and PPE have similar goals, these approaches differ in two keys ways. First, EDM seeks to use machine learning techniques to discover regularities in new and often large datasets. The enterprise is mainly data-driven. PPE, in contrast, is based on a psychological model of how various factors impact human learning and retention (Bahrick, Bahrick, Bahrick, & Bahrick 1993; Newell & Rosenbloom 1981). Thus, PPE is theory-driven. The insights and constraints afforded by a psychological theory may enable more effective use of educational data sets (Walsh & Lovett, *in press*). Second, PPE generates precise point predictions of future performance using measurements of accuracy, error rate or response time. Little attention in EDM has been placed on generating such precise performance predictions. Instead, these techniques mainly involve understanding existing data, or generating educational predictions at higher levels of aggregation (e.g., final grades, test scores).

Although these approaches may differ, EDM research highlights the usefulness of prior data when attempting to understand how a particular set of individuals will learn over time, based on how others behaved in the past. Currently, PPE has no way of incorporating any information from prior data into its predictions. However, taking such information into account could be useful when there is not yet enough historical performance data for calibration.

In this paper, we report the results of an evaluation of a new method for applying PPE to decisions about the timing of practice opportunities, using prior data to inform its learning and decay parameters when generating predictions of initial performance. We call these *prior-informed* predictions to distinguish them from *individualized* predictions based on calibration of PPE to an individual's own data, however sparse. Using real-world tutoring data, we compared the accuracy of PPE's prior-informed and individualized predictions of both initial performance (i.e., 2nd and 3rd event) and the first event after initial performance (i.e., 4th event), allowing for a comparison of the two methods when given both an inadequate and adequate amount of prior data for calibration.

Method

All of the data used in this report was obtained from Learnlab.com's DataShop (Koedinger, Baker, Cunningham, Skogsholm, Leber, Stamper, 2010), which is an online data

repository. Datashop contains a collection of publicly available datasets from different math, science, and English classroom and tutoring studies. For this paper, the data consists of performance measures from homework assignments of students from six classes, all from different semesters, using the ANDES tutoring system at the United State Naval Academy (USNA).[1] These datasets were chosen because they contain the largest number of individuals from multiple semesters collected from the same domain currently available on Datashop.

A single semester's worth of data from the USNA on DataShop is referred to as a dataset[2], which is composed of a record of the performance of individuals who attempted to solve a set of problems in a specific domain during a particular period of time. Each dataset contains a record of the performance of each individual student across that curriculum's content. The curriculum is made up of *problems*, defined as "a task [attempted by] a student usually involving several steps." An example of a problem would be comparing the difference in velocity between trains *A* and *B*. Successfully solving a problem involves completing a series of *steps*, which are "an observable part of a solution to a problem", such as finding the velocity of train *A*.

In the analyses presented here, we examined the aggregate performance of students as they attempted to complete an individual step over the course of the semester (i.e., sample). A sample consisted of a selection of two or more students from a single dataset who each had a minimum of four opportunities to attempt a particular step, and had an equivalent sample of students (i.e., two or more) from the remaining 5 datasets who also had a minimum of 4 opportunities to complete the same step. Using these criteria, 307 samples were identified and used for this analysis.

We examined the data at a step level for two reasons. First, steps were the smallest level of resolution of data available on Datashop. Second, each step isolates a particular knowledge component. Because learning occurs at the level of individual knowledge components (Anderson & Schunn, 2000), comparing analogous steps across problems is the proper way to observe the change in performance over time.

Procedure

For the analysis presented here, we systematically controlled which of the 5 datasets were used to inform predictions made about the performance of samples from the 6th datasets, selecting one dataset at a time (prediction dataset) and using data from the remaining 5 datasets as prior data to inform our prior-informed predictions. Next, from the predicted dataset, a single sample of students who completed the same step (prediction sample) were chosen. Then a second sample of individuals who completed the same step from the 5 remaining datasets (prior sample) was selected to inform predictions (prior sample) of the 2nd, 3rd, and 4th event. Finally, the error rate (performance measure) was calculated, as measured by the percent of incorrect attempts on their first opportunity to solve a step during an event for both the prediction and prior sample.

The various timing variables were then calculated for both samples based on the observed number and distribution of practice repetitions (Equation. 2). PPE then calibrated to the prior sample's performance (i.e., percent incorrect) on the first two events, obtaining a set of learning and decay rates (c_{prior} and d_{prior}) that were then generalized to the predicted sample. By calibrating PPE to the performance of the first event of the predicted sample, fixing the learning and decay rate to the parameters generated from the prior sample (c_{prior} and d_{prior}), allowing only the scalar to fluctuate, a prediction of the predicted sample's performance on the 2nd event was obtained.

This procedure was again repeated to generate predictions of performance on the 3rd and 4th event, by increasing the number of calibration events PPE used to generate learning and decay parameters from the prior sample and the event that PPE calibrated its scalar to before predicting the sample's performance on the next event.

Along with each of PPE's prior-informed predictions, predictions of each sample's performance on the 2nd, 3rd and 4th event were generated using the individualized PPE procedure, calibrating up to the event preceding the one being predicted (e.g., if predicting the 3rd event PPE would be calibrated using data from up to the 2nd event) and then generating a prediction for the remaining events.

Results

To evaluate the accuracy of PPE's predictions, we calculated the root mean square deviation (RMSD) between the sample's performance and both PPE's individualized and prior-informed predictions at each event (2nd, 3rd, and 4th). The R^2, a common metric for model assessment was not computed, because we only examined the prediction accuracy of a single event – thus, R^2 could not be computed. After the RMSD was calculated between each prediction and the sample's performance on each event, the average RMSD between the sample's performance and PPE's individualized and prior-informed predictions was computed (Figure 1).

Comparing the average RMSD between each of the model's predictions and the samples' performance on each event, we observe that there was little difference in the accuracy between individualized and prior-informed predictions on the 3rd and 4th event. A difference in the accuracy between the two methods of predictions emerges when predicting the performance of the 2nd event given the

[1] The exact studies which were exported from DataShop and used in this paper are cited in the acknowledgement section, per the guidelines on the Datashop website.

[2] All of the definitions listed in this paper came from the DataShop's online glossary and can be found on https://pslcdatashop.web.cmu.edu/help?page=terms

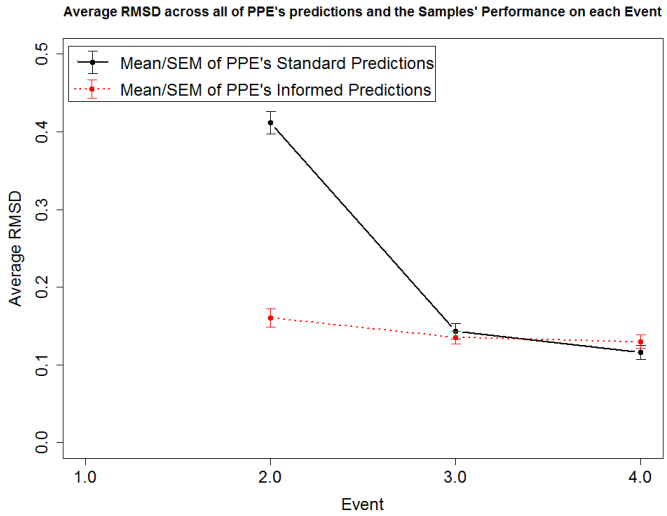


Figure 1. The average RMSD and plus and minus the standard error of the mean of the mean (SEM) between the samples' performance and the prediction of both PPE's individualized and prior-informed predictions.

performance of the first event (Figure 1). The average RMSD of PPE's individualized predictions increases dramatically, because calibrating to only the first event does not offer any information about the learning and decay rate of a sample. In comparison, the average RMSD of PPE's prior-informed prediction of the 2nd event increases only slightly, because it uses learning and decay rates informed by the prior data, allowing it to make an initial assumption based on prior data about the rate at which the sample will learn over time (Figure 2).

Calibrations Given Inadequate Prior Data

The first key question was whether PPE's prior-informed predictions were as accurate as the predictions generated by PPE's individualized approach when it had the opportunity to calibrate to three events before generating a prediction. To evaluate this question, we applied a paired t-test between the RMSD of each of PPE's individualized predictions and the sample's performance on the 4th event and the RMSD of each of PPE's prior-informed prediction and the sample's performance on the 3rd event.[3] We then applied the same test using PPE's individualized predictions and the sample's performance on the 4th event and the RMSD of each of PPE's prior-informed prediction and the sample's performance on the 2nd event. Both t-tests found that the predictions of the sample's performance on the 4th event, generated using PPE's individualized method, came from a different distribution than PPE's prior-informed predictions of both the 3rd ($t(306) = -1.91$, $p < .05$) and 2nd ($t(306) = -3.23$, $p < .001$) event (Figure 1).

These results reveal that PPE's prior-informed predictions of the 2nd and 3rd events were less accurate than when PPE calibrated to the performance of three events before generating a prediction. As predictions are generated earlier and earlier a loss of some prediction accuracy is expected and although PPE's prior-informed predictions were found to be less accurate as PPE's individualized predictions of the 4th event, the mean differences between the average RMSD between PPE's prior-informed and individualized predictions was not found to be extremely large.

Calibrations Given Adequate Prior Data

Finally, another paired t-test was run between the RMSD of the two methods when predicting a sample's performance on the 4th event. No significant difference was observed between the accuracy of these two predictions when predicting the sample's performance on the 4th event ($p > .05$). This result is not surprising in light of the overall average of PPE's free parameters used by these two methods as more data became available (Table 1). As PPE calibrates to additional data, the differences between parameters used by the individualized and prior-informed predictions begins to decrease.

Discussion

Our goal was to develop a method for PPE that could be used to generate predictions of a sample's initial performance when little historical performance data is available to calibrate the model. We accomplished this by calibrating PPE to a sample of individuals who performed

[3] All results were obtained using a Kolmogorov-Smirnov test, because the RMSDs were not normally distributed.

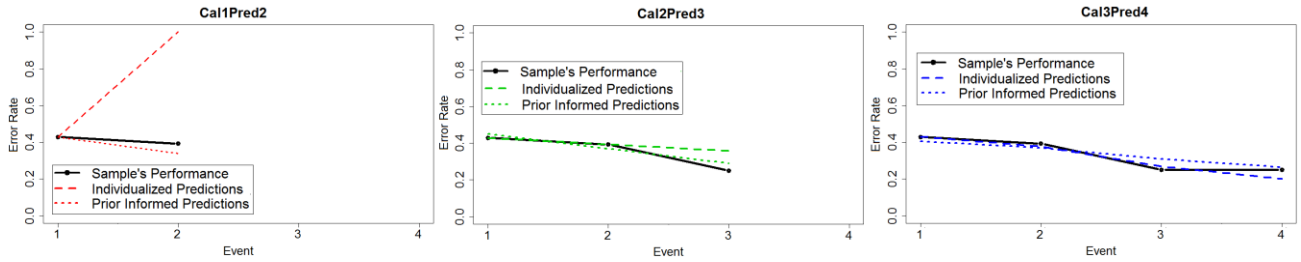


Figure 2: Predicted error rates based on prior-informed and individualized parameter estimates of a single sample from the USNA Dataset. Using the first event to predict the second event (Cal1Pred 2 – red line), two events to predict the third event (Cal2Pred3 – green line) and three events to predict the fourth event (Cal3Pred4 – blue line).

Table 1. The mean and standard deviation of the free parameters used to generate PPE's prior-informed and individualized predictions of each event.

	PPE Free Parameters								
	Learning Rate (*c*)			Decay Rate (*d*)			Scalar (*S*)		
Prediction Method	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)
Prior-informed Predictions	.68(.27)	.58(.29)	.52(.29)	.10(.15)	.11(.18)	.09(.15)	4.15(4.77)	2.92(3.77)	2.65(3.80)
Individualized Predictions	.40(.17)	.51(.32)	.44(.29)	.23(.08)	.06(.17)	.08(.19)	.56(.15)	3.05(4.53)	2.81(4.65)

a particular task and then generalizing the estimated learning and decay rates to a different sample of individuals who performed the same task. This method allows for performance predictions to be made earlier, which may be needed in some training and education scenarios, if waiting until data is available to fully calibrate the model is not practical, such as when educational opportunities are spaced far apart or only a few educational events will be completed.

In our analyses, we first wanted to compare the predictions of initial performance generated using the individualized approach and a novel approach informed by the data of other students. A comparison of the average RMSD of these two methods revealed little difference in the performance predictions of the 3rd and 4th events. The greatest difference was seen in the accuracy of the prediction methods during the 2nd event. Here, the benefit of using learning and decay rates from a prior sample was apparent.

We also examined if PPE's prior-informed predictions could maintain the same level of accuracy as those generated by the individualized approach on the 4th event, after it had calibrated to three previous events. A paired t-test found that there was a significant difference between the accuracy scores generated from PPE's individualized method of the 4th event and PPE's prior-informed predictions of the 3rd and 2nd event.

A decrease in accuracy of predictions of events early in the learning process would be expected, due to the greater uncertainty about the parameter values controlling the process. Still, the accuracy of PPE's prior-informed predictions of the 2nd event are a large improvement in comparison to those generated using PPE's individualized approach.

Comparing the accuracy of the two methods when predicting a sample's performance on the 4th event, after enough data were available for PPE to adequately calibrate its parameters, revealed no difference between the individualized and prior-informed predictions. As PPE calibrated to additional events the difference in the free parameters used by individualized and prior-informed predictions began to diminish leading to similar predictions, and decreasing the need to rely on learning and decay parameters generalized from prior data.

Conclusion

The notion that prior data can improve a model's ability to make predictions is not a new. However, until now, PPE has relied solely on calibrating to historical performance data to determine the values for its free parameters, which has limitations when little historical performance data are available. As shown by the analyses presented here, the benefits of using prior data are substantial. Using prior data to inform predictions of a sample's performance on the 2nd event improved the prediction accuracy by 25% compared to predictions generated through PPE's individualized approach. We find that the learning and decay parameters estimated using prior data add a significant benefit to the predictive ability of PPE, when used under conditions when little historical performance data is available and the individualized calibration approach is unlikely to find a set of parameters that will characterize the future performance of a sample.

Besides using the only the aggregate learning and decay parameters from the prior data other approaches exist for combining existing data with an individual's unique performance history in order to better calibrate parameter estimates for the individual. The hierarchical Bayesian method involves estimating population *hyperparameters* that define the distributions from which an individual's parameters are drawn (MacKay, 2003). For example, PPE's parameters (*S*, *c*, and *d*) may vary across individuals, forming normal distributions at the level of the sample. The mean and variance of the distribution of each parameter are its hyperparameters. This approach balances the tension between maximizing the fit of a model to an individual's data, and maximizing the fit to the group's data. The chief advantage of this approach is that gradually assigns greater weight to an individual's own performance as more data becomes available. This enables a smooth transition from group-based parameter estimates to personalized estimates as the length of an individual's training history increases. More work is needed in order to compare this to the approach explored in this paper.

In conclusion, methods which attempt to generate earlier performance predictions, such as the one discussed here, can provide instructors the ability to better gauge what the expected performance of sample might be compared to their expected performance given a particular training intervention. The ability to make these performance comparisons earlier and with less performance data from students has the potential to boost early learning and improve the overall educational outcome.

Acknowledgements

We would like to thank Carnegie Learning, Inc., for providing the Cognitive Tutoring data supporting this analysis. We used the "USNA Physics Fall 2006", USNA Physics Fall 2007", "USNA Introductory Physics Fall 2009", "USNA Physics Spring 2007", "USNA Physics Spring 2008", "USNA Introductory Physics Spring 2010", accessed via DataShop. The authors would also like to thank the Oak Ridge Institute for Science and Education (ORISE) who supported this research by appointing Michael Collins, to the Student Research Participant Program at the U.S. Air Force Research Laboratory (USAFRL), 711th Human Performance Wing, Cognitive Models and Agents Branch.

References

Anderson, J. R. (1995). *Learning and Memory*. New York: Wiley.

Anderson, J. R., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. *Advances in instructional psychology, Educational design and cognitive science*, 5. 1-33.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993) Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science* , 4, 316-321.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Collins, M. G, Gluck, K. A., & Jastrzembski, T. S. (2015). Datashopping for performance predictions. In: *Proceedings of Foundations of Augmented Cognition*, Los Angeles, CA, 12-23.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias⁄variance dilemma. *Neural Computation*, 4, 1-58.

Jastrzembski, T. S., Gluck , K. A., & Gunzelmann, G. (2006) . Knowledge tracing and prediction of future trainee performance. In*: Proceedings of the interservice/industry training, simulation, and education conference*. Orlando, FL, 1498 – 1508.

Jastrzembski, T. S., Addis, K., Krusmark, M., Gluck, K. A., & Rodgers, S. (2010) Prediction intervals for performance prediction. In: *Proceedings of 10th annual International Conference on Cognitive Modeling*, Philadelphia, PA, 109-114.

Jastrzembski, T. S., Juvina, I., & McKinley, A. (2013) Neurobiological extensions to a mathematical model for performance enhancement observed under conditions of noninvasive brain stimulation. In: *Proceedings of 12th annual International Conference on Cognitive Modeling*, Ottawa, Canada, 131-136.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A data repository for the EDM community: the pslc datashop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, England: Cambridge University Press.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6), 601-618.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition*, 1, 1-55

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2), 101.

Walsh, M. M., & Lovett, M. C. (*in press*). The cognitive science approach to learning. In S. Chipman (Ed.), *Oxford handbook of cognitive science*. Oxford: Oxford University Press.

Using Statistics to Learn Words and Grammatical Categories: How High Frequency Words Assist Language Acquisition

Rebecca L. A. Frost (r.frost1@lancaster.ac.uk)
Department of Psychology, Lancaster University,
Bailrigg, Lancaster, LA1 4YF, UK

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University,
Bailrigg, Lancaster, LA1 4YF, UK

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University,
Ithaca, NY, 14853-7601, USA

Abstract

Recent studies suggest that high-frequency words may benefit speech segmentation (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) and grammatical categorisation (Monaghan, Christiansen, & Chater, 2007). To date, these tasks have been examined separately, but not together. We familiarised adults with continuous speech comprising repetitions of *target words*, and compared learning to a language in which targets appeared alongside high-frequency *marker words*. Marker words reliably preceded targets, and distinguished them into two otherwise unidentifiable categories. Participants completed a 2AFC *segmentation* test, and a similarity judgement *categorisation* test. We tested transfer to a word-picture mapping task, where words from each category were used either consistently or inconsistently to label actions/objects. Participants segmented the speech successfully, but only demonstrated effective categorisation when speech contained high-frequency marker words. The advantage of marker words extended to the early stages of the transfer task. Findings indicate the same high-frequency words may assist speech segmentation and grammatical categorisation.

Keywords: statistical language learning, speech segmentation, grammatical categorisation.

Introduction

Learners must master (at least) two critical tasks prior to reaching linguistic proficiency; identifying individual words from speech, and discovering that these words belong to different grammatical categories. As speech contains no perfectly reliable acoustic cues to word boundaries (Aslin, Woodward, LaMendola, & Bever, 1996) or grammatical category membership (Monaghan, Christiansen, & Chater, 2007), learners must draw upon additional sources of information to accomplish these tasks.

It is well documented that learners can perform powerful computations on the statistical information contained in speech, which can help them to infer word boundaries (Saffran, Aslin, & Newport, 1996). Language learners have been shown to exploit transitional statistics to help identify word boundaries in both artificial (Aslin, Saffran, & Newport, 1998; Saffran et al., 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997) and natural languages (Pelucchi, Hay, & Saffran, 2009), and can do so from infancy onward, before knowing the meaning of a single word in the language (Saffran et al., 1996; Teinonen, Fellman, Naatanen, Alku, & Huotilainen, 2009). Furthermore, similar statistical information can help learners develop rule-like linguistic regularities (Gerken, 2010; Gómez, 2002; Lany & Gómez, 2008; Lany, Gómez, & Gerken, 2007), even while they are learning to segment speech (Frost & Monaghan, 2016).

Bearing in mind learners' aptitude for exploiting statistics, it follows that items appearing more frequently than others in speech might prove helpful for learning. Indeed, recent research has suggested that language acquisition may benefit from the presence of high-frequency words, with a variety of studies demonstrating that these may be advantageous for speech segmentation in particular (Altvater-Mackensen & Mani, 2013; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Mersad, & Nazzi, 2012).

One benefit of frequently occurring words for language acquisition lies in their ability to provide learners with helpful information about the boundaries of words that surround them in speech. In a recent study, Mersad and Nazzi (2012) demonstrated that 8-month-old French-learning infants could identify unfamiliar words from speech when they appeared in a stream containing the familiar word "maman" (mommy in French) but not when they appeared with non-word mãma. Similarly, Bortfeld et al., (2005) demonstrated that infants were better able to identify new words when they appeared in speech next to high frequency, words such as their own name and the word 'mommy', but not when they appeared alongside an unfamiliar item.

Monaghan and Christiansen (2010) examined the possibility that highly frequent words may assist with natural language acquisition through the PUDDLE model of speech segmentation, which they applied to natural language corpora of child-directed speech. Findings indicated that the model was able to quickly extract the high

frequency words in the speech input, which were then used for accurately segmenting the rest of the speech.

We can consider how this might work in practise by taking the sentence *youeatthecheeseyetyoudrinkthewine*. A learner hearing this sentence could recognise high frequency words *you* and *the,* and use these to discern information about the words that surround them in speech. Specifically, in this instance recognising the word *you* would help the learner to identify the way in which the preceding (yet) and succeeding (eat, drink) words end and begin, respectively.

In addition to helping with speech segmentation, highly frequent words may also help inform the formation of grammatical categories (Valian & Coulson, 1988): in the example given above, high frequency word *you* reliably precedes verbs, while *the* reliably precedes nouns, providing valuable information about the word's grammatical role as well as its identity. Interestingly Monaghan and Christiansen (2010) noted that there was substantial overlap between words that were useful for segmentation, and words that were useful for identifying grammatical categories in previous studies of child-directed speech (Monaghan, Christiansen, & Chater, 2007).

In the present study, we examined the way that the same high frequency words affected both of these tasks at the same time. Specifically, we tested whether presence of high frequency words helps learners to identify words that appear alongside them in speech (speech segmentation), and we examined the way that learners used the same high frequency words to learn that words belong to different categories (categorisation). In a subsequent test, we assessed the extent to which learners' pre-existing category knowledge affected their ability to use the language in a word picture mapping task that required them to use the language in a way that was either consistent or inconsistent with their training. We hypothesised that high frequency marker words would assist with speech segmentation, while also constraining the language by contributing to formation of grammatical categories. Further, we varied the ratio of marker words to category words in the study to determine whether variation in the marker words was helpful (Onnis, Monaghan, Christiansen, & Chater, 2004) or an impediment (Valian & Coulson, 1988) to learning.

Method

Participants

The experiment was completed by 72 adults (18 males, 54 females), all of whom were students at Lancaster University, with a mean age of 20.39 years (range = 18-48 years). All participants were native-English-speakers, with no known history of auditory, speech, or language disorder. Participants were paid £3.50, or received course credit.

Design

The experiment used a between subjects design, with three conditions of training type which varied the number of marker words per category from 0 to 2; *Marker0*, *Marker1,* and *Marker2*. Participants were randomly allocated to one of these conditions, with 24 participants receiving each type of training.

Materials

Stimuli

Speech stimuli were created using the Festival speech synthesiser (Black, Taylor, & Caley, 1990). The language contained 20 monosyllabic items (*no, ro, fo, to, li, gi, ni, ka, ma, sa, za, fe, te, re, de, ve, mu, zu, pu, bu*), which were combined pseudo-randomly to create eight bisyllabic *target words* (e.g., *samu, noli, nifo, fede, tero, buza, kato, mave),* and four monosyllabic *marker words (*e.g., *fo, pu, gi, re*), which preceded items in speech. Phonemes used for target and marker words contained both plosive and continuant sounds. There was no repetition of vowel sounds within target words. Each target word lasted approximately 500ms, and each marker word lasted approximately 250ms.

The eight target words were arbitrarily split into two equal categories (*A* and *B,* with four words in each). Category membership was denoted only by the co-occurrence with marker words in speech: in Marker1, one marker word reliably preceded words from each category; in Marker2, two words reliably preceded words from each category (so, in Marker2 markers appeared alongside targets half as often as in Marker1). Marker words preceded category words to reflect use of function words in English, though we acknowledge that information marking grammatical category membership can also occur after word stems. The speech stream for Marker0 contained target words only, so participants in this condition received no information regarding category membership, therefore we did not expect them to demonstrate such knowledge at test.

Eight transitions between words were omitted from the Marker0 condition (e.g., word4 never succeeded word1), and these omitted transitions were later used for the segmentation test stimuli. This was to ensure that all non-words had not occurred during training in any language condition.

To control for possible preferences for certain syllables in certain positions, and preferences for particular dependencies between syllables not due to the statistical structure of the sequences, six versions of the language were generated by randomly assigning syllables to positions within words and marker words (Onnis, Monaghan, Richmond & Chater, 2005). These were counterbalanced across each of the three conditions.

Training

A continuous stream of synthetic speech was created using the Festival speech synthesiser (Black et al., 1990) by concatenating target words and marker words (see Table 1). For Marker0, the speech stream comprised target words only, and lasted approximately 280 seconds. For Marker1, the speech stream comprised target words plus two marker words, and for Marker2, speech comprised target words plus four marker words. Streams for Marker1 and Marker2 both lasted approximately 420 seconds. The marker words for

each of these conditions were used equally in the speech stream in all instances.

Table 1: Example Speech Streams for Each Condition

	Speech
Marker0	*buza-noli-samu-tero-kapu*
Marker1	*ni-buza-zu-noli-zu-samu-ni-tero-zu-kapu*
Marker2	*fo-buza-zu-noli-ma-samu-ni-tero-zu-kapu*

For all conditions, speech was continuous, with no pauses within or between words. Speech streams had a 5 second fade in and out so that the onset and offset of speech could not be used as a cue to the language structure. Target words were each presented 150 times in all streams, with no immediate repetition.

Segmentation test

To test segmentation, we created a two-alternative forced-choice task to examine participants' preferences for words versus non-words (see Saffran et al., 1996 for a similar use of words and non-words). Non-words were bisyllabic items that comprised the last syllable of one target word and the first syllable of another. Non-words did not occur during the training stream, as certain syllable combinations were prevented from occurring: for Marker0, they were formed from the omitted word transitions, and for Marker1 and Marker2, non-words did not occur because a marker word intervened. Selecting a word over a non-word on this task would indicate that participants had successfully identified target words in the speech stream.

Eight test pairs were constructed by matching each target word with a corresponding non-word (e.g. /samu/ vs. mu/no). Items in each test pair were separated by a 1s pause. Test pairs were each presented twice, giving 16 test items in total.

Categorisation test

To test categorisation, we created a similarity-judgement task that contained pairs of target words. Half of all test pairs contained items from the same category (as determined by the marker words that preceded them in speech), with six test pairs containing two A words, and six test pairs containing two B words. Twelve additional mixed test pairs were created (so, one A word and one B word), giving 24 test pairs in total.

Transfer of category knowledge test

To test transfer, we created a word-picture mapping task which provided a grammatical category distinction onto which the distributionally defined category words could map (see Hay, Pelluchi, Graf Estes, & Saffran, 2011, for a similar experimental design testing transfer of segmented speech into a word meaning acquisition task). For this, we introduced eight images, each printed in black on a white background. Four of these images depicted an action, while the remaining four depicted objects. Each target word was paired with one of eight images, and participants were required to learn these pairings. Critically, for half of participants, word-picture pairings were *consistent* with the distributionally-defined categories heard during training, such that all A words appeared with actions and all B words appeared with objects. For the remaining participants, pairings were *inconsistent*: two A words and two B words were paired with objects, and two A words and two B word were paired with actions (see Figure 1). Each version of the language contained a different set of pictures, which were selected at random from an array of 8 object and 8 action images taken from The Peabody Picture Vocabulary Test (Dunn & Dunn, 1997).

In all conditions the order of trials on each test was randomised, and correct responses appeared an equal number of times in each position within pairs/arrays.

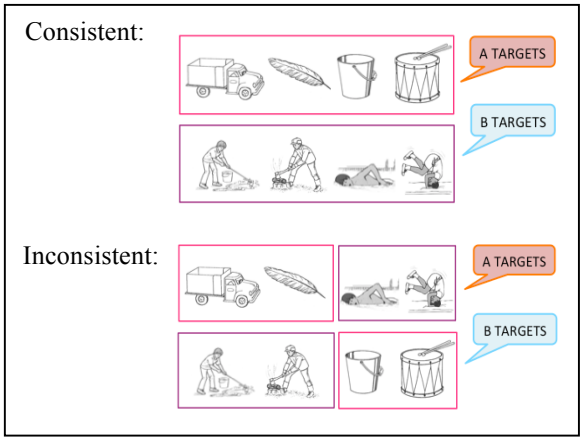


Figure 1. Consistent versus inconsistent word-picture mappings.

Procedure

Before hearing the familiarisation stream, participants were instructed to pay attention to the language and think about possible words it may contain. Participants were tested immediately after training. All participants received the tasks in the same order: participants completed the segmentation test first, followed by the categorisation test, then the transfer of category knowledge test. Tasks were programmed using EPrime 2.0, with instructions appearing onscreen before each task began.

During the segmentation test, participants were instructed to listen to each test pair then select which item best matched the language they had just heard, responding "1" for the first or "2" for the second sequence on a computer keyboard.

For the categorisation test, participants were instructed to listen to each test pair, then rate how similar the role of the items was in the familiarisation stream. Participants were required to respond on a computer keyboard using a 6-point Likert-scale, with 1 being extremely different, and 6 being very similar. If participants have formed categories based on the co-occurrence of target words and markers, then pairs of

items taken from the same category should receive higher similarity ratings than mixed pairs.

For the transfer of category knowledge task, participants heard a target word and saw its corresponding image onscreen for 2 seconds, presented in randomised order. After all eight word-picture pairs were presented, participants heard each target word in isolation and had to select the corresponding image from an array containing all 8 images, responding via keypress, with a number between 1 and 8. This exposure-test sequence was then repeated 3 additional times (so 4 times total). If prior category knowledge was influencing performance on this task, then participants should find it easier to use words from each of the categories consistently (i.e., all A words labelling objects) than inconsistently (i.e. some A words labelling objects, but some A words labelling actions). Thus, an effect involving consistency on this task would indicate transfer.

All speech was presented at a comfortable volume, through closed-cup headphones.

Results

Segmentation

One-sample t-tests were performed on the proportion of correct choices for the segmentation test to compare performance to chance. Performance was significantly above chance for Marker0 (M = .613, SE = .053, t (23) = 2.098, p = .047), for Marker1 (M = .68, SE = .034, t (23) = 5.571 p < .001), and also for Marker2 (M = .66, SE = .023, t (23) = 6.788, p < .001) conditions, indicating that all participants were able to extract individual words from the speech stream (see Figure 2).

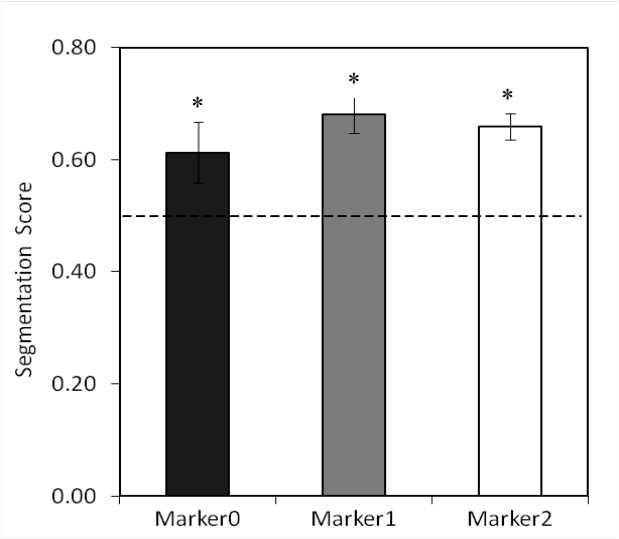


Figure 2: Mean segmentation scores (proportion correct), with standard error. Asterisks indicate above chance performance.

An ANOVA with condition and language version was performed on the segmentation data, with language version included to control for variation across the randomised assignments of phonemes to syllables (this factor was not further analysed in the results). There was no significant effect of condition, F (2, 54) = 1.032, p = .363, with participants in all conditions performing at a statistically similar level.

Categorisation

A repeated-measures ANOVA was performed on data from the categorisation test, with test-pair type (same or different category), condition (Marker0, Marker1, Marker2), and language version as factors. There was no significant effect of test-pair type, F (1, 54) = .968, p = .330, and there was no significant effect of condition, F (2, 54) = 1.327, p = .274. However, there was a significant interaction between test-pair type and condition, F (2, 54) = 3.316, p = .044, η_p^2 = .109), which was led by performance in the Marker1 group, who demonstrated a significant difference between ratings for test pairs containing items from the same (M = 3.882, SE = .179) versus different categories (M = 3.656, SE = .170, t (23) = 2.085, p = .048).

There was no significant difference between ratings for test-pairs containing items of the same (M = 3.451, SE = .118) versus different (M = 3.556, SE = .108) categories for the Marker2 condition (t (23) = 1.969, p = .061). There was no significant difference between ratings for Marker0 (same: M = 3.559, SE = .12; different: M = 3.525, SE = .125, t (23) = .339, p = .737), which was as expected due to the fact that there were no cues to category membership in this condition (see Figure 3).

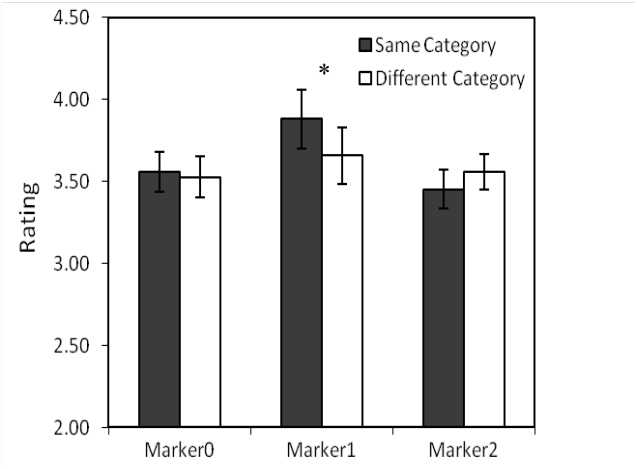


Figure 3: Mean similarity ratings, with standard error. Asterisks indicate significant difference.

Transfer of category knowledge

One-sample t-tests were performed on word-picture mapping data, collapsed across testing time, to compare performance to chance (taken as 12.5%, in accordance with the number of options available per trial). Performance was significantly above chance for Marker0 (M = .73 SE = .035, t (23) = 17.392, p < .001), for Marker1 (M = .69. SE = .039, t (23) = 14.505, p < .001), and for Marker2 (M = .68, SE =

.045, t (23) = 12.470, p < .001), indicating that all participants were able to learn the word-picture mappings.

A repeated-measures ANOVA was performed, with condition (Marker0, Marker1, Marker2), consistency (consistent, inconsistent), test time (test1, test2, test3, test4), and language version as factors. There was a significant effect of test time, F (3, 108) = 86.266, p < .001, η_p^2 = .706, with participants improving steadily across testing phases. There was no significant effect of condition (F < 1), and there was no significant interaction between test time and cue condition (F < 1). There was no significant effect of consistency (F < 1), and there was no significant interaction between consistency and condition, F (2, 36) = 1.890, p = .166.

A univariate ANOVA was performed on the data for the first test to examine whether participants' immediate responses were influenced by their training. There was no significant effect of condition, and there was no significant effect of consistency (both F < 1), but the interaction between these two variables approached significance, F (2, 36) = 2.554, p = .092, η_p^2 = .124 (see Figure 4), indicating that prior knowledge about grammatical categories may have influenced early performance. Trends in the means indicate that this interaction was driven by better performance for participants receiving consistent compared with inconsistent labeling in Marker1, t (22) = 1.661, p = .111, and better performance for participants receiving inconsistent compared with consistent labeling in Marker2, t (22) = -1.460, p = .158, however these differences were not significant.

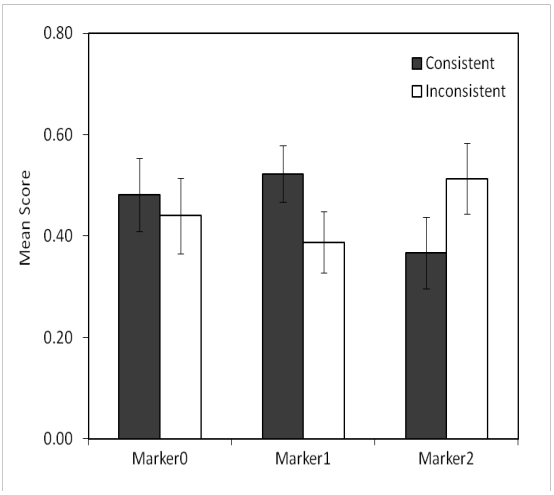


Figure 4: Mean word-picture mapping scores at Test1 (proportion correct), with standard error.

Discussion

This study aimed to investigate the possibility that high frequency words in speech may assist segmentation, while also simultaneously informing grammatical categorisation (Monaghan & Christiansen, 2010; Monaghan et al., 2007).

Data from the segmentation task indicate that all participants were able to identify target words in the speech, regardless of whether that speech contained bisyllabic target words only, or target words plus monosyllabic marker words. This is especially noteworthy given the complexity of speech in the Marker1 and Marker2 conditions. These findings document a rare demonstration of adults' ability to use statistical information to segment continuous speech that contains words of varying length (see Johnson & Tyler, 2010). That participants were able to segment around the high frequency marker words supports the possibility that learners were able to use these as anchors for segmentation (Altvater-Mackensen & Mani, 2013; Mersad, & Nazzi, 2012; Monaghan & Christiansen, 2010), although in this instance they did not significantly boost segmentation compared to a stream containing just targets. Perhaps increasing pre-exposure frequency to these marker words would result in better performance (see, e.g., Bortfeld et al., 2010, for benefit of prior exposure).

Critically, only participants in the Marker1 condition showed evidence of using the marker words to help form grammatical categories, demonstrated by their giving of higher similarity ratings for targets that came from the same grammatical category, and lower similarity ratings for items that came from different grammatical categories (as denoted by the marker words that preceded these targets in speech). These findings indicate the possibility that the same high frequency words that may assist with segmentation could also inform the formation of grammatical categories – providing crucial behavioral support to the claims of prior research (Monaghan & Christiansen, 2010; Monaghan et al., 2007).

While it was not expected that the Marker0 group would differentiate between the A and B categories due to the absence of category cues in this condition, the results of Marker2 demonstrated that too many high frequency words may be an impediment to learning. One potential reason for this is that due to the increased variability in Marker2: marker words and target words appeared alongside each other half as often in Marker2 than Marker1 speech – thus, it is possible that the category distinction may emerge with more training. However, Valian and Coulson's (1988) studies suggest that a high-ratio between marker words and category words is required to result in effective categorisation of an artificial language, consistent with our results. Representing the differentiation in frequency for function words and content words that is present in natural language is not possible in a small artificial language, but the Marker0 condition is the closest approximation to this of all the three conditions.

Results from the word-picture mapping task illustrate that the prior knowledge of grammatical categories for seen Marker1 participants may have influenced performance in the initial stage of the transfer test; demonstrated by greater performance for participants receiving consistent, compared with inconsistent, mappings. However, these effects are subtle and dissipate after further exposure to the word-

picture pairings, which could be due to the rapid learning that occurs in both consistent and inconsistent conditions of the study as exposure proceeds. Thus, the categorisation effect resulting from the same information sources as inform word segmentation may only be apparent at early stages of language development.

Interestingly, the data suggest that inconsistent, rather than consistent, word-picture pairings yielded better learning for the Marker2 condition at the initial test. This raises the possibility that the increased variability in this language may have led participants to prefer to use the language flexibly immediately after exposure. However, as with the transfer effect seen for the Marker1 group, these effects also dissipate across the remaining phases of the test.

To conclude, findings provide evidence to support the suggestion that the same high frequency words might be helpful for informing learners about word boundaries, and grammatical categories in continuous speech. Previous computational models of segmentation and of word learning have (separately) shown that these same high-frequency words prove useful to each task (Monaghan & Christiansen, 2010). The question then arises as to whether these information sources are used at different stages of language development first to solve the problem of identifying words, then later to discover the grammar. The view suggested by our current results is rather that segmentation and grammatical categorisation are not temporally distinct, but rather operate simultaneously (Frost & Monaghan, 2016).

Acknowledgments

This work was supported by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

References

Altvater-Mackensen, N. & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science, 16 (6),* 980-990.

Aslin, R. N., Saffran, J., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9,* 321–324.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. In J. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Black, A. W., Taylor, P., & Caley, R. (1990). The festival speech synthesis system. Edinburgh, UK: Centre for Speech Technology Research (CSTR), University of Edinburgh. http://www.cstr.ed.ac.uk/projects/festival.html

Bortfeld, H., Morgan, J.L., Golinkoff, R.M., & Rathbun, K. (2005). Mommy and me: familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16,* 298-304.

Dunn, L., & Dunn, L. (1997). *Peabody picture vocabulary test III.* Circle Pines, MN: American Guidance Service.

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition, 147*, 70-74.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition, 115,* 362–366.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431–436.

Hay, J.F., Pelucchi, B., Graf Estes, K., Saffran, J.R. (2011). Linking sounds to meanings: infant statistical learning in a natural language. *Cognitive Psychology, 63*, 93-106.

Lany, J., & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science, 19*(12), 1247–1252.

Lany, J., Gómez, R. L., & Gerken, L. (2007). The role of prior experience in language acquisition. Cognitive *Science, 31,* 481–507.

Mersad, K. & Nazzi, T. (2012). When mommy comes to the rescue of statistics: Infants combine top-down and bottom-up cues to segment speech. *Language learning and Development, 8*(3), 303-315.

Monaghan, P. & Christiansen, M. H. (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*, 545-564.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259–305.

Onnis, L., Monaghan, P., Christiansen, M. H. & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. Proceedings of the 26th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in speech processing. Journal of Memory and Language, 53, 225-237.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development, 80*(3), 674–685.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Teinonen, T., Fellman, V., Naatanen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. BMC *Neuroscience, 10*(21).

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71–86.

Modeling Impairments in Lexical Development

Michael Vinos (michael.vinos@gmail.com)
Graduate Program in Basic and Applied Cognitive Science, University of Athens, Greece

Angeliki Andrikopoulou (ang.andrikopoulou@gmail.com)
Department of Philosophy and History of Science, University of Athens, Greece

Christina F. Papaeliou (papailiou@rhodes.aegean.gr)
Department of Preschool Education and Educational Planning, University of the Aegean, Rhodes, Greece

Athanassios Protopapas (aprotopapas@phs.uoa.gr)
Department of Philosophy and History of Science, University of Athens, Greece

Abstract

We implemented the connectionist model of social-pragmatic word learning (Caza & Knott, 2012) to test the hypothesis that reduced joint attention between infant and mother would increase the difference in acquisition between nouns and verbs as observed in Autistic Spectrum Disorder (ASD). The ratio of objects to actions in the observed event stream was manipulated to create an original noun-verb asymmetry. Ten simulations were run for each of the combinations of three conditions of communicative reliability and two conditions of unfiltered random associative learning, which is regarded by some researchers as the primary mechanism of language learning in ASD. The simulations indicated that the reduction in the reliability of communicative actions does not lead to increased noun-verb asymmetry within the originally planned training epochs. A trend in the predicted direction appeared toward the end of training, suggesting that further simulations may help resolve the issue within the current architecture.

Keywords: connectionist model; lexical development; social-pragmatic approach; joint attention; bootstrapping; ASD; SLI; noun-verb asymmetry; associative learning

Introduction

The pronounced rate of vocabulary acquisition during the second year of life constitutes a major milestone in cognitive development (Ganger & Brent, 2004). Laboratory research as well as cross–linguistic studies of natural vocabulary acquisition demonstrate that verbs often lag behind nouns (Gentner, 2006; Bornstein et al., 2004; Childers & Tomasello, 2006). The differentiated learning rate may be attributed to conceptual differences, as verbs are considered less stable and their meaning depends on nouns (Waxman et al., 2013). Alternatively, input distributions may account for the discrepancy, as nouns typically predominate in infant-directed speech (Sandhofer, Smith, & Luo, 2000). It is also proposed that other factors may be involved, such as shape, individuation, concreteness and imageability (Gentner & Boroditsky, 2001).

This pattern of differentiated acquisition is also observed in children with Autism Spectrum Disorder (ASD) and Specific Language Impairment (SLI). However, the categorical asymmetry is greatly exaggerated in ASD (Blach & Chiat, 2003; Marshall, 2003; Ellis Weismer et al., 2011). The situation for SLI is less clear and there are relatively few studies comparing this population with children with ASD or typically developing (TD) children. There is some evidence to suggest that the frequency dependence and poor retention noted in children with SLI may particularly affect verb learning (Rice et al., 1994; Windfuhr et al., 2002). Weismer et al. (2011) found no significant differences between ASD and SLI in the number of action words, while other studies found significant differences between the two groups in the use of verbs describing cognitive and mental states (Ziatas et al., 1998).

In an attempt to provide a unified explanation of word learning, Hollich et al. (2000) proposed the Emergentist Coalition Model. This hybrid theory holds that children rely differentially on multiple cues over developmental time to map words onto referents – whether they are nouns or verbs. At first, infants are sensitive to perceptual cues, mapping a word to the referent that is most interesting or salient. They then use the social intent of a speaker, along with linguistic cues, to home in on word reference.

The ability to understand the intentions of others develops at the end of the first year and seems to constitute a crucial prerequisite for language development (Tomasello & Carpenter, 2007; Trevarthen, 1994). Consistent with this approach, over 50% of the variance in language production and comprehension can be attributed to factors such as the amount of time in joint attention between infant and mother (Carpenter et al., 1998). Deficits in joint attention have been implicated in ASD as predictors of "concurrent language ability" (Dawson et al., 2004). Moreover, recent studies suggest that language problems in SLI may be partially explained by observed deficits in joint attention (Papakalodouka & Papaeliou, 2016). Nevertheless, different processes are involved in acquiring nouns and verbs in joint attention episodes. For example, unlike what is the case with nouns, children benefit from hearing verbs if a verb is heard either before or after an event, and do not show a similar benefit if a new verb and event occur simultaneously (Tomasello & Kruger, 1992; Childers et al., 2007).

Yu and Ballard (2007) modeled social-pragmatic word learning by exploiting social cues such as joint attention and the prosody of the maternal speech. Despite excluding

"theory of mind" skills and prior knowledge of the characteristics of the language, their model outperformed mere statistical associative learning. To model simultaneous learning of individual word meanings and inferring the speaker's communicative intentions—the "chicken and egg" problem of child's early vocabulary acquisition—Frank et al. (2009) posited a computational structure of variables representing the task of word-learning linked together by a set of probabilistic dependencies corresponding to task assumptions. This *intentional model* outperformed several cross-situational word learning models (including the one by Yu & Ballard, 2007) in both aspects of the vocabulary acquisition process.

More recently, a connectionist model of the *social-pragmatic approach to word learning* (Tomasello, 2000) was proposed by Caza and Knott (2012; Knott, 2014). In this comprehensive approach, two subsystems bootstrap each other: One subsystem learns to identify favorable learning opportunities, that is, communicative events that involve reliable mappings between concepts and words. The other subsystem learns to associate words with observed concepts. This model implements the intuition that although plain co-occurrence of word forms and objects may be too noisy to permit successful learning, the reliability of co-occurrences is far from uniform across time. Specifically, in a social-pragmatic framework, co-occurrences are likely to be more reliable when an intent to communicate can be identified. However, in this model events do not come initially flagged as conducive to word learning or not. Rather, the reliability of learning itself is used to tune the network to the appropriate types of events. Therefore, as the model gradually learns that certain events, namely communicative actions of the mother, present favorable learning opportunities, communicative actions and intentions are increasingly recognized and vocabulary

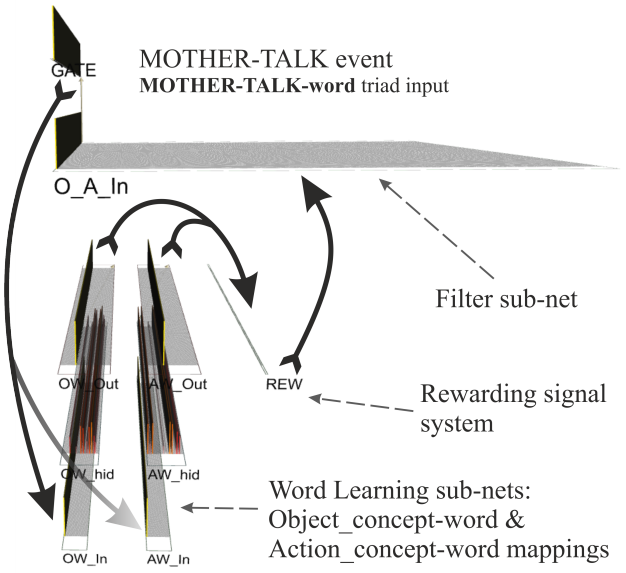


Figure 1: Network architecture of our version of the Caza & Knott (2012) model as implemented in Emergent.

acquisition is accelerated, in a mutually beneficial and reinforcing circle. In this way, development of joint attention enables and underlies efficient word learning.

Method

The separation between object and action words in the model of Caza and Knott (2012) makes it suitable for investigating the differential learning of nouns and verbs. In the present study we were interested in the effects of impaired joint attention on the noun-verb asymmetry. We developed our version of the model on the Emergent platform (Aisa, Mingus, & O'Reilly, 2008) using the Leabra algorithm for training the connection weights, which is more biologically plausible than backpropagation (O'Reilly & Munakata, 2000). The architecture of the model, depicted in Figure 1, consists of (a) a "filter" subnetwork, (b) two word learning subnetworks working in parallel to associate pairs of perceived object and action concepts to concurrent phonological representations of a word, and (c) a reward system, which updates the connections of the filter depending on whether a concept-word mapping has been determined by the model to be successful or not. The model assumes that reliable perception and identification of objects, actions, and phonological word forms is in place prior to the modeled process of lexical learning, that is, of the associations between these types of representations.

In this model, an action, an object, and a word form are perceived at each time frame. The word-learning subnetworks predict word forms based on the observed action and object, generating error signals by comparison to the word form that is actually perceived. At the same time, the action-object pair constitute an observed *event*, which is used by the filter subnetwork to gate a learning signal governing learning in the two word learning subnetworks. In this way, the filter subnetwork controls the extent to which each event leads to modification of the connection weights in the word learning subnetworks. In other words, it modulates the extent to which the current action/object-word pairings are allowed to affect current word knowledge.

Conversely, successful word learning rewards the filter, leading to learning of events that reliably predict correct concept-word pairs. That is, when an action/object-word pairing is predicted correctly by the model, the event preceding it is tagged as informative by a modification of the corresponding weight in the filter subnetwork. In this way, events (i.e., action-object pairings) that are followed by correct word predictions gradually become more influential in word learning. The activation threshold of the gate is set to 0.55 for all the simulations described below. The coordinated actions of these subsystems lead the model to recognize maternal communicative actions as ideal candidates for successful concept-word mapping.

Updates of the weights in the filter subnetwork were gradually disengaged from randomly occurring associative learning. Departing somewhat from the original simulations of Caza and Knott (2012), who let the probability of unfiltered associative learning diminish to zero, we retained

it minimally operative throughout the simulations as a more plausible approach to normal and impaired word acquisition. This is consistent with the literature showing that typically developing children as well as children with ASD do not rely solely on intention monitoring and pragmatic cues for mapping words to objects but also on an associative infrastructure as well (Preissler, 2008; Preissler & Carey, 2005). This particular feature of the model also permitted us to examine the effect of residual plain associative learning. Two settings were used for this minimal *residual associative learning* (RAL), one with probability of 0.001 in each trial (low RAL) and one with probability of 0.005 (high RAL). These values were reached in a designated number of trials from an initial value of 0.5 set at the onset of training. This manipulation goes beyond the demonstrations of Caza and Knott (2012) that bootstrapping is effective in getting the filter subnetwork to efficiently gate relevant training signals to the word learning subnetworks, and was aimed at testing the extent to which unfiltered training, which would likely include a large proportion of incorrect concept-word pairings, might be supportive or disruptive of overall word learning.

In the context of ASD, excessive associative learning is considered by some researchers to constitute the primary foundation of language learning (Luyster & Lord, 2009), not gated by the detection of communicative intent. When combined with an overall reduced rate of communicative event detection, as appropriate for ASD, an abnormally high rate of associative learning—compared to the baseline provided by the simulations of Caza and Knott (2012)—would be expected to track more closely the rate of occurrence of particular word types in the input. Therefore, it might interact with word class frequency in a way that can provide useful clues for understanding lexical development under such abnormal circumstances.

A mixture of training patterns constitutes the input stream. A baseline difference in noun-verb learning was modeled by setting the ratio of reliable object to action concepts to 4:3 in all simulations. Successful joint attention in this model amounts to correct identification of the action-object pair that is attended by the mother. In a communicative context, it is assumed for present purposes that the mother mostly speaks about the event that is currently in her attention. Following Caza and Knott (2012), a communicative action was represented as an event (object-action pair) composed of MOTHER (object) and TALK (action). Joint attention was modeled by the proportion of maternal communicative actions that were reliable, i.e., followed by a valid word-concept pair. Thus, successful joint attention naturally leads to more successful learning opportunities. Simulations tested the hypothesis that a decrease of communicative reliability would exaggerate the noun-verb asymmetry, approaching an ASD profile, rather than simply delay acquisition, resulting in an SLI profile.

There were three levels of joint attention, modeled as the ratio of reliable to nonreliable communicative events: high reliability (500:300), medium reliability (400:400), and low reliability (300:500). The latter condition amounts to inappropriate or unsuccessful detection of communicative intent. The two levels of RAL (low and high) were modeled as the probability of a currently observed action/object-word pairing to affect the word learning subnetworks regardless of the filter subnetwork gate. Ten simulations of 120 epochs were conducted for each combination of joint attention and RAL. Modules were created within Emergent to produce training sequences with the appropriate constraints, and to track the progress of training and network performance. An *endogenous* criterion of learning evaluated word prediction

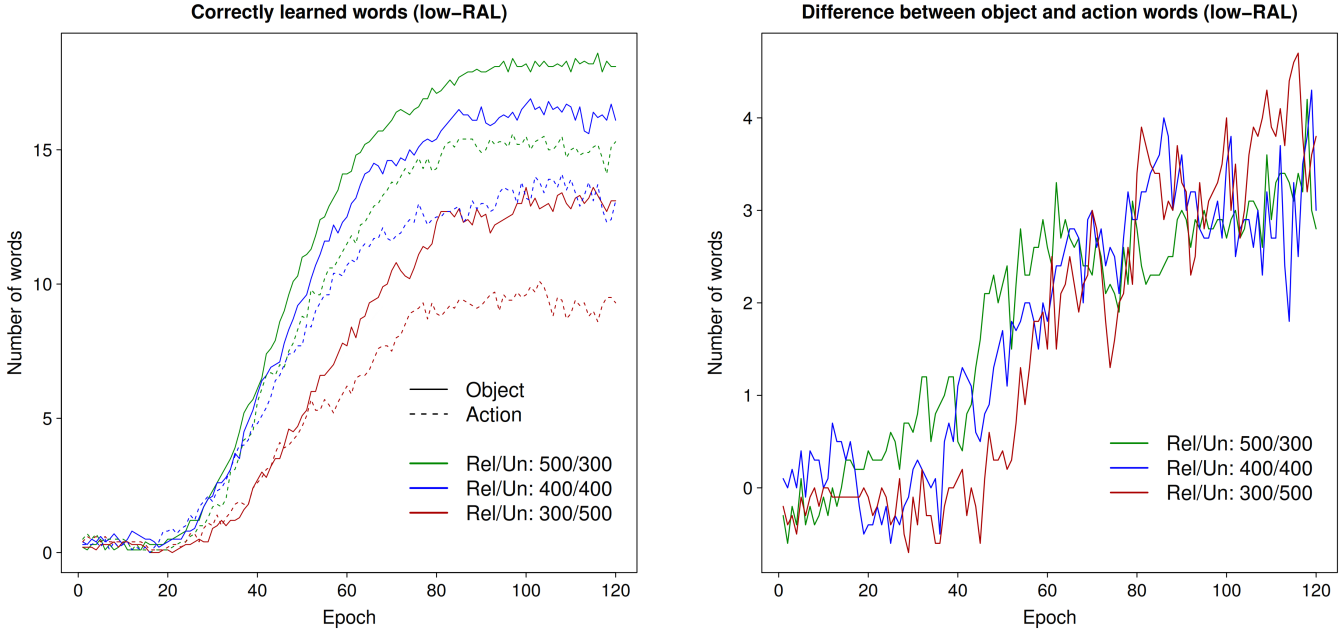


Figure 2: Number of correctly learned nouns and verbs (left) and difference between the number of correctly learned nouns and verbs (right) at the end of each epoch of training in the low-RAL condition.

success in every trial, causing a *reward* or *punishment* signal to the gating component regarding the previous event.

Results and Discussion

Results are displayed in Figures 2 (for low RAL) and 3 (for high RAL). In both RAL conditions the decrease in the reliability of maternal communicative events did not markedly increase the difference between nouns and verbs. Therefore, failure of joint attention did not produce an ASD profile. A tendency toward increased divergence between object and action words at the end of the training was noted but the situation is far from clear within the implemented training epochs. As this pattern of slightly differentiated asymmetry occurred far from ceiling, it invites further study adopting a more protracted training scheme.

Another effect seen mainly in the high-RAL condition (Figure 3) was an increase of the noun-verb asymmetry in the high reliability condition around the middle of training. It unclear whether this is amenable to a developmental interpretation or due to a ceiling effect. Additionally, in both RAL conditions a symmetry in learning action and object words was noted in the initial one-third of training epochs in the low and medium reliability conditions, despite their differential frequency of encounter in the training trials, suggesting perhaps that the associative learning mechanism, mostly activated in the beginning of training, made up for the frequency difference. Note that in the entire set of 60 simulations a low baseline gating threshold was adopted, corresponding to the mimimum gate unit activation required for gating the word learning subnetworks. This led to a large number of events other than MOTHER-TALK activating the word learning system throughout training , essentially constituting a noise-adding mechanism, possibly obscuring the model behavior. This choice (and other parameters) was found necessary for the bootstrapping process to kick in.

Under low-RAL, gating filter efficiency reached 100% at a mean number of 147k trials (*SD* = 23k) in the 300:500 condition, at 113k trials (*SD* = 19k) in the 400:400 condition, and at 114k trials (*SD* = 20k) in the 500:300 condition. Under high-RAL, 100% filter efficiency was reached at 130k (*SD* = 21k), 121k (*SD* = 26k), and 120k (*SD* = 18k) trials in the three conditions, respectively. Despite the high variability, the model seems to struggle more to overcome the low reliability burden (300:500 ratio) in the low-RAL condition compared to high-RAL, suggesting a contribution from the residual associative mechanism.

To further study the full extent of the model's range of behaviors, in a second set of simulations a (less plausible) *exogenous* criterion of learning was implemented, in which the reward signal was hardcoded into the training sequences, to closely track actual communicative reliability rather than correct word prediction. The resulting difference between object and action word learning in the two RAL conditions is displayed in Figure 4. Again, failure of joint attention did not produce an ASD profile, leading instead to a protracted rate of acquisition, with decreased difference between nouns and verbs, consistent with SLI. A possible exception may be noted under high RAL past the initial one-third of training.

Future Work

Following up on these observations, further simulations (currently in progress) can help us shed light on the role of social-pragmatic factors in learning word meanings. In addition to thorough exploration of the dynamic interplay between the ratio of communicative event types, the residual amount of associative learning, and the gating activation

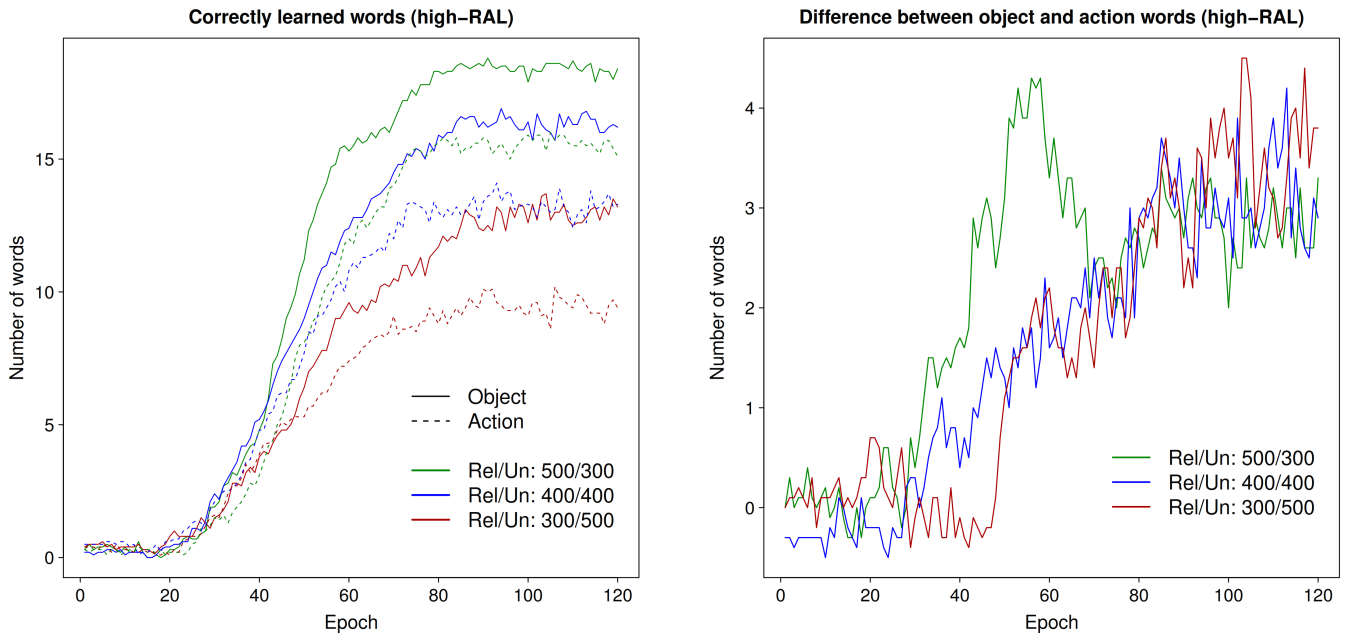


Figure 3: Number of correctly learned nouns and verbs (left) and difference between the number of correctly learned nouns and verbs (right) at the end of each epoch of training in the high-RAL condition.

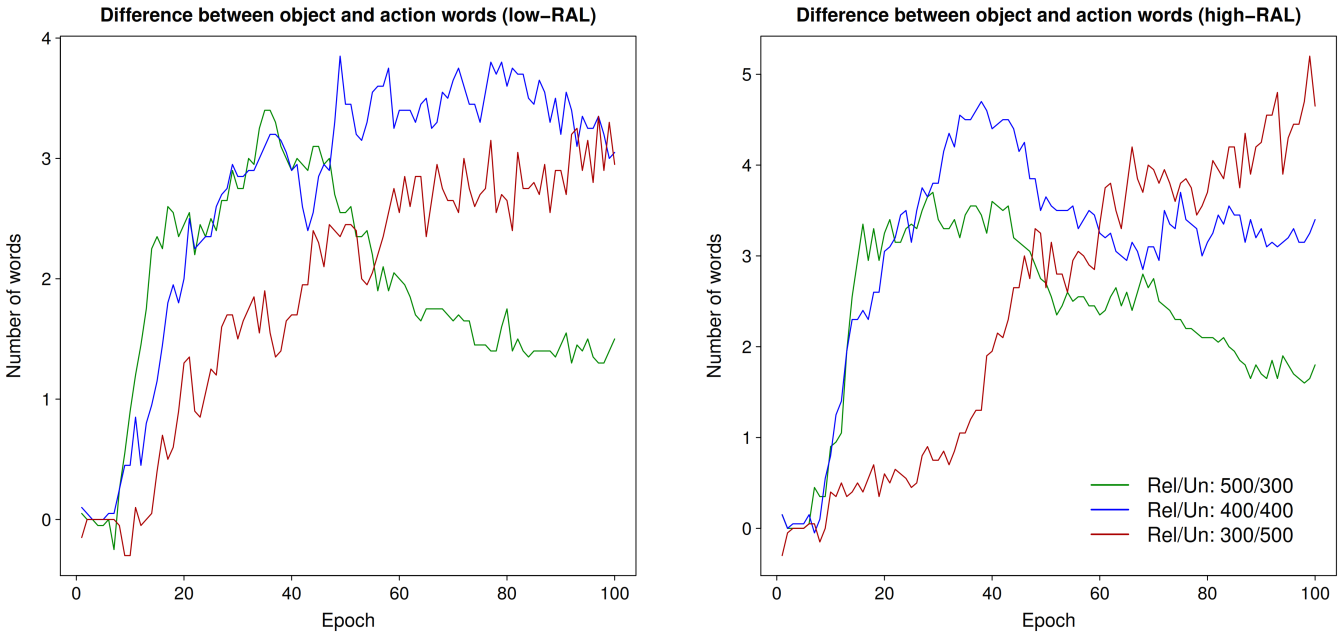


Figure 4: Difference between the number of correctly learned nouns and verbs at the end of each epoch of training in the low-RAL condition (left) and in the high-RAL condition (right) for the *exogenous rewarding* approach.

threshold, future simulations will address the role of specific network parameters on the function of the gating network and the resulting word learning efficiency. A parameter of interest concerns the amount of time for the associative learning mechanism to reach its designated residual value. In the simulations described above this was set to 5000 trials. However, a slower rate of decrease–accompanied by a higher baseline gating threshold—seems promising in the context of ASD (see Figure 5), consistent with the presumed abnormally increased importance of associative learning and reduced communicative reliability in this population.

Moreover, in a future revision of the model, the number of nouns and verbs could be increased substantially, to permit observation of their relative learning rate free from limiting ceiling effects. The small number in the current model may obscure patterns of differential acquisition, as the possibility for further improvement is rapidly curtailed.

In future work we also plan to incorporate additional types of communicative events into the training pattern sequences, as well as carefully studied chains of events abiding to specific constraints, in order to investigate the prospects of certain methods of interventions for impaired language learning, such as parental "synchronous undemanding talking" about the child's focus of attention (Yoder & McDuffie, 2006) and their interaction with the other modeled language acquisition parameters. Further follow-up work can enrich the lexical learning network with conceptual learning components in order to examine the combined effects of category learning and social interaction impairments typical of ASD.

Acknowledgments

This research has been co-financed by the European Union (European Social Fund) and Greek national funds through the Operational Program "Education and Lifelong

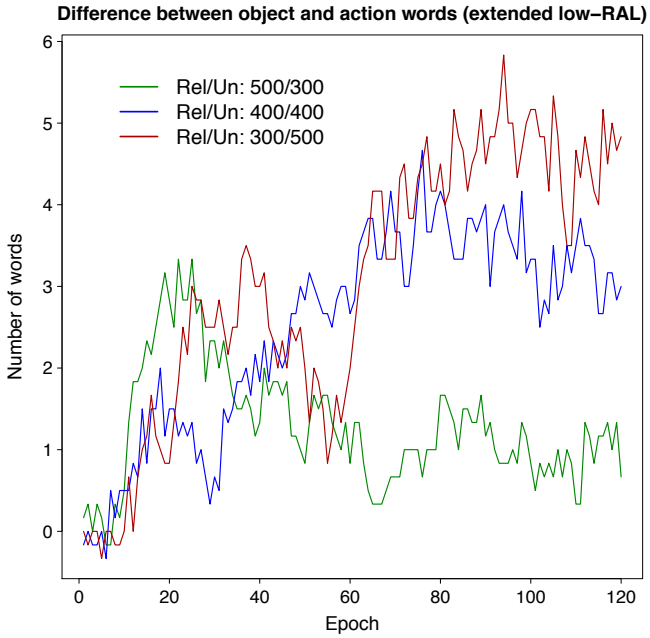


Figure 5: Difference between the number of correctly learned nouns and verbs at the end of each epoch of training in the low-RAL condition with a higher gating threshold and the time for RAL to reach its final value extended to 50000 trials.

Learning" of the National Strategic Reference Framework – Research Funding Programs: THALIS–UOA–COGMEK–375737 and ARISTEIA–BabyAffect–GSRT–3610.

References

Aisa, B., Mingus, B., & O'Reilly, R. (2008). The emergent neural modeling system. *Neural Networks*, *21*, 1146-1152

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., … Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: spanish, dutch, French, hebrew, italian, korean, and american english. *Child Development*, *75*(4), 1115–1139.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.

Caza, G. A., & Knott, A. (2012). Pragmatic bootstrapping: A neural network model of vocabulary acquisition. *Language Learning and Development*, *8*(2), 113–135.

Childers, J. B., & Tomasello, M. (2006). Are Nouns Easier to Learn Than Verbs? Three Experimental Studies. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 311–335). New York, NY, US: Oxford University Press.

Childers, J. B., Vaughan, J., & Burquest, D. A. (2007). Joint attention and word learning in Ngas-speaking toddlers in Nigeria. *Journal of Child Language*, *34*(2), 199–225.

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., … & Liaw, J. (2004). Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental Psychology*, *40*(2), 271.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621.

Gentner, D. (2006). Why Verbs Are Hard to Learn. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). New York, NY, US: Oxford University Press.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge University Press.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., … Rocroi, C. (2000). Breaking the language barrier: an emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65*, 1–123

Knott, A. (2014). How infants learn word meanings and propositional attitudes: a neural network model. In *Communicative Action* (pp. 107–124). Springer.

Luyster, R., & Lord, C. (2009). Word learning in children with autism spectrum disorders. *Developmental Psychology*, *45*(6), 1774.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press.

Papakalodouka, P., & Papaeliou, C. F. (2016). The Theory of Intersubjectivity may explain aspects of Specific Language Impairment. *Eleftherna*, *7*. University of Crete.

Preissler, M. A. (2008). Associative learning of pictures and words by low-functioning children with autism. *Autism: The International Journal of Research and Practice*, *12*(3), 231–248.

Preissler, M. A., & Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. *Cognition*, *97*(1), B13–B23.

Rice, M. L., Oetting, J. B., Marquis, J., Bode, J., & Pae, S. (1994). Frequency of input effects on word comprehension of children with specific language impairment. *Journal of Speech and Hearing Research*, *37*(1), 106–122.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: differential frequencies, different kinds of learning? *Journal of Child Language*, *27*(03), 561–585.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*(4), 401–414.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125.

Tomasello, M., & Kruger, A. C. (1992). Joint attention on actions: acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language*, *19*(2), 311–333.

Trevarthen, C. (1994). Infant semiosis. In W. Nöth (Ed.), *Origins of Semiosis: Sign Evolution in Nature and Culture*. Berlin: Walter de Gruyter.

Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., & Song, H. (2013). Are nouns learned before verbs? Infants provide insight into a long-standing debate. *Child Development Perspectives*, *7*(3), 155–159.

Weismer, S. E., Gernsbacher, M. A., Stronach, S., Karasinski, C., Eernisse, E. R., Venker, C. E., & Sindberg, H. (2011). Lexical and grammatical skills in toddlers on the autism spectrum compared to late talking toddlers. *Journal of Autism and Developmental Disorders*, *41*(8), 1065–1075.

Windfuhr, K., Faragher, B., & Conti-Ramsden, G. (2002). Lexical learning skills in young children with specific language impairment (SLI). *International Journal of Language & Communication Disorders*, *37*(4), 415–432.

Yoder, P., & McDuffie, A. (2006). Teaching young children with autism to talk. In *Seminars in speech and language* (Vol. 27, p. 161). Thieme Medical Publishers, Inc.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.

Ziatas, K., Durkin, K., & Pratt, C. (1998). Belief term development in children with autism, Asperger syndrome, specific language impairment, and normal development: Links to theory of mind development. *Journal of Child Psychology and Psychiatry*, *39*(05), 755–763.

Monolinguals' and Bilinguals' Use of Language in Forming Novel Object Categories

Sarah Fairchild (SarahF@Udel.Edu)
Anna Papafragou (APapafragou@Psych.Udel.Edu)
Department of Psychological & Brain Sciences, University of Delaware
105 The Green, Newark, DE 19716 USA

Abstract

Monolinguals and bilinguals differ along a number of dimensions, including way they label existing object categories (Pavlenko & Malt, 2011). In the present study, we ask whether English monolinguals, Spanish-English bilinguals, and English-Spanish bilinguals also differ in the way they use language when forming novel categories. Previous research with monolinguals shows that a shared label encourages children (e.g., Waxman & Markow, 1995) and adults (e.g., Lupyan, Rakison, & McClelland, 2007) to place objects together. Our results further demonstrate that when two objects shared a Licit Word label like "zeg," monolinguals and bilinguals alike are encouraged to group them together. Illicit Words like "gxz," on the other hand, only influence the categorization decisions of bilinguals. Thus, bilinguals appear to be more flexible in their use of linguistic information in categorization. Neither group made use of non-linguistic cues (patterned frames), suggesting a unique role for language in category formation.

Keywords: verbal labels; categorization; bilingualism

Introduction

The uniquely human faculty of language enables us to convey incredibly complex information to one another. However, language has more to offer than its primary function as a communication tool. Language can be used as a cognitive tool to aid in resource-intensive operations, such as spatial (Loewenstein & Gentner, 2005) and numerical reasoning (Dehaene et al., 1999). Some argue that language is a general computational tool, allowing us to simplify and more efficiently work with information in our environment through a recoding process (Clark, 1998). Here, we investigate how language specifically aids the formation of novel categories by highlighting deep conceptual relationships among objects.

People are more likely to believe that two objects belong together when they share a label (Gelman & Markman, 1986; Jones, Smith, & Landau, 1991; Waxman & Markow, 1995; Balaban & Waxman, 1997; Waxman, 1999; Nazzi & Gopnik, 2001; Sloutsky, Lo, & Fisher, 2001; Booth & Waxman, 2002b; Fulkerson & Waxman, 2007; Lupyan et al., 2007; Diesendruck & Peretz, 2013; Johanson & Papafragou, in press). For example, adults can learn a labeled category of 'aliens' faster than an unlabeled category (Lupyan et al., 2007). Likewise, labels guide children and adults in forming novel plant and animal categories when visual cues are ambiguous (Johanson & Papafragou, 2016). In this sense, language acts as a tool to form categories of novel entities (Wolff, 2011). Importantly, previous research suggests that non-linguistic cues are not as powerful as linguistic labels in influencing categorization. Boutonnet and Lupyan (2015) found that participants were faster to respond to label-object pairs (the word 'dog' – picture of a dog) than sound-object pairs (a dog bark – picture of a dog; cf. also. Lupyan and Thompson-Schill, 2012). Furthermore, linguistic labels are more effective than numbers or symbols in facilitating novel category formation (Gervits et al., submitted). Even for young infants, tones (Ferry, Hespos, & Waxman, 2010), content-filtered speech (Balaban & Waxman, 1997), and primate vocalizations (Ferry, Hespos, & Waxman, 2013) are less effective than labels in facilitating categorization. While it is clear that language can play an important role in category formation, to our knowledge all research to date has been conducted with monolingual participants. However, there are reasons to believe that bilinguals may differ from their monolingual counterparts in their use of labels in categorization. First, monolinguals and bilinguals naturally differ in terms of language abilities and experience. The regular use of two languages changes the way a person's first language is processed (e.g., Van Hell & Dijkstra, 2002) and may even lead to cognitive differences (e.g., Prior & MacWhinney, 2010). Second, we know that bilinguals describe existing categories differently from monolinguals in both their first and their second languages (Pavlenko & Malt, 2010). Additionally, bilingual children are more flexible than monolingual children in the way they apply labels to novel objects (they are less likely to apply the mutual exclusivity constraint when acquiring new words; Davidson, Jergovic, Imami, & Theodos, 1997), and in adulthood, bilinguals are better at acquiring novel labels (Kaushanskaya & Marian, 2009). It may be the case, then, that due to their specific linguistic experiences and advantages, language is a more available and/or flexible tool for bilinguals. For purposes of categorization, bilinguals may weigh linguistic labels more heavily than monolinguals; alternatively, or additionally, bilinguals may be more flexible in regards to what can be considered a label. In either case, bilinguals would not simply be "two monolinguals in one" (Grosjean, 1989) in terms of their use of language as a tool for categorization.

The Present Study

In the present study, we aim to fill a gap in the existing literature by investigating how monolingual and bilingual adults use labels to form new categories. Our goal is to determine to what extent well-formed labels are unique and

powerful category markers for monolinguals and bilinguals alike, as compared to other salient linguistic and non-linguistic cues. Unlike past studies, we take steps to ensure that all types of cues are equally discriminable and thus that any advantage of labels cannot be attributed to superficial features of the cues themselves.

Specifically, our experiment tests the effectiveness of the following cues on the categorization of novel objects: (a) novel Licit Words like "zeg" and "wob," (b) Illicit Words such as "xvxs" that possess linguistic features but are not phonologically possible labels (and as such resemble reverse speech or 'content-filtered speech' – see Balaban and Waxman, 1997), and (c) patterned Frames that may be used to indicate some degree of similarity between the objects contained in them. Following prior studies (Gelman & Markman, 1986; Sloutsky et al., 2001; Diesendruck & Peretz, 2013; Johanson & Papafragou, 2016, among others), we adopt a strong test of the effect of such cues on categorization by asking whether a shared Licit Word, Illicit Word or Frame can lead participants to group together objects that vary in perceptual similarity to each other, from extremely visually dissimilar to nearly identical.

Of interest is whether Licit Words are a more powerful cue compared to Illicit Words and Frames for both monolingual and bilingual speakers – especially for perceptually dissimilar Standard/Target pairings – and if so, whether the role of these cues is comparable across monolinguals and bilinguals. One prediction that one might make is that because language is a more available tool for bilinguals, they will be more influenced by the Licit Words. Alternatively, due to experience with two different linguistic systems, bilinguals may be more likely than monolinguals to make use of the Illicit Words. We do not predict any group differences in the use of the non-linguistic Patterned Frames. To further specify any potential effects of bilingualism, we compare bilinguals with a lifetime of experience using two languages and bilinguals who acquired a second language later in life.

Method

Participants

Thirty-two English monolingual adults (19 female) aged 18-21 (M = 18.53, SD = .76) and thirty-two English-dominant bilingual adults (23 female) aged 18-21 (M = 19.03, SD = .97) participated. In the bilingual group, there were sixteen Spanish-English bilinguals who learned Spanish as a native language but learned English in early childhood (before age 5) and spoke English in educational and professional contexts (i.e., heritage language speakers). The other sixteen participants in that group were English-Spanish bilinguals, who learned English as a native language and Spanish as a second language (L2) later in life (i.e., adult L2 learners). All reported English as being their dominant language, and the groups did not differ in terms of self-rated English proficiency. As might be expected, Spanish-English bilinguals rated themselves as having significantly higher Spanish proficiency than the English-Spanish bilinguals, $t(30) = 3.175$, $p = .003$. Bilingual participant characteristics are presented in Table 1. All monolingual and bilingual participants were undergraduate students at the University of Delaware, and received course credit for their participation.

Bilingual Type	Gender	Age	English Prof.	Spanish Prof.
English-Spanish	3 Male 13 Female	18.94 *1*	9.84 *.35*	6.84 *2.08*
Spanish-English	6 Male 10 Female	19.13 *.96*	9.42 *1.1*	8.64 *.88*
Total	9 Male 23 Female	19.03 *.97*	9.63 *.83*	7.74 *1.82*

Table 1: Bilingual participant characteristics (*SD* in italics).

Materials

Thirty-two grayscale photographs of objects chosen to be unfamiliar to participants were used as stimuli. These objects included arcane tools and pieces of other man-made devices. To ensure novelty, participants were asked at the completion of the experiment if they recognized any of the objects. Only one participant correctly identified a strawberry huller.

The 32 novel objects were then divided into 16 pairs, with one object in the pair designated as Standard A, and the other as Standard B. Each pair was morphed together using the Fantamorph program. For each pair, five morphed pictures (Targets) were created at 10%, 30%, 50%, 70%, and 90% similarity to Standard A (according to program specifications). See Fig. 1 for an example.

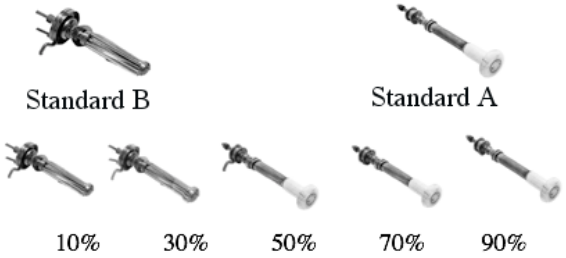


Figure 1: Example stimuli. Percentages reflect target similarity to Standard A.

Eight novel Licit Words were generated using the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). The Licit Words were between three and five letters in length, and consisted of only orthographically existing onsets, orthographically existing bodies, and legal bigrams (e.g., "zeg"). Additionally, all Licit Words were structured such that they could reasonably be English nouns. Eight Illicit Words that could not exist as English (or Spanish) words were also created. These Illicit Words were designed to be unpronounceable to English speakers, and were created by taking the reverse of the novel Licit Word and replacing the vowels with x's. For example, "zeg" would

become "gxz." Finally, eight different patterned Frames (e.g., solid, dashed) consisting of various geometric shapes (e.g., rectangle, hexagon) were created. Each of these frames was large enough to surround a Target or Standard.

Norming. Two norming studies were conducted to ensure that the stimuli were well-controlled. First, it was important to assess whether or not our participants would be sensitive to the similarity between Standards and Targets produced by morphing. A separate group of 10 monolingual English-speaking adults were presented with all possible combinations of each pair of Standards with a corresponding Target, and were asked to rate the similarity of the Target to one of the Standards on a 9-point scale, 10-90%. Participants were very accurate at rating the 30%, 50%, and 70% morphs; ratings did not differ significantly from actual similarity (all p's < .05). Ratings for the 10% and 90% Target stimuli differed significantly from actual similarity, but in the predicted direction; additionally, 10% objects were rated significantly lower than 30% objects, $t(9) = 7.003$, $p < .001$, and 90% objects were rated significantly higher than 70% objects, $t(9) = 10.902$, $p < .001$.

Second, we ensured that all three cue types were equally discriminable. A separate group of 10 monolingual adults from the same population were asked to rate the discriminability of pairs of Licit Words, Illicit Words, and Frames. These pairs are associated with pairs of Standards in the experiment. For each pair, they answered the question 'How similar are these two items?' on a seven-point scale. Ratings did not differ across cue types (p's > .1).

Procedure

Participants were tested individually or in pairs in a quiet room. All tasks were administered on a laptop with a 15.4" screen, approximately 24 inches away from the participant. The session lasted approximately one hour. Before the experiment began, participants completed a language history questionnaire to assess self-rated English (and Spanish, for bilingual participants) speaking, reading, and writing proficiency (from 1-10, 10 being the highest). All participants were tested in English.

Participants were presented with displays that contained triads of objects: Standards A and B from each of the 16 pairs of novel objects and one of the corresponding Targets. Within each triad, the Standards were presented at the top of the screen and the target was centered at the bottom of the screen, separated by a visible line from the Standards (see Fig.2). All three objects in the triad appeared at once and remained on the screen for 1 second. At that point, either one type of cue was introduced (Licit Words, Illicit Words, or Frames condition), or the display remained the same (No Cue condition). Licit and Illicit Words appeared below the objects and Frames surrounded the objects in a triad (see Fig.2). If a cue was present, the cue for the Target always matched the cue for Standard A. After 2 seconds, any cues disappeared and a red frame appeared around the whole display.

At the beginning of the task participants were given on-screen instructions that were reiterated verbally by the experimenter. They were told that they would see groups of three objects, and that they would be given time to look at the objects. Then, a red frame would appear on the screen. When the red frame appeared, they should press a key to indicate where they think the bottom object belongs, either with the object on the right or the object on the left. Participants were instructed to make a response as quickly as possible.

There were four stimuli blocks within the experiment, one for each cue type, counterbalanced across participants, and stimuli were fully rotated. Within each block, participants saw all Target/Standard triads in a random order, 20 trials per block, for a total of 80 trials. Assignment of Standard A to an object within a triad, as well the left/right position of Standard A within a display, was also counterbalanced. The task was administered using the OpenSesame experimental presentation software (Mathôt, Schreij, & Theeuwes, 2012).

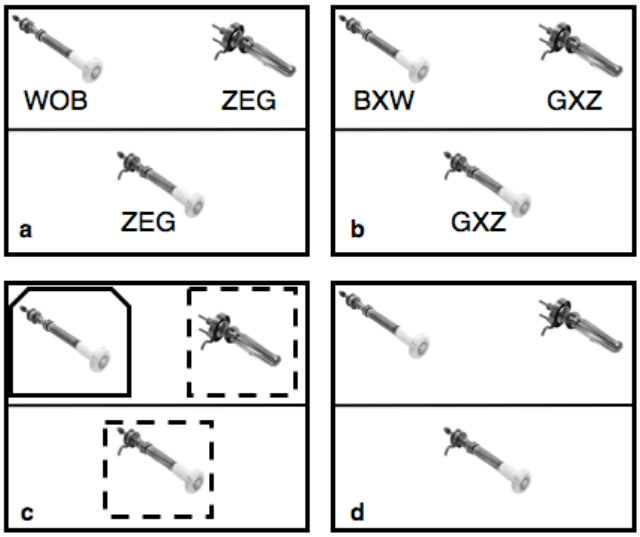


Figure 2: Example stimuli for (a) Licit Word, (b) Illicit Word. (c) Frame, and (d) No Cue blocks.

Following the categorization task, participants completed a lexical decision task as a measure of English proficiency. This task is widely used in both first and second language research, and has been shown to accurately predict language proficiency in various groups of monolinguals and bilinguals (Harrington, 2006), making it an excellent proxy for English language proficiency for our three participant groups. Twenty English words and twenty pseudowords were presented in a random order for 500 milliseconds each. Participants were asked to indicate by keypress, as quickly as possible, whether or not each stimulus was a real English word. Both accuracy and reaction times were collected. For the analysis of reaction times, incorrect responses and trials over 3000 milliseconds were excluded. This task was used to control for English proficiency when performing direct comparisons between the groups.

Results and Discussion

Monolingual Group

A 4 (Cue Type: Licit Word, Illicit Word, Frame, No Cue) by 5 (Perceptual Similarity: 10%, 30%, 50%, 70%, 90%) repeated measures ANOVA was performed with cue compliance as the dependent variable. We considered cue compliant responses to be those where the participant categorized the Target object with the cue-matched Standard A. Cue compliance in the No Cue block was calculated in the same way (i.e., in terms of responses to the predetermined Standard A).

The ANOVA yielded significant main effects of Cue Type, $F(3, 93) = 9.498$, $p < .001$, $\eta^2 = .063$, and Perceptual Similarity, $F(2.44, 75.64) = 709.086$, $p < .001$, $\eta^2 = .823$, as well as a significant interaction between the two, $F(6.14, 190.46) = 2.911$, $p = .001$, $\eta^2 = .043$. As expected, cue compliance increased as Perceptual Similarity increased. Furthermore, there was significantly greater cue compliance in the Licit Word condition ($M = .66$, $SD = .12$) as compared to the No Cue control ($M = .54$, $SD = .06$), $F(1, 31) = 24.144$, $p < .001$, $\eta^2 = .438$, but no significant difference between the Illicit Word ($M = .58$, $SD = .10$) or Frame ($M = .57$, $SD = .11$) conditions and the No Cue condition (both p's $> .1$). More specifically, in the 10% similarity condition, cue compliance in the Licit Word condition was significantly higher than in the Illicit Word ($p = .046$), Frame ($p = .005$), and No Cue ($p = .003$) conditions. The pattern continues into the 30% similarity condition, with higher cue compliance observed for Licit Words as compared to Illicit words ($p = .008$), Frames ($p < .001$), and the No Cue control ($p < .001$). Cue compliance was also significantly higher for the Licit Word condition as compared to the No Cue condition on 50% similarity trials ($p < .001$). No such differences were observed on the 70% and 90% trials, for which cue compliance in the No Cue condition was already at ceiling ($M = .95$ and $M = 1$, respectively). Results are shown in Figure 3.

These findings lead to three main conclusions. First, Licit Words – but not Illicit Words – were treated as labels by our participants. Recall that the cues were presented in a neutral fashion, without any explicit instruction or acknowledgment of their role or significance. Despite this fact, participants spontaneously recruited Licit Words like "zeg" to some degree during their categorization decisions but were not influenced by phonologically impossible Illicit Words like "gxz." Thus, only possible labels in the participant's native language is sufficient to facilitate categorization for monolingual adults.

Second, and perhaps most importantly, we can conclude that language is more powerful than both salient, equally discriminable linguistic (Illicit Words) and non-linguistic (Frames) cues. Labels influence categorization not because they are a simply a salient cue that serves to increase overall similarity, but because adults treat them as category markers.

Finally, we see that participants considered both linguistic and perceptual cues when forming novel categories in our task. When perceptual similarity was high (70% and 90% morphs), participants almost always chose the Standard that shared the same cue as the Target. When perceptual similarity and cues were at odds (10% and 30% morphs), however, participants generally grouped objects that were most similar visually. When the Target was perceptually ambiguous (50%) and could be grouped with either standard, participants were biased towards choosing the label-matched Standard.

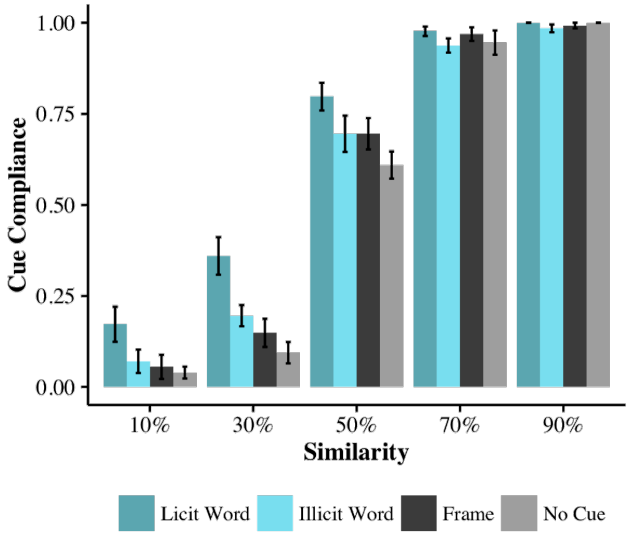


Figure 3: Cue compliance for the monolingual group. Error bars represent +/-1 S.E.M.

Bilingual Group

A mixed ANOVA was conducted with Cue Type (Licit Word, Illicit Word, Frame, No Cue) and Perceptual Similarity (10%, 30%, 50%, 70%, 90%) as within-subjects factors and Bilingual Type (Spanish-English, English-Spanish) as a between-subjects factor. The analysis revealed significant main effects of Cue Type, $F(3, 90) = 11.783$, $p < .001$, $\eta^2 = .081$, and Perceptual Similarity, $F(2.4, 72) = 409.207$, $p < .001$, $\eta^2 = .743$, as well as a significant interaction between the two, $F(6.6, 198) = 5.346$, $p < .001$, $\eta^2 = .069$. As has been observed in the monolingual group, cue compliance increased with Perceptual Similarity, and Licit Word ($M = .70$, $SD = .16$) cue compliance was significantly higher than No Cue ($M = .55$, $SD = .06$) cue compliance, $F(1, 31) = 25.97$, $p < .001$, $\eta^2 = .456$. Unlike the monolingual group, however, there was also significantly higher cue compliance in the Illicit Word condition ($M = .65$, $SD = .15$), as compared to the No Cue condition, $F(1, 31) = 12.130$, $p = .002$, $\eta^2 = .281$. Cue compliance did not differ significantly between the Frame and No Cue conditions, nor between the Licit Word and Illicit Word conditions (p's $> .1$).

Further analyses reveal that for 10% similarity trials, there was higher cue compliance in the Licit Word ($p < .001$) and Illicit Word ($p = .002$) conditions as compared to the No

Cue condition. Cue compliance was also significantly higher for Licit Words ($p < .001$) and Illicit Words ($p = .020$) than Frames in the 10% similarity condition. For 30% similarity trials, cue compliance was significantly higher in the Licit Word ($p < .001$) and Illicit Word ($p = .008$) conditions as compared to the No Cue condition. Furthermore, we observed higher cue compliance for both Licit Words ($p = .006$) and Illicit Words ($p = .020$) in comparison to Frames for 30% similar trials. On 50% similarity trials, again we saw that there was significantly higher cue compliance in the Licit Word ($p < .001$) and Illicit Word ($p = .003$) conditions as compared to the No Cue control. ($p < .001$) Licit Words ($p = .008$), but not Illicit Words ($p > .1$) had significantly higher cue compliance than Frames on 50% similarity trials. There were no differences between cues for the 70% and 90% trials. The main effect of Bilingual Type was not significant, nor were the interactions between Bilingual Type and Cue Type or Bilingual Type and Similarity (all p's > .1). Results from both groups combined are shown in Figure 4.

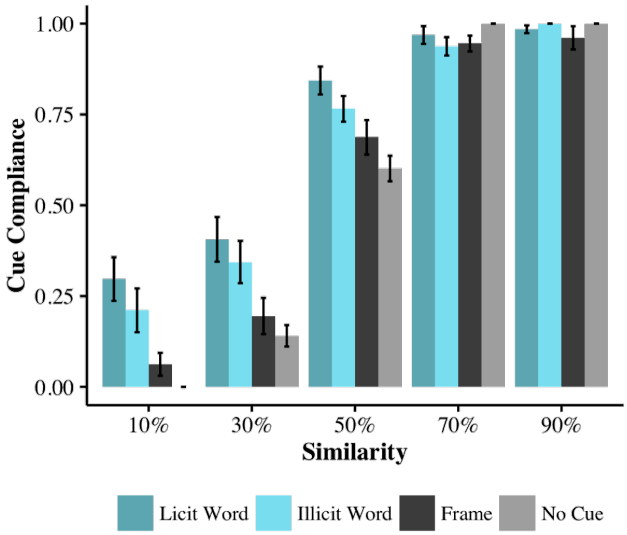


Figure 4: Cue compliance for the bilingual groups (combined). Error bars represent +/-1 S.E.M.

Interestingly, these results indicate that both Licit and Illicit Words are effective in facilitating categorization for bilinguals, and to the same extent. In contrast, only Licit Words influenced the categorization decisions of monolinguals. It is unlikely that the bilingual participants simply were unable to tell the difference between Licit and Illicit words, as they were all English-dominant bilinguals who did not differ from the monolingual participants on a separate test of English proficiency (lexical decision task, not reported here in detail for reasons of space). Rather, what this suggests is that bilinguals are able to use linguistic information more flexibly in category formation, treating linguistic cues that are not possible labels as category markers. Somewhat surprisingly, we also found that there was no difference between Spanish-English bilinguals and English-Spanish bilinguals. Thus, a lifetime of experience using two languages is not necessary to produce this kind of flexibility in the use of linguistic cues for categorization. Instead, a moderately high level of proficiency in a second language appears to be sufficient to cause participants to consider phonologically impossible words as labels. However, this flexibility did not extend to non-linguistic cues. Again, we saw that the patterned frames were not as effective as the linguistic cues, despite the fact that all cues were salient and equally discriminable. In this way, language appears to be a unique tool for promoting categorization. As was observed with the monolingual group, we found that bilinguals were sensitive to both linguistic and perceptual information when making categorization decisions. High visual similarity and a shared linguistic cue (Licit or Illicit Word) encouraged participants to group together two objects.

Conclusion

Previous research indicates that people are more likely to group together objects with a shared label. In the present study, we asked to what extent the influence of labels on categorization is unique for monolingual and bilingual adults by comparing Licit Words to salient, equally discriminable Illicit Words and Frames. Second, we further investigated the influence of bilingual language experience by comparing Spanish-English bilingual heritage speakers to English-Spanish late second language learners.

Our results first suggest that linguistic labels are unique in their ability to facilitate categorization for both monolinguals and bilinguals: two novel objects assigned the same Licit Word were more likely to be grouped together than two objects without linguistic cues. The same effect was not seen for very salient, equally discriminable non-linguistic cues, patterned Frames. Thus, it is not the case that language simply serves to highlight similarities and differences among objects. Rather, linguistic labels (but not shapes) are taken to be category markers. Additionally, we found that perceptual similarity is also taken as an important category marker that works alongside verbal labels. A shared label increased the likelihood of two objects being grouped together, but so did a high level of shared perceptual attributes. These findings contrast with some previous work which has shown that adults completely override perceptual similarity in favor of labels (Sloutsky et al., 2001). However, our results are consistent with more recent work by Gelman and Davidson (2013) which argues for a more nuanced relationship between language and perception in categorization. Because of the influence of similarity, the label advantage is most strikingly observed in the low-similarity conditions (10%, 30%, 50%) for both groups, as cue compliance is already at ceiling for the 70% and 90% similarity conditions.

Critical to our first goal, we found that Spanish-English and English-Spanish bilinguals were influenced not only by Licit Words like "zeg," but also by phonologically impossible Illicit Words like "gxz" when forming novel artifact categories. In this way, bilinguals may be more flexible in the way they integrate linguistic information in

categorization. What we do not know, however, is whether this flexibility comes from differences in the way monolinguals and bilinguals process language, represent object categories (Pavlenko & Malt, 2010), or learn novel words (Kaushanskaya & Marian, 2009).

Interestingly, we found that lifelong experience with two languages is not necessary for this flexibility to arise. We found no differences in behavior between Spanish-English bilinguals who grew up speaking both languages and English-Spanish bilinguals who learned a second language later in life. The flexibility observed here seems to stem, then, from a relatively high level of proficiency in two languages, regardless of age of acquisition.

In sum, we contribute to the existing literature by showing evidence that language is a unique tool for categorization that both monolinguals and bilinguals make use of. Bilingual language experience contributes to increased flexibility in the type of linguistic information that can be considered as a label, but further research is needed to understand the underlying mechanisms and test the limits of this flexibility with groups of bilinguals at varying levels of proficiency.

Acknowledgments

We thank Amelia Weiss for assistance with stimuli creation and data collection.

References

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*, 3–26.

Booth, A. E., & Waxman, S. (2002b). Object names and object functions serve as cues to categories for infants. *Developmental psychology*, *38*(6), 948.

Boutonnet, B., & Lupyan, G. (2015). Words Jump-Start Vision: A Label Advantage in Object Recognition. *The Journal of Neuroscience, 35,* 9329-9335.

Clark, A. (1998). Magic words: How language augments human computation. *Language and thought: Interdisciplinary themes*, 162-183.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, *284*(5416), 970-974.

Dessalegn, B., & Landau, B. (2013). Interaction between language and vision: It's momentary, abstract, and it develops. *Cognition*, *127*(3), 331-344.

Diesendruck, G., & Peretz, S. (2013). Domain differences in the weights of perceptual and conceptual information in children's categorization. *Developmental Psychology*, *49*, 2383–95.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. *Child Development*, *81*, 472–9.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 15231–5.

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition, 105,* 218-228.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Gervits, F., Johanson, M., & Papafragou, A. (submitted). Intentionality and the role of labels in categorization.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and language*, *36*(1), 3-15.

Johanson, M., & Papafragou, A. (2016). The influence of labels and facts on children's and adults' categorization. *Journal of Experimental Child Psychology, 144,* 130-151.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, *62*(3), 499-516.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315-353.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141,* 170.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking redundant labels facilitate learning of novel categories. *Psychological Science, 18,* 1077-1083.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44,* 314-324.

Nazzi, T., & Gopnik, A. (2001). Linguistic and cognitive abilities in infancy: when does language become a tool for categorization?. *Cognition*, *80*(3), B11-B20.

Pavlenko, A., & Malt, B. C. (2010). Kitchen Russian: Cross-linguistic differences and first-language object naming by Russian–English bilinguals. *Bilingualism: Language and Cognition*, *14*(01), 19-45.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology, 55A*, 1339-1362.

Sloutsky, V. M., Lo, Y. F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*, 1695-1709.

Waxman Sandra, R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*, 257-302.

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 253–265.

Learning How To Throw Darts
The Effect Of Modeling Type And Reflection On Dart-Throwing Skills

Janneke van der Loo (j.m.vdrloo@tilburguniversity.edu)
Eefje Frissen (e.r.frissen@tilburguniversity.edu)
Emiel Krahmer (e.j.krahmer@tilburguniversity.edu)
Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands

Abstract

In this study we investigate the effect of modeling type and reflection on the acquisition of dart-throwing skills, self-efficacy beliefs and self-reaction scores by replicating a study by Kitsantas, Zimmerman, and Cleary (2000). Participants observing a coping model were expected to surpass participants observing a mastery model who in turn were expected to outperform participants who learned without a model. Reflection was hypothesized to have a positive effect. Ninety undergraduate students were tested three times on dart-throwing skills, self-efficacy beliefs, and self-reaction scores. Contrary to what was expected, we found no main effects of modeling type and reflection. No interaction effects were found either. There was an effect of trial, indicating that participants improved dart-throwing skills, self-efficacy beliefs, and self-reaction scores over time. Furthermore, self-efficacy beliefs and dart-throwing skill were highly correlated. Our results suggest that learners do not benefit from observing a model and reflecting, but practice makes perfect.

Keywords: observational learning; modeling type; reflection; dart throwing; motor skills

Introduction

The ability to learn from observing others is a key constituent of human social cognition. Observational learning is the process of learning a new task by watching and/or listening how someone else performs this task. It relies on multiple functions: learners should be able to infer others' intentions from action observation, process others' action outcomes (i.e. successes and errors) and combine these sources of information in order to select behaviors leading to desired outcomes later on (e.g. Monfardini, et al., 2013; Rak, Bellebaum, & Thoma, 2013).

Various studies have shown the effectiveness of observational learning in different domains, such as writing (Braaksma, Rijlaarsdam, Van den Bergh, 2002), creative design (Groenendijk, Janssen, Rijlaarsdam, & Van den Bergh, 2013) and motor skill acquisition (Kitsantas, Zimmerman, & Cleary, 2000). A possible explanation for these successes is that the cognitive effort is shifted from executing certain tasks to learning (Braaksma et al., 2002).

Modeling Type

Previous research has shown that within observational learning the type of model can influence the learning process (Braaksma, Rijlaarsdam, & Van Den Bergh, 2002; Shea, Wright, Wulf, & Whitacre, 2010; Wesch, Law, & Hall, 2007; Zimmerman & Kitsantas, 2002). The current general belief is that observing a coping model (someone who gradually improves throughout the video) is more effective than observing a mastery model (a perfect example of someone performing a task) which on its turn is more effective than no model. For example, this effect was found in a study by Kitsantas, Zimmerman, and Cleary (2000). In their study, teenage girls were taught how to play darts by either watching a coping model, a mastery model or no model at all. The girls, who had never played darts before, significantly benefited from observing a model. Additionally, the girls who had observed the coping model outperformed the girls who had observed the mastery model on dart-throwing skills. Furthermore, they found that the girls in the coping model condition had more confidence in themselves, were more satisfied with their results, showed more pleasure and interest in dart-throwing, and attributed fewer of the errors to lack of ability, effort, and practice than the girls in the mastery model condition, who scored better than the girls in the no model condition. A possible explanation for the differences between the outcomes in the coping and mastery model is that a coping model provides more information about strategy implementation and error correction than a mastery model does (Zimmerman & Kitsantas, 2002). Kitsantas, et al. (2002) also found an effect of social feedback: participants being told by the experimenter when their technique was correct. Social feedback enhanced the girls' dart-throwing skill as well as their self-efficacy beliefs and self-reaction scores. The effects of modeling and social feedback were additive, since social feedback was equally effective with all modeling groups and no interactions were found.

Reflection

In most studies on observational learning, learning through modeling is compared with learning without a model, often referred to as 'learning-by-doing'. In learning-by-doing, learners put their own knowledge and beliefs on how a certain task should be done into practice, without observing someone else performing the task. They learn by recognizing and adapting to their own trials and errors (Bandura, 1977). The main difference between observational learning and learning-by-doing is the absence of a model in the latter. However, closer inspection of the methods suggests that there might be a confounding factor;

reflection. Reflection plays an important role within observational learning; the observer watches a model and transfers the provided information to his/her own acquisition by judging which parts could be beneficial and how he/she could use them. In observational learning, reflection is often encouraged by asking participants questions about the model's performance. In learning-by-doing, reflection is not an intrinsic part of the process and thus, reflection does not automatically take place and is often not encouraged within research. Adding reflection to learning-by-doing could therefore make these two types of learning more comparable.

This raises the question whether reflection could account for the differences found between observational learning and learning-by-doing. Reflection within observational learning has been studied by Braaksma, et al. (2002). In their experiment participants observed two models performing a writing task. One group started with the writing task immediately after the observation, whereas two other groups reflected upon the models. Both reflection groups outperformed the no reflection group on the writing tasks. However, this was not compared to learning-by-doing.

Additionally, the influence of reflection on learning in general has been studied. Reflection is found to be beneficial under certain circumstances (see i.e. Bulman & Schutz, 2013; Moon, 2004). Learners can be asked to monitor or record their performances during their learning process. They note their results and/or their own perception on the cause of success or failure. This process is said to enable a learner to detect, attribute and correct their errors (Ferrari, 1996; Zimmerman & Kitsantas, 2002).

Current Study

The common belief in educational research seems to be that both observational learning and reflection have a positive effect on learning. Additionally, the type of model within observational learning seems be of influence. However, how reflection interacts with observational learning is not clear yet. This raises the question whether reflection causes the differences between observational learning and learning-by doing and whether these differences can still be found when reflection is added to learning-by-doing. To study this we will attempt to replicate the previously mentioned study by Kitsantas, et al. (2000) in which the influence of modeling type and social feedback on the acquisition of dart skills was investigated. In this study we will however replace social feedback with reflection. Since Kitsantas, et al. (2002) found no interaction between modeling type and social feedback, we do not expect the replacement of social feedback with reflection to influence the effect of modeling type.

In line with Kitsantas, et al. (2000) we hypothesize that modeling type influences the acquisition of a novice motor skill, self-efficacy beliefs and self-reaction scores (H1). More specifically, we expect that observing a model leads to better acquisition of a novice motor skill, self-efficacy beliefs, and self-reaction scores than not observing a model (H1a) and that observing a coping model leads to better acquisition of a novice motor skill, self-efficacy beliefs and self-reaction scores than observing a mastery model (H1b).

Secondly, based on earlier findings concerning the influence of reflection (e.g. Braaksma, et al., 2002; Bulman & Schutz, 2013), we expect reflection to lead to better acquisition of a novice motor skill, self-efficacy beliefs and self-reaction scores than no reflection (H2)

Thirdly, we hypothesize that reflection reduces the effect of modeling type on the acquisition of a novice motor skill, self-efficacy beliefs and self-reaction scores (H3).

Method

Participants

Ninety undergraduate students at a Dutch University participated in this study. Six participants were removed because they had played darts more than twice in the last six months even though participants were informed about this restriction. This resulted in a sample of 84 participants, of whom 24 were male and 60 were female. The mean age was 21.4 years (SD = 3.27) and 82 participants indicated to have played darts once or zero times in the last six months. All participants were right-handed in order for the instruction videos to be accurate and useful.

Design

A 3 x 2 design was used in this experiment, with modeling type (coping model, mastery model and no model) and reflection (yes or no) as the independent variables. This resulted in six conditions to which the 84 participants were randomly assigned. In line with Kitsantas et al. (2000) all participants were tested on dart-throwing skills, self-efficacy and self-reaction. Whereas Kitsantas et al. (2000) measured these variables once, we did this in three trials within the experiment: at the start of the experiment, after the experimental treatment, and after a ten-minute practice session in order to detect possible improvements as a result of instructions, modeling, and/or practice.

Material

Identical to the study by Kitsantas, et al. (2000), a dart board with regular concentric circles was used instead of a regular dart board. In the study by Kitsantas, et al (2000) the dart board had seven circles, with each circle having a width of 2.54 centimeter, whereas in this study, the circles had the exact same width, but the dart board consisted of nine circles instead of seven. The dartboard was positioned at the official distance of 2.37 meters of the throwing line and with the bull's-eye at a height of 1.73 meters from the ground. Six regular darts were given to the participants in order for them to make six consecutive throws in the tests.

Two different videos were recorded, one of a coping model playing darts (see Figure 1) and one of a mastery

models playing darts. The same model was used in both videos. In the coping model condition, the model started out by making several errors in his dart-throwing technique. These errors were made by not following the instructions to dart-throwing that were used, for instance by holding the dart with all five fingers instead of the three mentioned in the instruction. The model commented on his own performance and gradually improved his dart-throwing skills. In the mastery model video, the model started out with a flawless technique and good skills, and maintained this throughout the video. A right-handed model was used. In order to transfer the model's performance to the participants, all participants were right-handed, otherwise interpretational errors might have occurred. Participants were shown one of the videos if they were in a coping model or mastery model condition and after they had read the instructions.

An online question form was created in order to collect all scores. Participants were seated at a desk with a laptop on which the question form was presented to them and they could independently fill in the questions. All scores were collected and stored in an online database.



Figure 1. Screenshot of the coping model playing darts.

Measures

Dart-throwing skills These were measured by adding up the points for all six consecutive throws. The minimum score per dart was zero and the maximum score was ten. Dart-throwing skill was measured in all three trials. The average dart-throwing score per dart was calculated by dividing the total score by six.

Self-efficacy Before each dart-throwing test, participants were asked how confident they felt that they could (1) throw 9 points with one dart, (2) throw 7 points with one dart, (3) throw 5 points with one dart, and (4) throw 3 points with one dart. Their scores were measured on a scale from 0 to 100 with 10-point intervals, with 0 being not sure and 100 being very sure. This self-efficacy measure was used because the same measure was used by Kitsantas, et al. (2000), with a minor change because of the difference in dart boards. For all three trials, the average score of these four questions was used as their self-efficacy score.

Self-reaction Participants scored their satisfaction with their own performance per test on a 0 to 100 scale for all three trials, with 0 being not satisfied and 100 being very satisfied. As with the self-efficacy measure, this measurement is identical to that of Kitsantas, et al. (2000).

Procedure

Participants were individually taken into a room. First they filled in a consent form, a form with demographic information, and they performed a baseline test in which their self-efficacy, dart-throwing skills, and self-reaction were measured. Within this test, participants first filled in an online form with the question about self-efficacy, then they threw six consecutive darts, and after throwing the darts, they answered the self-reaction question.

As a next step, they were given the instructions to dart throwing on paper. The participants were asked to study these for a few minutes, until they felt confident that they knew what to do. After having read the instruction, the procedure depended on the condition participants were assigned to.

No model, no reflection Participants were tested again straight after reading the instructions. Then, they got to practice dart throwing for 10 minutes.

Coping model, no reflection Participants were shown a video of a dart-thrower who started out by making certain mistakes, but significantly improved over the rounds, the model threw 15 darts in total. After having watched the video, the participants performed the same test again, thus, their self-efficacy, dart-throwing skills, and self-reaction were measured. Then, they got to practice dart throwing for 10 minutes.

Mastery model, no reflection Participants were shown a video of a dart-thrower who made 15 good throws. After having watched the video, the participants performed the same test again, thus, their self-efficacy, dart-throwing skills, and self-reaction were measured. Then, they got to practice dart throwing for 10 minutes.

No model, with reflection Participants were asked to make their first fifteen practice throws. After that, they were asked to reflect upon their own performance according to the five basic skills (grip, stance, sighting, throw, and follow through). After the reflection, they were tested on self-efficacy, dart-throwing skills, and self-reaction and are free to continue practicing afterwards. Because, the practice time had to be equal across all conditions, the duration of the first fifteen throws of the participants in the no model, with reflections group, was part of their total ten minutes of practice. Thus, if a participant performed his/her first fifteen throws in three minutes, he/she had an additional seven minutes to practice after the reflection.

Coping model, with reflection After having read the instructions, participants were shown a video of a dart-thrower who started out by making certain mistakes, but significantly improves over the rounds, the model threw 15 darts in total. Participants were asked to reflect on the model on the aspects explained in the instruction and to report the improvements to be made. After having watched the entire video, they were provided with a form with the five basic skills (grip, stance, sighting, throw, and follow through) as a memory aid. Participants were asked to reflect upon all of these items and write down as much as they remembered. After having watched the video and having reflected on the

models performance, the participants performed the same test again, thus, their self-efficacy, dart-throwing skills, and self-reaction was measured. Then, they got to practice dart throwing for 10 minutes.

Mastery model, with reflection After having read the instructions, participants were shown a video of a dart-thrower who made 15 good throws. They were asked to judge the model on the aspects explained in the instruction and to report the improvements to be made. After having watched the entire video, they were provided with a form with the five basic skills (grip, stance, sighting, throw, and follow through) as a memory aid. Participants were asked to reflect upon all of these items and write down as much as they remembered. After having watched the video and having reflected on the models performance, the participants performed the same test again, thus, their self-efficacy, dart-throwing skills, and self-reaction were measured. Then, they got to practice dart throwing for 10 minutes.

After the ten-minute practice, all participants were tested for one last time.

Analysis

The data were analyzed with a 3 (type of modeling) x 2 (level of reflection) repeated measures multivariate analysis of variance, with dart-throwing skills, self-efficacy, and self-reaction as dependent measures. A repeated measures ANOVA was conducted in order to discover the changes in the dependent scores over the three trials that were conducted within the experiment. Mauchly's test for sphericity was used to test for homogeneity of variance, and when this test was significant we applied a Greenhouse-Geisser correction on the degrees of freedom.

Results

An overview of the mean scores and standard deviations for all three trials in all six groups are shown in Table 1.

Dart-throwing skills

A within-subjects main effect of trials was found for dart-throwing skills (F (2,156) = 32.96, $p < .01$), with the mean scores gradually increasing each trial (M_{trial1} = 4.63, *C.I.* [4.26 , 5.00], M_{trial2} = 4.91, *C.I.* [4.55 , 5.26], M_{trail3} = 5.94, *C.I.* [5.64 , 6.24]). The difference between trial 1 and 3 (M = -1.31) was significant, $p < .01$, *CI* [-1.73 , -.89] and so was the difference between trial 2 and 3 (M = -1.03), $p < .01$, *CI* [-1.46 , -.61]). However, the difference between trial 1 and 2 (M = -.28) turns out not to be significant, p = .29, *CI* [-.68 , .13].

No significant main effect of modeling type on dart-throwing skills was found, F (2,78) = 2.33, p = .10. Even though the data numerically suggest that best results are obtained by observing a mastery model (M = 5.55, *C.I.* [5.08 , 6.02]), followed by a coping model ((M = 5.13, *C.I.* [4.65 , 5.61]) and learning-by-doing (M = 4.80, *C.I.* [4.28 , 5.32]), this effect did not yield significance.

Reflection was discovered not to have a significant main effect on dart-throwing skill either, F (1,78) = 0.88, p = .35. No reflection even led to numerically higher scores on dart-throwing skills (M = 5.29, *C.I.* [4.89 , 5.70]) than reflection (M = 5.03, *C.I.* [4.64 , 5.42]), even though this difference was not significant.

In addition, none of the interaction effects were found to be significant ($F_s < 1.52$, $p_s > .2$).

Self-efficacy

A main effect of trials on self-efficacy was found, F (1.80, 135.10) = 145.52, p < .01. Self-efficacy scores on the first test (M = 36.86, CI [33.37 , 40.36]), the second test (M = 49.77, CI [45.92 , 53.62]), and the third test (M = 61.10, CI [57.20 , 64.99]) were significantly different. All the individual differences were significantly different at a p-value of < .01.

The main effect of modeling type did not yield significance, F (2,75) = 2.10, p = .13. Again, the data numerically suggest that best results are obtained by observing a mastery model (M = 53.45, CI [47.82, 59.07]), followed by observing a coping model (M = 49.44, CI [43.72 , 55.17]) and learning-by-doing (M = 44.84, CI [38.66 , 51.02]). However, this effect failed to reach significance.

The main effect of reflection was not found to be significant either, F (1,75) = 3.40, p = .07. Similar to the effect of reflection on dart-throwing skills, the data numerically suggest that best results are obtained by not reflecting (M = 52.37, CI [47.62 , 57.12]) instead of reflecting (M = 46.12, CI [41.32 , 50.92]).

Again, no significant interaction effects were found (F_s < 1.37, $p_s > .26$).

Self-reaction

The results for self-reaction were very similar to those of self-efficacy. A significant main effect of trial on self-reaction scores was found, F (2, 156) = 4.56, p = .01. Again, the scores on the first test (M = 65.93, CI [61.84 , 70.02]), were significantly different from scores on the second test (M = 59.24, CI [54.74 , 63.74]) and the third test (M = 67.02, CI [62.53 , 71.52]). The difference between the first test and the second test (M = 6.69, CI [0.89 , 12.48], p = .02), and that between the second test and the third test (M = -7.78, CI [-13.53 , -2.03], p = .01) were significant. The difference between the first and the third test, however, did not turn out to be significant (M = -1.09, CI [-6.18 , 3.99], p = .67).

No significant main effect of modeling type, (F (2,78) = 1.61, p = .21) or reflection (F (1,78) = 0.58, p = .45) was found, indicating that learning-by-doing (M = 60.33, CI [54.91, 65.75]), observing a mastery model (M = 66.79, CI

Table 1: Average Dart-throwing, Self-efficacy and Self-reaction Scores (with SDs) as a function of Group (Reflection, No reflection), Model (No Model, Mastery, Coping) and Trial (1, 2, 3)

	Trial nr	No Model			Mastery			Coping		
		1	2	3	1	2	3	1	2	3
Dart skills	No reflection	4.22	4.68	5.97	5.18	5.34	6.34	5.14	4.82	5.93
		(1.90)	(2.07)	(1.52)	(1.75)	(1.57)	(0.82)	(1.59)	(1.46)	(1.16)
	Reflection	3.78	4.55	5.60	4.98	5.53	5.92	4.48	4.82	5.89
		(1.61)	(1.50)	(1.35)	(1.43)	(1.67)	(1.37)	(1.84)	(1.91)	(1.84)
Self-efficacy	No reflection	35.73	44.31	56.69	44.57	59.88	67.65	41.38	55.71	65.39
		(17.05)	(17.42)	(17.19)	(15.79)	(16.09)	(13.80)	(16.69)	(19.81)	(17.47)
	Reflection	33.42	42.19	56.71	37.57	49.64	61.36	28.52	46.88	58.79
		(16.15)	(19.76)	(19.23)	(13.20	(13.76)	(19.05)	(15.51)	(16.89)	(18.38)
Self-reaction	No reflection	66.92	57.83	65.08	70.20	61.60	73.33	67.07	56.43	68.36
		(16.80)	(26.73)	(22.26)	(20.68	(11.83)	(14.72)	(21.14)	(17.78)	(18.64)
	Reflection	51.46	55.92	64.77	67.70	63.07	64.80	72.20	60.60	65.80
		(26.22)	(24.63)	(14.58)	(9.67)	(21.69)	(25.56)	(14.75)	(19.69)	(24.39)

[61.85 , 71.73]) and observing a coping model (M = 65.08, CI [60.04, 70.11]) do not lead to significantly different self-reaction scores, and neither do not reflecting (M = 65.20, CI [60.96 , 69.45]) and reflecting (M = 62.93, CI [58.79 , 67.07]). Again, the numerical differences suggest best results are obtained by observing a mastery model, followed by a coping model and learning by doing, and by not reflecting instead of reflecting.

For self-reaction, no significant interaction effect were found ($F_s < 0.87$, $p_s > .49$).

Correlations

To further explore the relationships between dart-throwing skills, self-efficacy and self-reaction, we used a Pearson's correlation test. In the first trial the results show a strong correlation between self-efficacy and dart-throwing skills, $r(81) = .60$, $p < .01$., and a medium strong correlation between self-reaction and dart-throwing skills, $r(84) = .47$, $p < .01$. The correlation between self-efficacy and self-reaction was less strong, but still significant, $r(81) = .25$, $p < .05$.

For the second trial, again the correlation between self-efficacy and dart-throwing skills, $r(81) = .52$, $p < .01$., and the correlation between self-reaction and dart-throwing skills, $r(84) = .49$, $p < .01$.are significant and of medium strength. However, the correlation between self-efficacy and self-reaction is not significant and weak in this second test.

The correlation scores of the dependent variables on the third trial show that self-efficacy and dart-throwing skill, $r(81) = .41$, $p < .01$. and self-reaction and dart-throwing skill, $r(84) = .67$, $p < .01$. are again strongly connected, whereas a weaker, but significant connection can be observed between self-efficacy and self-reaction, $r(81) = .15$, $p < .05$.

General Discussion and Conclusion

The aim of this study was to determine what the effects of modeling type and reflection are on the acquisition of dart-throwing skills, self-efficacy beliefs and self-reaction scores. First of all, modeling type was expected to have an effect, with participants observing a coping model surpassing participants observing a mastery model who in turn were expected to outperform participants who learn without observing a model. Even though we found numerical differences between the modeling type conditions, with participants in the mastery condition scoring higher on dart-throwing skills, self-efficacy and self-reaction scores, than participants in the coping model condition, who on their turn scored higher than the participants in the no model condition, these differences were not significant. This indicates that in this study it did not matter if the learner observed a coping model, a mastery model, or no model at all. This was the case for dart-throwing skill, self-efficacy beliefs and self-reaction scores. Thus, in this study we found no support for H1a and H1b: observing a model did not lead to a better learning process than not observing a model, and observing a coping model did not lead to a better learning process then observing a mastery model, which is not in line with what was found by Kitsantas et al. (2000).

Similar results were found for reflection. We expected reflection to have a positive effect, but we found no significant effects for reflection. The no reflection group even scored numerically higher than the reflection group. Thus, in this study H2 was not supported: reflection did not lead to a better learning process than no reflection.

No interaction effects were found either. It was hypothesized that reflection would reduce the difference between observational learning and learning-by-doing (H3), but as no such effect was found, reflection could not influence this effect. In addition, no other interaction effects between number of trials, modeling type, reflection were

found on dart-throwing skills, self-efficacy beliefs, or self-reaction scores.

It was however found that dart-skills and self-efficacy beliefs significantly improved with each trial. Thus, our data suggest that it did not matter how someone was taught how to play darts, as long as they practiced, their scores improved and they felt more confident. Additionally, dart-throwing skills and self-efficacy beliefs were highly correlated, meaning that how higher the self-efficacy beliefs of a person were, the higher their dart-throwing score was.

In the current study we were not able to replicate the findings by Kitsantas, et al. (2000) who found very clear effects of modeling type. Our test was sufficiently powered: the sample was larger than in Kitsantas et al. (respectively 90 and 60) and the test was sensitive enough to find statistical differences for trial, indicating that the participants did learn from the interventions. The question that can be raised therefore is why there are such substantial differences between the findings of this study and that of Kitsantas, et al. (2000).

Perhaps the change in population contributes to the differences between the current study and the Kitsantas et al. study. Whereas Kitsantas, et al. made use of 14 to 16 year old girls, both male and female students with an average age of 21.4 years were used as participants in this study. Because of the higher age, it is plausible that our participants have watched someone else play darts before, either on television or in real life. This observational experience could lead to them not being so inexperienced. Lee and White (1990) discovered that participants who merely observed someone performing a certain motor task, acquired some amount of skill themselves. Thus, participants who have taken part in the present study might unconsciously have been more experienced than expected.

The girls in the Kitsantas, et al. study (2000) had never played darts before. Our participants were more likely to have played darts before. We tried to control for this by using a dart-experience restriction. Since the majority of the participants had played darts once or more, we made a post hoc distinction between the participants who had never played darts (N = 32) and the ones who had played (N = 52). The repeated measures multivariate analyses of variance was run again, but now split in two separate groups, the group with no experience at all and the one with some experience. The analysis with the average number of point as the repeated measures variable showed no significant results for either group. Thus, even in the no experience group, no effects of modeling type or reflection were found. The same can be said for the analyses with self-efficacy and self-reaction as the repeated measures. The outcomes of those analyses were not found to be significant either.

In sum, the results of Kitsantas, et al. (2000) were not replicated in this study, even though the experimental procedures and measurements were very similar. Learners did not benefit from observing a model or reflecting. However, they did improve by practicing. The results of this study are also not in line with earlier findings by Braaksma, et al. (2002), Ferrari (1996), Shea, et al. (2010), Wesch, et al. (2007), and Zimmerman and Kitsantas (2002). Our results indicate that the effects of observational learning and the role of reflection are still not clear and therefore more research into this subject is needed.

References

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, N.J.: Prentice Hall.

Bulman, C., & Schutz, S. (Eds.). (2013). *Reflective practice in nursing*. Oxford, England: John Wiley & Sons.

Braaksma, M. A., Rijlaarsdam, G., & Van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology*, *94*(2), 405.

Ferrari, M. (1996). Observing the observer: Self-regulation in the observational learning of motor skills. *Developmental review*, *16*(2), 203-240.

Groenendijk, T., Janssen, T., Rijlaarsdam, G., & van den Bergh, H. (2013). The effect of observational learning on students' performance, processes, and motivation in two creative domains. *British Journal of Educational Psychology*, *83*(1), 3-28.

Kitsantas, A., Zimmerman, B. J., & Cleary, T. (2000). The role of observation and emulation in the development of athletic self-regulation. *Journal of Educational Psychology*, *92*(4), 811.

Lee, T. D., & White, M. A. (1990). Influence of an unskilled model's practice schedule on observational motor learning. *Human Movement Science*, *9*(3), 349-367.

Monfardini, E., Gazzola, V., Boussaoud, D., Brovelli, A., Keysers, C., & Wicker, B. (2013). Vicarious neural processing of outcomes during observational learning. *PloS one*, *8*(9), e73879.

Moon, J. A. (2004). *A handbook of reflective and experiential learning: Theory and practice*. New York, NY: RoutledgeFalmer.

Rak, N., Bellebaum, C., & Thoma, P. (2013). Empathy and feedback processing in active and observational learning. *Cognitive, Affective, & Behavioral Neuroscience*, *13*(4), 869-884.

Shea, C. H., Wright, D. L., Wulf, G., & Whitacre, C. (2000). Physical and observational practice afford unique learning opportunities. *Journal of motor behavior*, *32*(1), 27-36.

Wesch, N. N., Law, B., & Hall, C. R. (2007). The use of observational learning by athletes. *Journal of Sport Behavior*, *30*(2), 219.

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology*, *94*(4), 660.

On the Tragedy of Personnel Evaluation

Momme von Sydow (momme.von-sydow@lrz.uni-münchen.de)
[1]University of Munich (LMU), Munich Center for Mathematical Philosophy (MCMP), Ludwigstr. 31, D-80539 München
[2]University of Heidelberg, Department of Psychology, Social Psychology, Hauptstr. 47-51, D-69117 Heidelberg, Germany
(momme.von-sydow@psychologie.uni-heidelberg.de)

Niels Braus (n.braus@stud.uni-heidelberg.de)
University of Heidelberg, Department of Psychology, Hauptstr. 47-51,
D-69117 Heidelberg, Germany

In social-dilemma situations (public-good games) people may pursue their local, egoistic interests and thereby lower the global, overall payoff of their group and, paradoxically, even their own resulting payoff. One may also speak of intra-individual dilemmas, where people pursue local goals at the expense of their overall utility. Our current experiments transfer this idea to the context of personnel evaluation and personnel selection. In our experiments, participants were put in the position of a Human Resources manager, who should for instance select workers who optimize the overall payoff of the company, rather than those who optimize only their specific payoffs. The results of the experiments, however, suggest that most, albeit not all, participants tended to focus on directly comparing individuals without considering the overall contribution to a group. Thus employees with the best overall effects for a company or organization may be evaluated the most negatively. This possible 'tragedy of personnel evaluation' may be linked to maladaptive incentive structures (personnel evaluation), advancement of employees (personnel promotion) and job offers (personnel selection), and may have a substantial negative impact on the effectiveness of companies or organizations.

Keywords: intra-individual dilemmas; social dilemmas; personnel evaluation; personnel selection, altruism; causal induction; global vs. local optimization; less-is-more effect

Inner-Organizational Dilemmas

Adam Smith (1776) famously argued that no altruism is needed to promote the common good: "By pursuing his own interest he [the individual] frequently promotes that of the society more effectually than when he really intends to promote it."

In contrast, in the wider context of evolutionary biology, economics and psychology, it has long been noted that social dilemma situations often arise when groups of organisms or human agents act 'optimally' – in the sense of pursuing their egoistic interests only. People in such situations may often behave contrary to the best interest of the whole, which paradoxically often results in disastrous outcomes for each individual as well. For instance, it has been argued that the over-exploitation and destruction of finite public resources by 'rationally' acting, selfish individuals is inevitable. This has been called the "tragedy of the commons" and has been discussed with regard to environmental pollution and sustainable development (Hardin, 1968). Social dilemmas have widely been studied, theoretically and empirically, in psychology and behavioural economics. Recent decades have yielded lively debate of possible solutions of the perhaps not always inevitable tragedies. Moreover, the debates have eroded the explicit or tacit strict egoism assumption without ignoring the limits of altruism (e.g., Fehr & Fischbacher, 2003, 2004; Fehr & Schmidt, 1999; Fehr & Gächter, 2002; Henrich, 2005; Ostrom et al., 1999). Likewise, evolutionary biology in recent decades has shifted from an emphasis on individual egoism (or gene-egoism) to an acknowledgement of multi-level approaches that generally suggests one should find egoistic as well as altruistic behaviour tendencies in social groups (Nowak & Sigmund, 2005; Sober & Wilson, 1999; Wilson & Wilson, 2007; cf. von Sydow, 2011).

In organizations it is also plausible, then, that each member frequently faces tension between egoistic and group-serving behaviours. The issue is not constrained to moral questions of pursuing the common good of a society, but arises within companies, for managers and workers alike, even when they merely aim to optimize certain economic key figures (e.g., net sales or operating profit). Solving such intra-organizational dilemmas, therefore, is one of the most crucial tasks of building efficient systems of co-operation.

One obvious potential solution of inner-organizational dilemmas seems to install an institutionalized, neutral third-party that tries to access justly the contributions of egoistic versus group-serving individuals. This seems linked to good leadership or should at least be a major concern of management control systems or Human Resources managers.

Yet do people really have the capacity for "altruist detection" or "egoist detection"?

First, one must concede that basic biological multi-level models of group altruism do not necessarily require the ability of organisms to detect altruist or egoists (Sober & Wilson, 1999; Wilson & Wilson, 2007). However, there are refined models that refer to such abilities, can be evolutionary stable without necessitating cultural transmission (Nowak & Sigmund, 2005). Thus it seems implausible to deny a general ability of altruist detection in humans.

Second, a *prima facie* potential for egoist or altruist detection – or, put more cautiously, of co-operator and cheater detection – seems quite plausible (at least for human beings) if we consider the importance of the topics in daily contexts. One might object that, in times when multi-level models had a bad standing in evolutionary biology, there was likewise psychological evidence in simple rule-testing tasks (Wason Selection Tasks) that people are only

successful at detecting cheaters but not co-operators or altruists (Cosmides, 1989). However, further research has suggested that this greater facility for cheater-detection depends on the question asked and the goals pursued (von Sydow, 2006; von Sydow & Hagmayer, 2006).

Third, in the specific context of our experiments on Human Resources management and personnel evaluation, it seems even clearer, at least from first sight, that people in management roles should be able roughly to discern who contributed the most to a company's success (if provided appropriate quantitative information). One reason is that, from the vantage point of game-theoretical analysis, the social dilemma becomes an "*intra*-individual dilemma" (von Sydow, 2015) that is normatively to be solved globally; the personnel manager should at least normatively optimize globally and refrain from selecting or rewarding employees with good individual operating figures over those who overall best serve the interests of the work group, company or organization. However, even for intra-individual dilemmas, it has been shown that people may optimize locally, at the expense of the global level (von Sydow, 2014). For instance, an individual may pursue too many projects, each with a positive utility, but may thereby reduce overall utility by ignoring those projects' negative external effects on other projects. Transferring the idea that people may often not realize or acknowledge the positive externalities if local effects, may in the context of a human recource management task imply that people may ignore interactions with other employees or the distributed indirect effects of group-serving altruism. The use of a simplified one-level model – considering only individual outputs – might be due to issues of complexity. Even if plausibly – as we argued before - people in daily life can distinguish egoistic from group-serving individuals, this need not apply to economic personnel evaluation contexts were evaluation is based on given numbers, where reporting units often have no direct acquaintance with the persons and processes involved, and in such contexts one might perhaps more easily ignore questions of group-serving behavior. Additionally, the context of personnel evaluation may elicit a competitive individual frame of mind.

This prompted our hypothesis that people in charge of personnel-evaluation may ignore the positive or negative effects of individuals on the work of other members of a larger group (externalities), sticking only to a person's direct operating figures. This, however, may imply a "tragedy of personnel selection", since people with the best overall effects on group-performance (e.g., by helping others and indirectly contributing the most to the overall goals of a company), might be evaluated the most negatively.

Experiment 1

In Experiment 1, participants put themselves in the role of Human Resources manager evaluating the utility of employees in a local shop. The managers obtained information on direct earnings of individuals and those of the shop daily. The overall profit strongly positively depended on the presence of an individual employee in a shift, suggesting his or her importance to the overall earnings of the team. We here call this employee "the altruist," since although not demonstrating particularly positive earning figures individually, he or she enhances the values for all other employees working during the same shift. Note, that although a functionally defined altruism would lead to the shown data, the number-based evaluation context does provide evidence for any intentions or effort. Hence, a perhaps more neutral term for the worker would have been 'facilitator'. We here nonetheless use the term 'altruist' as one easily conceivable interpretation. We aimed to explore how far participants as Human Resources managers take the essential indirect monetary contributions of an altruist worker into account, when evaluating workers.

Design

Experiment 1 has a two (rounds: 10 working days versus 20 working days) by two (difference of earnings between a normal worker with versus without the presence of the altruist; see Table 1), within-subjects design.

Table 1: Mean earnings of normal workers (NW) and altruist; and overall earnings

	Condition 1 / 2	Condition 3 / 4
NW without altruist	2200 €	2000 €
NW with altruist	2800 €	3000 €
Altruist	1600 €	1600 €
Overall without altruist	8200 €	7600 €
Overall with altruist	10000 €	10600 €

Method

Participants 124 participants from the US volunteered for the experiment via MTURK, each obtaining a reward of $1. 120 participants finished the computer experiment (59% male, 41% female; mean age was 34 years; highest education: 48% Bachelor or Master degree; 40% high school degree). Participants were randomly assigned to one of the four conditions (cf. Design).

Procedure and Material Participants in a computer experiment should imagine being in the role of a Human Resources manager evaluating the staff of a particular snack bar. There were five staff members, but in each day only four are working. Compared to many real personnel evaluation situations one with only five persons is relatively simple (cf. von Sydow, 2015, for a plausible inner-individual dilemma with ten nodes). Participants were instructed to establish which workers contributed most to the overall profit of the company, based on data provided by the reporting unit of the larger company.

First, participants in all conditions read the same instructions and overview description. Then, in the main part, they obtained for each day transparent overview information about the individual earnings of each of the four employees (presented by a picture) working in the shift (at this day), and information on the overall earnings (Figure 1).



Figure 1: Example of shown earnings at the individual and group level on a particular day (Condition 1 or 2).

The shown earnings of the normal workers and the altruist worker were based mainly on the mean earnings shown in Table 1. Additionally, we added some noise to each value (a normal distribution with SD = 600 €).

We randomly assigned the altruist-role to one of the pictures. The altruist randomly appeared in 6 of 10 rounds (days). The four normal workers appeared randomly (on average, 7.5 times in 10 rounds; or 15 times in 20 rounds). We counterbalanced the presentation-order in each shift. Participants could view the overview panels for each day as long as they wanted and proceeded to the next round by clicking a button when ready. After the 10 (or 20) rounds, the 'Human Resources managers' evaluated the employees of the snack bar in four tasks (each on a separate page).

First they had to choose which of the five workers was of "the *greatest* total utility for your business". If unsure, they had to choose intuitively. Second, they chose which person represented the lowest utility for the business. Third, they ranked employees according to "their total utility for the company in the present setting" (not reported here). Fourth, they rated the earnings of the five workers (presented randomly) on a scale of 1 to 10.

Finally, they had to provide further comments and demographic data. Additionally, they were tested on the Cognitive Reflection Task (Frederick, 2005) – a short task measuring the tendency to override an initial automatic albeit incorrect response by a reflectively corrected one – and a attention test item checking how careful or superficial people read instructions.

Results and Discussion

Figure 2A shows that, in all conditions, a clear majority of participants did not judge the altruist as representing the highest utility for the company and there were hence no large differences between conditions, the differences were reliable (exact $Chi^2(3, 124) = 3.33$, $p < .05$), with an average of 7% detecting that the altruist had the highest utility in Conditions 3 and 4, with only 1.5% in Conditions 1 and 2. The preponderant selection of normal workers over altruists over all conditions was statistically even more clearly above chance ($Chi^2(1, 124) = 17.8$, $p < .001$). Figure 2B shows that, across conditions, a clear majority of participants assigned even the lowest utility to the one who influenced the overall earnings of the work group the most positively. Given the higher number of normal workers, these judgments are most clearly above chance-level, even for the most critical condition: $Chi^2(1, 32) = 50,8$, $p < .0001$).

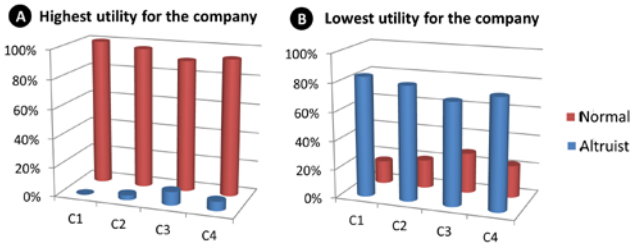


Figure 2: Percentages of choices of normal worker or altruist worker as representing the highest (Panel A) or lowest (Panel B) utility for the company (Condition 1 or 2).

The rating task (Figure 3) likewise showed that, in all conditions, the altruist was predominantly judged to be of lower utility to the company than the normal workers.

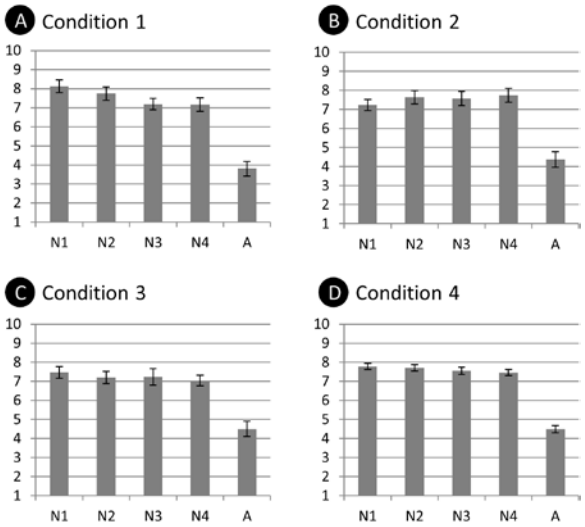


Figure 3: Average ratings (with SE) in Experiment 1 for the four normal workers (*N*) or altruist workers (*A*) of Conditions 1, 2, 3, and 4 (Panel A to D).

Overall, the results show that even if the 'altruist' worker causes a *decidedly* higher increase in the overall earnings than any 'normal' worker (Conditions 3 and 4) and this could be observed over 20 rounds (Condition 2 and 4) the altruist's overall contribution is not reflected by the evaluations of most participants (even though a small significant effect of conditions was found). Actually, the altruist is even evaluated most negatively by the vast majority of participants.

Experiment 2

Experiment 2 used the same personnel evaluation scenario as in Experiment 1, involving an altruist who is individually the lowest earner of the team but overall contributes the

most to the success of the group. However, Experiment 2 addressed several additional issues:

- *Number of days and repeated measurement*. We investigated whether most people's negative evaluation of the group-serving altruist changed, if people had to go through more learning-rounds. Additionally, the participants had to evaluate the workers not only once, but repeatedly.
- *Personnel selection*. We added a personnel *selection* task, wherein the manager determined the configuration of a shift. This should focus people more on the overall earnings of a group, allowing for easier detection of the altruist's highest overall contribution.
- *Selection of participants*. Due to bad results in the attention test, we aimed to rule out effects of unmotivated participants from MTURK. Hence we rigidly selected participants to exclude inattentive participants at the beginning of the experiment.
- *Sensitivity to distinguishing individual earnings*. Finally, we varied the earnings of the 'normal' workers (whose presence did not affect other workers' performance) to investigate whether participants were able to distinguish different individual earnings.

Design

Table 2 shows the varied average earnings of the normal workers in four conditions (C1 to C4). C1 has the same payoff-structure as conditions C3 and C4 of Experiment 1 (with the clearest altruist's contribution). Now in C2 additionally one normal worker (N1) stands out. In C3, N1 and N2 differ from N3 and N4, and in C4 all normal workers differ from one another.

Table 2: Mean earnings of normal workers (N1 to N4), the altruist worker, and the overall earnings, with and without the altruist in the four conditions

		C1	C2	C3	C4
Without Altruist	N1	2000€	2300€	2400€	2600€
	N2	2000€	1900€	2400€	2200€
	N3	2000€	1900€	1600€	1800€
	N4	2000€	1900€	1600€	1400€
With Altruist	N1	3000€	3300€	3400€	3600€
	N2	3000€	2900€	3400€	3200€
	N3	3000€	2900€	2600€	2800€
	N4	3000€	2900€	2600€	2400€
Altruist	(A)	1600€	1600€	1600€	1600€

In all conditions, the altruist's presence strongly affected the other workers' earning and correlates highly and consistently with group's high earnings. Nevertheless, the altruist individually had the lowest earnings. The overall average earnings were kept constant over the conditions.

Method

Participants and strict selection criteria 228 participants from the US checked the first page of the experiment in MTURK; 156 continued and passed a first participation-criterion (time spent on the first page > 20 sec. and < 6 min.). Only 140 (90%) passed Criterion 2, i.e. correctly rephrased the task. Of them, 120 (86 %) finished the experiment (52% male, 48 female; mean age: 33 years; highest education: 59% Bachelor or Master degree; 38% high school degree). The participants were volunteers, obtaining a reward of $1. Participants were randomly assigned to one of four conditions (cf. Table 2).

Material and procedure We used a similar task with a similar procedure and materials as in Experiment 1 (cf. Figure 1). However, instead of 10 or 20 rounds with information about the individuals' earnings and group-earnings, participants here obtained information about 40 working days (again each with four workers). For each day we required participants to stay a minimum time on the information screens (4 sec.), after which the 'continue' button could be pressed. We applied similar counterbalancing measures as in Experiment 1.

Additionally to the changed payoff-structure (Table 2) we now had *four* test-phases, where the workers should be evaluated, with the test-phases following every ten rounds. In the first three of them (following Rounds 10, 20 and 30) we presented only one evaluation task but we added a new personnel selection task. In the former, participants rated the workers' contribution to the group on a scale of 1 to 10. In the new personnel selection task, participants were asked which four of the five employees they would select to work in a further shift the next day. It was mentioned that all five employees wanted to work and that their choice should optimize the profit for the company on that day. In the fourth assessment phase we used all the evaluation measures used in the single test-phase of Experiment 1 as well as the described personal selection task.

Finally, we used (a) a brief Kimchi-Palmer–Test (similar to a Navon-Test) to explore global versus local perception preferences; (b) the attention test; and (c) participants supplied demographic data and comments on the task.

Results and Discussion

Since space precludes exhaustive treatment, we here present only the results for the fourth and final assessment phase of employees by the 'managers'. However, the results of all four phases are surprisingly similar.

Figure 4 presents the results of the rating task over all four rounds. It shows that the order of the average ratings reflects the individual mean differences shown between the normal workers. People seem to be sensitive to these individual differences, but not to the larger overall increase of group earnings occurring when the altruist was in the team. The ratings for the altruist in all conditions and phases were lowest from the first phase onwards. Moreover, the judgments appear quite stable over the time. Even in Phase 4 and the most critical condition, C4, the data was clearly at odds with a potential prediction linked to judgements based on the worker's overall contribution (A > N1 > N2 > N3 >

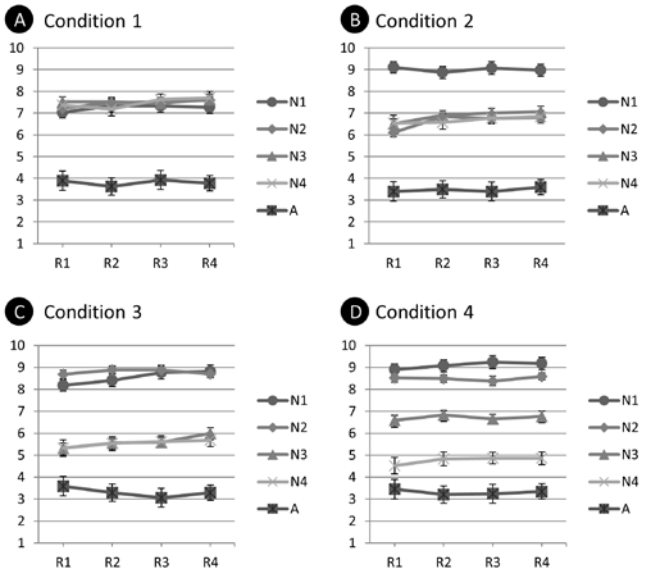


Figure 4: Average ratings (with SE) in Experiment 2 for the four normal workers (*N*) and altruist workers (*A*) in phases (R1 to R4) of Conditions 1, 2, 3, and 4 (Panel A to D).

N4) and contrasts showed significant results for all five mean differences predicted based on *individual* contributions (N1 > N2 > N3 > N4 > A): N1 > N2: $F(1, 28) = 5.5$; $p < .05$; N2 > N3: $F(1, 28) = 96.8$; $p < .001$; N3 > N4: $F(1, 28) = 34.5$; $p < .001$; N4 > A: $F(1, 28) = 37.9$; $p < 001$.

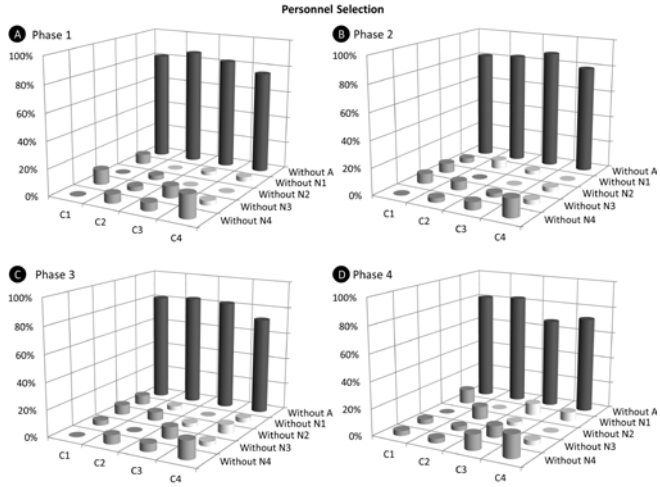


Figure 5: Results of the personnel selection task in the four test phases, showing the proportion of 'managers' choosing a team of four out of five, thus excluding worker N1, N2, N3, N4, or the altruist worker.

The added personnel-selection task (Figure 5) could have elicited better results, since it emphasizes the *overall* earnings of different group configurations. Participants may thus have realized the clearly lower outcome of configurations without the altruist relatively to the four other configurations. But Figure 5 shows that, even in this task, the participants tended to exclude the best player from the team (dark selections). This was quite stable over time. We also marked optimal selections that did not only involve the altruist, but the best normal workers (dark grey shading). But even without this distinction, and in the final Round 4, the altruist were excluded clearly more often (dark shading) than *all* other workers together, $\chi^2(1, N = 120) = 32.0$, $p < .001$), also in the most critical condition, C4 ($\chi^2(1, N = 29) = 5.83$, $p < .05$).

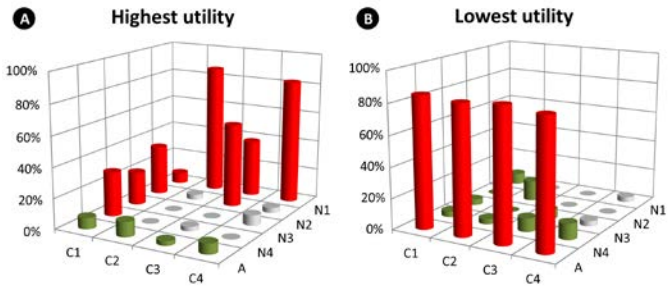


Figure 6: Percentage of 'managers' choosing either a normal worker (*N*) or the altruist worker (*A*) as of the highest (Panel A) or lowest (Panel B) utility for the company (Conditions C1, C2, C3, C4). The choices based on individual earnings are marked in red, those based on overall earnings in green.

Figure 6 presents the proportions of the judgments in the last test phase concerning which workers had the highest and which had the lowest utility for the company. A clear majority in all conditions did not assign the *highest* utility based on the overall impact on the earnings of the group (coloured green), but in line with individual earnings (coloured red) (C4: $\chi^2(1, N = 29) = 24.5$, $p < .001$). In all conditions, the majority of participants also assigned even the *lowest* utility to the 'altruist' (Figure 4B). This was the case in the most unclear condition (C4), since the red choices were significantly more frequent than all other choices taken together ($\chi^2(1, N = 29) = 24.5$, $p < .001$).

However, a small number of participants realized that the altruist should be on the team. This is strongly supported by ten comments of participants which showed an understanding of a difference between individual and overall performance. For instance, Participant 111 wrote: "The ones who sell the most aren't necessarily the top performers. The sales always increased on days when the girl in the green shirt was working although […] she didn't sell as well herself. She seemed to bolster everyone else's sales, making total sales increase by nearly 50%. This was consistent every day she worked, therefore I rated her as extremely important to the company." These participants were *all* among those who selected the altruist in the last round of the personnel selection task (Figure 5) and 80% of them made this selection already in Round 2 (only 30% in Round 1).

General Discussion

In sum, participants in all conditions of both experiments tended to focus on comparing the individuals' earnings without considering their overall contribution to the

earnings of the group. Although Experiment 1 showed a small effect of conditions, the illicit individualist judgments remained dominant both with a high-impact altruist and with 20 (instead of 10) rounds. Experiment 2 showed that participants were well able to distinguish even much smaller *individual* performance differences. Experiment 2 nonetheless corroborated the finding that people tend to wrongly evaluate the group-serving altruist lowest, despite the 40 rounds with four repeated test-phases. Even the personnel-selection task, which might have focused participants more closely on the overall earnings in different configurations of the team, the best team player was most often excluded from the team. As a conclusion, a "tragedy of personnel selection" may occur when people who contribute clearly the most to the overall performance are evaluated the most negatively due to their lower individual contribution.

The outcome was not 'tragic' for *all* participants, since some, but few, participant's selections and comments revealed that they clearly had detected the positive overall effect of the 'altruist'.

Nonetheless, the tragic results in our task with an ecologically rather low or at least common complexity may be due to various explanations. First, the tragedy may be tied to the abstract form of evaluation based on numbers only. But number-based evaluations play an increasing role in companies. Second, the evaluation-context might have elicited a kind of competition or individualism framing, preventing participants from considering helpful behaviour. But we conducted a further egoist-detection personnel evaluation experiment, which led to only partial improvements. Third, the tragedy of personnel selection may well be based on more general and cognitive mechanisms. For instance, it may be related to (a) a general difficulty to realize that many small externalities can add up to large payoffs (e.g., von Sydow, 2015, Dörner, 1993); (b) some general problems of understanding interaction effects (e.g., Novick & Cheng, 2004); or (c) general problems dealing with multilevel representations and the Simpson's paradox (Fiedler, Walther, Freytag, & Nickel, 2003; Waldmann & Hagmayer, 2001; von Sydow, Hagmayer, & Meder, 2016). Whatever the causes are, the practical importance of the problem of detecting group-serving employees and building most efficient teams (based on more than individual excellence) can hardly be overestimated, for both companies and other organizations.

Our findings suggest that there may be a tragedy of personnel selection in the real world as well, with implications for the sensibleness of incentive-structures (personnel evaluation), employee-advancement (personnel promotion) and job offers (personnel selection).

Acknowledgments

This work was supported by the grant Sy111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program *New Frameworks of Rationality* (SPP 1516). We are grateful to Maria Waltmann for support, to Klaus Fielder, Ulrike Hahn, Sandra v. Sydow, and Michael Waldmann for discussing related issues, and to Martha Cunningham for corrections. The pictures in Fig. 1 are from S. ter Burg, A. Hjartat, B. van Kleef, and T. Arbeider-parti, (creative commons licence; www.flickr.com/photos/ and ter-burg/8172735997, apotekhjartat/13406119904, mvon-ederland/12364694653, tromsarbeiderparti/4577244545). We thank anonymous reviewers for helpful comments.

References

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? *Cognition, 31*, 187-276.

Dörner, D. (1993). *Die Logik des Mißlingens. Strategisches Denken in komplexen Situationen.* Rowolt: Hamburg.

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin, 29*, 14–27. doi: 10.1177/0146167202238368.

Hardin, G. (1968). The Tragedy of the Commons. *Science*, 162 (3859), 1243–1248.

Fehr, E. & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics, 114*, 817–868. doi: 10.1162/003355399556151.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. doi:10.1038/415137a.

Fehr, E. & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Science, 8*, 185–190.

Novick, L. R., & Cheng, P. W. (2004). Assessing Interactive Causal Influence. *Psychological Review, 111*, 455–485.

Nowak, M. A. & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437/27*, 1291-1296. doi: 10.1038/ nature04131.

Ostrom, E., Burger, J., Field, C. B., Norgaad, R. B., & Policansky, D. (1999). Revisiting the Commons: Local Lessons, Global Challenges. *Science, 284* (5412), 278–282.

Henrich, J. et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*, 795-855.

Sober & Wilson (1999). *Unto Others. The Evolution of Unselfish Behavior*. Harvard University Press.

von Sydow, M. (2006). *Towards a Flexible Bayesian and Deontic Logic of Testing Descriptive and Prescriptive Rules*. Dissertation, Georg-August-Universität Göttingen.

von Sydow, M. & Hagmayer, Y. (2006). *Deontic Logic and Deontic Goals in the Wason Selection Task.* Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society (pp. 864-869). Mahwah, NJ: Erlbaum.

von Sydow, M. (2012). *From Darwinian Metaphysics towards Understanding the Evolution of Evolutionary Mechanisms*. Göttingen: Universitätsverlag Göttingen.

von Sydow, M. (2015). The Tragedy of Inner-Individual Dilemmas. In D. Noelle, et al. (Eds.), *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society* (pp. 2517-2522). Austin, TX: Cognitive Science Society.

von Sydow, M., Hagmayer, Y., Meder, B. (2016, in press). Transitive Reasoning Distorts Induction in Causal Chains. *Memory & Cognition*, 1-19. doi: 10.3758/s13421-015-0568-5.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*, 27–58. doi: 10.1016/S0010-0277(01)00141-X.

Wilson, D. S.; Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology, 82*(4), 2007, 327–348. doi: 10.1086/522809.

Are we *ON* the same page?
Monolingual and bilingual acquisition of familiar and novel relational language

Nathan R. George (nrg10@psu.edu)
Department of Psychology, Penn State University
University Park, PA 16802 USA

Junko Kanero (jkanero@temple.edu)
Department of Psychology, Temple University
Philadelphia, PA 19122 USA

Dorothee Chwilla (d.chwilla@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University
Nijmegen, The Netherlands

Daniel J. Weiss (djw21@psu.edu)
Department of Psychology, Penn State University
University Park, PA 16802 USA

Abstract

Verbs and prepositions pose significant challenges in second language learning, as languages differ in how they map these relational terms onto events. Second language learners must put aside their language-specific lens to uncover how a new language operates, perhaps having to rediscover semantic distinctions typically ignored in the first language. The current study examines how the acquisition of these novel mappings are affected by characteristics of the learner and of the language to be learned. English monolinguals and Dutch-English bilinguals learned novel terms that corresponded to containment and support relations of either English, Dutch, or Japanese. Results show that English distinctions are learned best across groups, potentially reflecting predispositions in human cognition. No differences were found between monolinguals and bilinguals in any language condition. The characteristics of the language to be learned appear to play a prominent role in the acquisition of novel semantic categories.

Keywords: Cognitive Semantics; Second Language Learning; Bilingualism; Event Perception

Introduction

Verbs and prepositions are fundamental components of language, conveying dynamic and static relations between objects in events (e.g., "He *kicked* the ball *over* the fence"). Despite their centrality, these relational terms prove challenging for both first and second language learners (Gentner, 2006). Part of the challenge arises from the fact that languages differ in the aspects of events they emphasize. For example, while English utilizes *in* and *on* to denote containment and support relations, respectively, other languages vary in terms of the granularity or foci of these distinctions (Gentner & Bowerman, 2009). Dutch makes finer-grained divisions, breaking support into three distinct categories: *op* (i.e., resting on), *aan* (i.e., point-to-point attachment), and *om* (i.e., encirclement with contact). In contrast, Japanese verbs require attending to degree-of-fit relations, sometimes in conjunction with the *in/on* distinction, as in *oku* (i.e., loose-fitting on) and *ireru* (i.e., loose-fitting in), but sometimes collapsing across it, as in *hameru* (i.e., tight-fitting on or in; see Figure 1). How do learners come to discern these categories in language?

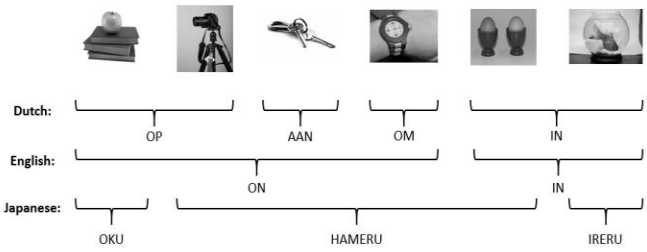


Figure 1: Containment and support in Dutch, English, and Japanese, inspired by Gentner and Bowerman (2009).

Over the first year and a half of life, infants learn to attend to a set of foundational components of events that support the structure of semantic categories across a wide array of languages (e.g., George, Göksun, Hirsh-Pasek, & Golinkoff, 2014; Göksun, Hirsh-Pasek, & Golinkoff, 2010; McDonough, Choi, & Mandler, 2003). With exposure to language, however, infants appear to focus on a subset of categories relevant to their native tongue. Language, in other words, has the function of orienting attention to some relations in events over others (George, et al., 2014; Göksun, et al., 2010). Heightening and dampening attention to early perceptual categories creates entrenched *lexicalization biases*, or strategies for word-to-world mapping. These biases are largely believed to influence event perception in children and adults, though there remains debate regarding the scope of these effects (e.g., Choi & Hattrup, 2012; George et al., 2014; Gleitman & Papafragou, 2013; McDonough et al., 2003; Papafragou, Hulbert, & Trueswell, 2008).

Bilingualism and Second Language Learning

The study of first language acquisition suggests that lexicalization biases affect the way in which adults process events for language. While beneficial for mapping language onto relations in events, lexicalization biases also potentially bear on challenges in second language learning. Learning a new language requires second language learners (SLLs) to not only acquire a new lexicon, but to identify how relational terms in the second language map onto events. Given that the overlap in the lexicalization patterns between languages is unknown, SLLs must be able to put aside their language-specific lens to uncover the way the new language operates, perhaps having to resurrect semantic distinctions typically ignored in the first language.

Research examining the endpoint of second language learning highlights plasticity in these biases. Hohenstein, Eisenberg, and Naigles (2006) found that even college students who did not begin second language instruction until after puberty can achieve native-like lexicalization biases in their second language. This process takes time, however. Song, Pulverman, Pepe, Golinkoff, and Hirsh-Pasek (2016) found that approximately seven semester-long courses are required to elicit lexicalization biases in a second language that do not differ from native speakers. Artificial language experiments suggest a higher degree of malleability. Havasi and Snedeker (2004) taught English-speaking adults nonsense verbs in English consistent with the Spanish bias to encode path in the main verb (e.g., crossing). At the beginning of the experiment, subjects preferred applying the novel verb to manner of motion, according with the English bias. As the experiment progressed, participants shifted towards a path interpretation for novel verbs, suggesting these biases can be changed with relatively short exposures.

While these biases are malleable, there remains a relatively impoverished understanding of the factors that contribute to the successful learning of novel semantic spaces. Most research in this area focuses on the added benefit of immersion over traditional classroom instruction (e.g., Song, et al., 2016). In the parallel field of phonetic learning, however, research suggests that learning features of a new language can be attributed to an amalgam of characteristics both of the learner and of the material to be learned (Antoniou, Liang, Ettlinger, & Wong, 2015). Here we investigate the impact of three features on the malleability of lexicalization biases: 1) the difficulty of the contrasts to be learned; 2) the degree of similarity between languages; and 3) bilingualism.

Ease of acquisition may be attributable in part to the characteristics of the language to be learned. Some lexicalization patterns may be easier to acquire, regardless of their relation to the learner's native tongue. Returning to work on development, Gentner and Bowerman (2009) note that the semantic categories underlying relational language emerge at different points across development. In their *Typological Prevalence Hypothesis*, they suggest that more cross-linguistically prevalent categories are more "natural" in perception and thus easier to learn. For instance, the English support category of *on* is more prevalent across languages than the Dutch category of *aan*, and correspondingly emerges earlier in English-speaking children's vocabulary than does *aan* in Dutch-speaking children's vocabulary (see also Beekhuizen, Fazly, & Stevenson, 2014). Extending this hypothesis to the study of second language learning, those categories that are more universal may also be those that are easier to rediscover regardless of the learner's native language.

The ease in acquiring a new language may also be proportional to the degree of similarity between that language and the language(s) known by the learner (e.g., Antoniou, et al., 2015). In the domain of containment and support, for example, we might expect that a monolingual English speaker would more easily acquire a novel language that shares the *in/on* distinction, when compared to a language such as Japanese, which sometimes requires collapsing across it. In this regard, we see a potential language-dependent advantage for bilingual speakers: using two languages with potentially disparate lexicalization patterns increases the likelihood that at least one of these will be reflected in any newly encountered language.

Finally, research on second language learning suggests that bilinguals acquire novel languages more efficiently than monolinguals overall (e.g., Kaushanskaya & Marian, 2009; Van Hell & Mahn, 1997). A common explanation regards bilingual advantages in several areas of cognition, including those associated with executive control (e.g., inhibition, working memory, etc.) among others (Bialystok, Craik, & Luk, 2012). These cognitive skills may lead to a more efficient restructuring of lexicalization biases. For example, inhibitory control may assist in the dampening of native biases when interpreting novel patterns.

The Current Study

The current study seeks to deepen our understanding of the factors underlying the successful learning of novel lexicalization patterns. We examine the ability of English monolinguals and Dutch-English bilinguals to learn novel terms that map onto either the English, Dutch, or Japanese semantic categories of containment and support. We make several hypotheses. First, if cognitive predispositions favor some lexicalization patterns over others irrespective of language experience, we would expect that both monolinguals and bilinguals would show the same patterns of learning, with the categories more central to cognition being learned more efficiently than those less prevalent. This pattern may be observed on the level of language (e.g., English more central than Japanese) or individual category (e.g., Dutch *in* more central than *aan*). Alternatively, if ease of acquisition is dependent upon the relation between the lexicalization patterns of language to be learned and those known by the learner, we would expect that monolinguals and bilinguals would perform equally well in acquiring the semantic categories of English (known by both) and equally poorly in acquiring the semantic categories of Japanese (known by neither), but that bilinguals would outperform

monolinguals in acquiring the semantic categories of Dutch (known only to bilinguals). We also expect variations of this pattern on the category level, reflecting differences in the degree of overlap between English and Dutch categories. For example, bilinguals should not outperform monolinguals in acquiring the Dutch category *in*, which is shared across English and Dutch. Finally, if there is a more general bilingual advantage in learning new lexicalization patterns, bilinguals should outperform monolinguals overall, even when learning Japanese, a language with which they have no experience.

Methods

Participants

One-hundred and six monolingual English-speaking adults were randomly assigned to one of three conditions of a word learning task: English mappings (N= 33, M = 18.76 yrs; 4 Male), Dutch mappings (N= 37, M = 19.69 yrs; 5 Male), or Japanese mappings (N= 36, M = 18.97 yrs; 3 Male). All participants' self-rated proficiency in a second language was five or less on a ten-point scale. An additional 41 participants were excluded from the current analyses due to bilingual status. Further, an additional six were excluded altogether for inattention (4) and technical error (2).

Fifty-four Dutch-English bilinguals living in the Netherlands were also randomly assigned to the English (M = 23.89 yrs; 6 Male), Dutch (M = 24.83 yrs; 7 Male), and Japanese (M = 22.89 yrs; 2 Male) conditions (18 in each). All bilingual participants' self-rated proficiency in English was six or higher on a ten-point scale. Two additional bilinguals were excluded for failure to understand the task.

Materials

For each condition, images were selected to represent four distinct semantic categories. In the English condition, these consisted of the two English containment/support categories (*on, in*) as well as two filler spatial relations (*beside, behind*) that are similar across the languages used in the experiment. In the Dutch condition, the semantic categories were made up of the four Dutch containment/support categories of *op* (resting on), *aan* (attachment), *om* (encirclement with contact), and *in*. In the Japanese condition, these categories corresponded to three Japanese containment/support categories of *oku* (loose-fitting on), *ireru* (loose-fitting in), and *hameru* (tight-fitting on/in) as well as one filler category that is similar across languages (*behind*). For each relation in each condition, 32 distinct images were chosen from the public domain (128 in total per condition). The assignment of images to their semantic categories was done in consultation with a native speaker for all conditions.

For each image, a recording was made that presented a novel word embedded within a syntactic frame. The use of syntactic frames allowed for the disambiguation of the meaning of the terms. For example, when viewing an apple resting on a pile of books, the recording might state, "The apple is *blick* the books." For each image, four sentences were recorded by a female native English speaker, each utilizing a different nonce word (*blick, frep, glorp, hirsh*).

A language history questionnaire (LHQ) assessed language proficiency (Li, Sepanski, & Zhao, 2006). The LHQ asks participants to self-rate their proficiency in each known language, among other aspects of language use.

Procedure

Participants were tested in a quiet room equipped with a computer and headphones. In all conditions, the experiment consisted of eight blocks. Each block consisted of eight training trials followed by eight test trials.

Training Each training trial presented an auditory sentence (e.g., "The ring is *blick* her finger") paired with a picture depicting the relation referenced. Within each block, eight trials were presented in succession, two for each of the condition's four categories. Each trial lasted five seconds and trials proceeded in a random order.

Test Each test trial presented a novel picture paired with a four-alternative multiple choice question. Questions required subjects to choose a word to fill in a sentence, mirroring the structure of the recorded sentences used during training with the blank corresponding to the position of the nonce word (see Figure 2). Answers consisted of the target word, plus the three remaining nonce words presented during the training phase. Within each block, eight trials were presented in succession, with two trials per category in the condition. Trials were untimed and randomized.



Figure 2: Example of a test trial.

Design No images were repeated during the experiment, ensuring responses to test trials were based on knowledge of the semantic category and not memory for word-image pairings. The assignment of images to training or test trials, as well as the mapping of nonce words to semantic categories were counterbalanced within each condition. Following the experiment, participants completed the LHQ.

Instructions were presented in the participant's native language; however, recordings and test sentences were presented in English for all participants.

Results

A series of ANOVAs were conducted. For all ANOVAs, the Greenhouse-Geisser correction was applied whenever sphericity was violated (Greenhouse & Geisser, 1959).

First, we conducted an ANOVA examining accuracy scores, with block as a within-subjects factor, and condition (English, Dutch, Japanese) and language background

(monolingual, bilingual) as between-subjects factors. Results showed a main effect of block (see Figure 3), reflecting improved mapping of words to referents across the experiment, $F(4.687, 745.710) = 150.839$, $p < .001$.

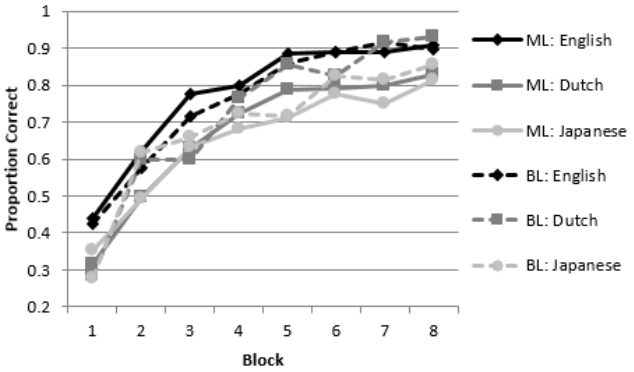


Figure 3: Learning by language background (monolingual vs. bilingual; ML and BL, respectively) and condition (English vs. Dutch vs. Japanese).

Further, there was a significant effect of condition, $F(2, 154) = 3.863$, $p < .05$. Bonferroni-corrected ($p < .017$) post-hoc contrasts suggested that English categories were learned significantly better than Japanese categories, $t(103) = 3.057$, $p < .017$. English categories were also learned better than Dutch categories, though this result was only marginally significant, $t(104) = 2.246$, $p = .027$. There was no difference between participants in mapping Dutch and Japanese categories, $t(107) = .899$, $p = .371$ (see Figure 4).

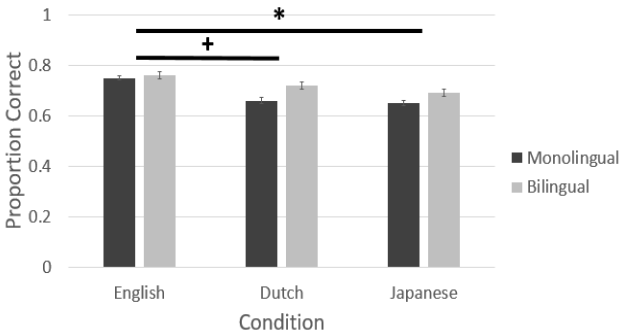


Figure 4: Proportion of correct responses by condition and language background. *$p < .017$ $^{+}p < .034$

There was no effect of language background, nor any interactions between block, condition, and language background, ps > .05. Thus, bilinguals did not show any advantage overall, nor within any language condition.

Because some categories are similar across languages (e.g., *in*), we next looked for differences on the level of categories, to determine whether certain semantic distinctions were driving the observed condition differences. For each condition, an ANOVA was conducted, with both category and block as within-subjects factors, and language background as a between-subjects factor (Figure 5).

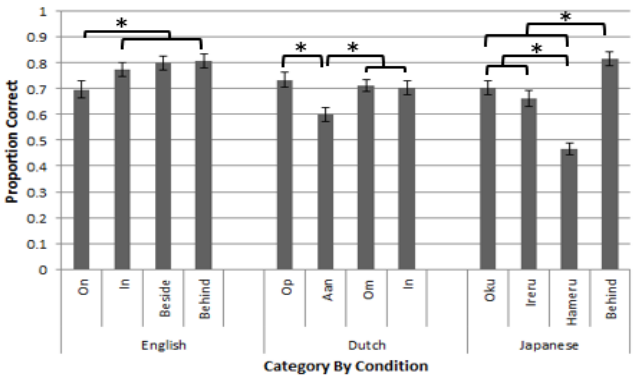


Figure 5: Proportion of correct responses broken down by condition and category. *$p < .008$

For the English condition, there was a significant main effect of block, $F(3.254, 162.750) = 45.535$, $p < .001$, and category, $F(3, 350) = 11.756$, $p < .001$. Bonferroni-corrected ($p < .008$) post-hoc contrasts were conducted to examine how learning differed among the categories. Participants performed worse on the category of *on* in comparison to *in*, $t(50) = 3.418$, $p < .01$, *beside*, $t(50) = 5.071$, $p < .001$, and *behind*, $t(50) = 5.360$, $p < .001$. Performance on all other categories was equivalent, ps > .008.

For the Dutch condition, there was a significant main effect of block, $F(4.839, 261.198) = 65.142$, $p < .001$, and category, $F(3, 378) = 10.475$, $p < .001$. Bonferroni-corrected ($p < .008$) post-hoc contrasts were conducted to examine how learning differed among the categories. Participants performed worse on the category of *aan* in comparison to *op*, $t(54) = 5.054$, $p < .001$, *om*, $t(54) = 5.180$, $p < .001$, and *in*, $t(54) = 4.365$, $p < .001$. Performance on all other categories was equivalent, ps > .008.

For the Japanese condition, there was a significant main effect of block, $F(4.804, 254.506) = 43.745$, $p < .001$, and category, $F(3, 371) = 49.386$, $p < .001$. Bonferroni-corrected ($p < .008$) post-hoc contrasts were conducted to examine how learning differed among the categories. Participants performed worse on *hameru* compared to *oku*, $t(53) = 8.370$, $p < .001$, *ireru*, $t(53) = 6.399$, $p < .001$, and *behind*, $t(53) = 12.983$, $p < .001$. Participants also performed worse on *oku* and *ireru* when compared to *behind*, $t(53) = 4.619$, $p < .001$ and $t(53) = 5.569$, $p < .001$, respectively. There was no difference between *oku* and *ireru*, $p > .008$.

Neither language background, nor any interactions between block, category, and language background were significant in any condition, ps > .05.

Discussion

Lexicalization biases present a challenge for learning a new language. To achieve native-like competency, learners must potentially put aside these longstanding biases to discover how a new language maps relational terms onto relations in events. Our results provide evidence that some lexicalization patterns may be easier to learn than others, a factor that has been found to impact other facets of second language acquisition (e.g., Antoniou, et al., 2015). Japanese

and to a lesser extent Dutch biases proved more difficult to learn than English biases. This finding is particularly striking given that both Dutch and Japanese conditions included categories native to English (e.g., *in* for Dutch, *behind* for Japanese) that may have inflated performance. Further, the bias for English lexicalization patterns is apparent on the level of individual categories. Categories most difficult to learn were those most typologically dissimilar from English. For instance, post-experiment debriefings suggest that monolinguals could rely on rough English equivalents to support learning *op* (prototypical *on*), and *om* (*around*) in Dutch, but that this strategy was less productive for *aan* (attached, hanging, on, etc.). Similarly, participants could simplify the categories of *oku* and *ireru* in Japanese to prototypical containment and support relations, whereas *hameru* had no clear English equivalent.

Interestingly, both monolinguals and bilinguals showed the same enhancement of learning English patterns over Dutch and Japanese, even though the native language of the bilingual group was Dutch. This pattern of results may support the notion that cognition is predisposed to attend to the categories of some languages more than others, an experience-independent effect that is immune to the language background of the learners tested in our study (though see below for an alternative explanation). As discussed by Gentner and Bowerman (2009), English containment and support categories are more universally represented across languages and appear early in children's vocabularies, reflecting their potential centrality to cognition. We mirror this pattern in the learning of additional languages. Thus, the superior learning of the English biases may be further reflection of the centrality of the *in/on* distinction, relative to the semantic categories of Dutch or Japanese. Our results are also in accord with recent work on action words, in which cognitive predispositions for encoding path information, but not adults' language backgrounds, predicted the ease with which they learned novel words (Emerson, Özçalişkan, & Frishkoff, 2016).

Bilingual Advantage?

While the results support experience-independent effects of the language to be learned, we find no evidence of additional experience-dependent effects. Dutch-English bilinguals did not outperform English monolingual speakers when learning the semantic categories of Dutch, a language with which monolingual participants had no prior experience. This result also holds on the level of individual categories, with Dutch-English bilinguals not outperforming English-speaking monolinguals on any Dutch category.

The lack of facilitative effect for Dutch-English bilinguals in learning Dutch lexicalization patterns is a departure from related research in phonology. Antoniou and colleagues (2015) found that English monolinguals, Mandarin-English bilinguals, and Korean-English bilinguals all learned novel words better when the words relied on Mandarin, as opposed to Korean phonetic contrasts, analogous to the cognitive predisposition for English categories here. Unlike the current results, however, there was also a language-dependent effect: Korean-English bilinguals outperformed the other groups on words that relied on Korean contrasts. Why might the results here be different? The lack of advantage may be due in part to the higher malleability of lexicalization biases. The phonetic space is drastically restructured over the course of development (Werker & Tees, 1984) and even advanced SLLs struggle to discern non-native phonetic contrasts (Pallier, Colomé, & Sebastián-Gallés, 2001). The heightening and dampening of lexicalization biases, however, is thought to be less dramatic, with non-native categories easily re-awakened in certain contexts (Choi & Hattrup, 2012).

We also found no evidence of a general bilingual advantage, again in contrast to research in other areas of second language learning (Antoniou, et al., 2015; Kaushanskaya & Marian, 2009; Van Hell & Mahn, 1997). Specifically, bilinguals were no better at mapping relational terms than monolinguals across both familiar and novel lexicalization patterns. Future research will examine performance relative to traditional measures of executive function to isolate whether these variables thought to underlie bilingual advantages in language learning are relevant to the learning of lexicalization patterns.

Limitations and Future Directions

When considering how language experience affects the learning of lexicalization patterns, there is another factor that must be considered: context. Despite written instructions in Dutch, the presentation of the experiment by an English-speaking experimenter and the use of English to frame nonce words may have biased bilinguals away from the lexicalization patterns of their native Dutch and towards those of their second language, English. Indeed, in research on motion conceptualization, Spanish-English bilinguals attend more to manner of motion when tested in English, which prominently marks this aspect of events, as opposed to Spanish, which does not (Kersten, et al., 2010). Future conditions will manipulate the language of presentation to determine whether the pattern of results in the current study is caused in part by a predominantly English context.

Further, while the patterns observed are consistent with a cognitive predisposition account, other variables warrant consideration. For instance, the pattern of performance may be attributable to a preference for English's semantically broader categories, which may not always be a feature of categories proposed to be cognitively central. In addition, the use of fillers, such as behind and beside, may have differentially helped participants in the English and Japanese conditions, as distinctions between these relations and those of containment and support may be more obvious than distinctions within containment and support relations.

Finally, a particularly promising avenue for future research is the use of this paradigm with native speakers of Japanese. The results suggest that Japanese is the most difficult of the three lexicalization patterns tested. If this is an experience-independent effect due to cognitive

predispositions, we would expect even Japanese speakers to show poorer learning of these categories relative to those of English and Dutch in an artificial language learning task.

Conclusion

Lexicalization biases provide a lens with which to process events for language. When learning a novel language, however, these biases must be put aside in favor of new ways of representing relations in events. The current study suggests that a primary influence on this process may be the extent to which cognitive predispositions align with the lexicalization patterns of a new language.

Acknowledgments

This research was supported by an NIH R01 grant to Daniel J. Weiss (HD067250) and an NSF PIRE grant (OISE-0968369) to Judith Kroll, Janet van Hell, and Giuli Dussias. We thank Carol Miller, Janet van Hell, Gerrit Jan Kootstra, Federica Bulgarelli, Sarah Miller, Johanna de Vos, Miriam Kos, Vanessa Engel, and Ronny Janssen for their invaluable contributions to this work. We also thank the Child Language and Cognition Lab and the Center for Language Science at Penn State University, as well as the Donders Institute for Brain, Cognition, and Behaviour at Radboud University for supporting this work.

References

Antoniou, M., Liang, E., Ettlinger, M., & Wong, P. C. (2015). The bilingual advantage in phonetic learning. *Bilingualism: Language and Cognition*, *18*(4), 683-695.

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning Meaning without Primitives. Typology Predicts Developmental Patterns. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240-250.

Choi, S., & Hattrup, K. (2012). Relative contribution of perception/cognition and language on spatial categorization. *Cognitive Science, 36,* 102-129.

Emerson, S. N., Özçalişkan, S., & Frishkoff, G. A. (2016). Effects of motion type and modality on word learning in English. *Applied Psycholinguistics*, *37*(3), 643-671.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544-564). New York, NY: Oxford University Press.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. In J. Guo, E. Lieven, S. Ervin-Tripp, N. Budwig, S. Özcaliskan, & K. Nakamura (Eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin* (pp. 465–480). New York, NY: Erlbaum.

George, N. R., Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Carving the world for language: How neuroscientific research can enrich the study of first and second language learning. *Developmental Neuropsychology*, *39*(4), 262-284.

Gleitman, L. R., & Papafragou, A. (2013). Relations between language and thought. In D. Reisberg (Ed.), *Handbook of Cognitive Psychology* (pp. 504-523). New York, NY: Oxford University Press.

Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2010). Trading spaces: Carving up events for learning language. *Perspectives on Psychological Science*, *5*, 33-42.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Havasi, C., & Snedeker, J. (2004). The adaptability of language-specific verb lexicalization biases. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hohenstein, J., Eisenberg, A., & Naigles, L. (2006). Is he floating across or crossing afloat? Cross-influence of L1 and L2 in Spanish-English bilingual adults. *Bilingualism: Language and Cognition, 9,* 249-261.

Kaushanskaya, M., & Marian, V. (2009). Bilingualism reduces native-language interference during novel-word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 829-835.

Kersten, A. W., Meissner, C. A., Lechuga, J., Schwartz, B. L., Albrechtsen, J. S., & Iglesias, A. (2010). English speakers attend more strongly than Spanish speakers to manner of motion when classifying novel objects and events. *Journal of Experimental Psychology: General*, *139*(4), 638.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaire: A Web-based interface for bilingual research. *Behavior Research Methods, 38*(2), 202-210.

McDonough, L., Choi, S., & Mandler, J. M. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology, 46,* 229–259.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science, 12,* 445–449.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*, 155-84.

Song, L., Pulverman, R., Pepe, C., Golinkoff, R. M., & Hirsh-Pasek, K. (2016). Does the owl fly out of the tree or does the owl exit the tree flying? How L2 learners overcome their L1 lexicalization biases. *Language Learning and Development, 12*(1), 42-59.

Van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning, 47,* 507–546.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7,* 49–63.

Dissociable effects of cue validity on bias formation and reversal

Angelo Pirrone (apirrone1@sheffield.ac.uk)
Department of Psychology and Department of Computer Science, Sheffield University, Sheffield, S102TN, United Kingdom
Department of Psychology, Peking University, Beijing, 100871, China

Qi Zhang (zhangqipsy@qq.com) and Sheng Li (sli@pku.edu.cn)
Department of Psychology and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, 100871, China.
PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, 100871, China.
Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, 100871, China.

Abstract

In two experiments we manipulated the prior probability of occurrence for two alternatives. After a first learning session, in a second session the cue to bias the decision was reversed. Our investigation shows that subjects are able to learn the reverse bias only when the bias of the first session is in line with their expected outcome. When, during the first session, the actual outcome of the bias is not in line with the expected outcome, there is an inhibition for the reversal bias learning in the second session. We investigate this phenomenon with computational models of choice showing that the inhibition of reversal is due to an increase in the rate at which subjects accumulate evidence for repeated, unexpected stimuli. We discuss a possible theoretical explanation that links this phenomenon to similar results found in the literature on reversal learning and to the effect of novelty on learning.

Keywords: bias; reversal learning; drift diffusion model; random dot kinematogram

Introduction

The Drift Diffusion Model (DDM, Ratcliff & McKoon, 2008) is a celebrated model of decision making that has been shown to fit data in a broad variety of - but not limited to - perceptual decision making tasks (Ratcliff & McKoon, 2008; Bogacz et al., 2006). In the DDM, the decision-maker integrates difference in evidence supporting two alternatives until one of two decision boundaries is reached and a decision is made in favour of that alternative.

The DDM is determined by seven parameters: (1) the boundary separation, *a*, which describes the distance between the two decision boundaries; (2) the drift rate, *v*, which refers to the rate at which noisy information is accumulated over time within a trial and reflects the difficulty of the task; (3) the starting point of evidence accumulation, *z*, that describes whether the decision-maker starts to integrate evidence near one of the two boundaries (when *z* is equidistant from the two boundaries, the decision process is *unbiased*); (4) a parameter that incorporates the non-decision time component of a reaction time, *ter*, which is the time to encode the stimulus and execute the motor response; (5) inter-trial variability in drift rate, *eta*, (6) inter-trial variability in the starting point, *sz* and (7) inter-trial variability in the non decision time component, *st*.

In a typical task known as the random-dot kinematogram (RDK), the decision-maker is asked to identify the direction of a coherent subset of dots on a computer screen; typically, a percentage of dots move randomly while the remaining dots move in a coherent direction (Ratcliff & McKoon, 2008). Mulder et al. (2012) used an RDK task and biased their participants toward answering 'left' or 'right' by either manipulating the prior likelihood of occurrence for one of the two alternatives or the reward associated with one of the alternatives. The likelihood manipulation consisted of an arrow pointing left or right that was presented before the occurrence of each trial. The reward associated with one of the two alternatives was manipulated by assigning different payoffs to the occurrence of the alternatives. Mulder et al. (2012) found that both manipulation of bias, either as prior probability or as potential payoff, can be mapped onto a change in starting point of evidence accumulation, while the rate at which the evidence is accumulated is not affected. In our investigation, we are interested on whether and how, once a perceptual bias is 'learned', it can be modified and which parameters of the decision process are associated with the modification of bias.

Experiment 1

We focused on bias reversal: the situation in which, after a bias is learned, subjects are exposed to the opposite bias. We are interested in subjects' behavioural performance and in which parameters are affected both in the first bias learning phase and in the bias reversal learning phase. In our first study, subjects were performing an RDK task similar to the one used in Mulder et al. (2012). For simplicity, we define 'congruent' the session where there is, most of the time, congruency between the arrow and the direction of the dots (as compared to the 'incongruent' session) and we define the single trials as 'valid' or 'invalid' as a function of the correspondence between the arrow and the direction of the dots. A group of subjects was presented with a 'congruent' session while a second group was presented with an 'incongruent' session. After a considerable number of trials during which subjects learned the bias, the bias was re-

versed. In this way, the group that in the first phase of the experiment was presented with a congruent bias was afterwards switched to an incongruent bias and vice versa. Our results show that only the group that switched from a congruent cue to an incongruent cue (for simplicity defined as the 'c to i' group) could learn to modify the bias while the second group that switched from an incongruent cue to a congruent cue (the 'i to c' group) did not modify the bias when it was reversed.

Subjects

Six healthy subjects (4 female, mean age = 23.3, SD = 0.6) performed an RDK task. For this and the following studies, subjects were recruited among Peking University students and had normal or corrected-to-normal vision. Their participation was voluntary and rewarded monetarily. The procedure was approved by the ethical review board at Peking University and informed consent was obtained from each subject.

Stimuli

The task consisted of an RDK task in which the prior probability of occurrence of one alternative was manipulated before the occurrence of each trial by presenting an arrow pointing left or right. Subjects were asked to fixate upon a cross on the centre of the screen, pay attention to the arrow that was presented before each RDK trial, and make their decision for each trial by pressing 'left' or 'right' on a keyboard. They were instructed to be as fast and accurate as possible in making a decision. The stimuli were similar to those used by Mulder et al. (2012): within a circle aperture of 5 deg, white dots with a size of 3x3 pixels were moving with a speed of 3 deg/s and a density of 16.7 $dots/deg^2/s$ on a black background. On each frame, noise dots followed a random, but constant direction. Signal dots had a limited lifetime of three frames after which they were redrawn in random locations. On each three consecutive frames (i.e. from 1 to 3, from 4 to 6 and so on) the dots constituting the signal were the same. The stimuli were generated on a personal computer using PsychoPy (Peirce, 2007) and presented on a 36 x 27 cm CRT screen with a refresh rate of 60 Hz at a viewing distance of 56 cm where the head of the subject was positioned on a chin rest. As done by Mulder et al. (2012), we matched the difficulty of the task across subjects. For this purpose, subjects performed a block of 200 trials of randomly presented stimuli with different coherence levels (respectively, 3, 10, 20, 30, or 40% coherence, 40 trials each) and by using the maximum-likelihood estimation procedure described in Palmer et al. (2005), for each subject, we computed the motion strength required for an expected accuracy level of 80%.

The cue to bias the decision was a white arrow presented on the centre of the screen, pointing to the left or to the right and indicating the correct response 80% of the times. One block consisted of 80 trials and after each block, subjects could take a self-paced break before continuing to the next trials. Subjects received 20 ¥ for their participation and were told that they could earn up to 50 extra ¥ if they would have reached 100% accuracy during the task. No penalty was introduced for a wrong answer. After each trial they were given a feedback ('correct' or 'wrong') and at the end of each block they were shown their average accuracy for the block. Subjects were also presented with neutral trials, in which instead of the arrow, they were presented only a rectangular shape equal to the rectangle constituting the body of the arrows of the biased trials.

Paradigm timing

Subjects performed 4 consecutive blocks of each biased session (i.e. the congruent or incongruent) and 4 consecutive blocks of neutral trials of an RDK task. At the beginning of each trial, subjects were presented a fixation cross for either 500 or 1500 ms after which the cue was presented for 1000 ms, followed by a fixation cross with a duration of 1000, 1500, 2000, or 2500 ms. Next, the RDK was presented for 1500 ms during which subjects had to make a decision by button press. After each trial, subjects were presented a feedback for 500 ms showing the number of points earned and whether the previous trial was correct or wrong. If subjects answered faster than 150 ms or slower than 1500 ms, they were presented with the words 'miss' (in Chinese) or 'too fast' (in Chinese) for 500 ms. Session-order was counterbalanced across subjects with a group performing a congruent-to-incongruent-cue manipulation and a second group performing an incongruent-to-congruent-cue experiment. The order of presentation of the neutral blocks was counterbalanced across subjects.

Behavioural analyses

A first inspection of the data showed that there was no cross-session learning and the performance of the subjects after the first block of each session was stable. For simplicity, in all the analyses we collapsed the data across blocks and we created five experimental conditions: (1) congruent session valid trials; (2) congruent session invalid trials; (3) incongruent session valid trials; (4) incongruent session invalid trials and (5) neutral trials.

To investigate the effects of our manipulation, we entered correct RTs, wrong RTs and accuracy levels, separately for each group, in three different 2 x 2 mixed-effect models with Session (congruent, incongruent), Validity (valid trials, invalid trials) and the interaction between Session and Validity as fixed factors and random effects for subject-specific constants and slopes, Figure 1. On

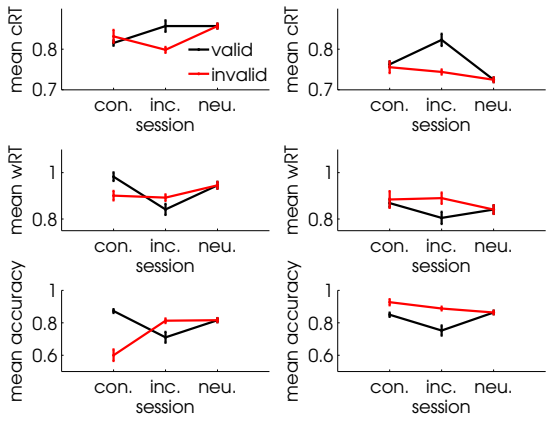


Figure 1: Mean correct RTs (first row), wrong RTs (second row) and accuracy levels (third row) for the 'c to i' group (left) and the 'i to c' group (right). Bars indicate standard errors of the group mean.

correct RTs, for the 'c to i' group, the interaction between Session and Validity resulted almost significant [F (1, 15.10) = 3.45, p = .08]. On correct RTs, for the 'i to c' group, no significant effect was found (p>.43). On wrong RTs, for the 'c to i' group, the main effect of Session resulted significant [F (1, 6.55) = 8.63, p = .02] and the interaction effect between Session and Validity was also significant [F (1, 26.13) = 7.56, p = .01]. Post-hoc tests with Bonferroni correction showed that, in the congruent session, subjects had faster errors for the invalid trials (M = .90, S.E.= .03) than for the valid trials (M = .98 S.E. = .02; p = .03). On wrong RTs, for the 'i to c' group, no significant effect was found (p>.46). On accuracy levels, for the 'c to i' group, the interaction between Session and Validity resulted significant [F(1,11) = 15.49, p =.002]. In the congruent session, subjects were less accurate for the invalid trials (M = .61, S.E. = .10), than for the valid trials (M = .87, S.E. = .05; p = .04). On accuracy levels, for the 'i to c' group, no significant effect was found (p>.30).

Model fitting

We fitted the diffusion model to RT distributions and proportions of correct and wrong responses using the Diffusion Model Analysis Toolbox (Vandekerckhove & Tuerlinckx, 2007) for MATLAB in which parameters are estimated by minimising a chi-square function. We decided to represent the reaction time distributions of correct and error responses in terms of five deciles, the 1st, 3rd, 5th, 7th and 9th decile of the RT distribution.

We fitted each subject separately, and for each subject we allowed the drift rate to vary freely across conditions and the starting point of evidence accumulation to vary as a function of the session (congruent vs. incongruent).

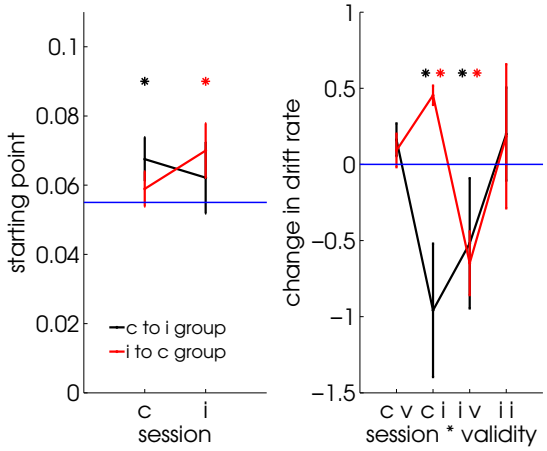


Figure 2: Effects of bias in DDM parameters. Average effects across subjects for starting point (left) and for change in drift rate with regards to the drift rate of the neutral trials (right). The blue horizontal line represents the unbiased level *a/2* (left) and the level at which the drift is zero (right). Asterisks show whether Wilcoxon Signed Rank tests resulted significant. X axis refers to (c v) congruent session valid trials (c i) congruent session invalid trials (i v) incongruent session valid trials and (i i) incongruent session invalid.

Figure 2 shows how the different parameters of the DDM were affected during the task; in particular we show how the starting point varied across sessions and what is the proportional change in drift rate as compared to the drift rate of the neutral trials.

Wilcoxon Signed Rank tests were performed to assess whether the starting point of evidence accumulation differed from the unbiased level (H0: median value of starting point = a/2, $p < 0.05$) and whether the proportional change in drift rate (with regards to the neutral trials) differed from zero (H0: median value of $\Delta v/v_{unbiased} = 0$, $p < 0.05$). Results are reported in Figure 2.

group	*a*	*ter*	*eta*	*sz*	*st*
c to i	0.11	0.63	0.15	0.04	0.30
i to c	0.11	0.63	0.32	0.06	0.32

Table 1: Parameter's values for the 'c to i' and the 'i to c' group.

In Table 1 are reported the mean values of the boundary separation, *ter*, *eta*, *sz* and *st* for the two groups; these parameters, being practically identical between the two groups, did not differ between groups ($p > .11$) as shown by mixed models with Group as fixed factor and random effects for subject-specific constants.

Experiment 2

In a second study, we tested the hypothesis that subjects in the 'i to c' group could not reverse the bias since they had to firstly learn a rule not in line with their expected outcome (i.e. that the arrow would point in the future direction of the dots) and this prevents them from overwriting the new rule when the bias is reversed.

We tested this hypothesis by manipulating the expectations of the subjects: all subjects performed an incongruent-to-congruent experiment as the 'i to c' group of the first study. However, before starting the experiment, one experimental group performed a training block in which there was always congruency between the arrow and the direction of the dots (congruent training group, for simplicity 'c training' group), and a second group performed a training block in which there was always incongruency between the arrow and the direction of the dots (incongruent training group,'i training' group). Our prediction was that only the group that had an incongruent training and for which the following incongruent session of trials was in line with their expected outcome (i.e. that the cue gives information about the *opposite* direction of the dots as experienced during the training) could modify their bias, as done by the 'c to i' group of the first experiment. We expected that the group that had the congruent training would instead exhibit a performance similar to that of the 'i to c' group of the first experiment, since they would not be able to reverse their bias, having learned a first bias that was in contrast with the expectations from the habituation manipulation. In this experiment, no neutral trials were presented.

Materials and Methods

Seven healthy subjects (4 female, mean age = 22.29, SD = 1.38) performed the second experiment. The apparatus and methods are as described for the first experiment. In this experiment, all subjects performed a congruent session after having performed the incongruent session. We manipulated the expected outcome of the first session by use of habituation. At the beginning of the experiment, subjects were presented two blocks of 80 trials and they were instructed that these trials were training trials to familiarise with the task. In this training phase, the arrow was indicating the valid or invalid direction of the trials 100% of the times. We created two groups: 4 subjects always had an invalid training while 3 subjects always

Behavioural analyses

performed a valid training.

We entered correct RTs, wrong RTs and accuracy levels, separately for each group, in three different 2 x 2 mixed-effect models with Session (incongruent, congruent), Validity (valid, invalid) and the interaction of Session and Validity as fixed factors and random effects for

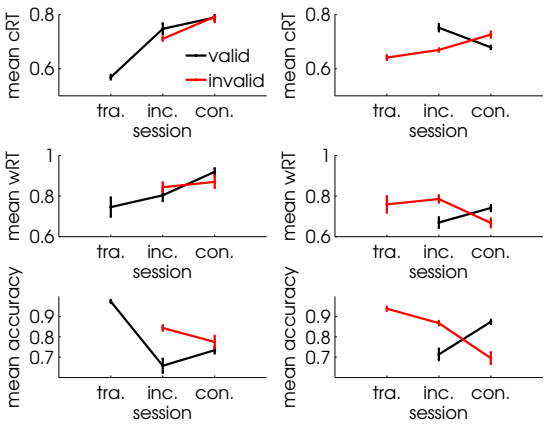


Figure 3: Mean correct RTs (first row), wrong RTs (second row) and accuracy levels (third row) for the 'c training' group (first column) and the 'i training' group (second column).

subject-specific constants and slopes, Figure 3. On correct RTs, for the 'c training' group, no significant effect was found ($p>.49$). On correct RTs, for the 'i training' group, the interaction between Session and Validity resulted significant [$F(1, 13.43) = 7.14$, $p = .02$]. In the incongruent session subjects were faster for the invalid ($M = .67$, S.E. = .05) than for the valid trials ($M = .76$, S.E. = .05; $p=.09$). On wrong RTs, for the 'c training' group, no significant effect was found ($p>.49$). On wrong RTs, for the 'i training' group, the interaction between Session and Validity resulted significant [$F(1, 27.03) = 7.6$, $p = .01$]. In the incongruent session, subjects had faster errors for the valid trials ($M = .67$, S.E. = .05) than for the invalid trials ($M = .79$, S.E. = .05; $p=.03$). On accuracy levels, for the 'c training' group, no significant effect was found ($p>.11$). On accuracy levels, for the 'i training' group, the interaction between Session and Validity resulted significant [$F(1,12) = 8.03$, $p =.02$]. In the congruent session subjects were less accurate for the invalid trials ($M = .75$, S.E. = .09), than for the valid trials ($M = .90$, S.E. = .04; $p = .03$).

Model fitting

We recovered DDM parameters for each subjects as done in the first study. Figure 4 shows how the different parameters of the DDM are affected during the task, both for the 'c training' and the 'i training' group. Wilcoxon Signed Rank tests were performed to assess whether the starting point of evidence accumulation differed from the unbiased level and whether the proportional change in drift rate (with regards to the training trials) differed from zero. Results are reported in Figure 2. The mean values of the other parameters of the DDM, that did not differ

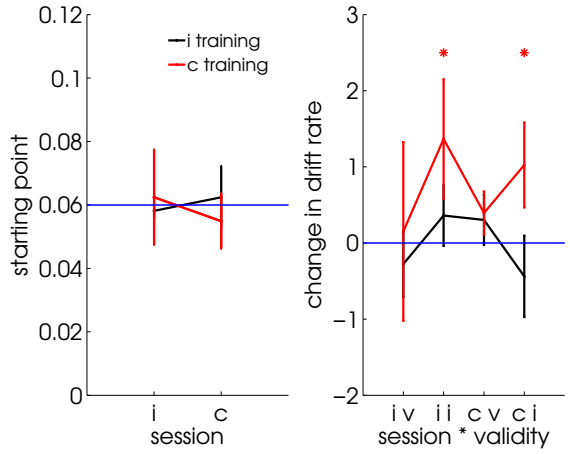


Figure 4: Effects of bias in DDM parameters. Average effects across subjects for starting point (left) and for drift rate as compared to the drift rate of the training trials (right).

between groups ($p > .79$), are reported in Table 2.

group	*a*	*ter*	*eta*	*sz*	*st*
i training	0.12	0.49	0.2	0.03	0.27
c training	0.12	0.53	0.14	0.02	0.31

Table 2: Parameter's values for the 'c training' and the 'i training' group.

Discussion

Our results show that when subjects learn a bias that is in line with their expectations, they are able to modify the bias in a second reversal session, but when they learn a bias that is not in line with their expectations, they are not able to modify the bias in a second reversal session. Only the group that firstly learned a bias in line with their expectations, either naive expectations (study 1) or controlled expectations (study 2), could become faster and more accurate for the trials for which they were slower and less accurate in the first session. A third investigation that we do not report in detail in this paper, showed that, for the 'i to c' group of the first study, even though the number of trials of the reversal session is doubled, subjects (n = 5) still cannot learn to modify the bias. In this way we rule out the hypothesis that what prevented subjects from learning the reversal bias was the length of the session that could be not long enough to give time to learn the reversal bias.

Similar results can be found in the literature on reversal learning showing inhibition of reversal dependant on the initial conditions of learning (Lindenblatt & Delius, 1988; Pace et al., 1980; Richman & Coussens, 1970; Reid, 1953). For example, Newman et al. (1980) have found that, for human subjects, learning a discrimination based upon the presence of a feature is easier than learning a discrimination based upon the absence of a feature. However, when there is a reversal, so that subjects who had learned a rule based upon the presence (absence) of a feature are then switched to a new session in which they have to learn the new rule based upon the absence (presence) of the rule, only subjects in the presence-to-absence group can master the reversal while subjects in the absence-to-presence group cannot master the reversal. This same result is replicated in multiple experiments on pigeons, monkeys, rats and humans by different methods and typology of stimuli for the discrimination (Lindenblatt & Delius, 1988; Pace et al., 1980; Richman & Coussens, 1970; Reid, 1953; Newman et al., 1980). In these studies, learning discrimination to a low criterion inhibits the acquisition of the reversal rule, while learning it to a high criterion does not interfere with learning the reversal, but instead facilitates the reversal in some cases (Lindenblatt & Delius, 1988; Reid, 1953).

We believe that our results can be due to the fact that different cue-stimulus correlations have, depending on their sign, different attention-enhancing effects in the first session of the experiment that are maintained throughout the task. When this correlation is positive, the attention-enhancing effect is low since the rule is easily understood and the decision is stimulus-driven in a bottom-up fashion allowing the possibility of overwriting the rule in the second session. However, when the cue-stimulus correlation is negative, the attention-enhancing effect is high and the decision is rule-driven in a top-down fashion; this leads to a stronger learning of the difficult rule, that as a consequence, inhibits overwriting this rule with the new reversal rule. The attention-enhancing effect represents a cuing advantage for unexpected, repeated stimuli in order to learn the rule and maximise the reward during the first session. The decision-maker, right after the presentation of the cue, pays attention to the direction predicted by the unexpected stimulus (i.e. in the first study, in the opposite direction of the arrow) and this means that more information can be obtained from such directions, resulting in faster and more accurate decisions when the prediction is matched compared to when it is not matched. However, during the reversal, the decision-maker is still more biased towards the opposite direction and hence the previous advantage becomes a disadvantage. The attention-enhancing effect elicited by a first cue not in line with the expectations could be revealed by the higher drift rates that subjects of the 'i to c' group of the first study and subjects of the 'c training' group of the second study show for the invalid trials of the second session. The rate at which information is accumulated to make a decision

can be modulated by attention, and an attentional gain resulting in a signal-to-noise increase (i.e. faster RTs and higher accuracy levels) has been described in perceptual decision making (Smith et al., 2004). This interpretation is in line with the behavioural results of the first and second study. Also, our explanation is in line with experimental evidence showing that the magnitude of the difference between the expected outcome and the actual outcome results in an increase in dopamine levels that is associated with attentional orienting (Daw et al., 2006).

In the literature, two mechanisms have been proposed for the effect of bias: one affects the starting point of evidence accumulation (Mulder et al., 2012; Diederich & Busemeyer, 2006), while the second affects the rate at which evidence is accumulated (Bogacz et al., 2006; Diederich & Busemeyer, 2006). An effect of bias on the starting point is in line with that found by Mulder et al. (2012). However, together with the variation in starting point, we show that the drift rate can also be affected by a bias.

It is not surprising that subjects find pointing in the direction of the arrow as the expected outcome. Behne et al. (2012) have shown that even 12-month old infants show an understanding of communicative pointing, and Butterworth (2003) has proposed that pointing gestures spontaneously emerge from other developmental achievements playing a pivotal role during development. Pointing in the opposite direction of a stimulus of interest is rather an experimental artefact than a behaviour naturally observed.

Future studies should investigate whether the theoretical explanation proposed is consistent across different sets of stimuli and tasks, and answer the broader general question of *why*, from an ecological perspective, the inhibition of reversal in various tasks and domains characterise human and animal cognition.

References

Behne, T., Liszkowski, U., Carpenter, M., & Tomasello, M. (2012). Twelve-month-olds comprehension and production of pointing. *British Journal of Developmental Psychology*, *30*(3), 359–375.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Butterworth, G. (2003). Pointing is the royal road to language for babies. *Pointing: Where language, culture, and cognition meet*, 9–33.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics*, *68*(2), 194–207.

Lindenblatt, U., & Delius, J. D. (1988). Preventing a feature-positive effect in pigeons. *The American journal of psychology*, 193–206.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience*, *32*(7), 2335–2343.

Newman, J. P., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 630.

Pace, G. M., McCoy, D., & Nallan, G. B. (1980). Feature-positive and feature-negative learning in the rhesus monkey and pigeon. *The American journal of psychology*, 409–427.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision*, *5*(5), 1.

Peirce, J. W. (2007). Psychopypsychophysics software in python. *Journal of neuroscience methods*, *162*(1), 8–13.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Reid, L. S. (1953). The development of noncontinuity behavior through continuity learning. *Journal of experimental psychology*, *46*(2), 107.

Richman, C. L., & Coussens, W. (1970). Undertraining reversal effect in rats. *Journal of Experimental Psychology*, *86*(2), 340.

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision research*, *44*(12), 1297–1320.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Acknowledgements

We thank Tom Stafford and Daniel A. Wilson for helpful comments on the analyses and manuscript. This work was supported by the National Natural Science Foundation of China (31230029, 31271081, 31470974) and by the Santander Research Mobility Award .

Effects of experience in a developmental model of reading

Ya-Ning Chang (y.chang5@lancaster.ac.uk)
Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK
Stephen Welbourne (stephen.welbourne@manchester.ac.uk)
Neuroscience and Aphasia Research Unit (NARU), University of Manchester, Manchester, M13 9PL, UK

Abstract

There is considerable evidence showing that age of acquisition (AoA) is an important factor influencing lexical processing. Early-learned words tend to be processed more quickly compared to later-learned words. The effect could be due to the gradual reduction in plasticity as more words are learned. Alternatively, it could originate from differences within semantic representations. We implemented the triangle model of reading including orthographic, phonological and semantic processing layers, and trained it according to experience of a language learner to explore the AoA effects in both naming and lexical decision. Regression analyses on the model's performance showed that AoA was a reliable predictor of naming and lexical decision performance, and the effect size was larger for lexical decision than for naming. The modelling results demonstrate that AoA operates differentially on concrete and abstract words, indicating that both the mapping and the representation accounts of AoA were contributing to the model's performance.

Keywords: age of acquisition; language development; reading; computational modelling; visual word recognition.

Introduction

Age of acquisition (AoA) effects refer to observations that stimuli learned early in life are processed more quickly and accurately than stimuli learned later in life. These AoA effects have been observed in a variety of language tasks including word naming, lexical decision, picture naming and semantic related tasks (Brysbaert, Van Wijnendaele, De Deyne, 2000; Cortese & Khanna, 2007; Ghyselinck, Lewis, & Brysbaert, 2004; Monaghan & Ellis, 2002). When the magnitude of AoA across tasks was compared in a review by Juhasz (2005), the results showed that the effect size is largest in picture naming (125 ms), followed by lexical decision (56 ms) and naming (31 ms). These findings indicate that AoA is a strong factor influencing lexical processing across several domains.

However, there has been scepticism about AoA effects because AoA is naturally confounded with other lexical semantic variables such as frequency and concreteness (or imageability) (Strain, Patterson, & Seidenberg, 2002; Zevin & Seidenberg, 2004). Nevertheless, when all these variables were considered in predicting large naming and lexical decision datasets, Cortese and Khanna (2007) showed unique AoA effects for both naming and lexical decision, indicating the AoA effects are not spurious findings.

Theories of AoA Effects

Several hypotheses have been proposed to explain the origin of AoA effects (Brysbaert et al., 2000; Ellis & Lambon Ralph, 2000; Lewis, Gerhand, & Ellis, 2001; Steyvers & Tenenbaum, 2005). One interpretation of the AoA effect is that early learned words have been encountered more times at the age that participants are tested than later learned words. This has been termed *the cumulative frequency hypothesis* (Lewis, 2001). On this view, cumulative frequency can be considered as a combined index by multiplying frequency and number of years that a stimulus is known to participants (i.e., age - AoA). However, most studies report additive effects of frequency and AoA (Ghyselinck, Lewis, & Brysbaert, 2004), suggesting the effects are distinct. In addition, the findings of the differential effect sizes of frequency and AoA in multi-task comparison studies (Ghyselinck, Lewis, & Brysbaert, 2004; Brysbaert & Ghyselinck, 2006) have been taken as evidence against the cumulative frequency hypothesis, because this theory would predict co-variance of frequency and AoA effects.

Another theory accounting for the AoA effect is *the representation mapping theory* (Ellis & Lambon Ralph, 2000). According to this computationally motivated account, the AoA effect is due to the gradual reduction in plasticity as more words are learned. Early learned words are privileged to easily adjust weight connections in the system; while later learned words can only cause small weight changes because of the reduced plasticity. Thus, an AoA effect is expected particularly when the mappings between inputs and outputs are arbitrary, because they require greater computational resources to resolve the mapping (Zevin & Seidenberg, 2002). This is also supported by behavioural data reported by Monaghan and Ellis (2002) where the AoA effect was stronger for low consistency words (e.g. break) than for high consistency words (e.g. block) in a word naming task.

The AoA effect also has been suggested to result from differences in semantic representations where early learned words have richer semantic representations than later learned words, termed *the semantic locus theory* (Brysbaert et al., 2000). Steyvers and Tenenbaum (2005) developed a semantic growth network to simulate the AoA effects in terms of the connections of words with others. In their network, early learned words have more connections with others and thus they have a more central role in the system, resulting in a faster access. The most direct evidence for the semantic locus theory comes from the observations of larger AoA effects in tasks that directly involve semantics, such as word-associate generation, picture naming, picture matching, and semantic categorisation (Brysbaert & Ghyselinck, 2006; Brysbaert et al. 2000; Catling & Johnston, 2009). In

addition, the magnitude of AoA effects can be related to the extent of involvement of semantics in the tasks (Ghyselinck, Lewis, & Brysbaert, 2004; Brysbaert & Ghyselinck, 2006).

These arguments have contributed to an emerging view that both the representation and semantic locus theories might contribute to the AoA effects (Catling & Johnston, 2009). For instance, the stronger AoA effects observed for low consistent words in naming might be explained by the semantic locus theory if one considers that semantics is differentially involved in naming according to the regularity of the orthography to phonology mapping (Strain & Seidenberg, 1995). On the other hand, the magnitude of AoA effects in different tasks is also compatible with the arbitrariness of the mappings between different representations, such that more arbitrary mappings elicit greater AoA effects. However, it remains unclear the extent to which AoA effects in a model of reading are required to be explained in terms of semantic locus or mapping effects.

AoA Models of Reading

Several computational models of reading have been developed to explore AoA effects (Ellis & Lambon Ralph, 2000; Monaghan & Ellis, 2010; Zevin & Seidenberg, 2002). Monaghan and Ellis (2010) developed a connectionist model that demonstrated clear AoA effects in naming in addition to cumulative frequency effects. The key for the model to capturing the AoA effects was that it was trained with a cumulative learning process. The model started to learn to read a small set of words, akin to a child beginning to learn to read, and gradually learned to build up an entire adult vocabulary. The process mimics the natural reading development that allows the model to capture the characteristics of AoA. Their findings provided evidence for the representation mapping theory. However, these models did not include semantic representations so they were limited in their ability to test the effect of the role of semantics in the size of AoA effects.

The primary aim of this study was to develop a large-scale developmental model of reading, trained cumulatively to simulate chronological language experience. The model comprised three key processing layers including orthography, phonology and semantics, and it was trained with a cumulative learning process to simulate different stages of reading development. We used the model to explore the AoA effects in both naming and lexical decision. In particular we attempted to examine the competing theories of AoA and investigate how semantic representations might implicate the emergence of AoA effects within the model.

Method

Network Architecture

The architecture of the model is shown in Figure 1. The model was based on the triangle model of reading previously implemented by Harm and Seidenberg (2004). The current model consisted of three processing layers including orthographic, phonological and semantic layers, one context layer, two attractor layers and five hidden layers for intermediation between the layers.

An attractor layer, which contained 50 units, was connected to and from the phonological layers. Similarly, there was a set of 50 attractor units for the semantic layer. The use of attractors was to help the model to reduce noise and develop stable phonological and semantic representations of words. There were also four context units connecting to the semantic layer via a set of ten hidden units. The context units provided additional information when presenting the model with homophones. One context unit was active for each homophone. But for words within the same homophone family, different context units were randomly assigned. In this way, each context unit was almost equally active across the training corpus. For non homophones, none of the context units were active. The semantic layer was connected to the phonological layer via a set of 300 hidden units, and the phonological layer was connected back to the semantic layer via another set of 300 hidden units. The orthographic layer was connected to both the phonological and semantic layers via different sets of 500 hidden units

Representations

The orthographic, phonological and semantic representations were similar to those used in Harm and Seidenberg's (2004) model. The training corpus contained 6229 monosyllabic words, which covered most monosyllabic words, including their inflected forms, in English. Frequency of each word was derived from the Wall Street Journal corpus (Marcus, Santorini, & Marcinkiewicz, 1993), and the score was log-transformed. For orthography, each word was represented by 14 letter slots and each slot comprised 26 units with one for each 26 alphabetic letters. Words were positioned with their first vowel aligned on the fifth slot. For words having two vowels, the second vowel was placed on the sixth slot; otherwise all the units in that slot were not active. Consonants preceding or following the vowel(s) were positioned in adjacent slots to the vowel(s) (so *yes* was represented as _ _ _ y e _ s _ _ _ _ _ _ _, and *great* as _ _ _ g r e a t _ _ _ _ _ _). For phonology, each word was represented by eight phoneme slots, with each slot consisting of a set of 25 phonological features. Each word was positioned with its vowel at the fourth phoneme slot. The first three slots were for onset consonants and the last four slots were for coda consonants (so *yes* was _ _ y E s _ _ _ and *great* was _ g r eI t _ _ _). The method of representing semantic knowledge for each word was adopted from that used in Harm and Seidenberg (2004). The semantic representation for each word consisted of 2446 semantic features, derived from WordNet (Miller, 1990). The presence of semantic features was encoded as 1 and the absence of semantic features was encoded as 0.

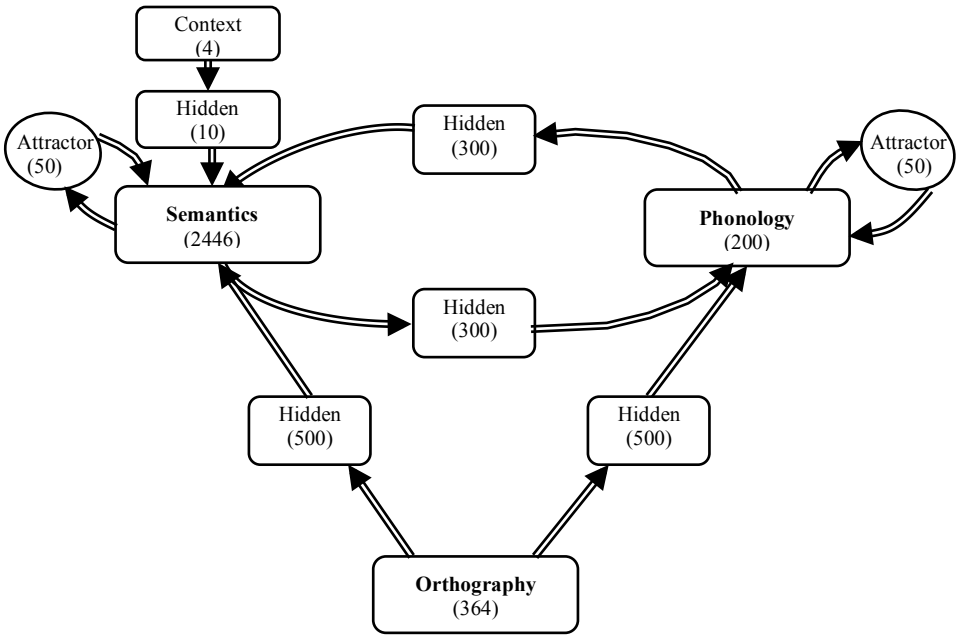


Figure 1. The architecture of the developmental model of reading.

Training Procedures

The training process had two phases. In pretraining, the model was trained with the mappings between phonology and semantics. This phase of training was an attempt to simulate the fact that children generally have developed some language skills (e.g. speaking and comprehension) before learning to read. In the reading development phase, the full reading model was trained.

In pretraining, the model was trained on both a speaking task (mapping from semantic to phonological representations) and a hearing task (mapping from phonological to semantic representations). The model also learned to develop a stable phonological attractor (mapping from phonological to phonological representations), and a stable semantic attractor (mapping from semantic to semantic representations). For both the speaking and hearing tasks, the input pattern of each word was clamped and presented for eight time steps, and in the last two time steps, the model was required to reproduce the target pattern of the word. Similarly, for both the phonological and semantic attractor training trials, the input pattern of each word was clamped for the first time step and in the last two time steps, the model had to reproduce the target pattern of the word. The input of context units was supplied only for the hearing task. During training, the four tasks were interleaved with 40% of trials for the speaking task, 40% of trails for the hearing task, 10% of trials for the phonological attractor and the remaining 10% for the semantic attractor.

During this stage of training, the model was trained on 2973 monosyllabic words, which were the most common words occurring in reading materials before age 18. Note that though several words in this set were unlikely to occur often in young children's language exposure, yet due to the training by frequency these words were rare during pretraining. The probability of a word being selected for training was determined by its logarithmic frequency. The model was trained with a learning rate of 0.05 using back-propagation through time algorithm. Error score was based on the cross-entropy error computed between the target and the actual activation of the output units. No error was recorded if the output unit's activation was within 0.1 of the target.

In the reading development phase, the model was trained on the reading task, which was to learn the mappings from orthography to both semantics and phonology, along with the four tasks in the pretraining phase. Following Monaghan and Ellis (2010), the model was trained to read cumulatively, to reflect 14 reading stages, one for each year. The reading stage was based on the educator's word frequency guide (WFG) by Zeno et al. (1995). The words in WFG were graded into 13 different grade levels by using readability measures, corresponding to the age range from 5-18 in the American and British schooling systems and the words appeared in adulthood were presented at the 14th stage. The model started to learn a small set of words and gradually more and more words were learned over time course of learning. The details of the staged training paradigm can be found in Monaghan and Ellis (2010) Table 1. For the reading task, the orthographic representation of a word along with its context layer representation were clamped and presented for 12 time steps, and for time steps six to 12, the model was required to produce the phonological and semantic representations for that word. All the five tasks were interleaved during training, but the training ratio for each task except the attractor tasks varied as the training proceeded. The training ratios for both the hearing task and speaking task gradually decreased from 40% to 20% in steps of 5%, while the training ratio for the reading task gradually increased from 10% to 50% in steps

of 10% to simulate greater exposure to reading versus listening and speaking with development. All the other training procedures remained the same as in pretraining.

Testing Procedures

After pretraining, the model was tested on both the speaking and hearing tasks. For the speaking task, the semantic representation of each word in the training set was presented and the activation of units at the phonological layer at the end of the eight time steps was recorded. Error score was measured by the sum of the squared differences between the activation of each input unit and its target activation. The accuracy of the model's phonological production was assessed by deciding whether for each phoneme slot the closest phoneme to the model's actual production was the same as the target phoneme. For the hearing task, the phonological representation of each word was presented and the activation of units at the semantic layer at the end of the eight time steps was recorded. Error score was measured by the sum of squared differences over the semantic layer. The semantic accuracy was measured by computing the Euclidean distance between the model's actual semantic representation and the semantic representation of each word in the training corpus. If the smallest distance was for the target representation then the model was correct.

After the reading training, the model's reading performance was tested. The orthographic representation of each word was presented and the activation of units at both the semantic layer and the phonological layer at the end of the 12 time steps were recorded. The measurement of error score and accuracy for both semantic and phonological output were the same as in the pretraining phase.

Results

Pretraining was halted after 2 million epochs where the model achieved an accuracy rate of 90.7% on the speaking task and an accuracy rate of 91.7% on the hearing task. After 0.8 million epochs of reading training, the model accurately produced 99.4% of phonological representations and 93.3% of semantic representations on the reading task.

Exploring AoA effects in the model

Behavioural naming data and lexical decision data (Cortese & Khanna, 2007) were simulated by mappings from orthographic to phonological representations (Chang, Furber, & Welbourne, 2012; Monaghan & Ellis, 2010), and by mappings from orthographic to semantic representations (akin to polarity measure in Plaut, 1997), respectively. According to the representation mapping theory, we would expect to obtain a larger AoA effect in lexical decision (semantics) than in naming (phonology) whereas the semantic locus theory predicts an AoA effect mainly in lexical decision (semantics), although if considering the role of semantics in naming (Strain et al. 1995), we might obtain a small AoA effect in naming (phonology) as well.

Multiple regression analyses were conducted on the model's phonological and semantic error scores to examine the AoA effects in the model. The predictor variables included: cumulative frequency (CF), orthographic neighbourhood size (OrthN), word length (Len), consistency (Cons), concreteness (Conc), and age of acquisition (AoA). Orthographic neighbourhood size was based on the number of words that can be made by changing one letter of the target word Coltheart (1977). The score of consistency was based on rime consistency, measuring the number of friends (sharing the same rime and pronunciation) divided by the total number of words sharing the same rime and weighted by their frequency values. The consistency score for each word was directly derived from the training corpus. The concreteness score was taken from Brysbaert, Warriner, and Kuperman (2014). AoA was taken as one of the 14 reading stages during training derived from the WFG.

All items in the training set were tested. Error items and outliers (3 standard deviations farther from the mean) were discarded and this removed about 3.7% of the items. In addition, words without measures for all psycholinguistic variables were removed, leaving 5272 words for analysis. Both the phonological and semantic error scores were log transformed to reduce the skew of performance distribution and all the predictor variables were centered in order to more clearly explore interaction terms.

Multiple Regression Results

Correlation analyses were conducted between the predictors. As expected, CF and AoA had a strong negative correlation, and AoA and Cons were positively correlated, suggesting early learned words tend to be high in frequency and inconsistent. OrthN was negatively correlated with Len, indicating that long words tend to have few neighbours.

To examine the unique contribution made by AoA to the model's performance, hierarchical regression analyses were conducted. For the word naming task, in step 1 all variables were entered into the regression model except AoA. The results showed CF, OrthN, Cons, and Len all made significant contributions. When AoA was entered into the regression model in step 2, it was a significant predictor (see Table 1). Similar analyses were conducted for the lexical decision task. In step 1, CF, Conc and Len were significant predictors. Again, in step 2, AoA was a significant predictor. These results showed that the AoA effects were found in both naming and lexical decision. Also the standardized beta value (β) was larger for the lexical decision than for the naming task, replicating behavioural studies showing a stronger AoA effect in tasks involving semantics than phonology (Table 1). For all these regression models, collinearity diagnostic analyses showed all variance inflation factors (VIFs) smaller than 4, confirming no serious multicollinearity problem.

Further regression analyses were conducted to examine the interaction terms. Three interaction terms were created: CF x Cons, to determine whether the model can replicate the widely observed consistency by frequency interaction

(Taraban & McClelland, 1987); AoA x Cons, to determine whether the model could reproduce the consistency by AoA effect (Monaghan & Ellis, 2002), thereby reflecting the mapping theory of AoA; and AoA x Conc, to determine whether AoA might have different effects depending on the semantic richness of the representation, reflecting the semantic locus theory of AoA.

In step 1, all the variables including AoA were entered into the regression model, and in step 2, the interaction terms were entered into the model separately. The results are summarized in Table 2. For the naming task, both CF x Cons and AoA x Cons were significant predictors, reproducing key behavioural effects on word naming. For lexical decision, only AoA x Conc was significant. Thus, consistency effects were less pronounced for lexical decision, consistent with the behavioural data. The AoA x Conc interaction term indicated that the AoA effect is modulated by the richness of the semantic representations, as measured by concreteness of the word.

Table 1. Results from a two-block regression analyses for the exploration of AoA in predicting both naming and lexical decision model performance.

		Naming	Lexical Decision
Step 1		β	β
	CF	-0.185***	-0.158***
	OrthN	-0.255**	0.019
	Cons	-0.247***	-0.015
	Len	-0.071***	-0.126***
	Conc	-0.001	-0.076***
	R^2(%)	21.94	24.14
Step 2			
	AoA	0.194*** $\Delta R^2 = 0.99\%$	0.406*** $\Delta R^2 = 4.33\%$

***p<.001; **p<.01; *p<.05; β is a standardized beta value

Table 2. Results from a two-block regression analyses for the exploration on three interaction terms in predicting both naming and lexical decision model performance.

	Naming	Lexical Decision
Step 1: Lexical Variables, R^2	22.93	28.47
AoA, β	0.194***	0.406***
Step 2: Interactions		
Model 1:		
CF x Cons, β	0.05***	-0.009
AoA, β	0.213***	0.403***
Model 2:		
AoA x Cons, β	-0.067***	-0.009
AoA, β	0.209***	0.408***
Model 3:		
AoA x Conc, β	0.004	-0.036**
AoA, β	0.192***	0.422***

***p<.001; **p<.01; *p<.05; β is a standardized beta value

General Discussion

This paper aimed to develop a large-scale computational model of reading including orthographic, phonological and semantic representations. Following Monaghan and Ellis (2010), the model was trained with a cumulative learning process. The model was able to produce correct phonological and semantic patterns for hearing, speaking, naming, and lexical decision tasks.

Multiple regression analyses on model performance demonstrated that the model was able to account for a range of standard word naming effects including cumulative frequency, orthographic neighbourhood size, consistency, concreteness and the interaction between cumulative frequency and consistency. More importantly, the results showed that AoA accounted for an additional 0.99% of variance in naming and 4.33% of variance in lexical decision, when other potentially confounding variables such as cumulative frequency and concreteness had been considered. Collectively, the regression results are consistent with the findings of previous regression analyses for behavioural (Cortese & Khanna, 2007) and computational (Monaghan & Ellis, 2010) studies.

So where in the model do the AoA effects derive? According to the representation mapping theory (Ellis & Lambon Ralph, 2000), the AoA effect could be observed when the mappings between input and output are more arbitrary. The significant interaction between AoA and consistency obtained in the regression analyses of naming in the model is consistent with the finding of Monaghan and Ellis (2002). In addition, the regression results also showed that the effect size of AoA (indexed by β) was larger for lexical decision than for naming (Table 1). This can be explained by the representation mapping theory in terms of different degrees of arbitrary mappings required for generating semantic versus phonological representations from orthography.

However, the current results cannot rule out the semantic locus theory (Brysbaert et al., 2000). This is because the semantic locus theory also predicts a larger AoA effect in lexical decision than in naming because it involves a greater role of the semantic representations themselves. Interestingly, there was a significant interaction between AoA and concreteness. This suggests that although the AoA effect in the model is due to mapping for word naming, for lexical decision it is likely a composite of effects in the

mappings between representations, and due to the semantic richness of the representations. So the present results provide evidence for the view that AoA effects arise from different sources according to task requirements.

The role of AoA in the reading system is profound, and effectively implementing these effects requires a computational model that can take into account the life history of the learner. We have shown that an implementation of the triangle model, involving orthographic, phonological, and semantic representations is able to take the chronology of experience and produce consequent effects in a mature reading system, resulting in AoA effects. We have replicated key behavioural data showing different sized effects of AoA depending on the lexical task (word naming or lexical decision), and linked this to the involvement of semantic representations in the task. For tasks primarily involving phonological representations, AoA effects are largely derived from the influence of experience on mapping between representations. For tasks that also involve semantics, the AoA effect is multicomponential. Experience affects mappings between representations but also the richness of the consequent representations. Happily, our model suggests that theorists with different views of the origin of effects of AoA are none of them wrong, but rather correct to varying degrees according to lexical task constraints.

Acknowledgments

This research was supported by ESRC grant RES-000-22-4049.

References

Brysbaert, M., Wijnendaele, I. V., & Deyne, S. D. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104*(2), 215-226.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition, 13*(7-8), 992-1011.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods, 46*(3), 904-911.

Catling, J. C., & Johnston, R. A. (2009). The varying effects of age of acquisition. *Q J Exp Psychol, 62*(1), 50-62.

Chang, Y. N., Furber, S., & Welbourne, S. (2012). "Serial" effects in parallel models of reading. *Cognitive Psychology, 64*(4), 267-291.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: an analysis of 2,342 words. *Q J Exp Psychol, 60*(8), 1072-1082.

Ellis, A. W., & Ralph, M. A. L. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology-Learning Memory and Cognition, 26*(5), 1103-1123.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica, 115*(1), 43-67.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*(3), 662-720.

Juhasz, B. J. (2005). Age-of-Acquisition Effects in Word and Picture Identification. *Psychological Bulletin, 131*(5), 684-712.

Lewis, M. B., Gerhand, S., & Ellis, H. D. (2001). Re-evaluating age-of-acquisition effects: are they simply cumulative-frequency effects? *Cognition, 78*(2), 189-205.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Comput. Linguist., 19*(2), 313-330.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-312.

Monaghan, J., & Ellis, A. W. (2002). What exactly interacts with spelling--sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 183-206.

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language, 63*(4), 506-525.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes, 12*(5-6), 765-805.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science, 29*(1), 41-78.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*(5), 1140-1154.

Strain, E., Patterson, K., & Seidenberg, M. S. (2002). Theories of word naming interact with spelling--sound consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 207-214.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language, 26*(6), 608-631.

Zeno, S. (Ed.). (1995). *The educator' s word frequency guide.* Brewster, NJ: Touchstone.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language, 47*(1), 1-29.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition*, *32*(1), 31-38.

Differentiating between Encoding and Processing during Diagnostic Reasoning: An Eye tracking study.

Anja Klichowicz (anja.klichowicz@psychologie.tu-chemnitz.de)
Technische Universität Chemnitz, Department of Psychology, Cognitive and Engineering Psychology, Wilhelm-Raabe-Strasse 43, D-09120 Chemnitz, Germany

Agnes Scholz (agnes.scholz@psychologie.tu-chemnitz.de)
Technische Universität Chemnitz, Department of Psychology, Cognitive and Engineering Psychology, Wilhelm-Raabe-Strasse 43, D-09120 Chemnitz, Germany

Sascha Strehlau (sascha.strehlau@yahoo.de)
Technische Universität Chemnitz, Department of Psychology, Cognitive and Engineering Psychology, Wilhelm-Raabe-Strasse 43, D-09120 Chemnitz, Germany

Josef F. Krems (josef.krems@psychologie.tu-chemnitz.de)
Technische Universität Chemnitz, Department of Psychology, Cognitive and Engineering Psychology, Wilhelm-Raabe-Strasse 43, D-09120 Chemnitz, Germany

Abstract

When finding a best explanation for observed symptoms a multitude of information has to be integrated and matched against explanations stored in memory. Although assumptions about ongoing memory processes can be derived from the process models, little process data exists that would allow to sufficiently test these assumptions.

In order to explore memory processes in diagnostic reasoning, 29 participants were asked to solve a visual reasoning task (the Black Box paradigm) where critical information had to be retrieved from memory.

This study focused on differentiating between processes that take place during the encoding and the evaluation of symptom information by comparing eye movement measures (the number of fixation and fixation duration per dwell).

Results will be discussed in light of existing theories on sequential diagnostic reasoning. Further, it will be discussed to which extent eye movements can be informative about memory processes underlying sequential diagnostic reasoning.

Keywords: diagnostic reasoning; eye tracking; process tracing; encoding-processing differences

Introduction

In sequential diagnostic reasoning multiple pieces of information have to be combined to find a best explanation for observed symptoms (e.g., Johnson & Krems, 2001). It is a complex cognitive process since the reasoner generates an undefined number of explanations for any number of observations (Johnson & Krems, 2001). Nevertheless, understanding this process is a major goal of research concerning reasoning and problem solving because of its high practical relevance. For instance, in the medical context, a complete understanding of diagnostic reasoning can help to save lives by improving the process of forming the right diagnosis (Mehlhorn, Taatgen, Lebiere, & Krems, 2011). But there are more applications such as finding the error in a technical system like a car or a computer (Johnson & Krems, 2001; Krems & Zierer, 1994; Mehlhorn et al., 2011). For instance, imagine you experience a loss in power of your car. Later, you witness some blue smoke coming from your exhaust pipe. Furthermore, you feel that recently, your car needs more oil as usual and your "check engine" light turns on. By combining these observations, you come up with the explanation that your car has an engine damage.

Diagnostic reasoning involves the processing of a number of observations and explanations. Often, the reasoner does not have all the necessary information available at once, but receives them in a sequential order. The reasoner then needs to integrate the symptom information into a situation model containing symptoms and explanations (Johnson & Krems, 2001; Johnson-Laird, Byrne, & Schaeken, 1992).

Besides this complexity, people are generally able to solve problems (Johnson & Krems, 2001), but how do they successfully engage in this demanding task? A number of process models (e.g., TAR: Johnson & Krems, 2001; TEC: Thagard, 1989; HyGene: Thomas, Dougherty, Sprenger, & Harbison, 2008) provide assumptions about ongoing memory processes. For instance, TAR assumes that *encoded* symptoms have to be *evaluated* concerning their fit with the current model of explanations (Johnson & Krems, 2001) and HyGene states that newly *encoded* information has to be judged concerning its implication for existing explanations (e.g., Thomas et al., 2008). Thus, in order to test these process assumptions, it is necessary to disentangle encoding and processing. Therefore, the process of diagnostic reasoning needs to be made visible. So far, different methodological approaches exist to trace memory processes in higher order cognitive tasks like judgement and decision

making (e.g., Glaholt & Reingold, 2011; Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011).

Recently, eye tracking is employed as a process tracing method to assess memory processes (see Glaholt & Reingold, 2011; Jahn & Braatz, 2014; Schulte-Mecklenbeck et al., 2011). Advanced hardware and improved understanding of its measures make it possible to get better insights (Jahn & Braatz, 2014; Renkewitz & Jahn, 2012; Scholz, von Helversen, & Rieskamp, 2015).

Monitoring eye movements allows the study of cognitive processes while participants interact with present objects without causing any restrictions and with minimal intrusions on the participant's behavior (Glaholt & Reingold, 2011). But interpreting eye tracking measures is challenging. The growing number of measures and their variety of applications complicate the assignment of specific measures to specific cognitive processes (Holmqvist et al., 2011).

Glaholt and Reingold, (2011) present a way to differentiate between encoding and processing in decision making by analyzing eye movement data. They asked participants to choose one picture out of a set of nine. They categorized dwells (defined as the sum of fixation durations of all fixations from the moment the gaze enters an area until it exits this area) into first-visit dwells and re-visit dwells, whereas re-visit dwells are the repeated viewing of an alternative. They found that *mean fixation durations per dwell* increases over time. The *number of fixation per dwell* is increasing between first-visits and re-visits as well. This indicates processing, such as evaluation of the current stimuli. The researchers interpreted these results as evidence that early in the process, participants merely screened the stimuli, moving to processing of the stimuli as they try to reach a decision. Even very early research finds evidence that the deliberate processing of information results in more and longer fixations (Loftus & Mackworth, 1978). As screening/*encoding* is assumed to happen implicitly without conscious control (Betsch, Hoffmann, Hoffrage, & Plessner, 2003) few fixations are plausible and in line with assumptions by Horstmann, Ahlgrimm, and Glöckner, (2009). *Increased* number of fixations, on the other hand, reflect processes such as the *evaluation* of an alternative which emerge later in the process of decision making (Horstmann et al., 2009). In contrast to Glaholt and Reingold (2011), Horstmann et al., (2009) did not find evidence that processing is associated with longer fixation durations.

Previous research on differences in eye movement patterns between encoding and processing of information in memory focuses on decision making. We assume that it is possible to use the knowledge about processes of decision making in the context of diagnostic reasoning as well. Encoding and processing differences are part of many models and assumptions. Whenever there is something to decide, we encode much information in a short period of time and subsequently evaluate and process this information (Glöckner & Betsch, 2008). As stated by TAR or HyGene (Johnson & Krems, 2001; Thomas et al., 2008), we assume that the same holds true when finding a best explanation for a set of symptoms.

Congruently, we assume that encoding and processing can be differentiated by the means of eye tracking measures, i.e. analyzing fixation duration and number of fixation per dwell. Following Glaholt and Reingold, (2011) and Horstmann et al., (2009) we assume that the number of fixation per dwell is increased during processes which appear later in the reasoning process such as symptom evaluation compared to the encoding of information.

In order to test our hypothesis, we used a task where information was presented in a sequential order. Remember the example with the engine trouble, people most likely witness one symptom such as the power loss at first and subsequently watch for more symptoms which are usually discovered one after another. In addition we need a task with the complexity comparable to that of everyday life problems. There were many symptoms pointing in the direction of engine damage, some of which might be related. The blue smoke is easily explained as soon as you know about the increased oil consumption. The car might burn some oil. On the other hand, we needed to control for prior knowledge of participants to get comparable results. Additionally, the task has to be learnable in an experimental setting. These requirements are met in Black Box task (BBX, Johnson & Krems, 2001). The Black Box task is a reasoning paradigm in which all information is, just as in our car example, *visuospatial* in nature. The symptoms are related to different areas of the car such as the exhaust pipe, the cockpit, or the dipstick in the motor compartment. Therefore, the Black Box task is especially suited to study memory processes during diagnostic reasoning with eye movements.

Method

Participants solved the Black Box task while their eye movements were recorded. In the Black Box paradigm, the participants' task is to determine a hidden state of a device using indirect evidence that needs to be combined following specific rules. Participants first learn and then apply the rules to get insight into the device by explaining observations with a combination of causes.

Participants

Twenty-nine students enrolled at Technische Universität Chemnitz took part in the experiment. One participant had to be excluded due to technical problems. Of the remaining participants 20 were female and 8 were male with a mean age of M = 22.3 (SD = 3). All participants had normal or corrected to normal vision.

Task and Apparatus

The participants' task is to locate hidden atoms in the Black Box which consists of a 10 x 10 grid by watching where light rays enter and exit the box. Participants do not see the path of the light rays. They only see the entrance and exit position of the light rays. As shown in Figure 1 each

atom has a field of influence (a circle around the atom). Hitting this field, the light rays get reflected and exit the box, depending on where in the box an atom is located.

Participants have to place the atoms based on the information drawn from the rays that were shot in the box. The participant has to place as many atom markers at the grid as needed to explain the ray pattern of one trial without placing more atoms as absolutely necessary. During one trial, the participant watches a fixed number of rays in sequential order one at a time displayed by numbers marking the entrance and exit of the current ray (Figure 2). Participants have to remember where the rays entered and exited the Black Box as well as the atom locations of already set atoms, because, just as in the car example, symptoms might be related. For instance, it is possible that a previous pattern needs to be remembered to explain the current one (as illustrated in Figure 2: Ray five can be explained by ray three.) Therefore, the paradigm allows to measure memory processes that take place in sequential diagnostic reasoning. The participant decides when to move on to the next ray by pressing the space bar. How many rays are left during one trial is shown at the upper left corner of each Black Box (cue). An exemplary trial is pictured in Figure 2.

The study recorded gaze data using a binocular IViewX RED eye-tracking system from SensoMotric Instruments with a sampling rate of 120 Hz. Data was analyzed with BeGaze 3.0, Microsoft Excel 2007 and IBM Statistics 23 (SPSS).

Stimuli were presented on a 22-inch computer screen using EPrime 2.0 software with a resolution of 1680 × 1050 pixels. All subjects were seated at a distance of 600 to 800 mm in front of the screen.

Procedure

Each participant was tested individually. After an initial instruction phase, in which participants were familiarized with the rules of the Black Box, two training phases followed. During the first training phase, rays and atoms were visible throughout the trial and therefore did not have to be remembered. In the second training phase participants solved the trial under test conditions (memory-based). That is, they only saw where the current ray of light entered and exited the Black Box and the cue in the upper left corner (showing how many rays are left in the current trial). They saw the atom and its field of influence as soon as they placed it by using the mouse until they demanded a new ray shot into the Black Box. No prior rays or atoms were present. Each training phase consisted of seven trials. A calibration and 48 test trials followed. The experiment ended after a short survey containing demographic questions. Participants needed 53 to 127 minutes to complete the entire experiment (M = 85.2 min; SD = 21.1).

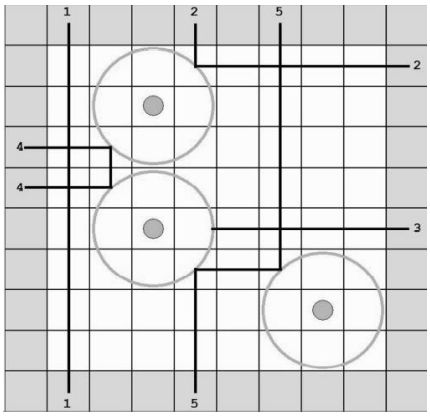


Figure 1: Rules in the Black Box.
If the ray does not hit any atom, it goes straight through the Black Box (1). If a ray hits the field of influence at an angle, it is reflected 90 degree. This results in a L-pattern (2). If a ray hits a field of influence of an atom straight forward, it is absorbed (3). The ray does not exit. Hitting a second field of influence after being already reflected the light ray path results in an U-pattern (4) or a Zick-Zack-pattern (5). Note, light ray paths and atom locations are not visible for the participant and have to be inferred. In this example all five observations can be explained by three atoms.

Analysis

To test our hypotheses, quadratic Areas of Interests (AOIs) were drawn around each square of the grid resulting in 100 separate AOIs, also called gridded AOIs (see Holmqvist et al., 2011). Furthermore, we coded the AOIs where the rays entered and exited the Black Box, where they hit the field of influence of an atom and where atoms had to be set.

Following Glaholt and Reingold, (2011) we identified the order of dwells in the AOIs and compared first-visit and re-visit dwells. First-visits are termed the first dwell in each AOI. Re-visits are all *repeated dwells in that AOI*. Re-visit dwells integrate all dwells from the second till the last dwell per AOI. In this point we slightly differ from the methods applied by Glaholt and Reingold, (2011) who defined *every* dwell as a re-visit that follows after any AOI is viewed a second time, even if the gaze hits an AOI not viewed yet. In decision making the transition between encoding the alternatives and their evaluation is believed to be the point in time where the first alternative is viewed a second time (Russo & Leclerc, 1994). Since our material is relatively complex we expect participants only to be able to encode information if they actually look at it. If the gaze shifts to a different AOI, information need to be encoded and subsequently evaluated as well. Therefore, for diagnostic reasoning, we expect that the first dwell into an AOI reflects encoding and that all subsequent dwells in that same area represent evaluation processes since participants test the implication of the information for the current explanation (e.g., Johnson & Krems, 2001).

We expect every first-visit of an area to map onto the encoding of that area. Even if a dwell consist of only one

fixation; we are confident that participants encoded this location sufficiently during their first visit since we defined a fixation with a minimum duration of 80 ms which is enough time to gather information since 50 to 60 ms are assumed enough time to encode even words (Rayner, 2009). During every re-visit of that same area we expect participants to process this location since it is already encoded.

For every dwell we calculated the mean number of fixations and the mean fixation duration. We aggregated first-visit dwells and re-visit dwells over every symptom presentation, trial and participant.

Results

Participants tended to look mostly at task relevant areas. Aggregated over symptom presentations, trials and participants 12.08 % of the fixation time was on average directed to the AOI containing the atom (*SD* = 9.05). The participants looked mostly to the AOI where the rays hit the field of influence of an atom (*M* = 21.16 %, *SD* = 14.76) and 14.63 % of the fixation time fell on average to the two AOIs covering the rays (*SD* = 7.00). 5.24 % of the time participants fixated on the AOI with the cue in the upper left corner (*SD* = 4.15). Taken together, participants looked more than 50% of the time to direct task-relevant AOIs which covered only about 5 % of the grid. Fixation duration weighted by the number of AOIs, participants looked significantly longer to the atoms ($t(27) = 19.82$, $p < .001$, $d = 3.81$), to the field of influence of the atom ($t(27) = 20.38$, $p < .001$, $d = 3.92$), to the rays ($t(27) = 20.45$; p = <-001, $d = 3.80$) and to the cue in the upper left corner ($t(27) = 12.86$, $p < .001$, $d = 2.47$) than to all the other AOIs.

They found correct solutions for most of the trials (*M* = 71.3%, *SD* = 0.13). Therefore, we conclude that participants have understood task instructions and engaged in diagnostic reasoning as we intended.

The calibration procedure reached a satisfying accuracy with a mean deviation of *M* = 0.54° (*SD* = 0.53).

We assumed that encoding and processing can be differentiated by the means of the eye tracking measures analyzing dwells concerning fixation duration and fixation frequency. More precisely, our hypothesis predicted that fixation duration and number of fixations per dwell is increased during re-visits compared to first-visits, assuming that first-visits represent encoding whereas re-visits map onto processing.

Over all, we find longer fixations during re-visits than first-visits ($M_{first\text{-}visits}$ = 267 ms, *SD* = 45, $M_{re\text{-}visits}$ = 279 ms, *SD* = 51, $t(27) = -1.94$, $p = .06$, $d = 0.37$) and we find significant more fixations during re-visit dwells compared to first-visit dwells ($M_{first\text{-}visits}$ = 1.25, *SD* = 0.09, $M_{re\text{-}visits}$ = 1.31, *SD* = 0.10, $t(27) = -4.11$, $p < .001$, $d = 0.79$).

An analysis of the different pattern separately shows that there are differences. To illustrate this, we exemplary report the L-pattern and the absorption in Table 1.

The increase in fixation duration and the significant increase in number of fixations per dwell during re-visits compared to first-visits support our hypothesis that encoding and processing can be differentiated by the means of the eye tracking measures as Glaholt and Reingold, (2011) assumed.

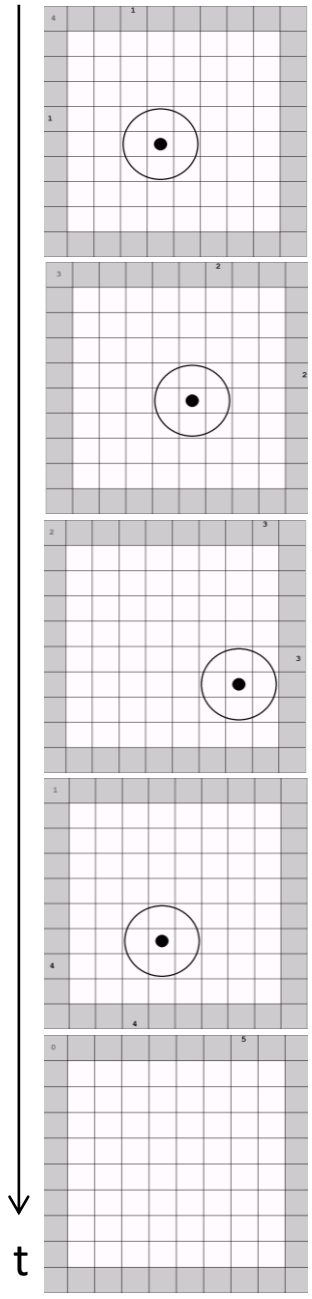


Figure 2: Exemplary trial
The trial consists of five rays shot into the Black Box sequentially. For each ray the participant can place atoms to explain them. When finished with one ray the participant decides when to move on to the next one. The number (cue) in the upper left corner shows how many more rays are to come in the current trial. In this example the participant should notice that no new atom need to be placed for ray five since this ray can be explained by the atom already set for ray three.

Table 1: Analysis of L-pattern and Absorption

		L-pattern	Absorption
Mean fixation duration per dwell	First-visit *M (SD)*	277ms (51)	255 ms (44)
	Re-visit *M (SD)*	279 ms (51)	280 ms (64)
	t-test	$t(27) = -0.27$, $p = .79$, $d = 0.05$	$t(27) = -2.31$, $p = .03$, $d = 0.44$
Number of fixations per dwell	First-visit *M (SD)*	1.26 (0.09)	1.19 (0.11)
	Re-visit *M (SD)*	1.34 (0.13)	1.27 (0.15)
	t-test	$t(27) = -4.46$, $p < .001$, $d = 0.83$	$t(27) = -2.59$, $p = .02$, $d = 0.50$

Discussion

This study aimed to find a measure which is able to describe reasoning processes differentiating between encoding and processing. Therefore, we employed eye tracking as a process tracing method because it is an objective and fine grained measure to human behavior. Following research by Glaholt and Reingold, (2011) and Horstmann et al., (2009) we assumed that encoding and processing can be differentiated by analyzing dwells concerning fixation duration and fixation frequency.

Horstmann et al., (2009) and Glaholt and Reingold, (2011) showed that a higher number of fixations might be an indicator for more deliberate thinking and processing. Since we assume encoding to be more intuitive and number of fixations increases during re-visits when processing takes place, our results support the research of Horstmann and colleagues (2009) as well as Glaholt und Reingold, (2011). Over all patterns as well as for L-pattern and absorption individually increases the number of fixation with medium to high effects. In line with the research by Horstmann et al., (2009) the analysis of fixation duration draws a less clear picture. It might be that there is not such a clear difference in fixation duration between (more intuitive) encoding and (deliberate) processing as Horstmann et al., pointed out. The researchers assume that intuitive and deliberate processes may have similar underlying mechanisms as indicated by no clearly distinct differences in fixation durations. The differences between encoding and processing might not be qualitative in nature but mainly quantitative regarding information sampling. However clear differences between L-pattern and absorption might indicate that both patterns are not strictly similar processed. Horstmann et al., (200) state that there may still be crucial differences regarding intuitive and deliberate processes, which cannot be shown by this measure. Since absorptions only identify the row or column of the atom location, a second ray of light is needed to determine the exact location in the Black Box, making the absorption a more complex pattern. It remains to future research to investigate the differences in processing of different rules. Therefore the analysis of U-pattern and Zick-Zack- patterns may be useful, as they are more complex as well. They are not included in the results of this paper since those patterns are not used frequently enough throughout the experiment to produce a reliable data basis.

We rule out the possibility that participants did not process during re-visit dwells but encoded new features of already encoded information. Since a single AOI contained simple material with not much features like colors or different shapes we consider this alternative very unlikely in this specific task.

Further, we are confident that participants actually encoded and processed what they were looking at. The fact that participants reached fairly good response accuracies in spite of a difficult and strongly memory based task speaks in favor of this assumption. There was no evident reason to separate eye movements and attention while solving the Black Box task.

Since looking back to associated, but emptied spatial locations can facilitate memory retrieval (e.g., Scholz, Mehlhorn, & Krems, 2016), it is an interesting questions whether there is a connection between gaze behavior and accuracy during diagnostic reasoning, too. Do people who solve the task successfully show a different gaze pattern? Since our participants reached a fairly high response accuracy in solving the trials, it is difficult to draw causal conclusions to this question from our data. Future research should address this question by careful manipulations.

We were able to replicate Glaholt and Reingolds (2011) results that an increased number of fixations to an already viewed area is an indicator for evaluation and processing of information. However, it is difficult to make a clear statement regarding the mean fixation duration per dwell.

In conclusion, eye movements show a high potential to assess memory processes during higher-order reasoning and thinking (see also Jahn & Braatz, 2014; Renkewitz & Jahn, 2012; Scholz et al., 2015). In particular, the number of fixations per dwell may be used as process tracing measures to asses processing during diagnostic reasoning. We found increased number of fixations during a dwell to indicate processing of information. The fact that we found a measure that indicates evaluation processes allows for a more detailed testing of process assumptions derived by process models on diagnostic reasoning such as TAR (Johnson & Krems, 2001), HyGene (Thomas et al., 2008) or TEC (Thagard, 1989). All these models make assumption about the manner in which the reasoner links observation or symptoms to a set of explanations. Thereby the models differ in their predictions to how this link is processed. TEC for instance, describes this process as a judgment to which extend symptoms and explanations are coherent (Thagard, 1989). Thereby, the encoding and evaluation of all information to reach a judgement is supposed to happen parallel. HyGene on the other hand, gives insight about the way explanations are generated and used to asses new information (Thomas et al., 2008). HyGene states that new information are encoded sequentially and have to be judged concerning their implication for existing explanations. TAR

describes diagnostic reasoning as a deliberate comprehension process in which observations are sequentially interpreted and integrated into a mental model (Johnson & Krems, 2001). TAR assumes that one information is encoded and evaluated before new information is gathered, making encoding and processes closely intertwined. All models have in common that they assume perception-action cycles with different encoding and processing phases. To describe reasoning, we need to know *how* people *use* information to reach causal conclusions. Therefore the first step is to know when people simply gather information and when they actually use it for evaluation, comparison or processing.

The process tracing measure presented in this paper may help to disentangle encoding and processing assumptions stated by different models giving more insight into diagnostic reasoning by providing a possibility to identify processing phases and test for precise research questions.

Acknowledgments

We thank Martin Baumann for helpful comments.

References

Betsch, T., Hoffmann, K., Hoffrage, U., & Plessner, H. (2003). Intuition Beyond Recognition: When Less Familiar Events Are Liked More. *Experimental Psychology*, *50*(1), 49–54. doi:10.1026//1618-3169.50.1.49

Glaholt, M. G., & Reingold, E. M. (2011). Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 125–146. doi:10.1037/a0020692

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks : Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Makingg*, *3*(3), 215–228.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures* (1rst Editi.). New York: Oxford University Press.

Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation ? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making*, *4*(5), 335–354. doi:10.2139/ssrn.1393729

Jahn, G., & Braatz, J. (2014). Memory indexing of sequential symptom processing in diagnostic reasoning. *Cognitive Psychology*, *68*, 59–97. doi:10.1016/j.cogpsych.2013.11.002

Johnson, T. R., & Krems, J. F. (2001). Use of current explanations of multicausal abductive reasoning. *Cognitive Science*, *25*, 903–939. doi:10.1016/S0364-0213(01)00059-3

Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418–439. doi:10.1037/0033-295X.99.3.418

Krems, J. F., & Zierer, C. (1994). Sind Experten gegen kognitive Täuschungen gefeit? Zur Abhängigkeit des confimation bias von Fachwissen. *Zeitschrift Für Experimentelle Und Angewandte Psychologie*, *XLI*(1), 98–115.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology. Human Perception and Performance*, *4*(4), 565–572. doi:10.1037/0096-1523.4.4.565

Mehlhorn, K., Taatgen, N. a., Lebiere, C., & Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1391–1411. doi:10.1037/a0023920

Rayner, K. (2009). *Eye movements and attention in reading, scene perception, and visual search. Quarterly journal of experimental psychology (2006)* (Vol. 62). doi:10.1080/17470210902816461

Renkewitz, F., & Jahn, G. (2012). Memory indexing: A novel method for tracing memory processes in complex cognitive tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1622–1639. doi:10.1037/a0028073

Russo, J. E., & Leclerc, F. (1994). An Eye-Fixation Analysis of Choice Processes for Consumer Nondurables. *Journal of Consumer Research*, *21*(2), 274. doi:10.1086/209397

Scholz, A., Mehlhorn, K., & Krems, J. F. (2016). Listen up, eye movements play a role in verbal memory retrieval. *Psychological Research*, 149–158. doi:10.1007/s00426-014-0639-4

Scholz, A., von Helversen, B., & Rieskamp, J. (2015). Eye movements reveal memory processes during similarity- and rule-based decision making. *Cognition*, *136*, 228–246. doi:10.1016/j.cognition.2014.11.019

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgement and Decision Making*, *6*(8), 733–739.

Thagard, P. (1989). Explanatory Coherence. *Behavioral and Brain Sciences*, *12*, 435–502.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185. doi:10.1037/0033-295X.115.1.155

Young children and adults integrate past expectations and current outcomes to reason about others' emotions

Desmond C. Ong[1] **(dco@stanford.edu)** **Mika Asaba**[1] **(masaba@stanford.edu)**
Hyowon Gweon (hyo@stanford.edu)
Department of Psychology, Stanford University, Stanford CA, USA

Abstract

Reasoning about others' emotions is a crucial component in social cognition. Here, we tested the ability of preschool children to reason about an agent's emotions following an unexpected outcome. Importantly, we controlled for the actual payoff of the outcome, while varying the prior expectation of the agents. Five-year-olds, but not four-year-olds, were able to correctly judge an agent's emotions following an unexpected outcome (Experiment 1). When explicitly provided with the agent's expectations, 4-year-olds were then also able to correctly judge the agent's feelings (Experiment 2). Our results suggest that the ability to reason about emotions given outcomes and prior expectations develops by 4 years of age, while the ability to spontaneously infer such prior expectations develops soon after. We discuss our results in light of the developmental literature on emotion understanding and counterfactual reasoning.

Keywords: Affective Reasoning; Affective Cognition; Theory of Mind; counterfactual reasoning

As much as the ability to feel is a fundamental part of what makes us human, the ability to *reason* about how *others* feel—what we call Affective Cognition (Ong, Zaki, & Goodman, 2015)—is also core to who we are. We have a rich and intuitive understanding of emotions that allows us to reason and explain how others felt in the past, predict what others will feel in the future, and even intervene by changing our own actions to influence how others will feel.

As we reason about others' emotions, we often have to rely on unobservable mental states that are internal to agents (e.g., goals, beliefs, desires), especially in the absence of explicit cues (e.g., facial expressions or the valence of events). Previous developmental studies have found surprisingly early signatures of understanding the relationship between people's mental states, goal-directed actions, event outcomes, and affective states. For instance, infants can differentiate between actions that are congruent or incongruent with the affect that agents previously expressed towards objects (Phillips, Wellman, & Spelke, 2002). Furthermore, infants distinguish between displayed emotion that is congruent or incongruent with the outcomes of agent's goal-directed actions (Skerry & Spelke, 2014). For instance, 10-month-olds showed more surprise when an agent successfully achieved its goal (e.g., jumping over a barrier) and expressed incongruent (negative) affect, as compared to when the agent displayed congruent (positive) affect. However, when the agent failed to achieve its goal, infants did not distinguish between congruent (negative) and incongruent (positive) affect (Skerry & Spelke, 2014). By preschool, children can also understand that failed outcomes lead to negative affect (e.g., Barden, Zelko, Duncan, & Masters, 1980; Wellman & Banerjee, 1991; Wu, Baker, Tenenbaum, & Schulz, 2014).

However, such competence might not necessarily suggest that children have a full-fledged, theory-like understanding of others' goals, actions, and affective states by preschool years. For instance, children's ability to link successes with positive emotions (and failures with negative emotions) might reflect mere statistical associations that young children have acquired from early social experience. As adults, we intuitively understand that the same outcome can have very different affective consequences depending on the agent's prior beliefs, or expectations, about what would happen. Importantly, these expectations cannot be explicitly observed from others' actions, but must be generated by reasoning about the agent's mental states, the structure of the physical world, and the agent's interactions with the physical world. In this study, we focus on children's ability to integrate an agent's prior expectation with an outcome when predicting their affect.

Imagine this scene at the bowling alley. Sally throws her ball: it heads towards the gutter but remarkably, curves back to finally knock down three pins. Annie's ball rolls straight towards the center pin, but it curves to the left and only hits three pins. Although Sally and Annie experienced identical, unexpected outcomes (i.e., hitting three pins), one might reasonably infer that Sally feels happier than Annie. Although seemingly intuitive, this conclusion requires a complex set of inferences, involving our understanding of the physical and social world. To judge Sally as happier than Annie, we must (1) understand where each ball *could* go and project this to our friend's expectations, (2) link these expectations with affective states, (3) calculate the discrepancy between the expected and the actual outcome, and (4) judge each character's emotional responses with respect to the expected and actual outcome. An adult-like prediction of the characters' emotions (e.g., Annie as disappointed and Sally, relieved) involves integrating information from various domains.

Decades of developmental research have shown how each of these components independently undergo their own developmental trajectory in early childhood. First, understanding where the ball could go requires an understanding of how objects move and how they respond to causal interventions from agents. It takes years to construct a coherent, intuitive theory of physics that supports accurate inferences and goal-directed actions from infancy (Spelke, Breinlinger, Macomber, & Jacobson, 1992) to the preschool years (Hood, Cole-Davies, & Dias, 2003; Keen, 2003). Second, children also show both surprising competence as well as a significant developmental

[1]These authors contributed equally to this work.

change in their understanding of others' emotions as reflected in their display cues (e.g. Gnepp & Hess, 1986) and consequences of their motives and behaviors (e.g. Yuill, 1984). More broadly, young children undergo significant changes in using others' mental states to predict and explain behaviors (Baillargeon, Scott, & He, 2010; Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Wellman, Cross, & Watson, 2001). It is an open question as to how children integrate inferences of others' mental states (e.g., goals and beliefs) with knowledge about the physical world to reason about affective states.

Across two studies, we examine whether preschool children utilize others' prior expectations (i.e., others' beliefs about what they expected to happen) to draw different conclusions about an agent's affective response to identical physical outcomes. Specifically, we adapted the bowling scenario introduced above into a simple task. In Experiment 1, we examined children and adults' inferences about others' emotions after having observed the full scenario. In Experiment 2, we break down the task into two parts and explicitly provide the character's expectations to uncover a more fine-grained understanding of the inferences involved in the final judgment.

Experiment 1

In Experiment 1[2], we examined how children and adults reason about agents' emotions following an unexpected outcome. Participants observed characters bowling and provided emotion ratings for each character.

Participants. Children: Twenty 4-year-olds (8 female, M_{Age}(SD): 4.53(.26), range: 4.19 - 4.98) and twenty 5-year-olds (13 female, M_{Age}(SD): 5.32(.16), range: 5.06 - 5.61) were recruited from a campus preschool and local museum. One additional child was tested and dropped from analysis due to failure to respond to test questions. Adults: Fifty nine adults (17 female, M_{Age}(SD): 32.8 (11.2), range: 20 - 68) were recruited through Amazon's Mechanical Turk (AMT). One additional adult participated but was dropped for providing an incorrect answer to a planned attention check question ("How many pins were knocked down?").

Materials. Children first played a warm-up game using a child-friendly toy bowling set: six colorful foam pins, a red foam ball, and a black tarp for the bowling lane. We ran the main experiment on a computer using cartoons animated in Keynote. The cartoon bowling alley consists of a wide beige rectangle flanked by two narrow grey rectangles (the "gutter"; see Fig. 1). We designed the cartoon pins and balls to match the bowling set that children had played with earlier. We used simple, generic cartoon characters for the practice trials, and Elmo and Cookie Monster for the test trials. In all trials, the characters' backs were facing the child such that no facial expressions were shown. We used the same stimuli with adults.

[2]All experiments, data, and analyses can be found at: https://github.com/desmond-ong/bowling

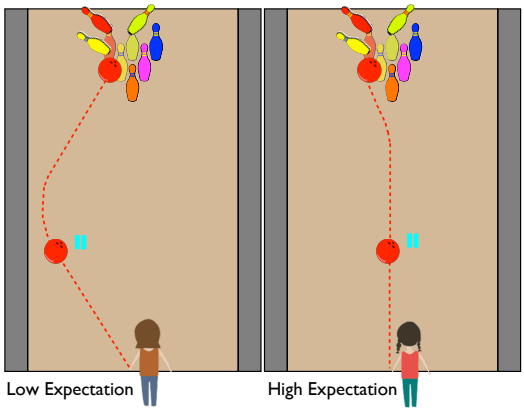


Figure 1: Sample stimuli used for Exp. 1 and 2. Left: Low Expectation video; Right: High Expectation video. Trajectories of the bowling balls are indicated by the dashed red lines. For Exp. 2, the position of the bowling balls at the midway pause are given next to the cyan "Pause" icons.

Procedures. Child Experiment: Children were tested in a quiet room inside of a campus preschool or local museum. The experimenter introduced the game of bowling by explaining that the goal is to use the ball to knock down as many pins as possible; children played with the bowling set for a few minutes. The experimenter then introduced the virtual bowling game on the computer. Children were asked how many pins were on the screen ("six") and what it meant when the bowling ball entered the gray areas ("it's out"). If a child could not count correctly or did not remember the answer, the experimenter repeated the explanation. Before the main experiment, all children were able to correctly report the number of pins and the meaning of the gray area.

There were four trial types: two practice trials presented in a fixed order (Gutter, Strike), and two test trials in a counterbalanced order (Low Expectation, High Expectation). In the Gutter trial, the character rolled the ball, which curved to the right, into the gray area, knocking down no pins. In the Strike trial, the character's ball went straight and knocked down all six pins. After each trial, the experimenter asked the child how many pins were knocked down and whether the character is feeling happy or sad. If the child said happy (sad), the experimenter asked, "kind of happy (sad), medium happy (sad), or really happy (sad)?" We converted children's verbal ratings to a 6-point scale, from "really sad" (1) to "really happy" (6), to be used as our Final Rating. After the practice trials, the experimenter asked: "Who do you think is happier?" (Forced Choice). If the child did not respond or said both, the experimenter asked the child to select one character.

The test trials had an identical sequence of prompts. In the Low Expectation trial, the ball initially headed sharply left towards the gutter but then curved back to hit three pins. In the High Expectation trial, the ball initially headed straight down the lane but then curved to the left to hit three pins. We counterbalanced the order and characters for the test trials.

Adult Experiment: Adult participants participated in a similar experiment with only the Low and High Expectation trials

(no practice trials). After each trial, they providing an emotion rating for the character on a 6 point Likert scale from "really sad" (1) to "really happy" (6). After both trials, participants were asked to choose who is happier; adults also had the option of reporting that the characters are equally happy.

Results. Children: All children correctly reported the number of pins knocked down in every trial. There were no significant effects of trial order or characters. For the practice trials, 4- and 5-year-olds provided higher ratings to the Strike character than the Gutter character (4 year-olds: Strike vs. Gutter: M(SD) = 5.95(.22) vs. 1.30(.57), $t(19) = 35.42$, $p < 0.001$; 5 year-olds: Strike vs. Gutter: M(SD) = 5.95(.22) vs. 1.70(.57), $t(19) = 29.76$, $p < 0.001$). When given a binary choice between the Strike and Gutter character, all children in both age groups reported that the Strike character was happier than the Gutter character (4-year-olds: 20 of 20, $p < 0.001$; 5-year-olds: 20 of 20, $p < 0.001$).

Next, we examined ratings for the test trials (Fig. 2). Recall that both trials have identical outcomes (3 pins knocked down) but they differ with respect to the balls' initial trajectories (and thus agents' prior expectations). Four year-olds' provided the same ratings for the Low Expectation and High Expectation characters (Low Exp. vs. High Exp.: M(SD) = 4.70(1.34) vs. 4.85(1.14), $t(19) = -.55$, $p = ns$), whereas five year-olds reported higher ratings for the Low Expectation character than for the High Expectation character (Low Exp. vs. High Exp.: M(SD) = 4.80(.95) vs. 4.30(1.22), $t(19) = 2.13$, $p = .046$). When given a forced choice between the Low Expectation and High Expectation character, neither group showed a strong tendency to choose the Low Expectation character (binomial test, $p = ns$ for both), but we note that more five-olds chose this character over the High Expectation character (13 of 20 5-year-olds vs. 9 of 20 4-year-olds).

Adults: Adults' responses were consistent with the 5 year-olds' ratings, providing higher emotion ratings to the Low Expectation character than the High Expectation character (Low Exp. vs. High Exp.: M(SD) = 4.39 (.74) vs. 3.86 (.80), $t(58) = 4.60$, $p < 0.001$). On the final forced choice question, 23 (of 59) participants chose the Low Expectation character as feeling happier, 2 chose the High Expectation character, and the remaining 34 said that both feel equally happy.

Our results revealed a difference between the age groups: Both adults and 5-year-olds provided more positive emotion ratings to the Low Expectation character and lower ratings to the High Expectation character. Four-year-olds were not able to differentiate between these trials. We consider two possibilities for 4-year-olds' failures. First, 4-year-olds might understand the physical trajectories of the ball and how it relates to the characters' expectations, but have difficulties inferring the characters' emotional responses with respect to the discrepancies between expectations and outcomes.

Another possibility is that 4-year-olds have difficulty using their physical understanding to spontaneously generate the characters' expectations. Additional pilot data suggests that 4-year-olds indeed have a weak understanding of the physics

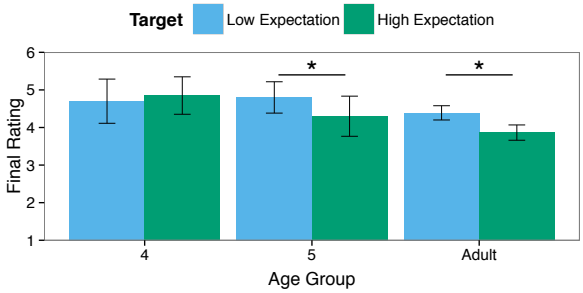


Figure 2: Exp. 1 Results. Participants' ratings of the characters' happiness after the test trials. Error bars indicate 95% confidence intervals. * indicates $p < .05$

involved in the experiment. We showed fourteen 4-year-olds (M_{Age}(SD): 4.66(.32), range: 4.12 - 4.99) the same videos, but with a pause in the middle. We asked children to predict where the ball would go and how many pins would be knocked down for the Low Expectation and High Expectation characters. Indeed, we found that children had great difficulty accurately predicting the final trajectory ("out" or "straight") and number of pins for the Low Expectation and High Expectation characters: only 7 of 14 children provided accurate responses for both characters. Thus, 4-year-olds' difficulty on this task may stem from a genuine inability to spontaneously generate the alternative outcome (where the ball is going to go and how many pins will be knocked down), which impairs their ability to utilize this alternative to infer the agent's emotion following the actual outcome.

Experiment 2

Experiment 2 tests the hypothesis that given explicit information about the agents' expectations, even 4-year-olds may be able to distinguish the two characters' emotions despite the same final outcomes. We used the videos from Exp. 1, but paused the videos halfway and explicitly provided participants with the characters' expectations about the ball's trajectory and how many pins will be knocked down.

Participants. Children. Eighteen 4-year-olds (8 female, M_{Age}(SD): 4.51(.32), range: 4.00 - 4.97), and eighteen 5-year-olds (9 female, M_{Age}(SD): 5.43(.27), range: 5.05 - 5.98), were recruited from a preschool and local museum. An additional 7 4-year-olds and 3 5-year-olds were tested and dropped due to incorrect responses to the initial emotion question (i.e., reporting the Low Expectation character to be happy or the High Expectation character to be sad; see Procedure). Adults. Fifty-eight adult participants were recruited through AMT (32 female; M_{Age}(SD) = 32.8(10.4), range: 18-64). An additional two participants were tested and dropped after failing planned check questions.

Materials. The stimuli were similar to those in Experiment 1, but we used the generic cartoon characters for the Low Expectation and High Expectation trials. The critical change was that the videos in the Low Expectation and High Expectation trials paused when the ball was halfway down the alley, and the experimenter explicitly provided the characters' ex-

pectations. To measure children's inferences about changes in emotion, we constructed a 10″ x 2″ physical scale with large arrows at each end. The scale was divided into eight equal sections with a large red sad face on the third section and a large green happy face on the fifth section (see Fig. 3).

Procedures. Child experiment: The procedures were similar to that of Experiment 1. After the warm-up game, children sequentially observed the Gutter and Strike practice trials and reported the number of pins that were knocked down. All children correctly reported the final outcomes. The experimenter then asked whether the character was feeling happy or sad (binary choice); children responded by putting a small magnet versions of the characters onto either the happy or sad face on the scale. After the Gutter and Strike trials, the experimenter showed both characters on the screen and asked the child to choose who is happier (forced choice).

Children then saw the Low Expectation and High Expectation test trials in a counterbalanced order. Halfway through the video, the video paused; this location is indicated by the ball and the cyan "Pause" icon in Figure 1 (the icon did not show up in the actual videos). The experimenter provided the character's belief about the ball's trajectory and the number of pins the ball would hit. In the Low Expectation trial, the video was paused when the ball was close to the gutter on the left side. The experimenter said, "Sally thinks that her ball is going to *go out* and hit *none* of the pins!" Children then rated how Sally feels right now ("happy or sad") by placing a Sally-shaped magnet on the scale (Initial Rating). When the video resumed, Sally's ball curved back and hit three pins. As an attention check, children were first asked how many pins were knocked down ("3"). The experimenter then reminded the child of their earlier emotion rating, and asked for a judgment of change, e.g., "Earlier you said she was feeling 'happy'. Do you think she is feeling better, or worse now?" (Verbal Report). The experimenter prompted the child to provide a final emotion rating for the character by saying, "Okay, can you show me? You can move her anywhere on the line". The marker's final position was used as the Final Rating.

The trial structure was identical in the High Expectation trial, except that the video was paused with the ball in the middle of the lane. The experimenter said, "Annie thinks that her ball is going to *go straight* and hit *all* of the pins!" After the child provided an Initial Rating by placing the Annie magnet piece on the scale, the video resumed to show Annie's ball continuing straight down the alley before curving left to knock down three pins. Children then answered how many pins she knocked down (Attention Check), indicated whether she feels better (Verbal Report), and provided a final emotion rating by moving the magnet piece on the scale (Final Rating). Last, after both test trials, the experimenter asked, "Who do you think is happier? Sally or Annie?" (Forced Choice). If the child did not respond or said both, the experimenter prompted the child to choose just one character.

Adult experiment: Adult participants participated in a similar experiment, with a few small changes. Adult participants were given the character's belief about the trajectory ("Annie thinks her ball is going straight / going out") and were prompted to predict how many pins the character expects to knock down (free-response from 0 to 6). Participants provided an initial happy or sad rating, observed the rest of the video, indicated whether the character feels better or worse, and provided a final rating for the character. Finally, after both trials, adult participants were asked to choose who is happier; they had the option of saying both are equally happy.

Results. All children correctly answered the number of pins knocked down in each trial. We observed no effect of trial order or character. For the Strike and Gutter practice trials, both 4- and 5-year-olds successfully reported that the Gutter character was sad (4 year-olds: 17 of 18; 5 year-olds: 18 of 18, $p < 0.001$ by binom. test for both) and the Strike character was happy (4 year-olds: 17 of 18; 5 year-olds: 17 of 18, $p < 0.001$ by binom. test for both). All children in both age groups chose the Strike character as happier than the Gutter character (4-year-olds: 18 of 18; 5-year-olds: 18 of 18, $p < 0.001$ by binom. test for both).

For the critical Low Expectation and High Expectation trials, we analyzed verbal responses ("better or worse?"), relative ratings (difference between the Initial Ratings and Final Ratings), and Final Ratings for each trial. Only children who correctly reported that the Low Expectation character was sad and the High Expectation character was happy in the Initial Ratings were included in our analyses. The vast majority of children reported that the Low Expectation character felt better after the outcome than at the pause (4 year olds: 16 of 18, $p = .001$; 5 year-olds: 17 of 18, $p < 0.001$), while children were at chance for reporting the High Expectation character to be feeling worse (4 year olds: 12 of 18; 5 year olds: 10 of 18). However, 5-year-olds showed a significant difference in their better/worse responses for the characters ($p = .02$, Fisher's exact) and 4-year-olds showed a marginal difference ($p = .09$). Next, we looked at children's relative ratings. Both age groups moved the Low Expectation character higher on the scale and the High Expectation character lower (4 year-olds: Low Exp. vs. High Exp.: M(SD) = 2.44(1.21) vs. -.83(1.91), $t(17) = -5.39$, $p < 0.001$; 5 year-olds: Low Exp. vs. High Exp.: M(SD) = 2.53(1.8) vs. -.86(2.39), $t(17) = -8.06$, $p < 0.001$). We found that 4-year-olds' final ratings for the Low Expectation and High Expectation characters did not differ (Low Exp. vs. High Exp.: M(SD) = 5.17(1.91) vs. 5.44(1.21), $t(17) = .79$, $p = ns$), whereas 5-year-olds rated the Low Expectation character as being happier than the High Expectation character (Low Exp. vs. High Exp.: M(SD) = 5.86(1.66) vs. 4.81(2.11), $t(17) = 2.60$, $p = .05$). Lastly, given a binary forced choice, neither group was able to report that the Low Expectation character was happier than the High Expectation character ($p = ns$ for both). However, 5-year-olds tended to choose the Low Expectation more often (12 of 18 5-year-olds vs. 8 of 18 4-year-olds).

Overall, adult participants responded as we predicted on

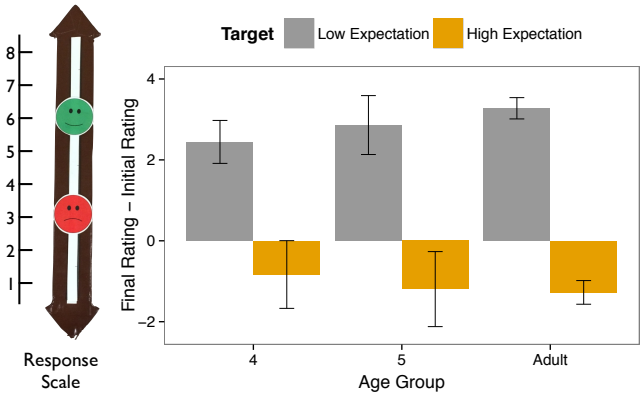


Figure 3: Left: Physical scale used in Exp. 2. Right: Participants' change in rating (before and after the outcome). Error bars represent 95% CIs.

all the questions. Participants reported that the Low Expectation character felt sad initially (57 of 58) and felt better after seeing the outcome (58 of 58); they thought the High Expectation character felt happy initially (56 of 58) and felt worse after the outcome (46 of 58). Adults' ratings were significantly higher for the Low Expectation character (Low Exp. vs. High Exp.: M(SD) = 6.33 (.98) vs. 4.62 (0.89), $t(57)$=10.46, p <0.001). On the forced choice question, 42 chose the Low Expectation character as feeling happier, and the remaining 16 said that both characters feel equally happy.

Given explicit information about characters' expectations, four- and five-year-olds showed an adult-like response, shifting their ratings in the appropriate directions after observing the final outcomes. Five-year-olds and adults, but not four-year-olds, providing higher final ratings for the Low Expectation character than the High Expectation character. Four-year-olds' relative differences between the Initial and Final ratings were not as robust as those of five-year-olds and adults, resulting in this lack of difference in the Final Rating score. One possibility is that younger children were simply rating the characters' emotions with respect to the final outcome (e.g., how each character feels after knocking down 3 pins), explaining the change in rating and the lack of difference between final ratings. However, 4-year-olds' "Better" and "Worse" verbal reports suggest that they are differentiating between these characters' emotions based on prior beliefs.

General Discussion

Across two experiments, we examined how children use an agent's prior expectations to judge how they would feel following identical, unexpected outcomes. In Exp. 1, We found a developmental shift between ages 4 - 5 in children's inferences about others' emotions. Given two characters with identical, unexpected outcomes, older children gave adult-like responses; a character who had a low expectation about the outcome would feel better than a character who had a high expectation. However, younger children (4-year-olds) were unable to distinguish the two characters. In Exp. 2, given explicit information about the characters' expectations, however, even the younger children appropriately adjusted their ratings about the character's feelings. Our results suggest that young children can already make use of an agent's prior expectations to reason about their relative emotions, but the ability to spontaneously infer those prior expectations continue to develop in preschool years.

Our study examined judgments about current emotions based on expectations, but we also often base these judgments from alternative states or *what could have happened* (contrary to what actually happened), a capacity often referred to as counterfactual reasoning. Five-year-olds can pass counterfactual thinking tasks (e.g., Beck, Robinson, Carroll, & Apperly, 2006) and factor alternate outcomes into judgments of their own emotions (Weisberg & Beck, 2010, 2012). However, previous studies suggest that counterfactual considerations do not factor into children's judgments of others' counterfactual emotions (*regret* and *relief*) until later in development, at 7 years of age (Ferrell, Guttentag, & Gredlein, 2009; Weisberg & Beck, 2010). Our findings provide evidence for a precursory capacity to reasoning about counterfactual emotions. Our next step is to more directly address children's understanding of alternative states and counterfactual emotions using a similar paradigm as the one used here.

An additional question to consider is exactly how and when children generate relevant alternative states. In our task, children's ability to reason about the physical states of the world was an important actor for generating these alternative states. Even though infants can make sophisticated predictions about physical outcomes (Spelke et al., 1992), children's understanding of the physical world still undergo significant changes in childhood (McCloskey, 1983). Previous work has suggested that relevant alternatives may be more easily generated, and have greater affective impact, the closer the outcome is to the goal (e.g., Ong, Goodman, & Zaki, 2015). We can specifically test this hypothesis in our bowling scenario by manipulating the ball's trajectory: imagine an agent who just barely misses getting a strike versus one who misses by a lot. Further, we can consider the possibilities that an agent's desires and prior experiences might also allow children to easily generate relevant alternative outcomes.

Finally, we note several connections between our results and prospect theory (Kahneman & Tversky, 1979). First, adults and children's ratings suggest that their inferences of others' emotions stem from a sensitivity to gains and losses, as opposed to just the final outcomes. Rather than waiting until the final outcome and making inferences based on the outcome, it is possible that people continuously keep track of people's expectations (and presumably their emotions) as the event unfolds, and dynamically adjust their inferences. It would be interesting to design a continuous measure of inference of affective states (e.g., Zaki, Bolger, & Ochsner, 2008) that we can use with children, to study how children make inferences over an extended period of time.

Our findings show an interesting asymmetry between the amount of change for the low expectation and high expecta-

tion characters; this might be potentially related to the asymmetric discounting of future losses and future gains. At first glance, prospect theory might seem to suggest the opposite; the gain curve is concave, while the loss curve is convex and steeper. However, this refers to people's predictions about their future losses and gains; for improbable outcomes that were unexpectedly realized, it might have exactly the opposite effect as seen in our data. Recent work has started to apply the concept of a naive utility calculus (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015) to understand how we reason about other minds, and it would be interesting to apply a more general notion of utility calculus in understanding how we reason about others' emotions.

In summary, affective cognition is a critical skill for appropriately and proactively responding to others' emotions, intervening on the world and other people to help them feel better, and critically, avoiding actions that would make others feel worse. These are all incredibly important social capacities, and we see impairments of these capacities in various developmental disorders or psychiatric disorders. We hope that our study will inspire more studies on children's affective reasoning and shed light on helping children develop these skills, better understand their own emotions, and build stronger relationships with others.

Acknowledgments

We thank Andrew McCabe and Willow Bailon for help with data collection and coding, Josh Tenenbaum and Jamil Zaki for helpful discussions, and Bing Nursery School, Palo Alto Junior Museum and Zoo, and the families who participated. This work was supported in part by an A*STAR National Science Scholarship to DCO and by a Varieties of Understanding grant from the John Templeton Foundation to HG.

References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.

Barden, R. C., Zelko, F. A., Duncan, S. W., & Masters, J. C. (1980). Children's consensual knowledge about the experiential determinants of emotion. *Journal of Personality and Social Psychology*, *39*(5), 968.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426.

Ferrell, J. M., Guttentag, R. E., & Gredlein, J. M. (2009). Children's understanding of counterfactual emotions: Age differences, individual differences, and the effects of counterfactual-information salience. *British Journal of Developmental Psychology*, *27*(3), 569–585.

Gnepp, J., & Hess, D. L. (1986). Children's understanding of verbal and facial display rules. *Developmental Psychology*, *22*(1), 103.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child development*, *83*(6), 1853–1868.

Hood, B., Cole-Davies, V., & Dias, M. (2003). Looking and search measures of object knowledge in preschool children. *Developmental psychology*, *39*(1), 61.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: an Analysis of Decision under Risk. *Econometrica*, *47*(2), 263-292.

Keen, R. (2003). Representation of objects and events why do infants look so smart and toddlers look so dumb? *Current Directions in Psychological Science*, *12*(3), 79–83.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324).

Ong, D. C., Goodman, N. D., & Zaki, J. (2015). Near-misses sting even when they are uncontrollable. In *Proceedings of the Thirty-Seventh Annual Meeting of the Cognitive Science Society.* (pp. 1775–1780).

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Phillips, A. T., Wellman, H. M., & Spelke, E. S. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, *85*(1), 53–78.

Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological review*, *99*(4), 605.

Weisberg, D. P., & Beck, S. R. (2010). Children's thinking about their own and others regret and relief. *Journal of Experimental Child Psychology*, *106*(2), 184–191.

Weisberg, D. P., & Beck, S. R. (2012). The development of children's regret and relief. *Cognition & emotion*, *26*(5), 820–835.

Wellman, H. M., & Banerjee, M. (1991). Mind and emotion: Children's understanding of the emotional consequences of beliefs and desires. *British Journal of Developmental Psychology*, *9*(2), 191–214.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, *72*(3), 655–684.

Wu, Y., Baker, C., Tenenbaum, J., & Schulz, L. (2014). Joint inferences of belief and desire from facial expressions. In *Proceedings of the Thirty-Sixth Annual Meeting of the Cognitive Science Society.* (pp. 1796–1801).

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology*, *2*(1), 73–81.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: the interpersonal nature of empathic accuracy. *Psychological Science*, *19*(4), 399–404.

Language does not explain the wine-specific memory advantage of wine experts

Ilja Croijmans[1,2]
(i.croijmans@let.ru.nl)

Asifa Majid[1,3,4]
(asifa.majid@let.ru.nl)

1 Centre for Language Studies, Radboud University, Nijmegen, NL
2 International Max Planck Research School for Language Sciences, Nijmegen, NL
3 Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL
4 Max Planck Institute for Psycholinguistics, Nijmegen, NL

Abstract

Although people are poor at naming odors, naming a smell helps to remember that odor. Previous studies show wine experts have better memory for smells, and they also name wine and wine-related smells differently than novices. This leads us to ask whether wine experts' odor memory is verbally mediated? In addition, does the odor memory advantage that experts have over novices generalize to all odors, or is it restricted to odors in their domain of expertise? Twenty-four wine experts and 24 novices smelled wines, wine-related odors and common odors, and were asked to remember these. Critically, half of the participants were asked to name the smells in addition to memorizing them, while the other half just remembered the smells. Wine experts had better memory for wines, but not for wine-related or common odors, indicating their memory is restricted to odors from their domain of expertise. Wine experts were also found to be more consistent and accurate than novices in their descriptions. But there was no relationship between experts' ability to name odors and their memory for odors. This suggests experts' odor memory advantage is not linguistically mediated, but may be the result of differential perceptual learning.

Keywords: expertise, wine experts, olfaction, language, memory, language and thought

Introduction

People from the West are poor at describing smells (Majid, 2015; Olofsson & Gottfried, 2015; Yeshurun & Sobel, 2010). When asked to name smells, they rarely exceed naming more than 50% of them correctly (Cain, 1979; Engen, 1987; Yeshurun & Sobel, 2010; Olofsson & Gottfried, 2015). In contrast, a large body of literature shows remembering smells is relatively easy, as demonstrated by an almost flat forgetting curve. Even after a month, people still recognize around 75% of previously smelled odors correctly (Cain, 1979; Engen, 1987; Herz & Engen, 1996; Schab, 1991).

Given the discrepancy between odor naming and odor memory, surprisingly having the correct label for a smell can increase memory for it. Herz and Engen (1996) concluded "the jury was still out" (p. 303) on whether odor memory is verbally mediated. However, many studies since their review have found language can improve memory for odors and flavors. A number of studies have tested participants' odor memory while providing them with a label generated by the experimenter during the initial presentation of the smell (i.e., during encoding). These studies have shown access to a meaningful label improves memory, and having access to a "veridical" label improves odor memory the most (Cessna & Frank, 2013; Jehl, Royet, & Holley, 1997; Olsson, Lundgren, Soares, & Johansson, 2009; Russell & Boakes, 2011). When participants self-generate a label for odors during encoding, this also leads to improved memory for those odors (Frank, Rybalsky, Brearton, & Mannea, 2011; Lehrner, Glück, & Laska, 1999; Lesschaeve & Issanchou, 1996). This holds true even when people describe complex flavor stimuli such as wines (Fiore et al., 2012; Hughson & Boakes, 2009). There is also some suggestive evidence for a causal role for language in odor memory. When an odor memory task is paired with a verbal interference task, subsequent odor recognition is poorer than when it is paired with a visual interference task (Annett, Cook, & Leslie, 1995; Annett & Leslie, 1996; Perkins & Cook, 1990). Taken together, these studies suggest language plays a critical role in how odors are remembered.

Through years of training and practice, experts acquire theoretical knowledge, perceptual experience, and train their verbal capacities in the domain of their expertise. Becoming an expert has effects on cognition, including memory and language. For example, chess experts have better memory for chess game layouts than novices (e.g., Chase & Simon, 1973; Gobet, 1998), while expert musicians are better able to judge the grammaticality of sentences than novices (Patston & Tippett, 2011).

To become an acknowledged wine expert, a person must study wine for many years. Even after becoming a professional in the field of wine, training and practice continues. An important part of what a wine expert does in their job is to recognize specific aromas and flavors in wine. Wine experts are better than novices at remembering the "flavors" of wine (i.e., a combination of mouth sensation and smell; Melcher & Schooler, 1996), as well as the (orthonasal) smells (i.e., "sniffing") of wines (Zucco, Carassai, Baroni, & Stevenson, 2011) and wine-related odors (Parr, Heatherbell, & White, 2002; Parr, White, & Heatherbell, 2004). So, it seems wine experts have better memory for stimuli salient in their domain of expertise.

For olfaction experts, smells and flavors are said to be more important in their daily life than for the average person (Royet, Plailly, Saive, Veyrac, & Delon-Martin, 2013). However, what consequences expertise has for the language

used to describe smells remains unclear. For example, Lawless (1984) found little agreement in the descriptions of wine from wine experts versus novices, and wine experts are apparently not more consistent than novices in their descriptions of wine-related smells (Parr et al., 2002).

On the other hand, in a recent study Croijmans and Majid (2016) found wine experts were more consistent than novices when describing the smell and flavor of real wines. Wine experts have also been found to be able to select the correct label for odors more often than novices (Bende & Nordin, 1997; Marino-Sanchez et al., 2010; Tempere, Hamtat, de Revel, & Sicard, 2015), and they seem to be more precise in their descriptions for wines and wine-related odors (Chollet & Valentin, 2000; Lehrer, 1983; Melcher & Schooler, 1996; Solomon, 1990, 1997; Zucco et al., 2011). Overall, then, wine experts appear to describe smells from their domain of expertise with more consistency and exactness than novices.

Brown and Lenneberg (1954) pioneered the idea that when a percept is expressed more consistently and concisely in language (i.e., when it is "codable"), it is remembered better. Can wine experts' aptitude for describing smells help their memory for odors? Previous studies do not give a satisfying answer. Melcher and Schooler (1996) found no difference when experts gave a verbal description of wines compared to a non-verbal condition, although experts were better than novices and intermediates in remembering the wines they tasted. Similarly, Parr and colleagues (Parr et al., 2002, 2004) found no significant relationship between the ability to label wine-related odors and their subsequent memory for those odors, although wine experts were again much better at remembering odors than novices. However, when inspecting the results more closely (Parr et al., 2004, Table 2, p. 416), a significant recognition memory difference between experts and novices was only found in the condition where participants labeled the stimuli (instead of rating the pleasantness of odors), leaving open the possibility that wine experts' memory is perhaps verbally mediated.

If language is used by experts to remember wines, we might predict wine experts would only be better at remembering the smells of wines, or perhaps wine-related odors, but not other odors (e.g., common household odors). This is because wine experts appear to be better at naming smells and flavors only when these came from their specific area of expertise (Croijmans & Majid, 2016). If language is used to remember wines, then wine experts should only show a memory advantage for wine (and, perhaps, wine-related odors), but not common odors. They should also be better at naming those odors, and there should be a clear relationship between naming and memory. To test these hypotheses, we asked wine experts and novices to remember odors from real wines, wine-related odors, and common household odors unrelated to wine. One group of participants was asked to name the stimuli (verbal condition), while another group smelled the odors without verbalization (baseline condition).

Methods

Participants

Forty-eight people participated in the experiment. Twenty-four were experts (6 women, M_{age} = 49, SD = 9, age range 29 – 60), and worked as qualified vinologists, sommeliers or wine producers. The other 24 people were novices (6 women, M_{age} = 47, SD = 13, age range 26 – 71). All participants were native speakers of Dutch, and were paid with a €15 voucher.

To confirm expertise, all participants completed a questionnaire assessing their knowledge of wine (following Hughson & Boakes, 2001; Lehrer, 1983; Melcher & Schooler, 1996). A non-parametric Mann-Whitney U test confirmed all wine experts had significantly higher scores than novices, $U = 0.0$, $p < 0.001$, $r = 0.86$.

Half the participants from each group were randomly allocated to the verbal condition, and half to the baseline condition.

Materials

Forty-eight stimuli were used in this study. There were 16 different red and white wines, selected for their distinctiveness. There were 16 wine-related smells from the "*Le nez du vin*" kit (Lenoir, 1995), i.e., aromas that can be found in wine, including wine faults. A further 16 were common household smells, using real objects, ranging from cleaning and beauty products, herbs and spices, to food. All smells were presented in small 30 ml brown screwtop jars. A small tuft of scentless polyester hollow fiber in each jar obscured the object inside so the participant could not see it.

Procedure

Twenty-four of the 48 stimuli (four white and four red wines, eight wine-related smells, and eight common smells; all chosen at random) were presented to the participant in random order during the encoding phase. Participants were informed there were three types of smells (i.e. wines, wine-related smells and common smells). In the verbal condition, participants were instructed to smell the odors and name them as quickly and precisely as they could. In the baseline condition, participants were just asked to remember the odors as best as they could. All participants were told they would be tested for their odor memory later.

After going through the 24 smells in the encoding phase, there was a 10-minute break in which participants completed the wine knowledge questionnaire, and two other questionnaires.

In the recognition phase, participants smelled all 48 stimuli one by one, and told the experimenter for each of the odors whether they had smelled the odors before. They named all the odors too.

Results

Odor memory

We first analyzed participants' recognition memory. Were experts better at remembering odors than novices? In addition, were they better in the verbal condition than in the

baseline? Following detection theory, hits and false alarms were first coded from participants' responses for each odor type. From these values, *d'* was calculated (Macmillan & Creelman, 1991; Parr et al., 2002). The larger the *d'*, the better the participants were able to distinguish between old and new odors. A three-way mixed ANOVA with expertise (wine expert vs. novice) and condition (verbal vs. baseline) as between-participant factors, and odor type (wine, wine-related, vs. common) as a within-participant factor was conducted, with *d'* as the dependent variable.

Experts were no better at remembering odors overall than novices, F (1, 44) = 0.64, p = .427, η_p^2 = .01, and naming the odors during encoding did not improve memory, F (1, 44) = 0.46, p = .503, η_p^2 = .01. There was, however, a main effect of odor type. Common odors and wine-related odors were remembered better than wines, F (2, 88) = 41.53, p < .001, η_p^2 = .49. There was no significant interaction between expertise and verbal condition, F (1, 44) = 1.40, p = .244, η_p^2 = .03, and no significant three-way interaction between verbal condition, expertise, and odor type F (2, 88) = 0.92, p = .402, η_p^2 = .02. So it appears experts do not benefit any more from verbalizing smells than novices.

There was no overall effect of expertise on odor memory, but we had specifically predicted experts' memory for odors would be restricted to smells in their domain of expertise (i.e., to wines and wine-related odors). The interaction between expertise and odor type was not significant, F (2, 43) = 2.35, p = .108, η_p^2 = .10, but Bonferroni corrected pairwise comparisons (as recommended by Hsu, 1996) demonstrated wine experts were significantly better at remembering the smells of wines (M = 0.23, SD = 0.56) than novices (M = -0.13, SD = 0.57), p = .036, d = .63. This was not true for wine-related smells, for which wine experts did not differ significantly from (M = 1.1, SD = 0.72) novices (M = 1.3, SD = 0.75), p = .394, nor was there a significant difference between experts and novices for common household smells (wine experts, M = 1.2, SD = 0.60, vs. novices, M = 1.0, SD = 0.97), p = .548.

Overall, these analyses show wine odors were particularly hard to remember for both experts and novices. Nevertheless, wine experts were better than novices at remembering whether they had previously smelled a specific wine. The results also suggest naming odors explicitly during encoding did not help improve odor memory, consistent with the suggestion experts' superior memory for wine odors is not verbally mediated. To test this proposal more directly, we conducted additional analyses examining the relationship between naming and memory.

Odor naming

Previous studies with novices suggest remembering an odor successfully depends on the ability to name that odor accurately and consistently (e.g., Cessna & Frank, 2013; Fiore et al., 2012; Frank, Rybalsky, Brearton, & Mannea, 2011). Therefore, we first examined the naming data more closely. Half the participants named odors twice during the experiment (verbal condition); i.e., during encoding and then again during recognition. The answers for those participants were coded for consistency (i.e., did they give the same label during encoding and recognition?). In addition, the labels of all participants during the recognition phase were coded for accuracy. An answer was considered correct if participants gave the same answer as the pre-determined "veridical" label. For wines it was considered correct if participants gave the correct color, grape type, or production country. Coding was done by the experimenter and one independent researcher.

A two-way mixed ANOVA with expertise and odor type as factors, and percentage of consistently named odors as the dependent variable, showed wine experts (M = 47.5, SD = 23.8) gave more consistent labels than novices (M = 29.3, SD = 19.0), F (1, 44) = 12.24, p = .002, η_p^2 = .36. There was also a main effect of odor type, F (2, 44) = 16.37, p < .001, η_p^2 = .42. Bonferroni corrected pairwise comparisons showed common odors (M = 46.2, SD = 23.9) were more consistently named than wine odors (M = 18.2, SD = 21.2), p = .018, d = .8. Wine-related odors (M = 50.9, SD = 24.8) were also more consistently named than wine odors, p = .009, d = 1.0. There was no difference between common odors and wine-related odors, p = .332. There was no interaction between expertise and odor type, F (2, 44) = 0.82, p = .448, η_p^2 = .04, but Bonferroni corrected pairwise comparisons showed wine experts gave more consistent answers for wine odors (M = 28.1, SD = 21.4) than novices (M = 8.3, SD = 16.3), p = .018, d = 1.0. Wine experts also gave more consistent answers for wine-related odors (M = 63.5, SD = 24.1) than novices (M = 38.3, SD = 19.0), p =.009, d = 1.2. But there was no difference between wine experts (M = 51.0, SD = 25.8) and novices (M = 41.4, SD = 21.8), p = .332, for common odors.

This analysis was also repeated with percentage of correctly named odors as the dependent variable. Wine experts (M = 33.0, SD = 17.0) named more odors correctly than novices (M = 24.8, SD = 19.2), F (1, 46) = 4.91, p = .032, η_p^2 = .10, and wines were more often correctly named than wine-related odors or common odors, F (2, 45) = 5.7, p = .006, η_p^2 = .20. There was no interaction between expertise and stimulus type, F (2, 92) = 1.6, p = .217, η_p^2 = .03, but Bonferroni corrected pairwise comparisons showed wine experts (M = 44.8, SD = 25.4) were more accurate in naming wines than novices (M = 29.7, SD = 26.1), p = .048, d = .59. This was not the case for wine-related odors (wine experts M = 26.8, SD = 12.7 vs. novices M = 20.3, SD = 14.4), p = .104, d = .48 or common odors (wine experts M = 27.6, SD = 13.0 vs. novices M = 24.5, SD = 17.1), p = .480, d = .20. In line with the results for consistently named odors, wine experts had a domain-specific advantage for accurately naming wine odors, but not other types of odors.

To summarize, experts were more consistent and accurate than novices when naming wine odors. In addition, experts were also more consistent (but not more accurate) than novices when naming wine-related odors. Odor naming for common odors was comparable across experts and novices.

Relationship between odor memory and odor naming

So, was naming consistency and accuracy related to experts' superior memory for wine odors? To test this, the correlations between *d'* and naming consistency, and *d'* and naming accuracy were calculated for the different stimuli, for wine experts and novices separately (see Table 1).

The correlation analyses replicated the previously established finding (e.g., Cessna & Frank, 2013) that novices have better odor memory for common odors they named correctly, $r = .523$, $p = .004$. They also have better memory for wine-related odors they named correctly, $r = .518$, $p = .005$. No other correlations were significant for this group.

For wine experts, there was a similar trend of better memory for correctly named common odors $r = .309$, $p = .071$, but this was not significant. Similarly, there was a positive correlation between naming accuracy and memory for wine-related odors, $r = .463$, $p = .011$. Critically, memory for wine odors and naming consistency and accuracy were not positively correlated (see Table 1); in fact, if anything, memory for wines and naming accuracy had a small (insignificant) negative relationship. Taken together, these results suggest the superior memory for wine odors displayed by wine experts is not verbally mediated, even though they seem to remember wine-related odors and common odors by their names, just like novices.

Table 1: Correlations between odor memory and naming consistency and accuracy for wine experts and novices (Pearson's correlation coefficients, reported *p*-values are one-tailed)

		Wine experts		Novices	
Naming	Odor type	*r*	*p*	*r*	*p*
Naming consistency	Wines	.041	.449	-.320	.155
	Wine-related odors	.381	.111	.130	.343
	Common odors	.385	.108	.328	.149
Naming accuracy	Wines	-.151	.241	-.036	.433
	Wine-related odors	.463	.011	.523	.004
	Common odors	.309	.071	.518	.005

Discussion

Wine experts were better than novices at remembering the odor of wines. This replicates a number of previous studies (Hughson & Boakes, 2009; Melcher & Schooler, 1996), and corroborates the proposal that wine experts superior memory for odors is restricted to wines (Zucco et al., 2011). Wine experts were no better than novices at remembering common household odors, or even wine-related odors.

We also found wine experts were more consistent and accurate than novices when naming wines, and this extended to wine-related odors. This is an interesting finding in its own right. Previous studies suggest wine experts' descriptions are highly idiosyncratic (Lawless, 1984), and experts are no more consistent in their descriptions for wine-related smells than novices (Parr et al., 2002, 2004). In this study, experts were more accurate and consistent than novices, but only for the odors from their domain of expertise. This replicates the finding of Croijmans and Majid (2016) that wine experts name wine odors, but not common odors, more consistently than novices. Note, the current study adds a new dimension to the issue. Croijmans and Majid (2016) found wine experts as a group were more consistent with each other in how they talk about wine odors than novices, whereas the current study shows experts are also more consistent with themselves when they name the same wine odors at two different times (cf. Brown & Lenneberg, 1954).

Since language influences odor memory amongst novices (Fiore et al., 2012; Frank et al., 2011; Lehrner et al., 1999; Lesschaeve & Issanchou, 1996), we asked whether it was also the basis of wine experts' superior memory for wine odors. But there was little evidence in support of this proposal. Experts were no better at remembering wine odors in the verbal condition than the baseline (i.e., non-verbal) condition. Moreover, there was no significant correlation between odor memory for wines and odor naming accuracy or naming consistency.

To recap, wine experts were only better at remembering the odors of wines, but they were more consistent and accurate in both their descriptions of wines *and* wine-related odors. During wine experts' training, they are taught to identify and name different aroma components of wines (Herdenstam, Hammarén, Ahlström, & Wiktorsson, 2009; Lehrer, 1983). This is exactly the sort of skill the *Le nez du vin* kit (Lenoir, 1995) is designed to help with. So, the greater consistency for naming wine-related odors probably comes from specific training to identify wine-related odors.

Wines are complex. They can contain up to 800 different volatiles (Ortega-Heras, González-SanJosé, & Beltrán, 2002). When remembering the odors of wines, experts appear to remember the whole gestalt rather than the individual components of a wine. An analogy can be made to memory for faces. Humans are excellent at remembering faces, yet perform poorly when having to recall individual features of faces, such as a nose, eye, or mouth (Tanaka & Farah, 1993). Expert memory for wines, similarly, appears to be holistic instead of featural.

Another analogy can be made to chess experts. Chess experts are better at remembering the layout of chess plays than novices (Frey & Adesman, 1976; Gobet & Simon, 1996). However, these layouts have to be possible configurations that are encountered during real chess games. Experts are no better than novices at remembering randomly assembled layouts (Frey & Adesman, 1976; Gobet, 1998). This suggests chess experts have learned to remember particular configurations of arrays.

These analogies suggest wine experts are perceptually processing wines in a different way to novices (e.g., Hughson and Boakes, 2009). As wine students become experts, it seems a shift occurs which encourages the holistic processing of wines. Aside from the perceptual gestalt, this representation likely includes knowledge about the specific wine from its particular region made from

particular grapes (Solomon, 1997). Wine-related odors, when presented out of context, seem to be processed similarly to common odors. This is exemplified by the relationship between language and memory for wine-related and common odors in both experts and novices; and also the absence of a wine-related odor memory advantage amongst wine experts.

The results from the current study suggest language is not directly involved in remembering wine odors. However, we cannot rule out the possibility that experts and novices recruited language sub-vocally during the baseline task: i.e., they could have been naming the odors silently. A better way to test whether odor memory is verbally mediated would be to test if experts are still better at remembering wine odors when performing a simultaneous verbal interference task. Chess grandmasters are apparently not hindered in their memory for chess board positions if they perform a verbal task at the same time as encoding a chess board layout (Robbins et al., 1996). However, novices' memory for odors is harmed if they have to perform an odor memory task simultaneously with a verbal interference task (e.g. Annett & Leslie, 1996). This raises the interesting question of what would happen with wine experts if they have to remember wine odors under verbal interference. If wine experts' odor memory is truly not verbally-mediated, then experts' odor memory should not be harmed by a verbal interference paradigm. If, on the other hand, they are using language (just like novices with common odors), their advantage for remembering wine odors should disappear.

In conclusion, our results show wine experts are better than novices at remembering wine odors, and at describing wines and wine-related odors. However, the results show little evidence that these two aspects of expert cognition are related. That is, wine experts' better memory for wine odors does not seem to be based on their ability to name wine odors.

Acknowledgments

This work was funded by The Netherlands Organization for Scientific Research: NWO VICI grant "Human olfaction at the intersection of language, culture and biology". Special thanks to Josje de Valk, Artin Arshamian and Laura Speed and all lab members, as well as all the participants involved in this study.

References

Annett, J. M., Cook, N. M., & Leslie, J. C. (1995). Interference with olfactory memory by visual and verbal tasks. *Perceptual and Motor Skills*, *80*, 1307–1317.

Annett, J. M., & Leslie, J. C. (1996). Effects of visual and verbal interference tasks on olfactory memory: The role of task complexity. *British Journal of Psychology*, *87*(3), 447–460.

Bende, M., & Nordin, S. (1997). Perceptual learning in olfaction: Professional wine tasters versus controls. *Physiology & Behavior*, *62*(5), 1065–1070.

Brown, R. W., & Lenneberg, E. H. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology*, *49*(3), 454–462.

Cain, W. S. (1979). To know with the nose: Keys to odor identification. *Science*, *203*, 467–470.

Cessna, T. C., & Frank, R. A. (2013). Does odor knowledge or an odor naming strategy mediate the relationship between odor naming and recognition memory? *Chemosensory Perception*, *6*(1), 36–44.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Chollet, S., & Valentin, D. (2000). Expertise level and odour perception: What can we learn from red burgundy wines? *Annee Psychologique*, *100*(1), 11–36.

Croijmans, I., & Majid, A. (2016). Not all flavor expertise is equal: The language of wine and coffee experts. *PLoS ONE: e0155845*. DOI: 10.1371/journal.pone.0155845.

Engen, T. (1987). Remembering odors and their names. *American Scientist*, *75*(5), 497–503.

Fiore, F., Cornoldi, C., De Beni, R., D'Urso, V., Eilertsen, D.-E., & Magnussen, S. (2012). Short-term memory for flavour. *Journal of Cognitive Psychology*, *24*(2), 134–139.

Frank, R. A., Rybalsky, K., Brearton, M., & Mannea, E. (2011). Odor recognition memory as a function of odor-naming performance. *Chemical Senses*, *36*(1), 29–41.

Frey, P. W., & Adesman, P. (1976). Recall memory for visually presented chess positions. *Memory & Cognition*, *4*(5), 541–547.

Gobet, F. (1998). Expert memory: A comparison of four theories. *Cognition*, *66*(2), 115–152.

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology*, *31*(1), 1–40.

Herdenstam, A. P. F., Hammarén, M., Ahlström, R., & Wiktorsson, P.-A. (2009). The professional language of wine: Perception, training and dialogue. *Journal of Wine Research*, *20*(1), 53–84.

Herz, R. S., & Engen, T. (1996). Odor memory: Review and analysis. *Psychonomic Bulletin & Review*, *3*(3), 300–313.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and methods*. London: Chapman & Hall.

Hughson, A. L., & Boakes, R. A. (2001). Perceptual and cognitive aspects of wine expertise. *Australian Journal of Psychology*, *53*(2), 103–108.

Hughson, A. L., & Boakes, R. A. (2009). Passive perceptual learning in relation to wine: Short-term recognition and verbal description. *The Quarterly Journal of Experimental Psychology*, *62*(1), 1–8.

Jehl, C., Royet, J. P., & Holley, A. (1997). Role of verbal encoding in short and long-term odor recognition. *Perception & Psychophysics*, *59*(1), 100–110.

Kellman, P. J., & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, *6*(2), 53–84.

Lawless, H. T. (1984). Flavor description of white wine by 'expert' and nonexpert wine consumers. *Journal of Food Science*, *49*(1), 120–123.

Lehrer, A. (1983). *Wine and conversation*. Oxford: Oxford University Press.

Lehrner, J. P., Glück, J., & Laska, M. (1999). Odor identification, consistency of label use, olfactory threshold and their relationships to odor memory over the human lifespan. *Chemical Senses*, *24*(3), 337–346.

Lenoir, J. (1995). *Le nez du vin*. Editions Jean Lenoir, France.

Lesschaeve, I., & Issanchou, S. (1996). Effects of panel experience on olfactory memory performance: Influence of stimuli familiarity and labeling ability of subjects. *Chemical Senses*, *21*(6), 699–709.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd edition). Mahwah, New Jersey; London: Lawrence Erlbaum Associates.

Majid, A. (2015). Cultural factors shape olfactory language. *Trends in Cognitive Sciences*, *19*(11), 629–630.

Marino-Sanchez, F. S., Alobid, I., Centellas, S., Alberca, C., Guilemany, J. M., Canals, J. M., … Mullol, J. (2010). Smell training increases cognitive smell skills of wine tasters compared to the general healthy population. The WINECAT Study. *Rhinology*, *48*(3), 273–276.

Melcher, J. M., & Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, *35*(2), 231–245.

Olofsson, J. K., & Gottfried, J. A. (2015). The muted sense: Neurocognitive limitations of olfactory language. *Trends in Cognitive Sciences*, *19*(6), 314-321.

Olsson, M. J., Lundgren, E. B., Soares, S. C., & Johansson, M. (2009). Odor memory performance and memory awareness: A comparison to word memory across orienting tasks and retention intervals. *Chemosensory Perception*, *2*(3), 161–171.

Ortega-Heras, M., González-SanJosé, M. ., & Beltrán, S. (2002). Aroma composition of wine studied by different extraction methods. *Analytica Chimica Acta*, *458*(1), 85–93.

Parr, W. V., Heatherbell, D., & White, K. G. (2002). Demystifying wine expertise: Olfactory threshold, perceptual skill and semantic memory in expert and novice wine judges. *Chemical Senses*, *27*, 747–755.

Parr, W. V., White, K. G., & Heatherbell, D. A. (2004). Exploring the nature of wine expertise: What underlies wine experts' olfactory recognition memory advantage? *Food Quality and Preference*, *15*(5), 411–420.

Patston, L. L. M., & Tippett, L. J. (2011). The effect of background music on cognitive performance in musicians and nonmusicians. *Music Perception: An Interdisciplinary Journal*, *29*(2), 173–183.

Perkins, J., & Cook, N. M. (1990). Recognition and recall of odours: The effects of supressing visual and verbal encoding processes. *British Journal of Psychology*, *81*, 221–226.

Robbins, T. W., Anderson, E. J., Barker, D. R., Bradley, A. C., Fearnyhough, C., Henson, R., … Baddeley, A. D. (1996). Working memory in chess. *Memory & Cognition*, *24*(1), 83–93.

Royet, J.-P., Plailly, J., Saive, A.-L., Veyrac, A., & Delon-Martin, C. (2013). The impact of expertise in olfaction. *Frontiers in Cognitive Science*, *4:*, 928.

Russell, A. M. T., & Boakes, R. A. (2011). Identification of confusable odours including wines: Appropriate labels enhance performance. *Food Quality and Preference*, *22*(3), 296–303.

Schab, F. R. (1991). Odor memory: Taking stock. *Psychological Bulletin*, *109*(2), 242–251.

Solomon, G. E. (1990). Psychology of novice and expert wine talk. *The American Journal of Psychology*, *103*(4), 495–517.

Solomon, G. E. (1997). Conceptual change and wine expertise. *Journal of the Learning Sciences*, *6*(1), 41–60.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *46*(2), 225–245.

Tempere, S., Hamtat, M.-L., de Revel, G., & Sicard, G. (2015). Comparison of the ability of wine experts and novices to identify odorant signals: A new insight in wine expertise. *Australian Journal of Grape and Wine Research*. doi: 10.1111/ajgw.12192

Yeshurun, Y., & Sobel, N. (2010). An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. *Annual Review of Psychology*, *61*, 219–241.

Zucco, G. M., Carassai, A., Baroni, M. R., & Stevenson, R. J. (2011). Labeling, identification, and recognition of wine-relevant odorants in expert sommeliers, intermediates, and untrained wine drinkers. *Perception*, *40*(5), 598 – 607.

Sub-Categorical Properties of Stimuli Determine the Category-Order Effect

Jordan Schoenherr (psychophysics.lab@gmail.com)
Department of Psychology, Carleton University
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

Robert Thomson (robert.thomson@usma.edu)
Army Cyber Institute, United States Military Academy
2101 New South Post Road, West Point, NY, 10996

Abstract

The category-order effect (COE) is observed when the categorical properties of items within the first half of a given list affect recall performance in a mixed-list serial-recall task. The present study examines whether the advantage is due to other sub-categorical properties (e.g., orthographic similarity and word frequency) rather than an artifact of stimuli used in previous studies (e.g., numbers vs. nouns). Participants were presented with numeric stimuli and nouns from a variety of semantic categories while their orthography and word frequency were systematically manipulated. The results suggest that a large portion of the COE can be attributed to the sub-categorical properties of the items.

Keywords: memory, category-order effect, recall

Introduction

In serial recall tasks, item information activates both short-term (e.g., order) and long-term (e.g., semantic) memory components (e.g., Healy, 1974; Lashley, 1951; Nairne & Kelley, 2004). These components affect the extent to which participants can rehearse and retrieve to-be-recalled items. In the present study, we examined the contribution of long-term memory representations by varying stimulus properties associated with category membership (e.g., semantic similarity) as well as those that are associated with individual stimuli (e.g., word frequency).

The Category-Order Effect

Brooks and Watkins (1990) referred to the dependency of recall performance on the order in which items from a particular category are presented in a list as the *category-order effect* (or COE). Category-order effects have been established across several word categories, including when higher-frequency words precede lower-frequency words (Watkins & Watkins, 1977), semantically-related words precede semantically-unrelated words, rhyming words precede non-rhyming words (Brooks & Watkins, 1990), and when numeric digits precede words (Greene & Lasek, 1994; Brooks & Watkins, 1990). Greene and Lasek concluded that the COE arises when a relatively smaller, homogenous category is presented before a relatively larger, heterogeneous category. Given that category set-size has proven to determine recall accuracy and response times, this appears to be a reasonable set of assumptions.

There is evidence that the category effect can be mediated by other factors such as rehearsal strategy. For instance, the COE can be eliminated by decreasing the presentation rate of the stimuli or by using articulatory suppression (Greene & Lasek, 1994). Greene and Lasek (1994) further explored whether the order of recall of the stimuli (i.e. input or output position) was a factor in the category-order effect. To test their hypothesis, participants were asked to either recall items in forward or backward order. In the backward condition, items were to be recalled in reverse order that they were presented in. Although recall order did affect overall accuracy and memory span, both forward and backward recall exhibited a significant input position COE, with improved full-list recall when the more readily categorized items were presented in the first half of the list.

In a series of experiments directed toward examining the subcategorical properties of the COE, Schoenherr and Thomson (2008) also found the effect using both successive and simultaneous split-list presentation. In their study, lists consisted of four non-repeating letters and four non-repeating numbers. Letters were from one of four categories: 1) four letter words, 2) four letter pseudowords (CVCVs), 3) four rhyming letter nonwords, or 4) four random letter sets. Their results indicated that a COE was exhibited when stimuli with more sub-categorical properties (i.e., letters making up words or pseudowords) were presented before stimuli that had fewer subcategorical properties (i.e., random letters, rhyming letters, and numbers).

Most interestingly, Schoenherr and Thomson conducted a serial-order analysis comparing within-category accuracy based on input position, and found evidence that only the accuracy for the category presented first improved rather than both presented categories, and only for the word and pseudoword categories. Further support for the contributions of subcategorical properties to the COE was evidenced in results from an oscillator-based connectionist model (Schoenherr & Thomson, 2009) wherein the manipulation of a distinctiveness parameter produced COE-like effects. Thus, while categorical information no doubt contributes to recall, a COE-like effect can be obtained by means of the manipulation of subcategorical information.

Semantic Similarity and Recall

Semantic similarity could help explain the COE as evidenced in the literature on immediate recall. For instance, Crowder (1979) presented participants with 10 item lists consisting of similar or dissimilar words. He observed

enhanced recall for similar words relative to dissimilar ones (see also Saint-Aubin, Ouellette, & Poirier, 2005). Moreover, research has also found that the probability of correctly recalling an item is inversely proportional to the number of associates for that item in a given list (for a review, see Nelson, 1989). A straightforward explanation of these findings is that once a word is presented, it activates both its representation and similar stimuli (with overlapping semantic or contextual information) in long-term memory. As activation increases, semantically related units also become active thereby becoming candidates for recall, creating proportionally higher interference for items with a greater category size or items seen in more contexts than items with a smaller category size; known as the fan effect (Anderson & Reder, 1999; West, Pyke, Rutledge-Taylor, & Lang, 2011).

Earlier evidence of the fan effect was observed by Crannell and Parrish (1957). They examined the effect of set sizes and semantic categories on immediate serial recall. More specifically, participants were asked to remember sets consisting of digits (1-9; set size = 9), letters (from the full set of letters), letters (from a limited set of letters 'a' to 'i'), three-letter words (from a set comprised of 286 members), and three-letter words (from a set comprised of 9 members). Overall, Crannell and Parrish found that digits led to the highest recall and that letters were recalled more accurately than words. Thus, it appears that the category set size explains only part of the advantage for digits relative to letters and words in general and that another stimulus property stored in long-term memory contributes to the recall advantage.

Sub-Categorical Properties and Recall

Despite the evidence for recall facilitation resulting from the availability information stored within long-term memory, it is not necessarily the case that this information need be categorical (West et al., 2011). Sub-categorical information can be used as the basis for grouping exemplars. As Howes and Solomon (1951) first observed, the frequency of words in a language corpus is negatively correlated with their response threshold. McGinnies, Comer, and Lacey (1952) later demonstrated that word length affects performance independently of word frequency (see also Postman & Adis-Castro, 1957). This finding is of considerable importance for COE experiments considering that word stimuli used in those studies (e.g., goose, dog, sheep, ox) have more characters than digit stimuli (e.g., 1, 6, 9) which has previously been shown to affect recall (Cowan et al., 1992). This could explain Schoenherr and Thomson's (2008/9) results whereby letters forming words or pseudowords exhibited relatively higher recall rates when compared against numeric digits.

Moreover, studies have also found that orthographic properties affect recall performance. For example, it has been found that lowercase words are reported more accurately than uppercase words (e.g., Perri et al., 1996; Jordan, Redwood, & Patching, 2003; cf. Smith et al., 1969). Given that words are reported in lowercase type set and numbers are presented as Indian-Arabic numerals, this difference might enhance recall performance resulting in an additional release from proactive interference solely from orthographic properties.

Present Study

Previous studies of the category-order effect and related phenomena have frequently used number-word lists as stimuli (Brooks & Watkins, 1990; Greene & Lasek, 1994; Young & Supa, 1941). Although this paradigm has been viewed as contrasting a small, homogeneous set of items (i.e., numbers) against a larger, heterogeneous set of items (i.e., animals), several properties of the stimuli prohibit such a direct interpretation. First, number stimuli have been presented as digits resulting in a decreased load during visual encoding of a single item (e.g., 4) relative to an equivalent word (e.g., four). Second, number stimuli have a higher frequency than the word stimuli used in the recall lists. Examining these values reveals that within the Brysbaert and New Corpus[1] there is a greater occurrence for the number words (WF_{BN} = 4.10) than animal words (WF_{BN} = 3.12) used in previous studies. Given this potential methodological confound, the present experiments use an immediate serial recall task to assess the effect of item (e.g., word frequency; orthography) and order (e.g., number-word vs. word-number) information on recall performance.

Experiment 1A

In Experiment 1A, we sought to replicate the Category-Order Effect with the materials used by Brooks and Watkins (1990) while also examining the effects of orthographic properties (i.e., letter case; numeric digits vs. words) on encoding.

Method

Participants

Twenty-three Carleton University students participated in the experiment receiving 1% toward their final grade in an introductory psychology class.

Materials

Stimuli consisted of monosyllabic numbers and animal names (nouns) used in previous studies of the COE (Brooks & Watkins, 1990; Greene & Lasek, 1994; Young & Supa, 1941). These include the related animal words *dog, horse, goose, cow, cat, ox, hen, pig and sheep*, in addition to the numeric digits 0-9 excluding the number 7 and their word equivalents (e.g., *zero, one*). Words were presented in both lowercase and uppercase formats.

Procedure

Similar to the procedure of Schoenherr and Thomson (2008), participants were told that an eight-item sequence of four letters and four numbers would be presented on the computer monitor. Each item was presented for 1s and was

[1] Word counts retrieved November, 20, 2011. WF_{BN} is given by adding one and taking the log10 of the SUBTLx word count.

immediately replaced with the subsequent item in the list. After all items were presented, the screen was cleared and a response cue of either FORWARD or BACKWARD was presented following a 250 ms inter-stimulus interval. If the cue indicated FORWARD, participants were instructed to write the items down in the order that they were perceived. Alternatively, if the cue indicated BACKWARD then participants were required to respond with the order of the categories reversed while preserving the order of the items within the category (see Figure 1). This was similar to the procedure used by Greene and Lasek (1994) to preserve within-category integrity between study and recall.

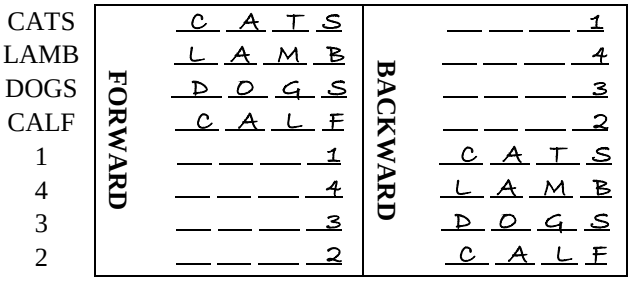


Figure 1. Sample of FORWARD and BACKWARD response for stimuli from Experiment 1.

Instructions emphasized both speed and accuracy. Participants completed 20 training trials and 48 experimental trials. No feedback was provided after training.

Results

A mixed ANOVA with 2 Category-Order (number-word vs. word-number) x 2 Number Format (digits vs. nouns) repeated-measures factors were conducted on the proportion of items recalled, with 2 Recall Order (forward vs. backward) x 2 Letter Case (uppercase vs lowercase) as between-subjects factors. There was no main effect of Letter Case on accuracy, nor was it present in any higher interaction. As such, the Letter Case factor was collapsed.

Replicating the findings of earlier studies, we obtained what has traditionally been the explanation of the COE, $F(1, 21) = 5.704$, $MSE = .013$, $p = .026$. Participants recalled more items in Number-Word lists ($M = .6812$, SD = .1806) than in Word-Number lists ($M = .6403$, $SD = .1706$). Importantly, we also observed a significant effect of Number Format, $F(1, 21) = 18.783$, $MSE = .006$, $p < .001$. Supporting our hypothesis that encoding fluency affected recall, we found that participants recalled more items in number lists written as digits ($M = .6863$) than number lists written as words ($M = .6352$). This suggests that the COE is influenced by sub-categorical properties of the stimuli such as orthography rather than solely categorical information.

Discussion

While we were able to replicate the category-order effect seen in prior literature (Brooks & Watkins, 1990; Greene & Lasek, 1994), we also identified a significant influence of sub-categorical properties (namely the orthography of numerical digits vs words) on recall performance. That said, the lack of a significant interaction between Category-Order and Number Format implies that the effect of sub-categorical properties on the COE might be limited.

Experiment 1B

Experiment 1B further examined the influence of word frequency on the COE. Experiment 1A found that number format influenced recall accuracy, however this only had a limited influence on the COE. However, number stimuli are represented at a higher frequency than word stimuli. If the COE was due to sub-categorical properties such as word frequency (rather than categorical properties such as categorical similarity) then one would predict that the COE might be largely attenuated or completely eliminated by controlling for word frequency. As such, lists of high- and low-frequency words were created from two categories based on word norms and these were paired with high- (e.g., *one* and *two*) and low-frequency (e.g., *twenty* and *thirty*) numbers printed as words.

Method

Participants

Twenty-two Carleton University students participated in the experiment receiving 1% toward their final grade in an introductory psychology class.

Materials

Two sets of high- and low-frequency words were created to exclude the possibility that one semantic category was more salient than another. One high frequency set consisted of terms pertaining to familial relatives (*son, mom, father, aunt, uncle, sister, brother*) whereas the other consisted of terms pertaining to units of time (*second, minute, hour, day week, month, year*). One low frequency set consisted of colour terms (*maroon, aqua, violet, tan, grey, purple, orange*) whereas the other consisted of items used in carpentry (*nail, wrench, wood, ruler, hammer, drill, screw*). This resulted in high- ($M = 4.10$) and low-frequency ($M = 2.62$) word sets.

Two sets of numbers were also created, high-frequency ($M = 4.10$; *two, three, four, five, six, eight, nine*) and low-frequency ($M = 2.64$; *twenty, thirty, forty, fifty, sixty, eighty, ninety*). The number one and ten were removed from the list to balance number-word frequency with category-word frequency. An important feature of the present study is that by constraining the items in the word list, the total set-size of the items available for participants to recall is reduced while category set-size increases for numbers relative to Experiment 1A. Sets were matched for mean word length.

Procedure

Experiment 1B replicated the procedure of Experiment 1A.

Results

Data from one participant was removed prior to analysis for failing to conform to task demands. A repeated-measures 2 Category-Order (number-word vs. word-number) x 2 Word Frequency (high-frequency vs low-frequency) x 2 Number

Frequency (high-frequency vs low-frequency) ANOVA was conducted on the proportion of items recalled with Recall Order (forward vs. backward) included as a between-subjects measure.

The most revealing effect of this study was the lack of an effect of Category-Order alone, $F(1,17) = 1.007$, $MSE = .009$, $p = .33$, nor was Category-Order significant in any interactions. This suggests that the COE might in fact be a result of sub-categorical properties that were not controlled in previous studies, such as the aforementioned set-size and word frequency. Supporting this interpretation, we obtained a significant main effect of Number Frequency, $F(1,17) = 27.63$, $MSE = .016$, $p < .001$. More items were recalled in lists containing high-frequency numbers ($M = .707$) relative to low-frequency numbers ($M = .598$).

In addition, we obtained a significant interaction between word frequency and recall order, $F(1,17) = 5.135$, $MSE = .003$, $p = .037$. Table 1 demonstrates that recall performance was highest (and significantly different, only) for forward recall. This suggests that items that have numerous traces stored in long-term memory benefit from relatively early activation as a result of presentation (i.e., input) order.

Table 1. Mean proportion recalled and standard error in forward and backward orders for high and low word frequency conditions.

Recall Order	Word Frequency	Proportion Recall
Forward	High	.695 (.053)
	Low	.662 (.052)
Backward	High	.623 (.062)
	Low	.630 (.061)

Interestingly, one effect that was not found in contrasts was that when high-frequency items were in the first input positions (e.g., high-frequency numbers followed by low-frequency words, or high-frequency words followed by low-frequency numbers) there was still no equivalent to a category-order effect ($M = .663$ for High-Low vs $M = .658$ for Low-High).

Discussion

Experiment 1B also provides evidence for an explanation of the category-order based on subcategorical properties. One reason for this lack of effect might be that only number words were used rather than digits, which, as we observed in Experiment 1A, provided the strongest COE. The increased performance for higher frequency words (regardless of whether they represented categories such as numbers, tools, or units of time) is consistent with the view that category-order effects were not necessarily due to category set-size or other categorical properties (e.g., similarity), but instead were driven in part by sub-categorical properties (orthography and word frequency). However, by allowing word length to vary, Experiment 1B might have eliminated another important orthographic feature associated with the expression of the COE.

Experiment 2

To better understand what categorical properties might underlie the COE, in the present study we control for category-relatedness and set-size, in addition to using numeric digits and number words as in Experiment 1A to replicate previous COE studies. One aspect of previous studies examining the category-order effect for numbers and words is that the categories used for comparison were very large and very small (animals vs numbers, respectively). In the present study, we identified natural categories used in the categorization literature (e.g., birds, tools) and used latent-semantic analysis (Dumais, 2004) to determine which words within these categories were highly (and equivalently, when possible) related.

Moreover, the fact that multiple lists were used from different categories should ensure long-term memory activation does not change considerably for any given list. To avoid the influence of the word-length effect, we only used four letter words. This procedure allowed us to create more principled categories and stimulus sets. Words were then divided into exemplar lists (e.g., *owls, hawk*) from the same category and related words (e.g., *wing, claw, beak*).

If category size facilitates performance in the initial positions of a list, both number words and number digits should produce highest recall performance when in the initial list positions. Similarly, if frequency of exposure to the stimuli facilitates performance in the initial positions of a list, then number stimuli regardless of format should produce highest recall performance.

Method

Participants

Thirty-two Carleton University students participated in the experiment receiving 1% toward their final grade in an introductory psychology class.

Materials

Stimuli sets were created by selecting six sets of words of similar frequencies ($M = 2.5$) using the same calculation as Experiment 1B, and inter-category relatedness ($M = .45$) using LSA co-occurrence compared to category label (*birds*, *fish*, *clothing*, *tools*, *trees*, and *vehicles*). For instance, from the set *bird* (features) was *peck, sing, flew, claw*; while *bird* (examples) was *jays, hawk, duck, kiwi*. The same number stimuli were used as in Experiment 1A.

Procedure

Experiment 2 replicated the procedure of Experiment 1A with the following differences. Participants completes both backwards and forwards recall conditions in separate blocks.

Results

Data from one participant was removed prior to analysis for failing to conform to task demands. A repeated-measures 2 Category-Order (number-word vs. word-number) x 2 Number Format (number digit vs word) x 2 Recall Order (forward vs. backward) was performed.

Unlike Experiment 1A but similar to Experiment 1B there was no main effect of Category-Order, $F(1, 30) = .034$, $MSE = .034$, $p = .855$, but similar to Experiment 1A there was a main effect of Number Type, $F(1, 30) = 39.07$, $MSE = .019$, $p = .001$. Again, lists containing numeric digits were recalled with higher accuracy than lists containing number-words ($M = .634$ vs $M = .557$). As seen in Figure 2, the interaction of Category-Order and Number Format was significant $F(1, 30) = 5.438$, $MSE = .018$, $p = .027$. While not significant, it was interesting that there was higher overall accuracy for the Word-Number condition with numeric digits (i.e., *1*), but the effect reversed when numbers were presented at words (i.e., *one*) which was the opposite pattern to Experiment 1A.

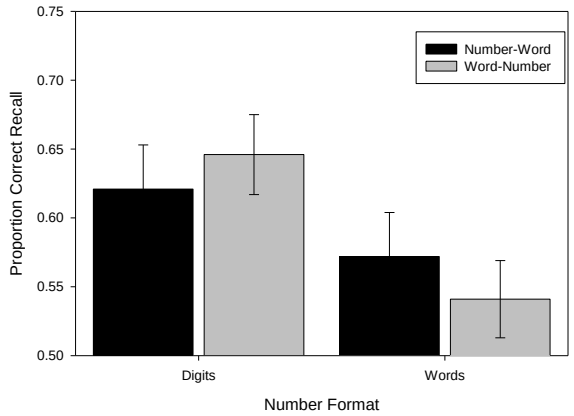


Figure 2. The effect of Number Format on the Category-Order Effect. Encoding fluency influences the nature of the effect rather than being due to numbers preceding words. Errors bars represent 1 SE.

The interaction of Category-Order and Recall Order was also found to be significant, $F(1,30) = 14.767$, $MSE = .02$, $p = .001$. As seen in Figure 3, recall was relatively higher for Word-Number lists over Number-Word lists when recalled in forward recall order compared to backward recall order.

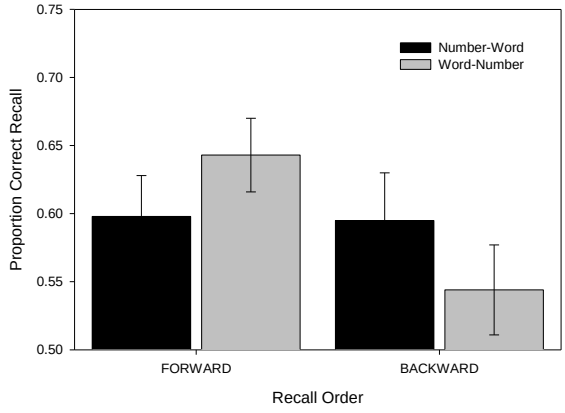


Figure 3. The effect of Recall Order on the Category-Order Effect. When words are recalled first (i.e., in the forward Word-Number or the backward Number-Word condition) recall performance is highest. Errors bars represent 1 SE.

Discussion

The present experiment did not replicate the main category-order effect from Experiment 1A, however, a pattern of improved overall recall was still present for numeric digits over numbers represented as words. Additionally, unlike prior studies, the effect of recall order did not show a category-order effect based on input position, but instead indicated that words exhibited relatively higher recall when recalled first, that is, for forward recall with the Word-Number category order and for backward recall with the Number-Word category order. The main difference in word stimuli in this study is that words were controlled for length, category-relatedness, and were from smaller category-sets than the stimuli from Brooks and Watson (1990).

One possible explanation for these findings is that, in the short-term, the lexical and phonotactic properties of the stimuli facilitate recall but when more time is allowed to pass, the activation of associated items in long-term memory creates competition during retrieval and decreases recall accuracy. If recall can proceed immediately, related words can be recalled with little interference. If participants are required to recall these items last, as more time passes a greater number of alternative candidates could be activated in long-term memory creating greater interference (Anderson, 1974).

General Discussion

Brooks and Watkins (1990) and Greene and Lasek (1994) both proposed the existence of a category-order effect whereby full-list recall is improved when numbers precede words in a mixed-list design. The original interpretation of the COE attributed this finding to the categorical properties of the number and word stimuli: number stimuli were drawn from a smaller, more homogenous category than word stimuli.

The results of two experiments revealed the conditions in which item and order information interact to increase recall performance in serial recall tasks in which the items belong to two different categories. In Experiment 1A, we replicated the findings of previous experiments. More items were recalled from lists which presented number items prior to words (Brooks & Watkins, 1990; Greene & Lasek, 1994; Young & Supa, 1941). Experiments 1A, however, extended these results. It was also observed that number format (digit vs word) and orthographic properties of list halves also contribute to the COE. In general, the greater the orthographic differences there were between the lists halves, the greater the increase in recall performance.

Experiment 1B further extended these results by eliminating the COE once word-frequency was controlled. Experiment 2 also eliminated the main effect of COE once category set-size, word length, and within category similarity was controlled. In fact, the trend evidenced a contrasting pattern of results: once the categorical and sub-

categorical properties of the word stimuli was more comparable to those of the number stimuli, the trend was for more items recalled when words preceded numbers. The fact that a COE-like effect was observed in Experiments 1A and 2 wherein word length was allowed to vary might suggest that orthographic properties reduce proactive interference giving rise to previous results interpreted as COE.

The relative gain for recall performance of digit-word lists relative to word-digit lists is only one demonstration of a category-order effect. This does, however, illustrate that what previous authors have construed as categorical properties might in fact be partly due to sub-categorical properties shared by the stimuli. For instance, the robust finding that high-frequency words are better recalled than low-frequency words in no way implies that high-frequency and low-frequency words are represented as contrasting categories in an individual's memory. Instead, categories would seem to require additional inter-item associations such as physical similarity of exemplars (e.g., whales and fish or whales and mammals), as seen in Experiment 2.

Traditionally, the COE was assumed to arise from an interaction between categorical properties of stimuli (numbers rather than words) and the order in which they were presented (numbers before words). When participants are presented with stimuli, a trace is created in short-term memory. Information associated with those stimuli is activated in long-term memory. When items share category membership, the associations between items in a list generally enhance recall performance (Saint-Aubin et al., 2005). However, when categories are large, the resultant spread of activation to other category members creates interference (Nelson et al., 1989).

The results of our experiments lead us to question previous assumptions about the COE. Namely, although order and item information do contribute to recall performance, and that categorical properties of the stimuli likely affect recall performance, the initial detection threshold of the stimuli appears to account for more recall performance once it has been controlled. This finding also has implications for studies in that perceptual effects appear to contribute more to recall performance than knowledge effects.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 5, 451-474.

Anderson, J. R. & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128, 186-197.

Brooks, J. O., & Watkins, M. J. (1990). Further evidence of the intricacy of memory span. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 1143-1141.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language, 31***,** 1-17.

Crannel, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology, 44*, 319-327.

Crowder, R.G. (1979). Similarity and order in memory. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 13, pp. 319-353). New York: Academic Press.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, *38*(1), 188-230.

Greene, R. L., & Lasek, R. (1994). Category-order effects in memory span. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1391-1395.

Healy, A. F. (1974). Separating item from order information in short-term memory. *Journal of Verbal Learning & Verbal* Behavior, 13, 644-655.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*, 401-410.

Jordan, T. R., Redwood, M., & Patching, G. R. (2003). Effects of form familiarity on perception of words, pseudowords and nonwords in the two cerebral hemispheres. *Journal of Cognitive Neuroscience, 15,* 537-548.

Lashley, K.S. (1951). The problem of serial order in behaviour. In L. A. Jefress (Ed.), *Cerebral Mechanisms in Behaviour: the Hixon Symposium.* Wiley.

McGinnies, E., Comer, P. B., & Lacey, O. L. (1952). Visual-recognition thresholds as a function of word length and word frequency. *Journal of Experimental Psychology*, *44*, 65-69.

Nairne, J. S., & Kelley, M. R. (2004). Separating item and order information through process dissociation. *Journal of Memory and Language, 50*, 113-133.

Nelson, D. L., Keelean, P. D., & Negrao, M. (1989). Word-fragment cuing: The lexical search hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 388-397.

Perri, R., Bartolomeo, P., & Silveri, M. C. (1996). Letter dyslexia in a letter-by-letter reader. *Brain and Language*, *53*, 390-407.

Postman, L., & Adis-Castro, G. (1957). Psychophysical methods in the study of word recognition. *Science*, *125*, 193-4.

Schoenherr, J. R. & Thomson, R. (2009). Modeling the Category-Order Effect with an Oscillator-Based Connectionist Network. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, ND.

Schoenherr, J. R. & Thomson, R. (2008). Category Properties and the Category-Order Effect. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington, USA.

Smith, F., Lott, D., & Cronnell, B. (1969). The effect of type size and case alternation on word identification. *The American Journal of Psychology*, *82*, 248-253.

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials? *Psychonomic Bulletin & Review*, *12*(1), 171-177.

Watkins, M. J. (1977). The intricacy of memory span. *Memory & Cognition, 5*, 529-534.

West, R. L., Pyke, A. A., Rutledge-Taylor, M. F., & Lang, H. (2010). Interference and ACT-R: New evidence from the fan effect. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 277-281). Philadelphia, PA: Drexel University.

Young, C. W., & Supa, M. (1941). Mnemonic inhibition as a factor in the limitation of the memory span. *The American Journal of Psychology*, *54*, 546-552.

Causal Learning With Continuous Variables Over Time

Kevin W. Soo (kevin.soo@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Abstract

When estimating the strength of the relation between a cause (X) and effect (Y), there are two main statistical approaches that can be used. The first is using a simple correlation. The second approach, appropriate for situations in which the variables are observed unfolding over time, is to take a correlation of the change scores – whether the variables reliably change in the same or opposite direction. The main question of this manuscript is whether lay people use change scores for assessing causal strength in time series contexts. We found that subjects' causal strength judgments were better predicted by change scores than the simple correlation, and that use of change scores was facilitated by naturalistic stimuli. Further, people use a heuristic of simplifying the magnitudes of change scores into a binary code (increase vs. decrease). These findings help explain how people uncover true causal relations in complex time series contexts.

Keywords: causal learning; causal reasoning; time

Introduction

Knowing the strength underlying cause-effect relationships (e.g. how strongly a drug suppresses a symptom) allows people to decide which causes to use to achieve desired outcomes (Hagmayer & Meder, 2013). Past research has uncovered various ways people infer the strength of a causal relation after observing the covariation between the cause and effect (e.g. Cheng, 1997; Griffiths & Tenenbaum, 2005; Hattori & Oaksford, 2007). Most research has focused on how people assess causal relations among *binary* causes and effects – usually 'absent' or 'present' (for exceptions, see Pacer & Griffiths, 2011; Rottman, in press; Saito, 2015). Additionally, the temporal order of the observations is often random and typically viewed to be an irrelevant factor by the researchers. The present study investigates causal learning with continuous variables that are observed in time series exhibiting increasing or decreasing trends. To anticipate the findings, we found that reasoning about continuous variables and reasoning about time series trends are interrelated processes; sequentially-presented continuous variables are treated as binary, which simplifies the learning process.

Learning with continuous variables

In the real world, variables are often *continuous* or at least ordinal in scale – e.g. a drug is not either 'present' or 'absent' but administered with a particular dosage. How do people infer causal strength when a cause and effect can each assume multiple levels?

One theory is that people simplify continuous variables by mentally dichotomizing them into binary variables, which would make it easier to summarize the values of the cause and effect for computing causal strength. One study found that people assimilate intermediate values and treat them as either the high or low value on the scale (Marsh & Ahn, 2009). However, the variables in that study were not really continuous; they primarily had a high or low value, and occasionally an intermediate value. It is unknown whether people mentally dichotomize variables when the variables take on values along a fuller range of possible levels, which would require arbitrarily choosing cutoff values.

Another theory is that when judging the causal strength of a continuous cause on a continuous effect, people perform a mental computation similar to Pearson's correlation coefficient *r*. In fact, in correlation-learning paradigms, peoples' correlation estimates are sensitive to many of the components that are used for calculating *r* like the slope, error variance and the variance of each variable (Lane, Anderson, & Kellam, 1985). A weakness of *r* as a model of human learning is that it is a computational-level theory that fails to explain how the learner actually processes the information in a tractable way. Computing *r* would be computationally intensive; all the X and Y values would need to be remembered and integrated into one score.

In the next section we introduce causal learning in time series contexts, and later discuss how causal learning over time and with continuous variables are interrelated.

Observations over time

There are two standard paradigms for causal learning: situations in which the order of the data is meaningful (e.g., perhaps X or Y undergoes a trend), and situations in which the data have no inherent temporal ordering and are presented randomly. We call these situations "longitudinal" and "cross-sectional", respectively; an analogy to the terms used to describe experimental research designs.

The correlation coefficient *r* is appropriate for causal learning in cross-sectional contexts, but can be misleading when used in longitudinal contexts. For example, consider the data in the "negative transitions" condition in Figure 1. The order of observations is denoted with the numbers 1-20 in the plot. Both X and Y gradually increase over time, and overall there is a positive correlation between X and Y. However, from one observation to the next there is a negative correlation between X and Y; when X increases, Y decreases, and vice versa. Which is the "right" way to interpret the causal relation between X and Y? Is the strength of the relationship positive or negative?

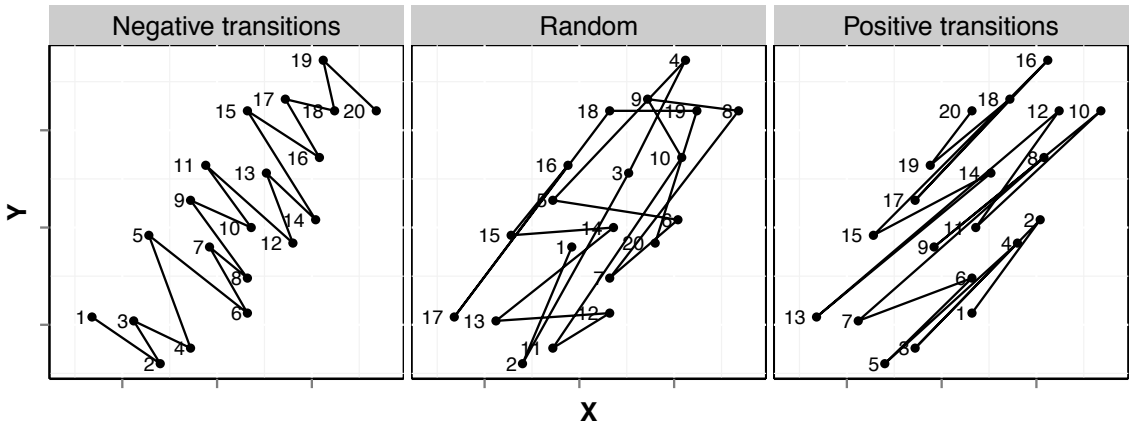


Figure 1: Positively correlated data presented in different orderings (denoted by numbers).

We assert that the latter interpretation based on how the variables change together is appropriate in a longitudinal context. For example, in Figure 1 (Negative transitions), X and Y both increase over time. It would be inappropriate to conclude that X and Y are positively related just because they both increase over time, as many variables exhibit temporal trends. For example, the US economy and the price of oil have generally increased over time (positive correlation), even though increases in the price of oil cause the economy to contract at a smaller time scale. Data sets like this arise naturally whenever 1) a cause and effect pair, X and Y, both increase (decrease) over time due to the influence of a common cause, Z; and 2) X has a negative (positive) causal influence on Y. Figure 2 shows a causal graph depicting this scenario. To estimate the true causal relationship between X and Y, we need to control for Z.

Table 1: Information from longitudinal observations of a cause (X) and effect (Y) relationship. The data are the same as in Figure 1 (negative transitions condition).

Time	States		Transitions			
			$\Delta_{Continuous}$		Δ_{Binary}	
	X	**Y**	**X**	**Y**	**X**	**Y**
1	17	27				
2	35	15	18	–12	+1	–1
3	25	26	–7	11	–1	+1
4	43	19	15	–7	+1	–1
5	32	48	–11	29	–1	+1
	•		•		•	
	•		•		•	
	•		•		•	
18	81	80	13	–3	+1	–1
19	78	93	–3	13	–1	+1
20	92	80	14	–13	+1	–1
	r_{States} = .70		$r_{\Delta Cont}$ = –.97		$r_{\Delta Binary}$ = –1	

Examining the *changes* in Y in response to changes in X accounts for secular trends in both variables (i.e. controls for the effect of the common cause influencing them over time). A way to account for secular trends in time series analysis is to take difference scores (Δ) of X and Y, and then compute a correlation on the difference scores; in Table 1 the Δ scores are denoted with $\Delta_{Continuous}$, from which $r_{\Delta Cont}$ is computed. Using Δ scores controls for first-order non-stationarity (linear trends) in a time series (Shumway & Stoffer, 2011).

The main question of this article is whether humans intuitively compute causal strength from the absolute or difference scores of the cause and effect (i.e. based on r_{States} or $r_{\Delta Cont}$). This is tested using datasets like those in Figure 1. The three plots have the same 20 data points so r_{States} = .70 for all three, but $r_{\Delta Cont}$ changes from very negative to very positive.

In addition to the fact that $r_{\Delta Cont}$ is better at uncovering the true causal relation in contexts with temporal trends, there is a second reason why $r_{\Delta Cont}$ may be a better model than r_{States} of how humans infer causal strength. Across many sensory modalities (e.g., sound, light, pain), humans encode relative changes rather than the absolute magnitudes of stimuli because our senses adapt to the current level of stimulation (Stewart, Brown, & Chater, 2005). This means that, even though Table 1 suggests that $r_{\Delta Cont}$ requires an additional step of computing Δ scores from raw scores, in naturalistic

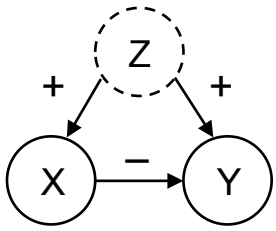


Figure 2: The relationship between X and Y when there is a common cause (Z) influencing both over time. X negatively influences Y. Z causes X and Y to increase over time.

settings Δ scores may be the primary form of data that we have access to through our sensory systems. For this reason, humans may naturally attend more to to Δ scores; making $r_{\Delta Cont}$ a better model of human learning than r_{States}.

Continuous vs. Binary Representation

A question raised above is how a learner could compute a correlation between two variables when it would require remembering all the values of X and Y as well as computationally integrating all those values; $r_{\Delta Cont}$ would be just as challenging to compute as r_{States}. Earlier, we proposed that in certain contexts learners may discretize continuous variables. In longitudinal situations, there is a natural cutoff that may ease discretization; whether the variable increased (coded as +1), decreased (–1), or stayed the same (0) from the previous time point (Δ_{Binary} in Table 1). Computing the correlation between Δ_{Binary} scores ($r_{\Delta Binary}$) only requires keeping track of the number of times that X and Y increase and decrease together and separately, raising the possibility of discretization as a plausible heuristic.

Experiment 1 tested whether learners infer causal strength based on the absolute values of the cause and effect (similar to r_{States}) in addition to how the cause and effect changed over time (similar to $r_{\Delta Cont}$ and $r_{\Delta Binary}$). Experiment 2 investigated the extent to which presentation format influences the reliance on transitions. Experiment 3 tested whether people use continuous or binary representations of change scores for inferring causal strength.

Experiment 1

Experiment 1 tested whether learners use information about transitions (Δ scores) in addition to information about states (raw scores) when inferring the causal strength between a cause (X) and effect (Y) in a time series context. We predicted a stronger effect of transitions relative to states on causal strength judgments.

Method

Subjects 50 subjects were recruited using Amazon Mechanical Turk (MTurk) and paid $0.60. The experiment lasted approximately 5 minutes.

Design and stimuli Subjects inferred causal strength from data sets consisting of 20 observations of X and Y, where X and Y could take on values ranging from 0 to 100. We manipulated the correlation between the states of X and Y as well as the transitions (see Figure 1) in a 2 (positive vs. negative r_{States}) × 3 (negative vs. random vs. positive transitions) within-subjects design.

Table 2: Means and SDs of $r_{\Delta Cont}$ in different conditions

	$r_{\Delta Cont}$ Mean (SD)		
	Negative transitions	Random transitions	Positive transitions
r_{States} = .7	–.96 (.02)	.73 (.07)	.97 (.01)
r_{States} = –.7	–.97 (.01)	–.73 (.07)	.96 (.02)

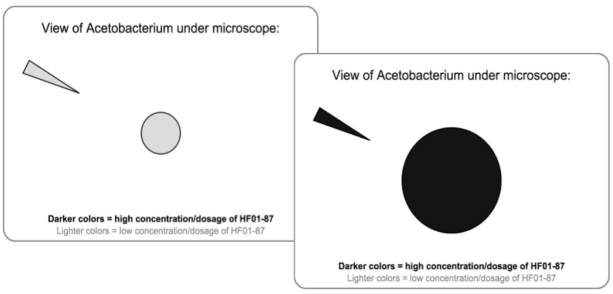


Figure 3: Stimuli viewed by subjects. Two consecutive observations are shown.

The manipulation of r_{States} involved data sets with either positive or negative r_{States}. Data sets with r_{States} = .7 were generated, and copies were made with the values of X flipped around the midpoint of the scale (X = 50), creating data sets with r_{States} = –.7.

The manipulation of transitions was achieved by reordering the observations to produce three conditions (as in Figure 1). In the negative transitions condition, the relationship between the Δ scores of X and Y at every time point is negative – increases in X are always accompanied by decreases in Y, and vice versa. In the positive transitions condition, increases in X are always accompanied by increases in Y. In the random transitions condition, the order of the 20 observations was randomized, resulting in a mix of positive and negative transitions. In the positive (negative) r_{States} condition, most of the transitions are positive (negative) (see Table 2).

20 data sets were created per condition. Each subject viewed one randomly chosen data set in each of the six conditions, and the order of the conditions was randomized.

Procedure Subjects were told they would evaluate how the dosage of a drug (X) affected the size of a microorganism (Y) over 20 observations ('days'). On each day, a new dosage of the drug was administered to the same microorganism and the size was observed under a microscope. The microorganism was represented using a circle (see Figure 3). Each scenario was presented as data of a different drug-microorganism pair.

The value of X was mapped onto the opacity of the circle, with darker shades representing higher doses. The value of Y was mapped to the diameter of the circle.

After 20 observations, subjects judged the causal strength, on a scale of 8 ("high levels of the drug strongly cause the microorganism to increase in size") to –8 ("high levels of the drug strongly cause the microorganism to decrease in size"). A rating of 0 indicated no causal relationship.

Results

Causal strength judgments were analyzed with a within-subjects factorial ANOVA (see Figure 4a for means). As expected, subjects rated causal strength higher in the positive r_{States} condition than in the negative condition; $F(1, 288) = 11.21$, $p < .001$, $\eta_p^2 = .04$. The primary question was

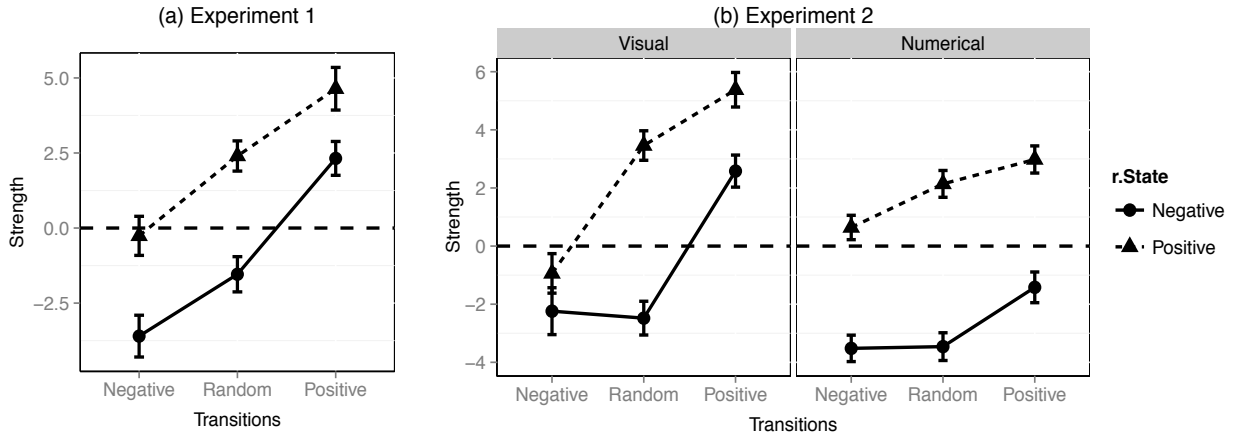


Figure 4: Means by Condition. Error bars represent standard errors.

whether the transitions had an influence on causal strength ratings above and beyond r_{States}. Indeed, subjects rated causal strength highest in scenarios with positive transitions, followed by random transitions, and negative transitions; $F(2, 288) = 20.38$, $p < .001$, $\eta_p^2 = .12$. The effect size of transitions was about three times larger than the effect of states (r_{States}). No interaction effect was observed. In the condition with negative r_{States} but positive transitions, subjects actually judge the causal strength to be positive; the effect of transitions 'overrides' the effect of states in this case (see Figure 4a). With positive r_{States} but negative transitions, subjects' judgments are not significantly different from zero; the negative transitions 'neutralized' the effect of states.

Experiment 2

In the introduction we argued that people may use transitions for estimating causal strength when the cause and effect are presented as naturalistic perceptual stimuli (such as in Experiment 1); adaptation of sensory systems to the level of stimuli results in an emphasis on changes in the environment. This feature of our sensory systems could be beneficial in helping us learn causal relations in time series contexts, which requires analyzing Δ scores to detect true causal relations. In Experiment 2 we test if the effect of transitions on causal strength judgments is larger for perceptual stimuli than for stimuli presented numerically. Such a finding would suggest that humans are better at uncovering causal relations in more naturalistic settings.

Method

Subjects 100 subjects were recruited using Amazon Mechanical Turk (MTurk) and paid $0.60. The experiment lasted about 5 minutes.

Design, stimuli and procedure The same method as Experiment 1 was used, except a between-subjects factor was added so that half the subjects viewed the data presented in a visual format (identical to Experiment 1 and shown in Figure 3), while half the subjects viewed the data with X and Y presented numerically.

Results

Replicating Experiment 1, there was a main effect of r_{States}, $F(1, 582) = 44.33$, $p < .001$, as well as transitions $F(2, 582) = 19.82$, $p < .001$. The main question was whether the influence of transitions was larger in the visual than numerical condition. Though the effect of transitions was obtained in both visual ($p < .001$) and numerical ($p < .001$) presentation formats separately, the effect of transitions is more dramatic (steeper slopes of the lines in Figure 4b) in the visual condition; $F(2, 582) = 8.10$, $p < .001$, $\eta_p^2 = .03$. Presenting stimuli in a numerical format attenuated the effect of transitions. When stimuli are presented in a naturalistic format (as in most causal learning contexts involving real-world stimuli), people more naturally attend to transitions, which helps uncover the true causal strength.

Experiment 3

Experiments 1 and 2 showed that people use transitions for inferring causal strength above and beyond states (r_{States}). However, computing a correlation on Δ scores is still a computational challenge. In the introduction we proposed that perhaps instead of using the continuous Δ scores ($\Delta_{Continuous}$), that people instead code changes as increasing or decreasing (Δ_{Binary}) (Table 1), which could simplify the memory and inference process. The stimuli in Experiments 1 and 2 conflated the two; e.g. in the negative transitions condition, $r_{\Delta Cont} \approx -1$, and $r_{\Delta Binary} = -1$ (because every transition was negative). In Experiment 3, we created data sets in which $r_{\Delta Cont}$ and $r_{\Delta Binary}$ diverged to test whether people discretize Δ scores for assessing causal strength.

Method

Subjects 50 subjects were recruited using Amazon Mechanical Turk (MTurk) and paid $1.50. The experiment lasted about 10-12 minutes.

Stimuli and Design We created datasets that hold r_{States} constant, and vary $r_{\Delta Binary}$ and $r_{\Delta Continuous}$. This is challenging because $r_{\Delta Binary}$ and $r_{\Delta Continuous}$ in a given data set are typically highly similar. This was accomplished in the following way:

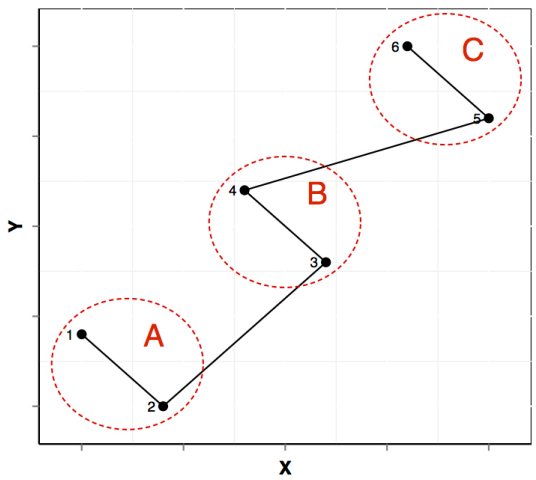


Figure 5: Sample observations from Experiment 3. Numbers reflect the order of the observations.

Each dataset had 12 observations – two observations of each of the following six joint states of (X, Y): (0, 20), (20, 0), (40, 60), (60, 40), (80, 100) and (100, 80). The r_{States} correlation was .83. The observations formed three clusters (A-C, see Figure 5) with low, medium, or high values. Transitions between two observations within a cluster necessarily involved a negative transition (X increasing and Y decreasing, or vice versa). Transitions across clusters necessarily involved a positive transition. Varying the ratio of within to between-cluster transitions changes both $r_{\Delta Cont}$ and $r_{\Delta Binary}$.

However, for a given ratio (and fixed $r_{\Delta Binary}$), $r_{\Delta Cont}$ is also influenced by the between-cluster transition path. Transition paths with large jumps between clusters (e.g. A, C, A…) result in higher $r_{\Delta Cont}$ values, whereas transition paths with small jumps between clusters (e.g. A, B, C…) result in lower $r_{\Delta Cont}$ values.

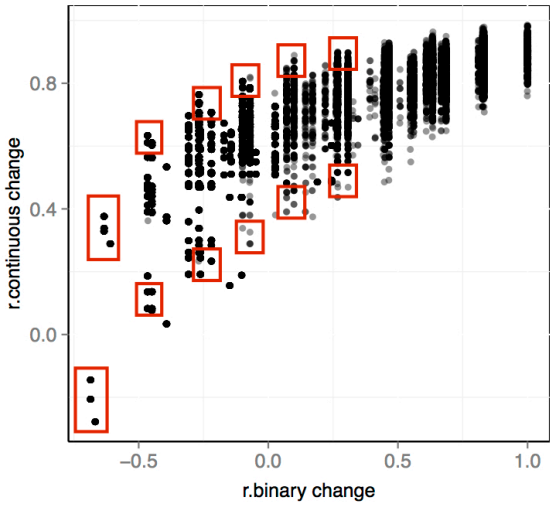


Figure 6: $r_{\Delta Cont}$ and $r_{\Delta binary}$ for generated data sets. Sampled regions of data sets have been marked with squares.

Ten thousand data sets were generated by randomly ordering the 12 observations, which produced data with varying degrees of $r_{\Delta Cont}$ and $r_{\Delta Binary}$ (Figure 6). In order to maximally discriminate $r_{\Delta Cont}$ and $r_{\Delta Binary}$, we selected data sets from 12 regions on the periphery of the two-dimensional space (Figure 6). Each subject worked with 12 data sets – one from each region. $r_{\Delta binary}$ values ranged from –.69 to .31, and $r_{\Delta Cont}$ values ranged from –.21 to .90.

The 12 data sets were presented in a random order. Randomly, half the time, the values of X were flipped around the midpoint so that r_{States} was either .83 or –.83; judgments from the flipped conditions were reverse-coded for analysis.

Procedure The procedure was the same as Experiment 1 (the data was only presented in the visual format), except that each data set had only 12 observations.

Results

The exact $r_{\Delta Cont}$ and $r_{\Delta Binary}$ values of a particular data set were used as predictors in a regression of subjects' final judgments of causal strength. The regressions had a by-subject random intercept for repeated measures, and by-subject random slopes for $r_{\Delta Cont}$ and $r_{\Delta Binary}$ to capture the possibility that some subjects might use $r_{\Delta Cont}$ or $r_{\Delta Binary}$ more strongly or weakly than other subjects.

A bivariate analysis found that $r_{\Delta Binary}$ was a significant predictor of the causal strength judgments (B = 2.42 and SE = .48, p < .001, R^2 = .037). However, somewhat surprisingly, $r_{\Delta Cont}$ was not a significant predictor even in a bivariate analysis (B = .76 and SE = .46, p = .10, R^2 = .006). The reason for the smaller effect size here compared to Experiment 1 is that the $r_{\Delta Cont}$ values here were less extreme. A multivariate analysis again found that $r_{\Delta Cont}$ was not significant (B = –.76 and SE = .57, p = .19, ΔR^2 = .002), and $r_{\Delta Binary}$ remained significant (B = 2.87 and SE = .59, p < .001, ΔR^2 = .032). $r_{\Delta Binary}$ is a significant predictor over and above $r_{\Delta Cont}$.

Overall, Experiment 3 suggests that people do discretize transitions as "positive" or "negative" for the purpose of estimating causal strength.

General Discussion

Past research on causal strength learning has focused primarily on binary variables, and variables that do not exhibit temporal trends (e.g., increasing or decreasing). The present research was concerned with how people judge causal strength from a continuous cause and effect when they are observed over time. We find that in longitudinal situations how the cause and effect change over time is crucial when judging the strength of the causal relationship.

In Experiment 1, we presented subjects with data sets with constant *state-based* information (r_{States}). By manipulating the order of observations to create all positive or all negative transitions (varying $r_{\Delta Cont}$), we created stimuli with state and transition-based information leading to diverging conclusions about causal strength. In conditions where they were in conflict, transition-based information

overcame (or at least neutralized) the effect of the state-based information.

Experiment 2 demonstrated that people attend more to transition-based information in causal learning contexts involving naturalistic visual stimuli rather than numerical stimuli. This finding suggests that our sensory systems, which are attuned to changes in the environment through the process of adaptation, may naturally code information in a way that helps us uncover causal relations in time series contexts. In this case, the visual system is involved.

Finally, Experiment 3 showed that people use the direction of change in a variable's state (increase vs. decrease) more than the magnitude of change for estimating causal strength. This result suggests that people discretize continuous changes into binary changes to determine if transitions are positive or negative, which may be a heuristic for calculating causal strength.

Soo & Rottman (2014) showed that people use transitions for inferring causal strength from binary variables. The present study generalizes this finding to the more complex (and one might argue, more real-world) problem of judging causal strength from continuous variables. As demonstrated in Figure 1, using transitions for estimating causal strength in a time series context is especially important when the cause and effect are continuous variables; only using the overall correlation between the states can result in concluding that there is a positive causal relation when it is in fact negative, or vice versa.

One important question to be answered is whether people use transitions for inferring causal strength even in situations when the order is statistically irrelevant – e.g. when each observation comes from different entities rather than consecutive observations of the same entity, or when there are no temporal trends. For example, in the random conditions in Experiments 1 and 2, if r_{states} was positive (negative), the transitions were mainly positive (negative). The way we tested for transitions involved introducing temporal trends, but this methodology cannot provide insight into situations when there are no temporal trends. Other important questions include understanding individual differences in the role of memory and attentional processes for keeping track of transitions, especially if observations are spaced out.

Overall, these experiments provide a positive view of human causal learning in longitudinal contexts; people are able to uncover causal relations despite complex temporal trends. Furthermore, this process appears to be fairly easy, and is facilitated by naturalistic presentations. It is possible that people also use changes over time for other sorts of learning processes such as learning correlations (as opposed to causal relations), or learning to discriminate between categories.

Acknowledgements

This research was supported by NSF 1430439.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–84.

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 33–50.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: model comparison and rational analysis. *Cognitive Science*, *31*(5), 765–814.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–60.

Lane, D. M., Anderson, C. A., & Kellam, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(5), 640–649.

Marsh, J. K., & Ahn, W.-K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 334–52.

Mooij, J., Janzing, D., Heskes, T., & Schölkopf, B. (2011). On Causal Discovery with Cyclic Additive Noise Models. *Advances in Neural Information Processing Systems*, *24*, 1–9.

Pacer, M. D., & Griffiths, T. L. (2011). A rational model of causal induction with continuous causes. *Advances in Neural Information Processing Systems*, *24*.

Rottman, B. M. (in press). Searching for the Best Cause: Roles of Mechanism Beliefs, Autocorrelation, and Exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Saito, M. (2015). How People Estimate Effect Sizes : The Role of Means and Standard Deviations. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2075–2079).

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal Contiguity and the Judgement of Causality by Human Subjects. *The Quarterly Journal of Experimental Psychology*, *41B*(2), 139–159.

Soo, K., & Rottman, B. M. (2015) Elemental Causal Learning from Transitions. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (2254-2259).

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute Identification by Relative Judgment. *Psychological Review*, *112*(4), 881–911.

Viewing time affects overspecification: Evidence for two strategies of attribute selection during reference production

Ruud Koolen (r.m.f.koolen@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Albert Gatt (albert.gatt@um.edu.mt)
Institute of Linguistics, University of Malta

Roger P.G. van Gompel (r.p.g.vangompel@dundee.ac.uk)
School of Psychology, University of Dundee

Emiel Krahmer (e.j.krahmer@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Kees van Deemter (k.vdeemter@abdn.ac.uk)
Department of Computing Science, University of Aberdeen

Abstract

Speakers often produce definite referring expressions that are overspecified: they tend to include more attributes than necessary to distinguish the target referent. The current paper investigates how the occurrence of overspecification is affected by viewing time. We conducted an experiment in which speakers were asked to refer to target objects in visual domains. Half of the speakers had unlimited time to inspect the domains, while viewing time was limited (1000 ms) for the other half. The results reveal that limited viewing time induces the occurrence of overspecification. We conjecture that limited viewing time caused speakers to rely heavily on quick heuristics during attribute selection, which urge them to select attributes that are perceptually salient. In the case of unlimited inspection time, speakers seem to rely on a combination of heuristic and more deliberate selection strategies.

Keywords: Definite reference; overspecification; heuristics; viewing time.

Introduction

In everyday language use, speakers often refer to objects in the world around them. Such references often take the form of a definite description that contains an article, one or more modifiers, and a head noun (e.g., "the brown chair", or "the large table"). Among the various reasons why speakers refer to objects, identification is perhaps the most obvious one: in many cases, speakers aim to distinguish one *target* from the *distractors* that are present in the context. Therefore, deciding what to say (or, in the terminology of Levelt, 1989, conceptualization) is a crucial part of referring: which attributes should be selected to make the target identifiable?

For this attribute selection process, previous research has shown that speakers tend to *overspecify* their object descriptions: they often use redundant attributes that are not strictly needed for unique identification of the target. For example, speakers might produce "the green chair" in a visual domain where only one chair is present. In general, one can say that overspecification is most likely to occur when speakers refer to target objects in rather complex visual domains, including cluttered domains (e.g., Clarke, Elsner & Rohde, 2013), and domains with a high amount of visual variation between the target referent and its distractors (e.g., Koolen, Goudbeek & Krahmer, 2013; Rubio-Fernández, 2016).

So why do speakers overspecify their object references so frequently? The answer to this question may be found in the incremental nature of speech production (Pechmann, 1989), and, in particular, in the recent suggestion that speakers rely on quick *heuristics* during attribute selection. Heuristics can be defined as "beliefs concerning the likelihood of uncertain events (...) that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (Tversky & Kahneman, 1974, p. 1124). Recent work in psycholinguistics has suggested that heuristics also affect attribute selection during reference production. In particular, it has been argued that speakers are in some cases prevented from making exact calculations about the shortest possible description in a referential domain (Van Deemter, Gatt, Van Gompel & Krahmer, 2012).

If speakers indeed use heuristics during attribute selection, this could explain their tendency to overspecify. We want to argue that heuristics cause speakers to include attributes that are salient in a given domain. In many referential situations, speakers are simply lacking time and / or cognitive effort to perform a careful object-by-object scan of the target and its distractors. Instead, they might rely on a heuristic, and select attributes that are easily and quickly processed by the visual system. The most notable example here is color, which is an absolute attribute that does not require comparison to any other object in the domain in order to be perceived. In other words, color "pops out" of the scene (Belke & Meyer, 2002; Treisman & Gelade, 1980). As a result, color is a preferred attribute, which is included irrespectively of the number of distractors that it rules out in a domain (Dale & Viethen, 2009). Also size can be salient, when size differences between the objects in the domain are sufficiently large (Van

Gompel, Gatt, Krahmer & Van Deemter, 2014), or due to speakers' general tendency to repeatedly include the same attributes during reference production (Tarenskeen, Broersma & Geurts, 2015).

The use of heuristics may serve as a plausible explanation for the occurrence of overspecification: if speakers include attributes based on salience and preference, they may end up with a description that contains attributes that are not strictly needed to identify the target referent. Furthermore, in some situations, also less preferred attributes might be required to rule out some last remaining distractors in the domain. Thus, given that speakers and listeners cooperate during referential communication (Brennan & Clark, 1996), it seems plausible to assume that attribute selection is usually determined by a combination of preference and discriminatory power (i.e., the number of distractors that an attribute excludes in a particular domain). This assumption implies that there are multiple influences on attribute selection, which has been modeled in various recent computational interpretations of the conceptualization process. For example, the *Visible Objects Algorithm* (Mitchell, Van Deemter & Reiter, 2013) models these influences as two consecutive stages: an early stage in which visually salient and preferred attributes are selected, and a later stage where attributes are included based on their discriminatory power. Also the algorithm proposed by Gatt, Goudbeek and Krahmer (2011a) distinguishes between heuristic and more stable strategies, but models them in parallel rather than consecutively.

Despite the above computational models, prior research in psycholinguistics that has tested the procedures proposed in these models is mostly lacking. Therefore, in this paper, our goal is to investigate if speakers are indeed guided by both heuristic and more deliberate processes during attribute selection. In order to reach this goal, we performed a reference production experiment where we manipulated *viewing time*: half of our speakers could take as much time as necessary to inspect the visual domains they were presented with, while the other half had limited viewing time. To see how viewing time influenced reference production, we analyzed speakers' tendency to overspecify their descriptions. Our expectations are as follows.

For speakers with unlimited viewing time, we expect that they will overspecify their referring expressions, but only to some extent. For one thing, heuristic viewing strategies will cause them to include properties that are perceptually salient for them. However, since there is unlimited time to inspect the visual domain here, we expect selection based on inherent salience to interact with selection based on discriminatory power and exact calculations. As a result, the amount of overspecification may be small: preferred attributes may be avoided in some cases, and speakers might rather select attributes to exclude as many distractors at once as possible. This strategy allows the listener to rule out the remaining distractor objects and identify the target (Olson, 1970).

For speakers who have limited time to inspect a scene, the situation might be different: in this case, they may not have enough time to calculate the shortest possible description and to perform an object-by-object scan of the visual scene. Instead, these speakers may base attribute selection heavily on inherent preferences for certain attributes, and might thus be guided primarily by heuristics rather than discriminatory power. For example, they might start uttering a description before they have scanned all distractor objects that are present in the visual scene. As argued by Pechmann (1989), this incremental process may in turn cause speakers to overspecify, since it makes them include salient attributes for which they are not sure if they are needed for identification or not. After all, the limited viewing time could prevent them from taking the listener perspective into account (Horton & Keysar, 1996), and from searching for attributes with the highest discriminatory power.

Method

To study the effect of viewing time on overspecification, we performed an experiment in which participants took part in a simple director-matcher task. In this task, one participant – the speaker – described a target referent from a group of seven objects, to a listener who saw the same objects but in a different configuration. The speaker was instructed to refer to target objects in such a way that the listener could identify the intended referents.

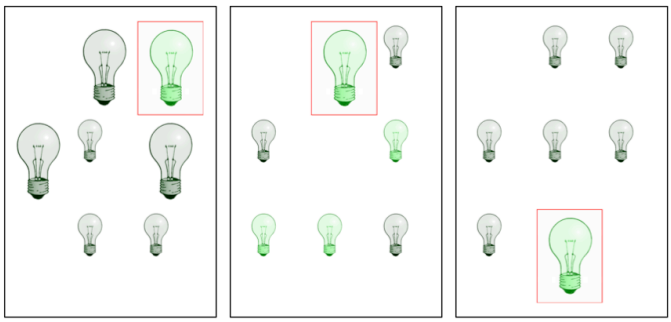
Participants

Participants were 36 undergraduate students from Tilburg University (19 female, 17 male, age range 17-34 years old, M = 21 years and 9 months), who took part as fulfillment of course credits. The participants were all native speakers of Dutch (the language of the study), and took part in the role of the speaker. A confederate – who was the same person for all participants – took part in the role of the listener.

Materials

The stimulus material consisted of thirty-six visual domains such as those shown in Fig. 1 on the next page. All domains depicted one target referent, which was marked with a red square, and six distractors. We constructed thirty-six domains that served as critical trials, and selected thirty-six different objects for the target objects. These objects were the same as those used in an earlier experiment by Gatt, Van Gompel, Krahmer, and Van Deemter (2011b). In all domains, the six distractors had the same type as the target referent. The seven objects were placed randomly in a 3x3 grid, such that the positions of the objects and the empty cells varied across domains.

Again following Gatt et al. (2011b), trials were manipulated in three *conditions*: (i) **C**, where *color* was required to identify the target referent; (ii) **S**, where *size* was necessary; and (iii) **C/S**, where participants could either use *color or size* to produce a distinguishing description. The use of size in the C condition, or color in the S condition, excluded only three out of six distractor objects. Example domains for the three conditions can be found in Fig. 1a-c. The thirty-six target objects were distributed across three lists, so that each



(a) Color needed (b) Size needed (c) Color or size needed

Figure 1: Experimental domains. The manipulations of color may not be visible in a black and white print of this paper.

target referent appeared in a different condition in each list. Participants were randomly assigned to one of the three lists, and the conditions always had twelve trials each.

Color contrast between the target and the distractors in the domains was manipulated in such a way that speakers could not use basic color terms to refer to a target's color. In Figure 1a-c, for example, speakers naturally used *light green* when their description included color, because some (or all) of the distractor objects were dark green. In half of the critical trials, the target's color shade was light, and in the other half it was dark. Target objects occurred either in red, grey, blue, or green.

The use of basic color terms was disabled in an attempt to balance the trade-off between speaker preferences for color and size attributes. As we have seen in the Introduction, it is known that speakers prefer to use color, even if this leads to overspecification (e.g., Pechmann, 1989; Belke & Meyer, 2002; Koolen et al., 2013, among many others). By contrast, attributes such as size tend to be used only when absolutely required. However, the inclusion of color becomes less likely when the color differences between the target and its distractors are small, and when basic color terms are not sufficient to identify the target (Viethen, Goudbeek & Krahmer, 2012). Both these conditions were met in the current experiment.

The second independent variable of the experiment, *pace*, manipulated viewing time. This variable was tested between participants. Half of the speakers took part in the *self-paced* condition, and could take as much time as needed for each trial to inspect the domain and describe the target. The other half of the speakers took part in the *system-paced* condition, Although the speakers in this condition could again take as much time as they needed to describe the target, the visual domains disappeared automatically after 1000 milliseconds for each trial. This means that the time that speakers had to inspect the domains and to find the distinguishing attributes for the targets was limited. The time window of 1000 milliseconds was decided upon with a pre-test. The main criterion here was that participants should experience pressure, but should still be given enough time to take a look at all objects in the domain, and thus to avoid underspecified descriptions that do not contain enough information to identify the target. From the pre-test, we learned that speakers were able to avoid underspecification when they were given 1000 milliseconds to inspect the domain; speakers confirmed that they could then indeed to take a look at all the objects that were visible. In the experiment itself, we found that only 1.3% of the descriptions were underspecified. The visual scenes for the confederate listener were always displayed for the duration of the whole trial, irrespective of condition.

In addition to the thirty-six domains for the critical trials, we created thirty-six filler domains. The fillers consisted of two abstract 3D 'Greebles' figures (Gauthier & Tarr, 1997), all purple, so that speakers were not primed with using color in the critical trials. One Greeble was marked as the target, and could be distinguished from the other Greeble by means of its main shape or by the direction in which its protrusions were pointing.

Procedure

The experiment took place in a quiet office room at Tilburg University. The average running time was approximately 15 minutes for each participant. After signing the consent form, participants were randomly assigned to either the self- or the system-paced condition, which resulted in eighteen speakers per condition. Participants sat at a table facing their listener, in front of a computer screen. The seventy-two trials (thirty-six critical trials and thirty-six fillers) were presented on the screen one by one, in randomized order for each participant. The visual domains for the confederate listener were shown on a laptop placed in front of him. The computer screen and the laptop were positioned such that eye contact between the speaker and the listener was possible. E-prime 2.0 was used to run the experiment.

The instructions emphasized that speakers had to describe the target objects so that the listener could uniquely identify them, and mark them on a paper answering form. Given that the listener was presented with the same visual domains as the speaker, but with the objects in a different configuration, the instructions for the speaker also mentioned that it would not make sense to refer to objects with location information. Irrespective of the condition that speakers were assigned to, they could take as much time as necessary for every trial to describe the target. Once the listener had identified a target, he pressed the space bar to continue to the next trial. Before each trial, a fixation cross was depicted in the middle of the screen for 1 second. There were three practice trials, including one with Greebles, and the speakers' descriptions were recorded with a microphone. The confederate listener never asked clarification questions, so the data presented here can be regarded as one-shot reference.

In order to measure if speakers in the system-paced condition indeed experienced more pressure than those in the self-paced condition, all speakers filled out a short questionnaire after finishing their referential task. This questionnaire consisted of five questions on a 10-point scale, derived from the NASA task load index (Hart & Staveland, 1988). We asked

speakers to estimate: the task load (1); the pressure that they experienced (2); how well they had succeeded in describing the objects (3); how hard they worked to describe the targets (4); and how frustrated they were during the experiment (5). After reversing the scores for the second question and calculating the mean for each speaker, we found that the speakers in the system-paced condition (M = 4.62; SD = 1.54) indeed experienced their task as more demanding than the speakers in the self-paced condition (M = 3.37; SD = 1.24), F(1,34) = 7.27; p = .011).

Data coding

In total, our 36 speakers produced 1296 object descriptions. There were 25 missing cases due to technical issues with the audio recordings. The remaining 1271 descriptions were all coded for the occurrence of referential overspecification. A description was either overspecified (score = 1) or not (score = 0). Overspecified descriptions always contained color *and* size: one attribute which was necessary for identification of the target referent, and one redundant attribute. Descriptions that were coded as not overspecified were mostly minimally specified, and contained either color *or* size. References that were underspecified (16 in total) were coded with 0 as well. We did not consider whether speakers used a type attribute or not in coding the references, because the seven objects in the domains were always of the same type. Naturally, type attributes were often mentioned by our speakers to produce proper noun phrases.

Research design and statistical analysis

The experiment had a 3x2 design, with *Condition* (levels: **C**, **S**, **C/S**) as a within-participants factor and *Pace* (levels: self-paced, system-paced) as a between-participants factor. The dependent variable was the proportion of object descriptions that was overspecified.

To test the effect of Pace and Condition on the occurrence of referential overspecification, we performed a logit mixed model analysis (Jaeger, 2008). In our model, Pace and Condition were included as fixed factors, and items and participants as random factors. The fixed factors were centered to reduce collinearity. The model had a maximal random effect structure and included random intercepts and random slopes for all within-participant and within-item factors, in order to ensure optimal generalizability (Barr, Levy, Scheepers & Tily, 2013). As such, the model contained random intercepts for participants and items, and random slopes for Condition and Pace at both the participant level and the item level. The p-values were estimated via parametric bootstrapping over 100 iterations.

Results

Figure 2 plots the proportion of descriptions containing only color, only size, or color-and-size, as a function of the three Conditions (**C**, **S**, **C/S**) and the manipulation of Pace.

Overall, 36% of the referring expressions produced by our speakers contained both a color and a size attribute and were thus overspecified. Our model showed a main effect of Pace

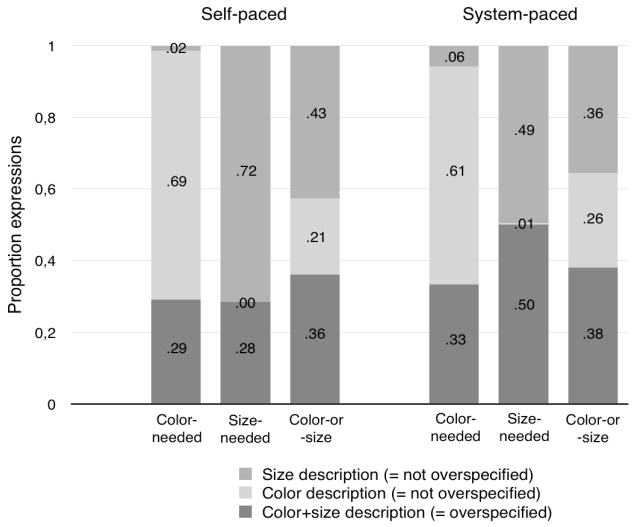


Figure 2: Proportion of color-only, size-only, and color-and-size descriptions as a function of the three Conditions and the Pace manipulation. Note that only the color-and-size descriptions (see the lower, darker bars) are overspecified.

on the occurrence of referential overspecification: redundant modifiers were more frequent in the system-paced condition (40.7%) than in the self-paced condition (31.3%), β = 0.72; SE = 0.57; $p < .05$. As can be seen in Figure 2, this effect of Pace was consistent across the three Conditions: the patterns for the **C**, **S**, and **C/S** conditions separately all show a higher proportion of overspecification for speakers who had limited time to inspect the scene. There was no main effect of Condition on the occurrence of overspecification (β = -0.01; SE = 0.36; n.s.).

Further inspection of the proportions in Figure 2 suggests that the effect of Pace is mediated by Condition. For instanc ple, in the **S** condition - where only size is needed to uniquely identify the target - the difference between speakers in the self-paced (28%) and the system-paced (50%) conditions is considerable in terms of overspecification, or at least bigger than in the **C** and **C/S** conditions. However, the interaction between Pace and Condition did not reach significance (β = 0.40; SE = 0.43; n.s.). This lack of interaction may be due to a substantial amount of speaker variation. We come back to this issue in the Discussion.

Discussion

The current paper investigated the effect of viewing time on attribute selection in the production of definite reference. In particular, the data revealed that viewing time influences the occurrence of referential overspecification: redundant modifiers were more frequent in the system-paced condition than in the self-paced condition. This effect did not interact with the type of visual domain, which we manipulated in terms of the attributes that were needed to uniquely identify the target referent (i.e., color, size, or color/size).

The main effect of *Pace* is in line with our expectations, and we regard it as converging evidence for the notion that speakers use quick heuristics when selecting the content of their referring expressions (e.g., Van Deemter et al., 2012; Dale & Viethen, 2009). In fact, speakers seem to particularly do so when their time to inspect the domain is limited, as was the case in our system-paced condition. The higher proportion of overspecified descriptions that we observed there suggests a heuristic, preference-based approach for speakers under pressure, where they select attributes merely based on inherent salience rather than discriminatory power. After all, these speakers may simply have lacked time and thus cognitive capacity (Horton & Keysar, 1996) to perform a deliberate object-by-object scan of the domain. Instead, they might have selected the content of their referring expressions in an incremental way (Pechmann, 1989). As we have seen in the Introduction section, this incremental process may eventually cause speakers to overspecify.

The lower proportion of overspecified object descriptions produced in the self-paced condition implies that a heuristic strategy is less dominant when speakers have unlimited time to inspect the visual scene. In this condition, there seems to have been sufficient time available for speakers to put effort in selecting the attribute that was most efficient in ruling out the distractor objects that were present. This way, speakers were – at least to a certain extent – able to avoid overspecification. However, it is important to note that if speakers in the self-paced condition had based the selection of attributes on discriminatory power alone, they may not have overspecified their descriptions at all, since there was in all domains one attribute that excluded all distractors at once. Hence, we take the finding that speakers did not avoid being redundant here as yet another argument for a model of human attribute selection where heuristic and more deliberate processes take place in parallel. As referred to in the Introduction section, such a model is similar to various computational models of attribute selection for definite reference (Gatt et al., 2011a; Mitchell et al., 2013).

The data did not show a significant effect of *Condition* on the occurrence of overspecification, in contrast to the related experiment by Gatt et al., 2011b, which manipulated similar conditions. This lack of main effect can be explained by the fact that in our study, the use of basic color terms was disabled, in an attempt to balance the trade-off between speaker preferences for color and size attributes. Normally, speakers have a strong preference to use color (e.g., Pechmann, 1989; Koolen et al., 2013), which could result in a high proportion of color-only descriptions in the **C** and **C/S** conditions, and a ceiling effect for overspecification in the **S** condition, even for self-paced speakers. In other words: speakers might then simply select color all the time, just because it is highly salient. However, with our subtle manipulation of color differences, we managed to discourage this strategy.

Close inspection of Figure 2 suggests the existence of an interaction between *Pace* and *Condition*: at least numerically, the effect of pressure was far most convincing in the **S** condition, where size was needed to distinguish the target. In this condition, the proportion of overspecifications almost doubled for speakers under pressure, while there was only a slight increase in the **C** and **C/S** conditions. However, the interaction between Pace and Condition was not significant. As mentioned earlier, we expect that this lack of interaction was due to a substantial amount of speaker variation, especially in the self-paced condition. The pattern that emerges there is that for almost half of the speakers, we find a proportion of 0 - 10% overspecified references, while the proportions for the other speakers range from 10 - 100%. In the system-paced condition, these proportions are more centered around the mean. This difference in consistency between the two conditions could explain why the interaction between pressure and condition was not significant.

Although the interaction effect was not significant, it remains of course interesting to speculate about the nature of the large numerical increase in the number of overspecified descriptions in the **S** condition. We reason that the inherent preference for color over size pays off here anyway, in spite of the subtle color differences in our domains. In cases with unlimited viewing time, color differences are generally detected faster than size differences (Belke & Meyer, 2002), because color is an absolute attribute that does not require comparison to other objects to be perceived, unlike the relative attribute size. In our experiment, this natural difference between size and color may explain the large increase in the proportion of overspecified descriptions in the **S** condition: speakers under pressure could simply have lacked sufficient time to compare the objects and sizes. Selecting color could provide a solution here, since it reduces the distractor set to three rather than six objects. This subset may be sufficiently small for speakers to detect if also size is needed to uniquely identify the target, perhaps even after the domain has disappeared, based on memory. In order to test this idea, it would be interesting to replicate our experiment in an eye-tracking paradigm, to test if speakers indeed rely on a subset of distractors before (and after) the domain has disappeared.

Related to the above, another next step could be to look at the way in which viewing time affects reference production at the surface, word level. In Dutch, which was the language of the experiment, color and size are usually realized before the head noun. However, if speakers indeed mentioned color to reduce the number of distractors, and 'decided' about the need for size later, one may expect that many size modifiers in the system-paced condition were produced *after* the head noun. Hence, there could also be more speech repairs in this case, with speakers producing descriptions such as "the light green light bulb, the large one". Third, there may be effects of pressure on speech onset times, which might be longer in the self-paced rather than the system-paced condition. These kinds of analyses can all be conducted with the current data; we are planning to do so in the near future. Hence, it would also be interesting to replicate our experiment in a language where post-nominal modification is the default, which is for example the case in Spanish or Maltese.

Other directions for future research include for example to run an experiment in which one of the attributes rules out no

distractors at all. In all conditions of the current experiment, both color and size ruled out at least one distractor object. In the system-paced condition, speakers might have included a redundant attribute because – during their quick scan of the scene – they realized that it excluded at least one distractor, although they were not sure how many distractors. If such a strategy is indeed applied, limited viewing time might not give rise to overspecification when certain attributes rule out no distractors at all, since during the quick scan, the speaker will not identify any distractor that these attributes would rule out. Second, one could conduct a follow-up experiment in which it is crystal clear for speakers that minimality (i.e., producing the shortest possible distinguishing description) is the aim, while giving them a very short inspection time. It is then the question to what extent their referential behavior resembles the behavior in the normal situation, where minimality is not particularly stressed. Finally, it would be interesting to explore the interplay between limited viewing time and overspecification from a more listener-oriented perspective, for example with a colorblind listener.

Conclusion

This paper explored the impact of viewing time on attribute selection in definite reference, and on referential overspecification in particular. We found more redundant attributes in the system-paced condition than in the self-paced condition, and conjectured that speakers rely heavily on fast heuristics when they have limited time to inspect the visual domain. In the case of unlimited inspection time, they seem to rely on a combination of heuristic and more deliberate strategies.

References

Barr, D., Levy, R., Scheepers, C. & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*, 255– 278.

Belke, E. & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during "same" "different" decisions. *European Journal of Cognitive Psychology, 14*, 237-266.

Brennan, S. & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*, 1482-1493.

Clarke, A., Elsner, M. & Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology, 4*: 329.

Dale, R. & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. *Proceedings of the 12th European workshop on natural language generation (ENLG)*. Athens, Greece, 58–65.

Gatt, A., Goudbeek, M. & Krahmer, E. (2011a). Attribute preference and priming in reference production: Experimental evidence and computational modeling. *Proceedings of CogSci 2011*. Boston, Massachusetts, 2627-2632.

Gatt, A., Van Gompel, R., Krahmer, E. & Van Deemter (2011b). Non-deterministic attribute selection in reference production. In *Proceedings of PRE-CogSci 2011*. Boston, Massachusetts.

Gauthier, I. & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX: results of empirical and theoretical research. In: Hancock, P.A., Meshkati, P. (Eds.), *Human Mental Workload* (pp. 139-183). Elsevier, Amsterdam.

Horton, W. & Keysar, B. When do speakers into account common ground? *Cognition, 59*, 91-117.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Koolen, R., Goudbeek, M. & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite object descriptions. *Cognitive Science, 37*(2), 395 – 411.

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Mitchell, M., Van Deemter, K. & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Association for Computational Linguistics.

Olson, D. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review, 77*, 257-273.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*, 89-110.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7*: 153.

Tarenskeen, S., Broersma, M. & Geurts, B. (2015). Overspecification of color, pattern and size: salience, absoluteness, and consistency. *Frontiers in Psychology, 6*: 1703.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science, 4* (2), 166-183.

Van Gompel, R., Gatt, A., Krahmer, E. & Van Deemter, K. (2014). Overspecification in reference: modelling size contrast effects. *Poster Presented at AMLaP 2014* (Edinburgh, UK).

Viethen, J., Goudbeek, M. & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. *Proceedings of CogSci 2012*. Sapporo, Japan, 1084-1098.

Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions

Jens Koed Madsen (j.madsen@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck, University of London, Malet Street, Bloomsbury, London, WC1E 7HX
School of Geography and the Environment, University of Oxford, Manor Road, OX1 3UJ

Abstract

The credibility of politicians is crucial to their persuasiveness as election candidates. The paper applies a parameter-free Baysian source credibility model (integrating trustworthiness and epistemic authority) in a real-life test predicting participants' posterior belief in the goodness of an unnamed policy after a named candidate has publically supported or attacked it.

Two studies test model predictions against policy support and attack of five presidential candidates from the USA. Model predictions were measured against observed posterior belief in the goodness of the policy.

The results strongly suggest the model captures essential traits of how participants update beliefs in policies given appeals to a candidates' support of attack. Further, individual differences suggest that people consider other factors than the ones elicited for the study. More studies into appeals to specific candidates are warranted to construct more accurate models of the influence of source credibility on political reasoning.

Keywords: Bayseian source credibility, trustworthiness, epistemic authority, political reasoning

Introduction

Source credibility influences a range of human cognitive phenomena such as the reception of persuasive messages (Petty & Cacioppo, 1984; Chaiken & Maheswaran, 1994; Tormala & Clarkson, 2007), the development of children's perception of the world (Harris & Corriveau, 2011), legal reasoning (Lagnado et al., 2013), decision-making (Birnbaum et al., 1976), adherence with persuasion strategies (Cialdini, 2007), and how people are seen in social situations (Fiske et al., 2007; Cuddy et al., 2011).

The normative function and role of source credibility in argumentation is still subject to debate. The dual-process-based Elaboration-Likelihood Model (Petty & Cacioppo, 1981) takes sources as heuristic rather than analytic cues (Petty & Cacioppo, 1984, Briñol & Petty, 2009). Comparatively, recent Bayesian models place source credibility within a rational paradigm (e.g. Hahn et al., 2012; Harris et al., in press). The latter predicts the convincingness of an appeal to expert opinion from a Bayesian perspective as a product of the perceived trustworthiness and epistemic authority of the source (Bovens & Hartmann, 2003; Hahn et al., 2009).

For the purpose of the present paper, source credibility is defined as an amalgamation of epistemic authority and trustworthiness[1]. Epistemic authority is "the authority of those with superior knowledge in a specific field – experts" (Harris et al., in press). This differs from administrative authority, which refers to people who have authority bestowed upon them. For example, a judge is an epistemic authority on legal matters whilst a police officer is an administrative authority. Trustworthiness refers to the likelihood that a person will not deliberately present wrong or misleading information. Thus, where epistemic authority refers to the *capability* of providing good information, trustworthiness refers to the *intention* to actually do so.

Other sources influence how we perceive the credibility of sources, e.g. facial configurations (Rezlescu et al., 2012), the gender of the candidate (see later), as well as a host of other factors. For example, a misogynist may believe that a woman is both trustworthy and capable, but may still refuse advice from the woman solely based on her gender. It is therefore important to stress that the scope of the model tested only considers trustworthiness and epistemic authority rather than a richer source credibility account.

Whilst the persuasive and behavioural influence of epistemic authority has not been studied extensively in politics (although it would be a highly interesting study given recent anti-governmenetal sentiments in many democracies), it is clear that trustworthiness is an important factor in politics. It increases compliance with public policy (Ayres & Braithwaite, 1992), influences the choice of political candidate (Hetherington, 1999), increases *intention* of voting (Householder & LaMarre, 2014), increases societal cooperation (Fukuyama, 1995), and lack of trust may instigate civic participation (see Levi & Stoker, 2000).

How to determine if a person is a trustworthy politician remains an open issue. In political science literature, the main facets cited to describe trustworthness are integrity, competence, fairness, flip-flopping, honesty, equitable, and being responsiveness to public needs (Miller & Listhaug, 1990; Levi & Stoker, 2000), which mirrors definitions in rhetoric (e.g. McCroskey, 1997), cognitive and social psychology (e.g. Bovens & Hartmann, 2003; Hahn et al., 2009; Fiske et al., 2007; Cuddy et al., 2011), and reasoning theory (Walton, 1997).

Given the influence of source credibility on politics, it is reasonable to assume that the perception of an election candidate impacts multiple factors for voters. First, it may modulate their beliefs and support for particular policies. Second, trust in a candidate or government correlates with a person's willingness to adhere with official policies. Third, low credibility spurs anti-incumbent voting behavior.

[1] Walton (1997) argues for six traits of source credibility. However, Harris et al (in press) suggest that, in absence of other interlocutors, two may suffice. The model employed here is similar to the one tested in Harris et al.

The studies in the paper test a Bayesian source credibility model that uses trustworthiness and epistemic authority to predict the persuasive potential of references to a candidate. All model parameters are elicited (prior beliefs *and* conditional probabilities), generating parameter-free predictions subsequently compared against posterior beliefs in the goodness of a policy. This ensures that there is no a posteriori model fitting to observed data. To the author's knowledge, this is the first study to apply the general Bayesian model in a parameter-free way to a political context to test the persuasive influence of appeals to specific political candidates' positions on policies.

Modelling source credibility

Bayesian approaches to reasoning take point of departure in subjective, probabilistic degrees of beliefs in propositions where the posterior degree of belief in a proposition is captured by Bayes' theorem (Oaksford & Chater, 2007; Howson & Urbach, 1996). The approach is suggested as an alternative to logicist approaches to reasoning (Oaksford & Chater, 1991) and has been applied to argumentation theory (Hahn & Oaksford, 2006; 2007, see also Oaksford & Hahn, 2004; Corner et al., 2011; Harris et al., 2012). The findings suggest Bayesian reasoning can account human information integration in practical reasoning.

Bovens and Hartmann provide a formal foundation for a Bayesian source credibility model (2003, see also Schum, 1981; Hahn et al., 2012). It integrates epistemic authority and trustworthiness and provides predictions for the posterior degrees of belief in the hypothesis (H) given the representation (Rep) by a source:

$$P(H|Rep) = \frac{P(H) \times P(Rep|H)}{P(H) \times P(Rep|H) + P(\neg H) \times P(Rep|\neg H)}$$ [2]

Harris et al. (in press) test the model using dialogous and show a good fit between predictions and observed convincingness. The current studies extend the model tested in Harris et al. (in press) by applying the model to predict the convincingness of appeals to specific and known experts in a political domain (election candidates).

Rather than a dichotomous description of an unknown source as completely trustworthy/untrustworthy, the current studies elicit prior trustworthiness and epistemic authority beliefs about five presidential candidates from the USA. The study makes use of real-life candidates and thus relies on participants' subjective beliefs about these rather than making use of fictitious and abstractly described sources. As such, the study is an extension of the empirical test of the Bayesian model, as it lodges the dialogue in a more natural setting.

Given the previous success of the Bayesian model in Harris et al. (in press), we hypothesise that the model can predict how convinced the electorate will be given their prior beliefs of each of the five candidates together with their conditional probabilities. The model captures different degrees of posterior convincingness depending on the prior beliefs regarding the candidate in question. Thus, if people have low prior beliefs in the trustworthiness and epistemic authority of a particular candidate (e.g. Donald Trump), the model predicts that the persuasiveness of that particular candidate will be low.

Givent previous empirical support for the model, we predict that the model will be able to account for a significant amount of the observed posterior degree of belief in the goodness of the policy on a population level. On an individual level, however, we predict greater noise due to the fact that, as mentioned in the above, estimations of other people involve other factors than the two central characteristics measured here and due to the fact that individual predictions are noiser than population estimations in general. Nonetheless, we predict the model to be significantly positively correlated with individual observations as well.

Study 1: Method, design, and respondents

Study 1 tests the predictive potential of the Bayesian source credibility model against appeals to specific candidates when they publicly endorse or attack a policy. This mimicks political discourse in which opinons are formed, not necessarily on the basis of evidence for or against a given policy, but on the basis that a politician that a person trusts and finds expert has supported the policy (mutatis mutandis if a person finds a candidate untrustworthy and inexerpt recommends a policy, it may be grounds for dismissal of the policy). Such reasoning is frequently observed in political debates and it is important to understand the persuasive potential of simply referring to the opinions of known political figures (e.g. a person who finds Bernie Sanders highly credible might believe a policy to be good because Sanders publically supported it).

Design and method

The candidates tested in the study were from the American race for presidential nomination for the 2016 election. Five candidates were included in the study. Candidate choice was influenced by prominence and contemporary attention. This yielded two Democratic (Hilary Clinton and Bernie Sanders) and three Republican candidates (Jeb Bush, Marco Rubio, and Donald Trump)[3].

In order to generate parameter-free predictions, prior beliefs for candidates as well as all conditional probabilities were collected from each participant. The soure credibility model merges epistemic authority, P(E), trustworthiness, P(T), and hypothesis, P(H) (see fig. 1).

[2] P(Rep|H) = P(Rep|H, exp, T) * P(Exp) * P(T) + P(Rep|H, ¬exp, T) * P(¬Exp) * P(T) + P(Rep|H, ¬exp, ¬T) * P(¬Exp) * P(¬T) + P(Rep|H, exp, ¬T) * P(Exp) * P(¬T); mutatis mutandis for P(Rep|¬H)

[3] Clinton, Sanders, Bush, and Rubio were expected to be front-runners. Trump was included, as he enjoyed a lot of attention at the time of the study (July 2015).

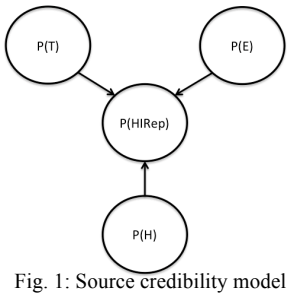


Fig. 1: Source credibility model

To elicit relevant prior beliefs relating to source credibility of each politician, participants were asked how trustworthy and politically expert they believe each candidate to be. Participants responded on a sliding scale, ranging from 0 to 100 (0 was complete disagreement with the statement, 100 was complete agreement with the statement). Results varied greatly between candidates with the highest average trust score given to Sanders (61.73) and the lowest to Trump (18.36) and the highest average expertise score given the Clinton (75.63) and the lowest to Trump (19.58)[4].

Conditional probabilities were elicited, which represent the likelihood that a person would represent something to be true if the person is trustworthy or not, expert or not, and if the hypothesis is actually true or false (e.g. for P(Rep|T, E, H) participants valued the likelihood that a completely trustworthy and politically expert person would support a policy in a world where that policy happens to be good). Table 1 presents the average estimations of the conditional probabilities later used to calculate posterior predictions.

H	H	H	H	¬H	¬H	¬H	¬H
T, E	T, ¬E	¬T, E	¬T, ¬E	T, E	T, ¬E	¬T, E	¬T, ¬E
80.38	58.21	34.63	18.04	22.59	42.3	59.90	71.26

Table 1: Conditional probabilities (study 1)

Model predictions were measured against posterior degrees of belief in the goodness of an unknown policy given public support or attack from a candidate. In line with previous studies on Bayesian argumentation (e.g. Harris et al., 2012; Harris et al., in press), P(H) was assumed to be 0.5. In order to convey this assumption, the interlocutor in the study explicitely states that she has "…no idea whether this policy is good or bad". Having heard this statement, the other interlocutor argues that the policy is good given the fact that "CANDIDATE has publically supported/attacked the policy". The participant was then asked how likely it is that the policy is good given the candidate's statement. In total, the participants read ten dialogues (attack and support for each candidate).

The posterior degree of belief in the goodness of the policy was compared against predictions from the model as described in the above. Given the fixation of P(H), there are no free parameters in the model and consequently no possibility for posterior model fitting.

[4] As discussed later, most participants identified as Democrats. As could be expected, prior beliefs divided across party lines with self-identified Republicans rating Republican candidates higher and mutatis mutandis for self-identified Democrats.

Respondents

252 participants were recruited from Mechanical Turk (see Paolacci et al., 2010 for validation of MTurk as a tool for social scientific research). Participants had to be American citizens eligible to vote (average age 34.11). 43.7% were women and participants mainly identified as supporting the Democratic Party (57.5% Democrats, 13.5% Republicans, 1.2% tea party, 9.9% libertarian, 17.6% undecided)[5].

Respondents were paid an equivalent of $9/hour to participate. Data was collected from the 30th of July to the 2nd of August 2015[6].

Study 1: Results

The results are divided into two main sections. The first section looks at averages across the participant population. This tests if the model is capable of predicting posterior degrees of belief in the goodness of policies given public support of attack from named election candidates. The second section looks at predictions from each individual.

Predicted posterior degrees of belief in the goodness of a policy given support were calculated via the above equation for P(H|Rep). Predictions for P(H|¬Rep) (i.e. if a candidate publically *attacked* the policy) P(H|¬Rep) was calculated from reversed conditional probabilities, e.g. P(¬Rep|H, E, T) = 1- P(Rep|H, E, T). The conditionals allow for this, as the full probabilistic range was elicitied for all conditions.

No gender or age effects were observed for prior beliefs of trustworthiness or expertise for any of the candidates or for conditional probability estimations. However, as will be discussed later, a one-way ANOVA shows that a gender effect *was* observed on posterior degrees of belief in the goodness of the policy for Hilary Clinton (but not for any other candidate), $F(1,251) = 4.984$, $p = 0.026$.

Population predictions

Posterior degrees of belief in the goodness of policies were non-trivial, as participants did indeed take the stance of the candidate into consideration when evaluating the likelihood that the policy was good given the appeal to the opinions of a specific candidate. The average posterior degrees of belief in the policy in the support condition ranged from 31.39 (Trump) to 65.32 (Sanders). Similarly, the average posterior degree of belief in the attack condition ranged from 36.84 (Sanders) to 63.51 (Trump). This suggests that, for preferred candidates, an endorsement is persuasive such that people update their beliefs in the likelihood that the policy will be good. Conversly, if disliked candidates

[5] For the purpose of the current study, the non-equal political distribution does not pose a problem, as the difference will be expressed through their prior beliefs in the trustworthiness and epistemic authority of each candidate. Indeed, the results show that participants favour candidates from their preferred party.

[6] The time of data collection is relevant, as developments in the candidates' campaigns could entail significant changes in estimations of candidates' trustworthiness and epistemic authority.

endorse a policy, people update their belief in the opposite direction and believe the policy to be bad.

A linear regression suggests the parameter-free Bayesian source credibility model captures essential characteristics of how the population update their beliefs given the support or attack of a particular candidate. The R^2 of the model predictions against observed average posterior belief is .824 ($p < 0.001$).

Individual predictions

Individual predictions are expected to be noisier. As mentioned previously, other factors than trustworthiness and epistemic authority influence how we see and react to other people. The gender difference for posterior beliefs in policies given Clinton's statements suggest that being a woman influences how men and women react to Clinton's statements. It is worth noting that this effect is not found for male candidates, suggesting that only being a female candidate influences the posterior degree of beliefs. It would be worth conducting a study on sexism in political reasoning. Other factors may be relevant for each candidate (e.g. Bush being the brother and son of former presidents, Trump being a Washington outsider, Sanders being a socialist, Rubio being Latino, Clinton being married to a former president, etc.). The model does not include these personal differences, but they will, to some degree, influence some voters regarding their persuasive potential.

Also, individual errors cannot be ameliorated given the parameter-free model. Thus, if a participant misunderstands the conditional probability questions, model predictions will be wrong. In some cases, this was observed, as some participants either flat-lined (50 for all conditionals) or reversed the scores presented in table 1. These have not been excluded for the current analyses, which reduces the predictive potential of the model on an individual level.

Despite the fact that individual predictions are difficult to model, the parameter-free Bayesian source credibility is able to capture a significant correlation in the support and attack conditions. For support, a linear regression between observed and predicted posterior beliefs yield an R^2 of .462 for support ($p < 0.001$) and .317 for attack ($p < 0.001$). As expected, the model accounts for less of the variance, but remains highly significantly correlated with observed posterior beliefs.

Study 2: Method, design, and respondents

Study 1 suggests the model can be applied successfully as a predictor of the persuasive impact of appeals to specific candidates for an unknown policy. The parameter-free predictions were strongest on a population level and, as may be expected, less strong on an individual level. Study 2 tests if the model can be applied to sequential political reasoning. First, the participant is told that one candidate has publically supported or attacked the policy (replication of study 1). Having provided their posterior belief, the participant is then told that another candidate has publically supported or attacked the policy.

Design and method

Study 2 follows the same method as study 1. Priors and conditional probabilities were elicited from participants (conditionals showed similar response patterns as study 1, see table 1 on p. 3). To limit sequential combinatorics, two candidates were chosen (Jeb Bush and Bernie Sanders). They were chosen, as they represented the highest scoring Democratic and Republican candidate from study 1.

It was a 2 (candidate #1 attack/support) x 2 (candidate #2 attack/support) x 2 (order of candidates) design. Half of the participants saw Bush as candidate #1 and Sanders as #2. Participants saw four dialogues with the order of presentation as a between subjects condition.

Model predictions were identical to study 1. However, rather than a one-off elicitation of belief in the policy after each dialogue, participants were asked for two posterior degrees of belief. The first posterior degree of belief was elicited after candidate #1 supported or attacked the policy. The second was elicited after candidate #2 was referenced. As in study 1, initial P(H) was assumed to be 0.5. In order to capture the dynamic element of the paradigm, $P(H|Rep_1)$ or $P(H|\neg Rep_1)$ was taken as P(H) for the subsequent representation ($P(H|Rep_2)$). This allowed for a dynamic belief updating mechanism to be tested against observed degrees of belief in the goodness of the policy.

Respondents

511 participants were recruited from Mechanical Turk from the 26th to the 31st of August. Respondents were paid an equivalent of $9/hour to participate. 48.0% were women; mean age was 37.14. The political affiliation was similar to study 1: 56.2% Democrats, 20.7% Republicans, 1.8% tea party, 4.9% libertarian, 16.4% undecided. No effect of gender or age was observed on prior beliefs, conditional probabilities, and posterior degrees of belief were observed.

Study 2: Results

Similar to study 1, population and individual observations were compared against model predictions. Responses and model predictions after candidate #1 were identical to the ones from study 1. Consequently, the subsequent analysis focuses on the fit between responses and predictions after candidate #2 has provided his opinion.

Population predictions

Notably, Harris et al (in press) report increased noise when introducing interlocutors who disagree with the advice given by the expert. This is further made complicated by the real-life nature of the dialogues. As opposing politicians tend to disagree on crucial political points (especially as political polarisation has increased in American politics), participants may have found the support-support and the attack-attack conditions confusing.

Despite a noisier dialogical structure in which two rather than one candidate offer their opinion on a particular policy, the model has a good fit with observed posterior

degress in the goodness of the proposed policy. A linear regression shows that the model predicts responses on a population level ($R^2 = .690$, $p < 0.001$)

A paired-sample t-test shows an order effect in three of four conditions: BS to SS v SS to BS: mean diff (12.44), $t = 5.405$, $p < 0.001$; BA to SS v SS to BA: mean diff (7.02), $t = 3.501$, $p = 0.003$; BA to SA v SA to BA: mean diff (-7.66), $t = 3.550$, $p < 0.001$)[7]. This may be due to a recency effect coupled with the estimation of the candidate, as the order effect was observed when Sanders ended the dialogue. As in study 1, most of the participants were Democrats and Sanders was rated higher in trustworthiness and expertise. Consequently, his opinion may have been given greater weight overall. Further studies in order effects for sequential political opinions should explore this further.

Individual predictions

As with study 1, individual predictions were expected to have a less good fit with observed data. As previous, some participants provided different answers to the conditional probabilities than expected. Further, order effects were observed, which further challenges individual predictions.

Despite the above considerations, the model had a good fit with individual responses. Linear regression analyses show the model accounting for 16.3% (Bush to Sanders, $p < 0.001$) and 19.3% (Sanders to Bush, $p < 0.001$).

General discussion

The studies applied a general Bayesian source credibility model to appeals to known and named political election candidates involved in the 2016 presidential election in the USA. This provides a real-life test if the model can predict the persuasive potential of stating that the candidate has publicly supported or attacked a particular policy (e.g. 'you shouldn't like this policy because Donald Trump has publicly supported it'). The model had no free parameters, as prior beliefs regarding trustworthiness and epistemic expertise, conditional probabilities, and posterior degrees of belief in the goodness of the proposed policy given the public support or attack were measured. Study 1 tests appeals to five known candidates (Clinton, Sanders, Bush, Rubio, and Trump) whilst study 2 tests sequential appeals to two known candidates (Bush and Sanders).

Overall, the results suggest that the Bayesian model enjoys a good fit with observed data on a population level and, to a lesser extent, on an individual level. On the population level, the model accounts for 82.4% and 69.0% of the variance in study 1 and 2 respectively. Individual differences were expected to be noisier, as variables that may influence the judgment of one person may be irrelevant to another. Despite greater noise, the model accounts for 31.7-46.2% of the variance in study 1 and 16.3-19.3% in the more complex study 2. This indicates that other variables influence the persuasive potential of election candidates (e.g. as a gender effect was observed for the persuasiveness of Clinton despite the fact that no differences were observed in the estimations of her trustworthiness and expertise, it is reasonable to assume that being a woman influences how persuaded people were of appeals to Clinton). Future studies should determine these additional variables more concretely (variables may vary depending on the culture in question).

Alongside a good model fit, study 2 tested sequential reasoning from appeals to specific candidates. Although only tentative, the results suggest that participants were influenced by a recency effect such that the statement of the final election candidate was more influential than the statement of the first candidate. Although displaying general tendencies that are predictable from the Bayesian model, this suggests that participants were experiencing heuristic biases and therefore are not perfectly Bayesian.

Descriptively, the data suggests that participants may have been expecting political disagreement between canddiates from different parties (given recent polarization in American politics, such an assumption may not be unreasonable). In the four cases in which politicians disagreed on the policy (e.g. participants were told that Sanders supported the policy while Bush attacked it), the summated differences between mean predictions and mean observations were 8.31. Comparitively, the summated differences when politicians agreed with each other were 32.51 across four conditions. Although descriptive, this suggests particiants were expecting disagreement between the candidates and agreement introduced additional noise into their responses.

Concluding remarks

The paper tested a Bayesian source credibility model integrating expertise and trustworthiness to determine the persuasive potential of an appeal to that particular source.

Overall, the model enjoys a good fit with with observed estimations of the likelihood of the goodness of an unknown policy given public support of attack of a candidate. As expected, the model has a better population than individual fit. However, results from the studies suggest that the Bayesian source credibility approach applicable to appeals to specific and known political candidates as a model to predict the persuasive potential of the person. The paper extends previous literature by applying a parameter-free Baysian source credibility model in political context when a named candidate has publically supported or attacked a policy.

More research is required to determine individual differences for integrating evidence from uncertain sources, but that is beyond the scope of the present paper.

Acknowledgments

The Danish Council of Independent Research, DFF – 1329-00021B, supported the research. I am grateful to Dr. Adam Harris for very useful comments on an earlier draft.

[7] BS deontes 'Bush supports', SA 'Sanders attack', etc.

References

Ayres, I. & Braithwaite, J. (1992) *Responsive Regulation*, Oxford University Press

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility, *Memory and Cognition 4,* 330-336.

Bovens, L., & Hartmann, S. (2003) *Bayesian epistemology*. Oxford: Oxford University Press.

Briñol, P. & Petty, R. E. (2009) Source factors in persuasion: A self-validation approach, *European Review of Social Psychology* 20, 49-96

Chaiken, S. & Maheswaran, D. (1994) Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement, *Journal of Personality and Social Psychology* 66 (3), 460-473

Cialdini, R. B. (2007) *Influence: The Psychology of Persuasion*, Collins Business

Corner, A., Hahn, U. & Oaksford, M. (2011) The psychological mechanism of the slippery slope argument*, Journal of Memory & Language* 64*,* 133-152.

Cuddy, A. J. C., Glick, P. & Beninger, A. (2011) The dynamics of warmth and competence judgments, and their outcomes in organizations, *Research in Organizational Behavior* 31, 73-98

Fiske, S. T., Cuddy, A. J. & Glick, P. (2007) Universal dimensions of social cognition: warmth and competence, *Trends in Cognitive Sciences* 11, 77-83

Fukuyama, F. (1995) *Trust*, New York: Basic Books

Hahn, U., Harris, A. J. L., & Corner, A. (2009) Argument content and argument source: An exploration, *Informal Logic 29,* 337-367.

Hahn, U., & Oaksford, M. (2006) A normative theory of argument strength, *Informal Logic* 26*,* 1-24

Hahn, U., & Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies, *Psychological Review 114*, 704-732

Hahn, U., Oaksford, M. & Harris, A. J. L. (2012) Testimony and argument: A Bayesian perspective*,* in Zenker, F. (Ed.) *Bayesian Argumentation*, Dordrecht: Springer, 15-38

Harris, A. & Corriveau, K. H. (2011). Young children's selective trust in informants, *Philosophical Transactions of the Royal Society B* 366, 1179-1187

Harris, A., Hsu, A. & Madsen, J. K. (2012) Because Hitler did it! Quantitative tests of Bayesian argumentation using *Ad Hominem*, *Thinking & Reasoning* 18 (3), 311-343

Harris, A., Hsu, A., Madsen, J. K. & Hahn, U (in press) The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach, *Cognitive Science*

Hetherington, M. J. (1999) The effect of political trust on the presedential election, 1968-96, *American Political Science Review* 93 (2), 311-326

Howson, C., & Urbach, P. (1996). *Scientific Reasoning: The Bayesian Approach (2nd Edition)*. Chicago, IL: Open Court

Householder, E. E. & LaMarre, H. L. (2014) Facebook Politics: Toward a Process Model for Achieving Political Source Credibility Through Social Medie, *Journal of Information Technology & Politics* 11 (4), 368-382

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning, *Argument & Computation* 4 (1), 46-63.

Levi, M. & Stoker, L. (2000) Political Trust and Trustworthiness, *Annual Review of Political Science* 3, 475-507

McCroskey, J. C. (1997) Ethos: A Dominant Factor in Rhetorical Communication, in McCroskey, J. C. (Ed.) *An Introduction to Rhetorical Communication* (7th ed.), Allyn and Bacon, 87-107

Miller, A. & Listhaug, O. (1990) Political performance and institutional trust, in Norris, P. (Ed.) *Critical Citizens: Global Confidence in Democratic Government* (pp. 204-216), Oxford, UK: Oxford University Press

Oaksford, M. & Chater, N. (1991) Against logicist cognitive science, *Mind and Language* 6, pp. 1-38

Oaksford, M. & Chater, N. (2007) *Bayesian Rationality: the probabilistic approach to human reasoning*, Oxford University Press, Oxford: UK

Oaksford, M. & Hahn, U. (2004) A Bayesian approach to the argument from ignorance, *Canadian Journal of Experimental Psychology 58*, 75-85

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010) Running experiments on Amazon Mechanical Turk, *Judgement and Decision Making* 5, 411–419

Petty, R. E. & Cacioppo, J. T. (1981) *Attitudes and persuasion: Classic and contemporary approaches*, Boulder, CO: Westview Press

Petty, R. E. & Cacioppo, J. T. (1984) Source Factors and the Elaboration Likelihood Model of Persuasion*, Advances in Consumer Research* 11, 668-672

Rezlescu, C., Duchaine, B., Olivola, C. Y. & Chater, N. (2012) Unfakeable Facial Configurations Affect Strategic Choices in Trust Games with or without Information About Past Behavior, PLoS One 8 (3), e34293

Schum, D. A. (1981) Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence, *Organizational Behavior and Human Performance* 27, 153-196

Tormala, Z. L., Clarkson, J. J. (2007) Assimilation and Contrast in Persuasion: The Effect of Source Credibility in Multiple Message Situations*,* Personality and Social Psychology Bulletin, 33 (4), 559-571

Walton, D. (1997) *Appeal to Expert Opinion: Arguments from Authority*. University Park, PA: Pennsylvania State University Press.

Expressive faces are remembered with less pictorial fidelity than neutral faces

Martina Lorenzino (martina.lorenzino@unifi.it)
Department of Neurosciences, Psychology, Drug Research, and Child Health (NEUROFARBA), University of Florence
via di San Salvi 12, Complesso di San Salvi, Padiglione 26, 50135 Firenze (FI), Italy

Giorgio Gronchi (giorgio.gronchi@unifi.it)
Department of Neurosciences, Psychology, Drug Research, and Child Health (NEUROFARBA), University of Florence
via di San Salvi 12, Complesso di San Salvi, Padiglione 26, 50135 Firenze (FI), Italy

Corrado Caudek (corrado.caudek@unifi.it)
Department of Neurosciences, Psychology, Drug Research, and Child Health (NEUROFARBA), University of Florence
via di San Salvi 12, Complesso di San Salvi, Padiglione 26, 50135 Firenze (FI), Italy

Abstract

A repeated finding in the literature of face recognition is that expressive faces are remembered better than neutral faces. However, a better facial-identity recognition may come at a cost of a reduced precision with which the pictorial facial features, irrelevant for identity recognition, are represented in memory. By means of a continuous-report task, we tested this hypothesis by measuring the memory precision of expressive and neutral faces. Commensurable face-identity and facial-expressions variations were generated with the method of Fechnerian scaling. The results confirm our hypothesis, but only under conditions of high memory load. We interpret the present findings as due to the effects of the categorical processes required for facial-identity recognition.

Keywords: emotion; faces; continuous-report procedure; visual working memory

Introduction

It has been often shown that expressive faces are easier to remember than neutral faces (Jackson et al., 2008; Langeslag et al., 2009). Some studies have found an advantage for the memory of positive expressive faces compared to negative faces (D'Argembeau & Van der Linden, 2007; D'Argembeau et al., 2003; Shimamura, Ross, & Bennett, 2006). However, the greatest evidence for a memory advantage for expressive versus neutral faces comes from studies comparing memory for facial expressions with a negative valence and with neutral faces (Jackson et al., 2008; Jackson, Linden, & Raymond, 2014; Jackson et al., 2009; Sessa et al., 2011). This memory advantage has been attributed to a more accurate encoding and memory maintenance for expressive faces, compared to neutral faces (Jackson et al., 2014; Ridout et al., 2003; Sergerie, Lepage, & Armony, 2005; Sessa et al., 2011).

It is important to point out that the majority of studies in this area of research has employed an old-new face recognition task, in which participants are asked to decide whether a test face had been previously presented in an array of expressive faces, or not. In this respect, it should be noted that a better performance in a old-new recognition task does not necessarily imply a better pictorial resolution for the memory representation of expressive faces compared to neutral faces. In fact, the memory representation of a face endowed of a high pictorial fidelity contains all those transient facial features which are irrelevant for face-identity categorization. On the other hand, what characterizes face-identity recognition is exactly the ability to isolate the identity-defining properties of a face, while disregarding its transient features (*e.g.*, pose, lighting, transient muscles variations, and so on). Thus, in general, one might expect that a better ability to recognize a face comes to the cost of a worst pictorial fidelity with which that face is represented in memory.

The aim of the present study is to measure the pictorial fidelity of the Visual Working Memory (VWM) representations of expressive and neutral faces. In this regard, we can make a specific prediction which is motivated by the following considerations. All the rest being the same, we should expect that is more important to remember the identity of a stranger if, in a casual encounter, she/he displays an angry expression (or a happy expression, for that matter) rather than a neutral face (Jackson et al., 2009; Jackson et al., 2014). This seems obvious, considering the implications of this situation for future social interactions with that person. On the top of this, we should also consider the fact that, as remarked above, the creation of a facial-identity category requires the selection of the identity-defining features and the omission of the category-irrelevant features. Specifically, the transient pictorial features of a face (the image properties that vary as a function of illumination, pose, and so on) are certainly irrelevant for a facial-identity category. Therefore, we hypothesize that (1) these transient pictorial properties are represented with a lower pictorial fidelity in memory when a facial-identity category is created, and (2) this process is stronger for expressive compared to neutral faces, because it is more important to recognize facial identity in the former case than in the latter.

A face can be represented in memory with different degrees of pictorial fidelity. For example, we can remember the gist of a face and disregard its pictorial details (Leal, Tighe, & Yassa, 2014), or we keep in memory a very detailed photographic representation of a face. By adapting the procedure of Wilken and Ma (2004), in the present study we asked participants to select from a morphing continuum the image of a learned face (*e.g.*, Bays & Husain, 2008; Lorenc et al., 2014; Zhang & Luck, 2008). We expect to find (1) a lower pictorial fidelity of VWM for expressive compared to neutral faces, and (2) a decrease in VWM fidelity with the increase of the set size of

to-be-remembered faces.

Experiment

The degree of pictorial fidelity with which a face is represented in memory can be measured with the continuous report task (Wilken & Ma, 2004). Although this task has been mostly employed with simple features (such as orientation or color, for example), it can also be used with faces. In the present study, participants were required to select a probe within a morphing continuum in order to match it to a learned face. This task was performed with face images belonging to two different kinds of morphing continua. A first continuum provided facial-identity variations with a neutral facial expression. A second continuum provided emotional expressions variations while keeping constant facial identity.

In general, it's not possible to directly compare distances along different morphing continua, because there is not a linear relation between the physical distance on such continua and the similarity space among the stimuli in the mental memory representations (Caudek, 2013; Danilova & Mollon, 2014). To deal with this problem, we generated the two morphing continua with the technique of Fechnerian scaling, which creates (for each participant) perceptually commensurable spaces for identity and expression variations. In this manner, the memory fidelity for the two classes of stimuli can be directly compared because, in both cases, the unit of measurement for each continuum corresponds to each subject's threshold of perceptual discrimination (Domini & Caudek, 2009; Dzhafarov & Colonius, 1999).

Method

Participants Eight subjects (n=8; 7 females; age range 26-30 years) with normal or corrected-to normal vision participated voluntarily to the Experiment. They were naïve to the purpose of the study and none of them had previously been exposed to the stimuli employed. The experiment was undertaken with the understanding and written consent of each participant. The experiment conformed with the institutional and national guidelines for experiments with human subjects and with the Declaration of Helsinki.

Stimuli Seven Caucasian faces (all males) were selected from the Karolinska Directed Emotional Faces database (Lundqvist, Flykt, & Öhman, 1998): Five images showed five different facial identities with a neutral expression (A, B, C, D, E, Figure 1a) and two images represented one of the previous facial identities with an happy and an angry emotional expression (100hA, 100aA, Figure 1b-c). All images were converted to grayscale, an oval aperture was superimposed on each face, and the ovals were placed on a transparent background. Each face image measured 280 × 329 pixels and all stimuli were matched for luminance and contrast values. We generated three different morphing continua: (1) a continuum between two face identities (indicated with A and B), both having a neutral expression (Figure 1a, top row); (2) a continuum between a neutral face A to an "happy" face (50hA)[1] with the same identity (Figure 1b, top row); (3) a continuum between a neutral face A to an "angry" face (50aA)[2] with the same identity (Figure 1c, top row).

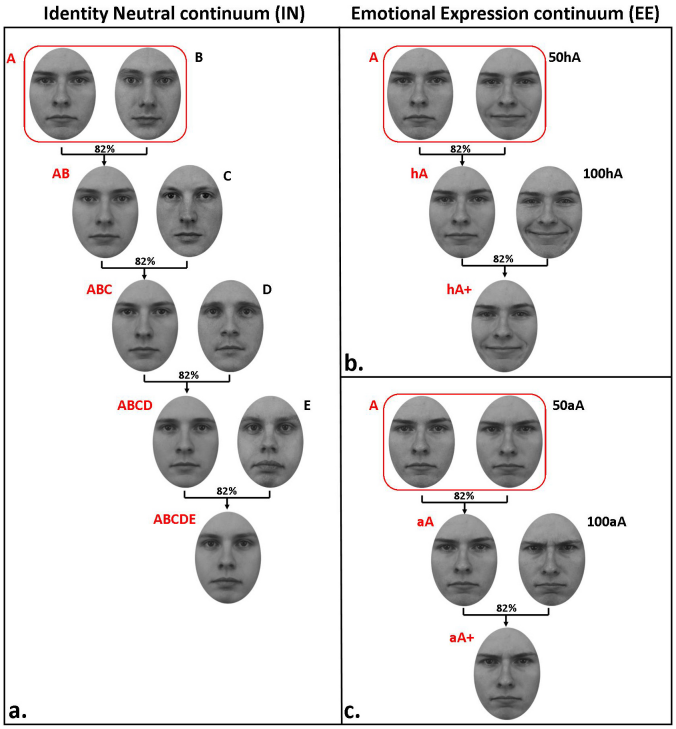


Figure 1: **(a)**. Creation of the Identity Neutral continuum. A 100-steps morphing was created between image A and image B. The image that produced an 82% rate of correct discrimination between A and B was employed as the new extreme (AB) for the creation of a new morphing continuum. This procedure has been repeated in order to obtain four thresholds of perceptual discrimination (AB, ABC, ABCD, ABCDE). **(b-c)**. Creation of Emotional Expression continuum. A 100-steps morphing was created between image A and image 50hA and between image A and 50aA. The images that produced an 82% rate of correct discrimination between A and 50hA and between A and 50aA were employed as the new extremes (hA and aA, respectively) for the creation of a new morphing continuum. This procedure has been repeated in order to obtain four thresholds of perceptual discrimination (hA, hA+, aA and aA+).

Procedure *Phase 1.* With the method of Fechnerian scaling, we generated the morphing continua that were then used in Phase 2. A 2AFC discrimination task was used for determining face discrimination thresholds between pairs of images belonging to each morphing continuum. For the fa-

[1]The 50hA image corresponded to the image numbered 50 of a morphing continuum between A and 100hA.

[2]The 50aA image corresponded to the image numbered 50 of a morphing continuum between A and 100aA.

cial identity variations, for example, we created a morphing continuum between two facial identities. We kept fixed one image (origin of the scale) and then, with psychometric methods, we selected a second face image which produced, for each participant, an 82% success rate in a perceptual-discrimination task. We then morphed this second image with a third face identity and repeated the procedure (Figure 1a). In this manner, we selected, for each participant, four face images (three pairs of images, each one characterized by one threshold of perceptual discrimination). We then linearly morphed each of the successive pairs of images, so as to generate the Fechnerian continuum for face-identity variations (Figure 2a). A similar procedure was used for generating the Fechnerian continuum for face-expressions variations (Figure 1b-c; 2b).

To determine each perceptual discrimination threshold, in each trial we presented the reference face (*i.e.*, the frame number 10 within the morphing continuum) and one of seven test faces corresponding to greater distances along the morphing continuum from the reference face. A 1.000 ms fixation cross indicates the start of each trial, followed by a 1.000 ms blank screen. A comparison face (frame 100 of the morphing continuum) was centrally presented for 1.000 ms. After a 1.000 ms blank screen, the reference and test faces were centrally presented side-by-side. Participants were asked to decide which of the two simultaneously presented faces was more similar to the comparison face presented before each trial. Stimuli were shown on the screen until the participant's response (Figure 3a).

A psychometric function was computed for each participant. The psychometric function was defined by the equation $P(x) = \gamma + (1 - \gamma - \lambda)p(x)$, where γ is the lower asymptote, λ is 1 - the upper asymptote, and $p(x)$ is the base psychometric function, varying between 0 and 1 (Wichmann & Hill, 2001)[3]. From this fitted psychometric curve we then derived the stimulus image corresponding to an 82% rate of correct discrimination.

Phase 2. Each trial started with a fixation point for 1.000 ms, then one or four faces were presented on the screen for 4.000 ms. After a 1.000 ms blank screen, participants were asked to select, from the Fechnerian continua generated in phase 1, the image that more closely matched the test stimulus. Participants could move along the Fechnerian continuum with a keypress, with only one image being shown on the screen at any time.

For set-size four, each trial started with a fixation point for 1.000 ms, then four faces were shown on the screen for 4.000 ms. After a 1.000 ms blank screen, four grey ovals were shown on the screen for 3.000 ms, in the same positions as the previously presented faces, with a red dot indicating the to-be-remembered face (Figure 3b).

Each session included 160 trials with 8 blocks of 20 trials each and lasted for about 1 hour. Each participant completed a total of 320 trials (160 $\times$ 2 set-sizes). The two set-sizes were tested in separate sessions, with the order of presentation of the two sessions being counterbalanced among participants.

[3]Psychometric functions were fitted by using `R` (R Core Team, 2013) and `psyphy` (Knoblauch, 2007)

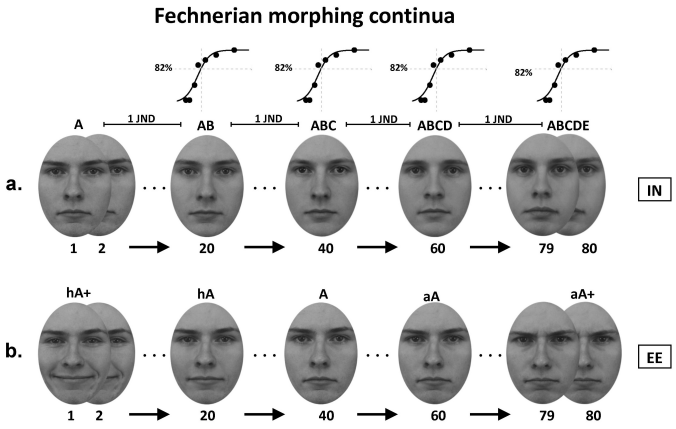


Figure 2: Each Fechnerian morphing continuum was created, for each participant, by morphing between the four images corresponding to the face discrimination thresholds measured in Phase 1 and consisted of 80 images. We generated two different types of Fechnerian continua: One Fechnerian morphing continuum of facial identity variations with a neutral expression (Identity Neutral [IN] continuum) and one Fechnerian morphing continuum of variations of emotional expression (Emotional Expression [EE] continuum). **(a)** IN continuum. The continuum was created by generating 20-steps morphing between faces corresponding to the discrimination thresholds obtained in the first phase (A/AB, AB/ABC, ABC/ABCD, ABCD/ABCDE). **(b)** EE continuum. The continuum was created by generating 20-steps morphing between faces corresponding to the discrimination thresholds obtained in the first phase (hA+/hA, hA/A, A/aA, aA/aA+).

Results

Figure 4 shows the average distance between the test face and participants' response, for each Fechnerian morphing continuum and for each set-size. The average response error was smaller than the average perceptual discrimination thresholds measured in Phase 1. The difference in memory fidelity between the IN and EE continua was not statistically significant, $\chi^2_1 = 0.44$, $p = .5075$. The effect of Set-size was statistically significant, $\chi^2_1 = 3.85$, $p = .049$. More importantly, the Continuum $\times$ Set-size interaction was statistically significant, $\chi^2_1 = 10.49$, $p = .0012$. We also divided the EE continuum into two segments ("happy faces," from frame 1 to frame 40, and "angry faces," from frame 41 to frame 80). However, the effect of face valence (positive, negative) was not statistically significant, $\chi^2_1 = 1.60$, $p = .2057$. The Face Valence $\times$ Set-size interaction was not statistically significant, $\chi^2_1 = 0.19$, $p = .6599$.

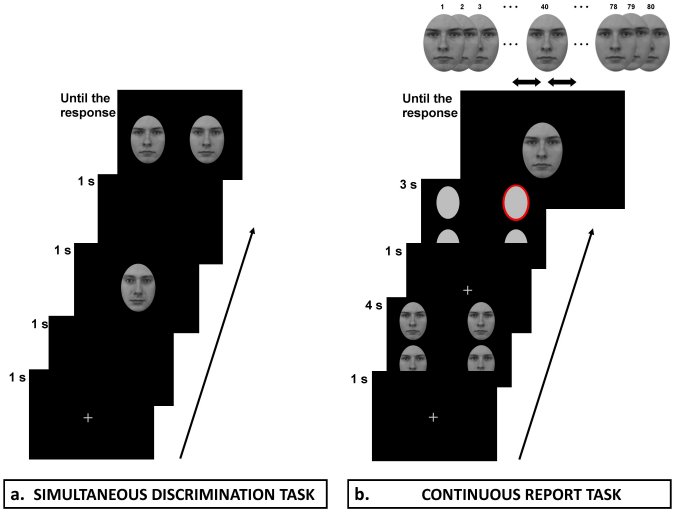


Figure 3: Experimental procedure. **(a)** In the simultaneous discrimination task (Phase 1) participants were forced to choose the face more similar to a comparison face previously shown. **(b)** In the continuous report task (Phase 2) a set (n = 1 or n = 4) of randomly chosen faces from a Fechnerian morphing continuum were shown (the figure reports the 4 items set-size condition). After 1 second interval, a cue indicated which image had to be remembered. In order to produce the response, participants manipulated a scalar probe that could vary along the morphing continuum.

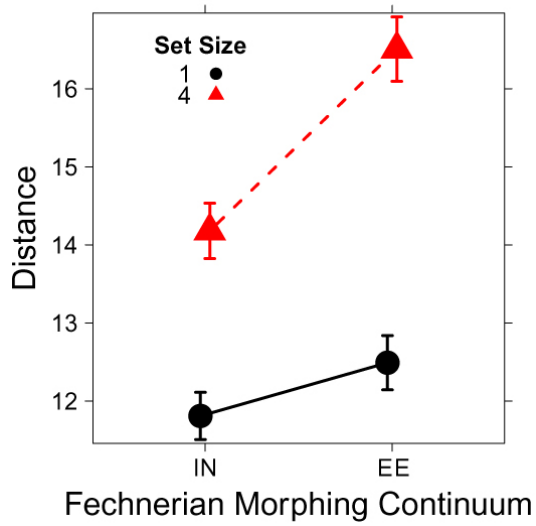


Figure 4: Average of the absolute value of the difference between the value of the Fechnerian morphed face reported by each participant and the value of the studied Fechnerian morphed face, for the IN and EE continua, as a function of the set-size.

Discussion

We asked whether the memory advantage for emotional compared to neutral faces, which has been found with the old-new recognition task, occurs at the expenses of the pictorial precision with which the expressive faces are represented in memory. The recognition of face identity is the consequence of an act of categorization. Each act of categorization weakens the representation of the category-irrelevant features (Caudek, 2013). We assumed that remembering face identity is more important for expressive than for neutral faces. As a consequence, we hypothesized that expressive faces should be represented in memory with a lower pictorial fidelity than neutral faces. Our results confirm this hypothesis, but only for set-size of four (not for set-size of one). We interpret this result as indicating that the negative impact of the categorization processes on the pictorial fidelity of the memory representation is stronger under a high memory load. With a low memory load, in fact, we found no difference in the precision of VWM for neutral and expressive faces, thus replicating the results of Bankó, Gál, & Vidnyánszky (2009).

Although it could be expected that face identity recognition is more important in the case of angry compared to happy faces (Jackson et al., 2009; Jackson et al., 2014), we did not find a difference between these two conditions. However, it should be noted that in the present study the level of emotional intensity was very low. Therefore, we could expect different effects on memory fidelity for angry compared to happy faces when emotional intensity is higher. Finally, in line with the previous literature, we found a decrease of VWM precision with an increase of set size (Alvarez & Cavanagh, 2004; Bays & Husain, 2008; Brady, Konkle, & Alvarez, 2011; Caudek, 2013).

A few limitations deserve mention. In particular, the present experimental design does not allow to determine if results are due to differences in memory for emotional versus neutral faces, or differences in memory for expression versus identity. To address this limit, in future experiments we plan to compare memory for emotional versus neutral faces by using the same type of response continuum, by producing stimulus variations, for example, by a change in the illuminant direction or in the orientation of the face, or by considering the subtle transformations of the morphological properties of the face during speech and language production. A second limit of the present study concerns the low statistical power due to small sample size. Finally, another issue to consider in future research has to do with the impact that face familiarity may have on VWM fidelity. We expect that familiar faces, by encouraging an act of individuation more strongly than unfamiliar faces, will show a lower VWM fidelity for transient facial features than unfamiliar faces. At the upper extreme of the familiarity spectrum, the self-face is the most important face (Kircher et al., 2001). Therefore, we speculate that VWM fidelity for irrelevant information should be lower for the self-face than for other personally-familiar faces.

Conclusions

The VWM advantage for expressive compared to neutral faces that had been found with the old-new recognition task

comes with a cost: Expressive faces are represented in VWM with a lower pictorial fidelity than neutral faces. We interpret this result as due to the effects of the categorical processes required for facial identity recognition.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15, 106-111.

Bankó, M., Gál, V., & Vidnyánszky, Z. (2009). Flawless visual short-term memory for facial emotional expressions. *Journal of Vision*, 9, 1-13.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, 321, 851-854.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, 11, 1-34.

Caudek, C. (2013). The fidelity of visual memory for faces and non-face objects. *Acta psychologica*, 143, 40-51.

D'Argembeau, A., & Van der Linden, M. (2007). Facial expressions of emotion influence memory for facial identity in an automatic way. *Emotion*, 3, 507-515.

D'Argembeau, A., Van der Linden, M., Comblain, C., & Etienne, A.-M. (2003). The effects of happy and angry expressions on identity and expression memory for unfamiliar faces. *Cognition and Emotion*, 17, 609-622.

Danilova, M. V., & Mollon, J. D. (2014). Symmetries and asymmetries in chromatic discrimination. *JOSA A*, 31, A247-A253.

Domini, F., & Caudek, C. (2009). The intrinsic constraint model and Fechnerian sensory scaling. *Journal of Vision*, 9, 1-15.

Dzhafarov, E. N., & Colonius, H. (1999). Fechnerian metrics in unidimensional and multidimensional stimulus spaces. *Psychonomic Bulletin & Review*, 6, 239-268.

Jackson, M. C., Linden, D. E., & Raymond, J. E. (2014). Angry expressions strengthen the encoding and maintenance of face identity representations in visual working memory. *Cognition and Emotion*, 28, 278-297.

Jackson, M. C., Wolf, C., Johnston, S. J., Raymond, J. E., & Linden, D. E. (2008). Neural correlates of enhanced visual short-term memory for angry faces: an FMRI study. *PLoS One*, 3, e3536.

Jackson, M. C., Wu, C. Y., Linden, D. E., & Raymond, J. E. (2009). Enhanced visual short-term memory for angry faces. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 363-374.

Kircher, T., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P., Bullmore, E. T., ... David, A. S. (2001). Recognizing one's own face. *Cognition*, 78, 1-15.

Knoblauch, K. (2007). psyphy: Functions for analyzing psychophysical data in R. R package version 0.0-5, URL http://CRAN. R-project. org/package= psyphy.

Langeslag, S. J., Morgan, H. M., Jackson, M. C., Linden, D. E., & Van Strien, J. W. (2009). Electrophysiological correlates of improved short-term memory for emotional faces. *Neuropsychologia*, 47, 887-896.

Leal, S. L., Tighe, S. K., & Yassa, M. A. (2014). Asymmetric effects of emotion on mnemonic interference. *Neurobiology of Learning and Memory*, 111, 41-48.

Lorenc, E. S., Pratte, M. S., Angeloni, C. F., & Tong, F. (2014). Expertise for upright faces improves the precision but not the capacity of visual working memory. *Attention, Perception, & Psychophysics*, 76, 1975-1984.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, 91-630.

Ridout, N., Astell, A. J., Reid, I. C., Glen, T., & O'Carroll, R. E. (2003). Memory bias for emotional facial expressions in major depression. *Cognition and Emotion*, 17, 101-122.

Sergerie, K., Lepage, M., & Armony, J. L. (2005). A face to remember: Emotional expression modulates prefrontal activity during memory formation. *NeuroImage*, 24, 580-585.

Sessa, P., Luria, R., Gotler, A., Jolicur, P., & Dell'Acqua, R. (2011). Interhemispheric ERP asymmetries over inferior parietal cortex reveal differential visual working memory maintenance for fearful versus neutral facial identities. *Psychophysiology*, 48, 187-197.

Shimamura, A. P., Ross, J., & Bennett, H. (2006). Memory for facial expressions: The power of a smile. *Psychonomic Bulletin & Review*, 13, 217-222.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63, 1293-1313.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, 4, 1120-1135.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453, 233-235.

Three barriers to effective thought experiments, as revealed by a system that externalizes students' thinking

Miki Matsumuro (muro@cog.human.nagoya-u.ac.jp)
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University
Fro-cho, Chikusa-ku, Nagoya, Japan

Abstract

This study aimed to develop a Thought Experiment Externalizer (TE-ext) and to apply it in order to observe barriers to problem solving. TE-ext enables students to visualize a problem situation. Users of TE-ext can implement changes in the situation and see the result as an animation. Experimental use of TE-ext identified three barriers to conducting an effective thought experiment (TE). First, participants tended not to change the situation from the original one; second, incorrect or inappropriate knowledge was applied to the situation; third, the participants did not apply the results of their TE to other situations. These factors prevented participants from rejecting their initial incorrect model and finding a new one through TEs.

Keywords: Thought experiment; scientific reasoning; science education system; problem solving

Introduction

Interest in thought experiments (TEs) has increased in recent decades (Matthews, 2014; Nersessian, 1992). A TE is an experiment conducted in one's mind. In this paper, we describe an effort to externalize TEs through a system that we call a Thought Experiment Externalizer (TE-ext). Initially, we review previous studies and identify research needs. Next, we introduce the TE-ext and our experiments using this system. Finally, we discuss the potential of TE-ext as an educational system.

Previous Studies

Previous studies can be divided into three categories based on their methods of collecting data.

Historical Research Some studies, particularly in the field of philosophy of science, have focused on specific historical cases (Brown, 1991, 2006; Nersessian, 2002). There are many renowned examples of TEs in history, such as those conducted by Galileo, Newton, and Einstein. Using these cases, the studies developed a definition of TEs and argued for their importance in the history of science.

Fieldwork Other studies have investigated the nature of TEs by collecting data in laboratories or classrooms (Lattery, 2001; Stephens & Clement, 2010). These studies have provided evidence that TEs are used not only by famous thinkers but by researchers, teachers, and students in their daily activities. Furthermore, they proposed cues to identify TE use and the types of situations wherein reasoners tended to conduct TEs.

Protocol Research Some researchers gave participants a physics problem and collected protocol data as the participants solved the problem (Clement, 2009; Kösem & Özdemir, 2014). They found differences between the TEs conducted by physics experts and novices. They also elucidated the TE process and its role in the scientific reasoning.

Summary of Knowledge of Thought Experiments Here, we briefly summarize the knowledge gained from these previous studies. There is no exact and agreed-upon definition of a TE, but Brown (2006) and Reiner (2006) proposed an operational definition. Their proposals were summarized by Kösem and Özdemir (2014) in terms of four steps, as illustrated below using the example of Galileo's TE that caused him to reject Aristotle's theory.

Step 1: Visualize a situation
(I am at the top of the tower with a heavy ball and a light ball)
Step 2: Carry out operation
(Connect each other ball and drop it from the tower)
Step 3: Apply background information
(Based on Aristotle's theory)
Step 4: See the result
(The theory leads to two contradictory results)

A TE involves visualizing a situation in one's mind and then imagining the result. It is a kind of mental simulation, which is a strategy in mechanical reasoning (Hegarty, 2004). The key step in the TE is the selection of the operation to be carried out in Step 2. Appropriate operations enable the reasoner to access implicit knowledge in Step 3, leading to a useful result in Step 4. Finally, a conclusion is drawn from the results of the TEs.

Justification for the Present Research

In this section, we describe what we consider to be the necessary research task from two perspectives.

Science Education Few studies have attempted to facilitate students' performance of TEs or to enable them to understand the benefits of TEs. Previous studies have been limited to examining the use of TEs by teachers and students in a daily classroom (Lattery, 2001; Stephens & Clement, 2010). Monaghan and Clement (1999) successfully facilitated mental simulation by means of a computer simulation. However, their face-to-face method would not be suitable for use with a large class.

Our TE-ext, which will be described below, has advantages for science education. Students can easily carry out operations because the available operations are visible. Additionally, the externalization of the TE reduces the cognitive load,

making additional cognitive resources available for reasoning.

Psychological Research Almost all previous studies involved no experimental operations, as they focused on reasoning in a naturalistic situation. Protocol analysis has been a prominent method in these studies, since TEs are conducted in the reasoners' mind. These studies have contributed to our knowledge of the process and nature of TEs. Based on this accumulated knowledge, we can learn more about TEs through experimental operations.

Our TE-ext is a useful tool for psychological experimentation because it provides us with quantitative data, such as sequences of operations. In this study, we use TE-ext and some experimental operations to investigate barriers preventing students from conducting an effective TE.

TE-ext

Material

We used the yoyo problem from Anzai and Yokoyama (1984). The problem was to predict the movement of the yoyo in Figure 1. In our study, students had to decide in which direction the yoyo would move (left, right, or staying in place) and in which direction it would rotate (clockwise, counterclockwise, or remaining still). The correct answer was that the yoyo would rotate clockwise and roll to the right. The direction of rotational momentum caused by the tension force in the string and the center of rotation at the yoyo's point of contact with the floor determine the movement.

The yoyo problem had three advantages. First, Anzai and Yokoyama (1984) defined students' internal models. In the students' naïve model, they believed that the yoyo's axis was the center of the rotation. Based on this model, they answered that the yoyo would rotate counterclockwise and roll to the left. The second advantage was that a cue capable of causing the students to shift to a correct model had already been demonstrated. When asked to think about a square object instead of a round yoyo, the students predicted that it would be dragged to the right and realized where the center of the rotation was (Anzai & Yokoyama, 1984). This insight led them to the correct answer. Therefore, to solve the problem through TEs, it was important to carry out the mental operation of changing the yoyo's shape into a square and pulling the string. Finally, the effects of the predicted barriers at each step could be associated with particular responses to the yoyo problem (see the discussion on barriers below).

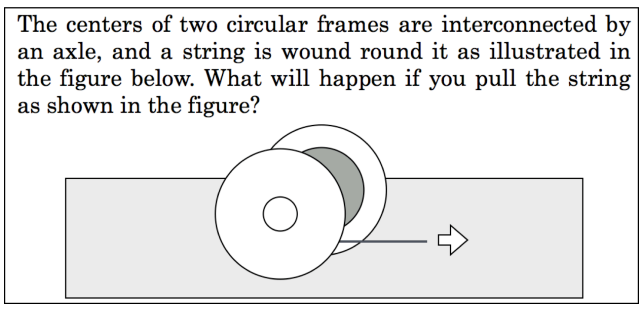


Figure 1: The yoyo problem.

System's Function

TE-ext externalizes Steps 1, 2, and 4 of the TE. Figure 2 shows sample screenshots of important parts of TE-ext. The properties of the yoyo shown in Figure 2(a) represented the problem situation, which was the initial state of TE-ext. The students could observe the yoyo from the side. Since TE-ext visualizes a problem situation on a PC monitor, users do not have to carry out the visualization in their mind in Step 1.

The students could carry out some operations on the yoyo in the TE-ext system, thereby externalizing Step 2 of the TE process. They could change the size and shape of the yoyo, the location of the end of the string, the width of the string, and the amount of string wound around the yoyo. The size and location of the end of the string could be changed by dragging the red and black rectangles, respectively. Other properties could be selected on the option panel (Figure 2(b)).

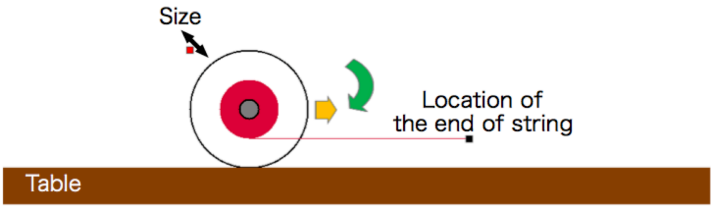
The students could then see the result (Step 4) in TE-ext by selecting a direction and clicking the "move" button. The direction of movement and rotation were selected using the animation buttons (movement: up, down, right, left; rotation: clockwise, counterclockwise). The yoyo in TE-ext moved in accordance with the selected buttons, even if such movement violated physical laws. The role of TE-ext was not to provide new knowledge but to externalize TEs; the animations were consistent with the fact that, in a TE, any prediction is possible. We then investigated barriers preventing students from conducting an effective TE using TE-ext.

Three Barriers

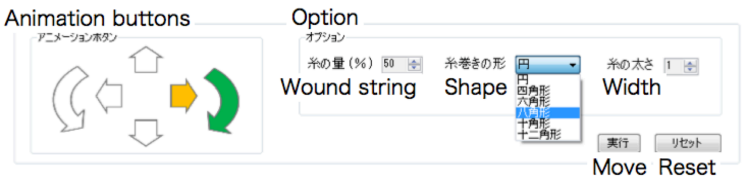
We defined an effective TE as one that causes students to shift their mental model. In scientific reasoning and problem solving, it is important to reject an incorrect model or concept and to find a new correct one (Vosniadou & Brewer, 1992). We predicted three barriers to an effective TE, or to achieving a shift in one's mental model through TEs, in alignment with the thinking steps described by Kösem and Özdemir (2014).

The first barrier would exist at Step 2, during which the students carry out operations. To conduct an effective TE, they have to change a situation into one invoking implicit knowledge that they could not apply to the initial situation. In the case of the yoyo problem, an operation to give the yoyo wheels a square shape is crucial. However, people are apt to avoid testing a situation that would violate their model (Wason, 1960). It would be difficult for most people to carry out an operation that makes the situation different from the initial one. Thus, students could be expected to conduct TEs in the original situation with round yoyos.

The second barrier would exist at Step 3. Even if students create a good experimental situation, they cannot always apply appropriate background information or knowledge. In this case, they may not be able to imagine the square-shaped yoyo being dragged to the right. It is easy for students to rely instead on the knowledge already in their mind. Therefore,



(a) The yoyo in TE-ext



(b) Option panel and animation buttons

Figure 2: Sample screenshots of TE-ext. Figure (a) shows the yoyo on the table; figure (b) shows the option panel and the animation buttons. The right and clockwise animation buttons are selected in the figure. Each shot captures important parts in TE-ext. The other part of the screen is used to display the yoyo's movement.

they may apply their incorrect model to the situation, leading to an inappropriate result.

The third barrier would exist at Step 4. The students have to derive a conclusion from the results of their TE. However, the results violating their current model are not easily applied to the solution of the original problem. In the yoyo problem, this would mean that the students, although they understand that a square-shaped yoyo is dragged to the right, would not be able to apply this result to the original round yoyo. Some previous studies have revealed that even when participants or scientists confronted contradictory data, they did not change their initial model (Chinn & Brewer, 1993; Dunbar, 1995). Instead, to retain their model, they reinterpreted the data or added peripheral explanations. The results of these studies suggest that the students cannot easily change their initial model, even if they see the appropriate results of a TE.

Using TE-ext, we investigated whether these barriers exist and whether we could eliminate them.

Methods

Participants

Nineteen undergraduate students participated in the experiment. Their major fields included literature, education, law, and economics. All students had basic knowledge of elementary physics from junior high school. Eleven of them had studied physics in high school; none had taken a physics course at the university level.

Procedures and Predictions

The experiments were composed of four phases. Procedures, aims, and predictions in each phase are summarized in Table 1. The students answered the same question in each phase. They observed a picture of the yoyo, read the problem as presented in Figure 1, and then selected the direction of movement and rotation. Additionally, they were asked to write down the reason for their decision.

Phase 0 included only the test. The aim here was to confirm that the students would answer "left and counterclockwise," which meant that they had the incorrect initial model. After Phase 0, the experimenter instructed the students on how to use TE-ext.

Following the introduction of TE-ext, Phase 1 was used to confirm the presence of the first barrier. During this phase, the students were asked to work on the yoyo problem for 10 minutes using TE-ext freely (i.e., thinking time) and then to answer the test question again. If the first barrier existed, it would prevent the students from carrying out suitable operations to change the situation from the original one. Therefore, they would conduct more animations with the round yoyo during their thinking time and therefore end up with the same result as in Phase 0.

We aimed to break the first barrier in Phase 2 by providing instruction along with thinking time and the test. First, the experimenter instructed the students, as a means of solving the yoyo problem, to change the properties in TE-ext to make the situation different from the original one. Additionally, she showed some examples of possible situations, including the square-shaped yoyo. After this instruction, the students worked on the yoyo problem for 10 minutes and then took the test again. If the instruction had overcome the first barrier, the students would conduct animations with yoyos of various shapes including a square one, during their thinking time. However, even if the first barrier was gone, the students would still be likely to select the answers of left and counterclockwise on the test due to the second and/or third barriers.

Phase 3 had two purposes. First, we investigated whether the second or third barrier prevented the students from reconsidering their thinking; second, we tried to break the third

Table 1: Procedures, aims, and predictions.

Procedure		Aim	Prediction
Phase	Step involved		
Phase 0	Test	Identify the students' initial model	Students will select "left and counterclockwise" based on their incorrect model.
Phase 1	Thinking time Test	Confirm the existence of the first barrier	If the first barrier exists: → Students will conduct more animations with the round yoyo and select "left and counterclockwise."
Phase 2	Instruction Thinking time Test	Break the first barrier	If the first barrier is broken: → Students will conduct animations with various yoyo shapes.
Phase 3	Additional test Instruction Thinking time Test	Confirm the existence of the second and third barriers Break the third barrier	If the second barrier exists: → The expected combinations, "right and still," will not be selected on the additional test. If the third barrier exists: → Students will not arrive at reach the correct model even if they do not have the second barrier. If the third barrier is broken: → Students will select "right and clockwise" on the test.

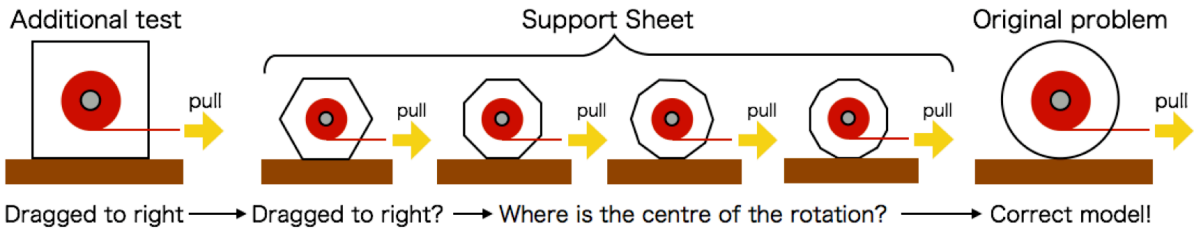


Figure 3: An image showing the mental process of solving the yoyo problem by using the additional test and the support sheet.

barrier. (Breaking the second barrier was outside the scope of our study, as the second barrier exists at Step 3, which TE-ext did not externalize.) Phase 3 included an additional test, instruction, thinking time, and the original test.

In the additional test, the students were asked to indicate the movement of the square-shaped yoyo (shown at the far left in Figure 3) just as they had done for the round yoyo on previous tests. We conducted this additional test to investigate the existence of the second barrier (i.e., the first purpose). If the second barrier prevented the students from using their knowledge appropriately, they would not select the expected choices (movement to the right, remaining still with regard to rotation).

To address the second purpose of Phase 3 (trying to break the third barrier), we targeted those students who selected incorrect combinations on the original test in Phase 2 and "right and still" on the additional test. They must have been prevented from shifting their model during Phase 2 by the third barrier. The additional test served as a cue for them, as suggested by Anzai and Yokoyama (1984).

We gave these students an additional "support sheet" including choices of movement for hexagonal, octagonal, 10-sided, and 12-sided yoyos. This bridged the gap between the square-shaped yoyo and the original round yoyo, as shown in Figure 3. We instructed the students to make predictions about the yoyos on the support sheet and then to solve the original yoyo problem based on their predictions for 4- to 12-sided yoyos. With this assistance, the students were given another 10 minutes of thinking time to use the TE-ext, the additional test, and the support sheet. Finally, they took the original test one more time. If we had successfully broken the third barrier by this point, they would imagine a correct model and select "right and clockwise."

Results

Model Shift

In Figure 4, we show the students' selections in each phase of the test. The correct answer is "right and clockwise." No student selected the correct combination, except for one student in Phase 0 who then changed his answer in Phase 1. In all phases, the selection was not equally distributed (Fisher's exact test $ps < .05$). The number of students who selected "left and counterclockwise" was significantly larger than the expected value in all phases ($ps < .002$). These results show that the shift to a correct model did not occur.

First Barrier

Figure 5 shows a ratio that indicates how frequently each shape was selected when the "move" button was clicked.

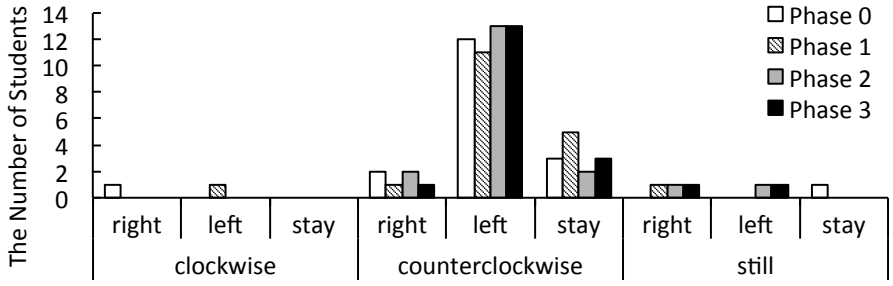


Figure 4: The student's selections on the test in each phase.

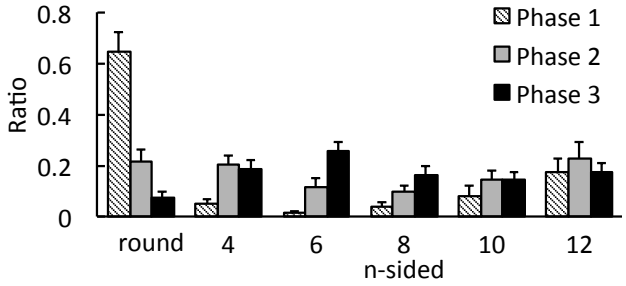


Figure 5: The ratio of the number of animations with each shape to all animations (bar represents standard error).

We conducted a 3 (phase) $\times$ 6 (shape) ANOVA. There was a significant interaction between the phase and shape factors ($F(10,180) = 13.84, p < .001$). Here we focused on the effect of the shape factor according to our purpose. In Phase 1, a simple main effect of the shape factor was significant ($F(5,90) = 26.412, p < .001$). The round yoyo was selected more frequently than other shapes were (ps $< .05$). In Phase 2, there was no significant effect of the shape factor ($F(5,90) = 1.376, p = .241$). In Phase 3, the simple main effect was significant again ($F(5,90) = 2.826, p = .020$). The only difference was that the hexagonal yoyo was selected more frequently than the round yoyo was (p = .050). A main effect of the shape factor was significant ($F(5,90) = 7.567, p < .001$). The preference for the round yoyo in Phase 1 means that the first barrier prevented the student from changing the situation from the original one. The instruction in Phase 2 broke this barrier, so that the preference disappeared. However, the various situations, including the square-shaped yoyo, did not lead the students to the correct model (Figure 4). This failure must be due to the second and/or third barriers.

Second Barrier

Figure 6 shows the selections made on the additional test and the support sheet. In the additional test, the selections were not equally distributed (Fisher's exact test p = .024). The number of students who selected the expected answer, "right and still," was significantly larger than the expected value (p = .002). This means that about half of the students who selected "right and still" did not have the second barrier. However, other students selected unexpected combinations. Although we cannot know why they did so, their incorrect model must have affected their selection. This result shows that half of the students could not arrive at the correct model because of the second barrier.

Third Barrier Half of the students did not have the second barrier. However, in the test in Phase 2, they were still unable to select the correct combination (Figure 4). This implies the existence of the third barrier.

In Phase 3, we attempted to break the third barrier using some supportive information. Despite these efforts, no student was able to select the correct combination on the original test (Figure 4). The reasons given for their decisions reveal that they added inappropriate or incorrect explanations to justify retaining their initial model. For example, a student who selected the expected combination on the additional test but stuck with "left and counterclockwise" on the original test wrote, "a round shape is easier to roll [to the left than other shapes] because of the small friction force." These results mean that our supportive assistance could not break the third barrier.

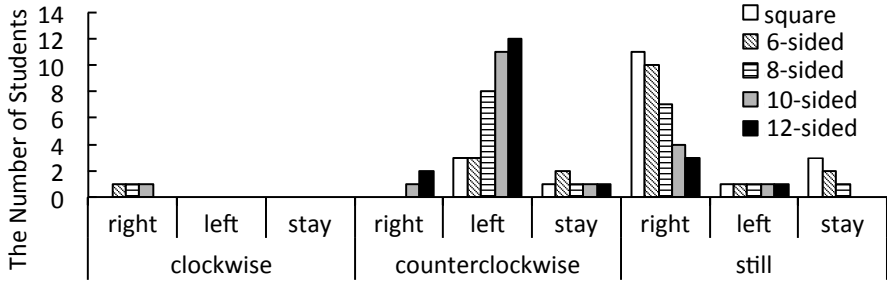


Figure 6: The number of students selecting each combination on the additional test and the support sheet.

General Discussion

In this study, we developed TE-ext, a system that permits observation of individuals' thought experiments, and used it to investigate the three barriers to an effective TE. We found that none of the students could make the shift to a correct model. All students experienced the first barrier, and half of them exhibited the second barrier. Even when students did not have the second barrier, the third barrier prevented them from shifting to a correct model. Although we successfully broke the first barrier by means of our instruction, the support sheet designed to break the third barrier did not have sufficient effects.

Almost all students continued to hold to their initial, incorrect model. They added inappropriate or incorrect peripheral explanations, as is typical of people dealing with information contrary to their beliefs (Chinn & Brewer, 1993; Dunbar, 1995). We need to break the third barrier to use TE-ext as an educational system.

One possible reason why our supports had no effect on the third barrier is that the students had incorrect models that differed from what we anticipated. Our students had different academic majors from the science and engineering students who participated in Anzai and Yokoyama (1984)'s study. The reasoning given for their selections on the test showed their belief that when the string was pulled, a force unwinding the wound string was present and the yoyo rolled in reaction to this force. Our supportive information in Phase 3 did not focus on this point, so it had little effect on this belief.

Regarding the potential educational use of TE-ext, we expected that the students would realize the benefits of TE by solving the yoyo problem successfully using TE-ext. Unfortunately, no one was able to shift to a correct model in our experiment. The following improvements, suggested by the results of our experiment, could make TE-ext a more powerful educational tool.

Improvement of TE-ext to help students visualize situations in daily life would be effective to break the second barrier. In such situations, they can draw on their own experience to see the result. Kösem and Özdemir (2014) showed that novices tended to use their own experiences as resources when conducting TEs.

Adaptive supports causing each student to conduct TEs suitable for his or her own incorrect model would help break the third barrier. In this study, our supports were not effective in modifying the students' initial model. If TE-ext could be improved by incorporating each student's cognitive model, it could adaptively provide TE situations suitable to instruct that student.

Additionally, TE-ext could work more effectively in a group setting. Students could share their TEs with each other by using TE-ext. TE-ext would serve as a hub connecting the students and their thought processes (Nersessian, 2009).

References

Anzai, Y., & Yokoyama, T. (1984). Internal models in physics problem solving. *Cognition and Instruction*, *1*, 397–450.

Brown, J. R. (1991). *The laboratory of the mind: Thought experiments in the natural sciences*. New York, NY: Routledge.

Brown, J. R. (2006). The promise and perils of thought experiments. *Interchange*, *37*, 63–75.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*, 1–49.

Clement, J. J. (2009). The role of imagistic simulation in scientific thought experiments. *Topics in Cognitive Science*, *1*, 686–710.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 365–395). Cambridge, MA: MIT Press.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*, 280–285.

Kösem, Ş. D., & Özdemir, O. F. (2014). The nature and role of thought experiments in solving conceptual physics problems. *Science & Education*, *23*, 865–895.

Lattery, M. J. (2001). Thought experiments in physics education: A simple and practical example. *Science & Education*, *10*, 485–492.

Matthews, M. R. (2014). *Science teaching, 20th anniversary revised and expanded edition*. New York, NY: Routledge.

Monaghan, J. M., & Clement, J. J. (1999). Use of a computer simulation to develop mental simulations for understanding relative motion concepts. *International Journal of Science Education*, *21*, 921–944.

Nersessian, N. J. (1992). In the theoretician's laboratory: Thought experimenting as mental modeling. In *Proceedings of the biennial meeting of the philosophy of science association* (pp. 291–301). Philosophy of Science Association.

Nersessian, N. J. (2002). The cognitive basis of model-based reasoning in science. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 133–153). Cambridge: Cambridge University Press.

Nersessian, N. J. (2009). How do engineering scientists think? model-based simulation in biomedical engineering research laboratories. *Topics in Cognitive Science*, *1*, 730–757.

Reiner, M. (2006). The context of thought experiments in physics learning. *Interchange*, *37*, 97–113.

Stephens, A. L., & Clement, J. J. (2010). Documenting the use of expert scientific reasoning processes by high school physics students. *Physical Review Special Topics-Physics Education Research*, *6*, 020122.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*, 535–585.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

Mind reading: Discovering individual preferences from eye movements using switching hidden Markov models

Tim Chuk (u3002534@connect.hku.hk)
Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong
Antoni B. Chan (abchan@cityu.edu.hk)
Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong
Shinsuke (Shin) Shimojo (sshimojo@caltech.edu)
Division of Biology and Biological Engineering, 1200 East California Boulevard, Pasadena, CA 91125.
Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract

Here we used a hidden Markov model (HMM) based approach to infer individual choices from eye movements in preference decision-making. We assumed that during a decision making process, participants may switch between exploration and decision-making periods, and this behavior can be better captured with a Switching HMM (SHMM). Through clustering individual eye movement patterns described in SHMMs, we automatically discovered two groups of participants with different decision making behavior. One group showed a strong and early bias to look more often at the to-be chosen stimulus (i.e., the gaze cascade effect; Shimojo et al., 2003) with a short final decision-making period. The other group showed a weaker cascade effect with a longer final decision-making period. The SHMMs also showed capable of inferring participants' preference choice on each trial with high accuracy. Thus, our SHMM approach made it possible to reveal individual differences in decision making and discover individual preferences from eye movement data.

Keywords: hidden Markov model; gaze preference; eye movement; face recognition.

Introduction

Gaze plays a crucial role in our social lives because it helps to signify the target of one's attention and interest in a complex environment. For example, studies have shown that gaze is an important cue for infants to infer what adults mean and therefore a precursor to their language development (Brunet, 1985). Gaze also has been shown to be an indicator of people's preferences. The phenomenon of preferential looking (Franz, 1964) suggested that the liked stimuli are usually looked at for longer time. Some researchers argue that gaze not only reflects individual attentions and preferences, it also helps to shape them. Shimojo et al. (2003) conducted a two-alternative-forced-choice (2AFC) preference task, in which participants were required to look at two face images and then to decide which one they liked more. The two face images were shown on the left and the right side of the screen. Their results showed that the participants spent significantly more time on the side that they were about to choose, starting from about 800ms before they made and indicated their decisions. They coined the term "gaze cascade effect" to distinguish it from the mere exposure effect, which suggested that people tend to prefer things that they are familiar with (Zajonc, 1968). Shimojo et al., (2003) argued that gaze shifts are essential to shaping preferences. In other words, it is not exposure to the stimuli alone that shaped the mere exposure effect; gaze shifts have to be involved (Shimojo et al., 2011).

In order to experimentally demonstrate this argument, in one of their follow-up studies (Simion & Shimojo, 2006), they adopted the same 2AFC settings but forced the participants to look at a fixation point located at the center of the screen. The participants' eye movements were therefore constrained. The two face images were sequentially superimposed in the foveal region for different time lengths. They found no bias that favored the longer-exposed images over the shorter ones, which argued for the role of gaze shifts in actively shaping one's preferences.

Although these studies discovered the role of gaze in reflecting and shaping preferences, the finding was based on group-level analysis and thus was not able to address individual differences in preference decision-making behavior. Some studies have shown that participants' eye movements in preference decision-making tasks may to some extent reflect their traits. For instance, it was found that when faced with happy faces and angry faces, old participants paid more attention toward the happy faces, while young participants paid more attention toward the angry faces (Isaacowirz, 2006). It was also found that when participants were shown unpleasant images, the optimistic participants paid significantly less attention toward the images than their pessimistic counterparts.

Studies have shown that people have substantial and persistent differences in their eye movements in cognitive tasks. For example, Castelhano and Henderson (2008) found that fixation durations and saccade amplitudes were highly consistent within-individuals. Participants showed similar fixation patterns when viewing different types of image. Peterson and Eckstein (2013) found that when looking at human faces, people had different preferred fixation locations and that these individual differences persisted over time and tasks.

In our previous studies (Chuk et al., 2014; Chan et al., 2015), we used hidden Markov models (HMMs) to address individual differences in eye movements. We conducted face recognition studies and found that individual differ-

ences in eye movements were useful indicators of participants' performance. We used an HMM to summarize the fixation locations and the scan-paths of an individual. We then clustered the individual HMMs into groups according to their similarities and found that while some people preferred to look at specific facial features (i.e. the analytic eye movement pattern), others preferred to look at the center of the faces (i.e. the holistic eye movement pattern). We found that people who showed analytic eye movement patterns performed significantly better than those who showed holistic eye movement patterns (Chuk et al., 2014). We also found that old people were significantly more likely to use holistic eye movement patterns than young people, and the more holistic their patterns, the lower their cognitive abilities (Chan et al., 2015). These findings were not possible with other existing methods that do not take individual differences into account.

The above findings suggested that one's eye movement pattern could be used to infer one's recognition performance, which also implied the possibility of using one's eye movement pattern to infer one's preference in decision-making tasks. Our HMM approach is more capable of capturing individual differences because it takes into account the spatial (i.e., fixation locations) and temporal (i.e., scan paths) information of one's eye movement simultaneously, when many of the alternatives only focus on the spatial information.

In this study, we try to read participants' mind through their eye movements. We aim to infer participants' preferences in a preference decision-making task by modeling their eye movements using HMMs. However, the standard HMMs may not be suitable for the purpose of this task. The gaze cascade effect showed that in these tasks, participants might go through at least two mental periods during a trial: exploration and decision-making. During the exploration period, they may switch their gaze between the two sides equally; during the decision-making period, they may look more often at the images to-be-chosen (Shimojo et al., 2003). This finding suggests that participants' patterns of switching their gazes between the two images could be different during the two periods. The standard HMM with one set of hidden states and a transition matrix is unable to capture this. Therefore, instead of using a standard HMM, we create a 'switch' for two HMMs, such that a high-level HMM consists of two low-level HMMs, each representing gaze patterns of a particular mental period (Figure 1).

Here we train the SHMMs on individual fixation data and cluster the models. The analysis separates participants into groups based on their eye movement differences. This helps to reveal common eye movement patterns when they make decisions, and in turn allows us to examine whether different patterns are associated with different decision-making behavior. We then use the models to infer their preferences on each trial and examine the models' accuracy.

Method

Materials and Procedure

We performed our analysis using the data collected in a preference decision-making study (Shimojo et al., 2011). A total of 12 participants were recruited for the 2AFC task. There were in total 60 trials. On each trial, two face images, one on the left and one on the right, were shown on the screen for the participants to make their choices. There was no time limit. Participants were allowed to move their eyes to compare the two images. They were told to press a button to indicate which image (left or right) they preferred once they had made their decisions. Eye movements were recorded using an Eyelink 2 eye tracker.

Switching hidden Markov model

A standard hidden Markov model (HMM) contains a vector of prior values, a transition matrix, and a Gaussian emission for each hidden state. The prior values indicate the probabilities of the time-series data to begin with the specific hidden states; the transition matrix indicates the transition probabilities between any two hidden states; the Gaussian emissions indicate the probabilistic associations between the time-series data and the hidden states. In our current context, the hidden states correspond to the regions of interest (ROIs), which were learned from the fixation locations, and the emissions are the fixation locations.

In contrast to a standard HMM, a switching HMM (SHMM) contains two levels of HMMs; the high-level HMM indicates the transitions between the low-level HMMs (Figure 1). In our implementation, the high-level hidden states represent the current gaze strategy (i.e., exploration or decision-making), whereas the low-level hidden states correspond to ROIs over the stimuli. Each high-level state has its own low-level prior values and transition matrix. The low-level states (ROIs) are shared among the high level states (gaze strategies). The high-level HMM has its own transition matrix, which governs the switching between gaze strategies. The high-level state sequences and the low-level state sequences are both 1st-order Markov chains.

In practice, the SHMM can be turned into a standard HMM by combining the high-level and the low-level hidden state variables into a single hidden state variable, whose values are the Cartesian product of the low- and high-level state values. In the current case, since the low-level states are shared among the high level states, the number of low-level states (K) is the same for each high-level state. Hence, the equivalent HMM has S*K hidden states, where S is the number of high-level hidden states; the transition matrix has block structure. Because the Gaussian emissions are shared among the high-level hidden states, they are independent from the high-level switches and are only attached to the low-level hidden states (Figure 1).

We performed the EM algorithm to estimate the SHMM parameters. In the E-step, the responsibilities were calculated using the standard forward-backward algorithm with the block transition matrix, initial state vector, and emission densities. In the M-step, the prior and pairwise responsibilities were summed over the high-level and the low-level states respectively to yield the parameter updates for both the high-level states and the low-level states.

For instance, the prior responsibilities were summed over the low-level hidden states for each of the high-level state in order to yield the parameter updates for the low-level states, and then they were summed over the high-level states to yield the parameter updates for the high-level states. Similarly, the pairwise responsibilities were summed over the low-level hidden states for each high-level state to yield the transition matrix updates for each transition matrix, and then were summed over the high-level hidden states to yield the updates for the switching matrix.

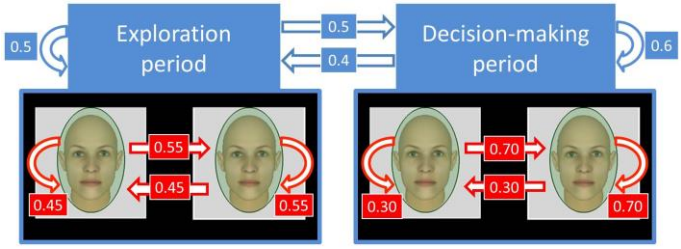


Figure 1: An illustration of the SHMM used in the current study. The high-level HMM states consisted of two gaze strategies: exploration and decision-making. The blue arrows indicate the transitions between them, and the numbers indicate transition probabilities. Eye movements within each period were modeled with a low-level HMM. The red arrows represent transitions between ROIs. The two periods have the same ROIs but different transition probabilities.

Training and clustering individual SHMMs

For each participant, we trained an SHMM using the fixation locations collected from all trials. In order to simplify the analysis, we used only two Gaussian emissions per model; one on each side. This helped to focus the analyses on the transition information between the two face stimuli (Figure 1). The advantage of using Gaussian rather than discrete emissions is that it can be easily extended to analyses that explore the ROIs on each face. We used two high-level hidden states to reflect that the participants may switch between exploration and decision-making periods during a trial.

For SHMM estimation, we initialized the transition matrices based on the actual data. The exploration period usually happened at the beginning of a trial and the decision-making period usually happened at the end of a trial. Therefore, we initialized the exploration and the decision-making transition matrices based on the first 5 and the final 5 fixations of each trial.

Since participants' exploration and decision-making may influence each other during a trial (e.g., Shimojo et al., 2003), here we assumed that participants may switch back to exploration after decision making before they made the final decision. To reflect this, the high-level transitions between the exploration and the decision-making periods were set to [0.5, 0.5; 0.4, 0.6] (Figure 1). That is, once the participant had switched from the exploration period to the decision-making period, the chance of switching back to the exploration period (0.4) was slightly lower than the chance of him or her staying in the decision-making period (0.6).

To reveal a common eye movement pattern shared by all individuals during the exploration period, we created an HMM using the exploration transition matrix and Gaussian emissions for each individual, and clustered these HMMs into a group using the VHEM algorithm (Coviello et al., 2014). To reveal whether participants had different fixation patterns when they were making decisions, we clustered their HMMs for the decision-making period into two groups and examined their differences.

To infer when the participants were in the exploration or decision-making period, we used the Viterbi algorithm to find the most likely hidden state sequence for each trial. In this sequence, a change in the value of the high-level hidden state indicated a switching point. Next, we examined whether SHMMs can be used to infer individual preferences on each trial. We used only the fixations in the last decision-making period to perform our inferences. That is, the inference was performed using only the fixations after the final switch from exploration to decision-making. We split the trials into two sets: one for all the trials on which the left side images were chosen, and one for all the trials on which the right side images were chosen. For each set, we used all but one trial to train an SHMM and used the held-out trial for testing. If, for example, on the test trial, the participant selected the image on the right, the SHMM trained on the trials in which the right image was chosen should fit the trial better than the SHMM trained on the left-selected trials. To test this, we compared the log-likelihoods generated by the two SHMMs (Chuk et al., 2014). We hypothesized that the SHMM for the right-selected trials should produce a higher log-likelihood than the SHMM for the left-selected trials. This was repeated over all the trials and all the participants. The accuracy of our inferences was evaluated using one-sample t-tests. We hypothesized that we can infer participants' decisions at a level that is significantly above chance.

Results

Categorization of individual SHMMs

First, we clustered the 12 participants' exploration HMMs into one representative HMM using the VHEM algorithm. The output showed that participants did not look more often at the side that they were going to choose during the exploration period. This is evident from the fact that both the not-chosen images and the chosen images had a prior value of roughly 0.5. The transition probabilities were also around the chance level for both sides, indicating that at the exploration stage of the trials, participants paid roughly equal attention to the two images. Table 1 below shows the transitional information of the representation HMM.

For the decision-making period, we clustered their decision-making HMMs into two groups. The two groups had an equal number of participants. Each participant had a probability of being associated with each group. The probabilities showed a clear-cut division between the two groups,

indicating that the two groups were clearly separated, t (11) = 3.32, p = .007. Tables 2 and 3 below show the representative transition matrices of the two groups generated by the VHEM algorithm. Participants in group A showed apparent fixation bias toward the side that they were about to choose, whereas those in group B showed only mild bias. Although participants in group B also paid more attention to the images that they were about to choose, the transition matrices suggested that their fixation pattern were not that different from the exploration period.

Table 1: The exploration period for all the participants

	to not-chosen	to chosen
Priors	.50	.50
from not-chosen	.55	.45
from chosen	.53	.47

Table 2: The decision-making period for group A.

	to not-chosen	to chosen
from not-chosen	.32	.69
from chosen	.32	.69

Table 3: The decision-making period for group B.

	to not-chosen	to chosen
from not-chosen	.44	.56
from chosen	.43	.57

Cascade plots

In order to visually compare the two groups in terms of the magnitude of their gaze cascade effects, we generated the gaze cascade plots as that seen in Shimojo et al. (2003). The plots show the proportion of time that the participants spent on looking at the chosen images, and span 2.5 seconds (see Figure 2). It can be seen from the plot generated from all participants that as it got closer to the end, participants spent more time on inspecting the side that they were about to choose. The proportion of time spent on the chosen sides went from about chance level (0.5) steadily up until it reached almost 0.9. Chi-square test showed that the trend began at around 900 ms before the end of the trials, $\chi^2(1) = 3.92$, $p = .05$. The plots of the two groups show some more interesting differences. It can be found that although the general trend was shared between the two groups, participants in group A showed a more stable gaze cascade effect. Chi-square test showed that the cascade effect occurred at around 1100 ms before the end of the trials, $\chi^2(1) = 4.74$, $p = .03$, which was earlier than the general trend. The proportion of time on the chosen image reached about 0.95 by the end, which was also higher than the general trend. Group B, however, showed a decrease in the proportion of time spent on the chosen images between 1.5 seconds and 1 second.

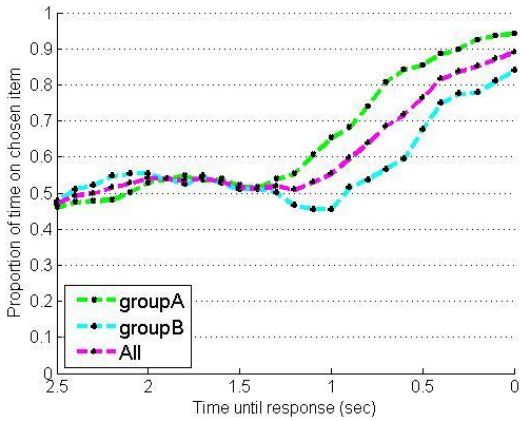


Figure 2: The cascade plots.

Chi-square test showed that the cascade effect occurred at around 600 ms before the end of the trials, $\chi^2(1) = 6.51$, $p = .01$, The proportion of time on the chosen image then accumulated continuously to about 0.85, which was lower than the general trend. Although both groups showed the gaze cascade effect, neither their magnitudes nor their onset times were the same. However, the observation that the two groups did not show qualitative differences in eye movement pattern suggests that there may be a continuum between the two groups.

Inference of individual preferences

We performed the inference of individual preferences using the fixations after the last time that the participant switched from the exploration to the decision-making period. The inference was done on every trial for every participant. The accuracy of the inferences is shown in Table 4 below.

Table 4: Inference of participants' preference.

participant no.	group	accuracies
01	A	.98
03	A	.94
07	A	.97
08	A	1
10	A	.98
11	A	.96
02	B	.78
04	B	.85
05	B	.93
06	B	.58
09	B	.94
12	B	.52

It can be seen that for participants in group A, the inference accuracies were both high and consistent ($M = 0.97$). One sample t-test showed that the inference accuracies were significantly above the chance level, $t(5) = 56.6$, $p < .001$. For group B ($M = 0.77$), while for some of the participants the inference accuracy was also high, there was a larger variance within the group. For instance, for participant 6 and participant 12, the models showed only around chance level

accuracies, but for participant 5 and participant 9, the models showed a high level of accuracies comparable to that of group A. One sample t-test showed that the inference accuracies of group B were also significantly higher than chance, $t(5) = 3.66$, $p = .01$. However, Group A's inference accuracies were significantly higher than that of group B, $t(10) = 2.79$, $p = .02$.

In addition, we examined the time length of the final decision-making period for each participant. Table 5 below shows the average time length of the decision-making period and the average number of fixations within the period.

Table 5: Average time length of the final decision-making period

participant	group	time length (sec)	number of fixations
01	A	1.32	4.76
03	A	1.39	4.89
07	A	1.78	5.88
08	A	1.89	5.5
10	A	6.09	17.22
11	A	2.79	7.17
02	B	2.93	9.14
04	B	3.87	9.75
05	B	3.71	12.72
06	B	4.82	16.95
09	B	3.92	12.52
12	B	5.26	13.2

It can be seen that on average, the participants switched to the final decision-making period at about 9.97 fixations prior to the end of the trials. The average duration of the period was 3 seconds. This was earlier than what the cascade plots showed, which suggested that our SHMM analysis was able to detect participants' preferences at an earlier stage. The final decision-making period for group B (4.09 secs) was longer than that for group A (2.54 secs). The difference between the two groups was marginal, $t(10) = 2.12$, $p = .06$. Note that the clustering of the two groups was completely based on the eye movement data, and thus the group difference in inference accuracy and in length of the final decision period emerged naturally as the results of the clustering.

Discussion

In this study, we revealed individual differences in decision-making behavior from their eye movements and also successfully inferred participants' preferences using their eye movements through the HMM-based approach. More specifically, by assuming a decision process involves an exploration and a decision-making period, and using SHMMs to model these two periods, we discovered two participant groups with different decision-making behavior automatically from the eye movement data. Although all participants showed a tendency of looking more often at the to-be-chosen images by the end of the trials, our SHMMs showed that this tendency was obvious only among 6 of the 12 participants. As shown in Table 2, for group A, the transition matrix showed a strong bias to look at the chosen images. The cascade plot showed that the cascade effect happened early, roughly 1.5 seconds before the end of the trials (Figure 2). They also had a stronger gaze cascade effect: and onset time was about 600 ms before the end of the trials, and the proportion of time spent on the chosen images reached almost 95%. For group B (Table 3), although they also showed the gaze cascade effect, it was both later and less obvious. The cascade plot showed that its onset time was about 1100 ms before the end of the trials, and the proportion of time on the chosen images was about 85%.

The fact that all the 12 participants showed the gaze cascade effect suggests its robustness. However, our SHMM analysis revealed that the 12 participants could be separated into two groups with substantial differences in their fixation patterns for preference decision making. This demonstrated the advantage of our approach: by summarizing participants' fixation patterns using the SHMM, those who had similar fixation patterns could be clustered together. The grouping was done automatically based on eye movement data alone, and participants' differences in decision making behavior emerged naturally as the result of the clustering.

In addition, the SHMM approach allowed us to infer individual preferences with very high accuracies. This was made possible because the structure of the SHMM allowed us to capture the eye movements during the final decision-making period, which were informative for such inferences. Both groups' inference accuracies were significantly above the chance level. The inference accuracy for group A was significantly higher than group B, suggesting that eye movement was a more accurate indicator of preference for participants in group A than those in group B. There was also a much larger individual difference among participants in inference accuracy in group B. It is worth noting that the two groups were discovered from the eye movement data without considering the participants' other behavioral differences. Therefore, although for some participants (e.g., participant 5 and 9), the inference accuracies were high, they were nevertheless clustered into group B because of their eye movement patterns. Having that said, the fact that some of these participants in group B also showed a high level of accuracies suggest that this group could be further divided into sub-groups, so that a more fine-grained categorization should be possible. This could be done in future studies.

The analysis of the time length of the final decision-making period showed a different pattern from the cascade plots (see Tables 2-3 and Figure 2). Our analysis results suggested that the final switch from the exploration period to the decision-making period happened at a much earlier time than that suggested by the cascade plots. The average number of fixations within the period, for all the participants, was 9.97 fixations, which translated to about 3 seconds of time. In contrast, in the cascade plots, the cascade effect emerged at around 900 ms before the trial end. The average time length of the final decision-making period in group B (4.09 secs) was longer than that in group A (2.54

secs; see Table 5). This is possibly because the transition matrix of group B showed a milder tendency towards the to-be-chosen side than group A, and hence it took longer for the gaze cascade effect to become noticeable. For group A, the transition matrix showed a more obvious tendency, so that the cascade effect appeared earlier even though the participants started the decision-making period later than group B.

Note that the cascade plots did not show a turning point which shifted a flat line to a steep slope. The accumulation of the proportion of time on the chosen image was gradual and slow. The accumulation only became apparent by the end of the trials, which therefore suggested that it would be hard to use the cascade plots to determine when the decision-making period had begun. Our analysis showed the advantage of being able to distinguish and isolate the fixations in the final decision-making period, which allowed us to infer participants' preferences with high accuracy using only their eye movements. Note however that the model assumes a separation of the two periods with the possibility to switch between the two periods during the decision-making process. It remains possible that there is an additional period between these two periods, in which exploration and decision making are mixed (e.g., Shimojo et al., 2013). This additional period could be represented by adding an additional high-level state to the SHMM. Future work will examine this possibility.

The finding that some people have a longer final decision period might imply some traits of these individuals. Previous studies demonstrated relationship between personality and response time when making decisions. For instance, extraverts were found to respond slower than introverts (Doucet & Stelmack, 2000). Reflective people have longer response time than impulsive people (Sternberg & Grigorenko, 1997). These findings point to the possibility that the two groups we found might also differ in terms of their personalities. In future studies, this should be explored.

In summary, here we used SHMMs to analyze eye movement data collected from a preference decision-making task. The SHMM assumes an exploration period and a decision-making period during the decision making process. By clustering individual models into groups in a data-driven fashion, we discovered two participant groups with different decision-making behavior: one group had a stronger and longer-lasting gaze cascade effect and a shorter decision-making period than the other group. These group differences emerged naturally as the results of the clustering based on eye movement data alone, demonstrating the power of the HMM approach of eye movement data analysis for the understanding of individual cognitive behavior. We also found that participants' preferences on each trial can be inferred from their fixation patterns with high accuracy. This result thus provides a strong evidence for the possibility of mind reading from eye movement behavior.

Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project 17402814 to J. H. Hsiao and CityU 110513 to A. B. Chan) and to JST.CREST (SS).

References

Bruner, J. (1985). Child's talk: Learning to use language. *Child Language Teaching and Therapy*, *1*(1), 111-114.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*,*62*(1), 1.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of vision*, *14*(11), 8.

Chan C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2015). Eye movement pattern in face recognition is associated with cognitive decline in the elderly. *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 321-326).

Coviello, E., Chan, A. B., & Lanckriet, G. R. (2014). Clustering hidden Markov models with variational HEM. *The Journal of Machine Learning Research*, *15*(1), 697-747.

Doucet, C., & Stelmack, R. M. (2000). An event-related potential analysis of extraversion and individual differences in cognitive processing speed and response execution. *Journal of personality and social psychology*, *78*(5), 956.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668-670.

Isaacowitz, D. M. (2006). Motivated gaze the view from the gazer. *Current Directions in Psychological Science*, *15*(2), 68-72.

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological science*, *24*(7), 1216-1225.

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature neuroscience*, *6*(12), 1317-1322.

Shimojo, S., Simion, C., & Changizi, M. A. (2011). Gaze and preference-orienting behavior as a somatic precursor of preference decision. In R, B, Adams (Ed.), *The science of social vision* (pp. 151-163). Oxford: Oxford University Press.

Simion, C., & Shimojo, S. (2006). Early interactions between orienting, visual sampling and decision making in facial preference. *Vision research*, *46*(20), 3331-3335.

Sternberg, R. J., & Grigorenko, E. L. (1997). Are cognitive styles still in style?. *American psychologist*, *52*(7), 700.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, *9*(2p2), 1.

Semantic Contamination of Visual Similarity Judgments

Layla Unger (LUnger@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15214 USA

Anna V. Fisher (Fisher49@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15214 USA

Abstract

The roles of semantic and perceptual information in cognition are of widespread interest to many researchers. However, disentangling their contributions is complicated by their overlap in real-world categories. For instance, attempts to calibrate visual similarity based on participant judgments are undermined by the possibility that semantic knowledge contaminates these judgments. This study investigated whether inverting stimuli attenuates semantic contamination of visual similarity judgments in adults and children. Participants viewed upright and inverted triads of familiar animals, and judged which of two test items looked most like the target. One test item belonged to the same category as the target, and one belonged to a different category. Test items' visual similarity to the target either corresponded or conflicted with category membership. Across age groups, conflicting category membership reduced accuracy and slowed reaction times to a greater extent in upright than inverted triads. Therefore, inversion attenuates semantic contamination of visual similarity judgments.

Keywords: semantic knowledge; visual similarity

Introduction

We perceive different things as related to each other in a variety of potentially overlapping ways. For instance, entities may be perceptually similar due to their shared perceptual features, or belong to the same semantic, "taxonomic" category. Many research endeavors and theoretical debates have focused on questions surrounding the influence of these relations on various facets of cognition, such categorization, inductive inference, memory encoding, and visual search, as well as the development and neural underpinnings of these processes (e.g., Deák & Bauer, 1996; Gelman & Markman, 1986; Konkle, Brady, Alvarez, & Oliva, 2010). However, perceptual and semantic relations commonly overlap amongst real-world entities. Moreover, these relations may interact during both learning and online processing of perceptual input. This interplay between semantic and perceptual relations severely complicates the study of questions about their respective contribution to cognition.

The complication of interest in the present paper is the fact that attempts to control for and manipulate perceptual similarity of real-world items independent of the semantic relations between them may be undermined by the contamination of similarity judgments by semantic category knowledge. Below we discuss this issue, and then present a study designed to provide a possible solution to this problem.

Measuring Perceptual Similarity

Researchers have used multiple approaches to measure and calibrate the perceptual similarity between stimuli, including using their own intuition (e.g., Fisher, 2011), collecting similarity judgments from adults (e.g., Deák & Bauer, 1996; Gelman & Markman, 1986), and, in developmental studies, collecting similarity judgments from children (e.g., Long, Lu, Zhang, Li, & Deák, 2012; Sloutsky & Fisher, 2004). The aim of these approaches is to assess visual similarity independent of semantic relatedness. For example, to calibrate stimuli in a match-to-sample task with triads consisting of a target, a perceptual match, and a semantic match, researchers may ask a separate sample of participants to judge the visual similarity of the target to each match item on a likert scale (e.g., Deák & Bauer, 1996; Gelman & Markman, 1986). Alternatively, researchers may calibrate such triads by asking participants to choose which match item looks most like the target in order to obtain ratios of the similarity of the perceptual match to the target versus the semantic category match to the target (e.g., Long et al., 2012; Sloutsky & Fisher, 2004).

The calibration of visual similarity based on participant similarity judgments is common to studies in many areas, such as memory, semantic knowledge, and semantic development (e.g., Blaye, Bernard-Peyron, Paour, & Bonthoux, 2006; Deák & Bauer, 1996; Gelman & Markman, 1986; Konkle et al., 2010). Intrinsic to this approach is the assumption that people can judge visual similarity without being influenced by the semantic knowledge.

However, this assumption may be unwarranted. Semantic knowledge may instead influence perceptual similarity judgments through any of multiple routes. Knowledge of semantic relationships between items may influence judgments of their similarity: 1) After perceptual similarity has been independently evaluated (Pylyshyn, 1999), 2) By feeding back into perceptual similarity evaluations (Lupyan, Thompson-Schill, & Swingley, 2010), or 3) By influencing the similarity of items' perceptual representations during prior learning (Goldstone, 1998; Goldstone, Lippa, & Shiffrin, 2001; O'Reilly, Wyatte, Herd, Mingus, & Jilk, 2013).

The degree to which any of these routes truly characterizes cognition is the subject of active research and debate (Chen & Proctor, 2012; Lupyan et al., 2010). Although an in-depth evaluation of this issue is beyond the scope of this paper, the brief overview presented here highlights the many ways in which semantic knowledge may contaminate perceptual

similarity judgments. The purpose of the present study is therefore to assess the contamination of perceptual similarity judgments by semantic category knowledge, and test whether this contamination is attenuated by inversion.

The Present Study

The choice to test whether inversion attenuates the influence of semantic category knowledge on perceptual similarity judgments was motivated by numerous findings that rotation away from a canonical orientation impedes the identification of the category to which a familiar item belongs (e.g., Jolicoeur & Milliken, 1989; Lawson & Jolicoeur, 2003). Specifically, increasing misorientation both slows and increases errors for identifying an item's category label. This effect has been attributed to a process in which a perceived item must be mentally normalized to its canonical orientation before its category membership can be retrieved from memory (Lawson & Jolicoeur, 2003). Consequently, inversion may interfere with access to semantic category knowledge, and therefore attenuate the influence of such knowledge on perceptual similarity judgments.

In the present study, participants performed a match-to-sample perceptual similarity judgment task in which they were asked to choose which of two test items "looks most like" a target item for triads of items that were presented in both upright and inverted orientations on different trials. All items were pictures of familiar animals, such as "dog" and "pig". One test item belonged to the same category as the target (e.g., both were pigs), whereas the other test item belonged to a different category (e.g., dog). The visual similarity of the test items to the target was manipulated such that it either corresponded or conflicted with the category membership of the target (see Fig. 1).

We predicted that if semantic category knowledge influences perceptual similarity judgements, participants should choose the visual similarity match less accurately and more slowly when category membership and visual similarity were in conflict. Moreover, if inversion attenuates the influence of category knowledge on perceptual similarity judgments, less accurate and slower responses on conflict versus no-conflict trials should manifest to a greater extent when triads are upright than when they are inverted. To test whether this predicted pattern manifests across ages to whom the categories are familiar, we conducted this study with both a kindergarten-age and an adult sample.

Method

Participants

The total sample of 42 participants included 24 participants in each of two age groups: Kindergarten (M_{age} = 5.45 years, *SD* =0.43 years), and Adults. Kindergarten participants were recruited from schools in a middle-class, metropolitan area in a Northeastern US city, and Adults were recruited via Amazon Mechanical Turk. Adults were compensated at a rate of $5/hour for their participation.

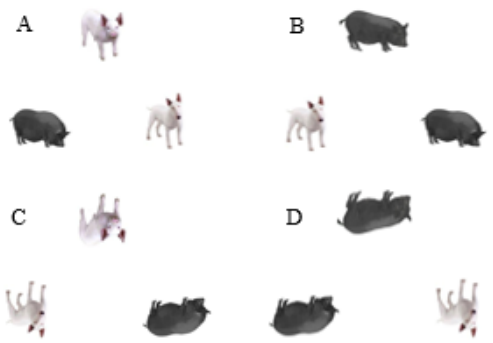


Figure 1. Examples of triads in each condition. A) Conflict, Upright; B) No Conflict, Upright; C) Conflict, Inverted; D) No Conflict, Inverted.

Materials

Both kindergarten and adult participants completed a Visual Similarity Judgment task, described below. Kindergarten participants viewed this presentation on a laptop computer, and made responses using a Cedrus RB-530 response box. To help kindergarten participants distinguish between the buttons they were instructed to use in the study (see Procedure), these buttons were given different-colored plastic covers. Adult participants viewed the presentation on Qualtrics, an online survey platform, and made responses using their personal keyboards. The Qualtrics version of the presentation was designed to record keyboard response times.

Visual Similarity Judgment Task This task consisted of triads of animal pictures presented on a computer screen that consisted of a Target item (e.g., a white pig) on the top, a Same Category test item (e.g., a black pig) on the bottom to one side, and a Different Category test item (e.g., a white dog) on the bottom to the other side. The pictures were photo-realistic images manipulated in graphics editing software to create two Semantic Conflict conditions: 1) No Conflict, in which the Same Category test item was visually similar to the target and the Different Category test item was visually dissimilar to the target, and 2) Conflict, in which the correspondence between category membership and visual similarity was reversed (see Fig. 1). Visual similarity was manipulated across several characteristics of the stimuli, including color, shape, and internal features.

Both No Conflict and Conflict triads were presented in two orientation conditions: Upright, and Inverted. The position of the two types of test items on the left or the right of the bottom of the screen was counterbalanced. Both, the Conflict condition and the Orientation condition were manipulated within participants.

The presentation included four practice trials, and 32 experiment trials. The practice trials consisted of a Conflict and a No Conflict triad presented in upright and inverted orientations, and the experiment trials consisted of eight Conflict and eight No Conflict triads presented in upright and inverted orientations. Each triad was presented once in

upright and inverted orientations. The order of the experiment trials was pseudo-randomized such that different versions of a given triad did not appear consecutively, and such that no more than two triads in the same combination of visual similarity and orientation conditions appeared consecutively.

Procedure

Children Participants were tested individually in a quiet space. To begin, participants were seated in front of the laptop and button response box, and the first practice trial was displayed. Participants were told that they were going to play a game in which they decide which of two animals on the bottom of the screen looks like the animal on the top. They were further asked to use the buttons on the response box to indicate which animal they chose. Participants were then allowed to proceed through the practice and experiment trials at their own pace. The button instructions were repeated on subsequent trials if participants either failed to make a response for several seconds, or started to press the buttons quickly and randomly.

Adults Participants completed the task via Qualtrics. The version adults completed was identical to the version children completed, with the exception that participants were instructed to use "z" and "m" keyboard keys rather than the left and right buttons of a response box.

Results and Discussion

First, to ensure that participants understood that the purpose of the task was to identify the similarity rather than the category match, the accuracy with which participants in each age group in each condition chose the similarity match was compared to chance (i.e., .5). All contrasts revealed significantly above chance performance (ps < .0001).

To test the prediction that conflict between semantic and visual similarity would decrease accuracy and slow response times (RTs) for Conflict versus No Conflict trials in the Upright and not the Inverted condition, we analyzed the effects of the Semantic Conflict and Orientation factors on accuracy and RT using repeated measures ANOVAs. Specifically, for each outcome measure, we calculated each participant's mean score for the four combinations of conditions produced by our Semantic Conflict and Orientation factors, and submitted these mean scores to separate repeated measures ANOVAs for each age group.

Accuracy Analysis

For adult participants, this analysis revealed a main effect of Semantic Conflict ($F(1,23)=10.122$, $p=.004$, $\eta^2=.306$), and a main effect of Orientation ($F(1,23)=6.457$, $p=.018$, $\eta^2=.219$). More importantly, both main effects were qualified by a significant interaction ($F(1,23)=15.826$, $p=.001$, $\eta^2=.408$). To explore this interaction, we conducted t-tests comparing Upright versus Inverted RTs separately for the Conflict and No Conflict conditions. In the Conflict condition, adults were more accurate on Inverted ($M_{accuracy}$=89.96%) than on Upright ($M_{accuracy}$=82.29%) trials ($t(23)=3.680$, $p=.001$, Cohen's $d=1.54$), whereas in the No Conflict condition, adult accuracy did not significantly differ on Inverted and Upright trials ($t(23)=1$, $p=.328$) (see Fig. 2).

Kindergarten participants exhibited similar patterns of accuracy. The repeated measures ANOVA for this age group also revealed a main effect of Semantic Conflict ($F(1,23)=6.4$, $p=.019$, $\eta^2=.218$), though unlike adults, the main effect of Orientation did not reach significance ($F(1,23)=1.15$, $p=.295$). Moreover, like adults, the analysis with this age group revealed a significant interaction between Semantic Conflict and Orientation ($F(1,23)=5.522$, $p=.028$, $\eta^2=.194$). T-tests comparing Upright and Inverted trials for each Semantic Conflict condition revealed that Kindergarten participants were marginally more accurate on Inverted ($M_{accuracy}$=84.90%) than on Upright ($M_{accuracy}$=77.08%) trials in the Conflict condition ($t(23)=2.01$, $p=.057$, Cohen's $d=.45$), and numerically more accurate on Upright than Inverted trials in the No Conflict condition ($t(23)=1.772$, $p=.09$) (see Fig. 2).

RT Analysis

Prior to calculating each participant's mean RT score, we filtered RTs to remove inaccurate trials and trials on which the participant responded either faster than 250 msec, or more than three standard deviations more slowly than the average RT for their age group.

For adults, this analysis revealed a main effect of Semantic Conflict ($F(1,23)=15.133$,$p=.001$, $\eta^2=.397$), and no main effect of Orientation ($F(1,23)=1.116$, $p=.302$). Critically, the main effect of Semantic Conflict was qualified by a significant interaction between this factor and Orientation ($F(1,23)=26.14$,$p=.000$, $\eta^2=.532$). In the Conflict condition, inverted trials yielded faster RTs than upright trials ($M_{Upright}$=1601ms, $M_{Inverted}$=1199ms, $t(23)=4.08$, $p<.0001$), whereas in the No Conflict condition, inverted trials yielded slower RTs ($M_{Upright}$=912ms, $M_{Inverted}$=1195ms, $t(23)$=-

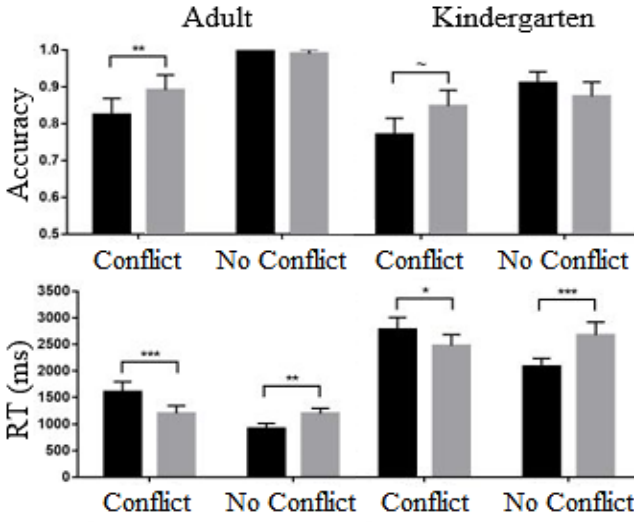


Figure 2. Accuracy and RT outcomes. Upright trials are depicted in black, and Inverted trials are depicted in gray. Error bars represent standard errors of the mean.

3.753,p=.001). These results are consistent with our prediction that semantic category knowledge influences perceptual similarity judgments such that participants are slower to judge perceptual similarity when it is in conflict with category membership, and that this influence is attenuated by inversion.

The pattern of results for kindergarten participants was similar to the pattern observed for adults. The repeated measures ANOVA revealed a main effect of Semantic Conflict (F(1,23)=6.096, p=.021, η^2=.210), and no main effect of Orientation (F(1,23)=2.072, p=.164). As in adults, the main effect of Semantic Conflict was qualified by a significant interaction between this factor and Orientation (F(1,23)=23.105, p<.0001, η^2=.501). Children's responses were also faster for inverted trials in the Conflict condition (M_{Upright}=2788ms, M_{Inverted}=2487ms, t(23)=2.153, p=.042), and slower for inverted trials in the No Conflict condition (M_{Upright}=2092ms, M_{Inverted}=2685ms, t(23)=-4.376, p<.0001).

The pattern of results in this experiment is broadly consistent with the prediction that decrements in accuracy and RT of visual similarity judgments due to conflict between category membership and visual similarity is attenuated by inversion. Both adults and children were less accurate and slower to respond on Upright than on Inverted trials in which category membership and visual similarity conflicted (though the effects on accuracy were marginally significant in children). In contrast, inversion did not improve performance in the No Conflict condition. Participants were instead similarly accurate on both upright and inverted No Conflict trials, and in fact slower on inverted than upright trials.

The observation of slower RTs in the inverted versus upright trials in the No Conflict condition was not specifically predicted in our semantic contamination hypothesis. This finding indicates that, in the absence of semantic conflict, inversion generally slows down responses in even non-semantic tasks such as the visual similarity judgment task used here. This possibility underscores the importance of the finding that inversion *speeds up* responses in the presence of semantic conflict.

General Discussion

Investigating the contributions of visual similarity and semantic relatedness is the focus of a wide range of research endeavors (Deák & Bauer, 1996; Gelman & Markman, 1986; Konkle et al., 2010; Lupyan et al., 2010; Sloutsky & Fisher, 2004). A critical component of this research is calibrating the degree to which the stimuli used in experiments are visually similar or semantically related. With respect to calibrating visual similarity, approaches taken to date in which researchers intuit or ask participants to judge visual similarity are undermined by the possibility that semantic knowledge contaminates visual similarity judgments. The purpose of this study was to examine this possibility and test whether semantic contamination is attenuated by inverting stimuli.

Our findings show that both adults and children show an effect of semantic conflict on visual similarity judgments that lowers accuracy and slows response times, and that is attenuated by inversion. Conversely, in the absence of semantic conflict, inversion slowed response times. Taken together, these findings suggest that semantic knowledge contaminates visual similarity judgments, and that inversion attenuates semantic contamination. Therefore, inverting stimuli for which visual similarity judgments are elicited provides a viable approach to calibrating stimuli for research investigating the role of visual similarity and semantic relatedness in various facets of cognition.

For example, recent cognitive neuroscience studies have investigated the degree to which knowledge about semantic relations is encoded as similar patterns of activity in brain regions involved in visual processing versus only in regions involved in subsequent amodal processing (Bruffaerts et al., 2013; Weber, Thompson-Schill, Osherson, Haxby, & Parsons, 2009). However, overlap between semantic relatedness and perceptual similarity renders it difficult to determine which of these sources is responsible for observed brain activity pattern similarities. For instance, a study conducted by Weber et al. (2009) found that the similarity of brain activity patterns in visual cortex evoked by viewing a set of animals correlated with behavioral judgments of their semantic similarity, but the fact that behavioral judgments of semantic similarity were in turn highly correlated with judgments of visual similarity renders it difficult to determine whether brain pattern similarity was related to semantic or visual similarity.

The approach introduced here to collecting visual similarity judgments that are relatively uncontaminated by semantic knowledge provides a route for attenuating such confounds. For example, the paradigm introduced in this study could be used to calibrate both items that are semantically but not visually similar, and items that are visually but not semantically similar, for studies that aim to disentangle the contributions of these forms of similarity to various facets of cognition and/or brain activity. The finding that this approach is similarly effective in both adult and child samples renders it viable for studies of both adult and developmental cognition. Therefore, the validation of the inversion approach demonstrated by the present study has the potential to support progress in a variety of lines of research.

Conclusions

This study demonstrated that conflict between visual similarity and semantic category membership slows and reduces accuracy of visual similarity judgments of upright, but not inverted stimuli for both adults and children. Therefore, this study suggests that semantic knowledge contaminates visual similarity judgments in adults and children, and that inversion attenuates this contamination.

Acknowledgements

This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education, Institute of Education Sciences (R305B040063) and by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition–

Scholar Award (220020401) to the second author. We thank children, parents, teachers, and participating institutions for making this project possible: the Children's School at Carnegie Mellon University, Sacred Heart Elementary School, and Phipps Conservatory. Finally, we thank Anna Vande Velde and Manon Sohn for their help collecting data.

References

Blaye, A., Bernard-Peyron, V., Paour, J.-L., & Bonthoux, F. (2006). Categorical flexibility in children: Distinguishing response flexibility from conceptual flexibility. *European Journal of Developmental Psychology, 3*, 163-188.

Bruffaerts, R., Dupont, P., De Grauwe, S., Peeters, R., De Deyne, S., Storms, G., & Vandenberghe, R. (2013). Right fusiform response patterns reflect visual object identity rather than semantic similarity. *NeuroImage, 83*, 87-97.

Chen, J., & Proctor, R. W. (2012). Influence of category identity on letter matching: Conceptual penetration of visual processing or response competition? *Attention, Perception, & Psychophysics, 74*, 716-729.

Deák, G., & Bauer, P. (1996). The dynamics of preschoolers' categorization choices. *Child development, 67*, 740-767.

Fisher, A. V. (2011). Processing of perceptual information is more robust than processing of conceptual information in preschool-age children: Evidence from costs of switching. *Cognition, 119*, 253-264.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Goldstone, R. L. (1998). Perceptual learning. *Annual review of psychology, 49*, 585-612.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27-43.

Jolicoeur, P., & Milliken, B. (1989). Identification of disoriented objects: effects of context of prior presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 200-210.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*, 558-578.

Lawson, R., & Jolicoeur, P. (2003). Recognition thresholds for plane-rotated pictures of familiar objects. *Acta Psychologica, 112*, 17-41.

Long, C., Lu, X., Zhang, L., Li, H., & Deák, G. O. (2012). Category label effects on Chinese children's inductive inferences: Modulation by perceptual detail and category specificity. *Journal of experimental child psychology, 111*, 230-245.

Lupyan, G., Thompson-Schill, S. L., & Swingley, D. (2010). Conceptual penetration of visual processing. *Psychological Science*682-691.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in psychology, 4*, 1-14.

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and brain sciences, 22*, 341-365.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: a similarity-based model. *Journal of Experimental Psychology: General, 133*, 166-187.

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia, 47*, 859-868.

Human Reinforcement Learning of Sequential Action

George Kachergis[2], **Floris Berends**[1], **Roy de Kleijn**[1], **& Bernhard Hommel**[1]
[1]Institute of Psychology / Leiden Institute for Brain & Cognition, Leiden University
Leiden, the Netherlands
[2]george.kachergis@nyu.edu
Psychology Department
New York University

Abstract

Learning sequential actions is an essential human ability, for most daily activities are sequential. We modify the serial reaction time (SRT) task, originally used to teach people a consistent sequence of button presses by cueing them with the next target response, to record mouse movements, collecting continuous response trajectories. Further, we introduce a reinforcement learning version of the paradigm in which the next target is not cued. Instead, learners must explore response alternatives, and receive a penalty for each incorrect response, as well as a reward for a correct response. Participants are not told that they are to learn a single deterministic sequence of responses, nor that it will repeat (nor how often), nor how long it is. Given the difficulty of the task, it is unsurprising that some learners performed poorly. However, many learners performed remarkably well, and some acquired the full 10-item sequence within 10 repetitions. We compare the high- and low-performers' detailed results in this reinforcement learning (RL) task with a cued trajectory SRT task, finding both similarities and discrepancies. Finally, we note that humans in this task outperform three standard RL models and have different patterns of errors that suggest future modeling directions.

Keywords: Sequence learning; serial reaction time task; sequential action; reinforcement learning; movement trajectory

Introduction

Traditionally, the bulk of cognitive psychology studies deal with single stimulus-response actions or decisions. However, much of human behavior can better be described as sequential action, consisting of partially-ordered hierarchies of simple actions–from cooking and cleaning to speaking and sports. More recently, several diverse lines of research have considered sequential action learning, stemming from various domains including linguistics (Elman, 1990; Saffran, Newport, & Aslin, 1996), implicit learning (Nissen & Bullemer, 1987; Cleeremans & McClelland, 1991), and everyday actions (Cooper & Shallice, 2000; Botvinick & Plaut, 2004).

Throughout these lines of research, a recurring topic of interest has been the process by which action-sequences are acquired, stored, and executed. Early work by Nissen and Bullemer (1987) focused on the modulatory role of attention, and introduced the Serial Reaction Time (SRT) task. Unbeknownst to the participants, the task utilized a 10-symbol sequence, with each successive symbol indicating which of four corresponding keys was to be pressed next. Participants showed improved reaction times across training, although the knowledge gains seemed to be implicit: when explicitly asked at the end of the experiment, they were unable to reproduce the sequence. Nissen and Bullemer (1987) concluded that attention is critical in developing awareness of learned behavior. The role of attention in the SRT task was further studied in Fu, Fu, and Dienes (2008), finding that reward motivation can improve the development of awareness of the sequence. Fu et al. (2008) reasoned reward motivation regulates the amount of attention paid to the stimuli, which in turn facilitates sequence learning.

Cleeremans and McClelland (1991) demonstrated that associative processes could account for the improvement in performance. They adapted the SRT paradigm to include a sequence derived from a 'noisy' finite-state grammar, and showed that the presence of grammatical structure facilitates sequence acquisition. Cleeremans and McClelland (1991) explained their findings with a simulation of the learning process. A Simple Recurrent Network (SRN; Elman (1990)) was able to produce results similar to human performance, findings that were later confirmed by Boyer, Destrebecqz, and Cleeremans (2005). The SRN demonstrates that associative processes are sensitive to the statistical structure of the training material, implying that rule-like behavior can emerge from networks trained on structured sequences. Indeed, Botvinick and Plaut (2004) showed that the SRN is capable of producing everyday hierarchical actions such as coffee- and tea-making after training on a set of valid examples that varied in order and complexity (e.g., sugar then milk or vice-versa, or only milk). Motivated by a study (Stadler, 1995) showing that introducing a longer delay (2000 ms instead of 400 ms) to a random selection of the response-stimulus intervals reduced sequence learning, Dominey (1998) proposed another recurrent network model that is able to account for both serial structure effects and temporal structure (i.e., patterns of delays).

To better reveal the mechanisms behind human sequential action learning, more information is needed than just the speedup of keypresses across an experiment. To measure real-time dynamics and uncertainty during learning, the SRT task was adapted to require and record movement trajectories in (Kachergis, Berends, de Kleijn, & Hommel, 2014b, 2014a). That is, instead of measuring discrete button-presses, continuous recordings were made during a mouse tracking task that replaced the original SRT's buttons with locations on a computer screen. The trajectory SRT paradigm not only replicated earlier findings, but also found evidence of sequential context effects (e.g., predictive movements towards the next response location), and unveiled changes in the movement dynamics (e.g., pre- and post-stimulus onset).

Paradigms such as artificial language learning tasks and the

SRT task have demonstrated that sequence learning is dependent upon the statistical structure within the training material, and attention towards the stimuli which is modulated by reward motivation. However, rewards do not only provide a source of motivation: in many situations they are an irreplaceable source of information–perhaps even the only feedback. Contrary to the SRT task, everyday human action learning is often not characterized by cued responses but by exploring the environment and learning which actions result in positive effects, and which result in negative effects. We believe that, in order to investigate human sequence learning in an ecologically valid manner, it is necessary to draw in another line of research: reinforcement learning (RL), a well-established paradigm in the field of machine learning (Sutton & Barto, 1998), which of course was originally motivated by much earlier behaviorist stimulus-response learning studies (e.g., Skinner (1950). RL paradigms allow learning agents to interact with a task solely through observations, actions, and rewards. The rewards validate the actions, without the need for explicit cueing or other forms of instruction. Thus, learning is exploratory, and accomplished via trial-and-error.

Although much reinforcement learning research is conducted in computer simulation, the inspiration for the approach and many algorithms is in fact rooted in animal behavior (Sutton & Barto, 1998) and there is evidence that similar processes play a role in human learning. For instance, the error-related negativity (ERN) event-related potential (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Gehring, 1992) has been studied extensively as a component of error processing. The ERN originates in the brain whenever task-relevant errors are committed. Holroyd and Coles (2002) links the ERN to the mesencephalic dopamine system, and proposes it is the result of a negative reinforcement signal which it conveys to the anterior cingulate cortex.

The current study adapts the trajectory SRT task to allow for free movement and limited instruction, allowing learners to explore and learn from trial-and-error. This RL sequence learning paradigm allows us to study the effect of rewards on sequence acquisition in more detail, yielding not only correct response times but also mistakes over time, which may be indicators of distinct mechanisms. For example, committing an error indicates incomplete knowledge (or a lapse in memory), whereas RT may be correlated with overall certainty or fluency at that point in the action sequence. Thus, we investigate the RL paradigm data both in terms of earlier trajectory SRT data and in comparison to three standard RL models.

Experiment

The goal of the current study is to examine sequence learning within the trajectory SRT paradigm, and to compare human performance to basic baseline reinforcement learning models. The trajectory SRT task was adapted to no longer cue participants with the next target position, forcing them to instead explore the response alternatives until the correct one was found. Moving the mouse cursor from the previous target to another response alternative resulted in a reward (+1) or penalty (-1) that was accumulated throughout the experiment and displayed continuously. Upon reaching a valid target, it would change color to green, add to the score by +1, and allow the participant to continue exploring. Reaching for an invalid target caused it to change to red, subtract from the score by 1, while the cursor was relocated to the previously occupied target, effectively resetting the participant's progress. Target validity was determined by a recurrent sequence, taken from the Nissen and Bullemer (1987) study, and adapted to fit the trajectory SRT paradigm. Designating the stimuli as numbers from left to right, top to bottom, the sequence read 4-2-3-1-3-2-4-3-2-1.

Methods

Participants Participants in this experiment were 13 Leiden University students and employees (age: M = 23.9, sd = 6.4) who participated in exchange for 3.5 euros or for course credit.

Procedure Participants were instructed that they would be presented with four target squares in the corners of the screen which they were to explore by moving the mouse, each time resulting in either a gain or loss of one point. Participants were told to try to maximize their score, which was displayed continuously at the top of the screen. Unbeknownst to the participants, only one of the four targets would be valid at any given moment, but all were colored blue, so the target could not be visually distinguished. Upon reaching a valid target, its color would change to green momentarily and the score would increase by one. The participant would be able to continue exploring for the next target. Arriving at an invalid target caused it to change to red momentarily and the score was decreased by one, while the cursor was relocated to the previously occupied target. Thus, although there were no instructions explicitly indicating it, participants likely inferred that they had chosen the incorrect stimulus, and should choose one of the remaining two–if they also assumed the same target was never repeated immediately, which was true. In the absence of a previous target (i.e., at the beginning of the experiment or after a rest break) the cursor was moved back to the middle of the screen.

Unbeknownst to the participants, each trial consisted of a series of 10 targets (labeled 1-4 left-to-right and top-to-bottom: 4-2-3-1-3-2-4-3-2-1) that repeated continuously, with no indication where one trial stopped and the next began. Participants completed eight blocks of 10 such trials, with a short rest break after every 2 blocks (i.e., 200 correct movements). A participant who somehow knew the sequence before entering the experiment and never made a mistake would therefore make 800 movements to valid targets, receiving a theoretical maximum of 800 points. At worst, a participant with no memory of even the previous target they had tried may make an infinite number of mistakes, and may never finish the experiment. Assuming enough memory to not repeat the same invalid target more than once when seeking each target (i.e., an elimination strategy), a participant using this elimination strategy would expect on average to score 0

points, as the expected value of completing one movement successfully is 0.[1] Note that participants were not told that there was a single deterministic sequence, let alone details such as how long the sequence was.

Results

The data from all 13 participants were analyzed. Figure 1 shows a histogram of the final score achieved by each participant. The distribution of scores is non-normal (Shapiro-Wilk's $W = 0.87$, $p < .05$), instead looking bimodal, with four participants collecting less than 300 points and all but one of the rest accumulating more than 500 points each. Given the bimodal score distribution, a median split was used to divide the participants into high-performing ($\geq$ 526; 7 people) and low-performing ($<$ 526; 6 people) groups. In the high-scoring group, participants achieved almost flawless performance after only approximately 30 trials, with a final mean score of 652 (max: 725), while the low-scoring group only gradually increased their score (final mean score: 287). The remaining analyses are carried out for each group in an attempt to understand the great variability in performance—and the impressive success of the high-scoring group.

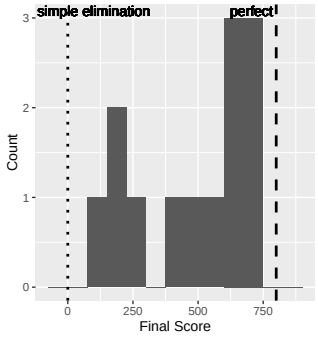


Figure 1: The histogram of participants' final scores after completing 80 sequence repetitions (800 targets) shows a bimodal distribution (lines: elimination strategy EV=0; perfect knowledge EV=800).

Response Times The overall median response time (RT) for all stimulus arrivals was 1,401 ms (sd: 4,980). Of 10,400 correct target arrival times (median: 1,078 ms, sd: 2,216), 317 (3%) were trimmed for being too slow (median + 2 · sd). Of the 4,117 incorrect stimulus arrival times (median: 2,397 ms, sd: 8,401), 100 were trimmed for being too slow (2.4%). Each subject's median RT for correct and incorrect movements was computed for each 10-trial block. Figure 2 shows the mean of subjects' median correct and incorrect RTs over the experiment, split into high- and low-performing group. RTs for correct movements improve in both groups during the first few blocks, but the high-scoring group speeds up more than the low-scoring group. Figure 2 also shows that the rare incorrect RTs for the high-performing group get slower over the course of the experiment, whereas the low-performing group's incorrect RTs only increase a bit. The strikingly slow mistakes of high-performing participants, compared to mistakes that are barely slower than correct movements for the low performers may indicate a different mode of behavior. A possible explanation is that low performers are simply not trying to learn a sequence, or do not expect it to to be deterministic, whereas high performers explicitly learn the sequence, and when they are uncertain they must pause to try to recall the next target.

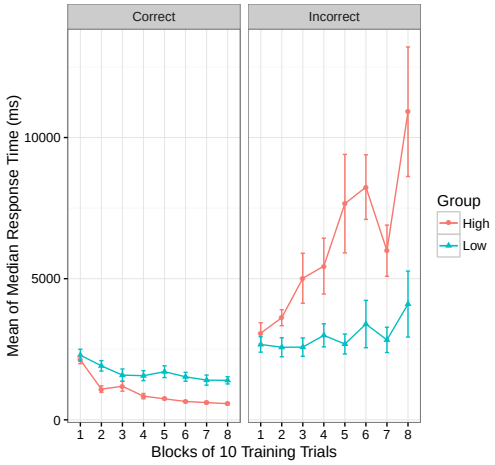


Figure 2: The mean of subjects' median correct RTs by block (left panel) shows that high-performers' RTs improved more than the low-performers' RTs over training. The mean of subjects' median incorrect RTs by block (right panel) shows that the high-performing group's incorrect RTs actually increased, whereas the low-performing group's stayed roughly the same across the experiment. Error bars show +/-1SE.

Accuracy The mean number of mistakes made over the entire experiment was 19.8 (sd: 21.3) for the high-scoring group, and 63.5 (sd: 11.9) for the low-scoring group. Over time, the number of mistakes decreased especially for the high scoring group. Examining the mistakes made by each group of participants according to where they were in the sequence revealed that for both groups the fifth stimulus was particularly challenging. This is reflected in the mean number of mistakes for each group (see Figure 4, as well as in the mean RT to the target by sequence position (see Figure 3).

Comparison to previous research The pattern we observe in the accuracy and response time data bears some resemblance to the pattern observed in a previous trajectory SRT study that used cues (Kachergis et al., 2014b). Although the task in this study (RL) was fundamentally different from the previous study (cued SRT), the same sequence was used in both experiments which enables us to compare the scaled response times from the former and the accuracy from high- and low-performers in the latter experiment. Shown in Figure 5, we see a similar pattern across experiments, and the overall mistakes per position in the RL experiment and the correct RTs in the cued experiment are significantly correlated ($r = 0.64$, t(8)=2.36, $p < .05$), with detailed comparisons below. We also compared the data to the Simple Condensator Model (SCM) proposed by Boyer et al. (2005), which implements a negative recency bias: expectation (activation) for every response builds at each step until a given response occurs,

[1] 33% of chance success in one try (+1), 33% chance of success in two tries (-1+1), and 33% chance of success in three tries (-1-1+1).

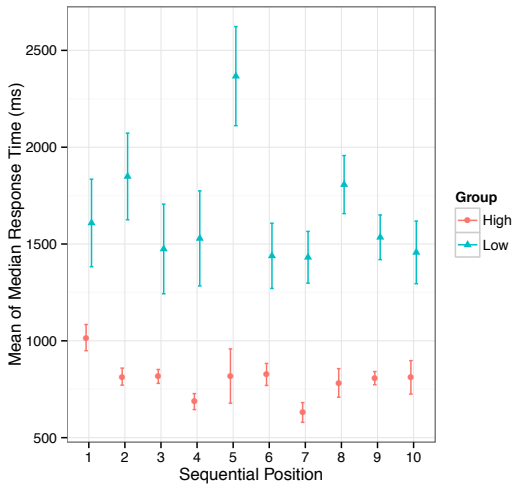


Figure 3: Mean of subjects' median correct response times by median split and sequential position. The correct RTs for the two performance groups were not significantly correlated ($r = 0.17$, t(8)=0.48, $p = 0.65$). Low-scorers were slowest at position 5, followed by 2 and 8, whereas high-scorers were worst at position 1, and almost consistently fast besides that. Error bars reflect +/-1SE.

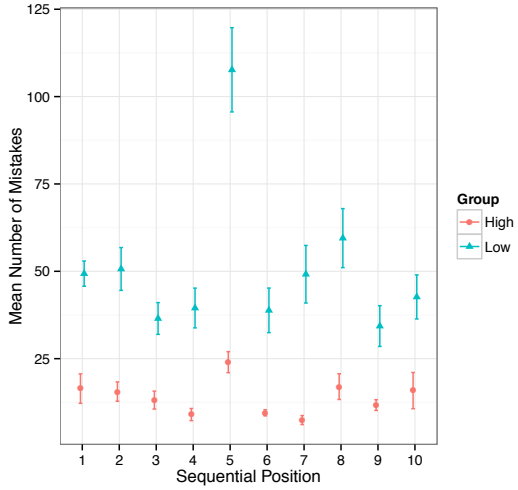


Figure 4: The mean number of mistakes made per block by position in the sequence split by performance group. The errors are highly correlated ($r = .79$, t(8)=3.68, $p < .01$), though note how much worse sequence position 5 was for the low-performing group relative to the next-worst position (8). Low-performers showed twice as many errors in position 5 as in 8, while the high-performing group showed only a 25% increase in errors. Error bars reflect +/-1SE.

and activation resets. Thus, stimuli that have been used least recently have highest activation—and fastest RTs. The SCM has previously been shown to correspond closely to human cued SRT responses, and Figure 5 shows it mirrors mistakes in both groups of the RL experiment quite well.

We examined mistakes and correct response times by their sequential position, and compared these to RTs from the previous cued SRT study. Overall, there is a significant correlation ($r = .88$, t(8)=5.37, $p < .001$) between correct RTs from the RL study and RTs from the cued SRT study. Comparing the cued RTs to the high- and low-scoring groups separately revealed a difference between the groups. The cued SRT RTs do not correlate significantly with the high-scoring group's RTs ($r = .51$, t(8)=1.68, $p = .13$), but do correlate significantly with the number of mistakes made in the RL study ($r = .83$, t(8)=4.18, $p < .01$). The low-scoring group shows the opposite pattern. The cued SRT RTs correlated significantly with the RL correct RTs ($r = .80$, t(8)=3.79, $p < .01$) but not with the RL mistakes ($r = .57$, t(8)=1.96, $p = .09$). Comparing the two groups with each other revealed a significant correlation in mistakes ($r = .79$, t(8)=3.68, $p < .01$), but no significant correlation in RT ($r = .17$, t(8)=0.48, $p > .05$).

The SCM was strongly correlated with the low-scoring RL group's RTs ($r = .93$, $t(8) = 6.93$, $p < .001$), but not with the high-scoring RL group's RTs ($r = .16$, $t(8) = 0.45$, $p = .67$). The low-scorers, failing to discern the repeating sequence, may have mostly used a negative recency bias. The SCM was correlated with the mistakes of both RL groups (low: $r = .89$, $t(8) = 5.52$, $p < .001$; high: $r = .74$, $t(8) = 3.12$, $p = .01$).

Models of sequence learning To compare human sequence acquisition with existing reinforcement learning models, we implemented the models using PyBrain (Schaul et al., 2010; see Figure 6 for an overview of the modeling experiment's

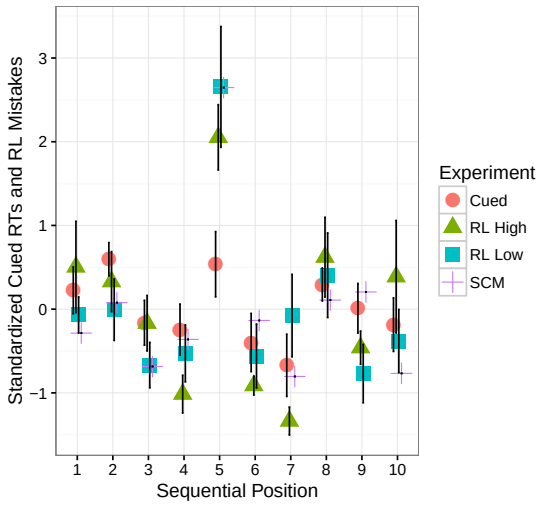


Figure 5: Scaled mean number of mistakes in the current experiment against scaled correct RTs from a cued SRT study Kachergis et al. (2014b) by sequence position. Error bars show +/-1SE.

setup). The environment contains all data regarding the targets, which it passes to the task, which in turn passes the current state of the environment to the agent, which selects the relevant action. The action is evaluated by the environment, which updates itself and passes a reward to the agent. The reward is used to update the agent's strategy, and the model continues with the next step. We defined the reinforcement learning SRT task in this framework for our simulations.

As in the human experiment, the data regarding the targets was only partially-visible to the agent. The task acted as a veil through which a certain state would be observable. To a human participant, the current position in the sequence would

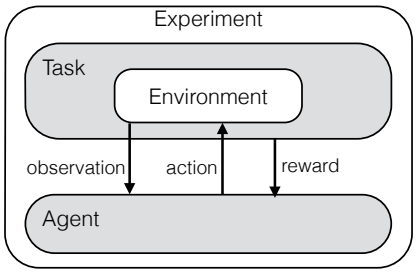


Figure 6: Overview of the experimental setup for the RL models. Each plated component is a PyBrain class, which interact with each other according to the arrows to simulate the same trial-and-error learning process that humans undergo.

be obvious, as it was colored differently from the other stimuli. At a minimum, the immediately prior occupied position was probably obvious as well, readily available in memory. Positions preceding that, however, might not be reliably accessible in memory. In the sequence we used (4231324321), following (Nissen & Bullemer, 1987), each position's identity is fully determined by the previous two positions. That is, one could perfectly predict the next position given only the two prior to it–assuming one has determined that there is a deterministic, periodically-repeating sequence. The RL models we use rely on a set of third-order observations, assuming that the models know their current position and the two prior positions.

The models differ in their learning component, which is contained within the agent and maintains a mapping between states and action-values. For each given input-state there are three action-values, corresponding to the number of movements that can be made by the agent. Upon receiving a reward, the agent updates the action-values using its learning algorithm. We tested three learning algorithms: SARSA (Rummery & Niranjan, 1994), standard Q-learning, and Q(λ)–Q-learning with eligibility traces (Watkins, 1989). Q-learning is an off-policy algorithm, learning about the greedy policy, updating old action-values using the maximum of all action-values for the current state, while it stochastically selects actions, sometimes exploring. SARSA is on-policy: instead of the maximum, it also takes into account the action it has selected for the current state. The eligibility traces in Q(λ) are temporary records of an event (e.g., an action or state) that help with temporal credit assignment by adding a trace to events that are eligible for learning updates. Theoretically, eligibility traces link RL temporal difference methods (like Q-learning and SARSA) to Monte Carlo methods.

These algorithms were chosen as simple baselines that differ somewhat in exploratory behavior and learning speed, and thus may be suitable to compare to human behavior which varied widely. As with the human participants, the simulated SARSA and Q-learners were tasked with iterating over the repeated sequence until the successful completion of 800 movements. For each model, a grid search over the parameters (learning rate α and discounting factor for future rewards γ) was used to find optimal values.

The best parameters found for the SARSA model ($\alpha = .01$, $\gamma = .98$) achieved a mean final score of 200 (sd=218). The best parameters found for Q-learning ($\alpha = .38$, $\gamma = .98$) yielded a mean final score of 290 (sd=116), while Q(λ) reached a mean final score of 451 (sd=34, parameters: $\alpha = .001$, $\gamma = .95$, $\lambda = .99$). However, despite considerable learning by the end of the experiment, none of the models performed as well as the high-performing human learners, who averaged a final score of 652. Even the *maximum* scores achieved by the models were below the high-scoring humans average or maximum (human= 725; Q-learning= 518, Q(λ)=557; SARSA= 546).

Although these common RL models were unable to reach human-level performance, we thought it worthwhile to examine whether their error patterns resemble those of people. The mistakes made by the SARSA and Q-learning algorithms did not vary much by sequence position, and while Q(λ) made more mistakes in the middle of the sequence (vaguely like humans), none of the models' error patterns were significantly correlated with humans.

Discussion

We adapted the trajectory SRT paradigm to be a reinforcement learning task. The task proved to be more challenging for some than for others, as indicated by a bimodal distribution of scores, and differences in the high- and low-performing groups' response times (RTs) and mistakes by sequence position. These data may suggest that participants adopt different strategies, discussed in greater detail below. Overall, the findings of the reinforcement learning paradigm are similar to a previous trajectory SRT experiment with cued targets: RT and accuracy were correlated across experiments. In particular, data from the high-performing participants compared remarkably well to the previous trajectory SRT study, despite the task differences. The most notable similarity was the difficulty participants experienced with the fifth stimulus position.

The better-performing half of participants made very few errors after as little as 10 repetitions of the length 10 sequence. Block-by-block analysis of the RTs by performance group showed a difference in speed-up across the experiment: the high-performing group already made faster correct responses in the second block of ten repetitions, and maintained this advantage. The difference in response times to incorrect targets suggests the two groups might have used different strategies. The rare but increasingly-slow mistakes in the high-performing group may indicate retrieval attempts and an awareness of their uncertainty as to the next step. In contrast, the persistent and relatively fast mistakes of the low-performing group suggest these participants may have adopted a probabilistic view of the task, randomly trying options instead of trying to learn a deterministic pattern. This was corroborated by the particular success of the parameter-free Simple Condensator Model (SCM) in matching the low-scoring group's RTs using a simple negative recency bias. A

distinction of stimulus-based and plan-based control (Tubau, Hommel, & López-Moliner, 2007) may capture the apparent differences between the low- and high-performing groups.

The results by sequence position showed that participants in both groups had more trouble with the target in sequence position five than any other target. This is similar to the pattern observed in previous studies (Nissen & Bullemer, 1987; Kachergis et al., 2014b), and has previously been taken to indicate that participants chunk the sequence into two parts: the initial 4-2-3-1 and the final 4-3-2-1, with positions 5 and 6 bridging the two chunks. We note that the only repeated transition in the sequence (3-2, at positions 6 and 9)–which might be expected to be worse due to the higher transition probability–shows neither slow correct responses nor more mistakes for either position it occurs in, somewhat unlike (Nissen & Bullemer, 1987). Models make the clearest predictions, but in the absence of a winning theory, a comparison to other paradigms can also be illuminating.

Despite the major difference of no cueing of the next response, performance in the RL experiment was quite comparable to performance in the cued SRT study. The correlations of mistakes and RTs by sequence position indicated a difference between the low- and high-performing groups that was not immediately obvious. Overall, the cued SRT response times are correlated to RTs and accuracy in the RL experiment, whereas this is not true for both the low- and high-performing groups separately. The low-scoring group closely matched the negative recency bias of the SCM (in mistakes, but especially in RT), but the pattern and strategy of the high-scoring group is less clear.

Hoping to better understand especially the high-performing group, we developed a simple reinforcement learning model and tested three different learning algorithms. High-performing humans were still far better than the models, which on average scored roughly as well as the low-performing humans. SARSA had quite variable performance, but was lowest on average, while Q-learning with eligibility traces fared the best. Examining the models' performance by sequence position showed they did not correspond well with errors in either group of humans. This suggests that simple model-free reinforcement algorithms do not capture the process by which humans learn action sequences, even though they eventually converge on a proper solution. One explanation for this is the fact that the task and models used in studies like this do not fully capture the essence of human action learning, which is goal-directed by nature. Future studies could shed light on the role of goals in the acquisition of such action sequences, as has been shown to exist for single-step action (see, for example, Hommel, Müsseler, Aschersleben, and Prinz (2001) for one proposed mechanism of goal-directed action). The process by which humans acquire action sequences is subtle, can yield quite variable performance, and is not easily captured by simple learning algorithms. However, studying it is important, as most of human behavior is essentially sequential in nature.

Acknowledgments

The preparation of this work was supported by the European Commission (EU Cognitive Systems project ROBOHOW.COG; FP7-ICT-2011).

References

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies. *Psychological Review*, *111*, 395–429.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (2005). Processing abstract sequence structure: Learning without knowing, or knowing without learning? *Psychological Research*, *69*, 383–398.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 2987–338.

Dominey, P. F. (1998). A shared system for learning serial and temporal sturcture of sensori-motor sequences? evidence from simulation and human experiments. *Cognitive Brain Research*, *6*(163), 163–172.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late {ERP} components. ii. error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455.

Fu, Q., Fu, X., & Dienes, Z. (2008). Implicit sequence learning and conscious awareness. *Consciousness and Cognition*, *17*, 185–202.

Gehring, W. J. (1992). *The error-related negativity: Evidence for a neural mechanism for error-related processing* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679 - 709.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*, 849–937.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014a). Reward effects on sequential action learning in a trajectory serial reaction time task. *IEEE Conference on Development and Learning / EpiRob 2014*.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014b). Trajectory effects in a novel serial reaction time task. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Rummery, G. A., & Niranjan, M. (1994). *On-line q-learning using connectionist systems* (Tech. Rep. No. CUED/F-INFENG/TR 166). Cambridge University.

Saffran, J., Newport, E., & Aslin, R. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., ... Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, *11*, 743–746.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, *57*(4), 193–216.

Stadler, M. A. (1995). The role of attention in implicity learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 674–685.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tubau, E., Hommel, B., & López-Moliner, J. (2007). Modes of executive control in sequence learning: From stimulus-based to plan-based control. *Journal of Experimental Psychology: General*, *136*, 43–63.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University.

The Role of the Prefrontal Cortex in Inductive Reasoning: An fNIRS Study

Layla Unger (LUnger@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Jaeah Kim (jaeahk@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Theodore J. Huppert (huppertt@upmc.edu)
University of Pittsburgh, Center for the Neural Basis of Cognition, Clinical Science Translational Institute, Departments of Radiology and Bioengineering
Pittsburgh, PA 15260 USA

Julia Badger (julia.badger@psy.ox.ac.uk)
University of Oxford, Department of Experimental Psychology, 9 South Parks Road
Oxford, UK OX1 3UD

Anna V. Fisher (Fisher49@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

This study examined neural activity associated with inductive inference using functional Near Infrared Spectroscopy (fNIRS). Induction is a powerful way of generating new knowledge by generalizing known information to novel items or contexts. Two key bases for identifying targets for induction are perceptual similarity, and rules that specify category-relevant features. Similarity- and rule-based induction have been argued to represent distinct mechanisms, such that only rule-based induction requires executive function processes associated with the prefrontal cortex (PFC), namely: active maintenance of representations and inhibition of salient but irrelevant features. Here, we address the lack of direct empirical evidence supporting this possibility by recording PFC activity using fNIRS while adult participants (n=24) performed an inductive inference task. We found that PFC activity during induction was greater when participants had been taught a category-inclusion rule versus when participants could only rely on overall similarity.

Keywords: inductive inference; fNIRS; PFC

Introduction

Inductive inference is a powerful component of learning because it allows us to use what we already know to derive new information. For instance, knowledge that one's cat has a four-chambered heart can be generalized to other entities, such as one's dog. However, in order for inductive inference to provide a useful source of information, targets for generalization must be identified based on bearing some relationship to the known entity.

Behavioral evidence from studies of adult cognition and developmental research suggest a distinction between similarity-based induction, in which targets chosen for inductive inference are those that are globally similar to the known entity, and rule-based induction, in which targets are those that share specific critical features (Sloutsky, Kloos, & Fisher, 2007; Yamauchi & Markman, 2000). This behavioral distinction may emerge because these two forms of inference involve qualitatively different processes. For instance, similarity-based induction may involve a global assessment of the degree to which known entities and potential targets share features, whereas rule-based induction may involve maintenance of the rule-relevant feature in memory and/or inhibition of rule-irrelevant features (Sloutsky, 2010).

One critical implication of this proposal is that rule-based induction should recruit prefrontal cortex (PFC) regions associated with active memory maintenance and inhibition of salient task-irrelevant information (Konishi, Kawazu, et al., 1999; Konishi, Nakajima, et al., 1999) to a greater extent than similarity-based induction. Indirect support for this possibility comes from evidence that rule- and similarity-based processes in other forms of reasoning recruit distinct brain circuitry (Grossman et al., 2002; Koenig et al., 2005; Nomura et al., 2007; Seger & Cincotta, 2002). However, no studies have yet directly tested this possibility.

Here, we first review behavioral evidence for a qualitative distinction between similarity- and rule-based induction, and neuroimaging evidence for distinct patterns of brain activity associated with similarity- and rule-based processes in other forms of reasoning. Then, we present a study investigating whether similarity- and rule-based induction is associated with distinct patterns of neural activity.

Similarity- versus Rule-Based Induction

Studies of both adult and developing cognition have yielded evidence for a qualitative distinction between similarity- and rule-based induction. For instance, in a number of

experiments with adults (e.g., Yamauchi & Markman, 2000) researchers taught participants two artificial bug categories that were each associated with typical anatomical features and a category label, then investigated the basis on which participants either inferred the value of a bug's occluded anatomical feature in the presence of its category label, or the value of a bug's label in the presence of all its anatomical features. In the presence of a label, adults tended to make inductive inferences consistent with the label regardless of other anatomical features in a rule-based manner, whereas in the absence of a label, adults made inferences consistent with the degree to which an item's anatomical features were typical of each category in a similarity-based manner.

Developmentally, several studies suggest that rule- and similarity-based induction emerge at different ages. For instance, recent research has shown that although 4- to 5-year-olds can learn a category-inclusion rule for a set of unfamiliar items, children often fail to use it as the basis for inductive inferences (Badger & Shapiro, 2012; Sloutsky, Fisher, & Kloos, 2015; Sloutsky et al., 2007) (cf. Gelman & Davidson, 2013). With age, children shift towards making rule-based inductive inferences (Badger & Shapiro, 2012). Therefore, whereas similarity-based inference is evident from early childhood, rule-based induction appears to develop gradually.

Neural Distinction between Similarity- and Rule-Based Reasoning

As previously noted, the behavioral distinctions between similarity- and rule-based inductive inference may be associated with a neural distinction between the brain circuitry recruited during reasoning. Specifically, it has been suggested that rule-based induction is associated with greater recruitment of processes associated with the PFC than similarity-based induction (Sloutsky, 2010). Indirect evidence for this possibility comes from observations of neural distinctions between rule- and similarity-based processes in other forms of reasoning.

The primary source of indirect evidence for this distinction comes from comparisons of brain activity observed across different studies and tasks. For instance, tasks that require rule learning, such as the Wisconsin Card Sorting Task, yield significant PFC activity (e.g., Konishi, Kawazu, et al., 1999). In contrast, similarity processing is associated with more posterior brain regions (de Beeck, Torfs, & Wagemans, 2008; Weber, Thompson-Schill, Osherson, Haxby, & Parsons, 2009). Similarly, tasks that require implicit extraction of a category prototype from similarities between exemplars involve visual regions that overlap with those associated with similarity judgments (Reber, Stark, & Squire, 1998; Zeithamova, Maddox, & Schnyer, 2008). Together, these findings suggest that rule-based reasoning recruits PFC, whereas processing visual similarity recruits visual cortex regions. However, this contrast between rule- and similarity-based processes is based on a comparison between studies that used very different paradigms.

Indirect evidence from comparisons made *within* studies comes from a smaller body of research that has primarily focused on novel category learning. These studies have also found neural distinctions between processes that are related to, but do not directly map onto the rule- versus similarity-based induction distinction of interest (Grossman et al., 2002; Koenig et al., 2005; Nomura et al., 2007; Seger & Cincotta, 2002). Many such studies compare rule-based reasoning to processes that do not involve similarity perception, such as learning categories by integrating perceptual information from multiple dimensions. The small subset of studies that have compared rule- to similarity-based reasoning have done so in domains that require additional processes, such as retrieving previously experienced exemplars or semantic knowledge from memory (Grossman et al., 2002; Koenig et al., 2005). Accordingly, the nature of this distinction varied as a result of the different processes evoked by different tasks.

Such distinctions support the possibility that rule- vs. similarity-based induction recruit distinct brain regions. At the same time, the fact that different tasks used across studies yielded different neural distinctions suggests that the nature of a potential neural distinction between similarity- and rule-based induction cannot be inferred from those observed for other forms of reasoning. Therefore, this review underscores the importance of obtaining direct empirical evidence to test the prediction that inductive reasoning, similar to other forms of reasoning, relies on neurally distinct mechanisms associated with rule-based and similarity-based processing.

Present Experiment

We focused on differences in PFC activity between similarity- and rule-based induction for two reasons. First, as noted above, the qualitative behavioral distinction between rule- and similarity-based induction may emerge because rule-based induction uniquely requires processes such as focusing on a specific feature to the exclusion of others and maintaining rules in working memory (Badger & Shapiro, 2012; Sloutsky, 2010) that are associated with PFC activity (Konishi, Kawazu, et al., 1999; Konishi, Nakajima, et al., 1999). Second, the most consistent neural distinction observed between rule- and similarity-based processes in other forms of reasoning is that rule-based processing involves PFC activity (Konishi, Kawazu, et al., 1999), whereas similarity processing involves activity in more posterior regions (de Beeck et al., 2008; Weber et al., 2009).

Accordingly, the present experiment tested whether rule-based induction is associated with greater PFC activity than similarity-based induction using functional Near Infrared Spectroscopy (fNIRS), a neuroimaging technology that uses cortical changes in infrared light absorption to measure brain activity. To test this prediction, we recorded PFC activity using fNIRS while adult participants completed one of two versions of an inductive inference task modeled on a paradigm introduced by Sloutsky et al. (2007) and updated with natural kind-like stimuli by Badger and Shapiro (2012). In this paradigm, participants are asked to infer which of two "match" items shares a property attributed to a "target". In

the "Rule-Given" version, participants were taught the category inclusion rule, whereas in the "No-Rule" version, participants were not taught the rule. We predicted that the Rule-Given version would yield high rates of rule-consistent match inference choices and significant PFC activity, whereas the No-Rule version would yield high rates of similarity-consistent match inference choices and no significant PFC activity.

Method

Participants

Participants were 24 adults (15 female, M_{age}=19 years) recruited from the undergraduate community at a Northeastern university who received partial course credit.

Materials and Apparatus

Stimuli were presented on a Dell computer screen with physical dimensions 60 cm x 34 cm and pixel dimensions 1920 x 1080. Participants were seated at a desk facing the screen with their heads about 2 feet away from the screen.

Neural activity was recorded using a continuous wave (CW6) real-time fNIRS system (TechEn, Inc.), with 4 light sources, each containing 690-nm (12 mW) and 830-nm (8 mW) laser light, and 8 detectors, to give oxygenation measures in 10 channels on the prefrontal cortex. The sensors were arranged in the layout depicted in Figure 1. Sensors were snapped into a cap strip built from foam sheet and plastic mesh, and connected to the fNIRS system by via optic cables. For each participant, the cap strip was positioned on the head, centered on position FpZ by the international 10-20 coordinate system standard, and extending over the Brodmann area 10 (anterior prefrontal cortex) and area 46 (dorsolateral prefrontal cortex) bilaterally. The strip was secured to the head using a neoprene scuba cap.

Induction Task

This task was presented using E-Prime (Psychology Software Tools, 2012). Stimuli were modeled on those created by Badger and Shapiro (2012), and consisted of bugs with five anatomical features that could each take one of two values: Head shape (pointy or round), body shape (triangular or round), body color (purple or green), spot color (brown or grey), and eye color (blue or orange). Of these features, body shape and color took up a larger proportion of the stimuli than the others. Following Badger and Shapiro, one of the smaller features, head shape, was selected as the category rule-inclusion feature. Specifically, bugs with pointy heads were "Rockbugs", and bugs with round heads were "Sandbugs". All other features were category-irrelevant.

We used these stimuli to create 16 induction trials, and 51 baseline trials. Each induction trial presented a triad of bugs consisting of a Target, a Rule Match, and a Similarity Match, arranged with the Target on top and the Matches on the bottom to either the right or left (Figure 2).

In half of the induction trials, the Target was a Rockbug, and in the other half, the Target was a Sandbug. The

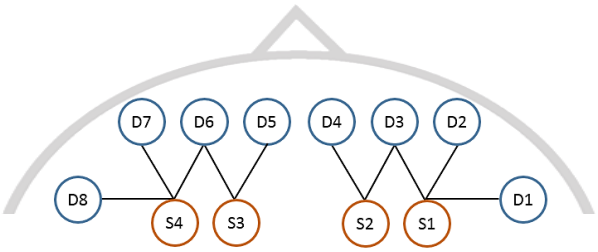


Figure 1. Probe layout including 4 sources (S1-4) and 8 detectors (D1-8), overlaid on overhead view of a head. Black lines represent source-detector channels. Source-detector distance was 2.8 cm.

assignment of Rule and Similarity Matches to the bottom right or left locations was randomized separately for triads with Rockbug and Sandbug Targets. In each triad, the Rule Match belonged to the same category as the Target but appeared dissimilar overall, whereas the Similarity match belonged to a different category but appeared similar. To accomplish this, the Rule Match had different values for all features from the Target except head shape and one of the smaller category-irrelevant anatomical features (eye or spot color), whereas the Similarity match had the same values for all features as the Target except head shape and one of the smaller features (see Figure 2). Independently for triads with Rockbugs and Sandbugs as Targets, we randomly assigned whether the small feature shared by the Rule Match and not the Similarity match was eye or spot color. All triads were pseudo-randomized such that no more than two triads with a Target bug from the same category or the Rule and Similarity Matches in the same locations appeared consecutively.

For baseline trials, we used the bugs to create a simple congruent Flanker task (Eriksen & Eriksen, 1974) (see Procedure). Specifically, each baseline trial presented three identical bugs that were all oriented to face either left or right. We approximately equated the number of times all bugs faced either left or right. Baseline trials were integrated with induction trials such that three baseline trials followed each induction trial, and one set of three baseline trials preceded the first induction trial. This ratio was used to ensure that there was a sufficient amount of baseline recording (i.e., approximately 5-10 s per each set of baseline trials, to mirror the length of time on the Induction trials). Baseline trials were pseudo-randomized such that, in each set of three baseline trials, no more than two featured bugs from the same category or bugs facing the same direction.



Figure 2. Example of induction trial. Top: Target, Bottom Left: Rule Match, Bottom Right: Similarity Match.

The Induction task was presented in two conditions: A "Rule-Given", condition and a "No-Rule" condition. The No-Rule condition consisted of only the inductive inference and baseline trials. The Rule-Given condition supplemented these trials with two Rule Demonstration slides, and 16 Categorization trials. Each Rule Demonstration slide depicted a pair of either Rockbugs or Sandbugs. Each pair of bugs had opposite values for all non-head shape features. Each Categorization trial presented a single Rockbug or Sandbug, with equal numbers of trials for each category. We assigned half of the Categorization trials for each category to appear before and half to appear after the baseline and induction trials. We pseudo-randomized each subset of Categorization trials such that no more than two bugs from the same category appeared consecutively.

Procedure

Participants were tested in a quiet space. One experimenter administered the Induction Task, and another managed fNIRS data collection (see details below).

Induction Task Participants were randomly assigned to complete either the Rule-Given or No-Rule condition of the task. The procedure for participants assigned to each condition was identical for Inductive Inference and Baseline trials. During Induction trials, participants were told that the Target possessed a novel biological property (e.g., "plaxium hormone", "tulvex nerve cells"), and asked to decide which of the two Match items shared the property. Baseline trials were modeled on the "congruent" version of the Eriksen Flanker Task, in which participants respond to some characteristic of a central stimulus in the presence of flanking stimuli that possess the same characteristic. This task was chosen as a baseline based on evidence that it elicits relatively little frontal activity (Bunge, Hazeltine, Scanlon, Rosen, & Gabrieli, 2002). In our version, participants were asked to point in the direction that the middle bug was facing.

Only Inductive Inference and Baseline trials were presented to participants in the No-Rule condition. In the Rule-Given condition, Inductive Inference and Baseline trials were presented in between an initial Rule Demonstration and Categorization phase, and a final Categorization phase. To demonstrate the rule, the experimenter showed the participant the two Rule Demonstration Slides, and provided the following descriptions of Rockbugs and Sandbugs: "These two pictures are of [Rockbugs]/[Sandbugs]. [Rockbugs live in rocks, and all have pointy heads that they use to dig in the rocks]/[Sandbugs live in sand, and all have round heads that they use to burrow in the sand]. [Rockbugs][Sandbugs] come in many different shapes and colors, but they all have [pointy heads that they use to dig in the rocks]/[round heads that they use to burrow in the sand]."

To test rule retention, participants were asked to identify the bug on each Categorization trial preceding and following the Induction and Baseline trials.

fNIRS Recording fNIRS data was recorded for each participant using custom data collection software described in Abdelnour and Huppert (2009). The fNIRS cap was first fitted on the head of the participant and the signal quality checked and adjusted if needed to make sure the fNIRS fiber optics made good contact with the scalp of the participant. After initial setup, the fNIRS data was collected at 20Hz at two wavelengths (690nm and 830nm). Following signal quality checking, the experimenter started the induction task. During the induction task, the timing of stimulus onset and offset as presented in Eprime were synced and marked in the fNIRS data by an automated analog signal sent from the computer port (of the stimulus presentation computer) to the fNIRS machine.

Results

Behavioral Results

We first determined that all participants in the Rule-Given condition successfully learned the category inclusion rule (i.e., all Rule-Given participants were 100% accurate on the initial and final Categorization Trials).Responses on the 16 Induction Trials were then analyzed to compare the degree to which participants in the Rule-Given and No-Rule conditions chose the Rule Match. The Rule Match was chosen significantly more often by participants in the Rule-Given condition (M=60.94%) than by than participants in the No-Rule condition (M=10.42%) ($t(22)=5.001$, $p<.0001$).

fNIRS Results

Pre-Processing The raw fNIRS data at the two wavelengths were converted into estimates of oxy- and deoxy-hemoglobin using the modified Beer-Lambert law (Cope et al., 1988) with a differential pathlength factor of 6 for both wavelengths. The data was resampled to 4Hz for statistical analysis using an autoregressively pre-whitened weighted least-squares regression model (Barker, Aarabi, & Huppert, 2013). In brief, the stimulus timing of the induction trials are used to construct a hypothesis of the timing of the expected response based on a canonical hemodynamic response. This model is then statistically compared against the data using a general linear model and brain activity is inferred from the statistical tests of the coefficients of this linear model. The iteratively whitened weighted least-squares regression described in Barker et al was used to solve this general linear model and had been previously shown to have substantially improved sensitivity and specificity and control of type-I errors compared to other regression methods for fNIRS data in the presence of physiological noise and potential motion-artifacts from slippage of the head cap. This analysis is similar to the general linear model and statistical parametric modeling methods commonly used in functional magnetic resonance imaging (fMRI) (e.g. SPM; Tak, Uga, Flandin, Dan, and Penny (2016)).

fNIRS Analysis Processed fNIRS data were analyzed to first compare activity during induction trials to baseline trials for

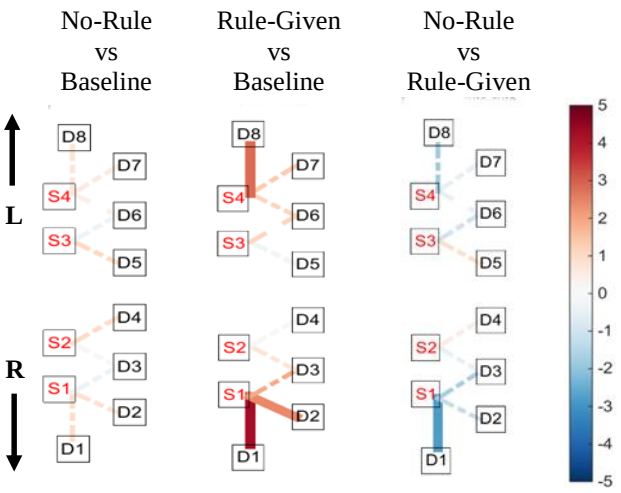


Figure 3. Group-level contrasts of oxy-hemoglobin signals for No-Rule minus baseline, Rule-Given minus baseline, and No-Rule minus Rule-Given. The color of the line for each source-detector represents the contrast t-statistic as marked on the color bar on the right. Solid lines represent significant t-statistics.

each condition, and then directly compare activity during induction trials in each condition (see Figure 3). The canonical general linear model for regression used in fMRI analysis, based on convolving the neural responses with the standard hemodynamic response function basis from SPM8, was used for statistical testing of neural activity between conditions (Friston et al., 1994). The comparison to baseline analyses revealed that Induction Trials in the Rule-Given condition were associated with significantly stronger activity in channels S1-D1 (Source 1 to Detector 1; $t(440)>4.249$, $p<2.616e-05$) and S4-D8 ($t(440)>2.827$, $p<0.005$), which corresponds approximately to Brodmann area 46, bilaterally, and channel S1-D2 ($t(440)>2.419$, $p<0.05$), which corresponds approximately to right Brodmann area 10. In contrast, no channels revealed significantly greater than baseline activity during induction trials in the No-Rule condition (all $ts<1.031$, all $ps>.065$). The direct comparison between induction trial-activity in each condition revealed significantly stronger activity in the Rule-Given than the No-Rule condition in channel S1-D1 ($t(440)>2.815$, $p<0.006$), which corresponds approximately to Brodmann area 46.

Discussion

The purpose of this study was to use fNIRS to investigate the possibility that rule-based induction recruits PFC, a brain region associated with executive functions, to a greater extent than similarity-based induction. Participants completed an inductive inference task in which they could infer that a novel property attributed to a Target item was shared by either a similar looking item from a different rule-based category, or a dissimilar looking item from the same rule-based category. Participants who were taught the category rule prior to the induction task both tended to choose the dissimilar looking same-category item, and revealed significant PFC activity in comparison to baseline. In contrast, participants who were not taught the rule tended to choose the similar looking different-category item, and did not reveal PFC activity above baseline. Finally, participants in the Rule-Given condition showed stronger PFC activity during induction than participants in the No-Rule condition.

These results support the proposal that rule- and similarity-based induction represent qualitatively distinct processes. Specifically, rule-based induction may uniquely require executive functions associated with PFC such as the active maintenance of rules in memory, and/or inhibition of rule-irrelevant input (Badger & Shapiro, 2012; Konishi, Kawazu, et al., 1999; Konishi, Nakajima, et al., 1999; Sloutsky, 2010). This distinction is consistent with similar distinctions observed between rule- vs. similarity-based processes in other forms of reasoning (Grossman et al., 2002; Koenig et al., 2005; Nomura et al., 2007; Seger & Cincotta, 2002). These results also support the proposal that the more protracted development of rule- vs. similarity-based induction implicates a greater role for the slow-maturing PFC in rule-based than in similarity-based induction. The findings presented here provide a foundation for further investigation into currently unresolved questions about inductive inference processes, as described below.

Limitations and Future Directions

The present study sets the stage for pursuing several questions that follow on from the present findings and remain currently unresolved. First, the past research that inspired the prediction that PFC activity should be evoked to a greater extent with rule- vs. similarity-based induction also predicts that activity in posterior regions associated with visual processing should be evoked to a greater extent with similarity- vs. rule-based induction. The present study investigated only a one-way dissociation; future research should investigate the predicted double-dissociation to provide further insight into the distinction between rule- and similarity-based inductive inference.

Second, although participants in the Rule-Given condition chose the category match more often than t in the No-Rule condition, they did not always do so. This may reflect variability in the degree to which different individuals perform rule-based induction. Future research should therefore test whether such variability both within and across individuals is associated with differences in PFC activity.

Finally, the current work provides a foundation from which to investigate the neural underpinnings of the previously observed distinction between the developmental trajectories of rule- and similarity-based induction. The present study was inspired in part by the possibility that rule-based induction emerges more gradually than similarity-based induction because the former uniquely recruits brain circuitry involving the slow-maturing PFC. However, no research has directly tested this possibility. Because the present study used a child-appropriate paradigm and imaging technology, the approach used here could be used to study the development of the role of the PFC in rule- versus similarity-based induction.

Conclusion

This study investigated whether a proposed qualitative distinction between rule- vs. similarity-based induction (in which the former uniquely involves memory maintenance and/or inhibition) corresponds with a neural distinction in which rule-based induction uniquely recruits PFC. The findings presented here are consistent with the proposed neural distinction, and lay a foundation for further research into the development of rule-based induction.

Acknowledgements

This work was supported by a grant from the Pennsylvania Department of Health's Commonwealth Universal Research Enhancement Program, a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education, Institute of Education Sciences (R305B040063), and by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition– Scholar Award (220020401) to the last author. We additionally thank Francois Ban and Kevin Long for their help collecting data.

References

Abdelnour, A. F., & Huppert, T. (2009). Real-time imaging of human brain function by near-infrared spectroscopy using an adaptive general linear model. *NeuroImage, 46*, 133-143.

Badger, J. R., & Shapiro, L. R. (2012). Evidence of a transition from perceptual to category induction in 3-to 9-year-old children. *Journal of experimental child psychology, 113*, 131-146.

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical optics express, 4*, 1366-1379.

Bunge, S. A., Hazeltine, E., Scanlon, M. D., Rosen, A. C., & Gabrieli, J. (2002). Dissociable contributions of prefrontal and parietal cortices to response selection. *NeuroImage, 17*, 1562-1571.

Cope, M., Delpy, D., Reynolds, E., Wray, S., Wyatt, J., & Van der Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data *Oxygen Transport to Tissue X* (pp. 183-189): Springer.

de Beeck, H. P. O., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *The Journal of Neuroscience, 28*, 10111-10123.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics, 16*, 143-149.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping, 2*, 189-210.

Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive psychology, 66*, 327-353.

Grossman, M., Smith, E. E., Koenig, P., Glosser, G., DeVita, C., Moore, P., & McMillan, C. (2002). The neural basis for categorization in semantic memory. *NeuroImage, 17*, 1549-1561.

Koenig, P., Smith, E. E., Glosser, G., DeVita, C., Moore, P., McMillan, C., . . . Grossman, M. (2005). The neural basis for novel semantic categorization. *NeuroImage, 24*, 369-383.

Konishi, S., Kawazu, M., Uchida, I., Kikyo, H., Asakura, I., & Miyashita, Y. (1999). Contribution of working memory to transient activation in human inferior prefrontal cortex during performance of the Wisconsin Card Sorting Test. *Cerebral Cortex, 9*, 745-753.

Konishi, S., Nakajima, K., Uchida, I., Kikyo, H., Kameyama, M., & Miyashita, Y. (1999). Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI. *Brain, 122*, 981-991.

Nomura, E., Maddox, W., Filoteo, J., Ing, A., Gitelman, D., Parrish, T., . . . Reber, P. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17*, 37-43.

Psychology Software Tools, I. (2012). E-Prime 2.0.

Reber, P., Stark, C., & Squire, L. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences, 95*, 747-750.

Seger, C. A., & Cincotta, C. M. (2002). Striatal activity in concept learning. *Cognitive, affective, & behavioral neuroscience, 2*, 149-161.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive science, 34*, 1244-1286.

Sloutsky, V. M., Fisher, A. V., & Kloos, H. (2015). Conceptual influences on induction: A case for a late onset. *Cognitive psychology, 82*, 1-31.

Sloutsky, V. M., Kloos, H., & Fisher, A. V. (2007). When looks are everything Appearance similarity versus kind information in early induction. *Psychological Science, 18*, 179-185.

Tak, S., Uga, M., Flandin, G., Dan, I., & Penny, W. (2016). Sensor space group analysis for fNIRS data. *Journal of neuroscience methods, 264*, 103-112.

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia, 47*, 859-868.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 776.

Zeithamova, D., Maddox, W. T., & Schnyer, D. M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *The Journal of Neuroscience, 28*, 13194-13201.

The Emergence of Linguistic Consciousness and the 'hard problem'

Abstract

Ray Jackendoff (2007) claims that most work on consciousness deals "almost exclusively with visual experience" and suggests to focus more on linguistic awareness. Jackendoff proposes that phonological ability – to divide utterances into words and syllables – is at the core of linguistic consciousness. This account can be supplemented by empirical research on language acquisition. Focusing on the step-by-step emergence of linguistic consciousness in infancy can offer new and potentially fruitful angles for investigating states of consciousness. In addition computational models of word segmentation and possible implications for linguistic consciousness are discussed.

1. Introduction

David Chalmers claims "Consciousness poses the most baffling problems in the science of the mind" (Chalmers, 2010, p. 3). The term 'consciousness' covers a range of phenomena, many of which, according to Chalmers, are easy problems (e.g., the ability to discriminate, categorize and react to stimuli, to report mental states, to focus attention, control behaviour etc.). 'Easy' does not imply that it is easy to uncover cognitive mechanisms underwriting these phenomena nor that a complete account of these mechanisms is on the horizon. Rather, it is assumed that, at least in principle, it is possible to solve the easy problems using the methods of cognitive science and neuroscience. Chalmers claims that once all questions regarding the easy problems are answered there will still be some phenomena left unaccounted for. These are the phenomena of subjective experience and they constitute the hard problem of consciousness. Philosophical thought experiments involving creatures real or imagined (e.g., Nagel (1974), Jackson's (1982) Chalmers (2002) are supposed to show that the 'what it is like' aspect of experience is inexplicable by the methods of neuroscience. Not everyone agrees with Chalmers that the hard problem may remain unsolvable by the cognitive sciences (e.g. Dennett (1991), van Gulick, (1993), McDermott, (2001), Carruthers & Schier (2014)). But many remain convinced that methods that work well for the easy problems may not be equally successful for addressing the problem Chalmers defines as follows: "an organisms is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be in that state" (Chalmers, 2010, p. 5).

Attempts to frame and to solve the hard problem have traditionally focused on sense experience in general and visual experience in particular. Here I shall focus on the emergence of linguistic consciousness during early language acquisition and suggest that doing so might illuminate 'what it is like to have linguistic experiences' and thus shed some light on the hard problem. I critically engage with Jackendoff's (2007) account and show how this account can be supplemented by empirical research on language acquisition.

2. Jackendoff 's account of linguistic consciousness

Many consciousness researchers share the position described by Zlatev here: "Language seems to be irrelevant for addressing the 'hard problem', unless one adopts a rather extreme position that it is language alone that in one way or another brings about the (illusion) of having qualitative experience" (Zlatev, 2008, 6). It may seem obvious that language has a distinctive function. If this is the case linguistic consciousness would fall under what Chalmers (2002, 2010) calls the easy problem of consciousness. However, Jackendoff (2007) suggested that linguistic consciousness may not reside in meaning but in phonology. Potentially, this proposal could provide a link between the easy and hard problems of consciousness.

According to Jackendoff the majority of consciousness researchers hold that the search for 'neural correlates' of consciousness is an important area of research because they believe that consciousness is "an emergent property of brains that are undergoing certain sorts of

activity" (Jackendoff, 2007, 77). However, Jackendoff argues that in most contemporary discussions we find "little description of how experience is actually structured – of how qualia are organized into the conscious field (Ibid., 79). Jackendoff claims that consciousness research focuses (almost) exclusively on vision. "But vision alone is perhaps too limited for an understanding of consciousness that cuts across modalities" (Ibid., 80). He suggests that considering aspects of linguistic consciousness will be illuminating.

The creative use of language presupposes a conscious state of mind and language allows reporting on the subjective experiences related to consciousness. Linguistic structure has three distinct components: phonologic, syntactic and semantic/conceptual structure. According to Jackendoff phonological structure underwrites linguistic consciousness: "When one is experiencing language, the forms of awareness -the qualia- most closely mirror phonological structure" (Ibid., 81). When experiencing language one is first aware of the sounds one perceives. It is not necessary to understand an utterance to become consciously aware of the fact that what one listens to is language (as opposed to 'unspecified noise'). When overhearing a conversation of foreigners one is aware that one listens to language without knowing what the conversation is about. On the other hand, tip of the tongue phenomena show that having meaning without phonological structure does not lead to linguistic awareness. Thus, "phonology is necessary and sufficient for the presence of linguistic qualia, and meaning is neither necessary nor sufficient" (Ibid., 82).

Jackendoff argues that conscious thought can have linguistic (phonology, words, inner speech) or non-linguistic (pictures, music) imagery. Only linguistic imagery can encode quantification, reference to absent objects, abstract concepts, etc. One becomes aware of one's thoughts through the associated phonological qualia or phonological images. Given that deaf people who use sign language do not experience phonological but visual images of hand movements Jackendoff suggests, "the form of thought itself is always unconscious" (Ibid.83). One only becomes conscious of thoughts "through the awareness of phonological structure associated with thoughts" (Ibid. 84) and Jackendoff claims that phonological structure reveals the content of these thoughts.

On Jackendoff's account phonological structure and abstract valuation features, which are bound to the structure of the linguistic percept, are essential for linguistic consciousness. The combination of phonological structure and valuation features affects attention and enhances the power of thought. The valuation dimension accounts for how the linguistic experience is related to the subject: is she the sender or receiver of linguistic signals, are the signals familiar or not, do they have affective content, are they meaningful, etc. It also provides a possible window in the 'what it is like' aspect of linguistic consciousness. Especially one of these valuation features, the external and self-initiated dimension plays an important role in the earliest stages of linguistic consciousness, during the process if language acquisition. Therefore this dimension could be directly relevant to addressing the hard problem.

Jackendoff distinguishes between linguistic percepts (+external) and linguistic images, (-external), and between self initiated and non-self initiated experiences. Hearing one's own voice while speaking is a +external, + self initiated experience. In complex situations (e.g. having a conversation with several people) one can experience several valuations simultaneously. This means "valuations are not characteristics of one's experience as a whole…they are attached to particular percepts and images" (p. 87). Jackendoff argues that these kinds of valuations cut across different sensory modalities. For example, in vision it is possible to distinguish between the perception of a blue square and the imagining of a blue square. The qualia of blue squareness would be categorized as +external/-self initiated and –external/+self initiated respectively. In acoustic perception and proprioception similar distinctions apply, suggesting that linguistic consciousness is based on similar neural correlates as other forms of consciousness. An account of consciousness that cuts across modalities certainly is attractive.

It might be possible to expand on Jackendoff's account to glean insights regarding the hard problem. Jackendoff discusses only cases that rely on subjective experiences of competent speakers who can provide detailed reports about these experiences (e.g. the phonological structure (content features) and the valuation features). Prelinguistic infants could not provide such reports but it does not follow that they have absolutely no conscious experience. It would seem that language acquisition relies on conscious experience and if linguistic consciousness is indeed based on phonological structure, then maybe it can offer insights regarding on how and when consciousness arises during infancy.

3. Language acquisition – from sub-conscious to conscious

Young infants are not able to report about their conscious states, emotions, and motivations. They cannot explain 'what it is like' to be them. However, infants process complex visual stimuli and look preferentially at faces. Initially an infant's visual acuity permits her to see only blobs, but the basic thalamo-cortical circuitry necessary to support simple visual and other conscious percepts is in place. It is likely that the infants has some basic level of unreflective, present-oriented awareness. Many of the neural circuit elements assumed necessary for consciousness are in place already by the third trimester (Koch, 2009).

Infants begin to receive rich information about their native language through exposure to spoken words when they are still *in utero*. Most fetuses begin to respond to sound at 22 to 24 weeks (Hepper & Shahidullah, 1994), and by the time babies are born their basic auditory capabilities are relatively mature (Lasky & Williams, 2005; Saffran et al., 2006). At the normal frequency of the human voice (125-250 Hz) there is little attenuation by the mother's skin, and tissues and the fetus can hear the mother talking (for reviews see Hepper, 2002; Lasky & Williams, 2005) and is affected by exposure to other external sounds (Lecanuet & Schaal, 1996). By the time infants are born they are able to discriminate the voice of their mother from those of other women (DeCasper & Fifer, 1980). Newborns also seem to be familiar with rhythmic properties of their native language. DeCasper & Spence (1986) further demonstrated that newborns showed a preference for passages of prose that their mothers had read aloud during the last six weeks of pregnancy. This preference persisted when the passages were read by another person to the newborn, suggesting that the child recognized not only the voice of the mother but also other acoustic properties of the prose (e.g., DeCasper & Fifer, 1980; Moon & Fifer, 2000). Seemingly infants learn from pre-natal input important details about the phonological structure of the prose read to them.

However, little is known about whether perception at this early stage involves conscious experience. Assuming Jackendoff's account regarding the role of phonological structure for linguistic consciousness, at this early stage the infant might already experience some of the sound-structure she perceives consciously. The fact that young infant prefer the acoustic prose they are familiar with, regardless of the voice reading the prose, might suggest that some of the valuation aspects of linguistic consciousness are not fully developed at this early stage. However, they could already be accessing the 'familiar' valuation aspect described by Jackendoff. If so this could allow some inferences regarding the earliest stages of conscious experience.

It is known that newborns can distinguish between utterances from languages that differ in rhythmic structure based on prenatal exposure to spoken language. Nazzi et al. (1998) showed that French newborns can discriminate between two unknown languages from different rhythmic classes (English versus Japanese), but they cannot discriminate between languages from the same rhythmic class (English versus Dutch). Over the course of several months infants learn to discriminate their own language from other languages in the same rhythmic class (Nazzi et al., 2000). At this early stage semantics or syntax presumably play no role in these discriminations. But phonological structure would be already accessible to the infant. Given that the learning results in increasingly fine-tuned discriminations it is possible that infants become gradually more aware of the relevant phonological features that allow these discriminations.

Between 6 and 12 months of age infants fine-tune the perception of the individual sounds that distinguish between words (or phonemes) in the language to which they are exposed. Werker and Tees (1984) found that 6- to 8-month-old babies distinguish between a wide range of sound differences that signal changes in meaning either in their native language or in non-native languages. And, while some of the 8- to 10-month-old infants were still able to discriminate non-native language contrasts, virtually all 10- to 12-month-old infants discriminated only native language contrasts. It has been suggested that these changes in perception reflect the growing ability of infants to focus their attention only on those acoustic dimensions that are relevant for their native language (Maye et al., 2002). The gradual development of early linguistic perception suggests that infants become increasingly aware of phonological structure. Given that some ability to discriminate important phonological dimensions is already present at birth it seems possible that at least some of these abilities are not dependent on consciousness. More research is

needed to confirm if and when these discrimination tasks involve conscious experience. One way to test for consciousness might be to focus on linguistic productions in early childhood.

Before producing the first words of their native language, young infants go through a phase of babbling during which they identify, acquire, and practice the sounds that are common in their language (for a review see Werker & Tees, 1999). The babbling child initially produces a wide range of sounds and later narrows this range to the sounds of her own language. For example the productions of 10-month-old infants exposed to one of four languages (French, English, Cantonese, and Swedish) are already acoustically significantly different and adults can reliably determine which productions were from languages other than their own (Boysson-Bardies & Vihman, 1991). Some researchers have shown that parental feedback can modify the phonological features of babbling (Goldstein & West, 1999; Goldstein et al., 2003) and over time the children succeed in refining their perception so as to categorize sounds along dimensions relevant to their native language.

Children need to master a complex skill set before they can produce linguistic expressions. They need to perceive their own vocalization as their own and monitor how their output matches the input they previously received. This process involves the valuations +self-initiated/+external (for their own voice), -self initiated/+ external (for the perception of someone else's voice), and +self-initiated/-external (for internally 'practicing' vocalizations and creating an inner template). Jackendoff's account offers a framework for testing which component comes 'on-line' when.

At the earliest stages of language acquisition consciousness arises at the level of linguistic phenomenology. At these early stages of language acquisition the child is not aware that what she perceives *is* language. Jackendoff claims that semantics is neither necessary nor sufficient for linguistic consciousness. Possibly, linguistic consciousness arises before the child has access to semantics. At birth the infant is able to perceive some sounds as 'linguistic' sounds and to focus her attention on them. During the first year of life she learns to fine-tune her perception and to produce the sounds of her language. This complex interplay of linguistic and non-linguistic components could provide a framework for testing hypotheses regarding linguistic consciousness.

4. Challenges from computational modeling

Computational models are used to simulate aspects of language acquisition such as speech segmentation. This is a task infants face when they learn to segment the continuous stream of language input into individual words. When infants accomplish this task they might be consciously aware of the input (though presumably not yet of the fact that this input is linguistic). However, computational models are almost certainly not conscious of any features of the linguistic input they receive. Thus if they succeed in the word-segmentation task this may indicate that consciousness is not required for this task.

It has been shown that the stochastic information contained in the linguistic input can assist speech segmentation. Monaghan and Christiansen (2010) used a computational model (PUDDLE) that closely resembles how children learn to break the stream of speech into individual words. They use input that is similar to input received by children. Like young children, PUDDLE builds its lexicon incrementally from the input. This 'strategy' does not require that the model makes multiple, simultaneous decisions about the match between a given utterance and the acquired lexicon. Just like young children, the model is initially unable to perform complex cognitive tasks simultaneously. PUDDLE thus simulates how children can take advantage of features that are readily accessible in child directed speech. The model performs like a child because "the memory resources and computational requirements are minimal" (Monaghan & Christiansen, 2010, 248). PUDDLE focussed especially utterance boundaries and the interspersal of high frequency words in speech (Ibid.). Combining these two cues resulted in very good results in the segmentation task. These results suggest that the stochastic information contained in the input might be sufficiently rich for speech segmentation and that this information might be extracted with relatively simple mechanisms. It further suggests that consciousness is not necessary to succeed in this task.

Other researchers use models that can access simultaneously several cues and combine the information to assist word-segmentation. Blanchard et al. (2010) propose that infants can learn individual words based on frequent occurrence (e.g., their own name, 'mom', frequent function words) and/or language specific phonotactic constraints (stress patterns, allophonetic variation, etc.). Frequently occurring words form the first tiny lexicon, which allows the learner to infer some phonotactic constraints. This information in turn can help to recognize additional words. The combination of these two cues solves one important problem that beginning language learners face: how can they know which phonotactic constraints apply before they know words and vise versa. Thus, "knowledge of familiar words, combined with increasingly refined phonotactic constraints, support and reinforce each other in speech segmentation" (Blanchard et al., 2010, 491).

The model PHOCUS relies on a few basic assumptions about language learning. Beginning with an empty lexicon, it incrementally adds items to the lexicon, based on phonemes that occur together (probabilistic and phonotactic cues). Phonemes that occur within words have high transitional probabilities while phonemes that cross word boundaries have low transitional probabilities (Saffran et al., 1996). In addition to these transitional probabilities, Blanchard et al.'s model could exploit phonotactic cues. Specifically, when the model encountered an unfamiliar word, it could rely on two kinds of phonotactic cues (phoneme combinations and occurrence of at least one syllabic sound per word). The combination of these two simple cues allowed the model to achieve a good performance for English input. Again, this performance was achieved in the absence of consciousness.

Computational models that are not conscious achieve good success-rates in the word-segmentation task. This suggests that consciousness is not necessary for this task. However, it does not suggest that children who succeed in the word segmentation task are *not* (at least on some level) conscious. And, successful word-segmentation is only one of many skills the child needs to master when she acquires language. While computational modeling has successfully simulated several aspects of language acquisition (from word-segmentation to complex syntactical regularities and recursion), so far no model has succeeded in simulating all aspects of language learning. If Jackendoff's overall account is correct, it might be the case that consciousness is required at a much 'earlier' stage of language acquisition than usually assumed, namely long before semantics and syntax are acquired. Maybe future research could test whether valuation features (+/- self-initiated) can be simulated by models and if so whether this leads to improved models. Given that some aspects of language acquisition can be simulated by non-conscious models it is also possible that consciousness does play an integrating role in language acquisition, linking the many different pieces of a very complex puzzle into a coherent whole.

5. Conclusions

Jackendoff's account of linguistic consciousness adds an interesting new perspective to consciousness research that has focused predominantly on visual perception. The proposal that phonology is the locus of linguistic consciousness puts the focus on phenomenology and the proposal of evaluation features grounds linguistic consciousness in personal experience. In combination phonological content and valuation features of linguistic awareness can provide access to the 'what is it like' aspect of linguistic consciousness. In addition, focusing on the step-by-step emergence of linguistic consciousness during infancy can add new and potentially fruitful angles for investigating states of consciousness. During early language acquisition infants might rely increasingly on the conscious experience of the linguistic input and model their output on some of the valuation features suggested by Jackendoff. Computational modeling work has shown that some aspects of language acquisition can be accomplished without consciousness. However, at this time no computational model of all aspects of language acquisition exists. This leaves open the possibility that linguistic consciousness is a necessary pre-condition for language acquisition.

References:

Blanchard, D., J. Heinz, & R. Golinkoff (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language, 37*, 487–511.

Boysson-Bardies B. & Vihman M. (1991). Adaptation to language: Evidence from babbling and first words in four languages. *Language, 67*, 297–319.

Carruthers, G. & Schier, E. (2014). Why are We Still Being Hornswoggled? Dissolving the Hard Problem of Consciousness. *Topoi*, 1-13.

Chalmers, D. J. (2002). Consciousness and its place in nature. In: D. J. Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings*. (pp. 247 - 272). Oxford: Oxford University Press.

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford: Oxford University Press.

DeCasper, A. & Fifer, W. (1980). Of human bonding: newborns prefer their mothers' voices. *Science, 208*, 1174–1176.

DeCasper, A. & Spence, M. (1986). Prenatal maternal speech influences on new- borns' perception of speech sounds. *Infant Behaviour and Development, 9*, 133–150.

Dennett, D.C. 1991. *Consciousness Explained.* Boston: Little-Brown.

Goldstein, M., & West, M. (1999). Consistent responses of human mothers to prelinguistic infants: The effect of prelinguistic repertoire size. *Journal of Comparative Psychology, 113*, 52–58.

Goldstein, M. King, A. & West, J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences, 100*, 8030–8035.

Hepper, P. & Shabidullah, B. (1994). Development of fetal hearing. *Archives of Disease in Childhood-Fetal Neonatal Edition, 71*, F81–87.

Jackendoff, R. (2007). *Language, Consciousness, Culture*. Cambridge, MA: MIT Press.

Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly 32*:127-136.

Lasky, R. & Williams, A. (2005). The development of the auditory system from conception to term. *Neoreviews, 6*, 141–152.

Lecanuet, J. & Schaal, B. (1996). Fetal sensory competencies. *European Journal of Obstestrics, Gynecology and Reproductive Biology, 68*, 1–23.

Maye, J., J. Werker & L. Gerken (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101–B111.

McDermott, D. (2001). *Mind and Mechanism*. Cambridge, MA: MIT Press.

Monaghan, P. & M. H. Christiansen (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*, 545–564.

Moon, C.M. & Fifer, W.P. (2000). The fetus: Evidence of transnatal auditory learning. *Journal of Perinatology, 20*, S37–S44.

Nagel, T. 1974. What is it like to be a bat? *Philosophical Review 83*:435-50.

Nazzi, T., J. Bertoncini, & J. Mehler, (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 756–66.

Nazzi, T., P. Jusczyk & E. Johnson (2000). Language discrimination by English-learning 5-month-olds: Effect of rhythm and familiarity. *Journal of Memory and Language, 43*, 1–19.

Saffran, J., R. Aslin, & E. Newport (1996). Statistical learning by 8-month old infants. *Science, 274*, 1926–1928.

Saffran, J., J. Werker & L. Werner (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In R. Siegler & D. Kuhn (Eds.). *Handbook of Child Development*. (pp. 58–108). New York: Wiley.

Van Gulick, R. 1993. Understanding the phenomenal mind: Are we all just armadillos? In: M. Davies & G. Humphreys (eds.) *Consciousness: Philosophical and Psychological Aspects*. Blackwell.

Werker, J. & Tees, R. (1984). Cross language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behaviour and Development, 7*, 49–63.

Zlatev, J. (2008). The dialectics of consciousness and language. *Journal of Consciousness Studies , 15* (6), 5-14.

An Empirical Evaluation of Models for How People Learn Cue Search Orders

Percy K. Mistry (pkmistry@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697-5100 USA

Ben R. Newell (ben.newell@unsw.edu.au)
School of Psychology, University of New South Wales, Sydney 2052, Australia

Abstract

We propose simple parameter-free models that predict how people learn environmental cue contingencies, use this information to measure the usefulness of cues, and in turn, use these measures to construct search orders. To develop the models, we consider a total of 8 previously proposed cue measures, based on cue validity and discriminability, and develop simple Bayesian and biased-Bayesian learning mechanisms for inferring these measures from experience. We evaluate the model predictions against people's search behavior in an experiment in which people could freely search cues for information to decide between two stimuli. Our results show that people's behavior is best predicted by models relying on cue measures maximizing short-term accuracy, rather than long-term exploration, and using the biased learning mechanism that increases the certainty of inferences about cue properties, but does not necessarily learn true environmental contingencies.

Keywords: learning; search order; predictive models; cue contingencies

Introduction

Making a decision requires people to search for information, decide when to terminate that search, and then make a decision based on the available information (Gigerenzer, Todd, & the ABC Group, 1999). Deciding which of two cities is larger might start with finding out whether each city is a state or national capital, whether it has an airport, and so on. At some point, the information gathering must stop, and a decision made on the basis of what is known about the available cues and their relationship to the decision criterion of population size. If people receive some sort of feedback—whether implicit or explicit, or immediate or delayed—about the accuracy of their decisions, then it also becomes possible to learn the usefulness of different cues. The field of decision making is full of models for learning how cues relate to criteria, based on principles like conditioning, reinforcement, and error correction. There are some models of how people learn when to terminate search, usually in the form of adaptive sequential sampling models, and based on principles like maximizing reward rates, controlling conflict, or maintaining confidence (Lee, Newell, & Vandekerckhove, 2014).

There are fewer models of how people learn the order in which to search. Many measures have been proposed as the basis for ordering search, including those that focus on immediate benefits like the current validity or success rate of a cue (Gigerenzer & Goldstein, 1999; Newell, Rakow, Weston, & Shanks, 2004), and those that take a longer view by focusing on information gain (Nelson, 2005). Central to calculating all of these measures are the *discriminability* and *validity* properties of a cue. Discriminability is the probability that a cue takes different values for two stimuli being compared. Validity is the probability that it identifies the correct stimulus, given that it discriminates. Despite their centrality, there are few models of how validity and discriminability are learned, and it is often simply assumed they are veridically available to people. This means, in turn, that there are few models of how people learn to order search. Exceptions are Todd and Dieckmann (2004) and Martignon and Hoffrage (2002). These, however, focus on simulation studies and lexicographic rules with one-reason decision making, which are not easily extended to cases where people search beyond one discriminating cue.

A process account of how people learn these cue contingencies and decide on search orders is relevant to any model of choice that employs sequential sampling and evidence accumulation, as well as heuristics that select cues based on learned contingencies. In this paper, we develop a modeling framework that allows for different assumptions about what cue measures are important for guiding search, and can use one of two simple learning mechanisms for ordering cue search. We evaluate the resultant 16 different models against previous experimental data measuring how people search.

Experimental Data

Our data come from experiments reported by van Ravenzwaaij, Newell, Moore, and Lee (2014) in which, on each trial, participants had to select which of two cities had a larger population. The names of the cities were not provided, but various cues—such as whether the city had an airport, a university, a sports team, and so on—were available for both cities. Participants had the option to select and view as many cues as they liked, in an order of their choosing. The cues were visually presented in a circular layout, with a random ordering for each participant, to control for presentation effects on the order of search. There were two experimental conditions: in the *known* condition, cue validities and discriminabilities were provided, while in the *unknown* condition, this information was not provided. In both conditions, participants received corrective feedback after every trial.

We focus on the *unknown* condition, within which $n = 24$ participants completed two environments: 100 trials each for Italian (9 binary cues) and USA (8 binary cues) cities. On average, about four cues are used on every trial, although there is large variability between trials and participants. Interestingly, the use of individual cues is very similar between the *known* and *unknown* conditions, with a correlation of 0.94.

Modeling Assumptions

All of the models we consider come from combining a measure of cue usefulness with a method for learning that measure, based on feedback over a sequence of trials. Search orders are determined by sampling from the learned distributions of the measure for each cue.

Measures of cue usefulness

The first four measures focus on immediate reward, and are simple to define. The remaining measures involve more detailed calculations, and we provide only the intuition behind these measures (see Nelson, 2005, for details).

Validity The probability v a cue identifies the correct choice, given that it discriminates between the two stimuli. It is the basis for search in the prominent fast and frugal heuristic known as take-the-best (Gigerenzer & Goldstein, 1996).

Discriminability The probability d that a cue takes different values for the two stimuli. Searching by discriminability is an extreme case of the linear family of measures considered by Lee and Newell (2011), Lee and Zhang (2012), and Ravenzwaaij, Moore, Lee, and Newell (2014).

Additive An average of cue validity and discriminability, $\frac{1}{2}(v+d)$. Searching using this average is a special case of the linear family of measures considered by Lee and Newell (2011) and Lee and Zhang (2012).

Success rate (SR) This is defined as $dv + \frac{1}{2}(1-d)$ by Newell et al. (2004). It measures the probability of making the correct choice by combining the probability of the cue discriminating and leading to a correct decision, with the probability of guessing correctly if it does not discriminate.

Information gain A measure of the expected reduction in uncertainty— the change in entropy of the choice options— from observing the value of the cue for the two stimuli.

Probability gain A measure of increase in the expected probability of making a correct guess.

Impact A measure of the average absolute change in the probabilities of each choice being correct, as a result of observing the value of a cue for the two stimuli, weighted by the probability of the cue providing this information.

Bayesian diagnosticity A measure of the expected weight of evidence of the cue measured in terms of likelihood ratios.

Learning mechanisms

We propose that people implicitly keep track of which cues are searched, and the success of each in discriminating, and indicating the correct choice (Lagnado, Newell, Kahan, & Shanks, 2006). Formally, after t trials, we assume people know they have searched a cue on γ_t trials, that it has discriminated α_t times, and indicated the correct choice β_t times.

Standard Bayesian Learning Given this information, a straightforward way to learn cue v and d is through Bayesian belief updating. We make the simplifying assumption that people have no strong prior beliefs (an assumption we revisit in the discussion), so that initially $v_0 \sim \text{beta}(1,1)$ and $d_0 \sim \text{beta}(1,1)$. After t trials, this means that;

$$\begin{aligned} v_t &\sim \text{beta}(1+\beta_t, 1+\alpha_t-\beta_t) \\ d_t &\sim \text{beta}(1+\alpha_t, 1+\gamma_t-\alpha_t). \end{aligned}$$

Biased Bayesian learning An alternative to standard Bayesian learning is motivated by the idea of confirmation bias, or positive test strategy, whereby people are prejudiced towards aspects that have previously produced positive results (Klayman & Ha, 1987), and the idea of selective attention, whereby people tend to focus on a limited set of attributes (Wilson & Niv, 2011). Le Pelley, Beesley, and Griffiths (2011) reported results that cues with a high level of predictive power resulted in a higher attentional bias. Beesley, Nguyen, Pearson, and Le Pelley (2015) suggested that attention bias towards the exploitation of predictive cues was more robust than an attention bias towards exploratory behavior arising from increasing uncertainty about cues. To account for these sorts of biases, we consider a second learning mechanism in which the cues that are not searched on a trial are assumed to have failed. This is contrary to standard Bayesian belief updating, which only considers information about validity and discriminability from cues actually searched. Formally, this means;

$$\begin{aligned} v'_t &\sim \text{beta}(1+\beta_t, 1+t-\beta_t) \\ d'_t &\sim \text{beta}(1+\alpha_t, 1+t-\alpha_t). \end{aligned}$$

Intuitively, biased Bayesian learning increases the tendency to persist with cues that have been successfully tested in previous trials, creating an attentional or confirmational bias for exploitation over exploration.

Determining search orders

Given a cue measure and learning mechanism, we propose that people determine a search on each trial by sampling from the learned distribution of the measure for each cue. The rank order of these samples determines the search order for that trial. Since all the cue measures we consider are deterministic functions of v_t and d_t, the learning results for these two measures presented above allow any measure to be sampled.

Model Demonstration

We illustrate our modeling framework with a simple example involving three cues. Table 1 shows the counts after $t = 30$ trials, and the corresponding validity v and discriminability d for each cue. On each trial, the order of search is determined by drawing a mental sample from the inferred distribution of the cue measure, which incorporates uncertainty. These illustrative distributions are shown in Figure 1, with the two left-most panels considering the validity measure, and the two right-most panels considering the success rate measure. For the standard learning model, uncertainty is reduced as the cue

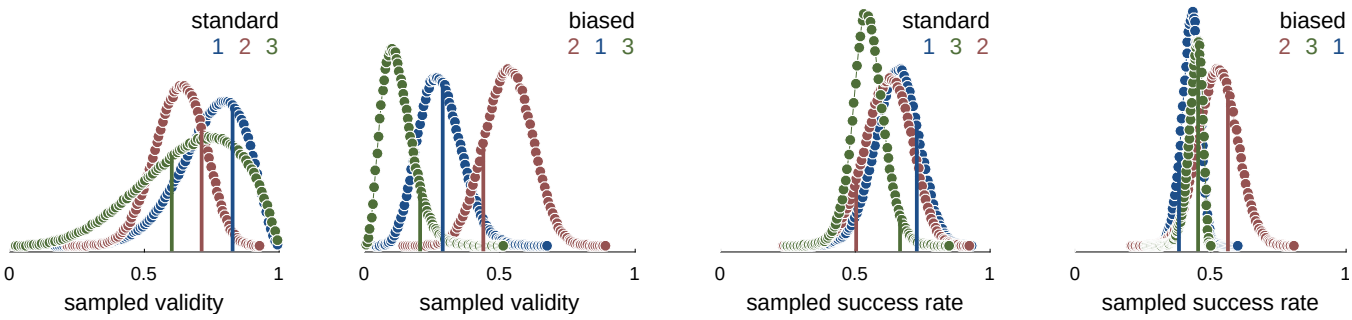


Figure 1: Demonstrations of determining search orders from learned inferences about cue measures. The two left panels relate to the validity measure, and the two right panels relate to the success rate measure. In each panel, the inferred distribution for three cues is shown, and the vertical lines indicate a sampled value that determines search order for a trial. For both the validity and success rate panel pairs, the left-most panel relates to standard learning, and the right-most panel relates to biased learning.

is sampled more often, leading to search orders more deterministically following the empirical estimates. Beliefs about v are updated only if the cue discriminates, which means the standard learning model will always result in lower certainty for v, compared to d, unless the cue always discriminates. In the biased learning model, uncertainty is reduced for both v and d on every trial, regardless of whether the cue is searched.

For this example, if search order deterministically followed the empirical rates, the validity measure would predict the cue order $[1,3,2]$, the discriminability measure would predict $[2,1,3]$, and the success rate measure would predict the order $[1,2,3]$. The other measures of usefulness might predict still different orders. The uncertainty associated with mental sampling means, however, that a distribution of search orders is predicted by each measure. Figure 1 illustrates the predicted search orders using samples from the distributions, shown by vertical lines. For the particular set of samples shown in Figure 1, the predicted search order is $[1,2,3]$ for standard learning using validity, $[2,1,3]$ for biased learning using validity, $[1,3,2]$ for standard learning using success rate, and $[2,3,1]$ for biased learning using success rate.

Table 1: Example situation after 30 trials, giving γ_{30} (number of times cue is searched), α_{30} (number of times cue is searched and discriminates), and β_{30} (number of times cue is searched, discriminates and is valid) counts for three cues.

Cue	γ_{30}	α_{30}	β_{30}	$d_{30} = \alpha_{30}/\gamma_{30}$	$v_{30} = \beta_{30}/\alpha_{30}$
1	16	10	8	0.63	0.80
2	26	25	16	0.96	0.64
3	15	4	3	0.27	0.75

Model evaluation

Generating model predictions

Because the model predictions depend on the mental samples drawn on each trial, they are inherently probabilistic. Accordingly, we generate 100 samples for each trial for each participant and measure. The predictions about cue order on each individual trial are made without the model making contact with behavioral data. The models are genuinely parameter-free, so there is no model fitting or parameter estimation involved, and thus no need to adjust for model complexity.

Evaluating model performance

A search order that involves different subsets of 8 or 9 different cues has many possible combinations (over 100,000 with 8 cues and over 900,000 with 9 cues). The actual set of unique search combinations used is fewer than 1% of these. We use a partial tau τ as a metric for the difference between observed and predicted search orders. This is a generalized version of Kendall's tau metric, and is a standard metric in statistics for the difference between two partially-ordered lists (Fagin, Kumar, Mahdian, Sivakumar, & Vee, 2006). Intuitively, τ is the number of pairwise swaps required to transform one search order into another, allowing for ties. Thus $\tau = 0$ when the observed order exactly matches the predicted order, but increases as the observed order becomes more different. The generalization of τ to the partial version we use allows rankings to have ties. This is important, because whenever a participant terminates search before examining all cues, they produce a partial order in which all of the non-searched cues can be considered as being ranked equal last.

Group-level results

We identified the cue measures, for each learning mechanism separately, that provided the best prediction (i.e. lowest τ) for each individual participant and on each trial. We then calculated $\Delta\tau$, the increase in τ of each model over this minimum τ value. The distribution of $\Delta\tau$ is shown in Figure 2. Blue (darker) lines and markers show the standard learning models, and the red (lighter) markers and lines the biased learning models. Since higher mass of the $\Delta\tau$ distribution closer to zero indicates better model performance, it is clear that validity, discriminability, additive, and success rate models, when combined with biased learning, are far better in predicting cue search orders. A series of paired sample Bayesian t-tests with

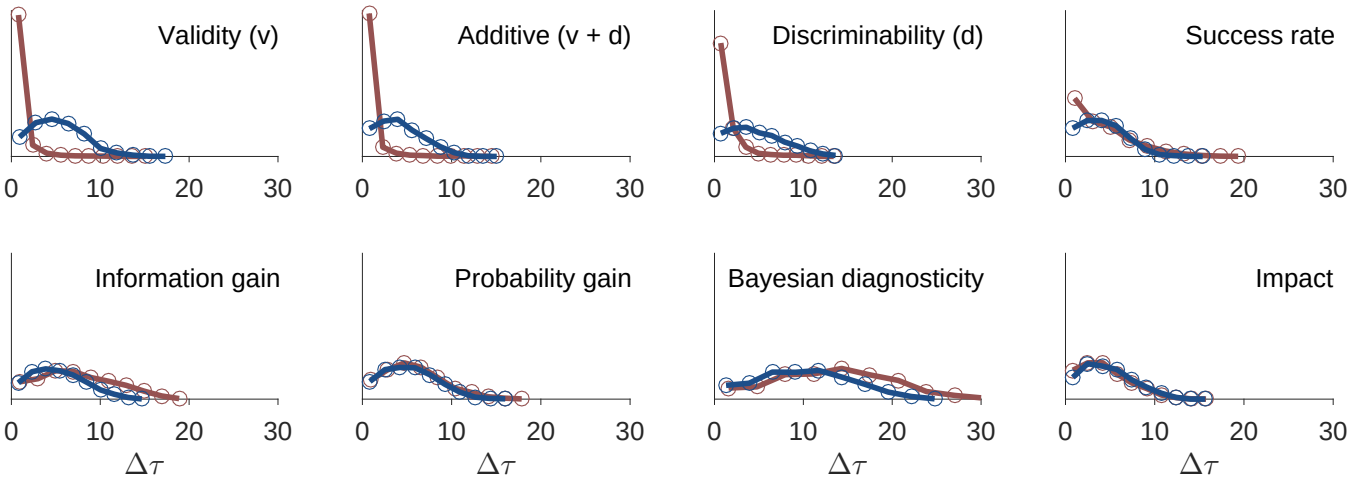


Figure 2: Distribution of $\Delta\tau$ across all trials and participants. Standard (blue) and biased (red) learning shown for each measure

default priors (Love et al., 2015) were carried out to test the one-sided alternative hypothesis that the population mean of partial taus generated by the biased-additive model is lower than for each of the remaining models. The log Bayes factors generated ranged from 6 to 28 for all pairwise comparisons for both USA and Italy datasets, except for the comparison with the biased-validity model, for which the Bayes factors were inconclusive. This provides evidence that the biased-additive and biased-validity models are significantly better than the others models with respect to the partial tau measures. Figure 3 shows τ across trials for the top four measures in the USA condition. The blue (darker) and red (lighter) circles show mean τ for standard and biased learning models, respectively. The error bars show the 95% credible interval across all participants and samples. The black line shows the mean for random sequences. In general, τ reduces across trials, suggesting behavior gradually becomes consistent with systematically predicted cue orders, either because the model becomes more accurate, people become more consistent, or both.

Individual-level results

The error bars in Figure 3 are large because of individual difference between participants. At the individual level, more confident evaluation is possible. Figure 4 demonstrates this, by showing the same analysis for an individual participant. This participant's search orders are best predicted by biased learning, and by the validity, discriminability, or additive combination of these two measures. It is not possible to display the same analysis for all conditions and participants, but the one in Figure 4 was chosen as prototypical. Results for all conditions (including Italian cities) and participants, are available as supplementary material at `www.osf.io/uqf5p`.

Effectiveness of cue learning

The prediction of cue search orders and evaluation of measures of usefulness depends on the learned v and d. Our models predict these for each individual cue at each trial. Figure 5 shows the predicted learning for 3 of the 8 different cues in the USA condition for a single participant. Cue 2 is whether the city has a sports team, cue 3 is whether the city has an airport, and cue 7 is whether the city is a national capital. The large circles show the mean learned values, that is, calculations based on counts of successes and failures, without taking into account uncertainty. The gray dots represent the set of 100 mental samples drawn from learned distribution, and make clear the associated uncertainty, which reduces across trials. For standard learning, the uncertainty reduces earlier for d than v, as expected. This results in greater difficulty in learning v for cues with empirically lower d (e.g., cue 7). The crosses at the top of each plot show the trials on which the cue was searched. As expected, cues accessed frequently show a greater reduction in uncertainty as well as higher accuracy (e.g. compare cue 3 to cue 7). Contingencies for cue 3 are learned accurately, as the cue is repeatedly selected.

A series of paired sample Bayesian t-tests with default priors to test the one-sided alternative hypothesis that the standard deviation of generated samples, across all cues, for the last trial for both the USA and the Italy environments, were lower for discriminability compared to validity revealed log Bayes factors in the range of 6 to 34; and lower for the biased compared to the standard model revealed log Bayes factors in the range of 15 to 26. This provides evidence that the uncertainty over discriminability is significantly lower than that about validity, and the uncertainty for both are significantly lower for the biased compared to the standard model. While the biased model shows lower uncertainty, cues not frequently accessed under this model are not learned accurately. For example, accuracy of learned v for cue 7 is less than cue 2 which in turn, is less than cue 3. Most cue validities and discriminabilities are learned quite effectively and quickly. For example, cue 3 is accurately assessed in the biased model by about the 10th trial.

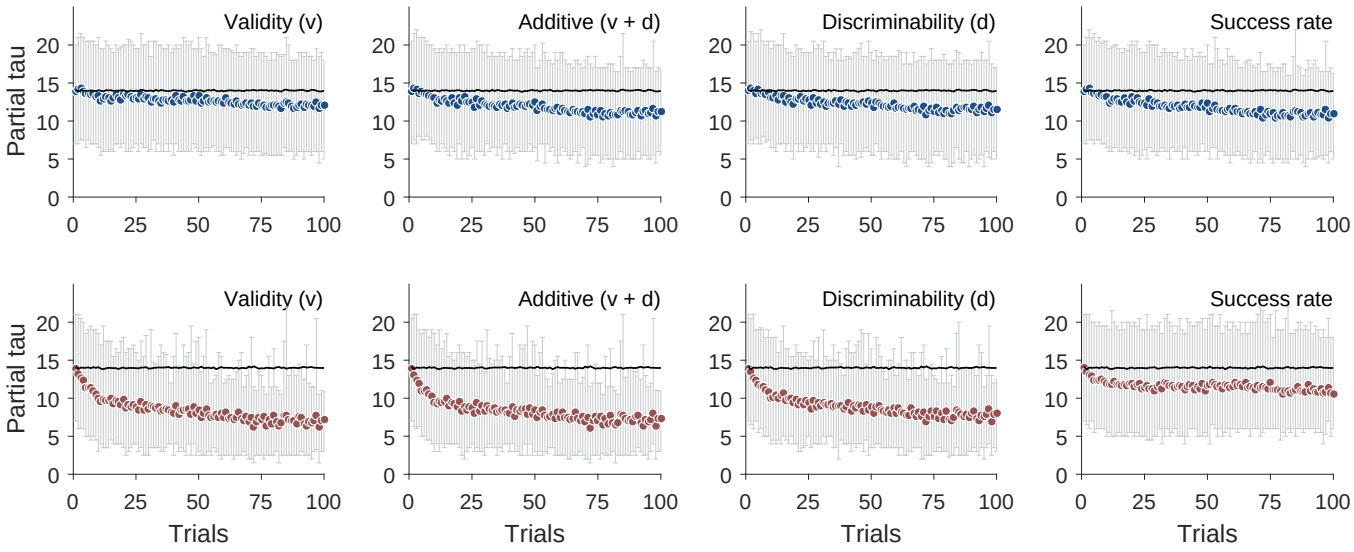


Figure 3: Each panel shows the mean and 95% interval for τ over all participants (USA sub-condition) for different cue measures. Top row: standard learning; Bottom row: biased learning. The black line is mean τ for random orders.

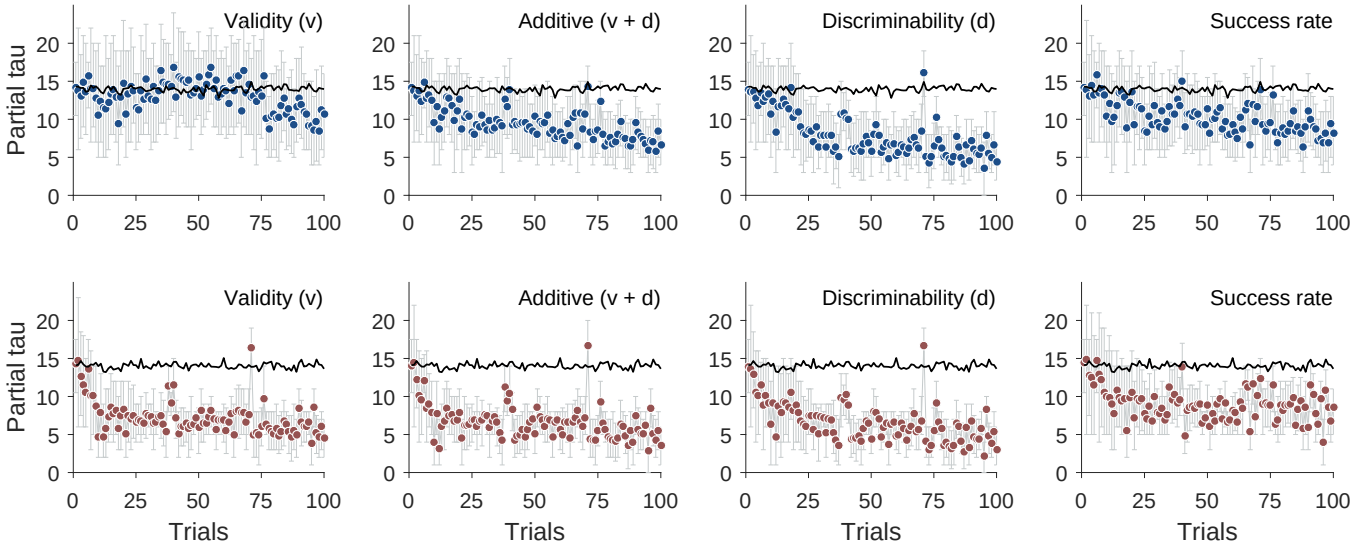


Figure 4: Each panel shows the mean and 95% interval for τ for a single participant (USA sub-condition) for different cue measures. Top row: standard learning; Bottom row: biased learning. The black line is mean τ for random orders.

Discussion

We have shown that simple parameter-free learning mechanisms make reasonable predictions about people's cue search orders. Our two key results are evidence for biased learning, and the demonstration that simple validity and discriminability (or additive combinations of them), make better predictors of cue search orders than more sophisticated measures of cue usefulness. We did, however, find that there were individual differences in use of the various measures. In future work, we propose examining parameterized models—such as a generalization of the additive model into a linear weighted model $wv + (1-w)d$, with w as a free parameter—to capture some of these differences.

Differences may also arise from memory, discounting phenomena, or sensitivity to cost and effort. All of these can be incorporated into extended learning mechanisms. For example, recency effects can be incorporated by using a decay rate for the counts, and cost sensitivity by appropriately weighting the process by which counts are updated. Finally, the biased learning model suggests under-exploration, but this could be on account of strong causal priors that people may have regarding the various cue attributes. Prior causal beliefs generated outside experimental settings can be difficult to measure, although appropriate parameterization of models could be used to infer such prior beliefs and improve the quality of predictions.

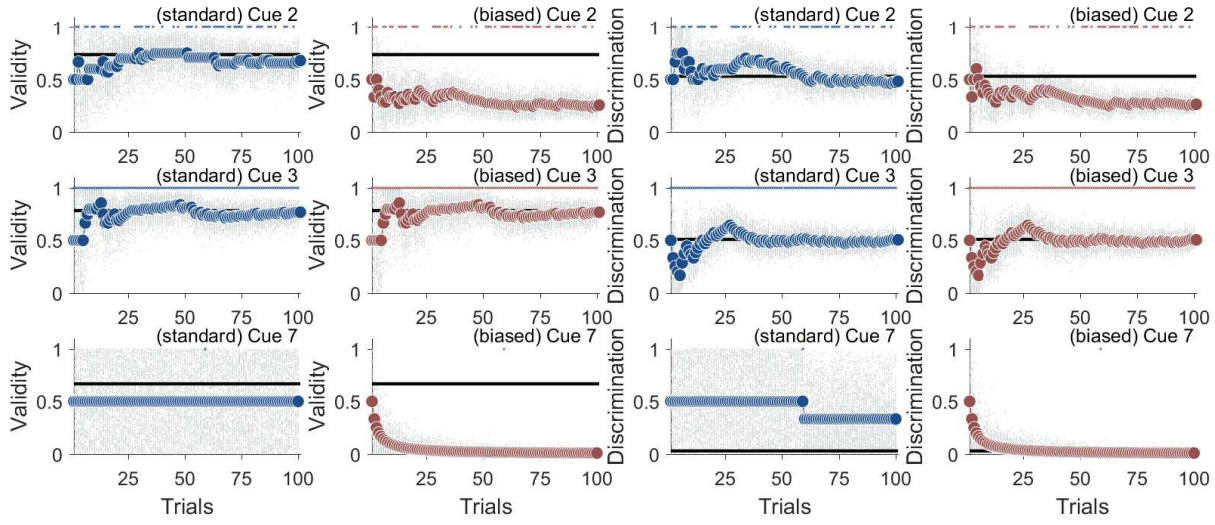


Figure 5: Learning predictions for a single participant, for 3 of the 8 cues in the USA environment. The first two columns show validities, and the last two columns show discriminabilities. Blue circles show standard learning; Red circles show biased learning. The thick black lines are the true values. Gray dots show the mental samples drawn from the learned distributions. The crosses at the top show the trials on which the cue was searched.

References

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *The Quarterly Journal of Experimental Psychology*, 1–25.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2006). Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, *20*, 628–648.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*, 650.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart* (pp. 75–95). Oxford University Press.

Gigerenzer, G., Todd, P. M., & the ABC Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211.

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*, 162.

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making*, *6*, 832–842.

Lee, M. D., Newell, B. R., & Vandekerckhove, J. (2014). Modeling the adaptation of search termination in human decision making. *Decision*, *1*, 223.

Lee, M. D., & Zhang, S. (2012). Evaluating the process coherence of take-the-best in structured environments. *Judgment and Decision Making*, *7*, 360–372.

Le Pelley, M., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 220.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A., & Wagenmakers, E. (2015). *Jasp (version 0.7)[computer software]*. Amsterdam, The Netherlands: JASP Project. Retrieved from https:// jasp-stats. org.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*, 29–71.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision making: The success of "success". *Journal of Behavioral Decision Making*, *17*, 117–137.

Ravenzwaaij, D., Moore, C. P., Lee, M. D., & Newell, B. R. (2014). A hierarchical Bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive Science*, *38*, 1384–1405.

Todd, P. M., & Dieckmann, A. (2004). Heuristics for ordering cue search in decision making. In *Advances in Neural Information Processing Systems* (pp. 1393–1400).

van Ravenzwaaij, D., Newell, B. R., Moore, C. P., & Lee, M. D. (2014). Using recognition in multi–attribute decision environments. *Proceedings of the 35th Annual Conference of the Cognitive Science Society Austin, TX: Cognitive Science Society*, *17*, 3627–3632.

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, *5*.

A Deep Siamese Neural Network Learns the Human-Perceived Similarity Structure of Facial Expressions Without Explicit Categories

Sanjeev Jagannatha Rao (sjrao@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Yufei Wang (yuw176@eng.ucsd.edu)
Department of Electrical and Computer Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Garrison W Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093 USA

Abstract

In previous work, we showed that a simple neurocomputational model The Model, or TM) trained on the Ekman & Friesen Pictures of Facial Affect (POFA) dataset to categorize the images into the six basic expressions can account for wide array of data (albeit from a single study) on facial expression processing. The model demonstrated categorical perception of facial expressions, as well as the so-called facial expression circumplex, a circular configuration based on MDS results that places the categories in the order happy, surprise, fear, sadness, anger and disgust. Somewhat ironically, the circumplex in TM was generated from the similarity between the categorical outputs of the network, i.e., the six numbers representing the probability of the category given the face. Here, we extend this work by 1) using a new dataset, NimsStims, that is much larger than POFA, and is not as tightly controlled for the correct Facial Action Units; 2) using a completely different neural network architecture, a Siamese Neural Network (SNN) that maps two faces through twin networks into a 2D similarity space; and 3) training the network only implicitly, based on a teaching signal that pairs of faces are in either in the same or different categories. Our results show that in this setting, the network learns a representation that is very similar to the original circumplex. Fear and surprise overlap, which is consistent with the inherent confusability between these two facial expressions. Our results suggest that humans evolved in such a way that nearby emotions are represented by similar appearances.

Keywords: facial expressions; similarity structure; deep siamese neural network; multidimensional scaling (MDS); facial expression circumplex

Introduction

According to Darwin, facial expressions of emotion evolved and adapted to prepare the organism to deal with its environment and to also serve to communicate the internal state of the organism (Darwin, 1872; Hess & Thibault, 2009). If facial expressions of emotion are an outward manifestation of an internal state, then similar internal states should lead to similar expressions, in order to make the outward manifestations consistent and easy to understand. At the same time, expression of different emotions should also be sufficiently distinguishable in order to make it possible to properly respond to them.

How are facial expressions represented in the brain? There are two competing theories. One theory is based on experimental evidence of categorical perception of expressions of emotion, suggesting that the representation of facial expressions is divided into discrete categories. Once an expression has been categorized, the subtleties of the expression are lost.

An opposing theory suggests that perception of facial expressions is not as discrete as suggested by data supporting categorical perception. This notion of facial expression perception suggests that while some facial expressions have full membership in one of the six basic emotion classes (happy, disgust, angry, sad, fear, surprise), that nevertheless there is an underlying similarity structure to the expressions. Russell is the strongest advocate of this view, and has presented results that support this notion of perception of facial expressions (Russell & Bullock, 1986; Russell, 1980; Russell, Lewicka, & Niit, 1989). This and other related research suggests that there is a continuous underlying multidimensional perceptual space in which there are clear neighborhood relationships between expression categories, where each expression is closer to some expressions than others.

Dailey et al. (2002) developed a neural network model trained to classify facial expressions into six basic emotions (this model is referred to as "The Model" (TM) in (Cottrell & Hsiao, 2011)). The model was able to fit data usually taken to support each of the two competing theories of facial expression recognition (Young et al., 1997). It displayed categorical perception as well as graded reaction times near category boundaries, and responses indicating that the model was sensitive to mixed-in emotions even on the opposite side of the category boundary.

Dailey et al. performed MDS on the human forced-choice responses published by (Ekman & Friesen, 1976) and on their model's responses to the same stimuli. These are shown in Figures 7 and 8, respectively. They showed that the ordering of emotions is the same in both the cases, a result that is unlikely to have occurred by chance (the probability of this outcome is 1/60, or 0.017). This reflects clear neighborhood relationships between facial expressions.

In this work, we aim to reproduce these results from MDS, albeit under more restrictive training conditions. In particular, the model is only told which faces are in the same category

and which faces are in different categories and is not explicitly given the categories themselves. To the best of our knowledge, this is the first time the circumplex has been shown to arise from facial expression data under such restrictive training conditions.

We design a siamese neural network and train it to learn a 2D representation of facial expressions. The network is trained on pairs of images with a binary label that indicates if the two images belong to the same or different facial expression categories. In essence, this model is not explicitly trained on the number of facial expressions categories in the underlying data, or on the relative relationships among the facial expressions. The low dimensional representation produced by the siamese network replicates to a large extent the circumplex found by Dailey et al. (Dailey, Cottrell, Padgett, & Adolphs, 2002).

Our results suggest that facial expressions of emotion have evolved to make their appearance easily discriminable, and that compatible inner states produce similar expressions. The similarity structure in the low-dimensional space discovered by the network indicates that human expressions of similar emotions are closer to each other when compared to the dissimilar ones. The inherent confusability between our perception of facial expressions is explained by the overlapping clusters in our representations. For example, the siamese network overlaps the Fear and Surprise clusters, which are known to be prone to confusion. The fact that we obtain distinct clusters in our similarity structure demonstrates our ability to express dissimilar emotions in a differentiable way.

Siamese Neural Network Model

Dimensionality Reduction and Siamese Neural Networks

Two classical methods for dimensionality reduction are Principal Component Analysis (PCA) and Multidimensional Scaling (MDS). PCA finds a linear projection of the input data to a low dimensional space that maximizes the explained variance. MDS arranges the data in the low dimensional space in a manner that best preserves the pairwise distances between input points. However, facial expression images pose several challenges, similar to those posed in any computer vision application. Changes in lighting can make images of dissimilar emotions more similar, and similar ones different. In emotion recognition, the identity of the individual is a confound; identity is noise with respect to expression, and vice-versa. This suggests that a nonlinear embedding is required. MDS provides this, but it does not provide a mapping of new data into the same space, so it is difficult to check for generalization.

We require a dimensionality reduction technique that is robust to these changes in input, and that provides a way to generalize to new images in order to check that the embedding is consistent. In this work, we aim to learn the low dimensional structure of facial expressions data without relying on the total number of categories in our data and without associating explicit category labels to each input data point. Siamese neural networks fit these modeling requirements perfectly.

Siamese neural networks are comprised of two neural networks that take a pair of images as input and share a common contrastive loss function. Like siamese twins who share organs, the two networks of a siamese neural network are identical to each other in their architecture, and they share the same weights.

Figure 1 shows the layout of our siamese neural network. It receives a pair of images that are resized to 227 x 227 as input in its first layer. Each of these inputs is then processed through a dedicated 6 layer feed forward network as shown in the Figure. The first three layers are convolutional and the last three layers are fully connected. The activations of the last layer from each network are used to compute the loss.

The loss function is an energy-based one that is designed to move the representations of pairs inputs that are supposed to be "the same" closer together, and ones that are supposed to be different farther apart. We use the loss developed in (Hadsell, Chopra, & LeCun, 2006). Let X_1 and X_2 be two images presented to the system, one to each network. Y is a binary label assigned to the pair, with $Y = 0$ if the images supposed to be similar, and $Y = 1$ if they are supposed to be different. G_1 and G_2 are the activation vectors of the last layer of each network, just before the contrastive loss function in Figure 1. Let $D_w = ||G_1 - G_2||$ be the Euclidean distance between these vectors, where the subscript indicates the dependence on the weights W of the network. Then the loss function is:

$$L = (1-Y)\frac{1}{2}(D_w)^2 + Y(\frac{1}{2}max(0, m - D_W))^2 \qquad (1)$$

where $m > 0$ is a margin. This loss function is inspired by an analogy to springs, where minimizing the first term corresponds to a spring pulling G_1 and G_2 closer together, and the second term corresponds to a repulsing spring, pushing G_1 and G_2 farther apart. This loss function can be optimized by gradient descent. In order to map the faces into a two-dimensional space, G_1 (and hence G_2) are composed of two units.

Siamese networks have been shown to work well in face verification (Chopra, Hadsell, & LeCun, 2005), where the categories are not known in advance, since there are an unbounded number of faces. The networks in this case map faces of the same person to nearby places in the representational space (the last layer), and faces of different people farther apart. Siamese neural networks have also been shown to work well at dimensionality reduction (Hadsell et al., 2006). We take our loss function from the latter publication. Both these models use deep convolutional neural network architectures to extract features from the input images.

Dataset

We use facial images corresponding to the six basic emotions, Happy, Surprise, Fear, Sad, Anger and Disgust from

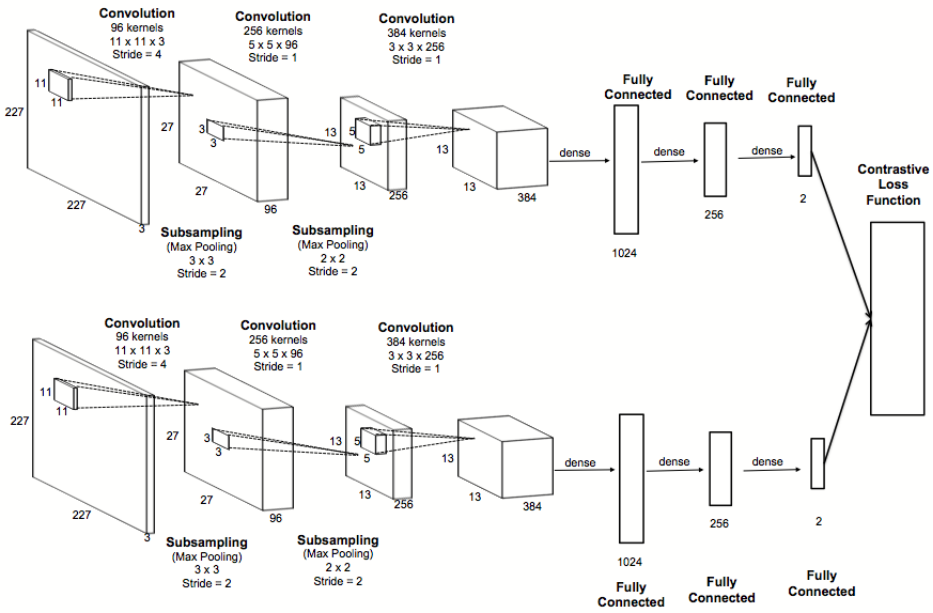


Figure 1: Siamese Neural Network Architecture

the NimStim dataset (Tottenham et al., 2009) for our analysis. We create all possible pairs from the images corresponding to these six basic emotions and use that as input to our siamese network. In all, we train on 126,756 pairs of images. The breakdown of category-wise pair counts is given in Table 1. Originally, we tried to balance the number of similar and dissimilar pairs, however, we ended up losing a significant amount of data and the model did not generalize well to unseen data. Hence we used all of the data, as shown in the table.

Approximately 10% of the subjects in the dataset are set aside for a validation set and 10% for a test set. The remaining 80 percent of the subjects contribute to the training set.



(a) Angry (b) Disgust (c) Fear

(d) Happy (e) Sad (f) Surprise

Figure 2: Sample Images from NimStim

Dealing with Limited Data

Deep convolutional neural networks (CNN) are trained on several hundred thousands of images. A large data set is required to learn the large number of parameters in the network. We are constrained by the relative small size of our dataset. A workaround for a small dataset is to initialize our model with a pre-trained model that will generalize to our problem.

The winning model of the ImageNet LSVRC-2012 contest (Krizhevsky, Sutskever, & Hinton, 2012) (dubbed "Alexnet") broke new ground in CNNs by using a 8 layer deep convolutional neural network. This model was trained to classify natural images into 1000 different categories. This model, along with its weights are publicly available. We use the first three convolutional layers as a starting point to build and train our siamese neural network.

Architecture

We experimented with several architectures, and we report the one that gave the minimum loss on the training set. Though we present only one architecture, all architectures that resulted in a significant reduction in loss during training yielded essentially the same representation. The network contains 6 layers, the first three are convolutional, and the next three are fully connected. Two such networks together form our siamese architecture. The output of the last fully connected layer serves as input to our loss function. We build our first three convolutional layers from the pre-trained weights on the ImageNet LSVRC-2012 dataset. We found this initialization to work really well for our purposes and has helped us cope with our limited dataset.

1. The first convolutional layer filters the $227 \times 227 \times 3$ input

image with 96 kernels of size $11 \times 11 \times 3$ with a stride of 4 pixels.

2. The second convolutional layer takes as input the (response-normalized and pooled) output of the first convolutional layer and filters it with 256 kernels of size $5 \times 5 \times 96$.

3. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$ connected to the (normalized, pooled) outputs of the second convolutional layer.

4. The fourth, fifth, and sixth layers are fully connected with 1024, 256, and 2 units, respectively.

The loss function takes its inputs from the 2 units of the sixth layer from each of the two individual networks. We have chosen to have two units in the last fully connected layer in order to extract a two dimensional representation of our data.

The siamese neural network was developed using the Caffe deep learning framework (Jia et al., 2014). We use ReLU (recitified linear units) activation functions throughout our architecture. We use a base learning rate of 0.0001, a step learning rate policy with a step size of 5000. The margin m in Equation 1 is set to 1. We use dropout after the fourth and the fifth fully connected layers. Training the model on a single GPU took around 5 hours. The layers were initialized with Xavier initialization and stochastic batch gradient descent was used during training.

Training

We started with pre-trained weights on the first three convolutional layers and trained only the subsequent layers. We stopped training when there was no additional improvement in the performance on the validation set. We then fine tuned the first three convolutional layers as well. Fine tuning was done for 5000 epochs at which point the loss did not reduce any further. The representations learned through the process of training are shown in figure 3. We plot the activations in the last layer of the network for each image in the training set to generate these plots.

Evaluation

Figure 3 shows how the different categories are organized by the network during the course of training. Each point in the plots represents one image in the NimStim dataset corresponding to one of the six basic emotions of happy, sad, angry, fearful, surprised and disgusted. At the start of training, the network is unable to differentiate the facial expressions as shown in Figure 3(a). The six basic emotions begin to form clusters around the 4000 epoch mark (Figure 3(d)), and become distinct after 5000 epochs as shown in Figure 4. Until this point the convolutional layers were fixed at their initial values, and at 5000 epochs the loss reached its minimum. At this point, we started fine-tuning the pre-trained layers and there is a further drop in loss. As expected, the representation becomes more distinct after fine tuning as seen in Figure 5.

The model generalizes to unseen data within the NimStim dataset. Its performance on test set, shown in Figure 6, is similar to that on the training set.

Table 1: Image Pairs by Emotion Catogory

Emotion	Angry	Disg.	Fear	Happy	Sad	Surpr.
Angry	3741	7134	6960	11049	7308	3828
Disgust		3321	6560	10414	6888	3608
Fear			3160	10160	6720	3520
Happy				8001	10668	5588
Sad					3486	3696
Surprise						946

The reader should compare the organization of the facial expressions in Figure 5 with that seen in Figures 7 (the MDS of human responses) and 8 (the MDS of The Model's responses). The human MDS is derived from the human subjects' averaged six-alternative forced choice responses for each face in the POFA dataset, as published by (Ekman & Friesen, 1976) and on The Model's responses to the same stimuli as reported by Dailey et al. (2002).

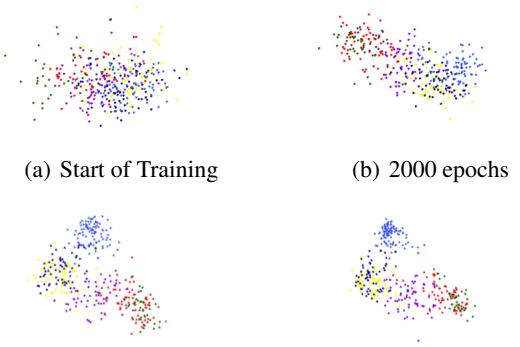
Of particular interest here is the ordering of the representations of facial expressions in two dimensions. The ordering of the emotions in the results reported by Dailey et al. and in the representations produced by the siamese network model are very similar. However, we do not get a perfect circumplex. Surprise and fear images completely overlap in our representation, which is consistent with the inherent confusability between them. Disgust and anger are just barely separated, and these too are expressions that are confused by human subjects.

The resulting order is unlikely to have happened by chance. The probability of a random ordering of six emotions matching the representation in Figure 7 is only 1/60: starting with any emotion, there are 5 to choose from next, 4 after that, etc. This gives 120 possibilities, but whether they are clockwise or counterclockwise does not matter, so there are 60 possible events.

To compute the probability of the current results, we can consider that we have a failure to separate two emotions, so the results are consistent with an ordering of Happy, Surprised, Fearful, Sad, Angry and Disgusted, or Happy, Fearful, Surprised, Sad, Angry and Disgusted. Since each of these ordering have a probability of 1/60, both together have a probability of 1/30, or 0.033.

Dailey et al. (2002) found that happy faces were the easiest to classify fear faces the most difficult to classify, consistent with the literature (Katsikitis, 1997; Ekman & Friesen, 1976; Matsumoto, 1992). The results of the siamese network model are consistent with these patterns. Happy images have been pushed into a tight cluster in the two dimensional representation, such that they are essentially linearly separable from the others, even in this very low dimensional space, and fear is completely overlapping with surprise, making it impossible to separate from the others.

The siamese network model has not been trained to classify the emotions into the six categories used by humans; rather it



(a) Start of Training (b) 2000 epochs

(c) 3000 epochs (d) 4000 epochs

Figure 3: Representations during training. Refer Figure 4 for legend.

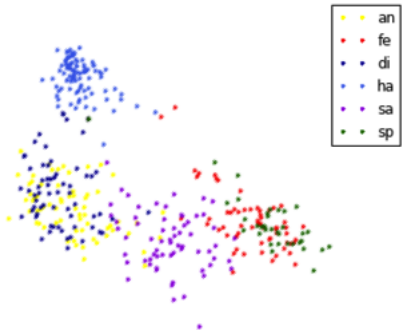


Figure 4: Representation after training for 5,000 epochs without fine tuning. Legends used in Figures 3-6: an: Angry, fe: Fear, di: Disgust, ha: Happy, sa: Sad, sp: Surprise.

has simply been trained on what humans consider "same" or "different" categories. These results, therefore, suggest that the similarity structure learned by the network is inherent in the similarity structure of the faces and the fact that some are different from others. We further hypothesize that, not only have the expressions evolved to be discriminable, but similar emotions have similar expressions.

Conclusions

We have presented a siamese neural network model that derives low dimensional representations of facial expressions under restrictive training conditions.

The network is only given same/different information about the images, it is not given any similarity information, so the structure of the clusters reflects the similarity between the expressions themselves. In Dailey et al., 2002, the circumplex was derived from the softmax output of the network, which also reflects confusability, but here, the network is deriving the similarity structure solely from the images and the information that some are the same, and some are different. It is never told which *categories* are similar to each other, so that is induced by the network from the similarity in the images. If Darwin is correct, and emotional expressions signal

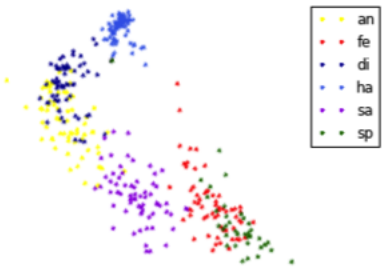


Figure 5: Representation after fine tuning. Refer Figure 4 for legend.

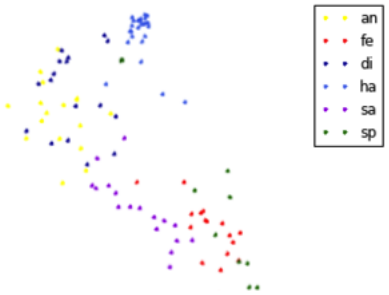


Figure 6: Test set representation. Refer Figure 4 for legend.

internal states, then the model predicts that anger and disgust have similar internal states, and fear and surprise also signal similar internal states.

Our results suggest that, through evolution, our facial expression of emotions and their perception have developed to communicate an inner state that is easily differentiable, and that associated emotional states are communicated similarly. Disgust and anger are often combined in everyday life, and in more exciting, if unfortunate, circumstances, fear and surprise are highly compatible and tend to co-occur. Our network has no access to these notions, no access to culture, yet it places these pairs of emotions either next to each other (as in disgust and anger), or overlapping (as in fear and surprise). The fact that we obtain relatively distinct clusters in our similarity structure suggests that our emotional expressions are inherently differentiable.

Acknowledgments

This work was supported in part by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center. We are also grateful to members of Gary's Unbelievable Research Unit (GURU) for their help.

References

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively with application to face ver-

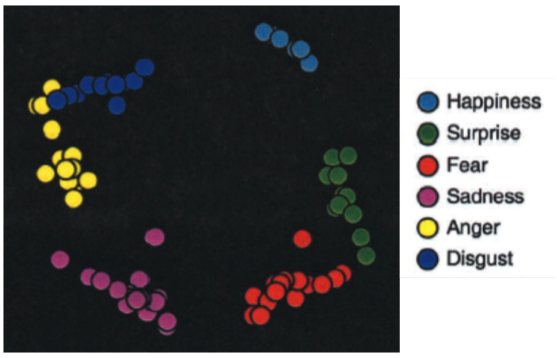


Figure 7: Human MDS representation of facial expressions.

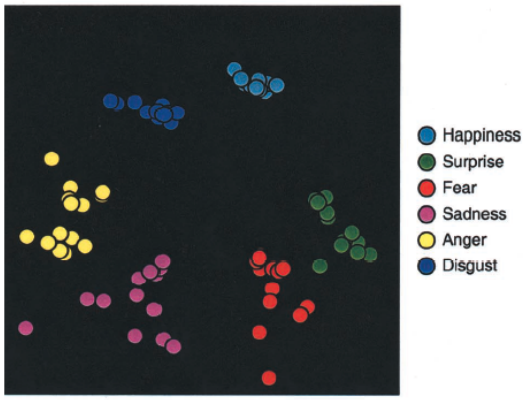


Figure 8: The Model's (Dailey et al., 2002) MDS Representation

ification. *Computer Vision and Pattern Recognition*, 539-546.

Cottrell, G., & Hsiao, J. (2011). Neurocomputational models of face processing. In *The Oxford Handbook of Face Perception.* Oxford, UK: Oxford University Press.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*(8), 1158-1173.

Darwin, C. (1872). *The expression of emotions in man and animals*. New York.

Ekman, P., & Friesen, W. (1976). Pictures of facial affect. *Consulting Psychologist Press*.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Computer Vision and Pattern Recognition*, *2*, 1735-1742.

Hess, U., & Thibault, P. (2009). Darwin and emotion expression. *American Psychologist*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: convolutional architecture for fast feature embedding. *Proceedings of the ACM International Conference on Multimedia*, 675-678.

Katsikitis, M. (1997). The classification of facial expressions of emotion: A multidimensional scaling approach. *Perception*, *26*, 613-626.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

Matsumoto, D. (1992). American–japanese cultural differences in the recognition of universal facial expressions. *Journal of Cross-Cultural Psychology*, *23*, 72-84.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161-1178.

Russell, J., & Bullock, M. (1986). Fuzzy concepts and the perception of emotion in facial expressions. *Social Cognition*, *4*, 309-341.

Russell, J., Lewicka, M., & Niit, T. (1989). A cross-cultural study of circumplex model of affect. *Journal of Personality and Social Psychology*, *57*, 848-856.

Tottenham, N., Tanaka, J., Leon, A., McCarry, T., Nurse, M., Hare, T., ... Nelson, C. (2009, August). *The nimstim face stimulus set.* Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N., Seth, A., & avid I. Perrett, D. (1997, June). Facial expression megamix: tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*, 271–313.

On the adaptive nature of memory-based false belief

Hidehito Honda (hitohonda.02@gmail.com)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902

Toshihiko Matsuka (matsukat@muscat.L.chiba-u.ac.jp)
Department of Cognitive and Information Science, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902

Abstract

Previous studies have shown that people's memories are changeable, and systematic incorrect memories (e.g., false memory) can be created. We hypothesize that people's beliefs about the real world can be changed similarly to the way systematic incorrect memories and systematic incorrect beliefs (which we call memory-based false belief) are generated. We also predict that since memory-based false beliefs are consistent with abstract knowledge that is consisted with prototypical patterns and organization found in the real world, false beliefs work adaptively in making inferences about environmental information in the real world. We conducted behavioral and simulation studies in order to examine our hypotheses on people's beliefs and inferences about the real world. The results showed that participants had systematic false beliefs about cities' attributes (e.g., whether they have a professional baseball team), and that such false beliefs worked adaptively in making inferences about population size.

Keywords: memory-based false belief; inference about real world; ecological rationality

Introduction

One of the most studied topics in human decision making has been how human cognitive characteristics affect adaptive decisions. Many studies have shown that people rely on heuristics, which result in various cognitive biases (Kahneman, Tversky, & Slovic, 1982; Kahneman, 2011). Contrarily, other studies have shown that human cognitive limitations do not necessarily lead to maladaptive decisions because people's cognitive limitations are systematic and people can take advantage of such systematic limitations (Gigerenzer, Todd, & The ABC Research Group, 1999; Todd, Gigerenzer, & The ABC Research Group, 2012). In the present study, we provide new evidence that systematic cognitive limitations work adaptively in making inferences.

One interesting finding on the relationship between cognitive limitations and adaptive inferences is that limitations of human memory do not always result in maladaptive inferences. For example, consider the following question: "Which city has a larger population, Tokyo or Chiba?" Who can make the most accurate inference to answer this question, people who know both Tokyo and Chiba or people who know only Tokyo? Intuitively, the amount of knowledge should correlate with the correctness of inferences. Thus, the more knowledge people have, the more accurate inferences people should make based on their knowledge. Thus, our intuition is that people who know both Tokyo and Chiba can make more accurate inferences than people who know only Tokyo. However, studies have shown that this is not always true: people who know only Tokyo can make more accurate inferences in this situation (e.g., Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002; Schooler & Hertwig, 2005). This superficially counter-intuitive phenomenon occurs because people have systematic cognitive limitations. For example, imagine the limitation in the amount of knowledge such that people do not know about some cities in Japan. Generally, missing knowledge about Japanese cities is systematic. For example, people are more likely to recognize cities that have larger populations than smaller ones. Thus, systematic limitations in the amount of knowledge do not always result in maladaptive inferences about cities' population sizes.

Previous studies have mainly examined how the amount of knowledge affects adaptive inferences. In other words, previous studies have focused on the limitation of quantity in memory. However, people have another limitation in memory; the quality of memory. Studies on human memory have shown that a simple experimental manipulation can make people have false memories (e.g., Roediger, & McDermott, 1995). The formation of false memories is not limited to experimental manipulations, and false memories can be generated in the real world (Loftus, 2005; Roediger & DeSoto, 2016). Based on these considerations, we predict that people have incorrect beliefs about real world information. We refer to such incorrect belief as *memory-based false belief*. Then, how do false beliefs affect inferences? Few previous studies have examined this issue. We conducted behavioral and simulation studies to examine this issue. In the following sections, we first introduce our hypotheses about memory-based false belief. Then, we report our behavioral and simulation studies for examining our hypotheses.

Memory-based false belief

Findings in memory research indicate that people can have false belief about the real world. There are noteworthy findings about false memories. One of the most straightforward manipulations for generating false memories (DRM paradigm, Roediger, & McDermott, 1995) is as follows. (1) Participants hear a list of words such as "table," "sit," "legs," "seat," "desk," "arm," "sofa," and so on. (2) After a short interval, the participants complete a recognition task where they are presented with a list of words. The list comprises "old" words that participants heard in (1) and some "new" words that participants did not hear in (1). The most typical results are that participants answer "old" for semantically associated new words (e.g., "chair") and seldom answer "old" for unrelated new words (e.g., "cat"). Although some factors are involved in the processes of generating a false memory, the findings suggest that false memory is generated with a systematic way (e.g., Roediger, Watson, McDermott, & Gallo, 2001). Specifically, the false recognition for the related new words occurs by means of a strong association between the memory for the old words and the related new words that were not actually presented in (1). According to these findings, we predict that people have false beliefs about the real world and that such false beliefs are systematic like false memory.

People are predicted to form abstract knowledge that is consisted with prototypical patterns and organization found in the real world (Rosch, & Mervis, 1975). Hereafter, we refer to this type of knowledge as the prototypical knowledge. For example, people may have prototypical knowledge about characteristics of big cities, such as "there is a professional football team," "there is an international airport," "it's a state capital," and so on. Imagine that there is no professional football team in city "X" which happens to have a large population. People may have a false belief such that there is a professional football team in city X based on prototypical knowledge. That is, a strong association between professional football teams and big cities may produce a false belief. Hereafter, we call this false belief *false positive belief* (FPB). In contrast, imagine a small city "Y" where there is a professional football team. For city Y, people may have a false belief such that there is no professional football team since the association between professional football teams and small cities is weak. Hereafter, we call this false belief *false negative belief* (FNB). We predict that people have two kinds of systematic false belief, which are generated by the associations between prototypical knowledge and the target object (in this case, a city).

How does systematic false belief affect inferences about the real world? Intuitively, false belief seems to deteriorate inferences about the real world. We predict that false belief deteriorates inferences if it is generated in a non-systematic way. However, the systematic nature of false belief described above may work adaptively in making inferences. When there is a correlation between an attribute of a city (e.g., whether it has a professional football team) and criterion for inferences (e.g., population size), inferences based on the attribute are generally accurate (e.g., when there is a professional football team in city X, but not in city Y, this implies that the population of city X is larger than city Y; e.g., Gigerenzer & Goldstein, 1996). Therefore, if false belief is systematically generated by prototypical knowledge about the real world, the false belief will function adaptively in making inferences.

In sum, our hypotheses about false belief are the followings:

Hypothesis 1: People have false belief consistent with prototypical knowledge. For the attributes associated with big cities, people have more FPB for cities with large populations than those with small populations. In contrast, people have more FNB for cities with small populations than for those with large populations.

Hypothesis 2: Systematic false belief functions adaptively in making inferences, although non-systematic false knowledge deteriorates accuracies of inferences. Thus, inferences about cities' population sizes based on systematic false beliefs (i.e., the belief people tend to possess) are more accurate than those based on non-systematic false beliefs.

Behavioral study: Examination of Hypothesis 1

In order to examine Hypotheses 1, we conducted a behavioral study about people's knowledge about cities, and examined the nature of false beliefs in the real world.

Method

Participants Japanese undergraduates (N = 25) from Aoyamagakuin University participated as part of course work**.**

Tasks, materials, and procedure Participants performed a knowledge task about Japanese cities. In this task, participants were presented with one city name and an attribute such as "professional baseball team." The question is "Is there a professional baseball team in this city?" The participants were asked to answer "yes," "no," or "I don't know" depending on their subjective knowledge.

We selected the top 100 Japanese cities ("shi") based on population size in 2011. The top five cities are "Yokohama," "Osaka," Nagoya," "Sapporo," and "Kobe" (see the Appendix for examples). We asked about five attributes of a city; "professional baseball team," "prefectural capital," "high court," "station of bullet train," and "capital area." We selected these attributes based on the following procedure. First, we conducted a pilot study about the cues used in making inferences about population size. In this study, 37 participants were asked "What is a valid cue when making an inference about which city has a larger population size for the presented two cities." Based upon the answers in this pilot study, we selected the five attributes showing actual validity for inferring population size. We examined the validity of the attributes in making inferences based on Gigerenzer and Goldstein (1999). Here, validity

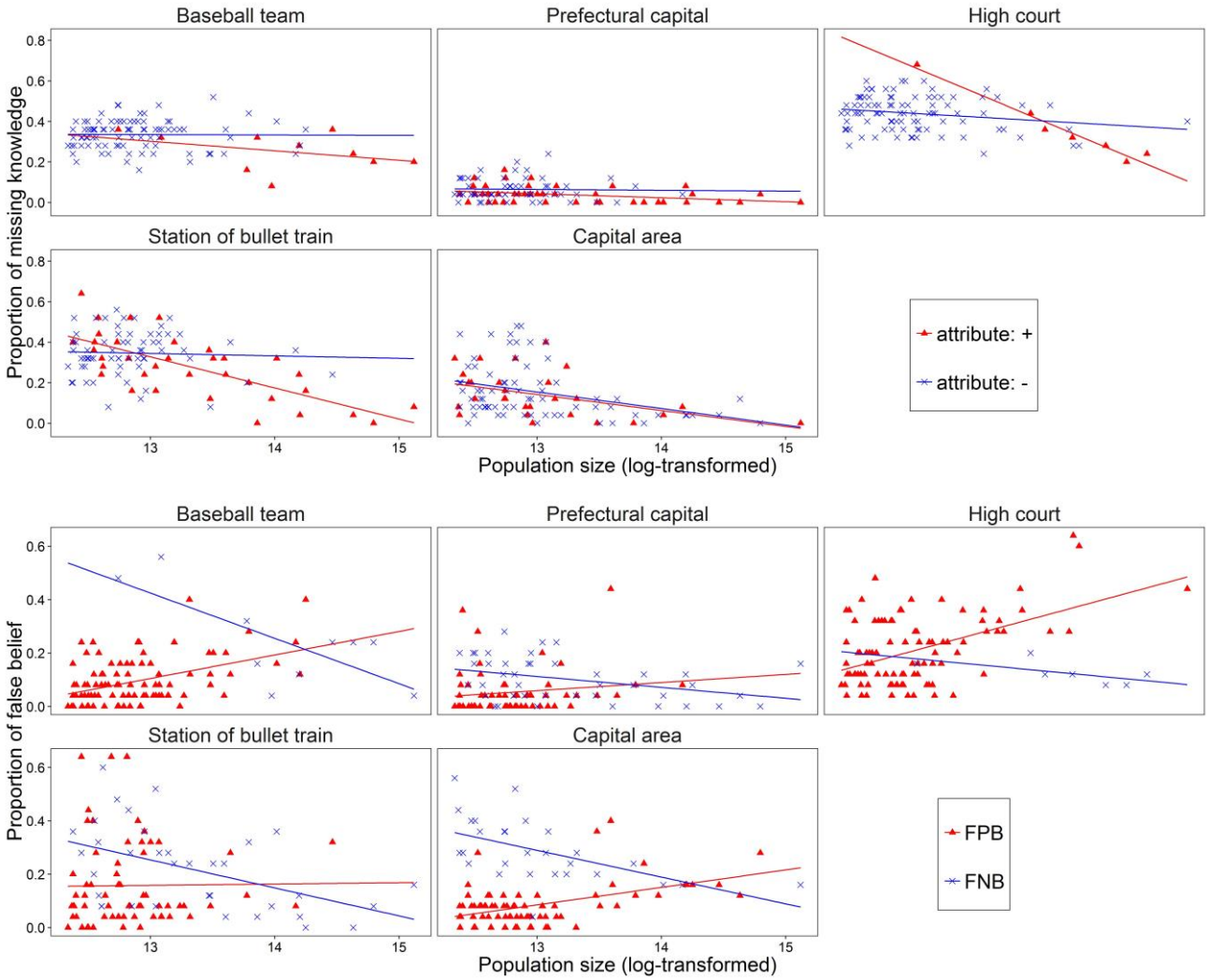


Figure 1. Relationship between population size and proportion of missing knowledge or FPB and FNB for five attributes.

Table 1. Correlation coefficients between proportion of missing knowledge, FPB, or FNB and population size.

Attribute	Missing knowledge				False knowledge			
	Attribute: +		Attribute: -		FPB		FNB	
Baseball team	-0.378		-0.012		0.444	***	-0.733	*
Prefectural capital	-0.345	*	-0.030		0.137		-0.372	*
High court	-0.981	***	-0.210	*	0.482	***	-0.642	
Station of bullet train	-0.718	***	-0.045		0.013		-0.488	**
Capital area	-0.445	*	-0.354	**	0.477	***	-0.354	**

* $p < .05$, ** $p < .01$, *** $p < .001$

reflects how often a cue leads to correct inferences. Using the 4950 pairs (100 * 99 / 2) for the 100 cities used in the behavioral experiment, the validity of the attributes for inferences was calculated by A_c / (A_c + A_i). A_c denotes the number of pairs for which an attribute could discriminate between two cities (e.g., in a pair of cities X-Y, X has the attribute and Y does not) and the use of the attribute information resulted in the correct inference (i.e., X actually has a larger population). A_i denotes the number of pairs for which the attribute could discriminate the two cities but the use of attribute information resulted in an incorrect inference (i.e., Y actually has larger population). The validities for the five attributes were 0.902 (professional baseball team), 0.671 (prefectural capital), 0.923 (high court), 0.710 (station of bullet train), and 0.503 (capital area). Thus, although people tend to believe that all these attributes are valid cues for making inferences, their actual validities vary from high to low.

Participants were tested individually using a computer. They were asked to answer the questions for all 500 attributes (5 attributes * 100 city names).

Results and discussion

We analyzed participants' subjective knowledge about the 100 cities from the relationship between the cities' population sizes and (1) missing knowledge (i.e., the response of "I don't know"), or (2) the FPB and FNB.

First, we examined the relationship between missing knowledge and population size, although this analysis was not our focus. Previous studies have shown that a city's population size is correlated with the number of times the city is referenced in media such as newspapers, and that the number of the city's references is correlated with recognition rate and familiarity (Goldstein & Gigerenzer, 2002; Honda, Matsuka, & Ueda, accepted). Thus, we predicted that the proportion of missing knowledge decreases as population sizes increase. In order to examine this prediction, we calculated the proportion of missing knowledge for every city and attribute, and conducted correlation analyses for every attribute. In particular, we calculated correlation coefficients between proportions of missing knowledge and log-transformed population sizes. Since an attribute is categorized into "+" (the city actually has the attribute) or "-" (the city does not actually have the attribute) and missing knowledge may differ between these attributes, we conducted a correlation analysis for each attribute. The upper panel of Figure 1 denotes the relationships between log-transformed population size and proportion of missing knowledge, and the left panel of Table 1 shows the correlation coefficients. Although the results depend on the attributes, there was a general relationship between the proportion of missing knowledge and population size: As population size becomes larger, the proportion of missing knowledge decreases. This result supports our prediction.

Next, we analyzed the relationship between FPB (or FNB) and population size in order to evaluate Hypotheses 1. As in the analysis of missing knowledge, we calculated the proportions of FPB and FNB for every city and attribute, and conducted correlation analyses for every attribute. In particular, we calculated correlation coefficients between proportions of FPB (or FNB) and log-transformed population sizes. The lower panel of Figure 1 denotes the relationship between log-transformed population size and proportions of FPB or FNB, and the right panel of Table 1 shows the correlation coefficients. Although the results varied among the attributes, as was also the case for missing knowledge, we found apparent tendencies. The proportion of FPB increases as the population size increases. In contrast, the proportion of FNB decreases as the population size increases. These results corroborated our Hypothesis 1. Figure 2 denotes the individual data on false beliefs and missing knowledge (proportion out of 500 attributions). We found that there were large individual differences.

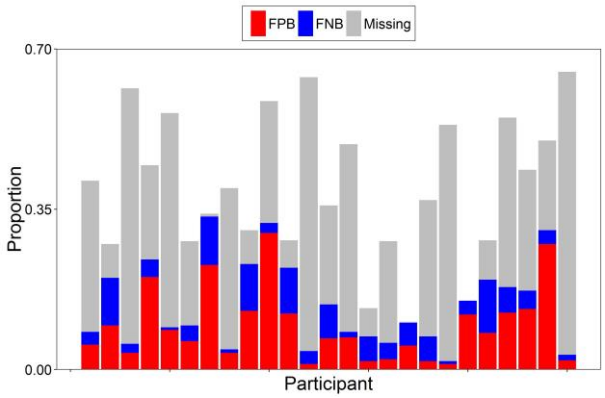


Figure 2. Individual differences in false belief and missing knowledge.

Taken together, people's knowledge about real world cities is correlated with their population sizes. The proportion of missing knowledge decreases as the population size increases. For false beliefs, the proportion of FPB (FNB) increases (decreases) as the population size increases. These results are consistent with Hypothesis 1.

Simulation study: Examination of Hypothesis 2

The results of our behavioral study showed that people have systematic false beliefs, which is consistent with Hypothesis 1. Hypothesis 2 predicts that such systematic false beliefs will function adaptively in making inferences in the real world. We examined this hypothesis by using computer simulations.

Method

We conducted computer simulations of binary choice population inference tasks based on the empirical data collected in our behavioral study. There were a total of 50 simulated participants. Among the 50 simulated participants, half of Systematic condition (hereafter Syst condition; Syst1, Syst2, …, and Syst25) was designed to possess exactly the same knowledge as the 25 participants in the behavioral study. That is, for example, simulated participant Syst1's knowledge about the five attributes for 100 cities was identical to that of empirical participant 1. The remaining 25 simulated participants in Non-Systematic condition (hereafter, Non-Syst condition; Non-Syst1, Non-Syst2, …, and Non-Syst25) possessed the same amounts of FPB, FNB, and missing knowledge as the empirical participants but the patterns were randomly reconstructed. Thus simulated participants in the Syst condition were assumed to possess systematic false beliefs while those in the Non-Syst condition were assumed to possess non-systematic false beliefs. The simulated participants were presented with two cities and asked to infer which city has a larger population. They answered all of the 4950 pairs (100 * 99 / 2) of the cities used in the behavioral study. They were set up to make inferences based on their knowledge. In the present simulation, we compared three inference strategies in order to examine the efficacy of systematic false beliefs regardless of what strategy was used (specific algorithms of inference strategies are shown in Table 2). Among the three strategies, two were knowledge-integration strategies and the third was a heuristic-based strategy. For the heuristic strategy, we

Table 2. Simulated inference strategies.

Strategy	Content of inference: Which city has a larger population, city A or B?
Take-the-best	Consider attributes in the order of their validities. Participant makes an inference based on the first attribute where one city has a positive value and the other has unknown or negative value.
Tally 1	Add up the number of positive cue values and subtract the number of negative values for Cities A and B. A participant makes an inference such that the city with the higher summation has a larger population.
Tally 2	Add up the number of positive cue values for Cities A and B. A participant makes an inference that the city with the higher summation has a larger population.

Note. Here, "positive (negative) value" means that a participant thinks a city has (does not have) the attribute. "Unknown" means that the participant does not have knowledge about the attribute (i.e., missing knowledge).

used "Take-the-best" (Gigerenzer & Goldstein, 1996). As integration models, we used Tally 1 and 2 (Marewski, & Schooler, 2011).

The correctness of an inference was defined as follows. We assigned 1 (or 0) to a pair where an inference strategy led to a correct (or incorrect) inference. When an inference strategy could not discriminate a pair, 0.5 was assigned to the pair. We calculated the mean correctness of inferences for 4950 pairs and regarded this value as the proportion of correct inferences.

For the simulated participants in the Syst condition, we calculated the proportion of correct inferences for each of the three inference strategies. For the Non-Syst condition, we made 100 sets of randomly reconstructed knowledge for every 25 participant, and we calculated the proportions of correct inferences using the three inference strategies for every 100 sets. We regarded the average proportion of correct inferences for the 100 sets as the proportion of correct inferences for each strategy.

Results and discussion

Figure 3 shows the performance on the inference task. As is apparent, the proportion of correct inferences was higher for the simulated participants in the Syst condition than for those in the Non-Syst condition for all of the three inference strategies. Since the amounts of missing knowledge, FPB, and FNB for Syst1, Syst2, …, and Syst25 corresponded to those for Non-Syst1, Non-Syst2, …, and Non-Syst25, we compared the proportion of correct inferences for each pair. Using the Take-the-best and Tally 2 strategies, 24 of the 25 participants in the Syst condition showed higher performance. Using the Tally 1 strategy, 23 of the 25 participants in the Syst condition did better. These results indicate that the systematic nature of human false beliefs about cities functioned adaptively in making inferences about cities' population sizes regardless of inference strategies. Furthermore, given that there were large individual differences in false beliefs and missing knowledge (see Figure 2), the effect of memory-based false belief on inferences is robust regardless of individual differences.

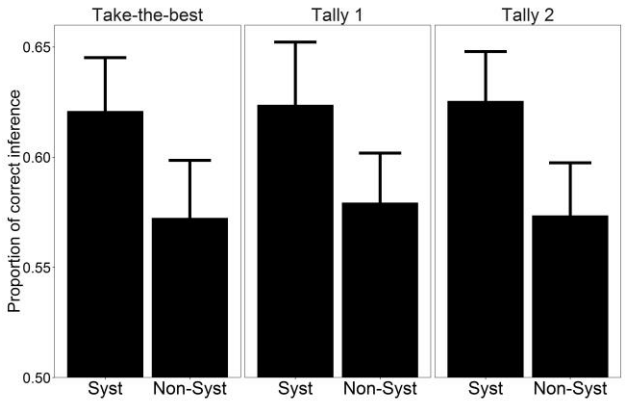


Figure 3. Proportion of correct inferences in the simulation task. Error bars denote standard deviation.

Taken together, the results of computer simulations for binary choice inference task show that the systematic nature of memory-based false beliefs functioned adaptively in making inferences, corroborating Hypothesis 2.

General discussion

In the present study, we examined the nature of people's false beliefs about the real world. The results of a behavioral study showed that participants had systematic false beliefs. In particular, the nature of their false beliefs correlated with environmental structure (i.e., cities' population sizes). For large cities, people tend to falsely recognize that there is an attribute which is associated with large cities (e.g., there is a professional baseball team). In contrast, for small cities, people tend to falsely recognize that there is no such an attribute. We also conducted a simulation study and examined how the memory-based false belief affected inferences about the real world. We found that the systematic nature of false beliefs functioned adaptively in making inferences.

Previous studies have discussed how limitations in the amount of knowledge affect adaptive inferences. For example, Gigerenzer and Goldstein (1996) examined two forms of limited knowledge. One was inability to recognize objects and the other was inability to recognize attributes of recognized objects. Schooler and Hertwig (2005) examined how forgetting aids inferences based on the recognition heuristic. These studies both showed that a limited amount of knowledge (i.e., the number of recognized objects or cues) does not necessarily lead to maladaptive inferences. Rather, a limited amount of knowledge can enhance adaptive inferences. In the present study, we examined the effect of limitations of memory in terms of systematic incorrectness. We found that the systematic nature of memory-based false belief can enhance accuracies of

inferences. This provides new evidence about the relationship between cognitive limitations and adaptive inferences.

Pleskac (2007) theoretically examined the recognition heuristic in terms of signal detection theory. He showed that false alarms and misses in recognition processes affected the performance of the recognition heuristic. Thus, he examined the relationship between accuracy of recognition and adaptive inferences. Here, we note two differences between Pleskac (2007) and our present study. First, Pleskac focused on recognition of objects and examined how false recognition affected the usage of the recognition heuristic. The present study focused on the nature of memory-based false belief (i.e., cities' attributes) and how false belief affected knowledge-based inferences. Hence, Pleskac and the present study examined basically different domains. Second and more importantly, we conducted not only a theoretical study (i.e., computer simulation), but also a behavioral study. In particular, we clarified the nature of false beliefs about the real world. We provided evidence that people had systematic false beliefs about the real world using a behavioral study, and showed how such memory-based false beliefs affected inferences by using computer simulations. Therefore, although our study and Pleskac were analogous in that they both examined how accuracy of memory affected adaptive inferences, our findings provide new insights about the relationship between limitations of memory and adaptive inferences.

In sum, we provided new evidence that limitations in human memory can enhance adaptive inferences. We believe that the present findings make a substantial contribution toward understanding the relationship between adaptive inferences and cognitive limitations.

Acknowledgments

This research was supported by Grants-in-Aid for Scientific Research (A, 16H01725; B, 25280049; C, 25330167) and Young Scientists (B, 16K16070).

References

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take-the-best heuristic. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 75–95). New York: Oxford University Press.

Gigerenzer, G., Todd, P.M., & The ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75–90.

Honda, H., Matsuka, T., & Ueda, K. (accepted). Memory-based simple heuristics as attribute substitution: Competitive tests of binary choice inference models. *Cognitive Science*.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan Magazines Ltd.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361–366.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*, 393–437.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin and Review, 14*, 379–391.

Roediger, H. L., & DeSoto, K. A. (2016). Recognizing the presidents: Was Alexander Hamilton president? *Psychological Science, 27*, 644–650.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Roediger, H., Watson, J., McDermott, K., & Gallo, D. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610–628.

Todd, P. M., Gigerenzer, G., & The ABC Group. (2012). *Ecological rationality: Intelligence in the world.* New York: Oxford University Press.

Appendix. List of city names and attributes used in the present study (Four examples).
"+ (-)" indicates that there is (is not) the attribute in that city.

	City name (shi)	Population size	Baseball team	Prefectural capital	High court	Station of bullet train	Capital area
1	Yokohama	3,689,603	+	+	-	+	+
2	Osaka	2,666,371	+	+	+	+	-
…	…	…	…	…	…	…	…
99	Chigasaki	235,140	-	-	-	-	+
100	Yamato	228,180	-	-	-	-	+

Causal Action: A Fundamental Constraint on Perception of Bodily Movements

Yujia Peng[1] yjpeng@ucla.edu
Steven Thurman[1] sthurmanucla@gmail.com
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology
[2] Departments of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Human actions are more than mere body movements. In contrast to other dynamic events in the natural world, human actions involve mental processes that enable willful bodily movements. We reported two experiments to demonstrate that human observers spontaneously assign the role of cause to relative limb movements, and the role of effect to body motion (i.e., the position changes of the body center of mass) when observing actions of others. Experiment 1 showed that this causal action constraint impacts people's impression on the naturalness of observed actions. Experiment 2a/b revealed that the causal constraint guides the integration of different motion cues within a relational schema. We developed an ideal observer model to rule out the possibility that these effects resulted from the learning of statistical regularity in action stimuli. These findings demonstrate that causal relations concerning bodily movements play an important role in perceiving and understanding actions.

Keywords: causation; causal asymmetry; biological motion; limb movement; body motion

Introduction

Human actions are more than mere body movements. In his *Philosophical Investigations*, Ludwig Wittgenstein (1953, p. 161) posed a famous question: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?" This question highlights the special status of the actions of a human being relative to other dynamic events in the natural world—human actions involve mental processes that cause willful bodily movements. For example, we see rhythmic stepping forward on the ground as causing the forward motion of the agent's body; we see lifting and quickly swinging arms as causing a basketball to launch by throwing it. These types of interaction between active body movements and the distal world gives us direct experience of cause-effect relations. Thus "to step up by lifting one's leg" is properly considered to include the objective content of an action: it involves a causal structure in which moving one's limbs in a certain way provides a means to cause position changes of the body such that the intentional goal of the action is fulfilled.

We often have a strong sense of causality as actions unfold. Observing certain limb movements triggers the expectation of changes in body position. The causal link between the two types of motion is manifested in most observed human actions in our daily life (Thurman & Lu, 2014), yet in some situations this relation can be violated. A striking example is the renowned "moonwalk" dance move, popularized by Michael Jackson decades ago, in which the dancer appear to be making the physical movement of walking forwards, but actually his body moves backwards. We hypothesize that the dance movement's impact on an observer reflects the surprise triggered by its violation of a fundamental expectation about causal actions—the assignment of the role of cause to relative limb movements, and the role of effect to body motion (i.e., the position changes of the body center of mass). Although this proposal appears intuitive, no direct evidence has established that this causal constraint on bodily movements influences perception of human actions (Thurman & Lu, 2013). It is by no means obvious that human observers are sensitive to the physical causal mechanisms that govern the actions of others. Demonstrating such sensitivity to human actions is challenging because in human actions, intention, causality and perceived body movements are entangled. To resolve this difficulty, the present study experimentally separated two types of motion cues involved in actions: relative limb movements (with reference to body-centered coordinates) and the position changes of the body center of mass (with reference to distal world coordinates). This separation makes it possible to determine whether people consider the systematic relations between the two types of motion cues to be causal, or merely an associative correlation.

The present study was inspired by a ubiquitous feature of causation highlighted by Hume (1739/1888): the temporal priority of a cause to its effect. The constraint that the causal relation is asymmetric, such that effects never (or almost never) occur before their causes, is considered a necessary condition to be explained by any adequate theory of causation (Price, 1992; White, 2006). In the context of action observation, if humans assign the cause role to limb movements, and the effect role to the position change of the body (i.e., body motion), we would expect that the relative temporal relationship between limb movements and body motion will be an important determinant of action perception. In the present paper, we report three experiments to investigate the role of causation in action perception.

Experiment 1

In accord with the general causal asymmetry based on temporal priority, we hypothesized that introducing a

temporal lag of limb movements relative to changes in body position (i.e., a situation in which the expected causal cue occurs after its expected effect) would signal a strong violation of the anticipated causal relation between the two motion cues. Observers are therefore likely to detect the inconsistency between the two motion cues and to judge the observed action as unnatural or artificial. In contrast, when limb movements (causes) are displaced forward in time so as to occur moderately early relative to body motion (effect), observers may show more tolerance to such misalignment (since the temporal relation between relative limb movements and body movements will still be qualitatively consistent with normal causal directionality). To examine the role of causal relation between relative limb movements and common body motion in perceiving the validity of observed actions, Experiment 1 was designed to measure how temporal offsets between limb movements and body motion impact the perceived naturalness of actions.

Methods

Participants

One hundred online participants were recruited through Amazon's Mechanical Turk.

Stimuli

Action stimuli were generated from the CMU motion-capture database (http://mocap.cs.cmu.edu) and processed using the Biological Motion Toolbox developed by van Boxtel and Lu (2013). We selected actions in which a person walked on an uneven surface with steps, with both horizontal and vertical body motion included in the sequence.

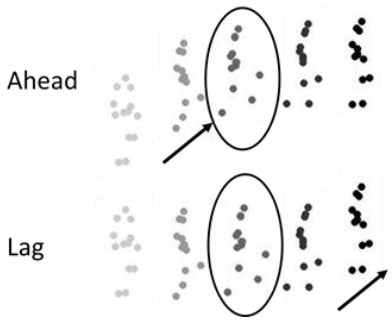


Figure 1: An illustration of different temporal relationships between body motion and limb movements.

The relationship between limb movements and body motion was manipulated by shifting the temporal sequence of common body motions forward or backward in time relative to the sequence of limb movements, as illustrated in Figure 1. Common body motions were made to either precede or lag behind the posture change resulting from the relative limb movements. The temporal sequence of body positions was shifted forward relative to limb movements in the "ahead" condition (i.e., effect precedes), and were shifted backward in time in the "lag" condition (i.e., cause precedes).

Procedure

Participants were first presented with a cover story as following: "Imagine you are viewing a walking sequence on the uneven surface with invisible steps through a slowly rotating camera. The rotation of the camera will help you perceive the 3D space. Look at the relative limb movements of the actor. Look at how the body position changes over time. Ask yourself if that could be a real person's motion in the environment. " Participants were asked to rate the naturalness of videos on a 1 (unnatural) to 5 (natural) scale.

On each trial, an action sequence was presented with a point-light walker on a checkerboard surface with invisible steps was presented. The viewpoint rotated clockwise with a speed of 3 degrees/second, aiming to facilitate a 3D perception of the biological motion stimulus in the environment. Each video lasted about 8 seconds and was played automatically with the start of each trial.

Six temporal offsets between limb movements and body motion were used: 0, ±0.5, ± 1 and 8.33 s. The 0 and very large offset (8.33 s) conditions served as extreme cases to help participants anchor the two ends of the rating scale.

The experiment consisted of 24 experimental trials, plus two attention check trials randomly placed in the experiment. An attention check trial involved a simple task, in which participants were presented with either a walking or jumping sequence, and were asked to identify the presented action. The purpose of including these two trials was to identify outlier participants who gave random responses in the online experiment. The total running time of the experiment was less than 10 minutes.

Results

Given that the participants were recruited online, we included three criteria to remove obvious outliers, including the failure of passing the two attention-check trials, or the average naturalness rating in the zero-offset condition (i.e., action sequence generated from the raw motion capture data) was lower than 2 standard deviations below the mean, or they gave the same naturalness rating value for all trials in the experiment. Nine out of 100 participants were removed according to the exclusion criteria.

The average naturalness rating for the zero-offset condition (i.e., perfect synchrony between limb movements and body motion) was the highest (M = 3.88, SD = 0.74); and the naturalness rating for the extreme offset condition of 8.3 seconds was the lowest (M = 2.28, SD = 1.02), suggesting that human observers utilized information about the magnitude of temporal offsets between limb movements and body position changes in their naturalness judgments.

To examine how temporal offsets between the two movement cues influenced the naturalness ratings, we conducted a repeated-measures ANOVA with two within-subject factors, temporal offset magnitude (.5 vs. 1 s) and offset direction (ahead vs. lag). As shown in Figure 3, the results revealed a significant main effect of temporal offset direction, $F(1,90) = 8.66$, $p = 0.004$, $\eta_p^2 = 0.83$. Observers judged actions to be more natural if the effect of body motion lagged behind the causal limb movements, relative to the corresponding ahead condition. In other words, when the temporal displacement was qualitatively consistent with the

causal direction (i.e., the causal limb movements preceded the effect of body motion), people showed more tolerance of temporal displacement than in the corresponding ahead condition (i.e., when the temporal offset was opposite to the causal direction).

We also found a significant main effect of offset magnitude, $F(1,90) = 29.13$, $p < 0.001$, $\eta_p^2 = 1.0$, indicating that people were sensitive to the temporal alignment between the two motion cues when assessing the validity of observed actions. In general, larger offsets resulted in lower ratings of naturalness. The two-way interaction between offset magnitude and temporal direction was not significant, $F(1,90) = 0.15$, $p = 0.70$.

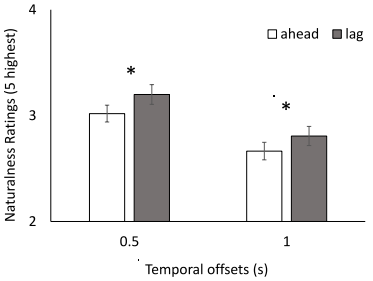


Figure 2: Naturalness ratings in the ahead condition (effect of body motion preceded) and in the lag condition (causal limb movements preceded). Error bars represent standard errors of the mean.

Experiment 2a

In Experiment 2, we aimed to create a reasoning task in which the two movement cues were represented by distinct visual entities in the display. This new reasoning task serves to assess whether people use the default causal relation to form an explicit binding between the two types of movements.

Methods

Participants

Twenty-two UCLA undergraduate students (mean age = 20.8; 16 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli

Four action sequences of an actor walking on an uneven surface were displayed from two viewing directions with orthogonal projection. The actor was made to appear to walk on a treadmill by maintaining a stationary position for the average location of two hip joints at the center of the screen. To depict the position change of the body over time, a gray dot moved according to the trajectory of body motion. Figure 3 provides an illustration of the stimuli.

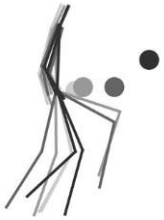


Figure 3: Illustration of stimulus for Experiments 2a and 2b. The stick-figure actor shows the posture change over time at a stationary location; the dot depicts the motion of the body. The figure and the dot change from a light color to a dark color in proportion to the elapsed time.

Procedure

Participants were first given a cover story along with illustration figures. "Imagine that you work for a specialized video analysis company and are given two sources of information: a processed video from a motion tracking system, which records a person's posture change over time and keeps the figure always at the center, and the location of the person reported from a GPS system." After the cover story, participants were presented with two video clips which illustrated how the posture change in a stationary location was separated from the position change of the body over time based on the original motion capture video. Note that our display explicitly isolated the effect cue, body motion (represented by the gray GPS dot) and the causal cue, limb movements (represented by the red figure). Participants were asked to decide whether the movements of the GPS dot matched the posture changes of the actor by pressing one of the two response buttons.

After receiving the cover story, participants were given two practice blocks with feedback to familiarize them with the task. Practice trials included stimuli with either perfectly aligned movements or with excessive temporal displacements (temporal offset of 8.33 s).

In the subsequent test session, 96 trials were presented to participants. The experiment included eight levels of temporal offsets between the position change of the body resulting from body motion and the posture change resulting from limb movements (±0.02s, ±0.5s, ±1s and ±1.5s, respectively). On each trial, the action stimuli lasted for 6.67 s. The first 100 frames (i.e., 1.67 s) presented only the walker, with the goal of encouraging participants to maintain fixation on the walking action. Then the GPS dot appeared at the center of screen in the 101th frame, and subsequently started to move according to the assigned trajectory of body motion. The test block included four viewpoints (45°, 135°, 225° and 315°), and each viewpoint was tested three times under each of the eight offsets. The order of conditions was randomized.

Results

The results of Experiment 2a are shown in Figure 4. A repeated-measures ANOVA with two within-subject factors (ahead vs. lag condition, and four temporal offset levels) revealed a significant main effect of offset magnitude, $F(3,19) = 37.83$, $p < 0.001$, $\eta_p^2 = 1.0$, indicating that participants were sensitive to the temporal offsets between the two motion cues in this explicit binding task. Importantly, the interaction of offset magnitude and temporal direction of

offset was significant, $F(3,19) = 4.51$, $p = 0.015$, $\eta_p^2 = 0.80$, indicating that the influence of causal direction of the two movement cues on the binding judgment depended on the magnitude of temporal displacement introduced into the stimuli. Specifically, when the temporal offset was very large (e.g., 1.5 second), observers presumably considered the posture change of the walker and the GPS dot movements to be generated from different sources, and therefore judged them as mismatched signals, regardless of the temporal direction of offset. Similarly, when the temporal offset was very small (e.g., .02 second), observers may not have detected the difference in temporal directions, and thus also did not show a temporal asymmetry effect. Hence, the critical testing offset conditions were the middle range and we did planned comparisons for ±.5 and ±1 second conditions respectively. Indeed, we found that with a 1 s offset, the lag condition (M = 0.44, SD = 0.29) yielded a significantly higher proportion of matched responses than did the ahead condition (M = 0.29, SD = 0.25), $F(1,21) = 9.69$, $p = 0.005$, $\eta_p^2 = 0.84$. Thus a causal asymmetry effect was observed within a middle range of the temporal window, when the two motion cues could be interpreted as originating from a single actor.

In summary, Experiment 2a used an explicit binding task to provide converging evidence that observers are sensitive to the temporal relation between limb movements and body motion. The two motion cues were more likely to be judged as matched when the causal limb movement preceded the effect of body motion, in comparison to when the effect cue preceded the cause. This temporal asymmetry effect supports the hypothesis that human observers naturally assign the role of cause to relative limb movements and the role of effect to body motion (i.e., they expect limb movements to cause changes of body position in the environment).

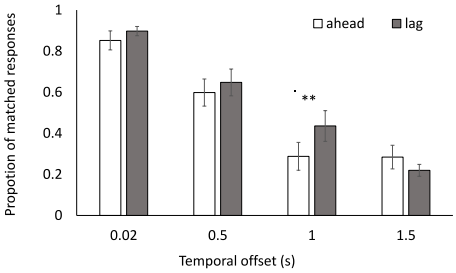


Figure 4: Results of Experiment 2a. The proportion of "matched" responses in the ahead condition (GPS dot shifted ahead of limb movements) and in the lag condition (GPS dot lagged behind of limb movements).

Experiment 2b

If the temporal asymmetry effect found in previous experiments was resulted from the observers' understanding of the causal relation between the two motion sources, then the effect should be radically altered when the causal relation is changed. Experiment 2b aimed to measure differences in people's pattern of judgments attributable to the influence of the causal interpretations conveyed by the cover story.

Participants

21 UCLA undergraduate students (mean age = 20.2; 19 female) participated in the experiment for course credit.

Stimuli and Procedure

We used stimuli identical to those of Experiment 2a, but changed the cover story so as to reverse people's understanding of the causal relation between posture change and dot motion, by specifying that the dot represented a moving laser spot that the person aimed to follow. In this situation, the two components of the stimuli are interpreted as representing two distinct entities, such that the laser spot is the cause that makes a person move in certain ways (i.e., the limb movements are the effect). Participants were instructed to judge whether the person succeeded in following the laser spot. All other procedural aspects were identical to those in Experiment 2a.

Results

The results of Experiment 2b are shown in Figure 5. A repeated-measures ANOVA revealed that the interaction of offset magnitude and temporal direction of offset was significant, $F(3,18) = 38.54$, $p < 0.001$, $\eta_p^2 = 1.0$. When the dot motion preceded body movements in the ahead condition, participants judged the actor to be successful in following the dot, regardless of the magnitude of temporal offset, $F(3, 18) = 6.71$, $p = 0.003$, $\eta_p^2 = 0.94$. However, when dot motion followed body movements in the lag condition, the temporal misalignment magnitude significantly impacted human judgments, $F(3, 18) = 62.13$, $p < 0.001$, $\eta_p^2 = 1.0$. When the proportion of "success" responses in the ahead condition was subtracted from the corresponding proportion in the lag condition, the difference was significantly below zero for temporal offsets of 0.5 s, 1 s, and 1.5 s (all ps < 0.001), indicating a higher proportion of success responses when the dot motion (cause) preceded the body movements (effect), relative to the corresponding condition in which the effect cue preceded the cause.

These results imply that when observers received a cover story in Experiment 2b that reversed their interpretation of the cause-effect relations between moving objects (i.e., moving dot as cause and limb movements as effect), their judgments changed dramatically. Observers were more likely to judge the action to be successful in following the laser spot when the dot motion was shifted forward relative to limb movements. Furthermore, this temporal asymmetry effect was maintained for a large range of offset magnitudes (from .5 s to 1.5 s). The strength and robustness of the effect is likely due to people's qualitative interpretation of a "following" action: this relation is granted as long as the movement of one object follows the same trajectory as that of another object. Whereas the causal relation between limb movements and body motion of a single agent (Experiments 1 and 2a) is closely coupled in time, the action of an agent that is following a separate object (Experiment 2b) can be much

more temporally variable, as long as the motion of the agent lags behind that of the object.

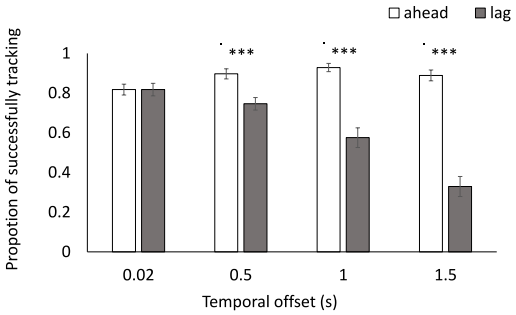


Figure 5: Results of Experiment 2b. The proportion of reported successful tracking responses in the ahead condition (laser dot was shifted ahead of limb movements) and in the lag condition (laser dot lagged).

An Ideal Observer Model

To further assess whether the temporal asymmetry effect could possibly be explained by statistical regularity in walking actions, rather than causal relations between two movement cues, we developed an ideal observer model solely based on visual statistics of action stimuli. The observer model is a hypothetical device that makes optimal decisions given available information based on natural statistics of the visual environment (Geisler 2011; Kersten, Mamassian & Yuille, 2004; Lu & Yuille, 2006). To capture the natural statistics in relevant action stimuli, we analyzed 20 walking actions (each consisted of 1500 frames) from the CMU motion-capture database, in which an actor explored an indoor environment with uneven surfaces. We generated a pool of point-light stimuli from 20 actions each viewed from six different directions with orthogonal projections. This stimulus set included a total of 180,000 posture frames (20*1500*6). To quantify body motion, the velocity was calculated as the position change of averaged hip joints from a frame to its neighboring frame. To reduce the number of postures resulting from limb movements in walking action, a K-mean algorithm (Jain, 2010) was employed to categorize the posture frames into a smaller number of key postures. Twenty key postures were selected since the sum of error reached a plateau after 20 clusters, indicating that adding more key postures did not improve the clustering performance. Each frame in the stimulus set was assigned to the most similar key posture as the corresponding label. Using all this information, we computed the histogram of velocity of body motion for each of the 20 key posture, and fitted the histogram using a 2D Gaussian distribution to estimate the mean and the covariance of common body motions given a key posture, as illustrated in Figure 6.

To simulate the judgment in Experiment 2a, each posture frame in an experimental trial was first assigned to the most similar key posture, and subsequently to the associated distribution of body motion. The displaced body motion in the input was mapped to the corresponding distribution to derive a likelihood. Finally, log-likelihood was calculated by summing up each frame's likelihood in the log scale for each stimulus. Figure 6 (right) shows the simulation results. Higher log-likelihood values indicate greater probability of considering the two motion cues as "matched". The ideal observer model consistently predicted more matched responses in the ahead conditions compared with the lag conditions, which is opposite to the temporal asymmetry effect observed in Experiment 2a. The ideal observer model (which lacks any causal constraint) thus failed to account for the pattern of human judgments. The failure of the model based solely on visual statistics shows converging evidence that causal understanding of body motion provides a critical constraint used by human observers in deciding whether perceived motion is natural.

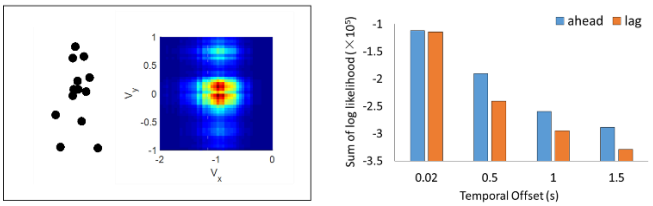


Figure 6: Results of an ideal observer model based on visual statistics. Left, body motion distribution associated with a key posture frame, derived from visual statistics from walking action observations. Right, the model simulation results for Experiment 2a, which is opposite from human results, suggesting that association between limb movements and body motion learned from visual statistics is insufficient to account for human performance.

General Discussion

The present study provides evidence for both perceptual and inferential processes involved in causal action judgments. Experiment 1 revealed that actions with preceding limb movements were rated as more natural than those with preceding body motion. This result suggests that people adopt the directionality of causal relation between limb movements and body motion when judging the naturalness of observed actions, despite this perceptual task appears not to explicitly require the sensitivity to causal relation between the two motion cues. The results of Experiment 2 collectively revealed that when considering movements attributed to a single actor (Experiment 2a), by default the brain binds bodily movements of articulated limbs (causes) to moving body locations in the environment (effects). However, when considering movements of one agent with respect to a separate distal object (Experiment 2b), the brain can flexibly assign the effect role to bodily movements of the agent, suggesting the involvement of the inferential process in causal action judgment. Thus people form a causal understanding of what drives perceived actions, which in turn influences their judgments of action naturalness.

The present findings both support and extend the large body of research on causal perception and inference. Seminal work on causal perception (Michotte, 1946/1963), coupled with contemporary developments (Scholl & Tremoulet, 2000; White, 2006), has illuminated the mechanisms by

which humans directly perceive the causal structure of the visual world. It is possible that a primitive psychological concept of causation may develop by extension from the fast, automatic and irresistible visual impression of causality arising from simple kinematic events. However, previous studies of causal perception have mostly involved simple displays of moving objects with rigid shapes. The present findings extend research on the interactions of objects (such as colliding balls) in the physical world to agent-related body movements. Thus, the paradigm introduced here (using action stimuli with whole-body movements) opens the door for further experimental investigations directed at the interface between causal perception and inference.

Actions afford privileged access to experience the role of agency, and hence provide a powerful tool to produce interventions that in turn help to discover causal relations in the physical and social environment (White, 1999). Numerous studies have shown that action perception is an active process, in which humans automatically predict actions to project the future course of an activity (Flanagan & Johansson, 2003; Graf et al., 2007; Prinz, 2006). Our study highlights the importance of causal interpretation in perceiving body movements. As relational binding in general enhances representational power (Lu, Chen & Holyoak, 2012), the perceived causal relation between the two movement cues enables people to understand why the body moves the way it does. The present study is limited to the simplest causal action, in which a single causal link exist between limb movements and body motion. The future investigations in causal action need to consider background cause (i.e., gravity) or other potential causal cues (i.e., body motion is caused by a moving skateboard). These situations open the door for.

Acknowledgments

This research was funded by a CSC Scholarship to YP and National Science Foundation grant BCS-1353391 to HL.

Reference

Abravanel, E., Levan-Goldschmidt, E., & Stevenson, M. B. (1976). Action imitation: The early phase of infancy. *Child development*, 1032-1044.

Buehner, M. J. (2012). Understanding the past, predicting the future causation, not intentional action, is the root of temporal binding. *Psychological Science, 23*(12), 1490-1497.

Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science, 20*(10), 1221-1228.

Choi, H., & Scholl, B.J. (2006). Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception. *Perception, 35*, 385-399.

Faro, D., Leclerc, F., & Hastie, R. (2005). Perceived causality as a cue to temporal distance. *Psychological Science, 16*(9), 673-677.

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature, 424*(6950), 769-771.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research, 51*(7), 771-781.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., & Prinz, W. (2007). Predicting point-light actions in real-time. *Neuroimage, 36*, T22-T32.

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature neuroscience, 5*(4), 382-385.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243-259.

Hume, D. (1888). *Hume's treatise of human nature* (L.A. Selby-Bigge, Ed.). Oxford, England: Clarendon Press. (Original work published 1739).Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol., 55*, 271-304.

Lu, H., & Yuille, A. (2006). Ideal observers for detecting motion: Correspondence noise. *Advances in neural information processing systems, 18*, 827-834.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological review, 119*(3), 617-648.

Ludwig, W. (1953). Philosophical investigations. *London, Basic Blackw*.

Masselink, J., & Lappe, M. (2015). Translation and articulation in biological motion perception. *Journal of vision, 15*(11), 10.

Michotte, A.E. (1963). *The perception of causality* (T.R. Miles, Trans.). London: Methuen & Co. (Original work published 1946).

Price, H. (1992). Agency and causal asymmetry. *Mind*, 501-520.

Prinz, W. (2006). What re-enactment earns us. *Cortex, 42*(4), 515-517.

Scholl, B. J., & Nakayama, K. (2002). Causal capture: Contextual effects on the perception of collision events. *Psychological Science, 13*(6), 493-498.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in cognitive sciences, 4*(8), 299-309.

Thurman, S. M., & Lu, H. (2013). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychological Science, 24*:1133-1141.

Thurman, S. M., & Lu, H. (2014). Perception of Social Interactions for Spatially Scrambled Biological Motion. *PLoS ONE. 9*(11), 1-12.

van Boxtel, J. J., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of vision, 13*(12), 7, 1-16.

White, P. A. (1999). Toward a causal realist account of causal understanding. *The American Journal of Psychology, 112*, 605-642.

White, P. A. (2006). The causal asymmetry. *Psychological review, 113*(1), 132-147.

A Computational Exploration of Problem-Solving Strategies and Gaze Behaviors on the Block Design Task

Maithilee Kunda (mkunda@vanderbilt.edu)
Mohamed El Banani (mbanani@gatech.edu)
James M. Rehg (rehg@gatech.edu)
School of Interactive Computing, Georgia Institute of Technology
85 Fifth Street NW, Atlanta, GA 30332 USA

Abstract

The block design task, a standardized test of nonverbal reasoning, is often used to characterize atypical patterns of cognition in individuals with developmental or neurological conditions. Many studies suggest that, in addition to looking at quantitative differences in block design speed or accuracy, observing qualitative differences in individuals' problem-solving strategies can provide valuable information about a person's cognition. However, it can be difficult to tie theories at the level of problem-solving strategy to predictions at the level of externally observable behaviors such as gaze shifts and patterns of errors. We present a computational architecture that is used to compare different models of problem-solving on the block design task and to generate detailed behavioral predictions for each different strategy. We describe the results of three different modeling experiments and discuss how these results provide greater insight into the analysis of gaze behavior and error patterns on the block design task.

Keywords: artificial intelligence; cognitive assessment; nonverbal intelligence; spatial reasoning; visual attention.

Introduction

The block design task is a cognitive assessment that is commonly used to measure nonverbal intelligence. In this task, a person has to reconstruct a given printed design using red and white blocks, as shown in Fig. 1. Originally devised in 1920 (Kohs, 1920), block design has been included in a multitude of standard neuropsychological test batteries, including every single edition of the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Adult Intelligence Scale (WAIS), which are among the most commonly used cognitive assessments around the world (Wechsler, 2003).

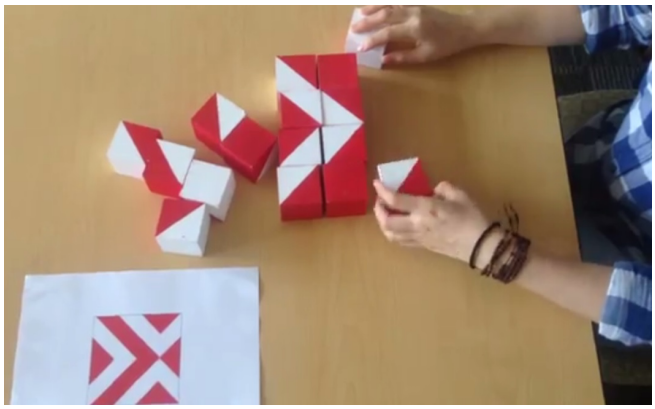


Figure 1: A person solving an example block design item. (To protect test security, actual test items are not shown.)

In the typical population, block design performance shows quantitative improvements with age in children (Wechsler, 2003) and a declining trend in the elderly (Ardila & Rosselli, 1989; Rönnlund & Nilsson, 2006). In addition to its value in the detection of brain damage (Lezak, 2004), block design is also sensitive to atypical development. Block design represents an area of strength for many individuals on the autism spectrum (Caron, Mottron, Berthiaume, & Dawson, 2006; Shah & Frith, 1993) and is an area of weakness in many individuals diagnosed with Williams syndrome (Farran, Jarrold, & Gathercole, 2001; Hoffman, Landau, & Pagani, 2003).

Kohs devised block design as a language-independent alternative to the Binet intelligence scale, which loaded heavily on verbal abilities (Kohs, 1920). The original scoring system takes into account the successful reproduction of the design as well as the time taken by the participant and the number of moves. Number of moves was dropped in later standardizations as being too cumbersome for daily practice (Hutt, 1932); current scoring schemes use final success and reaction time as the primary outcome measures. However, many researchers observe that obtaining more detailed measures of block design performance can add value by pinpointing specific cognitive processes and problem-solving strategies that an individual is employing to solve the task.

In particular, many studies have proposed that qualitative differences in block design performance are as important as quantitative differences in accuracy or reaction time. Several studies have observed that patterns of visual attention, i.e. eye gaze, between the original design and the design under construction may serve as markers for certain cognitive processes that are meaningful to the final outcome (Hoffman et al., 2003; Rozencwajg & Corroyer, 2002; Rozencwajg et al., 2005; Rozencwajg & Fenouillet, 2012).

Observing the errors that people make can also provide valuable information. Errors have been studied in terms of the particular sequence of moves that a participant makes (Joy, Fein, Kaplan, & Freedman, 2001; Toraldo & Shallice, 2004), any incorrect placements of blocks (Ben-Yishay, Diller, Mandleberg, Gordon, & Gerstman, 1971; Hoffman et al., 2003; Jones & Torgesen, 1981; Joy et al., 2001; Schatz, Ballantyne, & Trauner, 2000; Troyer, Cullum, Smernoff, & Kozora, 1994), the qualitative type or scale of errors that are made (Akshoomoff, Delis, & Kiefner, 1989; Joy et al., 2001), and also what is termed a "broken configuration" error, in which the participant places a block too far away from their cur-

rent construction, violating implied configural outlines of the given design (Akshoomoff et al., 1989; Akshoomoff & Stiles, 1996; Joy et al., 2001; J. Kramer, Blusewicz, Kaplan, & Preston, 1991; J. H. Kramer, Kaplan, Share, & Huckeba, 1999; Schatz et al., 2000; Troyer et al., 1994; Zipf-Williams, Shear, Strongin, Winegarden, & Morrell, 2000).

However, missing from these accounts is a precise description of the information processing mechanisms involved in solving block design items. As a result, we also know relatively little about how these mechanisms emerge in typical development or how they are altered in atypical development.

Obtaining a better understanding of the specific cognitive processes underlying performance on standard neuropsychological assessments, especially taking into account both quantitative and qualitative differences, provides value in understanding the basic science of cognition and development as well as the interpretation of these assessments in practical settings (Hunt, 1983; Kaplan, 1988; Keating & Bobbitt, 1978; Mislevy & Verhelst, 1990; Sternberg, 1988). Computational modeling is an excellent method for this kind of detailed scientific inquiry. Implementing a computational model forces precision in the theory being tested. Royer has argued for increased specificity in theories of block design performance through conceptual information processing models (Royer, 1984), and having a working, runnable computational model takes this desired level of specificity one step further.

We propose a new computational architecture for modeling the process of solving a block design item as a recurring interplay between attention, perception, memory, and action. We implemented this architecture using a custom block design simulation environment. We present the results from three computational experiments using this architecture and discuss how the architecture can be used to better understand the relationships between different problem-solving strategies and the gaze behaviors and errors that they generate.

There are several existing cognitive architectures used to model human visuospatial task performance, such as ACT-R (Gunzelmann & Lyon, 2011) and SOAR (Laird, 2008). However, these architectures were originally developed using frameworks of cognition that prioritize symbolic processing. Modules for representing and manipulating nonsymbolic information, for instance mental imagery, have been incorporated into these architectures relatively recently. In our own work, we study mechanisms of intelligent problem solving in which visual mental images form the primary (and often only) mode of representation, and we have developed a series of computational models that adopt this perspective to solve different neuropsychological assessments, such as the Raven's Progressive Matrices test (Kunda, McGreggor, & Goel, 2013) and the Embedded Figures test (Kunda & Ting, 2015).

We have built upon this prior work in the design of the block design architecture presented in this article. The main reason that we chose to construct a new architecture instead of using an existing framework like ACT-R or SOAR is that, because our research prioritizes the role of perceptual mental representations in problem solving, we have designed our architecture to likewise rely primarily on perceptual representations, instead of having a symbolic representation system that is common to a variety of perceptual representation modules.

The Simulation Environment

We have developed a computational architecture that can autonomously solve the block design task in a simulated environment. The current environment is simplified in many ways; ongoing refinements to the environment and architecture will gradually introduce more realistic representations.

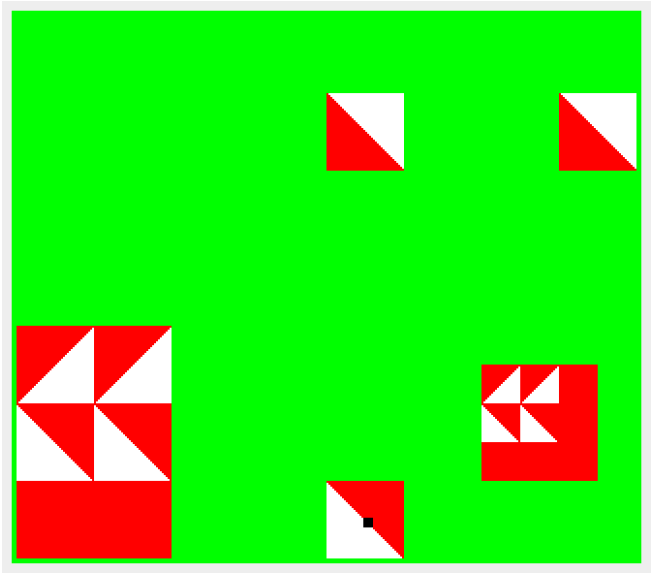


Figure 2: The simulated environment used by our architecture, representing a top-down view of the tabletop containing all of the materials for one block design item. The black dot in the image (bottom center) indicates the simulated gaze location of the architecture at the current time step.

The perceptual input available to the model consists of an overhead image of the table, as shown in Fig. 2. The table image is divided into the following regions:

1. **The target design**: The target design remains fixed for the duration of each test item. The design is represented by a single image scaled down to half the size of the actual constructed designs, consistent with standard block design administration procedures (Kohs, 1920; Wechsler, 2003).
2. **The construction area**: The construction area is located at the bottom left corner of the table. This area is empty at time 0, and has the ability to contain an assortment of blocks placed at various locations. This part of the environment is simplified by having an implied rectilinear grid that only allows blocks to be placed in slots that comply with this grid. As a result, block misalignment errors while performing the task cannot be replicated within the current generation of models. However, broken configuration errors can be replicated, as the construction area does not

have any explicit borders, allowing the model to assemble a design that extends beyond the implied boundaries.

3. **The block bank**: The block bank is the area located at the top right corner of the table and is similar to the construction area in that it contains an assortment of blocks placed at various locations and in various orientations. This area contains all of the available blocks at time 0.
4. **The hand area**: This area of the table represents the block being manipulated by the model and is located to the right of to the construction area on the table. At each time step, this area can either contain a block or be empty.
5. **The set of blocks**: The table contains N blocks, with at least as many blocks as are required to construct the target design. Each block has a 2-D position as well as a 3-D orientation to indicate the top-most face of the block.

The behavioral variables that can be captured from this environment include: success/failure, sequence of block moves, and gaze shifts to each area of the environment. While the simulated environment is capable of supporting time-based measurements of model performance, detailing timing information is being incorporated into the architecture as part of ongoing work and is not used in the current study.

The Computational Architecture

We designed the block design computational architecture to have six basic mechanisms, as shown in Fig. 3:

1. **The visual perception module** takes as input a visual scene representation from the simulation and outputs a new image at each time step, which we call the perceptual image. The content of this image varies depending on where the model is currently directing its gaze. The perceptual image has a fixed resolution but can "zoom" in or out relative to the table, which allows the model to capture different amounts of visual information from the scene, ranging from the general form of a large section of the table to a highly detailed image of a small section of the table.
2. **The visual attention module** specifies a gaze location at each time step, which can be directed to specific points within one of four broad locations: the target design, the available blocks area, the design being constructed, and the hand location. This module also includes mechanisms for performing visual search either for any block or for a block whose top-most face best matches the block face stored in the short-term memory buffer.
3. **The mental imagery module** stores the perceptual image in a short-term memory buffer for later use as a "mental image." The mental imagery module can compute a measure of visual similarity between the current perceptual image and the current mental image (Kunda et al., 2013).
4. **The short term memory buffer** stores the output of the mental imagery module as well as any relevant spatial information that is required to solve the task. In the current implementation, the only spatial information that is stored is a reference point on the current object/area of interest, in the form of Cartesian coordinates.
5. **The motor action module** takes actions that change the state of the environment. The available actions include: 1) selecting a particular block from the table and putting this block into the hand location, 2) rotating a block that is in the hand location, the block bank, or the construction area, or 3) placing a block currently held in the hand to a specific position inside the construction area or the block bank.
6. **The central executive module** is responsible for planning and decision making, and it coordinates the overall operation of the model. At each time step, this module can execute any action from a set of available actions. The available actions include changing the gaze location, performing a visual search for a specific image, storing, manipulating, or flushing the mental image, and selecting, rotating, or placing a block in the simulated environment.

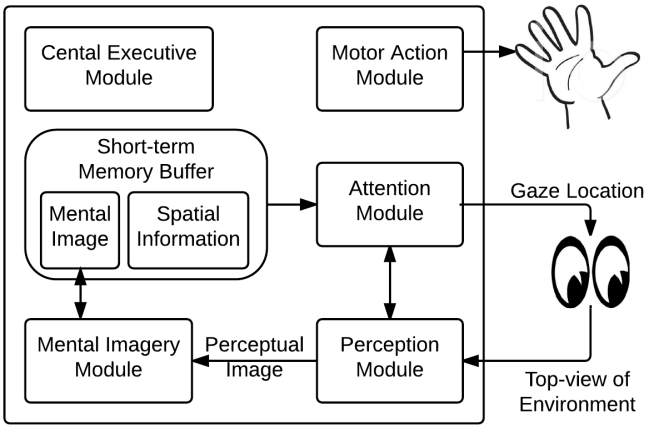


Figure 3: The overall model architecture, including inputs and outputs to the simulated block design task environment described in the previous section. The central executive module has bidirectional connections to the other five modules in the system; for clarity, these connections are not shown.

Experiments and Results

Using the simulation environment described above, we ran three computational experiments in which we compared the performance of different block design models built using the same underlying architecture. Each model was presented with 9 blocks and a single 3x3 target design, which appears in the lower right corner of Fig. 2. We present results from the model averaged over 1000 runs. Randomness effects within a single run stem from the initial random positioning of the blocks as well as randomness in the model's search strategies.

While there are likely many high-level strategies that can be used to solve block design items, the current models that we tested all follow what has been termed the "analytic strategy," in which "the displayed design is mentally segmented into units corresponding to block faces, then the blocks are directly placed, one by one, to match each unit" (Schorr, Bower, & Kiernan, 1982). Other high-level strategies, including the influence of Gestalt effects on perception, are not explicitly

implemented in the current family of models but are part of ongoing work to improve and expand the basic architecture.

Because there are random factors in the environment and in each model's processing, we measured the output of the models by constructing gaze transition graphs. These graphs show the average number of gaze transitions made by each model within or between four different areas of the simulation environment: the block bank, the construction area, the target design, and the model's "hand."

Experiment #1: Visual Working Memory

In solving a block design item, each model first looks at the design, stores some part of the design in memory, then selects and uses blocks to reconstruct this stored part of the design. Then, the model returns to the design to see and store the next part of the design. This experiment investigates the effects of the size of a model's visual working memory on its task performance, in terms of how much information from the design can be stored at each iteration.

Each model stores some part of the design in its mental imagery module. In the current implementation, the mental images are square and have a fixed size, defined by the number of blocks they can store. We created two different models, one with a 1x1 mental image size and one with a 3x3 mental image size. The mental images are stored at the same resolution at which they are perceived.

Each of the two models applied a guided search strategy to locate the blocks and used a visual similarity threshold of 0.95. (These two additional parameters are discussed in detail in the next two experiments.) The gaze transition graphs for each of these models are shown in Fig. 4.

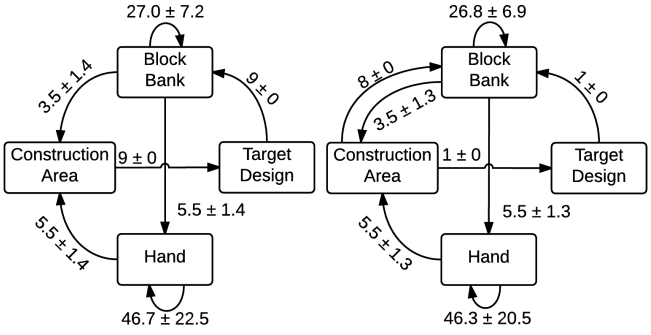


Figure 4: Gaze transitions made by models with a mental image size of 1x1 (left) or 3x3 (right). Results show *mean +/- standard deviation* over 1000 trials.

Two noticeable effects can be seen in this figure. First, the model with larger mental image capacity only has to look at the target design once, while the other model must return its gaze to the target design many times. Second, the model with larger mental image capacity shows more gaze transitions between the block bank and construction area, because it is working off a mental image instead of having to return to the design after placing each block.

Experiment #2: Search Strategy

Each model has a search strategy that determines which block from the block bank is chosen to add to the construction area. Because the blocks are identical, a random choice can always suffice. In the *random search strategy*, the model first randomly chooses a block from the block bank and moves it to the hand area. Then, the model segments the target design to determine the block face that should be added next to the construction area. If the top-most face of the block in hand matches the target block face, the block is immediately added to the construction area. Otherwise, the block is rotated to match the target before being added to the construction area.

Alternately, a model can be more strategic about its choice from the block bank. In the *guided search strategy*, the model first segments the target design to determine the block face that should be added next to the construction area. The model then proceeds to search the remaining available blocks in the block bank for the target block face. If an exact match is found, the matching block is added to the construction area. If not, the best match is picked up and rotated to match the target block face and then added to the construction area.

Both models have a mental image size of 1x1 and a visual similarity threshold of 0.95. The gaze transition graphs for each of these models are shown in Fig. 5.

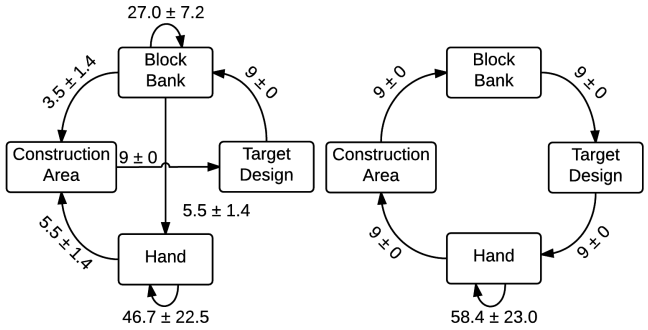


Figure 5: Gaze transitions made by models using a guided search strategy (left) versus a random search strategy (right). Results show *mean +/- standard deviation* over 1000 trials.

In this experiment, it is interesting to see that the random search strategy is, in a sense, more efficient. Gaze transitions follow a very definite pattern, and the only added gaze behaviors are those that monitor the rotation of each block while it is being held. The model following the guided search strategy spends much more time looking through the blocks in the block bank, and even then, the model still must monitor a large number of in-hand block rotations.

Experiment #3: Visual Similarity Threshold

The models use a measure of visual similarity to compare two images, for example to compare a mental image to a perceived block in the environment. Each model has a similarity threshold to determine when two images are considered to be the same. In the current implementation, visual similar-

ity is calculated simply as the proportion of matching pixels between two images (where 0.0 indicates no match and 1.0 indicates a perfect match). We varied this visual similarity threshold to be 0.95 in one model and 0.64 in another.

Both models have a mental image size of 1x1 and use a guided search strategy. The gaze transition graphs for each of these models are shown in Fig. 6.

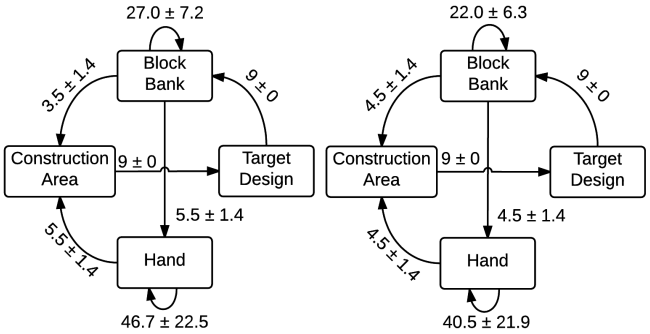


Figure 6: Gaze transitions made by models having a visual similarity threshold of 0.95 (left) or 0.64 (right). Results show *mean +/- standard deviation* over 1000 trials.

Both of these graphs show the same connectivity. However, the model with the lower similarity threshold value shows fewer gaze behaviors on both the block bank (while searching for a matching block) and on the hand (while rotating blocks as needed). This decrease in visual effort makes sense given that the model is more lenient with what it considers to be a visual match.

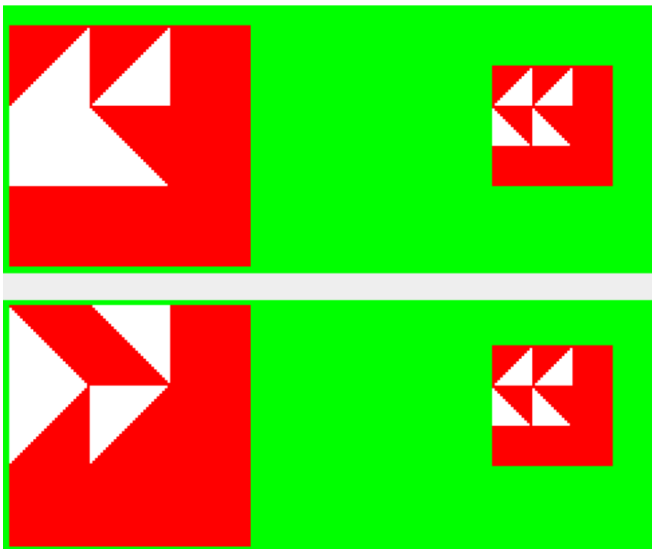


Figure 7: Examples of errors made by models with visual similarity thresholds of 0.64 (top) and 0.5 (bottom).

Lowering the similarity threshold can also lead to errors. While the model with the higher threshold makes no errors, the model with the lower threshold produces many incorrect reconstructions of the target design, with overall accuracy falling to 25.4%. The errors produced by decreasing the visual similarity threshold are specific to the incorrect placement of blocks that share high visual similarity with the correct block that should have been placed at that position. Two examples of the kinds of errors that occur at different threshold values are shown in Fig. 7. At a threshold value of 0.64, the errors take the form of placing a solid face in the place of a diagonal one. At a lower threshold value of 0.5, a model could replace a diagonal face block with a diagonal face block that has a different 2-D orientation. However, more acute errors, such as a solid red face being replaced by a solid white face or a diagonal face being replaced by its complement do not occur even with a similarity threshold of 0.5, as the visual similarity between such extreme pairs is much lower.

Discussion and Future Work

Despite the ubiquity of block design in clinical, scientific, and educational settings, we still do not know the precise nature of the mechanisms and strategies that people use to solve the task. Describing these mechanisms and strategies at a computational level provides an experimental platform on which to construct and flesh out theories of problem-solving on the block design task and to make detailed predictions about behavior, including both quantitative and qualitative variations.

We have shown how computational models of problem-solving on the block design task can be used to test the effects that different variations in strategy and other cognitive characteristics have on behavior, including accuracy, gaze patterns, and the types of errors that are made. While the models we have presented are admittedly simplified in comparison to human performance, these models give us a structured way to understand both the requirements of the task and how specific mechanisms relate to specific behaviors.

In future work, we will collect data from human participants to analyze their gaze transition graphs and error patterns, comparing them to those of our computational models, to gain more insight into the variability of strategy and behavior across our participant sample. These analyses will provide insight into cognitive development and individual differences in typically developing children as well as in individuals with different developmental or neuropsychological conditions.

Continued work in this direction will: 1) enhance our scientific understanding of how nonverbal intelligence develops and matures, 2) enable the generation of detailed behavioral predictions and research questions for future scientific inquiry, 3) inform the practical development and interpretation of neuropsychological assessments, and 4) provide a conceptual bridge to map behavioral observations onto measurements and models of neural activity.

Acknowledgments

We thank Y. Kang, M. Munch, E. Nkadi, and R. Stauffer for contributions to this project. This research was supported by NSF Expedition Award #1029679 and by the Georgia Tech Undergraduate Research Opportunities Program.

References

Akshoomoff, N. A., Delis, D. C., & Kiefner, M. G. (1989). Block constructions of chronic alcoholic and unilateral brain-damaged patients: A test of the right hemisphere vulnerability hypothesis of alcoholism. *Archives of Clinical Neuropsychology*, *4*(3), 275–281.

Akshoomoff, N. A., & Stiles, J. (1996). The influence of pattern type on children's block design performance. *J. International Neuropsychological Society*, *2*(5), 392–402.

Ardila, A., & Rosselli, M. (1989). Neuropsychological characteristics of normal aging. *Dev. Neuropsych.*, *5*, 307–320.

Ben-Yishay, Y., Diller, L., Mandleberg, I., Gordon, W., & Gerstman, L. (1971). Similarities and differences in block design performance between older normal and brain-injured persons: A task analysis. *J. Abnormal Psychology*, *78*, 17-25.

Caron, M., Mottron, L., Berthiaume, C., & Dawson, M. (2006). Cognitive mechanisms, specificity and neural underpinnings of visuospatial peaks in autism. *Brain*, *129*(7), 1789–1802.

Farran, E. K., Jarrold, C., & Gathercole, S. E. (2001). Block design performance in the Williams syndrome phenotype: A problem with mental imagery? *J. Child Psychology and Psychiatry*, *42*(6), 719–728.

Gunzelmann, G., & Lyon, D. R. (2011). Representations and processes of human spatial competence. *Topics in Cognitive Science*, *3*(4), 741–759.

Hoffman, J. E., Landau, B., & Pagani, B. (2003). Spatial breakdown in spatial construction: Evidence from eye fixations in children with Williams syndrome. *Cognitive Psychology*, *46*(3), 260–301.

Hunt, E. (1983). On the nature of intelligence. *Science*, *219*(4581), 141–146.

Hutt, M. (1932). The Kohs block-design tests: A revision for clinical practice. *J. Applied Psychology*, *16*(3), 298-307.

Jones, R. S., & Torgesen, J. K. (1981). Analysis of behaviors involved in performance of the block design subtest of the WISC-R. *Intelligence*, *5*(4), 321–328.

Joy, S., Fein, D., Kaplan, E., & Freedman, M. (2001). Quantifying qualitative features of block design performance among healthy older adults. *Archives of Clinical Neuropsychology*, *16*(2), 157–170.

Kaplan, E. (1988). The process approach to neuropsychological assessment. *Aphasiology*, *2*(3-4), 309–311.

Keating, D. P., & Bobbitt, B. L. (1978). Individual and developmental differences in cognitive-processing components of mental ability. *Child Development*, *49*(1), 155–167.

Kohs, S. (1920). The block-design tests. *J. Experimental Psychology*, *3*(5), 357-376.

Kramer, J., Blusewicz, M., Kaplan, E., & Preston, K. (1991). Visual hierarchical analysis of block design configural errors. *J. Clin. and Exp. Neuropsychology*, *13*(4), 455–465.

Kramer, J. H., Kaplan, E., Share, L., & Huckeba, W. (1999). Configural errors on WISC-III block design. *J. International Neuropsychological Society*, *5*(6), 518–524.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the ravens progressive matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47–66.

Kunda, M., & Ting, J. (2015). Looking around the minds eye: How internal deployments of attention can affect visual search performance. In *Proc. Third Annual Conf. Advances in Cognitive Systems (ACS).*

Laird, J. E. (2008). Extending the Soar cognitive architecture. In *Proc. 2008 Conf. Artificial General Intelligence (AGI).*

Lezak, M. D. (2004). *Neuropsychological assessment*. Oxford University Press, USA.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Rönnlund, M., & Nilsson, L.-G. (2006). Adult life-span patterns in WAIS-R block design performance: Cross-sectional versus longitudinal age gradients and relations to demographic factors. *Intelligence*, *34*(1), 63–78.

Royer, F. L. (1984). Stimulus variables in the block design task: A commentary on Schorr, Bower, and Kiernan. *J. Consulting and Clinical Psychology*, *52*(4), 700-704.

Rozencwajg, P., Cherfi, M., Ferrandez, A., Lautrey, J., Lemoine, C., & Loarer, E. (2005). Age related differences in the strategies used by middle aged adults to solve a block design task. *Int. J. Aging & Human Dev.*, *60*(2), 159–182.

Rozencwajg, P., & Corroyer, D. (2002). Strategy development in a block design task. *Intelligence*, *30*(1), 1–25.

Rozencwajg, P., & Fenouillet, F. (2012). Effect of goal setting on the strategies used to solve a block design task. *Learning and Individual Differences*, *22*(4), 530–536.

Schatz, A., Ballantyne, A., & Trauner, D. (2000). A hierarchical analysis of block design errors in children with early focal brain damage. *Dev. Neuropsychology*, *17*(1), 75–83.

Schorr, D., Bower, G. H., & Kiernan, R. J. (1982). Stimulus variables in the block design task. *J. Consulting and Clinical Psychology*, *50*(4), 479-487.

Shah, A., & Frith, U. (1993). Why do autistic individuals show superior performance on the block design task? *J. Child Psychology and Psychiatry*, *34*(8), 1351–1364.

Sternberg, R. J. (1988). Applying cognitive theory to the testing and teaching of intelligence. *Applied Cognitive Psychology*, *2*(4), 231–255.

Toraldo, A., & Shallice, T. (2004). Error analysis at the level of single moves in block design. *Cognitive Neuropsychology*, *21*(6), 645–659.

Troyer, A., Cullum, C., Smernoff, E., & Kozora, E. (1994). Age effects on block design: Qualitative performance features and extended-time effects. *Neuropsychol.*, *8*, 95-99.

Wechsler, D. (2003). *Wechsler intelligence scale for children (WISC-IV)*. The Psychological Corporation.

Zipf-Williams, E., Shear, P., Strongin, D., Winegarden, B., & Morrell, M. (2000). Qualitative block design performance in epilepsy patients. *Archives of Clinical Neuropsychology*, *15*, 149–157.

Preferring the Mighty to the Meek: Toddlers Prefer Novel Dominant Agents.

Ashley J. Thomas (ashleyjt@uci.edu)
Department of Cognitive Sciences,
3151 Social Sciences Plaza
Irvine, CA

Meline Abramyan (mdabramy@uci.edu)
Department of Cognitive Sciences,
3151 Social Sciences Plaza
Irvine, CA

Angela Lukowski (lukowski@uci.edu)
Department of Psychology and Social Behavior
4201 Social and Behavioral Sciences Gateway
Irvine, CA

Lotte Thomsen (lotte.thomsen@psykologi.uio.no)
Department of Psychology, University of Oslo, Norway
Department of Political Science, University of Aarhus, Denmark

Barbara W. Sarnecka (sarnecka@uci.edu)
Department of Cognitive Sciences,
3151 Social Sciences Plaza
Irvine, CA

Abstract

Every human society includes social hierarchies—relationships between individuals and groups of unequal rank or status. Recent research has shown that even preverbal infants represent hierarchical relationships, expecting larger agents and agents from larger groups to win dominance contests. However, to successfully navigate social hierarchies, infants must also integrate information about social rank into their own behavior, such as when deciding which individuals to approach and which to avoid. Here we demonstrate that two-year-old children (ages 21-31 months) preferred novel dominant agents to subordinates. That is, by the age of 21 months, toddlers not only use phylogenetically stable cues to predict the winner of dominance contests, they also like the dominant agents better. This finding suggests that young children use their ability to infer relative rank to selectively approach dominant individuals.

Keywords: social hierarchy, naïve sociology, infant cognition, social cognition

Introduction

Humans are ultra-social species. A person's ability to interpret, create, and maintain social relationships is key to his or her physical and mental well-being (Silk, 2007). One very common type of social relationship occurs between individuals who are ranked along some linear dimension such as age, military rank, physical size, etc. (Fiske & Schubert, 2012; Fiske, 1992; Kaufmann & Clément, 2014; Sapolsky, 2004). Such hierarchical relationships are found across species (Sapolsky, 2004), across human societies (Boehm, 1999; Fiske, 1992), and across human social settings (e.g., academic, domestic, recreational, professional) (Magee & Galinsky, 2008).

The ubiquity of social hierarchies means that it is important that individuals can recognize relative social rank. Indeed, natural selection has conferred this ability on many species, including fish, birds, chimps, wolves, and humans (Bond, Wei, & Kamil, 2010; Grosenick, Clement, & Fernald, 2007; Sapolsky, 2004).

In humans, the ability is present very early on: preverbal infants expect larger agents and agents with more allies to win right-of-way dominance contests (Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011; Pun, Birch, & Baron, 2016). They also expect rank to be transitive (i.e., they expect that if A outranks B, and B outranks C, then A will outrank C) (Gazes, Hampton, & Lourenco, 2015, see also Mascaro, Csibra, 2014) and expect that those who have won dominance contests in the past (e.g. over resources) will win new dominance contests (e.g. over territory) (Mascaro & Csibra, 2012.)

Such studies demonstrate that infants use dominance cues found across phylogenesis to *detect* hierarchical relationships. But to survive and thrive in human society, detection of rank is not enough; individuals must also use this information to motivate their own behavior toward others. For example, they must evaluate social situations to decide which individuals to approach and which to avoid, and which to form relationships with.

One reason to expect that children might behave differently toward high- vs. low-ranked individuals comes from observations in daycare centers (Sluckin & Smith, 1977). Toddlers are deferent to peers that outrank them, and assertive towards those they outrank. Furthermore, toddlers who take contested toys or thwart other children tend to be preferred playmates and are more likely to influence and be imitated by other children, and rank between two children often remains stable across time (see Hawley, 2015 for review; Sluckin & Smith, 1977).

While this indicates that children do integrate rank into their behavior, it could simply reflect a history of interactions between individual children (e.g. a child might think, "I fought with this kid yesterday over a toy and lost; I don't want to go through that again, so now I will just give him the toy.") This leaves open the question of whether children possess a 'default' core motivational system that specifies how to relate to novel individuals depending on their social rank.

For example, it has been argued that detecting social rank is adaptive because it allows infants (and animals) to 'size up the competition' and avoid dangerous conflicts they are unlikely to win (e.g. Pun et al., 2016). And in fact, the ability to avoid dominant individuals is associated with lower stress levels in other species (Sapolsky, 2005). For example, subordinate wolves in captivity, who cannot easily avoid dominant wolves, have elevated stress hormones compared to subordinate wolves in the wild. If avoiding dominant individuals is the main benefit of detecting rank, then we might expect children in the present study to avoid the novel high-ranking individual.

However, insofar as dominance rank reflects success in one's ecological context, affiliating with high-ranking individuals might bring benefits in the form of access to important know-how, (Chudek, Heller, Birch, & Henrich, 2012; Henrich & Gil-White, 2001), material resources, and control (Thomsen & Carey, 2013). Consistent with this proposal, macaques use rank to decide whom to align with when facing opponents, and they consistently choose allies who outrank themselves and their opponents (Silk, 1999).

The present study was a first step toward testing whether human children evaluate novel individuals according to their relative social ranks—motivating them to either approach or avoid them. Experiment 1 used a proven right-of-way paradigm (Thomsen et al., 2011, replicated in Pun et al, 2016) to establish that one individual in a dyad was ranked higher than the other. We then tested whether children ages 21-31 months prefer the dominant individual or the subordinate, using a well-established reaching paradigm (Hamlin, Wynn, & Bloom, 2007; Hamlin, Wynn, Bloom, & Mahajan, 2011). Experiment 2 followed up results from Experiment 1, showing children a scene that was similar but lacked any interaction between the characters. This ruled out a number of alternative explanations for the results of Experiment 1, including the explanation that toddlers may prefer the individual that reaches its goal, the individual that moves last or moves farthest, or the individual that stays upright throughout the scene.

Experiment 1

Methods

Participants

Participants were recruited at a children's museum. A total of 30 children participated in the experiment. Of these, 8 were excluded from the analysis for the following reasons: Refusing to choose a puppet (n=4); Choosing both puppets (n=2); Extreme fussiness (n=1); Distraction in the testing environment (a janitor entered the testing room and made loud noises, n=1). The remaining 22 children (9 girls, 13 boys) contributed data to the analysis. Their ages ranged from 21-31 months (M=24.95 months (SD=2.92 months). At the time of enrollment in the study, parents were asked to fill out a demographic questionnaire asking about the child's race and ethnicity, and their household income. Responses to the race/ethnicity questions included the following: White, Not Hispanic (n=11); Asian, Not Hispanic (n=6); Asian and White, Not Hispanic (n=2); White, Mexican-American (n=1) and Mexican-American, no race indicated (n=2).

We stopped at n=22 because we had a sufficiently strong Bayes Factor (see analysis below). Although frequentist statistical analyses do not allow for preferential stopping, Bayesian analyses do (Csibra et al., 2016; Dienes, 2011).

Materials & Procedure

The puppet stage used in all experiments was 75cm tall, 32.5cm deep, and 95cm long. It was placed on a folding table covered with black fabric. There were black curtains attached to the left and right side of the puppet stage, and a black curtain was used to cover the stage between scenes. Another black curtain behind the stage hid the experimenter who was manipulating the puppets. The puppets were 12.5cm tall and made of clay. They each had one plastic craft eye (with a fixed pupil so that the puppet always seemed to be looking straight ahead) and a black rectangle for a mouth. One puppet was a yellow oval and one was a red rectangle; the shapes were equal in diameter. The puppets were moved by means of black wooden dowels attached to their bases. The experimenter administering the puppet show wore black gloves and black clothing, and was hidden from the child's view. After the puppet show, two puppets identical to those used in the puppet show were presented to the child on a white foam board measuring 60 x 90cm. The puppets were placed on the board 75cm apart, and were attached to the board with magnets.

Participants were recruited from the floor of a children's museum during regular business hours. Parents

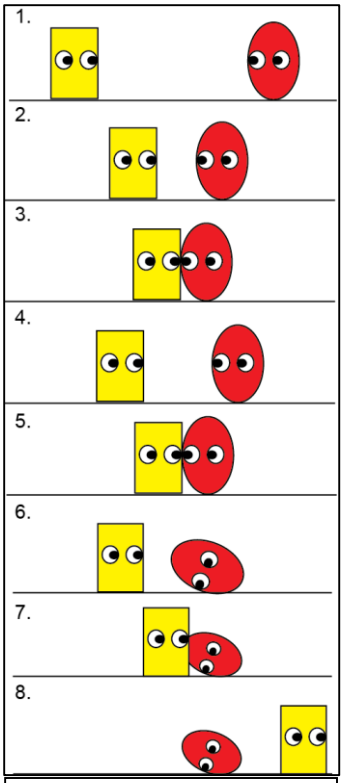


Figure 1: Puppet Show used in Experiment 1

were approached by an experimenter who greeted the parent and asked if they would like to hear more information about an experiment on children's understanding of social relationships. If the parent agreed, they were given the consent form and demographic questionnaire to fill out while the experimenter interacted with the child before leading the parent and child to a testing room. The testing room was a large room off the main floor of the museum. Before entering the testing room, parents were briefed about the procedure. They were asked to remain quiet during the puppet show and to close their eyes during the choice procedure. The participating child usually sat on the parent's lap. When this was not possible (e.g., because the parent was holding a younger sibling), the child sat in a chair next to their parent.

After the parent and child were seated in the testing room, the child was shown the puppet show. One experimenter (occluded from the child's view) acted as the puppeteer. A second experimenter who was blind to the condition (i.e. could not see which puppet was the dominant puppet and which was the subordinate) stood beside the stage and moved the curtain up and down between segments, saying "down goes the curtain" when pulling the curtain down, and "up goes the curtain" when pulling the curtain up.

Following Thomsen, et al, (2011), during the familiarization sequence, one puppet, alone on stage, crossed the stage two times (both times in the same direction). Then the other puppet, also alone, crossed the stage two times in the opposite direction. The direction that the higher and lower-ranked puppet traveled was counterbalanced, along with which puppet crossed the stage first and which puppet was higher ranked (i.e. yellow oval or red square), In the hierarchy display sequence the two puppets appeared on opposite sides of the stage and started across at the same time, meeting in the middle. Upon meeting, both puppets backed up and tried again, only to meet again in the middle. After this meet-and-retreat sequence was repeated three times, the puppets approached one another, this time without touching (so it did not look like the lower-ranked puppet was being knocked over). Next, the lower-ranked puppet bowed down and moved aside, allowing the other puppet (the high-ranked puppet) to pass by. The entire hierarchy display sequence was repeated three times.

Following Hamlin et al, 2007, during the choice procedure, the second experimenter (who was blind to the condition) reminded the parent to close his or her eyes. The experimenter then retrieved the board with the puppets attached to it, and held it so that the child could see (but not reach) the puppets. The experimenter looked at the child and said, "Hi!" then said, "Look!" and looked down at the board, fixing her gaze directly in the center of the board, between the two puppets. Finally, the experimenter said, "Which one do you like?" and pushed the board toward the child so that the child could reach it. The experimenter mentally counted off 30 seconds. If the child had not made a choice after 30 seconds, the experimenter (keeping her gaze fixed on the center of the board) encouraged the child by saying, for example, "Its ok to choose one," or "You can grab one.") If the child still made no choice, the experimenter returned the board back to its starting position and repeated the choice process. If the choice procedure was done three times and the child still made no choice, then the trial was coded 'no response.'

After the puppet show, each child was given a prize (i.e. a rubber duck). Parents were invited to ask questions about the study and were given information about the lab to take home. Two research assistants who were blind to condition coded each video for choice. There were no disagreements between the two coders.

Results and Discussion

Of the 22 children who contributed data to the analysis, 18 chose the 'high-ranked' or 'dominant' puppet (as established by the hierarchy-display sequence of the puppet show). To compare the observed distribution (i.e. 18/22) to an expected distribution (i.e. assuming that children choose the high-ranked 50% of the time), we used 'binom.test' in the statistical program 'R' (p=.004344, estimated probability of success = .8182). We also used a Bayes Factor Binomial calculator (http://pcl.missouri.edu/bf-binomial) and found that the odds of the alternative model (i.e. the children were choosing the high-ranked either more or less than 50% of the time) over the null model (i.e. the children were choosing the high-ranked 50% of the time) to be 24.93, which is considered to be strong evidence (Kass & Raftery, 1995).

Experiment 2

Experiment 2 was conducted to rule out five alternative explanations as to why children might prefer the puppet who successfully crossed the stage: (1) Children might prefer the puppet that reached its goal (in this case, crossing the stage) over the one that failed to reach its goal; (2) Children might prefer the puppet that remained visible the whole time (in Exp. 1, the low-ranked puppet was occluded when the high-

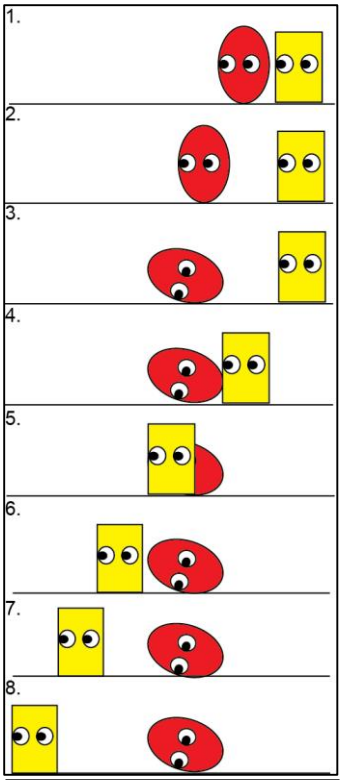


Figure 2: Puppet Show used in Experiment 2

ranked puppet passed in front of it); (3) Children might prefer the puppet that moved last; (4) Children might prefer the puppet that moved farther (i.e., traveled a longer distance); (5) Children might prefer the puppet that moved in the same way during the hierarchy-display sequence as it did during the familiarization sequence (crossing the stage instead of stopping in the middle). Experiment 2 replicated the features of Experiment 1, but without any social interaction (and therefore no show of relative social rank) between the puppets. The procedure was the same as Experiment 1 except for the following changes: During the familiarization sequence, the puppets again both crossed the stage twice, but this time they moved in the same direction. During the control-display sequence, the two puppets appeared on the same side of the stage. One puppet moved across the stage to the center, paused and bowed down (replicating the motion of the low-ranked puppet in Experiment 1). Then the other puppet crossed the stage, passing in front of the bowing puppet. The 'meet and retreat' sequence in Experiment 1 was taken out in Experiment 2. This sequence was repeated three times (See Figure 2).

Methods

Participants

Participants were recruited from the same children's museum using the same procedure as in Experiment 1. A total of 34 children participated in the experiment. Of these, 13 were excluded from the analysis for the following reasons: Choosing both puppets (n=7); Refusing to choose a puppet (n=3); Extreme fussiness (n=2); Experimenter error (the puppets were moved backward in the puppet show, n=1). The remaining 21 children (7 girls, 14 boys) contributed data to the analysis. Their age ranged from 21-31 months (M=26.24 months; SD=3.28 days). Parents were given the same demographic questionnaire as in Experiment 1. Responses to the race/ethnicity questions included the following: White, Not Hispanic (n=7); Asian, Not Hispanic (n=4); White, Mexican American (n=3); Asian and White, Not Hispanic (n=2); American Indian/ Alaska Native and Asian, Not Hispanic (n=1); Native Hawaiian/Other Pacific Islander, Not Hispanic (n=1); White, Hispanic (n=1). 2 participants declined to answer the race/ethnicity question.

Materials & Procedure: The puppets and stage were identical to those in Study 1. Procedures were the same as in Experiment 1 except for the following: During the familiarization sequence of the puppet show, both puppets moved across the stage in the same direction. During the Control-Display Sequence (analogous to the Hierarchy-Display Sequence in Experiment 1), the two puppets appeared on the same side of the stage. One puppet moved across the stage to the center, then bowed down and moved aside (replicating the motion of the low-ranked puppet in Experiment 1). Then the second puppet crossed the stage, passing in front of the bowing puppet. **(See Figure 2)**

Results and Discussion

Of the 21 children who contributed data to the analysis, 11 chose the puppet that completed its journey across the stage. To compare the observed distribution (i.e. 11/21) to an expected distribution (i.e. expected if the children chose the complete-crossing puppet 50% of the time), we utilized 'binom.test' in the statistical program 'R' (p=1, estimated probability of success = .524). We also used a Bayes Factor Binomial calculator (http://pcl.missouri.edu/bf-binomial) and found that odds of the null model (i.e. the odds that children were choosing the high-ranked 50% of the time) over the alternative model (that the children were choosing the high-ranked either more or less than 50% of the time) were 3.868, which is considered to be moderate evidence in favor of the null (Kass & Raftery, 1995).

General Discussion

Taken together, the results of these two experiments suggest that by 21 months, infants prefer novel high-ranking individuals. Results from Experiment 2 suggest that the children's choice of the dominant puppet in Experiment 1 could not be attributed to a preference for the puppet that accomplished its goal of crossing the stage, the puppet that remained visible throughout, the puppet that moved last or moved farther, or for the puppet that moved in the same way in the hierarchy-display sequence as it had during the familiarization sequence (i.e. remaining upright the whole time). When children saw a puppet show that maintained all of these features, but lacked the social interaction, which established the relative ranks of each puppet in Experiment 1, the preference for the puppet reaching its goal, disappeared. This suggests that the preference for the dominant puppet in Experiment 1 hinged on the fact that they directly interacted.

Taken together, this suggests that detecting relative rank does not only allow children to 'size up the competition' and avoid conflicts with dominant individuals. If this were the case, we would expect children to either avoid the high-ranking puppet, or to show no preference (if the children did not expect to have a conflict with the puppets).

There are at least two possible explanations for this preference. The first is that children may seek to affiliate with individuals who have more power and access to resources. In other words, humans may share the tendency previously shown in macaques (Silk, 1999) to seek alliances with higher-ranked individuals.

Another possible explanation for the preference, is that children may expect high-ranking individuals to act as leaders in what anthropologist Alan Fiske has called Authority Ranking relationships (Fiske & Haslam, 2005; Fiske & Rai, 2014; Fiske, 1992). Leaders in these types of relationships are motivated to guide and protect subordinates; who in turn are motivated to defer to leaders (and to punish other low-ranking individuals who don't defer). According to Fiske, subordinates *expect* leaders to provide this protection and guidance. Support for this view comes from a large body of ethnographic research that suggests human social hierarchy is often expressed through these mutually beneficial relationships between high- and low-ranked individuals.

When considering these results, a few points should be kept in mind. First, the puppets in the current experiment were much smaller than the children reaching for them. It is possible that children might avoid a dominant puppet if the puppets were as big as the children themselves, and further studies will investigate this possibility.

It is also worth noting that the children who participated in this study have likely had many positive experiences with high-ranking individuals such as parents, older siblings and other caregivers. Generalizing from these experiences, children may view higher-ranked individuals in a positive light, expecting them to provide protection and guidance. Further research could also explore whether children who have experienced abuse or neglect show the same preference, although such research presents many practical challenges. It is also worth noting that younger infants might not share this preference who have had less experience with high-ranking individuals. It could be that the 'default' preference is to avoid high-ranking individuals, but that through experience children form positive associations with them. Future studies will explore this possibility.

Finally, children (like adults) are complicated creatures, who can be fascinated by things that also scare them a little. Further studies are needed to establish whether children's preference for the 'dominant' puppet in the present study reflects unambiguously positive feelings (e.g., the expectation of guidance and protection), or some mixture of positive and negative (e.g., curiosity and fear).

In any case, it is interesting to compare the preference for high-ranking puppets shown in the present study to the established finding that infants prefer helpers over hinderers (Hamlin et al., 2007, 2011). In our study, one puppet bowed and moved aside which, while not actively helping, at least *allowed* the other puppet to reach its goal. (And by the same token, the non-bowing puppet could be seen as preventing or hindering the other from reaching its goal.) The findings by Hamlin et al. would seem to suggest that if children had looked at the scene primarily in terms of helping and hindering, they should have preferred the puppet who bowed (i.e. the helper). The fact that children actually showed a strong preference for the puppet who did not bow suggests that children did not view this social interaction primarily as a helping/hindering event. This may be because children view active helping (i.e., pushing another puppet up a hill, or opening a box for him) differently from passive helping (i.e., moving out of the way). Or it may be that social rank takes precedence over prosocial behavior. That is, children at this age may care more about affiliating with high-ranking individuals than with helpful ones.

The null result we found in Experiment 2 provides another point of contact with the broader literature. At least one study has shown that toddlers prefer competent individuals over non-competent individuals (Jara-Ettinger, Tenenbaum, & Schulz, 2015; see also Pasquini, Corriveau, Koenig, & Harris, 2007). This raises the question of why children did only preferred the puppet that crossed the stage in Experiment 1. One possibility is that in Experiment 2, where there was no meet-and-retreat sequence, and one puppet stopped halfway across the stage, children thought the puppet stopped of its own volition. Another possibility is that the two studies operationalized 'competence' differently enough to affect children's preferences. In the study by Jara-Ettinger et al., the puppets' goal was to play music. Children may have cared more about this goal than about seeing a puppet cross a stage in the present experiment. (That is, toddlers may only pay attention to competencies they care about, or may only prefer those individuals who are competent at doing something the toddler wants done.) In contrast, our experiment used the sparse goal of simply crossing the stage, which children presumably have no experience with. This makes it all the more striking that they preferred the high-ranking puppet in Experiment 1, since the puppets goal was both novel and irrelevant.

Future studies will investigate the effects that social rank, competence and prosocial behavior have on children's evaluations of individuals. For now, the present study demonstrates that very young children are not only sensitive to the relative social status of novel agents; but that they also prefer dominant ones.

References

Boehm, C. (1999). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Harvard University Press.

Bond, A. B., Wei, C. A., & Kamil, A. C. (2010). Cognitive representation in transitive inference: A comparison of four corvid species. *Behavioural Processes*, *85*(3), 283–292. http://doi.org/10.1016/j.beproc.2010.08.003

Chudek, M., Heller, S., Birch, S., & Henrich, J. (2012). Prestige-biased cultural learning: bystander's differential attention to potential models influences

children's learning. *Evolution and Human Behavior*, *33*(1), 46–56.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., Lengyel, M., Csibra, G., … Tatone, D. (2016). Developmental Psychology Statistical Treatment of Looking-Time Data Statistical Treatment of Looking-Time Data. http://doi.org/10.1037/dev0000083

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. http://doi.org/10.1177/1745691611406920

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689–723. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1454904

Fiske, A. P., & Haslam, N. (2005). The Four Basic Social Bonds. Structures for Coordinating Interaction. *Interpersonal Cognition*. Retrieved from http://books.google.com/books?hl=en&lr=&id=pbv35TUf1I8C&oi=fnd&pg=PP2&dq=Interpersonal+Cognition&ots=UJbUgC_qw-&sig=s90gD3hLIUszITnZDlZGV2JEIdI

Fiske, A. P., & Rai, T. S. (2014). *Virtuous Violence*. Cambridge University Press.

Fiske, A. P., & Schubert, L. (2012). How to relate to people: The extraterrestrial's guide to Homo sapiens. *Relationship Science: Integrating Evolutionary, Neuroscience, and Sociocultural Approaches. Decade of Behavior 2000–2010.*, 169–195. Retrieved from http://psycnet.apa.org/books/13489/009

Gazes, R. P., Hampton, R. R., & Lourenco, S. F. (2015). Transitive inference of social dominance by human infants. *Developmental Science*, n/a–n/a. http://doi.org/10.1111/desc.12367

Grosenick, L., Clement, T. S., & Fernald, R. D. (2007). Fish can infer social rank by observation alone. *Nature*, *445*(7126), 429–432. http://doi.org/10.1038/nature05646

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–9. http://doi.org/10.1038/nature06288

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 19931–6. http://doi.org/10.1073/pnas.1110306108

Hawley, P. H. (2015). Social dominance in childhood and its evolutionary underpinnings: why it matters and what we can do. *Pediatrics*, *135 Suppl*(Supplement), S31–8. http://doi.org/10.1542/peds.2014-3549D

Hawley, P. H., & Little, T. D. (1999). On winning some and losing some: A social relations approach to social dominance in toddlers. *Merrill-Palmer Quarterly*, *45*(2), 185 – 214.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196. http://doi.org/10.1016/S1090-5138(00)00071-4

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not So Innocent: Toddlers' Inferences About Costs and Culpability. *Psychological Science*, *26*(5), 633–640. http://doi.org/10.1177/0956797615572806

Kaufmann, L., & Clément, F. (2014). Wired for Society: Cognizing Pathways to Society and Culture. *Topoi*. http://doi.org/10.1007/s11245-014-9236-9

Magee, J. C., & Galinsky, A. D. (2008). Social Hierarchy: The Self Reinforcing Nature of Power and Status. *The Academy of Management Annals*, *2*(1), 351–398. http://doi.org/10.1080/19416520802211628

Mascaro, O., & Csibra, G. (2012). Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences*, *109*, 6862–6867. http://doi.org/10.1073/pnas.1113194109

Pasquini, E. S., Corriveau, K. H., Koenig, M. a, & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216–26. http://doi.org/10.1037/0012-1649.43.5.1216

Pun, A., Birch, S. A. J., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences*, 1–6. http://doi.org/10.1073/pnas.1514879113

Sapolsky, R. (2004). Social status and health in humans and other animals. *Annual Review of Anthropology*, *33*, 393–418. http://doi.org/10.1146/annurev.anthro.33.070203.144000

Sapolsky, R. (2005). The Influence of Social Hierarchy on Primate Health. *Science*, *308*(5722), 648–652. http://doi.org/10.1126/science.1106477

Silk, J. B. (1999). Male bonnet macaques use information about third-party rank relationships to recruit allies. *Animal Behaviour*, *58*(1), 45–51. http://doi.org/10.1006/anbe.1999.1129

Silk, J. B. (2007). The adaptive value of sociality in mammalian groups. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 539–559. http://doi.org/10.1098/rstb.2006.1994

Sluckin, A. M. (University of O., & Smith, P. K. (University of S. (1977). Two Approaches to the Concept of Dominance in Preschool Children. *Child Development*, *48*, 917–923. http://doi.org/10.2307/1128341

Thomsen, L., & Carey, S. (2013). Core Cognition of Social Relations. In *Navigating the Social World: What Infants, Children, and Other Species Can Teach Us* (p. 17). Oxford University Press.

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, *331*(6016), 26–29. =

Unifying Conflicting Perspectives in Group Activities: Roles of Minority Individuals

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University

Yugo Hayashi (yhayashi@fc.ritsumei.ac.jp)
School of Literature, Ritsumeikan University

Hitoshi Terai (teraihitoshi@gmail.com)
Faculty of Humanity-Oriented Science and Engineering, kinki University

Abstract

For drawing higher-level perspectives in group activities, resolving conflicts among group members is crucial. We investigated group activities with four members wherein one member had a different perspective from the other three. Four members engaged in a rule discovery task in which they were required to unify conflicts for the solution. Through two experiments, we investigated two hypotheses: 1) Innovative high-level perspectives are more likely to emerge from a minority individual than from the majority of group members, 2) Group members on the majority side might tend to have more egocentric perspectives than an individual on the minority side. Both hypotheses were supported.

Keywords: Emergence, Minority, Majority, Group problem solving.

Introduction

In group activities, conflicts often occur among group members. Indeed, such conflicts are important for bringing innovative ideas into a group by overcoming each member's fixed, old thoughts. A group found new ideas by producing a unified meta-perspective while resolving group members' contradictory local perspectives.

Many studies on divergent thinking have confirmed conflicts' important functions (Van Dyne & Saavedra, 1996; Dreu, 2002). In studies of creativity, many researches have indicated that unifying two conflicting concepts, which seem not to share any common properties and, therefore, are difficult to merge is important for finding new innovative ideas and designs (Finke, Ward, & Smith, 1992). Additionally, in insight problem solving, problem solvers' thought processes are tightly fixated by mental blocks that prevent them from achieving new solutions (Sternberg & Davidson, 1995). They face many unexpected counter examples that conflict with their viewpoints. To reach new solutions, they need to relax mental blocks and find a new perspective that resolves these conflicts. Similarly, the philosophy of science has indicated that anomalies are crucial for constructing new theories. Anomalies are defined as instances that break the symmetry of a classical theory, meaning that anomalies produce conflicts in current theory. Historical studies indicate cases of innovative scientific theory constructed through the resolution of such conflicts (Chinn & Brewer, 1993). Empirical studies also support the notionthat surprising results not predicted by a hypothesis play a crucial role in scientific discovery (Kulkarni & Simon, 1988).

Such conflicts more often emerge in a heterogeneous group with different perspectives. Preceding studies on pair problem solving have indicated that the important factor for finding new ideas is one member having a different perspective from the other (Shirouzu, Miyake, & Masukawa, 2002; Miwa, 2004). In group activities, organizational studies have indicated two types of conflicts: relationship and task conflicts (Simons & Peterson, 2000). Relationship conflicts come from differences in group members' personalities and task conflicts from differences in members' perspectives on and ideas for the task. Research has indicated that task conflict brings about positive effects by drawing out new solutions unifying the variable viewpoints that caused conflict (Tjosvold, Hui, Ding, & Hu, 2003). In such group activities, each member experiences difficulties in accepting other members' ideas and different perspectives. Such contradictions bring about conflicts in group interactions and improve interactive processes through which new innovative ideas emerge.

In this study, we investigated activities of groups comprising four members in which one individual had a different perspective from the other three members. The following is our research questions: In such group activities, with asymmetry among group members' opinions, does a high-level perspective emerge from a minority individual or from majority three members? Nemeth indicated that minority views foster greater thought about an issue, leading to divergent rather than convergent thinking (Nemeth & Wachtler, 1983; Nemeth, 1986). Meyers, et al., also confirmed that minorities produce significantly more arguable opinions, disagreement-relevant intrusions, and qualifiers than majority members do (Meyers, Brashers, & Hanner, 2000). These findings guide the following first hypothesis:

- Hypothesis 1: Innovative high-level perspectives are more likely to emerge from a minority individual than from group members on the majority side.

Conflicts are valuable, but also produce miscommunication and misunderstandings among group members, potentially

causing negative effects in group activities. Group members need to find a meta-perspective that unifies conflicted lower-level different perspectives. To produce a high-level perspective, each member must understand other members' perspectives in the initial stage of arguments. Studies in psycholinguistics have indicated that people tend to be biased by egocentrism (Keysar, Barr, Balin, & Brauner, 2000). Egocentric people are unable fully to understand other people's opinions and are attached to biased facts that can differ from reality. Previous research has consistently indicated that majorities generally exert a greater and more direct influence on group decisions that also leads to an unreflective acceptance of the majority position (Wood, Lundgren, Ouellette, Busceme, & Blackstone, 1994; Meyers et al., 2000). To overcome egocentric views, reflective and deliberate thinking is crucial. This determines our second hypothesis:

- Hypothesis 2: Group members on the majority side might tend to have more egocentric perspectives than an individual on the minority side.

This point, i.e., understanding other members' perspectives by escaping egocentric views, is crucial because the current study investigate group activities when very serious conflicts occur in a group. Many previous studies have also examined such activities. However, each member's misunderstandings of other group members in the current study were more serious compared to preceding studies. The conflicts that are investigated in the current study are characterized by incommensurability. An individual on one side had a completely different perspective that could not be understood by individuals on the other side. In the initial stage of discussion, arguments among them could not be resolved.

Further more, we investigate these group activities compared to collaborative activities by a pair of individuals. Specifically, if Hypothesis 1 is supported, we assume two possibilities: A minority individual's performance in finding a high-level perspective would be better than pair's performance as the baseline performance, or majority's performance would be worse than the baseline performance. No previous research was found, and we have no hypotheses about this research question.

Experimental design

We used a modified version of the experimental design developed by Hayashi and Miwa (2009), in which pairs of participants with different perspectives engaged in a rule discovery task.

How to produce conflicts

Figure 1 shows an example pair of two stimuli, one presented to a minority individual and the other to majority participants. The diagrammatic arrangements of "components" of both stimuli are identical within the dotted square, but one against the white background might be perceived differently from the other against the black background. In this case, one participant perceives seven black components (i.e., black perspective), but the other perceives five white components (i.e., white perspective), causing conflicts between the two participants. Participants were required to notice the other participant's perspective by finding the distinction of the background color. Before the experimental session, it was stressed that stimuli within the fixation (dotted) square were physically identical, but the background color outside the square was not mentioned. With this understanding of the experimental situation, participants could unify their contradictory perspectives and shift to a higher unified perspective that resolved the conflicts.

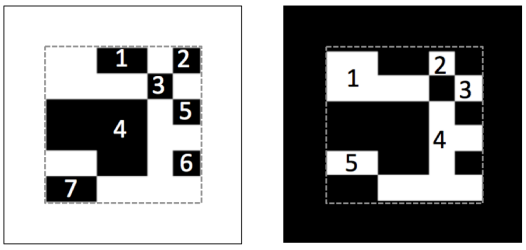


Figure 1: Example stimuli with two different perspectives.

Stimuli

In the present study, one participant was assigned to a minority role and the other three to a majority role. Either a white or a black perspective was given to each side by controlling the background color. Experimental stimuli were presented on a monitor placed in front of each participant. A workspace was provided for each member, separated from others by separator panels. Therefore, each member could not notice different background color on other members' monitors. Conversations among members were easily made.

For introducing a context that motivated participants to resolve conflicts, participants engaged in a rule discovery task. In the experiment, stimuli were presented sequentially (see Figure 2). For each stimulus, a square outer box was displayed for one second, followed by a stimulus picture presented inside the box frame. Participants were allowed to converse freely about the stimulus pattern presented on their monitor. When all members agreed to move to the next stimulus, or when 120 seconds had passed, another stimulus was presented.

Rule setting

Participants were required to find regularity in a sequential pattern of the number of components. In Figure 2, the pattern of the black components are 3, 4, and 5. The sequential rule does not exit in one color of components; regularity emerges across both colors of components. Specifically, Table 1 shows a stimuli presentation sequence in which regularity in a sequence of the sums of black and white components was manipulated (i.e., 6, 8, 10, and 12 in the bottom column).

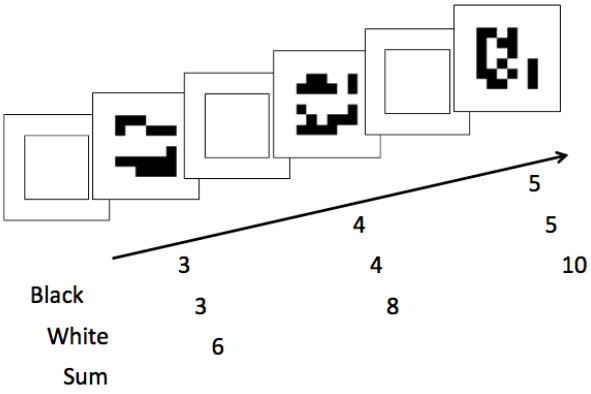


Figure 2: Series of stimuli presented on the black perspective side.

Table 1: Example sequences of the number of components.

	Introductory					Contradictory				
Black	3	4	5	6	...	2	2	6	5	...
White	3	4	5	6	...	4	6	4	7	...
Sum	6	8	10	12	...	6	8	10	12	...

Introductory phase: The introductory phase was established to let participants believe their own perspectives were correct. A sequence in white (or black) components was individually manipulated (i.e., 3, 4, 5, and 6) while regularity in a sequence of the sums of black and white components was maintained (i.e., 6, 8, 10, and 12). In the introductory phase, even though participants had different perspectives, no conflicts occurred because each continuously reported the identical sequence of the number (i.e., 3, 4, 5, and 6) to the others.

Conflict phase: After the 17th trial, regularity in the number of each color of components was broken: 2, 2, 6, 5 in black, and 4, 6, 4 and 7 in white, leading to conflicts among participants, but regularity in the sums of black and white components was still maintained (i.e., 6, 8, 10, and 12). The session was terminated after all members agreed on the regularity found in paired components (i.e., 6, 8, 10, and 12), or time for the problem-solving phase exceeded 60 minutes.

Post questionnaire

After the problem-solving phase, we conducted a questionnaire in which we asked participants why the conflicts in the problem-solving phase occurred. In particular, participants' understanding was tested on how precisely they understood the arrangement of components on other members' screens. An arrangement of components on their own screen was presented on the questionnaire sheet (Figure 3 (a)), and they were required to draw an arrangement of components from the other members' screens. The depicted figures were categorized into four types (Figure 3 (b)), each of which was characterized by background colors, and black and white assignment to components.

Previous studies found that typically, participants drew incorrect pictures as in Figure 3 (b2). In the picture, the background color was correctly identified but the colors of corresponding components, which should have been physically identical, were opposite. The picture contradicted the instruction that each diagrammatic pattern inside the fixation square for each participant was identical for the other members. This reversed figure was interpreted as evidence of participants' egocentric understanding.

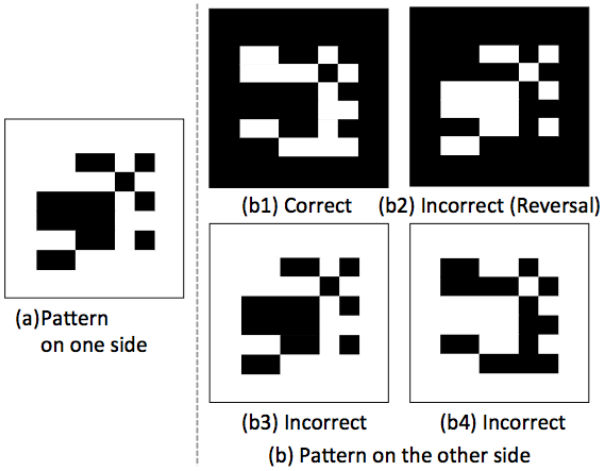


Figure 3: Four diagrammatic patterns depicted in post questionnaire.

In Experiment 1, we also recorded the participants' protocols. This secondary data will be examined in the discussion section.

Experiment 1

Participants

Thirty-eight undergraduates participated in Experiment 1. A total of 10 groups, i.e., eight groups consisting of four members (one as a minority individual and three as a majority group) and two groups consisting of three members (one as minority and two as majority), were constructed.

Results

Our main concern was whether the participants understood the other members' perspective for unifying and resolving the conflicts. Therefore, in the experiment, we did not record whether each individual found the sequential rule. Rules were not examined by individual members but rather discussed by groups. The result shows eight of the ten groups found the target rule: i.e., the sequential pattern of the sums of black and white components.

We scored the post questionnaire based on two criteria: first, whether the arrangement of components inside the fixation square was correctly drawn; second, whether the background color out of the fixation square was correctly drawn.

Table 2 shows the results. Fisher's exact analysis shows a marginally significant difference in the distribution between minority individuals and majority groups for the component color identification ($p < 0.10$) and a significant difference for the background color identification ($p < 0.05$).

The table shows that more individuals on the minority side drew correct figures in the questionnaire than did majority groups, thus confirming that individuals on the minority side understood the majorities' viewpoints and acquired a higher unified perspective that resolved the conflicts. This result supports our first hypothesis.

Table 2: Distribution of the number of participants who drew the correct and incorrect figures in Experiment 1. Parentheses show the ratios of participants.

	Components		Background	
	Correct	Incorrect	Correct	Incorrect
Minority	6 (.60)	4 (.40)	8 (.80)	2 (.20)
Majority	8 (.29)	20 (.71)	12 (.43)	16 (.57)

We also analyzed the degree to which the reversed figure (Figure 3 (b2)) was drawn on the questionnaire sheets. Table 3 shows the distribution of the number of participants who drew the reversed figure. Fisher's exact analysis shows a significant difference between minority individuals and majority groups in the distribution ($p < 0.05$). The table indicates that more majority members drew the reversed figure, thus implying that they tended to take egocentric viewpoints more often than did minority individuals. This result supports our second hypothesis.

Table 3: Distribution of the number of participants who drew the reversed figure in Experiment 1. Parentheses show the ratios of participants.

	Reversal	Others
Minority	1 (.10)	9 (.90)
Majority	15 (.54)	13 (.46)

Experiment 2

We conducted Experiment 2, adding a symmetric interaction condition in which a pair of participants solved the same experimental task as in Experiment 1. Additionally, we were interested in interaction processes in which the concern was whether the reversed figure, prominently observed in Experiment 1, increased or decreased from the initial stage to the final stage of problem solving. We expected that through the development of problem solving, misunderstandings would be gradually corrected, and then the reversed figure would decrease. If so, the reversed figure may function similarly as other incorrect figures do. However, if the reversed figure increased, they may have specific functions such as acting as a medium for bridging across misunderstanding to acquire the correct understanding.

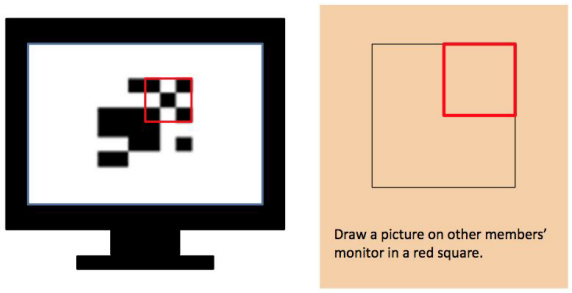
Participants

A total of 39 undergraduates were assigned to the asymmetry group conditions, in which nine groups consisting of four members (one as a minority individual and three as a majority group) and one group consisting of three members (one as minority and two as majority). Twenty undergraduates were assigned to the symmetry pair condition in which 10 pairs were constructed.

Procedures

The experimental procedures were almost identical to those in Experiment 1.

One procedure was added to capture participants' transition processes from misunderstandings to correctly noticing other members' perspective. Specifically, at the end of every trial (i.e., a set of the presentation of the outer square and the experimental stimulus), the participants were asked to answer an intermediate questionnaire. Figure 4 shows an example shot of the monitor and an experimental sheet from the intermediate questionnaire. A quarter area of the experimental stimulus was outlined within a red square. Participants were asked to draw the stimulus pattern in the same area of the other members' monitor on the experimental sheet.



(a) An example shot on participant's monitor (b) An example experimental sheet in intermediate questionnaire.

Figure 4: An example shot of the monitor screen and an experimental sheet from the intermediate questionnaire.

Results

In Experiment 2, eight of ten pairs in the symmetry pair condition found the target rule, but only three of ten groups in the asymmetry group condition did ($p < 0.05$: Fisher's exact analysis).

Table 4 shows the distribution of participants who drew the correct and incorrect figures in the post questionnaire. A chi-square test shows a significant difference in the distribution for the component color identification ($\chi^2(2) = 15.61$, $p < 0.01$), but does not for the background color identification

($\chi^2(2) = 4.47$, n.s.). We did not confirm that the performance of minority individuals is superior to that of pairs; on the other hand, the performance of majority groups is much worse than that of pairs.

Additionally, Table 5 shows the distribution of participants who drew the reversed figure in the post questionnaire. A chi-square test shows a significant difference in the distribution ($\chi^2(2) = 13.86$, $p < 0.01$). The result indicates that more majority members drew the reversed figure than pairs did. However, the results do not indicate that fewer minority individuals drew the reversed figure more than pairs did.

Table 4: Distribution of the number of participants who drew the correct and incorrect figures in Experiment 2. Parentheses show the ratios of participants.

	Components		Background	
	Correct	Incorrect	Correct	Incorrect
Pairs	17	3	17	3
	(.85)	(.15)	(.85)	(.15)
Minority	5	5	8	2
	(.50)	(.50)	(.80)	(.20)
Majority	8	21	17	12
	(.28)	(.72)	(.59)	(.41)

Table 5: Distribution of the number of participants who drew the reversed figure in Experiment 2. Parentheses show the ratios of participants.

	Reversal	Others
Pairs	3 (.15)	17 (.85)
Minority	5 (.50)	5 (.50)
Majority	20 (.69)	9 (.31)

Both results consistently confirm that majority group members were more egocentric and less accurate at understanding other members' perspective, but do not support that minority individuals were more sophisticated at unifying the multiple perspectives than were pairs as the baseline condition.

Next, we analyzed the intermediate questionnaire for capturing the transition process of correcting the misunderstandings. We divided the trials from the initial through final wherein the participants terminated the experiment into three stages: initial, middle, and final. Figures 5 (a) and (b) show the average rates of the reversed and other incorrect figures in the initial, middle, and final stages of problem solving.

First, for incorrect figures other than the reversed figure in Figure 5 (b), there were significant main effects of both the stage and the condition factors ($F(2, 112) = 23.06$, $p < 0.01$; $F(2, 56) = 6.92$, $p < 0.01$), but no significant interaction ($F(4, 112) < 1$, n.s.). The gradual decrease of the ratio of incorrect figures is reasonable because it means that misunderstandings were progressively corrected.

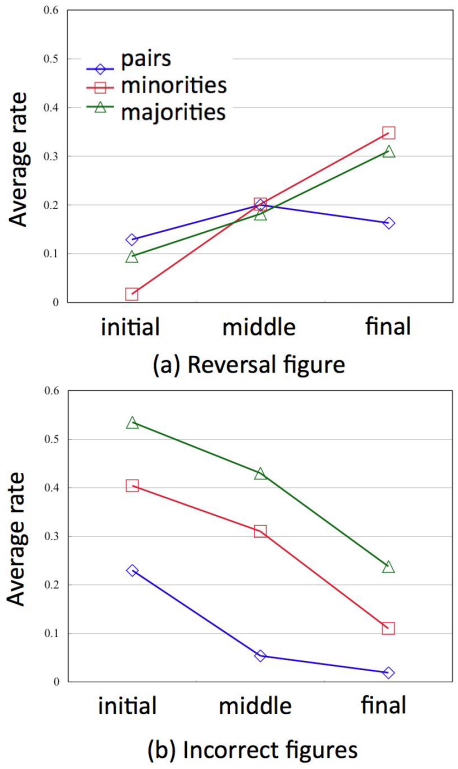


Figure 5: Transition of occurrence rates of reversed and other types of incorrect figures.

On the other hand, the gradual increase of the reversed figure is interesting. In Figure 5 (a), there was a significant main effect of the stage factor ($F(2, 112) = 7.52$, $p < 0.01$), but neither a significant main effect of the condition factor nor an interaction between the two factors ($F(2, 56) < 1$, n.s.; $F(4, 112) = 1.62$, n.s.). This result shows that the reversed figure gradually increased in the development of problem-solving. This point is discussed in detail later.

Discussion and conclusions

Overall results support our two hypotheses. Experiment 1 supported that minority individuals noticed a unified perspective that resolved the conflicts occurring among group members more often than did majority group members. Majorities tended to misunderstand the causes of the conflicts due to their egocentric viewpoint. These findings are consistent with preceding minority and majority studies (Nemeth & Wachtler, 1983; Nemeth, 1986; Wood et al., 1994; Meyers et al., 2000). Previous studies focused mainly on judgment patterns of third parties who were exposed to minority or majority opinions. Meanwhile, in the current study, by employing the deliberate experimental paradigm, we investigated how minorities and majorities formed their ideas through their interaction processes.

In Experiment 2, majorities showed poorer performances at understanding other members' perspective compared with

the pairs in the baseline condition. Studies in social psychology have indicated negative impacts of group activities such as the Ringelmann effect and social loafing (Karau & Williams, 1993). These phenomena indicate losses of motivation and coordination of group members, which in turn diminish the total performance of a group. These aspects that emerge in a homogeneous group may bring about negative impacts on members on the majority side.

In our experiments, the participants engaged in dual tasks. One task was to explore the reasons for conflicts and resolve them by unifying the conflicted multiple perspectives. In the second task, the participants also engaged in a rule discovery task for finding a regularity of the number of components in the experimental stimuli.

As we mentioned before, in Experiment 2, the performance of the asymmetry group in finding the regularity was much poorer that of the symmetry pair. In the seven unsuccessful groups in the asymmetry group condition, a total of twenty participants were assigned on the majority side, revealing very poor performance for noticing the minority individual's perspective. Specifically, only five of twenty participants noticed the background color of the minority perspective, and ten of twenty correctly draw the diagrammatic pattern of components.

Experiment 2 confirmed that with the development of problem solving, the participants gradually reported the reversed figure, meaning that the reversed figure was generated through group interaction processes. In both experiments, majorities drew more reversed figures. Through what processes did they misunderstand causes of the conflicts? To understand this point in more detail, we transcribed participants' verbal protocols recorded in Experiment 1. We focused on a set of terms relating to "reversed" such as invert, opposite, reverse, and turn over. These terms may be interpreted differently based on each problem-solving context. In correctly understanding situations, these terms may be understood as the figure/ground reversal of components caused by the change of the background color. However, in misunderstanding situations, these terms were construed wrongly as simple reversals of colors of the components, white to black and black to white, thus, causing the reversed figure. As a result of analysis, 13 out of 20 members on the majority side mentioned these terms, and 77 % (10 out of 13) of those drew the reversed figure. However, none of five minority individuals who mentioned those terms drew the reversed figure. We assume that the latter interpretation, i.e., simple changes of colors of the components, was more intuitive without deep consideration; whereas, the former, i.e., the figure/ground reversal by the background color, was more sophisticated and required deliberate thoughts and deeper understanding about causes of conflicts. Majorities who tend to be biased by their egocentric view may quit their deliberate thinking to correctly understand conflicting situations.

References

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. , *63*(1), 1-49.

Dreu, C. K. W. D. (2002). Team innovation and team effectiveness: The importance of minority dissent and reflexivity. *European Journal of Work and Organizational Psychology*, *11*(3), 285-298.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. MIT press Cambridge, MA.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, *65*(4), 681.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), pp. 32-38.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science*, *12*(2), 139–175.

Meyers, R. A., Brashers, D. E., & Hanner, J. (2000). Majority-minority influence: identifying argumentative patterns and predicting argument-outcome links. *Journal of Communication*, *50*(4), 3–30.

Miwa, K. (2004). Collaborative discovery in a simple reasoning task. *Cognitive Systems Research*, *5*(1), 41 - 62.

Nemeth, C. J. (1986). Differential contributions of majority and minority influence. *Psychological Review*, *93*(1), 23.

Nemeth, C. J., & Wachtler, J. (1983). Creative problem solving as a result of majority vs minority influence. *European Journal of Social Psychology*, *13*(1), 45–55.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*, *26*(4), 469 - 501.

Simons, T. L., & Peterson, R. S. (2000). Task conflict and relationship conflict in top management teams: The pivotal role of intragroup trust. *Journal of Applied Psychology*, *85*(1), 102 - 111.

Sternberg, R. J., & Davidson, J. E. (1995). *The nature of insight*. The MIT Press.

Tjosvold, D., Hui, C., Ding, D. Z., & Hu, J. (2003). Conflict values and team relationships: conflict's contribution to team effectiveness and citizenship in china. *Journal of Organizational Behavior*, *24*(1), 69–88.

Van Dyne, L., & Saavedra, R. (1996). A naturalistic minority influence experiment: Effects on divergent thinking, conflict and originality in work-groups. *British Journal of Social Psychology*, *35*(1), 151–167.

Wood, W., Lundgren, S., Ouellette, J. A., Busceme, S., & Blackstone, T. (1994). Minority influence: a meta-analytic review of social influence processes. *Psychological bulletin*, *115*(3), 323-345.

Blink durations reflect mind wandering during reading

Stephanie Huette (shuette@memphis.edu), Ariel Mathis, & Art Graesser
University of Memphis
Department of Psychology and the Institute for Intelligent Systems
369 Psychology Bldg
Memphis, TN 38152 USA

Abstract

Mind wandering is a prevalent but highly subjective phenomenon that is difficult to measure. Typically studies use probes at random points throughout at study that pop in and ask participants "Are you mind wandering" where they indicate yes or no, and then resume the study. This study investigated a method of extracting eye blinks from raw eye tracking data while participants were reading texts that varied in degree of engagingness on a similar topic. Blink durations were found to increase for less engaging texts. We hypothesize that eye blink durations may increase with mind wandering and discuss implications for mind wandering research.

Keywords: mind wandering, reading, eye tracking, consciousness

Introduction

Mind wandering is considered a state of moving from one thought to another internally, uncoupled from external stimuli (Smallwood & Schooler, 2006). There are task related thoughts (e.g. thinking about one's own experiences with making pasta while in a cooking class) and task unrelated thoughts (e.g. thinking about what to eat this evening while in a physics lecture). This is classically a difficult phenomenon to study, because we cannot directly observe the process of mind wandering unfolding, nor understand the contents of another's mind without directly asking about them. The process of asking participants to reflect on their mind wandering disrupts the process itself, interrupting the mind wandering and perhaps perturbing subsequent processing (Smallwood & Schooler, 2006). The focus of the present research is identifying a potential covert measurement of mind wandering that might globally index how engaged or disengaged a participant is with a given task.

Without focus on external stimuli we cannot process or learn new material from reading or listening nor can we do goal-oriented interactive tasks, all of which require attention, perceptual cognitive processing, and in some cases, action. This makes mind wandering important to understanding data from all domains of psychological research, because participants coming into a laboratory for many psychological experiment are likely to mind wander at some point during the task. However, mind wandering is not purely negative in its effects. More recently it has shown to play a role in aiding creative problem solving as well as autobiographical planning (see Mooneyham & Schooler, 2013 for a review). From a biological perspective it actually serves a functional role of resting the mind, even though it involves some of the same brain areas that are involved in tasks that require attention and goal-orienting.

Default Mode Network

The default mode network is hypothesized to be a network used when the mind is in a "resting state", supported by imaging research examining cortical regions involved in mind wandering (Mason, Norton, Horn, Wegner, Grafton & Macrae, 2007). Importantly the default mode network has been shown to be active during mind wandering using fMRI measurements, demonstrating a strong link between the phenomena of mind wandering and an underlying neural state (Christoff, Gordon, Smallwood, Smith & Schooler, 2009). This is a kind of auto-pilot the brain can use when it is not closely coupled to a stimulus, allowing the mind to process information internally. Allowing for a brief period of rest may allow subsequent performance on attention and focus-driven tasks to improve performance. The default mode network is active during many phenomena including mind wandering, daydreaming, thinking about the self, or thinking about the past or planning for the future (Spreng & Grady, 2010; Qin & Northoff, 2011). While the significance and meaning of the default mode network is a matter of debate (see Raichle, 2015), the literature provides hints into more covert measures of mind wandering that may be useful in detecting when a person becomes disengaged from a task. In the present work, we will demonstrate eye blink durations could be an indication the mind is wandering more often when reading difficult texts.

One recent study demonstrated eye blinks activating the default mode network temporarily (Nakano, Kato, Morito, Itoi & Kitazawa, 2013). Critically, this is not due to a temporary lack of stimulus from the eye closing. A condition where the video turned black for the duration of an eye blink did not produce the same pattern of activation as eye blinks. Thus, it can be inferred that blinks are linked to triggering the mind's resting state. When there is a task that involves close attention such as reading, the rate of blinking tends to decrease. Blinking is an underutilized resource for measuring engagement in a task. Part of this paper is to provide guidelines for extracting blinks from a continuous time series of eye movements.

Reading & Mind Wandering

Studies investigating text difficulty and mind wandering have been mixed. Self-report measures from people reading passages varying in difficulty demonstrated no differences in amount of mind wandering. Instead, differences in mind wandering were found in how interested the participant was in the topic (Giambra & Grodsky, 1989; Grodsky & Giambra,

1990). In more recent research, when text difficulty was varied (Feng, D'Mello, & Graesser, 2013), Difficult texts positively correlated with more self-reported mind-wandering. This is inconsistent with prior work that predicts simpler tasks to induce mind wandering (Smallwood & Schooler, 2006), presumably because there are more cognitive resources available for mind wandering. However, both overly easy and overly difficult tasks could both induce this kind of mind wandering. For the purposes of the present preliminary work, we are assuming there to be a correlation between mind wandering and how engaging a text is to read, in line with the Feng et al. findings (2013) where more mind wandering was correlated with more difficult to read texts. The second important link to the present work is between blinks and mind-wandering, which appears to be well-established.

Finding an observable behavioral characteristic of mind-wandering is a challenge. People are able to read entire pages of text while moving their eyes across the page while thinking about what they want to eat for dinner, making it difficult to predict or observe mind wandering without directly asking someone. Blinks have been investigated as an indicator of mind wandering. Smilek, Carriere, & Cheyne (2010) reported that blink frequency (blinks/sec) increased in periods prior to self-reported mind wandering during reading. However, this difference relied on separating the data according to self-report, whereas text difficulty was not varied.

Schad, Nuthmann, and Engbert (2012) proposed an attentional decoupling, finding that eye movements were predictive of overlooking text errors at different levels of language processing. The first level to be lost was the semantic/thematic information from text, followed by syntactic and lexical information, indicating mind wandering is not an on/off state of mind. This makes levels of language lost and degrees of mind wandering an important topic to pursue to understand what most impacts text comprehension and later recall (see also Uzzaman & Joordens, 2011). Other studies have also found that mind wandering is not a discrete state rather a gradual decoupling from external stimuli (Singh & Fawcett, 2008).

Study Motivation

Because prior research strongly points to the duration of blinks being linked to the default mode network, we hypothesized blink duration may also vary as a function of how engaging a text passage is during reading. This account is compatible with both the executive-resource hypothesis (Smallwood & Schooler, 2006) and a control-failure hypothesis (McVay & Kane, 2010), which both posit cognitive capabilities as important aspects to when and how mind wandering occurs. The present research asks if we can detect, using a covert measurement technique without probes, whether someone is mind wandering, and to what degree they are mind wandering? There is little research focusing on the temporal dynamics of how the mind transitions from a state of tight attentional coupling to the external stimulus to the internal train of thought that is detached from the external stimulus. This is phenomenologically one of the most interesting aspects of the experience of mind wandering, where in one moment you are listening to a lecture, and the next you realize you've missed the last ten minutes while thinking internally about the day's events, or future plans for the evening. Being able to detect when someone is about to begin mind wanderingis becoming of interest to more researchers and has been steadily progressing in recent years (Blanchard, Bixler, Joyce, & D'Mello, 2014; Bixler & D'Mello, 2014).

The goal of the current research was twofold: (1) to observe whether stimuli known to be less *engaging to different degrees* would produce blink duration differences, and (2) to observe this without the use of explicit probes. One of the challenges of mind wandering research is participants must be asked "Were you mind wandering?" at random points during experiments. While this approach typically yields interpretable results it poses problems that are readily acknowledged in the literature (Smallwood & Schooler, 2015). The first is it interrupts processing, perhaps perturbing subsequent mind wandering by bringing awareness and focus to the mind wandering aspect of the study. The second is participants may be engaging in mind wandering outside of their own awareness, or disengage to various degrees as a function of how engaging the current stimuli are. Asking if participants are mind wandering also does not reveal how deeply they are mind wandering. The blink frequency study of Smilek et al. (2009) used self-report, whereas we instead are relying on text characteristics to induce more or less mind-wandering. That is, mind wandering is predicted to decrease as a function of text difficulty, as we have shown in previous studies for these texts. We don't know whether or not participants in this particular study were indeed mind wandering but we rely on the established relation with text difficulty in this sample. Again, relying on self-report and probes also has its downsides in that it disrupts processing. Here we can rely on the assumption that more boring texts will induce more mind wandering (Feng et al., 2013).

Thus, instead of using a binary measurement of yes/no, eye blink duration affords a continuous measurement that could be sensitive to a spectrum of boring to engaging stimuli of all kinds. It also addresses the problem of interrupting processing because no overt response or knowledge of the study being related to mind wandering is required in order to collect eye movement data. This study may also help to address whether mind wandering occurs more frequently with engaging texts or with boring texts, helping to resolve conflicting results in prior studies.

This study includes one experiment and a description of methods on how eye blinks were extracted from a Tobii X2 system post-hoc. Eye trackers do not directly report whether samples are lost because of blinks, inaccuracy in calibration, head turns, or off-screen fixations. This poses a difficult computational problem for extracting blinks, and this algorithm could be used with any binocular eye tracking system. To determine a blink with high confidence, we used

only timepoints where both eye tracks were lost. We also used a reasonable duration of consecutive missing samples, because although eye blink duration has individual variability it is a relatively stable range. Single eye track loss is likely due to track loss from calibration error, and both eye samples being missing is likely to be looks off-screen. .

Educational outcomes are fundamentally tied to text comprehension and thus mind-wandering during reading is the focus of the present work. Three texts were given to all participants in randomized order, with similar thematic content and varying in degree of engagingness. We hypothesized blink durations would increase on the texts that were less engaging, regardless of the order in which they are read (which controls for fatigue effects). If the default mode network is activated with blinks, and less engaging texts induce more mind wandering, the duration of blinks will increase with less engaging texts, and decrease monotonically as text becomes more engaging.

Method

Participants

Thirty-nine undergraduate student participated in exchange for course credit in an undergraduate psychology course. Participants were right handed native English speakers with corrected to normal or normal vision by self-reported criteria. The study was not advertised as a mind wandering study to prevent self-selection.

Eleven participants were excluded from analysis due to low track quality, and one participant was excluded due to not finishing the experiment. Twenty-seven participants were included in the final analysis.

Design

The independent variable was text engagingness (3 levels). The data collected was a continuous stream of x,y pixel coordinates, sample presence or absence as indicated by the Tobii X2 system, as well as other standard data output from an eye tracker. The dependent measure to be derived was the average blink duration (in ms) per participant per condition. Other tasks were done after the reading, but this does not interfere with the results of the reading portion which is the only task to be discussed here.

Materials. The experiment was built using Tobii Studio and a Tobii X2-60 eye tracker was used to track eye movements. The experiment consisted of two segments: a blank screen and selected readings. Order of task type was kept constant, however the three levels within the reading task type were counterbalanced. The blank screen consisted of a dark gray screen which was used to gauge gaze behaviors in the absence of outside stimuli.

Three readings were selected to represent varying levels of engagement: high, medium, and low. Readings were historical in nature with topics including a selection of the United States Constitution, a reading on law and punishment in Plymouth colony, and an article on the Salem witch trials. Readings were presented one page at a time and were between seven and eight pages long.

Fixation points consisting of a white screen with a black and white target at the center were presented between tasks. Participants advanced past these by pressing the space bar on the keyboard.

Procedure. Participants were seated in a stationary chair at a Tobii X2-60 eye tracker. The research assistant briefly outlined the experiment and informed participants to use the space bar to advance after written instructions, fixation points, and through the readings. Participants were further instructed to hold their heads steady during the experiment and to not stare off screen. The eye tracker was then calibrated and the experiment was started. Written instructions indicating where the space bar should be used as well as the presence of the blank screen as the first task were shown on screen. Participants then stared at a blank screen for three minutes before the experiment automatically advanced.

The participants were next instructed to read through three selections and to press space bar to advance to each page. Readings were counterbalanced to control for order effects. Participants were not able to backtrack through the readings once they had advanced to the next page.

Data Processing. Both eyes were used and matched at each timepoint to see if one or both samples were present or missing. Timepoints where both eye tracks were lost, or one or both were present were calculated. If both eye samples were missing, these time points were counted as ones, and present samples were counted as zeros. The position of the eye is not relevant to the current analysis and so what remained was a vector indicating track loss for both eyes, continuously over the course of reading, linked with vectors of participant information and the text being read. Only samples where the participant was reading were used (e.g. instructions and calibration/drift correction screens were not used). The data was filtered so that only samples with both eyes missing for between 100 and 1000ms were counted. Prior research has demonstrated blink durations exhibit many individual differences but are within this range (Martens, Munneke, Smid & Johnson, 2006). For this reason a large range was chosen that is most likely to be a blink, and not sample loss due to track inaccuracies, head turns, or eyes closing deliberately for a longer duration.

Results. The text conditions were Low engagement, Medium engagement, and High engagement. Blink durations were an average of 320ms for Low engagement (StDev: 185ms), 258ms for Medium engagement (StDev: 94ms) and 224ms for High engagement (StDev: 51ms). A repeated measures ANOVA found a significant effect of text engagement level ($F(2,50)=4.6$, $p=.01$). This was in the direction of the more engaging the text, the shorter the blink durations. The results appear to be fairly consistent across participants, as seen in Figure 1.

Discussion

This study tracked eye movements over the course of reading three similar texts that varied in how engaging they were. We developed an algorithm to extract blinks from raw data, and found a relationship between blink duration and text engagingness, where longer blink durations were observed in less engaging texts. We hypothesize this reflects Default Mode Network activation, the brain areas and connections thought to be responsible for mind wandering.

However, this is very preliminary research that could have multiple interpretations. If blinks functionally trigger mind wandering as was observed in Nakano et al.'s work (2013), it could be that an increase in blinks slowly begin moving the mind deeper into the mind wandering state in a continuous fashion. That is to say, the reader may not be blinking because they are mind wandering, but instead, are mind wandering because they are blinking. While the current work cannot disentangle the directionality of what may be causing what, it is not impossible that blinks may be what allow the mind to wander more, and that the brain can anticipate this short rest period before the blink has even occurred (perhaps during the motor planning phase, Default Mode Network areas begin to activate). These are intriguing possibilities, and ones that might be addressed in follow-up research by incorporating imaging and EEG in conjunction with eye tracking.

Limitations

There are many limitations to the current work we readily acknowledge. The first is that we don't know if a given participant was mind wandering. It could be that some individuals did not mind wander and were very engaged in all the readings, while others were mind wandering for each passage. Follow-up research that includes probes or debriefing questions to measure mind wandering are necessary, and at this point we are assuming blink durations reflect mind wandering, when it could in fact be task difficulty, syntactic or lexical processing differences, or even something like word concreteness.

Other studies have quantified text characteristics using a tool called Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004). This could potentially be used to control for text characteristics and various measures of text cohesion that

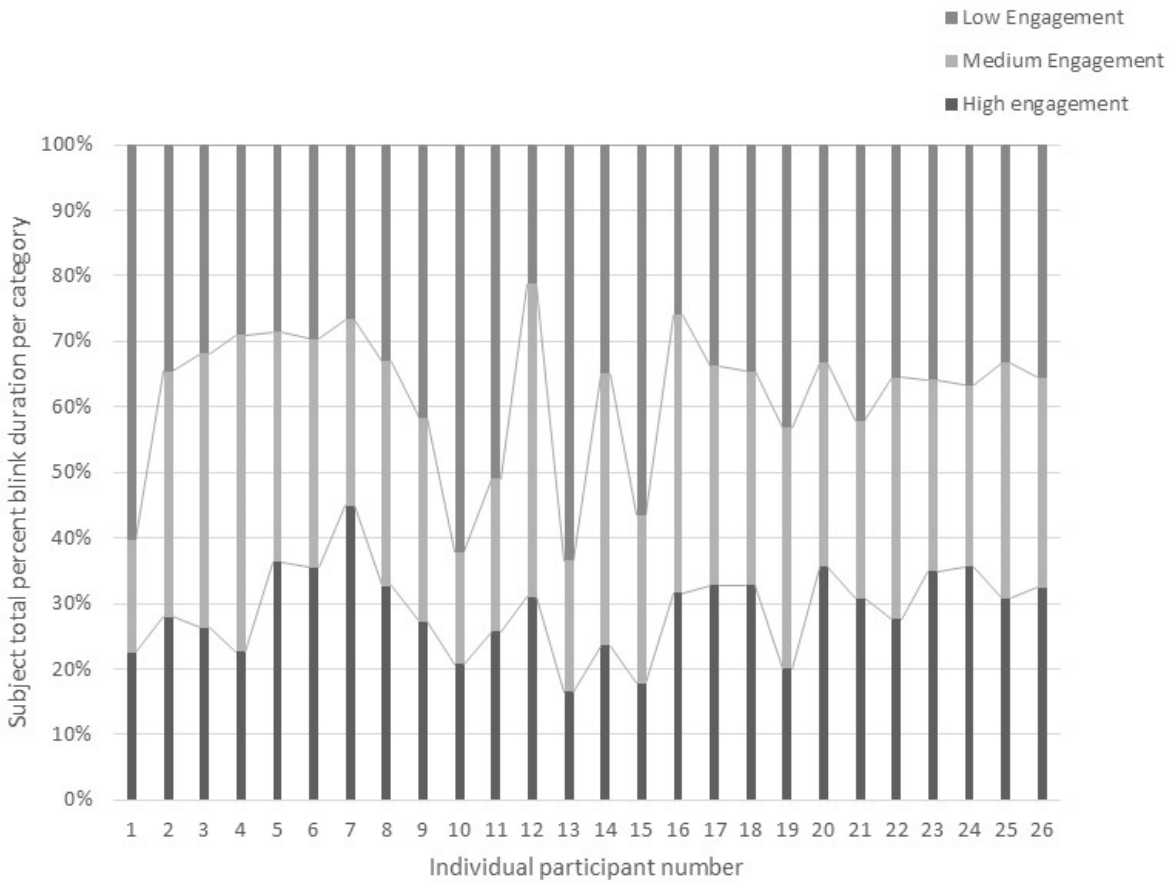


Figure 1 *Individual subjects blink duration by condition, as a total of all conditions.* Percentages are based on the individual's total average blink durations per condition. This shows the consistency of the effect, where it appears that most people show shorter blink durations for high engagement text, with few exceptions.

account for thematic language processing as well as other variables. It could be these characteristics that prompt one's mind to begin wandering, but this should be tested systematically.

The connection here between mind wandering and blink duration is tenuous but the data warrants further investigation. The literature and evidence showing a systematic difference between texts in this way does point toward this being an effect of mind wandering, but like all psychological measures, we have indirect evidence and more replication and validation of this paradigm will be required to establish this as an informative measure of engagement and mind wandering. We consider this first study to be observational rather than predictive, meaning

Conclusion

Despite many limitations, we have observed a connection between text engagingness and blink durations, meaning this is potentially a very useful measurement for future research in reading and mind wandering, and has the potential to quantify individual differences or may even be predictive of mind wandering starting to occur. With eye tracking technology becoming cheaper and more reliable, this could be used in conjunction with intelligent tutoring systems to help bring student's engagement back to the material. This could be generalizable to other stimuli and tasks as well, or even experiments without a task (such as passive listening to a story i.e. Huette, Winter, Matlock, Ardell & Spivey, 2014). This is the current direction of this research to generalize this to other tasks varying in the degree of engagingness, to help corroborate the current findings of differences in reading.

Acknowledgements

This work was supported by NSF grant # 1108845 to Sidney D'Mello and Art Graesser. Thank you to Sidney D'Mello and Andrew Olney for their helpful comments on this work.

References

Bixler, R. & D'Mello, S. (2014).Toward fully automated person-independent detection of mind wandering. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, G. J. Houben (Eds.), *Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization* (pp. 37-48). Switzerland: Springer International Publishing.

Blanchard, N.,Bixler, R., Joyce, T., & D'Mello, S. (2014). Automated physiological-based detection of mind wandering during learning. In *Intelligent Tutoring Systems* (pp.55-60). Springer International Publishing.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106(21),* 8719-8724.

Feng, S., D'Mello, S. K., & Graesser, A. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20*(1), 586-592.

Giambra, L. M. & Grodsky, A. (1989). Task-unrelated images and thoughts while reading. In *Imagery*. (pp. 26-31). Springer US.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers, 36*(2), 193-202.

Grodsky, A. & Giambra, L. M. (1990). The consistency across vigilance and reading tasks of individual differences in the occurrence of task-unrelated and task-related images and thoughts. *Imagination, Cognition and Personality, 10*(1), 39-52.

Huette, S., Winter, B., Matlock, T., Ardell, D. H., & Spivey, M. (2014). Eye movements during listening reveal spontaneous grammatical processing. *Frontiers in psychology*.

Martens, S., Munneke, J., Smid, H., & Johnson, A. (2006). Quick minds don't blink: Electrophysiological correlates of individual differences in attentional selection. *Journal of Cognitive Neuroscience, 18*(9), 1423-1438.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Response to Comment on" Wandering Minds: The Default Network and Stimulus-Independent Thought". *Science, 317(5834)*, 43-43.

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin, 136(2),* 188-197.

Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/ Revue canadienne de psychologie expérimentale, 67*(1), 11-18.

Nakano, T., Kato, M., Morito, Y., Itoi, S., & Kitazawa, S. (2013). Blink-related momentary activation of the default mode network while viewing videos. Proceedings of the National Academy of Sciences, *110(2),* 702-706.

Qin, P., & Nordhoff, G. (2011). How is our self related to midline regions and the default-mode network?. *Neuroimage, 57*(3), 1221-1233.

Raichle, M. E. (2015). The brain's default mode network. *Annual review of neuroscience, 38*, 433-447.

Schad, D. J., Nuthmann, A., & Engbert, R. (2012). Your mind wanders weakly, your mind wanders deeply: objective measures reveal mindless reading at different levels. *Cognition, 125*(2), 179-194.

Singh, K. D., & Fawcett, I. P. (2008). Transient and linearly graded deactivation of the human default-mode network by a visual detection task. *Neuroimage, 41*(1), 100-112.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigation the stream of

consciousness. *Annual review of psychology, 66*, 946-948.

Smilek, D., Carriere, J. S., & Cheyne, J. A. (2010). Out of mind, out of sight eye blinking as indicator and embodiment of mind wandering. *Psychological Science, 21(6)*, 786-789.

Spreng, R. N., & Grady, C. L. (2010). Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. *Journal of cognitive neuroscience, 22*(6), 1112-1123.

Uzzaman, S., & Joordens, S. (2011). The eyes know what you are thinking: Eye movements as an objective measure of mind wandering. *Consciousness and cognition, 20*(4), 1882-1886.

The Influence of Reputation Concerns and Social Biases on Children's Sharing Behavior

Haleh Yazdi (hyazdi@ucsd.edu)
UCSD Dept. of Psychology, 9500 Gilman Drive
La Jolla, CA 92093 USA

David Barner (dbarner@ucsd.edu)
UCSD Dept. of Psychology, 9500 Gilman Drive
La Jolla, CA 92093 USA

Gail D. Heyman (gheyman@ucsd.edu)
UCSD Dept. of Psychology, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

The present research builds on prior work on the social-contextual nature of children's generosity by systematically examining both observer effects and whether the recipient is an in-group or out-group member. Although previous research has examined these factors independently, no study to date has examined them in conjunction. We also extend prior research by including both measures of sharing behavior and children's evaluations of sharing scenarios, and by investigating a larger sample (N=164) with a broader age range than is typical of prior research (5- to 9-year-olds). We found that, across the entire age range tested, children were generous when observed and gave more to in-group members than out-group members, and that there was no interaction between these effects. We also found that children's own sharing behavior predicted their evaluations of sharing scenarios, with children rating in-group sharing as "nicer" than out-group sharing.

Keywords: sharing; prosocial behavior; in-group/out-group; reputation concerns; observer effects

Introduction

Humans share with one another for many reasons, including an expectation that others may reciprocate, and, in some circumstances, the knowledge that giving may enhance their own reputation among peers (e.g., when giving occurs publicly). Previous work suggests that children are sensitive to these factors relatively early in life, giving more to people who are similar to them, and making efforts to manage their reputations by acting more generously in the presence of onlookers. However, these studies have not explored how children weigh factors during individual acts of sharing, and whether, for example, they consider their reputation less when giving to people closer to them (e.g., friends, family, or other in-group members), or instead consider their reputation equally whether giving to strangers or to closer acquaintances. We look at this issue from a behavioral perspective, examining children's decisions about sharing their own resources in the presence or absence of an observer, and from a social evaluative perspective, examining how children evaluate the decisions others make about sharing resources.

A growing body of research demonstrates that, beginning early in life, children behave in a cooperative manner (Hay, 1979; Hay & Murray, 1979; Warneken & Tomasello, 2007) and become increasingly concerned with upholding social norms regarding fairness as they get older. Whereas young children (3- to 5-year-olds) verbally express strong preferences for equal outcomes (Ng, Heyman & Barner, 2011) and react emotionally to inequity (Geraci & Surian, 2011; Sutter, 2007), older children (7- to 8-year-olds) are more likely to display egalitarian behavior when allocating resources (LoBue, Nishida, Chiong, DeLoache & Haidt, 2011; Smith, Blake & Harris, 2013), particularly in situations where it is costly to do so (Blake & McAuliffe, 2011). This developmental pattern of stronger inequity aversion in later years may arise from a growing motivation for fairness and attention to the needs of others, or it may reflect a concern with displaying oneself as a cooperative and useful social partner (Warneken & Tomasello, 2009). In the latter case, older children may differ from younger children in that they are simply better at strategically displaying cooperative behavior (Shaw et al., 2014).

Numerous studies show that children, like adults, are more cooperative when they believe that their behaviors will be made known to others versus anonymously (Aloise-Young, 1993; Banerjee, 2002; Engelmann, Over, Herrmann & Tomasello, 2013; Harbaugh & Krause, 2000; Milinski, Semmann & Krambeck, 2002). Children steal less and help more in the presence of a peer observer (Engelmann at al., 2013), and are more generous with their resources when others are aware of their actions (Leimgruber, Shaw, Santos & Olson, 2012).

However, one primary limitation of these studies is that, by manipulating only audience effects, they provide only a single factor account of children's motivations to act altruistically. Yet, it is very possible that different social factors, such as the child's relation to the recipient, may give rise to different reasons for giving. For example, children's sharing of resources with parents or close friends may be motivated by concerns for the recipient's well being rather than by a desire to maintain a positive reputation. On the other hand, in cases where the recipient is a stranger, and future interaction with the recipient is unlikely, reputation concerns may play a larger role in determining children's sharing behavior. Additionally, there may be contexts where children may be motivated by both reputation concerns and concerns for the recipient.

In sum, by studying children's motivations for giving independently of social context, past studies have not considered the possibility that children's developing

concern with reputation may be mediated by the particular social relationship that they have with the recipient of their giving. To explore this, the current study manipulated both audience effects and the child's social relation to the recipient to address this possibility. Specifically, we examined whether children's sharing tendencies would differ as a function of whether or not (a) they were observed by a neutral onlooker and (b) the potential recipient shared a social identity with them.

Various studies show that children attend to the social identity of potential recipients, and often prioritize members of their own social group when allocating resources (Bernhard, Fischbacher & Fehr, 2006; Bigler, Jones & Lobliner, 1997; Dunham, Baron & Carey, 2011; Fehr, Bernhard & Rockenbach, 2008; Goette, Huffman & Meier, 2006). One means by which children establish their own social identity is via social groups, which promote cohesion among similar individuals (in-group members) while emphasizing the dissimilarities of others (out-group members) (Brewer, 1991). The categorization of social groups can be affected by meaningful shared identities such as gender, nationality, and race (Tajfel & Turner, 2004), or they can be determined by a mere categorical distinction (Tajfel, Billig, Bundy & Flament, 1971; Tajfel, 2001). Tajfel coined the term "minimal group" to refer to arbitrarily constructed social groups such as the "red" and "blue" group. These meaningless social groups are sufficient to induce children and adults' preferences for in-group members across a wide range of measures including resource allocation in both kids and adults (Dunham et al., 2011; Locksley, Ortiz & Hepburn, 1980; Spielman, 2000; Tajfel & Turner, 2004). Additionally, there is strong evidence of in-group biases when social groups are meaningful or familiar (Killen, Margie & Sinno, 2006; Rutland, Killen & Abrams, 2010; Smetana, Killen & Turiel, 1991). These findings demonstrate that regardless of how groups are established, children weigh considerations for fairness and cooperation in relation to their obligations towards in-group members.

The current study used minimal groups to establish a social relationship between the giver and recipients. Specifically, we examined if the basis for children's sharing tendencies would change based on whether the recipient was a member of their in-group or out-group. Our design eliminated the expectation of reciprocity by using "fake" recipients who children would never interact with. Furthermore, we used a neutral observer who was not a member of the child's in-group or out-group. Previous studies have shown that children behave more prosocially when the observer is an in-group member compared to an out-group member (Engelmann at al., 2013. Since we did not want the in-group/out-group status of the observer to confound with other factors, such as anticipation of punishment or rewards from the observer, we selected a neutral onlooker. This also allows us to test whether children are sensitive to reputation concerns when the observer does not have a group affiliation or vested interest.

By testing for audience effects in the context of intergroup sharing we sought to differentiate children's motivations for giving. If children's sharing behavior is not influenced by social biases, we would expect them to give equally to in-group and out-group members—both in the presence and absence of an observer. If children's giving is subject to reputation concerns, we would expect them to donate more resources overall in the presence of an observer. Additionally, if children's giving is motivated by a bias towards the potential recipient and their desire to maintain a positive reputation, we might expect their intergroup sharing to change as a function of whether or not they are observed.

Children of all ages may give more to in-group members when observed than not observed simply because they view this as the right thing to do, thereby enhancing their reputations. On the other hand, older children and younger children may differ in their in-group and out-group giving behavior. Adult literature suggests that, out-group giving is more motivated by the desire for favorable impression management than in-group giving (Levine, Prosser, Evans, & Reicher, 2005) Older children may have similar beliefs to adults and give more to out-group members than in-group members in the observer condition. If this is the case, we anticipate children's out-group sharing to change as a function of being observed. Furthermore, it is possible that older children, but not younger children, perceive sharing that occurs in public to be less commendable than sharing that occurs in private. The prosociality literature indicates that adults view private donations as more "altruistic" and less "egoist" than public donations (Padilla-Walker & Carlo, 2014). However, we do not know if children as old as 9-years-old have adult-like cognitions about out-group sharing that occurs in private versus public. If older children differ from younger children in their beliefs about in/out-group sharing and private versus public giving, then we expect their patterns of sharing to also diverge from the sharing behaviors of younger children.

We also included an evaluation task to test children's perceptions of giving. After completing the sharing task, children were asked to judge scenarios of sticker sharing between in-group and out-group members. This was to test whether (a) children act in accordance with what they believe to be "nice" giving behavior, and (b) if children believe it is nicer to share with out-group or in-group members when sharing takes place publicly vs. privately.

Past studies demonstrate a tension between children's desire for equal outcomes and their preferences for in-group loyalty when distributing resources. However, few studies have examined how this tension influences children's beliefs about how others should behave in cooperative contexts. Olson and Spelke (2008) found that when children acted on behalf of a third-party protagonist (a doll) they prioritized the doll's in-group members when allocating resources. More recently, DeJesus, Rhodes and Kinzler (2014) found that 4- to 10-year-old children expected others to behave in ways that would benefit their own group. Thus

children's evaluations about the way people *should* behave are often not congruent with how they expect them to actually behave. What has not yet been examined is whether children's evaluations about how others should allocate resources amongst in-group and out-group members mirror their *own* behavior in a parallel task. Additionally, we know little about the degree to which children consider social obligations when evaluating the amount of resources shared between individuals. Previous studies have found that children rate scenarios of sharing zero resources as a "mean" act, especially if resources were earned collaboratively (Ng, Heyman & Barner, 2011). This demonstrates that children consider social obligations between individuals when evaluating sharing behaviors, but we do not know if this extends to their beliefs about intergroup sharing.

Method

Participants A total of 164 children (82 females) were tested, with 84 children in the observer condition and 80 in the no observer condition. Children' ages ranged from 5 to 9 years with 33 five-year-olds, 32 six-year-olds, 35 seven-year-olds, 32 eight-year-olds and 32 nine-year-olds; 82% were Caucasian, 15% were Asian American, 3% were identified as Mixed/Other by caregivers.

Procedures Children first completed a resource allocation task and then took part in a social evaluation task in which they evaluated the resource allocation decisions of other children. Children randomly selected a green or orange block hidden behind the experimenter's back to assign them to one of two minimal groups (the orange or green group). Blocks were surreptitiously manipulated to ensure that an equal number of children from each age group were assigned to each minimal group. Children were told that their group included other children in the area who chose the same color block that they did. They then wore a wristband matching the color of their group for the duration of the task. Next, each participant heard four narratives describing a competitive relationship between the two groups (e.g., "The orange group *really* wants to win against the green group, and the green group *really* wants to win against the orange group"). The narratives emphasized competition between the groups without mention of rivalry over resources in order to prevent children from viewing the resource allocation task itself as a competitive task.

Resource allocation task: Stimuli for the resource allocation task were six full-color head and shoulder photographs of Caucasian females between the ages of 5 and 7 attached to manila envelopes. We chose all female recipients to control for gender as an extraneous variable in children's in-group/out-group sharing behavior. Photographs were taken from Dunham et al. (2011) and were edited using a photo editing software such that half the children wore orange t-shirts and half wore green t-shirts.

On each trial children were presented with a manila envelope with a picture of a child from the contrasting group (three trials) or their same group (three trials). Each time a picture of a target child was presented, the experimenter placed seven stickers on the table in a random arrangement. Children were told that the stickers belonged to them and they could distribute the stickers in any way that they liked. They were led to believe that the envelopes would later be mailed to each target child's house, and were told that the experimenter would not look inside of the envelopes. Children put the stickers that they wanted to donate inside the envelopes with the target child's picture on it, and put stickers that they wanted to keep in a separate envelope. Children completed seven trials in total including one practice trial. For counterbalancing purposes, there were two different orders for the presentation of target pictures where one order was the reverse of the other. The in-group or out-group status of the potential recipient was manipulated within subjects, and the presence or absence of an observer during the session was manipulated between subjects.

For both conditions (observer and no observer), the experimenter placed a display poster board between herself and the participant so that the experimenter's view of the participant was obstructed. In the observer condition, a research assistant sat next to the child and maintained a neutral expression while watching the child complete the resource allocation trials. Children were told that the observer cared about how nice they were during the task, and was watching to see how many stickers they gave away.

Evaluation task: Children were told that they would see images of how other children played a sticker game similar to the one they just took part in—except now, they would decide if the children in the pictures were being "nice" or "mean" when they played the game. Examples of a child performing a nice act (cleaning up the classroom) and performing a mean act (pushing another child down on purpose) were read to each participant prior to playing the game. Children answered whether the actions were "nice" or "mean" to ensure that they understood the meaning of each term.

Next, children saw eight PowerPoint slides with pictures of two children (a giver and a recipient) on each slide. The slides depicted scenarios of sticker sharing between in-group and out-group members in the presence or absence of a female observer. The in-group versus out-group status of the potential recipient was manipulated within subjects and whether or not an observer was shown was manipulated between subjects. Children were assigned to the same observer condition in the judgment task as they had been in the resource allocation task.

At the start of each trial, children were told which child donated their stickers and which child was the recipient. Target children in the evaluation task were 16 Caucasian females between the ages of 5 to 7. Four scenarios consisted of giving stickers to the out-group, where the ratios between giving and keeping were 7:0, 5:2, 2:5, 0:7, and four matching scenarios consisted of in-group giving. Two different orders for the presentation of PowerPoint slides

were used, where one order was the reverse of the other. For each trial, children judged whether the giver was nice or mean. They were then asked either "How nice?" or "How mean?" and presented with three smiling or frowning faces, respectively. A six-point pictorial Likert scale adopted from Ng, Heyman and Barner (2011) was transformed into values: 6 (very very nice) to 1 (very very mean) for data analysis.

Results

Resource Allocation task: Analyses were performed in R 3.0.2 (http://www.r-project.org) using the lme4 package (Bates, Maechler, Bolker & Walker, 2014). For the resource allocation task we used a linear mixed model with Participants treated as a random factor. P-values were obtained by comparing likelihood ratio tests of the full model with the effect in question against the basic model without the effect in question. The basic model included Age and Gender as predictor variables, and mean percentage of stickers donated by children as the dependent variable. The full model included Age, Observer Condition (Observer, No Observer), recipient's Group Status (In-group, Out-group), and the interaction between Observer and Group Status (Observer/No Observer x In-Group/Out-Group) as predictor variables.

The analyses revealed that adding Observer Condition and Group Status to the model significantly increased the goodness of fit as indicated by likelihood ratio tests, $x^2(1) = 26.55$, $p < .001$. Welch's two sample t-test indicated that children in the Observer Condition donated significantly more stickers (M = 54%), than children in the No Observer Condition (M = 38%), $t(944) = -9.57$, $p < .001$. A paired samples t-test showed that children donated significantly more stickers to In-group members (M = 49%) than to Out-group members (M = 43%), $t(163) = 5.03$, $p < .001$). **Figure 1** displays children' average percentage of giving by Observer Condition and Group Status.

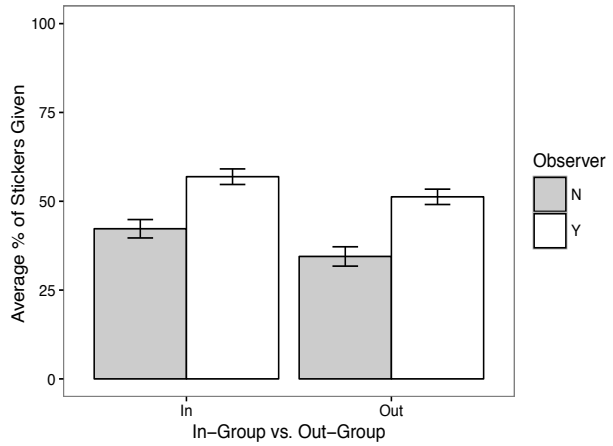


Figure 1: Percentage of stickers given by observer condition (Observer, No Observer) and group status (In-group, Out-group).

We did not find significant effects of Age or Gender in either of the analyses, nor did adding the interaction between Observer and Group Status improve the goodness of fit of our model. The difference between the Observer versus No-observer condition was slightly greater for out-group giving than for in-group giving, but this was not statistically significant.

Evaluation task: Multiple regression analysis was used to test if factors in the evaluation task, as well as children's' own giving behavior in the resource allocation task, predicted children's' explicit "niceness" ratings (1= very very mean…6= very very nice). For each participant, an average score of In-group/Out-group giving (referred to here as In-group sharing bias) was calculated by subtracting the mean number of stickers donated to Out-group members from the mean number of stickers donated to In-group members in the resource allocation task.

Data for the evaluation task were analyzed using a linear mixed model with Participants treated as a random factor. Correlation tests run prior to creating regression models revealed low collinearity among predictor variables. P-values were obtained by comparing likelihood ratio tests of the full model with the effect in question against the basic model without the effect in question. The basic model included Age and Gender as predictor variables and participants' mean "niceness" ratings as the outcome variable. A second model (Model 2) included Age, Gender, Observer Condition (Observer, No Observer) and the interaction between scenario allocation (7:0, 5:2, 2:5, 0:7) and inter-group sharing between target children (In-group, Out-group) as predictor variables. The full model included all predictor and outcome variables of Model 2, in addition to the interaction between participants' In-group sharing bias in the resource allocation task and inter-group sharing between target children in the evaluation task.

The analyses revealed that the full model accounted for significantly more variance in children's "niceness" ratings, than the basic model and Model 2, $x^2(1) = 7.07$, $p = .029$. The increased goodness of fit of the full model compared to Model 2 demonstrates a significant relationship between children's own giving behavior and their judgments of other children's intergroup giving. Children's ratings of different scenario allocations were dependent on the inter-group sharing that occurred. Children gave significantly different ratings for Out-group sharing scenarios where the giver shared 0 stickers ($\beta = 4.08$) versus 2 stickers (($\beta = 3.64$), 5 stickers ($\beta = 3.68$) and 7 stickers ($\beta = 4.01$). Children rated giving 0 stickers as meaner (M = 2.49) than giving any stickers: 2 stickers (M = 3.79), 5 stickers (M = 4.77), or 7 stickers (M = 5.29). Also, to our surprise, we found a significant main effect of group sharing such that children rated giving to Out-group members as significantly meaner (M = 3.77) than giving to In-group members (M = 4.39), $F(1,785) = 40.48$, $p < .001$ (see **Figure 2**). Additionally, Observer condition was a significant predictor of children's "niceness" ratings ($\beta = 3.94$), but age and gender were not.

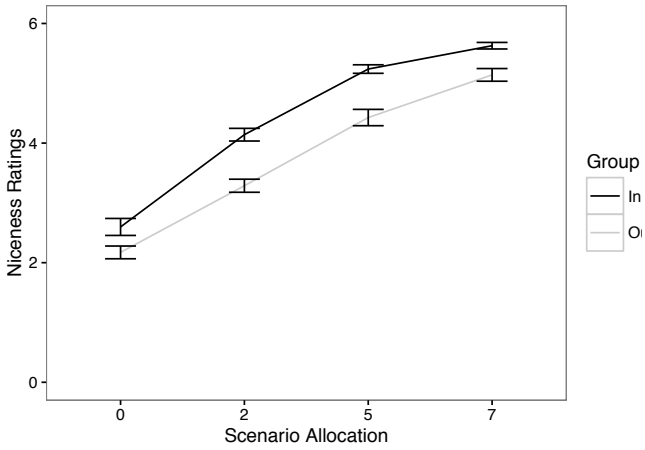


Figure 2: "Niceness ratings" (6 = very very nice, 1 = very very mean) by scenario allocation and inter-group sharing between target children (In-group, Out-group).

Discussion

We investigated how children's giving to others is affected by their desire to maintain a good reputation, and how this interacts with their relationship to potential recipients. We also examined whether children's evaluations of giving are reflective of their own sharing behavior, or if instead children judge other children negatively for exhibiting behaviors that they themselves perform.

Consistent with previous studies (Engelmann, Herrmann & Tomasello, 2012; Shaw et al., 2104), we found that children were significantly more generous when an observer was present. We also found that children donated more stickers to in-group members than out-group members across both observer conditions. However, surprisingly, children's giving to in-group and out-group members did not significantly differ between observer conditions: Children's giving to in-group members was just as influenced by reputation management as giving to out-group members, in contrast with our expectation that they might share for alternative reasons with in-group members—e.g., a desire to strengthen their group, or a heightened expectation of reciprocity (Bernhard, Fischbacher & Fehr, 2006; Sebastiàn-Enesco & Warneken, 2015).

Also surprising were children's explicit evaluations of giving behavior, and how they related to their own giving. Overall, children's own sharing behavior in the sticker allocation task predicted their "niceness" evaluations. While at first analysis this does not appear surprising, one consequence of this is that they rated scenarios of in-group giving as significantly nicer than scenarios of out-group giving. These findings contrast with DeJesus et al.'s (2014) study, which found that children evaluate out-group giving as nicer than in-group giving. In our study children did not rate giving 0 stickers differently for in-group or out-group giving, but they did rate giving all 7 stickers to an in-group member as significantly nicer than giving all 7 stickers to an out-group member. We see two potential reasons for this difference in findings. First, it is possible that children are more sensitive to group status in contexts where one must sacrifice all of his/her resources—e.g., that they are more likely to consider social obligations in scenarios where sharing incurs a high cost. Alternatively, children's evaluations of giving to in-group or out-group members may differ when they have themselves recently given, and themselves chosen to favor the in-group.

Future studies should examine this difference, and also the unexpected finding that children do not use different metrics for sharing with in-group and out-group members. Our intuition, beginning this study, was that in many cases, such as feeding one's family or providing clothing and shelter, reputation management is likely to be a secondary factor in giving behaviors. From this intuition, we reasoned that sharing might differ more generally between in-group and out-group members. While this second prediction appears to be incorrect, it remains unknown whether children's motives for giving to close relations might still be less sensitive to concerns about reputation, or whether here too, children give less when others are watching.

Future work should also examine why children seek to appear prosocial to a stranger. For example, it is unclear whether children were more generous in front of observers because they cared about their reputation or because they anticipated that their behavior could have resulted in a reward or punishment.

Also of interest is how suboptimal environmental conditions such as social instability and resource scarcity, which make sharing more costly for children, constrain their sharing behavior. Research in adults indicates that when faced with uncertainty of resource availability, individuals become less cooperative, especially across group boundaries (LeVine & Campbell, 1972). However, much remains unknown about how such factors across diverse societies impact children's sharing behaviors. Our sample consisted of predominantly Caucasian children of high socio-economic status living in a wealthy, urban community in the U.S. where resources are abundant and competition over scarce resources is minimal. This makes it unclear how well our findings generalize to children with different socioeconomic and ethnic backgrounds. How might children living in diverse areas of the U.S. or other countries compare to the children we tested? It is possible that with varying cultural and social norms regarding altruistic giving, we will find great variability in children's sharing tendencies. Future research in this area is needed to explore how cultural and socioeconomic differences may influence developmental trajectories of prosocial behavior.

Acknowledgments

We thank Del Mar Heights Elementary School for their involvement with recruiting participants, and Miriam Rubenson, Nestor Tulagan, Sydney Ghazarian and Kaitlyn Peck for their help in testing participants. We also thank Katie Wagner, Kevin Smith and Tim Lew for their help with data analyses.

References

Aloise-Young, P.A. (1993). The development of self-presentation: Self-promotion in 6-to-10-year-old children. *Social Cognition*, 11, 201–222.

Banerjee, R. (2002). Audience effects of self-presentation in childhood. *Social Development*, *11*, 487–507.

Banerjee, R., Bennett, M., & Luke, N. (2010). Children's reasoning about the self-presentational consequences of apologies and excuses following rule violations. *British Journal of Developmental Psychology*, *28*, 799–815.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). Lme4: Linear mixed-effects models using Eigen and S4.R package version 1.1-6.

Benenson, J.F., Pascoe, J. & Radmore, N. (2007). Children's altruistic behavior in the dictator game. *Evolution and Human Behavior*, *28*, 86–175.

Bernhard, H., Fischbacher, U. & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*, 912–915.

Bigler, B. S., Jones, L. C. & Lobliner, D. B. (1997). Social categorization and the formation of intergroup attitudes in children *Child Development*, *68*(3), 530–543.

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. Cognition, *120*, 215–224.

Blake, P.R. & Rand, D.G. (2010). Currency value moderates equity preference among young children. *Evolution and Human Behavior*, *31*, 210–218.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, *17*, 475–482.

DeJesus, J.M., Rhodes, M. & Kinzler, K.D. (2014). Evaluations versus expectations: Children's divergent beliefs about resource distribution. *Cognitive Science*, *38*(1), 178–193.

Dunham, Y., Baron, A. S., & Carey, S. (2011). Consequences of "minimal" group affiliations in childhood. *Child Development*, *82*(3), 793–811.

Engelmann, J.M., Over, H., Herrmann, E. & Tomasello, M. (2013). Young children care more about their reputation with ingroup members and potential reciprocators. *Developmental Science*, *16*, 952–958.

Fehr, E., Bernhard, H. & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*, 1079-1084.

Fehr, E., Fischbacher, U. & Gächter, S. (2002). Strong reciprocity, human co-operation, and the enforcement of social norms. *Human Nature*, *13*, 1–25.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science*, *14*, 1012–1020.

Goette, L., Huffman, D. & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. The American Economic Review, *96*(2), 212–216.

Harbaugh, W.T. & Krause, K. (2000). Children's altruism in public good and dictator experiments. *Economic Inquiry*, *38*, 95–109.

Hay, D.F. (1979) Cooperative interactions and sharing between very young children and their parents. *Developmental Psychology*, *15*, 647–653.

Hay, D.F. & Murray, P. (1982). Giving and requesting: Social facilitation of infants' offers to adults. *Infant Behavior and Developmen*t, *5*, 301–310.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). Data analysis: A model comparison approach. NewYork: Routledge/Taylor & Francis Group.

Leimgruber, K.L., Shaw, A., Santos, L.R. & Olson, K.R. (2012). Young children are more generous when others are aware of their actions. *PLoS ONE*, *7*, e48292.

Levine, M., Prosser, A., Evans, D. & Reicher, S. (2005) Identity and emergency intervention: How social group membership and inclusiveness of group boundaries shapes helping behavior. *Personality and Social Psychology Bulletin*, *31*, 443-453.

LeVine, R.A., & Campbell, D.T. (1972). Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior. New York: Wiley.

LoBue, V., Nishida, T., Chiong, C., DeLoache, J. S., & Haidt, J. (2011). When getting something good is bad: Even three-year-olds react to inequality. *Social Development*, *20*(1), 154–170.

Locksley, A., Ortiz, V. & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology*, *39*, 773–783.

Milinski, M., Semmann, D. & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, *415*, 424–426.

Moore, C. (2009). Fairness in children's resource allocation depends on the recipient. *Psychological Science*, *20*(8), 944–948.

Ng, R., Heyman, G.D., & Barner, D. (2011). Collaboration promotes proportional reasoning about resource distribution in young children. *Developmental Psychology*, *47*(5), 1230–1238.

Olson, K. R. & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*, 222–231.

Padilla-Walker, L.M. & Carlo, G. (2014). Prosocial Development: A multidimensional approach. New York: Oxford University Press.

Piazza, J., Bering, J.M., & Ingram, G. (2011). "Princess Alice is watching you": children's belief in an invisible person inhibits cheating. *Journal of Experimental Child Psychology*, *109*, 311–320.

Sebastiàn-Enesco, C. & Warneken, F. (2015). The shadow of the future: 5-year-olds, but not 3-year-olds, adjust their sharing in anticipation of reciprocation. *Journal of Experimental Child Psychology*, *129*, 40–54.

Shaw, A., Monitinari, N., Piovesan, M., Olson, K.R., Gino, F. & Norton, M.I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology*, *143*, 363–375.

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. PLos ONE, 8(3), e59510.

Spielman, D. A. (2000). Young children, minimal groups and dichotomous categorization. *Personality and Social Psychology Bulletin*, *26*, 1433–1441.

Sutter, M. (2007). Outcome versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*, 69–78.

Tajfel, H. (2001). Experiments in intergroup discrimination. In M. A. Hogg & D. Abrams (Eds.), Intergroup relations: Essential readings. London: Psychology Press (Original work published 1971).

Tajfel, H., Billig, M. G., Bundy, R. P. & Flament, C. (1971). Social categorization and
intergroup behavior. European Journal of Social Psychology, *1*(2), 149–178.

Warneken, F., Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology*, *100*, 455–471.

Whisman, M. A., & McClelland, G. H. (2005). Designing, testing, and interpreting interactions and moderator effects in family research. *Journal of Family Psychology*, *19*(1), 111–120.

Are There Hidden Costs to Teaching Mathematics with Incorrect Examples?

Min Kyung Hong (Min.Kyung.Hong@Vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University, Peabody College #552, 230 Appleton Place Nashville, TN 37203 USA

Darren J. Yeo (Darren.J.Yeo@Vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University, Peabody College #552, 230 Appleton Place Nashville, TN 37203 USA

Bethany Rittle-Johnson (Bethany.Rittle-Johnson@Vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University, Peabody College #552, 230 Appleton Place Nashville, TN 37203 USA

Lisa K. Fazio (Lisa.Fazio@Vanderbilt.edu)
Department of Psychology and Human Development, Vanderbilt University, Peabody College #552, 230 Appleton Place Nashville, TN 37203 USA

Abstract

This study aims to address potential costs of using incorrect worked examples in teaching mathematics. While such practice has been shown to be effective in educational research, previous findings in the memory literature suggest that exposure to an incorrect solution may lead students to later believe that it is correct due to increased familiarity. We designed a two-session experiment with 1-week delay in which students studied correct and incorrect worked out examples. We found only small changes in students' ability to successfully distinguish between correct and incorrect solutions over time. Students did rate the previously studied incorrect examples as being more correct after the 1-wk delay, but this did not affect their correctness ratings of new correct and incorrect worked examples or their problem solving accuracy. We conclude that the unique nature of mathematical problem solving may protect students from the dangers of using incorrect worked examples.

Keywords: incorrect examples; worked examples; problem solving; mathematics learning; illusory truth; source memory

Introduction

A common challenge in schools is to help students learn novel problem-solving techniques. For example, in mathematics students need to learn how to solve quadratic equations and calculate probabilities. Numerous studies have found that students learn more when they alternate between studying worked out examples of the problem and solution and solving the problems themselves, as compared to simply solving double the number of problems (Sweller & Cooper, 1985; Ward & Sweller, 1990; Kirshner, Sweller, & Clark, 2006; Renkl & Atkinson, 2010).

According to *Cognitive Load Theory*, studying a worked-out example is less burdensome for the learners' working memory than solving a problem, which leaves more room for deeper cognitive processing such as understanding and learning the steps for a solution (Sweller, 1999; Sweller et al., 2011; Paas, Renkl, & Sweller, 2003; Zhu & Simon, 1987) . While most studies have focused on studying correct worked examples, recent research suggests that studying incorrect examples can optimize learning (Große & Renkl, 2007; Tsovaltzi et al., 2010; Durkin & Rittle-Johnson, 2012; Adams et al., 2014). Presenting students with common incorrect ways to solve a problem can help students to confront and eliminate those common errors.

However, teachers are often resistant to use incorrect worked examples in the classroom, as they believe that repeated exposure to incorrect procedures might cause students to confuse incorrect and correct solutions. Research from the memory literature suggests that these teachers' fears may be correct. People rate repeated statements as being more true than those they have not encountered before (i.e. illusory-truth effect, Dechêne, Stahl, Hansen & Wänke 2009; Hasher, Goldstein, & Toppino, 1977), even when they have relevant prior knowledge to suggest that the statements are true or false (Fazio, Brashier, Payne & Marsh, 2015). Repetition increases the familiarity and processing fluency of a statement, which is thought to lead to increased perceived credibility of the information (Unkelbach, 2007; Begg, Anas, & Farinacci, 1992). The illusory truth effect has been shown to occur for trivia statements (e.g., Bacon, 1979), information about consumer products (e.g., Johar & Roggeveen 2007), and political opinions (e.g., Arkes, Hackett & Boehm, 1989), and it occurs both in the laboratory and in naturalistic settings (Gigerenzer, 1984; Boehm, 1994).

In addition, research has shown that people often remember factual information while forgetting the source of that information (Barber, Rajaram, & Marsh, 2008; Conway et al 1997; Dewhurst et al. 2009). This forgetting may be a natural consequence of information moving from episodic to semantic memory. For example, you likely know that George Washington was the first president of the United States, but no longer remember where you first learned that information. In related phenomenon such as the sleeper effect (Kumkale & Albarracin, 2004), false fame (Topolinski & Strack, 2010), and unconscious plagiarism (Bink, Marsh, Hicks & Howard, 1999), participants

remember previously presented information while forgetting its source or origin.

Taken together, these findings suggest that studying with incorrect worked examples may not yield benefits if students fail to accurately monitor the source of the information.

The present study was designed to examine if this forgetting of the source information occurs with correct and incorrect worked examples. That is, after a delay, would students continue to remember which worked examples were correct and which were incorrect or would they begin to confuse the two procedures? During the study phase, students saw both correct and incorrect worked examples and were asked to rate the correctness of the shown solution. They then received feedback about which examples were correct or incorrect and rated the examples again. Finally, they were asked to solve novel problems of the same type shown in the examples. One week later, the students returned and again solved problems and rated the correctness of the worked examples. Of interest was how students' ratings and problem solving accuracy would change over time.

If teachers' concerns are correct, then we would expect students to be less accurate at rating the correctness of the worked examples after the delay. In addition, we would expect students' problem solving accuracy to also decrease. If, however, students are able to remember which examples are correct and which are incorrect, then we would expect no changes in students' correctness ratings or problem solving accuracy over time.

Method

Participants

Thirty-four Vanderbilt University undergraduates participated in exchange for course credit (9 male; mean age 18.8 years). Three participants were excluded from the analysis because they failed to attend the second session of the experiment, leaving 31 participants in the analysis. Participants were tested individually or in small groups of up to four people.

Design

The experiment implemented a 3 (time: study, immediate, delay) × 2 (solution: correct, incorrect) within-subjects design (Figure 1).

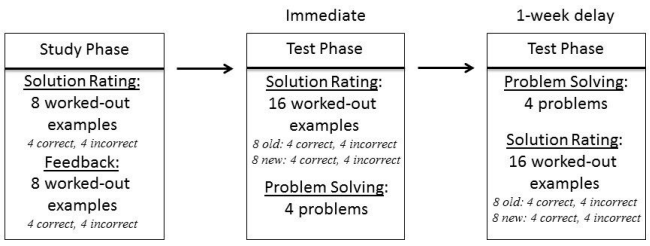


Figure 1. Study design.

Materials

We adapted 10 pretest problems, eight worked-out examples, and four post-test probability problems from Experiment 2 of Große & Renkl (2007) with minor word changes (e.g. "skat cards" to "poker cards"). The problems all dealt with probabilities concerning the intersection between two different events or event order. Event intersection problems differed on whether two different events were mutually exclusive or if they could occur together; event order problems differed on whether the sequence of events were relevant or irrelevant. We then created an additional set of eight worked-out examples that were structurally similar to the eight worked-out examples from Große & Renkl (2007). Therefore, participants were shown a total of 16 worked-out examples, half of which were presented with correct solutions and half with incorrect solutions. The incorrect worked examples contained accurate calculations, but the procedure used was not appropriate for the problem (e.g. failing to subtract the intersection between event A and event B when calculating the probability of event A or event B happening). Stimuli for the pretest, study phase, filler task, and solution rating task were presented on a computer using MediaLab and DirectRT software (Empirisoft Corporation, New York, NY). The problem solving task was administered through a paper-and-pencil test.

Procedure

Participants gave informed consent and solved 10 basic probability problems (pretest), which were used to assess their prior domain knowledge.

Participants then studied eight worked-out examples, four correct and four incorrect. Participants were told that the solutions were provided by other students and they would be asked to rate the accuracy of the student's solution. Each worked example was presented one at a time, and participants keyed in the intermediate answers when they were prompted with fill-in-the-blank boxes (shown in Figure 2). This ensured that participants paid attention to each of the steps in the solution and increased their engagement and depth of processing. Each step was revealed progressively, regardless of whether the participant's input for each blank was correct or incorrect. Thus, if a participant answered incorrectly, their answer would disappear and the correct answer would be presented along with the next step. At the end of each worked example, the full solution was presented and participants rated the correctness of the solution using a 7-point Likert scale with 1 labeled as "Definitely incorrect" and 7 indicating "Definitely correct".

Participants then solved visuo-spatial puzzles as a filler task for two minutes. After the filler task, participants were presented with the same set of fully completed worked-out examples, in the same order that they had seen previously. For each problem, they were presented with feedback about whether the worked example was correct or incorrect and were given at least 20 seconds to review the solution and the

feedback. After 20 seconds, their earlier correctness rating appeared at the bottom of the solution (e.g. "You rated the correctness of the solution of this worked-out example: 2"), and they were prompted to move onto the next worked example when they were ready. Participants then completed another set of visuo-spatial puzzles for two minutes before moving onto the test phase.

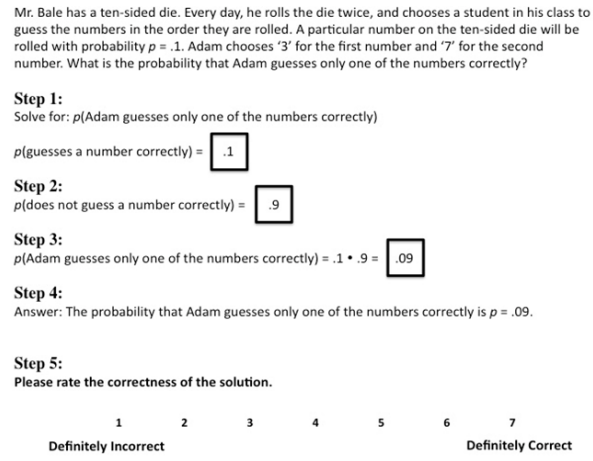


Figure 2. Illustration of a worked-out example. Each step was revealed progressively, and participants were prompted to type their answers in each fill-in-the-blank box, represented by the black boxes above.

During the test phase, participants again rated the correctness of the worked examples and they solved novel probability problems. Task order within the test phase was counterbalanced across participants. During the solution rating task, participants were given 16 worked-out examples, eight of which were *old* problem*s* shown during the study phase, and eight of which were *new* problems that were isomorphic to the *old* problems but with different cover stories. Participants rated each solution using the 7-point response scale. During the problem solving task, participants were given four post-test probability problems. Participants' performance on the probability problems was used to examine their ability to apply any gained knowledge from the study phase to novel problems. The order of the problems in the booklet was counterbalanced across participants.

At the end of the test phase, participants were asked if they had learned how to solve these specific types of probability problems prior the experiment (if so, what did they learn and where) and to list the math and statistics classes that they had taken in high school (advanced placement courses only) and college.

A week after the first session, participants came back to the lab and completed the solution rating and problem solving tasks again. The order of the tasks was reversed from the first session. At the end of the second session, participants were asked to give demographic information about their age, gender, and ethnicity.

Results

Solution ratings for old items across time

Recall that students rated the examples on a scale of 1 to 7. Ratings above 4 indicated believing that the solution was more correct than incorrect, and ratings below 4 indicated believing that the solution was more incorrect than correct. As shown in Figure 3, during the study phase, prior to feedback, students rated both correct and incorrect solutions as being more correct than incorrect, although their ratings for correct solutions were higher than for incorrect solutions. Thus, participants were having some difficulty identifying incorrect solutions as incorrect. On the immediate test, as expected, feedback allowed participants to more clearly distinguish between correct and incorrect solutions, and they rated the incorrect solutions as incorrect. Of primary interest were ratings on the delayed test. Did participants forget which solutions were incorrect, in line with the illusory truth effect and source forgetting? The answer seems to be somewhat; students rated incorrect solutions more highly on the delayed posttest than the immediate posttest, although not as highly as before feedback (during the study phase).

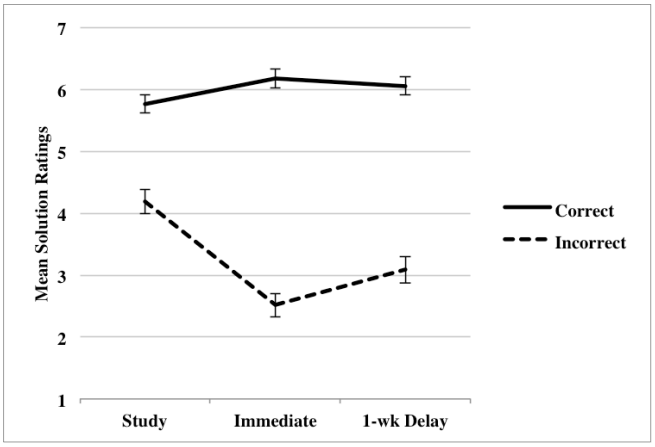


Figure 3. Mean solution ratings for correct and incorrect examples as a function of time. Error bars reflect the standard error of the mean.

To examine these patterns statistically, we conducted a 3 (time: study, immediate, 1-wk delay) × 2 (solution: correct, incorrect) repeated measures analysis of variance (ANOVA) on participants' solution ratings. This analysis was restricted to the eight problems that participants saw during all three phases of the experiment: prior to being told which solutions were correct and incorrect, immediately after the feedback, and one week later.

As shown in Figure 3, participants rated the correct solutions (M = 6.00) as being more correct than the incorrect solutions (M = 3.27), $F(1, 30) = 87.59$, MSE = 3.97, $p < .001$, $\eta_p^2 = 0.75$. This was true during the study phase before the students received feedback, $t(30) = 5.38$, $p < .001$, $d = 3.19$, and on the immediate, $t(30) = 10.39$, $p <$

.001, d = 4.90, and delayed tests, $t(30) = 7.57$, $p < .001$, d = 4.64. In addition, participants' ratings changed over time, $F(2, 60) = 6.46$, $MSE = 0.99$, $p = .003$, $\eta_p^2 = 0.18$, but this was qualified by a significant interaction between time and solution type, $F(2, 60) = 21.05$, $MSE = 0.83$, $p < .001$, $\eta_p^2 = 0.41$. While the ratings for correct solutions remained relatively constant over time, the ratings for incorrect examples changed dramatically. Follow-up t-tests showed no significant differences across the time points for correct solutions (study vs. immediate, $t(30) = 1.70$, $p = .10$; study vs. delay, $t(30) < 1$; immediate vs. delay, $t(30) = 1.31$, $p = .20$). However, there was a large decrease in the ratings for incorrect solutions after the participants received feedback (study M = 4.19; immediate M = 2.52; $t(30) = 5.99$, $p < .001$, $d = 2.58$). This change remained one-week later (study M = 4.19; delay M = 3.09; $t(30) = 3.98$, $p < .001$, $d = 2.20$), but the incorrect examples were rated as being slightly more correct as compared to immediately following the feedback ($t(30) = 2.99$, $p = .005$, $d = 0.38$). Overall, participants were able to successfully distinguish correct items from incorrect items, even after the delay. This pattern of results did not change when we included participants' pretest scores as a covariate.

Solution ratings for old versus new items

Given that students were able to successfully distinguish between correct and incorrect items, we were interested in whether they were able to transfer that knowledge to the new solutions. To more easily compare participants' knowledge over time, we conducted the analysis on the difference between participants' ratings for correct and incorrect items. Higher difference scores represent greater ability to distinguish correct solutions from incorrect solutions. The results are shown in Figure 4.

As noted above, for the old items, participants' ability to distinguish between correct and incorrect solutions was high immediately following the feedback, but this ability decreased after the one week delay. However, we did not see a similar pattern with the new examples. For the novel examples, the difference in rating of new examples was similar to the difference during the study phase for the old examples (mean differences = 1.96 vs. 1.58) and was lower than for the old examples on the immediate test. The difference in ratings for the new examples did not change from the immediate test to the 1-week delay. These findings suggest that participants were remembering particular examples of solution methods tied to individual problems and not a more general solution method that they used to rate the accuracy of new examples.

To confirm these observations, we conducted a 2 (time: immediate, 1-wk delay) × 2 (novelty: old, new) ANOVA on the difference scores. As shown in Figure 4, participants were better able to discriminate correct and incorrect solutions for old items (M = 3.31) than new items (M = 2.14), $F(1, 30) = 21.49$, $p < .001$, $SEM = 2.00$, $\eta_p^2 = 0.42$. There was no main effect of time, $F < 1$, but there was a significant interaction between time and novelty, $F(1,30) = 9.85$, $p = .004$, $SEM = 0.87$, $\eta_p^2 = 0.25$. The difference in knowledge between the old and new items was larger on the immediate test than one-week later. Difference scores for the old items decreased over time (immediate M = 3.66; delay M = 2.97; $t(30) = 2.37$, $p = .024$, $d = 1.84$), while difference scores for new items did not change significantly (immediate M = 1.96; delay M = 2.32; $t(30) = 1.54$, $p = .13$, $d = 0.15$). These results suggest that participants did not generalize what they learned from the feedback to the new examples.

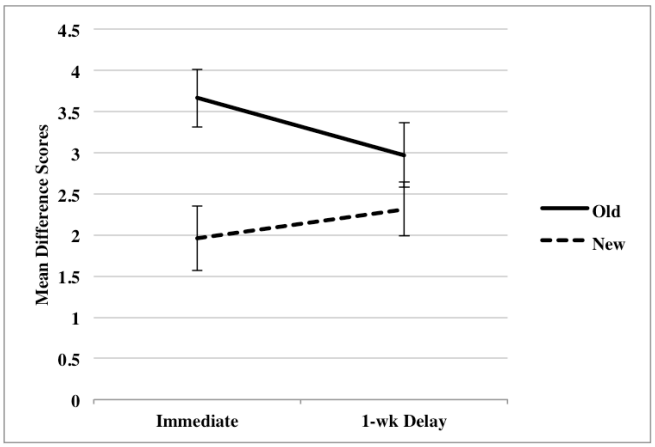


Figure 4. Solution rating differences across time for old and new problems. Error bars reflect the standard error of the mean.

Problem solving accuracy

Next, we examined how students' ability to solve the four booklet problems changed after a 1-week delay. There was no significant difference in the proportion of questions answered correctly on the immediate (M = .60) and delayed (M = .54) tests, $t(30) = 1.16$, $p = .26$, $d = 0.21$.

Problem solving strategies

We also examined the types of strategies used for each of four probability problems (shown in Figure 5) to see how often participants used the demonstrated strategies. Strategies were coded as *correct, incorrect,* or *other*. Incorrect strategies were the strategies presented in the incorrect worked-out examples, and other strategies were incorrect strategies that were not introduced through the worked-out examples. All the participants who used correct strategies solved the problems correctly.

Overall, students' strategy choice was consistent over time, and they used correct strategies more often than incorrect or other strategies. Students were equally likely to use correct strategies on the immediate and delayed tests, $t(30) = 1.23$, $p = .23$, $d = 0.22$. And there were also no changes in their use of incorrect, $t(30) = -1.36$, $p = .18$, $d = -0.25$, and other strategies, $t(30) = -0.15$, $p = .88$, $d = -0.03$.

Students did not seem to frequently adopt or revert to the incorrect strategies demonstrated in the incorrect examples.

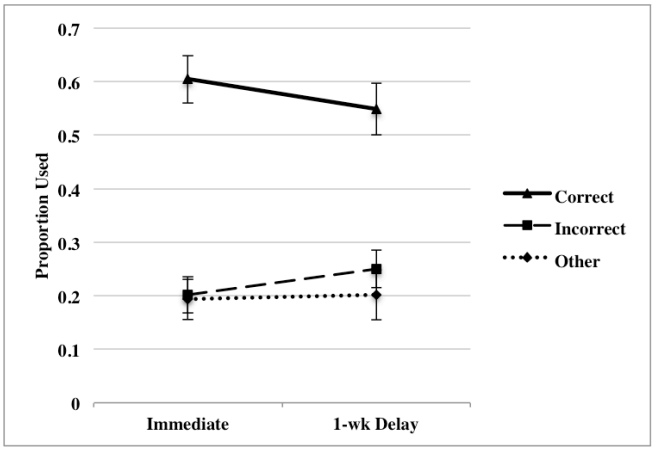


Figure 5. Proportion of correct, incorrect and other strategies used on the probability problems on the immediate and delayed test. Error bars reflect the standard error of the mean.

Discussion

The present study examined whether findings from the memory literature suggesting potential costs of learning through incorrect worked examples extended to mathematics problem solving. Overall, students showed little forgetting over time and were able to distinguish between the correct and incorrect worked examples, even after a one-week delay. For the repeated items, there was an increase in participants' ratings for the incorrect examples after a one-week delay, suggesting that students may have started to forget the "incorrect" tag. The fact that the studied incorrect examples were rated as more correct over time, while there was no corresponding decrease in the ratings for correct examples, suggests that the illusory truth effect may have played a role. Rather than simply forgetting which examples were correct or incorrect, participants may have believed all of the repeated examples to be more correct after the delay. The correct examples were already close to ceiling and thus could show no changes.

However, there was little evidence that participants generalized information from the studied examples. First, their ratings of new examples did not seem to be influenced by their improved knowledge of old examples on the immediate test or their decline in knowledge of old examples on the delayed test. Second, participants were not that likely to use the demonstrated incorrect strategies on the immediate test nor increase their use of it on the delayed test. So while the old incorrect examples gained a little bit of truth over the delay, this did not affect the students' ability to rate the new examples or to accurately solve the problems. Thus, our results provide some preliminary evidence for educators that their worries about the negative effects of presenting incorrect worked examples may be unfounded.

Worked examples differ from the stimuli used in previous memory studies in many ways. One key difference is that the correct and incorrect worked examples both support the learning of the same correct procedure. Thus, one does not need to remember both that this specific example was correct and that this specific example was incorrect. Instead, by learning the procedure, one can identify both which solutions are correct and which solutions are incorrect. In contrast, within the illusory truth paradigm, knowing that "Oslo is the capital of Finland" is false tells you nothing about the truth of "Marconi is the inventor of the wireless radio". If incorrect worked examples help students to gain an accurate representation of the correct solution, then it may not matter if students forget the specific label (correct or incorrect) that went with a given worked example.

One limitation of the current study was that the students did not fully learn the correct procedures. The participants correctly answered only a little more than half of the problems on both the immediate and delayed posttests. Our procedure was designed to match those used in the illusory truth and sleeper effect literatures and thus does not match how educators and educational researchers typically use incorrect worked examples. Future studies should examine students' memory after they experience a typical classroom lesson where the incorrect examples are directly compared to the correct examples. Such direct comparisons may increase learning of the correct procedures. It is also possible that students may have shown greater forgetting after a longer delay. Brown and Nix (1996) found that participants in an illusory truth experiment could remember which statements were labeled as true or false one week later, but after a month all repeated statements were rated as being more true, regardless of their initial labeling.

The current study represents a first step towards bringing together research on memory processes and research on learning to examine the long-term effects of different instructional techniques. We found small effects of forgetting on the rating scales typically used in the relevant memory research, but no negative effects on the students' actual problem solving skills. It remains to be determined whether the small amount of forgetting shown in this experiment would cause larger issues at a longer delay or if students can correctly learn the procedure even if they forget the initial correct and incorrect labels.

Acknowledgments

We would like to thank Meredith Sight and Octavia White for their help regarding this project.

References

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior*, *36*, 401–411.

Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*(2), 81–94.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 241-252.

Barber, S. J., Rajaram, S., & Marsh, E. J. (2008). Fact learning: How information accuracy, delay, and repeated testing change retention and retrieval experience. *Memory*, *16*(8), 934–946.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446–458.

Bink, M. L., Marsh, R. L., Hicks, J. L., & Howard, J. D. (1999). The credibility of a source influences the rate of unconscious plagiarism. *Memory*, *7*(3), 293–308.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, *20*(3), 285–293.

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088–1100.

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G. M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*, *126*(4), 393–413.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix me a list: Context moderates the truth effect and the mere-exposure effect. *Journal of Experimental Social Psychology*, *45*(5), 1117–1122.

Dewhurst, S. A., Conway, M. A., & Brandt, K. R. (2009). Tracking the R-to-K shift: Changes in memory awareness across repeated tests. *Applied Cognitive Psychology*, *23*(6), 849–858.

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, *22*(3), 206–214.

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002.

Gigerenzer, G. (1984). External Validity of Laboratory Experiments: The Frequency-Validity Relationship. *The American Journal of Psychology*, *97*(2), 185–195.

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634.

Hansen, J., & Wänke, M. (2009). Liking What's Familiar: The Importance of Unconscious Familiarity in the Mere-Exposure Effect. *Social Cognition*, *27*(2), 161–182.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behavior*, *16*(1), 107–112.

Jarvis, B. G. (2014). MediaLab (Version 2014.1.108) [Computer Software]. New York, NY: Empirisoft Corporation.

Jarvis, B. G. (2014). DirectRT (Version 2014.1.104) [Computer Software]. New York, NY: Empirisoft Corporation.

Johar, G. V., & Roggeveen, A. L. (2007). Changing false beliefs from repeated advertising: The role of claim-refutation alignment. *Journal of Consumer Psychology*, *17*(2), 118–127.

Kirschner, P.A., Sweller, J., and Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86.

Kumkale, G. T., & Albarracín, D. (2004). The Sleeper Effect in Persuasion: A Meta-Analytic Review. *Psychological Bulletin*, *130*(1), 143–172.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, *38*(1), 1–4.

Renkl, A., & Atkinson, R. K. (2010). Learning from worked-out examples and problem solving. (pp. 91–108). Cambridge University Press (New York, NY, US).

Sweller, J., & Cooper, G. A. (1985). The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. *Cognition and Instruction*, *2*(1), 59.

Sweller, J. (1999). *Instructional Design in Technical Areas*. *Australian Education Review, No. 43*. PCS Data Processing, Inc., 360 W. 31st, New York, NY 10001

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer New York

Topolinski, S., & Strack, F. (2010). False fame prevented: Avoiding fluency effects without judgmental correction. *Journal of Personality and Social Psychology*, *98*(5), 721–733.

Tsovaltzi, D., Melis, E., McLaren, B. M., Meyer, A.-K., Dietrich, M., & Goguadze, G. (2010). Learning from Erroneous Examples: When and How Do Students Benefit from Them? In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Sustaining TEL: From Innovation to Learning and Practice* (pp. 357–373).

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 219–230.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction*, *7*(1), 1–39.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction*, 4, 137-166.

Interaction of Instructional Material Order and Subgoal Labels on Learning in Programming

Laura M Schaeffer (lschaeffer6@gatech.edu)
Department of Psychology, 654 Cherry Street
Atlanta, GA 30332 USA

Lauren Margulieux (l.marg@gatech.edu)
Department of Psychology, 654 Cherry Street
Atlanta, GA 30332 USA

Richard Catrambone (richard.catrambone@psych.gatech.edu)
Department of Psychology, 654 Cherry Street
Atlanta, GA 30332 USA

Abstract

Subgoal labeled expository instructions and worked examples have been shown to positively impact student learning and performance in computer science education. This study examined whether problem solving performance differed based on the order of expository instructions and worked examples and the presence of subgoal labels within the instructions for creating applications (Apps) for phones. Participants were 132 undergraduates. A significant interaction showed that when learners were presented with the worked example followed by the expository instructions containing subgoal labels, the learner was better at outlining the procedure for creating an application. However, the manipulations did not affect novel problem solving performance or explanations of solutions,. These results suggest that some limited benefit can be gained from presenting a worked example before expository instructions when subgoal labels are included.

Keywords: instructional design; STEM education; programming.

Introduction

Learners have difficulty solving novel problems, or problems that require steps that are different from worked example problems they have already encountered (Catrambone, 1995; Reed, Dempster, & Ettinger, 1985; Ross, 1987, 1989). This difficulty stems from learners tending to fixate on superficial aspects of examples as opposed to the goal structure of the problem. When learners understand the goal structure of the example problems, they become more successful at solving novel problems (e.g., Catrambone, 1995).

Subgoals are part of the task structure and organize solution steps into a meaningful hierarchy; subgoals are specific to problems within a particular domain (Catrambone, 1994; Catrambone, 1998). Subgoal labels assist learners in noticing and learning the subgoals and organizing their problem solving knowledge. This organization is demonstrated when learners who received instructions with subgoal labels tended to explain their problem solutions using the subgoals (Catrambone, 1995; Margulieux, 2013). Subgoal labels within instructions have improved transfer in many domains, including computer programming, and have been shown to be most effective when provided in both expository instructions and worked examples (Margulieux, 2013).

Expository Instructions

Expository instructions usually consist of both declarative information, such as terminology, and procedural information (Trafton & Reiser, 1993). Procedural instructions describe and explain how to carry out a task (Eiriksdottir & Catrambone, 2011). Procedural instructions are often written at a more general level than worked examples, so they can be applied to a variety of situations. The learner is equipped with the high level concepts needed to solve novel problems within the domain (Catrambone, 1990). This allows students who master procedural instructions to be able to solve novel problems better than students who receive more specific instructions (Catrambone, 1990). However, because procedural instructions do not have the same level of detail as more specific instructions, such as a worked example, more detailed information must be inferred. This inferential process is quite challenging for many learners.

Worked Examples

Worked examples demonstrate how a specific instance of a task is performed (Eiriksdottir & Catrambone, 2011). Worked examples are generally structured as a problem statement followed by the steps needed to arrive at the solution. They provide a concrete application of the problem solution's abstract concepts, rules, and general directions (Charney & Reder, 1987; Pirolli & Recker, 1994; Wiedenbeck, 1989). This allows the learner to become familiar with the task and increase their understanding of how to carry out the task (Eiriksdottir & Catrambone, 2011). Because worked examples provide detailed information, learners are able to more easily apply the same

procedure to a similar problem than if they had been given more abstract information (Catrambone, 1990). Learners who use worked examples have also been shown to perform similar tasks more quickly than learners who used only procedural instructions (Catrambone, 1990).

One drawback of typical worked examples is that they do not inherently provide the learner with any general methods or reasoning behind decisions (Eiriksdottir & Catrambone, 2011). When given a worked example, the learner must infer information such as the nature of the task, the purpose of each step, rules governing the steps, subgoals, and organization (LeFevre & Dixon, 1986; Pirolli & Recker, 1994). In limited cases learners have been shown to infer general methods when several worked examples are presented, but usually guidance is needed for such connections to be made (Rumelhart & Norman, 1981). Presenting the learner with both procedural instructions and worked examples has been shown to produce the benefits associated with each type of instructional material while reducing the drawbacks. Catrambone (1995) showed that presenting procedural text with a worked example aided both initial performance and transfer.

There is reason to believe the order in which the instructions are presented might affect the learner's ability to process them. Several lines of research suggest that students perform and learn better when given a worked example followed by procedural texts (Alfieri, Nokes-Malach, & Schunn, 2013; Anderson, 1990; Dale, 1946). Dale (1946) argued that when learning math, students should first be introduced to concrete objects (e.g., five fingers as opposed to an abstract five), and then work up to semi-concrete ideas. If the material does not relate to a student's experience with the items in the equation, the formula will not mean anything (Dale, 1946). Dale (1946) concluded that the role of the teacher is to take the student from concrete experiences to significant and important generalizations. Other studies also suggest that it is better to give people principles for the concept or procedure that they are trying to learn after they view the cases (Alfieri, Nokes-Malach, & Schunn, 2013).

Another theory, from the inductive teaching research literature, suggests that worked examples provide the "why" behind the principles and procedure (Prince & Felder, 2006). The specifics from worked examples cause the learner to generate a need for more information, such as the rules, procedures, and principles. This curiosity then motivates the learner to incorporate and apply the instructions.

It has been noted that new information is best learned when the learner has a knowledge base to support the information, and they are unlikely to learn if the new information has few apparent connections to what they already know. Advance organizers have been used to provide such a foundation (Ausubel, 1968; Novak, 1977). Advance organizers can be used as an effective way to bridge the gap between the novice's knowledge and the basis on which the instructions function (Ausubel, 1968). When presented at a suitable level for the learner, advance organizers activate the learner's prior knowledge making the new information more familiar and meaningful, which decreases dependence on sheer memorization in favor of a meaningful understanding of the information.

A worked example might serve a similar function as an advance organizer because it gives the learner a base on which to apply the latter expository information. A worked example introduces the learner to the type of situation to which the expository information is applicable, mobilizing the learner's prior knowledge. Therefore, instructional materials might be more effective if the worked example is presented before the expository information.

Alternatively, presenting the worked example first might be disadvantageous. According to Ausubel (1968), instructions aid mental organization better when progressing from abstract ideas to specific details because this organization better fits our cognitive structure. Additionally, presenting specific details first, such as those found in the worked example, might cause the learner to focus on applying the expository instructions to problems that are very similar to the worked example. Consequently, the learner might have a more difficult time generalizing the instructions to other situations. Because of this, presenting the worked example first might hinder the learner's ability to use the abstract principles when solving novel problems. However, subgoal labels might help learners compensate for this effect because they explicitly provide the higher level functions found within the worked example and the expository instructions.

Present Study

The present study investigated the effects of instructional material order and subgoal labels in learning computer programming. Participants were taught how to use the programming language (Android App Inventor) to create a Fortune Teller application (app). The App Inventor programming environment uses a drag-and-drop interface to create apps for Android devices.

Drag-and-drop programming is ideal for novices because instead of writing code, the learners drag components from a menu and fashion them together like puzzle pieces. Creating code in this way has been shown to be easier for novices to comprehend than other types of programming environments (Hundhausen, Farley, & Brown, 2009).

Videos were used to convey the App Inventor instructions because videos have been shown to be a natural and efficient way for learners to gain knowledge of direct-manipulation interfaces (Palmiter & Elkerton, 1993; Palmiter, Elkerton, & Baggett, 1991). Participants also used a practice problem guide to practice creating the Fortune Teller app before being tested. Trafton and Reiser (1993) showed that learners who study and practice newly learned material are better able to apply the material than learners who are not given the opportunity to practice.

Method

Participants

Participants were 132 undergraduate students from the Georgia Institute of Technology compensated with course credit. The sample consisted of 68 females and 64 males. The mean age was 19.3 years with a standard deviation of 1.93. Participants were excluded if they had taken more than one computer science course or had experience with App Inventor. These qualifications were necessary because the instructional materials were designed for novices.

Design

The experiment was a two-by-two, between subjects, factorial design with 33 participants per cell. The first independent variable was the order participants received the instructional materials: expository followed by worked example or worked example followed by expository. The second independent variable was presence of subgoal labels: present or absent. The dependent variables consisted of performance on three assessment tasks to determine organization of domain knowledge and problem solving performance.

Procedure

Each session lasted between 60 and 90 minutes. Participants were randomly assigned to one of four conditions. All participants first completed the demographic questionnaire. Next, participants began the instructional period where they watched both instructional videos (the expository video and the worked example video) before using a practice problem guide to practice creating an app. The expository instructional videos contained general procedural instructions and declarative information, such as definitions, necessary for creating an app in App Inventor. The worked example video demonstrated how to create a specific app, the Fortune Teller app. Subgoals were created by Margulieux (2013) using the Task Analysis by Problem Solving (TAPS) method developed by Catrambone et al., (2012).

The videos used callouts to present the subgoal labels. These were text boxes containing the subgoal labels appearing on screen while the narration continued explaining the steps needed to achieve the subgoal.

The final instructional material was the practice problem guide, which was a scaffolded worked example. The stages of scaffolding can vary (Pea, 2004), but in the present study the practice problem guide provided learners with the steps necessary for creating the Fortune Teller app without giving them guidance on how to carry out the steps (e.g., where in the menus to find blocks). The scaffolded example used the same Fortune Teller app presented in the worked example video.

After the instructional period, the participants began the assessment period. During the assessment period, the participants were not able to use the materials from the instructional period. However, they were able to use the App Inventor website and refer to the app they created during the instructional period as an aid to problem solving (Margulieux, 2013). The first assessment consisted of four problem solving tasks in which participants were instructed to add or modify features of their Fortune Teller app. This assessment was broken into two parts where the participants were first asked to modify the app directly in App Inventor, and later asked to write down the necessary steps. The written portion allowed participants to demonstrate their knowledge of steps even if they did not know how to correctly execute the steps in App Inventor. This assessment measured participants' problem solving performance on novel tasks using App Inventor.

The second assessment was the explanation task. Correct solutions to the four problem solving tasks were given to the participants. Participants were asked to group steps of the problem solving task solution. They were then asked to label their groups by describing what goal was met for each grouping. This assessment measured how well participants could group steps based on structural similarity, and how well they could explain the solutions.

The final assessment was the generalization task that asked participants to describe the general procedure that they would use to create an app within a given set of constraints. A correct response to this task included the fundamental steps needed to make the app while excluding unnecessary details. This assessment was used to measure how well the participants could use abstract principles to outline the task procedure they learned earlier in the session.

Results

Demographic information such as age, GPA, college major, and experience with computer science were collected but were not correlated with performance on any of the following assessments.

General Procedure Task

The general procedure asked participants to describe the general process they would use to create an app. One point was awarded for each structurally necessary feature the participant described, for up to a maximum score of 6. ICC(A) for this assessment was .99. There was no main effect of instructional material order, F (1, 132) = 0.58, p = .45 There was also no main effect of subgoal labels, F (1, 132) = 1.31, p = .26 (see Table 1).

Table 1. Score on Task for Describing General Process to Create an App

Worked Example First		*Expository First*	
Subgoals	*No labels*	*Subgoals*	*No labels*
M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
2.85 (1.46)	2.03 (1.22)	2.12 (1.47)	2.40 (1.29)

Note: Score out of six possible points.

However, there was a significant interaction between the instructional material order and subgoal labeling, F (1, 132) = 5.49, p = .02. Simple main effects analysis showed that participants who received subgoal labels were able to provide more steps of the general process for creating an app than those who did not receive subgoal labels when presented with the worked example before the expository instructions, p = .02, but there were no differences between the subgoal labeled group and the group without subgoal labels when the expository instructions were presented before the worked example, p= .40.

Problem Solving Tasks

The following assessments were scored following the method developed by Margulieux et al. (2012), which has been shown to have high statistical power (due to partial scoring methods discussed later) and high interrater reliability. Two raters scored each of the assessments; interrater reliability was measured with an intraclass correlation coefficient of absolute agreement (ICC(A)).

Performance in App Inventor. For this task, participants were asked to modify or add different features of an app. They were awarded one point for each correct action in App Inventor taken towards the problem solutions for up to a maximum score of 22. ICC(A) for this assessment was .89. Visual inspection of the data revealed that the data were not normally distributed (see Figure 1). The residuals were not normally distributed, violating the normality assumption of the ANOVA. Therefore, a Kruskal-Wallis H test was used to determine if there were differences in the performance score among the four instructional groups. The mean rank of performance scores was not statistically significantly different among groups, $\chi^2(3) = .789$, $p = .852$ (see Table 2).

This was unexpected because prior research suggests that subgoal labels benefit problem solving by helping learners to represent their problem solving knowledge in a way that allows more flexible transfer (e.g. Catrambone, 1998; Margulieux, 2013). For the main effect of subgoal labels, the present study showed $\eta^2_p = 0.003$, and the observed power was 0.09 compared to $\eta^2_p = .38$ found in Margulieux's (2013) study. The present study saw a very small effect size that would have needed a much larger sample to reveal any significant differences.

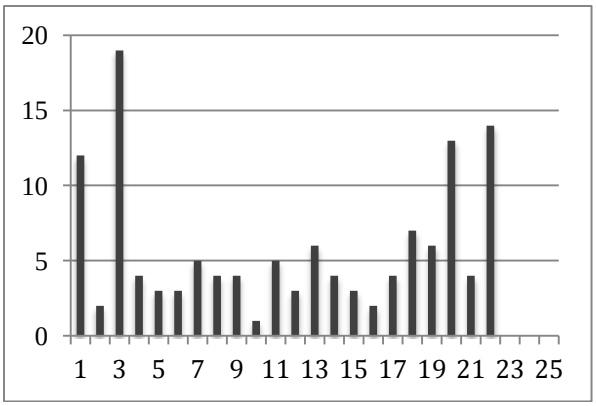


Figure 1. Distribution of scores for Problem Solving Task: Performance in AppInventor.

Written Performance. Participants were awarded one point for each correct step written towards achieving the problem solution for up to a maximum score of 22, and the ICC(A) for this assessment was .91. Visual inspection of the data revealed that the data were not normally distributed. The residuals did not have a normal distribution, violating the normality assumption of the ANOVA. Therefore, a Kruskal-Wallis H test was used to determine if there were differences in written performance score among the four instructional groups. The mean rank of the written performance scores was not statistically significantly different between groups, $\chi^2(3) = 1.64$, $p = .65$. These results did not support the hypothesis that instructional order and subgoal labels would affect the declarative knowledge concerning how to modify and add features to an app in App Inventor.

Explanation Task

In order to measure how well participants could organize and explain problem solutions, participants were given the solutions and instructed to meaningfully group and label the solution steps. Participants were awarded one point for each group that contained only structurally similar steps, for up to

Table 2. Descriptive Statistics for Problem Solving Task

	Worked Example First		*Expository First*	
	Subgoals	*No labels*	*Subgoals*	*No labels*
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
App Inventor Performance	11.41 (7.43)	10.25 (6.44)	10.75 (7.99)	10.37 (8.44)
Written Performance	10.91 (7.13)	10.15 (6.82)	8.67 (6.68)	10.42 (6.67)
Attempted Subgoals	6.41 (3.61)	6.06 (3.05)	5.75 (3.51)	6.22 (3.32)

Table 3. Descriptive Statistics for Explanation Task

	Worked Example First		*Expository First*	
	Subgoals	*No labels*	*Subgoals*	*No labels*
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Grouping	3.90 (1.88)	3.87 (1.78)	4.03 (2.01)	3.83 (1.91)
Explanations	1.46 (1.72)	1.24 (1.80)	1.26 (1.87)	1.45 (1.83)

Note: Scored out of a possible ten points.

a maximum of 10 points. ICC(A) for this assessment was .98. There were no significant differences on grouping structurally similar solution steps based on instructional material order, F (1, 116) = 0.02, p = .89, subgoal labels, F (1, 116) = 0.11, p = .74, or interaction , F (1, 116) = 0.06, p = .81 (see Table 3).Labels were scored for whether they described the function of the group of steps. For each label, participants earned one point if the explanation identified the purpose of the grouped steps. There were no significant differences based on instructional material order, F (1, 136) = 0.00, p = .98, subgoal labels , F (1, 136) = 0.00, p = .97, or the interaction , F (1, 136) = 0.47, p = .50. The hypothesis that the order the materials were presented, labeling of subgoals, and the interaction would effect performance on organizing and explaining problems solutions was not supported.

Discussion

The present study showed limited evidence that the instructional material order and subgoal labels affect a learner's performance in computer programming. This study suggests that similar learning occurs regardless of whether the worked example is presented before or after the expository instructions. The exception to this is that when asked to provide a general outline for creating an app, participants whose instructions contained subgoal labels and received the worked example before the expository instructions performed better than the other groups.

The reasoning behind presenting the worked example before the expository instructions was partly based on the literature about advance organizers. The benefit of an advance organizer lies on relating the new information to the existing cognitive structures. However, it is possible that the given instructions were not aligned with the participants' cognitive structures. The distribution of scores for the problem solving task in Figure 1 show that although some students did well, many performed poorly. It is plausible that the instructions might have been at an appropriate level for the high performers, but not for the low performers. For the participants who did not do well, the worked example might not have been able to bridge the gap between what the learners already knew and what they were about to learn. Instead, the instructions might have primarily been new information that was not easily anchored to existing cognitive structures. The inductive teaching literature shows that learners are unlikely to learn new information when there are few apparent connections to what the learner already knows. If the instructions were not at the proper level for the learner, then it follows that presenting the worked example first would have no added benefit.

Contrary to previous research such as Margulieux (2013), subgoal labels did not affect problem solving performance. There are several possible reasons that results in this study differed from results of previous research on subgoal labels. The main difference in research materials between this study and Margulieux (2013) is the media used for the expository instructions. Margulieux (2013) used a text document to convey this information, whereas the present study narrated the text document during a video. The use of a text document might have reduced the cognitive load as well as ambiguity of these instructions because the learner did not need to mentally transpose the text information to the App Inventor interface. Additionally, auditory information is more transient than text on a piece of paper; each piece of auditory information lasts for only a short period of time compared to text information that is continually present. Instructions presented through videos tend to be processed at a more superficial level than text instructions (Palmiter & Elkerton, 1993). Therefore, the subgoal labels in the videos might not have been processed to the same extent as when they were presented in a text document. As discussed previously, subgoal labels are thought to provide a framework for problem solving and aid in the creation of mental representations. However, if the information was not presented for a long enough duration, or processed to the necessary extent, the learner would not be able to form these connections..

Further Work

Further research should investigate the effectiveness of subgoal labels in videos compared to subgoal labels in text instructions, since the lack of a subgoal effect in the present study was surprising. Future research should also broaden the sample to include groups other than undergraduates to increase generalizability to other student groups. Additionally, this study focused on performance on the same day the task was learned. Testing after a delay would reveal how well the instructions were incorporated and applied long term. Much instruction aims to teach knowledge and skills that will be used not just on tasks on the day of instruction, but on future tasks. Investigating knowledge that is retained days and weeks after instruction

is more reflective of the real-world application of this type of instruction.

Acknowledgments

Thank you to Wendy Rogers and Frank Durso for their guidance, and to Lien Nguyen and Sarah Nay for their work on this project.

References

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: a meta-analytic review. *Educational Psychologist, 48*(2), 87-113.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart and Winston.

Catrambone, R. (1990). Specific versus general procedures in instructions. *Human-Computer Interaction, 5*, 49–93.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory & Cognition, 22*(5), 606-615.

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology, 87*(1), 5-17. doi: 10.1037/0022-0663.87.1.5

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127*, 355–376.

Catrambone, R., Gane, B. D., Adams, A. E., Bujak, K. R., Kline, K. A., & Eiriksdottir, E. (2012). Task Analysis by Problem Solving (TAPS): A Method for Uncovering Expert Knowledge. Unpublished manuscript, School of Psychology, Georgia Institute of Technology, Atlanta, GA.

Charney, D. H., & Reder, L. M. (1987). Initial skill learning: An analysis of how elaborations facilitate the three components. In P. Morris (Ed.), *Modelling cognition* (pp. 135–165). Chichester, UK: Wiley.

Dale, E. (1946). *Audiovisual Methods in Teaching.* Dryden Press, New York, NY.

Eiriksdottir, E., & Catrambone, R. (2011). Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors, 53*(6), 749-770. doi: 10.1177/0018720811419154

Hundhausen, C. D., Farley, S. F., & Brown, J. L. (2009). Can directed manipulation lower the barriers to computer programming and promote transfer of training?: An experimental study. *ACM Transactions in CHI, 16*(3). doi: 10.1145/1592440.1592442

LeFevre, J.. & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction, 3*, l-30. doi: 10.1207/s1532690xci0301_1

Margulieux, L. E. (2013) *Subgoal Labeled Instructional Text and Worked Examples in STEM Education.* Unpublished master's thesis, Georgia Institute of Technology, Atlanta, Georgia.

Margulieux, L. E., Guzdial, M., & Catrambone, R. (2012). Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. In *Proceedings of the Ninth Annual International Conference on International Computing Education Research*, 71-78. doi: 10.1145/2361276.2361291

Novak, J.N. (1977). *A Theory of Education*. New York: Cornell University Press.

Palmiter, S., & Elkerton, J. (1993). Animated demonstrations for learning procedural computer-based tasks. *Human-Computer Interaction, 8*(3), 193-216. doi:10.1207/s15327051hci0803_1

Palmiter, S., Elkerton, J., & Baggett, P. (1991). Animated demonstrations versus written instructions for learning procedural tasks: A preliminary investigation. *International Journal of Man-Machine Studies, 34*, 687-701. doi: 10.1016/0020-7373(91)90019-4

Pea, R. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences, 13*(3), 423-451. doi: 10.1207/s15327809jls1303_6

Pirolli, P., & Recker, M. M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235–275.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. *Journal Of Experimental Psychology: Learning, Memory, And Cognition, 11*(1), 106-125. doi:10.1037/0278-7393.11.1.106

Ross, B. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Ross, B. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 456-468.

Rumelhart, D. E., & Norman, D. A. (1981). Analogical processes in learning. In J. R. Anderson (Ed.), Cognitive skills and their acquisition (pp. 335-360). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Trafton, J. G., & Reiser, B. J. (1993). *Studying examples and solving problems: Contributions to skill acquisition*. Technical report, Naval HCI Research Lab, Washington, DC, USA.

Wiedenbeck, S. (1989). Learning iteration and recursion from examples. *International Journal of Man-Machine Studies, 30*(1), 1–22.

No Effect of Verbal Labels for the Shapes on Type II Categorization Tasks

Fotis A. Fotiadis (f_fotis@phs.uoa.gr)
Department of Philosophy & History of Science, University of Athens, Ano Ilissia Campus
GR-157 71, Athens, GREECE

Athanassios Protopapas (aprotopapas@phs.uoa.gr)
Department of Philosophy & History of Science, University of Athens, Ano Ilissia Campus
GR-157 71, Athens, GREECE

Abstract

Category learning is thought to be mediated—in at least some category structures—by hypothesis-testing processes. Verbal labels for the stimuli and stimulus individuation have been shown to facilitate the formation, testing, and application of rules of category membership (Fotiadis & Protopapas, 2014). We sought to replicate the phenomenon of facilitation due to verbal names for the stimuli by training participants for two consecutive days to either learn new names for abstract shapes, or learn shape-ideogram pairings; a third group was unexposed to the shapes. After training, participants were given a Type II categorization task—thought to be mediated by verbal processes of rule discovery—utilizing the trained shapes. We hypothesized that verbal labels for the shapes and shape individuation would provide facilitative effects in learning to categorize. Results revealed no effect of training on categorization performance. This study suggests that caution should be taken when generalizing findings across perceptual modalities or different experimental paradigms.

Keywords: Verbal labels; hypothesis testing; categorization; learning

Introduction

The ability to categorize spans a broad range of human capacities and behaviors. Researchers have examined the cognitive processes (Ashby & Maddox, 2005) and neural substrates (Poldrack et al., 2001) of category learning and have utilized computational modeling techniques in an effort to shed light on the nature of the underlying representations (Anderson, 1991).

The Multiple Memory Systems (MMS) hypothesis argues that human category learning is mediated by distinct learning systems (Ashby & Maddox, 2005; Poldrack & Foerde, 2008). A declarative, explicit, or verbal system is thought to be engaged in the learning of categories that can be characterized by a verbal rule. Hypothesis testing processes are thought to be recruited, and the knowledge acquired is thought to be available to consciousness. On the other hand, the learning of categories that defy a simple verbal description is thought to be accomplished through a procedural, implicit, or non-verbal system. Pre-decisional perceptual processes underlie learning, and the learned material is thought to be unavailable to consciousness. An on-going debate exists between the MMS theorists and single-system theorists arguing that a single, general learning mechanism suffices to account for behavioral data (e.g., Newell, Dunn, & Kalish, 2011).

In the context of this debate, growing empirical evidence suggests that verbal processes are important in the learning of rule-described categories. Ashby and colleagues (Ashby, Alonso-Reese, Turken, & Waldron, 1998) developed a computational theory suggesting that the verbal system mediates rule-based category learning. Verbal working memory interference has been found to impair the learning of rule-described categories (Miles & Minda, 2011), whereas experimental manipulations, such as using difficult-to-name stimuli (Kurtz, Levering, Stanton, Romero, & Morris, 2013), or verbal rehearsal of stimulus dimensions prior to learning (Minda, Desroches, & Church, 2008), have been shown to affect category learning.

Verbal Labels in Hypothesis Testing

Recently, Fotiadis and Protopapas (2014) provided evidence in favor of the hypothesis that verbal labels for the to-be-categorized stimuli facilitate hypothesis-testing processes underlying category learning. The authors utilized hard-to-name auditory stimuli and manipulated the availability of stimulus names by training separate groups of participants for three consecutive days to associate the auditory tones with pseudowords (label training condition) or with hard-to-name ideograms (ideogram training condition); or to associate tone intensity with colors (intensity training condition); a fourth group remained unexposed to the tones (no-training condition). On the fourth day all participants were administered the same auditory version of the Weather Prediction Task (Knowlton, Squire, & Gluck, 1994) utilizing the trained tones as cues. Results revealed a gradation in categorization performance in the order: label > ideogram > intensity > no-training. Thus, it was concluded that verbal labels, cue individuation, and exposure to the stimulus set each facilitated explicit hypothesis-testing processes underlying category learning.

The Shepard et al. (1961) Tasks

In their seminal paper, Shepard, Hovland and Jenkins (1961) revolutionized the study of category learning. They created six category tasks (Type I to Type VI) by manipulating category structure (categorization rule) while utilizing the same stimuli in each task and the same number

of exemplars in each category. In the most common implementation of the paradigm (Minda & Miles, 2010) categorization stimuli are comprised of three binary valued dimensions: Shape (square vs. triangle), color (black vs. white), and size (big vs. small).

The basic finding of the Shepard et al. (1961) study was that the order of difficulty of the six types (as assessed by participants' performance) cannot be accounted for by a simple stimulus-generalization theory. The key finding was that participants found it easier to learn Type II categories compared to Type IV categories, despite the reduced within-category similarity of the former (compared to the latter) category structure. The authors suggested that this Type II over Type IV advantage necessitates considering the mediation of executive attention mechanisms and the formulation and application of rules during category learning.

The Type II Task and Rule-Discovery Learning

The Type II task has a two-dimensional rule structure. Two out of the three dimensions are diagnostic of category membership[1], in an exclusive-or fashion. A simple verbal rule seems to be able to define category membership (e.g., "black triangles and white squares are category A"). Thus, the structures' processing demands are thought to be best met by explicit rule-learning processes (Minda & Miles, 2010).

This claim seems to be supported by empirical evidence. Minda et al. (2008) utilized the first four prototypical Shepard et al. (1961) category structures in an effort to examine rule-selection executive functions of children and adults. In their Experiment 2, Minda et al. assessed the effect of a concurrent verbal and a concurrent non-verbal task on categorization performance. The verbal secondary task—thought to occupy resources recruited by verbal processes of rule discovery—did impair performance in the Type II structure (compared to a control, no-task condition, and also compared to the non-verbal task condition). These results suggest that the Type II structure recruited the explicit system. Smith, Minda and Washburn (2004) studied category learning processes of human and non-human animals using the Shepard et al. tasks. Their results provided evidence in favor of the engagement of rule-discovery mechanisms in learning the Type II category structure. The evidence (**"**all-or-none learning") was only present for human subjects, whereas for non-human animals, lacking the faculty of language, there was no sign of rule discovery. This may be considered as evidence in favor of the engagement of rule-learning mechanisms in the Type II task.

Thus, theoretical reasons (Minda & Miles, 2010; Shepard et al., 1961) as well as empirical evidence (Minda et al., 2008; Smith et al., 2004) suggest that learning to categorize in the Type II category task is mediated by hypothesis testing processes of verbal rules.

[1] Depending on which dimensions are diagnostic, there can be three Type II subtypes: Shape-irrelevant, size-irrelevant, and color-irrelevant. See Love and Markman (2003) for evidence suggesting that performance varies systematically across these subtypes.

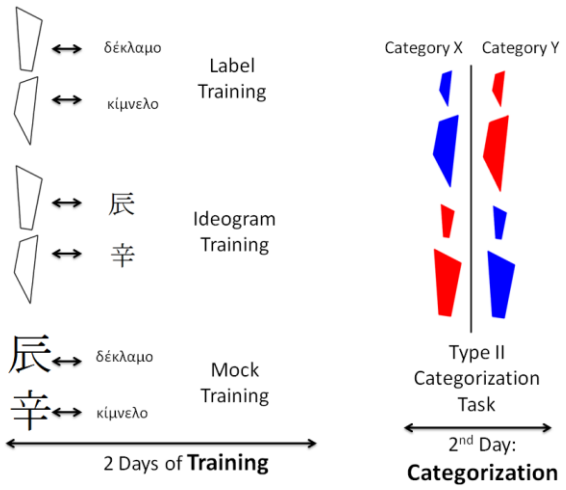


Figure 1: Design of the present study.

Rationale of the Present Study

The purpose of the present study was to further test the hypothesis that verbal labels for the to-be-categorized stimuli facilitate hypothesis testing processes recruited during category learning (Fotiadis & Protopapas, 2014). We specifically wanted to test our training manipulation in the visual modality, since there are reasons to suggest that learning across modalities is not governed by the same mechanisms (Conway & Christiansen, 2005). Moreover, given that in the Weather Prediction Task category membership is probabilistically defined (Knowlton et al., 1994) we sought to examine the effect of names for the stimuli in a task with a deterministic structure.

To manipulate the availability of names, separate groups of participants were trained for two consecutive days to associate abstract shapes with pseudowords (label training condition) or with hard-to-name ideograms (ideogram training). A third—control—group of participants remained unexposed to the shapes, and was trained to associate ideograms with pseudowords (mock training condition). On the second day, all participants were given the Type II categorization task. In this task, the two values in the shape dimension were the same shapes that were used in the training procedure. The category-diagnostic dimensions were shape and color, whereas size was non-diagnostic (see Fig. 1).

We reasoned that verbal labels for the shapes would facilitate verbal hypothesis-testing processes of rule formation, testing, and application. Therefore we predicted that participants in the label training condition would find it easier to discover the categorization rule, compared to the ideogram training group. We also hypothesized that familiarity with the stimuli and learning to associate the shapes to visual stimuli (ideograms) would help create individuated perceptual representations of the shapes and

therefore facilitate categorization. We therefore predicted that the ideogram training group would have an advantage in rule discovery compared to the mock training group.

Methods

Participants

Seventy-two students (16 male) of the University of Athens took part in the study and were randomly assigned (in groups of 24) to each training condition. Their mean age was 25.8 years (*SD* = 7.0). All were native speakers of Greek, reported normal or corrected-to-normal vision, no history of neurological illness, and no diagnosis of dyslexia.

Materials

Shapes. Two abstract shapes of low association value were selected from the collection of Vanderplas and Garvin (1959). The shapes have been previously used in experimental research and are considered to be hard to name (e.g., Hulme, Goetz, Gooch, Adams, & Snowling, 2007). The shapes were equated in size in a pilot experiment using the method of adjustment. Twelve participants took part in this psychophysical procedure, which was implemented in PsychoPy (Peirce, 2007), and the results provided the Points of Subjective Equality (P.S.E.s). For the training session, empty shapes with a black margin were created, with size corresponding to 75% of the P.S.E.s, whereas for the categorization session the shapes were filled with red or blue color. The categorization stimuli necessitated two levels of size for the stimuli, so for the "big" shapes the size corresponded to the P.S.E.s, and the "small" shapes were created by a 50% reduction in size.

Pseudowords. Ten pseudowords were created, equated in number of letters, syllables, phonemes, and stress position. They were also roughly equated in orthographic and phonological typicality using Levenstein distance of the 20 nearest neighbors (Protopapas, Tzakosta, Chalamandaris, & Tsiakoulis, 2012; Yarkoni, Balota, & Yap, 2008). To avoid name assignment biasing toward particular shapes, we administered an online questionnaire to 107 native speakers of Greek showing randomly one of the two abstract shapes along with the ten candidate pseudowords. Participants were simply asked to "choose a name" for the shape. We selected the two pseudowords that were selected as names for both shapes with roughly equal frequency, namely δέκλαμο (/'ðɛklamo) and κίμνελο (/'kimnɛlo).

Ideograms. Two Chinese characters were selected. These ideograms have been previously used and have been shown to resist a simple verbal description (Fotiadis & Protopapas, 2014): 辛 (U+8F9B), and 辰 (U+8FB0). To equate for number of strokes (and, thus, for perceptual complexity), a stroke was erased from the second character.

Procedure

Training comprised two sessions, administered on two consecutive days. On the second day, following training, the categorization task was administered. All following procedures were implemented in the DMDX display software (Forster & Forster, 2003).

Training

There were two training sessions, administered on consecutive days, aimed to allow overnight consolidation. Participants were given 160 trials in each training session, arranged in four blocks of 40 trials. At the beginning of a *Label Training* trial a fixation cross was presented for 500 ms at the center of the screen. Following that, one of the two shapes was randomly selected and presented for 2000 ms, and then the two pseudowords appeared in a vertical configuration. Participants were asked to respond by clicking on one of the two alternative responses (pseudowords). Upon response, feedback was provided (the word **"**Correct" or **"**Wrong") for 500 ms. The permutation of the two pseudowords was counterbalanced across trials, and each shape was presented equally often within a block of trials. On the first day of training, participants in the Label training condition were asked to read aloud the pseudoword of their choice before clicking on it, because we reasoned that learning a name necessitates the formation of an effective phonological component. This reading-aloud instruction was omitted on the second day, to equate task demands across training conditions as much as possible.

For the *Ideogram Training* condition pseudowords were replaced with ideograms. In the *Mock Training* condition participants were asked to learn to associate ideograms to pseudowords, so shapes were replaced by ideograms. No reading aloud took place in either the Ideogram or the Mock training conditions. Stimulus-response pairings (e.g., shape-pseudoword or shape-ideogram pairings) were counterbalanced across participants within a training condition. Each training session lasted approximately 20 minutes.

Categorization

In the Categorization session, which followed immediately after the second training session, participants were told that they would learn to classify stimuli into two categories, namely X and Y. They received a maximum number of 28 blocks of eight trials. In a categorization block each of the eight categorization stimuli was presented once. The beginning of a trial was signaled by the presentation of the stimulus at the center of the screen along with the category labels X and Y. The category labels were presented around the stimulus either horizontally or vertically in both permutations (i.e., X - Y or Y - X), providing four possible configurations. Participants responded by clicking on a category label. Following each response a smiling face was presented if correct or a frowning face if incorrect, for 1500 ms. If a participant did not respond within 10000 ms the

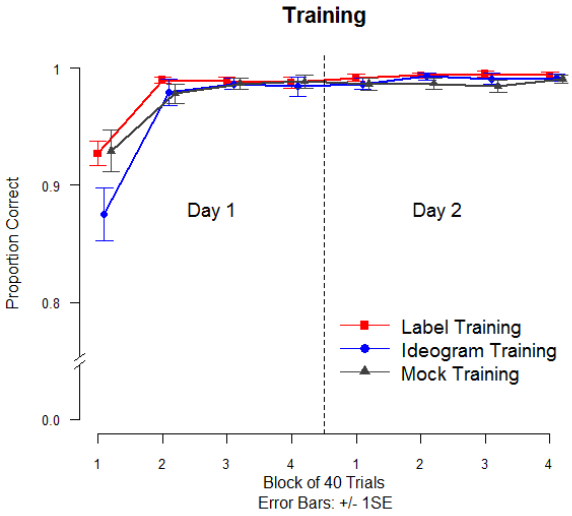


Figure 2: Learning curves in the three training conditions in blocks of 40 trials

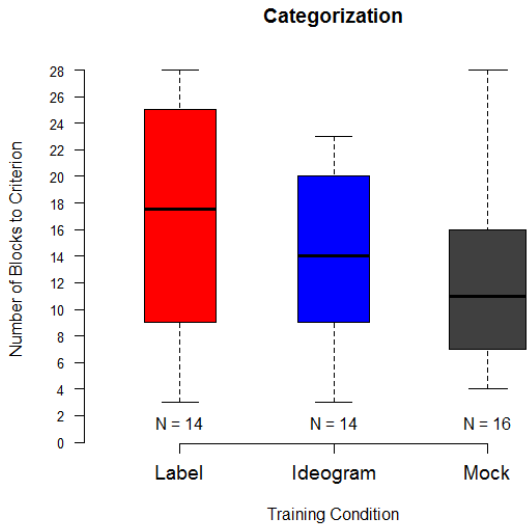


Figure 3: Number of blocks to reach learning criterion in the categorization task, per training condition. Data from learner participants only. Boxes denote interquartile range; thick lines mark the median; error bars extend to the full range; N denotes sample size.

trial was terminated and a prompt appeared on screen. The training session ended upon completion of all blocks, or upon two consecutive errorless blocks (following Mathy, Haladjian, Laurent, & Goldstone, 2013). The order of trials was pseudorandomized with MIX (VanCasteren & Davis, 2006) and was identical for all participants. Randomization constraints precluded (a) presentation of the same categorization stimulus on two consecutive trials, and (b) a lag between trials with the same configuration of category labels less than two. The assignment of category label (X-Y) to values of the diagnostic dimensions was counterbalanced across participants. A short break was provided every 56 trials. The maximum duration of the categorization session was 25 minutes.

Results

Training

Participants in all three training conditions mastered the training task by the third block of Day 1 and exhibited ceiling performance on Day 2. Across both training sessions, participants averaged 98.33% correct responses (*SD* = 1.16) in the label training condition, 97.33% (*SD* = 2.64) in the ideogram training condition, and 97.87% (*SD*= 2.29) in the mock training condition. The average performance per training condition and day, in blocks of 40 trials, is shown in Fig. 2.

The purpose of the analysis was to test if participants were equally successful in learning the shape-label and shape-ideogram pairings. Therefore, we only analyzed data from the label and ideogram training conditions on the second day of training. Participants' responses were analyzed in R (R Development Core Team, 2014) with a linear mixed-effects model including fixed effects of trial and training condition, as well as their interaction, and random effects of participants. By-participant random slopes of trial were included to model participants' individual learning rates (Baayen, Davidson, & Bates, 2008). In R notation, the model was specified as

```
accuracy ~ trial*condition+(1+trial|participant).
```

There was no simple effect of trial (β = .387, z = 1.317, p = .188), a result consistent with participants' ceiling performance on Day 2. There was also no interaction of trial by condition (β = −.031, z = −.113, p = .910) and—most importantly—no effect of condition (β = .369, z = .692, p = .489).

Categorization

Out of 72 participants in all training conditions, 44 (61.11%) managed to achieve two consecutive errorless blocks of trials, thus providing unequivocal evidence of having discovered the categorization rule. We refer to these participants as **"**learners.**"** There were 14 learners (58.33%) in the label training condition, 14 learners (58.33%) in the ideogram training conditions, and 16 learners (66.67%) in the mock training condition. A chi square test revealed that the percentage of learners in the categorization task did not differ significantly between training conditions (χ^2 = .468, df = 2, N = 72, p = .792).

However, using the percentage of participants reaching the learning criterion as a dependent variable has the disadvantage of disregarding the ease or difficulty with which participants in each training condition learned the rule. We therefore analyzed the number of blocks to reach criterion, for learner participants only[2] (see Fig. 3). An one-

[2] The exclusion of "non-learner" data is common practice in the categorization literature when analyzing number of blocks to reach criterion (e.g., Mathy & Feldman, 2009). The rationale is that a

way analysis of variance revealed that there was no effect of training condition on the number of blocks to reach the learning criterion, $F(2, 42) = 1.777$, $\eta^2 = .080$, $p = .182$.

Alternatively, categorization performance can be analyzed using accuracy as the dependent variable. This allows inclusion of all participants, under the assumption of errorless performance after the learning criterion is reached (e.g., Kurtz et al., 2013). A linear mixed-effects model with the same formula as above revealed an effect of trial ($\beta = 4.878$, $z = 5.33$, $p < .001$), reflecting an increase in accuracy as trials progressed, comparable learning rates among conditions (all βs < 1.35, $p > .14$), and—most importantly—no effect of condition on categorization accuracy (all βs < .13, $p > .32$).

Discussion

In this study we trained participants for two consecutive days to learn new names for shapes, or learn to associate shapes with hard-to-name ideograms. A third group of participants remained unexposed to the shapes. In a categorization task, administered immediately after training, we used the trained shapes to create the categorization stimuli. We predicted that names and familiarity with the shapes would each facilitate rule discovery in the categorization task. Our results revealed no effect of training condition of categorization, in contrast to previous findings (Fotiadis & Protopapas, 2014).

This discrepancy raises concerns about assuming that an effect manifesting itself in one modality would also be present in another modality. One purpose of the experiment was to replicate the effect of facilitation in learning to categorize due to names for the stimuli in the visual modality. The lack of an effect may be attributed to the change in modality per se, since there is reason to assume that learning processes may differ between modalities. Saffran (2002) showed that, for learning to take place, the temporal mode of presentation of stimuli in the visual and auditory modality should be different (concurrent vs. sequential respectively). Also, Conway and Christiansen (2005) implemented the same learning paradigm in different modalities and provided evidence in favor of a learning advantage in the auditory modality compared to the visual modality. Further empirical investigation is needed to assess whether learning to categorize in the auditory and the visual modality is mediated by the same processes.

Participants' ceiling performance during training complicates interpretation, insofar as potential differences between learning the shape-label and shape-ideogram pairings may be masked by the ease of the task. Thus, we cannot preclude the possibility that performance in the categorization task is affected by differences in training.

Alternatively, the lack of an effect of verbal labels in category learning may stem from methodological discrepancies between the present and our previous study, such as the structure of the categorization task used to reveal hypothesis learning processes. The Weather Prediction Task, previously shown to be affected by names for the stimuli, has a probabilistic structure, whereas the Type II task used in the present study has a deterministic structure. It remains to be investigated whether performance in a probabilistic category structure may be more easily affected by experimental manipulations, perhaps due to the uncertainty that is inherent in the task.

Further concerns stemming from the results of the present study are related to whether changing the surface structure of a paradigm affects the processing demands of a task. The result of no difference in categorization performance between the label and ideogram training groups might suggest that names for the stimuli do not facilitate rule discovery. An alternative explanation, however, may be related to the fact that our implementation of the Type II task utilized two abstract shapes whereas the canonical version uses two geometric shapes. It may be that the Type II task is learned through verbal processes of rule discovery only when the values of the diagnostic dimensions are highly familiar to participants. Mathy et al. (2013), who also used abstract shapes in implementing the Type II task, provided evidence in favor of the engagement of similarity-based processes (thought to reflect learning mediated by the implicit rather than the rule-based system) in learning to categorize. Thus, although the Type II task has been used to examine explicit processes (e.g., Minda & Miles, 2010), our version of the task may have recruited implicit processes that are not affected by verbal labels for the stimuli.

A final concern may be of representational nature. The finding that familiarity with the stimuli also failed to affect performance in the categorization task seems rather puzzling, given previous findings and current understanding in the field. For example, Folstein, Gaultier, and Palmeri (2010) provided evidence suggesting that mere exposure to the stimulus configuration may facilitate subsequent categorization performance. Our finding of no significant difference in performance between the ideogram and mock training conditions may be taken to indicate that learning processes involved in learning to categorize our version of the Type II task did not recruit the representations of the shapes that were presumably acquired during training. Indeed, informal reports of participants' strategies in debriefing revealed that participants mainly paid attention to the corners of the shapes and not to the shape forms in their entirety. Therefore, a plausible explanation for our findings is that the participants learned names for the entire shapes and formed individuated representations of them but then only used parts of the shapes in the categorization task. The representational mismatch undermined the potential of the verbal labels and the familiarity with the shapes to facilitate learning in the categorization task.

To conclude, we sought to replicate the effect of

value of 28 corresponding to a non-learner, perhaps responding at chance, and a value of 28 corresponding to a participant mastering the task at the last two blocks reflect qualitatively different behaviors that should not be aggregated.

facilitation in learning a verbal rule of category membership caused by having names for the stimuli. The results suggest that learning processes may operate differently across modalities or across categorization paradigms and that task processing demands may be significantly altered if the surface structure of a categorization paradigm is modified.

Acknowledgments

We thank Laoura Ziaka for help with collecting data.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442–481.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 24–39.

Folstein, J., Gauthier, I, & Palmeri, T. J. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology, 1*, 1–6.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*(1), 116–124

Fotiadis, F. A., & Protopapas, A. (2014). The effect of newly trained verbal and nonverbal labels for the cues in probabilistic category learning. *Memory & Cognition, 42*(1), 112–125.

Hulme, C., Goetz, K., Gooch, D., Adams, J., & Snowling, M. J. (2007). Paired-associate learning, phoneme awareness, and learning to read. *Journal of Experimental Child Psychology, 96*(2), 150–166.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory, 1*(2), 106–120.

Kurtz, K. J., Levering, K., Romero, J., Stanton, R. D., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 552–572.

Love, B. C., & Markman, A. B. (2003). The nonindependence of stimulus properties in human category learning. *Memory & Cognition, 31*(5), 790–799.

Mathy, F., Haladjian, H. H., Laurent, E., & Goldstone, R. L. (2013). Similarity-dissimilarity competition in disjunctive classification tasks. *Frontiers in Psychology, 4*(26), 1–14.

Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review, 16*(6), 1050–1057.

Miles, S. J., & Minda, J. P. (2011). The effect of concurrent verbal and visual tasks on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 588–607.

Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and nonrule- described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1518–1533.

Minda, J. P., & Miles, S. J. (2010). *The influence of verbal and nonverbal processing on category learning*. In B. H. Ross (Ed.), The Psychology of learning and motivation: Advances in research and theory, 52, 117–162. San Diego: Academic Press.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). *Systems of category learning: Fact or fantasy?* In B. H. Ross (Ed.), The psychology of learning and motivation (Vol. 54, pp. 167–215). San Diego: Academic Press.

Peirce, JW (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8–13

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso, J., Myers, C. E., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature, 414*, 546–550.

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory system debate. *Neuroscience and Biobehavioral Reviews, 32*, 197–205.

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources and Evaluation, 46*(3), 449–459.

R Development Core Team. (2014). R*: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47*, 172–196.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*(13), 1–42.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in Rhesus Monkeys: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General, 133*(3), 398–414.

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38*(4), 584–589.

Vanderplas, J. M., & Garvin, E. A. (1959). The association value of random shapes. *Journal of Experimental Psychology, 57*, 147–154.

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.

Context-dependent Processes and Engagement in Reading Literature

Miho Fuyama (miho02@sj9.so-net.ne.jp)
Graduate School of Media and Governance, Keio University
5322 Endo, Fujisawa, Kanagawa 252-0882, Japan

Shohei Hidaka (shhidaka@jaist.ac.jp)
School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

Abstract

It does not do the act of reading literature any justice to describe it as simply processing text to acquire information or knowledge. We enjoy reading stories, we become *absorbed* in them. Our absorption into stories is related to their contextual structure. We develop a statistical method for the analysis of reading time distributions which allows us to assess the context of a story rather than merely its text. This analysis detects statistically distinct distributions of reading times, with each distribution representing a distinct process or mode of reading. Our experiments support the hypothesis that the temporal change in these modes of reading are related to changes in the degrees of absorption of the subjects and also in the contextual structure of the stories being read.

Keywords: Reading; Literary; Reading-time analysis

Introduction

Reading literature is not just merely the process of attaching literal meaning to prose, but it also evokes various experiences and can affect the way the reader views the world around them. Past studies have argued the potential effects of literary works, and tried to empirically measure their effects on readers (Miall & Kuiken, 1994; Iser, 1976, and see also Miall 1999 for discussion about literariness).

In particular, it has been empirically shown that the context or story plot has substantial impact on readers (Miall, 1988; Rapp & Gerrig, 2006). This finding is related to an earlier argument by Aristotelis (Kassel, 1965). The present study follows this line of research, and investigates the effects of context by analyzing temporal changes of readers' responses in reading.

In past studies, hypothetical constructs such as *story grammar* or *script*, which readers supposedly process while reading, were employed to capture the contextual structures of stories (Thorndyke, 1977; Beaugrande, 1982). Introducing the concepts of story grammar and script, these studies claimed that narratives have their own internal structure like a grammar but at the discourse level, and these structures can be expressed as combinations of elements such as setting, theme, characters, or goals. However, using such frameworks allows us to analyze only limited classes of stereotypical stories, such as folk tales(Miall, 1989; Beaugrande, 1982). These limitations are due to the inflexibility of the constructed frameworks. There have been other approaches to capture the contextual structure of stories but none of them has yet offered a satisfactory description.

In the present study, rather than assuming a specific story grammar, we focus on temporal changes in reading process across different contextual structures of stories. Miall (1988) analyzed the relationship between readers' ratings and reading times of introductory sections of novels. He analyzed readers' responses by assuming two stages in the reading process. The first was called the "registration stage", in which readers form expectations about the likely meaning of narrative, and the second was called the "interpretation stage", in which readers use their expectations to comprehend the narrative. Miall supposed shifts between these stages would depend on both the context of the story and a reader's background knowledge. From his analysis, he concluded that the reading process form a cycle and repeatedly shifts between the two stages in the reading of a narrative.

Taking as a working hypothesis that there are separable stages in reading processes correlated to the contextual structure, as suggested by Miall, we investigate the stage-like changes involved in reading an entire story, not just an introductory portion.

In this study, we investigate the "modes" of reading processes, which are supposedly correlated to the context of the story or a reader's background knowledge. The "modes" are operationally defined by statistical properties of reading times of each unit of text. We will further discuss these modes in the next section.

Statistical analysis of reading time

We assume the reading processes is composed of several qualitatively distinct subprocesses, and we call such a subprocess *reading mode*. The question is, given reading data, how we can infer the number of reading modes reflected in the data? In previous studies about short texts and more rapid processes, differences in reading time alone have been interpreted as reflection of two qualitatively distinct processes (Miall, 1988; Gernsbacher, 1997). However, reading time may vary depending on multiple factors like frequency, familiarity, and the lengths of words (Inhoff & Rayner, 1986; White, 2008). We cannot, therefore, naively interpret reading time alone as an indicator of multiple reading modes.

This observation motivates the development of a new technique for the analysis of reading time. The analysis we present as an alternative is based on a statistical theory of processing time (Hidaka, 2013). In this theory, the presence of multiple different modes of processing can be detected by the statistical distribution of the processing time. If the reading

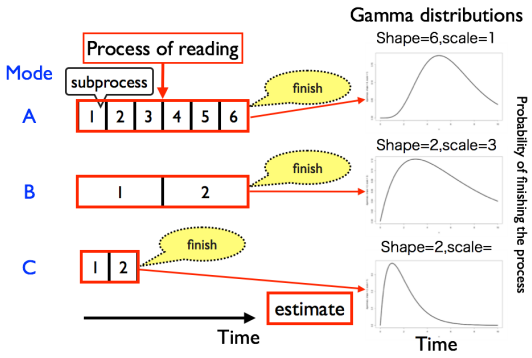


Figure 1: Three hypothetical processing modes, A, B, and C, which have different numbers of subprocesses with different average rates. In each mode, all the subprocesses run in the serial order, and its reading time follows a gamma distribution. Mode A: six subprocesses, each takes short time on average, Mode B: two subprocesses, each takes long time, and Mode C: two subprocesses, each takes short time. The average of both Mode A and B is the same, but their distributions (on the right hand side of the figure) are different.

process consisted of n subprocesses with the same constant processing rate over time, and finished only when all these subprocesses have finished, the reading time would follow a gamma distribution with shape parameter n. If, on the other hand, the reading process consisted of one subprocess with process rate t^k as a function of the process time t, then the process finished when at least one subprocess has finished, reading time would follow a Weibull distribution with shape parameter k.

This statistical analysis allows us to distinguish processes which have a same average speed of processes but have different number of subprocesses (Figure 1, A and B), and to distinguish processes which have same number of subprocesses but have different average speed of processes (Figure 1, B and C). This subprocess estimation gives an advantage over the previous studies analyzing differences in the reading time alone.

We adopt this statistical account of processing time in evaluating the number of reading modes based on reading time. If each observation in a reading time dataset follows essentially the same distribution as the others, we would treat this as an indicator of a single reading mode. If, on the other hand, the data set appears to have been generated by sampling from a mixture of distributions, we treat it as an indicator of multiple reading modes (Figure 2). Each dataset in question is composed of observations about a single subject. This technique therefore removes overall reading speed as a factor in the analysis.

Approach

When reading, one is generally also engaged in many other processes – eye movements, posture management, etc. If one were only lightly engaged in reading and more heavily preoccupied with a number of these other activities, it is entirely possible that their preoccupation could appear as distinct reading modes in our statistical analysis. To prevent the detection of such false modes, it would be valuable to have a measure of reading engagement independent from reading time. We could then test the results of our statistical analysis based on their correlation with that measure.

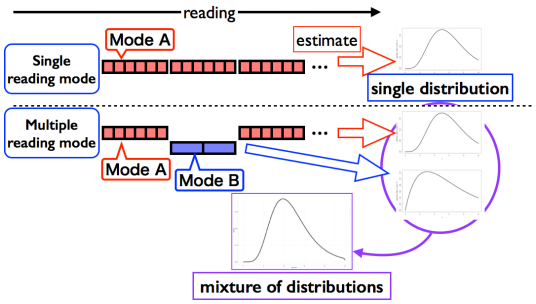


Figure 2: (top) if one type of reading mode is repeated across multiple pages, it would result in a gamma distribution. (bottom) if both two types of reading modes, A and B, take place across different pages, it would result in a mixture of two gamma distributions.

Since the analytic technique we will use is statistical in nature, it requires relatively large datasets in order to produce meaningful results. To this end, and although this is not typical of existing studies of reading, we use entire novels as the texts in our experiments.

Given the burden that reading such long texts places on the subjects of our experiments, our first experiment consisted of only one subject – Miho Fuyama, the first author of this paper. She is an avid reader, which suggests that she is generally easily engaged in reading as an activity. In Experiment 1, we studied her reading time and the degree of engagement in reading, in order to empirically establish the validity of our analysis and test whether her reading process has a single or multiple reading modes.

Having validated our statistical analysis, we adopted it in our second experiment to a cross sectional study of multiple subjects. In Experiment 2, we asked ten subjects to read a short novel. This experiment was designed to evaluate whether our findings from Experiment 1 hold in general. We also evaluated changes in reading modes could be related to the semantic structure of the text itself. To do so, we analyze the consistency of the dynamics governing the change of reading modes across subjects and treat the consistent dynamics as text-specific semantic effects in reading.

Experiment 1

The first author was the sole subject of several high-load reading tasks. We asked her to read 20 Japanese novels. Each session took one day including breaks. The set of samples from these 20 sessions of 20 novels was submitted to statistical analysis using the scheme described in the previous section, and we estimated the statistical distribution of her reading time for each two pages. For two of the novels (novels 17 and 18 in Table 1), she evaluated her degrees of absorption

Table 1: The novels read in Experiment 1.

No.	Title (Abbreviated)	Author	Page length
1	Shikisai	H. Murakami	370
2	Kamisama	H. Mori	314
3	Nameraka	H. Kawakami	189
4	Tenchi	T. Ubukata	474
5	Chinmoku	Y. Ogawa	308
6	Hikari	S. Miura	297
7	Kuchi	M. Banto	309
8	Mizuumi	B. Yashimoto	206
9	Kogoeru	A. Shino	401
10	Self-Reference	T. Enjo	308
11	Shi no izumi	H, Minagawa	427
12	Kisetsu no kioku	K. Hosaka	316
13	Eien no deguchi	E. Mori	313
14	Hokanaranu hito he	K. Shiraishi	295
15	Shorou tomurai dou	N. Kyogoku	498
16	Kodoku no utagoe	A, Tendo	312
17	Neko	Y. Ogawa	359
18	Ruto 225	C. Fujino	282
19	Yasashii uttae	Y. Ogawa	260
20	Burahuman	Y. Ogawa	146

each two pages as an indicator of her engagement to reading. Specifically we asked her how absorbed she was in reading every pair of pages in these novels. These absorption ratings were used to validate the statistical analysis.

Participant

The subject was the first author, Miho Fuyama, who was 30 years old when the experiment was conducted. She is a native Japanese speaker, is a regular reader, and has normal vision.

Material

We used 20 Japanese novels, which the first author read for the first time in this experiment. The titles, authors, and page lengths of books are listed in Table 1. We selected as texts books written by authors who have won Japan's prestigious literature prizes, such as the Naoki Prize or Akutagawa Prize.

Procedure

In each session of the experiment, the subject was asked to read a novel. Each session lasted several hours (including breaks), but was completed in one day. The subject reported her degrees of absorption for every two pages read in novels number 17 and 18. These reports were made approximately 100 days after the reading sessions. Her degree of absorption was measured on a five-level scale – "extremely bored", "bored", "normal", "absorbed", and "deeply absorbed". This scale was coded using the numbers $-2, -1, 0, 1$, and 2 respectively for each of the states. As the experiment required her to focus on and to become absorbed in such long texts, the subject was allowed to perform her readings at her home in order to minimize her tension. She was also allowed to have breaks whenever she wanted. The breaks were typically 5 to 15 minutes long, but there were also several hour-long lunch breaks. While reading, she sat at her desk and was videotaped with two small web cameras.

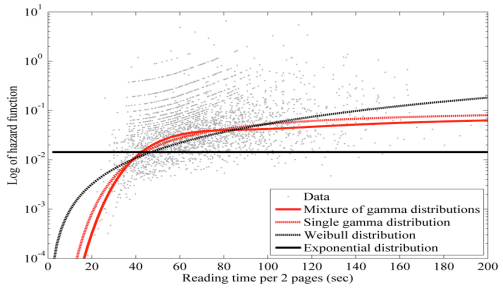


Figure 3: The hazard function for the sample (dots) and for the estimated probability distributions (lines) of reading time.

Analysis

From the videos, we transcribed the reading time for each pair of pages. These reading times were measured as the lengths of time between page turns, excluding time spent on break. The statistical analysis is performed on these transcribed reading times. We analyzed the aggregate of the data gathered across all the sessions of the experiment in order to increase the statistical power of our analysis.

In our analysis, we fitted mixtures of exponential distributions, those of Weibull distributions, and those of gamma distributions to the aggregate data. For each mixture distribution, ranging from 1 to 5 components, we estimated the parameters by maximizing likelihood. As these statistical models have different numbers of parameters, we chose the model with the smallest Bayesian Information Criterion (BIC) (Schwarz, 1978) as the one which best explains the data.

Results and discussion

We found that a mixture of two gamma distributions provides the best fit to the aggregate data amongst all the distributions considered. Figure 3 illustrates the differences between these various classes of distributions in explaining our data. It shows the hazard function $H(t)$ of the page-turn interval t. The hazard function $H(t)$ is the probability (density) to finish reading on condition of the reading being unfinited unitl t. Weibull, gamma and exponential distributions show qualitatively different hazard curves.

The exponential distribution, with a constant hazard function, did not fit the data well in Figure 3 (BIC = 29421.71). Likewise, the Weibull distribution has large deviation from the data at the tails of distribution ($t < 30$ and $140 < t$) (BIC = 26146.06). The single gamma distribution fits better than the exponential and Weibull distributions (BIC = 25722.64), but the mixture of two gamma distributions provides the best fit (BIC = 25655.29). In addition, mixtures of three gamma distributions (BIC = 25677.24) or more did not provide better fits than the two-component gamma distribution.

Figure 4 shows the probability density function of empirical reading intervals and the estimated probability density function, which is a mixture of two gamma distributions. One subcomponent, Distribution 1, has shape 13.80 and scale 4.24. The other subcomponent, Distribution 2, has shape 7.58

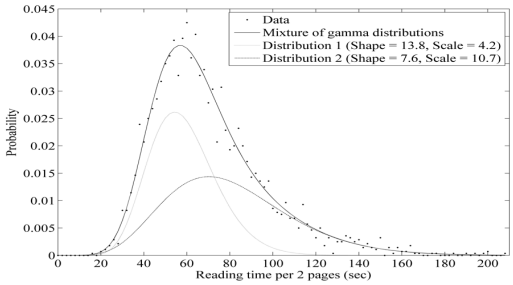


Figure 4: Sample (dots) and estimated (solid) probability distribution of reading time. The two curves under the fitting curve shows subcomponents of the gamma mixture distribution.

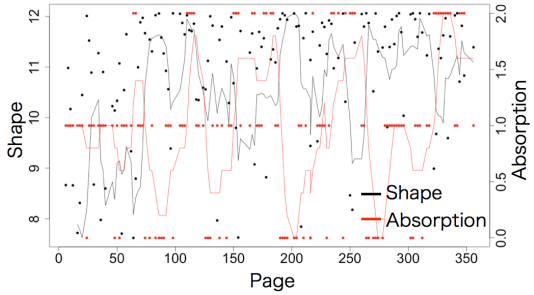


Figure 5: Page-based temporal profile of the statistical property (shape parameter) of reading time and the absorption ratings of the case No.17.

and scale 10.67. This result suggests that the subject shows of two distinct modes in her reading, with each mode involving different reading subprocesses.

Correlation to reading engagement We now address the question of whether the two distinct modes identified in our analysis are actually reflective of the text being read. In order to test this, we analyzed the correlation between the temporal change in mode and the degree of absorption reported by the reader. We obtained the reader's post-hoc report on engagement for each two pages of the books No. 17 and 18.

Taking the book No. 17 as a representative case, Figure 5 shows the temporal profile of the weighted-average of shape parameters (black dots) and the reader's degrees of absorption (red dots). The weights was given by the mixture of the two gamma distributions for each reading time of two pages. The corresponding moving average of the two over 5 data points are shown as black and red line, respectively.

We performed correlation analysis for a pair of the estimated shape parameters and the degrees of absorption. For the book No. 17 across 141 pairs of pages, we had correlation -0.284 ($p < 0.001$). For the book No. 18 across 118 pairs of pages, we had correlation -0.283 ($p < 0.01$).

In order to eliminate the possible effects of the number of letters, we additionally performed regression analysis on the book No. 17 with both the degrees of absorption and the number of letters as predictors. The results shown in the Table 2. The degree of absorption was significant variable as well as the number of letters to explain shape (both p <0.001). In sum, these analyses suggest that the temporal changes in the modes identified from our reading time analysis (Figure 5) does indeed reflect changes in reading engagement.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.318	0.803	19.07	<0.001**
Absorption	-0.515	0.134	-3.88	<0.001**
Letters	-0.004	0.001	-4.94	<0.001**

Table 2: Summary of multi-regression analysis with data of "Neko".

Remember that the shape parameter can be interpreted as the number of subprocesses involved in processing a given text, and the scale parameter can be interpreted simply as inverse of reading speed. Taking this theory into account, we conclude that the two modes estimated in this analysis are likely to represent a fast reading mode (Distribution 1) with a larger number of subprocesses and a slow reading mode (Distribution 2) with a smaller number of subprocesses.

Experiment 2

In Experiment 1, our statistical analysis detected two different modes of behavior in the reading data generated by the experiment. We further showed that the change in mode over time had a statistically significant correlation to the degree of absorption with the text reported by the subject. Our goal of Experiment 2 is to establish whether or not these findings are consistent across multiple subjects and, if so, to identify the factors involved with the reading modes detected in Experiment 1.

In order to meet these goals, we design a short experiment for multiple subjects. In our second experiment, we asked different subjects to read a short novel or a introductory part of long novel but kept the rest of the procedure the same as it was in No.17 and No.18 of Experiment 1. Specifically, subjects were asked to read these short text, and then they were asked to report their degrees of absorption for each two pages. The novel itself took less than an hour to read.

By analyzing their degrees of absorption, we evaluate the cross-experiment consistency, whether basic patterns in readers in the two experiments were consistent. We expect this analysis would reveal the significant factor of the reading modes, either/ both individual reading strategy or/ and the contextual structure of each novel. If the reading modes vary across readers but not within each reader, it would suggest that the reading modes depend on the readers' strategies. While, if the reading modes show similar time course in reading the same text across different readers, it would suggest that the reading modes depend on the context of each novel.

Participants

In Experiment 2, we employed ten participants, five male and five female undergraduate and graduate students at Keio University. Most of these subjects were not regular readers.

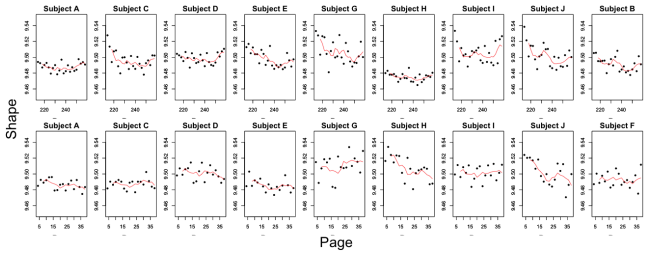


Figure 6: The page-based temporal profile of the estimated shape parameters for each subject. Top panels: "Kino", bottom ones: "Chinmoku".

Procedure

Each participant went through two reading sessions in random order. The procedure of each session was the same except for the text and the environment in which the reading took place. During each session, the participants read a 49-page-long short story "Kino" or 39-page-long introductory part of the story "Chinmoku Hakubutsukan" ("Chimoku" in short). Right after the reading session, the participant was asked to report their degrees of absorption in the same scale as Experiment 1 for each two pages. Five participants read the short story at first, the other five participants read the long introductory part at first. The participant took part again with more than two week interval, and read another text. The short story they read was "Kino" authored by Haruki Murakami. Another text is the introductory part of "Chinmoku" written by Yoko Ogawa. As the introductory partof the second book was chosen, it includes little change of context.

Analysis

For consistent comparison, we analyzed the aggregate of the reading time data across subjects by fitting to it a two-component mixture of gamma distributions. We fixed the class of distributions, instead of identifying it from data. This is due to the small sample size of our data at this point. Each participant provides reading time data for only 23 or 18 pairs of pages, and it did not sufficient statistical power to be conclusive even as its aggregation across subjects. Thus, we employed the statistical distribution estimated in Experiment 1.

Results

The data of two sessions were excluded from the analysis due to the the readers' irregular reading back and forth many times. This makes the completed dataset of 18 sessions in total.

Each panel of Figure 6 shows the page-based temporal profile of the modes estimated from reading time. In each panel, a dot shows estimated shape parameter for each reading time data point, and the line indicates its moving average.

The result shown in Figure 6 exhibit inter-subject consistency in temporal changes in reading modes. Each panel in Figure 6 shows the estimated shape parameter for each reader. The top panels show those of "Kino", and the bottom ones show those of "Chinmoku". We found the within-novel similarity in the shape parameter profile across pages: The readers of "Kino" showed similar U-shape profiles in common, and those of " Chinmoku Hakubutsukan" did similar flat profiles in common.

We performed correlation analysis on all the pairs of the subjects in order to test whether the readers of the same novel showed correlated temporal profiles of the shape parameters. The average correlation across all the reader pairs of "Kino" was 0.67 (from 0.46 to 0.87, $p < 0.02$ for every pair of readers), and that of "Chinmoku" was 0.51 (from 0.05 to 0.86, $p < 0.05$ for 23 out of 36 paris of readers). Thus, this result suggests that each novel showed the effects on the reading-mode profiles, with which the readers showed similar profiles, while individual reader showed little effects on them.

To analyze the factors on the reading times for each story, we performed a regression analysis on the shape parameters across readers and pair of pages with the degree of absorption and the number of letters as the predictors. The estimated regression coefficients and related statistics are summarized in Table 3. For the novel "Kino" on the top of Table 3, both absorption ratings ($p < 0.05$) and the number of letters ($p < 0.001$) were significant. For the novel "Chinmoku" on the bottom of Table 3, only the number of letters was significant ($p < 0.001$), but not the absorption ratings ($p = 0.404$).

Kino	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.430	0.012	773.26	<0.001**
Absorption	-0.0023	0.001	-2.04	0.042*
Letters	0.0001	<0.001	7.06	<0.001**
Chinmoku	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.4501	0.007	1327.48	<0.001**
Absorption	0.0008	0.001	0.84	0.404
Letters	0.0001	<0.001	6.40	<0.001**

Table 3: Summary of the regression analysis on the data of "Kino" and "Chinmoku". The single and double asterisk indicate $p < 0.05$ and $p < 0.01$, respectively.

Discussion

The results above suggests that there is the novel-specific effect on the statistical properties of reading time, which we can interpret mode switching profiles of reading process. As the page-based temporal profiles within the same novel were similar, it suggests that the contextual structure of each novel has major impact on switching of the reading modes. In addition, the temporal profiles could be interpreted as reflection of the semantic structure of each novel.

According to Miall's theory, a full novel is expected involved with both formation and exploitation of anticipation. Consistently with this theory, we found U-shaped patterns in reading mode switching for "Kino" with its full story, but we found flat patterns in for the introductory fraction of "Chinmoku", which indicated no clear distinct modes. This finding, no clear modes in reading an introductory part, is not well consistent with Miall's theory, but we interpret this that an introductory part of a full novel is likely to include only the formation of anticipation.

The regression analyses are consistent with this interpretation. The regression analysis of the novel "Kino", but not of "Chinmoku", showed a similar trend with that of the novel No. 17 analyzed in Experiment 1– for both cases, the absorption rating was significant predictor of the shape parameter. What was common between the novel "Kino" and No. 17 in Experiment 1, and different from "Chinmoku"? Two novels, which both showed the significant effect of absorption, were used as their full story, but the one, which did not show it, was used as its introductory part. Thus, one possible interpretation is that the full stories, which are supposed to have contextual coherence from the introduction to the end, may be related to temporal changes of the reading modes and the degree of absorption.

In sum, these findings could be treated as a supportive evidence that the temporal profiles in reading distributions relfect context-dependent reading modes.

General Discussion

Reading is an essentially mental and subjective experience. Its cognitive underpinnings have been difficult to characterize directly, and reading time is a major tool for drawing inferences about the underlying cognitive mechanism behind reading. This study offers a new approach to the analysis of reading time, an approach capable of identifying different modes of reading behavior from reading time data.

In Experiment 1, we collected an analyzed reading time data generated by a single subject reading several full novels in a natural situation. We observed significant correlation between subject's report of her engagement in reading and her reading modes inferred from the estimated reading time distribution. This experiment has three major implications:

1. In contrast to conventional studies on controlled, short readings, it is perhaps the first study on involving reading entire books in a more natural situation.
2. It establishes a new analytical technique for reading time data by associating the estimated modes with the subject's engagement in reading.
3. It provides a support evidence that there are at least two distinctive reading modes depended on the contexts in reading of whole novels.

In Experiment 2, we once again observed two distinct reading modes, and found that the mode switches across different subjects reading the same story were consistent with each other. This suggests that, to a large extent, the reading modes are dictated by the contextual structure of the text being read. This interpretation is also supported by the the regression analyses.

What is the contextual structure? We hypothesize that it is related to the predictability of the story. Perhaps, each of which we found as two reading modes can be called low- and high- predictability mode. Considering with reading of "Kino", the major shifts between two modes took place at the beginning and end of the story. At the beginning, any reader has little knowledge on the story, as also discussed in (Miall, 1988), and they need to build knowledge on the characters and stage where they play their roles. At the part approaching to the end, this story has a "twist", which is unexpected for most of readers. This is another place, in which the reader need to rebuild their knowledge on the story. Therefore, with these two parts at the beginning and end, well-predictable part in between them is considered to make the U-shaped temporal profiled in the reading time distributions.

Acknowledgements

The authors are grateful to Dr. Neeraj Kashyap for his proofreading of this manuscript. This study was supported by Keio University Doctorate Student Grant-in-Aid Program, Mori Grants, and Grant-in-Aid for Scientific Research B KAKENHI No. 23300099.

References

Beaugrande, R. D. (1982). the story of grammars and the grammar of stories. *Jornal of Pragmatics*, *6*, 383–422.

Gernsbacher, M. A. (1997). Two Decades of Structure Building. *Discourse processes*, *23*(3), 265–304.

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PLOS ONE*, *8*(10), e76242.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & psychophysics*, *40*(6), 431–439.

Iser, W. (1976). *Der Akt des Lesens: Theorie ästhetischer Wirkung*. München: Wilhelm Fink.

Kassel, R. (1965). *Aristotelis de arte poetica liber*. Oxford University Press.

Miall, D. S. (1988). Affect and Narrative: A Model of Response to Stories. *Poetics*, *17*, 259–272.

Miall, D. S. (1989). Beyond the schema given: Affective comprehension of literary narratives. *Cognition and Emotion*, *3*(1), 55–78.

Miall, D. S., & Kuiken, D. (1994). Beyond Text Theory: Understanding Literary Response. *Discourse Processes*, *17*(3), 337–352.

Rapp, D. N., & Gerrig, R. J. (2006). Predilections for narrative outcomes: The impact of story contexts and reader preferences. *Journal of Memory and Language*, *54*(1), 54–67.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, *9*(1), 77–110.

White, S. J. (2008). Eye movement control during reading: Effects of word frequency and orthographic familiarity. *Journal of experimental psychology: Human perception and performance*, *34*(1), 205–223.

Effects of Working Memory Training on L2 Proficiency and Working Memory Capacity

Gregory Colflesh (gcolflesh@casl.umd.edu)
Valerie Karuzis (vkaruzis@casl.umd.edu)
Polly O'Rourke (porourke@casl.umd.edu)
Center for Advanced Study of Language
University of Maryland, 7005 52nd Avenue
College Park, MD 20742 USA

Abstract

The current study examined the effects of working memory training on working memory capacity and second language ability in adult learners of Spanish. In order to maximize the effect of the training for language learners, the stimuli for the training tasks were Spanish words and sentences. While the training group did not show greater improvements on working memory assessments relative to controls, they did show more native-like patterns in a Spanish self-paced reading task. The combination of second language materials with working memory training may be helping users learn to cope with the increased processing demands associated with learning a new language, even if they are not necessarily improving their working memory.

Keywords: Working memory training, second language acquisition, self-paced reading, n-back.

Introduction

Working memory (WM) plays a role in many higher order cognitive functions, including the ability to learn a second language (L2). Indeed, one prominent theory of the acquisition of L2 morphosyntax proposes that the deficits exhibited by learners are largely due to WM limitations, given the high demand of operating in L2 (McDonald, 2006). Due to the importance of WM in a great many everyday activities, exploring the efficacy of WM training has been a very hot research topic in the last five to ten years. In the current study we examined the impact of WM training on second language processing ability.

Given the connection between working memory capacity (WMC) and second language acquisition (Linck, Osthus, Koeth & Bunting, 2014), the possibility of increasing WMC, and thereby enhancing second language learning outcomes, is an exciting prospect. Evidence from experimental research suggests that it may be possible to increase WMC with computer training in which individuals repeatedly perform tasks that tax attentional and executive control abilities (Baniqued et al., 2014; Holmes, Gathercole & Dunning, 2009; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Morrison & Chein, 2011; Novick et al., 2014; Olesen, Westerberg & Klingberg, 2004). Training typically targets domain-general mechanisms which "control attention, gate the flow of information into and out of WM buffers, reduce interference from irrelevant sources of information, and govern the engagement of domain-specific strategies." (Morrison & Chein, 2011, p. 47). These domain-general mechanisms drive the connection between WM and higher order cognition (Cowan et al., 2005; Lépine, Barrouillet, & Camos, 2005; Morrison & Chein, 2011). Training-related improvements in domain-general executive functions have been shown to be associated with increased general fluid intelligence (Jaeggi, et al., 2008) and language processing ability in the first language (Morrison & Chein, 2011; Novick, et al., 2014). Given the strong connection between WMC and second language learning, improving domain-general attentional and executive control abilities through training could also lead to improvements in second language proficiency. In the current study, we examined the impact of WM training on online syntactic processing and sentence comprehension in adult learners of Spanish. Additionally, in order to maximize the benefits of the WM training, the training tasks all contained Spanish language stimuli, thereby providing participants with increased exposure to the target language during training.

In the current study, the WM training consisted of adaptive versions of four standard WM tasks (Reading Span, *n*-back, Sorter, and Running Span). This multifaceted "kitchen sink" approach was implemented as a means of increasing the likelihood of achieving training-related gains by engaging similar WM mechanisms in a variety of different tasks (Morrison & Chein, 2011). In addition, the inclusion of multiple tasks was intended to increase participants' likelihood of enjoying and completing the training, as performing four tasks for 15 minutes each is more engaging than performing the same (difficult) task for 60 minutes. This is in contrast to studies that used a more targeted approach (Jaeggi, et al., 2008; Verhaeghen, Cerella, & Basak, 2004).

The key innovation in our study, which is one of the first to examine the impact of WMT on L2 proficiency, is that the stimuli in our WM training tasks consist of Spanish words and sentences. The motivation for this innovation is to (1) increase individuals' engagement in the training tasks by making the stimuli relevant to their real world learning goals and (2) to enhance the benefit of training on language learning by requiring participants to interact with L2 materials while performing WM demanding tasks.

We predicted that participants who completed the adaptive training tasks with a WM demand (i.e., WM training) would have greater improvement from pre-training to post-training for Spanish language and WM assessments compared to participants who completed the adaptive training tasks with minimal WM demands (i.e., active control).

Method

Participants

62 adult learners of Spanish (51 female, age = 20.26 (*SD* = 2.11), years education = 14.31 (*SD* = 1.27)) completed this study. Eligible participants were native English speakers who did not learn any non-English languages at home or during childhood, were currently or recently enrolled in a 200- or 300-level university Spanish class or similar experience, and passed the Spanish proficiency requirements. All participants were neurologically normal, began studying Spanish after age 11, and had not had an immersion experience in a Spanish-speaking country for longer than 6 weeks. Participants were semi-randomly assigned to the training and control groups; they were matched on working memory capacity (as indexed by the Shapebuilder task; see below) and comprehension accuracy of grammatical sentences in the self-paced reading task (see below). Post-hoc analyses confirmed that groups were similar for age, gender, education, self-rated Spanish ability (reading, writing, listening and speaking), and two other measures of Spanish vocabulary knowledge.

Pre-Post Assessment Tasks

Self-Paced Reading Task

In the self-paced reading task, participants read sentences, in Spanish, word-by-word, advancing at their own speed. Words appeared at the subject's prompt within a sentence frame with "-"s as place holders for each letter. Each word reverted to "-"s when the subject advanced to the next word. The sentence stimuli consisted of grammatical sentences and sentences containing subject/verb agreement errors (see sentence 1 below). There were 12 items in each condition. In addition to those 24 sentences, there were 67 additional sentences including fillers and conditions not reported herein. Half of the sentences were followed by a yes/no comprehension question. The timing of the task was based on Sagarra and Herschensohn (2010)'s methods.

1. La investigadora **cree**/*<u>**creen**</u> que la ciencia concreta es más importante que la teoría.
   The researcher [**believes**/**believe**] that concrete science is more important than theory.

Working Memory and Executive Function Assessments

N-back

For the *n*-back task used in the current study (Kane, Conway, Colflesh, & Miura, 2007), participants performed blocks of 2-back and 3-back, with lures (±1 and ±2, where appropriate). The *n*-back task indexes the ability to maintain, update and monitor information in the short-term store and resolve the conflict between what is familiar and recollected. While it is a frequently used index of WM ability, its cognitive underpinnings are distinct from those of complex span tasks (Kane et al., 2007).

Reading Span

The automated version of the reading span task (Redick, et al., 2012), a complex span task, was used to measure WMC. Complex span tasks measure the ability to store and recall information in the face of distracting information from an unrelated processing task.

Shapebuilder

For Shapebuilder (Atkins, et al., 2014) participants had to remember the shape and color of items that appeared sequentially on a 4x4 grid. Shapebuilder was used as an additional measure of WMC that did not mimic the reading span, which was also used as a training task.

Simon Task

Participants performed the Simon Task (Simon, 1990) as an index of inhibition control. The Simon Task, while not a WM assessment, was included because it indexes the ability to control attention via the inhibition of a prepotent, but goal-irrelevant response, which underpins the performance of many of the WM tasks.

Training Tasks

Spanish Reading Span

As the name suggests, the Spanish reading span training and control tasks were modeled off the standard reading span task (Daneman & Carpenter, 1980; Kane et al. (2004)). Participants were asked to recall a sequence of letters which appeared one-at-a-time on the screen. In between the presentation of each letter, they saw a sentence in Spanish (e.g., *El médico llegó para ayudar a la gente herido.)* and they had to decide if it was grammatical or not. Participants had to decide among five possible choices, one of which is "No Errors" and the remaining four were possible error types: spelling, accent use, subject/verb agreement, and gender agreement (gender error present in the adjective portion of the example above).

In the training version of the task, participants were presented with a sentence. They had 10 seconds to make a judgment on the sentence. After they made a judgment on the sentence (or did not make a response in the allotted time), they received feedback regarding their selection. Next, a letter was presented. The number of sentence-letter items in a trial started at 1 and increased as participants performed at or above leveling criteria. Leveling criteria was based on sentence and memory performance across four trials. Initially, the four trials were all set size 1. At the next level, three trials were set size 1 and one trial was set size 2. To level up, participants had to have 85% or greater accuracy on the sentences and had to have recalled 85% or greater of the letters presented for a later recall test in the correct serial positions. Participants remained at the same

level if they were between 70% and 85% on both, and they went down a level if they were below 70% on either component.

In the control version of the task, sentences and memoranda were presented separately within a trial, instead of being interleaved. All sentences presented for the grammatical judgment task were in an uninterrupted sequence. Memoranda letters were presented for the recall task in randomly selected sequences of 2-6 letters, independent of the sentence processing task. Participants went up a level if they achieved 85% accuracy on the sentences. As participants leveled up, there were more sentences to be judged in each block.

Spanish *n*-back

The Spanish *n*-back training and control tasks were modeled off of previous *n*-back tasks (Kane et al., 2007; Novick, et al., 2014). In the Spanish *n*-back task, participants saw a series of 20 + *n* words presented on the screen one at a time, with each presentation accompanied by an audio presentation of the word. The stimuli were presented in a mix of Spanish and English, with approximately 80% of the words in a list presented in Spanish. Participants responded to each word as it was presented. Participants pressed the "Yes" button if the current word matched the word *n* items back, either exactly (*manzana – manzana*; *apple – apple*), or by meaning (*manzana – apple*). If the current word did not match the word or translation of the word *n* items back, participants pressed the "No" button.

In the training version of the task, *n* increased as participants' level increased throughout the task. To go from 1-back to 2-back, participants had to meet leveling criteria for two steps: 1) *n* without lures, and 2) *n* ± 2 lures. Starting with 2-back, each level required three successful steps to achieve the next level: 1) *n* without lures, 2) *n* ± 2 lures, and 3) *n* ± 1 lures.

In the control version of the task, participants performed a 1-back task with the Spanish (and English) word stimuli. Participants were given the appearance that they were attaining different levels if they reached leveling up or down criteria, but the task always remained a 1-back task.

Spanish Running Span

The Spanish running span training and control tasks were modeled off of the running span task (Anderson, 1960; Bunting, Cowan & Saults, 2006; Pollack, Johnson, & Knaff, 1959). In the Spanish running span task, participants saw a series of 12 to 20 + K words, where K is the number of items that needed to be recalled, presented on the screen one at a time, accompanied by a simultaneous audio presentation of the word. The stream of stimuli was presented either entirely in Spanish or entirely in English. At the end of the presentation, participants needed to respond to what the last K words were. K started at 1 item and increased as participants leveled up. If the word list was presented in Spanish, participants saw a list of English words – targets and lures. If the word list was presented in English, participants saw a list of Spanish words – targets and lures.

In the training version of the task, participants had to drag the correct words and drop them into the correct positions. In the control version of the task, participants were given the translations of the last K words and only needed to drag the correct words to the appropriate boxes.

Spanish Sorter

The Spanish sorter training and control tasks were modeled off of the letter-number sequencing task used by Sprenger, et al. (2013). In the Spanish sorter task, participants saw a series of 2 words presented on the screen one-at–a-time, accompanied by a simultaneous audio presentation of the word. At the end of the presentation, participants had to drag the words that had just been presented to the correct semantic category, and the words within each category needed to be recalled in serial order and the correct serial position for that category. As participants leveled up, they were presented with more words to recall and sort.

The Spanish sorter control task was the same as the training task, except participants were provided with the translation equivalent of the words in the categories and they needed to drag the correct words to the correct translations. As participants leveled up, they were presented with more words.

Procedure

The study was comprised of 14 sessions. Sessions 1, 2, 13, and 14 were pre- and post-training sessions (1-2 hours long each), and sessions 3-12 were training sessions (45 minutes-1 hour long each). The pre-post sessions included the tasks described above and additional measures not reported herein. Participants completed pre- and post-training tasks on PCs, and training tasks on iPads. All participants completed the training sessions within 10-14 business days, completed session 13 1-5 days after training, and completed session 14 1-2 days after session 13. Multiple training sessions were not allowed to take place within the same day. Participants were compensated $5/session, with a balloon payment of $175 at the completion of session 14. Participants were also entered into raffles for 1 of 10 $200 gift cards following study completion. In addition to the participants analyzed here, 5 participants attritted from the study (7%), and 4 were found to be ineligible after session 1 (5.6%).

Results

WM Assessments

Data points more than 2.5 SD from the sample mean were excluded from analyses. Participants in the training group did not show greater improvements at post-test than the control group on the reading span or Shapebuilder tasks, indicating that WM training did not improve WMC. Similarly, there were no effects of training condition on the Simon task.

The 3-back trials in the *n*-back task were analyzed using overall d', a measure of signal detection. A 2 (training

condition) x 2 (assessment time) mixed factor ANOVA analysis of d' for the 3-back trials showed no significant effect of assessment time (F<1). There was, however, a significant effect of training condition (F(1, 51)=5.41, p<.05, η_p^2=.10), where the training group performed better than the control group. Simple comparisons at pre- and post-test showed no significant difference between training group and control group performance at pre-test (Ms = 1.97 & 1.28, respectively; t(57)=.44, p>.05) but at post-test, the training group performed significantly better than the control group, (Ms = 3.09 & 1.97, respectively; t(54)=2.11, p<.05). This suggests that WM training may have improved cognitive control ability, especially in the face of interference.

Self-Paced Reading Task

Reading times exceeding 3 SD from a given subject's mean were excluded from analyses, and any subject with a mean reading time over 800 ms per word was excluded from the analysis. This resulted in an N of 46, with 23 participants per training group. Separate 2 (assessment time) x 2 (training condition) x 2 (grammaticality) ANOVAs were conducted on the reading times at each of the four positions relative to the critical word: -1, 0 (critical word), +1, +2. Figure 1 shows the reading time data for the two groups at post.

At position -1, 0, +1 and +2 there were main effects of assessment time (position -1: F(1,44)=51.1, p<.001, η_p^2=.54; position 0: F(1,44)=54.2, p<.001, η_p^2=.54; position +1: F(1,44)=32.6, p<.001, η_p^2=.43; position +2: F(1,44)=28.7, p<.001, η_p^2=.40, respectively) such that reading times were faster at post. At position +1, there was also a three-way interaction of grammaticality, assessment time, and training condition (F(1,44)=4.89, p<.05, η_p^2=.10). Splitting across training condition revealed a main effect of assessment time (F(1,22)=8.83, p<.01, η_p^2=.29) and no effects of grammaticality in the control group. In the training group, there was an effect of assessment time (F(1,22)=27.4, p<.001, η_p^2=.56) and an interaction of assessment time and grammaticality (F(1,22)=6.60, p<.05, η_p^2=.23). This interaction was driven by increased reading times in the ungrammatical condition at post, evidenced by an effect of grammaticality at post (F(1,22)=5.38, p<.05, η_p^2=.20). There were no effects of grammaticality in the training group at pre. At position +2, there was a significant interaction of assessment time and grammaticality, driven by an effect of grammaticality at post (F(1,44)=9.16, p<.01, η_p^2=.17). While there were no interactions involving training condition at position +2, Figure 1 suggests the grammaticality effect was larger for the control group. Simple comparisons showed a significant effect of grammaticality in the control group (F(1,22)=9.11, p<.01, η_p^2=.29) but not in the training group.

Similar ANOVAs performed on accuracy data for the comprehension questions (Table 1) revealed no effects of assessment time, training condition or grammaticality.

Table 1: Mean Comprehension Accuracy (SD)

	Pre	Post
Control - Grammatical	0.78 (.19)	0.78 (.15)
Training - Grammatical	0.81 (.15)	0.79 (.15)
Control - Violation	0.80 (.12)	0.74 (.16)
Training - Violation	0.80 (.16)	0.76 (.14)

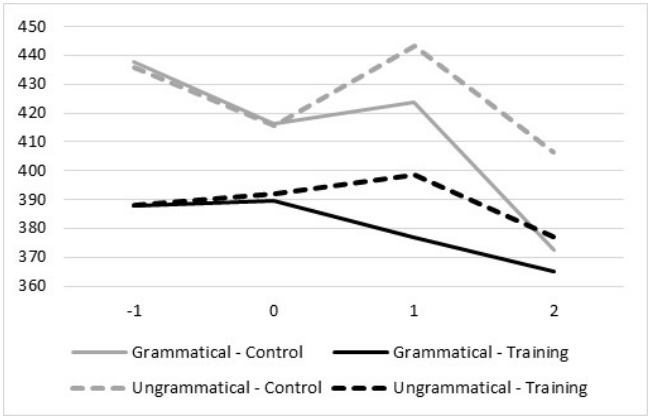


Figure 1. Mean reading times (ms) at post.

Discussion

Although many studies have examined the relationship between WMC and L2 acquisition, and many other studies have examined the impact of WM training on WM and general cognition, there has not been much research combining WM training with L2 acquisition. In this study, we explored whether completing WM training in a target language, Spanish in this case, would (1) improve performance on WM measures and (2) improve performance on an online processing task in the target language.

The WM training did have the predicted impact on *n*-back scores, such that the training group performed significantly better than the control group. However, the interaction between training condition and assessment time was not significant, despite a moderate effect size. Thus, further exploration may be warranted. It may be that the L2 WM training did improve some aspect of the WM system; however, it might be that participants in the training group developed better strategies to cope with the increased WM demand of the L2 *n*-back training task.

WM training did not have the predicted impact on the remainder of the battery of WM assessments, which included Reading Span, Shapebuilder, and the Simon task. Previous research found that complex span tasks and *n*-back measure different aspects of the WM system (Kane et al., 2007), so it is not surprising that *n*-back did not follow the same pattern as reading span. The current findings further support the distinction between *n*-back and complex span tasks.

It is important to note that prior research using assessment versions of complex span tasks in the first (L1)

and second language have shown strong correlations between L1 and L2 span tasks (Alptekin & Erçetin, 2010; Osaka & Osaka, 1992; Osaka, Osaka & Groner, 1993). This indicates a strong overlap between L1 and L2 versions of the tasks suggesting that the same resources are utilized. The lack of transfer between the L2 reading span and standard version in the current study, therefore, should not be an artifact of task language.

The WM training outcomes in the current study mimic what is generally found in the literature. The studies that have found improvements in WM and a transfer to other abilities tend to utilize the *n*-back task, or a variant of the *n*-back task (Jaeggi, et al., 2008; Jaeggi, Studer-Luethi, Buschkuehl, Su, Jonides, & Perrig, 2010). However, the methodology and outcome of these studies have been greatly scrutinized (Redick, et al., 2013; Shipstead, Redick, & Engle, 2012), specifically the lack of a contact control group. Redick and colleagues (2013) compared dual *n*-back WM training to that of a contact control group, and they did not find that WM training improved WM. It may be that the *n*-back training task allows for the development of strategies to circumvent the WM demands of the *n*-back task and those strategies may transfer to the assessment version of the *n*-back, but not the other assessments. If this is the case, then WM training is not actually leading to improvements in WM, but rather strategies to reduce the reliance on the WM, suggesting that WM is a stable trait.

However, even if WM training does not train working memory, it does not mean that the training is not useful. Perhaps instead of calling it WM training, it should be called cognitive training, as it appears training may improve some cognitive processes, which in turn may improve additional cognitive abilities, such as second language processing.

One interesting effect emerged out of the SPR reading time data: the training group showed an earlier sensitivity (at position +1) to subject-verb agreement violations, while the control group did not show an effect until the last position (+2). In this respect, the training group looked slightly more native-like, as native speakers (and higher proficiency learners) would show effects at the 0 and +1 position (Sagarra & Hernschensohn, 2010). The training, therefore, seemed to contribute toward enhanced online syntactic processing ability. The decomposition of inflected forms (i.e., utilization of morphological information) during online processing is perhaps one of the greatest obstacles for L2 learners. These difficulties are possibly due to reduced ability to engage in combinatorial rule application in L2 (Silva & Clahsen, 2008) or an inability to process the information due to the increased WM demands associated with operating in the L2 (McDonald, 2006). In McDonald (2006)'s account, increasing WMC via training should result in increased sensitivity to L2 morphosyntactic information. The training related changes in L2 morphosyntactic processing that we observed did not correspond to clear pre-post improvements in WMC, thus the findings from the training manipulation do not provide direct support McDonald (2006)'s account. Rather, linguistic exposure and interaction in a speeded and high WM demand context associated with the training may have accelerated the development of participants' morphosyntactic processing ability. Alternatively, the performance of the WM training may have led to improvement in underlying cognitive abilities other than WM that are relevant to language processing. A final possibility is that the training tasks were simply more engaging and, thus, participants were more motivated with respect to processing the Spanish training task stimuli.

In conclusion, working memory training in the second language did not transfer to performance on all of the WM tasks, yet it did lead to enhanced online morphosyntactic processing of the target language. This suggests that although WM training may not serve the laudable purpose of enhancing WMC, it may be useful in improving second language ability and/or other general cognitive abilities which underpin second language processing.

References

Anderson, N. S. (1960). Poststimulus cuing in immediate memory. *Journal of Experimental Psychology*, *60*(4), 216.

Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading*, *33*(2), 206-219.

Atkins, S. M., Sprenger, A. M., Colflesh, G. J. H., Briner, T. L., Buchanan, J. B., Chavis, S. E., . . . & Dougherty, M. R. (2014). Measuring Working Memory Is All Fun and Games. *Experimental Psychology, 61*(6), 417-438.

Baniqued, P. L., Kranz, M. B., Voss, M. W., Lee, H., Cosman, J. D., Severson, J., & Kramer, A. F. (2013). Cognitive training with casual video games: points to consider. *Frontiers in Psychology*, *4*(1010), 1-19.

Bunting, M., Cowan, N., & Scott Saults, J. (2006). How does running memory span work?. *The Quarterly Journal of Experimental Psychology*, *59*(10), 1691-1700.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42-100.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450-466.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science, 12*(4), F9-F15.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829-6833.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—

Implications for training and transfer. *Intelligence*, *38*, 625–635.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615-622.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Lépine, R., Barrouillet, P., & Camos, V. (2005). What makes working memory spans so predictive of high-level cognition? *Psychonomic Bulletin & Review*, *12*(1), 165–170.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861-883.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381-401.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46-60.

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language, Cognition and Neuroscience*, *29*(2), 186-217.

Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*, *7*(1), 75–79.

Osaka, M. & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society*, *30*, 287–289.

Osaka, M., Osaka, N. & Groner, R. (1993). Language-independent working memory: Evidence from German and French reading span tests. *Bulletin of the Psychonomic Society*, *31*, 117–118.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology*, *57*(3), 137-146.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164-171.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., . . . Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*, 359–379.

Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua, 120*(8), 2022-2039.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*, 628-654.

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, *11*(02), 245-260.

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R. W. Proctor & T. G. Reeve (Eds.). Stimulus-response compatibility: An integrated perspective (pp. 31–88). Amsterdam, The Netherlands: North Holland.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., . . . Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, *41*, 638-663.

Verhaeghen, P., Cerella, J., & Basak, C. (2004). A working memory workout: how to expand the focus of serial attention from one to four items in 10 hours or less. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1322-1337.

Acknowledgements

This research is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this article is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the author of the product, process, or service that is the subject of this article. Correspondence concerning this article should be addressed to Gregory Colflesh, Center for Advanced Study of Language, University of Maryland, 7005 52nd Avenue, College Park, MD 20742 (e-mail: gcolflesh@casl.umd.edu). Special thanks to Mike Bunting for supporting this line of research, and to Ashlyn Testut, Grace Goff, Kenneth Mai, and Mari Sakai for their help in setting up the lab for testing, running participants, and cleaning data.

Representation: Problems and Solutions

Nancy Salay (salay@queensu.ca)
Department of Philosophy, Watson Hall
Kingston, ON CANADA

Abstract

The current orthodoxy in cognitive science, what I describe as a commitment to deep representationalism, faces intractable problems. If we take these objections seriously, and I will argue that we should, there are two possible responses: 1. We are mistaken that representation is the locus of our cognitive capacities — we manage to be the successful cognitive agents in some other, non-representational, way; or, 2. Our representational capacities *do* give us critical cognitive advantages, but they are not fundamental to us *qua* human beings. As Andy Clark has convincingly argued, anti-representationalism, option one, is explanatorily weak. Consequently, I will argue, we need to take the second option seriously. In the first half of the paper I rehearse the problems with the current representational view and in the second half of the paper I defend and give a positive sketch of a two-systems view of cognition – a non-representational perceptual system coupled with a representational language-dependent one – and look at some consequences of the view.

Keywords: representation; representation-hungry problem; consciousness; animal cognition; perception; two-systems

Introduction

The current orthodoxy in cognitive science is that human beings are fundamentally intentional beings, that what makes us uniquely cognitive agents is our ability to think (and talk) about our world. It is precisely this capacity, we generally think, that is at the root of our cognitive superiority to even our closest genetic cousins. But there are deep problems with this view. If we take these objections seriously, and I will argue that we should, there are two possible responses: 1. We are mistaken that representation is the locus of our cognitive capacities — we manage to be the successful cognitive agents in some other, non-representational, way; or, 2. Our representational capacities *do* give us critical cognitive advantages, but they are not fundamental to us *qua* human beings.

The first response should be fairly familiar by now to anyone in cognitive science who has engaged with the arguments from the embodied/embedded camp influenced by the phenomenological work of people like Martin Heidegger and Maurice Merleau Ponty. There is, surely, something deeply right about these views. As Andy Clark has convincingly argued, however, anti-representational positions are explanatorily weak. Consequently, I will argue, we need to take the second option seriously.

Once all the dust has settled, we will see that both camps are partly right and partly wrong: our most fundamental relation with the world *is* an experiential, not a representational one; but, our paradigmatically cognitive capacities, the abilities that really set us apart from other animals, are our representational ones.

In the course of defending this view, I will argue that we cannot draw inferences about our cognitive hardware from the seemingly representational capacities of our minds, that there are more important lessons to be gleaned from extended mind debates besides metaphysical or conceptual ones about the boundaries of minds, and that clarifying matters and, consequently, carving a path for new, fruitful research, will require a narrowing of our understanding of *representation*, a narrowing that runs counter to recent suggestions from the embodied/embedded approach, with which I have deep sympathies.

In the first half of the paper, I rehearse the problems with the current representational view; in the second, I defend and give a positive sketch of a two-systems view of cognition – a non-representational perceptual system coupled with a representational language-dependent one – and look at some consequences of the view.

Deep Representationalism

The Representational Theory of the Mind (RTM) is the view that a large range of human behaviour is best explained by appeal to reasons rather than, for example, by descriptions of neural activity. Reasons, on this view, come in the form of propositional attitudes, beliefs, desires, fears, and the like, to mental representations. Why did the chicken cross the road? Because it wanted to get to the other side.

RTMs typically take it that *all* mental states are representational: abstract ideas such as *justice*, occurrent thoughts such as *the cat is on the mat*, and perceptions such as *my seeing this flower now*, all count as mental representations. Views might differ about how much conceptual content a given mental representation has, about the way in which perceptions get their representational content, but, the sphere of the mental is taken to be representational through and through. I will call this class of views, Deep Representationalism, DR. Despite much criticism, DR is still the prevailing view in cognitive science today.

One of the challenges facing DR is to produce a naturalistic account of the representation relation, of grounding it in some relation that can be specified in non-intentional terms. This has proved a difficult nut to crack. As C.S. Peirce pointed out in his comprehensive analysis of representation (Peirce, 1998), for something to be a representation it must stand in one sort of relation to the thing it represents and another sort of relation to the interpretation of the signifying relation. Smoke does not mean fire if there is no one around to recognise the relationship between smoke and fire. Giving a

naturalistic account would involve explaining this interpretant relation without appeal to intentionality at all.

Fred Dretske's information theoretic account (1988) is arguably the most comprehensive such attempt we have. On his account, Peirce's triadic relation is compressed into a dyadic one: a state R of a system S is a representation of kind O just in case R nomically covaries with the presence of O's and this co-variance relation explains R's functional role within S. On this view, the bi-metallic strip inside a thermostat *represents* the ambient temperature of the surrounding air because it is by virtue of the strip's physical properties, which nomically covary with the rise or fall in temperature, that it has its functional role, namely, to turn the furnace on when the temperature drops below a certain level and off when it rises above it. In other words, it is in virtue of its *information carrying properties* that it has its functional role. At least this is what Dretske wants to conclude.

As Ramsey (2003) convincingly argues, however, this account is not sufficient to justify our viewing such states as representations. Dretske's story would be a naturalistic account of representation only if R's functional role really were a consequence of the information it carried via the nomic co-variance relation actually used by the thermostat. But in none of the examples that Dretske gives to support his account does he successfully show this. The nomic co-variance relation *is* playing a functional role in these systems, but not *qua* information-carrying relations:

> The functionality of the strip is to cause something to happen in very specific conditions. But in this respect, it is no different than many other devices that we ordinarily treat as having no representational function. The firing pin in a gun similarly bridges a causal gap between the pulling of the trigger and the discharge of the round. … However, no one thinks the firing pin functions as some sort of representational device. (Ramsey, 2003)

More is needed to justify the judgement that one such relation is a consequence of the information it carries, while the other is merely a consequence of the causal conditions that obtain. The mistake, I think, is in supposing there is a difference between the two cases *at all*[1]. A central problem for DR views, then, is that thus far there has been no success in providing a naturalistic account of the representation relation. I take this to be a serious mark against such views.

There are positive reasons for being skeptical about DR accounts as well. DR is not entailed by RTM; one can consistently acknowledge that 1. RTM is true — a large range of human behaviour *is* best explained by appeal to mental representations, that 2. Not all human behaviour is best explained by appeal to mental representations, and, that 3. RTM is not reducible to a representational theory of the brain. In other words, one could hold a version of RTM while denying DR.

Or not. Eliminativist Materialism, EM, is a view that rejects RTM altogether, although many of the arguments made from that quarter are aimed at the reality and/or usefulness of higher-level mental representations such as beliefs, desires, and the like. More relevant to the current discussion are the arguments Daniel Dennett offers to fuel his more radical, low-level, eliminativism of representational perception.

Dennett (1991) points out that mainstream accounts of perception illusions, in which perception reports are temporally incongruous with the actual presentation of objects being perceived, all assume a Cartesian theatre in which the sub-processes of the brain come together and either rewrite the past, by altering our memories, or subpersonally reconcile contradictory experiences into one final account before they appear in consciousness. He argues that, since there is no evidence to support either account over the other, we should reject the central workspace hypothesis as explanatorily defunct. Dennett concludes that the real illusion is the Cartesian theatre itself; there are multiple processes ongoing at any given moment and those that win out or 'rise to the top' result in behaviour. Though we might balk at this radical elimination of conscious experience, Dennett's powerful arguments against the necessity of a representational view of perception cannot be ignored.

Finally, many from the embodied/embedded camp in cognitive science have been developing accounts of cognition that bottom out in completely non-representational relations with the world. Antony Chemero (2009), for example, argues that looking for and modelling the ways in which the environment constrains and directs cognitive behaviour is the best approach to developing a comprehensive account of cognition. Detailed dynamical systems analyses are on the rise, showing us that, contra RTM, many factors besides mental representations have behavioural explanatory force. (Noë,2010; Clark & Toribio, 1994; Beer, 2003; Haken, Kelso, Bunz, 1985)

These anti-representational approaches to cognition find philosophical support in the work of Merleau-Ponty (1945) and Heidegger (1927) both of whom argued that our most fundamental relation with the world is experiential, not representational. More recently, picking up these threads, Hubert Dreyfus (2007) contends that the intractability of a naturalistic account of intentionality should be a clue that any view founded on DR is doomed.

Finally, neuroscientists such as Walter Freeman (2000) argue that even were we to ignore the theoretical stumbling blocks that plague DR accounts, neural activity lacks a fundamental requirement of any representational vehicle: that it serve as a constant, consistent, relation to what is represented. Neural activity is so dynamic, he argues, that there is nothing at the neural level stable enough to play this role.

[1] Practical realists such as Lynne Rudder Baker (2001) take this approach as well.

RTM is Partly Right

Clark, who takes these embodied, embedded objections seriously, and, who agrees with the anti-representational view that RTM is not as explanatorily useful as proponents suppose, nevertheless argues that doing away with representational explanations entirely won't do either. (Clark 1997, 1994). Some problem situations are "representation-hungry:" to solve them agents require information that is not directly available from their immediate environment. Remembering one's credit card pin when paying for groceries, mentally running through possible future contingencies when making plans, feeling regret about past actions, are all activities that require, ostensibly, a capacity for mentally representing a situation to oneself. The anti-representationalist can't explain our ability to successfully solve these representation-hungry problems because those accounts explain all of our behaviour in terms of our direct, experiential relation with our environments. What happens when the relevant environment is not there to be experienced?

Motivated by these considerations, Clark (1997) and Wheeler (2007) have developed a hybrid view in which perception is a kind of action/representation mix, what they have termed action-oriented representation (AOR): sensory-motor loops that develop in certain sorts of situations over time as a result of experience. The fact that these are idiosyncratic to an agent, to its particular sensory capacities, and to its environment is meant to address many of the challenges that traditional, objectivist, representational views face. In particular, the problem of grounding these representations in some direct experience with the environment is met, on these accounts, via these base level, action-oriented, sensory-motor loops.

But, as Dreyfus has recently pointed out (2007), so long as perception is assumed to be a representation relation, whether objective or action-oriented, the problem of intentionality does not go away.

Language and Perception

The perceptual relation, then, must be non-representational. At the very least, the objections to DR militate against an out-of-hand dismissal of this view. But thorough going anti-representationalism, as we've just seen, is inadequate as a comprehensive theoretical framework, since it cannot account for our ability to think and reason and plan.

A nice middle ground view, that accommodates both of these insights, is the hypothesis that language is not a consequence of our innate, representational capacities – this would be a DR view; rather, the development of language makes those capacities possible. Many are developing such accounts.

Clark, for example, argues that language is best viewed as a cognitive niche construction, a tool that "transforms problem spaces in ways that aid thinking and reasoning." (2006). On this view, language is a cognitive resource that complements the capacities of our brains: since our brains are pattern completers, but language has logic-like structure, language allows us to formulate inferences, make plans, think about tasks in a linear fashion, and so on, all cognitive activities our brains are not very good at doing. "Words and linguistic strings are among the most powerful and basic tools that we use to discipline and stabilize dynamic processes of reason and recall." (Clark, 2006).

More generally, usage-based theories take language to emerge over time "through the interaction of cognition and use." (Ibbotson, 2013; Bybee, 2010; Tomasello, 2003)

From a rather different approach, Christiansen and Chater (2008) argue (indirectly) against DR by rejecting evolutionary accounts of language. Instead, they argue that we should see languages as "'organisms'; i.e., highly complex systems of interconnected constraints, that have evolved in a symbiotic relationship with humans."

Finally, comprehensive accounts such as Terence Deacon (2011) set up a wider theoretical framework in the context of which we can develop the concepts we will need in order to explain the dynamic co-development of language and our representational cognitive capacities more completely.

Precisely what mechanisms underwrite our capacity to develop/use language, how language use yields mental representations, indeed, whether/how consciousness emerges as a result of language use, are some of the core research questions of this new view. As we've just seen, there is already work being done on different aspects of these questions, but the hard problem of accounting for the shift from non-representational perception to language to conscious self-reflection is still more of a chasm than a gap. A good way to begin closing it is by clarifying the nature of non-representational perception, since this is precisely the point at which DR views begin to diverge.

Ironically, one of most compelling positive arguments *for* a non-representational view of perception is suggested by a debate firmly entrenched in the DR framework. In that context, a distinction is drawn between non-conceptual and conceptual representations: bare perceptions such as *my seeing a flower now* are instances of the former, they do not require a network of concepts in order to be had, while judgements of perception such as my thought *that flower is lovely* are instances of the latter, these do require conceptual scaffolding in order to be had. Various arguments are offered in support of this conceptual/non-conceptual divide, but the one I will mention here, since it also motivates the non-representational position, is the following set of observations and inferences.

The perceptual hardware of many animals with whom we share an evolutionary history is very like our own. It is plausible, consequently, to suppose that our underlying perceptual *experiences* are similar as well. In perception, things in our environment appear to us in certain ways, e.g. apples look red, taste sweet, and sound crunchy. For our evolutionary cousins, though the phenomenal details will vary with the sensory capacities of the animal in question, these perceptual experiences are likely similar. It is implausible, on the other hand, that animals share our concepts — our conceptual capacity, after all, is what sets us

apart, cognitively speaking. Perception, therefore, involves a re-presentation of the sensed features of an animal's environment, but these re-presentations do not require prior concepts: an animal can see a red apple as red without having the concept *red.* (Peacocke 2001A, 2001B, Bermúdez 1998, 1994)

The problem with this reasoning is that it begins the story too high up, at personal-level perception. Not only do humans *engage* in sub-personal perception, we know this from the myriad perception illusions that continue to be uncovered; it is likely that the *majority* of our perceptions are of this sub-personal sort. From this vantage point, taking the features of the relatively small set of personal-level perceptions as the baseline for theoretical analysis seems unwarranted. A more cautious conclusion of the reasoning above would be that we share a sub-personal, non-representational, perceptual capacity with our evolutionary cousins.

Non-Representational Perception

When I perceive something, I am merely experiencing. I am not, as the DR story goes, responding to a representation *of* my experience. Such an idea takes hold when we take perception to be of a representation on the Cartesian theatre of the mind, rather than a direct relation to something in the world. As Dennett says, "… the brain does not bother "constructing" any representations … That would be a waste of time and (shall we say?) paint!" (Dennett, 1991)

An example might help make the distinction clearer. Suppose I am faced by an angry, barking dog. Immediately I begin to move away. I don't move away because I've compared my mental representation of this dog with others I have in my memory, because I've concluded that this is a potentially dangerous dog, because I believe that I ought to move away, and so on. I move away from the barking dog because the barking dog is threatening me, or perhaps parts of me, because the situation is compelling me to move away. The experience of the barking dog, which includes my becoming aware of the dog through various sensory channels, the adrenaline rushing through my body, my increased heart rate, and so on, all happen while I am, *at the same time*, moving away.

Merleau Ponty develops many such examples, describing how, in moving, I am also re-orienting myself so that I can get a 'maximal grip' on my situation, that is, be in a position to more accurately perceive what is salient to me now. (Merleau-Ponty, 1945). Dreyfus sums up this non-representational alternative like this:

> According to Merleau-Ponty, as an agent acquires skills, those skills are "stored," not as representations in the agent's mind, but as the solicitations of situations in the world. What the learner acquires through experience is not represented at all but is presented to the learner as more and more finely discriminated situations. … For example, what we have learned from our experience of finding our way around in a city is "sedimented" in how that city looks to us. (Dreyfus, 2007)

Things are complicated in the human case, of course, because we are *also* language users. Thus, while we are responding to the world in a non-representational, experiential way, *at the same time*, we are responding with the representational capacities we have developed: we can label our experiences with words, we can describe to ourselves and to others various aspects of these experiences, that is, we can analyse them rather than just see them as experiential wholes, and we can reason and plan about better or worse ways of moving away from and toward situations.

Stated in this way, we can see that this account aligns closely with two-systems views of cognition (Butterfill & Apperly, 2013; De Bruin & Newen, 2012; Apperly & Butterfill 2009). Our cognitive capacity includes an older, quick-response, system that we share with many animals — instincts, perceptual experiences, and bottom-up motor responses are governed by this mechanism — and a newer, slow-response, system, unique to humans — our higher level conceptual capacities are underwritten by this system.

On this view, contra the anti-representationalist, we *do* reflect on our own experiences, that is, represent them to ourselves. But, contra the proponent of DR, this is a much more recent cognitive capacity that we've acquired, perhaps as a result of, and with the representational help of, language development.

Of course there are objections to the two-systems view. Peter Carruthers, for example, has argued in a number of different places (Carruthers 2011, 2013) against it, but he does so within a DR framework. For all the reasons against DR I've presented here thus far, we need to explore two-systems views before rejecting them out of hand.

Representation-Hungry Problems Revisited

If our representational capacities are grounded, at least partly, in something external like language, which is the view being suggested here, aren't language-naïve animals who are nevertheless capable of solving 'representation-hungry problems,' clear counter-examples to the claim?

Some ground-breaking studies of scrub jays (Correia, Alexis, Dickinson, Clayton, 2007; Clayton & Dickinson, 1998) show that they appear to be among the hallowed group capable of responding to features of the environment that were once present, but are no longer: they are able to access the "what", "when" and "where" of past experience and thus pass Endel Tulving's litmus test for episodic memory. Not only are they able to remember what kind of food they have cached in a specific location, but they seem to be sensitive to how old the food is as well. More recent studies seem to show, in addition, a capacity for future planning, another aspect of the episodic memory system. (Clayton, Russell, Dickinson, 2009) We know that when humans access this future-directed aspect of episodic memory, they mentally rehearse past actions in possible future situations. Intuitively this seems like a paradigmatically representational activity — we are mentally presenting *to ourselves* possible situations.

If the way in which scrub jays manage this future thinking is similar to this, then they must also have a capacity for representation. Or so such reasoning goes.

But on the view of perception I am exploring here, experience is *not* a representational activity. If that's the case, then re-experiencing past experiences shouldn't be seen as a representational activity either. That is, to re-experience what happened in the past or what might happen in the future is, at base, to experience, however the experiences are caused. A scrub jay is compelled to move this way rather than that way, to dig here rather than there. Nothing in this ability requires that we adduce a representational capacity, though it does of course require some form of memory, which need not be seen as a representational capacity either. When we talk of 'body memory', for example, we mean quite explicitly the sort of memory that does not require representation.

Representation-Hungry Problems Refined

Seen in this way, rather than serving as a counter-example to the two-systems view being explored here, the example of the scrub jay shows us that not all seemingly representation-hungry problems need be solved in representational ways. Consequently, we might rename the category more perspicuously to "experience or memory-hungry" problems and refine the original representation-hungry category to include only those problems whose solutions require the *explicit* use of representations, not just past learning. But if recalling past experiences does not require representation, what does we might wonder.

Arguably one of the greatest advantages of a capacity for representation is the ability it entails to reason in the absence of emotionally charged situations. Indeed, it is precisely when we seem incapable of rising above the emotional challenges of a particular situation, that we need such tools most. Representations, because they are stand-ins for the situations or things they represent, are stripped of the contextual details we experience in perception. In other words, representations, such as the thought, *the cat is on the mat*, are amodal in a way that perceptions are not: entertaining the thought, *the cat is on the mat* feels differently from actually seeing the cat on the mat. Some recent studies demonstrate that access to amodal representations allows an agent to make choices, maximising ones, that would not otherwise be possible.

Reverse Contingency Tasks are a set of problems that require agents to make a choice on behalf of some other agent; the choosing agent then receives whatever is left over. In standard tasks, a choice is offered between large and small groups of desirable objects, e.g. bananas or candies. When a very desirable selection is placed in front of an agent, it is very difficult for the agent to overcome the pull to pick the largest group. But picking the largest group is not maximising since the choosing agent ends up with whatever was *not* chosen, namely, the smallest group. Studies have shown that when the same task is repeated with tokens (that the subjects have been trained to associate with the relevant class of desirable objects) rather than the objects themselves, both chimpanzees (Boysen et al. 1996, 1999) and capuchin monkeys (Addessi and Rossi 2011) are able to maximise their own reward more consistently.

These studies demonstrate a clear cognitive advantage to being able to reason in this aloof way: agents are able to make better long-term decisions for themselves when they have a tool for over-riding their quick-response, perception-based, primary cognitive system. There are many other contexts we could describe in which this ability confers an advantage, but due to space considerations, I will leave that for another paper.

Conclusion

As with any new theory, old questions disappear and new ones emerge: here is a brief look at some of the interesting changes that flow from this view.

The biggest payoff of this two-systems account is that it makes the problem of intentionality go away. Perception is not representation, so there is nothing to *naturalise* there. Thinking about and talking *are* representational activities, but their representational aspect derives from language itself: words and sentences are, paradigmatically, symbols.

New theoretical scaffolding will be required at the neural level: if our fundamental relation with the world is an experiential, not a representational, one, neurons can no longer be seen to 'represent,' 'detect,' or 'mean' anything at all. This is a radical shift, at the level of description and interpretation of results, for neuroscience.

At the level of language, our understanding of how and why it developed will change as well. On the DR framework, it is natural to see language development as growing out of our desire to communicate our rich, internal representational lives to one another. But when we let go of this Cartesian picture, we also let go of this communication motivation for language. Researchers are already working on developing an alternative view of why and how language developed in the first place, but there are many directions we might fruitfully explore.

One sign of a good theory is that it yields new questions that open up new lines of research. There isn't space here to discuss any of these in detail, but here are just a few of the less obvious avenues of inquiry that seem to open up once we shift from DR to this hybrid view:

— What role, if any, does language learning play in the development of 'self'?

— Does level of literacy co-relate with amount of time spent in self-reflective thought?

— If language is a tool that makes thinking *about* and talking possible, how much of our mental life is a result of an *overuse* of this tool? We know that overuse of our modern social media technologies, for example, can decrease one's capacity for mental focus. Perhaps some of the features of the contemporary human mind are side-effects of overuse of the language tool? An inability to be present to experience, which in turn can lead to many emotional and psychological problems such as depression, existential angst, feelings of loneliness, might be fruitfully explored from this angle.

— Following from the previous thought, we might investigate the effects of meditation as a cognitive enhancing technique from a new perspective: mindfulness meditation might be such an effective skill to develop, not because it yields a new state of awareness, but because it brings us back into touch with our experiential relation with our environment, an awareness we are capable of pre-language, but that is generally over-ridden by our representational system. On this view, language is the boon that makes this deeper awareness possible, but it is also the bane that makes meditation as an ongoing practise necessary. Meditation teaches us how to keep the flood of words and thoughts where they are — out there — and frees us from their bondage.

References

Addessi, E., Rossi. S. (2011). Tokens improve capuchin …*Proc. of the Royal Society B: Biological Sciences*, 278, 1707, 849–854.

Apperly, I, & Butterfill, S. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116, 4, 953–970.

Beer, R.D., (2003). The Dynamics of active categorical perception in an evolved model agent. *Ad.Beh.* 11:209.

Bermúdez, J. (1998). *The Paradox of Self-Consciousness*. Cambridge MA: MIT Press.

Boysen, S., Mukobi, K., Berntson, G. (1999). Overcoming response bias…*An.Learn.& Behav.*,27, 2, 229–235.

Boysen, S., Berntson.G., Hannan,M., Cacioppo.J. (1996). Quantity-based interference and symbolic representations ….*Journal of Experimental Psychology: Animal Behavior Processes*, 22, 1, 76–86.

Butterfill, S., & Apperly, I. (2013). How to construct a minimal theory of mind. *Mind & Language*, 28, 5, 606–637.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press.

Carruthers, P. (2013). Animal minds are real, (distinctively) human minds are not. *Amer.Phil. Quarterly*, 50, 233–247.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge: MIT Press.

Christiansen, M. & Chater, N. (2008). Language as shaped by the brain." *Behav.and Brain Sciences*, 31.5, 489-509.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10 (8):370-374.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge: MIT Press.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, 101: 401–431.

Correia S., Alexis D., Dickinson A., Clayton. N. (2007). Western scrub-jays …. *Current Biology*, 17: 856–861.

Clayton, N., Russell, Dickinson, A. (2009). Are animals stuck in time or are they chronesthetic creatures? *Topics in Cognitive Science*, 1 (2009) 59–71.

Clayton, N., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature* 395: 272–274.

Deacon, T. *Incomplete Nature: How Mind Emerged from Matter*. New York: W.W. Norton & Company. 2011.

De Bruin, L., & Newen, A. (2012). An association account of false belief understanding. *Cognition*, 123(2), 240-259.

Dennett, D. (1991). *Consciousness explained*. Boston: Little, Brown.

Dretske, F. (1988*). Explaining Behavior*. Cambridge: MIT Press.

Dreyfus, H. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence* 171 (2007) 1137–1160.

Freeman, W. (2000). *How brains make up their minds*. New York: Columbia University Press.

Haken H, Kelso JA, Bunz H. (1985). A theoretical model of phase transitions in human hand movements. *Biol Cybern*, 51(5):347-56.

Heidegger, M. (1927). *Being and Time*, (Trans.) J. Macquarrie & E. Robinson. New York: Harper & Row. [1962].

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in Psychology*, 4, 255.

Merleau-Ponty, M. (1945). *Phenomenology of Perception*, (Trans.) C. Smith. London: Routledge and Kegan Paul [1962].

Noë, A. (2010). Vision without representation. In (Eds) N. Gangopadhyay, M. Madary, & F. Spicer. *Perception, action, and consciousness: sensorimotor dynamics and two visual systems.* New York: Oxford University Press.

Peacocke, C. (2001A). Does perception have a nonconceptual content? *JPhil.*, 98: 239–264.

Peacocke, C. (2001B). Phenomenology and nonconceptual content. *Phil.Phenomen.Research*, 62(3): 609–615.

Peirce, C.S. (1998). *The essential peirce. volume 2*. (Eds.) Peirce edition Project. Bloomington: Indiana University Press.

Ramsey, W. (2003). Are receptors representations? *Journal of Exp.& Theoretical Artificial Intelligence*, 15:2, 125-141.

Rudder Baker, L. (2001). Practical realism defended: replies to critics. In (Ed) A. Meijers. *Explaining Beliefs: Lynne Rudder Baker and her Critics*. Stanford: CSLI Publications, 183-218.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Wheeler, M., (2004). Is language the ultimate artefact? *Language Sciences*, 26, 693–715

Wheeler, M., (2007). *Reconstructing the cognitive world: the next step*. Cambridge, MA: A Bradford Book.

Lasting Political Attitude Change Induced by False Feedback About Own Survey Responses

David Sivén (david.siven@gmail.com)
Thomas Strandberg (thomas.strandberg@lucs.lu.se)
Lars Hall (lars.hall@lucs.lu.se)
Lund University Cognitive Science, Lund University, Box 192, S-221 00, Lund, Sweden.
Petter Johansson (petter.johansson@lucs.lu.se)
Lund University Cognitive Science, Lund University, Box 192, S-221 00, Lund, Sweden.
Swedish Collegium for Advanced Study, Linneanum, Thunbergsvägen 2, S-752 38 Uppsala, Sweden.
Philip Pärnamets (philip.parnamets@ki.se)
Lund University Cognitive Science, Lund University, Box 192, S-221 00, Lund, Sweden.
Division of Psychology, Karolinska Institutet, Nobels väg 9, S-171 77, Solna, Sweden

Abstract

False feedback on choices has been documented to induce lasting preference change. Here we extend such effects to the political domainand investigatethetemporal persistence of induced preferences, as well as, the possible role the length of confabulatory justifications may play. We conducted a two-day choice blindness experiment using political statements, with sessions being roughly one week apart. Changes in political preferences remained one week after initial responses, and were most prominent in participants who were allowed to confabulate freely. These findings, being the first to demonstrate lasting preference change using choice blindness, are discussed in light of constructivist approaches to attitude formation through a process of self-perception.

Keywords:political attitudes; attitude change; choice blindness; persuasion; confabulation

Central to human social interaction is not only to understand the attitudes and preferences of your interlocutors, but also how to influence and shape them.In democratic societies this can be the difference between, successfully running for office and affecting the future course of one's society, or facing electoral defeat and slowly recede into the margins of history.

Attitudes are most commonly defined as "psychological tendencies that are expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken, 1993), parts of larger interrelated networks of attitudes and values (Jost, Federico & Napier, 2009). Apartfrom evidence indicating that attitudes are constructed within the domain of a specific task in a specific context, there is also evidence that preferences sometimes arise not before an actual decision is made, but rather as a result of the choice made or action performed (Ariely & Norton, 2008).

In the domain of public opinion, political attitudes are often seen as deliberate mental states, central in societal life, that are motivating important behaviors such as voting, or donating money to charity etc. However, there is also a perspective that political attitudes are highly flexible, often constructed on the fly, and prone to contextual influence (Azjen et al., 2014; Bishop, 2005). So, if attitudes are guiding action and behavior, then understanding the functions of attitude change and contextual influence are crucial in deeply understanding the cognition involved in constructing political preferences and persuasive messages.

Choice Blindness and Preference Change

Choice blindness is the finding that participants are at times blind to false feedback about the outcome of previously made choices. In the classical experiment participants make binary preferential choices between pairs of faces, and, following a covert manipulation, are subsequently presented with the opposite face to their original choice (Johansson, Hall, Sikström & Olsson, 2005). The key findings are that participants only detected about 25% of the manipulations, and that they often constructed coherent arguments supporting the opposite of the original choice.

Choice blindness has recently been proposed as a viable alternative to the *free-choice-paradigm* (Brehm, 1956) as a method for studying the effect of choices on later preferences. If participants are asked to make a second round of choices, they are more likely to change their preference following false feedback about their choice compared to when they receive veridical feedback (Johansson, Hall, Tärning, Sikström & Chater, 2014). This has been interpreted as supporting self-perception view of the mechanisms underlying preferential change (Johansson et al., 2014). Similarly, following a choice blindness manipulation participants' source memory becomes distorted leading them to believe they have made choices in line with the false feedback (Pärnamets, Hall & Johansson, 2015).

However, the longevity of choice-induced preferences is uncertain. Using a choice blindness manipulation Taya,Gupta, Farber and Mullette-Gillman, (2014), investigated the temporal extent of choice induced preferences over a two day experiment using photographs of faces as stimuli. They found an indication of preference change as a result of the false feedback in the short-term, but no lasting effect. This could, however, be due to the fact that the participants were made aware of the manipulation

prior to the second round of choices.By contrast, using an alternative paradigm known as blind-choice, where participants are led to believe they have made subliminally guided ("blind") choices, Sharot and colleagues found an effect lasting 2-3 years for ratings concerning preferred holiday destinations (Sharot, Fleming, Yu, Koster & Dolan, 2012).

Here we are concerned with investigating whether false feedback about responses in a political survey can be used to induce both immediate as well as lasting changes in participants' political attitudes. Previous work has demonstrated the applicability of the choice blindness paradigm to both moral (Hall, Johansson & Strandberg 2012), as well as, political attitudes (Hall et al., 2013) using covert manipulations of ratings on paper-based scales. In Hall et al. (2013) participants' responses on a range of salient political issues were manipulated weeks before a general national election. The false feedback was issued so as to go consistently against the grain of participants' left/right-wing orientation. Correction rates were low: 22% of manipulated statements, 48% of participants failed to make any corrections. More strikingly, participants were asked pre- and post-test to rate their voting intention on a left- to right-wing scale, with findings that almost half the participants changed their voting intentions in the direction of the manipulation of their underlying attitudes. However, any effect of the false feedback on the individual political issues was not tested.

Attitudes and Reasoning

An important aspect of the choice blindness paradigm is that it not only involves false feedback about a previous choice or rating, but also an element of confabulation. When confabulating, a person is unconsciously expressing fabricated aspects about oneself, the world, or the reasons behind a choice, without any deceptive intentions (Fotopoulou, Conway, & Solms, 2007). If the false feedback in a choice blindness task is accepted, we know that whatever reasons are given are confabulations.

The literature on confabulation in everyday discourse is scarce, and most research has traditionally focused on confabulation in the clinical domain (Hirstein, 2010). Inspired by research on attitude strength and elaboration, we wanted to investigate the effect of confabulation as a moderator of the attitude change induced by the false feedback. There are studies on persuasion suggesting that elaboration is a causal mechanism responsible for attitude strength (Barden and Tormala, 2014). Following this view, more elaboration produces stronger judgments held with greater confidence. In a study on the effects of elaboration on attitude strength, Barden and Tormala (2014) found that attitude strength largely depended on people's perception of their own elaboration.

In the context of attitudes, according to the *elaboration likelihood model* of persuasion (ELM) (Petty & Cacioppo, 1986; Petty, Haugtvedt, & Smith, 1995), the strength of an attitude can be predicted by the amount of issue-relevant thinking a person has spent on the object (Petty, Haugvedt, & Smith, 1995). The ELM builds on literature concerning attitude persistence and suggests that attitude change can be seen as resulting from two different kinds of persuasion. The first is persuasion by perceptual cues, not involving any careful deliberation as to the merits of a specific argument. The second is careful and issue-relevant consideration of a specific argument (Petty & Cacioppo, 1986).Clarkson, Tormala, and Leone (2011) found that if participants get to think about some object for up to 300 s, their confidence regarding their own attitudes directed at this object will increase, and their attitudes will become more polarized (extreme). Shorter deliberation times (60 s) were shown to lead to lower confidence and attitude depolarization (Clarkson, Tormala, & Leone, 2011).

In this study we were interested if changing the amount of confabulatory response participants were asked to give in response to the false feedback would affect the amount of change in their ratings following the initial manipulation. We expected more elaborate confabulationsto enhance the self-perception mechanisms hypothesized to underlie choice induced preference change and increase both the size of the attitude change as well as its longevity.

Method

Participants

We recruited 140 participants (88 female, 52 male) from the student population at Lund University, with an average age of 22.7 (*SD* = 3.0). Participants received two cinema vouchers in exchange for their participation in two experimental sessions, roughly one week apart.

Materials

For registering answers and manipulating the participants' ratings, we used a Self-Transforming Suvey (STS) specifically developed for providing false feedback about attitude responses (Strandberg et al., in prep.). The STS presented political statements one at a time in a randomized order. In addition to the STS, two follow-up paper-based surveys were used for measuring any attitude change following the manipulation.

All three surveys consisted of 12 political statements.The political statements were very specific, and divided into the three subcategories of healthcare, school politics, and environmental issues. Of the statements in the STS, six were carried forward to the two follow-up surveys. Of these six statements, four were always the same target statements. For each participant two of the target statements would be manipulated. The four target statements concerned salient political issues in Sweden at the time of the experiment [spring 2015]. The issues were: higher gasoline taxes for urban denizens, introducing subsidies for energy-efficient household appliances, making the government rather than local municipalities responsible for public schooling, and, introducing free after-school homework programs. All statements were constructed so as to state a proposed policy

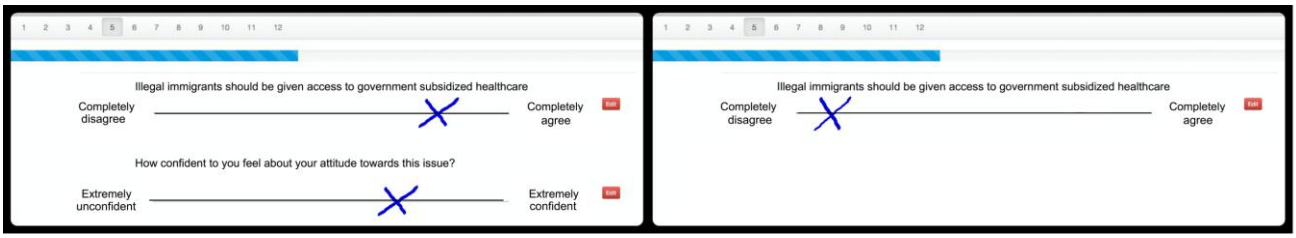


Figure 1 The STS. **Left:** Initial rating of a target statement. The X on the upper axis corresponds with agreeing with the statement. The X on the lower axis corresponds to neutral-to-high confidence**. Right:** False feedback following the completion of the survey. The application has now moved the X to correspond with disagreeing with the statement.

and give a brief explanation of that policy. One example full statement from the set of targets is:

The Swedish elementary school should be re-nationalized. Apart from the fact that local municipalities would lose much, albeit not all, influence, a re-nationalization would mean that the state becomes head of the school and assumes the responsibility for resource allocation and quality assurance.

The remaining items were considered fillers and had the same structure as the target statements. The filler items in each follow-up survey were unique to that survey.

Finally, an audio recorder was used to record the explanations of responses.

Choice Blindness and Confabulation Conditions

There were two independent variables in the experiment, one being the false feedback about responses, i.e. the choice blindness manipulation, the second being what confabulatory condition participants were assigned to.

The false feedback manipulation moved the participants rating, an 'X' they drew on the tablet across the midline of the bidirectionalvisual-analoguescale (see Fig. 1).The X ismoved by the STS to the other side of the axis andplaced either somewhere between 15% and 35% or between65% and 85%. The reason for not moving them closer to themiddle than this is to manipulate the participants intobelieving that they have a somewhat clear attitude for oragainst the statement.

In addition, participants were randomly assigned to one of two confabulatory conditions: Short or Long. In the Short condition, when participants were confronted with their ratings, they were asked to read aloud each statement as they appeared on the tablet, state where on the axis they had written their X, and if this meant that they agreed or disagreed with the statementas well as to what extent (e.g. whether the rating meant that they strongly agreed with the statement). In the Long condition, the participant was asked to, in addition to the above; also as thoroughly as possible account for the reasons behind their answer.

Procedure

The experiment consisted of two sessions, separated by roughly one week.

First Session (T1 & T2) The first session consisted of three subtasks: initial rating, interaction with manipulated answers, and follow-up rating.

First the participant was asked to answer 12 statements (T1) on a tablet, by marking an X along an axis in response to each statement. The endpoints were anchored as "Completely disagree" and "Completely agree". The experiment instructions emphasised that they were intended to represent extremes on a possible spectrum. Following each statement participants were asked to state their confidence in their attitude using the same scale anchored with "Extremely uncertain" and "Extremely certain". The participant was left to respond to the survey at her own pace.

Once the participant completed, the experimenter returned to the room and explained the interaction task. During this subtask, unbeknownst to the participants, the choice blindness manipulation took place. Each participant was told that the tablet would now randomly present four of the 12 statements and also the participants' responses to each of these statements. Instead the presented statements presented were the four target statements. Participants' responses to two of the presented statements had now been manipulated by the survey application running on the tablet. Participants were asked to read each of the presented statements and their rating, according to which condition, Short or Long, they had been assigned to (see above).

While performing the second subtask, all participants were knowingly and willingly recorded by an audio recorder. In the event of a participant in any way indicated that the response they saw did not correspond with their view, they were told that they could change their response if they wanted to, after which they could base their explanation on the position of the X in the now corrected position.

The third subtask (T2) consisted of the participant responding to a follow-up survey very similar to the one they had previously responded to on the tablet; only it was in the form of a traditional pen-and-paper survey. Six of the 12 statements were new to the participants, but the four possible target statements as well as two statements concerning healthcare, remained in the exact same form as in the STS. The participants were told that the motivation behind this additional survey was to measure if there was

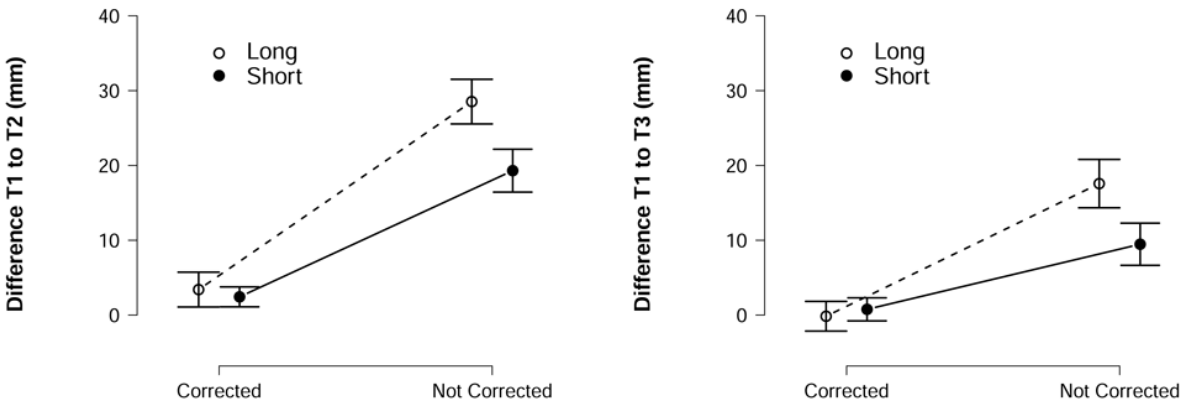


Figure 2 Differences in ratings during manipulated trials divided by correction and confabulatory condition. **Left:** Differences between original rating (T1) and first follow-up (T2). **Right:** Differences between original rating (T1) and second follow-up one week later (T3). Error bars are 95% confidence intervals.

any difference in responses if a survey was conducted electronically or physically. Further, the participants were told that it was likely that some of the statements that they had responded to on the tablet also would be included in the paper survey, since they were all randomly selected from the same bank of statements. Upon completion, the experimenter and the participant agreed on a time and date for the second session.

Second Session (T3) The participants were asked to return a week following the first session. The second session took place on average 6.3 days (*SD* = 1.8) later.During this second session (T3) participants were asked to fill out a survey of the political attitudes, similar to the one they had filled out the previous week. Once again, six statements were carried forward from the original and first follow-up survey, and six new statements were presented. Upon completion participants were debriefed in full, signed informed consent and data release statements and paid for their participation.

Analysis

All ratings were converted to a 0-100mm scale to facilitate comparisons between mediums (i.e. STS and paper-pen). We analyzed our data using linear mixed-effects models from the `lme4` package in R. Random-effects were modelled as per participant intercepts and slopes mirroring the full fixed-effects structure, or the maximally permitted structure that would converge. Significance of fixed-effects was assessed with likelihood ratio tests using the `car` package.

Results

Correction Rates

Of the 277 manipulated (M) trials, 134 (48.4%) were corrected by participants. Average by participant correction rate was 0.96 (*SD* = 0.78). 45 (32%) participants made no corrections, 56 (40%) one correction and 39 (28%) two corrections.

Overall Effect of Manipulation on Ratings

Rating differences comparing by manipulation We first analyzed differences in ratings for T2 and T3 compared to T1 (i.e. difference scores) based on manipulation. Time and manipulation were dummy-coded taking T2 and non-manipulated (NM) trials as the reference levels.

We found significant main effect of manipulation ($\chi^2_{(1)}$ = 39.23, $p = 3.7*10^{-10}$) as well as a significant interaction between time and manipulation ($\chi^2_{(1)} = 31.64$, $p = 1.9*10^{-8}$), but no main effect of time ($\chi^2_{(1)} = 2.41$, $p = .12$), with model conditional R^2 = .35. Interpreting the coefficients, participants were somewhat accurate in restating their original attitude in T2 during NM-trials ($b_{intercept}$ = 7.5mm, *SE* = 0.7) and this changed little from T2 to T3 (b_{T3} = 2.7mm, *SE* = 0.9). There was a large increase in distance from original rating for T2 manipulated (M) trials (b_M = 10.9mm, *SE* = 1.3) which decreased during T3 ($b_{T3_X_M}$ = -7.6mm, *SE* = 1.4).

Rating differences comparing by correction To compare the effect of correcting or failing to correct the manipulation, we performed the same analysis as above, subsetting the data on M-trials only. Differences scores were here transformed to be directional, meaning that a positive change is to be interpreted as a change in attitude in the direction of the manipulation (compared to T1). Time and correction were dummy-coded with T2 and corrected (C) trials as reference levels.

We found significant main effect of correction ($\chi^2_{(1)}$ = 96.87, $p = 2.2*10^{-16}$) and of time ($\chi^2_{(1)} = 30.85$, $p = 2.8*10^{-8}$), as well as a significant interaction between time and correction ($\chi^2_{(1)}$ =10.31, $p = .0013$), with model conditional R^2 = .47. Interpreting the coefficients, participants were displayed virtually no directional change in attitudes in T2 during C-trials ($b_{intercept}$ = 2.3mm, *SE* = 1.3) and this changed little from T2 to T3 (b_{T3} = -2.7mm, *SE* = 1.8). There was a large directional change in attitudes for T2 not

corrected (NC) trials (b_{NC} = 21.3mm, *SE* = 2.2) which decreased, but far from entirely, during T3 (b_{T3*NC} = -7.9mm, *SE* = 2.5).

Qualitative shifts by correction Given the shifts in attitudes post manipulation, we examined the proportion of these that crossed the mid-line of the attitude spectrum – hence implying a qualitative shift. In T2, 73% of responses represented such a shift for NC-trials, compared to 10% for C-trials. In T3, where the attitudinal effects of the manipulation were weakened, 41% of responses were still qualitatively shifted for NC-trials compared to 10% for C-trials.

Effect of Confabulation Condition

Rating differences comparing by manipulation To capture the effect of confabulation condition, we added confabulation as a fixed and random effect to the model. We analyzed all trials, with time, condition and manipulation being dummy-coded taking T2, short and non-manipulated (NM) trials as the reference levels.We found significant main effects of manipulation ($\chi^2_{(1)}$ = 44.14, p = 3.1*10^{-11}),and of condition ($\chi^2_{(1)}$ = 3.92, p = .048), but not of time ($\chi^2_{(1)}$ = 2.91, p = .088). These were qualified by significant interactions between manipulation and time ($\chi^2_{(1)}$ = 32.97, p = 9.4*10^{-9}), as well as between condition and time ($\chi^2_{(1)}$ = 4.15, p = .041). The interactions between manipulation and condition ($\chi^2_{(1)}$ = 0.040, p = .84) and the three-way interaction were not significant ($\chi^2_{(1)}$ = 0.21, p = .65). Model conditional R^2 = .37.

Interpreting the coefficients, participants were somewhat accurate in restating their original attitude in T2 during NM-trials in the Short condition ($b_{intercept}$ = 5.6mm, *SE* = 0.9) and the difference increased from T2 to T3 (b_{T3} = 4.3mm, *SE* = 1.3). There was a large increase in distance from original rating for T2, Short, manipulated (M) trials (b_M = 11.0mm, *SE* = 1.8) which decreased during T3 (b_{T3*M} = -8.2mm, *SE* = 1.8). The Long condition increased differences (b_{LONG} = 3.7mm, *SE* = 1.5), though this effect disappeared for T3 ($b_{LONG*T3}$ = -3.3mm, *SE* = 1.8). There was no specific effect of confabulation length for M-trials at T2 (b_{LONG*M} = -0.2mm, *SE* = 2.5) or at T3 ($b_{LONG*M*T3}$ = 1.2mm, *SE* = 2.7).

Rating differences comparing by correction To assess the effect of condition on directional change in ratings, we performed the same analysis as above, subsetting the data on M-trials only. Differences scores were here, again, transformed to be directional, meaning that a positive change is to be interpreted as a change in attitude in the direction of the manipulation (compared to T1). Time, condition and correction were dummy-coded with T2, Short condition and corrected (C) trials as reference levels. We found significant main effects of detection ($\chi^2_{(1)}$ = 120.51, p = 2.2*10^{-16}), condition ($\chi^2_{(1)}$ = 4.30, p = .038), and of time ($\chi^2_{(1)}$ = 28.88, p = 7.7*10^{-8}). These were qualified by significant interactions between detection and condition ($\chi^2_{(1)}$ = 7.32, p = .0068), as well as between detection and time ($\chi^2_{(1)}$ = 9.80, p = .0017). The interactions between condition and time ($\chi^2_{(1)}$ = 0.047, p = .49) and the three-way interaction were not significant ($\chi^2_{(1)}$ = 0.0078, p = .93). Model conditional R^2 = .42.

Interpreting the coefficients (see also Fig 2), participants displayed no directional attitude change in T2, C-trials for either Short ($b_{intercept}$ = 2.7mm, *SE* = 2.3) or Long (b_{LONG} = 0.3mm, *SE* = 3.2) conditions, with no changes for T3 (b_{T3} = -1.9mm, *SE* = 2.8). Importantly, there was a large directional attitude change for NC-trials (b_{NC} = 16.9mm, *SE* = 2.9), which was enhanced for Long condition ($b_{NC*LONG}$ = 8.7mm, *SE* = 4.1). Both directional effects due to NC-trials and Long condition diminished during T3 (b_{NC*T3} = -7.8mm, *SE* = 3.6; $b_{LONG*T3}$ = -1.5mm, *SE* = 3.7; $b_{NC*TALK*T3}$ = -0.5mm, *SE* = 5.1), though not sufficiently to remove the lingering directional effect (see Fig. 2).

Discussion

We investigated whether false feedback concerning specific ratings to political statements would influence later attitudes to the same statements. We found that participants' ratings, both immediately following false feedback and one week later, were shifted in the direction of the manipulation. These effects were large, representing considerable shifts in the direction away from the original attitude and towards the opposite. Notably there were no concomitant shifts for corrected trials, where participants had noticed the manipulation or shift for non-manipulated trials. A large portion of the shifts indicated qualitative shifts in attitudes. That there are shift in ratings immediately following the false feedback is in line with previous findings on choices between faces (Johansson et al., 2014; Taya et al., 2014). In those previous studies, however, the false feedback concerned preferential binary choices between pairs of faces. Here, in contrast, we influence political attitudes, a domain where we would expect higher resilience of participants underlying attitudes (cf. Druckman, 2004).

Importantly, the attitude shifts were not only present shortly following false feedback, but also when the same statements and rating task was administered one week later. This is the first demonstration of lasting preference shifts using a choice blindness paradigm (*contra* Taya et al., 2014), and indicate that the effects of false feedback might become integrated into the participants' set of attitudes. That choices and feedback about choices influence future preferences is in line with findings from other choice paradigms (Sharot et al., 2010). In the case of political attitudes, as studied here, the attitude shifts were obtained absent any reinforcement of the altered position. The participant only viewed the manipulation once, and was given one chance to state her opinion and [in the Long condition] give an explanation for it. After the experiment, the participants immersed themselves in their ordinary life for a full week, with their usual sources of information and political biases. It is therefore remarkable to find any lingering effect at all.

As such these findings provide support to inferential and constructivist accounts of preference and attitude formation, whereby the act of choosing has a determining effect on a

persons' preference set (Ariely & Norton, 2008). In previous work we have suggested that this is consistent with a self-perception account of preferences – i.e. that we infer our own preferences much like how we infer others' preferences; by observing and interpreting overt behavior (Johansson et al., 2014).

The change in ratings most marked in participants in the Long conditions, who had been asked to give longer explanatory statements concerning their manipulated responses. This is important, as it suggests that confabulation serves to consolidate choice induced preference change in the same way as veridical argumentation. The effect of the Long condition was especially salient in the one week follow-up.This supports theories of attitude construction involving the effects of elaboration and information processing on attitude change, such as the ELM (Petty, Haugtvedt, and Smith, 1995). For example, Barden and Tormala (2014) suggested that more elaboration produced better judgments, and that these in turn could be held with greater confidence. It has also been demonstrated that a person's attitude can strengthen if she perceives herself successfully defending it from persuasive attempts (Tormala and Petty, 2002; Knowles and Linn, 2004). These theories suggest that the perceived adjustment to contextual factors generate inferences about the metacognitive evaluation of the attitude. Following a similar logic, we could argue that observing yourself gives coherent arguments for a manipulated response, actually strengthens the induced attitude in the manipulated direction. And the more arguments supporting the induced attitude you hear yourself expressing, the more evidence you have to feel confident that this is what you truly believe.

In summary, our findings support constructivist accounts of attitude formation through a process of self-perception. Further, our findings support models predicting elaboration as underpinning strong attitude shifts. Together these results demonstrate a powerful influence of feedback in the moment with self-persuasion on attitudes in a domain of central importance to the functioning of democratic public life.

Acknowledgments

Many thanks to Anders Lindén at Andvision for implementing the design of the STS. PJ was supported by Bank of Sweden Tercentenary Foundation and The Swedish Research Council (2014-1371). LH was supported by Bank of Sweden Tercentenary Foundation (P13-1059:1) and The Swedish Research Council (2011-1795).

References

Ajzen, I., & Gilbert Cote, N. (2014). Attitudes and the prediction of behavior. In Crano, W. D., and Prislin, R. (Eds.), *Attitudes and attitude change* (pp. 289-311). New York, NY: Psychology press.

Ariely, D., & Norton, M. I. (2008). How actions create – not just reveal – preferences. *Trends in Cognitive Sciences*, *12*(1), 13–16.

Barden, J., & Tormala, Z. L. (2014). Elaboration and attitude strength: The new meta-cognitive perspective. *Social and personality psychology compass*, 8/1, (pp. 17-29).

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200.

Bishop, G. F. (2005). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Lanham; MD: Rowman and Littlefield publishers inc.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, *52*(3), 384–389.

Clarkson, J. J., Tormala, Z. L., & Leone, C. (2011). A self-validation perspective on the mere thought effect. *Journal of Experimental Social Psychology*, *47*(2), 449–454.

Druckman, J. N. (2004). Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects. *American Political Science Review*, *98*(04), 671-686.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Fotopoulou, A., Conway, M. A., & Solms, M. (2007). Confabulation: Motivated reality monitoring. *Neuropsychologia*, *45*(10), 2180–2190.

Hall, L., Johansson, P., & Strandberg, T. (2012) Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS One*, *7*(9), e45457.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS One*, *8*(4), e60554.

Hirstein, W. (2010). The misidentification syndromes as mindreading disorders. *Cognitive neuropsychiatry*, *15*(1-3), 233-260.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its Structure, Functions, and Elective Affinities. *Annual review of psychology*, *60*, 307-337.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making*, *27*(3), 281–289.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: *Attitude strength*, *4*, 93–130.

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*

Sharot, T., Fleming, SM., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting? *Psychological Science 23*(10), 1123–1129.

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. A. (2014). Manipulation Detection and Preference Alterations in a Choice Blindness Paradigm. *PLoS One 9*(9), e108515.

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me only makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of personality and social psychology*, vol. 83, no. 6, 1298-1313.

How should autonomous vehicles behave in moral dilemmas? Human judgments reflect abstract moral principles

Derek Powell[1] (derekpowell@ucla.edu)
Patricia Cheng[1] (cheng@lifesci.ucla.edu)
[1]Department of Psychology, University of California, Los Angeles,
Los Angeles, CA 90095 USA
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Self-driving autonomous vehicles (AVs) have the potential to make the world a safer and cleaner place. A challenge confronting the development of AVs is how these vehicles should behave in traffic situations where harm is unavoidable. It is important that AVs behave in ethically appropriate ways to mitigate harm. Ideally, they should obey a system of principles that both concur with human moral judgments and are ethically defensible. Here we compare people's moral judgments of AV programming with their judgments about the behavior of human drivers, with the goal of beginning to identify such principles. As many debates within ethics remain unresolved, empirical investigations like ours may guide the development of ethical AVs (Bonnefon et al., 2015). In addition, people's judgments about the behavior of AVs may serve as a window into the abstract principles people apply in their moral reasoning.

Keywords: Ethics; Moral Judgment; Robotics

Introduction

A number of auto manufacturers and tech companies are working to develop self-driving autonomous vehicles (AV). These developments hold great promise: creating safer roads (Gao et al., 2014), alleviating traffic congestion (Van Arem et al., 2006), reducing pollution (Spieser et al., 2014), and of course removing the tedious burden of driving through traffic. To be effective, autonomous vehicles must overcome a number of challenges: they must navigate to their destinations, steer properly to remain on roadways, and, perhaps most challengingly, identify objects and predict movements of pedestrians and other vehicles to avoid collisions. These challenges are rapidly being overcome, and experimental AVs like the Google Car have successfully driven for thousands of miles on public roadways (Waldrop, 2015).

Inevitably, AVs will face decisions with moral consequences—situations where an impending collision may injure or even kill human drivers or pedestrians. AVs will have to decide how to mitigate the harm caused by these situations in morally appropriate ways. Though they are by no means easily solved, navigation, steering, and object detection and avoidance are relatively well-defined problems. That is, it is clear what constitutes success: for example, arriving at the correct destination, staying on the correct side of the road, and braking before hitting a pedestrian, respectively. In contrast, moral judgments are sometimes ill-defined: individuals may differ in their moral judgments, and there is considerable disagreement within the field of ethics over how judgments should be made.

Traffic conditions may pose difficult choices similar to the classic Trolley dilemma. In the Trolley or Switch dilemma, a runaway Trolley threatens to kill five men working on a track unless it is redirected toward a side track with only one person working. The situation poses a dilemma: maximizing utility (saving the five) requires violating a moral rule (harming another person). Faced with the Switch dilemma, people generally make utilitarian judgments. The majority of people (85%; Hauser et al., 2007) judge that it is morally acceptable to flip the switch, redirecting the train away from the five and toward the one.

Vehicles in traffic might face dilemmas like the Switch dilemma when mechanical failures or road conditions prevent them from stopping or when other drivers act unpredictably. Should AVs be programmed to sacrifice lives if this produces better consequences overall? Bonnefon et al. (2015) presented participants with a switch-like scenario where an unavoidable deadly collision was about to occur with a group of 10 pedestrians unless an AV turned toward a single pedestrian. Participants approved of the AV turning to kill the one at rates very similar to those for human drivers. Participants were also generally willing to allow AVs to sacrifice the lives of their passenger by swerving to collide with a wall rather than a group of pedestrians. These researchers conclude that people are willing to accept AVs programmed for utilitarian sacrificial behaviors in at least some circumstances.

However, there are some moral situations where people refuse to sacrifice even when it would create the best

outcomes. For instance, although the vast majority of people approve of sacrificing one to save five in the classic Trolley dilemma, in the very similar Footbridge or Push dilemma a majority refuse to intervene. In the Push dilemma, a runaway trolley threatens to kill five workmen on a track. A large man is standing on a footbridge over the track, and the only way to save the five workmen is to push this man off the footbridge so that his body will stop the trolley. Approximately 88% of people refuse to push the man (Hauser et al., 2007). Whereas people are sometimes willing to trade-off between moral rules and better outcomes, at other times they are unwilling to do so—even for comparable outcomes. Will people approve of sacrificial behaviors from AVs in dilemmas that more closely resemble the Push dilemma, in contrast to their judgments about AVs facing switch-like traffic problems (Bonnefon et al., 2015)?

As it is impossible to anticipate all possible traffic scenarios in which an AV might find itself, AVs will need to respond to novel traffic situations according to abstract principles. Human moral judgments do not appear to perfectly adhere to any simple normative proposal, and researchers have offered divergent accounts of the principles underlying their judgments. For instance, some have argued that differences in Push and Switch judgments owe to affective reactions to the use of personal force (e.g. Greene et al., 2001) whereas others have argued that these judgments are the result of sophisticated deontological rules like the "doctrine of double effect" (DDE), which prohibits intentionally harmful actions, but may permit actions that produce a greater good for which harm is a foreseen but unintended consequence (Mikhail, 2011; Hauser et al., 2007). It is unclear how these proposals would apply to programmed AVs—it doesn't seem as if such programming involves "personal force" and the role of intentions is rather murky in this context. Can AVs even be said to have intentions in the sense of other moral agents?

Nevertheless, the development of moral AVs will require us to craft a system of principles that both satisfies human moral judgments in most cases and that is ethically defensible—even if imperfect.

Waldmann and colleagues (Waldmann & Dieterich, 2007; Wiegmann & Waldmann, 2014) have proposed that the causal structure of a moral situation is an important determinant of people's moral judgments. Their theory is related to the DDE, which claims that intentionality is inferred on the basis of the causal structure underlying the moral dilemma. However, rather than the intention of the agent, in Waldmann et al's theory it is the causal structure itself that drives moral intuitions. According to their theory, the locus of interventions in a causal system influences the attentional focus to different aspects of the moral dilemma, which in turn affects moral intuitions. In the switch dilemma, flipping the switch acts as a common cause of both the killing of the one on the side track and the saving of the five on the main track, highlighting both of these outcomes. In contrast, pushing the man off the bridge in order to stop the trolley represents a causal chain structure, and focuses subjects initially on the fate of the one victim.

This account leaves open the question of *why* causal models should be morally relevant. One possible answer, suggested by Kamm (2015), is built on the assumption that victims are endowed with rights (e.g., the right not to be harmed) and causal models highlight inter-victim relations. Victims stand in a *substitutive* relation if their roles in a situation could be arbitrarily swapped as, for instance, in the common-cause scenario. Shepard (2008) calls a similar concept the *symmetry* principle of *invariance under permutation of individuals*. Shepard regards this principle as a necessary overarching constraint for moral acts: An act is morally acceptable to such an extent as it would remain acceptable if individuals in a situation were permutated into different roles. In the trolley dilemma, sacrificing one to save five satisfies this criterion.

In contrast, in a causal chain scenario such as the footbridge, victims are in what Kamm (2015) terms a *subordinative* relation. Cause and effect, unlike effects of a common cause, are asymmetrical and not arbitrarily substitutable. Here Kamm argues that it violates our understanding of human rights to harm a person as a means of a later occurring greater good. Thus, the substitutability of agents' roles in the common-cause scenario, but not the causal chain scenario, could provide a justification for the seemingly opposite moral intuitions in a Switch and a Push dilemma.

The importance of causal structure (or likewise, of symmetry or substitutability) suggests that people's moral judgments may be influenced by the level of abstraction or concreteness at which they consider a situation. The Switch case represents an extreme—a rare case of perfect symmetry. In contrast, for situations with more complex causal structures—essentially any case that is not a simple artificial dilemma—there are many avenues by which varying degrees of asymmetry might be produced. If the intervention in the Switch scenario is considered concretely, the perfect symmetry of the scenario licenses the sacrifice of the one to save the five. However, if we consider the intervention in the abstract, other scenarios would be possible, and few of these will have perfectly symmetrical causal structures. On this account, encouraging abstract consideration of the dilemma may reduce approval for sacrificial actions. Making moral judgments about the programming of autonomous vehicles should encourage this sort of abstraction by introducing the possibility of other causal scenarios in which this programming will be applied. Thus, in considering moral judgments at a greater level of abstraction, we might expect people to less strongly endorse

sacrificial actions by AVs than in situations in which a human drives the vehicle.

The present experiments

To investigate how evaluating moral dilemmas at differing levels of abstraction would affect moral judgments, our experiments examined people's moral judgments of specific moral instances involving human drivers as compared with judgments about how AVs should be programmed to behave in these and other similar situations. In addition, the wording of the instructions were manipulated to encourage concrete or abstract consideration of the cases. We predicted that people would less strongly approve of sacrificial actions by programmed AVs than by human drivers, and when they are encouraged to think abstractly, in common-cause scenarios due to the possibility of the application of this judgment in other causal scenarios. However, no differences in judgments were predicted between judgments of AVs and human drivers in causal chain scenarios because these scenarios are already strongly aversive due to the asymmetrical inter-victim relation implied by the chain structure.

A secondary goal of the study was to examine people's moral judgments about AV programming in moral dilemmas with varied causal structures. Bonnefon et al. (2015) found that people are at least sometimes willing to allow AVs to sacrifice human lives to save the lives of others. However, their study only examined trolley-like cases with a common-cause structure, leaving open questions about the generality of their findings.

Experiment 1

Experiment 1 examined participants' moral judgments for dilemmas where a vehicle, either driven by a human or a driverless AV steered by a computer program, faced an unavoidable and deadly collision. We compared a car analog of the Switch and Push dilemma. In addition to manipulating the type of vehicle and dilemma, we also manipulated the abstractness of the problem description and the phrasing of the judgment probe for the AV condition, resulting in a concrete AV condition and an abstract AV condition. The main goal of the experiment was to study how abstractness of the steering mechanism (a driver facing a specific situation vs. a general program for the AV) would affect intuitions about Switch and Push cases.

Design. The experiment followed a 2 x 3 factorial between-subjects design (dilemma x abstraction): Participants were assigned either the Switch or Push dilemma, and to one of 3 levels of abstraction: either the human driver condition, the concrete AV condition, or the abstract AV condition.

Participants. Participants were 413 workers (237 female, median age = 32 years old) recruited from Amazon's Mechanical Turk (mTurk) work distribution website. Workers were paid $0.25 to participate in the study.

Materials. The Switch dilemma described a situation where the brakes of a truck have failed and the truck is headed toward a red car with three passengers. The driver (human condition) or the computer system directing the truck (AV conditions) must choose whether to continue on their present course, killing the three passengers in the red car, or to turn into a yellow car waiting at an intersection on a side street with only one passenger. Unfortunately, the truck cannot safely turn off the road onto a sidewalk as they are all full of pedestrians who will also be killed.

In the Push dilemmas, a runaway truck is again headed toward a red car with three passengers. However, in this dilemma the vehicle of interest is a blue car waiting at an intersection behind a yellow car with one passenger. The driver (human condition) or the computer directing the blue car (AV conditions) must decide whether to push the waiting yellow car with one passenger from the side street into the path of the runaway truck, saving the three passengers in the red car but killing the one in the pushed car. Each dilemma was accompanied by a diagram that depicted the situation, an example of which is shown in Figure 1.

The level of abstraction was manipulated between the concrete AV and abstract AV conditions by including additional wording in the abstract AV condition highlighting the generality of the principles employed by the AV:

"… In fact, there are thousands of possible dangerous traffic situations whose precise characteristics one cannot possibly anticipate. It is therefore crucial that the driverless car be equipped with a computer program that implements general rules designed to be applicable across a large variety

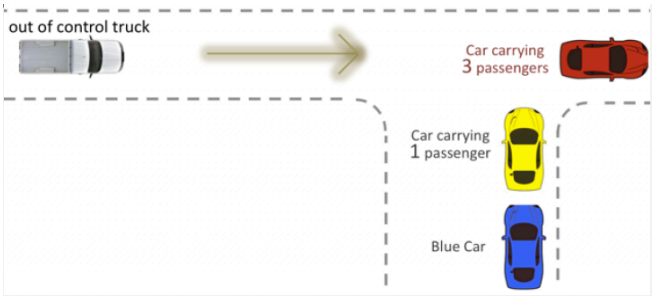


Figure 1: Diagram of traffic situation that was shown to participants for the Push dilemmas. The blue car is either an autonomous vehicle or is driven by a human driver, and must choose whether to push the yellow car into the path of the truck.

of possible scenarios. The program then directs how the car should best behave in these critical situations."

After reading the dilemmas participants were asked to make a moral judgment. In the human conditions participants were asked "Would it be appropriate for the [truck to turn onto the side street / blue car to push the yellow car into the crossing]?" for Switch and Push dilemmas, respectively. In the AV conditions participants were asked, "Would it be appropriate to design the steering function so that the [driverless truck turns onto the side street / blue car pushes the yellow car into the crossing]?"

The level of abstraction between the concrete AV and abstract AV conditions was also manipulated by additional language introducing the questions and problems. Before the "would it be appropriate …" question was asked, the abstract AV condition added instructions to "Imagine a world in which the steering function of a driverless car would decide that the blue car should push the yellow car into the crossing."

Procedure. Participants were recruited from mTurk and redirected to a Qualtrics survey website where study procedures were administered. Participants gave their consent to participate and answered some brief demographic questions before they were randomly assigned to a condition. Participants were then given some context about the topic of the study. Those assigned to the AV conditions read a brief explanation about the development of driverless cars, wherein it was explained that the cars would sometimes have to make decisions where an accident was unavoidable. Participants in the human driver conditions read a similarly worded introduction, but with the discussion of driverless cars omitted. Participants then considered their assigned dilemma and made a moral judgment about whether it was appropriate for the driver to take action in the dilemma. These judgments were made on a Likert scale ranging from 1 (completely appropriate) to 6 (completely inappropriate). Finally, participants answered some brief questions about how they made their judgments and whether they had ever seen the dilemmas before, two simple comprehension check questions (e.g., "two plus two is equal to what?") and whether they had paid attention and taken their participation seriously.

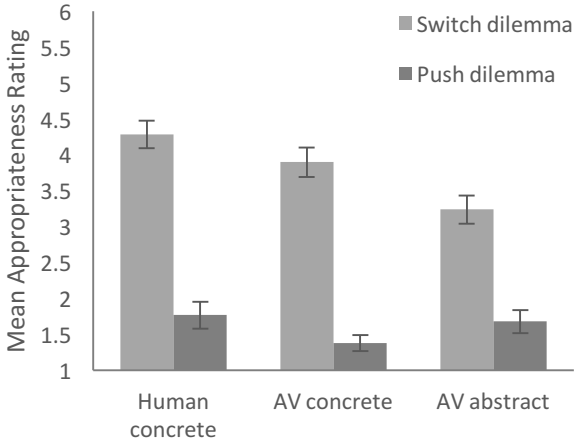


Figure 2: Participants' moral judgments in Experiment 1.

Results. Of the 413 participants recruited, 99 failed at least one comprehension check or indicated they had not paid attention. These participants were excluded, leaving 314 participants in the following analyses.

Participants' moral judgments are shown in Figure 2. These judgments were examined with a 2 x 3 (dilemma x abstraction) between-subjects ANOVA. Participants approved of acting in the Switch cases much more strongly than in the Push cases for all conditions, as indicated by a significant main effect of dilemma, $F(1, 308) = 231.5$, $p < .001$.

In support of our causal model explanation of the Switch and Push dilemmas, the main effect of abstraction was significant ($F(2, 308) = 5.359$, $p = .005$) and there was a significant interaction, $F(2, 308) = 4.898$, $p = .008$. The interaction appears to be driven by a difference between conditions for the Switch dilemma, and an absence of differences for the Push dilemma. Whereas significant differences were observed for the Switch dilemma between the abstract AV condition and the human condition ($t(103) = 3.886$, $p < .001$), as well as between the abstract AV condition and the concrete AV condition ($t(104) = 2.355$, $p = .02$), no significant differences were observed among the Push scenarios (all ps $> .05$). The contrast between the concrete AV and human condition was also non-significant, $t(101) = 1.39$, $p = .168$.

Experiment 2

To further test the effect of abstractness, we added an abstract human driver condition in Experiment 2. In other words, we added a manipulation of abstraction by wording. Two additional goals are to test the replicability of the results of the previous study and to evaluate a floor effect as an alternative explanation of the lack of influence of driver (human vs AV) for the Push dilemma.

Participants. Participants were 610 workers (351 female, median age = 32 years old) recruited from Amazon's Mechanical Turk (mTurk) work distribution website. Workers were paid $0.25 to participate in the study.

Design, Materials, and Procedure. The materials and procedure of this experiment were nearly identical to those of Experiment 1 save for the addition of an abstract human driver condition for Switch and Push dilemmas. This resulted in a 2 x 2 x 2 factorial design (dilemma x driver x wording).

In the abstract human driver conditions, participants were introduced to the dilemmas in the context of developing a training manual for delivery and ridesharing companies. The introduction was similar to the abstract AV conditions, stressing that the rules must be made to apply across many different scenarios. For the Switch and Push variants of these conditions, the moral judgment question read "Imagine a world where general guidelines in the transportation company manual prescribe that the driver of the [truck should turn onto the side street and hit the yellow car / blue car should push the yellow car into the crossing]. Would it be appropriate to write the general guidelines so that a company driver [would turn the truck onto the side street / in the blue car would push the yellow car into the crossing]?"

To examine the possibility of floor effects for the Push items, participants who were assigned to the Push dilemma conditions were also assigned to make a judgment about a Transplant dilemma. This dilemma asks participants to judge whether it is acceptable for a doctor to kill a patient in order to use his organs to save several other patients.

Results. Of the 610 participants originally recruited, 83 were excluded for failing at least one comprehension check or for indicating that they had not paid attention, leaving 527 participants in the final analysis.

Participants' moral judgments in Experiment 2 are shown in Table 1. These judgments were analyzed using a 2 x 2 x 2 (dilemma x driver x wording) between-subjects ANOVA. As in Experiment 1, approval was much lower for Push dilemmas than Switch dilemmas, indicated by a significant effect of dilemma, $F(1, 519) = 422.8$, $p < .001$. Replicating the results of Experiment 1, moral approval was lower in AV conditions than human conditions for the Switch dilemma but not the Push dilemma, as indicated by the two-way interaction between dilemma (switch or push) and driver (human or AV), $F(1, 519) = 4.351$, $p = .037$.

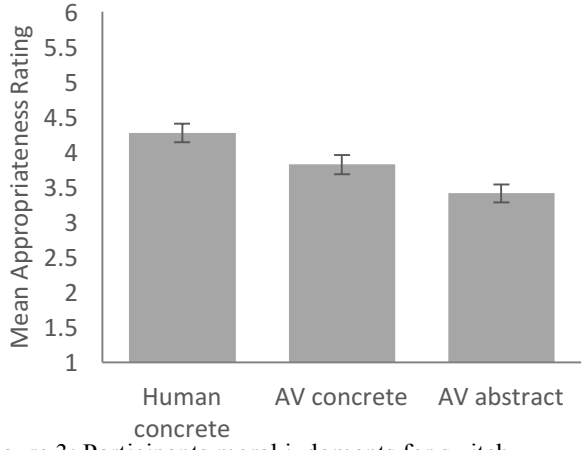


Figure 3: Participants moral judgments for switch dilemmas averaged from Experiments 1 and 2.

Also as predicted, moral approval generally appears to be lower for abstract wordings than for concrete wordings, although this effect was not significant, $F(1, 519) = 2.889$, $p = .09$. Although this manipulation was meant to introduce a greater level of abstraction in the same way as the AV conditions, the manipulation was unfortunately not fully equated. For example, the manual explicitly reminded subjects of the specifics of the dilemmas (Switch, Push), whereas the instructions for abstract AV condition did not. This may have made it more difficult for participants in the abstract human condition to invoke abstract principles.

Table 1. *Mean moral appropriateness judgments by dilemma (switch, push), driver (human, AV) and wording (concrete, abstract) from participants in Experiment 2.*

		Human	AV
Switch	Concrete	4.25 (1.49)	3.75 (1.55)
	Abstract	3.83 (1.65)	3.54 (1.48)
Push	Concrete	1.54 (1.18)	1.45 (1.04)
	Abstract	1.29 (0.85)	1.55 (1.14)

Note: Means with standard deviations in parentheses.

A further investigation of participants' Switch dilemma judgments in a 2 x 2 ANOVA (driver x wording) reveals a significant effect of driver ($F(1, 260) = 4.348$, $p = .038$). These findings qualitatively replicate the differences between conditions in Experiment 1, although only the contrast between the abstract AV and concrete human condition was significant, $t(122) = 11.36$, $p < .001$. All other effects were non-significant (all Ps > .06).

The absence of differences among the Push scenarios seems unlikely to be a simple floor effect, as indicated by the still lower approval for the Transplant dilemma (Mean = 1.29, SD = .819), $t(262) = 2.771$, $p = .006$.

Meta-Analysis of Experiments 1 and 2

With the exception of abstract human driver conditions, the materials and procedures of Experiments 1 and 2 are identical, allowing data from these experiments to be pooled for their shared conditions for increased statistical power.

Of particular interest are comparisons between the human, concrete AV and abstract AV conditions for the Switch dilemma, as qualitatively similar yet somewhat different patterns of results were observed in Experiments 1 and 2. Pooling these data reveals significant contrasts for all pairwise comparisons: lower approval was observed for the abstract AV condition as compared with the concrete AV ($t(240) = 2.125$, $p = .035$) and human conditions, $t(238) = 4.609$, $p < .001$. Lower approval was also observed for the concrete AV condition as compared with the human condition, $t(228) = 2.329$, $p = .021$.

Discussion

Across two experiments we found that people were generally willing to allow AVs to act to sacrifice one life in order to save three others in a common-cause scenario like the Switch but not in a causal chain scenario like the Push. However, as compared with judgments of human drivers, people were less willing to see AVs sacrifice in a common-cause dilemma. There was no analogous difference for a causal-chain dilemma. These findings are consistent with the role of causal structure in human moral judgments and with the constraint of symmetry based on Kamm's (2015) analysis of inter-victim relationships (see also Waldmann, Wiegman, & Nagel, in press).

Our main goal was to provide a theoretical account that both concurs with human moral judgments and is ethically defensible. We tested and confirmed the predicted interaction between the abstractness of construals of moral dilemmas and the causal structure of moral dilemmas. We acknowledge, however, that the experiments were not designed to test our account of the interaction against alternatives. It may be possible, for example, to derive predictions from two-system theories, although we do not see how they can make plausible predictions about abstractness. If anything, this account should seem to predict more utilitarian reasoning in abstract cases, which should not trigger emotional involvement, contrary to what we found.

Another factor that is often neglected in research on moral dilemmas concerns legal considerations. Trolley dilemmas describe rare situations which do not routinely happen. Therefore, there are no clear legal regulations about such accidents. By contrast, accidents involving cars are frequent. Traffic is strictly regulated by laws. In fact, articles about AVs typically focus on the possibility of accidents and on issues of liability. Thus, it seems plausible to assume that a programmer of an AV steering mechanism will try to prevent situations in which the AV either kills its owner or innocent bystanders. However, legal considerations seem less able to explain the lower moral approval we observed in the abstract AV condition compared to the concrete AV condition, both in Experiment 1 and in our meta-analysis. Both conditions involved the program in AVs operating across all traffic situations, and the company developing the program would be equally liable.

The development of ethical autonomous machines is an important and exciting application of the budding field of experimental ethics, to which we hope significant further inquiry will be devoted. Keeping pace with the rapid development of AI research and engineering certainly demands it.

References

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *arXiv*:1510.03346 [cs.CY]

Gao, P., Hensley, R., & Zielke, A. (2014). A road map to the future for the auto industry. *McKinsey Quarterly*.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–8.

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*, 1–21.

Kamm, F. M. (2015). *The trolley problem mysteries*. New York: Oxford University Press.

Shepard, R. (2008). The step to rationality: The efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, *32*, 3-35.

Spieser, K. et al (2014). Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In G. Meyer and S. Beiker (Eds.), *Road Vehicle Automation*, 229–245. Springer.

Van Arem, B., Van Driel, C. J., & Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, *7*, 429– 436.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science*, *18*, 247–53.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (in press). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Trémolière (Eds), *Moral inferences*. Hove: Psychology Press.

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature*, *518*, 20–23.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28–43.

Examining Search Processes in Low and High Creative Individuals with Random Walks

Yoed N. Kenett (yoed_kenett@brown.edu)
Joseph L. Austerweil (joseph_austerweil@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Providence, RI 02912 USA

Abstract

The creative process involves several cognitive processes, such as working memory, controlled attention and task switching. One other process is cognitive search over semantic memory. These search processes can be controlled (e.g., problem solving guided by a heuristic), or uncontrolled (e.g., mind wandering). However, the nature of this search in relation to creativity has rarely been examined from a formal perspective. To do this, we use a random walk model to simulate uncontrolled cognitive search over semantic networks of low and high creative individuals with an equal number of nodes and edges. We show that a random walk over the semantic network of high creative individuals "finds" more unique words and moves further through the network for a given number of steps. Our findings are consistent with the associative theory of creativity, which posits that the structure of semantic memory facilitates search processes to find creative solutions.

Keywords: Creativity; Semantic Networks; Random Walks; Cognitive Search

Introduction

How do people vary in their creative ability? While creativity is a multifaceted construct, here we focus on the role of representation and search processes in memory to creative ability. One relevant theory of what differentiates low and high creative individuals focuses on bottom-up factors: the structure of semantic memory (Kenett, Anaki, & Faust, 2014). An alternative theory focuses on the role of top-down, executive processes. This theory argues that top-down control is necessary to produce observed differences between low and high creative individuals (Benedek et al., 2014). In this paper, we demonstrate that a bottom-up, associative account is sufficient to produce some of these results.

Previous work on the relation between creativity and cognitive search has examined how different search processes over semantic memory (memory for knowledge and facts, Jones, Willits, & Dennis, 2015) are related to goal-directed creative tasks (Christoff, 2013; Sawyer, 2011). One major distinction between search processes is whether they are controlled (goal-directed) or uncontrolled (undirected; Christoff, 2013). Controlled search processes involve goal directed search for a specific creative solution (Mednick, 1962). Uncontrolled search processes wander across semantic memory, which can lead to novel combinations and insight (Baird et al., 2012). According to this account, differences in creative behavior are related to differences in memory structure that facilitates random mind wandering to connect distant associations, which are then evaluated for their appropriateness (Christoff, 2013; Sawyer, 2011).

Recent studies have used computational network science to analyze semantic memory structure as a semantic network (e.g., Baronchelli et al., 2013). A semantic network is a set of nodes and edges, where nodes correspond to words or concepts and edges connect pairs of nodes and signify some sense of association between them. Previous theoretical work has proposed that creative solutions and insight are the result of either sophisticated search processes and/or the creation of new edges in semantic networks (Schilling, 2005). However, previous work has not examined the nature of this search process in creativity, whether it be controlled or uncontrolled.

In this paper, we examine how well uncontrolled search on semantic networks can capture some of the differences in how low and high creative individuals conduct cognitive search. To this end, we simulate random walks over the semantic networks of low and high creative individuals. Based on the associative theory of creativity, we hypothesize that the structure of the semantic network of high creative individuals enables them to use simple search processes that reach further and to more weakly connected concepts, than low creative individuals. Specifically, we predict that random walks over the semantic network of high creative individuals will reach more nodes than the walks over the network of low creative individuals. Further, the similarity of the initial and final visited nodes in the walk will be weaker for the network of high creative individuals than for the network of low creative individuals.

The plan of the paper is as follows: First we describe previous work on uncontrolled search and differences in semantic memory structure between low and high creative individuals. Next, we investigate whether a purely uncontrolled random walk process on the two semantic networks captures differences in cognitive search performance between low and high creative individuals. We conclude with a discussion of the implications and limitations of our simulation results.

Previous work

Creative and Uncontrolled Search

Current neurocognitive research has progressed our understanding of the roles of specific cognitive processes (such as working memory and attention) and how they

interact to produce creative behavior (Sawyer, 2011). One of these processes is the ability to search through memory and connect seemingly unrelated concepts into something novel (Mednick, 1962). Mednick (1962) theorized that novel combinations of unconnected concepts are more creative the farther apart they are in memory (but see Sawyer, 2011 for an alternative viewpoint). He proposed that this will be evident in the type of associations low and high creative individuals generate to cue words: When presented with the cue word *table*, low creative individuals will generate responses that are mainly restricted to the common response *chair*. Conversely, high creative individuals will generate less frequent responses such as *leg* and *food* (Mednick, 1962). To test his theory, Mednick et al. (1964) had low and high creative individuals generate responses to cue words in a fixed amount of time. He found that high creative individuals generate more responses than low creative individuals. However, what can researchers infer about the differences in representation and processes of low and high creative individuals based on these results? To do so requires a formal account of the representation and processes responsible for producing responses to cognitive search tasks.

A classic uncontrolled search process is spreading activation (Collins and Loftus, 1975). According to this search process, activation spreads over links through words and quickly dissipates with time and distance. Spreading activation over networks can also capture similarity relations: Two words are similar to the extent that they activate each other. Computationally, spreading activation can be implemented as a random walk over a network. Starting at a particular node, a random walk selects an outbound edge with a probability proportional to the edge's weight and moves across it. As this process progresses, it explores more nodes in the network. Analogous to how spreading activation decays over a network, the probability that a walk moves from one node to another decays in their distance. Thus, the probability of moving from one node to another in a small number of steps captures their similarity.

Recent research has explored how random walk models can capture memory retrieval (Abbott, Austerweil, & Griffiths, 2015; Capitán et al., 2012; Griffiths, Steyvers, & Firl, 2007) and performance in creative tasks which require cognitive search (Bourgin, Abbott, Griffiths, Smith, & Vul, 2014; Smith & Vul, 2015). These studies investigated how well a random walk on a representation captures *general* performance on cognitive search and creative tasks. However, they have not examined whether *differences* in creative ability can be understood in terms of the same random walk process on different representations.

In the present study, we conduct naïve random walk simulations on the semantic networks of low semantic creative (LSC) and high semantic creative (HSC) individuals estimated in a previous study (Kenett et al., 2014). This study found that the semantic networks of low and high creative individuals had different structural properties. This was the case despite both networks having an equal number of nodes, edges and average number of edges per node.

Using Kenett et al.'s semantic networks of LSC and HSC individuals, we can test Mednick's (1962) associative theory of creativity using random walks. We formalize the search process proposed by the associative theory as an uncontrolled random walk, and predict that (on average) a random walk over the semantic network of HSC individuals will visit more nodes that are weaker in similarity than an equivalent length random walk over the semantic network of LSC individuals. This would reproduce previously reported differences in human performance by LSC and HSC individuals in generating and judging the strength of associative responses to cue words (Mednick et al., 1964; Rossman & Fink, 2010). We test these predictions via random walk simulations to see whether differences in representation are sufficient to produce observed differences in creative performance.

Semantic Networks of LSC and HSC Individuals

Here we describe how Kenett et al., (2014) estimated different semantic networks for LSC and HSC individuals.

Creativity Assessment - Participants (N = 140) were recruited as part of a larger research study on individual differences in creative ability (Kenett et al., 2014). They completed the following creativity tasks: The Remote Association Test, which measures associative thinking related to creativity (Nevo & Levin, 1978), Tel-Aviv University Creativity Test (Milgram & Milgram, 1976), a battery of divergent thinking tests (e.g., *what are all the things you can do with a brick*), frequently used in creativity research (Runco & Acar, 2012), and the Comprehension of Metaphors task (Faust, 2012), which compares processing of word-pairs that have different semantic relations (either literal, conventional metaphoric, novel metaphoric, or are meaningless). Participants were classified as LSC or HSC individuals using these scores. To do so, Kenett et al. (2014) used the JMP classification and regression tree approach (Galimberti & Soffritti, 2011), to predict performance on the Remote Association Test based on their divergent thinking performance (scores on fluency and quality of responses). Using the decision tree, they defined participants in the lower third of performance as low creative and those in the upper third as high creative. This classification was verified based on the performance of the two groups on the Comprehension of Metaphors task (not used to construct the decision tree).

Semantic Network Estimation - The semantic networks of the LSC and HSC groups were computed using the computational approach developed by Kenett et al. (2011). This approach consists of two parts. First, participants generated continuous free association responses to cue words. The LSC and HSC groups generated free associations to a list of 96 cue words. The cue words were based on fluency norms collected to a list of 36 categorical norms gathered by Henik and Kaplan (2005; e.g. fruits, trees, countries). The top four high frequency words from

each category were selected. These words were then tested for their degree of concreteness, on a seven point Likert scale, and only concrete words were selected (words scoring more than three points on the scale). The final pool of cue words consisted of 96 words from 24 categories.

Kenett et al. (2014) used the associative responses for these 96 cue words to estimate semantic networks for LSC and HSC individuals. This was achieved in the following steps: First, the data was preprocessed to standardize responses and fix any spelling mistakes. This resulted in a matrix where each row was a unique associative response, each column was a cue word and each cell contained the amount of participants generating response *i* to cue word *j*. Second, a *N* x *N* association matrix *A* was constructed, where element (*i, j*) was the Pearson's correlation of the response similarity between cue words *i* and *j*. Spurious correlations from the connectivity matrix were removed using a filter (planar maximal filtered graphs) and any non-zero cells were binarized to equal one. Thus, the resulting semantic network is unweighted (the weight of any existing edge is one) and undirected (symmetric relations). Although we plan to explore weighted networks in the future, some previous work has found qualitatively similar behavior for unweighted and weighted semantic networks (Abbott et al., 2015) and so, we used an unweighted network for simplicity and brevity. Importantly, the semantic networks for different groups (LSC and HSC) were comprised of the same nodes (96 cue words) and had an equal number of edges (282 edges). Further, the average degree, the average amount of edges per node in both networks was equal (average of 5.88 edges per node). Thus, differences in behavior between the random walks on the two networks cannot be explained as simply there being more connections in one of the networks.

Current Work

A naïve random walk approach was used to examine the theory that HSC individuals can search further in their semantic networks and reach weaker related nodes. Accordingly, we examine two properties of each walk: the amount of unique visited nodes (indicating the breadth of the search) and the distance between initial and final visited nodes (indicating the strength between initial and final nodes). The random walk analysis over the semantic networks of the two groups was conducted in the following steps: First, we computed the transition probability matrix for each group. As the networks are unweighted, and undirected, the transition probability T_{ij} of moving from node *i* to node *j* is

$$T_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{kj}} \quad (1)$$

where A_{ij} is the fully processed association matrix (1 if *i* is connected to *j*, and 0 otherwise) and the denominator is the number of nodes that node *j* is connected to.

Second, we choose an initial starting node (cue word) for both networks from which the walk initiates. For each walk simulation, the initial node (X_{IN}) is randomly chosen from a uniform distribution over the 96 nodes in the network. The random walk starts at $X_0 = X_{IN}$, and then at step n randomly generates the next state X_{n+1} according to the transition matrix *T*. In this model, the transition matrix is unweighted, meaning that the next state is uniformly chosen at random from all nodes connected to node X_n. Further, our model is a *non-jumping* model, which means that the cue word simply initiates the walk. We did not include jumps because previous work found that they did not substantially affect the pattern of first visits to each node by random walks (Abbott et al., 2015).

After running the random walk simulations on the LSC and HSC semantic networks, we examined differences in "search" behavior (the behavior of the random walks) due to the structure of the semantic networks. We computed two different measures of how "creative" a walk was to examine whether the random walks on the LSC and HSC networks capture differences in creativity.

The first measure of walk creativity was the number of unique nodes visited by the search. This measure indicates the breadth of the search achieved by the walk. The second measure of walk creativity was the similarity between the initial and final node visited by the walk. This was defined as

$$s = \exp(-D_{ij}) \quad (3)$$

where D_{ij} is the length of the shortest path between initial (*i*) and final (*j*) nodes through the network. The shortest path between two nodes is the smallest number of edges a walk could traverse to get from one node to the other node. Both measures were averaged over simulations per different numbers of steps and compared via Wilcoxon signed rank tests.

Results

We ran 10,000 random walks over both networks for a varying number of steps (10-200 steps in increments of 10 steps). Similar results held for 100,000 simulations. At the start of each simulation, the initial node was randomly chosen and identical for both networks. For each simulation, we computed the mean number of unique nodes visited during the walk and the similarity score between initial and final nodes per amount of steps for the two groups (as described in the subsequent subsections).

Unique Visited Nodes

For both walks, the average amount of unique visited nodes in the network increased with the number of steps of the walk. For the walks on the LSC network, the average number of unique visited nodes ranged from 7.21 (1.42)[1] for

[1] We report standard deviation in parentheses.

simulations taking 10 steps to 52.52 (10.35) for simulations taking 200 steps. For the walks on the HSC network, the average amount of unique visited nodes ranged from 7.24 (1.43) for simulations taking 10 steps to 54.94 (9.34) for simulations taking 200 steps. Overall, walks over the HSC network visited more unique nodes for a given number of steps than over the LSC network (Wilcoxon rank test significant at $p < 0.001$ for all walks longer than 20-steps).

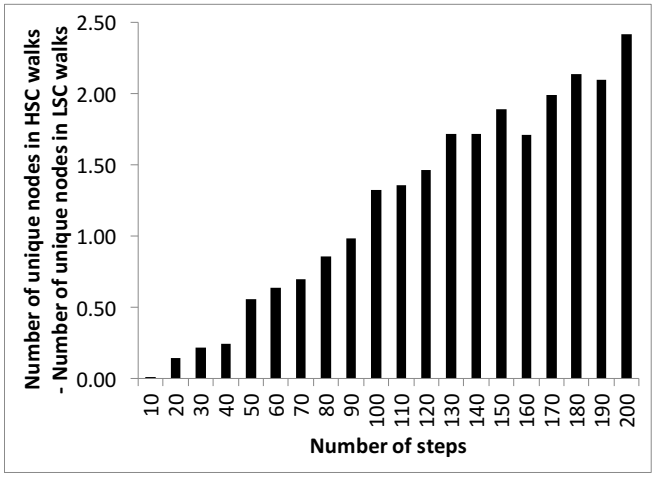


Figure 1: Difference score of the number of unique nodes visited by walks over the LSC and HSC networks varying the number of steps.

As shown in Figure 1, the difference between the number of unique visited nodes by the HSC and LSC walks increases with the number of steps. Generally, the increase is monotonic, but it is not always. In these cases, it is likely that increasing the number of steps allowed the LSC walk (on average) to reach nodes it had not reached with fewer steps, but were already reached by the HSC walk. It is not simply noisy – running more simulations led to similar results. These results are consistent with previous work showing that high creative individuals generate more associative responses than low creative individuals (Mednick et al., 1964).

Similarity Between Initial and Final Nodes

For walks on both networks, the average similarity score between the initial and final node in the walk decreased with the number of steps taken by the walk. For the simulations on the LSC network, the average similarity score ranged from 0.22 (0.22) for simulations taking 10 steps to 0.07 (0.14) for simulations taking 200 steps. For the simulations on the HSC network, the average similarity score ranged from 0.21 (0.22) for simulations taking 10 steps to 0.07 (0.13) for simulations taking 200 steps. Up to 110 steps, HSC walks resulted in lower similarity scores between the initial and final nodes (Wilcoxon rank test significance at $p < 0.1$ for 20-steps and $p < 0.01$ for all other number of steps). This reflects that until 110 steps, walks on the LSC network tended to remain closer to the initial node than walks on the HSC network. Related to this, Beaty & Silvia (2012) found that as more responses are generated in fluency tasks, the responses become more "creative" (as judged by subjective measures of the novelty and uniqueness). From 120 steps onwards, the results are less consistent and even reverse. For example, at 150-160 steps, the walk over the LSC network had a smaller average similarity score than the walk over the HSC network (Figure 2).

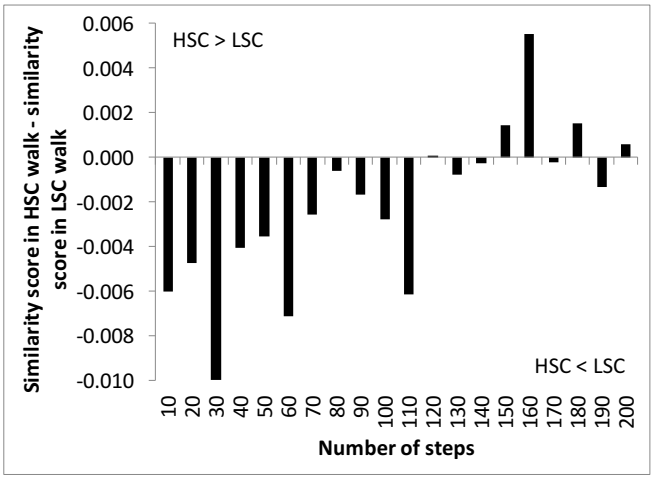


Figure 2: Difference score of the average similarity (shortest path) between initial and final visited nodes by walks over the LSC and HSC networks varying the number of steps.

This reversal in the difference score of the average similarity between initial and final visited walks by the two groups may be due to the large number of steps in the walk relative to the number of nodes (96), which enables the walk to traverse most of the network. Thus, the walk over the LSC network may reach farther in the network which will lower its average similarity score.

One question we had was whether the differences in similarity are simply due to the HSC network visiting more unique nodes. To control for this possible confound, we examined the similarity between the starting and ending node holding the number of visited nodes constant between walks on the two networks (rather than the number of steps). Due to the small amount of unique visited nodes for the 10 and 20 step walks, the truncated final node (the *k*th unique visited node after the initial starting node) was set to three for the 10-step walk, four for the 20-step walk and five for 30-step and onwards. Overall, walks over the HSC network resulted in a smaller similarity score between the initial and final truncated node compared to walks on the LSC network (Wilcoxon rank test significance at $p < 0.01$ for all steps; Figure 3). The reason that the relation between step size and similarity of initial and final node became unreliable in Figure 2 after 110 steps may be due to the HSC random walk traversing most of the network. We plan to investigate this in future work by examining whether the HSC random walk reaches stationarity before the LSC random walk.

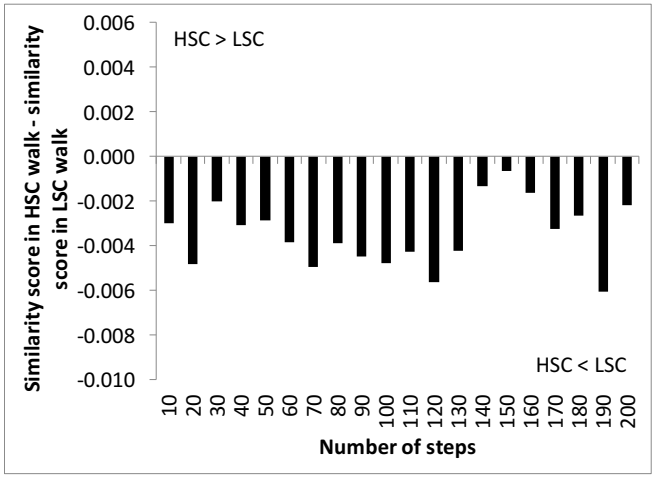


Figure 3: Difference score of the average similarity (shortest path) between initial and final truncated visited nodes by walks over the LSC and HSC networks varying the number of steps.

Discussion

In this paper, we examined how differences in semantic memory structure and search processes interact and relate to individual differences in creative ability. According to the associative theory of creativity, individual differences in creativity should be related to differences in semantic memory structure, which in turn facilitate cognitive search (Kenett et al., 2014; Mednick, 1962). To test this theory, we conducted random walk simulations and examined differences in how uncontrolled search behaves over semantic memory for HSC and LSC individuals.

Our work is situated with a growing amount of research studying uncontrolled thought processes (Christoff, 2013), and complements existing research by examining the search processes of LSC and HSC individuals using computational methods. We found that a random walk process visits more unique nodes and that the similarity strength between initial and final nodes visited by the walk is weaker for walks over the HSC network than the LSC network. These results were robust to the starting node and the number of steps taken by the walk. Thus, individual differences in thought processes between LSC and HSC individuals can be produced by an uncontrolled search process on differing semantic networks, providing support for Mednick's theory.

Notably, both networks have the same nodes, amount of edges, and average degree (number of edges per node). Thus, the differences between the random walks cannot be due to the HSC individuals merely having more connections in their semantic network. Rather, this reflects that HSC individuals have connections that enable a random walk to move quickly through the network. It is important to note that this demonstrates *sufficiency,* but not *necessity*: It does not rule out that different processes are being used by LSC and HSC individuals. Regardless, by showing how differences in representation can produce differences in search behavior, these findings support the associative theory of creativity, which posits that HSC individuals have a semantic memory structure that facilitates novel combinations (Kenett et al., 2014; Mednick, 1962).

There are limitations to this study. First, we treated creativity as a binary construct due to technical constraints (existing semantic network estimation techniques require large data sets). However, creativity is a continuous construct and not a categorical one. Future research should examine the search processes over semantic networks in individuals (see Zemla, Kenett, Jun, & Austerweil, 2016) to provide a better understanding of how search processes relate to creative ability. Second, we only explored undirected, unweighted random walks. Future research should also examine the effect of using directed and weighted random walks on the different semantic networks. Finally, the semantic networks of LSC and HSC individuals were constructed based on free associations. Some researchers have argued that explaining cognitive search behavior as a random walk over a semantic network estimated from free association data is potentially circular (Hills, Todd, & Jones, 2015). However, in our research, creative ability was measured independently with different tasks (divergent thinking) than those used for representing their semantic networks (free associations). Thus, our results are limited to the measures we used to classify participants into LSC and HSC individuals, as well as the way we constructed the semantic networks and measured the random walks. Future research should use different measures for estimating creativity, and the semantic networks of low and high creative individuals.

More generally, this work demonstrates how random walk models can be used to examine thought processes in different populations, such as LSC and HSC individuals. Random walks are a computational implementation of the theoretical concept of spreading activation. This enables researchers to use these models to investigate how well uncontrolled search processes can capture different cognitive processes operating on semantic memory, such as mind wandering or creative thought (Christoff, 2013). Specifically, such models can be used to understand the contribution of both bottom-up and top-down processes in creativity. We do not expect a completely uncontrolled search process (such as the naïve random walk) to explain creativity in all of its complexity. Rather, we intend to use it as a yardstick for determining whether controlled search processes are necessary or if randomness is sufficient to produce the observed variations in behavior (similar to studies of drift in computational models of evolutionary genetics; Reali & Griffiths, 2009). In future work, we also plan to incorporate controlled search processes into the model and see what sort of behavior it can produce that cannot be replicated by an uncontrolled search process (keeping the semantic networks constant). This will guide the development of experiments that provide stronger results capable of dissociating bottom-up and top-down processes contributing to creativity.

Acknowledgments

Data was collected as part of the PhD research of the first author in Bar-Ilan University, Israel.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review, 122*(3), 558-569.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science, 23*(10), 1117-1122.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in Cognitive Science. *Trends in Cognitive Sciences, 17*(7), 348-360.

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative over time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity and the Arts, 6*(4), 309-319.

Benedek, M., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014). Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence, 46*, 73-83.

Bourgin, D. D., Abbott, J. T., Griffiths, T. L., Smith, K. A., & Vul, E. (2014). *Empirical evidence for markov chain monte carlo in memory search.* Paper presented at the Proceedings of the 36th Annual Conference of the Cognitive Science Society, Boston, MA.

Capitán, J. A., Borge-Holthoefer, J., Gómez, S., Martinez-Romo, J., Araujo, L., Cuesta, J. A., & Arenas, A. (2012). Local-Based Semantic Navigation on a Networked Representation of Information. *PLoS ONE, 7*(8), e43694.

Christoff, K. (2013). Thinking. In K. N. Ochsner & S. M. Kosslyn (Eds.), *The Oxford Handbook of Cognitive Neuroscience* (Vol. 2, pp. 318-333). Oxford: Oxford University Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Faust, M. (2012). Thinking outside the left box: The role of the right hemisphere in novel metaphor comprehension. In M. Faust (Ed.), *Advances in the Neural Substrates of Language: Toward a Synthesis of Basic Science and Clinical Research* (pp. 425-448). Malden, MA: Wiley Blackwell.

Galimberti, G., & Soffritti, G. (2011). Tree-based methods and decision trees. In R. S. Kenett & S. Salini (Eds.), *Modern Analysis of Customer Surveys* (pp. 283-307). Chichester, UK: John Wiley & Sons, Ltd.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science, 18*(12), 1069-1076.

Henik, A., & Kaplan, L. (2005). *Content of categories: Findings regarding categories in Hebrew and comparison findings in the USA*. Ben-Gurion University.

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in Semantic Fields: How We Search Through Memory. *Topics in Cognitive Science, 7*(3), 513-534.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. Busemeyer & J. Townsend (Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232-254).

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*(407), 407.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. *PLoS ONE, 6*(8), e23912.

Mednick, M., T., Mednick, S., A., & Jung, C., C. (1964). Continual association as a function of level of creativity and type of verbal stimulus. *Journal of Abnormal and Social Psychology, 69*(5), 511-515.

Mednick, S., A. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232.

Milgram, R. N., & Milgram, N. A. (1976). Creative thinking and creative performance in Israel students. *Journal of Educational Psychology, 68*(3), 255-259.

Nevo, B., & Levin, I. (1978). Remote associates test: Assessment of creativity in Hebrew. *Megamot, 24*, 87-98.

Reali, F., & Griffiths, T. L. (2009). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society of London B: Biological Sciences*.

Rossman, E., & Fink, A. (2010). Do creative people use shorter association pathways? *Personality and Individual Differences, 49*, 891-895.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66-75.

Sawyer, K. (2011). The cognitive neuroscience of creativity: A critical review. *Creativity Research Journal, 23*(2), 137-154.

Schilling, M. A. (2005). A "small-world" network model of cognitive insight. *Creativity Research Journal, 17*(2-3), 131-154.

Smith, K. A., & Vul, E. (2015). The Role of Sequential Dependence in Creative Semantic Search. *Topics in Cognitive Science, 7*(3), 543-546.

Zemla, J. C., Kenett, Y. N., Jun, K.-S., & Austerweil, J. L. (2016). U-INVITE: Estimating Individual Semantic Networks from Fluency Data, *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Determining the alternatives for scalar implicature

Benjamin Peloquin
bpeloqui@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Successful communication regularly requires listeners to make pragmatic inferences — enrichments beyond the literal meaning of a speaker's utterance. For example, when interpreting a sentence such as "Alice ate some of the cookies," listeners routinely infer that Alice did not eat all of them. A Gricean account of this phenomenon assumes the presence of alternatives (like "all of the cookies") with varying degrees of informativity, but it remains an open question precisely what these alternatives are. To address this question, we collect empirical measurements of speaker and listener judgments about varying sets of alternatives across a range of scales and use these as inputs to a computational model of pragmatic inference. This approach allows us to test hypotheses about how well different sets of alternatives predict pragmatic judgments by people. Our findings suggest that comprehenders likely consider a broader set of alternatives beyond those logically entailed by the initial message.

Keywords: pragmatics; scalar implicature; bayesian modeling

Introduction

How much of what we mean comes from the words that go unsaid? As listeners, our ability to make precise inferences about a speaker's intended meaning in context is indispensable to successful communication. For example, listeners commonly enrich the meaning of the word *some* to *some but not all* in sentences like "Alice ate some of the cookies" (Grice, 1975; Levinson, 2000). These inferences, called *scalar implicatures*, have been an important test case for understanding pragmatic inferences more generally. A Gricean account of this phenomenon assumes listeners reason about a speaker's intended meaning by incorporating knowledge about A) alternative scalar items a speaker could have used (such as *all*) and B) the relative informativity of using such alternatives (Grice, 1975). According to this account, a listener will infer that the speaker intended to convey that Alice did not eat all the cookies because it would have been underinformative to use the descriptor *some* when the alternative *all* could have been used just as easily.

But what are the alternatives that should be considered in the implicature computation more generally? Under classic accounts, listeners consider only those words whose meanings entail the actual message (Horn, 1972), and these alternatives enter into conventionalized or semi-conventionalized scales (Levinson, 2000). For example, because *all* entails *some*, and hence is a "stronger" meaning, *all* should be considered as an alternative to *some* in implicatures. Similar scales exist for non-quantifier scales, e.g. *loved* entails *liked* (hence "I liked the movie" implicates that I didn't love it).

Recent empirical evidence has called into question whether entailment scales are all that is necessary for understanding scalar implicature. For example, Degen & Tanenhaus (2015) demonstrated that the scalar item *some* was judged less appropriate when exact numbers were seen as viable alternatives. And in a different paradigm, van Tiel (2014) found converging evidence that *some* was judged to be atypical for small quantities. These data provide indirect evidence that people may actually consider a broader set of alternatives when computing scalar implicatures. Since *some* is logically true of sets with one or two members, these authors argued that the presence of salient alternatives (the words *one* and *two*, for example) reduced the felicity of *some* via a pragmatic inference.

By formalizing pragmatic reasoning, computational models can help provide more direct evidence about the role that alternatives play. The "rational speech act" model (RSA) is one recent framework for understanding inferences about meaning in context (Frank & Goodman, 2012; Goodman & Stuhlmuller, 2013). RSA models frame language understanding as a special case of social cognition, in which listeners and speakers reason recursively about one another's goals. In the case of scalar implicature, a listener makes a probabilistic inference about what the speaker's most likely communicative goal was, given that she picked the quantifier *some* rather than the stronger quantifier *all*. In turn, the speaker reasons about what message would best convey her intended meaning to the listener, given that he is reasoning in this way. This recursion is grounded in a "literal" listener who reasons only according to the basic truth-functional semantics of the language.

Franke (2014) used an RSA-style model to assess what alternatives a speaker would need to consider in order to produce the typicality/felicity ratings reported by Degen & Tanenhaus (2015) and van Tiel (2014) for the scale *some*/*all*. In order to do this, Franke (2014)'s model assigned weights to a set of alternative numerical expressions. Surprisingly, along with weighting *one* highly (a conclusion that was supported by the empirical work), the best-fitting model assigned substantial weight to *none* as an alternative to *some*. This finding was especially surprising considering the emphasis of standard theories on scalar items that stand in entailment relationships with one another (e.g. *one* entails *some* even if it is not classically considered to be part of the scale).

We pick up where these previous studies left off, considering the set of alternatives for implicature using the RSA model. To gain empirical traction on this issue, we broaden the set of scales we consider. Our inspiration for this move comes from work by van Tiel, Van Miltenburg, Zevakhina, & Geurts (2014), who examined a phenomenon that they dubbed "scalar diversity," namely the substantial difference in the strength of scalar implicature across a variety of scalar

pairs (e.g., *liked* / *loved*, *palatable* / *delicious*). Making use of this diversity allows us to investigate the ways that different alternative sets give rise to implicatures of different strengths across scales.

In our current work, we use a computational framework to investigate the set of alternatives that best allow the model to predict human pragmatic judgments across a series of different scales. We begin by presenting the computational framework (RSA) we use throughout the paper. We next describe a series of experiments designed to measure both the literal semantics of a set of scalar items (used as input to the model) and comprehenders' pragmatic judgments for these same items (used for model comparison). These experiments allow us to compare the effects of different alternative sets on our ability to model listeners' pragmatic judgments. To preview our results: we find that standard entailment alternatives do not allow us to fit participants' judgments, but that expanding the range of alternatives empirically (by asking participants to generate alternative messages) allows us to predict listeners' pragmatic judgments with high accuracy.

Modeling Implicature Using RSA

We begin by giving a brief presentation of the basic RSA model. This model simulates the judgments of a pragmatic listener who wants to infer a speaker's intended meaning m from her utterance u. For simplicity, we present a version of this model in which there is only one full recursion: The pragmatic listener reasons about a pragmatic speaker, who in turn reasons about a "literal listener." We assume throughout that this computation takes place in a signaling game (Lewis, 1969) with a fixed set of possible meanings $m \in M$ and a fixed possible set of utterances $u \in U$, with both known to both participants. Our goal in this study is to determine which utterances fall in U.

In the standard RSA model, the pragmatic listener (denoted L_1), makes a Bayesian inference:

$$p_{L_1}(m \mid u) \propto p_{S_1}(u \mid m)p(m) \tag{1}$$

In other words, the probability of a particular meaning given an utterance is proportional to the speaker's probability of using that particular utterance to express that meaning, weighted by a prior over meanings. This prior represents the listener's *a priori* expectations about plausible meanings, independent of the utterance. Because our experiments take place in a context in which listeners should have very little expectation about which meanings speakers want to convey, for simplicity we assume a uniform prior where $p(m) \propto 1$.

The pragmatic speaker in turn considers the probability that a literal listener would interpret her utterance correctly:

$$p_{S_1}(u \mid m) \propto p_{L_0}(m \mid u) \tag{2}$$

where L_0 refers to a listener who only considers the truth-functional semantics of the utterance (that is, which meanings the utterance can refer to).

Alternative sets				
good / excellent	liked / loved	memorable / unforgettable	palatable / delicious	some / all
excellent	loved	unforgettable	delicious	all
good	liked	memorable	palatable	most
okay	felt indifferent about	ordinary	mediocre	some
bad	disliked	bland	gross	little
horrible	hated	forgettable	disgusting	none

Entailment items used in two alternatives condition
Top two alternatives added in four alternatives condition
Neutral item added in five alternatives condition

Figure 1: Alternative sets used in Experiments 2 and 3 (generated in Experiment 1). Green coloring denotes classic entailment items pulled from van Tiel et al. (2014), used in the two alternatives condition. Blue coloring denotes top-two alternatives added to entailment scales in the four alternatives condition. Red coloring denotes "neutral" items added in the five alternatives condition.

This model of the pragmatic speaker (denoted S_1) is consistent with a speaker who chooses words to maximize the utility of an utterance in context (Frank & Goodman, 2012), where utility is operationalized as the informativity of a particular utterance (surprisal) minus a cost:

$$p_{S_1}(u \mid m) \propto e^{-\alpha[-log(p_{L_0}(m|u))-C(u)]} \tag{3}$$

where $C(u)$ is the cost of a particular utterance, $-log(p_{L_0})$ represents the *surprisal* of the message for the literal listener (the information content of the utterance), and α is a parameter in a standard choice rule. If $\alpha = 0$, speakers choose randomly; as $\alpha \to \infty$, they greedily choose the highest probability alternative. In our simulations below, we treat α as a free parameter and fit it to the data.

To instantiate our signaling game with a tractable message set M, in our studies we adopt the world of restaurant reviews as our communication game. We assume that speakers and listeners are trying to communicate the number of stars in an online restaurant review (where $m \in \{1...5\}$). We then use experiments to measure three components of the model. First, to generate a set of plausible alternative messages in U, we ask participants to generate alternatives for particular scalar items (Experiment 1). Next, to measure literal semantics $p_{L_0}(m \mid u)$ we ask participants to judge whether a message is compatible with a particular meaning (Experiment 2). Lastly, to obtain human L_1 pragmatic judgments, we ask participants to interpret a speaker's utterances (Experiment 3).

Experiment 1: Alternative Elicitation

To examine the effect different alternative sets have on implicature, we needed a way of expanding alternative sets beyond

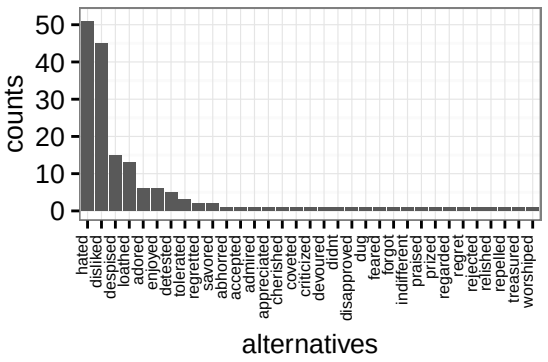


Figure 2: Combined counts for participant-generated alternatives for liked and loved in Experiment 1.

entailment pairs. We addressed this issue by adopting a modified cloze task to elicit alternatives empirically. This design was inspired by Experiment 3 of van Tiel et al. (2014).

Methods

Participants We recruited 30 workers on Amazon Mechanical Turk (AMT). All participants were native English speakers and naive to the purpose of the experiment.

Design and procedure On each trial participants were presented with a target scalar item from five scales, selected from the entailment items used in van Tiel et al. (2014). These were embedded in a sentence such as, "In a recent restaurant review someone said they thought the food was ____," with a target scalar presented in the blank. Participants were then asked to generate three plausible alternatives by responding to the question, "If they'd felt differently about the food, what other words could they have used instead of ____?" Participants saw each scalar item in random order. In total, each participant saw ten target scalar items (each entailment pair from the five scales selected from van Tiel et al., 2014).

Results and Discussion

Figure 2 shows an example alternative set for the scalar items *liked* and *loved* (combined). Figure 1 shows the complete list of alternative sets derived from Experiment 1.

Experiment 2: Literal Listener Task

Experiment 2 was conducted to approximate literal listener semantics—$p_{L_0}(m \mid u)$ in Equation (3)—for the same five pairs of scalar items used in Experiment 1. We included three conditions: two alternatives ("entailment"), four alternatives, and five alternatives. The two alternatives condition makes a test of the hypothesis that the two members of the classic Horn (entailment) scale (Horn, 1972) are the only alternatives necessary to predict the strength of listeners' pragmatic inference. The four and five alternative conditions then add successively more alternatives to test whether including a larger number of alternatives will increase model fit.[1] A secondary goal of Experiment 2 is to test whether the set of alternatives queried during literal semantic elicitation impacts literal semantic judgments. If it does we should see differences in these judgments for shared items between conditions.

Methods

Participants Conditions were run sequentially. In each condition we recruited 30 participants from AMT. In the two alternatives condition, 16 participants were excluded for either failing to pass two training trials or because they were not native English speakers, leaving a total sample of 14 participants.[2] In the four alternatives condition, 7 participants were excluded for either failing to pass two training trials or not being native English speakers, leaving a total sample of 23 participants. In the five alternatives condition, 3 participants were excluded for either failing to pass two training trials or not being native English speakers, leaving a total sample of 27 participants.

Design and procedure Figure 3, left, shows the experimental setup. Participants were presented with a target scalar item and a star rating (1–5 stars) and asked to judge the compatibility of the scalar item and star rating. Compatibility was assessed through a binary "yes/no" response to a question of the form, "Do you think that the person thought the food was ____?" where a target scalar was presented in the blank. Each participant saw all scalar item and star rating combinations for their particular condition, in a random order.

The two alternatives condition only included entailment pairs taken from van Tiel et al. (2014) for a total of 50 trials for each participant. The four alternatives condition included entailment pairs plus the top two alternatives generated for each scalar family by participants in Experiment 1 (100 trials per participant). The five alternatives condition included the four previous items plus one more neutral item chosen from alternatives generated in Experiment 1 (125 trials per participant).

Results and Discussion

Figure 4 plots estimated literal listener semantics for the three conditions. Each row shows a unique scalar family with items ordered horizontally by valence. Several trends are visible. First, in each scale, the alternatives spanned the star scale — there were scalar items that were highly compatible with both the lowest and highest numbers of stars. Second, participant compatibility judgments match intuitive literal semantics. That is, both weaker (e.g. *some* or *good*) and stronger (e.g. *all* or *excellent*) entailment items were seen as compatible with five-star ratings. This means participants' compatibility judgments reflect literal semantic intuitions, not

[1] Note that alternatives in the four and five alternative conditions were chosen on the basis of Experiment 1, which was run chronologically after the two alternative condition; all literal listener experiments are grouped together for simplicity in reporting.

[2] The majority of respondent data excluded from the two alternatives condition was caused by failure to pass training trials. We believe the task may have been too difficult for most respondents who may have been confused by the question wording. Training trials in later conditions included clearer language.



Figure 3: (Left) A trial from Experiment 2 (literal listener) with the target scalar 'liked'. (Right) A trial from Experiment 3 (pragmatic listener) with the target scalar 'liked'.

enriched pragmatic judgments. We see clear variability between scalar families (literal semantic judgments for *memorable* are considerably different from *palatable*), but also substantial consistency for individual scalar items across conditions (*good* in the two alternatives condition is similar to *good* in the four and five alternative conditions).

To examine this last issue of consistency across conditions (our secondary hypothesis) we fit a mixed effects model, regressing compatibility judgment on scale, number of stars and experimental condition, with subject- and word-level random effects, which was the maximal structure that converged. Results indicate no significant differences between two and five alternative conditions ($\beta = -0.05$, $z = -0.53$, $p = 0.59$) or between four and five alternative conditions ($\beta = -0.04$, $z = -0.52$, $p = 0.6$). The addition of condition as a predictor did not significantly improve model fit when compared to a model without the condition variable using ANOVA ($\chi^2(2) = 0.43$, $p = 0.81$). These findings suggest that literal semantic intuitions are not affected by the set of alternatives queried in our paradigm. This finding is important because literal semantic data generated in these three conditions are used as input to our model to simulate the effects of different alternative sets on implicature generation.

Experiment 3: Pragmatic Listener Task

Experiment 3 was conducted to measure pragmatic judgments. As in Experiment 2, we include several conditions to test inferences in the presence of different alternative sets. In the two alternatives condition, participants made judgments for items included in the entailment scales. In the four alternatives condition, participants made judgments for the entailment items and also the top two alternatives elicited for each scale in Experiment 1. Including two conditions with differing alternatives allowed us to rule out the potential effects of having a larger set of alternatives during the pragmatic judgment elicitation and also provided two sets of human judgments to compare with model predictions.

Methods

Participants We recruited 100 participants from AMT, 50 for each condition. Data for 9 participants was excluded from the two alternatives condition after participants either failed to pass two training trials or were non-native English speakers, leaving a total sample of 41 participants. In the four alternatives condition, data from 7 participants was excluded after participants either failed to pass two training trials or were not native English speakers, leaving 43 participants.

Design and Procedure Participants were presented with a one-sentence prompt such as "Someone said they thought the food was _____." with a target scalar item in the blank. Participants were then asked to generate a star rating representing the rating they thought the reviewer likely gave. Each participant was presented with all scalar items in a random order. Participants in the two alternatives condition gave a total of 10 pragmatic judgments. Participants in the four alternatives condition gave a total of 20 pragmatic judgments. The experimental setup is shown in Figure 3, right.

Results and Discussion

Figure 5 plots pragmatic listener judgment distributions for "weak" / "strong" scalar pairs (e.g. *good* / *excellent*). Several trends are visible. First, in each scale participants generated implicatures. They were substantially less likely to assign high star ratings to weaker scalar terms, despite the literal semantic compatibility of those terms with those states shown in Experiment 2. Second, the difference between strong and weak scalar items varied considerably across scales, consistent with previous work (van Tiel et al., 2014).

To rule out the potential effects of having a larger set of alternatives during the pragmatic judgment elicitation, we fit a mixed effects model. We regressed pragmatic judgments on scale and experimental condition with subject- and word-level random effects, which was the maximal structure that converged. There were no significant differences between the two alternative and four alternative conditions ($\beta = 0.05$, $t(150) = 1.04$, $p = 0.3$). The addition of the condition predictor did not significantly improve model fit when compared to a model without that variable ($\chi^2(1) = 1.13$, $p = 0.29$).

Model Simulations

Using literal listener data from Experiment 2, we conducted a set of simulations with the RSA model. Each simulation kept the model constant, fitting the choice parameter α as a free parameter, but used a set of alternatives to specify the scale over which predictions were computed. We considered four

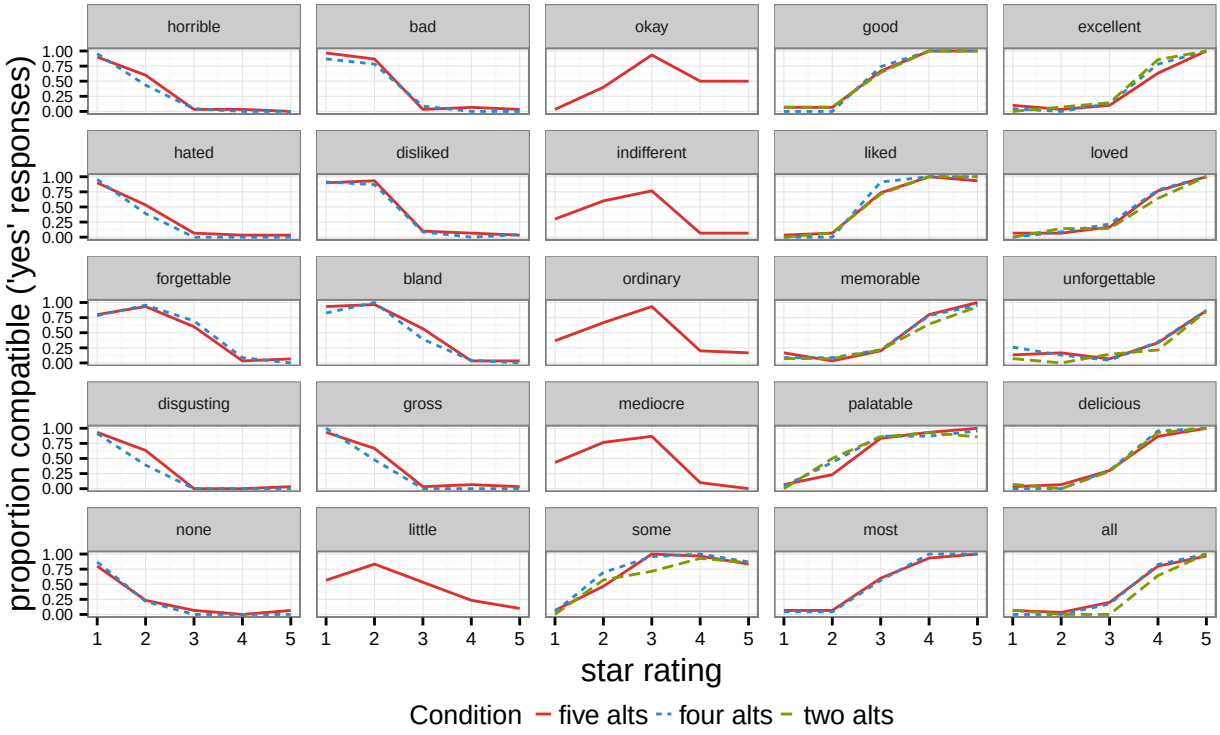


Figure 4: Literal listener judgments from Experiment 2. Proportion of participants indicating compatibility (answering "yes") is shown on the vertical axis, with the horizontal axis showing number of stars on which the utterance was judged. Rows are grouped by scale and items within rows are ordered by valence. Colors indicate the specific condition with conditions including different numbers of items.

different alternative sets, with empirical measurements corresponding to those shown in Figure 1: 1) the two alternatives in the classic entailment scales, 2) those two alternatives with the addition of a generic negative alternative, 3) the expanded set of four alternatives, and 4) the expanded set of five alternatives. Literal semantics for the generic negative alternative served as a baseline "none" semantics in which the scalar item was only compatible with 1 star.

Model fit with human judgments was significantly improved by the inclusion of alternatives beyond the entailment items (Table 1). The two alternatives model contained only entailment items, which, under classic accounts, should be sufficient to generate implicature, but fit to data was quite poor with these items. The addition of a generic negative element produced some gains, but much higher performance was found when we included four and five alternatives, with the alternatives derived empirically for the specific scale we used. Figure 6 plots model fit for the five alternatives model.

model	Empirical 2-alts	Empirical 4-alts
5-alts	r = 0.88 (alpha = 4.4)	r = 0.91 (alpha = 4.5)
4-alts	r = 0.85 (alpha = 4.4)	r = 0.89 (alpha = 4.7)
3-alts	r = 0.68 (alpha = 6.8)	r = 0.71 (alpha = 6.7)
2-alts	r = 0.54 (alpha = 8.9)	r = 0.56 (alpha = 9.4)

Table 1: Model performance with fitted alpha levels. Model fit assessed through correlation with human judgments in the two conditions of Experiment 3.

General Discussion

Pragmatic inference requires reasoning about alternatives. The fundamental pragmatic computation is counterfactual: "if she had meant X, she would have said Y, but she didn't." Yet the nature of these alternatives has been controversial. For a few well-studied scales, a small set of logically-determined alternatives has been claimed to be all that is necessary (Horn, 1972). For other, contextually-determined inferences, the issue of alternatives has been considered relatively intractable in terms of formal inquiry (Sperber & Wilson, 1995).

In our current work, we used the rational speech act framework to investigate the set of alternatives that best allowed the model to predict pragmatic judgments across a series of different scales. We found that the best predictions came when a range of scale-dependent negative and neutral alternatives were added to the implicature computation, suggesting the importance of considering non-entailment alternatives. This work builds on previous investigations, reinforcing the claim that negative alternatives are critical for understanding implicature (Franke, 2014), and replicates and extends findings that different lexical scales produce strikingly different patterns of inference (van Tiel et al., 2014).

While improvements in model fit were substantial as we

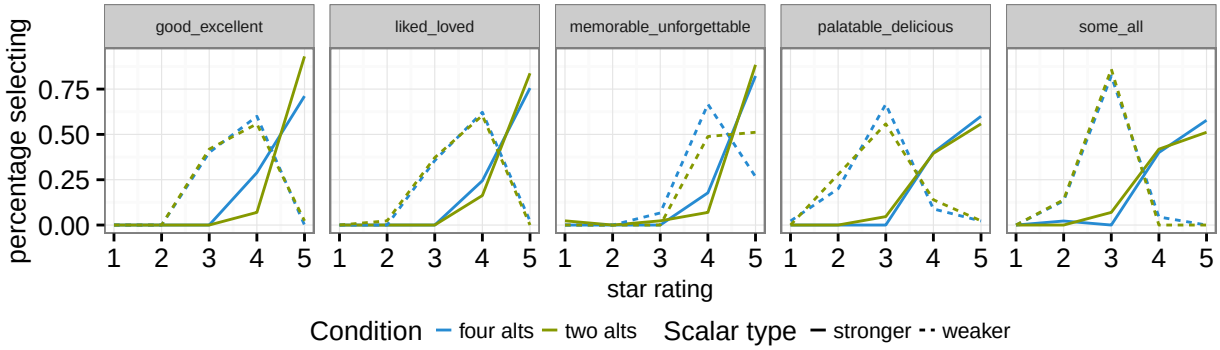


Figure 5: Pragmatic listener judgments from Experiment 3. Vertical axis shows proportion of participants generating a star rating. Horizontal axis shows number of stars on which the utterance was judged. Line type denotes condition and colors indicate the particular scalar items. Each panel shows one entailment scalar pair.

moved from two to four alternatives, we saw only a minor increase in fit from the move to five alternatives. One possible explanation is that alternatives are differentially salient in context, and in moving to larger sets we should consider weighting the alternatives differentially (as Franke, 2014 did). Preliminary simulations using weightings derived from Experiment 1 provide some support for this idea but would require further empirical work for confirmation.

The precise set of alternatives present during implicature is likely to be domain dependent. Our empirical paradigm elicited literal semantics, pragmatic judgments, and plausible alternatives all within the restricted domain of restaurant reviews. Our measurements might have differed substantially if we had instead grounded our ratings in a different context. Future investigations should probe the context-specificity of the weight and availability of particular alternative sets.

More broadly, considering the context- and domain-specificity of alternative sets may provide a way to unite what Grice (1975) called "generalized" (cross-context) and "particularized" (context-dependent) implicatures. Rather than being grounded in a firm distinction, we may find that these categories are simply a reflection of the effects of context on a constantly-shifting set of pragmatic alternatives.

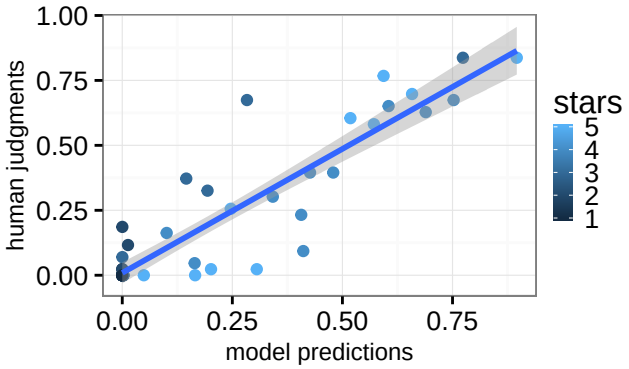


Figure 6: Judgments for entailment items from the four alternatives condition of Experiment 3 are plotted against model predictions using the five alternatives data from Experiment 2. Colors show the star rating for individual judgments.

Acknowledgements

Thanks to NSF BCS #1456077 for support, and thanks to Michael Franke, Judith Degen, and Noah Goodman for valuable discussion.

References

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

Frank, M., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 487–492).

Goodman, N. D., & Stuhlmuller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3). New York: Academic Press.

Horn, L. R. (1972). *On the semantic properties of logical operators.* (PhD thesis). University of California, Los Angeles.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Lewis, D. (1969). *Convention: A philosophical study*. John Wiley & Sons.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, UK: Blackwell.

van Tiel, B. J. M. (2014). Quantity matters: Implicatures, typicality, and truth.

van Tiel, B. J. M., Van Miltenburg, E., Zevakhina, N., & Geurts, B. (2014). Scalar diversity. *Journal of Semantics*, ffu017.

The Role of Similarity in Constructive Memory: Evidence from Tasks with Children and Adults

Georgi Petkov (gpetkov@nbu.cogs.bg)
Margarita Pavlova (margarita.velinova@gmail.com)

Department of Cognitive Science and Psychology,
New Bulgarian University,
21 Montevideo Street, Sofia 1618, Bulgaria

Abstract

Literature on memory research shows that when memorizing, people may blend two situations, i.e. when memorizing one story, they add elements from another story. Most of the cognitive models assume that the superficial similarity between two episodes is the primary factor for blending. However, there is evidence that people blend dissimilar stories as well, if these stories share the same relational structure. We contrasted the two factors in a single study and performed experiments with the same design and stimuli with adults and with 4-5-year-old children. The results show that there is no qualitative difference between the performance of adults and children. Also, both adults and children blend either pictures that have surface or structural similarity depending on the abstractness of the objects in them.

Keywords: constructive memory; development; analogy-making.

Evidence for the Constructive Nature of Memory

The constructivist approach became a leading theory for human memory during the last decades. According to it, people continuously create hypotheses about what has happened in the past and verify them according to their current knowledge and the current context (Neisser, 1967). The subjective experience of recalling is the final product of this process of continuous creation and verification of hypotheses. There is a huge amount of empirical evidence that suggests that people wrongly recall stimuli and events that never have occurred (Bartlett & Burt, 1933, Roediger, 1996, Roediger & McDermott, 1995, Schacter, 2002); blend episodes and misattribute the source of their memories (Loftus, 1997, 2003, Loftus & Ketcham, 1992, Berkowitz et al., 2008, Kokinov & Zareva, 2001).

However, the mechanisms of memory distortion are questionable. Most of the models of memory blending (Metcalfe, 1990; McClelland, 1995; Schacter et al, 1998; Nystrom & McClelland, 1992) assume that the surface similarity between two episodes is the only factor for their blending. According to these models, the more similar two episodes are, the larger the overlapped part of their representation is and hence, the probability to blend them in the course of retrieval is higher.

However, Kokinov (1998) reported a case of blending between episodes that do not share common features but share a relational similarity. Even completely dissimilar episodes could be blended, if they participate in a double analogy, i.e. the target partially maps to these two dissimilar episodes (Kokinov & Zareva, 2001, Zareva & Kokinov, 2003). Pavlova & Kokinov (2014) demonstrated that people tend to blend analogical pictures in a higher extend compared to superficially similar ones.

The Role of Relational Knowledge in Human Cognition

There is a lot of evidence that relational similarity is probably a hallmark of various constructive processes. Relational mapping is the fundament of various models of vision and recognition (Hummel & Stankiewicz, 1998; Chalmers et al., 1992; Petkov & Shahbazyan, 2007). People's understanding of social relationships and their theories of other's minds depend highly on their relational knowledge (Day and Gentner, 2007; Shahbazyan et al., 2014; Wittenbrink et al., 1997). Even emotions are probably constructs that are based on relational coherence of various signals (Russell, 2003; Gallagher, 2012, de Bruin et al., 2014).

Analogy-making may lie at the core of human intelligence and most of the other cognitive processes (Penn & Povinelli, 2012, Hofstadter, 2001). This point of research, together with the evidence for the role of relational structure of episodes and situations in memory distortions of them, naturally implies the hypothesis that the process of relational mapping is a basic factor for the process of construction of human memories. However, there is evidence that analogy-making ability is not innate, something happens during the 4-5th year of development, the so-called "relational shift" (Gentner, 1988). Having in mind that analogy-making develops during childhood, the question whether memory illusions develop in parallel seems interesting and promising.

Development of Analogy-Making in Children

A lot of experimental data during the last decades overturn the Structure theory of Piaget (Piaget, 1977) according to which children need to develop abstract operations in order to be able to make analogies and this cannot happen before 12 years of age. On the contrary, it seems that even 4-5-

year-old children can make analogies as long as they have the conceptual knowledge required by the task (Goswami & Brown, 1989; Goswami, 1991; Vosniadou, 1989). However, Gentner (1988) disagrees that accumulation of abstract knowledge is a sufficient prerequisite for complete, rich analogy-making. Instead, something in the basic cognitive mechanisms must change; the so-called *relational shift* should occur, i.e. the change of strategy children use to judge similarity - they first use shared objects to judge it before they use shared relational structure (Gentner, 1988, Rattermann & Gentner, 1998, Gentner & Toupin, 1986). Halford et al. (1998) agree that something dramatically changes during the 4-5 year of child development but assume it is just a quantitative increase of working memory capacity. Bulloch & Opfer (2009) also oppose to the necessity to generalize over types of similarity and assume that what develops with age is the subjective judgment of the predictive validity of different types of similarity. In other words, when reasoning, both adults and children are able to use relational as well as superficial similarity but sometimes prefer different strategies, depending on which similarity is relevant to the task.

Developmental changes reflect many cognitive abilities simultaneously: children's ability to manipulate relations increases; they master high language abilities; they acquire Theory of Mind, etc. The question what kind of memory distortions occur at this age and what type of situations children tend to blend is undervalued in the research and deserves higher regard.

Experiments

We conducted a comparison study using similar tasks and stimuli for testing adults and 4-5-year-old children. We asked the participants to memorize sets of four pictures. However, the sets were constructed in a special manner, so that two pairs of pictures were structurally similar to each other, another two pairs were superficially similar; and the last two pairs were dissimilar. After a short delay we performed a memory test about all pictures and measured the amount of blending, i.e. false recognition of pictures that were not presented but consist of combination of elements from the presented pictures. Finally, we compared the amount of blending between structurally similar, superficially similar, and dissimilar pictures.

In the first experiment we used "rich" pictures with a lot of salient elements that may attract attention, whereas in the second experiment we repeated the procedure with "poor" pictures with abstract geometric figures, and expected that both adults and children would pay more attention to the relational structure.

Experiment 1

The main aim of the experiment is to check whether there is a qualitative difference between blending in children and in adults. Specifically, will children blend the superficially similar items more often than analogical or dissimilar ones, whereas adults will blend the analogical items more often than superficially similar or dissimilar ones.

Design. The experiment had a 2×3 mixed design. The between-subjects variable was the age of the participants and had two levels: adults or children. The within-subjects variable was the type of distracter given during the memory test and it had three levels:

- **Analogous** – distracters designed by combining two bases that share the same relations, but do not share the same elements.
- **Superficially similar** – distracters designed by combining two bases that share the same elements and thus are superficially similar, but do not share the same structure, i.e. the relations between the elements in the two bases are different.
- **Dissimilar** – distracters designed by combining two bases that share neither the same structure, nor the same elements.

Procedure. The experiment was run on the E-prime software and consisted of four phases that immediately followed one after another.

First, the participants were presented sequentially with the four pictures (called "bases") from each set. Their task was to describe each picture.

Then, they saw the four pictures of each set again but this time simultaneously (see Figure 1). Adults had to observe them for one minute. Children observed the bases and received feedback from the experimenter about the objects and relations in each base. The same procedure was immediately repeated for the second set of four pictures, and then for the last, third set.

In the third phase the participants performed a filler task. Adults solved simple addition math problems for 5 minutes. Children solved a puzzle for 2 minutes. The purpose of this phase was to ensure a retention interval.

Finally, the participants performed a recognition test. They were presented sequentially with pictures and were asked to indicate for each one whether they had seen it in the first phase of the experiment. Pictures were either the same as the original ones (i.e. "bases"), or different, i.e. distracters, constructed by combining elements from two different bases. The recognition test included all the bases and half of the possible distracters (see Figure 2), presented one by one and fully randomized. After the vocal answer by the participants, the experimenter pressed a corresponding key on the computer keyboard. No feedback was given to either adults, or children.

Stimuli. Three sets of 16 pictures were developed following the procedure of Pavlova & Kokinov (2014) and used in phases one and three. In each set, 4 of the pictures were the bases, and the other 12 were the distracters. The 4 bases were specifically constructed so that they formed three pairs

as follows: two pairs were structurally analogous to one another (A1~A2 and A3~A4), two pairs were superficially similar to one another, but did not share the same structure (A1≈A3 and A2≈A4), and finally the last two pairs were dissimilar and shared neither the same structure, nor the same superficial elements (A1≠A4 and A2≠A3). An example is shown in Figure 1. The distracters are designed by combining elements from two bases and there were three types of distractors depending on these bases – analogous distracters, superficially similar distracters, and dissimilar distracters (see Figure 2).



Figure 1: An example of the bases. The bases compose the following pairs: analogous configurations (A1~A2, because in both the subject **holds** the object; and A3~A4, because in both the subject **looks at** the object), superficially similar configurations (A1≈A3, because in both there is a bear and a balloon and A2≈A4, because in both there is a rabbit and a kite), dissimilar configurations (A1≠A4 and A2≠A3).

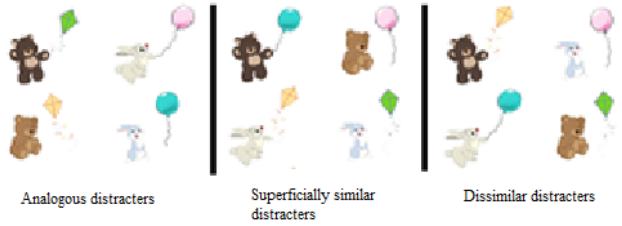


Figure 2: An example of the distracters designed for the bases in Figure 1. Each of the distracters combines elements from two of the bases.

Participants. The sample consisted of 32 participants – 12 adults (5 male; age 18 – 41 years, M = 26 years, SD = 7.5 years) and 20 children (10 male; age 48 – 63 months, M = 56 months, SD = 5.4 months). Children were recruited at a kindergarten in Sofia, Bulgaria; permission to participate was obtained from their parents prior to the study. Adults were students at New Bulgarian University and participated voluntarily.

Results and Discussion. The participants showed a very good performance on the recognition test with 88% correctly recognized bases for the adults and 93% for the children. Since we are interested in the source of blending, we analyzed only the answers for the distracters of each type. The data are presented in Figure 3.

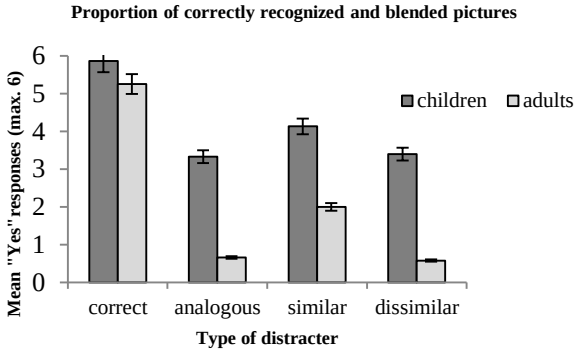


Figure 3: Distribution of "Yes" responses of adults and children among the various types of stimuli during the recognition test. The first column indicates the correct responses, whereas the next three columns indicate blending of various types.

Figure 3 illustrates the distribution of the "Yes" answers for all conditions, i.e. both the correct responses for the bases and the falsely recognized pictures for the three types of distracters. A repeated measures ANOVA showed that the main effect of type of stimulus was significant: $F(3,90) = 83.199$, $p < 0.001$, (partial $\eta^2 = 0.735$). The main effect of age was also significant: $F(1,30) = 12.707$, $p = 0.001$. Children are more likely to give a "Yes" answer irrelevant to what type the stimulus is. Analyzing adults' data for the distracters only, the pair-wise comparison showed that there was a significant difference between the superficially similar and the analogous distracters: $t(1,11) = 3.752$, $p = 0.003$, and between the superficially similar and the dissimilar distracters: $t(1,11) = 3.027$, $p = 0.012$. The same pattern was found for the children's data. There was a significant difference between the superficially similar and the analogous distracters: $t(1,19) = 2.904$, $p = 0.009$, and between the superficially similar and the dissimilar distracters: $t(1,19) = 2.774$, $p = 0.012$.

These results suggest that the responses of adults and children follow the same qualitative pattern: both adults and children falsely recognized more superficial distracters, which means that they blended the superficially similar bases more often compared to the analogous and dissimilar ones. Evidence in the literature suggests that object similarity plays an important role in similarity judgements for adults (Markman & Gentner, 1993) and children (Gentner & Namy, 1999).

However, the obtained results differ from those in Pavlova & Kokinov (2014) study. One possible explanation is that the stimuli used in the two experiments differ from each other. It is possible that the relations between the objects in this experiment are more subtle and thus harder to notice. Previous research shows that when stimuli are sparse both adults and children focus on relational similarity, whereas when stimuli are rich both adults and children focus on object similarity (Rattermann & Gentner, 1998;

Markman & Gentner, 1993; Bulloch & Opfer, 2009). Therefore, the more complex or rich the stimuli are, the more shared object similarities there are, the harder it is to notice the relations involved.

To test this assumption, a new experiment was conducted. Experiment 2 followed the same procedure as in Experiment 1, but with the stimuli used in Pavlova & Kokinov (2014). The main idea was to see whether by using more abstract stimuli we would obtain evidence for a qualitative difference between recognition responses in adults and children.

Experiment 2

Design and Procedure. Experiment 2 followed the same design and procedure as Experiment 1. The only difference was that we used the stimuli developed by Pavlova & Kokinov (2014) that consist of abstract geometrical relations instead of attractive pictures of toys.

Stimuli. Four sets (one more than in Experiment 1) of the original stimuli by Pavlova & Kokinov (2014) were used in phases one and three of Experiment 2. Each set consisted of 4 bases and 12 distracters. The bases formed three pairs: analogous, superficially similar, and dissimilar (Figure 4) and there were 3 types of distracters depending on the type of blending they represent: blending between analogous bases, blending between superficially similar bases, and blending between dissimilar bases (Figure 5).

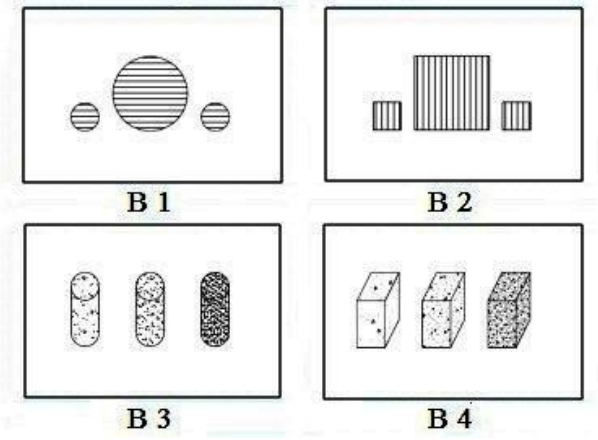


Figure 4: An example of the bases. The bases compose the following pairs: analogous configurations (B1~B2 and B3~B4), superficially similar configurations (B1≈B3 and B2≈B4), dissimilar configurations (B1≠B4 and B2≠B3).

Participants. The sample consisted of 30 participants – 11 adults (9 male; age 23 – 33 years, M = 27 years; SD = 3 years) and 19 children (10 male; age 51 – 71 months, M = 60 months; SD = 3.3 months). Children were recruited at a kindergarten in Sofia, Bulgaria; permission to participate was obtained from their parents prior to the study. Adults were students at New Bulgarian University and participated voluntarily.

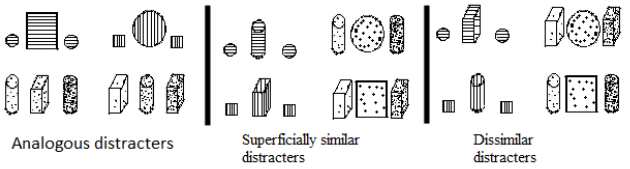


Figure 5: An example of the distracters designed for the bases in Figure 4. Each of the distracters combines elements from two of the bases.

Results and Discussion. The results showed a very good performance on the recognition test with 78% correctly recognized bases for the adults and 94% for the children. Since we are interested in the source of blending, we analyzed only the recognition for distracters of each type. The data are presented in Figure 6.

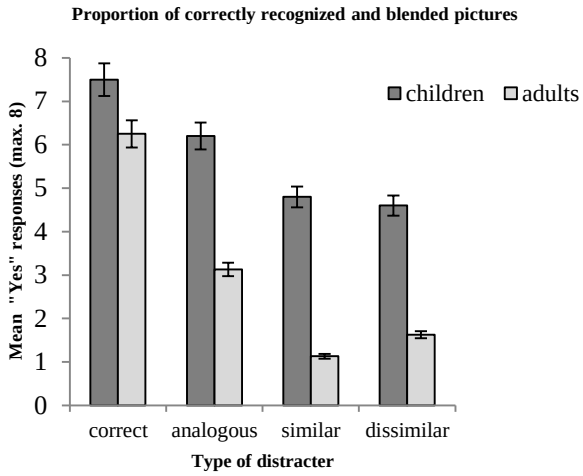


Figure 6: Distribution of "Yes" responses of adults and children among the various types of stimuli in during the recognition test in Experiment 2.

Figure 6 illustrates the distribution of both correct responses and falsely recognized pictures in the recognition task. A repeated measures ANOVA showed that the main effect of type of stimulus was significant: $F(3,84) = 53.352$, $p < 0.001$, (partial $\eta^2 = 0.656$). The main effect of age was also significant: $F(1,28) = 15.238$, $p = 0.001$, (partial $\eta^2 = 0.352$). Again, children are more likely to give a "Yes" answer irrelevant to the type of stimulus. Analyzing the adults' data for the distracters only, the pair-wise comparison showed that there was a significant difference between the analogous and the superficially similar distracters in favor of the analogous ones: $t(1,10) = 3.860$, $p = 0.003$, and between the analogous and the dissimilar distracters: $t(1,10) = 3.068$, $p = 0.012$. The same pattern was found for the children's data. There was a significant difference between the analogous and the superficially

similar distracters: $t(1,18) = 2.163$, $p = 0.044$, and between the analogous and the dissimilar distracters: $t(1,18) = 2.163$, $p = 0.012$. However, this time both adults and children more often gave a "Yes" answer for the analogous distracters, than for the superficially similar and the dissimilar ones.

Thus, again adults' and children's results share the same qualitative pattern but the results were exactly the opposite compared to the first experiment – this time the blending between analogous pictures was more frequent than between superficially similar ones.

General discussion

People blend similar pictures as well as superficially dissimilar pictures that share same relational structure. This result is in accordance with previous empirical evidence that blending cannot be fully explained by models that rely on superficial similarity alone. Most of the models of blending (Metcalfe, 1990; McClelland, 1995, Schacter et al, 1998, Nystrom & McClelland, 1992) assume that the overlap of the representations of the memories is the only factor for blending. Potentially, these models could account for the results; however, additional complications are necessary: the relational structure of the situations should be added to the representations, as well as mechanisms for weighting the different pieces of knowledge.

The most interesting result from the current experiments is that adults and children share the same qualitative pattern of results. It seems that there is no discrete difference between the cognitive mechanisms they use. These results do not support the relational shift hypothesis (Gentner, 1988, Rattermann & Gentner, 1998, Gentner & Toupin, 1986). Of course, one may assume that the children in our experiments have outgrown the stage of the relational shift (it is a complicated task to perform comparative study with same stimuli between adults and very young children). It should be mentioned that most of the evidence for the relational shift hypotheses emerge from explicit tasks, whereas memory blending is an implicit phenomenon. In addition, there is a small difference in the procedure in our experiments between adults and children – the latter received a feedback about the presented relations during the second phase of the experiment, when they observed the four pictures simultaneously. Additional research is required for a precise conclusion but the results from the second experiment definitely show that 4-5-year-old children are able to use relational structures in their representations. Moreover, relational structure may influence memory distortions.

During the recognition test children much more often say "Yes" compared to adults, no matter what the type of the stimuli is. This could be explained with a higher conformism, or trust to the experimenter, or searching for socially desirable answers in a higher extend.

Finally, it seems that the type of blending that occur more often depend not so much on the type of similarity (surface or structure) per se, but on the type of stimuli. The "rich", complex stimuli probably attract more attention to the concrete objects; whereas the "poor", abstract patterns make it easier for people to notice the relational structure. Of course, another possibility is that the accessibility of relations is better for the "poor" stimuli. The interesting fact is that both young children and adults seem to use the same cognitive mechanisms for this. Certainly, future work will shed more light on the mechanisms of memory construction and the role of different types of similarity in it.

Acknowledgements

This research was supported financially by the European Office for Aerospace Research and Development under grant FA9550-15-1-0510 (Anticipating Future by Analogy-Making). We would like to thank Milena Mutafchieva and Elena Andonova for the valuable discussions on an earlier draft of this manuscript.

References

Bartlett, F. C., & Burt, C. (1933). Remmembering: A study in experimental and social psychology. *British Journal of Educational Psychology, 3*(2), 187-192.

Berkowitz, S. R., Laney, C., Morris, E. K., Garry, M., & Loftus, E. F. (2008). Pluto behaving badly: False beliefs and their consequences. *American Journal of Psychology, 121*(4), 643-660.

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12*(1), 114-122.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation and analogy: A critique of artificial intelligence methodology. *Journal of Experimental and Theoretical Artificial Intelligence, 4*, 185-211.

Day, S. B., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory & Cognition, 35*(1), 39-49.

de Bruin, L., Strijbos, D., & Slors, M. (2014). Situating emotions: From embodied cognition to mindreading. *Topoi, 33*(1), 173-184.

Gallagher, S. (2012). Empathy, simulation, and narrative. *Science in Context, 25*(3), 355-381.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child development, 59*, 47-59.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development, 14*, 487-513.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277-300.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*, 1-22.

Goswami, U., & Brown, A. L. (1990). Melting chocolate and melting snowmen: Analogical reasoning and casual relations. *Cognition, 35*(1), 69-95.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 803-865.

Hofstadter, D. R. (2001). Epilogue: Analogy at the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from conitive science* (pp. 499-538). MIT Press.

Hummel, J. E., & Stankiewicz, B. J. (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition, 5*, 49-79.

Kokinov, B. (1998). Analogy is like cognition: dynamic, emergent, and context-sensitive. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences* (pp. 96-105). Sofia: NBU Press.

Kokinov, B., & Zareva, N. (2001). Episode blending as result of analogical problem solving. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.* Erlbaum, Hillsdale, NJ.

Loftus, E. F. (1997). Creating false memories. *Scientific American*(277), 70-75.

Loftus, E. F. (2003). Our changeable memories: Legal and practical implications. *Nature Reviews: Neoruscience, 4*, 231-234.

Loftus, E., & Ketcham, K. (1992). *Witness for the defense: The accused, the eyewitness and the expert who puts memory on trial.* New York: St. Martin's Griffin.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

McClelland, J. (1995). Constructive memory and memory distortions: A parallel-distributed processing approach. In D. Schacter (Ed.), *Memory distortions: How minds, brains, and societies reconstruct the past* (pp. 69-90). Cambridge, MA: Harvard University Press.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*(2), 145-160.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Nystrom, L., & McClelland, J. (1992). Trace synthesis in cued recall.*Journal of Memory and Language, 31*, 591-64.

Pavlova, M., & Kokinov, B. (2014). Analogy causes distorted memory by blending memory episodes. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Ed.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2729-2734). Austin, TX: Cognitive Science Society.

Penn, D. C., & Povinelli, D. J. (2012). The human enigma. In K. Holyoak, & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 439-452). Oxford University Press.

Petkov, G., & Shahbazyan, L. (2007). The RecMap model of active recognition based on analogical mapping. In R. Lewis, T. Polk, & J. Laird (Ed.), *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 293-298). Ann Arbor, Michigan.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures.* Oxford, England: Viking.

Rattermann, M. J., & Gentner, D. (1998). The effect of language on similarity: The use of relational labels improves young children's performance in a mapping task. In K. J. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data form the cognitive, computational, and neural sciences* (pp. 274-282). Sofia: New Bulgarian University.

Roediger, H. L. (1996). Memory illusions. *Journal of Memory and Language, 35*(2), 76-100.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memoeries: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145-172.

Schacter, D. L. (2002). *The seven sins of memory: How the mind forgets and remembers.* New York: Houghton Mifflin Harcourt.

Schacter, D. L., Norman, K. A., & Koutsaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*(1), 289-318.

Shahbazyan, L., Petkov, G., & Gurova, L. (2014). Analogical transfer of intentions. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Ed.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2907-2912). Austin, TX: Cognitive Science Society.

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge aquisition: a developmental perspective. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 413-437). Cambridge: Cambridge University Press.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit Level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology , 72*(2), 262-274 .

Zareva, N., & Kokinov, B. (2003). Blending of non-similar episodes as a result of analogical mapping with a third one. In R. Alterman, & D. Kirsh (Ed.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1299-1304). Erlbaum, Hillsdale, NJ: Cognitive Science Society.

Dual process theory of reasoning and recognition memory errors: Individual differences in a memory prose task

Giorgio Gronchi (giorgio.gronchi@unifi.it)
Department of Neuroscience, Psychology, Drug Research and Child's Health – Sect. Psychology
via di San Salvi 12, pad 26 - 50100 Firenze, Italy

Stefania Righi (srighi@unifi.it)
Department of Neuroscience, Psychology, Drug Research and Child's Health – Sect. Psychology
via di San Salvi 12, pad 26 - 50100 Firenze, Italy

Giacomo Parrini (giacomoparrini@hotmail.com)
Department of Neuroscience, Psychology, Drug Research and Child's Health – Sect. Psychology
via di San Salvi 12, pad 26 - 50100 Firenze, Italy

Lapo Pierguidi (lapo.pierguidi@gmail.com)
Department of Neuroscience, Psychology, Drug Research and Child's Health – Sect. Psychology
via di San Salvi 12, pad 26 - 50100 Firenze, Italy

Maria Pia Viggiano (mariapia.viggiano@unifi.it)
Department of Neuroscience, Psychology, Drug Research and Child's Health – Sect. Psychology
via di San Salvi 12, pad 26 - 50100 Firenze, Italy

Abstract

Cognitive factors can mediate the tendency to create false memory. We explored the role of the two systems of reasoning in the production of false memories. Such difference can be assessed through the Cognitive Reflection Test (CRT), a measure of the propensity to reflect rather than producing an intuitive response. By the use of a DRM-like paradigm in a prose recognition memory task, we measured CRT-related individual differences in producing false memories. We observed that intuitive thinkers are more likely to produce false memories.

Keywords: DRM, False memories, CRT, Dual process theory of reasoning.

Introduction

A long tradition of cognitive research has observed how people can develop false memories about past events (Bartlett, 1932; Deese, 1959; Loftus, 1975). This is a fundamental topic both from an applicative (e.g. eyewitness memory) and theoretical point of view (e.g. the nature of semantic knowledge, Collins & Loftus, 1975). The Deese-Roediger-McDermott paradigm (DRM) has become the standard way to investigate this phenomenon (Roediger & McDermott, 1995). In this paradigm, participants are presented with a list of semantically-related words constructed in a particular way. An initial word is chosen (for example, hospital) and then other semantically related words are selected (doctor, nurse, patient, etc). Such list of words (with the exception of the initial word) is presented for learning, followed by a recognition memory task. The key observation is that in the recognition task participants with high confidence report that the initial word "doctor" was actually in the list. Through such paradigm it is possible to observe how a false memory (a memory for an event that never occurred) can be created. This result is usually explained in terms of an associative model of memory. According to this perspective, the initial (and missing) word is strongly activated because is highly semantically correlated with the words presented in the list. However, specific theories (such as the fuzzy trace theory or the activation-monitoring theory) underlying false memory dynamics are still under debate (Brainerd & Reyna, 2002; Brainerd et al., 2001; Gallo & Roediger, 2002; Roediger, Balota, & Watson, 2001).

Many studies have examined how individual differences may influence the memory processes with regard to the false memory phenomena (Baird, 2001; Watson et al., 2005; Winograd, Peluso, & Glover, 1998). For example, a higher rate of false memory is observed in people with higher vividness of imagery (e.g. producing a photograph-like mental picture, Winograd et al., 1998) or with a higher expertise in the domain of the material to be learnt (Baird, 2001).

However, few studies have investigated individual differences based on thought and reasoning processes. Graham (2007) has investigated the role of the Need for Cognition (NFC) construct (Cacioppo & Petty, 1982) on false memories. The NFC consists of the tendency to engage and enjoy effortful cognitive activities. Some individuals consistently seek opportunities to engage in challenging cognitive activities whereas other individuals have little motivation and tend to avoid such tasks. Graham (2007) found that the NFC mediates false memories where high NFC individuals falsely recognized a greater amount of

semantic-related words than individuals with low NFC. Such result is explained in terms of high NFC individuals' tendency to greater elaboration of list resulting in stronger interconnections in memory (and thus an increase in the probability of activation of a word present in the same semantic network).

The NFC is somewhat related to the dual process theory of judgment (Epstein, 1994; Evans & Over, 1996; Sloman, 1996, 2014; Stanovich & West, 2000). Starting from the influential *Heuristics and biases* research program, many authors have sustained the distinction of two modalities of reasoning; A fast, effortless, associative, nearly-automatic reasoning process and a slow, resources-demanding, rule-base, controlled process. How to label those two processes and specific differences about their characteristics is still under debate. We will refer to the former as the intuitive system (Sloman, 2014) and to the latter as the analytical system. The Cognitive Reflection Test (CRT; Frederick 2005) is a predictive measure of the tendency to deliberate about the outputs of intuitive processes before responding in a cognitive task. The CRT is composed of three questions, each one has an (incorrect) obvious intuitive answer and a less accessible (correct) answer that needs some analytical system deliberation. In Sloman's (2014) view the slow and effortful analytical process attempts to inhibit (sometimes with success and sometimes not) the "obvious" response of the effortless and fast intuitive process. For example, given the Lily Pad problem of the CRT: *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?* The default, most obvious intuitive response that does not require a particular effort is 24; If 48 days is the time necessary for the patch to cover the entire lake, consequently 24 days should be intuitively the time necessary for half the surface. However, if the reader reflects a little bit more about the problem, can easily see that every day the patch doubles in size, so in the 47th day the lake was half covered and in the 48th and final day the lily pads finished to cover the entire surface. Such form of reasoning inhibits the obvious, intuitive response.

When considering the relationship between individual differences in CRT and the false memory phenomena, two possible predictions can be made. On the one hand, it is reasonable to predict that people who are more analytical on the CRT problems will tend to produce more false memories; people with high NFC tend to create more false memories and NFC has a significant (albeit small – about .22, Frederick, 2005) positive correlation with analytical responses on the CRT. However, Frederick (2005) says that CRT measures "the ability or disposition to resist reporting the response that first comes to mind" whereas the NFC measures the "tendency to engage in and enjoy thinking" (Cacioppo & Petty, 1982). Following this line of thought, a different prediction can be that high intuitive people on CRT (due to their inability to resist reporting the first response that comes to mind) will make more false memories because they are less effective in the inhibition of semantic-related words not present in the original list. Such prediction in a memory recognition task is also supported by the fact that high intuitive people on CRT can rely more on a familiarity-based judgment than an analytical more controlled strategy.

Experiment

We investigated individual CRT-related differences in a DRM-like paradigm employing a memory prose task. Participants read a brief story with ten target words and then answered a series of questions unrelated to this experiment. The filler tasks included the CRT. After ten minutes, the material was removed and an incidental recognition memory task was administered.

Method

Participants Two-hundred-fifty-three freshmen college students enrolled in the University of Florence (97 male) were recruited for course credits. The sample mean age (in years) was 22.9 (sd = 7.4).

Stimuli and procedure Two brief cover stories (in Italian) were developed on the basis of an Italian word database (Burani, Barca, & Saskia Arduino, 2001) reporting several measures (familiarity, imagery, concreteness, frequency of use). Two sets of ten target words were chosen (First Target Set: Mill, Magpie, Winter, Nun, Lead, Medal, Carpet, Table Cloth, Compass; Second Target Set: Swamp, Thrush, Autumn, Friar, Canvas, Novel, Tavern, Lamp, Soap, Tray). Employing the First Target Set of words, we wrote the text of Story A (target words are in bold): "*Alex passed through the **mill** pausing to watch a **magpie**. The **winter** was near. He continued for another quarter of an hour thinking about the **nun** met before and finally arrived. The atmosphere was quiet and immediately noticed that the **lead** was exactly where he had left it. Clutching the **medal** in his hand, he realized that something was wrong, but could not figure out what it was. He looked around for clues: The **lantern** on his right, the **carpet** on the floor, the **table cloth** fallen on the ground together with a shattered **compass**. Someone was here... he turned and saw his wife. Her face was worth a thousand words: It was then that he realized that he had become a millionaire*.". Story B was exactly the same (in particular from a syntactical and grammatical point of view), but the target words were substituted with the nouns from the Second Target Set in order to have a story with a similar meaning. In a preliminary experiment, we asked to a separate sample (n = 40) to rate the similarity of the target words to other lists of words (matched for the reported measures) in order to create two Related Sets of nouns (First Related Set: Thief, Wind, Cold, Church, Metal, Gold, Light, Dust, Lunch, North; Second Related Set: Mud, Bird, Leaf, The Habit, Painting, Book, Room, Light, Perfume, Silver). The Unrelated Sets of nouns were chosen

from the target words of the other scenario (Story B for Story A words and vice versa) and we verified that such words were judged dissimilar from the target words (a mean value less than 2.5 in a similarity Likert scale from 1 to 5). In the case of values higher than 2.5, we conducted another preliminary experiment. We select the words presented in the database that matched the Target words and we asked another group (n = 30) to rate their similarity with the Target words; the words with the least similarity values were chosen. To sum up, the First Unrelated Set was: Swamp, Dream, Music, Sandals, Canvas, Novel, Tavern, Living Room, Soap, Tray; the Second Unrelated Set comprised: Mill, Thief, Skis, School, Lead, Medal, Friend, Carpet, Table Cloth, Compass.

Each participant was randomly assigned to one of the two Stories. After reading the story, participants were required to respond to a series of questions about the emotional valence of the story. After this series of questions, participants were required to do other tasks for an unrelated experiment. Those tasks also included the CRT (Frederick, 2005). The three problems were administered in a free response format with a limit of two minutes to complete each problem. The three problems were: 1) Bat and Ball problem *A bat and a ball cost 1.10 Euros in total. The bat costs 1.00 Euro more than the ball. How much does the ball cost? 2)* Machine Problem *If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? 3)* Lilly Pad problem *In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?*

After about ten minutes from the moment in which the text with the story was taken away, participants were invited to open a closed envelope and to complete the recognition memory task inside it. The recognition memory task was composed of 30 words (Target, Related and Unrelated words associated to the corresponding Story, A or B) arranged in a random order. For each word, participants were required to say if they have read it in the story (yes or not).

Data Analysis For each participant, we computed three scores. A Target Score, a Related Score and an Unrelated Score were calculated giving one point for each word recognized respectively in the Target, Related and Unrelated set of words; each score had a minimum of 0 and a maximum of 10 points. With regard to the CRT score, in a first phase, each response was coded as an intuitive response or a normative response. Following Frederick (2005), for the Bat and Ball problem the intuitive response is 10 and the analytical response is 5, for the Machine problem the intuitive response is 100 and the analytical response is 5, and for the Lilly Pad problem the intuitive response is 24 and the analytical response is 47. Participants that gave other kinds of responses (or that left blank) were discarded (17 participants). The CRT score was computed giving 1 point for each intuitive response and 0 points for each analytical response, resulting in a score ranging from 0 (all three problems solved with the analytical strategy) and 3 (all the problems solved with the intuitive strategies). On the basis of the CRT score, four groups were created: The Intuitive Group (participants with a total of 3 points in the CRT score, n = 62), the Mild Intuitive Group (participants with a total of 2 points in the CRT score, n = 61), the Mild Analytical Group (participants with a total of 1 point in the CRT score, n = 56), and the Analytical Group (participants with a total of 0 points in the CRT score, n = 57).

Story-related differences for the three scores (Target, Related and Unrelated) were compared by means of a mixed-model two-way ANOVA with CRT as a 4-level between-subjects variable (Intuitive, Mild Intuitive, Mild Analytical, and Analytical Group), scores (Target, Related, Unrelated) as a 3 level within-subjects variable and recognition score (number of words) as dependent variable. Given a significant result of the ANOVA, Fisher's Least Significant Difference post hoc tests were conducted to compare measures among the groups.

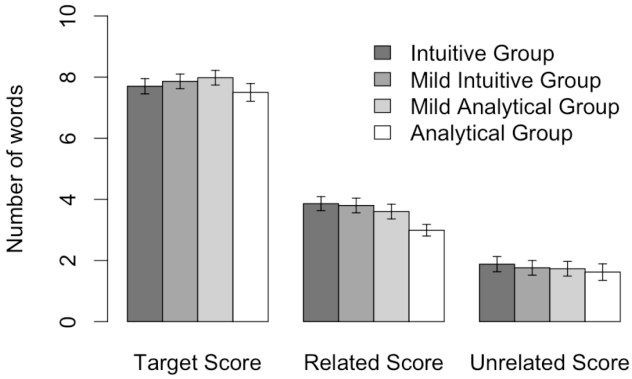


Figure 1: Mean scores of recognized Target, Related and Unrelated words of the four groups (Intuitive, Mild Intuitive, Mild Analytical, and Analytical Group).

Results

No statistically significant differences between the two Stories were found for Target Score, Related Score and Unrelated Score, so data were collapsed across the two texts.

The three scores were statistically different ($\chi^2(1) = 7.76$, $p = 009$). In particular, the Target Score was significantly higher than the Related Score ($\chi^2(1) = 13.25$, $p < .001$) and the Unrelated Score ($\chi^2(1) = 15.27$, $p < .001$). Moreover, the Related Score was statistically higher compared to the Unrelated Score ($\chi^2(1) = 10.22$, $p = .001$). No CRT-related differences were found for Target Score ($\chi^2(1) = 1.25$, $p = .264$) and Unrelated Score ($\chi^2(1) = 1.12$, $p = .291$). With regard to the Related Score, we found CRT-related

differences ($\chi^2(1) = 4.32$, $p = .038$). In particular, the Related Score for the Intuitive Group was significantly higher compared to the Normative Group ($p = .023$). Moreover, the Mild Intuitive Group scored higher than the Normative Group ($p = .041$).

Discussion

The aim of the present study was to explore the CRT-related individual differences on false memories. We found that those who scored high on the CRT (individuals who adopted an intuitive thinking style) created more false memories than those who scored low (individuals who adopted an analytical thinking style).

Given that the CRT measures mainly the ability to inhibit the first response that comes to mind, our findings can be explained in terms of a reduced tendency to inhibit the obvious response of recognizing an absent semantic-related word. Moreover, since the CRT intuitive responses are strictly associated with the use of heuristic reasoning (Toplak, West, & Stanovich, 2011), it can be hypothesized a stronger tendency to employ a familiarity heuristic to decide if a word was previously presented. Both interpretations can account for observed data. Further manipulations are needed to disambiguate these explanations. In particular, it may be desirable to measure false memories in a recognition task as well a free recall task and compare them with the CRT and an impulsivity test, such as the Barratt Impulsiveness Scale (BIS 11; Patton & Stanford, 1995). Following this procedure, it could be possible to observe if there is an interaction between the CRT individual differences and the type of task (recognition or free recall) where the use of a familiarity heuristic will induce more false memories only in the recognition task. At the same time, it could be interesting to explore the relationship between impulsivity, dual process of reasoning and false memories.

Moreover, it's important to note that the observed relationship between false memories and CRT is rather weak and it's necessary to better assess the role of potential confounding variables in future investigations. Indeed, an alternative explanation of our results is that the relationship between CRT and false memories is mediated by working memory capacity (Watson et al., 2005). Working memory capacity is correlated with analytical thinking and they positively correlate with the use of recollection processes (instead of familiarity) in memory retrieval. Thus it can be hypothesized that individuals who adopt an analytical thinking style may more rarely use familiarity processes producing less false memories. So, future research must include also a working memory capacity measure.

Contrary to the expectations based on the correlations among NFC, CRT and false memories, we find that people who adopted an analytical thinking style did not make more false memories. Such result is in line with the idea that individuals with both high and low NFC make use of their intuition. Research suggests that intuitions can influence judgments with a variable amount of thought effort. For example, an individual with high NFC will incorporate the intuition output together with other thoughts that are generated (so, the final judgment will be the result of a mediation between intuition and other thoughts). On the contrary, an individual with a low NFC will use the intuition output in a straightforward and direct way, where the final judgment will be determined almost completely by intuition. So, we can have false memories induced by the tendency to do greater elaboration of the material and false memories determined by the tendency to employ more intuitive thinking. Future research should measure both the NFC construct and the CRT in conjunction with a DRM task, in order to assess if there is an interaction between these two factors.

In conclusion, our work suggests that the ability to repress an intuitive response (as defined by the dual process theory of reasoning) is an important cognitive factor that may influence false memories.

Acknowledgments

Many thanks to Giacomo Fini, Alessio Petroni, Alessia Riela and Daniele Lari for collecting data. The English language was reviewed by Michela Gronchi. We thank the editor and three anonymous reviewers for their constructive comments, which helped us to improve the manuscript.

References

Baird, R. R. (2001). Experts sometimes show more false recall than novices: A cost of knowing too much. *Learning and Individual Differences, 13*(4), 349-355.

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge: Cambridge University Press.

Burani, C., Barca, L., & Saskia Arduino, L. (2001). Una base di dati sui valori di età di acquisizione, frequenza, familiarità, immaginabilità, concretezza, e altre variabili lessicali e sublessicali per 626 nomi dell'italiano. *Giornale Italiano di Psicologia, 28*(4), 839-856.

Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11*(5), 164-169.

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(*2), 307.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*(1), 17-22.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*, 709-724.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationally and reasoning*. Hove, UK: Psychology Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*, 25-42.

Gallo, D. A., & Roediger, H. L. (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language, 47* (3), 469-497.

Graham, L. M. (2007). Need for cognition and false memory in the Deese–Roediger–McDermott paradigm. *Personality and Individual Differences, 42*(3), 409-418.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*(4), 560-572.

Patton, J. H., & Stanford, M. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51*(6), 768-774.

Roediger, H. L., Balota, D. A., & Watson, J. M. (2001). Spreading activation and the arousal of false memories. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95-115). Washington DC: American Psychological Association.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803-814.

Sloman, S. A. (1996). The empirical case for two system of reasoning. *Psychological Bullettin, 119*(1), 3-22.

Sloman, S. A. (2014). Two systems of reasoning, an update. In J. Sherman, B. Gawronski, & Y. Trope (Eds.). *Dual process theories of the social mind*. New York: Guilford Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate. *Behavioural and Brain Sciences, 23*, 645-726.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275-1289.

Watson, J. M., Bunting, M. F., Poole, B. J., & Conway, A. R. (2005). Individual differences in susceptibility to false memory in the Deese-Roediger-McDermott paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 76-85.

Winograd, E., Peluso, J. P., & Glover, T. A. (1998). Individual differences in susceptibility to memory illusions. *Applied Cognitive Psychology, 12*(7), 5-27.

A neurocomputational model of the effect of learned labels on infants' object representations

Arthur Capelier-Mourguy (a.capelier-mourguy@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF United Kingdom

Katherine E. Twomey (k.twomey@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF United Kingdom

Gert Westermann (g.westermann@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF United Kingdom

Abstract

The effect of labels on nonlinguistic representations is the focus of substantial debate in the developmental literature. A recent empirical study (Twomey & Westermann, 2016) suggested that labels are incorporated into object representations, such that infants respond differently to objects for which they know a label relative to unlabeled objects. However, these empirical data cannot differentiate between two recent theories of integrated label-object representations, one of which assumes labels are features of object representations, and one which assumes labels are represented separately, but become closely associated with learning. We address this issue using a neurocomputational (auto-encoder) model to instantiate both theoretical approaches. Simulation data support an account in which labels are features of objects, with the same representational status as the objects' visual and haptic characteristics.

Keywords: connectionist model, label status, word learning

The nature of the relationship between labels and nonlinguistic representations has been the focus of recent debate. On one account, label representations are qualitatively different to object representations (Waxman & Markow, 1995). This *labels as invitations to form categories* (henceforth *label as invitations*) approach views labels as conceptual markers acting as abstract, top-down indicators of category membership, and assumes that labels are represented separately from their referents. In contrast, the *labels as features* view assumes that label representations are integrated into object representations (Gliozzi, Mayor, Hu & Plunkett, 2009; Sloutsky & Fisher, 2004). On this account, labels have no special status; rather, they contribute to object representations in the same way as other features such as shape and color. More recently, Westermann & Mareschal (2014) suggested a *compound representations* account in which labels are encoded in the same representational space as objects, and drive learning over time, but are not integrated within visual object representations. Rather, they become closely associated with object representations over learning. Importantly, although this view is superficially similar to the *labels as invitations* approach, it involves substantially different mechanisms. In the former, labels are qualitatively different from other features, and act in a top-down way to guide categorization by directing infants' attention to category-relevant exemplar features. In contrast, the compound representations view assumes that labels have the same status as other features with respect to how they are perceived. Specifically, they are not abstract guides to categories: like visual features, labels are low-level perceptual features. However, they are represented separately from visual features, and thus have equivalent representational status. In this sense, they are diagnostic – rather than deterministic – of categorization. This common status with other features allows labels to be embedded in object representations, rather than associated via an abstract link. However, despite substantial empirical work (e.g., Gelman & Coley, 1990; Gliga, Volein, & Csibra, 2010; Sloutsky & Fisher, 2004, 2012; Westermann & Mareschal, 2014) and a handful of computational investigations (Gliozzi et al., 2009; Mirolli & Parisi, 2005; Westermann & Mareschal, 2014), there is no current consensus as to the status of labels in object representations, and the debate goes on.

Nonetheless, the existence of a broader relationship between language and representation is not in dispute: multiple studies have demonstrated that language encodes perceptual distinctions (Boroditsky, 2001) and can influence representations on-line (Lupyan, 2016). While the effect of language on representation has been established in adults, however, when and how in development this relationship emerges is less clear. Studies demonstrate that labels can guide infants' online category formation in infants (Althaus & Westermann, 2016; Plunkett, Hu, & Cohen, 2008), and that learned, but unlabeled, category representations affect their in-the-moment behavior in the lab (Bornstein & Mash, 2010), but until recently the link between learned labels and representations had not been directly tested. Twomey & Westermann (2016; henceforth T&W) sought to trace the roots of this relationship to the earliest stages of language development, in 10-month-old infants. Infants were trained

by their parents over the course of a week with two objects via 3-minute play sessions. Critically, infants were taught a label for one of the objects, but not for the other. After the training phase, infants participated in a looking time task in which they were shown images of each object in silence. On the hypothesis that (previously learned) labels would affect infants' object representations, the authors predicted that infants should exhibit different looking times to the labeled and unlabeled objects. Their predictions were upheld: infants maintained interest for longer in the previously labeled than the unlabeled object. Infants of this age have been repeatedly shown to engage preferentially with novel stimuli when familiarized for sufficient time (for a review, see Houston-Price & Makai, 2004). Thus, infants' longer looking to the labeled object across familiarization was interpreted as a novelty response to the previously labeled object. The authors concluded that labels shape object representations from the very beginnings of language acquisition.

These data shed light on the status of labels debate. Specifically, they support both the labels as features and the compound representations theories: if a label is an integral part of an object's representation, there will be a mismatch between that representation and the object in the real world when the label is missing. In contrast, the labels as invitations account predicts that removing the label should have no effect on infants' responses to objects when those objects are presented in silence, because labels are not integrated into object representations. Thus, the behavioral data do not support the labels as invitations view. However, these empirical data cannot differentiate between the labels as features and compound representations views. Computational models, on the other hand, allow researchers to explicitly test the mechanisms specified by these theories against empirical data. Thus, in the current study we explored which of the labels as features and compound representations explains T&W's results by implementing both accounts in neural network models.

Model Architecture

We used a simple three-layer auto-encoder model to implement both the labels as features and the compound representations accounts. Auto-encoders are feed-forward connectionist neural networks consisting of an input layer, a smaller hidden layer and an output layer. These models have successfully captured data from infant categorization tasks (Cottrell & Fleming, 1990; Mareschal & French, 2000; Twomey & Westermann, 2015; Westermann & Mareschal, 2012, 2014). These models reproduce input patterns on their output layer by comparing input and output activation after presentation of training stimuli, then using this error metric to adjust the weights between units using back-propagation of error (Rumelhart, Hinton, & Williams, 1986). The sum of the square of these error values (SSE) is typically used as a proxy for looking time (Mareschal & French, 2000; Westermann & Mareschal, 2012, 2014), and we use this index in the current simulations. The network consisted of 19 input units, 15 hidden units, and 19 output units. Hidden units used a sigmoidal activation function while output units used a linear activation function, and weights were initialized randomly between -0.25 and 0.25. We used a learning rate of 0.1, a momentum of 0.1, and a Fahlman offset of 0.1.

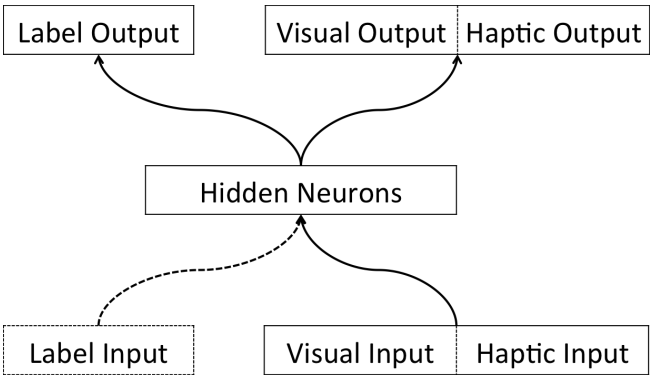


Figure 1. Network architecture

Labels as features Model (LaF)

Figure 1, including the dashed Label input, depicts the LaF model. To represent the label as a feature equivalent to all other features, we included it both at the input and the output level. Thus, the label had the same status as all other features in the model's representation.

Compound Representation Model (CR)

Figure 1, excluding the dashed label input, depicts the CR model. We based this model on the Westermann and Mareschal (2012) auto-encoder-type dual-memory model in which labels are encoded as separate outputs. Thus, when an object is presented as an input the label is retrieved as part of that representation, but the label does not act as an object feature at a perceptual level. Note that in this model, we used only 18 of the input units, making it a partial auto-encoder.

Stimuli

Simuli reflected the visual, haptic and label characteristics of T&W's 3D object stimuli.

Visual input. In T&W's empirical study stimuli were two small toys, one painted blue and one painted red. One toy was a castanet, and the other was two wooden balls joined with string. Thus, the stimuli were visually dissimilar, but both consisted of two wooden components connected with string/elastic. To reflect the partial overlap in visual appearance of these objects, we encoded the visual component of our stimuli as pseudorandom patterns of activation over 10 units. We kept the total number of active units constant for each object, but allowed distribution to vary; however, two out of the ten units overlapped to represent commonalities between stimuli (see Figure 2).

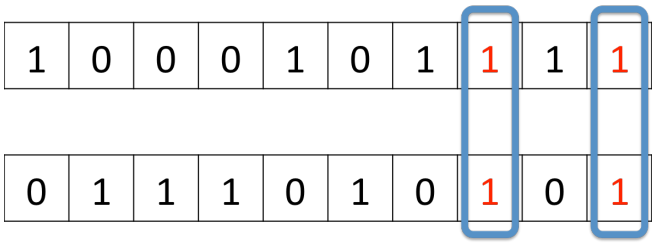


Figure 2. Visual stimuli presented to the network. Blue boxes represent overlapping features.

Haptic input. As well as visual experience, infants in T&W received haptic input when handling or mouthing the stimuli. We reasoned that the degree of overlap in this input would vary between infants. Because both objects were wooden and experienced simultaneously, infants would have experienced some minimum overlap in haptic experience of the objects. On the other hand, because the objects had different affordances, this overlap would never have been exact. Thus, we encoded haptic input over 8 units, with overlap varying randomly between 2 and 6 units between simulations. Haptic stimuli were presented to the model simultaneously with the visual stimuli and encoded in an identical fashion.

Label input. Label input consisted of a single unit, activated for the labeled object only.

Procedure

In line with T&W's experiment, our procedure consisted of two phases. First, to simulate the 3D object play sessions, we trained the model with both objects, one with a label and one without a label. Then, we tested the model in a familiarization task in which the label was absent, as in T&W. Specifically, we ran each architecture in a test condition in which the label unit was inactive for both stimuli.

Play sessions

To reflect the likely differences in playing time across children, and playing time with individual objects for each child, the total number of iterations for which the model received each stimulus during background training was selected randomly from a normal distribution of mean 500 and standard deviation 200, for each stimulus. Duration of presentation of both stimuli was randomised based on the same distribution. Presentation began with random selection of one of the stimuli. Then, each stimulus was presented 30 times and for a total number of presentations as determined before.

Familiarization Training.

Before familiarization training, we added noise to hidden-to-output weights (from a uniform distribution ranging from 0.001 to 1) to simulate the likely memory decay from infants' final play session, which had taken place the previous day. Then, we removed the label from the inputs and outputs for the LaF condition, and from the outputs only for the CR condition.

Familiarization then proceeded as follows: in line with T&W, stimuli were interleaved (with learning) for 100 iterations, or until the sum squared error (SSE) between the input and output fell below a threshold of 0.01. The threshold reflects infants' looks away from the screen as the trial proceeds (Westermann & Mareschal, 2012, 2014). Each stimulus was presented for eight trials; the familiarization phase therefore consisted of 16 trials in total. The initial stimulus was counterbalanced across simulations.

Infants' looking time on a given trial was indexed as number of presentations, either until error fell below threshold or until the maximum number of presentations was reached.

Results

Results from the CR and LaF models are depicted in Figures 3 and 4.

We submitted looking time (SSE) to a 2 (model; CR, LaF) x 2 (condition; label vs. no-label) x (trial; 1 - 8) mixed ANOVA. Overall, the CR and LaF models' looking time differed ($F(1, 67168) = 1166.73$, $p < .0001$, $\eta_p^2 = .017$), and decreased rapidly across trials ($F(7, 67168) = 28697.85$, $p < .0001$, $\eta_p^2 = .751$). There was also a small but robust difference in looking times to the labeled versus the non-labeled object, $F(1, 67168) = 55.07$, $p < .0001$, $\eta_p^2 = .001$). All two- and three-way interactions also contributed to the model (all Fs > 2.82, all ps < .004). To understand these interactions we conducted individual ANOVAs for each simulation. The CR model's looking time decreased rapidly across trial ($F(7, 33584 = 15307.88$, $p < .0001$, $\eta_p^2 = .761$), and there was a much smaller effect of condition ($F(1, 33584) = 4.18$, $p = .041$, $\eta_p^2 < .001$). However, unlike in T&W, there was no trial-by-condition interaction. Thus, although the CR model did show an effect of previously-learned labels on in-task looking times, it did not capture the nuanced pattern of results in the empirical study.

The LaF model's looking times also decreased across trial ($F(7, 33584) = 13940.44$, $p < .0001$, $\eta_p^2 = .74$), and this model showed a somewhat stronger effect of condition ($F(1, 33584) = 64.55$, $p < .0001$, $\eta_p^2 = .002$). Critically, the effect of trial interacted with the effect of condition (7,33584) = 8.67, $p < .001$, $\eta_p^2 = .002$). As depicted in Figure 4, post-hoc pairwise comparisons (Bonferroni-corrected) demonstrated that looking times initially decreased more rapidly in the no-label condition (all ps < .0001) but that this difference disappeared by the end of familiarization. Thus, while both models replicate T&W's overall finding that labels affect object representations, only the LaF model reflects the exact pattern of results reported by T&W: when all else is held equal, teaching the LaF model a label for one object but not another leads to a more rapid decrease in looking time to the unlabeled object in a subsequent, silent familiarization phase.

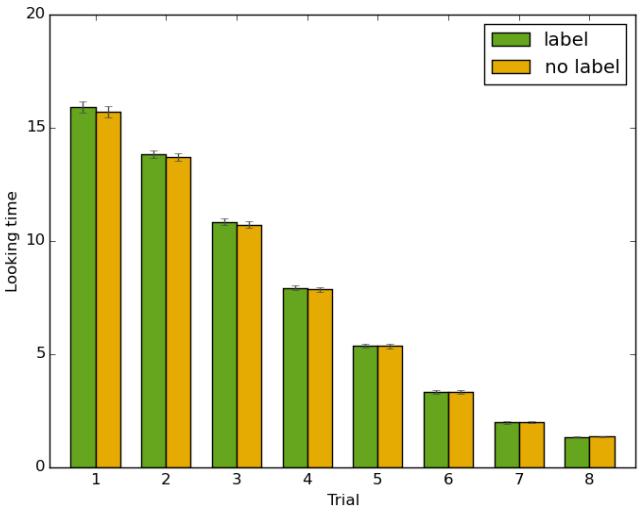


Figure 4. Results from the compound representations model.

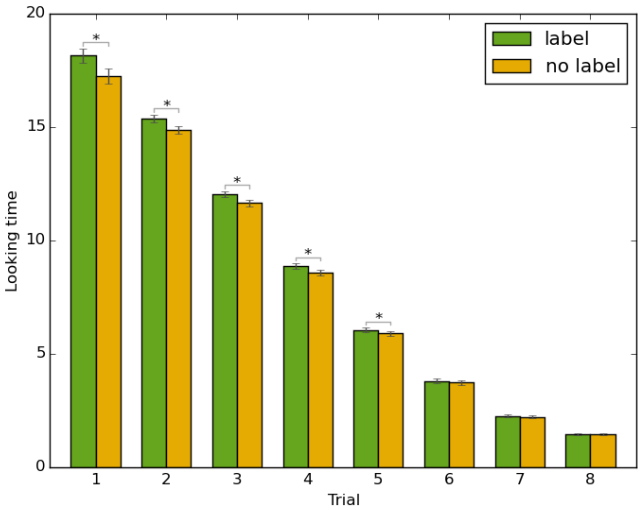


Figure 3. Results from the labels as features model. *p < .001

Discussion

In the current study we tested two possibilities for the relationship between labels and object representations using a neurocomputational model to capture recent empirical data (Twomey & Westermann, 2016). The target data showed that learned labels affect 10-month-old infants' looking times in a silent familiarization phase, suggesting that knowing a label for an object directly affects its representation, even when that object is presented in silence. As noted by T&W, both the compound representations and labels as features accounts predict some effect of labels on object representations, however the empirical data could not shed light on which of these two accounts best explained the pattern of results they observed. To untangle these two possibilities, we implemented both accounts in simple auto-encoder models (cf., Mareschal & French, 2000; Twomey & Westermann, 2015). In the compound representations model we instantiated labels on the output layer. This model learned to associate labels with inputs over time such that the presence of visual/haptic input for an object would consistently activate the label, but nonetheless, label representations were separate from visual and haptic object representations (Westermann & Mareschal, 2014). In the labes as features model, labels were represented on the input as well as on the output layer with the same status as the visual and haptic components of object representations (Gliozzi et al., 2009; Sloutsky & Fisher, 2004). Only the labels as features model captured the more rapid decrease in looking to the no-label stimulus exhibited by the infants in T&W's empirical study.

This work offers converging evidence that labels may have a low-level, featural status in infants' representations. In line with recent computational work (Gliozzi et al., 2009; Westermann & Mareschal, 2014) we chose to explore such low-level accounts to establish whether a simple associative model could account for the nuances of T&W's data. We did so for two reasons. First, the labels as invitations account assumes that in their communicative context, labels prompt infants to attend to the similarities between category exemplars (Ferguson & Waxman, 2016; Fulkerson & Waxman, 2007; Waxman & Markow, 1995). It is not clear how this view relates to T&W's data: although it predicts that representations of category exemplars will be more similar to one another in the presence of labels, it does not state whether this effect will be evident for a single object in the absence of a label or other communicative information, as in the familiarization task presented to infants in T&W's study (e.g., Ferguson & Waxman, 2016; Futó, Téglás, Csibra, & Gergely, 2010). In contrast, our labels as features model offers a parsimonious account of T&W's results, in which looking time differences emerge from a low-level novelty effect. Specifically, as argued by T&W, over background training the label is learned as part of the object representation. Thus, when the object appears without the label there is a mismatch between representation and external input. This mismatch leads to an increase in network error, our proxy for looking time, capturing the empirical data without a need for high-level communicative cues.

The current simulations also relate closely to the dual memory model presented by Westermann & Mareschal (2014). Our labels as features architecture suggests that if infants were to be taught a label for a *category* of objects – rather than a single object as described here – the absence of a label during familiarization should nonetheless provoke a similar novelty response. Interestingly, however, Westermann and Mareschal (2014) make the opposite prediction: familiarization time to new exemplars of a previously-learned category should be faster when those exemplars have previously been labeled. A key difference between the current model and that of Westermann & Mareschal is the latter's separation of memory into long- and short-term components. This allows the presence of a label during background training to actively restructure category representations, pulling exemplars closer to the

prototypical category member. From this perspective, new category members should be perceived as more similar to learned representations when that category has a label, irrespective of the presence of the label in-the-moment. In contrast, labels in the present model are simply shared features, the absence of which gives rise to a mismatch. It is of course possible that both predictions are correct, but come into play at different developmental stages. For example, early in language development infants could initially form simple, featural associations between labels and objects, as in the current model. Over time, however, label representations could become more deeply entrenched, leading to the "magnet"-type effect on representations predicted by Westermann & Mareschal (cf. Deng & Sloustky, 2015; Kuhl, 1991). An empirical test of this possibility is important for a detailed understanding of labels' status in object representations.

It is important to note that other computational work has explored the effect of labeling and representation in this age group. Gliozzi et al. (2009) used a self-organizing map (SOM; Kohonen, 1998) architecture to capture empirical data from a categorization task with 10-month-old infants. In this network, labes are represented as units in SOMs in the same way as visual features. This model could capture T&W's results for similar reasons to the success of our labels as features model, although this remains an empirical question. However, the two networks make very different assumptions about learning mechanisms, highlighting an important issue for both infancy research and computational work. Gliozzi and colleagues' model learns in an unsupervised way, strengthening associations between units in its SOMs using "fire together, wire together" Hebbian learning. In contrast, our model learns by comparing what it "sees" to what it "knows" and updating its representations in proportion to any discrepancy. Thus, the current results are compatible with an error-based learning account to development, in which infants learn by tracking mismatches between representation and environment (Heyes, 2015). Whether unsupervised learning, error based learning, or some combination of both drives early development is a profound theoretical issue outside the scope of the current paper; for now we highlight the importance of bearing in mind the link between the technical assumptions of a computational model and the implications for (developmental) theory. Taken together with T&W, however, the current work demonstrates how language can shape representation and even change behavior from the bottom up.

Acknowledgments

This work was supported by a Doctoral Scholarship Grant from the Leverhulme Trust and by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

References

Althaus, N., & Westermann, G. (2016). Labels constructively shape categories in 10-month-old infants. *In Press, Journal of Experimental Child Psychology*. http://dx.doi.org/10.1016/j.jecp.2015.11.013

Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development, 81*(3), 884–897.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology, 43*(1), 1–22. http://doi.org/10.1006/cogp.2001.0748

Cottrell, G. W., & Fleming, M. (1990). Face recognition using unsupervised feature extraction. In *Proceedings of the International Neural Network Conference* (pp. 322–325).

Deng, W. (Sophia), & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3).

Ferguson, B., & Waxman, S. R. (2016). What the [beep]? Six-month-olds link novel communicative signals to meaning. *Cognition, 146*, 185–189. http://doi.org/10.1016/j.cognition.2015.09.020

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition, 105*(1), 218–228.

Futó, J., Téglás, E., Csibra, G., & Gergely, G. (2010). Communicative Function Demonstration induces kind-based artifact representation in preverbal infants. *Cognition, 117*(1), 1–8. http://doi.org/10.1016/j.cognition.2010.06.003

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*(5), 796.

Gliga, T., Volein, A., & Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. *Journal of Cognitive Neuroscience, 22*(12), 2781–2789. http://doi.org/10.1162/jocn.2010.21427

Gliozzi, V., Mayor, J., Hu, J. F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science, 33*(4), 709–738.

Heyes, C. (2015). When does social learning become cultural learning? *Developmental Science.*

Kohonen, T. (1998). The Self-Organizing Map, a possible model of brain maps. *Brain and Values*, 207–236 568.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107. http://doi.org/10.3758/BF03212211

Lupyan, G. (2015). The paradox of the universal triangle: concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology*, (just-accepted), 1-69.

Mareschal, D., & French, R. (2000). Mechanisms of categorization in infancy. *Infancy, 1*(1), 59–76. http://doi.org/10.1207/S15327078IN0101_06

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in Psychology, 4*. http://doi.org/10.3389/fpsyg.2013.00491

Mirolli, M., & Parisi, D. (2005). Language as an aid to categorization: A neural network model of early language acquisition. *Progress in Neural Processing, 16*, 97.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition, 106*(2), 665–681. http://doi.org/10.1016/j.cognition.2007.04.003

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology-General, 133*(2), 166–188.

Sloutsky, V. M., & Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? *Journal of Experimental Child Psychology, 111*(1), 65–86. http://doi.org/10.1016/j.jecp.2011.07.007

Twomey, K. E., & Westermann, G. (2015). A neurocomputational model of curiosity-driven infant categorisation. *Proceedings of the 5th Joint IEEE International Conference on Development, Leaning and Epigenetic Robotics*. Providence, RI.

Twomey, K. E. & Westermann, G. (2016, August). A learned label modulates object representations in 10-month-old infants. *Poster to be presented at the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, PA*.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology, 29*(3), 257–302.

Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development, 27*(4), 367–382.

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120391.

A neural network model of hierarchical category development

Chris Gorman (cgorman@cs.otago.ac.nz)

Alistair Knott (alik@cs.otago.ac.nz)
Department of Computer Science, 133 Union St East
Dunedin, 9016 New Zealand

Abstract

Object recognition and categorization is a fundamental aspect of cognition in humans and animals. Models have been implemented around the idea that categories are sets of frequently co-occurring features. Out of these models a question has been raised, namely what is the mechanism by which we learn a hierarchically organized set of categories, including types and subtypes? In this paper we introduce such a model, the Dominant Property Assembly Network (DPAN). DPAN uses an unsupervised neural network to model an agent which develops a hierarchy of object categories based on highly correlated object features. Initially, the network generates representations of high-level object types by identifying commonly co-occurring sets of features. Over time, the network will start to use an inhibition of return (IOR) operation to examine the features of a categorized object that make it unusual as an instance of its identified category. The result is a network which, early in training, represents classes of objects using coarse-grained categories and recognizes objects as members of these general classes, but eventually is able to recognize subtle differences between subtypes of objects within the broad classes, and represent objects using these more fine-grained categories.

Keywords: categorization; computational modeling; prototype theory

Introduction

Humans and animals begin developing classifications of objects starting from very soon after birth (Quinn, Slater, Brown, & Hayes, 2001) and throughout the rest of our lives. We are able recognize tens of thousands of different objects (Biederman, 1987; Brady, Konkle, & Alvarez, 2011), which is an extremely important evolutionary tool. It allows us to rapidly determine whether a token object is, for example, edible, dangerous, friendly, and so on (DiCarlo, Zoccolan, & Rust, 2012). It is embedded in so many aspects of our lives that we often don't give it a second thought. It provides us a way to compress as much information as possible with as little cognitive effort as possible (Rosch, 1999). Humans and animals learn to represent token objects as instances of object categories. The structure of categories has been explored extensively in cognitive science (see e.g. Rakison & Yermolayeva, 2010). Categories reflect several types of structures. A central idea is that categories represent **correlations** between the features of objects. We recognize that, for example, a token object is a "dog" because it has features shared with other things we call dogs. But another key idea is that categories emerge to represent types of objects that have distinctive properties. Exactly how these two apparently conflicting criteria coexist is the subject of much debate (Tyler & Moss, 2001). In the current paper we will propose a novel architecture that reconciles them.

This category system is hierarchically organized. **Basic level categories** like "chair" or "dog" afford the agent a cognitively efficient representation of the object (Mervis & Rosch, 1981). Above the basic level we have **superordinate categories**. Superordinate categories, like "furniture" or "animal," are difficult for agents to visualize and they contain several intra-category differences. For example, a car and a boat can both be considered in the "vehicles" superordinate category, but they have far fewer shared properties. In addition, the important correlations may be determined as much by culture as by the objects themselves (Liu, Golinkoff, & Sak, 2001). Below the basic level we have **subordinate categories**. This level contains categories which may not contain many functionally actionable differences when compared against their basic level category, such as "computer chair" or "pug," but rather simply provide more detailed information. Developmentally, subordinate categories are learned after basic-level categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Finally, beneath the subordinate level we have the level of token individuals which is able to differentiate particular, unique instances of categories from one-another, such as differentiating "a dog" from "my dog, Charles Barkley." This level includes spatiotemporal information in conjunction with other sensory input and is beyond the scope of this work. This paper focuses on the subordinate level of categorization and, to a lesser degree, the basic level.

When a person has minimal experience with a type of object[1], they tend to focus on the major features shared between members of that type before investigating the differences. For example, a child may categorize any animal with four legs as "dog" (Rakison & Yermolayeva, 2010). A young child can recognize categories, such as "dog" and "cat," from a series of commonly occurring features. Within these categories, the child is able to pick up on more frequently occurring, albeit more subtle, combinations of features, such as those shared by pugs. They also notice other sets of subtle, frequently occurring features within the same category, like those shared by spaniels. Due to the subtle nature of these regularities, they weren't noticed until after the basic-level category was solidified. If we assume that the mechanism which learns basic-level categories does so by identifying the strongest correlations among object features, then by definition we must seek a different explanation for the emergence of subordinate-level categories.

[1]For the remainder of the paper we refer to a **category** or **type** as an internal representation of commonly co-occurring features and a **token object** as a particular instance of a category.

Most modern work on computational object categorization has focused on the basic and individual levels (Riesenhuber & Poggio, 2000; Winn, Criminisi, & Minka, 2005; Galleguillos, Rabinovich, & Belongie, 2008). However, in recent years, a significant amount of work has been done in an attempt to recognize subordinate categories as well (Zhang, Gao, Xia, Dai, & Li, 2015; Farrell, Oza, Morariu, Darrell, & Davis, 2011; Yang, Bo, Wang, & Shapiro, 2012; Chai, Lempitsky, & Zisserman, 2013). Most of these models recognise subordinate categories by analysing the sub-parts of objects and discovering regularities in the identity and configuration of these sub-parts. However, the models do not pay so much attention to the question of *when* an agent begins to learn subordinate-level categories within a given basic-level category - that is, to the developmental trajectory of subcategory learning. We propose that there are particular circumstances which lead to an agent starting to learn subcategories, and that the process of learning subcategories is implemented through overtly scheduled cognitive operations that have correlates in surface language.

Our model of subcategory learning is implemented within a network called the **dominant property assembly network** (DPAN). DPAN is presented with the visual features of a series of token objects and begins learning internal representations of those objects' basic-level categories by identifying the strongest correlations amongst these features. This might lead to the emergence of categories "dog" and "cat", for example. However, in order to learn more subtle correlations identifying subordinate-level categories, we propose that the strongest correlations should be actively inhibited. We accomplish this through a cognitive operation that is a type of **inhibition of return** (IOR).

When presented with a token object, DPAN first classifies the object to activate an internal representation of its category. This internal representation is associated with a certain collection of features which identify dogs, for example. Having activated this representation, DPAN then **inhibits** the associated features, allowing it to focus on what makes this particular token object different or unusual. This idea of "inhibiting the winner" is found in several neural circuits and is often referred to as "inhibition of return." It was originally shown in spatial attention (Posner, Rafal, Choate, & Vaughan, 2007), where agents were shown to "inhibit orienting towards visual locations which have been previously attended". DPAN, however, operates on the domain of properties rather than spatial locations. The IOR operation can be understood as an operation that identifies a *property* of the currently attended object. The process of identifying properties is one that is readily reported in language, in **predicative sentences**. For instance, in the sentence "The dog is brown", the dog is predicated as having the property brown. It is interesting that object categories can feature both referentially and as predicates: for instance, in the sentence "The dog is a pug", "The dog" is a referential expression, but "a pug" is simply a property that is predicated of the dog (Partee, 1987). The difference between referential nominals and predicating nominals is still a matter of debate for linguists; our model of IOR in category learning will make a suggestion about this difference. Once training is completed, the network will have developed representations of object classes as well as subtypes of those classes. In our example, it would contain category representations of "dog" alongside "pug," "beagle," "spaniel," and "corgi."

The other key novelty in DPAN is the use of **per-unit learning rates**. DPAN uses localist units to represent highly correlated features. The method by which the network learns these correlations is explained in detail in a later section. Once a localist unit emerges which strongly and sufficiently represents a token object's category, the unit stops learning to prevent further input from altering its encoded category. To achieve this end, the network employs per-unit learning rates. These learning rates are associated with a particular localist unit, monotonically decreasing, and are based on a measure of the rate of change of the connections for their respective unit.

The use of a per-unit local learning rate is well established in the field of neural networks (Thimm, Moerland, & Fiesler, 1996; Bengio, Simard, & Frasconi, 1994; Becker & Le Cun, 1988; Schiffmann & Geffers, 1993). The notion of a learning rate which changes based on its objective performance is also well established (Senior, Heigold, Ranzato, & Yang, 2013; Renals, Morgan, Cohen, & Franco, 1992). Since the learning rate is a monotonically decreasing function, it doesn't cause a feedback loop. For example, if the learning rate updated as a function of the gradient of change, then it would simply raise itself up or lower itself down because the learning rate directly affects the weights' gradient of change. By allowing the weights to update using a constant value, we are able to more accurately measure the amount of learning each unit is accomplishing. Then, when the weight change gradient for that unit is low, we can confidently say that it is because the unit has completed its learning process and no longer needs to be updated.

The remainder of this paper is organized as follows. First we describe DPAN in detail, followed by an introduction to our experiments and walk through of their results. In the final section we present our conclusions.

Architecture

DPAN is effectively organized into three separate layers (Figure 1). The first layer is the **rich property complex** (RPC) which contains the raw object properties provided by the sensorimotor system. The next layer is the **dominant property assembly** (DPA). The DPA essentially provides a workspace for computations to be done on the RPC without permanent modification. When the network is presented with a token object, it first copies the information directly from the RPC into the DPA. Above the DPA layer lies the **conditional principal component analysis** (CPCA) units, introduced by O'Reilly and Munakata (2000), which constitute the localist property

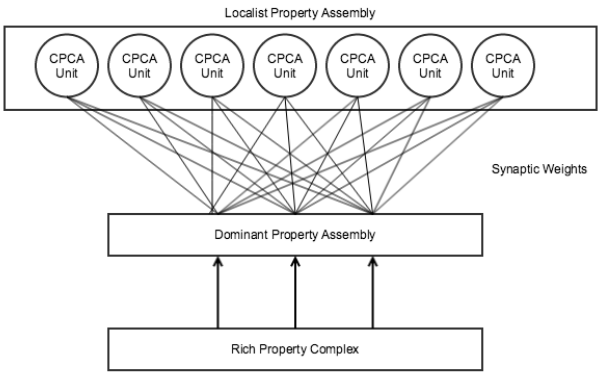


Figure 1: An overview of the DPAN architecture.

assembly (LPA) layer. The weights connect each feature in the DPA layer with each unit in the LPA layer. CPCA provides the core learning mechanism for the network and is explained in detail in the next section. The CPCA units in the LPA layer represent the basic level and subordinate categories.

CPCA

As we stated previously, the core learning of DPAN is accomplished via CPCA. CPCA is an unsupervised artificial neural network based on Hebbian learning which generates an internal model of strongly correlated features. The algorithm is explained in detail in (O'Reilly & Munakata, 2000), but we present a brief overview here as our method differs slightly from the original implementation. CPCA takes binary input and generates binary outputs using a competitive winner-take-all approach. It uses one layer of neurons fully connected to the input vector. The neurons produce output by calculating the weighted sum of of their inputs, selecting the one with the highest output, setting that value to 1 and the rest to 0. The weights are updated using Equation 1 where y_j is the activity of the unit, x_i is the input feature, w_{ij} is the weight between them, and α_j is a learning rate, between 0 and 1, of the unit. O'Reilly et al. provide a derivation proving that $w_{ij} = P(y_j = 1|x_i = 1)$. In a practical sense, the weight between a CPCA unit and an input vector increases when the unit is active and the input is 1, or decreases when the neuron is active and the input is 0. In DPAN, the weights are connected to the dominant property assembly rather than to the rich property complex. This allows us to manipulate the CPCA's input vector, during inhibition of return for example, while maintaining a reference to the original input data.

$$\Delta w_{ij} = \alpha_j y_j (x_i - w_{ij}) \tag{1}$$

The Training Algorithm

We start by defining a few terms. A **training item** is a combination of feature values originally presented to the RPC. A **training episode** encompasses all of the processing that is done on a single training item. If IOR is invoked, the episode may involve several **iterations**, i.e. an update of the network's weights. An **epoch** is a collection of training episodes such that each training item has been presented to the network once. The maximum number of epochs for each execution is represented by λ and the current epoch is represented by t. Training typically consists of multiple epochs where the list of training items has been shuffled each time. It's also worth mentioning that there should be at least enough CPCA units to represent the number of basic level and subordinate categories in the training data, but fewer than the number of features in the input vectors. The network is initialized with psuedorandom weights between 0.4 and 0.6. This allows each CPCA unit to have a decent chance of learning each of the objects without any major bias initially. At this point, we also initialize the learning rate for each unit. At the start of each training iteration, the RPC is copied into the DPA and the network calculates the activity of the CPCA units using Equation 2. The network then chooses the unit with the highest output $\max \vec{y}$, selects it as the winner, sets its output to 1 and the output of all other units to 0. Once the winner is selected, the weights for that unit $\vec{w}_j$ are updated using equation 1. The total change in weight is now measured to determine whether or not to begin IOR and to disable further learning on this unit. If the gradient of change is steeper than the threshold τ, the iteration is complete and the process begins again.

$$\vec{y} = \mathbf{W}^T \vec{x} \tag{2}$$

To calculate the gradient of change of the weight vector $\vec{w}_j$ at time t, DPAN allocates space for a temporary weight vector, $\vec{w}_j(t-1)$, updates the weights as normal, and then subtracts the current weight from the previous one, taking the absolute values (equation 3). $\vec{d}$ is now summed up to produce the scalar value of the total change the weight vector $\vec{w}_j$ underwent (equation 4). This sum is appended to a vector $\vec{\Omega}_j$ such that the contents of the vector are the total change in weight for each iteration. $\vec{\Omega}_j{}'$ is calculated such that it contains the gradient of $\vec{\Omega}_j$. Finally, the last two elements of $\vec{\Omega}_j{}'$ are subtracted from one another and if their difference is less than τ, we consider the gradient of change to be minimal.

$$\vec{d} = |\vec{w}_j(t) - \vec{w}_j(t-1)| \tag{3}$$

$$\sum_{k=1}^{|\vec{w}_j|} w_{ik} \tag{4}$$

If the network determines that there was minimal change, IOR begins. The DPA is now updated such that $\vec{x} = \mathbf{W}\vec{y}$, essentially copying the object prototype from the LPA into the DPA. The RPC is now subtracted from the DPA in the inhibition step, $\vec{x} := \vec{z} - \vec{x}$, and stored back into the DPA. At this stage, the DPA to now contains the difference between the prototype object and the actual token object. For example, if the token object was "pug" and the winning unit was "dog," the DPA would now contain the properties that pugs have which differentiate them from other dogs.

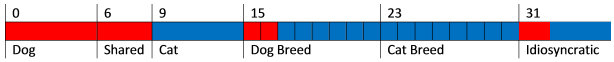
Since the IOR operation can run indefinitely, we must define stopping conditions. If, after the inhibition operation, the DPA doesn't contain anything "interesting," IOR ceases and no learning is done. That is, if there is minimal difference between the token object and the winning unit's property assembly, there is nothing for the network to learn, so it stops. For example, if the winning unit near-perfectly represented the token object there would be very little difference between the RPC and the prototype object. If there is something interesting left in the DPA, the network learns in the same way as before: the activity of each unit is calculated, a winner is selected, and that unit's weights are updated. The only differences this time are that the LPA layer performs a self-inhibition operation which prevents previous winners from this iteration to win again and that before the weights are updated, the contents of the DPA are replaced with the contents of the RPC; the network chooses a winner based on the unique properties, but then trains on the entire property complex. After the weights are updated, the gradient of change for the new winning unit is calculated. If the change was large, the inhibition loop finishes and the training iteration for this input is complete. Otherwise, the loop repeats, allowing more interesting features to bubble up.

The initial result of training (Stage 1) is that the coarse-grained object categories are learned. At this stage, when an object is presented to the network, the winning LPA unit represents a supertype. Now the network starts to systematically inhibit the units representing coarse-grained types after they are activated, allowing other units in the network to develop representations of finer-grained subtypes. When these are first learned (Stage 2), the network activates *first* a unit representing a supertype, and *then* (after IOR) a unit representing a subtype. After even more training, the subtype units learn better representations of the training objects than the supertype units, and the network activates a subtype representation *as its first response* (Stage 3). In linguistic terms, the response at Stage 1 could be rendered "that is a dog", at Stage 2, "That dog is a pug", and at Stage 3, "That is a pug".

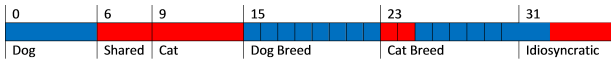
Experimental Setup

Input Data

DPAN was trained on a set of binary input vectors, each representing a token object containing 36 features. Each element of the vector encodes a distinct, abstract property of the token object. When a bit is set to 1 it indicates the presence of a property and when it is set to 0 it indicates the absence of that property. Each input vector represents one of four cat or four dog breeds. Figure 2 describes the layout of the input bits and provide examples of token individuals. All four breeds of dog shared the same set of "generic dog" properties and all four breeds of cat shared the same set of "generic cat" properties. There is an overlap of three bits between these two sets representing properties shared between all dogs and cats. The next 16 bits represent breed-specific properties, e.g. a wrinkled face and short snout for a pug. Each breed is identified



(a) A token pug



(b) A token tabby

Figure 2: Two vectors representing token individuals. Red indicates a value of 1 and blue a value of 0.

by which two of those bits are active, so there are eight bits for the dog breeds (pug, beagle, spaniel, and corgi) and eight bits for the cat breeds (tabby, maine coon, siamese, and persian). Finally, there is a set of five idiosyncratic property bits which are each uniformly randomly set to 0 or 1 for each token individual. These bits represent weak, uncorrelated properties, a unique color or a marking of some kind for example.

Training Runs

DPAN was trained on a set of 5000 input vectors. Each of the eight breeds were equally distributed in the training set. The network was trained using the following input parameters:

- $|y| := 25$
- $\alpha_j := 0.02 \forall j \in y$
- $\tau := 0.000001$
- $\lambda := 500$

These parameters were selected based on empirical evidence showing they provide the best results for this particular dataset. The network produces output at arbitrary epoch and input intervals. The output of the network is a heat-map of the weight matrix, again where red indicates a value of 1 and blue a value of 0. The column vectors of the weight matrix represent the CPCA units and the row vectors represent the input features.

Results

We present the results of the execution by examining three key stages of training. Early in training, DPAN learns localist representations of the basic-level categories for dog and cat, as shown in Figure 3a where units 23 and 16 represent dog and cat respectively. The basic-level units represent the highly correlated properties found in the dog and cat breeds while also maintaining the weak correlations of each subordinate-level category. For example, the network's localist dog unit encapsulates the common features associated with each dog it has been exposed to, but also maintains weak connections to the subtle correlations for each dog breed it has seen. At this stage of training, DPAN's each dog or cat it sees will activate its corresponding basic-level localist unit and reinforce these connections.

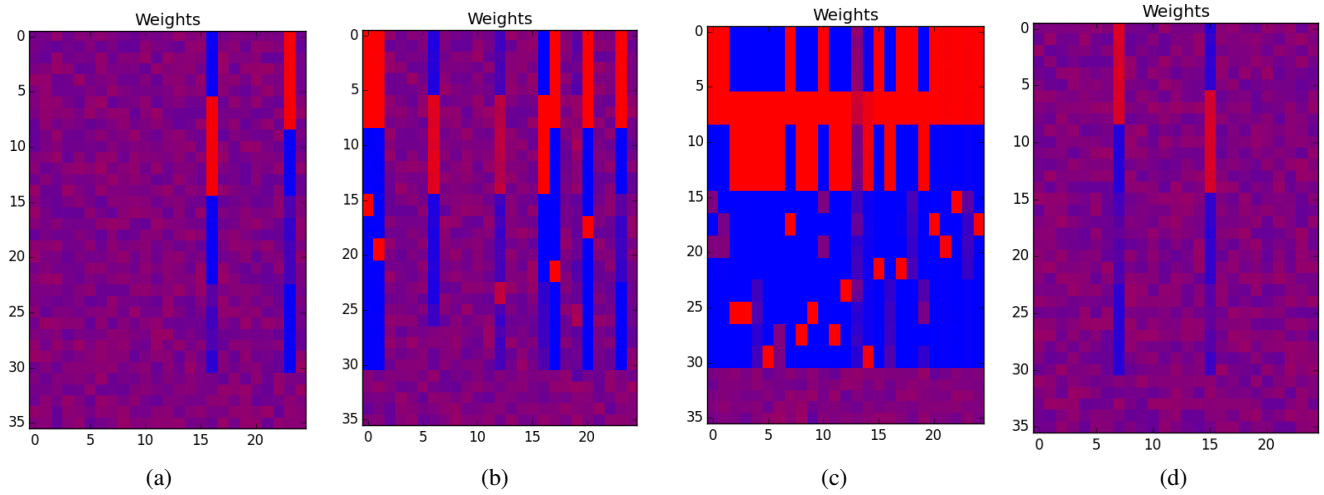


Figure 3: DPAN results: (a) aDPAN weight matrix during the first stage of training when basc-level categories are formed. (b) bDPAN weight matrix during the second stage of training. Subordinate-level units are beginning to emerge as a result of the IOR operation. (c) cDPAN weight matrix after the final stage of training. Each subordinate-level and basic-level category is well represented by at least one unit in the LPA layer. (d) dModified DPAN without IOR operation after 100 epochs.

The next stage of training occurs after learning in the 'dog' and 'cat' units stabilizes. At this point, whenever a cat or dog is presented, DPAN executes an IOR operation and chooses another unit to represent the 'unusual' features of the cat or dog that has just been classified. The results of this stage of training are illustrated in Figure 3b. At this point, there are CPCA units that represent subtypes of dogs and subtypes of cats. For example, unit 23 still represents the 'dog' category while unit 17 represents the 'corgi' subcategory.

Finally, after each subordinate-level unit has been exposed to enough token objects, DPAN reaches its last stage of training. During this stage, when the network is exposed to a token object it will simply activate the corresponding subordinate-level localist unit. Note as well that the basic-level units remain intact. If a new subordinate-level category is presented to the network at this point, i.e. one which has no localist representation, the basic-level unit will still activate.

In order to illustrate the importance of the IOR operation, an experiment was carried out wherein IOR was disabled during execution. The results of this experiment are presented in Figure 3d. As anticipated, after 100 epochs the network is still only able to learn the basic-level categories. Without the inhibition of return operation, the network is unable to learn the subtle differences that define the subordinate-level categories.

A Possible Account of the Difference Between Referential and Predicative Nominals

While describing the execution of DPAN, we separated it into three distinct stages. As discussed previously, these three stages of training model a human acquiring expertise in a given category. In the first stage, the network corresponds to referential uses of basic-level categories: "dog" and "cat." Stage 2 corresponds to predicative sentences, featuring subordinate-level categories as predicates: "The dog is a pug." We also posit that the use of the word "is" represents the inhibition operation that allows DPAN to learn these fine grained categories. That is, when describing the token object, a person first activates their internal representation of the object's basic level category ("The dog..."), then inhibits that ("..is a...") to focus on the subtle correlations of the subordinate-level category ("...pug."). Stage 3 corresponds to referential uses of subordinate categories. The agent no longer actively inhibits the basic-level category and instead initially activates its internal representation of the token object's subordinate-level category ("The pug.").

Conclusions and Future Work

Members of basic-level categories contain many highly-correlated features. After enough experience we can inhibit these more obvious connections, allowing us to hone in on more subtle correlations and to create finer-grained subordinate-level categories. The dominant property assembly network is able to learn basic-level and subordinate-level categories in the same manner. DPAN learns the strong correlations that exist in the data to create a representations of basic-level categories. Once these have been encoded, DPAN then uses its inhibition of return operation to learn the dataset's weaker correlations, generating new representations of subordinate-level categories. After enough experience, the subordinate-level categories win out over their basic-level counterparts, mimicking human categorization.

The next step for DPAN is to integrate it into a computational vision system, allowing it to train on real-world image data. DPAN would also benefit from an additional su-

pervised network to assign names to its category representations. DPAN could potentially be used in a large-scale image recognition system to create categories and subcategories of any number of objects.

References

Becker, S., & Le Cun, Y. (1988). Improving the convergence of back-propagation learning with second order methods. In *Proceedings of the 1988 connectionist models summer school* (pp. 29–37).

Bengio, Y., Simard, P., & Frasconi, P. (1994, January). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–66. doi: 10.1109/72.279181

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115.

Brady, T., Konkle, T., & Alvarez, G. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 1–34.

Chai, Y., Lempitsky, V., & Zisserman, A. (2013, December). Symbiotic Segmentation and Part Localization for Fine-Grained Categorization. In *Iccv 2013* (pp. 321–328). IEEE. doi: 10.1109/ICCV.2013.47

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012, February). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–34. doi: 10.1016/j.neuron.2012.01.010

Farrell, R., Oza, O., Morariu, V. I., Darrell, T., & Davis, L. S. (2011, November). Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *Iccv 2011* (pp. 161–168). IEEE. doi: 10.1109/ICCV.2011.6126238

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008, June). Object categorization using co-occurrence, location and appearance. In *Cvpr 2008* (pp. 1–8). IEEE. doi: 10.1109/CVPR.2008.4587799

Liu, J., Golinkoff, R. M., & Sak, K. (2001, January). One cow does not an animal make: Young children can extend novel words at the superordinate level. *Child Development*, *72*(6), 1674–94.

Mervis, C. B., & Rosch, E. (1981, January). Categorization of Natural Objects. *Annual Review of Psychology*, *32*(1), 89–115. doi: 10.1146/annurev.ps.32.020181.000513

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press.

Partee, B. (1987). Noun phrase interpretation and type-shifting principles. *Studies in discourse representation theory and the theory of generalized quantifiers*, *8*, 115–143.

Posner, M. I., Rafal, R. D., Choate, L. S., & Vaughan, J. (2007, August). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, *2*(3), 211–228. doi: 10.1080/02643298508252866

Quinn, P. C., Slater, A. M., Brown, E., & Hayes, R. A. (2001, June). Developmental change in form categorization in early infancy. *British Journal of Developmental Psychology*, *19*(2), 207–218. doi: 10.1348/026151001166038

Rakison, D. H., & Yermolayeva, Y. (2010, November). Infant categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 894–905. doi: 10.1002/wcs.81

Renals, S., Morgan, N., Cohen, M., & Franco, H. (1992). Connectionist probability estimation in the DECIPHER speech recognition system. In *Icassp-92* (Vol. 1, pp. 601–604 vol.1). IEEE. doi: 10.1109/ICASSP.1992.225837

Riesenhuber, M., & Poggio, T. (2000, November). Models of object recognition. *Nature neuroscience*, *3 Suppl*, 1199–204. doi: 10.1038/81479

Rosch, E. (1999). Principles of categorization. *Concepts: Core Readings*, 189–206.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976, July). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi: 10.1016/0010-0285(76)90013-X

Schiffmann, W. H., & Geffers, H. W. (1993, January). Adaptive control of dynamic systems by back propagation networks. *Neural Networks*, *6*(4), 517–524. doi: 10.1016/S0893-6080(05)80055-3

Senior, A., Heigold, G., Ranzato, M., & Yang, K. (2013). An empirical study of learning rates in deep neural networks for speech recognition. In *Icassp 2013* (pp. 6724–6728).

Thimm, G., Moerland, P., & Fiesler, E. (1996, February). The Interchangeability of Learning Rate and Gain in Backpropagation Neural Networks. *Neural Computation*, *8*(2), 451–460. doi: 10.1162/neco.1996.8.2.451

Tyler, L., & Moss, H. (2001, June). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, *5*(6), 244–252. doi: 10.1016/S1364-6613(00)01651-X

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. In *Iccv'05 volume 1* (Vol. 2, pp. 1800–1807 Vol. 2). IEEE. doi: 10.1109/ICCV.2005.171

Yang, S., Bo, L., Wang, J., & Shapiro, L. G. (2012). Unsupervised Template Learning for Fine-Grained Object Recognition. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 3122–3130). Curran Associates, Inc.

Zhang, L., Gao, Y., Xia, Y., Dai, Q., & Li, X. (2015, January). A fine-grained image categorization system by cellet-encoded spatial pyramid modeling. *IEEE Transactions on Industrial Electronics*, *62*(1), 564–571. doi: 10.1109/TIE.2014.2327558

Improving Visual Memory with Auditory Input

Scott R. Schroeder (schroeder@u.northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Viorica Marian (v-marian@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Abstract

Can input in one sensory modality strengthen memory in a different sensory modality? To address this question, we asked participants to encode images presented in various locations (e.g., a depicted dog in the top left corner of the screen) while they heard spatially *un*informative sounds. Some of these sounds matched the image (e.g., the word "dog" or a barking sound) while others did not. In a subsequent memory test, participants were better at remembering the locations of images that were encoded with a matching sound, even though these sounds were spatially uninformative – an effect that was mediated by whether the sounds were verbal or non-verbal. Because the sounds did not provide any relevant location information, better spatial memory cannot be attributed to auditory memory; rather, it is attributed to visual memory being strengthened by the matching auditory input. These findings provide the first behavioral evidence for cross-modal interactions in memory.

Keywords: Audio-Visual Integration; Memory; Multisensory Processing; Visual Spatial Memory

Introduction

We live in a multi-sensory world, where auditory, visual, and other sensory inputs merge together to form our experiences, many of which we later try to remember. For example, consider the experience of parking your car in an unfamiliar place and then later trying to find it. After parking your car, you will likely try to commit the location of the car to memory. At the same time as you are memorizing the car's location and preparing to walk away, you might lock the car's doors, and the car might make a sound to indicate that the doors have been successfully locked. Does hearing this auditory input (the sound of the car) help you encode and then later remember the important visual input (the car's location in space)? More generally, can input in one sensory modality strengthen episodic memory in a different sensory modality?

The answer to this question has basic-science implications for our understanding of how memory works in real-world, ecologically-valid contexts, as our everyday experiences are largely multi-sensory in nature. The answer may also have applied-science implications for educational programs, cognitive therapies, and human factors designs, as multi-sensory input might improve memory performance.

Despite its relevance to basic- and applied-science and despite evidence for cross-modal interactions in other cognitive domains, there is, to our knowledge, no persuasive behavioral evidence for cross-modal interactions in episodic memory. In several previous studies, hearing congruent auditory input during the encoding of a visual image (for example, hearing a dog's bark while encoding a picture of a dog) was found to aid later recognition of the previously-viewed visual image (the picture of the dog) (Lehmann & Murray, 2005). However, the finding that congruent auditory input improves visual item memory (also known as visual *what* memory) might not reflect auditory input changing visual memory. When participants were presented with a visual image (a picture of a dog) on the recognition memory test, they may have remembered hearing the image's congruent sound (the barking sound of a dog), which could be used to correctly indicate that they had previously seen the image, even if the participant forgot the image. Thus, the use of the helpful auditory memory trace may have led to better visual memory performance, even if hearing the auditory input did not actually strengthen visual memory.

To determine if auditory input can truly strengthen visual memory, we created a multi-sensory audio-visual memory task in which memory for the auditory input itself could not help participants perform the visual memory test. Specifically, we presented to-be-encoded visual objects in various spatial locations on the screen (for example, an image of a dog placed in the top left corner of the screen), along with spatially *un*informative auditory cues (for example, a barking sound played to both ears and thus not linked in any way to the location of the image). Then, in a later memory test, we assessed visual spatial memory (also known as visual *where* memory) for the previously-seen images. If *where* memory performance for a visual image is improved when the image is encoded with a spatially uninformative but congruent auditory cue (relative to a control condition), it would have to reflect better visual memory per se. It could not reflect the use of auditory memory to help participants perform the *where* memory test because the auditory memory trace does not contain any relevant location information and therefore would not help spatial memory performance. Thus, better visual *where* memory performance in this task would provide evidence that input in one modality can strengthen episodic memory in a different modality.

A cross-modal effect of auditory input on visual memory might depend on the type of auditory input. Verbal sounds

(i.e., spoken words like "dog") and non-verbal sounds (i.e., environmental sounds like a dog barking) are known to affect visual processing differently (Chen & Spence, 2011; Edmiston & Lupyan, 2015). Environmental sounds and spoken words differ in the location information they provide during visual spatial processing. Environmental sounds (such as a barking dog) provide helpful information about the location of the relevant object (the dog). In contrast, spoken labels (such as "dog") provide location information about the speaker but not about the relevant object (i.e., the dog), as the word "dog" can be uttered irrespective of the specific location (or even presence) of the dog (a feature of language known as displacement). Because environmental sounds and spoken words differ in their spatial informativeness for relevant objects, they may have different effects on visual *where* memory.

In the current study, we examined the effects of environmental sounds and spoken words on visual *where* memory (as well as visual *what* memory). In an environmental sounds experiment and a spoken words experiment, participants encoded a series of visual objects (for example, a dog) located in one of the four corners of the screen while hearing task-irrelevant, spatially uninformative auditory cues. The auditory cues were either *congruent* with respect to the visual object (the sound of a dog barking while seeing a dog), *incongruent* with respect to the object (the sound of a motorcycle's exhaust while seeing a trumpet), or *neutral* with respect to the object (a semantically-meaningless beep sound while seeing a helicopter). Participants then performed an item (*what*) and spatial (*where*) memory task to test memory for what pictures they saw and where they saw them. If hearing a congruent auditory cue were shown to help *what* memory performance relative to the neutral control condition, this finding would replicate previous research and demonstrate the benefits of having two sensory memory traces (or dual-codes) for memory performance; however, it would not provide evidence for cross-modal effects. The crucial test for cross-modal effects is *where* memory performance. A finding of better *where* memory for the congruent condition relative to the neutral control condition would provide evidence for cross-modal effects in memory.

Methods

Participants

Forty English-speaking young adults (median age = 21.5 years; 32 females, 8 males) were included in the study. Participants received monetary compensation or course credit for their participation. The experiment was approved by the Northwestern University Institutional Review Board.

Materials and Procedure

Participants completed the environmental sounds experiment and the spoken words experiment in a counterbalanced order, such that 20 participants completed the environmental sounds experiment first while the other 20 participants completed the spoken words experiment first. None of the auditory or visual stimuli that appeared in the first experiment also appeared in the second experiment.

Environmental Sounds and Visual Memory Experiment

Encoding. Participants viewed 60 pictures during the encoding task. In the set of 60 pictures, 20 were presented with their *congruent* environmental sound (i.e., the sound associated with that object), 20 with an *incongruent* environmental sound (i.e., the sound associated with a different object), and 20 with a *neutral* control sound (i.e., one of twenty tonal beep sounds). A semantically neutral sound (a tonal beep) was used for the control condition instead of no sound. The reason for using a semantically neutral sound as the control rather than no sound is that, in no-sound trials there would be a matching context between encoding and retrieval (i.e., both contexts would be silent), whereas in the congruent and incongruent trials, there would be a mismatching context between encoding and retrieval (i.e., there would sound at encoding but not at retrieval). The degree of match between encoding and retrieval context affects memory performance (Smith & Vela, 2001). By using a neutral sound (a tonal beep), the change in context from encoding to retrieval (a sound at encoding and no sound at retrieval) is the same for all trials, be they congruent, neutral, or incongruent.

To ensure that potential differences in the memorability of the pictures across congruent, neutral, and incongruent conditions did not affect the results, we created the stimuli in the following way. Four lists of 20 picture-sound pairs were compiled. Each of the four lists served in one of four positions – (1) as the 20 picture-sound pairs in the congruent trials, (2) as the 20 pictures in the neutral trials, which were paired with a tonal beep sound, (3) as the 20 pictures in the incongruent trials, or (4) as the 20 sounds in the incongruent trials. The four lists rotated, serving in all four positions an equal number of times across participants. The benefit of creating four lists and having them rotate positions is that if one list of pictures is easier or harder to remember than others, the results will not be affected because each list appears in each condition an equal number of times across participants and thus each condition is equally influenced by any discrepancies between lists. Nevertheless, care was taken to ensure that the lists were equivalent. The words associated with the picture-sound pairs (e.g., the word "cat" for a picture of a cat and the cat's meow sound) were matched across all four lists on English frequency, concreteness, familiarity, and imageability (MRC Psycholinguistic Database).

All pictures were selected to be similar in saturation and line thickness. Each sound was edited to be 1000 milliseconds in duration, so that congruent, neutral, and incongruent auditory cues were matched in duration. The 20 tonal beeps were sine waveforms ranging from 300 Hz to 2200 Hz with each tone being 100 Hz different from its nearest two tones. All sounds were peak-amplitude

normalized using Audacity. The sounds were presented using monophonic sound reproduction and played to both ears through headphones, so as to ensure that they did not provide any relevant spatial information.

Of the 60 pictures, 15 were presented in the upper left corner, 15 were presented in the upper right corner, 15 were presented in the lower left corner, and 15 were presented in the lower right corner.

Each trial started with a 200-millisecond fixation cross in the center of the screen. Following the fixation cross, a picture was displayed for 1000 milliseconds. Simultaneous with the onset of the picture, a sound was played for 1000 milliseconds. Figure 1 provides a visual representation of the encoding task.

In the instructions of the encoding task, participants were asked to try to remember the pictures for a later memory test but not to be concerned with remembering the sounds. After the encoding task, participants completed a five-minute filler task in which they performed a simple math test (participants determined which of two values is larger). The purpose of the filler task was to prevent recency effects (Cohen, 1989), which might lead participants to remember predominantly the last sequence of pictures, regardless of condition.

Retrieval. In the retrieval task, participants viewed 120 pictures: the 60 pictures they had seen in the encoding task ('old' pictures), plus 60 foil pictures that they had not seen before ('new' pictures). The retrieval task had two components – an item or *what* memory component and a spatial or *where* memory component. In each trial, a picture was displayed and participants had to click 'new' (indicating that they did not recognize the picture from the encoding task) or 'old' (indicating that they did recognize the picture from the encoding task). If participants clicked 'old' (indicating that they had seen the picture before), they then made a judgment about its spatial location. They did so by clicking in one of four boxes located in the four corners of the screen. A visual representation of the retrieval task is shown in Figure 1.

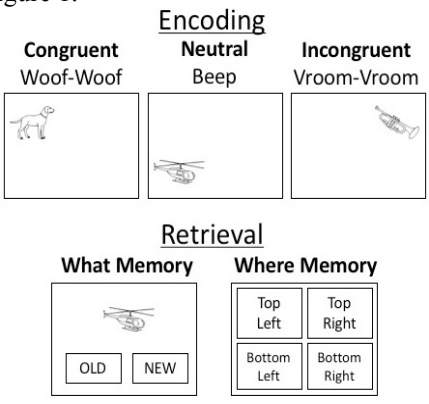


Figure 1: Top row depicts a congruent, neutral, and incongruent trial in the encoding phase of the *Environmental Sounds and Visual Memory* experiment. Bottom row depicts the *what* and *where* retrieval trials.

Spoken Words and Visual Memory Experiment

The spoken words experiment had the same methodology as the environmental sounds experiment unless noted below.

Encoding. In the spoken words experiment, participants were shown 64 pictures, of which 16 were presented with a *congruent* auditory cue (the English word for the visual object), 16 with an *incongruent* auditory cue (the English word for a different object), 16 with a *neutral non-linguistic* control auditory cue (a tonal beep sound), and 16 with a *neutral linguistic* control auditory cue (a pseudoword). The neutral *non-linguistic* control was included as in the environmental sounds experiment. In addition to the neutral *non-linguistic* control, a neutral *linguistic* control was also included for the purpose of having a linguistic control against which to compare the congruent linguistic condition. (The spoken words experiment had more pictures than the environmental sounds experiment in order to accommodate the additional condition in the spoken words experiment; note, however, that a pilot study that included an equal number of pictures and conditions in both experiments yielded the same results.)

Five lists of 16 picture-word pairs were compiled. The five lists served in one of the five positions – (1) as the 16 picture-word pairs in the congruent condition, (2) as the 16 pictures in the neutral non-linguistic condition, which were paired with a tonal beep sound, (3) as the 16 pictures in the neutral linguistic condition, which were paired with a pseudoword, (4) as the 16 pictures in the incongruent condition, or (5) as the 16 words in the incongruent condition. The pairings in the incongruent condition were created by matching a picture from one list (e.g., a trumpet) with a word from another list (e.g., the word "dog"). The five lists rotated, serving in every position an equal number of times across participants.

The words in the picture-word pairs were matched across all five lists on English frequency, English phonological neighborhood size, English biphone frequencies, number of English phonemes, concreteness, familiarity, and imageability (MRC Psycholinguistic Database; CLEARPOND; Marian, Bartolotti, Chabal, & Shook, 2012).

The pseudowords came from Colbertian, an artificial language (Bartolotti & Marian, 2012). Colbertian pseudowords were designed to conform to phonotactic rules of English and did not differ from the five lists of picture-word pairs in number of phonemes or in English biphone frequencies (CLEARPOND)

The spoken word stimuli were recorded at 44100 Hz by a female native English speaker. All words and pseudowords were equal to or shorter than 1000 milliseconds in duration. The tonal beep sounds were 1000 milliseconds in duration and ranged from 250 Hz to 1750 Hz, with each tone being 100 Hz different from its nearest two tones. None of the tones had the same frequency as the tones in the environmental sounds experiment.

After the encoding task, participants completed the five-minute filler math test, as in the environmental sounds experiment, but with different numbers.

Retrieval. In the retrieval task, participants viewed 128 pictures: the 64 pictures they had seen in the encoding task ('old' pictures), plus 64 foil pictures that they had not seen before ('new' pictures). As in the environmental sounds experiment, the retrieval task had two components – an item or *what* memory component and a spatial or *where* memory component. A visual representation of the encoding and retrieval phases is presented in Figure 2.

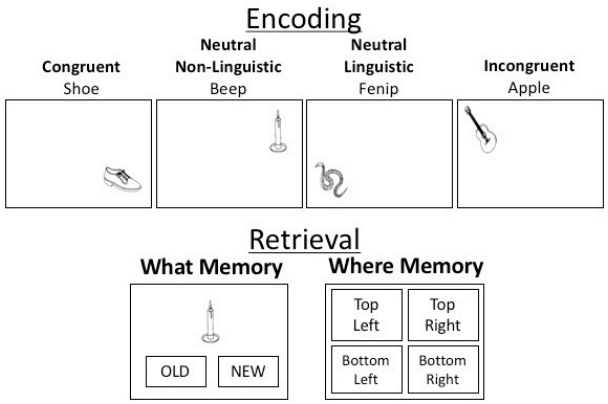


Figure 2: Top row depicts a congruent, neutral non-linguistic control, neutral linguistic control, and incongruent trial in the *Spoken Words and Visual Memory* experiment. Bottom row depicts the *what* and *where* retrieval trials.

Results

Where Memory

Environmental Sounds. The effects of environmental sounds on visual spatial or *where* memory were analyzed using a repeated-measures ANOVA with condition (Congruent, Incongruent, Neutral) as the independent variable and accuracy on the spatial memory task as the dependent variable. Accuracy rates by condition are displayed in Figure 3. The ANOVA yielded a significant main effect of condition, F (2, 78) = 11.72, $p < .001$, $\eta_p^2 = .23$. The significant main effect was followed up with contrasts between the experimental conditions (congruent and incongruent) and control condition (neutral). The contrasts indicated that the locations of pictures in the congruent condition were remembered significantly better than the locations of pictures in the neutral control condition (68.7% versus 56.9%), F (1, 39) = 14.74, $p < .001$, $\eta_p^2 = .27$. Conversely, the locations of pictures in the incongruent condition were not remembered significantly differently than the locations of pictures in the neutral control condition (57.8% versus 56.9%), F (1, 39) = 0.11, $p > .1$, $\eta_p^2 = .003$. These results suggest that visual spatial (*where*) memory was improved by hearing a congruent environmental sound. To determine whether this group-level effect was consistent across individuals, we computed the number of participants who remembered more locations in the congruent condition than in the neutral condition (and vice versa). Twenty-eight participants remembered more locations in the congruent condition, whereas only 12 remembered more locations in the neutral condition.

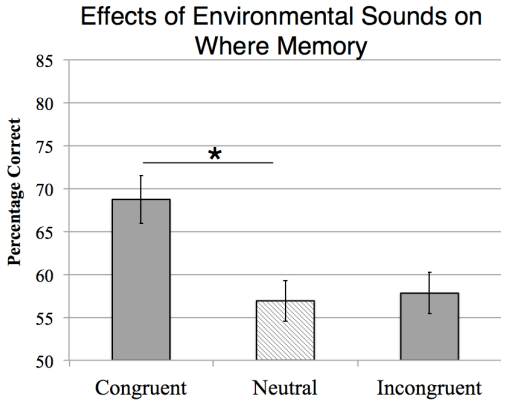


Figure 3: Memory accuracy on the spatial memory trials in the *Environmental Sounds and Visual Memory* experiment.

Spoken Words. To analyze the effects of spoken words on visual spatial or *where* memory, a repeated-measures ANOVA was conducted with condition (Congruent, Incongruent, Neutral Non-Linguistic, Neutral Linguistic) as the independent variable and spatial memory accuracy as the dependent variable. Accuracy by condition is presented in Figure 4. The ANOVA revealed no significant main effect of condition, F (3, 111) = 0.24, $p > .1$, $\eta_p^2 = .01$. These results indicated that visual spatial (*where*) memory was not improved by hearing a congruent spoken word. Consistent with these group-level results, individual-level results indicated that 18 participants remembered more locations in the congruent condition than in the neutral non-linguistic condition and 20 participants remembered more locations in the neutral non-linguistic condition than in the congruent condition.

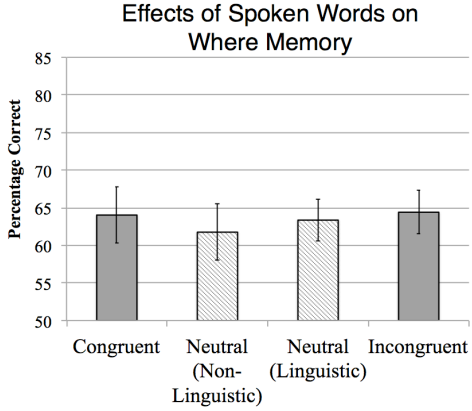


Figure 4: Memory accuracy on the spatial memory trials in the *Spoken Words and Visual Memory* experiment.

What Memory

Environmental Sounds. To analyze the effects of environmental sounds on visual item or *what* memory, a repeated-measures ANOVA was conducted with condition (Congruent, Incongruent, Neutral) as the independent variable and accuracy on item memory trials as the dependent variable. The accuracy rates for the three conditions are displayed in Figure 5. The ANOVA yielded a significant main effect of condition, F (2, 78) = 4.21, p < .05, η_p^2 = .10. Follow-up contrasts revealed that pictures in the congruent condition were recognized at a significantly higher rate than pictures in the neutral control condition (79.3% versus 73%), F (1, 39) = 5.47, p < .05, η_p^2 = .12. Similarly, pictures in the incongruent condition were recognized with significantly higher accuracy than pictures in the neutral control condition (77.9% versus 73%), F (1, 39) = 5.32, p < .05, η_p^2 = .12. These results indicate that hearing a congruent (or incongruent) environmental sound helped visual item (*what*) memory performance. These results at the group level were consistent with the results at the individual level, as 24 participants remembered more congruent pictures than neutral pictures, 13 participants remembered more neutral pictures than congruent pictures, and 3 participants remembered the same number of congruent and neutral pictures.

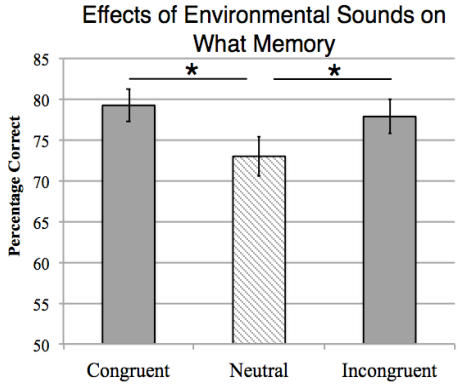


Figure 5: Memory accuracy on the item memory trials in the *Environmental Sounds and Visual Memory* experiment.

Spoken Words. The effects of spoken words on item or *what* memory were analyzed using a repeated measures ANOVA with condition (Congruent, Incongruent, Neutral Non-Linguistic, Neutral Linguistic) as the independent variable and accuracy on item memory trials as the dependent variable. Accuracy rates by condition are presented in Figure 6. The ANOVA yielded a significant main effect of condition, F (3, 108) = 4.56, p < .01, η_p^2 = .11. Follow-up contrasts revealed that pictures in both the congruent (77.5%) and incongruent (78%) conditions were recognized significantly better than pictures in the neutral non-linguistic control condition (71.2%) (F (1, 37) = 5.13, p < .05, η_p^2 = .12 and F (1, 37) = 10.31, p < .01, η_p^2 = .22, respectively). Moreover, congruent pictures were remembered marginally better and incongruent pictures were remembered significantly better than pictures in the neutral linguistic control condition (73.8%) (F (1, 37) = 3.00, p = .09, η_p^2 = .08 and F (1, 39) = 7.37, p < .05, η_p^2 = .17, respectively). These results provide evidence that hearing a congruent (or incongruent) spoken word improved visual item (*what*) memory performance. These group-level results were also reflected in the individual data, with 19 participants remembering more congruent pictures than neutral non-linguistic pictures, 13 participants remembering more neutral non-linguistic pictures than congruent pictures, and 6 participants remembering the same number of both.

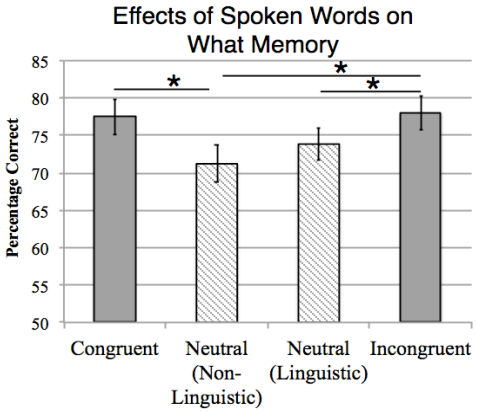


Figure 6: Memory accuracy on the item memory trials in the *Spoken Words and Visual Memory* experiment.

Discussion

We examined whether auditory input can impact visual memory. The results showed that hearing a congruent yet spatially uninformative environmental sound (for example, a barking sound played to both ears) improved memory for where a visual object was located (for example, a dog located in the top left corner of the screen). Because improved spatial memory in this case cannot be attributed to better auditory memory (memory for the environmental sound had no valid location information and therefore would not lead to a correct answer on the visual spatial memory test), the results are attributable to visual memory being improved by the environmental sound. To our knowledge, these results provide the first behavioral evidence for a cross-modal interaction in memory.

The effects of auditory input on *where* memory depended on the type of sound. While environmental sounds strengthened visual spatial memory, spoken words did not. These differences can be explained by theories positing that people unconsciously generate expectations on the basis of learned regularities (for example, predictive coding and schema theory). According to these theories, our cognitive system would expect an environmental sound, such as a dog's bark, to carry helpful location information because of the learned regularity that environmental sounds nearly always correlate with the location of the visual object (that

is, the dog's bark comes from the location of the dog). Because of this expectation, we may become more responsive to the relevant visual spatial information upon hearing an environmental sound. In contrast, our cognitive system would be unlikely to expect a spoken label, such as the word "dog", to carry helpful location information about the dog because spoken labels rarely correlate with the location of the visual object to which they refer. With a weaker expectation for valid spatial information about the referent, we might not be especially responsive to the relevant visual spatial information upon hearing a spoken label.

A second possible explanation for why congruent environmental sounds (but not congruent spoken words) enhanced *where* memory is that congruent environmental sounds may have elicited deeper processing, heightened arousal, or increased attention to the stimuli (relative to congruent spoken words). However, if congruent environmental sounds prompted deeper processing, heightened arousal, or increased attention to the stimuli, then congruent environmental sounds should have also enhanced *what* memory to a larger degree than congruent spoken words because deeper processing, heightened arousal, and increased attention to the stimuli are all known to produce stronger *what* memory (Craik & Tulving, 1975; Dolcos, LaBar, & Cabeza, 2004; Naveh-Benjamin, Guez, & Marom, 2003). Yet, relative to a neutral control sound, congruent environmental sounds did not enhance *what* memory more than congruent spoken words.

Still another explanation for why congruent environmental sounds (but not congruent spoken words) enhanced *where* memory relates to the ventriloquism effect. According to the ventriloquism effect (Slutzky & Recanzone, 2004), simultaneously hearing a sound and seeing an image can sometimes lead to the illusion that the sound is coming from the image. If the ventriloquism effect occurred for congruent environmental sounds (but not for congruent spoken words), it may have yielded helpful spatial encoding of the congruent environmental sounds. However, this explanation assumes that the ventriloquism effect depends on semantic congruence and on type of auditory cue, is precise enough to distinguish spatial locations within centimeters, and is reliable with headphones. At present, none of these assumptions has strong empirical support.

The finding that congruent environmental sounds and spoken words increased *what* memory replicates several previous studies (e.g., Lehmann & Murray, 2005). It is possible that these results are due in part to cross-modal interactions in memory, but they can also be attributed to dual-coding, where a memory is encoded in both the auditory modality and visual modality, and later visual memory performance is helped by remembering the encoded auditory cue.

In conclusion, we found that auditory input can strengthen visual episodic memory. The current results provide evidence for cross-modal interactions in memory and extend multi-sensory research on perception and attention by showing that audio-visual interactions are not just short-lived perceptual and attentional effects; they have longer-term consequences that persist in memory. Because most of our experiences are multi-sensory, these results capture how memory works in everyday situations. These findings may also carry practical implications, as cross-modal enhancements may be applied to increase memory in educational programs, cognitive therapies, and human factors designs.

Acknowledgments

This research was funded by grant NICHD R01 HD059858 to Viorica Marian.

References

Bartolotti, J., & Marian, V. (2012). Language learning and control in monolinguals and bilinguals. *Cognitive Science*, *36*(6), 1129-1147.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1554-1568.

Cohen, R. L. (1989). The effects of interference tasks on recency in the free recall of action events. *Psychological Research*, *51*(4), 176-180.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294.

Dolcos, F., LaBar, K. S., & Cabeza, R. (2004). Dissociable effects of arousal and valence on prefrontal activity indexing emotional evaluation and subsequent memory: An event-related fMRI study. *Neuroimage*, *23*(1), 64-74.

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93-100.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*(2), 326-334.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PloS one*, *7*(8), e43230.

Naveh-Benjamin, M., Guez, J., & Marom, M. (2003). The effects of divided attention at encoding on item and associative memory. *Memory & Cognition*, *31*(7), 1021-1035.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, *12*(1), 7-10.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203-220.

Measuring lay theories of parenting and child development

Emily Hembacher
ehembach@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Parenting practices are known to play an important role in shaping children's outcomes. For example, children whose parents engage them in high-quality conversations and who are given opportunities for free play are at an advantage for learning and later academic outcomes. However, communicating the results of relevant scientific findings to parents remains a challenge. One possible moderator of uptake of parenting information is the implicit theories parents hold with regard to child development and parenting. As a first step in investigating this possibility, the present work establishes a new measure of parenting attitudes including three subscales corresponding to attitudes about rules and respect, affection and attachment, and early learning. We then examine whether subscale scores predict uptake of new information about children's learning. Scores on the Early Learning subscale, but not the Rules and Respect subscale, predicted generalization from the article, providing initial evidence of the validity of this measure.

Keywords: Parenting attitudes; implicit theories

Child development research is constantly generating information that can be brought to bear on best practices for parenting. For example, research on children's learning has demonstrated that pedagogy can improve learning in some contexts and limit it in others, suggesting that allowing children to play freely and explore is critical for learning (Bonawitz et al., 2011; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011). Likewise, a great deal of research has demonstrated the importance of engaging young children in elaborative conversations for language development and future academic success (Hart & Risley, 1995; Hoff, 2003). A fundamental challenge we face is how to communicate the results of such scientific inquiry to a diverse public in a way that maximizes uptake and improves people's daily and long-term decision making.

One critical parameter that may moderate parenting behavior is parents' lay theories about child development and parenting. Lay theories reflect the core beliefs that people hold in different domains, which may or may not be explicitly articulated, but organize the processing of new information and decision-making (Dweck & Leggett, 1988; Ong, Zaki, & Goodman, 2015). For example, people with an entity theory of personality tend to interpret people's behaviors as stemming from fixed personality traits rather than situational factors such as needs, goals, or emotional states (Dweck, Chiu, & Hong, 1995).

There are two reasons to focus on parents' lay theories. First, parents' lay theories might be an important explanatory factor for many of the behaviors parents engage in with their child. For example, a parent who believes that building a strong emotional bond with their baby is one of the most important goals of parenting might have more physical contact with their child than a parent who does not hold this theory. Secondly, parents' lay theories may moderate the uptake of new information about parenting. It is well-established that people more easily encode new information that is consistent with an existing schema or mental model they hold (Bransford & Johnson, 1972). In addition, previous research has found that interventions on public health beliefs are more successful when they take into account people's existing belief structures in the domain (Kumar et al., 2015).

There is some evidence supporting the notion that parents' behaviors are mediated by implicit lay theories about child development, which vary by SES and across cultures. For example, cross-cultural studies have found profound differences in how parents interact with infants. Richman, Miller, & LeVine (1992) found that mothers in the Gusii community of Kenya primarily engaged with their children to soothe them when upset, but did not often speak to them with the goal of engaging or stimulating them, as did Caucasian parents in the United States. The authors attribute this behavior to cultural conventions stemming from the belief that there is no purpose in speaking to infants, as they will not understand what is being said (LeVine, 2004; Richman et al., 1992).

There are also important differences in how parents within western cultures interact with their children. Numerous studies have identified SES disparity in the amount that parents talk to their children, which in turn predicts children's language and academic outcomes (Hoff, 2003; Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002). In an effort to identify the source of this disparity, Rowe (2008) discovered that parents' knowledge of child development (as indexed by their scores on the Knowledge of Infant Development Inventory; KIDI) predicted their child-directed language, with more knowledgeable parents speaking to their children more even when controlling for the amount of speech directed at another adult. Although this study examined parents' knowledge, and not their lay theories *per se*, it provides evidence that people's domain knowledge has real consequences for their interactions with their children.

There are many other examples of parenting beliefs on which parents differ. For example, there is a large body of research based on Baumrind (1971)'s framework that identifies parents as authoritative, authoritarian, or permissive, according to their levels of responsiveness and control in their interactions with their children. In sum, parents' approaches to parenting appear to vary in predictable ways based on their knowledge and perceptions about children's learning and development.

Subscale	Full item
Rules and Respect	It is very important that children learn to respect adults, such as parents and teachers.
	It is important for young children to learn to control their impulses (e.g., waiting when told to wait).
	Children should be taught to be grateful to their parents.
	Children should not be punished for breaking small rules.*
	Parents should follow their children's lead rather than imposing structure in the form of rules.*
	Young children should be allowed to make their own decisions, such as what to eat for dinner.*
Affection and Attachment	Parents need to provide safe and loving environments for their children.
	Holding and cradling babies is important for forming strong bonds between parent and child.
	Children should be given comfort and understanding when they are scared or unhappy.
	Parents do not need to talk to their child about his or her emotions.*
	Children become spoiled if they receive too much attention from parents.*
	Too much affection can make a child weak.*
Early Learning	Children can learn about things like good and bad behavior from a very early age.
	Young children can teach themselves things by exploring and playing.
	Babies repetitive behaviors (e.g., banging a cup on the table) are a way for them to explore cause and effect.
	It is not helpful for adults to explain the reasons for rules to young children because they won't understand.*
	Children don't need to learn about numbers and math until they go to school.*
	Reading books to children is not helpful if they have not yet learned to speak.*

Table 1: Parenting Attitudes Scale items. *Indicates reverse coded items.

Although these previous studies provide preliminary evidence that parents' beliefs about parenting and child development affect their parenting behaviors, no previous research has attempted to identify the underlying *theories* that might organize their behavior and decision-making. Previous research has generally relied on observation of parent-child interactions or self-report of specific activities and behaviors. To our knowledge there is not an existing measure of parents' more general attitudes about parenting and child development, which might drive behavior and predict the uptake of interventions.

To address this gap, the present work establishes a self-report scale that captures adults' lay theories about child development and parenting. We generated a questionnaire measuring the degree to which parents endorse three potential lay theories: a "Rules and Respect" theory, an "Affection and Attachment" theory, and an "Early Learning" theory. As an initial test of the external validity of the questionnaire, we conducted an experiment to investigate whether parents' scores on these subscales would differentially predict their uptake of parenting information presented via a popular press article about children's early learning from free play (Gopnik, 2011). Higher scores on the Early Learning subscale predicted improved uptake in terms of both recall and generalization of information from the article about children's learning, but not a control article. Furthermore, the Rules and Respect subscale predicted improved uptake in terms of recall but not generalization of the presented information, suggesting that parents' implicit theories do affect how they process and use new information about child development, and that these differences may be detectable through self-report measures.

Scale Construction

To establish a new measure of parenting attitudes, we followed a structured plan based on psychometric best practices (Clark & Watson, 1995; Furr, 2011; Simms, 2008). We generated items corresponding to three hypothesized latent theories about parenting. The Early Learning theory corresponds to a view of children's early learning that is consistent with contemporary child development research, and includes the idea that young children can teach themselves by exploring and playing. The Affection and Attachment theory captures the notion that close parent-child relationships are important for development, and includes the ideas that parents should talk to their children about their emotions and that children are not spoiled by too much affection. The Rules and Respect theory corresponds to the idea that parents' primary role is to enforce rules and encourage behavior control. We generated hypothesized subscales and items based on a review of the literature on parenting attitudes, and conducted psychometric analyses on iterative samples of respondents collected on Amazon Mechanical Turk, both parents and non-parents (seven in total).

Item construction

In an initial phase of scale construction, we generated 42 statements that described attitudes consistent with one of three potential implicit theories about parenting: Early Learning (12 items), Affection and Attachment (10 items), and Rules and Respect (20 items). These statements were gen-

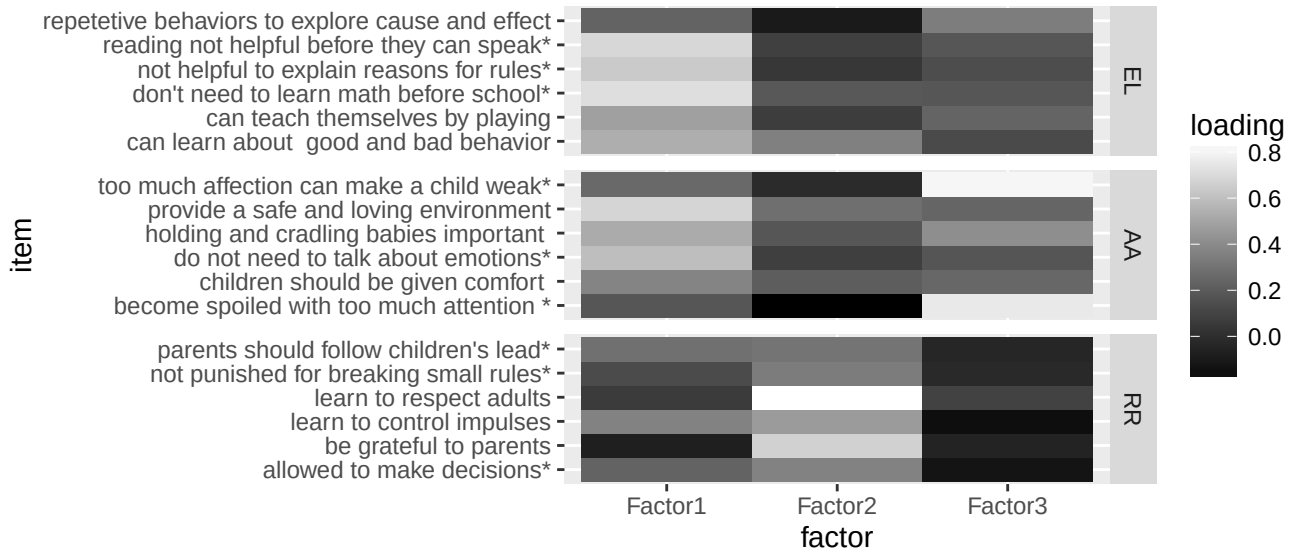


Figure 1: Factor loadings for subscale items. EL = Early Learning, AA = Affection and Attachment, RR = Rules and Respect. *Indicates reverse-coded items.

erated based on a literature review of parenting attitudes and behaviors. The Affection and Attachment and Rules and Respect subscales are related theoretically to the Authoritative and Authoritarian dimensions of Baumrind (1971)'s parenting framework, as well as theories of attachment parenting (Jones, Cassidy, & Shaver, 2014), but aim to assess beliefs about parenting rather than overt behaviors. The Early Learning subscale aimed to assess the extent to which adults believe that it is important to help infants and toddlers learn through play and conversation.

The initial 42-item scale was administered to 250 adults on Amazon's Mechanical Turk. Participants used a 7-point Likert scale to report the degree to which they agreed with each statement from 0 (Do not Agree) to 6 (Strongly Agree). Cronbach's alphas for the three subscales were .86 (Early Learning), .81 (Affection and Attachment), and .74 (Rules and Respect). We then conducted Exploratory Factor Analysis (EFA) to assess the dimensionality of the scale. Based on a parallel analysis (Horn, 1965), we retained 5 factors in this initial model. We subsequently dropped any items that had factor loadings less than .40 on the relevant factor, as well as any items that had factor loadings greater than .40 onto another factor. Items were also dropped if analyses revealed that Cronbach's alpha would be increased by dropping the item. We then dropped additional items with lower loadings until there were 6 items in each subscale. Some items were rephrased such that half of the items in each subscale were negatively worded to avoid response sets (Simms, 2008).

Revised questionnaire norming

The revised questionnaire was administered to a final group of 250 adults on Amazon's Mechanical Turk. Table 1 gives the full list of items[1]. For this sample, Cronbach's alphas were .76 (Early Learning), .75 (Affection and Attachment), and .69 (Rules and Respect). Because analysis of the previous sample identified 5 factors instead of the hypothesized 3, we again conducted EFA. This time, the parallel analysis identified 3 factors as predicted. We next examined the loadings of individual items onto the three factors (Figure 1). Items loaded onto the three factors roughly consistent with our a priori subscales, although some items from the Affection and Attachment subscale loaded onto both the Affection and Attachment and Early Learning factors.

Given the loadings of Early Learning and Affection and Attachment items onto a single factor, it is possible that participants' responses on these items were driven by a more general hands-on attitude towards parenting. Interestingly, when we conducted an exploratory analysis of factor loadings by gender, we found that female gender loaded highly positively onto the Affection and Attachment factor that did not include Early Learning items (.44), while the opposite was true for male gender (-.20). Thus, it is possible that these items reflect a theory related to Affection and Attachment that is separable from the hands-on approach and differentiated by gender.

As one test of the external validity of the subscales, we fit linear mixed-effects models predicting subscale scores based on gender and parenthood status (i.e., whether or not the participant reported having children) with participants as random effects. The model including gender revealed that females had higher scores on Early Learning, ($\beta = 0.41$, $SE = .19$, $p = .03$), and marginally higher scores for Affection and Attachment ($\beta = 0.34$, $SE = .23$, $p = .14$), whereas gender had no effect on Rules and Respect scores. The model including parenthood status revealed that parents had higher scores on the

[1]The most recent version of the questionnaire is available at https://github.com/langcog/parenting_proj

Uptake category	
Control recall	According to Article 2, what do the groups known for having many words for smells have in common? A) They train their children to be able to tell the difference between smells. B) They are hunter-gatherers.* C) They have unusually good hearing and vision.
Target recall	According to Article 1, children were more likely to find new and better ways to make a toy work when the experimenter: A) Pretended to be clueless about how the toy worked when playing with it.* B) Pretended to be an expert about the toy when playing with it. C) Pretended not to care about the toy.
Target generalization	Based on Article 1, preschool directors should plan to have: A) More structured time in which children are taught lessons and skills. B) More time spent memorizing things that will be helpful later, such as the ABCs. C) More time for children to play outside and with toys.*

Table 2: Examples of uptake questions. *Marks correct answer.

Affection and Attachment subscale compared to non-parents, ($\beta = 0.35$, SE = .13, p = .006), whereas parenthood status had no effect on Early Learning or Rules and Respect scores. These differences provide some intuitive support for the notion that subscale scores tap meaningful constructs that vary across groups.

Experiment

We next conducted an experiment to test whether scores on the three subscales would predict people's uptake of new information, as an initial test of the external validity of the scales. For this purpose, we had participants read two popular press articles: an article arguing that free play is beneficial to children's learning (Gopnik, 2011), and a control article about the language of smell (Yong, 2015). We operationalized uptake as accurate recall and generalization of the central message of the target article, and recall of the control article. We predicted that if people's subscale scores reflect coherent lay theories, they should differentially moderate uptake of the two articles. Specifically, we predicted that scores on the Early Learning subscale would be positively related to recall and generalization of the target article, but not recall of the control article. We predicted that scores on the Rules and Respect subscale would not predict uptake of either article. We excluded scores on the Affection and Attachment subscale from our analyses, since they are not orthogonal to scores on the Early Learning subscale. A sample analysis plan was pregistered at osf.io/6umza.

Methods

Participants Participants were 250 adults recruited from Amazon Mechanical Turk. Of these participants, 123 reported having children, 122 reported having no children, and 5 did not respond.

Procedure Participants first completed the 18-item parenting questionnaire described above. They were then instructed to read two popular press articles at their own pace. They first read a target article about children's early learning ("Why Preschool Shouldn't Be Like School," Alison Gopnik, Slate, 2011), and then read a control article about the language of smell ("Why Do Most Languages Have So Few Words For Smells?", Ed Yong, The Atlantic, 2015). After reading both articles, they first answered five 3-alternative forced-choice recall questions about the control article, and then answered five 3AFC generalization and five 3AFC recall questions about the target article. The recall questions assessed participants' memory for specific details of the text, and the generalization questions assessed participants' ability to generalize from the meaning of the text to new situations (Table 2). The alternative choices for recall and generalization questions were designed to reflect reasonable options that were nonetheless inconsistent with the articles.

Results and discussion

As a planned exclusion criterion, we excluded 48 participants who spent less than 30 seconds reading one or both of the articles, based on the assumption that they had not had time to read the whole article. Mean accuracy for the remaining sample was 0.67 (SD = 0.24) for control recall, 0.83 (SD = 0.19) for target recall, and 0.87 (SD = 0.22) for target generalization.

Figure 2 shows proportion correct on each of the three trial types, plotted by standardized subscale scores. Regression lines show the relationship between subscale scores and proportion correct. Qualitatively, there was not strong evidence of a relationship between Rules and Respect scores and performance (though target recall showed some trend). In contrast, we observed a stronger relationship between performance and Early Learning scores, especially for target generalization trials.

To quantify these trends, we fit a generalized logistic mixed-effects model with standardized Early Learning and Rules and Respect subscale scores, as well as interaction

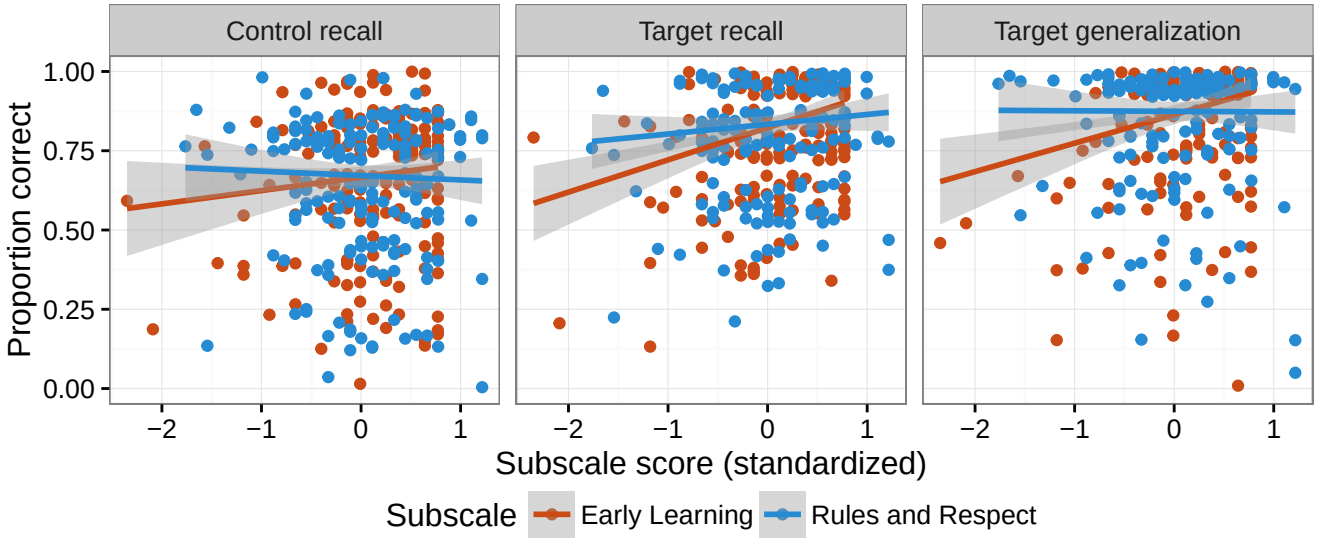


Figure 2: Relationship between subscale scores and uptake. Dots show individual participants. Lines show linear models and 95% CIs. Points are jittered to avoid overplotting.

Predictor	Estimate	Std. Error	z value	p
Cntl recall	1.11	0.37	2.98	0.00
Targ gen	1.57	0.49	3.19	0.00
Targ recall	1.28	0.49	2.59	0.01
Rules and Respect	-0.10	0.21	-0.48	0.63
Early Learning	0.27	0.22	1.22	0.22
Targ gen × R/R	-0.06	0.06	-0.94	0.35
Targ recall × R/R	0.32	0.06	5.62	0.00
Targ gen × EL	0.62	0.06	10.86	0.00
Targ recall × EL	0.49	0.05	8.97	0.00

Table 3: Fixed-effect coefficients for linear mixed effects model predicting task performance in Experiment 1. EL = Early Learning; RR = Rules and Respect.

terms for the subscale scores by question type (i.e., control recall vs. target recall vs. target generalization) as fixed effects, and participants and questions as random effects.[2] Coefficient estimates are shown in Table 3. Early Learning scores interacted positively with target recall and target generalization to predict correct performance, whereas Rules and Respect scores interacted positively with target recall but not target generalization, and coefficient magnitudes were relatively lower. Thus, Early Learning scores differentially predicted generalization from the target passage.[3] These results suggest that people's implicit theories about child development may organize the processing of new information in this domain.

[2]This random effects specification was the maximal convergent random effect structure.

[3]These results held when Affection and Attachment items were included in the analysis as well. Affection and Attachment scores also moderated generalization, though not as strongly as Early Learning scores.

One possible explanation for these results is that participants who scored higher on the Early Learning subscale would have been more likely to respond correctly to the target generalization and recall questions even without having read the article. For example, participants with high Early Learning scores might have pre-existing knowledge that is consistent with the information presented in the target article, or might intuitively answer the questions based on their implicit theories of children's learning. To address this possibility, we conducted a second experiment in which roughly half of the participants read the articles (n = 250), and the remaining half answered the recall and generalization questions without having read the articles (n = 229). We found that even among participants who had not read the articles, those who scored higher on Early Learning also scored higher on target recall, β = .49, SE = .04, z = 12.47, p<.001, and generalization, β = .30, SE = .04, z = 8.00, p<.001. However, we also found a three-way interaction such that accuracy was higher for target generalization trials for participants who had higher Early Learning scores and were in the reading group, β = .38, SE = .07, z = 5.61, p<.001. Thus, there is an advantage for uptake of an article about children's learning for participants who score higher on the Early Learning subscale of our questionnaire. Additionally, the fact that participants with higher Early Learning scores have higher accuracy for the target questions even without having read the article suggests that subscale scores are predictive of how parents reason about children's learning in general.

General Discussion

Understanding the sources of parents' behaviors and decisions with regard to their parenting is a critical step in improving children's welfare. There is a large literature outlining the activities and environments that can make a difference

in children's development, including their language (Hart & Risley, 1995) and executive functioning (Barker et al., 2014). Relevant research findings can sometimes be nuanced and unintuitive for those outside of academia, however. Thus, an important challenge is understanding the lay beliefs parents may have about parenting, and how they relate to parenting best practices as identified by developmental science. Interventions that aim to deliver new information may be more likely to succeed if they take into account parents' lay beliefs about child development (Kumar et al., 2015).

In the present work, we established a new scale to measure people's attitudes about parenting and child development in three categories: Rules and Respect, Affection and Attachment, and Early Learning. These subscales are meant to capture meaningful differences in how people view child development and the relative importance of different parenting behaviors. We subjected our new scale to psychometric testing, and found acceptably high correlations among subscale items, as well as the predicted factor structure across subscales. In addition, subscale scores meaningfully predicted responses to questions about children's learning, and uptake of new information on the topic. Specifically, participants with high scores on the Early Learning subscale were more likely to generalize the message of the target article about children's learning to new scenarios, whereas high scores on the Rules and Respect subscale did not predict generalization of the article.

In sum, this work provides initial evidence that meaningful differences in adults' attitudes about child development and parenting can be assessed by our new scale. This initial evidence for the reliability and validity of our parenting measure must be supplemented with further evidence in both areas across a broader range of participants. In addition, future work should target predictive validity by determining whether subscale scores differentially predict parents observable behaviors with their children, such as the quality of conversations they engage their child in, which would provide additional support for our scale.

Implicit theories are a powerful driver of human behavior. Sometimes, the interaction between interventions and their subjects' underlying beliefs can produce powerful, non-linear results (Medin & Bang, 2014). Thus, given both the variability in attitudes towards parenting across cultures and the importance of improving parenting outcomes, it behooves us to understand implicit theories of parenting. The current work takes a first step in this direction.

Acknowledgements

This work was supported by a gift from Kinedu, Inc. Thanks to members of the Language and Cognition Lab at Stanford for helpful discussion.

References

Barker, J. E., Semenov, A. D., Michaelson, L., Provan, L. S., Snyder, H. R., & Munakata, Y. (2014). Less-structured time in children's daily lives predicts self-directed executive functioning. *Frontiers in Psychology*, *5*, 323.

Baumrind, D. (1971). Current patterns of parental authority. *Developmental Psychology*, *4*(1p2), 1–103.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Childrens imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273.

Dweck, C. S., Chiu, C.-y., & Hong, Y.-y. (1995). Implicit Theories: Elaboration and Extension of the Model. *Psychological Inquiry*, *6*(4), 322–333.

Furr, M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE Publications Ltd.

Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H Brookes Publishing Company.

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337–374.

Jones, J. D., Cassidy, J., & Shaver, P. R. (2014). Parents Self-Reported Attachment Styles A Review of Links with Parenting Behaviors, Emotions, and Cognitions. *Personality and Social Psychology Review*, *19*(1), 1088868314541858–76.

Kumar, V., Kumar, A., Ghosh, A. K., Samphel, R., Yadav, R., Yeung, D., & Darmstadt, G. L. (2015). Enculturating science: Community-centric design of behavior change interactions for accelerating health impact. *Seminars in Perinatology*, *39*, 393–415.

LeVine, R. A. (2004). *Challenging Expert Knowledge: Findings from an African Study of Infant Care and Development.* Praeger Publishers/Greenwood Publishing Group.

Medin, D. L., & Bang, M. (2014). The cultural side of science communication. *Proceedings of the National Academy of Sciences*, *111*(4), 13621–13626.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Richman, A. L., Miller, P. M., & LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology*, *28*(4), 614–621.

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(01), 185–205.

Simms, L. J. (2008). Classical and Modern Methods of Psychological Scale Construction. *Social and Personality Psychology Compass*, *2*(1), 414–433.

Yong, E. (2015). Why do most languages have so few words for smells? *The Atlantic*.

Emotions in lay explanations of behavior

Desmond C. Ong (dco@stanford.edu)
Jamil Zaki (jzaki@stanford.edu)
Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, Stanford University, Stanford CA, USA

Abstract

Humans use rich intuitive theories to explain other people's behavior. Previous work in lay psychology of behavior have tended to treat emotion as causing primarily unintentional behavior (e.g., being sad causes one to cry), neglecting how people incorporate emotions into explanations of rational, intentional actions. Here, we provide preliminary explorations into integrating emotions into a theory of folk psychology. Specifically, we show that in the lay theory, people are willing to endorse emotions as causes of intentional actions. Moreover, people readily attribute beliefs and desires as explanations for emotional expressions. This work provides a first step in elaborating people's rich understanding of emotions as an important component of intuitive social cognition.

Keywords: Intuitive Psychology, Emotions, Affective Cognition, Explanations

"... men in rage strike those that wish them best" — Iago (Shakespeare, trans. 1996, 2.3.205)

In his plays, the Bard of Avon makes masterful use of the layperson's understanding of emotion: in *Othello*, the audience is privy to Iago's deliberate manipulation of Othello's jealousy and rage, and can effortlessly predict Othello's murder of Desdemona before it happens. Consider the following counterfactual scenario: would Othello still have killed Desdemona had he not been feeling those emotions? This seems unlikely. Intuitively, emotions played a key, irreplaceable role in Othello's decision making process.

We have a rich intuitive understanding of others and their emotions (Ong, Zaki, & Goodman, 2015), and extend this understanding to explain others' behavior. Reasoning about others' mental states and behaviors is often called folk, intuitive, or lay psychology (e.g., Heider, 1958; Malle, 2011). Many modern theories posit a "belief-desire psychology" (e.g., Bartsch & Wellman, 1995; Dennett, 1989; Gopnik & Meltzoff, 1997; Malle, 1999, 2011), in which an agent has a set of goals (**desires**), and a set of ideas about the world that help them understand how to achieve those goals (**beliefs**). The agent then forms an intention to act upon his beliefs to achieve his desires, resulting in **intentional action**. When laypeople are asked to explain an agent's behavior, they often appeal to both the beliefs and desires of the agent: "Sue went to the store, because she wanted a drink, and she thinks that the store sells drinks". **Unintentional behavior**, on the other hand, are often described as a result of **situational factors** via physical causality (or *impersonal causality*; Heider, 1958): "Sue slipped and fell because there was ice on the floor (not because she intended to)". This taxonomy[1] has proven fruitful in describing how laypeople explain behavior.

Previous theories relegate emotions to the bin of other, situational causes—feeling sad simply makes one cry. There is an implicit assumption that emotion-driven behavior is unintentional, or otherwise "irrational". See Figure 1a for an illustration, where we specifically distinguish **Emotions** (e.g., sadness) from **Situational Factors** (e.g., the ground is icy). Additionally, although most theories do not differentiate **Emotional Expressions** (e.g., crying, laughing) from **Unintentional Behavior** (e.g., slipping on ice, snoring), we separate them in this model. This model is likely an insufficient description as it does not account for how lay people use emotions in causal explanations of intentional behavior, such as how an agent's emotional state might influence the formation of intentions. Returning to the example of Othello, we argue that even his (false) belief (that Desdemona cheated on him) and his desire (to seek justice by punishing her) alone, without any rage, would not have led him to murder his wife.

How can we incorporate emotions into an intuitive psychology of behavior? In this paper, we explore two questions. First, are intentional actions actually independent of emotions? It is worth noting that many scientific theories of emotion include both "automatic" and "intentional" behavior as crucial parts of their definitions of emotion. On one hand, emotions cause characteristic "automatic" behavioral responses like facial expressions and vocalizations (e.g., Ekman, 1992). On the other hand, emotions also bias agents towards certain types of actions via *action tendencies* (Frijda, Kuipers, & Ter Schure, 1989; Fontaine, Scherer, Roesch, & Ellsworth, 2007) or approach/avoid motivations (e.g., Carver & Scheier, 2004). For example, being in a state of happiness predisposes one towards helping others (Isen & Levin, 1972) and taking risks (Isen & Patrick, 1983). Emotions have been incorporated into psychological (e.g., Schwarz, 2000), economic (e.g., Loewenstein & Lerner, 2003), and philosophical (e.g., Zhu & Thagard, 2002) theories of behavior, but they are still lacking from theories of lay psychology. Consider Figure 1b, which posits an intuitive theory in which emotions directly influence the intentional decision making process—in the Discussion, we return to how exactly this might occur.

Second, in the intuitive theory, are emotion-caused actions—specifically, emotional expressions—independent of beliefs and desires? Intuitively, this seems unlikely. For one, emotional display rules dictate what expressions are more appropriate in some contexts rather than other (e.g., Matsumoto, 1990). Display rules also subsume prosocial

[1]There are also other explanations that people give for intentional actions, such as upstream events that result in beliefs or desires (Malle, 1999's "Causal History of Reasons") and events that facilitated but not instigated the action (often called Enabling Factors). We do not discuss them in this paper.

or politeness considerations: presumably, one might want to force a laugh at a boss' joke, or hide their joy from a recent promotion when comforting a friend who has just lost their dog. Alternatively, crocodile tears and other deliberately deceptive strategies (e.g., DePaulo et al., 2003) are a common, if distasteful, intentional use of emotional expressions as a means to obtain some desired outcome. Indeed, emotional expressions form an interesting category of behavior that does not fall neatly into the "Intentional" vs. "Unintentional" dichotomy, as there are clear examples of both. Figure 1c allows for the former possibility of an agent's beliefs and desires influencing their emotional expressions. Under this model, beliefs and desires can impact expressions (thus, expressions do not fall into "Unintentional"), but at the same time are not required for expressions (thus, they do not fall into "Intentional" either). The place of emotional expressions in the lay theory deserves further attention, and Figure 1c provides one possible interpretation.

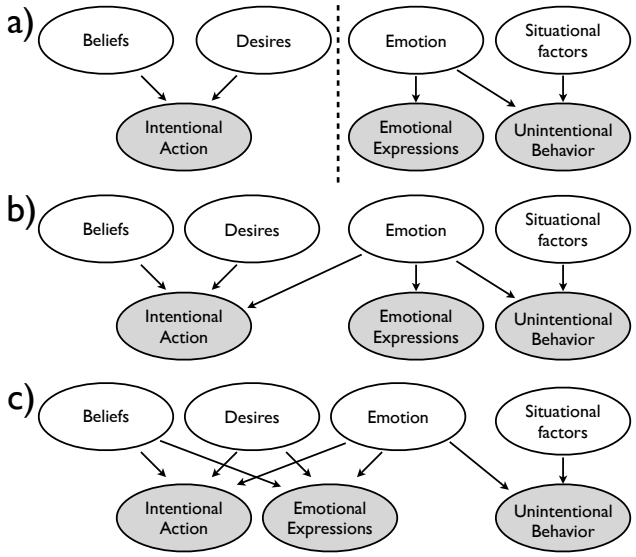


Figure 1: Possible lay theories of behavior. (a) Two distinct types of behavioral responses. Left: beliefs and desires influence an **intentional decision-making** process, resulting in intentional action. Right: unintentional behavior and emotional expressions are simply caused by situational factors and emotions, without the need for an agent's intentionality. (b) Emotions influence the intentional decision making process. (c) Beliefs and desires also influence emotional expressions.

In this paper we provide some preliminary explorations of how emotions are incorporated into intuitive belief-desire psychology, and specifically, how emotions are used or judged as explanations of intentional actions. The three possible models that we discussed in Fig. 1 each provides distinct, testable predictions. In Study 1, we study the types of actions that laypeople think emotions cause, and show that the top emotion-caused actions are judged to have been caused by emotions only about half of the time. In Study 2, we show that laypeople are willing to endorse emotions as causes of intentional actions. We also find that people are willing to endorse beliefs and desires as likely causes of emotional expressions (e.g., smiling), which have been treated as unintentional behavior in previous lay theories. This paper provides a first step in exploring how emotions are used by laypeople in explaining behavior, and we end by discussing future directions that are inspired by this research.

Study 1a: Emotions to Actions

In Study 1a, we asked participants to freely generate actions that would be likely caused by different emotions. We then obtained judgments of the counterfactual likelihood (i.e., how likely were the generated actions if the agent had not been feeling the given emotion). This allowed us to find the actions most likely to be "emotion expressions" (as described above) and to explore whether the emotions are causally necessary to bring about these actions (via the counterfactual judgments).[2]

Participants and Procedures. We recruited 100 participants (99% had English as their native language) through Amazon's Mechanical Turk (AMT). Participants saw statements of the form "Bob ____ because he was [**emotion**]", and were asked to give sentence completions. The presented **emotion** was one of: {happy, calm, angry, sad, surprised}[3], presented in a randomized order. Participants gave 5 different completions per emotion, for a total of 25 completions.

After participants had given completions to all 5 emotions, they were presented again with their answers, and asked to rate the likelihood of the counterfactual: "You wrote that '*Bob [cried] because he was [sad]*'. If Bob was not feeling [sad], would he still have [cried]?" Participants gave responses on a 7 point Likert scale from "Very Unlikely" to "Very Likely".

Results. Two coders (the first author and a naive coder) independently coded the free-responses into prototype actions, by removing adverbs and modifiers, and grouping by synonyms (e.g., "smiled", "smiled widely", and "beamed"). There was agreement on 99.6% of the responses; disagreements were resolved by consensus. The top 5 response groups for each emotion is given in Table 1. The 900 responses in this Table make up 36% of all responses. As expected, the majority of these modal responses would easily be judged to be emotional expressions: some notable exceptions are "killed himself", "punched the wall", "slept", and "sat down".

We turn next to the counterfactual rating task, to investigate necessity of emotion for the elicited actions. First, we note that the mean counterfactual ratings (i.e., how likely is the action to occur if the emotion was absent, $P(A|\neg E)$) for the modal responses (M(SD)=3.08(1.92)) was significantly lower

[2] All studies, data, and coding analyses are available at: https://github.com/desmond-ong/shakespeare/

[3] We chose a high-arousal positively valenced ("happy"), a low-arousal positively valenced ("calm"), a high-arousal negatively valenced ("angry"), a low-arousal negatively valenced ("sad"), and a high-arousal neutral valenced ("surprised") emotion.

Happy	**Calm**	**Sad**	**Anger**	**Surprised**
smiled: 73	relaxed: 52	cried: 97	"hit X": 69	jumped: 65
laughed: 52	slept: 46	frowned: 17	yelled: 56	laughed: 41
jumped: 42	sat down: 34	slept: 14	screamed: 27	screamed: 30
danced: 23	smiled: 26	killed himself: 13	cried: 16	yelled: 26
cried: 21	sighed: 10	yelled: 13	cursed: 14	smiled: 23

Table 1: Study 1a Results: Top 5 responses for each emotion, with frequency counts. The most common responses for anger were variants of "hit X", where X is an object or person (of these, the modal was: "punched the wall").

than that of the non-modal responses (M(SD)=3.32(1.85); t=3.14, $p < .001$): people judged the modal responses to be less likely to occur in the absence of the emotion, than the non-modal actions. Additionally, we hypothesized that the more often a particular action a is generated *across all participants* for emotion e, the more *necessary* is emotion e for action a to occur. To test this, we regressed participants' counterfactual likelihood ratings against the frequency of that action being generated across the sample, with random intercepts by participant and emotion. We find that, across all emotions, the more frequently an action is generated by all participants, the less likely participants rate it to occur in the absence of the corresponding emotion ($b = -0.0083, 95\%$ CI : $[-0.0106, -0.0060], t = -7.15, p < 0.001$; See Fig. 2). Thus, in participants' lay theories, the corresponding emotions are more necessary for these modal actions to occur: we shall use this result in Study 1b.

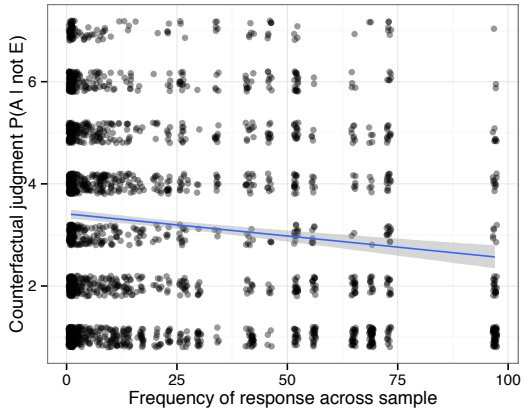


Figure 2: Study 1a Results. Individual counterfactual likelihood ratings ($P(A|\neg E)$) against frequency of an action's generation across the sample. Data points are rendered transparent and jittered for clarity.

Study 1b: Causes of Actions

In Study 1b, we recruited a separate group of participants to give explanations for the modal actions generated from Study 1a. This allows us to explore the sufficiency of emotions to cause the actions elicited above, and to test among causal models of "emotional expressions." If Model 1 (Fig. 1a) is correct then in Study 1b we should only find explanations due to emotions (and perhaps, situational factors), and not explanations that appeal to beliefs or desires.

Participants and Procedures. We recruited 100 participants through AMT (98% had English as their native language). Participants saw statements of the form "Bob [**action**] because ____", and were asked to complete the sentence. The presented **action** was one of the fifteen actions drawn from the most popular responses from Study 1a (the 15 unique actions in Table 1, with "punched the wall" used in place of "hit X"). Participants saw the actions in a random order and gave 1 completion per action.

Results. Two coders classified the free-response completion into one of five categories. The two main categories of interest are: (a) "**Emotion**" (if the explanation contained an explicitly mentioned emotion), and (b) "**Cause of Emotion**" (if there was a mention of an event that is very likely to cause an emotion, e.g., "his dog died"). We explicitly added a category of coding for Causes of Emotion, because laypeople often give events that caused emotions as explanations for emotional displays. For example, if someone asks, "why is he crying?", "because he is sad" is a somewhat unsatisfying explanation (because presumably, laypeople find it obvious), as compared to, "because his dog died (and hence he is sad)". We also coded for (c) "**Physical state**" (if the explanation referenced a physical state like tiredness or pain), (d) "**Mental state**" (if the explanation referenced a desire or a belief); (e) "**Situation**" (for other situation factors that does not cause emotions). The two coders achieved a high agreement (Cohen's $\kappa = 0.935$ over 1500 responses), and the responses with disagreements were discarded ($< 5\%$).

The distribution of coded free-responses is given in Figure 3. If the actions generated from Study 1a were characteristic of those emotions, and if they could only have been caused by emotions (i.e., Fig. 1a), then we should expect the vast majority of the explanations to be due to emotions or to upstream causes of emotions. However, only 39.8% of explanations appeal to an Emotion, and 50.9% appeal to either an Emotion or to a Cause of Emotion. There are many references to Physical states (17.7%; e.g., physical pain as a cause for "cursing" or "yelling"; fatigue as a cause of "sat down" or "slept") and Mental states (10.6%; e.g., a belief or a desire). Thus, it seems very likely that lay people's judgments of emotion-caused actions are not, in fact, predominantly caused only by emotions. The results of Studies 1a-b provides preliminary support for the model in Figure 1c, which does not distinguish intention-caused from emotion-caused actions. We test this model more precisely in Study 2.

Study 2: Rating explanations for actions

In order to further test the models in Figure 1, we explored a broader set of actions in Study 2. We chose a set comprising Intentional Actions, Emotional Expressions, and Unin-

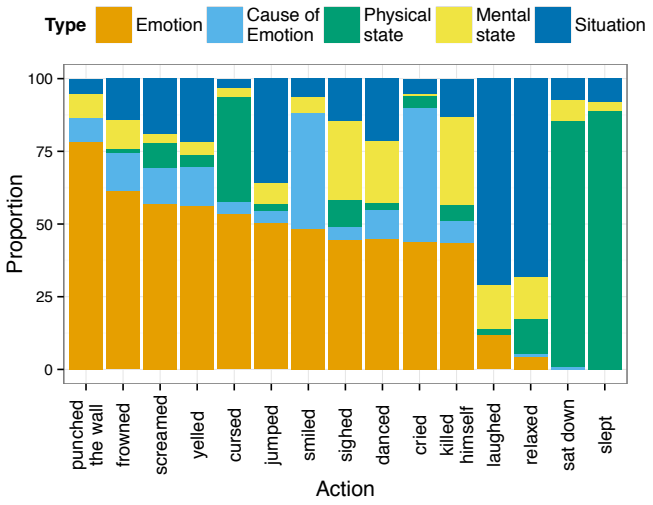


Figure 3: Study 1b Results: Distribution of explanation types.

tentional Behavior, and asked participants to rate how likely it was that each action was caused by Belief, Desire, Emotion, and Situational Factor explanations[4]. First, we aimed to determine if people judge emotions as suitable explanations for intentional behavior. Second, we aimed to confirm the finding from Study 1 that people will judge beliefs and desires as suitable explanations for emotional expressions. Specifically, we would predict, under the model in Figure 1c, that for Intentional Actions, beliefs and desires will be rated as the most likely causes, followed by emotion. For Emotional Expressions, emotion explanations will be rated as the most likely causes, followed by ratings of beliefs and desires. For Unintentional Behavior, we predict high ratings for situational factor causes and low endorsements for other causes.

Stimuli Selection. We selected a set of 20 actions, comprising 6 intentional actions, 6 emotional expressions, 5 unintentional behaviors, and 3 ambiguous behaviors (given in Fig. 4.) Of the 6 intentional actions, we chose 3 ("stole a pound of peaches"; "invited Sue to have lunch"; "watered his new plants") from Malle (1999) as they had been previously rated as highly intentional.We chose 6 emotional expressions from the modal responses of Study 1a. We used two unintentional behaviors ("yawned during a lecture", "won the lottery") from Malle (1999), as they had been rated as being low on intentionality. Finally, we chose 3 "ambiguous" behaviors ("interrupted his mother"; "ignored Greg's arguments"; "drove above the speed limit") from Malle (1999), which had been rated by some participants as being intentional, and others as being unintentional. We predicted that ambiguous behaviors would elicit uniform endorsements across all explanation types.

[4]Although the three models do not differentiate between the causes of Unintentional Behavior, we included Unintentional Behavior and Situational Factors in order to calibrate against baseline predictions that all three models should make.

Participants and Procedures. We recruited 100 participants through AMT (98% had English as their native language; 1% did not report). First, participants were told about different types of explanations, with examples of each. (1) People have **thoughts** or beliefs about the way the world is that make them behave so ("Bob moved to Iowa because he **thinks** people are nice there"); (2) People feel certain **Emotions** that make them behave so, ("Bob ran away because he was **feeling** scared"); (3) People behave that way to achieve certain **Aims** ("Bob kicked the ball because he **wanted to** win the game"); and (4) People behave that way because of the **Situation** ("Bob shivered because it was cold outside").

Next, participants saw statements of the form "Bob [**action**] because ...", and descriptions of four explanation types: ("... he was motivated by some thoughts or beliefs about the way the world is"; "... he felt some emotions"; "... he wanted to achieve some aims of his"; "... of some situational causes"). They then used continuous 100 point sliders to rate how likely it was that the behavior was caused by each type of cause, from "Not at all likely" to "Extremely likely". Participants saw the 20 actions in a random order.

Results. We used k-means clustering to cluster participants' responses into "explanation profiles". We took each participant's response to each action as a 4-dimensional vector corresponding to the 4 ratings they gave (**b**elief, **d**esire, **e**motion, **s**ituation). Although we had an *a priori* prediction that there would be $k = 4$ clusters, we ran bootstrap clustering stability analyses (Hennig, 2007) that revealed that solutions with 5 or more clusters were unstable across nonparametric resampling. Further analyses confirmed that $k = 4$ was the maximal stable solution. Indeed, visual inspection of the centroids of the 4 clusters (Fig 4, left column) verified that the stable cluster solutions correspond to our predicted explanation profiles.

The 4 cluster profiles look similar to our predictions: an Intentional profile (**b**=26.7, **d=79.5**, **e**=20.0, **s**=38.7) predominantly driven by Desire ratings, an Emotion Expressions profile (23.6, 20.5, **83.8**, 56.1) predominantly driven by Emotions, an Unintentional profile (4.8, 6.9, 7.2, **87.7**) driven by Situation causes, and an Ambiguous profile (57.1, 63.6, 66.6, 61.1) resulting from uniform ratings across all explanation types. First, we note that the Intentional profile seems to be driven primarily by only Desire, and that Belief seems to be only weakly associated with it. Second, we note that there is a sizable amount of rating of Emotions in the profile for Intentional Actions, with an average of **e**=20.0 points on a 100 point scale (just slightly behind Beliefs at **b**=26.7). We also note that there is a sizable endorsement of Beliefs (23.6) and Desires (20.5) in the Expressions profile. The Ambiguous profile is an interesting case that result from variations in participants' appraisal. Some participants may judge a given action as Intentional; others, Unintentional; and yet others, as Expressions.

Participants' ratings for the causes of different actions are shown in Figure 4. In each row, we show the actions that were assigned to the prototype cluster on the left. While the un-

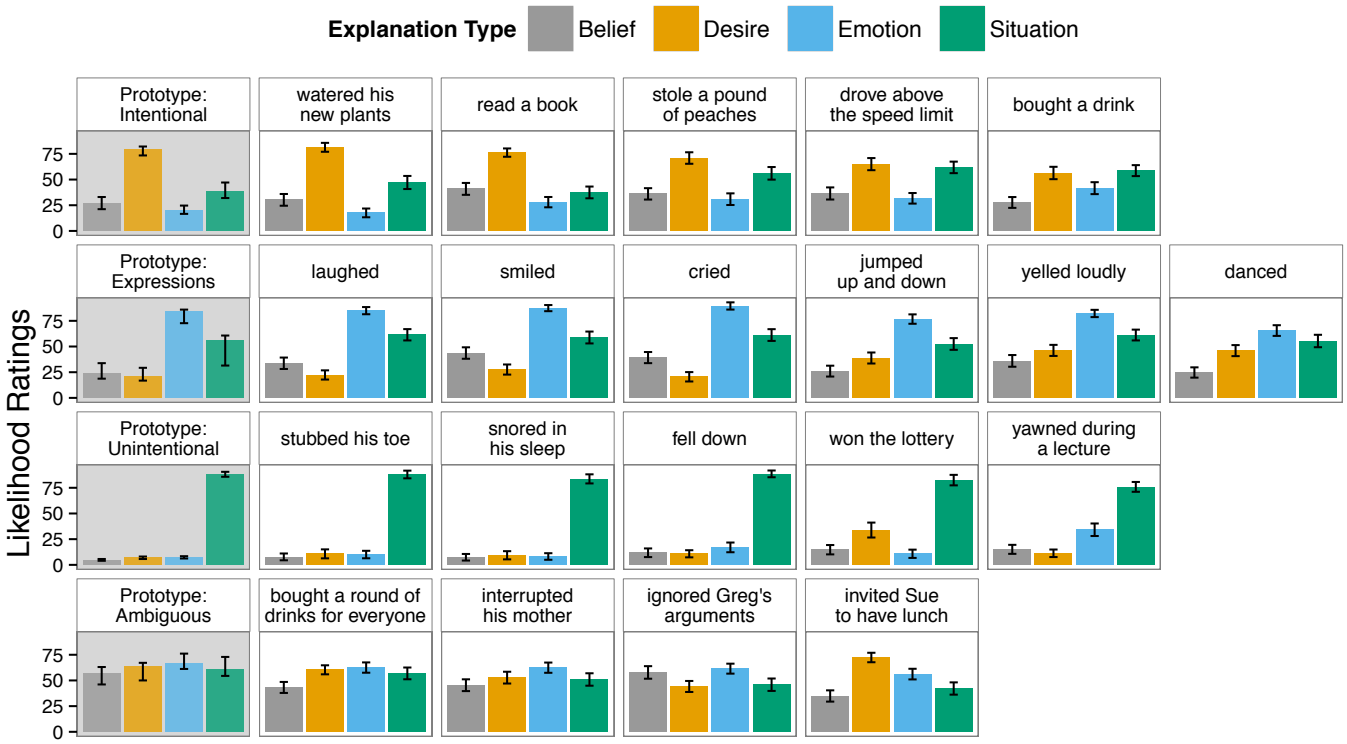


Figure 4: Study 2 Results. Mean ratings, with 95% confidence intervals (CIs), of how likely each explanation type was to have caused the action. Left most column: cluster prototypes from clustering analyses, with CIs obtained from 100 bootstraps.

supervised clustering for the Expressions and Unintentional Behavior matched our *a priori* predictions exactly, we note that there were a few actions that were not classified according to our initial hypothesized groupings. "Drove above the speed limit" fit more closely with the Intentional profile than the Ambiguous profile, while "bought a round of drinks for everyone", and "invited Sue to have lunch", more closely resembled the Ambiguous profile than the Intentional profile.

Overall, these results support our predictions of an intuitive theory with different types of actions, as evidenced by different clusters of explanation profiles. Importantly, there is a distinct profile for Emotional Expressions, and a distinct role of Emotion causes in Explanations. Finally, we visualized the explanation profiles using multidimensional scaling in two dimensions (Fig. 5). We see that the Intentional, Ambiguous, and Expressions clusters seem to fall along a spectrum (from top to bottom), suggesting some structure in a higher-dimensional semantic space that future work should verify and explore.



Figure 5: Multidimensional scaling of explanation profiles, with cluster centroids marked with dots. Clockwise from top: Intentional, Ambiguous, Expressions, Unintentional

Discussion

Laypeople have rich, intuitive theories of emotion that they use to explain many types of behavior. By contrast, emotion has tended to be neglected in work on intuitive theories. Using both an unstructured free-response, sentence completion task (Study 1a/1b) and a structured explanation rating task (Study 2), we show that people judge beliefs and desires to be explanations of emotional expressions, and endorse emotions as causes of intentional actions. Our results suggest the need to expand the belief-desire model of lay psychology to capture how people reason about emotion-driven and emotion-influenced actions, and provides a first step in that direction.

Measuring lay theories is difficult, as one must elicit par-

ticipants' judgments without imposing too much of the researcher's own bias. We tried to mitigate this by using two approaches, but limitations remain. The space of emotions and the space of actions (intentional, emotional expressions, and unintentional) are both large; by focusing on a small set of emotions and actions, we risk the chance of conclusions being driven by idiosyncratic emotion or action choices. We elicited a wide variety of actions in Study 1a, but had to focus on modal responses for analysis. Some of the modal actions that we observed (e.g., *anger* / "punched the wall") may be driven by cultural tropes, and may not be representative of real-life actions. The approach in Study 2 allowed us to examine what types of explanation profiles people attribute to a set of actions. Again, the set of actions was necessarily restricted, this time by experimenter selection. Future work should aim to achieve an unbiased yet representative sampling of actions. Extending the approaches taken here to may lead to new and more precise ways to measure lay theories.

Many questions remain about how exactly emotions affect intentional actions in the lay theory. Borrowing ideas from affective science, emotions might impact an agent's *beliefs*, by biasing their subjective judgments of probability (Wright & Bower, 1992) or influencing the processing of novel information (Forgas, 1995). Alternatively, emotions might influence *desires*, by introducing new goals via approach/avoid motivations (Carver & Scheier, 2004), or by introducing emotional states as regulatory goals in and of themselves (Gross, Richards, & John, 2006). Finally, emotions might have a direct impact on intentional *action* that is independent of beliefs and desires; this possibility might be needed for lay explanations of "rash" decisions (such as crimes of passion, like Othello). Any of these causal pathways—or all of them—may be part of the lay theory of psychology. Future work should address these possibilities.

We have focused on lay explanations of behavior, especially those that go beyond beliefs and desires. This work builds towards a larger research program on how humans use rich intuitive theories of emotion to reason about others—what we call Affective Cognition (Ong et al., 2015). These intuitive theories have broad and wide-reaching impact on all forms of social cognition, from understanding family, friends and colleagues, to making attributions in moral and legal judgments. And of course, they help us enjoy a little theatre.

Acknowledgments

This work was supported by an A*STAR National Science Scholarship to DCO, ONR grant N00014-13-1-0788, and a James S. McDonnell Foundation Scholar Award to NDG.

References

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford university press.

Carver, C. S., & Scheier, M. F. (2004). Self-regulation of action and affect. In *Handbook of self-regulation: Research, theory, and applications* (pp. 13–39).

Dennett, D. C. (1989). *The intentional stance*. MIT press.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, *129*(1), 74.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, *18*(12), 1050–1057.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (aim). *Psychological bulletin*, *117*(1), 39.

Frijda, N. H., Kuipers, P., & Ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of personality and social psychology*, *57*(2), 212.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. MIT Press.

Gross, J. J., Richards, J. M., & John, O. P. (2006). Emotion regulation in everyday life. In D. K. Snyder, J. Simpson, & J. Hughes (Eds.), *Emotion regulation in families: Pathways to dysfunction and health* (pp. 13–35). Washington DC: American Psychological Association.

Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of personality and social psychology*, *21*(3), 384.

Isen, A. M., & Patrick, R. (1983). The effect of positive feelings on risk taking: When the chips are down. *Organizational behavior and human performance*, *31*(2), 194–202.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In *Handbook of Affective Science* (p. 619-642). Oxford University Press.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and social psychology review*, *3*(1), 23–48.

Malle, B. F. (2011). Attribution theories: How people make sense of behavior. In *Theories in social psychology* (pp. 72–95). Wiley-Blackwell Malden, MA.

Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and emotion*, *14*(3), 195–214.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, *14*(4), 433–440.

Shakespeare, W. (1996). *Othello* (3rd ed.). Bloomsbury Arden Shakespeare. (Ed: EAJ Honigmann)

Wright, W. F., & Bower, G. H. (1992). Mood effects on subjective probability assessment. *Organizational behavior and human decision processes*, *52*(2), 276–291.

Zhu, J., & Thagard, P. (2002). Emotion and action. *Philosophical psychology*, *15*(1), 19–36.

But vs. *Although* under the microscope

Fatemeh Torabi Asr[1] and Vera Demberg[2]
[1]Indiana University, Bloomington (IN), USA
fatorabi@indiana.edu
[2]Saarland University, Saarbruecken, Germany
vera@coli.uni-saarland.de

Abstract

Previous experimental studies on concessive connectives have only looked at their local facilitating or predictive effect on discourse relation comprehension and have often viewed them as a class of discourse markers with similar effects. We look into the effect of two connectives, *but* and *although*, for inferring contrastive vs. concessive discourse relations to complement previous experimental work on causal inferences. An offline survey on AMTurk and an online eye-tracking-while-reading experiment are conducted to show that even between these two connectives, which mark the same set of relations, interpretations are biased. The bias is consistent with the distribution of the connective across discourse relations. This suggests that an account of discourse connective meaning based on probability distributions can better account for comprehension data than a classic categorical approach, or an approach where closely related connectives only have a core meaning and the rest of the interpretation comes from the discourse arguments.

Introduction

There exists a substantial body of research on the processing of discourse connectives like *because*, *but* or *nevertheless* and their role in facilitating the linking of text segments to one another (Kintsch & Van Dijk, 1978; Just & Carpenter, 1980; Millis & Just, 1994; Murray, 1995, 1997; Sanders & Noordman, 2000; Rohde & Horton, 2014), as well as a discourse connective's effect in helping comprehenders anticipate and integrate upcoming content (Köhne & Demberg, 2013; Xiang & Kuperberg, 2014; Drenhaus et al., 2014). Most studies have however compared connectives that substantially differ in meaning, e.g., additives vs. contrastives, or causals vs. concessives. The present study contributes to the existing body of experimental research as well as theoretical framing of the meaning of connectives by investigating the effect of two connectives, *but* and *although*, that can mark the same set of relations, but differ in their distribution of how frequently they are used to mark these relations.

A first question that this study addresses is therefore what effect these connectives have on the meaning of the overall discourse. Do they affect the interpretation of the arguments themselves (instead of just signaling the relation), and if yes, how can this be accounted for? We already know that connectives like *because* may trigger causal inferences, which would not be made solely based on the meaning of the related clauses (Noordman & Vonk, 1992; Millis et al., 1995; Traxler et al., 1997). Causal connectives like *because* have been argued to carry truth-conditional meaning, i.e., to extend to the set of statements entailed by the sentences they combine (in the case of *A because B*, this additional statement would consist of *B being the cause of A*, on top of the facts asserted by the arguments, A and B separately). Some researchers distinguish between connectives such as *because* and *before* that affect the truth-conditional state of what is said, and connectives regarded as non-truth conditional, such as *but* and *furthermore*.

Note though that non-truth-conditional connectives like *although* can trigger implications (Blakemore, 2002). For example, a sentence in the experiment run by Noordman & Vonk (1992), "Chlorine compounds make good propellants because they react with almost no other substances.", entails that *propellants must not combine with other substances*. A connective like *although* (instead of *because*) in the same sentence would imply that *propellants should combine with some substances*. [1] Our study investigates the effects of the connectives *but* and *although* on interpretation of the two arguments of a discourse relation, and discusses the resulting effects on the integration of upcoming content.

A second important point of this study is addressing the ambiguity of connectives *but* and *although*. We know from large discourse relation annotated corpora such as the PDTB (Prasad et al., 2008) that almost all discourse connectives are ambiguous (Asr & Demberg, 2012, 2013): while some connectives such as *since* are ambiguous between highly different relations (temporal vs. causal), others, like *and* can be present with almost any discourse relation. Existing proposals for describing the meaning of the discourse connectives *but* and *although* have considered sentence pairs such as (1), which seem to be fairly similar in meaning. Fraser (1998) suggests to describe each connective in terms of its *core meaning*, with further effects of inference being attributed to the discourse arguments. This approach boils down to assignment of an under-specified meaning to both connectives.

(1) She fried the onions, **but / although** she steamed the cabbage.

[1]Interested readers are referred to Grice (1975) and Bach (2006) for detailed discussions on entailment vs. implication (implicature).

Alternatively, the meaning of a connective can be described based on its usage, i.e., extending distributional semantics accounts to discourse connectives, such that each connective potentially has more than a single meaning, and inferences are subject to the probability distribution over the meanings.

The goals of the present paper are

- to better understand the effect of the connectives *although* and *but* on the interpretation of a text, and
- to test whether and how the distribution of different meanings of a connective affects comprehension.

We address these questions by a corpus study on the distribution of discourse relations with these connectives in PDTB as well as an offline coherence judgment task. Other experiments have shown that online and offline effects don't necessarily agree, due to shallow processing / lazy inferences during online comprehension (Noordman & Vonk, 1992). We therefore follow up the coherence judgment task with a reading study using eye-tracking.

Discourse relations marked by *but* and *although*

A first important difference to note between the connectives *but* and *although* is of course that they differ in syntax (*although* is subordinating while *but* is a coordinating connective). Additionally, they have been suggested to be subject to a *semantic asymmetry* (Blühdorn, 2008), which means that one argument is more salient or central to the discourse than the other one.

Previous work investigating the relations marked by the connectives *but* and *although* (e.g., König, 1991; Blakemore, 2002; Hall, 2004; Iten, 2000) has used different names to refer to the relevant discourse relations; in today's most well-known annotation schemes, the relation most typically expressed by *but* is known as "contrast", "antithesis" or "negative additive", depending on the scheme, while the relation most typically marked by *although* is variously known as "concessive", "negative causal", "concessive.expectation" or "violated expectation". We will therefore briefly define the relations we are interested in here, and will then proceed to a corpus analysis to assess their frequency of occurrence. A comparison between the interpretation of sentence pair (2-a) with that of (2-b) elaborates the inferences in which we are interested.

(2) a. **Although** she desired to have something savory with her drink, she took some cake from the fridge.
b. She took some cake from the fridge, **but** she desired to have something savory with her drink.
c. She took some cake from the fridge, **although** she desired to have something savory with her drink.

In (2-a), the first clause states that she desired something savory, which gives rise to the expectation that she'll take something savory from the fridge. The second clause however contrasts with this expectation, stating that she takes something sweet (cake). This segment can hence be classified as a `violated expectation` relation.

Definition: A `violated expectation` relation holds between two discourse segments wherever a discourse connective indicates that one of its arguments describes a situation *A* which causes *C*, while the other asserts or implies a state C' that contrasts with *C*. In other words, one argument of a `violated expectation` relation denotes a fact that triggers a set of potential consequences, while the other argument denies one or more of them.

In (2-b) on the other hand, a different reading is possible, where she takes the cake from the fridge, and then realizes that she wants something savory. In this case, there is no direct causal relation between the wanting something savory and taking something from the fridge. Hence, we analyze it as a `contrast` relation.

Definition: A `contrast` relation holds between two discourse segments when their arguments A and C contrast with one another in one or more respects. The difference to `violated expectation` relations is that there is no straightforward causal relationship between A and a (negated) C.

Finally, consider Example (2-c). Our hypothesis is that in this case, both interpretations are possible, i.e., the relation is ambiguous between a `violated expectation` and a `contrast` relation. Our experiment hence compares interpretations of sentences like (2-b) with a preferred `contrast` interpretation to ambiguous sentences like in (2-c).

Corpus study: PDTB

The Penn Discourse Treebank contains annotation of discourse relations in newspaper text for about one hundred connective types, including *but* and *although*. `Violated expectation` relations of the kind shown in (2-a) are annotated as COMPARISON.Concession.expectation relations in PDTB, see examples (3) and (4) below.

`Contrast` relations like the one in (2-b) are annotated as COMPARISON.Contrast relations in PDTB. They differ from Concession relations in that no (denied) direct causality is present between the arguments, see examples (5) and (6).

(3) You might find something, but the chances are low. (WSJ 21 54)

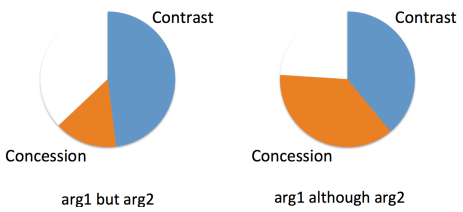


Figure 1: Distribution of discourse relations senses for *but* and *although* in PDTB.

(4) Oil prices haven't declined although supply has been increasing. (WSJ 02 31)

(5) He was on the board of an insurance company with financial problems, but he insists he made no secret of it. (WSJ 00 41)

(6) She didn't elaborate, although earlier US trade reports have complained of videocassette piracy in Malaysia [....] (WSJ 00 20)

We find that *but* and *although* frequently occur as markers in both COMPARISON.Contrast and COMPARISON.Concession relations. However, their distribution is different: *but* most frequently marks COMPARISON.Contrast relations while *although* most frequently marks COMPARISON.Concession relations. A closer look at the sentence initial vs. medial usage of *although* reveals that the distribution of discourse relations for sentence-initial vs. sentence-medial *although* also differs. In particular, the sentence medial use of *although* is divided half-half between Contrast and Concession (see Figure 1)[2]. Based on the corpus study, we hence predict that comprehenders will predominantly make a contrastive inference in the presence of *but*, but that both a contrastive and a violated expectation relation may be inferred when *although* in sentence-medial position is used.

Experiment 1

Our first experiment aims to investigate the difference between *but* and *although* in terms of their biases towards `contrast` and `violated expectation` inferences in an offline text comprehension setup.

Design and stimuli

We design short narrative texts like (7) embedding a discourse relation marked by *but* and *although*. Introduction and continuation are kept identical across conditions. Context is changed by alternating *cake/pizza* and *savory/sweet*. The final sentence is designed to disambiguate between the two alternative discourse relations: *eating pizza* will be consistent with a `contrast` interpretation in a setting where Mary took some cake from the fridge (8a) but then realized she wants something savory. In a `violated expectation` interpretation, *eating pizza* is consistent with taking pizza from the fridge (8b), despite having originally wanted something savory.

(7) **Introduction:** Mary was feeling tired and hungry when she came home yesterday evening.

a. She took some *cake* from the fridge, **but/although** she desired to have something *savory* with her drink.

b. She took some *pizza* from the fridge, **but/although** she desired to have something *sweet* with her drink.

Disambiguating sentence: She had a piece of *pizza* and went to bed earlier than usual.

In this experiment, people are asked to judge the coherence of the entire story after reading it carefully (no time pressure). We expect *but* to cause a strong bias for a `contrast` interpretation. Therefore, condition (a) with *but* should prepare the reader for accepting the continuation *eating pizza*, whereas condition (b) with this connective should result in a contradiction at the end of the story, thus incoherence. On the other hand, *although* should be more ambiguous, i.e., trigger either a `contrast` or a `violated expectation` inference. We expect the average coherence rating of the two *although* conditions to be more similar if the interpretation depends on learned prior usage in text.

As a pretest of the stimuli, we included four additional conditions that exclude the final disambiguating sentence, to make sure that the coherence of the texts up to the disambiguating sentence is matched. Thus, a total of 8 conditions: 2 (connective) * 2 (context) * 2 (with/without final sentence) are constructed for 24 items.

Procedure

We recruited 48 native speakers of English (25 female and 23 male aged between 22-68) on Amazon Mechanical Turk. Eight different lists of 24 stories in equally distributed conditions mixed with filler items were published as HITs. The worker were only allowed to do a single HIT (i.e., one list) of the experiment, so that nobody would see two conditions of the same item. Subjects scored the coherence of each story on a Likert scale from 1 (incoherent) to 7 (perfectly coherent). A compensation of $2.5 was paid per HIT.

[2]COMPARISON.Concession has another subtype in PDTB, called contra-expectation, which differs from the expectation subtype in terms of which of the arguments is the one that creates the expectation, and which one denies it. We found many similar instances of contra-expectation relations to Contrast, therefore, looked at the proportion of expectation vs. contra-expectations in the corpus (for *but* vs. *although*) as well and found very similar proportions to those illustrated in Figure 1.

Data treatment

In total 2400 samples (48 participant * 50 items including 26 fillers) were collected. Only 7 samples were left unanswered in total. Coherence judgment scores for the obviously incoherent and coherent filler items were checked to make sure that the participants understood the task and provided a sensible rating.

Results

Pretest (short versions): Coherence scores assigned to the stories excluding the final sentence were examined first as a pretest of local coherence. According to Anova and mixed-effect regression (with participant and item as random factors, and connective and context as fixed effects), no significant difference is observed between the local coherence of the stories containing either of the two connectives (with an mean rating of 5.14 for *although* and 5.34 for *but*). This shows that there was no difference in coherence between the usage of the connectives in our experimental items. Any difference in coherence judgments for the full texts can therefore be attributed to the final sentence which disambiguates the coherence relation to a `violated expectation` or `contrast` relation.

Main results (full stories): Table 1 presents the average ratings and standard deviations obtained for the complete versions of stories from different conditions.

Table 1: Coherence scores by context & connective

Condition	Mean score	SD
Contrast:but	**5.38**	1.71
Contrast:although	4.85	1.86
ViolExp:but	**3.31**	1.80
ViolExp:although	4.52	1.91

We fitted a linear mixed effects model with random intercepts and slopes for participant and item, using forward selection on fixed effects and backward selection on random slopes (in case the full model didn't converge). The final model (including random slopes for item and participant under connective and relation, as well as the interaction of connective and relation) shows a significant negative main effect of the connective *but* (reflecting the very low coherence judgments for the connective *but* in a `violated expectation` relation), as well as a significant interaction between connective and discourse relation. The presence of the interaction confirms that the two connectives affect interpretation of the discourse differently. Among the *but* conditions, stories with a last sentence consistent with a `violated expectation` inference were rated to be significantly less coherent than stories with a continuation consistent with a `contrast` inference. For texts including the connective *although*, completions consistent with `violated expectation` inferences and `contrast` inferences were both judged to be equally coherent. This finding is in line with our hypothesis, as it directly reflects the distribution of discourse relations that were observed in the corpus.

Experiment 2

The above results show that interpretation of readers wrt. connected sentences in short stories are affected by the fine-grained inferences triggered by the specific discourse connective. The results of the coherence judgments however do not necessarily imply that these inferences are drawn also during natural reading that doesn't include a task focussing on coherence judgments. In this section, we investigate whether readers pick up on the difference between *but* and *although* during online reading, and show effects of coherence on reading times of the final sentence.

Design and stimuli

Stimuli in the eye-tracking experiment are similar to (7), except we test each item also with an alternative final sentence to achieve a fully counterbalanced design. We hence have 4 conditions in this experiment: 2 (connective) * 2 (context). Disambiguating sentence B in (8) has the same function for conditions (8b) as disambiguating sentence A has for conditions (8a).

(8) **Disambig. sent. A:** She had a piece of **pizza** and went to bed earlier than usual.
Disambig. sent. B: She had a piece of **cake** and went to bed earlier than usual.

Items are mixed with filler stories, as well as items of two other experiments with similar length and narrative content. Every participant read 84 stories (including 12 *but* items and 12 *although* items) and answered a YES/NO comprehension question about the introduction part of the stories (not the part depending on the variable interpretation), e.g., "Was Mary at home the entire day yesterday?". The purpose of the questions was to make sure subjects read for comprehension, and to analyze the correlation between coherence of the story and the response time and correctness of the answers. Fillers had questions from all different parts of the text. For our analysis, we collected the total reading time of a story, response time to the question, and the reading time of a critical region in the disambiguating final sentence. The critical region (the word **cake/pizza**) is highlighted in example (8).

Procedure

The eye-tracking experiment was implemented within the Experiment Builder software for an EyeLink 2000 tracker, tracking at 500Hz on both eyes. All text material on the screen were shown in `Lucida Console` font (with

Table 2: Answers to the comprehension questions

Condition	Mean %correct	SD
Contrast:but	**0.85**	0.36
Contrast:although	0.83	0.38
ViolExp:but	**0.79**	0.41
ViolExp:although	0.81	0.40

Table 3: Critical region total RT

Condition	Mean RT	SD
Contrast:but	**296.28**	235.46
Contrast:although	331.90	346.72
ViolExp:but	**332.92**	266.05
ViolExp:although	328.48	263.06

same length characters), size 20 and triple line spacing. Subjects were asked to press the space key after reading a story to navigate to the question screen, and press J and F keys for YES and NO answers, respectively.

A total of 39 native English speakers were recruited for the experiment at the University of Edinburgh and received a compensation of 12 pounds for a two hour session. The eye-tracking experiment was followed by a standard memory test, to measure subjects' memory spans. People's memory span size can play a confound role in the correctness of their answers to the comprehension questions and also might affect reading patterns, e.g., when a sentence contradicts with a non-immediate but related sentence in the preceding context.

Data treatment

We had to discard data from 7 subjects because of frequent head movement, blinks or longer track losses during the experiment. The below results are thus based on a final set of 32 subjects. Standard outlier removal process was performed on the fixations before the RT analysis.

Results

Question answering correctness: Participants' question answering accuracies varied between 50% to 100% correct answers with mean and median of 82%. We observed that subjects with larger memory span size had better question answering performance ($p < 0.05$). Table 2 shows the proportion of correct answers to the comprehension questions across coherence conditions. A trend compatible with the results of the offline study can be observed. However, fitting a mixed-effect regression with all factors (connective, context and participant's memory span) as fixed effects plus participant and item as random effects only revealed a significant main effect of the memory span size. The best fit obtained through a forward model selection procedure showed only a marginal effect of the coherence condition for the subset of data including *but* ($p < 0.1$). This suggests that the coherence of the story as a whole (in terms of the congruence of the final sentence with the interpretation of the middle part) only slightly affected people's recall of the story. Total reading times of the stories and the questions did not correlate with correctness of the answers.

Interest area RT: Table 3 compares the total reading time of the critical area in the final sentence of the stories for every coherence condition. The critical area was chosen to be the only word at which the conditions differ, and which resolves the interpretation of the discourse relation expressed by *but / although*. That is, the word *pizza* vs. *cake* in (9) disambiguates whether the text segment *she desired to have something sweet with her drink* relates to the first segment by a `violated expectation` or `contrast` discourse relation.

Reading times on the critical region are consistent with corpus statistics and the offline study: *but* has a bias towards a `contrast` inference, hence the continuation consistent with this interpretation is processed faster than one confirming a `violated expectation` interpretation. For *although*, the critical word is read at a similar speed independent of whether it disambiguates in favour of a `contrast` or `violated expectation` relation.

We fitted a linear mixed effects regression model for total reading times on the critical region. While a complete model with random slopes for item and participant did not converge, the simpler model (excluding random slopes but including random intercepts) reveals an marginally significant interaction between discourse connective and discourse relation in total reading time of the critical area ($p < 0.1$). There is also a significant effect for this interaction in regressions out of this area to the preceding context ($p < 0.05$). This effect is consistent with the experimental result from the coherence judgment study. The two *but* conditions also differ in terms of the total reading time ($p < 0.1$) and regression-in ($p < 0.05$) at the area preceding the critical phrase in Arg2. Increased regression-in is also observed at the onset of Arg1 for the less coherent *but* condition ($p < 0.1$). Regressions from the final sentence to the previous context indicate that the subjects re-read the areas triggering the relational inference, after they encountered the unexpected continuation. The equivalent complete models to fit first-pass and second-pass duration measures did not converge.

Conclusion

We examined the very fine-grained inferences triggered by two discourse connectives *but* and *although*. Our findings show that these connectives trigger linguistic inferences (despite not being truth-conditional). Each connective in our experiment had a different bias for a specific semantic interpretation tested by a disambiguat-

ing statement in the final sentence, which would sound more or less coherent depending on that interpretation. Our results put into question the traditional perspective that generalizes an effect to a category of connectives, e.g., causal/additive/adversative in previous experimental studies. The probability distribution of a connective in the natural occurrences of various discourse relations (as we took from PDTB) seems to be a more accurate meaning representation for approximation of its effect on offline and online comprehension. This finding motivates future research on comprehension and production of discourse connectives as probability sensitive processes and the interaction between the two. Relevant open questions are why and how different distributional profiles emerge in a given language for closely related discourse connectives, and how such diversities can be explained from a cognitive perspective across languages.

References

Asr, F. T., & Demberg, V. (2012). Measuring the strength of linguistic cues for discourse relations. In *Proceedings of the Workshop on the Advances in Discourse Analysis and its Computational Aspects.*

Asr, F. T., & Demberg, V. (2013). On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 84–93).

Bach, K. (2006). The top 10 misconceptions about implicature. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, 21–30.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge University Press.

Blühdorn, H. (2008). Subordination and coordination in syntax, semantics and discourse. *Subordination versus Coordination in Sentence and Text. A Cross-Linguistic Perspective, Amsterdam*, 59–85.

Drenhaus, H., Demberg, V., Köhne, J., & Delogu, F. (2014). Incremental and predictive discourse processing based on causal and concessive discourse markers: ERP studies on German and English. In *Proceedings of the annual conference of the cognitive science society.*

Fraser, B. (1998). Contrastive discourse markers in english. *Pragmatics and Beyound New Series*, 301–326.

Grice, H. P. (1975). Logic and conversation. *Reprinted in Studies in the Way of Words 1985*, 2240.

Hall, A. (2004). The meaning of but: A procedural reanalysis. *UCL Working Papers in Linguistics*, *16*, 199–236.

Iten, C. (2000). Although revisited. *Working Papers in Linguistics*, *12*.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Kintsch, W., & Van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363.

Köhne, J., & Demberg, V. (2013). The time-course of processing discourse connectives. In *Proceedings of the 35th annual meeting of the cognitive science society.*

König, E. (1991). Concessive relations as the dual of causal relations. *Semantic Universals and Universal Semantics. Groningen-Amsterdam Studies in Semantics*, *12*, 190–209.

Millis, K., Golding, J., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes*, *20*(1), 29–49.

Millis, K., & Just, M. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language*.

Murray, J. (1995). Logical connectives and local coherence. *Sources of Coherence in Reading*, 107–125.

Murray, J. (1997). Connectives and narrative text: The role of continuity. *Memory and Cognition*, *25*(2), 227–236.

Noordman, L. G., & Vonk, W. (1992). Readers' knowledge and the control of inferences in reading. *Language and Cognitive Processes*, *7*(3-4), 373–391.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th international conference on language resources and evaluation* (pp. 2961–2968).

Rohde, H., & Horton, W. S. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, *133*(3), 667–691.

Sanders, T. J., & Noordman, L. G. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse processes*, *29*(1), 37–60.

Traxler, M., Sanford, A., Aked, J., & Moxey, L. (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 88.

Xiang, M., & Kuperberg, G. (2014). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, 1–25.

Tangible models and haptic representations aid learning of molecular biology concepts

Kristen Johannes (kjohann@wested.org)
WestEd, 300 Lakeside Dr., Oakland, CA, 94612

Jacklyn Powers (jpowers@wested.org)
WestEd, 300 Lakeside Dr., Oakland, CA, 94612

Lisa Couper (lcouper@wested.org)
WestEd, 300 Lakeside Dr., Oakland, CA, 94612

Matt Silberglitt (msilber@wested.org)
WestEd, 300 Lakeside Dr., Oakland, CA, 94612

Jodi Davenport (jdavenp@wested.org)
WestEd, 300 Lakeside Dr., Oakland, CA, 94612

Abstract

Can novel 3D models help students develop a deeper understanding of core concepts in molecular biology? We adapted 3D molecular models, developed by scientists, for use in high school science classrooms. The models accurately represent the structural and functional properties of complex DNA and Virus molecules, and provide visual and haptic feedback about biomolecular properties that are often implicit in traditional models. We investigated: 1) Can we measure conceptual growth on core concepts? 2) Do lessons with 3D models improve student outcomes on these measures?, and 3) What factors mediate learning? Model use yielded measurable gains in conceptual knowledge and the greatest gains were related to how actively models were used during a lesson and the facilitative role adopted by the teachers.

Keywords: Scientific models; science education; molecular structure; visual representations; haptic representations

Background

Molecular biology requires reasoning about molecules and processes too small to directly observe. As such, scientists use models to generate and test hypotheses about the structures of complex molecules such as proteins, viruses, and DNA (Berry & Baker, 2010). Many existing models primarily represent rigid structures of molecules. Flexible models that embed magnets into 3D-printed structures demonstrate how molecular structure relates to biomolecular functions such as DNA replication, viral assembly, protein folding, and enzyme catalysis. However learning from complex molecular models requires students to map between the model features and background knowledge about the domain. In the current study, we asked whether experience with flexible, tangible models helps students develop key conceptual knowledge in novice high school biology students and whether conceptual growth can be measured in real time. For the study, we iteratively developed and revised a series of 3D models and model-based activities. To assess conceptual knowledge, students completed identical pre/post measures organized around three targets areas: molecular structure, the relationships between structural properties and function, and the relationship between model representations and molecules.

The structurally-accurate, 3D-printed models are the product of technological development at the Scripps Research Institute (Höst, Larsson, Olson, & Tibell, 2013). The tangible models use embedded magnets to represent chemical attraction and repulsion, which allow students to "feel" bonds form in processes such as protein folding, nucleic acid synthesis, and viral self- assembly. As the structure and magnets of the 3D-printed models allow students to create “correct” models much more easily than incorrect models, the models provide formative feedback to the students, prompting them to address misconceptions, build on incomplete ideas, and confirm correct ideas.

In the field of biology, molecular models have played a critical role in transforming biochemistry from a descriptive science into a constructive one (de Chadarevian & Hopwood, 2004). We explore whether 3D physical models could be effective for promoting classroom learning by making explicit links between abstract and physical (visuospatial and haptic) representations (Bivall, Ainsworth, & Tibell, 2011). Prompting students to explore the affordances and limitations of models may scaffold students' development of metacognitive understanding of complex fields like molecular biology (Coll, France, & Taylor, 2012).

The Current Study

To examine whether interactions with flexible, 3D printed models embedded with magnets improve students conceptual understanding of molecular biology, we introduced two sets of interactive models and activities into high school biology classrooms. One set focused on DNA structure and replication, the other on the life cycle of a

virus and viral self-assembly. We were interested in three related questions:
1. Do our (pre-posttest) performance measures capture key concepts in molecular biology?
2. Do tangible models facilitate student learning of these concepts?
3. How does variation in classroom environment and model use relate to variation in learning gains?

We hypothesized that interactive models would lead to increased student learning, particularly improving students' understanding of concepts that are used regularly by working scientists but are left implicit in traditional biology curricula. Table 1 summarizes the targeted concepts. We discuss the concepts in the context of tangible models for our two featured molecular biology case studies: DNA replication and viral self-assembly. We further explored how naturally-occurring differences in classroom contexts influenced learning gains. As our models targeted concepts that are known to be challenging and typically absent from high school biology curricula, there is no clearly suitable business-as-usual condition for comparison. Thus, current aim was to establish whether our measures were reliable indicators of the concepts of interest (Table 1), and whether these newly developed models and activities improved student learning over time.

Table 1. Examples of target concepts in molecular biology

Concept	DNA replication examples	Viral assembly examples
Molecular structures	The size and structure of purines and pyrimidines forces complimentary base pairing with different numbers of H-bonds; The anti-parallel leading and lagging strands of DNA give rise to the "twist" of the double helix	The polio virus is assembled from identical pentameter subparts; Magnetic polarity (attraction bonds) allow subparts to assemble into full virus
Structure-function relationships	H-bonds between nucleotide bases are the weakest and first to break in the semiconservative process of replication	The viral replication cycle and self assembly processes are possible through repeating units and limited RNA
Model use and representations	Model component sizes and magnetic strengths are proportional to the relative sizes and bond strengths within the DNA polymer	Rate of shaking (virus assembly in a container) is analogous to the temperature of the system

Methods

Design. The study was designed as two model-based activities, described below, administered by teachers during their normal biology classes. Activities were preceded and followed by identical pre- and post-tests. Classroom observers for each activity measured students' interaction and use of models, the amount of time devoted to different parts of the activities, and teachers' roles (e.g., facilitating group discussion, answering questions, lecturing) across time points. Teacher interviews and questionnaires solicited feedback on teachers' use and perceptions of the activities and models.

Participants. Nine high-school biology teachers from three different California schools used the models and activities in their classrooms. Teachers received 4 hours of training on background content, features of the models and activities, and strategies for integrating the models into existing biology curricula and for giving student feedback. Class sizes ranged from 22 to 31, yielding final sample sizes of 850 for DNA models and activities and 675 for virus models and activities. The student sample was 59% female and included 33% 9th, 49% 10th, 8% 11th, and 10% 12th grade students.

Models and Activities

Model-based activities typically spanned two class periods. During this time, students interacted with the models with the goal of learning about specific phenomena. For each activity, teachers introduced the models and activities with a series of discussion questions designed to engage students with the relevant phenomena. Students were then given the opportunity to explore the models in small groups and to work through activity worksheets. After working in small groups, students shared their observations, evaluated their responses to the worksheet questions, and revisited the introductory questions in more detail. Finally, teachers elaborated on the properties and use of the models themselves, and helped students make connections between model representations and the structures, functions, and processes underlying the phenomena.

DNA model. A significant challenge for physical models of DNA is the representation of flexibility (that allows sections of DNA to be available for replication) and bonding that allows of both pairing and mis-pairing (that results in for sequences of DNA to code for different proteins). The multi-component DNA model (Figure 1) consists of multiple nucleotides, each with a base, sugar and phosphate. Hydrogen bonds on the bases set in plastic "plungers" that allow for a certain degree of rotation and accurately represent the Watson-Crick hydrogen bonding capacity. Individual bases have attachment sites for the sugars and the phosphates that make up of the backbone of DNA. Finally, the covalent bonds that hold the nucleotides together are represented by a flexible strand of plastic string. The model features a snap bead design so students can create strands of DNA with different lengths. In the current version of the model, one strand is fully connected, representing a template strand of DNA, whereas the other strand could be constructed nucleotide-by-nucleotide so students could engage in the process of replication. The 3D tangible DNA model allows students to investigate several challenging

concepts including the molecular structure of DNA and the semiconservative nature of replication.

DNA Labeling and Replication activities. After group discussion of DNA, students identified and labeled different parts of the DNA models on a worksheet and provided brief explanations of the evidence that they used to identify each part. Students discussed the component parts that make up DNA's structure; the ways in which different types of bonds and spatial configurations enable important functions of DNA; the advantages and limitations of models (including comparing tangible DNA models to a folded paper model that was distributed); and the ways in which the structure of DNA allow cells to replicate genetic information.

On the second day, students completed a DNA replication activity, building complementary strands for the DNA double-helix from the 3D model components. Students added components (e.g., nucleotides) to the model one at a time and, after each component was added, responded to a question on a worksheet. Students discussed: What processes occur during replication? How is DNA structure related to function in the process of replication? How does the process of replication ensure exact copies are created? And, how do the models help explain the relationship between structure and function?



Figure 1. Assembled interactive DNA model

Poliovirus model. The 3D model of the poliovirus (Figure 2), developed at Scripps Research Institute, highlights the dynamic process of self-assembly and the role of temperature. In self-assembly, subunits come together to form larger structures as a result of random motion and local attractions. Most biomolecular processes rely on self-assembly for the formation of complex molecules. However, self-assembly is a challenging concept because there are few analogs in everyday life. In the model, molecular interactions are represented by magnets on the edges of each viral subunit. When pieces are put into random motion, by shaking, the pieces collide and the magnetic attraction causes them to assemble into various intermediate structures and finally into a complete model of the viral capsid. The rate of shaking represents the temperature of the system and students can use the model to investigate the role of temperature on the formation of particles. The ability of the model to self-assemble emerges because the local structures and attractive forces of the poliovirus sub-units were used to design the model; this process of self-assembly was not specifically designed.

Viral life cycle and self-assembly activities. After a group discussion of viral replication, students demonstrated their understanding of the replication process by arranging steps starting with the virus binding to a cell and ending with many newly-formed viruses bursting from the cell. Students next discussed virus structure and replication (with a focus on correct vocabulary - e.g., *capsid, RNA*); the viral cycle compared to other cycles in biology; and the steps in the viral cycle represented by the cards and 3D model.

Students used the 3D poliovirus model to explore self-assembly: students were instructed to place the virus pieces into a closed container and continuously shake the container. As they observed the effect of the rate and force of shaking and the success of the viral self-assembly process, they answered questions about the model structure, stability, and the effect of increasing the energy to the system. Students also observed the interactions between the viral subunits and answered worksheet questions that prompted them to make inferences about the relationship between viral structure (multiple identical pentameter pieces) and function (replication and self-assembly).



Figure 2. Assembly process for Poliovirus model.

Measures

Pre/post tests. Students' understanding of content was measured before and after the activities with identical pre- and posttests designed separately for DNA and virus activities. Test items were designed to tap into core concepts, reasoning skills and spatial understandings that are central to molecular biology. Items were either researcher-developed, designed to directly address problem solving and model use, or sourced from AAAS (MLSCI:Wright & Hamilton, 2008) and other educational research studies related to molecular biology (e.g., Stieff, 2007). Tests aligned with the key concepts in Table 1. Sample items from the DNA and virus test are presented below in (1) and (2), respectively:

(1) Two major functions of DNA are replication and transcription. Name three ways these processes are the same and three ways they are different.

Explain how the following structures of DNA enable it to replicate: a. Hydrogen bonds between pairs; b. Covalent bonds between sugars and phosphates.

(2) Describe one way that self-assembly is needed for gene expression.

Describe how the following affect self-assembly:
a. The shapes of the parts that self-assemble; b. The energy available for self-assembly; c. The positive and negative charges on proteins

Classroom observations. Researchers observed each activity and recorded the amount of time spent using the models, the depth of model use, ranging from highly

interactive use to passive observation; and the teacher's role, (e.g., addressing the entire class, individually interacting with students, or doing unrelated activities.)

Teacher interviews. Teachers completed a brief questionnaire after each activity and participated in a one-on-one interview with researchers at the end of the study. These measures addressed teachers' implementation and perceptions of the activities and use of the models.

Results

1. Do our (pre-posttest) performance measures capture key concepts in molecular biology?

We identified three critical cross-cutting concepts as learning goals: Molecular Substructures, Structure-Function Relationships, and Model Understanding, (see Table 1). These concepts were used to design both sets of activities and the pre/post measures. To determine whether the items we developed tapped into the hypothesized targets, we computed Confirmatory Factor Analysis (CFA) models to test whether these three critical concepts were latent variables around which students' posttest performance on the pre/post measures could be organized (loading strength for the DNA and Virus concept models are shown in Figures 3 and 4, respectively). Specifically, we tested a model with factor loadings for our three concept-based factors against other two three-factor models, which were based on item type (e.g., multiple choice, diagram, open response) and on science curriculum content and objectives (e.g., providing information and constructing explanations), as well as an exploratory three-factor model in which the factors were not specified *a priori*. Models were computed using the `TAM` package in R (Kiefer, Robitzsch & Wu, 2015) and we compared BIC values as an index of model fit, with lower values indicating better fit. For both DNA and virus activities, the concept-based CFA model yielded the lowest BIC values of the four. Model fit BIC differences ranged from 65 to 120 in favor of the concept-model, strongly indicating a better fit to the posttest data (Rafferty, 1995). These models support the hypothesis that our models and activities target complex underlying concepts in molecular biology and not simply, for example, improving students' abilities to provide information or select the correct response on a multiple choice question.

2. Do tangible models facilitate student learning of these concepts?

After establishing organizing conceptual factors for the pre and post measures, we examined student responses to determine whether their proficiency had changed from pre- to posttest. For both the DNA and virus activities, we fit separate Rasch models to students' pre and posttest responses to compare gains in estimated student proficiency from pre- to post. We first used posttest scores to generate item difficulty estimates, the odds of students incorrectly answering an item vs. correctly answering the item, and Expected A Posteriori (EAP) estimates of student proficiency, the odds of answering an item correctly vs. incorrectly. We then used the posttest item difficulty estimates as fixed item parameters to estimate student proficiency at pretest, while equating item difficulty between pre- and posttest. Histograms of students' proficiency estimates are displayed in Figure 5 for DNA and Figure 6 for virus activities.



Figure 3. Loading strengths for the concept-based CFA model of DNA posttest items (on y-axis).

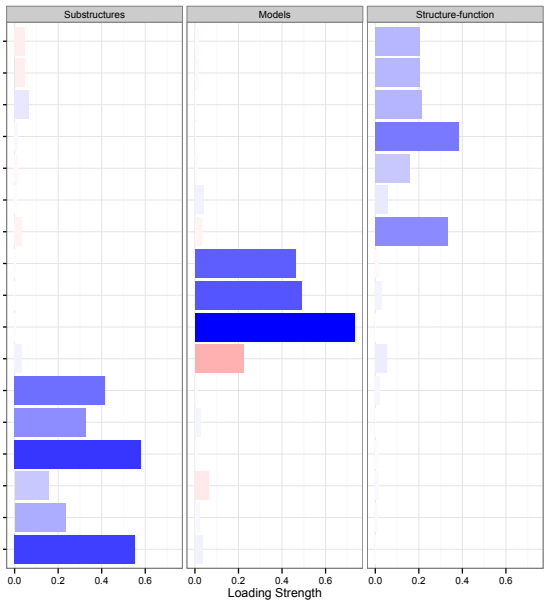


Figure 4. Loading strengths for the concept-based CFA model of Virus posttest items (on y-axis).

For both activities, student proficiency increased from pre- to posttest, following the model activities. The increase was deemed statistically reliable through Chi-square goodness of fit tests (DNA: $\chi 2(1, N=850) = 26.2$, $p<.01$; Virus: $\chi 2(1,$

N=675) = 40.1, p<.01). We take these findings as preliminary evidence that students using the tangible models make measurable and reliable gains in their understanding of key concepts in molecular biology.

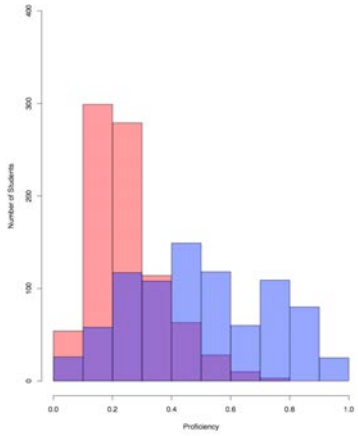


Figure 5. DNA activity student proficiency gains, evident as a distributional shift from pretest (pink) to posttest (blue).

.

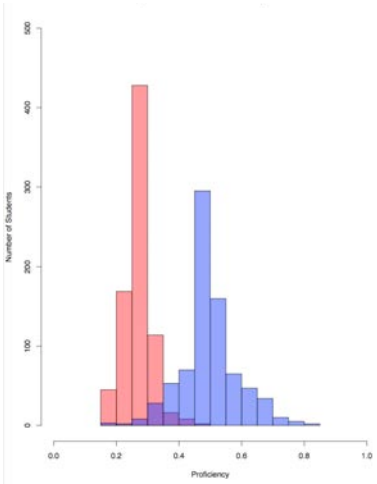


Figure 6. Virus activity student proficiency gains, evident as a distributional shift from pretest (pink) to posttest (blue).

3. What kind of instructional context and model usage leads to the greatest gains in learning?

Preliminary analyses (not reported here) revealed that, for both DNA and virus activities, item difficulty and teacher, together, predicted gains from pre- to posttest: test gains vary considerably by item and by teacher. We explored this variation further by examining classroom observations collected in each classroom for each activity. Specifically, we used mixed effects logistic regression models to test whether differences in gains from pre- to posttest were related to differences in aspects of classroom model use and the facilitating role adopted by teachers in the classroom

We fit separate models for DNA activities and virus activities. Each model included item and classroom random effects on the intercept as well as several fixed effects, explained in turn. The best fitting models for both DNA and virus test gains included item difficulty, the amount of time students spent engaging with the models in different ways, and the amount of time teachers spent actively facilitating students' activity use.

Students in the DNA activities made greater gains for more difficult items (β =0.2, p<.05) and several observed aspects of the classroom environment were also related to test gains, as shown in Figure 7. The amount of time spent using the models was not predictive of gains (β= 0.03, *ns*), however students who spent long periods of time passively using models showed smaller test gains (β= -0.31, p<.01). Additionally, the amount of time that teachers spent systematically monitoring and assisting students during the activities predicted greater test gains (β=0.22, p<.01).

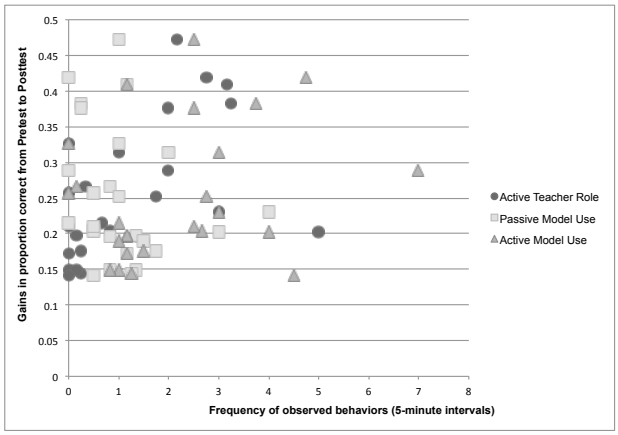


Figure 7. Relationships between DNA test gains (y-axis) and frequency of observations of active and passive model use and active teacher role predictors (x-axis). Passive model use and active teacher roles were reliable predictors.

In contrast to the DNA activities, students in the virus activities made greater gains for *less* difficult items (β =-0.28, p<.05). This reversal may be due to the fact that the virus test items were more difficult, overall, compared to the DNA items, leaving more room for improvement on easier items. Similar to the results from the DNA activities, the quantity of overall model use did not predict test gains. However, the amount of time students spent actively and passively using the model had opposite effects on test gains, as shown in Figure 8. Longer periods of active model use predicted greater test gains (β =-0.61, p<.01), while long periods of passive model use predicted smaller test gains (β =-0.12, p<.01). Finally, teachers' facilitative role during the virus activities was not related to students' test gains: the amount of time that teachers spent systematically monitoring and assisting students during the activities was not a reliable predictor of test gains (β=0.08, *ns*).

We also examined teachers' open-ended interview responses as qualitative evidence for the potentially facilitating effects of active model use and teacher role in student performance. Responses from teachers whose students showed the highest gains revealed common themes in the ways they structured the activities and used the models with their students. These teachers all reported starting their biology unit with the activities and structuring their class time to maximize model use.

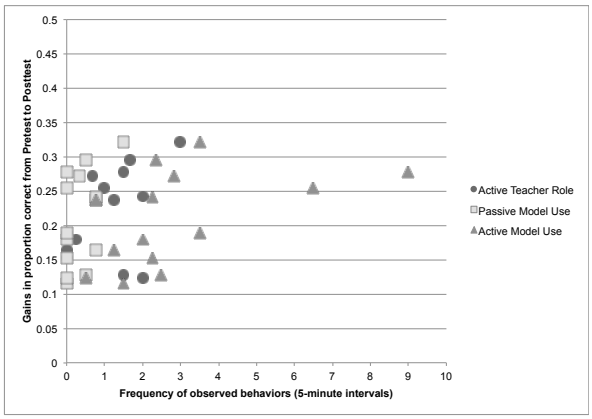


Figure 8. Relationships between virus test gains (y-axis) and frequency of observations of active and passive model use and active teacher role predictors (x-axis). Active and passive model use were reliable predictors.

Teachers used the models and activities to reinforce vocabulary, synthesize concepts learned across biology units and, especially, to highlight properties of molecules that are otherwise difficult to observe, as a quote from one teacher's interview reveals:

> *"[I think] having the models there allowed students to connect a physical shape and structure to something incredibly small that they don't really understand."*

Across high school classrooms, the conceptual gains were sensitive to students' opportunity to actively engage with the models and how teachers linked the interactive models to complex concepts and vocabulary. The 3D DNA and virus models offer spatial and haptic affordances that highlight important structural and functional features. Our data suggest that using these novel models leads to learning gains over the course of a few lessons.

Conclusions

In this study, we asked whether valid measures of key science concepts could be created and whether high school students that use flexible 3D models better understand key concepts in molecular biology. The tangible molecular models accurately represent the structural and functional properties of complex DNA and virus molecules. Model use helped high school students understand critical biology concepts that are often implicit in current DNA and viral assembly instruction. Active model use and teacher scaffolding were related to increased pre- to posttest gains. Our findings demonstrate that non-expert high school students benefit from model use in reasoning about molecular structures and processes: concepts that motivate experts' use of models in the lab.

Our results also suggest a role for modality-specific, grounded representations in conceptual learning (Barsalou et al., 2003; Mahon & Caramazza, 2008). We suggest that visuo-spatial and haptic representations of the models aid students in specifically encoding and understanding complex 3D molecular structure, and haptic representations of the forces and bonds between components enable students to further reason about how these structures influence biomolecular processes (Morris et al, 2007; White, 2012). This work highlights the role of learning tools that allow students to see 3D molecular structure and to use haptic feedback to "feel" molecular processes through models that accurately represent both 3D structure as well as chemical attraction and stability.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120047 to WestEd and the National Science Foundation under Grant (DRL-1108896, PI JD). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not represent the views of the Institute, the U.S. Department of Education or the National Science Foundation.

References

Bivall, P., Ainsworth, S., & Tibell, L.A.E. (2011). Do haptic representations help complex molecular learning? *Science Education*, *95(4),* 700-719.

Berry, C., & Baker, M.D. (2010). Multimedia in Biochemistry and Molecular Biology education: Inside protein structures. *Biochemistry and Molecular Biology Education, 38(8)* 425-429.

Coll, R.K., France, B., & Taylor, I. (2012). The role of models and analogies in science education: Implications from research. *International Journal of Science Education, 27(2),* 183-198.

de Chadarevian, S., & Hopwood, N. (Eds.). (2004). *Models the Third Dimension of Science,* Redwood City, *CA:* Stanford University Press

Höst, G.E., Larsson, C., Olson, A., & Tibell, L.E. (2013). Student learning about biomolecular self-assembly using two different external representations. *CBE- Life Sciences Education, 12,* 471-482.

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test Analysis Modules. R package version 1.13-0. http://CRAN.R-project.org/package=TAM

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. Sociological Methodology, 25, 111-163.

Stieff., M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction, 17(2),* 219-234.

Wright, T., & Hamilton, S. (2008). Assessing student understanding in the molecular life sciences using a concept inventory. In: Duff A, Green M, Quinn D, editors. *Proceedings of the ATN Assessment Conference: Engaging Students in Assessment.* Adelaide, Australia: ATN: American Society for Cell Biology. pp. 216–224.

Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors

Till Poppels (tpoppels@ucsd.edu)[1]
[1]Department of Linguistics
University of California, San Diego
San Diego, CA 92093 USA

Roger P. Levy (rplevy@mit.edu)[1,2]
[2]Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Previous research has found that comprehenders are willing to adopt non-literal interpretations of sentences whose literal reading is unlikely. Several studies found evidence that comprehenders decide whether or not a given utterance should be taken at face value in accordance with principles of Bayesian rationality, by weighing the prior probability of potential interpretations against the degree to which they are (in)consistent with the literal form of the utterance. While all of these results are consistent with string-edit noise models, many error processes are known to be sensitive to the underlying linguistic structure of the intended utterance. Here, we explore the case of exchange errors and provide experimental evidence that comprehenders' noise model is structure-sensitive. Our results add further support to the noisy-channel theory of language comprehension, extend the set of known noise operations to include positional exchanges, and show that comprehenders' noise models are well-adapted to structure-sensitive sources of signal corruption during communication.

Keywords: rational analysis; noisy-channel comprehension; non-literal interpretation;

Introduction

Evidence that comprehenders adopt non-literal interpretations when the literal meaning of a sentence is implausible or otherwise unlikely has been around for at least 15 years. For example, Christianson, Hollingworth, Halliwell, and Ferreira (2001) found that garden-pathed readers partially retain the thematic role assignment associated with their initial misinterpretation even when it is incompatible with the literal sentence once the garden path has been resolved. Ferreira (2003) subsequently showed that such tendencies arise not only from garden-path constructions, but also in ordinary sentences with implausible literal interpretations, particularly when implausible events are expressed using non-canonical linguistic forms (e.g., in passive voice). These observations raise the questions *when* and *how* comprehenders decide whether or not a given utterance should be taken at face value and receive a literal interpretation, and if not, what alternative interpretation should be adopted instead.

According to the noisy-channel (Shannon, 1949) theory of language comprehension, interpretations arise rationally and gradiently through probabilistic inference (Levy, 2008). As per usual in the tradition of Rational Analysis (Anderson, 1990; Chater & Oaksford, 1999), the noisy-channel theory takes as a starting point the hypothesis that comprehension is the statistically optimal solution to the problem of communicating under noise. The structure of this problem is schematically represented in Fig. 1: The speaker intends to convey meaning M by encoding it linguistically in a structured representation S with surface form w. For example, if S is the intended syntactic tree, w may represent the word string that corresponds to the intended yield of the tree. More generally, we take w to be an ordered, but otherwise unstructured, string of atomic elements (e.g., words)[1], and S to contain any additional information, such as the lexical category of words, their functional position in a phrase structure, etc. /Thus, S and w characterize the underlying structure and surface form of the speaker's *intended* utterance, and jointly determine the *actual* utterance u. Since this is where the speaker's intention takes physical shape, u can now be processed by the comprehender to produce I, the input to the interpretation process, which finally results in the comprehender's inferred meaning M'.

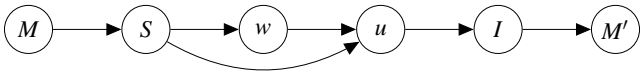


Figure 1: Schematic representation of the noisy channel.

Communication is successful when M' recovers M, which is threatened by the possibility of signal corruption at various points along this noisy channel. For example, phonetic reduction may cause the deletion of words between w and u, and interference during lexical retrieval may result in the accidental substitution of intended words with other words. These are two examples of (production-based) noise processes that may cause u to differ from w, but, and this is crucial for the present paper, these processes may be modulated by the structural information in S. For example, phonetic reduction is more common in function words than in content words (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), and accidental substitutions typically involve words of the same grammatical category (Harley & MacAndrew, 2001). Likewise, phonological exchange errors (e.g., "**c**oat-**thr**utting") frequently involve phonemes that occupy the same syllable position or share certain phonetic features (Ellis, 1980). More generally, then, the effect of S on u, captured by the curved edge in Fig. 1, is manifest in what we call "structure-sensitive" errors: it modulates the operation of noise processes between w and u, based on structural information about the atomic elements in w. The structure-sensitivity of $w \rightarrow u$ noise processes makes them qualitatively different from the noise processes that transform the actual utterance u into the input I to the comprehension system. The latter are physical and neural processes including environmental noise, precision limits of perceptual systems,

[1]In principle, these atomic units could also represent phonemes or orthographic characters, but the present paper focuses on word-based operations.

and imperfection in memory traces of previous input representations. These operate on the actual utterance u and thus must be conditionally independent of the upstream variables w and S, which have representational status only internal to production.

Given this representation of the communicative process, the comprehender's decision whether or not a given utterance should be interpreted literally, comes down to the conditional distribution $P(M'|I)$. We hypothesize that comprehenders infer this distribution by rationally integrating their prior expectations about meanings with the likelihood of the observed input under their model of the noisy channel:

$$P(M|I) \propto P(I|M)P(M) \tag{i}$$

This general characterization of comprehension as Bayesian inference underscores the importance of understanding the noise model, $P(I|M)$, that comprehenders use to reverse-engineer the process that may have generated the input they observed. In keeping with the tradition of Rational Analysis, we take as a starting point the noise model of an optimal comprehender (Fig. 2), which mirrors the production process and thus explicitly represents the possibility of various types of noise corruption, as discussed above.

The noisy-channel proposal has found general support in online reading-time studies (Levy, Bicknell, Slattery, & Rayner, 2009; Levy, 2011; Bergen, Levy, & Gibson, 2012), as well as in offline measures of comprehension accuracy (Gibson, Bergen, & Piantadosi, 2013). However, there are two important limitations in these previous studies, which are addressed in the present paper: First, the noise operations that comprehenders have been shown to detect and repair during interpretation so far only contain deletions, insertions, and substitutions, which represent only a small subset of the noise processes known to affect production. One prominent source of noise that is known to operate at various levels of linguistic representation involves the accidental exchange of elements (e.g., phonemes, affixes, or words), yet no previous study has found evidence that comprehenders expect and repair such errors (in fact, as discussed below, Gibson et al. (2013) report evidence that appears to be inconsistent with this possibility). A second, related limitation of previous research is that all of the extant evidence is compatible with structure-*in*sensitive noise models, since all of the noise processes that comprehenders have been shown to repair in these studies are string-based operations that do not require reference to, or knowledge of, the underlying structure of the input.

The central difference between such structure-*in*sensitive noise models and the structure-sensitive model we are testing in the present paper is captured by the dashed edge in Fig. 2: in a structure-*in*sensitive noise model, this connection does not exist and w' d-separates u' and S'. Consequently, the underlying structure of the intended word string has no bearing on the actual utterance except through the word string itself. If, however, as we hypothesize, comprehenders do consider the possibility of structure-*sensitive* noise operations,

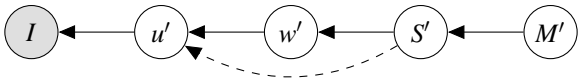


Figure 2: The comprehender's noise model ideally mirrors the production process. u', w', and S' are the comprehenders model of u, w, and S (cf. Fig. 1); the dashed edge corresponds to the structure sensitivity hypothesis pursued in the present paper. The shading emphasizes that only I is known to the comprehender and forms the basis for inferring M' (cf. Eq i).

their noise model completely mirrors the production process as represented in Fig. 1, and inferring the intended meaning of an utterance requires marginalizing over w', S', and u':

$$P(M'|I) \propto P(M') \int_{w',S',u'} P(I|u')P(u'|w',S')P(w'|S')P(S'|M')\,\mathrm{d}w'\,\mathrm{d}S'\,\mathrm{d}u' \tag{ii}$$

To test the structure sensitivity hypothesis and address limitations of previous research, we test whether comprehenders consider the positional exchange of words a potential source of signal corruption. Initial support for this hypothesis comes from the abundance of exchange errors in the phonological domain (MacKay, 1987), which comprehenders appear to repair routinely during comprehension. For example, upon encountering the phrase "coat-thrutting", the non-literal alternative "throat-cutting" appears to surface naturally (MacKay, 1987). It is important to note at this point that detecting exchange errors does not require structure sensitivity *per se*, since the exchange of two randomly sampled elements in a string is possible without reference to (or knowledge of) the underlying structure. However, if comprehenders expect the exchange of some elements but not others, that would support the structure sensitivity hypothesis because that distinction does require reference to the underlying structure of the surface string.

Gibson et al. (2013) found evidence that suggests that comprehenders do not consider the positional exchange of at least some words. They presented participants with sentences that varied in terms of semantic plausbility and syntactic structure and probed their interpretation through comprehension questions. Overall, comprehenders readily adopted non-literal interpretations when (a) the literal utterance was semantically implausible and (b) a more plausible alternative interpretation could be obtained by postulating the deletion or insertion of individual words. However, implausible transitive sentences, such as (1-a), were almost always interpreted literally, although a more plausible reading was available on the assumption that the two nouns had been exchanged by mistake (1-b). The finding that comprehenders retained the faithful but implausible reading in such cases suggests that the word exchange that would recover the more plausible reading is not a likely operation under the their noise model.

(1) a. The ball kicked the girl.
b. The girl kicked the ball.

However, the observation that comprehenders do not contemplate exchanges in such cases does not imply that their noise model precludes any possibility of exchange errors. For example, the noise model may place higher probability on the exchange of function words than content words or favor exchanges out of syntactic adjuncts compared to complements, which would make the noun-noun exchanges that are required for repairing (1-a) a low-probability noise operation.

The present paper provides evidence from applying the experimental paradigm used in the Gibson et al. to further explore the status of exchange errors under comprehenders' noise model. We presented participants with sentences whose thematic role assignment was either plausible or implausible on a literal interpretation and, crucially, could be reversed through the positional exchange of prepositions. If comprehenders are tempted by these non-literal interpretations, that would suggest not only that exchange errors *are* among the noise operations they consider, but also that their noise model is structure-sensitive in that it assigns greater probability to the exchange of some elements (e.g., prepositions) than others (e.g., nouns).

Methods

Participants were presented with sentences that were plausible or implausible on a literal interpretation, paired with comprehension questions that were designed to distinguish between literal and non-literal interpretations in a 2-alternative forced-choice task. The materials were constructed to afford alternative interpretations through word exchanges. To foreshadow the results, participants' response patterns strongly suggest that these exchanges were indeed among the noise operations they considered during comprehension.

Participants. 60 self-reported native speakers of English were recruited via Amazon.com's Mechanical Turk. 1 participant with less than 75% accuracy on filler items was excluded from the analysis.

Materials. Following a 2x2 within-subject design, the semantic plausibility of the literally described event was crossed with the order of prepositional phrase (PP) adjuncts, as illustrated in Table 1. More precise estimates of the semantic plausibility of the utterances were obtained in a separate norming experiment and the canonicality of PP orders was estimated in a corpus analysis (both described below).

Plausibility	PP order	The package fell...
plausible	canonical	...from the table to the floor.
plausible	non-canonical	...to the floor from the table.
implausible	canonical	...from the floor to the table.
implausible	non-canonical	...to the table from the floor.

Table 1: Example item in 2x2 within-subject design.

The literally encoded event is either plausible (package falling down) or implausible (falling up), and in each case the other reading is available by exchanging the prepositions while leaving everything else in place. Both plausibility and canonicality affect the prior probability of the literal interpretation: since the implausible event is less likely to occure, it is less likely to reflect the speaker's communicative event, but irrespective of her intended meaning, she is more likely to express it using the canonical PP order "from...to..." than the non-canonical "to...from...".[2]

Procedure. Participants were presented with 20 experimental items in a random order, interspersed with 48 filler items. In a full latin square, one of the four versions of each item was displayed together with a yes/no comprehension question that was designed to distinguish between plausible and implausible thematic-role assignments. For example, the sentences in Table 1 were followed by the question *Did something fall to the floor?*, to which completely literal comprehenders would respond "Yes" following the plausible sentences and "No" following the implausible ones. Optimal noisy-channel comprehenders, on the other hand, would weigh the possibility of an exchange error against the prior probability of the candidate interpretations, and thus may exhibit the reverse response pattern when the literal parse is semantically implausible and/or syntactically non-canonical. Across items, the plausibility of the comprehension question was counterbalanced to ensure that the literally correct answer was equally often "Yes" and "No".

Canonicality norming. To estimate the relative canonicality of PP orders, we performed a frequency analysis of the Brown corpus (Kucera & Francis, 1967) and the Wall Street Journal (Paul & Baker, 1992) sections of the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1994), using Levy and Andrew's (2006) `Tregex` software package for tree searches.[3] The relative frequency with which each preposition pair occurred in the more frequent order is summarized in Fig. 3. For the purpose of the 2x2 design, each utterance was categorized as canonical or non-canonical (based on the threshold indicated by the dashed line), but the analyses reported below made use of the continuous estimates.

Plausibility norming. To quantify differences in plausibility between sentences both within and across items, we conducted a separate norming experiment, in which participants ($n = 12$) rated the plausibility of the events described by the sentences in question (e.g., *The package fell from the table to*

[2]The corpus analysis described below found that in over 90% of sentences with PP adjuncts headed by "from" and "to" the former preceded the latter (cf. Fig. 3).

[3]The queries for this analysis were, for example, `__< (@PP <<# from $++ (@PP <<# to))` and `__< (@PP <<# to $++ (@PP <<# from))`, for preposition pair `from-to`.

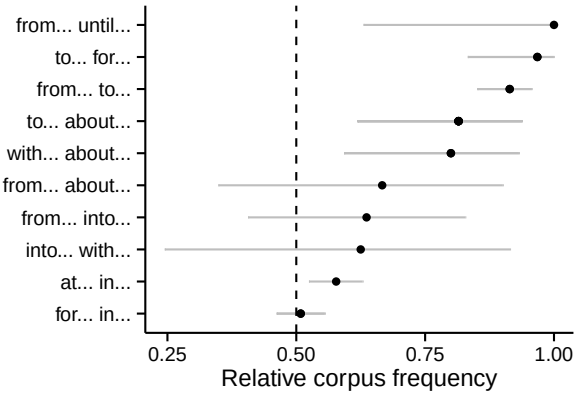


Figure 3: Relative frequency of canonical order of each PP pair. The dashed line indicates the threshold for categorizing orders as canonical vs. non-canonical. Errorbars represent Clopper-Pearson confidence intervals for proportions, calculated with the `PropCIs` R package (Scherer, 2014).

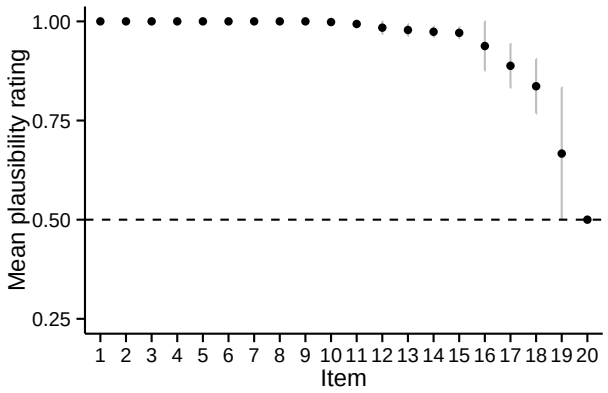


Figure 4: Mean normalized plausibility ratings of canonical PP exchange items. As in Fig. 3, the dashed line indicates the categorization threshold for establishing the 2x2 design. Errorbars represent SEs of the mean.

the floor vs. *The package fell from the floor to the table*), using two independent sliders, each ranging from "completely plausible" to "completely implausible". Only canonical item versions were used, so that the sentence pairs that participants rated differed only in their thematic role assignment, and not in the order of the prepositions. The two measures obtained for each sentence pair were normalized and scaled to range over the interval [0,1] by dividing each by their sum.[4]

Replication of Gibson et al. experiments. To situate our results within the context of the Gibson et al. (2013) findings, we replicated three of the experiments that were reported in that paper, each testing a different syntactic alternation. We chose one that according to Gibson et al. had received mostly non-literal interpretations (DO/PO benefactives), one that produced a balanced number of literal and non-literal interpretations (transitive/intransitive), and one that triggered almost no non-literal interpretations (active/passive). Example sentences from each experiment are shown below:

(2) Active/passive:
  a. The girl kicked the ball. [plausible]
  b. The girl was kicked by the ball. [implausible]

(3) Transitive/intransitive:
  a. The chemotherapy shrank the tumor. [plausible]
  b. The chemotherapy shrank from the tumor.[impl.]

(4) Direct-object/prepositional-object benefactives:
  a. The father bought his son a bicycle. [plausible]
  b. The father bought his son for a bicycle. [impl.]

[4]For example, if the plausible sentence was rated as "completely plausible" (corresponding to a score of 100), and the implausible alternative received the score 20, the normalized plausibility scores were $\frac{100}{100+20} \approx 0.83$ and $\frac{20}{100+20} \approx 0.17$, respectively.

The items used in these three experiments were included in the norming experiment described above, which presented all items in a within-subject design to ensure that plausibility norms were comparable across experiments (cf. Fig. 5).

Predictions. If accidental exchanges of prepositions are possible under comprehenders' noise model, we expect them to adopt non-literal interpretations at least on some trials. More specifically, the noisy-channel model predicts that the extent to which comprehenders perform such noise inferences should be inversely proportional to the prior probability of the literal utterance, and should therefore be driven by the semantic plausibility as well as syntactic canonicality of the displayed sentence.

If comprehenders' noise model is structure-sensitive, they may assign different probabilities to exchanges of some elements compared to others. Thus, we predict that comprehenders adopt non-literal interpretations in the case of prepositional exchanges more readily compared to the noun-noun exchanges that would permit the repair of implausible active/passive sentences. In other words, the rate of noise inference should be higher for sentences like *The package fell from the floor to the table* compared to sentences like *The ball kicked the girl*.

Results

Plausibility and canonicality norming. The mean normalized plausibility ratings of the PP exchange items are shown in Fig. 4, summarized alongsides the replication materials in Fig. 5 (since ratings for plausible sentences and their implausible counterparts summed to 1, only plausible versions are shown). Overall, sentences were rated as either highly plausible or highly implausible (all normalized mean ratings for plausible items were above 0.9), although the PP exchange materials were slightly less polarizing than the materials used

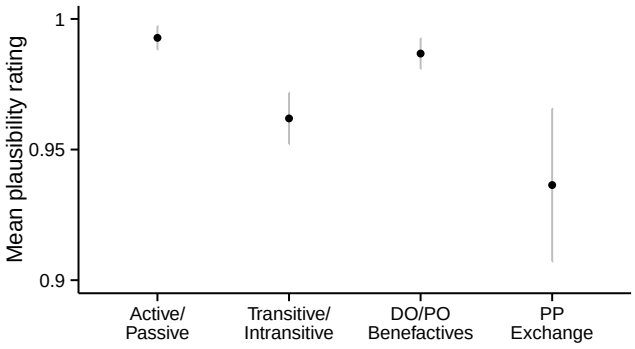


Figure 5: Mean normalized plausibility ratings for plausible items from all 4 experiments. Errorbars show SEs of by-item means.

in the Gibson et al. (2013) study, especially the active/passive sentences.

Noise inference. Fig. 6 shows the results from all four experiments, broken up by construction and binned plausibility. Consider first the replications of the Gibson et al. experiments in panels 1, 3, and 4 (from the left). As expected, accuracy on plausible items (light grey) was largely at ceiling for all of the experiments, and comparable to the mean accuracy on filler items (white), which is consistent with Gibson et al.'s results. Implausible items (dark grey) received non-literal interpretations to different extents across experiments. Most notably, implausible passives triggered noise inferences significantly more often than plausible passives ($p = 0.02$) and marginally more often than implausible actives ($p = 0.1$). This result is consistent with Ferreira's (2003) findings and slightly different from what Gibson et al. (2013) reported, who found passives to be interpreted literally on more than 95% of trials. In all other respects our results closely replicate those from the Gibson et al. study.

The central prediction of the present study was that implausible (and non-canonical) sentences would provoke non-literal interpretations that can be reached through the positional exchange of elements within the sentence. We replicated Gibson et al.'s finding that such exchanges appear to be impossible—or at least highly unlikely—in the case of active/passive constructions, which were overwhelmingly interpreted literally, even when they described highly implausible events (cf. panel 1 in Fig. 6). However, as illustrated in panel 2, implausible sentences with exchangeable PP adjuncts did receive non-literal interpretations on 31-37% of trials, while their plausible counterparts were interpreted literally more than 90% of the time, which is comparable to the mean comprehension accuracy on filler items. Consistent with this, a binomial mixed-effects regression analysis[5] revealed a main effect of plausibility ($\beta = 1.303$, $p < 0.0001$) as well as a small, but significant, main effect of canonicality ($\beta = 0.276$, $p = 0.038$). Both main effects were in the predicted direction: non-literal interpretations were more likely for utterances with implausible semantics and non-canonical form than for those that described plausible events and made use of the more canonical PP order.

Finally, the proportion of literal interpretations of implausible PP exchange items was significantly lower compared to implausible active/passive sentences ($p < 0.001$).

[5]The analysis was carried out using the `lme4` R package. The full formula was `response ~ plausibility * canonicality (1 + plausibility * canonicality || item) + (1 + plausibility * canonicality || subject)`.

Discussion

The results presented here demonstrate that word exchanges are among the noise operations comprehenders consider when interpreting implausible or otherwise unlikely utterances. Moreover, they suggest that comprehenders' noise model is structure-sensitive, since the possibility of exchanging two words for gains in plausibility was exploited in the case of PP adjuncts, but not active/passive constructions.

It is not clear from our results, however, exactly what elements comprehenders expect to be subject to exchanges. Notice that the plausible interpretation of

(5) [$_{PP}$ to [$_{NP}$ the table]] [$_{PP}$ from [$_{NP}$ the floor]]

can be achieved by exchanging either the prepositions or the nouns (or noun phrases). The significant effect of canonicality does suggest that prepositions were considered the object of the exchanges at least some of the time because exchanging nouns does not achieve a change in the order of prepositions and can therefore not explain the effect of canonicality. However, this does not imply that noun-noun exchanges are impossible. Although we have exemplified possible structure-sensitive constraints on exchanges in terms of lexical categories or content/function word status throughout the paper, it is also possible that the crucial difference between PP exchanges and noun-noun exchanges in actives/passives is that the former involve adjuncts whereas the latter affect elements in complement position. It is possible that either or both of these constraints exist, but since our data do not allow us to test either of these admittedly post-hoc explanations directly, they remain in the realm of speculation.

Conclusion

We have reported experimental evidence that comprehenders expect the communicative signal to be corrupted by accidental word exchanges, and that this expectation is structure-sensitive. These results build on previous findings suggesting that the language comprehension machinery chooses rationally from a range of possible interpretations. Going beyond previous formulations of comprehenders' noise model, we showed that it more completely mirrors the production process than previously assumed. Thus, our results add to a growing pile of evidence that language comprehension is well-adapted to the problem of communication across a noisy channel, and makes use of all available information in the

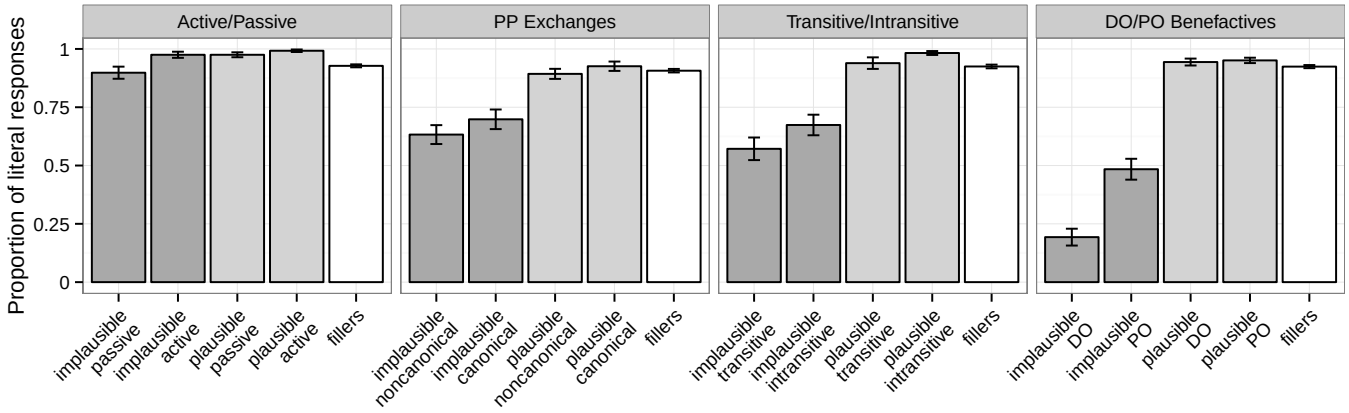


Figure 6: Proportion of literal responses across experiments. Errorbars represent SEs of by-subject means.

search for the interpretation that is most likely to reflect the speaker's communicative intent.

Acknowledgments

This work has benefitted from discussion with audiences at the Linguistics BrownBag at UCSD, AMLaP 2015, and the 2016 Annual Meeting of the LSA. We gratefully acknowledge support from NSF grant IIS-0953870, NIH grant HD065829, and an Alfred P. Sloan Research Fellowship to RPL.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the 34th annual meeting of the Cognitive Science Society* (pp. 1320–1325).

Chater, N., & Oaksford, M. (1999). Ten Years of the Rational Analysis of Cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407.

Ellis, A. W. (1980). Errors in speech and short-term memory: The effects of phonemic similarity and syllable position. *Journal of Verbal Learning and Verbal Behavior*, *19*(5), 624–634.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*(2), 164–203.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Harley, T. A., & MacAndrew, S. B. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research*, *30*(4), 395–418.

Kucera, H., & Francis, W. (1967). The Brown University Standard Corpus of Present-Day American English (Brown Corpus). *Providence: Brown University*.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the conference on empirical methods in natural language processing*, 234–243. doi: 10.3115/1613715.1613749

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In (pp. 1055–1065). Association for Computational Linguistics.

Levy, R., & Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *In 5th international conference on language resources and evaluation.*

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

MacKay, D. G. (1987). Constraints on theories of sequencing and timing in language perception and production. *Language perception and production: Relationships between listening, speaking, reading and writing*, *3*, 407.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. Retrieved from `citeseer.nj.nec.com/marcus93building.html`

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on speech and natural language* (pp. 357–362).

Scherer, R. (2014). *PropCIs: Various confidence interval methods for proportions.* R package version 0.2-5.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21.

No stereotype threat effect in international chess

Tom Stafford (t.stafford@sheffield.ac.uk)
Department of Psychology, University of Sheffield
Sheffield, S102TP, United Kingdom

Abstract

We examine data from over 6.6 million games of tournament chess between players rated by the international chess authority, FIDE. Previous research has focussed on the low representation of women in chess. We replicate and extend previous analysis (Chabris and Glickman, 2006) on an international level. We find no support for differential variability, differential drop-out between male and female players, or social context (in the form of proportion of female players at a national level) as drivers of drivers of male-female differences. Further, we examine games between mixed and same gender pairs for evidence of a 'stereotype threat' effect. Contrary to previous reports, we find no evidence of stereotype threat. Though this analysis contradicts one specific mechanism whereby gender stereotype may influence players, the persistent differences between male and female players suggests that systematic factors do exist and remain to be uncovered.

Keywords: learning; chess; skill acquisition; expertise;

Introduction

Chess has an illustrious history within cognitive science (Newell et al., 1958; Chase & Simon, 1973; Charness, 1992), providing a paradigmatic example of cognitive skill, and a testbed for theories of skill acquisition and performance. Aside from its worldwide popularity, and historical and cultural interest, chess has the advantage of being a skill with minimal perceptual or motor requirements, the upper-bound on an individual's performance being their cognitive capacity in planning, and reasoning through the complex space of possible moves. Chess also has the advantage that players are rated using the Elo system (Elo, 1978), which updates according to a player's success or failure in games against other rated players. This provides an objective measure of skill, which is not directly contaminated by the subjective perception of observers.

The chess playing community is predominantly male. Previous research has explored a number of possible competing explanations for the under-representation of women in chess (Chabris & Glickman, 2006; Bilalić et al., 2009). In their study, Chabris and Glickman (2006) looked at 250,000 US tournament players and found that men's ratings were not more variable than women's and that younger players did not learn at different rates according to gender. They found that men's ratings were higher than women's in areas where there were fewer women players (and not so in areas, defined by zip code, where at least 50% of younger players are female) — a result which is suggestive of an effect of social context on players' performance or skill acquisition which interacts with gender.

One notable psychological phenomenon which can influence performance is that of 'stereotype threat', whereby an individuals' awareness of a negative stereotype influences their performance. This was originally proposed for intelligence test performance and African Americans (Steele & Aronson, 1995), and has since been extended to other domains, most pertinently for our purposes to women and performance in non-stereotypically feminine domains of achievement, such as mathematics (Spencer et al., 1999).

In chess, both observational (Rothgerber & Wolsiefer, 2014) and experimental studies (Maass et al., 2008) appear to confirm the existence of stereotype threat. Rothgerber and Wolsiefer (2014) looked at 219 female chess players and found evidence for stereotype threat in the field — these players under-performed relative to their rating when they played male players. They report (p.79) that "Stereotype threat susceptibility was most pronounced in contexts that could be considered challenging: when playing a strong or moderate opponent". It should be noted that their sample was very young (5 – 15 years). Maass and colleagues (2008) ran an study using internet chess, where the perceived gender of opponents was experimentally manipulated with 84 participants (half male, half female, mean age 33.5). When believing they were playing an opponent of the opposite gender female players were less likely to win. If these findings generalise widely to chess performance they have the potential to systematically undermine the performance of female players.

A recent analysis suggests that publication bias has exaggerated the reality of the stereotype threat phenomenon (Flore & Wicherts, 2015) (see also (Stricker, 2008; Ganley et al., 2013)). In a similar vein, there is a possibility that a methodological confound is responsible for some positive replications of the effect, at least in the domain of mathematics and a gender stereotype threat effect (Stoet & Geary, 2012).

So although an obvious disparity exists in participation rates between men and women, there is uncertainty over the mechanisms by which this is perpetuated. In particular, the phenomenon of stereotype threat offers a specific psychological mechanism whereby cultural stereotypes and the existing relative paucity of female role models can interact with gender to hamper women's achieve-

ments in chess, but has not been convincingly established for a wide age range playing at the higher levels of the game. This is what this study set out to do.

Data and method

Our data comprise 6,641,158 games played by 461,637 FIDE rated players, of which 56,474 (12.2%) are women. Of these players, 176,583 were active during the 92 month period for which we have data, and for these players we have statistics on each game they played (and subsequent rating changes). The average birth year for these players was 1983, with a standard deviation of 19.78.

For each player the data consists of a unique player ID, date of birth, gender, nationality and a details of the games they played (including the piece colour they played as - White or Black - who they played against, the tournament this was part of, and the outcome). The data also contains all players' official FIDE rating calculated according to the Elo system. This system updates players' ratings according to game outcomes and acts both as a prediction system for the outcome of a match of any two rated players and as a way of ranking any player against the historical community of all players contained within the system. Because Elo ratings are updated after each game, it is possible to compare players who have never played, and may not even be contemporaneous.

Our analytic strategy is to first confirm for our international sample the differences found in the US sample studied by Chabris and Glickman (2006), and then to explore in greater detail the possible influence of social context — and particularly the availability of female role models - has on learning and performance. Specifically, this means to confirm a difference in ratings between male and female players, which cannot be attributed to differential variability in ratings or drop-out. Next, to see if drop-out and rating differences between young male and female players can be attributed to difference in proportion of female players at a country level. Finally, we investigate in detail the possibility of stereotype threat, a candidate phenomenon for reduced female performance in stereotypically male domains. The advantage of chess is that we are able to precisely gauge the challenge presented by individual games to each player, via comparison of player ratings.

Analysis scripts are available at `https://osf.io/aeksv`. For commercial reasons the full dataset is not available at the point of writing.

Results

Differences in ratings and drop-out

The average FIDE rating of men is 1880 (standard deviation 295, and for women 1700 (standard deviation 318). This difference is statistically significant, $t(176096) = 74.31, p < 0.0001$). For reference, a rating above 2500 is associated with Chess Grandmaster level.

The ratio of the standard deviation ratings for women to men was 1.08. Like Chabris and Glickman (2006), this shows higher variability in women's ratings, offering no support to the hypothesis that males are more variable in their ability.

One possible explanation for the male advantage in ratings is that women drop out of chess at higher rates than men. Following Chabris and Glickman (2006), we looked at young players (aged 5 to 25, birth years 1987–2007) who were active in the third year covered by our data, tracking them to see if they remained active in the remaining years covered by our data.

The average number of days until one of these players became inactive was 553 (for 28,190 men) and 560 (for 6,510 women, a significant difference ($t(34698) = 2.42, p = 0.016$). Clearly higher drop-out rate among potentially strong female players is not a plausible explanation for any male advantage in ratings.

Differences by country

One proposed mechanism by which a gender difference may be created, perpetuated or exacerbated is the minority status of women chess players. Considerable variation exists in the proportion of chess players who are female by nation (minimum: 2.0%, Denmark; maximum: 33.9%, Mozambique). The proportion of female chess players at competition level could conceivably influence younger female players, by providing role-models, mentors, or merely by lessening their perceived minority status. To investigate this issue, we compared the difference between male and female ratings, and between male and female time-until-drop-out, against the proportion of female chess players in each country.

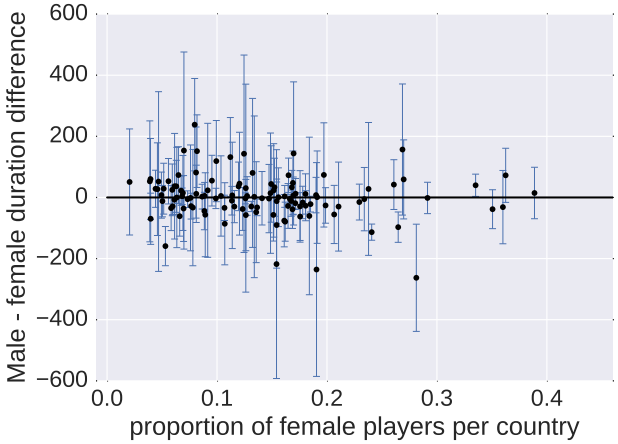


Figure 1: Difference in duration-until-inactive between young male and female players. 95% confidence intervals shown.

As shown in Figure 1, there is no discernible relation between proportion of female players in country and the difference in drop out rates for young male and female players (Pearson's $r = -0.15$, $p = 0.10$).

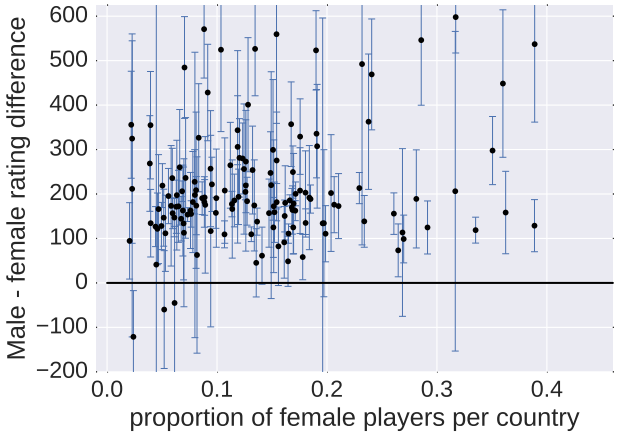


Figure 2: Difference in rating between young male and female players. 95% confidence intervals shown.

As shown in Figure 2, there is not a strong relation between proportion of female players in country and the difference in rating for young male and female players. If anything, countries with a higher proportion of female players see a larger disparity between male and female ratings (Pearson's $r = 0.21$, $p < 0.012$, compare to Chabris and Glickman, 2006, figure 4).

Differences in learning rate

So far our analysis has been restricted to using a player's best rating, but inspecting the change in players' ratings over time also allows the investigation of certain questions. Especially for younger players, we would expect performance to improve with practice, and so be reflected in an increase in rating over time. Previous research has asked if performance at chess is affected by gender, it is also possible to ask if learning is affected by gender. Anything that affects learning will have an influence on performance, by necessity.

Due to uncertainty over the precise model which best fits the learning curve (Gaschler et al., 2014; Gallistel et al., 2004) we fitted change in performance with a simple linear regression for each player, using their Elo rating at each time point. We restrict our analysis here again to young players (born 1987–2007) and to those who played at least 10 games during the period covered by the data.

The average slope was higher for males (0.101/day, n=21263) than for females (0.064/day, $n = 5265$); a highly significant difference $t(26526) = 15.95, p < 0.00001$. For reference, this represents a modest average yearly increment of 36.9 versus 23.4 Elo points.

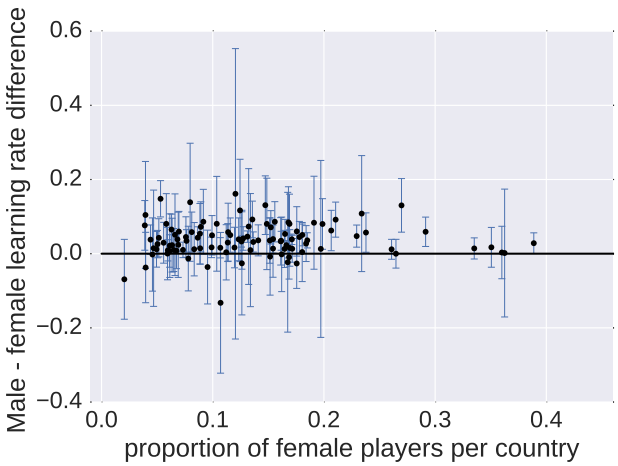


Figure 3: Difference in average slope of linear fit to ratings over time between male and female players. 95% confidence intervals shown but not visible at this resolution.

Figure 3 shows how the differing proportion of female players relates to the average difference in slope between male and female players. As with drop-out, there is no clear relation between the proportion of female players and how young female players learn with respect to young male players (Pearson's $r = 0.07$, $p = 0.48$), although we note that as with the Chabris and Glickman (2006) analysis of US zip codes, those countries with the highest proportion of female players ($> 30\%$) are conspicuous in showing no strong evidence of a difference between learning rates for young male and female players.

Differences in by-game performance

Our data also allows us to look at how individual game performance is affected by player characteristics. Figure 4 shows observed relationship between rating difference (rating of player playing White − rating of player playing Black) against outcome (coding a win for player playing White as 1; win for player playing Black as 0; draw as 0.5).

As we expect, there is a clear relationship between the relative player ratings and game outcome. Note that at around 0 difference in player ratings the average outcome is above 0.5 — showing, as is widely known, that the White player has an advantage.

We can ask how this relationship changes for higher and lower rated competitors. To do this we split the games into five groups (quintiles) according to the average rating of the two players, so that the 1st quintile contains games in which the average rating of the two players is in the top 20% of all games, and so on. To highlight how quintile affects outcome, it is helpful to

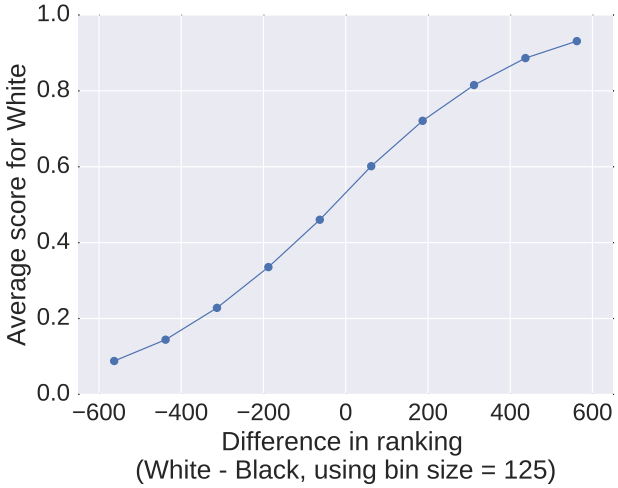


Figure 4: Difference in player rating against average game outcome (6,641,158 games). 95% confidence intervals shown.

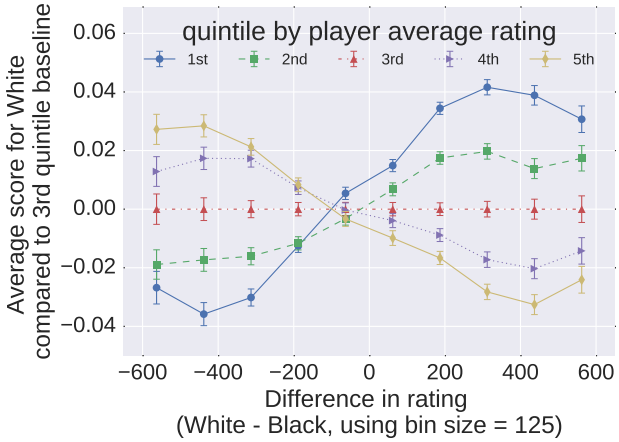


Figure 5: Effect of player average rating on game outcome (6,641,158 games). 95% confidence intervals shown.

'normalise' outcome against the effect of rating difference. To do this we plot the relative change in outcome compared to some baseline. In this case, the baseline we use will be the middle (3rd) quintile. The result is shown in Figure 5.

Figure 5 shows the variation around the basic pattern that is shown in Figure 4 which occurs as you move from the games between the lowest rated players (5th quintile) to the highest rated players (1st quintile). For the two higher quintiles, the curve is above the 3rd quintile when White's rating exceed Black's, and below that when Black's exceeds White's. This shows that even for the same absolute difference in rating, when the player's ratings are, on average, higher, the probability of the higher rated player winning is greater. The opposite is true for lower rated players – the same absolute difference in rating is less predictive of outcome. Alternatively put, the higher rated players can make more advantage of any absolute rating difference (note also that the effect shows at the 0 rating difference point, meaning that the probability of a White win when the players are evenly matched moves further away from 50% as the players' ratings goes up – higher rated players can make more of a small advantage).

Finally, we can look to see if gender affects the results. Previously female players have been shown to be at a relative disadvantage when they believed they were playing men (Maass et al., 2008), a finding which is in line with the literature on 'stereotype threat', whereby highlighting a person's membership of a minority group can affect their performance to be in line with stereotypes associated with that group (in this case, the stereotype that women are not as good at chess). In international chess the gender of your opponent is highly obvious. We code all the games in our data set according to whether they are played between two men ('MM'), two women ('FF') or mixed gender pairings, with a woman playing White ('FM') or Black ('MF'). A previous observational report of stereotype threat in chess (Rothgerber & Wolsiefer, 2014) suggested that this effect would be most likely in "challenging situations" and when playing someone of a higher grade. International chess tournaments are certainly challenging, and our previous analysis is suitable for showing how any effect changes with player rating relative to their opponent.

As with the analysis by player average rating, we normalise the outcome by comparing our results for player rating difference against a baseline. In this case we choose a baseline of the games when two men played ('MM'). A stereotype threat effect should reduce the probability of a woman winning when she plays a man, compared to when a man plays a man or a woman plays a woman. Graphically, this should appear as a lower curve for the 'FM' group (where White is a woman), and a higher curve for the 'MF' group (where White is a man). In particular, following Rothgerber's suggestions, we would expect that this effect would manifest most strongly when a woman plays a superior opponent (so in the negative portion of x-axis for the 'FM' group and the positive portion of x-axis for the 'MF' group). The results are shown in Figure 6

As can be seen in Figure 6, there is no stereotype threat effect observable in international chess. If anything, the opposite pattern is found – both mixed gender pairs show a relatively enhanced probability of White winning when White is the lower rated player, and a relatively diminished probability of White winning when

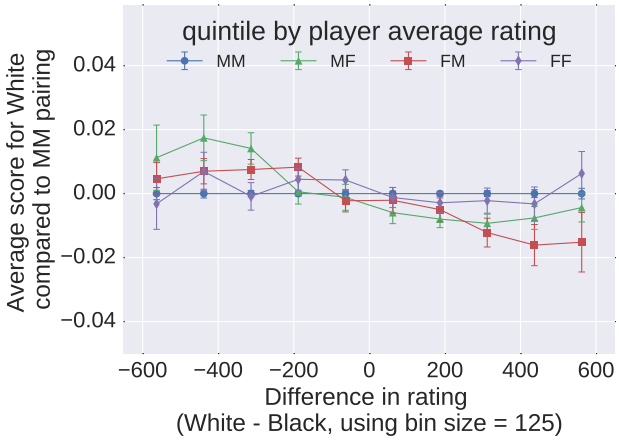


Figure 6: How player gender pairing affects game outcome (6,624,273 games). 95% confidence intervals shown.

White is the higher rated player, regardless of whether White is a man or a women. Recall that a stereotype effect should be the opposite of that seen here for the 'FM' and 'MF' pairings — diminished chance of White winning when White is a woman and the lower rated player (the 'FM' curve should move into the bottom left quadrant, which it doesn't), and enhanced chance of White winning when White is a man and the higher rated player (the 'MF curve should move into the top right quadrant, which it doesn't). By comparing this effect to Figure 4 you can see that it is the same as the effect of lower rated player-pairs overall (e.g. the 4th and 5th quintiles by pair average rating).

Discussion

We replicate and extend previous analyses of differences between male and female chess players. Like Chabris and Glickman we find no support for the idea that differential drop-out might explain sex differences in achievement in the sport. Those authors concluded that social context may be an important factor, based on a by-state analysis of US players. Our analysis used an international sample and examines by-country differences in proportion of female players. We find no evidence that this proportion influences the differences between men and women in ratings, drop-out or learning rate. Our analysis uses the international population of players, so we might expect greater cultural variation than between US zip codes (as in Chabris and Glickman's analysis). Further, our data allows us to explicitly test the idea of stereotype threat, one candidate mechanism by which social context may effect performance. This has not been done before for chess. We find no support for this phenomenon, contrary to previously published reports. We note that this is consistent with other studies of stereotype threat in high-stakes real-world settings (Stricker, 2008).

We use a large population for our analysis, rather than a sample of tens or hundreds (Rothgerber & Wolsiefer, 2014; Maass et al., 2008). It may be that the older age of our sample, the higher playing standard and/or the greater pressure of international competition induces a professionalism among players that protects against stereotype threat. Although this gives our results a strong validity in terms of the population of FIDE rated chess players, it does mean that we must recognise the unusually highly rated nature of our population compared to those used in some other studies of chess players, and particularly of younger chess players. Working with very large datasets introduces some new opportunities for the cognitive scientist (Stafford & Dewar, 2014; Stafford & Haasnoot, under review). Observational studies, however large, necessarily have reduced power of causal inference compared to experimental studies. Counterbalancing this is undeniable relevance of any phenomenon observable real-world data such as that used here.

The question of the under-representation of women in chess remains unsolved, we have merely provided evidence that stereotype threat is an unlikely mechanism for sustaining any difference in male-female ratings once players have achieved a standard that allows them to hold a FIDE rating. Some researchers (Bilalić et al., 2009; Charness & Gerchak, 1996) suggest that the gender difference at the top of the distribution is a natural consequence of different participation rates — in other words, that the low number of women in the highest echelons of chess is the simple result of the much larger number of men in the population of chess players from which the best players are drawn. It is certainly a problem that analysis of rated players limits the conclusions that can be drawn because we are in effect only looking at a subset of all possible players (Vaci et al., 2014). From this perspective the difference in participation between men and women in chess itself may be the primary factor to be explained, rather than any difference in ratings or maximal achievement (which may be explained sufficiently by differential participation).

Recently, chess has been a focus for large scale analytics. (Howard, 2006; Chassy & Gobet, 2015; Leone et al., 2014), and we see this study as part of that trend. Future work with this data has great potential for investigating differences in change in expertise, as well as performance. Highly relevant is the observation that tournament games are actually the most significant events in rating improvement (Howard, 2012, 2013). Future work on chess is sure to focus on within-game dynamics as well as the dynamics of ratings. To the end of promoting integration of existing work and further exploration of the rich data provided by FIDE chess ratings we are

happy to make the analysis scripts for the current analysis available immediately at `https://osf.io/aeksv`), and the full data available in time.

References

Bilalić, M., Smallbone, K., McLeod, P., & Gobet, F. (2009). Why are (the best) women so good at chess? participation rates and gender differences in intellectual domains. *Proceedings of the Royal Society of London B: Biological Sciences*, *276*(1659), 1161–1165.

Chabris, C. F., & Glickman, M. E. (2006). Sex differences in intellectual performance analysis of a large cohort of competitive chess players. *Psychological Science*, *17*(12), 1040–1046.

Charness, N. (1992). The impact of chess research on cognitive science. *Psychological research*, *54*(1), 4–9.

Charness, N., & Gerchak, Y. (1996). Participation rates and maximal performance: A log-linear explanation for group differences, such as russian and male dominance in chess. *Psychological Science*, 46–51.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chassy, P., & Gobet, F. (2015). Risk taking in adversarial situations: Civilization differences in chess experts. *Cognition*, *141*, 36–40.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? a meta-analysis. *Journal of school psychology*, *53*(1), 25–44.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(36), 13124–13131.

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental psychology*, *49*(10), 1886.

Gaschler, R., Progscha, J., Smallbone, K., Ram, N., & Bilalić, M. (2014). Playing off the curve-testing quantitative predictions of skill acquisition theories in development of chess performance. *Frontiers in psychology*, *5*.

Howard, R. W. (2006). A complete database of international chess players and chess performance ratings for varied longitudinal studies. *Behavior research methods*, *38*(4), 698–703.

Howard, R. W. (2012). Longitudinal effects of different types of practice on the development of chess expertise. *Applied Cognitive Psychology*, *26*(3), 359–369.

Howard, R. W. (2013). Practice other than playing games apparently has only a modest role in the development of chess expertise. *British journal of psychology*, *104*(1), 39–56.

Leone, M. J., Slezak, D. F., Cecchi, G. A., & Sigman, M. (2014). The geometry of expertise. *Frontiers in psychology*, *5*.

Maass, A., D'Ettole, C., & Cadinu, M. (2008). Checkmate? the role of gender stereotypes in the ultimate intellectual sport. *European Journal of Social Psychology*, *38*(2), 231–245.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development*, *2*(4), 320–335.

Rothgerber, H., & Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group Processes & Intergroup Relations*, *17*(1), 79–90.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4–28.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological science*, *25*(2), 511–518.

Stafford, T., & Haasnoot, E. (under review). Confirming and quantifying sleep consolidation in skill learning: a field study using an online game.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, *69*(5), 797.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102.

Stricker, L. J. (2008). The challenge of stereotype threat for the testing community. In *Presidential address to the division of evaluation, measurement, and statistics. 2007 american educational research association annual meeting.*

Vaci, N., Gula, B., & Bilalić, M. (2014). Restricting range restricts conclusions. *Frontiers in psychology*, *5*.

Acknowledgements

The "Sonas 92" dataset used in this analysis was prepared by Jeff Sonas of Sonas Consulting (jjeff@sonasconsulting.com). Without his generosity and advice this study would not have been possible. Thanks also to Alberto Ara, Stephen Want and to three anonymous reviewers and the editor for their detailed feedback.

‘Unlikely’ Outcomes Might Never Occur, But What About ‘Unlikely (20% Chance)’ Outcomes?

Sarah Jenkins (s.jenkins.12@ucl.ac.uk) and Adam J. L. Harris (adam.harris@ucl.ac.uk)
Department of Experimental Psychology, 26 Bedford Way
London, University College London, WC1H 0AP, UK

R.M Lark (mlark@bgs.ac.uk)
British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK

Abstract

A commonly suggested solution to reduce misinterpretations of verbal probability expressions in risk communications is to use a verbal-numerical (mixed format) approach, but it is not known whether this increases understanding over and above a purely numerical format. Using the ‘which outcome’ methodology (Teigen & Filkuková, 2013), we examined the effect of using verbal, numerical and mixed communication formats, as well as investigating whether marking outcomes as salient would alter the outcomes people perceived as ‘unlikely’ or having a 20% chance of occurring. We observed no effect of saliency, but replicated previous findings, with general preference for values at the high end of a distribution (including maximum/above maximum values) present in both verbal and mixed communication formats. This demonstrates the relevance of these findings for real-world consequential risk communication. Whilst the estimates differed between the mixed and numerical formats, we found that the mixed format yielded the more accurate estimates.

Keywords: verbal probability expressions; numerical probabilities; risk communication; geological hazards

Introduction

Effectively communicating information about risk and uncertainty remains an ongoing challenge for the scientific community. The process relies on recipients of risk communications both understanding the information, and also placing enough trust in it that it will be used in subsequent decision making. Most people do not have in-depth knowledge about, nor experience of, hazards and new technologies (Siegrist, Gutscher & Earle 2005). Individuals are therefore reliant on mediated information, which tends to be from an expert source (Sjöberg, 2000). Ensuring the audience understands the information as intended is a universal concern for scientific communications. Scientific forecasts are, however, typically probabilistic (at best). It is thus not possible to predict with certainty whether a destructive earthquake will occur in a certain place within the next month for example. A prediction that such an event is ‘unlikely’ does not imply that the event will not occur. Given that an estimate of ‘unlikely’ might be used to describe the likelihood of events with a 20% chance of occurrence (e.g., Theil, 2002), approximately 20% of the time, they will occur. As the prosecution of six experts following the L’Aquila Earthquake in 2009 attests (Cartlidge, 2012), such a lack of certainty is not always well received by the public, resulting in the potential for reduced trust in (and sometimes criminal proceedings against) the scientists who make such predictions.

Budescu and Wallsten (1995) proposed that the choice of risk communication format should be governed by the congruence principle: the precision of the risk communication should match the precision of the event in question, thus reflecting the nature of its uncertainty. Estimating the likelihood of a large earthquake might not be precisely quantifiable. In such instances, a specific numerical expression of the probability of this event might be perceived as overly precise. Using a verbal probability expression (VPE), however, better represents the uncertainty and underlying imprecision associated with the probability estimate. This characteristic is one reason for many organisations’ use of VPEs in risk communication (e.g., Intergovernmental Panel on Climate Change [IPCC], Mastrandrea, 2010).

There is, however, considerable variability in people’s usage and interpretations of VPEs (e.g., Budescu & Wallsten, 1985). In addition to ‘natural’ inter-individual variability, interpretations of VPEs are susceptible to contextual and cultural influences (e.g., Bonnefon & Villejoubert, 2006; Fischer & Jungermann, 1996; Harris & Corner, 2011; Harris, Corner, Xu, & Du, 2013; Juanchich, Sirota & Butler, 2012; Teigen & Brun, 1999, 2003; Weber & Hilton, 1990).

The abovementioned studies have typically used the ‘how likely’ translation approach to investigate interpretations of VPEs, whereby people are asked to translate VPEs to corresponding numerical probabilities. However, more recently, Teigen and colleagues have demonstrated that a ‘which outcome’ approach to understanding people’s interpretations of VPEs paints rather a different picture (Teigen, Juanchich & Riege, 2013; Teigen & Filkuková, 2013, Teigen, Juanchich & Filkuková, 2014). In this approach, participants are shown a distribution of outcomes and asked to complete probability statements (e.g. “It is unlikely that a battery will last… hours”) with a value they consider appropriate (see Figure 1 for example). This approach has highlighted a seemingly large qualitative disparity between the communicator’s intended meaning and the recipient’s understanding of the information. Teigen et al. (2013) found when the term ‘unlikely’ was used to describe outcomes which can be ordered on a unipolar dimension (e.g.,

battery life), participants interpreted the term as referring to outcomes from the higher end of the distribution and most often completed the sentence with a lifetime that *exceeded* the maximum time any sampled battery had lasted (Study 3a) – hereafter the 'which outcome effect'. This was despite a mean translation of around 40% in a pre-test. Furthermore, when a climate expert claimed a sea level rise of 100 cm was 'improbable', participants gave a much lower estimate of the expected sea level rise, suggesting they were aware 'improbable' is used by experts to describe values above the expected outcome (Study 5c).

These findings from this novel methodology potentially paint an even gloomier picture for the effectiveness of risk communications using VPEs. An expert who uses 'unlikely' to mean 20% will quickly lose the trust of an audience if they expect 'unlikely' to refer to outcomes which never happen. More immediately, there could be catastrophic consequences for those who choose not to evacuate their homes after hearing the chance of a tsunami is 'unlikely' and mistakenly believe a tsunami will therefore not occur.[1]

One commonly proposed solution to reduce mis-communication (observed from the 'how likely' methodology) is the use of a dual-scale, mixed format approach to express uncertainty, for example, 'It is unlikely (less than 33%)' (see Budescu, Broomell & Por, 2009; Budescu, Por & Broomell, 2012; Budescu, Por, Broomell & Smithson, 2014**,** Harris & Corner, 2011; Harris et al., 2013; Patt & Dessai, 2005;Witteman & Renooij, 2003**)**. Budescu and colleagues have demonstrated that such a 'verbal-numerical' format increased the differentiation of participants' interpretations of VPEs, an effect that replicated across 24 countries (Budescu et al., 2014). A question yet to be explored, however, is whether the addition of a VPE, over a purely numerical communication, might influence interpretations of probability estimates in a way that could be potentially harmful. Teigen and colleagues' findings using the 'which outcome' methodology suggest that they might. A citizen who hears that the chance of a volcanic lava flow extending as far as their village is 'unlikely' may discount the information, believing it will not happen, even if a numerical estimate is also provided.

This possible consequence, however, relies on the assumption that the same result from the 'which outcome' methodology will be obtained even when one potential outcome is of particular consequence. Previous research using the 'how likely' methodology suggests that such an assumption might not necessarily hold, as people's interpretations of VPEs are higher when those VPEs describe a severe outcome than a neutral outcome (Harris & Corner, 2011). More generally, making one outcome particularly consequential in the 'which outcome' methodology will enhance its salience. When considering the potential extent of a volcanic lava flow, for example, the location of a school a certain distance from the volcano might consume the attention of a communicator, such that all risk communications are assumed to be relevant to that particular location. Whether such an influence of saliency could override the effects reported in Teigen et al. (2013, 2014) and Teigen and Filkuková (2013) is a question addressed in the present paper.

The present paper therefore aims to further our understanding of the ramifications of Teigen and colleagues' previous work using the 'which outcome' methodology by testing the robustness of the effect across differing situations and communication formats. Ascertaining the effect of using different communication formats is instructive for designing future risk communication instruments. Furthermore, it enhances our theoretical understanding of the effect by determining whether it is primarily related to the pragmatics of linguistic communication, or linked to something more fundamental about people's understanding of probability and frequency. As well as examining whether different communication formats influenced people's outcome estimates, we tested whether marking certain outcomes as salient would alter the way in which people understood the risk communication. Owing to our underlying interest in consequential risk communications, the study also extended the existing evidence base by investigating scenarios featuring geological hazards.

Method

Participants

155 participants were recruited for this online study via Prolific Academic (www.prolific.ac). They were paid £0.85 upon completion of the study. 8 participants were excluded (6 due to duplicate IP addresses and 2 due to lack of consent) leaving a final sample of 147 (83 male) participants, aged 18-60 years (Mdn = 27).

Design

A 3×4×4 mixed design was employed with communication format (verbal/numerical/mixed) as a between subjects factor; scenario (volcano/flood/earthquake/landslide) and salient site (non-existent/close/far/multiple sites) as within-subjects factors. Scenario and saliency were randomised using the Latin Square Confounded method (Kirk, 1969), such that each participant only saw each scenario and each salient site once, but the combinations of these differed systematically across participants. Participants were required to type a numerical response which corresponded to the outcome that they believed was being described as either "unlikely" (verbal format), "there is a 20% chance" (numerical format), or "unlikely (20% chance)" (mixed format). We focused on the VPE 'unlikely' as it is an approved VPE of the IPCC. 20% is a plausible value for 'unlikely' given the IPCC's likelihood scale, as well as it being the average numerical translation of 'unlikely' in Theil's (2002) meta-analysis.

[1] This effect could occur over and above the effects of directionality (Teigen & Brun, 1995, 1999, 2003).

Materials and Procedure

Participants were first informed about the nature of the study and told they could withdraw at any time during the experiment. After consenting to participate, they were asked to indicate their age and gender, before reading the introductory text. The introductory text informed participants that they would see reliable projections of a model designed to predict future geological events and asked to make a series of judgements about these.

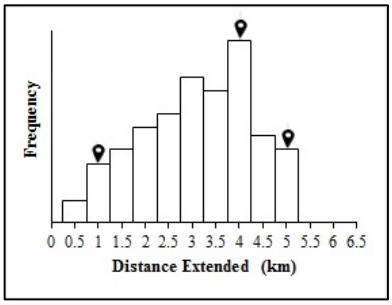
Each of the next four screens showed one of the four vignettes describing outcomes of how far lava flows, floodwater, earthquake tremors and debris flows would extend. These vignettes were developed in conjunction with geologists at the British Geological Survey to ensure they reflected plausible real-world situations. Each vignette was illustrated by a histogram with 10 bars, which reflected the number of times the model had produced the outcome. The shape of the distributions were similar and approximately normal across the scenarios, though the volcano and flood histograms had a slightly negative skew. The zero-frequency options were explicitly included in the histogram. The sentence completion task was presented at the bottom of each vignette (see Figure 1 for an example).

Saliency was manipulated through the inclusion of sites of particular scientific interest, which the event might extend as far as. These sites either homed rare plants or critically endangered animal species (e.g., the last habitat of 'white-spotted Antis' in Figure 1). There were four saliency conditions: No site of interest, one close site of interest (located in the second bin of the histogram), one far site of interest (last bin of the histogram), or multiple sites (second bin, modal bin and last bin, see Figure 1). After completing the study, participants were given a code to claim their reward, thanked and debriefed.

Reminder: The number of times the model has produced each outcome is a reliable indication of how likely that particular outcome is.

Mount Ablon has a history of explosive eruptions forming lava flows. An eruption has been predicted; the figure below shows the model's predictions of the distance extended by lava flows for this eruption, given the volcano's situation and recent scientific observations.

Due to the highly fertile soil and rich vegetation, multiple sites of special scientific interest home to the critically endangered 'white-spotted Antis', exist in the area surrounding the volcano. Sites A, B and C lie **1km**, **4km** and **5km** respectively away from the volcano (shown below). If lava flows reach any of these sites, the last surviving populations of 'white-spotted Antis' in the wild (at the site) would be lost.



Complete the sentence below with a <u>number</u> that seems appropriate in this context.

In the event of an eruption, it is **unlikely (20% chance)** that the lava flow will extend to a distance of ___ km.

Figure 1. Example vignette (volcano scenario, multiple salient sites, mixed format)

Results

Effect of Saliency

Because the different scenarios referred to different geological events, the x-axes on the histograms (see Figure 1) were all different. As it was predicted that responses would be pulled towards salient outcomes, to investigate the effect of saliency we standardised the outcomes across scenarios by 'binning' responses in accordance to where they were in relation to the salient points in the multiple site condition.

There were thus 8 response categories. Responses were similar across all four scenarios. Three Kruskal-Wallis tests were run to investigate if there was an influence of saliency in a) verbal b) numerical and c) mixed formats. These showed that outcome values were not significantly affected by saliency in either the verbal, χ^2 (3) = 0.482, p = .932, numerical, χ^2 (3) = 6.581, p= .087 or mixed format conditions, χ^2 (3) = 3.274, p = .351. In the following analyses,

Effect of Communication Format

Responses were similar across scenarios. Typical outcomes for 'unlikely' were chosen from the higher end of the distribution, from maximum and above maximum observed values. In contrast, typical outcomes for '20% chance' tended to correspond to lower values, primarily chosen from the intermediate values[2]. Results for the mixed format were in the middle of results for the verbal and numerical formats; outcomes tended to be chosen from the intermediate values, but this did not preclude a sizable proportion choosing from maximum and above maximum values (Bins 11 & 12). The contrasting patterns of responses are clearly evidenced in Figure 2.

The proportion of responses indicating high amplitude outcomes (the maximum value present in the histogram or above – bin 11 or 12) was highest in the verbal condition,

[2] Effects of communication format were unchanged if responses were binned into five categories (below minimum, minimum, intermediate, maximum and above maximum), as in Teigen et al. (2013).

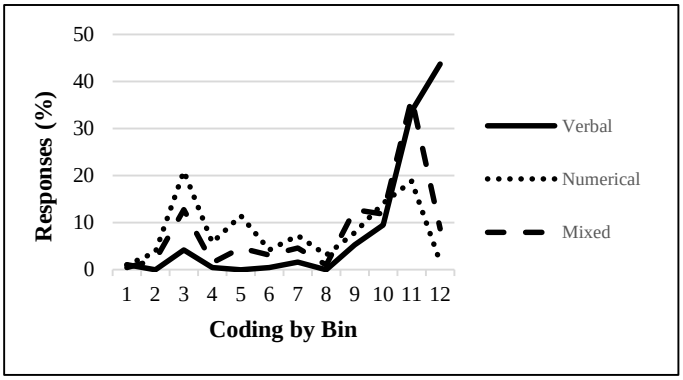


Figure 2. The distribution of responses by communication format.

followed by the mixed format condition. The numerical condition had the lowest proportion of responses indicating high amplitude outcomes, χ^2 (2) = 126.64, p < .001 (see Figure 3). A Kruskal-Wallis test showed a significant difference between the three communication formats χ^2 (2) = 163.29, p < .001. The verbal format yielded the highest estimates (mean rank = 402.61) followed by the mixed format (mean rank = 273.43), with numerical format yielding the lowest estimates (mean rank = 190.37). Three pairwise comparisons using Mann Whitney U-Tests were all significant (all ps < .001).

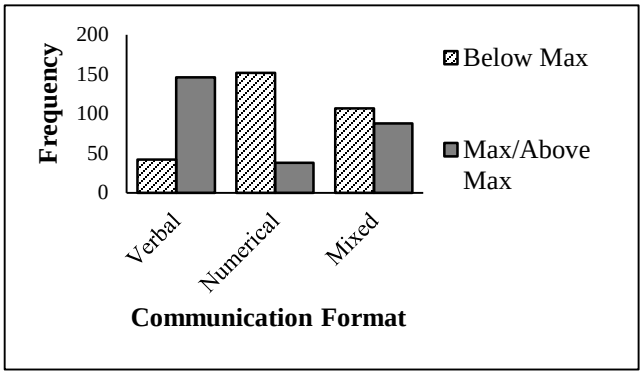


Figure 3. Frequencies of maximum/above maximum responses in the three communication format conditions (collapsed across saliency).

Accuracy of Estimates

We have demonstrated that the 'which outcome' effect is robust against a contextual influence of saliency. We have also shown that numerical and mixed format conditions produce different estimates, but is it possible to identify which format yields the most accurate estimates? The '20%' in the numerical and mixed communication formats enables the calculation of an objectively correct answer to the statement "there is a 20% chance that the *x* will extend to a distance of …" for the four scenarios by using the data the histograms were created with. The correct answer fell in the middle, or mid-to high saliency category in every scenario (either the 7th or 8th bin). It was thus possible to calculate the proportion of participants' estimates which were correct or not. Overall in the numerical and mixed format conditions, only 7% of all estimates (n= 386) were exactly correct. There was no significant association between communication format (numerical/mixed) and number of correct estimates, χ^2 (1) = 0.86, p = .426.

Given the low level of correct estimates, this simple, dichotomous categorisation of responses seemed rather crude. A more sensitive measure of accuracy was to therefore calculate a difference score (observed response – correct response) which was standardised across scenarios. We first calculated the mean and standard deviation of the outcomes implied by the histograms (see Figure 1) for each of the four scenarios, assuming that each histogram represented 200 datapoints. We consequently represented both the objectively correct response and the observed response as a z-score in this distribution. The difference score was subsequently calculated by subtracting the 'correct z-score' from the 'observed z-score.' Lower difference scores, indicating greater accuracy, were observed for the mixed format (M = -.003, SD = 1.4) than for the numerical format (M = -0.8, SD = 1.4), F (1, 377) = 34.45, p < .001, η^2_p = 0.08 (inferential statistics performed on reflect and square root transformed values to correct skew- results unchanged without transformation). There was no effect of scenario (p = .111), nor was there an interaction between communication format and scenario (F < 1).[3]

Discussion

We replicated Teigen et al.'s (2013) results and tested whether these would hold for numerical and mixed format expressions of probability. We also examined whether marking certain outcomes as salient would influence interpretations of the risk communication. We found evidence that the tendency to describe outcomes at the very end, or beyond the range of, a distribution as 'unlikely' generalised to consequential scenarios, wherever the word 'unlikely' was included - the verbal and mixed format conditions. This tendency was not apparent in the numerical condition. We found no evidence that responses were affected by increasing the saliency of the outcome.

Finding the 'which outcome' effect is generalisable to consequential scenarios in the applied domain of geological hazards has implications for organisations who are responsible for communicating the risk of such dangers. As previously suggested, there exists the real possibility that citizens will completely disregard the communication, believing that if the probability of a disastrous event is described as 'unlikely', it will not affect them. The potential for catastrophic consequences increases further given that

[3] Similar results to this study were also obtained using a voluntary, laboratory sample (n= 81).

marking outcomes as salient was not sufficient to attenuate the 'which outcome' effect.

Our findings also clearly demonstrate that format influences interpretations of risk communications. The fact that the 'which outcome' effect extends to mixed but not numerical formats indicates that the effect is related to the pragmatics of communication rather than people's understanding of probability and frequencies. Finding the effect occurred more in the mixed format than numerical conditions is of particular relevance to current literature, given the recent recommendations to use a dual scale, mixed format approach to express uncertainty (Budescu et al., 2009; Budescu et al., 2012; Budescu et al., 2014**,** Harris & Corner, 2011; Harris et al., 2013; Patt & Dessai, 2005;Witteman & Renooij, 2003). Our results suggest that this may not solve the problem of misinterpretations as much as first thought.

Given that this preference for values at the high end of the distribution was not present in the numerical condition, it was unexpected that estimates in the mixed format were more accurate than those in the numerical condition. This increase in accuracy was even more surprising given that the presence of 'unlikely' in the verbal condition was enough to significantly shift estimates towards outcomes at the higher end of the distribution and to those with a predicted 0% frequency of occurrence. There are two complementary explanations for this which stem from the proposition that VPEs contain an 'inbuilt hint' (Teigen & Brun, 1995) which provides the participant with extra information.

The first explanation focuses on the calculation of the objectively correct answer. A shallow reading of the scenario might have led those in the numerical condition to (wrongly) calculate 20% from the 'likely' end of the scale (given 'chance' focuses one's attention towards the occurrence of the event). This would shift estimates to the left of the scale, supported by Figure 2, which shows the numerical condition has two peaks, one to the left of the scale and one to the right, clustered around the correct answer's location. In contrast, those in the mixed format condition have additional information, with the inclusion of 'unlikely', which tells them to start from the 'unlikely' side of the scale, discouraging a shallow reading. This would shift their estimates towards the right of the scale, evidenced by the big peak to the right in Figure 2.

The extra information contained in VPEs is: "not only the probabilistic premise, but also the behavioural conclusion" (Teigen & Brun, 1999, p.163). VPEs can be positive or negative. If an outcome is described with a positive expression (e.g. 'possible'), the focus is on the fact that the outcome may occur (probability > 0), but if it is described with a negative expression (e.g. 'unlikely'), the focus is on the fact that it may not occur (probability < 1). Numerical probabilities tend to be presented in conjunction with other terms such as 'chance', which have positive directionality (as was the case in our study). Therefore, participants in the numerical condition may have interpreted chance as an indication that the event would indeed happen and thus given estimates closer to the likely end of the scale (the left). In contrast, those in the mixed format condition may have seen the 'unlikely' as extra information, interpreted it as pointing towards the non-occurrence of the outcome and thus given estimates closer to the 'unlikely' end of the scale.

Teigen and Brun (1999) suggest that neutrality may be achieved if two VPEs of opposing directionality are presented, but this was not achieved in the current study. Although the response pattern for the mixed format condition lies between the numerical and verbal formats, the general preference for values at the high end of the distribution in the mixed format condition is still present in around 45% of cases (see Figure 2) suggesting that participants are focusing more on the term 'unlikely', but why is this?

It is possible that people who focused more on the term 'unlikely' had lower numeracy levels and thus felt uncomfortable using the '20%' to form their estimates. Research has shown that less numerate decision makers are: "left with information that is less complete and less understood, lacking in the complexity and richness available to the more numerate" (Peters et al, 2006, p.412). Those lower in numeracy are also more likely to trust information presented in verbal form (Gurmankin, Baron & Armstrong, 2004). It would therefore be prudent to include a measure of numeracy in future studies to establish if the focus on 'unlikely' lessens as levels of numeracy increase.

Overall, although our results show the mixed format yielded more accurate responses, we would be hesitant to fully endorse calls to adopt such an approach as *the* solution to the problem of misinterpretations in risk communication research. Only a small proportion of responses were exactly correct. It is arguable that the increased endorsement of outcomes with a 0% frequency of occurrence as 'unlikely' in the mixed format over the numerical format (see Figure 2) represents a more consequential error, which is most critical to avoid. The degree to which the enhanced accuracy in the mixed format condition will generalise across different outcome distributions is also an open question. These are questions which future research can seek to address to maximise the effectiveness of future risk communication.

Conclusion

The present research provides an example of the disparity between what is communicated by scientists and what is understood by the audience in the context of geological risk communications. Whilst it is generally acknowledged that there is no 'optimal' presentation format and no one single 'fix' for risk communication (Budescu et al., 2012), identifying instances in which the format of uncertainty has a significant impact on audience's understanding is key to improving risk communication. Our findings show that the 'which outcome' effect extends to risk communications which use a mixed-format approach. Our study has significant practical implications for organisations responsible for communicating risk. Not only may people discount a hazard described as 'unlikely', the addition of a numerical translation may not be enough to prevent this disregard, leading to potentially catastrophic consequences.

Acknowledgements

SJ was supported by a UCL IMPACT studentship, half funded by the British Geological Survey. RML's contribution is published with permission of the Director, British Geological Survey.

References

Bonnefon, J. F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*(9), 747-751.

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*, 391-405.

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. In J. R. Busemeyer, R. Hastie, & D. L. Medin (Eds.), *Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 32: Decision Making from a Cognitive Perspective)* (pp. 275-318). San Diego, CA: Academic Press.

Budescu, D. V., Broomell, S., & Por, H-H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science, 20*, 299-308.

Budescu, D. V., Por, H.-H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change, 113*(2), 181-200.

Budescu, D. V., Por, H-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4*, 508-512.

Cartlidge, E. (2012, October 22). *Earthquake Experts Convicted of Manslaughter*. Retrieved from: http://www.sciencemag.org/news/2012/10/earthquake-experts-convicted-manslaughter?_ga=1.23557882.1178859578.1453738159 January 31st, 2016.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely=rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, *9*(3), 153-172.

Gurmankin, A. D., Baron, J., &, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication. *Med Decis Making*, *24*(3), 265-271.

Harris, A. J. L. & Corner, A. (2011). Communicating environmental risks: clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*(6), 1571-1578.

Harris, A. J. L., Corner, A., Xu, J., & Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK. *Climatic Change, 121*, 415-425.

Juanchich, M., Sirota, M., & Butler, C. L. (2012). The perceived functions of linguistic risk quantifiers and their effect on risk, negativity perception and decision making. *Organizational Behavior and Human Decision Processes*, *118*(1), 72-81.

Kirk, R., E. (1969). *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, CA: Brooks/Cole.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Held, H., et al. (2010). *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties.* IPCC cross-working group meeting on consistent treatment of uncertainties, Jasper Ridge,CA. [retrieved from: https://www.ipcc.ch/pdf/supporting-material/uncertainty-guidance-note.pdf January 31st , 2016.

Patt, A., & Dessai, S. (2005). Communicating uncertainty: lessons learned and suggestions for climate change assessment. *Comptes Rendus Geoscience*, *337*(4), 425-441.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*, 407-413.

Siegrist, M., Gutscher, H., & Earle, T. (2005). Perception of risk: the influence of general trust, and general confidence. *Journal of Risk Research, 8*(2), 145-156.

Sjöberg, L. (2000). Factors in Risk Perception. *Risk Analysis*, *20*(1), 1-11.

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica, 88*(3), 233-258.

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes*, *80*(2), 155-190.

Teigen, K. H., & Brun, W. (2003). Verbal probabilities: A question of frame? *Journal of Behavioral Decision Making*, *16*(1), 53-72.

Teigen, K. H., & Filkuková, P. (2013). Can > Will: Predictions of What Can Happen are Extreme, but Believed to be Probable. *Journal of Behavioral Decision Making*, *26*(1), 68-78.

Teigen, H. K., Juanchich, M., & Filkuková, P. (2014). Verbal probabilities: an alternative approach. *Q J Experimental Psychology*, *67*(1), 124-146.

Teigen, K. H., Juanchich, M., & Riege, A. H. (2013). Improbable outcomes: infrequent or extraordinary? *Cognition, 127*(1), 119-139.

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research*,*5*(2), 177-186.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 781–789.

Witteman, C., & Renooij, S. (2003). Evaluation of a verbal-numerical probability scale. *International Journal of Approximate Reasoning*, *33*, 117-131.

Visual Statistical Learning Deficits in Children with Developmental Dyslexia: an Event Related Potential Study

Sonia Singh (ssingh11@student.gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302 USA

Anne M. Walk (amcclur3@illinois.edu)
Department of Kinesiology and Community Health, University of Illinois, 405 N. 900 S. Goodwin Ave, Urbana, IL 61801 USA

Christopher M. Conway (cconway@gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302 USA

Abstract

A growing body of research suggests that individuals with developmental dyslexia perform below typical readers on non-linguistic cognitive tasks involving the learning and encoding of statistical-sequential patterns. However, the neural mechanisms underlying such a deficit have not been well examined. The aim of the present study was to investigate the ERP correlates of sequence processing in a sample of children diagnosed with dyslexia using a probabilistic visual serial learning paradigm. The behavioral results revealed that whereas age-matched typically developing children (n=12) showed learning in the task as reflected by their response times, the children with dyslexia (n=8) likely showed difficulty in learning. In conjunction with these behavioral results, the ERPs of the typically developing children showed a P300-like response indicative of this paradigm (Jost et al., 2015); whereas, the children diagnosed with a reading disorder showed no such ERP effects. These findings are consistent with the idea that differences in statistical-sequential learning ability might underlie the reading deficits observed in developmental dyslexia.

Keywords: Developmental dyslexia; statistical learning; sequential learning; implicit learning; ERPs.

Introduction

Statistical-sequential learning refers to the ability to learn statistically structured sequential patterns from the environment (Lashley, 1951; Saffran, Aslan, Newport, 1996). Statistical learning is thought to be important for the acquisition of language. For instance, in spoken language, linguistic units (e.g., phonemes, syllables, words) are organized in statistically structured sequences according to the specific language's phonology, phonotactics, semantics, and syntax. A growing body of research suggests that variations in statistical learning ability are associated with spoken language ability, in adults (Conway, Bauernschmidt, Huang, & Pisoni, 2010; Misyak, Christiansen, & Tomblin, 2010), children (Kidd, 2012), and infants (Shafto, Conway, Field, & Houston, 2012). Recent research also suggests that this same relationship between statistical learning and spoken language may hold true for statistical learning and written language. For instance, visual statistical learning was found to be related to reading ability in adults and children, even after controlling for age and attention (Arciuli & Simpson, 2012). Similarly, performance on a variant of the serial response time (SRT) task has also been shown to predict reading ability in a sample that included both healthy controls and adults diagnosed with dyslexia (Bennet, Romano, Howard, & Howard, 2008).

Developmental dyslexia (DD) is a learning disability that specifically impairs a person's ability to read despite having normal intelligence and ample opportunity for learning. Common characteristics among people with dyslexia are difficulty with phonological processing (the manipulation of sounds) (Bradley & Bryant, 1983; Snowling, 2000) and spelling (NINDS, 2011; Gabrieli 2009). In the standard view of dyslexia, individuals present with difficulty in reading but appear to process other information in a typical manner. In recent decades, however, it has become apparent that developmental dyslexia may be associated with impairments to other cognitive abilities such as motor functioning (Orban, Lungu, & Doyon, 2008), implicit learning (Du & Kelly, 2013), and cerebellar dysfunction (Nicolson, Fawcett, & Dean, 2001).

Of the many studies suggesting that there may be broader cognitive impairments underlying developmental dyslexia, a single commonality underlies many: sequential learning ability. For instance, individuals with dyslexia have been shown to perform below typical reading peers in variations of the SRT and other related sequence learning paradigms (Du & Kelley, 2013; Howard, Howard, Japiske, & Eden, 2006; Jiminez-fernandez, 2011; Vicari, Marotta, Menghini, Molinari, & Petrosini, 2003). Thus, there is mounting behavioral evidence for dyslexia to be associated with impairments in implicit statistical-sequential learning mechanisms. However, the neural mechanisms associated

with this proposed learning deficit have not been well explored. Furthermore, many of the previously mentioned tasks involved a motor component to learning; as such it is important to determine whether impaired sequential learning is due solely to difficulties with motor learning or whether deficits are also found using tasks that are not as dependent upon motor responses.

The Current Study

The purpose of the current study was to examine the neural correlates of statistical learning in children who have been diagnosed with DD compared to age-matched, typically-reading children. To this aim, we measured event-related potentials (ERPs) while children were engaged in a visual statistical learning paradigm previously used by Jost, Conway, Purdy, Walk, and Hendricks (2015). Eight children previously diagnosed with DD and twelve TD children participated. The learning task involved the presentation of a series of visual stimuli wherein target stimuli could be probabilistically predicted based on the preceding stimulus. ERPs to three different types of predictor stimuli reflecting high, low and zero probability of being followed by the target were compared across DD and TD groups. Based on the Jost, Conway, Purdy, Walk & Hendricks, (2015) study that demonstrated a P300-like ERP component that was associated with learning in the task, we investigated waveforms within the same 400-700ms time-window as used by Jost et al., (2015). We also examined behavioral correlates of learning as measured by response times (RTs) to the target stimuli. We predicted that if developmental dyslexia is associated with a deficit to general-purpose statistical learning mechanisms, the children with DD would be poorer at learning the predictor-target statistical patterns, as reflected by both the behavioral and ERP data, compared to the TD children. Furthermore, because the ERP effects are time-locked to the presentation of the predictor stimuli before any motor responses are made, any observed differences in the waveforms would suggest that the locus of impairment is at a perceptual or cognitive level, rather than only at the motor response level of processing.

Method

Participants

Twenty children (ages 8-12 years) were recruited from the greater Saint Louis region, eight of whom had a prior diagnosis of reading dyslexia (DD group) and the remaining had no prior diagnoses of cognitive, emotional, or learning disorders (TD group). The TD group was recruited through advertisements or by word of mouth and the DD group was recruited through a flier placed in the newsletter of a city school that specializes in teaching children with learning disorders. All participants' families were compensated $30.00- $40.00 for a 2 hour testing session and travel, and each child was given a small toy ($5.00-$10.00 in value). The resulting average age of the children across groups was similar, though the DD group [*M*=10.7] was slightly older than the TD [*M*=9.4] group on average. [*t*(18)=-1.707, *p*=.105] (and are still similar when one TD group participant was removed due to missing data on the cognitive tests[1]). Table 1 shows the demographic characteristics for the two groups.

Table 1: Descriptives for the children with developmental dyslexia (*DD*) and age-matched typically developing (*TD*) children [*see procedure below for explanation*].

	N	Gender	Mean Age
DD	8	5M, 3F	10.7 years
TD	12	8M, 4F	9.4 years

Table 2: Cognitive (raw) scores for both groups and Reading (raw) scores for the *DD* group

	DD*(N=8)*	TD*(N=11)*
BD	33.50	26.27
DS	15.38	13.91
WR	22.25	-
PD	11.00	-

Procedure

In addition to the statistical learning task (described below), we administered 2 sub-tests from the Wechsler Intelligence Scale for Children IV (WISC) (Wechsler, Kaplan, Fein, Kramer, Morris, Delis, & Maelender, 2003). We assessed participants' level of perceptual reasoning using the Block Design (BD) subtest. This test assesses visual-motor and visual-spatial skills by requiring children to recreate a 2-dimensional printed figure using 3-dimensional blocks, within a specific time frame. We also used the Digit Span (DS) sub-test as an index of short-term memory capacity as it provides a measure of a child's ability to retain new information, concentrate and manipulate input, thus demonstrating cognitive flexibility. The results of the BD and DS assessments are provided in Table 2. As one can see, the two groups of children obtained very similar scores, which were not statistically significant from one another for either subtest [BD: *t*(17)=-1.288, *p*=.215 and DS: *t*(17)=-1.259, *p*=.225].

[1] The resulting average age of the children across groups was similar, though the DD group [*M*=10.7] was slightly older than the TD [*M*=9.4] group on average. [*t*(17)=-1.678, *p*=.112]

The DD children were also administered the Pseudoword decoding and Word Reading sub-tests of the Wechsler Individual Achievement Test-II (WIAT) (Wechsler, 2005). The Word Reading (WR) sub-test assesses basic phonological skills and the Pseudoword Decoding (PD) sub-test assesses the child's ability to use phonetic decoding skills while reading from a list of nonsense words. The results of these two subtests are shown in Table 2. These scores indicate that the DD group performed at least two standard deviations below age matched peers on both WIAT sub-tests.

Statistical Learning Task

In the statistical learning task, children were presented with consecutive series of colored circles on a computer monitor one at a time (Figure 1). Stimuli were presented electronically using E-Prime 2.0.8.90 software (Psychology Software Tools, Pittsburgh, PA), on a Dell Optiplex 755 computer. Each visual stimulus was presented in the center of the screen on top of a dark background, displayed for 500ms Participants were instructed to press a button on a button box each time a target (T) color appeared. Every trial began with one to five occurrences of a standard (S) stimulus in the center of the screen. Following the presentation of these S stimuli, one of three possible predictor stimuli appeared, with each one predicting the target stimulus with varying levels of probability. When a high predictor (HP) color appeared, the target color followed 90% of the time; when a low predictor (LP) color appeared, the target color followed 20% of the time. When T did not appear (10% of the time for HP; 80% of the time for LP) the S followed instead. A zero-predictor (ZP) color was never followed by T, but always followed by S. The end of one trial was

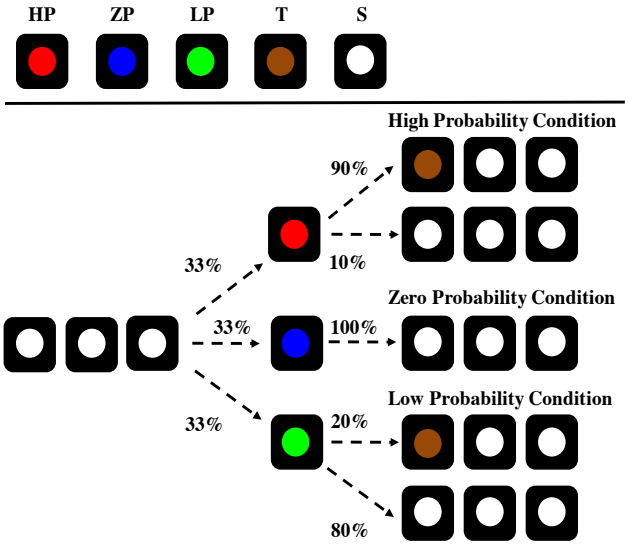


Figure 1. Diagram of the Statistical learning task

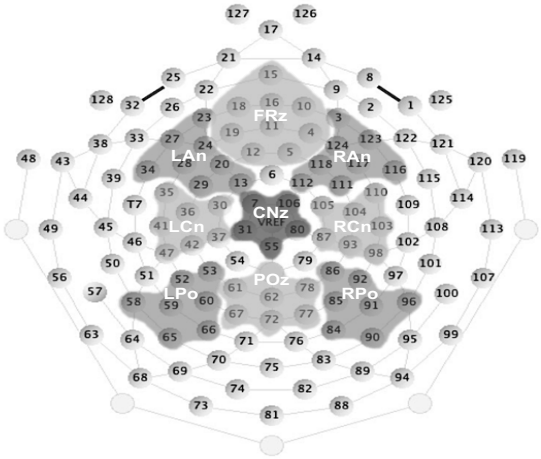


Figure 2: 128 sensors
EEG net with the highlighted nine regions of interest.

immediately followed by the next trial. Each of the three predictor colors occurred with the same frequency and the assignment of colors (red, blue, green, brown, white) to the three predictors, Ss, and Ts, was determined randomly for each participant. Note that Figure 1 shows an example depiction of the task with the HP being assigned to the red stimulus, ZP to the blue stimulus, etc.

Each predictor condition (HP, ZP and LP) was presented 50 times. Within each block, the trials were presented randomly and the end of one trial segued seamlessly into the next trial, so that participants could not distinguish the onset or offset of one trial from another. There were 50 trials of each predictor (for a total of 150 trials). Note that participants were given no instruction of the predictor-target statistical contingencies. Instead, the participant was expected to implicitly learn the statistical relationships between each predictor and the target, with learning to be expected to be observed through both response times and ERPs (as per Jost et al., 2015).

Electroencephalography Acquisition

The electroencephalograph (EEG) was acquired from 128 scalp sites using an Electrical Geodesic Inc. sensor net (Figure 2) and was pre-processed using Net Station Version 4.3.1 with subsequent processing using custom scripts written in Matlab (version R2012b 8.0.0783, The MathWorks) and the EEGLAB toolbox (version 10.2.2.2.4a; Delorme & Makeig, 2004). Electrode impedances were kept below 50 kΩ. The EEG was acquired with a 0.1 to 100 Hz band-pass at 250 Hz and then low-pass filtered at 30 Hz. The continuous EEG was segmented into epochs -200ms to +1000ms with respect to the predictor onset. ERPs were baseline-corrected with the 200ms prestimulus data and averaged-referenced. Individual ERPs were computed for each participant, probability condition, and electrode.

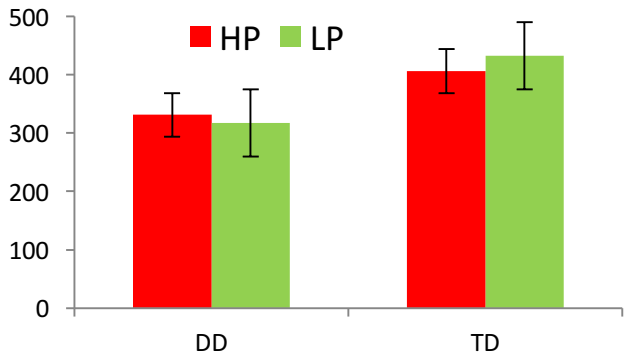


Figure 3: Response times (RT) across the two groups for the high (red) and low probability condition (green)

Statistical Analysis

Statistical calculations were performed on the individual mean amplitude ERPs within 400-700ms (as used by Jost et al, 2015). To analyze the effect of cortical topography, nine regions of interest (ROIs, Figure 2) were defined: left (LAn), middle (FRz), and right anterior (RAn); left (LCn), middle (CNz), and right central (RCn); and left (LPo), middle (POz), and right posterior (RPo) regions.

Behavioral analyses were conducted across DD and TD groups for response times. Pearson correlations were also performed on both groups with the WISC sub-tests as well as with the DD group and the WIAT sub-tests.

Results

RT data (Figure 3) was examined for only 10 TD children and on 6 of the DD children (due to missing data for 2 children in both groups). A 2 X 2 ANOVA with 2 levels of predictor (HP/ LP) and 2 groups (TD/DD) with RT as dependent variable revealed a significant main effect for predictor only *[*$F(1, 28) = 24.018$. $p < .001$]. The effect for group [$F(1, 28) = .086$. $p = .772$] and the group x predictor interaction were non-significant [$F(1, 28) = 1.03$. $p = .319$]. Even though the predictor X group interaction was non-significant (likely due to lack of statistical power), visually, it appears that RTs were quite different for the TD and DD groups, with the TD group showing the expected facilitation in RTs for the HP compared to the LP predictors, but the DD group not showing this same effect. This could indicate that for the DD children, the subtle differences between the HP/ LP were left undetected. However, a larger sample size is needed before making strong claims.

Figure 4 (left panel) displays grand average ERPs for the DD participants, within the 400-700ms range. Visual inspection indicates that the waveforms for the HP and LP predictors have the same amplitude but are larger than the ZP predictor. Figure 4 (right panel) displays grand average ERPs for the TD participants for each predictor condition, within the 400-700ms (across all 9 ROIs). Visual inspection indicates that the waveform for the HP predictor has a larger amplitude compared to the ZP and LP predictor waveforms.

Two one–way ANOVAs were performed separately for each group (DD, TD) within the 400-700ms window with EEG amplitude as the dependent variable and predictor conditions (HP, ZP and LP) as the independent variables. In the DD group, there were significant differences among predictor conditions [$F(2, 1941) = 24.762$, $p<.001$]. However, Tukey's post hoc-tests indicated that there were no significant differences between the HP and LP groups ($p=.865$) but there were significant differences between HP and ZP ($p<.001$) and LP and ZP ($p<.001$). Analyses for the TD group revealed that there were significant differences between the predictor conditions [$F(2, 2913) = 4.325$, $p=.02$]. Tukey's post hoc tests indicated significant differences between the HP and LP ($p=.01$), and HP and ZP ($p<.05$) but not between the ZP and LP groups ($p=.918$).

Pearson correlation analyses for all but one child with

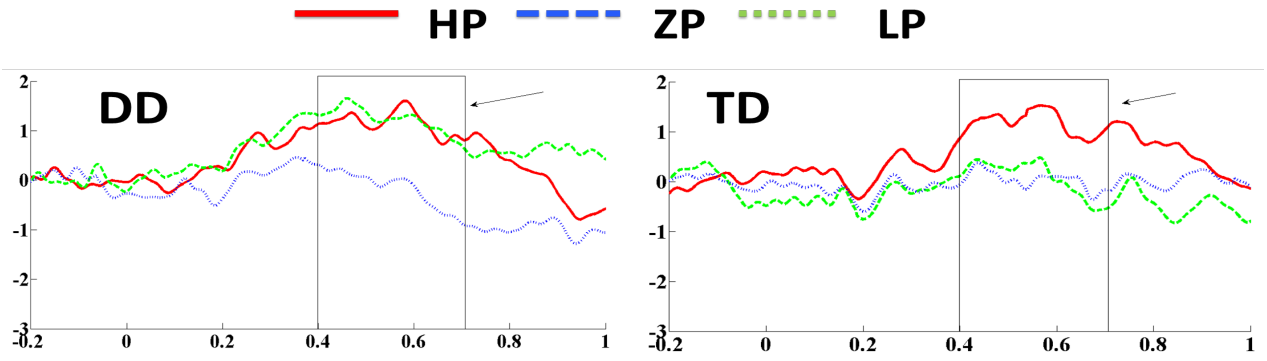


Figure 4: Grand averages are in response to the high predictors (*HP*, red solid line), zero predictors (*ZP*, green dotted line) and low predictor (*LP,* blue dashed line) (vertical axis: electric potential in μV, positivity upward; horizontal axis: time in seconds).

missing data (N=19) indicated that scores on the Digit span (DS) sub-test were strongly correlated with scores on the Block design (BD) sub-test [r(17)=.543; p=.02]. Pearson correlation analyses for only the DD group (N=8) children showed that scores on the Pseudoword Decoding (PD) sub-test were strongly correlated with scores on the Word Reading (WR) sub-test of the WIAT [r(6)=.967; p<.001]. Additionally, non-significant correlations were observed between: WR and statistical learning: [r(6)=.375, p = .360] and PD and statistical learning: [r(6)=.520, p = .186]. The lack of significant correlations is largely due to a small sample size issue.

Discussion

In this study, we examined the neural correlates of visual statistical-sequential learning in children who have been diagnosed with a reading disorder. The results showed that: (1) Relative to the TD group, typical learning patterns in RT data from children diagnosed with DD seemed less apparent, indicating unlikely encoding of the predictor-target relationships; and (2) the DD group showed atypical ERP waveforms within the 400-700ms time-window, compared to TD controls.

In a previous study using this same visual paradigm with a group of TD children, Jost et al. (2015) observed a P300-like component elicited by the HP but not the LP or ZP stimuli, similar to what we observed in the current sample of TD children. The P300 is regarded as an index of target detection and evaluation (van suijen et al., 2006) and has also been observed in other learning tasks (Baldwin & Kutas, 1997; Carrion & Bly, 2007; Russeler et al., 2003). Jost et al. (2015) suggested that the P300, typically observed during the occurrence of an infrequent target stimulus, "shifted" earlier in the input stream so that it now occurred in response to a stimulus that predicted the target with a high level of probability. That is, after sufficient exposure to the sequential statistics of the input array, the participants' brains treated the high predictor stimulus as if it were the target itself, displaying the prototypical P300 response. Regardless of the actual cognitive interpretation of the P300, it is clear that it reflects the participants treating the high predictor stimulus differently than both the low and zero predictor stimuli, presumably on the basis of having learned that this stimulus predicts the target with a high level of reliability.

In contrast, the DD group showed both a lack of facilitation of reaction times and an atypical ERP waveform pattern. Rather than showing the P300 effect to the HP stimulus, the DD group showed it for both the HP and the LP stimuli. This suggests that these children were unlikely to have encoded the statistical probabilities between predictors and target. Rather than learning that the HP stimulus was "special" in terms of its predictive power, it appears that these children learned that both the HP and LP were predictive of the target. On this account, they have not learned the subtle distinction in terms of the predictor-target probabilities that differentiate the HP from the LP stimulus (in terms of predicting the target with a 90% vs. 20% probability). This pattern of performance is also reflected in their RTs, by not responding faster to targets following the HP stimulus. Thus, it would appear that the children with developmental dyslexia were unable to learn even the most basic of statistical-sequential dependencies contained within this visual input stream.

How do the current findings relate to some of the prominent theories about the causes of developmental dyslexia, namely the phonological deficit and magnocellular deficit theories? The phonological deficit explanation is a prominent theory suggesting that dyslexia is a language-based disorder characterized by difficulties in single-word decoding (Orton, 1995) and phonological processing (Snowling, 2000). According to this theory, these individuals experience difficulties in perceiving and parsing phonemes, resulting in the inability to establish phoneme-grapheme connections. Although prominent, this theory does not explain other low-level visual, sensory, and motor coordination deficits that have also been associated with dyslexia. These shortcomings can instead be accounted for by the magnocellular deficit theory (Eden, Van- Meter, Rumsey, & Zeffiro, 1996). This theory postulates weaknesses in the perception of visual, rapid moving stimuli. Such degraded visual input is due to poor binocular fixation while reading. Its physiological manifestation is at the central nervous system level with impaired sensitivity of cells within the retinocortical magnocellular pathway (Stein, 2001). In the context of the present findings, the magnocellular theory would seem to be most relevant as it might help account for difficulties the DD children encountered while performing the visual statistical learning task. Future work will need to explore the specificity of this statistical learning deficit, that is, to what extent is it also apparent for other statistical learning tasks such as those incorporating auditory (non-linguistic) input streams or even visual-spatial patterns.

In conclusion, our findings suggest that children with DD show difficulty in implicitly learning statistical-sequential visual patterns. Because learning was indexed at the perceptual level by ERPs, and was not dependent upon a motor response, it appears that the learning deficit is not based on motor learning but reflects a more perceptual or cognitive learning problem. Additional research is required at both cognitive and neurophysiological levels in order to clarify the nature of this impairment and how it relates to or causes reading disability.

Acknowledgments

We thank the NIH (R01DC012037) for their financial support.

References

Arciuli, J. & Simpson, I.C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*, 286-304.

Baldwin, K. B., & Kutas, M. (1997). An ERP analysis of implicit structured sequence learning. *Psychophysiology*, *34*(1), 74-86.

Bennett, I.J., Romano, J.C., Howard, J.H., & Howard, D. V. (2008). Two forms of implicit learning in young adults with dyslexia. *Annals of New York Academy of Sciences, 1145,* 184-198. doi: 10.1196/annals.1416.006

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature*.

Carrión, R. E., & Bly, B. M. (2007). Event-related potential markers of expectation violation in an artificial grammar learning task. *Neuroreport*, *18*(2), 191-195.

Conway, C.M., Bauernschmidt, A., Huang, S.S., & Pisoni, D.B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356-371.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Du, W., & Kelly, S. W. (2013). Implicit sequence learning in dyslexia: a within-sequence comparison of first-and higher-order information. *Annals of dyslexia*, *63*(2), 154-170.

Eden, G. F., VanMeter, J. W., Rumsey, J. M., Maisog, J. M., Woods, R. P., & Zeffiro, T. A. (1996). Abnormal processing of visual motion in dyslexia revealed by functional brain imaging. *Nature, 382*(6586), 66-69

Gabrieli, J. D. (2009). Dyslexia: a new synergy between education and cognitive neuroscience. *Science*, *325*(5938), 280-283.

Howard, J. H., Howard, D. V., Japikse, K. C., & Eden, G. F. (2006). Dyslexics are impaired on implicit higher-order sequence learning, but not on implicit spatial context learning. *Neuropsychologia*, *44*(7), 1131-1144.

Jiménez-Fernández, G., Vaquero, J. M., Jiménez, L., & Defior, S. (2011). Dyslexic children show deficits in implicit sequence learning, but not in explicit sequence learning or contextual cueing. *Annals of dyslexia*, *61*(1), 85-110.

Jost, E., Conway, C.M., Purdy, J.D., Walk, A.M., & Hendricks, M.A. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. *Brain Research.* doi: 10.1016/j.brainres.2014.10.017.

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*, 171-184.

Lashley, K. S. (1951). The problem of serial order in behavior. 1951, 112-135.

Misyak, J.B., Christiansen, M.H., & Tomblin, J.B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*, 138-153.

Nicolson, R. I., Fawcett, A. J., & Dean, P. (2001). Developmental dyslexia: the cerebellar deficit hypothesis. *TRENDS in Neurosciences, 24*(9), 508-511.

NINDS Dyslexia Information Page: National Institute of Neurological Disorders and Stroke. http:// www. ninds. nih. gov/ disorders/ dyslexia/ dyslexia. htm. Accessed 20th November 2014.

Orban, P., Lungu, O., & Doyon, J. (2008). Motor sequence learning and developmental dyslexia. *Annals of the New York Academy of Sciences*, *1145*(1), 151-172.

Orton, S. (1995). Definition of dyslexia; re-port from committee of members. *Perspectives, 21*, 16-17.

Rüsseler, J., Hennighausen, E., Münte, T. F., & Rösler, F. (2003). Differences in incidental and intentional learning of sensorimotor sequences as revealed by event-related brain potentials. *Cognitive Brain Research*, *15*(2), 116-126.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Shafto, C.L., Conway, C.M**.**, Field, S.L., & Houston, D.M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy*, *17,* 247-271.

Snowling, M. J. (2000). Language and literacy skills: Who is at risk and why. Speech and language impairments in children: Causes, characteristics, intervention and outcome, 245-259.

Stein, J. (2001). The magnocellular theory of developmental dyslexia*. Dyslexia, 7*(1), 12-36.

van Zuijen, T. L., Simoens, V. L., Paavilainen, P., Näätänen, R., & Tervaniemi, M. (2006). Implicit, intuitive, and explicit knowledge of abstract regularities in a sound sequence: an event-related brain potential study. *Journal of Cognitive Neuroscience*, *18*(8), 1292-1303.

Vicari, S., Marotta, L., Menghini, D., Molinari, M., & Petrosini, L. (2003). Implicit learning deficit in children with developmental dyslexia. *Neuropsychologia*, *41*(1), 108-114.

Wechsler, D. (2005). Wechsler Individual Achievement Test 2nd Edition (WIAT II). London: The Psychological Corp.

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maelender, A. (2003). Wechsler intelligence scale for children: Fourth edition (WISC-IV) [Assessment instrument]. San Antonio, TX; Pearson.

The St. Petersburg Paradox: A Subjective Probability Solution

Hongbin Wang (hwang@tamhsc.edu)
Yanlong Sun (ysun@tamhsc.edu)
Jack W. Smith (jwsmith@tamhsc.edu)
Texas A&M University Health Science Center
2121 West Holcombe Blvd, Suite 1109, Houston, TX 77030 USA

Abstract

The St. Petersburg Paradox (SPP), where people are willing to pay only a modest amount for a lottery with infinite expected gain, has been a famous showcase of human (ir)rationality. Since inception multiple solutions have been proposed, including the influential expected utility theory. Criticisms remain due to the lack of a priori justification for the utility function. Here we report a new solution to the long-standing paradox, which focuses on the probability weighting component (rather than the value/utility component) in calculating the expected value of the game. We show that a new Additional Transition Time (AT) based measure, motivated by both physics and psychology, can naturally lead to a converging expected value and therefore solve the paradox.

Keywords: human judgment and decision making, probability, St. Petersburg Paradox,

Fate laughs at probabilities.
-- E. G. Bulwer-Lytton

Introduction

Suppose you are offered the following gamble:

- Toss a fair coin. If you get a head, you are paid \$1 and the game is over. Otherwise, toss again.
- If you get a head in the second tossing, you are paid \$2 and the game is over. Otherwise, toss again.
- If you get a head in the third tossing, you are paid \$4 and the game is over. Otherwise, toss again.
- … Game continues until you get a head. If you get a head in the nth tossing, you will be paid $\$2^{n-1}$.

How much are you willing to pay to play this gamble?

A simple calculation shows that the gamble's expected value, S, is infinite:

$$S = \$\sum_{n=1}^{\infty} p^n 2^{n-1} = \$\sum_{n=1}^{\infty} (\frac{1}{2})^n 2^{n-1} = \$(\frac{1}{2} + \frac{1}{2} + ...) \qquad \text{Eq. 1}$$

where n is the number of tosses to get the first head (i.e., after a steak of n-1 tails, one gets a head, and the game is over).

The question is, are you willing to pay any price for a right to play this game? Probably not. More than three hundred years ago, in 1713, Nicolas Bernoulli, a young Swiss mathematician, first proposed this problem and pointed out that a sensible person would only be willing to pay very little to play the game. This constitutes a contradiction, which nowadays is called the St. Petersburg Paradox (SPP).

A Little History

The SPP was so named after the eponymous Russian city, where Daniel Bernoulli, a mathematician and Nicholas Bernoulli's cousin, published his classical solution to the problem in 1738. However, the problem was initially proposed by Nicolas Bernoulli in 1713, who was clearly troubled by it. According to him, while the expectation of game gain was infinity, the player would be guaranteed to lose since it is "morally impossible" that one not achieve a head in a finite number of tossing.

In 1728, Gabriel Cramer, another Swiss mathematician, wrote to N. Bernoulli and suggested a solution. In Cramer's solution, money's quantity was replaced by its "moral value", representing the pleasure or sorrow money (or loss of money) could produce. In doing so Cramer showed the expectation would converge to less than \$3 if "one wishes to suppose that the moral value of goods was as the square root of the mathematical quantities".

N. Bernoulli was not entirely satisfied with this solution. In his reply to Cramer, N. Bernoulli wrote that the pleasure difference "does not demonstrate the true reason" for why one should not pay infinity to play the game. Even Cramer himself thought his square-root assumption about money and pleasure was not just.

Eventually in 1738, D. Bernoulli published his solution to the problem (Bernoulli, 1738). D. Bernoulli's solution was similar to Cramer's and based on the concept of utility, which measured the usefulness of values and was taken to be a logarithmic function of values. It was shown that while the expected value diverged the expected utility converged. D. Bernoulli's solution was seminal and extremely influential, and has since shaped the whole field of economics and of the psychology of decision making.

It was interesting to note that N. Bernoulli vigorously objected his cousin's approach. A series of communication showed that the two had engaged in serious arguments. To N. Bernoulli, the concept of utility, similar to the "moral value" of Cramer, was arbitrary and, to a certain extent, irrelevant. Rather, the concern here was to find a more general way to show if a game was fair, regardless of who was playing the game. "For example a game is considered fair, when the two players bet an equal sum on a game under equal conditions, although according to your theory, and by paying attention to their riches, the pleasure or the advantage of gain in the favorable case is not equal to the

sorrow or the disadvantage that one suffers in the contrary case".

N. Bernoulli had his own insights on how to solve the problem. To him, the true reason had to do with those "cases which have a very small probability, [which] must be neglected and counted for nulls, although they can give a very great expectation". Unfortunately, N. Bernoulli encountered great difficulty in deciding when a very small probability should be counted as zero. As a result, his approach was not fully developed at that time and was completely overshadowed by the utility-based solutions.

In this paper we argue that N. Bernoulli might be correct and we provide a new and complete mathematical treatment that is consistent with his insights. Before we dive in, however, we would like to briefly review existing solutions to the St. Petersburg Paradox.

The St. Petersburg Paradox

The key puzzle behind the SPP is how a gamble with an infinite expected return could be valued so little to a human player. Over 300 years many solutions have been proposed, which can roughly be divided to three categories.

The first line of solution attacks the realism of the gamble. Given that the expected gain for one player is infinite, the potential loss for the other player would be infinite as well. Since nobody has infinite payout, the game in its classical form is not realistic and will in no way be offered in the real world. Therefore, if we cap the potential payout in a revised game, the expected value would then converge.

Even D. Bernoulli had thought this issue was important. In 1731 he wrote to his cousin, "I have no more to say to you, if you do not believe that it is necessary to know the sum that the other is in position to pay". The wiki page on SPP (http://en.wikipedia.org/wiki/St._Petersburg_paradox) has a table listing the expected values with respect to a few interesting caps. For example, if the payout is capped with the total US GDP in 2007 (~$13.8 trillion), the expectation is merely $44.57.

The second category of solution focuses on the value component of Eq. 1. The argument is that people are not interested in the monetary value per se, but more in the utility, goodness, or pleasure the money brings forth, which can be represented by a utility function. Therefore, if the value is replaced by a utility function, which is typically assumed to be concave, it can be shown that the expectation would converge. Cramer's and D. Bernoulli's solutions belong to this category.

Utility is capable of capturing the time discounted value of wealth, and can simultaneously incorporate different risk preferences. The utility theoretical approach has since become a dominant solution to the game and has enjoyed a profound influence in the broad field of economics (von Neumann & Morgenstern, 1944). However, criticisms regarding the correct form of utility function remain. Note that both the square-root function and the logarithmic function, as well as many other concave functions, would work (in that the total sum converges). Should one solution be preferred to another? D. Bernoulli tended to believe the distinction was not important. He praised Cramer's solution in his paper, "Indeed I have found his theory so similar to mine that it seems miraculous that we independently reached such close agreement on this sort of subject" (Bernoulli, 1738).

Whereas it appears that it is an empirical issue and can only be answered by psychological investigations, Ole Peters, a physicist, recently proved, via a solid mathematical treatment, that the logarithmic function could be naturally derived based on the mathematics of time average of rate of return and was therefore necessary in the situation (Peters, 2011). Peters argued that D. Bernoulli accidentally chose the correct function form for utility even though he was not aware of its underlying physics. However, it is important to note that while Peter's treatment removes the arbitrariness in the utility functional form, it involves additional assumptions not in the original SPP such as repeated games (Buchanan, 2013).

The third line of solution focuses on the probability weighting component of Eq. 1 and argues that people simply regard those extremely large payout cases having so small a probability that render them impossible to occur. As we mentioned above, N. Bernoulli was first advocate of this idea. In a 1728 letter to Cramer, he wrote,

"with him a very small probability to win a great sum does not counterbalance a very great probability to lose a small sum, he regards the event of the first case as impossible, and the event of the second as certain. It is necessary therefore, in order to settle the equivalent justly, to determine as far as where the quality of a probability must diminish, so that it be able to counted null".

With this reasoning, he demonstrated, for example, that if one regarded probabilities less than 1/32 as null, then the expectation became merely $2. However, he recognized his difficulty and got stuck. He continued to write,

"but here is what which is impossible to determine, any assumption that one makes, one encounters always difficulties; the limits of these small probabilities are not precise".

On Subjective Probability of Rare Events

N. Bernoulli's difficulty has to do with the lack of a theory on how the human mind represents and processes rare events with very small probabilities. While it is certainly unsatisfying to set an arbitrary limit so that any smaller probabilities are treated as zero, it is important to recognize that the way in which the human mind represents probabilities may not be same as what the standard probability theory prescribes.

At human scales, very small probabilities are often linked to impossibility. Emile Borel, an eminent French mathematician, introduced what he called the single law of chance in his 1943 book. The law, which nowadays is simply called Borel's Law, says, "Events with a sufficiently small probability never occur". Borel goes on to clarify

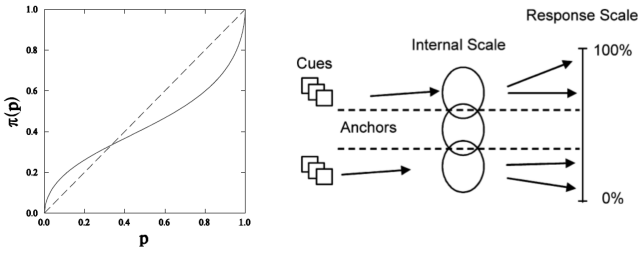
what he meant by "sufficiently small" probabilities, by distinguishing different scales, as shown in Table 1. According to Borel, therefore, probabilities smaller than 1 in a million should be regarded as small enough that for practical purposes a human can treat them as impossible. In the context of SPP, one in a million roughly corresponds to getting 20 tails in a row. By setting smaller probabilities to zero, the expectation for the gamble reduces to about $10.

Table 1: Negligible probabilities according to Borel's law

Scale	Limit	Example
Human	1 in 10^6	Any two chosen seconds in a year are same
Terrestrial	1 in 10^{15}	Any two chosen square foot on Earth are same
Cosmic	1 in 10^{50}	Any two chose atoms on Earth are same
supercosmic	1 in $10^{1,000,000,000}$	Any two particles in the universe are same

That people distort probabilities from their standard values in a systematic way is also a critical claim in prospect theory, a descriptive theory of human decision making (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). According to this theory, utilities should be weighted by the psychological correspondence of probabilities rather than the standard probabilities themselves, and the relationship between the two can be represented by a π function, as shown in Figure 1. Therefore, it seems that people are systematically underestimating large probabilities and overestimating small probabilities.

The π function has been used to explain many interesting phenomena such as the certainty effect. However, when applied in the SPP situation it works in the opposite direction - here small probabilities need to be further reduced rather than enlarged. Nevertheless, the π function is consistent with the general idea that subjective probabilities are not necessarily equal to objective probabilities.



A B

Figure 1: (A) The π function as described in prospect theory (based on Tversky and Kahneman, 1972); (B) Internal representations on a coarser scale have to be transformed to external probabilities on a finer scale, leading to distortion.

Such distortion, despite the difference in specific forms, may result from the limited resolution of the mind in representing uncertainty. We (Sun, Wang, Zhang & Smith, 2008) provided evidence supporting that the mind may adopt a probability scale much coarser than the one prescribed in probability theory. For example, the difference between p=0.0125 and p=0.00625 may be significant in probability theory (and in the standard SPP context), they may all be represented as "quite small", "unlikely", or "impossible" in the mind. This lack of resolution, in addition to other constraints such as anchor and adjustment, contributes to a distorted mental representational scheme of probabilities (see Figure 1B).

Probability and Time

It is desirable to develop a normative mathematical treatment that at the same time describes how the human mind perceives and represents probabilities. Hopefully, such a treatment can naturally lead to a solution to the SPP. Among other advantages, this treatment is consistent with N. Bernoulli's original insights to the problem when he first proposed it, and avoids the inherent arbitrariness of the utility-based approach.

In recent years we have advocated an approach that is based on the inherent connection between probability and time (Sun & Wang, 2010a; 2010b; 2013; Sun et al, 2015). A comprehensive treatment of the connection is still under development. In this paper we provide a concise and relevant narrative with a goal to demonstrate how it can be applied to solve the SPP.

Although the concept of probability is often formally defined using a set theoretical axiom system (e.g. Kolmogorov, 1965), there exists an intriguing relationship between probability and time. In general, for a given event, its probability describes the relative frequency of its occurrence in the long run. In the context of time, the probability of an event corresponds to the mean inter-arrival time (MT) of the event, describing how long it takes for the same event to occur again.

To facilitate, consider the situation of fair coin tosses, where Head (H) or Tail (T) can occur in each toss. We know,

P(T)=1/2, P(TT)=1/4, P(TTT)=1/8, P(TTTT)=1/16.

This is equivalent to say, in terms of the MT:

MT(T)=2, MT(TT)=4, MT(TTT)=8, MT(TTTT)=16.

That is, on average it takes 2 tosses for T to re-appear, 4 tosses for TT (2 Tails in a row) to re-appear, 8 tosses for TTT to re-appear, 16 tosses for TTTT to re-appear. In fact, it can be proved that the standard probability is simply the reciprocal of mean time:

$$p = 1/MT \qquad \text{Eq. 2}$$

It turns out that there is another time statistic that is also associated with uncertain events, waiting time (WT). WT describes how long one has to wait for an event to occur for the first time (rather than to re-occur). Intuitively, one would think a more frequent event (i.e., an event with a short mean time) would occur soon (i.e., with a short waiting time),

however, it is not generally so. The MT and WT can be dissociated, especially for random sequences with length longer than 1. In particular, it can be shown that streak patterns have the longest WT among sequences with equal lengths:

WT(T)=2, WT(TT)=6, WT(TTT)=14, WT(TTTT)=30.

That is, if one starts to toss a fair coin, on average it takes 2, 6, 14, and 30 tosses for T, TT, TTT, and TTTT to occur for the first time, respectively. This is different from these patterns' MTs (2, 4, 8, and 16, respectively). More complete treatments of WT can be found here (Sun & Wang, 2010a; 2010b; 2013).

An intuitive explanation for why the WT for streak patterns is the longest is that when a streak is interrupted, it takes longer to get back to that streak. Therefore, given a random sequence, streaks have larger variances than non-streaks. This statistical fact leads to another time-based statistic, which we call Additional Transition Time (AT).

Formally, a flipped coin can be viewed as a Bernoulli process with the probability of heads p_H (or probability of tails p_T). We can then associate it with a Markov chain, whose states are patterns consisting of outcomes generated by consecutive tosses (Figure 2). We consider the AT, $A_{i,j}$, as the transition time (i.e., the number of tosses it takes in this case) for the Markov chain to first reach pattern j, given the current pattern i.

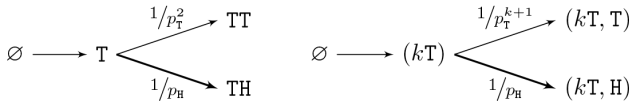


Figure 2: A binomial tree representing the expected AT when an existing streak is either continued or discontinued by a single additional trial. ϕ represents the very beginning of the process (i.e., start anew with a time stamp of zero). The left figure is a special case of the right figure with $k = 1$. For a fair coin $p_H = p_T = 1/2$, given the current state T, the process has the same probability to branch into either TT (continuation) or TH (discontinuation). However, from time's perspective, TH is more immediate (2 tosses away) than TT (4 tosses away). As the length of the initial streak k increases, the AT remains the same for a discontinuation of the streak ($1/p_H$), but grows exponentially for a continuation of the streak ($1/p_T^{k+1}$).

Figure 2 illustrates a quite striking result. Given that we have observed TTT, for example, how much additional transition time it takes to reach TTTT? Standard probability theory tells us there is an equal probability (1/2) of H or T in the fourth toss. However, on average, it would take 1/(1/2)=2 additional tosses to get TTTH, and $1/(1/2^4)$=16 additional tosses to get TTTT – *TTTT has a much longer AT*. Again, a simple explanation is that in waiting for the streak, if it goes awry, the wait would have to re-start from scratch. This is in contrast to TTTH, where a wrong outcome (i.e., expecting a final H but getting T) will not hurt much as the waiting (of a final H) can continue. Taken together, this is just another mathematical fact that justifies why streaks are rare and remarkable and why (kT,H) is more imminent than (kT,T).

It can be shown that the AT is just another manifestation of the waiting time. Both statistics are affected by the same start-anew effect due to the self-overlapping property of the streak pattern. Instead of treating each pattern as a whole (as in calculating the waiting time), the AT allows temporal prediction by breaking the waiting time statistics into two parts, as follows,

$$WT(kT,H) = WT(kT) + MT(H) = WT(kT) + 1/p_H$$
$$WT(kT,T) = WT(kT) + MT(kT,T) = WT(kT) + 1/p_T^{k+1} \quad \text{Eq. 3}$$

Thus, given a streak of k tails, the expected AT for the streak to be extended by one more tail is not the mean time of a single tail (MT(T)), but the mean time of k+1 tails (MT(kT,T)).

While it appears counter-intuitive, WT and AT capture an essential environmental statistic describing when an event is to occur. They are certainly relevant to human cognition. In many everyday situations, it is likely that the question of when an event is to occur is more important than the question of how often an event is to occur. Therefore, it is plausible that the brain and the mind have developed mechanisms to be sensitive to WT and AT statistics. We have previously argued that human perception of randomness in general and the gambler's fallacy in particular might be linked to the longer WT of streak patterns (Sun & Wang, 2010a; 2010b; 2013). More recently, we have shown how the brain could learn to capture the WT statistic through predictive neural learning (Sun et al., 2015).

Toward an AT Based Solution to SPP

It is therefore quite plausible that the AT captures, both normatively and descriptively, people's sensitivity to the rarity of streak patterns. In this account, TTTT is different from patterns such as TTTH not only in terms of the additional transition time it requires to complete the pattern, but also the fact that the difference increases quickly as the pattern length increases.

The SPP inherently involves streak patterns. For the player to win big, he/she would wish to delay the first occurrence of H as much as possible, that is, to get the streak of Ts as long as possible. According to N. Bernoulli, it is these long streaks that should be rendered as impossible to ever occur due to their very small probabilities. Unfortunately, the standard mean time based probability theory does not distinguish between streak and non-streak patterns and treats them equally likely – both as a function of pattern lengths. We argue that that the concept of AT offers a new perspective to resolve N. Bernoulli's difficulty and can lead to a more justified solution to the SPP.

More specifically, we have shown that having already observed TTT, while the probability of getting the fourth T

is ½, the AT for getting TTTT is 16. Given the genuine relationship between probability and time, it is possible to derive another probability measure based on the AT as follows:

$$p' = 1/AT \qquad \text{Eq. 4}$$

p', therefore, measures a type of uncertainty associated with obtaining the final outcome required to complete the entire sequence. Different from *p* (Eq. 2), *p'* is pattern structure dependent – different prefixes result in different *p'*. In essence, it is the local structures of patterns that differentiate patterns and make streaks special. Mean-time based *p* is blind to the structures and AT-based *p'* highlights the difference.

We have shown that the AT it takes for a streak of length (*k*-1) to be extended to length *k* is $1/p^k$. Thus, we can derive the following relationship:

$$p' = p^k \qquad \text{Eq. 5}$$

In the example above, to get the fourth T given TTT, we have p=1/2, p'=1/16.

We can then replace the context-free *p* in Eq. 1 with the context-sensitive *p'* to acquire another expectation measure as follows:

$$S' = \$\sum_{n=1}^{\infty}(\prod_{k=1}^{n} p'_k)2^{n-1} = \$\sum_{n=1}^{\infty}(\prod_{k=1}^{n}(1/2)^k)2^{n-1} \qquad \text{Eq. 6}$$

Different from S, which diverges, it can be proved that S' converges.

Figure 3 depicts S' converging behavior as the number of trials increases. For comparison, we also plot the logarithmic utility solution by D. Bernoulli. It is clear that S' converges very fast, to the asymptotic value as early as toss 5. In a sense this solution fits N. Bernoulli's original insights almost perfectly. As mentioned above, he suggested that a sensible man should treat those cases with probability less than 1/32 as impossible, which corresponds to streaks of TTTTT and longer. However, our solution avoids N. Bernoulli's difficulty and makes an otherwise arbitrary choice justifiable. According to this solution, long streak patterns occur with probabilities that decrease so fast that they over-compensate the rather large payouts in those situations. The net expectation converges, supporting the perceived limited value of the gamble.

Another feature of the new AT-based solution is that it does not need a concave utility function to resolve the puzzle. N. Bernoulli criticized that focusing on utility rather than value was not the true reason for why people valued the gamble less. Although the logarithmic utility function has received much support in the past as a reasonable way describing how people conceive goodness of value, there is a lack of a priori justification for the choice.

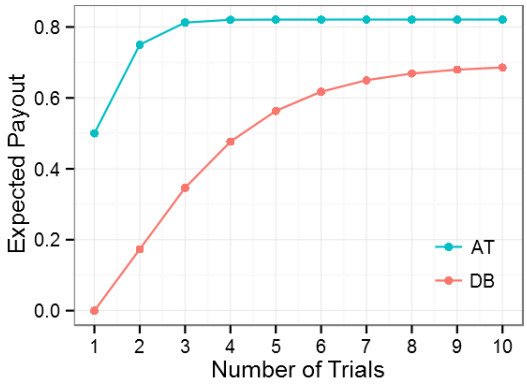


Figure 3. Converging expected payout in the SPP. **DB**: the logarithmic utility solution by Daniel Bernoulli. **AT**: new AT-based solution (Eq. 6).

Discussion

The St. Petersburg Paradox is an over 300-year old puzzle and still resides in the center of any formal understanding of human decision making. Why is a gamble with an infinite expected return valued so little? The classical solution, proposed by Daniel Bernoulli, resorts to the concept of utility. According to this account, humans do not value money at its face value, rather, its utility, measuring its usefulness or the pleasure it brings about, should be considered. Utility is apparently sensitive to, and therefore is capable of capturing the effect of, a range of factors, including individual difference, risk preferences, and time discount. The concept is so intuitively appealing and mathematically powerful that it has since become a cornerstone of modern economics. However, the lack of a priori analysis for choosing a specific form of utility function raises a problem. More recently, Ole Peters demonstrated that the logarithmic utility function is a mathematically necessary result if a probabilistic decision maker is assumed to maximize return over time.

Nicolas Bernoulli, the original proposer of the puzzle, hypothesized that the culprit was those cases that carry large payouts but with very small probabilities. He suspected that to a sensible human those cases should be rendered as impossible to occur. Equipped with standard probability theory, however, he encountered great difficulty in deciding when a probability was small enough.

In this paper we have suggested a different solution to the problem. The solution is consistent with N. Bernoulli's probability weighting idea, but avoids its difficulty. Essentially, we have derived a different probability measure based on the waiting time, an environmental statistic that describes how soon an uncertain event is to occur. It can be shown that in tossing a fair coin, while patterns of equal length all have the same mean time, that is, the same probability, they may have different waiting times. We show that via the concept of Additional Transition Time, a different probability measure can be derived for patterns. In

particular, we show that streak patterns have a much longer AT than non-streak patterns, rendering them to be very rare to occur. We demonstrate that this probability weighting leads to a converged expected return and therefore solves the SPP.

One advantage of our solution is that it eliminates arbitrary ad-hoc choices, for either the utility function form or the limit for sufficiently small probabilities. It is a normative solution based on mathematics, with fewer pre-assumptions. The brain's sensitivity to the waiting time has recently been demonstrated, which lends further credibility to the solution.

In sum, the treatment presented in this paper is in contrast to almost all existing theories attempting to explain and rationalize human biases in judgment and decision making (Falk & Konold, 1997; Gigerenzer & Hoffrage, 1995; Gilovich, Griffin & Kahneman, 2002; Griffiths, Chater, Kemp, Perfors & Tenenbaum, 2010; Ma, Beck, Latham & Pouget, 2006). In spite of different details, these theories are all based on the assumption that human mind encodes mean-time based probabilities or likelihoods. Here we argue that an accurate encoding of more complicated temporal structures (specifically, the wait time and additional transition time statistics) is at the core of how people represent uncertainty.

Acknowledgments

This work was partially supported by the Air Force Office of Scientific Research (AFOSR) grant number FA9550-12-1-0457, Office of Naval Research (ONR) grant number N00014-16-1-2111.

References

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis [exposition of a new theory on the measurement of risk, translated by louise sommer, 1954, published on econometrica, 22, 23-36]. *Commentaries of the Imperial Academy of Science of Saint Petersburg.*

Borel, E. (1962). *Probabilities and life [translated from probabilités et la vie, 1943, by maurice baudin]*. New York, Dover Publications.

Buchanan, M. (2013). Gamble with time. *Nature Physics*, *9*(1), 3-3.

Falk, R. & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301-318.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684-704.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of inuitive judgment.* New York: Cambridge University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences.*

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transimission*, *1*, 1-7.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432-8.

de Montmort, P. R. (1713). *Essay d'analyse sur les jeux de hazard [essays on the analysis of games of chance] (reprinted in 2006) (in french) (second ed.).* Providence, Rhode Island: American Mathematical Society.

Peters, O. (2011). The time resolution of the st petersburg paradox. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *369*(1956), 4913-31.

Sun, Y. & Wang, H. (2010a). Gambler's fallacy, hot hand belief, and the time of patterns. *Judgment and Decision Making*, *5*(2), 124-132.

Sun, Y. & Wang, H. (2010b). Perception of randomness: On the time of streaks. *Cognitive Psychology*, *61*, 333-342.

Sun, Y. & Wang, H. (2013). Finding structure in space and time: Active trace of patterns. In *Proceedings of the 35th annual conference of the cognitive science society.* Cognitive Science Society.

Sun, Y., O'Reilly, R. C., Bhattacharyya, R., Smith, J. W., Liu, X., & Wang, H. (2015). Latent structure in random sequences drives neural learning toward a rational bias. *Proceedings of the National Academy of Sciences of the United States of America, 112(12),* 3788-3792.

Sun, Y., Wang, H., Zhang, J., & Smith, J. W. (2008). Probabilistic judgment on a coarser scale. *Cognitive Systems Research*, *9*(3), 161-172.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk & Uncertainty*, *5*, 297-323.

von Neumann, J. & Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton, NJ: Princeton University Press.

A test of two models of probability judgment: quantum versus noisy probability

Fintan J. Costello (fintan.costello@ucd.ie)
School of Computer Science,
University College Dublin, Dublin 4, Ireland

Paul Watts (watts@thphys.nuim.ie)
Department of Mathematical Physics,
National University of Ireland Maynooth, Maynooth, Ireland

Abstract

We test contrasting predictions of two recent models of probability judgment: the quantum probability model (Busemeyer et al., 2011) and the probability theory plus noise model (Costello and Watts, 2014). Both models assume that people estimate probability using formal processes that follow or subsume standard probability theory. The quantum probability model predicts people's estimates should agree with one set of probability theory identities, while the probability theory plus noise model predicts a specific pattern of violation of those identities. Experimental results show just the form of violation predicted by the probability theory plus noise model. These results suggest that people's probability judgments do not follow quantum probability: instead, they follow the rules of standard probability theory, with the systematic biases seen in those judgments due to the effects of random noise.

Introduction

Researchers over the last 50 years have identified a large number of systematic biases in people's judgments of probability. These biases are typically taken as evidence that people do not follow the normative rules of probability theory when estimating probabilities, but instead use a series of heuristics (mental shortcuts or 'rules of thumb') that sometimes yield reasonable judgments but sometimes lead to severe and systematic errors, causing the observed biases (Kahneman and Tversky, 1973). This 'heuristics and biases' view has had a major impact in psychology (Kahneman and Tversky, 1982, Gigerenzer and Gaissmaier, 2011), economics (Camerer et al., 2003), law (Sunstein, 2000), medicine (Eva and Norman, 2005) and other fields, and has influenced government policy in a number of countries (Oliver, 2013, Vallgårda, 2012).

The existence of these systematic biases in people's probabilistic reasoning is incontrovertible. The conclusion that these biases necessarily demonstrate heuristic reasoning processes is, however, less sure. Recent research has shown that many of these biases can be explained if we assume that people estimate probability using formal processes that follow or subsume standard probability theory. Two such formal models are the quantum probability model proposed by Busemeyer and colleagues (Busemeyer et al., 2011, Busemeyer and Bruza, 2012), and our own probability theory plus noise model (Costello and Watts, 2014, 2016a,b). Both models can account for a number of well-known biases seen in people's probabilistic reasoning. Both models predict systematic bias away from standard probability theory for a range of probalistic expressions: importantly, the models make contrasting predictions about the occurence and direction of these biases. In this paper we test these contrasting predictions.

The probability theory plus noise model

In standard probability theory the probability of some event A is estimated by drawing a random sample of events, counting the number of those events that are instances of A, and dividing by the sample size. The expected value of these estimates is $P(A)$, the probability of A; individual estimates will vary with an approximately normal distribution around this expected value. The probability theory plus noise model assumes that people estimate the probability of some event A in exactly the same way: by randomly sampling items from their memory, counting the number that are instances of A, and dividing by the sample size. If this counting process was error-free then people's estimates would have an expected value of $P(A)$ just as in probability theory. Human memory is subject to various forms of random error, however. To reflect this the model assumes a minimal form of random error such that items have some probability $d < 0.5$ of being counted incorrectly (we assume that this error term is constant within a given participant, but varies across participants: some participants are more prone to random error than others).

Because of this random error, there is a chance d that an item which is not truly an instance of A will be incorrectly counted as A, and the same chance d that an instance of A will be incorrectly counted as not A. Given this error, a randomly sampled item can be counted as A in two mutually exclusive ways: either the item truly is an instance of A and is counted correctly (this occurs with probability $P(A)(1-d)$, since $P(A)$ items are truly instances of A, and items have a $(1-d)$ chance of being read correctly), or else the item truly is not an instance of A and is counted incorrectly as A (this occurs with probability $(1-P(A))d$, since $(1-P(A))$ items are truly not instances of A, and items have a d chance of being read incorrectly). The expected value for a noisy estimate for the probability of A is thus

$$\langle P_E(A)\rangle = P(A)(1-d) + (1-P(A))d = (1-2d)P(A) + d \quad (1)$$

and we expect individual estimates $P_E(A)$ to vary independently around this expected value.

According to the model, the pattern of bias due to random noise seen in Equation 1 is predicted for all types of events, but will be more pronounced for complex events (such as

conjunctions or disjunctions), which, because of their complexity, give more opportunity for error compared to simple events. For simplicity and clarity we ignore this effect here and assume the same rate of error for all events, so the expected value for conjunctive events $A \wedge B$ and disjunctive event $A \vee B$ are

$$\langle P_E(A \wedge B) \rangle = (1-2d)P(A \wedge B) + d \quad (2)$$

$$\langle P_E(A \vee B) \rangle = (1-2d)P(A \vee B) + d \quad (3)$$

We've previously shown how this model explains various biases frequently seen in people's probabilistic reasoning, such as underconfidence (people's tendency to overestimate the probability of low-probability events and understimate the probability of high probability events), subadditivity (people's tendency to give probabilities for the constituent events of a disjunction which, when added, are reliably higher than their probability for the disjunction as a whole) and the conjunction fallacy. The patterns of occurrence of these biases in experimental results match the model's predictions well (Costello and Watts, 2014, 2016a,b).

This model also predicts a range of other deviations from probability theory for people's probability estimates. Table 1 gives a series of probability theory identities: expressions which, in probability theory, are required to have a value of 0 for all events A and B. We can use our expected-value expression to obtain values for these identities as predicted by our model. For identity 5, for example, our model predicts an average value of

$$\begin{aligned}\langle P_E(A) \rangle &+ \langle P_E(\neg A \wedge B) \rangle - \langle P_E(A \vee B) \rangle \\ &= (1-2d)P(A) + d + (1-2d)P(\neg A \wedge B) + d \\ &\qquad - (1-2d)P(A \vee B) - d \\ &= d\end{aligned}$$

(with values varying randomly around that average). Since d (the rate of random error) is always positive in this model, the model predicts a positive value for this identity for all events A and B: in other words, it predicts that, if we substitute people's probability estimates for A, $\neg A \wedge B$, $A \vee B$ and so on, into this identity, the values obtained will, on average, be positive for all events. Similar predictions hold for the other identities in table 1. As we see below, the quantum probability model makes a different prediction.

The quantum probability model

The quantum probability model assumes that people's probabilistic reasoning follows the mathematical principles of quantum probability. The primary theoretical distinction between quantum and standard probability lies in the idea of 'compatible' or 'incompatible' observables. An observable defines the set of all possible distinct outcomes for a given measurement. For example, if we are checking to see whether some event A has occurred or not, we are, in the terminology of quantum theory, measuring an observable **A**, which returns one of two distinct outcomes: A (the event has occurred) and $\neg A$ (the event has not occurred).[1] Two observables **A** and **B** are compatible, in quantum theory, if the outcome of a joint observation (such as $P(A \wedge B)$) does not depend on the order in which **A** and **B** were measured. Two observables are incompatible if the outcome of such a joint observation does depend on the ordering of measurement of **A** and **B**. In other words, if two observables are compatible then the equality $P(A \wedge B) = P(B \wedge A)$ must hold, while if the observables are incompatible, equality need not hold. This distinction is fundamental both in standard quantum theory and in the quantum probability model (see Busemeyer et al., 2011, p. 199). We use this difference between $P(A \wedge B)$ and $P(B \wedge A)$ as a measure of compatibility in our experiment, with the idea that the greater the difference between $P(A \wedge B)$ and $P(B \wedge A)$, the more confident we are that the observables in question are incompatible.

If two observables are compatible, then quantum probability expressions for all possible outcomes of those observables (that is, for $P(A)$, $P(\neg A)$, $P(B)$, $P(\neg B)$, $P(A \wedge B)$, $P(A \wedge \neg B)$, $P(A \vee B)$, and so on) are exactly equivalent to the standard probability theory expressions for those outcomes. In other words, if two observables are compatible then all the probability theory identities given in Table 1 should have the value of 0, as required in standard probability theory.

If two observables are incompatible, in quantum theory those observables cannot both be measured simultaneously: instead they must be measured separately, one after the other. If two incompatible observables are measured in the order **A** then **B**, then quantum probability expressions for outcomes of the second observable can deviate from the requirements of probability theory, giving, for example

$$P(B) = P(\neg A \wedge B) + P(A \wedge B) + \delta_B \quad (4)$$

where $P(B)$ is the probability obtained when **B** is measured with no prior measurement of **A**, $P(\neg A \wedge B)$ and $P(A \wedge B)$ are the probabilities obtained when **A** and **B** are measured sequentially, and where $\delta_{\mathbf{B}}$ is a 'quantum interference' term for observable B. This quantum interference term arises because, contrary to the 'macroscopic realism' view of the world and thus to the assumptions of standard probability theory, in quantum theory if **A** is not measured, it is not necessarily in either state A or state $\neg A$: it may be in some 'superposition' of states. This means that the two probabilities $P(\neg A \wedge B)$ and $P(A \wedge B)$ do not necessarily cover all possible cases arising when estimating $P(B)$ with no prior measurement of **A**, and so $P(B) = P(\neg A \wedge B) + P(A \wedge B)$ does not necessarily hold. Note that quantum interference is not an error term here: for a given observable **B** (and a given participant, in

[1] In quantum theory, each of these outcomes A and $\neg A$ would be referred to as an *eigenvalue* of the observable, because the observable defines the set of orthonormal vectors of unit length (*eigenvectors*) in a multidimensional state space. We don't need to use this detailed view of quantum theory in our discussion here, and so we avoid this more complex terminology.

Table 1: Predicted values of the noise model and the quantum model for a series of probability theory identities. Standard probability theory requires these identities to have a value of 0. The probability theory plus noise model predicts that the average value of each these identities will be positive for all events, deviating from 0 by more than d (the chance of random error). The quantum probability model makes different predictions for three mutually-exclusive situations: when observables **A** and **B** are compatible; when observables are incompatible and measured in the order **A** then **B**, or when they are incompatible and measured in the order **B** then **A**.

label	identity	noise model	quantum model		
			compatible	incompatible **A** then **B**	incompatible **B** then **A**
1	$P(A)+P(\neg A\wedge B)-P(A\vee B)$	d	0	0	δ_A
2	$P(B)+P(A\wedge\neg B)-P(A\vee B)$	d	0	δ_B	0
3	$P(A\wedge B)+P(A\wedge\neg B)-P(A)$	d	0	0	$-\delta_A$
4	$P(A\wedge B)+P(\neg A\wedge B)-P(B)$	d	0	$-\delta_B$	0
5	$P(A\wedge B)+P(\neg A\wedge B)+P(A\wedge\neg B)-P(A\vee B)$	$2d$	0	0	0

Busemeyer et al.'s model) this quantum interference term δ_B has a fixed value that specifies the relationship between $P(B)$ and $P(\neg A\wedge B)+P(A\wedge B)$ for that observable. The quantum interference term δ_B can take on different values for different observables **B** (and different participants): in some cases positive, in some negative, and in some cases 0.

If two incompatible observables are measured in the order **A** then **B** then $P(A)$ has no such interference term, and so remain exactly equivalent to the corresponding standard probability theory expression, giving, for example

$$P(A)=P(A\wedge\neg B)+P(A\wedge B)$$

This is because measurement of **A** causes the observable to collapse out of superposition and take on either state A or state $\neg A$: the two probabilities $P(A\wedge\neg B)$ and $P(A\wedge B)$ arising from subsequent measurement of **B** do cover all possible cases arising after the outcome A, and so their sum equals $P(A)$. If incompatible observables are measured in the opposite order **B** then **A**, there is a parallel interference δ_A for observable **A**, and no interference term for observable **B**. Note that deviations from probability theory in the quantum model arise from single event probabilities ($P(A)$ or $P(B)$) because only these single-event probabilities have associated quantum interference terms: there are no such interference terms associated with combined probabilities such as $P(A\wedge B)$.

How does the quantum probability model impose an ordering on incompatible events? The model assumes that incompatible observables **A** and **B** are ordered in terms of causal links between those observables. If observables **A** and **B** are causally linked such that the occurrence of event A in some way influences the subsequent occurrence or non-occurrence of B, Busemeyer et al. (2011) assume that observable **A** is measured first and **B** is measured second. To quote Busemeyer et al. (2011, page 199)

> *when asked to judge the likelihood that "cigarette tax will increase and a decrease in teenage smoking will occur," it is natural to assume that the causal event "increase in cigarette tax" is processed first.*

This approximately matches the approach taken in standard quantum theory, where the observable that occurs first is, naturally enough, the first measured observable. In this case the quantum probability model would allow deviations from probability for the second observable (the caused event 'a decrease in teenage smoking') but not for the first observable (the causing event 'an increase in cigarette tax'). If **A** and **B** have no causal link, Busemeyer et al. (2011) take a different approach and assume that the most probable of the two events is measured first: that is, if $P(A)>P(B)$ then the conjunctive probability $P(A\wedge B)$ is measured, but if $P(B)>P(A)$ then the conjunctive probability $P(B\wedge A)$ is measured. This represents a significant deviation from quantum theory because it connects ordering not to observables but to outcomes of observations. This also introduces other problems to the model. One problem concerns the ordering of the 'decision' measurements of $P(A)$ and $P(B)$. To decide whether the conjunction is measured as $P(A\wedge B)$ or $P(B\wedge A)$, we must first measure $P(A)$ and $P(B)$ to see which is more probable. If **A** and **B** are incompatible, we will get different results if we measure $P(A)$ first and then $P(B)$, or $P(B)$ first and then $P(A)$. We cannot decide which ordering to use by assuming the most probable event is measured first: before measurement, we don't know which event is most probable. It is not clear how this issue is resolved in the quantum probability model; given this, we don't address this 'ordering by probability' approach here.

Predictions: probability theory identities

All of the identities in Table 1 have a value of 0 in standard probability theory. In the quantum model, probability judgments necessarily agree with standard probability theory when the observables **A** and **B** are compatible; and so in this

model, all of these identities should have a value of 0 for compatible observables and will only deviate from 0 for incompatible observables.

If observables are incompatible, then the values of these identities depend on the ordering of observables **A** and **B** and on whether the identity contains $P(A)$ or $P(B)$. If an identity contains a probability expression for the first measured observable, there is no interference term and the identity has a value of 0, as in standard probability theory. If an identity contains the second observable, however, there is an interference term for that observable, and the identity's value will be related to the value of that term. Consider, for example, identity 2 in Table 1. This identity contains $P(B)$. If observables are incompatible and **B** is measured first, then there is no interference term for **B** and identity 2 has a value of 0. If observables are incompatible and **A** is measured first, however, there is an interference term for **B** and identity 2 has the value

$$\begin{aligned}
&P(B)+P(A\wedge\neg B)-P(A\vee B)\\
&=[P(\neg A\wedge B)+P(A\wedge B)+\delta_B]+P(A\wedge\neg B)-P(A\vee B)\\
&=\delta_B+P(\neg A\wedge B)+P(A\wedge B)+P(A\wedge\neg B)-P(A\vee B)\\
&=\delta_B
\end{aligned}$$

(from Equation 4): this identity is predicted to have a value equal to the interference term δ_B.

Next consider identity 4 in Table 1. This identity again contains $P(B)$. If observables are incompatible and **B** is measured first, then there is no interference term for **B** and identity 4 has a value of 0. If observables are incompatible and **A** is measured first, however, there is an interference term for **B** and identity 4 has the value

$$\begin{aligned}
&P(A\wedge B)+P(\neg A\wedge B)-P(B)\\
&=P(A\wedge B)+P(\neg A\wedge B)-[P(A\wedge B)+P(\neg A\wedge B)+\delta_B]\\
&=-\delta_B
\end{aligned}$$

(from Equation 4): this identity is predicted to have a value equal to the negative of the interference term, or in other words, equal to that of identity 2 but with the opposite sign.

Parallel predictions hold for identities 1 and 3. These identities are expected to have values of 0 if **A** is measured first, while if **B** is measured first these identities have values equal to the interference term δ_A but with opposite signs. Note that, since identities 2 and 4 can only have values δ_B and $-\delta_B$ when **A** is measured first, while identities 1 and 3 can only have values δ_A and $-\delta_A$ when **B** is measured first, these two cases are mutually exclusive. This means that, if identities 2 and 4 have values significantly different from 0 for a given pair of events A and B, then the quantum probability model requires that identities 1 and 3 have a value equal to 0 (no interference term) for those events, and vice versa.

Finally, consider identity 5. This identity does not contain an expression $P(A)$ or $P(B)$, and so does not contain a quantum interference term. The quantum probability model thus predicts that, for any fixed ordering of observables **A** and **B** , this two identity will always a value of 0 irrespective of whether A and B are compatible or incompatible.

We can summarise the quantum model's predictions for the identities in Table 1 as follows. For a given pair of events A and B, there are three possible situations: First, **A** and **B** are compatible, in which case the quantum model predicts a value of 0 for all identities. Second, **A** and **B** are incompatible and measured in the order **A** then **B**, in which case the quantum model predicts a value of 0 for all identities but 2 and 4: these two identities are predicted to have opposite signs (one positive, one negative). Third, **A** and **B** are incompatible and measured in the order **B** then **A**, in which case the quantum model predicts a value of 0 for all identities but 1 and 3: these two identities are predicted to have opposite signs (one positive, one negative). The probability theory plus noise model, by contrast, predicts that every one of these identities will deviate from 0, and all will have positive values.

Experiment: incompatibility and probability theory identities

In this experiment we assess the role of incompatibility in the values of the probability theory identities in Table 1. Recall that the probability theory plus noise model predicts that values for these identities should be reliably positive for all pairs of events A and B. The quantum probability model, by contrast, predicts that for any pair of events, either all identities are 0, all are 0 but identities 2 and 4 (and these have opposite signs), or all are 0 but identities 1 and 3 (and these have opposite signs).

We tested these predictions using data from an experiment on conjunction and disjunction fallacies (Experiment 2 in Costello and Watts, 2014). This experiment gathered 68 participants' estimates for $P(A)$, $P(B)$, $P(A\wedge B)$, $P(A\vee B)$, $P(A\wedge\neg B)$ and $P(\neg A\wedge B)$ for 9 different pairs A,B of weather events (see Table 2). These pairs were selected so that they contained events of high, medium and low probabilities. Conjunctive and disjunctive weather events were formed by placing 'and'/'or' between the elements of each pair as required, generating weather events such as 'windy and cold', 'windy or cold', and so on. One group of participants ($N=34$) were asked questions in terms of probability, of the form

- What is the probability that the weather will be W on a randomly-selected day in Ireland?

for some weather event W. The second group ($N=34$) were asked questions in terms of frequency, of the form

- Imagine a set of 100 different days, selected at random. On how many of those 100 days do you think the weather in Ireland would be W?

where the weather events were as before. These two question forms were used because of a range of previous work showing that frequency questions can reduce fallacies in people's probability judgments; the aim was to check whether

Table 2: This table shows the degree of incompatibility of constituent events A and B in the Experiment, measured as the absolute difference between estimates for a conjunction in the ordering $A \wedge B$ and estimates in the ordering $B \wedge A$. Events are are listed in order of increasing incompatibility. This table also shows the average values for identities from Table 1, computed from individual participant's estimates for each event pair. The quantum model predicts that these identities should have a value of 0 when events A and B are compatible; when events are incompatible the quantum model predicts that most identities will have a value of zero, but some pairs of identities will have opposite signs (one positive, one negative) . The probability theory plus noise model predicts that all of these will have positive values, in all pairs of events.

A	B	$\lvert P_E(A \wedge B) - P_E(B \wedge A)\rvert$	Identity				
			1	2	3	4	5
windy	cold	0.005	0.56 (0.31)	0.25 (0.24)	0.26 (0.27)	0.29 (0.29)	0.55 (0.4)
icy	windy	0.007	0.17 (0.28)	0.21 (0.24)	0.2 (0.21)	0.16 (0.27)	0.38 (0.38)
rainy	windy	0.03	0.34 (0.32)	0.27 (0.31)	0.33 (0.27)	0.39 (0.35)	0.67 (0.47)
sunny	icy	0.03	0.11 (0.27)	0.16 (0.27)	0.27 (0.29)	0.22 (0.27)	0.39 (0.4)
cold	sunny	0.06	0.24 (0.28)	0.14 (0.3)	0.21 (0.32)	0.31 (0.29)	0.47 (0.44)
sunny	rainy	0.09	0.17 (0.28)	0.13 (0.25)	0.23 (0.24)	0.27 (0.3)	0.4 (0.38)
cold	cloudy	0.09	0.34 (0.27)	0.29 (0.27)	0.23 (0.3)	0.28 (0.28)	0.58 (0.41)
cloudy	rainy	0.09	0.13 (0.26)	0.35 (0.29)	0.43 (0.28)	0.2 (0.26)	0.57 (0.35)
cloudy	icy	0.11*	0.22 (0.3)	0.21 (0.29)	0.22 (0.27)	0.24 (0.27)	0.46 (0.46)

*Marginally significant difference between orderings in an unpaired t-test ($p = 0.06$). All other differences in ordering were unsignificant; all other differences gave evidence in favour of the null hypothesis (no difference between orderings) in a JZS Bayes Factor test for two-sample designs. All values for the identities were positive and significantly different from 0 in a one-sample t-test ($p < .0001$).

this question form could eliminate fallacy responses for everyday repeated events.

Half of participants in each group saw all conjunctions and disjunctions in the ordering $A \wedge B$, $A \vee B$, $A \wedge \neg B$ and $\neg A \wedge B$, the other half saw them in the reverse ordering $B \wedge A$ $B \vee A$, $B \wedge \neg A$ and $\neg B \wedge A$. This randomisation of the order of presentation of events in conjunctions and disjunctions was carried out purely to get a unbiased sample of probability estimates for these conjunctions and disjunctions, unaffected by any systematic influence of word order. Here, however, we use this ordering to estimate the degree of incompatibility of the twelve pairs of events used in the experiment, by considering the degree of difference between estimates given for $P(A \wedge B)$ and estimates for $P(B \wedge A)$ for each pair of events. The larger the difference between these standard and reverse-order estimates, the more incompatible the given pair of events.

Participants were given questions containing all single events and all conjunctive and disjunctive events, with questions presented in random order on a web browser. Responses were on an integer scale from 0 to 100 and were divided by 100 prior to analysis, and so probability estimates where given in units of 0.01.

Results

Two participants from the 'probability format' group were excluded (one because they gave responses of 100 to all but 4 questions and the other because they gave responses of 0 to all but 2 questions), leaving 66 participants in total. There was little difference in responses between the 'frequency format' and 'probability format' forms of question, so for simplicity we collapse the groups together in our analysis.

Each participant gave probability estimates for 42 distinct items (4 estimates $P(A \wedge B)$, $P(A \vee B)$, $P(A \wedge \neg B)$ and $P(\neg A \wedge B)$ for each of the 9 A, B pairs, and 6 estimates for the various constituents $P(A)$ and $P(B)$). As a consistency check we split the participants in into two random groups and calculated the average probability estimate in each group for each one of the 42 presented items. If participants were responding consistently we would expect there to be a reliable correlation between these average estimates across the two groups. The correlation was very high ($r = 0.97$, $p < 0.0001$), indicating consistent responses and showing that participants were not simply responding randomly in the probability estimation task.

For each participant we calculated the values of various identities from Table 1 for each of the nine pairs A, B. We also measured the degree of incompatibility between these event pairs in terms of the absolute difference between the average estimate for a conjunction presented in the order $A \wedge B$ and a conjunction in the order $B \wedge A$: the higher this difference, the more incompatible the events are.

Table 2 shows the average values obtained for the relevant

identities for each of the 9 event pairs in the experiment, and also shows the degree of measured incompatibility of those event pairs, in increasing order. The degree of incompatibility was low for almost all pairs. There was no statistically significant difference between conjunctive estimates for the ordering $A \wedge B$ and ordering $B \wedge A$ for any pair: only one pair approached significance. A JZS Bayes Factor test for two-sample designs gave evidence in favour of the null hypothesis (no difference between orderings) for all but one pair. There was no relationship between incompatibility and values of the probability theory identities: identity values were reliably positive even when the difference between $P(A \wedge B)$ and ordering $P(B \wedge A)$ was less than 0.01 (that is, less than 1 point on the 100 point rating scale used in the experiment). This is contrary to the predictions of the quantum model, which would predict values of 0 for compatible pairs.

Values for all identities were reliably positive, as predicted by the probability theory plus noise model. For each identity and each event pair, we carried out a single sample t-test comparing individual values for that identity against 0. Since there are 9 event pairs, and 6 identities, this gives 45 separate t-tests. All these t-tests were significant at at least the $p < 0.001$ level, supporting the probability theory plus noise model and going against the quantum model, which would expect either all or most of these values to be 0, and would expect some to be negative.

Finally, recall that the probability theory plus noise model predicts an average value of d for identities 1 to 4, and a value of $2d$ for identity 5. The results in Table 2 support this prediction: the average value for identity 5 was 2.02 times the average value for identities 1 to 4.

Conclusions

The results seen in this experiment appear to clearly contradict the quantum probability model's predictions about compatibility and values of probability theory identities. That model predicts that probability theory identities should have a value of 0 when events are compatible. That prediction did not hold. Even assuming the events are incompatible, the quantum model predicts a value of 0 for all identities but 2 and 4 (and these two identities are predicted to have opposite signs) or for all identities but 1 and 3 (and these two identities are predicted to have opposite signs). These predictions also did not hold: instead, values for all identities were reliably positive, for all events. The results support the probability theory plus noise account, which predicts reliably positive values for all the probability theory identities in Table 1.

The fundamental idea in the probability theory plus noise account is that people's process for estimating probabilities follows the requirements of probability theory, and that the systematic biases away from probability theory seen in people's judgments are simply the consequence of random error in that process. In other work we've shown that this model can explain biases such as conservatism, subadditivity, and binary complementarity. We've also shown that, for expressions in which this model predicts bias should be cancelled, people's probability estimates agree closely with the requirements of probability theory just as predicted by the model (Costello and Watts, 2014). Here we've shown further experimental evidence that supports this model and goes against a competing formal model based on quantum probability. Taken together, our results give evidence against the popular idea that people estimate probabilities using heuristics that do not follow the normative requirements of probability theory (Ariely, 2009, Gigerenzer and Gaissmaier, 2011, Kahneman, 2011, Shafir and Leboeuf, 2002).

References

Ariely, D. (2009). *Predictably irrational: the hidden forces that shape our decisions*. HarperCollins.

Busemeyer, J. R. and Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge University Press.

Busemeyer, J. R., Pothos, E. M., Franco, R., and Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, 118(2):193.

Camerer, C., Loewenstein, G., and Rabin, M. (2003). *Advances in Behavioral Economics*. Princeton University Press.

Costello, F. and Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3):463–480.

Costello, F. and Watts, P. (2016a). Explaining high conjunction fallacy rates: the probability theory plus noise account. *Journal of Behavioral Decision Making*. In press, available at http://dx.doi.org/10.1002/bdm.1936.

Costello, F. and Watts, P. (2016b). Probability theory plus noise: replies to Crupi and Tentori (2015) and to Nilsson, Juslin and Winman (2015). *Psychological Review*, 123(1):112–123.

Eva, K. W. and Norman, G. R. (2005). Heuristics and biases: biased perspective on clinical reasoning. *Medical Education*, 39(9):870–872.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62:451–482.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4):237.

Kahneman, D. and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Oliver, A. (2013). From nudging to budging: using behavioural economics to inform public sector policy. *Journal of Social Policy*, 42(04):685–700.

Shafir, E. and Leboeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, 53(1):491–517.

Sunstein, C. (2000). *Behavioral Law and Economics*. Cambridge University Press.

Vallgårda, S. (2012). Nudge: A new and better way to improve health? *Health Policy*, 104(2):200–203.

Influence of 3D images and 3D-printed objects on spatial reasoning

Akihiro Maehigashi (mhigashi@cog.human.nagoya-u.ac.jp)
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Masahiro Oda (moda@mori.m.is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University, Japan

Yoshihiko Nakamura (ynakamura@jo.tomakomai-ct.ac.jp)
Department of Computer Science and Engineering, National Institute of Technology, Tomakomai College, Japan

Kensaku Mori (kensaku@is.nagoya-u.ac.jp)
Information and Communications, Nagoya University, Japan

Tsuyoshi Igami (igami@med.nagoya-u.ac.jp)
Graduate School of Medicine, Nagoya University, Japan

Abstract

In this study, we experimentally investigated the influence of a three-dimensional (3D) graphic image and a 3D-printed object on a spatial reasoning task in which participants were required to infer cross sections of a liver in a situation where liver resection surgery was presupposed. The results of the study indicated that using a 3D-printed object produced more accurate task performance and faster mental model construction of a liver structure than a 3D image. During the task, using a 3D-printed object was assumed to reduce cognitive load and information accessing cost more than using a 3D image.

Keywords: External representation; 3D print; Spatial reasoning; Mental model

Introduction

Spatial reasoning and external representations

Spatial reasoning refers to inferring an object's shape and structure and the physical relationship between objects using spatial information (e.g., Byrne & Johnson-Laird, 1989). Spatial reasoning is ubiquitous in daily activities such as planning routes, inferring a road's slope angle, or arranging furniture in a room.

External representations such as figures, tables, and graphs are often used for spatial reasoning (Hegarty, 2011). Many studies on distributed cognition theory demonstrated the effects of using external representations on cognitive activity (e.g., Zhang & Norman, 1994). External representations can store information externally and reduce working memory load (Zhang & Norman, 1994). Physically manipulating external representations allows people to save mental rotation efforts (Kirsh & Maglio, 1994). Furthermore, spatially organized information on external representations could allow the offloading of cognitive processes onto perceptual processes (Scaife & Rogers, 1996).

Many studies on spatial reasoning have shown that different external representations of the same information have different effects on spatial reasoning (e.g., Hegarty, 2011). John, Cowen, Smallman, and Oonk (2001) experimentally investigated the effects of using two-dimensional (2D) and three-dimensional (3D) graphic images of the same spatial information for spatial reasoning and found that 3D images were more effective than 2D images in providing an understanding of shapes and layouts; 3D images integrate the multiple perspectives expressed by 2D images into a single perspective, provide supplementary depth cues, and display object features that would be invisible in 2D images. In contrast, they also found that 2D images were more useful than 3D images for an understanding of relative positions because 2D images display only necessary information and allow people to focus on it.

Other studies have shown that 3D images are more useful than 2D images only for people with high spatial ability (Hegarty, Keehner, Cohen, Montello, & Lippa, 2007; Nguyen, Nelson, & Wilson, 2011). Spatial ability is the ability to mentally store and manipulate spatial representations accurately (Hegarty & Waller, 2005). High spatial ability individuals can infer a structure's internal representation, whereas low spatial ability individuals cannot accurately construct a structure's internal representation and tend to depend on external representations (Kali & Orion, 1996). Because people with high spatial ability can recognize the complex spatial information in a 3D image with less difficulty, they can better take advantage of the information (Nguyen et al., 2011).

3D-printed objects

The recent prevalence of 3D printers has made it possible for people to replicate objects. 3D printers give people a totally new and unprecedented way of displaying information and have been used in various fields such as education, industrial manufacturing, and medicine. However, very few studies have investigated the influence of 3D-printed objects on spatial reasoning.

Some studies experimentally investigated human understanding of molecular structures using concrete models (Barrett, Stull, Hsu, & Hegarty, 2015; Stull, Barrett, & Hegarty, 2013). In their experimental tasks, participants learned molecular structures using 3D images or concrete models. After their learning the structures, they were required to orient the 3D images or concrete models in the same direc-

tions as the molecular structures depicted on paper. The results of these experiments demonstrated no difference in task accuracy between the use of 3D images and concrete models. However, the task completion time was shorter with the use of 3D images than with concrete models. Based on these results, they concluded that a 3D image was more useful than a concrete model for understanding physical structures. However, in their experiments, task accuracy rate was very high. Therefore, further investigations that consider situations requiring people to understand more complex structures with physical object models are necessary.

Furthermore, Maehigashi et al. (2015) investigated, using an ethnographic method, the influence of using a 3D-printed liver model on doctors during liver resection surgery. Protocol analyses results revealed that using 3D-printed models helped doctors elaborate their mental models of a patient's liver, mentally simulate the liver resection accurately, and share a similar mental model with other doctors. They also suggested the possibility that a 3D-printed model enhances mental model elaboration more than a 3D image.

In this study, we experimentally investigated the influence of 3D images and 3D-printed objects on a spatial reasoning task in which participants were to infer cross sections of a liver in a situation where liver resection surgery was presupposed. We tested the following two hypotheses: (1) spatial reasoning would be more accurate when a 3D-printed object was used than when a 3D image was used and (2) the learning time for constructing a mental model would be shorter when a 3D-printed object was used than when a 3D image was used.

Experiment

Participants memorized or referred to a liver's internal structure displayed by a 3D image or a 3D-printed object and inferred the locations of veins on a certain cross section of a liver and a tumor in the liver.

Method

Participants Forty-eight university students participated in this experiment.

Factorial design The experiment had a two-factor mixed design. The factors were (1) external representation (image and object) between participants and (2) task situation (memory and reference) within participants.

Material Two desks, a primary and a secondary desk, were used in the experiment. The primary desk (representing an operating table in a surgical setting) was set in front of a participant, and the secondary desk (representing a tool stand in a surgical setting) was set on the participant's right side. Three boxes were placed on the primary desk. Each box contained a 3D-printed model of a liver (target) (representing a patient's liver) and an answer sheet. On the secondary desk was either a computer on which a liver's 3D image was displayed or a box containing a liver's 3D-printed object created using the same manufacturing method as the 3D-printed model of the liver used for surgery. Figure 1 shows a 3D image, a 3D-printed object, and a target.



(a) 3D image (b) 3D-printed object (c) Target

Figure 1: (a) 3D image, (b) 3D-printed object, and (c) target

The 3D image was created with Pluto, a computer-aided diagnosis system developed at Nagoya University's Graduate School of Information Science, using data from a patient's liver measured by computed tomography (CT) (Figure 1a). In the 3D image, the thickest vein, an inferior vena cava (IVC), and five veins branching from the IVC were represented in blue, and a tumor was represented in white. The participants could rotate, zoom in on, and zoom out of the image using a mouse.

The 3D-printed object and the three targets were created with a 3D printer using the same CT liver data as the 3D image (Figure 1b, 1c). In particular, a 0.02-mm thick layer of acrylic resin was laid down in approximately 4,000 layers to produce the 3D-printed object and the target. The extra resin was then melted and removed, and the surface of the printed liver was polished. The 3D-printed object shows a liver's inside structure. In the 3D-printed object, the IVC, the five veins, and the tumor had the same relative scale and color as those in the 3D image. In contrast, the liver's inside structure was invisible in the target as the inside structure of a patient's liver is invisible during surgery. The target's surface was colored light gray. A line was drawn around each of the three targets created from the same CT liver data. Each line was drawn at a different location. Furthermore, on each target, the letters "A" and "B" were represented and indicated the two separated areas based on the drawn line. Two sets of 3D images, a 3D-printed object, and three targets were created from different CT liver datum.

Experimental task The experiment employed a spatial reasoning task in which participants were required to take a vein and a tumor location test for each target after examining a 3D image or a 3D-printed object. In the vein location test, participants were required to indicate the locations of the veins that appeared on the cross section resulting from cutting the target along the drawn line. In particular, participants were required to mark "O" for the IVC and "X" for the branching vein on the cross section's outer contour that was printed on the answer sheet (Figure 2). There were three types of cross sections: one with no IVC and two branching veins, another with one IVC and two branching veins, and the last with one IVC and three branching veins. In the tumor location test, the participants were to identify the area in the liver, A or B, where the tumor occurred.



(a) Contour of cross section of liver (b) Cross section of liver (c) Participant's answer

Figure 2: Figure 2: Vein location test. (a) Contour of cross section of liver, (b) cross section of liver, and (c) participant's answer. (a) shows the outer contour of a liver's cross section printed on the answer sheet. (b) shows an actual cross section of a liver. (c) shows a participant's answer, which provides the number of IVCs, O, and the branching veins, Xs, (drawn correctly here).

Procedure

Twenty-four participants were randomly assigned to the image condition where they conducted the experimental task with the 3D image, and the other participants were assigned to the object condition where they conducted the task with the 3D-printed object. First, the participants took anatomical and spatial ability tests. The anatomical test comprised five questions on the names of the liver's regions and veins. The spatial ability test was produced by Guay and McDaniels (1976) and comprised 24 questions requiring mental rotations. The participants were required to answer as many questions as possible in three minutes. Next, all the participants performed a practice task in the memory and reference task situations. In the practice task, the 3D image or the 3D-printed object representing one IVC and three branching veins were used. During the learning period, the participants used the 3D image or the 3D-printed object for one to three minutes to memorize the inside structure in a memory task situation and observe it in a reference task situation. After that, they took the vein and tumor location tests for one target.

After the practice, all participants conducted the experimental task in the memory and reference task situations. During the learning period, participants memorized the inner liver structure in the memory task situation and observed it in the reference task condition for three to five minutes using the 3D image or the 3D-printed object. When the participants had judged themselves prepared for the tests after at least three minutes had passed, or when five minutes had passed, the tests began. Participants took out the target and the answer sheet from one of three boxes on the primary desk and attempted the vein and tumor location tests. In the memory task situation, the computer display was switched off in the image condition, and the 3D-printed object was put into a box on the secondary desk in the object condition. In the reference task situation, the computer display stayed switched on in the image condition and the 3D-printed object remained on the secondary desk in the object condition. The participants could refer to the 3D image or the 3D-printed object freely while taking the tests in the reference task situation. Moreover, for the memory and reference task situations, different 3D images or 3D-printed objects created from the different CT liver datum were used.

The answer sheet provided for the vein and tumor location tests and two questionnaires. In the questionnaires, the participants rated their confidence toward their answers in the vein and the tumor location tests on a 7-point scale from (1) not confident at all to (7) extremely confident. After the participants completed the tests and questionnaires for one target, they returned the target and answer sheet to the box and took another set from another box. One of the two task situations was completed when they completed the tests and questionnaires for all three targets. A five-minute break was given between the task situations.

The order of the task situations was counterbalanced between the participants. The combinations of CT liver datum and task situations were also counterbalanced between the participants. Three sets of targets and answer sheets were randomly placed in the boxes on the primary desk. Participants were instructed to perform the tasks as accurately as possible. Furthermore, removing the target from the primary desk was forbidden during the experiment because it would be impossible for doctors to remove a patient's liver from the operating table during surgery. However, removing the 3D-printed object from the secondary desk was permitted in the object condition because doctors can place a liver's 3D-printed model right beside a patient's liver to confirm the interior structure of the liver during surgery (Maehigashi et al., 2015).

Results

None of the participants answered any of the questions on the anatomical test correctly, and no significant difference emerged between the image ($M = 9.08$) and the object ($M = 7.88$) conditions in the spatial ability test ($t(46) = 1.01, p = .32$). These results confirmed the homogeneity of the participants' anatomical knowledge and the homogeneity of spatial abilities between the conditions.

Next, we conducted 2(External representation: image and object) $\times$ 2(Task situation: memory and reference) analysis of variance (ANOVA) on the following dependent variables. First, the analysis was conducted on the learning time. The learning time was the mean time used by the participants to memorize or observe the inner structure of the 3D image or the 3D-printed object before attempting the tests in each condition (Figure 3). Results showed no significant interaction ($F(1,46) = 0.14, p = .71$). There was a significant main effect on the external representation factor, indicating that the learning time was shorter for the object condition than for the image condition ($F(1,46) = 7.72, p < .01$). The task situation factor also showed a significant main effect, indicating that the learning time was shorter for the reference condition than for the memory condition ($F(1,46) = 23.35, p < .001$).

Moreover, as the vein location test score, we calculated the mean absolute difference value between the correct number of veins and the number of drawn veins on the answer sheet in each condition for the IVC and the branching veins respec-

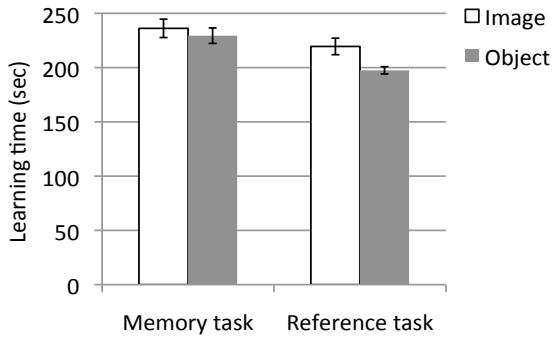


Figure 3: Learning time. The error bars indicate the standard error.

tively; the closer to zero the score, the more accurate the number of the drawn veins. First, for the IVC, analysis results showed no significant interaction ($F(1,46) = 1.73, p = .19$). There was no significant main effect on the external representation factor ($F(1,46) = 0.50, p = .48$). There was a marginally significant main effect on the task situation factor ($F(1,46) = 3.08, p = .09$). Next, for the branching veins, there was no significant interaction ($F(1,46) = 0.09, p = .77$) (Figure 4). There was no significant main effect on the task situation factor ($F(1,46) = 2.19, p = .15$). However, there was a significant main effect on the external representation factor, indicating that the the number of the veins was more accurately drawn for the object condition than for the image condition ($F(1,46) = 8.30, p < .001$).

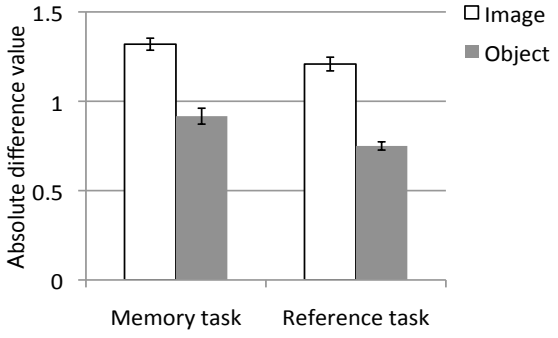


Figure 4: Absolute difference value for branching veins. The error bars indicate the standard error.

Furthermore, in each tumor location test, if the tumor location was correctly answered, a score of one was assigned. The tumor location test score was the mean total score of the tests for the three targets in each condition (Figure 5), meaning that the higher the score, the more accurate the answer. This analysis found a significant interaction ($F(1,46) = 18.98, p < .001$). Next, we conducted a simple main effect test on the external representation factor and found a marginally significant difference for the memory condition ($F(1,92) = 3.38, p = .07$) and a significant difference for the reference condition ($F(1,92) = 21.15, p < .001$), thus indicating higher scores in the object condition than in the image condition. We also conducted a simple main effect test on the task situation factor and found a marginally significant difference for the image condition, indicating that the score was lower in the reference condition than in the memory condition ($F(1,46) = 2.85, p = .10$). In contrast, no significant difference was observed for the object condition ($F(1,46) = 1.46, p = .23$). A significant main effect was observed on the external representation factor ($F(1,46) = 18.98, p < .001$), but not on the task situation factor ($F(1,46) = 0.12, p = .73$).

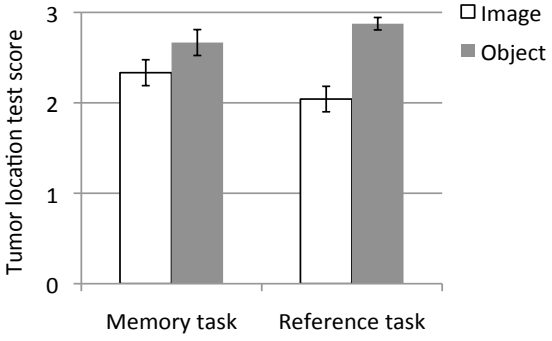


Figure 5: Tumor location test scores. The error bars indicate the standard error.

In addition, the analysis was conducted on the confidence ratings for the vein and the tumor location tests. No interaction was found in the rating for the vein location test ($F(1,46) = 1.34, p = .25$). No main effect was observed on the external representation factor ($F(1,46) = 0.40, p = .53$), but a significant main effect was observed on the task situation factor, indicating higher confidence ratings in the reference condition than in the memory condition ($F(1,46) = 47.06, p < .001$). Moreover, no interaction was found in the rating for the tumor location test ($F(1,46) = 1.13, p = .29$). No main effect was found on the external representation factor ($F(1,46) = 0.44, p = .51$), but a significant main effect was found on the task situation factor, thus indicating that the confidence rating was higher in the reference condition than in the memory condition ($F(1,46) = 36.27, p < .001$).

Finally, we conducted a correlation analysis on the relations between the spatial ability test score and the task performance, the learning time, and the vein and tumor location test scores in each condition (Table 1). In the memory task situation's object condition, a positive correlation was found between the spatial ability test score and the learning time. A negative correlation between the spatial ability test score and the vein location test score was also found for the branching veins. These results showed that participants with higher spatial ability tended to memorize the liver's inner structure more slowly and draw the number of branching veins accurately with the 3D image.

Table 1: Correlation matrices showing correlations between the spatial ability test score and the task performance, the learning time, and the vein and tumor location test scores in each condition. Values are correlation coefficients (r).

		Learning time	Vein location test score		Tumor location test score
			IVC	Branching vein	
Image	Memory task	.45**	-.14	-.44*	.30
	Reference task	-.22	-.19	-.13	-.19
Object	Memory task	.08	-.03	-.04	-.30
	Reference task	.17	.21	-.05	.18

$* p < .05$, $** p < .01$

Discussion

Accuracy of spatial reasoning

The vein and the tumor location test results indicated that the liver's inner structure was more accurately inferred when the 3D-printed object was used than when the 3D image was used, especially for the branching veins' structure in the vein location test. This result supported hypothesis 1 that stated that spatial reasoning would be more accurate with use of a 3D-printed object than with a 3D image.

It is possible that the participants in the object condition had a smaller cognitive load than those in the image condition. People perceive depth information in the real world more accurately than in the virtual 3D environment because the real world offers more depth cues (Kemeny & Panerai, 2003). This indicates that depth information is also missing from 3D images. Therefore, participants in the image condition might have to mentally complement or modify the 3D image's spatial information, temporarily storing this information in their memory and mentally resizing it to map the information to the target. Participants in the object condition, in contrast, were assumed to store the spatial information temporarily in their memory as they perceived it and map this information from the 3D-printed object directly to the target without internally complementing, modifying, or resizing it. Thus, participants in the object condition were assumed to have a smaller cognitive load and fewer errors from the internal manipulation of spatial information.

Moreover, it is also possible that the participants in the object condition incurred a lower information accessing cost than those in the image condition. Information accessing cost is incurred from acquiring information (Gray, Sims, Fu, & Schoelles, 2006). Participants in the image condition had to manipulate a computer mouse to acquire the required information, but participants in the object condition only had to pick up and physically rotate a 3D-printed object. Accessing information with a 3D-printed object was thus considered easier and less prone to errors or omissions than doing so with a 3D image.

In addition, in the vein location test, no difference was found between inferring the IVC's structure with the 3D image and the 3D-printed object. As the scores were very close to zero in all conditions, the test was considered easy and a ceiling effect was observed. Also, inferring the liver's inner structure was more accurate in the reference task situation than in the memory task situation. The participants in the memory task situation had to take the test using mental models constructed during the learning period. Participants in the reference task situation, in contrast, could continue updating their mental models and thus sustain more accurate models while attempting the test. As a result, the difference in the test score between the task situations was considered to be observed.

Furthermore, in the tumor location test, inference using the 3D image was more accurate in the memory task situation than in the reference task situation. This is because during problem solving, people tend to depend more on inaccurate memory with its low cost of accessing information than on seeking accurate external information, which has a higher cost (Gray & Fu, 2004). In our experiment, participants in the reference task situation using a 3D image might have avoided using the image because of the high cost of accessing information, depending instead on inaccurate memories, thus leading to lower test scores. The fact that the tumor location test was easier than the vein location test was also assumed to lead them to rely more on their inaccurate memory.

Learning time for mental model construction

Analysis of the learning time showed that the learning time was shorter when a 3D-printed object was used than when a 3D image was used. This result supported hypothesis 2 that stated that the learning time for constructing a mental model would be shorter when a 3D-printed object was used than when a 3D image was used.

This result could also be explained by the reduced cognitive load and information accessing cost when a 3D-printed object is used. Using the 3D image presumably required participants to mentally complement or modify the 3D image's spatial information. However, such internal manipulation was unnecessary when the 3D-printed object was used. Moreover, manipulating a computer mouse could involve more cost to access the required information than picking up and rotating a 3D-printed object. This could explain why learning times were shorter when a 3D-printed object was used than when a 3D image was used.

Moreover, the learning time was shorter in the reference task situation than in the memory task situation. In the mem-

ory task situation, participants had to construct mental models as accurately as possible during the learning period because they were not allowed to refer to the external representations during the tests. In the reference task situation, however, participants could refer to external representations during the tests, so they did not have to construct elaborate mental models during the learning period.

Effect of spatial ability and confidence on spatial reasoning

Participants with higher spatial ability tended to take a longer learning time and draw the number of branching veins accurately only when the 3D image was used in the memory task situation. Inferring the branching vein structure was the most difficult task in this experiment. Therefore, the effects of one's spatial ability were considered to be observed in the test score for the branching veins. Also, in the memory task situation, the participants had to construct mental models as accurately as possible and attempt the test without referring to any of the external representations. Therefore, the effects of one's ability to store and manipulate the spatial representation were considered to be prominent in the memory task situation. Because participants with higher spatial ability could store and manipulate the complex spatial representation, they were assumed to tend to take longer learning time to elaborate mental models and inferred the liver's inner structure accurately. Moreover, it is also possible that using a 3D-printed object might cancel the effect of using one's spatial ability to infer a physical structure, particularly raising the performance of participants with lower spatial ability, although we did not acquire sufficient data to support it in this experiment.

At last, the confidence ratings for the vein and tumor location tests showed an effect of the task situation and no effect of external representation. As people are sensitive to a task's cognitive load (Chandler & Sweller, 1991), participants reported themselves less confident in the memory task situation, which required a higher cognitive load than in the reference task situation. People are also sensitive to the cost of manipulating external representations (Gray et al., 2006). However, this cost is usually evaluated unconsciously (Walsh & Anderson, 2009). Therefore, the costs and effects of using external representations are considered difficult to be evaluated subjectively.

References

Barrett, T. J., Stull, A. T., Hsu, T. M., & Hegarty, M. (2015). Constrained interactivity for relating multiple representations in science: When virtual is better than real. *Computers and Education*, *81*, 69–81.

Byrne, R. M. J., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory and Language*, *28*, 564-575.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*, 293-332.

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, *28*, 359-382.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*, 461–482.

Guay, R., & McDaniels, E. (1976). *The visualization of viewpoints*. West Lafayette: The Purdue Research Foundation.

Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, *3*, 446-474.

Hegarty, M., Keehner, M., Cohen, C., Montello, D. R., & Lippa, Y. (2007). The role of spatial cognition in medicine: Applications for selecting and training professionals. In G. L. Allen (Ed.), *Applied spatial cognition* (p. 285-315). Mahwah, NJ: Erlbaum.

Hegarty, M., & Waller, D. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The cambridge handbook of visuospatial thinking* (p. 121-169). New York: Cambridge University Press.

John, M. S., Cowen, M. B., Smallman, H. S., & Oonk, K. M. (2001). The use of 2d and 3d displays for shape-understanding versus relative-position tasks. *Human Factors*, *43*, 79-98.

Kali, Y., & Orion, N. (1996). Spatial abilities of high school students in the perception of geologic structures. *Journal of Research in Science Teaching*, *33*, 369-391.

Kemeny, A., & Panerai, F. (2003). Evaluating perception in driving simulation experiments. *Trends in Cognitive Sciences*, *7*, 31–37.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513-549.

Maehigashi, A., Miwa, K., Terai, H., Igami, T., Nakamura, Y., & Mori, K. (2015). Investigation on using 3d printed liver during surgery. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 1476-1481.

Nguyen, N., Nelson, A. J., & Wilson, T. D. (2011). Computer visualizations: Factors that influence spatial anatomy comprehension. *Anatomical Sciences Education*, *5*, 98-108.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, *45*, 185-213.

Stull, A. T., Barrett, T., & Hegarty, M. (2013). Usability of concrete and virtual models in chemistry instruction. *Computers in Human Behavior*, *29*, 2546–2556.

Walsh, M. M., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology*, *58*, 416–440.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*, 87-122.

Making it Right: Can the Right-Hemisphere Compensate for Language Function in Patients with Left-Frontal Brain Tumors?

Ethan Jost (edj37@cornell.edu)
Cornell University, Department of Psychology
Ithaca, NY 13068 USA
Memorial Sloan Kettering Cancer Center, fMRI Lab
New York, NY 10065 USA

Nicole P. Brennan (brennann@mskcc.org), Kyung K. Peck (peckk@mskcc.org), Andrei I. Holodny (holodnya@mskcc.org)
Memorial Sloan Kettering Cancer Center, fMRI Lab
New York, NY 10065 USA

Morten H. Christiansen (christiansen@cornell.edu)
Cornell University, Department of Psychology
Ithaca, NY 13068 USA

Abstract

Both the degree to which the left-hemisphere is specialized for language and the relative ability of the right-hemisphere to subserve language function are underspecified. The present study sought to identify whether the right-frontal fMRI activation seen in a number of case studies in patients with left-sided brain lesions exists as a group-level trend in patients with left-frontal tumors. It also sought to examine the possible compensatory nature of this activation. Thus, a retrospective analysis of 197 brain tumor patients who had undergone pre-surgical fMRI language mapping was conducted. Patients with left-frontal tumors were found to be more likely to show right- or co-dominant fMRI activation during language mapping tasks compared to patients who had tumors elsewhere in the brain. Further, patients with left-frontal tumors who were identified as right- or co-dominant for language were found to possess more intact language function as measured by the Boston Naming Test.

Keywords: language; neuroplasticity

Introduction

Those studying the cognitive neuroscience of language have typically described the brain's left-hemisphere as specialized for language ever since the seminal work of Paul Broca and Carl Wernicke. These studies, along with the century and a half of research they inspired, have focused primarily on uncovering the roles of the left frontal and temporal lobes in the learning and processing of language (Dronkers, Plaisant, Iba-Zizen & Cabanis, 2007). This has led to an emerging characterization of the role that various regions, sub-regions, and pathways within the left-hemisphere's language network play in language processing (Hagoort, 2014; Friederici & Gierhan, 2013).

However, a lack of research on the role of the right-hemisphere in language has left the debate on specialization open. Even with the focus of research on left-hemisphere language function, some theories posit a complementary role for right-hemisphere processing (e.g., Jung-Beeman, 2005). In addition, there is much evidence that the right-hemisphere, specifically Broca's homologue, is able to take over function from the left-hemisphere following extensive early brain damage (Thal et al., 1991; Vicari et al., 2000; Tivarus, Starling, Newport & Langfitt, 2012). However, the degree to which adults with brain lesions retain this capacity is currently underspecified. Some recent smaller-scale work has suggested that increased right hemispheric activation may be correlated with better language outcomes in patients with left-frontal lesions, and in healthy patients who undergo targeted rTMS (Krieg et al., 2013).

Determining the degree to which the adult brain can reorganize language function, and under what conditions this occurs, promises to yield important insights that not only may inform clinical prognoses, but also lead to a better understanding of the brain's equipotentiality. Examining brain tumor patients may thus provide a fruitful source of data for uncovering the potentially more subtle plasticity that exists in adult populations than acute stroke and brain damage patients, as they may exhibit different compensatory mechanisms and competencies (Fisicaro et al., 2016).

Case studies with brain tumor patients have indicated that damage to traditional language cortex in both the frontal (Holodny, Schulder, Ybasco & Liu, 2002) and temporal (Petrovich, Holodny, Brennan & Gutin, 2004)

lobes can result in contralateral reorganization of the brain's language network, although how this affects patients' cognitive outcomes is not currently known. In addition, the fact that ipsilateral reorganization also occurs in this patient population (Brennan, 2008), creates an ideal setting in which to test for both the relative frequency of contralateral reorganization and its outcomes with respect to language. Does the right-hemisphere demonstrate compensatory activation when there is damage to the left-frontal lobe more often than the occasional case study would suggest? Also, do patient outcomes differ depending on whether the patient exhibits ipsilateral versus contralateral compensation? This set of questions led to the present study, in which a large database of patients with brain tumors of varying grade, size, location, and etiology was queried in order to answer these critical questions. We sought to identify whether or not the compensatory right-frontal activation seen in such case studies exists as a group-level trend among patients suffering from left-frontal tumors, and also to examine the possible compensatory nature of this activation.

Methods

Participants

A database was compiled involving all patients who underwent a pre-surgical fMRI language procedure at Memorial Sloan Kettering Cancer Center over a five-year time period, excluding patients with sub-optimal scans or incomplete patient information (n = 197, 95 female; mean age = 50.52 years, range: 10-83). Patients initially presented with a range of symptoms and tumor etiologies. Table 1 indicates the proportion of patients falling into a number of relevant categories.

Table 1: General patient information.

	Female	High Grade Tumor	Left-frontal tumor	Right-handed	Left-dominant language
Proportion of Patients	0.482	0.538	0.482	0.807	0.873

Data Acquisition

Data were acquired with a 1.5-T or 3.0 T scanner (General Electric, Milwaukee, WI) using an 8 channel head coil. Based on localizer images, a set of 26 T1-weighted (repetition time [TR], 600 ms; echo time [TE], 8 ms; thickness, 4.5-mm) and T2-weighted (TR, 4000 ms; TE, 102 ms; thickness, 4.5-mm) spin-echo axial slices, covering the whole brain, was obtained for the purpose of coregistration with the functional data. Functional images were acquired with a gradient-echo echo-planar imaging sequence (TR, 4000 ms; TE, 30/40 (for 3T/1.5T) ms; matrix , 128x128; field of view, 240 mm; thickness, 4.5-mm;flip angle, 90^0). Head motion was minimized using straps and foam padding.

Task Administration

Patients performed one or more covert block designed fMRI language tasks as part of the pre-surgical language task panel. Tasks consisted of a phonemic fluency task where they were required to generate words that began with a high frequency letter, semantic fluency where they were required to generate words that fit a category, verb generation where they were required to generate verbs to given nouns, or auditory responsive naming where they were required to answer simple questions. All tasks were delivered aurally (Ruff et al., 2008). The radiology team's reports were based on all scans collected for each patient, while reported laterality indices were taken from the task with the highest quality scan, with a preference for the phonemic fluency task.

There were 90 images in total for each patient, consisting of 5 activation images (20 sec) followed by 10 rest images (40 sec) repeated 6 times (6 min total). Subjects were monitored continuously while performing the task. Subject participation was confirmed using real-time imaging software, which provided real-time acquisition, processing, and display of functional results (Brainwave RT, GE Healthcare, Milwaukee, Wisconsin).
All language fMRI tasks were visually inspected for lateralization patterns and discrepancies in language lateralization by a board certified neuroradiologist. Clinical reports indicated language lateralization, language localizations in the peri-tumoral region, handedness, and if the Boston Naming Test (BNT) was performed, the patients score.

Results

Plasticity in the brain's language network

In order to determine the possible extent of both ipsilateral and contralateral reorganization following tumor infiltration of the left-frontal lobe, the researchers divided all right-handed patients in the original sample (n = 159, 72 female; mean age = 50.64, range: 10-83) into groups based on tumor location. Patients with tumors impacting the left-frontal lobe (n = 81, 37 female; mean age = 49.65 years, range: 11-75) can be considered the experimental group, while patients with tumors elsewhere in the brain – including other regions within the left-

hemisphere – (n = 78, 35 female; mean age = 52.35, range: 10-83) served as the control[1].

As can be seen in Table 2, patients with tumors localized in the left-frontal lobe were much more likely to be reported as right- or co-dominant for language function by the neuroradiology staff than would be expected given the control group ($\chi2 = 9.51$, $p = .002$, $\varphi = .245$).

This finding suggests that the frontal component of the typical left-hemisphere language network is able to shift to the right-hemisphere in patients with tumor impacting putative Broca's area in the left-frontal lobe. It also demonstrates that this contralateral reorganization happens with some frequency, and is not isolated to a small number of cases. However, it remains unclear whether the plasticity encountered in this patient population has any effect on language function. It is possible that the increased right-hemisphere activation is not compensatory in nature, and may even be deleterious.

Table 2: χ^2-table exhibiting the difference in report laterality between patients with left-frontal tumors and those with tumor elsewhere in the brain.

	Patients reported as left-dominant	Patients reported as right-or co-dominant	
Patients with left-frontal tumors	67 (72.85)	14 (8.15)	81
Patients with tumors elsewhere in brain	76 (70.15)	2 (7.85)	78
	143	16	159

Patterns of reorganization affect language outcomes

To elucidate the role that potential right-hemisphere reorganization plays in actual language outcomes, patients with left-frontal tumors who completed a BNT during their pre-surgical assessment (n = 28) were evaluated more closely. In order to determine the effect of laterality on language performance, the BNT scores of patients with left-dominant reports (n = 20; M = 48.30; SD = 14.71) were compared to those with right- or co-dominant reports (n = 8; M = 56.38; SD = 2.72)[2].

As depicted in Figure 1, this analysis found that patients identified as being right- or co-dominant in terms of language function had significantly more intact language abilities than did patients with left-dominant maps exhibiting ipsilateral, and often perilesional activation to language tasks ($t(22) = -2.36$, $p = .028$). This independent-samples t-test did not pass Levene's test for equality of variances ($F = 4.82$, $p = .037$), so corrected degrees of freedom were used.

This set of analyses demonstrates that the right-hemisphere is capable of taking over language function in adults who were likely left-dominant prior to tumor development. It also shows that therapeutic techniques which focus on transferring language function to the right-hemisphere in the face of left-hemisphere damage may result in better patient outcomes.

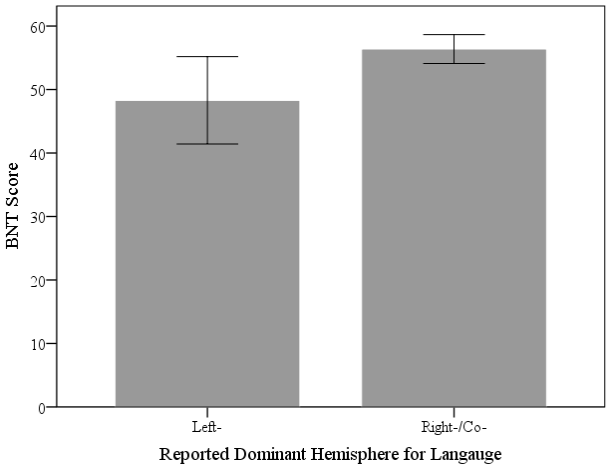


Figure 1: Bar graph displaying the difference in BNT performance between patients with left-dominant reports and those with right- or co- dominant reports. Error bars indicate 95% CIs.

Laterality index verification of report groups

In order to confirm the radiologist report data, a laterality index assessment also was conducted on the patients with BNT data. The patients in both the left-dominant and right-/co-dominant groups were also evaluated using laterality indices (LI), with interior frontal gyrus (IFG), middle frontal gyrus (MFG), frontal lobe, and hemispheric ROIs (see Figures 2-5).

Independent-samples t-tests confirmed that there was a significant difference between the report groups in laterality for each of the ROIs that were assessed, the IFG ($t(26) = 4.33$, $p < .001$), MFG ($t(26) = 4.08$, $p < .001$), frontal lobe ($t(26) = 4.44$, $p < .001$), and hemispheric

[1] There were no significant differences between patient groups in terms of age ($t(156) = -1.145$, $p = .254$), sex ($\chi2 = .01$, $p = .920$), or tumor grade ($\chi2 = 2.843$, $p = .092$).

[2] Again, no significant differences were found between groups for age ($t(156) = -1.145$, $p = .254$), or tumor grade ($\chi2 = 2.05$, $p = .152$), although sex and hemisphere of language dominance were not found to be independent ($\chi2 = 4.01$ $p = .045$, $\phi = .223$), as males were more likely to be right- or co-dominant than expected.

($t(26) = 5.78$, $p < .001$). See Table 3 for a report of means for each group's LI within each ROI.

This set of analyses corroborates the qualitative findings of the report data, and suggests that patients falling into the left- and right-/co-dominant groups were categorized correctly.

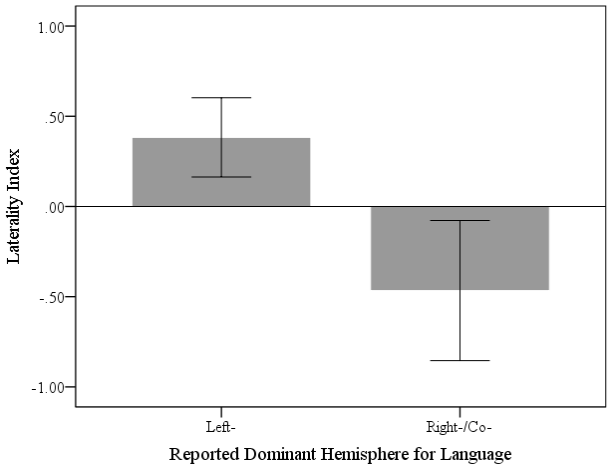


Figure 2: Difference in LI between patients with left-dominant reports and those with right- or co- dominant reports in the IFG ROI. Error bars indicate 95% CIs.

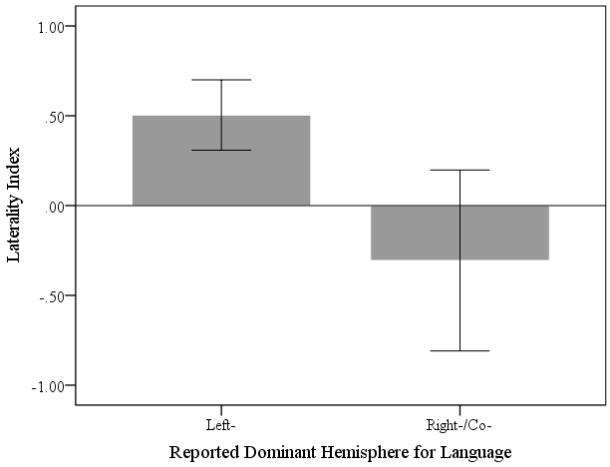


Figure 3: Difference in LI between patients with left-dominant reports and those with right- or co- dominant reports in the MFG ROI. Error bars indicate 95% CIs.

Discussion

This is the first large scale study investigating the complex relationship between reorganization in the brain's language network and behavioral outcomes as a result of brain tumor. While some past research with stroke patients has incorporated large datasets (Turkeltaub, Messing, Norise & Hamilton, 2011), case studies have made up the majority of past research on atypical language lateralization in adult patient populations.

Table 3: Group means for each ROI

	IFG	MFG	Frontal	Hemispheric
Left-dominant	0.383	0.504	0.472	0.389
Right-/co-dominant	-0.466	-0.306	-0.387	-0.153

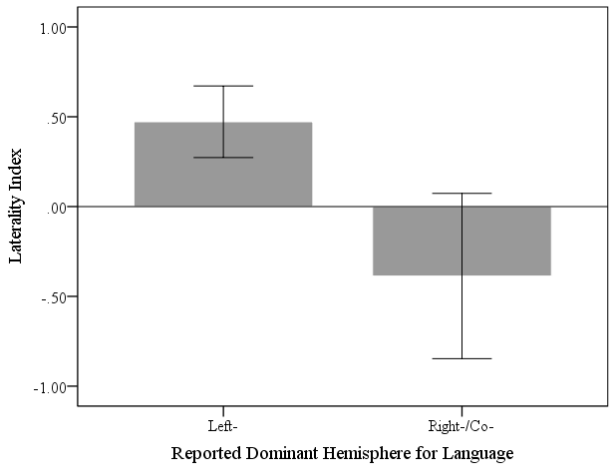


Figure 4: Difference in LI between patients with left-dominant reports and those with right- or co- dominant reports in the frontal lobe ROI. Error bars indicate 95% CIs.

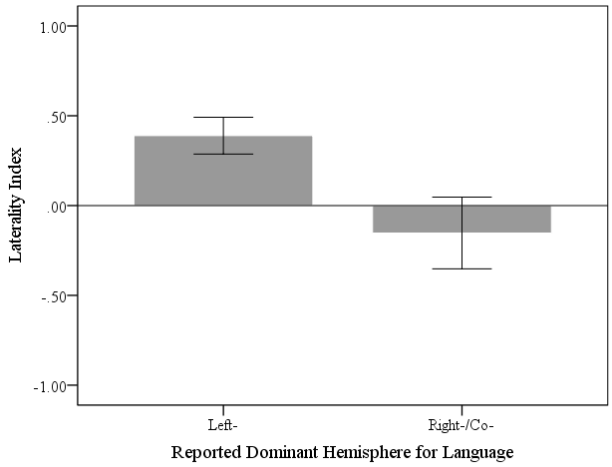


Figure 5: Difference in LI between patients with left-dominant reports and those with right- or co- dominant reports in the hemispheric ROI. Error bars indicate 95% CIs.

The findings of the present study point toward a degree of equipotentiality in the adult brain's language network that is somewhat unexpected. The frontal lobe, likely Broca's homologue, in the right-hemisphere seems to be able to take over language function when there is compromise of the left-hemisphere, and may even be a better candidate for reorganization than the ipsilateral cortex. It is also of note that the patients with left-hemisphere damage exhibiting right- or co-dominant activation maps fell within the normal-range of scores on the BNT, demonstrating the right-hemisphere's ability to maintain healthy levels of language performance. Patients with left-dominant maps indicating ipsilateral, and often peri-lesional compensation suffering from left-frontal lesions exhibited a much wider range of outcomes.

Such a pattern of results suggests that while typical development results in an overwhelming majority of adults having left-dominant language networks (Knecht et al., 2000), the right-hemisphere may also possess the requisite capacities for language processing. This falls in line with the work of Bates and colleagues (Bates & Roe, 2001; Thal et al., 1991; Vicari et al., 2000), whose work suggests that early language development can occur in children with severe damage to the left-hemisphere and in children who are lacking their left-hemisphere entirely. It is likely that the left-hemisphere is not uniquely suited or specifically adapted for language learning and processing. Rather, relatively ubiquitous and subtle biases in early development likely lead to the typically left-hemisphere language network. Future work determining the nature of these biases would be a fruitful addition to understanding the developmental intricacies of the brain's language network.

Clinical studies have also indicated that the right-hemisphere may be able to serve what is traditionally the role of the left-hemisphere in language in some circumstances (for a recent review, see Fisicaro et al., 2016). An investigation of a large number of stroke patients suffering from aphasia demonstrates that patients suffering from left-frontal damage recruit the right-hemisphere more often than healthy controls, although the findings of this study did not determine whether the compensatory activation is deleterious or beneficial (Turkeltaub et al., 2011).

Other case-study research has demonstrated an unclear and even conflicting relationship between the role of right-hemispheric activation and language outcomes in patients with aphasia at different points in recovery (Turkeltaub et al., 2012). Unlike this past research, however, the present study demonstrates a clear relationship between right-hemisphere activation during language tasks and positive patient outcomes in a large cohort. Determining whether or not this is true for stroke patients, or patients suffering from brain damage with other etiologies constitutes an intriguing question for future research.

Studies utilizing therapeutic techniques attempting to transfer language function to healthy right-hemisphere cortex after damage to the left-frontal lobe in two stroke patients have demonstrated that the right-hemisphere may be able to take over such function when the basal ganglia remains intact, and that this compensation leads to positive language outcomes for the patient (Crosson et al., 2005). Other recent research also seems to suggest that a rightward shift in language related function is related to the amount of tumor infiltration suffered by the basal ganglia (Shaw et al., submitted).

It seems likely that basal ganglia activation serves as the causal mechanism by which language function is able to transfer between hemispheres following left-hemisphere damage. Further research examining the trajectories and limitations of this transfer will hopefully not only elucidate mechanisms for contralateral reorganization in clinical populations, but also increase the understanding of how the hemispheres communicate during normal language processing.

The present study is limited in some regards. By utilizing radiologist reports for evaluations of laterality, the analyses are partially based on qualitative measures. A number of patient scans suffered from issues related to drop-out artifacts due to prior surgery and other common issues faced when scanning this patient population, rendering traditional quantitative analyses useless.

However, the limited laterality index findings reported here suggest that the radiologist report data overlaps quite a bit with more quantitative measures, at least at the group level, mitigating such concerns. In addition, the radiology team's reports were able to account for such issues in their analyses. For example, several patients whose scans exhibited multiple artifacts were difficult to categorize using traditional laterality indices due to low voxel counts in the affected hemisphere. By using the qualitative report data based on these scans, but not beholden to the limitations of traditional quantitative analyses, the study was able to include a larger number of patients.

Ideally future research will be able to control for such artifacts and issues in a more well-controlled prospective study, even if that means using a smaller sample. Future prospective research would also benefit from utilizing a broader range of neuropsychological assessments than were able to be accessed in the present study.

Conclusion

The present study is the first to indicate a group level trend in putative right-hemisphere reorganization as a result of damage to traditional left-frontal language cortex. It is also the first to demonstrate positive outcomes for patients as a result of such reorganization in a large sample. Such findings show that the left-hemisphere is

not likely to be uniquely adapted to subserve language function.

Rather, the right-hemisphere seems to be able to take over for left-hemisphere when the latter is damaged, and can maintain a normal-level of language abilities. This study also highlights the potential for therapies attempting to facilitate contralateral shifts in language function in patients suffering from language problems. As the right-hemisphere seems to be just as adept at subserving language function as the left-hemisphere, therapy targeted at shifting language function to healthy right-hemisphere cortex seems like an increasingly attractive option.

Acknowledgements

This research could not have taken place without the support of NIH grant 5T32HD055177. We would also like to thank all of the staff in the Radiology and Neurosurgery Departments at Memorial Sloan Kettering Cancer Center for their help at every stage of this project.

References

Bates, E., & Roe, K. (2001). Language Development in Children with Unilateral Brain Injury. In C.A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience.* Cambridge, MA: MIT Press.

Brennan N.P. (2008). Preparing the patient for the fMRI Study and optimization of paradigm selection and delivery. In A.I. Holodny (Ed.), *Functional neuroimaging: A clinical approach*. New York: Informa Healthcare USA, Inc.

Crosson, B., Moore, A.B., Gopinath, K., White, K.D., Wierenga, C.E., Gaiefsky, M.E., Fabrizio, K.S., Peck, K.K., Soltysik, D., Milsted, C., & Briggs, R.W. (2005). Role of the right and left hemispheres in recovery of function during treatment of intention in aphasia. *Journal of Cognitive Neuroscience, 17*(3), 392-406.

Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., & Cabanis, E. A. (2007). Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. *Brain, 130*(5), 1432-1441.

Fisicaro, R., Jost, E., Shaw, K., Brennan, N.P., Peck, K.K., & Holodny, A.H. (in press). Cortical plasticity in the setting of brain tumors. *Topics in Magnetic Resonance Imaging.*

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology, 23*(2), 250-254.

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current opinion in neurobiology, 28*, 136-141.

Holodny, A. I., Schulder, M., Ybasco, A., & Liu, W. C. (2002). Translocation of Broca's area to the contralateral hemisphere as the result of the growth of a left inferior frontal glioma. *Journal of computer assisted tomography, 26*(6), 941-943.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in cognitive sciences, 9*(11), 512-518.

Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., Ringelstein, E.B., & Henningsen, H. (2000). Handedness and hemispheric language dominance in healthy humans. *Brain, 123*(12), 2512-2518.

Krieg, S. M., Sollmann, N., Hauck, T., Ille, S., Foerschler, A., Meyer, B., & Ringel, F. (2013). Functional language shift to the right hemisphere in patients with language-eloquent brain tumors. *PLoS One, 8*(9), e75403.

Petrovich N.M., Holodny A.I., Brennan C.W., & Gutin P.H. (2004). Isolated translocation of Wernicke's area to the right hemisphere in a 62-year-man with a temporo-parietal glioma. *American Journal of Neuroradiology 25*, 130–133.

Ruff, I. M., Brennan, N. P., Peck, K. K., Hou, B. L., Tabar, V., Brennan, C. W., & Holodny, A. I. (2008). Assessment of the language laterality index in patients with brain tumor using functional MR imaging: effects of thresholding, task selection, and prior surgery. *American journal of neuroradiology, 29*(3), 528-535.

Shaw K., Brennan N.M., Woo K., Zhang Z., Young R., Peck K.K., & Holodny A.I. (submitted). Infiltration of the basal ganglia by brain tumors is associated with the development of co-dominant language function on fMRI.

Thal, D. J., Marchman, V., Stiles, J., Aram, D., Trauner, D., Nass, R., & Bates, E. (1991). Early lexical development in children with focal brain injury. *Brain and language, 40*(4), 491-527.

Tivarus, M. E., Starling, S. J., Newport, E. L., & Langfitt, J. T. (2012). Homotopic language reorganization in the right hemisphere after early left hemisphere injury. *Brain and language, 123*(1), 1-10.

Turkeltaub, P. E., Messing, S., Norise, C., & Hamilton, R. H. (2011). Are networks for residual language function and recovery consistent across aphasic patients? *Neurology, 76*(20), 1726-1734.

Turkeltaub, P. E., Coslett, H. B., Thomas, A. L., Faseyitan, O., Benson, J., Norise, C., & Hamilton, R. H. (2012). The right hemisphere is not unitary in its role in aphasia recovery. *Cortex, 48*(9), 1179-1186.

Vicari, S., Albertoni, A., Chilosi, A. M., Cipriani, P., Cioni, G., & Bates, E. (2000). Plasticity and reorganization during language development in children with early brain injury. *Cortex, 36*(1), 31-46.

Spatial Meaning is Retained in Emotion Metaphors: Some Evidence from Spanish

César Riaño (ce.riano557@uniandes.edu.co)
Department of Psychology, Universidad de los Andes
Bogotá, Colombia

Florencia Reali (f.reali96@uniandes.edu.co)
Department of Psychology, Universidad de los Andes
Cra 1a # 18a-12, Bogotá, Colombia
Centro de Investigación Básica en Psicología (CIBPSI) UdelaR,
Montevideo, Uruguay.

Abstract

Previous work has shown that the abstract use of the prepositions *in* and *on* retains spatial meaning, such as containment and support that includes the control relationship between a located object (the figure) and a reference object (the ground). We extend these ideas to the case of metaphorical descriptions of emotion in Spanish – some of them featuring the emotion as a located entity in the person´s body, and some of them featuring emotion as the ground in which the person´s body stands. Two rating experiments show that people judge emotions as more "controllable" when they are described as located entities (the figure) than when they are described as grounds.

Keywords: conceptual metaphors; emotion; spatial language; emotion; Spanish.

Introduction

Linguistic framing influences a range of cognitive domains such as time perception (Casasanto, 2010), social cognition (Landau, Meier & Keefer, 2010), memory (Fausey & Boroditsky, 2010; Fausey & Boroditsky, 2011), problem solving (Thibodeau & Boroditsky, 2011) and political attitudes (Landau, Sullivan & Greenberg, 2010; Matlock, 2012; Matlock et al., 2012). For example, the wording of political messages affects political attitudes (Fausey & Matlock, 2012) and the metaphorical framing of social concepts such as *crime*, affects people´s opinion about how to solve the problem (Fausey & Boroditsky, 2011). Emotion concepts are no exception: metaphorical framing has been shown to shape the way we evaluate them (e.g., Reali & Arciniegas, 2015; Richardson, Spivey, Barsalou & McRae, 2003).

We begin from the assumption that emotions are abstract concepts that are endowed with a complex conceptual structure grounded on basic perceptual schemas (Lakoff & Johnson, 1980). Certain schemas are embodied on physiological or physical experience – such as temperature rise or increased heart beat – providing grounding for the conceptual representation of emotion (e.g., Kövecses, 2010; Lakoff, 1987).

Linking emotions to image-like schemas provides a useful scaffolding for understanding them. If abstract emotions are grounded on concrete domains then their semantic structure will be aligned with the concrete features of perceptual schemas. For example, if emotions were represented in spatial terms, then spatial features – such as geometrical relationships – should be involved in affective evaluation (e.g., Crawford, Margolis, Drake & Murphy, 2006; Margolis & Crawford, 2008).

Evidence of the perceptual nature of emotion concepts has been recently provided. For example, people represent anger as being red (Fetterman, Robinson & Meier, 2012). Others have shown that positive and negative valence of emotion are associated to spatial metaphors (Crawford et al, 2006; Houser, Carter & Meier, 2009; Margolis & Crawford, 2008; Reali & Arciniegas, 2015), including up and down representations (Richardson, Spivey, Barsalou & McRae, 2003).

In this paper we assess the proposal that when emotion is metaphorically described in terms of a spatial relation between a located entity and a reference ground, the way people perceive certain aspects of the emotion – namely, control – depends on whether the emotion is described as a located entity in the person´s body (e.g., real happiness is in her) or, alternatively, as the ground in which the person´s body stands (she´s in a state of real happiness).

Spatial Meaning in Abstract Use of Prepositions

Certain linguistic expressions such as the prepositions *in* and *on* in English are used to establish spatial relationships between concrete entities (e.g., the fly is *on/in* the tailor´s hand). The prepositions *in* and *on* correspond to the preposition *en* in Spanish (e.g., la mosca está *en* la mano del sastre), which will be our case study here.

Abstract uses of prepositions are metaphorical because they go beyond the words´ concrete, literal meaning. Recent work has contributed to the understanding of abstract extensions of prepositions and how they relate to spatial meaning (e.g., Boers, 1996; Jamrosik & Gentner, 2015), including the use of prepositions in ontological metaphors in which events and other abstract notions are conceived as containers (e.g., she´s *in* a state of fury) (e.g., Lakoff & Johnson, 1980).

A number of studies have shown that the literal use of *in* and *on* to describe spatial events determines functional relations between the *figure* (e.g., the fly) and the *ground* (e.g., the tailor´s hand) (e.g., Coventry & Garrod, 2004; Feist & Gentner, 2003; Garrod et al., 1999).

The terms *figure* and *ground* (e.g., Talmy, 1983) are used to describe participants in the relationship named by the *in*

and *on* prepositions in English and the proposition *en* in Spanish. For example, the fly (the located entity, the figure) is *in* the soup (the ground).

Crucially, the use of the prepositions *in* and *on* is based on the principle of control of the figure by the ground. The basic idea is that functional relations, such as containment, entail inferences beyond geometric relations: if a figure is contained by a ground it can be inferred that if the ground moves, the figure moves with it (Garrod & Sanford, 1989; Jamrozik & Gentner, 2015). In English, because the use of *in* denotes geometric inclusion or containment, it involves greater ground control than does *on* (Coventry & Garrod, 2004; Garrod, et al, 1999).

Garrod and Sanford (1989) and, more recently, Jamrozik and Gentner (2015) proposed that the functional relations of containment and support might be extended to non-spatial (abstract) uses of *in* and *on* (e.g., she´s in a risky situation). They proposed that the metaphorical use of *in* and *on* helps establish the relative degree of control between an abstract object (the figure) and an abstract reference object (the ground).

A natural proposal that follows from these studies is that when the prepositions *in* and *on* are used to describe emotions certain control relations may be inferred. Garrod and Sanford (1989) even suggested that metaphorical framing of emotion (e.g., she´s *in* a bad mood), follows the same pattern. To our knowledge, however, this proposal has not been empirically tested for the case of emotion. The goal of this study is to assess empirically if spatial meaning – namely, the control relationship between the figure and the ground – is retained when metaphorical expressions of emotion are used.

Figure/Ground Relation in Metaphors of Emotion

The studies described above suggest that the control relationship between the figure and the ground are mapped from concrete to abstract domains. In this paper we use the Spanish preposition *en* to frame emotion metaphorically either as a located entity (the figure) in a ground (the body of the person´s experiencing the emotion) vs. as a place (the ground) containing a located object (the person´s body). We assess whether metaphor framing affects how controllable people judge emotion to be.

Lakoff (1987) proposed that emotions are grounded on primary metaphors such as THE BODY IS A CONTAINER FOR EMOTIONS and EMOTIONS ARE BOUNDED SPACES. The first metaphor conveys emotions as located entities (the figure) in a container (the person´s body), while the second conveys emotions as grounds in which the person stands. Figure 1 depicts a sketch of possible conceptual schemas associated with these two primary metaphors. Because of the concrete nature of spatial schemas, these are good metaphor examples to study the effect of image-like representations on the conceptualization of emotion.

Corpus analyses have shown that linguistic instantiations of these metaphors are quite common in Spanish (e.g., Reali & Arciniegas, 2015; Reali, Soriano & Rodriguez, in press). Consider the expression "no has de caer en la desesperación" (you must/should not fall into despair) (Reali & Arciniegas, 2015, p. 28). In this example, "despair" corresponds to the ground and "you" corresponds to the figure. Conversely, consider "Gracián corrió hacia la iglesia totalmente enloquecido lleno de odio" (Gracián ran towards the church totally insane, full of hate) (Reali & Arciniegas, 2015, p. 30). In this example hate corresponds to the entity (the figure) inside Gracián (the ground).

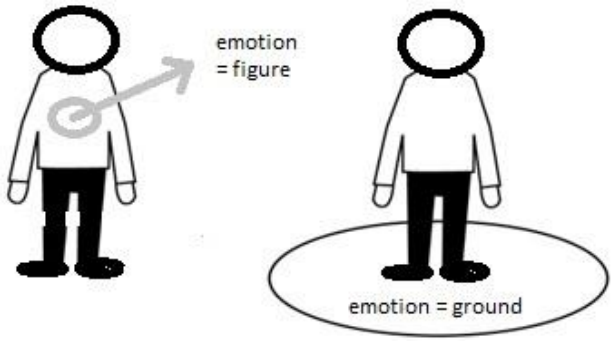


Figure 1. Emotions as located objects (the figure) in a ground (the person´s body) (left) and emotions as places (the ground) in which the person´s body (the figure) stands (right).

Expressions of this kind are common in Spanish, providing a nice case study to test the proposal that concrete properties of physical ground-figure settings – namely the control relationship between the figure and the ground– is preserved in the semantic structure of emotion. That is precisely the aim of this paper.

The prediction to be tested is the following: if spatial meaning is retained in metaphors of emotion, then when emotion is described as a located entity contained in the person´s body, a higher perception of emotion controllability should be found than when emotion is described as a ground in which the person´s body stands. Two questionnaire experiments were conducted to assess this hypothesis. To preview, the results show that metaphorical forms in which emotion is described as the located figure is associated to higher ratings of controllability.

Study 1

Study 1 was designed to explore whether the instantiations of emotion-as-ground vs. emotion-as-figure metaphors influence people's judgments of emotion control. The experiment was inspired by Jamrozik & Gentner (2015). Participants saw sentences describing situations involving a person experiencing a certain emotion as well as a cause or context. They were asked to rate the degree of control they thought the person had over the emotion. The metaphorical framing of emotion was manipulated so that emotion was described as the located figure (emotion-as-figure condition) or as the ground (emotion-as-ground condition).

Method

Participants A hundred and eight undergraduate students from Universidad de los Andes (Bogotá, Colombia) volunteered to participate in this Study or received extra course credits in exchange of their participation. All participants were 18 years-old or older and declared that their native language was Spanish.

Materials For both experiments the stimuli consisted of twelve sentence items gathered in one questionnaire. Each participant was presented with one of two version of each item, which consisted on a description of a fake situation of a person experiencing an emotion as well as a context for it. The two versions of each item differed only in the choice of metaphorical framing. In one condition, the emotion was framed as a located entity (the figure) in the human body (the ground), and in the second condition it was described as the place (the ground) in which the human body (the figure) is located. Each item included one of the following emotions: alegría (joy), ansiedad (anxiety), enamoramiento (infatuation), esperanza (hope), felicidad (happiness), furia (fury), ira (anger), locura (madness), miedo (fear), serenidad (serenity), tristeza (sadness), verguenza (shame). Example (1) illustrates a sentence item in the two conditions (a full list is included in the appendix).

(1)

a. La oscuridad del monte se acentuaba por la falta de luz de luna; había en Elena un gran miedo. (tr. The darkness in the woods was accentuated by the lack of moonlight; there was great fear in Elena)

b. La oscuridad del monte se acentuaba por la falta de luz de luna; Elena se encontraba en un estado de gran miedo. (tr. The darkness in the woods was accentuated by the lack of moonlight; Elena was in a state of great fear)

Sentence 1a corresponds to the first condition (emotion-as-figure condition) while sentence 1b corresponds to the second condition (emotion-as-ground condition).

Procedure Two questionnaire lists with twelve experimental items were created. Questionnaires were randomized across participants and the two conditions were counterbalanced across lists so that each participant only saw one version of each item. The order of items were pseudo-randomized across participants to avoid order effects.

Participants were asked to rate the situation described in each sentence on a scale from 1 to 7. They were instructed to judge, for each situation described, the degree of control the person had on the emotion. The scale was defined so that 1 indicated "the person has very little control over their emotion" and 7 "the person has total control over their emotion". The questionnaire began with a page of instructions asking participants to make their judgments based on first impressions without reading each sentence more than once. In the instructions, participants were given two practice items. The practice included two emotions not used in the experimental items and had a different (not metaphorical) structure as the twelve experimental items. The questionnaire was presented on paper and the data was annotated manually.

Results of Study 1

One of the participants was removed from the sample because he did not complete the questionnaire. This was a within subject design (each subject was exposed to both conditions) and conditions were counterbalanced across items. Each participant was exposed to one version of the item only. Ratings of perceived control were higher in the emotion-as-figure condition (M = 3.86; SD= .86) than in the emotion-as-ground condition (M = 3.68; SD=.85), and the difference was significant both by subjects (two-tailed $t(107) = 2.36$; p=.02) and by items (two tailed $t(11) = 2.33$; p =.04).

As predicted, the results of Study 1 suggest that participants rated emotions as more controllable when they are described as figures in grounds compared to when they are described as grounds holding figures.

Study 2

A possible caveat of Study 1 is that emotion valence was not manipulated in a systematic manner. It could be the case that evaluation of emotion controllability differs significantly depending on whether the emotion is positive or negative in valence. Moreover, Study 1 materials included 5 positive emotions and 7 negative emotions, thus hindering fair comparisons between ratings per valence. Study 2 was designed to address this problem: intended as an improved replication of Study 1, the set of materials included 12 positive and 12 negative emotions. Additionally, items were rated for naturalness in an independent control experiment.

Study 2 followed a 2x2 design; Factor 1 Metaphor Framing (emotion-as-figure/ emotion-as-ground) and Factor 2 Valence of Emotion (positive/negative). Valence of Emotion was varied between subjects, and Metaphor Framing was varied within subjects as in Study 1.

Method

Participants A hundred and twenty four undergraduate students from Universidad de los Andes (Bogotá, Colombia) volunteered to participate in this study or received extra course credits in exchange of their participation. All participants were 18 years-old or older and declared that their native language was Spanish.

Materials The stimuli consisted of twelve sentences each describing a person experiencing a negative emotion and twelve sentences describing a person experiencing a positive emotion. Sentences describing positive and negative emotions were gathered in two different questionnaires. Each participant was presented with one of these two questionnaires. As in Study 1, there were two versions of each sentence, varying on the choice of metaphorical framing. In one condition, the emotion was framed as a

located figure in a container (the person´s body), while in the second condition it was metaphorically described as a ground in which the person is. The twelve negative emotions were: angustia (anguish), ansiedad (anxiety), decepción (disappointment), desconsuelo (desolation), desesperanza (hopelessness), entusiasmo (enthusiam), furia (fury), ira (anger), locura (madness), miedo (fear), resignación (abandonment), tristeza (sadness) verguenza (shame). The twelve positive emotions were the following: alegría (joy), alivio (relief), enamoramiento (infatuation), esperanza (hope), exaltación (excitation), felicidad (happiness), júbilo (rejoice), orgullo (pride), satisfacción (satisfaction), serenidad (serenity), tranquilidad (tranquility). A full list of items is included in the appendix.

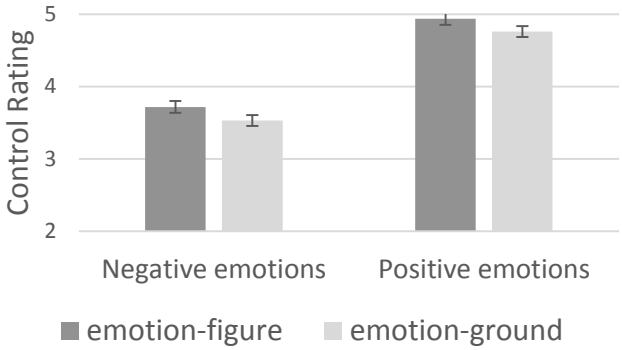
Additionally, Study 2 sentence items were controlled for naturalness of sentences in the two metaphor conditions using an independent sample. We collected naturalness ratings of the twenty four sentences in a separate test. One hundred and twelve participants (who did not participate in the experimental task) rated the naturalness of the 24 experimental sentences. Each participant saw one version of each sentence and conditions were counterbalanced across subjects. Participants rated each sentence's naturalness on a scale from 1 (not at all natural) to 7 (extremely natural). The sentences did not differ in their naturalness (M (emotion-as-figure) = 4.70, SD = 1.27, M (emotion-as-ground) = 4.74, SD = 1.28, t(111) = 0.83, p = .40).

Procedure Four questionnaire lists with twelve experimental items were created, two containing sentences featuring positive emotion and two containing sentences featuring negative emotions. Questionnaires were randomized across participants and the two conditions were counterbalanced across lists so that each participant only saw one version of each item. As in Study 1, the order of items were pseudo-randomized across participants to avoid order effects. Other aspects of the procedure and instructions were the same as in Study 1.

Results of Study 2

One of the participants was removed from the sample because he did not complete the questionnaire. As in Study 1, Metaphor Framing was a within-subject condition. Emotion valence was a between subject condition since each participant was exposed to one questionnaire containing either positive or negative emotions.

Results are shown in Figure 2. The main finding of Study 1 was replicated: ratings of perceived control over emotions were higher in the emotion-as-figure condition (M=4.33; SD=1.08) than in the emotion-as-ground condition (M= 4.15; SD=1.05). Also, ratings were higher for positive emotion (emotion-as-figure, M= 4.94; SD=0.93; emotion-as-ground, M=4.76; 0.94) than negative emotions (emotion-as-figure, M= 3.72; SD=0.87; emotion-as-ground, M=3.53; 0.80). A repeated measures ANOVA revealed that the effect of Metaphor-Framing condition was significant (F (1,121) = 8.97; p = 0.003; η^2= .07) and that the effect of Valence of Emotion was also significant (F (1,121) = 68.1; p <.0001; η^2 = .36). However, the interaction between Valence of Emotion and Metaphor Framing was not significant (F<.5; p=.9).



Taken together, the results of Study 1 and Study 2 suggest that participants rate emotions as more controllable when they are described as figures in grounds than when they are described as grounds. Additionally, people rate positive emotions as more controllable than negative, but this effect seems to be independent of the metaphorical mapping on figure/ground settings.

Discussion

The data suggests that the metaphorical framing of emotions affects the participants´ evaluation of degree of control. This is consistent with the proposal that the spatial relationships in figure/ground schemas are retained in abstract representations of emotion: when emotion is described as a located entity in the human body, the emotion is rated as more controllable.

One implication of these findings is related to the effect of the abstract use of prepositions on semantic structure. In line with studies of the abstract use of *in* and *on* in English (e.g., Jamrazik & Gentner, 2015), our data provides empirical evidence that the metaphorical use of *en* in Spanish preserves spatial meaning of functional relations such as control.

More generally, the data provides support to the view that perceptual schemas underlie the abstract concepts of emotion. This is consistent with accumulating work linking emotions to perceptual, image-like schemas (e.g., Crawford, Margolis, Drake & Murphy, 2006; Margolis & Crawford, 2008), adding to the evidence of the perceptual nature of emotion representation (Crawford et al, 2006; Fetterman, Robinson & Meier, 2012; Houser, Carter & Meier, 2009; Margolis & Crawford, 2008; Reali & Arciniegas, 2015; Richardson, Spivey, Barsalou & McRae, 2003).

Finally, this work supports the view that thought is shaped by language. It adds to the many papers showing that linguistic framing influences a range of cognitive domains from perception to social cognition (e.g., Landau, Meier & Keefer, 2010; Fausey & Boroditsky, 2010; Fausey & Boroditsky, 2011; Thibodeau & Boroditsky, 2011; Matlock, 2012).

In sum, different aspects of spatial meaning embedded in figure-ground relationships seem to be retained in the way

concepts of emotion are construed.

Acknowledgments

We thank Debora Shambo as well as other members from the Language and Cognition lab at Universidad de los Andes for their help on the data recollection process. This research was funded by the Fondo de Apoyo para Profesores Asistentes (FAPA), Vicerrectoría de Investigaciones, Universidad de Los Andes, Bogotá, Colombia.

References

Boers, F. (1996). Spatial prepositions and metaphor: A cognitive semantic journey along the up-down and the front-back dimensions. Tubingen, Germany: Gunter Narr Verlag.

Boroditsky, L. (2000). Metaphorical structuring: understanding special metaphors. *Cognition*, *75*(1), 1-28.

Boroditsky, L. & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science, 13*(2), 185-189.

Casasanto, D. (2010). Space for thinking. In B. Bergen, V. Evans & J. Zinken (Eds.), Language, cognition and space: State of the art and new directions (pp. 453–461). London: Equinox.

Crawford, L. E., Margolis, S. M., Drake, J. T., & Murphy, M. E. (2006). Affect biases memory of location: Evidence for spatial representation of affect. *Cognition & Emotion, 20*, 1153–1169.

Fausey, C. M., & Boroditsky, L. (2011).Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review, 18*, 150–157.

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review, 17(5),* 644–650.

Fausey, C. M., & Matlock, T. (2011). Can grammar win elections? *Political Psychology, 32,* 563–574.

Feist, M. I., & Gentner, D. (2003). Factors involved in the use of *in* and *on*. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the twenty-fifth annual meeting of the cognitive science society* (pp. 390–395). Hillsdale, NJ: Lawrence Erlbaum.

Fetterman, A.K., Robinson, M.D., & Meier, B.P. (2012). Anger as "seeing red": Evidence for a perceptual association. *Cognition & Emotion, 26*, 1445–1458.

Gentner, D. & Goldin-Meadow, S. (2003). Whither Whorf. In D. Gentner and S. Goldin-Meadow (Eds), *Language in mind: Advances in the study of language and thought*. Cambrige, MA: MIT Press.

Hauser, D. J., Carter, M. S., & Meier, B. P. (2009). Mellow Monday and furious Friday: The approach-related link between anger and time representation. *Cognition & Emotion, 23*, 1166–1180.

Jamrozik, A. & Gentner, D. (2015). Well-hidden regularities: Abstract uses of in and on retain an aspect of their spatial meaning. *Cognitive Science, 39,* 1881-1911.

Kövecses, Z. (2010). Metaphor and culture. *Acta Univesitatis Sapientiae Philologica, 2*, 197–220.

Lakoff, G. (1987). *Women, fire and dangerous things*. Chicago, IL: Chicago University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: Chicago University Press.

Landau, M. J., Meier, B., & Keefer, L. (2010). A metaphor-enriched social cognition. *Psychological Bulletin*, 136, 1045–1067.

Landau, M.J., Sullivan, D., & Greenberg, J. (2009). Evidence that self-relevant motivations and metaphoric framing interact to influence political and social issues. *Psychological Science, 20*, 1421–1427.

Lleras, M., Reali, F. & Alviar, C (2014). The metaphors we speak with affect how we think about time and space. In Bello P., Guarini, M., McShane M. & Scassellati B. (Eds.) *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1258-1264). Austin TX: Cognitive Science Society.

Margolis, S.O., & Crawford, L.E. (2008). Event valence and spatial metaphors of time. *Cognition & Emotion, 22*, 1401–1414.

Matlock, T. (2012). Framing political messages with grammar and metaphor. *American Scientist, 100,* 478–483.

Matlock, T., Sparks, D., Matthews, J. L., Hunter, J., & Huette, S. (2012). Smashing new results on aspectual framing: How people describe car accidents. *Studies in Language, 36*, 700–721.

Reali, F. & Arciniegas, C. (2015). Metaphorical conceptualization of emotion in Spanish: Two studies on the role of framing. *Metaphor and Social World*, 5(1), 20-42.

Reali, F., Soriano, T., & Rodríguez, D. (in press). How we think about depression: The role of metaphorical framing. To appear in *Revista Latinoamericana de Psicología*.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science, 27*, 767–780.

Thibodeau, P., & Boroditsky, L. (2011). Metaphors we think with: the role of metaphor in reasoning. PLoS ONE 6(2): e16782.

Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), Spatial orientation: Theory, research, and application. New York: Plenum Press.

APPENDIX

Experimental sentences used in Study 2. Those marked with a star (*) are the items used in Study 1.

emotion-as-ground condition:

X se encontraba en un estado de gran **Y**

X (person) was in a state of great **Y** (emotion)

emotion-as-figure condition:

Había un gran **Y** en **X**

There was great **Y** (emotion) in **X** (person)

Positive valence items

1. Las notas finales reflejaron su esfuerzo; María se encontraba en un estado de gran **satisfacción**/ había en María una gran **satisfacción**. X = María, Y= satisfacción (satisfaction)

2. La policía recuperó la mayoría de sus pertenencias; Alison se encontraba en un estado de gran **alivio**/ había en Alison un gran **alivio**. X = Alison, Y= alivio (relief)

3. Se acercaba su fiesta de cumpleaños por fin; Freddy se encontraba en un estado de gran **exaltación**/ había en Freddy una gran **exaltación**. X = Freddy, Y= exaltación (excitement)

*4. Llegaron buenas noticias por fin; Malena se encontraba en un estado de gran **alegría**/ había en Malena una gran **alegría**. X = Malena, Y= alegría (joy)

*5. Finalmente se confirmó el embarazo; Solange se encontraba en un estado de gran **felicidad**/ había en Solange una gran **felicidad.** X = Solange, Y= happiness (felicidad)

6. Su buen desempeño en la competencia rindió frutos; Mateo se encontraba en un estado de gran **orgullo**/ había en Mateo un gran **orgullo**. X = Mateo, Y= orgullo (pride)

7. Su difícil situación económica finalmente se solucionó; Gabriel se encontraba en un estado de gran **tranquilidad**/ había en Gabriel una gran **tranquilidad**. X = Gabriel, Y= tranquilidad (tranquility)

*8. Eran dos almas gemelas; Bernardo se encontraba en un estado de gran **enamoramiento**/ había en Bernardo un gran **enamoramiento**. X = Bernardo, Y= enamoramiento (infatuation)

*9. No se rendía frente a las circunstancias adversas; Selena se encontraba en un estado de gran **esperanza**/ había en Selena una gran **esperanza**. X = Selena, Y= esperanza (hope)

10. Se acercaban las vacaciones en la costa; Julián se encontraba en un estado de gran **entusiasmo**/ había en Julián un gran **entusiasmo**. X = Julián, Y= entusiasmo (enthusiasm)

11. Su hijo mayor le contó que se había comprometido; Ximena se encontraba en un estado de gran **júbilo**/ había en Ximena un gran **júbilo**. X = Ximena, Y= júbilo (rejoice)

*12. Tantos años de meditación valieron la pena; Ramiro se encontraba en un estado de gran **serenidad** había en Ramiro una gran **serenidad.** X = Ramiro, Y= serenidad (serenity)

Negative valence items

*1. Al subir al estrado se tropezó y se cayó frente al público; Alejandra se encontraba en un estado de gran **vergüenza**/ había en Alejandra una gran **vergüenza**. X= Alejandra, Y= vergüenza (shame)

*2. El delirio se comenzaba a manifestar en sus palabras; José se encontraba en un estado de gran **locura**/ había en José una gran **locura**. X=Alejandra, Y=locura (madness)

*3. La oscuridad del monte se acentuaba por la falta de luz de luna; Elena se encontraba en un estado de gran **miedo/** había en Elena un gran **miedo**. X= Elena, Y= miedo (fear)

4. Había sido víctima de un robo colosal; Roberto se encontraba en un estado de gran **desesperanza**/ había en Roberto una gran **desesperanza**. X=Roberto, Y=desesperanza (hopelessness)

5. Los resultados de sus exámenes de sangre no estaban listos; Lucía se encontraba en un estado de gran **angustia**/ había en Lucía una gran **angustia**. X=Lucía, Y=angustia (anguish)

*6. Faltaban cinco minutos para que salieran a luz los resultados del examen; Camilo se encontraba en un estado de gran **ansiedad**/ había en Camilo una gran **ansiedad**. X=Camilo, Y=ansiedad (anxiety)

*7. La muerte de su perro fue inesperada; Salomé se encontraba en un estado de gran **tristeza**/ había en Salomé una gran **tristeza**. X=Salomé, Y=tristeza (sadness)

8. Su novia mostraba desinterés frente a un posible futuro juntos; Jaime se encontraba en un estado de gran **decepción**/ había en Jaime una gran **decepción**. X=Jaime, Y=decepción (disappointment)

9. La falta de soluciones a sus problemas era evidente; Rocío se encontraba en un estado de gran **resignación**/ había en Rocío una gran **resignación** (abandonment).

*10. La infidelidad de su esposa salió a la luz; Marcos se encontraba en un estado de gran **furia**/ había en Marcos una gran **furia**. X=Marcos, Y=furia (fury)

*11. El fallo de la corte había sido injusto; Ximena se encontraba en un estado de gran **ira**/ había en Ximena una gran **ira**. X=Ximena, Y=ira (anger)

12. Sus enemigos lograron acabar con su compañía; Martín se encontraba en un estado de gran **desconsuelo/** había en Martín un gran **desconsuelo**. X=Martín, Y=desconsuelo (desolation)

Explanatory Judgment, Probability, and Abductive Inference

Matteo Colombo (m.colombo@tilburguniversity.edu)
Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Warandelaan 2, 5037 AB Tilburg
Tilburg, The Netherlands

Marie Postma-Nilsenová (m.nilsenova@tilburguniversity.edu)
Tilburg Center for Cognition and Communication (TiCC), Warandelaan 2, 5037 AB Tilburg
Tilburg, The Netherlands

Jan Sprenger (j.sprenger@tilburguniversity.edu)
Tilburg Center for Logic, Ethics and Philosophy of Science (TiLPS), Warandelaan 2, 5037 AB Tilburg
Tilburg, The Netherlands

Abstract

Abductive reasoning assigns special status to the explanatory power of a hypothesis. But how do people make explanatory judgments? Our study clarifies this issue by asking: (i) How does the explanatory power of a hypothesis cohere with other cognitive factors? (ii) How does probabilistic information affect explanatory judgments? In order to answer these questions, we conducted an experiment with 671 participants. Their task was to make judgments about a potentially explanatory hypothesis and its cognitive virtues. In the responses, we isolated three constructs: *Explanatory Value*, *Rational Acceptability*, and *Entailment*. Explanatory judgments strongly cohered with judgments of causal relevance and with a sense of understanding. Furthermore, we found that Explanatory Value was sensitive to manipulations of statistical relevance relations between hypothesis and evidence, but not to explicit information about the prior probability of the hypothesis. These results indicate that probabilistic information about statistical relevance is a strong determinant of Explanatory Value. More generally, our study suggests that abductive and probabilistic reasoning are two distinct modes of inference.

Keywords: Human Reasoning; Abduction; Explanatory Judgment; Explanatory Value; Probability.

Introduction

Explanatory judgments are central to abductive reasoning, that is: inferring to the available hypothesis that can best explain the evidence. Understanding what determines the explanatory power of a hypothesis is of crucial importance for researchers in different fields: for cognitive psychologists, who study the principles of human reasoning (e.g., Oaksford & Chater 2000); AI researchers, who use abduction for belief revision, knowledge representation and fault diagnosis (e.g., Paul 2000); for philosophers of science and epistemologists, who investigate the rationality of abductive inferences, and their relation to probabilistic inference (Lipton 2004; Douven 2011).

Expanding on a growing body of literature in the philosophy and cognitive psychology of explanation, we investigated (i) how explanatory power relates to other features of a hypothesis with respect to an explanandum, such the hypothesis' degree of confirmation, its causal relevance, logical relations, and the sense of understanding it provides; and (ii) how probabilistic information affects explanatory judgment.

Regarding the first question, empirical research in cognitive and developmental psychology has shown that explanation plays several intertwined roles in human cognition (Lombrozo 2011). Explanatory relationships, particularly causal relationships, guide categorization (Carey 1985, 201 ff; Lombrozo 2009), support inductive inference and learning (Holyoak & Cheng 2011; Legare & Lombrozo 2014), and calibrate metacognitive strategies involved in problem-solving (Chi, De Leeuw, Chiu, & LaVancher 1994).

Explanatory considerations can also contribute to the credibility of a hypothesis. Koehler (1991) reviewed much of the work on how explanation influences subjective probabilities, and argued that merely focusing on a hypothesis as if it were the true explanation of observed data is sufficient to boost the subjective probability assigned to that hypothesis. Explanation has also been demonstrated to influence how probabilities are assigned to one proposition in the light of another. Sloman (1994) found that a proposition boosted the probability assigned to another proposition if they shared an explanation.

In particular, cognitive virtues such as the simplicity (Lombrozo 2007; Bonawitz & Lombrozo 2012) and generalizability (Preston & Epley 2005) of a potentially explanatory hypothesis have been shown to be plausible determinants of explanatory judgment. For example, Bonawitz & Lombrozo (2012) and Lombrozo (2007) quantified the simplicity of an explanation in terms of the number of causes it cites, and provided pre-school and adult participants with information about the base rate of each cause. They found that both children and adults relied on the simplicity of a hypothesis as a cue commensurate to base rate information in the face of uncertainty. These results indicate that explanatory judgment is a function of both simplicity and probabilistic information.

While simplicity, and generalizability may be two determinants of explanation, the *sense of understanding* bestowed by a hypothesis seems to be a robust outcome of good explanations. Rozenblit & Keil (2002), for example, considered the phenomenon of the illusion of explanatory depth, whereby people believe they understand the world in greater detail, coherence, and depth than they actually do. A sense of understanding was found to be significantly stronger for explanatory knowledge relative to other knowledge domains. Along similar lines, the connection between inference and explana-

tory understanding has been criticized by Trout (2002), who argues that a sense of understanding is frequently deceptive, produced by overconfidence and hindsight bias, and in any case not a good indicator for an actually valuable explanation.

In this body of literature in cognitive psychology, explanations are typically presented as *causal* explanations. This choice coheres with philosophical literature on explanation that stresses that good explanations have to provide information about the causes that bring about the explanandum phenomenon (e.g., Salmon 1971/1984; Woodward 2003; Strevens 2008). However, other philosophical models, such as the covering-law model of explanation, focus on the logical relations between explanans and explanandum, such as logical implication (Hempel 1965). In order to have a better understanding of what determines explanatory judgments, it is therefore natural to compare them to judgments of causality, logical implication, and probability.

Regarding our second question about the relationship between probability and explanation, much of the attention has concentrated on the relationship between abductive and Bayesian inference. For example, van Fraassen (1989) argued that abductive and probabilistic inference are incompatible. If we believe more strongly in good explanations than warranted by their posterior probability, we will violate Bayes' rule and be subject to probabilistic incoherence. However, if abductive inferences have to agree with probabilistic inferences as governed by the axioms of probability and by Bayes' Theorem, it is unclear whether they have independent significance, or whether they are just redundant with respect to probabilistic inference.

To address van Fraassen's challenge, many authors have maintained that explanatory judgments *guide* probabilistic inference, or act as a heuristics for them. For instance, Lipton (2004) argued that explanatory judgments may give a good *descriptive* account of our inferential practices, or help to determine the ingredients of probabilistic inferences (e.g., Henderson 2014). Other authors have turned their attention to explicating explanatory power in probabilistic terms (McGrew 2003; Schupbach & Sprenger 2011; Crupi & Tentori 2012), which could help to better explore whether and how abductive reasoning can be embedded into an overarching probabilistic framework.

Notably, a substantial part of the aforementioned studies and recent literature in the psychology of explanation (Keil & Wilson 2000; Lombrozo 2011, 2012) has been concerned with the question of how explanatory power determines probability judgments. Instead, the reverse question has not gained as much attention. With our study, we hope to advance understanding of how probabilistic information determines explanatory judgments.

We constructed vignettes where participants were given information about the priors and likelihoods of a potential explanation, and were asked to make judgments on its explanatory power, posterior probability, and acceptability, as well as on its logical and causal relation to the explanandum. After eliciting these judgments, we used a principal component analysis to examine how they cluster together. We found that a small set of constructs, which we dubbed *Explanatory Value*, *Rational Acceptability*, and *Entailment*, accounted for most of the variation in the participants' judgments, and that explicit information about statistical relevance relations, but not about base rates, determined judgments of the **Explanatory Value** of a hypothesis. This result suggests that abductive and probabilistic reasoning are distinct modes of inference, and that abduction is often probabilistically incoherent.

In what follows, we first describe the methods we used in the experiment designed to collect our data. We then summarize our findings and put them into a broader perspective.

Experiment

Methods

Participants 744 students at Tilburg University (The Netherlands) participated in our study, and 671 completed it (383 male, M_{age} = 21.5 (SD = 2.3)). They were randomly assigned to one version of an experimental vignette. Participants were all proficient English speakers, and participated on a voluntary basis, either for free or in exchange for course credit.

Design and Material We employed three distinct types of vignettes with the same logical structure, but different content. Participants were presented with one version of a vignette, where two possible events were related to two possible explanations for that event.

> **Vignette 1:** *There are two urns on the table. Urn A contains 67% white and 33% black balls, Urn B contains only white balls. One of these urns is selected. You don't know which urn is selected, but you know that the chance that Urn A is selected is 25%, and that the chance that Urn B is selected is 75%. From the selected urn a white ball is taken at random.*
> *Please now consider the hypothesis that Urn A has been chosen.*

Vignette 2 and 3 had more concerete content, closer to cases of ordinary reasoning.

> **Vignette 2:** *Again and again, Ruud has knee problems when playing football. The doctors give him two options: knee surgery or a conservative treatment. If Ruud chooses to go into surgery, he cannot play football for half a year; if he chooses the conservative treatment, there is a 33% chance that he can play again after one month; otherwise (with a chance of 67%) he has to rest longer. You don't know which option Ruud chooses, but you believe that the chance that he chooses surgery is 75%—and that the chance that he chooses the conservative treatment is 25%. A month later a joint friend tells you that Ruud is still unable to play football.*

Please now consider the hypothesis that Ruud has chosen for the conservative treatment.

Vignette 3: *Louise arrives by train in Twin City. Twin city has two districts: West Bank and East Bank. In West Bank, there is only one taxi company, namely Green Taxi Ltd., and all their cabs are green. Green Taxi Ltd. also owns 67% of all cabs in East Bank. The other cabs in East Bank are owned by The Red Taxi Inc., all their cabs are red. Louise does not know which part of the city the train is entering, but judging from her knowledge of Twin City she assumes that there is a 75% chance that she is in West Bank (and a chance of 25% that she is in East Bank). At some point, Louise sees a green cab from the train.*
Please now consider the hypothesis that Louise is in East Bank.

After reading the vignette, participants assessed seven items (the construct names in italics were not provided to the participants) on a Likert scale ranking from 1 ("do not agree at all") to 7 ("fully agree"). For Vignette 1, the seven items read as follows:

Logical Implication The hypothesis logically implies that a white ball has been taken out.

Causal Relevance The hypothesis specifies the cause that a white ball has been taken out.

Confirmation The hypothesis is confirmed by the fact that a white ball has been taken out.

Posterior Probability The hypothesis is probable given that a white ball has been taken out.

Explanatory Power The hypothesis explains that a white ball has been taken out.

Understanding The hypothesis provides understanding why a white ball has been taken out.

Truth The hypothesis is true.

The distinction between explanatory power and explanatory value just serves to keep apart the name of a response variable and a broader cognitive factor associated with a hypothesis.

After filling in this questionnaire, the participants could provide information about the way they made their judgments, and we collected some demographic data.

Vignettes were, in a between-subjects design, varied in two dimensions, corresponding to two independent variables:

IV 1: Statistical Relevance The degree of statistical relevance between the explanans and the explanandum, with the four values "strong/weak disconfirm" and "strong/weak confirm".

IV 2: Prior Probability The prior probability of the hypothesis under consideration (.25, .5, or .75).

For all three types of vignettes, all possible $4 \times 3 = 12$ combinations of the values of these variables were realized in the experiment. For example, in the case of Vignette 1, Statistical relevance was manipulated by changing the color of the ball drawn from the urn and/or by changing the explanatory hypothesis (from Urn A to Urn B), leading to four different conditions ordered according to the degree of confirmation they provide.

Procedure Participants completed the questionnaire on a university PC or their own computer in the digital environment of LimeSurvey installed on a local server. The use of LimeSurvey guaranteed that the data could be protected and provided with a time stamp, and information about the IP address of the respondent. The experiment was self-paced, and took on average approximately 10 minutes to complete.

Results

Prior to the analysis of the effects of vignette manipulation, we explored the interdependencies of the seven items in the response questionnaire. To recall, the participants were asked to judge several features of the hypothesis with respect to the evidence: logical implication, causal relevance, explanatory power, increase in understanding, confirmation, posterior probability and truth. By analyzing the interdependencies with the help of the Pearson zero-order correlation coefficient, we determined whether some of these seven concepts tap on the same dimension and could be identified with each other.

Table 1: Zero-order correlations for 7 items (N = 671), all correlations with $p < .01$.

	1	2	3	4	5	6	7
1. *Log. Implication*	-	.38	.22	.32	.46	.30	.12
2. *Causality*		-	.45	.39	.56	.63	.37
3. *Confirmation*			-	.56	.35	.47	.63
4. *Post Probability*				-	.37	.51	.46
5. *Explan. Power*					-	.60	.28
6. *Understanding*						-	.36
7. *Truth*							-

The correlations are presented in Table 1. The analysis revealed that all of the variables correlated at least with .3 with several other variables, but at most .63. These medium-sized correlations show that the participants did not conflate cognate concepts (e.g., causal relevance and explanatory power) with each other, which would be reflected in correlation coefficients greater than .7. At the same time, the response variables were sufficiently related to each other to motivate a Principal Component Analysis: that is, a decomposition of the seven response variables into constructs that could account for most of the variation in the data.

Principal Component Analysis The factorability of the 7 items was examined with a Principle Component Analysis (PCA). The Kaiser-Meyer-Olkin measure of sampling adequacy was .82 and the Bartlett's test of sphericity was significant ($\chi^2(21) = 1790.77$, $p < .0001$). The initial eigenvalues showed 51% of variance explained by the first factor, 16% explained by the second factor, and 10% explained by the third factor. A visual inspection of the scree plot revealed a 'leveling off' of eigenvalues after the three factors, therefore, a three factor solution using the oblique rotation was conducted, with the three factors explaining 77% of the variance. All items had primary loadings over .7 Table 2 presents the factor loading matrix (loadings under .30 suppressed). In the remainder, we restrict our analysis to these three factors.

Table 2: Factor loadings and communalities based on a principle component analysis with oblimin rotation for 7 items (N = 671).

	1	2	3	Communality
Log. Implication			.94	.94
Causality	.86			.74
Confirmation		-.84		.77
Post Probability		-.72		.67
Explan. Power	.81			.73
Understanding	.87			.78
Truth		-.88		.75

The names for these factors were derived from the clustering shown in Table 2. Factor 1, **Explanatory Value**, clustered explanatory power together with related features of a hypothesis, such as causality and enhancement of understanding (Dieks & DeRegt, 2005; Strevens 2008). Factor 2, **Rational Acceptability**, captured those features that hang together with the acceptability of a hypothesis: probability, confirmation, and truth. The strong correlations between these features were not surprising: confirmation raises posterior probability, which is in turn an indicator of the truth of a theory. Finally, Factor 3 captured the strength of the logical relation between hypothesis and evidence. Since no other response variable was loaded on this factor, it figured as **Entailment**, showing the link to the response variable Logical Implication.[1]

Tests of Experimental Manipulation We conducted two analyses of variance (ANOVAs) to test the effects of the independent variables, Statistical Relevance and Prior Probability, on Explanatory Value, Rational Acceptability, and Entailment, respectively.[2]

Table 3: Estimated Marginal Means and SE of Explanatory Value by Statistical Relevance and Prior Probability (N = 671). SD = Strong Disconfirm, WD = Weak Disconfirm, WC = Weak Confirm, SC = Strong Confirm.

Explanatory Value				
	Statistical Relevance			
Prior	*SD*	*WD*	*WC*	*SC*
Low	2.02 (.22)	3.02 (.21)	4.72 (.20)	4.81 (.19)
Medium	1.95 (.21)	3.05 (.20)	4.66 (.21)	4.58 (.20)
High	1.88 (.24)	3.12 (.21)	4.63 (.22)	4.64 (.20)

First, we tested the effects of the experimental manipulation on *Explanatory Value*. There was a significant main effect of *Statistical Relevance*, $F(3,659) = 118.53$, $p < .001$, $\eta^2_p = .35$, but no effect of *Prior Probability*, $F(2,659) = 0.20$, $p = .822$. There was no interaction effect between *Statistical Relevance* and *Prior Probability*, $F(6,659) = 0.12$, $p = .994$—see Table 3 for the descriptives. A pair-wise comparison with Bonferroni correction of the levels of *Statistical Relevance* showed a significant difference between all levels ($p < .001$), with the exception of Weak and Strong Confirm.

Table 4: Estimated Marginal Means and SE of Rational Acceptability by Statistical Relevance and Prior Probability (N = 671). Abbreviations like in Table 3.

Rational Acceptability				
	Statistical Relevance			
Prior	*SD*	*WD*	*WC*	*SC*
Low	1.93 (.18)	2.93 (.18)	3.45 (.17)	5.39 (.16)
Medium	2.05 (.18)	3.10 (.17)	3.74 (.18)	5.62 (.17)
High	1.95 (.20)	3.20 (.18)	3.48 (.18)	5.82 (.17)

Second, we examined the effects of the experimental manipulation on *Rational Acceptability*. There was again a significant main effect of *Statistical Relevance*, $F(3,659) = 223.76$, $p < .001$, $\eta^2_p = .51$, but no effect of *Prior Probability*, $F(2,659) = 1.68$, $p = .188$. There was no interaction effect between *Statistical Relevance* and *Prior Probability*, $F(6,659) = 0.81$, $p = .463$—see Table 4 for the descriptives. A pair-wise comparison with Bonferroni correction of the levels of *Statistical Relevance* showed a significant difference between all levels ($p < .001$), with the exception of Weak Disconfirmation and Weak Confirmation ($p = .005$).

[1] The internal consistency for two of the three scales (the third scale only consisted of one item) was examined using Cronbach's alpha, resulting in alpha .82 for Factor 1 and .79 for Factor 2. Composite scores were calculated for each of the three factors using the mean of the items with primary loadings on each factor. The descriptive values for the newly constructed scales were $M = 3.65$, $SD = 1.91$ for Explanatory Value, $M = 3.63$, $SD = 1.87$ for Rational Acceptability, and $M = 3.75$, $SD = 2.40$ for Logical Implication.

[2] A prior analysis of the effect of the vignette on the three dependent variables revealed that Explanatory Value (but not Rational Acceptability and Entailment) was also affected by the vignette manipulation. For clarity of exposition, the statistics is not included here because the size and direction of the significant main effects and the interaction effect remained the same when vignette was included as a factor.

Table 5: Estimated Marginal Means and SE of Entailment by Statistical Relevance and Prior Probability ($N = 671$). Abbreviations like in Table 3.

Entailment				
	Statistical Relevance			
Prior	*SD*	*WD*	*WC*	*SC*
Low	2.37 (.27)	2.98 (.26)	5.92 (.26)	4.14 (.25)
Medium	2.33 (.26)	3.00 (.26)	5.93 (.26)	3.69 (.26)
High	1.88 (.30)	3.09 (.27)	5.92 (.27)	3.32 (.26)

Finally, we analyzed the effects of the experimental manipulation on *Entailment*. Similarly to the previous two dependent variables, there was again a significant main effect of *Statistical Relevance*, $F(3,659) = 105.40$, $p < .001$, $\eta^2_p = .32$, but no effect of *Prior Probability*, $F(2,659) = 1.23$, $p = .292$. There was no interaction effect between *Statistical Relevance* and *Prior Probability*, $F(6,659) = 0.76$, $p = .598$—see Table 5 for the descriptives. A pair-wise comparison with Bonferroni correction of the levels of *Statistical Relevance* showed a significant difference between all degrees ($p < .001$) with the exception of Weak Confirm and Weak Disconfirm ($p = .007$). Since Entailment measures the impact of the hypothesis on the explanandum rather than vice versa, it is logical that the order deviates from the previous two tables, with Weak Confirm obtaining the highest score.

Discussion

Our study investigated the coherence of judgments of explanatory power with other cognitive factors that affect the evaluation of a hypothesis in token explanations of singular events. With the help of a principal component analysis of participants' responses, we identified several constructs under which the response variables could be subsumed. We also investigated how probabilistic information affected these constructs, and *Explanatory Value* (the most salient construct) in particular. We observed a neat distinction between judgments of Explanatory Value and posterior probability. More generally, participants' judgments on the seven response variables (i.e., Logical Implication, Causality, Explanatory Value, Understanding, Posterior Probability, Confirmation, and Truth) were aligned along three dimensions: *Explanatory Value* (loaded with the response variables Causality, Explanatory Power, and Understanding), *Rational Acceptability* (Posterior Probability, Confirmation and Truth) and *Entailment* (Logical Implication). Participants' judgments on these three factors were strongly affected by changes in statistical relevance, specifically by manipulations of the likelihood of the target hypothesis. Instead, the prior probabilities of the candidate explanatory hypothesis, presented as objective base rates, affected participants' judgments in none of those three dimensions. These results suggest a number of conclusions, and further questions for research.

First of all, explanatory value is a distinct feature of a hypothesis that relates to several other processes such as deductive and inductive reasoning, causal reasoning, and reasoning about truth (Lombrozo 2012). Isolating "causality", "understanding" and "explanatory power" ($\rightarrow$ Explanatory Value) from "confirmation", "posterior probability" and "truth" ($\rightarrow$ Rational Acceptability), our principal component analysis coheres with previous results about the tight connections between explanatory power, causality, and a sense of understanding (Lipton 2004; Keil 2006 Lombrozo 2007; Trout 2002). It also suggests that reasoners distinguish between the concepts of *Explanatory Value*, *Rational Acceptability*, and *Logical Entailment*. However, the number of components behind *Explanatory Value* is obviously limited by the number of our dependent variables. One question for future research is whether *Explanatory Value* presents the same components when other dependent variables are added, or when vignettes with different content or structure are evaluated.

Our results bear on the relationship between abduction and probabilistic inference in at least two ways. On the one hand, they show that the rational acceptability of a hypothesis does not always track its explanatory value. This indicates that abduction cannot simply be reduced to Bayesian inference, at least not from a descriptive point of view. On the other hand, *Explanatory Value* is strongly determined by explicit, numeric information about likelihoods. This finding is consistent with the well-documented phenomenon of base rate neglect (Tversky& Kahneman 1982). It also indicates that in situations where causal detail is kept sparse and the explanandum corresponds to a singular token event, abductive and probabilistically reasoning may pull into different directions: the former is insensitive to information about base rates. This conclusion presents a challenge for those who want to defend the rationality of abductive inference as a way of coming closer to the truth. However, when base rates are presented in a different format, reasoners often take them into account (Gigerenzer& Hoffrage 1995), and it remains an open question whether in those circumstances, explanatory inference can be fully described in probabilistic terms. To explore this possibility, further empirical research is needed about the determinants of explanatory value across a broader range of situations, as well as further theoretical explications of explanatory power (Schupbach 2011; Pacer, Lombrozo, Griffiths, Williams, & Chen 2013). We believe that this exchange between normative theorizing about the nature of explanation and empirical evidence about the psychology of explanatory judgment is the key to our understanding of explanatory judgments and abductive reasoning (Colombo 2016).

Acknowledgments

The project was financially supported by the Netherlands Organisation for Scientific Research (NWO) under Vidi grant #276-20-023 (J.S), by the European Research Council (ERC) under Starting Investigator Grant #640638 (J.S.) and by the Deutsche Forschungsgemeinschaft (DFG) in the Priority Program 1516 "New Frameworks of Rationality" (M.C., J.S.). The authors wish to thank Leandra Bucher, Vincenzo Crupi,

Tania Lombrozo, Henrik Singmann, and the members of the aforementioned DFG program for helpful discussion and feedback.

References

Bonawitz, E.B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology, 48,* 1156.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Chi, M.T., De Leeuw, N., Chiu, M.H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science, 18,* 439–477.

Colombo, M. (2016). Experimental Philosophy of Explanation Rising. The case for a plurality of concepts of explanation. *Cognitive Science.* doi: 10.1111/cogs.12340

Crupi, V., & Tentori, K. (2012). A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science, 79,* 365–385.

De Regt, H., & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. *Synthese, 144,* 137–170.

Douven, I. (2011). Abduction. *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/abduction/>.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102,* 684–704.

Hempel, C.G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Free Press.

Henderson, L. (2014). Bayesianism and Inference to the Best Explanation. *British Journal for the Philosophy of Science, 65,* 687–715.

Holyoak, K.J., & Cheng, P.W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62,* 135–163.

Keil, F.C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227–254.

Keil, F.C., & Wilson, R.A. (2000). *Explanation and Cognition.* Cambridge, MA: MIT Press.

Koehler, D.J. (1991). Explanation, Imagination, and Confidence in Judgment. *Psychological Bulletin, 110,* 499–519.

Legare, C.H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of experimental child psychology, 126,* 198–212.

Lipton, P. (2004). *Inference to the Best Explanation.* Second edition. London: Routledge.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55,* 232–257.

Lombrozo, T. (2009). Explanation and categorization: How "why" informs "what?". *Cognition, 110,* 248–253.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6,* 539–551.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University.

McGrew, T. (2003). Confirmation, Heuristics, and Explanatory Reasoning. *British Journal for the Philosophy of Science, 54,* 553–567.

Oaksford, M., & N. Chater (2000). *Bayesian Rationality.* Oxford: Oxford University Press.

Pacer, M., Lombrozo, T., Griffiths, T., Williams, J., & Chen, X. 2013. Evaluating computational models of explanation using human judgments. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pp. 498–507.

Preston, J., & Epley, N. (2005). Explanations Versus Applications: The Explanatory Power of Valuable Beliefs. *Psychological Science, 16,* 826–832.

Rozenblit, L. R., & Keil, F. C. (2006). The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth. *Cognitive Science, 26,* 521–562.

Salmon, W. (1971/1984). Statistical Explanation. Reprinted in Salmon (1984): *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schupbach J. (2011). Comparing Probabilistic Measures of Explanatory Power. *Philosophy of Science, 78,* 813–829.

Schupbach, J., & Sprenger J. (2011). The Logic of Explanatory Power. *Philosophy of Science, 78,* 105–127.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52,* 1–21.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge/MA: Harvard University Press.

Trout, J.D. (2002). Scientific Explanation and the Sense of Understanding. *Philosophy of Science, 69,* 212–233.

Van Fraassen, B.C. (1989). Laws and Symmetry. New York: Oxford University Press.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

The impact of biased hypothesis generation on self-directed learning

Doug Markant (markant@mpib-berlin.mpg.de)
Center for Adaptive Rationality
Max Planck Institute for Human Development
Lentzeallee 94, Berlin, Germany

Abstract

Self-directed learning confers a number of advantages relative to passive observation, including the ability to test hypotheses rather than learn from data generated by the environment. However, it remains unclear to what extent self-directed learning is constrained by basic cognitive processes and how those limits are related to the structure of the to-be-learned material. The present study examined how hypothesis generation affects the success of self-directed learning of categorical rules. Two experiments manipulated the hypothesis generation process and assessed its impact on the ability to learn 1D and 2D rules. Performance was strongly influenced by whether the stimulus representation facilitated the generation of hypotheses consistent with the target rule. Broadly speaking, the findings suggest that the opportunity to actively gather information is not enough to guarantee successful learning, and that the efficacy of self-directed learning closely depends on how hypothesis generation is shaped by the structure of the learning environment.
Keywords: self-directed learning, category learning, active learning, information search, hypothesis generation

Self-directed learning (SDL) is typically characterized by an interaction between external and internal search processes. Active information collection and exploration of the external environment are hallmarks of SDL (e.g., a student deciding how to study or playing with a new toy to learn how it works). Information resulting from this external search then influences an internal belief updating process, often conceptualized as the sequential generation and evaluation of new hypotheses, which in turn drives subsequent information gathering. This ongoing interaction is central to theories of conceptual discovery, including scientific inquiry (Klahr & Dunbar, 1988), explanatory reasoning (Johnson & Krems, 2001), and sensemaking (Weick, 1995).

Although previous research has described a number of benefits of SDL relative to passive learning conditions, it is less clear how its efficacy is constrained by basic cognitive and perceptual processes (Gureckis & Markant, 2012). Previous work suggests that, in particular, SDL may be strongly limited by failures or biases in the hypothesis generation process. For instance, in concept learning tasks people make effective information search decisions when given a set of hypotheses to discriminate, but often fail to generate alternative hypotheses on their own (Tweney et al., 1980). Biased hypothesis generation may be especially impactful in real-world domains involving large or ill-defined hypothesis spaces. In education, some researchers have argued against pedagogical practices which emphasize self-directed discovery because students often fail to generate the target concept in the absence of guidance that narrows the hypothesis space (Mayer, 2004). Thus, understanding the hypothesis generation process is critical to being able to predict whether SDL will be effective depending on the nature of the learning problem.

Self-directed category learning

A recent study by Markant and Gureckis (2014) suggests that, indeed, the benefits of SDL depend closely on the structure of the target concept. The authors compared passive *reception* and self-directed *selection* in a perceptual category learning task (see stimuli in Figure 1, left), in which the target classification rule was either one-dimensional (1D, a criterion on a single feature dimension) or two-dimensional (2D, requiring the integration of both feature values). Reception-based learners observed stimuli generated from bivariate normal distributions corresponding to each category, whereas selection-based learners chose items to learn about by specifying the feature values. Selection led to higher accuracy than reception among those participants learning 1D rules. In contrast, in the 2D case selection was no better than reception, and participants in both conditions were well-described as responding according to simpler 1D hypotheses.

The goal of the present study was to examine whether this divergence between selection-based learning of 1D and 2D rules was caused by a biased hypothesis generation process. The experimental manipulations were motivated by an account of hypothesis generation as involving the sampling of salient cues and relations in the environment (Trabasso, Bower, & Gelman, 1968; Wallach, 1962). Accordingly, the success of SDL should depend on whether the feature representation facilitates the generation of hypotheses that are of the same form as the target rule. Poor learning of 2D rules in Markant and Gureckis (2014) may have been driven by stimuli with highly distinct dimensions that encouraged the generation of 1D hypotheses, rather than 2D hypotheses that involve integrating two dimensions. In contrast to existing theories of category learning that assume inherent differences in the difficulty of acquiring 1D and 2D rules (e.g., Ashby, Paul, & Maddox, 2011), under this proposal the ability to learn either type of rule through SDL depends on whether the feature representation encourages the generation of hypotheses of the same form.

Experiment 1

The first experiment examined the effect of perceptual feature representations on self-directed category learning. People were predicted to be more likely to generate 1D hypotheses when stimulus features are highly separable or it is difficult to combine information about their relative magnitudes, a bias that should facilitate the learning of 1D target rules.

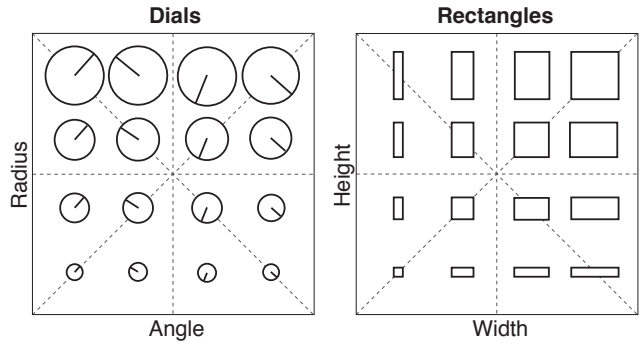


Figure 1: Example stimuli in the DIAL (left) and RECT (right) conditions in Experiment 1. Dotted lines indicate the set of target classification rules. Axis-aligned lines correspond to 1D rules, diagonal lines correspond to 2D rules.

Conversely, people may be more likely to generate 2D hypotheses when the two feature values are easily compared or combined, leading to faster learning of 2D target rules. In both cases, a mismatch between the feature representation and the form of the target rule (e.g., highly separable dimensions when learning a 2D rule) should lead to a longer internal search process and slower learning.

The task involved classifying shapes defined by two continuous feature dimensions into two categories, *A* and *B* (Figure 1). Each participant was assigned to learn either a 1D or 2D rule involving one of two types of stimuli, *dials* or *rectangles*. Dials were composed of a circle that varied in radius and a line segment that varied in orientation, whereas rectangles were defined in terms of their height and width. Participants learned to categorize stimuli through self-directed selection, such that on each trial they chose a combination of features and observed the true category label for that stimulus.

Previous studies involving dial stimuli have shown that the two feature dimensions (radius and angle) are highly separable (Shepard, 1964), whereas rectangle stimuli are perceived in an integral manner, with the dimensions of shape and size more salient than width and height (Macmillan & Ornstein, 1998). Based on this perceptual distinction, participants in the RECT conditions were expected to be more likely to generate 2D hypotheses that involved integrating both features, and as a result, to learn 2D rules more effectively than when stimuli are represented as dials.

Participants and materials

One-hundred twelve participants were recruited from the MPI subject pool (66 female, 44 male, 3 no gender given; $M_{age} = 25.1$, $SD = 5.2$) and were paid 7€ for participation in addition to a bonus based on their performance. Participants were randomly assigned to one of four conditions: 1D-DIAL, 1D-RECT, 2D-DIAL, or 2D-RECT. In addition, they were randomly assigned one of four possible variants of the target rule corresponding to rotations of the rule in the stimulus space.

Test items corresponded to a grid of 256 stimuli that uniformly tiled the feature space. This grid was partitioned into 8 blocks, each of which included 8 stimuli from every quadrant of the stimulus space, for a total of 32 items in each test block. The order of test items within each block and the order of the eight test blocks were randomized for each participant.

Procedure

The experiment was presented in a web browser using the *psiTurk* software (Gureckis et al., 2015). There were 8 blocks in total, each comprised of 16 learning trials followed by 32 test trials. Participants received .02€ for every correct classification during test trials, for a total possible bonus of 5.12€.

Learning trials On each learning trial a stimulus was randomly generated (with feature values drawn from a uniform distribution over their respective ranges) and displayed onscreen. The participant could then simultaneously vary the features by moving the mouse, with vertical mouse movement controlling one feature dimension and horizontal movement controlling the second feature dimension. The mapping between mouse directions and feature dimensions was randomized for each participant. After adjusting the feature values as desired, the participant pressed the spacebar to query the selected stimulus, after which the category label (*A* or *B*) was displayed until the participant pressed a button to complete the trial.

Test trials In each test trial a single test item was displayed at the center of the display. The participant categorized the item by pressing *A* or *B* on the keyboard at their own pace. No feedback was provided during test trials. At the conclusion of each test block, participants were told the proportion of items they classified correctly during that block.

Results

Classification accuracy Accuracy across blocks in shown in Figure 2A. A 2×2 ANOVA on overall classification accuracy (averaged over blocks and rule variants) was performed with stimulus (DIAL vs. RECT), rule (1D vs. 2D), and stimulus $\times$ rule interaction as factors. There was no main effect of stimulus type ($F(1,108) = 2.4$, $p = .12$), but there were significant effects of rule type ($F(1,108) = 22.6$, $p < .001$) and

Table 1: Classification accuracy.

		Pairwise comparisons (Tukey HSD)		
Experiment 1	*M* (*SD*)	1D-RECT	2D-DIAL	2D-RECT
1D-DIAL	.94 (.04)	* .13 [.07, .19]	* .23 [.17, .29]	* .06 [.01, .12]
1D-RECT	.81 (.08)		* .09 [.03, .15]	* .07 [.02, .13]
2D-DIAL	.71 (.09)			* .17 [.11, .23]
2D-RECT	.88 (.09)			
Experiment 2	*M* (*SD*)	1D-REL	2D-ABS	2D-REL
1D-ABS	.95 (.06)	* .08 [.02, .15)]	* .18 [.12, .25]	* .08 [.02, .16]
1D-REL	.86 (.11)		* .09 [.03, .17]	.00 [–.07, .06]
2D-ABS	.76 (.10)			* .09 [.02, .16]
2D-REL	.86 (.12)			

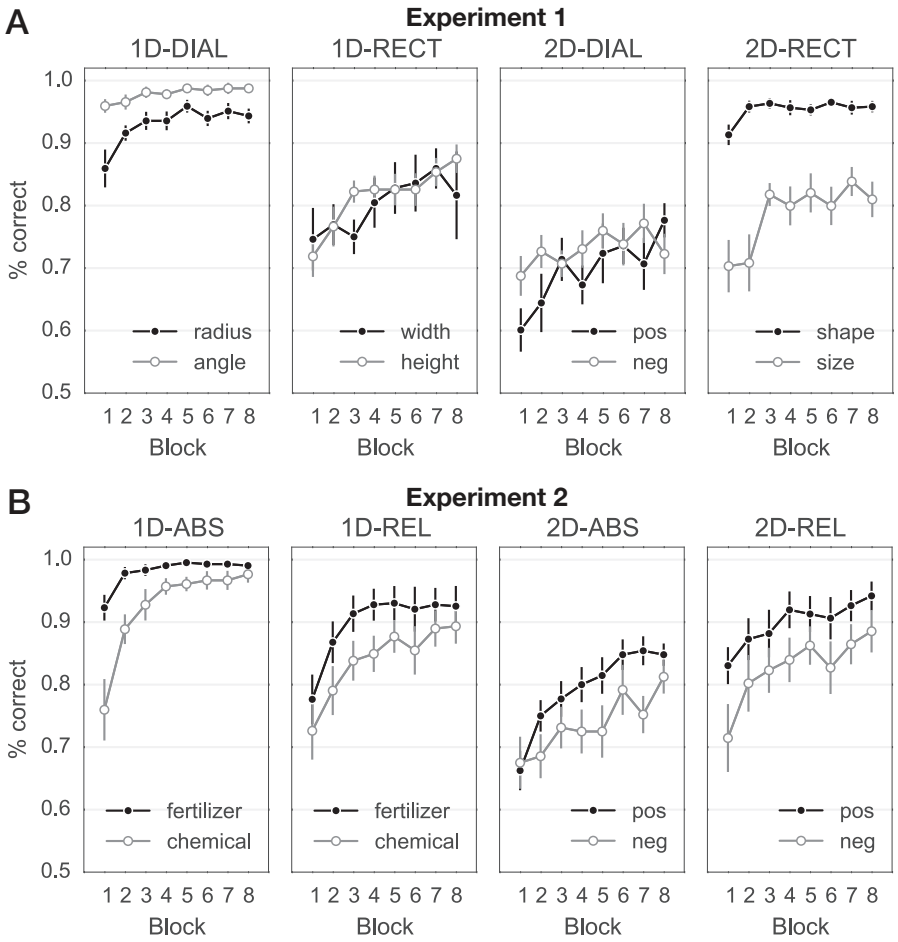


Figure 2: Classification accuracy in Experiment 1 (A) and Experiment 2 (B). Error bars indicate standard errors.

the stimulus $\times$ rule interaction ($F(1,108) = 94.4$, $p < .001$). Tukey HSD tests indicated significant pairwise differences in accuracy between all four conditions (see Table 1).

Two-sample t-tests were used to assess effects of rule variants within each condition. For 1D rules, participants were grouped according to the relevant feature dimension (i.e., radius, angle, width, or height). For 2D rules, participants were grouped according to whether the target rule was positively sloped or negatively sloped. In the 2D-RECT condition, positive and negative 2D rules are hereafter referred to as *shape* and *size* rules, respectively. There was no difference between rule variants in the 1D-RECT ($t(11.6) = .34$, $p = .74$) and 2D-DIAL ($t(23.9) = .92$, $p = .36$) conditions. Within the 1D-DIAL condition, participants learning a rule defined on the angle dimension were more accurate than those learning a rule on the radius dimension ($t(19.5) = 4.2$, $p < .001$). Finally, within the 2D-RECT condition, accuracy was higher for participants learning a shape rule as compared to those learning a size rule ($t(13.1) = 6.9$, $p =< .001$).

Modeling classification boundaries The goal of the second analysis was to relate classification performance to the form of hypotheses generated in each condition. Bayesian logistic regression was used to estimate linear decision bounds for each block of test responses, using the *bayesglm* function within the *arm* R package. Four models were estimated for each block: a 2D model with both features as predictors, two 1D models with a single feature as a predictor, and a baseline (intercept-only) model. Decision boundaries were then classified according to the model with the lowest AIC (thirty-two blocks, or 4%, were best-fit by the baseline model, indicating that there were few cases in which a linear decision boundary was not supported by participants' responses). The proportion of 2D boundaries across blocks is shown in Figure 3A.

Logistic regression on the proportion of 2D rules revealed significant effects of stimulus type (RECT: Wald $z = 6.0$, $p < .001$), rule type (2D: $z = 5.8$, $p < .001$), and the stimulus $\times$ rule interaction ($z = -4.7$, $p < .001$). As shown in Figure 3A, in the 1D-DIAL and 2D-RECT conditions, nearly all participants respond with a rule of the correct form throughout the task. This included 2D-RECT participants learning a *size rule* who responded with relatively low accuracy, indicating that the poor performance in that condition was not due to a failure to consider 2D hypotheses. In contrast, in both

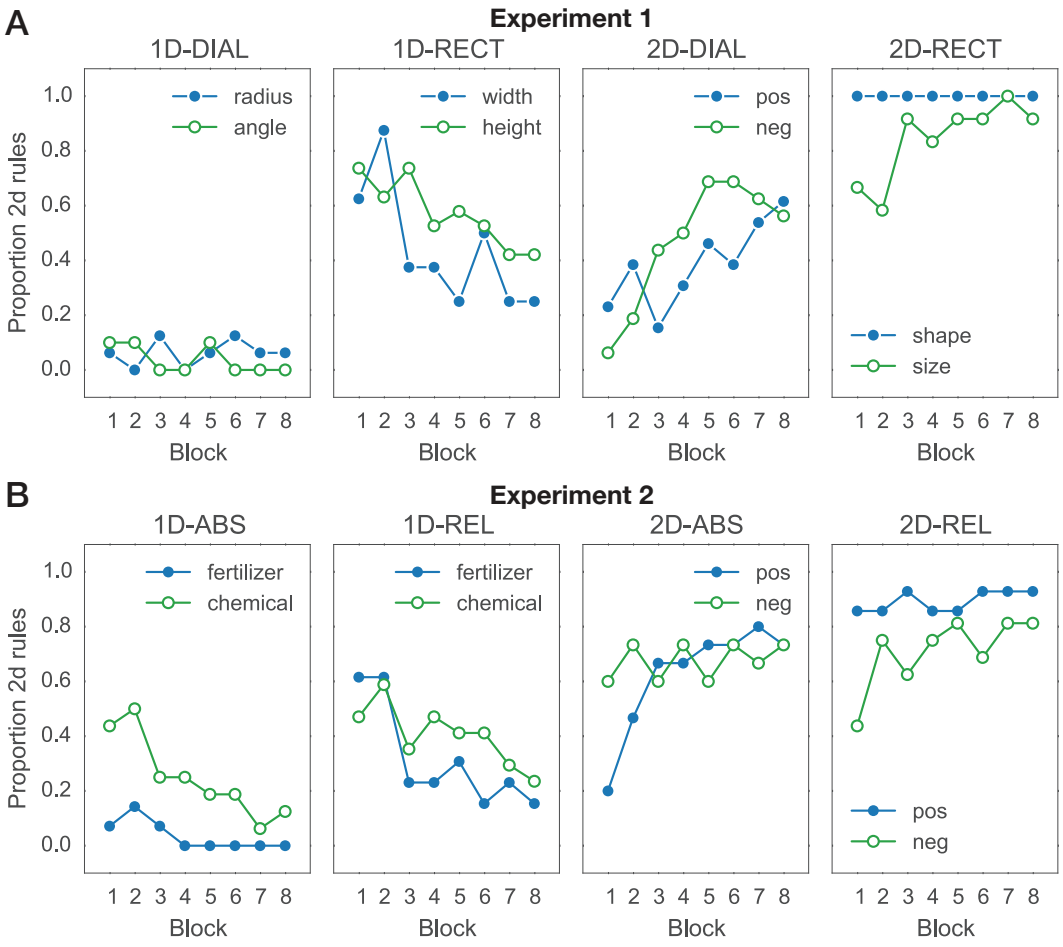


Figure 3: Proportion of best-fit decision boundaries classified as 2D. Within each condition, a separate line is shown for participants learning each type of target rule.

the 1D-RECT and 2D-DIAL conditions, decision boundaries were approximately equally divided between 1D and 2D hypotheses, suggesting that the manipulation of stimulus representation had the predicted effect on the kinds of hypotheses that were generated during learning.

Discussion

The results confirmed the hypothesis that the success of self-directed learning depends on a match between the target rule and salient features of the stimuli. In addition to replicating the difference in performance between 1D-DIAL and 2D-DIAL accuracy observed by Markant and Gureckis (2014), this gap in performance was reversed by changing the stimulus representation in the RECT conditions. The model-based analysis of rule use indicates that poor performance in the 1D-RECT and 2D-DIAL conditions was due, at least in part, to the generation of hypotheses of the wrong form.

In the 2D-RECT condition, performance diverged strongly depending on the form of the target rule, with higher accuracy among participants learning the *shape rule* than the *size rule*. However, rapid learning of the shape rule may be unsurprising given that the category boundary coincided with a simple relational comparison of the two features (i.e., whether the shape is taller than it is wide). The same general pattern was reported by Ashby and Gott (1988, Exps. 1 and 2) using similar stimuli (perpendicular line segments) under passive training. They found that participants were close to optimal performance when learning a positive, 2D classification rule that required comparing the lengths of the two line segments. When tasked with learning a 1D rule with the same stimuli, participants still responded according to 2D rules that integrated the two features. The authors concluded that the ease of comparing the two features led people to adopt 2D decision rules despite being able to separately attend to individual features. The present results show that this interference generalizes to other types of stimuli, and, more surprisingly, has a persistent effect under self-directed conditions despite learners' control over the training experience. This effect is particularly striking in the 1D-RECT condition in which, even on the last block of training, more than a third of participants responded according to some form of 2D hypothesis.

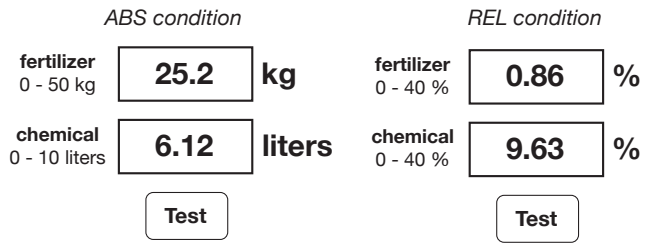


Figure 4: Example training trial displays in the ABS (left) and REL (right) conditions in Experiment 2.

Experiment 2

The goal of the second experiment was to evaluate the generalizability of the previous findings to an abstract domain in which perceptual biases were minimized. The task was designed to be structually equivalent to that of Experiment 1, while involving stimuli defined by abstract numerical features. As in Experiment 1, hypothesis generation was biased by manipulating the ease with which the two feature values could be integrated (see Figure 4). In the *relative* (REL) condition, the two features had the same range and were described in the same units. The common scale was predicted to cause participants to generate hypotheses based on integrating information about the two features. In the *absolute* (ABS) condition, the features had different ranges and were described in terms of different units. Like the DIAL condition in Exp. 1, the ABS condition was predicted to increase the likelihood of generating 1D rules involving a single dimension. Moreover, there should be an interaction between the stimulus representation and the target rule, such that SDL is most effective when the feature description facilitates the generation of the correct form of rule.

Participants and materials

One-hundred twenty people were recruited from the MPI participant pool (54 male, 62 female, 4 no gender given; age: $M = 25.2$, $SD = 3.3$) and were paid in the same manner as in Experiment 1. Participants were randomly assigned to one of four conditions: 1D-ABS, 1D-REL, 2D-ABS, or 2D-REL.

The goal of the task was to learn how the amount of two substances (a *Chemical* and a *Fertilizer*) affected whether a patch of virtual farmland would experience a successful (*S*) or failed (*F*) crop. Stimuli were defined by two continuous dimensions corresponding to the quantities of each substance. Within each target rule condition (1D or 2D), participants were assigned one of four possible variants of the target rule corresponding to different rotations of the rule in the stimulus space. In addition, the target rule was offset such that the classification boundary did not bisect the stimulus space.

In the ABS conditions, the two substances were defined in terms of different units (kg or liters) and had different ranges (one feature ranged from 0 to 50 while the second ranged from 0 to 10). In the REL conditions, both dimensions had the same range (0 to 40) and were labeled as percentages of a mixture applied to the soil (see Figure 4).

Test items A set of 32 items were generated by tiling the stimulus space at even intervals. Test sets were generated by randomly perturbing the location of each coordinate by a small amount (5% of the range on each dimension) while ensuring a constant number of items from each category. The order of test items and the order of the eight test blocks were randomized for each participant.

Procedure

The participants' goal was to learn how different feature combinations affected whether a patch of farmland would experience a successful or failed crop. Aside from the cover story and stimuli, the structure of the task was identical to that of Experiment 1. Participants completed 8 blocks, alternating between 16 learning trials followed by 32 test trials.

Learning trials On each learning trial an image of an empty plot of land appeared above two input boxes corresponding to the features. Each input was labeled with the name of the feature, the corresponding units ("kg" and "liters" in the absolute condition; "%" in the relative condition), and the possible range of each feature (Figure 4). Both features were initialized with values drawn from a uniform distribution over the corresponding ranges. Participants could then alter the value of either dimension by entering a new number within the allowed range. They then clicked a button to test the chosen combination of feature values. If the combination led to a successful crop, a new image appeared with fruit on the plot of land and the category label "Success!". If the combination led to a crop failure, a new image without fruit appeared along with the category label "Failure."

Test trials Each test trial began with the presentation of the image of an empty plot of land and a test item. Stimulus values were displayed in the same manner as in learning trials but could not be altered by the participant. Participants clicked on the outcome that they predicted to occur for the displayed feature combination. At the end of each block they were told their proportion of correct predictions.

Results

Classification accuracy Accuracy across blocks in shown in Figure 2B. Two participants (one in the 1D-ABS condition and the other in the 2D-REL condition) were excluded from further analysis because their overall accuracy was more than three standard deviations below the mean of their condition. A 2×2 ANOVA on overall classification accuracy (collapsed across rule variants) was performed with stimulus (ABS vs. REL), rule (1D vs. 2D) and stimulus $\times$ rule interaction as factors. There was no main effect of stimulus type ($F = (1, 114) = .09$, $p = .76$), but there was a significant main effect of rule type ($F(1, 114) = 25.5$, $p < .001$) and a significant stimulus type $\times$ rule type interaction ($F(1, 114) = 23.6$, $p < .001$). Tukey HSD tests indicated significant pairwise differences in accuracy between conditions, with the excep-

tion of 1D-REL and 2D-REL conditions (see Table 1).

Two-sample t-tests were used to assess effects of rule variants on overall accuracy within each condition. Overall accuracy was higher for learning rules on the Fertilizer dimension in the 1D-ABS condition ($t(19.9) = 3.2$, $p < .01$) but the differences within the remaining conditions were not significant.

Modeling classification boundaries The same method from Experiment 1 was used to categorize participants' decision boundaries as 1D or 2D. Logistic regression on the proportion of 2D rules revealed significant effects of stimulus type (REL: Wald $z = 5.0$, $p < .001$), rule type (2D: $z = 8.1$, $p < .001$), and a stimulus $\times$ rule interaction ($z = -3.6$, $p < .001$). The proportion of 2D boundaries are shown in Figure 3B, separated by target rule. The manipulation of feature representation had the predicted effect on the generation of 2D rules, albeit to a lesser extent than seen in Experiment 1. In the 1D-ABS condition the proportion of 2D rules was very low, whereas in the 2D-REL condition the proportion of 2D rules consistent with the target is high. Finally, in the 1D-REL and 2D-ABS conditions, there was a higher proportion of decision boundaries that were of a different form than the target rule.[1]

Summary

The present findings show that, in both perceptual and abstract domains, the efficacy of SDL is limited by biases in the hypothesis generation process. In addition to replicating the gap in performance between 1D and 2D rules found by Markant and Gureckis (2014) in both domains, this gap was eliminated through simple manipulations of the stimuli. When the stimulus representation facilitated the generation of hypotheses consistent with the true rule, self-directed learners were more likely to classify items using a hypothesis of the same form. When hypothesis generation was inconsistent with the target rule, self-directed learners were less successful at learning the category structure despite their ability to control the selection of training data. This impairment is perhaps most striking in the 1D-RECT and 1D-REL conditions, in which the learning of simple 1D rules suffered because participants were more likely to respond according to 2D hypotheses.

This study was motivated by one account of the hypothesis generation process, involving the sampling of salient cues or relations from the environment to form hypotheses about an underlying structure or concept (Trabasso et al., 1968; Wallach, 1962). The present findings suggest that the success of SDL depends on the way that hypothesis generation is shaped by the environment, including how materials guide attention and set the stage for the perception of relevant features or relationships (Goldstone, Landy, & Son, 2010). It is important to note, however, that a number of other hypothesis generation mechanisms have been proposed, including processes based on memory retrieval (Dougherty, Thomas, & Lange, 2010) and local adjustment of existing hypotheses (Bramley, Dayan, & Lagnado, 2015). An important goal of future work is to understand how these generation processes interact in order to guide information search during self-directed learning.

[1]The same analysis was performed on participants' selections during training (rather than test responses) to evaluate whether biased information sampling contributed to these results. However, the proportion of best-fit 2D boundaries indicated that participants' selections strongly supported the correct form of rule in both Exp. 1 (1D-DIAL: .08; 1D-RECT: .02; 2D-DIAL: .93; 2D-RECT: .91) and Exp. 2 (1D-ABS: .06; 1D-REL: .08; 2D-ABS: .86; 2D-REL: .88).

Acknowledgements

The author thanks Katja Munz, Annika Walinski, Jann Wascher, Chantal Wysocki, and Andrii Zakharov for help in translating the experiment and data collection.

References

Ashby, F., & Gott, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F., Paul, E., & Maddox, W. (2011). COVIS. In E. M. Pothos & A. J. Willis (Eds.), *Formal approaches in categorization* (pp. 65–87). Cambridge University Press.

Bramley, N. R., Dayan, P., & Lagnado, D. A. (2015). Staying afloat on neurath's boat–heuristics for sequential causal learning. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 262–267).

Dougherty, M., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. *Psychology of Learning and Motivation*, *52*, 299–342.

Goldstone, R., Landy, D., & Son, J. (2010). The education of perception. *Topics in Cognitive Science*, *2*(2), 265–284.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464-481. doi: 10.1177/1745691612454304

Gureckis, T. M., Martin, J., McDonnell, J. V., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 1–14. doi: 10.3758/s13428-015-0642-8

Johnson, T., & Krems, J. (2001). Use of current explanations in multicausal abductive reasoning. *Cognitive Science*, *25*(6), 903–939.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, *12*, 1–48.

Macmillan, N. A., & Ornstein, A. S. (1998). The mean-integral representation of rectangles. *Perception & Psychophysics*, *60*(2), 250–262.

Markant, D., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. doi: 10.1037/a0032108

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14–19.

Shepard, R. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

Trabasso, T., Bower, G., & Gelman, R. (1968). *Attention in learning: Theory and research*. Wiley.

Tweney, R., Doherty, M., Worner, W., Pliske, D., Mynatt, C., Gross, K., & Arkkelin, D. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, *32*(1), 109–123.

Wallach, L. (1962). The complexity of concept-attainment. *The American Journal of Psychology*, *75*(2), 277–283.

Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.

Representing Sequence: The Influence of Timeline Axis and Direction on Causal Reasoning in Litigation Law

Amy Rae Fox (amyraefox@gmail.com)
Department of Psychology, California State University - Chico
(current) Department of Cognitive Science, University of California San Diego
San Diego, CA, USA

Martin van den Berg (mvandenberg@csuchico.edu)
Department of Psychology, California State University - Chico
Chico, CA USA

Erica de Vries (Erica.deVries@univ-grenoble-alpes.fr)
Laboratory of Educational Sciences, Univ. Grenoble Alpes
Grenoble, France

Abstract

Can the representation of event sequence influence how jurors remember and reason in a legal case? We addressed this question by examining the interaction between an individual's preferred *spatial construal of time* (SCT) for an external (visual-spatial) representation and the SCT of a courtroom graphic. One hundred fifty three undergraduates played the role of jurors in a fictitious civil trial. The details of a case were recounted in a multimedia presentation featuring timelines animated in one of four orientations: left-right, right-left, top-bottom, and bottom-top. Participants were assessed on measures of comprehension and causal reasoning. Results indicated effects of timeline orientation and SCT choice behavior on comprehension and reasoning. We discuss these results in terms of the role of attention in temporal-causal reasoning, and implications for the design of multimedia materials for the courtroom.

Keywords: external representation; courtroom graphics; visualization; spatial construal of time; sequence; events; multimedia learning

Introduction

In litigation, lawyers often describe a situation while making an argument as to the cause of an alleged wrongdoing. Temporal order—a sequence of events—is the most basic requirement for causation. Increasingly, lawyers are turning to graphical representations such as animated PowerPoint presentations to support their courtroom arguments. In this study, we investigate the influence of the visual-spatial representation of event sequence on the comprehension and causal reasoning of jurors.

Visual Evidence

While computers and their multimedia artifacts are nearly ubiquitous in classrooms and boardrooms, their introduction to the American courtroom is a more recent phenomenon. A growing body of research is addressing the use of computer-generated exhibits as *demonstrative evidence*[1]. Park and Feigenson (2013) found that mock jurors remembered more information offered by attorneys using PowerPoint presentations than those offering the same information by oral argument alone. Participants also believed these attorneys were more credible, and decided in favor of their clients more often. The authors concluded that the use of visual aids influenced juror decision-making through both cognition (comprehension of the evidence) and persuasion (attitudes about the evidence). In a review of research addressing the effect of visualizations on courtroom decision-making, Feigenson (2010) noted a lack of convergent evidence on the effects of courtroom graphics, suggesting a nuanced role of visual evidence moderated by a number of factors, including: message format (i.e. still image, static-sequential animation, moving image animation, simulation), features of the case (e.g. complexity and familiarity of the scenario), and features of the case presentation (e.g. timing in case, differential use by opposing party). Feigenson also hypothesized a number of mediating variables, including: comprehension of the scenario, juror's ability to visualize the scenario, credibility, likeability, and emotional responses. While the results of experimental studies are mixed, this body of research suggests a range of effects that multimedia exhibits may have on comprehension, reasoning and subsequent decision-making. There is an urgent need for targeted research to explore the mechanisms by which these effects may occur, and the extent to which they are relevant to the high stakes arena of litigation law.

Sequence

We are interested in exploring how litigators might use multimedia displays to communicate about *causation*. Causation is a complex, multi-faceted construct in both law and philosophy, that has at least one basic requirement: the order of events. A cause can only be a cause if it occurs before an effect. In the auditory verbal medium, we can represent temporal order *in situ*, the first thing I say precedes the second thing I say, and so on. Similarly, in the visual-verbal medium, the first thing I write precedes the second. Research on text and listening comprehension has shown that comprehension for temporal order is significantly better when events

[1] Demonstrative evidence is offered to illustrate or clarify the testimony of a witness or the argument of a lawyer. Examples include a list of facts the lawyer refers to in supporting a claim of causation, or a timeline of events leading up to a crime.

are presented in chronological order in language (Mandler, 1985; Fenker, Waldmann, & Holyoak, 2005). Accordingly, if lawyers wish for their arguments to be understood and remembered, they would do well to instruct jurors about a sequence of events in the order they are alleged to unfold. To accompany such an oral argument, a lawyer might use a visual aid such as a *timeline* as demonstrative evidence.

A timeline is both a communication tool and a cognitive artifact. It consists of a chronological organization of events, most often depicted on a two-dimensional surface. Events may be represented by descriptions (i.e. text) and/or depictions (i.e. icons, pictograms or photographs). The *flow of time* unfolds along a linear path, often a horizontal or vertical axis. The positions of events serve as indicators anchoring their relative place in time. Depending on the granularity of detail, much information about the temporal relations of the events may be extracted from the representation, such as timing (a date/time) and duration (a quantity). Sequence information is always present, inherent in the spatial structure of the representation. In conjunction with an auditory narration, a pictorial timeline becomes a multimedia presentation. A timeline may also be animated to progressively elaborate a sequence of events *in situ*. Such a multimedia representation is potentially powerful, with both space and time, verbal and pictorial, auditory and visual mechanisms employed to communicate event sequence.

Learning About Sequence From Multimedia

When individuals experience a multimedia presentation, they process the stimuli through sensory mechanisms into working memory where stimuli are processed in accordance with their modalities (auditory, visual) and representational formats (descriptive, depictive) (Schnotz, 2014). The information is integrated with existing knowledge from long-term memory, the result of which is a mental model, representing the sequence of events in terms of spatial relations (Schaeken & Johnson-Laird, 1995). But what happens when the stimuli contain information about abstract concepts, such as time? A substantial body of research suggests that the embodied experience of space serves to structure our conceptualization of the abstract notion of time, such that certain properties of space (e.g. relative position, continuity) are imported into the domain of time (e.g. sequence, succession) (Lakoff & Johnson, 1980; Boroditsky, 2000). The mapping of time onto space is guided by conventions established through the habitual use of language and cultural artifacts (Núñez & Cooperrider, 2013). We call these mappings *spatial construals of time* (henceforth SCTs). Although multiple SCTs may be present in long-term memory (Núñez & Cooperrider, 2013), it is theorized that the import of mappings into working memory for task performance is constrained by a coherence-seeking mechanism[2]. According to the Coherent Working Models Theory, given a set of available mappings, only one is selected based on its adequacy to fulfill task demands (Torralbo, Santiago, & Lupiáñez, 2006; Santiago, Román, & Ouellet, 2011; Santiago, Ouellet, Román, & Valenzuela, 2012).

A task requiring the representation of temporal sequence on a two- dimensional surface brings attention to an allocentric frame of reference (Torralbo et al., 2006), thus activating a Reading/Writing Direction (RWD) consistent SCT of left-right for English speakers (Tversky, Kugelmass, & Winter, 1991). The result is the construction of a left-right oriented mental model in working memory, structuring knowledge of the sequence of events. If, however, a multimedia stimulus is presented in a different SCT, an individual must either import an alternative mapping into working memory, or perform a transformation of the incoming information into the SCT of the existing mapping. This raises two important questions: (1) Does the format of multimedia materials impact on the construction of the mental model? (2) If so, how does this impact reasoning operations on the mental model?

Present Investigation

We apply this conceptual framework to the domain of litigation law with three goals: (1) Preferences: replicate previous research on the relationship between SCTs and RWD, extended to computer-based stimuli. (2) Flexibility: test hypotheses derived from the Coherent Working Models Theory (Santiago et al., 2011) about the construction of mental models from inconsistent SCTs. (3) Stability: explore the stability of SCT preferences and potential impacts on mental model construction. We address these goals by focusing on the interaction between an individual's preferred SCT for a narrowly-defined task and the SCT of a stimulus.

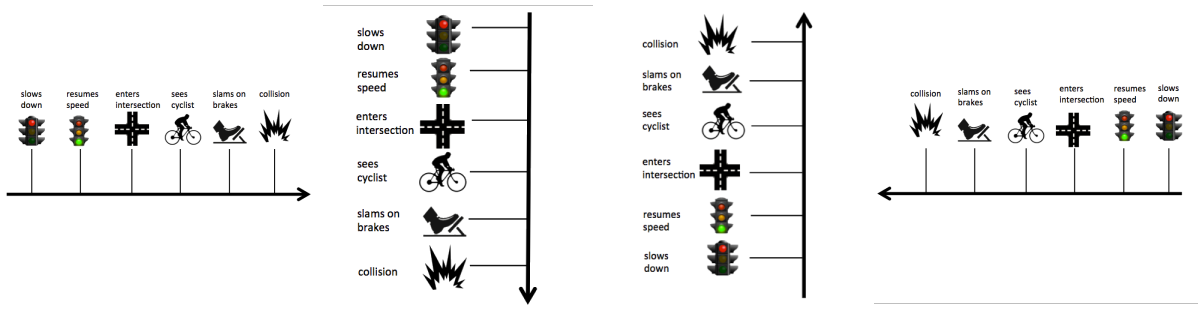
Hypotheses We hypothesize that individuals in the target population (English speakers of jury-eligible age) will prefer a RWD-consistent SCT (left-right) for a computer-based sequencing task. We predict the choice of this SCT will be *stable*: when asked to reconstruct a sequence of events after a stimulus, participants would likely persist. We hypothesize a limit to the flexibility of thinking with differing SCTs, such that the presentation of timelines oriented with different SCTs (same axis/opposite RWD or different axis) will impair the development of coherent mental models. Consequently, participants exposed to such stimuli will have poorer comprehension of the case and make more errors in causal reasoning.

Methodology

Design

Two factors, Timeline Axis and Consistency of Timeline Direction with Reading/Writing Direction (henceforth, direction) were fully crossed, yielding a 2-Axis (Horizontal/Vertical) x 2-Direction (Consistent/Inconsistent) design with four between-subjects conditions: (1)left-right, (2)top-bottom, (3)right-left, and (4) bottom-top.

[2]For example, it would be incoherent to simultaneously construe the flow of time as both Back-to-Front and Front-to-Back.



(a) $SCT_s = L - R$ (b) $SCT_s = T - B$ (c) $SCT_s = B - T$ (d) $SCT_s = R - L$

Figure 1: Stimuli - Timelines for Four Experimental Groups

Participants

One hundred fifty three undergraduates (63% female) sampled from a mid-sized American university were randomly assigned to the experimental conditions in exchange for course credit. The participants ranged from 18 to 64 years of age (Median=22). All reported fluency in English, with 89% reporting English as a first language, and 11% reporting Spanish. Native speakers of Mandarin, Arabic and Farsi were excluded from the sample. No significant differences were found between experimental groups with respect to other demographic variables, including involvement in traffic accidents and laterality.

Materials

Measure of Spatial Construal of Time A novel measure was developed as an indicator of participants' preferred spatial construal of time (SCT) for event sequence in the context of a well-defined task. Participants were asked to construct a timeline to indicate the order of a sequence of events. They were first instructed to choose an orientation (axis and direction) for the timeline (Figure 2). The orientation selected by the participant was recorded as the SCT and utilized for the subsequent sequencing task. The SCT measure was utilized twice during the experimental protocol, once before stimulus presentation (SCT_1), and once after (SCT_2).

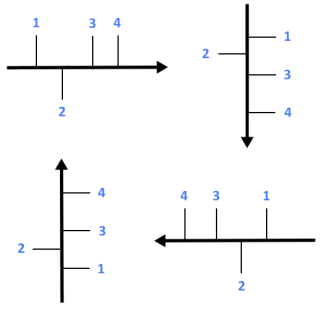


Figure 2: Measuring SCT preference for representational task. *The second item in each figure was positioned on the other side of the timeline to demonstrate that both sides could be utilized in the following task.*

Experimental Stimuli The experimental stimuli consisted of a fictitious civil litigation involving a motorist/cyclist traffic accident[3]. The scenario was developed such that verdicts in favor of the plaintiff or defendant were equally justified, depending on which of the conflicting witness statements a juror chose to believe. Measures of comprehension and reasoning were balanced for either verdict, mitigating the impact of bias toward either party.

The experimental manipulation was embedded in a multimedia presentation of testimony consisting of a PowerPoint presentation, ostensibly displayed in a courtroom accompanying a lawyer's examination of a witness. Participants heard an unidentified lawyer questioning a police officer who responded to a traffic accident. The police officer reported the events leading up to the accident as described by the defendent motorist and plaintiff cyclist. During each description, a timeline appeared on-screen, serving as the experimental manipulation. The axis and direction of the timeline depicted were different for each experimental group (Figure 1). Each timeline was animated in a sequential fashion and synchronized with the audio description.

Measure of Comprehension Comprehension was measured via twenty-five multiple-choice questions developed in accordance with the Meaning Identification Technique (MIT) for evaluating reading and listening comprehension (Marchant, Royer, & Greene, 1988).

Measure of Reasoning To assess temporal-causal reasoning, participants were asked to arrange twenty-eight events described in the testimony along a timeline (Figure 3). This task aims to capture both the *spatial structure* and *content* of a mental model. The structure of the model was determined by the chosen orientation of the timeline SCT_2. Causal reasoning was evaluated by scoring the submitted sequence in relation to its logical consistency with the verdict the participant rendered. For example, if a participant rendered a verdict in favor of the cyclist but indicated on the timeline that the cyclist ran the red light, the response would be logically

[3]Based on a scenario from the 2014 Colorado State High School Mock-Trial Program (Colorado Bar Association, 2014).

inconsistent. Two graduate students developed the scoring rubric which was applied via a computer algorithm.

Procedure

Participants were seated at computer workstations equipped with headphones and guided through the study under the guise of a mock-trial. First, participants completed the first measure of spatial construal of time (SCT_1). Then, they viewed the stimulus multimedia presentation. Following the testimony, they answered 25 questions testing their comprehension of the case. Then, they constructed a timeline of events mentioned in the testimony (SCT_2 and reasoning measures). Finally, they were asked to render a verdict. Participants were then debriefed. The total runtime averaged 43 minutes.

Results

Preferences for SCTs

A majority (76%) of participants selected a left-right SCT in the first sequencing task (henceforth SCT_1-constrained sample), followed by 12% selecting bottom-top, 10% top-bottom, and 2% right-left. In paper-based studies, participants face a stimulus oriented parallel to their sagittal axis, while in the present computer-based study, participants faced a stimulus oriented perpendicular to their sagittal axis (Figure 4). Our results are consistent with findings of studies conducted on children with paper-based stimuli (Tversky et al., 1991), suggesting that the influence of RWD is consistent across at least two spatial axes as well as a change in representational media.

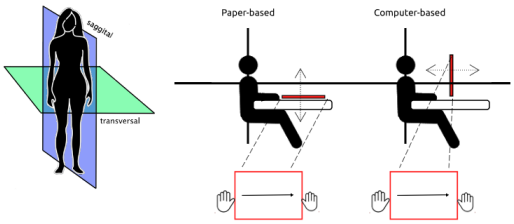


Figure 4: Orientation of Stimuli in Paper vs. Computer-based Studies.

Flexibility in SCTs

As hypothesized, a significant positive correlation was found between measures of comprehension and reasoning (r=.273, p<.01). A factorial MANOVA examining the effect of stimulus SCT on comprehension and reasoning yielded a significant main effect for direction, $\Lambda = .95$, F (2,111)= 3.08, p=.05, $\eta_p^2 = .05$. Univariate analyses revealed a significant effect of direction only on comprehension, F (1,112)=3.92, p=.05, $\eta_p^2 = .03$. Inspection of the estimated group-means revealed that the direction of the effect was *opposite* to our hypothesis. Timelines oriented inconsistent with RWD (bottom-top and right-left) were related to higher comprehension scores (see Figure 5). The data did not reflect a significant interaction between axis and direction, and no effects were significant for reasoning.

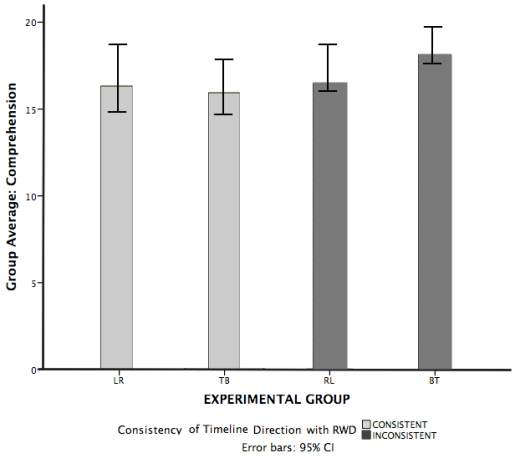


Figure 5: Group Means of Comprehension by Experimental Group.

Stability in SCTs

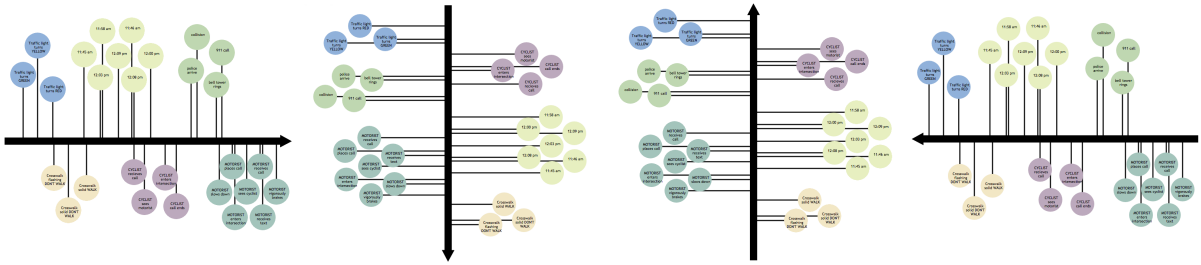
To explore the stability of SCT preferences while performing cognitive activities, we derived a new measure, *SCT Choice Behavior*[4], based on the post-stimulus SCT_2 in relationship to the SCT_s (experimental group) and pre-stimulus SCT_1.

Of the 116 participants in the SCT_1-constrained sample, seventy-one persisted with SCT_1 when presented with a differing SCT_s. Twenty-eight received the same stimulus as SCT_1 and persisted. Ten adapted to the SCT_s and seven chose a third, different SCT_2. We performed a multivariate Kruskal-Wallis test to examine the influence of SCT choice behavior on comprehension and reasoning, which revealed a significant effect on reasoning, $\chi_2(3, n=116)=10.7$, p=.013. Participants who chose SCT_2 different than both SCT_1 and stimulus SCT_s had significantly lower scores on the reasoning task.

Discussion

To explore preferences, flexibility and stability of spatial construals of time (SCTs) and effects on courtroom reasoning, we asked participants to assume the role of jurors in a mock-trial. We successfully replicated previous findings on the concordance of SCTs with RWD (Tversky et al., 1991) this time with computer-based interactive data visualizations. Concerning flexibility, our results shed light on the flexibility of SCTs and the role of attention as a coherence-seeking mechanism. Regarding stability, we offer new evidence as to the stability of SCT choices over sequential representational tasks, and discuss how these behaviors might affect operations on mental models.

[4]SCT Choice Behavior values: *persist* ($SCT_1 \neq$ stimulus, $SCT_2 \equiv SCT_1$), *adapt* ($SCT_1 \neq$ stimulus, $SCT_2 \equiv$ stimulus), *neither* ($SCT_1 \neq$ stimulus, $SCT_2 \neq$ stimulus, $SCT_1 \neq SCT_1$), *indeterminate* ($SCT_1 \equiv$ stimulus $\equiv SCT_2$).



(a) $SCT_2 = L - R$ (b) $SCT_2 = T - B$ (c) $SCT_2 = B - T$ (d) $SCT_2 = R - L$

Figure 3: Start Positions of Interactive Timelines for Reasoning Measure. *Text inside circles provide short description of events (e.g. CYCLIST enters intersection). Circles are organized into color coded categories (e.g. car events, bicycle events, traffic signals, etc.)*

Preferences for SCTs

As predicted, participants demonstrated a strong preference for the left-right SCT in the sequencing task. The observation that the preference was consistent across transversal and sagittal axes presents an interesting question for future research on SCTs. Might a change in axis be equally flexible when considering deictic time as sequential time? Or is this effect only observable for sequential relations, and on axes for which culturally derived SCTs exist? Answers to these questions may have practical applications in the realm of immersive virtual reality and 3D data visualization, as well as shed light on the complex relationship between external representations and temporal/spatial cognition.

Flexibility in SCTs

The most interesting results came when examining the influence of stimulus SCT_s on comprehension. We found that stimulus SCTs different from RWD resulted in superior comprehension. One explaination for this result is found by considering the role of attention in thinking with SCTs. By asking participants to construct a simple timeline prior to stimulus exposure, we brought attention to their preferred spatial construal for the task (SCT_1), ostensibly resulting in import into working memory. When presented with a different SCT_s (stimulus) individuals needed to import an alternative mapping, or transform the incoming information to the SCT_1. Rather than impair model construction, our data suggest this allocation of cognitive resources had an advantageous effect. If we assume that the discrepancy between SCT_1 and timeline orientation required additional attention be paid to the stimulus, then this increased attention may have resulted in a net increase in the cognitive resources dedicated to model construction. The harder the task was, the more the attention the participants devoted. As attention is a limited resource, however, we think it unlikely this effect would persist for increasingly complex tasks. We predict a threshold level of complexity, after which inconsistent SCTs would result in decreased performance.

The data failed to support our hypothesis that inconsistent timeline SCTs have a deleterious effect on causal-reasoning. We presumed that reasoning, an operation that manipulates a mental model, depends first on the fidelity of the contents of the model (Schaeken & Johnson-Laird, 1995). In this way, we expected that comprehension and reasoning measures would be strongly correlated. The actual correlation between measures was weak (r=.273, p<.001), and in fact only significantly correlated for the left-right stimuli condition (r=.403, p=.05). This suggests that participants in other groups may have found the reasoning task so challenging that they either substantially reduced their effort, or, through manipulation of the interactive data visualization, altered the contents of their mental model. To investigate these alternatives in the future, we need to compare the internal consistency of answers on the comprehension measure with arrangement of events on the reasoning measure. This will allow us to determine if participants indicated a different understanding of the sequence of events on the reasoning task than they indicated on the prior comprehension task.

We suspect that the task difficulty likely influenced the effort expended on the reasoning task (indicated by reasoning time). While there were no significant differences in reasoning time between groups, there was a strong correlation between reasoning time and reasoning scores only for the right-left group (r=.647, p<.001). It is likely that the lengthy manipulation of the interactive data visualization required by the reasoning task had the unintended consequence of altering participants' mental models, rather than reflecting their structure and content. A substantial body of literature supports the view that data visualizations are tools on which individuals may offload cognitive processing (see (Hollan, Hutchins, & Kirsh, 2000). We seek to refine the reasoning measure to more accurately reflect the content of participants' mental models without manipulating them.

Stability of SCTs

As predicted, most individuals persisted with their intial choice of SCT. Our analysis revealed that participants who chose a SCT_2 different from both their SCT_1 and SCT_s had

significantly lower scores on the reasoning task. This result may indicate a limit on the flexibility of SCTs during higher order cognitive activities; perhaps individuals can perform transformation between two SCTs without performance impairment, but not three. We plan a follow-up experiment to investigate the use of differing SCTs within the same stimulus presentation (i.e. one orientation for the defense, a different orientation for the prosecution). Alternatively, it is possible that the choice of a third SCT for the reasoning task was in itself indicative of a lack of effort on the part of the participant.

Limitations and Future Work

While we placed a high value on external validity in the design of our experimental materials, the participants' exposure to stimuli was not reflective of genuine litigation, where juries hear arguments over the course of several hours or days before group deliberation. Additionally, participants were not permitted to use external cognitive aids such as note taking, or review of testimony and transcripts. Any effect of graphics on real-life courtrooms must be considered in combination with the effects of persuasive argumentation and jury deliberation, and how these might operate in a more representative population.

In this investigation, we approached the courtroom as a classroom; before jurors can be persuaded, they must be instructed about the details of a case. Our approach was to apply research from the learning sciences to understand how jurors might integrate information from multiple sources and modalities. Our results add to the growing body of research on the influence of multimedia in the courtroom (see (Feigenson, 2010, 2011; Park & Feigenson, 2013) by providing evidence that differential presentations of temporal sequence *can* influence comprehension. To clarify these results, we plan subsequent studies that carefully control allocation of attention to the learning materials and measurement tasks. Our data suggest that the orientation of timelines can impact the comprehension of jurors, and that in situations where a lawyer wishes to capture jurors' attention, a left-to-right timeline may be preferable. We recommend testing the hypothesis that SCTs inconsistent with RWD improve comprehension by inducing increased allocation of attention, up to a threshold, at which point performance will begin to degrade. Answers to these questions will guide the designers of courtroom multimedia presentations on how to orient timelines to be maximally coherent for jurors; or alternatively, how to induce confusion for persuasive purposes.

Acknowledgments

The authors gratefully acknowledge the U.S. Department of Education for support of the International Cognitive Visualization program.

References

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, *75*(1), 1–28.

Feigenson, N. (2010). Visual evidence. *Psychonomic Bulletin & Review*, *17*(2), 149–154.

Feigenson, N. (2011). The Visual in Law: Some Problems for Legal Theory. *Law, Culture and the Humanities*, *10*(1), 13–23.

Fenker, D. B., Waldmann, M. R., & Holyoak, K. J. (2005). Accessing causal relations in semantic memory. *Memory & cognition*, *33*(6), 1036–1046. doi: 10.3758/BF03193211

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174–196.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Mandler, J. M. (1985). On the comprehension of temporal order. , *1*(4), 309–320.

Marchant, H. G., Royer, J. M., & Greene, B. (1988). Superior Reliability and Validity for a New Form of the Sentence Verification Technique for Measuring Comprehension. *Educational and Psychological Measurement*, *48*, 827–834.

Núñez, & Cooperrider. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences*, *17*(5), 220–9. doi: 10.1016/j.tics.2013.03.008

Park, J., & Feigenson, N. (2013). Effects of Visual Technology on Mock Jurors. *Applied Cognitive Psychology*, *27*(December 2012), 235–246.

Santiago, Ouellet, M., Román, A., & Valenzuela, J. (2012). Attentional factors in conceptual congruency. *Cognitive Science*, *36*, 1051–1077.

Santiago, Román, A., & Ouellet, M. (2011). Flexible foundations of abstract thought: A review and a theory. In T. W. Schubert & A. Maass (Eds.), *Applications of cognitive linguistics: Spatial dimensions of social thought*. Berlin: Walter de Gruyter.

Schaeken, W., & Johnson-Laird, P. N. (1995). Mental models and temporal reasoning. *Cognition*, *60*(96), 205–234.

Schnotz, W. (2014). Integrated Model of Text and Picture Comprehension. In R. Mayer (Ed.), *The cambridge handbook of multimedia learning* (2nd ed., chap. 4). Cambridge: Cambridge University Press.

Torralbo, A., Santiago, J., & Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. *Cognitive Science*, *30*(4), 745–57.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, *23*, 515–557.

Attractivity Weighting: Take-the-Best's Foolproof Sibling

Paul D. Thorn (thorn@phil.hhu.de)

Gerhard Schurz (schurz@phil.hhu.de)

DCLPS, Theoretical Philosophy, University of Duesseldorf,
Universitaetsstr. 1, Duesseldorf 40225, Germany

Abstract

We describe a prediction method called "Attractivity Weighting" (AW). In the case of cue-based paired comparison tasks, AW predicts a weighted average of the cue values of the most successful cues. In many situations, AW's prediction is based on the cue value of the most successful cue, resulting in behavior similar to Take-the-Best (TTB). Unlike TTB, AW has a desirable characteristic called "access optimality": Its long-run success is guaranteed to be at least as great as the most successful cue. While access optimality is a desirable characteristic, concerns may be raised about the short-term performance of AW. To evaluate such concerns, we here present a study of AW's short-term performance. The results suggest that there is little reason to worry about the short-run performance of AW. Our study also shows that, in random sequences of paired comparison tasks, the behavior of AW and TTB is nearly indiscernible.

Keywords: Bounded Rationality; Ecological Rationality; Attractivity Weighting; Take-the-Best; Meta-induction.

Prediction Games

The object of study within the present paper is a prediction method known as "Attractivity Weighting" (AW). AW was introduced (under the name "weighted meta-induction") as a possible response to Hume's problem of induction (Schurz, 2008) based on findings in mathematical learning theory (cf. Cesa-Bianchi & Lugosi, 2006). The formal properties that make AW an attractive prediction method were demonstrated in the context of so called "prediction games." Prediction games will serve as a framework for evaluating prediction strategies within the paper. The main aim of the paper will be to assess the short-term performance of AW. As a secondary result, we will observe the near indiscernability of the behavior of AW and another method, called "Take-the-Best," in random sequences of paired comparison tasks.

Informally, a prediction game consists of a sequence of events, along with a collection of participating prediction methods. In a series of *rounds*, the participating methods deliver predictions about the character of successive elements of the event sequence, predicting the character of the first event in the first round, the second in the second, etc. At the end of each round, the actual value of the just considered event is revealed.

Formally, a prediction game is defined as a pair, (E, M), where $E = (e_1, e_2, \ldots)$ is an infinite sequence of events, and M is a finite set of methods. For example, the elements of E may be the weather conditions (rainy or not rainy) for a sequence of days. For convenience, the elements of E are assumed to be real numbers in the interval [0, 1]. The elements of M are of two sorts. Methods of the first sort, called "object methods," make their predictions independently of the other methods. Methods of the second sort, called "meta-methods," make their predictions based on the predictions of the object methods. As a basis for making their predictions, it is assumed that each meta-method has 'access' to the present and past predictions of each object method.

Within a prediction game, the *prediction of a method*, m, of the value of the nth event, e_n, is denoted $P_n(m)$. The normalized *loss* for an individual prediction, $P_n(m)$, is a function of the distance between the prediction and the event's value, and takes a value in [0, 1]. The *score* for a method, m, for event n is denoted $S_n(m)$, and is defined as 1 *minus* the loss for the prediction. By default, we assume that losses are measured by *linear* distance, that is: $S_n(m) = 1 - |e_n - P_n(m)|$.[1]

The results of the following section also depend on the assumption that each method makes a prediction concerning each event in the event sequence. In order to accommodate naturally occurring situations where some methods do not make a prediction concerning some events, we treat non-predictions as *ersatz* predictions, as distinguished from *genuine* predictions. Ersatz predictions are recorded as a prediction of 0.5, and scored accordingly (cf. Martignon & Hoffrage, 1999).

Prediction games represent a relatively general framework. For example, the framework is apt for representing cue-based paired comparison tasks. In that case, (i) cues are treated as object methods, and (ii) methods, such as Take-the-Best (TTB), that make their predictions based on cues are treated as meta-methods: TTB proceeds by ordering the available cues (object methods) according to their 'ecological' validity, and imitates the prediction of the first cue (object method) in the ordering that delivers a genuine prediction. The *ecological validity* of a method m, as of round n, is equated with the average score of the method for the genuine predictions that it made within the first n rounds.

Attractivity Weighting

The main object of our discussion will be the meta-method AW. AW's predictions are formed by taking a weighted average of the predictions of the (accessible) object methods. The weights that AW assigns to the object

[1] More generally, the results of the following section depend on the assumption that the loss function, L, is *convex*, i.e., for all r, d_i (distance i), and d_j (distance j): $L((1-r)d_i + rd_j) \leq (1-r)L(d_i) + rL(d_j)$.

methods are called "attractivities." The attractivity of an object method, at a given round, is determined by comparing the average score of the object method with the average score of AW.[2] If the average score of the object method is less than or equal to AW's score, then the attractivity of the method is *zero*. If the average score of the object method is greater than that of AW, then the attractivity of the method is equal to the difference between the two averages. Thus the attractivity of a method m after the nth round is defined as follows, where $\bar{S}_n(m)$ denotes the average score of a method m for the first n rounds (i.e., $\bar{S}_n(m) = \sum_{i=1}^{n}(S_i(m)/n)$):

$$at_n(m) = \begin{cases} \bar{S}_n(m) - \bar{S}_n(AW) & \text{if } \bar{S}_n(m) - \bar{S}_n(AW) > 0, \\ 0, & \text{else.} \end{cases}$$

AW's prediction in round n is based on the attractivities assigned in round n–1. In the case where all object methods are assigned zero attractivity in round n–1, it is stipulated that AW imitates the prediction of the object method whose ecological validity is the greatest as of round n–1 (with ties broken by a randomized tie-breaker). The method imitated by AW in such cases is denoted "$maxVal_{n-1}$."[3] In round one, AW predicts 0.5. Formally, the predictions of AW are defined as follows, where m ranges over the set of accessible object methods, and n > 1:

$$P_n(AW) = \begin{cases} \frac{\sum_m P_n(m)\, at_{n-1}(m)}{\sum_m at_{n-1}(m)} & \text{if for some m: } at_{n-1}(m) > 0, \\ P_n(maxVal_{n-1}), & \text{else.} \end{cases}$$

In the following section, we explain the result that AW has a desirable characteristic called "access optimality."

The Virtues of AW

An important characteristic of AW is that it is access optimal: In the long-run, the mean score of AW is guaranteed to converge to the mean score of the best scoring object method to which it has access. In other words, where "$maxSuc_n$" denotes the average score of the best scoring object method as of round n, $\bar{S}_n(AW)$ goes to $maxSuc_n$, as n goes to infinity. In the short-term, the difference between the average score of AW and of the best scoring object method is bounded by $\sqrt{k/n}$, where k is the number of object methods (i.e., for all prediction games, and all n: $\bar{S}_n(AW) + \sqrt{k/n} \geq maxSuc_n$) (Schurz, 2008). If the number of accessible object methods, k, is large, then the *worst case* short-term performance of AW need not be very good. We will return to this point below.

Access optimality is an important characteristic, in the context of Hume's problem of induction. Hume's problem is easily illustrated within the framework of prediction games. Within prediction games, the character of the event sequence is unconstrained. This means that, regardless of the character of the preceding n–1 events, event n may take any value whatsoever (within [0, 1]). This implies that there is no sure-fire way to exploit the observation of past events in order to make accurate predictions about future unobserved events. Access optimality offers a means to mitigate this problem: Granted that there is no way to ensure good performance within a prediction game, applying AW ensures that one does no worse than the best scoring object method to which one has access.

To date, the only prediction methods that are known to be access optimal are variants of AW (e.g., variants that employ exponentially weighted attractivities). Beyond, this, it is demonstrable that all one-favorite meta-methods, including TTB, are not access optimal. That is, any method that forms its prediction, for each event, by imitating the prediction of a single object method (or cue) is not access optimal. It is also demonstrable that well-known weighting methods such as multiple linear regression and Franklin's Rule (see below) are not access optimal.

It is easy to see why one-favorite meta-methods are not access optimal. Consider a prediction game with ten object methods, and a one-favorite method called "Mono." Suppose that the predictions of the object methods are highly accurate when they are not imitated by Mono, and highly inaccurate when they are imitated by Mono (so that there is a negative correlation between the score of an object method and its being imitated). In such circumstances, the predictions of Mono will be highly inaccurate, while, for each event, the predictions of nine of the ten object methods will be highly accurate. Beyond the theoretical possibility of situations in which one-favorite methods fail to perform well (as illustrated by a simulation presented in Schurz and Thorn, 2016, fig. 3), there is a wide range of naturally occurring situations where one-favorite methods, such as TTB, perform poorly. The problem arises in situations where the payoff for performing a given action is an inverse function of the number of individuals who perform the action. Such cases may arise when the task is to determine where one should go in attempting to gather a seasonal resource (e.g., fish or berries). In such cases, widespread adoption of TTB applied to cues concerning the past productivity of given locations will drive each member of a population of TTBers to attempt to gather resources at the same location, resulting in a poor mean payoff for the TTBers.[4] Similar dynamics may be observed in a wide range of tasks, including market entry problems, career

[2] An alternative variant of AW, called "intermittent AW" in (Schurz & Thorn, 2016), may be formulated by identifying the attractivity of an object method with the difference between its ecological validity and the ecological validity of (intermittent) AW. We here investigate AW, since it is more frugal (cf. Newell, Rakow, Weston, & Shanks, 2004).

[3] The access optimality of AW holds regardless of how AW makes its predictions in cases where all object methods have zero attractivity.

[4] In action games, where actions take the place of predictions, a weighted average of several actions is interpreted as a probabilistic mixture, i.e., the weights are interpreted as the probabilities of emulating respective actions.

choice, cuing problems, route selection, departure time selection, etc. While TTB performs poorly in such tasks, the access optimality of AW ensures good performance. For empirical studies of human performance in such tasks, see Rapoport, Seale, Erev, and Sundali (1998), Rapoport, Stein, Parco, and Seale (2004), and Rapoport, Gisches, Daniel, and Lindsey (2014).

Despite its access optimality, it is possible to raise concerns about the short-term performance of AW, as we did in Schurz and Thorn (2016). In order to form a clearer picture of the concern, consider 'typical' environments where (i) the observed ecological validities of the accessible cues quickly approach their actual (long-run) ecological validities, as the number of observed items increases, and (ii) the average score of cues does not vary according to their use by meta-methods (contrary to the environments described in the preceding paragraph). We call environments meeting conditions (i) and (ii) "non-elusive."

It is possible to distinguish two sorts of non-elusive environment: *compensatory* (where there are methods of linear weighting that outperform TTB), and *non-compensatory* (where no method of linear weighting outperforms TTB) (cf. Martignon & Hoffrage, 1999). Concerns may be raised about the performance of AW in both sorts of environment. *First*, AW will perform worse than TTB, in the short-run, in non-compensatory environments. In such environments, TTB's one-favorite approach is appropriate, and while it is demonstrable that AW will adopt the behavior of a one-favorite method in the long-run (proceeding until only one object method has an attractivity greater than zero), this will take some time. The extent to which TTB will outperform AW, in the short-run, in typical non-compensatory environments is an open question. Using the simulation studies reported below, we attempted to answer this question. *Second*, AW (like TTB) will typically perform worse than *some* other weighting strategies (both in the long and short-run) in compensatory environments. For example, Schurz and Thorn (2016) show that Franklin's Rule (described below) outperforms AW in some compensatory environments. One possible solution to the present problem would be to include a variety of meta-methods that are known to perform well in compensatory environments among the set of methods to which AW has access, thereby permitting AW to emulate another weighting method, different from AW, in appropriate situations. We call this refined version of AW "vAW." (TTB could be adapted in a similar manner.) A residual worry may be raised regarding this proposal: Regardless of the capacity of vAW to emulate the behavior of a well-adapted weighting method in the long-run, the performance of the weighting method will probably exceed the performance of vAW in the short-run. This worry is analogous to the one that arose in the comparison of AW to TTB in typical non-compensatory environments: AW will lose out, in the short-run, in the midst of learning which method it should emulate in the long-run. Once again, we will use the simulation studies reported below to address the magnitude of this problem.

The Simulations

In order to evaluate the short-term performance of AW, we simulated prediction games using data sets characterizing natural environments. In particular, we tested the performance of AW, along with a number of other meta-methods, using the twenty data sets used by Czerlinski, Gigerenzer, and Goldstein (1999) in evaluating the performance of TTB. These data sets are heterogeneous, and representative of a wide range of environments, involving the prediction of city population, attractiveness of persons, high school dropout rates, homelessness rates, mortality rates, house prices, professor salaries, automobile fuel consumption, body fat, fish fertility, mammal sleep duration, biodiversity, rainfall, and atmospheric conditions.[5] As with Czerlinski, Gigerenzer, and Goldstein (1999), we used the data to formulate paired comparison tasks, i.e., tasks where a method must judge which of two objects (e.g., two German cities) has a greater *criterion value* (e.g., population) on the basis of a number of cues (e.g., whether a city has a university).

For each data set, we generated 1,000 prediction games (with finite event sequences).[6] Each prediction game, for a given data set, was generated by forming a random sequence of the set of all pairs of objects that the data set concerned. For example, for the data set that concerned the population of German cities, we generated a random sequence of the pairs of German cities (excluding repeats, e.g., if (Munich, Hamburg) was included (Hamburg, Munich) was not). The sequence of object pairs formed the basis for a sequence of paired comparison tasks. The value, e_i, for the ith element of event sequence was read off of the sequence of object pairs: If the first element of the pair had a greater criterion value, then $e_i = 1$, otherwise $e_i = 0$. The participating object methods were simply the cues for the corresponding data set. Cue values were determined as follows: (i) if the first element of the pair had a higher cue value, then the cue predicted that the first element has a higher criterion value (i.e., $P_i(\text{cue}) = 1$), (ii) if the second element of the pair had a higher cue value, then $P_i(\text{cue}) = 0$, and (iii) if the cue values were identical, then the result was an ersatz prediction, with $P_i(\text{cue}) = 0.5$. Notice that the preceding manner of defining cue predictions leaves open the possibility that any given cue's validity, for the set of all events, is less than 0.5. But if the validity of a cue is very low (and below 0.5), then a meta-method such as TTB may wish to form its predictions by predicting counter to the cue. To accommodate this possibility, we allowed meta-methods to consider the 'counter-cue', corresponding to any given

[5] We used the data sets with dichotomized cue values that are available at: http://www-abc.mpib-berlin.mpg.de/sim/Heuristica/

[6] Given the standard error of the mean (not reported) for each of the mean values reported below, samples of 1,000 were needlessly large, according to reasonable standards for reporting simulation results (Bindel & Goodman, 2009).

cue, that is: (i) if a cue predicts 0, its counter-cue predicts 1, (ii) if a cue predicts 1, its counter-cue predicts 0, and (iii) if a cue predicts 0.5, its counter-cue also predicts 0.5.

The participating meta-methods included AW and TTB, along with an array of other methods, including Franklin's Rule, the Minimalist, multiple linear regression (MLR), and Dawes' Rule. The latter three methods, along with TTB, were considered by Czerlinski, Gigerenzer, and Goldstein (1999).[7] Finally, we considered a variant of AW, called "vAW," that treats Franklin's Rule, MLR, and Dawes' Rule as if they were accessible object methods. The methods made their predictions as follows:

AW. The data sets employed in our simulations result in many cases where some of the relevant cues do not deliver genuine predictions. We introduced a slight modification of AW, in order to improve its performance in such situations: In determining AW's prediction (i.e., $P_n(AW)$), according to the equation described above, we assumed that, in any given round, m ranges only over those cues (and counter-cues) that delivered a genuine prediction in that round. In the case where no cue delivers a prediction AW predicts 0.5. AW remains access optimal with these modifications, so long as we restrict ourselves to prediction games where, for each cue, the average score of AW's predictions in cases where the cue was attractive but made no genuine prediction exceeds 0.5, in the long-run.

vAW. As AW, save that Franklin's Rule, MLR, and Dawes' Rule are also treated as accessible object methods

TTB. In each round n, TTB forms its prediction by (i) ordering the accessible cues and counter-cues by their observed ecological validity as of round n–1, and (ii) emulates the prediction of the first cue (or counter cue) in the ordering that delivers a genuine prediction. If no cue makes a prediction or if all cues have undefined ecological validities (as in round one), TTB predicts 0.5.

The Minimalist (Min). As TTB, save that the cue order is determined at random.

Franklin's Rule (FR). In each round n, Franklin's Rule predicts a weighted average of the genuine predictions of cues (or counter-cues) whose observed ecological validity is at least 0.5. The weight for each cue/counter-cue, in round n, is proportional to its ecological validity as of round n–1. As with TTB, if no cue makes a prediction or if all cues have undefined ecological validities (as in round one), Franklin's Rule predicts 0.5.

Dawes' Rule. As Franklin's Rule, save that each cue is assigned equal weight.

MLR. For each round n, the predictions of MLR were determined by finding the ordinary least squares multiple regression model for predicting the criterion values from the cue values, based on the objects that had been observed prior to round n. MLR's prediction about which of two objects has a higher criterion value was determined by which object had a higher predicted criterion value according to the regression model. For rounds where there is insufficient data to find a regression model, MLR predicts 0.5.

[7] We followed Czerlinski, Gigerenzer, and Goldstein (1999) in including MLR as a meta-method. In an extended study, we plan to also include logistic regression.

AW's Short-term Performance

Within our simulations, there was no considerable lag in the performance of AW in comparison to TTB, or in the performance of vAW in comparison to the best performing alternative weighting strategy (which was generally MLR). Figure 1 gives a sense of these results, showing the average scores of the considered meta-methods at various rounds of the simulated prediction games (i.e., the averages for the 20,000 prediction games, based on 1,000 simulations for each of the 20 data sets). For the sake of readability, we do not plot the values for Franklin's Rule, whose mean performance was similar to TTB and AW.

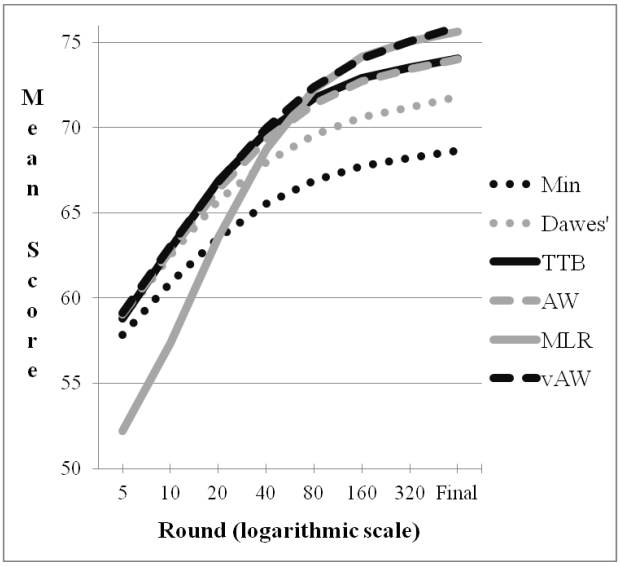


Figure 1: Average scores for the competing meta-methods, at various rounds.

In assessing the gravity of the concerns raised in the preceding section, it is important to note that the performance of AW closely matches that of TTB. While TTB usually performed slightly better than AW, there were only four data sets where its mean performance exceeded that of AW by more than 0.01, at any stage. This indicates that the access optimality of AW does *not* come at the expense of short-term performance, in comparison with TTB. Similarly, on average, vAW closely approximated the performance of the best performing method, from the early through the later stages of respective games. This indicates that a simple modification of AW enables performance that quickly matches the performance of alternative weighting methods, when appropriate.

For purposes of comparison with the results of Czerlinski, Gigerenzer, and Goldstein (1999), we recorded the average frugality of the competing meta-methods (i.e., the number of cue value pairs that the meta-methods needed to access before making a prediction). Table 1 shows the results.

Subsequent to observing the similarity of AW and TTB in terms of mean scores and mean frugality, we formed the hypothesis that, in non-elusive environments, the behavior of AW was very similar to that of TTB. In the following section, we report some results that support this hypothesis.

Table 1: Mean frugality of the competing meta-methods.

Method	Frugality
Min	2.2
TTB	2.3
AW	2.5
vAW	7.1
MLR	7.3
Dawes'	7.4
FR	7.4

The Near Indiscernibility of AW and TTB in Non-elusive Environments

In attempting to determine what judgment strategies humans use in paired comparison tasks, two sorts of data have been most important: *outcome patterns* and *process tracing* (cf. Bröder, 2012). Outcome patterns consist of data concerning the responses that subjects provide in the face of given paired comparison tasks (Bröder, 2003; Bröder & Schiffer, 2006; Rieskamp & Otto, 2006; Rieskamp, 2008). Rather than considering responses, process tracing monitors information acquisition patterns (Newell & Shanks 2003; Newell, Weston, & Shanks 2003). Information concerning response times (Bröder & Gaissmaier, 2007), and subject self-reporting (Walsh & Gluck, 2016), may also be relevant in determining what judgment strategies humans use.

By the analysis of outcome patterns, and (to a lesser extent) process tracing, some psychological studies are thought to corroborate the claim that TTB plays *some* role in human reasoning. Based largely on studies of Bröder (2003), Rieskamp and Otto (2006), and Rieskamp (2008) (that investigated human behavior in non-elusive environments), the received view is that human beings are adaptive in the strategies they use in making predictions, and that, in appropriate environments, subjects are disposed to (learn to) use TTB in making predictions. Note, however, that the theoretical conclusions of the above mentioned studies are based on maximum likelihood techniques (Bröder, 2003) and model fitting (Rieskamp & Otto, 2006; Rieskamp, 2008), and that these methods are based on comparative evaluations of candidate hypotheses. This means (as is acknowledged by the authors of the aforementioned studies) that data that strongly supports one hypothesis (e.g., that subjects are using TTB to accomplish a given prediction task) among a given pool of hypotheses, doesn't necessarily support the hypothesis over others that were not considered (e.g., the hypothesis that subjects are using AW to accomplish the task). Our general point is certainly not a new one, and it is not our intent to disparage the use of comparative methods of evaluation. Our only intent is to suggest that existing data does not tell in favor of the adaptive use of TTB over that of AW. In order to support the present claim, we report results bearing on the degree of confluence between TTB and AW, both in terms of their predictions (outcome patterns), and informational demands (process tracing).

The data reported in Tables 2 and 3 is from the simulations described above. Table 2 reports the percentage of trials in which the predictions of respective meta-methods agreed with the predictions of TTB. The averages reported here are the averages of the averages for 1,000 simulations for the 20 environments. In other words, we first collected the averages for each of the 20 environments. We then took the averages of those averages. Table 2 reports the average percentage of trials in which the informational demands of respective meta-methods (i.e., the profile of cues accessed) were identical to the informational demands of TTB. The averages reported are, again, the averages of the averages for 1,000 simulations for the 20 environments.

Table 2: Mean percentage of predictions identical with TTB.[8]

Method	Agreement w/ TTB
AW	98%
FR	93%
MLR	85%
Dawes'	83%
Min	83%

Table 3: Mean percentage of cases where cues accessed were identical with TTB.

Method	Agreement w/ TTB
AW	93%
FR	13%
MLR	12%
Dawes'	12%
Min	19%

As is evident from the data, the behavior of AW is very similar to that of TTB in non-elusive environments. It is of interest to note, for example, that the degree of overlap between AW and TTB is far greater than the degree of predictive fit between subject behavior and the best fitting models that have been offered in the literature, such as that of Rieskamp and Otto (2006). Absent the explicit intention

[8] If we require that all of AW's predictions are rounded to the nearest integer, then the percentage agreement with TTB is 99.5.

to create a prediction task in which TTB and AW produce different predictions, it is unlikely that one would produce a task that could be used as evidence in favor of the adaptive use of TTB over AW, and vice versa. Matters are complicated by the fact that human subjects appear to be adaptive in the methods that they deploy, employing weighting methods by default (Bröder, 2003; Rieskamp & Otto, 2006; Rieskamp, 2008).

Conclusions

The results of our simulations suggest that there is no reason to worry about the short-term performance of AW, in comparison to TTB. Similarly, a variant of AW can be formulated that performs well in both compensatory and non-compensatory environments.

Our simulations also suggest that the behavior of AW is nearly indiscernible from that of TTB, in non-elusive environments. It seems, then, that the possibility cannot be excluded that AW plays some role in human reasoning, inasmuch as there is data consistent with the hypothesis that subjects are adaptive users of TTB, in non-elusive environments. Considerations of 'prior intuitive plausibility' may favor the hypothesis that subjects use TTB rather than AW. On the other hand, the foolproof nature of AW, underwritten by its access optimality, suggests that AW is more adaptive than TTB: Inasmuch as we expect human cognition to be adapted to its environment (an environment in which the application of one-favorite methods can lead to catastrophic failure), there is some reason to expect that something like AW plays some role in human cognition. Such musings are, of course, an impetuous (rather than a substitute for) proper empirical studies, which have yet to be conducted.

Acknowledgments

Work on this paper was supported by the DFG Grant SCHU1566/9-1 as part of the priority program "New Frameworks of Rationality" (SPP 1516). For valuable help we are indebted to Leandra Bucher, Ralph Hertwig, Chris Lange, Laura Martignon, and Arthur Paul Pedersen.

References

Bindel, D., and Goodman, J. (2009). *Principles of Scientific Computing*. Retrieved from http://www.cs.nyu.edu/courses/spring09/G22.2112-001/book/book.pdf

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 611-625.

Bröder, A. (2012). The quest for take the best - Insights and outlooks from experimental research. In Todd, P., & Gigerenzer, G. (Eds.), *Ecological rationality: Intelligence in the world*, New York: Oxford University Press.

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladapative routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 904-918.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games.* Cambridge: Cambridge Uniersity. Press.

Czerlinski, J., Gigerenzer, G., & Goldstein, D.G. (1999). How good are simple heuristics? In: Gigerenzer et al. (Eds.) *Simple heuristics that make us smart*, (pp. 97-118).

Gigerenzer, G., Todd, P.M., & The ABC Research Group (Eds.). (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? In: Gigerenzer et al. (Eds.). *Name of the book* (pp. 119-140).

Newell, B.R., Rakow, T., Weston, N.J., & Shanks, D.R. (2004) Search Strategies in Decision Making: The Success of "Success." *Journal of Behavioral Decision Making*, 17, 117-137.

Newell, B.R., & Shanks, D.R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 53-65.

Newell, B.R., Weston, N.J., & Shanks, D.R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best." *Organizational Behavior & Human Decision Processes*, *91*, 82-96.

Rapoport, A., Gisches, E., Daniel, T., & Lindsey R. (2014). Pre-trip information and route choice decisions with stochastic travel conditions: Experiment. *Transportation Research Part B*, 68, 154-172.

Rapoport, A. Seale, D.A., Erev, I., & Sundali, J.A. (1998). Equilibrium play in large market entry games. *Management Science*, 44, 119-141.

Rapoport, A., Stein, W.E., Parco, J.E., & Seale, D.A. (2004). Equilibrium play in single-server queues with endogenously determined arrival times. *Journal of Economic Behavior and Organization*, 55, 67-91.

Rieskamp, J., & Otto, P. (2006). SSL: A theory of how people learn to select strategies. Journal of Experimental Psychology: General, 135(2), 207-236.

Rieskamp, J. (2008). The importance of learning when making inferences. *Judgment and Decision Making*, 3(3), 261-277.

Schurz, G. (2008). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science, 75,* 278-305.

Schurz, G., & Thorn, P. (2016). The Revenge of Ecological Rationality: Strategy-Selection by Meta-Induction Within Changing Environments. *Minds and Machines*, 26(1-2), 31-59.

Todd, P. M., & Gigerenzer, G. (Eds.). (2012). *Ecological rationality: intelligence in the world.* New York: Oxford University Press.

Walsh, M.M., & Gluck, K.A. (2016) Verbalization of Decision Strategies in Multiple-Cue Probabilistic Inference. *Journal of Behavioral Decision Making, 29,* 78-91.

A Perception-Based Threshold for Bidirectional Texture Functions

Banafsheh Azari (banafsheh.azari@uni-weimar.de)
CogVis/MMC, Faculty of Media, Bauhaus-Universität Weimar, Bauhausstraße 11, 99423 Weimar, Germany

Sven Bertel (sven.bertel@uni-weimar.de)
Usability Research Group, Faculty of Media, Bauhaus-Universität Weimar, Bauhausstraße 11, 99423 Weimar, Germany

Charles A. Wuethrich (charles.wuethrich@uni-weimar.de)
CogVis/MMC, Faculty of Media, Bauhaus-Universität Weimar, Bauhausstraße 11, 99423 Weimar, Germany

Abstract

For creating photorealistic images, Computer Graphics researchers introduced Bidirectional Texture Functions (BTFs), which use view- and illumination-dependent textures for rendering. BTFs require massive storage, and several proposals were made on how to compress them, but very few take into account human perception. We present and discuss an experimental study on how decreasing the texture resolution influences perceived quality of the rendered images. In a visual comparison task, observer quality judgments and gaze data were collected and analysed to determine the optimal downsampling of BTF data without significant loss of their perceived visual quality.

Keywords: Perceived image quality; realistic rendering; threshold in image perception; eye tracking.

Introduction

One of the main aims of Computer Graphics is the simulation of the complex reflection behaviour of real world materials. Among different types of materials, particular importance is given to fabrics. Graphical simulations of fabrics are used not only in interior design and architecture, but also increasingly in the clothing, car, film, and computer gaming industries.

Fabrics possess highly complex reflection behaviour, as reflection of the incoming light changes dramatically from material to material, depending, among other factors, on meso- and micro-structures of the thread, an example of which is shown in Figure 1, on the type of weaving, which influences the position of the thread in the fabric, on the interreflections between the components of the fabric, and on surface and subsurface scattering of light. Fabrics exhibit not only simple reflection characteristics, such as diffuse and specular reflection, but are characterised also by thread-dependent highlights and self shadowing, as well as anisotropic reflection.

To exactly simulate the correct reflection behaviour of fabrics, the paths of light within and on the surface of the material would have to be computed. Given the number of individual components of a thread, such task is computationally not feasible. Bidirectional Texture Functions (BTFs; Dana et al. (1999)) represent an alternative solution to this problem. A BTF contains reflectance information of points on a surface under fixed lighting and viewing conditions. It involves a function which depends on coordinates in texture space (x_p, y_p) on the surface of a simulated object, as well as the spherical angles of a vector from the viewer to the surface (θ_o, ϕ_o) and a vector to the illuminant (θ_i, ϕ_i).

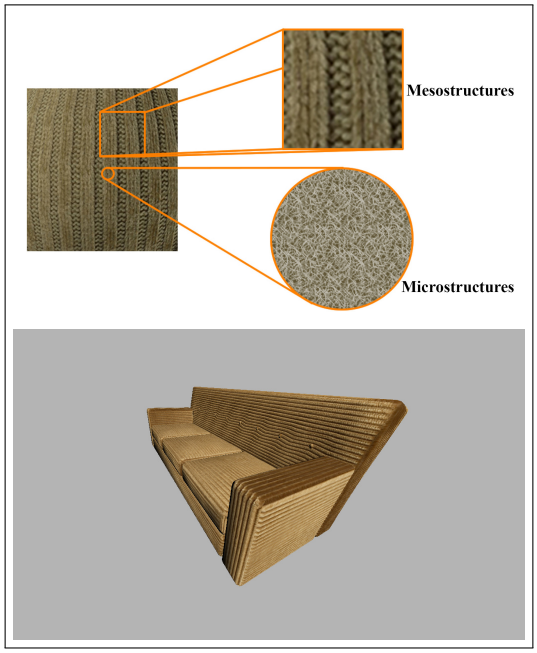


Figure 1: *Above:* Meso- and micro-structures of a woollen fabric. *Below:* An example of the stimuli used in our study.

In practical use, BTFs rely on large collections of digitally acquired pictures of a material (e.g., of a woollen fabric) that were taken for ranges of discretely varied illumination and viewing angles. When a simulation of the material needs to be computed to texturise a virtual object's surface, viewer and illuminant vectors are used to pick the matching pictures from this collection. The texture of an object then results from an interpolation of these (Nicodemus et al. (1977)).

A severe disadvantage of BTFs lies in the sheer size of picture collections needed, as they contain one photograph for each combination of viewing and illumination angles. The disadvantage is particular acute for the purposes of real time rendering, as the entire collection of pictures needs to be kept in the computer memory. Various past projects have therefore focussed on efficient compression methods for BTFs (including reflectance models based on linear factorization and pixelwise bidirectional reflection distribution functions, in short BRDFs, which are the general reflection model from which BTFs are derived, Filip and Haindl (2009)).

While the existing approaches are often technically well motivated, we believe that, before starting to choose how and how strongly to compress BTF data, it makes sense to first take a step back and see how the human observer perceives and judges compressed and non-compressed BTF textures in comparison tasks. Specifically, we look at BTF-based synthetic renderings of three–dimensional objects and ask: when does using high-resolution textures make sense because the high resolution leads to a perceived increase in texture quality? And when can one do just as well with lower resolution textures without perceived loss in quality? In this contribution, we present and discuss an investigation aimed at locating a robust threshold for downsampling BTF images without loosing perceptual quality. Information about the location of such a threshold is not only of importance to a better understanding of visual perception of textures, especially in object comparison tasks, but also of importance for developing novel data compression methods in synthetic rendering.

In the next section, we will review relevant related work. We will then describe our method of experimentation, which involves quality comparison tasks with pairs of texturized objects of varying BTF quality levels and varying exposure times. Gaze data was collected to aid visual comparison strategy detection. The presentation of study results is then followed by a discussion, conclusions, and an outlook.

Previous work

Compression methods for BTF data have been studied for many years in order to accelerate rendering and to compress data. However, only rarely the focus was on the perceived quality of the results of compression. Fleming et al. (2003) studied how humans perceive reflections on surfaces, while Lawson et al. (2003) demonstrated the importance of view changes in synthetic picture matching tasks. te Pas and Pont (2005a) showed that differences in the microstructure of a material are hard to distinguish from differences in the illumination, and that light source direction estimation depends on the material's bidirectional reflection distribution functions or BDRFs (te Pas and Pont (2005b)).

Work by Pellacini et al. (2000) introduced a new light reflection model for image synthesis based on experimental studies of surface gloss perception. Two experiments were conducted to explore the relationships between the physical parameters used to describe the reflectance properties of glossy surfaces and perceptual dimensions of glossy appearance. Psychophysical tests by Mcmillan et al. (2003) showed consistent transitions in perceived properties between interpolate and extrapolate BRDFs in the space of acquisition.

The accurate reproduction of material structures that can be achieved by using measured BTFs was investigated in Meseth et al. (2006), while Filip et al. (2008) performed a psychophysical study to optimize sparse sampling of BTFs data. In a further study, Filip et al. (2009) assessed different uniform reduced samplings of BTF data based on azimuthal angles of view and illumination as well as on elevation angles.

Few of the existing approaches include an investigation of viewers' gaze behavior while viewing the rendered images. A notable exception is the work by Filip et al. (2009) in which location, duration, and frequency of fixations were recorded. Fixation data was used to analyze strategies of the subjects over the course of the experiment (e.g., did locations and durations of fixations change as the study progressed? Both were found to be the case.).

In short, previous work focussed on the influence of light, viewing, material reflectance, shape, and angular sampling density of BTF data. In this contribution, we investigate the influence on perceived image quality that the *size of the individual BTF texture pictures* has, based on which a synthetic object's texture is interpolated. This variable has not been addressed previously. Our aim is to find a threshold for downsampling BTF resolution — that is, for reducing the image size of the individual BTF textures — without any perceived degradation in the quality of the rendered image. Similar to the procedure by Filip et al., we will collect gaze data to aid the detection of visual strategy and its change.

Method

In a pilot study using different self shadowing fabrics, like corduroy and wool, available in the BTF database of the University Bonn[1] we established that there are no differences in gaze behavior or perceived quality jugments between fabrics. We therefore decided to here focus on the corduroy dataset, which we will refer to as *Cord-256*, as its texture pictures are 256x256 pixels.

We then generated two new datasets by downscaling the *Cord-256* set through bilinear interpolation to respective resolutions of 128x128 pixels (*Cord-128*) and 64x64 pixels (*Cord-64*). For each of the three texture data sets, a three-dimensional textured model of a sofa was rendered through the standard BTF rendering method at a screen resolution of 1920x1084 pixels (compare Figure 1, below). The sofa model was oriented for display to the viewer to present textured parts across a large range of picture depth.

We chose a sofa object model for three main reasons: first, to present an everyday object that viewers are familiar with and instantly recognize. Second, to have an object with a structured surface and composition (e.g., individual buttons, cushions, etc.). This is important in order to ensure that a large set of fitting BTF pictures will be selected as basis for the object's texture, with widely varying illumination and viewing angles. And, third, a sofa is a type of object for which a cord texture would be commonly found.

Stimulus Pairs

Pairs of the rendered images displayed in full screen, native resolution mode were used as experimental stimuli. Each pair consisted of a sequentially presented rendering using two of the three texture resolutions as shown in Table 1. A total of 72 image pairs were shown to test subjects.

[1] http://btf.cs.uni-bonn.de/.

Table 1: Image pairs as experimental stimuli.

First Image	Second Image
Cord-256	*Cord-64*
Cord-256	*Cord-128*
Cord-128	*Cord-256*
Cord-128	*Cord-64*
Cord-64	*Cord-256*
Cord-64	*Cord-128*

Table 2: Answer posibilities.

1	First image has higher quality.
2	Second image has the higher quality.
=	The images have the same quality.
None	Subject is not sure.

The experiment was performed in three blocks of 24 image pairs each, between which image exposure time was varied. Exposure time per image was either 1000, 2000, or 3000 milliseconds (ms), respectively labeled as *short*, *medium* and *long* test conditions. Presentation order of the three blocks was balanced across subjects based on a Latin square design. Our rationale behind introducing variation of image exposure time was to test for effects it may have on comparative perceived image quality. It seems possible that, for pairs of different images, longer exposures could lead to higher frequencies of detecting that a difference exists.

Presentation of images in each pair was separated by 200 ms. After the presentation of the second image in a pair, subjects had 3000 ms to make a decision about the comparative image quality within the pair: was the first or second image of better visual quality? Or were the two images of the same visual quality? Responses were given on a three-key keyboard and were possible at any time after the start of the presentation of the second image. Subjects were also instructed that they could choose not to press any button if they felt unsure about the comparison. Please see Table 2 for an overview. When looking at the six image pairs in Table 1, it becomes clear that all pairs are different and that, consequently, any judgment that a pair shows that same image quality will be incorrect. However, subjects were not previously instructed that no same-quality pairs would be shown. After the decision time of 3000 ms had lapsed, the next image pair was automatically presented.

Experimental Setup

The images were presented on a 24-inch monitor with a resolution of 1920x1080 pixels at a distance of 70 cm from the viewer. The screen measured 22.35x15.80 inches and subtended approximately 33 degrees of visual angle. Due to the texture pattern, the minimal texture detail (i.e., for the parts of the sofa at the greatest depth in the image) had a cycle of 4 pixels, which means a subtended angle for a viewer of about 6 cycles per minute of a degree of arc.

An SMI RED250 remote eye tracking system was used in binocular mode with 250 Hz fixation detection, in order to record subjects' fixation behavior. *SMI BeGaze 2.4* software was used for subsequent analysis of gaze data.

Subjects. A total number of 20 subjects, 12 males and 8 females, participated in the experiment. Subjects were undergraduate or graduate students or department members in Computer Science or Civil Engineering, and they were not informed about the purpose of the experiment prior to conducting it. The age of the test subjects ranged from 22 to 39 years ($mean = 30.5$). Subjects had normal or corrected-to-normal visual acuity.

Procedure. Test subjects were seated in front of the monitor and eye tracker, introduced to the setup and to the experimental procedure, including the answer options. Before the start of the experiment, subjects were asked to read and sign a declaration of informed consent. Subjects could abort the experiment at any time and were guaranteed anonymous treatment of all collected data. They were familiarized with the used sofa images through a preliminary test round with eight image pair comparisons, the results of which were discarded for the subsequent analysis. Then, the subjects were calibrated on the eye tracker and the first of the three test blocks was presented. Calibration was repeated before each subsequent block. Each subject needed about 30 minutes to complete all three blocks.

Results

The section consists of two parts: an analysis of *subject performance* (i.e., the subjects' ability to judge image quality differences for the six pairs of Table 1) and an analysis of *gaze data* (locations, frequencies, and durations of fixations).

Subject Performance Analysis

The first three columns of Table 3 illustrate the numbers of correct and incorrect answers given for each of the six image pairs. Incorrect answers are provided as incorrect *equal* answers and as other incorrect answers. Looking at the numbers suggests that differences exist between the six pair conditions for numbers of correct answers. A Friedman ANOVA confirms the existence of significant differences ($\chi^2(2) = 41.989, p < 0.001, r = 0.952$). Two groups of pair comparisons exist, irrespective of presentation order: as a first group, *Cord-256* and *Cord-128* with lower performance, as a second group *Cord-256* and *Cord-64* as well as *Cord-128* and *Cord-64*, with higher performance. The same groups can be formed for the number of incorrect *equal* answers ($\chi^2(2) = 73.935, p < 0.001, r = 0.920$). The first group has many more incorrect *equal* answers than the second. A breakdown of performance and incorrect *equal* counts for the three exposure duration conditions (*short*, *medium*, *long*) revealed no significant differences.

In order to check for training effects, we compared numbers of correct answers for the three blocks (first: 268, sec-

Table 3: Frequencies of correct answers, incorrect equal-quality answers, and other incorrect answers (accumulated over all 20 subjects; sum of answers per pair: 240); average fixation durations and average fixation frequencies per image pair presentation.

	# correct	# equal (incorrect)	# other (incorrect)	av. fix. dur. [ms]	av. fix. freq.
Cord-256 / Cord-128	24	167	49	386.48	2.32
Cord-128 / Cord-256	39	156	45	398.51	2.25
Cord-64 / Cord-256	195	25	20	404.14	2.21
Cord-256 / Cord-64	193	24	23	412.01	2.20
Cord-64 / Cord-128	193	25	22	418.08	2.14
Cord-128 / Cord-64	196	19	25	407.59	2.20

ond: 274, third: 297). For each block, a total of 480 answers were collected across all 20 participants. A Friedman test revealed significant differences between the blocks ($\chi^2(2) = 6.195, p < 0.05, r = 0.952$). A comparisons of means shows a positive training effect.

Gaze Fixation Analysis

We next analyzed subjects' gaze fixation distributions across the sofa image in order to assess whether differences exist for different exposure durations and for different image pair comparisons. Fixation counts for cells in an overlaid 16x16 grid are shown in Figure 2 (upper part) for nine conditions. Fixation count patterns between any pair of these nine conditions are significantly correlated with all $r > 0.850$ and $p < 0.001$.

Table 4 shows average fixation duration (AFD, in ms) and fixation frequencies (FF). For the three BTF resolution conditions, a Friedman ANOVA shows significant differences in FF ($\chi^2(2) = 6.495, p = 0.039, r = 0.697$) and AFD ($\chi^2(2) = 7.777, p < 0.03, r = 0.649$). AFDs decrease and FFs increase from lower to higher resolution textures. For first and second images, a Wilcoxon test shows a significantly lower FF on the second image ($Z = 3.062, p < 0.003, r = 0.684$) and a longer AFD on the second ($Z = 2.420, p = 0.025, r = 0.541$). For the first, second, and third blocks we find an increase in AFDs ($\chi^2(2) = 8.527, p = 0.045, r = 0.623$) and a decrease in FFs ($\chi^2(2) = 8.954, p = 0.011, r = 0.608$).

In order to check whether the subjects' fixation location patterns were driven by visually perceivable differences between images in our BTF image pairs, we employed the Visible Difference Predictor (VDP) (Mantiuk, Daly, Myszkowski, and Seidel (2005)). VDP simulates low level human perception for known viewing conditions (in our case: a resolution of 1920x1080 pixels at an observer's distance of 0.7m). The last row of Figure 2 shows the visually perceivable differences per image pair (irrespective of presentation order) as predicted by VDP. Correlations between VDP results and respective fixation location patterns can be seen in Table 5 (as averaged over exposure durations; displayed in the columns above each VDP result in Figure 2). The results confirm the two groups of image pairs found in the subject performance analysis: (1) a weak correlation for *Cord-256* and *Cord-128* pairs and (2) strong correlations for the pairs within the group of *Cord-256* and *Cord-64* as well as *Cord-128* and *Cord-64*. Lastly, existence of the two groups is further supported by average fixation durations and fixation frequencies for the individual image pairs as seen in the right-hand part of Table 3. AFDs in the first group are significantly lower than in the second ($\chi^2(2) = 73.935, p < 0.001, r = 0.920$), while FFs are significantly higher ($\chi^2(2) = 41.989, p < 0.001, r = 0.952$).

Table 4: Average Fixation Duration[ms] (AFD) and Fixation Frequency (FF) for different image quality levels, first and second images, and for blocks.

	AFD[ms]	FF
Cord-256	429.78	2.24
Cord-128	436.45	2.23
Cord-64	444.90	2.17
First Image	422.43	2.38
Second Image	493.09	2.05
First Block	375.08	2.33
Second Block	405.25	2.30
Third Block	448.71	2.02

Table 5: Correlations between VDP results and fixations independently of exposure durations and presentation order.

	r	p
Cord-256 _ Cord- 64	0.808	0.0001
Cord-128 _ Cord- 64	0.753	0.0001
Cord-256 _ Cord-128	0.015	0.0175

Discussion

The results show that two groups of image comparisons exist in our study. The first group consists of comparisons between *Cord-256* and *Cord-128*. For this group, subjects are largely unable to perceive existing differences between the images. Instead, they frequently judge the pair to consist of the same image. The higher average FFs and lower AFDs in this group suggest more visual search for existing differences. The VDP model predicts few visually perceivable differences for image pairs in this group.

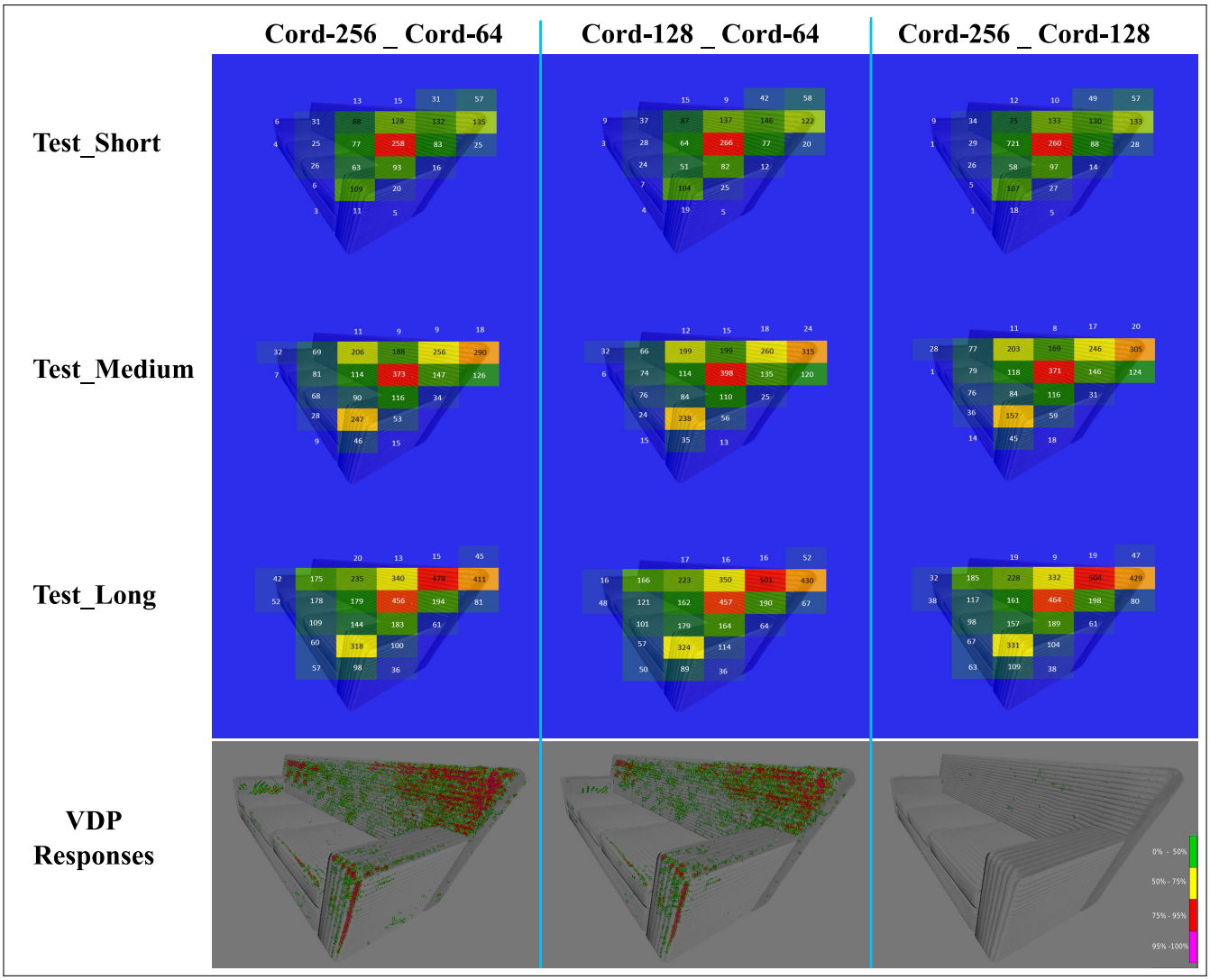


Figure 2: Fixation count in different test duration and responses of visual difference predictor for tested image pairs.

The second group consists of comparisons between *Cord-256* and *Cord-64* as well as between *Cord-128* and *Cord-64*. For this group, subjects are largely able to see the differences among the pairs. Occurrences of incorrectly labeling pairs as *equal* are few. The lower FF counts and higher AFDs suggest that subjects are better able to concentrate on informative locations (i.e., on locations at which the images within a pair differ). The VDP model predicts a higher number of differences which are also detectable with higher probability.

A comparison between the fixation location patterns between the first and the second group reveals that, irrespective of group, subjects seem to fixate on similar locations, and do so with similar frequencies. One conclusion is that they employ similar strategies while inspecting image pairs of any of the six types. VDP predictions differ markedly between the groups. We observed strong correlations between locations of predicted visually perceivable differences and observed fixation patterns only for the second group of comparisons. We interpret this as evidence for subjects' ability to pick up on differences in the second group and use information about the location of these differences for image comparisons. A significant, albeit very weak, correlation exists for the first group.

When comparing AFDs and FFs between the three BTF image qualities, it seems likely that low image quality leads to less visual search, suggesting that subjects are fast at discerning features that hint at low quality.

AFD was lowest for the first block and then increased over the course of the experiment, while the average FF decreased. This pattern is in line with the one presented in (Over et al. (2007)) and suggests that subjects may have applied a coarse-to-fine approach during visual search. Within the first comparisons, subjects may notice locations at which differences between images of different visual quality are located, leading to more fixations at them. This may differ for behavioral patterns in the beginning, when subjects spend more

time carefully searching for differences among image pairs, resulting in shorter and a larger number of fixations.

Longer AFDs in the second image in a pair compared to the first indicate that by the time subjects look at the second image they already have formed hypotheses about where to look for differences.

There were no differences in performance and gaze fixation for different exposure durations.

The main purpose of this study was to locate a threshold for robust, effective BTF compression based on a downsampling of BTF pictures. Above the threshold, differences between pictures are not visually perceivable by a human observer. Our results clearly indicate that differences between *Cord-256* and *Cord-128* lie above such threshold, while differences between *Cord-256* and *Cord-64* as well as between *Cord-128* and *Cord-64* lie below it.

The results are likely to apply to all self shadowing fabrics.

Conclusion

The results of our study narrowed the bracket in which the threshold is located that separates visually perceivable differences in BTF renderings from those that are not. Consequently, we can now suggest a perception-based criterion for downscaling BTFs. A result for image synthesis is that, above the threshold, the lowest texture resolution available can be used without visually perceivable degradation of image quality. This allows to significantly reduce computer memory usage in BTF rendering.

A logical next step would be to conduct a localized search within this established bracket, that is, between *Cord-128* on one side and *Cord-64* on the other, since our study showed that observers cannot distinguish between *Cord-256* and *Cord-128*.

In the future, we also plan to look for ability- and/or skill-dependent differences in the ability to distinguish BTFs at different quality levels. We have already conducted pilot studies with groups of engineers and artists.

In general, there are few studies on perceptual measures of rendering algorithms. This study is a first step in this direction.

Also, this study could open up new research insights for the field of perception of textures of real objects, especially in object comparison tasks. For example, future questions that can be addressed could relate to the categorization of textures in object perception, either general or with regard to group-dependent or individual differences, to effects of attention in object texture perception, or to effects of expertise which may be acquired through completing series of object texture comparisons similar to the ones employed in this study.

Acknowledgments

We would like to thank our test subjects for participating, Jakob Gomoll for helping in implementing and conducting the pilot study and Stefanie Wetzel for the constructive discussions.

References

Dana, K. J., Van Ginneken, B., Nayar, S. K., & Koenderink, J. J. (1999). Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, *18*(1), 1–34.

Filip, J., Chantler, M. J., & Haindl, M. (2008). On optimal resampling of view and illumination dependent textures. In *Proceedings of the 5th symposium on applied perception in graphics and visualization* (pp. 131–134).

Filip, J., Chantler, M. J., & Haindl, M. (2009). On uniform resampling and gaze analysis of bidirectional texture functions. *ACM Transactions on Applied Perception (TAP)*, *6*(3), 18.

Filip, J., & Haindl, M. (2009). Bidirectional texture function modeling: A state of the art survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(11), 1921–1940.

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, *3*(5), 3.

Lawson, R., Bulthoff, H. H., & Dumbell, S. (2003). Interactions between view changes and shape changes in picture-picture matching. *PERCEPTION-LONDON-*, *32*(12), 1465–1498.

Mantiuk, R., Daly, S. J., Myszkowski, K., & Seidel, H.-P. (2005). Predicting visible differences in high dynamic range images: model and its calibration. In *Electronic imaging 2005* (pp. 204–214).

Mcmillan, L., Smith, A. C., Matusik, W., & Matusik, W. (2003). A data-driven reflectance model. In *in proc. of siggraph.*

Meseth, J., Müller, G., Klein, R., Röder, F., & Arnold, M. (2006). Verification of rendering quality from measured btfs. In *Proceedings of the 3rd symposium on applied perception in graphics and visualization* (pp. 127–134).

Nicodemus, F. E., Richmond, J. C., Hsia, J. J., Ginsberg, I. W., & Limperis, T. (1977). *Geometrical considerations and nomenclature for reflectance* (Vol. 160). US Department of Commerce, National Bureau of Standards Washington, DC, USA.

Over, E., Hooge, I., Vlaskamp, B., & Erkelens, C. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, *47*(17), 2272–2280.

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the 27th annual conference on computer graphics and interactive techniques* (pp. 55–64).

te Pas, S. F., & Pont, S. C. (2005a). A comparison of material and illumination discrimination performance for real rough, real smooth and computer generated smooth spheres. In *Proceedings of the 2nd symposium on applied perception in graphics and visualization* (pp. 75–81).

te Pas, S. F., & Pont, S. C. (2005b). Estimations of light-source direction depend critically on material brdfs. *Perception ECVP abstract*, *34*, 0–0.

Influencing Categorical Choices Through Physical Object Interaction

Nicholas J. Shipp (n.j.shipp@herts.ac.uk)
Department of Psychology, University of Hertfordshire
Hatfield, AL10 9AB, UNITED KINGDOM

Frédéric Vallée-Tourangeau (f.vallee-tourangeau@kingston.ac.uk)
Department of Psychology, Kingston University,
Kingston upon Thames, KT1 2EE, UNITED KINGDOM

Susan H. Anthony (s.h.1.anthony@herts.ac.uk)
Department of Psychology, University of Hertfordshire
Hatfield, AL10 9AB, UNITED KINGDOM

Abstract

Recent research has shown that action knowledge influences categorical decisions (Borghi, Flumini, Natraj & Wheaton, 2012; Chao & Martin, 2000; Iachini, Borghi & Senese, 2008; Kalénine, Shapiro, Flumini, Borghi & Buxbaum, 2013). Shipp, Vallée-Tourangeau, and Anthony, (2014) showed that action influences categorisation in a forced-choice triad task when combined with taxonomic information and presented within a functional context. The present experiment examined whether participants would be more likely to match items in a triad task based on shared actions following priming with the functional actions of the objects. Participants engaged in the triad task used in Shipp et al. after a priming phase where they either interacted with a series of objects for their functional capacity (Action Priming), grouped them into categories (Taxonomic Priming) or moved them from one table to another (Movement Priming). Items within the triads were presented as an image either on a white background (context-lean condition) or as a functional scene with the object being used by an agent (context-rich condition). Consistent with Shipp et al. the results showed that action was primarily used to base choices on the triad task when the action choice also shared a taxonomic relation, and was presented in context. Additionally, participants were more likely to select the action related item when they had been primed with the functional action of the objects. The results are discussed in terms of the transfer effect from the object interaction task that facilitates how the objects are simulated (Barsalou, 1999, 2003; Yeh & Barsalou, 2006).

Keywords: Action; Triads; Categorisation; Priming

Introduction

Research into conceptual knowledge demonstrates that knowledge of action influences categorisation judgments (Borghi, Flumini, Natraj & Wheaton, 2012; Chao & Martin, 2000; Iachini, Borghi & Senese, 2008; Kalénine, Shapiro, Flumini, Borghi & Buxbaum, 2013; Shipp, Vallée-Tourangeau, & Anthony, 2014). For example Iachini et al. (2008) showed that when participants categorised cups varying in size, shape and grip method, they used the grip of the object as the primary strategy for categorisation. What is evident in such research is that the influence of action becomes particularly salient when participants are required to physically make the action associated with an object. For example participants in Bub, Masson and Cree (2008) performed a variant of the Stroop (1935) task where gestures were associated with colours. Objects were shown on a screen in varying colours and participants were instructed to make the gesture that they had previously associated with that colour. Participants were faster when the gesture they performed was congruent with the action that would normally be used to operate the object seen. Reaction times significantly increased when the action performed was incongruent with that normally associated with the object.

Shipp et al. (2014) found that action was not a primary source of categorising in a forced-choice triad task, but that it did have an additive effect. Participants were shown triads where a target was matched with a choice item sharing a taxonomic relation but no action, and a choice item sharing an action but no taxonomic relation. For example the target of *calculator* was presented with *set square* (taxonomically related, *stationary*) and *mobile phone* (action related, both operated by pressing buttons). In these Different Category Object (DCO) triads (see Fig. 1), participants were most likely to select the taxonomic choice when the objects were shown on a white background as *context-lean.* However in the *context-rich* condition where the objects were shown being used in a functional manner by an agent, participants were more likely to select the action-related item compared to the context-lean condition. Participants were also shown Same Category Object (SCO) triads where participants matched a target with one of two choice options which both shared a taxonomic relation to the target, but one also shared an action. For example *orange* was shown with *strawberry* (taxonomic, non-action) and *banana* (taxonomic/action, both require a peeling action). In both the context-lean and context-rich conditions participants were more likely to select the item sharing both a taxonomic and an action relation to the target. In order to rule out the possibility of perceptual features confounding the explanation (it is possible that participants chose *mobile phone* to go with *calculator* because they look similar) a series of Perceptual Object Category (PCO) triads were designed. In these PCO triads the target was presented with a choice sharing perceptual features but not taxonomic and not action relation, and a choice sharing an action but not perceptual or taxonomic relation. For example the target of *cocktail shaker* was presented with *vase* (perceptual choice) and *maracas* (action choice). The results showed that participants were significantly more likely to select the action choice when

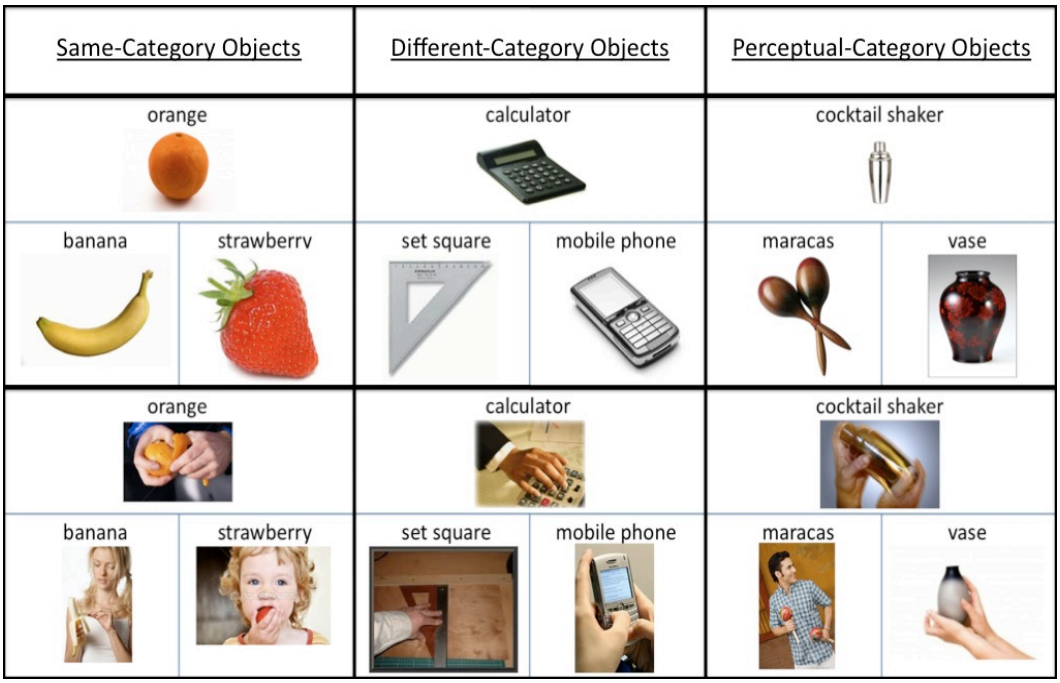


Figure 1. Examples of stimuli employed in the experiment. From left to right: Same Category Object triad, Different Category Object triad, and Perceptual Category Object triad in the context-lean condition (top panels) and in the context-rich condition (bottom panels).



Figure 2. Arrangement of the objects shown to participants in the Action Priming condition including additional resources as necessary.

shown in the context-rich condition supporting the notion that context enhances action choices. It further pre-empts the potential criticism that the action choices were being selected because of their perceptual similarities. Shipp et al. demonstrated that action is not used in categorisation when it stands alone but has an additive effect when matched with taxonomic relations, and its effect is heightened when shown in context.

The aim of the current research was to examine how participants perform on the forced-choice triad task used in Shipp et al. (2014) after they were primed with the physical actions that relate to the triad items. In addition research supports that differences exist between performing physical actions on objects compared to the intention to act (Jax & Buxbaum, 2010; Osiurak, Roche, Ramone & Chainay, 2013). Jax and Buxbaum showed

that when participants were instructed to put their hand on an object as though they would either use it or pass it to the experimenter, they were faster on the latter. Osiurak et al. not only replicated the results of Jax and Buxbaum, but also showed the reverse when participants were asked to actually pick up and use the object or pick it up and pass it to the experimenter. Borghi, Bonfiglioli, Lugli, Ricciardelli, Rubichi and Nicoletti (2007) showed primes of either a precision or a power grasp to participants followed by an object that would either be picked up with a precision or a power grasp. Previous research would suggest a strong compatibility effect (Ellis & Tucker, 2000; Tucker & Ellis, 1998, 2001) when asked to categorise the object as either natural or man-made, but this pattern was only found when participants had completed a motor training phase prior to this task.

The present experiment had two main aims. First to replicate the findings of Shipp et al. (2014) with action based choices being selected with higher frequency when shown in context, and when it also shared a taxonomic relation to the target. The second aim was to see if these action choices were more frequent when participants had previously interacted with the objects in a functional capacity. All of the items used were presented to participants prior to the triad task and participants completed either an action (using the objects), taxonomic (grouping the objects) or a movement (picking up the objects) priming phase. It was predicted that the action priming phase should lead to higher action choices on the triad task than the taxonomic or movement priming.

Method

Participants

Fifty-six undergraduate Psychology students (9 males, 47 females, M_{age} = 22.87, SD = 6.34) from the University of Hertfordshire took part in the experiment in return for course credit.

Materials

The triads used in Shipp et al. (2014) were considered for the present experiment consisting of ten of each triad type (see Fig. 1). It was important that all of the items in the triads could be interacted with prior to the triad task in the priming phase. Physical constraints of the items (e.g., *fax machine, bed, piano)* or to ethical considerations (*rifle, saw, axe*) meant that not all the original triads could be used. The final set of 16 triads consisted of five SCO triads (*pencil, glass, pin, orange,* leaflet), six DCO triads (*screwdriver, drink bottle, mug, calculator, book, paperclip, deodorant*) and five PCO triads (*USB pen, present, calculator, peppermill, handbag*). In the same manner as Shipp et al. context was manipulated: In the context-free condition the triads were presented as isolated items on a white background. In the context-rich condition participants saw the objects being used by an agent in a functional scene. All the 48 items that comprised the triads were then physically collected for use in the priming phase. The Action priming task was designed so that participants were presented with a list of tasks that involved using each item in its functional capacity such as writing their name with a pencil or reading a passage from a book.

Procedure

All participants upon entering the room saw the 48 items used in the triads presented on a table (see Fig. 2) and initially completed the priming phase. Participants were assigned to one of three priming task; action, taxonomic or movement priming. Those in the action priming completed a checklist task where the experimenter read out each task one-by-one which the participant had to complete before moving on to the next task. Each task involved using the object in its functional capacity and could either be completed on its own (e.g., *"tie the shoelace on the shoe", "open the book on any page and read out the top line"*) or were presented with additional resources (e.g., *"write your name on a piece of paper"* – where paper was provided to the participant though not included in any of the triads). Participants were instructed to move at their own pace and that there was no time limit to this task. Participants in the taxonomic priming were asked directly to sort the items into categories. They were told that they could sort them however they liked providing that each resulting group had a minimum of two members and that they would have to explain their groupings afterwards. They were also told that each sort must have a valid reason behind it and no items should be grouped together based on being remaining items that did not fit into other categories. Afterwards participants were asked to explain their decisions and describe what each category comprised. Again, no time limit was placed on the task and participants performed at their own pace. Participants in the movement priming saw two tables, one of which had all the 48 items on it, and were simply asked to pick up each item and move it to the next table in their own time. This was in order that the participants interacted with the object but not with respect to its intended function. Again no time limit was placed onto the task and participants performed at their own pace. After the priming phase concluded participants in each priming condition were allocated to either the context-lean or context-rich condition to undertake the triad task. This was completed immediately after the priming task took place with approximately two minutes between completion of the priming task and onset of the triad task. Thus the experiment employed a mixed design with three factors, Priming (action, taxonomic, movement), Triad Type (SCO, DCO, PCO), and Context (lean, rich), with triads as a repeated measures and priming and context as between subjects factors.

The triad task was presented Using Superlab on a 15" Macintosh laptop and began with a practice trial to show the participants the format of the triads on screen. Each trial began with a fixation cue presented at the top of the screen for 1000ms after which the cue disappeared and the target appeared consisting of the word and the appropriate picture depending on the experimental condition (context-lean or context-rich). After 1500ms the two choice options appeared beneath the target alongside the appropriate images. Using the same instructions as in Shipp et al. (2014) participants were instructed to "select the choice item that goes best with the target". Participants were

Table 1. *Mean percentage of action choices in the DCO, SCO and PCO triads across context and priming.*

Triad	Priming	Context	
		Lean	Rich
DCO	Action	.53	.58
	Taxonomic	.32*	.35
	Movement	.46	.40
SCO	Action	.48	.66*
	Taxonomic	.60	.54
	Movement	.48	.48
PCO	Action	.48	.78*
	Taxonomic	.50	.58
	Movement	.40	.65

Note. *Indicates those mean scores that significantly differed from a 50/50 chance ratio.

instructed to press the 'a' key to choose the item on the left-hand side of the screen and the 'l' key for the item on the right-hand side of the screen. The choice items were counterbalanced across the triads so that in half the triads the action choice appeared on the left hand side while in the remaining half the action choice appeared on the right. After they had made their choice the triad disappeared and the fixation cue appeared again for the next triad. Participants completed all 16 trials and were debriefed.

Results

The mean proportion of action responses was calculated for the SCO, DCO and PCO triads across context and priming. As was found in Shipp et al. (2014) participants showed a tendency to select the action choice more with the SCO (54%) and PCO (57%) triads than with the DCO triads (44%), and more so in the context-rich (58%) than in the context-lean (47%) condition. In addition Figure 3 shows that action choices were higher following the action priming (59%) than in the taxonomic (48%) and movement (48%) priming. A 3x2x3 mixed Analysis of Variance was used to analyze the mean action responses using Triads as a within subjects factor and Priming and Context as between subjects factors. The analysis revealed a significant main effect of Context with a higher number of action choices in the context-rich condition, $F(1, 50) = 5.25$, $p = .026$, $\eta^2 = .1$. The main effect of Triads was significant, $F(1.84, 92.67) = 6.74$, $p = .002$, $\eta^2 = .12$. Post hoc analysis using the Bonferroni adjustment found that the action responses on the DCO triads were significantly lower than both the SCO triads ($p = .046$) and the PCO triads ($p = .004$). No difference was found between the SCO triads and the PCO triads ($p = 1.$). The main effect of Priming was also found to be significant, $F(2, 50) = 3.84$, $p = .028$, $\eta^2 = .13$. Post hoc analysis using the Bonferroni adjustment revealed that the difference between action and taxonomic priming ($p = .063$) and the action and movement priming ($p = .066$) were marginally significant (see Fig. 3). The difference between the taxonomic and moving priming was not significantly different ($p = 1.$).

The two-way interaction effect between Context and Triads was significant, $F(1.84, 92.67) = 4.58$, $p = .012$, $\eta^2 = .08$. Post hoc analysis using the Bonferroni adjustment (see Fig. 4) found no effect of context on the DCO ($p = .92$) and SCO triads ($p = .45$), but action choices were significantly higher for the PCO triads in the context-rich than context-lean condition ($p < .001$). The two-way interaction between Priming and Triads was not found to be significant, $F(3.71, 92.67) = 2.02$, $p = .10$, $\eta^2 = .08$, nor was the two-way interaction between Priming and Context, $F(2, 50) = 1.72$, $p = .19$, $\eta^2 = .06$, or the three-way interaction between Context, Priming and Triads, $F < 1$.

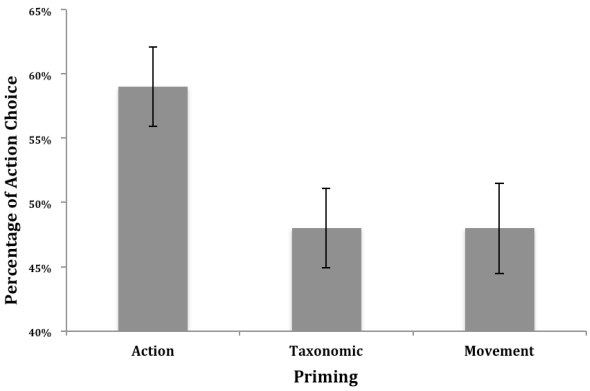


Figure 3. Mean percentage of action choices across the Action, Taxonomic and Movement Priming conditions. Error bars are standard errors of the mean.

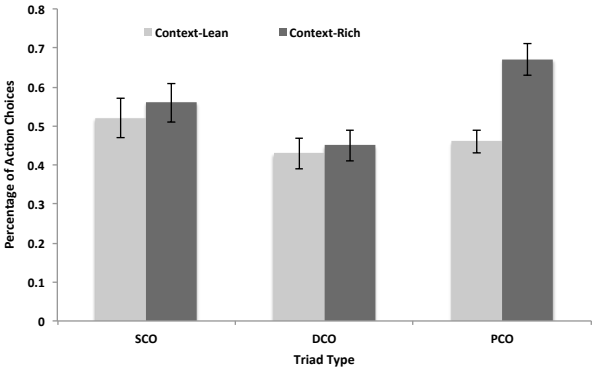


Figure 4. Mean percentage of action choices with Same Category Object (SCO), Different Category Object (DCO), and Perceptual Category Object (PCO) triads in the context-lean condition (light grey bars) and in the context-rich condition (dark grey bars). Error bars are standard errors of the mean.

Due to the dichotomous nature of the response variable it is possible that the mean action scores were no different

from selecting the action choice by chance. Therefore further post hoc analyses were conducted on the mean percentage of action choices across the conditions in order to assess if the means were significantly different from a 50/50 ratio. Single sample t-tests were conducted on the cell means shown in Table 1 using a theoretical mean of 0.5. Only the SCO and PCO triads following the action priming were significantly higher than chance ($p < .05$). The DCO mean following the action priming was higher than chance but not significantly so ($p = .21$). This partially supports the previous ANOVA analysis where higher action choices followed only the action priming. However this effect was not consistent for all the triad types. In addition the only action scores that were significantly lower than chance was on the DCO triads following the categorisation priming ($p = .03$).

Discussion

The results of the current experiment revealed three main findings. The first is the replication of the triad effect from Shipp et al. (2014). The action choice was least likely to be selected on the DCO triads when choosing between a taxonomic and an action choice. This shows that, as with the previous experiments, action information alone is less likely to be favoured as the basis for category membership when the alternative is taxonomic information. The action item was most likely to be selected with the SCO and the PCO triads. The results on the SCO triads support the notion of action having an 'additive' effect making such items sharing both action and a taxonomic relation as 'better' category members. The PCO triads further show that when no taxonomic relation is present as an alternative participants are more likely to use action than perceptual properties to group items together.

The second main finding is the partial replication of the context effect from Shipp et al. (2014). While previously it was found that across all three triads the action choice was most likely selected in the context-rich condition, this effect was limited with the PCO triads in this experiment. The most likely reason for this is the exclusion of certain triads from the original set used in Shipp et al. It is possible that these triads removed were more likely to lead to the action choice than other triads. An item analysis of the triads used in Shipp et al. (2014) supports this showing that certain items are more susceptible to context effects than others (Shipp, Vallée-Tourangeau, & Anthony, in prep). For example the *rifle/sword/water pistol* triad (DCO) was removed from the present experiment for practical reasons but has been shown to be strongly influenced by context.

The third main finding from this experiment is the effect of priming conditions. The results showed that participants were more likely to choose the action related item in the triads following the action priming where all the items were used in their functional capacity. The selection of action choices in the SCO and PCO triads were significantly higher than chance in the context-rich condition following the action priming. Selection of the action choice in the DCO triads was not significantly higher than chance. Such results are in line with notions of situated simulation (Barsalou, 1999; 2003; 2008; Yeh & Barsalou, 2006). Barsalou (2003) states that category features recently encountered become more readily available for subsequent processing. If someone recently peeled an orange then *peeling* should be temporarily more salient than other, more context-dependent, properties. Therefore prior functional engagement with the objects in the initial task could temporarily increase the saliency of the shared actions and have a direct influence on the simulations made, and hence on the choices in the triad task. Variations in how the participants interact with the objects would be expected to influence the simulations generated. Therefore participants who interacted with the items functionally should be more likely to simulate those recent actions and the salience of these could be further enhanced by the context shown. As such a transfer effect occurs in which the interaction with the objects directly facilitates the simulations generated. As the taxonomic and movement priming does not focus on the shared actions participants would be less likely to simulate the relevant actions associated with use and so less likely to select the action choice in the triads (as was found in the data).

An alternative explanation might be linked to functional actions, in comparison to structural actions, resulting in longer lasting neural representations (Bub & Masson, 2012; Jax & Buxbaum, 2010). The priming tasks used here can be separated into two distinctive sets of actions, functional actions related to functional use and volumetric/structural actions related to general movement and interacting with the objects in a non-functional manner (Bub & Masson, 2012; Bub et al., 2008; Jax & Buxbaum, 2010; Osiurak et al., 2013). Buxbaum and Kalénine's (2010) Two Action Systems (2AS) model proposes that the brain has separate action systems for the processing of functional and structural actions. The dorso-dorsal stream is specialized for acquiring information based on the structure of objects and their affordances whereas the ventro-dorsal system is specialized for the retrieval of conceptual representations. Functional activations tend to be long lasting and can cause interference effects on later actions (Bub & Masson, 2012; Jax & Buxbaum, 2010). Structural activations occur more quickly than functional, but decay rapidly and do not cause later interference effects. Jax and Buxbaum showed that when participants performed a functional action on an object and then later performed a grasp action, response latencies were significantly longer than when they performed the grasp action first. This interference effect as a result of functional activation lasted for approximately 20 minutes during the entire task. While this functional activation had an interference effect on the use/grasp task, it is possible that this has a facilitation effect on the triad task. The partial re-activation of the neurons during simulation of the objects in the triad should be facilitated by the current activation of the functional system. As such the simulation itself should make the action element more salient between the triad objects and participants will therefore be more likely to select the action choice. This is further amplified by the concurrent activation of conceptual knowledge with the functional system of the ventro-dorsal stream as proposed by Buxbaum and Kalénine (2010). The same facilitation

effect does not occur following the movement prime as the structural activations of the dorso-dorsal stream dissipate quickly. Hence only the long lasting functional activations as a result of the action priming should lead to higher action choices in the triad task. However it should be noted that such an explanation is purely speculative at this point as the data cannot be used to explicitly support such claims. However the ventro-dorsal facilitation effect could be tested and supported through using the full set of triads developed from Shipp et al. (2014). As explained above the full range of triads was not used in this experiment because of ethics and feasibility of priming how participants use certain objects such as *sword*. If it is the case that functional activations of the ventro-dorsal system facilitate simulations of the objects, then in the triad task used here this should result in higher action choices on the primed objects and lower action choices for those objects not primed. Future research should aim to compare performance on such triads in this task.

In conclusion the data reported here has shown that the triad effect found in Shipp et al. (2014) has been replicated. Participants were more likely to use action as a source for categorisation on the forced-choice triad task when it was shown in combination with taxonomic information, and when it was shown in context. The results further show that priming participants with the functional rather than structural actions of the objects led to increased action choices on the triad task. Such results are also in line with the view that a transfer effect occurs from the object-interaction task which facilitates how objects are mentally simulated (Barsalou, 1999, 2003; Yeh & Barsalou, 2006).

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18*, 513-562.

Borghi, A. M., Bonfiglioli, C., Lugli, L., Ricciardelli, P., Rubichi, S., & Nicoletti, R. (2007). Are visual stimuli sufficient to evoke motor information? Studies with hand primes. *Neuroscience Letters, 411*, 17-21.

Borghi, A. M., Flumini, A., Natraj, N., & Wheaton. L. A. (2012). One hand, two objects: Emergence of affordance in contexts. *Brain and Cognition, 50,* 64-73.

Bub, D. N., & Masson, M. E. J. (2012). On the dynamics of action representations evoked by names of manipulable objects. *Journal of Experimental Psychology: General, 141,* 502-517.

Bub, D. N., Masson, M. E. J., Cree, G. S. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition, 106,* 27-58.

Buxbaum, L. J., & Kalénine, S. (2010). Action knowledge, visuomotor activation, and embodiment in the two action systems. *Annals of the New York Academy of Sciences, 1191,* 201-218.

Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *NeuroImage, 12,* 478-484.

Ellis, R., & Tucker, M. (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology, 91,* 451-471.

Iachini, T., Borghi, a. M., & Senese, V. P. (2008). Categorization and sensorimotor interaction with objects. *Brain and Cognition, 67,* 31-43.

Jax, S. A., & Buxbaum, L. J. (2010). Response interference between functional and structural actions linked to the same familiar object. *Cognition, 115,* 350-355.

Kalénine, S., Shapiro, A. D., Flumini, A., Borghi, A. M., & Buxbaum, L. J. (2014). Visual context modulates potentiation of grasp types during semantic object categorisation. *Psychonomic Bulletin & Review, 21,* 645-651.

Osiurak, F., Roche, K., Ramone, J., & Chainay, H. (2013). Handing a tool to someone can take more time than using it. *Cognition, 128,* 76-81.

Shipp, N. J., Vallée-Tourangeau,, F., & Anthony, S. H. (2014). The context-dependent nature of action knowledge. In P. Bello, M. Guarani, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2925-2930). Austin, TX: Cognitive Science Society.

Shipp, N. J., Vallée-Tourangeau,, F., & Anthony, S. H. (in prep). Strategies for making forced-choice categories: An item analysis of Shipp, Vallee-Tourangeau, and Anthony (2014).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18 ,* 643-662.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance, 24,* 830-846.

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition, 8 ,* 769-800.

Yeh, W., & Barsalou, L. W. (2006). The situated nature of concepts. *American Journal of Psychology, 119*, 349-384.

Dynamics of Strategy Adaptation in a Temporally Extended Monty Hall Dilemma

Stephanie Petrusz (stephanie.petrusz@gmail.com) & Joshua Shields (joshua.n.shields@gmail.com)
Department of Psychology, Franklin & Marshall College
Lancaster, PA, 17603 USA

Theo Rhodes (theo.rhodes@oswego.edu) & Abisha Munroe (amunroe@oswego.edu)
Department of Psychology, SUNY Oswego
Oswego, NY, 13126 USA

Abstract

We present the results of two temporally extended experimental implementations of the Monty Hall dilemma in order to examine the dynamics of belief. In the first experiment, we used the standard three-door version of the dilemma, but biased the probability of the winning door positionally. Participants capitalized on the increased probabilities but did not discover the optimal switch strategy. In the second experiment, we increased the number of doors, in each case removing all but two doors. As the number of doors increased, participants converged on the optimal switch strategy, as well as increasing their confidence in their strategy. This suggests that the information relevant to the MHD is not win frequencies but how the different elements of the dilemma are related.

Keywords: Belief; Monty Hall Dilemma; Dynamical Systems

Introduction

The conditions under which an agent may be said to believe can be cashed out in different ways. Belief is essentially an agent's holding that a proposition or state of affairs is the case, and as such is difficult to externally assess. Two related but distinct ways suggest themselves: an agent may assent to the truth or falsity of a proposition, or an agent may be disposed to act as though a proposition is true—or, if one wishes to avoid speaking as though propositions are the contents of beliefs, to act as though a state of affairs obtains. If we focus on the second type of belief, that is, beliefs as underwriters of actions, two more things make themselves clear. First, the ability of beliefs to guide actions must be relatively robust to uncertainty. Second, beliefs must change in response to new information or evidence, so that actions may be adapted to a dynamic environment.

Formal models of belief therefore often treat beliefs as probabilities, and belief change or belief update as changes in those probabilities and in the connections between the beliefs they represent. Many of these models share certain features: the beliefs in question are framed as propositions and the dynamics are understood as state-to-state changes. Many models further share a Bayesian basis. However, a thread of argument in the epistemological literature claims that a basic species of Bayesian revision is probably inadequate to model belief change (see van Fraassen (1989), cited in Arló-Costa, unpublished). Evaluating the conditionals required by Bayesian update is problematic for reasons including the complexity of agents' existing commitments (see, for instance, the "web of belief" of Quine and Ullian, 1978) and the difficulty of evaluating the probability of some fact against an unenumerated backdrop of possibilities.

In our view the mathematical considerations particular to formal epistemological accounts of belief update are also appropriate to psychological accounts. For our empirical investigation we required sufficient data to construct a model and a task where we could plausibly discuss what the experimental participants believed about their actions. Because of its simplicity of execution and its popularity, we chose the Monty Hall Dilemma (e.g., vos Savant, 1990). This paradigm satisfied several aims. First, we wanted a task in which outcome probabilities were clear, though not necessarily obvious to the participants. Second, we wanted a task that involved beliefs as the basis for actions, not simply evaluating propositions with degrees of assent. Third, we wanted to offer participants the opportunity to vacillate while still making a desired response—in effect to say, "I'm responding this way but I don't know if I should", as a measure of degree of belief. We hoped that by evaluating participants' approaches to the MHD we would be able to address how people evaluate probabilities and choose between competing strategies for dealing with situations in which probabilities change. We hoped further to investigate whether a separation existed between participants' beliefs as the basis for actions and their propositional beliefs about the task.

The MHD is a good example of the difference between incorporating probabilistic evidence and reasoning probabilistically because it is an example of persistent failure to act in accordance with the rules of probability. In the classic version, a "contestant" is presented with three doors. Behind one of the doors is a car, and behind the other two are goats. The contestant chooses a door, and then one of the two remaining unchosen doors is opened—always revealing a goat. The contestant is then given the chance to stick with their initially chosen door and receive whatever lies behind it, or to switch to the remaining unchosen and unopened door, receiving instead whatever prize lies behind that door. Probabilistically, a contestant always maximizes their chances of winning the car by switching their pick, yet in practice people seldom choose to switch. In fact, many people resist the logic of switching even after it is

demonstrated or explained to them that switching is always the better bet.

The classic formulation can be extended in a variety of ways. Granberg (1999) tested a 4-door version of MHD with unequal probabilities, where the optimal strategy is to choose the lowest probability door initially and then switch to the highest probability door remaining after the reveal stage. Participants in Granberg's study were explicitly informed of the door probabilities on each trial, yet even so no participants reliably used the optimal strategy. Baratgin and Politzer (2010) showed that both the dilemma and the resistance to the winning strategy are rooted in the manner in which information is presented throughout the scenario. They presented several versions of MHD and concluded that when information is presented as being relevant only to the unchosen doors, people do not commit the typical MHD error. They argue that the opening of the "goat" door is taken to be relevant to the overall probabilities, and that players do not effectively isolate the new information to the real space of outcomes, which includes only the unchosen doors.

The MHD has been extensively studied in cognitive science but a full review is beyond the scope of this paper. Krauss and Wang (2003) discuss the psychological mechanisms that might underpin successful MHD strategies and suggest that correctly reasoning through the problem depends on the way in which information is represented throughout the scenario. Their treatment of the "natural frequencies" approach suggests that the solution is more apparent when the frequencies of win and loss events in the aggregate are considered, rather than the probability of winning or losing a single run. Saenen, Van Dooren, and Onghena (2015) used feedback to improve participants' performance on the MHD but noted that the increases in performance did not come with associated improvement in understanding of the dilemma. These results are of a piece with our considerations: the information needed to identify a correct strategy becomes available to be acted upon with practice, but does not become available to conscious knowledge in the same fashion.

We devised several variants of the MHD to explore whether manipulation of some of the dilemma's parameters would affect participants' exploration and adoption of different strategies. If this is the case then we may say participants are able to incorporate information that is probabilistic (that is, not fully reliable) as a basis for action. We also explored whether manipulations yielding more successful performance than in the classic case would translate to higher participant confidence (that is, explicitly held beliefs about the value of the new information or about the utility of an adopted strategy) in their attempts.

Experiment 1

In experiment 1, we use the standard 3-door MHD, biasing the probability of winner based on door position in order to test whether this bias would in fact be apparent to participants, and whether it would enhance or diminish their discovery of the switching strategy.

Methods

Participants 20 undergraduate students (mean age = 19.9, 13 male, 7 female) from Franklin & Marshall College participated in the experiment for course credit.

Design Experiment 1 used a 3-door MHD with color (shades of red) serving as the property distinguishing the "prize" (brighter shade) from the "goats" (darker shade). We chose color as the target stimulus property because it is both salient to participants and easy to program.

The experiment was divided into 8 blocks: an initial 100-trial block in which each of the three doors had an equal chance of concealing the "prize", followed by 4 100-trial blocks with unequal probabilities of the "prize" being behind each of the doors, and finally 3 50-trial blocks with equal probabilities. In the unequal-probabilities blocks either the leftmost or rightmost door had a higher chance of concealing the "prize". Participants were randomly assigned to either the left-biased (Group L) or right-biased (Group R) condition. Participants were asked both to select a door by clicking on it, and also to register their confidence in their choice, measured in how far from the center of the chosen square *in any direction* they clicked, while still making the response within the confines of the door. This lack of directionality was included specifically to avoid laterality effects, and also to simplify the verbal instructions to the participants. The center represented the highest possible confidence and the inside edge of the door represented the lowest confidence. In all there were 650 trials and 1300 distinct responses for each participant.

Procedure Each individual trial consisted of an instance of MHD, as follows: three "doors" (square boxes) were displayed on screen, along with the text prompt "Select a door". A participant would click on one box, whereupon the color behind a remaining non-prize door was revealed. The participant would then be asked to stick or switch, via the text prompt: "Switch your pick?" They would then either switch (by clicking on the unopened door they did not originally choose) or stick (by once again clicking on the originally-chosen door). The color behind the door they ultimately chose would then be revealed, along with text telling them either "You win" or "You lose" depending on the outcome of their final choice. At the end of the experiment participants were asked to briefly describe any strategy they might have employed in the course of the experiment.

Results

Because of the large number of data points, it is especially important to consider results in terms of both significance and effect size. We report only those results that meet both criteria.

A 2 (Group) × 8 (Block) ANOVA was conducted on the measures of door chosen, whether the correct door was selected, confidence, and whether a trial was a "switch" or "stick". Group significantly affected all four measures, but with a meaningful effect size only for door chosen, $F(1, 25984) = 1828.63$, $p < .001$, partial eta squared = .07. The "door chosen" measure was significantly higher for Group R ($M = 2.19$, $SD = .83$) than for Group L ($M = 1.67$, $SD = .77$), with doors coded as 1 (left), 2 (middle), and 3 (right). Block significantly affected all four measures, but with meaningful effect sizes for confidence, $F(7, 25984) = 83.65$, $p < .001$, partial eta squared = .02 and correct door, $F(7, 25984) = 73.95$, $p < .001$, partial eta squared = .02. Because lower values of the confidence measure indicate higher confidence (as measured in pixel distance from the center of the chosen door), we report the negative of the values of confidence. Confidence is lowest in Block 1 ($M = -48.57$, $SD = -31.97$) before increasing steadily until its maximum in Block 5 ($M = -35.5$, $SD = -24.15$) and then decreasing again until Block 8 ($M = -47.37$, $SD = -28.17$). The proportion of trials on which participants end up choosing the winning door is lowest in Block 1 ($M = .57$, $SD = .5$) and steadily increases until Block 5 ($M = .85$, $SD = .36$) before decreasing sharply in Block 6 ($M = .6$, $SD = .49$) and remaining relatively steady until the final block ($M = .63$, $SD = .48$). There was a significant Group × Block interaction for door chosen, $F(7, 25984) = 296.41$, $p < .001$, partial eta squared = .07.

Having characterized each trial as either a "switch" or "stick" trial, we conducted a second ANOVA using this strategy variable as an IV. Error degrees of freedom are smaller because the analysis is in terms of trials rather than responses. For the 2 (Group) × 2 (Strategy) × 8 (Block) ANOVA, Group significantly affected door chosen, $F(1, 12968) = 2710.61$, $p < .001$, partial eta squared = .17.
Group R chose the right-most door ($M = 2.32$, $SD = .85$) significantly more than Group L ($M = 1.45$, $SD = .75$). Strategy significantly affected correct door, $F(1, 12968) = 979.99$, $p < .001$, partial eta squared = .07. Trials that ended with a switch response had a higher proportion of correctly chosen doors ($M = .82$, $SD = .39$) than trials that ended with a "stick" ($M = .58$, $SD = .49$). Strategy also significantly affected confidence, but with an effect size so small as to be considered negligible. Block significantly affected door chosen, $F(7, 12968) = 20.33$, $p < .001$, partial eta squared = .01, confidence, $F(7, 12969) = 42.08$, $p < .001$, partial eta squared = .02, and correct door, $F(7, 12968) = 144.6$, $p < .001$, partial eta squared = .07. There were significant two-way interactions of Group × Block, Group × Strategy, and Strategy × Block. The three-way interaction of Group, Block, and Strategy was also significant.

Finally, we found that correct responses were only weakly correlated with confidence, $r = .31$, $p < .001$.

Discussion

The winning MHD strategy is switching, but as described above, a preponderance of players not only fail to discover or employ this strategy, but resist acknowledging that it is better even after having it explained. This fact is often leveraged as part of an argument about irrationality or about the difficulty of reasoning probabilistically. As Granberg (1999) reported, in an unequal-probabilities MHD the best strategy is even less obvious: to optimize the chances of winning a player should initially select a low-probability door and then switch to the highest remaining probability door after a "goat" has been revealed.

In Granberg's 4-door unequal-probabilities study participants were informed of the probabilities of finding the prize behind each door. In the present experiment participants were not informed of the probabilities at any point, yet were able over the course of the unequal-probabilities blocks to hone in on the location of the winning door (leftmost for Group L, rightmost for Group R). We take this to be evidence of participants' ability to *perceptually* coordinate with the structure of the experimental task, and believe this ability is conditioned on the availability of very many trials as a basis for learning, and on the fact that the unequal win probabilities made the presence of a winner more salient than in the standard case. We also believe that the availability of a tactic more obvious than switching (which, as shown by the second ANOVA, remains the better strategy) accounts in part for the strength of the correlation between performance and confidence, a subject to which we return in the general discussion.

In Experiment 1 we attempted to approach the Monty Hall paradigm in a way more streamlined than the classic case, but in a way that also allowed us to look at the effects of making the probabilities involved more obvious, and of how repeated exposures to the problem affected participants' approaches. However, further refinements were required.

Experiment 2

Experiment 2 was designed to not only manipulate the number of doors and the associated win probabilities, but also to further refine and streamline the experimental design.

Methods

Participants Twenty-eight undergraduate students (6 male, 22 female) from SUNY Oswego participated in the experiment for course credit.

Design In Experiment 2 we returned to an equal-probability design for the doors, but manipulated the *number* of doors in the Monty Hall scenario. Each block consisted of 30 MHD trials with between 3 and 11 doors, presented in ascending (from 3 to 11 doors), descending (from 11 to 3 doors) or random increments across 8 blocks. We also changed how we denoted the "prize" and "goat" doors, using a check symbol for a winning door and an x symbol for a losing door.

Procedure The procedure was the same as in Experiment 1, with the important exception that, after a participant made their initial choice, *all but one* of the remaining doors was opened, instead of a single door. This design was based on an experiment suggested in Spivey (2008), and was intended to approach the issue of making the winning *strategy* as salient as the winning door had been in Experiment 1.

Results

We conducted a 3 (Group) × 8 (Block) × 8 (Number of Doors) ANOVA on the dependent measures of correct responses, confidence (calculated for the second response in a trial), and the total number of "switch" trials per participant and block. As in Experiment 1, we report only those results that achieved both significance and meaningful effect size. Group significantly affected number of switches, $F(2, 7485) = 203.66$, $p < .001$, partial eta squared = .05. On average, the Descending group had the highest number of switches per block ($M = 27.21$, $SD = 2.74$), followed by the Ascending group ($M = 24.23$, $SD = 7$), with the random group having the lowest average number of switches ($M = 22.59$, $SD = 6.8$). Block significantly affected total number of switches (see Figure 1), $F(8, 7485) = 24.15$, $p < .001$, partial eta squared = .025. Switching was on average highest in Block 6 ($M = 26{,}57$, $SD = 3.77$) and lowest in Block 1 ($M = 19.64$, $SD = 6.99$).

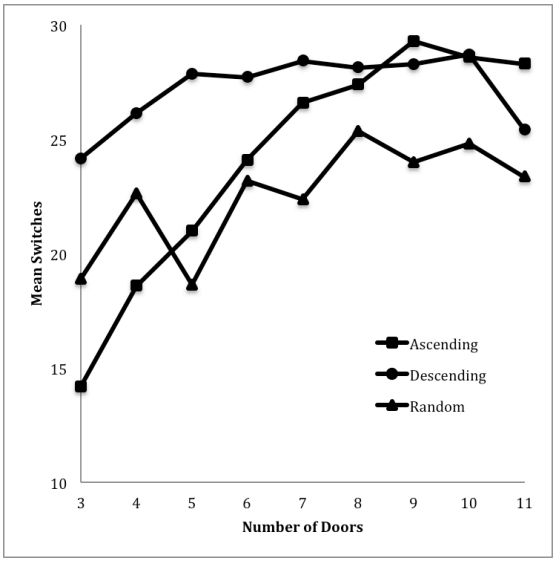


Figure 1: Mean switches made by participants by number of doors and condition.

Number of doors significantly affected correct responses (see Figure 2), $F(8, 7485) = 9.96$, $p < .001$, partial eta squared = .011 and total number of switches, $F(8, 7485) = 68.06$, $p < .001$, partial eta squared = .068. On average, the 10-door block had the highest proportion of correct responses ($M = 83.8\%$, $SD = 36.9\%$), while the 3-door block had the lowest ($M = 56.7\%$, $SD = 49.5\%$). On average, the 10-door block had the highest number of switches ($M = 27.14$, $SD = 3.96$) while the 3-door block had the lowest ($M = 18.54$, $SD = 7.37$).

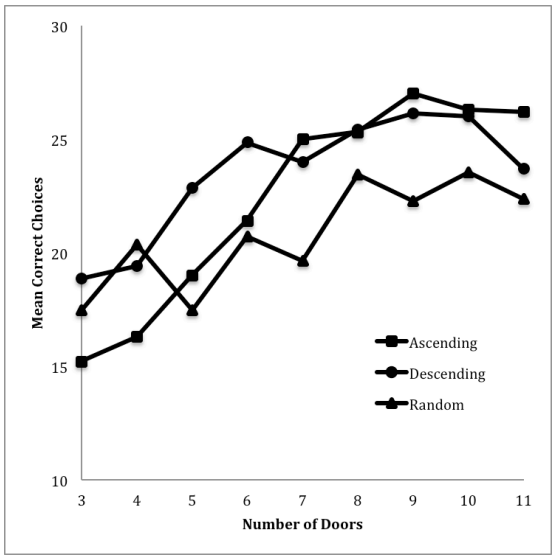


Figure 2: Mean correct choices made by participants by number of doors and condition.

Having calculated the number of switches, we included this variable in a second analysis. We regressed Condition, Block, Number of Doors, and Total Switches on correct responses. These predictors explained 12% of the variance ($R^2 = .12$, $F(4, 7559) = 252.76$, $p < .001$. The number of doors significantly predicted correct responses, $\beta = .09$, $p < .001$, as did total number of switches, $\beta = .3$, $p < .001$.

We then regressed Condition, Block, Number of Doors, and Total Switches on confidence (registered on the second response per trial). Confidence was normalized so that all values fell between 0 and 1. We did this to account for the fact that participants varied in their range of confidence values, and we were primarily interested in how confidence changed over the course of the experiment (see Figure 3).

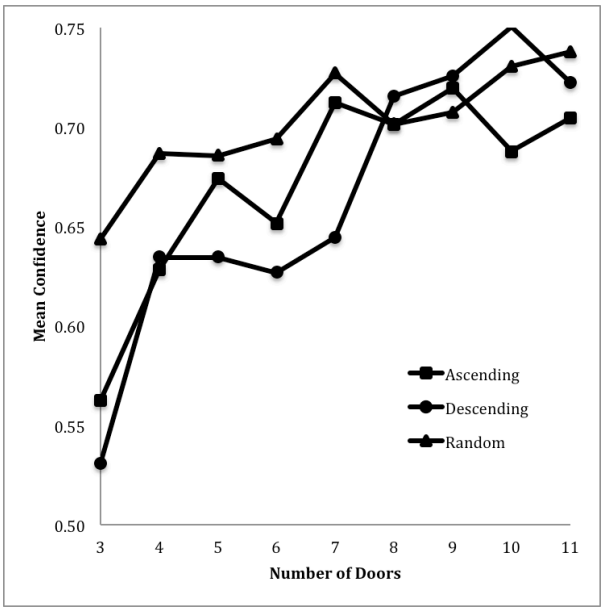


Figure 3: Mean normalized confidence estimates for by number of doors and condition.

These predictors explained only 3% of the variance in the normalized confidence (R^2 = .03, $F(4, 7559) = 65.45$, $p < .001$. All four predictors significantly affected normalized confidence: Condition ($\beta = -.06$, $p < .001$), Block ($\beta = .05$, $p < .001$), number of doors ($\beta = -.1$, $p < .001$), and total number of switches ($\beta = -.11$, $p < .001$).

Finally, we found that correct responses were only weakly correlated with confidence, $r = .24$, $p < .005$.

Discussion

In Experiment 2 we varied the MHD by manipulating the number of doors in each block. The presence of additional doors alters the win probabilities, though not the fact that switching still has a higher probability of winning than sticking. In theory, the additional doors create a more difficult problem of probabilistic reasoning than the 3-door case, but the removal of all but one door changes those calculations. As Spivey (2008) points out, seeing the subset of possibilities reduced so drastically may actually make switching a more obvious strategy. Our results bear this hypothesis out.

The significant effect of Group on number of switch trials means that the order of presentation of door numbers makes a difference to how participants approach the task—in other words, the task exhibits hysteresis. Here, the Descending group had the highest number of switch trials: the presence and subsequent removal of more doors makes it easier to see that the original door has a lower probability of being a winner than the one remaining out of all the rest, which persists as the number of doors is reduced over the following blocks. This interpretation is also supported by the fact that the 10-door block had the highest numbers of both switches and correct responses, while the 3-door block had the lowest of both. Although a case could be made that the 11-door block should have been even higher, the 11-door block was also the first block presented for a third of the participants, and therefore performance likely suffered because of the novelty of the task.

We attempted to better isolate the effect that switching has on performance and on confidence using linear regression, with mixed results. In particular, all of our independent variables together accounted for only 3% of the variance in participants' confidence. Taken together with the weak correlation between performance and confidence, it seems clear that we have not yet isolated a manipulation that primarily affects confidence, though this might also be a limitation of how confidence is measured.

General Discussion

The present set of experiments was designed to investigate belief from an empirical standpoint. To this end, a distinction was drawn between two different ways of thinking about belief. There is belief-as-proposition, which is perhaps the better-known way of considering belief, and there is belief as a guide to action. Showing that one *possesses* both species of belief but that they are not equivalent, then, takes a step toward an understanding of belief that is not beholden to its propositional character. Allowing participants to provide an overt measure of behavior (door choices) as well as a more covert measure of their own state (confidence) lets us see belief's different guises separately.

Our other aim was to consider why it is that the MHD presents such difficulty. We considered that participants might assume the initial symmetry between doors held throughout the scenario and attempted to introduce new information via the introduction of the positional bias in Experiment 1. We also considered that participants fail to see that the doors remaining after a reveal constitute a distinct subset of probabilities and attempted to make this fact more salient by manipulating the number of doors throughout Experiment 2.

The results of both Experiments 1 and 2 speak to the distinction between a belief on which action (e.g. choosing a door that might conceal a prize) is based and assent to the propositional description of that belief (e.g. "I believe that Door Number 3 is the winner"). Participants are able to accomplish the former regardless of their degree of belief in the latter. Performance and confidence hold together only weakly even when performance is successful. As acknowledged above, we do not feel that we have as yet hit upon a way of really manipulating confidence, but we do feel that we have demonstrated that successful action need not be conditional on any particular degree of belief. This finding is of a piece with the hypothesis that the aspects of the task about which a participant might have propositionally-construed beliefs are not necessarily the aspects that are used to guide behavior (Anderson & Runeson, 2008; Runeson, Juslin, & Olsson, 2000).

In Experiment 1 we hoped to shed light on the dynamics of beliefs in the face of changing evidence, by altering the

probabilities that a given door would be the winner. The results show that participants are able to perceive that one door is likelier to pay off and are able to adapt their strategy accordingly. This result, coupled with the understanding that the beliefs that guide actions are usually many more in number than those considered here, led us to believe that adaptation to changing information might be considered a dynamical system. Experiment 2 was designed to investigate this question more fully, by incorporating the changing number of doors. This provided not only a candidate control parameter for the MHD but also allowed us to test the hypothesis that failure to separate the probabilities associated with chosen doors versus unchosen doors (or, to put it another way, improperly conditioning beliefs about chosen doors on information about unchosen doors) is responsible for the difficulty of discovering and adopting the switching strategy.

The results of Experiment 2 support both contentions. The different conditions show signs of hysteresis, a hallmark of dynamical systems, and suggest further avenues of exploration in future manipulations within this paradigm. The results also support the hypothesis that information in one part of the dilemma is taken to affect other parts is a matter of framing. Additional manipulations regarding numbers of choices, distributions of probabilities, and physical versus virtual objects, are all natural extensions of the work presented here.

Our paradigm needs refinement, but works at a first past. Participants change their actions based on changing information, and these changes persist in some fashion. The actions are undertaken not wholly in keeping with degree of assent to the correctness of the action. Finally, there are large-scale dynamics with patterns of consistency that provide theoretical and methodological insights into investigating changes in belief.

References

Andersson, I. E., & Runeson, S. (2008). Realism of confidence, modes of apprehension, and variable-use in visual discrimination of relative mass. *Ecological Psychology,* 20(1), 1-31.

Arló-Costa, H. (unpublished). *Computational and probabilistic models of abduction*. Department of Philosophy, Carnegie-Mellon University.

Baratgin, J., & Politzer, G. (2010). Updating: A psychologically basic situation of probability revision. *Thinking & Reasoning*, 16(4), 253-287.

Granberg, D. (1999). A new version of the Monty Hall Dilemma with unequal probabilities. *Behavioural Processes*, 48(1), 25-34.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132*(1), 3-22.

Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief (Vol. 2)*. R. M. Ohmann (Ed.). New York: Random House.

Runeson, S., Juslin, P. and Olsson, H. (2000). Visual perception of dynamic properties: cue heuristics versus direct-perceptual competence. *Psychological Review*, 107: 525–555.

Saenen, L., Van Dooren, W., & Onghena, P. (2015). A randomized Monty Hall experiment: The positive effect of conditional frequency feedback. *Thinking & Reasoning*, *21*(2), 176-192.

Spivey, M. (2008). *The continuity of mind.* Oxford University Press.

Van Fraassen, B. (1989). *Laws and symmetry*. Oxford University Press.

Vos Savant, M. (1990). Ask Marilyn. *Parade*, 15.

Causal Contrasts Promote Algebra Problem Solving

Jian-Ping Ye (jpy2110@tc.columbia.edu)
Department of Human Development, 525 W 120th Street
New York, NY 10027-6625 USA

Jessica M. Walker (jewalker@chapman.edu)
Department of Psychology, 1 University Drive
Orange, CA 92866-1005 USA

Patricia W. Cheng (cheng@lifesci.ucla.edu)
Department of Psychology, 405 Hilgard Avenue
Los Angeles, CA 90095-1563 USA

Abstract

The causal-contrast approach is a new teaching method that recruits learners' implicit causal discovery process to improve math learning by juxtaposing contrasting information critical to discovering the goal of each solution step. Students often blindly memorize mathematical procedures and have difficulty transferring their knowledge to novel problems. By enabling learners to infer the goal of each step, the causal-contrast approach substantially improved high-school algebra problem solving compared to a traditional instructional control (Walker, Cheng & Stigler, 2014). The present study developed Walker et al.'s instructional materials into a computer-based teaching program and tested the new approach on community-college students, a population for whom the traditional approach is often ineffective. The study added two new conditions: a baseline that received no instruction and a condition using a teaching video from Khan Academy, a well-regarded online educational website representative of the traditional approach. A delayed post-test indicated that the causal-contrast condition produced dramatically greater success in solving transfer problems than the other three conditions.

Keywords: causal contrasts; causal induction; implicit learning; knowledge transfer; mathematics education

Introduction

Compared to previous years, U.S. students' ranking in science, technology, engineering, and math (STEM) subjects has been improving gradually; however, students' performance on the international mathematics assessments continue to fall below the international average (PISA, 2003, 2006, 2009, 2012). A massive amount of research has studied this issue, with the goal of improving students' mathematical learning and understanding.

One of the most common yet ineffective strategies that students use to learn mathematics is to blindly memorize the solution steps without truly understanding the reason behind each step (e.g., Stigler, Givvin, & Thompson, 2010). Without connecting the procedures to goals or concepts, repeatedly solving a large number of conventional problems may not be helpful for fostering the flexible use of mathematical knowledge (Cooper & Sweller, 1987; Schwartz et al., 2011). Previous research have suggested various ways to enhance mathematics education, such as emphasizing worked examples to support schema acquisition (Carroll, 1994; Sweller & Cooper, 1985), self-generating explanations during learning to improve understanding and knowledge integration (e.g., Chi et al., 1989; Chi, 2000), labeling subgoals to illustrate the reason why certain solution steps should be applied to promote learning and transfer (e.g., Catrambone, 1998) and using comparisons and contrasting examples in the instruction to improve the learning of concepts and procedures (e.g., Hattikudur & Alibali, 2010; Rittle-Johnson & Star, 2007; Richland & McDonough, 2010).

Fewer studies have examined the use of causal learning to support mathematical learning and problem solving. Whereas mathematics is traditionally taught using explicit instruction to convey analytic knowledge, the causal contrast approach is an instructional method that recruits an implicit empirical-learning process to help students discover the reasons underlying mathematical procedures (Walker et al., 2014). Humans have a natural capacity for learning cause-and-effect relations (Cheng, 1997; Gopnik et al., 2004; Leslie & Keeble, 1987). The causal-contrast approach hypothesizes that operation-outcome relations in mathematics can be thought of as causal. The approach 1) induces students to formulate the goal of an operation in a mathematical procedure by allowing them to fail to solve a challenging problem using their prior mathematical knowledge and 2) promptly provides the relevant *conditional contingency* information (Cheng & Holyoak, 1995) by adding a *contrasting problem*, one that controls for confounding factors by being as similar as possible to the challenging problem but with the features causing the difficulty removed. "No confounding" is a pre-requisite for causal learning. The goal is for students to readily discover the goal of each step in the solution. It is not the use of comparison *per se* that characterizes the causal-contrast approach, but rather the targeting of critical concepts by juxtaposing contrasting information designed to enable discovering the causes of outcomes. The causal-contrast approach is also distinct from subgoal-learning in that it does not explicitly mention what the subgoals are. Instead, it recruits the student's natural causal reasoning to construct

the goals and subgoals within a causal structure that supports the application of solution steps.

Traditional instructional approaches teach students analytically and explicitly the rules and steps for solving specific types of problems. Figure 1 illustrates how this approach teaches how to solve quadratic equations.

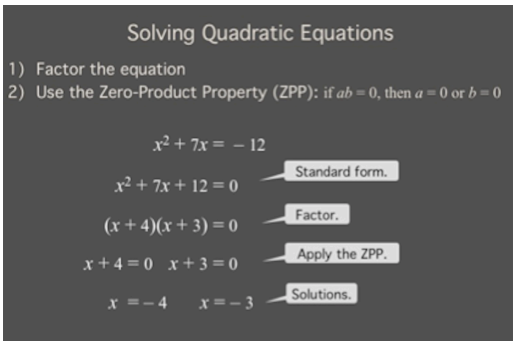


Figure 1. A cumulative static screen shot of sequentially presented and narrated traditional instructions.

For example, a student presented with the top equation in the figure is shown how to re-arrange the equation into standard form (second equation in the figure), factor the expression on the left-hand side, then determine the possible values of x, the desired unknown, using the zero-product property (i.e., if a • b = 0, then a = 0 or b = 0). For a substantial fraction of students, as evidenced by their failure to flexibly generalize their learning to novel problems, this approach does not lead to an understanding of the causal structure of the solution, the reason behind each step in the procedure and how the steps work together. For example, what is the purpose of factoring the expression on the left-hand side of the equation? And what is the relationship between factoring and the zero-product property?

Opportunities to compare worked examples (Rittle-Johnson & Star, 2007) to explain solutions (e.g., Chi et al., 1989; Chi, 2000) or to isolate subgoals (e.g., Catrambone, 1998) may enable understanding of the causal structure of a problem. However, even then a student may not identify all the concepts in the causal structure, or have all the requisite information to infer the causal relations when the concepts are identified. While Sfard (2007) emphasizes the use of teacher-student discourse (students compare their own solution to teacher's and explain the differences) to help students identify and correct misconceptions, the students may not discover the purposes of the concepts critical to transfer their success to novel problems.

The causal-contrast approach more directly targets specific links in the causal structure that make use of critical mathematical concepts essential to the solution of relevant types of problems. For example, students are first asked to try solving a quadratic equation (see top equation in Figure 2). If they fail to solve it, they are presented with the next two equations in the figure and asked to solve them. After students solve these, they are asked why these problems are easier for them to solve than the top equation. This comparison enables students to readily discover a cause of their failure: unlike the second and third equations, the top equation has both an x and an x^2 term, preventing one from isolating x by simply rearranging the equation. The newly formulated cause – namely, having both x and an x^2 terms – in turn becomes an "effect" for a subsequent operation, namely, factoring, to remove. Without their initial attempt to isolate x, learners would have no "effect" for which to discover its cause. This effect – failure to isolate x by rearranging the equation – is often not explicitly noted in traditional instruction. Similarly, without comparing the top equation with the next two equations, a learner who is asked to explain why the top equation is difficult may mention the effect alone, "I'm unable to rearrange the equation to isolate x", omitting to identify its "cause": having both x and x^2 terms in the same equation. Thus, each comparison is designed to direct attention to the accurate formulation of an essential causal relation in the structure of the solution.

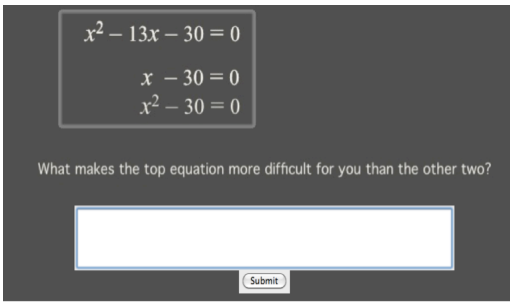


Figure 2. A screen shot of the causal-contrast intervention.

The process repeats as the student proceeds through the solution. In Figure 3, the pair of equations enclosed by the light gray rectangle illustrates the use of causal contrast for inferring a perceptual feature that causes a difficulty. Students are asked to solve the factored form of the quadratic equation (see top equation inside rectangle). If they fail to solve the problem, they are asked to solve a variant of the problem that has the difficulty removed (bottom equation inside rectangle), and to answer the question, "What values of x would make $x \cdot y = 0$ true, regardless of the value of y?" In this simple form, most students have no difficulty. They are then asked to compare the two equations. The comparison may allow them to discover why they initially failed, for example, that the perceptual complexity of the product prevented them from recognizing that the expression on the left-hand side is a product, that inside each pair of parentheses is "just a number". At this point, the perceptual complexity can switch causal roles, taking on the role of an effect that the student can prevent or remove in the future, for example, by pausing to take note of the perceptual complexity and to register the perceptual cues indicating when the zero-product property applies.

In more general terms, the comparisons we construct provide *conditional contingency* information (Cheng & Holyoak, 1995) consisting of the state of an effect (e.g., succeeding to solve a problem or not) in the presence of a candidate cause (a feature of the math problem, e.g., an equation having both an x and an x^2 term) and in its absence

(an equation not having both an x and an x^2 term), with alternative causes held constant.

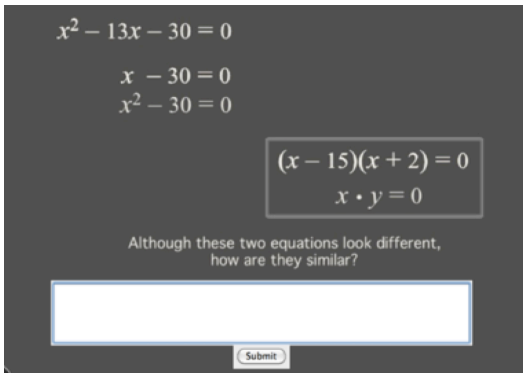


Figure 3. A later causal-contrast screen shot.

Comparing Causal-Contrast and Traditional Instruction

We conducted the present study on community-college students using a pretest/ intervention/posttest design to test a computer-based version of the Causal-Contrast instructional method (*CC*) against two computer-based traditional instructional methods and a baseline condition that received no instruction. The computer programs both animated and narrated the teaching materials for clearer and less attention-demanding instructions, reducing learners' cognitive load (Mayer & Moreno, 2003). By equating the instructions and feedback across conditions in the programs, we eliminated potential bias due to the human experimenter interacting with participants in Walker et al. (2014).

Although the causal approach was found to benefit both university students (at UCLA) and community-college students (Walker et al., 2014), the latter are less likely to benefit from traditional mathematics education. Causal induction is an implicit and evolutionarily old process (Hollis, 1997) and should be more uniformly available across the population than explicit reasoning processes.

One of the *Traditional* conditions (*T*) used materials identical to the causal contrast condition except for the instructional approach. The other traditional condition (*Khan Academy; K*) used an algebra teaching video from Khan Academy, a popular and well-regarded online educational website representative of traditional teaching. Khan Academy has become the largest school in the world; over 10 million students have watched its teaching videos online (Noer, 2012). Additionally, a *Baseline* condition (*B*) that did not receive any instruction was used as a control. All participants received information on the same mathematical concepts, solved the exact same set of problems with the same feedback.

The causal-contrast hypothesis predicts that students receiving the causal-contrast instructional intervention would have better performance in the post-test compared to students in the two traditional intervention conditions and in the baseline condition.

Participants

Sixty-eight community-college students recruited in Southern California and Downstate New York participated in our study. Participants (N=68) were randomly assigned into one of the three experimental conditions: Causal-Contrast (N=17), Traditional (N=18), Khan Academy (N=18), or the Baseline (N=15) control condition.

Study Design

This study consisted of two sessions, which were scheduled one to three weeks apart. In the first session, all participants were given a pretest measure assessing their knowledge of algebraic notation and of solving quadratic equations. Participants in the three experimental conditions received a lesson on solving quadratic equations on a computer followed by practice with an identical set of problems with identical feedback. For the baseline control group, immediately after the pretest, without receiving any instruction, participants were randomly assigned to receive one of two highly similar problem sets (Set A and Set B), which were the post-tests given to the experimental groups. The two sets consisted of analogous problems, and both assessed students' algebra problem solving mostly on quadratic equations. There was no time limit on the pretest and post-test, and the intervention lasted 25 minutes on average across the three experimental conditions. The groups did not differ in intervention time: $M_{\text{Causal-Contrast}}$ = 24.4, sd = 8.76, $M_{\text{Traditional}}$ = 23.3, sd = 7.55, and $M_{\text{Khan Academy}}$ = 29.5, sd = 12.3, F (2, 50) = 2.05, p = .14. After a one- to three-week delay, participants in the experimental conditions were given one or the other post-test set alternately in the second session, and those in the *Baseline* condition received their second post-test set.

Instructional Materials

Causal-Contrast Approach. The causal-contrast instructional materials are as explained earlier. Participants in the causal-contrast approach were given three challenging problems that students tend to fail to solve. If the participants failed to solve a problem, a feedback slide informed them that they solved the problem incorrectly. A branching function was used to identify the nature of the failure (e.g., lack of understanding vs. careless mistakes) to enable the appropriate feedback. Instead of showing the correct solution, a *contrasting problem* was then presented. If the student solved a challenging problem successfully, the next challenging problem was presented.

Traditional Instructional Method. The instructional materials used in this condition were designed to make use of techniques that are representative of traditional instruction and were based on a popular textbook (Sullivan & Sullivan, 2007). Participants in this condition were shown step-by-step procedures through *worked examples*, written solutions of example problems that provide justifications for each procedural step (see Figure 1). The examples were followed by practice problems. Combining worked

examples and problem solving has been proved to facilitate learning (Sweller & Cooper, 1985). The instructions stated the subgoals of the critical steps in the procedure; for example, subjects were told that a quadratic equation is rearranged to standard form (having a zero on one side and a polynomial on the other) *so that* it could be factored and solved using the zero-product property. Emphasizing subgoals in problem solving has been shown to promote learning and transfer (Eiriksdottir & Catrambone, 2011).

Khan Academy. The instructional material for this condition was an online video from Khan Academy's website: https://www.youtube.com/watch?v=uktzcTg_N7U. It makes use of animated digital technology designs. The video covers the necessary techniques for solving a quadratic equation, including factoring techniques and use of the zero-product property.

Measures

Post-test. The post-test included two types of problems: *instructed* and *transfer*. Instructed problems could be solved using the same solution procedures as the study problems. Transfer problems required generalization of concepts learned in the intervention. These problems included factorable quadratics in non-standard form. Table 1 lists the transfer problems:

Table 1. Transfer Problems

Q1	$4x^2 = 16x^4$
Q2	$(x-3)(x^2-4) = 0$
Q3	$\sqrt{x-3} + 5 = 9$
Q4	$(x-1)^2 = 4(x-1)$
Q5	$(x+3)(x-2) = 6$
Q6	$(2x+1)(2+x)x^2 = 0$
Q7	$\frac{4}{3}x^2 = 2x$
Q8	$(x-3)(x-7) + (x-3) = 0$

Results

Pretest

A one-way between-subjects ANOVA performed on participants' pretest performance shows that the four groups did not differ in their pretest performance: $M_{\text{Causal-Contrast}}$ = 73.2 (sd = 19.4), $M_{\text{Traditional}}$ = 76.1 (sd = 16.0), $M_{\text{Khan Academy}}$ = 68.6 (sd = 18.1), and M_{Baseline} = 70.3 (sd = 25.0), F (3, 64) = .497, p = .686).

Delay Interval

A one-way between-subjects ANOVA performed on time between instruction and post-test (in days) shows that the four conditions did not differ in their average delay time: $M_{\text{Causal-Contrast}}$ =14.5, sd = 5.37, $M_{\text{Traditional}}$ = 13.1, sd = 5.61, $M_{\text{Khan Academy}}$ = 12.9, sd = 4.83, and M_{Baseline} = 13.9, sd = 5.85, F (3, 64) = .318, p = .812.

Post-test

Explanation of Analyses. Before evaluating the effects of instruction type on participants' post-test performance, a correlation analysis was conducted to test whether the post-test performance was correlated with the pretest score. Pearson's correlation confirmed that the pretest and post-test scores were strongly correlated, r (68) = .711, p < .01, suggesting that the participants' post-test performance was in part due to their prior mathematical knowledge as indicated by their pretest scores.

Baseline Group's Post-test Sets A and B. To examine whether our two post-test sets in the *Baseline* group produced different performance, a one-way repeated measures ANCOVA with pretest as a covariate was conducted on the post-test scores, separately for the instructed problems and for the transfer problems. For neither type of problems was there a significant difference between the scores of Post-test Sets A and B, F(1,13) = .090, p = .768 ηp^2 = .007 and F (1,13) = 1.15, p = 0.303, ηp^2= .081 for the instructed and transfer problems respectively. Because the two sets produced highly similar performance, our analyses below collapsed across both sets.

Comparison across Conditions. Because the effect of mathematical instructions is strongly influenced by learners' prior knowledge (Clarke, Ayres & Sweller, 2005; Rittle-Johnson et al., 2009), to show a clearer picture of the effect of the interventions, we separated the participants into three groups based on their pretest scores: *low-pretest* (i.e., pretest < or = 50%), *medium-pretest* (i.e., 50% < pretest < 90%), and *high-pretest* (i.e., pretest > or = 90%). Tables 2, 3, and 4 show the respective means of pretest scores, instructed problems scores, and transfer problems scores in each subgroup for each intervention condition.

Table 2. Descriptive Statistics for Pretest Performance

		Causal-Contrast	Traditional	Khan Academy	Baseline
Low Pretest	Mean	37.5	50.0	43.8	31.7
	SD	17.7	0.00	2.50	20.2
	N	2	3	4	3
Medium Pretest	Mean	70.5	76.8	70.0	70.7
	SD	10.9	10.3	8.40	12.1
	N	10	11	11	7
High Pretest	Mean	93.0	93.8	96.7	93.0
	SD	4.47	2.50	2.89	2.74
	N	5	4	3	5

As shown in Table 2, there were relatively few low-pretest participants. Relative to other groups, these participants' pre-test scores varied substantially across conditions, making the assessment of the effectiveness of training in this group problematic. In view of the nature our pretest problems, these participants' pretest performance suggests that they lack sufficient arithmetic or algebra knowledge to benefit from the interventions. In contrast, the high-pretest participants can potentially still benefit because the transfer problems are considerably harder than the pretest problems.

In the following analyses, to better assess the instructional interventions in their effective range, we excluded the low-pretest participants. We combined the rest

of the participants because of our small sample sizes. Figure 4 shows these participants' post-test performance results.

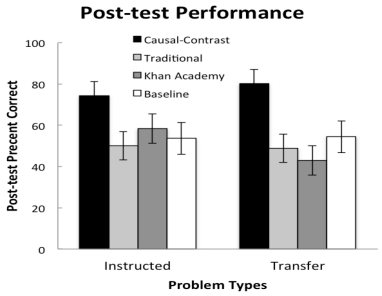


Figure 4. Estimated marginal means for medium- or high-pretest participants' Post-test scores.

Instructed Problems. A one-way ANCOVA on the post-test scores using pretest score as a covariate shows no significant difference in scores among the four conditions, F (3, 51) = 2.37, p = .081, ηp^2= .122, perhaps due to our small sample size.

Table 3. Descriptive Statistics for Instructed Problems

		Causal-Contrast	Traditional	Khan Academy	Baseline
Low Pretest	Mean	12.5	4.17	25.0	4.17
	SD	17.7	7.22	35.4	7.22
	N	2	3	4	3
Medium Pretest	Mean	62.5	48.8	44.3	46.4
	SD	22.8	35.1	33.2	35.9
	N	10	11	11	7
High Pretest	Mean	95.0	65.6	91.7	67.5
	SD	11.2	25.8	14.4	24.4
	N	5	4	3	5

Transfer Problems. A one-way ANCOVA with pretest score as a covariate conducted on the transfer problems shows a significant difference among the intervention conditions, F (3, 51) = 8.22, p < .001, ηp^2= .326. Follow-up pairwise comparisons using the Bonferroni correction indicate that the *CC* group outperformed the *T* group (p = .002), the *K* group (p < .001), and the *B* group (p = .023). The CC group transferred their knowledge about as well as UCLA students given similar training in our previous studies (Walker et al., 2014), with mean scores ranging from 80% to 82%. There was no statistically significant difference between the other three groups, p > .50.

Table 4. Descriptive Statistics for Transfer Problems

		Causal-Contrast	Traditional	Khan Academy	Baseline
Low Pretest	Mean	2.78	3.71	18.1	1.95
	SD	3.93	3.21	28.8	3.21
	N	2	3	4	3
Medium Pretest	Mean	75.0	44.4	32.3	38.9
	SD	14.1	32.3	29.30	34.8
	N	10	11	11	7
High Pretest	Mean	87.8	73.6	63.0	80.0
	SD	8.24	24.2	8.48	14.4
	N	5	4	3	5

Figure 5 shows performance on individual transfer problems for the four intervention groups. The CC group consistently outperformed the other three groups.

Comparing the frequency of participants who solved a problem correctly versus incorrectly, we see that the *CC* group greatly outperformed the other three groups on Transfer Problem 6, $\chi^2(1, N=56) = 8.13$, p=.004. This superiority is notable in that the CC instructions could have misled the participants into formulating a simplistic rule regarding the joint presence of x and x^2, one that ignores whether they are terms or factors. They could have been stymied, for example, because the factored equation still contains both x and x^2 terms, as they saw in the pair of contrasting problems in Figure 2. Instead, the intervention enhanced their correct flexible use of the relevant concepts and procedures.

Performance on Transfer Problem 3 is also notable in that this problem does not involve a quadratic equation. And yet the *CC* group outperformed the other three groups, $\chi^2(1, N=56) = 9.71$, p=.002. The *CC* group's deeper understanding of the causal structure of solving quadratic equations might have allowed them to reason more flexibly on another type of algebra problem. An intriguing possibility, which of course requires further research, is that the causal-contrast approach may awaken students' natural causal inference processes so that they create and test their own causal contrasts as they encounter new mathematical domains.

This possibility was collaborated by the CC group's superior performance on Transfer Problem 7, $\chi^2(1, N=56) = 6.21$, p=.013. This problem involved an x^2 term with a fractional coefficient, and there was no training on fractional coefficients in any of the interventions.

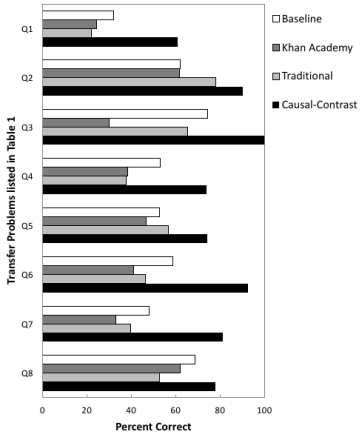


Figure 5. Performance on transfer problems in the four conditions.

Discussion

The causal contrast instructional approach invokes learners' natural causal induction process to identify the cause-and-effect relationships in a mathematical procedure. This approach decomposes the causal structure in a mathematical problem and accordingly caters the learning materials to allow learners to formulate the purposes of mathematical operations. These purposes are often not explicitly mentioned in traditional mathematical training.

Our current findings testing community-college students replicated and extended the results of previous studies

conducted on university students and community-college students (Walker et al., 2014). These studies show large improvements in solving algebra problems due to causal-contrast training. The improvements were especially dramatic in the community-college students.

Our comparison conditions now included a baseline condition and instructions from the Khan Academy as a benchmark. We eliminated the potential for experimenter-bias by incorporating all materials as animated computer programs, with no experimenter-participant interaction during the interventions.

Our results show that even when explicit analytic instruction focuses on teaching the reasons for mathematical procedures, students still often fail to learn in a way that promotes generalization to novel problems after a one-to-three week delay. In contrast, by allowing students to use an implicit, empirical learning process to discover the causal structure of solutions, students are able to fill in the missing links of their causal structure, need not rely on their rote memory of procedures, and become able to flexibly use their mathematical knowledge.

References

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology*, 86(3), 360.

Caspi, S. & Sfard, A. (2012). Spontaneous meta-arithmetic as a first step toward school algebra. *International Journal of Educational Research.* 51-52, 45-65.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology*: General, 127(4), 355.

Cheng, P.W.(1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367-405.

Cheng, P.W., & Holyoak, K.J. (1995). Complex adaptive systems as intuitive statisticians: causality, contingency, and prediction. *In Comparative Approaches to Cognitive Science, Ed*. HL Roitblat, J-A Meyer. Cambridge, MA: MIT Press.

Chi, M.H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In Glaser, R. (Ed.), *Advances in Instructional Psychology*. Mahwah, NJ: Lawrence Erbaum Associates.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: how students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182.

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development*, 53(3), 15-24.

Cooper G. & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79(4), 347-62.

Eiriksdottir, E., & Catrambone, R. (2011). Procedural instructions, principles, and examples how to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(6), 749-770.

Gopnik, A., Glymour, C. Sobel, D. Schulz, L. Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 1-31.

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help? *Journal of experimental child psychology*, 107(1), 15-30.

Hollis, K. L. (1997). Contemporary research on pavlovian conditioning: A "new" functional analysis. *American Psychologist, 52*(9), 956-965.

Leslie A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?" *Cognition*, 25, 265-288.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1), 43-52.

Noer, M. (2012, November). One man, one computer, 10 million students: How Khan Academy is reinventing education. *Forbes*. Retrieved from http://www.forbes.com.

PISA (2003, 2006, 2009, 2012). Program for International Student Assessment. International Association for the Evaluation of Educational Achievement.

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, 35(1), 28-43.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, 99(3), 561.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103(4), 759.

Sfard, A. (2007). When the rules of discourse change, but nobody tells you: Making sense of mathematics learning from a commognitive standpoint. *Journal for Learning Sciences,* 16(4), 567-615.

Stigler, J. W., Givvin, K. B., & Thompson, B. J. (2010). What community college developmental mathematics students understand about mathematics. *MathAMATYC Educator*, 1(3), 4-16.

Sullivan, M.I., & Sullivan, M. (2007). *College algebra: Concepts through functions*. New York: Pearson Education.

Sweller, J. & Cooper, G. (1985) The use of worked examples as a substitute for problem solving in learning Algebra. *Cognition and Instruction*, 2, 59-89.

Walker, J. M., Cheng, P.W., & Stigler, J.M. (2014). Equations are effects: Using causal contrasts to support algebra learning. In P. Bello, M. Guarini, M. McShane, & B. Scasellati. (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

A connectionist model for automatic generation of child-adult interaction patterns

Moinuddin M. Haque (m.m.haque@uvt.nl), Paul Vogt (P.A.Vogt@uvt.nl),
Afra Alishahi (A.Alishahi@uvt.nl), Emiel Krahmer (E.J. Krahmer@uvt.nl)
Tilburg Center for Cognition and Communication,
Tilburg University
PO Box 90153, 5000 LE Tilburg, the Netherlands

Abstract

This study introduces a neural network that models the social interactions from a video corpus. The corpus consists of recordings of naturalistic observations of social interactions among children and their environment. The videos are annotated multimodally including features like gestures. We explore how this video corpus can be utilized for modelling by training our model on a portion of the annotated data extracted from the corpus, and then by using the model to predict novel interaction sequences. We evaluate our model by comparing its automatically generated sequences to an unseen portion of the corpus data. The initial results show strong similarities between the generated interactions and those observed in the corpus.

Keywords: Neural Networks; Child Language Acquisition; Sequence Generation; Modelling Social Interactions.

Introduction

Children adapt to their environment and communication partners trough interaction. These interactions along with the linguistic information are richly augmented with social cues (such as eye gaze and gestures), and are suggested to facilitate child language developement (Tomasello & Todd, 1983; Iverson, Capirci, Longobardi & Caselli, 1999; Hollich, Hirch-Pasek & Golinkoff, 2000). Therefore, cognitive models of child language learning should take these cues into account and integrate such interaction-based features with linguistic input in the process of learning language.

One of the well-studied mechanisms for learning the meaning of words is cross situational learning (Quine 1960), which draws on word-referent co-occurrences that children observe from their environment. Many computational models have implemented (variations of) this mechanism using associative networks to predict a word form based on semantic features (e.g., Li & Farkas, 2002; Regier, 2005) or to discover statistical regularities in observations of linguistic labels and visual features or concepts (e.g., Siskind, 1996; Frank, Goodman & Tenenbaum, 2007; Fazly, Alishahi & Stevenson, 2010). Typically, such models treat learning as a unidirectional process where the learner focuses only on the linguistic cues and discards the social interactions. We refer to these models as *data-driven models of word learning*. To date, only a few models of word learning have incorporated interaction features. The data-driven models by Yu and Ballard (2007), and Frank, Tenenbaum and Fernald (2013) have incorporated information about eye gaze and pointing in a cross situational learning model. Another example of a data-driven model where multiple modalities are incorporated is the work Matusevych, Alishahi and Vogt (2013). In this model features, such as occurrence frequencies of utterances, utterance types, action types, action arguments along with participants and objects in the visual context, were used to simulate interactions in the context of playing with toys.

The language game model of Steels (2003) is an example of an *agent-based model* in which agents interact with each other, exchange utterances and can learn from each other. Typically, such models have been used to study language evolution. Various language game models have been used to investigate vocabulary development, also incorporating cross-situational learning (Smith, 2005; Steels & Loetzsch, 2008; Vogt & Haasdijk, 2010). However, these models tend to implement interactions between agents using toy languages and do not reflect naturalistic interaction patterns. Considering the restrictions of data-driven and agent-based approaches to word learning, the next natural extension is to integrate the two approaches, but the problem is finding large and rich datasets for training such models.

Large-scale corpora of child-adult conversations, such as CHILDES (MacWhinney, 2000), are available and provide us with naturalistic linguistic exchanges between children and adults. However, most of the corresponding audio and video files are not annotated with extra-linguistic (e.g., semantic information about the surrounding scene) or interaction-based cues (such as gaze and gesture) that are machine readable.

The CASA MILA corpus, which consists of longitudinal video recordings of 40 children interacting in naturalistic environments, is a corpus that has incorporated annotations for non-verbal social cues (Vogt & Mastin, 2013a). The video frames are richly annotated based on the observed interactions between children and their caregivers, and thus provides a valuable resource for modeling child-adult interaction. However, it only covers 1.5 hours of recording for each child, which is hardly enough for training a computational model of child language development. What we need is an automatic input-generation engine that can replicate the interaction patterns and their statistical properties observed in a corpus such as CASA MILA, without the quantitative limitations of such a corpus.

The current paper presents a study to generate novel interactions based on observations from the corpus. One approach to create more data can be simply by copying the already annotated data multiple times. The limitation of this approach is that there will be no new interactions present in

the data, and the rigidity of the interactions remains. Yet, the flexibility of having new interaction patterns in the data is one of our aims. We therefore present a method for discretizing the continuous video and annotation data, and using these data to train a neural network that generates new interactions. We evaluate this method by analysing the newly generated sequences against the original annotated sequences, as well as three different baseline models.

Methods

Data

The CASA MILA corpus contains video recordings of 40 different children at home, interacting with one or more communication partners. These recordings are from three different cultures: rural and urban Mozambique, and the Netherlands. Each recording contains naturalistic observations of interactions between infants and their communication partners at their home. The corpus is longitudinal in nature, with recordings taken at children's ages of 13-, 18- and 25-months old.[1] For this study we will only use the data from the 13-months old children from the Netherlands. This was done, because these mainly contained one-to-one interactions, thus simplifying our problem.

The corpus was annotated for a variety of tiers, five of which we use for the present study: child engagement, child-directed speech, child-directed gesture, child-speech and child-gesture. In the recordings we only focus on those parts in which the child interacted with someone. The annotated features are hierarchically organized. On the top layer is the child's joint engagement level (Mastin & Vogt, 2016), as described in Table 1.

Table 1. Joint engagement levels.

Name	Description	Example
Persons engagement	Infants interact with another person by responding to the other person or by trying to start an exchange.	The infant responds with a smile to the mother's voices; infant reaches toward the mother.
Passive joint attention	Infants play with an object that is also the focus of another person's attention but they do not acknowledge the other person's attention.	The infant plays with a toy car. The mother says: "What a nice car!", but receives no response at all from the infant.
Shared joint attention	Infants share the attention to an object or event with the interlocutor, but they do not share a mutual goal in the interaction.	The mother offers the infant a toy to play with, the infant looks from the mother to the toy, but does not respond otherwise.
Coordinated joint attention	Infants share the attention to an object or event with the interlocutor, and they clearly share a mutual goal in the interaction.	The mother offers the infant a toy to play with, the infant looks to the mother and the infant takes the toy and starts playing.

At the lower layers, speech and gestures (e.g., pointing, showing or reaching) are annotated for both caregiver and child.[2] In this study, we let the machine learn to predict the joint engagement level, and when the children and communication partners produce utterances and/or gestures, but not what words or gestures are actually used.

Feature selection

Figure 1, illustrates how the original annotations are transformed to provide the learning algorithm a simplified representation of the input. Figure 1(a) shows a small fragment of an actual timeline showing the engagement level (top row), child-directed speech (second row), child-directed gesture (third row), child speech (fourth row) and child gesture (bottom row). The highlighted regions show where speech or gesture was observed.

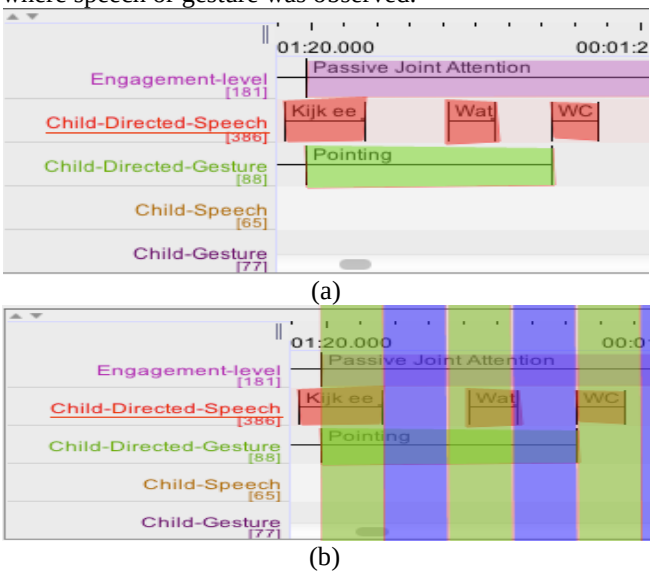


Figure 1. (a) Simplified annotated timeline with highlights. (b) Spliced timeline (not to scale).[3]

To process our data, the original time sequence is broken down into slices of 200 milliseconds duration, with the purpose of capturing the information in the annotations (Figure 1(b)). The duration of 200 milliseconds was determined after trial and error to capture significant information without having too many unchanged sufficient. These time slices allow us to represent the state of an interaction at time *t* as an 8-bit vector *x(t)*, where presence of activity is represented as 1 and absence as 0. To construct these vectors, engagement levels are represented with four bits, of which exactly one bit has the value of 1 at any given time (see Table 2, rows 1-4). Since we are only interested in predicting when someone speaks or gestures, the remaining four bits encode whether or not child-directed speech, child-directed gesture, child speech or child gesture was present at time *t*. This binary vector representation is then used to serve as input for our neural network.

[1] For more details on the recording procedures, consult Mastin and Vogt (2016) or Vogt, Mastin, and Schots (2015).

[2] Consult Vogt et al. (2015) for the transcriptions of speech, and Vogt and Mastin (2013b) for the annotation of gestures.

[3] Created with ELAN (Sloetjes & Wittenburg, 2008).

Table 2. The bit vector representation.

Position	*Represents*	*Possible values*
1	Persons engagement	0/1
2	Passive joint attention	0/1
3	Shared joint attention	0/1
4	Coordinated joint attention	0/1
5	Child-directed speech	0/1
6	Child-directed gesture	0/1
7	Child speech	0/1
8	Child gesture	0/1

Model

We developed a neural network model to generate novel interaction sequences based on the naturalistic patterns observed in the annotated corpus, and evaluated this on how well the model can recreate an unseen sequence. Treating the bit sequence as a time series with each vector as a given state, the model is trained to predict the next possible stage given a previous set of states.

We used a non-linear autoregressive neural network with external input (NARX) (Haykin, 1999; Lin, Horne, Tiňo & Giles, 1996; Gao & Er., 2005), which is a class of neural networks that is well suited for training nonlinear systems and time series. The NARX is a recurrent dynamic network with feedforward connections enclosing several layers of the network. The network is used to predict the next value of the input signal. The NARX architecture was chosen over others for its success with predicting time series.

Since the true output is available during the training of the network, one can create a 'series-parallel architecture', in which the true output is used instead of feeding back the estimated output, as a series-parallel architecture (Figure 2). During training the network is set in the series-parallel architecture. Once the network has finished training the network is changed into the closed loop, standard NARX 'parallel architecture' to make it usable for predicting the next state in the evaluation phase (Figure 3). During this prediction stage the output is fed back to the input of the feed forward neural network. The output of the NARX network is an estimate of the output of the nonlinear dynamical system that is being modelled. Making this distinction has two advantages: First, in the series-parallel architecture the input to the feedforward network is more accurate. Second, this series-parallel architecture has a purely feedforward network, and static backpropagation can be used for training.

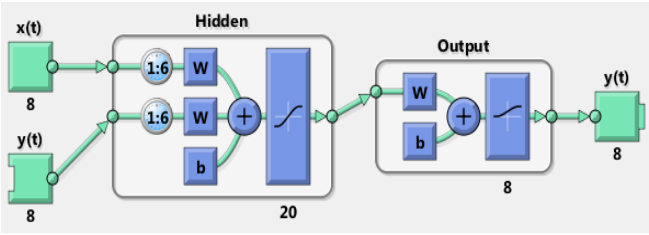


Figure 2: The Series-parallel Architecture of the NARX network in open loop.

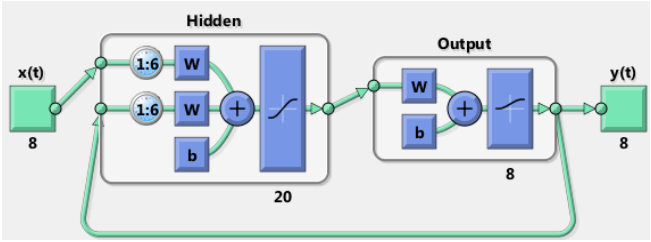


Figure 3: The Parallel Architecture of the NARX network in closed loop.

During training, the network receives two input vectors, *x* (input time series) and *y* (output time series) each represented as a vector with 8 bits. The input time series corresponds to the bit sequence we are training at time *t*. The output series corresponds to the observed output at t+1. To introduce a temporal memory the network delay is set to 6, meaning that each new vector is predicted based on the 6 previous input vectors. This parameter was derived empirically. The hidden layer has 20 neurons. For the transfer function in the hidden layer a hyperbolic tangent sigmoid transfer function was used. At the output layer a log-sigmoid transfer function is used to get the final result in the format of a matrix where each row corresponds to the target annotations.

The network is trained with a function that updates the weights and bias values according to Levenberg-Marquardt optimization (Levenberg, 1944). It minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well. The process is called Bayesian regularization. The error calculation is done using mean squared normalized error.

Experimental setup

From the corpus we constructed 2 different data sets. We will discuss each separately as study 1 and study 2. For both studies, data from 12 children in the 13-month age group of the Netherlands dataset was used.

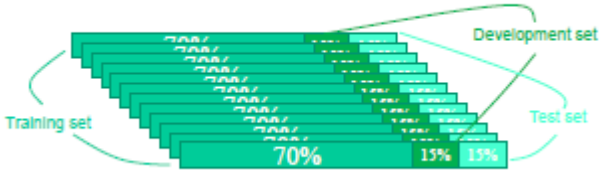


Figure 4: Division of data for study 1.

Study 1. Each of the children's data is broken into 3 distinct parts: a training section, a development section, and a test section, with a size of 70%, 15%, and 15% for each section respectively. The individual parts of the sections are joined together to form aggregated sets. The training set is used for training of the neural network. The development set is used

to fine tune the parameters of the network. The test set is used to evaluate the performance of the network. Figure 4 shows the division of data for study 1.

Study 2. The development set and training set are formed by selecting data from 11 of the 12 children. This data is broken into two sections per child. The development sections contained 15% of the data and, aggregated, these formed the development set. The training sections contained 85% of the data and formed the training set. Data from the 12th child is taken as the test set. Figure 5 shows the division of data for study 2.

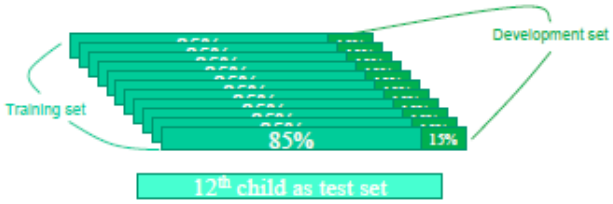


Figure 5: Division of data for study 2

Evaluation

Baselines. To evaluate the accuracy of the neural network generated sequences (NNG), we implemented three baseline models for predicting sequences. In each of these baselines the size of the generated sequence equals the length of the test set. The three baseline models are defined as:

- Complete random generation (CRG). This creates a sequence by putting together random slices taken from a set of all distinct slices observed in the training set.
- Attribute-based generation (ABG). This sequence is created by calculating the probability distributions for each of the attributes annotated in the individual slices (e.g., child gesture, mother gesture or engagement level). In this model it is assumed that the attributes are independent of each other and the predictions are based on their distribution probability only.
- Transition based generation (TBG). This is formed by calculating the transition probability for each of the unique slices, as observed in the training set. Once these probabilities are known a new discrete sequence is generated using the transition probabilities.

These three generations present us with a comparative baseline to test the effectiveness of the neural network generated sequences.

Transitions. To capture the patterns in interactions we will evaluate the transitions in interaction states. Transitions occur when at least one bit changes between two time steps *t* and *t+1*. For example, when a bit sequence "00010001" (representing child and mother having coordinated joint attention where the child is gesturing) is followed by "00011001" (representing child and mother having coordinated joint attention where the child is gesturing and mother is speaking).

Measures. To measure the performance of the neural network, we evaluated the sequences generated at a micro and a macro level. At the micro level, the generated sequence was aligned with the test set to check for matches. A match is said to occur when slices match each other exactly. We measured *Accuracy* as the percentage of the test set that had a perfect match.

For the macro level analysis we compared the generated *distributions* with the transition distribution observed in the test set. To do so we calculated the number of times a transition took place to create a probability distribution corresponding to each sequence. These probability distributions were then compared with the test set's transition probability distribution using the *Hellinger distance*, which was used to quantify the similarity between two probability distributions (Hellinger, 1909). The Hellinger distance forms a bounded metric on the space of probability distributions over a given probability space. Mathematically this is calculated by taking the square root of the distance between two vectors. The closer the Hellinger distance is to 0, the more similar two distributions are. The maximum distance of 1 is obtained when there is no overlap between the two distributions.

Results

Study 1

Accuracy. Figure 6 provides the accuracy of the different sequence generation models. As we can see, replicating the exact time series is difficult. The accuracy of the CRG (4%) is expectedly very low.

The low accuracy for ABG (11%) shows that the assumption that the attributes are independent of each other is too simplistic and that there are meaningful dependencies between them, which are useful for sequence generation.

The accuracy gain through NNG (21%) over the TBG (19%) is small. The TBG is very faithful to the training data and therefore makes few mistakes, but it cannot generalize beyond what it is has seen in the training data. To get a better understanding of the difference in the sequences generated by these two models we look at the Hellinger distance.

Time series distribution. Figure 7 shows the Hellinger distances calculated for each of the generated sequences. The macro level analysis displays a much better trend than accuracy measures. When looking at the macro-level analysis, the NNG has a Hellinger value of 0.30, which is considerable improvement over the baseline models. The Hellinger values for the baseline models are close to 1, meaning the compared distributions are dissimilar. In short the sequences generated by the NNG have more overall similarity with the test sets than those generated by the other baseline methods.

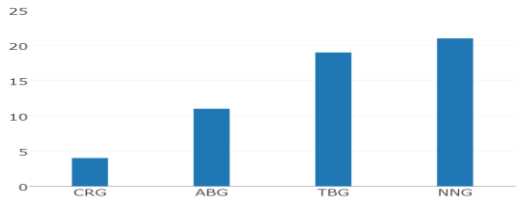


Figure 6. Accuracy measures for different generations.

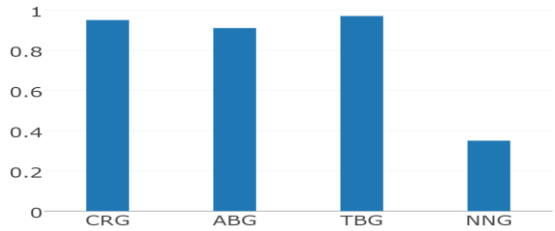


Figure 7: Hellinger Distance for different generations.

To continue our analysis, we provide a visualization of the transitions and their probabilities as observed. Figure 8 shows the probability distributions for the NNG sequence and the test set. We can see that the probabilities of the highly frequent transitions are closely replicated by the network, but it fails to replicate the low frequency transitions.



Figure 8: Transition probabilities in the test set (blue bars) and the NNG (yellow bars). The x-axis represents the different transitions, and the y-axis shows the probability.

Study 2

In the second study, we investigated how the model would behave when faced with data from a child it has not been trained on. Accuracy of the generated sequence for the NNG was approximately 5%, showing a much lower performance compared to study 1.

Doing the macro level analysis we found the Hellinger distance to be close to 1, which means that the generated sequences distribution is very different from the one observed in the test set. Although there are some matches for the slices produced, yielding accuracy greater than zero, the distribution of transitions differs.

Discussion

In this paper we took a novel approach to generating new multimodal interaction sequences based on corpus data, using neural networks.

The interaction sequences produced by the different generation techniques shows that the neural network generated patterns have higher accuracy and more similar transition distribution than the baseline methods. Although the TBG and NNG data have similar levels of accuracy, when it comes to the transitional distribution, we observe a large difference, showing that NNG generates interactions more similarity with the training data.

While the NNG yields the highest accuracy, it is still far from perfect. The reason for this is that the interactions the network is trained on represents a non-linear complex dynamical system, whose exact time series is extremely hard to replicate. Although neural networks have shown to be universal approximators, they have difficulty modelling time series. While NARX networks perform better than others when it comes to time series modelling, they still have problems learning long term dependencies due to vanishing gradients (Diaconescu, 2008). The behavior of the network is highly dependent on the size of the input sequence that represents the temporal memory of the model. However, the NARX model lacks a decent procedure for optimizing this size.

The Hellinger distance gives us a better performance, as it looks at the distributions of the series and not exact locations. For the NNG, the transitions observed in the generated sequence are more similar to the test set than the baseline sequences, but the distance obtained with the present method is still insufficiently close to zero. Highly frequent transitions in the interactions are fairly well replicated, but many transitions observed in the test set are not. Moreover, the network generates sequences that have not been observed. The reason for these discrepancies are likely due to the possibility that the test set contains transitions that have not occurred in the training data and vice versa. Further analysis is required to verify whether this is, indeed, the case.

For study 2, the Hellinger distance approaches 1, which means that the distributions of transitions are not well generalized for replicating interactions of an unseen child. One reason for this is that there are substantial individual differences between the interactions the children engage in (Vogt et al., 2015). So, although they interact using the same gestures in conjunction with speech, the frequencies with which gestures are used vary considerably, as do the sequences of interactions (Vogt & Mastin, 2013b). Since the network has a memory that helps generating sequences based on previous observations, testing the network on an unseen test set is likely the cause for the marked difference in the transition probabilities.

In order to reduce the complexity of the learning task, we chose to represent the speech and gesture in the annotated corpus as bitstrings indicating the presence or absence of speech or gesture. However, doing this reduced the amount of information contained in the training data. While maintaining all words and different gestures is likely too complex with the amount of data, using more informative categories of speech or gesture might improved the results.

To conclude, our initial attempt to replicate patterns of interactions between young children and their caregivers observed in a video corpus yields mixed results. Our method can replicate interactions of aggregated children reasonably well, but it cannot generalize to a previously unseen child, nor does it fail to replicate exact sequences in the time series well. In later stages we intend to introduce more features into the training along with exploring other modelling paradigms, such as incorporating the corpus data more directly in an agent-based model that can then be trained to interact following the patterns observed in the data.

Acknowledgements

This study was funded by the Netherlands Organization for Scientific Research (NWO) with a grant in the Natural Artificial Intelligence program.

References

Diaconescu, E. (2008). The use of NARX neural networks to predict chaotic time series. *Wseas Transactions on Computer Research*, *3*(3), 182-191.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017-1063.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1-24.

Frank, M., Goodman, N. D., & Tenenbaum, J. B. (2007, December). A Bayesian Framework for Cross-Situational Word-Learning. In *NIPS* (pp. 457-464).

Gao, Y., & Er, M. J. (2005). NARMAX time series model prediction: feedforward and recurrent fuzzy neural network approaches. *Fuzzy sets and systems*, *150*(2), 331-350.

Haykin, S. (1999). *Neural Networks, Second Edition*. Pearson Education.

Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210-271.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., ... & Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Child Language*, 43(2): 235-264.

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development*, *14*(1), 57-75.

Levenberg, K. (1944). A method for the solution of certain non–linear problems in least squares. 164-168.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *IEEE Transactions on Neural networks*, *17*(8), 1345-1362.

Lin, T., Horne, B. G., Tiňo, P., & Giles, C. L. (1996). Learning long-term dependencies in NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, *7*(6), 1329-1338.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database. *Computational Linguistics*, *26*(4), 657-657.

Mastin, J. D., & Vogt, P. (2016). Infant engagement and early vocabulary development: a naturalistic observation study of Mozambican infants from 1;1 to 2;1. *Journal of Child Language*, FirstView, 1-30.

Matusevych, Y., Alishahi, A., & Vogt, P. (2013). Automatic generation of naturalistic child–adult interaction data. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2996-3001).

Quine, W. V. O. (1960). *Word and object*: Cambridge University Press.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819-865.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39-91.

Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by Category: ELAN and ISO DCR. In *LREC*.

Smith, A. D. (2005). Mutual exclusivity: Communicative success despite conceptual divergence. *Language Origins: perspectives on evolution*, 372-388.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308-312.

Steels, L., & Loetzsch, M. (2008). Perspective alignment in spatial language. *Spatial Language and Dialogue*, 70-89.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, *4*(12), 197-211.

Vogt, P. & Haasdijk, E. (2010) Modelling social learning of language and skills. *Artificial Life, 16(4): 289-310*

Vogt, P., & Mastin, J. D. (2013a). Anchoring social symbol grounding in children's interactions. *KI-Künstliche Intelligenz*, *27*(2), 145-151.

Vogt, P., & Mastin, J. D. (2013b). Rural and urban differences in language socialization and early vocabulary development in Mozambique. In *Proceedings of the 35th annual meeting of the Cognitive Science Society*(pp. 3787-3792). Austin, TX: The Cognitive Science Society.

Vogt, P., Mastin, J. D., & Schots, D. M. (2015). Communicative intentions of child-directed speech in three different learning environments: Observations from the Netherlands, and rural and urban Mozambique. *First Language*, *35*, 341-358.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149-2165.

Connections between ACT-R's declarative memory system and Minerva2

Cvetomir M. Dimov (cvetomir.dimov@unil.ch)
Department of Information Systems, University of Lausanne
CH-1015, Switzerland

Abstract

As a first step towards applying ACT-R to problems of likelihood judgment, we draw parallels between ACT-R and HyGene. More specifically, in the spirit of theory integration, we demonstrate the relation between ACT-R's declarative memory system and the core of HyGene: Minerva2. We first start by transforming ACT-R's activation equations into what is in our view a more intuitive form. This form then allows us to more transparently see the correspondence of the effect of prior history between the two theories and of the current context between them. The results provide insights into the workings of the two theories and open an avenue for future attempts of theory integration, not only between the two theories, but also to related theories of memory. Moreover, we hope these results will be important steps toward testing ACT-R's capabilities of accounting for judgment phenomena.

Keywords: ACT-R, HyGene, Minerva2, declarative memory

Introduction

The cognitive architecture ACT-R (Anderson, 2007) is arguably the most advanced integrated theory of cognition at the moment. A testament to its generality is the broad spectrum of phenomena to which it has been applied. These include phenomena as diverse as analogy making (Salvucci & Anderson, 2001), past tense learning (Taatgen & Anderson, 2002) and solving the Tower of Hanoi puzzle (Anderson & Douglass, 2001).

One set of phenomena to which ACT-R has been applied are those in the field of judgment and decision making. Specifically, the memory system of this cognitive architecture has been used to explore the applicability of various decision strategies (Marewski & Schooler, 2011) and to investigate the properties of the fluency heuristic (Schooler & Hertwig, 2005). In addition, the full power of the architecture was used to construct 39 implementations of the recognition heuristic (Marewski & Mehlhorn, 2011) and identify those that best describe behavioral data.

A problem in the domain of judgment and decision making to which ACT-R was never, to our knowledge, applied is how people make probability judgments. Probably the most complete psychological theory of probability judgment is HyGene (Thomas, Dougherty, Sprenger, & Harbison, 2008). Thus, a first step in applying ACT-R to problems of probablity judgment would be to work towards the integration of ACT-R and HyGene by exploring the compatibility of the two theories.

HyGene extends MINERVA-DM (Dougherty, Gettys, & Ogden, 1999) by adding semantic memory and working memory storages. MINERVA-DM was first introduces to account for various likelihood judgment phenomena, such as frequency judgments and some judgment biases, and is itself based on the theory of memory Minerva2 (Hintzman, 1984). Thus, many of HyGene's properties stem from the very properties of Minerva2.

In the spirit of theory integration (Mischel, 2009, critisized the lack of theory integration efforts in psychology), we will demonstrate the close correspondence between ACT-R's declarative memory and Minerva2. We will first briefly introduce ACT-R's declarative memory equations and Minerva2. We will then present what is in our mind a more intuitive funcitonal form of activation - the exponent of activation. We will then derive the optimized learning equation in the exponentiated form and demonstrate the conditions under which it provides a good approximation to base-level activaiton. This will allow us to relate influence of prior exposure according to the two aforementioned theories of memory. We will then compare the effect of context on activation according to the two theories and again demonstrate the close correspondence between the two.

ACT-R's declarative memory system

ACT-R's declarative memory system integrates retrieval probability and retrieval time of a memory chunk through a single quantity: the memory activation of that chunk. Memory activation A_i of a chunk is modeled as a function of prior exposure to the object represented by the chunk (base-level activation), context relevance (spreading activation of chunk) and noise:

$$A_i = B_i + SA_i + \varepsilon = B_i + \sum_j W_j S_{ji} + \varepsilon, \qquad (1)$$

In this equation, B_i is the base-level activation, which reflects the recency and frequency of practice of the chunk i. ε is the noise value, which consists of permanent noise and instantaneous noise computed at the time of a retrieval request. We will only consider instantaneous noise below. W_j is the amount of activation from source j, while S_{ji} is the strength of association from source j to chunk i.

The equation describing learning through base-level activation for chunk i is:

$$B_i = \ln \sum_{j=1}^{n} t_j^{-d} \qquad (2)$$

where n is the number of presentations for a chunk i, t_j is the time since the jth presentation, and d is the decay parameter. According to the base-level equation, each time an item is presented, its activation increases. Then, with time, its activation decays, described with a power function .

The strength of association S_{ji} between two chunks j and i is 0 if chunk j is not the value of a slot of chunk i (also,

j and i should not be the same chunk). If the two chunks have common slots, the association strength is set using the following equation:

$$S_{ji} = S - ln(fan_j) \tag{3}$$

In this equation, S is the maximum associative strength and fan_j is the number of chunks in declarative memory in which j is the value of a slot. In addition, 1 is added for each chunk (in this case, chunk j) being associated with itself. Intuitively, the amount of spreading activation from a slot of the chunk currently in the focus of attention is distributed equally among all chunks with that slot in declarative memory.

A brief introduction to Minerva2

Minerva2 represents stimuli as feature-lists, which get stored as episodic traces in memory upon encoding. Each repetition procudes a new trace of that item. Upon cued retreaval, all episodic traces in memory are contacted and compared to the cue, and produce a cumulative output in working memory called an echo. In the echo, each trace is weighted by its activation.

Activation A(i)of an episodic trace in Minerva2 is a cubic function of its similarity to the cue. It is expected that the activation of a trace is close to 0 for a randomly chosen trace and it increases superlinearly to 1 when the cue and the trace are the same.

Exponentiating memory activation

We consider that for the current purposes the exponent of memory activation is a more intuitive quantity to work with than memory activation on its own:

$$A_i^e = e^{A_i} = e^{B_i + SA_i} = e^{B_i} e^{SA_i} = B_i^e SA_i^e \tag{4}$$

Base-level activation

In the following we will use the non-log-transformed version of base-level activation B_i^e (Anderson & Schooler, 1991; Anderson, Fincham, & Douglass, 1999), because it demonstrates the effect of prior exposure on activation more transparently:

$$B_i^e = \sum_{j=1}^{n} t_j^{-d} \tag{5}$$

Derviation of the optimized learning equation To derive the optimized learning equation, we suppose that we know the time since the chunk was created, called its lifetime L. Lets also suppose that it has been periodically strengthed n times since its creation with a period of $T = \frac{L}{n}$. Given these assumptions, we can derive a new form of the base-level activation:

$$B_i^e = T^{-d} \sum_{j=1}^{n} j^{-d} \tag{6}$$

We can approximate the sum above with an integral (Anderson et al., 1999), and arrive at the simplified equation:

$$B_i^e = T^{-d} \frac{n^{1-d}}{1-d} \tag{7}$$

At first sight, it seems that the base-level activation is a power function of n (Anderson et al., 1999). Note however that T is also a function of n. Taking this into account, we arrive at the final expression for B_i^e:

$$B_i^e = \frac{L^{-d}}{1-d} n \tag{8}$$

This is (the exponent of) the optimized learning equation, available, for example, in ACT-R's documentation. Note that it is a *linear function* of n, discounted only by the chunk's lifetime L.

Generality of the optimized learning equation Some might argue that equation 8 rests on the rather strict assumption of periodicity. To test how much of a restriction that is, lets start with the following "noisy" base-level activation equation:

$$B_i^n = \sum_{j=1}^{n} (jT + j^{1+d}\varepsilon_j)^{-d} = \sum_{j=1}^{n} (jT)^{-d} \left(1 + \frac{i^d \varepsilon_j}{T}\right)^{-d}, \tag{9}$$

where the noise j periods ago is $j^{1+d}\varepsilon_j$.

Using the approximation $(1+a)^n \approx 1+na$, we can separate the noise term from the rest:

$$B_i^n = \sum_{j=1}^{n} (jT)^{-d} \left(1 - \frac{j^d d \varepsilon_j}{T}\right) = B_i - dT^{-(1+d)} \sum_{j=1}^{n} \varepsilon_i \tag{10}$$

If ε has an expectation of 0, then our noisy base-level activation equation will be unbiased. Notice that the noise that we added grows as a *superlinear* function of the number of periods j.

Spreading activation

According to ACT-R the slots of the chunk present in the imaginal buffer spread activation to chunks in declarative memory. The exponentiated spreading activation is equal to:

$$SA_i^e = e^{SA_i} = e^{\sum_j \frac{w_{im}}{n_{im}} S_{ji}} = \prod_j e^{\frac{w_{im}}{n_{im}} S_{ji}}, \tag{11}$$

where w_{im} is the total amount of spreading activation from the imaginal buffer and n_{im} is the number of slots in the imaginal buffer's chunk. Intuitively, the total amount of spreading activation that the imaginal buffer can provide is *equally* distributed among its slots.

What remains to unpack is S_{ji}, which is the difference between the maximum possible amount of spreading activation S and the number of chunks in declarative memory with that slot value fan_j. For simplicity we set $w_{im} = n_{im}$ so that these two cancel each other out. The contribution to spreading activation from a slot is:

$$e^{S_{ji}} = \frac{e^S}{fan_j}, \tag{12}$$

Note that the total spreading activation from a chunk slot in the imaginal buffer is *equally* distributed among all chunks in declarative memory which contain that slot (Anderson & Reder, 1999, discuss the general equation for spreading activation).

The effect of prior history on activation according to ACT-R and Minerva2

In this section we will use equation 8 to draw parallels between memory activation according to Minerva2 and ACT-R. We will consider only the effect of prior history. For Minerva2 this means a constant similarity S. For ACT-R will assume that there is constant spreading activation.

Minerva2 encodes each observation as a separate trace. When memory is probed, each trace in memory contributes with a certain activation to the echo intensity and content. Given a similarity, the activation of a trace is $A = S^3$. For an occurrence frequency of an event n, we will have a total activation of the n traces related to that event of $A^{tot} = nA = nS^3$. That is, the relative total activation of all traces related to two events is simply a function of their relative occurrence frequencies:

$$\frac{A_1^{tot}}{A_2^{tot}} = \frac{n_1 S^3}{n_2 S^3} = \frac{n_1}{n_2} \tag{13}$$

According to ACT-R, using equation 8, the relative activation of two chunks is:

$$\frac{B_1^e}{B_2^e} = \frac{L^{-d}(1-d)}{L^{-d}(1-d)} \frac{n_1}{n_2} = \frac{n_1}{n_2} \tag{14}$$

It is important to emphasize that this result holds only for chunks with the same lifetime L. However, this is not necessarily a strong limitation to our result: In learning experiments often all pieces of information have the same lifetime.That is, learning usually starts at the same point - in the beginning of the experiment. And even if this is not the case, as demonstrated, there is some variability in the starting point allowed. In addition, it is often true for real-life memories such as general knowledge facts that they have similar lifetimes.

When there is considerable departure from this assumption, one can plug in the actual lifetime of memories to estimate their relative activation. For example, let us assume that one chunk of memory is 4 times older than another. That is, we assume that $L_1 = 4L_2$. Let us also take the standard value of the decay parameter $d = 0.5$. For these values the relative activation of the two chunks will be:

$$\frac{B_1^e}{B_2^e} = \frac{L_1^{-d}(1-d)}{L_2^{-d}(1-d)} \frac{n_1}{n_2} = \frac{n_1}{2n_2} \tag{15}$$

This means that the relative activation of the younger chunk will be boosted two times compared to that of the older chunk. Generally, with the default decay parameter, if a chunk is r^2 times youger, its relative activation will be r times higher. This is a testable difference between ACT-R and Minerva2 and could be another source of bias, which Minerva2 does not account for.

The effect of context on activation according to ACT-R and Minerva2

For simplicity, lets assume that there are a total of m slots in the imaginal buffer's chunk and that $w_k = m$. Also, we will assume that all slots have the same fan. If for chunk i a total of k slots spread activation, the total amount of spreading activation is:

$$SA_i^e = \prod_j e^{S_{ji}} = \left(\frac{e^S}{fan}\right)^k \tag{16}$$

Thus, spreading activation is power function of the number of matching slots. According to Minerva2, activation changes as the cube of the number of matching values:

$$A = S^3 = \left(\frac{k}{m}\right)^3 \tag{17}$$

Qualitatively, both expressions are superlinear functions of k, albeit with different functional forms. However, note that there is no strong justification for the latter expression. Its purpose is to allow "those items in memory that are most similar to the test probe to dominate the overall echo from secondary memory, while preserving the sign" (Dougherty et al., 1999, p. 183). ACT-R's spreading activation also affords this and is in fact consistent with other exemplar models such as Context Theory (Medin & Schaffer, 1978), where similarity of two stimuli follows a multiplicative rule. Such parallels can also be drawn between ACT-R and the more recent model of recognition memory REM (Shiffrin & Steyvers, 1997). In REM, the likelihood ratio λ_j plays the role of activation. Under the assumption that the values of a feature in a memory image V_{kj} are all equal, the expression for λ_j reduces to:

$$\lambda_j = (1-c)^{n_{jq}} \prod_{k \in M} \frac{c + (1-c)g(1-g)^{V_{jk}}}{g(1-g)^{V_{jk}}} \tag{18}$$

$$\lambda_j = (1-c)^{n_{jq}} \left(\frac{c + (1-c)g(1-g)^{V}}{g(1-g)^{V}}\right)^m, \tag{19}$$

where n_{jq} is the number of nonzero features, M are the indices of non-zero features that match, and m is the number of features that match. Generally, exemplar and memory models seem to converge to the notion that the relevance of a item of memory increases in a multiplicative fashion with the number of matching features.

Discussion and Conclusion

At first sight, ACT-R's declarative memory theory and Minerva2 seem quite different: ACT-R uses the notion of a memory chunk with slots, while Minerva2 uses vectors with binary values. In addition, ACT-R explicitly models memory decay, while according to Minerva2 the only way for information to be not perfectly recorded in memory is through imperfect encoding. Finally, ACT-R envisions a single internal representation of an external object, whose memory strength increases upon each encounter of that object or recall of the internal representation. Minerva2 on the other side saves separate traces upon each encounter.

Given these differences, it is not obvious at which points these two theories will be tangential to each other. Our results show that both as a function of prior history and current context the theories can make similar predictions. More

specifically, according to both theories, in many experimental conditions we expect relative activation to be a function only of relative presentation frequency. As for context, it rapidly increases the activation of memory traces similar to it. The correspondence between the functional form of this increase according to ACT-R and some exemplar models suggests an exploration of that very funcitonal form as a viable alternative to Minerva2's current cubic function.

More importantly, the results presented above also indicate the plausiblity of accounting for judgment phenomena with ACT-R. Future efforts should aim at fully implementing HyGene into ACT-R and repeating the very simulations that MINERVA-DM and HyGene have already presented. Of especial interest would be an exploration of those conditions, in which ACT-R's predictions differ from those of Minerva2, such as when chunks have diffferent lifetimes or different fans.

We conclude with a wish for more frequent theory integration efforts in psychology, as we hope that the aforementioned results demonstrated.

Acknowledgments

We thank the Swiss National Science Foundation for generously funding Cvetomir Dimov through Project 146702.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., & Douglass, S. (2001). Tower of hanoi: evidence for the cost of goal retrieval. *Journal of experimental psychology: learning, memory, and cognition*, *27*(6), 1331.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: a unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1120.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396–408.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). Minerva-dm: A memory processes model for judgments of likelihood. *Psychological Review*, *106*(1), 180.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

Marewski, J. N., & Mehlhorn, K. (2011). Using the act-r architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, *6*(6), 439–519.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: an ecological model of strategy selection. *Psychological Review*, *118*(3), 393.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Mischel, W. (2009). The toothbrush problem. *Association for Psychological Science Observer*, *21*(11), 1.

Salvucci, D. D., & Anderson, J. R. (2001). Integrating analogical mapping and general problem solving: The path-mapping theory. *Cognitive Science*, *25*(1), 67–110.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological review*, *112*(3), 610.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Remretrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say broke? a model of learning the past tense without feedback. *Cognition*, *86*(2), 123–155.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological review*, *115*(1), 155.

Distributed Cognition in the Past Progressive: Narratives as Representational Tools for Clinical Reasoning

Katherine D. Lippa (lippa2@gmail.com)
Department of Psychology, 3640 Colonel Glenn Hwy
Dayton, OH 45435 USA

Valerie L. Shalin (Valerie.shalin@wright.edu)
Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis)
Department of Psychology, 3640 Colonel Glenn Hwy
Dayton, OH 45435 USA

Abstract

Cognition may require access to past events, for example to understand undesirable outcomes or diagnose failures. When cognition is distributed between multiple participants, a particular representational challenge occurs because not all of the participants may have directly experienced the focal event. Language can transcend temporal and physical limitations on event accessibility. We suggest that people create complex linguistic constructs as tools to facilitate retrospective cognition. We illustrate this process by analyzing the use of a particular linguistic construct (narrative) in the domain of clinical reasoning. Results demonstrated that narratives support clinical cognition during practitioner-patient interactions. Narratives extended access to clinically relevant events providing information about circumstances, subjective experiences, patient functioning, and prior decisions. Whereas, the hermeneutic nature of narrative allowed collaborative hypothesis testing and creation of meaning. The use of narrative in clinical cognition challenges Bruner's (1991) distinction between narrative and paradigmatic reasoning and enriches the understanding of medical narratives.

Keywords: distributed cognition; medical cognition; narrative; doctor-patient interaction

Introduction

Retrospective cognition is often required to understand adverse events. Whether people are attempting to understand intelligence failures after a terrorist attack, the origin and significance of a technical error, or the evolution of a disease, the reasoning task creates representational challenges. The onset of the focal event may not have been recognized in real time and significant aspects of the environment may have been missed or misinterpreted. Crucially, those involved in retrospective analysis may not have directly experienced the events in question. All of these properties make retrospective cognition dependent on representations of past experience suitable for exchange with reasoning partners. Some of these representations take the form of physical tools, such as record systems. But others are ephemeral linguistic compositions. In this paper, we analyze how one representational medium, medical narrative, is used as a psycholinguistic tool for overcoming the challenges of retrospective cognition distributed between patient and practitioner.

Language as a Cognitive Tool

Research has established the cognitive utility of language for both individual and distributed cognition. On an individual level, the use of words facilitates myriad cognitive processes including perceptual discrimination, memory, and attention (Yoshida & Smith, 2005; Fulkerson & Waxman, 2007). Moreover, the semiotic functions of language provide an organizational structure for people to effectively parse their environment (Bloom & Keil, 2001).

The semiotic connection that language creates between thought and the environment (Clark, 2006) is especially significant during distributed cognition. Under this paradigm, the unit of analysis for cognitive functions includes environmental artifacts and other participants (Hutchins, 1995.) When cognition is distributed between participants, language acts synergistically with cognition, physical context and culture to create a space for interaction (Cowley, 2011). Whether an actor is in the next room or on the other side of the planet, the distribution of cognition across space requires language to extend the interaction space (Tylen, et al., 2010). Similarly, when actors have access to different portions of the environment (e.g. they are working on different displays or have specialized expertise that allows them to understand particular aspects of the environment) language allows them to create mutually accessible representations of key environmental elements (Bødker & Anderson, 2005).

Participants naturally track aspects of the environment under discussion, creating a joint focus of attention (Eberhard, et al, 1995). Moreover, language allows actors to monitor each other's understanding and coordinate task performance (Bangerter & Clark, 2003). However, these basic utilities of linguistic communication regard its function as the assertion of propositions. Larger linguistic constructs, such as narrative, are complex, intentionally constructed representations. In contrast to the fixed tools of a task environment such as displays, constructed representations are emergent, and are both shaped by and shape thought processes (Chandrasekharan & Nersessian, 2015; Nersessian, 2009).

Narrative[1] Cognition

Several scholars link narrative and cognition (Bruner, 1991; Pennington & Hastie, 1992; Bower, Black & Turner, 1979). Notably, Bruner has suggested that narrative constitutes a unique form of cognition divorced from paradigmatic (i.e. logical, scientific) thought. Under Bruner's conception, narrative cognition has 10 core properties: diachronicity, genericness, canonicity and breach, normativeness, particularity, intentional state entailment, hermeneutic composability, context sensitivity, referentiality, and narrative accrual. Despite the fact that Bruner proposed a dichotomous separation between narrative and paradigmatic thought, many of these narrative attributes may facilitate conventional cognitive functions.

Narrative facilitates the organization of individual pieces of information into a coherent whole. On a temporal level, the creation of a narrative entails the (re)configuration of events into a unified structure that orders individual events within a coherent 'plot' (Ricouer, 1980). Narrative genres guide attention with a template for the inclusion, structuring and presentation of particular elements (Barthes, 1975). Organizational advantages explain the benefit of narrative on memory (Bower, Black & Turner, 1979).

Narratives offer more than structure for representing events. Narratives facilitate event reasoning. The reasoning involved in narrative composition includes causal argument. The temporal nature of narrative facilitates abductive reasoning. Some theorists advocate that all narratives are causal arguments employing *post hoc ergo propter hoc* logic (Robinson & Hawpe, 1986). These intermingled processes of reasoning and narrative composition result from a single individual or emerge as part of a social dynamic as two or more actors co-construct a narrative (Ochs, Smith and Taylor, 1989). The co-construction of narratives suggests its utility for distributed cognition.

Research also suggests that narratives may suit reasoning about complex environments. The combination of particularity and genre conventions allows narrative to generate theory by filtering noisy data in recognizable patterns. Dubba et al. (2012) have shown this effect in an AI system for modeling an airport through spatio-temporal narratives. Their work demonstrates the understanding of a dynamic domain by combining inductive and abductive processes through a narrative framework.

Narratives in Medicine

The conventional medical decision making research paradigm focuses on the reasoning of individual physicians with factual data. Yet, patients share narratives with practitioners by recounting acute episodes and elements of their medical history (Haidet & Paterniti, 2003). Both practitioners and patients use narratives to understand illness related events following a clinical encounter. (Hunter, 1991; Crossley, 2000). Nevertheless, there is little understanding of the role narrative plays in distributed physician-patient clinical cognition. In this study, we look at the cognitive significance of narratives for physicians and patients managing Multiple Sclerosis (MS).

Methods

We observed 24 patients with MS interacting with 3 medical practitioners (2 neurologists and 1 nurse practitioner) at a university neurology clinic. Clinical sessions were audio recorded (n=10) or documented with field notes (n=14). Follow-up interviews with all patients were audio recorded within 2 weeks of their appointment. All audio recordings were transcribed using a literary transcription method. Field notes were elaborated immediately after each clinical session and original notes were retained for comparison.

Narratives (defined as speech describing one or more particular events and including temporal, evaluative and contextual components) were identified in the transcripts and notes. More than a year later approximately 25% of the corpus was recoded to verify coding reliability, resulting in Kappa = .82 (CI .78-.86).

All narratives were examined in successive rounds of analysis. Initial analyses verified the elements of narrative as defined by Bruner (1991) in this context. Subsequent analysis focused on identifying patterns concerning the cognitive and pragmatic utility of these narratives in the clinical context. Once patterns of use were identified, the set of narratives was reanalyzed for how those patterns manifested and varied. We used an additional focused coding scheme to identify causal reasoning and the use of functional descriptions, establishing reliability with a second independent rater (Kappa=.63 [CI .41-.85]). All of the patterns described below appeared in both the notes and the transcripts. To allow for greater depth of discussion specific examples below are taken from transcripts.

Results

Eighty-eight narratives were identified including one or more from each clinical session and interview. A sample of these appears in Table 1. In the discussion below, we analyze and provide examples of how narrative served as a tool to facilitate a variety of cognitive functions. As representations of external events, narratives provided joint access to clinically relevant content including contextual information. They helped focus attention on specific issues and relevant parameters for understanding those issues. Finally, they facilitated causal reasoning, and the integration of intentional and metacognitive components into clinical cognition.

[1] There are many definitions for narrative (Bruner, 1991; Ricoeur, 1980; Labov, 1997). In this study, narrative is defined as a recounting of events with temporal, evaluative, and contextual elements.

Table 1. Narrative Exemplars

Example 1 (male, 40s, minimal disability, college educated) Patient: Um, two weeks ago I had a um, um I guess an episode is what you'd call it. Um where I had a visual problem. I had kind of a backwards c shape blurry spot, you know in my vision, and it lasted about ten minutes… Practitioner: Did you have optic neuritis before in that eye? {Uh}[2] Like did you lose vision before? Or have blurry vision for some time before in that eye? Patient: Um, I've had a what people have told me it's called a like floater in my eye for a long {uh huh} time. Since like 93, 94. I have like uh like uh I haven't experienced that kind of loss of field in my eye before. And as I sat there I thought you know is it the right or the left, so {ok} then I you know closed my right and checked and then closed my left and checked and it seemed like it was in both, so it didn't seem like it was in one or the other in particular. And so then I closed both eyes and you could still like when you sit down and stare at a light bulb you still got that kind of greenish. Well it was still there… Practitioner: How big was it? Patient: Um, it was fairly large. I mean um I guess it was kind of like a field in my vision and I guess if I was looking at my computer there were certain portions where I was looking where I wouldn't be able to see what I was typing, because it would be obscured by the blurriness. Practitioner: [Do you get migraines? Patient: No, I do have headaches but I've never experienced what people told me is a migraine… Patient's Wife: He had me drive home it scared him that much… Practitioner: But it lasted only half an hour right? Patient: I don't even think it lasted that long? I know on Friday it only lasted 10 minutes because when it started I thought 'well this is unusual' and I looked at my watch and I wrote down in my notebook… Practitioner: Cause usually like MS attack would last it has to last more than 24 hours right, so we wouldn't consider it MS attack. What you describe to me is not really what sound like {but} MS related problem.
Example 2 (female, late teens, minimal disability, high school educated) Practitioner: Ok. Difficulty walking? Difficulty with balance? Patient: there was an episode that I had when I lived with my grandma. She has carpet floors and I got up for school. This was when I was still in school. I got up for school that morning in my bedroom to get dressed. And when I got up I felt fine, but when I started to walk I fell and I hit the carpet with my head. Practitioner: ok, oh really, wow, so you had a fall. Do you get dizzy?
Example 3 (male, 40s, significantly disabled, college educated) Practitioner: Which medication helps you the best? Pain medication? Patient: Hydrocodone, the Methadone, the Methadone seemed to help a lot. But when I tried to get off the Methadone, which I did…It was, it was the worst thing that's ever happened to me. Practitioner: What happened? Patient: Um, my body went into the worst type of shock. Just withdrawal and shock. Uh,uh,uh. I never want to take anything that my body becomes that reliant on. {yeah} I mean it hurt. It hurt so bad I would lay in home in bed. The sweats were so bad and I was freezing cold. I had cloths on, a down comforter on, I had a heater blowing on my body, and I could not get warm and I am just shaking so bad. I was seeing things, got a bad headache. When I go go to the restroom it was terrible. Didn't eat, dropped probably about 15 pounds in a week because I didn't eat any food.

Narrative Cognition and Representation

Medical decision making research paradigms notwithstanding, real clinical reasoning depends upon a great deal of information that the practitioner cannot directly access either because it is phenomenological or concerns events outside the clinic. Practitioners must know about the qualitative nature of symptoms and the context, sequence and timing of episodes for diagnostic purposes (Lippa & Shalin, 2015). The patient must bring these subjective and environmentally distributed elements into the clinic (Lippa, et al., submitted). The participants in this data used narratives as a tool for bringing external events into the clinic. This mutually accessible representation of events, helped guide joint attention and provided anchors for understanding subjective symptom reports.

Contextualization Narratives often began with a broad sketch of the environment, including both temporal and other contextual elements. Many of the narratives incorporated a reference to temporal context that was either calendar based, as in example 1 "two weeks ago" (43% of the corpus) or tied to key events in the participant's life, as in example 2 "when I lived with my grandma". One patient even used events from the news as a temporal referent. The specification of temporal attributes converges with a physician's interest in the recency of an event as an indicator of meaning. Older narratives indicate the

[2] {} denote minor interjections.

persistence of experience of increased import, while more recent narratives may reflect fortuitous and transient events of limited significance to disease. The provided contextual overviews varied a great deal including information about the patient's case history, personal background, emotional state, physical location, etc.

Event Classification Having situated events in context, narratives typically focused on a particular problem. This problem became the plot of the narrative to which all subsequent particulars and events needed to relate. Sometimes the patient began the narrative with a clear understanding that the events described constituted a problem and the function of the narrative was to draw attention to the focal problem and provide information to facilitate distributed problem solving. But in many cases, this information helped classify whether and to what extent the central events constituted a MS related problem. Example 1 includes both functions. The patient immediately concluded the symptom was disruptive and notable. This is clear from the fact that he wrote down the exact time of onset. We believe that he told his story hoping the practitioner would clarify the parameters of the problem and their relationship to MS. When she ended the discussion without a clear explanation, he was confused. He expressed this confusion during the follow up interview: "*I couldn't really understand what she's saying about it but what I got from her was that she didn't really think it was MS related so I don't know what to think now.*"

Inclusion of Detail In MS management narrative genres appear to be co-constructed by patients and practitioners (for a similar effect in psychotherapeutic settings see Sluzki, 1992). As patients and physicians interact, the physicians' questions help patients to define what aspects of their experience are relevant to MS management. We see this in example 1 when the practitioner asks first about the specifics of the symptom ("*How big was it?*") and later about the time course ("*But it lasted only half an hour right?)*. The patient could have chosen to respond to each question with a brief statement of facts, but instead he responded with extensive descriptions –mini-narratives by themselves. The practitioner's questions showed which aspects of the narrative she was attending to and directed the patient's attention to elements worth elaborating. This type of interaction signals to patients the kinds of information to include in clinical narratives. Over time they learn to tell stories that incorporate the specific information that physicians need (for a discussion of this process see Lippa & Shalin, 2015).

Functional Anchoring In addition to the basic leveraging of narrative structure described above, patients used narratives about functioning to anchor otherwise amorphous symptom descriptions. In most of the narratives in this study, including examples 1, 2 and 3, at some point patients directly reported their experiences (e.g. *I had kind of a backwards c shape blurry spot, you know in my vision; I felt fine; it hurt so bad)*. Such descriptions were often insufficient, employing ambiguous terms. A mutually understood anchor conveyed a more specific meaning.

To meet this challenge, patients often included supportive functional descriptions for subjective reports, or emphasized clinically relevant information in their narratives. Twenty nine functional descriptions appeared in 26 distinct narratives (30% of the corpus). In example 1, the patient addresses one of the practitioner's questions about the characteristics of the symptom by providing details concerning how the symptom affected his living activities, in this case, work at the computer. Other functional descriptions convey a sense of magnitude. In example 3, the patient emphasizes the severity of his pain by saying he couldn't function normally but had to "*lay in home in bed*" and even to "*go to the restroom was terrible*." Functional descriptions related symptoms to tasks with culturally grounded demands to impart meaning to the severity of the symptom.

Intentionality and Metacognition By design, medical records document physical information. They track disease processes (and support insurance reimbursement) rather than intentional information driving medical decisions. Narratives complement this content with information about phenomenological experiences, context and motivation that are not typically included in medical records.

In fact, practitioners asked patients to retell parts of their case history to recapture details or intentional elements contributed by the practitioner and/or patient during earlier interactions. Practitioners either asked patients to elaborate on brief descriptions from their records ("*What happened [when a particular medicine was discontinued]*?") or to recount why decisions noted in the record were made ("*I'll go check something {looks at computer} Did we stop the Ultram because it wasn't doing the trick anymore? And you liked the Tylenol?*). These questions suggest that the record noted that events occurred but not how or why. The practitioner had to solicit the patient's narrative recall to recapture significant particulars. Many of these cases involved recapturing the same kind of external and phenomenological content that seem to be a key affordance of narratives throughout the clinical encounter. Example 3 centers on this type of elaboration. This interchange was part of a long discussion of pain management. In this case the practitioner had access to a record of which medications had been given when but there was little information on why specific drugs were selected or discontinued. The example illustrates both the absence of such content in conventional medical records and the relevance of narrative content to link physical conditions, phenomenological experiences and decision processes.

Reasoning through Narrative

Narrative is a representational product, but it is also a hermeneutic process. Both the act of creating a narrative and the act of listening involve interpretation of the events depicted. Participants used the hermeneutic quality of narratives to facilitate collaborative reasoning.

At the most basic level, patients used narrative to present events of uncertain significance so that the physician could assign an appropriate interpretation. Example 2 follows this pattern. The physician asks a question. Rather than answering directly the patient describes an event of uncertain significance allowing the physician to interpret its significance.

A hermeneutic process also facilitates the hypothesis testing that is often considered a key part of diagnostic reasoning. Sometimes the practitioner would present one or more possible diagnoses during the course of the narrative and ask the patient questions to see if information either from the narrative or from past experience supported her hypothesis. In example 1, the practitioner proposes two possible diagnoses during the narrative (i.e. optic neuritis and migraines) and for each diagnosis asks the patient questions that might provide information to support her hypothesis. In cases like this one, the practitioner's questions and the patient's responses allowed them to test possible assignations of meaning to key events while co-constructing a narrative. Ideally, this led to a mutually understood and accepted interpretation of events. However, sometimes participants maintained discordant interpretations even as the narrative concluded. In example 1, the patient did not support the practitioner's hypothesis and by the end of the interaction began to object to her interpretation interjecting a 'but' into her summary.

Conclusions

The results of this study suggest that people use language to create artifacts, in this case, representations of events, that simplify cognition. These patients and practitioners faced a variety of cognitively demanding tasks that required reasoning about past events. In carrying out these tasks, they composed narratives whose composition and contents addressed a variety of cognitive challenges.

Narratives as Psycholinguistic Tools

Like physical tools, language extends our functional capacities. As tools become more complex, they facilitate the creation of more specialized tools. Thus if basic language is like an abacus, complex language is like a computer. It is a tool in itself but it also facilitates the creation of more specialized, context specific tools.

In this study, narratives served as tools to address the demands of ongoing clinical cognition. Participants constructed narratives as informational resources to represent external and/or prior events, support metacognition, coordinate care between medical practitioners, and provide detailed, functional accounts of symptom episodes. In addition, the interpretive aspects of narrative facilitated hypothesis testing and causal reasoning. This suggests physicians should view narratives not just as a socio-emotional concern but also as a cognitive resource that can be used (and recorded) more or less efficiently.

Whereas this study examined a single environment, transient, localized experience is common to many domains, suggesting that this psycholinguistic representational tool may support distributed cognition in other contexts.

Narrative Cognition

Narrative cognition has typically been discussed as divorced from other forms of cognition, especially cognition involved in problem solving, logical reasoning and hypothesis testing (Bruner, 1991). However, the narratives in this study were integrated ongoing cognitive processes including analytic reasoning. In fact, many of the properties of narrative that Bruner presented as definitional and distinct from paradigmatic reasoning served to facilitate classically defined cognitive processes, (see Table 2).

This suggests that paradigmatic and narrative cognition should not be considered dichotomous or opposing. But rather that they can, and are, productively intermingled. In

Table 2. Cognitive and Clinical Functions of Narrative Properties

Property	Description	Cognitive Functions	Use in Clinical Cognition
Narrative Diachronicity	Inherent temporality	Memory & Reasoning	Maintain understanding of case history
Genericness	Construction in relation to stylistic and plot conventions	Attention & Memory	Combination of episode based and case history narratives to recall clinically relevant information
Canonicity and Breach	Focus on a disruption in the normal flow of events	Problem Solving	Problem definition
Normativeness	Construction in relation to implicit norms	Problem solving	Problem definition
Particularity	Inclusion of highly specific details	Reasoning	Creation of functional representations & representation of relevant elements outside the clinical encounter
Intentional State Entailment	Incorporation of phenomenology and psychology of characters	Meta-cognition	Recording reasoning processes not included in official records
Hermeneutic Composability	Both creating and comprehension are interpretive activities	Hypothesis Testing	Co-construction of narratives and hypothesis testing

this study, the properties of narrative served critical functions in distributed practitioner-patient cognition by providing a vehicle for patients to create representations of key elements of the clinical problem space that would not otherwise be accessible to practitioners. In this data set, narrative was not opposed to paradigmatic reasoning but rather was a tool that facilitated clinical reasoning, including causal reasoning and hypothesis testing.

Acknowledgments

This work was funded by DGE-0504438, 2005-14 (J. Flach, PI). *IGERT: Technology Based learning with Disabilities.*

References

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science, 27*, 195-225;

Barthes, R., (1975). An introduction to the structural analysis of narrative. *New literary history*, 237-272.

Bloom, P., & Keil, F. C. (2001). Thinking through language. *Mind & Language, 16(4),* 351-367.

Bødker S., Anderson P.B. (2005). Complex mediation. *Human-Computer Interaction, 20*, 353–402.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11(2*), 177-220.

Bruner, J. (1991). The narrative construction of reality. *Critical Inquiry, 18(1)*, 1-21.

Chandrasekharan, S. & Nersessian, N.J. (2015). Building cognition: The construction of computational representations for scientific discovery. *Cognitive Science* 39, 1727-1763.

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences, 10(8)*, 370-374.

Cowley, S.J. (2011). Taking a language stance. *Ecological Psychology, 23*, 1–25.

Crossley, M. L. (2000). Narrative psychology, trauma and the study of self/identity. *Theory & Psychology, 10(4),* 527-546.

Dubba, K., Bhatt, M., Dylla, F., Hogg, D. C., & Cohn, A. G. (2012). Interleaved inductive-abductive reasoning for learning complex event models. In *Inductive Logic Programming* (pp. 113-129). Springer Berlin Heidelberg.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409–436.

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition, 105(1),* 218-228.

Haidet, P. & Paterniti, D.A. (2003). "Building" a history rather than "taking" one: A perspective on information sharing during the medical interview. *Archives of Internal Medicine, 163,* 1134-1140.

Hunter, K.M. (1991). Doctors' stories: The narrative structure of medical knowledge. Princeton, NJ: Princeton University Press.

Hutchins, E. (1995). How a cockpit remembers its speed. *Cognitive Science, 19*, 265-288.

Labov, W. (1997). Some further steps in narrative analysis. *Journal of Narrative and Life History, 7*, 395-415.

Lippa, K.D. & Shalin, V.L. (2015). Stepping up to the blackboard: Distributed cognition in doctor-patient interactions. *Proceedings of the Cognitive Science Society.*

Lippa, K.D., Feufel, M.A., Robinson, F.E. & Shalin, V.L. (submitted). Navigating the decision space: A study of medical decision making as distributed cognition

Nersessian, N.J. (2009). How do engineering scientists think? Model-based simulation in biomedical engineering research laboratories. *Topics in Cognitive Science, 1(4),* 730-757.

Ochs, E., Smith, R., & Taylor, C. (1989). Detective stories at dinnertime: Problem-solving through co-narration. *Cultural Dynamics, 2(2),* 238-257.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of personality and social psychology, 62(2),* 189.

Ricoeur, P. (1980). Narrative time. *Critical Inquiry, 7(1),* 169-190.

Robinson, J.A. & Hawpe, L. (1986). Narrative thinking as a heuristic process. In T.R. Sarbin (Ed.) Narrative Psychology: The storied nature of human conduct. (pp. 111-125). New York: Praeger.

Sluzki, C. E. (1992). Transformations: A blueprint for narrative changes in therapy. Family process, 31(3), 217-230.

Ty'len, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C.D. (2010). *Mind & Language, 25 (1),* 3–29.

Yoshida, H., & Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science, 16(2),* 90-95.

Predicting Overprecision in Range Estimation

Matthew Kaesler[1] (matt.kaesler@gmail.com)

Matthew B. Welsh[1] (matthew.welsh@adelaide.edu.au)

Carolyn Semmler[1] (carolyn.semmler@adelaide.edu.au)
1. The University of Adelaide, SA 5005, Australia

Abstract

Overprecision (overconfidence in interval estimation) is a bias with clear implications for economic outcomes in industries reliant on forecasting possible ranges for future prices and unknown states of nature - such as mineral and petroleum exploration. Prior research has shown the ranges people provide are too narrow given the knowledge they have – that is, they underestimate uncertainty and are overconfident in their knowledge. The underlying causes of this bias are, however, still unclear and individual differences research has shed little light on traits predictive of susceptibility. Taking this as a starting point, this paper directly contrasts the Naïve Sampling Model and Informativeness-Accuracy Tradeoff accounts of overprecision – seeing which better predicts performance in an interval estimation task. This was achieved by identifying traits associated with these theories – Short Term Memory and Need for Cognitive Closure, respectively. Analyses indicate that NFCC but not STM predicts interval width and thus, potentially, impacts overprecision.

Keywords: confidence; overprecision; need for cognitive closure; STM; informativeness; naïve sampling model.

Introduction

Overprecision, the form overconfidence observed on interval estimation tasks, has been described as the most robust yet least understood form of overconfidence (Moore & Healey, 2008). It occurs where people provide confidence intervals (lower and upper bounds between which they are confident, to a stated degree, that an unknown value lies). If, over a set of questions asking for (e.g.) 90% confidence intervals, objective accuracy levels are lower than 90%, this is deemed overprecision (Alpert & Raffia, 1982; Moore & Healey, 2008). It differs from overestimation, observed on point estimates when post-item confidence judgements exceed accuracy, as point estimate accuracy improves with task experience, while range estimate accuracy does not, implying different underlying processes (Hansson, Juslin & Winman, 2008). People show pronounced overprecision on interval estimation tasks - that is, far fewer ranges contain the true value than expected based on the stated level of confidence (Lichtenstein, Fischhoff & Phillips, 1982). Overprecision is of applied significance in engineering, mining and the oil and gas industry, which all make use of estimates delivered in this form (Moore, Tenney & Haran, 2015), with valuation errors of hundreds of millions of dollars resulting (Welsh, Begg & Bratvold, 2007).

The basic cause of overprecision is people generating intervals too narrow to reflect their degree of subjective uncertainty, failing to capture the true value as often as expected (Alpert & Raffia, 1982). The effect is resistant to debiasing, with participants exhorted to widen their intervals failing to do so enough to achieve good, much less perfect, calibration (e.g. Yaniv & Foster, 1995).

Factors contributing to overprecision, however, are poorly understood and, while task characteristics and debiasing techniques have attracted significant interest, individual differences in performance have received little attention, with only two recent studies relating individual differences to interval estimation performance (Haran, Ritov & Mellers, 2013; Hilton, Regner, Cabantous, Charalambides and Vautier, 2011). This study addresses this deficit by selecting individual differences relating to two promising theories of overprecision, the Naïve Sampling Model (NSM) and the Informativeness-Accuracy Tradeoff (IAT; Juslin, Winman & Hansson, 2007; Yaniv and Foster, 1995; 1997).

Naïve Sampling Model (NSM)

The NSM explains overprecision cognitively - as a consequence of short term memory (STM) capacity. The underlying concept being that, to create an interval, a person calls relevant examples from long term memory (LTM). For instance, if asked to set an interval around the population of Nigeria, one might call to mind populations of other African countries (or, failing that, non-African countries).

This sample, held in STM, is then used as the basis for creating the interval estimate – for example, by taking the 10th and 90th percentiles of that sample and using these as the low and high ends of an 80% confidence interval. The sample drawn from LTM, however, is limited by a person's STM to a small number of instances and, as sampling dispersion is a biased estimator of population dispersion, 80% coverage of the sample does not correspond to 80% coverage of the population – leading to too narrow ranges.

Individual Differences in NSM. NSM holds that better STM decreases overprecision due to increased interval width - as larger samples are less likely to underestimate population dispersion. Support for a reduction of overprecision due to STM was found in a study involving a learning and then a testing phase - but interval width was not explicitly measured (Hansson et al., 2008).

Informativeness-Accuracy Tradeoff (IAT)

The IAT is a motivational explanation of people's preferences when receiving and generating interval estimates (Yaniv & Foster, 1995; 1997). In interval

estimation terms, narrow intervals are more informative but, holding all else equal, less likely to be accurate, while wide intervals are less informative and more accurate. Thus, these objectives need to be traded off against one another.

Yaniv and Foster's (1995; 1997) participants showed a general preference for receiving informative (narrow) interval estimates that missed the true value rather than uninformative (broad) estimates containing it, stating that these were more 'useful' than the wider estimates.

Participants also produced intervals containing the true value far less often than the stated 95% confidence level, and when informed of the degree to which their intervals would need expanding to contain the true values often enough, opined that this would render the judgements useless to the receiver (Yaniv & Foster, 1995; 1997).

The regulation of informativeness-accuracy thus follows conversational norms (Yaniv & Foster, 1995; 1997) - in providing a judgement, estimators are attempting to help the receiver. The receiver's purpose in soliciting a judgement should guide the relative informativeness or accuracy of the estimator (Yaniv & Foster, 1995; 1997).

A key contribution of this theory is the decomposition of intervals into width (upper minus lower bound) and absolute error (distance from the interval's midpoint to the true value), with width thought to reflect strategy (how informative or precise an estimator is claiming to be) while absolute error reflects knowledge (Yaniv & Foster, 1997).

Individual Differences in IAT. The conversational norms explanation for regulation of informativeness-accuracy in interval estimation has received limited support, with attempts to manipulate receiver purpose not affecting overprecision. Nor is it clear why communications between experimenter and participant should induce informative rather than accurate responses (Haran, Radzevick & Moore, 2010, cited in Moore, Tenney and Haran, 2015).

A novel explanation is that people may be predisposed to informative or accurate judgments by innate thinking dispositions such as Need for Cognitive Closure (NFCC) - defined as the desire to answer a question rather than sustaining further uncertainty (Webster & Kruglanski, 1994). Wide intervals are necessarily ambiguous, and it may be that high-NFCC participants would produce narrow intervals to avoid ambiguity and attain a feeling of closure.

Additionally, low NFCC is qualitatively similar to high Actively Open-minded Thinking (AOT), which predicted reduced overprecision on a single interval estimate (Haran et al., 2013). NFCC is preferable, however, as it includes discomfort with ambiguity, which AOT does not, and the NFCC scales are better validated than those for the AOT.

Online Confidence

Given uncertainty regarding the causes of overprecision, online confidence (OC) – the average post-item confidence ratings from a 12-item form of the Raven's APM test (Arthur & Day, 1994) - was also used as a predictor in this study. OC is thought to reflect a stable confidence trait across domains, shown to be predictive of performance within a domain - such as, online confidence from an earlier English test predicting end-of-year English grades (Stankov, Lee, Luo & Hogan, 2012) - but not yet across domains.

Hypotheses

1. Better STM will result in wider intervals and reduce overprecision. (NSM hypothesis)
2.Higher NFCC will result in narrowed intervals and increase overprecision (IAT hypothesis).
3. Higher online confidence will predict reduced overprecision (confidence-overprecision hypothesis).

Method

Participants

Participants (n = 49, 29 females, mean age = 31.0 years, SD = 15.5) were drawn from University of Adelaide students and the general population. They were highly educated, with 87% of participants having attempted a Bachelor's degree. Psychology 1B students participated for course credit and others entered in a draw to win one of two $50 gift cards.

Materials

Digit Span (STM capacity) Forward digit span, adapted from the Wechsler Adult Intelligence Scales (Wechlser, 2008), measured STM capacity. To ease data collection, the task was administered in groups. Participants were read lists of numbers as per the standard digit span procedure but provided written rather than verbal responses. The two response methods have been shown to produce similar scores within subjects (Ryan, Townsend & Kreiner, 2014).

There were two trials at each span level, 16 trials in all, ranging from two digits to nine digits in length. Each trial number was read out, e.g. "Trial 1", followed by a four second break, after which the digits were read out at a rate of one per second. Once all participants had finished writing their response, the next trial was announced and the process repeated until all 16 trials had been administered. A participant's score was the last span level at which they were correct on at least one of the two trials.

Need for Cognitive Closure Scale 15-item. A fifteen-item scale derived from the revised 41-item NFCC scale (Roets & Van Hiel, 2007) was used to generate the general NFCC factor (Roets & Van Hiel, 2011). Participants rated statements such as "I don't like situations that are uncertain" on a six-point Likert scale from 1, strongly disagree, to 6, strongly agree. Scores summed from all items give scores from 15 to 90 with higher scores indicating greater NFCC.

Raven's Advanced Progressive Matrices 12-item. Raven's is a 36-item measure of fluid intelligence for use with highly-educated samples but was used here to derive OC. Arthur and Day's (1994) 12-item version was used, untimed, and a participant's score was simply their number of correct answers. After each item, participants provided

confidence ratings (0-100%) for the accuracy of their answer, which were averaged as the measure of OC.

General Knowledge Interval Estimates (Overprecision). Twenty general knowledge questions were used to assess overprecision in interval estimation. Participants made 80% confidence estimates in accordance with the oil and gas industry standard (Welsh, Bratvold & Begg 2005), and saw an example question to learn the required response format.

Four topic areas were used; geography, sport, big business and historical events, with five questions from each topic. In theory, question content is unimportant as optimum performance involves scaling intervals to accord with the precision of knowledge, not knowing exact answers to the questions. However, participants often react with frustration when they find questions too difficult or topics unfamiliar (Welsh et al., 2005) and, thus, a range of topics was used.

Bias Score. Bias score was used as the measure of overall overprecision, calculated as the given confidence level (i.e., 0.8) minus the proportion correct. Perfect calibration would result in a bias score of 0, overprecision a positive bias score and underprecision a negative bias score.

Interval Width. Interval width was measured by subtracting an estimate's upper from its lower bound. As width is relative to the scale of the correct answer, participant's ranges were ranked from widest to narrowest (i.e. widest = 1) for each item, using average ranks for ties. Participant's mean rank across the 20 questions was used in analyses. Scores closer to 1 thus indicate participant's intervals that were, typically, wider than other participants.

Absolute Error. This was calculated as the absolute difference between the true value and the interval midpoint. Absolute error was calculated as described for interval width but values were ranked smallest to largest (i.e. least absolute error = 1), ensuring higher interval width and absolute errors both correspond to greater overprecision.

Procedure

Participants first completed the digit span test and gave demographic information before undertaking the interval estimation task. They were told to remind themselves when reading each estimation question that they should be 80% sure that the true value was between the low and high bounds they provided. After the interval estimates, participants completed the NFCC scale and then the Raven's APM. All tests were conducted using pencil and paper and completion time was 30-45 minutes.

Results

Given the small sample, Efron's (1987) BCa bootstrap procedure with 2000 resamples was used to calculate statistics in addition to traditional methods.

Descriptive Statistics

Table 1 shows descriptive statistics for all continuous variables, revealing that participants displayed the expected pattern of overprecision in their interval estimates, with a mean bias score of .51. The degree of miscalibration is on the high end of that observed in other studies, with hit rates around 30% at a pre-stated confidence level of 80%.

Table 1: Descriptive statistics for all continuous variables.

	M	*SD*	*Min.*	*Max.*
Age	31.04	15.45	18	80
Digit Span	6.92	1.35	4	9
NFCC	57.06	10.80	38	82
Online Confidence	67.68	18.34	17.5	100
Raven's APM	7.88	2.76	2	12
Bias Score	.51	.17	.00	.80
Interval Width	25.00	6.64	11.33	39.55
Absolute Error	25.00	5.23	14.23	42.70

Table 2 present a correlation matrix for these variables - excepting between bias score, interval width and absolute error, due to their statistical dependency.

Table 2: Correlation matrix for all continuous variables

		1	2	3	4	5
1	Age	-				
2	DS	-.10	-			
3	NFCC	.04	-.20	-		
4	OC	**-.28**	.13	-.24	-	
5	RAPM	**-.36**	.20	**-.34**	**.67**	-
6	Bias Score	.09	-.06	**.32**	-.10	-.27
7	Int. Width	**.30**	-.27	**.40**	-.24	**-.38**
8	Abs. Err.	-.25	.13	.06	.12	-.01

Bold = $p \leq .05$. DS = digit span; NFCC = need for cognitive closure; OC = online confidence; RAPM = Ravens APM.

Looking at Table 2, one can see three major points relating to our hypotheses. The first is that online confidence is closely related to the people's scores on the Ravens (as would be expected) but not significantly related to any of the three overprecision measures (bias, width or error). The second is that digit span seems similarly unrelated – although the relationship between digit span and interval width, at -0.27, approaches significance. Finally, NFCC is significantly correlated with the bias score and more strongly with interval width but not absolute error.

Beyond our hypotheses, it is worth noting the significant negative relationship between participants Ravens scores and interval widths – due, perhaps, to the fact that age negatively predicts people's online confidence and Raven's scores and is positively related to interval width.

The above examination of correlations suggests that NFCC predicts overprecision bias and interval width whereas digit span and online confidence do not. This piecemeal approach, however, ignores potential relationships between the predictors. Therefore, multiple regressions were conducted as a means of examining all three potential predictors of overprecision simultaneously.

Interval Width

Table 3 displays the results of a multiple linear regression

conducted to test the hypotheses that: STM capacity, as measured by digit span, would be related to the production of wider intervals; that NFCC level would be related to the production of narrower intervals; and that online confidence would be related to the production of wider intervals.

The model reached significance, $F(3,45) = 4.17$, $p < .05$, $R^2 = .22$, with NFCC as the only significant predictor. The hypothesis that higher NFCC score would be related to the production of narrower intervals was supported, however the wide bootstrap CI^{95} for the regression coefficient makes the true strength of the relationship difficult to ascertain.

Table 3: Multiple linear regression analysis predicting interval width from digit span, NFCC and online confidence

	B	BCa CI^{95}	*SE*	β	*t*	*p*
DS	-.95	[-2.30, .30]	.72	-.19	-1.43	ns
NFCC	.20	[.07, .32]	.07	.33*	2.39	<.05
OC	-.05	[-.16, .09]	.06	-.13	-.99	ns

In contrast, the hypothesis that high STM would result in wider intervals was not supported. The (n.s.) relationship observed was, in fact, opposite to that predicted. Likewise, online confidence failed to predict interval width.

Overprecision

The above analyses support the supposition that NFCC predicts interval width but the degree to which this results in overprecision bias is, perhaps, more important to know.

Table 4 displays results of a multiple linear regression conducted to test the hypotheses that: STM capacity (i.e. digit span) would be related to reduced overprecision; that NFCC would be related to increased overprecision; and that online confidence would predict reduced overprecision.

The model did not reach significance, $F(3,45) = 1.75$, $p = .17$, $R^2 = .10$. Despite the significant correlation between NFCC and bias score in Table 2, NFCC was not a significant predictor of overprecision in this model - and none of the three hypotheses found support.

Table 4: Multiple linear regression analysis predicting bias score from digit span, NFCC, online confidence and Ravens

	B	BCa CI^{95}	*SE*	β	*t*	*p*
DS	.001	[-.04 .04]	.021	.01	.04	ns
NFCC	.005	[.001, .009]	.002	.32	2.16	ns
OC	.000	[-.003, .002]	.001	-.02	-.14	ns

Discussion

These results go some way in helping to decide between two promising theories of overprecision: the Naïve Sampling Model (NSM, Juslin et al., 2007); and the Informativeness-Accuracy Trade-off (IAT, Yaniv & Foster, 1995; 1997) with some support offered for the latter - at least, to the extent that one accepts that NFCC reflects an intrinsic tendency to prefer informativeness over accuracy.

A third possible predictor of overprecision – online confidence (Stankov et al., 2012) – showed no clear relationship with overprecision, suggesting that these forms of confidence and overconfidence are not closely related.

STM Capacity

The idea that superior STM causes the production of wider intervals and thus reduces overprecision is central to the Naïve Sampling Model (NSM: Juslin et al., 2007). Thus, this study showing no effect of STM on interval width and failing to replicate Hansson et al.'s (2008) finding that STM predicted reduced overprecision presents a challenge to the theory as described (although it does not rule out the possibility of other sampling processes providing a sound explanation of the interval estimation process).

These results appear not to be STM measurement issues, as the digit span test used here produced the expected 'seven plus-or-minus two' pattern of results (Miller, 1956). There may, though, be other explanations for why this study failed to support Hansson et al.'s (2008) assumptions and findings.

Firstly, forward digit span, as used in this study, measures only STM. The digit span task used in Hansson et al. (2008), however, is described as a composite (via an unstated formula) of a passive repeat-back task (presumably forward span) and a sequencing task, which are thought to reflect working memory in addition to capacity (Engle, Tuholski, Laughlin & Conway; 1999). Thus, Hansson et al.'s (2008) scale may have measured working memory and STM, affecting the relationship they observed. The positive correlation Hansson et al. (2008) observed between digit span and Raven's APM was stronger than seen herein, consistent with the notion that the more complex span task may have led to a stronger relationship with working memory, which is more closely related to Gf – as measured by Raven's APM (Wiley, Jarosz, Cushen & Colflesh, 2011).

Secondly, the general-knowledge interval estimation used herein differed significantly from the laboratory learning task used by Hansson et al. (2008) in that it did not control for prior knowledge. Hansson et al.'s (2008) task involved a learning phase of fictitious company data, followed by an estimates phase wherein participants made a point estimate, thought to reflect information successfully stored in LTM if correct, and an interval estimate at the 80% probability level, thought to reflect inference from STM if correct.

Their results suggested that STM capacity as assessed by digit span was negatively related to overprecision but proportion of correctly recalled values from LTM was not. This was interpreted as evidence for the importance of STM capacity as compared to information stored in LTM in generating correct interval estimates (Hansson et al., 2008). However, the strongest predictor of overprecision was the variance of the values correctly recalled from LTM, which could be assessed by comparing the values recalled to the distribution of values from the learning phase. Those in the low-variance group after a median split were almost twice

as overprecise as those in the high-variance group, pointing to the possibility of a role for information stored in LTM.

Controlling for prior knowledge is necessary to demonstrate the theorised dissociation between STM capacity and LTM information storage but such a task is not representative of real world estimation tasks.

Of interest - in light of this study's results - is the finding that low-variance LTM representations were associated with increased overprecision (Hansson et al., 2008). It is possible that when prior knowledge is not controlled, as in general knowledge questions as used in this study, between-subjects differences in LTM representations may obscure the relationship between STM capacity and overprecision. However, were STM's predictive power to be realized only in laboratory settings where prior knowledge is controlled, the utility of NSM for applied settings must be questioned.

Need for Cognitive Closure

The hypothesis that higher NFCC would correlate with production of narrower intervals was supported but the effect of this on a person's level of overprecision was mixed – supported in the correlation table but not in the multiple regression. This is, perhaps, unsurprising as overall bias is also affected by differences in individual knowledge and any effect on interval width is diluted by the inclusion of errors in estimation, which NFCC does not predict.

The results offer support for interpreting the NFCC trait as per Webster and Kruglanski (1994), which led to the presumption that it might be linked to a preference for informativeness over accuracy. They also support a qualitative similarity between low NFCC and high AOT, as low-NFCC participants in this study behaved like high-AOT participants in Haran et al. (2013). NFCC could be considered a complimentary construct to AOT given its significant relationship with interval width, which was not predicted by AOT in Haran et al. (2013).

Overall, the fact that a dispositional variable (NFCC) was the best predictor of both interval width and, despite the model not reaching statistical significant, overprecision, argues for an account of overprecision including motivational and strategic aspects rather than a purely cognitive one. It is, thus, broadly supportive of the principles of the Informativeness-Accuracy Tradeoff's overprecision explanation (Yaniv & Foster, 1995; 1997). The theoretical link drawn here between NFCC and IAT thus seems to have been sensible; although the mechanism underpinning this requires elucidation, these findings should stimulate further research into NFCC and, more generally, the role of intrinsic motivators as drivers of the IAT.

Future Directions

Few studies have examined overprecision on interval estimates from an individual differences perspective and, so, these results suggest multiple directions for future research. A first, drawing on results in Table 2, will be to nail down the relationships between age and the various confidence, intelligence and bias measures examined herein.

Naïve Sampling Model This study did not support NSM's ability to explain overprecision – at least in cases where estimates are made in contexts without controlled prior knowledge. Follow up research into the effects of task features on NSM is therefore necessary to clarify its utility. A future study could examine predictions of NSM in a within-subjects design, with conditions having different levels of control over prior knowledge, shedding light on the utility of NSM outside the laboratory. Additional work utilizing separate measures of working memory and STM could also help disentangle possible confounds, while investigating the ability of other sampling processes to explain overprecision would help clarify the worth of further developing sampling-based models of overprecision.

Need for Cognitive Closure As a new construct to the interval estimation field, many aspects of NFCC bear further investigation. Understanding the mechanisms underlying the association between high NFCC, the production of narrower intervals, and a potential relationship with increased overprecision is important. Thus, use of the 41-item revised scale - which includes subscales - (Roets & Van Hiel, 2007) in future research would be a sensible starting point.

Also relevant are moderators that might affect how those high or low in NFCC behave. The link between NFCC and interval estimates posited here was that wide intervals are ambiguous and high-NFCC participants would avoid ambiguity - providing narrower intervals. However, an alternative explanation could be drawn from the literature on NFCC and information search. Specifically, rather than trying to reduce affective discomfort caused by ambiguity, participants may simply not search for enough information to make credible estimates, as the perceived cost of information search is too high. Research on NFCC and information search suggests those high in NFCC search for less information and make faster decisions but the opposite pattern of behaviour can manifest under certain conditions (Choi, Koo, Choi & Auh, 2008; Van Hiel & Mervielde, 2002). Variables moderating the association between NFCC and information search should, therefore, be investigated.

For example, those low in ability to achieve closure, perceived ability to enact strategies to fulfil epistemic needs, and those with a low working memory capacity (WMC) may behave conversely to that suggested by their NFCC level, with high NFCC people prolonging information search and suspending closure and low NFCC people shortening it to achieve rapid closure (Czernatowicz-Kukuczka, Jasko & Kossowksa, 2014; Kossowska & Bar-Tal, 2013). The apparent relevance of WMC to the NSM and to the relationship between NFCC and information search may, thus, provide a path to integrating motivational and cognitive theories of overprecision.

The effect of moderators on the relationship between NFCC and information search paints a complex picture. A more nuanced view of the relationship between NFCC, interval width and overprecision may appear when moderators of NFCC behaviour are also measured.

Conclusion

The key finding of this paper is that NFCC, introduced as a possible measure of a person's preference for informativeness over accuracy in line with Yaniv and Foster's (1995;1997) IAT theory does predict overprecision in interval estimation – or, at least, the interval width aspect of this bias. Equally interesting is its failure to support the central assumption of the NSM account of overprecision. The results, therefore, point to the need to consider motivation and strategy in addition to the potential impact of cognitive processes when examining overprecision.

Acknowledgements

MBW acknowledges the support of the CIBP member companies: BG Group, ExxonMobil, Santos and Woodside.

References

Alpert, M. & Raiffa, H. (1982). A progress report on the training of probability assessors. In Daniel Kahneman, Paul Slovic & Amos Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press 294--305.

Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psych. measurement*, (54), 394-403.

Choi, J. A., Koo, M., Choi, I., & Auh, S. (2008). Need for cognitive closure and information search strategy. *Psychology & Marketing, 25*(11), 1027-1042.

Czernatowicz-Kukuczka, A., Jasko, K., & Kossowska, M. (2014). Need for closure and dealing with uncertainty in decision making context: The role of the behavioral inhibition system and working memory capacity. *Personality and Individual Differences, 70*, 126-130.

Efron, B. (1987). Better bootstrap confidence intervals. *J. of the American statistical Association*, *82*(397), 171-185.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *J. of experimental psychology: General*, *128*(3), 309.

Hansson, P., Juslin, P., & Winman, A. (2008). The role of short-term memory capacity and task experience for overconfidence in judgment under uncertainty. *J. of Experimental Psychology: L.M.C., 34*(5), 1027-1042.

Haran, U., Radzevick, J. R., & Moore, D. A. (2010). Audience effects on overprecision in judgment. *Unpublished Manuscript*.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*(3), 188-201.

Hilton, D., Regner, I., Cabantous's, L., Charalambides, L., & Vautier, S. (2011). Do positive illusions predict overconfidence in judgment? A test using interval production and probability evaluation measures of miscalibration. *J. of Behavioral Decision Making, 24*(2), 117-139.

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: A naive sampling model of intuitive confidence intervals. *Psych. Review, 114*(3), 678-703.

Kossowska, M., & Bar-Tal, Y. (2013). Need for closure and heuristic information processing: The moderating role of the ability to achieve the need for closure. *British J. of Psychology, 104*(4), 457-480.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*, (pp. 306-334). Cambridge, UK: Cambridge University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psych. Review*, *63*(2), 81.

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Keren & G. Wu (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley-Blackwell.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psych. Review, 115*(2), 502-517

Roets, A., & Van Hiel, A. (2007). Separating Ability From Need: Clarifying the Dimensional Structure of the Need for Closure Scale. *Personality and Social Psychology Bulletin, 33*(2), 266-280.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Ind. Diff., 50*(1), 90-94.

Ryan, J., Townsend, J., & Kreiner, D. (2014). Comparison of oral, written, and pointing responses to the WAIS-IV Digit Span. *Applied Neuropsychology: Adult*, *21*, 94-97.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747-758.

Van Hiel, A., & Mervielde, I. (2002). Effects of ambiguity and need for closure on the acquisition of information. *Social Cognition, 20*(5), 380-408.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *J. of Personality and Social Psychology, 67*(6), 1049-1062.

Wechsler, David. "Wechsler adult intelligence scale–Fourth Edition." *San Antonio, TX: NCS Pearson* (2008).

Welsh, M.B., Begg, S., & Bratvold, R. (2007). Modelling the Economic Impact of Common Biases on Oil and Gas Decisions. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Welsh, M.B., Bratvold, R., & Begg, S. (2005). Cognitive biases in the petroleum industry: Impact and remediation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *J. of Experimental Psychology: General, 124*(4), 424-432.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *J. of Behavioral Decision Making, 10*(1), 21-32.

A Dynamic Neural Field Model of Speech Cue Compensation

Gavin W. Jenkins (gjenkins@sfu.ca)
Paul Tupper (pft3@sfu.ca)
Department of Mathematics, 8888 University Drive
Burnaby, B.C., V5A 1S6 Canada

Abstract

Categorical speech content can often be perceived directly from continuous auditory cues in the speech stream, but human-level performance on speech recognition tasks requires compensation for contextual variables like speaker identity. Regression modeling by McMurray and Jongman (2011) has suggested that for many fricative phonemes, a compensation scheme can substantially increase categorization accuracy beyond even the information from 24 un-compensated raw speech cues. Here, we simulate the same dataset instead using a neurally rather than abstractly implemented model: a hybrid dynamic neural field model and connectionist network. Our model achieved slightly lower accuracy than McMurray and Jongman's but similar accuracy patterns across most fricatives. Results also compared similarly to more recent models that were also less neurally instantiated but somewhat closer fitting to humans in accuracy. An even less abstracted model is an immediate future goal, as is expanding the present model to additional sensory modalities and constancy/compensation effects.

Keywords: Speech recognition, concepts and categories, neural networks, dynamic systems modeling, psychology, linguistics, cognitive science

Constancy and Compensation in Perception

In most contexts, our senses provide more information than we require for a decision. This can make recognition tasks difficult when the undesired, noisy information is not just alongside but integrally mixed with desired information. As examples, the overall lighting of a scene as well as the reflectance or color of an object both affect the object's perceived lightness and hue; the actual shape of an object and viewing angle both affect perceived shape; and a speaker's gender and the content of his or her speech both affect sound pitch. Humans are adept at discounting noise and ambiguities, achieving location, shape, or speech cue constancy (Schneegans & Schöner, 2012; Rock, 1983; Bendor & Wang, 2005). Here, we present a neurally plausible, computational model potentially suitable for any type of constancy that relies on discounting dimensional feature information such as hue, shape cues, or speech cues.

Specifically, we test the model by identifying fricative consonants ('fricatives') from whole spoken syllables. We assume a speech cue to phoneme model, in contrast to a more purely acoustic approach (Graves, Mohamed, & Hinton, 2013; Pisoni, 1997), but our model could adapt to a raw acoustic approach with very similar architecture.

In the speech cue approach, few, if any, speech cues for fricatives are considered "invariant." That is, individual cues like vowel duration do not statically, cleanly, and conveniently correlate with phoneme categories (though there is some debate, see Stevens & Keyser, 2010). Rather, most or all cue information shifts contingently based on the contextual vowel sounds, speed of speech, or speaker.

Recent empirical and modeling evidence suggests that cue invariance can be overcome by considering very large numbers (dozens) of speech cues and by actively identifying contexts then normalizing incoming speech compared to other contexts. Jongman, Wayland, and Wong (2000) gathered human listener identification data for eight fricatives (f, v, θ, ð, s, z, ʃ, ʒ) or henceforth (f, v, th, dh, s, z, sh, zh) respectively, ranging from labiodental to post-alveolar place of articulation and including both voiced and voiceless fricatives at each place. Recordings of fricatives spanned over 20 speakers and 6 vowel contexts. McMurray and Jongman (2011) then tested this corpus of data using an abstract logistic regression model. They tested the model under several learning conditions, including a small collection of 10 fricative-only cues, a large set of 24 cues that also added vowel cues, and the same 24 cues, but with expectation-based context compensation for vowel and speaker, using the formal regression compensation model C-Cure (McMurray, Cole, & Munson, 2011). They determined that higher numbers of cues contributed to more accurate identification per fricative, as did expectation-based context compensation. Context compensation also allowed the model to fit human behavior more closely.

Neural Implementation of Cue Compensation

McMurray and Jongman's (2011) results are a compelling demonstration of the importance of a large number of cues and of cue compensation for fricative identification. However, the model is abstract and mathematical in both phoneme categorization and cue compensation. Such models are important, but a model in a neural framework offers a chance to discover and understand processes driving behavior that may derive from neural-level interactions not considered at an abstract level. Formal neural models also generate testable and informative neural level tests and predictions. Apfelbaum and McMurray (2015) presented a neural two layer PDP neural network for phoneme categorization, which performed impressively and comparably to McMurray and Jongman's (2011) results, but cue representation, context identification and context compensation were all still abstracted mathematically.

Here, we present a neural model to capture the same behavioral data as McMurray and Jongman (2011) and to further expand and supplement our understanding of speech cue compensation from a neural perspective. We use a

dynamic neural field (DNF) model for attention, memory, and storage of known speaker speech profiles instead of direct coding of these steps into the model, a DNF cue compensation mechanism instead of C-Cure, and a single layer neural network to ultimately decide phonemes.

The DNF architecture is described in detail below, but in general, DNF models involve fields of neural units whose receptive fields are systematically organized by dimensional information like space, size, color hue, or pitch. The DNF approach does not assume that all cognition is organized this way, only a subset of representations and processing that involve dimensional data. This includes attention to certain dimension values (attending to particular colors or points in space, for example), memory traces of those values, or in the current model, shifting cue values along a dimension to compensate for speaker identity. Beyond these processes, connectionist networks often take control. Hybrid models can include both in simulations, such as McMurray, Horst, Toscano, and Samuelson (2009) or the current model.

DNF models have been used for simulating processes ranging from word learning (McMurray, et al., 2009; Samuelson, Spencer, & Jenkins, 2013), to motor planning (Erlhagen & Schöner, 2002), to object recognition (Faubel & Schöner, 2008) and more, and the current model uses many of the same neural fields as do the above models in the same layouts. Mechanisms for different cognitive processes provide testable predictions for one another and can potentially be considered together as a coherent whole and unified model. This is not an advantage exclusive to neural models, but it is natural to them, since a shared, fundamental language is encouraged by common neural level simulation.

Our model of fricative perception utilizes several already-established DNF and connectionist mechanisms. The core neural field dynamics are common to all DNF models; the perceptual and memory portions of the model are common to DNF models that involve categorization; the phonetic cue compensation mechanism is inspired by a spatial transformation mechanism used in a DNF model of head and eye gaze spatial adjustment (Scheegans & Schöner, 2012); and the categorization step is performed by a sigmoidal connectionist network rather than logistic regression, similar to Apfelbaum and McMurray (2015).

Dynamic Neural Field Model Architecture

The Dynamic Neural Field (DNF) model consists of many 1- and 2-dimensional fields, shown as white rectangles in Figure 1. Units in fields are organized by one or two metric dimensions like cue values (such as voice onset time) or amount of adjustment needed to a cue. Each unit in a field has a receptive field maximally sensitive to one input value along its metric dimension(s) and less so to nearby values.

Units within a field interact with one another, sending close, strong local excitation and weaker, more diffuse lateral inhibition to neighbors. Both these interactions and input receptive fields create Gaussian "peaks" and "ridges" of activation when a field is given even a single value of input. These interactions keep peaks stable and localized yet robust against noise. Fields can be parameterized to have peaks collapse once input is removed (such as for attention) or to sustain themselves afterward (for memory).

Fields interact along shared dimensions. A field organized by pitch might send activation from above-threshold peaks to contribute to corresponding peaks in other pitch fields. A 1-dimensional field projecting to a 2-dimensional one projects a "ridge" of activation across all units in the larger field with the corresponding receptive fields and vice versa.

Fields also receive spatially correlated noise "input." This is insufficient to form peaks of activation alone, but is able to meaningfully influence other, stronger activity.

Figure 1 shows the full layout of *just one speech cue* in the model. That is, all of Figure 1 is repeated in the model for every speech cue. Initial input arrives at the input cue field (black star icon, blue region). In the example shown, the listener is hearing two different values of a cue. This could be due to hearing two speakers simultaneously, for example. This input projects to the adjacent attention field. Attention is a competitive field, where above-threshold units project global inhibition to the rest of the field, leading to a winner-takes-all activation pattern.

Both attention and input fields then project activation to a working memory (WM) field. This is a field with self-sustaining peaks, holding information for a time even after it has died away in temporary perception or attention fields. WM connects to a number of other fields of units representing long term memory (LTM) of speech cues. (right side of Figure 1). LTM fields are not dynamic; they are feed-forward and activated in a 1:1 correspondence with working memory. Long term memory information is held in Hebbian connection weights between WM and LTM fields. Whenever peaks are active in WM fields, LTM units represent the mathematical product of recent memory activation and LTM patterns. Thus, the total activation in LTM fields is effectively a similarity rating between recent speech cues and long term remembered patterns. If each LTM field and its Hebbian weights hold information of the history of cue values of an individual speaker, this similarity signal allows a listener to identify a known speaker's voice by competitively comparing the summed activation of LTM fields (top right of Figure 1).

Using speaker identity from above, the model activates a corresponding memory of the correct adjustment for that speaker's irregularities (green region, Figure 1). The 2-dimensional transform field (green region) accepts input from this adjustment value (top) and from the raw attended value (right) and adds them into a normalized cue value (lower left). Addition is performed by the overlap in activation between raw and adjustment ridges creating an intersecting diagonal ridge that projects to the summed value in the lower left diagonal neural field (red region of Figure 1). The same mechanism is suggested by Schneegans and Schöner (2012) for adding the angle of head rotation and angle of eye gaze to determine angles between body and objects in the visual environment. When an adjustment is

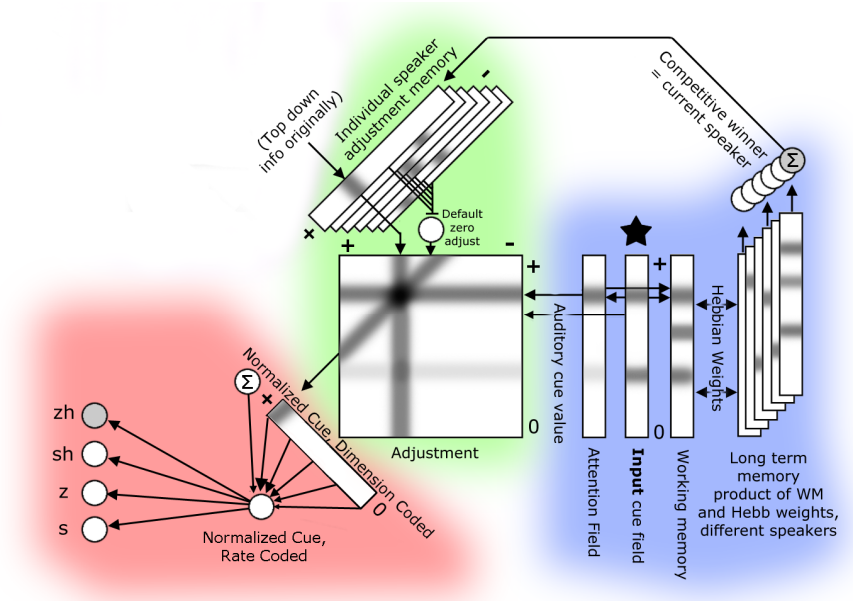


Figure 1. Architecture of a single cue in the DNF model. The full model consists of either 10 or 24 copies of this entire figure, except for just one set of output nodes shown in the lower left corner. See text for detailed description.

not known, a default adjustment of zero is used (suppressed otherwise).

The now-normalized cue information is transformed from dimensionally coded to rate coded format (red region of Figure 1). A gradient of connection weights projects stronger activation to the rate unit for peaks on one side of the cue dimension than the other. The sum of the dimensionally coded field also projects to the rate unit, allowing it to distinguish between no peak and a peak at the weak end of the scale. An almost identical neural circuit for place-to-rate code conversions is suggested by Groh (2001).

Finally, the normalized, rate-coded cue information feeds across a single layer network to determine the model's best guess at a phoneme. This final portion of the model equates to the neural portion of Apfelbaum and McMurray's 2015 PDP model of this data. Although only one cue node is shown, recall that all of the Figure 1 architecture is repeated *per cue*, creating a full [cues] x [phonemes] single layer network with one set of weights across speakers.

Experimental Design

We tested the DNF model using the fricatives phoneme dataset from Jongman, Wayland, and Wong (2000), including 8 fricatives spoken by 20 speakers each, in 6 vowel contexts (fricative + vowel syllables). McMurray and Jongman (2011) analyzed this data in several ways, but we focus here on neurally replicating three analyses in particular: phoneme categorization with 10 fricative-only cues and no compensation; with 24 cues to both fricative and accompanying vowel with no compensation, and with 24 cues and compensation. Together, these test the importance of number of cues and vowel information (10 vs 24) and compensation (24 without vs 24 with) for accurate (and human behavior-fitting) phoneme categorization.

Differences in DNF model architecture compared to McMurray and Jongman's (2011) model necessitated some lesser complexity in practical simulations. The DNF model compensated only for speaker context, not vowel context, due to the ability to only perform one compensation in a transform field at a time. Two or more compensations could simply and plausibly be performed in two or more fields in parallel, but would not better prove the concept initially and would take much longer to simulate. The neural transform field is also only capable of linear shifting adjustments along a dimension, which is less theoretically sophisticated than C-Cure compensation.

The neural network (red) portion of the DNF model was tuned first, without dynamic neural field input. Dynamic fields involve highly parallel processing, and are thus unrealistically slow to simulate on computers. Dynamic fields were therefore switched on during testing only. Training of the network used two out of every three syllable tokens from the same dataset categorized by humans, pre-coded for speech cues. In each epoch of training, the network received all training tokens once, blocked by speaker. The network received 2,000 epochs of training in each condition, with a learning rate of 0.3 and sigmoidal activation function. Cues were (mathematically during training) compensated during training for speaker prior to the neural network, for conditions including compensation.

The DNF long term memory fields were also pre-loaded with memory traces matching each speaker's cue profile (representing our memories of specific people's voices) and each speaker's adjustment value. Adjustments were chosen such that a linear shift would cause each speaker's mean value in a cue across recordings to equal the population mean value in that cue among the whole set of speakers. The model could capably establish this information itself with DNF during training, but this was impractically slow.

At test, the entire model was connected and used as a whole to categorize one epoch of the reserved generalization tokens, using the previously trained neural network.

We recorded accuracy across test trials using two different choice rules, as did McMurray and Jongman (2011). A discrete-choice rule always chose the phoneme with the highest activation in the output of the decision network. A probabilistic rule treated relative activation of each output node as relative probability of choosing that phoneme. The results figure (Figure 2) depicts colored regions bounded by these two different measures.

Results

Results of simulations are shown in Figure 2. Compared to McMurray and Jongman's (2011) mathematical model, the neural implementation of phoneme perception performed somewhat less absolutely accurately and somewhat less well fitted to human performance across conditions. However, the simulations were overall comparable while including many neurally-implemented mechanisms previously only abstractly implemented.

A large number of cues including vowel cues (24 cues, middle panel) provided moderate benefit compared to a medium number of fricative-only cues (10 cues, top panel) in terms of raw accuracy (+8% McMurray and Jongman [M&J], +11% DNF). Fit to human data worsened for M&J (RMSE from 0.077 to 0.099). Fit remained steady for the DNF model (RMSE of 0.156 in both cases), but results broadly shifted in line with human results in both models.

The addition of a speaker compensation system boosted raw accuracy in both models (+8% M&J, +3% in the DNF model) and fit human data more closely for M&J (RMSE from 0.099 to 0.061) and more closely for the DNF model (RMSE from 0.1563 to 0.1231).

The bottom panel in Figure 2 includes two additional lines, representing results from Apfelbaum and McMurray (2015) models. The grey line shows the performance of their PDP neural decision network (but otherwise not neurally implemented) model. The PDP model performed well quantitatively but with a flat performance across fricatives. The green line shows the performance of an exemplar model that stored every individual syllable token from the training set in memory for use in categorizing test items. The exemplar model performed very well, but had no neural implementation, and storage of every individual syllable in memory is likely unrealistic.

Qualitatively, the pattern of accuracy in the DNF model fits human data in shape about as well as the other models. Non-sibilant fricatives 'f' and 'v' were routinely low accuracy compared to humans for the DNF model, but the 'th, dh, s' dip shape is more accurately captured across conditions by the DNF model.

In an attempt to explore possible causes of our model's consistently low accuracy for 'f' and 'v', we tested the model with portions of the data including only 2 phonemes and 1 cue at a time. Two examples of 2-phoneme, 1-cue results are shown in Figure 3. The top row is data before speaker compensation, and the bottom row after. Horizontal position is value along the listed cue dimension (measured from the normalized DNF cue field) and each thin rows of dots is a different speaker. For some cues like spectral kurtosis (left), linear speaker compensation was able to actually *lower* accuracy in this analysis. One speaker may have had higher values for 'v' than 'zh' while another had not just a shifted but an *opposite* relationship between the same phonemes. Thus, when speakers' adjustments were chosen to match global means, some values for each phoneme shifted one direction, others shifted another, and confusion between the phonemes actually increased!

For other cues, like fricative duration (right side of Figure 3), there was a consistent relationship across speakers for one phoneme's values versus another, and compensation helped across all speakers.

Running the DNF model without the worst cues did not, however, increase model fits. The best results (by a small margin) were found when the worst nine cues were left uncompensated, but the overall improvement was not significantly better fitting than when compensating all cues, suggesting the network in the model is capable of discounting unproductive cues sufficiently. Low DNF accuracy for 'f' and 'v' fricatives remains unexplained.

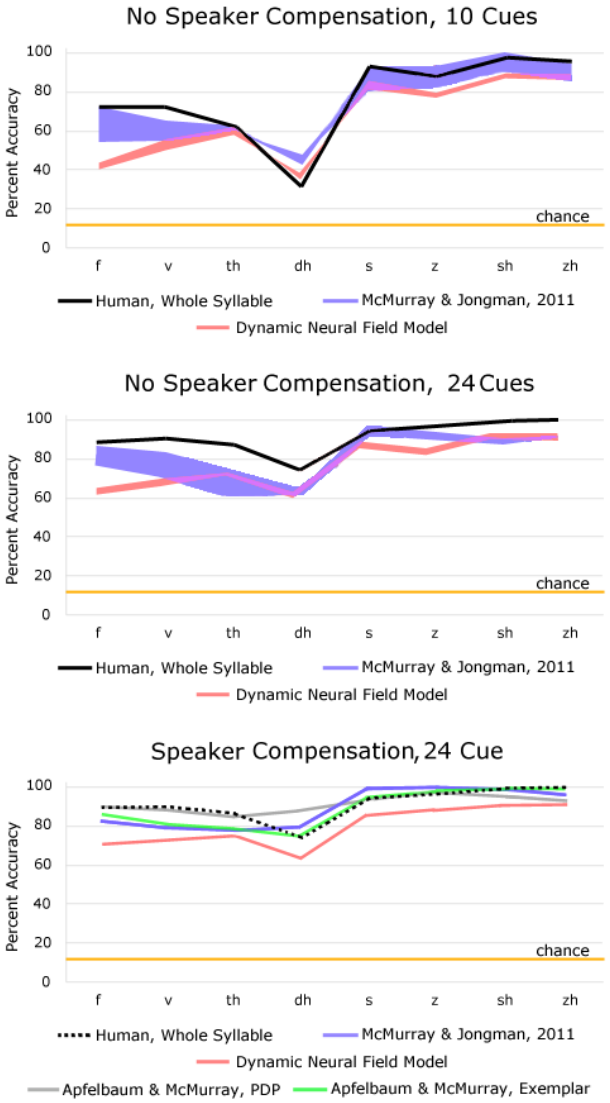


Figure 2. Simulation results. Shaded regions in the top two panels represent the accuracy range bounded by discrete-choice and probabilistic accuracy. The bottom panel shows only the mean of these two measures and dashed target line for easier reading.

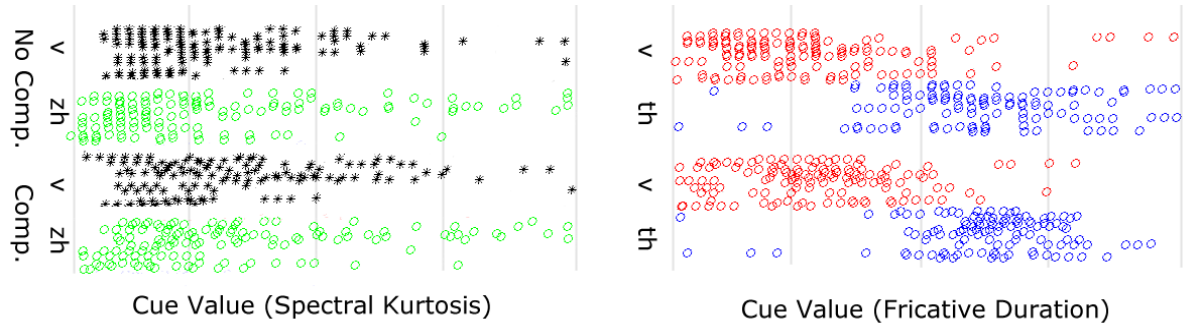


Figure 3. Cue values for test speakers before and after speaker compensation in the DNF model. Spectral kurtosis results in lower accuracy after compensation, while fricative duration results in higher accuracy.

Discussion

The DNF model establishes a set of plausible neural mechanisms for categorizing speech cues and compensating for variance between speakers. Accuracy was unsurprisingly lower than more abstract models, but only slightly so, and the qualitative pattern of DNF model results realistically follows that of human behavior.

The weakest portion of the DNF model quantitatively is its 'f' and 'v' accuracy. The possibility that some cues like spectral kurtosis were unhelpful for these phonemes when compensating for speaker was investigated and rejected as a hypothesis. An alternative explanation for this relative weakness is a high priority for future modeling.

Other directions for future work involve capitalizing on the rich representations that exist in a neural implementation and removing automated processes in model simulations.

The DNF model currently pre-loads some information into long term memory fields, like speaker adjustment profiles (as do competitor models). One high priority for model improvement is to remove this artificial seeding of information into the model and replace it with online learning of speaker adjustments. Speaker adjustments will be learned when context clues in the speech environment provide information about a speaker's intended phoneme beyond speech cues alone. Correct phoneme information can activate expected values for speech cues for that phoneme. These values can then be subtracted from the raw, perceived speech cue values of the speaker, using a transform field exactly like the one in the green region of Figure 1. The resulting adjustment value will then be stored for use later when context clues are unavailable (achieving the current starting point of the model).

The DNF model also currently blocks trials by speaker in order to conveniently activate one speaker memory profile at a time without rapid switching. Humans are able to utilize speaker information per syllable, however, so adapting the model to have this capability is a third modeling priority. This improvement requires only parameterization and gating of existing units to achieve more reliable timing.

Some of the DNF model's features suggest testable predictions for future investigation as well. For example, the DNF model naturally accounts for both prototype and exemplar memory representations, with no qualitatively different architecture than is described above. The dimension-coded LTM fields (right side of Figure 1) can store a profile of information about a specific speaker's history of one auditory cue's values, but could as easily store a profile for "males" or "females" in general, or for a specific moment of speech in time. Such stored information should be able to capture known human behaviors. For instance, Johnson, Strand, and D'Imperio (1999) presented discrimination results between speakers by gender that may be captured by DNF modeling. Artificial groups of phoneme tokens should also be able to be constructed grouped by a features beyond gender or speaker, such as visual scene context, arbitrary label, or otherwise. The DNF model would predict that active compensation for any such grouping may be feasible, useful, and actually utilized by humans in categorizing those phonemes.

The fact that the DNF model can represent 24-cue stimuli at all in a neurally plausible way is an advantage over some accounts of stimulus representation that will be explored further. McMurray and Jongman (2011) established the importance of considering many cues for accuracy. Many models have represented multidimensional stimuli, however, only in an abstract, n-dimensional "feature space" (Richardson, 1938; Nosofsky, 1986). Such a space cannot exist biologically, since a handful of dimensions requires more neurons than exist in the brain. The DNF model offers a solution to this problem: the architecture for a single cue (Figure 1) need only be replicated linearly for additional cues. 24-cue stimuli require only 24 times more neural resources than single cue stimuli.

This advantage of easily incorporating many new dimensions of information extends beyond auditory cues to dimensional information of nearly every other type. Visual dimensional features, for example, like size, orientation, spatial frequency, or color hue can be represented in much the same way as speech cues. Critically, just as voice features can be compensated and discounted, then, so can environmental lighting, distance, or angle of view. The current model can therefore potentially provide a unified explanation of effects like shape or size constancy as a

result of the same or a similar mechanism to phoneme constancy shown here. The model may similarly be relevant to aftereffects in "high-level" sensory processing, like face distortion adaptation (Köhler & Wallach, 1944), as a result of adjustment peaks requiring a short period of time to shift or die away when a stimulus is changed.

A final advantage of the DNF model is that it can process several feature or cue values at once through its compensation mechanism. Only one adjustment is used here for practical simulation time during initial modeling, but any number of parallel fields could be used with only linearly increasing neural investment, allowing many types of compensation at once. This is unlikely to help phoneme categorization further, but it would predict an advantage for compensating for context information in whole "scenes" of information in other modalities, such as distinguishing between "forest sounds" and "jungle sounds," benefitting from many parallel compensations along diverse dimensions at once. The DNF model not only predicts the capability of humans to perform *normalized* categorizations of this sort, but it predicts specific dynamics. If cue values are distinct, they should not interfere with one another, but if near one another along a dimension, lateral dynamics within the initial perceptual fields (blue region, Figure 1) fields should sharpen both peaks, or cause them to merge, etc., leading to distinct predicted categorization decisions.

Overall, compensation performance in the DNF model is promising. Future work will focus on increasing the model's self-sufficiency without pre-loaded information, investigation of novel behavioral predictions of the model, and expanding simulations to other sensory domains.

Acknowledgments

This research was supported in part by a Discovery Grant awarded by NSERC Canada and a Tier II Canada Research Chair. The authors would like to thank Bob McMurray and Allard Jongman for their prior model and consultation.

References

Apfelbaum, K., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. Psychon Bull Rev, *22*, 916-943.

Bendor, D. & Wang, X. (2005). The neuronal representation of pitch in primate auditory cortex, *Letters to Nature, 436*, 1161-1165.

Erlhagen, W. & S. & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review, 109*, 545-572.

Faubel, C. & Schöner, G. (2008). Learning to recognize objects on the fly: A neurally based dynamic field approach. *Neural Networks, 98* (3), 419-456.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psych. Rev., 105*, 251-279

Graves, A., Mohamed, A.-R., and Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6645-6649.

Groh, J. M. (2001). Converting neural signals from place codes to rate codes. *Biological Cybernetics, 85*, 159-165.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359-384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America, 108*, 1252-1263.

Köhler, W. & Wallach, H. (1944). Figural after-effects: An investigation of visual processes. *Proceedings of the American Philosophical Society, 88*, 269-357.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In C.N. Clements and R. Ridouane (Eds.), *Where do Phonological Features Come From? Cognitive, Physical, and Developmental Bases of Distinctive Speech Categories* (197-236). Amsterdan, John Benjamins.

McMurray, B. & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychogical Review, 2*, 219-246.

McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. K. (2009). Integrating connectionist learning and dynamical systems processing: Case studies in speech and lexical development. In Spencer, J. P., Thomas, S. C., & McClelland, J. L. (Eds.), *Toward a Unified Theory of Development* (218-249). Oxford University Press.

Miller, J. L. (2001). Mapping from acoustic signal to phonetic category: Internal category structure, context effects and speeded categorization, *Language and Cognitive Processes*, *16*(5/6), 683-690.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39-57.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (9-32). San Diego, CA: Academic Press.

Richardson, M.W. (1938). Multidimensional psychophysics. *Psychological Bulletin, 35*, 659-660.

Rock, I. (1983). *The logic of perception.* Cambridge, MA: MIT Press.

Samuelson, L. K., Spencer, J. P., & Jenkins, G. W. (2013). A dynamic neural field model of word learning. In Gogate & G. Hollich (Eds.), *Theoretical and computational models of word learning: Trends in psychology and artificial intelligence.* Hershey, PA, US: Information Science Reference / IGI Global.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics, 106*, 89-109.

Stevens, K. N. & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics, 38*,10-19.

The face-space duality hypothesis: a computational model

Jonathan Vitale (jonathan.vitale@student.uts.edu.au)
Mary-Anne Williams (mary-anne.williams@uts.edu.au)
Benjamin Johnston (benjamin.johnston@uts.edu.au)
University of Technology, Sydney
QCIS Centre - Innovation and Enterprise Research Lab, 15 Broadway - Ultimo NSW 2007

Abstract

Valentine's face-space suggests that faces are represented in a psychological multidimensional space according to their perceived properties. However, the proposed framework was initially designed as an account of invariant facial features only, and explanations for dynamic features representation were neglected. In this paper we propose, develop and evaluate a computational model for a twofold structure of the face-space, able to unify both identity and expression representations in a single implemented model. To capture both invariant and dynamic facial features we introduce the *face-space duality hypothesis* and subsequently validate it through a mathematical presentation using a general approach to dimensionality reduction. Two experiments with real facial images show that the proposed face-space: (1) supports both identity and expression recognition, and (2) has a twofold structure anticipated by our formal argument.

Keywords: face perception; face processing; face-space; duality hypothesis; dimensionality reduction

Introduction

As an explanation of findings in face perception, Valentine used formal models of concept representations to propose that faces are represented in a psychologically plausible multidimensional space, *i.e.* the *face-space* (Valentine, 1991). Faces are points of this space based on their perceived properties. Valentine's formal models have been used to explain results of many human experimental studies, as well as computational simulations (Calder, Burton, Miller, Young, & Akamatsu, 2001; Lee, Byatt, & Rhodes, 2000; Rhodes et al., 2011).

These models were initially designed to only account for coding *identity-related* features, such as sex, distinctiveness, age and attractiveness (Calder et al., 2001). For example, the feature 'eyebrows' can vary from marked to delicate, thus possibly being one of the perceivable features crucial for coding the sex of a face. However, dynamic aspects of faces, such as facial expressions, were neglected. Early brain lesion and neuroimaging studies suggested that face identity and expression are not integral dimensions[1], instead they are represented and processed by separate systems that process faces in parallel (Bruyer et al., 1983; Tranel, Damasio, & Damasio, 1988). However, contemporary understanding indicates that identity and expression are more closely connected than previously thought, suggesting that common-codings can respond to both of them and so processes of their perception can interact (Ganel, Valyear, Goshen-Gottstein, & Goodale, 2005; Kadosh et al., 2016; Rhodes et al., 2015).

[1]Here and for the rest of the paper we use the term 'integral' to define properties exhibiting inter-dependencies and not completely separable.

Can Valentine's framework support an integral understanding of identity and expression processing? Calder and colleagues (2001) demonstrated that a multidimensional space derived from a principal component analysis (PCA) (Turk & Pentland, 1991) can provide a set of components being either identity-independent, expression-independent or identity-expression-interdependent (Calder & Young, 2005). However, in order to perform identity and expression recognition, the authors used two distinct latent discriminant analysis (LDA) modules, one choosing the best components to support identity recognition, whereas the other choosing the best components to support expression classification.

In this paper, we show that a *single* face-space with a twofold representation supports both identity and expression recognition. The structure of this face-space can be realized in a parsimonious way by integrating both identity and expression in a single model. We demonstrate the computational validity of our hypothesis through a rigorous mathematical presentation and related experiments. This work further supports the integral nature of identity and expression, at least from a computational perspective.

Findings in Face Perception

Valentine's *'face-space'* framework (Valentine, 1991) is a notable cognitive model for face representation. According to this framework, facial representations are encoded in a multidimensional psychological space. The dimensions of this space are assumed to encode properties of the facial signals that better discriminate one face from another. The distance between two representations underlies their dissimilarity from a psychological perspective.

Identity and expression are two forms of facial information crucial for many social skills. Identity is considered an invariant feature of face, whereas expression a dynamic one. Traditional cognitive models of face perception suggest a complete separation of identity and expression systems after the completion of a structural encoding stage (Bruce & Young, 1986). Accordingly, Haxby *et al.* (2000) argue that invariant features, such as identity, and dynamic features, such as facial expression, are computed by separate regions of the brain.

However, new evidence from recent findings indicates that these systems operate interdependently (Pell & Richards, 2013). For example, Ganel and Goshen-Gottstein (2004) found that familiarity of faces increases the perceptual interdependence of identity and expression recognition. They suggested that differences between the facial configurations of individuals should lead to systematic differences in the

way emotions are expressed by these individuals. For this reason, every individual can express each facial expression in a unique way. Knowledge of the identity of the observed subject can therefore facilitate the process of his or her facial expression.

We found a similar effect in developing a computational model inspired by simulation-theories (Vitale, Williams, Johnston, & Boccignone, 2014). In this endeavour, we suggested to first pre-process the observed face so to reduce its identity-related information, while preserving motor components information (*i.e.* its dynamic features). This facilitates the recognition of facial expression based on feed-forward mechanisms of internal motor simulation. A similar account was recently supported by a human study of Ipser and Cook (2015).

Calder *et al.* (2001) demonstrated the validity of integral identity and expression representations from a computational perspective. They submitted digital images of faces showing different identities and facial expressions to PCA thus obtaining representations based on components of a low-dimensional space. Their results demonstrated that this common representation can support both identity and expression recognition and that the representations of identity and expression partially overlap.

The Face-Space Duality Hypothesis

In the previous section we provided studies supporting:

i A multidimensional spatial representation of faces as a plausible model for explaining many crucial effects in face perception;

ii An interdependence between representations of invariant and dynamic facial features;

iii That enhancing the separation between invariant and dynamic components of the face during face processing facilitates their classification.

Given this brief summary, we introduce the *'face-space duality hypothesis'*, suggesting that faces: (i) are encoded in a multidimensional face-space, (ii) under a common integral representation (iii) having a twofold structure: one supporting invariant features of the face (*e.g.* identity), whereas the other supporting dynamic features of the face (*e.g.* expression), thus contributing to their correct classification.

This hypothesis arises from the need to accommodate the apparently contrasting points (ii) and (iii) under a *single* representation based on a multidimensional space (i), which is an extension of the face-space framework proposed by Valentine (1991).

Consider a perceived face ϕ_i and a corresponding d-dimensional point y_i of a multidimensional psychological space. The spatial representation y_i encodes most of the original information of the input face ϕ_i and can be obtained through the mapping function $S(\phi_i) \mapsto y_i$. We introduce the functions $C^{\mathscr{E}}(\cdot) \rightarrow \mathbf{R}$ and $C^{\mathscr{I}}(\cdot) \rightarrow \mathbf{R}$, respectively providing the number of correctly classified facial expressions and identities from a set of observations encoding faces, and a permutation function σ over the coordinates $(y^1, y^2, \ldots, y^d)$ of a point y defined as:

$$\sigma = \begin{pmatrix} y^1 & y^2 & y^3 & \cdots & y^d \\ y^d & y^{d-1} & y^{d-2} & \cdots & y^1 \end{pmatrix} \tag{1}$$

We make use, here and for the rest of the paper, of the superscript $\tilde{\cdot}$ to denote a point or set of points to which is applied the permutation in (1). Then, given a set of perceived observations Φ and the associated multidimensional spatial representations Y, the mapping function S is defined such that:

1. $S(\Phi) \mapsto Y$;
2. $C^{\mathscr{E}}(Y) \gg C^{\mathscr{E}}(\Phi)$;
3. $C^{\mathscr{I}}(\tilde{Y}) \gg C^{\mathscr{I}}(\Phi)$;

In other words, the face-space duality hypothesis assumes a function S mapping the perceived faces onto a psychological multidimensional space having a twofold structure. This new representation allows the encodings $Y = \{y_1, y_2, \ldots, y_n\}$ and associated permutations $\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n\}$ to support significantly higher recognition rates than the original input representation Φ.

The rationale behind this idea is that the dual face-space, by maximising the separation between dynamic and invariant features of the face in a single multidimensional representation, will order the components of the resulting space in such a way that the first ones will mostly correlate with dynamic features of the face, whereas the latter will mostly correlate with invariant features of the face. Therefore, the resulting face-space would provide a single multidimensional representation (as per point (i)) where invariant and dynamic features of the face are interdependent (as per point (ii)), but preserving a certain degree of separation able to support subsequent classification processes (as per point (iii)).

We investigate our hypothesis from a computational perspective, validating it using a mathematical analysis of a general dimensionality reduction framework used in face recognition, by limiting the analysis to facial identity and expression only. We further confirm our approach with experimental results.

Dimensionality Reduction Models

The function S introduced above can be modeled as a dimensionality reduction function. A dimensionality reduction function maps a high-dimensional signal onto a point of a low-dimensional space. For example, consider an image of a face having resolution 100×100 pixels. This observed signal is represented by a set of pixels and can be posed as a column vector ϕ_i of dimension $\mathscr{D} = 10000$. Dimensionality reduction models provide a mapping function $S : \mathbf{R}^{\mathscr{D} \times 1} \rightarrow \mathbf{R}^{d \times 1}$, with $d \ll \mathscr{D}$, such that the low-dimensional representation

$y_i = S(\phi_i)$ is able to explain the observed data ϕ_i (Yan et al., 2007).

Linear dimensionality reduction techniques make use of a linear projection matrix $V \in \mathbf{R}^{\mathscr{D} \times d}$ in order to map the high dimensional observed sample onto the low-dimensional target space. So the projection y_i of an observation ϕ_i can be computed as $y_i = V^\top \phi_i$. When V is an orthogonal matrix, an approximation of the original observation can be reconstructed from its projection: $\phi_i \approx V y_i$. The projection matrix V can be estimated by solving an *objective function*. This objective function models desired constraints that the structure of the target low-dimensional space is required to satisfy.

For the purposes of this paper we limit our investigation to provide an implementation of our model through *linear* dimensionality reduction techniques.

Graph-based dimensionality reduction framework

Yan *et al.* (2007) provided a general framework for unifying many dimensionality reductions models. Since our approach makes use of this framework, we first briefly introduce it below.

Let $\Phi = [\phi_i, \ldots, \phi_n]$ be a matrix of the N observations represented as column vectors with dimension $\mathscr{D}$. The structure of the target low-dimensional space can be constrained by a similarity matrix W and a penalty matrix $W^{(p)}$. For each pair of samples (ϕ_i, ϕ_j), the similarity matrix W_{ij} encodes the associated non-negative similarity measure, whilst the penalty matrix $W^{(p)}_{ij}$ encodes the associated penalty measure. This penalty measure can be used as a repulsive force between pairs of samples to prevent samples with high similarity but belonging to different classes from being placed in close proximity in the low-dimensional space (Kokiopoulou & Saad, 2009).

Given these two graph structures, the optimal mapping matrix $V^\star$ can be found by solving the following objective function:

$$V^\star = \underset{V \in \mathbf{R}^{\mathscr{D} \times d}}{\arg\min} \frac{Tr(V^\top \Phi L \Phi^\top V)}{Tr(V^\top \Phi L^{(p)} \Phi^\top V)} \tag{2}$$

with $Tr(\cdot)$ denoting the matrix trace operator and L, $L^{(p)}$ respectively being the resulting Laplacian matrices computed from the similarity and penalty matrices W and $W^{(p)}$ (Yan et al., 2007).

Unfortunately, there is no closed-form solution to this optimization problem (Ngo, Bellalij, & Saad, 2012). However, the problem can be solved numerically with iterative algorithms whenever the matrix $\Phi L^{(p)} \Phi^\top$ is positive definite. The resulting optimal solution $V^\star$ is unique up to unitary transforms of the columns (Ngo et al., 2012).

To ensure that matrix $\Phi L^{(p)} \Phi^\top$ is positive definite, the process is usually split into two phases. First, given a dimension $\mathscr{D}' < N$, the observations Φ are provided in input to a PCA, and a first mapping matrix $V_{PCA} \in \mathbf{R}^{\mathscr{D} \times \mathscr{D}'}$ is estimated. Then, the samples $X = V_{PCA}^\top \Phi$ are provided as input to the objective function in (2) and the optimal mapping matrix $V^\star \in \mathbf{R}^{\mathscr{D}' \times d}$ is estimated. Finally, the overall mapping matrix able to reduce the dimensionality from dimension $\mathscr{D}$ to dimension d and performing the constraints specified in the objective function (2) is given by:

$$V_{overall} = V_{PCA} V^\star \tag{3}$$

Model implementation

Consider a set Φ of N observations of frontal faces. Each observation consists of $\mathscr{D}$ pixel values, represented by a $\mathscr{D} \times 1$ vector. In this work we limit the investigation of observations varying only in identity and facial expression.

We set a dimension $d < N \ll \mathscr{D}$ and we estimate a mapping matrix $V_{PCA} \in \mathbf{R}^{\mathscr{D} \times d}$ by submitting the samples Φ to a PCA, thus obtaining the corresponding PCA-encodings $X = V_{PCA}^\top \Phi$. We aim to estimate another mapping matrix $V^\star \in \mathbf{R}^{d \times d}$, such that the final overall matrix $V_{overall} = V_{PCA} V^\star$ validates our hypothesis.

We denote the identity class of the sample x_i with $\mathscr{I}(x_i)$ and the facial expression class of the sample x_i with $\mathscr{E}(x_i)$.

Considering the previously introduced scenario, we start by designing the appropriate similarity and penalty matrices. Invariant features of the face extend to more regions of the face than dynamic ones, thus explaining most of the variance in the considered dataset of observations (Turk & Pentland, 1991). This means that during facial expression classification the identity can potentially introduce a bias on the samples (the *identity-bias*), thus increasing their similarity and reducing their distance in the face-space even when they belong to a different class of facial expression (Sariyanidi, Gunes, & Cavallaro, 2015).

Therefore, as a first step we design our similarity matrix to encourage pairs of samples associated with the same facial expression to be in close proximity in the resulting space, and our penalty matrix to provide a repulsive force between pairs of samples belonging to the same identity. This would result in maximising the separation between dynamic and invariant components of the face, thus facilitating the classification of facial expression.

Accordingly, we define the similarity matrix $W^{\mathscr{E}}$ as:

$$W^{\mathscr{E}}_{ij} = \begin{cases} \frac{1}{n_{\mathscr{E}_i}}, & \text{if } \mathscr{E}(x_i) = \mathscr{E}(x_j) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $n_{\mathscr{E}_i}$ is the number of samples in X belonging to facial expression class $\mathscr{E}(x_i)$.

Similarly, we define the penalty matrix $W^{\mathscr{I}}$ as:

$$W^{\mathscr{I}}_{ij} = \begin{cases} \frac{1}{n_{\mathscr{I}_i}}, & \text{if } \mathscr{I}(x_i) = \mathscr{I}(x_j) \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where $n_{\mathscr{I}_i}$ is the number of samples in X belonging to identity class $\mathscr{I}(x_i)$.

By using the proposed similarity and penalty matrices, the resulting Laplacians becomes $L = I_N - W^{\mathscr{E}}$ and $L^{(p)} = I_N - W^{\mathscr{I}}$, with I_N a $N \times N$ identity matrix. Hence, these

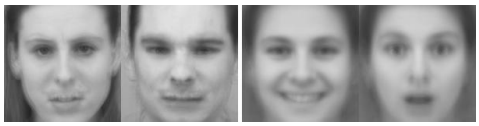


Figure 1: Some examples of prototypes. On the left are two prototypical identities (F05 and M07) in which expression-related features are reduced, whereas on the right are two examples of prototypical facial expressions (happiness and surprise).

Laplacians behave as block centring matrices. These matrices remove respectively the corresponding prototypical facial expression (*i.e.* an average identity showing the averaged facial expression) and the corresponding prototypical identity (*i.e.* the considered identity showing a neutral facial expression) from samples X (see figure (1)).

The objective function in (2) becomes:

$$\arg\min_{V \in \mathbf{R}^{d\times d}} \frac{Tr(V^\top X(I_N - W^{\mathscr{E}})X^\top V)}{Tr(V^\top X(I_N - W^{\mathscr{I}})X^\top V)} \tag{6}$$

We solve the objective function in (6) with the iterative algorithm proposed in (Ngo et al., 2012). Given the matrices $M^{\mathscr{E}} = X(I_N - W^{\mathscr{E}})X^\top$ and $M^{\mathscr{I}} = X(I_N - W^{\mathscr{I}})X^\top$ the optimal mapping matrix $V^\star$ can be found through the algorithm (1).

Data: Matrices $M^{\mathscr{E}}$, $M^{\mathscr{I}}$, a maximum number of iterations K and a tolerance ε.
Result: A mapping matrix V of dimension $D \times d$.
$V \leftarrow I_{D\times d}$;
for $i \leftarrow 1$ **to** K **do**
  $\rho \leftarrow \frac{Tr(V^\top M^{\mathscr{E}} V)}{Tr(V^\top M^{\mathscr{I}} V)}$;
  $\mathcal{G}(\rho) \leftarrow M^{\mathscr{E}} - \rho M^{\mathscr{I}}$;
  Compute the smallest (for minimisation) or largest (for maximisation) d eigenvalues $[\lambda_1, \ldots, \lambda_d] \equiv \Lambda$ of $\mathcal{G}(\rho)$ and associated eigenvectors $[v_1, \ldots, v_d] \equiv V$;
  if $|\sum_{j=1}^{d} \Lambda| < \varepsilon$ **then**
    break;
  end
end

Algorithm 1: Newton-Lanczos algorithm for optimization of objective function (6).

From the algorithm (1), it can be seen that the optimal mapping matrix $V^\star$ is the set of eigenvectors associated with the smallest eigenvalues of $\mathcal{G}(\rho^\star)$, with $\rho^\star$ being the result of the trace ratio in (6) when posing the optimal solution $V^\star$. If, instead of taking the eigenvectors associated with the smallest eigenvalues, we take the eigenvectors associated with the largest eigenvalues, we get the optimal solution for the following objective function:

$$\arg\max_{V \in \mathbf{R}^{d\times d}} \frac{Tr(V^\top X(I_N - W^{\mathscr{E}})X^\top V)}{Tr(V^\top X(I_N - W^{\mathscr{I}})X^\top V)} \tag{7}$$

Using simple properties of trace and eigenvalues, it follows that the objective function in (7) can be equivalently posed as:

$$\arg\min_{V \in \mathbf{R}^{d\times d}} \frac{Tr(V^\top X(I_N - W^{\mathscr{I}})X^\top V)}{Tr(V^\top X(I_N - W^{\mathscr{E}})X^\top V)} \tag{8}$$

By reminding that a centring matrix is symmetric idempotent and $Tr(AA^\top) = \|A\|_2^2$, it is possible to note that the objective function (6) attempts in minimising the distances between the encodings $Y = V^\top X$ and their corresponding prototypical facial expressions $Y^{\mathscr{E}}_{proto} = V^\top X W^{\mathscr{E}}$, while maximising their distances with respect to the prototypical identity $Y^{\mathscr{I}}_{proto} = V^\top X W^{\mathscr{I}}$, overall facilitating expression recognition. Conversely, the objective functions (7, 8) attempts in minimising the distances between the encodings $Y = V^\top X$ and their corresponding prototypical identities $Y^{\mathscr{I}}_{proto} = V^\top X W^{\mathscr{I}}$, while maximising their distances with respect to the prototypical facial expression $Y^{\mathscr{E}}_{proto} = V^\top X W^{\mathscr{E}}$, overall facilitating identity recognition.

Since the eigenvectors of $V^\star$ are estimated from the same matrix $\mathcal{G}(\rho^\star) = M^{\mathscr{E}} - \rho^\star M^{\mathscr{I}}$ in both the objective functions (6,7), the components of the two spaces are the same, but differing by order. In other words, given $V^\star$ as the optimal matrix resulting from objective function (6) we can easily get the optimal matrix of the objective function (7) $\tilde{V}^\star$, defined as a matrix with the same columns of $V^\star$, but arranged in an inverse order.

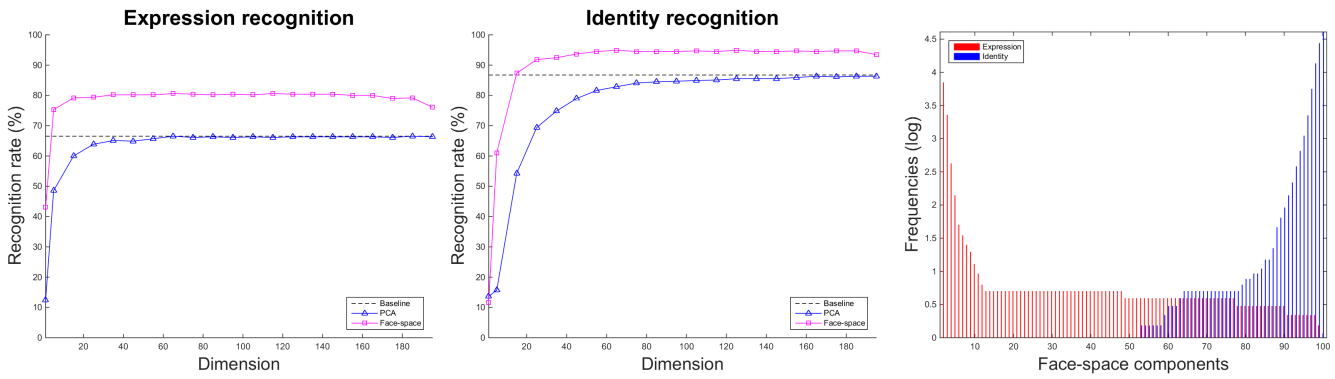
Finally, given the matrix $V_{PCA} \in \mathbf{R}^{\mathcal{D}\times d}$ and the matrix $V^\star \in \mathbf{R}^{d\times d}$ we can estimate the final mapping matrix $V_{overall}$ of the face-space through equation (3), which leads respectively to the mapping $Y = V^\top_{overall} X$ and the associated permutated mappings $\tilde{Y} = \tilde{V}^\top_{overall} X$ as suggested by our hypothesis.

Note that we are not claiming that this is the only way to implement the proposed mapping function S (for example it can be generalised to non linear models), and neither that human brain implement the suggested dual face-space in this way. However, this is a viable computational implementation of the proposed model able to support our hypothesis. In the remainder of this paper we further support the face-space duality hypothesis with experimental data.

Experiments

We further validate our hypothesis using images from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, Flykt, & Öhman, 1998). The dataset contains static images of 70 subjects—35 female and 35 male—exhibiting 7 different prototypical facial expressions of basic emotions (anger, disgust, fear, happiness, neutral, sadness and surprise). The pictures are taken in different face orientations and in two different sessions (A and B).

We used the frontal pictures taken in session A. The facial region was extracted from the images and its resolution



(a) Expression recognition task (b) Identity recognition task (c) Components used in recognition tasks

Figure 2: Results of the proposed experiments.

reduced to 80×80 pixels. Eyes and mouth were at approximately the same position. Illumination variations were reduced by applying a simple equalization process to the images.

We first pre-processed the data by submitting the pixels of the images in input to a PCA as explained previously. In the first experiment we retained the components able to explain 95% of the variance of the original data resulting in 200 components, while in the second experiment we retained the data explaining the 85% (so reducing the number of components and allowing better readability) resulting in 100 components.

We performed two experiments: the first to test the ability of the proposed face-space in supporting identity and expression recognition, and the second to demonstrate the twofold nature of the resulting face-space.

Support of identity and facial expression recognition

In the first experiment we test the ability of our model to support subsequent processes of identity and facial expression recognition. We used a 10-fold cross validation approach. For each iteration we divided the data by taking one fold as test and the rest for model training.

With each training data we estimated the mapping matrix $V_{overall}$ as per equations (6, 3). Then we mapped each test data onto the corresponding face-space, thus obtaining the encodings $Y^{\mathscr{E}} = V_{overall}^{\top}\Phi$ and $\tilde{Y}^{\mathscr{I}} = \tilde{V}_{overall}^{\top}\Phi$ respectively used during the expression and identity recognition tasks.

The classification was performed using the nearest neighbour algorithm. For each sample ϕ_i, x_i and y_i we computed the Euclidean distances with respect to the centroids of each class in the corresponding space (*i.e.* the prototypical identities in the case of identity recognition or the prototypical expressions in the case of facial expression recognition) and selected the label associated with the centroid having lower distance to the sample. We repeated this process for each dimension k by taking only the first k components of the encodings x_i and y_i. In classifying the raw observations ϕ_i for a baseline comparison, we considered all the pixel values of the input images as coordinates of points in a $\mathscr{D}$-dimensional space. The recognition rates in each dimension were averaged among each cross-validation test.

The results for facial expression and identity recognition are shown, respectively, in figures (2a) and (2b). It is clear that this face-space derived by a single integrated process can support both identity and expression recognition, whereas a simple PCA cannot overcome the baseline performance.

Validation of face-space twofold structure

We were able to confirm our hypothesis on the twofold structure of the face-space using data.

We estimated the mapping matrix $V_{overall}$ as per equations (6, 3) using the full dataset as training data.

Given the matrix $V_{overall} = [v_1, \ldots, v_d]$ the minimum set of expression components for a sample ϕ_i is the smallest set $V_{i_{min}} = [v_1, \ldots, v_k]$ such that $y_i = V_{i_{min}}^{\top}\phi_i$ is classified with the correct expression label $\mathscr{E}(\phi_i)$ through nearest neighbour algorithm, as in the previous experiment.

Similarly, given the matrix $\tilde{V}_{overall} = [v_d, \ldots, v_1]$ the minimum set of identity components for a sample ϕ_i is the smallest set $\tilde{V}_{i_{min}} = [v_k, \ldots, v_1]$ such that $y_i = \tilde{V}_{i_{min}}^{\top}\phi_i$ is classified with the correct identity label $\mathscr{I}(\phi_i)$ through nearest neighbour algorithm, as in the previous experiment.

For each sample ϕ_i we computed its minimum set of expression and identity components. Then, we set $n_k^{\mathscr{E}}$ the number of times the component k was included in the minimum sets of expression components and $n_k^{\mathscr{I}}$ the number of times the component k was included in the minimum sets of identity components. We computed the final results for expression and identity and for each component k as $f_k^{\mathscr{E}} = \log(\frac{n_k^{\mathscr{E}}}{N} \times 100 + 1)$ and $f_k^{\mathscr{I}} = \log(\frac{n_k^{\mathscr{I}}}{N} \times 100 + 1)$. Here N is the number of samples in the dataset (*i.e.* 490) and we used the logarithm for better readability of the results. The resulting log-frequencies are illustrated in Figure (2c).

From the results two peaks placed in the extremes of the face-space components are clearly visible. There are expression components clearly independent from identity ones (components #1 to #52), components shared among expression and identity classification tasks (components #53 to #99)

and just one identity component independent from expression ones (component #100). These results are in agreement with the study of Ganel and Goshen-Gottstein (2004), which suggest that expression is perceptually separable from identity, but identity is not perceptually separable from expression. This experiment further supports the proposed twofold structure of face-space as suggested by the face-space duality hypothesis.

Conclusions

In this paper we extended Valentine's face-space framework to better explain recent findings in face perception.

In alignment with recent studies in face processing, we suggested that the nature of identity and expression dimensions is highly interdependent. We proposed that the structure of the face-space can result from a single process integrating invariant and dynamic features of the face. From this process results a twofold structure. Reading the components of this space from the first to the last facilitates classification of dynamic features of the face, such as facial expression, while reading them from the last to the first better supports classification of invariant features, such as the identity.

We validated the face-space duality hypothesis, from a computational perspective, through a formal mathematical presentation, by considering a general framework of dimensionality reduction. The hypothesis was further supported by experimental data.

This framework might serve as the basis and inspiration for future empirical work on face processing and linked to neural approaches.

References

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*, 305–327.

Bruyer, R., Laterre, C., Seron, X., Feyereisen, P., Strypstein, E., Pierrard, E., & Rectem, D. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and cognition*, *2*(3), 257–284.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision research*, *41*(9), 1179–1208.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–651.

Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: the parallel-route hypothesis revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 583.

Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of the fusiform face area in processing facial expression. *Neuropsychologia*, *43*(11), 1645–1654.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223–233.

Ipser, A., & Cook, R. (2015). Inducing a concurrent motor load reduces categorization precision for facial expressions. *Journal of Experimental Psychology Human Perception & Performance*.

Kadosh, K. C., Luo, Q., de Burca, C., Sokunbi, M. O., Feng, J., Linden, D. E., & Lau, J. Y. (2016). Using real-time fmri to influence effective connectivity in the developing emotion regulation network. *NeuroImage*, *125*, 616–626.

Kokiopoulou, E., & Saad, Y. (2009). Enhanced graph-based dimensionality reduction with repulsion laplaceans. *Pattern Recognition*, *42*(11), 2392–2402.

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science*, *11*(5), 379–385.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91–630.

Ngo, T. T., Bellalij, M., & Saad, Y. (2012). The trace ratio optimization problem. *SIAM review*, *54*(3), 545–569.

Pell, P. J., & Richards, A. (2013). Overlapping facial expression representations are identity-dependent. *Vision research*, *79*, 1–7.

Rhodes, G., Jaquet, E., Jeffery, L., Evangelista, E., Keane, J., & Calder, A. J. (2011). Sex-specific norms code face identity. *Journal of Vision*, *11*(1), 1.

Rhodes, G., Pond, S., Burton, N., Kloth, N., Jeffery, L., Bell, J., ... Palermo, R. (2015). How distinct is the coding of face identity and expression? evidence for some common dimensions in face space. *Cognition*, *142*, 123–137.

Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *37*(6), 1113–1133.

Tranel, D., Damasio, A. R., & Damasio, H. (1988). Intact recognition of facial expression, gender, and age in patients with impaired recognition of face identity. *Neurology*, *38*(5), 690–690.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71–86.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, *43*(2), 161–204.

Vitale, J., Williams, M.-A., Johnston, B., & Boccignone, G. (2014). Affective facial expression processing via simulation: A probabilistic model. *Biologically Inspired Cognitive Architectures*, *10*, 30–41.

Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *29*(1), 40–51.

Fuse to be used: A weak cue's guide to attracting attention

Zara Harmon (zforough@uoregon.edu)
Department of Linguistics, 1290 University of Oregon
Eugene, OR 97401 USA

Vsevolod Kapatsinski (vkapatsi@uoregon.edu)
Department of Linguistics, 1290 University of Oregon
Eugene, OR 97401 USA

Abstract

Several studies examined cue competition in human learning by testing learners on a combination of conflicting cues rooting for different outcomes, with each cue perfectly predicting its outcome. A common result has been that learners faced with cue conflict choose the outcome associated with the rare cue (the Inverse Base Rate Effect, IBRE). Here, we investigate cue competition including IBRE with sentences containing cues to meanings in a visual world. We do not observe IBRE. Instead we find that position in the sentence strongly influences cue salience. Faced with conflict between an initial cue and a non-initial cue, learners choose the outcome associated with the initial cue, whether frequent or rare. However, a frequent configuration of non-initial cues that are not sufficiently salient on their own can overcome a competing salient initial cue rooting for a different meaning. This provides a possible explanation for certain recurring patterns in language change.

Keywords: Frequency; Inverse base rate effect; configural learning; Artificial language learning; cue salience; selective attention

Introduction

The inverse base rate effect was first documented by Medin and Edelson (1988) in a human contingency learning experiment. Medin and Edelson presented participants with pairs of symptoms co-occurring with diseases. One of the diseases was common (C), and one rare (R). One of the symptoms, the *imperfect predictor/cue* I, occurred whenever either disease was present. Another symptom was a perfect predictor of the common disease (PC), appearing whenever the common disease occurred and not otherwise, as shown in (1). The remaining symptom, PR, was a perfect predictor of the rare disease.

(1) I.PC →C
I.PR →R

Medin and Edelson's (1988) participants learned that PC predicts C, and PR predicts R. More interestingly, they inferred that the uninformative predictor I on its own predicts the common disease C, as does the combination of all symptoms, I.PC.PR. On the other hand, the combination of perfect predictors PC.PR was taken to predict the rare disease R. Two major accounts of these results have been proposed. According to Kruschke's (1996, 2001) explanation of the inverse base rate effect, participants learn about C first, acquiring moderate associations from both I and PC to C. The strengths of these associations are only moderate because they compete for associative strength. Participants then learn to predict R from I.PR, but at this point I predicts PC. I's presence in I.PR then forces the participants to assign high associative strength to PR→R in order to overcome the I→C association. PC→C is therefore weaker than PR→R, leading participants to respond with R when cued with PC.PR.

Juslin et al. (2001) instead propose that the IBRE can be explained by eliminative inference. Like Kruschke (1996), they maintain that participants learn well what predicts C but they learn little about R. PC.PR is then mapped onto R because it is unlike anything previously encountered, and nothing has been associated with R strongly: C is already associated with I.PC so it is reasoned not to *also* be associated with PC.PR, whereas R is unassociated. Therefore, C is eliminated from the options and PC.PR is associated with R. In response, Kruschke (2001) argues that eliminative inference does not explain part of the findings in the inverse base rate effect literature; namely, that I.PC.PR favors C less than I alone does, which is consistent with the PC.PR combination rooting for R.

Eliminative inference in form-meaning learning

In the present paper, we examine whether these effects are seen with learning of form-meaning mappings. Our interest in looking for an inverse base rate effect in this paradigm was motivated by results that suggested that eliminative inference plays a major role in form-meaning learning because forms and meanings are often assumed by learners to be in a one-to-one relationship (mutual exclusivity, Markman & Wachtel, 1989). In our own previous work, we exposed adult learners to a miniature artificial language in which the suffixes *–dan* and *–sil* were mapped onto the meaning [-dim(inutive);+pl(ural)] (multiple large creatures), while the suffixes *–nem* and *–shoon* were mapped onto the meaning [+dim;-pl] (one small creature). For some participants, the suffix *–nem* occurred more often than others in training; for other participants, the suffix *–dan* was more frequent than others.

Participants were asked to map words containing one of the four suffixes onto meanings. On each trial, they would

hear a word bearing one of the suffixes, and have to click on one of four pictures: either a picture of one large creature [-dim;-pl], multiple large creatures [-dim;+pl], one small creature [+dim;-pl] or, crucially, multiple small creatures [+dim;+pl]. This last meaning was novel, never presented in training. We found that, unlike rare suffixes, which were mapped onto the new meaning of [+dim;+pl] as often as the meaning they were paired with in training, the frequent suffixes were more likely to be mapped onto their familiar meanings than onto the new meaning.

Thus, when *–nem* was as frequent as *–shoon* both were equally likely to be mapped onto [-pl;+dim] and [+pl;+dim]. However, when the frequency of *–nem* increased, it became almost exclusively mapped onto [-pl;+dim]. Interestingly, when *–nem* was frequent, its competitor, *–shoon* was increasingly mapped onto [+pl;+dim]. Similarly, the frequent *–dan* pushed its rare competitor *–sil* out of the familiar meaning [-dim;+pl] into the novel meaning [+dim;+pl].

Our explanation for this finding is general-to-specific learning (Rogers & McClelland, 2004; Kapatsinski, 2013), also sometimes called *entrenchment* (Braine & Brooks, 1995), coupled with eliminative inference (Juslin et al., 2001; Ramscar et al., 2013). Participants start out with general, underspecified meanings: all suffixes could have any meaning at the onset of training. Gradually, the semantics of suffixes narrow. With some exposure, *–dan* becomes [+pl], and with extra exposure it becomes [+pl;-dim]. At this point, the learner is certain about the meaning of *–dan* but is less certain about the meaning of its competitor *–nem*. They therefore reason that 'if *–dan* means [+pl;-dim], *–sil* must mean something else'. There is only one meaning that has not been explicitly assigned to anything, [+pl;+dim], hence *–sil* comes to have that meaning. Interestingly, by being mapped onto the rare meaning, the rarer *–sil* finds a *niche*, in which it is somewhat protected from competition with highly frequent suffix *–dan*. This kind of semantic change has been documented in historical linguistics and described as a form *seeking* a niche (Aronoff, 1976). The Entrenchment+Elimination Hypothesis provides a non-teleological explanation for this phenomenon. It is possible that we see this effect at play with conflicting combinatorial cues that can be mapped onto more than one meaning.

Language forms as configural cues

While form-meaning learning seems to involve eliminative inference, it is different from classic work demonstrating the inverse base rate effect in that form-meaning learning is usually thought of as mapping *configurations* of cues onto meanings (Kapatsinski, 2009, 2013).

Kruschke (1996) assumes that the presented cue combinations are treated as sets of elemental cues. It is quite possible that instead an entire cue combination (I.PC or I.PR) forms a complex configural cue (e.g. Rescorla, 1973; Pearce, 1994). In that case, participants may form I.PC→C and I.PR→R associations. In the presence of such associations, it becomes an open question whether elemental I→C, PC→C, and PR→R associations form as well (cf. Pearce, 1994, vs. Rescorla, 1973).

Experiment 1

Participants

One hundred and eight native speakers of American English who were undergraduate students at the University of Oregon and were recruited from the Linguistics-Psychology Human Subjects pool participated for course credit. All had normal hearing and normal or normal-to-corrected vision and gave written consent.

The artificial language

The artificial language presented to the learners consisted of two constructions that were used to express the location of a creature relative to a table. The vocabulary in these languages consisted of one prepositional form, two postpositional forms and 25 nouns that were used to refer to different types of creatures. The prepositional form *bes* and postpositional form *zon* expressed the meanings of 'above' and 'below', respectively. The final and required form in each sentence of the language was the postpositional form *mik* that could be interpreted as referring to the 'table' or 'creature' or be treated as a grammatical morpheme such as English determiner *the.* The two resulting constructions are presented in (2).

(2) *bes* NOUN_i *mik* = CREATURE_i **ABOVE** TABLE
NOUN_i *zon mik* = CREATURE_i **BELOW** TABLE

We manipulated the frequency by which each of these constructions occurred in the language, creating the two conditions, BES and ZON. In each condition, the frequency of one of the two forms (*bes* or *zon* respectively) exceeded the frequency of the other by a factor of 3.

Tasks

The experiment consisted of an exposure stage (i.e., training without feedback) followed by a form-meaning mapping task and three questions probing the meaning of the forms.

Exposure Before exposure, participants were instructed that they are to learn two constructions expressing locations of creatures relative to a table in a miniature artificial language. Exposure stage proceeded as follows. Each trial consisted of the presentation of a picture of a creature above or below a table on the computer screen. Each picture was accompanied by a sentence that appeared at the bottom of the screen and described the picture using one of the constructions in (2). Picture background was black and sentence background was white. Each picture and its corresponding sentence appeared simultaneously and stayed on the screen for five seconds. Participants were instructed to use the cues in the sentence to predict the location of the creature.

Form→Meaning Mapping Task In this task, participants were asked to map a form onto one of two meanings by pressing the left and right arrow keys on the keyboard. At the beginning of each trial, the two target meanings of 'above' and 'below' were presented with pictures on the left and right side of the computer screen, respectively. At the same time, a sentence appeared at the bottom of the screen. Participants pressed the corresponding arrow to choose between the two meanings. As soon as they pressed a button, the experiment continued to the next trial.

Questions Once after training and again after judgment task, learners responded to three questions about the meanings of the forms *bes, zon* and *mik*: On each trial, 'What does *bes/zon/mik* mean?' appeared in the middle of the screen. At the same time a textbox appeared at the bottom of the screen for participants to type in their responses. The experiment advanced once the ENTER key was pressed.

Stimuli

Each participant experienced 80 exposure trials, with 20 trials allocated to the low-frequency construction and 60 trials allocated to the high-frequency one. Of the 25 nouns in the language, 20 appeared in both constructions during exposure. The form to meaning mapping test task consisted of 80 trials and 10 nouns, 5 of which were novel. Cue combinations presented to the learners in both tasks comprised 8 categories (10 items each): noun in isolation (N); isolated cues: *bes* N, N *zon,* N *mik*; cue combinations: *bes* N *mik,* N *zon mik*; and finally conflicting cue combinations: *bes* N *zon* and *bes* N *zon mik*. The experiment ended after the 3-trial question task about the meanings of *bes, zon,* and *mik*. Ninety-three participants also received the same questions immediately after training to determine if their beliefs about the form-meaning mappings changed during the test.

Procedure

Learners were tested one at a time. During training, the stimuli were presented on the computer screen and the subjects were asked to read aloud the sentences that accompanied each trial. Subjects did so while wearing a head-mounted microphone, although their speech was not recorded. The stimuli were presented using E-prime 2.0 Professional, which recoded the subjects' responses automatically during the test phase. The order of presentation of the stimuli was randomized separately for each learner.

Results

The data were analyzed using mixed effect logistic regression in R (lme4 package, Bates et al., 2015) with random intercepts for subjects and nouns, and random slopes for cue within both subjects and nouns. The binary dependent variable was meaning chosen: common ('above' in BES and 'below' in ZON) versus rare ('above' in ZON and 'below' in BES). These results are shown in Figure 1.

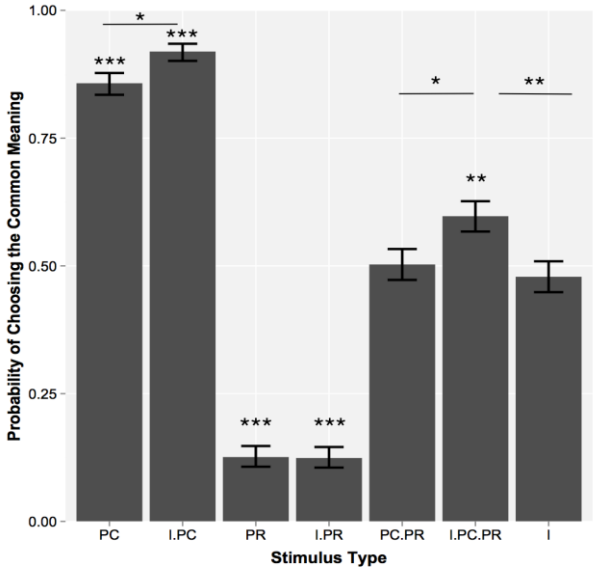


Figure 1: The probability of choosing the common predictor when presented with forms PC (*bes* N in BES and N *zon* in ZON), I.PC (PC + *mik*), PR (*bes* N in ZON and N *zon* in BES), I.PR (PR + *mik*), PC.PR (*bes* N *zon*), I.PC.PR (*bes* N *zon mik*), and I (N *mik*).

Participants showed a strong preference for mapping the common cue to the common meaning (z=-5.709, p<.001 for PC, z=-7.828, p<.001 for I.PC) as well as a strong preference for choosing the rare meaning in the presence of the rare cue (z=6.376, p<.001 for PR; z=7.221, p<.005 for I.PR). However, responses were at chance when both rare and common cues (PC.PR) were present (z=-0.141, p=.888) as well as when the only present cue was *mik* (z=0.507, p=.612) or no cues were present at all, i.e. the noun was presented in isolation (z=1.362, p=.173).

While *mik* in isolation was not associated with either of the two meanings, its presence together with PC in I.PC and I.PC.PR increased C responses compared to PC and PC.PR (for I.PC.PR vs. PC.PR, z=3.197, p=.001; for I.PC vs. PC, z=2.18, p=.029). There was no significant effect of adding I to PR (I.PR vs. PR, z=0.691, p=.49). In addition, the likelihood of mapping I.PC.PR onto C was also significantly higher (and not lower as in prior studies of the inverse base rate) than the likelihood of mapping I onto C (z=-1.993, p=.0462). Overall, these results suggest that I.PC behaved as a configural cue to C so that the combination I.PC elicited C better than you would expect from how well I and PC could elicit C on their own.

Note that the constructions in (2) place *zon* next to *mik* (I) while *bes* is separated from *mik* by the noun. Therefore, we might expect that *zon* may be more likely to fuse with *mik* into a larger configural chunk than *bes* is. If there is an effect of contiguity on cue chunking, we should see more

evidence of configural processing of I.PC in the ZON condition than in the BES condition.

Most of the results shown in Figure 1 hold across conditions. Thus, for example, I.PC is a better cue to C than PC alone is, whether PC is adjacent to I or not. Similarly, I is mapped onto C and R equally often in both conditions, suggesting some degree of configural processing of I.PC in both conditions.

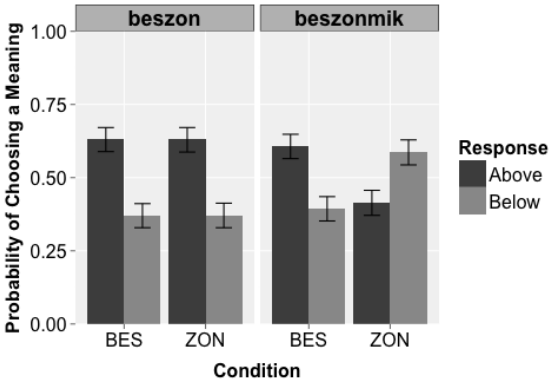


Figure 2: Participants' mapping of the two meanings of 'above' and 'below' to *bes* N *zon* (PC.PR) and *bes* N *zon mik* (I.PC.PR) in the two conditions BES and ZON.

However, as shown in Figure 2, the two conditions differ greatly in the effect of adding I to PC.PR (z=3.198, p=.001). There is no difference between PC.PR and I.PC.PR in the BES condition (z=-0.015, p=.99): participants' responses to both PC.PR and I.PC.PR are dominated by the PC *bes* (z=-2.756, p=.006), the cue that occurs first in the stimulus. In contrast, there is a large difference between PC.PR and I.PC.PR in the ZON condition (z=4.59, p<.0001): while the treatment of PC.PR is still dominated by *bes*, the initial cue (PR in this condition; z=-2.84, p=.005), I.PC.PR usually elicits the response appropriate for I.PC (*zon mik*). In other words, only *zon* and *mik* together were able to successfully compete with the early cue *bes*, and only when they were frequent in the input.

Experiment 2

The results of Experiment 1 suggested that participants mapped I.PC onto C, in addition to mapping PC onto C and PR onto R. In addition, in cue conflict situations, participants showed a bias to respond on the basis of the initial cue. Experiment 2 was designed to replicate these results without *mik*, the I stimulus. We were interested in whether *zon* may be able to override the influence of *bes* on its own in the ZON condition, now that it alone associates with 'below'.

The experiment was exactly the same as Experiment 1, except for the absence of *mik* in training and the absence of I, I.PC, I.PR, and I.PC.PR trials at test. On the closest equivalent to an I trial, a noun stem could appear on its own, without any other morphemes (e.g. after experiencing N *zon* ~ 'below' and *bes* N ~ 'above', the participants would be presented with N in isolation and with *bes* N *zon* at test). Twenty-nine additional participants (15 in BES and 14 in ZON) were recruited for this replication.

Figure 3 shows the results: As in Experiment 1, participants showed 50/50 guessing in the I (stem alone) condition (z=0.589, p=.56). They also continued to show reliance on the stimulus-initial cue on trials that involved conflicting cues (z=-2.061, p=.039). Interestingly, *zon* alone remained unable to override the influence of the initial *bes* (z=-0.167, p=.87), even though it was always final, rather than medial, during the training in this experiment, and there was no *mik* to distract attention from it, or to fuse with it. These results suggest that the inability of *zon* to overcome *bes* in Experiment 1 is not due to *zon* fusing with *mik* during training: participants continue to show reliance on the initial cue.

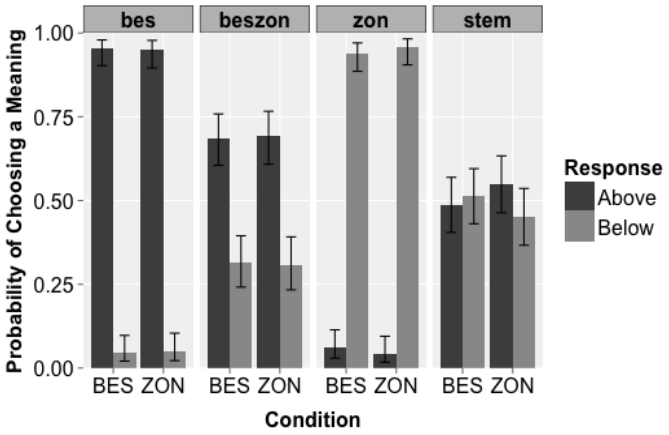


Figure 3: Participants' mapping of the two meanings of 'above' and 'below' to *bes* N, *bes* N *zon* (PC.PR), *zon* N and N stem alone (I) in the two conditions BES and ZON.

Discussion and Conclusion

A summary of our findings in shown in Table 1. These results differ substantially from much previous work on the inverse base rate effect and provide no support for the phenomenon.

Table 1: Participants' mapping of cue to meaning.

Stim/Cue type	Response
PC	C
PR	R
I (*mik*)	-
PC.PR	-
I.PC.PR (*bes zon mik*)	C
I.PC	C
I.PR	R

In particular, we fail to observe I→C and PC.PR→R (cf. Kruschke 1996, 2001). At the same time, as in previous studies, we do observe PC→C, PR→R and I.PC.PR→C. It

is worth noting that in previous studies the I.PC.PR→C association was weaker than PC.PR→R or I→C (Kruschke, 2001). Thus, our failure to find the latter two associations is unlikely to be a power issue. We now proceed to interpreting the findings.

First, the comparison of PC.PR and I.PC.PR detailed in Figure 2 suggests that the presence or absence of *mik* (I) adjacent to *zon* influences the meaning of the sentence. Without *mik*, *zon*, despite its higher frequency in ZON, is a weaker cue to its meaning than *bes* is. Thus, when *bes* and *zon* are placed in competition, *bes* wins, whether it is frequent or not. The stronger reliance on the initial cue is observed when the initial and postnominal cues are placed in conflict. Even when both cues are accurately mapped onto their meanings in the absence of conflict, the conflict situation reveals that one of the cues is stronger than the other (see also work within the Competition Model framework, e.g. MacWhinney et al., 1985, on similar findings from placing case and word order cues to agency in conflict).

The advantage of *bes* in this context is likely due to *bes*' position in the sentence (initial and prenominal). Early cues tend to be more important than later cues for word recognition due to incremental processing (e.g. Marslen-Wilson & Tyler, 1980). In addition, the meanings 'above' and 'below' are expressed by a preposition in English, thus *bes* is in the more expected location for an expression bearing this kind of spatial meaning. Given this, participants are likely to allocate selective attention to the beginning in searching for spatial expressions, which may make the initial cue more associable in this situation. Indeed, because of this bias, the subjects may learn the meaning of the initial, prenominal cue before they learn the meaning of the postnominal cue, allowing the initial cue to partially block other cues. If this happens, *bes* may block *mik* from associating with *bes*' meaning, 'above'.

Importantly, the decision about the meaning of a *bes*-initial sentence is influenced by later cues, just not as much as by the initial cue. The importance of the initial cue for this decision is therefore unlikely to be due to the participants failing to perceive other cues. Rather, the initial cue appears to develop a stronger association to its meanings than non-initial cues do or to result in an early commitment to an interpretation that is difficult to revise.To tease apart these possible contributing factors, future work should determine whether increased reliance on an initial cue during a test trial is due to its initial position during that test trial or its consistently initial position in training.

The presence of *mik* (I) in the context of *bes* N *zon* (PC.PR) reduced the likelihood of 'above' (R) responses in the ZON condition. This finding is consistent with *mik* being associated with the meaning 'below'. However, unlike in studies observing the IBRE (Kruschke, 1996), *mik* does not cue 'below' on its own: I is not associated with C, and I.PR does not show any evidence of cue competition. Thus, *mik* only cues 'below' in the presence of *zon*, suggesting that *zon* and *mik* form a configural cue.

Unlike PC.PR (*bes* N *zon*), I.PC.PR (*bes* N *zon mik*) does usually elicit the C response. This means that *zon mik* is able to compete with *bes* when it is more frequent than *bes.* Interestingly, Experiment 2 showed that the *mik*-less final *zon* is overpowered by the initial *bes* even if *zon* is the only cue to the meaning 'below'. The cue *zon* in Experiment 2 is as frequent as the cue combination *zon mik* in Experiment 1, and appears in the same position. Yet, a frequent *zon mik* can compete with *bes* but a frequent *zon* cannot. This finding suggests an influence of *cue salience*: the initial cue is much stronger than a later cue unless that cue is longer (hence, more salient).

Cue salience has previously been argued to be important for acquiring form-meaning mappings by MacWhinney et al. (1985) and Ellis & Sagarra (2011). The present study adds to this body of work by showing that salience matters under cue competition even when – in the absence of cue competition – the less salient cue can cue its meaning perfectly well on its own.

In the present study, *zon* is not lacking associations: when presented alone, it is almost always mapped onto 'below', whether it is frequent or not. This finding suggests that the emergence of the configural *zon mik* cue as a result of frequent *zon mik*→'below' exposures does not eliminate the *zon*→'below' associations: both the parts and the whole can become associated with outcomes (Rescorla, 1973). Frequent occurrence of I.PC→C causes both I.PC and PC to associate with C.

In contrast, the uninformative I (*mik*) is associated with neither C nor R. Responses to the questions about form meanings suggest a likely explanation. Because the sentences were paired with a visual world depicting the position of a creature relative to a table, most participants took *mik* (I) to mean 'table' or 'creature'. In fact, the proportion of participants who associated I with one of these referents (74% in BES, 70% in ZON) was identical to the proportion of respondents reporting that *bes* means 'above' and *zon* means 'below'. There was no difference in what *mik* was taken to mean across conditions. Nonetheless, *mik* behaved as if it reinforced *zon* in the form-meaning mapping task only when *zon* was the PC. This again suggests that it is the configuration I.PC and not I alone that was mapped onto C.

Given the reported meaning of *mik*, one might argue that the true cue-outcome structure presented to participants was more like (3) than like (1), allowing I to map onto A or B.

(3) I.PC →C A B
I.PR →R A B

Because of a strong preference for one-to-one form-meaning mappings in adults (giving rise to eliminative inference), the mapping of I onto A or B may have prevented it from *also* mapping onto C. This may be one reason we do not observe the IBRE in the present study. According to Kruschke (1996), the motivation for *strongly* associating PR with R is to counteract the influence of I→C.

Without such an influence, there is no motivation for PR→R to be stronger than PC→C.

An additional reason for not observing the IBRE in the present study may be that the difference in frequency between C and R was not as extreme as in some previous studies. In particular, Shanks (1992), was able to obtain the effect with a 7/1 ratio but not a 3/1 ratio (which we used here). Though Kruschke (1996) did find the effect with a 3/1 ratio, Kruschke's attentional explanation for the effect assumes that participants learn about C before they learn about R, despite examples of both being interspersed within a single training stage. This is more plausible with a more extreme frequency asymmetry. It may be that with our more moderate frequency ratio participants learn about both categories in parallel. If C is not learned before R, then there is little reason for I to become associated with C.[1]

At the start of this project, we thought that an I→C association may develop alongside, say, I→A because, in language change, a form that happens to frequently co-occur with another form seems to become associated with the meaning of that form over time. For example, in French, *pas* used to always mean 'step' but came to also mean 'not', by virtue of usually occurring in the construction *ne pas* 'not a step' (Bybee, 2003, 2015, p.126). Here, *pas* starts out as an imperfect cue (I) for 'not' but a perfect cue for 'step' (A), becomes associated with 'not' (C), and comes to subvert the erstwhile PC *ne*. How does this happen? One way would be for an I→C association (*pas*→'not') to develop despite the existence of an I→A association. Our results suggest that the I→C association does not easily develop alongside I→A. Rather, perhaps especially when PC is not very salient (like the unstressed *ne*), I.PC becomes associated with C. Thus, the development of the I→C (*pas*→'not') association may need to be preceded by the development of increasingly strong I.PC→C (*ne pas*→'not') followed by erosion of the low-salience *ne*. Indeed, the historical data appear consistent with this story: the use of *ne* with multiple nouns seems to have given way to use of *ne pas* for 'not' some time before *pas* could be used to mean 'not' without being accompanied by *ne* (Bybee, 2015, p.126). This historical development is thus one possible example of cues to meaning fusing together, and in so doing overcoming the otherwise low salience of the cue that perfectly predicts the meaning.

References

Aronoff, M. (1976). *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Bates, D., Maechler, M., Bolker, B, & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-9, URL: https://CRAN.R-project.org/package=lme4.

Braine, M. D. S., & Brooks, P. J. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In Tomasello, M., & Merriman, W. E. (Eds.), *Beyond names for things: Young children's acquisition of verbs* (pp. 353-376). Hillsdale, NJ: Erlbaum.

Bybee, J. (2015). *Language change*. Cambridge, UK: Cambridge University Press.

Bybee, J. (2003). Cognitive processes in grammaticalization. In Tomasello, M. (Ed.), *The new psychology of language, Vol.2*, 145-167. Mahwah, NJ: Erlbaum.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition. *Studies in Second Language Acquisition*, *33*, 589-624.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High-level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *JEP: LMC*, *27*, 849-871.

Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language*, *89*, 110-148.

Kapatsinski, V. (2009). Testing theories of linguistic constituency with configural learning: The case of the English syllable. *Language*, *85*, 248-277.

Kruschke, J. K. (2001). The inverse base-rate effect is not explained by eliminative inference. *JEP: LMC*, *27*, 1385-1400.

Kruschke, J. K. (1996). Base rates in category learning. *JEP: LMC*, *22*, 3-26.

MacWhinney, B., Pléh, C., & Bates, E. (1985). The development of sentence interpretation in Hungarian. *Cognitive Psychology*, *17*, 178-209.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121-157.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1-71.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *JEP: General*, *117*, 68-85.

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, *101*, 587-607.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, *24*(6), 1017-1023.

Rescorla, R. A. (1973). Evidence for "unique stimulus" account of configural conditioning. *Journal of Comparative and Physiological Psychology*, *85*(2), 331-338.

Rogers, T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science, 4*, 3-18.

[1] We should note that it is not the case that our frequency ratio is so small that participants do not learn which cues and outcomes are more frequent. Prior data from our lab and others have used 3/1 ratios and found that adult participants match the presented probabilities (e.g. Kapatsinski, 2013).

Stereotype-Based Intuitions: A Psycholinguistic Approach to Experimental Philosophy's 'Sources Project'

Eugen Fischer (E.Fischer@uea.ac.uk)
School of Politics, Philosophy, Language and Communication Studies, University of East Anglia
Norwich, NR4 7TJ, UK

Paul E. Engelhardt (p.engelhardt@uea.ac.uk)
School of Psychology, University of East Anglia
Norwich Research Park, Norwich, NR4 7TJ, UK

Abstract

Experimental philosophy's 'sources project' seeks to develop psychological explanations of philosophically relevant intuitions which help us assess their evidentiary value. This paper develops a psycholinguistic explanation of intuitions prompted by brief philosophical case-descriptions. For proof of concept, we target intuitions underlying a classic paradox about perception ('argument from hallucination'). We trace them to stereotype-driven inferences automatically executed in verb comprehension. We employ a forced-choice plausibility-ranking task to show that contextually inappropriate stereotypical inferences are made from less salient uses of the verb "to see". This yields a debunking explanation which resolves the philosophical paradox.

Keywords: Experimental philosophy; Sources Project; stereotype-driven inference; graded salience.

Introduction

Philosophical Research Context: This paper presents work that pioneers the use of psycholinguistic methods in experimental philosophy's 'Sources Project'. Experimental philosophy is a currently much-discussed and potentially transformative movement that imports methods from psychology into philosophy (for reviews see, Alexander, 2012; Knobe & Nichols, 2014). Philosophers frequently conduct thought experiments which revolve around the consideration of hypothetical cases. Intuitive judgments about such cases are elicited by verbal case-descriptions and play key roles in philosophy: A 'standard justificatory procedure' uses such intuitions as evidence for/against philosophical theories (Bealer, 1996; Cath, in press); clashes of intuitions with each other or background beliefs generate influential philosophical paradoxes and problems (Fischer, 2011; Papineau, 2009). Experimental philosophers employ surveys and experimental methods from psychology to elicit, explain, or assess such intuitions.

The strand of the movement that currently attracts most philosophical attention is the 'Warrant Project' (for a review see, Stich & Tobia, 2015). This project seeks to assess the evidentiary value of philosophically relevant intuitions and philosophers' warrant for accepting them. Up to now, the project mainly employs questionnaire-based surveys to study the sensitivity of intuitions to demographic parameters like gender, age, personality type, or cultural or socio-economic background (Weinberg et al., 2001), as well as questionnaire-based experiments to investigate order and framing effects (e.g., Petrinovich & O'Neill, 1996, Weinberg et al., 2012). The standard approach has experimental philosophers make inferences from such sensitivity or effects to lack of evidentiary value. Many of these inferences are philosophically problematic (for a review see, Fischer & Collins, 2015), and various studies have failed to replicate previous results about sensitivity to demographic factors (e.g., Nagel et al., 2013; Seyedsayamdost, 2015a,b). It is therefore timely to pursue the philosophically crucial aims of the Warrant Project with fresh approaches.

The most ambitious strand of the Warrant Project, known as 'Sources Project' (Pust, 2012) or 'cognitive epistemology' (Fischer et al., 2015), explores such approaches: It seeks to develop and experimentally test psychological explanations of intuitions that help us assess their evidentiary value. One promising approach is to look for explanations that trace intuitions back to largely automatic cognitive processes that are generally reliable but predictably engender cognitive illusions, under specific circumstances. The most prominent line of such research traces intuitive knowledge attributions back to a 'mind-reading' capacity subject to specific biases (Alexander et al., 2015; Gerken & Beebee, 2016; Turri, 2015).

Our research breaks new ground by examining whether philosophically relevant intuitions are generated or influenced by routine language processes, viz., by automatic inference processes that routinely occur in text comprehension and production. This approach seems worthwhile for the many philosophical relevant intuitions that are elicited by verbal descriptions of hypothetical cases. Elsewhere (Fischer et al., 2015; Fischer & Engelhardt 2016), we have argued that the routine process of stereotype-driven amplification (Levinson, 2000; cf. Garrett & Harnish, 2007) is generally reliable but engenders cognitive illusions, under specific vitiating circumstances. This paper explores the role of this process in generating intuitions at the root of influential philosophical paradoxes about perception, known as 'arguments from hallucination'. We show that vitiating circumstances obtain in their formulation. This finding debunks the intuitions and helps resolve the paradoxes.

Philosophical Application: Together with 'arguments from illusion', arguments from hallucination engender the classic 'problem of perception' (Smith, 2002), which has again become a focus of debate (Brewer, 2011; Crane, 2015; Fish, 2009; Robinson, 2001). Both arguments lead to the conclusion that when people use their five senses, they are (directly) aware only of subjective perceptions or sense-data. The philosophical problem is that of reconciling this conclusion with the common-sense conviction that we see and otherwise perceive physical objects and public events.

Philosophers typically formulate arguments from hallucination without world-knowledge about the phenomenon, assume merely that it is *possible* that 'one 'perceives' a physical object which is not there at all' (Ayer, 1956/1990, pg. 90), and proceed from brief descriptions of hypothetical cases. Analytic philosophers formulating the argument typically distinguish between different senses of perception-verbs like *see*: The relevant case-descriptions, they explain, use these verbs in a purely 'phenomenal' sense which serves merely to describe people's experiences and which lacks the existential and spatial implications *see*, etc. ordinarily have. Here is a classic statement:

> 'Let us take as an example Macbeth's visionary dagger: since we are concerned only with what is possible, the fact that this episode may be fictitious does not matter. There is an obvious [ordinary] sense in which Macbeth did not see the dagger; he did not see the dagger for the sufficient reason that there was no dagger there for him to see. There is another [viz., phenomenal] sense, however, in which it may quite properly be said that he did see a dagger; to say that he saw a dagger is quite a natural way of describing his experience. But still not a real dagger; not a physical object; not even the look of a physical object, if looks are open to all to see. If we are to say that he saw anything, it must have been something that was accessible to him alone, something that existed only so long as this experience lasted; in short a sense-datum.' (Ayer, 1956, pg. 90)

The second half of the argument then postulates that in this possible scenario the subject's experience is qualitatively indistinguishable from the experience of seeing a physical object; it assumes that qualitatively indistinguishable experiences involve awareness of the same kind of object; and it concludes that all cases of perception involve awareness of sense-data. This second half has been widely criticized (for a review see, Smith, 2002).

Already the argument's first half, however, involves a curious mistake: The verb *see* is explicitly used here in a phenomenal sense: *He saw a dagger* is to mean 'He had an experience as of / like that of / seeing a dagger'. An experience is being described by comparing it to that of seeing a certain physical object. This does not require that the object 'seen' be around, when the subject has the experience thus described. So we can say that Macbeth saw a physical dagger if his experience 'is like that of seeing a solid, physical dagger' (rather than, say, like that of seeing a strangely translucent dagger-image). The case-description explicitly makes this postulate (Ayer 1956, pg. 90), and the second half of the argument crucially depends upon it (above). Of course, only Macbeth 'sees' the dagger at this point, and he 'sees' it only in the phenomenal sense. But what he 'sees' in this sense is still 'a real dagger' – e.g., the very dagger now sticking in the king's corpse next door. What is 'special' is the sense in which the argument uses the verb *see*, but not the object 'seen'. So why did generations of competent philosophers (e.g., Price 1932, pg. 28-9; Ayer 1956, pg. 90; Smith 2002, pg. 194-5), including philosophers who stress that the use of *see* and its cognates is special (viz., phenomenal), conclude that the object seen is special (namely, non-physical)?

One key factor, we submit, is a spontaneous inference from 'Macbeth sees a dagger' to 'there is something for Macbeth to see in his vicinity, before his eyes'. This *Intuition (I)* has thinkers take for granted that the 'something' must be around for Macbeth to see, disregard that Macbeth is (in the phenomenal sense) aware of a real dagger (which currently is not in his physical vicinity), and think that the only way of avoiding the 'preposterous … claim that a hallucinating person is aware of *nothing whatever*' (Smith, 2002, 195) is to 'recognize a non-normal object of awareness' (ibid.), which is in the subject's range of vision, before his eyes: Thinking in terms of an intuitive dichotomy between external and internal perception (introspection), proponents of the argument immediately conclude that the thing seen must be before the subject's inner eye, in his mind. I.e.: We submit the argument rests on intuition (I), and (I), in turn, on inappropriate spatial inferences from the phenomenal use of *see*.

Psycholinguistic explanation: Both nouns (Hare et al., 2009) and verbs (Ferretti et al., 2001; Harmon-Vukic et al., 2009) are associated with stereotypes. Verbs can be associated with typical features of events, agents, and patients, which jointly form structured stereotypes (a.k.a. 'generalized situation schemas'). E.g., manipulators are typically cunning and shrewd, and their victims naïve and gullible, so the verb *manipulate* is associated with 'cunning' and 'shrewdness' as agent-properties, and 'naiveté' and 'gullibility' as patient-properties. Such associations guide spontaneous inferences from those expressions. Within a neo-Gricean framework, their interplay with integration processes (see below) is captured by the I-heuristic (Levinson, 2000; cf. Garrett & Harnish, 2007): In the absence of explicit indications to the contrary, infer that situations talked about conform to stereotypes associated with the nouns and verbs used!

Stereotypical inferences are supported by automatic activation processes in semantic memory (McRae & Jones, 2013). According to the well-supported graded salience hypothesis (Fein et al., 2015; Giora, 2003), a linguistic stimulus activates all semantic and stereotypical features

associated with the expression, in any of its uses or senses. The speed and strength of initial activation depends upon the 'salience' of the sense or use. This is a function of familiarity, conventionality, frequency, and prototypicality. Features associated with the expression's most salient use are activated regardless of context. E.g., the ambiguous stimulus *mint* activates the probe 'candy' rapidly and strongly, even where it is clearly used in a less frequent sense (prime: *All buildings collapsed except the mint*) (Simpson & Burgess, 1985; Till et al., 1988). This facilitates contextually inappropriate inferences from less salient uses of words. Such inferences need not go through, since processes including reinforcement and decay (Oden & Spira, 1983), and suppression (Faust & Gernsbacher 1996) mitigate initial preferential activation of contextually inappropriate stereotypical associates, namely, in the light of contextual cues, explicit indications of deviation from relevant stereotypes, and explicit marking of less salient uses (Givoni et al., 2013). Where one use is very much more salient than another, and the stereotypical association with certain features particularly strong, inappropriate stereotypical inferences may, however, go through, even in the face of explicit indications to the contrary (Giora 2003).

We hypothesize that this happens in the argument from hallucination. The verb *see* is used in at least three senses: an ordinary, literal, visual sense (*Bob saw the builders fixing the road*), a metaphorical, epistemic sense (*Jane saw Joe's point*), and an extended phenomenal sense (*Hitting his head, Jack saw stars*). According to our first hypothesis,

H_1 The literal, visual sense of *see* is much more salient than the metaphorical, epistemic sense; this, in turn, is much more salient than the phenomenal sense.

This non-salient sense is employed in the case-descriptions that serve as first premises of the argument from hallucination. We hypothesize further that proponents of the argument then make contextually inappropriate stereotype-driven inferences from those premises: We assume that the most salient visual sense of *see* has a strong stereotypical association with spatial patient-properties: *S sees X* is strongly associated with 'X is in front of S' and 'X is before S's eyes' (*directional implications*) as well as with 'X is around (S) to be seen' and 'X is within S's range of vision' (*proximity implications*). On this basis we hypothesize that

H_2 Competent speakers infer spatial patient-properties stereotypically associated with the visual sense of *see* also from less salient epistemic and phenomenal uses.

Such a stereotypical inference, we submit, leads proponents of the argument from hallucination from such initial premises as 'When hallucinating, Macbeth sees a dagger' to the intuitive judgment (I) that there is a dagger around for Macbeth to see, before his eyes (see above). H_1 and H_2 could jointly explain the intuition (I) we identified as the intuitive source of the argument from hallucination.

In following up H_1, we bear in mind that salience is a function of familiarity, conventionality, frequency, and prototypicality. We think it unlikely that the visual and epistemic uses of *see* will attract different familiarity- or conventionality-ratings, and suggest salience differences are due to differences in frequency and prototypicality. A corpus study to establish frequency is still ongoing. First evidence for prototypicality differences is provided by a production study. 13 participants were asked to provide up to 10 written completions of 4 sentence stems containing either *see* or *aware* (e.g. Jane sees____; Bob saw____). To discourage mono-topical response strategies, participants were instructed to 'try to give varied responses'. We coded responses as either perceptual (e.g. the handsome American) or non-perceptual (e.g. fresh opportunities). Results showed that completions of *see* were 94% perceptual use (by contrast, 51% for *aware*). This suggests that visual cases of 'seeing' are by far the most prototypical of this category, and that the literal, visual sense of *see* is by far the most salient. Non-perceptual completions involved mainly epistemic uses. Phenomenal uses were absent. This is consistent with H_1. We then used a forced-choice plausibility-ranking task to examine H_2 and, specifically, explore spatial (directional and proximal) inferences from metaphorical/epistemic uses of *see*.

Methods

Participants

Fifty undergraduate psychology students from the University of East Anglia were recruited through the participant pool and received course credit for participating.

Materials

We administered a paper-and-pencil questionnaire that had 120 minimal pairs. There were 48 critical items and 72 fillers. Twenty-four critical items contrasted *see* and *aware.* (*Aware* can be used in both perceptual and non-perceptual contexts and lacks directional implications in either.) The other 24 critical items contrasted *see* and *think of.* (*Think of* lacks proximity implications in all uses). The see-aware items consisted of single sentences, the see-think of items consisted of pairs of sentences (see Tables 1 and 2).

Table 1: Example items for testing spatial inferences and specifically, directional inferences (e.g. *in front of*).

Visual, Stereotype-Consistent
1a. Mona sees the drivers ahead of her in the queue.
1b. Mona is aware of the drivers ahead of her in the queue.
Visual, Stereotype-Inconsistent
2a. Ben sees the friend walking right behind him.
2b. Ben is aware of the friend walking right behind him.
Epistemic, Stereotype-Consistent
3a. Emma sees the challenges facing her.
3b. Emma is aware of the challenges facing her.

Epistemic, Stereotype-Inconsistent
4a. Jim sees the scheming going on behind his back.
4b. Jim is aware of the scheming going on behind his back.

Table 2: Example items for testing proximity inferences (i.e. *around to be see, within the range of vision*).

Visual, Stereotype-Consistent
1. Megan spent the week at home. She saw/thought of her parents a lot.
Visual, Stereotype-Inconsistent
2. Sitting in the office, John sees/thinks of Peter. Peter is on holiday in Turkey.
Epistemic, Stereotype-Consistent
3. Joan sees/thinks of fresh opportunities. Some opportunities are around.
Epistemic, Stereotype-Inconsistent
4. Joe sees/thinks of ways to solve the problem. There are no solutions around.

There were two manipulations. The first concerns the use of the verb *see* (literal/visual vs. metaphorical/epistemic). More specifically, it turns on whether the direct object of the main verb is a visible physical object (e.g. drivers, friends, etc., in the literal/visual condition) or an abstract object (e.g. challenges, scheming, etc., in the metaphorical/epistemic condition). The second variable manipulated was whether the remainder of the sentence was either consistent or inconsistent with the hypothesized stereotypical inferences, viz. directional and proximity inferences: In the stereotype-consistent condition, the patient of *see* and *aware* is said or clearly implied to be in front of the agent or before the agent's eyes; in the stereotype-inconsistent condition, the patient is placed behind the agent. Similarly, the patient of *see* and *think of* was placed in the vicinity or visual field of the agent, in the stereotype-consistent condition, and outside this field, in the stereotype-inconsistent condition. The critical verb in the *see/think of* items could occur in either first or second sentence (see Table 2).

Design and Procedure

The design of the study was 2 × 2 (visual/epistemic × s-consistent/s-inconsistent). The visual/epistemic variable refers to the direct object (visual vs. abstract) of the critical verb. *S-consistent/s-inconsistent* refers to the 'consistency/inconsistency with spatial (directional and proximity) implications from literal/visual use of *see*' and was contained in further contextual information in the sentence. Separate repeated-measures ANOVAs (with follow-up *t*-tests) were run on see/aware and see/think of separately. In line with our hypothesis H_2, we expected *see* to be preferred with visual direct objects in s-consistent contexts, and *aware* and *think of* to be preferred with both visual and abstract direct objects in s-inconsistent contexts. The critical condition occurred when the direct object was abstract and invoked epistemic sense and the context was s-inconsistent with directional and proximity inferences. For graphical purposes, we coded *see* responses as 1 and *aware* and *think of* responses as 0. Thus, high bars indicate greater preference for *see*.

Participants were presented with pairs of sentences or short two-sentence pairs, which differ only in one critical word (i.e. minimal pairs). Participants were asked to indicate which of the two versions they thought was more plausible, and to make a judgment even if they did not have a clear preference. By noting how often participants preferred one version over the other, we can measure the consistency of plausibility judgments, and which of the two verbs have stronger stereotypical associations with the targeted spatial properties

Results

See vs. Aware

Results showed that both main effects were significant (visual/epistemic $F(1,48) = 171.75$, $p < .001$ and s-consistent/s-inconsistent $F(1,48) = 662.78$, $p < .001$). The interaction was also significant $F(1,48) = 226.76$, $p < .001$ (see Figure 1). With visual direct objects, *see* was preferred over *aware* 96% of the time, when contexts were s-consistent, but only 5% of the time when contexts were s-inconsistent. With epistemic direct objects, *see* was preferred 44% of the time when contexts were s-consistent but only 8% of the time when contexts were s-inconsistent. The comparison of s-consistent and s-inconsistent was significant for both visual ($t(48) = 45.35$, $p < .001$) and epistemic ($t(48) = 9.26$, $p < .001$) senses. The preference for *aware* in the epistemic/s-consistent condition was not significantly different from chance $t(48) = 1.90$, $p = .06$.

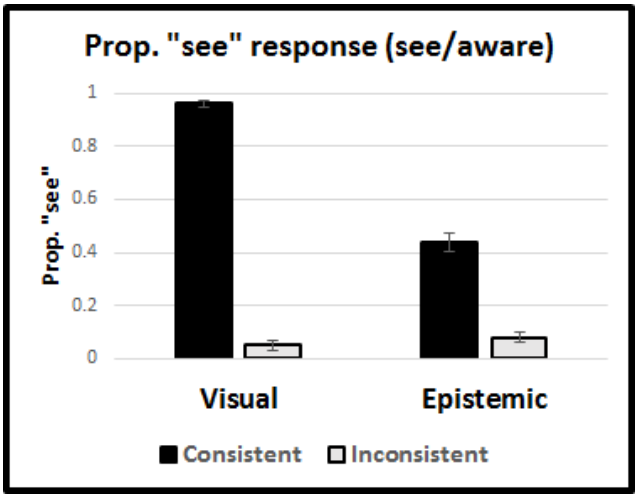


Figure 1: Results showing proportion of "see" responses with visual and epistemic objects.

See vs. Think of

Results showed that both main effects were significant (visual/epistemic $F(1,48) = 92.93$, $p < .001$ and s-consistent/s-inconsistent $F(1,48) = 817.33$, $p < .001$). The interaction was also significant $F(1,48) = 93.53$, $p < .001$

(see Figure 2). With visual objects, *see* was preferred over *think of* 95% of the time when contexts were s-consistent, but only 2% of the time when contexts were s-inconsistent. With epistemic objects, *see* was preferred 58% of the time when contexts were s-consistent, but only 5% of the time when contexts were s-inconsistent. The comparison of s-consistent vs s-inconsistent was significant for both visual $t(48) = 43.83$, $p < .001$ and epistemic $t(48) = 12.75$, $p < .001$ senses. The preference for *see* in the epistemic/s-inconsistent condition was significantly above chance $t(48) = 1.99$, $p = .053$. Thus, there was a marginal *see* preference.

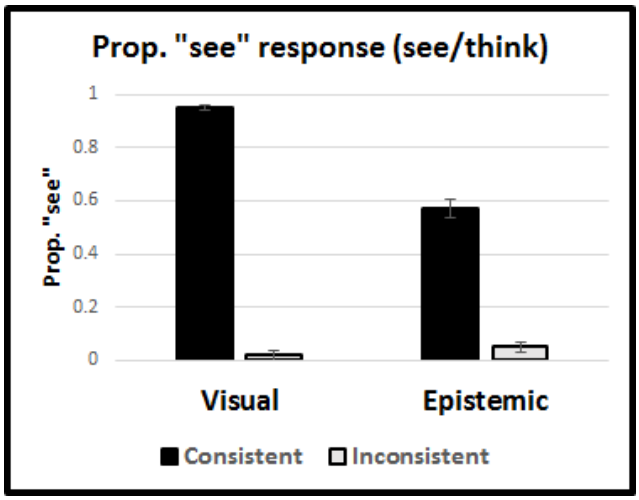


Figure 2: Results showing proportion of "see" responses with visual and epistemic objects.

Discussion

These results are consistent with our key hypothesis H_2 that competent speakers infer spatial patient-properties stereotypically associated with the visual sense of *see* also from epistemic and phenomenal uses. H_2 implies that in epistemic contexts which are s-inconsistent, competent speakers will reject *see* in favor of verbs that work equally well in epistemic contexts but lack spatial implications. This is precisely what we found: As long as items were s-consistent, participants had no pronounced preference for *aware* or *think of* over the evidently metaphorical use of *see* – all three verbs were deemed to work roughly equally well in these contexts. By contrast, participants had a very pronounced preference for *aware* and *think of* in epistemic contexts which were *in*consistent with the (evidently irrelevant) spatial implications of *see*. This suggests that participants made spatial inferences from the verb even in these inappropriate contexts, resulting in perceived inconsistencies with the context, which in turn lower subjective plausibility.

In conjunction with previous evidence that (H_1) visual uses of *see* have higher salience than epistemic and phenomenal senses, this is also consistent with the graded salience hypothesis. This hypothesis could explain our key finding (see Introduction) and let us extend it: Speakers/ hearers are the more likely to infer features stereotypically associated with the most salient use of a word also from its less salient uses, the greater the difference in salience is (Simpson & Burgess, 1985). Since the phenomenal use of *see* is even less salient than the epistemic use (see Introduction), it stands to reason that competent speakers will be even more prone to make inappropriate spatial inferences from phenomenal than from epistemic uses.

This conclusion can help resolve the targeted philosophical paradox (argument from hallucination), by debunking the intuition at its root, viz., that when Macbeth sees the dagger, there is something around for him to see, before his eyes. Our conclusion suggests that stereotype-driven inferences from the most salient visual use of *see* have philosophers leap to this intuitive judgment from prior case-descriptions which use *see* in a phenomenal sense. The intuition is thus due to a contextually inappropriate stereotypical inference. More generally, we have identified a vitiating circumstance: The generally reliable process of stereotype-driven amplification (Levinson, 2000) misfires where a word with strong stereotypical associates of a highly salient use is employed in a much less salient sense.

Future directions: We intend to garner further evidence for H_1 through a computer-based rating experiment asking participants to rate 'how good examples of *seeing* are afforded by the following cases', whose descriptions include the three uses distinguished. To collect further evidence for H_2 we intend to replicate the current results with a less artificial task, using pupillometry (Engelhardt et al., 2010). A more comprehensive explanation of the targeted intuitions will take further linguistic theories into account, such as head-driven phrase structure grammar (Pollard & Sag, 1994): When accessing a verb, we retrieve all associated syntactic information, including the grammatical roles (e.g., *see* requires a patient-role) and selection-restrictions on role-fillers (e.g., physical objects and public events). This may account for the intuition that Macbeth sees *something*. This paper explains how this 'something' gets endowed with spatial features – which exclude any absent physical objects as patients of *see*. The situation from which the present intuition arises is not rare in philosophy: Philosophers often give familiar words special uses which are less salient than those in ordinary language. It bears investigating to what extent these uses prompt contextually inappropriate inferences and unwarranted intuitions.

Conclusion

Psycholinguistic methods and findings can be profitably used to identify and explain automatic inferences that generate philosophically relevant intuitions. For proof of concept, this paper showed how psycholinguistic data can contribute to debunking explanations of intuitions that engender philosophical paradoxes.

References

Alexander, J. (2012). *Experimental Philosophy*. Cambridge: Polity.

Alexander, J., Gonnerman, C., & Waterman, J. (2015). Salience, and epistemic egocentrism. In J. Beebe (ed.), *Advances in Experimental Epistemology*. London: Bloomsbury.

Ayer, A.J. (1956). *The Problem of Knowledge*. Repr. 1990. London: Penguin.

Bealer, G. (1996). On the possibility of philosophical knowledge, *Philosophical Perspectives 10, Metaphysics*, 1-34.

Brewer, B.(2011). *Perception and Its Objects*. Oxford: OUP

Cath, Y. (in press). Reflective Equilibrium. In H. Cappelen und T. Gendler (eds.): *The Oxford Handbook of Philosophical Methodology*. Oxford: OUP.

Crane, T. (2015). The problem of perception. In N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Summer 2015.

Engelhardt, P.E., Ferreira, F., & Patsenko, E.G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology, 63*, 639-645.

Faust, M., & Gernsbacher, M.A. (1996). Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. *Brain and Language, 53*, 234-259.

Fein, O., Yeari, M., & Giora, R. (2015). On the priority of salience-based interpretations: the case of sarcastic irony. *Intercultural Pragmatics*, 12, 1-32.

Ferretti, T., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*, 516-547.

Fischer, E. (2011). *Philosophical Delusion and its Therapy*. New York: Routledge.

Fischer, E., & Collins, J. (2015) (eds.): *Experimental Philosophy, Rationalism and Naturalism. Rethinking Philosophical Method.* London: Routledge.

Fischer, E., & Engelhardt, P. E. (2016). Intuitions' linguistic sources: Stereotypes, intuitions, and illusions. *Mind & Language*, 31, 65-101.

Fischer, E., Engelhardt, P.E., & Herbelot, A. (2015). Intuitions and illusions. From explanation and experiment to assessment. In: Fischer and Collins (2015), pp. 259-292.

Fish, W. (2009). *Perception, Hallucination, and Illusion*. Oxford: Oxford University Press

Garrett, M., & Harnish, R.M. (2007). Experimental pragmatics: testing for implicatures. *Pragmatics and Cognition*, 17, 245-262

Gerken, M., & Beebe, J. (2016). Knowledge in and out of contrast. *Nous*, 50, 133-164.

Giora, R. (2003). *On Our Mind. Salience, Context, and Figurative Language.* Oxford: OUP.

Givoni, S., Giora, R., & Bergerbest, D. (2013). How speakers alert addressees to multiple meanings, *Journal of Pragmatics, 48*, 29-40.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009) Activating event knowledge. *Cognition*, 111, 151-167.

Harmon-Vukić, M., Guéraud, S., Lassonde, K.A. & O'Brien, E.J. (2009). The activation and instantiation of instrumental inferences.*Discourse Processes,46*,467-490

Knobe, J., & Nichols, S. (eds.) (2014). *Experimental Philosophy: Volume 2*. New York: OUP.

Levinson, S.C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*, Cambridge, Mass.: MIT Press.

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (ed.), *Oxford Handbook of Cognitive Psychology*, Oxford: OUP.

Nagel, J., Juan, V. S., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, 129, 652-661.

Oden, G.G., & Spira, J.L. (1983). Influence of context on the activation and selection of ambiguous word senses. *Quarterly Journal of Experimental Psychology 35*, 51-64.

Papineau, D. (2009). The poverty of analysis. *Aristotelian Society Supplementary Volume*, 83, 1-30.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17, 145-171.

Pollard, C., & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Price, H.H. (1932). *Perception*. 2nd ed. London: Methuen.

Pust, J. (2012). Intuition. In: E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/win2012/entries/intuition/

Robinson, H. (2001). *Perception*. London: Routledge.

Seyedsayamdost, H. (2015a). On Normativity and Epistemic Intuitions: Failure of Replication. *Episteme, 12*, 95-116.

Seyedsayamdost, H. (2015b). On gender and philosophical intuition: Failure of replication and other negative results *Review of Philosophy and Psychology*, *28*, 642-673.

Simpson, G.B., & Burgess, C. (1985). Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 28-39

Smith, A.D. 2002: *The Problem of Perception*. Cambridge, Mass: Harvard UP

Stich, S., & Tobia, K. (2015). Experimental philosophy and the philosophical tradition. In J. Sytsma and W. Buckwalter (eds.), *Blackwell Companion to Experimental Philosophy*. Wiley Blackwell: Malden.

Till, R.E., Mross, E.F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Journal of Verbal Learning and Verbal Behaviour, 16*, 283-298.

Turri, J. (2015). Skeptical appeal: the source-content bias. *Cognitive Science*, *39*, 307-24

Weinberg, J. S., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics,29*,429.

Weinberg, J. M., Alexander, J., Gonnerman, C., & Reuter, S. (2012). Restrictionism and reflection. *The Monist*, *95*, 200-222.

Mechanisms for storing and accessing event representations in episodic memory, and their expression in language: a neural network model

Martin Takac (takac@cs.otago.ac.nz)[1,2], **Alistair Knott (alik@cs.otago.ac.nz)**[1]
Dept of Computer Science, University of Otago, New Zealand[1], Centre for Cognitive Science, Comenius University, Bratislava[2]

Abstract

We present a neural network model of how events are stored in and retrieved from episodic long-term memory (LTM). The model is novel in giving an explicit account of the working memory (WM) medium mediating access to episodic memory: it makes a specific proposal about how representations of events and situations in WM interface with representations of events and situations in episodic memory. It also provides the framework for an account of how operations accessing temporally remote situations are reported in language.

Keywords: episodic memory; working memory; discourse models, neural networks

Introduction

In this paper, we present a new neural network model of episodic long-term memory (LTM). Like all computational models in cognitive science, its purpose is to make sense of a large body of experimental data, to provide a framework within which some of the questions raised by this data can be resolved. The influential network models of episodic memory have all done this in different ways. Moll and Miikkulainen (1997) clarified how episodic memories can be stored as pointers to first-order sensorimotor (SM) representations; Howard and Kahana (2002) showed how a recurrent network can account for the sequential structure of episodic memories; Norman and O'Reilly (2003) gave insights into the distinct roles of the hippocampus and cortex in storage and consolidation of episodic memories; Rolls and colleagues (e.g. Kesner and Rolls, 2015) hypothesised roles for specific circuits within the hippocampus and cortex. Our model, which builds on these earlier models, is intended to address two recent questions in the experimental literature.

One question concerns the relationship between episodic LTM and **working memory** (WM). There is good evidence that material to be stored in episodic memory is first maintained in WM (see e.g. Baddeley, 2000), as are queries to episodic memory, and the responses they retrieve (Fletcher and Henson, 2001). But there is debate as to whether material in WM occupies a *dedicated* neural medium (Baddeley, 2000; Shivde and Anderson, 2011), or whether the contents of WM are simply those components of the LTM system that are currently activated or attended to (Cowan, 1999; D'Esposito, 2007). There is strong experimental evidence for both positions; a model is needed to reconcile the two competing conceptions of WM.

The other question concerns the relationship between episodic memory and **language**. Episodic memory does not depend on language, but language is by far the most common medium for expressing its contents (Suddendorf and Corballis, 2007). For many recent theorists (e.g. Zwaan, 2008), the semantics of linguistic expressions should ultimately be given in terms of a model of episodic memory, as components of, or contributions to, episodic memory structures. Formal accounts of discourse structure have many striking similarities with models of episodic memory (van Lambalgen and Hamm, 2005): in particular they employ the notion of **discourse contexts**, that obtain at **reference times**. The context and reference time can be updated incrementally, or set to an arbitrary point in the past or future, by temporal adverbials (e.g. *in the afternoon*) or temporal subordinators (e.g. *when John arrrived*). However, there is no neural network model of episodic memory that provides a platform for investigating these similarities. What is needed is a model of how episodic memory interfaces with linguistic mechanisms, and in particular with linguistic devices for manipulating the context and reference time, so we can ask whether episodic memory provides a substrate for any components of the linguistic system.

Our model focusses on the storage of *events* in episodic memory. An event is a sentence-sized semantic unit centred around an action and its participants: we allow for an AGENT and optionally a PATIENT. Events can be represented from several different perspectives, which are conveyed linguistically by different aspectual types: for instance progressive (*John is arriving*) or perfective (*John has arrived*). We focus on the storage of events observed as wholes, which take time to occur and have a determinate endpoint (e.g. *John arrived*). We call these events *episodes*.

The main novelties of our model are in the way the LTM system interfaces with WM representations and with language. In the next section we describe the architecture of the model, considering these two issues in turn.

Architecture of the model

The WM system and its interface with LTM Our model of episode representations in LTM extends a model of episode representations in WM that is described in a companion paper (Takac and Knott, this volume). The WM model makes two important assumptions about how episodes in the world are *perceived*. The first of these is that experiencing an episode in the world involves a *well-defined sequence of discrete SM operations*: first attention to the agent, then attention to the patient (if there is one), then activation of a motor schema. Evidence for this is summarised in Knott (2012). Given this assumption, we propose that episodes are represented in WM as *prepared sequences of SM operations*. On this proposal, the WM representation of an episode is an executable structure, that allows the experience of an episode to be relived, or replayed. Our second assumption is that each individual par-

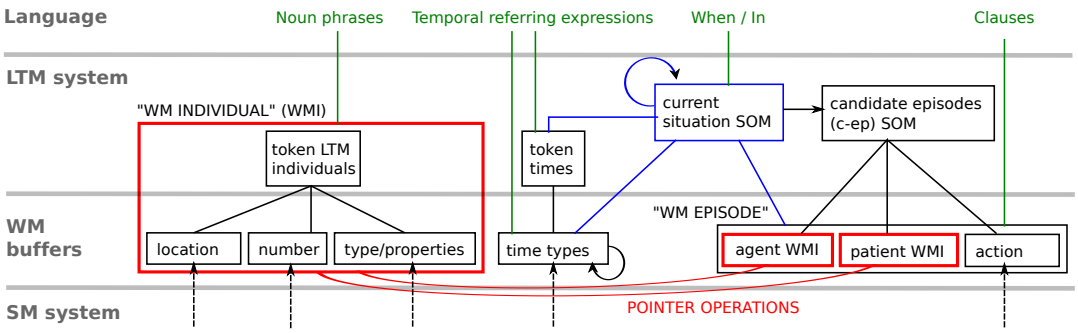


Figure 1: Architecture of the model of LTM and its interface with WM and language

ticipant in an episode is also perceived in a discrete sequence of SM operations: (i) attention to a spatial location, (ii) establishment of a cardinality (singular or plural), and (iii) identification of the object's type and intrinsic properties (Walles *et al.*, 2014). In our model, individuals are also represented in WM as prepared, replayable sequences of SM operations.

The architecture of our network is shown in Figure 1. We first focus on representations of individuals in WM and LTM. The prepared SM sequence associated with an individual is stored as a sustained pattern of activity in WM media holding location, number and type/properties (see the bottom left of the figure). LTM representations of token individuals are held in a separate medium, using a convergence-zone scheme (Moll and Miikkulainen, 1997). A **LTM individual** is represented as a sparse distributed assembly of units holding associations between location, number and type/properties, stored in long-term synaptic weights.

We now consider how episode representations make reference to individuals. Crucially, this reference must link individuals to roles such as AGENT and PATIENT. In our system, this role-binding is done in a WM buffer holding episode representations: the **WM episode** buffer (see the bottom right of Figure 1). Within this buffer, we posit separate fields for each role: the AGENT and PATIENT fields of a WM episode each hold a rich system of *pointers* back to the memory media representing an individual.[1] During experience of an episode, these pointers are initialised at different times: the AGENT representation is created when the agent is attended to, and later, the PATIENT representation is created when the patient is attended to. A complete WM episode representation is also an executable structure, that can be replayed. This replay process can provide top-down expectations for the different sequential stages of episode perception (see Takac and Knott, this volume). It is also involved in the interface to language, as we describe later.

Importantly, the AGENT and PATIENT fields of a WM episode point to *a mixture of WM and LTM representations*. When an individual is attended to, it is encoded by a pattern of activity across both WM and LTM media: the WM medium represents it as a type, and the LTM medium represents it as a token. We call this composite pattern a **WM individual**. The AGENT and PATIENT fields, which occupy a dedicated WM medium, each hold, and can recreate, a pattern in this composite WM/LTM area. Note this method of representing episode participants thoroughly blends the two competing conceptions of WM representations (as active LTM representations and as patterns in a dedicated WM medium): we see them as complementary, rather than as alternatives.

The three fields of the WM episode medium provide input to a LTM medium representing episodes, the **candidate episodes** (**c-ep**) medium. This is a self-organising map or **SOM**, whose units hold localist representations of particular episodes or episode types. When trained on a set of episodes, the SOM adapts its weights to represent the most frequently encountered episodes or episode types. If capacity is limited, it learns generalisations over episodes: for instance, if dogs often chase cats, it will learn to represent the generic episode 'a dog chases a cat', abstracting away from token dogs and cats. This learning happens gradually, through incremental alterations of the SOM's long-term synaptic weights, which is what makes it part of the LTM system rather than the WM system. At the same time, it is useful to think of the current *pattern of activity* in the SOM as a WM representation, using the activity-based conception of WM.

The WM episode also provides input to the other main component of the LTM system: a medium representing the **current situation**. The pattern of activity in this medium can also be thought of as being held in WM: during experience, it encodes something like the agent's 'cognitive set', creating expectation and/or readiness for some episodes over others. At the same time, the network that generates these expectations is the product of long-term learning: situations are complex, high-level representations, integrating information from many learning episodes. How to learn situation representations is a matter of considerable debate. Our network implements a novel method, made possible by the use of localist representations of episodes in the c-ep SOM. Since individual units in this SOM represent whole episodes, it is able to represent a large *probability distribution over episodes*, in a pattern of activity over all its units. In our model, the medium encoding the 'current situation' is the hidden layer of a net-

[1]The AGENT and PATIENT fields are isomorphic with their WM counterparts, with 69 units each. The WM episode also has 33 units representing the action. For details see Takac and Knott (2016).

work that is trained to predict *the next episode*, after the completion of each episode. This layer takes inputs from the WM episode medium (holding the completed episode), and from a copy of its previous state (see the blue arcs in Figure 1). It is implemented as a recurrent SOM, specifically a 'merge SOM' (Strickert and Hammer, 2005).[2] When trained on a sequence of episodes, its units come to encode localist representations of commonly-occurring *sequences* of episodes or episode types: these serve as representations of situations in our model. A separate layer of perceptrons is trained to predict the next episode in the c-ep SOM from activity in the situation SOM. After training, the combined networks generate a distribution of possible next episodes in the c-ep SOM. During experience, this distribution can be used to reconstruct patterns of activity in the WM episode medium (and subsequently in the WM individual medium), which guide the agent in her experience of the next event. This mechanism is discussed in Takac and Knott (this volume). In the present paper, our focus is on how the situation SOM can play a role in episodic LTM and its interface with language.

Episodic LTM model As shown in Figure 1, the current situation SOM also takes input from media representing *times*. There are two of these, holding representations of **token times** and **time types**. Token times are representations of unique times. Each token time is a sparse distributed code of 10 neurons (in a field of 20 neurons); a new token time is selected after each episode. Time types are localist representations of times of day (morning, afternoon and evening): they update more slowly, once every 20 episodes, in a cyclical fashion. In our model these updates happen using an internal timer, but in a full model, perception would also obviously play a role; we thus envisage the type-token relation for times is somewhat similar to that for physical individuals, hence their parallel representation in Figure 1.

Because the situation SOM takes the current token time/time type as input in addition to the current episode and previous situation, it does not only learn to make predictions which inform SM processing: it also creates memories about *episodes in the past*. Most concretely, it can learn associations between specific episodes and specific token times. Since these associations provide recurrent input to the SOM, its memories of specific episodes in the past are naturally organised into sequences. This sequential organisation is characteristic of episodic memory, and has frequently been modelled using recurrent networks; see e.g. Howard and Kahana (2002). Using a recurrent SOM has two particular benefits. First, it learns localist representations that support very flexible *queries*: the trained SOM can be presented with any partially-specified pattern of inputs, and the pattern of activity over SOM units can be used to reconstruct a complete pattern. Second, it can learn *generalisations* over token times. Just as the c-ep SOM can learn generalisations over episodes, the situation SOM can learn generalisations about the *types* of episode that occur at particular time *types*, and about the *types* of episode that typically follow one another. In fact, while these generalisations can be thought of as summary statements about past experiences, they are also the basis on which the agent makes predictions about forthcoming episodes, to inform SM experience: it is by generalising over token episodes that the agent can use knowledge of the past to make predictions about the present. In our model, therefore, the circuits responsible for establishing cognitive set during SM experience are identical to the circuits responsible for storing sequentially structured episodic memories.

This is a point of difference with most existing models. Cognitive set and episodic memory are normally seen as separate neural mechanisms, the former involving prefrontal cortex (PFC) (Miller and Cohen, 2001), the latter involving hippocampus and associated cortex (Kesner and Rolls, 2015). But there is recent evidence that hippocampal assemblies also hold stimuli during the delay period of WM tasks (e.g. Olsen *et al.*, 2012), and that PFC and hippocampal activity is tightly coupled when material is maintained in WM, in a manner predictive of retention success (Battaglia *et al.*, 2011). Similarly, while episodic memory is often distinguished from 'semantic memory', defined to include memory for generic episodes, recent evidence suggests the hippocampal region encodes generic episodes as well as specific ones (St-Laurent *et al.*, 2009; Ryan *et al.*, 2008). We envisage that the WM/LTM media in our model are all implemented in circuits *jointly* recruiting PFC and the hippocampal region.[3] Both regions encode stimuli using sparse distributed representations (Wixted *et al.*, 2014), which are similar to those evoked in our SOM media and WM media (see below). And there are monosynaptic connections linking PFC and hippocampus, allowing the formation of neural ensembles spanning the two regions (Dégenètais *et al.*, 2003).

Retrieval from episodic LTM A key property of episodic LTM is its support of 'mental time travel': a process whereby the agent re-establishes a cognitive state that was active at some time in the past, and relives episodes experienced at that time. This involves suspending SM experience, and entering a special 'retrieval mode' (Buckner and Wheeler, 2001). Our model has a very natural implementation of retrieval mode: an active representation in the situation SOM can be used to activate semantic material *without engaging SM processes*. Most directly, we can use a pattern of activity in the situation SOM to reconstruct material in the SOM's input media. We can reconstruct the episode that *led* to the situation in the WM episode buffer, and we can reconstruct an associated time in the token time/time type media. (We will call the reconstructed episode the **antecedent**, and the reconstructed time

[2]The c-ep and situation SOMs each have 400 units. These sizes were chosen in proportion to the number of object and action types used in the model. Full details are given in Takac and Knott (2016).

[3]The perirhinal cortex perhaps has a particular role in representing LTM individuals (Kesner and Rolls, 2015), and PFC has a recognised role in post-retrieval processes (Ranganath and Knight, 2003), which we will discuss below.

the **temporal reference**.) We can also retrieve the episode that happened 'in' the situation (which we will term the **consequent**). This is reconstructed from the probability distribution generated by the retrieved situation in the c-ep SOM.

In our model, the process of entering retrieval mode begins with a representation of the 'current situation' in the situation SOM. This is a pattern of activity over many localist SOM units, encoding a mixture of token situations and generic situations that are *similar* to the present situation (in that a similar sequence of episodes preceded them), with activity proportional to similarity. The next episode is predicted from the distribution of activity over all these units. But each *individual* unit represents a specific past situation or situation type, which is associated with specific episodes from the agent's past. We assume each situation SOM unit is associated with an emotional valence (not shown in Figure 1), as suggested by Labar and Cabeza (2006). In our model, if the summed valences of the active situation SOM units exceed a threshold, retrieval mode is established, and the situation unit with highest aggregate similarity/emotional valence is activated by itself. From this situation unit, we can reconstruct an antecedent unit, or a temporal reference, or a consequent unit, as described above.

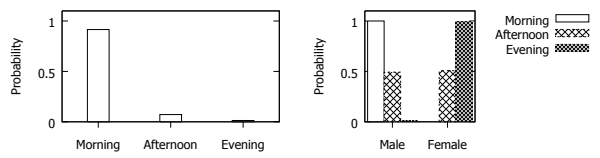
In this model, the process of remembering what happened in a retrieved situation is formally identical to the process of predicting what will happen in the current situation, in line with a constructivist account of LTM recall (Schacter, 1998). Note that what is retrieved in the c-ep SOM is in fact a distribution of possible episodes. The agent can use this distribution to reconstruct as best as possible the episode that *actually* happened in the remembered situation. But interestingly, she can also use it to simulate what 'might' have happened. In either case, she can use the recurrent circuitry of the situation SOM to play forward the real or imagined episode, and retrieve/imagine an arbitrary sequence of subsequent episodes.[4]

The network's interfaces to language The links between WM/LTM media and language are shown with blue lines in Figure 1. The key structures are the WM episode and the WM individual. Recall these WM media both encode prepared SM sequences, that can be actively replayed. In a model we developed earlier (Takac *et al.*, 2012; Takac and Knott, 2016) generating a clause involves *replaying a WM episode* and generating a NP involves *replaying a WM individual*, in a special mode where active SM/WM/LTM representations can trigger output phonology. In this model, NPs and clauses denote rehearsed cognitive routines rather than static mental representations. Our existing model focusses on generation of a single clause, from a single WM episode. The LTM/WM network presented in the current paper allows us to extend this model in several ways. Firstly, it allows us to model a *multi-sentence* discourse, reporting a *sequence* of episodes. In a standard model of discourse structure, sentences in a discourse add material to a temporally structured database of event representations, indexed by representations of discourse context and reference time, and each sentence updates the context and reference time (see e.g. van Lambalgen and Hamm, 2005). There are natural analogues of all these structures in our network: the situation SOM represents discourse contexts, the weights of its incoming and outgoing links hold the database of episodes, and the token time/time type media hold the reference times that index the database. Secondly, the network allows us to model sentences that *query* a database of episodes, or that respond to queries (i.e. questions and answers), as well as assertive sentences. This is because the medium holding sentence meanings, the WM episode, can also hold queries to the situation SOM, and responses to these queries (as we will show below).

Finally, the network lets us model linguistic devices that *reset the reference time* to some arbitrary point, such as those mentioned earlier, *in the afternoon* and *when John arrived*. It is easiest to approach these first from the perspective of sentence generation. Consider an agent who has just been reminded of a past situation, and wishes to convey this process in language. Recall the agent can retrieve the time associated with this situation (the temporal reference) or the episode which led to it (the antecedent). To communicate the retrieved situation, the agent can choose to retrieve either piece of information *and then express the retrieved information in language*, generating either a temporal referring expression or an antecedent clause. These are distinct communicative strategies, that initiate different LTM queries, and enable different linguistic interfaces. We suggest that the operations executing these strategies can trigger linguistic side-effects in their own right: specifically, the word *in* for the temporal reference strategy and the word *when* for the antecedent strategy. These words then naturally combine with words expressing a time, or an antecedent episode, to create a phrase like *in the afternoon* or *when John arrived*.[5] The hearer of such a phrase can use the strategy-signalling word to initiate a control process of his own, to enable an appropriate interface and build a representation in the relevant query medium, and then use this representation to retrieve a situation. The speaker can then produce a clause asserting (or querying) the consequent episode, that happened 'in' the retrieved situation, and the hearer can assert this episode in the retrieved situation, or execute a query about it, as appropriate. Note the speaker can generate the antecedent and consequent clauses *in either order* from a retrieved situation, since neither operation alters the situation representation: so our network naturally supports both preposed and postposed antecedent clauses.

[4] Of course it is important to distinguish between remembered and imagined situations. In our model, we take actual memories to be those with strong links to a token time, and which predict an episode distribution with low **entropy** (i.e. high confidence). These measures stand in for the 'feeling of familiarity' that accompanies actual memories in people. The measures are not fully reliable—but (notoriously) neither is the feeling of familiarity in people.

[5] On this account, temporal reference phrases like *in the afternoon* / *when John arrived* result from the execution of sequentially structured cognitive routines, so fit well with our general model of semantic representations as dynamic entities rather than static ones.



(a) (b)

Figure 2: (a) Distribution of type times for the query [Bird sings]. (b) Distribution of gender for the query [Person jogs] in the morning, afternoon, and evening.

Training and testing

To test the system's ability to answer questions and retrieve remote reference times, we exposed it to a stream of episodes with the following regularities. In the **morning**, 4 episode types were generated with equal probability: [Bird sings], [Man jogs], [Person sleeps], [Default]. [Person sleeps] was always followed by [(the same) Person sneezes]. [Default] episodes were of two types. One type featured a random combination of agent (Person/Dog/Cat/Bird), action (10 transitive, 6 intransitive, 4 causative), and patient (if appropriate: Person/Dog/Cat/Bird/Cup/Chair/Ball). The other type was the episode [Person kicks dog], followed by [(the same) Dog bites (the same) Person], or the episode [Person pats Dog], followed by [(the same) Dog licks (the same) Person]. In the **afternoon**, only [Default] episodes were generated. In the **evening**, 4 episode types were generated with equal probability: [Person lies-down], [Woman jogs], [Person sleeps], [Default]. [Person sleeps] was always followed by [(the same) Person snores]. Properties of episode participants not specified above (Number, Location, Colour, Gender) were generated stochastically as described in Takac and Knott (2016). The system was trained on a continuous stream of 20000 episodes, divided into 40 epochs of 500 episodes each. Then the situation SOM was presented with queries in its input media, encoding the semantics of questions, according to the scheme described above. A response was retrieved by propagating the query to the SOM (*ignoring its recurrent input*) to generate a pattern of activity in the SOM, then propagating this activity back into the input media as an activity-weighted linear combination of the weight vectors of all SOM units. Again see Takac and Knott (2016) for details.

When do birds sing? For this question, we presented the situation SOM with a partial query [AG=bird, PAT=empty, ACT=sing] and an unspecified time. After retrieval, we inspected the distribution of activities in the 'time type' medium. As shown in Figure 2a, the system correctly responds that birds sing in the morning.

Who jogs in the morning/evening? For this question, we presented the situation SOM with a partial query [AG=Person (unspecified for gender), ACT=jog], with the time-type set to either Morning, Afternoon or Evening. After retrieval, we inspected the distribution of activities in the gender part of the agent. The system correctly responds that morning joggers are men, and evening joggers are women (Figure 2b). Interestingly, it remains agnostic about the gender of joggers in the afternoon, when no-one jogged.

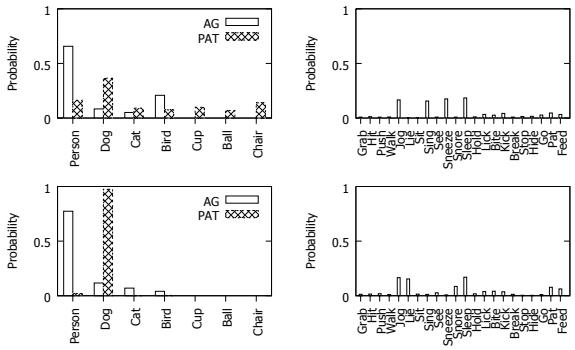


Figure 3: Distribution of agents and patients (left) and actions (right) for morning (top) and evening (bottom) episodes.

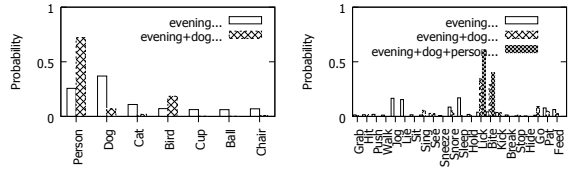


Figure 4: Change of distribution of patients (left) and actions (right) for progressively refined queries.

What happens in the morning/evening? For this question, we specified only a time type (morning or evening) and left the whole WM episode unspecified. The distributions retrieved in the WM episode are clearly distinct for the two times (see Figure 3). But in each case they are too broad to retrieve specific episodes, featuring particular combinations of agent, patient and action. However, given that episodes are experienced sequentially in our model (agent→patient→action), there is a natural way for a query to be *progressively refined*, by first selecting an agent from the agent distribution, then issuing another query *featuring this agent*, then iterating this process on the patient and action fields. (This kind of query refinement is exactly the kind of 'post-retrieval process' envisaged by Ranganath and Knight, 2003.) Figure 4 shows how patient and action distributions change when the query is progressively refined in this way: there are no binding errors in the episode eventually returned ([Dog licks/bites Person]), and this episode is indeed a common occurrence in the morning/evening.

When P kicks a dog, what happens next? This question targets the next episode prediction system. Recall that when a person kicks a dog, the dog always bites that person. For

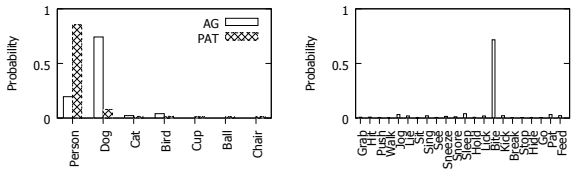


Figure 5: Distribution of agents and patients (left) and actions (right) in the episode predicted following [Person kicks dog].

this question, we presented the situation SOM with the (fully specified) episode [Person kicks dog], and computed the predicted distribution for the next episode in the c-ep SOM, from which we reconstucted distributions in the WM episode. As shown in Figure 5, these are correctly weighted towards [Dog bites Person]. As an extension of this, we asked a final question: **When P sleeps in the morning/evening, what happens next?** The system correctly answered [Person sneeze] for the morning query and [Person snore] for the evening query (along with other actions associated with morning/evening, at lower probabilities).

Summary and discussion

In this paper we presented a model of episodic LTM with two novel features. One regards the interface between LTM and WM. Our model reconciles the idea of a dedicated buffer for WM representations with the conception of WM representations as active LTM representations. The other regards the interface between LTM/WM and language. The model allows several elements of a linguistic theory of discourse structure (discourse contexts, reference times and temporally structured semantic representations) to be identified directly with components of the episodic LTM system. And it explains how sentences can query this LTM system, and use it to reactivate memories of temporally remote situations.

There are many issues to discuss about the design of the model—most importantly, its space requirements. The c-ep and situation SOMs must hold localist representations of a very large set of possible episodes and situations. At the same time, their ability to self-organise means they only need to represent those episodes/situations *that occur*, which are a very small subset of those that are *possible*. And their ability to learn *generalisations* over episodes/situations makes them efficient encoders. We provide an extended discussion of space requirements in Takac and Knott (2016).

Acknowledgements

This work was supported by the New Zealand Marsden Foundation through Grant 13-UOO-048. Martin Takac was also partially supported by Slovak grant VEGA 1/0898/14.

References

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *TICS*, **4**(11), 417–423.

Battaglia, F., Benchenane, K., Sirota, A., Pennartz, C., and Wiener, S. (2011). The hippocampus: hub of brain network communication for memory. *Trends in Cognitive Sciences*, **15**(7), 310–318.

Buckner, R. and Wheeler, M. (2001). The cognitive neuroscience of remembering. *Nature Reviews Neuroscience*, **2**, 624–634.

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake and P. Shah, editors, *Models of working memory*, pages 62–101. Cambridge University Press, Cambridge, UK.

Dégenètais, E., Thierry, A.-M., Glowinski, J., and Gioanni, Y. (2003). Synaptic influence of hippocampus on pyramidal cells of the rat prefrontal cortex. *Cerebral Cortex*, **13**, 782–792.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Phil. Transactions of the Royal Society B*, **362**, 761–772.

Fletcher, P. and Henson, R. (2001). Frontal lobes and human memory—Insights from functional neuroimaging. *Brain*, **124**, 849–881.

Howard, M. and Kahana, M. (2002). A distributed representation of temporal context. *Journal of Math. Psychology*, **46**, 269–299.

Kesner, R. and Rolls, E. (2015). A computational theory of hippocampal function, and tests of the theory: New developments. *Neuroscience and Biobehavioral Reviews*, **48**, 92–147.

Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA.

Labar, K. and Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, **7**(1), 54–64.

Miller, E. and Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167–202.

Moll, M. and Miikkulainen, R. (1997). Convergence-zone episodic memory. *Neural Networks*, **10**(6), 1017–1036.

Norman, K. and OReilly, R. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning systems approach. *Psychological Review*, **110**, 611–646.

Olsen, R., Moses, S., Riggs, L., and Ryan, J. (2012). The hippocampus supports multiple cognitive processes through relational binding and comparison. *Frontiers Hum. Neurosci*, **6**, Article 146.

Ranganath, C. and Knight, R. (2003). Prefrontal cortex and episodic memory. In E. Wilding *et al.*, editors, *Memory encoding and retrieval*, pages 1–14. Psychology Press, New York.

Ryan, L., Cox, C., Hayes, S., and Nadel, L. (2008). Hippocampal activation during episodic and semantic memory retrieval. *Neuropsychologia*, **46**(8), 2109–2121.

Schacter, D., Norman, K., and Koutstaal, W. (1998). Cognitive neuroscience of constructive memory. *Ann Rev Psych*, **49**, 289–318.

Shivde, G. and Anderson, M. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *JEP: Learning, Memory, and Cognition*, **37**(6), 1342–1370.

St-Laurent, M., Moscovitch, M., *et al.* (2009). Determinants of autobiographical memory in patients with unilateral temporal lobe epilepsy or excisions. *Neuropsychologia*, **47**, 2211–2221.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, **64**, 39–71.

Suddendorf, T. and Corballis, M. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, **30**, 299–351.

Takac, M. and Knott, A. (2016). A simulationist model of episode representations in working memory. Technical Report OUCS-2016-01, Dept of Computer Science, University of Otago. www.cs.otago.ac.nz/research/publications/OUCS-2016-01.pdf.

Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition*, **125**, 288–308.

van Lambalgen, M. and Hamm, F. (2005). *The Proper Treatment of Events*. Blackwell, Oxford.

Walles, H., Robins, A., and Knott, A. (2014). A perceptually grounded model of the singular-plural distinction. *Language and Cognition*, **6**, 1–43.

Wixted, J., Squire, L., Jang, Y., *et al.* (2014). Sparse and distributed coding of episodic memory in neurons of the human hippocampus. *PNAS*, **111**(26), 9621–9626.

Zwaan, R. (2008). Time in language, situation models, and mental simulations. *Language Learning*, **58**, 13–26.

What Makes You Feel You Are Learning: Cues to Self-Regulated Learning

Jessie Chin (chin5@illinois.edu), Elizabeth A. L. Stine-Morrow (eals@illinois.edu)
University of Illinois at Urbana Champaign, Department of Educational Psychology
Beckman Institute for Advanced Science and Technology, 405 N Mathews Ave
Urbana, IL 61801 USA

Abstract

While learning in a multitext environment increases with the rise of electronic environments, little is known about what makes learners feel that they should continue learning or already learn enough from one text. The current study aimed at examining what cues learners use to regulate their effort among multiple sources in a multitext environment. By manipulating the amount of new information and conceptual overlap across texts within a topic, we created three types of text environments to generate different trajectories of two cues to perceived learning, new information (measured by rating of perceived new information) and encoding fluency (measured by ratings of reading ease). Results showed that the dominant cue to gauge perceived learning was the perceived amount of new information. The study extended theories in animal foraging and metacognition, and established a novel paradigm to better investigate adult learning in the wild.

Keywords: information foraging, metacognition, self-regulated learning

Introduction

With the availability of information sources is exploding through the development of modern information technologies, learners have to take an active role to manage their learning in contemporary education settings. Since viewing all information is not possible given our limited processing capacity (Simon, 1956), choosing the important information, as well as deciding the amount of effort to allocate to particular sources of information, is necessary for promoting learning. However, how learners determine if they would like to continue or stop learning remains unknown. The current study aimed at investigating the cues learners used to gauge their perceived learning, which would in turn influence when they stop learning.

Research in metacognition has examined the way in which the learner takes an active role to regulate study. Research has suggested that learners allocate their study time according to their perception of how well they are learning the materials (e.g., judgments of learning (JOLs), Dunlosky & Connor, 1997; Metcalfe & Kornell, 2005). Koriat (1997) has argued that learners monitor different sorts of cues in accessing their learning to generate the JOLs. Intrinsic cues are the characteristics of the study items, such as inherent difficulty (e.g., word frequency, familiarity of the items). Extrinsic cues are the conditions of study (e.g., presentation rate or the encoding process available for the learners). Mnemonic cue reflects learners' own assessment of learning, such as the assessment of the outcome of previous recall (e.g., JOLs or recall accuracy from the previous round of study) or the assessment of processing (e.g., ease of processing), which is generated from the experience from participants.

Monitoring for text learning has usually been found to be less accurate than word-pair learning (e.g., Dunlosky, Baker, Rawson & Hertzog, 2006). One sort of metacognitive knowledge that readers often rely on to make judgments about their comprehension performance is processing ease, suggesting that readers who perceive a more fluent reading experience tend to believe that they have better comprehension (e.g., Dunlosky, et al., 2006; Hertzog, Dunlosky, Robinson & Kidder, 2003; Maki, 1998). Processing ease has often been studied as the amount of disruptions in reading. Dunlosky and his colleagues (2006) proposed that processing ease was one of the major cues for readers to judge their subsequent performance on recall or comprehension. However, this cue was often misleading for learners to estimate their learning (Hertzog, et al., 2003).

Information foraging theory also attempts to explain how people manage the acquisition of information in text environments, such as from the internet (Pirolli & Card, 1999). It suggested that information seekers would adapt to the payoff structure (information values and search costs) of the environment to accumulate information gain. Information value is estimated by "information scent" representing a proximal cue of profitability of information (Pirolli & Card, 1999). For example, in the web search, information scent is defined as the semantic relevance between a proximal cue (such as the hyperlinks, titles of the website, or the text snippets of a webpage) and the distal information (such as the linked website).

To answer the question, what makes learners feel they are learning, we adopted theories from both metacognition and information foraging to investigate the cues of perceived learning. Following Koriat (1997), two kinds of proximal cues were examined. First, like information scent of a site (Pirolli & Card, 1999), people may rely on the amount of new information (i.e., new ideas) in the text to determine the amount of knowledge they are able to gain from a text. Accordingly, when individuals perceive that there is more to learn from the texts (an intrinsic cue) and use this as a cue to perceived information gain, they would perceive that they could learn more.

Alternatively, as demonstrated in the literature in metacognition (Dunlosky et al., 2006), people may rely on encoding fluency (i.e., ease of processing, a mnemonic cue) to determine the amount of knowledge they are able to acquire from a text. Individuals who perceive that they have less difficulty learning information from the texts will perceive they learn more. Therefore, either one or both of the cues, the perceived amount of new information and

Table 1. Three types of text patches

	New Information	Encoding Fluency
The HI-LCO condition. High information and low conceptual overlap	High	Low
The HI-HCO condition. High information and high conceptual overlap	High	High
The LI-HCO condition. Low information and high conceptual overlap	Low	Very High

perceived encoding fluency may contribute to the perceived learning. Hence, the current study used three kinds of multitext environments to operationalize the availability of these proximal cues. By using the analogy of "food patches" in foraging theories, a collection of articles under a topic was operationalized as a text patch in the study.

Specifically, the manipulation focused on the relationship across articles on a topic (which operationalized the "text patch") to induce different trajectories of fluency and availability of new information to examine their relative contributions to perceived learning. To create three patch types varying in the amount of new information and the stimulation of perceived encoding fluency, the current study followed Kintsch's theory on how information is encoded from texts (Kintsch 1994). Kintsch and his colleagues (1975) suggested that although introducing new concepts in a sentence creates demands for comprehending, introducing new propositions (i.e., ideas) about the same concepts is less effortful relative to introducing new ideas that contain new concepts. In other words, while introducing more new ideas (propositions), there was a way to mitigate the effort of encoding by using the same concepts to build up new relationships among old concepts. They demonstrated this empirically by showing that while controlling the number of propositions and sentence length in the short sentences, learners spent longer time and recalled less for the items with more new concepts than the items with fewer new concepts. This finding suggested a way to differentially foster encoding fluency while increasing the equivalent amount of new information (ideas) in text, through the repeated usage of concepts across articles, called conceptual overlap. By manipulating the amount of new information and conceptual overlap across articles, the three conditions were summarized in Table 1.

In the HI-LCO condition (high information and low conceptual overlap), there were more and more new ideas and more new concepts introduced as participants read the articles in a topic. As shown in Figure 1a, participants would be expected to experience an increasing amount of new information if they continued to exploit this text patch. In addition, because the conceptual overlap among articles was low in this condition, participants would be expected to experience low encoding fluency across articles (as in Figure 1b). In the HI-HCO condition (high information and high conceptual overlap), there were more and more new ideas but few new concepts introduced as participants read the articles in a patch. The conceptual overlap among articles would be expected to increase coherence across articles even though the information load was high. Therefore, as shown in Figure 1a and 1b, participants would be expected to experience high information gain, but also high encoding fluency over the articles in this condition. In the LI-HCO condition (low information and high conceptual overlap), texts introduced virtually no new information and were, of course, high in conceptual overlap. As shown in Figure 1a, participants would be expected to perceive little new information because the content of articles was repetitive. At the same time, because of the use of paraphrase, participants would have high encoding fluency (Figure 1b).

The differential patterns of change in these cues across the articles in three conditions allowed us to test how these cues affect perceived learning in three conditions (Figure 1c and 1d). By investigating the functions of perceived learning in the three conditions, we were able to differentiate the cues learners use to estimate their perceived learning.

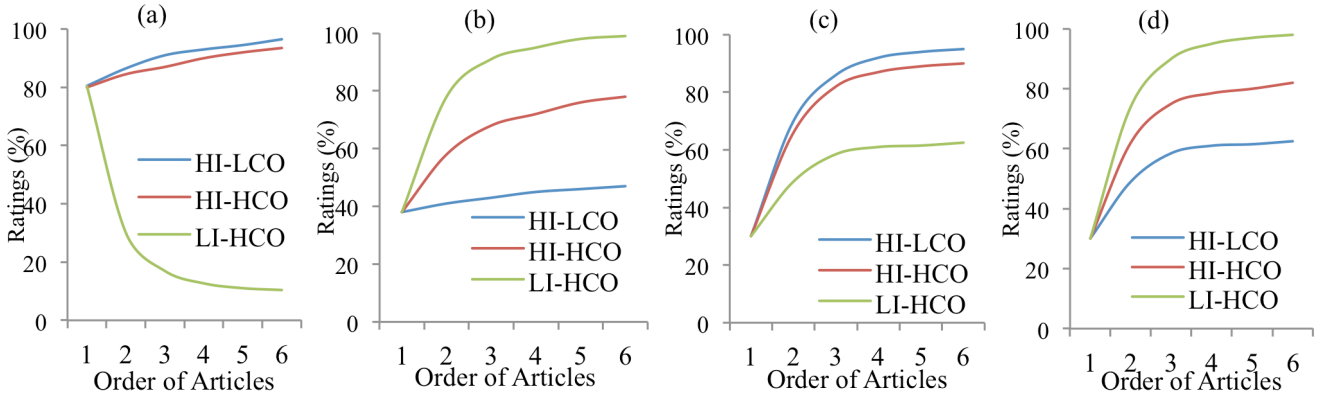


Figure 1. The expected (a) perceived amount of new inforamtion and (b) reading ease as a function of the sequence of articles in each condition; The expected perceived learning if adults adopt the cue of (c) the perceived amount of new information or (d) reading ease as a function of the sequence of articles in each condition

Method

Participants

Seventy-nine participants were recruited from Amazon Mechanical Turk. Data from 27 people were excluded due to technical problems. Data from the remaining fifty-two participants (Mean age =38.9, *SD*=10.9, age range=23-69; 46% female) were used in the analysis. Most (60%) of the participants completed college (Mean years of education =15.3, *SD*=1.9, range=12-20).

Table 2. Descriptive statistics of the text properties among three conditions

	HI-LCO	HI-HCO	LI-HCO
Number of words	220.28 (3.06)	220.56 (3.06)	220.28 (3.06)
Number of sentences	13.22 (0.43)	13.39 (0.43)	12.61 (0.43)
Sentence length	16.91 (0.50)	16.69 (0.50)	17.75 (0.50)
Log word frequency (WF)	2.88 (0.03)	2.90 (0.03)	2.91 (0.03)
WF of the content words	2.01 (0.03)	2.03 (0.03)	1.99 (0.03)
Flesch-Kincaid grade level	10.07 (0.22)	10.19 (0.22)	9.50 (0.22)
Number of unique concepts in all the articles	172~190	100~106	56~80

Materials

In the study, adults were asked to learn about "Transplantation and Donation" by reading a set of short articles about three topics, Bone Grafts, Blood Donation and Corneal Transplants. These text materials were adapted from articles extracted from dozens of health information websites with good credibility, such as MedlinePlus, WebMD, Mayo Clinic and others. For each topic, three sets of six sequential articles were created, with the three sets constituting the primary manipulation of the experiments, such that sets varied in the amount of new information and degree of conceptual overlap across articles (i.e., repeated use of the same concepts in multiple articles). *New information* in an article was operationalized as the ideas (propositions) that had not appeared in any of the previous articles under the same topic. *Conceptual overlap* was operationalized as the proportion of the same content words that were appeared in the consecutive articles. This measure was obtained from Coh-Metrix, a text analysis database developed by Grasser and his colleagues (Graesser, McNamara, Louwerse & Cai, 2004).

The first article for each topic was always an introductory article to provide an overview of the topic and was the same across the three sets. Subsequent articles in each set were

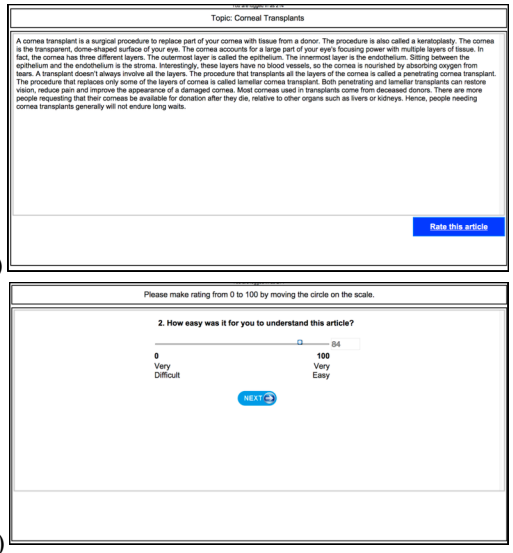


Figure 2. Example layouts: (a) article page, (b) rating page

constructed as follows.

The HI-LCO condition was the high information - low conceptual overlap condition. Following the introduction, the subsequent articles were constructed so that each subsequent article Article$_N$ contained relatively more new information (i.e., propositions) and new concepts than Article$_{N-1}$. This was accomplished by elaborating on five different areas (subtopics) of the topic to create the five subsequent articles. In this condition, new ideas as well as new concepts were consistently introduced across the six articles in a topic.

The HI-HCO condition was the high information - high conceptual overlap condition. Subsequent articles were created by introducing new ideas using about the concepts that had been introduced in prior articles within the same topic. There was always more new information but few new concepts in Article$_N$ relative to Article$_{N-1}$ in this condition. Subsequent articles were constructed by elaborating on existing concepts with new ideas. For example, for the topic of "bone graft," the first article briefly introduced concepts such as allografts and autografts, as well as more basic concepts related to donors, locations of grafts (hips, spines), tissue banks, materials, procedures, and so forth; subsequent articles elaborated, for example, on allografts and autografts, in part by grounding the explanation in these more basic concepts.

The LI-HCO condition was the low information - high conceptual overlap condition. This was a paraphrase condition that introduced minimal new ideas or concepts across the five subsequent articles. There was almost no new information or new concepts in Article$_N$ relative to Article$_{N-1}$. In this condition, the information in the first article was paraphrased in the five subsequent articles without adding new concepts or new ideas. (Because it was not practically feasible to have the low information - low conceptual overlap condition in which introducing no new

information with new concepts, there were three conditions in the study).

To verify the differences in conceptual overlap across conditions, CohMetrix (Graesser et al., 2004) was used. Conceptual overlap was defined as the proportion of content words that occurred in the current article and any of the previous articles on the same topic ($M_{HI\text{-}LCO}$=0.28, $SE_{HI\text{-}LCO}$ =0.01; $M_{HI\text{-}HCO}$=0.47, $SE_{HI\text{-}HCO}$=0.01, $M_{LI\text{-}HCO}$ =0.70, $SE_{LI\text{-}HCO}$=0.01). ANOVA was conducted to examine the effects of condition on conceptual overlap. There was a significant effect of condition on conceptual overlap (F(2,42)=367.20, p<.0001). Post-hoc tests further confirmed that conceptual overlap in the HI-LCO condition was lower than the HI-HCO condition (d=-0.18, p<.001) and the LI-HCO condition (d=-0.42, p<.001) and the conceptual overlap in the HI-HCO condition was lower than the LI-HCO condition (d=-0.23, p<.001).

Despite the fact that conceptual overlap was different across conditions, all other linguistic properties of the texts in three conditions were carefully controlled, including the number of words, number of sentences, sentence length, word frequency (Balota, et al., 2007) and readability (descriptive statistics in Table 2). There were no differences in the number of words (F(2,51)=0.003, p=0.99), number of sentences (F(2,51)=0.93, p=0.40), sentence length (F(2,51)=1.27, p=0.29), average log word frequency (F(2,51)=0.23, p=0.80), log word frequency of all content words (F(2,51)=0.50, p=0.61) and Flesch-Kincaid grade levels (F(2,51)=2.95, p=0.6) across three conditions. Therefore, except the differences in new propositions and conceptual overlap, the texts used in three conditions were largely equivalent.

Experimental Design

The study followed a within-subject design with condition being the within-subject variable (The HI-LCO condition, HI-HCO condition and LI-HCO condition). The order of condition was counterbalanced. Participants would answer three multiple-choice questions about the shared contents across three conditions (the first article) at the end of the each topic. The purpose was to motivate participants to pay attention to the study, and make sure that participants read and understood the main ideas in each topic.

Procedure

When a participant logged into Mechanical Turk, s/he was presented with a consent form, and then completed a demographic questionnaire. For the main task, participants were asked to learn health information from multiple texts, with the goal to learn as much as possible about topics related to Donation and Medical Transplants. All participants were asked to read 18 articles, six articles in each of the three conditions. The assignment of condition to each topic was counterbalanced.

To begin, participants viewed a button corresponding to one of the topics, Bone Grafts, Blood Donation or Corneal Transplants. Once clicking on the topic button, the first article was presented (Figure 2a). After reading each article, participants pressed a button, which initiated the presentation of three rating scales (Figure 2b) to operationalize perceived encoding fluency, the perceived amount of new information and perceived learning. The perceived amount of new information was assessed with the item, "Taking into account the other articles that you have read today about this topic, how much new information was in this article?" which participants rated on a scale of 0 to 100, where 0 meant that this article did not have any new information, and 100 meant that this article had completely new information. Encoding fluency was assessed with the item, "How easy was it for you to understand this article?" which participants rated on a scale of 0 to 100, where 0 meant that it was very difficult to learn from this article, and 100 meant that it was very easy to learn from this article. In addition to the hypothesized proximal cues used to determine information gain, perceived learning was assessed by the item, "Taking in to account the other articles that you have read today about this topic, how much new information did you learn from this article?" which participants rated on a scale of 0 to 100, where 0 meant that they did not learn anything new from this article, and 100 meant that they learned everything from this article.

After providing these ratings for the article, participants pressed the "NEXT" button to read the subsequent article. After reading all six articles about one topic, participants were directed to a new topic and continued in the same way. Articles within a topic were presented sequentially. After reading one topic, participants answered three multiple-choice questions about the main concepts in the introductory article. Lastly, participants completed two cognitive tasks. Reading time on each page was recorded.

Table 3. Descriptive statistics for ratings of perceived reading ease, the perceived amount of new information, perceived learning and comprehension in three conditions

	HI-LCO	HI-HCO	LI-HCO
Reading ease	79.64 (2.75)	81.89 (2.40)	84.33 (2.13)
The perceived amount of new information	84.29 (1.97)	76.85 (1.92)	34.93 (2.96)
Perceived learning	81.35 (2.01)	75.76 (1.84)	38.51 (3.38)
Comprehension (Accuracy scores)	0.83 (0.04)	0.82 (0.04)	0.94 (0.03)

Results

The analysis focused on answering two questions; (a) whether participant's ratings of new information and ease of processing varied across conditions as predicted, and (b) to what extent did learners use these cues to judge their perceived learning across the articles.

The Effects of Condition on Monitoring

Table 3 summarizes the descriptive statistics for ratings of perceived ease of processing, the perceived amount of new

information, and perceived learning; and for accuracy scores of the comprehension questions in the three conditions. Repeated measures ANOVA was used to examine the effects of condition on encoding fluency showing a significant main effect of condition ($F(2,51)=3.88$, $p<.05$). Post-hoc tests suggested that adults thought articles in the HI-LCO condition were more difficult than articles in the HI-HCO condition ($d=-2.25$, $SE=1.15$, $p=0.05$) and the LI-HCO condition ($d=-4.69$, $SE=1.92$, $p<.05$). However, $SE=3.61$, $p<.001$).

Finally, there was a significant effect of condition on comprehension performance ($F(2,36)=3.66$, $p<.05$). Adults tended to perform better in the LI-HCO condition than the HI-HCO condition ($d=0.12$, $SE=0.05$, $p<.05$) and the HI-LCO condition ($d=0.11$, $SE=0.04$, $p<.01$), in part because of the repetition of ideas across articles. There was no difference in performance between the HI-LCO condition and the HI-HCO condition ($d=0.01$, $SE=0.05$, $p=0.87$).

Table 4. Estimates of parameters of the mixed effects models - effects of condition on ratings (reading ease, perceived amount of new information, perceived learning) across order of articles

	Reading Ease		New Information		Perceived Learning	
	Estimate (*SE*)	*t*	Estimate (*SE*)	*t*	Estimate (*SE*)	*t*
Intercept	78.72 (2.50)	31.51*	72.35 (3.05)	23.71*	69.56 (2.95)	23.62*
Order	0.91 (0.34)	2.65*	-8.95 (0.72)	-12.5*	-7.20 (0.65)	-11.01*
Con 1	1.10 (1.09)	1.01	7.33 (2.27)	3.23*	7.24 (2.07)	3.49*
Con 2	1.37 (0.97)	1.41	-2.46 (2.01)	-1.22	-3.42 (1.84)	-1.86†
Con 1 x Order	-0.86 (0.28)	-3.05*	4.99 (0.58)	8.56*	4.06 (0.53)	7.62*
Con 2 x Order	-0.11 (0.25)	-0.44	-0.13 (0.52)	-0.25	0.34 (0.47)	0.72

Note. Contrast coding was used to test the effects of conditions on ratings across the order of articles. Con 1 was to examine the differences between high information (averaging HI-LCO and HI-HCO) and low information (LI-HCO); Con 2 was to examine the effects of conceptual overlap in high-information conditions (HI-LCO and HI-HCO).

* $p<.05$; † $p<0.10$

articles in the HI-HCO condition were equivalent with articles in the LI-HCO condition in terms of encoding fluency ($d=2.44$, SE=1.87, $p=0.20$). There was also a significant effect of condition on the perceived amount of new information ($F(2,51)=123.45$, $p<.001$). Post-hoc tests further showed that participants thought there was more new information in the HI-LCO condition than the HI-HCO condition ($d=7.44$, $SE=2.00$, $p<.001$) and the LI-HCO condition ($d=49.36$, $SE=4.12$, $p<.001$). Adults also rated that there was more new information in the HI-HCO condition than the LI-HCO condition ($d=41.92$, $SE=3.66$, $p<.001$). Condition also had an effect on perceived learning ($F(2,51)=96.69$, $p<.001$). Post-hoc tests further showed that participants thought that they learned more in the HI-LCO condition than the HI-HCO condition ($d=5.59$, $SE=1.96$, $p<.01$) and the LI-HCO condition ($d=42.84$, $SE=4.10$, $p<.001$), and adults thought that they learned more in the HI-HCO condition than the LI-HCO condition ($d=37.25$,

The Effects of Condition on the Changes in Monitoring

Mixed effects models (PROC MIXED in SAS) were used to analyze how the ratings changed across articles in three conditions article by article (Estimates of parameters are in Table 4). The same contrast coding was used to examine the effects of condition.

Results showed that reading ease increased across articles in the LI-HCO condition more than in the other two conditions ($t=-3.05$, $p<.01$). However, there was no difference in the changes in reading ease between the HI-LCO and HI-HCO conditions ($t=-0.44$, $p=0.66$) (See Figure 3a for the average ratings of encoding fluency as a function of articles in three conditions). On the other hand, the perceived amount of new information decreased across articles more in the LI-HCO condition than the other two conditions ($t=8.56$, $p<.001$). There was also no difference in the changes of the perceived amount of new information

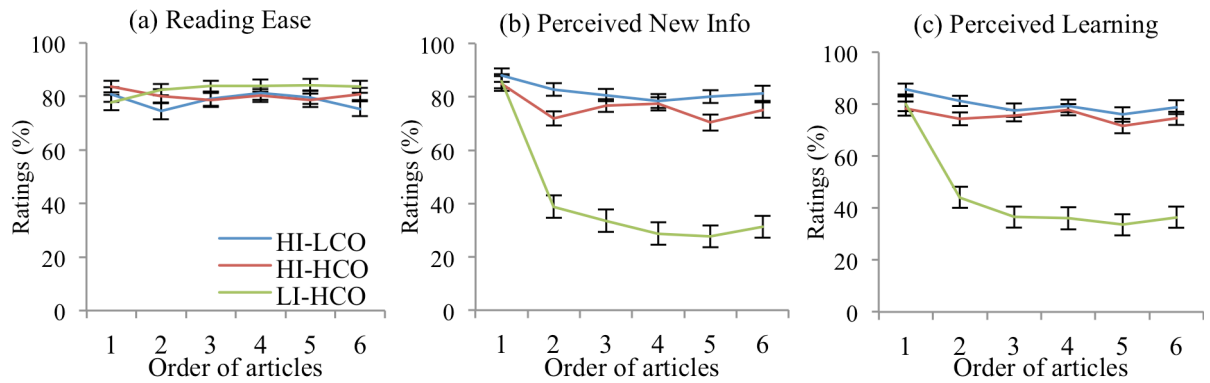


Figure 3. (a) Perceived ease, (b) the perceived amount of new information, (c) perceived learning from articles on the same topic, as a function of condition

between the HI-LCO condition and HI-HCO condition (t=-0.25, p=0.80) (See Figure 3a for the average ratings of the perceived amount of new information with progression through the articles in the three conditions). Similarly, perceived learning also decreased across the articles more in the LI-HCO condition than the other two conditions (t=7.62, p<.001). The changes in perceived learning did not differ in the HI-LCO condition and HI-HCO condition (t=0.72, p=0.47) (See Figure 3a for the average ratings of perceived learning as a function of articles in three conditions). So generally it appeared that perceived learning tracked the changes in new information and not fluency (as measured by reading ease).

The Cues to Perceived Learning

To further investigate the cues used to judge perceived learning, an analysis was done to test how changes in reading ease and the perceived amount of new information related to the changes in perceived learning article-by-article using the linear mixed effects models. Results showed that the perceived amount of new information was the dominant cue used to judge perceived learning (*Est*=0.78, *SE*=0.01, t=61.48, p<.0001), but that reading ease was not (*Est*=-0.04, *SE*=0.03, t=-1.24, p=0.21).

Discussion

Our findings successfully validated the text manipulation in three conditions such that adults perceived that the LI-HCO condition had less new information than the HI-HCO and HI-LCO conditions did. Also, averaging ratings from six articles in one condition, adults tended to perceive that the LI-HCO and HI-HCO conditions were easier than the HI-LCO condition. Interestingly, averaging six articles, adults even rated articles in the HI-HCO condition overall as easy as those in the LI-HCO condition. Even if there was much more new information in the HI-HCO condition than the LI-HCO condition, conceptual overlap induced encoding fluency for learners to comprehend information in the HI-HCO condition and the LI-HCO condition equivalently. Therefore, as expected, when manipulating the amount of new information and conceptual overlap across articles, adults perceived more new information and lower encoding fluency in the HI-LCO condition; more new information and higher encoding fluency in the HI-HCO condition, and less new information and higher encoding fluency in the LI-HCO condition. Results also suggested that amount of new information was more central for learners to judge how much they were learning than fluency.

Some previous studies have shown that learners relate encoding fluency to their judgments of learning, although encoding fluency is a misleading cue for estimating actual learning performance (Dunlosky et al., 2006; Rawson et al., 2002). Interestingly, in our study, although some learners did relate encoding fluency to their learning to a modest degree, the dominant cue to perceived learning was not encoding fluency. Encoding fluency did not contribute to perceived learning alone. It appears that encoding fluency is only a meaningful cue for adult learners to gauge their learning when they also perceive new information is available. The dominant cue for judging perceived learning in this study, which has been neglected in the previous literature, is the perceived amount of new information available. Therefore, the main contribution of the current study was to provide the evidence that the perceived amount of new information has accounted for most of the variance in perceived learning relative to encoding fluency. However, regarding the probes we used to measure the cues and perceived learning, there were similar phrases used in the probes for both the perceived amount of new information and perceived learning, follow-up studies would avoid using similar terms across probes for cues and perceived learning.

In addition to the use of different probes for the cues and perceived learning, follow-up studies would use this paradigm to examine how learners use these cues to regulate their study and their actual learning outcome. Given that the goal of the study was to understand what makes learners feel that they are learning, we did not measure the learning outcome deliberately. The multiple-choice questions used in the study were mainly used to ensure that participants captured the major ideas in the topic (with the accuracy scores exceeding 80%), instead of to measure the individual differences in learning outcome. Hence, follow-up studies would focus on how learners use these cues to regulate their study for optimizing their learning in the multitext environments.

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition, 25,* 691–700.

Dunlosky, J., Baker, J. M. C., Rawson, K. A. & Hertzog, C. (2006). Does aging influence people's metacomprehension? Effects of processing ease on judgment of text learning. *Psychology and Aging, 21,* 390-400.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36,* 193-202.

Hertzog C., Dunlosky, J. Robinson, A. E. & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22-34.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49(4),* 294-303.

Kintsch, W., Kozminsky, E., Streby, W.J., McKoon, G. & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. *Journal of verbal learning and verbal behavior, 14,* 196-214.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349 –370.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds). *Metacognition in educational theory and practice* (p. 117-144), Mahwah, NJ: Lawrence Erlbaum Associates.

Metcalfe, J. & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52,* 463-477.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106,* 643-675.

Rawson, K. A., Dunlosky, J. & McDonald, S. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 55*(2), 505-524.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychology Review, 63, 129-138.

A Tale of Two Disasters: Biases in Risk Communication

Matthew B. Welsh[1] (matthew.welsh@adelaide.edu.au)
Sandy Steacy[2] (sandy.steacy@adelaide.edu.au)
Steve H. Begg[1] (steve.begg@adelaide.edu.au)
Daniel J. Navarro[3] (daniel.navarro@unsw.edu.au)

1 - Australian School of Petroleum, University of Adelaide, Australia
2 – School of Physical Sciences, University of Adelaide, Australia
3 – School of Psychology, University of NSW, Australia

Abstract

Risk communication, where scientists inform policy-makers or the populace of the probability and magnitude of possible disasters, is essential to disaster management – enabling people to make better decisions regarding preventative steps, evacuations, etc. Psychological research, however, has identified multiple biases that can affect people's interpretation of probabilities and thus risk. For example, availability (Tversky & Kahneman, 1973) is known to confound probability estimates while the description-experience gap (D-E Gap) (Hertwig & Erev, 2009) shows low probability events being over-weighted when described and under-weighted when learnt from laboratory tasks. This paper examines how probability descriptions interact with real world experience of events. Responses from 294 participants across 8 conditions showed that people's responses, given the same described probabilities and consequences, were altered by their familiarity with the disaster (bushfire vs earthquake) and its salience to them personally. The implications of this for risk communication are discussed.

Keywords: description-experience gap; risk communication; decision making; availability; bias.

Introduction

People make decisions based on their perception of risks. They also demand that others (e.g., policy makers) propose methods to mitigate risks. Such proposals are often costly to implement and there may be financial or other costs should a risky event occur. It is, therefore, important that risk perception matches reality. Psychological research, however, indicates that people are subject to a range of biases that lead to mis-perception of probabilities leading to inconsistent trade-offs and sub-optimal decisions (see, e.g., Kahneman & Tversky, 1979; Lichtenstein, Fischhoff, & Phillips, 1982; Tversky & Kahneman, 1973, 1974).

The L'Aquila verdict, which saw 6 scientists convicted of manslaughter for failing to adequately communicate earthquake risk prior to the event, shocked the scientific community and prompted urgent re-appraisal of methods used to convey information about low-probability, high-impact events. Prior to the trial, an international commission convened to investigate the earthquake's predictability - and the communication of its risk - recommended that risk assessments be formalised through authoritative operational earthquake forecasting (OEF). It also suggested that social scientists develop methods for communicating risk to the public and decisions makers (Jordan et al., 2011).

Research on seismic aspects of OEF is progressing with new models being developed and tested (see, e.g., Marzocchi, Lombardi, & Casarotti, 2014; Steacy et al., 2014). However, progress on communication remains a major topic of discussion (Jordan et al., 2014).

Biases in Risk Perception

Given the above, it seems valuable to review psychology's findings about how people interpret risk (i.e., the probability of an undesirable outcome). A variety of known effects seem relevant to how a person, faced with information being provided by an expert, might interpret that information and thus react. Some central examples are discussed below.

Prospect Theory and the Description-Experience Gap

Prospect Theory (Kahneman & Tversky, 1979) describes how the majority of people respond when facing risky choices in simple gambles such as: 'Take $100 now' or 'Win $250 if this coin toss comes up heads and nothing otherwise'. In such instances, people are risk averse when dealing with gains and risk seeking for losses. They also act as if low probabilities are higher than they actually are and as if high probabilities are lower.

Many of these observations, however, are shown to reverse in situations where, instead of being told the probabilities, people have to learn them from an environment – the so-called Description-Experience Gap (hereafter, D-E Gap; Hertwig & Erev, 2009).

Risk communication, of course, involves both described probabilities (from the scientists or government) and the probability for risky events that each member of the 'audience' has learnt from their environment.

Availability

Availability (Tversky & Kahneman, 1973) describes the tendency of people to weight the likelihood of events according to how *available* such events are to memory. This makes sense in that, when operating in a natural environment, the number of instances of an event recalled should depend on how many such events have been seen.

It can leads to biased estimates of probability, however, where an event's availability differs from its actual rate of occurrence. For instance, media coverage focusses on 'interesting' – and often rare – events; the result being that people overestimate the likelihood of events (e.g., terrorist

attacks) that receive coverage while underestimating probabilities for events that are less often reported upon.

Key aspects of availability are the familiarity and salience of events – that is, how well known they are to the person and how memorable (for a discussion of this, see Sunstein, 2005). For example, most people are aware of (i.e., familiar with) the risks of house fires but these are more salient to someone who has seen such a fire than to people who have simply heard about them. Both familiarity and salience affect the availability of people's memories of particular events and thus their estimates of their probabilities.

Given this, it seems likely that people's recalled probabilities will differ markedly from the actual probabilities of occurrence for the disasters.

Format Changes

Another effect known to alter interpretation of probabilities is changes to presentation formats. For example, there is evidence that people prefer natural frequencies (e.g., '1 in 100') to percentages (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995) and even those who dispute this (Sloman et al., 2003) agree that the way probabilities are presented can enhance or impede people's understanding of them.

Relative vs Absolute Change

Finally, it is possible that people are more sensitive to relative than absolute change, leading to the expectation that people's reaction to a disaster risk may (depending on the presentation of the information) be influenced by the degree to which the probability has *changed* as well as by the final probability itself (Stone, Yates, & Parker, 1994).

Aims and Objectives

Given the above, the central aim of this paper is to determine what effects play a role in predicting behaviour when the described probability of a disaster needs to be integrated into a person's pre-existing knowledge – learnt from their experiences with such disasters in the real world.

A key question, then, is whether people's behaviour will be better predicted by Prospect Theory or the D-E Gap? Prospect Theory predicts that low probability events will be overweighted while the D-E Gap predicts the opposite.

Availability, meanwhile, suggests that people's learnt probabilities will be biased by their familiarity with and direct experience of the disaster they are being asked about.

Given this, the aim is to interrogate people's experience with disasters and use this and their responses to illuminate the probabilities they have 'learnt' from the environment.

Method

Participants

Participants were 294 residents of Adelaide (152 female, 133 male and 9 identifying as other) ranging from 13 (see below) to 63 (M = 25.8, SD = 8.7).

Initial recruitment was undertaken via fliers placed around the University of Adelaide, which indicated that only Adelaide residents over 18 could participate. A significant amount of snowball sampling occurred as participants forwarded the survey link to friends and family, however, leading to the inclusion of one respondent who was only 13.

Of the participants, 12 had not completed high school and 46 were high school graduates. The remainder included: 99 who had attended but not yet graduated university; 87 with bachelor degrees; and 50 with higher degrees. Participants received a $20 gift voucher for completing the experiment.

Materials

An online survey was designed, which asked participants: for demographics; to select their response to a hypothetical disaster; and then describe their experience with disasters of that type. The survey questions are expanded on below.

Demographics

Participants provided their age, identified gender, level of education and length of residency in Adelaide.

Disaster Questions

The experiment used a 2x2x2 between-subjects design. (NB: the sample was limited by funding constraints to being slightly underpowered for a 2x2x2 design, it is sufficient for 2x2 interactions.) The scenario described to participants that they had inherited a $400,000 house in the Adelaide Hills and, following this, scientists had revised the probability of a disaster (that would destroy the house) occurring in that area up to 1% per annum for the next 10 year period.

Familiarity. The first condition was familiarity with disaster type. Two disasters were included: earthquakes and bushfires as these reflect different experiences for Adelaide residents. Bushfires, while rare in an absolute sense (i.e., the probability of an individual being affected in any given year is very low) are familiar to Adelaide residents, with major events occurring at a rate of 6-7 per decade, historically (Luke & McArthur, 1978), and multiple, less severe fires occurring every year – with Country Fire Service (CFS) data recording ~12/year in the 2001-2015 period across Greater Adelaide's peri-urban area (CFS, 2015).

Earthquakes, by comparison, are unfamiliar, with only 30 perceptible quakes since records began 132 years ago (i.e., ~0.23 events per year) and none having caused damage since a M5.5 event in 1954. By these data, bushfires are ~50 times as frequent in Adelaide as earthquakes.

Format. The second condition was the format used to express the disaster probability – either a percentage (e.g., 1%) or natural frequency (e.g., 1 in 100). Otherwise, probabilities were identical across conditions with a 1% p.a. chance of the disaster occurring across the next decade.

Magnitude. The final condition was the magnitude change in probability and was either 10 or 100. That is, while the final (annual) probability of the disaster was 1%, this was described as having been increased from either 0.1% or

0.01% - or the natural frequency equivalents.

Options. In all cases, participants selected from the same set of four responses. These were: 1) Do nothing; 2) Spend $10,000 on preparing the house to better cope with the named disaster and reduce the probability of it being destroyed by half; 3) Spend $20,000 on preparing the house to better cope with the named disaster and reduce the probability of it being destroyed by three-quarters; and 4) Sell the house at a $40,000 loss.

These options thus reflect increasingly extreme *immediate* responses to the revised disaster probabilities and, in that sense, are ordinal. They also, however, form (approximate) pairs in expected value terms, with the EV of options 1 and 4 both being approximately -$40,000 when considered across the 10 year period of increased risk – while options 2 and 3 are both valued at approximately -$30,000. (These approximate values were used as it seemed unlikely that participants would use probability theory to calculate the precise values; i.e., that a 1% p.a. risk equates to a 9.56% chance of disaster across the decade). Given this, the choice data is more appropriately regarded as nominal.

Experience Questions.

In addition to the demographic and main experimental questions, five questions captured participants' real world experience with the disasters they were asked about. The first two of these asked: how many such events the participant recalled having occurred in Adelaide; and how many of those recalled events caused significant damage.

Three questions asked about the disaster's salience to the participant. Specifically, whether: 1) the participant; 2) their friends/family; or 3) people in their local area, had suffered injury or property loss from such a disaster.

Procedure

Participants accessed the survey in their own time using the URL and an identifier code on tear-off tabs at the bottom of the fliers. After an initial page describing the basics of the experiment, participants were sorted into groups according to which of the 8 identifier codes they entered (corresponding to the possible condition combinations).

They were then shown the disaster question for that condition and asked to select their response. After this they were asked the five follow-up questions and four demographic questions. Finally, they were asked to enter their email address (to arrange gift voucher pick-up). Most participants completed the task in less than 10 minutes.

Results

Initial analyses indicated that the *magnitude* manipulation was having no effect (main or interaction), therefore results are discussed without further mention of this. Table1 shows the percentage of participants selecting each of the four options – overall and, separately, divided by the *familiarity* and *format* conditions.

Looking at Table 1, one sees three instances where the odds of a participant selecting a particular option seems to be affected by the manipulations with people about twice as likely to select to "Do nothing" rather than something in the low familiarity (Quake) and percentage formats conditions, Odds Ratio = 2.45 and 2.25, respectively. Participants in the frequency format conditions were, similarly, twice as likely to choose to sell the house as not, Odds Ratio= 1.98.

Table 1. Percentages of participants selecting each option overall and by condition.

		Do nothing	$10K prep	$20K prep	Sell house	Total
Overall		20.1	42.2	20.1	17.7	100
Familiarity	Quake	13.6	18.7	9.5	8.8	50.7
	Fire	6.5	23.5	10.5	8.8	49.3
Format	Perc.	13.3	21.8	8.8	6.5	50.3
	Freq.	6.8	20.4	11.2	11.2	49.7

To assist in seeing interaction effects, Figure 1 shows the proportion of participants selecting each of the options, divided simultaneously by both familiarity and format.

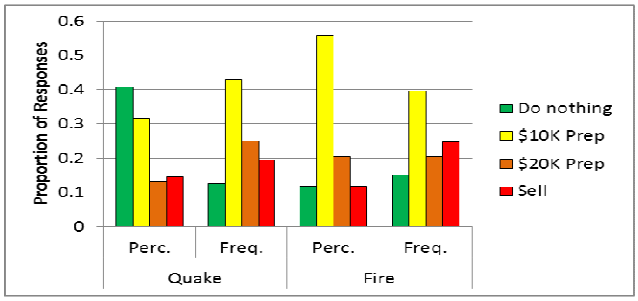


Figure 1: Participant responses by familiarity and format.

As was seen in Table 1, both manipulations affect responses and Figure 1 suggests an interaction effect – with participants seeing the earthquake version of the task with percentage probabilities far more likely (41%) to do nothing than any other group (~13% across the three other groups).

A 2x4 Chi-square test confirmed that familiarity was significantly associate with participant choice, $\chi^2 = 9.2$, $p = .036$. Analysis of residuals indicated the familiarity effect was driven by more participants selecting option 1 (do nothing) in the unfamiliar (Earthquake) condition relative to the familiar (Bushfire) condition.

The effect of format was also significant, with people presented probabilities in natural frequency format tending towards more active responses. Specifically, the format change equated to a 'upward' shift of half a category on average. Again, a 2x4 Chi-square test indicated that format was significantly associated with participant choice, $\chi^2 = 10.8$, $p = .012$. The residuals of this test told a similar story, with differences in the proportion of people choosing to do nothing in the percentage format being significantly higher than in the natural frequency format – although here there was also a significant difference for option 4 (sell house),

selected by more people in the frequency format. That is, the frequency format discouraged doing nothing and encouraged the most extreme response (sell house).

Table 2. Log-linear analysis of 4x2x2 contingency table.

	Interactions	G^2	df	p
1	Choice*Format*Familiarity	38.6	10	<.001
2	Choice*Format	12.6	3	.006
3	Choice*Familiarity	12.0	3	.007
4	Format*Familiarity	.06	1	.807
5	Choice*Format (No Familiarity)	26.5	6	<.001
6	Choice*Familiarity (No Format)	25.9	6	<.001
7	Format* Familiarity (No Choice)	13.9	4	.008

Note: rows 2-4 and 5-7 differ according to whether G^2 is calculated after collapsing across the 3rd condition or after removing its effect. Note 2: $G^2 \approx \chi^2$.

To assess the interaction effect, a 4x2x2 (choice x familiarity x format) log-linear analysis was conducted. Looking at its results in Table 2, one sees that most associations are significant – excepting between Format and Familiarity (row 4, which reflects the near-equal numbers in the conditions). Rows 2 and 3 correspond to the above 2x4 tests and provide similar results. The key point from the table, however, is the highly significant 3-way interaction (row 1) – that is, format and familiarity interact with choice both individually and when considered together.

Availability of Experience

Recalled Events.

Peoples' recollection of disasters was tested simply by asking them to indicate how may such events they recalled from their residency in Adelaide and how many resulted in "widespread" damage. As no participants had resided in Adelaide long enough to recall the 1955 earthquake, no participants should have recalled earthquakes causing damage, making this comparison with the bushfire group redundant. Given this, Figures 2 and 3 show only the total number of disasters recalled plotted against the predicted numbers based on participant's length of residency in Adelaide and the observed base rates of occurrence.

Looking at Figures 2 and 3, one can see marked differences in the numbers of events recalled, with 7 being the highest value provided by a participant for number of earthquakes recalled whereas, in the bushfire condition, several values were close to 100 and the highest was 1000 (requiring the log scale seen in Figure 2).

A similarity between the two figures, however, is seen in the consistent *under*-recollection of disasters. That is, people recalled fewer bushfires and earthquakes than the base rates suggest they have experienced – and this effect was far stronger in the bushfire condition where the typical number recalled was less than 10 but the prediction was more than 200. Calculation of the proportion of recalled over experienced events across the two conditions shows a clear difference, with average proportion recalled being 0.54 (median = 0.22) for earthquakes and 0.05 (median = 0.02) for bushfires. That is, the lower probability events (earthquakes), while recalled less often than the base rate predicts, are disproportionately better remembered than the more common bushfires – by a factor of 10.

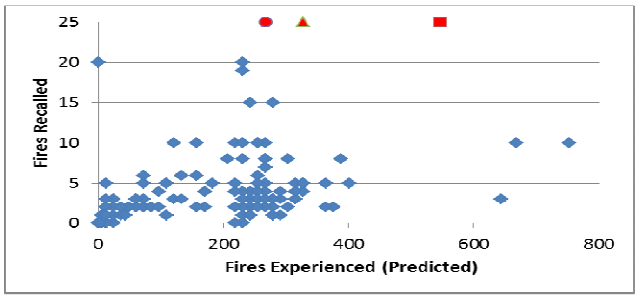


Figure 2: Scatterplot of predicted (residency x base rate) vs actual recalled number of bushfires. Red datapoints are outliers with (from left) 50, 80 and 1000 recalled fires.

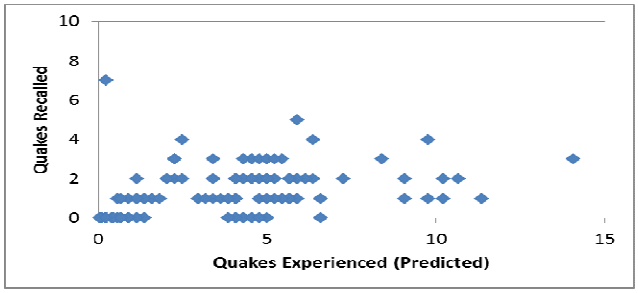


Figure 3: Scatterplot of predicted (residency x base rate) vs actual recalled number of earthquakes.

Interestingly, the number of recalled events seems largely unrelated to the participants' choices by in the experimental task – as shown in the mean rank data in Table 3.

Table 3. Mean rank of proportion of events recalled.

Choice	Earthquake	Bushfire
Do nothing	123.6	176.1
$10K preparations	113.1	170.4
$20K preaprations	134.6	169.5
Sell House (-$40K)	125.6	170.9

Analysis of differences between the groups in Table 2 was undertaken via a 2x4 (familiarity x choice) ANOVA, run - due to the skewed data - on the ranked proportion of events recalled (as per Conover & Iman, 1981). This confirmed the difference in proportion of events recalled between the earthquake and bushfire familiarity groups, $F(1, 283) = 21.3$, p <<.001 but indicated no significant difference between the proportion of events recalled by participants making different choices, $F(3, 283) = 0.27$, $p = .848$, and no interaction effect between choice and disaster type.

Salience.

The salience, to an individual, of the disaster type they were asked about was scored on a 0-3 scale according to how many of the three, salience questions they answered 'yes'.

The data for bushfires and earthquakes are shown in Figures 4 and 5, respectively. Comparing these figures, one sees, firstly, that salience differs markedly between the bushfire and earthquake groups – as expected, given our sample. Most (59%) of the bushfire group had non-zero salience, compared to only 14% of the earthquake group.

Looking at the Figures together, though, one can see hints of the same pattern of results – increased salience seems to push people away from the two (financially) worse options (doing nothing or selling the house) and towards paying for preparations to reduce the risk to their house.

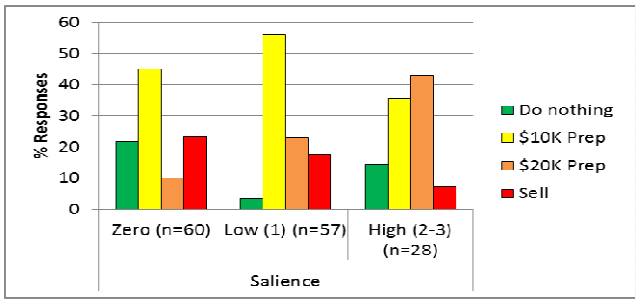


Figure 4: Participant responses by salience (bushfire).

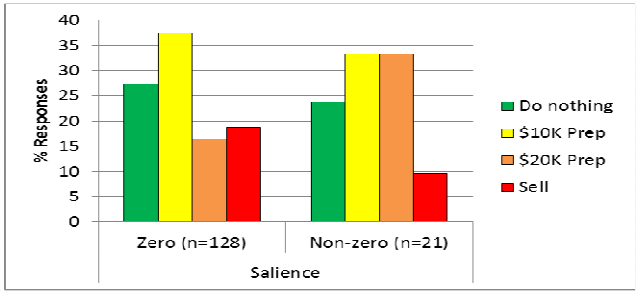


Figure 5: Participant responses by salience (earthquake).

Given only two people scored 3 on salience in either disaster condition, scores of 2 and 3 were grouped as 'high' salience for a 3x4 (salience x choice) chi-square test examining data from both disaster types simultaneously. This showed a significant association between salience and participants' choices, $\chi^2(6) = 22.1$, $p = .002$. Examination of residuals indicated this result was driven by an increased tendency for people with salience zero to elect to 'do nothing' and a decreased tendency for them to select '$20K preparations', whereas people with salience 1 disproportionately selected '$10K preparations' and those with higher salience '$20K preparations'.

Running the same analysis on just the bushfire data produced near-identical results but a 2x4 (salience x choice) chi-square test conducted on the earthquake data alone failed to reach significance – despite the similarity in patterns noted above, $\chi^2(3) = 3.8$, $p = .300$ – indicating that the overall effect is primarily driven by the bushfire data.

Discussion

These results provide a number of insights into difficulties facing people interested in the communication of disaster risk. Practical points relating to risk communication are discussed in the two following sections, prior to a discussion of theoretical implications in terms of Prospect Theory and the Description-Experience Gap.

Familiarity and Format

Two of our three manipulations showed clear effects: familiarity with disaster type; and the probability presentation format. The first is perhaps unsurprising, as one might expect that familiarity with a disaster might affect how one responds to it - and knowing how one *should* respond is likely to reduce the tendency (seen in our data) of people to do nothing when faced with an unfamiliar disaster.

The second observation, that people respond more strongly when shown probabilities in a natural frequency format than percentages seems more surprising but accords with Gigerenzer and Hoffrage's (1995) suggestion that people better understand this format. That is, it suggests that percentages may be an added level of abstraction that detracts from the immediacy of the risk communication.

Availability: Recalled Disasters and Salience

Interestingly, despite the clear effect of the disaster *familiarity* condition, the number/proportion of disasters recalled by participants had no bearing on their responses. That is, people who remembered more or fewer disasters relative to the ground truth did not differ from one another in terms of their responses to our questions. The clearest trend from this data was that, for both types of disaster, people tended to recall fewer than the base rate of occurrence suggested they had experienced.

By contrast with recollection, salience clearly affected responses with people having had more direct experience of disasters inclined to take stronger precautions but also to avoid overreaction, as selling the house in our example would be classified (in expected value terms, at least). The implications of this for risk communication are awkward in that it suggests that messages need to be tailored according to whether a person has had *salient* experience of a disaster – as those without such may overreact or fail to act when presented with the same information to which a more experienced person will react appropriately.

Theoretical Implications

The recalled experience and salience data have implications for how best to approach risk communication. That people tended to underestimate the occurrence of disasters fits with the Description-Experience Gap understanding of learnt, low probabilities – tending to be lower than the true probability – rather than the Prospect Theory assumption that low probability events will be overweighted.

However, the fact that the less common of the two disasters (earthquakes in the Adelaide context) shows *higher* rates of recall suggests that earthquakes are rare enough to be more memorable than bushfires – making them more available and thus causing their occurrence to be underestimated *less* than the more common bushfires.

This has implications for situations where smaller scale disasters occur relatively frequently – specifically, that people will tend to markedly underestimate the rate of occurrence of these, possible as a result of memory limitations – in both encoding numbers of similar, unremarkable events and then in retrieval when asked to recall a number of such events.

Future Research

The results suggest a number of directions. First is an attempt to more rigorously examine experience – that is, the roles of familiarity, recalled events and salience. How these relate to one another may shed further light on when and how to communicate risk. A first step would be to examine the different salience questions separately as it seems unlikely that these will be equally predictive (e.g., personal loss may have more impact than losses in the local region).

A second step would be to conduct the reverse of the familiarity condition used herein – that is, test a sample in a location where earthquakes are more common and bushfires rarer to ensure that it is the manipulation and not differing characteristics of these disasters driving the results.

Finally, future work should pay closer attention to the question of what is the *right* thing to do? Despite a brief discussion of expected value, the decision here was complex enough that no cut-and-dried correct answer existed. Given the goal of improved risk communication, it is important, however, to know if people are reacting appropriately – which requires a correct answer for comparison.

Conclusions

A person's familiarity with a disaster - and its salience to them - impacts how they interpret and respond to events with the same risks. The unfamiliar earthquake risk in our experiment disproportionately produced inaction, while low salience did the same or pushed people to overreact. People with personal experience of disasters, by comparison, tended to prepare and try to mitigate their effects.

Also of interest was the observation that presenting probabilities as natural frequencies resulted in people taking more active steps to mitigate or avoid the disaster.

Finally, our results offer insights into how people learn the probabilities of disasters from the real world, the biases that tend to affect these and how these interact with biases known to affect described probabilities.

Acknowledgments

This research was funded by a University of Adelaide interdisciplinary research fund (IRF) grant. MBW and SBH acknowledge support from the CIBP group's industry sponsors: B.G. Group, ExxonMobil, Santos and Woodside. DJN was supported by the ARC Future Fellowship scheme.

References

CFS (Producer). (2015, 22/01/2016). Fires in the peri-urban area of Greater Adelaide. [Map] Retrieved from www.cfs.sa.gov.au

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35*(3), 124-129.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102*(4), 684-704.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences, 13*(12), 517-523.

Jordan, T., Chen, Y.-T., Gasparini, P., Madariaga, R., Main, I., Marzocchi, W., . . . Zschau, J. (2011). Operational Earthquake Forecasting: State of Knowledge and Guidelines for Implementation. *Annals of Geophysics*.

Jordan, T., Marzocchi, W., Michael, A., & Gerstenberger, M. (2014). Operational earthquake forecasting can enhance earthquake preparedness. *Seismological Research Letters, 85*(5), 955-959.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: an analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Luke, R. H., & McArthur, A. G. (1978). Bush fires in Australia. *Bush Fires in Australia.*

Marzocchi, W., Lombardi, A. M., & Casarotti, E. (2014). The establishment of an operational earthquake forecasting system in Italy. *Seismological Research Letters, 85*(5), 961-969.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes, 91*(296-309).

Steacy, S., Gerstenberger, M., Williams, C., Rhoades, D., & Christophersen, A. (2014). A new hybrid Coulomb/statistical model for forecasting aftershock rates. *Geophysical Journal International, 196*(2), 918-923.

Stone, E. R., Yates, J. F., & Parker, A. M. (1994). Risk communication: Absolute versus relative expressions of low-probability risks. *Organizational Behavior and Human Decision Processes, 60*(3), 387-408.

Sunstein, C. R. (2005). Precautions against What-The Availability Heuristic and Cross-Cultural Risk Perception. *Alabama Law Review, 57*, 75.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Individual Differences in Pupil Dilation during Naming Task

Kaidi Lõo (kloo@ualberta.ca)
Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada

Jacolien van Rij (vanrij.jacolien@gmail.com)
Center of Language and Cognition, University of Groningen
P.O.box 716, 9700AS Groningen, The Netherlands

Juhani Järvikivi (jarvikiv@ualberta.ca)
Department of Linguistics, University of Alberta
4-55 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada

Harald Baayen (harald.baayen@uni-tuebingen.de)
Department of Linguistics, University of Tübingen
Wilhelmstrae 19 D-72074 Tübingen, Germany

Abstract

The present study investigates individual differences in pupil dilation during standard word naming. We looked at (i) how individual subjects' pupil size changes over the course of time and (ii) how well pupil size is predicted by the frequency of the stimuli. The time course of the pupil size was analysed with generalized additive modeling. The results show large individual variations in the pupil response pattern in this very simple task. Although, we see a pupil response to both stimulus onset and articulation onset and offset, both the amplitude of change and the direction of change differ substantially between subjects. This raises the question of what makes the pupil response functions so diverse, and one factor indicated by the frequency effect or the lack thereof might be shallow reading versus reading for content.

Keywords: pupillometry; word naming task; individual differences; word frequency; lexical processing

Introduction

The eye's pupil diameter size reflects changes in luminance, emotional state, but also cognitive processes (Ahern & Beatty, 1979; Hess & Polt, 1960; Kahnemann & Beatty, 1966; Young & Biersdorf, 1954). For example, Kahnemann and Beatty (1966) conducted a recall task to show that pupil size can be used to measure mental effort. The recall of more complex number strings triggered a stronger pupil response compared to the recall of less complex strings. Pupil dilation has also been found to reflect *linguistic* processing, such as the complexity of a linguistic task (Hyönä, Tommola, & Alaja, 1995), sentence intelligibility (Zekveld, Kramer, & Festen, 2010), sentence complexity (Ben-Nun, 1986; Just & Carpenter, 1993; Schluroff et al., 1986) and spoken language comprehension (Engelhardt, Ferreira, & Patsenko, 2010).

Finally, pupil diameter has been used to investigate lexical processing costs (Geller, Still, & Morris, 2015; Kuchinke, Võ, Hofmann, & Jacobs, 2007; Papesh & Goldinger, 2012). Kuchinke et al. (2007) measured pupil dilation in high and low frequency words and found that high frequency words receive higher peak pupil dilation compared to low frequency words. Papesh and Goldinger (2012) replicated the effect, using delayed naming paradigm. They delayed the time of the naming response up to 2000 ms and observed that frequency affects pupil size, even after the naming responses were issued. Geller et al. (2015) reported for a masked priming study that higher-frequency words elicited an earlier dilation of the pupil compared to low-frequency words. They also observed that dilation increased when words had more lexical competitors, both for low- and high-frequency words.

The present study investigates changes in pupil size while participants read out loud words presented to them one by one on a computer screen, using an inter-stimuli interval of around 4000 ms in order to allow the pupil to contract to baseline after each trial. The present study addresses the question of the extent to which the response to cognitive load in lexical processing varies across individual participants. The response of the pupil to a cognitive task has been described mathematically by Hoeks and Levelt (1993) as single smoothly increasing and then slowly decreasing function of time. Wierda, van Rijn, Taatgen, and Martens (2012) argued that a time series of pupil dilation values can be a superposition of several such functions, arising as a consequence of several cognitive events taking place within the window of time under investigation.

If in a controlled task such as word naming the pupil indeed responds in a fixed and consistent way to the demands of the task (reading the word, preparing for articulation, controlling the articulation itself, and preparation for the next trial), then one would expect the same pupil dilation function for all participants, possibly with minor variations in shape, similar to minor variations in intercept and slope that one may detect for linear response functions with the help of (generalized) linear mixed models. However, when reading styles differ substantially across participants, a general response function may not be appropriate. In what follows, we address these questions using generalized additive mixed modeling (Wood (2006); see also Baayen, van Rij, de Cat, & Wood, 2016). First, however, we introduce the details of our naming experiment.

Word Naming Experiment

Participants

Thirty-three native speakers of Estonian (18 women; 22-69 years; mean age 38), with normal or corrected to normal vision and no diagnosed speech impairments, participated in the experiment. The participant's dominant eye was tracked (30 right eye, 3 left eye). Participants received 15 euros for their participation.

Materials and Design

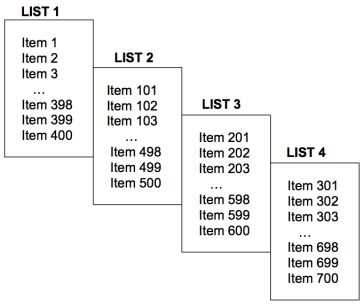


Figure 1: Outline of the experimental design.

A total of 2800 case-inflected nouns were randomly selected from the Balanced Corpus of Estonian. The frequency distribution resembled the distribution of the complete corpus, and ranged between 1 and 1000 per million (mean 14.27). The length of stimuli varied between 2 and 19 characters (mean 7.88; sd 2.62 characters). Twenty-eight experimental lists were created from the randomized set of 2800 items, each with 400 words. To maximize the number of items in the experiment, an overlapping design was used (see Figure 1), with a 300-word overlap between successive lists. A given word occurred four times in the experiment, once in each of four lists.

Apparatus

The experiment was conducted in a medium illuminated sound-attenuated booth. Eye movements and pupil size were recorded using the head-mounted EyeLink II eye tracker by SR Research Ltd. EyeLink II is a video-based tracking system with a resolution of 500 Hz, ca 3.0 ms delay, and an average spacial accuracy of approximately 0.5 degrees of arc.

The naming data was recorded separately from the stimulus presentation program ExperimentBuilder by SR Research with a Marantz PMD670 digital recorder, using a supercarioid condenser table top microphone by Beyerdynamic, placed approximately 10 cm from the participant's mouth.

Procedure

Participants were tested individually. First, they were familiarized with the procedure: reading aloud, as naturally as possible, words presented one by one on the computer screen. They were asked to start speaking as soon as the word appeared on the screen. As participants were wearing a head-mounted eye tracker, they were instructed to move their head as little as possible.

Participants were seated in front of the computer screen at a distance of approximately 60 cm. The experimenter placed the headband of the eye tracker over participant's head and adjusted it such that the eye position could be correctly tracked on the computer screen. Further, the eye tracking system was calibrated. Adjustments were made until the spacial accuracy of the eye location measurement was smaller than 0.5 degrees of arc. The stimuli were presented on a 21-inch Dell grey background computer screen in black lower case 26-point Courier New Bold font. The screen resolution was 1024x768 pixels.

Each trial started with a drift correction on the left of the screen, after which the target appeared in the center. The target stayed on the screen for 1500 ms and was then replaced by a fixation cross that remained on the screen for 2500 ms. Thus, in total each trial lasted 4000 ms. We extended the length of each trial to ensure that the pupil size was recorded with enough time delay for the pupil to contract to the baseline. The experiment started with ten practice trials, which were followed by the 400 experimental trials. Every 100th trial was followed by a short break. At the end of the experiment participants filled out a language background questionnaire. The whole procedure lasted approximately 90 minutes.

Data Preparation and Analysis

Naming latency and articulation duration were calculated directly from the audio recordings using Matlab (version 8.5.0, (MATLAB, 2015). Prior to the analysis, data for two participants were removed due to technical problems during tracking. Thus, the analysis was conducted on the data from 31 subjects.

Trials with misspoken stimuli, eye movement spikes and saccades due to eye-blinks were removed from the analysis using R (version 3.2.1, R Development Core Team, 2015, 6.3% of the trials).

The statistical analysis was performed with R (version 3.2.1, (R Development Core Team, 2015), using the *mgcv* package, version 1.8.6 of 2015, for generalized additive mixed regression (GAMM) modeling (Wood, 2006); see also Baayen et al., 2016, for visualization, we made use of the *itsadug* package (version 1.0.1, van Rij, Baayen, Wieling, and van Rijn (2015).

The dependent variable of interest was log-transformed *Pupil size*, measured in the standard (arbitrary) units delivered by the eye tracking system. The main predictors were *Time* in milliseconds and *Frequency*. Frequency was transformed to normality using the Johnson transformation (version 1.3, R package *Johnson*, Fernandez (2014)). *Gaze coordinates* (x- and y-axis position on the screen in pixels) were added to account for changes in measured pupil size due to the location on the screen.

For each participant, we fitted a separate GAMM to the 400

time series of pupil dilation values (resulting in 31 models). In addition to a general smooth for a subject's pupil dilation curve, we also included, for each word, a nonlinear random effect curve in time, using shrunk factor smooths. The X and Y coordinates of the fixation position were included as controls using a tensor product smooth, and as we anticipated the effect of frequency to vary over time, we also included a tensor product smooth for frequency and time. The model was checked for autocorrelation in the residual error, and an AR (1) autocorrelation parameter was then added to the model to remove, as far as possible, autocorrelative structure from the residuals (see Baayen et al. (2016) for a detailed discussion).

In what follows, we first discuss the estimated pupil dilation functions for the ensemble of 31 subjects. We then provide more details on three subject groups resulting from a clustering analysis.

Individual Patterns over Time

Figure 2 presents the estimated pupil dilation functions provided by GAMMs that focused on the main effect of time (leaving out frequency as covariate). The x-axis represents time and the y-axis presents pupil size. The stimulus was presented at time 0 ms, and a trial ended after 4000 ms. The first black dotted vertical line in a panel represents the median onset of articulation and the second dotted line the median offset of articulation.

Figure 2 shows that the relation between the Pupil size and the Time differs substantially between subjects. The different dilation functions fall into five groups, obtained with a divisive hierarchical clustering method using Manhattan distance applied to the first three principal components of a principal component analysis of the correlation matrix of the empirical first derivative of the subject time smooths, and indicated by different color coding.

Group 1. The first group is the largest (12 subjects: s02, s04, s05, s07, s08, s09, s18, s19, s21, s26, s30 and s31). For these subjects, the pupil dilation function shows a first peak shortly after stimulus onset and a second peak at or shortly after the onset of articulation. However, some subjects show a slightly different pattern from this general trend. For subject s02, s21, s26 and s30 there is no clear first peak, and for subject s04 and s09 the second peak much later after the onset of articulation.

Group 2. The second group includes eight subjects (s03, s11, s13, s15, s22, s25, s32, s33). Here only one clear pupil response is visible, which occurs slightly after speech onset and peaks after speech offset. Somewhat different are subject s11 and s15 who also show a slight peak after stimulus onset. However, the main difference is a well-differentiated initial peak in the pupil dilation function which is present in group 1, but absent in group 2.

Group 3. The third group also includes eight subjects (s06, s12, s14, s23, s24, s27, s28 and s29). Compared to the first two groups, this group is more heterogeneous. Also here, only one pupil peak is visible and it occurs after speech onset. However, unlike in the second group, in this group pupil size is declining instead of increasing from stimulus onset. Some subjects again deviate from the general group pattern. For example, for subject s14 there is no peak after speech onset, but the pupil stays at high plateau even after the articulation offset. Furthermore, pupil fluctuation patterns of subject s27 and s29 are somewhat similar to the first group as also here two clear pupil responses are present. However, for these subjects the second peak is much smaller than the second peak in the first group. Finally, unlike the first two groups, the relative heights of the pupil maxima are quite diverse.

Other dilation curves. The last two groups obtained by the hierarchical clustering technique clearly stand out from the rest (see the last row in Figure 2: subject s10, s16 and s17). Subject s10 and s16 start with a high pupil size, but it declines significantly after the stimulus onset. The pupil size increases also only a little after the articulation onset and stays relatively constant without a clear second peak. Finally, the last cluster only included subject s17. This subject seems to be similar to the first group and also has an initial pupil peak. However, like subject s10 and s16, this subject has no clear second peak. Because the two last groups include only three subjects, we excluded them from further group analysis.

In the present experiment, there are three key events to which the pupil appears to be sensitive: the presentation of the stimulus, the onset of articulation, and the offset of articulation. The first three groups show an increase in pupil size around the onset of articulation. The first group also shows an increase in pupil size after the word is presented, whereas it is less so for the rest of the four groups. The difference between the second and third is group is that although for both a second peak is present, subjects in the second group have an increasing pupil size and subjects in the third group decreasing pupil size. Finally, subject s10, s16 and s17 (group 4 and 5) start with a decline in pupil size from stimulus onset and show only a small increase at articulation onset.

The subjects-specific pupil dilation functions indicate that even though participants engage in exactly the same task, with exactly the same procedure, they apparently engage in this task in cognitively significantly different ways. Subjects who show little or no dilation following stimulus onset might be experienced readers, and subjects who show a decreasing pupil dilation function might be highly skilled talkers.

Next, we included a smooth for frequency and a tensor product smooth for frequency and time to the GAMM model. We compared the frequency effect differences between the three largest subject groups.

Frequency Effects in Subject Groups

Figure 3 presents the main partial effect of frequency (top panels) and the way frequency over time modifies the pupil dilation function for the three main subject groups (bottom panels). As in Figure 2, the first vertical dotted line represents median articulation onset and the second line median articulation offset. The frequency measure, normalized (and thus centered), ranges from -2 to +3; pupil size was log-transformed.

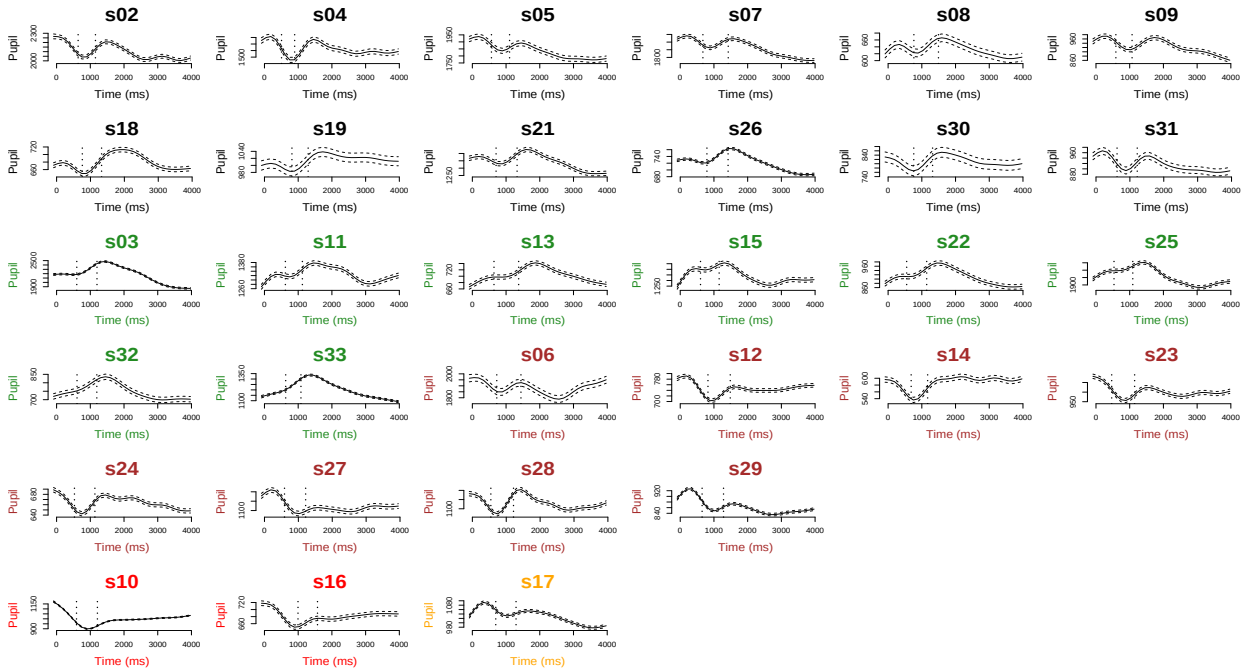


Figure 2: The fitted effects of time smooths for 31 subjects, five groups are indicated by different color-coding. The first vertical dotted line indicates the median articulation onset and the second dotted line the median articulation offset.

Group 1. The summary of the statistical model for the first group indicated a significant main effect of Frequency (t(898846)=-4.93; p-value < 0.0001). Pupil dilation decreases linearly with increasing frequency, which is consistent with high-frequency words being cognitively less demanding, typically affording shorter responses latencies in the word naming task (Forster & Chambers, 1973).

Frequency entered into a significant nonlinear interaction with time (F(14.02, 896425.4)=8.72; p-value < 0.0001). The way in which frequency modulates the pupil dilation is presented in the bottom-left panel of Figure 3. The contour plot can be read like a topographic map with peaks and valleys, yellow indicates the highest and blue the lowest elevation. The contour lines show the slope of the Pupil size as a function of Time and Frequency. The lines that are closely spaced represent steep slope and contour lines further apart gentler slopes, i.e., slower changes in Pupil size.

When the value on the y-axis is kept constant at zero, we see a similar pattern to the simple time smooths for members of the first group in Figure 2. The pupil starts off with a slight peak in pupil size shortly after the stimulus onset, contracts before the articulation, dilates again during articulation and finally, contracts at the end of the trial. This is color-coded by changes from green to yellow, from yellow to green, from green back to yellow and from yellow back to green and then blue. However, as we can also see, the pupil dilation changes differently over time dependent on the frequency of the word. The pupil size increases earlier in time and more with lower frequency words than higher frequency words (see the peak around 0 ms and the yellow peak around 1500 ms in the bottom-left panel of Figure 3).

It is noteworthy that the effect of frequency on pupil size is the largest for this group *after* speech offset. This can be seen by considering the gradient before and after articulation. Before articulation, we find three contour lines, after articulation, we find five. Even around 3000 ms after stimulus onset, frequency still has a strong effect. This suggests the effect of frequency is not restricted to the early stages of information uptake, and is likely to have a strong semantic component.

Group 2. As the middle panels of Figure 3 show, the second group has no main effect of Frequency (F(1, 578743.7)=1.17; p-value=0.28. However, the frequency and time interaction is significant (F(15.034, 578743.7)= 9.90; p-value < 0.0001). The contour lines in the bottom-middle panel are all fairly vertical, indicating an effect of time and hardly any modulation by frequency, except a weak effect before speech onset at high frequency range.

This pattern of results suggests that the second group might not be semantically engaged. It is well known that experienced readers can read out text while thinking about other issues. We suspect that the members of the second group are rather 'mechanically' performing the task, but are not deeply engaged in interpretation, possibly because isolated words in a word naming experiment are out of context and have no communicative value.

Group 3. This group also shows no significant main effect of Frequency (F(1.255, 591182.6)=8.95; p-value=0.42), but a non-linear interaction (F(14.388, 591182.6)=7.52.; p-

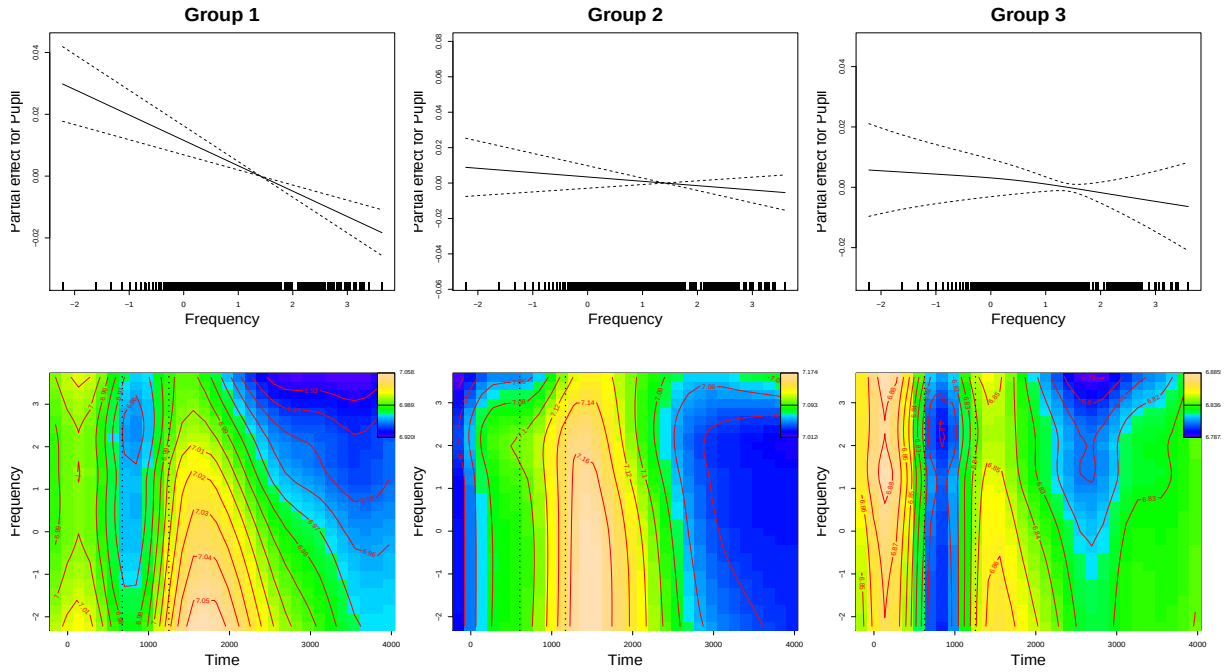


Figure 3: The partial effects of Frequency (the upper panels) and the tensor product smooth for Frequency and Time for three subject groups without random effects. The first vertical dotted line is the median articulation onset, the second vertical dotted line is the median articulation offset.

value <0.0001). The interaction is presented in the bottom-left panel of Figure 3.

The interaction effect of the third group is somewhat similar to the first group. However, compared to the first group, the effect is more shallow and gradient in particular during and after articulation. At stimulus onset, there is no frequency effect, but there is a gradual decline in pupil size across frequency span until the speech onset. Further, compared to the first group, we see fewer contour lines and a more gradual change from yellow to green after the end of the articulation. Also for this group the pupil dilation peak is weaker after speech offset. The results suggests more engagement with words compared to the second but less compared to the first group.

General Discussion

The results of this study can be summarized as follows. Inspection of the pupil dilation function for 31 participants revealed substantial variation in not only the magnitude of dilation in response to stimulus onset, speech onset, and articulation offset, but also in the direction of change, with some subjects showing contraction and others dilation. In the light of such substantial inter-subject variability, it makes little sense to try to extract a 'population dilation curve' from this kind of data. Such a curve would not come close to characterizing the pupil response function for many of the participants. This seems to apply in particular to pupillometry data, but also cognitive research in general (see e.g., Roehm, Bornkessel-Schlesewsky, Rösler, & Schlesewsky, 2007, for similar conclusions in EEG data). What this shows is either that different subjects engage in exactly the same task in very different ways or that subjects are all engaged in exactly the same way, but their engagement is manifested differently in their pupil dilations. However, based on the group differences in frequency, we argue for the first option.

Given substantial variability in subjects' reading abilities (see e.g., Kuperman & Van Dyke, 2011), loquaciousness, amount of education, social status and responsibilities, as well as differences in age, gender, and motivation for participating in a psycholinguistic experiment, these differences are perhaps unsurprising. But these differences clearly indicate that general statements about loci of processing effects in the 'population' based on pupillometry data are potentially hazardous, if the present pattern of results, revealed by detailed investigation with generalized additive mixed models, turn out to be replicable in future experiments.

This conclusion is supported by an examination of the frequency effect of three subject groups. Two of these subject groups (group 1 and 3) showed a frequency effect that was the strongest after the offset of articulation. One of the groups (group 1) also showed a weak frequency effect immediately after stimulus onset. The second group showed only a weak effect of frequency, even though the members of this group, just as the others, read the words and produced them correctly. The weak frequency effect, in combination with a relatively shallow pupil response, may be indicative of short but se-

mantically shallow lexical processing (e.g., Baayen and Milin (2010), who observed the absence of a word frequency effect for fast readers, and the strongest effect of frequency for the slowest readers in self-paced reading of continuous text).

We think that for a proper understanding of lexical processing, in all its currently bewildering variability, it will be essential to consider in much more detail the vast differences in experience, motivation, socio-cultural background, as well as differences in brain morphology, that subjects bring into an experiment.

Acknowledgments

We would like to thank Dr. Fabian Tomaschek for his help with extracting the naming latencies. This research was supported by the national scholarship program Kristjan Jaak, funded and managed by Archimedes Foundation in collaboration with the Estonian Ministry of Education and Research and by an Alexander von Humboldt research award to the first and last author, respectively.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*(4412), 1289–1292.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. *arXiv preprint arXiv:1601.02043*.

Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language*, *28*(1), 1–11.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*(4), 639–645.

Fernandez, E. S. (2014). Harrell miscellaneous [Computer software manual]. (R package version 1.4)

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627–635.

Geller, J., Still, M. L., & Morris, A. L. (2015). Eyes wide open: Pupil size as a proxy for inhibition in the masked-priming paradigm. *Memory and Cognition*, *15*, 1-11.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3423), 349–350.

Hoeks, B., & Levelt, W. J. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, *25*(1), 16–26.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, *48*(3), 598–612.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 310.

Kahnemann, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, 1583–1585.

Kuchinke, L., Võ, M. L. H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*(2), 132–140.

Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of memory and language*, *65*(1), 42–73.

MATLAB. (2015). *version 8.5.0 (r2015a)*. Natick, Massachusetts: The MathWorks Inc.

Papesh, M. H., & Goldinger, S. D. (2012). Pupil-blah-metry: Cognitive effort in speech planning reflected by pupil dilation. *Attention, Perception, & Psychophysics*, *74*(4), 754–765.

R Development Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Roehm, D., Bornkessel-Schlesewsky, I., Rösler, F., & Schlesewsky, M. (2007). To predict or not to predict: Influences of task and strategy on the processing of semantic relations. *Journal of Cognitive Neuroscience*, *19*(8), 1259–1274.

Schluroff, M., Zimmermann, T. E., Freeman, R., Hofmeister, K., Lorscheid, T., & Weber, A. (1986). Pupillary responses to syntactic ambiguity of sentences. *Brain and language*, *27*(2), 322–344.

van Rij, J., Baayen, R. H., Wieling, M., & van Rijn, H. (2015). *itsadug: Interpreting time series, autocorrelated data using GAMMs.* (R package version 1.0.1)

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences*, *109*(22), 8456–8460.

Wood, S. N. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

Young, F. A., & Biersdorf, W. R. (1954). Pupillary contraction and dilation in light and darkness. *Journal of comparative and physiological psychology*, *47*(3), 264.

Zekveld, A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. *Ear and hearing*, *31*(4), 480–490.

Modeling Commonsense Reasoning via Analogical Chaining: A Preliminary Report

Joseph A. Blass (joeblass@u.northwestern.edu)
Kenneth D. Forbus (forbus@northwestern.edu)
Northwestern University, 2133 Sheridan Road,
Evanston, IL 60208

Abstract

Understanding the nature of commonsense reasoning is one of the deepest questions of cognitive science. Prior work has proposed analogy as a mechanism for commonsense reasoning, with prior simulations focusing on reasoning about continuous behavior of physical systems. This paper examines how analogy might be used in commonsense more broadly. The two contributions are (1) the idea of *common sense units*, intermediate-sized collections of facts extracted from experience (including cultural experience) which improves analogical retrieval and simplifies inferencing, and (2) *analogical chaining*, where multiple rounds of analogical retrieval and mapping are used to rapidly construct explanations and predictions. We illustrate these ideas via an implemented computational model, tested on examples from an independently-developed test of commonsense reasoning.

Keywords: Analogical Reasoning; Commonsense Reasoning; Analogical Abduction

Introduction

Consider three situations. (1) A person throws a crumpled-up piece of paper the size of an egg at another person's head. (2) is like (1), but the item thrown is an actual egg. (3) is like (1), but the item is a small, white stone of the same size. Despite these situations' similarities, you would likely interpret the first as a playful action, the second as emotionally aggressive but perhaps not too harmful, and the third as a serious act of aggression. These conclusions come quickly and easily to us, without conscious pondering. Such extremely rapid construction of explanations and predictions is a hallmark of commonsense reasoning.

Several models for commonsense reasoning have been proposed, ranging from logical reasoning using general, first-principles axioms (e.g. Davis, 1990, Lenat, 1995) to numerical simulation (e.g. Battaglia *et al*., 2013). We take analogical reasoning as a promising approach for explaining commonsense reasoning, for three reasons. First, analogical reasoning can work with partial knowledge: we may not have a fully articulated general theory of how much harm being hit by something might have, but if we have examples, we can still work with those. Second, analogical generalization provides a potential mechanism for learning probabilistic generalizations to represent experience. Third, analogy can allow a system to generate multiple inferences by importing whole relational structures from a single case, rather than requiring separate rules for each inference.

Our prior work on exploring analogy in commonsense reasoning focused on reasoning about the behavior of continuous systems (e.g. Forbus & Gentner, 1997; Forbus 2001). Here we explore how analogy might be used for commonsense more broadly. We have argued that much of human abduction and prediction might be explained by analogy over experiences and generalizations constructed from experience (Forbus, 2015). This paper explores in more detail how that might work. Specifically, we propose that multiple analogical retrievals are used to quickly elaborate a situation, providing a set of plausible explanations and predictions. We call this process *analogical chaining* (AC). The units that are retrieved might be specific situations or larger structures, such as traditional scripts (e.g. Schank & Abelson, 1977) and frames (e.g. Minsky, 1974), if they are good matches for the situation. However, we also propose that experience is factored into *Common Sense Units*, cases in the case-based reasoning sense, that are typically larger than single facts and smaller than frames or scripts[1]. A CSU consists of several facts that connect, for example, an event of a particular type with a precursor or with a potential outcome. Such cases can be useful for prediction when the precursor matches the current situation, and for explanation when the outcome matches the current situation (Forbus, 2015). Because they are smaller, they should be more easily transferrable to a wider range of situations, because they contain less non-overlapping information.

We begin by reviewing the structure-mapping models that this model is built upon, and the Cyc-derived ontology used. We describe our model, including our hypotheses about the nature of CSUs and the computational issues raised by AC. We present an experiment where a pool of CSUs are used to answer questions from the Choice of Plausible Alternatives (COPA, Roemmele *et al*., 2011) test of commonsense reasoning. We close with related and future work.

Background

Analogy is an important tool for reasoning and decision-making; we use past experiences to understand and make decisions in new situations (Markman & Medin 2002). We use Gentner's (1983) structure-mapping theory of analogical reasoning, which argues that analogy involves finding an alignment between two structured descriptions. The Structure-Mapping Engine (SME, Forbus et al. 2016) is a computational model of analogy and similarity based on structure mapping theory[2]. SME takes in two structured, relational cases (a *base* and a *target*) and computes up to three *mappings* between them. These include the correspondence between the two cases, *candidate inferences* suggested by the

[1] We are inspired by the Goldilocks Effect argument for analogical reasoning (Finlayson & Winston, 2005).

[2] See (Gentner & Forbus, 2011) for a survey of other models.

mapping, and a *similarity score* that serves as a measure of how good the mapping is. If a candidate inference involves an entity not present in the other case, that entity is hypothesized as a *skolem entity*.

Running SME across every case in memory would be prohibitively expensive, and implausible for human-scale memories. MAC/FAC (Forbus et al. 1995) retrieves cases that may be helpful for analogical reasoning from a case library, without relying on any indexing scheme. It takes in a *probe* case like those used by SME, as well as a case library of other such cases. MAC/FAC efficiently generates remindings, which are SME mappings, for the probe case with the most similar case retrieved from the case library. MAC/FAC proceeds in two stages: first, it computes dot products between content vectors of the probe and each case in the case library, a coarse approximation for scalability. Up to the three most similar cases are passed to the second stage, which uses SME to calculate similarity based on structured descriptions. Typically only one, but up to three if close, retrievals are output by MAC/FAC.

We use the Cyc ontology (Lenat, 1995) as a source of representations. The subset of contents of ResearchCyc that we use for our knowledge base contains over 110,000 concepts and over 32,000 relations, constrained by over 3.8 million facts. We extend this knowledge base to support qualitative reasoning, analogical reasoning and learning, and additional lexical and semantic information. The knowledge is partitioned into over 58,000 *microtheories*, which can be linked via inheritance relationships to form logical environments to support and control reasoning. The MiddleEarthMt or other microtheories representing fictional worlds, for example, are rarely useful in commonsense reasoning (although the converse is not true). Microtheories simplify the consideration of alternatives during reasoning.

Using ResearchCyc representations allows us to leverage the several person-centuries of work that has gone into its development, but also reduces the risk of tailorability. By using natural language inputs and someone else's representations, we reduce the chance that our results are an effect of having spoon-fed answers to our systems.

Common Sense Units

People are spontaneously reminded of similar prior situations. We further hypothesize that experience, both direct and culturally transmitted (e.g., reading, watching videos) is carved up into smaller pieces as well, and combined via analogical generalization to create probabilistic structures (via SAGE, McLure et al. 2015). These lack logical variables but can behave like rules when applied by analogy, and serve as grist for analogical reasoning about novel situations. Because they include fewer statements they are less specific (in the model theory sense), and more likely to match to a wide range of cases.

We think of CSUs as the set of facts surrounding a particular common plausible inference. For example, a CSU for reciprocity might encode simply that one agent performs a positive deed for another, which causes the other agent to perform a positive deed for the first at some future time. CSUs are intended to be smaller than situations, hence making them more compositional. We have not yet explored learning CSUs, because we first want to establish that they can be useful. The current paper provides evidence for this.

Analogical Chaining (AC)

Many prior computational models of analogical reasoning have treated analogy as a one-shot process, where a single analog is retrieved and used, or perhaps replaced with another if the first is not satisfactory. We go beyond that here by chaining analogies, i.e. using the elaboration of a situation via analogical inferences to retrieve yet more analogs, similar to how chaining in logical inference works. This is conceptually similar to Derivational Analogy (Veloso & Carbonell, 1993; differences discussed below).

AC proceeds as follows. The case library of CSUs is a stand-in for some of the commonsense knowledge a human gains over their lifetime. The current situation (the *target*) is used as a probe for MAC/FAC over the case library. If no mapping is produced, the program seeks another reminding, without cases that were rejected or previously used. If a mapping is found, any candidate inferences are asserted into an inference context, along with statements indicating what category any skolems belong to. Inferences are placed in a separate context from the case because there is no guarantee that they are correct. Another retrieval is then performed, with the probe being the union of the target and the inference context. If no information was added to the case, the previously retrieved analog is suppressed, to prevent looping. When information is added to the inference context, previously rejected CSUs are freed up to be retrieved against in case they might build off the inferences made. The process repeats until an answer has been found (for a question-answering task) or there are no more inferences to carry over into the target case (Figure 1).

There are several potential advantages to this model. Cases can be dynamically added to the case library, and can be used immediately. AC enables both inference about what is present in the case (filling in implicit relational links) as well as abductive explanations for what caused an event or predictions about what might happen next.

The strongest advantage of analogical reasoning is that, unlike logical inference, it does not require a fully articulated logically quantified theory. The difficulty in creating such theories is well-known, and seems to stem from two reasons. First, people have difficulty articulating a complete, accurate account of their reasoning. Second, their models tend to be full of gaps and unintended consequences. By contrast, reasoning by analogy from experience does not require a complete axiomatic theory of, for example, causality or human actions. It only requires examples with explanations specific to those cases. Analogical reasoning moreover is guided by what has happened, rather than what might be logically possible. To be sure, analogy can go awry as well – no powerful reasoning system with imperfect information and finite resources can always guarantee valid results. AC

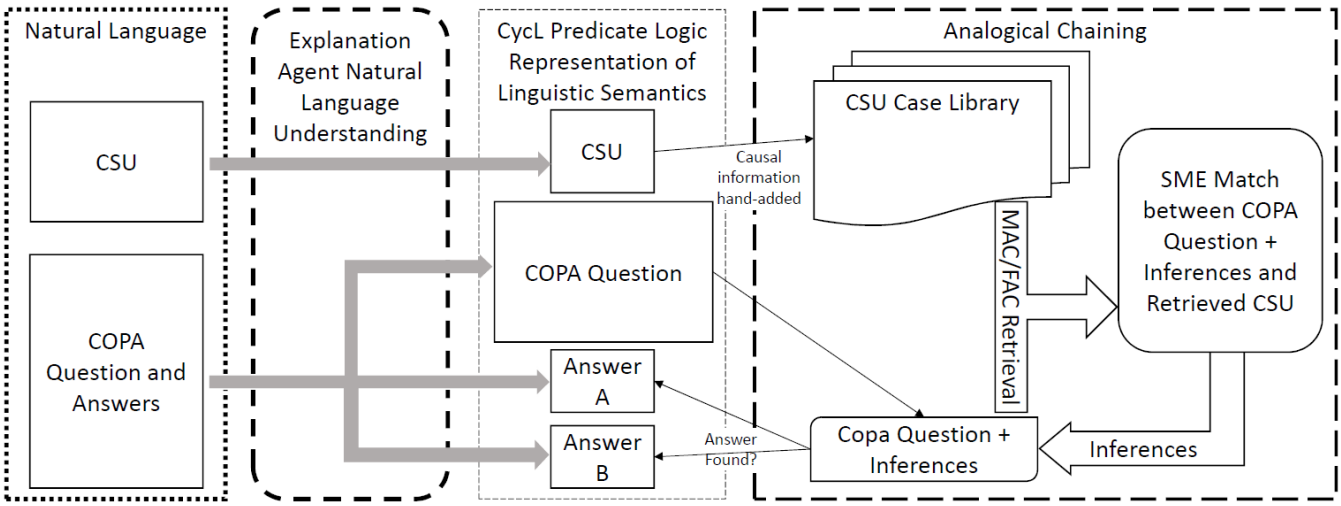


Figure 1: Analogical Chaining Workflow for answering COPA questions

should provide a compression of the inference space, both in terms of the number of inferences completed per step and fewer inappropriate branches explored, compared to logical chaining.

Simulation

Method

To explore the plausibility of these ideas, we implemented AC using the Companion cognitive architecture (Forbus *et al.* 2009). For testing, we focused on a small subset of COPA training set questions[3], and automatically encoded the questions and the majority of the CSUs via natural language understanding capabilities built into the architecture. These include six questions involving the causes and consequences of situations involving violent impacts, and a seventh question involving boiling water. These questions are shown in Figure 2. This paper uses question 461 for illustration.

```
214: The vandals threw a rock at the window.
What happened as a RESULT?
    The window [cracked / fogged up].
294: The egg splattered. What was the CAUSE of
this?
    I [dropped / boiled] it.
347: The boy got a black eye. What was the CAUSE
of this?
    The bully [mocked /punched] the boy.
370: The water in the teapot started to boil.
What happened as a RESULT?
    The teapot [cooled / whistled].
390: The truck crashed into the motorcycle on
the bridge. What happened as a RESULT?
 [The motorcyclist died / The bridge collapsed].
461: The mother called an ambulance. What was
the CAUSE of this?
 Her son [lost his cat / fell out of his bed].
496: My ears were ringing. What was the CAUSE of
this?
    I went to a [museum / concert].
```

Figure 2: The Selected COPA Questions and Answers

[3] We use training set questions here because publishing test set questions would violate the security of the test.

We created 32 CSUs designed to be relevant to the topics of the questions, plus distractors. These CSUs ranged in size from 2 to 8 facts. COPA questions are designed so that both answers are actually plausible, but one answer is always more plausible than the other. Consequently, CSUs that would contribute to incorrect answers were encoded as part of the set, as well as several CSUs irrelevant to answering the specific questions chosen (for example, that a system with a faulty component may malfunction). Sample CSUs are shown in Figure 3. Representations for 19 CSUs were almost

```
When a loved one is hurt, you call an ambulance.
(loves caller6829 person6293)
(senderOfInfo call22246 caller6829)
(communicationTarget call22246 ambulance22371)
(isa ambulance22371 Ambulance)
(isa call22246 MakingAPhoneCall)
(causes-PropProp
 (and (objectHarmed someHarm1523 person6293)
      (loves caller6829 person6293))
 (and (isa call22246 MakingAPhoneCall)
      (senderOfInfo call22246 caller6829)
      (communicationTarget call22246 ambulance22371)
      (isa ambulance22371 Ambulance)))
                          *********
Mothers love their sons (similar CSUs cover mothers & daughters)
(isa mother22349 HumanMother)
(sons mother22349 son26849)
(loves mother22349 son26849)
(causes-PropProp
 (and (isa mother22349 HumanMother)
      (sons mother22349 son26849))
 (loves mother22349 son26849))
                          *********
Falling out of bed hurts.
(isa bed2498 Bed-PieceOfFurniture)
(isa fall24789 FallingEvent)
(isa impact1953 ViolentImpact)
(objectHarmed impact1953 person22386)
(primaryObjectMoving fall24789 person22386)
(causes-PropProp
 (and (isa fall24789 FallingEvent)
     (from-UnderspecifiedLocation bed2498 person22386)
      (isa bed2498 Bed-PieceOfFurniture)
      (primaryObjectMoving fall24789 person22386))
 (and (isa impact1953 ViolentImpact)
      (objectHarmed impact1953 person22386)))
```

Figure 3: CSUs required to solve COPA question 461.

```
(communicationTarget call22246 ambulance22371)
(isa ambulance22371 Ambulance)
(isa call22246 MakingAPhoneCall)
(loves caller6829 person6293)
(senderOfInfo call22246 caller6829)
(causes-Underspecified love9172 call22246)
                        *********
(causes-PropProp
 (and (objectHarmed someHarm1523 person6293)
      (loves caller6829 person6293))
 (and (communicationTarget call22246 ambulance22371)
      (isa ambulance22371 Ambulance)
      (isa call22246 MakingAPhoneCall)
      (senderOfInfo call22246 caller6829)))
```

Figure 4: Top: the NLU-output CSU about calling an ambulance when a loved one is hurt. Bottom: the causal fact after manual editing (other facts the same)

entirely automatically generated by our NLU system, with only causal representations manually edited (described next). The remaining 13 CSUs also began as automatically processed natural language, but required more significant manual changes, due to various limitations of the NLU system, mostly involving words unknown to the system.

Causal representations automatically generated from natural language were modified by hand. The NLU system generates structurally flat causal representations, which are difficult for SME to operate over. For example, saying that someone you loved being hurt leads to calling them an ambulance results in an underspecified causal relation. That information was automatically extracted by the NLU system but not connected to the causal fact; we edited those facts to connect those relevant automatically generated facts (Figure 4). Additionally, we added facts to several CSUs indicating that a particular event was an instance of a ViolentImpact, (a new concept for our system), and removed facts which made the CSU overly specific (i.e., information that would be worn away via analogical generalization).

Since AC involves within-domain analogies, we use required partition constraints (Forbus *et al*., 2016) to restrict matching entities to be within the same categories. For example, matches with the CSU in Figure 3 had the constraint that ambulance and phone call could only be placed in correspondence with an ambulance and a phone call, respectively.

```
(causes-PropProp
 (and (isa rock2942-skolem StoneStuff)
      (isa throw2912-skolem ThrowingAnObject)
      (objectActedOn throw2912-skolem rock2942-skolem)
      (target throw2912-skolem person6293-skolem))
 (and (isa someHarm1523-skolem ViolentImpact)
      (objectHarmed someHarm1523-skolem
                    person6293-skolem)))
                        *********
(causes-PropProp
 (and (from-UnderspecifiedLocation bed2498-skolem
                                   person6293-skolem)
      (isa bed2498-skolem Bed-PieceOfFurniture)
      (isa fall24789-skolem FallingEvent)
      (primaryObjectMoving fall24789-skolem
                           person6293-skolem))
 (and (isa someHarm1523-skolem ViolentImpact)
      (objectHarmed someHarm1523-skolem
                    person6293-skolem)))
```

Figure 5: Plausible inferences for question 461: the person was hit by a rock; the person fell out of bed (correct)

For each question, the Companion read the question and answers into separate microtheories. The system read and understood the questions without human intervention. The Companion automatically filtered out the phrase asking for cause or effect, since we found that for most COPA questions only one answer is plausible regardless of which is asked for. The Companion then used AC to flesh out the question, storing the inferences in a separate microtheory.

After each extension, the Companion would check whether it had reasoned its way to one of the answers, using SME. The base normalized score (i.e. the similarity of the base/target divided by the similarity of the base with itself) measures how much of the base is covered by the match. Here, an answer is used as the base and the union of the question and inferences microtheories are used as the target. If the base normalized score of the comparison is above 0.999, all the facts in the answer have identical (but for entity tokens) corresponding facts in the reasoning microtheory, and the model selects that answer as correct.

Results

Of the seven questions selected, a Companion using AC was able to answer six correctly. Most inferences generated through chaining either helped the system find the answer or were perfectly plausible (Figure 5), although in some cases it considered at least one strange possibility before finding the right answer (Figure 6). Answering five of these six questions correctly involved chaining through the same CSU expressing that a violent impact causes harm, demonstrating that AC can use the same CSU in different contexts.

Question 461 was the only question which was not answered correctly from its raw NLU output, which included representations that prevented SME from detecting success. Specifically, in the correct answer "her son fell out of his bed," the multiple possessives "her" and "his" were interpreted as (possesses mother son) and (possesses his bed), which are reasonable. However, the coreference system did not resolve "his" to "son" (i.e., (possesses son bed)), and the CSU did not contain the first "possesses" fact (another fact in the CSU expressed the mother/son relationship), so the base normalized score of the match was not quite high enough to detect success. However, there was still the information that the son fell out of *a* bed, if not *his* bed. To verify this analysis, we removed these extra "possesses" facts, and the system was able to correctly find the answer.

Was analogical chaining necessary? Yes, since every question required two or three analogies to reach the correct answer. For example, after amending question 461 as noted

```
(causes-PropProp
    (and (isa someHarm1523-skolem Concert)
         (isa go-to35116-skolem AttendingSomething)
     (toLocation go-to35116-skolem someHarm1523-skolem)
      (performedBy go-to35116-skolem person5082-skolem)
         (loudnessOfEvent someHarm1523-skolem Loud)))
    (and (isa someHarm1523-skolem ViolentImpact)
         (objectActedOn someHarm1523-skolem
                        ear2942-skolem)
         (isa ear2942-skolem Ear)))
```

Figure 6: Implausible inference for question 461: the harm was a concert which hurt their ears

above, the system was able to find the answer only after retrieving and applying the three CSUs in Figure 2. It first postulated that a loved one had been hurt, then that it was her son, and from a fall. We suspect that most COPA questions should be answerable within three AC steps, but confirming this remains future work.

Related Work

As of this writing, three other systems have been tested on COPA, all focused on text analysis. Gordon *et al*. (2011) used Pointwise Mutual Information to evaluate how often words in the questions co-occurred with words in the answer. While their system performed significantly above chance (65.4% accuracy), it only slightly gained in accuracy as the training corpus dramatically increased, from 10^6 to 10^7 stories. Goodwin *et al*. (2012) achieved a similar performance with other textual analysis techniques (63.4% accuracy), but found that using multiple components in their analysis did not significantly improve accuracy over using only bigram co-occurrence. Luo *et al.* (2016) achieved higher accuracy (70.2%) using a large corpus to automatically extract causal relationships between concepts, then using this extracted information to determine the 'causal strength' between a question and each of its answers. While the extracted causal information appears more effective than the other two techniques, it still requires that information to be represented in the training corpus, which much of commonsense knowledge is not. Together these findings suggest that there are upper limits to text-based techniques, which argues for investigating approaches like ours that use conceptual representations. Of course, all three of these techniques were able to attempt the entire COPA test. AC will require a large case library of CSUs before we test it on the full COPA test.

Derivational analogy, as implemented in the PRODIGY architecture, similarly chains together previously known cases to derive solutions to problems (Veloso & Carbonell, 1993). It plans for a goal by analogy to plans that previously achieved a similar goal, with subgoals recursively planned for by analogy. Stored cases are indexed by and retrieved via their justifications, initials states, and goal states. Derivational analogy differs from AC as we have described it in three important ways. First, our cases are stored and retrieved without requiring any information about what they previously allowed the system to conclude. Although this means that sometimes a highly dissimilar yet nonetheless relevant case may not be retrieved by MAC/FAC, it also circumvents issues with indexing and retrieval, and enables AC to use a relevant case even when it has not been useful in similar past situations. Second, derivational analogy was specifically used to create plans to achieve goals, rather than to explain a state of affairs or predict future outcomes. It is not clear whether derivational analogy could be used for tasks that cannot be easily framed in terms of planning or problem-solving (although answering COPA questions could be framed in such a way). Finally, PRODIGY made use of both case-based and first-principles reasoning. AC does not use any first-principles reasoning at any stage.

Much AI research on commonsense reasoning has relied on formal logic and deductive inference (see Davis, 1990 and Mueller, 2014). All such approaches rely on using large numbers of logically quantified axioms. We have noted several problems with this approach, including the difficulty of constructing correct logically quantified axioms. Analogy only requires acquiring relevant cases and refining them via analogical generalization, rather than complete and correct domain theories. Furthermore, reasoning using formal logic must proceed serially: each inference rule asserts only its consequences. AC also proceeds serially, but a highly relevant case can lead to several inferences (not necessarily derivable from the same logical rule) being asserted at once, potentially reducing the number of inference steps needed.

In Explanation-Based Learning (EBL) (DeJong, 2006), a human provides a formal domain theory and examples from a domain to a system, which it uses to refine its own formal theory of that domain. AC differs in that it only requires the human to provide (in simplified natural language) cases that illustrate an underlying principle, rather than the logic of that underlying principle, which is simpler for non-experts. Also, the domain theories generated through EBL are still in logic and as such face the same drawbacks listed above.

AnalogySpace (Speer, Havasi & Lierberman 2008) used a large knowledge base of commonsense assertions in natural language to make predictions about concepts, which could then be compounded with further predictions. However, they define similarity as a linear operation over feature vectors, using a reduced-dimensionality approximation of MAC's dot-product of content vectors to retrieve relevant concepts, and do not include any measure of structural similarity. Furthermore, this work was only used to predict features of individual concepts, and it is unclear how it would scale up to explain or predict larger cases.

Though AC generates possible explanations for situations, it differs from using logic for abduction (e.g. Hobbs, 2006) since it does not require a logically quantified domain theory, and does prediction as well as explanation.

The importance of the Goldilocks Principle, using cases that are neither too small or too large in analogical matching, was highlighted by Finlayson and Winston (2005), which helped inspire our thinking about CSUs..

Conclusions and Future Work

There is already evidence that analogy is widely used in human cognition (Gentner, 2003), so it would be surprising if it were not used for commonsense reasoning. This paper has explored how that might work. We proposed Common Sense Units, intermediate-sized representations, closer to rules in size than raw experiences, but still without logical variables, as a means of encoding experience that supports flexible analogical processing. We proposed analogical chaining, the repeated use of analogies to rapidly construct explanations and predications, as a means of performing commonsense reasoning. While AC is serial at the level of applying an analog, it is parallel with respect to the application of candidate inferences within a step, thereby

being more efficient than traditional chaining with logical axioms. The bundling of common patterns of facts via CSUs also provides more focus for each inference step. CSUs and AC were used to answer COPA questions, demonstrating its potential as a model of commonsense reasoning.

We plan to explore several directions of future work. First, we plan to expand our NLU capabilities to support fully automatic construction of CSUs from natural language, rather than mixing automatic generation with some manual editing, both to reduce tailorability and to expand coverage, including crowdsourcing CSUs (c.f. Li *et al*., 2013). Extracting CSUs from larger stories via analogical generalization is, we think, a promising approach. Second, we plan to expand the reasoning techniques used for checking the validity of retrieved cases, skolem resolution, and determining when sufficient reasoning has been done. Answers to multiple-choice questions can also help guide chaining. Finally, we plan to test the expanded model on the entire COPA and other commonsense reasoning tests, such as Winograd schemas[4].

Acknowledgements

This research was supported by the Socio-Cognitive Architectures for Adaptable Autonomous Systems Program of the Office of Naval Research, N00014-13-1-0470.

References

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *PNAS*, *110*(45), 18327-18332.

Davis, E. (1990, 2014). *Representations of commonsense knowledge*. Morgan Kaufmann.

DeJong, G. (2006). Toward robust real-world inference: A new perspective on explanation-based learning. In *Machine Learning: ECML 2006* (pp. 102-113). Springer.

Finlayson, M. A., & Winston, P. H. (2005) Intermediate Features and Informational-level Constraint on Analogical Retrieval. In *Proceedings of the 27th Annual Meeting of the Cog. Sci. Society (CogSci 2005)*, Stresa, Italy. 666–671.

Forbus, K. 2001. Exploring analogy in the large. In Gentner, Holyoak, and Kokinov (Eds) *The Analogical Mind: Perspectives from Cog. Sci.* Cambridge, MA: MIT Press.

Forbus, K. (2015). Analogical Abduction and Prediction: Their Impact on Deception. *AAAI Fall Symposium on Deceptive and Counter-Deceptive Machines.*

Forbus, K., Ferguson, R., Lovett, A., & Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science.*

Forbus, K. & Gentner, D. 1997. Qualitative mental models: Simulations or memories? *QR 1997*, Cortona, Italy.

Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*(2), 141-205.

Forbus, K., Klenk, M. and Hinrichs, T. (2009). Companion Cognitive Systems: Design Goals and Lessons Learned So Far. *IEEE Intelligent Systems*, 24(4), 36-46.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Sci.*, *7*(2), 155-170.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science*, 2. 266-276.

Gentner, D., Ratterman, M. J., & Forbus, K. (1993). The roles of similarity in transfer. *Cog. Psych.,* 25, 524–575.

Gordon, A., Bejan, C., and Sagae, K. (2011) Commonsense Causal Reasoning Using Millions of Personal Stories. *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11),* August 7–11, 2011, San Francisco, CA

Goodwin, T., Rink, B., Roberts, K., and Harabagiu, S. (2012) UTDHLT: COPACETIC System for Choosing Plausible Alternatives. *Procs of the 6th Int'l Workshop on Semantic Eval.* (SemEval 2012), June 7-8, 2012, Montreal, Canada.

Hobbs, J. (2006) Abduction in Natural Language Understanding. In Horn & Ward (Eds.)., *The Handbook of Pragmatics* Blackwell Publishing Ltd, Oxford, UK.

Lenat, D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Comm. of ACM*, *38*(11), 33-38.

Li, B., Lee-Urban, S., Johnston, G., & Riedl, M. (2013). Story Generation with Crowdsourced Plot Graphs. In *Proceeding of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, Washington.

Luo, Z., Sha, Y., Zhu, K., Hwang, S., & Wang, Z. (2016). Commonsense Causal Reasoning between Short Texts. *15th Int'l Conference on Principles of Knowledge Representation and Reasoning*, Cape Town, South Africa.

Markman, A. B., & Medin, D. L. (2002). Decision making. *Stevens' Handbook of Experimental Psychology*.

McLure, M. D., Friedman, S. E., & Forbus, K. D. (2015). Extending Analogical Generalization with Near-Misses. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Austin, TX (pp. 565-571).

Minsky, M. (1974). A Framework for Representing Knowledge. Reprinted in *The Psychology of Computer Vision,* P. Winston (Ed.), McGraw-Hill, 1975.

Mueller, E. T. (2014). *Commonsense Reasoning: An Event Calculus Based Approach*. Morgan Kaufmann.

Roemmele, M., Bejan, C., & Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. *AAAI Spr. Symp. on Log. Form. of Commonsense Reasoning*, Stanford University

Schank, R.C. & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale , NJ: Earlbaum Assoc.

Speer, R., Havasi, C., & Lierberman, H. (2008). AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, Chicago, IL

Veloso, M., & Carbonell, J. (1993). Derivational analogy in PRODIGY: Automating case acquisition, storage, and utilization. In *Case-Based Learnin*g (55-84). Springer US.

[4] http://commonsensereasoning.org/winograd.html

Social Cues Modulate Cognitive Status of Discourse Referents

Kara Hawthorne (khawthor@olemiss.edu)
Department of Communication Sciences and Disorders, 304 George Hall
University of Mississippi, University, MS 38677 USA

Anja Arnhold (anja.arnhold@uni-konstanz.de)
Department of Linguistics, Post Box D 186
University of Konstanz, 78457 Konstanz GERMANY

Emily Sullivan (esulliva@ualberta.ca)
Department of Communication Sciences and Disorders, Corbett Hall
University of Alberta, Edmonton, AB T6G 2G4 CANADA

Juhani Järvikivi (jarvikiv@ualberta.ca)
Department of Linguistics, 4-55 Assiniboia Hall
University of Alberta, Edmonton, AB T6G 2E7 CANADA

Abstract

We use visual world eye-tracking to test if a speaker's eye gaze to a potential antecedent modulates the listener's interpretation of an ambiguous pronoun. Participants listened to stories that included an ambiguous pronoun, such as "The dolphin kisses the goldfish… He…." During the pre-pronominal context, an onscreen narrator gazed at one of the two characters. As expected, participants looked more at the subject character overall. However, this was modulated by the narrator's eye gaze and the amount of time the participant spent looking at the gaze cue. For trials in which participants attended to the narrator's eye gaze for > 500ms, participants were significantly more likely to interpret the pronoun as referring to the object if the narrator had previously looked at the object. Results suggest that eye gaze – a social cue – can temper even strong linguistic/cognitive biases in pronoun resolution, such as the subject/first-mention bias.

Keywords: Ambiguous pronoun resolution, visual world paradigm, eye-tracking, reference, social cues, eye gaze.

Introduction

In this paper, we test if a social cue – the speaker's eye gaze to potential referents – impacts the listener's interpretation of ambiguous pronouns in a discourse context where there are two characters who could serve as the pronoun's antecedent. For example, in "The dolphin kisses the goldfish behind the lake. He…" *he* could refer to either the dolphin or the goldfish. Unlike previous work on the effect of a narrator's attention on offline pronoun interpretation (Nappa & Arnold, 2014), we manipulate the narrator's eye gaze during the pre-pronominal context, which is the time period in which speakers naturally look at their intended referent (Griffin & Bock, 2000). We test if the narrator's eye gaze modulates the listener's assumptions about the narrator's focus of attention by using the visual world eye-tracking paradigm to monitor online processing of ambiguous pronouns.

Linguistic and Cognitive Biases

Pronoun resolution is constrained by the relative salience or prominence of potential antecedents in the discourse representation (e.g., Gordon, Grosz, & Gilliom, 1993; Gundel, Hedberg, & Zacharski, 1993). Salience modulates the listener's attention to potential referents and the degree to which s/he expects each referent to be talked about in the upcoming discourse (Ariel, 1990; Arnold, 1998).

There are several linguistic and cognitive biases that rapidly affect referent salience during online pronoun processing. Linguistic factors include pronoun gender (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000), verb type (Koornneef & Van Berkum, 2006; Pyykkönen & Järvikivi, 2010), and parallel syntactic structure and syntactic function. Subjecthood, in particular, strongly increases the salience of a discourse entity (Järvikivi, Hyönä, Bertram, & Van Gompel, 2005; Kaiser & Trueswell, 2008).

Perhaps the most widely known cognitive bias concerns the order in which the referents are introduced in the pre-pronominal context: first-mentioned entities are preferred over later-mentioned entities. This finding has been replicated often since Gernsbacher & Hargreaves (1988; e.g., Carreiras, Gernsbacher, & Villa, 1995), including in visual world studies of pronoun resolution (Arnold, et al., 2000; Järvikivi et al., 2005). Unlike in languages with freer word order (Järvikivi et al., 2005; 2014), subjecthood and first-mention are difficult to tease apart in English; we will primarily use the term *subject bias* in this paper, acknowledging that order-of-mention and syntactic function both contribute to the observed effects.

Social Biases

In addition to linguistic and cognitive effects, social cues impact discourse processing (e.g., Van den Brink et al., 2012; Jiang & Zhou, 2015) and conversational success. For

example, two people engaged in a cooperative task are much slower when they are not able to use social cues such as pointing, eye gaze, and head nodding (Clark & Krych, 2004). Social cues also impact pronoun resolution: a coreferential gesture at pronoun onset tempers the subject bias when each of the potential antecedents has previously been associated with a particular location in the speaker's gestural space (Goodrich & Hudson Kam, 2012).

Eye gaze is a salient way for a speaker to signal their attention to the listener (Langton, Watt, & Bruce, 2000). Interlocutors attend closely to each other's faces (Argyle & Cook, 1976), and a speaker will even restart an utterance if the listener is not visually attending (Goodwin, 1981). In production, speakers fixate characters before they name them when describing an image (Griffin & Bock, 2000), so attending to the speaker's eye gaze could have processing payoffs. For example, Hanna and Brennan (2007) report that listeners can use a speaker's gaze to figure out which object in a hidden array the speaker is referring to even before the speaker has reached the point of linguistic disambiguation. Moreover, the attentional effects of eye gaze are reflexive and occur even when eye gaze is manipulated to not be an informative cue (e.g., Firesen & Kingstone, 1998).

Only recently have researchers begun to explore the role of eye gaze in pronoun resolution. Nappa and Arnold (2014) tested the influence of several social cues in the moment when the pronoun is heard. They found that when a narrator turns her head and looks at a character while producing an ambiguous pronoun (with or without a pointing gesture), listeners were more likely to interpret the pronoun as the character being deictically cued.

When a speaker turns to look at a character right at the moment of pronoun production, it is perhaps not surprising that it influences pronoun interpretation, since the speaker is directly highlighting a potential referent. However, gaze cues are typically less overt, and they usually occur during the preceding discourse context rather than during the pronoun itself. Griffin and Bock (2000) report that speakers look at a character close to a full second before referring to it with a full noun phrase. Speakers look at the referent before producing a pronoun as well, though at a somewhat reduced rate (van der Meulen, Meyer, & Levelt, 2001).

While eye gaze serves as a visual cue that increases the salience of one potential referent, recent evidence suggests that not all visual cues have such an effect on ambiguous pronoun interpretation. Arnold and Lao (2015, Exp. 2) had participants listen to stories of two characters, e.g., "Birdy picked apples with Doggy near the farmhouse. He…." while they briefly (200ms) flashed a halo around one of the characters. This cue did not affect participants' gaze behaviour or their antecedent selection preferences. A further study (Järvikivi & Pyykkönen-Klauck, submitted), shows that absence of one of the referents at the pronoun onset (i.e., if a potential referent had walked out of the visual scene) did not affect adult listeners' pronoun resolution preferences. This suggests that visual cues that are coincidental with linguistic information but that are not social – in other words, cues that are not indicative of the speaker's intentions or attention to discourse referents – do not automatically impact language comprehension.

In this paper, we ask if a speaker's eye gaze during the pre-pronominal context influences how listeners interpret an ambiguous pronoun.

Methods

Participants

Participants were 86 native English speakers. Data from additional participants were excluded because of poor calibration (n = 27), corrupted results file (n = 5), or experimenter/equipment/participant error (n = 1/8/2).

Materials

Experimental stimuli consisted of 20 mini-stories involving one animal character performing an action on the other at a particular location (Table 1), as well as a visual display (Figure 1). The visual display contained both characters, the location, and a hedgehog narrator, who introduced herself as such at the beginning of the experiment (cf. Staudte & Crocker, 2011 for evidence that listeners attend to an artificial speaker's gaze similarly to a human's gaze).

Table 1: Example mini-story.
The letters (a)-(d) represent the four experimental conditions.

Sentence	Audio	Narrator's Gaze
Intro (a,b)	There are the dolphin and the goldfish.	forward
Intro (c,d)	There are the goldfish and the dolphin.	forward
Action (a)	The dolphin kisses the goldfish	**dolphin (*subject*)**
Action (b)	The dolphin kisses the goldfish	**goldfish (*object*)**
Action (c)	The goldfish kisses the dolphin	**dolphin (*object*)**
Action (d)	The goldfish kisses the dolphin	**goldfish (*subject*)**
Location	behind the lake.	forward
Pronoun	He wants to play on the playground.	forward
Probe	Who wants to play on the playground?	n/a

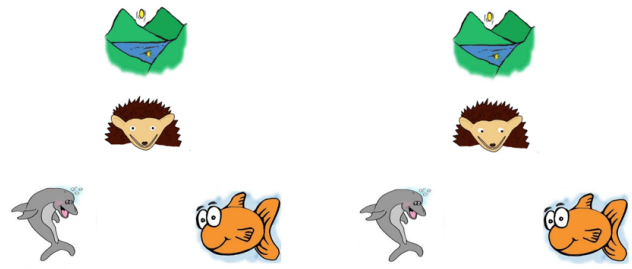


Figure 1. Example visual display. The animal in the center of each image is the narrator throughout the experiment. The image on the left appeared during the intro, location, and pronoun sentences (see Table 1). The image on the right, in which the narrator is gazing at the goldfish, appeared only during the action sentence. During the probe question, only the two animal characters (e.g., the dolphin and the goldfish) appeared on the screen.

Each story began with an introduction to the two characters. The subject of the action sentence was always named first to control for effects of the first-mention and subject biases. Next was the action sentence, in which one animal performed an action on the other. This sentence was

manipulated in two ways – which animal was the subject of the sentence (e.g., dolphin or goldfish) and which animal was gazed at by the narrator (subject or object) – to create four versions of the story (a-d, Table 1). After the action, the narrator's gaze returned to the front for the location sentence; the location was mentioned to draw participants' eyes away from the animals before the ambiguous pronoun in the pronoun sentence. Finally, the participants heard a new voice asking for an overt judgment on the referent of the pronoun (the probe sentence).

An additional ten mini-stories were recorded as filler items. The fillers were structurally the same as the experimental stories, except they did not have an ambiguous pronoun. Instead, one of the characters was referred to by name. The named animal was the subject of the action sentence half of the time and the object of the action sentence half of the time.

The stories were recorded by a 21-year-old female, and the probe questions were recorded by a 19-year-old male. Both were native English speakers. The speakers were asked to read the stories in a happy, animated voice, and care was taken that neither animal character was prosodically more prominent. Recordings were done in a sound-attenuated booth using a head-mounted CountryMan microphone and Korg MR-2000S Studio Recorder. Each story was recorded individually, and a 1-second pause was inserted at the sentence boundaries.

The visual displays were created using Adobe Photoshop CS5.1 software. Most of the images were previously used in Pyykkönen et al. (2010) and Järvikivi et al. (2014) and others were drawn by hand to match the style of the existing images. Each image fit into an area of 426 by 341 pixels. The images were counterbalanced, such that the subject appeared on the left side of the screen half of the time.

Procedure

The session began with a brief familiarization to the animals using Microsoft Office PowerPoint. Each animal was displayed one at a time, and the participant was asked to label it. If the participant provided a label different from the one used in the mini stories, s/he was told the label that would be used in the experiment.

The experiment used eye-tracking and the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Arnold et al., 2000). The experimenter first calibrated the eye-tracking equipment. Next, the narrator hedgehog appeared at the center of the screen and introduced herself by saying "Hi, my name is Hailee! I'm going to tell you some stories about the animals you just saw. Are you ready?" This was done so that it was clear that the hedgehog was the one telling the stories.

A drift correction was performed before each trial to ensure that the equipment remained properly calibrated. Then the participant heard one of the stories while his or her eye gaze was tracked. After the probe question, the participant was asked to indicate which animal was referred to in the pronoun sentence by pressing a key on the keyboard.

Equipment

Participants were tested on either an EyeLink 1000 or 1000+ eye-tracker. The experiment was run using Experiment Builder (SR-Research Ltd) and a 500 Hz sampling rate. Stimuli were played through Bose SoundLink Mini speakers, and testing was done in a quiet room.

Design

Each participant was assigned to one of four experimental lists, with twenty experimental and ten filler trials per list. Each list included each story in one of the four conditions (e.g., 2a-d, Table 1). The lists were counterbalanced, so that the narrator's gaze was on the subject and object of the action sentence for an equal number of trials. In order to maximize the number of trials in which the participant noticed the narrator's change of eye gaze during the action sentence, we included all eligible participants, even though the number of participants on each list varied from 20-25.

Results

Offline responses and eye-tracking data from all 1718 trials were analyzed with linear mixed effects modelling, using *R* (R Core Team, 2013) and the *glmer* and *lmer* functions from the package *lme4* (Bates, Maechler, & Bolker, 2012). For all analyses, the best fit model was determined using backwards stepwise model comparisons. Models were compared with likelihood ratio tests; only factors that significantly improved the model fit at a $p < .05$ level were retained for the fixed and random intercept effects (see Baayen, 2008, Bates et al. 2015). Random slopes were checked but omitted from our final analyses due to convergence errors.

Gazed at role (whether the narrator looked to the subject or object character) and earlier attention to the narrator's gaze (how long the participant looked at the narrator during the action sentence) were the factors of interest. Trials were considered to have no attention to the narrator if the participant looked at the narrator < 200ms during the action sentence (57% of trials). If the participant looked at the narrator between 200 and 500ms during the action sentence, this was considered short attention (25% of trials), while looks longer than 500ms were considered long attention (18%). Results from offline responses and eye-tracking analyses are summarized in Tables 2 and 3. For the intercept, "no," "short," and "long" indicate duration of fixation to the narrator and "S" and "O" indicate whether the narrator was gazing at the subject or object animal. The intercept values were releveled and models were re-run to examine all possible comparisons. Redundant comparisons have been omitted.

Offline Responses

The dependent variable for the offline analysis was whether the participant selected the subject or object character in response to the probe question. Responses are presented in Figure 2.

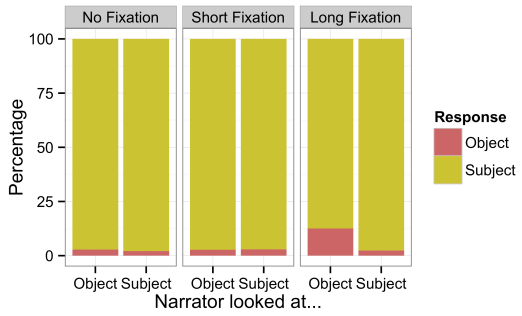


Figure 2. Percentage of object and subject responses, split by participants' earlier attention to the narrator.

Data were evaluated with binomial generalized linear mixed effects models. A model containing an interaction between gazed at role (subject versus object) and earlier attention to narrator's gaze (long, short, no) significantly outperformed a model containing only main effects ($\chi 2(1) = 7.43$, $p = .024$), so the full model was retained as the final model. When comparing the short versus no attention to narrator trials, there were no interaction or simple effects (p-values > .1). Results from the long versus short attention to narrator trials are presented in Table 2. If the participant had previously attended for a long time to the narrator while the narrator was looking to the object, s/he was significantly more likely to make an "object" response than if the narrator was looking at the subject or if the participant had only paid short attention to the narrator.

Table 2. Response results for the model: response ~ GazedAtRole* EarlierAttentnToNarr + (1|participant) + (1| item) + (1|trial).

Offline Responses: Short vs. Long Attention to Narrator				
Fixed Effect	**Intercept**	**Estimate (*SE*)**	***z*- value**	***p*-value**
2-way interaction	short, S	-2.05 (.87)	-2.34	.019
GazedAtRole	short, S	.060 (.62)	-.10	n.s
GazedAtRole	long, S	-1.99 (.62)	-3.22	.0013
EarlierAttentnToNarr	short, S	.34 (.69)	.49	n.s
EarlierAttentnToNarr	short, O	-1.71 (.55)	-3.14	.0017

Eye-Tracking Analysis

Eye-tracking data for each trial were aggregated into 300ms windows, starting at the onset of the ambiguous pronoun through 1500ms post-onset. Aggregation mitigates the auto-correlation that is inherent to time-course data. Time window was included as a fixed factor in the models, along with gazed at role (subject vs. object) and previous attention to the narrator's gaze cue (long, short, no). The dependent variable was the logit transformed proportion of looks to the subject divided by the proportion of looks to the subject plus object. This allows us to look at the strength of the subject bias while controlling for different amounts of time spent looking at the narrator and location. Data points in which the participant did not look at either the subject or object during the time window were dropped.

There were no three way interactions of time by gazed at role by earlier attention to narrator's gaze (p-values > .1). We next tested a model with all 2-way interactions. There were no gazed at role by attention to narrator's gaze or time by gazed at role interactions, so these were checked and removed one at a time (p-values > .1).

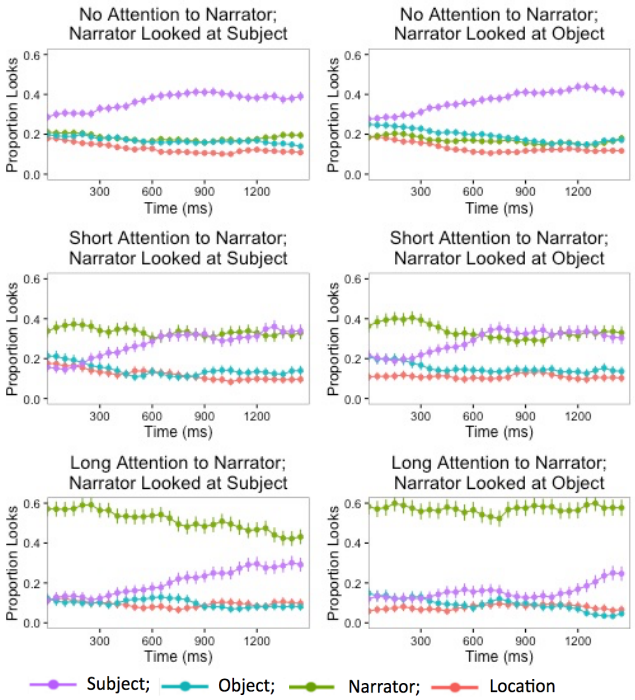


Figure 3. Proportion of participant looks to the interest areas from pronoun onset.

The final model contains a main effect of gazed at role, with participants spending less time looking at the subject relative to the subject plus object for trials in which the narrator had previously gazed at the object (Figure 3). This suggests that the narrator's earlier eye gaze cue influenced participants' eye movements after the onset of the ambiguous pronoun. The model also contains an interaction of time by earlier attention to the narrator's gaze, suggesting that the participants' previous attention or lack of attention to the narrator influenced the timing with which they settled on a referent for the ambiguous pronoun.

The effects of time for no, short, and long attention to narrator's gaze are presented in Table 3. Simple effects of attention to narrator are not included, since they were not significant at any time window. However, looking behavior across time was different across these conditions. For trials with no earlier attention to the narrator, participants looked increasingly more at the subject through the 300-599ms window, then leveled off. For trials with short attention to the narrator, looks to the subject leveled off even earlier, with no significant differences in proportion looks to the subject for 300-599ms with any subsequent time windows. The proportion of looks to the subject increased most slowly

for trials with long earlier attention to the narrator, with significant differences between 600-899ms and the subsequent time windows.

Table 3. Response results for the model: Looks to S/(S+O) ~ GazedAtRole + EarlierAttentnToNarr *Time + (1|participant) + (1| item) + (1|trial).

Eye-Tracking Results: Effects of GazedAtRole for adjacent time windows, split by degree of attention to the narrator's earlier gaze cue			
Fixed Effect	**Intercept**	**Est. (*SE*)**	***t*-value**
GazedAtRole	long	-.18 (.09)	-2.03
NO ATTENTION TO NARRATOR'S GAZE CUE			
Time (vs 300-599ms)	0-299ms	.50 (.17)	2.85*
Time (vs 600-899ms)	0-299ms	.78 (.17)	4.51*
Time (vs 900-1199ms)	0-299ms	.94 (.17)	5.46*
Time (vs 1200-1500ms)	0-299ms	.99 (.17)	5.71*
Time (vs 600-899ms)	300-599ms	.39 (.17)	1.68
Time (vs 900-1199ms)	300-599ms	.45 (.17)	2.65*
Time (vs 1200-1500ms)	300-599ms	.50 (.17)	2.92*
Time (vs all later windows)	600-899ms	--	*all < 2*
Time (vs 1200-1500ms)	900-1199ms	.05 (.17)	0.30
SHORT ATTENTION TO NARRATOR'S GAZE CUE			
Time (vs 300-599ms)	0-299ms	.88 (.30)	2.96*
Time (vs 600-899ms)	0-299ms	1.25 (.29)	4.35*
Time (vs 900-1199ms)	0-299ms	1.16 (.29)	3.99*
Time (vs 1200-1500ms)	0-299ms	1.21 (.28)	4.24*
Time (vs all later windows)	300-599ms	--	*all < 2*
Time (vs all later windows)	600-899ms	--	*all < 2*
Time (vs 1200-1500ms)	900-1199ms	.05 (.28)	0.18
LONG ATTENTION TO NARRATOR'S GAZE CUE			
Time (vs 300-599ms)	0-299ms	.56 (.43)	1.32
Time (vs 600-899ms)	0-299ms	.85 (.42)	2.02*
Time (vs 900-1199ms)	0-299ms	1.66 (.41)	4.02*
Time (vs 1200-1500ms)	0-299ms	1.82 (.41)	4.48*
Time (vs 600-899ms)	300-599ms	.28 (.41)	.69
Time (vs 900-1199ms)	300-599ms	1.10 (.40)	2.71*
Time (vs 1200-1500ms)	300-599ms	1.25 (.40)	3.15*
Time (vs 900-1199ms)	600-899ms	.81 (.39)	2.06*
Time (vs 1200-1500ms)	600-899ms	.97 (.39)	2.49*
Time (vs 1200-1500ms)	900-1199ms	.16 (.38)	0.41

Visual inspection of Figure 3 reveals a potential effect of earlier attention to the narrator on the proportion looks to the narrator after the pronoun onset, despite the fact that the location was mentioned last before the pronoun. To examine this effect, we conducted a secondary analysis in which the dependent variable was logit transformed proportion looks to the narrator, aggregated into 300ms windows.

There were several three way interactions of time by gazed at role by earlier attention to the narrator (*t*-values > 2), so the 2-way interactions of gazed at role by earlier to narrator were tested separately for each time window. There were no 2-way interactions or effects of gazed at role for the 0-1199ms time windows (all *t*-values < 2). The proportion of looks to the narrator during each of those time windows was greater the longer the participant had attended to the narrator during the action sentence (long > short > no). For the 1200-1500ms window, there was a 2-way interaction of gazed at role by earlier attention to the narrator. Participants who had previously paid long attention to the narrator while the narrator gazed at the object looked more to the narrator during the 1200-1500ms window than if the narrator had looked at the subject.

Discussion

Results suggest that a narrator's eye gaze to a character impacts the listener's resolution of a subsequent ambiguous pronoun, but that the effect is modulated by the listener's attention to the narrator's gaze. As predicted, when the narrator looked at the character that served as the grammatical object, the listener was more likely to interpret the pronoun as referring to the object, but only if the participant had attended closely to the narrator's gaze cue. In contrast to previous work, in which the narrator both looked at and turned her head toward one of the referents during production of the pronoun itself (Nappa & Arnold, 2014), we found (1) that eye gaze alone and (2) that an eye gaze cue that occurs during the pre-pronominal discourse context are sufficient to temper the subjecthood and first-mention biases. The latter is an important finding, since the timing with which the narrator gazed at one of the potential referents in the present study closely approximates the time period in which speakers are likely to provide such cues in a real discourse context (Griffin & Bock, 2000). Thus, a social cue – the speaker's gaze to a potential referent *before the listener even hears the pronoun* – can modulate other linguistic and cognitive biases at the point of pronoun disambiguation.

Another notable finding relates to attention to the narrator herself after the pronoun was produced. The longer the participant attended to the narrator during the time when her eyes were cueing one of the potential referents, the more the participant continued to look at the narrator after the onset of the pronoun, perhaps anticipating further informative eye movements that could help disambiguate it. Indeed, for trials in which the participant paid long attention to the narrator's gaze cue, the participant looked even longer at the narrator after pronoun onset if the narrator had previously looked at the object (vs. the subject), even though the location was the last item mentioned before the pronoun. Thus, participants appear to particularly look to the narrator for 'help' when she had gazed at the object and the participant was therefore more likely to be entertaining the object as a potential referent for the pronoun.

The present results, together with other work on social and other visual cues to pronoun resolution (Arnold & Lao, 2015; Järvikivi & Pyykkönen-Klauck, submitted; Nappa & Arnold, 2014) suggest that visual context impacts discourse processing, but only when that visual information is relevant to language comprehension. In that vein, many studies have demonstrated that the visual environment can help to rapidly resolve temporary referential ambiguities in sentences, even overriding linguistically-based parsing preferences (e.g., Tanenhaus et al., 1995; Chambers, Tanenhaus, & Magnuson, 2004; Knoeferle & Crocker, 2006). The present study shows that listeners are sensitive to visual cues in reference resolution as well, at least when the cue is social (i.e., eye gaze) and is therefore potentially informative about the speaker's intentions. Social visual cues can increase the salience of a potential referent even when the cue occurs in the pre-pronominal discourse context.

Acknowledgments

Thank you to University of Alberta Support for the Advancement of Scholarship (SAS/EFF) and Undergraduate Research Initiative (URI) for funding that supported this project. Thank you also to Lauren Rudat and the other members of the Centre for Comparative Psycholinguistics.

References

Argyle, M. & Cook, M. (1976). *Gaze and mutual gaze.* Cambridge, MA: Cambridge University Press.

Arnold, J. E. (1998). *Reference form and discourse patterns* (Doctoral dissertation). Stanford University, California.

Arnold, J.E., Eisenband, J.G., Brown-Schmidt, S., & Trueswell, J.C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition, 76*, B13–B26.

Arnold, J. E. & Lao, S. Y. C. (2015). Effects of psychological attention on pronoun comprehension. *Language, Cognition and Neuroscience, 30*(7), 832-852.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious Mixed Models, Retrieved from http://arxiv.org/abs/1506.04967

Carreiras, M., Gernsbacher, M. A., & Villa, V. (1995). The advantage of first mentioned in Spanish. *Psychonomic Bulletin Review, 2,* 124-129.

Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *J. of Psycholinguistic Research, 19,* 245-264.

Friesen, C. K. & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin and Review*, *5*(3), 490-495.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *J. of Memory and Language, 27*, 699–717.

Goodrich Smith, W. & Hudson Kam, C. L. (2012). Knowing 'who she is' based on 'where she is': The effect of co-speech gesture on pronoun comprehension. *Language and Cognition, 4*(2), 75-98.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science, 17,* 311-347.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psych. Science*, *11*(4), 274-279.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69,* 274-307.

Hanna, J. E. & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *J. of Memory and Language*, *57*(4), 596-615.

Järvikivi, J., Van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. *Psych. Science, 16,* 260-264.

Järvikivi, J., & Pyykkönen- Klauck (submitted). Visual context influences the representation of discourse referents in 4-year-olds.

Järvikivi, J., Pyykkönen-Klauck, P., Schimke, S., Colonna, S., & Hemforth, B. (2014). Information structure cues for 4-year olds and adults: Tracking eye movements to visually presented anaphoric referents. *Language, Cognition and Neuroscience*, 29, 877- 892.

Jiang, X. & Zhou, X. (2015). Who is respectful? Effects of social context and individual empathic ability on ambiguity resolution during utterance comprehension. *Frontiers in Psychology*, *6*, 1588.

Kaiser, E. & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes, 23,* 709-748.

Konopka, A. E. (2012). Planning ahead: How recent experience with structures and words changes the scope of linguistic planning. *J. of Memory and Language*, *66*(1), 143-162.

Koornneef, A. W. & Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *J. of Memory and Language, 54*, 445-465.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2015). lmerTest: Tests in linear mixed effects models. R package version 2.0-29.

Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50-59.

Nappa, R. & Arnold, J. E. (2014). The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension. *Cognitive Psych.*, *70*, 58-81.

Pyykkönen, P. & Järvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psych., 57, 5-16.*

Snodgrass, J. G. & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Behavior Research Methods, Instruments, & Computers*, *28*(4), 516-536.

Staudte, M. & Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human–robot interaction. *Cognition*, *120*(2), 268-291.

Van den Brink, D., Van Berkum, J. J., Bastiaansen, M. C., Tesink, C. M., Kos, M., Buitelaar, J. K., & Hagoort, P. (2012). Empathy matters: ERP evidence for inter-individual differences in social language processing. *Social Cognitive and Affective Neuroscience*, *7*(2), 173-183.

Van der Meulen, F. F., Meyer, A. S., & Levelt, W. J. (2001). Eye movements during the production of nouns and pronouns. *Memory & Cognition*, *29*(3), 512-521.

Hysteresis in Processing of Perceptual Ambiguity on Three Different Timescales

Marieke M. J. W. van Rooij (m.vanrooij@bsi.ru.nl)
Behavioural Science Institute,
Radboud University, Nijmegen, the Netherlands

Harald Atmanspacher
Collegium Helveticum, ETH Zürich,
University of Zürich, Switzerland

Jürgen Kornmeier
Institute for Frontier Areas of Psychology and Mental Health,
Eye Center, Medical Center - University of Freiburg, Freiburg, Germany
Faculty of Medicine, University of Freiburg, Germany

Abstract

Sensory information is a priori incomplete and ambiguous. Our perceptual system has to make predictions about the sources of the sensory information, based on concepts from perceptual memory in order to create stable and reliable percepts. We presented ambiguous and disambiguated lattice stimuli (variants of the Necker cube) in order to measure a hysteresis effects in visual perception. Fifteen healthy participants observed two periods of ordered sequences of lattices with increasing and decreasing ambiguity and indicated their percepts, in two experimental conditions with different starting stimuli of the ordered sequence. We compared the stimulus parameters at the perceptual reversal between conditions and periods and found significant differences between conditions and periods, indicating memory contributions to perceptual outcomes on three different time scales from milliseconds over seconds up to lifetime memory. Our results demonstrate the fruitful application of physical concepts like hysteresis and complementarity to visual perception.

Keywords: perception, ambiguous figures; hysteresis, memory-effects

Introduction

The environmental information available to our senses is incomplete and to varying degrees ambiguous. As a result, our perceptual system needs to disambiguate and interpret this information in order to construct fast and reliable solutions to all types of sensory input of whatever quality (the 'visual inference problem', von Helmholtz, 1911). Accordingly, perception has been widely discussed as a weighting of sensory (bottom-up) information with memorized (top-down) concepts in terms of Bayesian probability estimation (e.g., Kersten, Mamassian, & Yuille, 2004). The contribution of sensory and memory information to the perceptual outcome may thus change as a function of the quality of the sensory information and probably also of the availability of fitting memory contents (e.g., Kornmeier & Mayer, 2014). However, the term "memory" encompasses information from the immediate perceptual history up to long-term memory and it is unclear exactly how memory contents from different time scales contribute to the perceptual outcome.

In the present study, we investigated perception of the ambiguous "Necker lattice" (Kornmeier, Heinrich, Atmanspacher, & Bach, 2001; Kornmeier & Bach, 2004), a variant of the classical Necker cube (Necker, 1832), and lattices variants with different degrees of disambiguation in order to quantify the contributions of sensory information and memory from three different time scales.

Hysteresis and Ambiguous Figures

During prolonged observation of an ambiguous figure, such as the Necker lattice, our perception becomes unstable and suddenly alternates between different interpretations (e.g. the two different 3D perspectives of the lattice), although the stimulus information stays unchanged. Luminance changes of the apparent lattice back layer can disambiguate the stimulus and stabilize the observers' percepts. Ambiguous figures are popular stimuli to study the respective bottom-up and top-down contributions to the resolution of sensory ambiguity (e.g., Long & Toppino, 2004) and their physiological correlates (e.g., Kornmeier & Bach 2012).

In the present study we focus on *hysteresis* effects in the context of ambiguous figure perception. Hysteresis is an elementary form of history-dependent behaviour that is frequently observed in physical systems (first demonstrated by Ewing, 1881). It describes *multistable* system-behaviour where, depending on the direction of parameter change, the system can be in one of several states. Specifically, transitions back to a

previous state lag behind the change in parameters. Hysteresis behaviour can be found in visual perception, e.g., in the perception of ambiguous figures and disambiguated variants thereof (e.g., Fender & Julesz, 1967; Fisher & Ciuffreda, 1989; Hock, Kelso, & Schöner, 1993).

Observation of hysteresis depends on preservation of history in parameter change and the presence of perceptual reversals in two directions. Figure 1 displays a series of stimuli that is designed to elicit two perceptual reversals. It starts with an unambiguous "front-side right bottom" view lattice "S_{FRB}" with maximally reduced luminance m of the apparent lattice back layer (Figure 1, left). Linearly increasing m with each stimulus results in the most ambiguous lattice with iso-luminant front and back layers (Figure 1, third to left). Again linearly decreasing m to a maximally reduced back-layer luminance, results in the alternative unambiguous "front-side left top" view lattice "S_{FLT}" (Figure 1, middle). This procedure is then reversed via the ambiguous lattice (Figure 1, third to right) back to the unambiguous front-right-bottom view S_{FRB} (Figure 1, right).

Different Time-Scales of Memory

Classic Hysteresis Effect Presenting a series of lattices as described above, perceptual reversals are expected to take place around the ambiguous lattice. Assuming the absence of any type of influence of perceptual history on perception, the order of stimulus presentation should have no effect on the critical stimulus parameter for perceptual reversals. However, the presence of hysteresis implies that perception will reverse from P_{FRB} (front-right-bottom view, cf. stimulus S_{FRB}) to P_{FLT} (front-left-top view, cf. stimulus S_{FLT}) at a certain back-layer luminance value $m_{critical,1}$, which will be different from value $m_{critical,2}$, when moving back from P_{FLT} to P_{FRB}. In other words, the second perceptual reversal depends on the first perceptual reversal. We plan to quantify this classical hysteresis effect with a measure called "hysteresis distance" = $m_{critical,1} - m_{critical,2}$, and hypothesise this distance to be significantly different from zero.

Additionally, If multiple series are presented to a single person, his/her aggregated stimulus perception can be expressed as a sigmoidal probability function of the stimulus morphing parameter m, and the function's inflection point at the probability $p = 0.5$ can be used as an estimate of the aggregated critical morphing parameter value $\bar{m}_{critical}$ (Figure 3; see method section for more details about the fitting procedure). These critical values can subsequently be used to understand memory effects on three different timescales.

Immediate and Intermediate Memory Effects Hysteresis is a hallmark characteristic behaviour of nonlinear dynamical systems (with its own set of implications for understanding the behaviour of such systems, e.g., see Kaplan & Glass, 2012). In the context of visual perception, measures of hysteresis can quantify the influence of perceptual memory on the current percept. However, contrasting two reversals in a series of stimuli containing two periods of increasing and decreasing ambiguity (e.g., S_{FLT} => S_{Amb} => S_{FRB} => S_{Amb} => S_{FLT}), may reflect a mixture of memory effects on two different time scales: (1) An *immediate memory effect*, reflecting the influence of immediately preceding stimuli within one period and (2) an *intermediate memory effect*, reflecting the influence of period 1 (e.g., S_{FLT} => S_{Amb} => S_{FRB}) on the perceptual dynamics in period 2 (e..g, S_{FRB} => S_{Amb} => S_{FLT}). In order to fully disentangle immediate from intermediate memory effects, we tested several other contrasts.

Long-Term Memory Effects Long-term perceptual memory can also have an influence on the perception of an ambiguous stimulus. In the case of the Necker cube perception is biased towards the front-side right bottom perceptual interpretation (e.g. Sundareswara & Schrater, 2008; see also data from page 485 in Washburn, Mallay, & Naylor, 1931), reflecting a general view-from-above preference, which is also demonstrated in other contexts (e.g., Troje & McAdam, 2010) and which may be based on long-term perceptual memory. We will quantify the influence of this long term memory in the following way: The hysteresis effect should be manifested as a linear shift of the two sigmoidal perceptual probability functions coming from the two stimulus presentation orders in opposite directions along the abscissa (representing the morphing parameter). Moreover, if the a priori probabilities of the two Necker lattice interpretations are equal, i.e. in the absence of any perceptual bias, we should expect equal distances of two sigmoids' inflection points from m_{Amb} (representing the Necker lattice with iso-iluminant layers). Any deviation from this symmetry should indicate inequality of the two Necker lattice interpretations, which would allow quantification of long-term memory effects.

Method

Participants Fifteen participants (9 women, 5 men, 1 unknown, $M = 27.4 \pm 7.9$ years old) participated in an ambiguous figures experiment designed to measure memory effects such as hysteresis on perceptual reversals. Previous to participating, all participants were asked for their consent and given instructions about the procedure. The study was performed in accordance with the ethical standards laid down in the Declaration of Helsinki (World Medical Association, 2000) and approved by the ethics committee of the University of Freiburg.

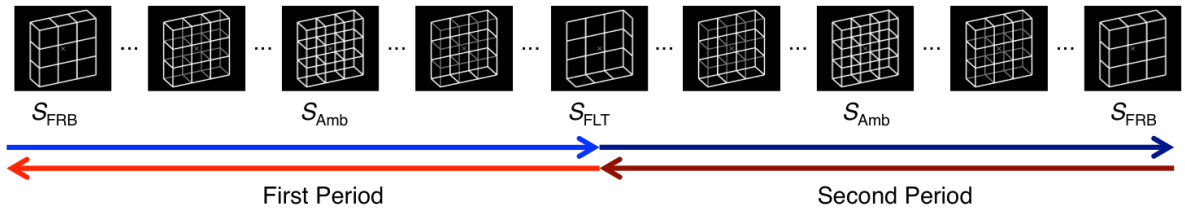


Figure 1: Necker lattices stimuli used in the experimental paradigm. From left to right, in the S_{FRB} => S_{FLT} => S_{FRB} series, we start with a fully disambiguated front-side right bottom view (S_{FRB}), then ambiguity increases to full ambiguity (S_{Amb}), then ambiguity decreases again to a front-side top left view (S_{FLT}), and then the process is reverted back. Around the 9th and 26th trial (S_{Amb}), we expect a perceptual reversal due to the changing orientation of the lattice. In the S_{FLT} => S_{FRB} => S_{FLT} series, the stimuli are presented in opposite order. S: Stimulus; Amb: Ambiguous; FRB: Lattice **f**ront-side to the **r**ight **b**ottom; FLT: Lattice **f**ront-side to the **l**eft **t**op.

Stimuli We presented an ambiguous "Necker lattice" consisting of 3x3 Necker cubes (Figure 1, S_{Amb}) and 16 disambiguated lattice variants $S_{FRB;1}$ - $S_{FRB;8}$, and $S_{FLT;1}$ - $S_{FLT;8}$, with "FRB" = front-side right bottom and "FLT" = front-side left top, and S = stimulus (Figure 1). Stimulus disambiguation and thus percept stabilization was achieved by reducing the luminance of the virtual back layers of the lattices. Luminance was 0.8 cd/m^2 for the ambiguous Necker lattice and the front layers of all disambiguated lattices. Luminance values in cd/m^2 for the apparent back-layers were in descending order: 0.71, 0.59, 0.48, 0.37, 0.27, 0.18, 0.07, and 0.04. All lattices subtended an area of 5.5° x 6.5° visual angle. A cross in the centre of the lattices served as fixation target.

Procedure Each participant was presented with ten repetitions of a series of 33 Necker lattices with varying orientation[1]. The experiment consisted of two conditions, each consisting of two periods. In Condition 1 the lattice orientations were varied step-wise from S_{FRB}, over the fully ambiguous lattice S_{Amb} to S_{FLT} (first period) and back again over S_{Amb} to S_{FRB} (second period). In Condition 2 we presented the lattices in the reversed order: S_{FLT} => S_{Amb} => S_{FRB} => S_{Amb} => S_{FLT}.

Both conditions were presented five times to each participant in randomized order. After stimulus onset, participants indicated their perceived lattice orientation (either P_{FRB} for the front-side right bottom view or P_{FLT} for the front-side left top view) by pressing one of two buttons on a hand-held response box during the presentation of the stimulus. After participants' responses the stimulus remained on the screen for another 1000 *ms* (but maximal 4000 *ms* in total) and was replaced by a blank screen with the fixation target for 400 *ms*.

[1] Due to a coding error, the first eight participants were not presented with the 33th stimulus. In order to allow comparison across the entire group of participants, the resulting missing response in each

Results

Participants responded to more than 99% of the presented stimuli, indicating a sufficiently long response time window and ensuring an unequivocal stimulus-response relation.

Classic Hysteresis Within Each Series

A total of 150 series (across participants and repetitions) containing 33 perceptual responses each were collected. Each series contained two potential perceptual reversals, i.e., there were 300 potential reversals in total. Of these, 102 were clearly identifiable with each response being preceded by a continuous sequence of P_{FLT}'s (or P_{FRB}'s, depending on the condition) and followed by a continuous sequence of P_{FRB}'s (or P_{FLT}'s). For these reversals, the particular stimulus for which the reversal took place was determined by averaging the respective stimulus values around the clear division between the two continuous sequences (see Figure 2, "clear reversal"). For the remaining 198 reversals, where participants recorded reversals more than once per period, we estimated this value, by first reordering the responses between the last of the preceding continuous sequence and the first of the following continuous sequence, in order to mimic a "clear reversal". Then, we again took the average of stimulus order values before and after the resulting "clear reversal" (see Figure 2, "estimated reversal"). For 18 reversals this procedure did not work, the corresponding series were too messy and were excluded from the analysis, resulting in 274 reversals in 137 series (i.e. one reversal per period).

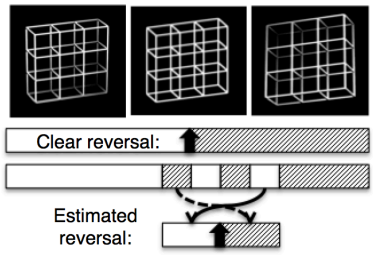


Figure 2: Two response pattern examples for a P_{FRB} (in white) => P_{FLT} (in grey) reversal. The first pattern represents a "clear reversal". The second pattern requires reordering the responses between the continuous sequences of P_{FRB}'s and P_{FLT}'s.

The hysteresis distance was computed by numbering the stimuli by presentation order and subtracting the stimulus order values N_i for which the two reversals in that series took place: $N_{diff} = N_{rev1} - N_{rev2}$. On average, the resulting difference values N_{diff} were significantly smaller than zero (*Mean* N_{diff} = -1.35, *SD* = 2.33, $t(136) = -6.76$, $p < .001$), indicating a *negative hysteresis* effect; rather than the second reversals lagging behind and occurring during more strongly disambiguated stimuli, they occur at less strongly disambiguated stimuli. In particular, the second reversals result 1.35 "disambiguation steps" sooner than the first. Comparing conditions showed no significant difference between the S_{FLT} => S_{FRB} => S_{FLT} and S_{FRB}=> S_{FLT} => S_{FRB} conditions ($t(135) = -1.84$, $p = .07$).

Aggregated Response curves

After collecting all responses, each P_{FLT} (i.e., the perception of a lattice in the front-side left top view) was coded as "0", and each P_{FRB} as '1'.[2] Then, for each of the two conditions, i.e., taking 5 repetitions together, individual responses for each half of the series were fitted with a sigmoid curve as described by equation 1, where m is the point of inflection and s is the slope.

$$f(x) = \frac{1}{1+\exp\left(\frac{m-x}{s}\right)} \qquad (1)$$

Of the resulting fits, we excluded 11 response curves with R2 goodness-of-fit values below a predefined criterion of 0.5. Figure 3 displays the remaining aggregated response curves containing the first perceptual reversal of a series for each participant. About half of these response curves are P_{FLT} => P_{FRB} reversals, i.e., the first reversal in S_{FLT} => S_{FRB} => S_{FLT} series (displayed in red), and the other half P_{FRB} => P_{FLT} reversals, i.e., the first reversal from S_{FRB} => S_{FLT} => S_{FRB} series (displayed in blue). In addition, the means of all red and blue curves respectively are displayed in bold. For the remainder of the paper, we present the results using mean curves only.

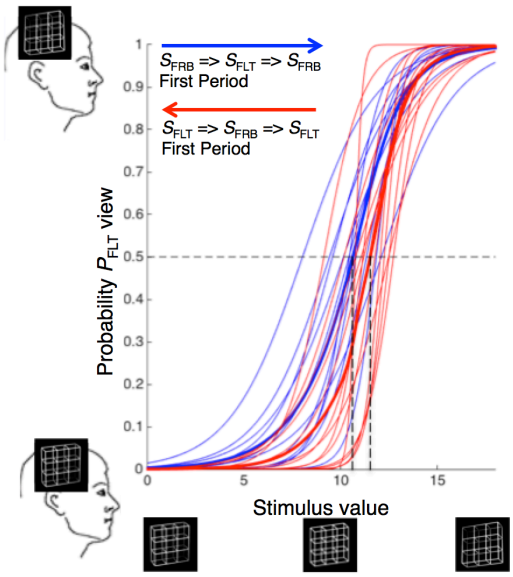


Figure 3: Response curves for each participant for the perceptual reversals in the first period of a series, from P_{FRB} => P_{FLT} (in blue), and P_{FLT} => P_{FRB} (in red). The aggregated response curves are displayed in bold.

Immediate and Intermediate Effects

The sigmoids' inflection points indicate where, on average, the perceptual reversals took place. Looking only at the perceptual reversals that occurred in the first period (see Figure 3), gives insight into the effect of immediate memory. On average, these first reversals occurred for the second stimulus to the right of the fully ambiguous lattice, (*Mean* N_{rev} = 10.97, *SD* N_{rev} = 1.16) meaning for the stimulus that has been disambiguated towards S_{FLT} with two steps. Furthermore, the two bold lines (aggregated data) indicate a difference between the P_{FLT} => P_{FRB} and P_{FRB} => P_{FLT} reversals. When the immediate predecessors of the more ambiguous stimuli are S_{FRB}, the corresponding percepts lose stability later (i.e., at *Mean* N_{rev} = 10.44, *SD* N_{rev} = 1.16) than when the immediate predecessors are S_{FLT} (i.e., at *Mean* N_{rev} = 11.39, *SD* N_{rev} = 1.01).

Figure 4 displays the response curves for both the first and second period of the S_{FRB}=> S_{FLT} => S_{FRB} condition. Again we see that the perceptual reversals do not occur for the same values of back-layer luminance. The results we found when comparing the periods within conditions: Both perceptual reversals occurred at stimuli with luminance values favouring percepts P_{FLT}, whereas the reversals in the first period occurs closer to the

[2] Of the fifteen participants, three had misunderstood the instructions and consistently responded exactly opposite to the remaining participants so their responses were recoded.

fully ambiguous stimulus (N = 9) than in the second.

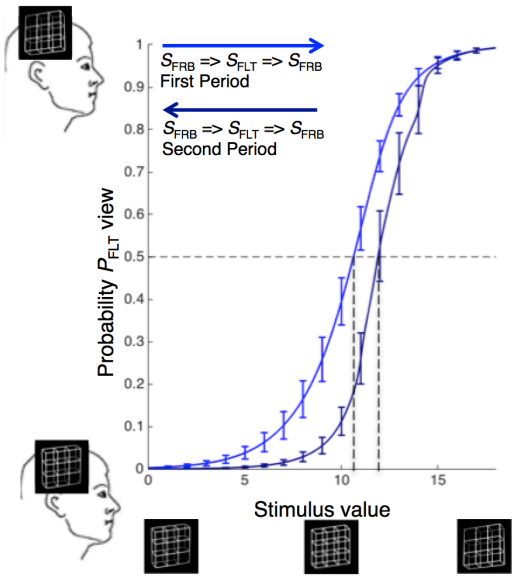


Figure 4: Aggregated response curves for the two periods within the same S_{FRB} => S_{FLT} => S_{FRB} condition. Perceptual P_{FRB} => P_{FLT} reversals occurring in the first (blue) and P_{FLT} => P_{FRB} reversals in the second (dark-blue) period.

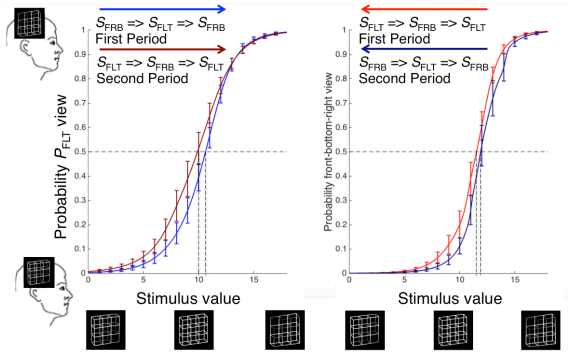


Figure 5: Aggregated response curves for the same type of reversal, P_{FRB} => P_{FLT} (left) occurring in the first period of Condition 1 (blue) and in the second period of Condition 2 (dark-red); and P_{FLT} => P_{FRB} (right) in the first period of Condition 2 (red) and in the second period of Condition 1 (dark-blue).

Long-Term Memory Effects

Figure 5 displays the response curves together for the same *type* of reversal from different periods, and it is clear that regardless of the timing, the P_{FLT} percept loses stability "sooner" (Fig. 5 right, i.e. at lattices with more luminance contrast) than the P_{FRB} percept (Fig. 5 left), although both perceptual reversals occur past the ambiguous stimulus at luminance values in support of percepts P_{FLT}.

Altogether, using a Two-Way Analysis of Variance, we found a main effect for the factor *Condition* (i.e., S_{FRB} => S_{FLT} => S_{FRB} versus S_{FLT} => S_{FRB} => S_{FLT}) on the variable *inflection point position* ($F(1,45) = 5.11$, $p = .029$) but no effect for the factor *Period* (i.e., first versus second period, $F(1,45) = .14$, p = .74). Furthermore, we found an interaction effect between condition and period ($F(1,45) = 27.24$, $p < .001$).

Discussion

Presenting a series of lattice stimuli with two different orders of increasing and decreasing ambiguity results in significant hysteresis effects around the resulting perceptual reversals. These hysteresis effects can be interpreted in relation to memory contributions to perceptual outcomes on three different time scales: (1) The critical stimulus parameter (back-layer luminance) for a reversal between the two different 3D percepts of the lattice stimuli differed between the two presentation orders, reflecting an *immediate memory effect*, i.e. the influence of the immediate perceptual history on a milliseconds time scale on the current percepts. (2) This immediate memory effect is stronger in the first time period of increasing and decreasing lattice ambiguity compared to the second time period, indicating the influence of the first period on the second and thus an *intermediate memory effect* on a time scale of seconds on perception. (3) The inflection points of all sigmoidal perceptual probability functions of any experimental condition are all located at luminance values favouring the lattice front side top left perspective S_{FLT}, reflecting a *long-term memory effect*.

At each moment in our everyday life our perception results from the disambiguation and interpretation of incomplete and to varying degrees ambiguous sensory information. Ambiguous figures are perfect stimuli to study the principles underlying these disambiguation processes, because we can compare perception at maximal stimulus ambiguity and during a stepwise parametrical ambiguity reduction. The literature offers plenty of evidence for top-down influences on the perception of an ambiguous figure (e.g., Long & Toppino, 2004) and studies about priming and adaptation effects in particular already indicate different memory effects on different time scales (e.g., Long, Toppino & Mondin, 1992).

The present experimental paradigm further allows a distinction of memory effects on three different time scales. Recent predictive coding approaches (e.g., Friston, 2012) postulate that the perceptual system makes predictions about the sources of the sensory information, based on the immediate perceptual history and stored concepts from perceptual long-term memory. The present results confirm this perfectly; any specific perceptual situation, like a specific experimental paradigm, can have a strong influence on the perceptual process and outcomes. This observation is highly relevant for any type of experiment on perception and interpretation of results.

The present experiment together with similar, earlier studies (e.g., Hock, Bukowski, Nichols, Huisman, & Rivera, 2005) demonstrate the fruitful application of the physical hysteresis concept to visual perception. Our results show patterns of another important concept from physics, namely complementarity or non-commutativity, which has already been successfully applied in the context of multistable perception and cognition (e.g., Atmanspacher, Bach, Filk, Kornmeier, & Römer, 2008).

References

Atmanspacher, H., Bach, M., Filk, T., Kornmeier, J., & Romer, H. (2008). Cognitive time scales in a Necker-Zeno model for bistable perception. *Open Cybernetics & Systemics Journal*, *2*, 234-251.

Ewing, J. A. (1881). Effects of Stress on the Thermoelectric Quality of Metals. Part I. *Proceedings of the Royal Society of London (1854-1905) 32*, 399–402.

Fender, D., & Julesz, B. (1967). Extension of Panum's Fusional Area in Binocularly Stabilized Vision*. *JOSA*, *57*, 819-830.

Fisher, S. K., & Ciuffreda, K. J. (1989). The effect of accommodative hysteresis on apparent distance. *Ophthalmic and Physiological Optics*, *9*, 184-190.

Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology*, *83*, 248-252.

Hock, H. S., Bukowski, L., Nichols, D. F., Huisman, A., & Rivera, M. (2005). Dynamical vs. judgmental comparison: hysteresis effects in motion perception. *Spatial vision*, *18*, 317-335.

Hock, H. S., Kelso, J. A. S., Schöner, G. (1993). Bistability and Hysteresis in the Organization of Apparent Motion Patterns. *Journal of Experimental Psychology: Human Perception and Performance 19*, 63-80.

Kaplan, D., & Glass, L. (2012). *Understanding nonlinear dynamics*. Springer Science & Business Media.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271-304.

Kornmeier, J., & Bach, M. (2004). Early neural activity in Necker-cube reversal: Evidence for low-level processing of a gestalt phenomenon. *Psychophysiology*, *41*, 1-8.

Kornmeier, J., Heinrich, S. P., Atmanspacher, H., & Bach, M. (2001). The reversing 'Necker Wall' - A new paradigm to the reversal of orientation reveals an early EEG correlate. In T. Arnold & C. Herrmann (Eds.). *KogWis 2001: abstracts of the 5th Meeting of the German Cognitive Science Society* (Vol. 2). Leipziger Universitätsverlag.

Kornmeier, J., & Mayer, G. (2014). The alien in the forest OR when temporal context dominates perception. *Perception abstract*, *43*, 1270-1274.

Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: alternating views of reversible figures. *Psychological bulletin*, *130*, 748.

Long, G. M., Toppino, T. C., & Mondin, G. W. (1992). Prime time: Fatigue and set effects in the perception of reversible figures. *Perception & psychophysics*, *52*, 609-616.

Necker, L. A. (1832). LXI. Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid.

Sundareswara, R., & Schrater, P. R. (2008). Perceptual multistability predicted by search model for Bayesian decisions. *Journal of Vision*, *8*(5), 12.

Troje, N. F., & McAdam, M. (2010). The viewing-from-above bias and the silhouette illusion. *i-Perception*, *1*, 143.

von Helmholtz, H. V. (1911). *Handbuch der Physiologischen Optik*. Hamburg/Leipzig: Leopold Voss.

Washburn, M. F., Mallay, H., & Naylor, A. (1931). The influence of the size of an outline cube on the fluctuations of its perspective. *The American Journal of Psychology*, 484-489.

World Medical Association. (2000). Declaration of Helsinki, ethical principles for medical research involving human subjects. *52 nd WMA General Assembly, Edinburgh, Scotland*.

Critical Features of Joint Actions that Signal Human Interaction

Tianmin Shu [1] **Steven Thurman** [1] **Dawn Chen** **Song-Chun Zhu** **Hongjing Lu**
{tianmin.shu, sthurman}@ucla.edu sdawnchen@gmail.com sczhu@stat.ucla.edu hongjing@ucla.edu
Departments of Psychology and Statistics, University of California, Los Angeles

Abstract

We examined the visual perception of joint actions, in which two individuals coordinate their body movements in space and time to achieve a joint goal. Animations of interacting action pairs (partners in human interactions) and non-interacting action pairs (individual actors sampled from different interaction sequences) were shown in the experiment. Participants were asked to rate how likely the two actors were interacting. The rating data were then analyzed using multidimensional scaling to recover a two-dimensional psychological space for representing joint actions. A descriptive model based on ordinal logit regression with a sparseness constraint was developed to account for human judgments by identifying critical features that signal joint actions. We found that identification of joint actions could be accomplished by assessing inter-actor correlations between motion features derived from body movements of individual actions. These critical features may enable rapid detection of meaningful inter-personal interactions in complex scenes.

Keywords: joint action; feature selection; human interaction

Introduction

Humans actions provide critical information for understanding other people's intentions and reacting accordingly. Beyond recognition of individual actions, the ability to engage in *joint* action (i.e., when two or more individuals coordinate their actions in space and time) paves the way for other social interactions (Sebanz, Bekkering, & Knoblich, 2006; Thurman & Lu, 2014). In everyday life, we constantly coordinate our own actions with those of others to achieve a joint outcome involving a change in the environment (e.g., lifting a box together and moving it to a different location), or to achieve a social goal through human interaction (e.g., walking towards one another and giving a "high-five" as a greeting). Carpenter (2009) provided evidence that the ability to participate in joint action is already fairly developed after the first year of life.

However, it is by no means a trivial task to identify whether two persons are interacting in a meaningful manner solely from visual input. There are many circumstances in which individuals incidentally intersect the orbit of other people's body movements, creating a scene that might be confusable with potential interactivity coordinated by the two people. Hence, it is important to examine what specific information in a visual input signals engagement in joint action, and to determine the conditions under which joint action emerges from body movements of individuals.

In the literature, both reasoning-based and feature-based mechanisms have been proposed to support the recognition and planning of actions when interacting with the environment or other people. In particular, there have been several recent efforts to interpret an agent's behavior via the intentional stance required to understand joint actions. For example, Baker et al. (2009) developed a computational model for reasoning about intentions within a sprite world inspired by the seminal work of Heider and Simmel (1944), in which simple shapes (e.g., red circles or blue triangles) move around in a constrained environment. Although these studies illustrate the potential fruitfulness of applying high-level constraints to reason about the goal underlying observed actions, the investigations have been limited to simplified environments and movements of rigid objects. Relatively few psychological studies have used whole-body movements of humans as the visual input to examine the mechanisms underpinning joint actions.

On the other hand, Marsh et al. (2009) proposed that social interaction through joint actions can be understood as emerging from dynamical principles across individuals, rather than relying on explicit reasoning on intention, at least in some situations. For example, Richardson et al. (2007) showed that connections between humans can arise through synchronization of action patterns (e.g., rocking together when sitting on rocking chairs side-by-side), rather than mental state attribution. This line of work suggests the possibility that bottom-up feature-based mechanisms may play an important role in signaling joint actions. However, the related studies have been limited to uninstructed coordination of simple incidental rhythmic movements, which only cover a small range of possible human actions, and did not aim to probe the important features underpinning meaningful interactions between actors. In the literature of action recognition, there is rich evidence showing that humans are sensitive to some signature movements to enable efficient action detection (Casile & Giese, 2005; Troje & Westhoff, 2006; van Boxtel & Lu, 2015, 2012). Hence, it is conceivable that some critical features of coordinated movements between actors can facilitate the perception of joint actions.

The present study investigated people's mental representations of joint actions and how well features measuring different types of movement coordination between individual actors can capture human judgments of interaction. We collected human ratings of the degree to which two actors seem to be interacting with each other for different video clips. These data were then used to derive a psychological space representing joint actions between two individuals. Computational models were developed to pinpoint critical features of coordinated movements used in constructing a representation space for joint actions by fitting the models to human interactivity ratings.

[1] These two authors contributed equally.

The Experiment

The experiment was designed to investigate perception of joint actions between two human actors using naturalistic stimuli. Two human actors were recorded engaging in various forms of social interaction, such as shaking hands, passing a water bottle, and salsa dancing. We then generated the full set of stimuli by pairing each actor not only with their true interaction partner, but also with each of the other actors involved in different recorded joint actions (e.g., a salsa dancer on the left with a person shaking hands on the right). We sought to determine the extent to which subjects would recognize true interpersonal interactions through joint action, and also whether the stimulus properties of non-interacting stimulus pairs would result in coincidental visual cues that signal attribution of social interaction in joint action.

Method

Stimuli

Action stimuli were generated from the CMU Mocap database (`http://mocap.cs.cmu.edu`) and processed by the Biological Motion Toolbox (van Boxtel & Lu, 2013). We selected ten paired interactions in which two human agents were engaged in various social interactions (see Table 1). A total of 100 stimuli with paired actions were presented in the study, including 10 truly interacting with coupled partners and 90 not interacting. The point-light stimuli were rendered as stick figures with lines connecting the joints, and videos lasted 3.67 seconds (110 frames presented at 30 fps).

Procedure

Participants first viewed two videos, one of a single stick figure walking and another running, and were asked to write a description of what each person was doing. This was to ensure that they understood the format of the videos and were sufficiently competent in English to complete the remainder of the experiment. Participants then viewed 25 videos, 5 of which were chosen from the 10 interacting pairs and 20 of which were the non-interacting pairs formed from the remaining 5 actors. For each video, participants were asked to "rate the degree to which the actors appear to be interacting" on a scale from 1 (Definitely NOT) to 7 (Definitely).

In addition to the 25 videos of interest, participants viewed two videos that were presented on randomly selected trials among the other videos in order to check whether they were paying attention to the task. These were videos of a single actor walking or running. Participants were asked to use a slider (as in the rating questions) to choose the action depicted in each video. Participants who failed either of these attention checks were excluded from the data analysis.

After the video rating task, participants completed the Autism Quotient (AQ) questionnaire (Baron-Cohen et al., 2001), where an attention check was also included.

Results

195 participants were recruited and paid $0.80 for participation in the 5-10 minute experiment on Amazon's Mechanical Turk (Mturk). We excluded 26 participants because they failed one or more of the attention checks during the experiment. Furthermore, we included two additional exclusion criteria to remove participants who did not pay attention to the task. The first measure was the standard deviation of each participant's interactivity ratings across all the testing trials. We excluded 5 participants based on the standard deviation of their ratings being less than 2 z-scores below the mean standard deviation for all participants. Finally, 4 participants were excluded for mistakenly providing higher average interactivity ratings for the non-interacting pairs than for the interacting pairs. Data from the remaining 160 participants were analyzed.

As shown in Figure 1b, human participants yielded significantly higher mean ratings for interacting pairs (mean $= 6.21 \pm 0.82$) than for non-interacting pairs (mean $= 4.38 \pm 0.27$), demonstrating that Mturk participants were sensitive to true joint actions supporting meaningful human interactions

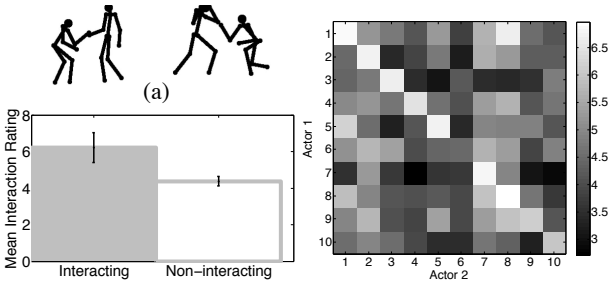


Figure 1: (a) Frames of two confusing stimuli: 1) arm wrestling and shaking hands, 2) threatening and arm wrestling. (b) Mean interactivity ratings for truly interacting and non-interacting stimuli, represented by collapsing across action types. (c) The matrix of mean interactivity ratings for all pairs, highlighting the variable ratings given to different non-interacting stimuli.

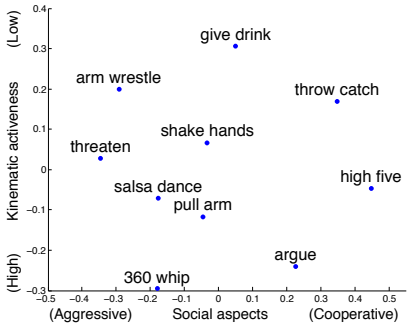


Figure 2: MDS solution derived from human interactivity ratings. The vertical dimension captures activeness of body movements involved in joint action, whereas the horizontal dimension associates with social aspects of joint actions with the exception of "argue" sequence, in which the social content in the point-light display is ambiguous (as this joint action can be interpreted either as an argument or as a normal conversation between two people).

(see Figure 1b). Figure 1c showed the mean interactivity rating for each of the 100 pairs of actors. These results can be visualized as an interactivity-rating matrix in which the diagonal elements represent mean ratings for the true interaction stimuli, and the off-diagonals represent mean ratings for non-interacting stimuli. There was substantial variability in interactivity ratings for the set of non-interacting stimuli, with some pairs (e.g., two examples in Figure 1a) yielding high ratings, and other pairs (e.g., giving a high-five and arm wrestling) yielding relatively lower ratings.

We analyzed the matrix of mean interactivity ratings using multidimensional scaling (MDS). MDS is usually applied to similarity data, and the rated interactivities for different pairs of actions can be viewed as analogous to similarities. The MDS algorithm requires that the input matrix is symmetric and that its diagonal elements are all ones (i.e., items are maximally similar to themselves). We therefore divided every element in the rating matrix by the maximum rating along the diagonal, then set all diagonal elements to 1, and finally took the average of the upper and lower triangles of the matrix to form a symmetric input matrix.

Figure 2 shows the results of applying multidimensional scaling to the rating matrix. Because we used interactivity rather than similarity ratings, actions that are closer in this space are more likely to be rated as interactive when paired together. Note that "shake hands", "salsa dance", and "pull arm", which are close to one another, are all actions with sustained touch. In addition, the two dimensions appear to reveal distinct psychological variables. The vertical dimension corresponds to the levels of activity involved, ranging from highly active actions (e.g., "play 360 whip", in which two people hold hands and jump while rotating 360 degrees) to less active actions (e.g., "give drink", in which one standing person passes a can of soft drink to another person standing next to him). The horizontal dimension appears to be associated with social aspects of joint actions, ranging from aggressive, threatening actions (e.g., one person holds an invisible object and appears to threaten another person), to more friendly and cooperative joint actions (e.g., two people walking towards each other to give a high five). One exception was the "argue" joint action, in which two people presumably in argument raise their arms, but never touch each other. When this action sequence is converted into a point-light display, the social content becomes ambiguous (as this joint action can be interpreted either as an argument or as a normal conversation between two people). This is confirmed by the result that the interactivity rating for the "argue" joint action was also the lowest among all the true interactions.

In addition, we found a statistically significant correlation between AQ score and discrimination score, $r = -0.17$, $p = .024$. This relationship suggests that participants with more autistic traits were less able to distinguish between interacting and non-interacting stimuli than participants with fewer autistic traits. This finding is consistent with other evidence of impaired social cognition in autism.

The Model

In order to determine what visual features play important roles in guiding human judgments regarding meaningful interactions between individuals, we developed a descriptive model based on ordinal logit regression, coupled with a sparseness constraint to encourage selection of a relatively small number of discriminative features. We first describe the model used to predict interactivity judgments for experimental stimuli, and its selection of critical features. We will then provide details on the computation of five different types of features from body movements.

Rank Prediction and Feature Selection

The model predicts a rank order of interactivity judgments for each pair of actions shown in the stimulus based on the 2D coordinates of joints involved in individual actions. To train the model, human interactivity ratings were converted to rank order. We clustered average human ratings from the entire experiment into four rank levels, with equal numbers of cases assigned to each level. After training, the model predicts the rank order of the interactivity judgment y for a given unseen pair of actions, $y = 1, \cdots, 4$, based on the visual features $\boldsymbol{x}$ derived from the input video. Based on the logit link, we can write the conditional probability for a rank $j = 1, \cdots, 4$ as

$$\pi_j(\boldsymbol{x}) = p(y = j \mid y \leq j, \boldsymbol{x}) \;=\; \frac{\exp(\psi_j + \boldsymbol{x}^\top \beta)}{1 + \exp(\psi_j + \boldsymbol{x}^\top \beta)}, \quad (1)$$

where $\{\psi_j\}_{j=1,\cdots,4}$, thresholds for the rank j, and β, the coefficients for the features are the parameters to be learned. We use a dummy variable $y_i^j \in \{0,1\}$ to indicate whether the rank of the i-th instance in the data is j. Then for each instance, we define a vector $\boldsymbol{y}_i = (y_i^1, \cdots, y_i^4)^\top$. Let $Y = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^\top$ be the reconstructed responses of the training instances. For n training instances, the log-likelihood is defined as

$$\begin{aligned} & \ell(\beta, \{\psi_j\}_{j=1,\cdots,4} | X, Y) \\ = & \sum_{i=1}^{n} \sum_{j=2}^{4} y_i^j \log \pi_j(\boldsymbol{x}_i) + (1 - \sum_{j'=j}^{4} y_i^{j'}) \log(1 - \pi_j(\boldsymbol{x}_i)). \end{aligned} \quad (2)$$

Since a large number of features are included in the analysis, the model selects the critical features according to a sparseness constraint that includes an ℓ_1 norm penalty for the coefficients β. Hence, for n instances with p-dimensional feature vectors X and labels Y, we maximize

$$\ell(\beta, \{\psi_j\}_{j=1,\cdots,4} | X, Y) - \lambda \sum_k |\beta_k|. \quad (3)$$

Finally, given learned model parameters (i.e., β and ψ) and the features of a new test stimuli $\boldsymbol{x}$, the estimated rank $\hat{y}$ of interactiveness for the test stimuli can be obtained based on the following probability for each $j = 1, \cdots, 4$:

$$p(y = j | \boldsymbol{x}) = \begin{cases} \pi_4(\boldsymbol{x}) & \text{if } j = 4 \\ \prod_{j'=2}^{4} (1 - \pi_{j'}(\boldsymbol{x})) & \text{if } j = 1 \\ \pi_j(\boldsymbol{x}) \prod_{j'=j+1}^{4} (1 - \pi_{j'}(\boldsymbol{x})) & \text{otherwise} \end{cases} \quad (4)$$

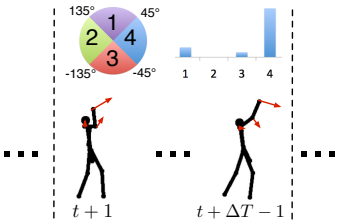


Figure 3: Illustration of motion entropy for the right limb. In a time window of ΔT, we compute the moving directions of the three joints on the right limb at each frame, and compute the histogram of moving directions, which is re-weighted by the magnitude of the velocities. The entropy of the histogram is the motion entropy of the right limb within the time window.

Features of Coordinated Body Movements

A total of 172 features were provided for the model to select. The features are derived from the coordination of limb movements between two actors. We grouped the features into five types according to their functional roles.

Type I: Touching. When viewing a pair of actions, one actor touching the other actor could signal potential joint action with interactive activity (e.g., *Shake Hands*, *Give Drink*, etc.). The feature of touching can be quantified using spatial distance of two joints, each from one actor in the stimulus. The spatial distance of joints can be denoted as d^t_{ij}, where i and j denote index of body joints from the two individuals respectively, and t indicates the frame number. An auxiliary variable $T_{ij} = \sum_t \mathbb{1}(d^t_{ij} < D)$ is introduced to count the number of frames when a pair of joints between the two actors are closer than a threshold, i.e., $D = 8$ pixels. The two joints are considered to be "touching" each other if T_{ij} is longer than $\tau = 30$ frames (i.e., 1 second), which can be denoted as a dummy variable $s_{ij} = \mathbb{1}(T_{ij} > 30)$. Then we can obtain the total number of "touching" pairs of joints by $S = \sum_{ij} s_{ij}$. To eliminate accidental spatial overlap of joints when viewing the action pairs from a particular viewpoint, the two actors are defined to be touching at certain point if and only if one or two pairs of joints between them are considered to be touching, i.e., $1 \leq S \leq 2$. The touching variable is set as 1 if the two actors' joints satisfy this criterion.

Type II: Passing. In some joint actions supporting human interactions, two actors can pass each other, producing spatial overlap of the two actors. The second type of features aim to capture spatial relations between two actors in the stimulus. Similar to the touching feature, we can also count S using the same method as defined in the previous paragraph with thresholds $D = 13$ and $\tau = 30$. The binary passing variable is coded as 1 indicating the presence of this feature if $S \geq 3$.

Type III: Temporal correlation of body movements. The third type of features aims to capture the variability of limb movements over time. Here we define four limbs for an actor (left arm, right arm, left leg and right leg), with each limb including 3 joints. For example, if a person only moves the upper body in joint action (e.g., passing an object while sitting on a chair), there is no variability of leg movements over time, but the arm movements change when actions unfold over time. To quantify this, for each limb we compute entropy based on the histogram of moving directions of joints in two-dimensional space. We first equally quantize joint motion directions into 4 bins as shown in Figure 3, and then count the frequency of moving directions of each joint on that limb at each frame in the given temporal window $\Delta T = 6$. To ensure sensitivity to the influence of joint movement speed, the frequency counts are weighed by velocity magnitude of joint movements. If the joints move in a similar direction over time or remain static, the entropy measure during this period is low. In contrast, the entropy is high when the joints move in varied directions and speeds over time. Because actions progress over time, we obtain a time series of motion entropy for each limb of individual actors in stimuli. To capture the coordination of body movements in joint actions, we compute the correlation of motion entropy sequences between the two actions for each limb pair, resulting in 16 features (i.e., all possible combination of four limbs in each actor). In addition, we include another feature to capture the temporal correlation of whole-body motion entropy sequences for the two actors, where entropy is calculated by pooling motion from all joints. In total, set III has 17 features.

Type IV: Correlation of motion trajectories. Two actors engaged in joint action often perform the same movements at the same time, e.g., raising arms together to give a high-five. To capture this kind of limb movement coordination, we calculated the inter-actor correlation of the motion trajectories for the centers of mass of the four limbs and for the center of mass of the entire body, yielding a total of 17 features.

Type V: Motion correlation with temporal shifts. A characteristic of joint actions is that one person acts in response to the other person's actions. In some situations, limbs move in a synchronized manner (e.g., each actor lifting arms in synchrony to give a high-five greeting), which is capture by feature set IV. In other situations, however, one person moves their limbs first to initiate an interaction (e.g., a person passing an object to the other person). To capture these temporal relations in coordinated movements, we introduced a range of relative temporal shifts (-1.6s, -0.8s, 0.8s, 1.6s) of motion trajectories between two actors to calculate the correlation indices defined in sets III and IV, yielding 136 features in set V.

Finally, all features were standardized to fix the mean at 0 and the variance at 1 across all training stimuli. Any feature values within 1.2 standard deviations of the mean were set to the mean value in order to remove insignificant feature values and to facilitate the feature selection process.

Model Results

Rank Predictions for Unseen Interactions

To evaluate whether the model can generalize its learned features to new joint actions, we split the 100 stimuli into two sets for each of the 10 joint actions: a training set with 81 pairs formed from the other 9 actions, and a testing set with the remaining 19 pairs, one of which is the truly interacting

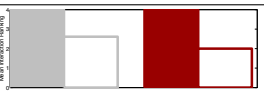
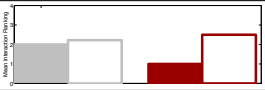
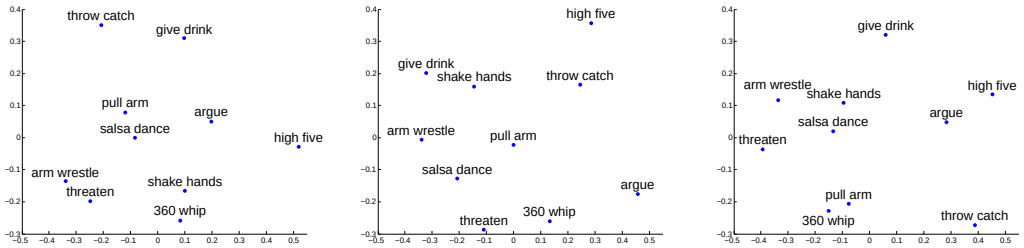
ID	Name	Illustrative Frame	Mean Interaction Ranks	ID	Name	Illustrative Frame	Mean Interaction Ranks
1	Shake Hands			2	Pull by Arm		
3	High Five			4	Give Drink		
5	Arm Wrestle			6	Argue		
7	Play 360 Whip			8	Salsa Dance		
9	Threaten			10	Play Catch		

Table 1: Comparison between mean human interactivity ranks and model predictions for each type of interaction. An illustrative frame from each interaction video is also shown. The gray and red bars are the average ranks based on human ratings and model predictions respectively. Solid bars denote the ranks for interacting pairs and blank bars denote the average ranks for the non-interacting pairs in that testing set.



(a) The first three feature sets. (b) The first four feature sets. (c) All features.

Figure 4: MDS results of fitted ranks from models trained with different feature sets.

pair performing the joint action of interest and 18 of which are non-interacting pairs formed from the partners in the joint action of interest coupled with actors from different joint actions. The final model for each of the 10 joint actions is fitted to select 22 critical features, which yielded the highest correlation between predicted ranks and human ranks.

Table 1 depicts the model's prediction when tested on each interactive pair, in comparison to rank orders from human ratings. The overall correlation between model predictions and human judgments is strong, with $r = .85, p < 10^{-5}$. The average root-mean-square error (*RMSE*) of rank order was 0.47. The best fitted joint actions are *Play Catch*, *Play 360 Whip*, and *Arm Wrestle*, which yielded low *RMSE* (0.08, 0.20, and 0.24 respectively), and the worst fitted actions were *High Five*, *Argue*, and *Threaten* with high *RMSE* (0.90, 0.73, and 0.71 respectively). These differences in *RMSE* across various joint actions suggest that the descriptive model using features of coordinated movements is able to predict the rank order of joint actions which can be defined with strong visual cues, while the model shows its limitation to other joint actions which are rich in social content (such as threaten) but less informative in terms of coordinated movements.

The five features commonly selected across different training sets are 1) touching, 2) correlation between the motion entropies of the right leg of actor 1 and the left arm of actor 2 with an 0.8s temporal shift, 3) correlation between the trajectories of the left leg of actor 1 and the left leg of actor 2 with a -1.6s temporal shift, 4) correlation between the whole-body motion entropies of the two actors, and 5) correlation between the trajectories of the right arm of actor 1 and the left arm of actor 2, which were all chosen on at least 8 test runs with positive coefficients. In particular, touching was selected on all test runs, showing its generalization to a large range of joint actions. These results regarding the features commonly selected by the model are robust, as feature rankings do not change significantly across different sets of model parameters.

MDS from fitted ranks

We trained the model with all 100 action pairs and let the model predict the interactiveness of any pair of actions so that we can derive the model-fitted rating matrix, which was used to recover the psychological space of joint actions by the multidimensional scaling algorithm. The joint action space derived from the model-simulated matrix makes it possible to examine how different types of features impact similarity between different joint actions. We fit the rating matrix using the first three feature sets, first four feature sets and

all features in the five sets, respectively. The resulting MDS results are shown in Figure 4. The comparison of model-derived joint action space with human results shows that set V features play the most important role in mimicking human judgments, as one dimension corresponds to the amount of visual motion information in observed action pairs, and the other dimension is associated with social aspects of joint action. Given that the model is solely based on visual information, without explicitly modeling reasoning of social content beyond the observed joint actions, the recovery of a socially-relevant dimension is intriguing. Note that this dimension only emerges after introducing the set V features with temporal shift between movement of the two actors. This dimension may be strongly associated with the degree of synchronization of body movements between two actors. For example, the friendly impression of the joint action of "walking towards each other to give a friendly high-five greeting" results from the harmonious body movements created by two actors in cooperation. In contrast, the perceived aggression of the joint action of "a person holding an object moves toward the other actor who follows with an avoiding move" results from the temporal relation of movements between the two actors.

Conclusions

The experiment reported in this paper showed that human interactivity ratings provide a tool to gauge the psychological space for representing joint actions. A similar approach has been employed in previous research on perceiving emotion from arm movements in individual actions (Pollick et al., 2001). Our study shows that humans represent major two dimensions of joint actions, one concerning visual motion information and the other concerning social aspects of joint actions. To further understand the representational space of joint actions, we assessed the role of critical features of coordinated movements in signaling human interactions. Research on recognition of individual actions has identified critical features involved in the processing of action stimuli (Casile & Giese, 2005; van Boxtel & Lu, 2015). However, few previous studies have systematically examined critical features of coordinated movements in joint actions. Accordingly, the present study fills an important gap in research on action perception.

Humans can perceive activities jointly performed by two actors from very impoverished stimuli such as point-light displays (actions depicted by discrete joints in a motion sequence). Observers can identify whether actors interact in a meaningful way to achieve a shared outcome involving a change in the environment or the fulfillment of a social goal. An important question concerns how the visual system can generalize action perception to point-light stimuli, which are rarely observed in the visual world. The present findings suggest that a robust ability to identify joint actions may involve the extraction of critical features of coordinated body movements between two actors, and making relational connections between these motion features based on inter-actor correlations. Our findings shed light on how humans achieve efficient detection of meaningful inter-personal interactions in complex scenes, paving the way for a deeper understanding of social cognition.

Acknowledgments

This research was funded by a NSF grant BCS-1353391 to HL and DARPA MSEE project FA 8650-11-1-7149 and ONR MURI project N00014-16-1-2007 for SZ.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). he autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders,*, *31*, 5-17.

Carpenter, M. (2009). Just how joint is joint action in infancy? *Topics in Cognitive Science*, *1*(2), 380-392.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, *5*(4), 6.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2), 243-259.

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, *1*(2), 320-339.

Pollick, F. E., Paterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition*, *82*(2), B51-61.

Richardson, M. J., Marsh, K. L., & Isenhower, R. W. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, *26*(6), 867-891.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70-76.

Thurman, S., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLOS ONE*, *9*(11), 1-12.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: evidence for a life detector? *Current Biology*, *16*(8), 821-824.

van Boxtel, J., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. *PLOS ONE*, *7*(5), e37085.

van Boxtel, J., & Lu, H. (2013). A biological motion toolbox for reading, displaying and manipulating motion capture data in research settings. *Journal of Vision*, *13*(12), 1-16.

van Boxtel, J., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. *Journal of Vision*, *15*(1), 1-17.

Visual constraints modulate stereotypical predictability of agents during situated language comprehension

Alba Rodríguez (arodriguez@cit-ec.uni-bielefeld.de)[1]
Michele Burigo (mburigo@cit-ec.uni-bielefeld.de)[1]
Pia Knoeferle (pia.knoeferle@hu-berlin.de)[2]

[1]Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Bielefeld
[2]Department of German Language and Linguistics, Humboldt University, Berlin

Abstract

We investigated how constraints imposed by the concurrent visual context modulate the effects of prior gender and action cues as well as of stereotypical knowledge during situated language comprehension. Participants saw videos of female or male hands performing an action and then inspected a display showing the faces of two potential agents (one male and one female face) as they listened to German OVS sentences about stereotypically female or male actions. Unlike previous experiments (Rodriguez et al., 2015), the display concurrent with the sentence also showed a picture of the object of the videotaped action and a 'competitor object' (with opposite stereotypical valence) that had not appeared in the video but could be mentioned in the sentence. We measured eye movements to the faces of the agents during comprehension. The design manipulated the match between the videotaped action and the action described by the sentence (*action-verb(phrase) match*) and the match between the stereotypical valence of the verbally described action and the *gender* of the agent of the previous video (conveyed only by the hands; *stereotypicality match*). We replicated the results obtained in Rodríguez et al. (2015): an overall target agent preference (i.e. the agent whose gender matched that of the hands seen in the previous video), reduced by action-verb mismatches. However, unlike in their study, mismatch effects emerged earlier. In addition, stereotypicality effects emerged in the verb region. The earlier mismatch effects and added stereotypicality effects suggest that the visual availability of the objects, perhaps jointly with the verbal input, facilitated the activation of representations from the recent videos (speeding up mismatch effects) and the consideration of alternative representations, favoring stereotypical expectations.

Keywords: Visual constraints; situated language comprehension; gender; stereotypes; eye-tracking.

Introduction

It is a well-established finding that looks towards objects are closely time-locked to their referring expressions (Tanenhaus, Spivey-Knowlton, Eberhard & Sedivy, 1995). Moreover, comprehenders can even make anticipatory eye-movements to potential referents during incremental sentence comprehension if semantic restrictions or world-knowledge impose a sufficiently constraining context. The incremental comprehension of sentences can activate mental representations based on our experiential, long-term knowledge, and these can guide our attention in a concurrent scene (Altmann & Kamide, 1999; Altmann & Kamide, 2004; Kamide, Altmann & Haywood, 2003). However, when prior visual information also acts as a cue during situated language comprehension (e.g., previously seen agents acting upon patients or objects), comprehenders appear to preferentially rely on these recent actions rather than on their stereotypical thematic role knowledge to predict which entity is mentioned next. If prior visual events are available, comprehenders tend to visually anticipate the characters that had been recently seen as involved in those events, rather than guessing who should have been involved based on other types of knowledge (Knoeferle & Crocker, 2007; Rodríguez, Burigo, & Knoeferle, 2015). Entities from previous events are preferentially looked at during comprehension even when verbal information is at odds with the previously inspected actions (Abashidze, Knoeferle; 2015; Rodríguez, Burigo & Knoeferle, 2015).

In eye-tracking experiments on situated language comprehension, Rodríguez, et al. (2015) contrasted gender and action cues from previously seen events with knowledge of gender stereotypes. Participants inspected videos of female or male hands performing an action. After inspecting the videos, they listened to non-canonical German OVS sentences about stereotypically male or female actions while inspecting another display showing the picture pairs of two potential agents (one male and one female face). The non-canonical, yet grammatical word order permitted us to first mention the event theme (i.e. object) and the verb, and observe participants' anticipation of an upcoming agent by monitoring eye movements to the agents' faces. One of the agents on the display was the 'target agent' (i.e. the agent whose gender matched that of the hands seen in the video), while the other character was the opposite gender competitor. Participants' task was to verify whether the content of the sentence matched the immediately preceding event via a button press. In their first experiment, Rodríguez et al. (2015) manipulated two factors. First, whether the videotaped action matched or mismatched the action described by the sentence, i.e. *action-verb (phrase) match*; and second, whether the gender-stereotype associated with the verbally described action matched or mismatched the *gender* of the agent in the

previous video, only conveyed by their hands (i.e. *stereotypicality match*, e.g., congruous case: female hands performed the action in the video and then the sentence was about a stereotypically female action; incongruous: female hands performed the action and then the sentence was about a stereotypically male action, see Table 1). Experiment 2 was similar to Experiment 1, but instead of a mismatch in action-verb reference, the subject of the sentence either matched or mismatched the gender of the agent in the video. These results showed that visual (gender) cues were easily extracted from prior events (i.e. the pair of hands acting upon an object) and successfully integrated with information from a subsequent scene depicting potential agents. Participants responded correctly on more than 90% of the trials in both experiments and they showed an overall preference for inspecting the target agent vs. the competitor during comprehension, regardless of the stereotypical content of the sentence. This preference for the target agent was however reduced in cases of action-verb mismatches at the verb (Exp1) and final noun (Exp2) regions.

One possible account for these results and those obtained in other studies (Knoeferle & Crocker, 2007; Abashidze, Knoeferle, 2015) might be that a preceding event gains relevance by virtue of being recent in memory (Morrison, Conway & Chein, 2014). In that sense, mental representations of events during language comprehension might be preferentially derived from recent perceptual information, rather than experiential knowledge in long-term memory. The mismatch effects observed in our previous experiments (Rodríguez et al. 2015) suggest that when the event described by the utterance does not match a previously seen event, participants still show a preference for it(s agent) even if reduced, and do not seem to rely on the described events and long-term stereotypical knowledge in directing their attention. Incongruence in our experiments thus only led to a disengagement of attention from the target agent. How could we encourage event representations that rely more on long-term knowledge and less on the immediate context, especially where language is at odds with the recently perceived events?

Visual constraints on language comprehension

Both a recent visual context and co-present scenes can affect how rapidly listeners understand language and to which extent they can anticipate upcoming referents. On the one hand, prior scenes can permit anticipation of upcoming entities even when the scene is no longer present during language comprehension and when listeners must rely on previous cues (as was the case in Rodríguez et al., 2015) or a mental representation of a scene. Altmann (2004), for example, had participants inspect a visual context showing a man, a woman, a cake, and a newspaper. After participants had inspected the scene, it was replaced with a blank screen. Next a spoken sentence was played ('The boy will eat the cake'). Participants directed their looks in the blank screen to where the cake had been placed while listening to the word 'eat' while hearing 'the man will eat the cake', and this happened as efficiently as with a concurrent visual context.

Much like a recent visual context can constrain anticipation, listeners' language comprehension and visual attention is also influenced by how the concurrent scene is configured (i.e. more or less constrained). One influential study tested how changes in the concurrent visual context affected the resolution of syntactic ambiguity in sentences like *Put the apple on the towel in the box* (Tanenhaus et al. 1995). An example scene contained either just one referent for an apple (i.e. an apple on a towel) and an empty towel, or it contained a second apple (i.e. on a napkin). When only one apple was present ('one referent' condition), participants tended to incorrectly fixate the empty towel as the goal for the apple after hearing the modifier *on the towel*. However, when two apples were present ('two referent' condition) participants rarely looked at the empty towel (the incorrect goal). Thus, the presence of two referents in the scene prompted participants to interpret *on the towel* as the modifier of the noun *apple*, thus resolving the syntactic ambiguity against their preferred analysis (attachment into the verb phrase).

In terms of the configuration of the visual context in the experiments by Rodríguez et al. (2015), the scene might have been too *unconstrained*: The presence of only the pictures of the potential agents may have biased participants to recruit one type of information (i.e. the gender of the agent in the previous video) when verifying the content of the sentence. While the agent could be easily linked to the video, information about the objects (and their corresponding actions) was no longer available in the concurrent scene. Consequently, participants may have been discouraged from exploiting other sources like long-term knowledge of stereotypes during comprehension. The appearance of objects, together with verbally expressed actions (i.e. the verb phrase), may impose additional constraints on participants' representations, both of recent events as well as of new alternative representations derived merely from language (especially when the sentence mismatches the preceding events). At the same time, world-knowledge associated with the representation of actions performed upon those objects could also be enhanced (Blair, & Banaji, 1996; Bargh, 1999), motivating participants' inferences and expectations towards the upcoming agents.

In summary, previous research has shown that listeners' expectations of soon-to-be-mentioned entities during language comprehension can be constrained by multiple sources. The visual context is one of them, sometimes in the form of prior scenes that are no longer present during language comprehension. This is reflected in anticipatory shifts towards the relevant characters or objects, or the locations where such entities had been previously presented. However, manipulating the configuration of the concurrent scene (i.e. which objects it contains) can also impose contextual constraints on real time language processing.

The current study

The findings from Tanenhaus et al. (1995) and Spivey et al. (2000), which support the idea that manipulating visual constraints can modulate real time language comprehension, motivated a modification in the design employed by Rodríguez et al. (2015, Experiment 1). Participants first inspected a videotaped event that showed a pair of hands interacting with objects (e.g., baking a cake or building a toy model). Recall that in the studies by Rodriguez et al. (2015), the target display following the video only contained photographs of the faces of a male and a female character. By contrast, in the present experiment, we added two pictures: one photograph showed the object that had been acted upon in the preceding video; the other was a photograph of an object that had not been seen before, but that could potentially be mentioned. This latter object was part of an action with opposite stereotype valence from that of the object in the preceding event. For example, after showing a video of a pair of female hands baking a cake, the visual display would contain a female face (the target agent), a male face (the competitor agent), the cake (from the stereotypically female action in the preceding video) and a toy model (a competitor object, part of a model building action which would be stereotypically male, see Figure 1). We measured visual attention to the agent picture pairs during OVS sentence comprehension. Like in our previous experiments, participants answered via button press whether the sentence matched the video they just saw before ("yes" or "no").

The design in the present experiment was the same as in Rodríguez et al. 2015, Experiment 1: action-verb (phrase) match was manipulated together with stereotypicality. Accordingly, we predicted to replicate their findings of shorter response times to action-verb mismatches compared with matches. Additionally, the presence of objects in the target display, together with verbally expressed actions could help boost the representations of action events. If those representations enhanced the activation of gender-related stereotype knowledge, reaction times might be modulated by the stereotypicality of the described actions. This would extend the findings by Rodriguez et al. who failed to observe clear stereotypicality effects in the response times.

For the eye-movements, we initially expected that participants would prefer to inspect the target agent, in line with previous results (Rodriguez et al., 2015) and supporting accounts of visually mediated language comprehension, (Tanenhaus et al., 1995; Knoeferle & Crocker 2007). This preference should decrease when the action described by the sentence mismatches the previously depicted event. However, if changes in the scene that people inspect during comprehension facilitate access to the representation of the recently inspected event as well as the formation of new alternative representations, then action-verb(phrase) mismatch effects could occur earlier than in Rodríguez et al. (e.g., by the end of the first noun or during the verb).

Mismatches in stereotypicality could cause a decrease in preference for the target agent (in line with the gender bias instantiated by the object depictions), particularly for action-verb mismatches. For these, the content of the sentence would not support a representation based on the recently inspected event but rather new representations of action events involving objects that had not been previously seen. The stereotypical beliefs associated with those representations could then be used to anticipate the alternative agent (Blair, & Banaji, 1996; Bargh, 1999). For example, if participants saw female hands backing a cake (a stereotypically female action), but the following sentence described a model building action (stereotypically male), then the presence of an object such as a toy model might bias participants towards looking away from the female face and towards the male agent during sentence comprehension.

Experiment

Participants

32 participants took part in the experiment (16 females, 18-32 years). All were German native speakers and had normal or corrected to normal vision. They all gave informed consent before the experiment.

Materials and Design

We used the materials from our previous studies, which were selected from a norming study: a total of 128 videos of actions (both stereotypically female and male actions were recorded with a female and a male actor; two female actors and two male actors were used). Videos were close-ups of pairs of hands acting upon objects on a table, from external perspective and centered on screen.

From the videos and target displays we constructed 32 experimental items. Each item consisted of a video pair (a pair contained one stereotypically female action, e.g. baking a cake; and one stereotypically male action, e.g. building a model) with their corresponding German sentence pairs with a NP1(object)-V-ADV-NP2(subject) structure. We matched within an item the number of syllables of the words, and the onsets of the different constituents were synchronized. Sentences were presented via speakers, and had a relatively neutral intonation. For the target displays (shown during sentence presentation) we took screenshots of two male actors and two female actors (see Fig. 1) uniformly dressed in black. Additionally for this experiment, the target display also showed snapshots of the objects that belonged to the same item or action pair. One would be the target object (the one that appears in prior events), while the other object would be the competitor object, part of the unseen action with the opposite stereotypical gender valence.

The experimental manipulation followed that of Rodríguez et al. (2015, Experiment 1). We manipulated: a) action-verb congruence (the action described by the sentence matched or mismatched the action previously seen in the video) and b) stereotypicality congruence (the action

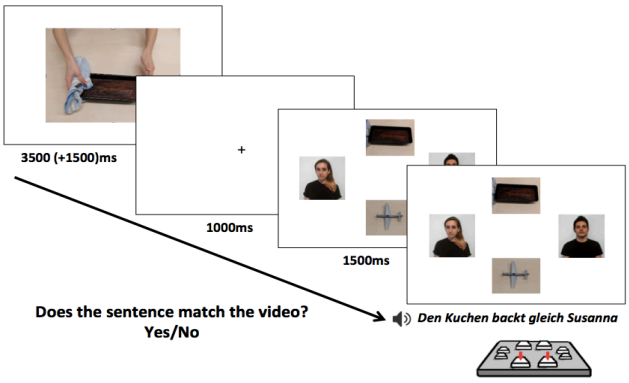


Figure 1. Example of an experimental trial.

described by the sentence either matched or mismatched stereotypically with the gender implied by the hands performing the action in the video. The sentence-final subject in the experimental items always matched the agent of the video in terms of gender. The manipulation of the above-mentioned two factors gave rise to four conditions (see Table 1), which were counterbalanced across 8 lists in a Latin Square manner.

Fillers contained videos of actions which were not classified as stereotypically female or male (e.g. filling out a crossword puzzle), with the same sentence structure as the experimental items (18); videos showing two pairs of hands engaged in an action with dative constructions (18: 9 dative-first and 9 dative-middle sentences); and pictures of objects with sentences of different structures (34). Half of the fillers contained video-sentence mismatches of different types (e.g. action, agent gender, and color).

Procedure

An Eyelink 1000 Desktop Mounted Eye-Tracker recorded participants' eye movements. Viewing was binocular but only the right eye was tracked. Participants completed 10 practice trials before starting the experiment. Trials started with a video of the action for 3500 ms, then the video stopped and the final frame (displaying both the hands in resting position and the object) stayed for another 1500 ms. After that a cross appeared for 1000 ms and then a target screen was shown, with one picture of a female face and another of a male face along the horizontal axis. Along the vertical axis (position was counterbalanced across trials), the target screen showed two objects (one object had featured in the video and the other was a competitor with opposite gender stereotypicality). After a 1500 ms preview, the sentence was presented and eye-movements to the pictures recorded. Participants verified whether the video they just saw matched the sentence that they listened to ("yes" or "no") via button press (Cedrus RB 834). The position of the response buttons was counterbalanced across participants (Figure 1).

Analysis

Sentences were divided into four time regions, consisting of the first noun phrase region (NP1, object region), the verb region (V), the post-verbal adverb region (ADV) and the final noun phrase region (NP2, subject). Each time region extended from its onset to the onset of the next region except for NP2, which finished at the end of the sentence. For analysis, each time window was shifted forward by 200ms, to account for saccadic planning (Matin, Shao & Boff, 1993; Ferreira et al., 2013).

Because looks to one of the characters implied fewer looks to the other character in the scene, we computed the mean log-gaze probability ratio for each region to measure the bias of inspecting the target agent (i.e. the picture of the character whose gender matched that of the hands in the previous video) over the competitor ($ln(P(\text{target agent})/P(\text{competitor}))$). Values above zero reflect a target agent preference, while values below zero reflect a preference for the competitor. These scores are suitable for parametric tests such ANOVAs (Arai, Van Gompel & Scheepers, 2007; Knoeferle et al., 2011). Mean log probability ratios were calculated (both by subject and by item) for each of the regions separately and then subjected to ANOVAs. Reaction times (RTs) were calculated from sentence onset and accuracy was computed per condition. Missing and incorrect responses were excluded from both eye-tracking and the RT analyses.

Results

Accuracy and response times: Participants responded correctly on 97% of the trials with no significant differences in accuracy between conditions. Response times were faster for action-verb mismatches than for matches ($p < .001$).

Eye-movement data: Similarly to previous experiments

Table 1. Example item with literal translations

Video	Sentence				Action-verb match	Stereotypicality match
Female hands baking a cake	***Den Kuchen***$_{NP1}$ the cake	***backt***$_{V}$ bakes	***gleich***$_{ADV}$ soon	***Susanna***$_{NP2}$ Susanna	Yes	Yes
Female hands building a model	***Das Modell***$_{NP1}$ the model	***baut***$_{V}$ builds	***gleich***$_{ADV}$ soon	***Susanna***$_{NP2}$ Susanna	Yes	No
Female hands building a model	***Den Kuchen***$_{NP1}$	***backt***$_{V}$	***gleich***$_{ADV}$	***Susanna***$_{NP2}$	No	Yes
Female hands baking a cake	***Das Modell***$_{NP1}$	***baut***$_{V}$	***gleich***$_{ADV}$	***Susanna***$_{NP2}$	No	No

(Rodriguez et al., 2015), most mean log ratios remained above zero throughout the sentence, suggesting an overall preference for the target agent. However, an action-verb congruence effect emerged as early as in the NP1 region ($p < .01$, Figure 2A). Participants were more likely to inspect the target agent when the object mentioned in the sentence matched (vs. mismatched) the depicted events. At the verb, main effects of both action-verb ($p = .01$, Fig. 2B) and stereotypicality ($p = .01$) congruence emerged, which prevailed until the ADV region. As during NP1, the target received more looks compared to the competitor in action-verb matches compared to mismatches;; additionally, the target was looked at more when the action described by the sentence was stereotypically congruent in terms of gender. For the final, NP2 region, there was a main effect of action-verb congruence ($p < .01$), and an interaction between verb-action congruence and stereotypicality congruence (in the by subject analysis, $p < .05$).

.

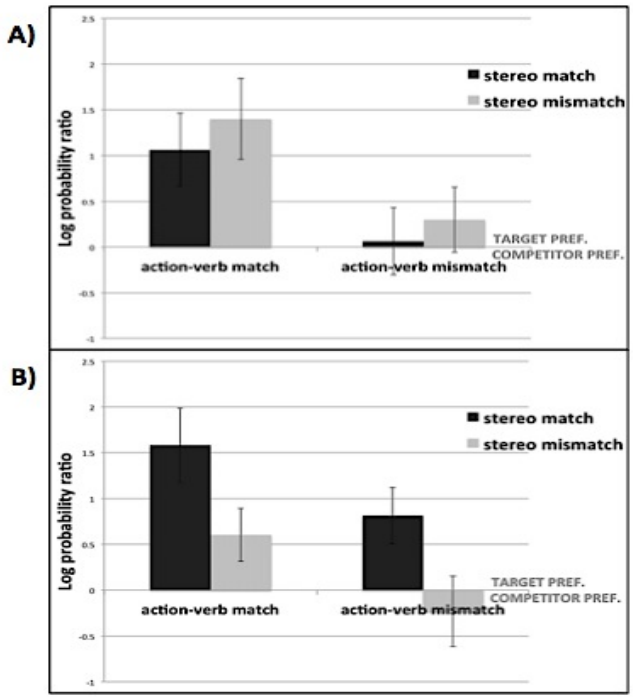


Figure 2. By-subject mean log probability ratios (SE) in the NP1 region (A) and verb region (B)

Discussion

Previous research has contrasted the effects of gender and action cues from a recently inspected event with those of gender stereotype knowledge in situated language comprehension (Rodriguez et al., 2015). These studies have observed no clear effects of gender stereotype knowledge. Instead, analyses of the data from these studies have revealed participants' reliance on gender and action cues derived from the preceding visual context. In the present experiment, we assessed to which extent constraints imposed by the *concurrent* scene might affect the use of recent visual cues as well as boost reliance on knowledge of gender stereotypes. Based on prior research (Rodriguez et al., 2015, Experiment 1), we manipulated the match between the previous event and the sentence (i.e. *action-verb(phrase) match*), as well as between the stereotypicality of the action described and the gender of the previously seen agent (i.e. *stereotypicality match*). Participants saw one video of either female or male hands performing an action and then listened to a German OVS sentence about a stereotypically female or male action, while inspecting a target scene. Unlike in prior research (Rodríguez et al., 2015), participants inspected not just photographs of the faces of two potential agents in the target scene, but also a photograph of an object from the previously seen event and a so-called 'competitor' object (with opposite gender stereotypicality valence). We reasoned that the additional object presentation together with the sentential verb phrase would facilitate the representation of the recently perceived event as well as of another alternative event when language mismatched the prior event (i.e. action-verb mismatches). The object and event representations could enhance (long term) stereotypical knowledge, potentially encouraging expectations of the alternative competitor (Blair, & Banaji, 1996; Bargh, 1999).

The added contextual constraints gave rise to subtly different results compared with the results reported by Rodríguez et al. (2015). We still did not find differences in the behavioral results for the current experiment compared with the first experiment reported by Rodríguez et al. (2015). Moreover, the overall preference for looking at the target agent (vs. the competitor agent) replicated.

Crucially, however, we did observe more rapid effects of action-verb congruence manipulations in the present study (already in the NP1 region). In addition, the present analyses confirmed effects of stereotypicality which had been absent in Rodriguez et al. (2015). These effects emerged as early as in the verb region of the sentence for both action-verb matches and mismatches: participants' preference for inspecting the target agent was more pronounced when the verbally expressed action matched in terms of gender stereotypicality (e.g. when female hands had performed an action, participants preferred to look at the female agent more when the following sentence mentioned a cake baking action compared to a model building action). Also worth mentioning is that the fully mismatching condition (where the action described mismatched previous events and was stereotypically incongruent with the gender cued by the hands did experience a subtle shift of attention towards the competitor agent.

To date, prior visual events have shown a virtually invariant influence on visual attention during situated language comprehension. These events provide detailed information about actions and their associated target objects, as well as individuating information (e.g. about gender of an agent's hands) that can serve to identify associated agents. Upon the encounter of linguistic information, this recent information, rather than world-knowledge derived from language, predominantly guides people's attention over

entities in a referential manner (e.g. a verb phrase can identify the agent that was involved in prior events, and a referent that matches those features will be fixated during the unfolding of the sentence). We have seen this even in cases in which language was at odds with prior events (Abashidze, Knoeferle; 2015; Rodríguez, Burigo & Knoeferle, 2015). However, even in the presence of those recent events, constrains in the concurrent visual input can potentially boost a greater *in situ* visuo-linguistic interaction.

When it is sufficiently constraining, the concurrent visual context seems to facilitate the detection of visuo-linguistic mismatches early during the sentence. Perhaps the concurrent visual context served to boost the representation of the prior event and this resulted in more rapid integration of the prior event representation with representations of the unfolding sentence, eliciting rapid mismatch effects. Additionally, verbal information together with photographs of related objects seems to allow for the consideration of alternative events, enabling further inferences more in line with world-knowledge. Indeed, when language described a different event from the one that participants had just seen, and when world-knowledge derived from the sentence pointed towards a different agent from the one cued by the preceding event, then the impact of the recent events can be greatly diminished.

The results obtained in this experiment support the idea that manipulating visual constraints in situated language comprehension may not only make a difference in terms of how comprehenders resolve structural ambiguity (Tanenhaus et al. 1995; Spivey et al., 2000). Constraining the visual environment via additional (object) referents (as implemented in the present study compared with Experiment 1 by Rodriguez et al., 2015) can boost concurrent visuo-linguistic interactions and facilitate event representations of preceding events, but also new representations of unseen events on site. This facilitation can modulate the time course with which mismatch effects emerge. The presence of additional entities in the concurrent scene during language comprehension seems to also allow for a greater use of inferences based on world-knowledge, which can modulate the extent to which prior perceptual information is used.

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n. 316748 and from the Cognitive Interaction Technology Excellence Cluster 277 (DFG).

References

Abashidze, D., & Knoeferle, P (2015) Do people prefer to inspect the target of a recent action? The case of verb-action mismatches. *Proceedings of the 21st Conference on Architectures and Mechanisms for Language* Valletta, Malta.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank' screen paradigm. *Cognition, 93*, 79-87.

Arai, M., Van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54,* 218-250.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology*, 361-382.

Blair, I. V., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142-1163.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language L, 69*, 165-182.

Kamide, Y., Altmann, G. T. M. & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,*133-159.

Knoeferle, P. & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language, 57*(4): 519-543.

Knoeferle P., Carminati, M.N., Abashidze D. & Essig K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. *Frontiers in Psychology, 2*:306, 1-12.

Matin, E., Shao, K.C. & Boff, K.R. (1993). Saccadic overhead: information processing time with and without saccades. *Perception & Psychophysics, 53*, 372–380.

Morrison, A.B., Conway, A.R.A. & Chein, J.M. (2014) Primacy and recency effects as indices of the focus of attention. *Frontiers in Human Neuroscience*, *8*:6.

Rodríguez, A., Burigo, M., & Knoeferle, P. (2015). Visual gender cues elicit agent expectations: different mismatches in situated language comprehension. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science* , 234–239.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. E., & Sedivy, J. C. (2000). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45, 447-81.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632-1634.

Biases and Benefits of Number Lines and Pie Charts in Proportion Representation

Michelle Hurst (hurstm@bc.edu)
Boston College Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Charlotta Relander (relander@bc.edu)
Boston College Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Sara Cordes (cordess@bc.edu)
Boston College Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Abstract

In two experiments, we investigate how adults think about proportion across different symbolic and spatial representations in a comparison task (Experiment 1) and a translation task (Experiment 2). Both experiments show response patterns suggesting that decimal notation provides a symbolic advantage in precision when representing numerical magnitude, whereas fraction notation does not. In addition, pie charts may show some advantages above number lines when translating between representations. Lastly, our findings suggest that the translation between number lines and fractions may be particularly error-prone. We discuss what these performance patterns suggest in terms of how adults represent proportional information across these different formats and some potential avenues through which these advantages and disadvantages may arise, suggesting new questions for future work.

Keywords: Fractions; Decimals; Number lines; Pie charts Rational numbers; Proportion

Introduction

How we learn and understand the relationship between numerical symbols (i.e., number words or Arabic numerals) and the quantities they represent is a critical component of numerical cognition research. While the acquisition of symbols representing discrete, countable sets (e.g., 5 ducks) is well-studied, less is known about the acquisition of symbols representing proportional information in our environment, namely, fractions and decimals. The mapping between symbolic fractions and the underlying quantity they represent seems to be particularly difficult for both children and adults (Hurst & Cordes, 2015; Ni & Zhou, 2005), potentially due to the complicated nature of the symbolic representation (e.g., bipartite structure of fractions, etc.), as well as the variety of ways in which proportional information is spatially depicted (e.g., number lines, pie charts). In the current study, we investigated how adults map between symbolic (fractions, decimals) and spatial (number lines, pie charts) representations of proportional quantity.

Evidence from infants (e.g., Denison & Xu, 2009; McCrink & Wynn, 2007), children (e.g., Boyer, Levine, & Huttenlocher, 2008), and adults (Fabbri, Caviola, Tang, Zorzi, & Butterworth, 2012; Matthews & Chesney, 2015) suggests that even by a young age, we can understand proportional information when presented non-symbolically. However, mapping between these non-symbolic representations and symbolic representations (fractions, decimals) is not a trivial task. For example, the bipartite structure of fraction notation can lead people to treat fractions as two distinct whole numbers, rather than a coherent unit (e.g., Ni & Zhou, 2005) and superficial similarities between whole number and decimal notation can lead children to make place-value errors, like "longer decimal = larger value" (e.g., 0.313>0.43; Desmet, Gregoire, & Mussolin, 2010). Despite the presence of whole number biases in both fraction and decimal notation, differences have been noted in the affordances of these distinct representations. For example, the bipartite format of fraction notation has been shown to better convey discrete, part-whole information (DeWolf, Grounds, Bassok, & Holyoak, 2014; Rapp, Bassok, DeWolf, & Holyoak, 2015). Decimal notation, on the other hand, has been shown to better convey continuous numerical magnitude information (DeWolf et al., 2014; Hurst & Cordes, 2015), making it more closely align with continuous quantities (Rapp, et al., 2015). However, how these properties of the symbolic representations (i.e., continuous magnitude versus discrete part/whole information) align with conventional spatial representations and in turn impact adults' ability to use and manipulate proportional information in both symbolic and spatial forms is an open question.

Spatial representations, particularly pie charts and number lines, are commonly used in educational instruction and in every day communication of proportional information (e.g., pie charts in investment portfolios). Thus, investigating how these spatial, non-symbolic representations are interpreted and manipulated could shed light on numerical representation and educational practices. Although magnitude information in both pie charts and number lines can be represented in a continuous fashion (meaning, not broken up into unit pieces), these spatial representations are perceptually distinct and may offer different advantages and

disadvantages for relating to decimal and fraction notation. For example, substantial research suggests that number lines may be best for representing continuous magnitude as it aligns with the manner in which we are posited to represent number – along a mental number line (e.g., Wang & Siegler, 2013). If so, rational number magnitudes presented in decimal notation may more naturally translate to number lines than those presented in fraction notation. On the other hand, pie charts may be better at conveying part-whole structure, since the "whole" refers to the complete circle ("whole" may not be as spatially defined in a number line), highlighting an alignment between fractions and pie charts.

In the current study we address how adults map between symbolic and spatial representations of proportional magnitude using a magnitude comparison task (Experiment 1) and a direct translation task (Experiment 2). In particular, we address the question of rational number magnitude representation in two ways: (1) the ease of magnitude access in various forms and (2) the representational flexibility offered by particular representations.

Experiment 1

Methods

Participants Fifty-four Boston College students (18 to 24 years, M=19.2 years, 39 Female) participated in exchange for course credit. An additional nine adults were excluded based on our exclusion criteria (see Data Analysis).

Stimuli The magnitude comparison task stimuli were fractions, decimals, pie charts, and number lines representing proportions between 0 and 1. On each trial, participants were presented with two proportion stimuli: Fraction vs. Fraction (FvF), Fraction vs. Pie Chart (FvP), Fraction vs. Number line (FvL), Decimal vs. Decimal (DvD), Decimal vs. Pie Chart (DvP), and Decimal vs. Number line (DvL). Each type of trial was presented in a separate block. The two proportion stimuli presented on each trial differed by one of two approximate ratios: Small (approximately 1.125, ranging from 1.08 to 1.14) and Large (approximately 1.5, ranging from 1.43 to 1.52).

The proportion magnitudes used in the fraction and decimal trials were approximately matched (e.g., 1/3 would be converted to 0.33). The magnitudes used in the symbolic comparisons (FvF and DvD) were identical to the magnitudes in the symbolic versus non-symbolic comparisons (FvP, FvL, DvP, and DvL). However, in the pie chart (FvP and DvP) and number line (FvL and DvL) blocks, one of the values in each stimulus pair was represented using a pie chart or number line (respectively) instead of a symbol. The choice of which stimuli were represented using a spatial representation was determined so that the spatial representation conveyed the larger magnitude on half the trials.

The symbolic fraction comparisons were created so that on the FvF trials, the two symbolic fractions (4.7 cm high x 3.1 cm wide) were made up of four distinct positive integers (e.g., 2/3 vs. 3/4 would not occur), in order to avoid the use of denominator or numerator based strategies. On the DvD trials one decimal value included digits to the thousandth position (i.e., three digits after the decimal point; e.g., 0.635; 5.5 cm wide) and one decimal value included digits to the hundredth position (i.e., two digits after the decimal point; e.g., 0.76; 4cm wide). The longer decimal was larger on half the trials in order to make length an unreliable strategy. Pie chart stimuli were white circles (radius=3.4cm) with the corresponding proportion filled in black (clock-wise). The number line stimuli were 8 cm lines extending from the end points of 0 to 1 (labeled under the left and right end points, respectively) with a location on the line indicated by a 0.7 cm vertical line.

Procedure All participants completed a magnitude comparison task in which they were shown two values and asked to choose which was larger as accurately and quickly as possible. There were 8 set orders of the six blocks and the order was counterbalanced across participants with an approximately equal distribution of participants in each order. Each block contained four unique trials from each ratio bin shown twice (once with the largest on the right and once with the largest on the left) in a random order, resulting in 96 total trials (4 comparisons x 2 shown twice x 2 ratios x 6 blocks).

Each stimulus remained on the screen until the participant selected an answer by pressing the left or right arrow on the keyboard corresponding to their response (left or right quantity, respectively). Between each stimulus presentation a fixation cross (0.5cm x 0.5cm) was presented in the center of the screen for 1000ms. Each block started with an instruction screen and two practice trials with feedback. The experimenter remained quietly in the room with the participant throughout the task and answered any questions about the procedure prior to each block.

Data Analysis Reaction time (RT) was the primary dependent variable because accuracy was fairly high with low variability. Only RTs from correct trials and those that were within 3 standard deviations of the individual's mean RT of that trial type were included in analyses. Only data from those blocks in which the participant scored at or above chance (4/8 questions correct) were included. Participants who had missing data based on these criteria were excluded from all analyses (N=9). At the group level, average RTs for each cell that were more than 3 standard deviations away from the group mean were replaced with the next highest value that was not considered an outlier. This resulted in 18/1080 data points being replaced (~1.7%).

Results and Discussion

Comparison Performance First, we used a Repeated Measures ANOVA to investigate whether RT differed across Symbolic Notation (2: Fractions, Decimals), Comparison Type (3: Same Symbol, Pie, Line), and Ratio

(2: Small, Large). See Figure 1 and Table 1 for the descriptive statistics.

Table 1: RT in milliseconds (standard error) for each type of trial in fraction (top) and decimal comparisons (bottom)

	FvF	FvP	FvL	Avg
Small Ratio	3479 (177.8)	3186 (159.1)	3424 (161.9)	3363 (138.0)
Large Ratio	2609 (134.1)	2490 (130.0)	2050 (89.7)	2383 (96.5)
Avg	3044 (145.6)	2837 (125.8)	2737 (114.2)	2873 (109.6)

	DvD	DvP	DvL	Avg
Small Ratio	1098 (30.6)	2446 (114.1)	2792 (136.2)	2112 (76.5)
Large Ratio	976 (23.3)	1586 (45.7)	1618 (43.2)	1394 (30.2)
Avg	1038 (25.9)	2016 (71.3)	2205 (80.9)	1753 (48.9)

There was a main effect of Ratio $F(1,53)=161.8$, $p<0.001$, *partial* $\eta^2=0.75$, Ratio X Symbol interaction, $p=0.003$, *partial* $\eta^2=0.15$, Ratio X Comparison interaction, $p<0.001$, *partial* $\eta^2=0.4$, and Ratio X Symbol X Comparison interaction $p<0.001$, *partial* $\eta^2=0.15$. Pair-wise comparisons investigating whether there were ratio effects (significantly slower RTs for Small ratio trials than Large ratio trials) in each block separately, found significant ratio effects in each of the six blocks (all p's<0.001). However, comparisons involving number lines (DvL and FvL) had significantly larger ratio effects than those involving pie charts (DvP and FvP; p's<0.007) and than those involving only symbols (DvD and FvF; p's<0.03).

Thus, the existence of ratio effects in our data suggests that adults did access the approximate magnitudes of proportional values represented in decimal, fraction, pie chart, and number line form. However, the size of the ratio effects varied depending on the representational form. In particular, those comparisons involving number lines had the highest ratio effects, above those involving pie charts or only symbols. This supports the general idea that number lines are thought to communicate magnitude information better than other representations (Cramer, Post, & DelMas, 2002; Wang & Siegler, 2013). Since number lines are continuous, ordered, and approximate, adults may have been more inclined to use magnitude-based strategies, leading to those comparisons being more dependent upon the particular magnitudes (i.e., higher ratio effects).

There was an overall main effect of Symbol, $F(1,53)=158.8$, $p<0.001$, *partial* $\eta^2=0.75$, with Fraction comparisons taking longer than Decimal comparisons. This finding is consistent with other work suggesting that magnitudes represented in fraction notation take longer to access (Hurst & Cordes, 2015). However, there was also a Symbol X Comparison interaction, $F(2,106)=55.8$, $p<0.001$, *partial* $\eta^2=0.5$. Follow up tests indicated that participants were much faster when comparing two decimals than comparing a decimal with either a pie chart or a number line (p's<0.001). Conversely, comparisons involving two fractions were slower than those involving a fraction and a number line ($p=0.024$) or a fraction and a pie chart ($p=0.07$, *marginal*). Thus, while decimal notation seemed to offer an advantage over spatial representations (that is, performance on trials involving two decimals was better than when a spatial representation was involved), fraction notation appeared to present a symbolic *disadvantage* when processing numerical magnitude. Processing proportional information in fraction notation may not only be more difficult than decimal notation, but it may also be more difficult than processing proportional information via conventional, analog spatial representations. Given that the purpose of numerical symbols is to provide a precise way to communicate numerical magnitudes, it is counter-intuitive that discrete fraction notation does not provide more precise magnitude information than analog spatial representations.

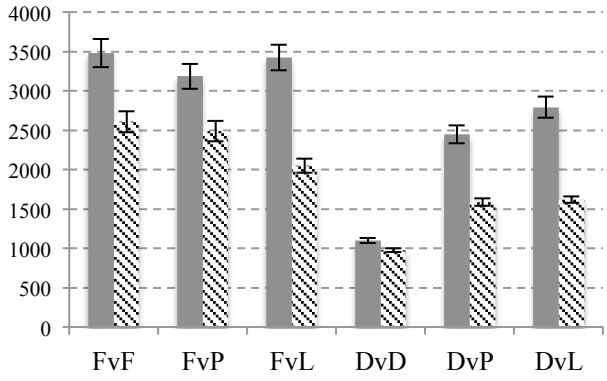


Figure 1: RT on correct trials for the Small (left, grey bars) and Large (right, striped bars) Ratio across each of the six comparisons

Comparison Biases Next, we were interested in whether these distinct symbolic and spatial formats may give rise to biases in the representation of proportional information. That is, did performance differ depending on if the larger value (i.e., the correct response) was symbolic or spatial? We conducted a 2 (Spatial: Pie Chart (PC), Number line (NL)) X 2 (Symbol: Fraction, Decimal) X 2 (Format of Largest: Spatial, Symbolic) repeated measures ANOVA on average RT. There was a main effect of Symbol, $F(1,53)=57.1$, $p<0.001$, *partial* $\eta^2=0.5$ and a Symbol X Spatial interaction, $F(1,53)=10.0$, $p=0.003$, *partial* $\eta^2=0.16$, showing the same pattern of findings reported earlier. However, the critical tests were those involving the format of the largest value. There was a main effect of Format,

$F(1,53)=29.6$, $p<0.001$, *partial* $\eta^2=0.4$, suggesting a bias toward indicating that the symbol was smaller than the spatial representation. However, this was was qualified by two further interactions. A Symbol X Format interaction, $F(1,53)=4.1$, $p=0.048$, *partial* $\eta^2=0.07$, indicated that this bias (faster responses when the symbol was smaller) was greater when comparisons involved fractions (when fraction symbol was the largest value: 2934ms vs. when the fraction was smallest: 2579ms, $p<0.001$) than when the comparisons involved decimals (when decimal symbol was largest: 2136ms; vs. when decimal was smallest: 1973ms, $p<0.005$). Second, there was a Spatial X Format interaction, $F(1,53)=97.6$, $p<0.001$, *partial* $\eta^2=0.65$, which indicated that this bias only held when comparing a symbol to a pie chart. That is, participants were quicker to respond when the pie chart represented the larger value (2097ms) than when it was smaller (2747ms; $p<0.001$); but the reverse pattern was found for number lines, where participants were quicker to respond when the number line depicted a smaller value than the symbol (2323ms vs. 2455ms; $p=0.028$). Thus, adults may be biased toward thinking that pie charts are larger than fractions and decimals, but that number lines are smaller than fractions and decimals. Moreover, given that the symbol bias was greater for fractions, it may be that adults have a tendency to think a fraction represents a smaller value than the equivalent decimal. However, previous studies that had adults directly compare fractions and decimals did not find evidence of this pattern (Hurst & Cordes, 2015). Thus, it may be that these biases only arise when adults are directly thinking about number lines and pie charts in relation to the fractions and decimals.

Thus, in Experiment 1, we show that although adults accessed magnitude information in all comparisons (as evidenced by ratio effects), number lines in particular may encourage magnitude-based comparison strategies (as evidenced by largest ratio effects). Most notably, comparisons involving only symbols were privileged (i.e., faster) over those involving spatial representations, but only for comparisons involving decimal notation; that is, decimals provided a symbolic advantage for conveying magnitude, but fractions did not. Furthermore, although participants were asked to compare across formats, it is unclear how they went about performing these comparisons. Did they explicitly translate magnitude information across representations in order to make the comparison? In Experiment 2, we investigate performance when translating across representational formats (spatial to symbolic and vice versa) in order to shed light on the accuracy with which proportional information is represented in each form.

Experiment 2

Methods

Participants Forty-one adults (18 to 21 years, M=19.2 years, 36 Female) were included in all analyses. Three additional adults participated but were excluded for not following instructions properly. All participants received course credit and none participated in Experiment 1.

Stimuli The same pie chart and number line stimuli from Experiment 1 were used. The fractions and decimal values were equivalent to the pie chart and number line magnitudes so that the magnitudes used in all eight blocks were the same (up to rounding error for decimal magnitudes). All other aspects of the stimuli were identical to Experiment 1.

In order to input their answer, a text box was provided for fractions and decimals. When translating to a number line, a number line (7cm) from 0 to 1 (marked under the left and right endpoints respectively) was shown and participants could move a small vertical line (0.5cm) along the line to select their response. When translating to a pie chart, an empty circle (radius=3cm) was shown on the screen with a 3cm line extending from the top of the circle to the center. Another line extended from the center of the circle to the edge of the circle and moved around the circle corresponding to the location of the participant's cursor. When the participant clicked with the mouse, the pie chart filled in the portion between the top vertical line and the participant's adjustable line black.

Procedure Participants completed a Translation task in which they were given a quantity represented using either a fraction, decimal, pie chart, or number line, and asked to estimate that quantity using a different representation. The task included eight distinct blocks: Pie to Fraction (PtoF), Fraction to Pie (FtoP), Line to Fraction (LtoF), Fraction to Line (FtoL), Pie to Decimal (PtoD), Decimal to Pie (DtoP), Line to Decimal (LtoD), and Decimal to Line (DtoL). There were 8 trials per block, making the task 64 trials (8 trials per block x 8 blocks). The blocks were presented in 8 set orders, counterbalancing across participants.

Prior to each block, participants were shown an instruction screen and the experimenter showed them how to input their response in the correct format. On each trial, the target value was displayed on the left and the empty response (text box, empty line, or empty pie chart) was displayed on the right. Participants could fill in their answer by clicking in a location (number line and pie chart) or typing in a response (decimal and fraction). Participants pressed a button to move on to the next question. The experimenter remained quietly in the room the entire time.

Data Analysis Mean Absolute Error (MAE), calculated as the absolute value of the difference between the correct proportion and the response, was the primary dependent variable.

Exclusion criteria and outlier treatment was identical to that of Experiment 1. At the group level, 8/328 data-points (~2.4% of the data) were considered outliers and replaced with the next highest non-outlier.

Results and Discussion

We conducted a 2 (Direction: Symbol to Spatial vs. Spatial to Symbol) X 2 (Symbol Type: Fraction (F) vs. Decimal (D)) X 2 (Spatial Type: Pie Chart (PC) vs. Number Line (NL)) repeated measures ANOVA on MAE (see Table 2 for descriptive statistics involving PCs (top) and NLs (bottom)).

Table 2: MAE (standard error)

	Translating		
	to PC	*from PC*	Avg
F	0.039 (0.004)	0.035 (0.006)	0.033 (0.003)
D	0.033 (0.002)	0.032 (0.003)	0.036 (0.002)
Avg	0.036 (0.002)	0.033 (0.003)	0.035 0.003

	Translating		
	to NL	*from NL*	Avg
F	0.044 (0.002)	0.050 (0.008)	0.041 (0.004)
D	0.044 (0.003)	0.031 (0.002)	0.044 (0.002)
Avg	0.044 (0.002)	0.041 (0.004)	0.042 0.003

First, there was a main effect of Symbol Type, $F(1,40)=4.5$, $p<0.05$, *partial* $\eta^2=0.1$, revealing adults were more accurate when translations involved a decimal (compared to those involving a fraction). This finding aligns with those of Experiment 1 and previous work (e.g., Hurst & Cordes, 2015) suggesting that decimals provide more accurate magnitude information than fractions. There was also a main effect of Spatial Type, $F(1,40)=9.76$, $p<0.004$, *partial* $\eta^2=0.2$, revealing, in contrast to previous research highlighting the benefits of number lines (Wang & Siegler, 2013), that adults were more accurate when translations involved a pie chart compared to a NL.

Furthermore, there was a three-way interaction between Direction, Symbol, and Spatial Type, $F(1,40)=5.0$, $p<0.05$, *partial* $\eta^2=0.1$. We investigated this three-way interaction further by conducting two 2 X 2 repeated measures ANOVAs, looking at the effect of Direction (2) and Symbol Type (2) separately for PCs and NLs.

The 2 x 2 ANOVA on data from trials involving PCs revealed no main effects or interactions (p's>0.2) suggesting that performance was very similar regardless of whether the translation involved a fraction or a decimal or whether the PC was the target or the initial value.

However, the pattern was not the same when we looked at the data from trials involving NLs. The 2 x 2 ANOVA on data from NL trials revealed a main effect of Symbol, $F(1,40)=4.23$, $p<0.05$, *partial* $\eta^2=0.1$, again suggesting that trials involving decimals resulted in lower error than those involving fractions. There was no main effect of Direction, however, there was a significant Symbol X Direction interaction, $F(1,40)=6.01$, $p<0.02$, *partial* $\eta^2=0.1$. Follow up t-tests revealed that when translating proportional information *to* a NL, performance was equally accurate regardless of whether the starting value was a fraction or a decimal ($p=0.8$). However, when translating proportional information *from* a NL, performance was significantly better when converting into a decimal compared to a fraction ($p=0.023$). That is, adults were particularly inaccurate when converting from a NL into a fraction, a finding that may be attributed to two factors: (1) the fact that fractions are particularly inaccurate for representing magnitude information (Hurst & Cordes, 2015) and (2) a mismatch in the way magnitude information is represented in fraction form (part-whole) and in NL form (linear, continuous).

General Discussion

In this study, we investigated adults' representation of proportional magnitudes across common symbolic (fractions, decimals) and spatial (number lines, pie charts) formats. In line with the exact precision offered by symbols, decimal notation provided adults with the greatest level of precision when comparing magnitudes in Experiment 1 and when translating between different representations of magnitude in Experiment 2. Thus, decimals seemed to offer the symbolic advantage that is expected. In contrast, fractions did not; comparisons involving exclusively fraction notation took the longest, even compared to those involving spatial, non-symbolic representations. These findings of a symbolic magnitude advantage for decimals adds to a growing literature (DeWolf et al., 2014; Hurst & Cordes, 2015) by further suggesting that decimals are also more accurate than analog spatial representations of proportion and, conversely that fractions are potentially less precise at conveying magnitude information than spatial representations.

In addition, Experiment 2 suggested some advantage for estimating proportional magnitude using pie charts over using number lines. Given the literature suggesting that teaching fractions using number lines is beneficial, this finding is counterintuitive (e.g., Cramer et al., 2002; Wang & Siegler, 2014). However, it is important to note that these adults were not receiving instruction on number lines and pie charts, and what's more, in line with instructional practices in the U.S. over the past 15 years, these adults likely received a curriculum that relied heavily on pie charts. Keeping this in mind, there are at least two potential explanations for these findings. On the one hand, adults may have attempted to engage in a partitioning strategy (i.e., dividing the image into the total number of parts) for both pie charts and number lines, but executing that strategy may have been easier with a pie chart. For example, since both pie charts and number lines were presented continuously (i.e., un-partitioned), it may be easier to visibly estimate

partitions in a pie chart because it is symmetric through the center of the circle. Thus, the same partitioning strategy may not be equally accurate across the two representations. On the other hand, adults may have opted not to engage in partitioning with number lines but instead invoked an altogether different strategy when faced with a number line trial. For example, given its continuous nature, participants may have attempted to estimate the proportion of the line that was marked using a magnitude-based strategy. This hypothesis (that number lines evoke more approximate strategies than pie charts) is also consistent with the relatively high ratio effects found in Experiment 1 when comparing number line magnitudes. Given that fractions are particularly poor conveyors of magnitude information (Hurst & Cordes, 2015), this poor strategy selection may have led to lower response precision particularly when translating between number lines and fractions. Since fractions are more aligned to discrete representations (Rapp et al., 2015), it may be that the approximate strategy evoked by the number line is particularly difficult to translate into fraction form. Further research could investigate how these performance differences arise by investigating specific strategies invoked for different representations (i.e., pie charts and number lines) as well as in different tasks. Although we found that number lines and fractions may be misaligned in some translation contexts in Experiment 2, this response penalty (slower RTs) was only found when translating *into* a fraction. This finding suggests that adults in Experiment 1 may not have mentally converted spatial representations into fractions when comparing across formats, but instead did the reverse - converting fractions into spatial representations - to make the comparison.

In conclusion, the current study adds to the growing literature investigating the advantages and limitations of proportion representation in various forms. Results suggest that the mapping between symbolic and spatial representations of proportion may heavily depend on the specific symbolic notation involved and on the type of spatial representation of quantity. Future work should attempt to isolate how these different mappings are learned in young children and what impact these distinct representational formats may have on their understanding of proportional information.

Acknowledgments

Funding provided to MH through a post-graduate fellowship from the Natural Sciences and Engineering Research Council of Canada.

References

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: where young children go wrong. *Developmental psychology*, *44*(5), 1478.

Cramer, K. A., Post, T. R., & delMas, R. C. (2002). Initial fraction learning by fourth-and fifth-grade students: A comparison of the effects of using commercial curricula with the effects of using the rational number project curriculum.*Journal for Research in Mathematics Education*, 111-144.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The Mental Representation of Parity and Number Magnitude. *Journal of Experimental Psychology: General, 122*(3), 371-396.

Denison, S., & Xu, F. (2010). Twelve-to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13*(5), 798-803.

Desmet, L., Grégoire, J., & Mussolin, C. (2010). Developmental changes in the comparison of decimal fractions. *Learning and Instruction*, *20*(6), 521-532.

DeWolf, M., Bassock, M., & Holyoak, K. J. (2015). Conceptual Structure and the Procedural Affordanes of Rational Numbers: Relational Reasoning With Fractions and Decimals. *Journal of Experimental Psychology: General, 144*(1), 127-150.

DeWolf, M., Grounds, M., Bassok, M., & Holyoak, K. (2014). Magnitude Comparison With Different Types of Rational Numbers. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 71-82.

Faulkenberry, T., & Pierce, B. (2011). Mental Representations in Fraction Comparison. *Experimental Psychology, 58(6),* 480-489.

Gallistel, C., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44,* 43-74.

Hurst, M., & Cordes, S., (2015). Rational Number Comparisons Across Notation: Fractions, Decimals, and Whole Numbers. *Journal of Experimental Psychology: Human Perception and Performance. DOI:* 10.1037/xhp0000140

Kallai, A.Y., & Tzelgov, J. (2014). Decimals are not processed automatically, not even as being smaller than one. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 962-975.

Matthews, P. G., & Chesney, D. L. (2015). Fractions as percepts? Exploring cross-format distance effects for fractional magnitudes. *Cognitive psychology*,*78*, 28-56.

McCrink, K., & Wynn, K. (2007). Ratio abstraction by 6-month-old infants.*Psychological science*, *18*(8), 740-745.

Moyer, R. S. & Landauer, T. K. (1973). Determinants of reaction time for digit inequality judgments. *Bulletin of the Psychonomic Society, 1,* 167-168.

Ni, Y., & Zhou, Y. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40*(1), 27.

Rapp, M., Bassok, M., DeWolf, M., & Holyoak, K. J. (2015). Modeling discrete and continuous entities with fractions and decimals. *Journal of Experimental Psychology: Applied*, *21*(1), 47.

Schneider, M., & Siegler, R. S. (2010). Representations of the magnitudes of fractions. *Journal of Experimental Psychology, 36*(5), 1227.

Wang, Y., & Siegler, R.S. (2013). Representations of and translations between common fractions and decimal fractions. *Chinese Science Bulletin, 58*(36), 4630-4640.

Exploring the Relationship between Adolescents' Interest in Algebra and Procedural Declines

Natalie Corbett (Natalie.Corbett@temple.edu)
Temple University, College of Education, 1301 Cecil B. Moore Avenue,
Philadelphia, PA, 19122 USA
Julie L. Booth (Julie.Booth@temple.edu)
Temple University, College of Education, 1301 Cecil B. Moore Avenue
Philadelphia, PA 19122 USA
Christina Barbieri (Barbieri@udel.edu)
University of Delaware, College of Education and Human Development,
113 Willard Hall Education Building,
Newark, DE 19716 USA
Laura K. Young (Laura.Young@temple.edu)
Temple University, College of Education, 1301 Cecil B. Moore Avenue
Philadelphia, PA 19122 USA

Abstract

Algebra I is considered a gatekeeper course for higher education, high-paying jobs, and access to STEM careers, yet many students find themselves struggling to learn algebra. Prior research links intrinsic motivation for learning math with mathematics achievement, particularly during adolescence. The current study measured middle school students' interest in algebra and their procedural skills across the span of an algebra unit to determine whether students who show declines in algebraic problem-solving also show a decline in a particular type of intrinsic motivation – interest in algebra. Pre-test and post-test scores were used to categorize participants into those who showed declines in problem-solving skills and those who did not. Of the overall sample (N = 367), a group of 25 students showed declining skills over the course of the unit. These students also showed significant declines in interest in mathematics from pre- to post-test in comparison to students who did not show procedural declines. Our findings support the relationship between performance and motivation in the classroom, particularly in algebra class. Educational implications are discussed.

Keywords: algebra; motivation; achievement; procedural skills; procedural skill decline; education

Introduction

Prior research has examined mathematical cognition and problem solving in an attempt to understand and, thus, improve students' mathematics abilities. Problem solving skills can be defined as using the cognitive tools needed to solve a given mathematical problem (Schoenfeld, 1992). These problem-solving abilities become more important as students advance in mathematics, particularly when they are in algebra classrooms (Mayfield & Glenn, 2008). An understanding of how to apply procedural skills is also known to contribute to mathematical performance (Star, 2005). Failure to develop these skills leads to poor mathematics performance, specifically in middle school (Lein et al., 2015). Prior research has also suggested that a student's attitude towards mathematics can affect their performance in the classroom (Middleton & Spanias, 1999). It has been well established that achievement and motivation in mathematics are closely linked during adolescence (Wang, 2013).

In early motivation research, being *intrinsically* motivated, or motivated to engage in a task for the sake of the task itself, was deemed more positive than being *extrinsically* motivated, or motivated to engage in the task for an external reward (Deci & Ryan, 1985). This line of research also found that students whose motivational orientations were more aligned with an intrinsic motivation (e.g., mastery orientation, value, interest) tended to persist and succeed more academically. However, the distinction between and value given to intrinsic and extrinsic motivation is not always accepted (Cameron & Pierce, 1994; Ryan & Deci, 1996). Researchers today tend to focus on particular motivation constructs that may be more useful in explaining students' academic behavior than general categories or dichotomies such as intrinsic and extrinsic motivation.

Wang and Pomerantz (2009) found that students who initially had more positive achievement motivation (e.g., mastery orientation) in the fall of the seventh grade had better academic performance in the spring of eighth grade than students whose motivation-orientation would traditionally be considered less positive. They also found that over time students in the United States showed general declines in positive achievement motivation orientations, such as becoming less mastery oriented and valuing mathematics less. Ma (1997) suggests that there is a reciprocal relationship between the development of attitudes towards mathematics and the development of math ability. Fisher, Dobbs-Oates, Doctoroff, and Arnold (2012) found that high interest in mathematics coincided with strong math skills in preschool aged children. However, Fisher and colleagues did not examine adolescents, which makes it

difficult to ascertain the applicability of these findings to the relationship between adolescent student achievement and interest in math. According to the 2012 Programme for International Student Assessment (PISA), adolescents across the United States struggle with mathematics, ranking below average nationally in comparison to other industrialized nations (Kelly, Nord, Jenkins, Chan, & Kastberg, 2013). Further, this same data revealed that 50% of students in the United States have low levels of interest in mathematics. While we know that math achievement and math motivation are related (e.g., Wang & Pomerantz, 2009), little research has assessed the relationship between adolescents' *interest* in mathematics and their mathematical problem-solving abilities.

It is likely that interest in a particular topic or subject area develops over time with experiences; for example, students' various experiences with mathematics, over the course of their schooling, may contribute to interest in mathematics. It is also possible that struggling to develop skills important to success in a particular domain, such as procedural skills in mathematics, may relate to declines in interest in that particular domain.

Achievement in mathematics is linked with success in future employment (Lein et al., 2015). Algebra I in particular is often considered a gatekeeper course for higher-level math courses as well as select science courses (Matthews & Farmer, 2008). When learning algebra, mathematical problem-solving becomes particularly more challenging than problem-solving with arithmetic. Not only do concepts become more difficult (Booth, 2011) but procedures do as well. Procedural knowledge is an important component of mathematical cognition (Star, 2005). Being able to overcome this challenge and maintain or fine-tune one's problem-solving skills may be related to maintaining or increasing levels of interest in mathematics. Experiencing a loss or decrement in problem-solving skills may be related to declines in interest in math as well.

The current study focuses on students who demonstrate this depreciation in procedural problem-solving skills over the course of an algebra unit. In particular, the current analyses explore whether students who show declines in algebraic problem-solving also show a decline in interest in algebra.

Methods

Participants

Middle school students enrolled in Algebra I (N = 367) from classrooms of eight different teachers across three public school districts in the US Midwest participated in the current study. One of the school districts could not release demographics data; this school district made up 41.1% of the sample (n = 151). Of the two school districts that did provide demographic data 50.8% were female, 29.1% were underrepresented minority (URM) students, 14.6% were classified as receiving special education, and 7% were classified as Limited English Proficient by their districts.

Procedure

The current data comes from a larger experimental study that assessed the effectiveness of a worked examples intervention. However, for the purposes of the current research question, the data was collapsed across conditions to increase power due to the small sample size. It is important to note that results were consistent when running the same analyses by condition.

Each classroom completed the study over the course of one of three possible math units. The unit was selected by the teacher to coincide with their syllabus. These units included pre-algebra (58.9%), properties of exponents (15.3%), and solving quadratic equations (25.9%).

Students individually completed a 20-item motivation survey using a seven point Likert scale. This survey assessed various motivation constructs including *interest in algebra*. Interest in algebra is defined as the intrinsic value or enjoyment that one experiences when working on an algebra task (see Simpkins, Davis-Kean, & Eccles, 2006 for a similar interest in math measure). After completing the motivation survey, students completed a unit pre-test. Unit tests were created to reflect material students should have learned within the unit regardless of condition; the problems were based on items from standardized math tests used in public schools. Students were given the entire math period to finish the test. After the pre-test, students then completed four worksheets that coincided with teacher instruction on the particular unit. The motivation survey was re-administered after completion of the unit, followed by a post-test that was identical to the pre-test.

This data was collected to answer the following research question:

Do students who show declines in algebraic problem-solving show a decline in interest in algebra as well?

Measures

Interest in Algebra Participants completed a survey that measured several motivation constructs. The items for interest in algebra asked students to answer using a seven point Likert scale. One represented "Not at all" and seven represented "Yes, definitely." Interest in algebra was measured with three items, which were: "(1) Do you enjoy doing algebra very much?", "(2) Do you think algebra is fun?", and "(3) Do you think algebra is boring? [reverse scored]". Higher means represent high levels of interest in algebra.

Decline in Procedural Skill Students were categorized into two groups based upon whether they showed declines in procedural skill or not. This was done using their performance on the unit tests taken pre- and post-instruction. Procedural problem-solving skill was

demonstrated by knowing how to solve a problem, such as solving a quadratic equation using graphing. Assessments differed based upon whether students were currently studying a pre-algebra unit, a quadratics unit, or a unit on properties of exponents. Pre- and post-test scores were computed by percent of problems solved correctly. Then, a difference score was computed for each student based on subtracting their pre-test score from their post-test score. Most students maintained their skill or demonstrated increases (n = 342). However, 25 students declined from pre- to post-test; these students were categorized as the Declining group. These two categories of students were then compared on interest in algebra from pre- to post-test.

Results

Table 1: Descriptives.

	Declining Skill Group (n = 25)		Non-Declining Skill Group (n = 342)	
	Mean	*SD*	Mean	*SD*
Pre-test Procedural	.542	.291	.190	.191
Post-test Procedural	.371	.265	.437	.249
Pre-test Interest	4.28	1.61	4.03	1.52
Post-test Interest	3.83	1.74	4.14	1.52

Interest in algebra at pre-test was positively and significantly correlated with procedural scores at pre- (r = .258, p < .001) and post-test (r = .235, p < .001). Interest in algebra at post-test was also positively and significantly correlated with procedural scores at pre- (r = .142, p = .007) and post-test (r = .211, p < .001). However, the purpose of the current study was to examine differences in interest for students who showed declines in procedural skills rather than the normative learning pattern. Thus, to answer our research question, each participant was first classified into either the Declining Skill group or the Non-Declining Skill group based on whether their performance in solving procedurally-based problems decreased from pre- to post-test. Descriptives are presented in Table 1. Then, a 2 (Time) x 2 (Group) repeated measures ANOVA was conducted to assess whether changes in interest from pre- to post-test differed by group.

There was no significant main effect of group (F[1, 365] = .014, p = .906, MSE = 4.033, η_p^2 < .001), meaning that groups did not differ by interest collapsed across time points. There was no significant main effect of time (F[1, 365] = 2.009, p = .148, MSE = .643, η_p^2 = .006), meaning that interest did not differ from pre- to post-test collapsed across the two groups. There was a significant group by time interaction (F[1, 365] = 5.814, p = .016, MSE = .643). Participants who did not show declines in procedural skill did not show a decline in math interest from pre-(M = 4.03, SD = 1.52) to post-test (M = 4.14, SD = 1.52). Participants who *did* show declines in procedural skill also showed a decline in math interest from pre- (M = 4.280, SD = 1.61) to post-test (M = 3.827, SD = 1.74).

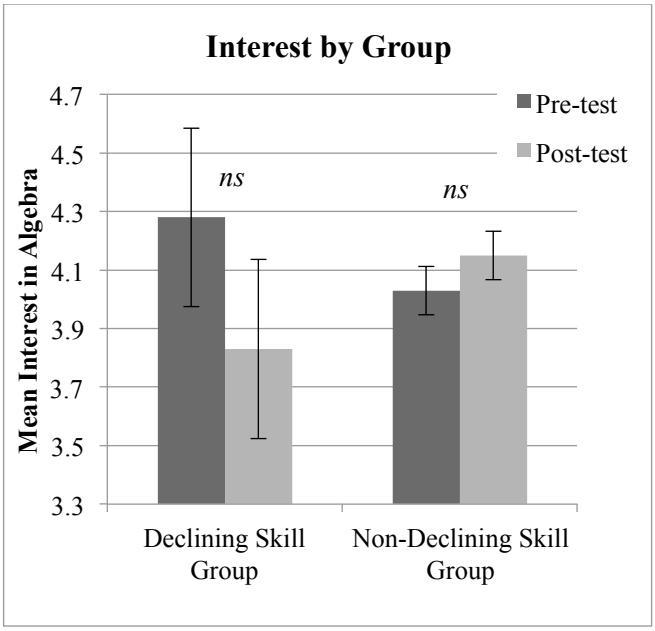


Figure 1: Interest by group.

To test simple effects, two repeated-measures t-tests comparing pre- and post-interest, one per group, were used. The apparent decrease in interest from pre- (M = 4.28) to post-test (M = 3.83) for students in the Declining Skills group was trending toward significance (t [24] = 1.905, p = .069). The apparent increase in interest from pre- (M = 4.03) to post-test (M = 4.15) for students in the Non-Declining Skills group was trending toward significance (t [341] = -1.851, p = .065). Figure 1 displays this finding.

Discussion

It was expected that students who showed declines in interest would also show declines in procedural skills. Our data was consistent with this hypothesis and expands upon prior research. For instance, Wang and Pomerantz (2009) found that positive motivation-orientation is a predictor of math achievement; it is also known that math performance is strongly correlated with interest (Fisher et al., 2012). The results of this study suggest that as a student declines in procedural skills over time, they may also decline in interest. It is important to note that the current data is correlational in nature and directionality of the findings cannot be established. However, we speculate that the findings may be attributed to self-efficacy beliefs, which have been found to have an impact on achievement (Lee, 2009). Low self-efficacy is associated with low levels of other more positive motivation-orientations and achievement. High self-efficacy is associated with high

achievement and more positive forms of motivation (Pajares, 1996).

It is possible that as a student declines in math skill, their self-efficacy beliefs are lowered which in turn lowers their interest in math. After repeated failure experiences, a student may feel less efficacious as well as less interested in the domain. This could perhaps also lead to the student exerting less effort-- leading to even more failure. Moreover, a student's lowered expectations from one failure experience may create a self-fulfilling prophecy and from that, a cycle of below average achievement. However, due to the correlational nature of the data, these speculations need to be explored and validated in future work.

Because of the intricate nature of the development of achievement motivation as well as the quasi-experimental design of the study, we cannot determine the direction of the relationship between interest in math and algebraic procedural skills. Further research should explore the relationship between declines in middle school students' algebraic problem-solving skills and their interest in math. In the interim, it is clear that both students' motivation for math, such as interest in algebra, and procedural skills should be fostered in the classroom as they are both important components of mathematics achievement.

Acknowledgements

Thanks are due to the participating school districts. Research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100150 to Strategic Education Research Partnership Institute. Christina Barbieri's continued contribution to this project was supported by Grant R305B130012 from the U.S. Institute of Education Sciences to the University of Delaware. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Booth, J. L. (2011). Why can't students get the concept of math. *Perspectives on Language and Literacy, Spring*, *2011*, 31-35.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research*, *6* (3), 363-423.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Pantheon.

Fisher, P.H., Dobbs-Oates, J., Doctoroff, G.L., & Arnold, D.H. (2012). Early math interest and the development of math skills. *Journal of Educational Psychology, 104*(3), 673-681.

Kelly, D., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. (2013). *Performance of U.S.15-Year-Old students in mathematics, science, and reading literacy in an international context-first look at PISA 2012(NCES 2014-024)*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences, 19(*3), 355-365.

Lein, A. E., Jitendra, A. K., Starosta, K. M., Dupuis, D. N., Hughes-Reid, C. L., & Star, J. R. (2015). Assessing the relation between seventh-grade students' engagement and mathematical problem solving performance. *Preventing School Failure: Alternative Education For Children And Youth, 60*, 1-7.

Ma, X. (1997). Reciprocal relationships between attitude toward mathematics and achievement in mathematics. *The Journal of Educational Research*, *90*(4), 221-229.

Matthews, M. S., & Farmer, J. L. (2008). Factors affecting the algebra I achievement of academically talented learners. *Journal of Advanced Academics*, *19*(3), 472-501.

Mayfield, K. H., & Glenn, I. M. (2008). An evaluation of interventions to facilitate algebra problem solving. *Journal of Behavioral Education, 17*(3), 278-302.

Middleton, J. A., & Spanias, P. A. (1999). Motivation for achievement in mathematics: Findings, generalizations, and criticisms of the research. *Journal for Research in Mathematics Education, 30*, 65-88.

Pajares, F. (1996). Self-efficacy beliefs and mathematical problem-solving of gifted students. *Contemporary Educational Psychology*, *21*(4), 325-344.

Ryan, R. M., & Deci, E. L. (1996). When paradigms clash: Comments on Cameron and Pierce's claim that rewards do not undermine intrinsic motivation. *Review of Educational Research*, *66*(1), 33-38.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36,* 404-411.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In Grouws, D. *Handbook of research on mathematics teaching and learning* (pp. 334-370). New York: Macmillan.

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology, 42*(1), 70.

Wang, Q., & Pomerantz, E. (2009). The motivational landscape of early adolescence in the United States and China: A longitudinal investigation. *Child Development, 80*, 1272-1287.

Wang, X. (2013). Why students choose STEM majors motivation, high school learning, and postsecondary context of support. *American Educational Research Journal, 50*(5), 1081-1121.

Factors Influencing Categorization Strategy in Visual Category Learning

Sujith Thomas (sujith@cse.iitk.ac.in)
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
India 208016

Harish Karnick (hk@cse.iitk.ac.in)
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
India 208016

Abstract

Studies in visual category learning show that participants use different category generalization strategies. Some studies report a preference for a rule-based strategy, while others report a preference for a similarity-based strategy. We conducted category learning experiments in which we varied three variables — family resemblance of a category, saliency of the defining rule and presentation of transfer stimulus after a delay. Our results show that these factors influence the choice of category generalization strategy. Our study offers a possible explanation for the divergent results in the literature.

Keywords: visual category learning; supervised learning; generalization; family resemblance; defining rule saliency

Introduction

Studies in visual category learning investigate how humans learn and generalize visual categories. Let us say that the items shown in Figure 1 are true instances of a category. The true instances have a common defining feature — the blue five pointed star. Now consider the transfer stimuli shown in Figure 2. Stimulus A has the defining feature (blue star). Stimulus B does not have the defining feature, but is more similar to the true instances shown in Figure 1. For the transfer stimuli in Figure 2, two generalization strategies are possible: defining rule-based strategy and similarity-based strategy.

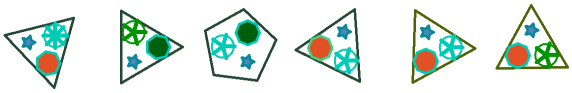


Figure 1: True instances of a target category. All the instances have a common defining feature (blue five pointed star).

In this article, we refer to categorization based on a defining feature as rule-based strategy. If a rule-based strategy is preferred, then participants will generalize stimulus A to the target category, but not stimulus B. On the other hand, if a similarity-based strategy is preferred, then stimulus B will be generalized to the target category, but not stimulus A.

Observational learning is a supervised learning paradigm where the participants are informed about the correct category of the training stimuli. Participants are then asked to categorize the transfer stimuli. No feedback is provided. Some studies in the observational learning literature report a preference for a rule-based strategy; while others report a preference for a similarity-based strategy. In our study, we try to identify the variables that influence the choice of categorization strategy.



Figure 2: Two transfer stimuli for the target category in Figure 1. Stimulus A has the defining feature, but stimulus B is more similar to the target category instances.

Studies on Supervised Learning

In the studies conducted by Xu and Tenenbaum (2007) and Schmidt (2009), observational learning paradigm was used. In each trial, participants were informed about the correct category of three stimuli. The participants were then asked to pick other items that might belong to the same category. Participants preferred a similarity-based generalization. Thomas and Karnick (2014) showed that participants preferred similarity-based generalization even when the target category had a defining feature.

Feldman (2000) showed that categories based on a simple rule (conjunction of boolean features) were learned more easily. Ashby, Maddox, and Bohil (2002) also showed that a rule-based category structure was easier to learn. In (Conaway & Kurtz, 2014; Rabi, Miles, & Minda, 2015; Deng & Sloutsky, 2015), a category could be generalized using either a rule-based strategy or a similarity-based strategy. The results showed that participants preferred a rule-based strategy over a similarity-based strategy.

The studies discussed above indicate that participants use different strategies for visual category generalization. In this study, we try to identify the variables that influence the choice of categorization strategy. Kloos and Sloutsky (2008) showed that participants were able to learn the defining rule when the family resemblance was low, but not when the family resemblance was high. So, family resemblance could be one factor that influences the choice of categorization strategy.

The study conducted by Yim, Castro, Wasserman, and Sloutsky (2014) showed that a category is easier to learn when the defining rule can be identified more easily. So, the ease with which the defining rule can be identified (rule saliency) might also affect the categorization strategy.

In the studies that report similarity-based generalization, the training and transfer stimuli are usually presented simultaneously (Xu & Tenenbaum, 2007; Schmidt, 2009; Thomas & Karnick, 2014). In (Feldman, 2000), training stimuli were presented simultaneously, but the transfer stimuli were presented in a separate testing block. In (Conaway & Kurtz, 2014; Rabi et al., 2015), the training stimuli were presented one by one, and the transfer stimuli were presented in a separate testing block. The results in (Feldman, 2000; Conaway & Kurtz, 2014; Rabi et al., 2015) show a preference for rule-based generalization, and in these studies the transfer stimuli were presented in a separate testing block. So, presenting the transfer stimuli separately could be another factor that might be influencing the choice of categorization strategy.

So the three variables that could be influencing the choice of categorization strategy are: family resemblance of a category, saliency of the defining feature (we will refer to this as rule saliency) and delayed presentation of transfer stimuli. If manipulating the above variables affects categorization strategy, then that would explain the divergent results in the visual category learning literature.

In Experiment 1, we trained a regression model using behavioral data to predict the rule saliency. In Experiment 2, we studied the effect of family resemblance and rule saliency on category generalization. In Experiment 3, we studied the effect of presenting the transfer stimulus after a delay. In the next section, we explain how we quantified the notion of family resemblance and rule saliency. Quantification allows us to study the effect of the independent variables as they change in a continuous manner.

Family Resemblance and Rule Saliency

In Figure 1, the family resemblance between the true instances is high. But, the defining rule (blue star) is not easy to spot. In other words, the rule saliency is low.

Figure 3 shows another set of target category instances. Here the family resemblance between the instances is low. But, the defining rule (red polygon) stands out from the other features; therefore, the rule saliency is high.

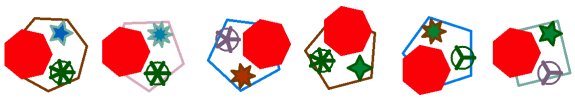


Figure 3: True instances of a target category. All the instances have a common defining feature (red polygon).

In our study, we use abstract figures having ten dimensions, where each dimension can take six distinct values. The stimuli dimensions are — number of outer boundary sides, outer boundary color, number of points in the star, star boundary color, star fill color, number of inner polygon sides, inner polygon boundary color, inner polygon fill color, number of spike-wheel points and spike-wheel color.

We use the concept of entropy to quantify family resemblance of a category. We estimate the entropy using the instances of a category. First, the entropy for each of the ten stimuli dimensions is calculated as follows:

$$dimEnt(j) = \sum_{k=1}^{6} -P(x_{jk}) \log_2 P(x_{jk}) \quad (1)$$

where $dimEnt(j)$ is the entropy of the jth dimension, x_{jk} are the distinct values that the jth dimension can take and $P(x_{jk})$ is the probability with which the value x_{jk} occurs among the instances of a category. The entropy of a category is found by averaging the entropy across all the dimensions as follows:

$$entropy(C) = \frac{1}{10} \sum_{j=1}^{10} dimEnt(j) \quad (2)$$

where $entropy(C)$ is the average entropy of the ten dimensions for category C. In our study, the value returned by Equation 2 is considered to be the category entropy. Category entropy will be low when there is less variability (high family resemblance) among the category instances. Category entropy will be maximum when the non-defining feature values are unique for every instance of the category. The maximum possible value of category entropy in our study is 2.07, because we have six distinct values for each dimension and there can be at most eight non-defining features.

We trained a regression model using behavioral data to quantify the notion of rule saliency. We explain this in more detail in the next section.

Experiment 1

In our experiments, the defining feature is the part of the figure that is common across all the instances of a category. Visual saliency is defined as the perceptual quality that makes a feature stand out from other features and grab our attention (Itti, 2007). In order to quantify visual saliency, we have followed the intuition that participants would take less time to detect the defining feature when it is more salient.

Participants were shown six stimuli like the ones depicted in Figures 1 and 3. Their task was to correctly identify the defining feature as quickly as possible. Participants could choose the defining feature from among the four options that appeared on the screen: boundary, star, polygon and spike-wheel.

In order to change the rule saliency, the area covered by the defining feature, and its color were randomly varied. The time taken by a participant to detect the defining feature was noted. This time was then used to quantify the visual saliency as follows:

$$saliency(i) = \frac{max_p - time_p(i)}{max_p - min_p} \quad (3)$$

where $saliency(i)$ is the saliency of the defining feature for the ith trial, max_p is the maximum response time for the pth participant, min_p is the minimum response time for the pth participant, and $time_p(i)$ is the response time for the ith trial

by the pth participant. In Equation 3, $saliency(i)$ will give a value of one for the trial in which a participant took minimum time to respond. Saliency would be zero for the trial in which a participant took maximum time. The saliency value given by Equation 3 will lie in the range $[0,1]$ for all trials.

Participants and Procedure

There were 22 adult participants (Males=16, Females=6, Mean age=24.27, S.D.=4.33). Participants were volunteers from the student community. Each participant had to respond to 35 trial questions. The first 15 questions were the practice trials, and the remaining 20 were the experimental trials. So the experiment gave us 440 (22 participants $\times$ 20 trials) data points to train our regression model.

In each trial, six abstract figures were shown. These figures were randomly generated. A part of the figure was randomly selected to be the defining feature, and was given a random size and color. Participants were expected to correctly identify the defining feature across the six abstract figures. The response times for only the correct responses were recorded. The maximum time permitted for the participants to identify the common defining feature was 20 seconds.

Modeling Rule Saliency

In this section, we describe the features that were used to train the linear regression model. First, the color value of the different parts of figures were converted from the RGB space to the Lab space. The advantage of Lab space is that the perceptual difference between two colors is proportional to the Euclidean distance between them.

The L value in Lab space corresponds to the lightness component of the color, and the a and b values correspond to the color information. We have used the following six properties of the different parts of the figure: L value, a value, b value, color-saliency, area and perimeter. The color-saliency is estimated using the technique mentioned in (Gelasca, Tomasic, & Ebrahimi, 2005).

For each of the above properties the difference (d_i) between the average defining feature values and the average non-defining feature values was found as follows:

$$d_i = \begin{cases} |v_i - w_i| & \text{if } i = 2 \text{ or } i = 3 \\ v_i - w_i & \text{otherwise} \end{cases} \quad (4)$$

where i ranges from 1 to 6 for each of the six properties mentioned above, v_i is the average value of the ith property for the defining feature and w_i is the average value of the ith property for the non-defining features. For each of the six property values Equation 4 indicates how different the defining feature is from the non-defining features.

In Equation 4, the absolute value of the difference is taken for properties 2 and 3. These two properties correspond to the a and b values in the Lab color space. If the a and b values of the defining feature are very different from the non-defining features, then it might make the defining feature more salient. Here the sign of the difference should not matter. But the sign of the difference matters for the remaining four properties.

For each trial the d_i value for each of the six properties is found. Then the maximum and minimum values for each d_i across all the trials is found. The features for training a linear regression model for saliency is calculated as follows:

$$f_i = e^{\frac{d_i - min(d_i)}{max(d_i) - min(d_i)}} - 1 \quad (5)$$

where i ranges from 1 to 6, and $min(d_i)$ and $max(d_i)$ are the minimum and maximum values of d_i respectively. The main function of Equation 5 is to scale the d_i values so that the corresponding features f_i lies in the range $[0, e-1]$. Features similar to those described in this section have been previously used to detect salient objects and regions of interest in images (Osberger & Rohaly, 2001; Tian, Wan, & Yue, 2010).

Results

The features obtained using Equation 5 were used to train a linear regression model for predicting the saliency values calculated using Equation 3. The linear regression model was trained using the gradient descent algorithm.

After training, the rule saliency values predicted by the saliency model for the category instances in Figures 1, 3 and 4 were .05, .98 and .41 respectively. A value close to zero means that rule saliency is low, and a value close to one means that the rule saliency is high. The stimuli in Figures 1, 3 and 4 were not part of the training or the test data. This shows that the values predicted by the regression model are reasonable.

Experiment 2

In this experiment, we studied the effect of rule saliency and category entropy on the generalization behavior. Figure 4 shows a sample screenshot. In each trial, six true and six false instances of a category, and a transfer stimulus were shown. Participants had to decide whether the transfer stimulus was a true instance of the category. A participant could respond by clicking Yes, No or Can't decide.

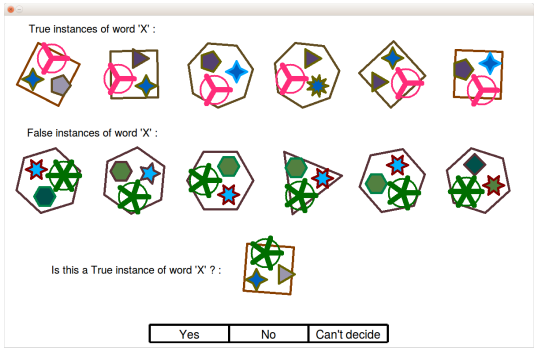


Figure 4: Sample screenshot for Experiment 2. The defining feature is the pink spike-wheel.

In each trial, the target and contrast category instances were generated randomly. A part of the stimuli (boundary, star, polygon or spike-wheel) was randomly selected to be the defining feature. A multinomial distribution was used to assign values to each of the non-defining dimensions. A

Table 1: The four types of transfer stimuli.

	Generated From	Defining Feature
Transfer type 1	Target Category	Present
Transfer type 2	Target Category	Absent
Transfer type 3	Contrast Category	Present
Transfer type 4	Contrast Category	Absent

multinomial distribution associates a probability with each of the six values that a dimension can take. An instance was generated by assigning to each dimension a value that was sampled from the multinomial distribution. In each trial, a random multinomial distribution was assigned for both target and contrast categories.

Category entropy was calculated based on the category instances that were generated in each trial. Category entropy captures the variability present in the instances that were shown in each trial. So, once the instances are generated, the multinomial distribution from which the instances were generated is not important. In each trial, the saliency of the defining feature was varied by changing its color and size.

Four types of transfer stimuli were used in our experiment. Transfer types 1 and 2 were generated using the target category multinomial distribution. Transfer types 3 and 4 were generated using the contrast category multinomial distribution. For the transfer types 2 and 3, the defining features were from the opposite category (i.e. the category from which they were not generated). Table 1 gives the details of the four types of transfer stimuli that were used in our experiments.

The transfer types 1 and 4 were the true and false instances of target category respectively. Transfer type 2 stimuli were similar to the true instances, but did not have the defining feature. Transfer type 3 stimuli had the defining feature, but were not similar to the true instances. If participants used a rule-based strategy, then transfer type 3 stimuli would be generalized to the target category, but not transfer type 2 stimuli. The opposite would happen if participants used a similarity-based strategy.

Participants and Procedure

The 40 adult participants (Males=26, Females=14, Mean age=22.00, S.D.=3.21) in Experiment 2 were volunteers from the student community. Each participant responded to 10 practice trials followed by 20 experimental trials. No feedback was given to the participants during the trials. 800 data points (40 participants $\times$ 20 trials) were obtained from the experiment. There were 640 datapoints corresponding to type 2 and type 3 transfer stimuli. Participants were made to respond to more type 2 and type 3 stimuli because these reveal more about participant generalization behavior. The maximum response time limit was 40 seconds. The generalization strategy of the participants was the dependent variable.

Results

Our results show that participants were able to correctly identify the type 1 stimuli as true instances, and type 4 stimuli as false instances of the target category. The overall percentage of correct, incorrect and can't decide responses for type 1 and type 4 stimuli were 78.75%, 11.25% and 10.0% respectively.

The 3D bars in Figure 5 show how the categorization strategy varied for transfer stimuli types 2 and 3. There were 640 datapoints. The median values of category entropy and rule saliency across Experiments 2 and 3 were used to divide the 640 datapoints into four sets. The four sets corresponded to the high and low values of rule saliency and category entropy. The percentages of similarity-based responses and rule-based responses for each set was found. Figure 5 shows that similarity-based responses (black bar) were preferred over rule-based responses (green bar).

The rule saliency values were obtained from the saliency model, and lie in the range $[0,1]$. Figure 6 shows the effect of rule saliency on rule-based responses (green line). We divided the 640 datapoints into five equal sized bins (each having 128 datapoints). In Figure 6, the numbers in brackets denote the datapoints in each bin. Figure 6 plots the percentage of rule-based responses for each of the five bins. In Figure 6, the x error bars denote the range of rule saliency in each bin.

Figure 6 also shows the effect of rule saliency on similarity-based responses (black line). The x error bars for similarity-based responses (black line) align with the x error bars for rule-based responses (green line) because they correspond to the same five bins. Figure 6 shows that similarity-based responses were preferred over the rule-based responses for different values of rule saliency.

In our study, rule saliency and category entropy are continuous variables. For each trial, rule-based response is a binary variable. Point-biserial correlation coefficient gives the correlation between a continuous variable and a binary variable. We use Pearson product moment correlation because it is equivalent to the point-biserial correlation. The Pearson correlation between rule saliency and rule-based responses was found to be not significant $(r(638) = .09, ns)$ [1].

Figure 7 shows the effect of category entropy on rule-based responses (green line), and similarity-based responses (black line). The figure shows that similarity-based responses are preferred over rule-based responses; except when entropy is very high. When entropy is very high, the non-defining feature values of category members are very different from each other. In this scenario, similarity will be mostly determined by the defining feature. Hence, for high values of category entropy there will be no difference between similarity-based categorization and rule-based categorization.

The Pearson correlation between category entropy and rule-based responses was found to be significant $(r(638) = .12, p < .01)$. The maximum value that the category entropy took in Experiment 2 was 1.85 (maximum possible value is 2.07).

[1] Here, $df = 638$ because there were 640 datapoints.

The results of Experiment 2 show that participants preferred similarity-based responses over rule-based responses. Category entropy had a significant effect on categorization strategy; but rule saliency did not.

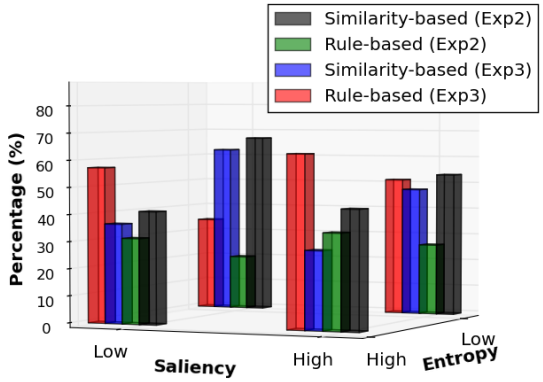


Figure 5: Rule-based and similarity-based responses for transfer stimuli types 2 and 3 in Experiments 2 and 3.

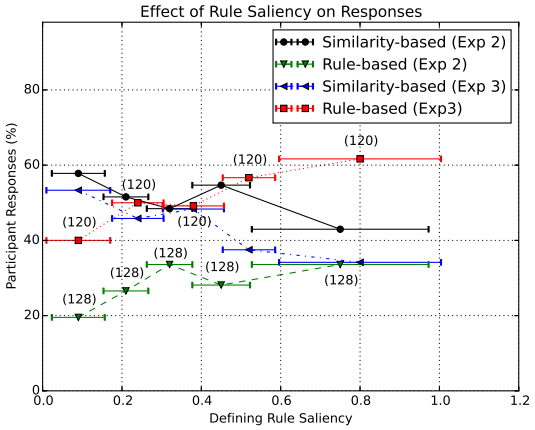


Figure 6: Effect of rule saliency on participant responses for transfer types 2 and 3.

Experiment 3

In Experiment 2, all stimuli were shown simultaneously. In Experiment 3, we wanted to study whether showing the transfer stimulus after a time delay will lead to a preference for rule-based responses.

In Experiment 3, participants were shown true and false instances of a category for a maximum duration of 40 seconds — just as in Experiment 2. When participants clicked the next button the screen would go blank for six seconds, and then the transfer stimulus would be presented. As before, a participant could select one of the three options — Yes, No or Can't decide. Apart from the delayed presentation of the transfer stimulus, every other detail remained the same as that in Experiment 2.

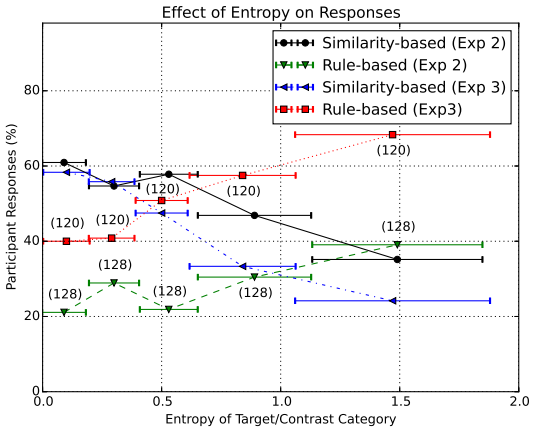


Figure 7: Effect of category entropy on participant responses for transfer types 2 and 3.

Participants and Procedure

We had a new set of 50 adult participants (Males=30, Females=20, Mean age=21.88, S.D.=3.73). The participants were volunteers from the student community. Each participant responded to 10 practice trials followed by 16 experimental trials. No feedback was given to the participants during the trials. We obtained 800 data points (50 participants $\times$ 16 trials) from the experiment. Out of 800 data points, 600 corresponded to type 2 and type 3 transfer stimuli.

Results

Our results show that participants were able to correctly categorize the type 1 stimuli as true instances, and type 4 stimuli as false instances of the target category. The overall percentage of correct, incorrect and can't decide responses for type 1 and type 4 stimuli were 87.5%, 10.0% and 2.5% respectively.

Figure 5 shows how the categorization strategy varied for transfer stimuli types 2 and 3. There were 600 datapoints. The median values of rule saliency and category entropy were used to divide the 600 datapoints into four sets, that corresponded to the high and low values of rule saliency and category entropy. Figure 5 shows that rule-based responses (red bar) were greater than similarity-based responses (blue bar) when either rule saliency or category entropy was high.

Figure 6 shows the effect of rule saliency on categorization strategy. The 600 datapoints were divided into 5 equal sized bins (each having 120 datapoints). Figure 6 shows that for higher values of saliency rule-based responses (red line) were preferred over the similarity-based response (blue line). The rule-based responses were found to be correlated with saliency $(r(598) = .15, p < .001)$ [2].

Figure 7 shows that rule-based responses (red line) were preferred over similarity-based responses (blue line) for high values of entropy. The rule-based responses were found to be correlated with entropy $(r(598) = .20, p < .0001)$.

[2]Here, $df = 598$ because there were 600 datapoints.

In Experiment 2, the number of rule-based and non-rule-based responses for type 2 and type 3 stimuli were 181 and 459 respectively (total 640 datapoints). In Experiment 3, the number of rule-based and non-rule-based responses for type 2 and type 3 stimuli were 309 and 291 respectively (total 600 datapoints). The difference between the observed frequencies of responses across the two experiments was found to be significant ($\chi^2(1, N = 1240) = 68.88, p < .0001$).

The results of Experiment 3 show that presenting the transfer stimulus after a delay caused a significant increase in rule-based responses.

Discussion and Conclusion

The aim of our study was to identify the variables that cause participants to prefer one categorization strategy over another. When the training and transfer stimuli were presented simultaneously, participants preferred similarity-based generalization. When transfer stimuli was presented after a six seconds delay, participants preferred rule-based responses for higher values of rule saliency and category entropy.

Our Experiment 2 results are consistent with the results in (Xu & Tenenbaum, 2007; Schmidt, 2009; Thomas & Karnick, 2014), that report a preference for similarity-based generalization. In (Kloos & Sloutsky, 2008), the transfer stimuli was presented in a separate testing block. Our Experiment 3 results show that participants prefer similarity-based generalization when category entropy is low, but rule-based generalization is preferred when entropy is high. This result is consistent with (Kloos & Sloutsky, 2008). In (Feldman, 2000; Conaway & Kurtz, 2014; Rabi et al., 2015), transfer stimuli was presented in a separate testing block, and participants preferred rule-based generalization. The Experiment 3 results can explain the above results as well.

When transfer stimulus is presented after a delay, participants use their short term memory to remember the training stimuli. Visual short term memory has a limited capacity (Todd & Marois, 2004). So, remembering all the features of the training stimuli would become difficult. This might be causing the participants to use a simple rule-based strategy, which puts less demand on the short term memory.

We have quantified family resemblance using category entropy, which does not take into account the saliency of different features. This allows us to independently vary category entropy and rule saliency. But, when defining rule saliency is increased it should lead to greater family resemblance. Category entropy does not take this effect into account.

Many studies in the literature make use of artificial categories. Artificial stimuli allow us to study the learning behavior in the absence of prior knowledge about the categories.

In this study, we have manipulated three variables: family resemblance, rule saliency and delayed presentation of transfer stimulus. Our results show that these variables influence the choice of categorization strategy. Our study offers possible explanations for the divergent results in the visual category learning literature.

References

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & cognition*, *30*(5), 666–677.

Conaway, N., & Kurtz, K. J. (2014). Now you know it, now you dont: Asking the right question about category knowledge. In *Proceedings of the 36th annual meeting of the Cognitive Science Society* (pp. 2062–2067).

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental psychology*, *51*(3), 392–405.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Gelasca, E. D., Tomasic, D., & Ebrahimi, T. (2005). Which colors best catch your eyes: a subjective study of color saliency. In *First international workshop on video processing and quality metrics for consumer electronics.*

Itti, L. (2007). Visual salience. , *2*(9), 3327. (revision 72776)

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52–72.

Osberger, W. M., & Rohaly, A. M. (2001). Automatic detection of regions of interest in complex video sequences. In *Photonics west 2001-electronic imaging* (pp. 361–372).

Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of experimental child psychology*, *131*, 149–169.

Schmidt, L. A. (2009). *Meaning and compositionality as statistical induction of categories and constraints*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Thomas, S., & Karnick, H. (2014). Is word generalization for novel concepts modelled by similarity or by formal concepts? In *Cognitive 2014, the sixth international conference on advanced cognitive technologies and applications* (pp. 274–280).

Tian, M., Wan, S., & Yue, L. (2010). A color saliency model for salient objects detection in natural scenes. In *Advances in multimedia modeling* (pp. 240–250). Springer.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751–754.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in bayesian word learning. *Developmental science*, *10*(3), 288–297.

Yim, H., Castro, L., Wasserman, E. A., & Sloutsky, V. M. (2014). The interactions of category structure and supervision in category learning: a comparative approach. In *Proceedings of the 36th annual meeting of the Cognitive Science Society* (pp. 1814–1819).

Simulating Developmental Changes in Noun Richness through Performance-limited Distributional Analysis

Daniel Freudenthal[1], Julian M. Pine[1], Gary Jones[2] and Fernand Gobet[1]
[1]Department of Psychological Sciences, University of Liverpool
[2]Division of Psychology, Nottingham Trent University

Abstract

In this paper we examine how a mechanism that learns word classes from distributional information can contribute to the simulation of child language. Using a novel measure of noun richness, it is shown that the ratio of nouns to verbs in young children's speech is considerably higher than in adult speech. Simulations with MOSAIC show that this effect can be partially (but not completely) explained by an utterance-final bias in learning. The remainder of the effect is explained by the early emergence of a productive noun category, which can be learned through distributional analysis.

Keywords: Language Acquisition; Learning biases; Productive noun use.

Introduction

A key question in the study of language acquisition is how children build word classes such as noun and verb. One source of information that children might employ in this process is the distributional properties of the language they are exposed to – the lexical environments in which words occur can act as a cue to a word's grammatical category. English words that are preceded by determiners such as *a* and *the* and followed by (auxiliary) verbs such as *can* and *will* tend to be nouns, while words that are preceded by (pro)nouns like *I* and *You* and followed by determiners such as *a* and *the* tend to be verbs.

Several approaches to this problem have been proposed, but they tend to focus on building large word classes with high accuracy rather than developing mechanisms that can be plausibly applied by, and fit developmental data from language-learning children. Thus, mechanisms for distributional analysis routinely collect data from large corpora and entire utterances, and make limited contact with the (developmental) child data. Recent work by Freudenthal et al. (this volume) has shown that it's possible to perform distributional analysis developmentally by embedding it in an existing model of language acquisition that learns to represent increasingly long utterances. In this paper we examine how the developmental emergence of the classes of verb and noun interacts with the model's learning biases to explain actual developmental child data – the decreasing ratio of nouns to verbs in child speech.

Two influential approaches to learning word classes from distributional information are those of Redington, Chater and Finch (1998), and Mintz (2003). Mintz (2003) introduces the notion of a *frequent frame* – a combination of two words with one word intervening (e.g. *You* X *A*). Mintz restricts his analysis to the 45 most frequent frames for a given corpus, and finds that the words that co-occur within a given frame tend to belong to the same category. The relative simplicity of Mintz's approach suggests it is well within the capabilities of language-learning children. However, it has been argued that it doesn't work very well in languages with relatively free word order, such as Dutch (Erkelens, 2008), or German (Stumper et al. 2011). It has also been argued that the approach is too sensitive to noise, and relatively poor at classifying nouns (Freudenthal et al. 2013).

In Redington et al.'s approach, the contexts for words that are to be classified (*target words*) are expressed as vectors containing counts for a number of high frequency *context words* in preceding and following position. Target and context words are typically restricted to the 1000 and 150 most frequent words for a given corpus. Similarity is expressed as the (rank-order) correlation between these vectors and the matrix of correlations is used as input for a cluster analysis. Redington et al.'s approach appears more robust. However, it has been criticized for having a high computational overhead (St. Clair et al. 2010). The fact that the approach requires the child to track frequencies of large numbers of words makes it less plausible as a mechanism that might be used by language-learning children.

While both approaches are intended as mechanisms that language-learning children use, neither approach incorporates a developmental component. That is, both approaches collect statistics from large corpora and complete utterances, and thus ignore the fact that children's early utterances are no more than one or two words long, and only gradually increase in length over a period spanning several years. While children may well attend to longer utterances than they produce, it seems unlikely that they process entire utterances from a young age. A mechanism that collects statistics across complete utterances may thus use information that is not available to language-learning children in the early stages of development.

The importance of development is further underscored by experimental evidence which suggests that children's productive use of words develops at different speeds for different word classes. In particular, data from production studies suggest that children develop a highly productive noun category before they develop a fully productive verb category (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1999; Tomasello & Olguin, 1993).

Recent work by Freudenthal et al. (this volume) has shown that Redington et al.'s approach can be amended in a way that makes it far more plausible as a mechanism used by language-learning children. Specifically, Freudenthal et al. show that it is possible to add a developmental

component to the mechanism by gradually expanding the contexts to which it is exposed in a developmentally plausible way. Freudenthal et al. did this in the context of MOSAIC (Freudenthal et al. 2007, 2010, 2015), an existing computational model that has been used to successfully simulate a number of key phenomena in language acquisition. MOSAIC is a simple learning model that takes as input orthographically transcribed child-directed speech. Training in MOSAIC takes place by feeding the input corpus through the model multiple times. A key feature of MOSAIC's learning mechanism is that it builds up its representation of the utterances to which it is exposed by starting at the right edge of the utterance and slowly working its way to the left. Thus, with each exposure to the input MOSAIC represents increasingly long utterance-final phrases that become increasingly adult-like. The utterance-final bias is MOSAIC's key mechanism for simulating cross-linguistic differences in children's early speech. MOSAIC's utterance-final bias is therefore independently motivated, and MOSAIC provides a natural setting for investigating developmental variants of mechanisms that build word classes on the basis of distributional information. A detailed description of MOSAIC and the way it is trained is provided in Freudenthal et al. (2015).

Freudenthal et al. (this volume) also show that it is possible to significantly simplify the mechanism of Redington et al. Redington et al. collect counts for a fixed number of context words in preceding and following position, and express similarity as the rank-order correlation between vectors containing these counts. Since the rank-order correlation is non-parametric in nature, the frequency information contained in the counts is largely ignored. Freudenthal et al. therefore test the performance of a mechanism that does away with token frequency information altogether and expresses the context for a word as a simple list of words (types) in preceding and following position. Similarity between words is then expressed as the amount of overlap between these lists of words. It is shown that, given the corpora employed, this simplified version performs as well as the rank-order correlation.

The fact that Redington et al.'s approach can be applied in a developmental setting and can be simplified considerably greatly increases its plausibility as a mechanism that can be used by language-learning children. This plausibility would be enhanced even further if the mechanism could be shown to provide a good fit to child data. Freudenthal et al. provide a first attempt at this by tracking the mechanism's developing ability to classify nouns and verbs. Using a measure of 'noun richness', they show that, when the mechanism is embedded within MOSAIC, it tends to link together nouns in the early stages, with verbs coming in in the later stages. This is due to the fact that nouns frequently appear in utterance-final position, and MOSAIC therefore registers contexts for (and thus classifies) nouns earlier than it registers contexts for verbs. The number of items that are classified in the early stages, however, is rather low, and Freudenthal et al. argue that the inclusion of utterance boundaries as framing elements could potentially increase the number of classified items.

The main aim of this paper is to further investigate the possibility that an utterance-final learner that builds word classes on the basis of distributional information can successfully simulate aspects of child speech. To this end, the noun richness score used by Freudenthal et al. is developed further to provide a descriptive measure of child speech, and thus a target for simulation. Our analyses are organised as follows.

A first analysis will focus on the relative numbers of nouns and verbs in the speech of children and their mothers. If, as hypothesized, children show highly productive use of nouns and less productive use of verbs in the early stages, one would expect the ratio of nouns to verbs in child speech to exceed that in adults.

A second analysis will focus on the extent to which the utterance-final bias instantiated in MOSAIC can account for a potential advantage for nouns in child speech. While a noun advantage is consistent with early productivity around nouns, it is also possible that it reflects the operation of an utterance-final bias in learning. This possibility will be investigated by taking the child-directed speech from the corpora used in the first analysis, and using them as input for MOSAIC models. Rote[1] output from the models will be generated at various MLU (Mean Length of Utterance) points, and the noun richness scores at various MLU points will be compared with those of the actual children.

A third analysis will determine whether the mechanism for building word classes as described by Freudenthal et al. (this volume) can improve MOSAIC's fit to the children's noun richness data. This will be done by taking the MOSAIC models at the different MLU points and linking together words that are distributionally similar based on the contexts that are represented in the models. The development of noun and verb classes will be tracked by counting the relative numbers of nouns and verbs that are classified together. Next, the rote output files from analysis 2 will be used to generate novel utterances by substituting words that have been linked on the basis of the distributional analysis, and the resulting noun richness scores will be compared to the relevant child data.

To preview the results, the analysis of child speech will show that, early in development, children's noun richness scores are more than 30 percentage points higher than those of adults. Across the MLU range studied here, children's noun richness scores drop substantially until they approximate those of adults. Simulations with rote output from MOSAIC (analysis 2) show that an utterance-final bias can account for roughly half of this decrease in noun richness (a drop of around 15%).

Substitution of distributionally similar words (analysis 3) can increase initial noun richness scores by an additional 10 percentage points, without affecting the fit to the later stages. However, this requires a substantial early noun class,

[1] Rote output contains no substitutions, and is learned directly from the input.

something that can be achieved by including utterance endings as framing elements.

Taken together, these findings suggest that children's early noun richness is jointly determined by utterance-final learning and increased productivity around nouns relative to verbs. They also confirm the viability of distributional analysis for building word classes in a developmental setting. However, they suggest that a plausible distributional analysis mechanism must show an early advantage for nouns, and thus cast doubt on mechanisms like that of Mintz (2003), which has been shown to be biased towards classifying verbs (Freudenthal et al. 2013).

Child Noun Richness

A first analysis was aimed at determining the relative rates of verb and noun use by children and their mothers. If children show productivity around nouns earlier than around verbs, one would expect children to show high rates of noun use in the early stages of development. The analysis was carried out on all 12 children from the Manchester corpus (Theakston et al. 2001). Corpora for individual children consist of 34 fortnightly recordings of interactions between child and mother, starting at a child age of approximately 2 years. Recordings are transcribed and include a MOR tier, which contains morphological coding, including Part of Speech (PoS) information for individual words.

For the current analysis, the MOR line was used to identify nouns and main verbs in the child and maternal speech. Noun richness was then expressed as the number of nouns divided by the number of nouns plus verbs. The analysis thus disregards function words and modal/auxiliary verbs (which are frequently omitted by children), and focuses on the two main classes of content words. Figure 1 presents the children's and mothers' noun richness scores. Data points represent averages for the 34 individual tapes contained in each corpus, with MLU computed for the child speech. Data reported in Fig. 1 was computed on the basis of utterance types – duplicate utterances were removed on a tape-by-tape basis. Fig. 1 shows a clear decrease in the children's noun richness scores (from .78 to .45), while the mothers' scores range between .42 and .46. Fig. 1 thus shows a clear advantage for nouns in the children's (early) speech, a finding that is consistent with the notion that young children show greater productivity around nouns than they do around verbs.

The data in Fig. 1 are consistent with greater early productivity around nouns. However, it is also possible that they reflect a processing bias for utterance-final phrases. Nouns tend to occur in utterance-final position, and the children's high early noun richness scores could potentially be explained by an utterance-final bias in learning. This possibility was examined by measuring noun richness in the output of MOSAIC models trained on the maternal speech for the 12 individual children in the Manchester corpus. The input was fed through the model multiple times, and output of increasing average length was generated after each exposure to the input. Average noun richness and MLU for the models over a range of runs are shown in Table 1. The input was fed through the models a total of 50 times, and (declarative) output was generated for runs 30 through 50. Since there is no morphology tier in the model output, words in MOSAIC's output were assigned to their most frequent PoS tag based on the MOR-lines across the Manchester corpus. As with the child analysis, duplicate utterances were removed from the analysis. As can be seen in Table 1, the model output shows a decrease in noun richness of approximately 15 percentage points between MLUs 1.5 and 4.7. The data in Table 2 therefore indicate that, while the utterance-final bias instantiated in MOSAIC can account for some of the early advantage for nouns, it is not sufficient to explain the size of the effect found in children.

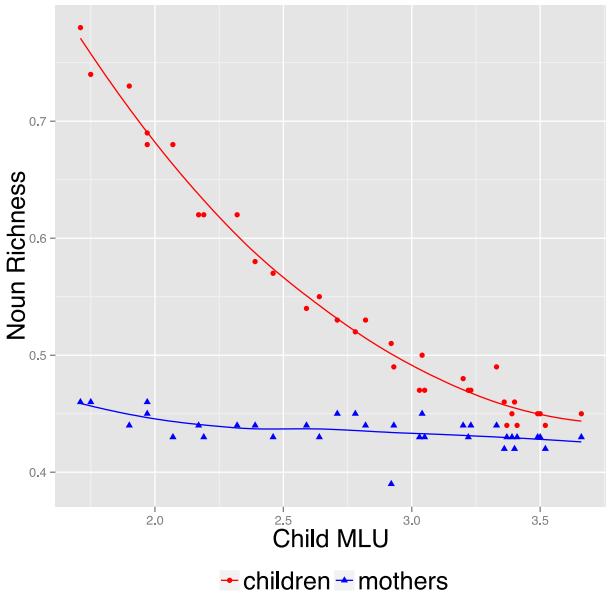


Fig. 1: Average noun richness scores for children and mothers in the Manchester corpus.

Table 1: Summary data and noun richness for rote output from MOSAIC models ranging from 30 to 50 exposures to the input. Data are averaged over 12 models

Run	MLU	Lines	Nouns	Verbs	Noun Richness
30	1.49	173	34	18	0.65
32	1.67	429	108	56	0.65
34	1.81	960	304	168	0.64
36	2.01	2,081	799	455	0.64
38	2.23	4,158	1,812	1,070	0.63
40	2.51	7,551	3,596	2,356	0.61
44	3.34	16,052	9,070	7,438	0.55
50	4.69	18,702	12,976	13,696	0.49

Building a Noun Category

In this section we investigate if early productivity around nouns can provide an additional source of noun richness. This will be done by linking together items based on the

similarity of the contexts in which they occur, and substituting distributionally similar words in MOSAIC's output. Early work by Redington et al. has shown that such a distributional analysis can result in accurate word classes. However, Redington et al.'s method, has been criticized for carrying a high computational overhead (St. Clair et al. 2010). Moreover, it has not been applied in a developmental setting. Freudenthal et al. (this volume) show that a substantially pared down version of Redington et al.'s method can perform as well as the original mechanism. Specifically, the mechanism used by Freudenthal et al. does away with token frequency information. Rather than expressing the context for a given word as token counts for (150) context words in preceding and following position, context is presented as a simple list of words (types) that occurred in preceding and following position. Similarity is then expressed as a measure formally known as the *Jaccard* distance: the length of the intersection of the contexts divided over the length of the union. Two words that were preceded by {*a*, *the*} and {*a*, *green*} respectively thus have a similarity of 1/3.

Freudenthal et al. also show that the mechanism can work developmentally by applying it to the contexts represented in MOSAIC models at different points in training. More specifically, using the input corpora from the Manchester corpus (as used above), they find that the models tend to link together nouns in the early stages of development, with a verb category emerging in the later stages: Noun richness, defined as number of noun-noun links over noun-noun plus verb-verb links decreases over the model's development. The number of items that is classified, however, is relatively small, particularly early in development. Over runs 36 to 50 the average number of linked items ranges from under 30 to around 700 (see Table 2).

Table 2: Number of links and accuracy scores for Jaccard distance at different points in development. Scores are averaged over all children in the Manchester corpus. Two words are linked together if the Jaccard distance exceeds 0.2, both for preceding and following contexts. The analysis is limited to the 1000 most frequent words of each corpus

Run	# of links	Acc.	Noun Acc.	Verb Acc.	Noun Richness
36	27	0.78	0.83	0.50	0.72
38	140	0.83	0.84	0.55	0.80
40	370	0.87	0.88	0.68	0.80
44	648	0.89	0.86	0.82	0.64
50	717	0.91	0.88	0.87	0.57

However, Freudenthal et al. only consider lexical items as contexts. Earlier work has shown that the inclusion of utterance boundaries can greatly increase the number of items that are classified together. For the current simulations it was therefore decided to include utterance endings as context elements. Initial simulations showed this results in increased linking of words, but also in a decrease in accuracy scores, particularly for verbs (the highest verb accuracy achieved was .65 for run 50). Given the non-parametric nature of the Jaccard distance, this is not surprising, since an utterance boundary in the shared context can override several non-matching lexical items. It was therefore decided to give precedence to lexical context items by disregarding utterance boundaries if they were the only shared context despite lexical items being available. This essentially gives greater weight to lexical context items, if present, but still allows linking on the basis of utterance endings if no lexical items are present.

Table 3: Average number of links and accuracy scores for Jaccard distance at a threshold of 0.2. Utterance endings are included unless lexical items are present in following context but these show no overlap.

Run	# of links	Noun links	Acc.	Noun Acc.	Verb Acc.	Noun Richness
30	1	0	0.24	0.04	0.13	0.08
32	11	2	0.39	0.51	0.08	0.70
34	95	72	0.76	0.84	0.21	0.96
36	554	464	0.86	0.90	0.21	0.98
38	1,278	1,035	0.84	0.86	0.24	0.98
40	1,475	1,126	0.83	0.85	0.43	0.94
44	1,305	926	0.89	0.88	0.74	0.81
50	1,452	976	0.91	0.89	0.81	0.75

As can be seen in Table 3, inclusion of the utterance boundary greatly increases the number of linked items, particularly for the early stages (by a factor of 20 for run 36). Noun accuracy scores are comparable to those in Table 2, though verb accuracy scores are lower and only exceed .6 from run 42 onwards. However, the data in Table 3 show clear emergence of an early noun category, and a steady decrease in noun richness scores from run 36 onwards. They therefore confirm that distributional analysis can be meaningfully applied in a developmental setting and can result in reasonable accuracy scores, even when including utterance endings as framing elements.

Simulating Child Noun Richness

A final analysis was aimed at investigating whether the word classes that were built using distributional analysis are sufficient to explain the pattern of noun richness displayed by the children. This was done by taking the output files from the MOSAIC models (reported in Table 1) and using the links reported in Table 3 to substitute words that are distributionally similar. Substitution of distributionally similar words allows the model to produce novel utterances that were not present in its rote output. The early emergence of a large noun class is likely to result in many substitutions of nouns in novel contexts[2], and therefore in higher noun richness scores than for the rote output. Table 4 gives noun richness scores for the models in runs 34 through 50 with

[2] Substitution can also lead to duplicate utterances. However, since analyses are conducted on utterance types, duplicates are removed from the analysis.

substitutions based on the relevant links for a given run. Substitution can increase the number of utterances that are generated considerably. For output files containing more than 3,000 utterances, substitution was therefore carried out on a random sample of 3,000 output utterances.

It is apparent from Table 4 that (early) noun richness scores are increased relative to those in Table 2. Across the range of runs, noun richness decreases from .73 to .45, a decrease of nearly 30 percentage points, and provides a better fit to the child data. It is also apparent, however, that MLU scores are increased relative to the rote output, particularly for the later stages. This increase in MLU is caused partly by more words being linked in the later stages of development. However, it also reflects the fact that longer utterances provide more opportunities for substitution, and thus contribute more novel utterances than short utterances[3].

Table 4: Average noun richness scores for MOSAIC output with substitutions

Run	MLU	Utterances	Nouns	Verbs	Noun Richness
34	2.03	1,435	612	267	0.67
36	2.37	4,503	2,798	971	0.73
38	2.58	8,169	5,679	2,191	0.72
40	2.98	11,346	8,933	4,342	0.67
44	4.26	25,571	23,939	19,299	0.55
50	5.44	35,578	32,628	38,854	0.45

Table 5: Noun richness scores for MOSAIC output with substitutions sampled back to rote MLU distribution

Run	MLU	Utterances	Nouns	Verbs	Noun Richness
34	1.81	960	387	169	0.68
36	2.01	2,081	1,179	405	0.74
38	2.23	2,979	1,953	686	0.74
40	2.52	3,000	2,135	945	0.69
44	3.34	3,000	2,305	1,612	0.59
50	4.70	3,000	2,330	2,540	0.47

Table 5 presents the results of a noun richness analysis that controls for this increase in MLU by sampling from the generated output to match the MLU profile of the rote output. That is, for every utterance of length N in the original output, 1 utterance of length N is sampled from the output with substitutions. As can be seen in Table 5, this procedure has minimal effects on the noun richness scores, indicating that the substitution procedure results in a genuine increase in (early) noun richness scores.

[3] In fact, allowing all substitution quickly makes the output file too large to manage. The likelihood of individual substitutions is therefore gradually decreased across runs. Additionally, to prevent them from dominating the output, no substitutions are made in utterances longer than 6 words.

Conclusions

Several authors have argued that children build word classes on the basis of distributional information. The main proposed mechanisms have either been computationally expensive or too sensitive to noise, and fail to incorporate a developmental component. Evaluation has also tended to focus on the standard mechanisms of accuracy and completeness rather than empirical child data.

Freudenthal et al. (this volume) have shown that it is possible to take Redington et al.'s (1998) mechanism and apply it in a developmental setting. They have also shown that it is possible to substantially reduce its computational complexity without affecting its performance. In terms of evaluation, the revised mechanism shows a pattern of noun and verb linkage that appears qualitatively plausible, but may not build classes that are sufficiently large.

The current work builds on the work of Freudenthal et al. (this volume) by introducing a novel evaluation metric and target for simulation. The analysis of noun richness scores confirms that children have an early preference for the production of nouns relative to verbs. To our knowledge, this is the first demonstration of such a preference in corpus analyses. This finding is consistent with the notion that children show earlier emergence of a noun than a verb category, but may also reflect the operation of an utterance-final bias in learning.

The analysis of MOSAIC's rote output suggests that approximately half of children's early noun preference can be explained through utterance-final learning. This finding is encouraging as it provides independent support for the utterance-final bias instantiated in MOSAIC. However, it is also apparent that the utterance-final bias as instantiated in MOSAIC is insufficient to explain the size of children's early noun preference, and therefore suggests that some of the early noun advantage may be explained by differences in children's early productivity around nouns and verbs.

This suggestion was investigated by performing a distributional analysis on the basis of the contexts encoded in MOSAIC in different stages of development. In line with suggestions by Freudenthal et al. (this volume), utterance endings were included as framing elements. This was shown to increase the number of classified items, but had a detrimental effect on classification accuracy, in particular for verbs. It was argued that this effect could be countered by disregarding utterance boundaries when lexical context items are present but show no overlap, effectively weighting contexts for their lexical content. Doing so greatly increased the number of links for the early stages of development, whilst maintaining reasonable accuracy scores, particularly for the later stages of development.

Finally, it was shown that, with the exception of the very early stages, substitution of linked items resulted in higher noun richness scores and a better fit to the child data. Thus, the distributional analysis implemented in MOSAIC was able to build an early noun class that was sufficient to raise early noun richness to levels close to those displayed by young children.

Taken together, the results described here provide converging evidence that the productivity of children's early noun category develops more quickly than the productivity of their early verb category (Tomasello, 1992). They also suggest that distributional analysis is a viable mechanism for building word classes, even at early stages of development, when relatively few contexts may be available. The analyses reported here suggest that, provided utterance endings are included as framing elements, a variant of Redington et al.'s mechanism can form an initial noun class that is large enough to simulate children's early noun richness scores.

The findings reported here also cast doubt on the feasibility of the frequent frames approach advocated by Mintz (2003). Freudenthal et al. (2013), show that, unlike Redington et al.'s mechanism, frequent frames tend to classify together verbs, rather than nouns. Mintz's approach is thus unlikely to be successful in simulating (early) child noun richness scores. It is also difficult to see how Mintz's approach might be modified to incorporate a developmental component. Freudenthal et al. (2013) show that, while inclusion of utterance boundaries increases the numbers of nouns classified by frequent frames, this has dramatic negative effects on classification accuracy. While the analyses reported here suggest that inclusion of utterance endings may negatively affect accuracy for Redington et al.'s mechanism, they also show that this can be remedied by weighting framing elements for lexical content.

The analyses reported here further illustrate the strength of our approach, which embeds a mechanism for learning word classes in an existing model of language acquisition to simulate developmental variation in children's production of verbs and nouns. By comparing rote and productive output we were able to show that children's early noun richness is jointly determined by an utterance-final bias in learning and early productivity around nouns. We were also able to show that the developmental biases (i.e. contexts encoded) in the model are sufficient for a variant of Redington et al.'s mechanism to provide such an advantage for nouns, thus providing evidence for both these developmental biases and the feasibility of distributional analysis.

Acknowledgements

Daniel Freudenthal, Julian Pine, and Fernand Gobet are members of the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool, for which support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. This research was supported by ESRC Grant ES/J011436/1.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*, 952-965.

Erkelens, M. A. (2009). *Learning to categorize verbs and nouns.* Unpublished PhD Thesis, Universiteit van Amsterdam, Amsterdam.

Freudenthal, D., Pine, J. M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, 311-341.

Freudenthal, D., Pine, J. M. & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language*, *37*, 643-669.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet, F. (2013): Frequent frames, flexible frames and the noun-verb asymmetry. In: M. Knauf, M. Pauen, N. Sebanz E I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 2327-2332). Austin, TX: Cognitive Science Society.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet. F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61-76.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet, F. (this volume). Developmentally plausible learning of word categories from distributional statistics. *Proceedings of the 38th annual meeting of the Cognitive Science Society.*

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk (3rd Edition)*. Mahwah, NJ: Erlbaum.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91-117.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.

Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A powerful cue for acquiring syntactic structures. *Cognitive Science*, 22, 425-469.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). Frequent frames in German child-directed speech: A limited cue to grammatical categories. *Cognitive Science, 35*, 1190-1205.

St. Clair, M.C. Monaghan, P., & Christiansen, M.H. (2010). Learning grammatical categories from distributional cues. Flexible frames for language acquisition. *Cognition*, *116*, 341-360.

Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of Verb-Argument structure: An alternative account. *Journal of Child Language*, *28*, *127-152*.

Tomasello, M. (1992). First verbs: A case study of early grammatical development. Cambridge: CUP.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development, 8*, 451-464.

The Comprehension of English Garden-path Sentences by Mandarin and Korean Learners of English as a Second Language

Zhiying Qian (zqian3@illinois.edu)
University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Eun-Kyung Lee (eunkyunglee@yongsei.ac.kr)
Yonsei University, Korea

Dora Hsin-Yi Lu (doralu@tea.ntue.edu.tw)
National Taipei University of Education, Taiwan

Susan M. Garnsey (sgarnsey@psych.illinois.edu)
University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

Abstract

How the properties of a first language (Mandarin, Korean) influence the comprehension of sentences in a second language (English) was investigated in a series of self-paced reading time studies. Native Mandarin- and Korean-speaking learners of English were compared with native English speakers on how they resolved a temporary ambiguity about the relationship between a verb and the noun following it in a sentence (e.g., *The club members understood [that] the bylaws would be applied to everyone.*). Frequency biases of verbs' subcategorization structure (direct-object-bias vs. sentential-complement-bias) was manipulated in Experiment 1. Results showed that L1-Mandarin learners of L2-English were able to use both the verb bias and the complementizer cue, and their usage of these cues was not modulated by proficiency. L1-Mandarin learners' use of the verb bias cue contrasts with previously reported findings with L1-Korean learners of L2-English, who showed sensitivity to verb bias only in higher proficiency learners (Lee, Lu, & Garnsey, 2013). The difference between L1-Mandarin and L1-Korean learners suggests that L1 word order (Mandarin & English, SVO; Korean SOV) influences how quickly L2 learners learn word-order-dependent cues about structures in the L2. Experiment 2 added plausibility manipulation (e.g., *The club members understood the bylaws/the pool*…). Neither the native speakers or the L2 groups (L1-Mandarin L2-English & L1-Korean L2-English) used plausibility to disambiguate sentences, challenging the claims that L2 learners rely more heavily on plausibility than syntactic cues during sentence processing.

Keywords: verb bias; plausibility; garden-path sentences, L2 sentence processing

Introduction

Consider (1),

(1) *The scientist read the article…*
(a)……*at lunch time.*
(b)……*had been published two months ago.*

The syntactic role of *the article* is temporarily ambiguous between being the direct object of the preceding verb *read* or the subject of an upcoming embedded clause. In (1a), *the article* turns out to be the direct object of *read*, and the scientist did read the article. In (1b), however, *the article* turns out to be the subject of the embedded clause, and the scientist read something about the article, but not necessarily the article itself. Such temporary ambiguity at the ambiguous noun (*the article*) arises because English allows the complementizer *that* to be dropped. Sentences like (1b) can be disambiguated by adding a complementizer *that* after the main clause verb, as shown in (2).

(2) *The scientist read <u>that</u> the article had been published two months ago.*

Readers typically slow down at reading *had been* in (1b) than in (2), because they have initially analyzed *the article* as the direct object of *read*. At the position of *had been*, they realize that this analysis is incorrect, and thus start to revise that interpretation. Such slowing down in reading time is termed garden-path effect, and has been taken to reflect reanalysis processes.

Another cue native English speakers have been found to rely on to avoid garden-pathing is verb bias, which refers to the frequency with which a particular verb takes a particular structure, such as direct object (DO) or sentential complement (SC) (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell & Kim, 1998; Wilson & Garnsey, 2009). Consider (3) and (4),

(3) *The club members <u>understood</u> the bylaws would be applied to everyone.* (DO-bias verb)
(4) *The ticket agent admitted the mistake <u>might be</u> hard to correct.* (SC-bias verb)

Understand biases towards taking direct objects, and *admit* towards sentential complements. In (3), the parser anticipates a direct object after encountering *understood*, and thus experiences garden-path effect at *would*, where such interpretation turns out to be incorrect. In contrast, the parser experiences less difficulty processing *might* in (4),

because it anticipates an embedded clause after reading the sentential-clause-biased verb *admit*, and the sentence turns out to have the embedded clause structure.

Previous studies have shown that verb bias has a rapid effect on the processing of direct object/sentential complement (DO/SC) ambiguous sentences, and that either the complementizer *that* cue or the verb bias cue alone is sufficient for native English speakers to avoid garden-pathing. For instance, Garnsey et al. (1997) found that native speakers were slower at reading the disambiguating verb (e.g., *would*) after DO-bias verbs (e.g., *understand*), but not after SC-bias verbs (e.g., *admit*). After SC-bias verbs, the reading times at the disambiguating verb in sentences without the complementizer *that* (i.e., ambiguous) were just as fast as those with the complementizer *that* (i.e., unambiguous). This is the optimal efficient pattern of using the two cues. How about non-native speakers? Are they able to learn to use verb bias that is specific to the L2? Are they able to learn the complementizer *that* cue if such cue is not available in their native language?

A prevailing view in second language sentence processing literature claims that while L2 learners are capable of using lexical-semantic information during online parsing, they cannot use syntactic information in the way that native speakers do (i.e., the Shallow Structure Hypothesis; Clahsen & Felser, 2006). What has not been considered on this view is L2 learners' use of lexically-associated syntactic cue, such as verb bias. On the one hand, verb bias is lexically-associated information that is stored in the lexicon and retrieved when words are recognized. Such information might be considered to be part of the lexical information the Shallow Structure Hypothesis claims that L2 learners rely on. On the other hand, verb bias is about structure, so L2 learners may not use it to the extent that native speakers do. Studies thus far have revealed that L2 learners are able to learn L2-specific verb bias cue and use it fast enough to guide on-line parsing in the L2 (Dussias & Cramer Scaltz, 2008; Dussias, Marful, Gerfen, & Bajo Molina, 2010; Frenck-Mestre & Pynte, 1997), even if such information cannot be used in the same way in their L1 because L1 and L2 use different word orders (Lee, Lu, & Garnsey, 2013).

English follows SVO word order to place verbs early in the sentence, and therefore verbs provide useful information about the upcoming syntactic structure. In Korean, however, since the word order is SOV, verbs appear at the ends of clauses, and therefore are not useful in the same way as in English. In addition, unlike in English, where the complementizer *that* is optional, a clause-final complementizer particle *ko* is obligatory in Korean. Thus L1-Korean speakers do not have L1 experience with predicting upcoming structure based on either verb bias or the complementizer, but they do have experience with an end-of-clause complementizer that is a perfect cue to an embedded clause. Since the complementizer is a perfect cue on its own, it is possible that Korean speakers would never learn to associate structural biases with verbs, which is a much less reliable cue. Lee et al. (2013) compared L1-Korean L2-English speakers of higher proficiency with those with lower proficiency on their use of verb bias and the complementizer in reading ambiguous and unambiguous DO/SC ambiguous sentences that contained DO-bias or SC-bias verbs. Results showed that lower proficiency L2-learners must rely on the presence of the complementizer to use the verb bias cue, but higher proficiency group could use the two cues interactively, just like native speakers. However, higher proficiency learners did not achieve the optimal efficient pattern seen in the native speakers.

The fact that higher proficiency L1-Korean learners did not achieve the optimal native pattern might well be true for any L2 learners, simply because they don't have as much experience as native English speakers. Alternatively, however, it is possible that the fact that verb bias is not available early enough in the sentence to base predictions on in Korean is responsible for the failure to achieve the native pattern, making it important to test L2-English learners whose L1 has a word order placing verbs earlier in the sentence. Mandarin places verbs early in the sentence, with the same SVO order as English, and native speakers of Mandarin has been found to use verb bias to develop expectations about the upcoming structure (Qian, 2015). Mandarin has no complementizer in the type of DO/SC sentences used in this study. These differences between Korean and Mandarin suggest that L1-Mandarin learners of L2-English might use verb bias and complementizer cues differently from L1-Korean learners. L1-Mandarin learners may find it easier to learn and use the biases of English verbs earlier. As for their ability to use the complementizer *that*, it is not clear what to predict. Given that *that* is a frequent word, it may be easy for them to learn. Alternatively, the fact that the English cue *that* has many other usages besides a complementizer (pronoun, demonstrative, relative pronoun,…) may make it a difficult cue for Mandarin-L1 learners to rely on. Experiment 1 aims to test these predictions.

Previous research comparing the use of verb bias and plausibility cues in resolving DO/SC ambiguity by native English speakers found that plausibility did not have a chance to influence parsing in the presence of verbs with strong biases (Garnsey et al., 1997; Trueswell, 1996). It is possible that the same would not be true for L2 learners because L2 learners have been argued to rely heavily on lexical-semantic information rather than structure. Lee et al. (2013) has already shown that L1-Korean learners of L2-English did learn to make use of verb bias in such sentences. However, they did not also consider the role that plausibility might play. In Experiment 2, verb bias and plausibility were both manipulated and pitted against each other to examine the relative importance of plausibility and verb bias in L2-English sentence processing. Native English speakers rely more heavily on verb bias than plausibility, but the opposite might be true for L2-English learners.

Experiment 1

Participants

32 native English speakers (22 male, mean age 20) and 78 L1-Mandarin learners of L2-English (26 male, mean age 24) participated in Experiment 1. The native English control group was the same group as in Lee et al. (2013).

L2 proficiency was assessed using a cloze test (i.e., fill-in-the-blanks) that contained 40 blanks. The L2 group was divided into higher and lower proficiency groups based on median split (lower<32, higher≥32). Additional background information of the L2 group is summarized in Table 1.

Table 1. Language background information of the L1-Mandarin L2-English group in Experiment 1.

	All Learners	Lower Proficiency Group	Higher Proficiency Group
# of Participants	78	40	38
Age	24(18-37)	23(18-37)	24(18-35)
Proficiency score	31(21-37)	28(21-31)	34(32-37)
Age of classroom instruction	10(4-16)	10(5-16)	10(4-16)
Age start residence in English countries	21(15-33)	21(15-33)	22(15-30)
Duration of L2 country residence (months)	30(6-60)	28(6-60)	32(6-60)
Daily use of English (%)	50(5-95)	47(10-85)	4(5-95)

Materials and Design

Ten DO-bias and ten SC-bias verbs were each used four times to create 80 sets of sentences, with each set containing ambiguous and unambiguous versions of the same sentence (disambiguated by using *that*), as shown in (5). All disambiguating verbs and the words immediately following them were auxiliary verbs, so that the properties of the disambiguating words did not differ between items with DO-bias and SC-bias verbs. All critical items ended with embedded clauses.

(5) Example stimuli in Experiment 1:

DO-bias verb *The club members understood (that) the bylaws would be applied to everyone.*

SC-bias verb *The ticket agent admitted (that) the mistake might be hard to correct.*

DO- and SC-biased verbs were selected based on a norming study reported in Garnsey et al. (1997). DO-bias verbs were followed at least twice as often by direct object completions as by sentential complement completions in a sentence production task. The reverse was true for SC-bias verbs. DO- and SC-biased verbs used in the present study were matched on the number of letters, $F<1$, and frequency (Francis & Kucera, 1982) $F<1$. Verb properties are summarized in Table 2.

Table 2. Properties of the verbs used in the experiments.

	DO bias strength (%)	SC bias strength (%)	Mean length	Mean log frequency
DO-verb	76	13	8.1	1.9
SC-verbs	17	59	7.9	1.7

To ensure that any effect found at the disambiguating region was caused only by the biases of the verbs, two plausibility norming tasks were conducted to examine whether the ambiguous nouns were equally plausible as the direct object of the preceding verb and as the subject of the embedded clause between DO-bias and SC-bias items. The plausibility of the ambiguous noun as the direct object was rated by asking a separate group of 56 native speakers of English to judge the plausibility of the subject, verb and ambiguous noun combinations on a 1 (very implausible) to 7 (very plausible) scale, as shown in (6).

(6) *The club members understood the bylaws.*
The ticket agent admitted the mistake.

(7) *The club members understood that the bylaws were…*
The ticket agent admitted that the mistake was…

Ambiguous nouns following DO-bias verbs were rated as slightly more plausible than those following SC-bias verbs (6.5 vs 6.2, $F(1,78)=5.4$, $p<.05$). This replicated previous findings that plausibility ratings of the ambiguous noun as the direct object were affected by biases of the verbs (Garnsey et al., 1997). However, the small difference in plausibility is unlikely to have a detectable effect in sentences with strongly biased verbs, given previous findings from studies specifically manipulating plausibility (Garnsey et al., 1997).

A separate norming study that assessed the plausibility of the ambiguous noun as the subject of the embedded clause was conducted with 12 native English speakers, who rated on a 1 (very implausible) to 7 (very plausible) scale the plausibility of sentence fragments such as (7) as the beginning of a sentence. The ambiguous nouns for both DO- and SC-items were both rated as highly plausible (mean DO: 6.1; mean SC: 6.1) and did not differ between verb types, $F<1$. The properties of the ambiguous nouns in Experiment 1 are summarized in Table 3.

Table 3. Properties of the ambiguous nouns used in Experiment 1.

	Mean length	Log frequency	Plausibility as the direct object	Plausibility as the clause subject
DO-items	7.4	1.3	6.5	6.1
SC-items	7.1	1.4	6.2	6.1

Critical sentences were distributed over two lists according to a Latin Square design. 80 distractors were added to each list for a total of 160 trials per list. A comprehension question was asked following each sentence, and the question targeted the content of various parts of the sentences (e.g., *Were the bylaws applied fairly?*).

Procedure

Participants sat in a dimly lit and sound-attenuated booth in front of a 23-inch LCD monitor. 160 sentences were presented one word at a time in white 26-point Arial font on a black background in a non-cumulative moving window self-paced reading paradigm. Each time participant pressed a button on a Cendrus-830 response box, the next word appeared and the previous word reverted to the mask character. Following each sentence, a comprehension question was presented and participants press one of two buttons to indicate *yes* or *no* answers. Feedback about question accuracy was given. A "too slow" feedback message was presented if participants did not make a response within four seconds. The sentences were divided into four blocks with forty sentences each. The experiment began with a practice block of five sentences, and the entire session took 30-45 minutes to complete. L2 learners completed the cloze test after the self-paced reading experiment.

Results

Comprehension Accuracy On average, the accuracy rate for native English speakers was 92% (range 87%-96%) and for L1-Mandarin learners of L2-English was 86% (74%-96%). Higher proficiency learners answered comprehension questions to critical trials more accurately than lower proficiency learners (88% vs. 85%, $p<.01$).

Reading Times To remove individual differences in reading speed, statistical results reported below were based on length-corrected residual reading times computed separately for each participant. Reading times in the disambiguating region were averaged across two words (e.g., *would be*) and were then analyzed using linear mixed-effect models with maximal random effects structure (Barr, Levy, Scheepters, & Tily, 2013). $T>2$ is interpreted as significant. Raw reading times are summarized in Table 4.

Table 4. Raw reading times at the disambiguating region in Experiment 1.

	DO-bias verbs		SC-bias verbs	
	Ambig	Unambig	Ambig	Unambig
L1-English	362	343	341	342
L1-Mandarin	469	454	444	436

The disambiguating region. The multi-level model at this region that included ambiguity, verb bias, language group (native vs. non-native, high proficiency L2 group vs. low proficiency L2 group) and their interactions as fixed effects, and random intercepts and random slopes of the ambiguity factor for subjects and items revealed a main effect of ambiguity, with ambiguous sentences being read slower than unambiguous sentences (425 vs 414 ms; β=10, SE=3, $t>3$), an interaction between verb bias and ambiguity, with the ambiguity effect of DO-items (15 ms) being larger than that of SC-items (5 ms; β=11, SE=5, $t>2$), and an interaction between verb bias and language group (native vs. non-native; β=15, SE=8, $t>2$), with the difference between DO- and SC-items being bigger in the non-native speakers group (21 ms) than the native speakers group (10 ms).

Since there was no effect associated with the comparison between the higher proficiency and lower proficiency L1-Mandarin L2-English groups, further analyses did not break down into higher and lower proficiency groups. Analysis with the native speakers group showed a main effect of ambiguity (ambiguous 352 ms; unambiguous 342 ms; $t>2$) and an interaction between verb bias and ambiguity, which resulted because the disambiguating region was read slower in ambiguous than in unambiguous sentences only when the main clause verb had DO bias (ambiguity effect 19 ms; β=18.53, SE=5.61, t=3.30), but not when it had SC bias (ambiguity effect -1 ms; β=.43, SE=4.39, $t<1$), as shown in Figure 1.

Analyses on the L1-Mandarin L2-English group revealed a main effect of verb bias, with DO items read slower than SC items (461 vs 440 ms; β=23, SE=11, $t>2$) and a main effect of ambiguity (455 vs 445 ms; β=11, SE=4, $t>2$). Although the disambiguating region of ambiguous sentences were read slower than unambiguous sentences only after DO-bias verbs (t=2.74) but not after SC-bias verbs (t=1.55), this difference was not big enough to produce an interaction between verb bias and ambiguity. These results indicated that L1-Mandarin learners of L2-English were able to use verb bias and complementizer cues, but that neither cue alone was sufficient in the way that it is for native speakers.

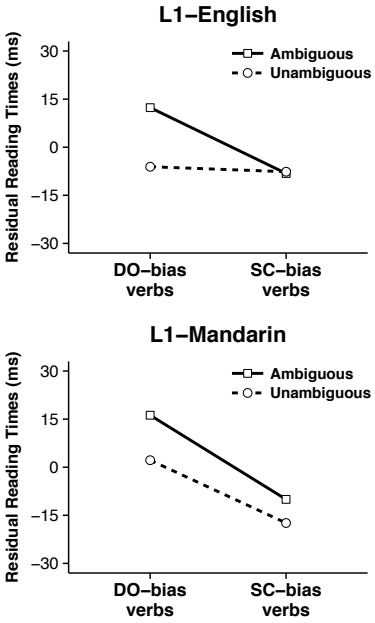


Figure 1: Reading time at the disambiguating region in Experiment 1.

Discussion

Experiment 1 manipulated verb bias and the presence of the complementizer *that* to compare the use of these two cues by L1-Mandarin speakers of L2-English and native English speakers. Native speakers showed the usual interaction between verb bias and ambiguity, suggesting an optimal, efficient and interactive use of the two cues. L1-Mandarin learners have learned to use verb bias, and their use of verb bias information was not modulated by proficiency, suggesting that even lower proficiency L1-Mandarin learners of L2-English have learned to use verb bias information. Perhaps, this is because they already do the same in their L1. Given that Mandarin does not have a complementizer that functions similarly to the complementizer *that* in English to signal an upcoming embedded clause, the prediction was that L1-Mandarin learners of L2-English might not be sensitive to such a cue. Contrary to that prediction, L1-Mandarin learners did use the complementizer cue. The use of complementizer was also not modulated by proficiency. Presumably, this is because the complementizer *that* cue is a salient cue and therefore is easy to learn.

Experiment 2

Experiment 1 showed that L1-Mandarin learners of L2-English were able to use both verb bias and complementizer cues to anticipate upcoming syntactic structure. The next question is how their usage of those cues compares with the usage of the kinds of lexical-semantic cues that have been proposed to be especially important for second language sentence processing. To evaluated that, Experiment 2 added plausibility manipulation, and tested native English, L1-Mandarin L2-English, and L1-Korean L2-English speakers.

Participants

65 native English speakers, 70 L1-Mandarin speakers of L2-English, and 69 L1-Korean speakers of L2-English participated in Experiment 2. L2 groups were divided into higher and lower proficiency groups based on median split in each group (L1-Korean: lower proficiency group<33, higher proficiency group≥33; L1-Mandarin: lower<35; higher≥35). Additional language background information are not provided here due to limit of space.

Materials and Design

The same 10 DO-bias and 10 SC-bias verbs were each used four times to construct 80 sets of sentences that fully crossed plausibility and ambiguity, as shown in (8). Plausible sentences were identical to sentences used in Experiment 1, for the most part.

As in Experiment 1, the ambiguous nouns were rated for their plausibility as the direct object of the main clause verb and as the subject of the embedded clause. Properties of the ambiguous nouns are summarized in Table 5. Plausible nouns in sentences with DO-bias verbs did not differ from those in SC-bias sentences in the number of letters (F<1) and log frequency (F<1). The same was true for implausible nouns (Fs<1). As in Experiment 1, 80 distractors were added to each list for a total of 160 trials per list. The procedure for Experiment 2 was exactly the same as Experiment 1.

(8) Example stimuli for Experiment 2:

DO-bias verb

Plausible: *The club members understood (that) the bylaws would be applied to everyone.*

Implausible: *The club members understood (that) the pool would be closed on Mondays.*

SC-bias verb

Plausible: *The ticket agent admitted (that) the mistake might be hard to correct.*

Implausible: *The ticket agent admitted (that) the kiosk might be difficult to find.*

Table 5. Properties of the ambiguous nouns used in Experiment 2.

	Plausibility as the direct object	Plausibility as the clause subject
	DO-items	
Plausible Noun	6.5	6.1
Implausible Noun	2.3	5.2
	SC-items	
Plausible Noun	6.2	6.1
Implausible Noun	1.9	5.3

Results

Comprehension Accuracy On average, the accuracy rate to comprehension questions was 93% for native speakers, 87% for L1-Mandarin group and 85% for L1-Korean group. Higher proficiency L1-Mandarin group was more accurate than lower proficiency L1-Mandarin group (89% vs 85%, p<.01), and higher proficiency L1-Korean group was more accurate than lower proficiency L1-Korean group (87% vs 83%, p<.01).

Reading Times At the disambiguating region, analysis on the residual reading times for native speakers revealed a main effect of ambiguity, with ambiguous sentences being read slower than unambiguous sentences (356 vs 349 ms; β=8, SE=3, t>2), and an interaction between verb bias and ambiguity (β=13, SE=5, t>2). The interaction resulted because ambiguous sentences were read slower than unambiguous sentences only after DO-bias verbs (364 vs 351 ms, β=20, SE=7, t>2) but not after SC-bias verbs (348 vs 347 ms; β=2, SE=4, t<1), as shown in Figure 2. There was no effect involving the plausibility factor (ts<1.6). This is consistent with a previous finding that plausibility of the noun as a direct object had no effect on disambiguation region reading times in native English speakers when verbs were strongly biased (Garnsey et al., 1997).

For L1-Mandarin group, there was a main effect of ambiguity, with ambiguous sentences being read slower than unambiguous sentences (447 vs 425 ms; β=21, SE=4,

t>5), a main effect of proficiency, with the higher proficiency group reading faster than the lower proficiency group (421 vs 451 ms; *β*=16, *SE*=5, *t*=3), and an interaction between verb bias and ambiguity (*β*=18, *SE*=8, *t*>2), which was caused by the ambiguous sentences being read slower than unambiguous sentences after DO-bias verbs (462 vs 432 ms; *β*=40, *SE*=9, *t*>4), but not after SC-bias verbs (433 vs 419 ms; *β*=13, *SE*=7, *t*<2), as shown in Figure 2. There was no effect involving the plausibility factor (*t*s<2).

For L1-Korean group, there was a main effect of ambiguity (482 vs 460 ms, *t*>5), and a main effect of verb bias (DO 479 ms vs SC 463 ms, *t*>2). There was no interaction between verb bias and plausibility, nor were there any effects involving the proficiency and plausibility factors (*t*s<2).

There was a numeric pattern in both the native speakers and the L2 learners' data that reading times were slower on the disambiguation following a DO-bias verb and an implausible noun. This tendency was not significant in any language groups, and this effect was hypothesized to be spill over from reading the implausible noun itself. Due to the length limit of this paper, data analyses at the ambiguous noun region were not included.

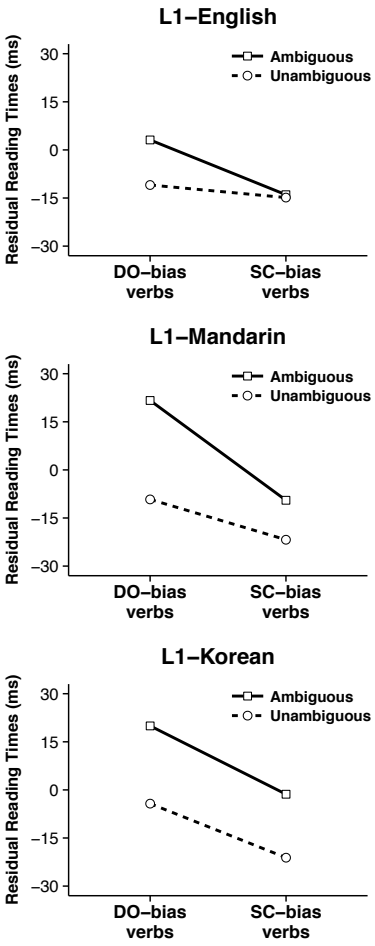


Figure 2: Residual reading times at the disambiguating region, collapsing over plausibility, in Experiment 2.

Conclusion

The present study investigated L2 learners' use of verb bias, complementizer *that*, and plausibility cues to predict upcoming syntactic structure. Results showed that L1-Mandarin speakers combined the verb bias and complementizer cues interactively, though they did not show the optimally efficient pattern seen in native speakers. In addition, even lower proficiency L1-Mandarin learners have learned to use verb bias and the complementizer, suggesting that verb bias and complementizer were not hard to learn for L1-Mandarin learners, perhaps because they use verb bias in the same way in their L1, and the complementizer *that* cue is a salient cue that is easy to learn. Moreover, just like native speakers, both L1-Mandarin and L1-Korean learners of L2-English did not use plausibility in their processing of DO/SC ambiguous sentences, contrasting the claims that L2 learners rely more heavily on semantic cues. Perhaps it may simply take too long, even for native speakers, to put the verb and noun meanings together in the way that is required for it to influence parsing decisions.

References

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*, 3-42.

Dussias, P.E., & Cramer Scaltz, T.R. (2008). Spanish-English L2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta psychologica, 128*, 501-513.

Dussias, P.E., Marful, A., Gerfen, C., & Bajo Molina, M.T. (2010). Usage frequencies of complement-taking verbs in Spanish and English: Data from Spanish monolinguals and Spanish-English bilinguals. *Behavior Research Methods, 42*, 1004-1011.

Francis, W.N., & Kucera, H. (1982). *Frequency and analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *Quarterly Journal of Experimental Psychology, 50*, 119-148.

Garnsey, S.M., Pearlmutter, N.J., Myers, E., & Lotocky, M.A. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language, 37*, 58-93.

Lee, E-K., Lu, H-Y., & Garnsey, S.M. (2013). L1 word order and sensitivity to verb bias in L2 processing. *Bilingualism: Language & Cognition, 16*, 761-775.

Qian, Z. (2015). *The reanalysis and interpretation of garden-path sentences by native speakers and second language learners.* (Dissertation), University of Illinois at Urbana-Champaign.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35*, 566-585.

Trueswell, J.C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language, 39*, 102-123.

Wilson, M.P., & Garnsey, S.M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language, 60*, 368-392.

Choice adaptation to increasing and decreasing event probabilities

Samuel Cheyette (sjcheyette@gmail.com)
Dynamic Decision Making Laboratory, Carnegie Mellon University
Pittsburgh, PA. 15213
Emmanouil Konstantinidis (em.konstantinidis@gmail.com)
School of Psychology, University of New South Wales
Sydney, NSW 2052, Australia
Jason L. Harman (jharman@lsu.edu)
Department of Psychology, Louisiana State University, Baton Rouge, LA. 70803
Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory, Carnegie Mellon University
Pittsburgh, PA. 15213

Abstract

A constant element of our modern environment is change. In decision-making research however, very little is known about how people make choices in dynamic environments. We report the results of an experiment where participants were asked to choose between two options: a dynamic and risky option that resulted in either a high or a low outcome, and a stationary and safe option that resulted in a medium outcome. The probability of the high outcome in the risky option decreased or increased linearly over the course of the task while the probability of the medium outcome stayed the same throughout. We find that adaptation to change is related to the direction of that change, and that the way people adapt to changing probabilities relates to their willingness to explore available options. A cognitive model based on Instance-Based Learning Theory reproduces the behavioral patterns.

Keywords: Change; Dynamic Decisions; Adaptation; Instance-Based Learning Theory; Decisions from Experience

Introduction

More than ever before the world around us seems to be changing rapidly. Technology has contributed to increasing availability of information and connectedness that contribute to a sense of rapid change. We make decisions in constantly changing situations and our ability to detect and adapt to those changes may determine the success of our choices. For example, a broker in the stock market must be sensitive to the changes in the interest rates in an attempt to maximize the long-term investments gains. In the context of reinforcement learning and restless bandit tasks, researchers have investigated change in similar settings, such as adaptation and detection of change, and exploration-exploitation tradeoffs in dynamic environments (e.g., Gureckis & Love, 2009). Yet, relatively little is known about how humans detect change when the change occurs gradually, and particularly when making decisions from experience while aiming at maximizing long-term gains.

Dynamic decision theory was first introduced by Ward Edwards (1962) who argued for the study of dynamic situations in which decision makers confront a sequence of decisions, and in which the environment changes while a decision maker is evaluating possible courses of action. Yet, to this date behavioral work on how humans behave under changing conditions and how we adapt to change is relatively limited.

About a decade ago, research started to shift from the overwhelmingly popular study of one-shot decisions to the study of repeated and consequential choice. In repeated choice conditions, early decisions produce payoffs and information that may influence future choices. This is one of the reasons that researchers have focused on experience and cognitive processes such as learning, memory, and recognition as key psychological processes of dynamic decision-making (Gonzalez, Lerch, & Lebiere, 2003).

The study of *decisions from experience* has expanded considerably in the past years, perhaps due to the development of simple experimental paradigms (e.g., "clicking paradigms") which have been used to study choice in its most essential form: in binary conditions (Baron & Erev, 2003; Hertwig et al., 2004). These paradigms involve the selection between two options, in the absence of descriptions of possible outcomes and probabilities. For example, in a repeated consequential choice paradigm (Baron & Erev, 2003), participants select between two buttons a fixed number of times (e.g., 100 times). After each selection, an outcome is displayed (i.e., feedback). This outcome is the realization of a probability distribution assigned to the button selected, which is unknown to participants. This paradigm is illustrated in Figure 1.

Using this experimental paradigm, researchers have investigated a number of issues relevant to how humans adapt to change and make choices in dynamic settings. For example, Rakow and Miler (2009) investigated repeated binary choice in which the associated probabilities of the outcomes could change over the sequence of trials. Specifically, the probability of one option would gradually change over a set of trials. The information given to participants was manipulated by providing the outcomes associated with each option or seeing a summary of the outcomes of previous trials. They observed a rapid adaptation (quick identification of the best option) when the probability changed, but a slow adaptation when only the outcome changed. The historic feedback helped but only in

early choices and not in later choices. Additionally, over all of their experiments, Rakow and Miler found some evidence that people react more quickly to negative changes than to positive changes. Their studies concluded with the importance of the adaptive nature of human memory and speculated how forgetting and recency of information can play an important role in adaptation.

Lejarraga, Dutt, and Gonzalez (2012) used Rakow and Miler's (2009) data to demonstrate how an Instance-Based Learning (IBL) cognitive model (Gonzalez et al., 2003) could account for that data. They compared the predictions from the IBL model to observed human choices suggesting that adaptation occurs through the reliance on recent outcomes. More recently, Lejarraga, Lejarraga, and Gonzalez (2014) investigated whether groups make better choices than individuals in dynamic tasks using similar problems with changing probabilities over time. They found that decisions made in groups were better than individual decisions in stable conditions, but groups were not superior to individuals after a sudden change had occurred in the probabilities. That is, groups had more difficulty in detecting and adapting to a sudden change compared to individuals. They also used an IBL model and a Bayesian updating model with "perfect memory" to explain why groups were slower at changing their policies compared to individuals.

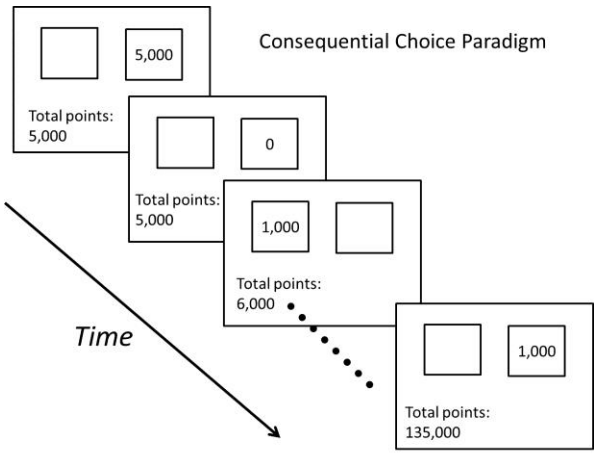


Figure 1: Repeated choice, consequential paradigm. The example shows that selecting the right button may result in an outcome of 1,000 with probability p=.2, 0 with p=.6 and 5,000 with p=.2, while the left button results in 1,000 with certainty (p=1). The probabilities are unknown to the participants.

We advance this line of research in four ways: First, we investigate how individuals adapt to *gradual* and *continual* change rather than to sudden changes. This is relevant to test the role of human memory in adjusting to gradual changes by the slightly altering past experiences. Second, we are also interested in looking at the *direction* of change. In research related to control of dynamic systems, researchers have found that adapting to change and being able to control a dynamic system in the long term, depends on whether the external environmental changes occur in a positive (i.e., increasing amounts) or negative (i.e., decreasing amounts) way (Gonzalez & Dutt, 2011). We present results from an experiment in a repeated choice paradigm in conditions of gradual positive and negative change. Third, we analyze exploratory behavior to evaluate how individuals explore the environment in order to detect gradual change. Fourth, we demonstrate the effects of human memory in these changing situations with an Instance-Based Learning model (Gonzalez et al., 2003), which relies on the ACT-R architecture's memory decay function (Anderson & Lebiere, 1998).

In addition to exploring gradual and continual change, the paradigm we implement involves high reward outcomes that change from very rare to near certain; analogous to a foraging animal in a changing environment once rich in resources that gradually deplete or vice versa (Mehlhorn et al., 2015). This design not only extends the results of Rakow and Miler (2009) to gradual and continual change but presents a test of boundary conditions in which exploration between options could be abandoned before change is detected due to very rare or very frequent early experiences. Accordingly, we expected better adaptation to change when the high risky outcome changes from very frequent to very rare.

Methods

Participants

Two-hundred and forty participants (88 Female, M_{age} =31.32) were randomly assigned to one of three conditions: increasing dynamic condition (N=76), decreasing dynamic condition (N=83), and stationary condition (N=81). Participants were recruited from Amazon Mechanical Turk for a "choice" experiment. They were paid $0.50 for participating and an additional bonus payment based on the points they accumulated over the course of the 100 trials, at a rate of 1 cent per 1000 points. The average bonus payment gained for the duration of the experiment of about 10 minutes was $0.26.

Design

The experiment asked participants to choose repeatedly between two options, with the goal of maximizing their long-term earnings that accumulate from each of the choices they make over 100 trials. The two options available in each choice include one risky option that could result in a high outcome (500 points) with probability p or a low outcome (0 points) with probability $1–p$, and a safe option that could result in a medium outcome (250 points) all the time. The presentation of the safe and risky options (left/right) was counter-balanced.

The main treatment involved the function of p which linearly increased, linearly decreased, or stayed stable as a function of time (choice trial number). In the increasing condition the probability of obtaining 500 points began at 0.01 and increased by 0.01 each trial, up to probability 1 at trial 100. This condition represents an environment where rewards change from extremely rare to certain as a function of time. In the decreasing condition the probability of obtaining 500 points started at 1 and decreased by 0.01 each trial, ending at 0.01 at trial 100. This condition represents an environment where rewards change from certain to extremely rare as a function of time. In the stationary

condition, the probability of obtaining 500 points stayed stable at 0.5 throughout.

Importantly, the cumulative expected value (EV) for all the options in all the conditions remains equal to 25,000 points over the course of the experiment. Thus, effectively selecting the risky option consistently over the course of 100 trials would result in approximately 25,000 accumulated points, the same overall accumulated outcome from selecting the safe option consistently over the course of 100 trials. However, the relative value of the two options changes over time. In the decreasing condition, the risky option is better than the safe option in the first 50 trials (i.e., it results in 500 points more often than 0 points) and then it becomes worse than the safe option in the last 50 trials. In the increasing condition, the risky option is worse than the safe option in the first 50 trials and it becomes better than the safe option in the last 50 trials. In the stationary condition the probability of getting 500 or 0 points in the risky option is always the same (0.5), so effectively the risky option is relatively as valuable as the safe option over the 100 trials.

Procedure

After providing consent and answering demographic questions, participants were given instructions for the game, and then they performed the choice task for 100 trials. Upon completion of the task, participants were given a final debriefing to determine whether they were aware of the changing probabilities and the direction of change.

Participants were given their total number of points accumulated, translated into a monetary bonus they earned, and then thanked for their participation.

Results

As a first step, we compared the proportion of risky choices (P-Risky) (see Figure 2, left panel). We analyzed the data using a generalized logit mixed-effects model with condition and block (blocks of 20 trials) as fixed effects, and subject-specific random intercepts. We found a significant difference in the proportion of risky choices participants made across conditions, $\chi^2(2) = 33.11$, $p < .001$. The P-Risky was higher in the decreasing condition (M = 0.48), followed by the stationary condition (M = 0.36), and the increasing condition (M = 0.25). There was also a significant effect of block, $\chi^2(4) = 229.09$, $p < .001$, and a significant interaction between condition and block, $\chi^2(8) = 1{,}289.77$, $p < .001$.

Looking at Figure 2 (left panel), the trends over time reveal a decrease in the P-Risky for the stationary condition, suggesting a general tendency to gradually select the safe option over time, even when the options were objectively equal. This is explained by risk aversion (Kahneman & Tversky, 1979), which has been investigated in decisions from experience paradigms through IBL models (Lebiere, Gonzalez, & Martin, 2007). The frequency and recency of occurrence of the low outcome in the risky option creates an imbalance of preference towards the safe option (Lebiere et al., 2007).

Second, the patterns suggest that although the P-Risky in the increasing condition was the lowest, the overall proportion of risky choices moved in the direction of the increased probability in the non-stationary option. However, this adaptation seems to be slow. Initially, participants quickly favored the safe option as we observe from the immediate drop of the P-Risky in the first 10 trials, but they moved slowly towards preferring the risky option as per the increase in the probability of the high outcome.

Third, the P-risky in the decreasing condition reduced rapidly over the course of 100 trials. Initially, choices quickly favored the risky option but they started to favor the safe option more quickly as the probability of the high option decreases, suggesting probability matching behavior (Erev & Barron, 2005). We observe that people were faster to cross the P-Risky = 0.50 threshold in the decreasing condition compared to the increasing condition. In the increasing condition, P-Risky did not reach 0.50 until the 98th trial, whereas in the decreasing condition, P-Risky reached the 0.50 mark on the 50th trial, essentially tracking the probability function throughout (i.e., a demonstration of probability matching behavior; see also Rakow & Miller, 2009).

Participants seem to select the risky option that matches the probability of the high outcome. That is, participants seem to select the *maximizing* option more accurately in the decreasing rather than the increasing condition. To test this, we calculated the proportion of maximization choices (P-Max) before and after the objective change of the relative goodness of the options (e.g. trial 50; see Figure 3, left panel). We found a significant difference in the P-Max across increasing and decreasing conditions, $\chi^2(1) = 15.81$, $p < .001$. The P-Max was higher in the decreasing condition (M = 0.66) than the increasing condition (M = 0.57). In addition, participants maximized more in the first period of the task, $\chi^2(1) = 537.50$, $p < .001$, and the interaction was also significant, $\chi^2(1) = 1{,}916.19$, $p < .001$. The P-Max for the increasing condition was significantly higher in the first half (M = 0.83) than the second half (M = 0.33; $\chi^2(1) = 1{,}717.80$, $p < .001$), whereas the order is reversed in the decreasing condition but to a lesser degree ($M_{first\ half}$ = 0.61, $M_{second\ half}$ = 0.64; $\chi^2(1) = 12.34$, $p < .001$). The contrast is quite stark: the maximizing rate never drops below 0.50 in the decreasing condition, but it is on average 0.30 in the second half in the increasing condition. This is consistent with the observation that people are adapting significantly more rapidly in the decreasing condition than in the increasing condition.

A possible explanation for the different degrees of adaptation between the increasing and decreasing conditions is the lack of exploration of the options. As we observe in the stationary condition, participants' choices drift towards the safe option over time, even when there is no change in probabilities and values. In the increasing condition, people might also have this tendency given that the safe option appears to provide higher payoffs more often than the risky option in the first few trials. This might prevent participants form exploring the risky option in later trials. In contrast, in the decreasing condition, since the risky option provides higher payoffs than the safe option in early trials, it is possible that people are more aware of the changes in the probability given that they are already selecting the risky option more often.

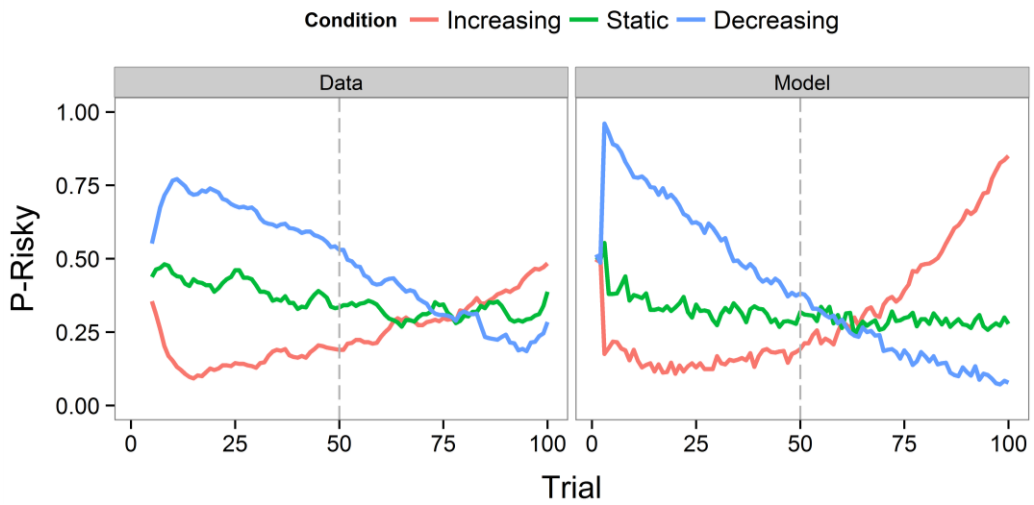


Figure 2: The left panel shows human responses and the right panel shows model predictions for each of the three conditions, marked in different colors, where the measure is the proportion of choosing the risky option across trials.

To test for exploration, we used a measure proposed in past research: the alternation rate (A-rate; Gonzalez & Dutt, 2011; 2012). This is the proportion of switches from one option to another in consecutive trials. We performed a similar analysis (mixed-effects logit model) and we found a significant effect of condition, $\chi^2(1) = 16.79$, $p < .001$, as participants switched more between options in the decreasing ($M = 0.30$) than the increasing condition ($M = 0.21$). The effect of period (before or after trial 50) was not significant, $\chi^2(1) = 0.43$, $p = .51$, but the interaction between condition and period was significant ($\chi^2(1) = 10.47$, $p = .001$): while there was a difference in A-rate between periods in the increasing condition ($p = .003$), this was not the case in the decreasing condition ($p = .12$; Figure 4, left panel).

Instance Based Learning Model

An IBL model designed to account for over-time effects of binary choice (Gonzalez & Dutt, 2011) is a generalist (it applies to a wide variety of tasks) instead of a specialist (a model that is made for one particular task; Lejarraga, et al., 2012) and it builds on the ACT–R cognitive architecture (Anderson & Lebiere, 1998). It proposes that a choice is a function of the accumulated value (*blended value*) for each of the two options, through experience. This value is a function of the outcomes observed and the associated probability of retrieving the corresponding instances from memory. Memory retrieval depends on the activation of a value that reflects how readily available this information is in memory. In this IBL model, activation reflects the *frequency* (how many times an outcome has been observed in the past), *recency*, and *noise* of the experience. The formulation of this model appears in multiple past publications (e.g., Gonzalez & Dutt, 2011; Lejarraga, et al., 2012), but for completeness we reproduce it here. A choice between the two options is made by using the blended value V which represents the value of option j in a particular trial t:

$$V_{jt} = \sum_{i=1}^{n} P_{it} x_i$$

where x_i refers to the payoff obtained in each option stored in memory as instance i for the option j, and p_i is the probability of retrieving that instance from memory, which is relative to the activations of other instances in option j:

$$P_{jit} = \frac{e^{A_{it}/\tau}}{\sum_j e^{A_{jt}/\tau}}$$

where τ is random noise defined as $\tau = \sigma\sqrt{2}$, and σ is a free noise parameter. The activation of instance i represents how readily available the information is in memory:

$$A_{jit} = \sigma \ln\left(\frac{1-\gamma_{jit}}{\gamma_{jit}}\right) + \ln \sum_{t_p \in \{1,\ldots,t-1\}} (t - t_p)^{-d}$$

The activation is higher when instances are observed frequently and more recently. When an instance is not observed often, the memory will decay with the passage of time (the parameter d, the decay, is a non-negative free parameter that defines the rate of forgetting). The noise component σ is a free parameter that reflects noisy memory retrieval, γ is a random sample from a uniform distribution (between 0 and 1), and t_p denotes all the previous trials that outcome i was observed.

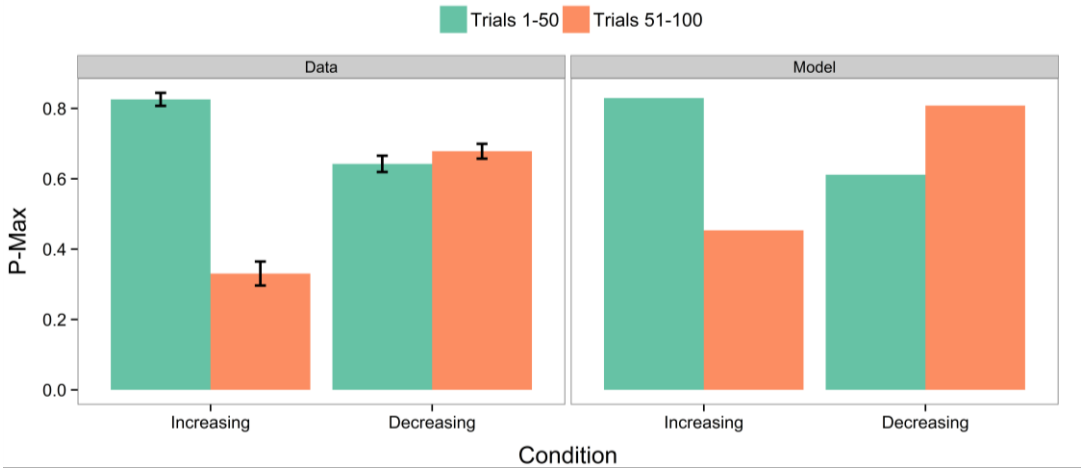


Figure 3: The left panel shows human data and the right panel shows model predictions of average maximization rates (P-Max) in the changing probability conditions (increasing and decreasing) before and after the halfway point (50th trial).

We produced predictions via simulations using this IBL model with parameters from past research (d = 5 and σ = 1.5; see Lejarraga, et al., 2012). We simulated the choices over time of 500 participants in each of the experimental conditions. The right panels in Figures 2, 3, and 4, show the model predictions that correspond to each of the results observed from the human data.

The model made predictions of choice behavior across time that reflected similar trends in human data (right panel, Figure 2). The model differed in how often it chose the risky option on average across the three conditions. In the decreasing condition, the mean P-Risky over the 100 trials was 0.40; in the increasing condition the mean P-Risky was 0.32; in the stationary condition it was 0.33. This was largely due to worse adaptation in the increasing condition, in which P-Risky does not reach 0.50 until the 77th trial (so 27 trials after it would have been beneficial to do so). On the other hand, in the decreasing condition, the model begins choosing the safer alternative in advance and P-Risky crosses the 0.50 mark on the 34th trial. In the stationary condition, P-Risky drops to around 0.30 and remains around that level. In agreement with the observation in human participants, although the two options have the same EV, the model chooses the safe option about two times more on average than the risky option.

The P-Max between conditions in the first and second half of the experiment is shown in Figure 3 (right panel). The average P-max in the first half (M = 0.83) is higher than in the second half (M = 0.48). However, in the decreasing condition we observed a trend in the opposite direction. In the decreasing condition we find that the average P-Max in the first half (M = 0.60) is lower than the second half (M = 0.81). Regarding A-rate (Figure 4), the model accurately predicts more switching in the decreasing condition (M = 0.28) compared to the increasing condition (M = 0.23).

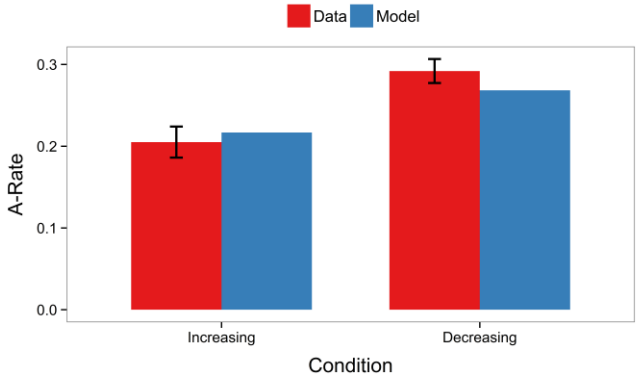


Figure 4: Average observed (Data) and predicted (Model) alternation rates (A-rate) in the changing probability conditions (increasing and decreasing).

In contrast to human participants, the predictions of the model are more extreme: adapting better in the decreasing than in the increasing condition and doing better in terms of maximizing choices in the second half than in the first. However, these are good predictions given that this is an "out of the box" model prediction, where the simulated data were produced in the complete ignorance of human data. In fact, when we calculated the mean squared difference (MSD) for each condition, we found that the predictions in the stationary condition were the closest to observed data, with an MSD of 0.006; the decreasing condition was the next closest with an MSD of 0.018, and the increasing condition was the next one with an MSD of 0.020.

Discussion

The main purpose of the current investigation was to examine how people adapt their choices to gradual and continual change of event probabilities. Specifically, we were interested in whether the direction of change

(increasing or decreasing probabilities of maximum outcomes) would have an effect on people's choice behavior. According to economic theory one should expect no difference in choice across conditions, since the cumulative EV of all options was *the same* in all experimental conditions. In contrast, as expected by cognitive theories, results show that people are sensitive to the dynamics of experienced outcomes and to the direction of change of the associated probabilities to these outcomes.

Three main phenomena emerged from this investigation: 1) risk aversion in experience, 2) slow adaptation to increasing probabilities, and 3) fast adaptation to decreasing probabilities. These patterns of risky choice are reinforced by two observed behaviors: the maximization and the alternation behavior. In the increasing condition, people chose the best option in the first half of their experience but they fell far below the average optimal behavior in the second half; while in the decreasing condition participants stayed above average optimal behavior throughout. The results suggest that participants explore the two options more in the decreasing than the increasing condition.

We observe that the IBL model provides close predictions to the observed behavior. Even though the performed simulations were not exhaustively in line with what was observed in the task, the model provided insightful observations into the mechanisms of adaptation to change. Introspecting into the IBL model's mechanisms we observe, and it is also discussed in more detail in Lebiere et al. (2007), that the model naturally develops a preference for the safe option rather than the risky option. This is due to the experiences of extreme outcomes in the risky option and the blending choice mechanism of the model that "blends together" these outcomes, giving rise to a slight preference for the safe option (i.e., the stationary option). These predictions emerge from the activation of instances that reflects the frequency and recency of the occurrence of outcomes. The stationary outcome develops initially a higher probability of retrieval and a slightly higher blended value. In the increasing condition, this tendency prevents the model from exploring the risky option, resulting in "lack of awareness" of the change. In contrast, in the decreasing condition, because the low outcome of the risky option has an extreme low probability of occurrence early on, the model develops a preference for the risky over the safe option. This results in "awareness" of the change in the probability which helps the model being more successful at adapting to the probability decrease.

In conclusion, the area of dynamic experience-based decision-making has remained largely unexplored and this study attempted to provide a deeper understanding of the factors that are involved in the adaptation to continuous dynamic change. We found that people were slower at adapting to changes in the outcome probability when a high outcome changes from rare to frequent compared to a high outcome changing from frequent to rare. People are slow at switching to a risky choice in the increasing condition and fast at switching to a safe option in the decreasing condition. Differences in exploration of the available options, joined with the dynamics of experience and the cognitive effects involved (frequency and recency of experiences) provide an explanation of this behavior.

Acknowledgements

This work was supported by the National Science Foundation Award Number: 1154012 to Cleotilde Gonzalez.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making, 16*, 215-233.

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review, 69*, 109-135.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112*, 912-931.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*, 591-635.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. P*sychological Review, 118*, 523-551.

Gonzalez, C., & Dutt, V. (2012). Refuting data aggregation arguments and how the IBL model stands criticism: A reply to Hills and Hertwig (2012). *Psychological Review, 119*, 893-898.

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition, 113*, 293-313.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-539.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, *47*, 263-291.

Lebiere, C., Gonzalez, C., & Martin, M. (2007). *Instance-based decision making model of repeated binary choice.* Paper presented at the 8th International Conference on Cognitive Modeling, Oxford, UK.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25*, 143-153.

Lejarraga, T., Lejarraga, J., & Gonzalez, C. (2014). Decisions from experience: How groups and individuals adapt to change. *Memory & Cognition, 42*, 1384-1397.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2*, 191–215.

Rakow, T., & Miler, K. (2009). Doomed to repeat the successes of the past: History is best forgotten for repeated choices with nonstationary payoffs. *Memory & Cognition, 37*, 985-1000.

A Robust Implementation of Episodic Memory for a Cognitive Architecture

David Henri Ménager (dhmenager@ku.edu)
Department of Electrical Engineering and Computer Science
1520 W. 15th Street, Lawrence, KS 66045 USA

Dongkyu Choi (dongkyuc@ku.edu)
Department of Aerospace Engineering
1530 W. 15th Street, Lawrence, KS 66045 USA

Abstract

The ability to remember events plays an important role in human life. People can replay past events in their heads and make decisions based on that information. In this paper, we describe a novel extension to a cognitive architecture, ICARUS, that enables it to store, organize, generalize, and retrieve episodic traces that can help the agent in a variety of manners. After discussing previous work on the related topic, we review ICARUS and explain the new extension to the architecture in detail. Then we discuss four architectural implications of the new capability and list some future work before we conclude.

Keywords: episodic memory; cognitive architectures; virtual sensing; expectations; impasse resolution

Introduction

Episodic memory is one of the cornerstones of human cognitive ability (Tulving, 2002). It is responsible for enabling one to remember the events of his or her life. This remembering, however, is not simply a recollection of personal facts. Rather, it is a *relived experience* made possible by three necessary tenents: a subjective sense of time; a sense of self; and autonoetic consciousness (Tulving, 1985, 2002). These allow humans to go back in time in one's head using the subjective sense of time without confusing the events as happening right now. This ability impacts many different aspects of human cognitive capabilities.

Despite the fundamental importance of the episodic memory, however, computational models of episodic memory in the context of cognitive architectures are not discussed very frequently aside from some recent work (Nuxoll & Laird, 2007; Faltersack, Burns, Nuxoll, & Crenshaw, 2011; Bölöni, 2011). These systems have demonstrated how episodic memory aids in problem solving, reinforcement learning, narration, and so on. In our work, we aim to build a system that provides these and other capabilities in a single implementation. We built a psychologically plausible episodic memory module and integrated it within a cognitive architecture, ICARUS (Langley & Choi, 2006). The initial implementation quickly resulted in three new or improved capabilities in our system that we believe are important.

In the sections below, we first describe the background in the literature that serves as basis for our work. After a brief review of ICARUS that follows, we provide a detailed description of how an episodic memory has been implemented in the architecture. We also discuss some architectural implications of the new extension. Then, we conclude after a discussion on future work.

Background

Researchers of memory have held that any biological or computational model of episodic memory must support encoding and retrieval of experience (Tulving, 1983). Encoding is the process of recording and organizing experiences into the episodic memory, and retrieval is the process of using a retrieval cue to find episodes in this memory. These two functionalities have been the subject of much discussion amongst the experts of psychology because it is rather difficult to characterize the processes that govern the interactions between episodic and semantic memories.

Researchers realized that the nuances of memory between implicit and explicit memory and the nature of knowing and remembering are related to the relationship between semantic and episodic memories (Schachter, 1987; Tulving, 1985). We believe that a psychologically plausible model of episodic memory should have an account for these nuances.

As for the details of how the episodic memory works, there is psychological evidence suggesting that it is an index-based long-term memory that supports cue-based retrieval (Hellerstedt, 2015; Tulving, 1983). Researchers also studied the three technicalities of episodic memory to define the notion of index-based memory, episodes, and cues, as summarized below.

Index-based Memory Schiller et al. (2015) argue that episodic memory receives many of its characteristics from the hippocampus, one of which is the ability to create cognitive maps, or a hierarchical network of memories. As memories come and go, the hippocampus is believed to dynamically change the structure of this network in order to preserve the similarity relationship between connected memories. With this structural representation, the index of an episode may be seen as the path from one of the top-level episodes in the network to the episodes of interest.

Episodes Previous research suggests that there exists an episodic buffer that interfaces between the episodic memory and the central executive of the working memory (Baddeley, 2000). This buffer is responsible for accepting diverse sets of data from the agent's sensors and creating a common representation of the data. This representation is what is used to create episodes. Hellerstedt (2015) further states that the creation of these episodes occurs in an on-line fashion.

Cues To retrieve episodic memories, semantic patterns, or cues, are used. These patterns can match against elements in episodic memory even if there is some missing information. Assuming that episodes are organized by similarity, the retrieval process is about finding the episode that contains the most similar situation to the cue. While retrieval is in progress, semantic memory is responsible for generating, specifying, and storing cues (Hellerstedt, 2015). A cue may return more than one episode, in which case the system must use some conflict resolution strategies to determine which episode to present.

Based on our understanding of episodic memory as described above, we implemented our extension to ICARUS. Before we describe the details of our implementation, however, it will be useful to review the architecture briefly to facilitate our discussions.

ICARUS Review

As a cognitive architecture, ICARUS provides a framework for modeling human cognition and programming intelligent agents. The architecture makes commitments to its representation of knowledge and structures, the memories that store these contents, and the processes that work over them. ICARUS shares some of these commitments with other architectures like Soar (Laird, 2012) and ACT-R (Anderson & Lebiere, 1998), but it also has distinct characteristics like the architectural commitment to hierarchical knowledge structures, teleoreactive execution, and goal reasoning capabilities (Choi, 2011). In this section, we provide a brief review of the architecture to facilitate our discussion on the new episodic memory module afterwards.

Representation and Memories

ICARUS distinguishes two main types of knowledge. One is its *concepts* that describe certain aspects of the situation in the environment. They resemble Horn clauses (Horn, 1951), complete with a predicate as the head, perceptual matching conditions, tests against matched variables, and references to any sub-relations. For example, the first two in Table 1 are concept definitions. The first one, `(on ?o1 ?o2)`, describes a *primitive* situation where a block is on top of another block, using only perceptual matching conditions for two blocks and tests against the matched objects and their attributes. The second one, `(clear ?block)`, depicts a *complex* situation where there is nothing on top of a block, using another concept as a sub-relation in addition to perceptual matching conditions for a block.

The other type of knowledge in ICARUS is its *skills* that describe procedures to achieve certain concept instances in the environment. These are essentially hierarchical versions of STRIPS operators (Fikes & Nilsson, 1971) with a named head, perceptual matching conditions, preconditions that need to be true to execute, direct actions to perform in the world or any sub-skills, and the intended effects of the execution. For instance, the last entry in Table 1 shows a *primitive* skill definition, `(look-right ?robot)`, that describes a procedure to get the robot to look right. The skill has a named head, perceptual matching against a robot and its attributes, the precondition of the robot not already looking right, the action to perform in the world, and the intended effect. *Complex* skills have a similar syntax, except that they include subskills instead of direct actions in their body.

Table 1: Two sample ICARUS concepts and a skill for the modified blocks world.

```
((on ?o1 ?o2)
 :elements (?o1 is (block ?o1 ^x ?x1 ^y ?y1 ^len ?len1)
            ?o2 is (block ?o2 ^x ?x2 ^y ?y2 ^len ?len2
                           ^height ?height2))
 :tests ((*overlapping ?x1 ?len1 ?x2 ?len2)
         (= ?y1 (+ ?y2 ?height2))))

((clear ?block)
 :elements (?block is (block ?block))
 :conditions ((not (on ?another ?block))))

((look-right ?robot)
 :elements (?robot is (robot ?robot ^looking ?looking
                                 ^holding ?holding))
 :conditions ((not (eq ?looking 'right)))
 :actions ((*look-right ?robot))
 :effects (?robot is (robot ?robot ^looking right
                                 ^holding ?holding)))
```

To store these knowledge and other structures, ICARUS employs a handful of distinct memories. The concept and skill definitions are stored in conceptual and procedural long-term memories, respectively. The short-term instances of these definitions are stored in a belief memory and a goal / intention memory. The former maintains the current state of the world, while the latter houses the agent's current goals and intentions for execution.

Inference and Execution

The ICARUS architecture operates in cycles (see Figure 1). At the beginning of each cycle, the system receives sensory input from the environment as a list of objects with their attribute–value pairs. Based on this information, the architecture infers all the concept instances that are true in the current state by matching its concept definitions to perceived objects and other concept instances.

Once all the beliefs are inferred, the system finds all the relevant skill definitions for the current goal(s) that are executable in the current belief state. ICARUS then chooses one or more of them and execute them in the world. The architecture will continue its cycles in this manner until all of its goals are achieved or its operations are terminated for any other reasons. With this brief review, we now continue our main discussion on the new episodic module in ICARUS.

Episodic Module in ICARUS

In the context of episodic memory, ICARUS shares some architectural assumptions with Soar (Laird, 2012). More specifically, both the architectures assume that episodic memory is a long-term, cue-based system that maintains cues in the agent's working memory, and the agent deliberately en-

codes and retrieves episodes. However, there are also significant differences. Each time a Soar agent takes action(s), the architecture records a snapshot of its current state as an episode, which is stored statically in the episodic memory without any generalization (Nuxoll & Laird, 2007). An individual episode in Soar, therefore, does not capture the notion of a relived experience. Rather, it is necessary to compile multiple episodes to describe an event. Furthermore, the Soar agent does not remember events where it did not perform any action and stayed as an observer. In contrast, ICARUS has these unique characteristics in its episodic memory:

- Episodic memory is organized as a compound structure composed of an episodic cache, an episodic generalization tree, and a concept frequency tree.
- Episodic generalization tree is organized by similarity.
- Episodes represent durative experiences of variable length.
- Episodic memory is a dynamic structure that supports generalization among similar episodes.

In this section, we describe the new episodic memory module in ICARUS in detail. We start with the representation of episodes, and then explain the encoding, retrieval, and generalization processes.

Episodic Representation

The episodic generalization tree is the main data structure that organizes and stores episodes. The episodic cache acts as a storage for the agent's unprocessed experience. Since our agents do not run for days or years just yet, we assume that the agent has sufficient memory to store the complete state–action sequence. The concept frequency tree records the number of times ICARUS has seen each concept instance. The episodic cache and the concept frequency tree provide a mechanism for recognizing and explaining significant events.

Episodic Processes

On every cycle, ICARUS records the current state and executed actions into the episodic cache and updates the concept frequency tree, as indicated by the arrows going into the episodic memory in Figure 1. When the agent perceives or infers one or more *significant* events, it begins to encode a new episode. The architecture tries to explain each significant event by analyzing information stored in the episodic cache and finding a logical process that causes the event to happen. If the explanation attempt is successful, the state–action sequence that explains the significant event will be stored as the new episode. Otherwise, the significant event(s) and the time when the event(s) occurred are stored as the new episode. This is to make it possible for the agent to return to the particular episode and try to explain again after it accumulates more knowledge about the world.

The episodic memory supports episodic generalization, and the resulting hierarchy is stored in the episodic generalization tree. The root (top-level) node of this tree is the most general episode and is allowed to have an arbitrary number of children. Each child is a k-ary tree where $k \in \mathbf{N}$. Episodes become more specific at each decreasing level of the tree toward the leaf nodes, and there are fully instantiated episodes at the bottom of the tree. This structural organization reflects our understanding of Schiller et al. (2015), and the notion of episodes representing durative events is consistent with all the literature on episodic memory we are aware of. Next, we explain the processes for encoding, retrieving, and generalizing episodes in ICARUS.

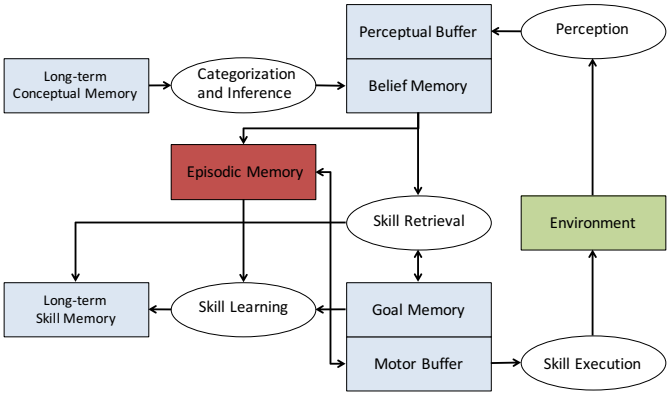


Figure 1: A block diagram that illustrates ICARUS's processes including the new episodic module marked in red. Arrows represent the direction of information flow.

Encoding Encoding is the process that consumes a raw state–action sequence stored in the episodic cache and inserts a fully specified episode into the episodic generalization tree. This process is triggered by the noticing of one or more significant events. Currently, ICARUS considers the following four cases as significant:

1. A rarely seen concept predicate
2. A rarely seen partial set of bindings for a concept predicate
3. A rarely seen full set of bindings for a concept predicate
4. The absence of a concept instance that the agent expects

Once a new episode has been created in this manner, the architecture inserts it to the episodic generalization tree using a slightly modified level order search. In level order search, a node's children are added to the search queue after having visited that node. Rather than always doing this, a similarity test is done first to ensure that the episode already in the tree unifies to the episode to insert if and only if the two episodes are structurally similar. Figure 2 shows this procedure graphically. The numbers in curly braces represent the order in which the insertion happens. After the similarity test is passed, then and only then are the similar episode's children added to the search queue. If the episode to insert cannot be unified with any of the episodes on a given level, then that episode becomes a sibling of those episodes. A new episode has successfully been encoded into the episodic memory. If

the episode to insert is a copy of one of the leaf episodes, then the counter for the number of times that particular episode has been observed goes up by one and the episode to insert is not inserted. All episodes are guaranteed to be instantiated by the general episode at the root node.

Retrieval ICARUS supports cue-based retrieval of the episodic memory. Given a cue, ICARUS performs the level order search described in the encoding process except that, instead of the similarity test, the system checks to see if the episode contains a matching fully specified copy of the cue or if the episode contains a generalized version of the cue that unifies with the cue. If so, the episode that matches with the cue is returned. Figure 3 shows the retrieval process. More than one episode can match the cue, so retrieving from a cue returns a forest of trees. One major advantage of this is mechanism, is that the agent only has to search the episodes directly under the root to discover if a situation is completely new. The reason is that, if the cue matches, the cue will either match directly, or at least one of the root's children will contain some generalized version of the cue and the generalization match test will pass. If all match tests fail then there does not exist a matching episode in the tree. This is known, in the worst case, at level two of the tree.

Generalization The ability to generalize knowledge is a key cognitive ability. One of the ways ICARUS supports this is by performing generalization in the episodic tree at encoding. The state–action trace in the episode represents a demonstration of how a significant event came to be. Therefore it is possible to learn from that trace. For example, if person x drops a glass on the ground and it breaks, and person y drops a glass on the ground and it breaks as well, ICARUS will generalize that knowledge to say that if anyone drops a glass on the ground, it will break (assuming $x \neq y$). This may not be true in general, but the number of times ICARUS makes such generalizations forms the conditional probability, or the confidence with which ICARUS believes a glass breaks when someone drops it. The ability to gain knowledge in this way seems quite central to general intelligence. Sibling generalization is implemented by iterating through the newly inserted episode's siblings. ICARUS makes a generalized episode, and for each sibling it checks to see if it can unify the sibling and the new entry by variablizing the bindings where conflicts are found. If a consistent variablization has been made, ICARUS tests to see if the generalized episode is still more specific than the parent of the newly inserted episode. If so, then the the generalized episode's parent becomes the freshly inserted episode's parent and the generalized episode's sub-episodes becomes the freshly inserted episode and the sibling episode that generalized with it. The count for the generalized episode is the summation of the count of its children.

Architectural Implications

The modified blocks world we use in this work is a partially observable world with two tables. When the agent is looking

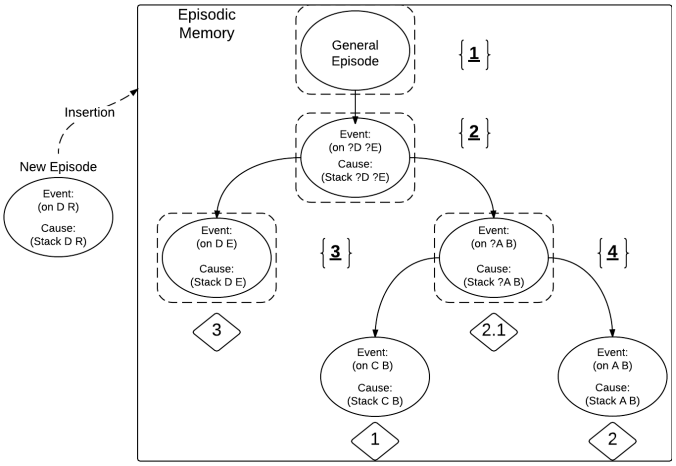


Figure 2: The insertion into the episodic generalization tree. The numbers in the diamonds represent the order in which the existing episodes were experienced.

at one table it cannot see the contents of the other table. Both tables contain three boxes, a red, a blue, and a green, and one box on each table has a block inside of it. Though the contents of the tables are the same, the tables themselves are labelled differently and the positioning and names of the boxes are different. Specifically, one table is labeled *Rainbow* and the other is named *Cloud*, and the boxes are named `Box1` through `Box6`. This is done so that the agent knows that there are two distinct environments in this domain.

We aim to demonstrate many interesting applications from our work, but for the moment we use the modified blocks world domain to supply examples for learning from observation, impasse resolution and making expectations.

Learning from Observations

Despite episodic memory being the memory for events, constantly reflecting on passed experiences may cause the agent to disproportionally slow down as a result of the amount of resources required to search the episodic memory for an episode that answers or responds to a query. When an agent enters a new environment it may not know how to characterize it, and as a result may be unsure of what to do and relies on its episodic memory. As it collects more experiences it may come to knowledge of specific patterns about the environment and formally characterize them in terms of rules.

Since episodes in the episodic memory have a count associated with them, we can capture this ability in ICARUS. If the agent experiences an episode a sufficient number of times ICARUS will try to formalize it into a rule. In our domain, if the agent experiences 10 times that an extra box appears when it stacks two boxes on top of each other, it would have a sufficient amount of evidence to logically relate a stacking two boxes with producing a new box. This mechanism also ties into remembering and knowing because even if ICARUS forgets the stacking experience it would still maintain the rule. Of course there is the possibility that the agent overfits rules

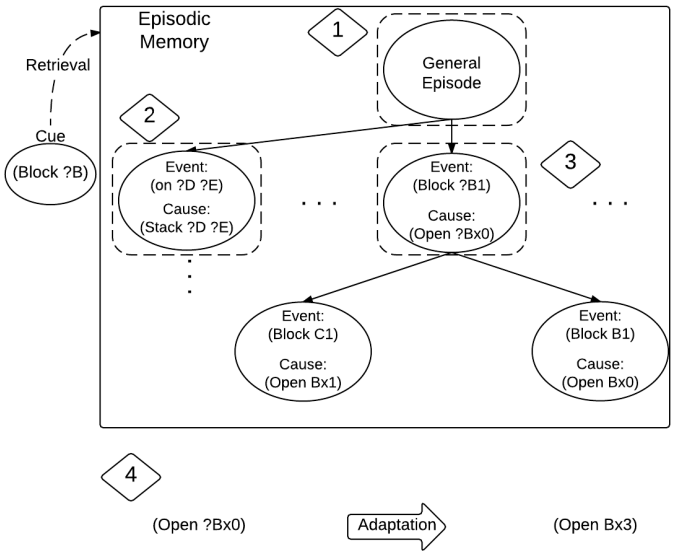


Figure 3: Retrieval and impasse resolution in ICARUS.

so a mechanism for modifying rules needs to exist. Also, since ICARUS keeps a complete state–action sequence it can always review old experiences to learn something new.

Impasse Resolution

ICARUS encounters an impasse when it does not know how to reach its goal in the current state. One way to overcome an impasse is to problem solve through means-ends analysis. Note, however, that a plan generated by this method is dependent on the agent's conceptual knowledge of how the world works. So, if the agent does not know how to logically relate the goal state to skills or concepts the agent fails. In such cases it would be advantageous to recall a similar experience to the current one and repeat the actions that produced the goal in the previous experience. Figure 3 shows the order in which ICARUS matches against the cue. Note that it is possible to match against generalized episodes. In stage 4, the environment's current bindings are applied to the retrieved solution. Another interesting aspect of impasse resolution is that it is one of the means for which ICARUS gathers support to formalize experiences into a rules as discussed above.

ICARUS spends some time at the Rainbow table exploring the world. During this time it tries a number of different actions and eventually it opens all the boxes and realizes that one of them has a block inside of it. The agent goes to the Cloud table and sees a new set of boxes on the table. The task is to find the block inside the box. Even though ICARUS knows that opening a box might reveal an item contained inside it, there is no logical reason why the box has to contain a block so the problem solver cannot reason properly about what to do in this case. When ICARUS uses `(block ?b1)` as a cue to the episodic memory, it remembers that it found a block at Rainbow by, say, opening `Box0`. This solution is adapted to the current situation and the agent decides to open all the boxes because `Box0` is of type `box`.

Expectations

Our system uses the concept frequency tree to create expectations of what beliefs should be true relative to a given environment. As an agent collects more experiences it collects information about how often it sees certain beliefs over the total number of times it has been in a certain environment. Over time these conditional probabilities give the agent an idea of what to expect when it enters an environment.

In our example ICARUS spends several cycles at the Rainbow table. As it is exploring and acting in the environment, the beliefs change. At a later time when the robot thinks about being at the Rainbow table it may realize things like "Box Box0 is always on the table" and "There was a block in the box for about half the time I was there". In subsequent interactions, unless proven otherwise, ICARUS would assume that `Box0` is on the Rainbow table. This *virtual sensing* is a process for discovering hidden or unseen facts about a given environment. Incorporating this ability into ICARUS is important because often times, agents act in partially observable worlds. So being able to recall information that pieces together a more complete view of the world may allow an agent to operate in a more natural and efficient manner.

Remembering and Knowing

Many psychologists and students of memory have discussed the nature of *remembering* and *knowing*. That is, once someone recognizes an item, is that item recognized because the person had a recollective experience of the item, or did he or she somehow know about the item without having any recollective experience? This nuance has encouraged researchers to characterize knowing and remembering responses and the protocols that govern the interaction of the two.

It seems that remembering is based on episodic memory while knowing responses are based on semantic memory (Gardiner, 1988; Rajaram, 1993; Knowlton & Squire, 1995). Specifically, in a remembering situation, responses are heavily influenced by the conditions present at the time of the encoded episode and by the amount of resources available to spend on remembering. For example, the more distracted a person is while performing a task, the less likely the person is able to recollect on what happened. On the other hand, knowing responses are automatic and influenced neither by the conditions at encoding nor the amount of resources available (Gardiner, 1988; Jacoby, 1991). While remembering involves searching through the episodic memory, knowing simply involves the state of the semantic memory, thus giving rise to the automatic property of know responses.

In our work, we take this research into account and aim to provide a computational theory that is consistent with these results. For instance, remembering responses in ICARUS utilize the episodic memory retrieval mechanism and are thus susceptible to encoding conditions as suggested in the literature. Knowing in ICARUS is facilitated by the semantic memory and does not involve extensive search, thus giving rise to seemingly automatic performance.

Discussions

We added an episodic memory to our theory of cognition in order to better capture the full range of human cognitive abilities for intelligent computational agents. One of the novel features of this episodic implementation is the ability to generalize knowledge at encoding. This serves a two-fold purpose. It provides an ordering to the episodes that lends itself to efficient search and secondly it reduces the demand during retrieval time to adapt a previous solution to the current situation because solutions can be arbitrarily specific. Since ICARUS learns from observations at the level of episodes, the ability to generalize knowledge at encoding also implies that the agent will be able to learn generalized models of how the world works.

We plan to build on this work in a number of different directions. We mention here three of them which, we believe, are most relevant. The strength of our system largely depends on ICARUS's ability to explain significant events. Towards that end, we would like to tightly integrate an explanation mechanism for the creation of episodes and to augment learning from observations within the context of the episodic memory. We will also expand the notion of expectations in ICARUS to use both the concept frequency tree and the primitive skills. This will enable the architecture to learn from different types of surprises. Another interesting direction is to implement a grammar for self-cueing. We expect that this will facilitate the agent's use of the episodic memory for a number of tasks like virtual sensing and other applications of case-based reasoning.

Conclusions

In this work, we detailed a computational model of episodic memory within the context of a cognitive architecture, ICARUS. We founded our approach on psychological evidence concerning the nature of episodic memory, the mechanisms of remembering and knowing, and the distinct features of implicit and explicit memory. We believe episodic memory is a fundamental component of human cognitive ability, and the extended architecture serves as an important basis for future research. We showed that the extension provides ICARUS with at least three new or improved cognitive functions. We plan to continue our research in this promising direction and hope to report in the near future the results of our evaluations using both qualitative and quantitative measures.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bölöni, L. (2011). An investigation into the utility of episodic memory for cognitive architectures. In *AAAI fall symposium: Advances in cognitive systems.*

Choi, D. (2011). Reactive goal management in a cognitive architecture. *Cognitive Systems Research*, *12*, 293–308.

Faltersack, Z., Burns, B., Nuxoll, A., & Crenshaw, T. L. (2011). Ziggurat: Steps toward a general episodic memory. In *AAAI fall symposium: Advances in cognitive systems.*

Fikes, R., & Nilsson, N. (1971). STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, *2*, 189–208.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*(4), 309–313.

Hellerstedt, R. (2015). *From cue to recall: The temporal dynamics of long-term memory retrieval*. Unpublished doctoral dissertation, Lund University.

Horn, A. (1951). On sentences which are true of direct unions of algebras. *Journal of Symbolic Logic*, *16*(1), 14–21.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 699.

Laird, J. E. (2012). *The soar cognitive architecture*. Cambridge, MA: MIT Press.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence.*

Nuxoll, A. M., & Laird, J. E. (2007). Extending cognitive architecture with episodic memory. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence* (pp. 1560–1565).

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, *21*(1), 89–102.

Schachter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 501–518.

Schiller, D., Eichenbaum, H., Buffalo, E. A., Davachi, L., Foster, D. J., Leutgeb, S., & Ranganath, C. (2015). Memory and space: towards an understanding of the cognitive map. *The Journal of Neuroscience*, *35*(41), 13904–13911.

Tulving, E. (1983). *Elements of episodic memory.* Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1), 1–25.

Wallace, S. A., Dickinson, E., & Nuxoll, A. (2013). Hashing for lightweight episodic recall. In *AAAI spring symposium series.*

Consistency and credibility in legal reasoning: A Bayesian network approach

Saoirse Connor Desai[1] (saoirse.connor-desai@city.ac.uk), **Stian Reimers**[1] (stian.reimers.1@city.ac.uk), **David Lagnado**[2] (d.lagnado@ucl.ac.uk), [1]Psychology, City University, London, EC1R 0JD UK, [2]Experimental Psychology, UCL,
London, WC1H 0AP UK

Abstract

Witness credibility is important for establishing testimonial value. The story model posits that people construct narratives from evidence but does not explain how credibility is assessed. Formal approaches use Bayesian networks (BN) to represent legal evidence. Recent empirical work suggests people might also reason using qualitative causal networks. In two studies, participants read a realistic trial transcript and judge guilt and witness credibility. Study 1 varied testimonial consistency and defendant character. Guilt and credibility assessments were affected by consistency but not prior convictions. Study 2 constructed a BN to represent consistency issues. Individual parameter estimates were elicited for the corresponding BN to compute posterior predictions for guilt and credibility. The BN provided a good model for overall and individual guilt and credibility ratings. These results suggest people construct causal models of the evidence and consider witness credibility. The BN approach is a promising direction for future research in legal reasoning.

Keywords: Legal reasoning, Evidential reasoning, Bayesian networks, Evidence, Reliability, Credibility

Introduction

Evaluating witness credibility is crucial to establishing the inferential value of testimony in criminal trials. The juror's task is to assess the truthfulness, reliability, and accuracy of the witnesses whose evidence is at issue, and evaluate how well the evidence supports the claims of the prosecution and defence in order to reach a verdict (Crown Court Bench Book, 2010). Lawyers can use many strategies to undermine the credibility of witnesses thereby challenging the reliability of their evidence. Drawing out testimonial inconsistencies under cross-examination, introducing contradictory testimony by other witnesses, disclosing information of previous convictions considered relevant to issues of credibility, and evidence relevant to a witness' reputation or truthfulness, are all important methods of assessing credibility (Spellman & Tenney, 2010). Legal reasoning studies confirm that people perceive inconsistent prosecution eyewitnesses as less accurate and credible, reducing the likelihood of conviction and increasing defendant credibility (Berman & Cutler, 1996; Berman, Cutler & Narby, 1995). Further, guilt judgments are sensitive to whether evidence contradicting an alibi shows a witness has been intentionally deceptive or made a genuine error (Lagnado & Harvey, 2008; Lagnado, 2011; Lagnado, Fenton & Neil, 2013). Mock juror studies also show that disclosing similar previous convictions affect judgments about the testifying defendant's credibility making them appear more likely to lie under oath and/or more likely to have committed the alleged crime (Lloyd-Bostock, 2000; Wissler & Saks, 1985).

Empirical studies of how people reason about legal evidence show that people reason about different types of evidence in complex ways (Pennington & Hastie, 1992). How people represent the credibility of witnesses and reliability of their evidence is important for understanding legal reasoning and determining what inferences are permissible given these representations. This paper presents a framework for analyzing the integration of testimonies whose sources vary in credibility and reliability. This approach builds on extant descriptive accounts of juror decision-making and employs the Bayesian network framework to model inferences about witness credibility and evidential reliability. Though intended mainly as a normative and prescriptive model of evidential reasoning this paper will show that the framework also captures peoples' ability to draw probabilistic conclusions from interrelated bodies of evidence.

Legal reasoning The story model of juror decision making is the leading cognitive model of how people reason about legal evidence (Pennington and Hastie, 1986, 1992). According to the story model, jurors organize and interpret the mass of evidence presented during the trial by constructing narrative explanations from the evidence. They use causal schemas – such as scripts of typical human thought and behavior – to fill gaps in the evidence and develop a causal 'situation' model of what transpired. The story ultimately adopted is the one that provides the best 'fit' for the evidence and is most plausible, complete and coherent.

The story model has achieved broad empirical support and has considerably advanced understanding of juror decision making. It qualitatively describes the constructive nature of people's explanations and emphasizes the interdependencies between trial evidence. One weakness of the story model is that it does not model how people reason about the credibility and reliability of different types of evidence and how this affects their story evaluation. In addition to reasoning about the crime itself, jurors (or fact-finders in general) must also reason about how well the evidence presented supports the arguments put forward by the prosecution and defence (i.e., what is the evidential support for a given story).

A further shortcoming of the story model is that it rejects the idea that people reason probabilistically. Even if people cannot estimate the precise probabilities of events, they can

often draw probabilistic conclusions from the evidence they hear in the courtroom (e.g. the presence of suspect at crime scene increases the likelihood that he is guilty).

Causal models of witness evidence are constructed in a similar way to causal models of the crime itself. Causal schemas about the nature of witnesses are used to draw inferences about the motivations and beliefs of people giving testimony; these inferences can also modify beliefs about the crime. Without a way to represent the relations between different types of evidence and how they interact it is difficult to elicit and test individual causal models constructed from the evidence. Further, without a formal analysis of how evidence items relate it is impossible to ascertain which inferences are permissible given the evidence.

Modeling evidence reliability The Bayesian network framework provides a potential solution to this problem. This approach makes it possible to test people's causal models of the reliability and credibility of witnesses giving testimony and compare inferences to a normative standard.

Bayesian networks (BN) use graph structures to represent the probabilistic relations between hypotheses and uncertain evidence, showing what inferences are rationally permitted from a given model of the evidence (Pearl, 1988, 2000). BNs have proved valuable for modeling relations in bodies of uncertain evidence in forensic contexts (e.g., Garbolino & Taroni, 2002) and have also been applied to legal contexts (Fenton et al., 2014; Lagnado et al., 2013). Fenton et al., (2014) claim that fact-finders (e.g., jurors) could use small-scale causal building blocks (*legal idioms* corresponding to common inference schemas) that make it possible to reason about complex and interrelated bodies of evidence. These idioms are customized to the legal context, capture generic patterns of legal inference, and can be re-used to make large-scale inferential problems tractable.

The basic idiom consists of the relation between a hypothesis and an item of evidence, corresponding to the relation between the legal proposition that needs to be proved (e.g., the defendant is guilty) and the submitted evidence. The evidence idiom can be supplemented with a reliability idiom enabling the modeling of potential causes of an evidence report that are vital for establishing the reliability of evidence from human sources. Reliability can be separated into issues of: i) observational sensitivity, ii) objectivity, and iii) veracity (Schum, 1994). The graph can be used to represent the fact that these different causes serve to explain the evidence. For example, the victim's testimony in an assault case depends both on whether or not the defendant assaulted the victim, and whether or not the victim is trustworthy and/or inaccurate. These factors are directly related to assessments of witness credibility.

Graphical models have been successfully applied to research in a number of areas of causal cognition (for review see Danks, 2014). Recent empirical work also suggests that people may reason about testimonies using qualitative causal networks (Lagnado, 2011; Lagnado, Fenton & Neil, 2013).

Study 1

Study 1 investigates some of the conditions under which witness credibility can be challenged, and how this impacts assessment of guilt. In particular, we explored how the consistency of the victim's testimony with other key pieces of evidence, and evidence disclosing the defendant's previous convictions, affected judgments of guilt and credibility, and how these factors combine and interact.

Real trial dialogue (*R v. Capel)* including cross-examination of witnesses was used to accentuate issues of credibility. These factors were specifically chosen because: 1) they address issues of witness credibility and reliability, and were raised in closing arguments and judge's directions, and 2) could be subtly manipulated in order to maintain ecological validity. The stimulus case has been used in previous mock jury research and typically results in a hung jury.

There were two main aims of Study 1: to establish whether people's judgments were affected by changes in witness consistency and whether it is feasible to model these changes using causal schemas of witness credibility and reliability. Given the findings of previous legal decision making research it was hypothesized: 1) that consistent testimony would result in greater belief in guilt, weaken the credibility of the defendant, and bolster the credibility of the victim, relative to inconsistent testimony, and 2) that disclosing a similar prior conviction to the current crime would result in greater belief in guilt, weaken the defendant's credibility and bolster the victim's credibility.

Methods

Participants 126 U.S. and U.K. based participants (64 female, mean±SD age 29.12±10.80, randomly split between six conditions) were recruited from https://www.prolific.ac/ a site for recruiting participants for web-based studies, and were paid £2.40($3.49). Average completion time was 31 minutes.

Design, materials and procedure We investigated the impact of prior character (three levels: no prior, different prior, similar prior) and inconsistency (two levels: consistent, inconsistent) on evidential reliability. Participants read one of six versions of a realistic courtroom transcript. Consistency was manipulated by varying the consistency of key pieces of testimony with the victim's testimony (i.e., consistent or inconsistent). Prior conviction evidence was manipulated by substituting evidence of good character (i.e., revealing the fact that the defendant has no prior convictions) with evidence of bad character (i.e., disclosure of a prior conviction).

The transcript was divided into 20 evidential statements and judgments about the probability of guilt and credibility of the victim and defendant were elicited after each

statement. All participants saw the evidential statements in the following order: the charge and plea, the prosecution's opening statement, direct and cross-examinations of three prosecution and three defence witnesses, prosecution and defence closing arguments, the judge's summary, and instructions on the law. They read through the evidential statements at their own pace and updated their ratings the probability of guilt, and credibility of the victim and defendant, in light of each statement. After reading all the evidence, participants indicated their final judgments about guilt, and credibility of the victim and defendant.

Table 1: Manipulation of Key Pieces of Evidence

	Victim's Friend	Police Officer	Bartender
Inconsistent	8 pints	JD drunk	Saw push
Consistent	4 pints	JD not drunk	No push

Case summary and consistency manipulations The defendant (SC) is charged with assaulting the victim (JD), and pleads not guilty. The prosecution argues that SC punched JD in an unprovoked attack and calls three witnesses: JD, JD's friend, and a police officer who was at the scene. The defence argues that the punch was an act of self-defence; JD was drunk, aggressive and had pushed SC first. The defence also calls three witnesses: SC, SC's friend, and a local bartender who was also at the scene. JD makes three important claims in his testimony: 1) that he had only had 4 pints to drink, 2) he was not drunk, and 3) he did not provoke SC first. In the inconsistent version, these claims were contradicted by three witnesses: JD's friend testifies that JD drank 8 pints, the police officer testifies that JD appeared drunk, and the bartender states he saw JD push SC first. In the consistent version, these claims were not contradicted: JD's friend corroborates JD's claim that he only drank 4 pints; the police officer states that JD did not appear drunk, and the bartender does not mention that he saw the push (see Table 1). Inconsistencies between the victim and subsequent testimonies regarding these key issues could make the victim appear likely to have lied about the incident because he was drunk and aggressive or show that his recollection of events was inaccurate.

Study 1: Results

Probability of guilt Mean probability of guilt judgments given at the end of the trial, across all six conditions, are shown in Fig. 1. A prior conviction for a similar crime (M = 67.60, SD = 28.27) increased belief in guilt relative to prior conviction for a different crime (M = 65.16, SD = 29.95), or no prior convictions (M = 64.32, SD =27.58), but these differences were not significant F (2, 98) = 1.90, p = .16. Introducing key pieces of evidence that were inconsistent with the victim's testimony (M = 58.98, SD = 29.03) resulted in significantly diminished belief in guilt relative to a consistent testimony (M = 72.20, SD = 26.49), F (1, 98) = 12.35, p < .001.

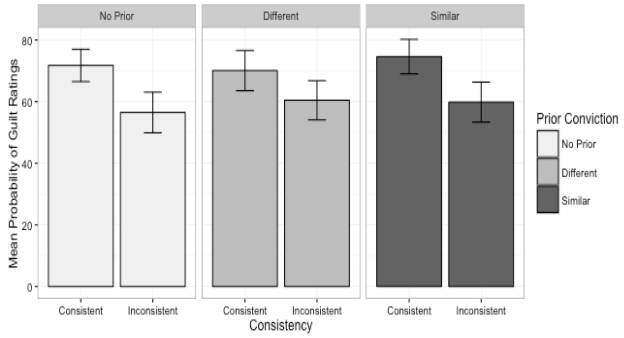


Figure 1: Mean final probability of guilt judgments (± standard error) for each condition of the study. Range from 0 -100.

Defendant credibility The defendant was considered to be more credible after an inconsistent testimony (M = 44.69, SD = 22.77), than a consistent one (M = 35.55, SD = 21.40), but this difference was not significant, F (1, 98) = 1.94, p = .17. Defendant credibility ratings were in the expected direction, defendant appeared more credible when he had no prior convictions (M = 43.59, SD = 21.20), than a prior conviction for a different (M = 39.40, SD = 26.37), or similar (M = 37.26, SD = 19.17) crime, but the differences were not significant, F (2, 98) = .08, p = .92.

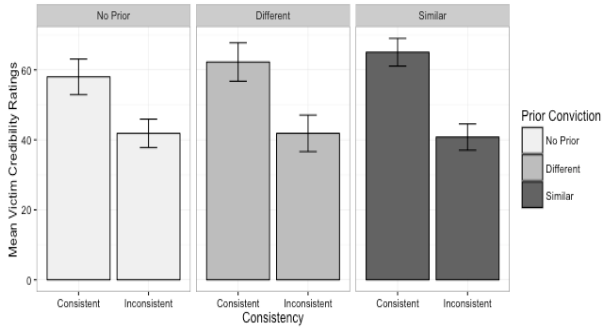


Figure 2: Mean victim credibility judgments at the end of the trial (± standard error) for each condition of the study. Range from 0 -100.

Victim credibility Mean victim credibility judgments across the six conditions are shown in Fig. 2. The victim appeared more credible when the defendant had a prior conviction for a similar (M = 53.50, SD = 21.40), than different (M = 51.81, SD = 26.60) offence, and rated least credible when the defendant had no prior convictions (M = 50.12, SD = 22.26), but these differences were not significant, F (2, 98) = .14, p = .87. The victim appeared more credible when evidence was consistent with his testimony (M = 61.81, SD = 22.31) than when it was inconsistent (M = 41.52, SD = 19.93), F (1, 98) = 16.97, p < .001. Evidence relating to the defendant's credibility did not influence the perceived credibility of the victim.

Inconsistencies between the victim's testimony evidence mentioned in other witness testimonies substantially undermined the victim's credibility.

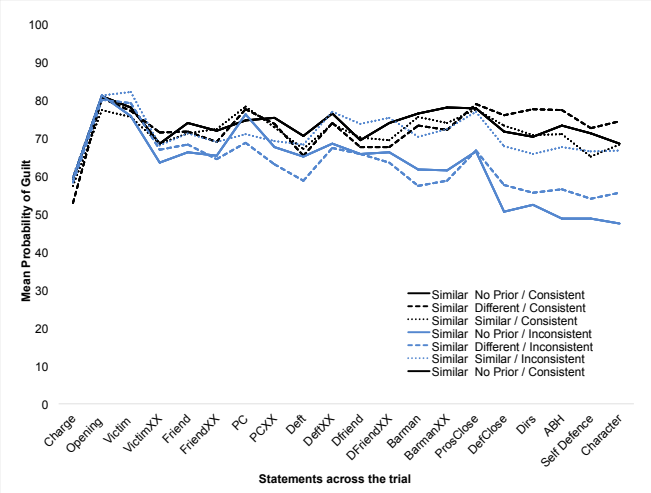


Figure 3: Mean probability of judgments across statements for each condition of in Study 1. Range from 0 -100. The graph shows that the bartender testimony is important for establishing probability of guilt.

Discussion

The results suggest that people draw inferences about witness credibility from subtle inconsistencies between their testimony and the testimonies of other witnesses, which in turn influence their beliefs in the defendant's guilt. In this study, there was no clear impact of evidence aimed at undermining the defendant's character, and therefore his credibility, or inferences about guilt. One reason that disclosing prior convictions had no impact on guilt in the current study could be due to methodological differences between current and previous studies. In previous studies participants read descriptions of hypothetical cases in which prior conviction evidence outweighed other evidence in the case. In this study prior conviction information was balanced with other issues in the case, which could account for the lack of difference between conditions. These results suggest that people do more than constructing a plausible story from the evidence; also factoring in the consistency and credibility of witnesses. The key effect of consistency on people's judgments of credibility and guilt was replicated in a laboratory study too.

Study 2

Study 2 was designed to replicate the consistency effect and to examine whether the idiom-based model provided a good fit for participants' guilt and credibility ratings. We constructed a BN to represent the key pieces of testimony that were varied in Study 1 (see Fig. 4). The BN captures the impact of the three witness testimonies that were varied on the probability that JD was drunk, is a credible witness, and provoked SC. Individual parameter estimates were elicited for the corresponding CPT in the BN in order to compute posterior predictions for each of the aforementioned probabilities.

Methods

Participants 137 participants (65 female, mean ±SD age 32.52 ± 10.67, randomly split between two conditions) were recruited from https://www.prolific.ac/ and completed the study for monetary compensation (£2.40/$3.49). The average completion time was 25 minutes.

Design, materials and procedure We manipulated the consistency of key evidence with the victim's testimony in the same way as Study 1 (Table 1), thus the materials were identical except character evidence (i.e., defendant prior convictions not manipulated). In Study 2 we also took participants' estimates of the conditional probabilities to complete individual parameterizations of the BN model, after they provided posterior guilt and credibility ratings. These questions were asked in the following format:

a) Suppose that JD had drunk 8 pints.
What's the probability that the policeman would testify that JD appeared drunk/did not appear drunk?
b) Suppose that JD had drunk 4 pints.
What's the probability that the policeman would testify that JD appeared drunk/did not appear drunk?

Conditional probability ratings were indicated using a slider ranging from 0 = *exceptionally unlikely* to 100 = *virtually certain*, and pre-set to the midpoint. Participants in both Experimental groups completed these questions, however, questions that related to specific details in the testimonies of the three witnesses were adapted according to condition (e.g., JD appeared OR JD did not appear drunk).

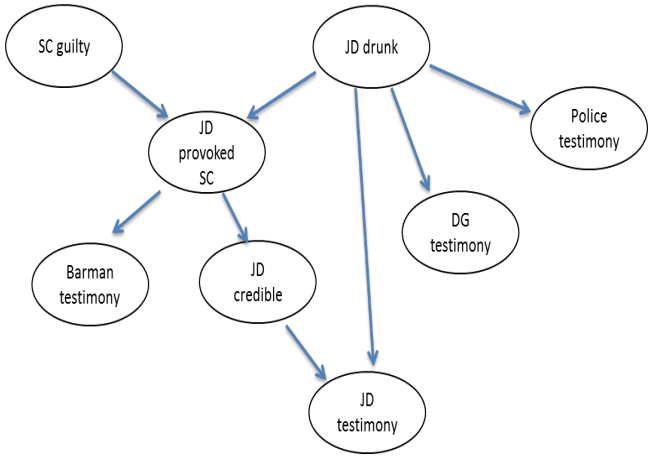


Figure 4: A Bayesian network of *R v. Capel* using evidence- reliability idiom to capture JD credibility.

Study 2: Results

Observed posterior probability judgments After considering all the evidence, a consistent testimony resulted in greater belief in guilt (M = 73.87, SD = 27.42) than an inconsistent testimony (M = 68.34, SD = 29.13), but this difference was not significant, t (135) = -1.41, $p > .05$. When key pieces of evidence were inconsistent with the victim's testimony (M = 44.67, SD = 24.19), the defendant appeared more credible than when evidence was consistent (M = 35.58, SD = 22.98), t (135) = 2.52, $p < .05$. The victim was also considered less credible after an inconsistent testimony (M = 50.14, SD = 22.78) than a consistent one (M = 64.58, SD = 24.83), t (135) = -3.55, $p < .001$. The results almost replicate the effect of consistency in Study 1 and were in the same direction. However, observed posterior guilt judgments showed an overall stronger tendency toward guilt than in Study 1, which could be explained by variation in people's prior expectations and assumptions about criminal proceedings or defendants in general.

Modeling participants' inferences using the BN To test the BN model more formally, we analyzed model predictions using the conditional probabilities provided by participants. The conditional probabilities concerned key issues relating to the reliability of JD's testimony (shown in Table 1)[1].

First, participants' mean probability judgments were used to parameterize separate graphs for each condition. The posteriors generated by the model showed that the probability JD was drunk was higher for the inconsistent (.79) than the consistent condition (.24), likewise the probability that JD provoked SC was higher for the inconsistent (.77) than the consistent (.17) condition, and the probability that JD is a credible witness was lower for the inconsistent (.42) than consistent condition (.70). This shows that the BN provided a good model for participants' judgments of guilt and credibility. The discrepancy between the predicted probabilities estimates and observed posterior probability judgments also suggest that participants are reasoning about factors other than consistency.

To examine whether the BN captured participants' judgments at an individual level, we used each participants' unique conditional probabilities to parameterize the BN. We then used this model to compute posteriors for JD credibility, the probability JD provoked SC, and the probability JD was drunk, and compared these model predictions with participants' actual posterior judgments of guilt and credibility. Observed posterior judgments of guilt and credibility were correlated with the model posterior prediction for the probability that JD provoked SC. In this instance, the model prediction for 'JD Provoke' served as a proxy for the probability of guilt. More precisely, a higher posterior probability that JD provoked SC, would be associated with lower observed guilt ratings, lower ratings of JD's credibility, and higher ratings of SC's credibility, and vice versa.

The predictions derived from the BN were supported by the data. The model predictions for provoke were negatively correlated with observed posterior probability of guilt ratings, r = -.37, $p < .001$, R^2 = .14. Provoke was also negatively correlated with victim credibility ratings r = -.40, $p < .001$, R^2 = .16, and positively correlated with defendant credibility ratings r = .56, $p < .001$, R^2 = .45. The model predictions for JD credible were also positively correlated with ratings for guilt and JD credibility and negatively correlated with SC credibility. The model prediction for JD drunk was positively correlated with SC credibility but was not correlated with observed posteriors for guilt or JD credibility.

These results show that the BN model provides a good fit for individual participants (despite the simplicity of the BN model). The model predicts that the more credible JD is considered to be, the more unlikely it was that he was drunk and provoked SC, and this was upheld by the data. The R^2 value shows that the posterior predictions for provoke share some of the variability in the guilt and credibility judgments. The relatively small R^2 is most likely due to noisy participant' judgments, because we used a rich set of materials, and because provoke was used as a proxy for guilt. In addition, we only predicted the impact of consistency of selected pieces of evidence with the victim's testimony and did not model other elements of the case. Other factors were identified in closing arguments (e.g., the force of the punch) that might have been equally important for reasoning about the defendant's guilt. Nonetheless, the results of Study 2 demonstrate that it is possible to construct causal BNs to represent evidence that includes critical aspects of witness credibility, and then compare the model predictions with actual inferences.

General Discussion

The studies reported in this paper examined how the consistency of testimony affects inferences about witness credibility and judgments of guilt in a realistic legal reasoning task. Our results suggest two main conclusions. Firstly, that the consistency between key pieces of evidence and the victim's testimony, as identified in closing arguments, affect judgments about witness credibility and the degree of belief in the defendant's guilt. More specifically, belief in guilt was lower when key pieces of evidence mentioned in the testimonies of subsequent witnesses was inconsistent with details given in the victim's testimony. These inconsistencies damaged the credibility of victim and strengthened the credibility of the defendant. Secondly, the pattern predicted by the BN model captured qualitative patterns of inference displayed by participants at a general and individual level. In the consistent version of the case, participants thought the victim was credible, unlikely to be drunk, and unlikely to have provoked the defendant. Inconsistent statements lead to the opposite

[1] A prior of drunk =.50 was used for all the modeling.

pattern, the victim appeared less credible, more likely to be drunk and more likely to have provoked the defendant. Study 2 provides direct support for the reliability model described earlier.

The results of this study strengthen the claim that people's reasoning is sensitive to interrelations between testimony, credibility, and reliability that are predicted by the qualitative aspects of Bayesian network models. The BN captures people's intuitions that credibility is related to whether or not we think the victim's testimony is caused by the fact that he was indeed assaulted as he claimed or that he was drunk and provoked the defendant into a fight. It is therefore possible to model the strength of the links between the different items of evidence using the Bayesian network supplemented with legal idioms.

These results complement the story model and show that in addition to constructing a causal model of the crime people also construct causal models of the witnesses giving evidence and this affects their story evaluation. The results show that people do more than construct a plausible story to explain the evidence, but also take into account issues of credibility and reliability, such as the consistency of testimony. These results suggest that it is possible to extend the story model to include issues relating to credibility and reliability of evidence (Lagnado et al., 2013). In fact, these results strengthen Pennington and Hastie's claim that evidential reasoning is not a straightforward updating process as has been proposed by belief-adjustment models (Hogarth & Einhorn, 1992). This methodology fits with previous Bayesian approaches to modeling belief updating (e.g., Hahn & Harris, 2009) and by using Bayesian networks adds a richer structural account of people's reasoning.

The findings reported in this paper demonstrate that it is possible to model testimony integration using BNs. One potential shortcoming of the approach used here is that we used the model posterior probability for whether the victim provoked the defendant as a proxy for inferences about guilt. This could explain why the correlations between the model predictions and the observed posterior judgments were not stronger. Another reason that the model was not better able to capture people's inferences is that we only modeled some of the evidence in the case. More specifically, the BN focused on issues relating to the victim's testimony and did not include critical evidence relating to the defendant's testimony (e.g., the defendant was seen running away after the assault). Furthermore, the observed posterior judgments showed that consistency also influenced perceptions of the defendant's credibility, which was also not included in the BN. The BN model could readily be extended to include these factors.

Conclusions

In sum, this research shows that it is possible to model legal arguments using the BN framework and that this approach describes the qualitative patterns of inference exhibited in legal reasoning. Although it appears that people use network-like structures to reason about evidence it is still necessary to develop a fuller psychological model of how people represent hypotheses and reason about uncertain evidence. Building upon the story model, and the insight that people use causal representations and inference, is a promising direction for further research.

Acknowledgements

We thank Prof. Cheryl Thomas who provided the transcript used in these studies, and Jacob de Zoete who assisted with modeling the data.

References

Berman, G. L., & Cutler, B. L. (1996). Effects of inconsistencies in eyewitness testimony on mock-juror decision making. *Journal of Applied Psychology*, *81* (2), 170.

Berman, G. L., Narby, D. J., & Cutler, B. L. (1995). Effects of inconsistent eyewitness statements on mock-jurors' evaluations of the eyewitness, perceptions of defendant culpability and verdicts. *Law and Human Behavior*, *19* (1), 79.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. MIT Press.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Garbolino, P., & Taroni, F. (2002). Evaluation of scientific evidence using Bayesian networks. *Forensic Science International*, *125*(2), 149-155.

Harris, A. J., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1-55.

Judicial Studies Board (2010). *Crown Court Bench Book: Directing the Jury, London, UK*

Lagnado, D. (2011). Thinking about Evidence. In *Proceedings of the British Academy*, 171, 183-223.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46-63.

Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic bulletin & review*, *15*(6), 1166-1173.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Palo Alto, CA: Morgan Kaufmann.

Pearl, J. (2000). Causality. New York: Cambridge University Press (2nd edition).

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. Journal of personality and social psychology, 62(2), 189.

Schum, D.A. (1994), *The Evidential Foundations of Probabilistic Reasoning*, Northwestern University Press.

Spellman, B. A., & Tenney, E. R. (2010). Credible testimony in and out of court. *Psychonomic bulletin & review*, 17(2), 168-173.

Attention and the Development of Inductive Generalization: Evidence from Recognition Memory

Tracey Miser (miser.4@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH43210 USA

Vladimir Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH43210 USA

Abstract

Induction, the ability to generalize knowledge from known to novel instances, is essential for human learning. This study investigates how attention allocation during category learning and induction affects what information is represented and encoded to memory. In Experiment 1 5-year-olds and adults learned rule-based categories. They were then presented with an Induction-then-Recognition task. Similar to previous results with familiar categories, children exhibited better memory for items than adults. In Experiment 2, adults learned similarity-based categories and then were presented with an Induction-then-Recognition task. In this condition, adults' memory was as good as children's memory in Experiment 1. These results indicate that the way categories are represented affects the way induction is performed.

Keywords: Induction; Learning; Memory.

Introduction

By means of induction humans have the ability to generate new knowledge, reason about new objects, and learn in new situations. Induction enables the generalization of properties from the familiar to the novel. For example, if one learns that honey bee exoskeletons are composed of chitin, this knowledge could be extended to other types of bees and wasps, to all hymenoptera, or even to all arthropods. There is considerable evidence that inductive generalization appears early in development (Gelman & Markman, 1986; Sloutsky & Fisher, 2004a), with infants as young as 9-months exhibiting this ability (Baldwin, Markman & Melartin, 1993; Graham, Kilbreath, & Welder, 2004; McDonough & Mandler, 1998; Mandler & McDonough, 1998), and further evidence that the development of induction is protracted to at least 11 years of age (Fisher& Sloutsky, 2005). Despite considerable literature describing the mechanisms of induction, it is still not understood *what*, if anything, develops. Thus, the goal of the current study is twofold: (1) to understand *what* changes, and (2) to determine how those changes are reflected in the use of different mechanisms of induction.

In order to describe the mechanisms of induction, two theories have been proposed: the knowledge-based and the similarity-based views. Each theory presents arguments for how induction is performed and describes what developmental changes may occur. According to the knowledge-based view, induction is always based on category information, even in young children. In knowledge-based induction, the category of a novel item is identified and properties are inferred based on whether the novel item belongs to the same category as the familiar target. Proponents contend that children are able to make knowledge-based inferences because they hold many a priori assumptions about categories (Gelman & Markman, 1986). For example, assumptions that natural kind categories share many properties, including internal structure (Gelman, 1988; see Murphy, 2001 for a review). Considering children are most familiar with basic-level categories, this view posits that children are more likely to infer properties at the basic-level.

According to the similarity-based view, knowledge-based induction is a product of development and not reliant on a priori assumptions, while early induction is similarity-based. In similarity-based induction, multiple features of the novel input are assessed for similarity to a familiar entity. Properties are then inferred based on how similar the novel item is to the known target. There is evidence that, in addition to visual features (Sloutsky & Fisher, 2004a), linguistic labels (Deng & Sloutsky, 2013) and salient motion (Deng & Sloutsky, 2012) may factor in the similarity computation. According to this view, in knowledge-based induction, a novel item is labeled (either externally or by self-generating the label), then the label of the novel item is compared to the label of the familiar target and properties are inferred based on whether the entities belong to the same category. As such, this mechanism depends on the knowledge that items from the same category share many properties, including shared labels (Deng & Sloutsky, 2013), but also causal (Badger & Shapiro, 2012Bulloch & Opfer, 2009), and ontological (Gelman & Davidson, 2013) properties.

A critical difference between the above accounts is whether children use a different mechanism than adults when performing induction. In order to test this, Sloutsky and Fisher (2004a) developed the Induction-then-Recognition paradigm (ITR) in which participants perform an induction task followed by a surprise memory test of items that were presented during induction and items that were not. The ITR paradigm is based on the "level-of-processing effect" in which deeper semantic processing

increases recognition accuracy (i.e., "hits"; Craik & Lockhart, 1972; Craik & Tulving, 1975), while at the same time also increasing the proportion of false recognitions of non-presented, but categorically related, items (i.e., "false alarms"; Koutstaal & Schacter, 1997; Rhodes & Anastasi, 200; Tharpar & McDermott, 2001). Thus, the overall result (i.e., hits – false alarms) is negative. In contrast, shallower perceptual processing may yield similar proportions of hits, but false alarms will be much lower (Marks, 1991). The result in this case is more accurate memory.

Following the above logic, Sloutsky and Fisher (2004a) proposed that memory for items presented in an induction task may reveal the level of processing used when performing induction. That is, use of a similarity-based mechanism, which is shallower than the semantic level of processing, will result in more accurate memory following induction. This pattern of memory was predicted for young children with either familiar or novel categories and for adults with novel categories. And while the above account has received much empirical support (Sloutsky & Fisher, 2004a; 2004b; Fisher & Sloutsky, 2005), many questions remain. For example, if the knowledge-based theory is correct and children use the same mechanism as adults, what explains age related differences in memory for items following induction (Sloutsky & Fisher, 2004a; Fisher & Sloutsky, 2005)? However, if the similarity-based view is accurate and children use a different mechanism than adults, what accounts for high induction performance in tasks where perceptual information is in contrast with category label (Gelman & Markman, 1986)? To address these questions, each view posits specific predictions about *what* develops.

The knowledge-based view posits that what develops is category knowledge (Gelman, 1988). This view holds that the mechanism driving induction does not change, and any changes in how induction is performed can be explained by changes in children's knowledge. Such age related gains include a shift from domain general to domain specific knowledge, knowledge regarding the hierarchical structure of categories, and scientific knowledge (Gelman & E. Markman, 1986; Gelman & Heyman, 1999; Gelman, 2004). For example, Gelman and O'Reilly (1988) found that school age children were able to successfully perform induction when superordinate category labels were provided, while preschool children did not. According to the authors, this finding demonstrates developmental gains in understanding of, and the ability to use, hierarchical category structure.

On the other hand, the similarity-based view holds that what develops is a different mechanism of induction (Sloutsky & Fisher 2004a). That is, changes in induction performance are due to a shift from an early similarity-based mechanism to a mature knowledge-based mechanism. If this is the case, then Gelman and O'Reilly's (1988) findings could result from children failing to make inferences within the superordinate category because the exemplars provided for the broad category of "animal" were too perceptually different to promote similarity-based induction. Importantly, according to this theory, a similarity-based mechanism would still be available for use by adults with novel categories, not that one mechanism replaces the other.

Adding to the above arguments regarding *what* develops, there is growing evidence in the category learning literature indicating that developmental changes in attention contribute in key ways to changes in category learning and representation. In an eye-tracking study by Best, Yim, and Sloutsky (2013), it was found that adults, but not 6- to 8-month-old infants, demonstrated a cost of attention when switching from categorizing learned categories to novel categories. Adults continued to focus on the previously learned category rule feature even when new categories were introduced. Infants, however, did not show a similar cost. It was argued that the difference between infants' and adults' switching was due to adults' more developed selective attention and resulting focus to a single predictive feature. Similar results with adults were found by Hoffman and Rehder (2010). In that study, participants demonstrated the same cost of attention such that they continued to attend to a category rule feature even when new categories were introduced.

The changing role of attention in categorization was also demonstrated by Deng and Sloutsky (2015). In that study, children and adults were trained to perform either item classification or feature inference, it was found that older children and adults demonstrated an asymmetry between their responses during item classification and feature inference such that participants relied on a single deterministic feature when classifying, but on multiple probabilistic features when inferring a missing feature. The same asymmetry was not found in younger children. The authors argued that younger children's representations remained similarity-based regardless of training, while older children's and adults' representations were flexible (i.e., capable of being knowledge- or similarity-based) depending on the demands of the task.

Considering that categorization, like induction, is a generalization process, it is possible that developmental changes in attention also contribute to changes in how inductive generalization is performed. Distributed attention may promote forming a similarity-based category representation as well as a similarity-based mechanism of induction. If this is the case, then memory for specific items would be high. That is, attention to multiple features would result in encoding the overall appearance of an item (Marks, 1991). Alternately, attention focused to a single feature, such as a rule feature or label (Deng & Sloutsky, 2013), would promote forming a knowledge-based representation and use of a knowledge-based mechanism of induction. It then follows, use of this mechanism would result in poor memory discrimination for individual entities because encoding would be limited to a single feature.

Overview of the Current Study

The current experiments were designed to further examine the mechanism of induction across development. In Experiment 1, children and adults were taught rule-based categories. After testing their categorization, participants were given an Induction-then-Recognition task with one of the studied categories after which their memory for items was tested. If children perform induction on the basis of similarity, and adults on the basis of the rule, then children should exhibit better memory for items than adults. However, even if children's memory is better than adults', it could be argued this effect stems from adults being more efficient in their induction. This issue was addressed in Experiment 2 where adults learned similarity-based categories and were then presented with the ITR task. If adults are merely more strategic at induction than young children, then memory for items in this condition should remain as low as in Experiment 1. In contrast, if Experiment 1 reflects differences in how induction is performed, then adults in this experiment should exhibit better memory for items after induction than in Experiment 1.

Experiment 1: Deterministic Instructions

Method

Participants Thirty-one adults (M_{age} = 20.3 years) and thirty-four children (M_{age} = 5.4 years) participated in the experiment. One additional child and two additional adults were dropped from analysis due to failure to follow directions during part of the experiment. Two children were dropped from analysis for not responding above chance to High and Medium Similarity items at Category Testing.

Materials Both experiments presented here used visual stimuli consisting of two categories of artificial insects. The category structure was similar to that previously used by Deng and Sloutsky (2012, 2013). Each category was provided a novel label: dax (Category D), and fep (Category F). Items in both categories were composed of seven features (head, body, legs, wings, antennae, claws, and tail), which differed between categories on the shape and color of each feature. All the default features of Category D were given the value of 1 (e.g., head1, body1, etc.), and all default features of Category F were given the value of 0 (e.g., head0, body0, etc.; see Table 1 for example stimuli structures). Each category had one prototype that was assembled from all of the default feature values for that category (Prototypes are pictured in Figure 1). The remaining stimuli were constructed by exchanging probabilistic features between categories or introducing new probabilistic features.

Of the seven features present on each stimulus, one was deterministic and was never exchanged between categories. That feature (the tails) remained constant within a category and perfectly predicted category membership (i.e., the category rule feature). The remaining six features were probabilistic and varied from one exemplar to the next so that the overall appearance of the items was also predictive of category membership. Subsets of the primary stimuli described above were created for use during different phases of the experiment.

Procedure The experiment consisted of four phases: (1) Instructions and Category Training, (2) Category Testing, (3) Induction, and (4) Recognition Memory. Instructions and feedback for all phases and both age groups were presented in text on the computer monitor. Adults read the instructions at their own pace, while children were read the instructions aloud by an experimenter. Adults made keyboard responses, and children made verbal responses which were then logged by the experimenter via the keyboard.

Instructions and Category Training: In Instructions prior to Training, participants were shown the default deterministic feature for Category D, then the default deterministic feature for Category F. Features were presented one at a time along with the feature label and a statement directing attention to that feature (e.g., "Daxes always have this kind of tail."). During Training, corrective feedback was provided after every trial with statements directing attention to the deterministic feature (e.g., "Correct! That one is a dax. It has a dax tail."). This phase contained three blocks of twelve trials each. Only High Similarity items were presented in this phase (see Table 1 for stimuli structure).

Category Testing: Participants were told to continue categorizing, as in Training phase. This phase contained two blocks. Block one presented the same twelve High Similarity items from Training, plus twelve new Medium Similarity items. Block two presented twelve Switch items (see Table 1 for stimuli structures). No feedback was provided in this phase.

Induction: Participants were shown an example of a High Similarity Category F exemplar and were informed that "This fep has beta-cells in its body." Participants were then asked to view other animals and decide whether they also have this property. Corrective feedback followed each trial indicating that only the target category (Category F) exemplars had beta-cells (e.g., Incorrect. That one does not have beta-cells). Items included six Probabilistic New items from each category, six Probabilistic Switch items from each category (see Table 1 for stimuli structures), and twelve Distracter items. This phase contained one block of 36 trials. The upcoming recognition memory test was not mentioned.

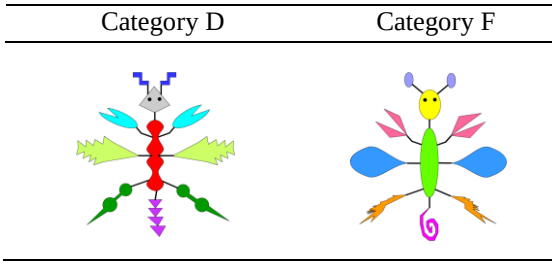
Recognition Memory: Participants were instructed they would be presented "old" items from the Induction phase as well as "new" items that had not been presented at any time during the experiment. Subjects were asked to determine

Table 1: Example of stimulus structures used in Experiments 1-2

	Category F								Category D							
	Head	Body	Legs	Wings	Antenna	Claws	Tail		Head	Body	Legs	Wings	Antenna	Claws	Tail
Prototype	0	0	0	0	0	0	0	Prototype	1	1	1	1	1	1	1
High Sim.	0	0	1	0	0	0	0	High Sim.	1	0	1	1	1	1	1
Med Sim.	1	1	0	0	0	0	0	Med Sim.	0	0	1	1	1	1	1
Switch	1	1	0	1	1	1	0	Switch	0	0	0	0	1	0	1
Prob. New	0	0	0	0	N0	0	0	Prob. New	1	N1	1	1	1	1	1
Prob. Switch	0	0	0	N1	0	0	0	Prob. Switch	1	1	N0	1	1	1	1
Low Sim.	N1	1	1	N2	N0	0	0	Low Sim.	1	1	N1	N2	N0	1	1

Note: 0 = dimension corresponding with Category F features; 1 = dimension corresponding with Category D features. Nn = new feature and the corresponding feature identifier.

Figure 1: Prototypes of Category F and Category D



Note: Prototypes were not presented during the experiment.

which items were old and which were new. Items included the same twelve target category items from the Induction Phase (old Category F exemplars), twelve novel Low Similarity target category items (new Category F exemplars, see Table 1 for stimuli structures), six non-target category items from the Induction Phase (old Category D exemplars,), and six novel Distracters.

Results and Discussion

Accuracy in Category Training was high for both children (M = 95%) and adults (M = 97%) with both groups performing significantly better than chance (both ps < .001). Proportions of rule-based responses to High Similarity and Switch items in Category Testing are reported in Table 2 along with results of Experiment 2. Both children and adults made significantly more rule-based responses than would be expected by chance (both ps < .001). Induction accuracies were also above chance for children (M = 85%, $t(30)$ = 13.14, p < .001) and adults (M = 91%, $t(30)$ = 18.88, p < .001). The remaining analysis focused on results of the Recognition Memory Phase.

Memory discrimination was analyzed using signal detection d' scores calculated from the Z-score normalized proportions of hits (H) and false alarms (FA). If participants do not discriminate old from new items, d' is at or below 0. Experiment 1 d' scores are presented in Table 3 alongside d' scores for Experiment 2. Children's memory (H(.77) – FA(.31) = .46, d' =1.48) was significantly better than that of adults (H(.80) – FA(.45) = .36, d' = 1.06), $t(60)$ = 1.92, one tailed p=.03.

Results of Experiment 1 indicate that both children and adults ably learned the categories, generalized learning to new exemplars, and correctly inferred properties to the target category. Furthermore, children's proportion of rule-based responses on Switch items was very high, suggesting that they (like adults) did learn a rule-based category. However, children exhibited better memory than adults, indicating they did not use their knowledge of the rule to perform induction. This finding suggests that children were likely attending to more than just the category defining feature during induction. These findings are consistent with the theory that children rely on overall similarity when performing induction. It could be argued however, that these memory differences simply reflect developmental differences in the efficiency of allocating attention during induction. Experiment 2 addresses this question.

Experiment 2: Probabilistic Instructions

This experiment was similar to Experiment 1, with one exception: adults were trained on similarity-based (rather than rule-based) categories. If Experiment 1 simply reflects differences in attentional efficiency during induction, adults' memory after induction should remain low in Experiment 2. In contrast, if memory for items is reflective of the mechanisms of induction, adults' memory in Experiment 2 should be higher than that in Experiment 1.

Method

Participants Thirty-seven adults (M_{age} = 20.3 years) participated. Four additional adults were dropped for failing to follow directions (N = 2) or not meeting category learning criteria described in Experiment 1, N = 2).

Materials, Design and Procedure Visual Stimuli were identical to those used in Experiment 1. Procedures for this experiment were similar to those described in Experiment 1, with two exceptions: (1) Instructions preceding Category Training, and (2) feedback during Category Training. In this experiment, instructions preceding Category Training introduced each probabilistic feature, one feature at a time, and one category at a time, for both categories. Each feature was presented with text indicating the feature label and a statement directing attention to that feature (e.g., "Most feps have this kind of head."). Feedback during this phase re-directed attention to overall appearance of the items (e.g., "Correct! This is a fep. It looks like a fep.").

Results and Discussion

Accuracy in Category Training was high (M = 86%) and significantly better than chance ($p < .001$). Proportions of rule-based responses to High Similarity and Switch items presented in Category Testing are reported in Table 2. The proportion of rule-based responses on High Similarity items was significantly better than chance ($p < .001$), however, Switch item responses in this experiment were lower than adults in Experiment 1 ($F(1, 62) = 497.65$, $p < .001$) and lower than chance ($t(32) = -10.36$, $p < .001$). Induction accuracy was high (M = 81%), perhaps somewhat lower than that in Experiment 1 but still significantly higher than chance ($t(32) = 10.21$, $p < .001$). The remaining analysis focused on memory discrimination in the Recognition Memory Phase.

Table 2:Mean Proportions of Rule-based Responses during Category Testing Phase

	High Similarity	Switch
Adults- Experiment 1	.98 (.03)	.97 (.07)
Children- Experiment 1	.92 (.09)	.91 (.17)
Adults-Experiment 2	.82 (.10)	.16 (.19)

Note: Experiment 1 (Deterministic Instructions), Experiment 2 (Probabilistic Instructions). Standard deviations are in parentheses.

Memory discrimination d' scores for Experiment 2 are presented in Table 3, alongside d' scores for Experiment 1. To further examine adults' memory discrimination, *d'* scores from both experiments were submitted to a One Way ANOVA. The analysis revealed a significant difference between the experiments, $F(1, 60) = 10.28$, $p = .002$, with adults in Experiment 2 having higher memory sensitivity scores (H(.79) – FA(.23) = .56, $d' = 1.80$).

Results of Experiment 2 clearly indicate that post-induction memory is reflective of how induction was performed. Adults' high memory in this experiment indicates that results of Experiment 1 do not stem from adults' having more efficient attention allocation than young children.

Table 3:Mean *d*' scores for Experiments 1 and 2

Age Group	*d'*
Adults – Experiment 1	1.06 (.88)
Children – Experiment 1	1.48 (.84)
Adults – Experiment 2	1.80 (.80)

Note: Experiment 1 (Deterministic Instructions), Experiment 2 (Probabilistic Instructions). Standard deviations are in parentheses.

General Discussion

The two experiments reported here aimed to further understand the development of induction, and specifically, the role of attention in the development of inductive mechanisms. To do so, novel categories were introduced as a means to test how adults and children attend to entities during category learning, how they represent those categories for purposes of generalization, and how well they discriminate between entities encoded to memory.

In Experiment 1, adults focused on a single rule feature, resulting in a constricted, knowledge-based representation as evidenced by high proportions of rule-based responses to Switch items and low memory discrimination. However, in Experiment 2 adults' attention was broadened to the overall appearance, thus their representation included more detail as evidenced by low proportions of rule-based responses to Switch items and greater memory discrimination.

Importantly, adults' responses to Switch items and their memory scores in Experiments 1 and 2 correspond with the instructions and feedback provided during Category Training. Low memory in Experiment 1 was not due to more efficient induction, but rather a result of training. And, while children had high proportions of rule-based responses to Switch items in Experiment 1, their memory discrimination was higher than that of adults who were provided the same training on the category inclusion rule. It is clear that children followed the category rule when categorizing but seems very likely that their attention remained distributed across multiple features of the stimuli. As such, children formed similarity-based representations during category learning, used a similarity-based mechanism during induction, and had higher memory discrimination in the recognition test.

Together, these findings demonstrate that adults can use either a knowledge- or similarity-based mechanism of induction, while children use only a similarity-based mechanism. The question being addressed here was *what* underlies the development of the mature knowledge-based mechanism? Results presented here implicate attention as the catalyst of representational and mechanistic change.

That is, as children develop the ability to attend selectively to relevant visual input and filter out the irrelevant (Enns & Cameron, 1987), the kinds of representations formed during category learning and the inductive mechanisms used develop in kind. In sum, the inductive mechanism is directly related to the pattern of attention during category learning and the subsequent representation formed.

Acknowledgments

This research was supported by a grant from the NSF (BCS-0720135) to V. M. Sloutsky.

References

Badger, J. R.,& Shapiro, L. R. (2012). Evidence of a transition from perceptual to category induction in 3- to 9-year-old children. *Journal of Experimental Child Psychology*, 113, 131-146.

Baldwin, D. A., Markman, E. M.,& Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64*, 711-728.

Best, C. A., Yim, H., &Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, 116, 105-119.

Bulloch, M. J.,& Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science*, 12, 114-122.

Craik, F. I. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.

Craik F. I. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294.

Deng, W.,& Sloutsky, V. M. (2012).Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, 23, 178-186.

Deng, W.,& Sloutsky, V. M. (2013). The role of linguistic labels in inductive generalization. *Journal of Experimental Child Psychology*, 114, 432-455.

Deng W., & Sloutsky, V. M. (2015). The development of categorization: Effects of classification and inference training on category representation. *Developmental Psychology*, 51, 392-405.

Enns, J. T., & Cameron, S. (1987). Selective attention in young children: The relations between visual search, filtering, and priming. *Journal of Experimental Child Psychology*, 44, 338-63.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development, 76*, 583-597.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65-95.

Gelman, R. (2004). Cognitive Development. *Stevens' Handbook of Experimental Psychology.*

Gelman, S. A.,& Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology*, 66, 327-353.

Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10*, 489-493.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Gelman, S. A.,& O'Reilly, A. W. (1998). Children's inductive inferences within superordinate categories: The role of language and category structure. *Child Development*, 59, 876-887.

Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, 75, 409-427.

Hoffman, A. B.,& Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology General*, 139, 319-340.

Koutstaal, W., & Schacter, D. L. (1997).Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language, 37*, 555-583.

Mandler, J. M., & McDonough, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology*, 37, 60-96.

Marks, W. (1991). Effects of encoding the perceptual features of pictures on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 566-577.

McDonough, L.,& Mandler, J. M. (1998). Inductive generalization in 9- and 11-month-olds. *Developmental Science*, 1, 227-232.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Rhodes, M. G., & Anastasi, J. S. (2000). The effects of a levels-of-processing manipulation on false recall. *Psychonomic Bulletin & Review, 7*, 158-162.

Sloutsky, V. M., & Fisher A. V. (2004a). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133*, 166-188.

Sloutsky, V. M., & Fisher A. V. (2004b). When Development and learning decrease memory. *Psychological Science, 15*, 553-558.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory and Cognition, 29*, 424-432.

Predicting Decision in Human-Agent Negotiation using functional MRI

Eunkyung Kim, Sarah I. Gimbel, Aleksandra Litvinova, Jonas T. Kaplan and Morteza Dehghani
{eunkyung; sgimbel; alitvino; jtkaplan; mdehghan}@usc.edu
University of Southern California, Los Angeles, CA 90089 USA

Abstract

The importance of human-agent negotiation, and the role of emotion in such negotiations, have been emphasized in human-agent interaction research. Thus far, studies have focused on behavioral effects, rather than examining the neural underpinnings of different behaviors shown in human-agent interactions. Here, we used a multi-round negotiation platform, instead of the more common single-shot negotiation, and were able to find distinct brain patterns in emotion-related regions of the brain during different types of offers. Using multi-voxel pattern analysis to analyze brain imaging data acquired during functional MRI scanning, we show that it is possible to predict whether the negotiator concedes, does not change, or asks for more during the negotiation. Most importantly, we demonstrate that left dorsal anterior insula, which is known to be an emotion-related brain region, shows a different pattern of activity for each of the three offer types.

Keywords: Decision-making; Negotiation; fMRI

Introduction

We encounter artificial agents frequently in our daily lives. From movie ticket vending machines to rescue robots in a disaster, more and more artificial agents are interacting with people. As interactions between humans and agents become common, effort has been made to make these interactions more effective. For example, relevant lines of research have investigated how to make users more engaged in the interactions with agents. Moreno, Mayer, Spires, and Lester (2001) showed that interactive pedagogical agents who communicate with students via speech promote meaningful learning in multimedia lessons, and agent's visual presence itself does not provide any cognitive advantage for students' learning. Castellano, Pereira, Leite, Paiva, and McOwan (2009) demonstrated that social interaction-based features, such as user behavior and contextual information, predict user engagement with a robot companion better than non-verbal behavior features. There has also been research into physiological responses to artificial agents, for example examining heart rate or stress detection during human-agent interaction. Nenonen et al. (2007) showed that a user's gaming experience was enhanced when the user's heart rate was used to control an interactive game. Zhai, Barreto, Chin, and Li (2005) also reported that users' physiological signals, such as pupil diameters, could be used to detect changes in stress level while interacting with a computer. The goal of these studies is to achieve the most effective human-agent interaction, and these approaches have expanded to various branches of human-agent interaction research.

Human-agent negotiation is among the most studied branches of human-agent interaction. Several studies have attempted to simulate human behavior in human-agent negotiation. For example, Byde, Yearworth, Chen, and Bartolini (2003) designed a system that automates price negotiation between buyers and sellers, using a belief function that estimates the distribution of prices and is updated by prior negotiations. The system passed the limited version of the Turing test; human negotiators could not distinguish between a software agent and a human negotiator. Lin, Kraus, Wilkenfeld, and Barry (2008) developed an agent that is capable of negotiating with human counterparts under conditions of incomplete information. Instead of a traditional quantitative decision-making model, a new model that learns information about the counterpart was used to build the agent so that it can achieve better negotiation agreements than human counterparts. These systems have shown the possibility of artificial agents that can reach better negotiation agreements with humans than human negotiators.

In order to make human-agent negotiations more effective, it is important to consider the role of emotions in human-agent interaction. There is now considerable amount of research showing that various emotions expressed by agents can either enhance or hinder interactions (Beale & Creed, 2009; Kim, Dehghani, Kim, Carnevale, & Gratch, 2014). One example of enhanced interactions is demonstrated by Brave, Nass, and Hutchinson (2005), who showed that agents showing emphatic emotion (e.g., showing a happy face when the user won the game and showing a sad face when the user lost the game) resulted in greater positive responses from users compared to when agents showed self-oriented emotion. Also, Van Kleef, De Dreu, and Manstead (2004) showed that people tend to concede more to an angry counterpart than to a happy counterpart.

The majority of previous research work has focused on replicating human-human negotiation results, or achieving better outcomes in human-agent negotiations. We believe that by examining the neural factors underlying behavior, we could advance our understanding about the cognitive processes involved in human-agent negotiations. To examine the neural underpinnings of human-agent negotiations, it is important to understand the functional organization and neurological processes in the human brain involved in such interactions. The use of functional Magnetic Resonance Imaging (fMRI) allows the investigation of the neural substrates of these human-agent negotiations.

Previous fMRI studies have identified specific emotion-related brain regions that play a significant role in social functions: the insular cortex (Damasio et al., 2000; Ruiz et al., 2013) and the amygdala (LeDoux, 2003). Increased insula activation was reported when cooperators see their partner defect (Rilling, Dagenais, Goldsmith, Glenn, & Pagnoni, 2008) or when people are presented with unfair offers (Sanfey,

Rilling, Aronson, Nystrom, & Cohen, 2003). Haruno and Frith (2010) showed that fairness of the negotiation outcome can be predicted based on patterns of activity in amygdala. Activity in these brain regions is believed to affect human behaviors during various types of negotiations, as emotion plays a key role in social interactions that are vital during negotiations (Hess & Bourgeois, 2010). For example, it has been shown that getting an unfair offer or rejecting an offer in the Ultimatum game triggers negative emotions, such as anger (Pillutla & Murnighan, 1996). However, these neurological studies were mostly limited to single-shot negotiations such as the Ultimatum game (Sanfey et al., 2003).

In this study, we build on and extend this line of research by studying the relationship between brain patterns and decision-making during general multi-round human-agent negotiations, which are more similar to real-world negotiations than single-shot negotiations. Investigating brain activity during general multi-round human-agent negotiation is vital to understanding how humans react in interactions with agents. Moreover, investigating the neural systems that are active during different types of interactions would allow us to better understand the underlying cognitive processes responsible for the resultant behavior. As far as we know, our work is the first study that uses brain imaging data to predict the course of general multi-round human-agent negotiations.

We hypothesize that activity in emotion-related brain regions plays a key role in decision-making paradigms during human-agent negotiation. This is supported by previous studies showing that emotion plays an important role in decision-making (Hegtvedt & Killian, 1999; Loewenstein, Weber, Hsee, & Welch, 2001). We hypothesize that emotions arising during human-agent interaction are an indicator of what type of decision the person is going to make, potentially predicting one's negotiation behavior.

The paper is structured as follows: First, we introduce the negotiation task we used in our fMRI experiment. Next, we describe our experimental design and the procedure of the experiment. Then, we explain the methods we used for analyses. Finally, we discuss our results and future work.

Objects Negotiation Task

Dehghani, Carnevale, and Gratch (2014) introduced the Objects Negotiation Task, a web-based multi-round negotiation task where a participant and a computer agent can make a proposal in turn to negotiate about the distribution of different types of items. Because this task was originally designed for behavioral studies, some changes were made to optimize the task for use in the fMRI scanner. The version of the Objects Negotiation Task used in this fMRI experiment is composed of six phases (figure 1).

1. The participant proposes an offer.
2. The participant waits for the computer agent to accept or reject the offer.
3. If accepted, negotiation ends. If rejected, the participant waits for the computer agent to propose an offer.
4. The participant reviews the computer agent's offer for five seconds.
5. The participant decides whether to accept or reject the computer agent's offer.
6. If accepted, negotiation ends. If rejected, the participant waits for five seconds and then is redirected to the first phase.

Each negotiation can be as long as six rounds. On the sixth round, a message is displayed to the participant to indicate that it is the last round. The participant then can either accept the computer agent's offer or toss a coin. If a participant chooses to toss a coin, he/she gets two of each item if the coin lands on heads or nothing if the coin lands on tails.

Experiment

Participants

Ten participants (seven female), recruited from the online campus bulletin board at the University of Southern California, took part in the experiment. Participants' mean age was 27.5 years ($\pm$ 4.93) and all participants had normal or corrected to normal vision. Each participant was provided with a written informed consent according to the guidelines of the USC Institutional Review Board. All participants were screened to rule out medication use, head trauma, history of neurological disorders, and other serious medical conditions.

Procedure

Participants were instructed to read the hypothetical scenario shown below.

> You are a restaurant owner in a small town. There has been a major fire in the market providing the necessary fruits for your restaurant and as a result only a limited number of fruits are available. Because of this you have to split the available fruits with another restaurant owner. You and the other owner value each fruit differently. In order to run your restaurant you need to get as many fruits as possible.
>
> In the task that follows, you will negotiate about how to distribute the fruits between you and the other restaurant owner.

Participants learned the rules of the Objects Negotiation Task and performed a practice task. The practice negotiation task had an identical interface to the task performed in the scanner, but with a different set of negotiation items and payoffs. During the practice negotiation task, participants were provided with a trackball mouse similar to the one used in the scanner in order to get used to operating it and moving items around the task interface. After completing the practice negotiation task, participants were checked with a metal detector to ensure that they were safe to go inside the fMRI scanner.

Participants then performed a total of six negotiations in the scanner, each including up to six rounds. Different sets

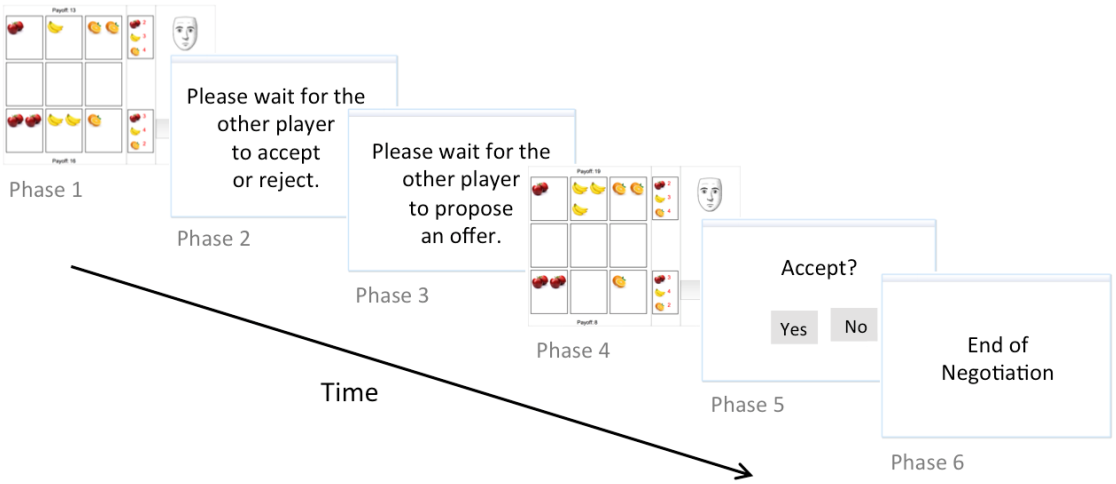


Figure 1: Timeline of the Objects Negotiation Task used in the fMRI experiment. Here, we show the timeline of an example in which the computer agent rejected the participant's offer (phase 3) and then the participant accepted the agent's offer (phase 6).

of negotiation items were used in each negotiation. After the fMRI scan, the participants received a short survey to acquire background information including gender, age, and handedness. All participants were paid $30 for their participation.

Participants were not told that they would be playing with artificial agents. To simulate playing against other humans, we added randomized delays when the computer agent proposed an offer.

Analysis

In our analyses, we aimed to find the best predictor of participants' offer types based on their brain activities. To achieve this goal, we labeled our data first, and then performed general linear model (GLM) analysis. Lastly, we used the GLM analysis results as input to multi-voxel pattern analyses (MVPA) using two types of feature selection methods.

Data Labeling

All of the participants' offers were labeled with one of three categories based on their payoff changes from the previous offer: (1) positively-changed offer, (2) not-changed offer, and (3) negatively-changed offer. Participants' offers on the first rounds of each negotiation were labeled 'not-changed offer,' and their offers on all other rounds were compared with their offer on the previous round based on payoffs and then labeled according to the comparison. For example, if the participant's payoff in the first round was 20, his/her payoff in the second round was 23, and in the third round 21, the first round would be labeled as a 'not-changed offer,' the second round as a 'positively-changed offer' (because his/her payoff was increased by 3), and the third round as a 'negatively-changed offer' (because his/her payoff was decreased by 2). Here, 'positively-changed offer' means that the participant concedes more to the other player, 'not-changed offer' means that the participant holds, i.e., he/she retains the offer on the previous round, and 'negatively-changed offer' means that the participant asks for more.

Before analyzing the fMRI data, we had to exclude data from three participants. The first exclusion was due to incomplete fMRI scan data. In the second case the participant made less than five offers in each offer category, leading to a too small number of offers in each offer category. In the third case we excluded the data because of the participant's left-handedness, since more than 70% of left-handed people have different functional brain structure compared to right-handed people such as for language processing (Warrington & Pratt, 1973).

General Linear Model Analysis

To compare brain activity between the offer-making period and the non-offer-making period, we ran a general linear model (GLM) analysis using tools from the FMRIB's Software Library (FSL) (Smith et al., 2004). Data pre-processing for GLM analysis included following steps. First, all participants' fMRI data were motion-corrected using FSL's MCFLIRT tool to fix head motion artifacts during scans. Then, non-brain such as a scalp was removed from the data using FSL's Brain Extraction Tool. Next, spatial smoothing using a 5mm full width at half maximum Gaussian kernel was applied to increase statistical power by improving the signal to noise ratio. Also, slice timing correction for interleaved acquisitions was used to compensate for timing difference between slices of functional images. Finally, high-pass temporal filtering was performed to let high frequencies (containing activities relevant to decision-making) pass and to remove low frequencies such as signal drifts.

After completing data pre-processing, we modeled brain activity during offer-making with a double gamma hemodynamic response function. Brain activity during all other time points were considered baseline.

Multi-Voxel Pattern Analysis

To analyze brain patterns in decision-making during negotiations, we used multi-voxel pattern analysis (MVPA) (Norman, Polyn, Detre, & Haxby, 2006), a machine learning approach for investigating patterns of brain voxels. Instead of analyzing each voxel separately, MVPA takes multiple voxels into account together. This is useful because activity in one voxel cannot be separated from neighboring voxels.

We used GLM analysis results as input to MVPA. As data pre-processing steps for MVPA, we de-trended the data to remove any bias resulting from scanner drift over the acquisition time. Then, we converted the data to z-scores to normalize the range of each voxel. We used leave-one-participant-out cross validation for MVPA, in which a classifier is trained on six participants' data and then tested with the last participant's data. We repeated this seven times, leaving each participant out once, then averaged the results to calculate prediction accuracy. A balancer was also included to keep the chance level the same (33%) throughout our analyses because every participant has a different set of offers. Since a balancer chooses a new set of offers in each category whenever it runs, we ran MVPA five times and averaged the results.

Lastly, we applied feature selection methods for choosing the voxels used in our analysis. Feature selection is a common approach to reduce the number of features (voxels) by selecting only relevant features as input to a classifier. Classification performance improves with feature selection because it picks features that vary significantly between categories (Guyon & Elisseeff, 2003). To validate our hypothesis that brain activities on emotion-related regions are closely related to decision-making during negotiations, we used two different feature selection methods for MVPA. For both approaches, we used a linear Support Vector Machine (SVM) classifier to perform classification. In the following sections, we explain each the two feature selection method we used for MVPA: Region of Interest and Searchlight analysis.

Region of Interest Analysis: Insular Cortex In the region of interest (ROI) analysis, we attempted to predict the offer categories based only on the voxels in the insular cortex, which is known as an emotion-related brain region. The insular cortex on each side of the brain can be divided into three subregions with distinct patterns of connectivity: dorsal anterior insula, connected with dorsal anterior cingulate cortex; ventral anterior insula, connected with pregenual anterior cingulate cortex; and posterior insula, connected with primary and secondary somatomotor cortices (Deen, Pitskel, & Pelphrey, 2011).

We trained the classifier using voxels from each of the six regions separately with feature selection. In our analyses, we used GLM analysis results to compute F-score per each voxel, and then used an analysis of variance measure to select the top 5% of features with the highest F-scores. Each participant's brain was first transformed into standard MNI space (Evans et al., 1993) to minimize differences from individual brains. After performing this process for all participants, individual-level analyses were combined for a group-level analysis.

Searchlight as a Feature Selection Method In the searchlight analysis (Kriegeskorte, Goebel, & Bandettini, 2006), a map of classification accuracies is generated by measuring the information in small spheres (radius = 5 voxels) centered on every voxel in brain. As activities on one voxel are inevitably influenced by activities on neighboring voxels, searchlight analysis is a preferable approach to capture local spatial areas that show significantly different activity patterns on each experimental condition. To confirm the strong correlation between brain activities on emotion-related regions and negotiation offer types, and to find other brain regions that show notable pattern differences on each offer type, we used searchlight analysis as a feature selection method for MVPA.

Steps for using searchlight analysis as a feature selection method are as follows. We first generated searchlight maps from each participant's brain. Then we transformed individual searchlight maps into the standard MNI space, and merged six of them to generate seven merged searchlight maps in total. Next, we generated t-maps using t-tests across six participants versus chance, and thresholded the top 5% of the t-maps and binarized them. After that, we transformed all participants' functional images (EPIs) into individual EPI space and ran MVPA using thresholded and binarized t-maps as masks for each participant. Finally, we averaged MVPA results to calculate overall prediction accuracy.

RESULTS

ROI MVPA: Left Dorsal Anterior Insula

We hypothesized that activity in the insular cortex would be highly correlated with different types of offers. We performed ROI MVPA with each of the six insula regions (left/ right ventral anterior insula, left/right dorsal anterior insula, and left/right posterior insula), and found that left dorsal anterior insula to be the best predictor of offer types (positively-changed offer, not-changed offer, or negatively-changed offer). The prediction accuracy of ROI MVPA using voxels from left dorsal anterior insula is 43.88%, with a standard error 1.30%. Interestingly, insular regions other than left dorsal anterior insula show chance level performance.

A binomial test shows that the performance of ROI MVPA with left dorsal anterior insula is significantly higher than chance level ($p = 0.0058$), indicating that activity in left dorsal anterior insula predicts offer types in negotiations.

MVPA with Searchlight as a Feature Selection Method

To map the spatial distribution of information, we performed MVPA with searchlight as a feature selection method. Figure 2 shows the overlaid accuracy map for all seven participants. Before overlaying accuracy maps, we thresholded the top 5% of seven t-maps each across six participants and binarized them. Consistent with the finding from the ROI analysis, we find that left dorsal anterior insula is included in six t-maps.

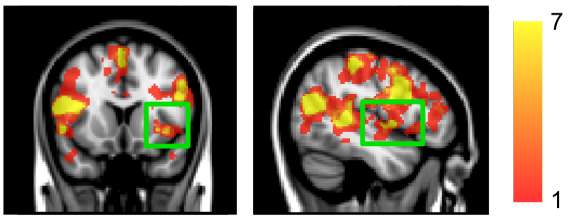


Figure 2: Combined t-maps that were used as masks for MVPA with searchlight as a feature selection method. Each of seven t-maps were generated across six participants' searchlight results, and then thresholded and binarized. Areas marked with red indicate that the area is included for one t-map, and areas with yellow indicate the area is included for all t-maps. Left dorsal anterior insula is included in six t-maps (green boxes).

The prediction accuracy for MVPA with searchlight as a feature selection method is 47.45% with a standard error 0.017%. Compared to ROI MVPA, MVPA with searchlight as a feature selection method has the prediction accuracy improved by 3.57% and has much smaller standard error. Two sample t-test results revealed that there is a significant difference in prediction accuracy between MVPA with searchlight as a feature selection method and ROI MVPA with left dorsal anterior insula, $t(23.124) = 1.6272$, $p = 0.058$.

DISCUSSION

In this experiment, we investigated the neural correlates of decision-making during negotiations. We classified negotiation offers into three categories: conceding, holding, or asking for more. We then analyzed brain imaging data by offer category to see if we could predict the category of negotiation based on the brain activation. We used functional MRI to capture participants' brain activity during negotiations, and used MVPA to analyze the fMRI data. ROI MVPA and MVPA with searchlight as feature selection methods were used and both methods resulted in significantly better prediction accuracies than the chance level of 33%. Most notably, MVPA with searchlight as a feature selection method yielded the higher prediction accuracy of 47.45%, indicating the importance of analyzing neighboring voxel clusters together.

Our results reveal that there are distinct brain patterns across participants for each type of offer. More specifically, activity in left dorsal anterior insula, which is a well-known emotion-related brain region, was found to play a key role in distinguishing offer types. This is in line with our hypothesis that activations in emotion-related brain regions would be closely related to decision-making. Emotions provoked while the participant is interacting with the computer agent during negotiations mediate the participant's negotiation behaviors, and these processes are captured in fMRI data as a form of increased blood flow in emotion-related brain regions. Thus, this confirms not only the importance of the role of emotion in human-agent interaction, but also the possibility of interpreting the underlying processes during negotiations with fMRI data.

Furthermore, the results indicate that we can predict negotiation offers based on brain activities. Offer type prediction results from MVPAs with two types of feature selection methods support the feasibility of successful predictions of negotiation behaviors. This has implications for human-agent negotiation research, allowing us to perform more detailed predictions of negotiation behavior based on fMRI data as compared to predictions based on behavioral research. For example, we can conduct an fMRI study on how engaged participants are during negotiations under various conditions. In addition to non-verbal features that were found to be related to user engagement by Castellano et al. (2009), introducing brain imaging data would give us extra information on degree of user engagement. With these approaches, we can come close to achieving the most effective human-agent interaction.

Also, our results suggest that there are more brain regions correlated with negotiation offer types. For instance, right posterior supramarginal gyrus and right hippocampus were included in all t-maps that were generated with searchlight maps. Generally, decision-making during negotiations are affected by one's empathy towards others (Loewenstein & Lerner, 2003), which is related to brain activities on right supramarginal gyrus, and remembering previous offers to update one's knowledge (Zeng & Sycara, 1998), which is related to brain activities on right hippocampus. Thus, these findings would allow us to track down more details of decision-making during negotiations.

In this study, we looked at only one previous round of negotiation, but it is possible that negotiators actually take into account the whole history of negotiation when proposing the next offer. In that case, studying a longer history would result in better prediction of negotiators' offers. As a first approach studying general multi-round human-agent negotiations, we were able to show that we can do offer type prediction during negotiations with one previous round using fMRI data. In future studies, we plan to take longer history into consideration. Also, we plan to replicate our results using larger sample sizes. Even though we had only seven participants in this study, we expect to see the same results in a large number of participants because we were able to do a significantly better offer type prediction than the chance level with large effect sizes. In addition, we would like to note that the probability of replication is dependent not on the sample size but on the p value (Killeen, 2005).

In conclusion, we examined the relationship between brain patterns and types of negotiation offers. Unlike previous fMRI studies that were limited to single-shot negotiations, we studied general multi-round human-agent negotiations. Our results demonstrate that we can predict behaviors from multi-round negotiations using fMRI data. We believe this finding could help enrich human-agent negotiation, and hope to find more work on this topic.

Acknowledgments

This research is supported by AFOSR FA9550-14-1-0364.

References

Beale, R., & Creed, C. (2009). Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, *67*(9).

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, *62*(2).

Byde, A., Yearworth, M., Chen, K.-Y., & Bartolini, C. (2003). Autona: A system for automated multiple 1-1 negotiation. In *Ieee international conference on e-commerce, 2003.*

Castellano, G., Pereira, A., Leite, I., Paiva, A., & McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on multimodal interfaces.*

Damasio, A. R., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L., Parvizi, J., & Hichwa, R. D. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature neuroscience*, *3*(10).

Deen, B., Pitskel, N. B., & Pelphrey, K. A. (2011). Three systems of insular functional connectivity identified with cluster analysis. *Cerebral Cortex*, *21*(7).

Dehghani, M., Carnevale, P. J., & Gratch, J. (2014). Interpersonal effects of expressed anger and sorrow in morally charged negotiation. *Judgment and Decision Making*, *9*(2).

Evans, A. C., Collins, D. L., Mills, S., Brown, E., Kelly, R., & Peters, T. M. (1993). 3d statistical neuroanatomical models from 305 mri volumes. In *Nuclear science symposium and medical imaging conference, 1993.*

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*.

Haruno, M., & Frith, C. D. (2010). Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nature neuroscience*, *13*(2).

Hegtvedt, K. A., & Killian, C. (1999). Fairness and emotions: Reactions to the process and outcomes of negotiations. *Social Forces*, *78*(1).

Hess, U., & Bourgeois, P. (2010). You smile–i smile: Emotion expression in social interaction. *Biological psychology*, *84*(3).

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological science*, *16*(5).

Kim, E., Dehghani, M., Kim, Y. K., Carnevale, P. J., & Gratch, J. (2014). Effects of moral concerns on negotiations. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10).

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology*, *23*(4-5).

Lin, R., Kraus, S., Wilkenfeld, J., & Barry, J. (2008). Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, *172*(6).

Loewenstein, G. F., & Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, *619*(642).

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological bulletin*, *127*(2).

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching. *Cognition and Instruction*, *19*(2), 177–213.

Nenonen, V., Lindblad, A., Häkkinen, V., Laitinen, T., Jouhtio, M., & Hämäläinen, P. (2007). Using heart rate to control an interactive game. In *Proceedings of the sigchi conference on human factors in computing systems.*

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, *10*(9).

Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational behavior and human decision processes*, *68*(3).

Rilling, J. K., Dagenais, J. E., Goldsmith, D. R., Glenn, A. L., & Pagnoni, G. (2008). Social cognitive neural networks during in-group and out-group interactions. *Neuroimage*, *41*(4).

Ruiz, S., Lee, S., Soekadar, S. R., Caria, A., Veit, R., Kircher, T., ... Sitaram, R. (2013). Acquired self-control of insula cortex modulates emotion recognition and brain network connectivity in schizophrenia. *Human brain mapping*, *34*(1).

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... others (2004). Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, *23*.

Van Kleef, G. A., De Dreu, C. K., & Manstead, A. S. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of personality and social psychology*, *86*(1).

Warrington, E. K., & Pratt, R. (1973). Language laterality in left-handers assessed by unilateral ect. *Neuropsychologia*, *11*(4).

Zeng, D., & Sycara, K. (1998). Bayesian learning in negotiation. *International Journal of Human-Computer Studies*, *48*(1).

Zhai, J., Barreto, A. B., Chin, C., & Li, C. (2005). Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. In *Southeastcon, 2005.*

Searching large hypothesis spaces by asking questions

Alexander N. Cohen (alxcoh@gmail.com)
Hunter College High School

Brenden M. Lake (brenden@nyu.edu)
Center for Data Science
New York University

Abstract

One way people deal with uncertainty is by asking questions. A showcase of this ability is the classic 20 questions game where a player asks questions in search of a secret object. Previous studies using variants of this task have found that people are effective question-askers according to normative Bayesian metrics such as expected information gain. However, so far, the studies amenable to mathematical modeling have used only small sets of possible hypotheses that were provided explicitly to participants, far from the unbounded hypothesis spaces people often grapple with. Here, we study how people evaluate the quality of questions in an unrestricted 20 Questions task. We present a Bayesian model that utilizes a large data set of object-question pairs and expected information gain to select questions. This model provides good predictions regarding people's preferences and outperforms simpler alternatives.

Keywords: Bayesian modeling; active learning

People seem to ask rich and probing questions when faced with uncertainty. Whether someone is learning a new task, meeting a new person, listening to a presentation, or attending a press conference, people ask questions to better understand the state of the world – a form of "active learning" (Gureckis & Markant, 2012). Question asking is associated with several computational challenges, including selecting a question from a possibly infinite set of allowable options, and evaluating its quality against a large hypothesis space of possible world states. The scope of these challenges raises several key questions: Do people ask good questions? And if so, how do people effectively search over large numbers of questions and hypotheses?

The classic game of 20 Questions (20Qs) provides a window into this broader human ability. A round of 20Qs is played between a "game-master" and a "question-asker." The game-master thinks of an object and the question-asker asks up to 20 questions before guessing the identity of the object. The game-master answers each question with either "yes" or "no," with additional options such as "probably" or "probably not" available in variants of the game. Over the course of playing a game, an ideal question-asker would consider thousands of possible objects in the hypothesis space and select questions from an infinite set of options. Moreover, an effective player must continually update the plausibility of the hypotheses with each new piece of information. How do people manage to play this game? And do they ask effective questions as measured by normative metrics?

Previous work has found that people are surprisingly effective question-askers in modified 20Qs-style games with a limited set of possible objects (Eimas, 1970; Denney & Denney, 1973; Thornton, 1982; Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2015). People generally prefer to ask "constraint-seeking" questions such as "Is it alive?" rather than more specific "hypothesis-scanning" questions such as "Is it a frog?," especially when faced with large amounts of uncertainty. This ability seems to develop in childhood (Eimas, 1970; Ruggeri et al., 2015), decline in elderly populations (Denney & Denney, 1973), and can be disrupted when the feature statistics of the game do not match the real world (Nelson et al., 2014). In most cases, the preference for constraint-seeking questions is consistent with maximizing stepwise Expected Information Gain (EIG), an information theoretic measure with theoretical and empirical motivations. EIG is a measure of informativeness under a Bayesian framework, where higher scoring questions provide a larger reduction in the posterior entropy (effectively, providing information about the correct object). People ask questions predicted by this metric when reasoning in small (Eimas, 1970; Nelson et al., 2014; Ruggeri et al., 2015) or highly visual (Gureckis & Markant, 2009; Rothe, Lake, & Gureckis, 2016) hypothesis spaces, but the generality of this account remains an open question.

In this paper, we study how people evaluate questions in a larger and more naturalistic 20Qs task, which includes many broad classes of common objects (e.g., animals, artifacts, household objects, foods, etc). Using a large set of objects and questions as a rough proxy for semantic knowledge, we propose a Bayesian model for computing posterior belief over object hypotheses and the EIG of candidate questions. Human question preferences in both "complete" and "one-shot" 20Qs games are compared with this model along with various alternatives, providing an important test of how the EIG metric generalizes to richer and more naturalistic domains.

Model of 20 Questions

A Bayesian framework is developed to reason and ask questions in a large hypothesis space of possible objects.

Data set of 1000 objects. We used a data set from Palatucci, Pomerleau, Hinton, and Mitchell (2009) of 1000 objects and 218 questions to construct the model. Objects span a variety of broad semantic classes including animals, insects, food, household items, tools, clothing, vehicles, sports, buildings, and other tangible nouns. It excludes specific people, specific places, proper nouns, ideas, verbs, etc. The set of questions concerns higher-level semantic categories (e.g., "Is it an animal?" or "Is it furniture?"), color ("Is it Yellow?"), shape and texture ("Is it long?," "Is it fuzzy?," "Is it bigger than a microwave oven?"), parts ("Does it have ears?"), actions ("Can it cause you pain?"), uses ("Can you play with it?"), common locations ("Does it live above ground?"), and emotions ("Does it make you happy?"). Each

of the 218,000 object-question pairs was evaluated on Amazon's Mechanical Turk using a five point scale from "definitely not" (coded as -1) to "definitely yes" (1). There were many different participants and each question was only answered once, necessitating the use of a noisy response model described in a subsequent section.

Bayesian framework. A Bayesian framework for 20Qs has been developed in prior work (Nelson et al., 2014; Ruggeri et al., 2015), and we extend this framework to model our large scale 20Qs task.

We use the following notation. The data set $\mathcal{D}$ contains 1000 objects $o \in O$ and 218 questions $f \in \mathcal{F}$ (which can also be viewed as "features"). The response for a particular object-question pair is a value on the five point scale $\mathcal{D}_{of} \in \mathcal{A} = \{-1, -0.5, 0, 0.5, 1\}$. During a specific game of 20Qs, the same questions and response scale as in $\mathcal{D}$ are used. The first question and its response are denoted $\{f_1, a_1\}$, the second question and response $\{f_2, a_2\}$, and so on where the information revealed so far is denoted as $\mathcal{K} = \{\{f_1, a_1\}, \{f_2, a_2\}, \ldots, \{f_n, a_n\}\}$ with $a_j \in \mathcal{A}$ and $f_j \in \mathcal{F}$. We also use the notation $\mathcal{K}_f$ and $\mathcal{K}_a$ to separately indicate the set of questions $\mathcal{K}_f = \{f_1, f_2, ..., f_n\}$ and their corresponding answers $\mathcal{K}_a = \{a_1, a_2, ..., a_n\}$.

During a game of 20Qs, Bayes' rule can be used to reason about the probability of each object given the questions and their responses so far,

$$P(o|\mathcal{K}_a; \mathcal{K}_f) = \frac{P(o)\prod_{j=1}^{n} P(a_j|o; f_j)}{\sum_{o' \in O} P(o')\prod_{j=1}^{n} P(a_j|o'; f_j)}. \quad (1)$$

For simplicity and for consistency with the behavioral study, a flat prior is used over the objects $P(o) = 1/1000$. Ideally, the likelihood $P(a_j|o; f_j)$ would be known and modeled separately for each feature-object pair to capture variability across different people in how they answer the same question. However this would require having many responses for each of the 218,000 object-question pairs, for which only one response is provided and collecting many more is unfeasible. Instead, we ran a separate Mechanical Turk experiment to estimate a shared noise model to use as proxy, described in a later section. The semicolon notation indicates that $\mathcal{K}_f$ is a parameter rather than random variables like $\mathcal{K}_a$ and o.

Expected Information Gain (EIG). In conjunction with the Bayesian model, EIG was used as the metric for deciding what question to ask next. The goal is to ask the question f_{n+1} that maximizes the expected reduction in uncertainty (measured as Shannon entropy $H[\cdot]$ in the posterior distribution). The expectation is a weighted average over all possible answers a_{n+1}, such that

$$EIG(f_{n+1}) = \sum_{a_{n+1} \in \mathcal{A}} P(a_{n+1}|\mathcal{K}_a; f_{n+1}) \Big[H[P(o|\mathcal{K}_a)] - H[P(o|a_{n+1}, \mathcal{K}_a; f_{n+1})] \Big]. \quad (2)$$

Note that we dropped an implicit dependence on $\mathcal{K}_f$ in each distribution. The posterior predictive distribution is,

$$P(a_{n+1}|\mathcal{K}_a; f_{n+1}, \mathcal{K}_f) = \sum_{o \in O} P(a_{n+1}|o; f_{n+1}) P(o|\mathcal{K}_a; \mathcal{K}_f).$$

Likelihood model and response noise. The likelihood $P(a|o; f)$ from Eq. 1 (dropping the subscript j for convenience) requires modeling variability in how different people answer (a) the same question (f), in relation to a particular object-question pairing. This is needed because the gamemaster (as captured by the data set $\mathcal{D}$) may not entirely agree with the question-asker (a human participant or the Bayesian model) on how to answer a question. For instance, people may answer ambiguous questions differently, such as the pairing of "dog" and "Is it bigger than a loaf of bread?"

A Mechanical Turk experiment was used to fit a response model by querying a subset of object-question pairs multiple times with different participants. The results were used to fit the model of response noise, estimated using a separate Bayesian analysis. Assume that the response model for any cell $\mathcal{D}_{of} \in \mathcal{A}$ is a unknown multinomial distribution h_{of} over the five point scale. Let $h_{of} \in \mathcal{H}$ denote the possible set of multinomial distributions, where $\mathcal{H}$ is approximated by the empirical set of multinomials collected from Mechanical Turk. We assume that multinomials for new cells are drawn from the set of existing cells. The likelihood model for a question response then becomes

$$\begin{aligned} P(a|o; f) &= P(a|\mathcal{D}_{of}) \\ &= \sum_{h' \in \mathcal{H}} P(a|h_{of} = h') P(h_{of} = h'|\mathcal{D}_{of}) \\ &= \sum_{h' \in \mathcal{H}} P(a|h_{of} = h') \frac{P(\mathcal{D}_{of}|h_{of} = h') P(h_{of} = h')}{P(\mathcal{D}_{of})} \end{aligned}$$

which marginalizes over the uncertainty regarding the latent multinomial, given just a single sample from that multinomial $\mathcal{D}_{of}$. Both terms $P(a|h_{of} = h')$ and $P(\mathcal{D}_{of}|h_{of} = h')$ are just the probability of that response given the multinomial distribution represented by h'. A uniform prior $P(h_{of} = h')$ is used across the possible multinomials $h' \in \mathcal{H}$.

The set of multinomials $\mathcal{H}$ collected from Mechanical Turk consisted of 500 unique feature-question pairs. Fifty participants in the USA were asked to answer 100 questions randomly chosen from this set of 500 pairs, resulting in approximately 10 observations per cell. The corresponding set of 500 empirical multinomials was used to create $\mathcal{H}$. The fitted likelihood model is shown in Table 1.

Table 1: Fitted response model $P(a|\mathcal{D}_{of})$.

		$\mathcal{D}_{of}$				
		-1.0	-0.5	0.0	0.5	1.0
a	-1.0	0.704	0.311	0.185	0.126	0.062
	-0.5	0.098	0.272	0.129	0.082	0.039
	0.0	0.097	0.214	0.380	0.201	0.099
	0.5	0.056	0.117	0.172	0.313	0.176
	1.0	0.044	0.086	0.133	0.278	0.623

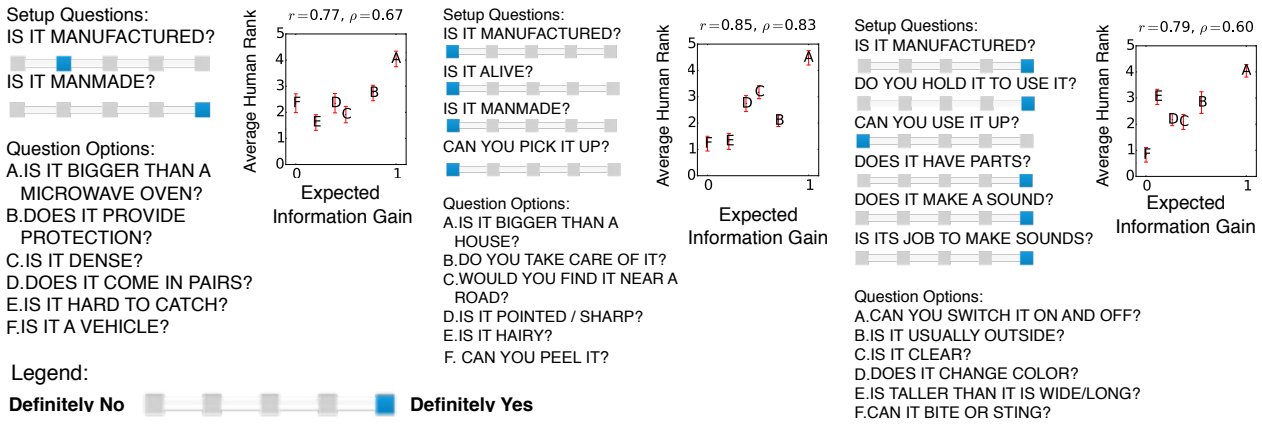


Figure 1: Three examples of one-shot 20 question games. Participants observed the Setup Questions with answers on a five point scale (see legend). They then ranked the Question Options in order of preference. In the scatter plots, the average human rank (error bars show ± 1 s.e.) is compared with Bayesian expected information gain, which was normalized to fall between 0 and 1. The object chosen by the game-master for each game were Kitchen, Ground, and Tuba (from left to right).

Experiment

A behavioral experiment was run on Amazon's Mechanical Turk using Psiturk (Gureckis et al., 2015) to compare human question asking with the Bayesian model and alternative models. Data was collected from 25 participants that reside in the USA, providing a base pay of $6 with an additional performance-based bonus. Two participants were not analyzed due to technical difficulties.

Participants acted as the question-asker while the computer plays the role of the game-master. Participants played two forms of 20Qs, including first a set of "complete games" and second a set of "one-shot games," each with a massive hypothesis space (any object from the set of 1000) and a fixed set of question choices. Complete games involved full games of 20Qs where each participant played unique games. One-shot games involved introducing participants to partially-finished 20Qs games (with some questions already answered) and then asking participants for their preferences regarding the next question (see examples in Fig. 1). Unlike the complete games, the one-shot games and associated question options were identical for each participant.

A number of checks were put in place to ensure that participants were engaged in the experiment. Before completing any of the full games, participants were asked to play a tutorial full game of 20Qs with artificial visual objects with four binary features. Participants were also given a short quiz after each segment of instructions, and correct answers were needed to proceed to the next section. This included instructions meant to convey the semantic space of possible objects (similar to the description in the section "Data set of 1000 objects"), although the precise set of 1000 objects remained unknown to participants. Participants were also told they should expect noise – that is, the computer responses are based on other people's answers (as usual in 20Qs) and thus they should expect occasional disagreements.

Complete games. Participants played five full games of 20Qs: one practice game and four real games. In each game, an object was randomly chosen by the computer from the set of 1000. At each step, participants were shown six questions they could ask (in random order), selected to evenly span a range of EIG values according to the Bayesian model.[1] Participants chose a question and then received the answer, after which they received a new set of questions to choose from, etc. Before receiving the answer choices at each step, they were also asked to type out their ideal question (which wasn't analyzed). In order to mitigate the possibility of multiple exposures to the same question, questions were presented as an option only once per full game. Also, the design discouraged multi-step planning strategies, as future question options were always unknown.

At the completion of the game, participants were asked to "guess" the identity of the hidden object. To make the task feasible, rather than selecting the correct object from the entire set of 1000 possibilities, people were shown a set of 20 objects that included the correct answer and 19 randomly chosen distractors. The 20 options were not shown until the game was completed, meaning no more questions could be asked (although the previous questions and answers could still be viewed). Participants were rewarded with a bonus for choosing the correct object from the set.

The potential bonus started at $0.50 and was decremented by $0.05 for each question they asked. Participants could choose to enter the guessing phase at any time instead of selecting another question. After nine questions the guessing phase began automatically. After guessing, they were shown the correct object.

[1] Possible questions were ranked by the model from best (1) to worst (218). Six questions were chosen based on their position in the ranked list (1, 37, 73, 109, 145, 181). If a question was presented previously as a possible choice, it was excluded from the ranking.

One-shot games. After the full games, participants completed ten one-shot games of 20Qs. Each of these games was pre-generated and the same games were shown to each participant in random order. These one-shot games followed the same structure as the full games, except that participants were incrementally shown a number of questions and answers, as if someone else was playing the game for them. To make sure they were following along, they were asked to reconstruct the answers to two questions as a quiz before continuing to the next stage.

After the quiz, participants were asked to choose between six questions (see Fig. 1). Rather than selecting only the best question as in the full games, they were asked to rank the questions in order of preference. After ranking, the game was over, the top question was answered, and an analogous guessing phase occurred. A bonus of $0.20 was rewarded for correctly guessing the object.

The one-shot games varied in depth: the number of previous questions shown before participants ask a new question. There was 1 trial of depth 0 (participants chose the first question), 3 trials of depth 2, 3 trials of depth 4, and 3 trials of depth 6 (see Fig. 1 for examples). Deeper games were not included, since we expected people would have trouble processing that much information on each trial. The one-shot games were generated by recursively querying the model. Each question option appeared only once across the set of one-shot games.

Alternative Models

In addition to the Bayesian model that maximizes EIG (henceforth "full Bayesian model" or "EIG"), we evaluated a range of alternatives.

Expected Utility (EU). The EU model uses the full Bayesian machinery to choose the question that maximizes expected "reward gain" rather than "information gain," metrics that often make similar predictions (Markant & Gureckis, 2012). EU maximizes the expected bonus in the guessing phase, assuming it occurs immediately after the next question is answered. EU is defined like EIG except that the entropy function $H[\cdot]$ (Eq. 2) is replaced by a function that computes the expected bonus (and multiplied by -1). In the bonus phase, we assume the EU model selects the object (of 20) with highest posterior probability. The probability that its choice is correct (according to model belief), marginalizing over all possible correct objects and all possible sets of random distractors, can be computed exactly with combinatorics.

Context Insensitive (CI). The CI model is a simplified Bayesian model that ignores the current game state (previous questions and answers). It chooses questions to maximize EIG while assuming the current knowledge state is the prior (uniform). The CI model tests whether people ask context-sensitive questions or just choose questions based on a set of pre-computed "good" questions.

Random Subset (RS). The RS model is a simplified Bayesian model that does not consider the whole hypothesis space. Instead, RS considers only K naively chosen hypotheses, related to the 20Qs model of Navarro and Perfors (2011). Rather than full Bayesian updating (Eq. 1), all but K randomly chosen objects are assigned zero prior probability (and thus zero posterior probability). As with the full Bayesian model, questioned are selected to maximize EIG but with this simplified posterior. A simulated experiment with RS used 25 simulated participants, each with a different random subset of the hypotheses (also re-sampled for each trial). We ran a total of 100 simulated experiments, and the reported statistics (e.g., correlation coefficients) were computed and then averaged over simulated experiments. Different values of K were explored where increasing K converges to the full model.

Familiarity. Although steps were taken to mitigate question repetition in the behavioral experiment, including using unique question options within each full game and not repeating options across one-shot games, a potential design issue is that some questions were seen by participants more than once (either as choices or previously revealed information). For instance, the high-value early question "Is it manufactured?" appeared many times as revealed information (Fig. 1). The Familiarity model attempts to explain the value of a question by the number of times people previously saw it.

Results and Discussion

Analysis of one-shot games. We analyzed the ability of each model to predict human question preferences on the one-shot games. For the human data, a question quality score was computed as the average rank position across the 25 participants, where 0 is the worst question and 5 is the best question. The correlation between these human quality scores and the model scores were computed for each game individually (both Pearson (r) and Spearman rank (ρ) are reported).

Overall, full Bayesian EIG provides the best account of people's preferences. The model predictions for each one-shot game are shown in Fig. 2, and several individual games can be examined in detail in Fig. 1. The average correlation between EIG and human quality scores was $r = 0.777$ and $\rho = 0.718$. At first glance, it appears that games at depth 6 show a weaker correlation on average, but this may be due to one game with an unconventional selected object (a "brake").

The full results and the fits for the alternative models are shown in Table 2. EU has the second best fit with an average $r = 0.717$ and $\rho = 0.643$. While similar to EIG, its performance can degrade with depth when it becomes confident it can get the bonus, leading to weak preferences for subsequent questions. The CI model dramatically degraded with depth (as expected), achieving an average correlation of $r = 0.367$ and $\rho = 0.366$. This suggest that people are not just memorizing a set of "good questions" and asking those regardless of context. Finally, the RS model showed a strong average correlation with the human ratings ($r = 0.712$ and $\rho = 0.650$) when considering $K = 20$ objects but fared poorly when con-

Table 2: Pearson (r) and Spearman (ρ) correlations between human and model preferences in the one-shot games.

Game Number	Depth	Object	Full Bayesian r	Full Bayesian ρ	Expected Utility r	Expected Utility ρ	Context Insensitive r	Context Insensitive ρ	Random Subset k=20 r	Random Subset k=20 ρ	Random Subset k=5 r	Random Subset k=5 ρ
1	0	Blanket	0.874	0.829	0.723	0.771	0.874	0.829	0.864	0.827	0.847	0.805
2	2	Kitchen	0.767	0.667	0.825	0.667	0.597	0.522	0.737	0.669	0.690	0.640
3	2	Step	0.825	0.829	0.825	0.826	0.815	0.829	0.835	0.822	0.816	0.792
4	2	Almond	0.839	0.812	0.820	0.812	0.876	0.812	0.840	0.806	0.843	0.802
5	4	Scarecrow	0.834	0.812	0.907	0.812	0.679	0.734	0.765	0.809	0.709	0.753
6	4	Ground	0.850	0.829	0.796	0.829	0.238	0.486	0.630	0.626	0.118	0.162
7	4	Cricket	0.928	0.943	0.927	0.943	-0.480	-0.314	0.854	0.816	-0.315	-0.198
8	6	Brake	0.320	0.257	0.526	0.257	0.216	0.257	0.356	0.326	0.267	0.213
9	6	Tuba	0.795	0.600	0.064	-0.086	-0.044	-0.371	0.558	0.302	-0.061	-0.140
10	6	Mop	0.739	0.600	0.763	0.600	-0.104	-0.143	0.687	0.507	0.457	0.300
Average	2		0.810	0.769	0.823	0.768	0.763	0.721	0.804	0.766	0.783	0.745
Average	4		0.871	0.906	0.877	0.861	0.146	0.308	0.750	0.750	0.171	0.239
Average	6		0.618	0.486	0.451	0.257	0.023	-0.086	0.534	0.378	0.132	0.124
Average	All		0.777	0.718	0.717	0.643	0.367	0.366	0.712	0.650	0.437	0.412

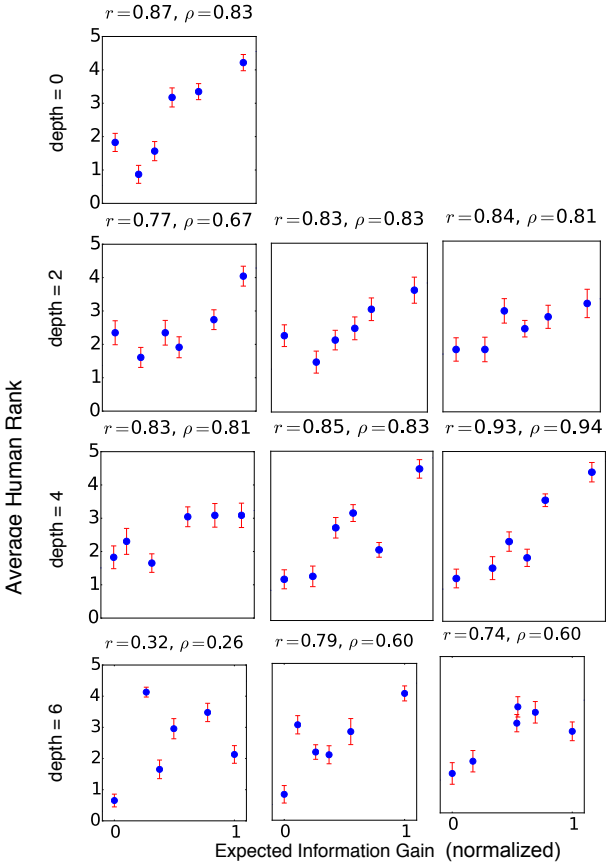


Figure 2: Bayesian model fits for one-shot games, with the different games organized by depth. The average human rank score (error bars show ± 1 s.e.) is compared with expected information gain (normalized between 0 and 1).

sidering $K = 5$ objects ($r = 0.437$ and $\rho = 0.412$), suggesting that while people may approximate the full Bayesian inference by considering a random subset of hypotheses, the subsets may need to be of considerable size. RS performance across a wider range of subset sizes is shown in Fig. 3.

The Familiarity model had to be analyzed differently. Since different participants saw each question a different number of times, the model correlation was computed separately for each participant's ranking. The correlation values averaged across participants as well as one-shot games were low (average $r = 0.0799$, $\rho = 0.0806$), but it varied substantially by depth since early questions are more likely to be repeated (depth 0, $r = 0.454$; depth 2, $r = 0.168$; depth 4, $r = -0.0479$; and depth 6, $r = -0.005$). While this statistic lacks the power of the previous calculations, the EIG model correlations computed in the same way yielded a higher average fit ($r = 0.384$, $\rho = 0.383$), suggesting that question familiarity was not the driving factor in question choice.

Analysis of complete games. We also analyzed the ability of the models to predict participant choices in the complete 20Qs games. The same analyses used in the one-shot games could not be repeated for these complete games for several reasons. First, participants chose to ask just one question at each step rather than rank a list of options. Second, each participant played a unique set of games, unlike the pre-generated one-shot games for which responses could be aggregated. In an accuracy analysis, the models selected the question with maximum score, and these choices were compared to the participant choices. The EIG model performed best at 28.4% correct where chance is 16.7%, followed by EU (27.0%) and CI (23.6%). The RS models had accuracy levels of 21.5% ($K = 5$) and 25.8% ($K = 20$). All of the models were above baseline, but no model succeeded in predicting a majority of choices. These results point to the challenges of predicting individual responses in the complete 20Qs games.

Conclusions

People ask questions when faced with uncertainty, seemly undaunted by very large hypothesis spaces. Previous work has used Bayesian modeling and the Expected Information Gain (EIG) to explain how people assign value in 20 Questions (20Qs), but these studies have considered restricted hypotheses spaces that can be presented all at once to participants, such as the children's game "Guess who" (Nelson et

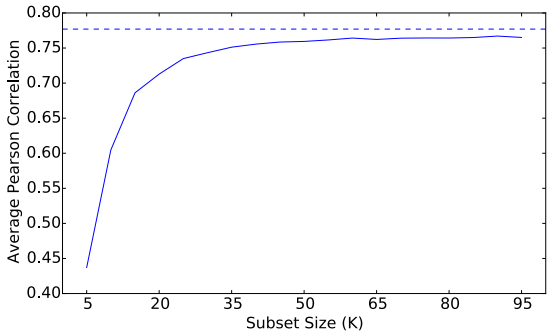


Figure 3: Average correlation (r) between human preferences on the one-shot games and the Random Subset model for different subset sizes K (solid line; same as Table 2 last row). For comparison, the full Bayesian model is shown as the dotted line.

al., 2014). It is unclear how people assign value to questions in very large hypothesis spaces that they are unlikely to represent explicitly and in their entirety.

Here we studied 20Qs in a very large hypothesis space – unbounded, from the perspective of participants – by relying on people's pre-existing semantic knowledge. While this creates difficulties for formal modeling, we presented a Bayesian model that incorporated a large data set of objects and question responses. In a series of one-shot and full games, people's ranking and selection of questions were best predicted by the full Bayesian model and expected information gain. On the one-shot games, this model achieved relatively strong average correlations with the aggregate human preferences ($r = 0.78$ and $\rho = 0.71$), providing a better explanation for the data than a range of alternatives, including a simple form of approximate Bayesian inference that considers small random subsets of the hypothesis space (Navarro & Perfors, 2011). Our results are consistent with the view of people as "good question-askers," although predicting individual responses in the full 20Qs games remains a challenge.

Many questions remain for future work. How does the Bayesian model perform on unusual games when a rare object is selected (e.g., brake or scarecrow) instead of a more common object (e.g., blanket or apple)? Our framework could also be used to explore how changes to the prior influence choice, comparing scenarios where the secret object was selected by different processes (e.g., by a child, an adult, or a computer). An additional challenge is to explain how people generate questions in more natural, free-form tasks rather than choosing from a pre-selected set (Rothe et al., 2016).

While our computational-level analysis will help constrain future algorithmic accounts of this ability (Marr, 1982), it does not itself offer a concrete algorithm for how people evaluate question quality in large hypothesis spaces. While our analyses suggest people consider more than just a few hypotheses, it seems implausible they are considering (even implicitly) hundreds or thousands of hypotheses when less demanding and still accurate inference strategies may exist.

An intriguing possibility is that people may aim to discover the right higher-level category rather than the specific item, at least during early stages of a 20Qs game. People could use pre-existing semantic categories such as animals, plants, living things, artifacts, vehicles, tools, etc. as these intermediate goals. Compared to the full Bayesian account, this alternative algorithm predicts that people devalue questions that cross-cut the high-level categories they are considering. For instance, while "Does it move?" is often a good question according to the full Bayesian model, it applies to subsets of both living things and artifacts, cutting across the salient hierarchical structure in the hypothesis space and thus potentially devaluing it relative to other questions. The predictions of this account can be tested empirically.

Future work will explore this and other algorithms for approximate Bayesian inference, as well as individual difference data, to further our understanding of how people search large hypothesis spaces by asking questions.

Acknowledgments. We thank M. Palatucci and Intel for their data set, and Todd Gureckis for helpful comments on an earlier draft. This research was supported by the Moore-Sloan Data Science Environment at NYU.

References

Denney, D. R., & Denney, N. W. (1973). The use of classification for problem solving: A comparison of middle and old age. *Developmental Psychology*, *9*(2), 275–278.

Eimas, P. D. (1970). Information processing in problem solving as a function of developmental level and stimulus saliency. *Developmental Psychology*, *2*(2), 224–229.

Gureckis, T. M., & Markant, D. B. (2009). Active Learning Strategies in a Spatial Concept Learning Game. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Markant, D., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Marr, D. C. (1982). *Vision*. San Francisco, CA: W.H. Freeman and Company.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. (2009). Zero-Shot Learning with Semantic Output Codes. In Y. Bengio, D. Schuurmans, & J. Lafferty (Eds.), *Advances in Neural Information Processing Systems (NIPS).*

Rothe, A., Lake, B. M., & Gureckis, T. M. (2016). Asking and evaluating natural language questions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Thornton, S. (1982). Challenging "Early Competence": A Process Oriented Analysis of Children's Classifying. *Cognitive Science*, *6*, 77–100.

Using a smartphone game to promote transfer of skills in a real world environment

Inge Doesburg (doesburg.inge@gmail.com)
Department of Artificial Intelligence, University of Groningen, The Netherlands

Niels Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen, The Netherlands

Abstract

This article presents an experiment in which participant's working memory, tasks-switching and focusing skills are trained in a game called Wollie on a smartphone. Before and after the training period they performed three task (a recall, Stroop and task-switching). The goal of this research was to see how the participants, from the test group, learn within the game and how this affects the three tasks. Only in the Stroop results a clear difference between the two groups was found. However, we found that participants who had the most trouble in playing Wollie, improved the most on Stroop and task-switching, indicating that these participants still lacked the relevant skills for all these tasks.

Keywords: transfer; smartphone; working memory; task-switching; games

Introduction

The goal of this paper is to test if transfer can be found between a game on a smartphone and multiple cognitive tasks. The phenomenon of using certain skills learned in one task, while performing a different task is called transfer of skills. Taatgen (2013) describes transfer as an overlap between tasks. The author explains that skills needed to perform a task are broken down into small elements, called primitive information processing elements (PRIMs). While learning a specific skill, more general skills can be learned as well, which consist of a combination of PRIMs. When two tasks both need one or more of the same combination of PRIMs (units), training on one task can also improve the performance on the other task. When there is improvement in the second task, this means that there was transfer of the combination of PRIMs.

An important point to make is that in this study we focus on the assumption, as stated by Taatgen (2013), that the main motor of transfer is growth of skills and strategies. Many studies on transfer see brain-training as if training a muscle, but this would not explain how transfer occurs. While the primitive elements (PRIMS) theory shows how skills and strategies can be more easily understood to transfer between tasks.

Research on transfer often points to the field of education as the field that can gain the most from knowledge of transfer. Transfer can help to learn new skills quickly and most learning opportunities occur in educational facilities. Transfer is not only important within educational facilities, but also between a student's education and future workplace (Eraut, 2009; McKeough et al., 2013).

By making people train cognitive skills by using a game on their phone it will feel less like training and more like playing a fun game. This will help with keeping them excited and motivated (Prins et al., 2011). There have already been successful transfer studies using games to train certain skills (Anguera et al., 2013), but these were still played on a computer in an experimental setting. Dufau et al. (2011) have shown how cognitive research that would normally take years, can by using smartphones, be done in a couple of months.

We propose an experiment to see if transfer can occur between a game on a smartphone and three tasks played on a computer. Two groups of participants will be participating, a test and a control group. Both groups perform three tasks (recall, Stroop and task-switching) on a computer before and after training on their smartphone game. The control group receives the game Tetris, which trains no cognitive skills. The test group receives the game Wollie, which trains working memory, task-switching and focusing skills. We expect the test group to show more improvement on the task, due to transfer of skills. However, we expect improvement only to occur in those participants who do not yet posses the skills and strategies needed in the particular tasks.

Methods

Participants

54 participants (31 women, 23 men; mean age 23.4 years, range 19-39 years) participated in this study. Participants were compensated for their time.

Procedure

Participants came to the university to perform session 1 of the experiment. They needed an Android phone, with a screen size between 3.7 and 5.5 inch and have an Android version of 3.0 or higher. The first 9 digits of the IMEI number was used as a subject number, to make sure that it was easy to match the pre and post data to the data send to the server by the smartphone application. Participants performed three task, taking about 15 minutes each. The first task was a recall task, the second a Stroop task and the third a task-switching task.

After performing these three tasks, the participants were assigned to one of two groups. This was done semi-randomly to ensure that the two groups would, on average, not differ too much on their scores in the first session. For each new participant we would take the average of the three pretest scores, and compare this to the running averages of the participants that were already tested. On the basis of that comparison the

participant would be assigned to the precondition that would reduce the difference between the running averages most.

Stroop task

In the second part of the experiment, the participant performed a Stroop task. This task was created in PsychoPy2 (Peirce, 2007) and is based on the similar Stroop task used by Juvina & Taatgen (2009). In each trial the participant would see three words. The word in the middle had a particular color: red, blue or green. The participant was asked to press Z if the left word described the color of the word in the middle, or M if the right word described the color of the word in the middle. The critical measure in the Stroop task is the Stroop interference: the difference between reaction times on congruent trials (color and word are the same) and incongruent trials (color and word are different).

There was a practice block containing 12 trials, followed by 4 blocks of 2x36 trials. The entire task used a list with 36 possible trials of which each was randomly chosen once in each part of each block. Between each block participants were allowed to take a short break.

Task-switching task

The final task was created in MATLAB (MATLAB 8.5, The MathWorks Inc., Natick, MA, 2015). The type of task was based on the similar task used by Karbach & Kray (2009), while the design is based on the similar task by Rogers & Monsell (1995). In this task, the participant is shown 4 squares. Every trial, a picture will appear in one of these squares. The first picture will always appear in the upper left square. In later trials, the picture will appear in the next square in a clockwise direction. Whenever a picture appeared in the upper two squares, the task was to judge whether the object in the picture was small or large. Whenever a picture appeared in the lower two squares, the task was to judge whether the object in the picture was a fruit or a vegetable. This means that in half of the trials the task is the same as in the previous trial, and in the other half the task switches.

The measure in task switching is the difference in reaction times between repetition trials (task remains the same), and switch trials (task switches).

Recall task

The recall task was created in PsychoPy2 (Peirce, 2007) and is based on the similar recall task used by Chein & Morrison (2010). In this task the participant was presented with a number of words and letters. Their goal was to recall only the letters. When presented with a letter, they did not need to do anything, but when presented with a word they needed to perform a decision task. They were asked to press yes, the right ctrl key, if they thought the word referred to themselves. If they did not think the word referred to themselves they were asked to press no, the left ctrl key. Examples of words presented in this tasks are: lazy, caring, arrogant and tidy.

The task consisted of seven blocks of which the first was a practice block. Each block contained three types of trials, one with 4 letters, one with 5 and one with 6 (span 4, 5 and 6). Each trials consisted of the presentation of the first letter, followed by 4 seconds of the decision task. After this the next letter was presented, again followed by 4 seconds of the decision task. This continued until the length of the span was fulfilled. Each letter was presented for one second. When all letters where presented followed by the decision task, the participant had to recall the letters in the correct order.

Wollie

The game Wollie was created for the test group to train the subjects working memory, task switching and focussing skills. It is based on a smartphone game called Rules, which can be found in the App Store. See figure 1 for multiple screen shots of the game Wollie.

The game consists of 9 levels, which can be played in two difficulty modes, beginner and expert. The goal of this game is to complete each level, by remembering and applying rules. Each new level is a bit harder than the previous one, because the higher the level the more complex the rules. In each level you will eventually have to remember 10 rules. See table 1 for the rules in level 1. Each level starts with a presentation of the first rule, for instance: Tap in descending order. After being presented with this rule, a new screen appears with a 4 by 4 grid of blocks, each block containing a picture and a number.

The first rule, needs to be applied to these blocks. Applying this particular first rule, means tapping the blocks in order of high to low numbers. When the correct block is tapped, it disappears. When an incorrect block is tapped it will expand for half a second after which it returns to its former size. When all blocks are removed from the field within the given time frame the player has successfully completed the first rule. In the second part of the level, the player is presented with the second rule, for instance: Tap all green things. After the presentation of the second rule, a new 4 by 4 grid appears. Now that we have more than one rule, the newest rule is always applied first. In this case this means, that all blocks containing an image which is green will be tapped first, when there are no more green images, the first rule must be applied, until all blocks are removed from the field. Again, when the player empties the screen within his playing time, he has successfully completed the second part of this level. Whenever one rule is applied successfully, meaning that there are no more blocks for this rule and the player needs to switch to an older rule, the screen flashes yellow for a second. In case the smartphone is on vibrate, the phone will also vibrate on a switch moment. If the smartphone is on normal mode, it will not vibrate but it will play a short sound. All of this is to help the player realize that he needs to switch to a different rule. This process described above repeats itself until the player

has learned and applied 10 rules or fails to tap all images before the time is up.

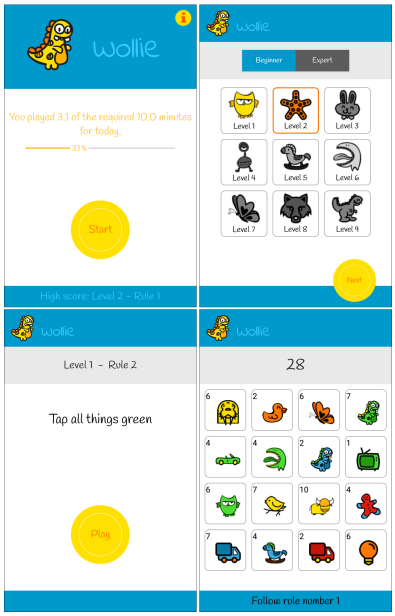


Figure 1: Four screen shots of the smartphone game Wollie. Top left: the home screen. Top right: screen to chose which level to play. Bottom left: screen showing the newest rule to apply. Bottom right: game play screen with 4x4 grid.

Table 1: Wollie rules level 1.

Rules level 1 Wollie	
1	Tap numbers in descending order.
2	Tap all things green.
3	Tap odd numbers.
4	Tap nines.
5	Tap animals.
6	Tap walruses.
7	Tap monsters.
8	Tap green monsters.
9	Tap birds.
10	Tap tens.

The rules that need to be applied to the 4x4 grid have to be recalled from memory. Learning to store and recall these rules will help the participants train their working memory. Every time they have tapped all the images for one rule, they need to switch to the previous rule and tap the images for this rule. These switching moments will help train the participant's task-switching skills. Learning to focus on the rule to come and therefore thinking about which attributes of the images to focus on will train their focusing skills. To help the participant in remembering to play the game everyday a notification will appear on their phone everyday at 7 o'clock at night.

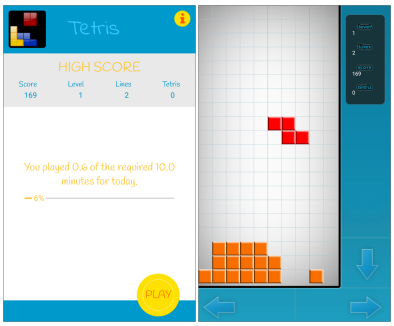


Figure 2: Screen shots of the smartphone game Tetris. Left: the home screen, right: the game play screen.

Tetris

The game Tetris was created for the control group. The game Wollie requires subjects to switch between tasks and recall rules, while Tetris only requires subjects to focus on one object at a time. Making it a good control task. Because Tetris is a well known game, which is reproduced for all kinds of platform we used some basic Java code found on the open-source website Github (www.github.com/tdan94/Tetris, at the moment of publishing this Github repository was no longer available) and added what we needed to use it in our experiment. See figure 2 for some screen shots of our version of Tetris.

Both applications show the participants how much they had to play on that day, to make sure they reached their total playing time of 140 minutes within two weeks. The applications also presented a reminder at seven o'clock at night, which told them if they had already played enough and if not, how much time they still needed to play that day. Data about their training behavior was stored on the participant's smartphone and send to a server. After two weeks of training the participants came back for the second session. They performed the same three tasks as performed in the first session, namely recall, Stroop and task-switching.

Results

For each task we will present an array of results. First, we test the difference in improvement between the control and test group with an ANOVA. After which we dig a little deeper into the influence of each group's smartphone application on the scores of the task. The influence of the condition (Wollie/Tetris) on the pre/post scores will be given.

For each task a table will be presented showing the correlation between the following variables and the accuracy score of the task: Time, Level20, Level120, DiffLevel and HighScore. *Time* is the total amount of time a subject has trained with the application. This tests whether more training leads to better transfer. *Level20* is the highest level reached after 20 minutes. This tests whether initial progress on the tasks can predict the amount of transfer. *Level120* is the highest level reached after 120 minutes. This tests whether progress made in the minimum amount of time subjects had to train predicts transfer. *DiffLevel* is the difference between *Level20* and *Level120*. This tests more or less the same as just *Level120*, but focusses on the learning gain. *HighScore* is the highest level reached. Because Tetris has no levels we use the highest score at the specific time points instead of the level when looking at the influence of Tetris on the accuracy score of the task.

Stroop

In figure 3 the average pre and post scores on the Stroop task are presented. The ANOVA showed a weak effect of group on improvement in accuracy $F(1,52)=3.91$, $p=0.053$. To see if the game variables had some impact on the improvement on the task, the correlations were calculated. Table 2 shows the influence of previously discussed variables on the improvement in the Stroop task. In the Wollie group both *DiffLevel* and *HighScore* have a negative significant influence on the improvement. This means that the lower the *DiffLevel* or *Highscore*, the higher the improvement on the task. For Tetris no significant correlations were found. In figure 4 the significant negative correlation between *DiffLevel* and the improvement in accuracy can be seen. The correlation of -0.42 ($p=0.03$) means that 17.6% of the improvement on the task can be assigned to the *DiffLevel* variable. This figure also shows us that subjects who improved greatly in Wollie, were also likely to score high on the first Stroop task session.

Task Switching

In the Task Switching data we excluded one person in the control group who had a switch-cost of 2182 ms. Figure 5 shows the average pre and post switch-costs for each group. The ANOVA showed no effect of group on improvement in accuracy $F(1,51)=0.996$, $p=0.32$. However, in table 3 it can be seen that there are a few variables in the Wollie group that had a significant correlation with the amount of improvement in switch-costs, namely *Level120*, *DiffLevel* and *HighScore*. Like in the Stroop data these correlations were negative, meaning that the faster subjects progressed through the game, the lower the improvement on task switching.

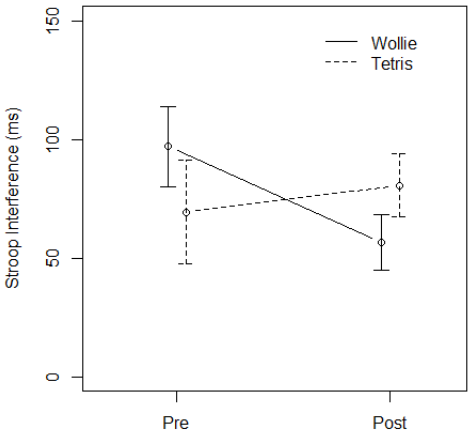


Figure 3: The influence of the two different smartphone applications on the Stroop Interference.

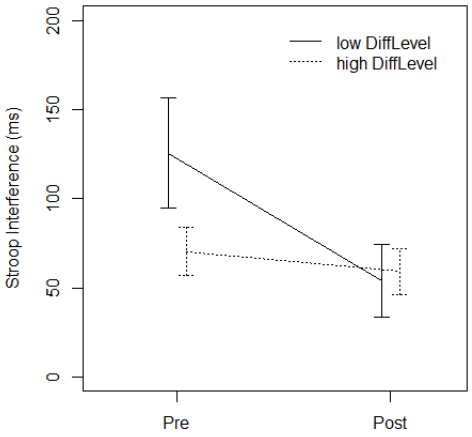


Figure 4: The improvement on the Stroop Interference for subjects with a low and subjects with a high DiffLevel in the test group.

Table 2: Correlations between performance measures in the training and improvement in Stroop interference. The p-values are presented within parenthesis. The * means that the result was significant ($p<0.05$).

Stroop correlation values					
Group	Time	Level20	Level120	DiffLevel	HighScore
Wollie	-0.07 (0.74)	-0.20 (0.92)	-0.37 (0.057)	-0.42 (0.03*)	-0.42 (0.03*)
Tetris	0.16 (0.44)	-0.20 (0.32)	0.05 (0.82)	0.29 (0.14)	0.004 (0.99)

Table 3: Correlations between performance measures in the training and improvement in switch costs in task switching. The p-values are presented within parenthesis. The * means that the result was significant ($p<0.05$).

Task Switching correlation values					
Group	Time	Level20	Level120	DiffLevel	HighScore
Wollie	0.05 (0.80)	-0.18 (0.37)	-0.42 (0.029*)	-0.43 (0.023*)	-0.42 (0.031*)
Tetris	0.0 (1.0)	0.08 (0.71)	0.18 (0.38)	0.20 (0.32)	0.03 (0.87)

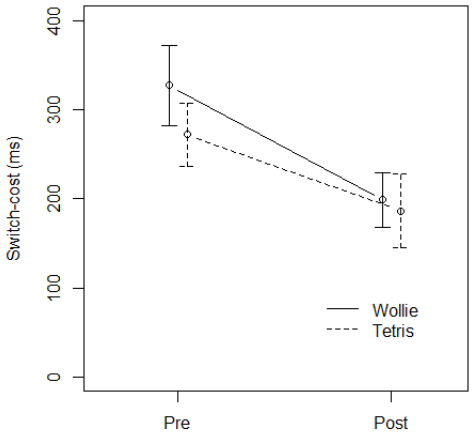


Figure 5: The influence of the two different smartphone applications on the switch-costs in the Task Switching task.

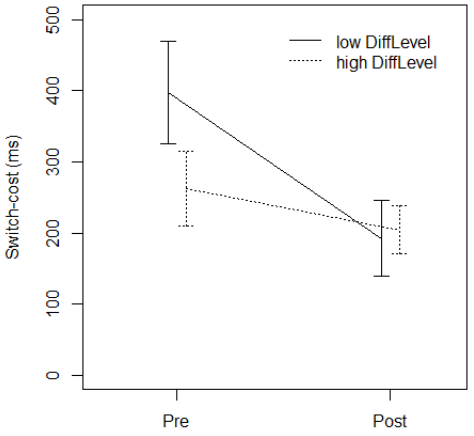


Figure 6: The improvement on the switch-costs between subjects with a low and subjects with a high DiffLevel in the test group. A low DiffLevel means that subjects only had a small improvement in reached levels between 20 and 120 minutes of training.

In figure 6 a graph with a visualization of the correlation between subjects' DiffLevel and the amount of improvement on the switch-costs is shown. The correlation of -0.43 (p=0.023) means that 18.5% of the improvement on this task can be assigned to the *DiffLevel* variable. The graph is very similar to figure 4, where the improvement for the low and high DiffLevel group was also presented, but then for the Stroop task. Just like figure 4, figure 6 indicates that subjects who managed to progress through the Wollie levels at a fast pace (a high *DiffLevel*) were also likely to score high on the first task-switching session.

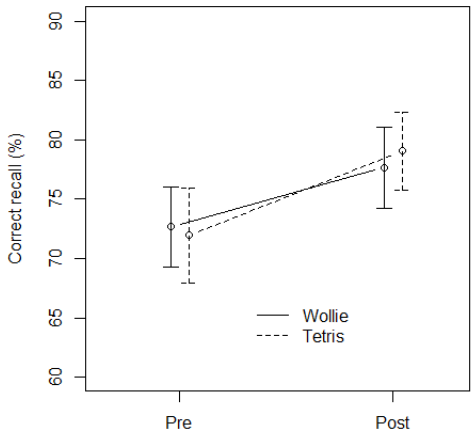


Figure 7: The influence of the two different smartphone applications on the correct amount of recalled letters in the Recall task.

Recall

Figure 7 shows the average pre and post recall scores of the two smartphone application groups. Looking at the graph it is not surprising that when performing an ANOVA it showed no effect of group on improvement in accuracy $F(1,52)=0.412$, $p=0.524$. Both group seem to improve an equal amount. In table 4 the correlations between the game variables (mentioned above) and the improvement on the recall task are presented. There is only one significant correlation, which is between *Level120* and the improvement on the task in the Tetris group.

Discussion

The goal of this experiment was to test whether a game in which skills can be identified that correspond to well-known

Table 4: Correlations between performance measures in the training and improvement in recall score. The p-values are presented within parenthesis. The * means that the result was significant ($p<0.05$).

Recall correlation values					
Group	Time	Level20	Level120	DiffLevel	HighScore
Wollie	-0.20 (0.32)	0.10 (0.61)	0.18 (0.36)	0.18 (0.36)	0.10 (0.64)
Tetris	-0.14 (0.48)	-0.27 (0.17)	-0.39 (0.046*)	-0.31 (0.12)	-0.38 (0.051)

experimental paradigms can improve performance on these paradigms. The evidence we found is mixed: a comparison to a Tetris control condition does not result in a consistent benefit of Wollie over Tetris, with only a weak effect on the Stroop task.

However, when we correlate the improvement on the game to the tests, we see that subjects that progress through the Wollie game at a fast pace are already good at both Stroop and task switching, and therefore show little improvement. On the other hand, subject that struggle to improve on the game show considerable improvement on both Stroop and task switching. A possible explanation is that some subjects has already mastered the appropriate control skills before they came into the experiment. These subjects had an easy time on the game, but therefore also had no benefit from playing it. Other subjects were not yet strong on the relevant control skills, and therefore still had something to learn.

A limitation of the research presented here is that the control condition (Tetris), despite its use in the past as a control condition in gaming studies, may have cognitive benefits after all. This may be the case for the improvement on recall, because there subjects that made slow progress with Tetris made more progress on Recall. However, any benefits from Tetris are more diffuse, but may still have affected the direct comparisons with Wollie.

Future research on this topic should focus on those people who do not yet posses the skills and strategies needed to succeed in particular areas of their lives. Those people are the ones who will eventually benefit most from knowledge on transfer. They could be children in an educational facility, but also adults needing to improve some skills to apply in their workspace. This research has shown that people can benefit from even a small amount of training with a smart phone application. Nowadays, most people own a smart phone and carry it with them all day. Making this an easier and fun way to practice new skills.

References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... others (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*(7465), 97–101.

Chein, J. M., & Morrison, A. B. (2010). Expanding the minds workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*(2), 193–199.

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., ... others (2011). Smart phone, smart science: how the use of smartphones can revolutionize research in cognitive science. *PloS one*, *6*(9), e24974.

Eraut, M. (2009). 2.1 transfer of knowledge between education and workplace settings. *Knowledge, values and educational policy: A critical perspective*, 65.

Juvina, I., & Taatgen, N. A. (2009). A repetition-suppression account of between-trial effects in a modified stroop paradigm. *Acta psychologica*, *131*(1), 72–84.

Karbach, J., & Kray, J. (2009). How useful is executive control training? age differences in near and far transfer of task-switching training. *Developmental science*, *12*(6), 978–990.

McKeough, A., Lupart, J. L., & Marini, A. (2013). *Teaching for transfer: Fostering generalization in learning*. Routledge.

Peirce, J. W. (2007). Psychopypsychophysics software in python. *Journal of neuroscience methods*, *162*(1), 8–13.

Prins, P. J., Dovis, S., Ponsioen, A., Ten Brink, E., & Van der Oord, S. (2011). Does computerized working memory training with game elements enhance motivation and training efficacy in children with adhd? *Cyberpsychology, behavior, and social networking*, *14*(3), 115–122.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, *124*(2), 207.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439.

Acknowledgments

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen.

Solution of division by access to multiplication: Evidence from eye tracking

Shawn Tan (shawn.tan@carleton.ca)
Institute of Cognitive Science, Carleton University

Kasia Muldner (kasia.muldner@carleton.ca)
Institute of Cognitive Science, Carleton University

Jo-Anne LeFevre (jo-anne.lefevre@carleton.ca)
Institute of Cognitive Science, Carleton University

Abstract

People report solving division problems by mentally recasting division problems as multiplication (e.g., 72 ÷ 8 → 8 × [?] = 72). Mediation of division by multiplication occurs mainly on larger problems. Eye tracking data was used to determine whether patterns of gaze durations on division problems provided support for mediation. Adults solved division problems in two formats: traditional (e.g., 72 ÷ 8 = [ ]) and recasted (e.g. 72 = 8 × [ ]). Processing of individual problem elements was compared across formats. Results provide support for mediation. Processing patterns for traditionally-formatted problems were more similar to those for traditional division in earlier work (72 ÷ 8) whereas problems in recasted format (72 = 8 × [ ]) were more similar to patterns found when participants solved multiplication problems (e.g., 8 × 9). These findings provide a novel source of support for differential processing of problems across presentation formats.

Keywords: mental arithmetic; strategies; numerical cognition; mathematical cognition; division; eye tracking; gaze duration

Introduction

Mental arithmetic involves the coordination of memory retrieval, strategy selection, and decision processes. Much of the work on mental arithmetic involves basic problems such as 9 + 6, 18 ÷ 6 or 7 × 5. Even for these basic problems, participants report using a variety of solution processes. Although memory retrieval is the most commonly reported strategy, addition and subtraction problems may also be solved using counting (LeFevre et al., 1996a; 2006), and multiplication problems may be solved via reference to a related fact (e.g., solving 9 × 6 as 10 × 6 – 6; Campbell & Xue, 2001; LeFevre et al., 1996b). Our goal in the present research was to use eye tracking to provide a detailed description of the processes involved in solving division problems. We focused on simple division, that is, the set of problems which are defined by the inverse of multiplication problems with operands ranging from 2 to 9 (i.e., 2 × 2 to 9 × 9), such as 12 ÷ 4, 28 ÷ 7, and 54 ÷ 9.

Participants report using two strategies to solve simple division problems, that is, *retrieval* and *recasting*. LeFevre and Morris (1999) found that when participants were presented with division problems (e.g., 72 ÷ 9), they reported first mentally recasting these as multiplication problems (e.g., 9 × [ ] = 72), and subsequently retrieving the answer associated with the implied multiplication problem (e.g., 9 × 8).

In general, people solve problems with smaller operands more quickly and accurately than they solve problems with larger operands (Zbrodoff & Logan, 2005). Although this *problem-size effect* is extensively documented, the causes of it are not yet fully understood (Ashcraft & Guillaume, 2009). One possibility is that the problem-size effect arises, in part, from individual differences in strategy use across different types of problems. Accordingly, LeFevre and Morris (1999) found that strategy reports varied with the size of the operands. For division problems with small operands, direct retrieval was the most frequently reported strategy, whereas recasting was mainly reported for division problems with dividends larger than 25 (referred to as *large*) and was less frequently reported for problems with dividends of 25 or less (referred to as *small*). The additional step of recasting could explain why division problems took longer to solve. However, LeFevre and Morris (1999) did not explicitly manipulate the format of division problems.

Manipulating the format of basic arithmetic problems offers insight into the representations of arithmetic facts and the types of cognitive processes used by problem solvers. Mauro, LeFevre and Morris (2003) found that participants took longer to solve problems in division formats than in recasted multiplication formats, particularly for large problems. Based on these results, Mauro et al. proposed the *mediation hypothesis*: Presenting division in recasted formats (e.g., 42 = 7 × [ ]) provided problem solvers with a visual representation that is more compatible with the process of activation and retrieval of associated multiplication facts than the traditional division formats (e.g., 42 ÷ 7 = [ ]). They also proposed that solving division-format problems produces longer latencies than solving the corresponding problem in multiplication format because solvers recast the problem before retrieving the solution. This recasting stage is not needed for division problems that are shown in recasted formats or for division problems in traditional format that have small operands because they are solved through direct retrieval. Similar patterns have been identified for subtraction problems, which participants report recasting into addition formats to access their knowledge of addition (Campbell, 2008).

Although Mauro et al. (2003) found evidence for the mediation hypothesis, they relied on broad behavioral

Table 1: Examples of formats, elements and spatial positions of division problems used in the present experiment. Of primary interest are the recasted format (b) and the traditional format (d), shown in bold.

	Spatial Positions				
	1	2	3	4	5
			Recasted format		
(a)	72	=	[ ]	×	9
(b)	**72**	**=**	**9**	×	**[ ]**
			Traditional format		
(c)	72	÷	[ ]	=	9
(d)	**72**	÷	**9**	**=**	**[ ]**

measures (reaction time, error rate). We propose that more fine-grained information about how attention is allocated across formats could shed further light on the mediation hypothesis. Accordingly, in the present work we used eye tracking to record participants' gaze patterns as they solved division problems in traditional division and recasted multiplication formats. In general, eye tracking allows the identification of potential interest areas in a visual display that capture participants' attention as they process the display in response to a specific stimulus. This method has been used extensively to study participants' focus of attention during performance of tasks such as reading or scene processing (Rayner, 2009). However, use of eye tracking techniques is still novel in the field of numerical cognition (reviewed by Hartmann, 2015), and thus there are relatively few studies that have used eye tracking to study simple arithmetic (cf. Moeller, Klein, & Nuerk, 2011; Zhou, Zhao, Chen, & Zhou, 2012). Curtis, Huebner, and LeFevre (in press) compared patterns of eye tracking on single-digit addition and multiplication problems, and the corresponding inverse subtraction and division problems. For division problems such as 56 ÷ 8, they found that participants spent more time processing the dividend (i.e., 56) compared to the divisor (i.e., 8). In contrast, on the inverse multiplication problems (e.g., 8 × 7), processing time was divided evenly between the two operands.

In the present study, to further investigate the mediation hypothesis, participants' gaze information was recorded as they solved problems in either a traditional division format (e.g., 72 ÷ 8 = [ ]) or in a recasted multiplication format (e.g., 72 = 8 × [ ]). Interactions of problem elements and elements' location within the structure of the problem can influence dwell time durations and resulting gaze patterns. Given that changing the location of one problem element entailed a change in location of other elements in the problem, it was therefore necessary to control the location of a specific element across all problem formats. Since prior work suggests that problem size is a factor in determining whether participants use recasting (LeFevre and Morris, 1999; LeFevre, Mauro and Morris, 2003), we controlled for confounds between problem size and spatial location by maintaining the dividend at the first spatial position across formats (note that the dividend determines problem size). Gaze times upon each problem element and overall gaze patterns were analyzed to obtain information on the processing of problems in recasted vs. traditional formats and over large and small problems. We subsequently compared eye-tracking patterns from division problems in our study with eye-tracking patterns from division and multiplication problems reported in Curtis et al. (in press) consisting of three elements (e.g., 72 ÷ 9, 8 × 9).

Methods

Stimuli

As shown in Table 1, we created division problems in two formats: traditional (c and d) and recasted (a and b). The dividend (i.e., the larger number) was always located in the first position whereas the missing element was located at either the third or fifth position. Thus, there were four possible problem formats. Although participants solved problems presented in all four formats, we were most interested in comparing the most familiar traditional format, that is format (d), to the recasted format (b). This comparison matched the positions of the dividend and the divisor, such that only the location of the symbols varied across the two formats.

For division formats, we created 128 division problems by combining divisors ranging from 2 to 9 with quotients ranging from 2 to 9 and varying the location of the missing element at either the third or the fifth position. In this set, there were 112 non-tie problems in which divisors and quotients are not equal. These were divided into two sets of 56 such that reciprocal problems appeared in different sets (e.g., 6 ÷ 2 = 3, 6 ÷ 3 = 2). There were 8 tie problems with equal divisors and quotients (e.g., 9 ÷ 3 = 3). Tie problems were presented twice in each set. Thus, there were 56 non-tie and 16 tie problems totaling to 72 problems in division format for each set. The procedure was repeated for multiplication formats. This resulted in a total of 288 problems. We divided the problems into two experimental sets, counterbalanced by format, problem size, and location of missing element. Participants were randomly assigned to one of the experimental sets.

We also created 16 practice problems by combining dividends ranging from 2 to 9 with a divisor of 1, located at either the 3rd or 5th position respectively. The missing element was located at the alternative location (5th or 3rd position).

Problems were presented in black 60 point Arial font on a white background. Five interest areas were defined, one around each problem element. Each interest area measured 6 cm by 4 cm and was centered on the problem element (see Figure 1). The borders of each interest area were connected but did not overlap.

Participants

Thirty-three students (13 males, 20 females) from Carleton University participated in the experiment for course credit.

All participants had normal or corrected-to-normal vision, and ranged in age from 16 to 33 years (*Mdn* = 20).

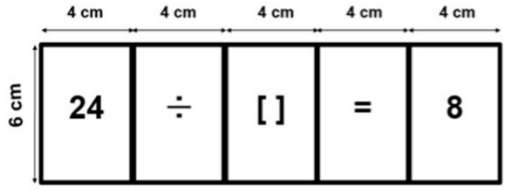


Figure 1: Dimensions of interest regions centered on problem elements for a typical problem in the experiment.

Procedure

Participants were seated at a distance of approximately 60cm in front of a desktop computer with a 15-inch by 13-inch monitor. The computer was linked to an SR-Eyelink 1000 eye tracker. The experiment began with the practice block followed by the 144 problems (36 in each of the four formats) divided into four blocks, with 36 problems in each block. Participants were allowed to take a break after completing each block. Each block started with a calibration phase followed by presentation of the problems.

At the start of each trial, a 1.25 cm by 1 cm rectangle randomly appeared at one of the four corners of the screen. Participants were instructed to look at this rectangle, and once the eye tracker detected a fixation upon the rectangle, the problem appeared. This procedure was used to ensure that participants' attention would not be focused on the central interest area upon initial presentation of the problem, because this could result in artificial inflation of fixation times on the central interest area. Participants were asked to verbally state their answer into a microphone placed directly in front of them and their responses were connected to an ASIO voice trigger accurate to +/- 1 ms. Problems remained on the screen until the microphone recorded a response. Once this occurred, the problem disappeared and the experimenter recorded the response using a keyboard. Eye gaze information was recorded by the eye tracker. The experiment lasted for about 50 minutes.

Results

Trials were discarded if participants made an irrelevant sound before providing a solution or if the apparatus failed to detect a response. This resulted in 338 (7.1%) of the trials being discarded. Participants made errors on a further 300 (6.8%) of the remaining trials. These were not included in the analyses of response latencies.

Overall Performance

To get a sense of overall performance, we analyzed median latencies and percentage of errors for all participants in two separate 2(format: multiplication vs. division) by 2(problem size: large vs. small) by 2(position of missing element: 3rd vs. 5th) repeated-measures ANOVAs. Means (of medians) across participants are shown in Table 2. Small problems had dividends of 25 or less; large problems had dividends larger than 25.

As expected, participants were faster at solving small problems than large problems (1393 vs. 1760 ms), $F(1,32) = 40.6$, $p < .001$, $\eta_p^2 = .56$. The location of the missing element also had an effect: Participants responded more slowly when the third element was missing than when the fifth element was missing (1605 vs. 1547 ms), $F(1,32) = 5.61$, $p = .024$, $\eta_p^2 = .15$. Although participants were faster on recasted than on traditional formats (1560 vs. 1593 ms), the difference was not significant, $F(1,32) = 1.6$, $p = .21$, $\eta_p^2 = .05$.

Mirroring the results for response time, participants made fewer errors on small than on large problems (4.7% vs. 9.4%), $F(1,32) = 16.8$, $p < .001$, $\eta_p^2 = .34$. No other main effects or interactions were significant.

In summary, the overall analyses were consistent with our expectations. Participants were faster and more accurate on small than on large problems. We also replicated the trend for recasted formats to be solved more quickly than traditional formats. However, the location of the missing element also influenced solution times. Thus, further analyses were focused on the comparison of the most familiar traditional format, (d) in Table 1, to the recasted format with the same structure, (b) in Table 1. For this pair, large problems in recasted format were solved more quickly than those in traditional format (1689 vs. 1735 ms).

Table 2: Mean latencies, standard errors, and percentages.

Format	*M* (ms)		*SE* (ms)		*Error* (%)	
	Small	Large	Small	Large	Small	Large
Traditional Format						
72÷[ ]=9	1444	1808	74.3	128.3	4.2	10.4
72÷9=[ ]	1385	1735	66.2	119.8	5.0	9.6
Recasted Format						
72=[ ]×9	1364	1805	62.4	109.7	3.6	7.7
72=9×[ ]	1380	1689	69.6	97.1	5.9	9.8

Eye gaze patterns

Gaze duration. Gaze duration, defined as the sum of all fixation times in an area of interest was calculated and analyzed for each of the five interest areas (as shown in Figure 1). Analyses of the total number of fixations in each area of interest produced similar results and so are not described in further detail.

To obtain information on participants' attention patterns for the interest areas in the traditional format (e.g., 72 ÷ 9 = [ ]) and the corresponding recasted format (e.g., 72 = 9 × [ ]), we analyzed median gaze durations for each participant in a given interest area in a 5(interest area: left number, symbol-1, middle number, symbol-2, missing element) ×

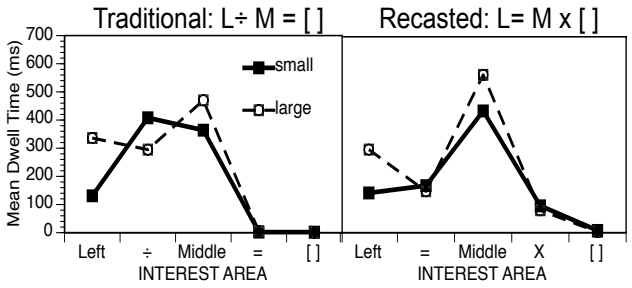


Figure 2: Mean gaze durations (total dwell time) for each interest area by format and problem size.

2(format: traditional, vs. recasted) × 2 (problem size: small vs. large) repeated-measures ANOVA. The analysis showed main effects of problem size $F(1,32) = 33.69$, $p < .001$, $\eta_p^2 = .513$, and interest area, $F(4,29) = 105.34$, $p < .001$, $\eta_p^2 = .936$, interactions between format and interest area, $F(4,29) = 15.63$, $p < .001$, $\eta_p^2 = .683$; interest area and problem size, $F(4,29) = 17.23$, $p < .001$, $\eta_p^2 = .704$; and among format, interest area, and problem size, $F(4,29) = 4.09$, $p < .01$, $\eta_p^2 = .361$ (see Figure 2).

As shown in Figure 2, participants spent very little time on the last two interest areas, presumably because these contained no important information (e.g., the blank was always in the last position in the two formats that are the target of this analysis). In the traditional division format, participants looked longest at the division sign and the middle number (the divisor). Compared to small problems, processing time on large problems was also allocated to the left number (the dividend). In contrast, in the recasted format, participants spent more time looking at the middle interest area (i.e., the divisor), with relatively little time spent looking at the equal sign or the left number. On large problems, somewhat more time was spent looking at the left number, but this was still considerably less than gaze durations on the middle number in this format.

In some respects, these results are similar to those found by Curtis et al. (2016), who used traditional division problems but presented these without the equal sign or the symbol for the missing element (e.g., 72 ÷ 9). Specifically, in both studies, gaze durations were longer for the dividend when solving large as compared to small problems, accounting for the overall effect of problem size. However, in contrast to Curtis et al., in our data there was a tendency for fixations in the division format to be shifted rightwards toward the middle number, possibly because the additional symbols in our problems resulted in an overall shift of gaze to the centre of the display. Curtis et al. observed longer gaze durations to the operator, which in their stimuli appeared in the middle of the display, on small problems for all operations. Although these results suggest that processing of arithmetic problems is directed to the centre of the display, regardless of the underlying symbols, this pattern does not always hold. Specifically, Curtis et al. found that on large division problems, participants spent longer fixating on the left operand (the dividend) than on either the operation sign or the divisor. As stated above, this was not the case in the present research, where gaze durations to the dividend were longer on large traditional problems than in the other conditions, but participants nevertheless fixated most on the middle operand (the divisor). Thus, there were similarities between the traditional division format in this research and in Curtis et al., and between recasted formats and multiplication problems in Curtis et al.

The pattern of results we obtained, coupled with previous findings, provides some support for the mediation hypothesis. However, gaze durations for arithmetic problems also varied depending on the specific elements included in the problem display. Even though participants did not spend much time processing the equal sign or the missing element in the traditional format, the presence of these elements may have influenced the overall processing patterns. Such findings suggest that gaze patterns may need to be interpreted carefully across different display conditions.

Distribution of attention over time. To explore gaze patterns over the course of solving a division problem (i.e., gaze profiles), we normalized the distribution of processing times for each problem. This normalization was necessary because participants varied in terms of the total time spent to solve a given problem. For each problem, the total time taken to solve the problem was divided into six equal time intervals. Within each time interval, gaze duration on each interest area was calculated by summing the duration of all fixations for that interest area. We then calculated the median gaze duration for each interest area and time interval for each participant, and then calculated means across all participants. Figure 3 shows these mean distributions of gaze durations in each interest area across all time intervals for small and large problems in traditional and recasted formats.

As shown in Figure 3, problem format and problem size influenced participants' distribution of attention over time. However, in comparing the graphs in Figure 3, some common patterns of attention can still be observed. First, there were longer mean durations on the first symbol than the first and middle operands during time interval 1 for all four graphs, suggesting that initial processing on all problems was directed more towards the first symbol (dashed red line in Figure 3) than towards either operand. This pattern may indicate that participants initially fixate to the left of center (see Schneider, Maruyama, Dehaene, & Sigman, 2012, for a similar finding). The left of center area (i.e., the second area of interest) contains the symbol relevant for discriminating between the formats: A division sign in the second area of interest indicates a traditional problem whereas an equal sign indicates a problem in recasted format. Gaze duration on this interest area declines

quickly in the next two intervals for all except small problems in traditional formats (e.g., 10 ÷ 2 = [ ]; 18 ÷ 3 = [ ]). For all problem types, gaze duration in the middle of the display (i.e., the single-digit divisor, see green line in Figure 3) increases in time interval 2.

For the recasted format, problem size did not have much effect on gaze profiles. Following focus of attention on the division sign during the early stages of the trial, gaze duration on the middle interest area dominated the remainder of the trial (green line in Figure 3, lower panels). There was stable amount of processing to the larger number (i.e., the dividend, blue dotted line in Figure 3) that was greater on large than on small problems. On both large and small problems, the problem size effect appeared to be distributed equally over the left and middle operands (see also Figure 2). The similarity in patterns of processing between small and large problems observed here for the recasted format was also reported by Curtis et al. (in press) for traditional multiplication problems (e.g., 3 × 9, 8 × 7). There were longer gaze durations on the dividend for large problems than small problems but this did not influence distribution patterns over time.

In contrast to the findings for the recasted format, problem size did influence distribution of attention patterns over time for the traditional format. On small problems, gaze durations were similar for the middle operand and the division sign after time interval 2. Gaze duration on the left operand (the dividend) decreased after time interval 1, suggesting that participants allocated less attention to the dividend after the beginning of the trial. In contrast, for large problems, gaze duration on the operation sign declined after the first interval. The left operand (dividend) and middle operand had similar gaze durations for the first half of the trial, with a gradual shift toward a focus on the middle operand in the last half of the trial. This pattern may reflect recasting of the problem in the first half of the trial, followed by retrieval based on multiplication knowledge. Curtis et al. found that the dividend was processed longer on large than on small division problems. We also found a similar result with longer gaze durations for large than small problems over all time intervals.

In summary, the analyses of gaze durations across the course of the trial support the mediation hypothesis in that gaze distribution patterns over time varied across formats. We observed similar patterns of gaze distribution over time for large and small problems in recasted format as compared to different patterns of gaze distribution over time across large and small division problems in traditional format. These differences occurred even though the location of the numerical information was identical in the two formats.

Discussion

In this research we used eye tracking measures to assess whether strategy differences reported in LeFevre and Morris (1999) on division problems (i.e., retrieval and recasting) would be reflected in gaze patterns. Based on earlier studies (LeFevre and Morris, 1999; Mauro et al., 2003), retrieval

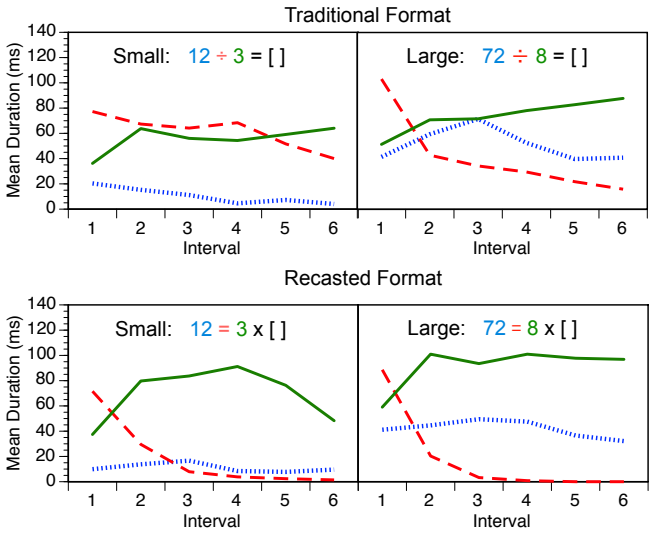


Figure 3: Gaze durations normalized across time intervals. Processing time on the dividend (i.e., the largest number) is shown by the blue dotted line, the first symbol (i.e., the division sign in the traditional format and the equal sign in the recasted format) is shown by the red dashed line, and the divisor (the smaller number) is shown by the green solid line. Because very little time was spent on the second symbol or the missing element (see Figure 2), these were not included in the graphs.

from memory was the most common strategy for small problems, as well as for problems in recasted formats, whereas for large problems in traditional formats, participants reported mentally transforming the problem into a multiplication format before retrieving the solution (e.g., mentally changing 72 ÷ 8 into 8 × [ ] = 72). We hypothesized that eye patterns would reflect, at least in part, the mental activities going on during problem solution (Rayner, 2009) and that differences observed in eye-tracking measures across formats would be associated with a change in strategy that would influence gaze patterns. The present research supports this hypothesis by showing that gaze patterns revealed by eye tracking data vary both across formats and problem size.

Gaze patterns for problems presented in recasted format (i.e., 72 = 8 × [ ]) were similar to patterns obtained from participants who solved traditional multiplication problems (i.e., 8 × 9) reported by Curtis et al. (in press), suggesting that in our study, participants were recasting the division problems into a multiplication format. In both cases, processing time was concentrated on the central element of the presented stimulus, with relatively less time spent processing the other problem elements. Curtis et al. suggested that participants' focus on the centre of the display (i.e., the operation sign) reflected mental processing occurring during retrieval of the answer, because it continued after the two operands had been processed. In the present research, the focus on the center of the display (the single-digit divisor) may also indicate that participants are

using this number to access multiplication knowledge, as part of mental processing involved in retrieval of an answer. In contrast, when the same problems were presented in the traditional division format, processing time was distributed more evenly across the first, second, and third interest areas, with additional processing time on the dividend (first interest area) for large as compared to small problems. Curtis et al. also observed this problem-size effect on the dividend for division problems in the traditional format.

Notably, we also identified some factors that may be important for interpreting gaze patterns. Comparisons between the results of Curtis et al. (in press) and the present research suggest that varying the way problems are displayed may influence processing patterns. Division problems in Curtis et al. were presented without equal signs or missing answers (e.g., 72 ÷ 8). Although there were similarities in the distribution of processing time across that format and the format used in the present research (e.g., 72 ÷ 8 = [ ]), there were also differences that appear to be related to the presence of the additional problem elements. These findings suggest that a larger range of formats that balance the position of specific elements and control for format familiarity will be necessary in studies that use eye tracking to study mental arithmetic.

Despite this qualification, the present research demonstrates the utility of using eye-tracking measures to extend and corroborate findings from more traditional dependent variables. We also showed the importance of comparisons across experiments with varying problem structures and formats. Comparing gaze profiles over time further extends inferences from dwell time analyses to mental activity over the time course of the problem solving process. In summary, gaze durations and time-course profiles provided theoretically and empirically relevant information about mental arithmetic.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada through Discovery Grants to J. LeFevre and K. Muldner.

References

Ashcraft, M. H., & Guillaume, M. M. (2009). Mathematical cognition and the problem size effect. In B. Ross (Ed.), *Psychology of Learning and Motivation, Vol. 51* (Vol. 51, pp. 121–151). Burlington: Academic Press.

Campbell, J. I. D. (2008). Subtraction by addition. *Memory & Cognition*, *36*(6), 1094–102.

Campbell, J. I. D., & Xue, Q. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General*, *130*(2), 299–315. doi:10.1037//0096-3445.130.2.299

Curtis, E., Huebner, M. G., & LeFevre, J. (in press). The relationship between problem size and fixation patterns during addition, subtraction, multiplication, and division. *Journal of Numerical Cognition*.

Hartmann, M., Mast, F. W., & Fischer, M. H. (2015). Spatial biases during mental arithmetic: evidence from eye movements on a blank screen. *Frontiers in Psychology*, *6*, 1-8.doi: 10.3389/fpsyg.2015.00012

LeFevre, J.A., Sadesky, G. S., & Bisanz, J. (1996a). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 216–230. doi:10.1037//0278-7393.22.1.216

LeFevre, J.A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996b). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*(3), 284–306. doi:10.1037/0096-3445.125.3.284

LeFevre, J. A., DeStefano, D., Penner-Wilger, M., & Daley, K. E. (2006). Selection of Procedures in Mental Subtraction. Canadian Journal of Experimental Psychology, 60(3), 209–220. doi:10.1037/cjep2006020

LeFevre, J. A., & Morris, J. (1999). More on the relation between division and multiplication in simple arithmetic: Evidence for mediation of division solutions via multiplication. *Memory and Cognition, 27,* 803-812.

Mauro, D. G., LeFevre, J. A., & Morris, J. (2003). Effects of problem format on division and manipulation performance: Division facts are mediated via multiplication-based representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 163-170.

Moeller, K., Klein, E., & Nuerk, H.-C. (2011). Three processes underlying the carry effect in addition--evidence from eye tracking. *British Journal of Psychology*, *102*, 623–645. doi:10.1111/j.2044-8295.2011.02034.x

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Schneider, E., Maruyama, M., Dehaene, S., & Sigman, M. (2012). Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas. *Cognition*, *125*(3), 475–490. doi:10.1016/j.cognition.2012.06.015

Zbrodoff, N. J., & Logan, G. D. (2005). What everyone finds: The problem-size effect. In *Handbook of mathematical cognition* (pp. 345–508). Psychology Press, New York, NY.

Zhou, F., Zhao, Q., Chen, C., & Zhou, X. (2012). Mental representations of arithmetic facts: Evidence from eye movement recordings supports the preferred operand-order-specific representation hypothesis. *Quarterly Journal of Experimental Psychology*, *65*(4), 661–74. doi:10.1080/17470218.2011.61621.

The role of word-word co-occurrence in word meaning learning

Abdellah Fourtassi (a.fourtassi@ueuromed.org)
The Euro-Mediterranean University of Fes
Department of Psychology, Stanford University

Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Ecole Normale Supérieure / PSL Research University / EHESS / CNRS, France

Abstract

A growing body of research on early word learning suggests that learners gather word-object co-occurrence statistics across learning situations. Here we test a new mechanism whereby learners are also sensitive to word-word co-occurrence statistics. Indeed, we find that participants can infer the likely referent of a novel word based on its co-occurrence with other words, in a way that mimics a machine learning algorithm dubbed 'zero-shot learning'. We suggest that the interaction between referential and distributional regularities can bring robustness to the process of word acquisition.

Keywords: word learning; semantics; cross-situational learning; distributional semantic models; zero-shot learning.

Introduction

How do children learn the meanings of words in their native language? This question has intrigued a lot of scholars studying human language acquisition. Quine (1960) famously noted the difficulty of this process. In fact, every naming situation is ambiguous. For example, if I utter the word *gavagai* and point to a rabbit, you may possibly infer that I mean the rabbit, the rabbit's ear, or its tail or color,...etc. A popular proposal in the language acquisition literature suggests that, even if one naming situation is ambiguous, being exposed to many situations allows the learner to narrow down, over time, the set of possible word-object mappings (e.g., Pinker, 1989). This proposed learning mechanism has come to be called Cross-Situational Learning (hereafter, XSL). Laboratory experiments have shown that humans are cognitively equipped to learn in this way. For example, L. Smith and Yu (2008) presented adults with trials that simulated real world uncertainty: each trial was composed of a set of words and a set of objects, in such a way that no single trial had enough information about the precise mappings. However, after being exposed to many of such trials, participants were eventually able to name the objects with a better-than-chance performance. Many experiments replicated this effect with adults, children and infants (Yu & Smith, 2007; Suanda, Mugwanya, & Namy, 2014; Vlach & Johnson, 2013). Subsequent research tried to characterize the algorithmic underpinnings of XSL. Some experiments suggested that learners accumulate in a parallel fashion all statistical regularities about word-object co-occurrences, and they use them to gradually reduce ambiguity across learning situations (McMurray, Horst, & Samuelson, 2012; Vouloumanos, 2008; Yurovsky, Yu, & Smith, 2013). Other experiments suggested that learners maintain, instead, a single hypothesis about the referent of a given word. New evidence either corroborate this hypothesis or contradict it (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013). Yurovsky and Frank (2015) proposed a synthesis of both accounts, whereby the learner's choice to adopt one of the two learning strategies depends on the complexity of the learning situation.

This being said, XSL is unlikely to be the unique mechanism of word learning at work. First, real learning situations are much more ambiguous than typical simulated situations used in laboratory experiments. When subjects are tested in a more realistic learning context, the load on memory increases and, therefore, the ability to make use of the available visual information diminishes (Medina et al., 2011; Yurovsky & Frank, 2015). Second, XSL assumes a perfect covariance between words and their referents. This assumption does not take into account the fact that words –in real situations– are sometimes uttered in the absence of their referents (e.g. when talking about past events, "remember that cat?"). In this experiment, we propose a statistical learning mechanism that purports to complements XSL, through relying on cues from the concomitant linguistic information, and more precisely on word co-occurrence.

Form word co-occurrence to semantic similarity

Typical XSL settings assume that words occur in isolation. In real learning contexts, however, words are embedded in natural speech, and have consistent distributional properties. In particular, semantically similar words tend to co-occur more often than semantically unrelated words. For example, the word "ball" and "play" tend to co-occur more often than "ball" and "eat". This fact is documented in linguistics under the name of the 'distributional hypothesis' (hereafter, DH) (Harris, 1954), and has been popularized by Firth's famous quote "You shall know a word by the company it keeps" (Firth, 1957). The distributional hypothesis is also the basis for distributional semantics, the sub-field of computational linguistics that aims at characterizing words' similarity, based on their distributional properties in large text corpora. Tools from the field of distributional semantics such as Latent Semantic Analysis (Landauer & Dumais, 1997), Topic Models (Blei, Ng, & Jordan, 2003), or more recently Neural Networks (Mikolov, Karafiát, Burget, Cernocký, & Khudanpur, 2010) have proved to be effective in modeling human word similarity judgement (Landauer, McNamara, Dennis,

& Kintsch, 2007; Griffiths, Steyvers, & Tenenbaum, 2007; Fourtassi & Dupoux, 2013; Parviz, Johnson, Johnson, & Brock, 2011).

Zero-shot learning

Models that learn through DH typically require a large corpus, especially if nothing is known about the language. Here, we explore the case where some words are already known and only one word is learned through DH. This corresponds to the so-called 'zero-shot learning' situation.

An interesting example of this situation has been given by Socher, Ganjoo, Manning, and Ng (2013). They built a model that can map a label to a picture even when the label has not been used in training! More precisely, using the CIFAR-10 dataset, the model was first trained to map 8 out of the 10 labels ("automobile", "airplane", "ship", "horse", "bird", "dog", "deer", "frog") in the dataset, to their visual instances. The remaining labels ("cat" and "truck") were omitted and reserved for the zero-shot analysis. Second, they used a distributional semantic model (based on Neural Networks) to obtain vector representations for the entire set of labels (i.e., including "cat" and "truck") based on their co-occurrence statistics in a large text corpus (Wikipedia text). When tested on it ability to classify a new picture (a cat or a truck) under either the label of "truck" or "cat", the model performed with a high accuracy, using only the patterns of co-occurrence among labels, and the semantic similarity between the new and old pictures. For example, when presented with the picture of a cat, the model has to classify it as "cat" or "truck". The models makes the link between the picture of the cat and that of a similar picture (e.g. dog), and chooses the label that is more related to the label of this similar picture, i.e., "cat". In fact, "cat" co-occurs more often with "dog" than with, say, "airplane". Therefore the label "cat" is favored over the alternative label (i.e., "truck").

The conditions of zero-shot learning are often met in the context of word acquisition. For instance, this corresponds to the (rather ubiquitous) situation where an unknown word is heard in the absence of its visual referent. Therefore, we suggest that human learners can go about it in a way that mimics the mechanism of zero-shot learning. In the following, we test this hypothesis with adults, following closely the spirit of the model developed by Socher et al. (2013).

Method

The experiment consists of 4 steps:

1. Referential familiarization
2. Learning consolidation
3. Distributional familiarization
4. Semantic generalization

The referential familiarization and consolidation consists in explicitly teaching subjects the association between words in an artificial language and their referents. In the distributional familiarization, participants hear 'sentences' made of words from this artificial language without visual referents; some of these words were familiar (introduced in the referential familiarization), and others were novel words. Crucially, the novel items co-occur consistently with words of the same semantic category. Finally, the semantic generalization phase tests whether subjects can rely on distributional information *alone* to infer the semantic category of the novel words, without any prior informative referential situation. Below is a detailed description of each step of the experimental procedure.

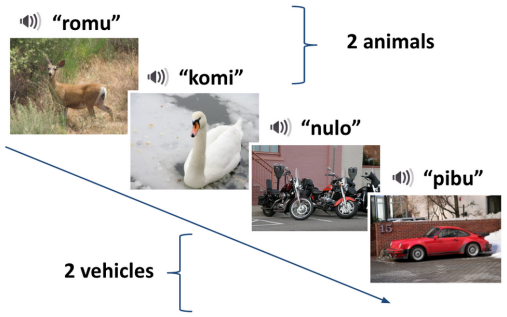


Figure 1: Referential familiarization. Participants are presented with multiple series of word-objects pairings. The objects belong to the category of animals or the category of vehicles.

Step 1: referential familiarization In this phase of the experiment (Figure 1), participants are taught the pairing of 4 words in an artificial language[1] with 4 objects. The objects belong to either the category of vehicles (car, motorcycle) or the category of animals (deer, swan). Participants see a picture of the referent on the screen and hear its label simultaneously. There are 3 trials, each consists of a randomized presentation of the series of 4 pairings.

Step 2: learning consolidation The purpose of this phase is to consolidate and strengthen the participants' knowledge about the 4 word-object pairings (Figure 2). Participants are tested using a Two Alternative Forced Choice paradigm (2AFC). They are presented with a series of trials where they hear a label (*pibu*, *nulo*, *romu* or *komi*) and are shown two objects; one of which is the correct referent, and the other belongs to the other semantic category. Crucially, after they have made a choice, they get a feedback on their answers ("correct"/"wrong"). Participants are presented with 16 questions of this sort, which correspond to the combinatorial possibilities of forming pairs of items from one semantic category with items from the other category (4 cases), in conjunction with the order of the visual presentation of the referents (4×2 cases) and the item being labeled ($4 \times 2 \times 2 = 16$ cases in total).

[1] The audio stimuli were graciously provided by Naomi Feldman.

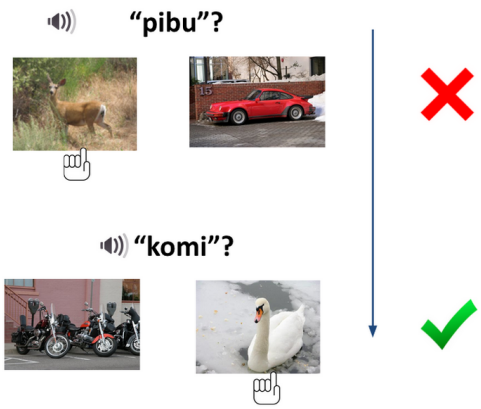


Figure 2: Learning consolidation. Two-Alternative Forced Choice paradigm (2AFC), with feedback.

Step 3: distributional familiarization Distributional familiarization follows the referential training and consolidation. Participants listen to 'sentences' made of words from this artificial language without any visual referent. As explained in Figure 3, each sentence consists of 3 words. Two of which are familar words from one semantic category, i.e., either *romu* and *komi* (animals) or *pibu* and *nulo* (vehicles). The third word is a new artificial word that consistently co-occur with them. The new words are *guta* and *lita*. The way *guta*/*lita* are distributed with either (*romu*, *komi*) or (*pibu*, *nulo*) was counterbalanced across participants so as to avoid different sorts of linguistic and perceptual biases that may arise from the way the stimulus is organized. There is a 750 ms pause between words, and 2500 ms pause between sentences. There are 16 sentences in total, 8 for each semantic context; (*romu*, *komi*) and (*pibu*, *nulo*). Words within sentences are randomized and the semantic context is alternated during the exposure.

Step 4: testing semantic generalization Participants are presented again with a two alternative forced choice. As explained previously in the learning consolidation phase, they hear a label and they are asked to choose between two objects, but here participants do not get feedback on their answers. We are particularly interested in how participants respond in the situation where they hear the novel item (*guta* or *lita*) and are presented with two new objects that represent a new animal (squirrel) and a new vehicle (trolley). Participants have never been shown the referential mapping of the new words, so their answer would reveal whether distributional learning alone had helped them infer semantic knowledge about the word (i.e., the semantic category of the referent). This test phase is composed of 4 questions about the novel labels/objects, varying the visual order of the objects (1×2) and the object being named ($1 \times 2 \times 2 = 4$ cases in total), in addition to 4 selected questions about the familiar words/objects used in the referential training. We eliminated any overlap between questions about novel items and questions about familiar items so as to avoid any form of cross-situational learning during the test phase.

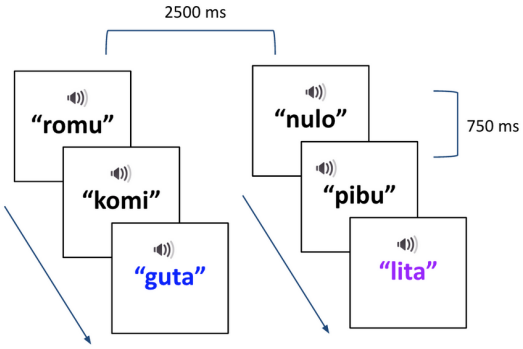


Figure 3: Distributional familiarization. Sequences of words are presented with no visual referents. Two new words ("guta" and "lita") are introduced and co-occur consistently with the words corresponding to one of the two semantic categories ("romu" and "komi" for the category of animals, and "nulo" and "pibu" for the category of vehicles)

Procedure As shown in Figure 4, participants are first trained on the pairing between 4 artificial words and their referents (part 1 and part 2). Then they are exposed to 2 blocks of distributional familiarization (part 3), and they are tested 3 times (part 4): before any exposure to distributional information (baseline) and after the first and the second block of distributional exposure (respectively session 1 and 2).

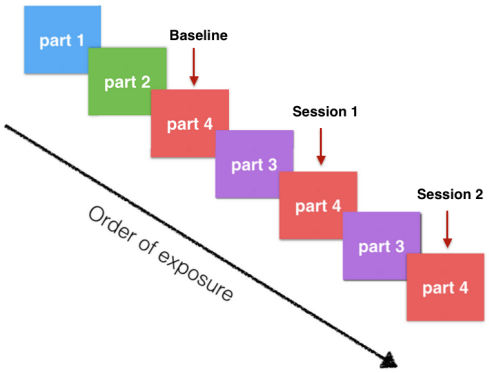


Figure 4: Order of exposure in the experiment. Participants are trained referentially once (part 1 and part 2), distributionally twice (part 3). They are tested in three sessions (part 4): before and after each block of distributional learning

Population and rejection criterion 50 Participants were recruited online through Amazon Mechanical Turk. We included in the analysis participants whose total score on the familiar word-object questions during the testing phases (i.e., part 4) were above chance level. This is a way to select only subjects who paid attention during the training parts. 2 par-

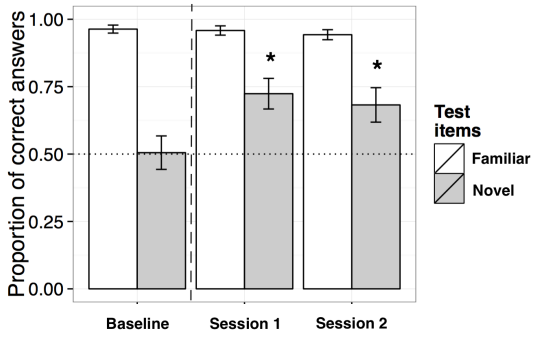


Figure 5: proportion of correct answers for familiar and novel test items, before any distributional exposure (baseline) and after the first and second block of exposure (session 1 and 2)

ticipants were excluded based on this criterion.

Results and Analysis

Figure 5 shows the proportion of correct answers on both familiar and novel items, as a function of the testing session. In the familiar condition, answers were almost perfect in the three sessions (before exposure, after one block, and after two blocks of exposure to part 3). This shows that participants have reliably learned the association between words and their referents during the training phase, and that this learning was not affected by subsequent exposure to distributional information. In the novel condition, and before distributional training (i.e., baseline), subjects were at chance level ($M = 50.5\%$ of correct answers). A one sample t-test comparing the mean against chance (i.e, 50%) gives a $t(47) = 0.083$ with p-value $= 0.93$. The absence of learning is a predictable result since participants had no prior cue about the relevant object mapping. However, after one and two blocks of distributional training, subjects were significantly above chance level. A one sample t-test gives, respectively, for session 1 an average of correct answers $M = 72.4\%$, with $t(47) = 3.94$ ($p < 0.001$), and for session 2, an average of $M = 68.2\%$, with $t(47) = 2.85$ ($p = 0.006$). In order to compare the behaviour of the participants before and after distributional training, we performed a paired t-test. For baseline vs. session 1, there was a significant change, the difference mean is equal to $M = 0.218$, with $t(47) = 2.99$ ($p < 0.01$). Similarly, for baseline vs. session 2, the difference mean is $M = 0.177$, with $t(47) = 2.24$ ($p = 0.029$). However, between session 1 and session 2, the difference mean $M = 0.041$ was not significant, $t(47) = 0.662$, $p = 0.51$. This shows that most of the learning occurred during the first block of distributional exposure. Additional training did not significantly improve learning (if anything, it seems to slightly decrease the average of correct responses).

Discussion

The results show that, when learning the meaning of words, people are sensitive, not only to the co-occurrence of words and objects (as suggested in XSL), but also to co-occurrence statistics between words themselves (as suggested in the DH). More importantly, we showed that these two sensitivities interact in a way that mimics a machine learning mechanism called zero-shot learning. In fact, participants in our experiment were able to guess the semantic category of a novel word whose visual referent was never presented through the semantic properties of the words with which it co-occurred consistently. Participants knew beforehand that they would be introduced to an artificial language and that they would have to learn the meaning of words in this language, but they were not explicitly instructed about the fact that words that co-occur in the same sentences are supposed to have similar meanings. Participants have spontaneously turned to co-occurrence in order to cue semantic similarity, and infer the category of the ambiguous words.

Although we used an artificial language whose 'sentences' fall short, on many aspects, of real speech, this work provides evidence for the *cognitive plausibility* of this learning mechanism, much in the spirit of the statistical learning literature (e.g., L. Smith & Yu, 2008; Saffran, Aslin, & Newport, 1996). If it scales up to real languages, this word-word co-occurrence mechanism would prove crucial in complementing word-object co-occurrence mechanisms. In fact, most word-object co-occurrence learning strategies (e.g. XSL) assume that words covary perfectly with their referents. This assumption is not always correct. For example, when talking about a past event, the conversation may not match the immediate visual context. In contrast, words used in a given conversation, be it about present, past or future events, normally co-occur in a coherent fashion. The learner can rely on this intrinsic property of speech to bring about robustness to the learning process. For example, suppose the learner, while at home, hears a discussion about the last visit to the "zoo". XSL learning, if operating alone, would be confusing. In contrast, if XSL operates in concert with DH, the learner would tend, if in doubt, to link a new word (e.g., "zoo") not to some surrounding object, but to other co-occurring words, which are likely to be zoo-related words (such as "animals", "bird" and "monkey"). Further work is needed to characterize the precise conditions under which learners would rather switch to the word-word co-occurrence cue to infer meaning.

Moreover, the proposed mechanism can help learners develop an early semantic representation for words with a rather abstract meaning. Abstract words (like "eat" and "good") are learned later in development than words with salient concrete referents (such as "ball" and "shoe") (e.g., Bergelson & Swingley, 2013). They are presumably harder to learn because there is no obvious or/and lasting correspondence between the word and the physical environment. Bruni, Tran, and Baroni (2014) developed a model which extends purely word-word co-occurrence learning strategies (such as LSA model) to also encompass co-occurrence with the visual context. They assessed the contribution of textual and visual information in approximating the meaning of abstract vs. con-

crete words. They found that visual information was mostly beneficial in the concrete domain, while it maintained an almost neutral impact on the abstract domain where most learning was based on word-word co-occurrence. Future work will investigate the extent to which this finding squares with psychological behaviour. For instance, an interesting question would be to test whether human learners switch from word-object cue to word-word cue when the potential abstractness of the target word increases.

Finally, during the write-up of this paper, it came to our knowledge that Ouyang, Boroditsky, and Frank (in press) conducted an experiment that shared many similarities with ours. However, it also presented interesting differences both in terms of the experimental setup and the results. Ouyand at al. exposed adult participants to auditory sentences from a MNPQ language. It is an artificial language where sentences take the form of "M and N" or "P and Q". Ms and Ps are used as context words, whereas Ns and Qs are target words. We believe there are two crucial differences between the two experiments. First, the context words (M and P) were composed of a mix of various proportions of real English words or non-words. In our experiment, they were all non-words. Second and more important, Ouyang et al. (in press) followed the spirit of MNPQ's paradigm in keeping constant the order of the words in the sentences, that is, M and P always occurring first in the sentence, and N and Q always occurring last. Our experiment was more faithful to the hypothesis of *bag-of-words*, which is crucial in distributional semantic models: order within a particular semantic context (e.g., a sentence) is irrelevant. It was therefore randomized across trials. Interestingly, although none of the context words we used were known words, we obtained a high learning rate. In contrast, Ouyang et al. (in press) obtained successful learning only when most of the context words were familiar English words. A plausible explanation for this difference is that, in the case of MNPQ language, participants have two possible learning dimensions: learning the positional patterns (what word comes first, and what words comes last) and learning the co-occurrence patterns (what couple of words co-occurred with each other). In fact, it has been shown that when both positional and co-occurrence cues are available, participants tend focus on the first ones (K. Smith, 1966). By using familiar words, Ouyang et al. (in press) showed that participants were more likely to learn co-occurrence patterns, probably through alleviating part of the memory constraint. In our case, the positional patterns were random, which left participants with only one learning dimension (i.e., co-occurrence pattern).

To conclude, this experiment provided a cognitive proof of principle to the zero-shot learning mechanism, according to which (early) semantic knowledge can be learned through sensitivity to word co-occurrence in speech. Future work will focus on exploring properties of this new learning and how it interacts with cross-situational learning.

Acknowledgments

This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, the Fondation de France, the Region Ile de France (DIM Cerveau et Pensée), and an AWS in Education Research Grant award.

References

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bruni, E., Tran, N., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, *49*.

Firth, J. R. (1957). A synopsis of linguistic theory 19301955. In *Studies in linguistic analysis.* Oxford, Blackwell.

Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. In *Proceedings of ACL.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146-162.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of INTERSPEECH.*

Ouyang, L., Boroditsky, L., & Frank, M. C. (in press). Semantic coherence facilitates distributional learning of word meaning. *Cognitive Science*.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the n400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop.*

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT press.

Quine, W. (1960). *Word and object*. The MIT Press.

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

Smith, K. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. Y. (2013). Zero-Shot Learning Through Cross-Modal Transfer. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational learning. *Cognitive Psychology*, *66*.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*(2), 729–742.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Yurovsky, D., & Frank, M. C. (2015). An Integrative Account of Constraints on Cross-Situational Learning. *Cognition*, *145*.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*.

Where Should Researchers Look for Strategy Discoveries during the Acquisition of Complex Task Performance? The Case of Space Fortress

Marc Destefano and Wayne D. Gray
Rensselaer Polytechnic Institute

Abstract

In complex task domains, such as games, students may exceed their teachers. Such tasks afford diverse means to trade-off one type of performance for another, combining task elements in novel ways to yield method variations and strategy discoveries that, if mastered, might produce large or small leaps in performance. For the researcher interested in the development of extreme expertise *in the wild*, the problem posed by such tasks is "where to look" to capture the explorations, trials, errors, and successes that eventually lead to the invention of superior performance. In this paper, we present several successful discoveries of methods for superior performance. For these discoveries we used *Symbolic Aggregate Approximation* as our method of identifying *changepoints* within score progressions in the venerable game of *Space Fortress*. By decomposing performance at these changepoints, we find previously unknown strategies that even the designers of the task had not anticipated.

Keywords: expertise, performance, Space Fortress, SAX, changepoint analysis, skill acquisition, plateaus, dips, leaps, strategy discovery, method invention

Introduction

In studying the behavior of people who become expert performers, we must look beyond group measures, and focus on the behavior of individuals; that is, at their explorations, failures, and successes as they strive to become experts. Ours is a variant of Ericsson and Ward's (2007) expert performance approach with the twist that we wish to capture our performers in the act of discovering the methods that result in extreme expertise.

For example, we had people play the complex game of *Space Fortress* (Mané & Donchin, 1989; Donchin, 1995) across 31 sessions of 8 games per session, or 248 games total. Averaging the data across hours and across players produces the classic performance curve shown at the bottom of Figure 1, which shows few exceptions to the story (Fitts & Posner, 1967; Newell & Rosenbloom, 1981) of steady, incremental increases in performance with practice. Unfortunately for this story, as the upper plot shows, this smooth average represents none of players' actual performance. Although each of our 9 players show improvements with time, these improvements are not smooth and each individual curve is more notable for its plateaus, dips, and leaps (Gray & Lindstedt, 2016) than for smooth and steady improvement with practice.

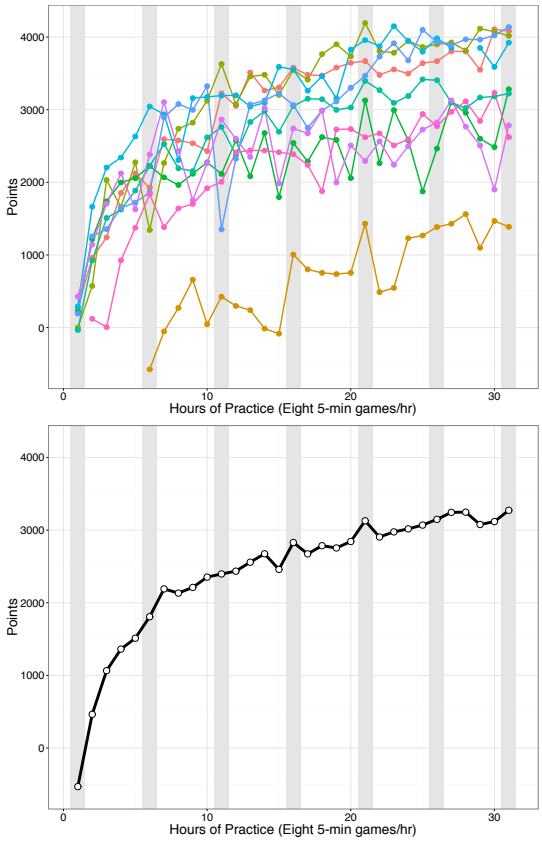


Figure 1. *Space Fortress Skill Acquisition Curves.* The top plot shows the scores of 9 Players who played Space Fortress for 31 hours each. To keep the plots compact, the early games for the lowest scoring players are truncated for hours 1-4. The bottom plot shows the average performance per hour for all 9 Players. The average (lower plot) is textbook perfect. As is clear, this average performance does not represent the progress of any of our 9 individual players.

Why so much fluctuation? Certainly, there are many external factors which could lead to such irregularity: time of day, stress, motivation, etc. Averaging across these factors has enabled the field to formulate general laws which capture what our subjects have in common. However, we argue that,

especially for people at the limits of expertise, it would be a mistake to dismiss individual variations as "statistical noise."

We believe that something meaningful is going on within individuals during these variations, and that is it important to study them as individuals. Different subjects of different skill levels may discover or invent different strategies. Different methods for implementing the same strategy may be easier for some individuals than others, and the nature of such differences would not be apparent in aggregate data.

Extreme expertise in tasks such as Space Fortress requires highly complex strategies that demand tremendous cognitive resources. More specifically, they are not simply "reactive" – they require strategic choices as well as the invention of methods that can be executed in the time alloted by a fast paced game, situations in which, "even hesitating requires a decision"[1] and in which a temporary decrease in performance may be the cost of exploring a different way of doing things. Gray and Lindstedt tell us that:

> ...dips in performance [are] not simply a concern due to their possibility of demotivating learners ...but should be viewed as periods of experimentation, discovery, trial & error, and successive approximations to developing league-stepping habits. The capture and study of these dips may well be the most important and overlooked task of the past 120 years of Experimental Psychology. (Gray & Lindstedt, 2016)

Thus, we propose that some of these fluctuations in performance are caused by strategy discovery, testing, and exploitation. To measure these fluctuations, we settled on SAX (Lin, Keogh, Lonardi, & Chiu, 2003) as our Change Point Algorithm to identify time periods to target for further investigation. Here, we present a case study of an expert user discovering and adopting strategies in *Space Fortress*, a game that allows incremental progress in low-level perceptual-action skills, while also allowing for planning and strategy formation. In doing so, we have been able to determine specific moments when our focal players implemented radically advanced strategies to maximize their performance.

Space Fortress

Space Fortress (Mané & Donchin, 1989) is a cognitive-task-oriented video game with a long history in the Cognitive Science community. Perhaps the most efficient way to describe it is that it's similar to *Star Castle* (Cinematronics, 1980) combined with a Sternberg-style memory task (Sternberg, 1966) with an AX-CPT task thrown in (Cohen & Servan-Schreiber, 1992). There have been many different versions of the game developed over the past 30 years, with some evolution of the ruleset over that time.

In our version (Destefano, 2010) the player operates a small spaceship via second-order (acceleration-based) thrust

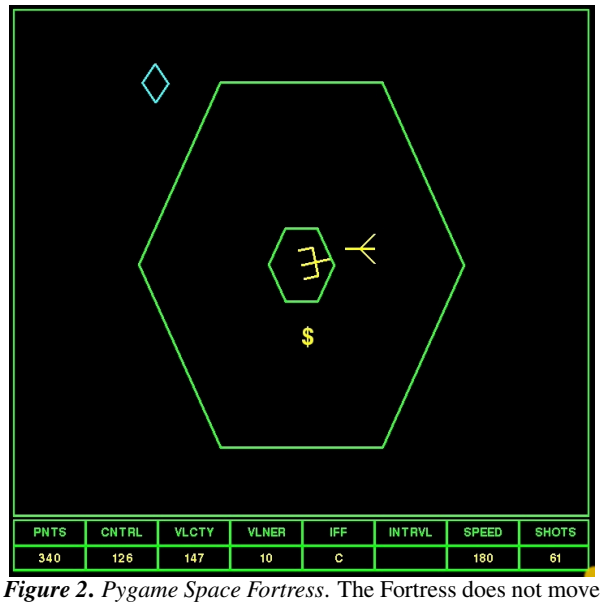


Figure 2. *Pygame Space Fortress*. The Fortress does not move from the center of the screen, but will rotate to face the player's ship (here at the right of the Fortress). The blue diamond is a mine, which is actively 'chasing' the player.

in frictionless space within a 2D plane, with first-order (velocity-based) rotation. If the ship exits the world space by flying off the "edge" of the world, it immediately reenters on the opposite side, giving the game world the topography of a torus. Within this plane are two nested hexagons: a small one to determine the "bounce zone" of the Fortress (if you pass its boundaries, you incur a control point penalty and your velocity vector is reversed), and a large one to encourage the strategy of keeping near the Fortress. By staying within the boundaries of the large hexagon, you gain a small control point bonus once per second. The goal of the game is to maximize *Total* score.

There are many ways of scoring points, each of which entails complex methods involving some combination of memory, timing, and perceptual-motor coordination. For example, *killing* the Fortress requires staying alive while shooting at least ten times with a delay of at least 250 msec between shots. Once the Fortress has been hit 10 times, it is *vulnerable* and can be destroyed by firing two shots within 250 msec of each other. However, if the inter-hit-interval is > 250 msec, then the Fortress is not killed rather, its vulnerability is reset from 10 to zero.

While the Fortress is spinning around trying to shoot the player's ship, mines are appearing every few seconds at random locations and actively chase the ship. The ship cannot destroy the Fortress while a mine is on the screen. However, the Fortress can continue trying to destroy the ship. To make things more difficult, there are two types of mines, friends or

[1] Lec (1962): Auch zum Zögern muß man sich entschließen – "Even the hesitation you have to decide."

foes. When a mine appears, a 3-letter sequence appears in the *IFF* (identify friend or foe) slot at the bottom of the screen (see Figure 2). If these three letters are one of the three 3-letter sequences memorized before the game started, the mine is a "foe mine." If not, then the mine is a "friend mine." When either mine is on the screen the Fortress cannot be damaged and the player runs the risk of having their ship destroyed by bumping into the mine. Killing a friend mine is easy. Just shoot at it. Killing a foe mine is harder as, before it can be shot, it first must be tagged as a "foe" by tapping a special key twice at a interkeystroke interval between 250–400 ms.

While the above is going on, *bonus symbols* appear and disappear below the Fortress at regular intervals. Whenever two '$' symbols appear one after the other, the player has the option of collecting a bonus. If the player presses the "bonus missiles" key, the number of available missiles increases. Alternatively, if the player presses the "bonus points" key, a score increase is received.

With this scoring system, the optimal strategy is thus to always keep flying (but not too fast), shoot the Fortress as often as possible (but not too quickly), stay within the hexagons at all times, always memorize the three 3-digit sequences that appear before each game, deal with mines as quickly as possible, and be sure to grab all the bonuses you can (either points or missiles, as needed).

Pygame Space Fortress (PSF) (Destefano, 2010), which allows for a finer-grained resolution of logged data than prior versions, was used in a longitudinal study to measure the performance of players as they progressed from novices to high-scoring experts. Specifically, Rensselaer students were recruited to play PSF for 31 'hours' each. As in previous research, a *Space Fortress* hour is defined as eight 5-min games, with arbitrary break times between each game. (*Space Fortress* has never had victory conditions — skill was determined by how many points a player could achieve in a fixed amount of time).

In the first session, players played an hour of PSF on a lab computer and had a copy of PSF installed on their laptop. They then played the next five hours, at home, in 1-hr (8 game) increments. At the end of the fifth hour, the home version of the game would lock up. This required the player to return to the lab to play an 'hour', while the experimenter collected log files from the last 5 "home" hours and integrated them into the timeline of data for each player. Prior to leaving the lab, the player's laptop was reset to play and collect data for an additional 5 hours. Hence, each of our 9 players, had a total of 31 hours of play divided into 6 lab and 25 home sessions.

Symbolic Aggregate Approximation

As we were looking for plateaus, dip, and leaps in individual performance, we needed a *Change Point Algorithm* (CPA) to identify points at which scores increased or decreased faster than would be expected by chance. Although many types of CPAs have been developed, we use the Symbolic Aggregate approXimation method (SAX) (Lin et al., 2003). SAX discretizes time-series data by first dividing the time in question into equal segments (in our case, 31 – one for each 'hour' of play), then normalizes the dependent variable and divides it into an arbitrary number of quantiles along the normal distribution function. We chose four quantiles as an appropriate balance to demonstrate possible effects of plateaus, dips, and leaps in performance, without cluttering our analysis with an oversensitive measure. See Figure 3 for an example of applying the SAX algorithm to notional data.

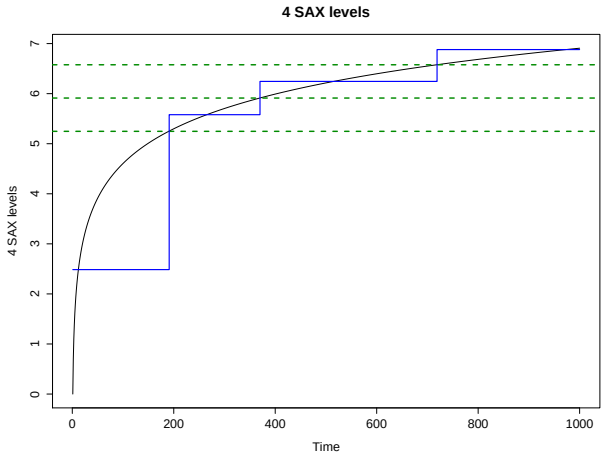


Figure 3. The application of the SAX algorithm to notional data that increases logarithmically over time. The figure superimposes a "perfect" performance curve with the result of analyzing the curve through SAX using an "alphabet size" of 4: The dashed green lines separate the 4 quantiles, or bands of categorization, and the solid blue line indicates which band the current data point falls into.

Player 3534

We begin our dive into the relationship between performance plateaus, dips, and leaps with strategy discovery with a close look at data from Player 3534. Figure 4 shows that player's total score over 31 clusters of eight 5-minute PSF games. For this example, we focus on the transition from the plateau between hours 15 through 18, to the dip in hour 19, to the leap in game 20 (indicated in the figure by the gray bar). Is this merely statistical noise, the player having an "off-day"? A cursory look at the data suggests an affirmative answer, but closer investigation reveals a strategy discovery.

Many measures of performance drop in hour 19 — not only the total score, but *all* of the individual scores: PNTS, VLCTY, CNTRL, and SPEED. Not only that, but bonus captures drop in that hour, and even the accuracy of the player's shots decreases into a lower SAX band. Seeing all these measures of performance falter in this particular hour, it is tempt-

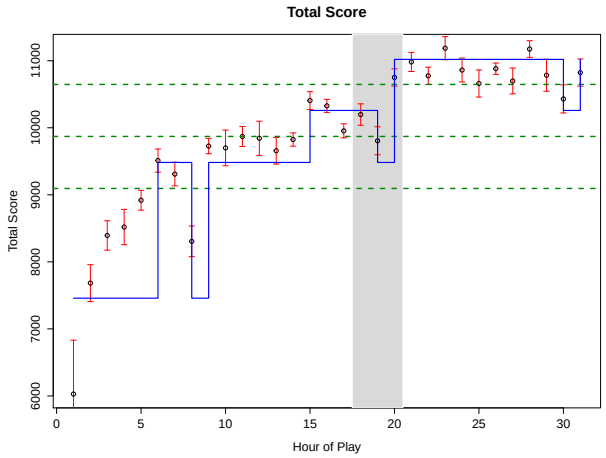


Figure 4. Player 3534's total score over 31 nonconsecutive hours of Space Fortress play. The figure shows a general trend of logarithmic growth, to which we apply SAX analysis to help us identify plateaus, dips, and leaps. As shown in 3 we then subdivide the figure into 4 quantiles. The blue line shows the band that each data point falls into, and the vertical gray bar highlights the three hours of play that is the focus of subject 3534's strategy implementation — a plateau transitioning into another plateau, via a dip and a leap in performance.

ing to dismiss these results as a simple consequence of a poor night's sleep, low motivation, or a distracting environment. Indeed, were we to cease the investigation at this point, we would write off this situation as an anomaly, a simple random fluctuation in performance. However, there are other measures of skill that complicate this simple conclusion.

As shown in Figure 5, Player 3534 was just as adept at destroying the Fortress and shooting mines in hour 19 as in hour 18. This might seem contradictory — given that these are the primary goals of the game, how could it be that the player flew more poorly and was less accurate in shooting, yet still maintained high counts in destroying both mines and Fortresses? The answer is connected to an often overlooked rule of *Space Fortress*:

> The mine appears five seconds after the destruction of either the Fortress or another mine.

Another way of phrasing this rule is that the mine does not and will not appear within five seconds of destroying the Fortress or another mine. This being the case, what happens when the player becomes skilled enough at the game to shoot the Fortress ten times (with at least 250 milliseconds between each shot!), and then double-shoot it to destroy it, *all within five seconds*? The player would then be able to destroy the Fortress over and over again, constantly resetting the internal timer that controls the mine's appearance, and thus *preventing the mine from ever appearing*.

This is a legitimate strategy, and one that might even seem optimal at first glance — after all, it takes a high degree of

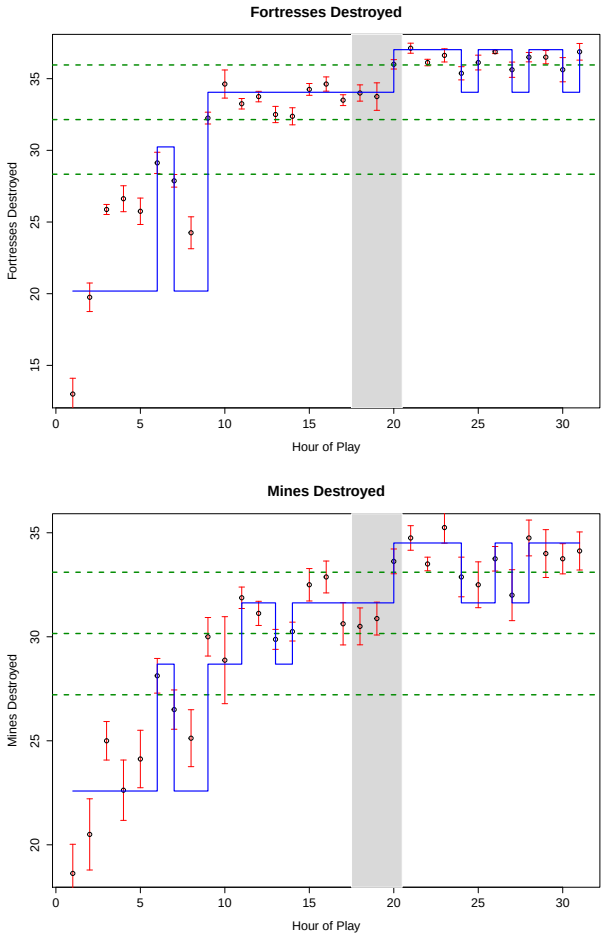


Figure 5. Player 3534's performance over time in destroying Fortresses (top figure) and mines (bottom figure). Unlike other measures of performance (score, accuracy, etc.), 3534's performance in these measures did not drop in hour 19.

skill to pull it off, and the player can now spend all their time and effort interleaving just two subtasks (the Fortress and the bonus symbol) instead of three. Indeed, Player 3534 experimented with it briefly, as shown in Figure 6 (note, in Figure 6, the unit of the x-axis is individual games, not hours). By preventing the mines from appearing, the player misses out on getting points in both the PNTS and SPEED scores, and the trade-off turns out to not be worth it.

What then? It is still desirable to destroy the Fortress as often as possible, but experts will also allow mines to appear. The best strategy our players have found, including 3534, is the following:

1. Fly slowly and as closely to the Fortress as possible without bouncing into the small hexagon.
2. Shoot the Fortress as quickly as possible to increase its vulnerability to the point where it becomes destructible, without shooting it so quickly that its vulnerability resets.

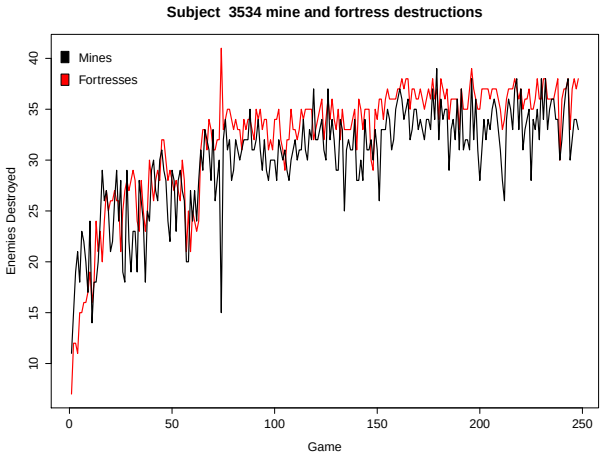


Figure 6. Number of mines and Fortresses destroyed by subject 3534 for each of their 248 games. Note the spike in Fortress kills in Game 74, with its corresponding drop in mine kills. In this particular game, 3534's score dropped by approximately 2000 points (roughly a 20% drop from recent previous games), which was harsh enough for 3534 to abandon the strategy immediately.

3. Wait for the mine to appear, while aiming your ship towards the most probable location for that to happen (typically away from the closest screen edges to the ship).
4. Manage the mine as normal, either shooting it right away if it's a friend, or tagging it as a foe and then shooting it.
5. Double-shoot the Fortress as quickly as possible.
6. Repeat, while also interleaving in the bonus symbol task when appropriate.

Step 3 here is the insight that allows the best players to pull ahead of the rest of the pack. This strategy:

- Requires the player to be fast and accurate enough to be able to shoot Fortresses continuously, *but choose not to do so.*
- Requires a different flight pattern than the one used to follow the strategy of destroying the Fortress as quickly and as often as possible.

To further investigate this phenomenon, we developed a new performance measure: the average number of mines per game that were destroyed *within one second of shooting the Fortress*, shown for player 3534 in Figure 7. Critically, this measure does not drop in hour 19, and then skyrockets into hours 20 and 21. We are now faced with the following collection of premises for our player in regards to hour 19, the dip that precedes the leap:

1. An equivalent number of Fortresses and mines were destroyed.

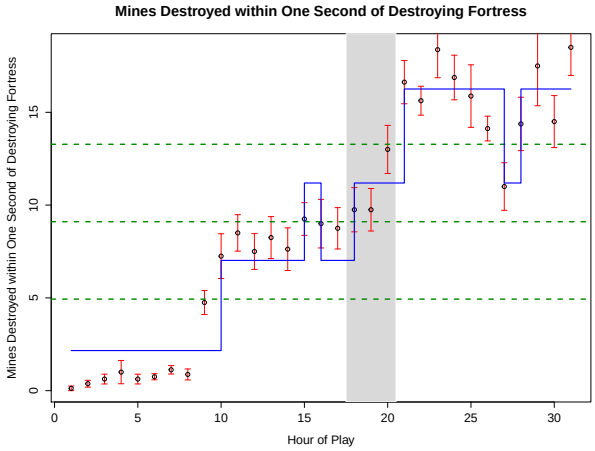


Figure 7. The average number of mines per game that player 3534 destroyed within one second of destroying the Fortress. This is an advanced strategy that takes skill and coordination to perform effectively.

2. The Fortress was destroyed within one second of destroying the mine an equivalent number of times.
3. The player flew less well (lower VLCTY and CNTRL scores).
4. Accuracy dropped — more targets were missed.
5. Then in hour 20, the player shows a dramatic increase in the use of the one-second strategy, with a corresponding leap in total score.

We thus infer from this collection of premises that our player *invented, tested, and implemented* a new flight pattern that allowed them to better prepare for mine onset once the Fortress was vulnerable, and with this new pattern the player was able to exploit the one-second strategy more often.

VLNER to 9

There is an even-more-advanced addendum to this strategy that was discovered by Player 4171. This strategy requires discovering another minor rule of *Space Fortress*:

> Shooting a friendly mine will increase the vulnerability of the Fortress by 1

This rule becomes relevant when combined with two other pieces of information. First, once the initial stock of 100 shots is depleted, firing a shot will incur a 3-point penalty. It can thus be reasoned that each shot is worth some small number of points (slightly less than 3). Second, most mines (70% of them, in fact) are friendly. With these three facts, Player 4171 discovered that approximately three points could be saved with every mine appearance by implementing the following strategy: Shoot the Fortress to a vulnerability of 9 (instead of the typical 10), wait for the mine, and the 70% of the time that the mine is friendly, shooting it will increase

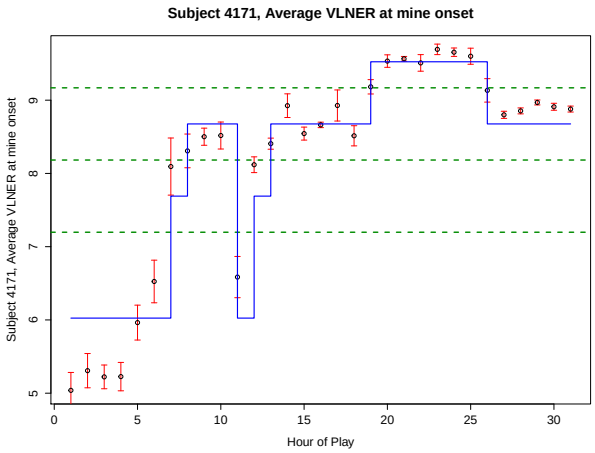


Figure 8. The average vulnerability rating of the Fortress when the mine appears for Player 4171. Note the clear transition from a plateau of 10 to a plateau of 9 once the Player has logged 25 "hours" playing the game. This is a hyper-advanced strategy, one which is well above and beyond the wait-for-the-mine-to-appear strategy detailed previously.

the vulnerability of the Fortress to 10, meaning that it can be immediately destroyed. For the remaining 30% of mines that appear, once they are destroyed it is then necessary to shoot the Fortress once more before destroying it with a double-shot. Figure 8 shows Player 4171 adopting this strategy 25 hours into the study.

Conclusion

We addressed the problem of "where to look" to capture the explorations, trials, errors, and successes that result in new strategies and methods for complex, task performance. This work should provide a strong complement to recent efforts to rethink the study of skill acquisition (e.g., see Anglim & Wynton, 2015; Tenison & Anderson, 2015).

Strategy discovery and method invention become "blurred out" in averaged data, and subtle but critically important signals are lost, such as the signs of strategy invention and the markers for strategy shift that we see by applying SAX analysis to individual *Space Fortress* players.

The plateau, dips, and leaps approach (Gray & Lindstedt, 2016) maintains that extreme expertise involves both slow accretion of low-level skill and bursts of exploration which sometimes leads to new strategies or methods. Such explorations are often punished by drops in performance and abandoned. Other times, as in the cases documented here, success may leap performance to new levels of proficiency.

Acknowledgements

Address correspondence to Wayne Gray – grayw@rpi.edu. The work was supported, in part, by grant N000141310252 to Wayne D. Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

References

Anglim, J. & Wynton, S. K. A. (2015). Hierarchical bayesian models of subtask learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 957–974.

Cinematronics. (1980). Star Castle. Arcade.

Cohen, J. D. & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, *99*, 45–77.

Destefano, M. (2010). *The mechanics of multitasking: the choreography of perception, action, and cognition over 7.05 orders of magnitude* (Doctoral dissertation).

Donchin, E. (1995). Video games as research tools - the space fortress game. *Behavior Research Methods Instruments & Computers*, *27*(2), 217–223.

Ericsson, K. A. & Ward, P. (2007). Capturing the naturally occurring superior performance of experts in the laboratory toward a science of expert and exceptional performance. *Current Directions in Psychological Science*, *16*(6), 346–350.

Fitts, P. M. & Posner, M. I. (1967). *Human performance*. Belmont, CA: Brooks Cole.

Gray, W. D. & Lindstedt, J. K. (2016). League-stepping habits as an escape route (with plateaus, dips, and leaps) from stable suboptimal performance. *Cognitive Science [accepted pending revisions]*.

Lec, S. J. (1962). *Unfrisierte gedanken (unkempt thoughts)* (J. G. (translator), Ed.). New York: St. Martin's Press.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery* (pp. 2–11). DMKD '03. San Diego, California: ACM.

Mané, A. & Donchin, E. (1989). The Space Fortress Game. *Acta Psychologica*, *71*(1-3), 17–22.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652–654.

Tenison, C. & Anderson, J. R. (2015). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Developmentally plausible learning of word categories from distributional statistics

Daniel Freudenthal[1], Julian M. Pine[1], Gary Jones[2] and Fernand Gobet[1]
[1]Department of Psychological Sciences, University of Liverpool
[2]Division of Psychology, Nottingham Trent University

Abstract

In this paper we evaluate a mechanism for the learning of word categories from distributional information against criteria of psychological plausibility. We elaborate on the ideas developed by Redington et al. (1998) by embedding the mechanism in an existing model of language acquisition (MOSAIC) and gradually expanding the contexts it has access to in a developmentally plausible way. In line with child data, the mechanism shows early development of a noun category, and later development of a verb category. It is furthermore shown that the mechanism can maintain high performance at lower computational overhead by disregarding token frequency information, thus improving the plausibility of the mechanism as something that is used by language-learning children.

Keywords: Word class acquisition; Distributional analysis

Introduction

A key issue in understanding how children acquire language is how they build word class categories such as verb or noun. Recent computational work has shown that there is a great deal of information in the distributional properties of different languages that can be used to distinguish between instances of different word class categories in the input. For example, in English, words that are preceded by determiners such as *a* and *the* and followed by auxiliary verbs such as *can* and *will* tend to be nouns, whereas words that are preceded by nominative pronouns such as *I* and *You* and followed by determiners such as *a* and *the* tend to be verbs.

Several approaches to this problem have been proposed, but they tend to focus on building large word classes with high accuracy rather than developing mechanisms that can be plausibly applied by, and fit developmental data from language-learning children. Thus, mechanisms for distributional analysis routinely collect data from large corpora and entire utterances, and make limited contact with the (developmental) child data. In this paper, we explore a more developmentally plausible mechanism by embedding it in an existing model of language acquisition that learns to produce increasingly long utterances, and thus gradually expands the contexts over which statistics are computed in a way that is consistent with children's processing biases.

The two dominant approaches to distributional analysis are those of Mintz (2003) and Redington et al. (1998). Mintz introduces the notion of a 'frequent frame' - a combination of two words with one word intervening (*You* X *A*). Mintz identifies the most (typically 45) frequent frames for a given corpus, and finds that the words that co-occur within a frame tend to belong to the same grammatical category. The approach of Mintz seems intuitively appealing since its relative simplicity means it is well the capabilities of language-learning children. However, it has been argued that it works less well for languages with relatively free word order such as Dutch (Erkelens, 2009) and German (Stumper et al. 2011). It has also been argued that its all-or-none nature makes it overly sensitive to noise - the presence of a small number of items that do not fit the dominant word class for a frame has large effects on classification accuracy (Freudenthal et al. 2013).

Redington et al. (1998) express the contexts in which words occur as vectors containing counts of frequent context words in preceding and following position. Similarity between words is then expressed as the (rank order) correlation between these vectors, and the matrix of correlations is used as input to a cluster analysis. The probabilistic nature of Redington et al.'s approach makes it naturally resistant to noise, and thus less prone to error. However, Redington et al.'s approach requires children to track the frequency of large numbers of words and has been criticized for carrying a high computational overhead (St. Clair et al. 2011), thus making it less appealing as a mechanism that language-learning children might employ. The plausibility of this approach would thus be greatly increased if it could be shown that its computational overhead can be reduced significantly without affecting performance.

While both approaches are intended as mechanisms that language-learning children employ, neither considers the fact that word classes are likely to be gradually built up over time. Thus, they collect statistics from complete utterances and across large corpora, and focus on building large word classes with high accuracy. However, it is debatable whether young children represent statistics that reflect all of the input they have encountered. Children's early utterances are just one or two words long, and only gradually increase in length over a period of years. While children may well attend to longer utterances than they produce, it seems unlikely that they would process entire utterances from a young age. A mechanism that tracks complete utterances may therefore employ statistics that are not available to language-learning children in the early stages of development. A developmentally more plausible mechanism would build word classes gradually by slowly expanding the contexts from which statistics are collected.

This suggestion is further supported by experimental evidence which suggests that children's productive use of words develops at different speeds for different word classes. In particular, production studies suggest that children develop a category of noun earlier than they develop a category of verb (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1999; Tomasello & Olguin, 1993).

The main aim of this paper is to investigate if such developmental effects can be simulated by gradually

expanding the contexts from which statistics are collected. This will be done by embedding variants of Redington et al.'s mechanism in an existing computational model of language acquisition (MOSAIC; Freudenthal et al. 2007, 2010, 2015), that successfully simulates a number of key phenomena in language acquisition. Full details of MOSAIC and the way in which it is trained are provided in Freudenthal et al. (2015).

MOSAIC is a simple learning model that takes as input orthographically transcribed corpora of child-directed speech. Training in MOSAIC takes place by feeding the input corpus through the model multiple times. A key feature of MOSAIC is that it builds up its representation of the utterances to which it is exposed by starting at the right edge of the utterance and slowly working its way to the left. Thus, with each exposure to the input MOSAIC represents increasingly long utterance-final phrases that become increasingly adult-like. The utterance-final bias is MOSAIC's key mechanism for simulating cross-linguistic differences in children's early speech and is thus independently motivated. MOSAIC thus provides a natural framework for testing the notion that word classes develop gradually as the contexts from which statistics are collected are expanded. MOSAIC employs a mechanism for distributional analysis that borrows from the ideas of Redington et al., and thus allows us to investigate the performance of different implementations of the mechanism in a developmental setting.

A first analysis will investigate how well Redington et al.'s mechanism performs when statistics are collected from increasingly long utterance-final phrases. This will be done by training MOSAIC on corpora of child-directed speech and producing output at several stages of development. The output (of increasing average length) from MOSAIC will then be used to derive the counts of context words on which the mechanism is based, and thereby test the performance of the mechanism in a developmentally more plausible setting.

A second analysis will investigate how well a substantially simplified version of Redington et al.'s approach performs. For all words that are to be classified (the *target* words), Redington et al.'s mechanism collects counts for a number (typically the 150 most frequent words for a corpus) of *context* words in preceding and following position. The vectors of counts are then concatenated and similarity between words is expressed as the rank order correlation between concatenated vectors. The mechanism thus collects counts for large numbers of context words, but only uses a limited amount of this frequency information. Here, we investigate the performance of a variant of Redington et al.'s mechanism that disregards token frequency information altogether. Rather than collecting counts for target words in preceding and following position, the approach simply notes the identity of the words in preceding and following position. The context for a given target word is therefore expressed as a list of words (types) rather than a vector of counts (tokens). Similarity is then expressed as a measure formally known as the Jaccard distance: the length of the intersection of two contexts divided by the length of the union - for two words that have been preceded by {a, the} and {the, green} respectively, the similarity is 1/3.

Disregarding token frequency reduces the computational overhead associated with the mechanism considerably: there is no need to collect large numbers of counts, and computation of the Jaccard distance is a mathematically simpler operation than the computation of a rank order correlation. While the simplified mechanism uses less information than the original approach, this may only have limited effects on performance, since the rank order correlation used by Redington et al. only utilizes a limited amount of frequency information.

A final question concerns how well the mechanism is capable of capturing developmental effects in the building of word classes. If the mechanism could be shown to display early emergence of a noun class and late emergence of a verb class, this would increase its developmental plausibility. Such a developmental pattern may arise naturally from MOSAIC's utterance-final bias. Nouns frequently occur near the end of utterances, whereas verbs tend to occur in medial position. Contexts for nouns may therefore register earlier than for verbs, resulting in the emergence of a noun category before a verb category.

In summary, the main aim of this paper is to develop a psychologically plausible mechanism for learning word classes from distributional information by 1. assessing the performance of Redington et al.'s mechanism when gradually extending its access to input in a developmentally plausible way, 2. comparing the performance of the original mechanism with a greatly simplified one that disregards token frequencies, and 3. determining if the developmental variation results in a developmentally plausible pattern of noun and verb linkage. These questions are investigated by applying the two variants of Redington et al.'s mechanism to the increasingly long phrases encoded by MOSAIC models in different stages of development. The main dependent variables are the standard measures of accuracy (number of correct classifications), and completeness (number of classifications), both overall and for nouns and verbs separately. Additionally, a measure of noun richness is computed to track the emergence of noun and verb categories.

The simulations

The first set of analyses was aimed at determining the performance of Redington et al.'s approach in a developmental setting. To this end, MOSAIC was trained on the child-directed speech for each of the 12 children in the Manchester corpus (Theakston et al., 2001). Each model was exposed to the input a total of 50 times. The average length of the utterances that MOSAIC represents increases from zero to approximately five words over training. The model's ability to classify words was assessed at several points in the model's training. As in Redington et al.'s original formulation, target and context words were restricted to the 1000 and 150 most frequent words for each corpus (based on

corpus-wide counts). Target and context words were fixed throughout development.

At each point in development, context vectors for the target words were generated by determining how often the context words occurred in the position directly before or after the target words. A complicating factor here is that MOSAIC does not represent duplicate utterances, and may thus underestimate how often a context and target word co-occur. This was remedied by taking each utterance in the input corpus, determining the largest utterance-final phrase from that utterance that was represented in the model, and adding this utterance-final phrase to the pool of utterances over which statistics were computed. Thus, if the input corpus contained three instances of "it's a dog", and two instances of "that's a dog", and the model only represents the utterance-final phrase "a dog", then 5 instances of the phrase "a dog" were added to the pool of utterances. The pool of utterances was then searched for the target words, and any occurrence of context words in preceding and following position noted. The rationale behind this procedure was to generate accurate counts for an input corpus based on the utterance-final fragments from that corpus that were represented in the model. As training proceeds, and MOSAIC represents more and longer utterances, the counts generated in this manner will become a closer approximation of the counts that would be generated from a corpus-wide analysis.

Vectors containing counts of content words in preceding and following position were concatenated and rank order correlations were computed for every pair of target words. For the current analysis, two words were considered of the same class if the rank order correlation exceeded a certain threshold. That is, the cluster analysis performed by Redington et al. was omitted for ease of interpretation. Results are reported for two levels of the threshold: 0.5 and 0.6. The accuracy of the resulting classification was scored against the (most common) grammatical class assigned to each word based on the morphology (MOR:) line of the transcripts. The main dependent variables were the overall accuracy of the classification and the number of items classed together (number of links). Within class accuracy for nouns and verbs was computed separately by dividing the number of noun-noun (or verb-verb) pairs over the number of pairs containing at least one noun (or verb).

Table 1 shows the results of these analyses, averaged over the 12 different models. The top rows present the data for a threshold of 0.5, and the lower rows a threshold of 0.6. Results are reported for 5 developmental stages, ranging from 36 to 50 exposures to the input. For completeness, the rows labeled 'all' provide data for a corpus-wide analysis, that includes all complete utterances. In line with the current practice in MOSAIC, only declarative utterances were included in the analysis. As can be seen in Table 1, the mechanism manages to maintain high overall accuracy throughout, which tends to be higher in the later stages. The number of links is higher for the threshold of 0.5 than for 0.6, though accuracy scores are not much different.

The main developmental effect in these simulations is that the number of linked items in the corpus-wide analysis is lower than it is for most earlier developmental stages, which link more items with lower accuracy. This pattern seems implausible. Children's early language use has been characterized as relatively rote and lexically specific, and is thought to become more productive with age – the opposite of the pattern in Table 1. It is also apparent from Table 1 that, while the mechanism manages relatively high accuracy for nouns, it is far less accurate in linking verbs, particularly for the early stages – verb accuracy exceeds 0.6 for just 4 out of the 12 cells in Table 1. Taken together, these findings suggest that the mechanism is not sufficiently constrained in the early stages of development.

Table 1: Performance of Redington et al.'s approach at thresholds of 0.5 and 0.6, averaged over 12 models.

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
0.50				
36	1,954	0.68	0.74	0.37
38	4,580	0.73	0.77	0.41
40	5,219	0.72	0.75	0.43
44	3,258	0.79	0.80	0.57
50	2,454	0.81	0.81	0.71
All	2,382	0.82	0.81	0.73
0.60				
36	1,328	0.71	0.76	0.39
38	2,645	0.73	0.78	0.42
40	2,493	0.71	0.74	0.42
44	1,132	0.79	0.79	0.50
50	708	0.82	0.81	0.69
All	667	0.83	0.81	0.70

The reason for the mechanism's initial lack of constraint is that, early in development, contexts for the target words are derived from short, utterance-final phrases, that may only be two words long. This means not only that a limited number of contexts may be available for a given target word, but also that the available contexts may be biased towards preceding or following words. Since vectors for preceding and following contexts are concatenated (and zeros are effectively ignored), this means that the mechanism initially links items on the basis of just preceding or following contexts. With increasing exposure to the input, MOSAIC will represent longer phrases that extend further to the left of the utterance. As a result, the mechanism registers not only more contexts, but a better mix of preceding and following contexts. The mechanism thus becomes more constraining, and links (fewer) items, with higher accuracy – in particular for verbs.

These observations suggest that the practice of concatenating vectors for preceding and following contexts may be inappropriate from a developmental perspective. They also illustrate the value of embedding the mechanism in an existing developmental model of acquisition as the

constraint provided by developmental data may lead to insights that remain hidden in corpus-wide analyses.

The data in Table 2 show that a more plausible developmental pattern results when preceding and following contexts are separated. For this analysis, the rank order correlation was computed separately for preceding and following contexts, and two items were considered of the same category only if *both* correlations were sufficiently high. Table 2 reports results for thresholds of 0.4 and 0.5[1]. The pattern of results in Table 2 differs starkly from that in Table 1. Rather than becoming more constraining with development, the mechanism shows a steady increase in the number of links over development, a pattern that is consistent with children's language use becoming more productive with age. The mechanism also achieves higher accuracy, in particular for verbs – verb accuracy is lower than 60% for just two out of twelve cells.

Table 2: Performance of Redington et al.'s approach with separated vectors at thresholds of 0.4 and 0.5.

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
0.40				
36	50	0.65	0.64	0.33
38	254	0.81	0.85	0.61
40	756	0.81	0.83	0.60
44	1,426	0.84	0.84	0.77
50	1,658	0.86	0.84	0.83
All	1,691	0.86	0.85	0.84
0.50				
36	21	0.74	0.73	0.29
38	107	0.85	0.89	0.60
40	267	0.83	0.83	0.62
44	405	0.86	0.85	0.78
50	433	0.89	0.87	0.86
All	442	0.89	0.87	0.88

The data in Table 2 thus suggest that, when similarity is computed separately for preceding and following contexts, Redington et al.'s mechanism can be applied in a developmental setting, which increases its plausibility as a mechanism that could be employed by language learning children. However, separating preceding and following contexts does not alter the basic statistics over which correlations are computed: counts for the 150 most frequent context words. As was argued in the introduction, the need to collect these counts increases the complexity of the mechanism considerably. Thus, once a child has identified the most frequent context words, it needs to track their frequency before and after all target words. However, since the similarity between words is expressed using a non-parametric measure (i.e. a rank order correlation), much of the frequency information is discarded. The following section explores the mechanism's performance when frequency information is disregarded completely.

Disregarding Token Frequencies

Table 3 shows the results for 12 new MOSAIC models trained on the individual corpora for children in the Manchester corpus. Similarities between words were computed on the basis of the Jaccard distance or the rank order correlation. The rank order correlation was computed as in the previous analysis: based on counts for the 150 most frequent contexts words in (separated) preceding and following contexts. The Jaccard distance is defined as the length of the intersection divided over the length of the union of two sets. As with the rank order correlation, preceding and following contexts are considered separately. The Jaccard distance disregards token frequencies, and thus greatly reduces the computational complexity of the mechanism, as it is no longer necessary to collect counts for context words. Since context is represented as a simple list of word types, there is also no need to restrict context words to the most frequent words in the corpus. Thus, while in practice most context words will be contained in the 150 most frequent words, in principle, any word can act as a context word.

Table 3: Performance of Jaccard distance and rank order at thresholds of 0.2 and 0.45

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
Jaccard				
36	27	0.78	0.83	0.50
38	140	0.83	0.84	0.55
40	370	0.87	0.88	0.68
44	648	0.89	0.86	0.82
50	717	0.91	0.88	0.87
All	738	0.91	0.88	0.87
Rank order				
36	35	0.68	0.76	0.46
38	181	0.8	0.83	0.53
40	488	0.82	0.83	0.63
44	846	0.85	0.84	0.77
50	913	0.87	0.85	0.84
All	925	0.87	0.85	0.85

As in previous analyses, the criterion for considering two words to be of the same category was fixed: 0.2 for the Jaccard distance and 0.45 for the rank order to allow for a meaningful comparison of accuracy and completeness. As can be seen in Table 3, results for the two measures are quite similar, although the rank order measure tends to have lower accuracy – so may be more appropriately set at a slightly higher threshold. Data from Table 2, however, suggests that this may cause completeness to fall below that for the Jaccard

[1] The value of these thresholds differs from those used for the first analysis, and was chosen to enable a meaningful overall comparison in terms of accuracy and completeness.

distance. Nevertheless, the conclusion that can be drawn from the data in Table 3 is that the Jaccard distance performs as well as (if not slightly better than) the rank order measure, and thus that the computational overhead of the mechanism can be reduced significantly without affecting performance.

Noun Richness

The previous analyses showed that Redington et al.'s approach can be applied in a developmental setting provided similarity is computed for preceding and following contexts separately. Its computational overhead can also be significantly reduced without affecting performance by disregarding frequency information. These changes increase the mechanism's plausibility as something that could be employed by language-learning children. This plausibility would be further enhanced if the relative emergence of word classes displayed by the mechanism corresponded to that found in language-learning children. Several authors have argued that children show evidence of a productive noun category before a productive verb category. The relative emergence of noun and verb classes was investigated using a measure of noun richness. Noun richness is a measure of the relative size of the noun and verb category and is computed by dividing the number of noun-noun links over the number of noun-noun plus verb-verb links. A model that shows early emergence of a noun category and late emergence of a verb category would thus show decreasing noun richness. Table 4 shows the development of noun richness for the two models reported in Table 3 – the Jaccard distance at a threshold of 0.2 and the rank order correlation at a threshold of 0.45. For completeness, noun richness for the rank order measure for concatenated contexts (at a threshold of 0.6, data from Table 1) are reported as well. Table 5 lists the number of noun-noun and verb-verb links) for the Jaccard distance.

Table 4: Noun richness scores for Jaccard distance, separated rank order and concatenated rank order.

Runs	Jaccard, 0.2	Rank Order, 0.45	Concatened Rank Order, 0.6
36	0.72	0.68	0.87
38	0.80	0.80	0.89
40	0.80	0.82	0.88
44	0.64	0.71	0.87
50	0.57	0.65	0.82
All	0.57	0.68	0.81

Table 5: Number of noun-noun and verb-verb links for Jaccard distance

Runs	Nouns	Verbs
36	13.75	3.58
38	93.33	16.92
40	254.83	53.67
44	353.5	191.25
50	357.75	256.92
All	364.42	266.92

As can be seen in Table 4, the measures show the highest noun richness score for the second or third developmental phase (run 38/40). Noun richness scores subsequently decrease for all three measures. This decrease is smallest for the concatenated rank order (0.07), intermediate for the separated rank order (0.17) and largest for the Jaccard distance (0.23). Table 5 lists the number of noun-noun and verb-verb links for the Jaccard distance and shows that both the number of noun and the number of verb links increase with development, but verb links are added at a greater rate in the later stages. Results of these analyses thus confirm that computing similarity separately for preceding and following contexts aids verb learning, and leads to a more plausible developmental pattern of noun and verb linkage. This pattern is most pronounced for the computationally simplest measure - the Jaccard distance.

Conclusions

The main conclusion to be drawn from the analyses reported here is that Redington et al.'s (1998) approach can be adapted in such a way that it provides a developmentally plausible account of how children build grammatical word classes on the basis of distributional information. Thus, the mechanism is able to yield developmentally plausible results when the length of the utterances over which it computes statistics is gradually increased. In addition, it was shown that, when statistics are computed separately for preceding and following contexts, the mechanism is able to reach high levels of accuracy for both nouns and verbs, even in early stages of development when few contexts may be available. When similarity is computed for joint (concatenated) preceding and following contexts, the mechanism is too liberal in the early stages, when it tends to link items based on just preceding or following contexts.

It was also shown that disregarding token frequencies of context words substantially reduces the computational overhead of the mechanism without affecting its performance. Thus, rather than representing contexts as vectors of counts for context words, the mechanism simply represents a list of word types that have appeared in preceding and following position. This latter finding is significant, as it has been suggested that the computational complexity of the original mechanism makes it implausible as a mechanism used by language-learning children.

Finally, it was shown that the mechanism can simulate a plausible pattern of noun and verb linkage, with a noun category emerging ahead of a verb category, as is evident in language-learning children. This pattern was least pronounced in the variant that employed concatenated vectors, and most pronounced for the simplest variant (Jaccard distance), thus providing additional evidence that computing similarity separately for preceding and following contexts is particularly helpful for classifying verbs.

Taken together, the analyses reported here suggest that distributional analysis can be a viable approach to the learning of word classes from the earliest stages of

development. Thus, while Redington et al. applied their method to complete utterances and corpora containing millions of word tokens, current analyses suggest that it's possible to accurately classify words on the basis of just a few contexts, provided these are drawn from both the preceding and following position. That is: two words that have overlap in both the preceding and following contexts are very likely to be of the same category, particularly if these are the only known contexts for these words.

The current approach could be argued to combine the best features of Mintz (2003) and Redington et al.'s (1998) approaches. The great strength of Mintz's approach is its ability to classify (large numbers of) items on the basis of very little information (meaning it can be plausibly used by language-learning children), while the probabilistic nature of Redington et al.'s approach means it is naturally resistant to noise, and thus likely to be more accurate. The analyses reported here suggest that a system of intermediate complexity can quickly begin to classify items with high accuracy on the basis of a small number of contexts and gradually expand these contexts as they become available. This necessarily means that the number of items classified in the early stages will be considerably lower than that classified by Mintz's approach. However, while the great strength of Mintz's approach is its ability to classify large numbers of items on the basis of very little information, it is typically applied to corpus-wide statistics and not subject to the developmental manipulations that were studied here, and which provide support for the amended mechanism of Redington et al. Freudenthal et al. (2013) also suggest that the number of classified items may be increased by including utterance boundaries as framing elements.

While disregarding token frequencies for context words makes the mechanism considerably simpler (and hence more psychologically plausible), it retains many of the desirable properties of Redington et al.'s approach. Since similarity is expressed probabilistically, it is naturally resistant to noise and appears better placed to deal with languages whose word order is more flexible than English.

The results presented here also illustrate the value of evaluating distributional learning mechanisms against developmental criteria. Using an established model of language acquisition we show that gradually expanding the contexts over which similarity is computed can result in a developmentally plausible pattern of noun and verb linkage with high accuracy, even when frequency information is discarded. However, the fact that few contexts are available early in development means that distributional statistics can become overly liberal unless similarity in preceding and following contexts is considered separately, a finding, that is likely to remain hidden in analyses that compute statistics over complete utterances and large corpora.

Acknowledgements

Daniel Freudenthal, Julian Pine, and Fernand Gobet are members of the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool, for which support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. This research was supported by ESRC Grant ES/J011436/1.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*, 952-965.

Erkelens, M. A. (2009). *Learning to categorize verbs and nouns.* Unpublished PhD Thesis, Universiteit van Amsterdam, Amsterdam.

Freudenthal, D., Pine, J. M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, 311-341.

Freudenthal, D., Pine, J. M. & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language*, *37*, 643-669.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet, F. (2013): Frequent frames, flexible frames and the noun-verb asymmetry. In: M. Knauf, M. Pauen, N. Sebanz E I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 2327-2332). Austin, TX: Cognitive Science Society.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet. F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61-76.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk (3rd Edition)*. Mahwah, NJ: Erlbaum.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91-117.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.

Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A powerful cue for acquiring syntactic structures. *Cognitive Science*, 22, 425-469.

St. Clair, M.C. Monaghan, P., & Christiansen, M.H. (2010). Learning grammatical categories from distributional cues. Flexible frames for language acquisition. *Cognition*, *116*, 341-360.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). Frequent frames in German child-directed speech: A limited cue to grammatical categories. Cognitive Science, 35, 1190-1205.

Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of Verb-Argument structure: An alternative account. Journal of Child Language, 28, 127-152.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, *8*, 451-464.

A Learned Label Modulates Object Representations in 10-Month-Old Infants

Katherine E. Twomey (k.twomey@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF UK

Gert Westermann (g.westermann@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF UK

Abstract

Despite substantial evidence for a bidirectional relationship between language and representation, the roots of this relationship in infancy are not known. The current study explores the possibility that labels may affect object representations at the earliest stages of language acquisition. We asked parents to play with their 10-month-old infants with two novel toys for three minutes, every day for a week, teaching infants a novel word for one toy but not the other. After a week infants participated in a familiarization task in which they saw each object for 8 trials in silence, followed by a test trial consisting of both objects accompanied by the trained word. Infants exhibited a faster decline in looking times to the previously unlabeled object. These data speak to the current debate over the status of labels in human cognition, supporting accounts in which labels are an integral part of representation.

Keywords: representation, word learning, language acquisition, linguistic relativity

Since Whorf (1940), scientists have been amassing evidence for a link between how we use language and how we represent the world. Much of this evidence comes from crosslinguistic studies which demonstrate that different (grammatical) structures interact with speakers' range of conceptual representations, for example space (Majid, Bowerman, Kita, Haun, & Levinson, 2004), time (Boroditsky, 2001) and color (Roberson, Davidoff, Davies, & Shapiro, 2005). The effects of language on representation have also been observed within a single language (e.g., Johanson & Papafragou, 2016; Son, Doumas, & Goldstone, 2010). For example, Lupyan (2016) asked adults to draw either "a triangle" or "a three-sided polygon" and found differences in the size and orientation of the shapes drawn by the two groups. While the strength and direction of this relationship has long been debated (for a review, see Slobin, 1996), there is consensus that language and representation interact in both perceptual and abstract domains.

Perhaps not surprisingly given the ubiquity of the relationship between language and representation, there is evidence that it begins early in development (for a review, see Robinson, Best, Deng, & Sloutsky, 2012). Recent work indicates that the relationship may begin at least at the same time as the onset of language learning. For example, looking time studies demonstrate that the presence of a label can direct 12-month-old infants' attention to commonalities between category exemplars (Althaus & Plunkett, 2015), and that labels can guide online category formation in infants in the first year of life (Althaus & Westermann, 2016; Plunkett, Hu, & Cohen, 2008). This work illustrates the effect of labels on online processing in these very young infants; however, whether such in-the-moment attention translates to differences in longer-term, learned representations in infants of this age is not clear.

While to our knowledge no study has explicitly examined the effect of learned labels on early representation, there has been a wealth of studies in early label learning and categorization. Infants show their first evidence of word learning from as early as six months (Bergelson & Swingley, 2012) and show signs of mapping novel words to objects in looking time tasks from around 10 months (Mather & Plunkett, 2010; Schafer & Plunkett, 1998; Werker, Cohen, Lloyd, Casasola & Stager, 1998). In the categorization literature, multiple studies show that long-term experience with everyday objects such as pets affects the category representations infants bring to bear in the lab (e.g., Kovack-Lesh, McMurray, & Oakes, 2014). In particular, in a carefully controlled training study Bornstein & Mash (2010) asked caregivers to train their 3-month-old infants with multiple exemplars of a novel object category via daily booklet-reading sessions. After two months, trained infants were invited to take part in a 3D object examining/categorization task, together with another group of untrained infants. The training changed infants' online behavior: relative to untrained infants with no previous experience of the novel category, trained infants showed no learning during a familiarization phase in which they were presented with previously-seen exemplars. Thus, evidence suggests that (a) word learned over time affect infants in-task responses and (b) category representations formed over time also affect these responses. Here, we ask about the relationship between the two; specifically, whether longer-term learning of labels in very early development can shape object representations.

Thus, in a training/familiarization study we asked parents of 10-month-old infants to train their children with two novel toy objects at home over a week (seven sessions), labeling one object (*labeled* object) but not the other (*non-label* object). We then recorded infants' looking times in a familiarization/preferential looking task using an eye-tracker. First, images of the two objects were presented individually, in silence, on a computer screen. We hypothesized that if object labels affect early object representations, we should observe differences in infants'

looking times to the previously labeled and the non-labeled stimuli. We did not anticipate that these beginner word learners would show robust label-object associations (c.f., Horst & Samuelson, 2008). However, given that even partial knowledge of the mapping could influence infants' looking times (Yurovsky, Fricker, Yu, & Smith, 2014), after familiarization infants also saw a single preferential looking test trial in which both images appeared simultaneously, accompanied by the label.

Methods

Participants

Twenty-four 10 month-old infants (12 girls; M = 10m, 23d; SD = 14.15d, range = 9m, 26d – 11m, 13d) participated. All infants were typically developing and monolingual English learning with no family history of color blindness. Data from an additional four infants were excluded for fussiness (as defined by failure to start or complete the eyetracking task due to excess movement and/or crying); 2), experimenter error (1), and low eyetracker sample rate (< 35%; 1). All but three participants had completed all seven training sessions (7 sessions: 21; 6 sessions: 3). All participants returned for the second session approximately a week after the first (6 days: 2; 7 days: 19; 8 days: 3). Families were recruited by contacting caregivers who had previously indicated interest in participating in child development research. Caregivers' travel expenses for both visits were reimbursed and infants were given a storybook for participating.

Stimuli

Play sessions. 3D stimuli are depicted in Figure 1, and consisted of two age-appropriate wooden toy objects (castanets and two wooden balls joined with string), chosen because they are novel to 10-month-old infants (Fenson et al., 1994). Objects were approximately equal in size and were painted either red or blue using non-toxic paint. The label was *tanzer*, a pseudoword selected because it is plausible in English and was used in a previous developmental study (Horst & Twomey, 2012).

Looking time task. Familiarization stimuli consisted of digital photographs of the individual training objects presented centrally on a white background. Test stimuli consisted of photographs of the training objects presented side-by-side on a white background. The auditory stimulus for the test phase consisted of the phrase *Look! A tanzer!* spoken by a female speaker from the local area and recorded and edited for timing and clarity in Audacity 2.0.6. The phrase onset was at 4000ms, label onset at 5127ms, and label offset at 6000ms. Calibration and attention getter stimuli were a short video of a bouncing cartoon bird, accompanied by a jingling sound.

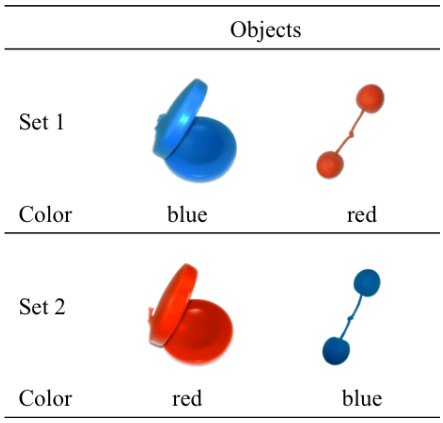


Figure 1: Stimuli used in the current study.

Procedure and Design

Visit 1: Play session. Objects' color and label were counterbalanced between participants such that for each object, half the infants received a blue exemplar and half received a red exemplar, with the constraint that each infant received one red and one blue object. Each exemplar served as the label object for half of the infants (*label* condition) and the no-label object for the other half of the infants (*no-label* condition).

First, the experimenter showed the caregiver the two objects and asked them whether their child had similar toys at home. Substitute items were available, however no child had prior experience of the objects. The experimenter then explained that she would demonstrate a play session, with the goal of teaching the infant a word for one of the objects. She then asked the caregiver to conduct a similar play session for three minutes, every day for a week and explained that they would be invited to return to the lab after seven days to take part in a simple looking-time study, which would involve the child watching static images on a computer screen.

The play session took place in a quiet, infant-friendly room with the caregiver present at all times. Before the session began, the experimenter emphasized that caregivers should not invent a name for the no-label object: only the label *tanzer* should be used, and only in reference to the labelled object. The experimenter then sat opposite the infant on the floor and introduced both toys by holding them in front of the infant and allowing the infant to take the toys in their own time. While the infant was looking at the toy the experimenter referred to them using a label or a pronoun as appropriate, for example "*Look, a tanzer!*" (label), "*Look at this*!" (no-label). During the course of the play session the experimenter explained to the caregiver that they should encourage their child to interact with both toys for an approximately equal amount of time, and that their child should be allowed to play with both toys at the same time (to encourage comparison; Oakes, Kovack-Lesh, & Horst, 2009). Infants heard the label approximately twice every fifteen seconds. After the play session caregivers were given

the toys, written instructions and a sticker chart on which to record their play sessions.

Visit 2: Looking time task. The looking time task took place in a quiet, dimly-lit testing room. Children were seated on their caregiver's lap 50-70 cm in front of a 21.5" 1920 x 1080 computer screen at a 60 cm viewing distance. A Tobii X120 eyetracker located beneath the screen recorded the child's gaze, and a video camera above the screen recorded the caregiver and child throughout the procedure. Caregivers were instructed not interact with their child or look at the screen during the task to avoid biasing their child's behavior.

The eyetracker was first calibrated using a five-point infant calibration procedure. We displayed an attention-grabbing animation in the four corners and center of a 3 x 3 grid on a grey background accompanied by a jingling noise, and recorded infants' orientation to it with a key press. Calibration accuracy was checked and repeated if necessary (n = 1).

The attention-grabbing stimulus then appeared in the center of the screen. Immediately after the infant oriented towards the attention-grabbing stimulus, the experimenter began the familiarization phase using a keypress. Familiarization stimuli were presented individually in silence for 10 s. Infants saw eight identical images of the previously-labeled stimulus and eight identical images of the no-label trials, interleaved. Whether the label or no-label image was shown first was counterbalanced between children. Each trial was immediately followed by the attention-grabbing stimulus. Subsequent trials were advanced manually by the experimenter once the infant had reoriented to the screen, or began automatically after 5s.

Immediately following the familiarization trials a single test trial was presented in an identical manner. Left-right positioning of the objects (castanet/ball and label/no-label) was counterbalanced between children. The test trial was 12 s long, with auditory stimulus beginning at 4000 ms, label onset at 5172 ms and offset at 6000 ms. The test trial continued for 6 s post label offset to allow sufficient time to exhibit a response (Bergelson & Swingley, 2012).

Coding and data cleaning

Timestamps for which the eyetracker failed to reliably detect either eye were excluded. Familiarization stimulus AOIs were centered on the single image and measured approximately 950 by 700 pixels. Test AOIs were centered on the two images and measured 766 by 435 pixels. Individual gaze samples were numerically coded (1 = AOI look, 0 = background look) to create a raw looking time measure.

Results

Familiarization phase

Individual looking time samples (numerically coded; 1 = AOI look, 0 = background look) were submitted to a binomial mixed effects model using the R package lme4 (version 1.1-10; Bates, Mächler, Bolker, & Walker, 2015). The model included fixed effects of trial (1 – 16) and label (centered; 0.5 = label, -0.5 = no-label), a trial-by-label interaction term, and random intercepts for participant and stimulus (this was the maximal model that converged; Barr, Levy, Scheepers, & Tily, 2013).

As is typical in looking time studies, target looking decreased across familiarization (main effect of trial: beta = -0.032, SE = 0.0036, z = -8.83, p < .0001), and did so fastest across training for no-label stimuli (trial x label interaction: beta = 0.014, SE = 0.0072, t = 1.99, p < .05). Specifically, mean looking per trial decreased from 5422ms on the first label trial to 2887ms on the last label trial, but 5696ms on

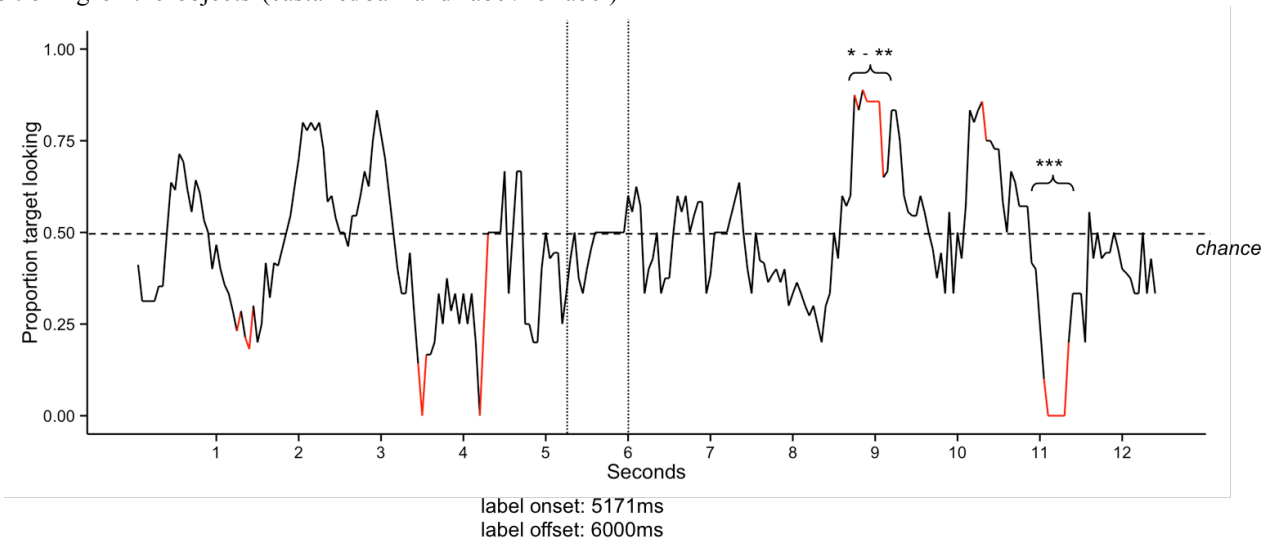


Figure 3: Proportion looking to the labeled target per 50 ms time bin during the test trial. Looking significantly different from chance (0.50) is indicated by a red line; systematic looking is indicated by asterisks; * p < .05, ** p < .01, *** p < .001.

the first no-label trial to 2733ms on the last no-label trial. Whether the stimulus has been previously trained with a label had no independent effect on overall looking times (main effect of label: beta = -0.033, *SE* = 0.069, *t* = -0.47, *p* = .63; cf. Capelier-Mourguy, Twomey & Westermann, under review).

Test phase

To establish whether infants responded to the auditory label we calculated the proportion of target looking out of all AOI on the test trial. Proportion target looking for each 50 ms time bin is depicted in Figure 3. Systematic fixation of target/distractor was defined where looking differed significantly from chance (0.50) for at least 150 ms (Holmqvist et al., 2011).

We defined the prelabeling phase from 0 – 6367ms, which includes 367 ms from label offset for preparation of the saccade (Swingley, 2009), and the postlabeling phase as 6368 ms to 12,000 ms. Infants showed no overall systematic preferences during either the prelabeling ($t(23) = -0.64$, $p = .53$; one-sample, chance = 0.5; all tests two-tailed) or postlabeling ($t(21) = -0.13$, $p = .90$; note two participants did not provide data in the postlabeling phase), and no difference between these phases ($t(21) = -0.65$, $p = 0.52$). However, as can be seen in Figure 3, fine-grained inspection of the unfolding of looking over time reveals that infants showed a late response to the label, increasing their target looking to above-chance levels 2.5 s after the label offset, in line with recent work examining 10-month-old infants' label disambiguation (Mather & Plunkett, 2010).

Discussion

In the current study, 10-month-old infants were trained over a week with two 3D objects and taught a novel label for just one of them. Infants were subsequently familiarized in silence with these objects in a looking-time study, and finally presented with a preferential looking trial in which they saw both objects accompanied by the label. Overall, infants maintained interest for longer in label than no-label stimuli. Critically, given that all 3D object training and familiarization for each object was identical except for the presence of the label during the play sessions, this study demonstrates that a learned label can affect infants' object representations. For the first time, then, this study traces the start of the relationship between language and representation to infants less than a year old.

Several studies have shown that infants' experience outside the lab affects their in-task performance. For example, Hurley & Oakes (2015) showed that 4-month-old infants with experience of pet cats and/or dogs at home paid more attention to the diagnostic features of cat and dog stimuli (i.e., the head) than their peers without pets at home. However, there was no difference in visual inspection strategies when stimuli were faces or vehicles. In these "pet" studies, stimuli were drawn from infants' everyday environment. Participants therefore came to the lab with substantial and prolonged experience (see also Perry, Samuelson, & Burdinie, 2014). However, when infants were trained with novel stimuli (e.g., Bornstein & Mash, 2010) similar effects of prior experience were found. Our results offer converging evidence that prior experience affects in-task behavior, showing that even a relatively small amount of additional linguistic experience – in this case, as little as 21 minutes over seven days – can affect infants' responses in-task.

However, infants in our study faced an added challenge. Not only did we ask them to encode two new objects, but we also asked them to attach a new label to one of them. Infants of this age are at the very beginnings of language acquisition (Fenson et al., 1994) and exhibit more fragile word learning than older children; for example, in a recent preferential looking study this age group increased looking towards the referents of frequently-heard words around 2.5 – 5.0 s after hearing labels (Mather & Plunkett, 2010). We replicate this finding: infants looked to the target 3 s after label offset. However, both objects were novel before training: our results demonstrate that 10-month-olds will treat a novel word as familiar after only a week of experience. Importantly, infants were not reminded of the novel word during their second visit, and did not receive a training session on that day. Thus, we can be confident that they responded to the label based on retrieval of the association from long-term memory; to our knowledge, this is the first study to demonstrate long-term novel word learning in infants of this age.

Mechanism

These results have several possible interpretations. Perhaps the most straightforward explanation is that the prior experience during training may have increased infants' attention to the labeled object at test. That is, despite our request to the contrary, infants may have spent more time engaging with one of the two objects during training. Infants of this age have a well-documented novelty preference (Fantz, 1964); that is, they will engage preferentially with items that are new (and therefore interesting), relative to items that are familiar (and therefore less interesting). The object infants had played with the most would be more familiar, and would therefore elicit less attention (lower looking times) during familiarization. Although we cannot rule out this possibility, we consider this unlikely. On this account, the fact that caregivers were teaching their child a label could unconsciously – or indeed consciously – bias them towards the labeled object, particularly as object names dominate maternal speech (Fernald & Morikawa, 1993). If anything, then, we would expect infants have more experience of the labeled than the no-label object. If so, the no-label object would have been most novel and should have elicited the longer looking times. In fact, we found the opposite.

Rather than resulting from the amount of experience infants received with each object, the patterns in looking time may stem from differences in infants' representation of the labeled versus the no-label objects. In fact, looking time

differences were driven by a drop in looking to the no-label stimulus just over halfway through familiarization. This finding suggests that relative to the label stimulus, the no-label stimulus was more familiar to infants, begging the question of why the label stimulus should be most novel, and speaks to the recent debate concerning status of labels in object representations.

On the *invitations-to-form-a-category* (henceforth *invitations*) account, infants have an innate bias towards linking word representations with category representations. Importantly, the two are represented qualitatively differently. While object representations are learned from perceptual experience, labels play a top-down, supervisory role, serving as high-level, conceptual markers of category membership (Waxman & Markow, 1995). When a label becomes associated with an object it acts as an "invitation to form a category" indicating that it should also be applied to new items providing they are sufficiently similar; however labels are not integrated into object representations. In sharp contrast, the *labels-as-features* account assumes that labels are simply features of perceptual object representations (Gliozzi, Mayor, Hu & Plunkett, 2009; Sloutsky & Fisher, 2004). Labels are integrated into object representations in the same way as color or shape; thus, the word *strawberry* and the color red will have the same qualitative status in a speaker's representation of the fruit. More recently, Westermann & Mareschal (2014) argued that labels are represented separately from objects, but critically, in the same representational space. On this *compound-representations* account, labels are neither simple features of object representations nor a different type of structure altogether. Instead, labels are perceptual representations which become associated with object representations over time.

In the context of the current study, the invitations account predicts that there should be no difference in looking to the labeled and no-label stimuli, because labels do not affect individual object representations[1]. This is not what we found. On the labels-as-features account, if an object appears consistently with a label, the label is incorporated into the object representation. When the object appears without the label, then, what infants experience is different from the learned representation. Thus, infants' greater looking to the labeled stimulus is in line with the labels-as-features account. However, the compound representation account also makes this prediction: when a compound label-object representation is formed, a missing label will increase the novelty of the object relative to its representation. Note that we are not claiming that the learned label renders the representation more robust, but rather that the label becomes incorporated into (invitations account) or associated with (compound representations) the representation over learning. Given the present data, then, we must remain agnostic as to whether labels are encoded as part of or separately from object representations. However computational work is underway to tease apart the labels-as-features and compound-representations accounts (e.g., Capelier-Mourguy et al., 2016).

Overall, the current study provides converging evidence that long-term learning experience has a marked effect on in-lab behavior – a critical point for our interpretation of the wealth of existing infant studies. We also demonstrate word learning from limited exposure in infants at the earliest beginnings of the language acquisition journey. Most importantly, however, we show for the first time that language has long-term effects on nonlinguistic representation before the end of the first year, demonstrating that language affects the way we see the world right from its very beginnings.

[1]Although this account does predict that learned labels shape *category* representations; we are currently testing this prediction.

Acknowledgments

This work was supported by the ESRC International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

References

Althaus, N., & Plunkett, K. (2015). Categorization in infancy: labeling induces a persisting focus on commonalities. *Developmental Science*.

Althaus & Westermann (2016). Labels constructively shape categories in 10-month-old infants. *Journal of Experimental Child Psychology,* doi:10.1016/j.jecp.2015.11.013

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1 – 48.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bornstein, M. H., & Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, *81*(3), 884–897.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

Capelier-Mourguy, Twomey & Westermann (2016). A neurocomputational model of the effect of learned labels on infants' object representations. *38th Annual Cognitive Science Society Meeting, Philadelphia, Pennysylvania, USA.*

Fantz, R. L. (1964). Visual experience in infants: Decreased attention familar patterns relative to novel ones. *Science*, *146*(668-670).

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early

communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), R5–+.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*(3), 637–656.

Gliozzi, V., Mayor, J., Hu, J. F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

Horst, J. S., & Twomey, K. E. (2012). It's taking shape: shared features influence children's novel noun generalizations. *Infant & Child Development*.

Hurley, K. B., & Oakes, L. M. (2015). Experience and distribution of attention: Pet exposure and infants' scanning of animal images. *Journal of Cognition and Development*, *16*(1), 11–30.

Johanson, M., & Papafragou, A. (2016). The influence of labels and facts on children's and adults' categorization. *Journal of Experimental Child Psychology*, *144*, 130–151.

Kovack-Lesh, K. A., McMurray, B., & Oakes, L. M. (2014). Four-month-old infants' visual investigation of cats and dogs: Relations with pet experience and attentional strategy. *Developmental Psychology*, *50*(2), 402–413.

Lupyan, G. (2016). The paradox of the universal triangle: concepts, language, and prototypes. *The Quarterly Journal of Experimental Psychology*, *0*(ja), 1–69.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108–114.

Mather, E., & Plunkett, K. (2010). Novel labels support 10-month-olds' attention to novel objects. *Journal of Experimental Child Psychology*, *105*(3), 232–242.

Oakes, L. M., Kovack-Lesh, K. A., & Horst, J. S. (2009). Two are better than one: Comparison influences infants' visual recognition memory. *Journal of Experimental Child Psychology*, *104*(1), 124–131.

Perry, L. K., Samuelson, L. K., & Burdinie, J. B. (2014). Highchair philosophers: the impact of seating context-dependent exploration on children's naming biases. *Developmental Science*, *17*(5), 757–765.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681.

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*(4), 378–411.

Robinson, C. W., Best, C. A., Deng, W., & Sloutsky, V. M. (2012). The role of words in cognitive tasks: What, when, and how? *Frontiers in Psychology*, *3*.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." *Rethinking Linguistic Relativity*, *17*, 70–96.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology-General*, *133*(2), 166–188.

Son, J. Y., Doumas, L. A., & Goldstone, R. L. (2010). When do words promote analogical transfer? *The Journal of Problem Solving*, *3*(1), 4.

Swingley, D. (2009). Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language*, *60*(2), 252.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391.

Whorf, B. L. (1940). *Science and linguistics*. Bobbs-Merrill.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*(1), 1–22.

An experimental study on the observation of facts in explanation reconstruction

Hitoshi Terai (terai@fuk.kindai.ac.jp)
Faculty of Humanity-Oriented Science and Engineering, Kindai University,
Kaya no Mori 11-6, Iizuka, Fukuoka, 820-8555, Japan

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa, Aichi, 464-8601, Japan

Naohiro Toyama (toyama@cog.human.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa, Aichi, 464-8601, Japan

Abstract

Explanation reconstruction performs a crucial role not only in the progress of science but also in educational practices and daily activities, including the comprehension of phenomena. In this study, we conducted experiments to examine the factors that facilitate shifts in explanations. We focused on the transition of attention on a key fact that contradicts an initial explanation and has a central role in its reconstruction. We used a short story as an experimental material in which participants first constructed an initial explanation and then reconstructed it. In the experiment, we controlled the time of presentation of the key fact (bottom-up condition), reflective thinking (top-down condition), and the two together (bidirectional condition) to facilitate understanding of the explanatory shift. The experimental results are summarized as follows. First, when the prior explanation was rejected, attention to the key fact was inhibited although a new explanation was required. Second, the successful group increased their attention on the key fact just before the explanatory shift. Third, protection of the preceding explanation with unobserved facts was inhibited by guiding the participants ' attention toward the key fact. Finally, although the initial explanation was not completely shifted, a explanatory pre-shift was achieved by activating reflective thinking with attention to the key fact.

Keywords: eye-movement analysis; explanation reconstruction; explanatory shift; reinterpretation

Introduction

Scientific activities aim to understand the world in two ways: descriptive and explanative (Simon, 2000). Descriptive understanding describes the nature and characteristics of phenomena by observations and experiments, whereas explanative understanding reveals the mechanisms behind the phenomena and the reasons why such phenomena occur.

Through the history of science, first descriptive understanding is usually established and then explanative understanding is investigated.

On the basis of historical facts, we can confirm many cases where the explanation for a certain phenomenon was completely changed because the structures of the explanation and the concepts of objects were essentially shifted. Such cases are generally observed in the history of science, e.g., the shift from the caloric theory to the oxygen theory and the transition from Newton's traditional theory to Einstein's relative theory(S. F. Mason, 1953).

Explanatory Shift

In the current study, we call the phenomena that is an explanation fundamentally reconstructed from a preceding explanation "an explanatory shift." We define the explanatory shift as a reconstruction of an original explanation that includes changes to both the structure of an explanation and the meaning of facts related to the explanation. Such changes neither partially expand a preceding explanation nor refine it. A preceding explanation is defined as "an initial explanation." We also call a reconstructed explanation "a shifted explanation." The explanatory shift is characterized by the perception of "a key fact" that must be fundamentally revised from the initial explanation. Namely, in the before and after stages of an explanatory shift, the meaning(s) of the key fact are fundamentally different.

As an example, consider the change of the caloric theory to the oxygen theory. Initially, in the caloric theory, burning was explained as the release of calories. After an inconsistency was observed in this explanation, i.e., the increase in weight after burning, a new explanation was required. The change of the explanation from the caloric theory, i.e., burning releases calories, to the oxygen theory, i.e., burning is associated with oxygen, was established by reinterpreting the key fact, i.e., the increase in weight by burning(S. F. Mason, 1953).

The Difficulty of an Explanatory Shift

It is difficult to shift the mind to a new explanation from an initial explanation by reinterpreting a key fact when it requires overcoming an accepted fact that in the past seemed to be common sense. For example, it has been repeatedly demonstrated that a confirmation bias leads to gathering positive instances that confirm an established hypotheses (e.g., Wason, 1960) and initial hypotheses are typically maintained against anomalous data(e.g., L. Mason, 2001). Therefore, even when an initial explanation is rejected and a new explanation is required, the initial explanation is protected by local modification and slight expansion with non-existent facts rather than the acceptance a new understanding by an essential change of the initial explanation.

In the oxygen theory, the connection of burning with oxygen explains a key fact; however, in the caloric theory, there

was only an assumption to support the release of phlogiston that had negative weight, which was part of the accepted theory even though the negative-weight phlogiston had not yet been observed or proved. Moreover, in the oxygen theory, as the explanation shifted from the caloric theory, the relationship of oxygen to burning could be accurately reinterpreted by overcoming an accepted common sense fact that had been accepted since the times of ancient Greece: "fire is an element."

Difficulties exist in such essential reconstructions of explanations. However, few studies have examined the process of the explanatory shift.

Transition of Attention toward an Explanatory Shift

In our previous research(Terai, Miwa, & Matsubayashi, 2012), we used eye-movement analysis to examine the transition of attention to a key fact that was contradictory to an initial explanation.

In the experiment, a short story is set in a town with only two barbershops (Figure 1). A character is looking for a barber and must pick one of the two shops. Barber A's staff has unkempt hair, while barber B's staff has hair that is beautifully cut. Initially, participants typically select barber B on the basis of the following naïve explanation provided as the initial explanation: "A barber with beautiful hair is very skilled." However, a new explanation is now required by the character for selecting barber A. The fact "barber A's hair is messy and barber B's hair is neat" contradicts the initial explanation. Therefore, in the story, the key fact that must be reinterpreted is "barber A's staff has unkempt hair and barber B's staff has neat hair." The reconstructed explanation from Gardner (1978) is "each does the other's hair because there are only two shops in town; therefore, barber A's staff, who did barber B's staff's hair, is more skilled." In the shifted explanation, the key fact becomes evidence for selecting barber A, but in the initial explanation, it is evidence for selecting barber B. The meaning of the key fact has completely shifted with the transition from the initial to the shifted explanation.

The experimental results are summarized as follows. First, the attention to the key fact, which is inconsistent with the initial explanation, declines. Second, the successful group increases their attention on the key fact just before the explanatory shift. Finally, we could facilitate the explanatory shift by both highlighting the key fact to lead the participants' attention to the key fact and instructing participants that the highlighted fact was crucial for finding the right explanation.

Relation between the Observation and the Explanatory Shift

The results of our previous research (Terai et al., 2012) show that the explanatory shift is facilitated by leading the participants' attention to a key fact by highlighting it and then instructing that the highlighted fact is crucial for finding the right explanation. However, it is unclear whether it is enough to only increase the chance of gathering information about the key fact or, in addition to this, there must be an intention to reconsider the meaning of the key fact in order to facilitate an explanatory shift.

Grant and Spivey (2003) discussed the relations between achieving insight and attention by inducing gaze in participants. In the experiments, the X-ray problem (Duncker, 1945), consisting of a tumor, healthy tissue wrapping around the tumor, and outer skin, was presented to the participants. They were required to solve how to destroy the tumor without harming the healthy tissue[1]. The results of the experiments revealed that problem-solving performance was increased by inducing the participants' attention (gaze) to the area in which the successful participants increased attention. Therefore, it is suggested that an explanatory shift is facilitated by having the opportunity to obtain the key fact from the bottom-up.

On the other hand, taking time to only see the key fact might be insufficient to shift an initial explanation. It is known that seeing is affected by knowledge held by the investigator, even in scientific activities where objectivity is required, i.e., the theory-ladenness of observation(Duhem, 1914; Hanson, 1958; Kuhn, 1962) . For example, in astronomy, Tycho and Simplicius saw a mobile sun, whereas Kepler and Galileo saw a static sun(cf., Hanson, 1958). This means that facilitation of an explanatory shift might require reflective thinking as a top-down process, which tries to reconsider a phenomenon from a different perspective.

Purpose

Our experimental purpose is to reveal the factors that facilitate an explanatory shift. We focused on the observation of a key fact, which is required for reinterpretation to achieve explanatory shift.

In our previous study, the attention on a key fact was inhibited, and then attention on the key fact was increased just before the explanatory shift as discovered through eye-movement analyses. However, the previous study did not sufficiently discuss differences of attention on the key fact between successful and unsuccessful groups. The degree of accuracy of eye-movement measurement is one of the reasons. Therefore, in experiment 1, we capture differences in the amount of attention on the key fact between successful and unsuccessful groups through more precise eye-movement measurement. Moreover, we reconfirm that moving the attention to the key fact is a robust process.

Then, in experiment 2, we discuss the factors that facilitate an explanatory shift by controlling both the bottom-up and the top-down processes: the opportunity to obtain the key fact and reflective thinking, which involves the reconsideration of a phenomenon again from a different perspective, respectively.

[1]The correct solution is the simultaneous convergence of low intensity X-rays on the focal tumor.

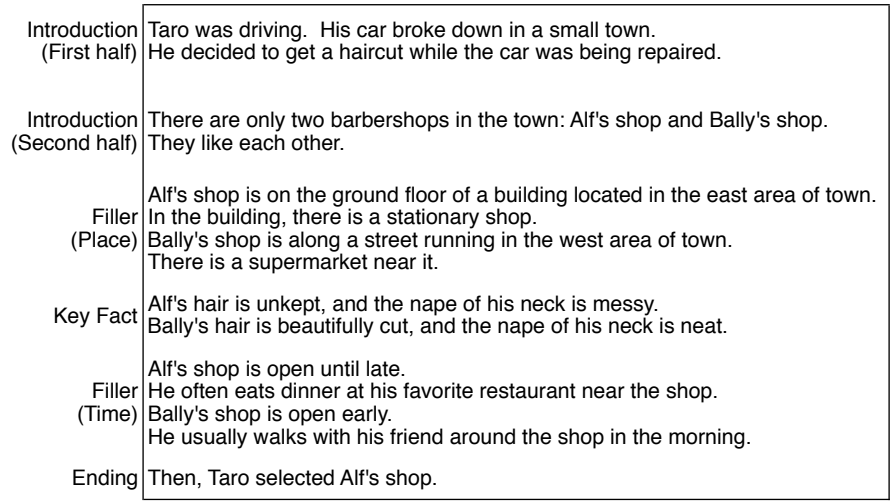


Figure 1: The barber task (modified from the original barber task in Gardner (1978))

Note. The story in the above box was presented to the participants in experiment 1. The story consists of an introduction, facts of regarding the place, barbers, and time, and an ending.

Experiment 1

Methods

Participants Forty-two undergraduates participated in this experiment.

Task Figure 1 shows the barber story task used in this experiment as introduced in our previous research(Terai et al., 2012).

Procedure Experiment 1 consists of two phases: the initial explanation phase and the reconstruction phase. In the initial explanation phase, the participants read the story (Figure 1) without the ending on a computer screen while thinking about which barbershop to select. Their task was to construct an explanation for their decision. They reported their constructed explanation by using a computer keyboard. The initial explanation phase was followed by the reconstruction phase in which the initial explanation was rejected, and they were required to reconstruct their explanation of the story. They reported their initial explanation and reconstructed explanation at their own pace. After reporting their explanations, they received the following message: "Since there is another reasonable explanation for this quiz acceptable to all, please reconsider another creative explanation." They were told that there was an evaluator in another room connected by the Internet, even though no such evaluator existed, and the same message was always returned. The reconstruction phase continued for 15 min, and then the data were analyzed until they reached the shifted explanation.

The experiment was conducted individually, and participant eye movements were recorded using a Tobii T60 eye tracker. In the experiment, we used a chin rest and required the participants to inhibit their head movement to increase the precision of gathering eye movement.

Results

Ten of 39 participants (three participants were excluded because of equipment trouble) constructed the initial explanation in the initial explanation phase and then constructed a shifted explanation in the reconstruction phase (the successful group). On the other hand, twelve participants could not construct the shifted explanation in the reconstruction phase, although they constructed the initial explanation in the initial explanation phase (the unsuccessful group).

In the following analysis, we focus on fixation ratios of each of the facts (introduction, place, key fact, time, and ending) in both the successful and the unsuccessful groups. Average fixation ratios of each fact were normalized by the number of characters in each paragraph corresponding to each fact.

Unsuccessful Group Figure 2 (a) shows the average fixation ratios in the first 5 min of the reconstruction phase in the unsuccessful group. To examine differences in the fixation ratios among each fact, a one-way analysis of variance (ANOVA) showed a significant main effect of the facts. The results of the multiple comparisons using Holm's method showed that the fixation ratio of the key fact was significantly lower than those of the place and time facts ($MSe = 0.00, p <$ .05).

Next, we examined whether there were any differences between the fixation ratios of each fact and the baseline calculated with assumption that each fact was seen in unbiased way. A t-test indicated that the fixation ratio of the key fact was lower than the baseline ($t(11) = -5.36, p < .01$). Moreover, a t-test indicated that the fixation ratio of the place fact was greater than the baseline ($t(11) = 2.24, p < .05$).

Successful Group An average time of constructing the shifted explanation in the successful group was 219.3 s ($SD =$ 179.0). Figure 2 (b) shows the average fixation ratios up to 30 s before constructing the shifted explanation in the successful group. A one-way ANOVA showed no significant main effect of the six facts ($F(5,40) = 1.36, ns$) whereas a t-test indicated that the fixation ratio of the key fact was lower than the baseline ($t(8) = -4.39, p < .01$).

Figure 2 (c) shows the average fixation ratios during 30 s until constructing the shifted explanation. A one-way ANOVA showed a significant main effect of the six facts ($F(5,45) = 7.48, p < .01$). The results of the multiple comparisons using Holm's method showed that the fixation ratio of the key fact was significantly higher than those of other facts ($MSe = 0.02, p < .05$). A t-test also indicated that the fixation ratio of the key fact was higher than the baseline ($t(9) = 4.70, p < .01$), whereas the fixation ratio of the first half of the introduction was lower than the baseline ($t(9) = -5.19, p < .01$).

The results of the experiments are summarized as follows: (1) The attention of the key fact, which is inconsistent with the initial explanation, declined and (2) the successful group increased their attention on the key fact just before the explanatory shift. These results are consistent with the experimental results in Terai et al. (2012).

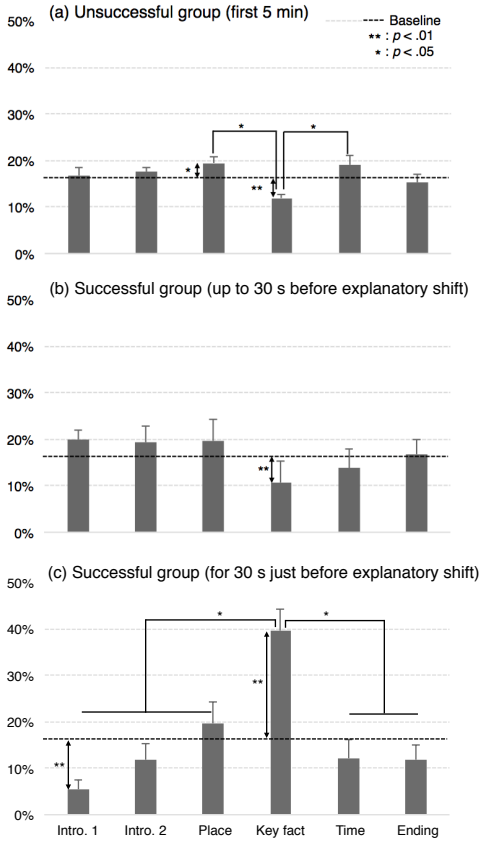


Figure 2: Average fixation ration of each fact.
Note. Each bar chart shows the average fixation ratios in the reconstruction phase. Error bars indicate standard errors.

Experiment 2

In experiment 2, we discuss factors to facilitate an explanatory shift by controlling both the bottom-up and the top-down processes: the opportunity to obtain the key fact and reflective thinking, which tries to consider a phenomenon from a different perspective.

Method

Participants Eighty-eight undergraduates participated in this experiment.

Task We used a barber task that was slightly modified from the one used in the previous experiment. In addition, in the reconstruction phase, we changed the procedure of the presentation of the story to control the rates of seeing each fact. In particular, each primal fact (place, key, and time facts) was repeatedly presented rather than having the whole story displayed at one time.

Conditions In addition to the control condition, we introduced three experimental conditions (bottom-up, top-down, and bidirectional). In the bottom-up condition, we controlled the observation time of each fact. In particular, the observation time of the key fact was twice (20 s) as long as that in the control condition, where the observation time of each fact in each period was 10 s.

In the top-down condition, we instructed the participants to reconsider a new explanation with reflective thinking. In particular, the participants were instructed to listen to instructions for reflective thinking through a set of headphones. Sentences in the instructions are as follows: (1) "Try to consider a new explanation without being bound to the initial explanation." (2) "Try to consider a radically new explanation." (3) "Try to consider a new explanation only on the basis of the presented facts." (4) "Try to consider a new explanation without adding any unobserved fact(s)."

The order of these instructions was randomized. During the vocal playback of each instruction, each fact was temporarily hidden. In the control and bottom-up conditions, these instructions were not given, while each fact was hidden at the same time in the bottom-up condition.

In the bidirectional condition, experimental manipulation was the combination of the bottom-up and top-down conditions. In particular, the participants observed the key fact for twice as long as the others and they were instructed to listen to the directions for reflective thinking through a set of headphones.

We conducted the experiment as a between-participants design (bottom-up: n = 20, top-down: n = 21, bidirectional: n = 23).

Procedure The procedure in the experiment followed that of the previous experiment. In the reconstruction phase, the observation time of each fact (place, key, and time) and the instructions for reflective thinking were controlled.

Results

Table 1 shows the results of success and failure of the explanatory shift. In the analyses, we excluded the participants who did not construct the initial explanation during the initial explanation phase. There was no significant difference among conditions where the successful participants who could shift their explanation in the bidirectional condition were greater than the ones who did in the other condi-

Table 1: Performance

Conditions	*n*	Shift	No shift
Control	17(24)	4	13
Top-down	17(20)	4	13
Bottom-up	15(21)	3	12
Bidirectional	20(23)	6	14

Note. Numbers in () refer to all the participants, including those who did not construct the naïve explanation.

tions.

Although there was no significant difference in the performance among the conditions,the experimental conditions might have influenced the process toward the explanation shift. Thus, we focused on the unsuccessful group (who could not form an explanatory shift). In the following analyses, we classified reconstructed explanations on the basis of their description to realize whether the participants tried to reinterpret the key fact. In particular, we defined the explanations that were based on the fact of the hair as "a pre-shifted explanation" in which participants were trying to reinterpret the key fact: e.g., "Because he had a full schedule and no time to fix himself up" and "Because Alf has hair and Bally is bald." On the other hand, we also defined explanations that avoided to reinterpreting the key fact were based on place and time facts and other facts not described in the story as "non-shifted explanations." In the following analyses, we compare each experimental condition with the control condition as the baseline.

Bottom-up vs. Control Figure 3 (a) shows the differences in the number of reconstructed explanations between the bottom-up condition and the baseline. In each explanation type, t-tests indicated no significant difference ($t(24) = -0.65, n.s.; t(24) = 0.96, n.s.; t(24) = -0.90, n.s.; t(24) = 1.24, n.s.$).

Top-down vs. Control Figure 3 (b) shows differences in the number of reconstructed explanations between the top-down condition and the baseline. In the non-shifted explanations based on the place and time facts, t-tests indicated no significant difference ($t(23) = 1.42, n.s.; t(23) = 1.43, n.s.; t(23) = 0.88, n.s.$). On the other hand, In the non-shifted explanations based on other facts, the number of reconstructed explanations in the top-down condition were lower than those in the baseline ($t(23) = -2.17, p < .05$).

Bidirectional vs. Control Figure 3 (c) shows differences in the number of reconstructed explanations between the bidirectional condition and the baseline. In non-shifted explanations based on the place and time facts, t-tests indicated no significant difference ($t(25) = 2.02, n.s.; t(25) = 0.28, n.s.$). On the other hand, In non-shifted explanations based on other facts, the number of reconstructed explanations in the bidirectional condition were lower than those in the baseline ($t(25) = -3.35, p < .005$). Moreover, in the pre-shifted explanations based on the key fact, the number of reconstructed explanations in the bidirectional condition were higher than those in the baseline ($t(25) = 3.20, p < .005$).

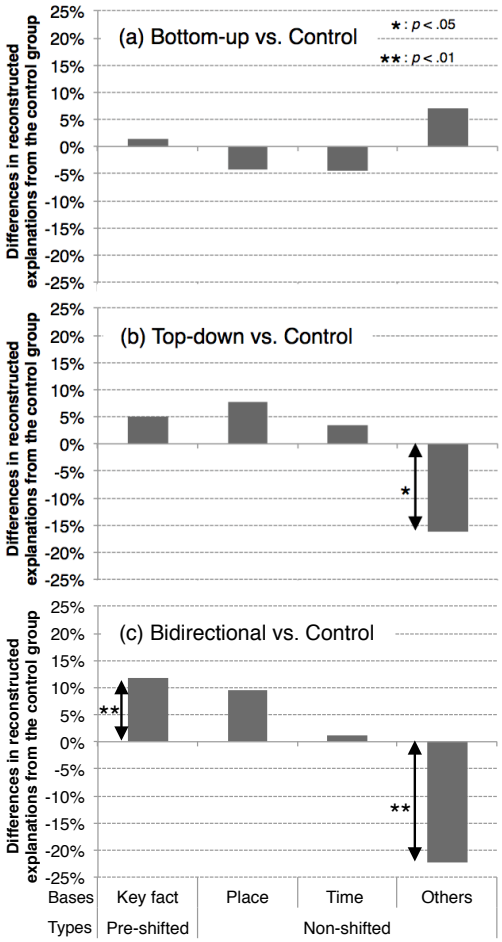


Figure 3: Types of reconstructed explanations

Discussion

Summary

The experimental results in experiment 1 are summarized as follows. First, when the initial explanation was rejected, a new explanation was required after attention on the key fact was inhibited. Second, the successful group increased their attention on the key fact just before the explanatory shift. These results are consistent with the results of Terai et al. (2012).

The experimental results in the experiment 2 are also summarized as follows. First, protection of the initial explanation with unobserved facts, i.e., constructing a non-shifted explanation, was not inhibited by only guiding the participants ' attention toward the key fact. Second, although the initial explanation was not completely shifted, a explanatory pre-shift was achieved by activating reflective thinking with attention on the key fact.

Explanatory Shift and Seeing

The experimental result that increasing the opportunity to obtain the key fact from bottom-up was insufficient to the explanatory shift is consistent with the theory-ladenness of observation. Namely, even if the chance to see the key fact in-

creases, the explanatory shift might not be facilitated because the perspective remain unchanged.

In problem-solving research, the Einstellung (set) effect is known to occur when the first idea that comes to mind, triggered by familiar features of a problem, prevents a better solution from being found(Bilalić, McLeod, & Gobet, 2008; Luchins & Luchins, 1950). Moreover, functional fixedness is well known as one of the cognitive biases that limits people to using a daily object only in the way it is traditionally used(Duncker, 1945; Scheerer, 1963).

Therefore, an underlying initial explanation leads a person to see a key fact with a bias. Then, the initial explanation will not be shifted even if the opportunity of seeing the key fact is increased.

Explanatory Shift and Reflective Thinking

The results of experiment 2 revealed that the focus of participants on the presented facts rather than non-existent facts facilitated by the reflective thinking (the top-down condition). Moreover, reflective thinking coupled with increasing the opportunity of gathering information about the key fact led the participants to construct a pre-shifted explanation (the bidirectional condition).

The dual process theory considers that high-order cognition consists of different types of systems, i.e., System 1 and System 2(Evans, 2003; Kahneman, 2011; Stanovich, 2004). System 1 is fast, parallel, automatic/unconscious, and intuitive, whereas System 2 is slow, serial, temporal, and requires conscious attention. Facilitation of reflective thinking might affect System 2. Stanovich's tripartite model of the mind based on the dual process theory suggests that System 1 corresponds to the autonomous mind and that System 2 is divided into two subsystems, the algorithmic mind and the reflective mind. The reflective mind inhibits an automatic response by the autonomous mind and activates the algorithmic mind to start calculating an alternative solution. The instruction that requires reflective thinking in experiment 2 might correspond to making an effort by the reflective mind to inhibit an automatic response to the key fact and to reinterpret the key fact in the algorithmic mind.

Explanatory Shift and Cues

In an insight problem-solving study, Grant and Spivey (2003) induced the participants' attention to the skin by subtly pulsing the skin area of a diagram in the X-ray problem. The thickness of its outer edge increased and decreased by one pixel three times per second. Such experimental manipulation might not only increase the opportunity of information gathering from the bottom-up but also provide a cue, which is essential for problem solving.

Similarly, in our previous research(Terai et al., 2012), we could facilitate the explanatory shift by both highlighting the key fact to lead the participants' attention toward it and instructing the participants that the highlighted fact was crucial for finding the right explanation. The bidirectional condition in experiment 2 of the present study lacked a cue that gave an information about where to direct reflective thinking. Such cues might be crucial to facilitate an explanatory shift.

References

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect. *Cognition*, *108*(3), 652–661.

Duhem, P. (1914). *La théorie physique son objet et sa structure* (2nd ed.). Paris: Chevalier et Rivière. (English Translation Phillip Wiener, The Aim and Structure of Physical Theory, Princeton: Princeton University Press, 1954)

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*(270), 1-113.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *TRENDS in Cognitive Sciences*, *7*(10), 454–459.

Gardner, M. (1978). *Aha! insight*. New York: W. H. Freeman & Co.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, *14*(5), 462–464.

Hanson, N. R. (1958). *Patterns of discovery*. Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. The university of Chicago Press.

Luchins, A. S., & Luchins, E. H. (1950). New experimental attempts at preventing mechanization in problem solving. *Journal of General Psychology*, *42*, 279–294.

Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. *Learning and Instruction*, *11*(6), 453–483.

Mason, S. F. (1953). *A history of the science – main currents of scientific thought*. London: Lawrence & Wishart Ltd.

Scheerer, M. (1963). Problem solving. *Scientific American*, *208*, 118–128.

Simon, H. A. (2000). Discovering explanations. In R. A. Keil F. C. & Wilson (Ed.), *Explanation and cognition* (pp. 21–59). Cambridge, MA: MIT Press.

Stanovich, K. E. (2004). *The robot 's rebellion: finding meaning in the age of darwin*. Chicago, IL: University of Chicago Press.

Terai, H., Miwa, K., & Matsubayashi, S. (2012). Explanation reconstruction through reinterpretation of key facts. In *Proceedings of 34rd annual conference of the cognitive science society* (pp. 2411–2416).

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Gesture reveals spatial analogies during complex relational reasoning

Kensy Cooperrider (kensy@uchicago.edu)
Department of Psychology, 5848 S. University Avenue
Chicago, IL 60637 USA

Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208 USA

Susan Goldin-Meadow (sgm@uchicago.edu)
Department of Psychology, 5848 S. University Avenue
Chicago, IL 60637 USA

Abstract

How do people think about complex relational phenomena like the behavior of the stock market? Here we hypothesize that people reason about such phenomena in part by creating spatial analogies, and we explore this possibility by examining people's spontaneous gestures. Participants read a written lesson describing positive and negative feedback systems and then explained the key differences between them. Though the lesson was highly abstract and free of concrete imagery, participants produced spatial gestures in abundance during their explanations. These spatial gestures, despite being fundamentally abstract, showed clear regularities and often built off of each other to form larger spatial models of relational structure—that is, spatial analogies. Importantly, the spatial richness and systematicity revealed in participants' gestures was largely divorced from spatial language. These results provide evidence for the spontaneous use of spatial analogy during complex relational reasoning.

Keywords: analogy; relational reasoning; gesture; complex systems; spatial cognition

Introduction

Ecosystems in flux. Seesawing financial markets. Shifting climate patterns. What these diverse phenomena have in common is that they are all examples of complex relational systems: they involve multiple causal factors that change over time and bring about changes to other factors in the system. Such systems underlie phenomena throughout the natural and social world, in all domains and at all scales. Yet, despite the ubiquity and importance of these systems, much remains to be learned about the cognitive processes involved in understanding them.

Current evidence suggests that complex relational reasoning presents challenges even for adults. For example, undergraduates have considerable difficulty detecting higher-order causal patterns such as *positive feedback* and *negative feedback,* the focus of the present paper (Rottman, Gentner, & Goldwater, 2012). Expertise in identifying such patterns does develop, either through exposure to the same patterns across a range of domains (Rottman, Gentner, & Goldwater, 2012) or through a scaffolded process of comparing examples (Goldwater & Gentner, 2015). An interesting open question, however, concerns the nature of the representations that people form as they develop such abilities. What are these representations like and how do they differ from people's representations of other kinds of systems?

Possible clues may come from research on how people understand systems more generally. Much previous research has investigated how people understand mechanical processes with multiple causal components, such as sets of gears and pulleys. A major finding of this line of work is that people often develop mental models of the system that are visuospatial in nature (Hegarty, 2004). One line of evidence for the visuospatial character of these models is that when reasoning about such systems, people often produce diagrams (Forbus, Usher, Lovett, & Wetzel, 2011; Novick, 2001; Tversky, 2011) or gestures (e.g. Schwarz & Black, 1996; Nathan & Martinez, 2015). Based on such observations, it seems plausible that people develop mental models of other types of complex systems, such as the causal patterns under consideration here. However, there is a crucial difference between mechanical systems and positive and negative feedback systems. Positive feedback and negative feedback are consummate abstractions. They are relational patterns that may sometimes be instantiated in mechanical or concretely spatial systems—e.g. a flush valve toilet is an example of a negative feedback system—but their relational essence transcends any one concrete instantiation. It might thus seem unhelpful, or even counterproductive, to recruit visuospatial reasoning processes when thinking about such pure abstractions.

At the same time, a separate line of research has investigated how people recruit space when talking and thinking about purely abstract ideas. This tendency can be seen in everyday language, for instance in the spatial words and grammatical structures people draw on to talk about time (e.g. Clark, 1973; Traugott, 1978), or in the extension of spatial prepositions to describe abstract relations of other kinds (Jamrozik & Gentner, 2015). In fact, evidence has now accumulated that this is not just a linguistic phenomenon—people use spatial representations when reasoning online about abstract concepts, whether or not language is involved (Boroditsky, 2001; Casasanto & Bottini, 2014). One clear source of evidence for the use of

space in abstract reasoning comes from the gestures people produce (Cienki, 1998). To date, the best-studied cases of abstract spatial gesture have involved relationally simple concepts, such as the representation of a temporal sequence as a line (Cooperrider, Núnez, & Sweetser, 2014). Nonetheless, such findings raise the intriguing possibility that people might create more complex spatial structures in gesture to represent more complex relational structures.

The above observations lead us to the following hypothesis about how people reason about complex relational patterns like positive and negative feedback: they may do so, at least in part, by creating abstract spatial models of the relational structures involved—that is, spatial analogies. Furthermore, if this hypothesis is correct, then gesture should provide a powerful window onto this phenomenon. Gesture is well suited to the expression of spatial ideas (Alibali, 2005), and it has been shown to reveal implicit aspects of understanding that people have difficulty verbalizing (Goldin-Meadow, 2003; Broaders, Cook, Mitchell, & Goldin-Meadow, 2007). Moreover, the spatial information revealed in people's abstract gestures often goes beyond what is found in the language co-produced with those gestures (Cienki, 1998).

In the present study, we explore this spatial analogy hypothesis by having people read a lesson contrasting two types of complex relational patterns—positive and negative feedback—and then explain the key differences between them. The most interesting possibility is that gesture might reveal spatial analogies—that is, systematic spatial models of relational patterns that are not inherently spatial. We also considered other possible outcomes, however. For one, people might not spatialize much of anything in their explanations. After all, gesture is thought to stem from vivid visuospatial or motoric imagery (e.g. Hostetter & Alibali, 2008), which our lesson lacks. Another possibility is that people might spatialize in gesture, but in a piecemeal fashion. That is, they may occasionally produce abstract spatial gestures (e.g. an upward gesture when describing "increasing") but these gestures will not cohere into a larger model.

Methods

Participants

23 adults from the University of Chicago community participated for course credit or cash. Four participants were excluded from the analyses: three because their gestures were largely occluded on the video; one for producing no gestures at all. In all, data from 19 participants (10 female; mean age = 20.8 years) are reported in the analyses.

Materials and procedure

After giving consent to participate and to be videotaped, participants carried out a series of activities that served both to familiarize them with causal systems and to assess their understanding of them. First, participants completed an adaptation of the Ambiguous Sorting Task (AST) used previously by Rottman, Gentner, & Goldwater (2012) and Goldwater & Gentner (2015). In this type of sorting task, participants are given a set of vignettes printed on index cards and are asked to sort them into categories. Each vignette is an example of one of several types of causal systems (e.g. positive feedback) instantiated in one of several domains (e.g. economics). Participants are also given *seed cards*—vignettes just like those that need to be sorted but which serve as anchors for the categories to be used. A key feature of the task is that the seed cards leave the relevant categories open to interpretation: a participant may categorize the vignettes according to the type of causal system described or, more superficially, by the domain in which that system is couched. In the adaptation of the AST used here, participants were presented with three seed cards: a first unlabeled card describing the phenomenon of stock market bubbles (a positive feedback system), a second unlabeled card describing predator-prey relationships (a negative feedback system), and a third card simply labeled 'other.' Participants were then given 11 new vignettes and were given 5 minutes to sort them.

After the sorting was complete, the experimenter removed the materials and prompted the participant to explain the main difference between the different categories involved in the sorting task. This phase is the *pre-lesson explanation*. Together the sorting task and the pre-lesson explanation serve to familiarize participants with causal systems, which they will go on to learn more about and explain.

Next, participants were given a one-page written lesson ('Causal Systems Lesson') explaining the differences between positive and negative feedback systems (though without using those labels). The lesson was grounded in the seed cards used in the sorting task. It explained how the stock market vignette exemplifies one type of causal system and how the predator-prey vignette exemplifies a different type. The lesson also moved beyond the particular examples, characterizing in more abstract terms how each type of system involves different relationships between causal factors. Importantly, the lesson used no concrete spatial imagery and very little spatial language. Participants were instructed to study it for 3 minutes and were told that they would later be asked to explain it to another participant.

When the 3 minutes were up, the experimenter removed the lesson and brought in the other participant (who was actually a confederate). The experimenter then prompted the participant as follows: "Please explain the lesson you just read. Go into as much detail as possible, but focus on the differences between the two types of causal systems." The instructions made no mention of gesture. This phase is the *post-lesson explanation*, and it is the focus of our analyses.

Analysis

Participants' performance on the sorting task was analyzed but is not discussed in the present report. Videos of participants' pre- and post-lesson explanations were transcribed and analyzed using ELAN video annotation software (https://tla.mpi.nl/tools/tla-tools/elan/). The gesture

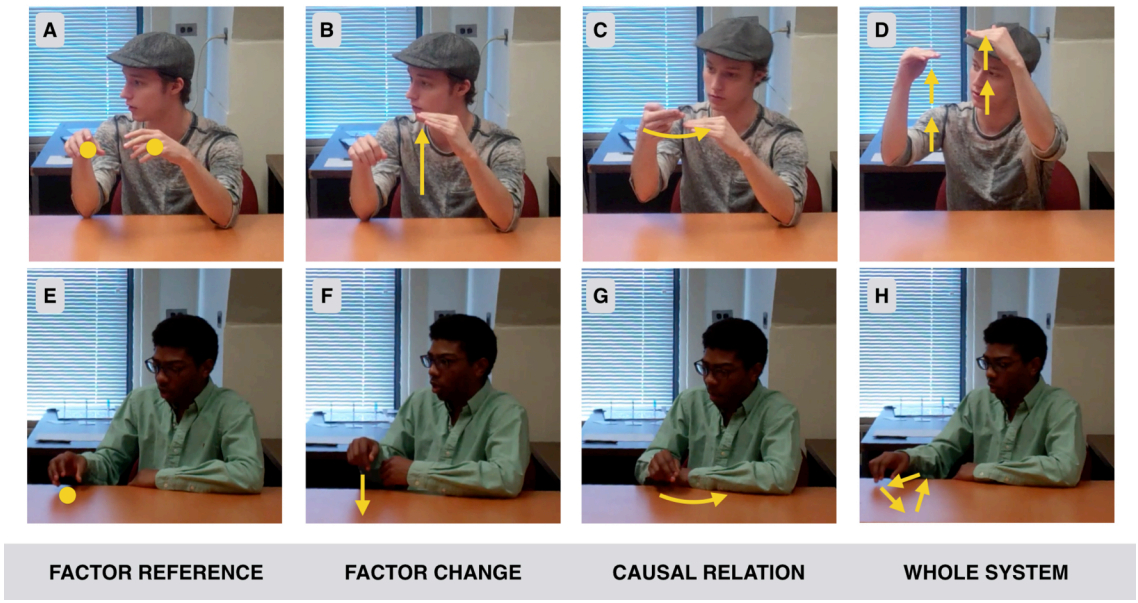


Figure 1: Examples of the different gesture types, taken from two participants' explanations. *Factor reference* gestures (A, E) represent the factors as locations in space, depicted by the yellow circles. *Factor change* gestures (B,F) represent increases and decreases as movements, depicted by the straight yellow arrows. *Causal relation* gestures (C,G) represent causation as movement, depicted by the curved arrows. *Whole system* gestures (D, H) represent the behavior of the system as a whole and often involve multiple movement phases, as depicted by the multiple arrows.

analyses reported here focus on participants' post-lesson explanations.

A first step in the analysis was to identify all gestures in the explanations that were "representational" (e.g. Chu et al., 2013). Representational gestures depict some property of a referent (commonly called "iconic" or "metaphoric" gestures), or point to a referent's location (commonly called "deictic" gestures). In the present data, the representational gestures were abstract in nature—that is, they used location, movement, and spatial arrangement to depict ideas that themselves had no concrete location, did not actually move, and had no visible spatial properties. Once representational gestures were identified, they were then categorized into gesture types (see below). Reliability was assessed by having a second coder categorize the representational gestures in 5 randomly selected explanations (26% of the data). The coders agreed 83% (N=220) of the time in whether a gesture fit into the categorization system (i.e. belonged to one of the four categories). For those gestures that both coders agreed fit into the system, they assigned the gesture to the same category in 85% (N=142) of cases. Finally, we analyzed the gestures' spatial properties and relationships to other gestures in the same explanation, as well as the language that was co-produced with them. Each of these analyses is described in more detail as the results are presented.

Results

Gesture rates and types

Participants produced a mean of 24.12 (SD=4.39) representational gestures per minute speaking. The abstract, textual nature of the Causal Systems Lesson thus did not stand in the way of eliciting representational gestures.

Based on pilot studies involving similar materials, a system was developed for categorizing the recurring ways people gesture to represent elements of feedback systems. First, people locate the factors (e.g. the predator and prey populations in the negative feedback example) by placing their gestures in space or by pointing to locations. These we call *factor reference gestures* (see Fig. 1). Second, people represent changes to the factors (e.g. an increase in the predator population) as movements. These we call *factor change gestures*. Third, people represent causal relations in the system (e.g. how the change in the predator population causes a change in the prey population) as movements, sometimes between previously established locations. These we call *causal relation gestures*. Fourth, people use movements to characterize the behavior of the system as a whole (e.g. the equilibrium that is reached in the predator-prey system). These we call *whole system gestures*. The majority (71%) of participants' representational gestures fell into one of the above four types. However, not all participants produced all four gesture types: 19 (100%) produced factor reference gestures, 18 (95%) produced

factor change gestures, 13 (68%) produced causal relation gestures, and 9 (47%) produced whole system gestures.

Spatial properties

Spatial axes We next analyzed the spatial characteristics of people's gestures, considering each gesture type separately. 96% of factor reference gestures located the factors on the left-right axis, most often with the first-mentioned factor on the left and the second-mentioned factor on the right. Factor change gestures were more variable, with 21% depicting increases and decreases as movements along the left-right axis, 29% along the front-back axis, 26% along the up-down axis, and the rest involving either some combination of these axes or a more complex movement. Causal relation gestures most commonly (75%) depicted causation as movement along the left-right axis. Whole system gestures varied considerably across participants in their use of space and in other qualitative characteristics, but they tended to use multiple movement phases (see Fig. 1, panels D and H) and often involved two hands.

Spatial consistency We next examined how consistent participants' gestures were in their spatial properties over the course of the explanation. To assess this kind of within-participant consistency, we used a measure developed in the study of spatial grammatical devices in signed languages (Senghas & Coppola, 2001). For every gesture, we asked whether it represented a system element (e.g. a particular factor) for the first time or represented a system element that had been previously represented. If the gesture repeated an element, we coded whether it used space in the same way (consistent) or a different way (inconsistent) as the immediately preceding gesture. For this analysis, we focused on factor reference and factor change gestures because causal relation and whole-system gestures, when they occurred, often only occurred once or twice in an explanation. Overall, participants were highly spatially consistent: the majority of factor reference gestures were spatially consistent (mean percentage=87%), as were the majority of factor change gestures (mean percentage=69%).

Model integration Finally, we analyzed whether participants' gestures were integrated with a larger spatial model built up over the explanation, or were more piecemeal in nature. Use of model-integrated gestures varied across participants. If, for instance, a participant produces a gesture representing an increase to a factor that incorporates the previously established location of the factor involved, the gesture would be considered *model-integrated*. As an example of a model-integrated gesture, a participant may locate the first factor on the left and then later show an increase in that factor as an upward movement in left space (see Fig. 1, panel B). If, on the other hand, the increase was depicted in neutral space or right space, the gesture would not be considered model-integrated. Similarly, if a participant produces a gesture representing a causal relation between two factors as a movement between the previously established locations of the two factors, the gesture would be considered model-integrated (see Fig. 1, panel C). Alternatively, if the gesture represented causation as a movement in neutral space, it would not be considered model-integrated. Note that factor reference gestures, which serve to establish such locations in the first place, cannot be model-integrated and are excluded from the analysis. Further, we did not expect whole system gestures, which depict a high-level summary of the whole system, to be closely integrated with the detailed causal patterns depicted in the other gestures. Overall, 50% of participants' factor change gestures were model-integrated, as were 72% of their causal relation gestures. 84% of participants produced at least one model-integrated factor change gesture, and 47% produced at least one model-integrated causal relation gesture.

Relationship to language

Finally, we analyzed the relationship between participants' gestures and the language with which they were co-produced. Most often, in 93% of cases, the gestures represented aspects of the system that were simultaneously mentioned in speech. For example, a participant would produce a factor reference gesture while referring in speech to "the first factor" or a factor change gesture while mentioning an "increase." Interestingly, however, gestures sometimes filled in where speech left off—especially when characterizing the behavior of the system as a whole. For example, a speaker describing a positive feedback system said "it's sort of…" trailing off in speech but providing a complex spatial characterization in gesture. These cases may stem from the difficulty of verbalizing the overall system dynamics.

Finally, we investigated how common it was for gestures to be co-produced with overtly *spatial* language. Table 1 provides examples of both spatial and non-spatial language that was co-produced with the different gesture types. Note

Table 1: Examples of spatial and non-spatial language co-produced with gestures

	non-spatial language	spatial language
factor reference	"first factor" "certain variable"	"external variable"
factor change	"increase in" "change"	"rise" "go up"
causal relation	"influences" "causes"	"rebounds" "turns around"
whole system	"self-correcting" "regulate each other"	"negative loop" "building on each other"

that words such as "increase" which have a general definition that is not specifically spatial, were not considered spatial. Strikingly, overall, only 17% of the gestures were co-produced with overtly spatial language. For example, factor reference gestures, though consistently exploiting the left-right axis of space, were not once co-produced with a reference to left or right.

Discussion

We investigated the possibility that people would spontaneously use spatial analogies when reasoning about positive and negative feedback, relational patterns that are complex, widespread, and fundamentally abstract. As a potential window into such hypothesized analogies, we examined the gestures people produced as they tried to articulate the main characteristics of these patterns and the differences between them. Despite the paucity of concrete spatial imagery or language in the lesson we provided, when people explained the patterns, they gestured at strikingly high rates. These gestures did not represent the actual locations or movements of objects—rather, the gestures used space abstractly to represent the different factors in the system, the changes to those factors, the causal relations between them, and the overall dynamics of the systems being described. Over the course of people's explanations, the gestures were highly consistent in where they were placed in space, and they were often integrated into larger spatial models that were built up over time. Finally, the spatial richness we observed in gesture was largely divorced from spatial language, and sometimes divorced from language altogether. In sum, the gestures we observed provided vivid evidence that people draw on spatial analogies during complex relational reasoning, evidence that would have been scarce in a verbal transcript.

One limitation of the present study is that, although the lesson was largely devoid of rich imagistic content, it did include a sprinkling of abstractly spatial words. For example, the phrase "opposite *direction*" was used to describe the change from increasing to decreasing. It remains possible that subjects took these words as cues to build larger spatial models. However, the scarcity of spatial language overall in participants' explanations makes this possibility somewhat doubtful. Nonetheless, further study will be needed to determine whether excluding such words would have a significant impact on the extent to which people create spatial models in gesture.

As we have argued, the gestures we observed revealed the spontaneous use of sophisticated spatial analogies—that is, spatial models of relational structure. Spatial analogy is likely a ubiquitous process in human reasoning. Perhaps the best-studied examples to date have involved reasoning about maps and scale models (e.g. Uttal & Wellman, 1989). In such cases, a set of concrete spatial relations in the world is mapped in schematic fashion to some spatial representation of that world. The analogical mapping is thus between one spatial format and another spatial format. By contrast, in the spatial analogies under examination here, the base concept is a purely abstract set of entities and relations—factors, changes, and causation—that is mapped to a set of spatial relations—locations, movements, and movements between locations. Prior work has demonstrated that people are able to understand and reason with spatial analogies of this abstract type (Gattis, 2004), but little work to date has examined whether the spatial analogies are spontaneously created or recruited on the fly. Informal observations, in addition to our own data, suggest that spatial analogies may constitute a powerful strategy in both cognition and communication. The ubiquity of abstract spatial models like Venn diagrams, family trees, and cladograms, for example, hints at the wider utility of spatial analogy in relational reasoning, far beyond our chosen test case of positive and negative feedback patterns (Novick, 1996; Tversky, 2011). Interestingly, the phenomenon of model-integrated gestures we have described also resembles a phenomenon in established signed languages sometimes described as "spatial modulation" (Senghas & Coppola, 2001). In American Sign Language, for example, a verb may be said to be "spatially modulated" if it incorporates spatial information that was previously established for one or more of its arguments. As our data show, hearing gesturers do something very similar under the right circumstances (see also So, Coppola, Licciardello & Goldin-Meadow, 2005).

Analogy is often thought of as an effortful process in which someone, struggling to capture a new idea, alights on an apt comparison. However, empirical work has shown that this formulation is, at least in some cases, misleading: analogical mapping can occur unintentionally, without any effort (Day & Gentner, 2007). We suggest a similar unintentional deployment of analogy may be at work here. Participants very rarely referred to their gestures—or to the spatial information contained therein—explicitly (e.g. "Imagine the system is like this"). Nor did they show signs of engaging in an effortful process of design and development, as might be signaled by restarts or amendments to the spatial structure. Rather, we suggest that participants constructed these spatial models fluidly and more or less unconsciously as they articulated the relational structure they were describing.

A related issue is whether the abstract gestures we observed were helpful to the speaker over and above any role they may have served in communication. Prior work has shown that gesturing can help speakers by reducing cognitive load (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001). To our knowledge, though, cognitive benefits of this sort have not been shown for abstract spatial gestures of the type described here. In fact, if spatial analogy is an effortful process, as is sometimes assumed, then producing gestures like those documented here would actually increase cognitive load rather than lightening it. Testing these possibilities is a direction for future work.

Complex relational patterns underlie diverse phenomena across the natural and social worlds. While earlier work has demonstrated the difficulties of reasoning about such patterns, less is known about the kinds of representations

people bring to bear during such reasoning. Here we provide evidence for the role of spontaneous spatial analogy in this kind of reasoning, and for gesture as one means of externalizing such analogies. Though we have barely scratched the surface of this arena, it remains plausible that spatial analogies will prove to be a central ingredient in the human ability to understand complex relational phenomena.

Acknowledgments

Funding for this study was provided by the NSF Spatial Intelligence and Learning Center (SBE 0541957, Gentner and Goldin-Meadow are co-PIs) and NICHD (R01-HD47450 to Goldin-Meadow). We thank Ishara Ruffins and Erin Richard for research assistance.

References

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information Information. *Spatial Cognition & Computation*, *5*(4), 307–331.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539–50.

Casasanto, D., & Bottini, R. (2014). Spatial language and abstract concepts. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(2), 139–149.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2013). Individual Differences in Frequency and Saliency of Speech-Accompanying Gestures: The Role of Cognitive Abilities and Empathy. *Journal of Experimental Psychology. General*.

Cienki, A. (1998). Metaphoric gestures and some of their relations to verbal metaphorical expressions. In J.-P. Koenig (Ed.), *Discourse and cognition: Bridging the gap*. Stanford, California: CSLI.

Clark, H. H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

Cooperrider, K., Núñez, R., & Sweetser, E. (2014). The conceptualization of time in gesture. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & J. Bressem (Eds.), *Body-Language-Communication (vol. 2)*. New York: Mouton de Gruyter.

Day, S. B., & Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory & Cognition*, *35*(1), 39–49.

Forbus, K., Usher, J., Lovett, A., & Wetzel, J. (2011). CogSketch: Sketch understanding for Cognitive Science Research and for Education. *Topics in Cognitive Science,* 1-19.

Gattis, M. (2004). Mapping relational structure in spatial reasoning. *Cognitive Science*, *28*, 589–610.

Goldin-Meadow, S., Nusbaum, H. C., Kelly, S. D., & Wagner, S. M. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*, 516–522.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard U. Press.

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137–153.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–5.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Jamrozik, A., & Gentner, D. (2015). Well-hidden regularities: Abstract Uses of in and on retain an aspect of their spatial meaning. *Cognitive Science*.

Novick, L. R. (2001). Spatial diagrams: Key instruments in the toolbox for thought. In D. L. Medin (Ed.), *The psychology of learning and motivation*, *Vol*. 40. San Diego, CA: Academic Press.

Nathan, M. J., & Martinez, C. V. J. (2015). Gesture as model enactment: the role of gesture in mental model construction and inference making when learning from text. *Learning: Research and Practice*, *1*(1), 4–37.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*(5), 919–32.

Schwartz, D. L., & Black, J. B. (1996). Shuttling Between Depictive Models and Abstract Rules: Induction and Fallback. *Cognitive Science*, *20*(4), 457–497.

Senghas, A., & Coppola, M. (2001). Children Creating Language: How Nicaraguan Sign Language Acquired a Spatial Grammar. *Psychological Science*, *12*(4), 323–328.

So, C., Coppola, M., Licciardello, V., & Goldin-Meadow, S. (2005). The seeds of spatial grammar in the manual modality. *Cognitive Science, 29*, 1029-1043.

Traugott, E. C. (1978). On the Expression of Spatio-Temporal Relations in Language. In *Universals of Human Language*. *Volume 3: Word Structure*. Stanford, CA: Stanford.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499-535.

Uttal, D. H., & Wellman, H. M. (1989). Young children's representation of spatial information acquired from maps. *Developmental Psychology*, *25*(1), 128–138.

What Determines Human Certainty?

Louis Martí (lmarti13@gmail.com)
Francis Mollica (mollicaf@gmail.com)
Steven Piantadosi (spiantadosi@gmail.com)
Celeste Kidd (celestekidd@gmail.com)
Department of Brain and Cognitive Sciences, University of Rochester,
Rochester, NY 14627 USA

Abstract

Previous work on concept learning has focused on how concepts are acquired without addressing metacognitive aspects of this process. An important part of concept learning from a learner's perspective is subjectively knowing when a new concept has been effectively learned. Here, we investigate learners' certainty in a classic Boolean concept-learning task. We collected certainty judgements during the concept-learning task from 552 participants on Amazon Mechanical Turk. We compare different models of certainty in order to determine exactly what learners' subjective certainty judgments encode. Our results suggest that learners' certainty is best explained by local accuracy rather than plausible alternatives such as total entropy or the maximum a posteriori hypothesis of an idealized Bayesian learner. This result suggests that certainty predominately reflects learners' performance and feedback, rather than any metacognition about the inferential task they are solving.

Keywords: Concepts; metacognition; learning; human experimentation; symbolic computational modeling; certainty; ideal learning model

Introduction

Most of us are certain that we landed on the moon, but many of us are far less certain about who will win this year's presidential election. Ideally, our certainty would be a direct reflection of the evidence we observe, but is our sense of certainty actually calibrated to reality? Several studies have demonstrated that individuals presented with disconfirming evidence can become even more entrenched in their original beliefs. Tormala and Petty (2004, 2011) found that when individuals were confronted with messages that they perceived to be strong (e.g., from a expert source) but went against their existing beliefs, their certainty regarding those beliefs *increased* instead of decreased. In contrast, within the visual domain, there is evidence that individuals not only calculate their own subjective measure of visual uncertainty, but that their subjective uncertainty is predictive of objective uncertainty (Barthelme & Mamassian, 2009). In other words, individuals' certainty of visual stimuli reflects veridical probabilities.

The Dunning-Kruger effect provides further evidence of a miscalibration between reality and certainty (Dunning & Kruger, 1999). Dunning and Kruger demonstrated a metacognitive inability of unskilled individuals to recognize their own incompetence. This results in an inflated sense of certainty regarding their own performance and aptitude. The inverse was also found, in which highly competent individuals would be less certain regarding their own abilities in relation to others.

Here, we test whether individuals' subjective certainty while acquiring novel concepts is driven by objective probabilities. Are learners as certain as they should be given the data, or is their subjective sense of certainty driven by other factors (e.g., accuracy, quantity of observed data)? This question has implications for understanding the subjective experiences that accompany concept discovery, which may themselves interact with future learning.

Boolean concept learning as a prototypical domain

Historically, Boolean concept-learning tasks have been used to study concept acquisition because they allowed researchers to study the mechanisms of the learning process in a simplified domain with a known, limited hypothesis space (e.g., Bruner, Goodnow, & Austin, 1967; Feldman, 2000; Goodman et al., 2008; Shepard, Hovland, & Jenkins, 1961).

As an example, consider a classic Boolean concept-learning paradigm. In this task, participants are asked to respond *yes* or *no* to a series of images and are given feedback after each response. Each image has a shape, size, and color that has one of two values, resulting in a total of eight different images. The accuracies of the participants' *yes* or *no* responses are determined by an unstated concept such as "large black square" or "triangle" that subjects must discover. This latent concept is a Boolean rule which can easily be stated in logic, meaning that it is straightforward to quantify representations that learners are likely to be using. Feldman (2000) tested participants on 41 different concepts spread across six families corresponding to the number of positive examples and feature dimensions used in each stimulus. His results showed that performance in learning decreased as Boolean complexity increased, indicating that concept difficulty was directly proportional to the number of Boolean operators in the shortest logically equivalent expression. For example, the concept "large red triangle" should be more difficult than "red triangle" since the former incorporates an additional feature. The idea of simplicity-driven concept learning was put into a probabilistic setting by Goodman et al. (2008), who constructed a Bayesian learner that tried to acquire concepts h from data d according to an idealized model of $P(h \mid d)$. The prior in this model favored simplicity, and the likelihood favored hypotheses that explained the observed labels. In this way, it was able to combine a formalization of a simplicity preference with Bayesian inference from the observed data, providing a close fit to empirical learning.

These tasks provide an ideal domain for us to study cer-

tainty since there exist well-tested models and theories of the processes that guide learning in such simple domains. We collect subjective measurements of learners' certainty or confidence throughout a novel concept-learning task. We then compare learners' self-reported certainty throughout the task to a set of models we constructed to represent several a priori plausible, formal theories about what quantitative measure subjective certainty might reflect. We constructed a rule learning model similar to Goodman et al. (2008) in order to provide an idealized quantification of formal measures of certainty, which we then compare to human judgements.

Methods

We tested participants in a novel concept-learning task during which we measured their knowledge of the concept (via *yes* or *no* responses) and their certainty throughout the learning process.

We recruited 552 participants on Amazon Mechanical Turk. Participants clicked to consent to the study before viewing the task instructions. The instructions explained that the participant's task was to figure out the meaning of a word that represented a certain concept. Participants practiced on eight practice trials to ensure that they understood the task before proceeding to the actual study. During the practice trials, participants saw either a cat or a dog, and had to guess whether each item fit the undisclosed concept for a novel word or not by responding *yes* or *no*. In addition to guessing, participants had to report whether or not they were certain about the concept for the novel word. After each guess, participants received feedback about whether or not their guess was correct. For the practice trials, the novel word always referred to the concept of "cat".

For the experimental trials (see Figure 1), participants saw one of ten conditions, each composed of of 24 trials. Each condition represented one unique concept, such that each participant made judgements for only one concept. In a partial replication of the Shepard et al. study (1961), the observations spanned three binary dimensions: shape (square or triangle), color (red or green), and size (large or small). A total of eight images were used across all conditions which exhaustively spanned the space. Across all conditions participants would see these eight images in blocks of three with the ordering of the images assigned randomly per condition. Each condition tested for a different concept with varying complexity (see Table 1).

Concepts 1, 5, 6, 7, 8 and 9 (Table 1) were identical to concepts used in both the Shepard et al. (1961) and Feldman (2000) experiments. These concepts exhaustively spanned across the concept family consisting of three features and four positive examples. Additional conditions were added to test for potential differences between operators. Concept 9—"*(green and large and triangle) or (green and small and square) or (red and large and square) or (red and small and triangle)*"—is predicted to be the most difficult as it essentially transforms that condition into a rote memorization task.

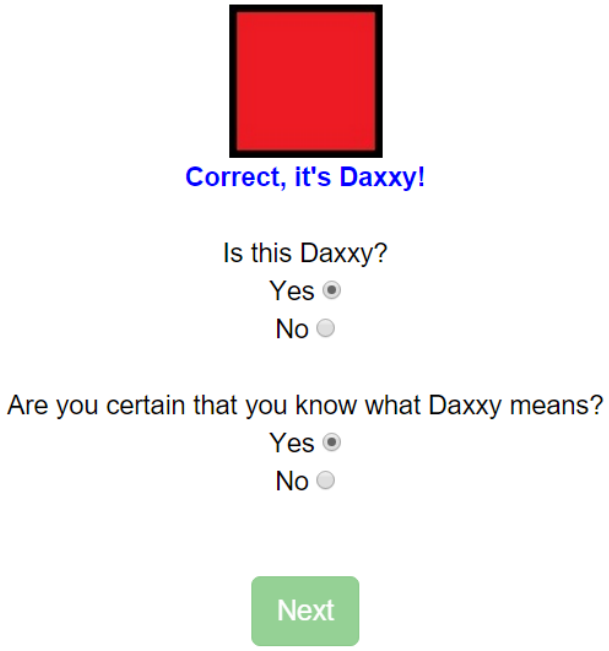


Figure 1: Participants saw 24 trials (as above) in sequential order, randomized between-conditions. After responding, feedback was displayed for one second (if correct) or two seconds (if incorrect) before the next stimuli was shown. Previous feedback was not displayed at any time.

Like the practice trials, each participant gave a "yes" or "no" answer as to whether the current image was part of the concept. On each trial, each participant also indicated whether they were certain regarding their answer, providing a binary forced choice judgment. After their responses, participants received feedback on their responses. Correct responses received one second of feedback before the next trial commenced. Incorrect responses were penalized with a slower two seconds of feedback before the next trial to incentivize attention to the task.

Ideal learning model

We aim to address the question of whether learners' subjective sense of certainty reflects veridical probabilities. In other words, do learners feel as certain as is justified by the observed data? Addressing this question requires us to use an ideal learning model in order to determine how confident an ideal learner should be (given the uncertainty of the model). Here we use an ideal learning model that has already been used to formalize concept learning in a probabilistic setting in which notions of certainty and uncertainty (e.g., Shannon, 1948) are well defined. Though ideal learning models of this type have been used before to understand novel concept acquisition, they have not previously been applied towards understanding learners' subjective sense of certainty.

The ideal learning model was developed using Python

		Concept
1	RED	red
2	AND	red and small
3	OR	red or small
4	XOR	red xor small
5	AND OR AND	(red and small) or (green and large)
6	Complex 1	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and large and square)
7	Complex 2	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and large and triangle)
8	Complex 3	(green and large and triangle) or (green and large and square) or (green and small and triangle) or (red and small and square)
9	Memorization	(green and large and triangle) or (green and small and square) or (red and large and square) or (red and small and triangle)
10	XOR XOR	red xor small xor square

Table 1: Concepts presented to subjects in the experiment.

Rule
START → PREDICATE
START → TRUE
START → FALSE
PREDICATE → and(PREDICATE, PREDICATE)
PREDICATE → or(PREDICATE, PREDICATE)
PREDICATE → not(PREDICATE)
PREDICATE → red(x)
PREDICATE → green(x)
PREDICATE → triangle(x)
PREDICATE → square(x)
PREDICATE → large(x)
PREDICATE → small(x)

Table 2: Grammar used to generate logical rules in the idealized learning model. The variable x is the current object.

and the Language Of Thought library, *LOTlib* (Piantadosi, 2014). This model defines a probabilistic context-free grammar (PCFG) with a set of primitives: red, green, triangle, square, large, and small, and logical operations (shown in Table 2). The PCFG serves as a prior over hypotheses and specifies an infinite hypothesis space.

To establish a tractable hypothesis space, the model drew 1,000,000 samples from the posterior distribution of hypotheses (i.e., hypotheses scored by simplicity and fit to the data) using tree-regeneration Metropolis-Hastings (Goodman et al., 2008) and stored the best 1,000 hypotheses at each data amount that subjects saw. The model incorporated parameters for the noise in the data (alpha) and a power law memory decay on the likelihood of previous data[1] (beta), best fit as 0.64 and 0 respectively.

Analysis

We considered and compared several different models of what might drive uncertainty. Perhaps the most natural is that learners might use the uncertainty of an idealized learning model, as quantified by the posterior **entropy** (Shannon 1948). This provides a measure of the number of bits of information learners have yet to discover about which hypothesis is the true generator of the data. However, it also may be the case that learners tend to pick probable hypotheses, and their uncertainty reflects only the probability of the best hypothesis. We refer to this as the **MAP** model. We may also consider **maximum likelihood** in which the prior is ignored, corresponding to a maximum likelihood model over the structured, compositional hypothesis space. Beyond these idealized model-based theories, it is critical to include a variety of trial-level alternatives. It could be for instance that participants just become more confident as they complete more of the experiment, a model we refer to as **Trial**. Alternatively, certainty may just reflect a measure of their performance so far, reflecting a lack of objective self-awareness of how much certainty they should have. **Total Accuracy** quantifies performance on all previous trials of the experiment. We also include **Local Accuracy** measures of how well subjects have done on the previous N trials ($N = 2,3,4,5$), potentially incorporating their performance on the current trial (e.g. the one they are responding to), called **Local Accuracy Current**. If this predictor beat out the others, it would indicate that learners are only certain when they anticipate being able to guess accurately on the current trial. The **Current Accuracy** model is a baseline that simply quantifies whether participants were right on the next trial. Its performance as a predictor shows whether subjective certainty is well-calibrated to true accuracy on the next item, regardless of the underlying computational processes.

Logarithmic transformations are common in psychophysics (Stevens, 1957). Therefore, each of these predictors was considered in its standard form, as well as under a logarithmic transformation, yielding a total of 32 models. The accuracy predictors used a $log(1+x)$ transformation to avoid problems with zeroes.

Results

Certainty and accuracy by concept

We composed and evaluated plots of participants' certainty and accuracy over the course of the experiment for each concept in order to determine *(1)* whether certainty and accuracy improved over the course of the experiment, *(2)* whether the-

[1] Weighting the log likelihood of an example n back by $(n+1)^{-\beta}$.

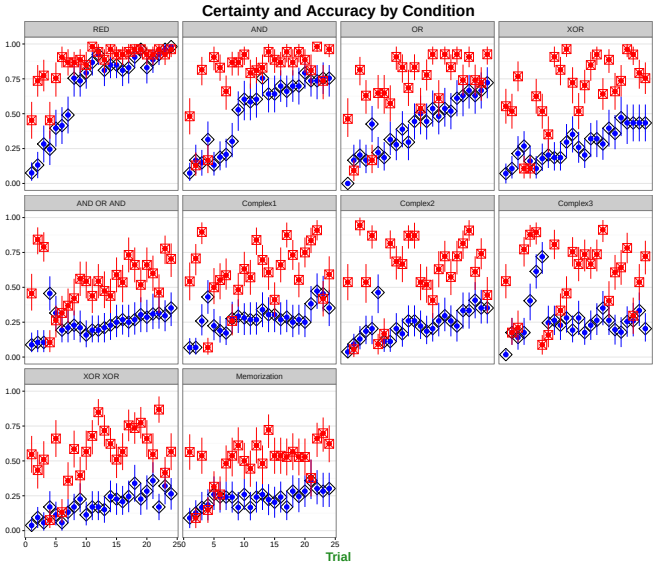


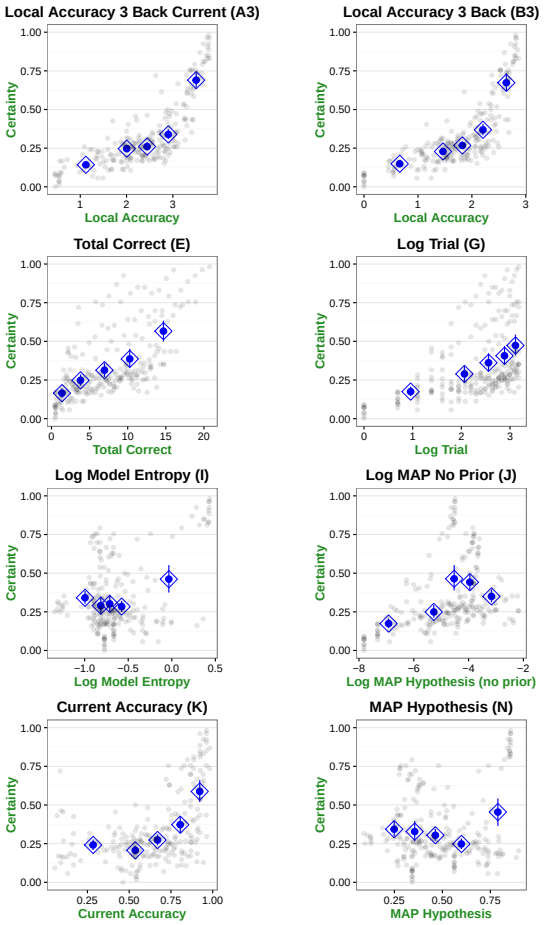
Figure 2: Mean certainty (blue) and mean accuracy (red) across concept conditions

oretically harder concepts (according to Feldman 2000) were, in fact, more difficult for participants, and *(3)* whether participants' certainty correlated with their accuracy in general.

Figure 2 shows participants' certainty and accuracy (*y*-axis) over trials of the experiment (*x*-axis). The increasing trend of the accuracy curves reaches ceiling for some concepts, indicating that participants successfully acquired them. In other conditions, participants did not reach ceiling, indicating that they did not acquire the target concept. This is actually beneficial to our analysis as it allows us to analyze conditions and trials in which participants should have high uncertainty. The certainty curves follow a generally increasing trajectory, but only reach high values (ceiling probability of a participant reporting being certain) in conditions in which participants also achieved high accuracy. The increasing trend of certainty in conditions for which accuracy does not go above 50% may be reflective of overconfidence.

Predictors of certainty

Figure 3 shows certainty (*y*-axis) over several different key predictors of certainty (*x*-axis). Local accuracy models have low dispersion, meaning that individuals with low local accuracy have low certainty and individuals with high local accuracy are highly certain. There are no cases where an individual is highly accurate and highly uncertain and no cases where an individual has low accuracy and is highly certain. On the other hand, total correct and log trial are highly linear but have high dispersion. This is likely due to condition effects. In conditions where the concept is extremely simple (e.g. "red") participants might reach high certainty extremely quickly and, due to a low level of negative feedback (incorrect responses), remain highly confident. Current accuracy



(a) (b)

Figure 3: Visualizations of several key model fits, giving the participant response means for each concept and trial (gray) and binned model means in each of five quantiles (blue) for certainty rating (*y*-axis) as a function of model (*x*-axis). Straight lines with low variance correspond to models which accurately capture human performance.

has a similar shape and dispersion but is not a good predictor due to situations in which the participant is very uncertain but happens to get the trial correct by chance. Finally, log model entropy and the MAP plots perform very poorly. Data points are completely scattered and the predictors perform poorly because of this. For example, although there are many likely MAP hypotheses for which participant certainty is high, there are just as many for which certainty is low. This could be due to complicated concepts for which the ideal learner model does well but participants do not.

Model comparison results

Before examining how well the certainty models predict human certainty, it is important to verify that the ideal learning model predicts human accuracy during the task. A logistic regression predicting behavioral accuracy from model accuracy

	Model	AIC	Pseudo R^2	Log Likelihood	Beta	Standard Error
A3	Local Accuracy 3 Back Current	14753.0	0.11	-7374.5	1.11	0.03
A4	Local Accuracy 4 Back Current	14778.8	0.11	-7387.4	0.85	0.02
B3	Local Accuracy 3 Back	14786.6	0.11	-7391.3	1.33	0.04
B4	Local Accuracy 4 Back	14803.0	0.11	-7399.5	0.99	0.03
A5	Local Accuracy 5 Back Current	14816.3	0.11	-7406.1	0.68	0.02
B5	Local Accuracy 5 Back	14838.8	0.11	-7417.4	0.76	0.02
A2	Local Accuracy 2 Back Current	14872.6	0.11	-7434.3	1.44	0.04
B2	Local Accuracy 2 Back	14892.2	0.11	-7444.1	1.91	0.05
C3	Log Local Accuracy 3 Back Current	14948.6	0.10	-7472.3	3.40	0.10
D3	Log Local Accuracy 3 Back	14990.2	0.10	-7493.1	3.24	0.10
C4	Log Local Accuracy 4 Back Current	15010.9	0.10	-7503.5	2.92	0.09
C2	Log Local Accuracy 2 Back Current	15018.9	0.10	-7507.4	3.78	0.11
D2	Log Local Accuracy 2 Back	15037.6	0.10	-7516.8	3.83	0.11
D4	Log Local Accuracy 4 Back	15051.6	0.10	-7523.8	2.73	0.08
C5	Log Local Accuracy 5 Back Current	15078.1	0.09	-7537.0	2.53	0.08
D5	Log Local Accuracy 5 Back	15121.0	0.09	-7558.5	2.32	0.07
A1	Local Accuracy 1 Back Current	15215.5	0.09	-7605.8	1.88	0.05
C1	Log Local Accuracy 1 Back Current	15338.3	0.08	-7667.1	3.94	0.12
B1	Local Accuracy 1 Back	15346.5	0.08	-7671.3	2.92	0.09
E	Total Correct	15389.5	0.08	-7692.7	0.14	0.00
D1	Log Local Accuracy 1 Back	15430.1	0.07	-7713.1	4.32	0.14
F	Log Total Correct	15525.0	0.07	-7760.5	0.77	0.03
G	Log Trial	15911.6	0.04	-7953.8	0.70	0.03
H	Trial	16014.1	0.04	-8005.1	0.07	0.00
I	Log Entropy	16205.5	0.03	-8100.8	-1.05	0.05
J	Log Maximum likelihood	16207.4	0.03	-8101.7	0.30	0.02
K	Current Accuracy	16339.6	0.02	-8167.8	0.69	0.04
L	Log Current Accuracy	16339.6	0.02	-8167.8	1.00	0.06
M	Entropy	16389.6	0.02	-8192.8	-0.49	0.03
N	MAP	16545.0	0.01	-8270.5	0.93	0.10
O	Log MAP	16593.9	0.00	-8295.0	0.29	0.04
P	Maximum likelihood	16609.7	0.00	-8302.8	5.02	0.90

Table 3: Performance of predictors in determining subjective certainty. All were significant at $p < .001$.

results in a significant relationship, B = 3.206, $p < .001$.

Table 3 shows the full model results, giving the performance of each model in predicting certainty ratings. These have been sorted by the primary measure of performance, AIC, which quantifies the fit of each model penalizing its number of free parameters (closer to $-\infty$ is better). In this case, the AIC is simply the AIC score of a logistic regression including the variable of interest. This table also provides a pseudo R^2 measure, giving a rough measure of the "amount of variance" accounted for by each model (this is not literally an R^2 since amount of variance does not have a clear analog in logistic models). The reported numbers also include a β giving the regression coefficient, its standard error, and a two-tailed p-value comparing it to zero.

As this table makes clear, the **Local accuracy** models outperform any of the alternatives, a pattern which is robust to the way in which local accuracy is quantified (e.g. the number back that are counted or whether the current trial is included). The quantitatively best model A3 tracks accuracy over the past three trials and includes the future accuracy on the next trial. One possible explanation for this is that individuals are simply deciding their own certainty based on recent performance and whether or not they think they know the answer to the current item.

Interestingly, model E, the **Total Correct** count of responses, is the second best predictor of certainty outside of the local accuracy models. The high performance of this and local accuracy models implies that people's certainty is largely influenced by their own perception of how well they are doing on the task. It is important to note that all the models which use either local accuracy or total correct outperform all other models.

The relative poor performance of the number of **Trial**s in predicting performance (model G) indicates that participants are not simply becoming more certain over time regardless of performance. It also excludes the possibility that learners are waiting for exhaustive data before becoming certain. If they were, a sudden spike of certainty would be visible at trial 8, when in our experimental design they have seen all possible feature dimensions and outcomes.

Strikingly, the poor performance of of the **Entropy** and **MAP** models rules out that subjective certainty is calibrated with an ideal learner. The poor performance of these models is consistent with the theory that learners are likely not maintaining more than one hypothesis in mind—perhaps they store a sample from the posterior, but do not have access to the full posterior distribution. Such a failure of metacognition is consistent with the poor performance of **Current accuracy**, a measure of whether or not the participant got the next trial correct. Subjective certainty does not accurately predict accuracy on the current example, or vice versa.

While entropy is not a major predictor of certainty, a generalized linear mixed model fit by maximum likelihood provides evidence that it is a significant factor when controlling

for the most predictive model (**Local Accuracy 3 Back Current**), B = 0.097, p = .005. However, despite being significant, the effect of entropy is small, especially when compared to the effect of local accuracy, B = 1.264, p $<$.001. This evidence does not rule out the possibility of unknown factors fully mediating the relationship between entropy and certainty.

Conclusions and Discussion

Our analyses revealed that local accuracy is the best predictor of certainty in our simple concept-learning task. Further, this effect is robust to exactly how accuracy is computed. This means that participants seem to be basing their certainty on their immediate performance—*inferring certainty from their own behavior and feedback.* Specifically, participants seem to be assessing their performance on the past several items along with a guess on whether or not they know the current item. This general pattern is consistent with metacognitive studies showing that often subjects do not understand—or perhaps even remember—the causes of their own behavior (Johansson et al., 2005; Nisbett & Wilson, 1977). Subjects don't directly observe their own cognitive processes and are often blind to their internal dynamics. This appears to be true in the case of subjective certainty reports. They do not appear to reflect an awareness of how much certainty subjects *should* have.

The analyses also help inform us about which factors *do not* drive certainty, and several of these results are surprising. For example, one reasonable theory of certainty posits that participants could be basing their certainty off of their confidence in the MAP hypothesis under consideration (as in hypothesis-testing accounts of learning). Our analyses do not support this account. If participants were basing their certainty off of the MAP hypothesis that they were considering, the MAP predictors would perform much better. Since the MAP predictors do not form well, it is unlikely that learners' certainty relies on internal estimates of the probabilities that most Bayesian learning accounts assume.

Our analyses also reveal that there is still a lot that we do not understand about human certainty. We tested many major, reasonable hypotheses about the factors that drive human certainty, yet the proportion of variance explained by the highest performing predictor here is only 11%. Thus, although local accuracy performs better than other predictors, it cannot be the whole story. The low performance of the proposed predictors here is surprising given that our hypotheses spanned major hypotheses involving metacognitive awareness of both uncertainty and task performance—factors that most people have previously assumed are major drivers of certainty in learners. Further, these results hint that non-metacognitive factors may play a surprisingly substantial role in influencing human certainty. As an example, neurochemical changes in the brain induced by stimulants such as Adderall and Ritalin are known to robustly influence self-reported confidence on performance in cognitive tasks without actually boosting objective measures of task performance (Smith & Farah, 2011). It is possible that a large component of certainty could reflect factors that are almost entirely removed from the veridical probabilities, such as the context of the judgement or differences in individual learners' overall self-confidence or mood.

Acknowledgments

The authors would like to thank the Kidd Lab, and the Computation and Language Lab for providing feedback.

References

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS Comput Biol*, *5*(9), e1000504–e1000504.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Piantadosi, S. T. (2014). *LOTlib: Learning and Inference in the Language of Thought.* available from https://github.com/piantado/LOTlib.

Shannon, C. (1948). A mathematical theory of communities. bell.. 8)/st. *Techn. J*, *27*, 379–423.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Smith, M. E., & Farah, M. J. (2011). Are prescription stimulants smart pills? the epidemiology and cognitive neuroscience of prescription stimulant use by normal healthy individuals. *Psychological bulletin*, *137*(5), 717.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Tormala, Z. L., Clarkson, J. J., & Henderson, M. D. (2011). Does fast or slow evaluation foster greater certainty? *Personality and Social Psychology Bulletin*, *37*(3), 422–434.

Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, *14*(4), 427–442.

Moral Judgments: Studying People with Helping Professions

Maurice Grinberg (mgrinberg@nbu.bg)
Evgeniya Hristova (ehristova@cogs.nbu.bg)
Veselina Kadreva (vkadreva@nbu.bg)
Research Center for Cognitive Science, Department of Cognitive Science and Psychology
New Bulgarian University
21 Montevideo Str., Sofia 1618, Bulgaria

Abstract

While a considerable amount of research is done in the field of moral psychology, to our best knowledge, no systematic study of moral judgments made by professional groups who make moral decisions as part of their occupational duties exists (e.g. firefighters, medical doctors, midwives, police officers). By their training and practice, such professionals are expected to exhibit differences in moral judgment compared to the general population. Here we report data about moral judgments of firefighters and midwives using moral dilemmas in which one person must be sacrificed in order to save more people. The study reveals that midwives and firefighters are considerably less utilitarian compared to a control group of students. Midwives almost never find the utilitarian action to be permissible. This striking result demonstrates that further understanding of the specific mechanisms involved in special professional groups' moral judgment is needed.

Keywords: moral dilemmas; moral judgment; professional biases in moral judgment

Introduction

Experimental research on moral judgment extensively has used the 'Trolley problem' (Foot, 1978). It presents a situation in which a runaway trolley will kill five people if not diverted on a side track, where it will kill one person. Another widely used dilemma – the 'Footbridge dilemma' – describes a similar situation, but in which a bystander should be pushed from a footbridge in front of the trolley in order to save several people whose life is endangered by the moving trolley (Thomson, 1985).

From a utilitarian point of view (based on the comparison of the utility of saving one or several human lives), the situations are identical. However, when people are asked to judge the moral appropriateness of the suggested resolutions, most of them find the resolution permissible in the 'Trolley problem' but not in the 'Footbridge dilemma' (e.g. Greene et al., 2001).

This behavioral dissociation reveals that moral judgment does not universally follow deontological rules or utilitarian calculations, but is shaped by various factors, and probably several cognitive mechanisms stand behind the process. For instance, Greene et al. (2001) claimed that the 'personal' infliction of harm in the 'Footbridge dilemma' as opposed to the 'impersonal' one in the 'Trolley problem,' leads to higher emotional level, which is responsible for the greater proportion of deontological judgments in the 'Footbridge dilemma.'

Further research, using a set of modified scenarios, established that two other important factors also influence moral judgment. Harm is judged as more permissible when the death of the potential victim is inevitable compared to when it is avoidable (the 'inevitability of death' factor) and when it is inflicted intentionally rather than as a side effect (the 'instrumentality of harm' factor). These two factors are considered to be related to rational deliberations (e.g. Hauser et al., 2007; Moore et al., 2008) and therefore subject to training and rules.

On the other hand, life-or-death decisions are common for professionals from different occupational groups (e.g. judges, pilots, medical doctors, public health professionals, etc.). A central question needing extensive exploration is about the role played by special training and professional experience.

Typically, such decisions have significant consequences for many people but no systematic study of the factors underlying them is available. Here, a first step in this direction is made with the participation of representatives of two professional groups – midwives and firefighters.

Goals and Hypotheses

Presently, many experiments on moral judgment are conducted on-line thus including diverse participants. However, to the best of our knowledge, no specific information about the participants' occupation has been used in relation to their moral judgments.

The present study provides the first results of a series of experiments, started recently, focusing on moral judgments of representatives of occupations directly related to moral choice or involving ethical issues (military officers, firefighters, medical doctors, midwives, etc.). Here, we report the results about firefighters and midwives.

Both groups face situations in which they have to make decisions related to life and death. For instance, Chappa, Gonzales, & Stinger (2013) consider the question of measuring firefighters' 'moral courage', which they define as 'making moral decisions despite the circumstances'. One of the items of their questionnaire, aimed at evaluating moral courage, is about the readiness to 'stop and help people' in dangerous situations. According to Sekerka, Bagozzi, & Charnigo (2009), moral courage is a managerial competency and has to be supported by 'proactive organizational ethics'. Even if firefighters have no specific training in moral courage, the corresponding behavior is part of the code of conduct of this profession (Cooper, 1995) and as such is potent to implicitly bias moral judgment outside the professional context.

Similarly, midwives facing termination of pregnancy for fetal abnormality or situations in which birth can endanger the life of the mother, have to get to terms with a moral decision and shared responsibility (Cignacco, 2002; Garel, Etienne, Blondel, & Dommergues, 2007). Some of the difficulties they report are related to the responsibility felt when supporting the patients and coping with their own distress. In Garel et al. (2007), 75% of the midwives reported concern about the child being alive in late termination of pregnancy and 25% of them reported moral conflicts related to personal, cultural, or religious background. Some midwives believed that it was illegal to kill a living unborn child. At the same time, these problems were less pronounced in experienced midwives. The latter draws attention to a factor common to professions involving stress and decision making concerning human life – experience and training can lessen the emotional impact of the responsibility and risks and promote a more utilitarian viewpoint.

Following this brief analysis of specificities related to job experience, we discuss several important factors to be taken into account but which sometimes lead to different hypotheses about the outcomes of the experiments.

The first factor is the practical real-life experience with similar situations involving life-or-death decisions. Both groups are experienced in making judgments and taking action in life-threatening situations. This could make them more tolerant towards intervention in general and hence, more utilitarian.

The second factor is their repetitive experience with highly emotional situations and their practice and ability in emotion management. That, according to the dual-process theory of Greene et al. (2001), should increase the number of utilitarian judgments.

A third factor, specific to the artificial settings of moral dilemma scenarios, could be the underestimation of the presented task based on comparison with real life situations encountered in the professional life of the participants. If this is the case, the judgments of firefighters and midwives will not differ from those of ordinary participants.

A fourth factor may be the fact that firefighters and midwives rarely face situations in which they have to make choices involving saving the life of one person or the lives of five. More often, they are involved in decisions and actions concerning the life of a single person. That factor will imply non-utilitarian judgments based on the awareness of the value of each life.

As noted above, it is well established that physical directness of harm (personal vs. impersonal) and instrumentality of harm (instrumental vs. incidental) are important factors in moral judgments (see e.g. Moore et al., 2008). However, due to the professional experience of both groups, e.g. the physical contact for midwives, it is expected that these factors would be less important for them.

Although the presented considerations apply to both groups, some important differences should also be noted. Midwives usually do not make the moral decisions themselves but only take part in their execution (Cignacco, 2002). Additionally, midwives' professional obligations imply responsibility for individual patients while firefighters are often responsible for a number of people, e.g. when they participate in rescue missions. This difference might contribute to more utilitarian judgments for firefighters than for midwives.

Moreover, firefighters are more regularly exposed to situations that resemble the hypothetical scenarios. This lack of experience in similar situations for midwives could lead to underestimation of the presented task based on comparison with real life situations encountered in their professional life. It is possible that this underestimation would produce less utilitarian judgments for midwives compared to firefighters.

This paper undertakes a systematic comparative exploration of the above considerations based on a standardized set of artificial moral dilemmas that has been extensively used in other contexts and are carefully controlled for uniformity, structure, and content (Hristova, Kadreva, & Grinberg, 2014a, 2014b).

Method

Stimuli and Design

Three groups of participants with different occupations were studied – ***midwives***, ***firefighters***, and ***students***.

Using hypothetical artificial moral dilemmas, three factors significant in moral judgment (Greene et al., 2001, Moore et al., 2008) were explored:

- ***physical directness of harm*** – requiring physical contact (*personal* harm) or mediated through other means (*impersonal* harm);
- ***inevitability of death*** – harm leading to death is inflicted either to a person that is going to die anyway (*inevitable* death), or to a person that is not endangered by the described situation (*avoidable* death);
- ***instrumentality of harm*** – in order to save endangered people, harm is inflicted as an instrument to saving others (*instrumental* harm) or as a byproduct of the undertaken action (*incidental* harm).

Two avoidable and two inevitable situations were used. In order to manipulate *physical directness of harm* and *instrumentality of harm*, only the resolution paragraphs were modified. As it is not possible to have *personal/incidental* resolution, there are three resolution versions for each situation – *personal/instrumental*, *impersonal/instrumental*, and *personal/incidental*. In all situations, there is a constant tradeoff between killing one person and saving five other persons. It is also important to note that all of the endangered and potentially sacrificed persons in the scenarios are adults.

Each situation is followed by the same question: 'Is it permissible to act as described?' with two possible responses – 'Yes' and 'No'.

The following is an example of an *avoidable* dilemma with three possible resolutions:

You are in a factory. You are standing on a platform above a railway track. Some loaded trolleys are moving along the

rails. One heavy loaded trolley is speeding towards five workers as its breaks had suddenly stopped working. There is no time for them to run away and they are going to die.
The trolley could be stopped only if a heavy object is set on its way.
Personal/Instrumental resolution: The only thing that you can do is to push the worker standing next to you on the platform. He is going to fall down on the rails. Together with the tools that he is equipped with, the worker is heavy enough to stop the moving trolley. He is going to die but the other five workers will be saved.
Impersonal/Instrumental resolution: The only thing that you can do is to activate a control button and to release the safety belt of a worker hanging from a platform above the rails. The worker will fall onto the rails of the trolley. Together with the tools that he is equipped with, the worker is heavy enough to stop the moving trolley. He is going to die but the other five workers will be saved.
Impersonal/Incidental resolution: The only thing that you can do is to activate a control button and to release a large container hanging from a platform. It will fall onto the rails of the trolley. The container is heavy enough to stop the moving trolley. On the top of the container there is a worker who will also fall on the rails. He is going to die but the other five workers will be saved.
Is it permissible to act as described? YES/NO

Participants and Procedure

Fourteen firefighters took part in the study. At the time of the study, all of them worked in the same firefighter unit. All of them are male, aged between 25 and 47 (M = 36). They had work experience between 2 and 27 years (M = 11).

Sixteen midwives took part in the study. At the time of the study, all of them worked in an obstetrics and gynecology hospital in the intensive care and post-intensive care units taking care of babies. All of them are female, aged between 32 and 61 (M = 36), having work experience between 10 and 38 years (M = 23).

A group of 30 students – 14 male, 16 female, with age between 18 and 35, (M = 23) – also took part in the study.

Each participant was presented with 12 dilemmas – 4 situations (2 avoidable and 2 inevitable) each with 3 possible resolutions (*personal instrumental*, *impersonal instrumental,* and *impersonal incidental*). For each participant, the dilemmas were presented in a pseudo-randomized order ensuring that the same situation never appears in two consecutive dilemmas. In total, twelve pseudo-randomized orderings were used.

For the midwives and the firefighters, data was collected using paper-and-pencil questionnaires on their workplaces.

Students participated in lab settings, using computerized administration of the stimuli and recording of responses.

Results

Physical Directness of Harm and Inevitability of Death

In the first set of analyses, we used the responses for 8 of the 12 dilemmas. In those 8 dilemmas *physical directness of harm* and *inevitability of death* factors are varied resulting in 2 dilemmas for each combination of the factors' levels. All of those 8 dilemmas are *incidental* dilemmas.

First, we analyzed data for all participants. Then, we performed separate analyses for female participants (midwives and female students) and for male participants (firefighters and male students) in order to control for possible gender differences.

All participants. The mean number of responses 'permissible' was analyzed in a 2x2x3 mixed ANOVA with *physical directness of harm* (*personal* vs. *impersonal*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-subjects factors and *occupation* (*firefighters* vs. *midwives* vs. *students*) as a between-subjects factor.

The analysis revealed a main effect of the *occupation* ($F(2, 57) = 52.06$, $p < 0.001$, $\eta_p^2 = .2$). Midwives (8.6% responses 'permissible') were less utilitarian than both firefighters (28.6%) and students (49.6%) ($p = .068$ and $p < .001$, respectively). Firefighters were less utilitarian than students were ($p = .013$). The results are presented in Figure 1.

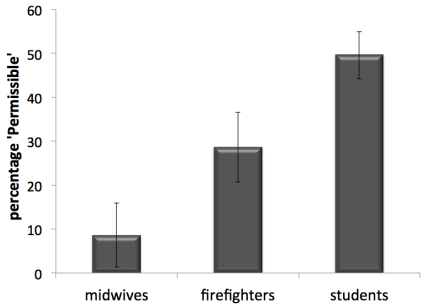


Figure 1: Mean percentage of responses 'permissible' (*YES*) for each occupational group. Error bars stand for SEM.

There was also a main effect of *physical directness of harm* ($F(1, 57) = 28.13$, $p < .001$, $\eta_p^2 = .33$) – *impersonal* harm was judged as more permissible than *personal* harm (40.0% vs. 17.5% 'permissible' responses). There was also a main effect of *inevitability of death* ($F(1, 57) = 20.89$, $p < 0.001$, $\eta_p^2 = .27$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (36.1% vs. 21.8% responses 'permissible'). The results are presented in Figure 2. No significant interaction effects were found.

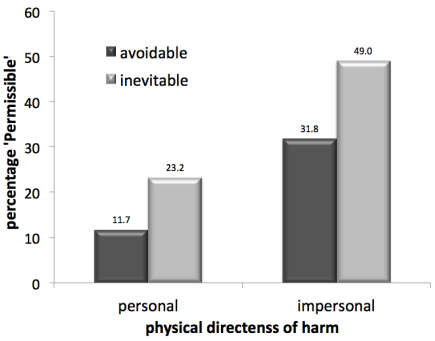


Figure 2: Mean percentage of responses 'permissible' for each dilemma type.

Females only (midwives and female students). The mean number of responses 'permissible' was analyzed in a 2x2x2 mixed ANOVA with *physical directness of harm* (*personal* vs. *impersonal*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-subjects factors and *occupation* (*midwives* vs. *female students*) as a between-subjects factor.

The analysis revealed a main effect of the *occupation* ($F(1, 30) = 12.5$, $p = 0.001$, $\eta_\rho^2 = .29$). Midwives gave fewer responses 'permissible' (8.6%) than the female students (39.8%). Data is presented in Figure 3.

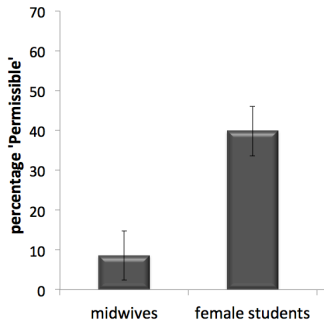


Figure 3: Mean percentage of responses 'permissible' for midwives and female students.

There was also a main effect of *physical directness of harm* ($F(1, 30) = 16.3$, $p < .001$, $\eta_\rho^2 = .35$) – *impersonal* harm is judged as more permissible than *personal* harm (34.4% vs. 14.1% 'permissible' responses) and a main effect of *inevitability of death* ($F(1, 57) = 15.7$, $p < 0.001$, $\eta_\rho^2 = .34$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (31.3% vs. 17.2% responses 'permissible'). The results are given in Figure 4. No significant interaction effects were found.

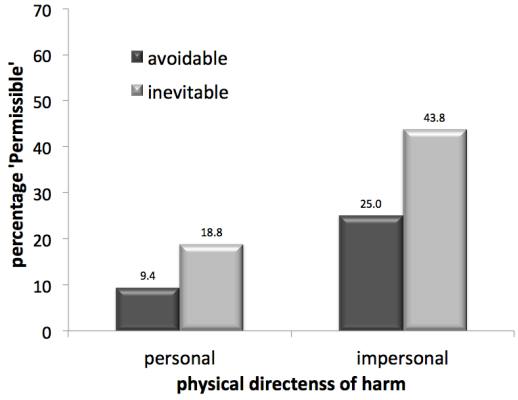


Figure 4: Midwives and female students. Mean percentage of responses 'permissible' for each dilemma type.

Males only (firefighters and male students). The mean number of responses 'permissible' was analyzed in a 2x2x2 mixed ANOVA with *physical directness of harm* (*personal* vs. *impersonal*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-subjects factors and *occupation* (*firefighters* vs. *male students*) as a between-subjects factor.

The analysis revealed a main effect of the *occupation* ($F(1, 26) = 6.9$, $p = 0.014$, $\eta_\rho^2 = .21$). Male students gave more responses 'permissible' (60.7%) than the firefighters (28.6%). The results are presented in Figure 5.

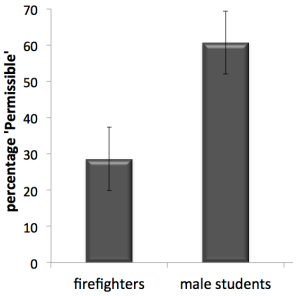


Figure 5: Mean percentage of responses 'permissible' for firefighters and male students.

There was also a main effect of *physical directness of harm* ($F(1, 26) = 14.1$, $p = .001$, $\eta_\rho^2 = .35$) – *impersonal* harm was judged as more permissible than *personal* harm (57.1% vs. 32.1% 'permissible' responses) and a main effect of *inevitability of death* ($F(1, 26) = 8.4$, $p = 0.008$, $\eta_\rho^2 = .24$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (51.8% vs. 37.5% responses 'permissible'). The results are shown in Figure 6. No significant interaction effects were found.

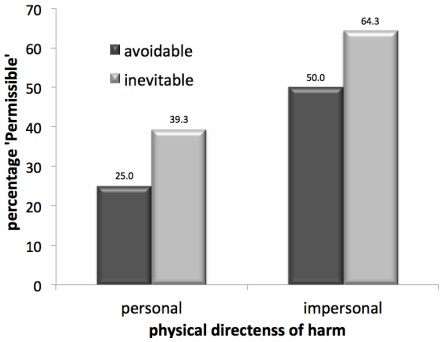


Figure 6: Firefighters and male students. Mean percentage of responses 'permissible' for each dilemma type.

Instrumentality of Harm and Inevitability of Death

For the second set of analyses, we used the responses for 8 of the 12 dilemmas. In those 8 dilemmas *instrumentality of harm* and *inevitability of death* factors are varied resulting in 2 dilemmas for each combination of the factors' levels. All of those 8 dilemmas are *impersonal* dilemmas.

First, we analyzed data for all participants. Then we performed separate analyses for female participants (midwives and female students) and for male participants (firefighters and male students) in order to control for possible gender differences.

All participants. The mean number of responses 'permissible' was analyzed a 2x2x3 mixed ANOVA with *instrumentality of harm* (*instrumental* vs. *incidental*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-

subjects factors and *occupation* (*firefighters* vs. *midwives* vs. *students*) as a between-subjects factor.

The analysis revealed a main effect of the *occupation* ($F(2, 57) = 12.38$, $p < 0.001$, $\eta_p^2 = .30$). Post-hoc tests showed that midwives (23.4% responses 'permissible') are less utilitarian than both firefighters (57.1%) and students (67.5%) ($p = .007$ and $p < .001$, respectively, Bonferroni correction applied). There was no significant difference in the responses 'permissible' between firefighters and students ($p = .81$). The results are presented in Figure 7.

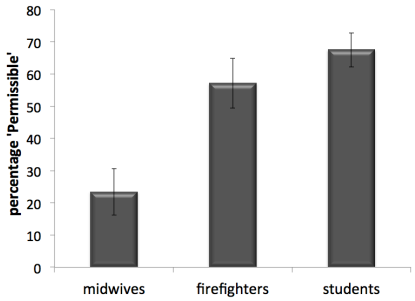


Figure 7: Mean percentage of responses 'permissible' for each occupational group.

There was also a main effect of *instrumentality of harm* ($F(1, 57) = 12.68$, $p < .001$, $\eta_p^2 = .18$) – *incidental* harm was judged as more permissible than *instrumental* harm (58.4% vs. 40.4% responses 'permissible') and a main effect of *inevitability of death* ($F(1, 57) = 22.25$, $p < 0.001$, $\eta_p^2 = .28$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (58.3% vs. 40.4% responses 'permissible'). The data is presented in Figure 8. No significant interaction effects were found.

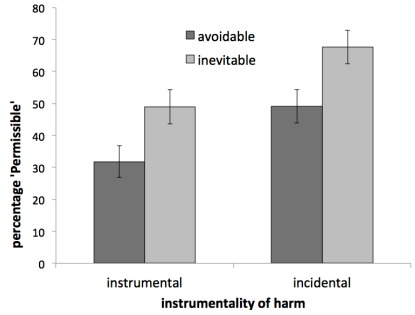


Figure 8: Mean percentage of responses 'permissible' for each dilemma type.

Females only (midwives and female students). The mean number of responses 'permissible' was analyzed in a 2x2x2 mixed ANOVA with *instrumentality of harm* (*instrumental* vs. *incidental*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-subjects factors and *occupation* (*midwives* vs. *female students*) as a between-subjects factor.

The analysis revealed a main effect of the *occupation* ($F(1, 30) = 17.3$, $p < 0.001$, $\eta_p^2 = .37$). Midwives gave fewer responses 'permissible' than female students (23.4% vs. 64.8%). The results are given in Figure 9.

There was also a main effect of *instrumentality of harm* ($F(1, 30) = 16.89$, $p < .001$, $\eta_p^2 = .36$) – *incidental* harm was judged as more permissible than *instrumental* harm (53.9% vs. 34.4% responses 'permissible') and a main effect of *inevitability of death* ($F(1, 30) = 14.4$, $p = .001$, $\eta_p^2 = .32$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (53.1% vs. 35.2% responses 'permissible'). The results are shown in Figure 10. No significant interaction effects were found.

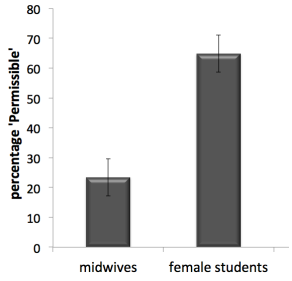


Figure 9: Mean percentage of responses 'permissible' for midwives and female students.

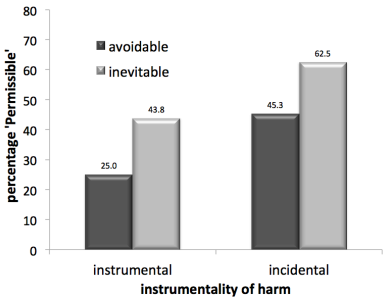


Figure 10: Midwives and female students. Mean percentage of responses 'permissible' for each dilemma type.

Males only (firefighters and male students). The mean number of responses 'permissible' was analyzed in a 2x2x2 mixed ANOVA with *instrumentality of harm* (*instrumental* vs. *incidental*) and *inevitability of death* (*avoidable* vs. *inevitable*) as within-subjects factors and *occupation* (*firefighters* vs. *male students*) as a between-subjects factor.

The analysis revealed a main effect of *inevitability of death* ($F(1, 26) = 7.7$, $p = .010$, $\eta_p^2 = .23$) – killing someone whose death is *inevitable* was judged as more permissible than harming a person whose death is *avoidable* (71.4% vs. 56.3% responses 'permissible'). The data is presented in Figure 11. No other main effects or interactions were significant.

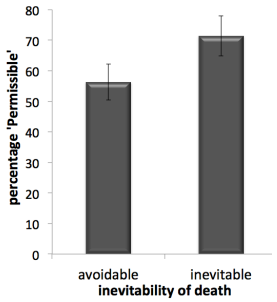


Figure 11: Firefighters and male students. Mean percentage of responses 'permissible' for avoidable and inevitable scenarios.

Summary and Discussion

For the first time in this paper we investigated systematically moral judgments made by representatives of two special professions – midwives and firefighters.

The results show a striking difference in the proportion of utilitarian responses given by midwives, firefighters, and students in Trolley-type moral dilemmas. Midwives are the least utilitarian: they judge the utilitarian action of sacrificing one person to save five others to be morally permissible in less than ten percent of the scenarios. At the same time, students choose the utilitarian response in almost half of the dilemmas. The firefighters are more utilitarian than the midwives are, but they are still much less utilitarian than the students are.

These results demonstrate a large effect of the participants' occupation on moral judgments. As noted above, these professionals are regularly exposed to life threatening situations and they are more likely to encounter and deal with real life moral dilemmas as part of their job-related duties. Because of that and the related coping and emotion management, one possible expectation was that they would be more utilitarian in their judgments. However, the results support the opposite hypothesis based on other factors related to the fact that both occupations aim at helping people, and therefore people from these professions share higher moral values, respect more strongly individual rights and thus, restrain from utilitarian judgements related to life-or-death.

Overall, firefighters gave a lot more utilitarian judgments than midwives. This could be related to the similarity of the artificial moral dilemmas to situations firefighters encounter in their professional activities.

The exploration of physical directness, instrumentality and inevitability of death factors for the two professional groups replicated the well-established effects in the general population. Impersonal and incidental harm was judged as more permissible than personal and instrumental harm respectively. Killing someone whose death is inevitable was judged as more permissible than harming a person whose death is avoidable. No significant interactions with occupation were present, questioning possible interference of professional experience with these factors.

The analyses, controlling for gender, revealed a more complex pattern of moral judgment. For female participants (midwives and female students), all three factors mentioned above shape their moral judgments (physical directness of harm, inevitability of death, and instrumentality of harm). Midwives gave fewer utilitarian responses than female students did. For male participants (firefighters and male students), only the physical directness of harm and inevitability of death had the expected effects on their moral judgments. Instrumentality of harm had no effect on judgments. Firefighters gave fewer utilitarian responses than male students did.

In order to shed more light on possible mechanisms of moral judgment, future research would benefit from more detailed measurement of different factors related to occupational experience and training as well as exploration of more diverse professional groups and better control of age and gender of participants.

Acknowledgements

We gratefully acknowledge the financial support by New Bulgarian University and the invaluable help in data collection of the following people: Iva-Maria Dineva (the midwives' data), and Georgi Manchev (the firefighters' data).

References

Chappa, Olga, Gonzalez, Rebecca, & Stinger, Donna. (2013). The path of measuring moral courage in the workplace. SAM Advanced Management Journal, 78(2), 17.

Cignacco, E. (2002). Between professional duty and ethical confusion: midwives and selective termination of pregnancy. Nursing Ethics, 9(2), 179–191.

Cooper, R. (1995). The fireman: Immaculate manhood. Journal of Popular Culture, 28(4), 139.

Garel, M., Etienne, E., Blondel, B., & Dommergues, M. (2007). French midwives' practice of termination of pregnancy for fetal abnormality. At what psychological and ethical cost? Prenatal Diagnosis, 27(7), 622–628.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5–15.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., Cohen, J. D. (2001). An fMRI investigation of Emotional Engagement in Moral Judgment. Science, 293, 2105–2108

Hauser, M., Cushman, F., Young, L., Kang-Xing, J., & Mikhial, J. (2007). A Dissociation Between Moral Judgments and Justifications. Mind & Language, 22(1), 1–21.

Hristova, E., Kadreva, V., & Grinberg, M. (2014a). Emotions and Moral Judgment: A Multimodal Analysis. In Recent Advances of Neural Network Models and Applications (pp. 413-421). Springer International Publishing.

Hristova, E., Kadreva, V., & Grinberg, M. (2014b). Moral Judgments and Emotions: Exploring the Role of 'Inevitability of Death' and 'Instrumentality of Harm'. In Proceedings of the Annual Conference of the Cognitive Science Society (pp. 2381–2386). Austin, TX:.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. Psychological Science, 19(6), 549–557.

Sekerka, L. E., Bagozzi, R. P., & Charnigo, R. (2009). Facing ethical challenges in the workplace: Conceptualizing and measuring professional moral courage. Journal of Business Ethics, 89(4), 565–579.

Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395–1415.

Children consider others' expected costs and rewards when deciding what to teach

Sophie Bridgers[1] **(sbridge@stanford.edu), Julian Jara-Ettinger**[2] **(jjara@mit.edu) & Hyowon Gweon**[1] **(hyo@stanford.edu)**
[1]Department of Psychology, Stanford University, Stanford, CA 94305
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139

Abstract

Humans have an intuitive sense of how to help and inform others even in the absence of a specific request. How do we achieve this? Here we propose that even young children can reason about others' expected costs and rewards to flexibly decide what is best for others. We asked children to choose one of two toys to teach to another agent while systematically varying the relative costs and rewards of discovering each toy's functions. Children's choices were consistent with the predictions of a computational model that maximizes others' utilities by minimizing their expected costs and maximizing their expected rewards. These results suggest that even early in life, children draw rational inferences about others' costs and benefits, and choose to communicate information that maximizes their utilities.

Keywords: social cognition, prosocial decision-making, communication, pedagogy, naïve utility calculus

Introduction

Life is a series of small and big decisions. While some of these decisions involve choosing what is best for ourselves, frequently they are made with other people in mind. Humans, even in early childhood, often deviate from egoistic behavior and act to benefit others (e.g., Fehr & Fischbacher, 2003; Warneken & Tomasello, 2006).

Concerns for others not only motivate commendable sacrifices but also underlie some of the most mundane communicative acts. Imagine, for instance, that a guest at your house is going to the restroom to wash her hands. To complete this goal, you know she will need to turn on the lights and operate the faucet, and you can predict whether she will have any trouble with these tasks. Based on this, you might prioritize communicating the things that are trickier to figure out (e.g., an unusual faucet, "first turn the lever and pull it towards you"), while leaving simpler things untold (e.g., a normal light switch, "flip up the switch to turn on the lights"). Despite the apparent simplicity of this decision, choosing what to communicate based on what you think others need to know and what they can figure out by themselves is far from trivial.

There are many ways to help and inform others, but our time and resources are limited. Therefore, we must prioritize what to share and how to help. How do we constrain such decisions? Offering help or information can be straightforward when it is clear what our social partners want or need. Under these circumstances, even preverbal infants help others by handing objects out of reach (Warneken & Tomasello, 2006), providing unknown information (Liszkowski, Carpenter, & Tomasello, 2009), and fulfilling others' desires (Repacholi & Gopnik, 1997). However, naïve agents might not know what kind of help they need, or that they need help at all. For instance, your guest has no way of knowing that your faucet is unusual until she attempts to turn it on. As a knowledgeable host, you can anticipate and pre-empt the trouble she might have by telling her how the faucet works in advance. As such, communicative prosociality often relies on our ability to proactively anticipate, select, and provide information others need before they even realize they needed it.

Our example above illustrates an important consideration in proactively informing others: what others are capable (or incapable) of achieving on their own. Though humans learn a great deal through self-guided exploration (e.g., Bonawitz et al., 2011; Schulz, 2012; Sim & Xu, 2014), truly naïve agents (e.g. your guest in the bathroom) may tediously struggle through trial-and-error only to arrive at a poor solution. In these cases, learning from knowledgeable others is particularly valuable. They not only can provide new information, but also can prioritize information that would be too difficult, time-consuming, or perhaps even impossible for the learner to acquire by herself.

Prospectively reasoning about the information others need requires a sophisticated use of Theory of Mind; one must reason about their social partner's knowledge, goals, and competencies, to infer what would be useful for them to know (Strauss, Ziv, & Stein, 2002; Tomasello & Carpenter, 2007). Research suggests children readily consider others' mental states to selectively provide information that is useful for others. For instance, 5- to 7-year-olds adjust the amount of information they teach based on the learner's prior knowledge (Gweon, Shafto, & Schulz, 2014) and goals (Gweon, Chu, & Schulz, 2014), omitting unnecessary or irrelevant information to communicate just enough to support accurate inference. Children's tendency to communicate what is useful and relevant might reflect sensitivity to the value of information; their resistance to provide unnecessary information might reflect their desire to reduce the overall costs of information transfer.

In these studies, however, providing less information meant lower costs both for the children themselves as teachers (in terms of demonstrating the evidence) and for the learner (in terms of processing the evidence). It is thus unclear whether children were simply driven by the desire to reduce their own costs of teaching, or by the desire to reduce the learner's costs of learning. If the latter is true, children should be able to prioritize teaching information that decreases the learner's costs of learning even when children's own costs of teaching are equated.

Here, we show that children decide how to help others by considering the expected costs and rewards from the helpee's perspective. More specifically, we propose that children's choices of potential prosocial actions reflect the use of a *naïve*

utility calculus (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015); children reason about others' costs and rewards and incorporate them into a single concept of utility to prioritize actions that have the highest net utility for others.

From offering physical aid to providing useful information, helping others can take various forms (Tomasello, 2009). Here, we focus on children's information sharing to test our hypothesis. Prior research on children's inferences and evaluations in pedagogical contexts suggests that children hold strong expectations about the information knowledgeable informants ought to select for naïve learners (e.g., Bonawitz et al., 2011; Gweon, Pelton, Konopka, & Schulz, 2014; Gweon, Shafto, & Schulz, 2014), and even act as helpful informants themselves (Gweon, Chu, & Schulz, 2014; Gweon, Shafto, & Schulz, 2014). Thus, children's information selection as teachers is an ideal case study for examining children's prosocial communicative decisions. In the following sections, we first describe our behavioral experiment, then present our theoretical predictions alongside their formal instantiation in computational models, and finally compare these predictions to our empirical data.

Experiment

In the current study, children first learned about two novel, causal toys. They were then told that a naïve learner wanted to figure out both toys, and were asked which toy would be best to teach her if they had to choose just one. We systematically manipulated the relative difficulty for discovering how to activate the toys (costs) and the attractiveness of the toys' causal effects (rewards) across four experimental conditions. Critically, the costs for demonstrating (teaching) each toy were approximately the same, because once children knew how they worked, both toys were easy to activate. Additionally, unlike previous studies, children never actually demonstrated the toy they selected. We made these decisions in order to minimize the effect of children's own costs and rewards on their decisions.

We selected 5 – 7 years as our target age range for two reasons. First, children this age can consider others' mental states to decide what to communicate as teachers (Gweon, Shafto, & Schulz, 2014). Second, they can infer and integrate the expected costs and rewards of others' actions in their reasoning (see Jara-Ettinger, Gweon, et al., 2015). Thus, we expected that children this age might be capable of simulating another person's utility to make decisions on their behalf.

Methods

Participants We recruited 126 5-, 6-, and 7-year-olds (M(SD) = 6.36(0.96) yrs, 54% female) from a local museum. An additional 7 children were excluded from analysis due to difficulty understanding English (5) and unknown date of birth (2). We randomly assigned children to the Play condition or one of four experimental (teach) conditions: (1) Easy-Dull v. Hard-Cool; (2) Easy-Cool v. Easy-Dull; (3) Easy-Dull v. Hard-Dull; (4) Easy-Cool v. Hard-Dull.

Stimuli We constructed four toys that differed in their relative cost of discovery (the difficulty of spontaneously figuring out how to activate the toy; Easy vs. Hard), and the relative reward of their causal effect (its salience; Dull vs. Cool). Easy (low-cost) toys had a single button that generated an effect; figuring out these toys was nearly immediate (2*s* on average) and usually involved one action. Hard (high-cost) toys had one large button and 6 identical small buttons, 5 of which were inert and one that had to be pressed simultaneously with the large button to generate an effect; figuring out these toys took a long time (82*s* on average) and usually involved many failed actions. However, after discovery, both toys were quite easy to activate. Dull (low-reward) toys played music, and Cool (high-reward) toys had a plastic orb that lit up different colors and spun around. The names of the experimental conditions describe the two toys used. Figure 2d provides a schematic of the toys.

Procedure All conditions began with the *Discovery Phase*. The experimenter introduced the participant to two novel toys. She said that she did not know how the toys worked, and needed the child's help to figure them out. Children explored both toys until they activated them. For Hard toys, if children stopped exploring before successfully activating the toy, the experimenter encouraged them to continue with a series of prompts (e.g., "Hmm, I wonder what would happen if you pressed two buttons."). The experimenter delivered these prompts as spontaneous suggestions and with uncertainty to prevent children from thinking she knew how the toys worked. Critically, the experimenter never explicitly told or showed children how to activate the toys. Once children figured out both toys, they were asked to demonstrate each toy twice (order counterbalanced).

Next came the *Choice Phase*. In the experimental conditions, the experimenter told children that she had a friend who knew nothing about the toys but was going to play with them later all by herself. The experimenter said that before her friend played alone with the toys, she would teach her friend how *just one* of the toys worked. Children were then asked: "How should I help her? Which toy should I teach her?" Children selected a toy and were asked to explain their choice. In the Play condition, the experimenter asked children to choose a toy with which they wanted to play (there was no mention of a friend), and allowed the child to play with it briefly.

Last was the *Manipulation-check Phase*. Here, children were asked (1) which toy was *harder* to figure out, and (2) which toy had a *cooler* effect, with the option of answering "same" for both questions.

Computational Modeling and Predictions

To help us interpret our results, we implemented a set of formal computational accounts of our main and alternative hypotheses that capture how children should behave under different considerations. Following recent theoretical and empirical work on social reasoning via a *naïve utility calculus*, these models are based on the assumption that people reason about others in terms of their utilities (Jara-Ettinger, Gweon, et al., 2015; Jara-Ettinger, Tenenbaum, & Schulz, 2015). In

the context of our experiment, we assume that participants estimate each teaching plan's utility, and then choose the plan with the highest net utility for the learner. Here, a *teaching plan* refers to the participant's decision of which toy the experimenter should teach to a naïve learner; since the learner will play with both toys, deciding what is best to teach involves a consideration of both the toy that will be taught and the toy that the learner will figure out on her own.

A learner's expected utility can be described as the difference between her expected rewards and costs. In our experiment, a helpful agent would maximize the learner's overall utility by making a choice that maximizes her expected rewards (showing her a novel toy with high value) and minimizes her expected costs (showing her a novel toy that would require a lot of time and effort to figure out alone).

If children prioritize providing information that would maximize the learner's expected utility, their teaching decisions should reflect the relative costs and rewards of the toys under consideration. More specifically, we test our predictions across four conditions. First, when one toy is both more rewarding than the other and expected to be more difficult (i.e., incur higher costs) for the learner, children should show a strong preference for teaching this toy and leave the other toy (low reward, low cost) for the learner to discover on her own (Easy-Dull v. Hard-Cool condition). When the two toys are equally costly to figure out but differ in reward, children should prefer the toy with a higher reward (Easy-Cool v. Easy-Dull). If the two toys are equally rewarding but differ in cost, children should choose the toy that is higher in cost (Easy-Dull v. Hard-Dull). Finally, when one toy is higher in cost and the other higher in reward, it is unclear which teaching decision would have the higher net utility for the learner, and so children might be split between the two toys (Easy-Cool v. Hard-Dull).

Model Implementation In computing the utilities, our main (full) model considered the learner's expected utilities under each teaching plan as the linear sum of the following components: the cost the learner would incur and the reward she would obtain when activating the toy she was taught (activation cost, C_A, and activation reward, R_A, respectively), the expected cost the learner would incur to discover how the other toy works and the reward she would obtain upon discovery (discovery cost, C_D and discovery reward, R_D, respectively), and the reward the teacher would obtain from activating the toy she teaches (here we assume that the cost of teaching each toy is negligibly different since both toys are easy to activate, but the results are the same if this assumption is removed). That is, given two toys, *Toy X* and *Toy Y*, our model compares two teaching plans: $Teach_X$ (teach how *Toy X* works and leave *Toy Y* for the learner to figure out) and $Teach_Y$ (teach how *Toy Y* works and leave *Toy X* for the learner), and computes the utility of each plan through the equations shown in Table 1.

The cost functions were determined by assuming that each button press results in a constant cost set to 1. (Using a different value does not change our results because all other parameters are defined in terms of this basic cost unit). Thus, a toy's activation cost equals the number of button presses required to activate it. A toy's expected discovery cost (as estimated by participants) was set to the expected number of buttons the learner would need to press to discover how to activate the toy, assuming that the learner would first try simple activation sequences (pressing one button at a time) and only try more complex and costly sequences (trying combinations of two buttons) once the simpler hypothesis space was depleted. Indeed, when participants interacted with the high-cost toys, most children began by pressing each button once and only attempted combinations of two buttons once they had pressed each button individually.

The toys' rewards were treated as variables, allowing for the possibility that different participants assign different rewards to each toy. The music reward was set within a range that captured rewards large enough that could surmount the expected costs of discovery of either toy, as well as rewards that were too small for a teacher to deem the effect worthy of discovery (from 1 to 45; as the expected discovery cost of the more costly toy is ~ 22. Our model results are robust to different reward ranges). For each possible music reward, we set a uniform distribution over the range of rewards for the light-up toy. This range was selected so that the probability that the lights' reward surpasses the music's reward matched the empirical data obtained from children's choice of which effect was "cooler" in the *Manipulation-check Phase*.

Alternative Models While we assume that children consider both the expected costs and rewards of each teaching plan (which includes both the toy that will be taught and the toy that will be left untaught), it is also possible that children instead use simpler strategies and only reason about a subset of these variables.

Thus in addition to the full model, we implemented simpler ("lesioned") models that correspond to four plausible alternative hypotheses. First, children might only be concerned about the relative difficulty of the toys, and so only consider the costs the learner will incur when discovering the untaught

	Activation reward	Activation cost	Discovery reward	Discovery cost	Teacher reward
$U(Teach_X) =$	$+R_A(Toy_X)$	$-C_A(Toy_X)$	$+R_D(Toy_Y)$	$-C_D(Toy_Y)$	$+R_A(Toy_X)$
$U(Teach_Y) =$	$+R_A(Toy_Y)$	$-C_A(Toy_Y)$	$+R_D(Toy_X)$	$-C_D(Toy_X)$	$+R_A(Toy_Y)$

Table 1: R_A and R_D determine the toy's positive reward from activation and discovery, respectively (which are the same), C_A determines the negative activation cost for a knowledgeable agent, and C_D determines the negative discovery cost for a naïve agent.

toy, and when activating the taught toy (costs-only model). Alternatively, once children know how each of the toys work, they might only focus on teaching what is cool and just consider the rewards the learner obtains after successfully activating each toy (rewards-only model). Yet another possibility is that children consider both the costs and rewards but are only concerned with what is being taught and do not think about the experience the learner will have trying to activate the untaught toy (taught-only model). Finally, children might take what is being taught for granted and focus instead on the costs and rewards of having to figure out the other toy (untaught-only model). Table 2 summarizes the different considerations of the full and alternative models. We can assess the relative performance of these different models by comparing how well each model's predictions correspond to children's choices.

		Model Parameters				
		Activation reward	Activation cost	Discovery reward	Discovery cost	Teacher reward
Models	Full Model	✓	✓	✓	✓	✓
	Costs-only	✗	✓	✗	✓	✗
	Rewards-only	✓	✗	✓	✗	✓
	Taught-only	✓	✓	✗	✗	✓
	Untaught-only	✗	✗	✓	✓	✗

Table 2: Model space. Each model captures different considerations that participants can use to decide what to teach.

Results and Discussion

In the *Manipulation-check Phase*, most children thought Hard toys were harder than Easy toys, Cool toys were cooler than Dull toys, and that they were equally hard or cool when these dimensions were held constant; this confirmed that our manipulation of the toys' relative costs and rewards was successful. However, compared to the difficulty manipulation, there was less agreement about the toys' relative "coolness" (i.e., some children thought the music was cooler than the lights). Figure 1 provides a summary of children's responses.

Children's choices of which toy was best to teach varied systematically with respect to the relative costs and rewards of the toys. When one toy was higher in both discovery cost and reward (Easy-Dull v. Hard-Cool), children strongly preferred the experimenter teach that toy (21/25; two-tailed binomial test, $p = 0.001$). When the toys differed only in reward (Easy-Cool v. Easy-Dull), children showed a mild preference for the experimenter to teach the cooler toy (18/26; two-tailed binomial test, $p = 0.076$). [1] When the toys differed only in discovery cost (Easy-Dull v. Hard-Dull), children selected the harder toy (21/25; two-tailed binomial test, $p = 0.001$), and finally, when cost and reward were pitted against each other (Easy-Cool v. Hard-Dull), children's choices were split (10/25; two-tailed binomial test, $p = 0.424$).

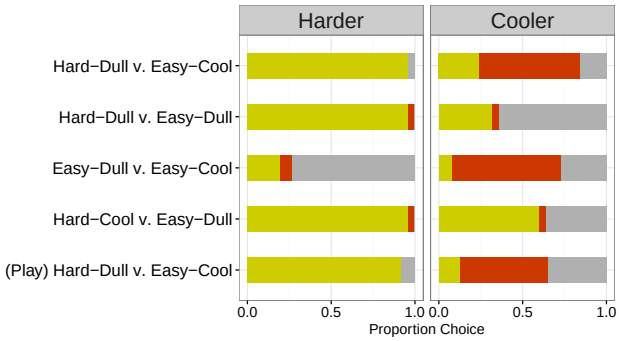


Figure 1: The proportion of children who selected one toy as harder (left panel) and one toy as cooler (right panel) in each condition. Order of toy names per condition reflects order of bars. 'Same' judgments are in grey.

One might think children were at chance in the Easy-Cool v. Hard-Dull condition not because one toy was higher in rewards and the other higher in cost, but rather because the two toys were equally rewarding. Perhaps, adding buttons to the "Dull" toy made it more interesting, leading more children to choose the Hard-Dull toy. To address this possibility, we compared the results from this condition to the Play condition. In this condition, children overwhelmingly preferred to play with the Easy-Cool toy (20/25; two-tailed binomial test, $p = 0.004$), showing a significant difference from their tendency to teach this toy in the Easy-Cool v. Hard-Dull experimental condition (two-tailed Fisher's Exact test, $p = 0.009$). These results confirm that children's responses in the Easy-Cool v. Hard-Dull condition were driven by a consideration of both the expected rewards and costs for the learner.

The pattern of children's choices across conditions suggests they were indeed reasoning about the relative expected costs and rewards of figuring out how different toys worked for a naïve learner and considered these factors when deciding what should be taught. Figure 2a provides a summary of children's responses in the Choice-phase.

Model Fit To evaluate model performance we calculated the likelihood of each model generating the empirical data. Since our models are idealized formalizations, we added a small probability $\alpha \in (0,1)$ that any given participant may get distracted and respond at random. Thus, for each model we assumed that children sample a response from the model's distribution with probability $1-\alpha$ and that they give a random response with probability α. Here we present results with the noise parameter $\alpha = 0.1$ (see Model Implementa-

[1] Children's explanations for why they selected the Easy-dull toy revealed that many considered the music ("Dull" effect) cooler than the lights ("Cool" effect), indicating that they were also choosing the toy with the higher expected rewards.

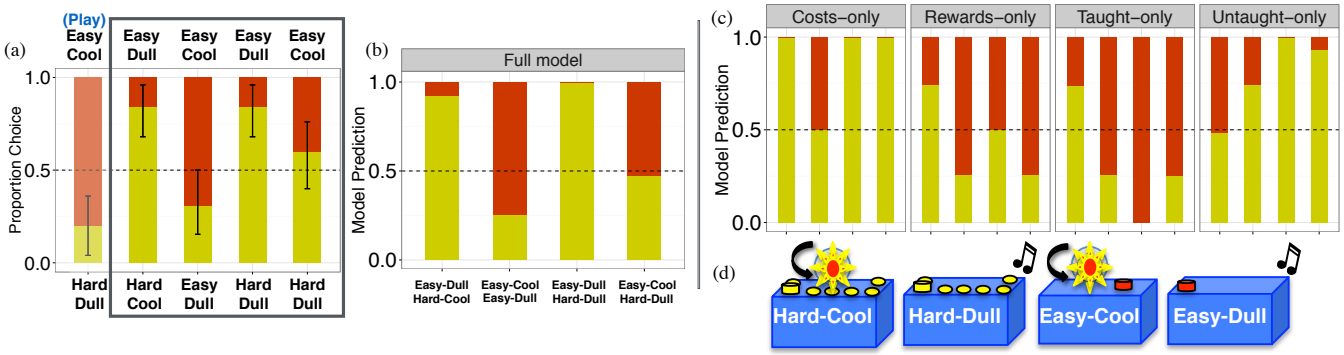


Figure 2: (a) Proportion of children selecting each toy in the *choice phase* (teaching conditions outlined in black and play condition indicated in blue). Error bars are bootstrapped 95% CI. (b) The full model predictions. (c) The predictions of (from left to right): Costs-only, Rewards-only, Taught-only, and Untaught-only models (d) Schematic of the four toys used varying in cost (Easy/Hard) and reward (Cool/Dull). Condition names reflect the two toys used.

tion section). Our qualitative conclusions, however, hold for a wide range of noise and persistence parameters (see `https://osf.io/p29zr/` for model code and predictions).

The full model had the highest likelihood of generating the observed data across conditions. After computing each model's likelihood, we computed the ratio between the full model's likelihood and each alternative model's likelihood. Thus, for each alternative model, a value below 1 indicates that it explains the data better than the full model, and a value over 1 suggests that the full model better explains the data. All model comparisons favored the full model by at least 3 orders of magnitude (all likelihood ratios $>$ 1e4). In short, our empirical data provided compelling support for the full model over the four alternative models.

To better understand why the data are strong evidence for the full model, we consider the alternatives. The alternative models failed to predict participant performance in various conditions. Here we focus on their most salient limitations, which shed light on why the full model captures participants' judgments.

First, the *costs-only* model failed to capture data in the Easy-Cool v. Hard-Dull condition. According to this model, participants should always select whichever toy is most difficult for the learner to discover alone. If this were true, then participants should have preferred to teach the high-cost toy, but instead, their choices were split, as the full model predicted. The *rewards-only* model's critical failure was in the Easy-Dull v. Hard-Dull condition, where it predicted that participants should have no preference, when participants in fact preferred to teach the high-cost toy. The *taught-only* model also failed to predict the Easy-Dull v. Hard-Dull condition, as it predicted that the easier toy should be taught (even for a knowledgeable agent, the low-cost toy is slightly easier to activate than the high-cost toy). Last, the *untaught-only* model most notably failed to predict responses in the Easy-Cool v. Easy-Dull condition. Because the costs were matched, this model predicted that children should teach the less rewarding toy (so as to let the learner discover the high-reward alone). Instead, participants preferred to teach the more rewarding toy, as the full model predicted. Altogether, the full model was the only model that accurately predicted responses in all four teaching conditions. Figure 2b-c provides a summary of each model's predictions.

General Discussion

We propose that children consider others' expected rewards and costs of learning when deciding what to helpfully communicate to a naïve leaner, in a way consistent with a *naïve utility calculus*. Children's behavior in our teaching conditions was unambiguously most consistent with a model that took into account the costs and rewards of both the taught and untaught toys, suggesting children think about what the learner ought to know and what would be more difficult for her to discover on her own. When the toys only differed in value, children selected to teach the cooler toy, maximizing the utility for the learner by increasing her expected rewards. Similarly, when the toys only differed in discovery cost, children selected the harder toy, maximizing the utility for the learner by decreasing her expected costs. When cost and reward were pitted against each other, children were split, suggesting a desire to simultaneously increase the learner's rewards (teaching the cooler toy) and reduce her costs (teaching the harder toy).

It is possible that children who taught the Hard toys did so not because the learner would find them more difficult but simply because children received help on these toys (i.e., experimenter prompts). However, we think this is unlikely. If children were using this simpler strategy, they should prefer the Hard toy in all conditions where it was used, but in the Easy-Cool v. Hard-Dull condition, children were at chance. Children's explanations provide additional support that the learner's expected costs and rewards were indeed informing their choices: Children spontaneously appealed to differences in cost and reward, and their tendency to do so varied appro-

priately with the dimensions of contrast across the particular toys used. For example, they appealed more to reward in the Easy-Cool v. Easy-Dull condition but more to cost in the Easy-Dull v. Hard-Dull condition. In sum children in our experiment did appear to reason about the relative utilities of teaching one toy or the other and selected the teaching plan with the higher overall utility for the learner.

Simulating an unknown person's expected costs and rewards across two goals (toys) and integrating these considerations to determine the best teaching plan requires complicated perspective taking with many unobservable variables. We did not find any age differences in our experiment, but it is possible that younger children would perform differently on this task. Perhaps, a consideration of reward comes online before a consideration of cost, or perhaps the ability to integrate these variables improves across development. Comparing younger children's teaching decisions to older children's and the predictions of our different models would shed light on how a *naïve utility calculus* might develop.

Additionally, it is an open question how much of children's ability to reason about the rewards and costs of another person's exploration depended on their own experience interacting with the toys. Even adults struggle to take the perspective of others who know less or believe something different than they do (Sommerville, Bernstein, & Meltzoff, 2013), so one might expect that after successfully activating the toys, children would fail to recognize that a toy would be hard for a naïve person to figure out. We are currently examining whether children's teaching selections differ when they are explicitly taught how the toys work rather than discover how to activate them through play.

In our experiment and model, the teacher is assumed to take on the learner's utility function and to select what is best to teach based on the expected costs and rewards for the learner. However, teaching is a dyadic situation wherein the teacher's own costs and rewards are presumably also a consideration. Critically in our experiment, we aimed to see if children understood that teaching can reduce the costs and increase the rewards of learning for another person. In order to examine this, we controlled the potential costs and rewards of teaching by making the toys, once known, simple to activate and having children make decisions about what someone else should teach, rather than act as teachers themselves. Exploring children's behavior as actual teachers, as well as situations where certain toys are more rewarding or costly to teach will give us a deeper understanding of how we reason about the utility of communicating certain things over others in real-world social interactions.

Deciding how to act prosocially on behalf of others is no trivial feat. By considering the expected costs and benefits for others of the aid we could provide, we can constrain our decisions of how best to help and inform others. While learning from knowledgeable teachers is useful for acquiring accurate beliefs about the world, not all knowledge is equally useful; some information leads to higher rewards (e.g., knowing where to forage for food), and some leads us to bypass significant costs or even risks (e.g., learning how to hide from predators). The results of the present study provide compelling evidence that children are able to reason about the expected rewards and costs of learning from the perspective of the learner, and that these considerations inform their decisions of what is most useful to communicate.

Acknowledgments

We thank Avani Singh for help with data coding, and Ellen Markman, Noah Goodman, and Josh Tenenbaum for their helpful feedback. We also thank the Palo Alto Junior Museum & Zoo and the families who participated. This work was supported by the Varieties of Understanding grant from the John Templeton Foundation to H.G. and a NSF Graduate Research Fellowship to S.B.

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. E. (2011). The double-edged sword of pedagogy: Instruction limits spontanesous exploration and discovery. *Cognition*, *120*.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*.

Gweon, H., Chu, V., & Schulz, L. E. (2014). To give a fish or to teach how to fish? children weigh costs and benefits in considering what information to transmit. In *Proceedings of the 36th annual conference of the cognitive science society.*

Gweon, H., Pelton, H., Konopka, J., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when agents fail to tell the whole truth. *Cognition*, *132*.

Gweon, H., Shafto, P., & Schulz, L. E. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers inferences about costs and culpability. *Psychological science*, *26*(5).

Liszkowski, U., Carpenter, M., & Tomasello, M. (2009). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*.

Sim, Z., & Xu, F. (2014). Acquiring inductive constraints from self-generated evidence. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Sommerville, J. A., Bernstein, D. M., & Meltzoff, A. N. (2013). Measuring beliefs in centimeters: Private knowledge biases preschoolers' and adults' representation of others' beliefs. *Child Development*, *84*.

Strauss, S., Ziv, M., & Stein, A. (2002). Teaching as a natural cognition and its relations to preschoolers developing theory of mind. *Cognitive Development*, *17*.

Tomasello, M. (2009). *Why we cooperate*. MIT Press.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*.

Disfluency production in speech and gesture

Niloofar Akhavan (nakhavan13@ku.edu.tr)
Department of Psychology, Koç University
Rumelifeneri Yolu Sariyer 34450, Istanbul – TURKEY

Tilbe Göksun (tgoksun@ku.edu.tr)
Department of Psychology, Koç University
Rumelifeneri Yolu Sariyer 34450, Istanbul – TURKEY

Nazbanou Bonnie Nozari (nozari@jhu.edu)
Department of Neurology, Department of Cognitive Science,
Johns Hopkins University
1629 Thames Street, Suite 350, Baltimore, MD 21231, USA

Abstract

The cognitive architecture and function of co-speech gesture has been the subject of a large body of research. We investigate two main questions in this field, namely, whether language and gesture are the same or two inter-related systems, and whether gestures help resolve speech problems, by examining the relationship between gesture and disfluency in neurotypical speakers. Our results support the view of separate, but inter-related systems by showing that speech problems do not necessarily cause gesture problems, and on many occasions, gestures signal an upcoming speech problem even before it surfaces in overt speech. We also show that while gestures are more common on fluent trials, speakers use both iconic and beat gestures on disfluent trials to facilitate communication, although the two gesture types support communication in different ways.

Keywords: gesture, speech production, disfluency

Introduction

People spontaneously produce gestures when they talk. It has been widely accepted that gesture and speech are semantically and temporally coordinated. However, there is mixed evidence on whether speech and gesture form a tightly integrated communication system originating from the same representational system, or whether they are two separate, but interrelated systems (Butterworth & Hadar, 1989; McNeill, 1992, 2005; Alibali, Kita, & Young, 2000; Kita, 2000; Krauss, Chen, & Gottesman 2000; Goldin-Meadow, 2003; Kita & Özyürek, 2003; de Ruiter, 2007; Hostetter & Alibali, 2008; Goldin-Meadow & Alibali, 2013; Pouw et al., 2014). McNeill (1992, 2005) argues that the coordination between speech and gesture arises from the fact that the two emerge from the same system. Others have proposed language and gesture as separate, but interdependent systems. Gesture may influence language (Krauss et al., 2000; de Ruiter, 2000), or gesture can be influenced by language (Hostetter &Alibali, 2008, 2010; Kita & Özyürek, 2003). Even in the absence of overt speech, phonological representations can influence gestures (Nozari et al., 2015).

The current study investigates the interaction between language and gesture production systems, but within a different framework than that of the previous studies. We examined the relationship between disfluency and gesture production in healthy adults to address two questions: 1) Do gesture and language reflect operations of a single system or two separate systems? 2) Does gesture support the language system? In order to better understand the relationship between gesture and speech, we also analyze the temporal relationship between speech and gesture problems.

Do gestures help language production?

One way to examine whether gestures benefit language production is to look at individuals with aphasia who have a variety of difficulties in speech production. The results are mixed. Some report that individuals with Broca's aphasia do not necessarily produce gestures to clarify their incomplete speech (Goodglass & Kaplan, 1963; Cicone et al., 1979; McNeill, 1985; Glosser et al., 1986). In contrast, others demonstrated that they produce more meaning-laden gestures when they have trouble retrieving words than when their production is fluent (Hadar et al., 1998; Lanyon & Rose, 2009; Raymer et al., 2006; Rose & Douglas, 2001; Göksun et al., 2013, 2015).

A second population for studying the benefits of gesture for language production is people who stutter. Gesture production is halted during bouts of stuttering and fluent gesture production is linked to fluent speech production (Mayberry & Jaques, 2000). This close correspondence between fluency and gesture production in this population suggests that gestures simply accompany speech and do not have a compensatory role when speech is problematic. Finally, the potential role of gestures for helping language production has been studied in healthy individuals, using gesture prevention paradigms. Performance on language production tasks usually deteriorates when gestures are prevented (Hostetter, Alibali, & Kita, 2007, but see Beattie & Coughlan, 1999). However, these paradigms require active inhibition of gestures that would have otherwise been

naturally produced. Implementing such inhibition may divert attention from the main task of speaking and cause the increased error rates under circumstances when gesturing is prohibited.

In summary, investigation of the potential benefits of gesture for language production has yielded inconclusive results. Here, we propose a new approach to the same issue by exploring the relationship between disfluency and gesture in neurotypical speakers. Disfluency is an excellent tool for this purpose, because (a) it is a surefire of a glitch in the language production system, and (b) unlike overt errors that are infrequent in speech of neurotypical adults, disfluency rate is estimated to be between 6 (Bortfel et al., 2001) to 26 (FoxTree, 1995) per 100 spoken words in healthy individuals. These characteristics allow us to investigate the relationship between language and gesture in healthy adult speakers without imposing unusual demands on either language or gesture production.

Maclay and Osgood (1959) originally classified disfluency into four categories: (1) Filled pauses are verbal interruptions that do not relate to the proposition of the main message (e.g. *uh* and *um*, *er* and *ah*), (2) Silent pauses are periods of silence longer than the pauses in an equivalent fluent utterance, (3) Repeats are unmodified repetitions of a word, a part of a word, or a string of words (e.g., The girl is running around the *around the* tree). 4) Fillers and Comments that can be in the form of question asking the listener to rehearse the missing words (e.g., *what is this called?*). Disfluency can also entail corrections or repairs to produced words or phrases. Levelt (1989) divides repairs into two main categories: (1) Error repairs, where the original utterance was wrong (e.g., error = "dog", repair = "cat", or error = "cap", repair = "cat"), and (2) appropriateness repairs, where the original utterance was not wrong, but was considered by the speaker to be incomplete or ambiguous (e.g., original utterance = "the pen", repair = "the red pen"). Using a task that elicits production of sentences and gestures, we examine the relationship between these types of disfluency and gesture.

We also examine the timeline of gesture interruption compared to speech interruption. Seyfeddinipur and Kita (2001) found that gesture suspension and resumption points took place before speech disfluency. They interpreted the early interruption of gestures compared to speech as a sign that speakers knew that there was a problem in speech but delayed the interruption until a repair was available.

Iconic vs. beat gestures

Co-speech gestures are classified into several categories. Two of these categories that have received attention for their role in compensating for speech problems are *iconic* and *beat* gestures. Iconic gestures are used as symbols to reenact actions (e.g., drawing a circle in the air to represent doing cartwheel) or to represent concrete objects (e.g., bending the index finger to represent a hook). Beat gestures are flicks of the hand that follow the speech prosody without the gesture conveying semantics (McNeill, 1992).

It has been suggested that producing an iconic gesture, for instance drawing a circle in the air, helps speakers produce the word (cartwheel in this case). Different hypotheses have been proposed for how iconic gestures facilitate production. Some have proposed that gesture helps lexical retrieval (Krauss, 1998), some have posited that gesturing helps packaging of conceptual information (Alibali, Kita, & Young, 2000), and some have argued that it helps create a mental image of the word's referent during lexical search (Wesp, Hesse, Keutmann, & Wheateon, 2001). Regardless of the exact mechanism, all of these accounts maintain that iconic gestures benefit language by helping the retrieval of information from the lexical-semantic system.

On the other hand, beat gestures are believed to be free of semantics and as such, are unlikely to be directly involved in retrieving information from the lexical-semantic system (Krauss & Hadar, 1999, but see Lucero, Zaharchuk, & Casasanto, 2014). Instead, they may have a communicative role by engaging the listener until the speech problem is resolved. Thus, beat gestures may "hold the conversational floor" during speech problems, similar to what has been proposed for filled pauses (e.g., Maclay and Osgood, 1959; but see Clark & Fox Tree, 2002). Given the different roles that iconic and beat gestures play, in examining the role of gesture in compensating for speech problems, we inspect these two types of gesture separately.

Predictions

Regarding our two main questions we have the following predictions. (1) If speech and gesture arise from the same system, we would expect problems in the two to co-occur. But, if gesture and speech are separate systems, at least past the conceptualization point, speech problems may emerge in the absence of gesture problems. (2) To explore the role of gesture for resolving speech problems, we posed two questions: (a) is the primary role of gestures to help when there is a problem in speech? If so, we would expect gestures to mainly arise during disfluent -- compared to fluent -- trials. (b) If gestures' primary role is not to help with speech problems, do they play any role in resolving language problems? If yes, we would expect more disfluent trials with than without gestures. Note that for (a) we divide trials primarily based on the presence or absence of gesture and then inspect which proportion of gesture trials also included a disfluency. On the other hand, in (b) we divide the trials primarily based on the presence of absence of disfluency, and then inspect which proportion of disfluent trials contained a gesture. We also analyze the temporal relationship between speech problems and gesture addresses the question of whether the two systems are highly synchronized (McNeill, 1992; Mayberry & Jaques, 2000) or whether gesture problems foreshadow speech problems (Seyfeddinipur, & Kita, 2001).

Methods

Participants

Twenty monolingual native Farsi speakers (9 females) between the ages of 18 and 30 were tested. Participants lived in Iran, were all right-handed, had normal hearing and vision. All participants gave their written consent for participation in accordance with the guidelines approved by the IRB committee of the Koç University. Two participants' data was excluded. One of these person's gestures were out of the camera frame and the other person's video recording crashed during the coding.

Materials

Participants watched 20 dynamic movie clips, depicting different motion events. Each movie lasted for 3–4 seconds. The clips were previously developed and used in English and Persian (Göksun et al., 2015; Akhavan et al., 2015). All actions were performed by a woman in an outdoor area.

Procedure

All participants were tested individually in their home environment in a silent room. They were instructed to watch each clip and then describe what they saw. No explicit instruction regarding gesture use was provided. Test stimuli were displayed on a Dell laptop in three different randomized orders across participants. Participants received no feedback throughout the testing sessions. The testing sessions were both audio- and videotaped. The camera was set in a position to capture the hands and the body of the participants but not the heads.

Coding

Speech. Participants' speech was transcribed by a native Farsi speaker (first author), and coded for disfluency. The following were coded as disfluency: a) filled pauses (e.g., *uh* and *um*, *er* and *ah*), b) repetitions (e.g., the girl is *running, running* around the tree), 3) fillers and comments (e.g., *What's the word I need …* when used in the middle of a sentence while searching for a word), 4) appropriateness repairs when the speaker repaired an utterance to make it more complete (e.g., *the girl is running around the tree …* elaborating this by saying, *she is running very fast around the tree*).

Gesture. Gestures were coded as iconic or beat gestures. Gesture coding was done manually by the first author of the study using the ELAN software package (Brugman & Russel, 2004). Gesture abnormalities were coded as interruption (suspension of an ongoing gestural unit), repetition (immediate repetition of a gestural unit) and change (suspension of an ongoing gestural unit with an immediate initiation of a new gestural phase).

Speech-gesture relation. Parallel to the speech start-stop, gestures' re-start and interruption points were identified. The start point was coded when the hand started to move. The suspension was coded when an ongoing gestural unit was interrupted at the time of preparation or at the time of a stroke before the action was completed (holding the hand or being retracted back into its preparation position). The resume point was coded when the gesture that was at the static-hold position started a dynamic phase and was completed. Last, the time gap between the onset of a speech problem and gesture, as well as the gap between resumption of fluent speech and gesture were coded in milliseconds.

Results

A total of 356 trials were included in the analyses. Participants produced 307 iconic and 61 beat gestures, and 174 instances of disfluency. Each subject produced at least one gesture of one instance of disfluency. There were only three instances of overt errors and repairs. The rest of the disfluency cases comprised 53.4% filled pause, 27.0% appropriateness repairs, 12.6% comments, and the 7.0% repetitions. On 126 occasions, disfluency was accompanied by 105 iconic gestures and 21 beat gestures.

To test whether gesture and language reflect operations of a single system or two separate systems, we examined how often disfluent speech was accompanied by gestures that also showed a problem, as opposed to problem-free gestures. This analysis was conducted only on iconic gestures for which a problem can be objectively defined. We found that disfluent speech was accompanied by gestures that showed no problems on many occasions (45 instance or 42.9% of the time; X^2 (1, N = 105) = 2.14, p = 0.14). This finding shows that speech problems can occur without any problems in gesture, supporting the view that the two arise from different systems, at least past the conceptual level.

To investigate whether gesture can help resolve speech problems, we first asked if the primary goal of gestures was to help repair speech problems. If true, we would expect reliably more gestures when speech was disfluent, compared to when speech was fluent. Results indicated otherwise; there were significantly more gestures when speech was fluent than disfluent, (242 vs. 126; X^2 (1, N = 368) = 36.57, p < .001). These findings imply that gestures' primary function may not to resolve speech problems.

We then asked whether gesture has any role in resolving speech problems. To answer this question, we looked only at trials where there was speech disfluency. If speakers use gestures to resolve speech problems, we would expect significantly more disfluent speech trials that contain a gesture than those in without a gesture. This was the case for disfluent speech, (126 vs. 48, X^2 (1, N = 174) = 34.97, p < .001).

Next, we examined the pattern of iconic and beat gestures separately on disfluent cases. Of the 126 disfluent cases that were accompanied by gestures, 105 (i.e., 83.4%) were iconic, and the rest (16.6%) were beat gestures. Thus, speakers mainly produced iconic gestures with speech problems, X^2 (1, N = 126) = 56.0, p < .001. A closer examination of the data showed that beat gestures were produced mostly with filled pauses and comments (80.9%), and only in a few cases with

repetitions and repairs. These findings show that on the majority of disfluent trials, speakers employed iconic gestures that, as reviewed earlier, have links to the lexical-semantic system. The use of beat gestures, on the other hand, was confined to cases where the nature of the disfluency implied that the speaker was cueing the listener that they are not done speaking. Together, these results suggest that while the primary role of gestures may not be to resolve problems in speech, speakers often use them when such problems arise.

In the following analysis, we examined the temporal pattern for the 126 cases of disfluent speech accompanied by the gestures, where 60 of the gestures manifested an abnormality and the other 66 gestures emerged normally at the time of the speech disfluency. First, we looked at the 60 cases when the gesture showed abnormality and coincided with the speech disfluency. We had 22 interrupted and 38 repeated gestures. Of the interrupted gestures, 91% co-occurred with filled pauses and fillers. We examined the temporal relation between the gesture interruption point and the disfluency starting point. The results revealed that gestures stopped either simultaneously or before the starting point of speech disfluency. There were significantly fewer cases, in which speech disfluency starting point preceded gesture interruption, X^2 (2, N = 20) = 23.72, p < .001 (see Table 1)

Next, we examined the temporal relationship between when speech was resumed and when gesture was resumed. Gestures were either resumed the same time as the speech was resumed (7 cases) or before speech resumption (7 cases) by an average of 237 ms. Importantly, gestures were never resumed after speech resumption.

Table 1: The timeline of speech disfluency start point and gesture interruption point

	Percentage	**Average time gap (in ms)**
Speech disfluency precedes gesture interruption	5.0%	893.33
Gesture interruption precedes speech disfluency	40.0%	380
Speech disfluency and gesture interruption occur simultaneously	55.0%	0

It should be noted that out of 38 cases of gesture repetition, 95% co-occurred with repetition type of disfluency or appropriateness repairs. Further, the repeated units of the gesture occurred either simultaneously or before the repeated or repaired section of the speech (average of 104 ms gap), (20 vs. 14, X^2 (1, N = 34) = 1.06, p = 0.30). There were only 2 cases where speech repetition preceded gesture repetition. In sum, the analysis of both interrupted and repeated gestures showed that in the majority of cases, changes in the gesture occurred before or simultaneously with problems in the speech.

Recall that among the 126 cases of gestures accompanying disfluent speech, 45 of the gestures were iconic and problem-free. These iconic gestures preceded their semantically targeted speech with the average of 1036 ms gap. Finally, out of 126, the remaining 21 were beat gestures, where 70.3% were produced before the starting point of the speech disfluency. The rest were simultaneously produced with speech disfluency.

Discussion

To our knowledge, this is the first controlled study examining the speech and gesture interaction in the light of disfluency in neurotypical adult speakers. We first asked if language and gestures come from the same of separate -- but related -- systems. Our results supported the latter: on many occasions when there was a problem in speech, gestures showed no trace of the problem. This finding is incompatible theories that propose a tight co-expression of gesture and speech (McNeill, 1992, 2005), and is better aligned with separate but interrelated models of speech and gesture (Krauss et al., 2000; de Ruiter, 2000; Hostetter & Alibali, 2008, 2010; Kita & Özyürek, 2003).

We then turned to a question that has been a focus of many past studies: Do gestures have a supporting role for speech production? This questions has been addressed by examining various populations (e.g., individuals with aphasia, individuals who stutter), but the results have been mixed. The only investigation of this question in neurotypical adults have been through the use of the gesture inhibition paradigm, which poses unusual cognitive demands on the speaker, making the interpretation of the results difficult. By examining the relationship between disfluency during normal speech production, we were able to address this issue from two angles: we first showed that the main role of gesture was *not* to remedy speech problems: gestures were reliably more prevalent on the fluent than disfluent sentences. Without independently assessing the fluency of speech, it is not possible to predict the exact link between gesture and fluency from correlational data. However, from this data, we can conclude that when speakers did encounter problems, they showed evidence of using gestures to help resolve speech problems: significantly more disfluent trials were with than without gesture. Moreover, the majority of these gestures were iconic gestures that are linked to lexical-semantic system. Thus, it is likely that speakers used these gestures to increase the activation of that system and to facilitate lexical retrieval (see also Cook, Yip, & Goldin-Meadow, 2012).

Another explanation could be that gesturing may lighten the verbal working memory (VWM) load (Goldin-Meadow et al., 2001). Speakers were better able to remember verbal items when they gestured during intervening speech than when they did not. Additional evidence in support of this hypothesis showed that speakers with lower VWM capacities

produce gesture more often than those having higher VWM capacities (Gillespie et al., 2014).

Although beat gestures were not the most common type of gestures to accompany disfluent speech, they accompanied some of the trials with filled pauses and comments. These are cases where speakers most clearly signal to the listener that the speech problems are temporary and that speech will be soon resumed. As such, beat gestures, while they may not play a direct role in supporting lexical retrieval (e.g., Lucero et al., 2014), seem to play a social role in communication. This finding explains the prevalence of such gestures during speech (e.g., Beattie & Coughlan, 1999), in the absence of semantic meaning (see also Krauss & Hadar, 1999, Lucero, Zaharchuk, & Casasanto, 2014).

Finally, we examined the temporal relationship between speech and gestures, and found that in most cases gesture interruption of repetition preceded or coincided with the onset of the speech problem. Critically, there were only two instances where gesture problem manifested after the speech problem. These findings imply that speech and gestures are not always temporally synchronized as suggested (McNeill, 1992, 2005; Mayberry et al., 2000) Moreover, it suggests that the two systems are closely connected, such that speech problems can be reflected in gestures even before they surface in speech.

Collectively, these results support a model of separate, but interrelated systems for speech and gesture. In addition, it sheds light on the functional role of gestures: while not specialized for resolving overt problems in speech, both iconic and beat gestures are used by speakers when speech shows problems, although in different capacities.

References

Akhavan, N., Nozari, N., & Göksun, T. (2015). Motion event expressions in language and gesture: Evidence from Persian. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 60–65). Austin, TX: Cognitive Science Society.

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation*, *5*(4), 307-331.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593-613.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90(1),* 35-56.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, *44*(2), 123-147.

Brugman, H., & Russel, A. (2004, May). Annotating Multimedia/ Multi-modal Resources with ELAN. In *LREC*.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic research*, *4*(1), 75-87.

Butterworth, B., and Hadar, U. (1989).Gesture, speech, and computational stages: a reply to McNeill. *Psychol.Rev.* 96,168–174.doi:10.1037/0033-295X.96.1.168

Cicone, M., Wapner, W., Foldi, N., Zurif, E., & Gardner, H. (1979). The relation between gesture and language in aphasic communication. *Brain and Language*, *8*(3), 324–349.

Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73-111.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and cognitive processes*, *27*(4), 594-610.

de Ruiter, J. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

de Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, *7*(1), 21–38.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34, 709–738.

Fraundorf, S. H., & Watson, D. G. (2014). Alice's adventures in um-derland: psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, *29*(9), 1083-1096.

Gillespie, M., James, A. N., Federmeier, K. D., & Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, *132*(2), 174-180.

Glosser, G., Wiener, M., & Kaplan, E. (1986). Communicative gestures in aphasia. *Brain and Language*, *27*(2), 345–359.

Göksun, T., Lehet, M., Malykhina, K., & Chatterjee, A. (2013). Naming and gesturing spatial relations: Evidence from focal brain-injured individuals. *Neuropsychologia*, *51*(8), 1518–1527.

Göksun, T., Lehet, M., Malykhina, K, & Chatterjee, A. (2015). Spontaneous gesture and spatial language: Evidence from focal brain-injury. *Brain and Language, 150*, 11-13.

Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., & Wagner, S. (2001). Explaining math: Gesture lightens the load. *Psychological Science*, *12*, 516-522.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences, 105(27),* 9163 - 9168.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*, 257.

Goodglass, H., & Kaplan, E. (1963). Disturbance of gesture and pantomime in aphasia. *Brain*, *86*(4), 703–720.

Hadar, U., Burstein, A., Krauss, R., & Soroker, N. (1998). Ideational gestures and speech in brain damaged subjects. *Language and Cognitive Processes*, *13*(1), 59–76.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Hostetter, A. B., & Alibali, M. W. (2010a). Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language*, *63*(2), 245-257.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). Does sitting on your hands make you bite your tongue? The effects of gesture prohibition on speech during motor descriptions. In *Proceedings of the 29th annual meeting of the Cognitive Science Society* (pp. 1097-1102). Mawhah, NJ: Erlbaum.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162-185). Cambridge: Cambridge University Press.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16-32.

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261-283). Cambridge: Cambridge University Press

Krauss, R. M., & Hadar, U. (1999). The role of speech related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, 93-116.

Lanyon, L., and Rose, M.L.(2009).Do the hand shave it? The facilitation effects of arm and hand gesture on word retrieval in aphasia. *Aphasiology* 23,809–822. doi:10.1080/02687030802642044

Levelt, W. J. (1989). *Speaking: From intention to articulation. ACL*. MIT Press Series in Natural- Language Processing. MIT Press, Cambridge, Massachusetts.

Lucero, C., Zaharchuk, H., & Casasanto, D. (2014). Beat gestures facilitate speech production. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 898-903). Austin, TX.

Maclay, H., Osgood, Charles E. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15, 19-44.

Mayberry, R., & Jaques, J. (2000). Gesture production during stuttered speech: insights into the nature of speech-gesture integration. *Language and Gesture*, 199-215.

McNeill, D. (1985).So you think gestures are nonverbal? *Psychol. Rev*. 92,350–371. doi:10.1037/0033-295X.92.3.350

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2005). *Gesture and Thought*. Chicago, IL: University Of Chicago Press.

Nozari, N., Göksun, T., Thompson-Schill, S.L., & Chatterjee, A. (2015). Phonological similarity affects production of gestures, even in the absence of speech. *Frontiers in Psychology, 6* (1347). doi: 10.3389/fpsyg.2015.0134

Pouw, W. T., de Nooijer, J. A., van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2014.00359.

Raymer, A. M., Singletary, F., Rodriguez, A., Ciampitti, M., Heilman, K. M., & Rothi, L. J. G. (2006).Effects of gesture+ verbal treatment for noun and verb retrieval in aphasia. *Journal of the International Neuropsychological Society*, *12*(06), 867–882.

Rose, M., & Douglas, J. (2001). The differential facilitatory effects of gesture and visualization processes on object naming in aphasia. *Aphasiology*, *15*(10-11), 977–990.

Seyfeddinipur, Mandana and Kita, Sotaro (2001) 'Gesture as an indicator of early error detection in self-monitoring of speech.' *Proceedings of the ISCA (International Speech Communication Association) Tutorial and Research Workshop. DiSS'01:Disfluency in spontaneous speech' University of Edinburgh, Scotland* .

Wasp, R., Hesse, J., Keutmann, D., & Wheaton, K. (2001). Gestures maintain spatial imagery. *The American journal of Psychology, 114*, 591-600.

Predictions with Uncertain Categorization: A Rational Model

Elizaveta Konovalova (elizaveta.konovalova@upf.edu)
Department of Economics and Business,
Universitat Pompeu Fabra, Barcelona, Spain

Gaël Le Mens (gael.le-mens@upf.edu)
Department of Economics and Business,
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

A key function of categories is to help predictions about unobserved features of objects. At the same time, humans often find themselves in situations where the categories of the objects they perceive are uncertain. How do people make predictions about unobserved features in such situations? We propose a rational model that solves this problem. Our model complements existing models in that it is applicable in settings where the conditional independence assumption does not hold (features are correlated within categories) and where the features are continuous as opposed to discrete. The qualitative predictions of our model are borne out in two experiments.

Keywords: Feature inferences, Categories, Concepts, Predictions, Judgments, Rational Analysis, Bayesian Model

Introduction

According to J. Anderson, 'The basic goal of categorization is to predict the probability of various unexperienced features of objects' (Anderson, 1991). At the same time, humans often find themselves in situations where the categories of the objects they perceive are uncertain. In this article, we propose a computational model of feature prediction under uncertain categorization. We consider settings where an individual perceives some feature(s) of an object that belongs to a particular domain and makes a prediction about the value of an unobserved feature of the object. We assume that the individual has organized her knowledge of the domain into categories. We propose that the decision maker makes predictions according to a posterior distribution derived by application of Bayes' theorem. As such, our model falls into the rational analysis tradition (Anderson, 1991; Marr, 1982).

A number of prior papers have studied feature prediction under uncertain categorization (e.g., Murphy & Ross, 1994, 2010a; Griffiths, Hayes, & Newell, 2012; Papadopoulos, Hayes, & Newell, 2011). They led to interesting insights about whether and how people use categories in making predictions about unobserved features. Most of the existing studies have considered settings in which features are discrete-valued. As we explain below, a limitation of such settings is that, in this context, it is difficult to distinguish whether people do not use categories at all, or make an optimal use of the categories. By contrast, we study a setting where features are continuously-valued. In our setup, the predictions of a model that makes optimal use of the categories (our rational model) and a model that ignores categories altogether sharply differ.

Our model can be seen as an extension of the 'inference component' of Anderson's rational model of categorization (Anderson, 1991). Just as in this landmark model, the decision maker first relies on her knowledge of some feature of the object to derive the posterior probabilities that the object comes from each candidate category. Then, the decision maker uses her knowledge of the structure of each category to make predictions about the value of the unobserved feature. Anderson's model assumed that the within-category feature correlation was zero – an assumption known as *conditional independence*. Our model generalizes Anderson's model to settings where this assumption is relaxed. This extension expands the relevance of the rational model as there are many settings where it does not hold (e.g., Murphy & Ross, 2010a). In virtually all the settings where people believe that there is a causal relationship between two variables (e.g. educational achievement and income, quality and price of consumer goods), the corresponding mental representation invokes a within-category correlation (Rehder & Hastie, 2004).

Existing Paradigm

In the experimental paradigm used in the vast majority of experiments that focused on feature prediction with uncertain categorization, participants are shown a set of items of various shapes and colors divided into small number of categories, typically 4 (Murphy & Ross, 1994). Then they are told that the experimenter has a drawing of a particular shape and were asked to predict its likely color (or similar questions about the probability of an observed feature given an observed feature). An important characteristic of this paradigm is that the categories are shown graphically to the participants. The idea was to avoid complications related to memory and category learning by participants.

Suppose the two features are X and Y and there are 4 categories. Participants are asked to estimate $P(y \mid x)$, the proportion of items with $Y = y$ out of items with $X = x$. There is some evidence participants' predictions that are the same as those implied by a model that focuses on just the 'target' category, that is, the most likely category given the observed feature (Murphy & Ross, 1994). There is also some evidence that participants sometimes make predictions that are the same as those implied by a model that takes into account multiple categories (Murphy & Ross, 2010a). Still, other experiments have found evidence that participants do not pay attention to categories at all but instead are sensitive to the overall feature correlation (Hayes, Ruthven, & Newell, 2007; Papadopoulos et al., 2011; Griffiths et al., 2012).

A limitation of this paradigm pertains to the fact that the features are discrete-valued. This implies that the predictions of a model that ignores categories altogether or makes optimal use of the categories are exactly the same. This is a consequence of the law of total probability. In this case, we have

$$P(y \mid x) = \sum_{c=1}^{4} P(c \mid x) P(y \mid cx), \tag{1}$$

where $P(c \mid x)$ is the proportion of items belong to c out of all the items such that $X = x$, and $P(y \mid cx)$ is the proportion of items with $Y = y$ out of the items that both are in c and have $X = x$.

In order to estimate $P(y \mid x)$, a participant that would ignore the categories would consider all objects with $X = x$ and would respond with the proportion of objects with y among all objects with x. A participant that would consider all 4 categories would compute the proportion of items with y among the items with x *in each category* and then would compute the weighted average by multiplying each of these numbers by her estimates of $P(c \mid x)$. The responses given by the two participants would be *exactly the same*. It is therefore difficult to assess whether the participants use multiple categories (but see Murphy and Ross (2010a) for an attempt to do so using post-prediction questions). When features are continuous, however, the predictions of these two strategies differ.

A Rational Model for Predictions in Continuous Environments

By contrast to the existing paradigm, we consider a setting where the values of the two features are not discrete, but real-valued random variables X (first feature) and Y (second feature). We assume the individual has organized her knowledge of the domain of objects in a set of categories $\mathcal{C}$. Following recent work, we model mental categories using probability distribution functions (*pdfs*) on the feature space (Ashby & Alfonso-Reese, 1995; Sanborn, Griffiths, & Shiffrin, 2010). Let $c \in \mathcal{C}$ be a category. We denote by $f_c(x,y)$ the value of the associated *pdf* at position (x,y) in the feature space, where x denotes the value of the first feature and y denotes the value of the second feature. This *pdf* denotes the prior belief of the individual over positions given that she knows that an object is from category c.

Now suppose that the individual observes that the first feature has value x and predicts the value of the second, unobserved feature. We assume her predictions are driven by her posterior on the value of the second feature given her observation of the first value of the first feature: $f(y \mid x)$. How does the individual compute this quantity, assuming that she does not have a pre-existing mental representation of the (probabilistic) relation between x and y?

We propose that the individual relies on her mental representation of the categories to make the prediction. That is, she will make use of the category *pdfs* she has in memory. More precisely, we propose that the individual's posterior belief on the value of the second feature is a weighted sum of the posteriors obtained for each possible category:

$$f(y \mid x) = \sum_{c \in \mathcal{C}} p(c \mid x) f_c(y \mid x), \tag{2}$$

where $p(c \mid x)$ is the subjective probability that the object comes from category c given the observed feature value x on the first dimension and $f_c(y \mid x)$ is the marginal distribution of value of the second feature, conditional on the fact that the object is in category c and that its first feature has value x.

This model is realistic to the extent the agent can compute the components of the RHS on the basis of her mental representations. Here, we assume she does so by applying the rules of probability calculus. First, consider the posterior distribution of Y given x and c. We have:

$$f_c(y \mid x) = \frac{f_c(x,y)}{\int_v f_c(x,v)dv}, \tag{3}$$

Second, we assume the agent also computes the probabilities that the item comes from each candidate category in a way that is consistent with Bayes' theorem. We have

$$p(c \mid x) = \frac{P(c) f_c(x)}{f(x)} = \frac{P(c) \int_v f_c(x,v)dv}{\sum_{c \in \mathcal{C}} P(c) \int_v f_c(x,v)dv}, \tag{4}$$

where $P(c)$ is the prior on the category. This is the probability that an object about which the individual has no information comes from category c. In the category learning literature, this term is frequently called the 'category bias'.

The predictions of our model follow the rules of probability calculus. Thus, our model makes rational predictions (given the constraints imposed by the mental representation of the categories). Next, we illustrate how the model works by analyzing what happens when the category *pdfs* are bi-variate normal distributions.

Suppose we have two categories ($\mathcal{C} = \{1,2\}$) and that categories can be represented by bi-variate normal distributions as follows:

$$\begin{pmatrix} X_c \\ Y_c \end{pmatrix} \sim N\Big(\begin{pmatrix} \mu_{xc} \\ \mu_{yc} \end{pmatrix}; \begin{pmatrix} \sigma_{xc}^2 & \rho_c \sigma_{xc} \sigma_{yc} \\ \rho_c \sigma_{xc} \sigma_{yc} & \sigma_{yc}^2 \end{pmatrix} \Big), \tag{5}$$

where μ_{xc} and μ_{yc} are the category means on the two features, σ_{xc} and σ_{yc} are the standard deviations on the two features and ρ_c is the within-category correlation for category c. We first assume that there is no within-category correlation, consistent with the conditional independence assumption (Anderson, 1991). Then we consider the general case.

Model Predictions - with Conditional Independence

Assuming conditional independence amounts to assuming $\rho_{c1} = \rho_{c2} = 0$. Some algebra leads to

$$f(y \mid x) = p(c_1 \mid x) f_{\mu_{y1}, \sigma_{y1}}(y) + p(c_2 \mid x) f_{\mu_{y2}, \sigma_{y2}}(y), \tag{6}$$

where f_{μ_y, σ_y} denotes the density of a normal distribution with mean μ_y and standard deviation σ_y, $p(c_2 \mid x) = 1 - p(c_1 \mid x)$, and

$$p(c_1 \mid x) = \frac{1}{1 + e^{ax^2 - bx + c}}, \tag{7}$$

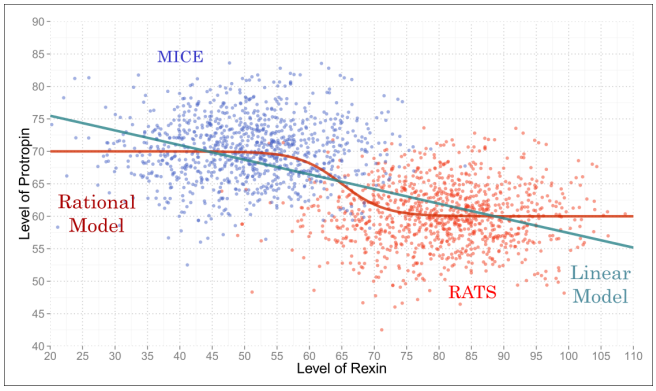


Figure 1: Categories used in Experiment 1. The solid lines represent the mean of the posterior implied by our model and the linear model. Participants were shown the level of 'Rexin' (x-axis) and were asked to predict the level of 'Protropin' (y-axis). See column 'True' in Table 1 for parameter values.

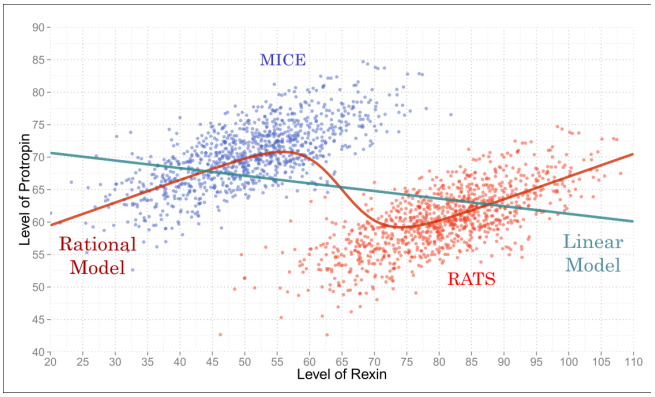


Figure 2: Categories used in Experiment 2. The solid lines represent the mean of the posterior implied by our model and the linear model. The parameters are the same as for Figure 1, except for the within-category correlations: $\rho_R = \rho_M = 0.7$.

with

$$a = \frac{\sigma_{x2}^2 - \sigma_{x1}^2}{2\sigma_{x2}^2\sigma_{x1}^2}, b = \frac{\sigma_{x2}^2\mu_{x1} - \sigma_{x1}^2\mu_{x2}}{\sigma_{x2}^2\sigma_{x1}^2}, c = \frac{\sigma_{x2}^2\mu_{x1}^2 - \sigma_{x1}^2\mu_{x2}^2}{2\sigma_{x2}^2\sigma_{x1}^2} + log\frac{\sigma_{x2}}{\sigma_{x1}}.$$

See Figure 1 for an illustration. The predictions made by the model are sensitive to the relative positions of the categories and make a 'smooth' transition from one category to the other. Due to the fact that it is essentially a version of Anderson's rational model ('AM'), we will refer to this model as 'Anderson's model' in subsequent discussions. Its predictions are different from the predictions of other existing models.

Ignoring the Categories: Linear Model (LM) A simple model that would ignore categories altogether would make predictions according to a regression line with negative slope (see Figure 1 for illustration).

Single Category - Independent Features (SCI) We refer to the most likely category given the observed feature (x) as the 'target' category (this is category 1 if $p(c_1 \mid x) > .5$, as per eq. 7). This is the same as the rational model, but with all the weight on the target category (c^*). In this case, $f(y \mid x) = f_{c^*}(y \mid x)$, where $f_{c^*} = f_{\mu_{y1},\sigma_{y1}}$ if the target category is category 1, and $f_{c^*} = f_{\mu_{y2},\sigma_{y2}}$ otherwise. The mean of the posterior implied by this model follows a 'step function' where the two steps are at $y = \mu_{y1}$ and $y = \mu_{y2}$ and the jump is situated where x is such that $p(c_1 \mid x) = .5$ (with the experimental parameters, this is obtained for $x = 65$).

Model Predictions - General Case

When the conditional independence assumption does not hold, the posterior is given by

$$f(y \mid x) = p(c_1 \mid x) f_{\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}}\rho_{c_1}(x - \mu_{xc_1}), \sigma_{y1}}(y) + p(c_2 \mid x) f_{\mu_{yc_2} + \frac{\sigma_{yc_2}}{\sigma_{xc_2}}\rho_{c_2}(x - \mu_{xc_2}), \sigma_{y2}}(y), \quad (8)$$

where $p(c_1 \mid x)$ is given by the same equation as before (equation 7). See Figure 2 for an illustration. In this case, the prediction of the second feature is influenced by the positions of the categories as well as by the internal structure of the categories (the within-category correlation between X and Y). We will refer to this model as the rational model with possible correlation (RMC).

In the setting of the Figure 2, the mean of the posterior implied by the linear model would be a downward slopping line. The mean of the posterior implied by the Single Category - Independent Features model would be a step function, just as before. Another relevant comparison model is a model that uses just the target category but is sensitive to within-category feature correlations.

Single Category - Correlated Features (SCC) Let c^* be the most likely category given the observation of the first feature. We have $f(y \mid x) = f_{c^*}(y \mid x)$, where $f_{c^*} = f_{\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}}\rho_{c_1}(x - \mu_{xc_1}), \sigma_{y1}}$ if the target category is category 1, and $f_{c^*} = f_{\mu_{yc_2} + \frac{\sigma_{yc_2}}{\sigma_{xc_2}}\rho_{c_2}(x - \mu_{xc_2}), \sigma_{y2}}$ otherwise.

Experiment 1: With Conditional Independence

Participants faced a feature prediction task that closely matches the setting of the previous section. They learnt two categories in a two dimensional feature space and then made a series of predictions about the value of the second feature on the basis of the value of the first feature of an item.

Design

To avoid the influence of unobserved prior knowledge, our experiment used artificial categories. Participants had to assume they were biochemists who studied the levels of two hormones in blood samples coming from two categories of animals. (see Kemp, Shafto, and Tenenbaum (2012) for a similar setup). The hormones were called 'Rexin' and 'Pro-

tropin' and the two categories of animals were 'Mice' and 'Rats.' Similarly to prior literature on feature prediction with uncertain categorization and in order to avoid issues related to memory, we provided the participants with visual representations of the categories in the form of scatter plots of the exemplars of the two categories. The data was generated on the basis of eq. 5 and the parameters in the 'True' column of Table 1 (see Figure 1).

The flow of the experiment was as follows: After reading general instructions, participants were told their lab had a collection of Rat blood samples. They were shown a scatter plot of the levels of Rexin and Protropin in these blood samples and given as much time as they wanted to study it. Then they were asked to make a series of 13 successive predictions of the likely level of Protropin given the level of Rexin found in a blood sample (the scatter plot was visible on the computer screen while participants made the predictions). We asked participants to make these predictions so that they would become familiar with the relation between the levels of Rexin and Protropin (a positive within-category correlation). No feedback was provided about the predictions. Then participants went through a similar procedure for the Mouse blood samples. To conclude this learning stage, participants were shown a graph with the Rat sample data and Mouse sample data (the scatter plots of Figure 1, without the prediction of the rational model).

In the next stage, participants were told that a batch of new blood sample had just arrived at their lab and that these blood samples had already been tested for Rexin. They were also told that the 'label on the blood sample has been erased and thus you do not know if it belongs to a rat or a mouse.' Participants were asked to predict the likely level of Protropin for 48 blood samples. The question was 'What is the likely level of Protropin in this blood sample?'. Participants answered using a slider scale with minimal value 40, maximal value 90, and increments of 1 unit.

Participants

29 participants were recruited via Amazon Mechanical Turk. 1 participant was eliminated due to a technical error that occurred during the experiment.

Results

Models were estimated on a participant-by-participant basis and evaluated in terms of the BIC criterion (minus the log-likelihood minus a penalty increasing in the number of free parameters). Table 1 reports the mean estimated parameter values (across participants). They are close to the true parameter value, which suggests that collectively, participants understood the task and behaved in a way generally consistent with the predictions of the rational model.

We proceed to two model comparisons (see Table 2). In the first comparison, we compare Anderson's rational model (AM), the single category independent feature model (SCI) and the linear model (LM). Anderson's model provides the best fit for 64% of the participants, whereas the two other models provide the best fit for just 18% of the participants. This suggests that most participants generally took into account the two categories when predicting the value of the second feature. Also, they displayed the "smooth" transition between the categories predicted by Anderson's model.

Table 1: Estimated model parameters. Parameters were estimated separately for each participant. The values are the mean estimated parameters across participants. AM: Anderson's rational model, SCI: the single category independent feature model; RMC: rational model with possible within-category feature correlation; SCC: single category model with possible within-category feature correlation; LM: linear model.

Experiment 1						
Param.	**True**	**AM**	**SCI**	**RMC**	**SCC**	**LM**
$\mu_{x,R}$	80	78.5	77.5	81.5	82.1	
$\mu_{y,R}$	60	60.4	60.4	60.9	61.0	
$\mu_{x,M}$	50	48.2	47.5	49.2	52.0	
$\mu_{y,R}$	70	68.8	69.2	68.6	68.1	
σ_x	10	7.1	7.4	7.5	12.5	
σ_y	5	0.2	1.7	2.5	3.3	
$\rho_R = \rho_M$	0	NA	NA	-0.1	-0.2	
α	78					74.7
β	-.2					-0.2
σ^2	5.7					4.1
BIC	NA	233.3	257.7	231.5	260.3	273.6
Experiment 2						
$\mu_{x,R}$	80	79.3	79.9	79.6	80.1	
$\mu_{y,R}$	60	60.8	60.9	61.2	60.8	
$\mu_{x,M}$	50	47.4	49.9	47.1	50.4	
$\mu_{y,R}$	70	69.3	69.1	68.7	68.7	
σ_x	10	3.3	9.9	5.8	10.1	
σ_y	5	5.8	5.0	3.0	4.9	
$\rho_R = \rho_M$	0.7	NA	NA	0.5	0.5	
α	73					66.8
β	-.12					-0.03
σ^2	6.7					6.7
BIC	NA	316.1	317.1	246.2	285.9	326.6

In the second comparison we included two additional models: a version of the rational model with possible within-category feature correlation (RMC) and a version of the single category model with possible within-category feature correlation (SCC). If people behave rationally, according to the task environment, these models should perform more poorly than their equivalents with 0 within-category feature correlation. This is because the true correlation is 0, and these models have the correlation coefficient as one more free parameter. They should suffer some penalty in term of the BIC. We find that Anderson's model provides the best fit for 36% of the participants, about half as many as in the previous comparison. The rational model with within-category correlation (RMC) provides the best fit for 46% of the participants. The other models are the best fitting model for very few participants. These numbers suggest that a number of participants behaved as if there were some within-category feature corre-

lation despite the fact there was none. This could be because it is hard for people to grasp the concept of randomness or the absence of a pattern (Nickerson, 2002). Moreover, Grice's maxim of quantity suggests that participants might not expect the experimenter to show a graph that communicates an absence of relation (Grice, 1975).

Despite this pattern of behavior, this analysis suggests that most participants considered the two categories, because the two models with smooth transitions between categories (AM and RMC) provide a much better fit than the models that focus on a single category of the linear model (LM). Next, we adapt this design to a setting with positive within-category correlation. In this case, our rational model (RMC) and Anderson's model (AM) make distinct predictions.

Experiment 2: With Positive Within-Category Correlation

In addition to testing our model in a setting without conditional independence, we wanted to see whether we could manipulate the propensity of participants to rely on just the target category or the two candidate categories. Many of the studies that had found that participants tend to rely just on the target category included a question that asked participants about the most likely category of the stimulus before making their predictions. There is evidence that the wording of this question affects the propensity to rely on one or multiple categories when making predictions (Murphy & Ross, 2010b; Murphy, Chen, & Ross, 2012; Hayes & Newell, 2009). We included a similar manipulation in our study.

Design

The design of the experiment was the same as in Experiment 1, but with a within-category feature correlation of .7 (see Figure 2). There were 3 conditions. In the control condition, participants were just asked to predict the second feature value upon seeing the value fo the first feature, as in Experiment 1. In the 'MC condition', participants were asked about the most likely category before predicting the value of the second feature. This was multiple choice question: 'From which animal did the blood sample come from?'. The choices were 'Mouse' and 'Rat'. In the 'SL condition', participants were asked the same question, but answered using a continuous slider that went from 'Definitely a Mouse' (left) to 'Possibility a rat or a mouse' (middle) to 'Definitely a Rat' (right). We predicted that the SL condition would make people more aware of the uncertainty about the category of the item and thus increase their propensity to rely on two categories, at least as compared to what happens in the MC condition.

Table 2: Percentage of participants whose feature predictions were best fit by each of the candidate models.

	Experiment 1 Comparison		**Experiment 2** Condition		
Model	(1)	(2)	MC	SL	Control
AM: Anderson	64%	36%	0%	0%	4%
SCI: Single Cat. Indep. Features	18%	4%	0%	0%	0%
SCC: Single Cat. Corr. Features		0%	17%	24%	25%
RMC: Rat. Mod. Corr. Features		46%	72%	69%	63%
LM: Linear	18%	14%	10%	7%	8%
Nb part.	28	28	29	29	24

Participants

102 participants were recruited via Amazon Turk. 20 participants were eliminated from the analysis because their responses seemed very inconsistent with the stimuli.[1]

Results

We fitted the 5 candidate models on a participant-by-participant basis and compared them in terms of the BIC criterion (see Tables 1 & 2). The rational model (RMC) provided the best fit to the data in all 3 conditions (it is the best fitting model for 60 to 70% of the participants). For about 20% of the participants, the best fit is a model that focuses on the most likely category (SCC or SCI).

Comparisons of the percentage of participants for whom the best fit is the rational model or a single category model do not show meaningful differences across categories. In order to uncover differences, we estimated a quasi-rational model with a free parameter that characterizes the propensity to rely on multiple categories. This model assumes that the posterior is given by

$$f(y \mid x) = p(c_1 \mid x)^{\alpha} f_{\mu_{yc_1}+\frac{\sigma_{yc_1}}{\sigma_{xc_1}}\rho_{c_1}(x-\mu_{xc_1}),\sigma_{y1}}(y) + p(c_2 \mid x)^{\alpha} f_{\mu_{yc_2}+\frac{\sigma_{yc_2}}{\sigma_{xc_2}}\rho_{c_2}(x-\mu_{xc_2}),\sigma_{y2}}(y), \quad (9)$$

where $\alpha > 0$. When $\alpha = 1$, the model reduces to the rational model (eq. 8). When α is high, the model becomes close to the single-category correlated feature model (SCC), and when α is close to 0, the model gives about equal weight to both categories, irrespective of the value of the observed feature.

Maximum Likelihood Estimations on a participant-by-participant basis give the following proportions of participants with α higher than 1: MC condition: 83%, SL condition: SL: 62% and control condition: 71%. These proportions are all significantly higher than 50% but not significantly different from each others (one-sided binomial tests with level of .05). Although the differences are not large, the ranking of the three proportions is consistent with our expectations. Maybe more significantly, the fact that α is higher than 1 for

[1] For each participant, we regressed the Y values (the predictions) on X (the values of the first feature, shown to them). Those participants for whom the regression coefficient was significantly positive were dropped from the analysis. The reason is that the rational model can fit such pattern of predictions well with $\mu_{y,R} > \mu_{y,M}$. But such pattern can hardly be considered rational given given the true values of $\mu_{y,R}$ and $\mu_{y,M}$ ($\mu_{y,R} < \mu_{y,M}$ - see Figure 2). Ancillary analyses with all the participants lead to similar results.

most participants in all conditions suggests that most participants gave more weight to the target category than what is prescribed by the rational model.

Discussion

Model comparisons suggest that the rational model provides an appropriate characterization of the behavior of a large proportion of the participants, when compared to other models. This implies that most participants consider the two candidate categories when making predictions about the unobserved feature. This might seem surprising in light of the existing evidence gathered by Murphy, Ross and colleagues that the majority of participants tend to rely on just the target category (in their experiments about 25% of the participants rely on multiple categories). But our analyses with the quasi-rational model suggest that most participants in fact give too much weight to the target category, at least compared to the prescription of the rational model. Seen with this lens, our results are not inconsistent with the prior findings, but rather refine them and extend these to a different experimental paradigm.

Discussion & Conclusion

Our rational model implies that when the category of the item is uncertain, participants should give some weight to the predictions implied by membership in the two candidate categories. This should be the case both under conditional independence or when there is within-category feature correlation. Our empirical results suggest that a majority of participants behaved according to this qualitative prediction. At the same time, most but not all participants gave too much weight to the most likely category. This is broadly in line with prior empirical findings in the literature on category-based feature prediction.

Our model is a computational model and, as such, it does not specify how people might perform the computations that lead to these predictions (Marr, 1982). Nosofsky (2015) recently proposed an algorithmic model that achieves such predictions when the features are discrete-valued. Adapting this exemplar model to the case where feature values are continuous is an interesting avenue for further research.

Finally, a potential limitation of our experiments is that people were watching the data (the scatter plots) when making the predictions. A possible interpretation of our findings is thus that people engaged in some elaborate form of curve fitting on the basis of what they were looking at. A natural next step is to run similar experiments where participants first learn the categories and then make feature predictions on the basis of memorized categories.

Acknowledgments

We appreciate the discussion with participants in the Behavioral and Management Breakfast at UPF, the ConCats seminar at NYU and comments by Mike Hannan and Robin Hogarth. This research was funded by a GSE Seed Grant, MINECO grants #PSI2013-41909-P and #RYC-2014-15035 to Gaël Le Mens.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics speech arts* (pp. 41–58). New York.

Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 576–595.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, *37*(6), 730–743.

Hayes, B. K., Ruthven, C., & Newell, B. R. (2007). Inferring properties when categorization is uncertain: A feature-conjunction account. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 209–214).

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1), 35–73.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. NewYork: Freeman.

Murphy, G. L., Chen, S. Y., & Ross, B. H. (2012). Reasoning with uncertain categories. *Thinking & Reasoning*, *18*(1), 81–117.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive psychology*, *27*(2), 148–193.

Murphy, G. L., & Ross, B. H. (2010a). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, *63*(1), 1–17.

Murphy, G. L., & Ross, B. H. (2010b). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 263–276.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological review*, *109*(2), 330–357.

Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1929–1941.

Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2011). Noncategorical approaches to feature prediction with uncertain categories. *Memory & cognition*, *39*(2), 304–318.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113–153.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010, March). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106.

Numeric Competencies and Anchoring Biases

Sangsuk Yoon (ssyoon@temple.edu), Nathan Fong (nmfong@temple.edu)
Department of Marketing, 1801 Liacouras Walk
Philadelphia, PA 19122 USA

Abstract

Two experiments were conducted to examine the role of three facets of numeracy (objective (ONS), subjective (SNS), and symbolic number mapping (SMap)) in three anchoring tasks (experimenter-given, self-generate, and valuation). We found that the three numeric competencies were associated with different anchoring tasks. SMap was associated with none of the three anchor tasks, while ONS consistently predicted stronger susceptibility to self-generated anchoring. The role of ONS and SNS in experimenter-given and valuation tasks were inconsistent. In Experiment 1, where the direction of adjustment from an anchor is specified, ONS and SNS were positively associated with anchor susceptibility in a valuation task, while they were not in an experimenter-given anchor task. On the other hand, in Experiment 2 where the direction of adjustment from an anchor is uncertain, ONS and SNS were positively associated with anchor susceptibility in an experimenter-given anchor task, while they were not in a valuation task.

Keywords: anchoring effect; numeric competencies; individual differences; decision biases; symbolic number mapping; anchor susceptibility

Introduction

Anchoring refers to a tendency for people's numeric judgments to incorporate an initially considered standard, regardless of its relevance to the given numeric judgment tasks (Tversky & Kahneman, 1974). Anchoring has been examined in diverse domains, including factual knowledge, negotiations, time estimation, physical length estimation, math calculation, medical decisions, and legal sentencing (for a recent review, see Furnham & Boo, 2011). A recent project on replicability of diverse findings in psychological science tested anchoring effects at 36 different labs around the world, and found that anchoring is a robust phenomenon, with an effect size stronger than in the original study (Klein et al., 2014). Moreover, anchoring has been known to be hard to debias (Wilson, Houston, Etling, & Brekke, 1996).

Despite its robustness, the role of individual difference factors in anchoring effect is inconsistent and inconclusive (Bergman, Ellingsen, Johannesson, & Svensson, 2010; Bodenhausen, Gabriel, & Lineberger, 2000; Brandt, Evans, & Crawford, 2014; Furnham, Boo, & McClelland, 2012; Welsh, Delfabbro, Burns, & Begg, 2014). For example, McElroy and Dowd (2007) investigated the role of personality traits on anchoring, and found that *openness-to-experience* to be positively associated with the anchoring effect. Furnham et al. (2012), however, did not find the same result from their study. Studies on the role of cognitive abilities in anchoring have also shown inconsistent results. For example, Bergman et al. (2010) showed that participants with higher cognitive reflection task (CRT) score were less influenced by an anchor than participants with lower CRT score, while Oechssler, Roider, and Schmitz (2009) and Stanovich and West (2008) did not.

Surprisingly, even though most anchoring tasks involve numeric estimation or judgments, the literature on individual differences and anchoring has not shown any significant effects of numeracy on anchoring bias. Numeracy an important cognitive ability, and is a separable construct from other cognitive abilities (Peters & Bjalkebring, 2015). Moreover, previous studies showed that numeric competencies are associated with diverse decision tasks (Burns, Peters, & Slovic, 2012; Dieckmann, Slovic, & Peters, 2009; Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Miron-Shatz, Hanoch, Doniger, Omer, & Ozanne, 2014; Peters, 2012), but only a handful of studies have investigated the role of numeracy on anchoring bias, all finding null effects. For example, Stanovich and West (2008) showed a null effect of numeracy (partially captured from SAT score) on anchoring, and Welsh et al. (2014) investigated the effect of numeracy skill on anchoring susceptibility using the Numerical Ability Test (Bennett, Seashore, & Wesman, 1947). Both studies, however, did not find any significant association between numeracy and anchoring biases.

In this study, we aimed to investigate the role of numeric competencies on anchoring bias by improving two limitations from the previous studies on the role of numeracy in anchoring. First, the previous studies focused on only one facet of numeracy, objective numeracy - the ability to understand and utilize mathematical concepts. However, Peters and Bjalkebring (2015) recently showed that numeracy is not a single construct. They introduced three aspects of numeracy (objective, subjective, and approximate numeracy), and that they are distinct from other. Subjective numeracy is a person's self-perception of their own ability to use and understand numbers, which might be formed by their actual ability to utilize numbers and emotional responses about numbers. Approximate numeracy is the ability to accurately classify numeric magnitude, which is known to be associated with informal math skills in early developmental stages. Peters and Bjalkebring (2015) also showed that each numeric competency is associated with different cognitive tasks. For example, objective numeracy is associated with decision bias in an attractiveness rating task (Bets task; showing lower preference for a non-loss gamble ($9 in 7/36; nothing in 29/36) over a loss gamble ($9 in 7/36; lose 5 cents in 29/36), even though the non-loss gamble has higher

expected value), while subjective and approximation numeracy were not. In another task, where participants were asked to state monetary amounts that made them indifferent to risky gambles with high gains, they showed that objective and approximation numeracy were associated with closer valuation of the indifferent point for the sure thing to the expected value of the risky gamble, while subjective numeracy was not. Based on the finding that different numeric competencies are associated with different decision tasks, in this study, we used three different facets of numeracy to investigate the role of numeracy in anchoring biases.

Second, we investigated the role of numeracy on three different types of anchoring tasks: experimenter-given anchors, self-generated anchors, and valuation tasks. The previous literature on the relationship between numeracy and anchoring has focused on only experimenter-given anchors for factual questions. In the experimenter-given anchor task, participants are asked to answer a series of factual questions (e.g., the length of Mississippi river), preceded by a comparison (anchoring) question whether the answer for the given question is lower/higher than an anchor (Jacowitz & Kahneman, 1995). Self-generated anchor tasks (e.g., the freezing point of vodka) also uses factual questions, but a comparison question is not provided (Epley & Gilovich, 2001). They can use self-generated anchors by retrieving relevant quantities (e.g., the freezing point of water), and previous studies show that the mechanisms are different between experimenter-given and self-generated anchor tasks (Epley & Gilovich, 2001, 2005). Valuation tasks can also be regarded as a type of experimenter-given anchor task, but an individual's preferences plays into an important role. Indeed, valuation studies showed weaker and inconsistent results compared to other anchoring tasks. For example, Ariely, Loewenstein, and Prelec (2003) and Bergman et al. (2010) showed that a positive linear relationship between numeric anchor and willingness-to-pay, while Fudenberg, Levine, and Maniadis (2012) and Maniadis, Tufano, and List (2014) failed to replicate the effects. Based on the fact that different mechanisms are associated with different anchoring tasks, we tested the role of the three facets of numeracy in these three different types of anchoring tasks in two experiments.

General Method

Numeracy Measures

For measuring numeric competencies, we used the same measures and protocol used in Peters and Bjalkebring (2015). For the objective numeracy scale (ONS), we used an eight-item scale developed by Weller et al. (2013). An example of an ONS item is: "If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1,000?" We used the number of correct answers as a single score for ONS. The Cronbach's alpha of Experiment 1 was .74, and that of Experiment 2 was .66.

For the subjective numeracy scale (SNS), we used an eight-item scale developed by Fagerlin et al. (2007). An example of a SNS item is: "how good are you working with fractions?" We obtained a single score by adding up the ratings. The Cronbach's alpha of Experiment 1 was .85, and that of Experiment 2 was .86.

For measuring approximate mapping competence, we used a symbolic number mapping task, and followed the same procedure introduced in Peters and Bjalkebring (2015). Participants were asked to make marks corresponding to six numbers (4, 6, 18, 71, 230, and 780) on a 165-mm horizontal line with endpoints labeled 0 and 1,000, where each question was presented on a separate sheet. For the scoring, we followed the same two-step procedure used in prior studies (Peters & Bjalkebring, 2015; Schley & Peters, 2014). First, we obtained absolute deviation from the target length for each question, and summed across all the absolute deviations. Next, we log-transformed the summed absolute deviation to reduce positive skew issues, and multiplied by -1 so that a higher score indicates higher symbolic mapping ability.

The three numeric competency scales were positively correlated with each other, consistent with the literature (Table 1), but the correlation was not extremely high. This indicates that they are related but separable constructs (Peters & Bjalkebring, 2015).

Table 1: The Correlations Between Numeracy Scales (***p < .001)

	Experiment 1		Experiment 2	
Variable	1	2	1	2
1. ONS	-		-	
2. SNS	0.41***	-	0.46***	-
3. SMap	0.50***	0.34***	0.32***	0.35***

Anchoring Tasks

For the experimenter-given anchor task, we used the same eight stimuli used in Study 2 of Brandt et al. (2014) for Experiment 1, while we used six items used in Jacowitz and Kahneman (1995) for Experiment 2. For the self-generated anchor task, we used the six questions used in Epley and Gilovich (2006). For the valuation task, we used the six market goods used in previous studies on valuation anchoring (Ariely et al., 2003; Yoon, Fong, & Dimoka, 2013). An example of each task is presented in Table 2. In the valuation task, participants were asked to state the maximum amount they would be willing to pay (WTP). We calculated individual-level anchor susceptibility with two procedures: first, we calculated the distance from the given anchor (anchor distance) and rank-transformed the anchor distance to correct right skewedness (Brandt et al., 2014; Klein et al., 2014). After rank-transformation, we rescaled the score to be within a range from 0 to 1. This rescaled

anchor distance was a dependent variable in both experiments. In Experiment 1, we explicitly presented the direction of adjustment (e.g., the length of Mississippi River is longer/shorter than 70 miles), as used in previous studies on anchor susceptibility (Brandt et al., 2014). In Experiment 2, we used classic anchors in which the direction of adjustment is uncertain (e.g., do you think the length of Mississippi River is longer or shorter than 70 miles?), since it has been shown that anchoring effect is stronger when the direction of adjustment is uncertain than when it is certain (Simmons, LeBoeuf, & Nelson, 2010).

Table 2: An Example of Experimenter-Given, Self-Generated, and Valuation Anchor Task

Experimenter-Given Anchor Task
What is your best estimate for the length of the Mississippi River? (low: 70 miles; high: 2,000 miles)
Self-Generated Anchor Task
What is the freezing point of vodka? (self-generated anchor: 32F, the freezing point of water)
Valuation Task
What is the most you would be willing to pay for this cordless mouse?

Experiment 1

Methods

Participants: A total of 216 participants (M_{age} = 35.75, *SD* = 10.80, Female = 42%) were recruited from Amazon MTurk, receiving 50 cents in exchange for participating in a 15-minute study. Participants were recruited with the criteria that location is U.S. only, approval rate is greater or equal to 97%, and the number of times approved is greater or equal to 1,000 times.

Procedure: Participants were randomly assigned to one of the three conditions: experimenter-given anchor (N = 75), self-generated anchor (N = 70), or valuation tasks (N = 71). For experimenter-given anchor and valuation tasks, half of the questions were given with high anchors, and the other half was given with low anchors (a within-subject design). The order was counterbalanced across participants. After completing anchoring task, participants responded to the three numeracy tests in a randomized order. At the end of the study, participants were asked to provide demographic information (e.g., age and gender). To test the anchoring effect, we regressed numeric estimates (or WTP) on anchoring condition (categorical variable, coded 0 for low anchor and 1 for high anchor) and question fixed effects. To examine the role of numeracy in anchoring, we used hierarchical linear regression analysis with random coefficients for each participant, using Stata 14 software.

Results

Experimenter-Given Anchor Task: We found a significant anchoring effect: the answer for high anchor questions was significantly higher than that for low anchor questions (b = 0.16, p < .001). Similar to previous studies (Welsh et al., 2014), ONS was not associated with anchor distance. Indeed, the other numeracy scales (SNS and SMap) were not significantly associated with anchor distance in experimenter-given anchor task (Table 3, column 1).

Self-Generated Anchor Task: The result shows that ONS is negatively associated with anchor distance, while SNS and SMap were not. This implies that participants with higher score in ONS are more susceptible to self-generated anchoring (Table 3, column 2).

Valuation Task: For the valuation task, similar to previous valuation tasks, we found a weaker anchoring effect compared to the experimenter-given anchor task (b = 0.06, p = .090). The results of hierarchical linear regression analysis for the effect of numeracy on anchoring effect showed that ONS and SNS were significantly associated with anchor distance (positively), while SMap was not (Table 3, column 3). This implies that high ONS and SNS predict lower anchor susceptibility in valuation task.

Table 3: Multilevel Regression Results in Experiment 1 (each numeracy scale was run separately)
(* p < .05, ** p < .01, standard errors are in parentheses)

	(1) Experimenter-Given	(2) Self-Generated	(3) Valuation
ONS	0.119	-0.185**	0.152**
	(0.064)	(0.055)	(0.055)
SNS	0.131	-0.145	0.211**
	(0.077)	(0.078)	(0.072)
SMap	0.170	-0.132	0.038
	(0.116)	(0.072)	(0.079)
N	575	420	374

Discussion

In Experiment 1, we explored the relationships between three facets of numeric competencies and three types of anchoring tasks. For the experimenter-given anchor task, we found a null effect of the three numeracy scales on anchoring biases. This result is in line with the previous literature on the role of ONS on anchoring effect (Stanovich & West, 2008; Welsh et al., 2014). We additionally found that SNS and SMap are not significant predictors of anchoring biases. For the self-generated anchor task, ONS significantly predicted anchor distance, but the direction was toward self-generated anchor, which indicates higher anchor susceptibility. This might be due to the fact that high ONS participants may have considered self-generated

anchors more and may have been influenced by self-generate anchors, similar to the result of higher decision bias in the Bets task in Peters and Bjalkebring (2015). Peters and Bjalkebring (2015) explained that the stronger bias for high ONS participants in the bets task might be driven by increased processing of available meaningful numerical information (comparing \$9 gain and losing 5 cents) for the high ONS participants. We conjecture that more active use of self-generated anchors by high ONS participants might contribute to their higher susceptibility to self-generated anchors. For the valuation task, both ONS and SNS were positively associated with anchor distance, which indicates less anchor susceptibility, while SMap was not a significant predictor of anchor distance.

In Experiment 1, we explicitly specified the direction of adjustment for the comparison question (anchor) in experimenter-given and valuation tasks. Simmons et al. (2010) showed that participants were less susceptible to anchors when the direction of comparison question is explicitly specified. Brandt et al. (2014) also showed a stronger or similar anchoring effect when the direction of adjustment is uncertain than when it was certain. Moreover, the directed anchor in valuation task was informative to the given task (e.g., the market price of the item higher than \$10) by informing participants about market price or by having participants consider market price of the market goods used in this study. In Experiment 2, we examined the role of numeracy in anchoring biases using classic uncertain direction comparison questions as anchors and employing a between-subject design for high/low anchors to rule out the possibility that the results were contributed from the characteristics of within-subject design.

Experiment 2

Participants and procedure: A total of 353 (M_{age} = 35.73, *SD* = 11.77, Female: 46%) participants were recruited from Amazon MTurk with the same recruitment criteria as in Experiment 1. Overall procedures were similar to Experiment 1, except that high/low anchor was a between-subject factor and the direction of adjustment for the anchor question was not explicitly specified. Therefore, participants were randomly assigned one of the five conditions: experimenter-given low anchor (N = 69), experimenter-given high anchor (N = 75), self-generated anchor (N = 64), valuation low anchor (N = 72), and valuation high anchor (N = 73).

Results

Experimenter-Given Anchor Task: We found a significant anchoring effect (b = 0.58, p < .001): the answer for the high anchor group was significantly greater than that for the low anchor group. We found that both ONS and SNS are positively associated with anchor distance, while SMap was not (Table 4, column 1). This indicates that high ONS and SNS participants may be less susceptible to experimenter-given anchor.

Self-Generated Anchor Task: Consistent with Experiment 1, only ONS was negatively associated with anchor distance, while the other two numeracy measures were not associated with anchor distance (Table 4, column 2). This result replicates the finding that higher ONS predicts higher susceptibility to self-generated anchor.

Valuation Task: In Experiment 2, we found a significant positive effect of anchoring on WTP (b = 0.32, p < .001). In Experiment 1 where anchor was market price, we found a marginally significant anchoring effect, while in Experiment 2 where anchoring was their willingness-to-buy for a random price, we found a significant anchoring effect. This might imply that providing price information or leading participants to think about the market price of an item seem to reduce anchoring effects in valuation tasks. For the effect of numeracy on anchoring, none of the numeracy measures were associated with anchor distance (Table 4, column 3).

Table 4: Multilevel Regression Results in Experiment 2
(each numeracy measure was run separately)
(* p < .05, standard errors in parentheses)

	(1) Experimenter-Given	(2) Self-Generated	(3) Valuation
ONS	0.161*	-0.196*	-0.020
	(0.058)	(0.091)	(0.083)
SNS	0.187*	-0.095	0.149
	(0.072)	(0.095)	(0.083)
SMap	0.071	-0.010	0.158
	(0.074)	(0.089)	(0.116)
N	848	384	788

Discussion

In Experiment 2, we examined the role of three facets of numeracy on anchoring biases, with an uncertain anchor where the direction of adjustment was not clear. While participants generally adjusted their estimates in the direction specified in the description in Experiment 1, participants could adjust both higher or lower than a given anchor in Experiment 2. Previous literature has shown that anchoring effect is stronger when a comparison question is uncertain than when the comparison anchor specifies a certain direction to be adjusted (Brandt et al., 2014; Simmons et al., 2010). Therefore, in Experiment 2, we tried to replicate findings in Experiment 1 with a study design wherein experimenter-given anchor shows stronger effect. Compared to Experiment 1, we found slightly different patterns regarding the role of numeracy on anchoring biases. Consistent with Experiment 1, we found that only ONS significantly predicted higher susceptibility to self-generated anchor. In contrast, we found the opposite pattern in experimenter-given anchor and valuation tasks. ONS and SNS significantly predicted lower anchor susceptibility in

experimenter-given anchor task, which is inconsistent with previous studies (Stanovich & West, 2008; Welsh et al., 2014) and Experiment 1. For the valuation task, unlike Experiment 1, we did not find significant effects of ONS and SNS on anchor distance. The results seem to imply that different types of anchor comparison might interact with ONS and SNS in experimenter-given anchor and valuation tasks.

General Discussion

In this study, we examined the distinctive role of three facets of numeracy (objective, subjective, and symbolic number mapping) in anchoring biases by conducting two experiments. In Experiment 1, we examined three different anchoring tasks with a certain anchor where the direction of adjustment was explicitly stated, and found that ONS was associated with higher anchor susceptibility in the self-generated anchor task, and was associated with lower anchor susceptibility in the valuation task. SNS was associated with lower anchor susceptibility in the valuation task, but was not associated with the other two anchoring tasks. Finally, SMap was associated with none of the anchoring tasks used in this study.

In Experiment 2, we tried to replicate the findings of Experiment 1 with an uncertain anchor where the direction of adjustment was not stated. Since previous literature has shown that the anchoring effect is weaker when the direction of adjustment is clear (Brandt et al., 2014; Simmons et al., 2010), to rule out the possibility that the findings were contributed by the characteristics of the specified direction of adjustment, we employed the classic anchoring paradigm where the direction of adjustment is uncertain. We found the same result for the self-generated anchor task: only ONS significantly predicted higher susceptibility to self-generated anchoring, replicating the findings of Experiment 1. Therefore, we conclude that ONS is associated with higher anchor susceptibility in self-generated anchor tasks. For the other two tasks, however, we found the opposite patterns. Contrary to Experiment 1, ONS and SNS significantly predicted lower anchor susceptibility in the experimenter-given anchor task, while they did not predict anchor susceptibility in the valuation task. The conflicting results seem to imply that there could be a possible interaction between informativeness of an anchor (e.g., direction of adjustment, market price) and the two numeracy scales (ONS and SNS). Future research may be required to clarify the relationship between informativeness of an anchor and numeracy competencies in experimenter-given and valuation tasks.

In this study, the results of two experiments showed that the three different facets of numeracy were associated with different anchoring tasks, but there are still several limitations. First, even though we showed that ONS predicted higher anchor susceptibility to self-generated anchors, it is not clear that participants actually used the self-generated anchor in the task. Future research may be needed to test how actively participants use self-generated anchor depending on numeracy. Second, we did not find any association between SMap and the three anchoring tasks. The main distinctive feature of approximate numeracy is non-symbolic number-related capability, but the SMap task we used in this study is involved in symbolic number processing (mapping symbolic numbers to non-symbolic magnitudes). Even though SMap is highly correlated with other types of approximate numeracy tasks (e.g., dot-discrimination, dot-line, dot-ratio tasks), a recent study showed that they are separable (Chesney, Bjalkebring, & Peters, 2015). Future research is needed to test the relationship between approximate numeracy and anchoring biases using other approximate numeracy tasks where only non-symbolic mapping is involved. Indeed, a previous study showed that higher SMap was associated with a less concave shape in value function (Schley & Peters, 2014) but we did not find a significant effect in the valuation task. One limitation of the valuation task in this study is that the task was not incentive compatible. Non-incentive compatible methods might reduce participants' commitment to the valuation task. Future research may be required to more precisely investigate whether SMap moderates the relationship between anchoring and WTP using incentive compatible methods as other anchoring valuation studies used (Ariely et al., 2003; Fudenberg et al., 2012; Yoon et al., 2013). Indeed, further research is needed to test whether SMap is associated with other anchoring tasks using informal math questions (e.g., physical length estimation (LeBoeuf & Shafir, 2006), or time estimation (Thomas & Handley, 2008)).

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics, 118*(1), 73-105. doi:Doi 10.1162/00335530360535153

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1947). Differential Aptitude Tests.

Bergman, O., Ellingsen, T., Johannesson, M., & Svensson, C. (2010). Anchoring and cognitive ability. *Economics Letters, 107*(1), 66-68.

Bodenhausen, G. V., Gabriel, S., & Lineberger, M. (2000). Sadness and susceptibility to judgmental bias: The case of anchoring. *Psychological Science, 11*(4), 320-323.

Brandt, M. J., Evans, A. M., & Crawford, J. T. (2014). The unthinking or confident extremist? Political extremists are more likely than moderates to reject experimenter-generated anchors. *Psychological Science*, 0956797614559730.

Burns, W. J., Peters, E., & Slovic, P. (2012). Risk perception and the economic crisis: a longitudinal study of the trajectory of perceived risk. *Risk Analysis, 32*(4), 659-677.

Chesney, D., Bjalkebring, P., & Peters, E. (2015). How to estimate how well people estimate: Evaluating measures of individual differences in the approximate number

system. *Attention, Perception, & Psychophysics, 77*(8), 2781-2802.

Dieckmann, N. F., Slovic, P., & Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. *Risk Analysis, 29*(10), 1473-1488.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*(5), 391-396.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making, 18*(3), 199-212.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic Why the adjustments are insufficient. *Psychological Science, 17*(4), 311-318.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672-680.

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the robustness of anchoring effects in WTP and WTA experiments. *American Economic Journal: Microeconomics, 4*(2), 131-145.

Furnham, A., Boo, H. C., & McClelland, A. (2012). Individual differences and the susceptibility to the influence of anchoring cues. *Journal of Individual Differences, 33*(2), 89.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21*, 1161-1166.

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, Š., Bernstein, M., & Nosek, B. (2014). Investigating variation in replicability: The "many labs" replication project. Retrieved from Open Science Framework.

LeBoeuf, R. A., & Shafir, E. (2006). The long and short of it: Physical anchoring effects. *Journal of Behavioral Decision Making, 19*(4), 393-406.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361-381.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects. *The American Economic Review, 104*(1), 277-290.

McElroy, T., & Dowd, K. (2007). Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and Decision Making, 2*(1), 48-53.

Miron-Shatz, T., Hanoch, Y., Doniger, G. M., Omer, Z. B., & Ozanne, E. M. (2014). Subjective but not objective numeracy influences willingness to pay for BRCA1/2 genetic testing. *Judgment and Decision Making, 9*(2), 152-158.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization, 72*(1), 147-152.

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science, 21*(1), 31-35.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of personality and social psychology, 108*(5), 802.

Schley, D. R., & Peters, E. (2014). Assessing "economic value" Symbolic-number mappings predict risky and riskless valuations. *Psychological Science*, 0956797613515485.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: do people adjust from provided anchors? *Journal of personality and social psychology, 99*(6), 917.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology, 94*(4), 672.

Thomas, K. E., & Handley, S. J. (2008). Anchoring in time estimation. *Acta Psychol (Amst), 127*(1), 24-29.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science, 185*(4157), 1124-1131.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making, 26*(2), 198-212.

Welsh, M. B., Delfabbro, P. H., Burns, N. R., & Begg, S. H. (2014). Individual differences in anchoring: Traits and experience. *Learning and Individual Differences, 29*, 131-140.

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*(4), 387.

Yoon, S., Fong, N. M., & Dimoka, A. (2013). The Robustness of Anchoring Effects on Market Good Valuations. *Available at SSRN 2352692*.

A perspective on all cognition? A study of everyday environments from the perspective of distributed cognition.

Nils Dahlbäck (nils.dahlback@liu.se)
Mattias Kristiansson (mattias.kristiansson@liu.se)
Department of Computer and Information Science, Linköping University
SE-58183 Linköping, Sweden

Abstract

Distributed cognition is a perspective that primarily has been applied to complex socio-technical systems such as flight decks of commercial airliners, or operating rooms where professionals perform cognitive tasks in environments specifically designed for this. For some scholars distributed cognition is exactly this kind of specialized cognitive system. On the other hand it has been claimed by some workers in the field that distributed cognition is not a kind of cognition but a perspective on all cognition. We have therefore studied an environment very different from the systems previously studied, namely single people's homes. We find that there are many similarities between the home and the specialized socio-technical environments. To us this suggests that the specially designed complex environments can be seen as specialized cases of the general principles of distributed cognition which are not reflections of "particular work practices" but of general features of human cognition.

Keywords: everyday cognition; distributed cognition; memory practices.

Introduction

Situated and distributed cognition is not one but many closely related views on or approaches to cognition, which all have in common that cognition is not viewed solely a process residing in the head of the agent, but instead cognitive processes exist in, or at least are influenced by, the agents physical and social environment (Hollan, Hutchins, Kirsh, 2000). There seems however not to be any consensus between workers in the field on how these closely related but different approaches relate to each other. For Robbins and Aydede (2009, p 3) "situated cognition is the genus, and embodied, enactive, embedded, and distributed cognition and their ilk are species", though they note that this usage is not standard, and it is not difficult to find competing views. For instance, Zhang & Norman (1994) claim that distributed cognition has three key components: (1) Embodiment of information that is embedded in representations of interaction, (2) Coordination of enaction among embodied agents, and (3) Ecological contributions to a cognitive ecosystem, which suggests that in their view at least embodied cognition is a sub-aspect of distributed cognition. And Sutton (e.g. 2006) takes a still wider perspective on distributed cognition, which includes embodied and situated aspects among also others.

When it comes to empirical research, almost all the studies on distributed cognition that we are familiar with are detailed studies of complex socio-technical systems, such as flight decks of commercial airliners (Hutchins, 1995a, Hutchins & Klausen, 1996; Dekker, Nyce, & Myers, 2013, navigation teams on the bridges of large vessels (Hutchins, 1995b, Lützhöft, 2004, Lützhöft & Dekker, 2002) operating theatres (Hazlehurst, Gorman, & McMullen 2008; Hazlehurst, McMullen, & Gorman 2007). This is by some scholars taken as a defining feature of the field, as when Rogers and Ellis (1994), to take one example, writes that distributed cognition is an approach which takes as the fundamental unit of analysis "a collection of individuals and artefacts and their relations to each other in a particular work practice". And this characterization seems also to fit studies of very different settings and in very different historical periods, such as Tribble's work on actors' memory for their role and lines in early modern theatre companies, such as Shakespeare's plays when they were first performed (Tribble, 2005; Tribble & Sutton, 2011)

There are a number of features that these environments have in common. First, they consist of teams of many persons with specialized tasks and who have specialized training in performing these tasks. Second, they take place in or rather also consist of a specialized technical environment, which is specially designed to support the tasks being performed by the operators as individuals and the system as a whole.

In contrast to the views on distributed cognition as a kind of cognition particular to specific environments mentioned above, Hutchins claims that distributed cognition is not a kind of cognition but a perspective on all cognition (2013), despite the fact that almost all empirical work within the framework of distributed cognition has been on professional teams working to solve tasks which require the coordination of many agents and their tools to succeed.

One feature that characterizes this approach to distributed cognition is that it keeps the 'cognitivist' view of cognition as computation. "I do believe that the computation observed in the activity of the larger system can be described in the way cognition has been traditionally described – that is, as computation realized through the creation, transformation, and propagation of representational states (Hutchins, 1995b, p.49). This definition is further specified: "…the actual implementation of many interesting computations is achieved by other than symbolic means. For our purposes, 'computation' will be taken, in a broad sense, to refer to the propagation of representational state across representational media. This definition encompasses what we think of as prototypical computations (such as arithmetic operations), as well as a range of other phenomena which I contend are

fundamentally computational but which are not covered by a narrow view of computation" (Hutchins, 1995, p.118) Computation is here thus seen as something wider than formal computation in a strict sense. Our interpretation is that this concerns representations that have some kind of combinatorial syntax, and that the transformations of the representations are made partly based on the agents' understanding of also the semantics of the representations.

Another feature is that cognitive systems can exist on many levels which makes it similar to Cognitive Systems Engineering (Woods and Hollnagel 1983, 2005). Systems consisting of one or more agents (persons) and their physical environment including specially designed cognitive tools can have cognitive properties in their own right which are not the same as the cognitive properties of e.g. the agents in the system. One clear example of this is in Hutchins (1995a) analysis of the flight deck of a commercial airliner, where he states that "To call speed bugs a "memory aide" for the pilots is to mistake the cognitive properties of the reorganized functional system for the cognitive properties of one of its human components. Speed bugs do not help pilots remember speeds, rather they are part of the process by which the cockpit system remembers speeds" (Hutchins, 1995a, p.283).

Another feature of this version of distributed cognition is its emphasis on the close connection between cognition and culture. "I am proposing an integrated view of human cognition in which a major component of culture is a cognitive process (…) and cognition is a cultural process" (Hutchins, 1995b, p 354). A corollary of this is that the historical and cultural development of cognitive systems and functions is important for understanding its current workings, which for instance forms an important part of the analysis of the cockpit as a cognitive system.

But, as mentioned above, most if not all empirical work on distributed cognition has been in complex socio-technical environments, and it is clear that this perspective has helped us understand hitherto unobserved aspects of these environments. The question that we have sought an answer to is then: what can be seen from this perspective when studying agents and environments very different from those previously studied. To do this, we have conducted a cognitive ethnography in an environment possibly maximally distant from the ones mentioned above, namely home environments and single individuals. Previous work in everyday environments by e.g. Kirsh (1995) has primarily focused on theoretical aspects of distributed cognition in such environments with examples illustrating the theoretical points made. There are also empirical studies from home environments which have focused on the management of particular tasks such as medication management (see Palen & Aaløkke, 2006). But to our knowledge no previous study has been conducted and presented of homes as distributed cognitive systems in their own right.

On first appearances, it would seem that an apartment with one person living there, is very different from the complex specialized socio-technical environment previously studied, and that therefore many of the features mentioned above will not be present there. The aspects in focus in this study are that an apartment is in a sense more of a multi-purpose device than an environment for one or a few cognitive tasks, and it is not designed to support cognitive tasks per se. Further, there is no professional training in solving the cognitive tasks for the person living there. And finally, while there certainly is a long historical tradition behind the design of a home and how it is furnished today, this historical development has as far as we can tell, not been driven primarily by a concern for successful performance in cognitive tasks.

So what can we actually see when viewing this kind of environment from the perspective of distributed cognition**?** We are of course not claiming that one perspective is in any absolute sense better than another, but as with any changes in perspective, some aspects become more visible and some become less visible. There are two closely related questions we address here: First, what can be seen in the home as a cognitive system when viewed from the perspective of distributed cognition. Second, which are the differences between the previously studied socio-technical environments and the home.

The Cognitive Ethnography

Eight homes have been studied during a period of two years. The total number of hours in the field and to a lesser extent on the phone with the informants equals roughly 70 hours distributed across 48 occasions of observations, video recordings, and interviews. The time for real-life meetings varied from one and a half hour to four hours. Telephone chats were no longer than 10 minutes.

In-home interviews, telephone chats, and photographs have been used to study the physical settings more broadly. Some activities have been studied more closely. These include leaving home, grocery shopping, cooking and keeping track of the near future. The data collection has been explorative and has therefore not been absolutely balanced across the homes. Some homes have been studied in more detail and across more occasions than others. All homes except from one are single individual households, and all residents have been studied doing individual activities. Despite this, as in any ethnography, the data collection is conducted in a social context.

Two of the homes have been studied specifically in relationship to when residents are about to leave home. For these homes video-recordings have been the primary source of data. The participants wore a head mounted camera (GoPro Hero 2) for about one hour before leaving their home, but with no researcher or other person present during this period. In total the video analysis is based on three hours of video distributed across six occasions of leaving home (Kristiansson, Wiik & Prytz, 2014).

The residents' age span from 70 to 88, and they all manage everyday life by themselves. All except two live in apartments in central areas of a medium-sized city in Sweden. Two live in the outskirts of the same municipality.

Five live in one to two-room apartments. Two in large four room apartments and one in a house. Below the participants are anonymized and referred to by an alias.

The analytical work can be described as a combination of bricolage (Kvale & Brinkmann, 2013) and as following a funnel-approach (Agar, 2008). Bricolage is not a completely systematic approach. Instead analytical tools are used freely across the material to note patterns and themes in the material. The funnel-approach is Agar's name for the inductive approach of over time focusing on indicative aspects of the material. This consequentially means that the analysis focus on some aspects. For this article the funnel-approach has been used to zoom in on indicative aspects of the material in relationship to principles and characteristics of distributed cognition. Distributed cognition has therefore been used as the primary analytical tool. For more details on the ethnographic method see Kristiansson (2016).

Notes from the field study

In this section we will give some examples from the field work, which in the following section will be analyzed with respect to the features of distributed cognition previously presented. In this paper we concentrate on a subset of potential aspects previously used in research on distributed cognition: functional spaces, cognitive tools, routines for cognitive tasks, and the cultural historical development. In Kristiansson (2016) additional aspects are discussed.

Functional spaces supporting cognitive tasks

All participants have spaces for which they have more or less deliberately assigned functionality. All have for instance a spot for their home keys, but the particular solutions differ. Felicia has a metal cup on a bench in the hallway where she drops the keys every time coming home, while Greta puts the keys into a pocket of the jacket when coming home, and Charles always puts the keys on the high bench next to the apartment door. Despite the fact that they all have routines for this, these routines are not adhered to as strict as in professional environments. The routines can perhaps in some cases be seen as ideals for how they should do. For instance, on one occasion the observer points out that the keys are not on the bench where he claims they always are, he replies with a joking comment meaning that a non-adherence to a rule or ideal was revealed ("Nu kom du på mig").

Both the hallway and the kitchen commonly have designated areas that serve cognitive tasks. In some cases an area such as a kitchen table is used both for placing things to be remembered to bring when leaving the home and having meals. One participant, Felicia, has instead of using the kitchen concentrated written information and reminder notes regarding future events to a room, "the office", in the back of her large four-room apartment. In this office she has notes on several spots but mostly on a notice-board. The particular solutions across the studied cases differ but they all either have specific spaces used for memory purposes, or transform multi-purpose spaces when performing specific activities.

Cognitive tools

In all environments we find standard tools for remembering future tasks, such as calendars and post-it notes, but also here we see a large variability in how these are used.

In Felicia's office she has notes on several spots but mostly on a notice-board. On a couple of occasions when the observer does walking interviews in her apartment she is cued by notes on and around the board. For instance at one occasion there is an information sheet on a table below the board. When passing she stops and says that she has forgotten to tell about this event, something that she had planned to do. Putting written information and reminders on open places is one way she uses external information to remind herself. Another way appears at a sit-down discussion with Felicia about her recent activities. While Felicia browses through her calendar she suddenly finds a sticky note far in the back. She reads it and quickly says that she has already finished this intention. Putting information in less prominent places in this way is another way Felicia aims to remind herself. Also, it seems that when intentions become really important or urgent for Felicia objects or notes related to the target activity are placed in the hallway area which is normally stripped from explicit information.

Therefore, despite that Felicia uses written information a lot and to some extent concentrates it in one room her practices of using it is not confined to one kind. This can be contrasted with the case of Beatrice's home.

Beatrice is the participant that seems to use the largest quantity of cognitive tools that hold symbolic informational content. These uses of cognitive tools are the type of processes that the most resembles distributed cognition that has been observed in professional settings. Except from the card-index system that will be described below Beatrice has a pocket calendar, a wall calendar, a list of people that have invited her which she should invite back, a diary, a catalogue over read books, a to-do paper with five categories, a work-in-progress shopping list and more. For specifics of a few of them see below.

The pocket calendar presents a new week on each page turn, and the calendar is located next to the phone in the kitchen. This is a deliberate strategy because this is where most entrances are made. The calendar is almost full every day roughly two weeks onward. The calendar has two bookmarks. One silk-ribbon to mark where she last time ended moving information to her diary and one red card to signify the present.

She has a well-organized social life which is managed with a set of cognitive tools. For instance, at the back of the calendar she has a page with names ordered in a list. The order of the list signifies in which order she owes her friends an invitation. A cross next to the name means that they have managed to invite her two times before. Crossed out names means that she has invited them back and that they therefore are even. Next to the list she has a space

where she sketches on constellations over soon-to-come invitations. Currently there are two clusters of names that corresponds to the list of names. Some names have arrows to both constellations. Most often she only keeps track of her own social debts. She also mentions some aspects of some individuals on the list. Some are more tired and cannot invite back in the same pace as others. Some cannot visit at all and are handled with the tool below. One couple only comes once a year and need to be planned accordingly.

Placed next to the phone in the kitchen is a to-do-list that hold six categories: to buy (no groceries since they are managed separately), to visit, to phone, to write, to invite and, to fix.

This paper with categories leads to the fact that there are few reminder notes spread across her home. Also, because every type of future intention has a designated external spot spatially close to the other kinds, the sheet of paper works as a good overview tool. First, this shows just as previous research within work settings, that the specifics of external representations shape coordination of internal processes and external structures. Second, it also shows that the specifics of how reminder notes are used can partly determine their functionality. Altogether, what is apparent from the observations of Felicia and Beatrice is that in terms of note-taking the inter-individual differences between these distributed systems are large.

Routines for cognitive tasks

Cognitive tools or spaces are of course not enough in and of themselves. They need to be put into use, preferably through routine practices which we see many instances of. Moa, for instance, has a number of functional spaces in the kitchen and hallway where important objects are located. Before leaving home objects are moved from one of these spaces to another as a preparation for leaving. These spaces are also iteratively browsed almost every time when she passes them before leaving home, which leads to that important items are being moved, sometimes in many steps, to the bag-to-bring that is always located on the kitchen table. The kitchen table is also the functional space located farthest away from the exit door, which means that the other spaces will be scanned a final time just before leaving the apartment.

Cultural historical development

Many practices come from people's experiences of other people's practices. Consider again Beatrice. One of her cognitive tools is a card-index system over all social dinners she have made during the last 50 years. The index includes dates, invited guests, what she (and her husband) served, and comments. The idea to the card-index system, she says she remembers clearly, comes from the mother of a friend from her youth that used a similar card-index system. "I found it really smart", she says, and also says that she started using the card index shortly after she married her husband, with the incentive to not serve the same thing twice to the same guest.

An extended example

We will in the final part of this section present a longer illustration of the use of the cognitive tools and spaces. This is based on an analysis of a head mounted video recording made without any observer present.

At one occasion Moa five minutes before leaving home for an exercise session searches for something. The object(s) turns out to be two cards, one that looks like a bus card and one that looks like an exercise card, both of which she needs for the current occasion of leaving home. This search is interesting because it highlights functional relationships between residents and the home environment that involve physical resources out of plain sight that in the past have been shaped as a consequence of previous occasions of similar activities, where Moa in a stressful situation needs to use a combination of opportunistic actions and different deliberate mental resources to gain knowledge of the her own cognitive systems.

A number of objects and spaces were involved in the search: (a) the pockets of a small handbag located on a chair in the kitchen, (b) the pockets of a medium sized shoulder bag for exercise located on the kitchen table, (c) pockets on a number of jackets located on a rack in the hallway, (d) the top of a larger kitchen bench, (e) the top of a portion of the kitchen sink, (f) the top of a smaller bench in the hallway, and (g) the top of a larger bench in the hallway. At a quick glance the search started with the shoulder bag, then past the spaces over to the rack, and back to the kitchen past the spaces, then she searches handbag and finally the shoulder bag once again where she finds the cards. The entire search takes about two minutes. The top surfaces were during this short episode quickly visually looked at when passing by.

If we look closer on how the specific search of the shoulder bag and the jackets evolves we can see something that can be interpreted as a trade-off between fast and opportune actions and deliberate guiding decisions. When Moa searches the jackets she does not search all the jackets equally. Six jackets are searched with relatively distinct hand movements while the pockets of one jacket is searched with quicker in-and-out movements. The last two jackets on the racket (as seen from the entrance) she stops and stares at and determines not to search. These are the two thickest jackets on the racket. Moa likely has a more or less rough idea of when the last time she used the cards was. This knowledge conjoined with the knowledge of that it is currently spring makes it unlikely that the cards should be found in jackets intended for winter. There is therefore what can be seen as a trade-off of the opportune actions sequences of searching jackets and the ongoing guiding thoughts of when to stop the sequence of search. Moa uses what seems to be deliberate remembering processes in combination with physical search. For the jackets in the case of Moa this combination of resources appear to be an efficient search heuristic for this occasion. But if we look at the search of the shoulder bag from a similar perspective we instead see something that from an outside perspective is inefficient.

The black shoulder bag consists of six pockets: a large pocket, a large side pocket and four small side pockets. Small side pocket 2 is closed from start. When Moa searches the shoulder bag the first time she has previously, about five to ten minutes ago, interacted with the bag several times for other reasons related to the packing of the bag. Interactions these times have been with the large pocket and the large side pocket. The fact that Moa has interacted with the bag previously we can take into account when we view how she searches the bag. The first time Moa searches the bag she starts by looking into the large side pocket, stretching the opening with her hands. The missing thoroughness here is not strange given she has had her hand into that pocket before. Then she tilts the bag so that the small side pocket 2 is turned towards the head of Moa. (This is where the two cards later will be found.) After searching other pockets with her hands she will again interact with this pocket by quickly touching the pocket. It appears that she takes micro-decisions as she goes along similar to the way she made decisions about what jackets to search.

We have seen that Moa makes decisions as she goes along. We have also seen that decisions can sometimes be detrimental. We can have ideas of why Moa skipped searching some jackets given what we know about external conditions of weather and jackets. It is harder from the video to know why she does not initially searches small side pocket 2 even though she appears to consider it. But something we can see in the video is that the target pocket also contained a pair of glasses. The glasses was not a flat object. When she takes that object out from that pocket she does that quickly and puts it in another bag next to the shoulder bag. Here we can think of two paths of why she initially decided to not search this pocket. The first one is about her previous interaction with the content of the pocket (not captured by the video). This could be the last time she used the cards but it could also be the moment when she put the unidentified object into that pocket. A possibility is that this created an idea of that this space already has an object which therefore would override the pocket as a space for the two cards. Another decision path is about the physical features of cards. If we view the shape of the pocket, small side pocket 2 is the only side pocket that clearly have a content, and that content is for certain not a flat object. This is because it is closed and the ratio of the size of the pocket and the content of the pocket creates a wrap around the non-flat object. This suggests that the reason for Moa skipping searching small side pocket 2 could have nothing or little to do with remembering but instead about interpretative processes of the pocket's features.

Discussion

We asked in the introduction two questions. First, what can be seen in the home as a cognitive system when viewed from the perspective of distributed cognition. Second, which are the differences between the previously studied socio-technical environments and the home.

Our findings suggest that even though the home is not an environment specifically designed for solving one specific (professional) cognitive task or function, when looked at from the perspective of distributed cognition a number of interesting similarities with the more commonly studied complex socio-technical environments are revealed. Parts of the environments are designed or modified to serve one or more cognitive tasks, in many cases retrospective but more commonly prospective memory, and similar to the socio-technical environments previously studied, the use of these is often done through particular routines where the place, tool, and the routine together serve a cognitive function.

It should be noted that we see a large variation in the concrete implementation of these functions in the eight environments we have studied. For instance, one informant (Beatrice) has developed a large number of specialized tools for several cognitive tasks. Since the studies of professional environments in most cases concern only one instance of these, we however cannot say whether this is a difference between professional and non-professional environments.

It must also be noted that there are instances of cognition in home environments such as visits by home healthcare professionals which creates a combination of professional and non-professional environments (see Kristiansson, 2013, Dahlbäck, Kristiansson, Stjernberg, 2013 & Palen & Aaløkke, 2006). Therefore the contrast we make here between the two kinds should not be seen as two discrete cases, but rather two endpoints on a continuum.

There are further reasons for not seeing professional and non-professional environments as fundamentally different kinds. All participants display levels of expertise in how they manage their daily chores. How they for cognitive tasks shape, use and interpret their physical environments is for instance based on their knowledge of their own practices. Of course such mechanisms establishes a cognitive connectivity between the agent and aspects of their environment that equals to some sort of expertise. Further, despite the fact that participants for most daily chores have not received professional training there are cases when participants have experiences (of cooking, financial management, healthcare etc.) from their previous working lives that feeds expertise into their management of daily life. Thus, this lends support to Kirsh's (2009) claim that most people become experts or near experts in dealing with their everyday environments.

Similar to professional environments, we also found examples of a kind of cultural knowledge accumulation of the cognitive elements in the home environments, where the informants tell us how some of their routines and tool designs are learned from or influenced by older generations. What we did not see in this study is any form of socially distributed memory, which of course is a natural consequence of the fact that we have studied environments with only one person. But it is for instance interesting to note that one informant, Beatrice, decided to develop a complex memory tool for remembering social events when her husband has passed away, which can be interpreted as a

transformation of a distributed memory system initially comprising of two persons to a one person system with the same function when one of them is no longer there. So the details of the distributed memory system changes with changing circumstance while keeping the basic functionality.

Despite the differences between the two prototypical cases of professional and non-professional environments we have described, the pattern that emerges is one of a continuum in cognitive functioning between the everyday environment and the complex socio-technical environments previously studied from the perspective of distributed cognition. We have also as yet unpublished studies of e.g. families with children, which, while of course differing in details, show no fundamental differences with what we present here. To us this suggests that the specially designed complex environments can be seen as specialized cases of the general principles of distributed cognition which are not reflections of "particular work practices" but of general features of human cognition.

Acknowledgments

We thank Corinna Kruse, Fredrik Stjernberg and Lars-Christer Hydén for help with the analysis of the field study, and Richard Wiik and Sofia Lindvall for help with the fieldwork and initial analysis. We also thank the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Agar, M. H. (2008). *The Professional Stranger: An informal introduction to ethnography* (Second Edi). Bingley: Emerald Group Publishing.

Dahlbäck, N., Kristiansson, M., & Stjernberg, F. (2013). Distributed Remembering Through Active Structuring of Activities and Environments. *Review of Philosophy and Psychology, 4(1), 153-165*.

Dekker, S. W. A., Nyce, J. M., & Myers, D. J. (2013). The little engine who could not: "rehabilitating" the individual in safety research. *Cognition, Technology and Work, 15*, 277–282.

Hazlehurst, B., Gorman, P. N., & McMullen, C. K. (2008). Distributed cognition: An alternative model of cognition for medical informatics. *International Journal of Medical Informatics, 77(4)*, 226–234.

Hazlehurst, B., McMullen, C. K., & Gorman, P. N. (2007). Distributed cognition in the heart room: how situation awareness arises from coordinated communications during cardiac surgery. *Journal of Biomedical Informatics, 40*(5), 539–51.

Hollan, J., Hutchins, E. & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, 7*(2), 174-196.

Hollnagel, E., & Woods, D. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Boca Raton: CRC Press.

Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: new wine in new bottles. *International Journal of Man-Machine Studies, 18*, 583–600.

Hutchins, E. (1995a). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265–288.

Hutchins, E. (2013). The cultural ecosystem of human cognition. *Philosophical Psychology*, (September), 1–16.

Hutchins, E., & Klausen, T. (1996). Distributed cognition in an airline cockpit. In Y. Engeström & D. Middleton (Eds.), *Cognition and Communication at Work* (pp. 15–34). Cambridge: Cambridge University Press.

Kirsh, D. (1995). The intelligent use of space. *Artifical Intelligence, 73*, 31-68.

Kirsh, D. (2009) Problem solving and situated cognition. In: Robbins, P., & Aydede, M. (Eds.). *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.

Kristiansson, M. (2013). The case of cognitive ecology for cognitive processes in everyday life situations. *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2778–2783). Berlin: Cognitive Science Society.

Kvale, S., & Brinkmann, S. (2013). *Den kvalitativa forskningsintervjun* (Andra uppl). Studentlitteratur AB.

Lützhöft, M. (2004). *"The technology is great when it works": Maritime Technology and Human Integration on the Ship's Bridge*. Doctoral Disseration, Linköping University.

Lützhöft, M. H., & Dekker, S. (2002). On Your Watch: Automation on the Bridge. *Journal of Navigation, 55(1)*, 83–96.

Palen, L. & Aaløkke, S. (2006). Of Pill Boxes and Piano Benches: "Home-made" Methods for Managing Medication. *Proceedings of 20th anniversary conference on Computer Supported Cooperative Work* (pp. 79-88). Banff: ACM.

Robbins, P., & Aydede, M. (Eds.). (2009). *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press.

Rogers, Y., & Ellis, J. (1994). Distributed cognition: An alternative framework for analysing and explaining collaborative working. *Journal of Information Technology, 9(2)*, 119–128.

Sutton, J. (2006). Distributed cognition: Domains and dimensions. *Pragmatics & Cognition*, *14*(2), 235–247.

Tribble, E. B. (2005). Distributing Cognition in the Globe. *Shakespeare Quarterly*, *56*(2), 135–155.

Tribble, E. B., & Sutton, J. (2011). Cognitive Ecology as a Framework for Shakespearean Studies. *Shakespeare Studies*, *39*, 94–104.

Zhang, J., & Norman, D. (1994). Represesentations in Distributed Cognitive Tasks. *Cognitive Science*, *18*, 87–122.

Temporal Structure Modulates ERP Correlates of Visual Sequential Learning

Kimberly M. Ross (Kross25@student.gsu.edu)
Department of Psychology, P.O. Box 5010, Atlanta, GA 30302 USA

Christopher M. Conway (Cconway@gsu.edu)
Department of Psychology, P.O. Box 5010, Atlanta, GA 30302 USA

Abstract

Sequential learning (SL) refers to the ability to learn the temporal and ordinal patterns of one's environment. Whereas research on the learning of ordinal patterns is common, the learning of temporal patterns within sequential events has been far less studied. The current study examines the effects of synchronous and asynchronous temporal patterns on visual sequential learning. We hypothesize that entrainment (i.e. exposure to a regular rhythmic pattern) allows for better processing of the ordinal structure of sequential events. Twenty healthy adult participants (11 females, 18–34 years old) performed two versions of a visual sequential learning paradigm while event-related potentials (ERPs) were recorded. The SL task involved the visual presentation of colored circles, wherein a target circle was embedded that was partially predictable based on preceding predictor stimuli. One version of the task incorporated synchronous temporal presentation of the stimuli whereas the other version involved asynchronous presentation of stimuli using a randomized ISI on every trial. Reaction time data demonstrated that learning occurred in both temporal conditions. On the other hand, the mean ERP amplitudes between 350 and 750ms post-predictor onset in the posterior regions of interest revealed that learning of the statistical contingencies between stimuli was disrupted for the asynchronous temporal condition but intact for the synchronous condition. These neurophysiological data suggest that the brain processes regular and irregular timing events differently, with statistical learning of ordinal visual patterns being improved by a synchronous temporal structure, possibly a result of heightened attention to the stimuli due to entrainment.

Keywords: Sequential learning; statistical learning; temporal processing; entrainment; ERPs

Introduction

An important question in cognitive science is what facilitates learning of structured events under different contexts. When events are not fully random but contain a degree of temporal or ordinal regularity, our brain is able to extract these regularities to facilitate processing using predictive mechanisms, that is, by learning to predict future stimuli in the sequence (Selchenkova et al., 2014). In this manner, we are able to generate expectancies about future events, an ability that that allows us to improve our responses and execution of actions (Rohenkohl et al., 2012).

Sequential learning (SL) is the ability to learn incidentally the ordinal patterns of one's environment (i.e., sequences of items that unfold in time, such as a melody or a gymnastics routine). Surprisingly, whereas SL studies of the ordinal structure of sequences are common, research on SL of temporal patterns is scare. Most SL studies focus on learning sequences of stimuli with identical durations and identical inter-stimuli intervals (e.g. Conway & Christiansen, 2005; Jost et al., 2015). However, SL of sequences containing temporal irregularity, that is with stimuli of varying durations or with different inter-stimuli intervals, is of fundamental importance to human cognition because these types of sequences are frequent in our environment. We often process irregular temporal patterns that help us make decisions and influence our future behavior, including when we perform motor movement coordination, and when we process language or music (Brandon et al., 2012).

To understand the human ability to process statistical sequential information better, one must examine both the ordinal and temporal information available in any given context. The Dynamic Attending Theory (DAT) proposes that events with a regular pattern are processed easier than events with an irregular temporal pattern. This theory assumes that the brain creates internal oscillations, or attending rhythms, that entrain to external rhythms, creating a type of attention to future events, which facilitates learning (Jones & Boltz, 1989). The internal oscillations of the DAT are thus adaptive, allowing for enhanced processing of regular temporal structures. Therefore, the DAT implies that attention to individual items of a sequence is stimulus-driven (Jones et al., 2002). In this respect the temporal structure of a sequence becomes useful or salient to the attender.

SL can be observed behaviorally (response times) as well as indexed neurophysiologically with event-related potentials (ERPs, for a recent review, see Daltrozzo & Conway, 2014). The ERP technique has been used to study SL using variations of the classic auditory oddball sequence task, the serial reaction time task (SRT) and with artificial grammar learning (AGL) paradigms (Brandon et al., 2012; Karabanov & Ullen, 2008; Schmidt-Kassow et al., 2009; Schwartze et al., 2011; Selchenkova et al., 2014; Selchenkova, Jones & Tillman, 2014). The ERP technique has the advantage of a high temporal resolution at the millisecond scale, allowing for the exploration of neural events with precise timing. Thus, ERPs are particularly well suited for examining temporal cognition, including SL of temporal and ordinal regularities.

The Current Study

The current study uses a probabilistic visual serial learning task (Jost et al., 2015) with a manipulation of the temporal

synchronicity as a way to explore the effects of temporal regularity on SL. Whereas the literature on the relationship between SL and temporal processing is scarce, there is some evidence that temporal regularity modulates SL. The main conclusions of this research is that a regular temporal pattern allows for better processing of the sequence, with faster reaction times and a larger peak amplitude and shorter latency onset of a P3-like component (Rohenkohl et al., 2012; Schmidt-Kassow et al., 2009). However, because most of the studies exploring the effect of the temporal regularities on SL have been conducted in the auditory domain, it is important to test this relationship in the visual domain (Rohenkohl et al., 2012). The results of previous SL literature have shown that perceptual modality (i.e., visual vs. auditory patterns) influences SL (Conway & Christiansen, 2005). Thus, it is important to explore whether entrainment to a temporal structure can facilitate SL independently of the modality of the to-be-learned sequence, that is not only in the auditory but also in the visual domain.

We hypothesize that: (1) Reaction time data will show better learning in the synchronous condition compared to the asynchronous condition; (2) ERP data will show that the synchronous sequences yield larger peak amplitudes of the P300 than asynchronous sequences, suggesting a facilitative effect for processing statistical regularities in temporally regular events.

Method

Participants

Twenty adult participants (11 females, 18 right-handed, 18-34 years old, average age = 20.5) without reported language, cognitive, neurological, or psychological deficits and who were native English speakers participated in this experiment. Participants were recruited through Georgia State University's SONA system, receiving course credit for their participation. All participants provided written informed consent, which was approved by the Institutional Review Board of Georgia State University. Participants were asked to fill out a brief demographic questionnaire and the Edinburgh Handedness Inventory (Oldfield, 1971).

Procedure

The SL paradigm, based on Jost et al. (2015), involved the presentation of a sequence of colored circles (brown, blue, grey, pink, orange, red, purple, yellow, green, white) in the center of a computer screen with a black background (Figure 1). Participants were asked to press a button whenever they saw a circle of a specific color (the "target"). Each trial consisted of one to five "filler" circles, followed by one of the three predictor circles (high, low, and zero predictor-target probability, chosen randomly on each trial). Depending on which predictor stimulus was presented, the next stimulus was either the target circle or the filler circle. The target circle followed the "high predictor" on 80% of the trials, with a filler circle following 20% of the time. The target circle followed the "low predictor" 20% of the time, with a filler circle following 80% of the time. The target circle never followed the "zero predictor" circle. After the target or final filler was presented at the end of the trial, the sequence repeated itself by starting off again with one to five filler circles and then the randomly chosen predictor stimulus. The color assigned to the target, predictors, and filler circles was randomly chosen for each participant at the beginning of the task and the selection of colors for each stimulus type remained constant throughout the task for each participant. In the synchronous temporal condition, the stimuli appeared on the screen for 500ms with a 1000ms (black screen) ISI. In the asynchronous temporal condition, stimuli appeared on the screen for 500ms, with a randomized ISI, ranging between 600ms and 1400ms so that on average across trials the ISI remained the same as in the synchronous condition (1000ms).

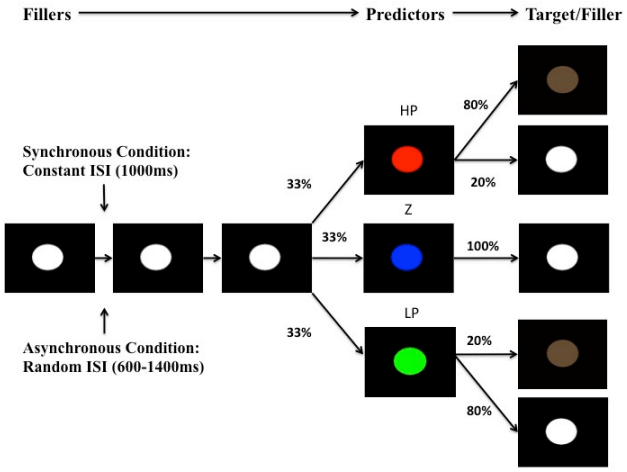


Figure 1: Visual SL task layout [high probability, HP; low probability, LP; zero probability, Z]. In this example, three filler circles precede the predictor stimuli, but this number could range from one to five. After the appearance of either a target or filler at the end of the trial, a new sequence begins. In this example, the target stimuli are green, but in reality the colors of the standard, predictors and target stimuli were randomly assigned for each participant.

The experimental conditions were separated into two separate tasks, synchronous and asynchronous. The assignment of set colors as well as order of each task were counterbalanced across participants, so that each participant received a different set of colors for each task. Each task lasted approximately 25 minutes, and included the presentation of 180 trials through 6 blocks of 30 trials each. Compared to the Jost et al. (2015) paradigm that included the presentation of 150 trails, the overall number of trials was increased to 180 trials to increase the signal to noise ratio, allowing us to better compare ERP effects between the first and second half of the SL task. After participants completed both sequential learning tasks, they completed a pattern consciousness inventory, a measure that tested the

overall level of consciousness of the probabilistic structure of the sequence.

Recording Technique

The electroencephalograph (EEG) was measured from 256 scalp sensors using an Electrical Geodesic Inc. (EGI) EEG net (Figure 2). Net Station Version 4.3.1 tools were used to process the EEGs. Active electrode impedances were kept below 50 kΩ. Recordings were made with a 0.1 to 30 Hz bandpass filter and digitized at 250 Hz. The EEG was segmented into epochs -200 to +1000ms with respect to the predictor onset. An artifact detection operation removed trials containing noise from eye blinks and other movements. Separate ERPs were computed for each participant, experimental condition, electrode and block. All experimental sessions were conducted in a 132 square foot double-walled, sound-deadened acoustic chamber.

Statistical Analysis

Statistical calculations were performed on the individual mean amplitude ERPs within the time-window of interest (350-750ms post predictor onset), estimated from previous research and visual inspection (Jost et al., 2015)), using Net Station Version 4.3.1. To analyze the effect of cortical topography, nine regions of interest were defined (ROIs, Figure 2): left (LAn), middle (FRz), and right anterior (RAn); left (LCn), middle (CNz), and right central (RCn); and left (LPo), middle (POz) and right posterior (RPo) regions. Based on previous research, we expected the ERP effects of learning to be focused in the posterior central (POz) region (Jost, et al., 2015). Visual inspection of the grand averages confirmed an SL effect in both the left posterior (LPo) and central posterior (POz) regions, and so all analyses were conducted on these two combined regions (posterior medial-left). Repeated-measure ANOVAs on the individual mean amplitudes were conducted with the following within-participant factors: Predictor ("high predictor" or HP, "low predictor" or LP, and "zero predictor" or Z), temporal regularity (synchronous, asynchronous) and block (first three blocks vs. last three blocks). One participant was excluded from the ERP statistical calculations due to a high percentage (over 60%) of trials containing artifacts (Synchronous, 94%; Asynchronous, 82%).

Response times to target stimuli were analyzed with repeated-measure ANOVAs with predictor, temporal regularity, and block as within-participant factors. One participant was also excluded from response time calculations, due to a computer error.

All statistical analyses were conducted with SPSS (PAWS Statistics 18 – Release 18.0.3 September 9, 2010). All reported p-values were adjusted with the Greenhouse–Geisser correction for non-sphericity, when appropriate. Partial eta-squared is reported as a measure of effect size for all ANOVAs (Cohen, 1988; Olejnik & Algina, 2003). Reported p-values of the posthoc tests were Šidák corrected.

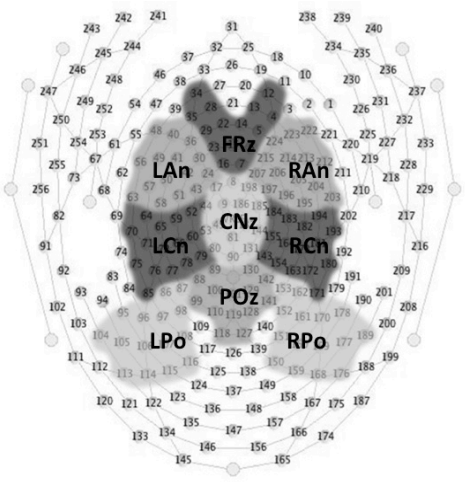


Figure 2: 256 sensors EEG net with the highlighted nine regions of interest.

Results

Table 1 displays the mean reaction time data for both the synchronous and asynchronous tasks, separated by the first half and second half of each in order to observe effects of learning that might be present following a certain amount of exposure to the patterns. A 2x2x2 repeated measures ANOVA revealed a significant main effect of Predictability [$F(1,18) = 11.79$; $p = .003$, $\eta_p^2 = .40$] and a main effect of Block [$F(1,18) = 6.63$; $p = .02$, $\eta_p^2 = .27$], but no significant interactions. These data indicate that the HP condition was responded to significantly faster than the LP condition across both halves of both tasks. The main effect of block also suggests that participants improved on their performance of the task in the second part of the tasks, regardless of predictor type and task synchronicity. Overall, these results suggest that participants showed facilitation with responding to targets when the HP stimulus was present, indicating learning of the ordinal structure in both the synchronous and asynchronous versions of the task. Note that numerically the synchronous condition displayed improved RTs for the HP condition from the 1st half to 2nd half of the task (379.9 vs. 357.1 ms); although the asynchronous condition also showed a similar facilitation, the effect was smaller (387.2 vs. 372.0 ms). However, this difference in effects was non-significant, likely due to lack of statistical power [$F(1,18) = .112$; $p = .74$, $\eta_p^2 = .006$].

Table 1: Mean (SD) reaction time scores by timing condition, predictor and block.

	Synchronous		Asynchronous	
	1st Half	2nd Half	1st Half	2nd Half
HP	379.9 (59.4)	357.1 (66.0)	387.2 (43.1)	372.0 (63.0)
LP	385.7 (45.9)	379.9 (50.7)	406.7 (40.9)	389.3 (46.8)

The ERP data, however, presents a more nuanced picture. Figure 3 displays the grand averaged ERPs for each task across all participants, time-locked to the three predictors (HP, LP, & Z) at the posterior medial-left regions of interest used for topographic analyses during both the first and second half of the task. Visual inspection suggests a larger positivity between about 350ms and 750ms for the HP predictor compared to the LP and Z predictors in the second half of the task for the synchronous but not asynchronous conditions.

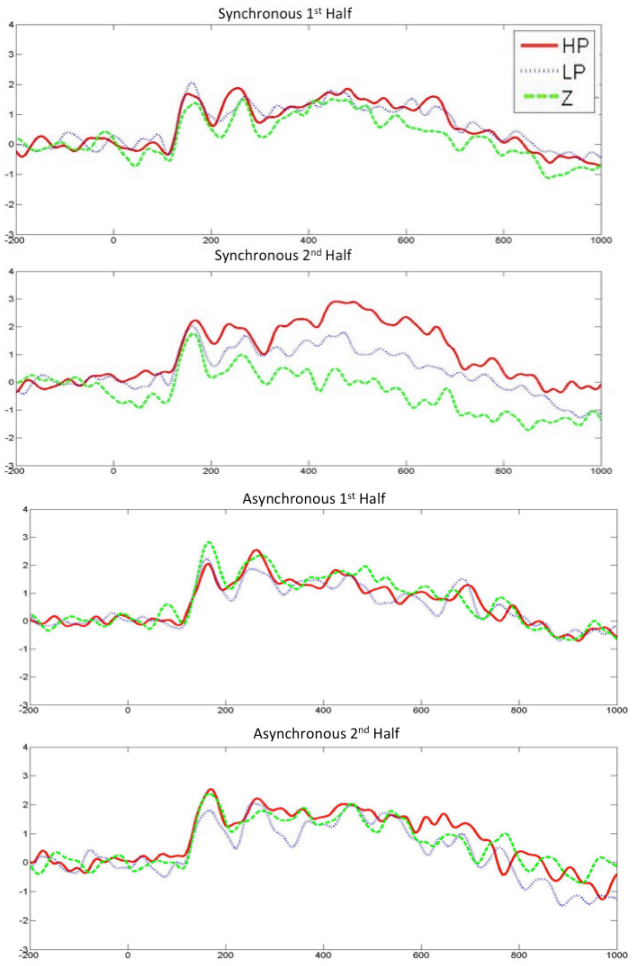


Figure 3: Grand average ERPs observed in the posterior medial and left regions of interest in response to the high probability condition (HP, red line), low probability condition (LP, blue line), and zero probability condition (Z, green line) (vertical axis: electrical potential in μV, positivity upward; horizontal axis: time in milliseconds) in the first and last three blocks of each task. The synchronous task is shown in the upper panels and the asynchronous task in the lower panels.

Figure 4 shows the means for the posterior medial-left region for each of the three predictors in the first three blocks of the task versus the last three blocks of the task in both timing conditions. From visual inspection, it is very clear that the timing condition appears to be affecting the ERPs elicited by each predictor type in different ways. Specifically, whereas in the first half of both tasks, the ERP effects do not appear to differ, they do differ in the second half for the synchronous but not the asynchronous task, presumably reflecting differences in participants' learning of the varying predictor-target probabilities.

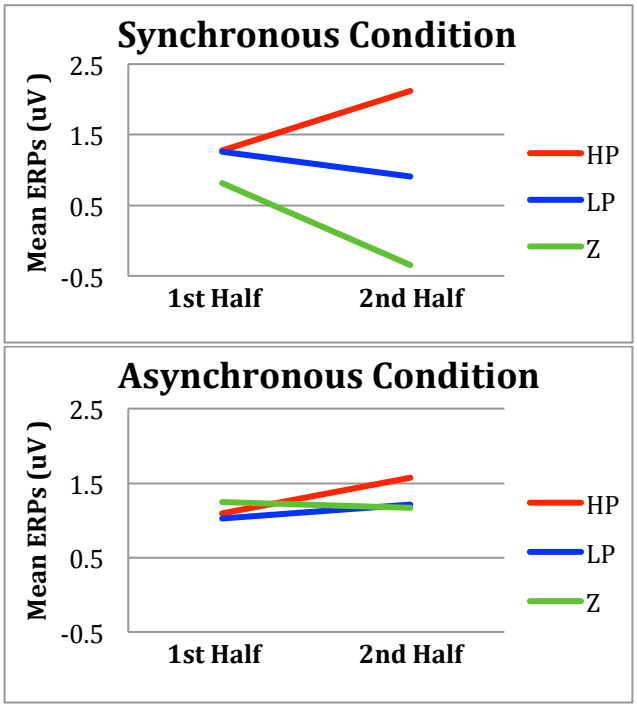


Figure 4: Line graph depicting the means in microvolts (μV) for the posterior medial-left region for each of the three predictors in the first half versus the second half of the task 350-750ms post-predictor onset.

A 2x2x3 repeated measures ANOVA confirmed that there was an interaction between Timing and Predictability 350ms to 750ms poststimulus onset [$F(2,36) = 6.01$; $p = .006$, $\eta_p^2 = .25$] indicating a significant difference between HP and Z in the synchronous condition but not the asynchronous condition ($p = .009$). There was also a significant Block and Predictability interaction [$F(2,36) = 4.13$; $p = .040$, $\eta_p^2 = .19$] 350ms to 750ms poststimulus onset, indicating that the difference between HP and LP ($p = .03$) and HP and Z ($p = .02$) was larger in the second half of the experiment, regardless of timing condition.

While there was no significant three-way interaction between Predictability, Timing and Block, two 2x3 repeated measures ANOVAs indicate an effect only in the synchronous condition. We found a significant Block and Predictability interaction in the synchronous condition [$F(2,36) = 5.77$; $p = <.001$, $\eta_p^2 = .24$], indicating differences between means from the first to the second half of the task. Posthoc tests revealed significant differences between HP and LP from the first half to the second half ($p = .013$) as well as significant differences between HP and Z from the first half to the second half ($p = .01$). A repeated measures

ANOVA found no significant interaction between Block and Predictability in the asynchronous condition.

The consciousness inventory produced an average rating per participant, for each temporal condition. A paired-sample t-test revealed no significant differences between timing groups ($t(18)= -.304$, $p =.77$). However, the consciousness scores were significantly correlated with the mean ERP amplitudes for each predictor condition in the second half of the synchronous task: HP ($r=.608$, $p=.006$), LP ($r=.614$, $p=.005$) and Z ($r=.603$, $p =.006$) from 350-750ms post-predictor. This suggests that for the synchronous condition, in the second half of the task, there was a positive relationship between one's level of consciousness of the probabilistic patterns and the ERP amplitudes.

Discussion

This study explored the effects of temporal regularity on the neural correlates of visual SL using neurophysiological evidence (ERPs). The main findings of this research are that (1) reaction time data showed that the target was responded to faster following the presentation of the high predictor compared to the low predictor, regardless of temporal condition, indicating learning for both conditions; (2) under the synchronous temporal conditions only, the ERPs indicated a significant effect of predictor type in the last 3 blocks of the task, with greater P300-like amplitudes for the HP condition.

The ERP results of the synchronous condition mirrors those of Jost et al. (2015), who also observed a P300-like ERP component for the HP predictor, reflecting the learning of the probabilistic contingencies between stimuli. The fact that this P300 effect was not seen in either the first three or the last three blocks of the asynchronous task highlights the fact that processing was enhanced during trials that had highly regular rhythms. This shows that variability in timing may influence the P3b, which is typically seen 300-500ms over central and parietal electrode sites (Schmidt-Kassow, Schubotz & Kotz; 2009). According to the predictions made by the DAT, events with highly regular temporal rhythms produce entrainment of oscillatory waves, so that perception and encoding are enhanced because stimuli are being presented during the highest point in the wave of attention (Jones & Boltz, 1989). Our findings fit well within the expectation of the DAT that temporal regularity provides an opportunity to direct attention to salient information, in this case, the onset of the stimuli being presented, which led to improved encoding of the statistical regularities.

Whereas the classic oddball paradigm is used to explore the effect of deviant stimuli in a stream of input, the modified oddball paradigm that was applied in the present study (based on Jost et al., 2015), that includes predictor-target statistical contingencies, allows for the exploration of the extraction of sequential probabilities out of a serial input stream. While most of the research conducted on SL and temporal processing has been conducted in the auditory domain, this study shows that SL of visual stimuli is also sensitive to temporal regularities. The P300 may reflect expectations about when the target stimulus occurs, and therefore stimuli that occur at expected time points are processed more efficiently. Since we do not always experience events in a regular temporal fashion, understanding how we process events structured with a varying temporal regularity has important implications for human cognition, especially for incidental and implicit learning.

Interestingly, reaction time data showed learning effects in both timing conditions in the last 3 blocks of the task, which seems to stand in contrast to the ERP findings. One interpretation of this discrepancy is that reaction times could represent the implicit learning of the patterns, while the ERP effects index attention-dependent processes that were affected by entrainment. The consciousness scores revealed a significant positive correlation with ERP means in the synchronous condition in the last half of the experiment, suggesting that as one's awareness of the sequence increased, so did their neurophysiological responses. The P300 is known to be affected by attentional manipulations (Polich, 2007), so taken together, these findings suggest that temporal regularity results in increased attentional processing of the patterns while leaving implicit learning more or less unaffected (for a similar argument that sequential learning relies upon both implicit and explicit learning processes, see Batterink et al., 2015) On the other hand, the means of the reaction times go in the direction that would be expected if the synchronous condition led to improved learning; thus, there simply may not be enough power to detect these effects behaviorally.

In addition to using larger sample sizes, future studies might explore different ways of varying the temporal structure of input sequences. For instance, in line with previous research (Brandon et al., 2012; Selchenkova et al., 2014), a metrical framework might be adapted and tested using this predictor-target paradigm. Selchenkova et al. (2014) manipulated the temporal structure of sequences by using both metrical and isochronous structures in an artificial grammar-learning paradigm. They found that the highly metrical condition showed a larger P300 component in the exposure phase and an earlier N2 component in the test phase, in comparison to the isochronous condition. Studies like these suggest that a complex interplay between metricality and temporal regularity can have a dramatic effect on SL, and thus it may be advantageous to further explore these dimensions.

Finally, the current research on temporal processing and SL is expected to have implications for our understanding of certain pathologies. Language, cognitive and motor impairments, such as Specific Language Impairment (SLI), attention-deficit hyperactivity disorder (ADHD), and dyslexia, as well as Parkinson's Disease and Schizophrenia, appear to be associated with temporal and entrainment deficits that in turn could lead to difficulties in sequence processing (Basu et al., 2010; Hsu & Bishop, 2014; Davalos et al., 2011; Harrington et al., 2011; Noreika et al., 2013).

For example, SLI is thought to stem from an impairment of the procedural memory system, which in turn may affect both sequential and temporal processing (Hsu & Bishop, 2014). Future research ought to explore entrainment and SL in typical and atypical participants in order to better characterize the nature of the deficits that these individuals are experiencing. One possibility is that SL is impaired in these pathological populations because of a lessened ability to dynamically attend to stimuli, leading to inefficient processing of both auditory and visual stimuli. This research approach is expected to advance our comprehension and assessment of several types of cognitive impairments affecting language, attention, motor coordination, and more generally a wide range of cognitive systems. By exploring SL in healthy adults and its relationship with temporal processing, this research could pave the pathway towards a better understanding of the cognitive impairments of these pathological populations.

Acknowledgments

We thank the NIH (R01DC012037) for their financial support and J. Daltrozzo for his help in designing the study.

References

Basu, M., Krishnan, A., & Weber-Fox, C. (2010). Brainstem correlates of temporal auditory processing in children with specific language impairment. Developmental Science, 13, 77-91.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language, 83*, 62-78.

Brandon, M., Terry, J., Stevens, C., & Tillmann, B. (2012). Incidental Learning of Temporal Structures Conforming to a Metrical Framework. *Frontiers in Psychology, 3*, 1-10.

Cohen J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale (NJ): Erlbaum Associates.

Conway, C., & Christiansen, M. (2005). Modality-Constrained Statistical Learning Of Tactile, Visual, And Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 24-39.

Daltrozzo, J., & Conway, C. (2014). Neurocognitive mechanisms of statistical-sequential learning: What do event-related potentials tell us? *Frontiers in Human Neuroscience, 8*, 1-17.

Davalos, D., Rojas, D., & Tregellas, J. (2011). Temporal Processing In Schizophrenia: Effects Of Task-Difficulty On Behavioral Discrimination And Neuronal Responses. Schizophrenia Research, 127, 97-98.

Harrington, D., Castillo, G., Greenberg, P., Song, D., Lessig, S., Lee, R., & Rao, S. (2011). Neurobehavioral Mechanisms of Temporal Processing Deficits in Parkinson's Disease. PLoS ONE, 6(2), 1-14.

Hsu, H., & Bishop, D. (2014). Sequence-specific procedural learning deficits in children with specific language impairment. *Developmental Science*, 17, 352-365.

Jones, M.R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. Psychological Review, 83, 323-335.

Jones, M., & Boltz, M. (1989). Dynamic attending and responses to time. Psychological Review, 96(3), 459-491.

Jones, M., Moynihan, H., Mackenzie, N., & Puente, J. (2002). Temporal Aspects of Stimulus-Driven Attending in Dynamic Arrays. *Psychological Science*, 13, 313-319.

Jost, E., Conway, C., Purdy, J., Walk, A., & Hendricks, M. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. Brain Research, 1597, 95-107.

Karabanov, A., & Ullen, F. (2008). Implicit and Explicit Learning of Temporal Sequences Studied With the Process Dissociation Procedure. Journal of Neurophysiology, 100, 733-739.

Noreika, V., Falter, C., & Rubia, K. (2013). Timing deficits in attention-deficit/hyperactivity disorder (ADHD): Evidence from neurocognitive and neuroimaging studies. Neuropsychologia, 51, 235-266.

Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsycholgia 9(1), 97-113.

Olejnik S, Algina J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods, 8:434–447.

Polich, J. (2007). Updating P300: An Integrative Theory Of P3a And P3b. *Clinical Neurophysiology, 118*, 2128-2148.

Rohenkohl, G., Cravo, A., Wyart, V., & Nobre, A. (2012). Temporal Expectation Improves the Quality of Sensory Information. Journal of Neuroscience, 32, 8424-8428.

Schmidt-Kassow, M., Schubotz, R., & Kotz, S. (2009). Attention and entrainment: P3b varies as a function of temporal predictability. *Neuroreport, 20*, 31-36.

Schwartze, M., & Kotz, S. (2015). The Timing of Regular Sequences: Production, Perception, and Covariation. Journal of Cognitive Neuroscience, X, 1-11. doi:10.1162/jocn_a_00805

Schwartze, M., Rothermich, K., Schmidt-Kassow, M., & Kotz, S. (2011). Temporal regularity effects on pre-attentive and attentive processing of deviance. Biological Psychology, 87, 146-151.

Selchenkova, T., Francois, C., Schon, D., Corneyllie, A., & Perrin, F. (2014). Metrical Presentation Boosts Implicit Learning of Artificial Grammar. PLOS ONE, 9(11), 1-9.

Selchenkova, T., Jones, M., & Tillmann, B. (2014). The influence of temporal regularities on the implicit learning of pitch structures. The Quarterly Journal of Experimental Psychology,67(12), 2360-2379.

Knowledge and use of price distributions by populations and individuals

Timothy F. Lew (tflew@ucsd.edu) and Edward Vul (evul@ucsd.edu)
University of California, San Diego
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA 92093 USA

Abstract

How much do individuals, compared to the population, know about the distribution of values in the world? Participants reported the prices of consumer goods such as watches and belts and we compared how accurately individuals vs. the overall population knew the mean and dispersion of prices. Although individuals and the population both knew objects' average prices and relative standard deviations, the population was more sensitive to the absolute standard deviation of prices. In a second experiment, we examined whether individuals' impoverished distribution knowledge impairs their ability to interpret advertisements. Consistent with people using Bayesian inference, the higher an object's actual price dispersion, the more participants relied on advertisements; however, this effect is considerably smaller than a simple proportional offset, suggesting again that individuals underestimate dispersion. Thus, despite having a sense of the distribution of real world quantities, individuals tend to know only a fraction of the world distribution.

Keywords: Probabilistic inference; decision-making; behavioral economics; prior knowledge

Introduction

How much does a television cost? How about a television endorsed by Quentin Tarantino? Chances are you can accurately guess the prices of these and many other everyday objects. People appear to use their knowledge of how values are distributed (e.g. the mean and variance of a value) to infer the prices of objects or other real-world values like people's life spans and cake baking times (Griffiths & Tenenbaum, 2006). This behavior can also arise from people only knowing a few examples of values (Mozer, Pashler & Homaei, 2008). For example, rather than know the average time it takes to bake a cake and how much baking times vary, each person may only know one or two cake baking times. Aggregating over multiple people in a population each with a few sample values can give the appearance of individuals representing complex distributions.

However, people only reporting a few car costs or baking times may actually reflect them using knowledge of distributions. Over multiple trials a single individual's responses resemble learned distributions (Lewandowsky, et al., 2009), suggesting that people do represent the distribution of real-world values. In fact, people appearing to rely on only a few samples may actually reflect people efficiently approximating complex distributions by taking a few samples (Vul, et al., 2014). Here, we examined how well individual people vs. populations know the distribution of real world values—the prices of everyday objects—and how they use that prior distributional knowledge to update price estimates given new information (i.e. advertisements).

Although individuals appear to represent distributions of values, different sources of errors and biases may impede their ability to estimate the true mean and variance of values as accurately as populations. First, individuals are likely to experience an imperfectly representative fraction of the world, resulting in idiosyncratic biases in their expectations. Anchoring effects may bias individuals' judgments, such that hearing a high price causes people to overestimate the prices of objects (Tversky & Kahneman, 1974). People might estimate the price of an object by comparing it to other objects they have recently encountered (Ungemach, et al., 2011; Vlaev, et al., 2011). Additionally, rather than independently sample possible prices of objects, people may be biased by their memories of previous responses (Vul & Pashler, 2008; Hourihan & Benjamin, 2010). When aggregating the judgments of a population of independent individuals instead, many of these biases will wash out allowing the population to represent the distribution more accurately.

In turn, the limitations of people's distributional knowledge may determine how they infer the values of new objects and how they integrate new sources of information. For instance, how should a person infer the price of an object when it is presented with an advertisement? People may use Bayesian inference to determine how to use ads, similar to how listeners infer the meaning of utterances in the domain of pragmatic inference (Frank & Goodman, 2012). When referring to an object, listeners rely more heavily on speakers' utterances when no objects are particularly salient (have a low prior probability of being referred to). Similarly, people should rely on ads more heavily when they are more uncertain about the price of an object (i.e. when the price of a type of object varies greatly between instances). Consequently, if people do not know the dispersion of objects' prices, they will be unable to determine how much to rely on advertisements.

In the current study, we measured individuals' and populations' knowledge of the prices of everyday objects. In Experiment 1, we selected a set of object categories and asked participants to make multiple guesses about the price range of an object from each category. To measure participants' accuracy, we used Google Product search to obtain the true prices of objects from each category. In Experiment 2, participants performed the same price estimation task but we presented the object categories with positive or negative advertisement modifiers.

Participants accurately estimated the average prices of the objects. Aggregating all of our participants' responses together revealed that the population was very well calibrated to the dispersion of prices. However, while individual participants knew which objects' prices had greater or lower dispersion, they severely underestimated how much prices' dispersions varied overall. In Experiment 2, advertisements influenced participants' price estimates. Participants guessed objects were cheaper when presented with negative ads and more expensive when presented with positive ads and weighted ads in proportion to the dispersion of objects' prices. But, consistent with participants underestimating the dispersion of prices, the interaction between advertisements and price dispersion was relatively weak. Although people possess some distributional knowledge, that knowledge is highly limited compared to the knowledge of the population and falls far short of the true distribution of values in the world.

Experiment 1

We evaluated how much individuals vs. the population know about real-world price distributions by asking them to guess the prices of objects multiple times.

Methods

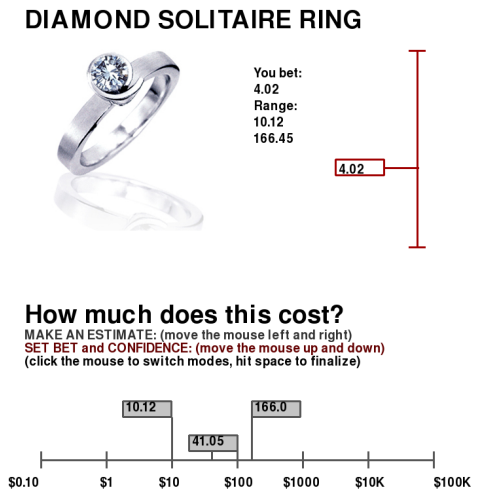


Figure 1. Example trial. Participants clicked to guess a price using the horizontal slider and then set a confidence range using the vertical slider. As participants adjusted the confidence slider, their current bet was displayed. Here, the participant has guessed the ring costs $41.05 with a range from $10.12 to $166.00. If the object's price falls in this range, the participant will earn 4.02 points. The larger the range, the fewer points. Participants could alternate between setting the guess and range. Exact bets and ranges were displayed as participants made their choice.

Participants. People from the Cambridge, MA community participated as part of a paid, daylong behavioral economics battery (the other study included in this battery is irrelevant for the present purposes—it investigated delayed discounting in behavior). Participants completed a varying number of blocks, depending on how quickly they completed each block. There were initially 29 participants; 23 returned for Session 2, 14 for Session 3 and 10 for Sessions 4 and 5. By spacing out responses across sessions, we hoped to decrease dependence between responses and ascertain participants' full distributional knowledge.

Stimuli. We selected 50 object categories (e.g. "Diamond solitaire ring", "Printer"), attempting to cover a wide span of empirical log price means and variances while minimizing the correlation between the empirical means and variances. All of our subsequent analyses use the log prices. We manually obtained object prices from Google Products searches for the categories. Despite our efforts, variance in log prices remained weakly, but significantly, correlated with mean log price ($r=.28,\ p=.048$). Indeed, participants appeared to expect a strong correlation, such that the means and standard deviations of their responses were highly correlated ($r=.55,\ p<.001$).

Procedure. Each trial, participants saw the name of an object category and a generic picture of that category (Figure 1). We told participants that each trial we had selected the price of a specific object from that category using an Internet search; this discouraged subjects from making repeated guesses or focusing on values like the mean. Participants guessed the price of the object by clicking on a log scale and then selecting a confidence interval around their guessed price. Participants earned points if the correct price was in the confidence interval. As the size of the confidence interval increased, participants' potential reward decreased and vice versa. Thus, participants were incentivized to set a sufficiently narrow confidence interval so that getting the correct answer was rewarded, but a sufficiently broad confidence interval such that they were likely to get a correct answer. To keep participants motivated, every few trials we told participants their current score.

Each block, participants guessed the price of objects from a random subset containing 35 of the categories, block randomized.

Results

Did participants know the mean price of objects? We compared the empirical mean prices calculated from the Google Product search results to subjects' estimated mean prices (Figure 2). Reported mean price was strongly correlated with the actual mean price ($r=.85,\ p<.001$), demonstrating that subjects knew which products were more or less expensive. To evaluate how well participants knew the exact prices of goods, we found the best fitting linear regression. A slope of 1 would indicate that on average participants knew the exact prices of objects. The regression line was close to one ($m=.81,\ 95\%\ confidence\ interval=.67—.96$), suggesting that together subjects knew the mean prices of objects fairly accurately.

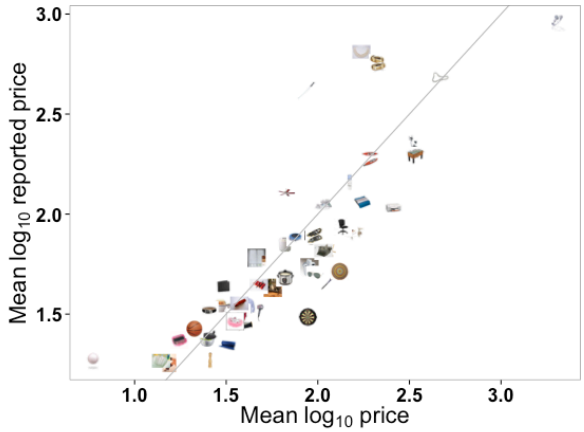


Figure 2. The mean reported price of objects as a function of their true price. The grey line denotes equality. Participants possessed very accurate knowledge of the average prices of objects.

Did the population know the dispersion of prices? We compared the true dispersion of prices to the dispersion of prices across all participants' responses (Figure 3A). For each object, we calculated the standard deviation of all the guesses that all the participants made. The true and estimated variances were strongly correlated (*r=.68, p<.001*), demonstrating that the group knew which prices were more or less variable. The slope of the linear regression was .48 (*95% confidence interval=.33—.63*), suggesting that the group accounted for roughly half of the actual variability in prices.

Did individual participants know the dispersion of prices? We next compared the true dispersion of prices to the dispersion of individual participants' responses (Figure 3B). For each participant, we calculated the standard deviation of their responses for each object category. We then found the average standard deviation of responses. Within-participant dispersion was significantly correlated with the true dispersion (*r=.51, p<.001*), demonstrating that participants knew which objects had more or less variable prices. However, the slope of the linear regression was much smaller (*.12, 95% confidence interval=.064—.19*) than the across-participant slope, indicating that despite knowing the relative dispersion of prices, individual participants were not as well calibrated to the absolute dispersion of prices as the overall population.

Did individual participants know the range of prices? Confidence interval width was strongly correlated with the true price variance (*r=.63, p<.001*) (Figure 3C) indicating that participants possessed explicit knowledge about the dispersion of prices. However, participants' reported confidence intervals and the dispersion of their responses explained comparable amounts of the true dispersion of prices. The 95% confidence interval of the confidence interval regression slope (*.11, 95% confidence interval=.068—.15*) overlapped with the within-participant dispersion regression slope, indicating that they captured similar levels of objects' price dispersions. This suggests that participants used the same impoverished knowledge of price distributions when generating confidence intervals and repeated guesses.

Experiment 2

Experiment 1 demonstrated that individuals represent the mean and dispersion of object prices, but are not as well calibrated on the dispersion of prices as the population. In Experiment 2, we examined whether participants' limited knowledge of prices' dispersions impaired their ability to use advertisements to infer the prices of new objects.

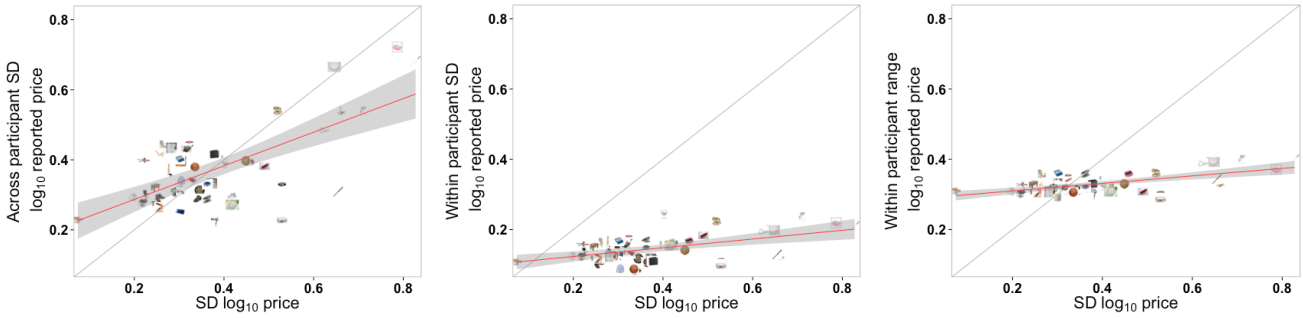


Figure 3. Across and within participant dispersion estimates as functions of the true price dispersions. A) The standard deviation of responses, aggregated across participants. B) The average standard deviation of each participant's responses. C) The average range of participants' confidence ranges. Grey lines indicate equality. Red lines with grey shading indicate linear regression fits with confidence intervals. Participants were generally sensitive to the dispersion of prices, but the population was much better calibrated than individuals.

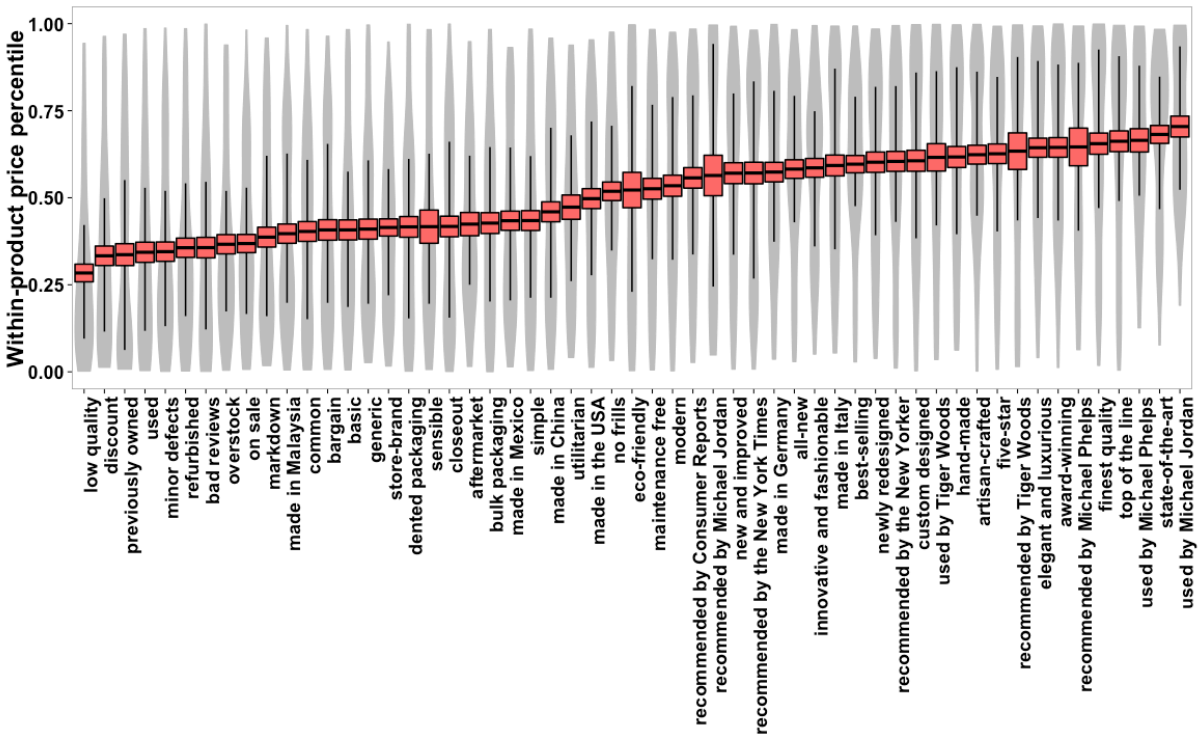


Figure 5. The influence of advertisement modifiers on price estimates. The percentile ranks of objects when accompanied with each type of modifier. The centers of the red boxplots denote the average percentile ranks, the tops and bottoms of the boxes denote ±1 SEM and the error bars indicate the 25th and 75th percentiles. The grey violin plots indicate the frequency of different percentile ranks for a given advertisement. The effectiveness of ads greatly varied: Very negative modifiers (left side) like "low quality" caused participants to think objects were cheaper, neutral modifiers (center) like "utilitarian" had minimal effect and very positive modifiers (right side) like "used by Michael Jordan" caused participants to report more expensive prices.

Methods

Participants. The participant pool was identical to Experiment 1. There were initially 25 participants; 24 returned for Session 2, 17 for Session 3 and 4 for Session 4 and 3 for more than 5 sessions (8, 14, and 24 sessions).

Stimuli. The stimuli were identical to Experiment 1. However, each stimulus was now accompanied by one of 55 ads. 28 of the ads were positive (e.g. "award winning", "five-star") and 27 were negative (e.g. "bargain", "overstock"). Each trial, we presented a random modifier alongside the object category image, with the constraint that objects in the same block could not have the same modifier.

Procedure. The procedure was identical to Experiment 1.

Results

Did the valence of ads affect participants' judgments? We first examined whether negative and positive advertisements caused participants to judge objects as cheaper and more expensive, respectively. For each object category, we rank ordered participants' responses and then converted the ranks to percentiles, such that the lowest and highest estimates were the 0th and 100th percentiles, respectively. We then found the average percentile for objects that had been accompanied by each modifier. Transforming responses into percentiles allowed us to compare people's price estimates despite the wide variance in objects' prices.

The efficacy of ads varied greatly (Figure 5). The most negative ad, "low quality", resulted in a mean rank of the 28th percentile (SEM=2.5) whereas the most positive ad, "used by Michael Jordan" resulted in a mean rank of 70th (SEM=3.0). These patterns demonstrate that participants used advertisements to infer the price of objects.

Did uncertainty about the price of an object increase the efficacy of advertisements? Participants may have relied on not only the advertisements but also their own knowledge about the distribution of objects' prices. In particular, we expected that participants would rely more heavily on advertisements when objects' prices varied more, as a way to compensate for their uncertainty. Given that individuals underestimated the dispersion of prices, however, we were uncertain whether participants would be

able to use objects' true price dispersions to determine how much to rely on advertisements.

To test how people inferred the prices of objects, we used the *lme4* package for R to design a suite of mixed effects models that utilized different types price and advertisement information. First, participants may have solely relied on their knowledge of objects' average prices. We wrote an average price model that treats objects' true mean prices as a fixed effect and participants and objects as random effects. Second, to test whether participants used advertisements to infer the prices of objects, we designed an advertisement model that builds upon the average price model by treating advertisements as a random effect. Finally, we evaluated whether participants relied on ads more when the dispersion of an objects' prices was high. We designed an advertisement-dispersion model that also accounts for the standard deviation of prices by adding an interaction between advertisements and the true price standard deviation. Thus our full model is defined by:

$$\log_{10} estimate_{p,i,m} = \beta_0 + \beta_1 Mean[\log_{10} price_i] + \gamma_p + \delta_i + \zeta_m + \theta_m SD[\log_{10} price_i]$$

$\log_{10} estimate_{p,i,m}$ is participant p's price estimate for object i when it is accompanied by advertisement modifier m. The second line corresponds to the nested average price model. β_0 is the intercept and β_1 is the coefficient for $Mean[\log_{10} price_i]$, the true mean price of the objects. γ_p is the offset for each participant and δ_i is the offset for each object. The third line corresponds to the additional advertisement modifier coefficients for the advertisement model. ζ_m is the offset for each ad. The fourth line corresponds to the interaction between true price variance and the advertisements in the advertisement-dispersion model. θ_m indicates how much each modifier affects estimates as the standard deviation of prices, $SD[\log_{10} price_i]$, changes.

Advertisements influenced participants' responses, such that the advertisement model fit (*AIC=3757*) participants' behavior much better than the average price model (*AIC=4336,* $\chi^2(1)=580$, $p<.001$). The advertisement-dispersion model had a lower but comparable AIC (*AIC=3753*) and provided a slightly better fit than the advertisement model ($\chi^2(2)=8.4$, $p=.015$). This pattern is consistent with participants using Bayesian inference to integrate prior knowledge about the prices of objects with new advertisements to determine value of novel objects. Nevertheless, the small improvement from the advertisement-dispersion interaction suggests participants' still failed to fully account for the variability of prices.

To more directly evaluate the effect of price dispersion on the use of ads, we used the full model to test whether the influence of ads increased with the dispersion of prices. We extracted the object-advertisement coefficient and then separately averaged the coefficients for positive valence and negative valence ads for each object. This yielded an average shift for positive ads and an average shift for negative ads for each object. We then calculated the absolute difference between the average positive ads and average negative ads. Larger absolute differences indicate that advertisements had a greater influence on the object's price (that is, the price was more malleable to the influence of advertisements).

There was a significant positive correlation between objects' price dispersions and the difference between positive and negative price estimates ($r=.44$, $p=.0014$) (Figure 6), indicating that participants used price distribution knowledge and advertisements to infer the prices of novel objects. However, consistent with the small improvement from the advertisement model to the full model, the magnitude of the interaction between price dispersions and ads was small. We compared the magnitude of the ad random effects to the ad-dispersion random effects and found that the ad-dispersion interaction accounted for only 3% of the overall variance attributed to advertisements. Thus, although participants relied more heavily on advertisements when objects had highly variable prices, participants' impoverished knowledge limited their ability to effectively weigh the value of ads.

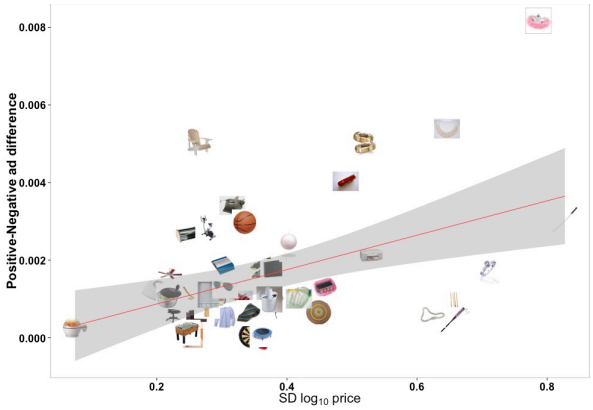


Figure 6. The influence of advertisements on price estimates given the dispersion of objects' prices. Advertisements had a larger impact on guesses when objects had more disperse prices.

Discussion

We asked people to guess the prices of objects from different categories and compared how well individuals vs. the population knew the distribution of values. Although individuals knew the mean price of objects well, they greatly underestimated the dispersion of prices compared to the population. Participants were also able to use their knowledge of price distributions to determine how much to rely on advertisements, but their behavior indicated that they still only knew a fraction of the overall dispersion of prices.

Both individuals and the overall population exhibited knowledge of the mean and the dispersion of prices in the world, though our population was much better calibrated to the absolute dispersion of prices. This deficiency may have

arisen from individuals being biased by their idiosyncratic experiences and contexts. Seeing more or less expensive objects could have resulted in participants anchoring their responses and respectively making higher or lower guesses (Tversky & Kahneman, 1974) or comparing our stimuli to objects they had recently encountered (Ungemach, et al., 2011; Vlaev, et al., 2011). In the future, accounting for people's histories, such as their socioeconomic backgrounds, may reveal differences in their price inferences.

Our experiments may have also been limited in their ability to reveal people's full knowledge of distributions. Although we tried to increase the independence of responses by spacing out sessions over hours, increasing delay intervals to weeks or months could reveal more extensive distribution knowledge. Vul & Pashler (2008), for example, found that responses made even after 3 weeks were relatively similar. Participants' responses may have also been influenced by decision biases. Participants, for instance, might have been risk averse and reported the mean to minimize their maximum loss or risk seeking and set small confidence intervals. These strategies could have limited participants' apparent distributional knowledge.

Additionally, people most likely possess more complex categorical knowledge about real-world values than just individual object categories (Hemmer & Steyvers, 2009; Hemmer & Persaud, 2014). The prices of cell phones, televisions and computers, for instance, may all fall under the category of electronics. People's judgments about the price of a new cell phone then might be constrained by both their knowledge of how much cell phones cost and also how much electronics in general cost. Furthermore, individual differences in the categorization of objects (is a smartwatch an electronic device or a fashion accessory?) may lead to distinct biases in how people estimate prices.

People used their prior knowledge of price distributions with advertisements to infer the prices of new objects, relying more heavily on advertisements when objects had highly variables prices. This behavior may reflect processes comparable to linguistic pragmatic inference (Frank & Goodman, 2012). In the case of pragmatic inference a speaker's utterance can help a listener select an object out of a crowd. Similarly, if a person buying a car is uncertain about the value of the car, an advertisement can help the buyer infer whether it's a steal or a lemon. Moving forward, using the framework of pragmatic inference to examine the role of advertisements in decision-making may help us learn why the effectiveness of our modifiers varied so much and how to craft more influential advertisements. More broadly, this approach may give insights into questions like how ads are interpreted in different contexts (Barner & Snedeker, 2008) and when people decide that ads are informative (Frank & Goodman, 2014).

People have some knowledge about the distribution of values in the world. Despite knowing the mean and the relative dispersion of prices, they have a poor idea of how much prices actually vary compared to the population. Furthermore, their impoverished distribution knowledge impairs their ability to appropriately weigh new information like advertisements to infer the prices of novel objects. Our future work will examine the sources of the idiosyncratic limitations on people's distribution knowledge.

References

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development*, *79*(3), 594—608.

Frank, M.C., & Goodman, N.D. (2012). Predicting pragmatic reasoning in language games. *Science*. *336*(6084), 998-998.

Frank, M.C., & Goodman, N.D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80-96.

Griffiths, T.L., & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767-773.

Hemmer, P., & Persaud, K. (2014). Interaction between categorical knowledge and episodic memory across domains. *Frontiers in psychology*, *5(584)*, 1-5.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189-202.

Hourihan, K.L., & Benjamin, A.S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(4), 1068-1074.

Lewandowsky, S., Griffiths, T.L., & Kalish, M.L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*(6), 969-998.

Mozer, M.C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*(7), 1133-1147.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Ungemach, C., Stewart, N., & Reimers, S. (2011). How incidental values from the environment affect decisions about money, risk, and delay. *Psychological Science*, *22*(2), 253-260.

Vlaev, I., Chater, N., Stewart, N., & Brown, G. D. (2011). Does the brain calculate value? *Trends in Cognitive Sciences*, *15*(11), 546-554.

Vul, E., Goodman, N., Griffiths, T.L., & Tenenbaum, J.B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599-637.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. Psychological Science, *19*(7), 645-647.

A Unified Framework for Bounded and Unbounded Numerical Estimation

Dan Kim (kim.3839@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH 43215 USA

John Opfer (opfer.7@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH 43215 USA

Abstract

Representations of numerical value have been assessed using bounded (e.g., 0-1000) and unbounded (e.g., 0-?) number-line tasks, with considerable debate regarding whether one or both tasks elicit unique cognitive strategies (e.g., addition or subtraction) and require unique cognitive models. To test this, we examined 86 5- to 9-year-olds' addition, subtraction, and estimation skill (bounded and unbounded). Against the measurement-skills hypothesis, estimates were even more logarithmic on unbounded than bounded number lines and were better described by conventional log-linear models than by alternative cognitive models. Moreover, logarithmic index values reliably predicted arithmetic scores, whereas model parameters of alternative models failed to do so. Results suggest that the logarithmic-to-linear shift theory provides a unified framework for numerical estimation with high descriptive adequacy and yields uniquely accurate predictions for children's early math proficiency.

Keywords: cognitive development; numerical cognition; number-line estimation; psychophysical function

Introduction

In this paper, we sought to resolve a debate on what gives rise to developmental changes in numerical estimation and provide a unified framework for understanding seemingly irreconcilable data regarding the psychophysical functions that link numbers to their magnitude estimates (Barth & Paladino, 2011; Cohen & Sarnecka, 2014; Opfer, Thompson, & Kim, 2016; Siegler & Opfer, 2003; Slusser, Santiago, & Barth, 2013).

Conventionally, developmental changes in numerical estimation have been viewed as following a logarithmic-to-linear shift in representations of numeric magnitude (Siegler & Opfer, 2003; Siegler, Thompson, & Opfer, 2009). This shift was first observed on a number-line task, in which a target number was estimated on a line flanked by a number at each end (Fig. 1A). On this task, young children's placement of numbers typically follows an approximately logarithmic function (Siegler & Booth, 2004; Siegler & Opfer, 2003; Opfer & Siegler, 2007; Opfer, Siegler, & Young, 2011; Thompson & Opfer, 2008), but this logarithmic pattern changes to a linear one later with age and experience, with timing depending on the number ranges tested (Siegler, Thompson, & Opfer, 2009). For example, on a 0-100 number line, where estimates of kindergarteners are logarithmic, second graders produce linear estimates (Booth & Siegler, 2006), while they estimate numbers on a log scale on a 0-1000 number line (Siegler & Opfer, 2003). That logarithmic-to-linear transitions appear at different times in development suggests that logarithmic and linear representations can co-exist and compete in the same child. Thus, a simple model of numerical estimation is thought to be a mixed log-linear model (MLLM), in which estimates are predicted as a weighted sum of logarithmic and linear transforms of the number to be estimated (Anobile et al., 2012; Opfer et al., 2016)

Recently, two related challenges to the logarithmic-to-linear shift theory have been raised. The first challenge argued that children's estimates reflect one of three proportional reasoning strategies (Barth & Paladino, 2011; Cohen & Blanc-Goldhammer, 2011; Cohen & Sarnecka, 2014; Slusser, Santiago, & Barth, 2013), which had been modeled in adults using three different cyclic power models (Hollands & Dyre, 2000). By comparing multiple extensions of the cyclic power model (CPM) to a log or linear model, Slusser et al. (2013) showed that one of the power models showed better fits for a majority of 5- to 10-year-olds' estimates than did the logarithmic model.

However, a recent study by Opfer and colleagues (2016) provided evidence against this account, showing that the fit of the cyclic power models was an artifact of an unusual anchoring procedure used by Slusser et al. (2013) that did not characterize estimates using the standard "free" numerical estimation procedure. Further, Opfer et al. (2016) found that both free and anchored estimates were better fit by MLLM than a mixed cyclic power model (MCPM) that included all variants of the CPMs proposed. Within this MLLM, the effect of anchoring could be traced to decreasing the logarithmicity of estimates without any need

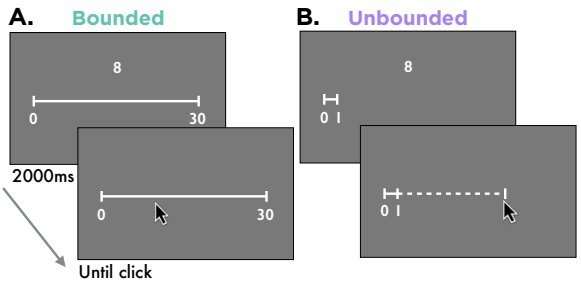


Figure 1. Illustration of the bounded and unbounded number-line task.

for positing unique estimation strategies.

A second challenge to the logarithmic-to-linear shift account was the measurement-skills account proposed by Cohen and colleagues (2011, 2014). Here, log-like patterns in estimates are proposed to arise because conventional number-line tasks are "bounded" by two numbers, thereby requiring sophisticated arithmetic skills, like subtraction, which young children may not have acquired (Fig. 1A). In this account, when a target number such as 70 is given on a 0-100 number line, a participant may need to consider both the number 70 and the difference 30 after subtracting 70 from 100. In contrast, estimation of numbers on a number line "unbounded" by the larger endpoint is thought to require only addition (Fig. 1B), is thought to more accurately reveal numeric magnitude judgments, and is thought to be best modeled using three more variants of power models called "scallop power models" (SPM) (Cohen & Blanc-Goldhammer, 2011).

To test this second challenge, Cohen and Sarnecka (2014) tested the effects of boundedness in number-line tasks by giving both bounded and unbounded tasks to 3- to 8-year-olds. Model fits of a log or linear function were compared to those of multiple variations of the CPM for the bounded estimates and to those of multiple SPMs for the unbounded estimates. In the bounded condition, the extensions of cyclic models provided better fits to estimates, and the response bias parameters (βs) changed with age. For the unbounded number-line estimates, in contrast, one of three scallop power models predicted the estimates better, but the estimation bias parameters (βs) stayed constant across age groups. These findings led the researchers to conclude that there were no such things as log-to-linear shifts in numerical representation, but these changes stemmed from the poor use of arithmetic strategies. Also, they suggested that numerical estimation be made on an unbounded number line for a better assessment of number representations.

The Current Study

In this paper, we tested three rival hypotheses regarding bounded and unbounded numerical estimation. The primary hypothesis is that the cognitive process of estimation in the two tasks is essentially similar and does not require positing six estimation strategies associated with six unique psychophysical functions. From this perspective, bounded and unbounded numerical estimates -- like free and anchored numerical estimates -- are best viewed as reflecting children's representations of numeric magnitude and best modeled using the same mixed log-linear model. The second hypothesis is that both tasks are equally well suited for characterizing numerical magnitude estimates, with *the parameters of the MLLM derived from each task providing better predictors of children's use of numbers in other contexts (such as addition and subtraction) than even the models that allegedly track addition and subtraction skill*. Finally, because number-line estimation tasks do assess children's representations of numeric magnitude, providing children with more numbers against which to anchor their estimates will result in improved performance, leading to bounded numerical estimates being more accurate and more linear than unbounded estimates.

To test these hypotheses, we nearly replicated the procedure used by Cohen and Sarnecka (2014) with the only exception being to use a fully-balanced design that could detect order effects and the administration of a battery of math tests (addition and subtraction). Then, we pit the MLLM tested by Opfer et al. (2016) against the MCPMs for bounded number-line tasks and against the mixed scallop power model (MSPM) for unbounded number-line tasks.

The MLLM consists of logarithmic and linear components and is defined as:

$$y = a\left(\lambda \frac{U}{\ln(U)} \ln(x) + (1-\lambda)x\right), \quad (1)$$

in which y indicates an estimate of number x on a 0-U number line. Also, a denotes a scaling parameter, and λ is a logarithmicity index that measures the degrees of logarithmic compression in estimates. If estimation is perfectly linear, a λ value converges to 0, whereas the value of the logarithmicity index gets close to 1 as estimation shows more logarithmic compression.

The MCPMs were formulated as proposed by Hollands and Dyre (2000) (also see Opfer, Thompson, & Kim, 2016, for details). The first MCPM (MCPM1) is formalized based on Slusser et al. (2013)'s study that hypothesized that the number of reference points used for estimation changes in development: children with poor proportion skills would only use a single reference point, i.e., the lower bound, (0 cyclic power model), and then learn to use the lower and upper bounds (1 cycle power model) and the middle point with the two endpoints (2 cycle power model) as they become more familiar with the number range. The MCPM1 is defined as:

$$y = w_1 \cdot 0\text{CPM} + w_2 \cdot 1\text{CPM} + w_3 \cdot 2\text{CPM}, \quad (2)$$

where each of w_1, w_2, and w_3 denotes a weight for each variant of the CPM respectively. Each weight and the sum of weights are constrained to be between 0 and 1, so that contribution of three models in a response can be assessed individually. The MCPM2 is identical to the MCPM1 except that 0CPM is replaced with the subtraction bias cyclic model (SBCM) in the MCPM2 as proposed by Cohen & Sarnecka (2014). The SBCM was similar to 1CPM, but includes an additional parameter (s) that is associated with the subtraction bias.

The MSPM for unbounded number-line tasks is also formed in the same manner based on Cohen and Blanc-Goldhammer (2011)'s assumption that the unbounded tasks are solved using an addition strategy. The following is the formalization of the mixed model:

$$y = w_1 \cdot 1\text{SPM} + w_2 \cdot 2\text{SPM} + w_3 \cdot \text{MSPM}. \quad (3)$$

In the model, 1SPM indicates the single scallop model, 2SPM the dual scallop model, and MSPM the multiple scallop model. The same constraints set on weights in the mixed CPM are set in this model.

After obtaining model fits and parameter estimates from each of the models, we next compared fits of the MLLM with the MCPMs for the bounded condition and with the MSPM for the unbounded condition to examine the best-fitting model for each number-line task. Whereas the alternative functions are task-specific, the MLLM is theoretically applicable to both bounded and unbounded conditions. Therefore, if the MLLM is an unifying, generalizable function that captures numerical representations regardless of number-line boundedness, the MLLM should not only describe the bounded estimates better (Opfer et al., 2016), but also predict data from the unbounded condition better. Also, we examined whether model parameters, such as subtraction bias (s) in the MCPM2 and logarithmicity (λ) in the MLLM, actually predict addition and subtraction accuracy. If compressive estimates appear due to a lack of subtraction skills in the bounded number-line tasks, significant correlation between subtraction bias (s) and actual subtraction performance should be observed. On the other hand, if compression in estimates reflects logarithmic representations of numbers, the logarithmicity index (λ) should predict both addition and subtraction achievement better than other arithmetic-strategy or estimation bias parameters.

Experiment

Methods

Subject Thirty 5- to 6-year-old kindergarten, 30 first grade, and 26 second grade students were recruited in Columbus, OH (kindergarteners: 17 female children, M = 5.90 years, SD = .32 years; first graders: 19 female children, M = 6.74 years, SD = .39 years; second graders: 21 female children, M = 7.91 years, SD = .46 years).

Materials and Procedure Participants completed both bounded and unbounded number-line tasks given in a counterbalanced order. In the bounded condition, a number was shown for 2,000 ms above a number line flanked by 0 and 30/100/1000 (Fig. 1A). On every trial, the mouse cursor was reset to be located at the 0 point and moved only horizontally on a number line. Participants were instructed to estimate a given number on a number line with the following instruction:

> Now we're going to play a game with numbers. This is a number line. In this game, each number line will have a 0 at one end and 30/100/1000 at the other end. There will be a number up here. Your job is to show me where that number goes on a number line like this one. When you decide where the number goes, you have to drag this little mark to where the number should go. When you're ready to go again, press the green bar (spacebar with a green sticker on) on the keyboard.

The unbounded number-line task was identical to the bounded one except that a single-unit line (0-1) was presented instead of the full range number line (Fig. 1B). An experimenter introduced the task with the following instruction that was created based on Cohen and Sarnecka (2014):

> Now we're going to play a game with numbers. This is a number line. In this game, each number line will have a 0 here and 1 over here. All the other numbers go after 1. There will be a number up here. Your job is to show me where that number goes on a number line like this one. When you decide where the number goes, you have to drag this little mark to where the number should go. When you're ready to go again, press the green bar on the keyboard.

The length between 0 and 1 was adjusted based on the ranges of number lines. In other words, a displayed length for 0-1 intervals was the shortest for a 0-1000 number line, whereas it was the longest in the 0-30 number line task.

Based on Slusser et al. (2013), different number ranges that would elicit compressive estimates were used depending on children's grades. 5- and 6-year-old kindergarteners were given 0-30 number-line tasks with to-be-estimated numbers sampled evenly from a 0-30 range: 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 30. For first graders, a 0-100 number line was used with 25 numbers between 0 and 100: 3, 4, 6, 8, 12, 17, 21, 23, 25, 29, 33, 39, 43, 48, 52, 57, 61, 64, 72, 29, 81, 84, 90, 96, 100. Second graders were asked to estimate numbers chosen between 0 and 1000 on a 0-1000 number line: 2, 5, 18, 34, 56, 78, 100, 122, 147, 150, 163, 179, 246, 366, 486, 606, 722, 725, 738, 754, 818, 938, 1000. In both bounded and unbounded conditions, the same target numbers were randomly presented. To keep children's attention on the tasks, a neutral sound was produced once a stimulus was displayed or a response was made by mouse click. The tasks started after an instruction without any practice, and there was no feedback provided over trials.

Upon completion of the two number-line tasks, arithmetic performance was assessed with paper-and-pencil addition and subtraction tests. The addition test consisted of 50 one-digit addend problems, such as 1+1 and 5+3. The subtraction items were generated by rearranging the addition. To be specific, the first addends in the addition test were used as subtrahends, while the second addends became the answers. For the questions like 1+1 = 2 and 5+3 = 8, they were rearranged to be 2-1 and 8-5 in the subtraction. Participants were given the tests in a random order and asked to solve as many problems as they could within 1 minute for each test.

Results

Whether cognitive process of estimation is qualitatively different between bounded and unbounded tasks was examined using psychophysical functions. For the bounded condition, individual data were fit by the MLLM, MCPM1, and MCPM2, whereas estimates on unbounded number lines were fit with the MLLM and MSPM. The model fits were then assessed by comparing AICc values.

Against the measurement-skills account, none of the MCPMs was the best fitting model for bounded number-line estimation. For 0-30 and 0-100 number-line estimates, most

children's performance was better explained by the MLLM (Table 1). Estimates of 50% of participants on 0-1000 number lines were also best fit by the MLLM. In the unbounded condition, across all the three number ranges, 100% of children produced estimates that were better described by the MLLM than the MSPM. The results suggest that logarithmically distorted estimates in a number-line task do not stem from a lack of proportion, subtraction, or addition skills, but come from representations of numerical magnitude that are well predicted with the mixed log-linear model.

Would variants of the power models based on arithmetic strategies predict children's actual arithmetic performance? To address this question, we next correlated each child's addition and subtraction performance with the best-fitting parameter values from the models. If response patterns in the bounded condition came from poor subtraction skills, parameters, such as βs of the MCPM1 & MCPM2 and s of MCPM2, would be expected to correlate with arithmetic scores. As presented in Table 2, however, none of the βs from the MCPMs was a reliable predictor for either addition or subtraction scores: only $\beta_{1\text{CPM}}$ of MCPM1 correlated with subtraction scores, but such correlation was not found in the subtraction bias parameter s of MCPM2. This correlation also remained insignificant if adjusted parameter values (absolute values of β-1 or s-1) were used for analyses. In contrast, the logarithmicity parameter λ of the MLLM reliably predicted both addition and subtraction performance, $r(84) = -.42$, $p < .001$ for addition; $r(84) = -.36$, $p < .01$ for subtraction. The finding suggests that the more logarithmic their estimates were in bounded tasks, the worse performance in arithmetic tests was observed.

We also examined relations between the parameter βs of the MSPM and arithmetic achievement in the unbounded tasks. According to the measurement-skills account, these parameters in the scallop models should reflect representation of numeric magnitude most accurately of all because the task is supposedly pure of the sins of boundedness and the model is correct. Against this idea, however, none of the parameter βs were significantly correlated with arithmetic performance, whereas the MLLM parameter λ again showed a significant correlation with addition and subtraction, $r(84) = -.33$, $p < .01$ for addition; $r(84) = -.34$, $p < .01$ for subtraction (Table 2). These findings again support the idea that logarithmic estimates on a number line should be viewed as a reflection of logarithmic representation of numbers, which interferes with use of numbers in other contexts.

Table 1. Percent of participants best fit by each model.

	Bounded			Unbounded	
	MLLM	MCPM1	MCPM2	MLLM	MSPM
0-30	96.67	.00	3.33	100	.00
0-100	80	10	10	100	.00
0-1000	50	3.85	46.15	100	.00

The final issue we examined was to quantify the effect of boundedness on numerical estimation. If the bounded number-line tasks required more challenging measurement skill, the bounded tasks would be expected to elicit more erroneous and compressive estimates. On the other hand, if the two endpoints provide better anchors against which to judge large numbers, estimates in the bounded condition would be expected to yield less logarithmic and more accurate estimates. Consistent with the latter account, on all the 0-30/100/1000 number-line tasks, percent absolute errors (PAE) in median estimates were always greater in the unbounded than bounded conditions (bounded $M = 12\%$, $SD = 5\%$, unbounded $M = 17\%$, $SD = 7\%$ for 0-30 range; bounded $M = 12\%$, $SD = 4\%$, unbounded $M = 14\%$, $SD = 7\%$, for 0-100 range; bounded $M = 20\%$, $SD = 12\%$, unbounded $M = 24\%$, $SD = 14\%$, for 0-1000 range).

Additionally, as shown in Figure 2A, in all the three number ranges, the values of the logarithmicity parameter (λ) were greater for median estimates in the unbounded condition than those in the bounded condition (bounded $\lambda = .30$, unbounded $\lambda = .76$ for 0-30 range; bounded $\lambda = .42$, unbounded $\lambda = .55$ for 0-100 range; bounded $\lambda = .73$, unbounded $\lambda = .75$ for 0-1000 range). These results do not support the measurement-skills prediction that unbounded number-line tasks are easier and require less advanced mensuration skills than conventional bounded tasks.

Finally, the same analyses were repeated on individual children's data. For the 0-30 number-lines, greater PAEs were again observed in the unbounded than bounded condition ($M = 17\%$, $SD = 7\%$ for bounded; $M = 21\%$, $SD = 7\%$ for unbounded), $t(29) = 3.82$, $p < .001$. For the 0-100 number-lines, greater PAEs were also observed in the unbounded than bounded condition ($M = 15\%$, $SD = 7\%$ for

Table 2. Correlation between parameter values and arithmetic performance after number-line ranges were controlled for.

			Addition	Subtraction
Bounded	MLLM	λ	-.42***	-.36**
	MCPM1	β_{0CPM}	-.01	-.10
		β_{1CPM}	.17	.27*
		β_{2CPM}	-.08	.10
	MCPM2	s	.13	.06
		β_{SBCM}	.02	-.09
		β_{1CPM}	.17	.14
		β_{2CPM}	-.06	-.13
Unbounded	MLLM	λ	-.33**	-.34**
	MSPM	β_{1SPM}	-.05	.00
		β_{2SPM}	.19	.01
		β_{MSPM}	.16	.15

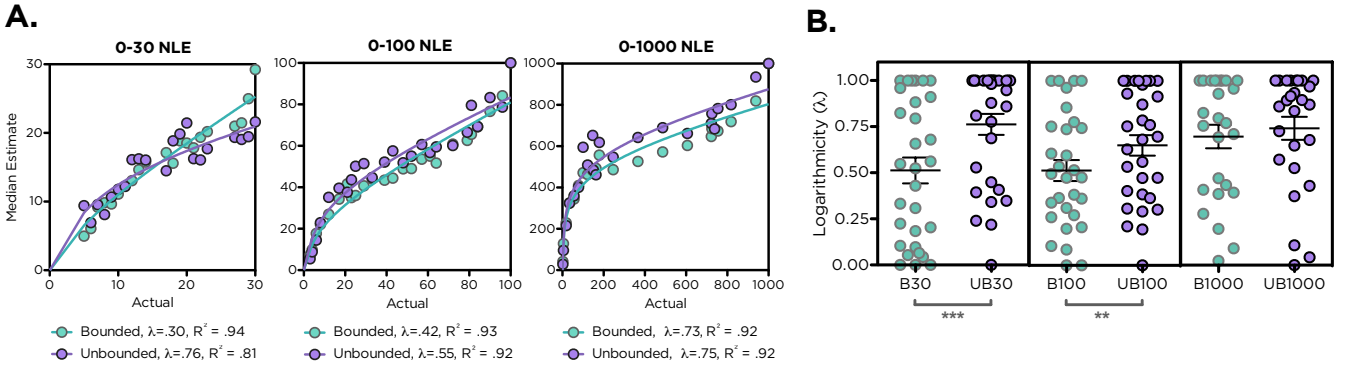


Figure 2. A: Median estimates in 0-30, 0-100, and 0-1000 number lines with and without bounds. B: Individual logarithmicity (λ) across the three number-line tasks.

bounded; M = 19%, SD = 9% for unbounded), $t(29) = 3.51$, $p < .01$. And for 0-1000 number-lines, greater PAEs were again observed in the unbounded than bounded condition (M = 20%, SD = 8% for bounded; M = 27%, SD = 11% for unbounded), $t(25) = 5.21$, $p < .001$. Therefore, for all the three number ranges, conventional number lines with two bounds elicited more accurate responses than unbounded number lines. These results are not at all consistent with the measurement-skills account.

Although number-line tasks without an upper bound produced more erroneous estimates, it is not clear whether the unboundedness increased random noise or logarithmic compression in estimates. To address this issue, individual children's degrees of logarithmicity were obtained with the MLLM fitting.

As shown in Figure 2B, in all the three number ranges, the values of the logarithmicity measure were greater for a number line without bounds. In 0-30 and 0-100 number-line tasks, estimates in the unbounded condition showed significantly greater logarithmic compression in the 0-30 range (M = .51, SD = .38 for bounded; M = .76, SD = .31 for unbounded), $t(29) = 4.33$, $p < .001$, and in the 0-100 range (M = .51, SD = .31 for bounded; M = .65, SD = .30 for unbounded), $t(29) = 3.37$, $p < .01$. Although logarithmicity was slightly greater for the unbounded 0-1000 number-line task than for the bounded one, the difference was not statistically significant (M = .70, SD = .32 for bounded; M = .74, SD = .32 for unbounded), $p > .05$. Against the arithmetic strategy hypothesis, negatively accelerating patterns in estimates were observed in both bounded and unbounded number-line tasks, and these patterns were stronger when the number line did not have a right-end bound. Even if the number-line task without the right endpoint revealed more accurate representation of numbers as the measurement-skills account claimed, what it revealed in the unbounded task was more logarithmic compression.

Discussion

In this paper, we attempted to address the nature of developmental changes in numerical estimation and to provide a unified framework for seemingly irreconcilable data regarding the psychophysical functions that link numbers to their magnitude estimates (Barth & Paladino, 2011; Cohen & Sarnecka, 2014; Opfer, Thompson, & Kim, 2016; Siegler & Opfer, 2003; Slusser, Santiago, & Barth, 2013). Specifically, we wished to see whether the logarithmic-to-linear shift existed on both bounded and unbounded number lines. As Cohen and Sarnecka (2014) stated, "For those children whose data are best fit by the logarithmic function, we are skeptical that this reflects a logarithmically organized quantity representation. If it did, we should expect to see the same pattern on the unbounded task…." (p. 1650).

In the teeth of this skepticism, we found robust evidence for a logarithmically organized quantity representation on the unbounded task, as well as the bounded task. Evidence came from multiple sources. First, the MLLM—which had previously unified discrepant data from free and anchored numerical estimation (Opfer et al., 2016)—predicted data from bounded and unbounded tasks better than rival psychophysical models. Additionally, for both tasks, the logarithmicity parameter was significantly greater than zero in each of the three age groups tested. This success of a single simple model over many competing complex models suggests there may be a simple cognitive process lurking behind many varieties of numerical estimation.

Second, individual differences were stable across the two tasks: children whose estimates were most logarithmic in the bounded task were also most logarithmic patterns in the unbounded task, $r(84) = .73$, $p < .001$. This wouldn't be expected if the two tasks elicited radically different estimation strategies or if only one of the tasks provided an accurate picture of children's numeric magnitude judgments.

Third, although the two tasks differed in overall accuracy, it was not in the direction predicted by the measurement-skills account: PAEs and logarithmicity values of both median and individuals' estimates were always greater for the unbounded condition across all three number ranges. Therefore, even if the more difficult unbounded tasks provide a better picture of the internal mental number line, what they show is *stronger* evidence for logarithmic representation of numbers.

Finally, the parameters expected by the measurement-skills account to track arithmetic biases unique to each task never predicted actual arithmetic performance, calling into question the psychological meaning of these parameters (Table 2). In contrast, the logarithmicity component of the MLLM reliably predicted children's arithmetic proficiency regardless of whether bounded or unbounded number lines were used, which is wholly inconsistent with the idea that bounded number lines require greater arithmetic skills than unbounded ones.

Why might our results have differed so strongly from those of Cohen and colleagues (2011, 2014)? Two reasons—methodological and analytical—seem likely. First, unlike Cohen and Sarnecka (2014), we counterbalanced the order of bounded and unbounded tasks so that any effects of task order would not influence our overall results. Counterbalancing also made it possible to test the presence of order effects. And we found that the order of tasks sometimes influenced estimates. For example, on the 0-1000 task, more erroneous estimates were produced in *both* tasks if the unbounded task was first presented, $F(1, 24) = 9.79$, $p < .01$, $\eta_p^2 = .29$, suggesting that the greater difficulty of the unbounded task resulted in fatigue or confusion for the second task. This is critical because it suggests that unbounded tasks are not intrinsically easier for children – quite the opposite.

Another important difference came from the analytic strategies employed. Cohen and Sarnecka (2014) tested the log *or* linear model against the 0CPM *or* 1CPM *or* 2CMP model. Because the number of *or*'s corresponds to greater overall model complexity (i.e., greater degrees of freedom), a poorer fit for one of *two* family A models than one of *three* family B models is ambiguous. Our own analytic strategy, in contrast, was to explicitly include all of the same family of models in the same equation (e.g., MLLM, MCPM1&2, MSPM), and thereby include the correct number of degrees of freedom. This strategy allowed us to improve the validity of our model comparison. Additionally, it yielded two unique insights: that individual differences in the logarithmicity of estimates were stable across tasks and that individual differences in logarithmicity of estimates predicted math skills better than alternative models. These findings directly contradicted the predictions of the measurement-skills account and could not be tested using the analytic strategy of Cohen and Sarnecka (2014).

In conclusion, the present study shows that the log-to-linear shift theory provides a framework for numerical estimation that is high in descriptive adequacy for both bounded and unbounded number lines and is reliably predictive of children's mathematics performance. An interesting question for future studies is whether the same framework provides a good account of magnitude estimation more broadly.

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2012). Linear mapping of numbers onto space requires attention. *Cognition, 122*(3), 454-459.

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: evidence against a representational shift. *Developmental science, 14*(1), 125-135.

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental psychology, 42*(1), 189.

Cohen, D. J., & Blanc-Goldhammer, D. (2011). Numerical bias in bounded and unbounded number line tasks. *Psychonomic Bulletin & Review, 18*(2), 331-338.

Cohen, D. J., & Sarnecka, B. W. (2014). Children's number-line estimation shows development of measurement skills (not number representations). *Developmental psychology, 50*(6), 1640.

Hollands, J.G., & Dyre, B. (2000). Bias in proportion judgments: the cyclical power model. *Psychological Review, 107*, 500–524.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*(3), 169-195.

Opfer, J. E., Siegler, R. S., & Young, C. J. (2011). The powers of noise‐fitting: reply to Barth and Paladino. *Developmental Science, 14*(5), 1194-1204.

Opfer, J. E., Thompson, C. A., & Kim, D. (2016). Free versus anchored numerical estimation: A unified approach. *Cognition, 149*, 11-17.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development,* 428-444.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation evidence for multiple representations of numerical quantity. *Psychological science, 14*(3), 237-250.

Siegler, R. S., Thompson, C. A., & Opfer, J. E. (2009). The Logarithmic‐To‐Linear Shift: One Learning Sequence, Many Tasks, Many Time Scales. *Mind, Brain, and Education, 3*(3), 143-150.

Slusser, E. B., Santiago, R. T., & Barth, H. C. (2013). Developmental change in numerical estimation. *Journal of Experimental Psychology: General, 142*(1), 193.

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effects of context on age and sex differences in symbolic magnitude estimation. *Journal of Experimental Child Psychology, 10*1(1), 20-51.

Temporal Causal Strength Learning with Multiple Causes

Cory J. Derringer (cjd78@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Abstract

When learning the relation between a cause and effect, how do people control for all the other factors that influence the same effect? Two experiments tested a hypothesis that people focus on events in which the target cause changes and all other factors remain stable. In both four-cause (Experiment 1) and eight-cause (Experiment 2) scenarios, participants learned causal relations more accurately when they viewed datasets in which only one cause changed at a time. However, participants in the comparison condition, in which multiple causes changed simultaneously, performed fairly well; in addition to focusing on events when a single cause changed, they also used events in which multiple causes changed for updating their beliefs about causal strength. These findings help explain how people are able to learn causal relations in situations when there are many alternative factors.

Keywords: causal learning; causality; causal strength; conditionalizing

Introduction

Learning the strengths of causal relationships is essential for successfully manipulating the environment. Consider a student deciding whether to take a GRE prep course. Knowing the extent to which taking the course would improve or harm her score would help the student decide whether the course is worthwhile.

However, causal relationships do not exist in isolation; often multiple causes influence the same effect in different ways (e.g., whether the student has been taking practice tests or drinking at the local pub). Further, the causes could be confounded; one more hour studying may mean one less hour sleeping. In the same way statisticians use regression to estimate the effect of one variable above and beyond another, individuals should attempt to control for alternative causes to estimate the unique strength of a target cause.

In the next section we briefly review the dominant theory for how people control for alternative causes. We then propose a new theory called the Informative Transitions heuristic, and report two experiments that tested this theory.

Controlling for Alternatives Using Focal Sets

Focal set theory (Cheng & Novick, 1990; Cheng & Holyoak, 1995) proposes that when assessing the strength of a target cause (A) on an effect, people examine a subset of the data in which the other factor(s) are constant. For example, given the data in Figure 1, a learner could choose the subset in which B=0 (the gray shading). Within this subset a learner could calculate causal strength. (There are multiple theories for how people calculate causal strength within a focal set; see Hattori and Oaksford (2007) for a summary.)

Time	1	2	3	4
Cause A	1	1	0	0
Cause B	0	1	0	1
Effect	1	1	0	1

Figure 1: A focal set (gray shading) for cause A conditional on cause B's absence

It has been widely demonstrated that people often control for alternative causes when estimating the strength of a target cause (Spellman, 1996; Spellman, Price, & Logan, 2001; Waldmann & Holyoak, 1992; Waldmann, 2000; Waldmann & Hagmayer, 2001). However, the precise mechanism(s) is less clear. We first review two limitations of focal sets, and then discuss other options.

First, when the alternative cause B is a binary variable, two focal sets may be used: the presence or absence of B. Using these different focal sets, the subject could come to different conclusions about the strength of the target cause. In Figure 1, the focal set B=0 implies that A causes the effect (A is correlated with the effect in this focal set), but the focal set B=1 implies that it does not.

The second problem is that it is unclear exactly how people would use focal sets when there are multiple alternative factors. Imagine a situation in which there are four possible causes (A-D) of an effect. When assessing the causal strength of A, a learner could choose a focal set such as (B=0, C=0, D=0); however, using this focal set ignores the other 7 possible combinations of B, C, and D, which means ignoring 7/8ths of the potential data. If there are 8 causes (Experiment 2), choosing any one particular focal set (e.g., when all the other factors are 0), involves ignoring 127/128ths of the potential combinations. If a learner only experiences a small number of observations, there may not be observations in which A=1 and A=0, while all the other factors are 0; the relation between A and the effect cannot be inferred. Thus, selecting one focal set for each cause seems an inefficient strategy. In sum, there is a tension between using just one focal set, which involves discarding data, and using multiple, which requires integrating the conclusions in some yet-unspecified way.

Controlling for Alternatives in Other Ways

Here we list a couple alternative approaches to controlling for alternative causes aside from focal sets. First, in many learning situations, associative theories (e.g., Rescorla &

Wagner, 1972) asymptote to focal sets (see Spellman, 1996). One prediction of the standard version of Rescorla and Wagner's learning rule (RW) is that the associations between a cue and outcome get updated only on trials when the cue is present, which we will assess in the following experiments.

Second, when there are too many alternatives, people may stop controlling for alternative causes due to a working memory overload (Goedert, Harsch, & Spellman, 2005), in which case the causal strength estimates would resemble the bivariate relation between each cause and the effect.

Third, though we are not proposing it as a descriptive theory, it is useful to consider multiple regression as a gold standard computational-level theory, especially when there are many alternatives that need to be controlled for.

Using Informative Transitions

The current experiments test an alternative way that people may calculate conditional causal strength judgments as they experience data over time, which we call the Informative Transitions heuristic. We propose that people leverage the fact that causes in the environment are often autocorrelated—that they often remain in the same state for extended periods of time[1]. For example, if someone is in a bad mood in the morning, they are likely to be in a bad mood at lunchtime. Some degree of stability occurs for many other variables (the economy, one's health etc.). The data in Table 1a are an example in which each of the four causes is fairly stable over time. When causal factors change fairly infrequently, causes often change in isolation (provided that they are independent).

These transitions from one timepoint to the next in which only one potential cause changes and the others remain stable (hereon "informative transitions", or "IT"s) can give the learner unique insight into the relationship between the changing cause and the effect. Any change or lack of change in the effect during an IT informs the learner about the relationship between that cause and the effect holding the other causes constant. If the cause and effect change in the same direction (positive IT), it provides evidence that the cause is positive (generative) such as Table 1a Time 6-7 for Cause A. If they change in the opposite direction (negative IT), then the cause is more likely to be negative (inhibitory) (e.g., Table 1a, Time 4-5 for Cause D). If the cause changes but the effect stays constant (neutral IT), it may not be a cause at all (Time 3-4 for Cause B in Table 1a).

According to the IT heuristic, transitions in which multiple causes change simultaneously (Table 1b) are less informative; a change in the effect cannot be attributed to a single cause. If people use the IT heuristic, they will learn about causal strength more from ITs (Table 1a) than when multiple causes change (Table 1b). By contrast, focal set theory predicts no differences based on trial order.

The IT heuristic retains the most compelling aspect of the focal set theory: it simplifies scenarios with many causes, allowing the subject to examine the bivariate relationship between each cause and the effect. In a sense, the IT heuristic can be viewed as dynamically using different focal sets at different times (e.g., for Time 1-2 in Table 1a, the target cause is B, and the focal set is A=1, C=0, D=1; for Time 2-3 the target cause is C, and the focal set is A=1, B=0, D=1, etc.).

Table 1a: Dataset wherein all transitions are informative

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cause A	1	1	1	1	1	0	1	1	1	1	1	1	1	0
Cause B	1	0	0	1	1	1	1	1	1	0	0	0	1	1
Cause C	0	0	1	1	1	1	1	1	0	0	1	1	1	1
Cause D	1	1	1	1	0	0	0	1	1	1	1	0	0	0
Effect	0	0	0	0	1	0	1	0	0	0	0	1	1	0

Table 1b: Data from Table 1a, rearranged to remove ITs

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Cause A	1	1	0	1	1	1	1	1	1	1	1	1	0	1
Cause B	1	0	1	1	0	1	1	0	1	0	1	0	1	1
Cause C	1	0	1	1	1	0	1	1	1	0	1	1	1	0
Cause D	1	1	0	1	0	1	0	1	0	1	0	1	0	1
Effect	0	0	0	0	1	0	1	0	1	0	1	0	0	0

The current experiments test the IT heuristic by changing the order of the learning data to manipulate the number of ITs. Tables 1a and 1b are an example; they have the same 14 trials, but in a different order. If learners use the IT heuristic, they will be more successful at detecting the underlying causal relations in datasets like Table 1a than 1b.

Experiment 1

Method

Participants For each experiment, 100 participants located in the United States were recruited through Amazon's Mechanical Turk service. Each participant had previously completed at least 100 tasks on MTurk and had an approval rate of at least 95 percent. Participants were paid two dollars, with accuracy bonuses as an incentive. The study lasted approximately 15 minutes.

Design We examined how participants learn about positive causes, negative causes, and neutral factors (non-causes). We created three types of datasets so that participants would not have a strong expectation of the number of each type of cause: one positive, one negative, two neutral; two positive, one negative, one neutral, and one positive, two negative, one neutral. Twenty datasets were created for each type. Each participant viewed six datasets (two of each type), randomly selected and presented in random order.

[1] In reality, the number of causes that change on average from one observation to the next depends on how many causes are being observed and how frequently the causes change on average. If there is a very large number of causes, then typically more than one will change. However, if the causes are very stable, then it is possible for only one (or even less than one) cause to change on

The independent variable (between subjects) was the order of the trials within a dataset: whether each transition was informative (Table 1a) or not (Table 1b).

The datasets were created manually in the following way. On each trial, one cause was selected to change, and the effect was changed in the appropriate direction. For example, if a positive cause was selected to change, the effect would change in the same direction. By contrast, when a negative cause changed, the effect changed in the opposite direction. When a neutral cause changed, the effect did not change. Each cause changed either three or four times overall, and there were 14 trials for each dataset.

With a logistic regression predicting the presence/absence of the effect, this scheme guaranteed that positive and negative causes always had strong logistic regression weights (51.13 and -51.13 respectively), while neutral causes would have weights approaching zero[2] (Table 5).

We created datasets for the No-IT condition by reordering the datasets from the IT condition to remove all of the informative transitions (e.g., Table 1b). For any given transition either no causes change, or two or more change. Still, the regression weights remained the same. For example, in Tables 1a and 1b, Cause A is positive, B and C are neutral, and D is negative.

Procedure Participants were told to imagine that they worked in a nursing home, and were in charge of making sure each patient slept well. Their task was to examine the medications taken by each patient to find out which ones enhanced or interfered with sleep. Participants observed 14 sequential days of data for each patient. For each day, participants were shown which of four medications the patient took that day. Medications were represented using animated pictures of pills, which were either full-color (cause present) or visibly faded (cause absent). These presence/absence cues for each cause remained on-screen for the duration of the trial. Participants predicted whether the patient would sleep well or poorly that night, and received feedback immediately. They were then allowed to adjust sliding scales indicating their beliefs about the causal status of each medication. When they were satisfied with their current judgments, they clicked a button to proceed to the next trial, and the on-screen presence/absence cues changed to reflect the changes during that transition (pictures either faded or came back to full color for causes that changed).

The sliders used the scale -10 (strong sleep inhibitor) to 10 (strong sleep enhancer). The numerical values on the scale were given to the participants during the instruction phase. After viewing data from 14 timepoints, participants submitted final judgments about the influence of each medication using the same scale. This procedure was then repeated for five more patients, using different datasets. After completing the study, participants' bonus amount was calculated and paid. Bonus amounts were calculated based on participants' final judgments for each cause. Five cents were paid for final judgments that were sufficiently accurate, and the maximum possible bonus was $1.20.

[2] For a technical and non-obvious reason, our particular data generation process guaranteed that if all the causes were entered into a linear regression to predict the effect, positive, negative, and neutral causes always had regression weights of 1, -1, and 0 respectively; the positive and negative causes were very strong.

Results

Trial-By-Trial Updating of Causal Strength Beliefs Recall that participants had the opportunity to update their current beliefs about each drug's effectiveness after every trial. If transitions are important for causal strength learning, participants would be more likely to update causes that change. For this analysis, we focused on the first five transitions, because most of participants' learning took place toward the beginning of each scenario; the probability of updating a given cause was .27 after Trial 1 and .09 after Trial 14.

We tested whether participants were more likely to update their beliefs about causes that had just changed vs. causes that did not change using a logistic regression, which included a by-subject random intercept and a by-subject random slope for whether the cause changed. For this model, we analyzed a subset of the data that only included positive and negative causes on trials in which the effect changed. Participants in both the IT ($B = 1.10$, $SE = 0.09$, $p < .001$) and No-IT conditions ($B = 0.27$, $SE = 0.09$, $p < .01$) were significantly more likely to update their causal strength beliefs for causes that changed than causes that did not change (Table 2).

Table 2: Trial-by-Trial Probability of Updating Causal Strength Based on Whether the Cause Changed.

	Experiment 1		Experiment 2	
Cause	*IT*	*No-IT*	*IT*	*Random*
Changed	.35	.21	.38	.27
Did Not Change	.17	.18	.09	.13

Participants in the No-IT condition were still more likely to update their beliefs about causes that changed on the most recent trial, even though multiple causes changed simultaneously. This effect highlights one reason we expected people to perform better in the IT condition: participants in the No-IT condition received worse information about the relationships between the causes and the effect from transitions, but were still more likely to update their beliefs to reflect that information.

We used similar regressions to test the RW prediction that participants would be more likely to update present causes than absent ones (Table 3). Participants in the IT ($B = 2.66$, $SE = 0.26$, $p < .001$) and No-IT conditions ($B = 2.36$, $SE = 0.22$, $p < .001$) updated more often for causes that were present.

Table 3: Trial-by-Trial Probability of Updating Causal Strength Based on Whether the Cause was Present.

	Experiment 1		Experiment 2	
Cause	*IT*	*No-IT*	*IT*	*Random*
Present	.28	.27	.23	.35
Absent	.06	.04	.03	.02

Trial-By-Trial Accuracy of Predictions of the Effect Given that participants were more likely to update their beliefs about the causes that changed in the most recent trial compared to those that did not, it is clear that transitions play a role in temporal causal strength learning. Next we analyzed the differences in accuracy for predicting the effect on each trial. Recall that on each trial, participants were asked to predict whether or not the patient would sleep well. We predicted that participants in the IT condition would be more accurate than those in the No-IT condition.

We tested this using a logistic regression, predicting accuracy on a given trial using trial number, condition (IT vs No-IT), and an interaction between the two. We included a by-subject random intercept and by-subject random slope for trial number, allowing for individual differences in how accuracy changes over the course of the trials. Figure 2 summarizes our findings: participants in the IT condition improved more over time than participants in the No-IT condition (B = 0.05, SE = 0.01, p < .01). This interaction seems to come from the surprising finding that accuracy in the IT condition started out lower than the accuracy in the No-IT condition; this pattern was not found in Experiment 2, so we do not dwell on it. The overall accuracy difference between participants in the IT condition (58.5% correct) and those in the No-IT condition (59.8% correct) was nonsignificant.

Final Judgments of Causal Strength If ITs help participants learn causal strengths, then participants' final judgments of causal strength in the IT condition would be more positive for the positive causes, and more negative for the negative causes, relative to the No-IT condition. Because the data were heavily skewed, with many judgments near the positive extreme for positive causes and the negative extreme for negative causes, we used a Gamma GLM. For positive causes, we transformed the data by multiplying each judgment by -1 and adding 11. For negative causes, we added 11 to each judgment. This ensured that the shape of each distribution would not change, but that they could be mapped onto a gamma distribution, which required all values to be positive. We incorporated a by-subject random intercept to allow for individual differences in subjects' baseline judgments. Participants in the IT condition judged the positive causes to be more positive (B = 0.04, SE = 0.01, p < .01), and the negative causes to be more negative (B = -0.05, SE = 0.02, p < .001), relative to the No-IT condition (Table 4).

Table 4: Mean (SD) of Final Causal Strength Judgments.

	Experiment 1		Experiment 2	
Cause	*IT*	*No-IT*	*IT*	*Random*
Pos.	5.4 (4.5)	3.5 (5.3)	5.4 (5.1)	3.2 (5.8)
Neg.	-5.0 (5.1)	-2.7 (5.5)	-5.2 (5.8)	-3.9 (5.6)

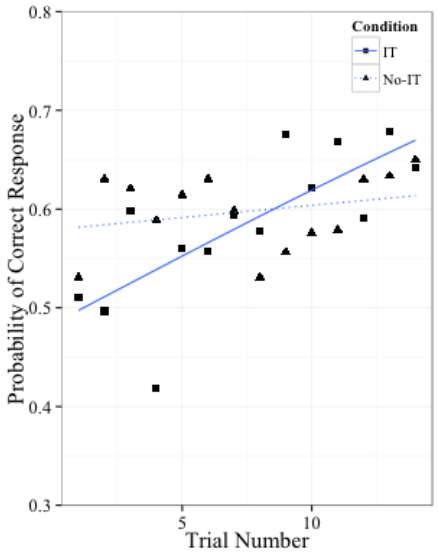


Figure 2: Accuracy of Trial-By-Trial Predictions of Effect in Experiment 1.

Discussion

Participants formed and updated their causal strength beliefs based on which causes changed during transitions. Additionally, participants' predictions of the effect improved more in the IT condition than in the No-IT condition, and participants in the IT condition made more accurate final judgments of causal strength. Together, these findings support the idea that transitions facilitate causal learning.

Interestingly, although participants in the No-IT condition were less accurate in their final judgments, they were clearly learning, and the judgments were significantly different from 0 in the correct directions for both positive and negative causes (p's < .001). This pattern of results raises the question of how participants in the No-IT condition were able to perform as well as they did. As outlined in the introduction, if participants used focal sets, it is not clear exactly what focal sets they were using. One possibility is that the bivariate correlations between each cause and the effect in our datasets were moderately strong (Table 5), so it is possible that participants in the No-IT condition used the bivariate relations and did not control for alternative causes.

Table 5: Regression weights and average bivariate correlations for causes in Experiments 1 and 2

Analysis	*Positive*	*Neutral*	*Negative*
Experiment 1			
Log. Regression Weight	51.13	< 0.001	-51.13
Bivariate Correlation	0.48	0.00	-0.47
Experiment 2			
Log. Regression Weight	51.13	< 0.001	-51.13
Bivariate Correlation	0.23	0.01	-0.22

Experiment 2

In Experiment 2, we attempted to test the limits of causal strength learning by having participants learn causal strengths in a scenario with eight causes. Having eight causes also reduces the bivariate correlations relative to Experiment 1 (Table 5), thereby making the task harder and potentially allowing for a bigger difference between the two conditions. However, the multiple regression weights were still just as strong for the positive and negative causes.

Another advantage of adding more causes in Experiment 2 is that an eight-cause scenario is even more difficult to explain in terms of focal sets. As previously discussed, as the number of causes increases, a smaller fraction of the potential data would be included in any particular focal set.

Importantly, the order of the trials does not matter for the focal set account. By contrast, if participants use the IT heuristic, they will perform better with more ITs.

Method

Participants One hundred workers were recruited through Amazon's Mechanical Turk service using the same qualification criteria from Experiment 1. Participants were paid two dollars, with accuracy bonuses as an incentive. The study lasted approximately 15 minutes.

Design and Procedure There were several differences compared to Experiment 1. Each participant made judgments about eight causes instead of four, over 25 days instead of 14. To compensate for the increased length of each scenario, participants only completed three scenarios instead of six. The IT datasets were created using a script rather than manually, and each cause changed three times. Additionally, rather than creating No-IT datasets, we randomized the order of the trials as a comparison for the IT condition. This means that sometimes there were informative transitions (when only one cause changed) in the Random condition. The modal number of ITs per dataset in the Random condition was four—the maximum number was nine—compared to 24 in the IT condition. On average, 3.36 causes changed for each transition in the Random condition.

We used three types of datasets. The first type had three positive causes, three negative causes, and two neutral causes. The second type had two positive, three negative, and three neutral causes. The third had three positive, two negative, and three neutral causes. Each participant viewed one dataset from each type, with the order of the datasets being randomized. Aside from these changes Experiment 2 was the same as Experiment 1.

Results

Trial-By-Trial Updating of Causal Strength Beliefs Experiment 2 replicated the pattern of results from Experiment 1 (Table 2). Participants updated their beliefs about causal strength more when a given cause changed compared to when it did not change. Furthermore, this occurred in both the IT condition (B = 2.03, SE = 0.14, p < .001) and the Random condition (B = 1.00, SE = 0.10, p < .001). In addition, they also were more likely to update their causal strength judgments when a cause was present in the IT (B = 3.01, SE = 0.30, p < .001) and Random conditions (B = 5.49, SE = 0.79, p < .001) (Table 3).

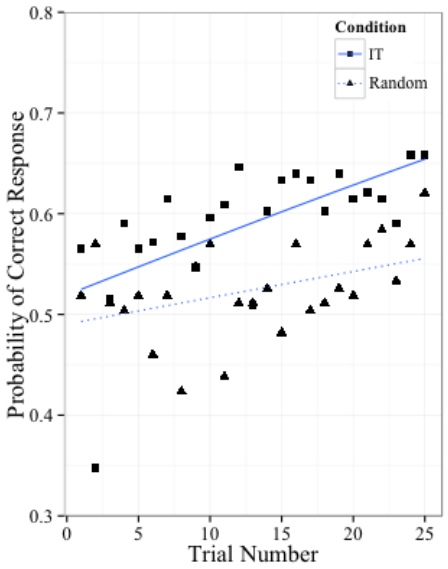


Figure 3: Accuracy of Trial-By-Trial Predictions of Effect in Experiment 2.

Trial-By-Trial Accuracy of Predictions of the Effect We hypothesized that if participants are better able to infer causal strength in the IT condition, they would also be better able to predict the effect. Figure 3 plots the probability of correctly predicting the effect.

A logistic regression with by-subject random intercepts found that participants were more accurate in the IT condition than in the Random condition (B = 0.27, SE = 0.06, p < .001). The interaction from Experiment 1 was marginal in Experiment 2 (B = -0.01, SE = .01, p = 0.06).

Final Judgments of Causal Strength As in Experiment 1, we examined the differences in participants' final judgments of positive and negative causes in the IT and Randomized conditions (Table 4). A mixed effects Gamma regression with random intercepts for each participant found stronger

causal strength judgments in the IT condition for both positive (B = 0.08, SE = 0.02, p < .001) and negative causes (B = -0.07, SE = 0.02, p < .01). Participants in the Random condition performed significantly above chance (p's < .001). Overall, the results were similar to Experiment 1.

General Discussion

Our main hypothesis was that participants learn causal strengths better in stable environments (IT condition). Indeed, participants' final causal strength estimates were stronger—closer to the normative regression weights—in these environments, suggesting that participants capitalize on the stability of the environment to learn causal strengths.

More generally, participants were more likely to update their beliefs about a cause immediately after that cause changed, even in learning environments in which more than one cause changed from one observation to the next. This finding builds upon research concerning how people learn cause-effect relations in single-cause environments; that research also found that people are more likely to update beliefs about a cause after it has changed (Soo & Rottman, 2015). The current paper generalizes this finding to situations with multiple causes and demonstrates the effectiveness of this learning habit in stable environments.

Our finding that participants update their beliefs more often after a cause changes parallels another finding: participants updated their beliefs more for present causes than absent ones. RW predicts this as well (cf. Van Hamme & Wasserman, 1994).

Several questions are as yet unanswered. Participants in the No-IT and Random conditions could infer causal direction, despite low bivariate relations in Experiment 2. Understanding this may yield other causal learning insights.

Additionally, there are some differences in the results of Experiment 1 vs. 2, particularly in the accuracy of the trial-by-trial predictions of the effect. In Experiment 2 the accuracy in the IT condition was generally better than the Random condition (Figure 3). By contrast, in Experiment 1, the accuracy in the IT condition started lower than in the No-IT condition, and rose more quickly (Figure 2). This raises the possibility that informative transitions may, in certain situations, temporarily impair learning. Aside from the main difference of 8 vs. 4 causes, there are other differences between the two experiments that could explain these different patterns: weaker bivariate relations and more learning trials in Experiment 2, and the fact that the Random condition in Experiment 2 did not exclude informative transitions. Investigating these factors can elucidate whether situations with many informative transitions may temporarily result in worse causal learning.

Individuals often face situations in which multiple causes can influence an effect. One way people may cope with this complexity is by focusing on times when a cause changes, especially in situations with fairly stable causes over time.

Acknowledgments

This research was supported by NSF 1430439.

References

Cheng, P. W., & Holyoak, K. J. (1995). Complex Adaptive Systems as Intuitive Statisticians: Causality, Contingency, and Prediction. In H. L. Roitblat & J. -A. Meyer (Eds.), *Comparative Approaches to Cognitive Science*. Cambridge, MA, US: The MIT Press.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology, 58*(4), 545-567.

Goedert, K. M., Harsch, J., & Spellman, B. A. (2005). Discounting and conditionalization: Dissociable cognitive processes in human causal inference. *Psychological Science*, *16,* 590-595.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science, 31*(5), 765-814.

Rescorla, R., & Wagner, A. R. (1972). A Theory on Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Theory and Research* (pp. 64-99). New York, NY: Appleton-Century-Crofts.

Soo, K., & Rottman, B. M**.** (2015). Elemental causal learning from transitions. In R. Dale, et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7*(6), 337-342.

Spellman, B. A., Price, C. M., & Logan, J. M. (2001). How two causes are different from one: The use of (un)conditional information in Simpson's paradox. *Memory & Cognition, 29*(2), 193-208.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation, 25*, 2. 127-151.

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology, 26*(1), 53-76.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*(1), 27-58.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*(2), 222-236.

Document Cohesion Flow: Striving towards Coherence

Scott A. Crossley (scrossley@gsu.edu)
Department of Applied Linguistics/ESL, Georgia State University
Atlanta, GA, 30303 USA

Mihai Dascalu (mihai.dascalu@cs.pub.ro)
Stefan Trausan-Matu (stefan.trausan@cs.pub.ro)
Department of Computer Science, University Politehnica of Bucharest
Bucharest, 060042 Romania

Laura K. Allen (laurakallen@asu.edu)
Danielle S. McNamara (dsmcnamara1@gmail.com)
Department of Psychology, Arizona State University
Tempe, AZ, 85287-2111 USA

Abstract

Text cohesion is an important element of discourse processing. This paper presents a new approach to modeling, quantifying, and visualizing text cohesion using automated cohesion flow indices that capture semantic links among paragraphs. Cohesion flow is calculated by applying Cohesion Network Analysis, a combination of semantic distances, Latent Semantic Analysis, and Latent Dirichlet Allocation, as well as Social Network Analysis. Experiments performed on 315 timed essays indicated that cohesion flow indices are significantly correlated with human ratings of text coherence and essay quality. Visualizations of the global cohesion indices are also included to support a more facile understanding of how cohesion flow impacts coherence in terms of semantic dependencies between paragraphs.

Keywords: Cohesion Flow; Natural Language Processing, Computational Models; Cohesion Network Analysis; Coherence; Writing Quality

Introduction

Writing is an important aspect of communication because it provides the opportunity to articulate ideas and synthesize perspectives in a persuasive manner that is often independent of time and space constraints (Crowhurst, 1990). Learning how to convey meaning competently in written texts is a crucial skill for academic and professional success. Indeed, the writing skills of college freshmen are among the best predictors of academic success (Geiser & Studley, 2001). Importantly, there are a number of attributes of high-quality writing that are directly related to the linguistic features of the written text.

One important element of writing quality is text *cohesion*. Halliday and Hasan (1976) established this term to characterize the cohesive links or ties between the sentences from the same text or fragment of text. Cohesion is a crucial element for easing the understanding of texts, particularly for challenging texts that present knowledge demands to the reader (McNamara, Kintsch, Songer, & Kintsch, 1996).

In addition, there is a general sense that essay quality is highly related to the cohesion of a text, which is reflected in the literature about writing as well as textbooks that teach students how to write (Collins, 1998). However, this assumption is compounded by differences in cohesion types (i.e. local and global cohesion). Local cohesion refers to text features that link short text segments such as sentences together, while global cohesion refers to text features that link larger segments of texts (i.e., paragraphs). Previous studies have indicated that local cohesion indices are generally negatively associated with essay quality (Crossley & McNamara, 2010, 2011) while global cohesion indices are generally positively correlated with essay quality (Crossley, Kyle, & McNamara, in press; Crossley & McNamara, 2011).

The primary difference between local and global cohesion indices is that global cohesion indices tap into text *coherence*, while local cohesion does not (Crossley et al., in press). The distinction between cohesion and coherence is that cohesion refers to the presence or absence of explicit textual cues that allow the reader to make connections between the ideas in the text. On the other hand, coherence refers to the understanding that the reader derives from the text, which may be more or less coherent depending on a number of factors, such as textual features, prior knowledge, and reading skill (McNamara et al., 1996). Thus, coherence plays a central role in terms of the sense that a text creates while addressing a deeper function of discourse (i.e., illocutionary acts; Widdowson, 1978).

The goal of this study is to further research links between global cohesion devices and both text coherence and essay quality to better understand the cognitive underpinnings of language comprehension. We do this by developing new automated indices of global cohesion (termed *cohesion flow* indices) and demonstrating how these indices can be used to predict human judgments of text coherence and essay quality. These indices are provided in the *ReaderBench* tool (RB; Dascalu, Trausan-Matu, McNamara, & Dessus, 2015).

Cohesion and Cohesion Flow

Cohesion at both the local and global levels helps to develop a text that is unified and connected as a result of individual sentences and paragraphs "hanging" together and relating to

one another (Celce-Murcia & Olshtain, 2000). In addition to text segments "hanging" together, a well structured document should have a clear and logical movement of ideas highlighted by the topic flow between sentences and paragraphs (O'Rourke, Calvo, & McNamara, 2011).

Knowing that cohesion is a central element of a text's structure, we are interested in examining how flow and cohesion might be structurally constructed in a text (i.e., derived from the proper sequencing of text elements). Thus, we operationalize *cohesion flow* as a measure of a document's structure derived from the order of different paragraphs and of the manner in which they combine to hold the text together. A text that demonstrates strong cohesion flow by linking ideas between paragraphs will likely be a more coherent text. This would allow ideas to bond together and flow smoothly from one paragraph to another, creating a text that readers can more easily comprehend. In addition, the ability to illustrate and automatically assess cohesion flow would enable researchers to observe how text segments in a document fit together and examine how the sequencing of these segments may affect readers' comprehension.

Cohesion Features and Text Quality

Previous research has placed an emphasis on the relationship between cohesive devices in text and human judgments of that text's overall quality and coherence. In terms of essay quality, McNamara, Crossley, and McCarthy (2010) reported that local cohesion indices did not demonstrate significant differences between low and high scored essays, nor did they correlate with essay scores. In a second study, Crossley, Roscoe, et al. (2011) found that two indices of global cohesion (semantic similarity between initial and middle paragraphs, and semantic similarity between initial and final paragraphs) significantly correlated with essay quality. More recently, Crossley et al. (in press) reported that while local features of cohesion (i.e., sentence overlap indices) were negatively related to essay quality, global indices of cohesion (i.e., paragraph overlap indices) were strongly predictive of essay quality. In another study, Crossley and McNamara (in press) found that modifying the cohesion structure in a text at the global level led to increased essay quality scores (as assigned by human raters). They also reported strong links between global cohesion features and the essay quality scores.

Similar findings have been reported in terms of text coherence. In three recent studies, Crossley and colleagues (Crossley et al., in press; Crossley & McNamara, 2010, 2011) examined the degree to which judgments of text coherence were predicted by automated indices of local and global cohesion reported by a number of computational tools. Crossley and McNamara (2010) reported that cohesion indices calculated at the local level demonstrated significant but negative correlations with human ratings of coherence. A follow up study (Crossley & McNamara, 2011) reported similar results for local cohesion devices. However, this study also examined a number of global cohesion indices that calculated overlap between initial, middle, and final paragraphs. These global cohesion indices were positively correlated with judgments of text coherence. Two recent studies found that while local features of cohesion were negatively related to text coherence, global indices of cohesion were strongly and positively associated with text coherence (Crossley et al., in press; Crossley & McNamara, in press) and that modifications to structure of a text in terms of global cohesion led to increased text coherence scores (Crossley & McNamara, in press). Overall, these studies provide support for the notion that expert ratings of text coherence and essay quality are best explained by global and not local cohesion devices.

Current Study

The current study develops and tests automated measures of global cohesion flow based on the Cohesion Network Analyses available within RB. We also demonstrate how visualizations based on these network analyses can be used to illustrate cohesion trends within a text. Our approach is unlike previous operationalizations of global cohesion in that prior studies (e.g., Crossley et al., in press) focused on assessing lexical and semantic overlap between adjacent paragraphs, and have not assessed the flow of cohesion throughout a text. In this study, we apply automated indices of cohesion flow to a corpus of essays written by college level students and examine how well these measures predict human scores of text coherence and overall writing quality. Thus, our goals are to operationalize new measures of cohesion flow and to use these measures to better understand how these features relate to text coherence and writing quality as judged by human raters. Such an approach gives us insight into the textual properties that drive human perceptions of coherence in language. In addition, we examine how visualizations extracted the cohesion flow indices can be used to illustrated text coherence.

Method

Quantifying cohesion flow

Measuring cohesion. In the current study, we limit the perspective of coherence to *semantic global cohesion* that captures text organization in terms of paragraph links (Lapata & Barzilay, 2005). We focus on paragraph links over sentence links because previous studies have reported stronger associations between global cohesion and text coherence and quality compared to local cohesion. We do this by evaluating cohesion flow using three different semantic models: a) semantic distances in ontologies (i.e., Wu-Palmer; Budanitsky & Hirst, 2006), b) cosine similarity in Latent Semantic Analysis (LSA) vector spaces (Landauer, Foltz, & Laham, 1998) and c) Jensen Shannen dissimilarity of Latent Dirichlet Allocation (LDA) topic distributions (Blei, Ng, & Jordan, 2003).

Building the cohesion flow graph. Our cohesion flow evaluation relies on Cohesion Network Analysis (CNA) (Dascalu et al., 2015), a computational model that uses a

cohesion-based representation of discourse. CNA combines the semantic models discussed above with Social Network Analysis (SNA) to build a multi-layered cohesion graph (Dascalu et al., 2014). Starting from the general CNA approach, a pruned graph consisting of *cohesion*[*i*, *j*] edges, where *i* and *j* are paragraph indexes, is constructed at inter-paragraph level. This cohesion flow graph uses the chronological indexing of text elements (in contrast to the initial cohesion graph that uses undirected edges) and two building criteria:

- *maximum value*: for each paragraph we determine, out of the following paragraphs, which is the closest one in terms of the selected similarity measure (maximum cohesion to following text elements).
- *above threshold*: out of all possible links between subsequent paragraphs, we select only those graph edges that exceed the imposed threshold of average + stdev of all inter-paragraph similarity measures.

Measuring cohesion flow. Based on the previous cohesion graph, a *topological sort* (Cormen, Leiserson, Rivest, & Stein, 2009) is performed in order to observe the referential sequencing of paragraphs. Afterwards, six cohesion flow indices are calculated in order to evaluate cohesion flow at the document level:

- *Absolute position accuracy*: number of paragraphs that, after performing the topological sort on the cohesion flow graph, are in the correct position (the ordered paragraph index is the same as the initial index).
- *Absolute distance accuracy*: the absolute value of the difference of ordered and initial paragraph indexes. A value closer to 0 characterizes a more cohesive text in terms of adjacency links.
- *Adjacency accuracy* determines how many paragraphs follow the idea of adjacency maximum flow: sum of absolute values of (j-i-1) where cohesion[i, j]>0.
- *Average flow cohesion* is determined as the average cohesion in our cohesion flow graph, i.e., average of all cohesion[i, j];
- *Spearman correlation* between the ordered paragraph index and the initial sequence index.
- *Max order sequence* determines how many ordered paragraph indices follow an increasing trend to determine if flow moves forward in a document (i.e., what is the longest sequence that follows an ascending trend).

All the above indices are normalized in [0; 1] by relating to the overall size of the document. In total we developed 36 indices based on the six feature categories above with each feature computed for topological sort and for LSA, LDA, and Wu-Palmer (36 features in total). Documents having fewer than three paragraphs cannot be not considered in our cohesion flow analysis.

Visualizing cohesion flow. In addition to computing automated indices of cohesion flow, we are able to visualize the process in the *ReaderBench* (RB) tool. Starting from the cohesion flow graph, we apply specific SNA metrics in order to obtain an in-depth perspective of the paragraph connectivity (i.e., the strength of paragraph associations). This flow network visualization and modeling plays an important role in understanding and interpreting the obtained results. We have selected *force-based layouts* for visual representation because they provide a comprehensive view of the social network as a planar graph. In this layout, paragraphs gravitate by having their own mass proportional to their relatedness to other paragraphs. Edges are proportional to the inverse semantic relatedness of paragraphs, while elasticity coefficients are used to obtain a more realistic visualization that minimizes edge crossings and the overall network energy. The size of each paragraph can be proportional to its betweenness centrality (i.e., the number of times it acts as a bridge along all shortest paths between pairs of two other paragraphs from the input text).

Validation Corpus

We used the corpus described in Crossley and McNamara (2011), which comprises 315 timed essays collected from undergraduate students at Mississippi State University. During the collection process, students were given 25 minutes to write an essay and no outside referencing was allowed. Such an environment better represents high stakes testing (i.e., SAT writing tests). Two SAT prompts were used and students were randomly assigned one prompt to which they responded. All students were native speakers of English and were in either Composition One or Composition Two (i.e., freshmen composition courses). Each essay was read and scored by two trained raters using both an analytic and a holistic rubric.

The rubric used to score the essays contained one analytic feature (*organization*: the body paragraphs follow the plan set up in the introduction) that evaluated semantic based, global cohesion (i.e., text coherence). A holistic grading scale based on a standardized rubric commonly used in assessing Scholastic Achievement Test (SAT) essays was also included in the rating rubric. The holistic scale and all of the rubric items had a minimum score of 1 and a maximum score of 6. Details on the rubric used can be found in Crossley et al. (in press).

Eight raters with advanced degrees in English and at least 3 years experience teaching university composition classes rated the 315 essays. The raters were informed that the distance between each score was equal. After the raters were trained, each rater then evaluated a selection of the 315 essays. In all cases, each essay was scored by at least two raters. Once final ratings were collected, differences between the raters were calculated. If the difference in ratings on a survey feature was less than 2, an average score was computed. If the difference was greater than 2, a third expert rater adjudicated the final rating. Correlations between the raters (before adjudication) were acceptable with the average correlations across all essay feature judgments and raters at $r = .72$.

Statistical Analysis

We conducted correlation and regression analyses to examine relations between our document flow indices and

human judgments of coherence and essay quality. We first conducted correlational analyses to examine associations between the indices and the human scores. Those indices that demonstrated a statistical ($p < .05$) and meaningful (at least a small effect size, $r > .1$) relation were then used in a regression analysis. Indices that were highly collinear ($r > .899$) were flagged, and the index with the strongest correlation with human scores was retained while the other indices were removed. The remaining indices were included as predictor variables in a stepwise multiple regression to predict both human scores of coherence and overall essay quality. The model from the stepwise regression was then used to assign scores for the essays in the test set using a leave-one-out-cross-validation (LOOCV).

Results

Organization Scores

Each of the 36 document flow indices demonstrated a significant correlation with the organization scores. However, after controlling for multicollinearity, only three indices remained. These three indices demonstrated medium effect sizes with human ratings of text coherence (see Table 1) and were used in the subsequent regression analysis.

Table 1: Correlations between RB cohesion flow indices and raters' organization scores

Variable	*r* value	*p* value
Adjacency accuracy (Maximum value based on LDA)	.399	< .001
Spearman correlation of flow versus initial ordering (Above mean+stdev based on LDA)	.382	< .001
Absolute distance accuracy on topological sort (Maximum value based on Aggregated score)	.381	< .001

A LOOCV linear regression analysis was conducted including the three RB indices. These three variables were regressed onto the raters' coherence evaluations for the 315 writing samples. Of these three variables, two were significant predictors in the regression: Absolute distance accuracy on topological sort (Maximum value based on Aggregated score) and Adjacency accuracy (Maximum value based on LDA). The linear regression using the two variables yielded a model that reported $r = .398$ (MAE = .779). The model accounted for 16% of the variance in the human evaluations of coherence.

Essay Quality Scores

Each of the 36 document flow indices demonstrated a significant correlation with the essay quality scores. However, after controlling for multicollinearity, only three indices remained. These three indices demonstrated medium effect sizes with human ratings of essay quality (see Table 2) and were used in the subsequent regression analysis.

Table 2: Correlations between RB cohesion flow indices and raters' scores of essay quality

Variable	r value	*p* value
Adjacency accuracy (Maximum value based on LSA)	.356	< .001
Average document flow cohesion (Above mean+stdev based on Aggregated score)	.317	< .001
Absolute distance accuracy on topological sort (Maximum value based on LDA)	.310	< .001

A LOOCV linear regression analysis was conducted including the three RB indices. These three variables were regressed onto the raters' essay score for the 315 writing samples. Of these three variables, one was a significant predictor in the regression: Adjacency accuracy (Maximum value based on LSA). The linear regression using the variable yielded a model that reported $r = .334$ (MAE = .763). The model accounted for 11% of the variance in the human evaluations of essay quality.

Discussion and Conclusion

While text coherence can be influenced by a reader's prior knowledge and/or reading skill (McNamara et al., 1996), it also depends on the features of the text. Understanding how elements of the text can increase human judgments of coherence is thus an important area of research in discourse processing and in theories of writing because it can provide us with information about the impact of linguistic features on the cognitive processes involved in text comprehension. Previous research has supported the notion that both text coherence and essay quality are associated with global, but not local cohesion features in the text. The findings from this study further this previous research and extend it by introducing new computational operationalizations of global cohesion that are available in the *ReaderBench* (RB) tool. These operationalizations, which are based on cohesion flow, provide new evidence for how text features can combine and interact to create more coherent text that leads to stronger evaluations of writing quality. The findings provide us with a more in-depth understanding about how text cohesion can lead to text coherence and the effects such cohesion has on essay quality.

The primary goal of this research was to develop and test new indices of global cohesion. The indices developed are based on cohesion flow (i.e., the hierarchical topic progression among paragraphs in a text). The purpose of the new indices was to allow for the examination of the unity and connectedness of paragraphs in a text, but not simply through assessing semantic or lexical overlap among the paragraphs. Rather, our indices examine if paragraphs are sequenced appropriately in reference to one another. Thus, the indices do not measure lexical and semantic overlap between paragraphs, but rather overlap among paragraphs.

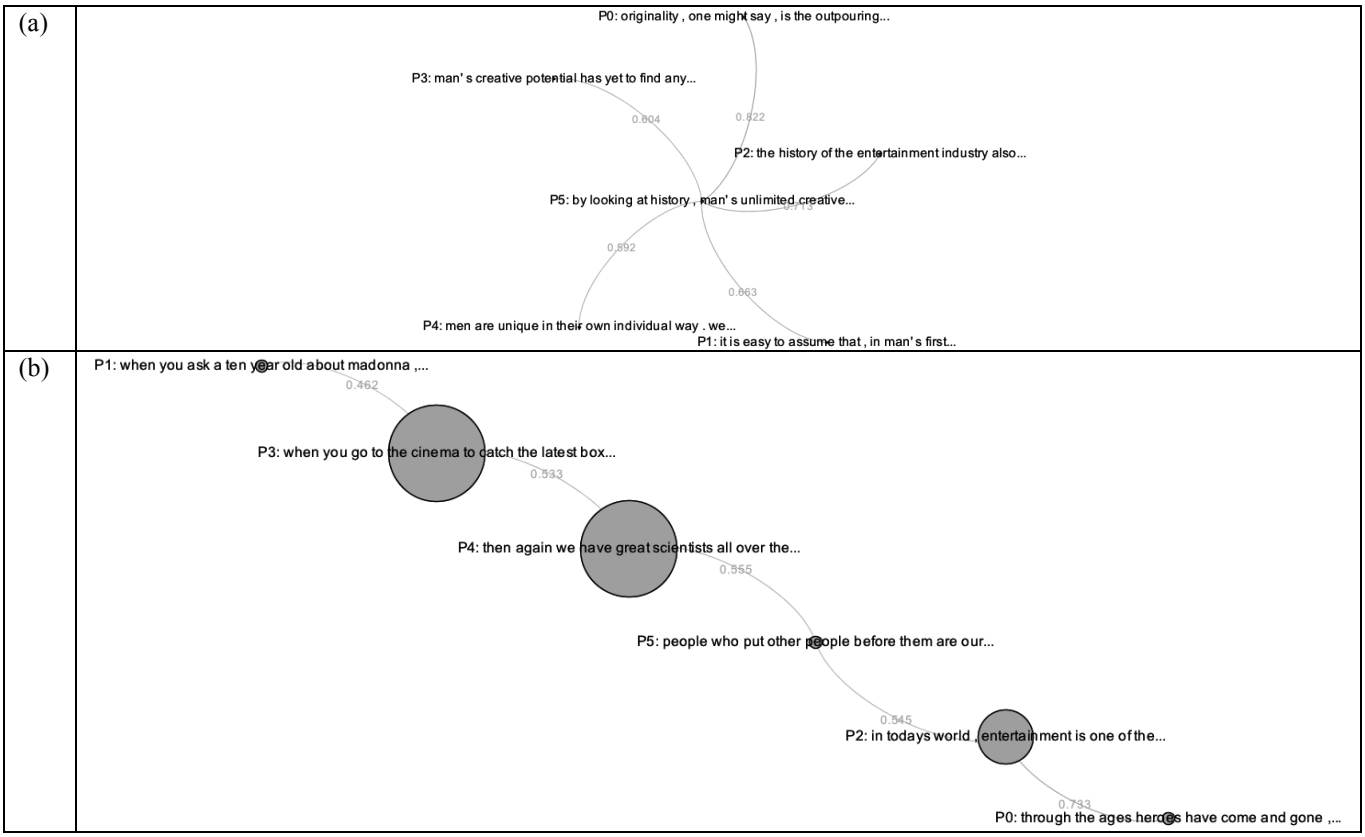


Figure 1: Differences between (a) an essay scored high in organization and (b) an essay scored low in organization.

To do this, we conducted a cohesion flow evaluation derived from a multi-layered cohesion graph that afforded chronological indexing of text elements. Using three different semantic models (LSA, LDA, and Wu-Palmer), we developed 36 indices of cohesion flow that calculated referential sequencing among paragraphs. Our hypothesis was that essays that contained strong cohesion flow among paragraphs would be judged as more coherent and of higher quality. We found this to be true in both cases, providing additional support that automated indices that measure global cohesion through cohesion flow are related to text coherence and essay quality.

We found that each of the 36 indices correlated with human ratings of text coherence and essay quality. Because these indices were measuring a very similar construct, most of the indices were multi-collinear. However, in both test sets, measures based on adjacent accuracy were the strongest predictors of human ratings. This result indicates that texts with the greatest number of paragraphs that flowed together were the most coherent and of the highest quality. These indices were followed by Spearman correlations of flow versus initial ordering, absolute distance accuracy on topological sort, and average document flow cohesion. These indices indicated that a) texts with paragraphs that showed strong correlations or absolute values between the ordered paragraph index and the initial sequence index received higher ratings, and b) texts with high average flow cohesion also received higher scores.

Beyond their predictive power in machine learning models, an important component of the cohesion flow indices is that they are able to be visualized in the RB. As an example, Figure 1 depicts the cohesion flow of two six-paragraph essay (P0…P5) using LSA as semantic similarity function and maximum value as building criteria. The essays presented in this figure were scored 5.5 (essay *a*) and 2.0 (essay *b*) on the organization rating. The figure demonstrates an SNA perspective of the cohesion flow graph in which the paragraphs are the nodes, the edges represent semantic similarity values (LSA) that meet the building criteria, and the size of each node is proportional to its betweenness centrality. A well-organized essay (*a*) summarizes all previous ideas in the conclusions; therefore, the visualized star shape links all paragraphs to the conclusions. Moreover, it is expected that all nodes have a size equal to 0 because we have only in-edges towards the last paragraphs, so no shortest paths exist between pair of nodes excluding the conclusion. In contrast, essay (*b*) highlights an essay that lacks proper sequencing in that the cohesion flow is disrupted by the ordering of paragraphs. In short, the visualization process allows the complex mathematical equations that underlie our cohesion flow indices to illustrate how cohesion flow operates within a text. Such a visualization is helpful in understanding trends in global cohesion patterns within a text and how these patterns are related to text coherence and writing quality.

In conclusion, we have developed new indices of global cohesion that are based on cohesion flow. These indices show significant correlations with human ratings of text coherence and essay quality providing evidence that the coherence of a reader's mental representation is influenced by links among the paragraphs in the text. However, these indices predict only a small amount of the variance in human ratings of text coherence (about 16%) and essay quality ratings (about 11%). These numbers are on par with previous studies examining global cohesion indices and human ratings of text quality (Crossley & McNamara, 2011; Crossley et al., in press), but do indicate that human ratings of text coherence and essay quality are only partially explained by cohesion flow among paragraphs. Future iterations of this research will combine the cohesion flow indices introduced in this study along with other global indices of cohesion and indices related to characteristics of writing such as lexical sophistication, fluency, spelling, and syntactic complexity. Such analyses are beyond the scope of the current study, but we expect them to provide a deeper understanding of how linguistic and semantic elements are collectively related to text coherence and text quality, both crucial elements of discourse processing. In addition, we will explore the degree to which the visualizations are useful to practitioners and writers in order to improve document structure. The overall aim is to provide users with data to maximize cohesion flow between adjacent paragraphs and to develop conclusions with coherent summaries.

Acknowledgment

This research was partially supported by the EC H2020 project RAGE Grant agreement No 644187, as well as by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02) and NSF (1417997 and 1418378).

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics, 32*(1), 13–47.

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and Context in Language Teaching: a Guide for Language Teachers*. New York, NY: Cambridge University Press.

Collins, J. L. (1998). *Strategies for struggling writers*. New York, NY: The Guilford Press.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (Eds.). (2009). *Introduction to Algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion. *Behavior Research Methods.*

Crossley, S. A. & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of CogSci'10* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of CogSci'11.* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (in press). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research.*

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011) Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas et al. (Eds.), *Proceedings of AIED 2011*. (pp. 438-440). New York: Springer.

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education, 15*, 155-172.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learners productions and strategies with ReaderBench. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 335–377). Switzerland: Springer.

Dascalu, M., Trausan-Matu, S., McNamara, D.S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *ijCSCL, 10*(4), 395–423. doi: 10.1007/s11412-015-9226-y

Geiser, S., & Studley, R. (2001). UC and SAT: *Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Oakland: University of California.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion In English*. London: Longman.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 259–284.

Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: models and representations. In *19th international joint conference on Artificial intelligence* (pp. 1085–1090). Edinburgh, Scotland: Morgan Kaufmann Publishers Inc.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27* (1), 57-86.

McNamara, D.S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.

O'Rourke, S., Calvo, R. A., & McNamara, D. S. (2011). Visualizing topic flow in students' essays. *Journal of Educational Technology & Society, 14,* 4-15.

Widdowson, H.G. (1978). *Teaching Language as Communication*. Oxford: Oxford University Press.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *ACL '94* (pp. 133–138). New Mexico, USA: ACL.

The Relationship Between Inhibitory Control and Free Will Beliefs in 4-to 6-Year-Old-Children

Adrienne O. Wente (adrieorchid@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Titus Ting (tripletxtreme@gmail.com)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Rosie Aboody (raboody@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Tamar Kushnir (Tk397@cornell.edu)
Department of Human Development, MVR Hall
Ithaca NY, 14850

Alison Gopnik (gopnik@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Abstract

This study explores the relationship between beliefs about self-control and the ability to exercise self-control in 4- to 6-year-old children. Sixty-eight children were asked a series of questions to gauge whether they believed that they could freely choose to act against their desires or inhibit themselves from performing desired actions. Children were also asked to provide qualitative explanations for why they could or could not exercise free will, and to complete two inhibitory control tasks: forbidden toy and day/night. Choice responses were negatively correlated with performance on the forbidden toy task, when children performed that task first. There was also a negative correlation between a belief in an internal locus of control, and success on the forbidden toy measure. Refraining from touching a forbidden toy appears to be correlated to less belief in free will. Though this may appear counter-intuitive, it is consistent with cross-cultural research.

Keywords: cognitive development; executive function; social cognition; choice; free will; inhibitory control

Background

Previous research suggests that a belief in one's ability to freely exercise self-control emerges between the ages of 4 and 6. (Kushnir, Gopnik, Chernyak, Seiver, & Wellman, 2015). During this same period, children also become better at practicing self-control (Carlson, 2010). The present study explores the relationship between self-control abilities and corresponding free will beliefs.

A concept of free will may have several different components, which is part of why defining free will can be philosophically problematic. Under one interpretation, free will is viewed as the ability to act free from external physical constraints. Other definitions specify the ability to act against external social or normative constraints, or even internal motivational constraints, such as one's own desires.

In regards to motivational constraints, Kushnir *et al.* (2015) asked 4-to 6-year-old children a series of questions about the freedom to act against or refrain from acting on one's own immediate desires. Children were asked if people can simply 'choose to not' perform a desirable action, such as to not eat a tasty cookie. They were also asked if they could 'choose to' engage in an undesirable action, such as eat a disgusting cracker. Six-year-olds endorsed the freedom to act against desires, including the ability to refrain from desired actions (i.e. to exercise self control). Four-year-olds, on the other hand, tended to believe that people's behavior had to be consistent with their desires, especially in cases that required self-control.

Kushnir *et al.* (2015) argue that this finding reflects a conceptual change in children's folk psychology or theory of mind. Two-year-olds view actions as stemming from desires (Wellman & Woolley, 1990). A belief in free will, or choice, can be viewed as an added component to this causal chain. With this added link, actions need not stem directly from desires, but can be altered through choice.

A theory theory framework suggests that children's beliefs take the form of intuitive theories that undergo continual revision as new evidence is encountered (Gopnik & Wellman 2012). Under this framework a child might initially hold a belief that actions stem from immediate desires, and thus self-control (refraining from desirable actions) is not possible. As they encounter additional evidence they may gradually come to believe that the relationship between desires and actions is mediated by choice.

What types of additional evidence might children receive? One possibility is that children are simply taught a theory of choice and free will by other people, one which includes the message that acting against or refraining from acting on one's own desires is a free choice. Another possibility is that beliefs about choice and related beliefs about self-control stem from first person experiences. Perhaps children witness themselves or other people practicing self-control, and develop a belief in free choice to explain why this is so. Related to this second possibility, young children are continually developing increasing abilities to delay gratification and practice self-control (Carlson, 2010). Endogenous changes in inhibitory control could spur the conceptual change in children's beliefs about free will.

Several lines of research support the possibility of a link between beliefs about free will and inhibitory control. First, researchers have found a positive correlation between the development of inhibitory control and theory of mind (Carlson & Moses, 2001; Oh & Lewis, 2008). As free will beliefs are a subcomponent of theory of mind beliefs, it is possible that this relationship extends to free will beliefs.

Other research suggests that holding a belief in an internal locus of control is positively correlated with self-control (Mischel, Zeiss, & Zeiss, 1974). Children who believe their actions (and thus their choices) are guided by internal processes, like desires, may do better in tasks that require self-control than children who believe their actions are externally guided (by the situation, or by others).

On the other hand, cross-cultural research suggests the opposite relationship. People from interdependent cultures hold a stronger belief in an external locus of causation than do people from independent cultures (Choi, Nisbett & Norenzayan, 1999), and a recent study found that Chinese children hold weaker beliefs about free will than U.S. children (Wente *et al.*, in press). Nevertheless, children from interdependent cultures, like China, tend to outperform U.S. children on inhibitory control tasks (Sabbagh, Xu, Carlson, Moses & Lee, 2006; Oh & Lewis, 2008). This may seem counter-intuitive from a Western perspective. It is possible, however, that children in interdependent cultures think that self–control reflects the direct influence of external social norms, rather than an internal choice to follow those norms.

To explore the relationship between self-control and free will beliefs, the current study tested 4- to 6-year-olds on the forbidden toy and day/night inhibitory control tasks, and measured their beliefs about free will. Approximately half of the children completed the free will task first, and half completed the inhibitory control tasks first. This allowed us to assess the possibility of carryover effects between tasks. To further explore the relationship between locus of control and self-control, children were asked to explain why they could or could not practice free choice. In Kushnir *et al.* (2015), most explanation types could be classified as internal or external. The present study uses a slightly modified coding scale.

Method

Four- to 6-year-old children were asked a series of questions to gauge whether they believed that people could inhibit themselves from performing desired actions, or act against their desires. To measure inhibitory control, each child also completed the forbidden toy and day/night tasks. Task order was counterbalanced so that some children completed the inhibitory control measures first and others answered the free will questions first.

Participants

Participants were 68 4- 5- and 6-year-olds (*M*=5.37, *Range*= 4.01-6.97 years, 26 males) recruited from a study participant database or local preschool. Additionally, 14 children were excluded from the study. Of these, 8 failed to complete the full study (most because they expressed fear of being alone in the testing room), 5 were excluded due to experimenter error, and 1 due to low English language proficiency. All participants were tested in Berkeley, CA. The sample was predominantly middle and upper middle class and represented the diversity of the local population.

Stimuli

Free Will Task The experimenter used white index cards to draw different activities and foods children suggested.

Day Night Task The day/night images featured in Gerstadt, Hong, & Diamond (1994) were photocopied, and replicas were created. The night card featured a white moon and stars on a black background. The day card featured a white background with a sun outlined in black. Children were tested on 8 of each type of card.

Forbidden Toy Task A marble run and several small marbles were used. The marble run was approximately two feet tall and consisted of a series of tracks the marbles could roll down when placed in the opening at the top.

Procedure

All children were tested in a small testing room located in the Institute of Human Development at the University of California Berkeley, or in an empty room at their preschool.

The order of the free will questions and inhibitory control testing as well as which inhibitory control task was presented first was counterbalanced across participants.

Free Will Task The free will task was modeled after Kushnir *et al.* (2015). All children answered 2 physically impossible control questions, 2 action (desire) questions, and 2 inhibition (desire) questions and 2 possibility control questions taken from Wente *et al.* (in press).

Action (desire) questions gauged whether children believed that they could choose to act against their desires. Children suggested an activity and food that they really

disliked. Then the experimenter drew the activity or food for the child and used the drawings to ask the experimental questions. For example, if the child indicated that they did not like raw onions, the experimenter drew a raw onion and said, "Let's pretend this is a raw onion sitting on the table in front of you. You really do not like raw onions; you really think they taste yucky. Your mom says that it is ok to eat the onion or not eat the onion. Can you choose to eat the onion, or do you have to not eat the onion because you don't like it?"

Inhibition (desire) questions gauged whether children believed that they could choose to inhibit themselves from performing a desired action. Children were asked to name a favorite food and activity. The experimenter drew the food or activity and asked the child if they believed they could choose to inhibit themselves from eating the food or engaging in the activity. For example if the child said they really enjoyed eating ice cream, the experimenter drew a picture of ice cream and said, "Lets pretend that this is ice cream sitting on the table in front of you. You really like ice cream; you really think it tastes yummy. Your mom says it's ok to eat the ice cream or not eat the ice cream. Can you choose to not eat the ice cream, or do you have to eat it because you like it?"

Impossible control questions used similar language, but instead asked children if they believed they could choose to do impossible things. They were, "If you really wanted to could you choose to {float in the air/ walk through a wall} or do you have to {come down/ walk around the wall}?"

Possibility control questions, on the other hand, asked children if they believed they could choose to do possible and desirable things. These questions were, "Can you choose to {walk into the living room/ step down off a step} or do you have to {stay in the kitchen/ stay on the step}?"

Children were also asked to provide a qualitative explanation following their choice response. For example, if the child said they could "choose to not eat the cookie" the experimenter asked, "And why can you choose to not eat the cookie?"

The order that the 'choose to' and 'have to' options were first presented was counterbalanced within and across participants. Question order was counterbalanced across participants. Control and desire questions were alternated.

Day/Night Task The day night task was modeled after Gerstadt, Hong, & Diamond (1994). Children were instructed to say "day" when the experimenter showed them the night card, and "night" when the experimenter showed them the day card. There was a training phase where children had to provide 6 correct answers. The test phase consisted of 8 of each type of card.

Forbidden Toy Task This task was modeled after the forbidden toy procedure in Carlson (2005). The experimenter introduced children to an enticing toy, a marble run, and demonstrated how it worked. The experimenter then said that s/he had forgotten something and had to leave the room to go and get it. The child was told not to touch the toy while the experimenter was gone. Children waited 5 minutes for the experimenter to return.

Results

Free Will Task

Children answered 4 question types: possible control, impossible control, inhibition desire and action desire. For each of these question types, children answered 2 questions. Response patterns did not differ across the 2 questions, and thus the data were combined for further analysis. For each question type, children received a 'choose to' score that ranged from 0-2. A 0 indicates that the child provided no 'choose to' responses, whereas a 2 indicates that they provided 2.

An initial set of independent samples t-tests were used to examine whether children's answers differed as a result of task order (free will first vs. inhibitory control first). Children were more likely to provide inaccurate 'choose to' answers to the impossible control questions when the free will task was conducted first, $t(66)= 2.54$, $p=.013$. However children's answers were well below chance for both orders of presentation: free will first, $t(32)= 3.72$, $p=.001$; inhibitory control first, $t(34)= 16.23$, $p<.001$. There was no effect of task order for the other 3 question types. There was also no effect of task order when examining just the children who passed the forbidden toy task.

Overall, children provided 'choose to' responses significantly less than chance for the impossible control questions, $t(67)=-15.93$, $p<.001$, and significantly more than chance for the possible control questions, $t(67)= 33.04$, $p<.001$. This indicates that children believed they could not choose to perform impossible actions, yet did believe that they could choose to perform unconstrained actions. It also suggests that children understood the language and structure of the questions.

A 4(Question type: Action vs. Inhibition vs. Impossible Control vs. Possible Control) x 3(Age: 4 vs. 5 vs. 6) Repeated Measures ANOVA on mean 'choose to' responses indicated a main effect of question type, $F(3, 195)= 99.49$, $p<.001$, a main effect of age, $F(2, 65)= 10.84$, $p<.001$, and a interaction between age and question type, $F(6, 195)= 5.54$, $p<.001$.

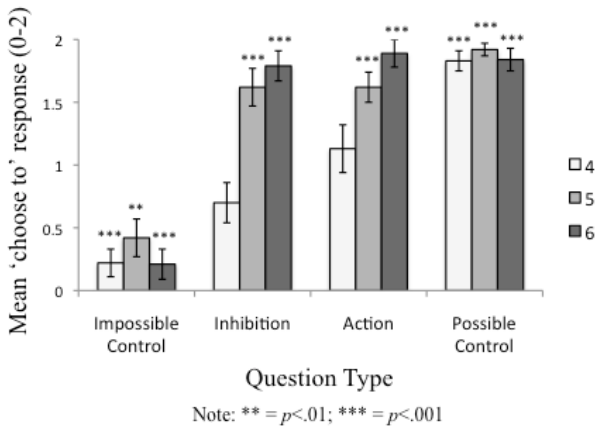
Paired sample t-tests were used to explore the main effect of question type. Children provided more 'choose to' responses for both the inhibition desire questions, $t(67)= 9.09$, $p<.001$ and the action desire questions, $t(67)= 10.09$, $p<.001$ than the impossible control questions. They also provided fewer 'choose to' responses for both the inhibition desire questions $t(67)=4.96$, $p<.001$ and action desire questions $t(67)= 3.4$, $p=.001$ than the possible control questions. There was a trending difference in responses between the action and inhibition questions, $t(67)=1.99$, $p=.051$, suggesting that children might have provided more choose to responses for the action questions.

Independent samples t-tests also revealed that 4-year-olds provided significantly fewer 'choose to' responses than 5-year-olds, $t(47)$= 3.76, p<.001. There was no difference between 5- and 6-year-olds.

To explore the age by question type interaction, four Univariate ANOVAs looked at the effect of age on 'choose to' responses for each of the 4 question types. There were no developmental differences for the impossible control and possible control questions, but there was a difference for the action questions, $F(2, 65)$= 6.38, p=.003. Independent samples t-tests indicated that 5-year-olds provided more 'choose to' responses than 4-year olds, $t(47)$= 2.17, p=.035, but that there was no difference between 5- and 6-year-olds. There was a similar finding for the inhibition questions, $F(2, 65)$= 15.5, p<.001, the difference between 4- and 5-year-olds was significant $t(47)$= 4.23, p<.001, but the difference between 5- and 6-year-olds was not.

Overall, children provided more 'choose to' responses for desire questions than for impossible control questions, and fewer 'choose to' responses for the desire questions than for possible control questions. Children's 'choose to' responses for the desire questions increased between ages 4 and 5. Means and comparisons to chance (1) are presented in Figure 1.

Figure 1: Mean 'choose to' response



Note: ** = p<.01; *** = p<.001

Qualitative Explanations

To explore children's beliefs about locus of control, children were also asked explain their answers to the choice questions. For example, if a child answered that they had to eat a cookie, the experimenter asked, "Why do you have to eat the cookie?" All answers were coded as internal, external, or other. See Table 1 for further information about explanation coding.

Each child was assigned an internal and external score. These were asymmetrical because some answers were coded as 'other.' Scores indicated the proportion of internal or external explanations given. For example, if the child provided 3 internal explanations and 1 external explanation they received an internal score of .75 and external score of .25. Overall, the mean internal score was .411, SD= .358, and mean external score was .564, SD= .364. A one-way ANOVA revealed that there were no significant developmental differences. Independent samples t-tests did reveal that children who completed the free will task first trended towards providing more internal explanations than children who completed the inhibitory control portion first $t(66)$=.1.92, p=.059. The ratio of internal to external explanations was similar when the free will questions were asked first. However, children who completed the inhibitory control tasks first provided significantly more external than internal explanations, $t(34)$= -2.77, p=.009.

Table 1: Explanation Coding

Explanation Type	Definition	Example Explanations
Internal	References to mental states or processes, or choice itself	*"I don't want to."* *"Your brain says what you want to do."* *"I can choose to go anywhere or not go anywhere."*
External	References to factors external to the child's mind	*"Because it's not fair."* *"My mommy is ok with it."* *"It's so good."* *"What if you have a tummy ache."*
Other	*"I don't know"*/ no answer/ irrelevant answer/ combination of internal and external	

Day/Night Task

All children received a day/night score ranging from 0-16. A score of 0 meant that the child provided no correct answers whereas a score of 16 meant that the child provided all correct answers. Means and standard errors are presented in Table 2.

Table 2: Day/Night Scores

	N	Mean	SE
4-year-olds	23	9.17	.87
5-year-olds	25	10.4	1.01
6-year-olds	19	13.11	1.03

An independent samples t-test indicated that children's answers did not differ based on task order. A 3(Age: 4 vs. 5 vs. 6) Univariate ANOVA on day/night score revealed that children's performance improved with age, $F(2, 64)$= 3.88, p=.026. Independent samples t-tests indicated that the difference between 4- and 6-year-olds was significant, $t(40)$= 2.93, p=.006. No other age differences reached significance.

Forbidden Toy Task

Children waited 5 minutes for the experimenter to return to the room, and received a score ranging from 0-300 based on the number of seconds they waited without touching the toy. For example, if the child touched the toy 18 seconds into the delay period, they received a score of 18. An independent samples t-test indicated that children's wait time did not differ as a result of task order. A 3(Age: 4 vs. 5 vs. 6) Univariate ANOVA did not reveal age differences. Children also received a pass/fail score. Exactly half the children, 34, passed, and half the children failed. Fisher's exact tests confirmed that age differences in the pass/fail variable were not significant.

Relationship Between Measures

Next we explored the relationship between measures. Children's answers to the inhibition and action desire questions were combined and all children received a free will score that ranged from 0-4. A higher score indicates that the child provided more choose to responses.

Table 3: Correlations between measures

	Internal	D/N	External	Forbidden
(a) Partial correlations all data combined (n=68)				
Free Will	.06	.01	-.08	-.11
Internal Score		-.07	-.89*	-.31*
Day/ Night			.12	.01
External				.25*
(b)Partial correlations for free will first (n=33)				
Free Will	.03	-.17	-.03	.21
Internal Score		-.3	-1*	-.35
Day/ Night			.3	.00
External				.35
(c) Partial correlations for inhibitory control first (n=35)				
Free Will	.14	.14	-.19	-.43*
Internal		.27	-.73*	-.15
Day/ Night			-.14	-.02
External				.06

Partial correlations controlling for age were calculated using the following variables: free will score, internal score, external score, day/night score and forbidden toy wait time. Results indicate that internal score was negatively correlated with forbidden toy wait time, $r(64)=-.31$, $p=.012$, whereas external score was positively correlated, $r(64)= .25$, $p=.042$. Running a similar analysis using the pass/not pass variable confirmed this finding. A One-Way ANOVA on mean external score revealed that children who passed the forbidden toy task provided more external explanations than children who failed, $F(1, 67)= 5.37$, $p=.024$.

Splitting data by task order again revealed that children who provided external explanations were more successful at the forbidden toy task, but only when they answered the free will questions first. There was a trending negative correlation between internal score and forbidden toy wait time, $r(29)=-.35$, $p=.051$, and a trending positive correlation between external score and forbidden toy wait time, $r(29)=.35$, $p=.051$. Again, a One-Way-ANOVA confirmed that children who passed provided significantly more external explanations if they completed the free will task first, $F(1, 32)= 5.195$, $p=.03$.

For children who completed the inhibition tasks first, there was a negative correlation between children's free will score and wait time during the forbidden toy task, $r(32)=-.433$, $p=.01$. A Univariate ANOVA on free will score, controlling for age, and using the pass/not pass variable confirmed this finding. When the inhibitory control tasks were completed first, children who passed the forbidden toy task provided significantly fewer 'choose to' responses than children who failed, $F(1, 32)= 7.133$, $p=.012$.

Discussion

Children's answers to the free will questions were consistent with results of previous studies (Kushnir *et al.*, 2015). Older children ascribed more choice than younger children for the desire questions, and they trended towards ascribing more choice in cases of action than inhibition. There were no developmental differences for the control questions. There were developmental differences on the day/night task. Older children scored higher than younger children. Age did not predict performance on the forbidden toy task. Half the children passed this task, and half touched the toy.

Neither the forbidden toy task nor the day/ night task was positively correlated with free will beliefs. Previous research found a positive relationship between the development of theory of mind and executive functioning. In particular, studies have found a correlation between cool executive functioning tasks, such as the day night task, and theory of mind (Carlson & Moses, 2001, Sabbagh *et al* 2006). The present results do not suggest that free will beliefs are related to executive functioning in the same way that theory of mind beliefs are, although they do not necessarily rule out this possibility.

Both qualitative explanations and choose to responses were related to the forbidden toy task. Children who answered the free will questions first, and provided external responses, were more likely to later pass the forbidden toy task. Also, children provided more external explanations if they answered the free will questions after completing the inhibitory control measures.

At first, these results seem at odds with previous research showing that holding a belief in an internal locus of control enhances self-control (Mischel, Zeiss & Zeiss, 1974). However, this prior work assumes a strong internal motive; the child decides if they want persist and work towards a reward. The forbidden toy task, on the other hand, has a strong external motive – to obey the experimenter's rule not

to touch the toy. No reward is given for those who pass, other than the reward of following the rule.

Analysis also revealed a negative relationship between self-control abilities and beliefs about free will. Of those who completed the inhibition tasks first, children who preformed better on the forbidden toy measure provided fewer 'choose to' responses to the free will questions than those who failed.

Ego depletion (Baumeister, Bratslavsky, Muraven, & Tice, 1998) could explain the negative correlation found in this study. Children may have depleted their willpower practicing self-control, and then subsequently answered that they could not further practice self-control. However, there were no order effects for the desire free will questions. The experience of control itself did not make children less likely to endorse free will, suggesting that ego depletion cannot explain these results. Instead, children who passed the task were less likely to endorse free will than those who did not.

These results are consistent with cross-cultural results where inhibitory control is correlated with an external locus of control and lesser belief in free will. One explanation is that people perceive acts of self-control differently. There is evidence that some actions, particularly actions based on external moral norms, may not be viewed as freely chosen. Furthermore, the extent that they are tends to differ across cultures (Chernyak & Kushnir, 2014; Chernyak et al, 2013). Some children may have not touched the toy because they did not believe that they had the choice to touch the toy. In order to fail, on the other hand, children must first conceive of the choice to touch the toy. Their subjective experience during the forbidden toy task may have subsequently influenced answers during the free will task. In this sense, the subjective first person experience of practicing self-control could, over time, shape beliefs about free will.

These findings suggest a number of ways that the practice of self-control may relate to beliefs about self-control. First, we found that children who provided external explanations performed better on a forbidden toy self-control task. However, this relationship is in the opposite direction than previous findings. Second we found that children provided more external explanations after completing the inhibitory control tasks. Finally, we found a relationship between the experience of self-control and free will beliefs, but this relation was that children who exercised self-control were less likely to endorse free will. Future research should consider how beliefs about free will relate to both internally and externally motivated acts of self-control, and how culturally variable experiences influence children's abilities and beliefs.

Acknowledgments

The authors would like to thank the John Templeton Foundation for funding this research, as well as Sophie Bridgers and Elyanah Posner for help with data collection.

References

Baumeister, R., Bratslavsky, E., Muraven, M., & Tice, D. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74,* 1252-1265.

Carlson, S. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental neuropsychology, 28*, 595-616.

Carlson, S. (2010). Developmentally sensitive measures of executive functioning in preschool children. *Developmental Neuropsychology, 28,* 595-616.

Carlson, S. & Moses, L. (2001). Individual differences in inhibitory control and children's theory of *mind. Child Development,* 1032-1053.

Chernyak, N., Kushnir, T., (2014). The self as a moral agent: Preschoolers behave morally, but believe in the freedom to do otherwise. Journal of Cognition and *Development, 0,* 1-12.

Chernyak, N., Kushnir, T., Sullivan, K.M., & Wang, Q. (2013). A comparison of Nepalese and American children's concepts of freedom of choice and social constraint. *Cognitive Science, 37*, 1343-1355.

Choi, Nisbett & Norenzayan (1999). Causal attribution across cultures: Variation and Universality. *Psychological Bulletin, 125*, 47-63.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3.5-7 years old on a Stroop-like day-night test. *Cognition, 53*, 129- 153.

Gopnik, A., & Wellman, H. M., (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085-1108.

Kushnir, T., Gopnik, A., Chernyak, N., Seiver, E., & Wellman, H. M. (2015). Developing intuitions about free will between ages four and six. *Cognition, 138,* 79-101.

Mischel, W., Zeiss, R., & Zeiss, A. (1974). Internal-external control and persistence: Validation and implications of the Stanford preschool internal-external scale. *Journal of Personality and Social Psychology, 29,* 265-278.

Oh, S., & Lewis, C. (2008). Korean preschoolers' advanced inhibitory control and its relation to other executive skills and mental state understanding. *Child Development, 79*, 80-99.

Sabbagh, M. A., Xu, F. Carlson, S. M., Moses, L. J. & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U.S. preschoolers. *Psychological Science, 17*, 74-81.

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition, 35*, 245-275.

Wente, A., Bridgers, S., Zhao, X., Seiver, E., Zhu, L., & Gopnik, A. (In press). How universal are free will beliefs? Cultural differences in Chinese and U.S. 4- and 6- year-olds. *Child Development.*

Explanatory Biases in Social Categorization

Samuel G. B. Johnson[1], Haylie Shestle Kim[2], & Frank C. Keil[1]
(samuel.johnson@yale.edu, hayliekim@berkeley.edu, frank.keil@yale.edu)
[1]Department of Psychology, Yale University, New Haven, CT 06520 USA
[2]Department of Linguistics, University of California, Berkeley, CA 94720 USA

Abstract

Stereotypes are important simplifying assumptions we use for navigating the social world, associating traits with social categories. These beliefs can be used to infer an individual's likely social category from observed traits (a *diagnostic* inference) or to make inferences about an individual's unknown traits based on their putative social category (a *predictive* inference). We argue that these inferences rely on the same *explanatory logic* as other sorts of diagnostic and predictive reasoning tasks, such as causal explanation. Supporting this conclusion, we demonstrate that stereotype use involves four of the same biases known to be used in causal explanation: A bias against categories making unverified predictions (Exp. 1), a bias toward simple categories (Exp. 2), an asymmetry between confirmed and disconfirmed predictions of potential categories (Exp. 3), and a tendency to treat uncertain categorizations as certainly true or false (Exp. 4).

Keywords: Social categorization; inductive reasoning; stereotyping; explanation; causal reasoning.

Introduction

Stereotypes help us to navigate the social world. Like other categories, social categories allow us to use a subset of an individual's properties to predict that individual's other properties (Murphy, 2002). If a cat has stripes, it may be a tiger and therefore aggressive. If a person has a law degree, she may be a lawyer—and therefore litigious.

However, stereotypes often underdetermine what inferences to draw, even in the absence of individuating information. Angie, for instance, is both a woman and a Texan. What should we predict about her pool hustling skill? Our stereotype of Texans suggests she should do well, whereas our stereotype of women suggests she should do poorly. That is, social categories are *cross-classified* (e.g., Ross & Murphy, 1999): Individuals can belong to multiple categories simultaneously, so the outcome of stereotype-based inference can be ambiguous.

A related complication is that it is often unclear what social categories individuals belong to. This causes ambiguities in using stereotypes both in diagnostic (trait-to-category) and predictive (category-to-trait) reasoning. For instance, Sarah is socially awkward and fascinated by sci-fi novels. While these traits fit well with our stereotype of engineers, it is also possible that she is (for example) a lawyer. That is, the diagnostic inference about Sarah's occupation is uncertain. This uncertainty also propagates to predictive inferences about her other traits—she is more likely to be good at math if she is an engineer and more likely to sue you if she is a lawyer.

A great deal is known about *when* stereotypes are applied to individuals (e.g., Hilton & von Hippel, 1996). However, less is known about *how* people use stereotypes to make inferences about individuals when multiple social categorizations are plausible—a problem exacerbated by the prolific cross-classification of human beings.

We propose that stereotype-based inferences follow *explanatory logic*—a set of heuristics used to perform diagnostic and predictive inferences across various types of psychological processes (e.g., Johnson, Rajeev-Kumar, & Keil, 2015; Johnson, Jin, & Keil, 2014). Put differently, social categories are not merely statistically associated with stereotypical traits; instead, people seek out the category that best *explains* an individual's traits, and do so using a specific set of domain-general heuristics. This proposal is a specific version of the broad theoretical orientation that emphasizes attributional processes (e.g., Hamilton & Sherman, 1996; Pettigrew, 1979), explanatory theories (McGarty, Yzerbyt, & Spears, 2002; Wittenbrink, Hilton, & Gist, 1998), and intuitive schemas (Fiske, 1993; Hilton & von Hippel, 1996) in stereotyping. (See also Tversky & Kahneman, 1983 for related evidence of heuristic processing in stereotype use.)

Here, we look for signature biases of explanatory logic in stereotype-based social categorization. Experiments 1–3 look at diagnostic reasoning biases (the role of *inferred evidence*, a *simplicity heuristic*, and an *asymmetry in positive vs. negative evidence*). Experiment 4 looks at a predictive reasoning bias (*digitizing* uncertain beliefs).

Experiment 1: Inferred Evidence

Suppose you have been recruited by your friend to help determine the identity of a cocktail she has been served, which is spicy, orange-flavored, and hazel-colored. Is it mixed with triple sec (orange-flavored and clear) or Grand Marnier (orange-flavored and hazel-colored)? The hazel color of the orange-flavored liqueur would be the deciding factor. Unfortunately, the spicy flavor means that the drink contains spiced rum—which is also hazel-colored. In the absence of further evidence, there is just no way to tell which is the right liqueur.

However, people do not settle for ignorance in such cases—they try to *infer* what color the liqueur was, and use erroneous cues to do so, such as the base rate of liqueurs that are clear versus hazel-colored. They may conclude that most liqueurs are clear, therefore this particular liqueur is likely clear, therefore triple sec is the probable culprit (Johnson, Rajeev-Kumar, & Keil, 2015). This logic is incorrect because it relies on a reference class that is too broad: People ought to consider only the

drinks made with either triple sec or Grand Marnier (of which 50% would be made with hazel-colored liqueurs); instead, they consider *all* drinks (of which most are made with clear liqueurs). This bias typically leads people to infer explanations that make fewer unverified predictions (Khemlani, Sussman, & Oppenheimer, 2011).

Such reasoning extends far beyond the bar. It appears in causal explanation (Khemlani et al., 2011), classification (Sussman, Khemlani, & Oppenheimer, 2014), and causal strength judgments (Johnson, Johnston, Toig, & Keil, 2014), and emerges early in childhood (Johnston, Johnson, Koven, & Keil, 2015). Thus, inferred evidence appears to be used across diverse explanatory tasks.

Do people also use this strategy when using stereotypes to reason about individuals? For instance, if you meet someone highly educated, is she likelier to be an engineer or a lawyer? If you focus on her math skills (unclear from the conversation), you may conclude she is likely a lawyer. But if you focus on her argumentativeness (also unclear), you may conclude she is likely an engineer.

To test this bias, participants were told about either the personality traits (Exp. 1A) or physical traits (Exp. 1B) of an individual from an unfamiliar culture. They were told that the individual had one trait (e.g., hard-working) but that it was unknown whether the individual had another trait (humorous). Participants were told about stereotypes for two groups, one of which was associated with both the known and unknown trait (a religion whose members are known to be hard-working and humorous) and the other associated with only the known trait (an occupation category whose members are known to be hard-working). If people infer that unknown traits are likely to be absent (based on low base rates), then they should categorize her (non-normatively) into the one-trait group.

Method

We recruited 200 participants from Amazon Mechanical Turk for Experiment 1A (N = 100) and Experiment 1B (N = 100); 18 were excluded from analysis due to poor performance on check questions (see below).

Participants were instructed to:

> Suppose you are visiting a foreign country called Gazda. The people of Gazda belong to many different kinds of groups, including different religions, occupations, and ethnicities.
>
> During your visit, you hear about some citizens of Gazda, but you aren't sure which groups they belong to. However, you do have some information from your friend, a native Gazdan. Your task will be to try to figure out which groups these citizens belong to.

Next, participants completed three items, concerning inferences based on either personality (Exp. 1A) or physical (Exp. 1B) stereotypes. For example:

> You've heard that Taylor is hard-working, but no one has told you whether or not Taylor is humorous.
>
> There are two groups of people who are known to be hard-working:
>
> About 1 in 10 people belong to the religion of Ghalism, and they have a reputation for being hard-working and humorous.
>
> About 1 in 10 people have the occupation of Chener, and they have a reputation for being hard-working.

Participants then categorized the individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Chener occupation") scale. The category order was randomized.

The personality traits (Exp. 1A) were either positive (hard-working, humorous), negative (dishonest, arrogant), or neutral (traditional, emotional). The physical traits (Exp. 1B) were not valenced (brown bracelets, blue shoes, tall hats, black clothes, left ear piercings, white bottomed shirts). Items were presented in a random order.

At the end of each study, a set of check questions were asked. Participants incorrectly answering 30% or more of these questions were excluded from analysis.

Results and Discussion

Measures were recoded so that negative scores (-5 to 0) correspond to the one-trait category and positive scores (0 to 5) correspond to the two-trait category.

Participants in Exp. 1A showed a significant bias toward the one-trait categories [M = -0.91, SD = 1.28; $t(93) = 6.92$, $p < .001$, $d = 0.71$, $BF_{10} > 1000$], which was consistent across personality traits that were positive [M = -0.97, SD = 1.72], neutral [M = -0.81, SD = 1.71], and negative [M = -0.95, SD = 1.65]. Participants in Exp. 1B showed a similar bias for physical traits [M = -0.84, SD = 1.63; $t(88) = 4.84$, $p < .001$, $d = 0.52$, $BF_{10} > 1000$]. The effect size did not differ across experiments [$t(180) = 0.46$, $p = .64$, $d = 0.05$, $BF_{01} = 7.8$].

These results demonstrate a non-normative bias in social categorization, stemming from a more general explanatory heuristic. Social stereotypes are rich with associated attributes (Andersen & Klatzky, 1987) so that some potentially diagnostic traits are sure to be unknown at a given time for an individual. Thus, future work might test whether inferences about natural social categories depend (erroneously) on which unknown traits are salient.

Experiment 2: Simplicity

Your friend is on her second drink, which is nutty and strong. It could be amaretto (both nutty and strong), or it could be Frangelico (nutty but not strong) mixed with rum (strong but not nutty). You probably find the amaretto explanation more plausible, because it is simpler: Before knowing anything about the taste, it is more likely. You would not be alone: For deterministic causal systems, people are biased toward simple explanations, because people use simplicity as a heuristic for estimating prior probabilities (Lombrozo, 2007). That is, people assume that simple explanations are more often true than complex explanations, even overriding the actual base rates.

However, while complex explanations typically have lower prior probabilities, they are often better fits to the

data. Keep in mind that rum, Frangelico, and amaretto do not *always* taste nutty or strong. Thus, we must estimate not only the prior probability of each explanation, but also how likely the actual taste is, given each explanation. A combination of multiple ingredients may be more likely to lead to a strong and nutty taste than one ingredient alone.

Hence, the *complexity* of an explanation can be used to estimate its fit to the data, even as its *simplicity* is used to estimate its prior probability (Johnson, Jin, & Keil, 2014). One piece of evidence for this *opponent heuristic* model is that the simplicity bias is stronger for deterministic than for stochastic causal systems. The goodness-of-fit is perfect for deterministic causal systems, so complexity does not lead to an explanatory benefit; but the goodness-of-fit is imperfect for stochastic causal systems, affording a benefit to the better-fitting complex explanation.

Some stereotypes are relatively homogeneous (e.g., race and skin color), whereas others are far more heterogeneous (e.g., race and driving habits), in parallel to the distinction between deterministic and stochastic causal systems (e.g., Park & Hastie, 1987). Would we also find a difference in stereotype-based categorization between homogeneous and heterogeneous stereotypes?

Method

We recruited 100 participants from Amazon Mechanical Turk; 24 were excluded from analysis.

Participants read about four sets of traits. For two, the categories were *heterogeneous* relative to the traits:

> You've heard that Taylor is greedy and impatient.
> There are three groups of people who are known to be greedy and/or impatient:
> Most (50% of the people) who believe in the religion of Ghalism have a reputation for being greedy and impatient.
> Most (70% of the people) who have the ethnicity of Folian have a reputation for being greedy.
> Most (70% of the people) who have the occupation of Chener have a reputation for being impatient.

Given this information, the conditional probability of the traits is equal given the simple category (Ghalism) and complex category (Folian and Chener). For the other two items, the categories were *homogeneous* relative to the traits (e.g., "All (100% of the people) who…").

Participants then categorized the individual ("Which of the following groups do you think Taylor is more likely to belong to?") on a 0 ("Ghalism religion") to 10 ("Folian ethnicity and Chener occupation") scale. The category order and content was randomized. Items were presented in a random order, and counterbalanced with condition.

Results and Discussion

Measures were recoded so that negative scores (-5 to 0) reflect simple categorizations and positive scores (0 to 5) reflect complex categorizations.

When the categories were homogeneous, participants were biased to categorize the individuals into simple categories [M = -0.96, SD = 2.59; $t(75)$ = -3.23, $p < .001$, d = -0.37, BF_{10} = 11.3], consistent with previous work using deterministic causal systems (Lombrozo, 2007). But when the traits were heterogeneous, participants had no significant bias [M = 0.41, SD = 2.31; $t(75)$ = 1.55, p = .13, d = 0.18, BF_{01} = 3.4], leading to a difference between conditions [$t(75)$ = 3.66, $p < .001$, d = 0.53, BF_{10} = 39.9].

Although this experiment used personality traits, one might expect that there are generalized expectations about the homogeneity of personality versus physical traits. Indeed, in follow-up experiments we have shown that people prefer simpler categorizations based on physical traits (which are more homogeneous) compared to personality traits (which are more heterogeneous).

These findings have potential implications for intergroup bias, in light of work documenting greater perceived homogeneity in one's outgroup relative to one's ingroup (Park, Ryan, & Judd, 1992). For instance, one may prefer to explain the behavior of an outgroup member in terms of only their outgroup category, whereas people may be more willing to explain ingroup members' behavior using sets of overlapping categories.

Experiment 3: Positive vs. Negative Evidence

It's time for your friend's third drink, and you are assessing whether it is a martini (which is dry) or a G&T (sparkling and dry). If she reports that it is both sparkling and dry, it is probably a G&T, because that explanation can account for both pieces of evidence, whereas a martini could account for only one piece of evidence—the G&T explanation makes more *confirmed* predictions, or has more *positive* evidence in its favor. Conversely, suppose the drink is dry but *not* sparkling. Then, it is probably a martini, because the G&T explanation makes the *disconfirmed* prediction that it is sparkling—the G&T explanation has more *negative* evidence against it.

In general, negative evidence is weighed more heavily than positive evidence (Johnson, Merchant, & Keil, 2015a) in a way that appears to be non-normative. That is, in the case where the drink is dry and sparkling, a G&T will be a somewhat better explanation than a martini (because G&T has more *positive* evidence); but in the case where the drink is dry but *not* sparkling, a martini will be a *much* better explanation than a G&T (because G&T has more *negative* evidence).

Exp. 3 tests whether this bias persists when judging the probability of social categories based on stereotypes, by measuring the relative weight placed on positive and negative evidence. We also sought to replicate Exp. 1 by measuring the weight placed on unknown evidence.

Method

We recruited 196 participants from Amazon Mechanical Turk for Experiment 3A (N = 99) and Experiment 3B (N = 97); 10 were excluded from analysis.

Participants completed four items, concerning religious, ethnicity, occupation, and social class categories. The

stereotyped traits were either personality traits (Exp. 3A) or physical traits (Exp. 3B). Items were assigned to the *Y*, *YY*, *YN*, and *YD* conditions, using a Latin square.

For the *Y* item, participants were given one piece of positive evidence for a categorization:

You've heard that Jamie is traditional.

People who have the occupation of Chener have a reputation for being traditional.

For the *YY* item, an additional piece of *positive* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional and humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

For the *YN* item, an additional piece of *negative* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional but that Jamie is not humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

For the *YD* item, an additional piece of *unknown* evidence was included, relative to the *Y* item:

You've heard that Jamie is traditional but no one has ever told you whether or not Jamie is humorous.

People who have the occupation of Chener have a reputation for being traditional and humorous.

After each item, participants were asked to rate the probability that the individual belonged to the category ("How likely do you think it is Jamie belongs to the occupation of Chener?") on a 0–10 scale.

Results and Discussion

Positive evidence scores were calculated by subtracting ratings of Y from ratings of YY, *negative evidence* scores were calculated by subtracting ratings of YN from ratings of Y, and *unknown evidence* scores were calculated by subtracting ratings of YD from ratings of Y.

For personality traits in Exp. 3A, the effect of negative evidence [$M = 2.82$, $SD = 2.96$; $t(82) = 8.69$, $p < .001$, $d = 1.53$, $BF_{10} > 1000$] was larger than the effect of positive evidence [$M = 0.63$, $SD = 1.60$; $t(82) = 3.57$, $p = .001$, $d = 0.44$, $BF_{10} = 30.3$], leading to a significant difference [$t(82) = 5.30$, $p < .001$, $d = 1.40$, $BF_{10} > 1000$]. Similarly, for physical traits in Exp. 3B, the effect of negative evidence [$M = 3.00$, $SD = 2.35$; $t(85) = 11.87$, $p < .001$, $d = 1.61$, $BF_{10} > 1000$] was larger than the effect of positive evidence [$M = 1.07$, $SD = 2.01$; $t(85) = 4.91$, $p < .001$, $d = 0.61$, $BF_{10} > 1000$], leading to a significant difference [$t(85) = 4.89$, $p < .001$, $d = 0.97$, $BF_{10} > 1000$].

We anticipated significant effects of latent evidence as well, given the bias in Exp. 1. Confirming this prediction, there was a significant detrimental effect of adding unknown evidence both for personality traits [$M = 1.36$, $SD = 1.61$; $t(82) = 7.70$, $p < .001$, $d = 0.74$, $BF_{10} > 1000$] and for physical traits [$M = 1.16$, $SD = 1.98$; $t(85) = 5.45$, $p < .001$, $d = 0.62$, $BF_{10} > 1000$].

These results are consistent both with Exp. 1 and with work on causal explanation and category-based induction (e.g., Johnson, Merchant, & Keil, 2015a). This reaffirms the idea that social categorization relies on the same heuristics as other diagnostic reasoning processes. That said, negative evidence often prompts social perceivers to *subtype* an individual, or create a new subordinate category to accommodate disconfirmations of a stereotype (e.g., Brewer, Dull, & Lui, 1981). Future work might address the interaction of these two inference strategies.

Experiment 4: Belief Digitization

Your friend is *pretty* sure her fourth drink is a Long Island Iced Tea—a sure harbinger of a hangover. But she's not *completely* sure—it could instead be a rum and coke. Suppose the two of you want to engage in the morbid exercise of calculating the probability of a hangover. Let's say she is 70% sure the drink is a Long Island Iced Tea (in which case there is an 80% chance of a hangover) and a 30% chance it is a rum and coke (with a 20% chance of a hangover). If we do the math, there is a 62% chance of a hangover (70% * 80% + 30% * 20%).

However, even when fully sober, people do not reason in this normative way, using *graded* probabilities that respect the fact that it is uncertain which drink she has. Instead, people tend to *digitize* (Johnson, Merchant, & Keil, 2015b; Murphy & Ross, 1994), treating explanations as though they are certainly true or certainly false (i.e., a 100% chance of a Long Island Iced Tea). In that case, the probability of a hangover seems to be much higher (80%).

Exp. 4 tested whether people would similarly use social categories in a 'digital' manner to make predictions about a target individual: If one believes that a person probably, but not definitely, belongs to some social category, does one then behave as though that categorization is *certain*?

Method

We recruited 198 participants from Amazon Mechanical Turk for Experiment 4A ($N = 100$) and Experiment 4B ($N = 98$); 38 were excluded due to poor performance on check questions and 51 due to ratings of P(*A*) and P(*B*,*C*) that did not sum to 100%. Analyses including the latter participants lead to the same conclusions.

Participants completed three items, each including information about the traits associated with three categories (similar to Exp. 2):

People who believe in the religion of Ghalism have a reputation for being business-minded and liberal.

People who have the ethnicity of Folian have a reputation for being business-minded.

People who have the occupation of Keader have a reputation for being liberal.

Then, participants were told that two potential categorizations make different predictions about another trait. For one item, in the *high/low* condition, the simple (hence, likelier) categorization has a high probability of that trait, while the complex (less likely) categorization has a low probability of that trait:

When people believe in the religion of Ghalism, they

are usually formal.

When people have the ethnicity of Folian and the occupation of Chener, they are occasionally formal.

In the *low/low* condition, both categorizations have a low probability of the trait:

When people believe in the religion of Ghalism, they are occasionally formal.

When people have the ethnicity of Folian and the occupation of Chener, they are occasionally formal.

In the *low/high* condition, the simple categorization has a low probability of the trait, while the complex categorization has a high probability:

When people believe in the religion of Ghalism, they are occasionally formal.

When people have the ethnicity of Folian and the occupation of Chener, they are usually formal.

Participants were told that an individual had both of the relevant properties for making a categorization ("You've heard that Taylor is business-minded and liberal").

Following this information, participants completed a *diagnosis* question and a *prediction* question, appearing on separate pages. For the diagnosis question, participants rated the probability of the simple categorization ("Taylor believes in the religion of Ghalism") and complex categorization ("Taylor has the ethnicity of Folian and the occupation of Chener"), and asked to ensure their probabilities added up to 100%. For the prediction question, participants estimated the probability of the additional trait ("What do you think is the probability that Taylor is formal?") on a 0 to 100 scale.

Results and Discussion

For the diagnosis questions, participants in both experiments favored the simple over the complex categorizations. When the categorizations were based on personality traits in Exp. 4A, participants judged the simple category [M = 61.5, SD = 14.8] more likely than the complex category [M = 38.5, SD = 14.9]. Likewise, when the categorizations were based on physical traits in Exp. 4B, participants again judged the simple category [M = 64.0, SD = 18.0] more likely than the complex category [M = 36.0, SD = 18.0]. These simplicity preferences are consistent with Exp. 2 and with previous work on simplicity preferences (Lombrozo, 2007). Importantly, however, the probability of the complex category was non-negligible—38.5% in Exp. 4A and 36.0% in Exp. 4B. These would be quite large probabilities to ignore.

Nonetheless, responses to the prediction questions revealed that participants used only the conditional probability of traits given the simple (high-probability) category, ignoring the possibility that the complex (low-probability) category was correct. Participants in Exp. 4A rated the probability of the additional trait (e.g., formality) higher in the high/low than in the low/low condition [M = 65.4, SD = 20.6 vs. M = 57.1, SD = 24.2; $t(48) = 2.08$, p =.043, d = 0.41, BF_{01} = 1.2]. That is, people used the feature likelihoods given the high-probability category when making predictions about that feature, since manipulating that likelihood (high/low vs. low/low) influenced predictions. However, participants did not use the likelihoods given the low-probability category: The low/high and low/low conditions did not differ [M = 56.6, SD = 23.6 vs. M = 57.1, SD = 24.2; $t(48) = -0.15$, p = .88, d = -0.02, BF_{01} = 8.8]. That is, manipulating the feature likelihood given the low-probability category (low/high vs. low/low) did not influence predictions. Thus, people tacitly 'digitize' high-probability categorizations, treating them as certainly true when making predictions.

This pattern was replicated in Exp. 4B, with physical traits. Again, participants distinguished between the high/low and low/low conditions when making feature inferences [M = 68.2, SD = 20.9 vs. M = 58.8, SD = 27.0; $t(59) = 2.74$, p = .008, d = 0.46, BF_{10} = 3.3], indicating that they used the feature likelihood given the high-probability category. However, they did not distinguish between the low/high and low/low conditions [M = 59.0, SD = 22.5 vs. M = 58.8, SD = 27.0; $t(59) = 0.07$, p = .94, d = 0.01, BF_{01} = 9.8], indicating that they ignored the feature likelihood given the low-probability category.

This finding qualifies any claim that social categories are adaptively useful due to their inductive potency: To the extent that social categories help to make predictions about individuals, they lead us to be overconfident in those predictions. This does not mean that categories are not helpful for navigating the social world, just as the corresponding findings in causal explanation do not show that it is useless to explain anything. Nonetheless, this result helps to show how cognitive mechanisms can contribute to prejudice: Even if a stereotype has a grain of truth, people apply it too zealously, failing to take into account other potential categories that may apply.

General Discussion

We often simplify the social world by using stereotypes, assuming that an individual's traits are consistent with their social category. Yet, it is often unclear what social category an individual belongs to. How, if at all, could we rely on stereotypes in such cases?

Here, we have shown that people are subject to a variety of biases in thinking about uncertain social categorizations. People use erroneous cues to *infer evidence* when diagnostic evidence is missing, leading to a bias against categorizations predicting unknown features (Exp. 1 and 3). People prefer *simpler* categorizations (belonging to one category) over more complex categorizations (belonging to multiple categories), but this tendency is eliminated when the stereotypical features are only heterogeneously linked with their categories (Exp. 2). People weigh *disconfirmed* predictions (negative evidence) more heavily *against* a category than they weigh *confirmed* predictions (positive evidence) in its favor (Exp. 3). And when people categorize an individual as *likely* belonging to a category, they treat that individual as *certainly* belonging to that category, when making

inferences about other features (Exp. 4).

These results help to clarify the mechanisms underlying social judgments and allow us to make two new claims about stereotype use: First, stereotypes act as *explanations* in much the same way that intentions explain behavior and causes explain effects; second, people use a set of *heuristics* to evaluate these explanations, which are shared across superficially distinct psychological processes.

These two claims are linked, and rely on the same underlying logic. Many inferential processes share a common informational structure, wherein hypotheses must be evaluated with respect to some body of data. In principle, these problems could be solved through Bayesian updating, accounting for a hypothesis's prior probability and fit to the data. But in practice, people use a variety of heuristics that (at best) *approximate* Bayesian reasoning, and these heuristics are highly similar across tasks such as classification, causal reasoning, and even some visual tasks. We take these heuristics to be a *signature* of explanatory reasoning and the *mechanism* by which these inferences are made. Thus, finding these heuristics at play in stereotype use is strong evidence that this process is both explanatory and heuristic.

We are currently expanding on this work in two ways. First, we are testing other explanatory biases (e.g., a bias to judge explanations with major behavioral implications to be more likely) in the social domain. Second, we are extending these findings from novel stimuli to more realistic, enriched situations. We look forward to the possibility that this work can help reveal how stereotypes are used in realistic settings, and potentially guide interventions to reduce bias and prejudice.

References

Anderson, S.M., & Klatzky, R.L. (1987). Traits and social stereotypes: Levels of categorization in person perception. *Journal of Personality and Social Psychology*, *53*, 235–46.

Brewer, M.B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, *41*, 656–70.

Fiske, S.T. (1993). Social cognition and social perception. *Annual Review of Psychology*, *44*, 155–194.

Hamilton, D.L., & Sherman, S.J. (1996). Perceiving persons and groups. *Psychological Review*, *103*, 336–55.

Hilton, J.L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–71.

Johnson, S.G.B., Jin, A., & Keil, F.C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Johnston, A.M., Toig, A.E., & Keil, F.C. (2014). Explanatory scope informs causal strength inferences. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015a). Argument scope in inductive reasoning: Evidence for an abductive account of induction. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015b). Predictions from uncertain beliefs. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2014). Inferred evidence in latent scope explanations. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015). *Sense-making under ignorance*. Under review.

Johnston, A.M., Johnson, S.G.B., Koven, M.L., & Keil, F.C. (2015). Probability versus heuristic accounts of explanation in children: Evidence from a latent scope bias. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Khemlani, S.S., Sussman, A.B., & Oppenheimer, D.M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–35.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–57.

McGarty, C., Yzerbyt, V.Y., & Spears, R. (Eds.) (2002). *Stereotypes as explanations*. Cambridge, UK: Cambridge University Press.

Murphy, G.L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–93.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance- versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, *53*, 621–35.

Park, B., Ryan, C.S., & Judd, C.M. (1992). The role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*, 553–67.

Pettigrew, T.F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*, 461–76.

Ross, B.H., & Murphy, G.L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*, 495–553.

Sussman, A.B., Khemlani, S.S., & Oppenheimer, D.M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Wittenbrink, B., Hilton, J.L., & Gist, P.L. (1998). In search of similarity: Stereotypes as naïve theories in social categorization. *Social Cognition*, *16*, 31–55.

Are children flexible speakers?
Effects of typicality and listener needs in children's event descriptions

Myrto Grigoroglou (Mgrigor@Udel.Edu)
Department of Linguistics and Cognitive Science, University of Delaware
125 East Main Street, Newark, DE 19716 USA

Anna Papafragou (APapafragou@Psych.Udel.Edu)
Department of Psychological and Brain Sciences, University of Delaware
105 The Green, Newark, DE 19716 USA

Abstract

Do children take into account their addressees' needs in spontaneous production? Developmental evidence for speaker adjustments is mixed. Some studies show that children are often under-informative when communicating with ignorant addressees but other studies demonstrate successes in children's ability to integrate another person's perspective. We asked whether children adapt their event descriptions depending on (a) the typicality of event components, and (b) the listener's visual access to the events. We found that children's ability to use information about the listener's visual perspective to make specific adjustments to event descriptions emerged only in highly interactive contexts, in which participants collaborated towards mutual goals.

Keywords: referential communication; event cognition; language production; instruments; perspective-taking; pragmatics

Introduction

According to a widely shared perspective, communication is a collaborative effort governed by rational expectations (Grice, 1957). On Grice's theory (1975), a collaborative speaker is reasonably expected to be as informative as required by the purpose of the communicative exchange (maxim of quantity), truthful (maxim of quality), relevant (maxim of relation), and perspicuous (maxim of manner). Within this framework, production is often considered as an addressee-oriented process in which speakers flexibly adjust their utterances to their listeners' informational needs in a given context (e.g., Clark & Marshall, 1981).

However, in a seminal study, Brown and Dell (1987) suggested that not all adjustments in production are oriented towards a specific addressee. They proposed that speakers make two types of adjustments in production. *Generic* adjustments are geared towards an unspecified comprehender. For instance, speakers are more likely to mention atypical (unusual) event components (as opposed to typical ones) because any comprehender would find atypical components harder to infer (e.g., 'Adolph stabbed the man *with an icepick*' is more felicitous compared to 'Adolph stabbed the man *with a knife*'). Additionally, speakers also make adjustments to *specific* addressees in a particular communicative context. For instance, speakers are more likely to offer a greater amount of information to an ignorant addressee, but less information to a knowledgeable addressee (to avoid redundancy).

In the present study, we are interested in whether children adjust their utterances to the needs of their addressees and– to the extent that they do–whether these adjustments are oriented towards a generic or specific interlocutor.

Children's referential communication

Previous research on children's ability to adjust to the informational needs of their addressees has focused on nominal reference. The usual paradigm used to elicit referential expressions from children involves the child asking a partner to move objects on a visual display. The crucial manipulation concerns the partner's knowledge state: The partner is either knowledgeable (has full visual access to the same objects as the child) or partially ignorant (she has limited visual access to the objects on display). This line of work has led to contrasting views concerning children's referential communication abilities.

On an *egocentric* view, young children's early egocentrism prevents them from adjusting their production to their addressees' informational needs. This view is supported by experimental evidence showing that children are often under-informative when communicating with ignorant addressees. For instance, 3-year-olds fail to adjust their utterances appropriately when addressing ignorant vs. knowledgeable interlocutors (Perner & Leekam, 1986) and 5-year-olds are often underinformative when describing one of two objects in a contrast set to an ignorant addressee (Davies & Katsos, 2010). Furthermore, 2-, 3- and 4-year old children produced ambiguous utterances before receiving any referential training (Matthews, Lieven, & Tomasello, 2007). Even older children (6- to 8-year-olds) have been shown to produce many ambiguous utterances in their referential communication (Deutsch & Pechmann, 1982; Girbau, 2001; Sonnenschein, 1982).

On an *audience-design* view, children are able to adjust their production to the needs of their addressees, as long as these needs are sufficiently transparent. Experimental evidence from certain types of tasks supports this view. Children seem to be sensitive to a partner's perspective when they are engaged in a task that has a clear communicative purpose (e.g., when children give their

partner instructions about how to manipulate objects on a visual display; see Bahtiyar & Küntay, 2009; Nadig & Sedivy, 2002). Children also become more informative when they are given specific feedback in the form of clarification questions (e.g., "[Do you need] the *girl eating an ice-cream* or the *girl swimming*?"; see Matthews, Lieven & Tomasello, 2007; Matthews et al., 2012). Additionally, children seem to be taking into account the knowledge state of their interlocutors when they are familiar with them: for instance, 2-year-old children are more likely to name a hidden toy when their mother has not witnessed the hiding (O'Neill, 1996).

The egocentric and audience-design views make specific predictions concerning the types of adjustments that children might make in production. Both accounts predict that children should make 'generic' adjustments, since these are guided by broad comprehension constraints and do not involve true perspective-taking. However, they make different predictions concerning children's ability to perform addressee-specific adjustments. According to the egocentric account, children might not make adjustments to specific addressees, because of limitations in the ability to take into account another person's perspective. According to the audience-design account, children might make listener-specific adjustments in contexts where listeners' informational needs are sufficiently transparent.

Current study

In this study, we compare the predictions of the egocentric and audience-design views by asking whether adults and preschoolers adapt their event descriptions depending on (a) the typicality of event components (generic adjustment), and (b) the listener's visual access (specific adjustment). Both factors have been argued to play a role in adults' early syntactic choices in production (Brown & Dell 1987; Lockridge & Brennan 2002). Unlike prior work that has focused on children's nominal reference, we elicit descriptions of events to explore children's adjustments. Describing events is much more complex than referring to single objects and requires more advanced syntactic structure. Specifically, we focus on instrument phrases in event descriptions, which are typically encoded in a non-obligatory adjunct (i.e., Ving *with a Y*).

Additionally, we explore the communicative circumstances under which children can make successful adaptations. Previous experimental evidence suggests that in certain types of tasks children are able to take into account the informational needs of their addressees, but more research is required to clarify exactly which factors contribute to children's success. In Experiment 1, we test whether the presence of an addressee with specific informational needs affects typical and atypical instrument mention. In Experiment 2, we test whether instrument mention is affected by presenting stimuli in contrastive pairs of typical/atypical events. In previous research, contrastive contexts have been shown to facilitate unique identification of referents (e.g., Brennan & Clark, 1996; Brown-Schmidt & Tanenhaus, 2006; Deutsch & Pechmann, 1982). In Experiment 3, we explore how instrument mention is affected by introducing a clear communicative goal to the task.

Experiment 1

In Experiment 1, participants watched short video clips depicting different events and described them to listeners who either saw or could not see the events. Half of the events contained typical and half atypical instruments (e.g., watering plants with a watering can/a hat used as a container). Both the egocentric and audience design accounts predict that participants should make generic adjustments in production by mentioning atypical instruments more frequently than typical instruments. In terms of addressee-specific adjustments, on the audience-design view, participants are expected to offer more information about instruments when their addressee has no visual access to events; on the egocentric view, instrument mention should not differ depending on the knowledge state of the addressee.

Methods

Participants Twenty-four 4- to 5-year-old children (range: 4;4-5;2, mean: 4;11) and twenty-four adults participated in the experiment. The children were recruited from daycares in the Newark (DE) area. Adults were undergraduate students at the University of Delaware and received course credit for their participation.

Materials Test items included 12 events depicting an agent performing an action. For each of the test events, we created two short video clips, one showing the agent performing an action using a typical instrument (e.g., watering plants with a can) and the other showing the agent performing the same action using an atypical instrument (e.g., watering plants with a hat). Typicality of instruments was pre-rated by a group of 14 adults and 16 children. Overall, adults mentioned the selected typical instruments in 73% of their responses and children in 63% of their responses. For atypical versions of the same events, we chose instruments that were either not mentioned at all or mentioned very infrequently (less than 6% of the time) by both children and adults. We chose these novel (i.e., highly atypical but still possible) instruments because we wanted to ensure that there would be a large difference between typical and atypical instruments in the test events, noticeable even by our younger participants. We also created a set of 6 filler clips showing various everyday actions that did not involve instruments (e.g., watching television, running). The same (male) agent performed all actions in the test and filler clips.

Procedure Participants were informed that they would be watching a set of short video clips and that, at the end of each video, they would have to describe what they saw.

They were also introduced to a 'friend' of the experimenter's (the confederate listener) who had not seen the videos and wanted to know what participants would see. Participants were randomly assigned to one of two conditions. In the *Visual Access* condition, the listener sat next to the participant and also watched the clips. In the *No Visual Access* condition, the listener sat behind an opaque barrier so that she was unable to see the videos (or the participants as they were describing them). No restrictions were placed on participants' productions**.**

Coding Participants' descriptions were tape-recorded and transcribed. Descriptions were coded for the explicit mention of instruments (either within the same clause or in a separate clause) before or after the main verb (e.g., "The guy is eating soup with a big spoon", "A man getting the knife and cutting something") or incorporated into the verb (e.g., "The man is hammering a fence"). We also coded for implicit mention of instruments, in cases where the instrument was not mentioned but simply inferred by the systematic choice of locution (e.g., "He is trying to open the door", for an event in which the agent was using a hanger to try and break into the room).

Results and Discussion

We conducted an ANOVA with Age (Children, Adults) and Visual Access (Visual Access, No Visual Access) as between-subjects factors and Typicality (Typical, Atypical) as a within-subjects factor. The analysis revealed a main effect of Age ($F(1, 43) = 89.56$, $p < .001$), a main effect of Visual Access ($F(1, 43) = 461.84$, $p < .001$), and a main effect of Typicality ($F(1, 43) = 224.31$, $p < .001$). These effects were qualified by an interaction between Age and Visual Access ($F(1, 43) = 9.60$, $p = .003$): adults were much more likely to add instrument information when the events were not visible to their interlocutor (M=.64) compared to situations where the interlocutor also had visual access to the events (M=.47; $p < .05$); in children, however, this difference was not significant (M=.20 vs. .23 respectively). The analysis also revealed an Age by Typicality interaction ($F(1, 43) = 29.04$, $p < .001$): adults were much more likely to mention instruments for Atypical compared to Typical versions of events (M=.87 vs. .22), while in children this difference was smaller (M=.36 vs. .06 respectively). No other interactions were found.

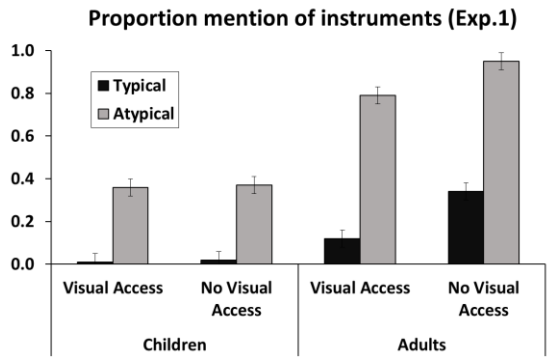


Figure 1: Proportion of mention of Typical and Atypical instruments by age group and type of Visual Access in Experiment 1

These results show that adults made both typicality-based adjustments and more specific adjustments to the informational needs of their addressee.[1] Children, however, performed only typicality-based adjustments by mentioning only the most unusual event component, a result in accordance with the egocentric view.

There are several explanations for the fact that children may have ignored the needs of their addressee. One possibility is that children had difficulty estimating the goals of the exchange: asking children to simply describe events for a passive listener may not have provided the necessary communicative goal that would highlight the listener's specific needs. To explore this possibility, we conducted a second experiment that clarified the goals of the exchange.

Experiment 2

In Experiment 2, we explored a new paradigm with the goal of making the addressee's needs more prominent for children: we asked whether instrument information can be identified and used by children to unambiguously single out and describe an event within a pair of closely matched alternatives. In such contrastive contexts, adults might be expected to produce instruments regardless of typicality or visual access (since the goal of the task is to disambiguate the right referent) but children might show both effects of typicality and visual access. We reasoned that such a contrastive context might highlight the need to clearly distinguish between two almost identical events for the sake of an uninformed addressee. In order to seek developmental changes in the ability to make use of perspective information, we compared two age groups of children. Recall that the egocentric account predicts that–despite the different manipulations–children should not show effects of visual access. The audience-design account, however, predicts that children should use instruments more

[1] It should be noted that adults' specific adaptations emerged despite our listener being a confederate and not a naïve participant (see Lockridge & Brennan, 2002 for discussion).

frequently when the addressee lacks visual access to the events.

Methods

Participants Sixty children and thirty adults participated. The children ranged between 4;0 and 6;0 and fell into two age groups: a younger group (n = 30) with a mean age of 4;7 and an older group (n = 30) with a mean age of 5;6. All children attended daycares in the Newark (DE) area. Adults were undergraduate students at the University of Delaware and received course credit for their participation.

Materials Materials consisted of pairs of events constructed out of clipart pictures. Each pair was displayed on a computer screen in Powerpoint format. There were 8 pairs of test events. Within each pair, the same event was depicted with a typical vs. an atypical instrument (e.g., a woman sweeping the floor with a broom vs. a tree branch). The position (left-right) of the typical and atypical versions were counterbalanced within the stimulus set. The test events were arranged in two presentation lists. In each list, one version of each event was placed within a red circle (half of the time, the circle was placed around a typical and the other half around an atypical event). For each event, the version that was circled was different between the two lists. Each presentation list also contained 8 pairs of control events.

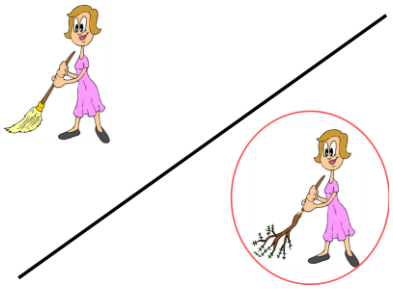


Figure 2: Example stimulus from Experiment 2 depicting a pair of typical and atypical events. The circle indicates the target event

Procedure Participants were assigned to either a *No Visual Access* or a *Visual Access condition*. In the No Visual Access condition, the experimenter introduced participants to her 'friend' (a confederate). Then the experimenter showed participants the display of events on a computer screen and told participants: "These are two twins. They are each doing something different. Look at both twins and tell [the confederate] what the twin inside the circle is doing. She has a picture of the twins too, but she doesn't know which one we are talking about." Participants saw that the confederate had a binder which contained color printouts of the pictures on their computer screen but lacked circles around the target pictures. The confederate was then seated across from participants so that she could not see the computer screen. Throughout the experiment, the confederate avoided eye contact with the participants but kept looking into her binder and followed the descriptions of the events turning the pages as appropriate. In the Visual Access condition, participants and confederate were seated next to each other so that they both had visual contact with the pictures described. The confederate looked at the screen as participants were describing the events and followed along by turning the pages in her binder.

Coding Participants' descriptions were tape-recorded and transcribed. For test items, responses were coded following the coding scheme of Experiment 1.

Results and Discussion

We conducted an ANOVA with the proportion of explicitly mentioned instruments as the dependent variable, Age (Younger, Older, Adult) and Visual Access (Visual Access, No Visual Access) as between-subjects factors and Typicality (Typical, Atypical) as a within-subjects factor. The analysis revealed a main effect of Age ($F(2, 84) = 103.60$, $p < .001$), with adults being more likely than either group of children to include instrument information overall (M_{ad}=.91 vs. M_y=.17 and M_o=.32; ps $< .001$), and older children being more likely to include more instrument information than younger children (p =.026). There was also a main effect of Typicality ($F(1, 84) = 55.07$, $p < .001$), qualified by an interaction between Age and Typicality ($F(2, 84) = 6.76$, $p = .002$): matched-pairs comparisons revealed that both older and younger children were much more likely to mention instruments for Atypical compared to Typical versions of events (younger: $t(29) = -6.71$, $p < .001$, M_A=.27 vs. M_T=.08; older: $t(29) = -5.64$, $p < .001$, M_A=.43 vs. M_T=.19), while in adults this difference was not significant ($p > .05$, M_A = .93 vs. M_T = .88). There were no other main or interaction effects.

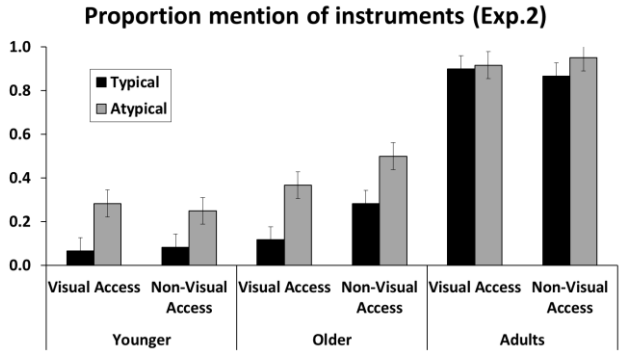


Figure 3: Proportion of mention of Typical and Atypical instruments by age group in Experiment 2

Thus, in these contrastive contexts, adults did not perform any type of adjustment (generic or specific). The reason is that instrument mention in adults was very high, regardless

of typicality or visual access, since the goal was to uniquely identify the correct referent. However, children made only generic (typicality-based) adjustments and did not adapt to specific addressees. This result provides support to the egocentric view of children's adjustments.

Experiment 3

Experiment 3 was a version of Experiment 2 with two main modifications: first, the addressee was no longer a confederate who acted as a passive listener but was introduced as a 'naïve' listener who was actively involved in the task; second, the task had a specific communicative purpose (guessing game). We reasoned that in a highly interactive paradigm, in which participants interacted with a 'real' addressee, children's ability to take into account the needs of their interlocutor might be more likely to arise.

Methods

Participants Thirty-two children and thirty adults participated[2]. The children ranged between 4;0 and 6;0 and fell into two age groups: a younger group (n = 19) with a mean age of 4;4 and an older group (n = 13) with a mean age of 5;2. All children attended daycares in the Newark (DE) area. Adults were undergraduate students at the University of Delaware and received course credit for their participation.

Materials and Procedure Materials were identical to Experiment 2. The procedure included the following modifications. The adult listener (the experimenter's confederate) was presented to the participants as a 'naïve' addressee. The experimenter told the participant and the confederate that they would play a guessing game together, in which the participant had to help the addressee guess the right event. At the beginning of each trial, the addressee said: "I can see two pictures. Which one is it? Tell me about it!" At the end of each trial, the addressee said: "I hope I got it right!" and placed a sticker next to the picture that best matched the participant's description. Feedback was provided in one practice trial at the beginning of the task. During the main test phase, the participants could not see where the addressee put the sticker so that their production was not affected. As in Experiment 2, participants were assigned in either a *Visual Access* or a *No Visual Access* condition.

Results and Discussion

We conducted an ANOVA with Age (Younger, Older, Adult) and Visual Access (Visual Access, No Visual Access) as between-subjects factors and Typicality (Typical, Atypical) as a within-subjects factor. The analysis yielded a significant effect of Age ($F(2, 56) = 37.319$, $p < .001$), with older children mentioning instruments more frequently than younger children, but less frequently than adults. Crucially there was a main effect of Visual Access ($F(1,56) = 5.029$, $p = .020$), indicating that participants of all age groups used more instrument information when the addressee did not have visual access to the events.

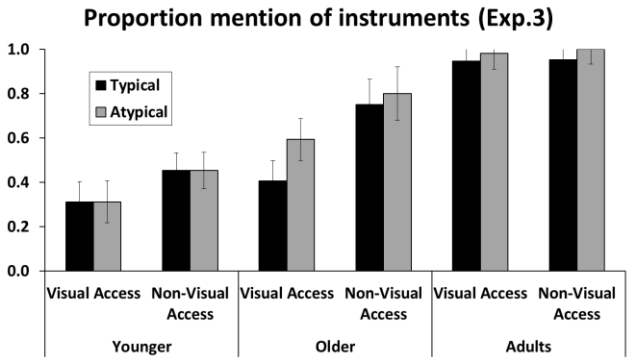


Figure 4: Proportion of mention of Typical and Atypical instruments by age group in Experiment 3

Therefore, in this more interactive paradigm, children showed sensitivity to their addressee's informational needs by mentioning instruments more frequently when the addressee could not see the events, in accordance with the audience-design view. In the context of contrastive presentation of the stimuli, children–just like adults–did not perform generic adjustments, since both types of instruments (typical, atypical) were important for disambiguating the correct referent.

General Discussion

The present study investigated referential communication patterns in children and adults, focusing on event reference. We showed that adult speakers, similarly to Brown and Dell (1987) and Lockridge and Brennan (2002), performed both 'generic' adjustments (adding information about atypical, i.e., generally unpredictable, instruments) and more specific adjustments to addressees' needs (mentioning instruments more often when addressees could not see the events). Children, however, often made only generic (typicality-based) adjustments. Their ability to use information about the listener's visual perspective to make specific adjustments to their event descriptions emerged only in contexts where the addressees' needs were made particularly transparent.

What were the precise factors that made addressees' needs accessible to children in Experiment 3? An important difference between Experiments 1-2 and Experiment 3 was the role of the addressee. In Experiment 3, the addressee was a "real" interlocutor, who had more genuine informational needs that children could easily identify. In fact, in studies that show early successes in children's tendency to make addressee-specific adjustments, the addressees are either the children's parents (O' Neill & Topolovec, 2001; O'Neill, 1996) or confederates of the experimenter with an active role in the task (Bahtiyar &

[2] The results presented here are preliminary.

Küntay, 2009; Nadig & Sedivy, 2002). By contrast, in studies where the addressee is either a static picture on a computer screen or imaginary (Davies & Katsos 2010; Girbau 2001), children fail to make adaptations. Even adults seem to be more likely to make addressee-specific adjustments when the addressee is naïve as opposed to a confederate (see Brown & Dell, 1987; Lockridge & Brennan, 2002).

A second important difference between Experiment 3 and the previous experiments is the communicative purpose of the task. In Experiment 3, the speaker and the addressee had to engage in a collaborative process (guessing game) to achieve a mutually pursued goal (finding the 'right' picture). This conclusion is supported by previous findings that demonstrate children's sensitivity to other people's perspective in tasks that require collaboration between interlocutors (e.g., Deutsch & Pechmann, 1982; Matthews, Lieven, & Tomasello, 2010; O' Neill & Topolovec, 2001; O'Neill, 1996).

Overall, our findings suggest that children do not seem to be egocentric communicators. However, they need extra communicative cues to facilitate their assumptions about what type of information is relevant for the purpose of the exchange and how much information their addressee needs.

Acknowledgement

We would like to thank the children and the parents in the preschools of Newark, DE area. Special thanks to Dan Curtin, Sarah Fairchild and Hanna Schwartz for their assistance in data collection. This research has been partially supported by NSF grant BCS-0749870 to A.P.

References

Bahtiyar, S., & Küntay, A. C. (2009). Integration of communicative partner's visual perspective in patterns of referential requests. *Journal of Child Language*, *36*(3),

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*(4), 441–472.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. H. Joshi, B. I. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, *120*(8), 1956–1972.

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*(2), 159–184.

Girbau, D. (2001). Children's Referential Communication Failure: The Ambiguity and Abbreviation of Message. *Journal of Language and Social Psychology*, *20*(1-2), 81–89.

Grice, H., P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388.

Grice, H., P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics: Speech acts*, Volume 3 (pp. 41–58). New York: Academic Press.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*(3), 550–557.

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and Four-Year-Olds Learn to Adapt Referring Expressions to Context: Effects of Distracters and Feedback on Referential Communication. *Topics in Cognitive Science*, *4*(2), 184–210.

Matthews, D., Lieven, E., & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, *78*(6), 1744–1759.

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Developmental Psychology*, *46*(4), 749–760.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336.

O' Neill, D. K., & Topolovec, J. C. (2001). Two-year-old children's sensitivity to the referential (in)efficacy of their own pointing gestures. *Journal of Child Language*, *28*, 1–28.

O'Neill, D. K. (1996). Two-Year-Old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*(2), 659.

Perner, J., & Leekam, S. R. (1986). Belief and quantity: three-year olds' adaptation to listener's knowledge. *Journal of Child Language*, *13*(2), 305–315.

Sonnenschein, S. (1982). The Effects of Redundant Communications on Listeners: When More Is Less. *Child Development*, *53*(3), 717.

From embodied metaphors to metaphoric gestures

Margot Lhommet (m.lhommet@neu.edu)
Stacy Marsella (s.marsella@neu.edu)
College of Computer and Information Sciences
Northeastern University
Boston, MA 02115 USA

Abstract

Humans turn abstract referents and discourse structures into gesture using metaphors. The semantic relation between abstract communicative intentions and their physical realization in gesture is a question that has not been fully addressed. Our hypothesis is that a limited set of primary metaphors and image schemas underlies a wide range of gestures. Our analysis of a video corpus supports this view: over 90% of the gestures in the corpus are structured by image schemas via a limited set of primary metaphors. This analysis informs the extension of a computational model that grounds various communicative intentions to a physical, embodied context, using those primary metaphors and image schemas. This model is used to generate gesture performances for virtual characters.

Keywords: embodied cognition; gesture; metaphor; nonverbal behavior; human-computer interaction

Introduction

Metaphoric gestures turn abstract ideas and discourse structures into the visual and the embodied. For example, holding or weighting a large object can suggest the importance of an idea. Metaphoric gestures also structure the discourse, for example by putting ideas in distinct locations in the physical space to allow referring to them later on or to emphasize their difference. The chosen locations can have a metaphorical meaning as well: for example, events located on the left are understood as being in the past while events on the right are in the future (Calbris, 2011).

When modeling how speakers select gestures to realize a communicative intention, a key challenge arises: how can gestures, that are actions inherently specified in physical terms such as size, location or path, communicate meaningful information about abstract elements that do not have physical features? In other words, where does the semantic relation between abstract referents (such as an important idea) and their gesture portrayal (a big object) comes from?

There is evidence that the human conceptual system is embodied and structured by metaphors (Tversky & Hard, 2009). We understand abstract concepts by mapping them to image schemas (embodied experiential concepts) (Lakoff & Johnson, 1980). Reasoning processes are actions taken on these image schemas (Barsalou, 2009). For example, we make sense of the sentence "the price rises" by our understanding that an increase in quantity often correlates with an increase in height. Lakoff and Johnson (1980) and other researchers have studied how conceptual metaphors are reflected on the verbal channel via verbal metaphors. One outcome of their research is lists of conventional metaphors that link abstract domains to concrete domains (see for example Grady's list of Primary Metaphors (1997).

There is evidence that these conceptual metaphors also shape gestures (see (Cienki, 2008) for a review). Our previous work proposed a computational model that maps communicative intentions to two highly expressive image schemas (CONTAINER and OBJECT) that are common to a wide range of metaphoric gestures (Lhommet & Marsella, 2014). Although it uses only two image schemas and a restricted set of metaphors to guide the mapping, this model supports the generation of gestures that convey a wide range of communicative intentions. In particular, gestures communicate information about the referent (e.g. depicting a big object to suggest an important idea) and structure the discourse itself (e.g. contrasting facts by assigning them opposite locations in space). This suggests that it is possible to model a large range (if not the whole range) of gestures using a restricted set of image schemas and primary metaphors.

More specifically, such a model of gesture generation:

- allows for a large space of communicative intentions to be mapped to a comparatively small space of concrete elements (image schemas).
- can convey complex communicative intentions via composition over this small set of image schemas.
- guides how properties in abstract propositions (such as "important idea") can be conveyed by manipulations of the gesture property (size of the gesture).

In this paper, we systematically investigate and extend the coverage of our previous model by using a corpus to study how communicative intentions are mapped to gesture elements via primary metaphors. The first section gives an overview of the computational model. The second section presents the analysis of the corpus. We then describe the implementation by focusing on two examples. Finally, we conclude by mentioning the advances and limits of this approach as well as discussing future work.

Model

This work builds on a previous model of gesture generation that maps communicative intentions (CIs) to a mental space structured by image schemas (Lhommet & Marsella, 2014). This mapping is guided by Grady's list of primary metaphors (1997), referred to as PMs in the rest of this paper.

This process is illustrated in Figure 1. First, a CI is grounded, i.e. mapped to image schemas that have physical properties using PMs. These properties then inform the generation of a gesture plan that conveys the desired meaning.

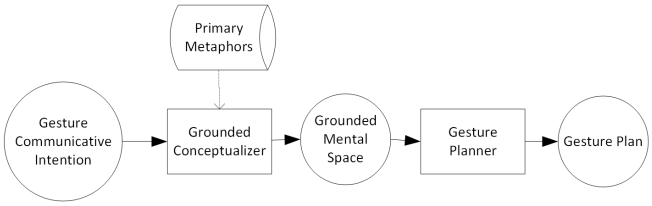


Figure 1: Our model maps CI to concrete elements using PMs to generate gestures.

Communicative Intentions Gesture can express a wide range of information that can complement, reinforce or contradict the information communicated via other modalities (Kendon, 2000). Our model takes as input a CI that describes the meaning that a speaker wants to convey via gesture. This CI contains the minimal set of information required to generate a gesture performance that communicates the intended meaning. For example, a speaker wants to give information about the social status of an individual.

Grounded Conceptualizer The *Grounded Conceptualizer* maps the elements of the CI to image schemas. PMs systematically project the objects, properties and relations from one domain to another[1]. For example, the PM SOCIAL STATUS IS VERTICAL ELEVATION links the social status of an individual (or entity) to a location on a vertical scale. Individuals with a lower social status will have a lower location in space, while climbing up the ladder means that they improve their social status.

Grounded Mental Space A grounded mental space is structured by image schemas and actions taken on them. This is in line with the work of Barsalou (2009) and others, that show that the brain regions responsible for perception and action coordinate during meaning creation and comprehension to create "embodied simulations" of linguistic content. This suggests that thought and reasoning processes are actually actions taken on the objects of the grounded mental space. Using our previous example, we can move individuals up and down the social status scale and infer how it impacts their social status.

Since the grounded mental space informs the generation of gesture, it should contain elements that have gesture correlates. For example, the representation of an individual on a social status scale suggests the existence of a concrete OBJECT with a physical elevation in space, and actions of moving up or down.

Gesture Planner Finally, the *Gesture mapper* combines elements of a grounded mental space into a gesture plan. This FML-like output (Heylen, Kopp, Marsella, Pelachaud, & Vilhjálmsson, 2008) can be interpreted by a nonverbal behavior generator (such as (Marsella et al., 2013)) to generate a multimodal performance.

[1]This process can be seen as a simplified blending (Fauconnier & Turner, 2008)

Corpus

To help quantify the PMs and image schemas that play a significant role in the generation of metaphoric gestures, we created an annotated corpus of human gesturing. Several criteria were taken into account: 1. the gesturers should be "good gesturers", 2. have both of their hands visible and free, 3. the discussion topic should be abstract to elicit metaphoric gestures and 4. the discussion should be improvised instead of rehearsed.

Description We used a video[2] from the footage of the Working Families Summit (Washington D.C., June 23rd 2014). 6 female speakers (a journalist, a politician, two professors, a CEO and an activist) discuss abstract concepts such as time, flexibility, income and social status. The 50 minutes video was chunked into segments that portray only one speaker at a time. Pauses and segments where the journalist holds a pen and a notebook were discarded, leaving a total of 22 videos with a mean duration of 1min 42s (SD=50s), for a total of 37min 32s.

Annotations 2 coders annotated the corpus with VideoAnt[3]. Coders could freely annotate what they consider as a gesture, then select the CI reflected by the gesture from the following list:

- Generic reference: simple reference to an object or fact
- Specialized reference: reference to an object or fact and depiction of one or several of its properties
- Action: reference to an action
- Discourse structure: enumeration, contrast or causal relationship
- Other: none of the previous categories seems appropriate

They also annotated which element(s) of the gesture convey the desired meaning (e.g. the size of the object depicted, the shape of the motion) and selected the primary metaphor(s) that underlies this association (from Grady's list of 100 primary metaphors).

The coders were trained using a video segment that was discarded from the corpus. They both carried the analysis on the whole dataset. Inter-coder agreement was .69 using Cohen's kappa (a kappa value above .61 indicates substantial agreement (Landis & Koch, 1977)). Further analysis indicated that most disagreements were caused by two situations: one coder annotated a movement as a gesture while the other did not consider it as a gesture, or gestures were not perfectly aligned in time. After discussing these cases, a kappa value of .91 was obtained.

[2]https://www.youtube.com/watch?v=cQBlciBr_3w

[3]VideoAnt is a web-based annotation tool developed by the College of Education & Human Development at the University of Minnesota, available at http://ant.umn.edu

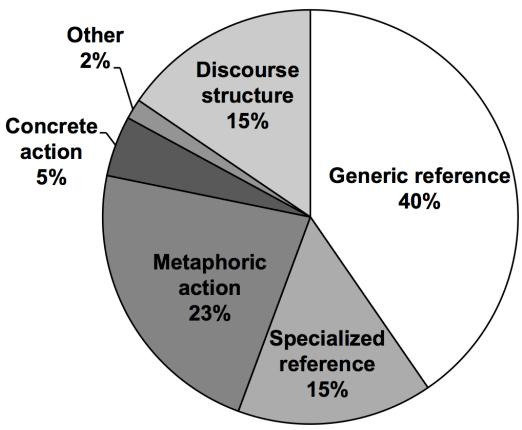


Figure 2: Distribution of CIs in the corpus.

Analysis

The final dataset consists of 740 gestures (with an average of one gesture every 3 seconds). Figure 2 describes the list of CIs that results from the analysis of the corpus.

Generic references 40% of the gestures are known to gesture researchers as "Conduits" (Reddy, 1979). The PM ABSTRACT IS CONCRETE instantiates an object in space to represent a concrete or an abstract object or an element of the discourse. The hand, facing up with an open palm, presents an immaterial object for the viewer to see. The type of the referent has little impact on the gesture shape (McNeill, 2005).

Specialized references 15% of CIs consist in illustrating abstract properties of referents. The following PMs are used in the corpus to map abstract properties to physical properties expressed with gestures:

- *Object location* (46%): location of the object in the physical space
  - SOCIAL STATUS IS VERTICAL ELEVATION (25%)
  - MOMENT IN TIME IS LOCATION (15%)
  - KNOWLEDGE IS LOCATED IN THE HEAD (6%)
- *Object size* (25%): e.g. the distance between two hands or the size of the gap between two fingers
  - QUANTITY IS SIZE (15%)
  - IMPORTANCE IS SIZE (10%)
- *Object shape* (29%): the shape of the hands reflects the shape of the referent
  - ESSENTIAL IS INTERNAL (23%): e.g. palms oriented towards the speaker
  - CERTAIN IS FIRM (6%): e.g. hand shape is a fist

Depicting actions 25% of the gestures represent actions. 20% are metaphoric actions, i.e. prototypical actions that have meaning based on a underlying metaphor (such as depicting improving one's social status by moving an object up in the physical space). Figure 3 represents the distribution of primary metaphors that underly the generation of metaphoric gesture in the corpus. 5% of the CIs are concrete actions that mimic an actor acting in the physical space (such as a woman lifting a brick over her head).

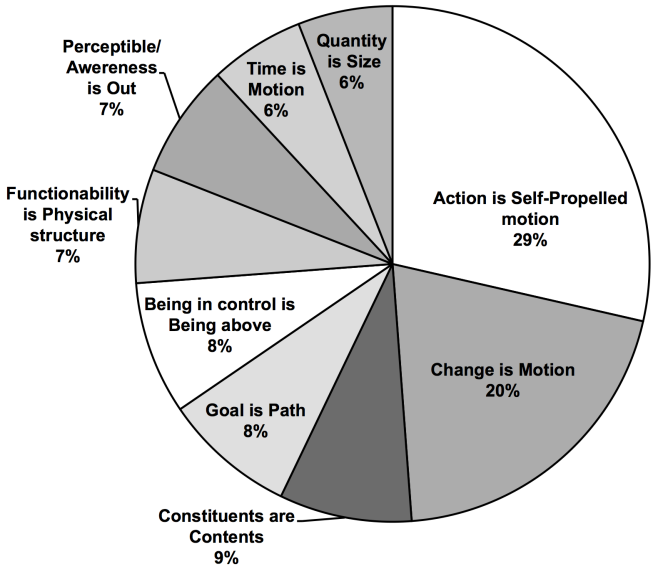


Figure 3: Distribution of the PMs underlying metaphoric actions in the corpus.

The reader familiar with gesture studies may notice that the distribution between concrete and abstract actions differs from what is typically reported, with comparatively few concrete actions and a lot of metaphoric actions. Our view is that this difference is largely due to the nature of the corpora used. Most research on gesture have used corpora about physical phenomena (e.g. retelling a scene from a cartoon (McNeill, 1992) or explaining how to navigate a city (Bergmann & Kopp, 2009)). Therefore, gestures in these corpora reflect concrete actions. Our corpus focuses on abstract topics that do not have concrete features, so most gestures depict metaphoric actions.

Discourse structures 15% of the gestures structure and organize the discourse. Among them, enumerations, contrasts and expression of causality are equally distributed. Half of the enumerations in the corpus are represented as objects sequentially taken out of a container. The other half by counting on fingers. Expression of causality relies on the primary metaphor EFFECTS ARE OBJECTS WHICH EMERGE FROM CAUSES. Contrasting objects over a property relies on the metaphor SIMILARITY IS PROXIMITY where the distance between objects represents how much they differ regarding this property. The property itself can influence elements of the gesture; for example, comparing the social status of two individuals uses the vertical scale while comparing events in time uses the horizontal scale. Our previous work offers additional detail on discourse structures and their relation to primary metaphors (Lhommet & Marsella, 2014).

Others 2% of the gestures communicate intentions that are not covered by the annotation scheme (6 occurrences over 740 gestures). These are gestures that express uncertainty



(a) A continuing action (b) The shape of an action

Figure 4: Gestures can depict actions at two levels

(shrugs combined to stereotyped facial expressions) as well as emblem gestures (in particular, the corpus counts two occurrences of "quotes" traced in the physical space).

Implementation

This computational model is implemented into a framework that leverages the Cyc architecture[4]. Cyc embeds a first-order logic reasoning engine that runs forward and backward inferences. The knowledge is hierarchically organized so properties and rules can be propagated along the inheritance links. The previous version of the framework, presented in (Lhommet & Marsella, 2014), can derive gesture performances for several communicative intentions: (a) depicting generic referents, (b) depicting properties of object using elaborate metaphors and (c) realizing enumerations and (d) contrasts .

In the rest of this section, we extend this framework to generate gestures that communicate information about actions. Our corpus analysis showed that gestures can communicate two kinds of information about actions: (a) the status of an action: the speaker on Figure 4a says "In this country we have to continue to do that" while making a loop in the physical space. (b) the physical shape of an action: another speakers says "a lot of countries have horrible cultural mores that are suppressing women" while making the gesture depicted by Figure 4b. This gesture suggests a force applied downwards that represents the control applied on women. It seems driven by the primary metaphor BEING IN CONTROL IS BEING ABOVE.

Communicative intentions are specified using Cyc's first-order logic declarative language. Script 1 represents the CIs associated to the gestures depicted in Figures 4a-4b, using pseudocode for clarity.

Cyc's high-level term *Action*, and specializations of this term with more refined meanings, are used to model the CIs. In Script 1(a), *Continuation* specifies that an action previously initiated continues. In Script 1(b), the action is typed as *ExercisingAuthoritativeControl*, a specialization of *ControllingSomething*, which itself inherits from *PurposefulPhysicalAction*. The actor (the mores) and object (the women) of the action are associated to the action using predicates.

[4] http://www.cyc.com

Script 1 Communicative intentions to depict actions

(a) Depict a continuing action: "We have to continue to do that"

```
(intention depictAction a)
(isa Continuation a)
```

(b) Depict the shape of an action: "Mores are suppressing women"

```
(intention depictAction b)
(isa ExercisingAuthoritativeControl b)
(performedBy b mores) (objectControlled b women)
```

Primary metaphors are modeled as inference rules that map terms from the CIs to concrete terms that represent image schemas, using Cyc's forward chaining engine. During the grounding phase, all primary metaphors rules are tested against the contents of a given CI. If the condition side of the rule (i.e. the tuples before the '->' symbol) matches the input, then the predicates in the action side of the rule (i.e. the tuples after the '->' symbol) are set as true. The grounded mental space is created with all the predicates that are true when quiescence occurs (i.e. no rule matches anymore).

Script 2 details the implementation of the primary metaphor BEING IN CONTROL IS BEING ABOVE. Applying this rule to the CI defined by Script 1b) results in adding to the grounded mental space two *Concrete Objects* that represent the mores and the women, and assigning them locations on a vertical scale such as the object representing the mores is located above the object representing the women. Another rule, not depicted here, matches with the fact that the action is a *PurposefulPhysicalAction* and adds a `(shape act forceful)` predicate to the grounded mental space.

Script 2 Primary metaphor: BEING IN CONTROL IS BEING ABOVE

```
(isa ControllingSomething act)
(performedBy act actor) (isa actor Agent)
(objectControlled a object) (isa object Thing)
->
(isa ConcreteObject actor2)
(isa ConcreteObject object2)
(location actor2 locA) (location object2 locO)
(> locA locO s) (isa s VerticalScale)
```

Gesture plans are derived by another set of inference rules. They convert the grounded mental space into a gesture plan that reflects the physical properties using a FML-like formalism (Heylen et al., 2008). The gesture plans for the mentioned examples are described by Script 3. The system proposed by Xu, Pelachaud, and Marsella (2014) converts this formalism into the standard BML format (Kopp et al., 2006) to be rendered by the SmartBody animation system (Thiebaux, Marsella, Marshall, & Kallmann, 2008).

Script 3 Gesture plans

(a) Depict a continuing action

```
<goal=depictShape shape=cycle/>
```

(b) Depict the shape of an action: "Mores are suppressing women"

```
<goal=depictShape shape=force source=locA target=locB
scale=vertical constraints=[locA>locB]/>
```

Related Work

Researchers have explored several techniques to automate the generation of virtual humans' nonverbal behaviors that realize communicative intentions. Earlier systems used manual annotations of the information to convey nonverbally (e.g. (Kopp & Wachsmuth, 2002)). Some systems learn the mapping from speech input to specific classes of nonverbal behaviors (e.g. prosody to beat gestures (Levine, Krähenbühl, Thrun, & Koltun, 2010), text to head movements (Lee & Marsella, 2010) or text to gesturing style (Neff, Kipp, Albrecht, & Seidel, 2008). Other approaches rely on expert rules that infer information from the speech. BEAT infers rheme and theme from the text to generate intonation and emphasis (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001). NVBG detects communicative intentions in the text (e.g. affirmation, emphasis, disfluencies) using a keywords mapping (Lee & Marsella, 2006). Cerebella integrates acoustic, syntactic and semantic analyses to infer communicative intentions and elements of the mental state (emotional state, energy, emphasis,...) (Marsella et al., 2013; Lhommet & Marsella, 2013). Approaches that take speech as input generate nonverbal behavior that is limited in the range of what can be inferred from the speech utterance only.

Some work address the production of speech and gesture from a joint representation. Bergmann, Kahl, and Kopp (2013) studied how linguistic and cognitive constraints impact the coordination of speech and gesture. Lascarides and Stone (2009) formalize the relation of gesture and speech with a logical form of multimodal discourse, in particular between discourse elements and deictic gestures. In the Gestures as Simulated Action framework, perceptual and motor representations automatically become active during language production and, under certain conditions are sources of gestures (Hostetter & Alibali, 2008).

Discussion

In this paper, we presented a computational model of gesture generation informed by embodied cognition that turns various communicative intentions into gesture by grounding them in a physical, embodied context. Using the analysis of a video corpus, we showed that most CIs present in the corpus can be conveyed using a very limited set of PMs (at the exception of stereotyped gestures that could easily be integrated by providing a direct mapping from specific CIs to these emblem gestures.)

A possible application of this model is the automatic generation of multimodal performances for virtual humans. Virtual humans engage users in face-to-face interactions, ideally using the same verbal and nonverbal behaviors as humans (Cassell, 2000) have proven to be effective in a wide range of applications such as health to training simulations (e.g. (DeVault et al., 2014)). Metaphoric gestures improve message understanding and impact how a speaker is perceived, in particular in terms of persuasiveness and competence (Beaudoin-Ryan & Goldin-Meadow, 2014). This may be another reason why metaphoric gestures dominate in this corpus since all the speakers are professional public speakers. Given that good communication skills, persuasiveness and competence are critical in health interventions and training, metaphoric gestures should be an important capability of virtual humans designed for these applications.

Furthermore, this computational model provides a more controlled yet flexible methodology to experiment with social and psychological constructs; for example, virtual humans can serve as confederates in psychology and social psychology experiments to study the impact of nonverbal behaviors.

A limit to the broad application of this work is the need to manually specify the gesture communicative intent of the speaker. A promising avenue here is the Embodied Construction Grammar (ECG) framework (Bergen & Chang, 2005) that represents a speaker's intended meaning based on image schemas, along with the mental simulation of these representations using executing schemas (S. S. Narayanan, 1997). Our future work will investigate the integration of our computational model into the ECG framework, in particular applying the work of S. Narayanan (1999) on inferring and reasoning on conceptual metaphors from speech onto gesture.

References

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281-1289.

Beaudoin-Ryan, L., & Goldin-Meadow, S. (2014). Teaching moral reasoning through gesture. *Developmental Science*, *17*(6), 984-990.

Bergen, B. K., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction grammars: Cognitive grounding and theoretical extensions* (p. 147-). John Benjamins Publishing.

Bergmann, K., Kahl, S., & Kopp, S. (2013). Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *Intelligent virtual agents* (p. 203-216). Edinburgh, UK.

Bergmann, K., & Kopp, S. (2009). Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of the 8th international conference on autonomous agents and multiagent systems - volume 1* (p. 361-368). Richland,

SC: International Foundation for Autonomous Agents and Multiagent Systems.

Calbris, G. (2011). *Elements of meaning in gesture* (Vol. 5). John Benjamins Publishing.

Cassell, J. (2000). *Embodied conversational agents*. MIT Press.

Cassell, J., Nakano, Y., Bickmore, T., Sidner, C., & Rich, C. (2001). Annotating and generating posture from discourse structure in embodied conversational agents. In *Workshop on representing, annotating, and evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents.*

Cienki, A. (2008). Why study metaphor and gesture. In A. Cienki & C. Müller (Eds.), *Metaphor and gesture* (p. 5-25). Amsterdam; Philadelphia: John Benjamins Pub. Co.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., ... Morency, L.-P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on autonomous agents and multi-agent systems* (p. 1061-1068). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Fauconnier, G., & Turner, M. (2008). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., & Vilhjálmsson, H. (2008). The next step towards a function markup language. In *Intelligent virtual agents* (p. 270-280). Tokyo, Japan.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review*, *15*(3), 495-514.

Kendon, A. (2000). Language and gesture: Unity or duality. In D. McNeill (Ed.), *Language and gesture* (p. 47-63). Cambridge University Press.

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., ... Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents* (p. 205-217). Marina del Rey, CA.

Kopp, S., & Wachsmuth, I. (2002). Model-based animation of co-verbal gesture. In *Computer animation, 2002. proceedings of* (p. 252-257). Geneva, Switzerland.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by* (2003rd ed.). Chicago: University of Chicago Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lascarides, A., & Stone, M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*, *26*(4), 393-449.

Lee, J., & Marsella, S. C. (2006). Nonverbal behavior generator for embodied conversational agents. In *6th international conference on intelligent virtual agents.* Marina del Rey, CA.

Lee, J., & Marsella, S. C. (2010). Predicting speaker head nods and the effects of affective information. *Multimedia, IEEE Transactions on*, *12*(6), 552-562.

Levine, S., Krähenbühl, P., Thrun, S., & Koltun, V. (2010). Gesture controllers. In *Acm siggraph 2010 papers* (p. 124:1-124:11). New York, NY, USA: ACM.

Lhommet, M., & Marsella, S. (2014). Metaphoric gestures: Towards grounded mental spaces. In *Intelligent virtual agents* (Vol. 8637, p. 264-274). Boston, MA: Springer International Publishing.

Lhommet, M., & Marsella, S. C. (2013). Gesture with meaning. In *Intelligent virtual agents* (Vol. 8108, p. 303-312). Edinburgh, UK: Springer Berlin Heidelberg.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., & Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th acm siggraph/eurographics symposium on computer animation, anaheim, ca* (p. 25-35). New York, NY: ACM.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

Narayanan, S. (1999). Moving right along: A computational model of metaphoric reasoning about events. In *In proceedings of the national conference on artificial intelligence (aaai'99.* Orlando, FL.

Narayanan, S. S. (1997). *Knowledge-based action representations for metaphor and aspect (karma)*. Unpublished doctoral dissertation, UNIVERSITY of CALIFORNIA.

Neff, M., Kipp, M., Albrecht, I., & Seidel, H. P. (2008). Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics (TOG)*, *27*(1).

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (Vol. 2, p. 164-201). Cambridge: Cambridge University Press.

Thiebaux, M., Marsella, S., Marshall, A. N., & Kallmann, M. (2008). Smartbody: behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems* (Vol. 1, p. 151-158). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, *110*(1), 124-129.

Xu, Y., Pelachaud, C., & Marsella, S. (2014). Compound gesture generation: A model based on ideational units. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Intelligent virtual agents* (p. 477-491). Springer International Publishing.

Relation between bimanual coordination and whole-body balancing on a slackline

Kentaro KODAMA (kkodama@kanagawa-u.ac.jp)
Faculty of Economics, Department of Economics, Kanagawa University
3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa, Japan

Yusuke KIKUCHI (ykikuchi@fun.ac.jp)
Future University Hakodate
116-2, Kamedanakano-cho, Hakodate-shi, Hokkaido, Japan

Hideo YAMAGIWA (somatolearning@gmail.com)
Tokyo Metropolitan Tobu Medical Center
3-3-25, Shinsuna, Koto-ku, Tokyo, Japan

Abstract

To reveal the fundamental skills involved in slacklining, this study examined a hypothesis regarding single-leg standing on a slackline. In the field of practice, instructors teach learners how to maintain balance on a swinging flat belt (slackline), such as by moving their hands in parallel. We hypothesized that bimanual coordination in the horizontal direction might contribute to dynamic balancing on a slackline. In our pilot study, two participants at different skill levels were asked to maintain their balance on a slackline as long as possible. The dynamic stability of bimanual coordination was assessed by a nonlinear time series analysis (cross recurrence quantification analysis), then compared among the participants. Bimanual coordination stability was higher in the experienced player than in the novice player. The results suggest that the single-leg standing skill might be correlated with bimanual coordination stability. Further investigations are expected to clarify this notion in the future.

Keywords: slackline; balance sport; dynamic stability; whole-body coordination; self-organization

Introduction

Slacklining

A slackline is a flat belt that is fixed between two anchor points. Because it bounces and swings in all directions, it is very difficult for beginners to maintaining balance on it. Slacklining exists in two main forms, as balance training and as a competitive balance sport. The latter form first started around 2007 in Europe and has spread to other parts of the world. International contests have been held in which the competitors display various acrobatic skills. Slacklining has also received much attention as a method of balance training, not only for top-level athletes (e.g., Olympic ski jumpers), but also for the elderly or those with motor disorders and are participating in rehabilitation (one of the present authors has used slackline training for rehabilitation). Compared with traditional balance training with a balance ball or beam, we consider, slackline training is enjoyable as a sport or leisure pursuit and can be a sustainable lifelong activity.

Although slacklining has progressed in both its training and competitive forms, academic research on the topic has been limited. Most existing research has examined the effect of slackline training on balancing ability (e.g., Granacher et al., 2010). Although these studies are important to obtain evidence on the effects of slackline training on balance abilities, the skills involved in slacklining remain unclear. Beginners in slackline training experience considerable difficulty in understanding how to start and develop their skill. One study addressed questions of skill but only through a case study that investigated the specific situation of balance recovery after perturbation (Huber & Kleindl, 2010). The fundamental skills required for slacklining have not yet been investigated rigorously.

In this study, we assumed that single-leg standing on the slackline is an essential skill for successful slacklining. This is the first ones used when being instructed in slackline training (e.g., Keller et al., 2012). In practical training, after mastering this ability without support, beginners are encouraged to proceed to the next step, such as walking on a slackline. Therefore, in the present study we chose to experimentally investigate the single-leg standing task as a fundamental skill for slacklining.

Whole-body dynamic balancing

With regard to maintaining postural balance, one might argue that a static balancing strategy, involving the absolute straining of muscles and fixing of joints as if one's body were an inanimate object consisting of rigid materials, is one possible approach. In fact, traditional postural balance studies have assumed such a static model and have often used a *quiet standing task* to evaluate balancing ability (e.g., Winter, 1995). Those studies have also applied static indexes such as the trajectory length of the center of pressure, concluding that shorter lengths are more stable. According to such a static model and indexes, less movement is interpreted as indicating greater stability of the human postural system (e.g., Horak, 1989).

This static balance model, however, is not in accordance with the fact that living systems always fluctuate in various time scales at different levels, from the microscopic level (e.g., the cell) to the macroscopic level (e.g., the skeleton). Such an intrinsic fluctuation can be observed even in a quiet

standing task (Balasubramaniam & Wing, 2002). Recent studies of human postural balance have demonstrated the relevance of dynamic balancing (e.g., Delignières, Torre, & Bernard, 2011; Michael a Riley & Turvey, 2002). In these studies, human postural sway is regarded not as meaningless random noise at the musculoskeletal level but as having a meaningful structure emerging from interactions among components in the body-environment system.

Furthermore, it is difficult to investigate whole-body dynamic balancing by means of the quiet standing task paradigm, because quiet standing does not require participants to actively use their whole-body including the upper limbs. In our daily behavior, we often stand and walk on various unstable environments (e.g., sloped ground, bumpy roads, suspension bridges, and so on). Such unstable environments require us to maintain balance dynamically with whole-body coordination, including bimanual coordination of the upper limbs. While one is on a slackline, dynamic balancing with whole-body coordination is more critical because the fluctuation generated by one's own body movement can easily cause amplification of the line's fluctuation, causing one to fall off. Explaining how one maintains balance on a slackline by a static balance strategy is difficult, given the intrinsic fluctuation of an embodied system consisting of so many non-rigid components at different timescales.

Emergence of stable coordination pattern

The human body has a vast range of components (i.e., degrees of freedom: DoFs), from the microscopic cell level (10^{14}) to the macroscopic joint level (10^{2}) (Bernstein, 1967; Turvey, 1990). The *DoF problem* suggested that the large number of independently controllable DoFs places a computational burden on the central nervous system (Turvey, 1990). This indicates the limitations of the unidirectional top-down motor control model, represented metaphorically in terms of computer information processing. In addition, with regard to whole-body movement, solving the DoF problem becomes more difficult because many DoFs need to be considered in a whole-body system. Bernstein, who proposed the DoF problem, suggested as a possible solution that each component (DoF) is coordinated and coupled with other components to organize a functional structure (i.e., *synergy*) rather than being controlled separately (Bernstein, 1967).

Subsequent to Bernstein's suggestion of synergy, an alternative derived from self-organization theory, known as the *dynamical systems approach*, has been widely applied to human movement studies and in other areas of cognitive science. Compared to the traditional cognitive approach that assumes internal computation and prescription in the brain, dynamical systems approach focuses more on interactions between the body (including the brain), environment, and task. Movement or coordination patterns are then theorized as emerging through interactions among several constraints within organism, environment, and task (Davids, Button, & Bennett, 2008; Newell, 1986). The large number of DoFs in an embodied system can be reduced in order to satisfy these constraints. Thus, the functional structure (i.e., *synergy*) to achieve the task emerges as a particular coordination pattern. The individual organism regulates its behavior in order to satisfy the task demands that must be performed in a specific environment.

Indeed, in the framework of *synergetics* (Haken, 1978), a self-organization theory, a system's low-dimensional spatial temporal patterns at the macroscopic level emerge through interactions among components (DoFs) at the microscopic level under constraints from the system's environment or embedded context. The macroscopic pattern also constrains the components' behavior at the microscopic level to keep the pattern stable (Kelso, 1995). Kelso and his colleagues applied *synergetics* to the modelling of bimanual coordination (Haken, Kelso, & Bunz, 1985). The model can describe the qualitative change of pattern within a system by using the concepts of *synergetics*. Rhythmic coordinated behaviors, such as inter-limb coordination, can be modeled as a motion equation (Haken et al., 1985). Recently, dynamical systems approach has not only provided a theoretical framework but also obtained evidence by analytical tools (e.g., fractal analysis and recurrence analysis) based on nonlinear dynamics theory (Holden, Riley, Gao, & Torre, 2013; Van Orden & Riley, 2005). The present study also applies these frameworks and techniques to investigate dynamic balancing skill in slacklining.

Current hypotheses

Based on slackline instructors' experience, here, we propose three hypotheses related to the skill of single-leg standing on a slackline. To maintain whole-body balance in an unstable environment (i.e., on the slackline), the overall task is regarded as keeping the center of mass (COM) above the base of support (i.e., the place where one's foot contacts the line)(Shumway‐Cook & Woollacott, 2013). To satisfy this task demand, components of the embodied system are supposed to self-organize. The hypothesized control strategy proceeds as follows:

1) in a horizontal direction, one should raise both hands high and coordinate them in parallel in order to regulate the COM's position above the line;
2) in the vertical direction, one should flexibly bend his/her knee to compensate for the line's fluctuations;
3) in the anteroposterior direction, one should maintain a straight back to keep the center of gravity vertically balanced over the sole of the foot standing on the line.

These three hypotheses are depicted by the three pictures in Figure 1. In the current article, we focused on the first hypothesis. As shown in Figure 1 (4), we hypothesize that fluctuations from the line might be canceled or absorbed by compensating strategy between both hands and the COM positions. According to this hypothesis, if the position of

COM moves to right, coupled both hands move to left so as to counter-balance. Such strategy enables one to regulate whole-body balance flexibly using two DoFs (both hands and COM positions) against the line fluctuations. We suppose that such a compensating relation realizes a functional unit as a *synergy* (Latash, 2008). Furthermore, at this moment, bimanual coordination is supposed to play an important role, because if left and right hands move individually, both hands' position cannot compensate the COM's position efficiently. So we consider bimanual coordination in terms of coupled oscillator that can be regarded as a single system and its dynamic stability as the hypothesized variable to test the first hypothesis. The current article reports partial results of testing it.

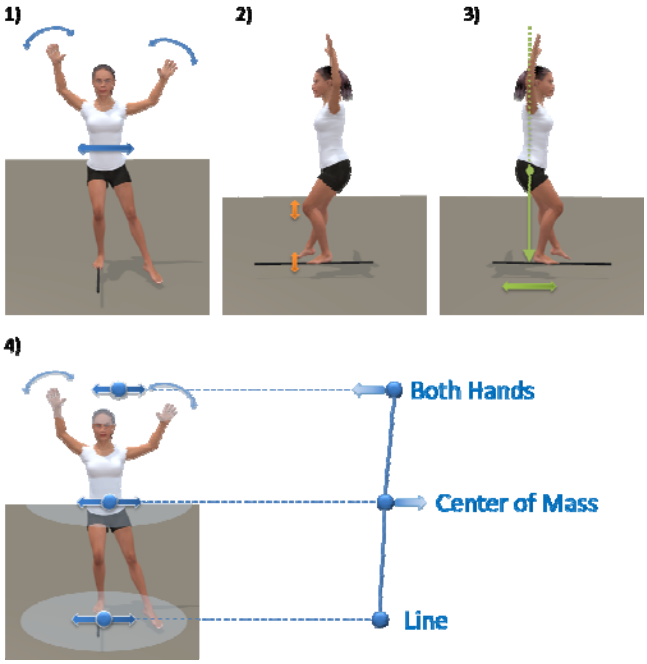


Figure 1 Current hypotheses: regulating whole-body posture in the 1) horizontal, 2) vertical, 3) anteroposterior direction, 4) compensational relation between both hands and COM against the line.

Method

Participants

Two participants, an experienced player (age 40, male) and a novice player (age 30, male), were recruited to make possible comparisons between slacklining performances at different skill levels. The experienced player had more than three years of experience in slacklining, whereas the novice player had just started slacklining. The experimental procedures were approved by the research ethics committee of Kanagawa University, where the experiment was conducted. Each participant provided informed consent to participate in this study.

Apparatus

Slacklining was performed on SLACKRACK300 (GIBBON SLACKLINES, 300 cm length, 30 cm height). A 3D motion capture system (OptiTrack V120: Trio, NaturalPoint, Inc.) was used to measure participants' body movement (sampling frequency: 120 Hz). Nine reflective markers ware attached to the top of the head, the front of the head, the tips of the index fingers, the COM (around the hip), the tops of the knees, and the ankles. Three-dimensional time series data of each marker were smoothed by a second-order Butterworth low-pass filter with a 12 Hz cutoff frequency.

Procedure

The experimental task was to perform single-leg standing while on a slackline. Participants were required to maintain balance on the slackline for as long as possible, using their preferred leg. Each trial lasted for three minutes, including rest breaks. Participants underwent five trials.

Data Analysis

To evaluate single-leg standing skill, we calculated persistence time, or how long the participant could stay on the line before falling off, by the following procedure. If he maintained the task for at least five seconds, it was regarded as a *success*; if he fell off before five seconds had passed, that effort was a *failure*. For each instance of success, the length of time on the line was measured. It was expected, of course, that the experienced player would have a longer persistence time than the novice.

To analyze 3D motion capture data quantitatively, time series data were processed as follows. Among the instances of success, those when the player stayed on the line for at least 15 seconds were defined as *steady*. The first five and the last five seconds of each steady instance were omitted because participants were more likely to have balance difficulties in the initiating and ending phases of each attempt, with the result that the data from these phases were often not steady. Finally, the remaining time series data on the steady attempts were divided into five seconds section. After we had collected all data satisfying the above conditions, the data were smoothed and analyzed by the following nonlinear time series analysis method.

To quantify stability of movement (i.e., bimanual coordination), *cross-recurrence quantification analysis* (CRQA; Zbilut, Giuliani, and Webber, 1998) was applied. This is a nonlinear time series analysis method that captures the recurring properties and patterns of a dynamical system that results from two streams of information interacting over time (Zbilut et al., 1998) and quantifies how similarly the two observed data series unfold over time (Shockley, 2005). Recurrence quantification analysis was originally developed to uncover subtle time correlations and repetitions of patterns, and it is relatively free of assumptions about data size and distribution (Zbilut & Webber, 1992). In CRQA, two time-delayed copies of the original time series were used for embedding the data in higher dimensional space,

reconstructing the phase space, to analyze the recurrent structure between them (Zbilut et al., 1998).

Within the framework of dynamical systems approach, for inter-limb rhythmic coordination, two CRQA measures are regarded as significant indexes of stability of movement (Pellecchia, Shockley, & Turvey, 2005; Shockley, 2005). The percent recurrence (%REC) in CRQA corresponds to the ratio of the actual number of shared locations to the number of possible shared locations in phase space. It provides an index of the magnitude of stochastic noise in the system (Pellecchia et al., 2005); a higher %REC indicates less noise in the system. The other measure was related to the line structure calculated from the recurrence plot, Maxline (MAXL). MAXL is the longest shared trajectory in phase space and the length of the maximum diagonal line on the plot (Webber & Zbilut, 2005). MAXL is a measure of the stability of the shared activity (Shockley, 2005). It is supposed to provide an index of the system's sensitivity to perturbations (i.e., the strength of the attractor against perturbations) (Pellecchia et al., 2005).

We performed CRQA using the R package "crqa" (version 1.0.6) (Coco & Dale, 2014) after determining the optimal values for the input parameters (e.g., time delay, embedding dimensions, radius) with reference to the standard guidelines of the RQA method (Webber & Zbilut, 2005).Consequently, we chose parameters of 20 for time delay, 3 for embedding dimensions, and 25 for radius (with z-score normalization and maximum distance rescaling; Webber & Zbilut, 2005).

Results

Figure 2 presents the persistence time in the single-leg standing task for each participant. The average persistence time was longer for the experienced player (107.25 sec) than for the novice player (20.39 sec), reflecting the experienced player's far greater skill.

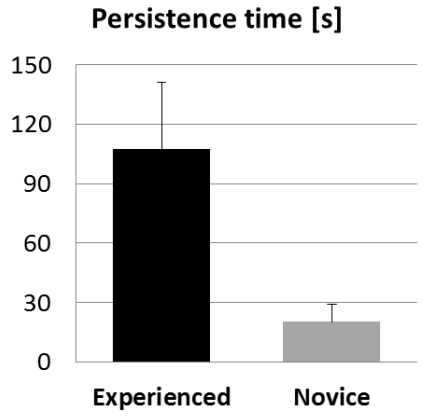


Figure 2 Persistence time in the single-leg standing task (Error bar: Standard deviation)

Figure 3 represents sample raw data from 20-second time series of both hands in the horizontal direction for each participant: top, experienced player, bottom, novice player. The x- and y-axes represent time [sec] and position [m], respectively. Black rigid and dashed lines represent left- and right- hand motion, respectively. Blue rigid line shows the COM's motion.

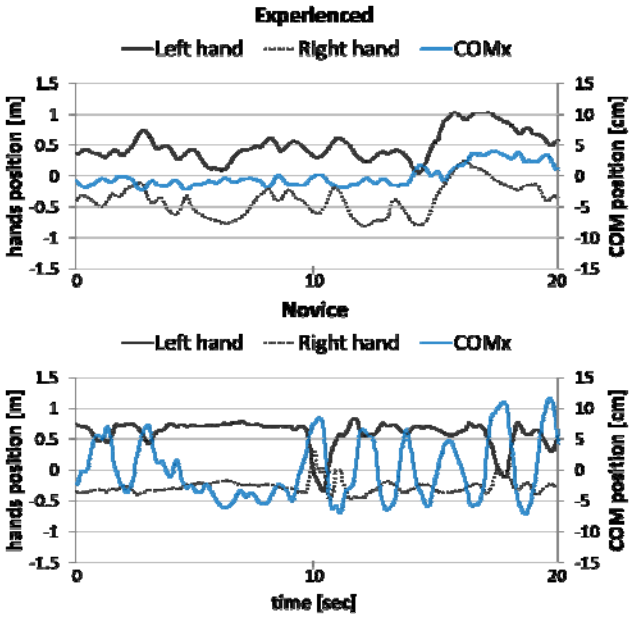


Figure 3 Sample 20 sec. time series of both hands for the two participants (top: experienced player, bottom: novice)

Figure 4 shows %REC for each participant in bimanual coordination. %REC was higher for the experienced player (22.95%) than for the novice player (17.01%). Figure 5 shows MAXL for each participant in bimanual coordination. MAXL was longer for the experienced player (126.37) than for the novice player (70.67).

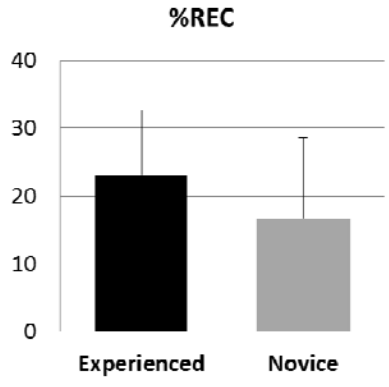


Figure 4 Percent recurrence for each player (Error bar: Standard deviation)

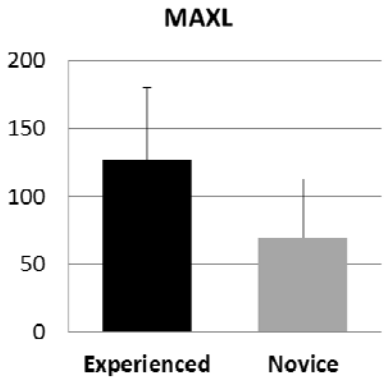


Figure 5 Maxline for each player
(Error bar: Standard deviation)

Discussion

As shown in Figure 2, the experienced player had greater single-leg standing skill than the novice player. This result is consistent with the prediction that persistence time should be longer for high-level than for low-level players. It confirms that the two players were at different skill levels.

As shown by the sample time series of both hands (Figure 3), it seems that the experienced player could maintain a more stable bimanual coordination pattern in parallel than the novice player. In other words, the experienced player can move their hands in the same way keeping a constant relation between hands. Whereas their hands move in the opposite way to the COM motion as if compensating each other or counter-balancing. Although bimanual movement of the experienced player seems to be apparently larger than that of the novice player, it moves more rhythmically or oscillatory than that of the novice player. In other words, bimanual movement of the experienced player seems to be more dynamically stable. On the other hand, the novice player could not consistently maintain a particular bimanual coordination pattern; in fact, he broke his bimanual coordination pattern and crossed the hands twice. Bimanual movement of the novice player seems to move less rhythmically and be less dynamically stable. The COM position of the novice player is also more variable than that of the experienced player. As a result of comparing two participants' behaviors, our hypotheses on compensating relation between both hands and COM positions (Figure 1) seems to be supported as far as observed in sample time series of Figure 3.

The CRQA for bimanual coordination found that %REC and MAXL were greater in the experienced player than in the novice player. These results indicate that bimanual coordination during the single-leg standing task was more stable in the experienced player than in the novice player in terms of both the magnitude of noise in the system and the system's sensitivity to perturbations. This fact supports our first hypothesis. Further experiments should be conducted to obtain more samples and test the hypothesis quantitatively. It will also be necessary to investigate the relation between bimanual coordination and whole-body dynamic balancing by quantifying the coupling between bimanual coordination and the line as well as the actual COM horizontal position. Although the causal relation between bimanual coordination and whole-body balancing is also important and should be examined, it might not be so simple. It is difficult to reveal whether bimanual coordination causes whole-body balancing and otherwise because such a phenomenon might have an emergent property that cannot be reduced to a particular factor simply.

Although our pilot study is a case study recruiting only two participants, the results suggest that dynamic balancing skill on a slackline is correlated with bimanual coordination stability. Thus, our hypothesis seemed to be partially supported. In the future, we intend to examine a specific relationship among the skill and hypothesized variables quantitatively. That is, we wish to investigate whether bimanual coordination contributes to dynamic balancing on a slackline and, if so, how it contributes and how experienced players acquire this skill. We plan to conduct an experiment comparing two groups of different skill levels (i.e., expert vs. novice) with more samples. The future experiment should conduct not only statistical test comparing two groups but also analysis of relation between bimanual coordination and whole-body balancing.

In the practical fields of sports training and rehabilitation, slacklining is expected to improve dynamic balancing skill (Kodama, Yamagiwa, & Kikuchi, 2015). Slacklining requires one to activate inner muscles, that supports a fundamental level of action and involves the regulation of muscle tone (Bernstein, 1996), rather than outer muscles. We suppose that slackline balance training might improve the sensitivity to dynamic changes in the environment and in one's own body in terms of haptic perception based on muscle sense (Carello, Silva, Kinsella-Shaw, & Turvey, 2008). If so, slacklining can contribute to improvement of haptic perception system through the body. Even though this notion is a matter of speculation, if the present study can reveal the skill for slacklining, we believe it can have implications for the practical fields regarding how to develop dynamic balancing skill based on sensitive haptic perception system.

Conclusion

The purpose of this study was to reveal the fundamental skills required for slacklining. We examined a hypothesis regarding the dynamic balancing skill for slacklining. In the practice of slackline training, instructors impart their knowledge of how to develop this skill to learners, such as by moving their hands in parallel. We hypothesized that bimanual coordination in the horizontal direction might

contribute to dynamic balancing on a slackline. In our pilot study, participants at different skill levels were required to maintain single-leg standing on a slackline for as long as possible. The dynamic stability of bimanual coordination was assessed by nonlinear time series analysis (i.e., cross-recurrence quantification analysis) and compared between participants. As expected, persistence time on the line was longer in the experienced player than in the novice player. Also, bimanual coordination stability was higher in the experienced player than in the novice player, in terms of both the magnitude of stochastic noise in the system (i.e., %Recurrence) and the system's sensitivity to perturbations (i.e., Maxline). These results suggest that skilled performance of single-leg standing on a slackline is correlated with bimanual coordination stability. We suppose that bimanual coordination contributes to whole-body dynamic balancing on a slackline. Further investigations are expected to clarify this notion in the future.

References

Balasubramaniam, R., & Wing, A. M. (2002). The dynamics of standing balance. *Trends in Cognitive Sciences*, 6(12), 531–536.

Bernstein, N. A. (1967). *The Co-ordination and regulation of movements*. Pergamon Press Ltd.

Bernstein, N. A. (1996). *Dexterity and Its Development*. Psychology Press.

Carello, C., Silva, P. L., Kinsella-Shaw, J., & Turvey, M. T. (2008). Muscle-based perception: theory, research and implications for rehabilitation. *Revista Brasileira de Fisioterapia*, 12(5), 339–350.

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*, 5, 510.

Davids, K., Button, C., & Bennett, S. (2008). *Dynamics of Skill Acquisition: A Constraints-led Approach*. Human Kinetics.

Delignières, D., Torre, K., & Bernard, P.-L. (2011). Transition from persistent to anti-persistent correlations in postural sway indicates velocity-based control. *PLoS Computational Biology*, 7(2), e1001089.

Granacher, U., Iten, N., Roth, R., & Gollhofer, a. (2010). Slackline training for balance and strength promotion. *International Journal of Sports Medicine*, 31(10), 717–723.

Haken, H. (1978). *Synergetics: an introduction : nonequilibrium phase transitions and self-organization in physics, chemistry and biology*. Springer-Verlag.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51(5), 347–356.

Holden, J. G., Riley, M. A., Gao, J., & Torre, K. (Eds.). (2013). *Fractal Analyses: Statistical and Methodological Innovations and Best Practices*, Frontiers E-books.

Horak, F. (1989). Components of postural dyscontrol in the elderly : a review . Neurobiol Aging Components of Postural Dyscontrol in the Elderly : A Review. *Neurobiology of Aging*, 10(6), 727–738.

Huber, P., & Kleindl, R. (2010). A case study on balance recovery in slacklining. *ISBS-Conference Proceedings Archive*, (1990), 1–4.

Keller, M., Pfusterschmied, J., Buchecker, M., Müller, E., & Taube, W. (2012). Improved postural control after slackline training is accompanied by reduced H-reflexes. *Scandinavian Journal of Medicine and Science in Sports*, 22(4), 471–477.

Kelso, J. A. S. (1995). *Dynamic Patterns: The Self-organization of Brain and Behavior*. MIT Press.

Kodama, K., Yamagiwa, H., & Kikuchi, Y. (2015). Effects of slackline training on dynamic postural balancing. *Proceedings of Second International Workshop on Skill Science*, 48–49.

Latash, M. L. (2008). *Synergy*. Oxford University Press,.

Newell, K. M. (1986). Constraints on the development of coordination. *Motor Development in Children: Aspects of Coordination and Control*, 34, 341-360.

Pellecchia, G. L., Shockley, K. D., & Turvey, M. T. (2005). Concurrent cognitive task modulates coordination dynamics. *Cognitive Science*, 29(4), 531–57.

Riley, M. a, & Turvey, M. T. (2002). Variability of determinism in motor behavior. *Journal of Motor Behavior*, 34(2), 99–125.

Shockley, K. D. (2005). Cross recurrence quantification of interpersonal postural activity. In M. Riley & G. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 142–177).

Shumway‐Cook, A., & Woollacott, M. H. (2013). *Motor Control: Translating Research Into Clinical Practice*. Lippincott Williams & Wilkins.

Turvey, M. T. (1990). Coordination. *American Psychologist*, 45(8), 938–953.

Van Orden, G. C., & Riley, M. A. (Eds.). (2005). *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. National Science Foundation.

Webber, C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. In M. Riley & G. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 26–94).

Winter, D. (1995). Human balance and posture control during standing and walking. *Gait & Posture*, 3(4), 193–214.

Zbilut, J. P., Giuliani, A., & Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A*, 246(1-2), 122–128.

Zbilut, J. P., & Webber, C. L. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3-4), 199–203.

Causality, Normality, and Sampling Propensity

Thomas F. Icard (icard@stanford.edu)
Philosophy & Symbolic Systems, Stanford University

Joshua Knobe (joshua.knobe@yale.edu)
Cognitive Science & Philosophy, Yale University

Abstract

We offer an account of the role of normality—both statistical and prescriptive—in judgments of actual causation. Using only standard tools from the literature on causal cognition, we argue that the phenomenon can be explained simply on the assumption that people stochastically sample (counterfactual) scenarios in a way that reflects normality. We show that a formalization of this idea, giving rise to a novel measure of causal strength, can account for some of the most puzzling qualitative patterns uncovered in recent experimental work.

Introduction

Judgments of *actual causation*—concerning the extent to which a given event or factor *caused* some outcome on a particular occasion—have been at the center of attention in work on causal cognition. One intriguing phenomenon that has long been recognized is that people's judgments of actual causation can be influenced by the degree to which they regard certain events as *normal*. In recent years, this effect has been explored both in experimental studies (Kominsky et al., 2015; Phillips et al., 2015) and in formal models (e.g., Halpern and Hitchcock 2015).

In this paper we propose a novel explanation of the role of normality in causal cognition by appeal to the idea that many cognitive processes—including those underlying causal judgments—can be understood as involving *probabilistic sampling* from some underlying distribution. In short, causal strength is assessed in part by stochastically generating alternative ("counterfactual") scenarios and using these scenarios to determine the extent to which some event is causally relevant to a given outcome. This hypothesis, together with a further assumption—that the distributions from which these samples are drawn directly reflect prescriptive normality (as well as statistical factors)—forms the core of our explanation.

To explore this possibility, we begin by summarizing the effects to be explained and describing at an informal level how these effects could be explained in terms of sampling. Then we turn to the details of a natural sample-based algorithm and show that the *measure of actual causal strength* corresponding to this algorithm, unlike other causal strength measures offered in the literature, would generate precisely the effects observed in existing studies.

Three Effects of Normality on Actual Causation Judgments

To begin with, we need to distinguish two kinds of norms. First, there are purely *statistical norms*. For example, it is a statistical fact that winter months in Oregon tend to be cloudy and overcast, so if Oregon ever had a sunny winter, that weather could be said to be violating a statistical norm. Second, there are *prescriptive norms*. These norms are constituted not by purely statistical tendencies but by the way things ought to be or are supposed to be. Suppose we believe that the police ought to accord criminal defendants certain rights. Even if we do not believe that the police actually do tend to accord defendants these rights, we might think that failing to do so is a violation of a prescriptive norm.

A question then arises as to which of these two types of norms impacts people's judgments of actual causation. As we will see below, existing research suggests that actual causation judgments are influenced by both kinds of norms. More strikingly, the impact of these two kinds of norms shows precisely the same pattern. As a result, researchers have suggested that it might be helpful to posit a single undifferentiated notion of normality that integrates both statistical and prescriptive considerations (Halpern and Hitchcock, 2015; Kominsky et al., 2015). On this approach, an event counts as "abnormal" to the extent that it either violates a statistical norm or violates a prescriptive norm, and as "normal" to the extent that it follows both of these types of norms. Difficult questions arise about precisely how statistical and prescriptive considerations are integrated into an undifferentiated notion, but we will not be resolving those questions here. Instead, we focus on three ways in which normality impacts people's intuitions about actual causation.

First Effect: Abnormal Selection

We will eventually be introducing a formal framework to describe this effect more precisely, but for the moment, we can offer the following rough characterization:

> *In cases where an outcome depends on a causal factor C, people will be more inclined to say that C caused the outcome when they regard C as abnormal than when they regard C as normal.*

This basic effect appears to arise both for statistical norms and for prescriptive norms.

First, it has been known for decades that actual causation judgments can be influenced by statistical norms (Hilton and Slugoski, 1986). Suppose that a person leaves a lit match on the ground and thereby starts a forest fire. In such a case, the fire would not have begun if there had been no oxygen in the atmosphere, and yet we would not ordinarily say that the oxygen caused the fire. Why is this? The answer appears to be that it is so normal for the atmosphere to contain oxygen. (Our intuitions would be very different if matches were struck

on a regular basis but there was never a fire except on the very rare occasions when oxygen was present.)

Strikingly, this same effect arises for prescriptive norms. Consider the following case:

The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly e-mailed them reminders that only administrators are allowed to take the pens.

On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk.

Faced with this case, participants tend to say that the professor caused the problem (Knobe and Fraser, 2008; Phillips et al., 2015). But now suppose that we change the first paragraph of the case in such a way as to make the professor's action not violate a prescriptive norm:

The receptionist in the philosophy department keeps her desk stocked with pens. Both the administrative assistants and the faculty members are allowed to take the pens, and both the administrative assistants and the faculty members typically do take the pens. The receptionist has repeatedly e-mailed them reminders that both administrators and professors are allowed to take the pens.

Faced with this latter version, participants are significantly less inclined to say that the professor caused the problem (Phillips et al., 2015). Yet the two cases do not appear to differ from the perspective of purely statistical normality; the difference is rather in the degree to which the agent violates a prescriptive norm. The result thereby suggests that prescriptive norms impact causal judgments.

This phenomenon has been explored in a wide range of studies (Cushman et al., 2008; Roxborough and Cumby, 2009; Samland et al., 2016), and the results strongly suggest that the effect really does involve prescriptive considerations and cannot be reduced to a matter of purely statistical norms. First, one can explicitly pit the prescriptive against the statistical. In one study, participants were told that administrative assistants were allowed to take pens and faculty members were not (a prescriptive norm) but that in actual fact administrative never did take pens while faculty members always did (a statistical norm). People's judgments ended up being affected more by the prescriptive than by the statistical, with participants tending on the whole to say that the administrative assistant did not cause the problem but the faculty member did (Roxborough and Cumby, 2009). Second, one can look at cases in which different people have different prescriptive judgments. For example, one paper looked at controversial political issues (abortion, euthanasia) and found that people who had opposing moral judgments about these issues arrived at correspondingly opposing causal judgments about people who performed the relevant actions (Cushman et al., 2008).

At this point, it is widely agreed that prescriptive considerations do indeed impact people's causal judgments. The remaining questions are about how to explain this. Existing accounts invoke everything from conversational pragmatics to motivational bias (see Livengood and Rose 2016 for an overview of the literature). Rather than introducing some additional component above whatever is used to account for causal judgments, our account will derive the normality effects as a straightforward consequence of the way counterfactuals are generated to determine causal strength.

Second Effect: Supersession

Supersession is an effect whereby the normality of one factor can actually influence the degree to which other factors are regarded as causes. The effect can be characterized roughly as follows:

Consider cases in which an outcome depends on two different factors C and A, such that the outcome will only occur if both C and A occur. Then people will be less inclined to say that C caused the outcome if A is abnormal than if A is normal.

In other words, it is not just that a given factor is regarded as more causal when it is abnormal; a factor will be also be regarded as more causal when other factors are normal. This effect too arises for both statistical and prescriptive norms. Turning first to statistical norms, consider the following:

Alex is playing a board game. On every turn of the game, two six-sided dice are rolled and a coin is flipped. Alex will either win or lose the game on his next turn.

Alex will only win the game if the total of his dice roll is greater than 2 AND the coin comes up heads. It is very likely that he will roll higher than 2, and the coin has equal odds of coming up heads or tails.

Alex flips the coin and rolls his dice at exactly the same time. The coin comes up heads, and he rolls a 12, so as expected, he rolled greater than 2. Alex wins the game.

Now contrast that with a case in which the second paragraph is slightly modified:

Alex is only win the game if the total of his dice roll is greater than 11 AND the coin comes up heads. It is very unlikely that he will roll higher than 11, but the coin has equal odds of coming up heads or tails.

The difference between these two cases is solely in the normality of the dice roll. (The success of the dice roll is statistically normal in the first case, statistically abnormal in the second.) Yet this difference actually leads to a change in the degree to which people regard the coin flip as a cause. Participants were significantly less inclined to say that Alex won because of the coin flip when the dice roll was abnormal than when it was normal (Kominsky et al., 2015).

This same effect then arises for prescriptive norms. In one study, participants were asked to imagine a motion detector that goes off whenever two people are in the room at the same

time. Suzy and Billy enter the room at the same time, and the motion detector goes off. In one condition, Billy is supposed to be in the room, while in the other condition, he is specifically not supposed to be in the room. Suzy was judged to be significantly less a cause of the detector going off when Billy violated the prescriptive norm than when he acted in accordance with the prescriptive norm (Kominsky et al., 2015).

Third Effect: No Supersession with Disjunction

The supersession effect described above arises in cases where the causal structure is conjunctive. For disjunctive structures we find a quite different pattern:

> *Consider cases in which an outcome depends on two different factors C and A, such that it will only occur if either C or A occurs. Then people are just as inclined to say that C caused the outcome when A is abnormal as they are when A is normal.*

Existing sudies have put this claim to the test by comparing disjunctive cases to conjunctive cases and looking for an interaction whereby manipulations of normality do not have the impact in disjunctive cases that they do in conjunctive ones. This interaction arises both for statistical norms and for prescriptive norms (Kominsky et al., 2015).

For statistical norms, we can see the effect by looking at the case of the coin flip and dice roll described above. One can simply modify the rules described in that case so that Alex wins if he succeeds either on the coin flip or on the dice roll. When the rules are changed in this way, the supersession effect disappears. Participants are just as inclined to see the coin flip as causal in the case where the dice roll is abnormal as they are in the case where the dice roll is normal.

Precisely the same result then arises for the prescriptive norm case with the motion detector. When participants are told that the motion detector will go off if at least one person is in the room, the supersession effect again disappears. Participants are just as inclined to see Suzy as the cause when Billy's act violates a prescriptive norm as when it does not.

Summary

Across three different effects, prescriptive norms appear to be having the same impact as statistical norms. If there had only been an impact of statistical norms, one obvious approach would have been to explain that impact in terms of something specific to the statistical case in particular—though even that would be difficult (see below)—but given that these effects appear to arise for both kinds of norms, it seems that we need a unified explanation that can be applied to both.

Sampling Propensities

The idea that various mental operations can be described in terms of probabilistic sampling processes has generated much excitement over the past several years (see, e.g., Griffiths et al. 2012; Icard 2016 for overviews). Our focus in this paper is on the role of sampling in causal cognition. We follow a long line of work that proposes, in judging whether some event C caused E, we must consider various counterfactual scenarios involving C and E (Lewis, 1973). In short, the sampling hypothesis in the causal domain proposes that these counterfactual scenarios will be selected and evaluated stochastically by a sampling-like process.

Different researchers have spelled out the counterfactual approach to causality in quite different ways. The most important point to emphasize is one that is common to many such implementations. This is that, in assessing any given causal claim, a wide variety of different counterfactual scenarios will be relevant. It is generally agreed that, at a minimum, one must check whether the cause C was in some sense *necessary* for effect E to happen. Necessity is usually thought to correspond to a counterfactual claim of the form: Had C not occurred, E also would not have occurred. Clearly there are many possible ways an event C might not occur, and theories of causation will often attempt to specify exactly which such non-C scenarios are relevant to the causal claim.

Dual to necessity, a number of researchers have proposed that judgments of actual causation also involve a notion of *sufficiency* (Woodward, 2006; Lombrozo, 2010): Given that C in fact occurred, the outcome E still would have occurred even if background conditions had been slightly different. To the extent that this counterfactual sufficiency claim does not hold, that generally counts against the causal strength of C on outcome E. Again, there are clearly many ways the background conditions might have been different.

Thus, independent of any particular proposal about exactly which counterfactuals are relevant, it is clear that there may in general be very many relevant counterfactual scenarios to consider, certainly more than any person could plausibly assess in real time. We propose that people solve this intractability problem by probabilistically *sampling* possible states of the world (for related ideas see Lucas and Kemp 2015, Gerstenberg et al. 2014). It is not as though there is some fixed set of counterfactual scenarios one must evaluate; rather, on any given occasion one will stochastically consider various alternative sequences of events, and thereupon make a judgment of causal strength.

Given therefore that a person will generate counterfactual scenarios with different probabilities, and will evaluate those scenarios probabilistically, the interesting question is how to understand these *sampling propensities*. In as far as sampling algorithms are typically used to approximate probabilistic calculations, it would make sense for these counterfactual sampling propensities to reflect the perceived environmental statistics. Indeed, one can interpret much of the recent work on sampling in cognition, including in causal cognition, as proposing that some species of subjective probability is indirectly represented by sampling propensities. However, the distinctive feature of our proposal in this paper is that sampling propensities are also proportional to *prescriptive* normality. Thus for instance, in the detector scenario when judging whether Suzie caused the alarm, if Billy is prohibited from being in the room, people will be more likely to con-

sider counterfactual situations in which he did not enter.

It bears emphasis that sampling propensities, on this view, need not (only) encode the subject's uncertainty. It has already been suggested recently that people's sampling distributions in some contexts might come apart from reasonable judgments of probability or likelihood (see Lieder et al. 2014; Icard 2016), e.g., in a way that favors practical rationality (making the right choice) at the expense of theoretical rationality (having true or accurate beliefs). But some aspects of sampling propensity might be altogether divorced from probabilistic judgment. Consider, e.g., what might happen if we are observing an agent going through a series of steps in an attempt to solve a difficult puzzle. As we observe her taking each step, we consider other possible steps she might have taken. In such a case, we might be especially likely to consider other possible steps that would have been *good* ways of solving the puzzle (Kahneman and Miller, 1986). We consider these possibilities not because we explicitly represent the agent as having a high probability of taking them but, perhaps, because we are interested in finding the solution and are drawn to consider the best ways of accomplishing that task.

Similar remarks apply to the case we originally introduced to illustrate the concept of a prescriptive norm. Suppose we believe that the police should accord criminal defendants certain rights but that, as a purely statistical matter, they almost never do accord those rights. Now suppose we are observing a case in which the police fail to accord a defendant her rights. In such a case, we might consider possibilities in which the police do accord the defendant these rights. We would be drawn to consider those possibilities not because we explicitly regard them as probable but, perhaps, rather because we see them as instantiating a moral ideal.

Existing research provides some support for this hypothesis. When participants are given a vignette and asked to provide a counterfactual, they are more likely to mention possibilities they regard as statistically frequent (Kahneman and Miller, 1986). Furthermore, they are also more likely to mention possibilities they regard as prescriptively good (McCloy and Byrne, 2000). In addition, when participants are given a counterfactual and asked to rate the degree to which it is relevant or worth considering, they are more inclined to rate a possibility as relevant to the extent that it conforms to prescriptive norms (Phillips et al., 2015). These findings provide at least some initial evidence in favor of the claim that people are drawn to consider possibilities that do not violate prescriptive norms. The idea that such inclinations would similarly play into the hypothesized sampling propensities underlying causal and other judgments does not seem implausible.

Significantly, the mix of statistical and prescriptive norms that we find in people's actual causation judgments has also arisen in a number of other areas. For example, one finds a similar effect in people's intuitions about intentional action and about freedom (Phillips et al., 2015). Existing attempts to explain these effects have suggested that they might be due in some way to people's tendency to regard possibilities as more relevant when those possibilities accord with norms, though this basic approach has been spelled out within a number of different formal frameworks (e.g., Knobe and Szabó 2008; Halpern and Hitchcock 2015). If all of these phenomena can be elegantly explained in terms of sampling, this would provide strong evidence in favor of the present hypothesis.

Ultimately, however, the real test of this hypothesis is whether, when combined with further assumptions, it can accurately predict the patterns of people's causal judgments.

The Account

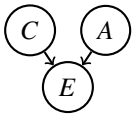
We formalize the notions of sufficiency and necessity in a straightforward way using Bayes nets and causal interventions (Pearl 2009), where we assume that the underlying probabilities reflect normality judgments as discussed in the previous section. That will affect not only the sufficiency/necessity judgments themselves, but also the probabilities with which sufficiency and necessity are assessed, effectively giving a weighting between them. This plays an important role in our account of abnormal selection.

Intervening on a Bayes net $\mathcal{N} = \langle \mathcal{G}, P \rangle$ involves setting some variable X to a specific value x. This gives rise to a new mutated Bayes net $\mathcal{N}_{X=x} = \langle \mathcal{G}_{X=x}, P_{X=x} \rangle$, where in the graph $\mathcal{G}_{X=x}$ we cut all links to the node representing X, so that it has no parents, and in $P_{X=x}$ we have $P_{X=x}(X = x) = 1$, leaving everything else the same. We can then use this to infer what would have happened under various counterfactual suppositions, including in cases where we nonetheless want to keep some observations from the actual situation fixed. That is, we can also use $P_{X=x}(Y \mid \vec{Z} = \vec{z})$ to determine counterfactual probabilities of another variable Y, holding fixed that we actually observed $\vec{Z} = \vec{z}$, for example.

We adopt standard notation for interventions and conditionalization. Given a network $\mathcal{N} = \langle \mathcal{G}, P \rangle$ we will write $P\big(Y = y \mid do(X = x)\big)$ for $P_{X=x}(Y = y)$ and similarly for intervention with observations, $P\big(Y = y \mid do(X = x), \vec{Z} = \vec{z}\big) = P_{X=x}(Y = y \mid \vec{Z} = \vec{z})$. We will also use somewhat nonstandard notation for negative intervention: $P\big(Y \mid do(X \neq x)\big) = P_{X \neq x}(Y)$, where $P_{X \neq x}$ is just like P, except that $P_{X \neq x}(X = x) = 0$ and $P_{X \neq x}(X = x')$ is the renormalized probability of $X = x'$, i.e., $\frac{1}{Z} P(X = x')$, with $Z = \sum_{x' \neq x} P(X = x')$.

Desiderata

All of our motivating examples involve a simple 3-node graph (known in the literature as an "unshielded collider" structure):



Assuming the random variables A, C, E are all binary—taking on values 0 and 1—and given a distribution P that factors over this graph, we are interested in two special cases:

CONJUNCTIVE: $P(E \mid C, A) = \min(C, A)$

DISJUNCTIVE: $P(E \mid C, A) = \max(C, A)$

In other words, the conjunctive version has E on (value 1) if both C and A are on, off (value 0) otherwise. This kind of model would describe the scenario with the pens, for example: the receptionist has a problem ($E = 1$) just in case both the administrator takes a pen ($C = 1$) and the professor takes a pen ($A = 1$). By contrast, the disjunctive version has E on if at least one of C or A is on. This describes the disjunctive scenarios: e.g., the motion detector goes off just in case either Billy enters the room or Suzy enters the room.

To account for the effects we need a definition of *causal strength*. Suppose we have a functional $\kappa_P(C,E)$ that measures the strength of cause $C = 1$ on effect $E = 1$ under distribution P. We can then rewrite our desiderata in a slightly more formal manner. Suppose that P_1 and P_2 are two distributions such that $P_1(C) = P_2(C)$, but that $P_1(A) > P_2(A)$. Then depending on whether P_1 and P_2 are conjunctive or disjunctive, we should expect different patterns:

ABNORMAL SELECTION: In the conjunctive case, $\kappa_{P_1}(A,E) < \kappa_{P_2}(A,E)$.
SUPERSESSION: Again, in the conjunctive case, $\kappa_{P_1}(C,E) > \kappa_{P_2}(C,E)$.
NO SUPERSESSION WITH DISJUNCTION: In the disjunctive case, $\kappa_{P_1}(C,E) = \kappa_{P_2}(C,E)$.

Again, we assume that probability in a Bayes net positively correlates not only with perceived statistical likelihood, but also with prescriptive normality. Thus, if event A is considered more normal in situation 1 than it is in situation 2, then we would expect $P_1(A) > P_2(A)$. So, under our interpretation of the probabilities as sampling propensities, these renderings capture the more informal descriptions of the effects

Necessity and Sufficiency

To determine *actual necessity* of $X = x$ for $Y = y$, we identify a path—a so called *active path*—from X to Y, and freeze all other variables $\vec{Z}$ outside the path to specific values $\vec{z}$. We then intervene to set $X \neq x$ and check whether nonetheless $Y = y$.

There are various proposals in the literature for how to choose $\vec{Z}$ and $\vec{z}$ (see, e.g., Halpern and Pearl 2005). For our purposes concerning the unshielded collider, all variables outside the (single) path will be fixed to their actual values. But in the general case, supposing we adopt some method for selecting $\vec{Z}$ and $\vec{z}$ in arbitrary models, we propose:

NECESSITY STRENGTH: $P\big(Y \neq y \mid do(X \neq x), \vec{Z} = \vec{z}\big)$

In the simple cases of interest here, the proposal amounts to setting C to 0, holding fixed $A = 1$, and checking whether this is enough to make $B = 0$; and likewise for assessing necessity strength of A. Thus, in the conjunctive model, the necessity strength of $C = 1$ is just $P\big(B = 0 \mid do(C = 0), A = 1\big) = 1$, and the necessity strength of $A = 1$ is also 1. In the disjunctive model, the necessity strengths are both 0 if in fact the other was present, while they are both 1 if the other was absent.

To determine *sufficiency* of $X = x$ for $Y = y$, we intervene to set $X = x$, forgetting everything about the actual situation, and sample forward to determine whether $Y = y$. That is:

SUFFICIENCY STRENGTH: $P\big(Y = y \mid do(X = x)\big)$

Again, in the unshielded collider model this value is quite simple. In the conjunctive case, the sufficiency strength of C is $P(A)$, for A it is $P(C)$. In the disjunctive case both are 1.

A Measure of Actual Causal Strength

Suppose X and Y are binary variables.[1] Then we can define a simple algorithm for determining causal strength of $X = 1$ on $Y = 1$ as in Figure 1. Intuitively, we simulate an X-situation,

1. Initialize $N = 0$, and for $k \leq K$:
   (a) Sample a value $X^{(k)}$ from $P(X)$.
   (b) If $X^{(k)} = 1$, draw $Y^{(k)}$ from $P\big(Y \mid do(X = 1)\big)$. Let $N = N + Y^{(k)}$.
   (c) If 0, draw $Y^{(k)}$ from $P\big(Y \mid do(X = 0), \vec{Z} = \vec{z}\big)$. Let $N = N + (1 - Y^{(k)})$.
2. Return N/K.

Figure 1: Algorithm for Determining Causal Strength

and depending on the value we sample for X, we either test for necessity or suffiency by simulating a Y-situation. It is easy to see that as $K \to \infty$, the fraction N/K converges to the following, which we take to be our measure of causal strength:

$$\begin{aligned}\kappa_P(X,Y) \quad = \quad & P(X = 1)\,P\big(Y = 1 \mid do(X = 1)\big) \; + \\ & P(X = 0)\,P\big(Y = 0 \mid do(X = 0), \vec{Z} = \vec{z}\big).\end{aligned}$$

The causal strength of $X = 1$ is simply the weighted sum of its sufficiency strength and necessity strength, these being weighted by $P(X = 1)$ and $P(X = 0)$, respectively.

It is possible for $\kappa_P(X,Y)$ to be arbitrarily close to 0, e.g., if the necessity strength of X is 0 but $P(X)$ is very small. But it cannot be 0 in any case where in fact $X = Y = 1$, since that would mean $P(X) > 0$ and that X clearly has some sufficiency strength. In general, $\kappa_P(X,Y) \in (0,1]$.

How does this account treat the three effects discussed earlier? Consider supersession. In the scenario with the coin and dice, the robust sufficiency of the coin coming up heads in the conjunctive case depends on how likely the dice are to sum to greater than the threshold. With a low threshold (2) this is quite likely, and hence robust sufficiency and overall causal strength are great; with a high threshold (11) it is unlikely, lessening the overall causal strength. Thus the statistical example goes through easily. The prescriptive case actually has exactly the same structure. Judgments of sufficiency strength for a given agent will depend on how often scenarios are sampled in which the other agent behaves differently. Under the assumption that this is more likely when the other agent violates a prescriptive norm we predict exactly the same pattern.

The other effects are explained similarly. In fact we have:

[1] The generalization is straightforward, if only cumbersome.

Fact 1. Given κ as a measure of causal strength, ABNORMAL SELECTION, SUPERSESSION, and NO SUPERSESSION WITH DISJUNCTION are all guaranteed.

While the precise numerical values will of course depend on the exact sampling propensities, Fact 1 shows that the patterns discussed above will be borne out so long as these probabilities reflect normality in the way that we have proposed.

It is worth emphasizing that no existing accounts of causal strength (e.g., among those surveyed by Fitelson and Hitchcock 2011) satisfy these desiderata. Not one of them can capture NO SUPERSESSION WITH DISJUNCTION. Pearl's (2009) measure of *Probability of Necessity and Sufficiency*, which is most similar to our measure κ, only captures SUPERSESSION. This is significant even if one is only interested in capturing people's causal judgments concerning purely probabilistic/statistical patterns, ignoring prescriptive normality altogether. The sampling view we propose naturally suggests a measure, κ, that easily captures all of these effects.

It is also worth mentioning that this account makes a striking further prediction about the simple 3-node causal structure we have been considering: namely, in disjunctive scenarios, the causal strength of a factor *C* should *decrease* with abnormality. Further work (together with Jonathan Kominksy) has confirmed this prediction in both the statistical and prescriptive cases. See Icard, Kominsky, and Knobe (2016) for details and further discussion of the present hypothesis.

Conclusion

We have offered a concrete model of actual causation judgments that accounts for the effect of normality—both statistical and prescriptive—by appeal only to elements and operations already present in the probabilistic graphical model representation of causal knowledge. While we believe the details of this particular algorithm and strength measure are of independent interest, our broader hypothesis, and central claim, is that the effect of normality on causal judgments can be explained by appeal to two main assumptions: (1) that in judging causal strength people probabilistically sample counterfactual states of the world, and (2) that these sampling probabilities are directly related to normality.

References

Cushman, F., Knobe, J., and Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, 108(1):281–289.

Fitelson, B. and Hitchcock, C. (2011). Probabilistic measures of causal strength. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 600–627. Oxford University Press.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psych. Science*, 24(4):263–268.

Halpern, J. Y. and Hitchcock, C. (2015). Graded causation and defaults. *Brit. J. for Phil. Sci.*, 66(2):413–457.

Halpern, J. Y. and Pearl, J. (2005). Causes and explanations: A structural-model approach. Part 1: Causes. *British Journal for the Philosophy of Science*, 56(4):843–887.

Hilton, D. J. and Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1).

Icard, T. F. (2016). Subjective probability as sampling propensity. *Rev. of Phil. and Psych.* forthcoming.

Icard, T. F., Kominsky, J. F., and Knobe, J. (2016). Causality, normality, and sampling propensity. Manuscript.

Kahneman, D. and Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psych. Rev.*, 94:136–153.

Knobe, J. and Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psych.*, 2:441–448.

Knobe, J. and Szabó, Z. (2008). Modals with a taste of the deontic. *Semantics and Pragmatics*, 6:1–42.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., and Knobe, J. (2015). Causal superseding. *Cognition*, 137:196–209.

Lewis, D. (1973). Causation. *J. of Phil.*, pages 556–567.

Lieder, F., Hsu, M., and Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the 36th Annual Meeting in Cognitive Science*.

Livengood, J. and Rose, D. (2016). Experimental philosophy and causal attribution. In Sytsma, J. and Buckwalter, W., editors, *A Companion to Experimental Philosophy*. Blackwell. Forthcoming.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4):303–332.

Lucas, C. G. and Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, 122(4):700–734.

McCloy, R. and Byrne, R. (2000). Counterfactual thinking and controllable events. *Memory and Cog.*, 28:1071–1078.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Phillips, J., Luguri, J. B., and Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, 145:30–42.

Roxborough, C. and Cumby, J. (2009). Folk psychology concepts: Causation 1. *Philosophical Psych.*, 22(2):205–213.

Samland, J., Josephs, M., Waldmann, M. R., and Raboczy, H. (2016). The role of prescriptive norms and knowledge in childrens and adults causal selection. *Journal of Experimental Psychology: General*. forthcoming.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, pages 1–50.

Explanations in Causal Chains: Selecting Distal Causes Requires Exportable Mechanisms

Jonas Nagel (jnagel1@uni-goettingen.de)
Simon Stephan (sstepha1@uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

When A causes B and B causes C, under what conditions is A a good explanation for the occurrence of C? We propose that distal causes are only perceived to be explanatory if the causal mechanism is insensitive to inessential variations of boundary conditions. In two experiments, subjects first observed deterministic $A \rightarrow B \rightarrow C$ relationships in a single exemplar of an unknown kind. They judged A to be crucial for C by default. However, when they subsequently learned that the causal mechanism fails to generate the $A \rightarrow C$ dependency in other exemplars of the same kind, subjects devalued A as a crucial explanation for C even within the first exemplar. We relate these findings to the idea that good explanations pick out portable dependency relations, and that sensitive causes fail to meet this requirement.

Keywords: explanation; causal mechanisms; causal chains; sensitivity; portability

Introduction

Causal relationships are implemented by causal mechanisms (Machamer, Darden, & Craver, 2000). If we say that one event (A) causes another event (C), we generally assume that there is some process leading from A to C that can principally be discovered, even if we do not yet know what the actual mechanism is. Accordingly, the causal arrow in notations such as $A \rightarrow C$ is sometimes interpreted as a "mechanism placeholder" (Pearl, 2000).

When mechanism knowledge about a particular causal relationship (e.g., $A \rightarrow C$) is made explicit, the resulting causal model takes the form of a causal chain (e.g., $A \rightarrow B \rightarrow C$, where B is an intermediate cause implementing the mechanism). The current research asks how such integration of mechanism knowledge affects people's conceptualization of the original causal relationship. More specifically, the question is which properties of mechanism B can affect people's impression of the importance of A for explaining C.

Intuitively, there are two ways to interpret the causal role of B in causal chains. First, B could be seen as a *mediator* implementing the causal influence of A on C. Under this reading, A continues to be seen as explanatory for C, even though its causal influence is completely mediated via B. Second, it could be seen as an *alternative explanation* for the occurrence of C, screening off the influence of A on C. The fact that, if we know the value of B, A adds nothing to explaining C, provides a reason to devalue A as appropriate explanation of C under this interpretation.

Nagel and Stephan (2015) showed that both interpretations can arise when subjects learn that different mechanisms implement one and the same type-level causal relationship. They had their subjects learn a strong dependency of grades in a gym class (C) on the pupils' gender (A) from fictional covariation data. Subjects indicated strong agreement with the claim that, within the observed sample, a pupil's gender was crucial for his or her grade. In a second learning phase, half of the subjects learned that the $A \rightarrow C$ relationship was mediated by a genetically determined physiological process (B_1), while the other half learned that it was mediated by the gender preferences of the teacher (B_2). Afterwards, subjects in the physiological mechanism condition continued to endorse the statement that a pupil's gender was crucial for this pupil's grade (indicating that B_1 was interpreted as a mediator of the original $A \rightarrow C$ relationship), while subjects in the teacher mechanism condition devalued gender as crucial explanation for the grades (indicating that B_2 was interpreted as an alternative explanation for C).

Both conditions were equivalent in terms of objective causal structure and observed dependency patterns, so it seems the difference in interpretation results from some aspect of the manifold content-related differences between both contrasted mechanisms. One salient hypothesis is that intentional agents, like the teacher, might be seen as initiators of causal sequences (unmoved movers) and therefore always screen off the influence of upstream physical preconditions of their actions from downstream effects. Blind physical processes like genetics, by contrast, may not have this quality and may thus be regarded as mere mediators of the influence of upstream root causes. In a second experiment, Nagel and Stephan (2015) ruled out this hypothesis. They presented their subjects with a scenario in which a physical signal (A) was picked up by a human agent who deliberately reacted to the signal (B) to produce a final outcome (C). Subjects repeatedly observed this deterministic causal chain in all conditions. Half of the participants learned that the agent had a benign motivation in implementing the $A \rightarrow C$ chain, while the other half learned that he had a malevolent motivation. Afterwards, they were asked to what extent it was appropriate to state that the original signal (A) vs. the agent's reaction (B) was crucial for the occurrence of the outcome (C). It turned out that in case of the benign agent, both the distal signal and the proximal action were judged to be equally crucial for the occurrence of C. The distal physical cause thus retained its explanatory power and was seen to bring about the outcome *by means of* human agency. By contrast, if the agent had a morally dubious motive, the distal cause was devalued as explanation despite a perfect dependency relation with the outcome in the observed sample; proximal human agency served

as *alternative explanation* for the outcome instead. This finding strongly suggests that it is not an agent's intentionality per se that leads to devaluation of upstream causes as explanations for downstream effects, but rather some other property that is related to the motivation of the agent. In the remainder of this paper, we will outline and test the hypothesis that this property does not reside exclusively in moral qualities of intentional agents, but quite generally reflects inferences about whether the mechanism can be expected to generalize to other, inessentially different contexts.

Sensitivity and Explanatory Relevance

Woodward (2006) investigated the human practice of making causal claims. He noted that causal claims require not only that the effect be counterfactually dependent on the cause, but also that this counterfactual dependence continue to hold under varied boundary conditions. Causal relationships that do not fulfill this second requirement are called *sensitive*, and Woodward (2006) argues that sensitive causal relationships are regarded as deficient despite a strong dependence relationship under the conditions in which they do obtain. Good causes are those that not only bring about their effects in the narrow context of actually observed circumstances, but would continue to do so in different contexts. The requirement for good causes to be insensitive resonates with philosophical and psychological accounts of explanation. Many theorists have argued that good explanations tend to pick out factors that are generalizable beyond the concrete set of observations that presently is to be explained (Garfinkel, 1981; Hitchcock, 2012; Lombrozo & Carey, 2006). This ensures that the generated explanation will be useful for making predictions and for planning interventions in future similar situations (Lombrozo, 2010).

One reason for high sensitivity of causal relationships is that the mediating mechanism works reliably only under quite specific boundary conditions, but is easily disturbed in other, similar situations. We propose that whenever people find out that an observed causal relationship is implemented by a mechanism that is highly sensitive in this sense, they devalue the distal cause as explanation for the terminal effect. To illustrate, consider again the scenarios used by Nagel and Stephan (2015). If the influence of gender on grades is mediated by a genetically determined physiological mechanism, this implies that the relationship will continue to hold in future observations with different samples of pupils, which makes the $A \rightarrow C$ relationship insensitive. The teacher mechanism, by contrast, implies that this relationship depends on the presence of highly peculiar boundary conditions which will rarely be met in other, similar situations (as most other teachers, hopefully, will not exhibit the same bias). This makes the observed $A \rightarrow C$ relationship highly sensitive—it will break down as soon as we leave the narrow context of the class that was actually observed in the sample. It is recognized that gender will not *generally* influence grades and is therefore considered a poor explanation for the grades *even within the observed sample*.

The mechanisms compared by Nagel and Stephan (2015) differed on many dimensions other than sensitivity, including intentional agency and moral abnormality. Our goal in the present experiments is to isolate the sensitivity of the mediating mechanism. We created new experimental material from the domains of biology and physics in which we manipulated a given mechanism's sensitivity purely in statistical terms. We first presented subjects with a single exemplar of an unknown natural kind or artifact and let them discover a deterministic $A \rightarrow B \rightarrow C$ chain within this entity. In a subsequent learning phase, we showed subjects the same exemplar again, but this time in the company of several other exemplars of the same kind with identical appearance. One half of the subjects saw that the new exemplars behaved just like the first exemplar in terms of the $A \rightarrow B \rightarrow C$ dependency pattern. The other half saw that only the first exemplar once again showed the same dependency pattern, while in all other exemplars the presence of A failed to lead to the presence of B (and, hence, C). Subjects in both conditions were then asked how appropriate it was to say that A was crucial for the presence of C *within the first exemplar only* which had constantly displayed perfect dependency relations in both conditions. We predicted reduced appropriateness ratings in the condition in which the dependence of C on A did not generalize to other exemplars of the same kind.

In the first experiment, we tested and confirmed these predictions in the domain of biology. In the second experiment, we replicated the findings in the domain of artifacts and additionally controlled for the relative complexity of the potentially explanatory variables A and B.

Experiment 1

Participants

The experiment was conducted as an online study. A total of 150 subjects were recruited from a panel (*www.pureprofile.com*), 44 of which (29%) were removed prior to the analyses because they did not complete the survey or failed to solve a simple attention check question at the end. The mean age of all included subjects ($N = 106$, 80 women) was 37 years ($SD = 8.16$). Included subjects received a payment of £ 6 per hour.

Design, Materials, and Procedure

Subjects were randomly allocated to the conditions (Sensitivity: Sensitive vs. Insensitive). They were asked to take the perspective of a marine biologist who discovered a single exemplar of deep sea fish and called it "Fish #1". Their task was to test whether noise (A) leads to activity in the brain of the fish (B) and finally to an illumination of the fish's antenna (C). We then presented our subjects an animation of Fish #1 (see Figure 1). When participants pressed the "Play" button, they saw sound waves coming out of the speaker. About half a second later, the brain activity device's monitor displayed a flickering amplitude moving across the screen. Finally, about

one second later, the fish's flash bulb turned from blue to yellow. If participants hit the stop button, all variables returned to their initial state. Subjects could (de-)activate the loudspeaker as often as they wished.

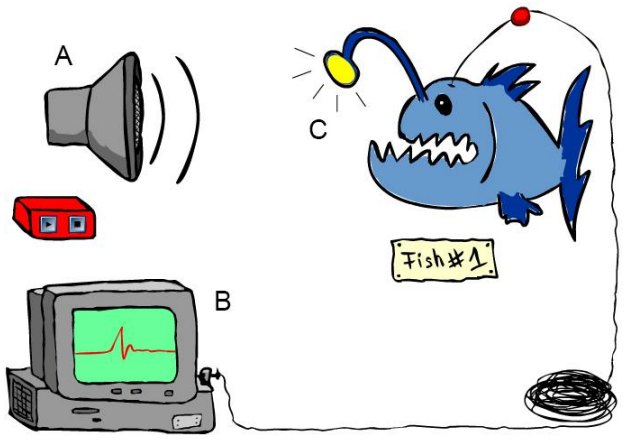


Figure 1: Screenshot of the animation used in the first learning phase of Experiment 1 while all three variables are active. Letters A, B, and C were not shown to participants.

When participants felt they had learned enough about the relationships, they proceeded to the first question screen on which they were asked how appropriate the following statements were for describing the observations they had just made of Fish #1. The first statement was "The presence of sound waves is crucial for Fish #1's antenna to lighten up" (appropriateness rating $[A \rightarrow C]_{pre}$), and the second statement was "Activity of the brain area is crucial for Fish #1's antenna to lighten up" (appropriateness rating $[B \rightarrow C]_{pre}$). We used two independent rating scales to allow participants to judge both causes as equally explanatory, while at the same inviting them to see both statements as contrastive alternatives by using the word "crucial". Participants provided their judgments on 11 point rating scales ranging from "0 = not at all appropriate" to "10 = very appropriate" for each statement. We expected equally high ratings for both statements, indicating that brain activity is seen as a mediator and sound waves are seen to be explanatorily relevant.

The experimental manipulation was applied after subjects had given their baseline ratings. All participants read that they had caught nine additional exemplars of the same kind of fish. On the next screen, they saw a large animation showing all ten fish (consecutively labelled Fish #1 to Fish #10), each in the same set-up as shown in Figure 1. Below the ten fish, there was a device with a play- and stop-button which they could use to (de-)activate all ten loudspeakers simultaneously. In both conditions, Fish #1 again reacted exactly as in the first learning phase. The crucial difference between both conditions was the behavior of the additional nine fish exemplars. In the insensitive condition, all other fish behaved exactly like Fish #1, while in the sensitive condition, none of the other fish reacted to the activation of the loudspeaker with brain activity or antenna lightning. This manipulation was intended to confirm in the insensitive condition the assumed prior expectation that the $A \rightarrow C$ dependence is exportable from Fish #1 to the whole kind, while subjects in the sensitive condition should conclude that the $A \rightarrow C$ dependence is not exportable beyond the narrow context of Fish #1.

After having made these additional observations, subjects were asked to reconsider the results of their previous experiment *with Fish #1 only* and to answer the same two questions again in light of their new knowledge about the whole swarm of fish. The appropriateness ratings $(A \rightarrow C)_{post}$ and $(B \rightarrow C)_{post}$ were measured exactly as the baseline ratings described above. Our central prediction for the sensitive condition was that the $(A \rightarrow C)_{post}$ ratings should drop considerably because the $A \rightarrow C$ dependency is not exportable beyond the context of Fish #1. The $(B \rightarrow C)_{post}$ appropriateness ratings, by contrast, should not be reduced by the swarm information. Brain activity remains a good predictor of antenna flashing across the whole swarm. In the insensitive condition, of course, neither of the ratings should be affected by the swarm data.

Finally, we wanted to make sure that participants encoded the contingencies between the variables accurately both within Fish #1 and across the whole swarm of fish. On a first screen, participants were prompted to recall what they had learned about Fish #1 and were asked six conditional probability questions concerning Fish #1 only. For example, $P(C|A)$ was assessed with the question "How likely is it for Fish #1's antenna to lighten up given that sound waves are present?" and an 11 points rating scale ranging from 0 (impossible) to 100 (certain). Analogous questions were asked for $P(C|\neg A)$, $P(B|A)$, $P(B|\neg A)$, $P(C|B)$, and $P(C|\neg B)$. On a second screen, the same six questions were repeated with the whole swarm as reference class. These twelve estimates were used to compute contingency estimates (ΔP) for each of the three causal relationships both at the exemplar level and at the swarm level. Equally high contingency estimates at the exemplar level for $A \rightarrow C$ and $B \rightarrow C$ would demonstrate that the expected effects on the appropriateness ratings would not be due to different dependency assumptions within the exemplar, but rather due to sensitivity of the $A \rightarrow C$ relationship across the whole kind.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 2a. We conducted a three-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) $\times$ 2 (Relationship: $A \rightarrow C$ vs. $B \rightarrow C$, within-subject) $\times$ 2 (Rating Position: Pre vs. Post, within-subject) mixed ANOVA. We obtained a significant two-way Sensitivity $\times$ Position interaction, $F_{1,104} = 8.61$, $p < .01$, $\eta_g^2 = .014$, indicating that the swarm information affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was qualified by a

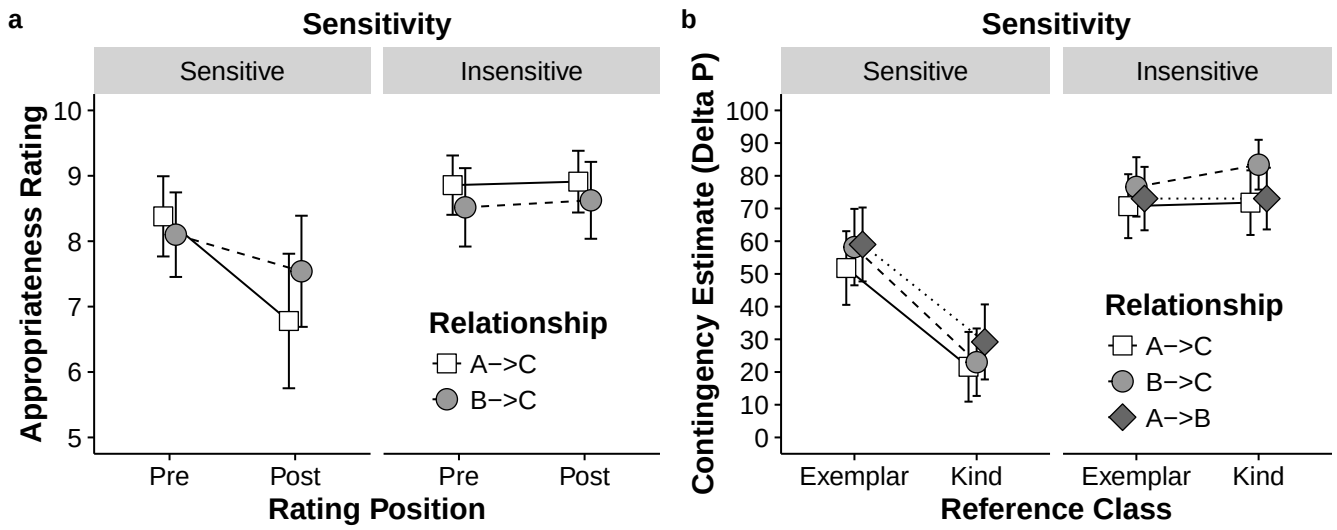


Figure 2: Group means (error bars = 95% CI) of appropriateness ratings (a) and contingency estimates (b) in Experiment 1.

marginally significant three-way Condition × Position × Relationship interaction, $F_{1,104} = 3.64$, $p < .06$, $\eta_g^2 = .003$, indicating that the selective decrease of ratings from pre to post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, as predicted by our account.

To assess subjects' contingency estimates for the three relationships, we calculated each subject's ΔP estimate for each relationship both within Fish #1 and across the whole swarm of fish. For example, the contingency estimate for the $A \rightarrow C$ relationship within Fish #1, $\Delta P(A \rightarrow C)_{\text{Exemplar}}$, was calculated by subtracting each subject's $P(C|\neg A)_{\text{Exemplar}}$ rating from the same subject's $P(C|A)_{\text{Exemplar}}$ rating. The contingency estimates across the whole swarm were calculated analogously using the conditional probability judgments for the whole swarm. The descriptive results are summarized in Figure 2b. Most importantly, the $\Delta P(A \rightarrow C)_{\text{Exemplar}}$ estimates in the sensitive condition were not lower than the $\Delta P(B \rightarrow C)_{\text{Exemplar}}$ estimates. This finding rules out the alternative explanation that the selective drop in the $(A \rightarrow C)_{\text{post}}$ appropriateness ratings in the sensitive condition results from selectively decreased dependency estimations within Fish #1 for this particular relationship.[1]

[1]The surprisingly low $\Delta P(B \rightarrow C)_{\text{Kind}}$ estimate in the sensitive condition resulted from very low $P(C|B)_{\text{Kind}}$ ratings. The observation that in Fish #1 (the only exemplar in which brain activity was ever recorded) brain activity reliably led to antenna illumination did not suffice to make subjects generalize this relationship across the whole kind. Hesitance to generalize from sparse data, however, is different from gathering positive evidence for sensitivity. The finding that subjects continued to regard brain activity as the crucial explanatory factor within Fish # 1 despite low kind-general contingency estimates thus does not directly disprove our hypothesis, but is certainly a finding that needs further investigation.

Discussion

In this experiment, we have demonstrated that sensitive mechanisms reduce the explanatory relevance of distal causes in causal chains. Our subjects first learned that, within the narrow context of a single exemplar of a biological kind, cause A deterministically produced effect C, and that this causal relationship was always mediated by mechanism B. At this point, they interpreted the B to be a mediator of the observed $A \rightarrow C$ relationship and correspondingly found that A and B were equally crucial for the occurrence of C. However, if they subsequently found out that cause A failed to activate mechanism B in all other exemplars of the same kind, this affected their representation of the perfect dependency relation within the initially observed exemplar. Even within this exemplar, they now judged A to be less crucial for the occurrence of C than the more proximal B, indicating that A became deficient as explanation for C, despite the perfect $A \rightarrow C$ dependency relation that was observed throughout for this exemplar.

Experiment 2

In the second experiment, our main goal was to replicate the findings from Experiment 1 in the domain of artifacts. Furthermore, we made the contents of causes A and B more similar to each other in order to rule out alternative explanations for their differential treatment in the post-ratings. In Experiment 1, A was a simple, physical variable external to the system under study, while B was a complex, physiological variable internal to the system. In Experiment 2, we used only internal, physical variables that varied in complexity. We counterbalanced the assignment of the simple and the complex variable to the positions of distal cause A and proximal cause B. We expected the same pattern of results as in Experiment 1 under both assignments, showing reduced explanatory relevance of the distal cause results from mechanism sensitiv-

ity per se, regardless of its surface characteristics.

Participants

We recruited and compensated 331 subjects as in Experiment 1. 115 (35%) were removed prior to analysis according the same criteria as above. The mean age of all included subjects ($N = 216$, 109 women) was 40 years ($SD = 8.36$).

Design, Materials, and Procedure

Subjects were randomly allocated to one of four conditions that resulted from a 2 (Sensitivity: Sensitive vs. Insensitive) $\times$ 2 (Complexity of Mechanism: Complex vs. Simple) between-subjects design. They encountered an exemplar of an unknown machine, "Machine #1", with three visible devices: a single rack wheel, a complex system of rack wheels, and a fan (see Figure 3). Subjects in the complex mechanism condition saw the three devices in the arrangement shown in Figure 3, while for subjects in the simple mechanism condition, the positions of the single wheel and the complex system of wheels were reversed. The procedure was analogous to Experiment 1. In a first learning phase subjects set the leftmost device in motion and observed that this was followed by movement of the device in the middle, which was in turn followed by movement of the fan on the right. Switching off the leftmost device resulted in subsequent inertia of all variables. On the next screen, we assessed appropriateness ratings $(A \rightarrow C)_{pre}$ and $(B \rightarrow C)_{pre}$ as in Experiment 1. In the second learning phase, subjects were shown Machine #1 again together with five additional machines with identical surface features, consecutively labelled "Machine #2" to "Machine #6". They could separately intervene on each machine's leftmost device as often as they wished. In the Insensitive condition, all six machines behaved just as Machine #1 in the first learning phase. In the Sensitive condition, Machine #1 also worked just as before, but in none of the additional machines did the device in the middle or the fan ever turn on. On the following screens, subjects again indicated their $(A \rightarrow C)_{post}$ and $(B \rightarrow C)_{post}$ appropriateness ratings as well as their exemplar-specific and kind-general conditional probability judgments analogous to Experiment 1.

Results

The descriptive results for the appropriateness ratings are displayed in Figure 4. We conducted a four-way 2 (Sensitivity: Sensitive vs. Insensitive, between-subjects) $\times$ 2 (Relationship: $A \rightarrow C$ vs. $B \rightarrow C$, within-subject) $\times$ 2 (Rating Position: Pre vs. Post, within-subject) $\times$ 2 (Complexity of Mechanism: High vs. Low, between-subjects) mixed ANOVA. We again obtained a significant two-way Sensitivity $\times$ Position interaction, $F_{1,212} = 13.02$, $p < .001$, $\eta_g^2 = .01$, indicating that the information about the additional machines affected appropriateness ratings in the sensitive condition more than in the insensitive condition. More importantly, this interaction was again qualified by a significant three-way Sensitivity $\times$ Position $\times$ Relationship interaction, $F_{1,212} = 12.23$, $p < .001$, $\eta_g^2 = .003$, indicating that the selective decrease of ratings from pre to post in the sensitive condition was more pronounced for the $A \rightarrow C$ relationship than for the $B \rightarrow C$ relationship, just as in Experiment 1. The assignment of the single wheel and the complex system of wheels to distal cause (A) vs. proximal cause (B) did not affect the results, nor did this factor interact with any of the other variables in the design. The data shown in Figure 4 is therefore collapsed across this factor.

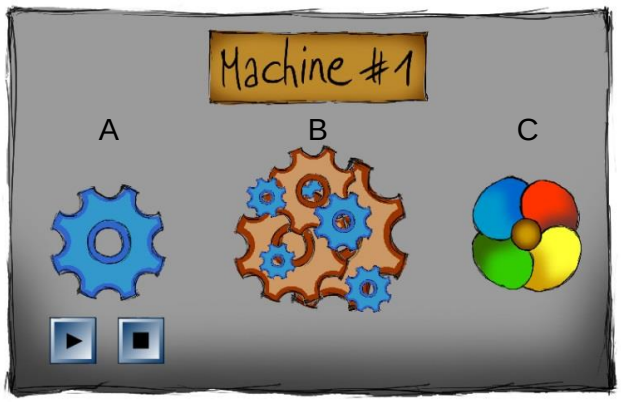


Figure 3: Screenshot of the animation used in the first learning phase of Experiment 2. Devices A and B were reversed in the Simple Mechanism condition. Letters A, B, and C were not shown to participants.

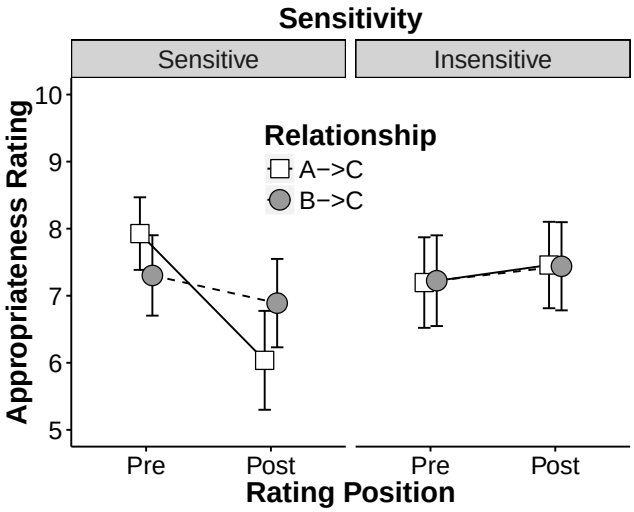


Figure 4: Group means (error bars = 95% CI) of appropriateness ratings in Experiment 2.

Subjects' contingency estimates were analyzed analogous to Experiment 1 and yielded analogous effects. The $\Delta P(A \rightarrow C)_{Exemplar}$ estimates in the sensitive condition were again as high as the $\Delta P(B \rightarrow C)_{Exemplar}$ estimates. As before, this rules out that the selective drop in $A \rightarrow C$ appropriateness ratings results from decreased $A \rightarrow C$ dependency estimates within the focal entity.

Discussion

In this experiment, we have shown that the finding that sensitive mechanisms lead to devaluation of distal causes generalizes to the domain of artifacts. If setting a physical device in motion (A) leads to movement of a second device (B), which in turn sets in motion a third device (C), both A and B are seen as equally crucial for the movement of C. However, if it is later learned that this dependency does not generalize to other exemplars of the same kind of machine, people revise their interpretation of the chain even within the first exemplar. They now judge the distal cause to be less crucial for the occurrence of the effect than before, and as less crucial than the more proximal cause that mediates the relationship. The effects were even cleaner than in Experiment 1, despite the fact that we made causes A and B more similar to each other and even counterbalanced their contents. This supports our hypothesis that high sensitivity of mechanisms *per se* leads to an interpretation of the mechanism as alternative explanation rather than as a mediator of the original relationship.

General Discussion

In two experiments, we have demonstrated that sensitive mechanisms tend to be seen as alternative explanations of their effects rather than as mediators of the causal influence of a distal cause. When people learn about a new indirect causal relationship in a single context, their default intuition seems to be that the distal cause is a crucial contributor to the terminal effect. However, if they afterwards realize that the mechanism generating the dependency between distal cause and terminal effect breaks down in most other similar contexts, they revise this intuition and devalue the causal contribution of the distal cause. The information that the mediating mechanism requires highly specific, uncommon boundary conditions makes clear that the observed A $\rightarrow$ C dependency is highly sensitive (Woodward, 2006). Sensitive causes, in turn, tend to be regarded as somewhat deficient, presumably because they fail to support future predictions and interventions in similar cases (Lombrozo, 2010; Lombrozo & Carey, 2006). As Garfinkel (1981) put it, if we want to explain the occurrence of a particular outcome, our real object of explanation is never just the occurrence of *that particular* outcome. Instead, we search for stable causes that explain the occurrence of a whole equivalence class of inessentially different outcomes. If we find out that a mechanism relates a cause to a to-be-explained effect in only a small subset of cases within the relevant equivalence class (e.g., it only works reliably in a small number of exemplars of an otherwise apparently homogeneous kind), the cause does not provide a stable explanation for this kind of effect. The discovered mechanism then turns into an alternative explanation for the outcome that screens off the influence of the distal cause from the explanandum. This explanation captures not only the present data, but also previous findings in which sensitivity of the mechanism was not directly manipulated in statistical terms, but rather implied by qualitative characteristics of the described mechanism (e.g., moral abnormality; see Nagel & Stephan, 2015).

The fact that sensitive mechanisms lead to a decrease in explanatory relevance of the distal cause A (rather than to an increase in relevance of the proximal cause B) suggests that the following psychological process might underlie the observed phenomenon. When sensitive mechanisms are observed, it becomes necessary to assume the influence of an additional, latent variable that interacts with the distal cause A to produce proximal cause B in a few but not all contexts (e.g., an abnormal preference structure of the teacher, or a genetic abnormality in Fish #1) in order to capture the structure of the complete situation. The apparent necessity of this additional variable for producing B (and, hence, C) makes it obvious that A is not sufficient in producing B (and, hence, C) even within the focal entity. Sufficiency, in turn, has been shown to be closely linked to explanatory relevance (e.g. Hilton, McClure, & Sutton, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). Future studies might aim to test this hypothesis more specifically.

Acknowledgments

We thank York Hagmayer, Jana Samland, and Michael Waldmann for helpful discussions.

References

Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.

Hilton, D. J., McClure, J., & Sutton, R. M. (2009). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*, 383–400.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*, 942–951.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 1–25.

Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1691–1697). Austin, TX: Cognitive Science Society.

Pearl, J. (2000). *Causation: Models, reasoning and inference*. Cambridge, MA: Cambridge University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*, 1–50.

The Variable Relationship Between On-Task Behavior and Learning

Karrie E. Godwin (kegodwin@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Howard Seltman (hseltman@cmu.edu)
Carnegie Mellon University, Department of Statistics
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Ma. Victoria Q. Almeda (victoriaalmeda@gmail.com)
Department of Human Development, Teachers College Columbia University
525 W. 120th St., New York NY 10027 USA

Shimin Kai (shimin.kai@gmail.com)
Department of Human Development, Teachers College Columbia University
525 W. 120th St., New York NY 10027 USA

Ryan S. Baker (ryanshaunbaker@gmail.com)
Department of Human Development, Teachers College Columbia University
525 W. 120th St., New York NY 10027 USA

Anna V. Fisher (fisher49@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

The *Time-on-Task hypothesis* asserts that learning is a function of time one allocates to a learning task. Thus, time off-task reduces learning opportunities and is therefore thought to be detrimental to learning. To date, the available research suggests a positive relationship between time on-task and achievement; however, the strength of the correlation fluctuates dramatically. One potential explanation that has been put forth to account for the mixed results is differences in the operational definition of *time*. The present study tests this hypothesis by examining whether a more stable relationship between on-task behavior and learning can be obtained if *time* is operationalized in a uniform way. The results of the present study indicate that while on-task behavior was positively correlated with learning outcomes overall, marked variability was still found across classrooms suggesting that the divergent results obtained in previous research are not driven solely by differences in how *time* is measured.

Keywords: On-task behavior; Attention; Learning; Achievement

Introduction

In education a common assumption is that the more time children attend to something, the better they should learn the material. This tenant of conventional wisdom was formalized by Carroll (1963) and has become known as the *Time-on-Task hypothesis*. The *Time-on-Task hypothesis* asserts that learning is a function of the amount of time one allocates to a particular learning task. Thus, time off-task reduces learning opportunities and is therefore thought to be detrimental to learning (Carroll, 1963). Carroll's *Time-on-Task hypothesis* spurred a great deal of research attempting to show that learning is directly related to the amount of time one spends on a particular task (Cobb, 1972; Lahaderne, 1968; McKinney, Mason, Perkerson, & Clifford, 1975; Samuels & Turnure, 1974; Frederick, Walberg, & Rasher, 1979).

Given that predictors of achievement are often intractable (see Karweit & Slavin, 1980 for discussion), identifying factors that are malleable is of particular importance for practitioners. Thus, *time* is a factor that is of great interest as it is largely a malleable factor. If time is shown to be a meaningful determinant of learning, then interventions can be created that increase time or that optimize how existing instructional time is utilized.

Despite considerable work in this area, the understanding of the relationship between *time* and learning remains limited, as the existing literature has yielded conflicting results. To date, the available research suggests a generally positive relationship between the amount of time spent on-task and achievement. However, the strength of the correlations fluctuate widely, ranging between 0.10 and 0.70 (Karweit, 1984; for more recent examination of the relationship between on-task behavior and learning see: Baker, Corbett, Koedinger, & Wagner, 2004; Fisher, Godwin, & Seltman, 2014; Gobel, 2008; Godwin & Fisher, 2014; Kovanovic et al., 2015).

While a full account of the factors driving the variability in the relationship between time and learning is needed,

some potential explanations have been put forth in the literature. One possible explanation for the mixed findings is the variety of ways in which *time* has been operationalized. Indeed, Karweit and Slavin (1982) suggest that differences in operational definitions, as well as methodological and procedural differences, may contribute to the mixed results across the existing literature. Although establishing precise operational definitions is important for all research, it may be particularly pertinent in this context given that *time* has been defined in many different ways across a multitude of studies (for discussion see Caldwell, Huitt, & Graeber, 1982; Frederick & Walberg, 1980; Goodman, 1990; Karweit & Slavin, 1981; Wiley & Harnischfeger, 1974). For example, in the prior literature *time* has been equated with attention and thus it has been defined at the student level as looking time, on-task behavior, and engaged time (e.g., Choudhury & Gorman, 2000; Cobb, 1972; Fisher, Godwin, & Seltman, 2014; Godwin & Fisher, 2014; Lahaderne, 1968; Lee, Kelly, & Nyre, 1999). Other researchers have defined *time* at a classroom level in which *time* is indexed by the amount of time allocated, or the amount of time actually spent on an instructional activity or a particular subject area (e.g., Arlin & Roth, 1978; Baker, Fabrega, Galindo, & Mishook, 2004). Still other researchers have equated *time* with the amount of schooling and operationalized *time* as the length of the school day, length of the school year, total number of school days attended, or even years of schooling (e.g., Agrawal, Smith, & Wick, 1977; Coleman et al., 1966; Cooper, Allen, Patall, & Dent, 2010; Hough & Bryde, 1996; Hyman, Wright, & Reed, 1975; Karweit, 1973; Wiley & Harnischfeger, 1974). With such a wide variety of definitions of *time*, it is perhaps not surprising that discrepant results have been obtained.

In an influential study with elementary school students, Karweit and Slavin (1981) tested the hypothesis that the manner in which *time* is operationalized (among other factors) influences the relationship between time and learning. A small subset of children from 18 elementary school classrooms participated in the study. The sample consisted of two age groups: second/third graders and fourth/fifth graders. In this observational study, several measures of time were collected during mathematics instruction in order to assess whether time (in any form) was a significant predictor of children's achievement scores on the Mathematics Computation and Mathematics Concepts and Applications subscales of the *Comprehensive Test of Basic Skills* (CTBS). The authors focus on four central measures of *time* which include: Total scheduled time (i.e., the amount of time allocated to math instruction), Total instructional time (i.e., the amount of time actually spent on math instruction subtracting time for procedural activities), Engaged time (i.e., the amount of time students spent engaged with the instructional activity/ time-on-task), and rate of engagement (engaged time/total instructional time).

Karweit and Slavin (1981) found mixed results as a function of how *time* was measured and as a function of grade level. For second and third grade students, only engaged time and rate of engagement were found to be significant predictors of children's CTBS post-test scores. All other measures of *time* (i.e., total scheduled time and instructional time) were not significantly related to second and third grade children's CTBS scores. In contrast, for older children none of the measures of *time* (i.e., total scheduled time, total instructional time, engaged time, nor rate of engagement) were found to be significant predictors of their CTBS scores.

Karweit and Slavin (1981) concluded that measures of *time* which focus more closely on the amount of time students spend on-task or measures that are more indicative of how students utilize instructional time are stronger predictors of achievement compared to measures which operationalize *time* more generally such as time allocated for instruction or even total instructional time. Thus, the manner in which *time* is operationalized is thought to influence whether a relationship between *time* and learning is found as well as the strength of that relationship. Additionally, the authors point out that the inconsistency in the relationship between *time* and learning across grade levels has been observed in other studies and highlights the need to examine individual difference factors (e.g., interest in the subject matter, student aptitude) which will likely contribute to the amount of *time* needed to master course content.

Although the work of Karweit and Slavin (1981) suggests that using a uniform measure of *time*, and in particular using measures such as time on-task or engaged time (compared to more global measures) should reduce variability and increase the strength of the relationship between *time* and learning, this issue remains unresolved and continues to be an important area of inquiry today (see Kovanovic et al., 2015). The present study continues to explore the relationship between *time* and learning by examining whether the variability in the prior literature is merely an issue of measurement inconsistency. Specifically, we investigate whether a positive relationship between on-task behavior and learning is *consistently* obtained when utilizing a uniform measure of *time* (i.e., for all participating classrooms the fraction of students' on-task behavior was recorded) in a large sample of elementary school students. The present work makes an important contribution to the literature as it examines this question with a larger sample of elementary school children and examines possible individual characteristics (i.e., gender, grade level) and school based factors (i.e., school type: private, public charter schools) which may contribute to the observed variability in the prior literature.

Method

Participants

Twenty classrooms participated in the present study. However, one classroom was excluded from analyses for reasons discussed below. The classrooms were recruited

from 7 schools which included 4 charter schools and 3 private schools. Participating schools were located in or near a medium-sized city in the Northeastern United States of America. In order to obtain a representative sample, different grade-levels were recruited. Four grade-levels participated in the present study: kindergarten, first-grade, second-grade, and fourth-grade. The distribution across the four grade-levels was as follows: 5 kindergarten classrooms, 5 first-grade classrooms, 7 second-grade classrooms, and 3 fourth-grade classrooms. Despite efforts to recruit a balanced sample, third-grade teachers did not volunteer to participate in the present study. The sample consisted of 375 children (177 males, 198 females). The average number of children observed in a single observation session was 17.6 children. The number of children observed per session ranged from 10 to 23 children. These children were also part of a larger study examining patterns of attention allocation in elementary school. The results of that study are reported elsewhere (Godwin et al., In Press).

Design & Procedure

Each classroom was observed two times in order to obtain more stable estimates of children's on-task behavior. The average delay between observation sessions was 3.05 calendar days (the delay ranged from 1 day to 7 days). Each observation session lasted approximately one-hour. All observations occurred during mathematics instruction. The average number of observations per session was 249.6 and the average number of observations per child within a session was 15.5. The classroom observations occurred between October 2012 and December 2012.

Coding Behavior Coders utilized the *Baker-Rodrigo Observation Method Protocol* (BROMP) for coding behavioral data in field settings (Ocumpaugh, Baker, & Rodrigo, 2015). Training consisted of coding videotapes and live observation sessions. In order to establish inter-rater reliability Cohen's Kappa was calculated. Values ranged from 0.79 to 0.84 which exceeds the 0.75 threshold which Fleiss (1981) refers to as "excellent" in field settings.

A round-robin coding strategy was utilized in order to reduce the tendency of observers to attend to salient instances of off-task behavior. Prior to beginning the observation session, the observation order was determined. Each observation period lasted up to 20 seconds. The first unambiguous behavior observed during the 20 second period was recorded. Then the coder would observe the next child in the rotation. This process repeated for the duration of the observation. Thus, every child was observed multiple times throughout the observation session. During observations peripheral vision was utilized in order to make it less apparent to the child that he or she was being observed. This procedure has been employed successfully in prior research to reliably code middle and high school students' behavior and affect (e.g., Ocumpaugh et al., 2012).

Coders classified children's behavior as on- or off-task using the direction of the child's eye gaze. Contextual clues, such as teacher instructions, were also considered in order to disambiguate between on- and off-task behaviors. If the child was looking at the teacher or the instructional materials they were categorized as on-task. If the child was looking elsewhere, they were categorized as off-task. For each child the fraction of children's on-task behavior was then calculated by taking the number of times each child was on-task divided by the total number of observed behaviors (i.e., on-task and off-task behaviors).

Learning Measures All of the learning measures were administered by the teachers as part of their standard practice. As a consequence of collecting genuine learning measure that are used in practice, there was variability in the types of learning measures that were administered in each classroom (an issue we return to in the Discussion Section). Learning measures included: quizzes, report cards, and/or fall and winter MAP (*Measure of Academic Progress*) scores from a computerized adaptive assessment program.

The distribution of available learning outcomes was as follows: 4 students had learning outcomes which consisted of fall and/or winter MAPs, 203 students had learning outcomes which consisted of quiz scores and fall and winter MAPs; 132 students had learning outcomes which consisted of report cards and quiz scores; and 17 students had learning outcomes which consisted of report cards. Quiz scores were excluded from analysis due to low information content. One classroom utilized letter grades for their report cards (as opposed to percentages); consequently, this classroom was dropped from the analysis. As a result, both classroom observations and learning outcomes were available for 356 students from 19 classrooms.

The learning measures were converted into Z-scores and averaged together to create the composite variable *Total Learning Outcome* for each student. The composite variable *Total Learning Outcome* was computed separately for each classroom in an attempt to correct for different grading practices across schools and classrooms. This approach obviates the need for hierarchical modeling. It is important to note that hierarchical linear models were also tested; however, the estimate of the random intercept was 0 indicating that a hierarchical approach was not necessary. This finding is expected given that Z-scores were created separately for each classroom.

Results

On-Task Behavior

Consistent with estimates reported in the prior literature (50-75%, Karweit & Slavin, 1981), children were largely on-task. On average .75 (SD = .13) of children's behaviors were categorized as on-task. The fraction of children's on-task behavior ranged widely with some children exhibiting minimal rates of on-task behavior while other children were

consistently on-task (fraction of on-task behaviors ranged from: .27 to 1.0).

Learning Outcomes

Recall that children's learning scores (report cards and MAP scores) were converted into Z-scores and averaged together to create the composite variable *Total Learning Outcome* for each student. This procedure was repeated for each classroom. Learning Z-scores ranged from -2.82 to 2.03.

Effect of On-Task Behavior on Learning Outcomes

Linear models were fit to assess whether students with a greater fraction of on-task behavior tended to perform better on the learning outcomes relative to their classmates. *Total Learning Outcome* was entered as the dependent variable. Four predictors were included in the model: fraction of on-task behavior, gender, school type (private, charter), and grade level (Kindergarten, First, Second, Fourth grades).

On-task behavior was found to be a significant predictor of children's *Total Learning Outcome,* while controlling for gender, school type, and grade level (β = 1.00, t = 2.60, p = .01). However, the relationship between on-task behavior and learning was weak (see Figure 1), with *Total Learning Outcome* increasing by 0.20 *SD* for every 20% rise in on-task behavior. Furthermore, on-task behavior only accounted for 1.8% of the variability in children's learning scores (R^2 = 0.018, p =.012).

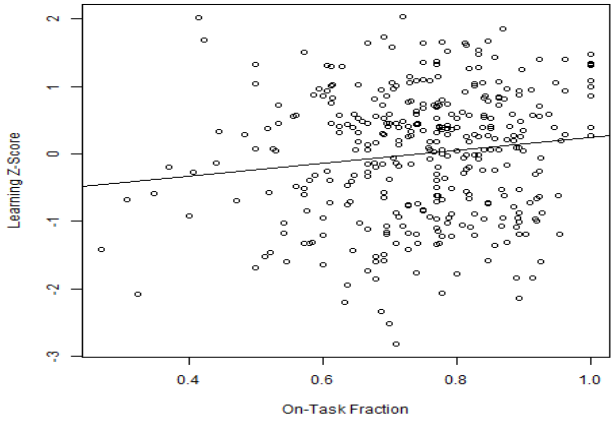


Figure 1. Scatter plot depicting the relationship between the fraction of on-task behavior and the composite variable *Total learning Outcome* (Z-score).

Moderation Analyses

Next, we examined whether the relationship between children's on-task behavior and learning was moderated by gender, grade level, and school type. The results of the moderation analyses are reported below.

Effect of Gender Prior research has found that in elementary school, females typically exhibit more on-task behavior than males (e.g., Godwin et al., In Press; Marks, 2000). Consistent with prior research, we found that females exhibited significantly higher rates of on-task behavior (M = .76, SD = .13) compared to males (M = .73, SD = .13); t(355) = 2.40, p = .017. However, the relationship between on-task behavior and *Total Learning Outcome* scores was not moderated by gender (t(352) = .90, p = .367).

Effect of Grade In order to examine whether the relationship between on-task behavior and learning varied as a function of grade-level, a moderation analysis was performed. The relationship between on-task behavior and children's *Total learning Outcome* score was moderated by grade-level, although the effect was marginally significant (F(3, 348) = 2.25, p = .08). The relationship between fraction of on-task behavior and learning was smaller among first graders (β = -1.41) than it was for kindergartners (β = 1.87, p =.014), second graders (β = 1.11, p = .033), or fourth graders (β = 1.29, p = .036) [all p-values are for slope differences compared to first grade]. In future research it will be important to test the generalizability of this finding by sampling a larger number of classrooms per grade-level from a larger sample of schools in order to determine if the observed grade-level effects are consistent across samples.

Effect of School Type Lastly, a moderation analysis was conducted to investigate whether the relationship between on-task behavior and learning varied as a function of the type of school children attended (i.e., private school or public charter school). The results mirrored those of gender; the relationship between on-task behavior and *Total Learning Outcome* scores was not moderated by school type (t(352) = -1.32, p = .19).

Effect of Classroom Despite measuring *time* in the same manner across each classroom (i.e., for all classrooms in the present study, *time* was operationalized as the fraction of on-task behaviors), only a weak relationship between on-task behavior and learning was found (r = .13). Additionally, there was considerable variability across classrooms (see Figure 2). Indeed the slope estimates obtained when modeling the relationship between on-task behavior and learning were positive for 11 classrooms, while eight classrooms had negative slope estimates (see Figure 2). Furthermore, only three of the 19 classrooms had slope estimates that reached significance: a kindergarten classroom, second grade classroom, and a fourth grade classroom ($ps \leq .025$). All three of the classrooms were from charter schools. It is important to note that for all three classrooms, the direction of the effect is positive (kindergarten class β = 4.54, second grade class β = 2.46, fourth grade class β = 3.51); thus, higher rates of on-task behavior were associated with better learning outcomes. After correcting for multiple comparisons using the Bonferroni correction, the relationship between fraction of on-task behavior and *Total Learning* remained significant for one classroom, the kindergarten classroom (p = .0013).

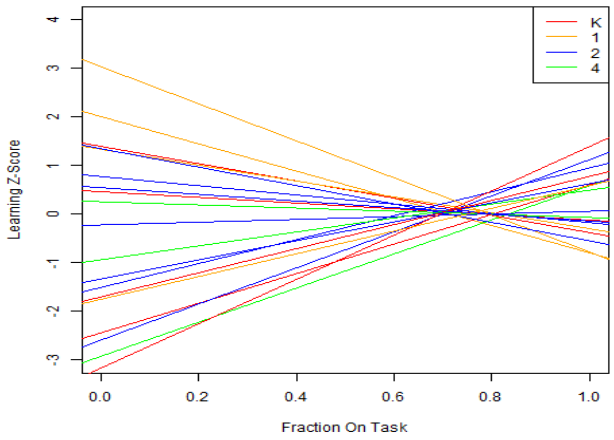


Figure 2. Displays the regression estimates for the relationship between the fraction of on-task behavior and *Total Learning Outcome* Z-scores for each classroom (each line represents a classroom).

The considerable variability observed across classrooms in the strength and the direction of the effect between on-task behavior and learning points to the possibility that the relationship between *time* and learning may be determined in part by classroom level factors. Teacher effectiveness may be one such factor. For example, high levels of on-task behavior and high learning outcomes might be indicative of classrooms with very experienced or high quality teachers, whereas high levels of on-task behavior but low learning may reflect lower quality instruction in which teachers might be engaging, but perhaps not very effective. Additionally, the composition of students within a classroom might influence the relationship between *time* and learning. For example, students who are struggling academically may need to spend large quantities of time on-task in order to obtain average learning scores whereas their high achieving peers might be able to excel academically even with minimal amounts of time spent on-task. To the extent that certain classrooms might have differing compositions of low and high achieving students one might expect that the relationship between on-task behavior and learning may be muted or even negative (for discussion see Frederick & Walberg, 1980).

Discussion

In the present study the variability in the strength of the relationship between on-task behavior and learning as well as the direction of the effect suggest that the divergent results obtained in the previous literature are not driven solely by differences in how *time* is measured. Instead, it appears that differences in the observed relationship between on-task behavior and learning may arise from other factors at the classroom level (e.g., teacher quality, years of experience, classroom size, etc.) as well as from other pertinent individual differences factors (e.g., student aptitude). However these possibilities remain to be explored further in future research.

While the present study provides important insights into the nuanced relationship between on-task behavior and learning several limitations should be noted. First, the learning measures are variable across classrooms and schools. Additionally, we were able to obtain standardized learning measures for a small subset of the participating schools - gathering standardized learning measures should be a focus of future research. Due to the inherent nature of the available learning measures some of the learning measures are more closely linked to the classroom observations (e.g., quiz scores vs. grades and MAP scores). One possibility that can be tested in future research is to assess if a stronger relationship between on-task behavior and learning can be obtained when using learning measures that are closely yoked to observations of student behavior (e.g., class assignments, unit test scores, homework grades) compared to more general measures of achievement (e.g., standardized tests). Some insights into this question have already been obtained. For example, prior research has begun to investigate this question by comparing the relationship between *time* and a subset of dependent learning measures (e.g., standardized test scores vs. chapter tests; Karweit & Slavin, 1981). Additionally, recent research has aimed to reduce potential confounds introduced by delays between observations of student behavior and the collection of students' learning outcome data by yoking observations of students' on and off-task behavior with their *immediate* learning outcomes (Godwin & Fisher, 2014). However, more systematic research is needed. Future research should also incorporate students' pre-test scores as a covariate in order to take students' level of prior knowledge into account, particularly if the lesson content is not entirely novel. As discussed previously, it is likely that students' familiarity with the material and aptitude will influence the relationship between *time* and learning. Lastly, the generalizability of the present results should be assessed by examining the relationship between *time* and learning across multiple subject areas.

In conclusion, the present work suggests that obtaining a consistent relationship between *time* and learning is likely not merely an issue of measurement and highlights the need for future work to be conducted that is able to identify the circumstances in which increasing time on-task actually results in improved learning outcomes.

Acknowledgments

We thank Megan Miller, Laura Pacilio, and Jessica Meeks for their help collecting data. We also thank the children, parents, and teachers who made this project possible. This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (R305B090023) and by the Institute of Education Sciences, U.S. Department of Education (R305A110444). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Agrawal, K., Smith, J., & Wick, J. W. (1977). Achievement in relation to school characteristics. Chicago Department of Research and Evaluation, Board of Education.

Arlin, M. & Roth, G. (1978). Pupils' use of time while reading comics and books. *American Educational Research Journal*, 15, 201-216.

Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.

Baker, D. P., Fabrega, R., Galindo, C., & Mishook, J. (2004). Instructional time and national achievement: Cross-National evidence. *Prospects,* 34(3), 311-334.

Carroll, J. (1963). A Model of School Learning. Teachers College Record, 64, 723-733.

Cobb, J.A. (1972). Relationship of discrete classroom behavior to fourth-grade academic achievement. *Journal of Educational Psychology, 63,* 74-80.

Cooper, H., Allen, A. B., Patall, E. A., & Dent, A. L. (2010). Effects of full-day kindergarten on academic achievement and social development. Review of Educational Research, 80(1), 34-70.

Caldwell, J.H., Huitt, W.G., & Graeber, A.O. (1982). Time spent in learning: Implications from research. *The Elementary School Journal,* 82, 470-480.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley.

Frederick, W. C. & Walberg, H. J. (1980). Learning as a function of time. *Journal of Educational Research*, 73, 183-194.

Frederick, W.C., Walberg, H.J., & Rasher, S.P. (1979). Time, teacher comments, and achievement in urban high schools. *Journal of Educational Research*, *73* (2), 63 - 65.

Gobel, P. (2008). Student Off-task behavior and motivation in the CALL classroom, *International Journal of Pedagogies and Learning*, 4(4), 4-18.

Godwin, K. E. & Fisher, A.V. (2014). Selective sustained attention, the visual environment, and learning in kindergarten-age children: Preliminary results of an individual difference study. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society (pp. 541-546). Austin, TX: Cognitive Science Society.

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (In Press). Off-task behavior in elementary school children. *Learning & Instruction.*

Goodman, L. (1990). *Time and learning in the special education classroom.* SUNY Press.

Hough, D. & Bryde, S. (1996). The effects of full-day kindergarten on student achievement and affect. Paper presented at the annual meeting of the AERA, New York.

Hyman, H. H., Wright, C. R., & Reed, J. S. (1975). The enduring effects of education. Chicago: University of Chicago Press.

Karweit, N. (1973). Rainy days and Mondays: An analysis of factors related to absence from school. Report 162. Baltimore: Center for Social Organization of Schools. The Johns Hopkins University.

Karweit, N. (1984). Time-On-Task reconsidered: Synthesis of research on time and learning. *Educational Leadership,* 41(8), 33-35.

Karweit, N. & Slavin, R. E. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal*, 18(2), 157 - 171.

Karweit, N. & Slavin, R. E. (1980). Time-On-Task: Issues of timing, sampling, and definition. Baltimore: Center for Social Organization of Schools. The Johns Hopkins University.

Karweit, N. & Slavin, R. E. (1982). Time-On-Task: Issues of timing, sampling, and definition. *Journal of Educational Psychology, 74(6),* 844-851.

Kovanovic, V., Gasevic, D., Dawson, S., Joksimovic, S., Baker, R.S., Hatala, M. (2015) Penetrating the Black Box of Time-on-task Estimation. *Proceedings of the 5th International Learning Analytics and Knowledge Conference*, 184-193.

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of Educational Psychology, 59(5),* 320-324.

Lee, S. W., Kelly, K. E., & Nyre, J. E. (1999). Preliminary report on the relation of students' on-task behavior with completion of school work. *Psychological Reports*, 84, 267 - 272.

Marks, H. (2000). Student engagement in instructional activity: patterns in elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153-184.

McKinney, J.D., Mason, J., Perkerson, K., & Clifford, M. (1975). Relationship between classroom behavior and academic achievement. *Journal of Educational Psychology, 67*, 198 - 203.

Ocumpaugh, J., Baker, R.S., & Rodrigo, M.M.T. (2015). *Baker Rodrigo Ocumpaugh Method Protocol (BROMP) 2.0. Technical and Training Manual*. Technical Report. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Samuels, S. J. & Turnure, J. E. (1974). Attention and reading achievement in first-grade boys and girls. *Journal of Educational Psychology,* 66(1), 29 - 32.

Wiley, D. E., & Harnischfeger, A. (1974). Explosion of a myth: Quantity of schooling and exposure to instruction, major educational vehicles. *Educational Researcher,* 3,7-12.

The effect of "mood": Group-based collaborative problem solving by taking different perspectives

Yugo Hayashi (y-hayashi@acm.org)
College of Comprehensive Psychology, Ritsumeikan University,
2-150 Iwakura-cho, Ibaraki, Osaka, 567-8570, Japan

Abstract

Collaborative problem solving based on different perspectives is an effective strategy for constructing new knowledge and discoveries. It remains unclear what kind of interaction process underlies development of an abstract or integrated perspective upon experiencing conflict with different perspectives in a group. The present study investigates two factors in an experimental setting: (1) groups with a single opposing perspective (maverick) would hold an advantage over groups and (2) groups with positive moods would hold an advantage over groups with negativity. We investigate the factors influencing perspective taking in problem-solving groups using conversational agents. Results showed that (1) a single different perspective in the group can be accepted for perspective taking compared to several members with an opposing perspective, and (2) positive mood generated by group members facilitating perspective taking compared to negative mood.

Keywords: Collaborative problem solving; minority influence; emotion and cognition; perspective taking; conversational agents.

Introduction

Studies in collaborative problem solving have found that the use of meta-cognition can be promoted by externalizing the problem from different perspectives (Chi, Leeuw, Chiu, & Lavancher, 1994; Lombrozo, 2006; Miyake, 1986) and asking reflective questions (Okada & Simon, 1997). Dunbar (1995) investigated the use of inductive reasoning in a scientific research group and developed the concept of distributed reasoning, in which group members achieve their goals by taking charge of different types of inferences. New perspectives and knowledge can be discovered or generated by integrating different views and abstraction (Schwartz, 1995). The present study focuses on collaborative problem solving based on different perspectives in groups where individuals can discover a new solution by combining others' knowledge and perspectives. Past studies have shown that in such situations, conflicts occur upon emergence of a different perspective in the group, making problem solving difficult (Hayashi, Miwa, & Morita, 2006). It is important to understand how individuals adopt the different perspective and generate a higher abstract understanding to solve the problem. However, it remains unclear the cognitive process underlying development of and motivation towards this abstraction of different perspectives in collaborative problem solving. Accordingly, this study focuses on the cognitive and affective factors, such as the presence of a single different perspective in the group and the affective mood, that are critical in perspective taking and integration of perspectives.

A single member with a different perspective vs. multiple members

Studies in group problem solving have shown that conflict may occur at several levels, such as task conflict, interpersonal conflict, and process conflicts (de Wit, Greer, & Jehn, 2011). In collaborative problem-solving settings where a member encounters members with a different perspective, it is likely s/he will experience conflict (Hayashi et al., 2006). Studies from social psychology have shown that negative moods brought on by conflicts can emerge when encountering members with different perspectives (De Dreu & Weingart, 2003). Group members become dissatisfied when they interpret challenges to their viewpoints by other group members as a negative assessment of their own abilities and competencies (Swann, Polzer, Seyle, & Ko, 2004). Research in social psychology has also discovered that a "maverick" can encourage other group members to contemplate different perspectives (Elliston, Keenan, Lockhart, & Van Schaick, 1985; Near & Miceli, 1987). A maverick provides dissent and opposing views about the organization and the social system. S/he can be seen as a person who may cause trouble and confusion, or can be interpreted as a person that can bring innovative ideas as a reformer in the group. Research on group dynamics shows that such mavericks can influence other group members to reconsider different views. Members with little power and anti-normative positions can influence the majority of group members, a phenomenon called "minority influence" (Moscovici, Lage, & Naffrechoux, 1969; Nemeth, Brown, & Rogers, 2001). This study shows that a different perspective can be effective when it is presented in a small quantity. Such a minor opposing perspective can provide important information but with less conflict, thereby producing a better outcome. Hayashi (2012) has shown that the use of a maverick can play an important role in perspective taking in a group problem-solving task. The study investigated these cognitive processes through long-term verbal protocol analyses by controlling the number of partners and their utterances in laboratory-based experiments. This previous study found that a mere single other perspective will provide less conflict and better accessibility to the different source of information. However, the cognitive mechanisms involved when individuals pay attention to these mavericks and use their perspectives remains unclear. Further, in the previous study, a significant difference was not found at several levels of perspective-taking performance between members with the same number of different perspectives. By conducting a follow-up experiment, we may be able to clarify our hypothesis that maver-

icks, who produce less conflict and are accessible, bring constructive thoughts toward development of abstract perspectives(integration). Here, we conduct a new experiment by focusing on the effect of the maverick along with new aspects that relate with cooperative thinking, such as group mood.

The effect of mood in the group

In collaborative problem solving, others with different perspectives are liable to bring conflict, thereby producing negative impressions among others. Small group research has shown that negative mood in the group suppresses productive interaction. On the other hand, despite conflict during interaction, if members retain a positive mood within the group, it may function to facilitate active behaviors, such as directing more attention towards different perspectives and motivating consideration of others' perspectives. Therefore, even if one experiences conflict with others with different perspectives, a positive mood may aid in considering an alternative perspective. Previous studies investigating positive effects on task performance have found that partners' positive utterances facilitate more task engagement (Martenes, Bradley, & Eckert, 1997) and flexible thinking (Bless, 2000). Pleasant feelings towards others may also facilitate perspective taking (Weinstein, Hodgins, & Ryan, 2010). It can be predicted that even though members encounter others with opposing views, if members are motivated to interact in a way that promotes greater cooperative behavior, they may try to integrate different perspectives and abstraction. Such motivational effects that are generated during interaction may play an important role and could be a strong factor, over and above the effects of interacting with a member with a mere different perspective.

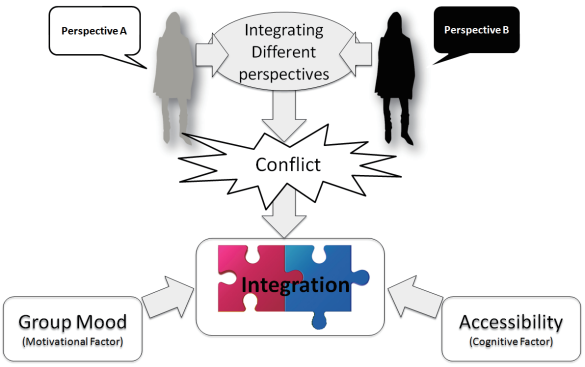


Figure 1: Research framework

Given the above, this study investigates two factors: (1) groups with a single opposing perspective (maverick) and (2) groups with a positive mood. We investigated factors influencing perspective taking in problem-solving groups.

Study goals

The first goal of the present study is to discover the power of groups with a single different perspective (maverick) and examine the strength of this influence on groups with several members. Second, based on the implication above, we examine whether groups with positive feelings provide an opportunity for the whole group to think more flexibly and motivate others to more frequently consider differing perspectives. Our goal is to investigate the next two points:

1. Investigate the effect of a maverick(5:1) in comparison with members with same number(3:3) of perspectives.
2. Investigate how emotional states shared by the members affect the perspective-taking process and impressions among others.

We predict that a mere member with a different perspective will produce conflict compared to groups with several members, and that it will be easier to access the dissenting information. We also focus on the motivational factors based on previous studies showing that affective mood may play a role in facilitating cooperative behavior. Our hypotheses are as follows.

H1: A mere different perspective in the group will facilitate perspective taking compared to several members with an opposing perspective.

H2: Positive mood generated by group members will motivate higher cooperative states, thereby facilitating perspective taking compared to negative mood.

Method

The rule discovery task

The current study uses an experimental paradigm designed by Hayashi et al. (2006). In this paradigm, two participants engage in a rule discovery task, and each of them engages in the task with a different perspective. While solving the task, each of the participants confronts conflict about a different perspective proposed by the other member and has to develop an abstract perspective of the self and other to discover the rule. In this task, the participants' goals were to count a series of objects presented on a computer display and discover its sequential rule. However, as will be explained in the following section, these other "members" were manipulated by the conversational agent to control their behavior. Materials for the objects were several sets of stimuli with white and black unit squares randomly arranged on a 6×6 grid (colored black or white; see Figure 2). In each set, a pattern consisting of combined square blocks was shown against either a black or white background. Using the Gestalt effect from perception psychology, the background color was controlled to change in the problem-solver's perspective (Koffka, 1935).

Each set consisted of several "objects"(or patterns) in black or white, each of which consisted of a single block or multiple blocks. As shown in Figure 2, one of the paired objects has a total of 10 "components" (4 black and 6 white). When a participant focuses on the white components inside a black

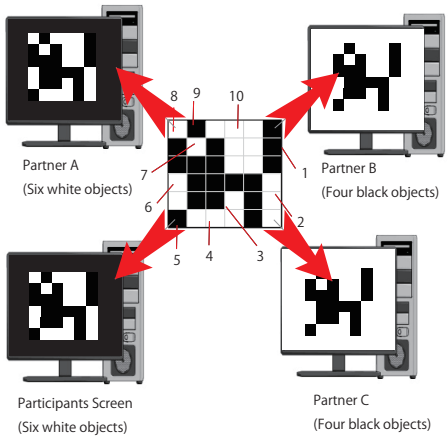


Figure 2: Experimental stimuli used in the study(from Hayashi (2013)).

background, the white components become the figure, and the six components pop out. Conversely, when a participant is focusing on black components inside a white background, it becomes the figure. Participants, collaborating through computer terminals, were separated so that neither could see the partner's display. They collaboratively worked on the task through computer terminals connected via a local network. All members exchanged their opinions about the number of objects and the sequence of the target rule by text-based chat messages. Participants were informed that they would conduct the task with real humans inside the room and were told not to chat about anything else related to the task. As shown in Figure 2, six members(including the conversational agents) saw the objects with either a black background (white perspective) or a white background (black perspective). Three members saw a black and the other three saw a white background, and each member counted the number objects that appeared as a figure. For each trial/object, a square outer box was shown on the display for 1 second, which was followed by a stimulus picture presented inside the box frame. Participants were required to send one message per trial to the other members. Participants were told that they had to count the number of objects inside this box frame. Based on the experimental settings of Hayashi (2012, 2013), the number of white and black components was controlled so the total number of components presented to the participants varied between 6 and 12. The sequential pattern (target rule) of the sums of black components and white components was repeatedly presented during each trial (i.e. 6, 8, 10, 12 or 6, 8, 10, 12; see Figure 3). In the initial stage of the task, the local black/white numbers were controlled to be the same. Experiments from a previous study showed that participants are fixated to the figure perspective, and participants try to search for the target rule based on their perspective (Hayashi et al., 2006; Hayashi & Miwa, 2009; Hayashi, 2012, 2013). On trial 9, we controlled the trials for differentiation in responses so that participants would report different numbers. Through this manipulation, participants experience conflict. To discover the rule for this task, the participant has to look at the different colored objects and integrate the other two members' perspectives.

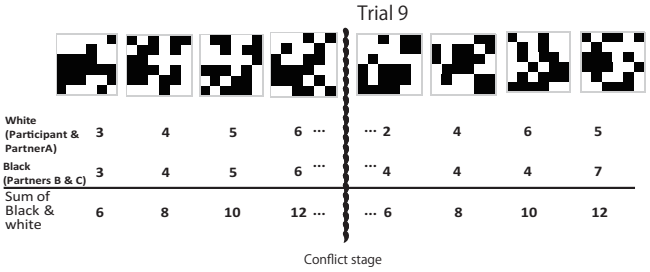


Figure 3: Sequence of objects used in Hayashi (2013).

Experiment design

Controlled factors

The experiment has a 2 × 2 between-subjects factorial design. The two factors were number of different perspective (single(5:1) vs. multiple(3:3)) and the mood(positive vs. negative). The first factor was controlled by arranging the number of partners(agents) looking at the different perspective. The second factor was controlled by manipulating partners' (agents') conversations. In the positive mood condition, the partner was manipulated to include messages that provides positive phrases such as, "I like this group" and "It is interesting to see others with different perspectives." On the other hand, in the negative mood condition, the conversational agents were manipulated so that they would include negative phrases such as "I don't like this group" and "It is quite frustrating to work with others with different perspectives."

Participants

Undergraduates in a psychology class participated in the experiment. All participants were randomly assigned to each condition and participants who recognized their partner's (agent's) purpose and those who did not complete the task or refused to continue the task were excluded from analysis. After this exclusion, the following number of participants were arranged to each condition: 5:1/positive: 40, 5:1/negative: 50, 3:3/positive:50, and 3:3/negative:37.

Procedure

Conversational agents

A text-based chat communication platform using one server and clients, including one chat engine and five conversational agents(Figure 4), was utilized. The system developed by Hayashi (2012, 2013) was programmed in Java. On the server side, a broadcast mechanism was used to distribute all the messages simultaneously. When messages were sent to the server, they were re-distributed to all clients and/or agents.

The conversational computer agents were designed by a typical rule-based system. Based on pre-defined rules, the system can respond to sentences that were inputted by the participants (Figure 4).

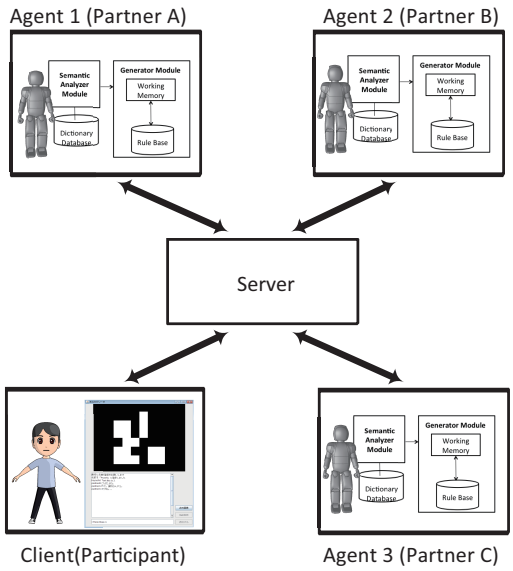


Figure 4: Communication platform(from Hayashi (2013)).

All three agents were implemented by the rule shown below. All agents autonomously responded to each other's text messages as independent interlocutors, and had different expressions. The conversation agent extracts keywords from the sentences that were distributed by the participant. The most frequent keywords used during this task are related to the (1) number of objects, (2) colors of objects, and (3) rules about the sequence. The agent contains a temporary working memory storage to represent the current status of the input messages from the (1) keywords of the participant, (2) keywords from the agents, and (3) objects presented on the screen. A rule base, in an "if-then" format, defined all responses from the agents. When the agent detects keywords of (a) numbers, (b) colors, and (c) the hypothesis, working memory is updated. Then, a pattern-matching strategy is used for binding the rules. In the present study, the agents were programmed to respond based on the numbers or colors of the objects that were set for its perspective (Agents 1 and 2 responded based on the black objects, which was the same as the participant. Agents 3, 4, and 5responded based on white objects, which was different from the participant). The following shows a simple example of some of the basic rules utilized in the study:

$< Rules >$
$Trials < 8$
$(agent\ 1,2)$
$-> numbers(white\ objects)$
$(agent\ 3,4,5)$
$-> numbers(black\ objects)$
$Trail > 8$
$(agent\ 1,2)$
$IF : "numbers(any)" -> numbers(white)$
$ELSE\ IF : "colors" -> colors(white)$
$ELSE\ IF : "rules" -> rules$
$ELSE -> numbers(white)$
$(agent\ 3,4,5)$
$IF : "numbers(any)" -> numbers(black)$
$ELSE\ IF : "colors" -> colors(black)$
$ELSE\ IF : "rules" -> rules$
$ELSE -> numbers(black)$

For example, if a sentence by the participant inputs "I think this is four objects," indicating four as black objects, agent 1 will respond based on the number of objects such as "Well, I think this is four, too." If the inputs such as "Is it the black objects you are counting ?", agent 1 will respond "I count white objects." As noted in the previous section, the agents in the positive mood conditions responded with additional positive phrases, and negative phrases were used in the negative mood condition.

Dependent variables

Two variables were used to investigate the effects of the two factors. The first index is the evaluation towards their partners(agents) to evaluate the effectiveness of the manipulation of the agents producing positive/negative mood. It assessed the degree of pleasantness and the importance of developing a solution. They were asked to rate the following statement on a 5-point scale, ranging from 1 (extremely unpleasant/inefficient) to 5 (extremely pleasant/efficient): "The degree of pleasantness felt while working with this member." The second index asked for the description of the target rule on an answer sheet. If their answers were in some way related to integrating the number of black or white components of the objects, they were judged as 'integrated' (e.g., The sum of black and white repeat by 6, 8, 10, and 12). The difference in the number of the black and white components is between zero and two.) The ratio of participants in each condition was collected.

Results

Evaluation of others

Figure 5 indicates the results of the evaluation. A $2 \times 2 \times 5$ mixed-effects ANOVA was conducted on the average scores of evaluation of the partner, with perspective (5:1 vs. 3:3) and emotion (positive vs. negative) as between-subject factors and evaluation (partner 1 vs. 2 vs. 3 vs. 4 vs. 5) as a

within-subject factor. The second-order interaction was not significant ($F(4, 680) = 2.06$, $p = .036$). There was significant interaction between perspective and evaluation ($F(4, 680) = 2.92$, $p < .05$). Consequently, an analysis of the simple main effect was conducted on each evaluation. Focusing on the affective state, the rating score towards partners 1, 2, 3, 4, and 5 in the positive condition was significantly higher than it was in the negative conditions($F(1, 850) = 42.65$, $p < .001$; $F(1, 850) = 18.40$, $p < .001$; $F(1, 850) = 14.86$, $p < .001$; $F(1, 850) = 24.90$, $p < .001$; and $F(1, 850) = 18.73$, $p < .001$, respectively).

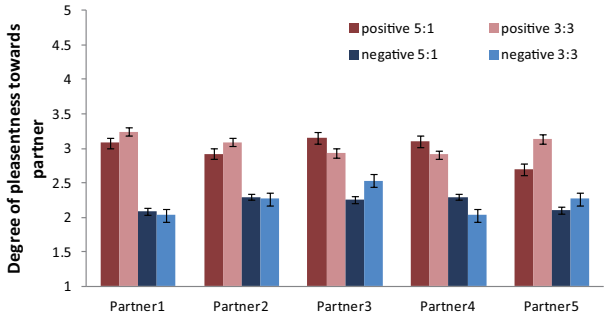


Figure 5: Results of the evaluation of others.

These results show that only the emotional mood factor influenced evaluation of group members. This indicates that the manipulation of collaborative members' positive mood produced a strong effect on making the participants feel pleasant towards their partners. Next, we examine how the two factors affected problem-solving performance and determine whether the two hypotheses are confirmed.

Performance

Figure 6 describes the problem-solving performance. The vertical axis represents the ratio of participants who succeeded in integrating the perspective on their final answers, and the horizontal axis represents each experimental condition. Our interest is to investigate how the two factors influence performance. Therefore, we conducted an ANOVA using the χ^2 distribution based on the arcsine transformation method. This method enables detecting both the main effects and interaction of the two experimental factors. The analysis was performed by a 2×2 ANOVA with the perspective (5:1 condition vs. 3:3 condition) and emotion (positive condition vs. negative condition) factors as between-subjects variables. There was a main effect of both perspective and emotion ($\chi^2(1) = 7.65$, $p < .001$; $\chi^2(1) = 6.77$, $p < .001$). The performance in the experience condition was better than was that in the no experience condition, and the performance in the conversation condition was also better than was that in the chat condition. In addition, there was no interaction between the two factors ($\chi^2(1) = 0.04$, $p = .85$).

The main effects of the two factors indicate that the single perspective and positive mood contribute to successful problem solving. This result indicates that both **H1** and **H2** were supported.

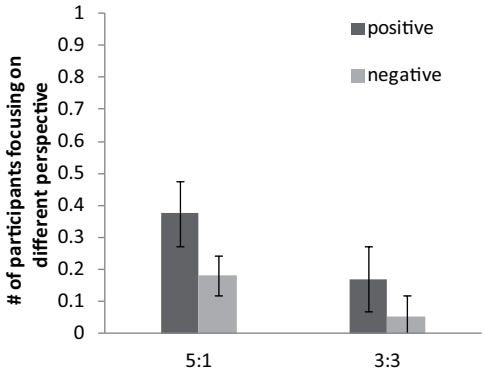


Figure 6: Results on perspective-taking performance.

Discussion and Conclusion

The present study investigated the two factors that influence perspective integration in a group-based collaborative problem-solving task. The results from the questionnaire show that participants were influenced by the emotional state that was constructed in the group. That is, participants engaging in a group with positive members felt more pleasant to partners compared to those with negative members. This was consistent even if the number of members with different perspectives increased in the group. Interestingly, even when participants received the same amount of information with the same rule base, with the only difference being positive/negative phrases, the impressions towards others changed drastically. From the point of system development for using conversational agents as experimental tools, it shows new implications that mood can be controlled when using multiple agents. This provides advanced implications from the work by Hayashi (2012, 2013). For the main results of the study, performance results show that both factors were effective. Participants performed better when they (1) encountered members with different perspectives less in the group(5:1) compared to those with several members, and (2) engaging in a group with positive members tended to integrate different perspectives as compared to those with negative mood. For (1), this provides stronger evidence compared to Hayashi (2012), where there was no clear difference between the two conditions(5:1 vs. 3:3) in terms of problem-solving performance. This is interesting in the sense that participants have more opportunity to interact with members with the different perspective in the 3:3 condition as compared to those with a single member in the 5:1 condition, but integration is better in the latter. A phenomenon found by attending to a fraction of the available information is the discovery of "anomalies," which are crucial for developing and generating new theories under stalemated conditions (Chinn & Brewer, 1993). The present findings could be interpreted in that individuals could

have considered the potential of a maverick functioning as an anomaly in problem solving. Either way, it can be pointed out that the mere single individual with a different perspective provides less conflict and motivates others to consider it as a potential, and that this can be related with emotional states. For the results of (2), this indicates that under situations where members meet with conflicting perspectives, the positive mood created by group members facilitates active behaviors and motivates consideration of their perspectives. Such results are consistent with previous work in group studies showing the effects of group emotion (Martenes et al., 1997; Bless, 2000; Weinstein et al., 2010). The current results add new insight in terms of the effect of affect over conflictive situations demonstrated in this study. This result shows how the power of social influence plays an important role in cognitive activities and thus could influence performance. Further, taken together with the results of the mere single member's effect, it is possible that two factors can produce a synergetic effect. That is, if problem-solvers perceive the mere individual, it may produce positive mood, which could then facilitate better performance. Focusing on this point, further analysis could be conducted on verbal inputs and processes. Such attempts could provide implications on how the emotion in groups moderates the problem-solving process and its role in social cognition.

Acknowledgments

This work was supported (in part) by The Ministry of Education, Culture, Sports, Science, and Technology, Japan (MEXTGrant), Grant No. 16K00219.

References

Bless, H. (2000). *The interplay of affect and cognition: The mediating role of general knowledge structures.* (J. P. Forgas, Ed.). New York: Cambridge University Press.

Chi, M., Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439-477.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1-49.

De Dreu, C. K., & Weingart, L. R. (2003). Task versus relationship conflict, team effectiveness, and team member satisfaction: A metaanalysis. *Journal of Applied Psychology*, *88*(4), 741-749.

de Wit, F. R. C., Greer, L. L., & Jehn, K. A. (2011). The paradox of intragroup conflict: a meta-analysis. *Journal of Applied Psychology*, *97*(2), 360-390.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In J. R. Sternberg & E. J. Davidson (Eds.), *The nature of insight* (p. 365-395). MIT Press.

Elliston, F., Keenan, J., Lockhart, P., & Van Schaick, J. (Eds.). (1985). *Whistleblowing research: Methodological and moral issues.* New York: Praeger.

Hayashi, Y. (2012). The effect of "maverick":a study of group dynamics on breakthrough in collaborative problem solving. In *Proceedings of the 34th annual conference of the cognitive science society(cogsci2012)* (p. 444-449).

Hayashi, Y. (2013). The effect of "trust dynamics": Perspective taking during collaborative problem solving. In *Proceedings of the 35th annual conference of the cognitive science society(cogsci2013)* (p. 2500-2505).

Hayashi, Y., & Miwa, K. (2009). Prior experience and communication media in establishing common ground during collaboration. In *Proceedings of the 31th annual conference of the cognitive science society(cogsci2009)* (p. 526-531).

Hayashi, Y., Miwa, K., & Morita, J. (2006). A laboratory study on distributed problem solving by taking different perspectives. In *Proceedings of the 28th annual conference of the cognitive science society(cogsci2006)* (p. 333-338).

Koffka, K. (1935). *Principles of gestalt psychology*. Routledge and Kegan Paul.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 1454-1463.

Martenes, B. K., Bradley, T. A., & Eckert, T. L. (1997). Effects of reinforcement history and instructions on the persistence of student engagement. *Journal of Applied Behavior Analysis*, *30*(3), 569-572.

Miyake, N. (1986). Constructive interaction and the interactive process of understanding. *Cognitive Science*, *10*(2), 151-177.

Moscovici, S., Lage, E., & Naffrechoux, M. (1969). Influence of a consistent minority on the responses of a majority in a color perception task. *Sociometry*, *32*(4), 365-380.

Near, J. P., & Miceli, M. P. (1987). Whistle-blowers in organizations: Dissidents or reformers? *Reserch in Organizational Behavior*, *9*(4), 321-368.

Nemeth, C., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: stimulating quatity and quality. *European Journal of Social Psychology*, *31*(6), 707-720.

Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, *21*(2), 109-146.

Schwartz, L. D. (1995). The emergence of abstract representation in dyad problem solving. *Journal of the Learning Science*, *4*, 321-354.

Swann, J., W. B., Polzer, J. T., Seyle, D. C., & Ko, S. J. (2004). Finding value in diversity: Verification of personal and social self-views in diverse groups. *Academy of Management Review*, *29*(1), 9-27.

Weinstein, N., Hodgins, H. S., & Ryan, R. M. (2010). Autonomy and control in dyads: Effects on interaction quality and joint creative performance. *Personality and Social Psychology Bulletin*, *36*(1), 1603-1617.

The Impact of Interactivity on Simulation-Based Science Inquiry with Variable-Setting Controls

Jung Aa Moon (jmoon001@ets.org)
Educational Testing Service
Princeton, NJ 08541 USA

Michelle LaMar (mlamar@ets.org)
Educational Testing Service
San Francisco, CA 94105 USA

Carol M. Forsyth (cforsyth@ets.org)
Educational Testing Service
San Francisco, CA 94105 USA

Madeleine Keehner (mkeehner@ets.org)
Educational Testing Service
Princeton, NJ 08541 USA

Abstract

The current study investigated how interactivity of simulation controls affects data collection in science inquiry. A chemistry simulation was designed to allow either low or high interactivity in setting experimental variables. Adult participants were randomly assigned to one of the interactivity conditions and solved a series of assessment items. The results from the first item indicated that the highly interactive controls posed challenges in conducting a thorough investigation. Performance in the last item which is a repetition of the first item suggested that the participants were able to overcome the initial challenges over the course of their investigations. The results provide implications for designing educational simulations for learning and assessment.

Keywords: interactivity; simulation; science inquiry; education; assessment

Introduction

Traditional science assessments conducted on a large scale have until recently been limited mostly to the paper-and-pencil format (e.g., Shavelson, Carey, & Webb, 1990). These assessments often measured test-takers' prior knowledge of science facts and principles by collecting their final responses to multiple choice items. In the last few decades, however, criticisms of the traditional assessments and advances in educational technology have spurred a growing interest in using simulation-based tasks for science assessments (Gobert, Sao Pedro, Raziuddin, & Baker, 2013; Wieman, Adams, & Perkins, 2008).

The simulation-based assessments typically provide test-takers with interactive tools that support multiple science inquiry steps such as designing experiments and collecting data to test a scientific hypothesis. The interactive nature of these simulations raises a question of how simulation interactivity impacts cognition and educational outcomes. Answering this question is important for designing simulation environments where test-takers or learners have a sufficient level of freedom to conduct science inquiry. It is also important for ensuring that simulation interactivity does not hinder valid measurements of knowledge and skills.

Interactivity

The topic of interactivity has been actively researched in the field of multimedia learning. While there are various approaches to how interactivity can be interpreted (Domagk, Schwartz, & Plass, 2010), a number of studies have associated interactivity with the amount of control learners have over various aspects of the learning system (Kalyuga, 2007; Moreno & Mayer, 2007). For instance, studies have manipulated whether learners can control the sequence of the learning materials (Swaak & de Jong, 2007) and whether they can select their own answers rather than to receive the correct answers selected by the system (Moreno & Mayer, 2005). While some studies have found that learner-controlled learning results in better outcomes than system-controlled learning (Hasler, Kersten, & Sweller, 2007; Moreno & Valdez, 2005), others suggest negative outcomes associated with the greater learner control (e.g., Moreno & Valdez, 2005; Tuovinen & Sweller, 1999).

Cognitive load theory (Sweller, 1994) provides accounts for those seemingly mixed findings on interactivity. The main claim of the theory is that cognitive load irrelevant to learning results in negative learning outcomes as it takes away limited cognitive resources that could otherwise be used for learning. On the other hand, cognitive load put into creating schemas (i.e., cognitive structures that allow organization of information) is beneficial for learning. One implication of the cognitive load theory for interactivity research is that interactive systems can have positive outcomes if they engage learners in deeper cognitive processing germane to learning (Hasler, Kersten, & Sweller, 2007; Kalyuga, 2007).

Interactivity in Science Inquiry

The prior research on interactivity raises a question of how those findings apply to science simulations which typically provide individuals with a fair amount of freedom to conduct science inquiry. In particular, one of the distinctive features of simulation-based science assessments is that they allow test-takers to collect their own data for scientific hypothesis testing. Test-takers can make their own decisions on various aspects of data collection such as how much data to collect and how exhaustively to search the experimental space. This contrasts to traditional assessments in which test-takers are typically asked to draw a conclusion based on the data prepared by assessment developers.

Collection of adequate data is one of the critical inquiry skills involved in scientific investigations (Kuhn, Schauble, & Garcia-Mila, 1992). Prior research found that people often exhibit suboptimal data collection behaviors. For instance, people often jump to a conclusion based on limited observations without sufficiently sampling variables (Harrison & Schunn, 2004; Kuhn et al., 1992). To understand how simulation interactivity impacts data collection behaviors and inquiry outcomes, we manipulated the design of slider/button controls in a chemistry simulation. While there is a fair amount of human-computer interaction research on how design of slider controls impacts user behaviors (e.g., Roster, Lucianetti, & Albaum, 2015), few studies (e.g., Renken & Nunez, 2013) assessed its impacts on science inquiry behaviors.

Concentration Simulation

Concentration simulation developed by PhET (Wieman, Adams, & Perkins, 2008) was modified for the purpose of our study. The simulation is an HTML5 application written in JavaScript and delivered through a standard web browser. The simulation (Figure 1) allows one to design and run simulation trials to investigate the relationship among solute, water, and concentration level. The top left panel displays the slider and button controls used to set the amounts of solute and water. Once a trial is run, the solute and water are mixed to form a solution. The resulting concentration level of the solution and the variable settings for the trial appear in the data table located in the bottom left panel. In a typical screen, the right panel displays an assessment item and response options and/or text entry.

Method

Participants

Adult participants (N = 308) recruited through Amazon Mechanical Turk completed the study for monetary compensation ($5). There was no particular entry condition for participation. Background information about participants was collected in a separate survey to which 248 of the participants responded (111 females and 137 males, mean age 35, age ranged from 20 to 61). Participants were randomly assigned to the interactivity conditions.

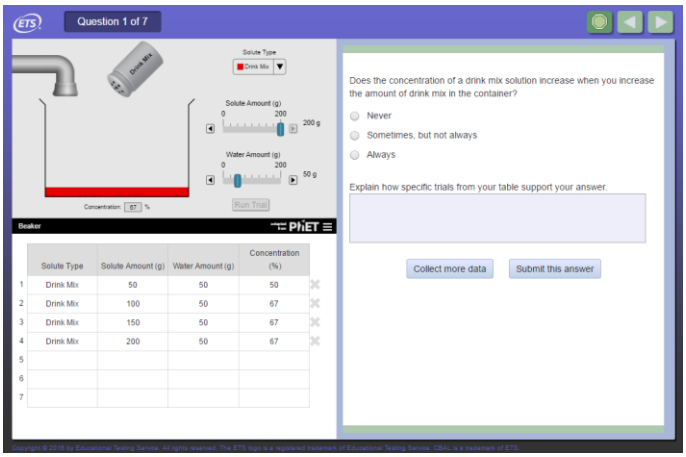


Figure 1. Concentration simulation.

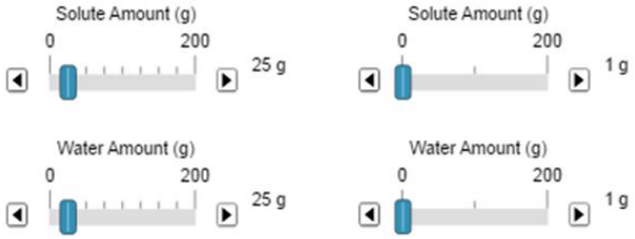


Figure 2. Variable-setting controls in the low (left) and high (right) interactivity conditions.

Design

The study involved a between-subject design in which three features of the simulation were manipulated to allow different levels of interactivity. Those features were: 1) the slider and button controls, 2) the amount of the data one could keep in the data table, and 3) the capability to re-order the rows in the data table. Due to the space limit, the current paper focuses on discussing the results of the slider/button control manipulation. These controls (Figure 2) were manipulated to allow different amounts of control over the values one could choose for the solute and water variables. While the range for each variable was the same in both conditions (0-200 g), the smallest amount of adjustment one could make using the controls was 25 g in the low interactivity condition and 1 g in the high interactivity condition. This manipulation offered a far greater number of choices for the high condition than for the low condition. While the high condition participants could set any of the 201 integer values for each variable, the low condition participants' choices were limited to 9 values (0, 25… 200).

In order to adjust a variable, participants could drag the slider (blue rectangle in Figure 2) to a desired location on the scale and release it. Once released, the slider automatically snapped to the closest value available in the relevant condition. Participants could also click the tweak buttons (buttons with an arrow sign) located next to each scale. A button click incurred 25 g of change (left button for decrement, right button for increment) in the low condition

and 1 g of change in the high condition. In the low condition, a tick mark was placed at every available value on the scale. In the high condition, only three tick marks were placed (0, 100, and 200 g) due to programming constraints and space limit. We intentionally chose not to make the visuals of the sliders the same for the two conditions because of a concern that doing so would bias the participants in the high condition prefer those particular values with tick marks.

Experiment

The experiment consisted of a brief tutorial session followed by a main session. During the tutorial session, participants were introduced to the terminology used in the simulation and were familiarized with various simulation features including the slider and button controls. During the main session, participants solved seven assessment items presented one at a time in a self-paced manner. Participants were asked to use the simulation to conduct investigations necessary for answering each item. The order of the seven items was the same for the two conditions.

These items were designed to assess understanding of the impacts of solute and water amounts on the concentration level. All the items were designed around the concept of saturation (i.e., concentration[1] does not further increase once maximum is reached). The first and last items were identical except for their relative locations in the item series. These two items were different from the five items in the middle in two major aspects. First, some of the middle items directed participants to investigate the impacts of the variables on the concentration level under a specified condition such as with a larger range of solute values or with a smaller amount of water. These directed investigations were expected to facilitate observation of saturation. Second, the first and last items involved a solute type ("drink mix") different from the one used in the middle items ("chemical A"). While the two solutes are different, we expected that some knowledge gained from investigations with one solute could be transferred to investigations with the other solute. In a sense, the first and last items can be viewed as pre- and post-tests that allow assessment of learning gained through investigations. Due to space limit, the current paper focuses on the results in these two items. Participants' performance in the two conditions did not significantly differ in any of the middle items.

Task

In the first and last items, participants were asked to test the following statement: "Does the concentration of a drink mix solution increase when you increase the amount of drink mix in the container?" Participants were asked to choose one of the three response options ("never", "sometimes, but not always", and "always") and explain how specific trials from the data table support their answer by typing their responses to the text entry. The correct answer is "sometimes, but not always" because adding more drink mix does not further increase the concentration level once the saturation point is reached.

Success in these items largely depends on the ability to sample sufficient ranges of the variables. Insufficient sampling is likely to lead to either "always" (concentration keeps increasing before the saturation point is reached) or "never" response (no further increase in concentration past the saturation point). In order for the saturation point to be observed, the amount of water has to be less than or equal to 75 g. The amount of drink mix that results in saturation depends on the amount of water.

Results

We describe the results on response choice (Table 1) and multiple measures of inquiry behaviors (Table 2). The numbers in the parentheses are adjusted standardized residuals in Table 1 and standard deviations in Table 2. The F-stats in Table 2 are from the repeated measures ANOVA with condition (between) and item (within) as variables.

First Item

Choice of Response Table 1(A) shows how many participants in each condition chose each of the response options. Only 37 of the low and 26 of the high participants chose the correct "sometimes" response. The Chi-square test was significant (χ^2 (2, N = 308) = 8.4, p = .015), indicating the significant effect of being in the interactivity condition on the choice of response. The high participants had a greater tendency to choose the "always" response compared with the low participants. When the incorrect responses ("never" & "always") were combined together, the Chi-square test was not significant (p = .107).

Time on Task The amount of time spent on solving the item did not significantly differ between the low (143.6 s) and high (156.5 s) conditions, t(305) = 1.4, p = .158.

Number of Trials The average number of trials run was significantly greater in the low condition (low: 5.0, high: 4.3, t(303) = 2.5, p = .012).

Table 1: Choice of response

(A) First item

	Never	Sometimes	Always	Total
Low	16 (2.2)	37 (1.6)	100 (-2.7)	153
High	6 (-2.2)	26 (-1.6)	123 (2.7)	155
Total	22	63	223	308

(B) Last item

	Never	Sometimes	Always	Total
Low	6 (1)	90 (1)	57 (-1.4)	153
High	3 (-1)	82 (-1)	70 (1.4)	155
Total	9	172	127	308

[1] Before saturation is reached, concentration is mass of solute divided by mass of solution (i.e., solute plus water).

Table 2: Results on inquiry measures.

	First item		Last item		F-stat		
	Low	High	Low	High	Condition (C)	Item (I)	C * I
Time on task (second)	143.6 (71.7)	156.5 (87.6)	143.6 (82.6)	151.6 (96.5)	1.1	.1	.4
Number of trials	5.0 (2.8)	4.3 (2.1)	7.5 (4.2)	6.9 (4.7)	5.1^{*}	96.7^{*}	.0
Solute sampling range	104.2 (54.3)	91.0 (52.4)	139.7 (50.7)	138.8 (50.5)	2.8^{+}	108.8^{*}	2.9^{+}
Water sampling range	32.3 (50.0)	24.3 (46.2)	59.8 (61.6)	58.4 (63.6)	.9	56.0^{*}	.9

*: $p < .05$, $^{+}$: $p < .10$

Sampling Range For each participant who ran at least one trial, the sampling range of each variable was obtained by getting the minimum and maximum values sampled across all trials and subtracting the former from the latter. The average sampling range of the solute variable (black bars in Figure 3) was significantly greater in the low participants (104.2) than in the high participants (91.0) for the first item, $t(303) = 2.2$, $p = .032$. The average sampling range of the water variable (white bars) did not significantly differ between the low (32.3) and high (24.3) conditions, $t(303) = 1.5$, $p = .145$.

The results above suggest that while overall performance did not significantly differ between the conditions, the high participants collected less data and searched a narrower experimental space. A further analysis was performed to understand what might have contributed to their suboptimal data collection behaviors.

Preference for Round Numbers Nowhere in the item were participants asked to try any particular amount of solute or water. The solute and water amounts set by the high participants revealed that even though they could select any of the 201 integer values, they preferred "round" numbers such as multiples of 50 or 100. Across all the trials run by the high participants, 50, 100, and 200g were the three most frequently selected values for both solute and water. These three values accounted for 28% of the solute values and 70% of the water values, which is much higher than chance (.5%). The low participants also frequently selected the three values (43% in solute, 74% in water).

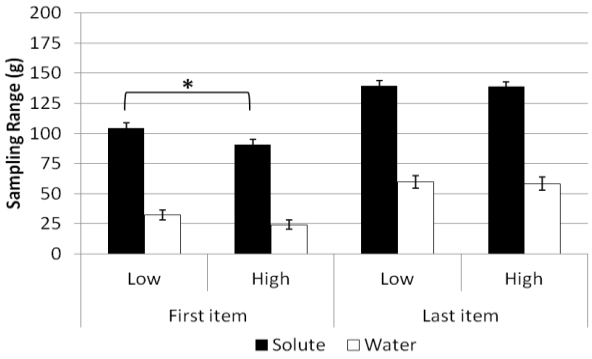


Figure 3. Average sampling ranges of solute and water. The error bars are the standard error of mean.

Interactions with Simulation The preference for the round numbers cost the high participants frequent interactions with the simulation. To adjust solute or water, participants could drag the slider, click the tweak buttons, or combine the two actions. While the two conditions did not differ much in the average number of drag actions (low: 5.4, high: 5.7), the high participants clicked the tweak buttons much more frequently (low: 2.5, high: 22.8). The condition variable was a significant predictor of tweak button click counts in a negative binomial regression (Wald Chi-square ($df = 1$) = 66.4, $p < .001$). Due to the limited interface space, it was almost impossible for the high participants to make very fine adjustments with the slider alone. A likely strategy for setting a round number is to drag the slider close to the desired value and click the tweak buttons to make small adjustments.

Did the high participants' tweak button use influence their sampling range? We categorized participants into two groups based on the type of variable-setting actions: "drag-only" group who used the sliders only and "tweak" group who used the tweak buttons at least once. The numbers of participants in the drag-only and tweak groups were 107 and 43 respectively in the low condition, and 57 and 98 respectively in the high condition. The results of ANOVA on the solute sampling range indicated the significant interaction between condition and action type ($F(1, 301) = 5.9$, $p = .016$) and the main effect of condition ($F(1, 301) = 3.9$, $p = .048$). The sampling range did not differ much between the two groups in the low condition (Figure 4). In the high condition, the tweak group (83.5) sampled a significantly narrower range than the drag-only group (103.8), $t(153) = 2.4$, $p = .020$.

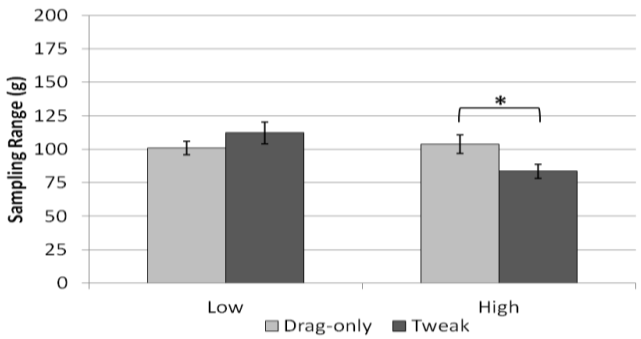


Figure 4. Sampling ranges of solute for the first item. The error bars are the standard error of mean.

The above results suggest that the high participants' preference for the round numbers led to greater workload involved in making frequent interactions with the simulation. One potential interpretation of these results is that experiencing the greater amount of workload deterred the high participants from searching a sufficient experimental space. One may question why the high participants preferred the round numbers despite the workload. A prior study from our laboratory (Koster van Groos & LaMar, 2016) suggests that people choose them because they are easy to remember and keep track of.

Last Item

Choice of Response Compared with the first item, a greater number of participants in both conditions (low: 90, high: 82) chose the correct "sometimes" response in the last item (Table 1B). An exact McNemar's test indicated that the performance improvement was statistically significant ($p < .001$). There was no significant association between the interactivity condition and response choice (χ^2 (2, N = 308) = .98, p = .32).

Time on Task The participants spent about the same amount of time as before (first: 150 s, last: 148 s). There was no significant difference between the low (143.6s) and high (151.6s) conditions, $t(300) = .76$, $p = .443$.

Number of Trials Compared with the first item, the participants ran on average 2.5 more trials in the last item (first: 4.7, last: 7.2). The difference between the conditions was not significant (low: 7.5, high: 6.9, $t(301) = 1.2$, $p = .233$).

Sampling Range The sampling range increased from the first to the last item for both solute (first: 97.5, last: 139.3) and water (first: 28.2, last: 59.1). The difference between conditions was not significant in neither solute (Figure 3, low: 139.7, high: 138.8, $t(289) = .1$, $p = .884$) nor water (low: 59.8, high: 58.4, $t(289) = .2$, $p = .844$) in the last item.

Preference for Round Numbers The participants showed a continued preference for the round numbers. The three most preferred values of 50, 100, and 200 g accounted for about 31% of the solute values and 78% of the water amounts in the high condition. These values accounted for 44% of solute and 73% of water amounts in the low condition.

Interactions with Simulation While the two groups did not differ much in the number of drag actions (low: 7.0, high: 7.5), the high participants used the tweak buttons more frequently (low: 4.4, high: 26.5). The condition variable was a significant predictor of tweak button click counts (Wald Chi-square ($df = 1$) = 39.85, $p < .001$).

In the last item, participants in both conditions showed improved performance in various aspects. A greater number of participants in both conditions selected the correct response. The improved correctness of response choice is consistent with positive changes in their data collection behaviors. While spending about the same amount of time as before, participants collected more data and sampled greater ranges of solute and water.

The results on the sampling range (Figure 3) suggest that the difference between the two conditions became smaller in the last item[2]. The high participants initially sampled a smaller range of solute, yet they later sampled approximately the same range as the low participants. Some other patterns of behaviors, however, did not change much between the two items. The high participants continued to prefer the round numbers despite the greater workload associated with setting those numbers. While the low participants also selected those numbers frequently, their workload is likely to be low because they could set those numbers with a relatively smaller number of button clicks and/or drag actions.

Discussion

The current study investigated how interactivity in variable-setting controls impacts simulation-based science inquiry. Our results suggest that the greater simulation interactivity had initially negative impacts on inquiry performance. Our high participants preferred values that were easier to work with despite the additional workload involved in setting those values. They also ran fewer trials and sampled the experimental variables less exhaustively, likely because experiencing the greater workload hindered thorough scientific investigations.

However, the results in the last item suggest that the initial challenges imposed by the simulation interface became less important over time. On various measures, the low and high participants achieved an equivalent level of performance. It appears that the participants who were initially penalized by the highly interactive simulation interface were able to overcome their challenges over the course of their investigations. One plausible explanation is that observing saturation in the middle items made them realize the importance of sufficiently sampling variables. It is also possible that observing saturation led them to intentionally look for a saturation point in the last item by sampling more exhaustively. The current results alone do not tell us whether the better performance in the last item is due to the improvement of skills, content knowledge, or both. Identifying the contributions of skills and knowledge would be a potentially interesting topic for future research.

Compared with the high condition, the low condition offered much more limited options for variable amounts. One can view that the limited options served as scaffolding for decisions on data collection such as what amount to sample and how exhaustively to sample. The high participants who did not have such scaffolding presumably

[2] Not all participants sampled more than one solute amount in both items. The condition*item interaction in Table 2 is significant when we exclude participants whose sampling range is zero in any of the items, $F(1, 274) = 4.9$, $p < .05$.

had to make those decisions on their own. Their decisions on data collection were less than optimal at least in the first round of their investigations.

While our results appear to suggest negative outcomes of high simulation interactivity, the question of which interactivity level is a better choice cannot be answered without fully understanding the nature of the workload associated with high interactivity. Based on the cognitive load theory (Sweller, 1994), a more appropriate question is whether the extra workload had any relevance to inquiry skills and knowledge. It is possible that our highly interactive controls engaged individuals in deeper cognitive efforts. They had to figure out what is the optimal grain size of data and what is the appropriate sampling range through trials and errors. Their performance improvement suggests that they achieved some learning on those aspects. If the low participants passively followed guidance of the system instead of thinking on their own, their seemingly thorough data collection behaviors may be an overestimation of their genuine inquiry skills and knowledge.

The current discussion focused on the impacts of the variable-setting controls on data collection behaviors. An ongoing analysis is in progress to investigate the impacts of the other two interactivity manipulations (data sorting capability & the amount of data one could keep in the data table) on data collection and organization behaviors. Despite the limited scope, this study provides several implications for designing educational simulations for the purpose of learning and assessment. First, interactive features of simulations can serve as scaffolding that aids cognitive systems to achieve better performance. From the assessment point of view, however, the availability of scaffolding may make it harder to measure genuine knowledge and skills of test-takers. Second, learning and assessment design needs to consider the challenges imposed by simulation interactivity. Especially with highly interactive simulations, providing multiple learning and assessment opportunities seems necessary due to the initial challenges individuals may experience. Overall, the current research suggests the importance of considering how simulation interactivity impacts cognition in learning and assessments.

Acknowledgments

We thank the PhET Interactive Simulations project at the University of Colorado Boulder. Their contributions were funded by the Gordon and Betty Moore Foundation.

References

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, *26*, 1024–1033.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. J. D. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*, 521–563.

Harrison, A. M., & Schunn, C. D. (2004). The transfer of logically general scientific reasoning skills. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 541–546). Erlbaum, Mahwah, NJ.

Hasler, B. S., Kersten, B., & Sweller, J. (2007). Learner control, cognitive load and instructional animation. *Applied Cognitive Psychology*, *21*, 713–729.

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review*, *19*, 387–399.

Koster van Groos, J. M. & LaMar, M. M. (2016). *Separating inquiry goal and skill during analysis of student performance on a simulation-based science assessment.* Poster presented at the annual conference of the American Educational Research Association, Washington, DC.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, *9*, 285–327.

Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*, 117–128.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*, 309–326.

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development*, *53*, 35–45.

Renken, M. D., & Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learning and Instruction*, 23(1), 10–23.

Roster, C. A., & Lucianetti, L. (2015). Exploring Slider vs . Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice*, 11(1), 1–15.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan*, *71*, 692–697.

Swaak, J., & de Jong, T. (2007). Order or no order: System versus learner control in sequencing simulation-based scientific discovery learning. In F. E. Ritter, J. Nerb, T. M. O'Shea, & E. Lehtinen (Eds.), *In order to learn: How the sequence of topics influences learning*. New York, NY: Oxford.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295–312.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*, 334–341.

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. *Science*, *322*, 682–683.

Causal Reasoning in Infants and Adults: Revisiting backwards-blocking

Deon T. Benton (dtbenton@andrew.cmu.edu)
Department of Psychology, 1098 Morewood Avenue
Pittsburgh, PA 15232 USA

David H. Rakison (rakison@andrew.cmu.edu)
Department of Psychology, 1098 Morewood Avenue
Pittsburgh, PA 15232 USA

Abstract

Causal learning is a fundamental ability that enables human reasoners to learn about the complex interactions in the world around them. The available evidence with children and adults, however, suggests that the mechanism or set of mechanisms that underpins causal perception and causal reasoning are not well understood; that is, it is unclear whether causal perception and causal reasoning are underpinned by a Bayesian mechanism, associative mechanism, or both. It has been suggested that a Bayesian mechanism, rather than an associative mechanism, underpins causal reasoning because such a mechanism can better explain the putative backward-blocking finding in children and adults (e.g., Sobel, Tenenbaum, & Gopnik, 2004). In this paper, we report two experiments to examine to what extent infants and adults exhibit backward blocking and whether humans' ability to reason about causal events is underpinned by an associative mechanism, a Bayesian mechanism, or both.

Keywords: causality; infants; adults; causal reasoning; causal learning; causal perception; infant and child development

Introduction

The emergence of causal perception

Of the skills that enable humans to understand the events they experience in the world, perhaps few are as important as the ability to learn about causality in the world. This is a key cognitive ability that enables infants, children, and adults to perceive and encode cause-and-effect relationships and reason about the effects of interventions on those relationships. However, before humans can reason explicitly about causal events through interventions, they must be able first as infants to perceive simple cause-and-effect relations between objects in the world. The ability to perceive cause-and-effect relations has been studied most extensively using Michottian launching events. The reason Michottian launching events have been used in particular is because they are among the simplest events in which to observe the causal relationship between two objects.

In studies that use these sequences, infants typically are habituated to and tested on either a direct-launching, a delayed-launching, launching-without-collision, or a delayed-plus-no-collision event (Figure 1). In the direct-launching event, one object travels across a stage, ostensibly makes contact with a second, stationary object located mid-screen, at which point the first object stops moving and the second object begins to move. The delayed-launching event is similar to the direct-launching event except the second objects moves after a brief delay upon contact from the first object. The launching-without-collision event is also similar to the direct-launching event except that a small spatial gap is inserted between when the first object stops moving and when the second object begins to move. Finally, in the delayed-plus-no-collision event, the second object begins to move after a brief delay and in the presence of a spatial gap.

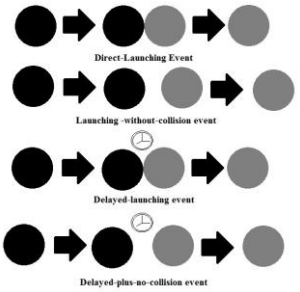


Figure 1: Launching events

Research that has employed launching-event sequences has established that causal perception emerges in development between 4½ and 10 months of age (e.g., Cohen et al., 1998; Rakison & Krogh, 2012). In a classic study on infants' ability to perceive causality in launching-event sequences, for example, Leslie and Keeble (1987) found that 6½-month-old infants who were habituated to a direct-launching sequence—in which a red cube caused a green cube to move through contact—dishabituated to the reversal of the event—in which the green cube now caused the red cube to move. Subsequent work by Cohen and Amsel (1998) showed that causal perception undergoes a developmental transition, whereby 4- and 5½-month-olds responded to the continuity of motion and spatiotemporal relations between the objects in the events, whereas 6¼-month-olds responded on the basis of causality. In a separate study, Schlottman and Surian (1999) found that 9-month-olds will perceive causality in launching events that incorporate a gap only when the objects involved in those events are imbued with animacy cues. Finally, Oakes and Cohen (1990) found that 10-month-olds, but not 6½-month-olds, responded on the basis of causality when realistic stimuli were used instead of simple geometric figures, and Rakison and Krogh (2012) found that 4½-month-olds showed evidence of causal perception when provided with real-world causal-action experience using Velcro-covered

mittens. Taken together, this research suggests that causal perception emerges between 4½ and 10 months of age.

The emergence of causal reasoning

In contrast to causal perception, the ability to reason about causal events emerges later in development. Although developmental researchers have used a variety of tasks to study causal reasoning (for a review see Bullock, Gelman, & Baillargeon, 1982), we restrict our focus to research that has used the blicket-detector design. This is because (1) the blicket-detector task has been used most frequently to study causal reasoning in young children, and (2) an ultimate aim of ours is to examine to what extent the abilities to perceive causality in launching-event sequences and distinguish blickets from non-blickets in a blicket-detector task are underpinned by the same or different mechanisms (e.g., Gopnik et al., 2004; McClelland & Thompson, 2007).

In a typical blicket-detector study, children are introduced to a machine called the "blicket detector" and told that the machine lights up and plays music only when objects labeled blickets are placed on its surface. Children are then asked to determine which objects are blickets and which are not and to intervene to make the machine go. Research that has used this design has shown that 3- to 5-year-olds can make causal inferences with blicket-like objects that span the biological and psychological domains (Schulz & Gopnik, 2004) and that 18- to 30-month-olds can use higher-order relations between objects to make causal inferences (Walker & Gopnik, 2014). Of these findings, perhaps the most relevant from the perspective of the present experiments is the finding that children 2 years and older can use screening-off and backward-blocking reasoning (hereafter BB) to make inferences and generate interventions about the causal status of blicket objects (e.g., Gopnik et al., 2001; Sobel et al., 2004). For example, Gopnik et al. (2001) showed that when 2-, 3-, and 4-year-olds were shown an indirect screening-off (hereafter IS) event—in which together two objects, objects A and B, caused the detector to activate (i.e., AB+) and then an event in which object A alone fails to activate the detector (i.e., A-)—they categorized only object B as the cause. This ability to use IS reasoning—which children ostensibly share with adults—is a hallmark of causal reasoning that enables human learners to distinguish objects associated with an effect from those that produce an effect.

In addition, previous research showed that children use BB reasoning and base rates to reason about blickets. For example, Sobel et al. (2004) found that 4-year-old children who are shown a BB event—in which together objects A and B produce the effect (i.e., AB+) and then an event in which object A produces the effect (i.e., A+)—are less likely to categorize object B as a blicket compared to same-age children who are shown the IS test event if blickets are rare than common. Together, the BB and IS findings are important abilities because it has been suggested that contemporary associative models such as the Rescorla-Wagner model (henceforth, the RW model) fail naturally to account for base rates and why object B is not treated equivalently across the BB and IS conditions. Indeed, in terms of the RW model (Rescorla & Wagner, 1972), object B should be treated equivalently across both the IS and the BB trials because the associative strength between B and the activation of the detector is the same in both cases. Given the failure of contemporary associative models to account for the BB finding, some researchers have proposed that causal learning is underpinned by a Bayesian-inference mechanism (discussed below). Putting this debate aside, the research on causal reasoning suggests that it emerges between 18 months (cf., Sobel & Kirkham, 2006) and 4 years of age.

Development gap and theoretical debates

Despite extensive work on causal perception and causal reasoning, little is known about the relation between these two abilities. For instance, it is unclear whether causal perception and causal reasoning are underpinned by the same or different mechanisms. That is, is causal perception underpinned by one mechanism and causal reasoning, another? Or, is causal perception and causal reasoning underpinned by the same mechanism? According to proponents of the domain-general view of causal learning—where the same all-purpose mechanisms govern learning—early causal perception and later causal reasoning are (1) the emergent consequences of continually enriching perceptual and cognitive systems and (2) are abilities that are underpinned by an associative learning mechanism. This position garners support from behavioral and computational research that showed that infants, young children, and computational (PDP) models use the correlational structure and the predictive statistics of causal events to process and encode their causal relations and that this ability develops over time (e.g., Cohen et al., 1998; McClelland & Thompson, 2007).

In contrast, according to proponents of the domain-specific position—where specific mechanisms process specific kinds of inputs—humans possess specialized modules or mechanisms that are designed specifically to process causal events (Leslie, 1995; Gopnik et al., 2001). For example, within the domain-specific position, some have posited that humans use a simple form of Bayes' rule to reason about the conditional probabilities in causal events and that this ability may be present from birth or shortly thereafter (e.g., Gopnik & Wellman, 2012). This argument is ostensibly supported by research by Sobel and Kirkham (2006, 2007) that showed (1) that infants 5 months of age and older use IS and BB reasoning in a modified habituation version of the blicket detector study, and (2) by research by Griffiths et al. (2011) that showed that adults engage in BB reasoning to reason about super pencils. However, these findings should be considered cautiously because (1) the evidence was mixed about whether infants in Sobel and Kirkham (2006, 2007) processed the events associatively or based on BB and IS reasoning, (2) the habituation task itself failed to preserve the conditional probabilities of previous

blicket studies, and (3) BB reasoning was observed only in one condition of four in Griffiths et al. (2011).

Based on the above limitations, a primary goal of the current set of experiments was to examine causal reasoning in infants and adults and examine to what extent 6-month-olds (Experiment 1) and adults (Experiment 2) use BB and IS reasoning to process causal events. We chose 6-month-olds because this is the age at which they begin to process launching-events causally. Experiment 1 used a novel task design—where the conditional probabilities between this and previous blicket-detector studies were identical—in which infants were habituated to an AB+ A+ event and then tested on A+, B+, A-, B- events (Figure 2). This design choice was important in two ways: (1) it bore closer resemblance to, and allowed direct comparisons with, the conditions presented in the blicket detector studies with children and (2) enabled us to test whether infants were processing the events associatively—in line with predictions made by the RW model—or according to BB reasoning (Figure 2). These predictions derive from previous research that examines infants' use of Bayesian inference (Sobel & Kirkham, 2006, 2007). Note that although the RW model has been ruled out as an informative model of causal reasoning in adults, it is still possible that it can predict the performance of young infants in a habituation task in which there are multiple trials. The specific aims of Experiment 2 were (1) to replicate with adults the conditions presented to children in previous blicket studies and (2) to examine to what extent adults engage in BB and IS reasoning. An additional aim of Experiment 2 was to explore whether adults process the causal events along a causal gradient (see the Conclusion section) in which some adults might process the events associatively, whereas others might process the events according to BB reasoning. An important strength of Experiment 2 was that the design enabled us to assess to what extent adults engage in BB reasoning by comparing pre- and post-ratings of B. With the exception of one study (Griffiths et al., 2011), most compare children's ratings of B in the BB and IS conditions; and the one study that conducted pre- and post-comparisons, a drop in the rating of B was observed in only 1 of 4 conditions.

Experiment 1

Methods

Subjects. Nineteen 6-month-old (*M* = 6 months; range: 5 months 15 days to 6 months 24 days) infants participated in the experiment.

Stimuli and Design. The habituation and test stimuli were computer-animated events that were presented on a computer-generated stage (Figure 2). In each of the two habituation movies, the red and blue circles entered the stage from the right and left (counterbalanced) and moved horizontally across the stage until they abutted a square that was located mid-screen at which point a sun appeared from the square. The second habituation movie was identical to the first except that only one of the two objects (object A) moved horizontally across the stage. Following the habituation phase, infants were shown 4 test events (Figure 3).

Procedure. Each infant sat on their caretaker's lap facing the television monitor. The parent was instructed to abstain from any form of communication with the infant and to remain neutral to avoid biasing the infants' natural response to the habituation and test events. The caregiver was also naïve to the hypotheses and predictions of the experiments to eliminate the chance that the caregiver could reliably influence the infant's looking behavior during the experiment.

	Associative (RW model)	Bayes (BB reasoning)
A-	Longer looking	Longer looking
B+	Equal looking	Longer looking
B-	Longer looking	Equal looking

Figure 2: Looking-time predictions to the A-, B+, and B- test events relative to the A+ habituation event.

During the habituation phase of the experiment, infants were presented with the two habituation events, as outlined in the previous section. In other words, infants were presented with the AB+ event, in which objects A and B caused the sun to appear from the box, and the A+ event, in which object A singly caused the sun to appear from the box. Whether the object (A or B) entered the stage from the right or from the left was counterbalanced across infants.

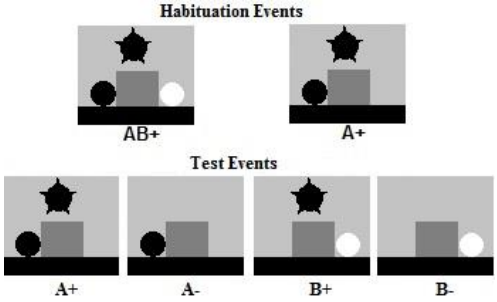


Figure 3: The habituation and test movies

When the habituation criterion was reached or when 16 trials had been presented, the four-trial test phase began. Infants were excluded from the final analyses if they did not reach the criterion within 16 trials. Two infants were excluded for this reason. The test events (i.e., A+, A-, B+, B-) were presented using a Latin square to ensure that event presentation was counterbalanced.

Results

The first analyses compared $\log_{10}$ looking times to the last three habituation trials to $\log_{10}$ looking times to the familiar A+ test event. The rationale for this analysis was to examine whether infants had, in fact, habituated to the familiar (A+) habitation event. Infants' fixation times were entered into a one-way between-subjects ANOVA, which revealed that infants' looking times to the familiar (A+) test event (*M* =

0.81 , *SD* = 0.39) did not reliably differ from their looking times to the last three habituation events (*M* = 0.87, *SD* = 0.06), $F(1, 36) = 0.35$, $p = 0.56$. This analysis revealed that infants had indeed habituated to the familiar event (A+).

The second analysis examined whether there was an effect of sex on looking time to the four test events. In particular, We examined whether male or female infants differed in the amount of time they looked at the four test events. Thus, looking times were entered into a 2 (Sex: male vs. female) x 4 (Test type: A+, A-, B+, B-) mixed-design ANOVA, with sex as the between-subjects variable and test type as the within-subjects variable. Neither the main effects for sex, $F(1, 17) = 0.14$, $p = 0.71$, or test type, $F(3, 51) = 0.71$, $p = 0.55$, nor the interaction (sex × test type) was significant, $F(3, 51) = 1.55$, $p = 0.21$. The data for sex was subsequently collapsed.

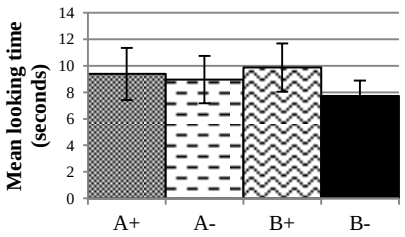


Figure 4: Results from Experiment 1: Infants' mean looking times as a function of test trial.

The primary analysis examined whether infants differed in their looking times to the four test events. The rationale for this analysis was that if infants showed differential looking to the four test events, then it would be possible to examine whether infants were processing the events associatively or according to Bayesian inference (as outlined in the Introduction). A repeated measures ANOVA with test type (A+, A-, B+, B-) as the within-subjects factor revealed that infants did not differ reliably in their looking times to the four test events, $F(3, 54) = 0.28$, $p = 0.84$. This analysis suggests that infants were processing the events neither associatively nor in a way that is consistent with previous BB findings.

Despite the fact that the main analysis failed to yield a significant finding, it is possible that infants processed the events based on associative learning or Bayesian inference. In particular, by comparing looking times to the A+ and the A- test trials for infants who received these two trials first and looking times to the B+ and B- test trials for infants who received these trials first, it is possible to determine whether infants processed the events based on Bayesian inference or associative learning. Thus, separate paired-sample t-tests were used to compare infants' looking times to the A+ and A- test trials and looking times to the B+ and B- test trials. The analysis that compared looking times to the A+ and A- test trials revealed that infants looked equally long at the A+ test trial (M = 15.64 s, SD = 11.84 s) and A- test trial (M = 12.72 s, SD = 5.81 s), $t(4) = 0.44$, $p = 0.68$. The analysis that compared looking time to the B+ and B- test trials revealed that infants looked equally long at the B+ test trial (M = 9.58 s, SD = 5.45 s) and B- test trial (M = 11.17 s, SD = 7.22 s), $t(5) = 0.39$, p = 0.72. Considered together, this set of analyses reveals that infants were not processing the events based on associative learning or Bayesian inference.

Discussion

The null results from Experiment 1 indicated that it is unclear whether infants processed the events associatively or according to previous BB findings; that is, the results neither showed that infants looked longer at the A- and B- test events relative to the A+ and B+ test events as would be predicted from an associative (PDP) model nor did they show that infants looked longer at the A- relative to the remaining three test events as would be predicted from a Bayesian perspective. One possible explanation for this finding is that the events may have been too complex for infants to process compared to those in previous causal-perception studies. Indeed, standard Michottian launching events are typically simpler and involve far less dynamic cues—which can be difficult for 6-month-olds to process, much less according to BB reasoning—than the events used in the present experiment.

Nonetheless, the results from Experiment 1 suggest tentatively that 6-month-olds in the current design cannot solve (or reason) about BB events in associative or Bayesian way despite previous research by Sobel and Kirkham (2006, 2007) that suggests that Bayesian inference is present by at least 5 months (c.f., Shultz, 2007). Despite the fact that the results from Experiment 1 provided inconclusive evidence about whether 6-month-olds used BB reasoning to process the causal events, it is still possible that in the context of a standard blicket-detector study, adults will use BB and IS reasoning to process causal events. This was the goal of Experiment 2.

Experiment 2

Methods

Subjects. Sixty college students were recruited from Carnegie Mellon University to participate in Experiment 2.

Stimuli and Design. A device similar to the blicket detector in previous studies was used in this study. The device was 5" x 7" x 3" and was made of wood (painted black) with a white lucite top. The machine operated via a remote control that was attached to the end of an electric wire that was attached to the side of the machine. When the button was depressed and the object predetermined to be the "blicket" was placed on the surface of the detector, the music and the lights began to play and flash. The button was not pressed, and hence the music and lights did not play or flash, when the object predetermined not to be the blicket was placed on the detector's surface.

Eight cube and cylinder objects, each of different color and approximately 1" in diameter, were used. No objects of

the same shape were used to demonstrate the effect of blicketness and the object that was designated as the blicket was counterbalanced across subjects. Two unrelated objects were used in the pretest phase of the experiment.

Procedure. Participants were tested in a quiet testing room. Participants were introduced to the machine and told that it was called a blicket detector that activated only when blickets were placed on it at the beginning of the experiment. They were instructed also that their job was to determine which objects were blickets and which were not. Participants were then given two pretest trials to ensure that they understood the purpose of the experiment. In one of the pretest trials, one of the two unrelated objects activated the machine and was labeled the blicket, and the other of the two objects (both randomly determined) did not activate the machine.

Following this pretest phase, participants were then given four test trials (counterbalanced). These test trials paralleled those in previous blicket-detector studies and included the one-cause (1C), two-cause (2C), indirect screening-off (IS), and backward-blocking (BB) trials. The 1C and 2C trials served as the controls to ensure that participants understood the test events. Participants were instructed to rate on a scale of 0 (definitely not) – 100 (definitely is) that each object in the pair was a blicket both before and after a trial. In the 1C trial, object A activated the machine when placed alone on the detector but object B did not when placed on the detector alone. Both objects were then placed on the machine twice, which activated. In the 2C trial, object A activated the detector 3 of the 3 times it was placed alone on the detector, whereas object B activated the machine 2 of the 3 times it was placed alone on the detector. In the BB trial, both objects A and B were placed on the detector twice, which activated both times. Object A was then placed on the detector by itself and the detector activated. The IS trial was identical to the backward-blocking trial except that object A did not activate the detector.

Results

To analyze whether adults' ratings of objects A and B differed for the BB, IS, 1C and 2C test trials, a repeated-measures ANOVA with ratings of objects A and B for each test trial as the within-subjects factor revealed that participants' ratings of both objects differed between each test trial, $F(7, 413) = 69.58$, $p < .001$. To examine whether the pre-ratings of A and B differed from the post-ratings of A and B for each test trial, paired samples *t* tests with Bonferroni corrections were conducted

The first paired-samples t-test for the 1C test trial revealed that the post-rating of A ($M = 94.92$, $SD = 15.58$) increased significantly from the pre-rating of A ($M = 49.62$, $SD = 16.96$), $t(59) = -15.34$, $p < .006$. In contrast, the post-rating of B ($M = 10.33$, $SD = 21.68$) decreased significantly from the pre-rating of B ($M = 50.08$, $SD = 15.66$), $t(59) = 11.03$, $p < .006$. These results replicate the 1C condition in previous blicket-detector studies.

The second paired-samples t-test for the 2C test trial revealed that the post-rating of A ($M = 94.75$, $SD = 11.33$) increased significantly from the pre-rating of A ($M = 50.60$, $SD = 18.25$), $t(59) = -15.78$, $p < .006$. Likewise, post-rating of B ($M = 79.10$, $SD = 19.42$ significantly increased from the pre-rating of B ($M = 48.75$, $SD = 16.46$), $t(59) = -10.79$, $p < .006$. These results also replicate previous 2C results.

The third paired-samples t-test for the IS test trial revealed that participants' post-rating of A ($M = 12.03$, $SD = 27.06$) decreased significantly from their pre-rating of A ($M = 55.77$, $SD = 18.39$), $t(59) = 10.20$, $p < .006$. In contrast, the post-rating of B ($M = 89.48$, $SD = 16.60$) increased significantly from the pre-rating of B ($M = 49.83$, $SD = 17.32$), $t(59) = -12.12$, $p < .006$. This result suggests that adults used IS reasoning.

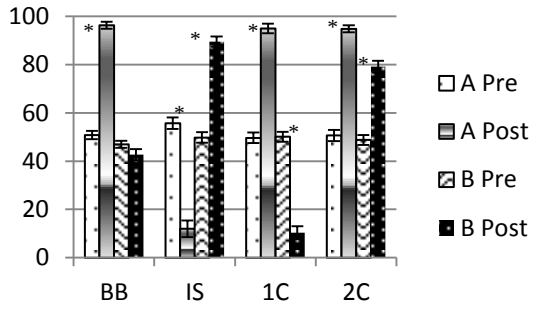


Figure 5: Results of Experiment 2: Participants pre- and post-ratings of objects A and B as a function of condition. Asterisks indicate significance at $p < .05$ between A pre and A post and B pre and B post pairs.

The final paired-samples t-test for the backward-blocking test trial revealed that participants' post-rating of A ($M = 96.25$, $SD = 11.07$) increased significantly from their pre-ratings of A ($M = 50.83$, $SD = 13.38$), $t(59) = -19.73$, $p < .006$. The pre-rating of B ($M = 46.92$, $SD = 11.28$), however, did not differ from the post-rating of B ($M = 42.67$, $SD = 18.14$), $t(59) = 1.84$, $p = .07$, demonstrating the absence of BB.

Given the absence of the BB effect, we conducted an additional analysis to examine whether the effect was moderated by a tendency for some participants to use BB reasoning and others to use associative reasoning. A repeated-measures ANOVA revealed that pre-ratings of B that varied between 0 and 50 ($M = 42$, $SD = 1.87$) did not differ reliably from post-ratings of B ($M = 46.83$, $SD = 3.25$), whereas pre-ratings of B that varied between 50 and 100 ($M = 51.83$, $SD = 1.87$) were significantly higher than post-ratings of B ($M = 38.5$, $SD = 3.25$), $F(1, 58) = 20.52$, $p < .001$. These same participants did not differ in their use of IS reasoning. This provides preliminary (but speculative) evidence that participants processed the causal events along a causal gradient; that is, not all participants showed BB reasoning. This is an important finding because, in contrast to our finding, previous research suggests that BB reasoning is a fundamental human ability.

Discussion

The results from Experiment 2 indicated that participants' post-ratings of objects A and B differed reliably from their pre-ratings of objects A and B in all test trials except for the BB test trial where no significant difference was observed. This BB result is particularly interesting because it contravenes the prediction about BB made by the Bayesian perspective; that is, if participants engaged in BB reasoning, then a significant decrease in the rating of B between the pre- and post-rating phases should have been observed according to this view. However, associative models (e.g., the RW model) make no such prediction about a drop in the rating of B as a blicket. The results of Experiment 2 seemed to support both perspectives in which some adults engaged in BB reasoning whereas others engaged in associative reasoning. In general, the results from Experiment 2 replicated those with children in previous blicket-detector studies.

Conclusions

The null results from Experiment 1 indicated that the 6-month-olds processed causal events neither associatively or in terms of BB. This result suggests that BB reasoning and, to a lesser extent the ability to use a simple form of Bayes' rule as has been assumed, may not be present from birth or shortly thereafter. Instead, this ability may develop as infants learn about causal events in the world. For example, infants 6 months of age and younger may process the features of the objects or the paths that the objects take independently, whereas older infants may process the relations between the objects and begin to parse the events in an associative- or Bayesian-like way. Ongoing research is testing this hypothesis with older infants.

The results from Experiment 2 replicated three of the four test-trials given to children and showed that adults use IS reasoning to reason about blicket events. For the BB trial, however, adults did not rate B differently between the pre- and post-rating phases. Note that we tested the same number of adults in this study as in Griffiths et al. (2011). Nonetheless, there was evidence that participants whose pre-ratings of B varied between 0 and 50 appeared to process the events associatively, whereas participants whose pre-ratings of B varied between 50 and 100 appeared to use BB reasoning. This effect also appeared to be restricted to the BB trial, which suggests that the different modes of reasoning apply only to BB trials. This result suggests that adults may process causal events along a gradient; that is, some adults may process the events associatively, others may process the events according to Bayesian inference, and still others may use a combination of both to process causal events.

In summary, the results of the present experiments reveal that (1) 6-month-olds processed BB events neither associatively nor in terms of Bayesian inference and (2) that adults may process causal events along a causal gradient rather than in a strict Bayesian or associative way.

Acknowledgments

This research is based upon work supported by the National Science Foundation Graduate Fellowship under Grant No. DGE-1252522 to Deon T. Benton.

References

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. *The developmental psychology of time*, 209-254.

Cohen, L. B., & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant behavior and development*, *21*(4), 713-731.

Cohen, L. B., Amsel, G., Redford, M. A., & Casasola, M. (1998). The development of infant causal perception. *Perceptual development: Visual, auditory, and speech perception in infancy*, 167-209.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental psychology*, *37*(5), 620.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, *111*(1), 3.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407-1455.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?. *Cognition*, *25*(3), 265-288.

McClelland, J. L., & Thompson, R. M. (2007). Using domain-general principles to explain children's causal reasoning abilities. *Developmental Science*, *10*(3), 333-356.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event.*Cognitive Development*, *5*(2), 193-207.

Rakison, D. H., & Krogh, L. (2012). Does causal action facilitate causal perception in infants younger than 6 months of age?. *Developmental science*, *15*(1), 43-53.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In*Classical conditioning: Current research and theory*.

Schlottmann, A., & Surian, L. (1999). Do 9-month-olds perceive causation-at-a-distance?. *Perception-London-*, *28*, 1105-1114.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains.*Developmental psychology*, *40*(2), 162.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Developmental psychology*,*42*(6), 1103.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive science*, *28*(3), 303-333.

An Analysis of Frame Semantics of Continuous Processes

McFate, Clifton (c-mcfate@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL, 60208, USA

Forbus, Kenneth (forbus@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
2133 Sheridan Road, Evanston, IL, 60208, USA

Abstract

Qualitative Process theory provides a formal representation for human-like models of continuous processes. Prior research mapped qualitative process elements onto English language constructions, but did not connect the representations to existing frame semantic resources. Here we identify and classify QP language constituents through their instantiation in FrameNet frames to provide a unified semantics for linguistic and non-linguistic representations of processes. We demonstrate that all core QP relations can map to FN, though larger QP evoking phrasal constructions do exist outside of this mapping. We conclude with a corpus analysis showing that these frames occur in natural text involving a variety of continuous processes.

Keywords: Frame Semantics; Qualitative Reasoning

Introduction & Background

Daily life requires interacting with, and reasoning about, continuous processes. They can be as common as coffee flowing into your mug or as abstract as economic growth. Despite their mathematical complexity, people rapidly generate predictions based on mental models of these situations. Forbus' (1984) qualitative process (QP) theory provides a formal language for representing mental models of continuous systems. QP theory is domain general and there is evidence supporting qualitative models as a mental representation. For example, Friedman *et al* (2011) use qualitative representations to simulate conceptual change, and demonstrate intermediate states of knowledge found in middle-school students (Sherin *et al*, 2012).

An important issue is bridging the gap between these cognitive models of change and purely linguistic models. Doing so illuminates the semantics of continuous processes and lays groundwork for systems that learn from and reason with natural language (McFate, Forbus, & Hinrichs, 2014).

Kuehne (2004) developed QP frames, a frame semantic representation inspired by Fillmore *et al*'s (2001) FrameNet. This approach was expanded by McFate *et al* (2014). While useful, both approaches had limited coverage and did not connect QP frames to existing frame semantic resources. We bridge that gap by providing a QP mapping of specific process types in FrameNet as well as their constraints, e.g. *limit points*, which mark the boundaries of qualitative states. We evaluate our mapping on science texts, but expect our approach to be domain general.

Qualitative Process Theory

In QP theory, changes within a continuous system are always the result of processes. Causality starts with direct influences, which express the relationship between the rate of a process and the constrained quantity. A direct influence provides partial knowledge of a differential equation, where the set of direct influences must be combined to determine derivatives. Indirect influences propagate the direct effects of processes through the rest of a system, by providing partial information about instantaneous (e.g. algebraic) causal relationships. Consider water flow into a tub. A direct influence holds between the flow rate and the amount of water in the tub. An indirect influence holds between the amount of water and the water level. Processes are represented by *model fragments* which describe the participating entities, the conditions under which instances of it are active, and what consequences hold when active. The conditions typically include ordinal relationships involving a quantity and one of its limit points.

A system that acquires model fragments from text could reason about real-world scenarios, predicting, for example, that our tub of water may overflow. However, the incremental nature of language makes extracting complete models from text difficult. We turn to frame semantics to provide flexibility, and view this research as a step towards systems that learn by reading about the continuous world.

Frame Semantics

Semantic frames are conceptual schemas that relate lexical items in a sentence to their role in a semantic description (Fillmore, Wooters, & Baker, 2001). Fillmore *et al's* (2001) FrameNet is a frame semantics for English. FrameNet frames are evoked by a frame-bearing lexical unit (LU). The dependent structures in the sentence form arguments to that frame's frame elements (FEs). For example, the `Motion` frame includes frame elements for the `Source`, `Goal`, and `Theme`. It is instantiated in a specific construction by a frame evoking LU such as the word *fly* in "The bird *flew* to Florida". Here, the NP subject fills the FE of `Theme` and the prepositional phrase fills the FE of `Goal`. The specific grammatical instantiation of these roles is called a *valence pattern* for that lexical unit.

Baker, Fillmore & Cronin (2003) present inter-frame

relationships including inheritance and subframe. A frame that inherits from a parent has a corresponding frame element for each element of the parent and can introduce others. The subframe relation allows frames to act as ordered arguments to another frame, forming a kind of script. These relations create a frame-lattice with top-level frames like `Event` and specialized inheritors like `Motion`.

QP Language

Kuehne (2004) recast QP theory in a frame-semantic representation to better handle compositionality in language. Quantity frames fill the argument slots of direct and indirect influence frames. The influences participate in quantity transfer descriptions and process frames which capture the results and activation conditions of a process. We adopt a representation closer to FrameNet's, with influences and quantities related through shared lexical units (see Figure 1) This change from previous work (McFate *et al*, 2014) facilitates extracting partial information when grammatical constructions leave out required roles (e.g. an agent).

Kuehne (2004) identified several patterns that instantiated QP frames, and used them to extract QP frames from text, using neo-Davidsonian lexical representations from the Cyc KB[1]. McFate *et al* (2014) extended this approach and introduced narrative functions (Tomai, 2009) to guide disambiguation. This system was limited by the coverage of Cyc's semantic templates, and it also became evident that a finer-grained set of distinctions would be useful. Integrating QP frames with FrameNet helps solve both problems by providing valence patterns by frame type. It benefits frame semantics by providing representations for mental models.

Next we provide FrameNet mappings for each QP frame: *direct influence*, *process*, *indirect influence*, *quantity*, and *ordinal*. We also discuss how *limit points* are represented.

Unifying FrameNet with QP Frames

Continuous processes are process verbs and nominalized verbs in English. Since direct influences are only allowed within processes, we start with them. The direct influence (DI) frame has `constrained` and `constrainer` entities, constrained and constrainer quantities and quantity types, as well as a `sign`. The constrainer entity is the process itself, and the `sign` indicates the direction of contribution for the rate. We introduce an `agentive causer` FE motivated differences between causative and inchoative constructions, as discussed below.

FrameNet has many frames that instantiate continuous processes and DIs. A straightforward mapping for these frames aligns QP elements to potential FN elements. Figure 1 provides an example of `Fluidic_motion`.

We now walk down part of the FN hierarchy for `Event` verbs, and examine the additional valence patterns at each layer. We begin with a broad distinction between causative (with an agent) and inchoative constructions. For example,

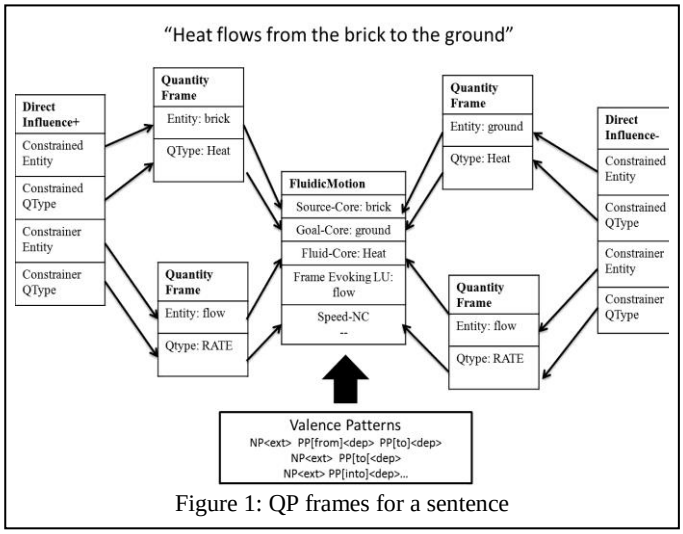


Figure 1: QP frames for a sentence

spill can be causative whereas *flow* is only inchoative e.g.

"The boy *spilled* the juice."

*"The boy *flowed* the juice."

"The juice *flowed* from the box."

While our analysis covers many physical processes, it is not intended to be exhaustive. Instead, it is exemplary of the productivity of mapping these resources.

Inchoative

There is no root frame for all events that don't require an agent. We examine an inheritor of `Event`, `Motion`. We also examine non-agent-requiring state-change frames.

Motion

Basic motion in is captured with the `Motion` frame, which has many inheritors. We use `Fluidic_Motion` as an example. Tables in the following sections illustrate valence patterns for target frames. For example, Table 1 summarizes `Fluidic_Motion` patterns and their alignment to DI frames.

Table 1

Valence Pattern		Example	FrameNet Frames
DI+ NP-V-PP		Water **flows** to the basin.	Fluidic Motion: *Core:* Goal, Source, Path, Fluid, Area *Non-core:* Speed
Cnd-Qtype	Cnd-Ent		
NP<ext>	PP[to]<dep>		
	PP[into]<dep>		
	PP[in]<dep>		
DI- : NP-V-PP		Water **flows** from the basin	
Cnd-QType	Cnd-Ent		
NP<ext>	PP[from]<dep>		
	PP[out]<dep>		

The left-hand column illustrates valence patterns that instantiate the QP elements (here the constrained quantity type and entity). The patterns are presented with FrameNet grammatical functions. The most common are `ext`, `dep`, and `obj` which indicate an external argument (subject), verb dependent, or object. The top left cell of Table 1 says that a positive DI can appear with an NP subject and PP dependent instantiating the quantity type and entity.

For motion, positive direct influences correspond to frame instantiations that include the `Goal` while negative direct

[1] http://www.cyc.com/platform/researchcyc, v4.0

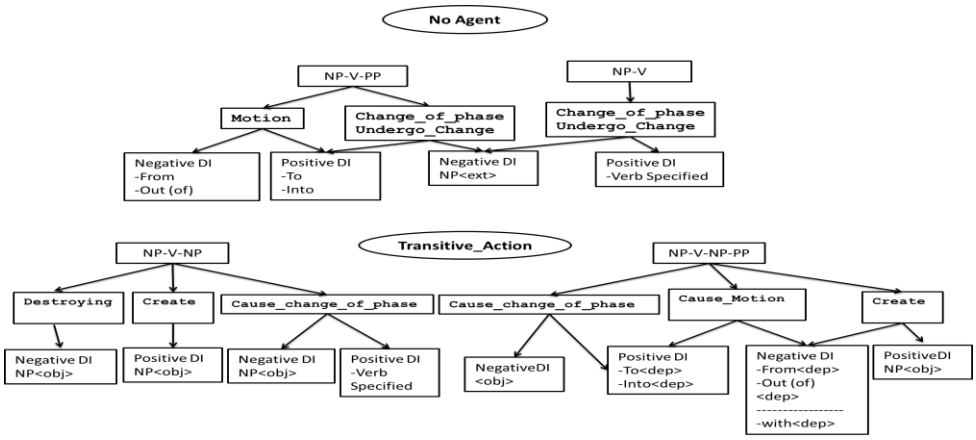


Figure 2: Agentive and Non-Agent Mapping

influences correspond to the `Source` FE. The constrained quantity corresponds to the `Theme` in the `Motion` frame and its corresponding element in the inheritors (ie. `Fluid`).

These can also appear in a single ditransitive prepositional form e.g. "X flows from Y to Z". Each of the sub-frames that evoke this pattern evokes a DI, but not all possible valence patterns for motion do. For example, `Motion` can have `Path` or `Area` FEs that can occur in isolation e.g.

"Joe ran through the long tunnel"

"Joe ran around the room"

These valence patterns instantiate a motion process, but would require additional sentences to specify a DI.

State_Change & Conversion

State change processes can be represented as a pair of direct influences representing the increase in substance at one phase and decrease in substance at another (e.g. steam and water). This is an important representational decision in because it separates state-changes from their preconditions.

Thus the same NP-V-PP structure in state-change can introduce two influences based on the semantics of the verb (Table 2). FrameNet captures inchoative state-change with the `Change_of_phase`. The core element, `Undergoer`, is the changing entity. This differs from prior analyses of flow and motion, because they are referring to an event without referring to the details of what happened during it (e.g. "The water boils away").

Table 2

Valence Pattern		Example	FrameNet Frames
DI+ : Resultative-State		The water **froze** to ice.	Change_of_phase: *Core*: Undergoer *Non-core*: Result, Speed, Place
Cnd-QType	Cnd-Ent		
PP[to]<ext>	NP<ext>		
DI- : Resultative-State			
Cnd-QType	Cnd-Ent		
NP<ext>	NP<ext>		
DI- : NP-V		The water **froze**.	
Cnd-QType	Cnd-Ent		
DEN	NP<ext>		

This is clearest in the intransitive where the resulting state, *ice*, is indicated only by the verb, *freeze*. In contrast, intransitive motion may indicate a process but does not specify the predicates (from/to) needed for a full DI. Our conversion interpretation is supported by multiple PP attachments as in "The water froze from liquid to ice". Thus, our mapping extends the `Change_of_phase` frame to include required initial and final states. This puts our interpretation closer to the `Undergo_Change` and `Cause_Change` frames which include verbs such as *change*, *convert*, and *turn*.

Causative

Causative verbs appear in additional constructions. Many verbs appear in both the causative and inchoative.

We begin with `Transitive_Action` which includes causative frames such as `Cause_Motion` and `cause_Change_Of_Phase`. Each of the verbs under `Transitive_Action` inherits its required `Agent` FE.

`Cause_motion` verbs, with their mandatory `Agent` or `Cause`, appear in additional patterns. For example, the subject of the sentence can be the entity: "The plant creates energy". Here, the plant does not necessarily gain the energy. It is not the constrained entity of the DI, but its role as agent is vital to the model. This motivates adding an `agentiveCauser` FE which allows for these entity-less causal constructions. These DIs can also appear with prepositional attachments that specify the entity as well as in a passive. (Table 3).

Table 3

Valence Pattern		Example	FrameNet Frames
DI +/- NP V NP PP		New York **pumps** water from the subway.	Cause_fluidic_motion: *Core:* Goal, Source, Path, Fluid, Area, Agent/Cause
Cnd-QType	Cnd-Ent		
NP<obj>	PP[from]<dep>		
	PP[to]<dep>		
	…		
DI+: Passive		Water is spilled from the bucket.	
Cnd-QType	Cnd-Ent		
NP<obj>	PP<from>		

Again, differences in semantic meanings across verbs results in different interpretations of the same construction. `Create` for example, with the preposition *for* creates a benefactive where the entity becomes the PP dependent

rather than the NP subject. This same construction leaves the entity underspecified when used with `cause_Motion` or even the `Destroying` Frame. The mappings discussed so far are represented graphically in Figure 2.

Constrainers

So far we have ignored rates. Kuehne (2004) found that English frequently left rates implicit. Non-causative frames above contain the non-core `Speed` FE which includes rate constructions (e.g. at a rate of X per Y). For process frames lacking the `Speed` element, we implicitly encode the rate as the `constrainer quantityType`.

Indirect Influences

The indirect influence frame has a `consequent` and `antecedent quantity`, `quantityType`, and `entity`, as well as a `sign`. It maps to the `Contingency` and `Objective_Influence` frames**,** both of which have a frame element corresponding to a causal (antecedent) and dependent (consequent) entity. These LUs can be modified with an adverb such as *negatively* to reverse directionality.

`Contingency` includes verbs like *depend* which use a prepositional attachment to indicate the independent variable. `Objective_Influence` appears in a transitive and passive construction. A similar mapping holds to the `Actor` and `Affected` of the `Causation` frame, as in: "Deforestation causes less carbon to be taken out of the atmosphere." This sentence indicates a constraint on carbon absorption tied to amount of deforestation. Similar patterns tie the rates of two processes together: "Heating water causes it to boil." Disambiguating whether the rate or governed quantity is the antecedent frequently requires domain knowledge.

Several phrasal constructions can indicate covariance and thus evoke an indirect influence (Table 4). Construction 1 is the comparative correlative (Culicover & Jackendoff, 1999). Here the first and second phrases map to the antecedent and consequent roles. The sign is given by the directionality of the adjectives. A similar mapping exists for correlative conjunction constructions where, the conjunct phrases both involve change verbs (e.g. increase/decrease).

Table 4

1. Comparative Correlative	The higher the water level in the bucket, the greater the water pressure in the bucket.
2. Correlative Conjunction	Both temperature and pressure increase.
3. As X, Y	As the temperature in the boiler increases, the pressure in the boiler increases.
4. Changes with Y	The pressure of the boiler changes with the temperature.

Quantity Frames

Quantity frames have the required FEs `entity`, `quantity`, and `quantityType`. `QuantityType` relates the frame to the Cyc collection denoting the quantity of the sentence. An example would be *heat* or *pressure*. The `entity` is the object that the quantity attribute pertains to. An optional `quantityValue` and `quantityUnit` FE relate the quantity to numerical data. The optional `signOfDerivative` can indicate a direction of change.

Quantity evoking units include words such as "*heat*" or "*volume*". These fit best in the `Measurable_attribute` frame, though it has thus far only been applied to gradable adjectives (e.g. the hot brick). Furthermore, in FrameNet the frame explicitly evokes deviation from a norm.

Quantity frames often rely on possession to indicate the entity. These instantiations map to the `Possession` frame and include possessive verbs (e.g. *have*) and genitive constructions, though only a subset of `Possession` verbs is suitable. A counterexample is "The brick owns mass."

Containment also links quantities to entities. One can describe the "energy contained in the boiler" or separate out quantities with "the air and water pressure in the container". These constructions fit into the `Containers` and `Containing` frame.

The `QType` can be compositional with the unit. This occurs in constructions with measurement phrases (Table 5).

Table 5

QValue	Qunit	Qtype			
Num.Quant	DEN	PP[of]<dep>		5 **liters** of water	**Measures** *Core*: Count, Entity, Unit

QValue	Qunit	Qtype	Entity		
Num.Quant	N<dep>	DEN	NP<ext>	The wall is 6 feet **tall**	**Dimension** *Core*: Dimension, Measurement, Object
QValue	Qunit	Qtype	Entity		
Num.Quant	N<dep>	DEN	NP<obj>	The 6 foot **tall** wall	

Qunit	Qtype		
DEN	PP[of]<dep>	**Amount** of water	**Quantity** *Core*: Entity, Quantity

The first example consists of a measurement expression modifying an "of" PP. The next two take a measurement expression and adjective phrase and return either a noun-modifier or a predicate (Fillmore *et al*, 2012). Additionally, *'amount of'* can be used to explicitly define a quantity.

Fillmore *et al* (2012) also identify a measurement phrase construction specific to rates which consists of a numerator, the definite NP, and a denominator, the indefinite np (e.g. "miles per hour"). Rates can also link to a nominalized process verb (e.g. 'The rate of flow'). These often anaphoric phrases map to the `Rate_quantification` frame.

Ordinals

Inequalities frequently drive processes. An ordinal frame has the elements, `entity1` and `entity2` as well as a `relation` FE that defines the relation (>, <, =, negligible) between them. Entities in an ordinal share a `qtype`.

The most direct way these appear in language is through gradable adjectives in a comparative construction (e.g. 'cooler than' or 'more cool than'). In FrameNet, these fit best into the `Evaluative_comparison` frame, though it

does not include comparative adjectives.

Ordinal frames are also evoked by non-comparative `Measurable_attribute` expressions such as *hot* or *cold*.

There are also several ways to indicate a difference without specifying directionality. These are often referential and fall under the `Similarity` frame (Table 6).

Table 6

<table>
<tr><td>Ent1</td><td>Ent2</td><td>Qtype</td><td>The temperature difference between the objects</td><td rowspan="10">Similarity
Core: Differentiating_fact, Dimension, Entities, Ent1, Ent</td></tr>
<tr><td>PP[between]<dep></td><td>PP[between]<dep></td><td>N<dep></td></tr>
<tr><td>PP[between]<dep></td><td>PP[between]<dep></td><td>PP[in]<dep></td></tr>
<tr><td colspan="3">Compound-NP-V</td><td rowspan="3">The temperatures of A and B differ.</td></tr>
<tr><td>Ent1</td><td>Ent2</td><td>Qtype</td></tr>
<tr><td>PP[of]<dep></td><td>PP[of]<dep></td><td>NP<ext></td></tr>
<tr><td colspan="3">NP-V-PP</td><td rowspan="5">The temperatures differ between the bricks.</td></tr>
<tr><td>Ent1</td><td>Ent2</td><td>Qtype</td></tr>
<tr><td>PP[across]<dep></td><td>PP[across]<dep></td><td rowspan="3">NP<ext></td></tr>
<tr><td>PP[between]<dep></td><td>PP[between]<dep></td></tr>
<tr><td>PP[from]<dep></td><td>PP[from]<dep></td></tr>
</table>

One significant difference between our representation and FrameNet's is that FrameNet provides the `Entities` FE which groups multiple individuals to fill one role. We separate them since they indicate different quantity frames.

Limit Points, Transitions, and Constraints

Limit points are quantities that define a value where a model fragment changes status (e.g. *boiling point*). Frequently they occur as a compound noun consisting of a constrained process and a barrier. They can also include numerical values, and participate in possession and containment. Limit points are also evoked with `Extreme_value` adjectives such as maximum or minimum as well as verbs that signify arrival at a point as in the `Arriving` frame.

Many limit points are left implicit or referred to only as a deviation from the norm as in: "The water gets cold which causes condensation" These limit points can be made explicit with a modifier such as *enough* (`Sufficiency`)

FrameNet's `Process` frame has sub-frames indicating different states. Starting conditions are captured with patterns from the `Process_start` frame which includes verbs such as 'begin'. Similar frames exist for stopping, continuing, pausing, and resuming. When used in conditionals, lexical units evoking sub-frames and limit points define the constraints of a model fragment (Table 7).

Many frames in FrameNet can indicate their own activation conditions. Consider the `Cause_Motion` valence patterns in Table 2. The required `Agent` can be replaced with a non-animate `Cause` and viewed as preconditions:

"A temperature difference drives heat to the brick."

These constructions can also be used to elaborate on previous instantiations of frames that they are causative of.

"A temperature difference drives heat flow."

Table 7

<table>
<tr><td>Condition</td><td>Process</td><td>Once the submarine reaches crush depth, compression begins.</td><td rowspan="5">Process_Start
Core Unxp: Event
Non-Core: Time

Arriving
Core: Goal, Theme</td></tr>
<tr><td>PP[after]<dep></td><td rowspan="4">NP<ext></td></tr>
<tr><td>PP[when]<dep></td></tr>
<tr><td>PP[if]<dep></td></tr>
<tr><td>PP[once]<dep></td></tr>
<tr><td>Condition</td><td>Process</td><td rowspan="5">After reaching 2,070 degrees, the steel begins melting.</td></tr>
<tr><td>PP[after]<dep></td><td>Vping<dep></td></tr>
<tr><td>PP[when]<dep></td><td rowspan="3">VP[to]<dep</td></tr>
<tr><td>PP[if]<dep></td></tr>
<tr><td>PP[once]<dep></td></tr>
</table>

Explicit `Causation` verbs and modifiers can also indicate a process constraint (e.g. 'because the temperatures differ…').

Finally, quantities within a process can be constrained at certain values using correspondence statements such as:

"The force of the spring is zero when the block is at zero."

This temporal correspondence is captured by the frame `Temporal_collocation`. This frame is vast and includes indexical terms such as "today". We constrain our mapping to patterns that relate two events (e.g. the adverb '*when*').

Corpus Analysis

We have demonstrated that the core elements of QP theory each correspond to FrameNet frames. Next we use a corpus analysis to evaluate the frequency of these patterns in natural language descriptions. We predict that descriptions of continuous processes are prevalent in natural language, especially in explanatory texts. Furthermore, if frame-semantics captures core QP representations, then we would expect that a large number of these descriptions conform to our frame-semantic analysis.

Our corpus consisted of grade school science topics from the Simple English Wikipedia: full articles on the water cycle, condensation, and Bernoulli's principle as well as the first 6 sections about the sun and introduction of the global warming article. There were a total of 90 sentences. Each sentence was annotated for any QP frames. For each process evoking LU (e.g. *flow*) we further evaluated its FrameNet entry. An LU did not have to result in a DI to be counted; it could be an elaboration of a process (e.g. "condensation occurs when…"). We counted FN as having the valence pattern if the specific LU in the correct frame had the complete pattern annotated either alone or as a part of a larger pattern with the same core elements. We also evaluated if the frame-type was analyzed in this paper or one of its children. The results are summarized by article in Table 8. Out of all 90 sentences, 56 (62.2%) had QP material. We identified 53 total process evoking lexical units (e.g. *flow*).

These results suggest a substantial number of sentences in science texts convey QP information, consistent with our prediction. Furthermore, we found that 43.4% of process evoking units already had their specific valence pattern annotated in FrameNet. Thus, mapping QP theory, and possibly other non-linguistic representations, to FrameNet

Table 8

Corpus Analysis Results				
	QP Evoking	Process Evoking LUs	Valence in FN	Type Analyzed
Water (15)	9 (.6)	8	6	4
Sun (28)	15 (.54)	20	10	11
Bernoulli(14)	11 (.79)	5	2	3
Global-warming (21)	13 (.62)	9	4	3
Condensation (12)	8 (.67)	11	1	8

can reveal how these models are expressed linguistically.

Finally, 29 of the 53 process evoking units either directly evoked or evoked inheritors of the set of process frames analyzed above (see Figure 2). The remaining verbs evoked: `Giving`, `Receiving`, `Gathering_up`, `Arriving`, `Expansion`, `Emanating`, `Emitting`, `Departing`, `Soaking_up`, `Using_resource`, `Removing`, and `Fire_burning`. Extending to these additional frames is future work.

Like in Kuehne's (2004) analysis, we found that reference to an explicit rate was rare, only occurring in four sentences of the Bernoulli article. Furthermore, we found no compound-noun limit points (e.g. boiling point) but did find constraints based on deviation from the norm (e.g. when it gets cold…). In part this was due to choice of articles, e.g. articles on boiling or phase changes per se do mention them.

Related Work

Our work dovetails nicely with work in semantic role labeling. General frame semantic parsers such as Das *et al*'s (2014) SEMAFOR could provide FrameNet parses to be generalized using our mapping. Furthermore, knowledge of constraints on qualitative models could resolve ambiguities in language processing as in McFate *et al* (2014).

Ovchinnikova *et al* (2010) used a data-driven analysis to cluster and enrich FN frames about medical treatment. A similar approach could be used to extend our mapping.

Finally, while we've focused on lexicographic descriptions, we identified multi-word and phrasal patterns for indirect influences and quantities, and expect more. These are beyond the initial goal of FrameNet. Fillmore, Lee-Goldman and Rhodes (2012) propose a method for extending FN to include such structures.

Conclusions and Future Work

Much remains to be done. Our corpus analysis suggested additional frames that could be mapped to QP theory, and a larger corpus would both illuminate new frames and establish their frequency. We also hope to use this analysis to enrich the coverage of McFate *et al*'s (2014) system by automatically extracting valence patterns from FrameNet. Future work must also examine how pragmatic and narrative constraints influence QP frame instantiation.

Qualitative Process Theory provides a formalism for representing mental models of continuous processes. While preliminary, by linking this formalism to frame semantic resources we enrich the linguistic representations with a higher-order inferential model and provide a resource that facilitates interactive learning of these models in the future.

Acknowledgments

This research was supported by the Intelligent and Autonomous Systems Program of the Office of Naval Research.

References

Allen, J. (1995). Natural language understanding (Vol. 2). Redwood City, CA: Benjamin/Cummings.

Baker, C. F., Fillmore, C. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281-296.

Culicover, P. W., & Jackendoff, R. (1999). The view from the periphery: The English comparative correlative. *Linguistic Inquiry*, 30(4), 543-571.

Das, D., Chen, D., Martins, A. F., Schneider, N., & Smith, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1), 9-56.

Fillmore, C. J., Wooters, C., & Baker, C. F. (2001). Building a Large Lexical Databank Which Provides Deep Semantics. *Proceedings of PACLIC-2001*.

Fillmore, C. J., Lee-Goldman, R., & Rhodes, R. (2012). The FrameNet Constructicon. *Sign-based Construction Grammar. CSLI, Stanford, CA*.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*, 85–168.

Friedman, S., Forbus, K., & Sherin, B. (2011). How do the seasons change? Creating & revising explanations via model formulation & metareasoning. *Proceedings of the 25th International Workshop on Qualitative Reasoning.*

Hinrichs, T. and Forbus, K. 2012. Learning Qualitative Models by Demonstration. *Proceedings of QR2012*.

Kuehne, S. E. (2004). *Understanding natural language descriptions of physical phenomena.* Doctoral Dissertation, Northwestern University, Evanston, Illinois.

McFate, C.J., Forbus, K. and Hinrichs, T. (2014). Using Narrative Function to Extract Qualitative Information from Natural Language Texts. *Proceedings of AAAI-2014*

Ovchinnikova, E., Vieu, L., Oltramari, A., Borgo, S., & Alexandrov, T. (2010). Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. *In LREC-2010*.

Sherin, B. L., Krakowski, M., & Lee, V. R. (2012). Some assembly required: How scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching*, 49(2), 166-198.

Tomai, E. (2009). *A Pragmatic Approach to Computational Narrative Understanding.* Doctoral Dissertation, Northwestern University, Evanston, Illinois.

Inferring Generic Meaning From Pragmatic Reference Failure

Phil Crone
pcrone@stanford.edu
Department of Linguistics
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Generic sentences (e.g., "birds lay eggs") express generalizations about kinds, in contrast to non-generic sentences that express facts about specific individuals or sets of individuals (e.g., "all birds lay eggs"). Although generics are pervasive in natural language, there is no unique linguistic marker of genericity, making the identification of generics a challenge. We investigate the morphosyntactic cues that listeners use to identify whether a sentence should receive a generic interpretation or not. We find that two factors – the definiteness of a sentence's subject NP and the tense of the sentence – are extremely important in guiding intuitions about whether a sentence should receive a generic interpretation. We argue that the importance of these factors can be explained by taking generic interpretations to arise due to a failure to ground expressions as referring to specific entities or events.

Keywords: Psycholinguistics; pragmatics; generics

Introduction

Generic sentences express generalizations about kinds and are an important tool for the transmission of knowledge (Gelman, 2003). A key difference between generic and non-generic statements is that generics allow for exceptions: "birds fly" expresses a generalization about the kind *bird* and is true despite the fact that some birds do not fly, while "all birds fly" is false, because, e.g., emus do not fly (Prasada, 2000). Despite their importance, generics are not consistently marked by any particular lexical, morphological, or syntactic convention. This presents a challenge for speakers who intend to communicate generic meanings and listeners who must correctly interpret speakers' intended meaning.

Prior work suggests that morphosyntactic features, pragmatic factors, and world knowledge all serve as cues that guide the interpretation of sentences as generic or not. In English, the subject noun phase (NP) of a generic sentence is often a bare plural ("birds fly"), but it can also be an indefinite singular ("a bird has feathers") or definite singular ("the bird is a warm-blooded animal"). In contrast, definite plural subject NPs ("the birds have beaks") are generally thought to force non-generic interpretations. Tense and aspect also as cues to generic meaning; simple present tense sentences ("birds fly") are more associated with generic meanings than, e.g., present progressive ("birds are flying overhead") or past tense ("birds flew by my window") sentences (Carlson, 1977; Krifka et al., 1995; Lyons, 1977).

Pragmatics and world knowledge also influence a sentence's interpretation as generic or non-generic. For example, if a unique bird is present in the context of an utterance of a sentence with the subject NP "the bird," this NP is likely to be interpreted as referring to the bird in the context, giving rise to a non-generic interpretation. World knowledge about properties shared by members of a kind may influence the interpretation of potentially generic sentences. For instance, the sentence "a bird does not fly" is likely to be interpreted as being about some particular bird, given world knowledge that, in general, birds fly.

Previous experimental work has confirmed the influence of these three factors—morphosyntax, pragmatic context, and world knowledge—on the interpretation of sentences as generic. Adults and children can both use the definiteness of subject NPs, as well as tense and aspect, to identify generics, and prefer generic interpretations when the subject NP has no available referent in context (Gelman & Raman, 2003; Cimpian, Meltzer, & Markman, 2011). By age 3, children are less likely to assign a generic interpretation to a sentence when its subject NP has a possible referent in the preceding linguistic context and can also use knowledge about the generalizability of properties to kinds in identifying generics (Cimpian & Markman, 2008). World knowledge about whether the subject of a sentence is animate or inanimate also affects interpretations of genericity (Brandone & Gelman, 2009).

In the current paper, we propose that these findings can be synthesized under a single explanation: generic interpretations are driven by a failure to find reference for entities or events described by the sentence in question. We hypothesize that listeners consider two possible interpretations of potentially generic statements: the speaker is either trying to be informative about a particular entity or set of entities, or else the speaker is trying to be informative about a kind. If reference to the particular fails conspicuously, then kind reference becomes likely.

This hypothesis—which we refer to as the "pragmatic reference failure hypothesis"—is inspired by the finding that the referential status of subject NPs in a particular context affects judgements of genericity (Gelman & Raman, 2003). It also provides a candidate explanation for morphosyntactic effects. Like referential context, features like tense and definiteness influence the probability that a speaker was referring to a specific entity. The definite determiner presupposes the existence of a unique, salient entity in the common ground, suggesting that the speaker intends to refer to a particular entity. Certain tenses and aspects, such as simple past and progressive aspect, also have a referential character that English simple present tense lacks.

In the current investigation, we build on the prior work described above by assessing the impact of morphosyntactic factors on adults' judgments about genericity across a wide variety of sentences. Most previous work on the identifica-

tion of generics has focused on children's abilities, perhaps stemming from an assumption that children face challenges in identifying generics that are less relevant for adults. However, research on the probabilistic nature of language comprehension suggests that adults face a similar problem (Levy, 2008; Frank & Goodman, 2012). In the case of identifying generics, we take adults to reason about the likelihood that an utterance is generic given morphosyntactic features of the sentence, features of the context, and the their own world knowledge. The current study also differs from previous work in that we collect naturalistic examples of generic and non-generic sentences generated by study participants, rather than rely on artificially constructed examples. This allows us to consider a realistic representation of genericity in natural language. In this way, we follow recent work on the interpretation of naturally-occurring uses of quantifiers involved in scalar implicature patterns, rather than constructed examples of scalar implicatures (Degen, 2015).

The paper is organized as follows. We first introduce an experimental paradigm that allows us to gather a large number of naturalistic examples of generic and non-generic sentences from study participants. We then collect judgments from a second, independent set of participants, as to whether these sentences should receive generic or non-generic interpretations (Experiment 1). We next use several machine learning techniques to identify which factors are most useful in classifying sentences as generic or not. We find that just two factors—definiteness of the subject NP and tense—prove extremely successful in classifying sentences. Finally, we validate the importance of tense for generic interpretation with human subjects (Experiment 2). Taken together, the results of these experiments show that judgments of whether a sentence is generic or non-generic can largely be explained by considering whether the sentence can be referentially grounded.

Experiment 1

It has previously been argued that the number and definiteness of a sentence's subject NP influence its interpretation as generic or non-generic. Previous work investigating these cues to genericity have generally fixed the number of the subject NP as either singular (Cimpian et al., 2011) or plural (Gelman & Raman, 2003) and only manipulated definiteness. In Experiment 1, we considered the effects of number, definiteness, animacy, and their interaction on the interpretation of generics. A set of participants performed a sentence completion task in which the subject NP was provided. A second set of participants then classified these sentences as generic or non-generic.

Method

Participants We recruited two sets of participants through Amazon's Mechanical Turk website to complete two independent tasks. We restricted participants to individuals within the United States and paid all participants 50 cents. The sets consisted of 100 and 94 participants, respectively. The first task took approximately 14 minutes to complete, while the second task took approximately 6 minutes. We excluded from the analysis 4 participants from the first set and 6 participants from the second set who indicated that their native language was not English.

Stimuli In task 1, the first set of participants was presented with a sequence of NPs followed by a single-line text box. Participants were instructed to "write a sentence starting with the phrase below." The NPs were generated from a set of 48 nouns, split evenly between animates and inanimates. For each participant, each noun was randomly assigned morphosyntactic features using a 2×2 factorial design crossing number (singular, plural) with definiteness (definite, indefinite). Full NPs were created from the base noun and the randomly chosen number and definiteness features. For example, if the noun "panda" received values *plural* and *definite*, the full NP would be "the pandas."

The sentences produced by self-reported native English speakers were then split into random sets of approximately 50 sentences each. For task 2, each set of sentences was presented to one individual from the second set of participants. For a sentence whose subject NP was *noun*, participants indicated whether they thought the sentence was about "*nouns*," "a specific *noun*" (for sentences with singular subject NPs), or "a specific group of *nouns*" (for sentences with plural subject NPs).[1]

Procedure In task 1, we first presented participants with four example NPs. After providing a sentence completion for each example item, participants were shown an example sentence completion that they could have written for the given NP. These example sentences were constructed to favor non-generic interpretations for all NP types. Next, all 48 subject NPs were presented in pseudorandom order, counterbalanced so that no two consecutive NPs matched in both number and definiteness. We required each sentence completion to be at least six characters in length. After providing sentence completions for all NPs, participants reported their native language.

In task 2, participants were instructed that they would be viewing a sequence of approximately 50 sentences that had been produced by other Mechanical Turk workers. Participants began by judging 4 example sentences constructed by us for which the participants received feedback. The remaining sentences were presented in random order. After completing task 2, participants reported their native language. We measured reaction times for each item, measured from the time the item was presented until the time a response was submitted. Reaction times were used to exclude from the analysis responses in whose reaction times were greater than 2 standard deviations from the mean, but were not further analyzed.

[1] Participants included in task 1 also judged whether the sentences they produced were generic or non-generic in a similar manner. Their responses are not reported here, but were similar to the results obtained in task 2 ($r = 0.67$). We also ran a separate rating experiment using the phrasing "*nouns* in general" for the generic question, and found consistent ratings ($r = 0.99$, Fleiss' $\kappa = 0.81$).

Table 1: Example Productions from Experiment 1.

Definiteness	Number	Genericity	Examples
Indefinite	Singular	Generic	"An elephant is large.", "A car is a form of transportation."
Indefinite	Singular	Non-generic	"A dolphin swam alongside the boat.", "A guitar kept me up all night."
Indefinite	Plural	Generic	"Foxes chase rabbits.", "Computers have made communication easier."
Indefinite	Plural	Non-generic	"Owls were perched in the barn.", "Marbles rolled all over the floor."
Definite	Singular	Generic	"The peacock is a noisy bird.", "The chair has four legs."
Definite	Singular	Non-generic	"The pigeon landed on the car.", "The trumpet was out of tune."
Definite	Plural	Generic	"The kangaroos jump a lot.", "The guitars are stringed instruments."
Definite	Plural	Non-generic	"The pandas ate bamboo.", "The fences were strong."

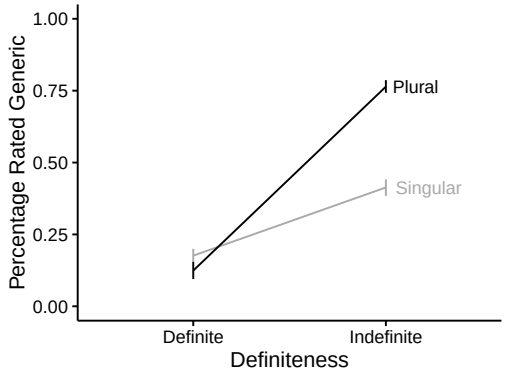


Figure 1: Percentage of sentences rated generic in Experiment 1 by definiteness and number. Error bars show 95% confidence intervals.

Results and Discussion

Both definiteness and number affected judgments (Figure 1). Sentences with indefinite subject NPs—especially bare plural subjects—were more likely to be rated generic (see Table 1 for examples). We fit a logistic mixed-effects model to predict a sentence's classification as generic or non-generic from the interaction between definiteness, number, and animacy of the subject NP.[2] The model identified significant main effects of definiteness such that sentences with indefinite subject NPs were more likely to be rated generic ($\beta = 3.29, z = 14.41, p < 0.001$) and animacy such that sentences with inanimate subjects were less likely to be rated generic ($\beta = -1.29, z = -3.33, p < 0.001$). There were also significant interactions between animacy and definiteness such that sentences with inanimate, indefinite subjects were more likely to be rated generic ($\beta = 0.86, z = 2.52, p < 0.05$) and between definiteness and number such that sentences with indefinite, singular subjects were less likely to be rated generic ($\beta = -1.84, z = -7.59, p < 0.001$).

The results are consistent with previous findings that indefinite singulars and bare plurals facilitate generic interpretations compared to definite singulars and definite plurals, respectively (Cimpian et al., 2011; Gelman & Raman, 2003). The interaction between definiteness and plurality reveals a superadditive effect by which indefinite, bare plurals were rated more generic than would be predicted by the main effects of definiteness and plurality. This is unsurprising, as bare plurals are often taken to be the canonical subject type for English generics. The effect of animacy is also consistent with previous findings that both children and adults produce more generic statements when describing animals than when describing artifacts (Brandone & Gelman, 2009).

Comparing the effect sizes across significant predictors, we find that definiteness had the greatest influence on participants' responses. This finding is straightforwardly interpretable in terms of the pragmatic reference failure hypothesis, which states that generic interpretation arises from reference failure. Since the definite determiner *the*, unlike indefinite determiners, presupposes the existence of a unique, salient entity in the common ground, it follows that definite subject NPs are more likely to refer to some particular entity. Therefore, sentences with definite subjects are less likely to have an intended generic interpretation.

Note, however, that there were sentences with definite subject NPs that were judged generic. Thus, we cannot conclude that sentences with definite subject NPs categorically receive non-generic interpretations. Rather, we argue that the presuppositions of the definite determiner are most easily satisfied if the NP can be grounded in a concrete referent present in the common ground. In principle, these presuppositions can be satisfied by taking the NP to be kind-referring, but such an interpretation requires an additional interpretative step to conclude that the speaker does not intend to refer to particular individual in the common ground.

One unexpected finding is that across all subject NP types, including plural definites, sentences were occasionally judged to be generic. This seems to clash with the general view that definite plurals do not allow for generic interpretations in English. This result forced us consider whether our methodology measured some property other than genericity. However, inspection of definite plurals that were judged generic suggested that these ratings were at least *prima facie* appropriate (Table 1). Moreover, recent work has suggested that sentences with definite plural subject NPs in En-

[2]All data and code can be viewed at `https://github.com/langcog/generic_ref`.

glish can receive generic interpretations, at least in certain circumstances (Farkas & de Swart, 2007), but that such uses may come with additional social meaning (Acton, 2014).

Classifying Generics

In Experiment 1, we directly manipulated the definiteness, number, and animacy features of subject NPs and observed how these manipulations influenced genericity judgments. However, the sentences produced in Experiment 1 varied with respect to many morphosyntactic, lexical, and semantic respects that we did not directly control. It is plausible that these factors also contributed to judgments about the classification of sentences as generic or not. To probe the influence of these factors on participants' responses, we extracted additional linguistic features from the sentences produced in Experiment 1 and then employed a number of classification techniques to determine which were most important in determining whether a sentence would be judged generic.

Method

We first used the Stanford part-of-speech (POS) tagger (Toutanova, Klein, Manning, & Singer, 2003) to obtain POS tags and tense/aspect features for verbs for all sentences judged in Experiment 1. Using these POS tags, we automatically extracted the following features for each sentence: tense, aspect, presence of main verb *be* or *have*, and presence of a modal. We also calculated the total sentence length minus the length of the subject NP for each sentence.

We next split the sentences from Experiment 1 into training and test sets, with the training set consisting of 75% of all sentences produced. Using the training set, we trained a variety of models to classify sentences as generic or non-generic. All models included definiteness, number, and animacy of the subject NP, the factors extracted using POS tags, and sentence length as predictors. Models included generalized linear models, basic decision trees, boosted decision trees, and random forests. Model parameters were optimized using 10-fold cross-validation on the training set. The models were then evaluated against the 25% held-out test set.

Results & Discussion

All models achieved classification accuracy on the test set between 85% and 87%. The most straightforwardly interpretable of these methods was a basic decision tree pruned to have three terminal nodes (Figure 2). This decision tree considers only the definiteness of the subject NP and the tense of the sentence and achieved a test accuracy of 86.7%.

To assess the reliability of human judgments (and hence the theoretical ceiling for the classifiers), an independent set of participants re-rated the sentences in the held-out test set. The experimental details were the same as those for task 2 of Experiment 1. We recruited 41 participants, all of whom reported that their native language was English. The agreement between these judgments and those collected in Experiment 1 was substantial: the judgments were in agreement 87.8% of the time, virtually identical to the accuracy of the machine classification techniques we employed (Cohen's $\kappa = 0.73$).

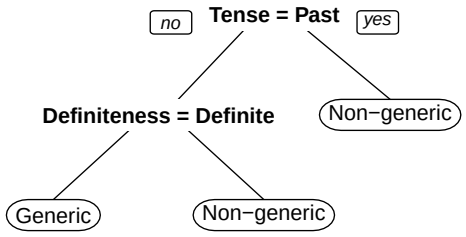


Figure 2: Decision tree with three terminal nodes for classification of sentences as generic or non-generic.

The most striking finding in our attempt to classify sentences as generic was that the vast majority of factors considered did not prove useful in this classification task. In the end, we were able to achieve human-level accuracy in the classification task using only the definiteness of a sentence's subject NP and the tense of the sentence. As discussed in the context of the results of Experiment 1, the definite article presupposes the existence of a unique, salient entity. Thus, a listener can infer that the speaker intended to refer to some particular individual, rather than a kind.

The importance of tense can be understood in a similar manner. Partee (1973) observes from the interaction of tense and negation that the past tense in English has a referential character. "I didn't turn off the stove," neither means that there is no time in the past when the speaker turned off the stove (which is false) nor that there is any time in the past when the speaker did not turn off the stove (which is trivially true). Rather, the speaker refers to a specific interval of time during which he/she did not turn off the stove. In general, English simple past tense clauses refer to particular intervals of time during which particular events occurred (Heim, 1994). In contrast, the vast majority (92%) of non-past sentences produced in Experiment 1 used simple present tense, which is employed in English to express habits or ongoing states, rather than strictly temporally delimited events. While kinds are unlikely to participate in specific events, habits or states may express general properties of kinds.

We propose that the influence of definiteness and tense on the interpretation of sentences as generic or non-generic can be explained by the pragmatic reference failure hypothesis. Listeners attempt to referentially ground the expressions in particular entities or events. Failure to identify referents for these expressions gives rise to generic interpretations.

Experiment 2

The results of the classification task described above suggest that tense plays a critical role in guiding judgments as

to whether a sentence should be interpreted as generic. In order to confirm this finding, we ran an experiment in which we directly manipulated the tense of sentences to see if this would change interpreters' judgments about genericity. This experiment is partial replication of Cimpian et al. (2011), in particular their findings regarding the role of tense in guiding generic interpretations. However, our study differs from previous work in that we compare the magnitude of the effect of tense to those of other effects. In addition, we consider whether tense meaningfully interacts with other factors in driving generic interpretations.

Method

Participants Recruitment details were similar to those of task 2 in Experiment 1. We recruited a total of 50 participants, all of whom were self-reported native English speakers.

Stimuli We chose a set of 48 sentences produced in Experiment 1, one for each base noun used in the experiment. All subject NP types, crossing definiteness, number, and animacy, were equally represented in this set. None of the selected sentences contained main verb *be* or *have*, a modal, or a passive construction. The selected sentences were evenly split between those with simple present and simple past tense verbs. In several cases, sentences produced in Experiment 1 were slightly altered to avoid awkward phrasing or to change the definiteness or number feature of the subject NP to ensure that all pairs of number and definiteness features were equally represented.

Once all 48 sentences were collected, we created alternate versions of each sentence in which the tense of the main verb was changed from simple present to simple past or vice versa.

Procedure The procedure was similar to that of part 2 in Experiment 1. Participants began by judging four example sentences. After this, participants viewed all 48 sentences in random order and indicated whether the sentence was about "*nouns*" or "a specific *noun*/group of *nouns*." Half of sentences judged by each participant contained simple present tense verbs, while half contained simple past tense verbs. Items were presented in random order; the order of tenses was also randomized.

Results and Discussion

The effects of definiteness and number of the subject NP were broadly similar to the effects found in Experiment 1. In addition, the tense manipulation greatly influenced judgments, with present tense sentences being more likely to be judged generic across all subject NP types (Figure 3).

We fit a logistic mixed-effects model to predict sentences' genericity classifications from the interaction of tense, definiteness, number, and animacy. The model revealed only two significant effects; sentences with indefinite subject NPs were more likely to be judged generic ($\beta = 2.81, z = 3.04, p < 0.01$), as were present tense sentences ($\beta = 2.31, z = 3.26, p < 0.01$). The effects of tense and definiteness are similar in magnitude and provide further confirmation of the findings of the classification task.

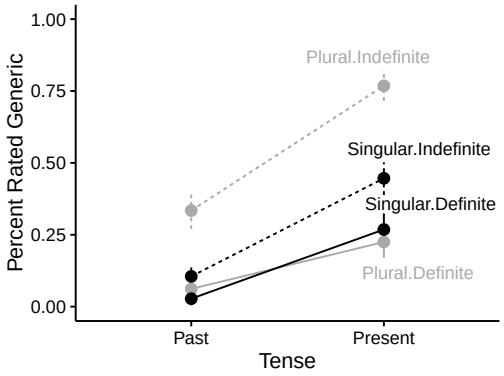


Figure 3: Percentage of sentences rated generic in Experiment 2 by tense, definiteness, and number. Error bars show 95% confidence intervals.

Note also that the consistency of the effect of tense across all subject NP types further confirms that definite subject NPs are not precluded from expressing generic meanings. Rather, as discussed above, it is more difficult to arrive at a generic interpretation of a sentence with a definite subject NP. However, this difficulty can be overcome if other factors, such as tense, support a generic interpretation.

General Discussion

We set out to explore the question of what gives rise to generic interpretations of sentences. The results here replicate previous findings about cues to genericity: Indefinite subject NPs support generic interpretations more than definite subject NPs (Cimpian et al., 2011; Gelman & Raman, 2003), present tense sentences are more likely to be judged generic than past tense sentences (Cimpian et al., 2011), and sentences with animate subjects tend to be more generic than those with inanimate subjects (Brandone & Gelman, 2009). Overall, we found that the most important factors for identifying generic statements were the definiteness of the sentence's subject NP and the tense of the sentence. We have argued that the importance of these two factors can be explained by viewing generic interpretations as the result of listeners' failure to ground expressions as referring to particular entities or events.

As noted above, there are exceptions to the generalizations discussed here. Even past tense sentences with definite plural subjects may be interpreted generically in certain cases (e.g. "The dinosaurs went extinct.") How can we account for such cases? Judgments regarding generic or non-generic meanings must be viewed as graded inferences about speakers' intentions. The factors that we have discussed here influence the likelihood that listeners will interpret a sentence as generic or not, but none of these factors can be viewed as categorically creating or precluding a generic interpretation.

We leave open for future research the question of the extent to which the pragmatic reference failure hypothesis is generalizable to cover generic interpretations in other lan-

guages. Some existing work supports the idea that tense and aspect influence genericity judgments in Spanish in a similar manner as reported here (Pérez-Leroux, Munn, Schmitt, & DeIrish, 2004). However, the role of definiteness in signaling generic and non-generic meanings is not expected to be the same across languages. For example, Romance languages canonically express genericity with sentences whose subjects are definite plural NPs, the subject NP type that is least associated with generic meanings in English. The theoretical linguistics literature offers several potential explanation for this cross-linguistic difference. It has been argued that languages that canonically express genericity with definite plural subjects do so because determiner-less NPs, like English bare plurals, cannot occupy arguments positions in these languages (Chierchia, 1998). Alternatively, languages may simply differ in whether their definite determiners may lexicalize reference to kinds or not (Dayal, 2004).

Our explanation regarding the source of generic interpretations has the potential to explain other factors that have previously been shown to serve as cues to generic meaning. For example, although we did not find aspect to play a crucial role in classifying sentences as generic or not, previous work has found that both children and adults are more likely to interpret present progressive sentences as non-generic than simple present sentences (Cimpian et al., 2011). Our failure to identify aspect as a crucial cue to generic meaning is possibly due to the fact that only 6.8% of the sentences produced in Experiment 1 used progressive or perfect aspect. However, our hypothesis is amenable to taking aspect to play a role in determining generic meaning. Present progressive sentences in English refer to events that are occurring at the speech time. A listener can infer from a sentence using the present progressive that a speaker intended to refer to a particular event.

Our proposal also straightforwardly explains previous findings that the contextual availability of a referent for the subject NP of a sentence makes it more likely that the sentence will be judged non-generic (Gelman & Raman, 2003). The presence of a referent for the subject NP greatly reduces the difficulty on the part of a listener to referentially ground this expression. We therefore predict that in such cases, listeners will be more likely to arrive at a non-generic interpretation.

We look forward to future work that tests the extension of the pragmatic reference failure hypothesis to explain additional factors influencing generic and non-generic interpretations and to explain cues to generic meaning in languages other than English.

Acknowledgments

Thanks to Daniel Lassiter, Rose Schneider, and the members of the Language and Cognition Lab. We gratefully acknowledge the support of NSF BCS Grant 1456077.

References

Acton, E. (2014). *Pragmatics and the social meaning of determiners*. Unpublished doctoral dissertation, Stanford University, Stanford.

Brandone, A. C., & Gelman, S. A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts: A window onto a conceptual divide. *Cognition*, *110*(1), 1-22.

Carlson, G. N. (1977). *Reference to Kinds in English.* Ph.D. dissertation. University of Massachusetts, Amherst.

Chierchia, G. (1998). Reference to kinds across languages. *Natural Language Semantics*, *6*(4), 339–405.

Cimpian, A., & Markman, E. (2008). Preschool children's use of cues to generic meaning. *Cognition*, *107*(1), 19–53.

Cimpian, A., Meltzer, T. J., & Markman, E. (2011). Preschoolers' use of morphosyntactic cues to identify generic sentences: Indefinite singular noun phrases, tense, and aspect. *Child Development*, *82*(5), 1561–1578.

Dayal, V. (2004). Number marking and (in)definiteness in kind terms. *Linguistics and Philosophy*, *27*(4), 393–450.

Degen, J. (2015). Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, *8*(11), 1–55.

Farkas, D. F., & de Swart, H. (2007). Article choice in plural generics. *Lingua*, *117*, 1657–1676.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S. A., & Raman, L. (2003). Preschool children use linguistic form class and pragmatic cues to interpret generics. *Child Development*, *74*(1), 308–325.

Heim, I. (1994). Comments on Abusch's theory of tense. *Ellipsis, tense and questions*, 143–170.

Krifka, M., Pelletier, F. J., Carlson, G. N., ter Meulen, A., Link, G., & Chierchia, G. (1995). Genericity: An introduction. In *The Generic Book.* Chicago: The University of Chicago Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lyons, J. (1977). *Semantics: Volume 1*. New York: Cambridge University Press.

Partee, B. (1973). Some structural analogies between tenses and pronouns in English. *Journal of Philosophy*, *70*, 601–609.

Pérez-Leroux, A. T., Munn, A., Schmitt, C., & DeIrish, M. (2004). Learning definite determiners: genericity and definiteness in English and Spanish. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *Supplement to the proceedings of the 28th Annual Boston University Conference on Language Development.*

Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Science*, *4*(2), 66–72.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network.*

Effect of Aging on Inhibitory Attentional Mechanisms

Maegen Walker (maegenw@hawaii.edu)
Margeaux Ciraolo (ciraolo@hawaii.edu)
Andrew Dewald (adewald@hawaii.edu)
Scott Sinnett (ssinnett@hawaii.edu)
Department of Psychology, University of Hawai'i at Mānoa
2530 Dole Street, Honolulu, HI 96822 USA

Abstract

The ability to inhibit the processing of irrelevant information declines as adults age (Hasher & Zacks, 1988; Lustig, Hasher and Tonev, 2006; Mayr, 2001). However, previous research investigating inhibitory control in older adults has not evaluated the *extent* to which irrelevant information is processed and later recognized. Using a dual task paradigm with young adults, Dewald, Sinnett, and Doumas (2011) demonstrated inhibited recognition for previously ignored words, provided they had appeared infrequently with targets in the primary task, compared to words that did not appear with targets. The current study adapted this paradigm to examine inhibitory mechanisms in a sample of older adults. Here, older adults exhibited inhibited recognition for all words while young adults continued to show greater inhibition for words that had appeared with targets compared to words that had not. This finding suggests that older adults may experience a decline in the *selective* inhibition of irrelevant information.

Keywords: Aging; Attention; Dual Task Paradigms; Inhibition; Inattentional Blindness

Introduction

One critical mechanism for information processing is the ability to *inhibit* irrelevant information from being processed (Tipper, 1992). This is accomplished by directing attention toward the desired target or task while ignoring potential distractors. For example, in order to drive a car one must be able to focus on driving (e.g., minding roads and traffic signals) while ignoring potential distractors, such as billboards or an incoming text message. However, attentional capabilities change over the course of the human lifespan and it is well established that performance decreases as adults progress in to old age (Campbell, Grady, Ng, & Hasher, 2012; Mayr, 2001; Rabbitt, 1965; Störmer, Heekeren, & Lindenberger, 2013). This age related cognitive decline in attention has been largely attributed to deficiencies in inhibitory control, with older adults unsuccessfully inhibiting the processing of irrelevant information (Hasher & Zacks, 1988; Kramer, Hahn, & Gopher, 1999; Lustig, Hasher and Tonev, 2006; Madden, Pierce, & Allen, 1996; Mayr, 2001; Plude & Hoyer, 1986).

Illustrative of this cognitive decline, older individuals exhibit difficulties in tasks involving selective attention (i.e., selecting a stimulus of interest and focusing attentional resources toward it while inhibiting irrelevant information). A study conducted by Farkas and Hoyer (1980) used a card-sorting version of a visual search task and found that, compared to younger individuals, older adults were slower to respond when presented with a distractor card that was very similar, compared to cards that were dissimilar, to the target card. This finding suggests that older adults are more likely to be distracted by these items, indicating a reduced ability to inhibit irrelevant items from re-orienting attention.

Additional studies of selective attention demonstrate that, as adults progress in to old age, top-down control over attentional allocation may become impaired when viewing displays containing moving items (Folk & Lincourt, 1996; Watson & Maylor, 2002). In this case, older individuals appear to have more trouble visually identifying the target stimulus (Watson & Humphreys, 1997; 1998). This suggests that they may be less able to inhibit processing of previously viewed distractor objects such that they continue to capture attention during the visual search task. Therefore, being unable to inhibit processing for the irrelevant items appears to negatively impact perception of the target stimulus in moving displays.

The most compelling evidence supporting the perspective that the ability to inhibit the processing of irrelevant information declines as adults age has come from work with the classic Stroop task. During this task, participants are presented with a series of written color words appearing in a variety of different colors (i.e., the word "yellow" written in the color red). Participants must inhibit reading the word (e.g., yellow) and report only the color (e.g., red). When comparing performance on this task between younger and older adults, the tendency to incorrectly report the word rather than the color (i.e., Stroop interference) is significantly more pronounced for older participants (Brink & McDowd, 1999; Hartley, 1993; Spieler, Balota & Faust, 1996). That is, the ability to inhibit the incorrect responses appears to be compromised in older adults.

Milham et al. (2002) conducted further work exploring inhibitory processes with the Stroop task in aging adults using fMRI. Along with higher instances of Stroop interference in older participants, results from this study suggest decreased responsiveness in brain regions believed to be associated with attentional control and working memory, in particular the dorsolateral prefrontal cortex (DLPFC) and the parietal cortex (Banich et al., 2000a, 2000b; MacDonald, Cohen, Stenger, & Carter, 2000; Sinnett, Snyder, & Kingstone, 2009). These particular brain regions are associated with the modulation of neural activity

by facilitating processing systems that contain task-relevant information while inhibiting systems that contain task-irrelevant information. Through this early inhibition of processing irrelevant information, the DLPFC suppresses activation of semantic and phonological associations as well as potential actions related to the irrelevant information (Banich et al., 2000b).

The Stroop task is favored among those investigating the modulation of inhibitory attentional mechanisms with an aging population because participants must simultaneously facilitate the processing of the written word *color* (i.e., say "red" if the word "yellow" is written in the color red) and inhibit the processing of the semantically written word (i.e., "yellow"). However, this, and previously used behavioral methods, have two important limitations. First, when using the Stroop task it is difficult to isolate facilitatory from inhibitory mechanisms because this paradigm does not offer an effective method for dissociating these two processes. Thus, when Stroop interference occurs it is difficult to determine if the attentional system is failing to facilitate processing of the presented color (i.e., the relevant information) or if it is failing to inhibit processing of the written word (i.e., the irrelevant information). Second, other previously utilized behavioral methods evaluate the function of inhibitory mechanisms by looking at rates of distraction induced by the irrelevant stimuli. That is, the ability to inhibit processing of the irrelevant item is assessed by examining accuracy or reaction time to a target that is presented at the same time as the distracting item, with higher rates of distraction and slower reaction times indicating reduced inhibitory control. While this is useful information, the *extent* to which these ignored items are actually processed and subsequently stored in long-term memory is presently undetermined.

To address these gaps in the literature, we adapted a dual-task paradigm (see Dewald, Sinnett, & Doumas, 2011; 2012) for use with older adults. This paradigm overcomes the ambiguity presented in the Stroop task by varying the frequency with which irrelevant distractor (i.e., ignored) items are presented simultaneously (i.e., paired) with attended target items. This method allows for the isolation and examination of both inhibitory and facilitatory mechanisms separately. Previous research using this paradigm with young adults has shown that infrequently pairing the ignored distractor items with targets in the attended task leads to *inhibited* processing of these distractor items over ignored distractor items that are *not* paired with targets (see Dewald et al., 2011)[1]. The extent to which the irrelevant information may have been processed is evaluated through the use of a surprise recognition test for the previously ignored distractor items.

The current study focuses on evaluating the proposed inhibitory effect in an older adult population and compares performance on this task to a sample of younger adults. Based on findings from previous research suggesting that older adults experience an overall decline in their ability to inhibit processing of irrelevant information (Hasher & Zacks, 1988; Kramer et al., 1999; Lustig et al., 2006; Madden, et al., 1996; Mayr, 2001; Plude & Hoyer, 1986), we predict that older adults will be less able to inhibit processing the distractor items during the primary task compared to the young adults. This decline in inhibitory control should lead to the ignored items being processed to a greater extent. As a result, older adults should have overall higher recognition scores than the younger adults on the surprise recognition test as younger adults tend to inhibit these items (Dewald et al., 2011). Specifically, older adults should recognize all distractor items at or around chance levels while younger adults should exhibit greater inhibition for distractor items that were paired with task targets when compared to distractor items that were not paired with task targets. This finding would suggest that older adults are less able to inhibit processing of the irrelevant information resulting in greater amounts of the ignored information capturing attention, being stored in long-term memory, and subsequently recognized more often later.

Methods

Participants

Thirty-nine young adults (23 female, mean age of 20.7) were recruited from undergraduate courses at the University of Hawai`i at Mānoa in exchange for course credit. The results from one participant were excluded from the analyses due to a failure to complete the surprise recognition task. The final analyses were conducted with the remaining 38 young adults (22 female, mean age of 20.8).

Twenty-six healthy older adult participants (18 female, mean age of 72.2) were recruited, on a voluntary basis, from local retirement communities around Honolulu, Hawai`i, as well as from continuing education programs for seniors at the University of Hawai‘i at Mānoa. This target age group (>60 years old) was chosen based on criteria set by the World Health Organization (2014) designating 60 as the generally accepted age at which an individual is considered to be 'elderly'. All participants were naïve to the experiment, provided informed consent, and had normal, or corrected to normal, vision and hearing.

Stimuli

A total of 50 pictures (on average 5 to 10 cms) were selected from the Snodgrass and Vanderwart (1980) picture database (i.e., attended stimuli). Each picture was superimposed with a single English word (i.e., ignored distractor items) selected from a pool of high frequency words retrieved from the MRC psycholinguistic database

[1] This same body of literature has also demonstrated facilitated processing of ignored information using a variation of this paradigm, (Dewald & Sinnett, 2012, 2013; Walker, Dewald, & Sinnett, 2014; Seitz & Watanabe, 2003; Watanabe, Nàñez & Sasaki, 2001). However, the current study is only concerned with conditions of inhibited processing.

(Wilson, 1988). All words had an average length of 5 letters (range 4-6) and average frequency of 120 per million (range 28-686). The words were superimposed over the pictures in bold, capitalized letters and presented in Arial font (24 points). Care was taken to ensure that picture-word combinations did not have any semantic relationship.

Attended Stimuli The 50 pictures were duplicated resulting in two copies of each picture (i.e., picture pairs). All pictures were randomly rotated +/-30 degrees from their original orientation to ensure that the task is sufficiently demanding in each version of the experiment (see Rees et al., 1999). To create an experimental block, half (25) of the picture pairs were presented as immediate picture repetitions. These immediate picture repetitions served as the identification targets for the attended task. The remaining 25 pairs did not occur as immediate repetitions in the same block. Instead, they were separated from their duplicate and randomly inserted in between occurrences of repeating picture pairs creating the non-repeating pictures in the visual stream. This process was repeated to create the second experimental block using the same stimuli. Critically, the 25 picture pairs that served as immediate repetitions in the first block did not repeat in the second block and those that did not repeated in the first block did repeat in the second block (i.e., picture pairs that were identification targets in the first block were non repeating pictures in the second block and vice versa). Therefore, each of the original 50 pictures was presented four times, once as a target repetition pair in the first block, then again as a non-repeating picture pair in the second block.

Ignored Distractor Items 100 words were randomly selected and superimposed on the pictures. Half (50) of the words appeared superimposed on immediate picture repetitions (i.e., targets), serving as the *target-aligned* (TA) words, and the other half were superimposed on the non-repeating pictures in the stream, serving as *non-aligned* (NA) words. This created a block size of 100 picture-word combinations with 25 immediate target picture repetitions and accompanying superimposed TA words for each of the two blocks. The same 100 words and 50 picture pairs used in the first block were used in the second block and the same procedure was applied. Like the pictures, the 25 words that were TA in the first block were NA in the second block and vice versa. Therefore, all of the words were presented twice in the experiment. Twenty-five of the 50 TA words were presented as TA in the first block and were then presented as a NA words in the second block and vice versa. The NA words were always presented on top of non-repeating pictures in both blocks (i.e., always NA). This method was used to create six different versions of the experiment (see Dewald et al., 2011).

Surprise Recognition Test The later surprise recognition test for the ignored words was administered after participants had completed the primary attended task. This test consisted of 50 words from the experiment along with 50 never before seen foil words, selected from the same database (Wilson, 1988). Because of the high number of words presented, two types of surprise recognition tests were created for each version of the experiment. One tested only the 50 TA words along with 50 foil words and the other tested only the 50 NA words along with 50 foil words for a total of 100 words in each recognition test. Each participant was randomly assigned to get one type of test only (TA words or NA words). The recognition tasks were randomized and presented one word at a time. The words were written in bold, capitalized letters in Arial font at a size of 24 points (i.e., identical to their initial presentation in the repetition detection task).

Procedure

Participants were seated in front of a computer with the screen a comfortable distance away. They were then shown a rapid serial visual presentation (RSVP) of the picture-word stream, using DMDX software (Forster & Forster, 2003). In the primary repetition detection task participants were instructed to ignore the superimposed words and attend only to the pictures. Participants were required to respond when they noticed a picture immediately repeat by clicking the left mouse button with their preferred hand. Each item in the picture-word presentation was presented for 500ms with a 150ms inter-stimulus interval (ISI; blank screen) between each item for a stimulus onset asynchrony (SOA) of 650ms (see Figure 1). Before the first experimental block, a training block of eight trials was given and repeated until participants were familiar and comfortable with the task. Immediately after the primary picture repetition detection task, the surprise word recognition test was administered to all participants. Words were presented, one at a time, on the computer screen, again using DMDX software (Forster & Forster, 2003). Each word remained on the screen until a response (key press) was given. Participants were instructed to press the "B" key if they felt that they had seen the word during the repetition detection task or, instead, the "V" key if they felt that they had not seen the word before. Response buttons were counterbalanced across participants.

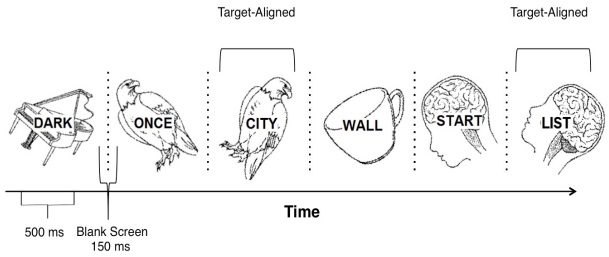


Figure 1: Schematic representation of the task. Immediately repeated pictures serve as the attended task targets while superimposed words are the ignored distractor items. Words appearing with attended task targets are the TA words (i.e., "City" and "List"); all other words are NA words.

Results

Because our interest was focused on determining if recognition rates for TA and NA words differ between young and old adults, statistical analyses include a two-way ANOVA as well as pre-planned t-tests within each age group. These analyses were designed to assess performance both within and across age groups.

Immediate Repetition Accuracy Overall performance accuracy for both younger and older adults on the repetition detection task revealed that participants had an equal number of hits and misses on the primary task, [hit rate: M = 0.50, SE = 0.025, miss rate: M = 0.50, SE = 0.025, $t(63)$ = 0.030, p = 0.488]. Performance on this task was significantly above chance [$t(63) = 16.67$, $p < 0.01$], which was taken to be an indication of the successful detection of picture repetitions on the primary task. A target appears, on average, in one of every 15 trials. Therefore chance is calculated as the probability of obtaining a hit in any given presentation of 15 trials (i.e., 7%).

Older adults accuracy (hit rate: M = 0.37, SE = 0.032) was significantly lower than young adults (hit rate: M = 0.58, SE = 0.031) when detecting picture repetitions on the primary task [$t(62) = 4.64$, $p < 0.001$], indicating that the repetition detection task may have been more difficult for them. There was no significant difference in false alarm (FA) rates for older adults (M = 0.01, SE = 0.003) compared to young adults (M = 0.007, SE = 0.001) [$t(62) = 0.783$, p = 0.218].

Overall Recognition Performance In order to assess whether age modulated overall word recognition, a two-way ANOVA was conducted on surprise recognition test performance with age (young vs. old) and target-alignment (TA vs. NA) as between subject factors, and accuracy as the dependent variable. There was a marginal main effect for age [$F(1, 64) = 3.35$, $p = 0.07$], suggesting that older adults recognized fewer words on the surprise recognition test compared to young adults. There was no main effect for target alignment indicating that overall TA word recognition (M = 0.35, SE = 0.030) was not significantly lower than NA (M = 0.40, SE = 0.026) [$F(1, 64) = 1.75$, $p = 0.190$], and no interaction [$F(1, 64) = 0.495$, $p = 0.485$] (see Figure 2).

Although an interaction was not observed, in order to assess any possible influence of age on later surprise recognition rates pre-planned t-tests were conducted on accuracy performance for each age group.

Young Adult Accuracy Performance Overall word recognition on the surprise recognition test was M = 0.40 (SE = 0.022), which was significantly different from chance [$t(37) = 4.15$, $p < 0.001$]. Recognition for TA words (n = 19, M = 0.37, SE = 0.033) was significantly different from chance [$t(18) = 3.85$, $p < 0.001$], while recognition for NA words (n = 19, M = 0.44, SE = 0.027) was only marginally significantly different from chance [$t(18) = 2.02$, $p = 0.06$]. Recognition for TA words was significantly lower than NA words [$t(36) = 1.71$, $p = 0.04$] (see Figure 3).

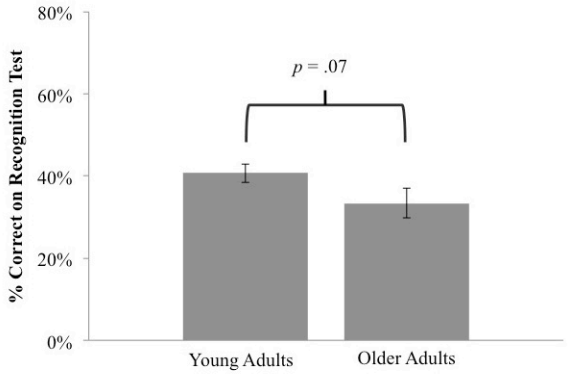


Figure 2: Recognition rates and SE for all word types (TA and NA combined) between young and older adults. The ANOVA revealed a marginal main effect for age (p = 0.07) suggesting older adults recognized fewer words, overall, compared to young adults.

Older Adult Accuracy Performance Overall word recognition on the surprise recognition test was M = 0.33 (SE = 0.036), which was significantly different from chance [$t(25) = 4.64$, $p < 0.001$]. Recognition for TA words (n = 13, M = 0.32, SE = 0.055) was significantly different from chance [$t(12) = 3.18$, $p < 0.008$], and recognition for NA words (n = 13, M = 0.34, SE = 0.047) was also significantly different from chance [$t(12) = 3.29$, $p < 0.007$]. Recognition for TA words was not significantly different from NA words [$t(24) = 0.253$, $p = 0.401$] (see Figure 3).

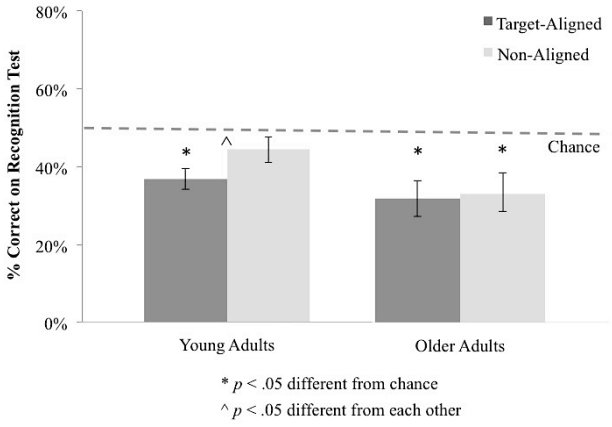


Figure 3: Recognition rates and SE for TA words (dark grey bar) compared to NA words (light grey bar) for young and older adults. Young adults recognized TA words significantly less often than NA words; NA words were recognized around chance levels. Older adults showed no significant difference in recognition rates between word types, however all words were recognized at rates below chance.

Discussion

This experiment directly assessed the ability to inhibit the processing of irrelevant information within an elderly population. In the past, researchers have evaluated inhibitory processes in younger and older adults by measuring rates of distraction from irrelevant items during an attention-demanding task (see Brink & McDowd, 1999; Farkas & Hoyer, 1980; Folk & Lincourt, 1996; Hartley, 1993; Milham et al., 2002; Spieler et al., 1996; Watson & Maylor, 2002). While informative, this approach does not allow one to evaluate the *extent* to which ignored information is processed and subsequently recalled or recognized later. The paradigm employed here allows for such analyses to be made by testing memory of the irrelevant information via a surprise recognition test.

Using this paradigm with young adults, Dewald and colleagues (2011) revealed inhibited processing for TA words, as they were recognized less often than NA words during the surprise recognition test. This finding was replicated in the current study with our young adult sample. Previous research has demonstrated that older adults may have difficulty inhibiting the processing of irrelevant information (Hasher & Zacks, 1988; Kramer, Hahn, & Gopher, 1999; Lustig, Hasher and Tonev, 2006; Madden, Pierce, & Allen, 1996; Mayr, 2001; Plude & Hoyer, 1986). Therefore, it was reasonable to predict that this age group would show a decline in the ability to inhibit irrelevant information on this task as well, resulting in similar recognition rates for TA and NA words. Consistent with our prediction, we found that older adults showed no difference in recognition rates between TA and NA words. This finding suggests that older adults processed all irrelevant words to a similar extent regardless of target-alignment.

Due to a reduction in the ability to execute inhibitory control, we expected that older adults would have overall higher recognition rates during the surprise recognition test compared to the younger adults. Specifically, we expected that older adults would recognize all word types at or around chance levels, while younger adults would demonstrate below chance recognition rates for TA words and at chance recognition for NA words (as seen in Dewald et al., 2011). However, we found that older adults actually recognized significantly fewer words overall, compared to younger adults. Therefore, it appears that older adults inhibited all types of irrelevant information in order to complete the task, rather than just those items appearing with targets, as the young adults seem to do.

It is possible that older adults experienced a reduced ability to *selectively* inhibit word processing while attending to the pictures in the RSVP stream. This reduction in selective inhibitory control may have resulted in a more global inhibition of processing for all presented irrelevant words during the exposure stage, regardless of target alignment. This may be due to the fact that the primary task may have been more difficult for the older adults. Indeed, older adults exhibited overall lower accuracy scores when detecting picture repetitions during the primary attended task. Lavie (2005) has demonstrated that distracting information has less influence on task performance when task difficulty is increased. This is likely due to the fact that increased task difficulty requires additional attentional resources (i.e., increased cognitive load). As a result, processing of irrelevant information is more likely to be inhibited. Therefore, if the primary task was more difficult for the older adults, larger amounts of attentional resources may have been required in order to identify and respond to targets during this portion of the experimental session. Therefore, rather than selectively filter the most intrusive irrelevant information (i.e., TA words), as young adults appear to do, older adults seem to employ a more global inhibitory control leading to more extensive filtering of all irrelevant information.

Alternatively, the observed lower recognition rates for the older adults may be due to a reduced ability to attend to the words during the primary task. This may also be reflected by older adults' lower performance in the repetition task when compared younger adults. Thus, it could be argued that lower recognition rates in the surprise task may be due to a lack of attentional allocation rather than a decline in selective inhibitory control. However, we would expect older adults to exhibit chance performance on the surprise recognition test if this were the case, which would suggest that the information failed to be processed rather than undergoing active inhibition.

Taken together, the findings tentatively demonstrate that as we progress in old age, inhibitory control may diminish, resulting in an inability to execute selective inhibition over irrelevant information presented in attention-demanding tasks. Additional research is necessary in order to fully understand how these mechanisms may operate in old age. Future studies will investigate this further by systematically increasing task difficulty, through faster presentation rates, in a young adult population. We predict that increased presentation speed of the RVSP stream will result in reduced performance on the primary task (as seen in older adults here). If young adults continue to show preferential inhibition for TA words under these circumstances, this may provide additional support for a decline in selective inhibitory control in an older adult population.

References

Banich, M., Milham, M., Cohen, N. J., Wszalek, T., Kramer, A., Liang, Z.-P., Gullet, D., Shah, C., & Brown, C. (2000a). Prefrontal regions play a predominant role in imposing an attentional ''set'': Evidence from fMRI. *Cognitive Brain Research*, 10, 1–9.

Banich, M., Milham, M., Atchley, R., Cohen, N.J., Wszalek, T., Kramer, A., Liang, Z. P., Gullet, D., Shah, C., & Brown, C. (2000b). FMRI studies of Stroop tasks reveal unique roles of anterior and posterior brain system in attentional selection. *Journal of Cognitive Neuroscience*, 12, 988–1000.

Brink, J. M., & McDowd, J. M. (1999). Aging and selective attention: an issue of complexity or multiple mechanisms?. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54*(1), P30-P33.

Campbell, K. L., Grady, C. L., Ng, C., & Hasher, L. (2012). Age differences in the frontoparietal cognitive control network: implications for distractibility. *Neuropsychologia*, *50*(9), 2212-2223.

Dewald, A. D., & Sinnett, S. (2012). Enhanced Performance for Recognition of Irrelevant Target-Aligned Auditory Stimuli: Unimodal and Cross-modal Considerations. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society* (pp. 294-299).

Dewald, A. D., & Sinnett, S. (2013). Speed Facilitation In The Absence Of Enhanced Recognition For Target-Aligned But Irrelevant Stimuli Under Cross-modal Presentations. In *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society* (pp. 2183-2188).

Dewald, A. D., Sinnett, S., & Doumas, L. A. (2011). Conditions of directed attention inhibit recognition performance for explicitly presented target-aligned irrelevant stimuli. *Acta psychologica*, *138*(1), 60-67.

Farkas, M. S., & Hoyer, W. J. (1980). Processing consequences of perceptual grouping in selective attention. *Journal of Gerontology*, *35*(2), 207-216.

Folk, C. L., & Lincourt, A. E. (1996). The effects of age on guided conjunction search. *Experimental Aging Research*, *22*(1), 99-118.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116-124.

Hartley, A. A. (1993). Evidence for the selective preservation of spatial selective attention in old age. *Psychology and Aging*, *8*(3), 371.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 22, pp. 193-225). New York: Academic Press.

Kramer, A. F., Hahn, S., & Gopher, D. (1999). Task coordination and aging: Explorations of executive control processes in the task switching paradigm. *Acta Psychologica*, *101*(2), 339-378.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences*, 9(2), 75-82.

Lustig, C., Hasher, L., & Tonev, S. T. (2006). Distraction as a determinant of processing speed. *Psychonomic bulletin & review*, *13*(4), 619-625.

MacDonald, A. W., III, Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science, 288*(5472); 1835–1838.

Madden, D. J., Pierce, T. W., & Allen, P. A. (1996). Adult age differences in the use of distractor homogeneity during visual search. *Psychology and aging*, *11*(3), 454.

Mayr, U. (2001). Age differences in the selection of mental sets: the role of inhibition, stimulus ambiguity, and response-set overlap. *Psychology and aging*, *16*(1), 96.

Milham, M. P., Erickson, K. I., Banich, M. T., Kramer, A. F., Webb, A., Wszalek, T., & Cohen, N. J. (2002). Attentional control in the aging brain: insights from an fMRI study of the stroop task. *Brain and cognition*, *49*(3), 277-296.

Pashler, H. E. (1998). *The psychology of attention* (Vol. 15). Cambridge, MA: MIT press.

Plude, D. J., & Hoyer, W. J. (1986). Age and the selectivity of visual information processing. *Psychology and Aging*, *1*(1), 4.

Rabbitt, P. (1965). An age-decrement in the ability to ignore irrelevant information. *Journal of Gerontology*, *20*(2), 233-238.

Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science*, *286*(5449), 2504-2507.

Seitz, A. R. & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature*, 422, 36.

Sinnett, S., Snyder, J. J., & Kingstone, A. (2009). Role of the lateral prefrontal cortex in visual object-based selective attention. *Experimental brain research*, 194(2), 191-196.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 461.

Störmer, V. S., Li, S. C., Heekeren, H. R., & Lindenberger, U. (2013). Normal aging delays and compromises early multifocal visual attention during object tracking. *Journal of cognitive neuroscience*, *25*(2), 188-202.

Tipper, S. P. (1992). Selection for action: The role of inhibitory mechanisms. *Current Directions in Psychological Science*, *1*(3), 105-109.

Walker, M., Dewald, A. D., & Sinnett, S. (2014). The role of modality congruence in the presentation and recognition of task-irrelevant stimuli in dual task paradigms. In Bello P., Guarini M., McShane M. & Scassellati B. (Eds.) *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1736-1741). Austin TX: Cognitive Science Society.

Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature*, *413*(6858), 844-848.

Watson, D. G., & Humphreys, G. W. (1997). Visual marking: prioritizing selection for new objects by top-down attentional inhibition of old objects. *Psychological review*, *104*(1), 90.

Watson, D. G., & Humphreys, G. W. (1998). Visual marking of moving objects: A role for top-down feature-based inhibition in selection. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 946.

Watson, D. G., & Maylor, E. A. (2002). Aging and visual marking: Selective deficits for moving stimuli. *Psychology and Aging*, *17*(2), 321.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6-10.

World Health Organization (2014). *Definition of an older or elderly person.* United Kingdom: WHO press.

Wright, R. E. (1981). Aging, divided attention, and processing capacity. *Journal of Gerontology*, *36*(5), 605-614.

Analytic Eye Movement Patterns in Face Recognition are Associated with Better Performance and more Top-down Control of Visual Attention: an fMRI Study

Cynthia Y.H. Chan (cynchan@hku.hk)
Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong

J.J. Wong (jjwong@hku.hk)
Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong

Antoni B. Chan (abchan@cityu.edu.hk)
Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong

Tatia M.C. Lee (tmclee@hku.hk)
Laboratory of Neuropsychology, University of Hong Kong, Pokfulam Road, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong

Abstract

Recent research has revealed two different eye movement patterns during face recognition: holistic and analytic. The present study investigated the neural correlates of these two patterns through functional magnetic resonance imaging (fMRI). A more holistic pattern was associated with more activation in the face-selective perceptual areas, including the occipital face area and fusiform face area. In contrast, participants using a more analytic pattern demonstrated more activation in areas important for top-down control of visual attention, including the frontal eye field and intraparietal sulcus. In addition, participants using the analytic patterns had better recognition performance than those showing holistic patterns. These results suggest that analytic eye movement patterns are associated with more engagement of top-down control of visual attention, which may consequently enhance recognition performance.

Keywords: eye movement; functional magnetic resonance imaging (fMRI); face recognition; analytic patterns; Hidden Markov Model (HMM); top-down visual attention.

Introduction

Studies have revealed distinct eye movement patterns in face processing. For example, compared with young adults, older adults fixate more on the lower half of a face (Wong, Cronin-Golomb, & Neargarder, 2005). Some differences seem to be culturally influenced: Caucasians demonstrate more of an analytic pattern, focusing on characteristics of face parts, whereas Asians process faces with a more holistic pattern, with most fixations landed around the face center (i.e., the nose; Kelly et al., 2011). However, it remains unclear whether people adopting different eye movement patterns process faces differently. Findings on the relationship between eye movements and face recognition performance are still mixed. For example, Kelly et al., (2011) found no performance differences between Asians and Caucasians who adopted different eye movement patterns. Mehoudar et al. (2014) reported that individual eye movements were not predictive of performance in face recognition. In contrast, Henderson, William, and Falk (2005) showed that eye movements facilitate face learning since restricting eye movements during face learning was found to impair recognition performance. Goldinger, He, and Papesh (2009) showed that during face learning, eye movements in trials that eventually led to a miss during recognition were more suppressed, with fewer regions visited, shorter scanning distance, and fewer fixations, as compared with those that eventually led to a hit.

This inconsistency may be due to limitations in eye movement data analysis methods for discovering common patterns from individuals. In view of this, Chuk, Chan and Hsiao (2014a) proposed a Hidden Markov Model (HMM) based approach to analyze eye movement data; the model takes individual differences in both spatial and temporal dimensions of eye movements into account. Through clustering participants' eye movement patterns according to their similarities, both holistic and analytic patterns in face recognition were discovered in Asians (Chan, Chan, Lee, & Hsiao, 2015; Cheng, Chuk, Hayward, Chan, & Hsiao, 2015; Chuk et al., 2014a) as well as in Caucasians (Chuk et al., 2014b). In addition, it was found that people who adopted analytic patterns had better recognition performance than those with holistic patterns (Chuk et al., 2014b; Chan et al., 2015). These results demonstrate that the HMM based approach is a powerful analysis tool for discovering common eye movement patterns and their relationship with recognition performance.

Nevertheless, it remains unclear why analytic eye movement patterns lead to better face recognition performance than holistic patterns. Since analytic patterns involve more spread out regions of interest (ROIs) targeting at facial features, and more transitions among them, people with analytic patterns may engage more top-down control of visual attention, which in turn enhances recognition performance.

Thus, here we aim to investigate the neural correlates of the observed association between eye movement patterns and face recognition performance via functional magnetic resonance imaging (fMRI).

In face recognition, neuroimaging research identified two face-selective perceptual areas, including the occipital face area (OFA) in the inferior occipital gyri and the fusiform face area (FFA) in the lateral fusiform gyrus. OFA is the entry point of the face network and is sensitive to facial features. FFA is related to invariant aspects of face perception such as face identity and is sensitive to the configuration and spacing between facial features (Hoffman & Haxby, 2000; Pitcher, Walsh, Yovel, & Duchaine, 2007). When a task was switched from face to object recognition, functional connectivity decreased significantly between the OFA and FFA (Zhen, Fang, and Liu, 2013), suggesting an important role of these two areas in face recognition.

Brain regions beyond the face-selective network have also been identified, namely the primary visual cortex (V1), frontal eye field (FEF), intraparietal sulcus (IPS), posterior cingulate cortex (PCC), and prefrontal cortex (PFC; Zhen et al, 2013; Leube et al., 2003; Phillips et al., 1998). Some of these regions are greatly engaged in top-down control of visual attention in cognitive tasks in general (Gilbert & Li, 2013; Noudoost, Chang, Steinmetz & Moore, 2010; Leech & Sharp, 2014). For example, the FEF, IPS and PCC are related to saccadic eye movement (Schall, 2004), attention shift (Corbetta et al., 1998) and internally directed attention (Leech & Sharp, 2014) respectively. The PCC was also found to play an essential role in facilitating or monitoring working memory loaded tasks (Hampson, Driessen, Skudlarski, Gore, & Constable, 2006), and early top-down control of visual attention has been shown to be advantageous to performance on working memory and perceptual tasks (Rutman, Clapp, Chadick, & Gazzaley, 2010). The prefrontal cortex (PFC), especially the dorsolateral prefrontal cortex (DLPFC), plays an important role in memory (Braver et al., 1997; Kane & Engle, 2002), goal maintenance (Colvin, Dunbar, & Grafman, 2001), planning and execution (Fincham, Carter, van Veen, Stenger, & Anderson, 2002), and modulates other brain regions by conveying top-down signals in controlling visual attention.

Accordingly, we hypothesized that analytic eye movement patterns in face recognition may be associated with more top-down control of visual attention, as reflected in higher activations in the IPS, FEF, PCC, and DLPFC. To test this hypothesis, here we examined participants' eye movement patterns and brain activations in face recognition and examine the association between them.

Method

Participants

A total of 20 Chinese participants (aged 18 – 24; M = 21.7; SD = 2.36; 11 females) were recruited from the University of Hong Kong. All participants were right-handed. They either had normal or corrected visual ability. Informed consent was collected from each participant; the research protocol was approved by the Ethics Review Board at The University of Hong Kong.

Materials

A total of 120 young Chinese face images (60 females) with neutral facial expressions were used and split into two sets of images. Sixty of them (30 in each set) were used as target faces to be remembered and the other 60 (30 in each set) were used as distractors in the test phase. All face images were 270 × 360 pixels (8 visual degrees) in grayscale, and were adjusted to match for luminosity, contrast, and quality. The distances between the eyes and the mouth were standardized. Each face was a frontal view and was cropped according to the original shape of the face such that hair, ears and the neck were removed, leaving only the face visible.

Experimental Design

Apart from the face recognition task, a verbal and spatial working memory test and an executive functioning test were administered in order to examine possible correlations between eye movement patterns and cognitive abilities.

Face Recognition task. The task included three runs, each of which consisted of a study and a test phase. In each study phase, participants were shown 10 face images one at a time at a rate of 3s each. The face was displayed either at the upper center or at the lower center (see Figure 1) of the screen randomly. Participants were instructed to fixate at a cross "+" located at the center of the screen between the presentation of each image. The stimulus onset asynchrony (SOA) between images varied from 6 to 16s. Participants were instructed to remember all the faces in the study phase. In each test phase, participants were shown 20 face images individually, which consisted of 10 old images and 10 new images. They were asked to judge whether they recognized the displayed images from the study phase previously, within 3s per face. Individual A-prime was computed.

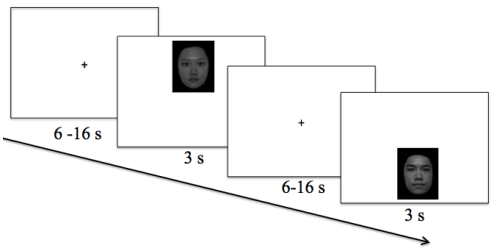


Figure 1. Display of images in the study and test phases.

Working memory test. Both verbal and spatial working memory abilities were measured via a computerized two-back test (Lau, Ip, Lee, Yeung, & Eskes, 2013). For verbal two-back, a sequence of 100 single digit numbers was displayed at the screen center. Starting from the third number, participants had to judge whether each displayed number was identical to the one displayed two items ago. For the spatial two-back task, a sequence of 100 symbols was displayed at several different locations around the screen center. Starting from the third symbol, participants had to judge

whether each displayed symbol was in the same position as the one displayed two items ago. Accuracy (%) was calculated for verbal and spatial tasks separately.

Executive functioning test. Executive functioning, specifically planning and execution ability, was measured via a computerized Tower of London (TOL) test (e.g., Phillips, Wynn, McPherson, & Gilhooly, 2001). Three beads and three pegs of different heights were shown on screen. Participants had to move the beads from peg to peg one at time to match the bead positions on a target board while adhering to certain rules. Participants were told to complete it as quickly as possible using the fewest number of moves. The planning and execution ability was measured by performance scores (a total of 10 trials; 1 mark for every trial finished with minimal moves) and number of extra moves (actual minus minimal moves). .

Procedure

The study consisted of three parts: an eye tracking experiment, a set of cognitive tests, and an fMRI experiment. In the eye tracking experiment, participants sat in front of a computer screen with their head on a chinrest and their eye level adjusted via a chinrest to approximately the mid-level of the screen. Prior to the experiment, 9-point calibration was conducted repeatedly until the errors were lower than 0.3° and 0.5° visual degree for the dominant eye and the non-dominant eye respectively. Then, they performed the face recognition task while their eye movement was recorded. After the eye tracking experiment, participants performed cognitive tests including the verbal and spatial two-back, and TOL tasks, followed by the fMRI experiment. Throughout the whole fMRI experiment, participants had to perform the same face recognition task during the eye tracking experiment but with a different set of images.

Eye Data Acquisition, Processing and Analysis

Eye movements were recorded with an SMI REDn eye-tracking system (60 Hz), which was connected to a 17" monitor with screen resolution 1280 × 768 pixels (with a 60 cm eye-monitor distance). Both eyes were tracked with the 'Smart Binocular' tracking mode; only data from the dominant eye was used in analysis. We analyzed participants' eye movement data in the test (recognition) phase.

A Hidden Markov Model (HMM) based approach was used to analyze eye movement data. First, each participant's eye movement data was modeled by an HMM with a variational Bayesian approach. Each HMM included three Gaussian components (i.e. ROIs), as indicated by different colors in Figure 2. The overlapping area of two or more ROIs indicates that the fixations around that area have similar probabilities of belonging to those ROIs. The prior values in the matrices represent the probability that an initial eye fixation is located at each of the ROIs. The rest of the matrix represents the transition probabilities among the three ROIs. Next, we applied a variational hierarchical EM algorithm (VHEM; Coviello, Chan, & Lanckriet, 2014) to cluster the individual HMMs into two subgroups according to their similarities. According to our previous studies (e.g., Chuk et al., 2014), a holistic pattern and an analytic pattern would be shown. Finally, for each individual HMM, we calculated the difference between the log-likelihood of being classified as the holistic pattern and the log-likelihood of being classified as the analytic pattern, to represent the degree of similarity of one's eye movements to the two patterns (named "H-A Scale" in later sections).

Image Data Acquisition, Processing and Analysis

Imaging was performed on a 3-Tesla Phillips scanner head scanner with a standard eight-channel head coil. A total of 580 volumes (190 × 3 runs) of functional data were collected as echo-planar images (128 × 128 matrix; 40 slices with in-plane resolution of 3 × 3 mm^2 and slice thickness 3.5mm; TE 30 ms; TR 2000 ms; FOV 230 × 230 mm^2; flip angle 90°). A T1-weighted spin-echo pulse sequence with a spatial resolution of 1 × 1 × 1 mm^3 data set was collected for anatomical reference.

The FSL software package was used for image preprocessing. Functional images were corrected for head motion with FSL's intra-modal motion correction tool (MCFLIRT), high-pass filtered at 100s and spatially smoothed with a Guassian filter (full-width-half-maximum = 5mm). T1 anatomical images were co-registered at participant level and to a standard brain Montreal Neurological Institute template (MNI152; 2 mm). A general linear model was used to analyze functional imaging data incorporating predictors, which corresponded to the particular experimental conditions. The onset times were set as the stimulus onset and the durations were set as 1s for all predictors.

To examine the neural correlates of different eye movement patterns, we examined the correlations between participants' H-A Scales and the event contrasts of non-missed test trials (i.e. trials with either correct or incorrect responses) against the fixation baseline. Significant signal changes were reported if they exceeded a p-value of 0.05 corrected at a whole-brain level.

Result

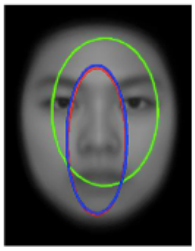
Cognitive Test and Behavioral Performances

The average face recognition performance in A' was .85 (SD = .07). The average number of fixations per trial was 2.56 (SD = .85) and the average reaction time was 1.64s (SD = .85s). The mean accuracy of the verbal 2-back task was 76.6% (SD = 18.0%) and that of the spatial 2-back task was 77.5% (SD = 11.5%). The average correct score and number of extra moves in the TOL test were 5.3 (SD = 2.6) and 24.3 (SD = 16) respectively.

Eye Movement Pattern Analysis

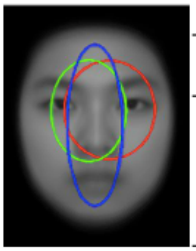
Participants' eye movement patterns were clustered into two groups via the HMM-based approach, as shown in Figure 2. Each group consisted of 10 participants. H-A Scale was computed, with a positive value representing a holistic pattern (M = .04; SD = .56; ranging from -.65 to 1.31).

In Figure 2, the top figure shows a typical holistic eye movement pattern. The ROIs were around the face center. The first fixation was located most commonly at the vertical and center of a face (red; prob. = .54). The subsequent eye fixations usually stayed at the same position (red to red = .56), although sometimes they also transited to the surrounding region (red to green = .24; blue to green = .23). In contrast, the bottom figure shows a typical analytic eye movement pattern. The ROIs were relatively more clearly located at different facial features, the left eye (green), the right eye (red), and the center/mouth (blue) of a face. The first fixation was most commonly located near the left eye (green; prob. = .56). Most eye gaze transitions appeared between the two eyes (red to green = .36; green to red = .65, see e.g. Chuk et al., 2014a).



Holistic Strategy's Representative HMM

	To red	To green	To blue
Prior values	.56	.09	.35
From red	.59	.24	.17
From green	.23	.17	.60
From blue	.62	.23	.15



Analytic Strategy's Representative HMM

	To red	To green	To blue
Prior values	.05	.56	.39
From red	.37	.36	.27
From green	.65	.06	.29
From blue	.50	.35	.15

Figure 2. The representative HMMs of the two subgroups, holistic (top) and analytic (bottom) patterns, and the corresponding transition probability matrices.

Relationship between Eye Movement Pattern and Recognition Performance/Cognitive Ability

With regards to the relationship between eye movement pattern and face recognition performance, a negative correlation was found between the H-A Scale and face recognition A-prime scores, r = -.47, p = .036 (Figure 3). People who used a more analytic (holistic) strategy yielded a better (worse) face recognition performance. This finding is consistent with previous studies (e.g., Chuk et al., 2014b; Chan et al., 2015; Cheng et al., 2015). H-A Scale was neither correlated with average number of fixations per trial (r = .05, n.s.) nor reaction time (r = .30, n.s.). Thus, participants with holistic patterns did not necessarily make fewer fixations or had shorter viewing times on the faces. When examining the relationship between eye movement patterns and cognitive abilities, we found that spatial working memory performance was marginally correlated with the H-A Scale, r = -.41, p = .073, indicating that people who possessed higher (lower) spatial working memory tended to employ a more analytic (holistic) eye movement pattern. However, the results were not significant for verbal working memory (r = -.08, n.s), and executive functioning (scores: r = -.26, n.s.; extra moves: r = .15, n.s.).

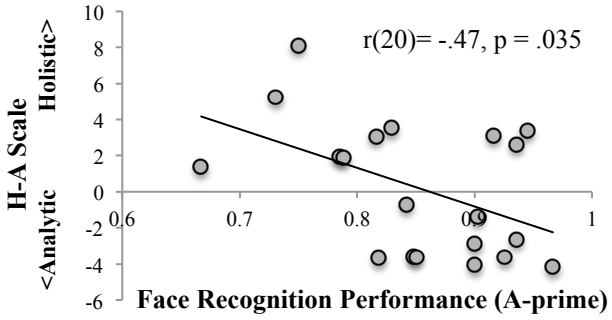


Figure 3. Correlation of H-A Scale and face recognition performance.

Relationship between Strategy and Neural Activity

Correlation analysis between participants' H-A Scales and the event contrast of non-missed test trials against fixation baseline was conducted for all 20 participants. It was found that the more holistic the eye movement patterns, the higher the brain activation in some face-selective perceptual areas (Figure 4), including the left OFA ([-34, -76, -10]; $z > 2.3$, $p < .05$) and the right FFA (BA 37; [42, -55, -18]; $z > 2.3$, $p < .05$). Higher activation was also found in the left pars triangularis of the inferior frontal gyrus (BA 45; [-55, 26, 6]; $z > 2.3$, $p < .05$).

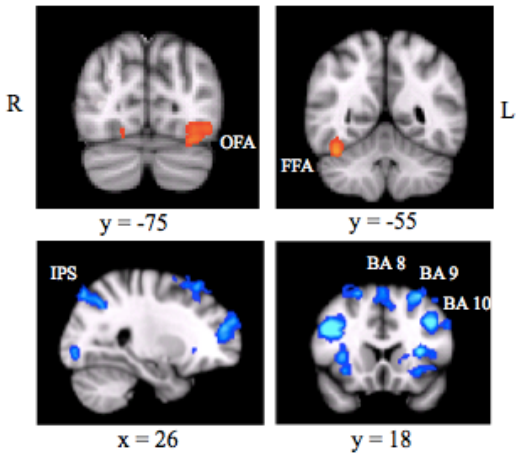


Figure 4. Holistic patterns: more activation in the left OFA (top left) and the right FFA (top right). Analytic patterns: more activation in bilateral IPS (bottom left) and the prefrontal cortex (bottom right)

When people had more analytic eye movement patterns, there was significantly higher activation in the V1 ([±14, -92, 4]; $z > 2.3$, $p < .05$), eye movement planning and attention control related areas including the IPS ([±30, -64, 46]; $z > 2.3$, $p < .05$), PCC ([±2, -30, 40]; $z > 2.3$, $p < .05$), and FEF (encompassed in DLPFC; BA 8; [±30, 18, 58]; $z > 2.3$, $p < .05$), and executive functioning related areas including the DLPFC (BA 9; [±46, 44, 30]; $z > 2.3$, $p < .05$) and the

frontalpolar prefrontal cortex (FPPFC; BA 10; [±24, 52, 18], $z > 2.3$, $p < .05$).

Discussion

The current study investigated the neural correlates of different eye movement patterns in face recognition. Through the HMM-based approach for eye movement data analysis, we identified holistic and analytic eye movement patterns in the participants (e.g. Chuk et al., 2014a). By calculating participants' H-A Scale according to the likelihood of each individual's eye movement being classified as a holistic or an analytic pattern, our neuroimaging results demonstrated that a more holistic pattern is associated with increased activation in face-selective perceptual areas OFA and FFA. In contrast, a more analytic pattern is associated with increased activation in areas related to top-down control of visual attention, including FEF, IFS, and PCC. In addition, consistent with previous studies, we found that people using analytic patterns recognized faces better than those with holistic patterns (Chan et al., 2015; Chuk et al., 2014a).

The regions whose activations are associated with analytic eye movement patterns included the V1, IFS, FEF, PCC, DLPFC (BA 9), and FPPFC (BA 10). The V1 is one core region of visual attention. The DLPFC (BA 9) is involved in planning (Fincham et al., 2002) and delivering top-down signals to other brain regions. The FEF was previously reported to be active during voluntary saccades of eye movement (Schall, 2004). The PCC was reported to monitor visual attention shift during post-saccadic eye movement (Olson, Musil & Goldberg, 1993) and resetting the visual attention plan in anticipating the next attentional shift (Small et al., 2003). The Right IPS is shown to be involved in serial attention processes and the left IPS acts as an attentional modulator to maintain the activities of initial perception and processing of selective information in the working memory as well as long-term memory network (Majerus et al., 2007; Noudoost et al., 2010). Taken together, the elevated activation of these regions suggests that people adopting analytic patterns may be engaging active eye movement planning and top-down control of visual attention during face recognition. This is consistent with the behavioral finding that analytic patterns involved more spread out ROIs and transitions among ROIs than holistic patterns. Since top-down visual attention allows us to filter out irrelevant information and improve the efficiency of information processing by selecting useful visual information, it has been found to be associated with better cognitive task performance (Rutman, Clapp, Chadick, & Gazzaley, 2010). Thus, it is possible that participants adopting analytic eye movement patterns were engaging more top-down control of visual attention, and consequently had increased accuracy in face recognition.

On the other hand, some face-selective perceptual areas, including the left OFA and right FFA, were found to be more active when people adopted holistic patterns. The OFA plays a major role in face part processing while the FFA is concerned with the processing of face configurations (e.g. Pitcher et al., 2007). Thus, our finding suggested that people who adopted holistic eye movement patterns may have relied more on perceptual areas for face recognition with less engagement of eye movement planning and visual attention control. This is consistent with their eye movement behavior, which focused on the face center and lacked transitions among facial features.

In a previous study, Chan et al. (2015) showed that in face recognition, more old adults adopted holistic eye movement patterns, and more young adults exhibited analytic patterns. In addition, they found that the likelihood of old adults' eye movements being classified as holistic was negatively correlated their cognitive status: the higher the likelihood, the lower the cognitive ability. Our current results suggest that this effect may be due to ageing-related cognitive decline in old adults, affecting their ability to actively engaging top-down control of visual attention. Our results also suggest that the adoption of different eye movement patterns may be related to limitations of cognitive capacity, rather than merely a matter of preference or motivation.

Note that the term "holistic" here refers to the specific eye movement pattern described here, and is different from the term "holistic processing" frequently described in the face perception literature (e.g., Taubert et al, 2011). Thus, having a holistic eye movement pattern does not necessarily leads to more engagement of holistic face processing. Indeed, although here we found holistic eye movement patterns are associated with high activations in the OFA and FFA, recent research has suggested both holistic and featural representations may be found in these areas (e.g., Goffaux, Schiltz, Mur, & Goebel, 2013). Using the Expanding Spotlight technique, Miellet et al. (2013) showed that Westerners, who exhibited analytic eye movement patterns, used local high-spatial-frequency information of facial features for recognition, whereas Easterners, who had holistic eye movement patterns, used global low-spatial frequency information. Future work will examine whether similar results can be obtained between the analytic and holistic patterns identified here through the HMM approach.

In summary, here we showed that holistic eye movement patterns in face recognition are associated with increased activation in the face-selective perceptual areas, whereas analytic patterns are associated with increased activation in areas involved in top-down control of visual attention. As previous studies have shown that analytic patterns lead to better performance in face recognition, the current results suggest that this advantage in performance may be a consequence of more engagement of active eye movement planning and visual attention control during the task. This finding thus provides strong evidence for a close relationship between eye movement patterns and cognitive performance.

Acknowledgments

We thank Mr. Tim Chuk for his technical support in using the HMM-based model, Dr. Bolton Chau for his advise on fMRI data analysis, and Ms. Kit Yan Pang for data collection. We are also grateful to the Research Grant Council of Hong Kong (project code 17402814 to J.H. Hsiao and

CityU 110513 to A.B. Chan), and the University of Hong Kong May Endowed Professorship in Neuropsychology to Professor Tatia M.C. Lee.

References

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage, 5*, 49-62.

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2015). Eye movement pattern in face recognition is associated with cognitive decline in the elderly. In Noelle, D. C., et al. (Eds.), *the Proceeding of the 37th Annual Conference of the Cognitive Science Society* (pp. 321-326). Austin, TX: Cognitive Science Society.

Cheng, Z., Chuk, T., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2015). Global and local priming evoke different face processing strategies: evidence from an eye movement study. *J. Vis., 15*, 154.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014a) Understanding eye movements in face recognition using hidden Markov models. *J. Vis., 14*, 8.

Chuk, T., Luo, X., Crookes, K., Hayward, W., Chan, A. B., & Hsiao, J. H. (2014b). Caucasian and Asian eye movement patterns in face recognition: A computational exploration using hidden Markov models. *J. Vision, 14*, 1212.

Colvin, M. K., Dunbar, K., & Grafman, J. (2001). The effects of frontal lobe lesions on goal achievement in the water jug task. *J. Cogn. Neurosci., 13*, 1129-1147.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., ... & Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*, 761-773.

Coviello, E., Chan, A. B., & Lanckriet, G. R. G. (2014). Clustering hidden Markov models with variational HEM. *J. Mach. Learn. Res.*, *15*, 697-7

Fincham, J. M., Carter, C. S., van Veen, V., Stenger, V. A., & Anderson, J. R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proc. Natl. Acad. Sci., 9*9, 3346-3351.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nat. Rev. Neurosci., 14*, 350-363.

Goffaux, V., Schiltz, C., Mur, M., & Goebel, R. (2013). Local discriminability determines the strength of holistic processing for faces in the fusiform face area. *Front. Psychol.*, *3*, 604.

Hampson, M., Driesen, N. R., Skudlarski, P., Gore, J. C., & Constable, R. T. (2006). Brain connectivity related to working memory performance. *J. Neurosci, 26*, 13338-13343.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Mem. Cognition, 33*, 98-106.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat. Neurosci., 3*, 80-84.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychon. Bull. Rev., 9*, 637-671.

Kelly, D. J., Jack, R. E., Miellet, S., De Luca, E., Foreman, K., & Caldara, R. (2011). Social experience does not abolish cultural diversity in eye movements. *Front. Psycho, 2,* 1-11.

Lau, E. Y. Y., Ip, M. S. M., Lee, T. M. C., Yeung, A. W. M., Eskes, G. A. (2013) Neurocognitive and psychosocial outcomes of obstructive sleep apnea in Hong Kong Chinese: Similar to or different from western populations? *J. Sleep. Disord. Treat Care,* 2:3.

Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain, 137*, 12-32.

Majerus, S., Bastin, C., Poncelet, M., Van der Linden, M., Salmon, E., Collette, F., & Maquet, P. (2007). Short-term memory and the left intraparietal sulcus: focus of attention? Further evidence from a face short-term memory paradigm. *Neuroimage, 35*, 353-367.

Miellet, S., Vizioli, L., He, L., Zhou, X., & Caldara, R. (2013). Mapping face recognition information use across cultures. *Frontiers in Perception Science*, *4*, 34.

Noudoost, B., Chang, M. H., Steinmetz, N. A., & Moore, T. (2010). Top-down control of visual attention. *Curr. Opin Neurobiol, 20,* 183-190.

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London task. *Q. J. Exp. Psychol. - A, 54*, 579–597.

Phillips, M. L., Bullmore, E. T., Howard, R., Woodruff, P. W., Wright, I. C., Williams, S. C., ... & David, A. S. (1998). Investigation of facial recognition memory and happy and sad facial expression perception: an fMRI study. *Psychiatry Res-Neuroim, 83*, 127-138.

Pitcher, D., Walsh, V., Yovel, G., Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Curr. Biol., 17*, 1568–1573

Rutman, A. M., Clapp, W. C., Chadick, J. Z., & Gazzaley, A. (2010). Early top–down control of visual processing predicts working memory performance. *J. Cogn. Neurosci., 22*, 1224-1234.

Schall, J. D. (2004). On the role of frontal eye field in guiding attention and saccades. *Vision Res., 44*, 1453-1467.

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vis. Res., 51*, 1273-1278.

Wong, B., Cronin-Golomb, A., & Neargarder, S. (2005). Patterns of visual scanning as predictors of emotion identification in normal aging. *Neuropsychol., 19,* 739–749.

Zhen, Z., Fang, H., & Liu, J. (2013). The hierarchical brain network for face recognition. *PloS one, 8*, e59886.

Establish Trust and Express Attitude for a Non-Humanoid Robot

Mei Si (sim@rpi.edu)
Department of Cognitive Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Joseph Dean McDaniel (mcdanj2@rpi.edu)
Department of Computer Science, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

In recent years, there has been an increasing interest in designing social robots to interact with people to provide therapy and companionship. Most social robots currently being used are light-weight and much smaller in size compared to people. In this work, we investigate designing interactions for larger and more physically capable robots as they have more potential to assist people physically. A modified version of Baxter robot was used, by sitting Baxter on top of an electronic wheelchair. Two experiments were designed for studying the role of facial expressions and body movements in establishing trust with the user and for expressing attitudes. Our results suggest that the robot is capable of expressing fine and distinguishable attitudes (proud vs. relaxed) using its body language, and the coupling between body movements and speech is essential for the robot to be viewed as a person.

Keywords: Robot Human Interaction; Gesture; Trust

Introduction

In recent years, there has been an increasing interest in designing social robots to interact with people, as a tutor or a companion. For example, Sony's AIBO robotic dog has been used for improving autistic children's play, reasoning, and affective skills (François, Powell, & Dautenhahn, 2009). Autom is a humanoid robot, which is capable of establishing eye contact with the user and making small talk. It was used for helping people keep track of their weight loss history (Kidd & Breazeal, 2008). Nao and Hanson robots have also been widely used for research, tutoring, and entertainment purposes.

Most social robots currently being used are light-weight and much smaller in size compared to people. This makes them naturally look non-threatening. On the other hand, they have limited movability and physical strength, which limits their potential for physically interacting with and assisting people. In this work, we want to investigate designing interactions for robots, which are larger and more physically capable. We start this exploration using a modified version of Baxter, a dual-arm robot by Rethink Robotics. The robot by itself is 3'1" (93.98 cm) in height. With modification, the robot sits on top of an electronic wheelchair. Its actual height is similar to an adult sitting on a chair. The robot's arm can reach 41" (104 cm). Its body looks sturdy, with a weight of 165 lbs (75 kg).

To enable long-term natural interaction between robots and users, we strive to not only rely on safe, dependable robotic movements but also leverage existing research studying human-computer interaction and human social interaction to create a positive experience with trust and comfort. Furthermore, we want to have the capacity of creating different "personalities for the robot. Here we do not necessarily need to build an active or intriguing character, but rather to present meaningful and consistent behavior patterns that the user can infer from the robots actions for enhancing the believability of the robot as a human-like character. Breazeal studied the requirements to "promote the illusion of a socially aware robotic creature and found that to socially engage a human, its behavior must address issues of believability such as conveying intentionality, promoting empathy, being expressive, and displaying enough variability to appear unscripted while remaining consistent" (Breazeal, 2000). Similar arguments have been given for creating digital companions (Bickmore & Rosalind, 2005), and for assistant robot. It has been found that even though assistant robot does not need to represent a social character, having a "personality helps the user to understand and predict its behaviors (Severinson-Eklundh, Kerstin and Green, Anders and Httenrauch, Helge , 2003).

Both facial expressions and non-verbal behaviors play important roles in social interactions. In particular, it has been shown that for human-like robots, showing appropriate non-verbal behaviors such as gaze, head nod, and gestures can improve people's performance in a collaborative task because people can better understand the robot's intention (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005). Many robots used by rescue teams, law enforcement, and the military are designed with functionality as a higher priority, and thus, they do not appear humanlike and have limited means to support natural human-robot interaction. While these robots are often modified to have a facial display to express affect, researchers have found that their body movements may play a more important role (Bethel & Murphy, 2008). Our modified Baxter robot belongs partially to this category. We face the challenge of the robot not having an exact human-like shape. Because of the robot's physical strength and size, potential negative misinterpretations of the robot's intentions will become particularly problematic.

In this work, we designed two experiments for studying the expression of attitudes using the robot's body movements. In the first experiment, we examined how using human-informed gestures and social dialogue, and having a digital face can make people feel more comfortable to interact with the robot. In the second experiment, we further investigated expressing two different attitudes relaxed and proud using

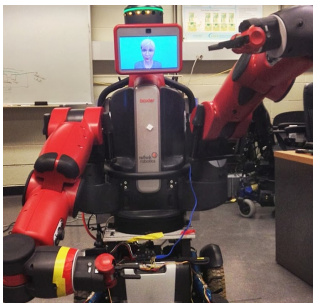


Figure 1: Baxter with Virtual Human Face

the robot's body movements.

Experiment 1

In this experiment, we examined how placing Baxter's movements into a social context with its dialogue affects people's perceptions of its safety and friendliness. We studied the importance of having meaningful gestures and a face display. In particular, it is natural that people feel more comfortable interacting with a novel device after seeing how it operates or moves. To study whether there is a difference between this phenomenon and the effect of viewing the robot as a person, we designed an experimental condition in which the robot's arm movements are not coupled with its speech and compared the participants' responses in this condition against other conditions.

Experiment Design and Subjects

Forty-three undergraduate students from Rensselaer Polytechnic Institute were recruited to participate. Their participation was compensated by giving course credits. The independent variable is the way the robot interacted with the participants, which has four conditions.

Common to all the conditions, we designed an interaction routine in which the robot talked about the research projects it is involved with and its various capabilities. Baxter's introductory statements consisted of fifteen sentences. Baxter then asked the participants ten questions about his/her personal interests, e.g. what kind of music do they like (to which they replied verbally on 7-point Likert scales) and commented on the participants' responses after all questions had been answered. The interaction with each participant lasted approximately five minutes. Afterward, Baxter prompted the participant to approach it and shake its right hand but assured the participant that he/she could decline to shake hands if he/she felt uncomfortable doing so. Then as the participant approached, Baxter extended its arm towards him/her. Before this gesture, the participant had not physically touched or been touched by the robot. Therefore, though extending one's arm towards the other person is natural for handshakes, it could be perceived as an unexpected event from the robot. We expected the participants to feel less apprehension – measured by how much hesitation they had to continue the handshake – if they treated the robot like another person.

Guided by research in proxemics (Hall et al., 1968), we positioned the participants within the social zone of Baxter, between 4' and 12' away: specifically, 9' (2.7 m) away. When the participants approached Baxter to shake hands, they left the social zone and entered the intimate zone. For safety reasons, the experimenter was in the same room during the interaction.

Two control groups were used: one in which participants saw no arm movement during the routine (No Gestures, $N = 9$) and one in which participants saw arbitrary arm movements at the start of the routine, not tied to any particular behavior or gesture (Arbitrary Gestures, $N = 8$). Neither control group saw a virtual face on Baxter. The first control group is a baseline to compare against all other groups that add a form of gesture to the routine. The second control group is used to determine whether any arm movement elicits the same response as arm movements that are performed to couple with the dialogue.

Two experimental groups were used: one in which participants saw meaningful gestures without a face (Meaningful Gestures, $N = 13$), and one in which participants saw meaningful gestures accompanied by a virtual human face (Meaningful Gestures with Face, $N = 13$).

We adopted a between-group design, and the participants were randomly assigned to one of four different groups. For dependent variables, we measured a categorical dependent variable, "hesitation," which represents how much the participant hesitated before approaching Baxter when prompted to shake hands. A score of "low" was given to the participants who waited less than one second to approach Baxter. A score of "medium" was given to the participants who waited between one and two seconds, and a score of "high" was given when the participants waited more than two seconds or looked to the experimenter for clarification that they could approach Baxter if needed. This measurement was taken by the experimenter.

For the participants' subjective reports about the robot, we used three dependent variables. During the interaction, one of the questions asked by Baxter was how friendly it was ("friendliness"). After the interaction, all participants responded to two additional questions asked by the experimenter during the debriefing phase: how comfortable would the participant be interacting with Baxter at a close distance ("comfortability closeness"), and how comfortable would the participant be letting Baxter touch them with one of its arms ("comfortability touch"). Similar to answering Baxter's questions, the participants answered these two questions using a 7-point Likert scale.

Implementation

Figure 2 shows the system architect we used to create the Baxter "character" in this experiment. The robot's gestures were created manually to complement each line of the dialogue, informed by human nonverbal behaviors, in particular, McNeill's work on hand gestures (McNeill, 1992). Baxter uses deictic gestures such as pointing to the participant

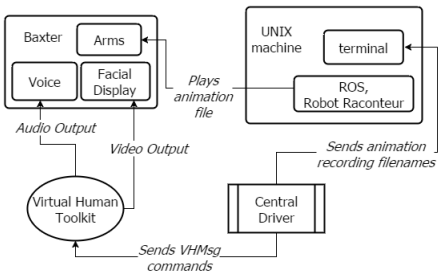


Figure 2: System Architect

or pointing to the general gesture space. Baxter uses iconic gestures when talking about activities. For instance, Baxter asks the participants, "To what extent do you enjoy working on pieces of art, or writing?" and moves its left hand in small circles as if it is writing on imaginary paper. It uses metaphoric gestures when discussing concepts like creating and expanding by bringing its hands together and then moving them apart. Finally, beat gestures are used with greetings and exclamations to punctuate Baxter's emotional intent.

We recorded Baxter's arm movements using ROS "Robot Operating System" and Robot Raconteur, a communication library for robotic systems (Wason & Wen, 2011). A ROS service records the joint positions of Baxter's arms as we manually manipulate the positions of the arms. We then systematically reduced the recording files down to keyframes to reduce any jittery movement, and stored the result of interpolating joint positions between key frames to be used for driving the robot's movements during the interaction.

In the condition which included a virtual face on Baxter's display screen, we utilized the Virtual Human Toolkit (VHT) (Hartholt et al., 2013). We used the face of a female character named Rachel, shown on Baxter as in Figure 1. During the interaction, Rachel's facial expressions remained mostly neutral. Rachel's voice was generated using Text-to-Speech.

We created a custom Windows Forms application written in C#/.NET as a driver for VHuman and Baxter during the interaction. It sent arm movement, speech, and facial expression commands to their respective destinations for each segment of the routine. This system also allowed the experimenters to utilize a "Wizard of Oz -style approach. When the participants were asked questions by Baxter, the experimenters could enter their answers into the application for Baxter to comment on later in the routine.

Results

A chi-square test of independence was conducted to examine the relation between the experiment condition and amount of time that participants hesitated before approaching Baxter to shake hands. The test showed a significant relationship between which version of the routine participants observed and their amount of hesitation to approach Baxter, $X^2(6) = 15.46$, $p = 0.02$. No participants declined to shake hands with Baxter, but they differed significantly in the amount of time to approach Baxter after the request.

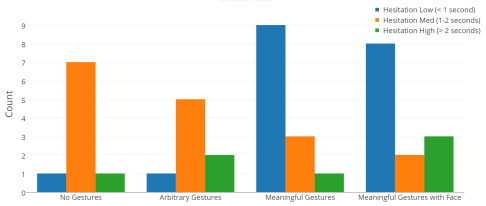


Figure 3: Hesitation before Handshake

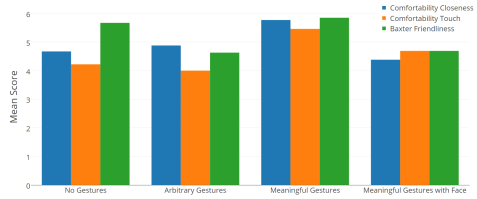


Figure 4: Mean Ratings of "Comfortability Closeness", "Comfortability Touch", and "Friendliness"

Of the standardized residual values, shown in Table 1, the differences were significant at the 0.05 level between the "Meaningful Gestures" condition and "No Gestures" and "Arbitrary Gestures" conditions for participants in the low and medium hesitation categories. Overall, participants in the "Meaningful Gestures" condition hesitated least in approaching Baxter, and participants in the control groups "No Gestures" and "Arbitrary Gestures" hesitated more. A comparison of hesitation measures of all conditions is shown in Figure 3.

One-way analysis of variance(ANOVA) tests were conducted to evaluate the differences between all groups on three ordinal measures: comfortability interacting with Baxter at a close distance ("comfortability closeness"), comfortability letting Baxter use its arm to touch the participant ("comfortability touch"), and how friendly the robot seemed during the routine ("friendliness"). A summary of the ANOVA tests can be found in Table 2, and the mean scores for each condition in all measures can be seen in Figure 4.

There was a significant difference among groups on how comfortable the participants would be interacting with Baxter at a close distance after having seen the routine, $F(3,39) = 3.154, p < 0.05$. Post hoc comparisons using the Fisher LSD test indicated that the mean score for the "No Gestures" group ($M = 4.67, SD = 1.12$) was significantly lower than the "Meaningful Gestures" group ($M = 5.77, SD = 1.09$), $p < 0.05$. The mean score for the "Meaningful Gestures" group was significantly higher than the "Meaningful Gestures with Face" group ($M = 4.38, SD = 1.33$), $p < 0.01$.

Groups differed significantly with how comfortable the participants would be letting Baxter touch them with one of its arms after having seen the routine, $F(3,39) = 3.865, p < 0.05$. Post hoc comparisons using Fisher LSD indicated that the mean score for the "Meaningful Gestures" group ($M = 5.46, SD = 0.88$) differed significantly from the "No Ges-

Table 1: Standardized Residual Values of Participants Hesitation in Experiment 1

	Low	Medium	High
No Gestures	**-1.5**	**1.8**	-0.4
Arbitrary Gestures	**-1.3**	1.0	0.6
Meaningful Gestures	**1.4**	-0.9	-0.8
Meaningful Gestures w. Face	0.9	**-1.4**	0.6

Table 2: ANOVA Tests for Experiment 1

Measure	*df*	*MS*	*F*	*p*
comfortability closeness	3	4.54	3.15	.035
comfortability touch	3	4.51	3.86	.017
friendliness	3	4.45	3.74	.019

tures" group ($M = 4.22, SD = 0.67$), $p = 0.01$, and from the "Arbitrary Gestures" group ($M = 4.00, SD = 0.76$), $p < 0.01$.

The final ANOVA test revealed a significant difference among groups with how friendly participants perceived Baxter to be, $F(3,39) = 3.742, p < 0.05$. Post hoc comparisons using Fisher LSD indicated that the mean score for the "No Gestures" group ($M = 5.67, SD = 1.00$) was significantly higher than the "Meaningful Gestures with Face" group ($M = 4.69, SD = 1.18$), $p < 0.05$. Additionally, the mean score for the "Meaningful Gestures" group ($M = 5.85, SD = 1.07$) was significantly higher than the mean scores for both the "Arbitrary Gestures" group ($M = 4.63, SD = 1.06$), $p < 0.05$, and the "Meaningful Gestures with Face" group, $p = 0.01$.

Discussion

In this study, the participants were more likely to approach Baxter without hesitation when they witnessed gestural arm movement beforehand, with or without the addition of a virtual human face on Baxter's display screen. Introducing irrelevant, arbitrary arm movement in the second control group demonstrated either no impact (when comparing to the first control group) or a negative impact (when compared to the two experiment groups) on the participants' hesitations, suggesting that meaningful arm movement was key in reducing the participants' hesitations. This confirms our hypothesis that gestures associated with a social context are most effective in increasing people's trust and feeling of comfort in interacting with the robot. We want to point out that the positive impact of accompanying speech with meaningful gestures took place after just a few minutes of interaction. This provides strong evidence for the benefit of embodying a robot as a social character.

Although the "Arbitrary Gestures" group reported a mean score for "comfortability closeness" similar to the experimental groups, this behavior negatively impacted ratings of how likely participants would be to let Baxter make physical contact with them. Participants in both control groups reported lower scores for "comfortability touch" than participants in the "Meaningful Gestures" group.

Finally, the participants rated Baxter as being friendlier when they saw meaningful gestures instead of arbitrary gestures or when the virtual face was not present. There are many possible explanations for this result. The digital human like face might introduce the uncanny valley effect. Moreover, because the rest of the robot's body is not human like, the participants might experience a disconnection between the face and the body. Combining with the fact that displaying the virtual face did not help reduce the participants' hesitations to approach Baxter, this result suggests that we should use caution when using a 3D realistic virtual face for this robot.

Experiment 2

This experiment evaluates whether people can differentiate and appropriately label the robot's attitudes based on the robot's body movements. Unlike in Experiment 1, we utilized the software CrazyTalk Animator by Reallusion to create the facial animations. An original face was drawn as a boxy, outlined appearance that is much less realistic than the human face in Experiment 1.

Experiment Design and Subjects

We modeled two routines for Baxter: a Proud routine (sterner and more bragging) and a Relaxed routine (more relaxing and soothing). We picked these two attitudes because they are likely to be useful in health care, and educational domains. Moreover, the difference between these two attitudes is subtler than more commonly perceived emotions like happiness or anger, and therefore, the success of this experiment will provide us more confidence in using the modified Baxter robot to express emotions and attitudes in the future. Similar to Experiment 1, in these routines, Baxter talked about student life on campus and the many things that the university offers.

Twenty-four undergraduate students were recruited to participate. This time, we used a within-group design. The participants observed the dual-arm Baxter robot equipped with a cartoonish face discuss a topic twice, using different nonverbal behaviors in each routine. Each participant saw both routines in random order. Each routine lasted approximately one and a half minutes. Baxter's speech was controlled to be similar between routines, with the main differences being the displayed nonverbal behavior and facial expressions.

For each routine, the participants need to pick the most appropriate emotion/attitude label from a list. HUMAINE Emotion Annotation and Representation Language (EARL) consolidates emotion labels into ten categories (Schröder, Pirker, & Lamolle, 2006). Pride and Relaxed belong to the "Positive Thoughts" and "Quiet Positive" categories respectively. We also picked one relevant label from each other categories. The ten options for the participants are anger, fear, relaxation, frustration, sadness, shock, happiness, affection, pride, and surprise.

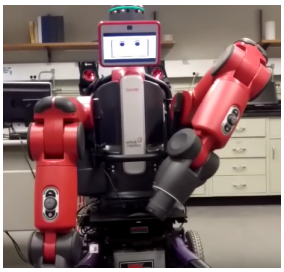


Figure 5: Baxter with CrazyTalk Cartoonish Face

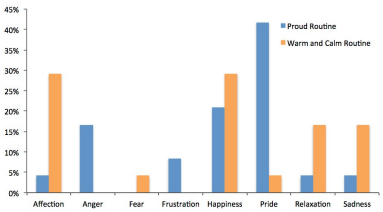


Figure 6: Emotion Labels for Each Routine

Implementation

As in Experiment 1, we recorded Baxter's arm movements using ROS and Robot Raconteur (Wason & Wen, 2011). When modeling pride, we want the robot to seem highly aroused and actively engaging the participant during most of the routine (Ekman & Friesen, 1967). Therefore, in the Proud routine, gestures using Baxter's arms were more focused toward the observing human participant. For example, during the initial introduction, Baxter says, "Hello, I am Baxter, one of the amazing robots in the robotics department, raises its right arm, and pats its torso in a highly precise, confident manner. In contrast, the Relaxed routine was associated with positivity (Ekman & Friesen, 1967). Movements for the Relaxed routine included sweeping the arms in slow arcs and keeping the hands rotated inwardly toward Baxter's body, with more frequent pauses and less overall movement than the Proud routine. For example, the initial greeting includes a small, tentative wave of the hand.

Similarly as in Experiment 1, we identified keyframes for each recorded file and interpolated the joint positions for creating smoother animations. Figure 5 shows Baxter with one of the faces from CrazyTalk. This face has limited capacity of expressing emotions with subtle differences. Also, we wanted to prevent the participants from forming their judgments mainly based on the robot's face. Therefore, the facial expressions remained mostly neutral for both routines. We used a male Text-to-Speech voice for the robot.

Results

A chi-square test of independence was performed to examine the relation between the designed routine and the attitude associated with it by participants. The labels shock and surprise were not chosen by participants for either routine, so we do not include them in the results.

The test showed a significant difference for which emotions were chosen for each routine based on the Pearson Chi-Square value, $X^2(7) = 22.80, p = 0.002$. Of the standardized residual values, the differences were significant at the 0.05 level among the three labels – pride, affection, and anger. Pride accounted for the largest amount of difference in the distribution, with affection second and anger third.

Figure 6 shows the distribution of chosen emotions for each routine among the twenty-four participants. The participants mostly matched the Proud routine to the intended proud emotion at 41.70% (10 participants), with happiness second at 20.80% (5 participants) and anger coming in third at 16.70% (4 participants). For the Relaxed routine, participants mostly matched the routine to both happiness and affection at 29.20% (7 participants) each, with relaxation and sadness at 16.70% (4 participants) each.

Discussion

The spread of participant labels for Baxter's emotional intent was more even for the Relaxed routine than for the Proud routine. In the Proud routine, the gestures were firmer and more aggressive and were more tightly coupled with precise moments in the dialogue, which may have contributed to this result. In post-experiment interviews, many participants reported that Baxter spoke about itself and the university with high praise and that the audience should agree with Baxter's comments – as opposed to the perceived inclusive, friendly tone of the Relaxed routine. Also, being relaxed can often be a secondary element of another perceived primary emotion such as joy (Ekman & Friesen, 1967). Relaxed and calmness is conveyed through slow, flowing movements, often associated with voice tone in ratio to the face and body (Ekman & Friesen, 1967, 1969). Pride, on the other hand, can be displayed more readily through use of nonverbal behavior. Therefore, the relaxed attitude may have been too vague to be successfully identified. In our previous work using a TurtleBot, we have observed that the noise made by the robot when it is moving can intensify the expression of negative and high dominance emotions such as anger, and affect the expression of low dominance emotions such as shyness (Barron & Si, 2013). Though the participants did not specifically comment on the noise as a factor affecting their impressions about the robot, we suspect the noise played a similar role in this study, i.e. lead some participants to believe the robot was angry.

We also observed an interesting ordering effect. The participants were more likely to see a greater contrast when they saw the Relaxed routine first and then labeled the Proud routine with a negative emotion. Feedback from these participants mentioned that Baxter seemed less friendly in the second routine. Labels applied to the Relaxed routine were more consistent regardless of the ordering of the routines.

Another interacting observation was that several participants mentioned hearing a change in the voice inflection between the two routines even though we used the same Text-to-Speech engine. Though not exactly the same as the McGurk effect, this also demonstrates what people see can affect what

they hear, even in such a brief and simple social interaction scenario.

Conclusions and Future Work

In this work, we examined the effects of designing meaningful arm movements for a modified version of Baxter robot. Our results suggest that it is feasible to create a social character using this robot even though it does not have an exact human-like shape. We demonstrated that designing body movements that are timely coupled with the robot's speech makes people more likely to trust the robot and feel comfortable in its presence comparing to demonstrating the robot's arm movements independently from its speech or not demonstrating how the robot movements at all. Moreover, this difference takes place in just a few minutes. Finally, the Baxter robot can use body language in a similar way as humans for expressing distinguishable attitudes, more specifically, being either proud or relaxed. However, its proud manner has the danger of being misinterpreted as anger, and its relaxed attitude is very easily confused with other positive emotions such as happy and affection.

Future work would further examine how using meaningful arm and face movements to express emotional intent affects people's willingness to approach and interact with Baxter in close proximity and over a longer period. Right now the robots body movements are planned independently from the participants. In real life, when two conversational partners have established rapport, their body movements are often coupled (Cassell, Gill, & Tepper, 2007). For example, one will often nod when the other pauses in his/her speech. One of our future directions is to enable such interactions for the robot. Secondly, we want to observe how people's attitudes toward the robot change over longer terms of interaction, and at moments when the robot makes a seemingly threatening action. In the latter case, we want to investigate whether the same techniques people use for recovering trust can be applied to the robot. Finally, compared to digital characters, when working with a physical robot, there is always the additional challenge of synchronizing the movements of its various parts, such as the two arms, or between the face and the arms. Even with the same set of commands, it is possible the movements become out of sync for a variety of reasons, which may break the image of the robot as a social character. We are interested in developing a software framework for helping with this challenge. When the robots body movements are delayed, we can also slightly delay its facial expressions and speech, or replan for another set of body movements, facial expressions, and speech.

References

Barron, M., & Si, M. (2013). Augment interactive storytelling with cognitive robot. In *Proceedings of the 6th digital games research association (digra) conference.*

Bethel, C. L., & Murphy, R. R. (2008). Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *38*(1), 83–92.

Bickmore, T., & Rosalind, J. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293-327.

Breazeal, C. (2000). *Sociable machines: Expressive social exchange between humans and robots*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent robots and systems, 2005.(iros 2005). 2005 ieee/rsj international conference on* (pp. 708–713).

Cassell, J., Gill, A., & Tepper, P. (2007). Coordination in conversation and rapport. In *Proceedings of the workshop on embodied language processing* (p. 41-50).

Ekman, P., & Friesen, W. V. (1967). Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, *24*(3), 711–724.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, *1*(1), 49–98.

François, D., Powell, S., & Dautenhahn, K. (2009). A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking. *Interaction Studies*, *10*(3), 324–373.

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold Jr, A. R., Durbin, M., ... Vayda, A. P. (1968). Proxemics [and comments and replies]. *Current anthropology*, 83–108.

Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., ... Gratch, J. (2013). All together now. In *Intelligent virtual agents* (pp. 368–381).

Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *Intelligent robots and systems, 2008. iros 2008. ieee/rsj international conference on* (pp. 3230–3235).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Schröder, M., Pirker, H., & Lamolle, M. (2006). First suggestions for an emotion annotation and representation language. In *Proceedings of lrec* (Vol. 6, pp. 88–92).

Severinson-Eklundh, Kerstin and Green, Anders and Httenrauch, Helge . (2003). *Social and collaborative aspects of interaction with a service robot.* (Tech. Rep.). Royal Institute of Technology (KTH).

Wason, J. D., & Wen, J. T. (2011). Robot raconteur: A communication architecture and library for robotic and automation systems. In *Automation science and engineering (case), 2011 ieee conference on* (pp. 761–766).

Similarity-Based Reasoning is Shaped by Recent Learning Experience

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
David Jaz Myers (david.myers@oberlin.edu)
Oberlin College, Department of Psychology
120 W. Lorain St., Oberlin, OH 44074, USA

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
SUNY Purchase College, Department of Psychology
735 Anderson Hill Road, Purchase, NY 10577, USA

Abstract

Popular approaches to modeling analogical reasoning have captured a wide range of developmental and cognitive phenomena, but the use of structured symbolic representations makes it difficult to account for the dynamic and context sensitive nature of similarity judgments. Here, the results of a novel behavioral task are offered as an additional challenge for these approaches. Participants were presented with a familiar analogy problem (A:B::C:?), but with a twist. Each of the possible completions (D1, D2, D3), could be considered valid: There was no unambiguously "correct" answer, but an array of equally good candidates. We find that participants' recent experience categorizing objects (i.e., manipulating the salience of the features), systematically affected performance in the ambiguous analogy task. The results are consistent with a dynamic, context sensitive approach to modeling analogy that continuously updates feature weights over the course of experience.

Keywords: similarity; analogy; statistical learning; relational reasoning; categorization

Introduction

> *But soft! What light through yonder window breaks?*
> *It is the east, and Juliet is the sun.*
>
> Romeo & Juliet, Act 2, Scene 2

It is one of the most recognizable metaphors in the Western canon: a love-struck Romeo spies Juliet at her window and compares her to the star that nourishes the earth with light and heat. Though it seems like a straightforward, if not clichéd figure of speech, there are in fact several commonalities between Capulet's daughter and the sun; they are both rising in the east, they are both beautiful, and they are both golden (for the sake of illustration, we assume Juliet is wearing a yellow dress). When we are confronted with several possible matches in a similarity-based comparison such as this, what drives us to select one interpretation in particular[1]?

In the past three decades, a great deal of research has examined the nature and development of similarity-based reasoning, which is believed to play a major role in everything from problem solving to creativity to scientific discovery (Holyoak & Thagard, 1996). Cognitive scientists have argued that *analogy* is at the heart of this process, and computational models of analogy have successfully captured a range of cognitive and developmental phenomena that require similarity-based reasoning (Gentner & Forbus, 2011; Hummel & Holyoak, 2005). The dominant approach to modeling analogy assumes that highly structured symbolic (or hybrid) representations and a particular suite of cognitive machinery are necessary for analogical mapping (e.g., SME: Falkenhainer, Forbus, & Gentner, 1989; LISA: Hummel & Holyoak, 2003; Hummel, 2010).

One key observation is that children undergo a perceptual-to-relational shift in reasoning over the course of development: younger children typically match based on perceptual similarities ("they're both yellow!"), while older children may start to match based on relational similarities ("they are both *rising* in the east!"; Gentner, 1988; Piaget, 1952). This developmental trajectory could be taken as evidence for structured models of analogy like SME and LISA as these approaches are well equipped to model this phenomenon. Once children learn the important relations in the world (e.g. that the sun appears in the east as the earth rotates during its orbit each day), they represent these structures symbolically (Gentner et al., 1995). Such representations would allow a mapping mechanism to operate reliably over structured symbols for consistent and effective similarity-based reasoning.

However, several empirical findings may present a challenge for this approach. For instance, similarity-based reasoning does not always follow a perceptual-to-relational shift. Adults are sometimes lured towards perceptual matching (Morrison et al., 2011) and even children flexibly generalize based on perceptual *or* relational similarity depending on the task context and their prior experiences (Bulloch & Opfer, 2009; Opfer & Bulloch, 2007).

Recently, we showed that these patterns of behavior spontaneously emerge over the course of development in a connectionist model that relies on domain-general statistical learning and fully distributed internal representations (Thibodeau, Tesny, & Flusberg, 2014). Indeed, context-sensitive, similarity-based reasoning is one of the key strengths of the connectionist framework (Flusberg & McClelland, 2014), and while critics have commonly

[1] With all due respect to our high school English teachers, who convincingly argued that Shakespeare's brilliance lies in the fact that his metaphors are intrinsically multifaceted, and thus no single interpretation is really "correct."

assumed that this only applies to perceptual or feature-based similarity (see, e.g., the accompanying commentary to Rogers & McClelland, 2008), we have argued – and demonstrated – that fully distributed neural networks can in fact reason and draw inferences based on shared relational structure (Flusberg et al., 2010; Thibodeau et al., 2013). A critical feature of these models is their ability to extract and represent the statistical structure of their inputs over the course of training. By treating (relational) language as part of the fabric of experience, we can see the emergence of sophisticated metaphorical and analogical abilities.

One unique prediction that follows from this type of model is that what counts as "similar" in comparisons among novel stimuli will dynamically change as a result of recent experience. This is because internal representations of environmental structure are updated relatively quickly during early epochs of training, which will cause the object features that are prominent in experience to have greater weight in object representations (and therefore in emergent similarity relationships). The current study represents a novel test of this prediction by investigating whether recent experience attending to particular feature dimension really does lead that feature to figure more prominently in analogical reasoning. Finding support for this possibility would present a challenge for models of analogical reasoning that do not naturally accommodate dynamic, experience-based updating of distributed object representations.

To bring it back to Shakespeare, if an audience member has a great deal of (recent) experience making similarity judgments based on color, we would expect her to interpret Romeo's remark as a comment on Juliet's outfit. If, on the other hand, she has been attending more to the attractiveness of those around her, we would expect her to interpret the metaphor as a comment on Juliet's radiant beauty.

In the present study, we tested the hypothesis that getting participants to attend to and learn about a particular facet of experience would influence their subsequent similarity-based reasoning. Participants first completed a category-learning task where they had to figure out the meaning of a novel category label that referred to one of several possible features (size, shape, or brightness). They then completed an ambiguous analogy task and a similarity-rating task. Our hypothesis was that experience in the training phase would influence subsequent performance in the similarity-based reasoning tasks.

Experiment

Methods

Participants We sampled 400 participants from Amazon's Mechanical Turk. Of these, eight submitted incorrect completion codes and were excluded from analysis, leaving data from 392 for analysis.

Materials and Design The experiment consisted of four between-subjects conditions and three tasks: a category label training task, an ambiguous analogy task, and a similarity rating task.

Training Task Participants were randomly assigned to one of four training groups. Three of them differed in how they provided feedback to participants learning about a novel category label. A fourth, control condition omitted the training phase altogether.

On each of the 16 trials of the training task, participants were presented with two shapes and asked to choose "Which is the {truffet/lodi}?" (see Figure 1 for an illustration of a trial of the training phase). For half of the participants, the target label was "truffet" and for the other half, the target label was "lodi." The two shapes always differed along three dimensions: *size*, *shape*, and *brightness*. There were two levels of each dimension (i.e. large vs. small, circle vs. square, and bright vs. dim). Participants were forced to choose between the shapes and were given accuracy feedback ("correct" or "incorrect"). One training group was given feedback indicating that the novel label referred to the dimension of size (either meaning "bigger" or "smaller", counterbalanced across participants), another group was given feedback indicating it referred to shape ("square" or "circle"), and the last training group was given feedback indicating it referred to brightness ("dimmer" or "brighter"). While the feedback participants received always consistently mapped on to one of these relations, they were never explicitly told what the target label meant.

The pairs of shapes shown in the 16 training trials were identical across conditions, though the order in which they were presented was randomized.

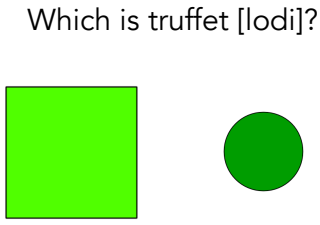


Figure 1. An example trial from the training phase.

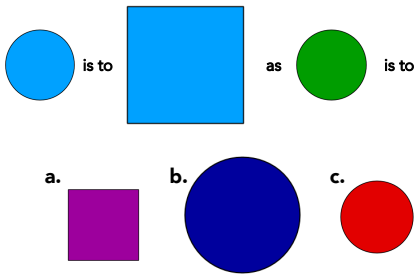


Figure 2. An example trial from the ambiguous analogy task.

Ambiguous Analogy Task Following the training phase, participants completed eight trials of an ambiguous analogy task. The instructions for this task read: "On the following screens, you will see a series of analogy questions, as in *A is*

to B as C is to what? You will see three items that could potentially complete the analogy. Choose the one that you think best completes the analogy." Figure 2 shows an example trial of this task.

The analogy task is ambiguous because none of the forced choice responses are a perfect match to the sample relation. For instance, in Figure 2 the best answer for the analogy would be a large, green, square. A large green square would be a *different* size and shape than the small green circle but the *same* brightness; this would mirror the relationship between the large blue square and the small blue circle, which are *different* sizes and shapes but the *same* brightness.

This "perfect" analogy was not provided as an option to participants, though. Instead, the three forced choice options were designed to be relationally similar to the sample in one of three ways: by shape, size or brightness. For instance, in Figure 2 (see also Table 1), (*a*) was a relational match to the sample by shape (because, like the sample objects, these two were *different* shapes) but not by size or brightness (because, unlike the sample objects, these two were the *same* size and *different* in brightness); (*b*) was a relational match to the sample by size (because, like the sample objects, these two were *different* sizes) but not by shape or brightness (because the two objects were, unlike the sample, the *same* shape and *different* in brightness); (*c*) was a relational match to the sample by brightness (because, like the sample objects, these two were the *same* brightness) but not by size or shape (because, unlike the sample, were the *same* size and the *same* shape).

Table 1. Illustration of the relationship between the sample objects and potential analogical matches for one of the ambiguous analogy trials. In this case, option a is the best analogy on the basis of shape, option b is the best analogy on the basis of size, and option c is the best analogy on the basis of brightness.

Option	Objects	Shape	Size	Brightness
Sample	:	Different	Different	Same
a.	:	**Different**	Same	Different
b.	:	Same	**Different**	Different
c.	:	Same	Same	**Same**

Note, however, that although we have used the terms *same* and *different* to refer to the relations in the table, the relationships themselves are dimension-specific. For instance, objects that *differ* in size, *differ* because one is *larger* (*smaller*) than the other; objects that *differ* in brightness differ because one is *brighter* (or *dimmer*) than the other.

As with the training phase, the stimuli for the ambiguous analogy task were identical across conditions. No feedback was given on this task. The order of response options and trials were randomized between participants.

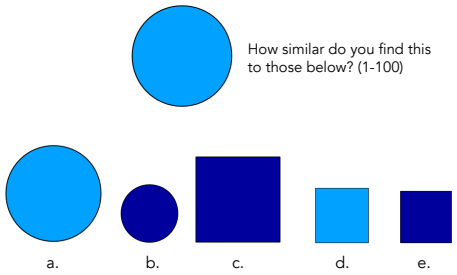


Figure 3. An illustration of the similarity-rating task.

Similarity Rating The final task consisted of four trials where participants made similarity ratings. On each trial, a sample object was shown at the top of the screen and five others were shown below (see Figure 3). Of these five items, one was identical to the original (e.g., option *a* in Figure 3) and one was maximally different, given the similarity space (i.e. option *e* in Figure 3, which is different in size, shape and brightness from the original). The other three items matched the original along one of the three target dimensions but not the other two. For instance, in Figure 3, *b* was a shape match, *c* was a size match and *d* was a brightness match.

Participants indicated how similar they viewed the original to each of the five items by using a 101-point slider bar (0 = lowest similarity, 100 = highest similarity). The order of the five comparison objects was randomized between participants, as was the order of the trials.

Results

Training The results of the training phase indicated that participants reliably learned the "meaning" of the target word they were assigned. A repeated-measures logistic regression with a predictor for trial (1-16) revealed that participants significantly improved (i.e. changed their response patterns to be more consistent with the feedback that they were given) over the course of training, $\chi^2(1)=80.053$, $p<0.001$ (AIC_1, a model with an intercept only, =4442.1; AIC_2, in which a predictor for trial number was added, =4364.1), $B=0.078$, $SE=0.008$, $p<0.001$.

A repeated-measures logistic regression with separate predictors for trial by condition (shape, size, brightness) revealed that participants in each of the three conditions significantly improved over the course of training, but it revealed that they did so at different rates, $\chi^2(2)=267.57$, $p<0.001$ ($AIC_1=4364.1$, $AIC_2=4100.5$). People improved the fastest in the "shape" condition, $B=0.438$, $SE=0.035$, $p<0.001$. People improved more gradually in the "size", $B=0.072$, $SE=0.012$, $p<0.001$, and "brightness" conditions, $B=0.028$, $SE=0.011$, $p=0.009$ (see Figure 5).

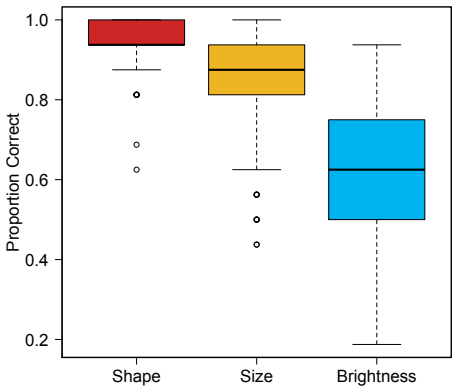


Figure 5. Boxplot illustrating differences in the degree to which participants were attuned to the feedback in the Shape, Size, and Brightness conditions.

By design, the training phase required participants to test hypotheses about the meaning of the target word. Does "truffet" (or "lodi") mean *larger*, *smaller*, *square*, *circle*, *brighter*, or *dimmer*? The differences in the rates of change across the conditions suggest that participants were most likely to test the possibility that the novel label described the shape of the object. This is consistent with research showing that children typically have a shape bias in word learning (Landau, Smith, & Jones, 1988).

Data from the final trial of the training phase support this hypothesis: People answered the final trial in a way that was congruent with their condition 98.0% of the time in the shape condition, 83.2% of the time in the size condition, and 70.0% of the time in the brightness condition, $\chi^2(2)=29.333$, $p<0.001$.

Ambiguous Analogy Task As shown in Table 2, the basic pattern of data from the ambiguous analogy task are consistent with our prediction: After training that focused on shape, people were more likely to choose the shape-match (69.7% compared to 61.9% in the control condition); after training that focused on size, people were more likely to choose the size-match (54.8% compared to 32.7% in the control condition); after training that focused on brightness, people were more likely to choose the brightness-match (14.5% compared to 5.3% in the control condition).

Table 2. Percentages of responses by training condition.

	Shape	Size	Brightness
Control (n = 89)	62	33	5
Shape (n = 102)	**68**	29	3
Size (n = 101)	40	**55**	5
Brightness (n = 100)	50	34	**15**

To analyze these data, we fit a series of mixed-effect logistic regression models to predict participants' choices. In these models, there were two fixed effects: response type (shape-match, size-match, and brightness-match), and training condition (shape, size, brightness, and control). Participant and trial number were included as random effects (c.f. Clark, 1973; Jaeger, 2008;).

We first tested whether there were significant differences in the likelihood of selecting the shape-, size-, and brightness-matches by comparing a model *without* a predictor for response type to a model *with* a predictor for response type, $\chi^2(2)=1890.7$, $p<0.001$ ($AIC_1=11983$, $AIC_2=10096$). This model revealed that people were more likely to choose the shape-match (54.8% overall) than the size-match (38.1% overall), $B=0.677$, $SE=0.051$, $p<0.001$, or the brightness-match (7.0% overall), $B=2.773$, $SE=0.0785$, $p<0.001$. People were also more likely to select the size-match than the brightness-match, $B=2.096$, $SE=0.079$, $p<0.001$.

We then tested whether there were significant differences in response patterns by training condition by adding interaction terms between condition and response type, $\chi^2(6)=369.43$, $p<0.001$ ($AIC_1=10096$, $AIC_2=9744.5$). The main effects of response in this model were consistent with the previous model (i.e. people were most likely to choose the shape-match and least likely to choose the brightness-match). In full model, the control condition served as a baseline (see Figure 5, which shows differences from the control condition). Critically, the model revealed the predicted congruence effects.

Compared to those in the control condition, participants in the shape condition were more likely to choose the shape-match ($B=0.262$, $SE=0.108$, $p=0.015$) and less likely to choose the brightness-match ($B=-0.538$, $SE=0.260$, $p=0.038$), though they were no less likely to choose the size-match ($B=-0.179$, $SE=0.111$, $p=0.108$).

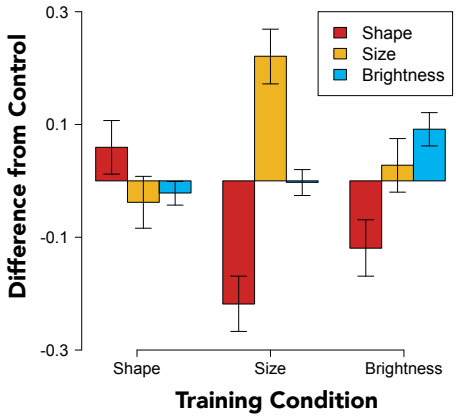


Figure 5. Results of the ambiguous analogy task. The difference in the proportion of shape-matching, size-matching, and brightness-matching choices from the control condition are shown for each condition. Error bars denote standard errors of the means.

Compared to participants in the control condition, those in the size condition were more likely to choose the size-match ($B=0.914$, $SE=0.107$, $p<0.001$) and less likely to choose the shape-match ($B=-0.888$, $SE=0.105$, $p<0.001$), though they were no less likely to choose the brightness-match ($B=-0.053$, $SE=0.231$, $p=0.818$).

Compared to those in the control condition, those in the brightness condition were more likely to choose the brightness-match (B=1.101, SE=0.195, p<0.001) and less likely to choose the shape-match (B=-0.487, SE=0.105, p<0.001, but were no less likely to choose the size-match (B=0.124, SE=0.109, p=0.256).

These analyses reveal that training systematically influenced behavior on the ambiguous analogy task. When the training phase encouraged people to attend to the shapes of the objects, they were subsequently more likely to use shape information to complete the analogy task. Parallel effects were found for the size and brightness conditions.

Individual Differences We can further investigate the effect of training by testing whether people who performed better on the training task showed more pronounced effects on the ambiguous analogy task. We found such a relationship for each of the three conditions in three separate repeated-measures logistic regressions. In the shape condition, people who selected more shape-matches in training chose more shape-matches on the ambiguous analogy task, B=7.578, SE=3.837, p=0.048. Similarly, people in the size and brightness conditions who chose more size- and brightness-matches in training chose more size- and brightness-matches in the ambiguous analogy task, B=7.399, SE=1.792, p<0.001 and B=3.533, SE=1.568, p=0.024, respectively. This shows that when we take individual differences in performance on the training task into account, we see even more nuanced and systematic effects on the ambiguous analogy task.

Similarity-Rating Results from the similarity-rating task mirrored those of the ambiguous analogy task, showing a systematic effect of training. A mixed-effect ANOVA with condition and comparison object as fixed effects and participant and trial number as random effects revealed a main effect of comparison object, F[4, 30952]=18159.528, p<0.001 and an interaction between comparison object and training condition, F[12, 30952]=33.627, p<0.001. There was a marginal main effect of condition, F[3, 388]=2.605, p=0.052.

Comparing the fit of nested models confirmed these effects: a model with a predictor for a main effect of comparison object provided a significantly better fit to the data than a model without such a predictor, $\chi^2(4)=37127$, p<0.001 (AIC_1= 310610, AIC_2=273524). Including interaction terms between comparison object and condition provided an even better fit to the data, $\chi^2(15)=408.93$, p<0.001 (AIC_1=273524, AIC_2=273270).

The full model revealed that people rated the identity match as the most similar to the comparison object (M=97.00, sd=7.85), followed by the shape-match (M= 51.56, sd=18.95), brightness-match (M=36.32, sd=18.38), size-match (M=32.34, sd=19.05), and, finally, the object that differed from the original on each of the three dimensions (M=14.04, sd=15.34). All pairwise comparisons were significant (ps<0.001).

To investigate the interaction between training condition and comparison object, we treated similarity ratings from the control condition as a baseline (see Figure 6). We found that people in the shape condition rated the shape-match as marginally more similar to the comparison object than people in the control condition, B=2.463, SE=1.422, p=0.083. These participants rated the size-match as significantly less similar, B=-3.721, SE=1.636, p=0.023. Similarity ratings of the anchors and brightness match did not differ. That is, people in the shape condition considered shape as a marginally more important dimension for determining the similarity of objects, and size as a less important dimension, after being forced to attend to shape in the training phase.

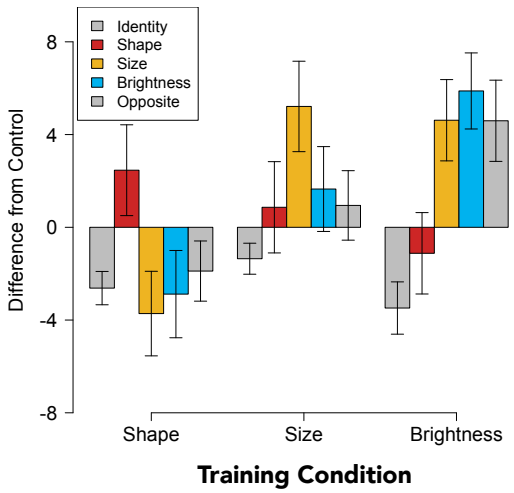


Figure 6. Results from the similarity-rating task. The difference in the mean similarity of the identity-match, shape-matching, size-matching, brightness-matching, and maximally different objects from the control condition are shown by shape, size, and brightness condition. Error bars reflect standard errors of the means.

We found that people in the size condition rated the size-match as significantly more similar to the comparison object than people in the control condition, B=5.213, SE=1.640, p= 0.001. Similarity ratings of the anchors, shape-match, and brightness-match did not differ. In other words, people in the size condition viewed size as a more important dimension for determining the similarity of objects after being forced to attend to size in the training phase.

Finally, we found that people in the brightness condition rated the brightness-match as significantly more similar to the comparison object than people in the control condition, B=5.881, SE=2.070, p=0.004. These participants also rated the size-match as more similar, B=4.618, SE=1.643, p=0.005, and the maximally different object as marginally more similar, B=4.595, SE=2.606, p=0.078. They rated the identity-match as less similar to the comparison object, B=-3.481, SE=1.517, p=0.022. Their ratings of the shape-match did not differ from the control condition. In this condition too, we see the predicted congruence effect, but we also see that people seem to shift their conception of similarity more globally as well.

General Discussion

Is Juliet the sun because she is rising in the east, because she is beautiful, or because she is yellow? The answer, of course, is "yes." However, our experiment has shown that the similarity match people make depends in no small part on their learning history and which aspects of experience they have been attending to. In our study, participants first completed a category training task where they had to determine the meaning of a novel category label by choosing which of two objects on a given trial fit the label (or they completed no training task at all). Feedback was given that was consistent with a label that mapped onto the size, shape, or brightness of objects. Participants then completed an ambiguous analogy task and a similarity-rating task.

A series of analyses demonstrated that responses to the analogy and similarity-rating task were systematically influenced by the particular category-training feedback participants received. Participants who learned that a novel category label referred to size, for example, were more likely to choose an object on the ambiguous analogy task where size was the critical dimension, even though options based on shape or brightness were equally valid. They also rated objects that matched in brightness as relatively more similar in the similarity-rating task.

These findings support a dynamic view of similarity. The salience of object features can change as a function of experience, and, in turn, affect similarity-based reasoning and inference. This finding has important implications for theories of analogical reasoning, metaphor, and similarity. In particular, popular approaches to modeling analogy and other types of similarity-based reasoning would have difficulty accommodating these results because they have no natural way of dynamically updating feature weights as a result of ongoing experience. Conversely, this is exactly the sort of finding that we would expect based on models of analogy based around principles of statistical learning and distributed representation (Thibodeau et al., 2013; 2014).

References

Bulloch, M.J., & Opfer, J.E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12,* 114-122.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335-359.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1–63.

Flusberg, S. J. & McClelland, J. L. (2014). Connectionism and the emergence of mind. *Oxford University Press Handbook of Cognitive Science* (online edition)

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., & Glick, J. J. (2010). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology*, *1*(0), 1-11

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child development*, pages 47–59.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *WIREs Cognitive Science, 2,* 266–276.

Gentner, D., Rattermann, M. J., Markman, A. B., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 263-313). Hillsdale, NJ: LEA.

Gentner, D., Simms, N., & Flusberg, S. (2009). Relational language helps kids reason analogically. *Proceedings of the 31st annual meeting of the Cognitive Science Society, Amsterdam.*

Holyoak, K. J., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought.* Cambridge, MA: MIT Press.

Hummel, J. E. (2010). Symbolic versus associative learning. *Cognitive Science, 34,* 958-965.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Hummel, J. E., & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. *Current Directions in Psychological Science, 14*(3), 153–157.

Jaeger F. T. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434-446.

Landau, B., Smith, L.B., & Jones, S.S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3,* 299-321.

Morrison, R.G., Doumas, L.A.A., & Richland, L.E. (2011). A computational account of children's analogical reasoning: blanacing inhibitory control in working memory and relational representation. *Developmental Science, 14,* 516-529.

Opfer, J. E., & Bulloch, M. J. (2007). Causal relations drive young children's induction, naming, and categorization. *Cognition*, *105,* 207-217.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*, 689–749.

Thibodeau, P. H., Flusberg, S. J., Glick, J. J., & Sternberg, D. A. (2013). An emergent approach to analogical inference. *Connection Science, 25*(1), 27-53

Thibodeau, P. H. & Tesny, E. M., & Flusberg, S. J. (2014). Accounting for the relational shift and context sensitivity in the development of generalization. *Proceedings of the 36th annual meeting of the Cognitive Science Society, Quebec, Canada.*

Curiosity and Its Influence on Children's Memory

Haley Walin (hwalin@berkeley.edu)
Shaun O'Grady (shaun.ogrady@berkeley.edu)
Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California,
Berkeley, CA 94720 USA

Abstract

Curiosity has a tumultuous past. Originally curiosity was considered a vice of excess leading to misconduct and disaster. Recently, curiosity has transformed into a virtue of self-expression resulting in success and better performance. In classrooms, educators try to find ways of eliciting curiosity from their students: allowing them to pick their own research topics and books, including pop culture references in lecture, and many more strategies. Recent adult studies have revealed better memory for trivia facts that elicit more curiosity. The current study modifies the methods used in previous adult studies in order to make them more appropriate for children. Results from a sample of 24 7- and 8-year-olds reveal that by age eight curiosity significantly affects memory for trivia facts. This research may shed light on the cognitive advantages of curiosity and legitimatize the encouragement of curiosity in classrooms for school age children.

Keywords: curiosity; children; Information Gap Theory

Introduction

Philosophers and psychologists have attempted to define curiosity for generations with no real consensus. In 1994, George Loewenstein, taking pieces of each previous theories of curiosity, proposed the Information Gap theory (Loewenstein, 1994). Loewenstein saw curiosity not as a drive, reaction or motivation but rather a state of deprivation. Inspired by the idea that curiosity begins with prior knowledge and leads to new knowledge, Lowenstein argues that curiosity is induced by information gaps which can explain the variation in the situational determinants that lead to a state of curiosity. Every one has some prior knowledge that is organized into networks (Chi & Koeske, 1989). When an information gap in a particular knowledge network is made salient, curiosity is induced.

Information Gap theory can explain why curiosity seems to have varied intensity. Loewenstein (1994) believes that the size of the information gap predicts the intensity of the resulting curiosity. Larger gaps lead to low curiosity and small gaps lead to high curiosity. Larger gaps suggest that the knowledge network is not as extensive and therefore probably does not contribute to a person's self-concept, making that piece of information less rewarding. Smaller gaps suggest that the knowledge network is extensive and therefore probably does contribute to a person's self-concept, making that piece of information potentially more rewarding.

Recently, Kidd & Hayden (2015) have argued that we should move beyond a mere definition of curiosity and investigate the behavior in light of Tinbergen's (1963) four questions for explaining the behavior of an organism: (1) what function does the behavior serve?, (2) what is the mechanism or causal explanation for the behavior?, (3)how did the behavior evolve over phylogeny?, (4) how does the behavior develop over ontogeny? Indeed, Loewenstein's Information Gap theory can provide insight into curiosity's function and mechanism.

Once a salient information gap produces curiosity, people engage in exploratory behaviors in order to find the desired information. Exploratory behaviors and the eventual retrieval of desired information will vary depending on the size of the gap, the expected reward of filling the gap and one's perceptions of their ability to fill the gap. If the size of the gap is too large and the reward of filling it is not substantial or important to one's self concept people are less likely to go through with acquiring the desired information. In addition, if someone believes they are unable to acquire the information they will be less likely to do so. If one does eventually obtain the information, she will then put that new information into her knowledge networks and in the process modify them. The filling of the gap or rather the modifying of the network is the reward. Curiosity, therefore, can act as an intrinsic motivation for people to continue to build up and expand their knowledge networks.

Machine learning has been using a process similar to Information Gap theory to create artificial curiosity. In the case of a robotic arm, a reinforcement learner with an adaptive world model uses the expected improvement of the world model as intrinsic motivation and reward (Ngo, 2012). Essentially, the robotic arm has knowledge networks that it aims to improve and experiences reward when that expected improvement, filling of information gaps, is completed. For machine learning, artificial curiosity has been an effective way for machines to learn about the world for themselves, a form of active learning.

Curiosity in fact can be considered a mechanism for active learning since the desire, initiative and execution of learning originates purely from the learner. Active learning has been shown to lead to a better quality of learning (Benware & Deci, 1984). One question remains: can curiosity lead to better learning? Scientists have shown that information itself is rewarding (Marvin & Shohamy, 2016) and that curiosity may lead to better memory retention in

adults (Gruber et al., 2013; Kang et al., 2009), however, the mechanism by which it does remains unknown.

Recent research in reward learning may illuminate the process by which curiosity can influence memory retention. Research on reward began with the simple idea that dopamine was the pleasure/reward neurotransmitter (Wise, 1985). It was soon found, however, that rats depleted of dopamine still somehow could feel pleasure (Berridge & Robinson, 1998). Interestingly, dopamine depleted rats do show a decrease in exploratory behaviors (curiosity) in the presence of novel objects and reward seeking (Ungerstedt et al., 1971; Heffner et al., 1972). Dopamine, in recent years, has been shown to regulate reward anticipation rather than reward itself. In reward anticipation it is the predictive cue that leads to a spike of dopamine (Simansky et al., 1985; Philips et al., 1991; Phillips et al., 1992; Phillips et al., 1993, Blackburn et al., 1989; Schultz et al., 1995). The reason curiosity is affected by dopamine depletion might be that curiosity is a form of reward anticipation. The information gap of Lowenstein's theory could act as a predictive cue for reward. Experiments have shown that the indicated level of curiosity induced in participants by trivia questions affected the level of activity in the substantia nigra and ventral tegmental areas (Gruber et al., 2013; Kang et al., 2009), the main centers of reward, before the answer to that question was given. The SN/VTA is potentially highly connected with the hippocampus. Some hypothesize that the VTA and hippocampus form a functional loop and VTA increased activity can lower the threshold needed for a long-term potentiation (Lisman & Grace, 2005). Curiosity, therefore, by activating the reward centers can make later encoding of information easier. Is this true for children as well?

The relationship between curiosity and exploration has been studied extensively. Using a Bayesian perspective, Laura Szhulz and her colleagues (Schulz, 2012) have investigated children's exploratory behavior when confronted with ambiguous evidence (Schulz & Bonawitz, 2007; Gweon & Schulz, 2008; Cook, Goodman & Schulz, 2011;) and when evidence violates their expectations (Bonawitz, van Schijndel, Friel & Schulz, 2012). Yet, most of the research on curiosity and memory has been conducted on adults. This phenomenon has yet to be explicitly studied in children.

In this study a modified form of the Gruber et al. (2013) methods will be used to elicit and measure curiosity as well as measure memory retention in children. Child participants were asked to rank trivia questions based on how much curiosity the questions induce relative to each other on a scale presented on a game board. The game will not use any questions a child determines she knows, therefore all the questions should incite some level of curiosity since every question highlights a particular information gap. After ranking the questions, the children will learn the answers and will then be tested on these questions to see which answers they remember and how those questions relate to the child's curiosity rankings. We hypothesize that children will be more likely to remember questions they ranked on the high end of the scale (i.e., high curiosity questions rather than the questions they ranked as on the low end of the scale (i.e., low curiosity questions). In addition to answering pivotal questions about the development of curiosity, the present study will further emphasize the importance of peaking a student's curiosity in the classroom.

Methods

Participants

Nine seven-year-old children (Mean age = 7.43; SD = 0.22; 4 female) and 16 eight-year-old children (Mean age = 8.43; SD = 0.30; 8 female) participated in this study. Participants were recruited using the Berkeley Early Learning Lab database as well as during testing sessions at the Lawrence Hall of Science. Informed consent was obtained from the participants' parents. Children received a small prize for their participation.

Materials

Thirty-four trivia questions were previously piloted to ensure that on average 60% of children ages 6-7 did not know their answers. The trivia questions included content related to living animals, dinosaurs, astronomy, geography, and miscellaneous (e.g., How big is the brain of a stegosaurus? How many planets are in the solar system?). Questions were printed on 2 inch by 3 inch laminated cards and the answers were printed on the back of each card.

Children were asked to rate 6 randomly-selected questions based on the amount of curiosity the questions elicited. For each question the child was asked if they believed they knew the answer to the question. If the child stated that they did know the answer or if the the child guessed either correctly or incorrectly, the question was discarded and a new one was drawn. This was done in order to insure that any differences in recall were due to curiosity ratings rather than prior knowledge.

The participants were given a board with a scale of 1-6 (1-3 representing no to little curiosity and 4-6 representing substantial to extreme curiosity). On the board above each number of the scale was an associated box where the child could place one question card.

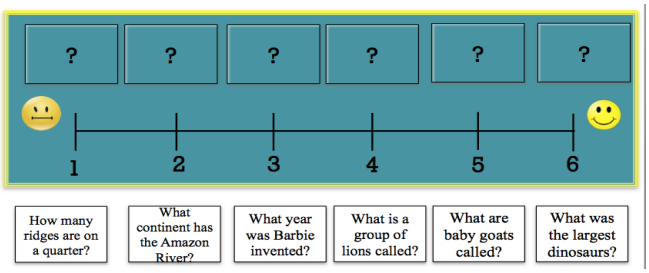


Figure 1: Scale and Questions

Procedure

Each testing session consisted of 4 trials in which the child was presented with 6 questions per trial. Children therefore received a total of 24 questions out of the original 34 questions. Extra questions were included in the case that a child may know some answers to the 24 questions presented to them. Before each trial the experimenter told the child to indicate to the experimenter if she knew the answer to a presented question. If a child thought they knew the answer to a question or if the child guessed either a correct or incorrect answer, a new question was chosen from the top of the deck.

At the beginning of each trial the experimenter shuffled the deck and drew the first 6 question cards from the top of the deck. The experimenter read the cards and placed each question card in front of the child. The experimenter instructed the participant to use the board to rate the questions based on their level of curiosity and guided participants through the rating scale: First the experimenter would ask the child which question they were the most curious to know the answer to and then which question they were the second most curious to know the answer to. Those questions were placed in boxes 6 and 5. The experimenter then asked the child which question she was the least curious about and which question she was the second least curious about. These questions were placed in boxes 1 and 2. For the remaining two questions, the experimenter would ask the child which question she was the most curious about of the two, that question was then put into box 4 and the last question was put in box 3.

After all the questions were rated, the experimenter, starting at one end of the scale, would reread the question and then flip the card over and read the answer. Once all of the answers had been revealed the experimenter put the 6 questions off to the side for later testing. The trial was then repeated and the order of presentation of questions and answers was counterbalanced across trials in order to reduce an possible recency or primacy effects. After two trials, the 6 questions from trial 1 and the 6 questions from trial 2 were combined and shuffled randomly. The child was then asked to answer each question from memory. The child underwent two more trials and one more testing block. All ratings and responses to questions were recorded with a video camera.

Answers were coded as correct if the child provided the correct answer and incorrect if the child did not provide the correct answer or stated that they did not know or that they forgot the answer. Since we used child-friendly trivia questions most of the answers consisted of one word however some of the answers contained two words such as 'argentine lizzard' or 'wild prairie rose'. If a child responded with a key word or phrase such as 'argentine' or 'prairie rose' the response was coded as correct. However, if the child responded with a word from the answer that was not a key word or phrase such as 'lizzard' or 'wild rose' the response was coded as incorrect.

Results

Data analysis showed that there was a significant difference in memory retention between age groups with 8-year-olds showing greater recall than 7-year-olds but age was not a significant factor within either of the two age groups. On average, 7-year olds recalled 60.2% (SD = 0.13) of the answers while 8-year-olds remembered 77.6% (SD = 0.13) and this difference was significant ($t = 2.304$, $df = 12.026$, $p = 0.039$).

Generalized Linear Models with Mixed effects (GLMMs) were used to predict recall based on age group and curiosity rating[1]. Since each child rated multiple questions for each curiosity rating participant ID was entered as a random effect in order to control for individual differences in recall. Results from model comparisons revealed that model with the best fit predicted recall from curiosity ratings, age group and the interaction of curiosity ratings and age group. This model outperformed a null model ($\chi^2 = 12.12$, $df = 3$, $p = 0.006$), as well as models using curiosity ratings alone (($\chi^2 = 10.966$, $df = 2$, $p = 0.004$), age group alone (($\chi^2 = 6.009$, $df = 2$, $p < 0.049$), or the combination of curiosity ratings and age group without interactions (($\chi^2 = 4.855$, $df = 1$, $p < 0.028$).

The coefficients provided in the output of the GLMMs are in the form of the logarithm of the odds that a response is correct at a given value or level of the independent variable. By taking the exponent of the log odds we can see the odds ratio of a correct response. Table 1 provides the odds ratio of coefficients for each of the variables in the model as well as their corresponding 95% confidence intervals.

Analyses of the coefficients from the model with interactions revealed that for 7-year-olds the odds ratio of giving a correct answer increased by only 0.156 for each unit increase in curiosity rating while for 8-year-olds, each unit increase in curiosity rating led to a 1.17 increase in the odds of recalling the correct answer. Children's percentage of correct responses by age group and curiosity rating are presented in figure 2 on the following page.

	Odds ratio	Lower limit	Upper limit
Intercept	1.903	0.0008	4416.062
Curiosity	0.156	0.028	0.864
Age_group	1.015	0.367	2.805
CuriosityXAge	1.288	1.028	1.613

[1] Analyses revealed no significant effects of trial or block order indicating that children in this study did not show an effect of receny or primacy.

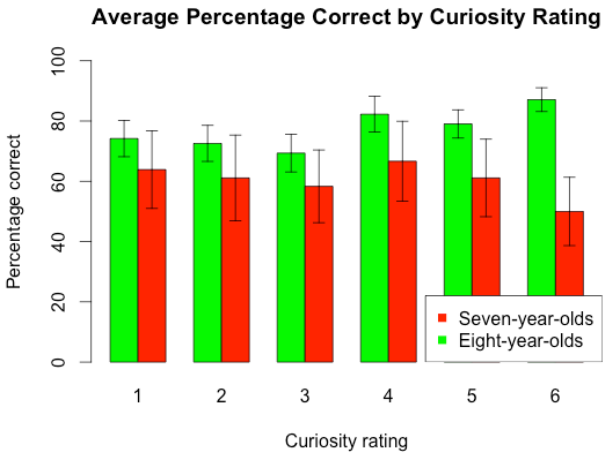


Figure 2: Percentage correct by curiosity rating and age group.

Discussion

Our results showed that by age 8, children's own ratings of curiosity predicted their memory recall. The more curious they are about a trivia fact, the more likely they are to remember the answer. For 7-year-olds, however, no such effects were found. The sample size is small for this study, so we need to interpret our results with caution. On-going work aims to increase the sample size of each age group.

Multiple factors could be causing the difference between the age groups seen in this study. First, memory and attention both improve with age and it is possible that the eight-year-olds' higher performance is due to improvements in these more basic cognitive functions. Secondly, younger children may lack the metacognitive skills necessary for curiosity to effectively influence their learning patterns. Curiosity requires one to first recognize when there is an information gap and then selectively focus attention and use cognitive resources to fill that gap. Although it has been shown that young children have metacognitive skills (Brown, 1977; 1979; Flavell, 1979), they develop with age and therefore, some children may not have the ability to recognize an information gap effectively and then use those metacognitive skills to focus attention on appropriate information. Whatever the reason may be, this experiment while hinting at developmental differences cannot determine the underlying cause of the differences.

In addition, the effect of curiosity on memory may not be very strong due to a variety of other factors. One of them being that children may be more likely to remember answers that have words which are more familiar to them. Some of the questions children are curious about may have answers with words that are not commonly used (for example: argentine lizard), while other less curiosity inducing questions have easier to remember answers (for example: lion). For this reason, future studies will investigate the familiarity of content knowledge as well vocabulary knowledge.

Lastly, some answers to the questions were in the form of numbers. When considering numbers, children are also more likely to forget answers to questions they were highly curious about if the number is hard to remember, such as 119 or 1959. Other answers to lower curiosity questions may be much easier to remember, such as 8.

With a much larger sample size, we hope to be able to reduce the noise level in the data. Since different children choose different questions as high or low curiosity, based on their own prior knowledge and experience, we hope that the effects of these idiosyncratic factors such as familiarity with specific words or numbers will wash out. It also remains to be seen if the age difference we have found will hold up with a larger sample size, and whether other less verbal methods may produce the same results.

Conclusions

In the past researchers and philosophers have struggled to define curiosity in any sufficient manner. With Loewenstein's (1994) information gap theory we can describe it not just as a desire to learn but as a psychological concept with positive results for learning and memory. This study has shown that the benefits of curiosity are not just for adults.

This study has shown that 8-year-old children do seem to have higher memory retention of facts that elicit more curiosity, whereas, 7-year-olds apparently do not. This is important not only for our own curiosity as developmental cognitive scientists but also for others involved in child development such as teachers and parents.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE 1106400) to S.O.

References

Benware, C. A., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*(4), 755-765.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience?. *Brain Research Reviews*, *28*(3), 309-369.

Blackburn, J. R., Phillips, A. G., Jakubovic, A., & Fibiger, H. C. (1989). Dopamine and preparatory behavior: II. A neurochemical analysis.*Behavioral Neuroscience*, *103*(1), 15.

Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning.*Cognitive psychology*, *64*(4), 215-234.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child development*, 1-8.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*,*120*(3), 341-349.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit.*Neuron*, *84*(2), 486-496.

Gweon, H., & Schulz, L. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play. In *Proceedings of the 30th annual meeting of the Cognitive Science Society* (pp. 570-574).

Heffner, T. G., Zigmond, M. J., & Stricker, E. M. (1977). Effects of dopaminergic agonists and antagonists of feeding in intact and 6-hydroxydopamine-treated rats. *Journal of Pharmacology and Experimental Therapeutics*, *201*(2), 386-399.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2009). The wick in the candle of learning epistemic curiosity activates reward circuitry and enhances memory.*Psychological Science*, *20*(8), 963-973.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron*, *46*(5), 703-713.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, *116*(1), 75.

Marvin, C. B., & Shohamy, D. (2016). Curiosity and Reward: Valence Predicts Choice and Information Prediction Errors Enhance Learning. *Journal of experimental psychology*.

Ngo, H. Q., Luciw, M. D., Foerster, A., & Schmidhuber, J. (2012, June). Learning skills from play: Artificial curiosity on a Katana robot arm. In *IJCNN*(pp. 1-8).

Phillips, A. G., Pfaus, J. G., & Blaha, C. D. (1991). Dopamine and motivated behavior: insights provided by in vivo analyses. *The mesolimbic dopamine system: From motivation to action*, 199-224.

Phillips, A. G., Coury, A., Fiorino, D., LePiane, F. G., Brown, E., & Fibiger, H. C. (1992). Self-Stimulation of the Ventral Tegmental Area Enhances Dopamine Release in the Nucleus Accumbens: A Microdialysis Studya.*Annals of the New York Academy of Sciences*, *654*(1), 199-206.

Phillips, A. G., Atkinson, L. J., Blackburn, J. R., & Blaha, C. D. (1993). Increased extracellular dopamine in the nucleus accumbens of the rat elicited by a conditional stimulus for food: an electrochemical study. *Canadian journal of physiology and pharmacology*, *71*(5-6), 387-393.

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, *43*(4), 1045.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, *16*(7), 382-389.

Simansky, K. J., Bourbonais, K. A., & Smith, G. P. (1985). Food-related stimuli increase the ratio of 3, 4-dihydroxyphenylacetic acid to dopamine in the hypothalamus. *Pharmacology Biochemistry and Behavior*, *23*(2), 253-258.

Ungerstedt, U. (1971). Postsynaptic supersensitivity after 6-hydroxy-dopamine induced degeneration of the nigro-striatal dopamine system. *Acta Physiologica Scandinavica*, *82*(S367), 69-93.

Wise, R. A. (1985). The anhedonia hypothesis: Mark III. *Behavioral and Brain Sciences*, *8*(01), 178-186.

The Effects of Discourse Cues on Garden-path Processing

Ana Besserman, Elsi Kaiser {pianibes, emkaiser} @usc.edu
Department of Linguistics, University of Southern California, Los Angeles, CA, 90089

Abstract

We report a self-paced reading study that investigated garden-path sentences like *While the boy washed {a/the} dog barked loudly* and *While the man hunted {a/the} deer ran into the woods.* In such sentences, the critical noun phrase (*dog, deer*) tends to be misparsed as an object of the preceding verb, and has to be re-analyzed as a subject of the following clause when the disambiguating verb (e.g. *barked, ran*) is encountered. To better understand how discourse level information guides real-time processing, we build on earlier corpus work in linguistics which found a relationship between syntactic function and information status: Entities in subject position tend to be already-mentioned (old/given) information and definite, while entities in object position are typically new information and indefinite. We investigated whether the information status of the ambiguous noun influences the extent of processing difficulty, and whether this effect also depends on the argument structure of the first verb. Results from self-paced reading showed that information status matters when processing the ambiguous NP after optionally transitive verbs (e.g. *hunt*) but not after reflexive absolute verbs (e.g. *wash*). These results suggest that access to discourse-level representations during re-analysis of the noun phrase is modulated by verb argument structure.

Keywords: garden-path, information status, definiteness, givenness, verb argument structure, sentence processing

Introduction

Research in sentence processing shows that parsing occurs incrementally and is guided by both bottom-up and top-down information. Indeed, interactive models of sentence processing often assume a principle of immediacy, namely "the idea that every source of information that constrains the interpretation of an utterance (syntax, prosody, word-level semantics, prior discourse, world knowledge, knowledge about the speaker, gestures, etc.) can in principle do so immediately" (Hagoort & Van Berkum, 2007:802). We report a self-paced reading study that aims to shed light on which sources of information constrain utterance interpretation, with a focus on the interplay of discourse-level information and syntactic information. Specifically, we use garden-path sentences to test whether the information status of nouns that are temporarily ambiguous between subject and object influences how they are parsed, and if this is modulated by verb argument structure.

Garden-Path sentences: Lingering misparses

Temporarily ambiguous sentences such as *While the boy washed the dog barked loudly* have been found to cause comprehension difficulties (e.g., Frazier & Rayner, 1982; Ferreira & Henderson, 1991). That is because the noun phrase "the dog" is often initially parsed as the object of the verb "wash" (as in *While the boy washed the dog*). Yet, when hearers reach the verb "barked", they must readjust to this new piece of information – the second verb – by (i) re-analyzing "the dog" as the subject of the second clause (i.e., *the dog barked loudly*) and (ii) re-analyzing the verb "washed" as part of a reflexive structure (*While the boy washed (himself), the dog barked loudly*). The initial misparse of the noun 'the dog' occurs because when comprehenders encounter a noun phrase that can be processed as a direct object of the preceding verb, they have a preference for parsing it as such instead of treating it as the subject of a new clause. That is, if possible, hearers prefer to use available input to continue clauses they are currently processing rather than beginning a new clause. This parsing strategy has been termed "late closure" (Frazier & Rayner 1982). Although there are cues that can help disambiguate the syntactic role of the temporarily ambiguous noun phrase – such as prosody in spoken language or a comma in the written modality (*While the boy washed, the dog barked loudly,* e.g., Kjelgaard & Speer, 1999; Christianson, Hollingworth, Halliwell, & Ferreira, 2001) – extensive research has shown that in the absence of these cues the human parser has a strong bias towards late closure in these kinds of contexts, i.e., towards treating the ambiguous noun as the object of the verb.

Over the last couple of decades studies have also shown that the initial misparse (e.g. *The boy washed the dog)* has persistent effects that linger even after the sentence has been disambiguated (Christianson et al., 2001; Patson, Darowski, Moon, & Ferreira, 2009; Slattery, Sturt, Christianson, Yoshida, & Ferreira, 2013). Data from comprehension questions shows that, even after processing the whole sentence, participants often provide incorrect responses that are compatible with the first (and incorrect) analysis. For example, after reading the sentence "While the boy washed the dog that was white and furry barked loudly", participants were asked "Did the boy wash the dog?" and over 65.6% responded with 'yes' (Christianson et al., 2001), even though the correct answer is 'no.'

These results have been taken to show that the interpretation from the initial parse (i.e. *While the boy washed the dog*…) still lingered after reanalysis. Similar results have been seen in various types of tasks, such as paraphrasing (Patson et al., 2009) as well as with different verb types: Both (i) verbs like 'wash' which are temporarily ambiguous between a transitive interpretation (*The boy washed the dog*) and a reflexive interpretation (*The boy washed (himself)*) and (ii) verbs like 'hunt' which can optionally have an unmentioned/unspecified object (*The man hunted the deer* vs. *The man hunted*) can seemingly result in lingering misinterpretations (e.g. Christianson et al., 2001).

Verb Argument Structure

One of the key issues that our experiment investigates is the contribution that verbs make to real-time parsing – in particular, how and whether differences in verbs' argument structure interact with discourse-level, information-structural information in guiding the parsing of temporarily ambiguous structures. Verbs have a pivotal role in sentence structure: They are responsible for connecting syntactic representations (through grammatical roles, such as subjects and direct objects) to semantic representations (thematic roles, such as agents and patients). Boland (1993:134) notes that "More than any other word in a sentence, the matrix verb defines the situation that sentence describes." Indeed, prior research has shown that verb-specific argument structure information is used immediately during online processing (e.g., Trueswell, Tanenhaus, & Kello, 1993).

Verbs differ in how they can and/or must combine syntactically with other words and phrases. Crucial for our purposes is the observation that some verbs – such as "bathe" and "hunt" – can appear either with an overt object that is pronounced and visible in the surface structure (e.g. *The man hunted the deer*. *The boy washed himself.*) or without an overt object (e.g. *The man hunted. The boy washed.*). In this second case, we refer to the object as *covert*, because it is not pronounced but the existence of an object is still present in the meaning of the verb. Indeed, the fact that some verbs allow both overt and covert objects is what allows a garden-path to emerge in sentences like *While the boy washed the dog barked*, as discussed above.

Prior research on these kinds of garden-path sentences has tested two different types of verbs: (i) Optionally Transitive verbs (henceforth referred to as OPT, e.g. "hunt") and (ii) Reflexive Absolute Transitive verbs (RAT, e.g. "wash"). While both verb types can cause garden-pathing, they differ in fundamental ways in regards to their argument structure, particularly when there is no overt object expressed in the sentence. In such cases, RAT verbs are interpreted reflexively: The interpretation of "The boy washed" is that he washed *himself*. In this case, the object is known: It is coreferential with the subject of the clause. In contrast, OPT verbs without an overt object make reference to an unspecified object. Thus, the interpretation of "The man hunted" is that he must have hunted *something* that was left out of the clause. This unspecified object could be interpreted as an entity in prior or following discourse. Thus, in sentences without overt objects, in the case of RAT verbs the referent of the object is nevertheless known (because it is determined by the syntactic property of reflexivity), but OPT verbs have unspecified objects.

One of the goals of this work is to investigate the consequences of these differences between RAT verbs and OPT verbs for real-time sentence processing. In particular, during reanalysis of garden-path sentences like *While the boy washed the dog barked* or *While the man hunted the deer ran into the woods,* comprehenders must not only re-analyze the ambiguous noun phrase, they must also re-interpret[1] the first clause so that the verb does not have an overt object.

Crucially, we suggest that RAT verbs and OPT differ with respect to how the covert (unpronounced) object is processed. When RAT verbs (e.g. "wash") have covert objects, the verb is interpreted reflexively and its meaning is determined by the syntactic structure of the clause: The object of "wash" is whatever the subject of this verb is. For OPT verbs (e.g. "hunt"), the verb has no specified object: the hearer doesn't know who/what is being hunted, unless they can find a suitable reference in the discourse.

Given this difference between RAT and OPT verbs, we hypothesize that (i) processing the patient/object of RAT verbs operates on the syntactic and semantic levels and can be largely independent of discourse representations, but (ii) processing the patient/object of OPT verbs makes reference to discourse representations and is sensitive to information status.

Information Status

Give our hypothesis that processing the covert object of OPT verbs makes reference to discourse representations, in this section we briefly review existing work on the question of how discourse-level information guides sentence processing. Earlier work has shown that there are many instances in which the parser makes use of top-down information about the discourse to guide processing (e.g. Kaiser & Trueswell, 2004; Boland, 2005; Delong, Urbach & Kutas, 2005).

One of the most important types of discourse-level information has to do with *information status*, e.g., whether a noun refers to an entity that is new information/being mentioned for the first time, or an entity that has already been mentioned in the preceding discourse (old/given information). The garden-path ambiguity presented here makes for an interesting testing ground of the availability of discourse cues during parsing because the temporary ambiguity is in the grammatical role domain: a noun phrase that is first interpreted as a direct object must be re-assigned to the role of subject. Crucially, studies have shown that there is *a strong relationship between syntactic function and information status*: Entities realized in subject position tend to be already-mentioned (old/given) information and definite, while entities realized in object position are more typically new information (being mentioned for the first time) and indefinite (Comrie, 1988; Prince, 1992). This is not a surprising pattern: first, definiteness and givenness correlate highly with one another, i.e., noun phrases that are definite (e.g. *the deer*) have usually been already mentioned in the discourse (discourse-old), whereas noun phrases that are indefinite (e.g. *a deer*) are usually being introduced into the

[1] There has been discussion regarding whether the first clause remains incomplete or not, particularly due to lingering misinterpretations. Slattery et al. (2013) argue hearers achieve complete interpretations consistent with reanalysis of the garden-path, and lingering misinterpretations are likely due not to incomplete parsing but to incomplete *erasure* of the first erroneous parsing. These distinctions are not central for the claims we are making.

discourse for the first time (discourse-new).[2] Secondly, there is a widely recognized bias in English and other languages for old information to precede new information sentence-internally (e.g. Firbas, 1966; Halliday, 1967). Crucially, English has a relatively fixed word order and a strong preference for subjects to occur before objects (Prince, 1981). Consequently, and following the old-before-new bias, subjects are often old and definite, and objects often new and indefinite.

Since syntactic function (subjecthood and objecthood, in this case) correlates with information status (old+definite information and new+indefinite information), one might wonder whether the ease of re-analyzing a noun from object to subject is sensitive to the noun's information status.

Previous experimental work has found contrasting evidence regarding how rapidly definiteness and givenness information is accessed and utilized online. An ERP study by Kirsten, Tiemann, Seibold, Hertrich, Beck, & Rolke (2014) investigated whether participants reacted to unfelicitous uses of definites and indefinite determiners in German (e.g., a definite article introducing an entity that is not uniquely identifiable and had not been previously mentioned) and found *immediate* N400/P600 complex responses as participants read the determiner. On the other hand, an ERP study by Schlueter, Williams & Lau (2015) suggests that hearers do *not* use information about definiteness online to predict the upcoming noun based on previous mention in English, even though there is a strong correlation between definiteness and givenness of the noun. They only found later effects of definiteness information during subsequent integration processes. Thus, the question still remains as to when and how these discourse cues are accessed and utilized during parsing.

Aims of this work, predictions

Using garden-path sentences where the critical noun phrase is temporarily ambiguous between an object and a subject, we used self-paced reading to investigate (i) whether the information status of this ambiguous noun influences the extent of processing difficulty, and (ii) whether this is different for RAT verbs and OPT verbs.

In light of the relationship between subjecthood and definiteness and givenness, we predict that re-analysis from object to subject will be harder – as shown by reading time slowdowns – for nouns that are indefinite and new information than for definite/given nouns. This prediction is based on the corpus finding that objects tend to be new information and subjects to be given information.

However, we predict the impact of information status will be modulated by verb type. Given the differences between RAT and OPT verbs, we hypothesize that (i) processing the object of RAT verbs operates on the syntactic and semantic levels and can be largely independent of discourse representations, but (ii) processing the object of OPT verbs makes reference to discourse representations and is sensitive to information status. In other words, we expect OPT verbs to show more sensitivity to the nouns' information status.

Experiment

Methods

Participants Forty-eight college-aged English native speakers participated for course credit.

Design & Stimuli We conducted a word-by-word self-paced reading experiment using Linger[3]. There were 24 target items, 12 with OPT verbs and 12 with RAT verbs. For consistency, targets were adapted versions of the same items used in previous studies (Christianson et al. 2001, Patson et al. 2009). We used a 3x2x2 design where we manipulated (i) the information status of the critical noun phrase, (ii) verb type and (iii) whether or not the sentence was temporarily ambiguous.

To manipulate givenness (whether the ambiguous NP has been mentioned in prior discourse or not), we added a context sentence that either introduced the critical entity (1b and 2b) or didn't introduce any entities (1a, 1c and 2a, 2c) before the critical sentence.

Definiteness of the ambiguous NP was manipulated to create an indefinite+new condition (1a and 2a), a definite+old condition (1b and 2b), and a definite+new condition (1c and 2c). The definite+new condition is equivalent to prior garden-path studies (Christianson et al. 2001, Patson 2009, Slattery

Figure 1. Example items

Optionally Transitive Verbs (OPT verbs)

(1a) Indefinite + New
It was a beautiful afternoon. While the man hunted (,) **a deer** ran into the woods near the house.

(1b) Definite + Old
A deer was drinking water by the lake. While the man hunted (,) **the deer** ran into the woods near the house.

(1c) Definite + New
It was a beautiful afternoon. While the man hunted (,) **the deer** ran into the woods near the house.

Reflexive Absolute Transitive Verbs (RAT verbs)

(2a) Indefinite + New
It was just another Sunday morning. While the boy washed (,) **a dog** barked surprisingly loudly near the window.

(2b) Definite + Old
A dog was sitting in the yard. While the boy washed (,) **the dog** barked surprisingly loudly near the window.

(2c) Definite + New
It was just another Sunday morning. While the boy washed (,) **the dog** barked surprisingly loudly near the window.

[2] For a more in-depth analysis of definiteness and information status, see Birner & Ward (1993).

[3] D. Rohde, http://tedlab.mit.edu/~dr/Linger/

2013), which used definite NPs without prior mention. (Indefinite+old was not tested, as it is infelicitous/unnatural.)

Lastly, ambiguity of the noun phrase was manipulated so that every condition appeared either with a comma (unambiguous) or without one (ambiguous, the true garden-path). Our use of the comma for disambiguation follows Christianson et al. (2001), Patson et al. (2009) and Slattery et al. (2013). Fig. 1 shows examples of items in all conditions.

In addition to the 24 target items, there were 50 filler items. After every target and filler item, participants saw a yes/no question. For all the targets, the question probed whether the initial misparse was still lingering in comprehenders' minds. For example, for "While the boy washed the dog that was white and furry barked loudly", participants were asked "Did the boy wash the dog?". If participants successfully re-analyze and 'over-write' the initial misparse, the correct response to every target question is "no". Questions about fillers were designed to balance out the number of 'yes' and 'no' responses across the entire experiment.

In sum, we manipulated (i) information status of the ambiguous NP type (Indefinite+New, Def+Old, Def+New), (ii) verb type (RAT/OPT) and (iii) ambiguity (presence vs. absence of a comma). The critical region was the disambiguating verb (e.g. *ran/barked*) which immediately followed the critical NP, and the four words after the disambiguating verb (to detect spillover effects).

Data analysis

We analyzed reading times and question answer accuracy. Reading times faster than 100ms or +/- 3 standard deviations from the mean for any given position were excluded from analysis. This affected 0.07% and 1.95% of the data, respectively. Reading times (continuous data) and answer accuracy (categorical data) were analyzed with mixed-effects regression using R, with random slopes and intercepts for subjects and items when supported.

Results and discussion

Self-paced reading time data

We analyzed reading times in order to see which conditions resulted in relatively higher processing load, in particular at the disambiguation point when re-analysis occurs. Planned pairwise comparisons (using effects coding) of the three information status conditions in RAT verb conditions show a significant ambiguity slowdown at all five critical positions, starting with the disambiguating verb (|t|>2). This is the expected garden-path effect, which can be seen by comparing the solid lines (ambiguous) to the dotted lines (unambiguous) in the boxed region of Fig. 2: Ambiguous sentences, which allowed readers to be garden-pathed, showed a relative slowdown in reading times at and after the disambiguating verb (compared to unambiguous sentences). This slowdown persists for several words. However, we found no significant effects of the critical noun's information status.

Conditions with OPT verbs were analyzed in a parallel way and reveal a more complex pattern. Again, there is a significant ambiguity effect (|t|>2) at all five critical positions: Ambiguous conditions have longer reading times than unambiguous conditions (solid vs. dotted lines in Fig. 3). In addition, there are significant effects involving the information status manipulation on the second word of the spillover region: In particular, conditions with indefinite nouns (triangles in Fig.3) seem to show a bigger ambiguity-related slowdown than conditions with definite nouns. Specifically, at the second word in the spillover region, comparisons of Indefinite+New with both Definite conditions reveal a significant main effect of Ambiguity, a significant main effect of Information Status, as well as a significant interaction (|t|>2): Conditions with indefinite nouns are slowed down by ambiguity more than conditions with definite, previously mentioned nouns.

Figure 2. Reading times: RAT verbs

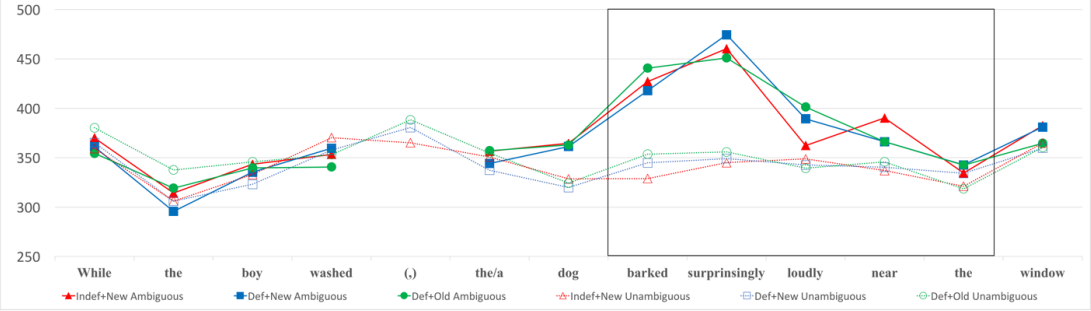
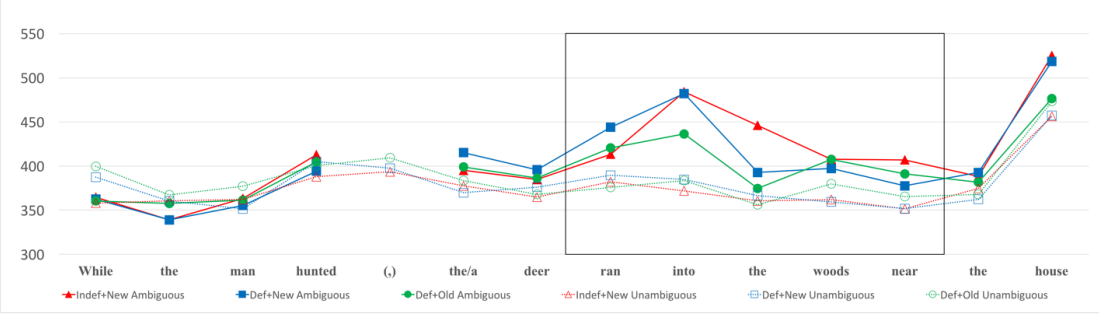


Figure 3. Reading times: OPT verbs



In addition, at the first word in the spillover region, the Definite+Old condition is read numerically faster than both Indefinite+New and Definite+New conditions, although the comparison did not reach significance.

Generally speaking, indefinite nouns seem to suffer a bigger slowdown due to ambiguity than definite nouns – a pattern which could also be rephrased as sentences with definite nouns recovering faster from the garden-path. This fits with our prediction that re-analysis from object to subject will be harder for nouns that are indefinite and new information than for definite/given nouns, which we derived from the finding that definites are more frequent subjects than indefinites (Prince, 1992).

Comprehension questions

In addition to reading time, we also analyzed response accuracy to the yes/no comprehension questions for target items (e.g. Did the man hunt the deer? Did the boy wash the dog?). Following Christianson et al. (2001), we coded 'no' as correct, as it signals that the initial misanalysis was appropriately corrected/abandoned. Unlike reading times, response accuracy does not provide a measure of how easy (or hard) it is re-analyze the critical noun as a subject and is best regarded as a measure of the extent to which the initial noun-as-object misparse persists (or doesn't) in participants' final interpretation of the sentence.

In both the RAT and OPT conditions (Figures 4 and 5), response accuracy is significantly higher with unambiguous than ambiguous sentences (p's<.001), as is to be expected. Furthermore, with OPT verbs, accuracy is significantly higher in the Indefinite+New condition than in the other two conditions (p's<.001). With RAT verbs, accuracy is close to ceiling, especially in the unambiguous condition. Still, accuracy in the Indefinite+New condition is significantly higher than in the Definite+Old condition (p<.01) and numerically higher than in Definite+New (p=.14).

Thus, with both verb types, nouns' information status influences response accuracy, i.e. whether people incorrectly say that the ambiguous noun is the object of the critical verb. (These effects cannot be attributed to garden-pathing; they also appear in unambiguous conditions.) With OPT verbs, why do definites result in more (incorrect) object interpretations than indefinites? We suggest this is because definites imply familiarity and are more likely to be interpreted as coreferential with the covert object of the OPT verb (the man hunted *something*). But indefinites suggest newness and are consequently less likely to be interpreted as referring to the covert object, causing higher accuracy.

The accuracy data for RAT verbs is harder to interpret due to potential ceiling effects, but the general pattern seems to follow a similar trend as the OPT verbs, with higher accuracy in the Indefinite+New condition. These findings again suggest that a noun's information status matters. Although for RAT verbs, the lack of an overt object should not give rise to a search for a referent in the discourse (as they are interpreted reflexively), the error rate in the *unambiguous* condition suggests that this does occur (similarly to other unambiguous RAT conditions in Christianson et al. 2001).

Figure 4. Response accuracy: OPT verbs

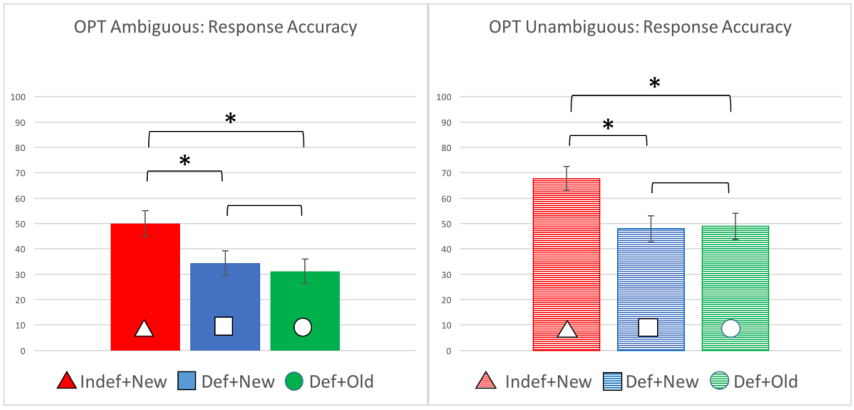
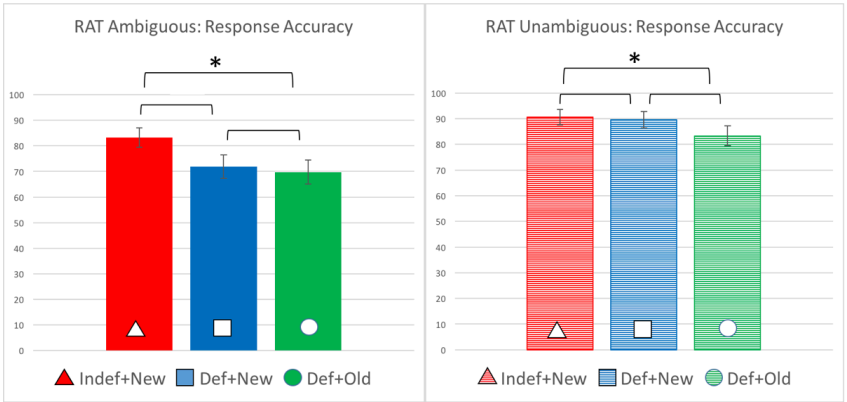


Figure 5. Response accuracy: RAT verbs



General Discussion

We conducted a self-paced reading study using garden-path sentences where the critical noun phrase is temporarily ambiguous between an object and a subject, in order to test whether the information status of this ambiguous noun influences the extent of processing difficulty, and whether this effect is modulated by verb argument structure.

As predicted, re-analysis from object to subject was more difficult – as shown by reading time slowdowns – for nouns that are indefinite and new information than for definite/given nouns. This pattern is a reflection of earlier corpus findings showing that objects tend to be new information and subjects to be given information.

Also in line with our predictions, the influence of information status was modulated by verb type: due to the differences between reading times after RAT and OPT verbs, we found that processing the object of RAT verbs operates on the syntactic and semantic levels and can be largely independent of discourse representations; on the other hand, processing the object of OPT verbs makes reference to discourse representations and thus is sensitive to information status.

Theories of sentence processing must account for the bottom-up and top-down sources of information utilized during parsing. The differences found between the two verb types (OPT and RATs) support the idea that the verbs' lexical and subcategorization information play a crucial role during sentence processing, and modulate which sources of information the processor utilizes in real time. More specifically, verb differences guide access to different levels of linguistic representation: in some instances the discourse level is much more readily employed than in others. This reinforces the idea of an economical parser, that only resorts to information that is relevant in constraining sentence interpretation.

Moreover, the response accuracy data contributes to a fruitful line of research on lingering misinterpretations, suggesting that misinterpretations can also result from discourse-level coreference relations. Definiteness and givenness cues influenced the offline interpretation of the *while*-clause, which we interpret as evidence that discourse representations affect the likelihood that a noun will be interpreted as coreferring with a preceding covert object.

References

Altmann, G. T. M. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30, 191-238.

Birner, B., & Ward, G. (1993). Uniqueness, familiarity, and the definite article in English. *Berkeley Linguistics Society*, 20, 93–102.

Boland, J. E. (1993) The role of Verb Argument Structure in Sentence Processing: Distinguishing between Syntactic and Semantic Effects. *Journal of Psycholinguistic Research*, 22(2), 133-152.

Boland, J. E. (2005). Visual arguments. *Cognition*, 95, 237–74.

Christianson, K., Hollingworth, A., Halliwell, J.F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology,* 42(4), 368-407.

Comrie, B. (1989). *Language Universals and Linguistic Typology*, 2nd ed. Blackwell, Cambridge.

Delong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. Nature Neuroscience 8.1117–21.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30, 725–745.

Firbas, J. (1966). Non-thematic subjects in contemporary English. *Travaux Linguistiques de Prague*, 2, 239-56.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178–210.

Givon, T. (1976). Topic, pronoun, and grammatical agreement. In Li, C., (Ed.) *Subject and topic*. (pp. 149-58). NY: Academic Press.

Hagoort, P., & Van Berkum, J. J. A. (2007). Beyond the sentence given. *Philosophical Transactions of the Royal Society*. Series B: Biological Sciences, 362, 801-811.

Halliday, M. (1967). Notes on transitivity and theme in English, Part 2. *Journal of Linguistics*, 3, 199-244.

Kaiser, E., & Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94(2), 113–147.

Kirsten, M., Tiemann, S., Seibold, V. C., Hertrich, I., Beck, S., & Rolke, B. (2014). When the polar bear encounters many polar bears: event-related potential context effects evoked by uniqueness failure. *Language, Cognition and Neuroscience*, *29*(9), 1147–1162.

Kjelgaard, M. M., Speer, S. R. (1999). Prosodic Facilitation and Interference in the Resolution of Temporary Syntactic Closure Ambiguity. *Journal of Memory and Language*, 40(2), 153–194.

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science*, 18, 226-228.

Patson, N. D., Darowski, E. S., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(1), 280–285.

Prince, E. F. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223-233). New York, NY: Academic Press.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness and information status. In S. Thompson & W. Mann (Eds.), *Discourse description: Diverse analyses of a fundraising text* (pp. 295-325). Philadelphia, PA: John Benjamins.

Rohde, H. & Horton, W. (2014). Anticipatory looks reveal expectations about discourse relations. *Cognition*, 133(3), 667-691.

Schlueter, Z., Williams, A., Lau, E. (2015*). How quickly is definiteness information incorporated into comprehender expectations?* CUNY poster. University of Southern California, Los Angeles.

Slattery, T., Sturt, P., Christianson, K., Yoshida, M., and Ferreira F. (2013). Lingering Misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*, 69, 104-120.

Trueswell, J.C., Tanenhaus, M.K., Kello, C. (1993). Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3), 528-553.

Children's Use of Orthographic Cues in Language Processing

Takayo Sugimoto (takayosg@vega.aichi-u.ac.jp)
Aichi University
1-1 Machihatakemachi, Toyohashi,
Aichi, 441-8522 Japan

Abstract

Rendaku, or sequential voicing, is a morphophonemic process in Japanese in which the voiceless word-initial consonant of the second element (=E2) of a compound word becomes voiced (e.g., ori + **k**ami →ori**g**ami, 'folding' + 'paper' → 'paper folding', /k/ becomes /g/). In adult grammar, *rendaku* is subject to two conditions: It applies if and only if (a) E2 is a *Yamato* word (native vocabulary) in the lexicon and (b) it contains no voiced consonant (e.g., b, d, & g). Recent psycholinguistic studies have revealed that Japanese-speaking preschoolers do not follow adult's grammar; they develop their original prosodically-based *rendaku* processing strategy (preschooler-specific *rendaku* strategy). Their strategies qualitatively change in the early middle childhood to be adult-like *rendaku*, creating a discontinuity in children's word-processing strategies. This study investigated factors responsible for this developmental discontinuity. We conducted an experiment using cross-modal linguistic stimuli (prosody & orthography) to see whether children's orthographic knowledge affects their *rendaku* strategy or not. Our results showed that orthographic cues affected literate children's *rendaku* processing. They were aware the correspondence between types of orthography and word categories in Japanese.

Keywords: preschoolers; *rendaku*; orthography; word category; pitch accent;

Introduction

Rendaku is a morphophonemic process in Japanese that the initial voiceless consonant of the second element (E2) of a compound becomes voiced (Vance, 2015).

(1) Compound Word Formation and *Rendaku*

E1	+	**E2**	→	Compound Word	
ori	+	**k**ami	→	ori**g**ami	**/k/ →/g/**
'folding'		'paper'		'paper folding'	

Rendaku has long been studied based on the data from adult subjects (Labrune, 2013). Neuropsychological studies have shown that *rendaku* is not just a matter of pronunciation such as English consonantal assimilation (e.g., tran**s**cript [s] vs. tran**s**action /s/→[z]); *rendaku* has functions of signaling the syntactic (grammatical), semantic (=meaning) aspects of E2 within a compound (Ogata, E., Hayshi, R., Imaizumi, S., Hirata, N., & Mori, K., 2000).

Thus, studying children's *rendaku* acquisition should lead to our deeper and comprehensive understanding of language development and cultural influences on our language faculty.

In adult grammar, *rendaku* is generally subject to 2 conditions (Ito & Mester, 1986, 1995, Fukuda & Fukuda, 1999, Fukuda, 2002) [1]:

(2) *Rendaku* conditions
Rendaku applies if
(a) E2 is a *Yamato* morpheme (native vocabulary)
and
(b) E2 contains no voiced obstruent in it (Lyman's Law).

The *rendaku* conditions seems complex and difficult for children to acquire since in order to be able to apply *rendaku* properly, children must know the lexical strata and Lyman's Law. Lexical strata are word categories stratified in the Japanese lexicon: *Yamato*, Sino-Japanese, foreign loans, and onomatopoeia (Ito & Mester, 1995, McCawley, 1968). Lyman's Law (Ito & Mester, 1995, Vance, 2015, and many others), which prohibits more than one voiced consonant within a morpheme (an element or unit of a word), is active only in the native vocabulary. This means that children must know which word belongs to the native vocabulary.

Children also have to parse the E2 and decide the applicability of *rendaku* immediately. How do children acquire this seemingly complex knowledge about *rendaku*? Do they know these two conditions from the beginning? Or do they learn them in some order?

Rendaku acquisition studies have assumed that children learn the adult's *rendaku* conditions (Fukuda & Fukuda, 1999, Fukuda, 2002). However, recent psycholinguistic studies have revealed that children do not acquire *rendaku* along with the adult's grammar. Instead, Japanese-speaking preschoolers develop a prosodically-based *rendaku* strategy (preschooler-specific *rendaku* strategy, Sugimoto, 2013a).

(3) Preschooler-specific *rendaku* strategy
Apply *rendaku* if E2 is an unaccented word.

The preschooler-specific *rendaku* strategy is also observed in English-Japanese simultaneous bilinguals (Sugimoto, 2015a). Sugimoto (2015b) reported that

[1] These conditions are not strict restrictions and there are some exceptional cases. For the purpose of conducting experiments, we assume and start from these conditions.

preliterate blind children also show the prosodically-based preschooler-specific *rendaku* strategy.

Japanese is a language with a pitch accent system. Pitch accent is a prosodic feature of a word and it differentiates the meaning of each Japanese word. Pitch accent can be divided into two types: Accented and unaccented. Accented words have a tonal rise followed by a sudden fall. Unaccented words have no such a tonal (rise & fall) contour.

Preschooler-specific *rendaku* processing reflects the distribution of pitch accent types of native vocabulary, to which *rendaku* applies (Table 1). Children are aware that *rendaku* does not apply every word; it applies a particular category of words. They actively construct their own *Rendaku* rule, making best use of their knowledge: Apply *rendaku* if E2 is an unaccented native word.

Table 1: Word categories and pitch accent

Pitch accent / Lexical strata	Accented words[2] (examples)	Unaccented words
Yamato words (=native vocabulary)	29% (ka'rasu 'crow')	71% (sakana 'fish')
Sino-Japanese words (vocabulary of Chinese origin)	49% (hu'dan 'daily')	51% (saihu 'purse')
Loan words (excluding Sino-Japanese)	93% (ke'eki 'cake')	7% (pijano 'piano')

(Sugimoto, 2016, data originally from Kubozono, 2006, p180)

The commonly observed children's prosodically-based *rendaku* processing strategy changes in quality around in the early middle childhood to be adult-like *rendaku*, creating a discontinuity in children's *rendaku* strategies (Sugimoto, 2015b). School age children no longer depend on prosodic information of E2. Their *rendaku* patterns become similar to those of adults being independent from the prosody of E2. These qualitative change in children's *rendaku* processing strategy create a developmental discontinuity.

It is easy to imagine that some developmental changes and learning outcomes during these years might motivate children's *rendaku* strategy at around their entry into elementary school.

What kind of development or learning motivates and determines such changes or a discontinuity?

One possibility would be the influence of literacy. Since three types of Japanese characters represent lexical strata, it could be likely that children become well aware of Japanese lexical strata and word categories along with usages of Japanese *hiragana*, *katakana*, and kanji characters. If literacy affects children's *rendaku* processing strategy, how and why?

The purpose of this study was to investigate factors responsible for this developmental discontinuity. We conducted an experiment using cross-modal linguistic stimuli (prosody & orthography) to see whether literacy affects children's *rendaku* processing strategy or not.

[2] The apostrophe in Japanese accented words (e.g., ka'rasu (=crow)) indicates the position of pitch accent.

(4) Research questions

Q1. Do literate preschoolers know the correspondence between types of orthography and word category?

Q2. How do they use their orthographic knowledge in their *rendaku* processing?

(5) Working Hypothesis

If literate preschoolers use their orthographic knowledge in *rendaku* processing, then their *rendaku* patterns should differ, depending upon types of orthography given.

Method

We used a compound noun formation task (Nicoladis, 2003, Sugimoto, 2013a&b, 2016) to see Japanese-speaking preschoolers' language processing strategy described below.

Participants and ethical considerations

The total number of 73 six-year olds with *hiragana* & *katakana* literacy living in the Tokyo dialect area participated in our study. 40 children were assigned to the no orthography condition (control group); 16 children were in the *hiragana* condition; 17 children were in the *katakana* condition.

Prior to the study, we obtained written form of parental permission from all the participants.

Procedure

Children were tested individually in a quiet room. We went through three trials: 4 warm-up trials, 4 comprehension trials, and finally the 16 test trials (E1=*hima'wari*). In each trial, children were shown 3 types of pictures on a laptop computer, E1 & E2 in a random order. The experimenter read aloud E1 and E2 and then asked the children to name picture C (compound noun).

(6) Compound noun formation task (Sugimoto, 2016)

E1	+	E2	→	C (Compound noun)
hima'wari	+	ka'rasu	→	himawari**ga**'rasu
'sunflower'		'crow'		'sunflower crow'



The experimenter read aloud the following statements followed by a question as in (7).

(7) Instructions in the experiment

E1: Kore-wa himawari-desu. 'Here's a *hima'wari*'

E2: Kore-wa karasu-desu. 'Here's a *ka'rasu*.'

C: Koreni namae-o tuketekudasai.
'How would you name it?'

Compound noun formation task with cross-modal linguistic stimuli : Visual(orthography) & auditory (speech) information

We used a modified version of compound noun formation task, cross-modal stimuli, which used two types of visual stimuli (picture and orthography) and auditory stimuli (experimenter read aloud the stimuli).

Design and material

The experiment used a 3 factor inter-subject design: 2(pitch accent)×2(word)×2(orthography). First, the pitch accent of E2 was controlled to two types: unaccented vs. accented (Kubozono, 2006). We used known words and novel words. As for novel words, we counterbalance poss

In order to counterbalance possible phonotactic effects, we divided each condition group into two subgroups and switched types of pitch accent assignment between the subgroups of each condition (see Tables 2 &3 in Appendix).

We also used three types of Japanese orthography conditions (visual stimuli) as in (8).

(8) 3 orthography conditions
A. No orthography (control)
B. *Hiragana* condition
C. *Katakana* condition

Visual stimuli for each of these three conditions are illustrated in Figures 1a-c, respectively.

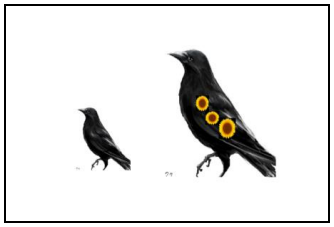


Figure 1a: No orthography condition (control group)

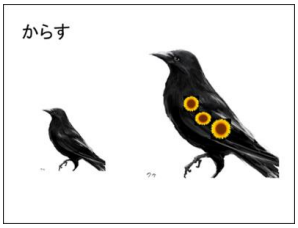


Figure 1b: *Hiragana* condition (からす= 'crows' in *hiragana*)

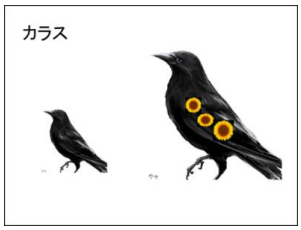


Figure 1c: *Katakana* condition (カラス＝ 'crows' in *katakana*)

Measures

During the experiment, we used a SONY IC recorder (ICD SX-1000) and recorded the children's utterances. After the experiment, two people, one of whom was the author, listened separately to the recordings and transcribed them, judging whether or not the children voiced the target consonant. The reliability of the judgments (Cohen's kappa coefficient (Omura, 2000)) was calculated. The agreement ratio was κ =.96, which is considered highly reliable.

Our scoring was the following. All of the 16 compound nouns in the test trial are subject to *rendaku* (see Table 2 in Appendix). When a subject voiced the morpheme-initial obstruent of E2, then we judged that he/she had correctly applied *rendaku*; for each compound, one point was added to the score. We calculated the total scores and subtotals by condition (pitch accent, orthography, and word type).

Results

Descriptive statistics is shown in Table 4. A three-way ANOVA found a significant interaction of pitch accent and orthography [$F(2,70)=15.061$, $p<.001$, $\eta^2=.301$]. The simple effect of pitch accent was significant in the control group and in the *hiragana* condition [$F(1,70)=93.092$, $p<.001$, $\eta^2=.571$], [$F(1,70)=4.750$, $p=.033$, $\eta^2=.064$], respectively. But the simple effect of pitch accent was not significant in the *katakana* condition [$F(1,70)=.20$, $p=.888$, $\eta^2<.0001$].

Table 4: Descriptive statistics of *rendaku* score

	Condition	Mean	SD	N
Unaccented words（８）	Control	6.2	1.244	40
	Hiragana	3.38	2.391	16
	Katakana	1.82	1.286	17
	Total	4.56	2.444	73
Accented Words（８）	Control	3.58	1.893	40
	Hiragana	2.44	2.366	16
	Katakana	1.76	1.888	17
	Total	2.9	2.122	73

(8 pts. are the maximum score for each pitch-accent type)

Multiple comparisons of within-condition show significant differences between unaccented words and accented words in the control and the *hiragana* conditions (Figure 2). In the two conditions, children used the prosodically-based preschooler-specific *rendaku* strategy. On the other hand, the *katakana* condition show no significant difference between the two pitch accent types ($p<.001$, 95%*IC*: .891-774). Thus, children in the *katakana* condition did not used the preschooler-specific *rendaku* strategy. Compared to the other two conditions, their *rendaku* processing rates in the *katakana* conditions were inhibited, but when we look at the within-group difference, children in the *katakana* condition applied *rendaku* to accented words more often than unaccented words.

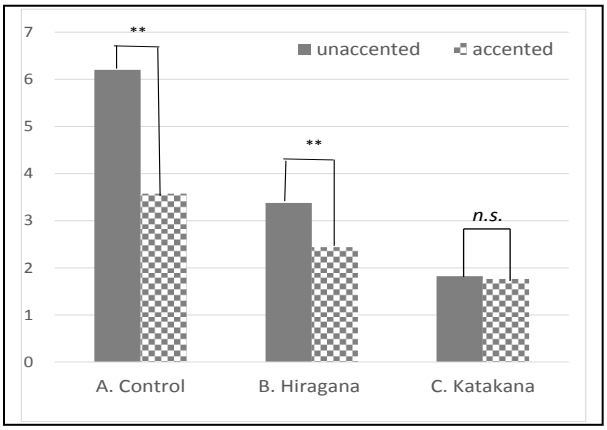


Figure 2: *Rendaku* Score by Condition (8pts.)

Discussion

The significant interaction of pitch accent and orthography conditions, which indicates that literate preschoolers were affected by orthographic cues and showed different *rendaku* processing. Children in the no orthography condition (control) and in the *hiragana* condition showed similar *rendaku* patterns: apply rendaku to unaccented E2s. But children in the *katakana* condition showed a different *rendaku* pattern: they did not apply rendaku to unaccented words, contrary to the other two groups. When *katakana* is given, which represents foreign loan words in Japanese, children's *rendaku* processing was inhibited. Children in (c) did not to apply *rendaku* to non-native vocabulary as much as those in the other conditions did.

It follows from this that the *katakana* condition made children assume the word stimuli were nonnative vocabulary. *Katakana* is used for foreign loan words in the Japanese writing system. Children seem to know *katakana* is used for non-native vocabulary, and judged that the stimuli (both existent native vocabulary and novel words were given) were non-native vocabulary, to which *rendaku* does not apply.

Our results suggest that literate preschoolers used orthographic cues in *rendaku* processing. Preschoolers used different strategies, depending on types of orthography provided in the stimuli.

Preschoolers in the 'no orthography' and the *hiragana* conditions showed the similar tendency: they applied rendaku to unaccented words more often than accented words while those in the *katakana* condition seemed reluctant to apply it. It follows from these that children attend to orthographic cues but children still use prosodically-based rendaku strategy.

Table 5: Orthographic cues and children's application of *rendaku*

Condition	Application of *rendaku*	Use of preschooler-specific strategy
No orthography (pictures only)	Yes	Yes. U > A
Hiragana	Yes	Yes. U >A
Katakana	Yes, but highly inhibited	**No. U≒A**

(U = unaccented words, A= accented words)

Why, then, did preliterate preschoolers in the *katakana* condition were not willing to apply rendaku to unaccented words more often than accented words? We can think of the following possibilities. Preliterate children may be aware of some correspondence between types of orthography and word categories. That is, they may know *hiragana* is used for a certain category of words and *katakana* for another. They may also be aware that *katakana* is used to represent non-native Japanese vocabulary, to which rendaku does not apply.

Preschoolers first define the *rendaku* word category based on prosodic information, that is, pitch-accent (preschooler-specific *rendaku* strategy). We only found that our literate preschoolers can differentiate *rendaku* strategies based on a rough *katakana* vs. non-*katakana* distinction, not *hiragana* vs. *katakana* distinction. We need their longitudinal data to consider orthographic effects on the developmental changes in children's language processing.

Conclusion

Japanese speaking preschoolers, when they have no orthographic knowledge, first define the *rendaku* category based on pitch-accent, i.e., preschooler-specific *rendaku* strategy. After acquiring orthography, children seem to gradually change their *rendaku* processing strategy, using information such as the relationship between orthography and word category. Something more is needed for their redefinition of the *rendaku* category, which may cause the developmental change in their *rendaku* processing strategy. We need developmental paths of individual children.

Acknowledgments

The author is grateful for all the participants of this study.
This work was financially supported in part by JSPS KAKENHI Grant #26580080 and "The Japanese Lexicon: A *Rendaku* Encyclopedia" (Project leader: Prof. Timothy J. Vance, National Institute for National Language and Japanese Linguistics).

References

Ambridge, B. & E. Lieven. (2011). *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge: Cambridge University Press.

Bauer, L. (2011). Typology of compounds. In Lieber & Štekauer (eds.), *The Oxford Handbook of Compounding*. Oxford.

Clark, E. et al. (1985). Compound Nouns and Category Structure in Young Children. *Child Development* 56, 84-94.

Clark, E. & Berman, R. (1987). Types of linguistic knowledge: interpreting and producing compound nouns. *Journal of Child Language*, 14, 547-567.

Devescovi, A. et al. (2005). A Cross-linguistic study of the relationship between grammar and lexical development. *Journal of Child Language* 32, 759-789.

Di Sciullo, A. (2009). Why are compounds a part of human language? A view from asymmetry theory. In Lieber & Štekauer (eds.), *The Oxford Handbook of Compounding*, 145-177. Oxford: Oxford University Press.

Endo, T. (2000). Observation methods. In Omura, A. (ed.), *Research methods in educational psychology (=kyouikusinrigaku-no gihou)*. Tokyo: Fukumura Shuppan.

Fukuda, S.E. & Fukuda, S. (1999). The Operation of *Rendaku* in the Japanese Specifically Language Impaired: A Preliminary Investigation. *Folia Phoniatr Logop* 51, 36-54.

Fukuda, S.E. & Fukuda, S. (2001). An Asymmetric Impairment in Japanese Complex Verbs in Specific Language Impairment. *Cognitive Studies*, 8-1, 63-84.

Ito, J. & Mester, A. (1986). The phonology of voicing in Japanese: Theoretical consequences for morphological accessibility. *Linguistic Inquiry*, 17, 49-73. MIT.

Ito, J. & Mester, A. (1995). Japanese Phonology. In Goldsmith, J. (ed.) *The Handbook of Phonological Theory*. Oxford: Blakwell.

Ito, J. & Mester, A. (2001). Alternations and distributional patterns in Japanese phonology. *Japanese Journal of Phonetics*, 5-2, 54-60.

Labrune, L. (2013). A cross-linguistic approach to *rendaku*-like compound-markers, with special reference to Korean and Basque. Talk presented at 3rd International Conference on Phonetics & Phonology, Tokyo.

Kubozono, H. (2006). Where loanword prosody come from? A case study of Japanese loanword accent. *Lingua* 116, 1140-1170.

Kubozono, H. (2011). Japanese pitch accent. In van Oostendrop, M., Ewen, C., Hume, E., & Rice, K. (eds.) *The Blackwell companion to phonology*. Vol.5, 2879-2907. Oxford: Blackwell.

McCawley, J. (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.

Mckean, C. et al. (2012). Functional reorganization in the developing lexicon: Separable and changing influences of lexical and phonological variables on children's fast-mapping. *Journal of Child Language* ,39, 1-29.

Nakamura, C, M. Arai, & R. Mazuka. (2012). Immediate use of prosody and context in predicting a syntactic structure. *Cognition* 125: 317-323.

Nicoladis, E. (2003). What compound nouns mean to preschool children. *Brain & Language*, 84, 38-49.

Nicoladis, E. (2007). Preschool children's acquisition of compounds. In Libben & Jarema (eds.), *The Representation and Processing of Compound Words*. Oxford: Oxford University Press.

Ogata, E., Hayshi, R., Imaizumi, S., Hirata, N., & Mori, K. (2000). Neural Processing Mechanism for *Rendaku* and Accent Rules in Compound Word Recognition. *Onsei*, 99 (678), 17-24, IEICE.

Ota, M. (2003). The Development of Lexical Pitch Accent System: An Autosegmental Analysis. *Canadian Journal of Linguistics/ Revue canadienne de linguistique*, 48(3/4), 357-383.

Sato, Y., Y. Sogabe, R. Mazuka. (2010). Development of hemispheric specialization for lexical pitch accent in Japanese infants. *Journal of Cognitive Neuroscience, 22*, 2503-2513.

Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. *Journal of Child Language* 38, 1-34.

Sugimoto, T. (2013a). The role of pitch accent in the acquisition of *Rendaku*. Poster presented at the 3rd International Conference on Phonetics and Phonology, Tokyo.

Sugimoto, T. (2013b). The acquisition of *Rendaku*: The lexical strata and the Lyman's Law. *Proceedings of the 146th Annual Meeting of the Linguistic Society of Japan*, 67-72.

Sugimoto, T. (2015a). Development of preschooler-specific language processing strategy: A longitudinal study. *Proceedings of the 25th Annual Meeting of Japan Society of Developmental Psychology*, 238. University of Tokyo.

Sugimoto, T. (2015b). Children's language processing during the transition from early childhood to middle childhood. *Proceedings of the Annual Meeting of Japan Society of Educational Psychology*, 332. Niigata University, Japan.

Sugimoto, T. (2015c). Children's acquisition of *rendaku*. Talk presented at *Rendaku* Workshop. University of Bordeaux, France.

Sugimoto, T. (2015d). Developmental aspects of preschoolers' language processing and their acquisition patterns of *rendaku*. *Journal of University of Tokyo Graduate School of Education* Vol. 54. 261-270.

Sugimoto, T. (2016). Acquisition of *rendaku* by bilingual children: A Case study. *NINJAL Research Papers11*, 1-9.

Vance, T. (2015). *Rendaku*. In H. Kubozono (ed.), *Handbook of Japanese Phonetics and Phonology*, 397-441. Berlin: Walter de Gruyter.

Appendix

Table 2: List of stimuli for 16 E2s used in the test trial

	8 known words		8 novel words (Old Japanese)	
	Accented	Unaccented	Accented/Unaccented[#]	Unaccented/Accented [#]
1	ta'nuki 'racoon dog'	sakura 'cherry blossom'	tokama 'reaping hook '	tekona 'virgin'
2	ka'rasu 'crow'	tsukue 'desk'	hikime 'long arrow'	hokai 'food container'
3	ho'uki 'bloom'	hatake 'field'	sasara 'an old musical instrument'	tatara 'bellows'
4	ho'taru 'light bug'	kuruma 'car'	koromo 'Kimono'	hokora 'god's palace'

(E1: himawari 'sunflower'; E2: 16 words listed in Table 2)

[#] For 8 novel words, we created two types of pitch accent assignment patterns to counterbalance possible phonotacitic effects (see Table 3 below). In particular, we divided children in each of three conditions into two groups and used different pitch accent assignment for each group in each condition.

Table 3: Two types of pitch accent assignment

Groups / Novel word stimuli (E2)	Group A of each condition	Group B of each condition
tekona, hokai, tatara, hokora	Accented e.g., te'kona (antipenulte)	Unaccented e.g., tekona (no pitch accent)
tokama, hikime, sasara, koromo	Unaccented e.g., tokama (no pitch accent)	Accented e.g., to'kama (antipenulte)

Modeling language discrimination in infants using i-vector representations

M. Julia Carbajal [1] **(carbajal.mjulia@gmail.com)**
Radek Fér [2] **(ifer@fit.vutbr.cz)**
Emmanuel Dupoux [1] **(emmanuel.dupoux@gmail.com)**

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS; 29, rue d'Ulm
75005 Paris, France
[2]Speech@FIT, Faculty of Information Technology, BUT; Božetěchova 2
612 66 Brno, Czech Republic

Abstract

Experimental research suggests that at birth infants can discriminate two languages if they belong to different rhythmic classes, and by 4 months of age they can discriminate two languages within the same class provided they have been previously exposed to at least one of them. In this paper, we present a novel application of speech technology tools to model language discrimination, which may help to understand how infants achieve high performance on this task. By combining a Gaussian Mixture Model of the acoustic space and low-dimensional representations of novel utterances with a model of a habituation paradigm, we show that brief exposure to French does not allow to discriminate between two previously unheard languages with similar phonological properties, but facilitates discrimination of two phonologically distant languages. The implications of these findings are discussed.

Keywords: language discrimination; speech; acoustics; computational models; habituation

Introduction

When infants acquire their first language, they meet the formidable challenge of dealing with massive variability and ambiguity at all levels of acoustic and linguistic structure. Infants growing up in a multilingual environment must face an additional level of variability due to the presence of two (or more) languages with independent yet partially overlapping acoustic and structural properties. Although the task may seem hard, a large number of studies show that the ability to discriminate spoken languages is present early on in life (Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998; Nazzi, Jusczyk, & Johnson, 2000; Bosch & Sebastian-Galles, 2001; Ramus, 2002; Byers-Heinlein, Burns, & Werker, 2010). For example, using a habituation paradigm, Mehler et al. (1988) showed that French newborns, in spite of their brief experience with language, are able to discriminate their native language from a foreign one (in this case, Russian) as evidenced by an increase in their arousal following a switch from Russian to French utterances. This discrimination was still observed when infants were presented with low-pass filtered speech, and a preference for their native language was suggested by an asymmetry in the arousal depending on the language presented during habituation.

Further research extended these findings, supporting the claim that newborns can distinguish any two unheard languages if they belong to different rhythmic classes, such as Japanese and English, but that they fail to do so if they belong to the same rhythmic class, e.g., English and Dutch (Nazzi et al., 1998; Byers-Heinlein et al., 2010). These results point at prosody as a strong cue for language discrimination at an early developmental stage. However, languages often differ in many other dimensions, such as their phonemic inventories and phonotactic rules. These cues may become relevant through further exposure to one or more languages and thus facilitate their discrimination: by 4 to 5 months of age, both monolingual and bilingual infants can discriminate two languages even within the same rhythmic class, such as Spanish and Catalan, if they were exposed to at least one of them before (Nazzi et al., 2000; Bosch & Sebastian-Galles, 2001).

While these studies suggest that language distance plays an important role, the specific acoustic features and mechanisms that may allow language discrimination throughout the first year of life, and the impact of prior exposure to one or more languages, are not yet fully understood. In the present study we explore how state-of-the-art speech technology tools can help us understand this feat. As a first step in the application of these novel techniques to the study of infant perception, we propose the use of *i-vectors*, a method to represent any given utterance as a pattern of deviations from a previously constructed background acoustic distribution, to implement an unsupervised model of language discrimination. The i-vector representation, in combination with discriminative classifiers, was originally developed for automatic Speaker Recognition (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010), and in recent years has been adapted to Language Identification systems showing excellent performance (Martínez, Plchot, Burget, Glembek, & Matějka, 2011). These models are typically trained on large datasets containing many different speakers/languages to capture all possible sources of variability. Here, we simplify the model to represent the brief experience of an infant exposed to a single speaker of French, and then test the system's ability to discriminate new unheard utterances of two languages that differ in many phonological dimensions, such as rhythm, syllabic structure and phonemic repertoire (French and English), and two languages with largely overlapping phonologies (Spanish and Catalan). As most studies of language discrimination have made use of habituation paradigms, we also propose a computational model of the habituation task, which will allow us to compare the performance of our system with what has been observed in young infants.

The remainder of the paper unfolds as follows. We first introduce the concept of Universal Background Model and i-vector representation, discussing how these representations can be adapted to model infants' experience. Next, we describe the datasets that we selected for the modeling of the background space and the language discrimination tests. Then, we present a model of the habituation task that uses the extracted i-vectors as input, and two additional measures of discriminability. Finally, the results are described and discussed with respect to current experimental data, followed by a perspective on future work.

Methods

Universal Background Model and i-vectors

The first step of the modeling consists in constructing a representation of the acoustic space formed through the infant's exposure to a given linguistic environment, i.e., their "native" language. To model the distribution of speech features, speech technologies typically use Gaussian Mixture Models (GMM). With a sufficient number of mixture components, GMMs can model any arbitrarily complex distribution. The typical number of components for a Language Identification (LID) system is around one thousand.

The parameters (weights, means and covariances) of the model can be estimated by Maximum Likelihood using an Expectation-Maximization algorithm (Bishop, 2006). A GMM trained on a large database of several hundred hours of speech containing many different speakers, languages and other sources of variability, can be used to represent the overall feature distribution. In the context of speaker and language recognition, this is called the *Universal Background Model* (UBM). Evidently, young infants cannot count on such a large and variable amount of data to build their representations of the acoustic space, however, nothing prevents UBMs from being trained on a much smaller dataset. In the present study, we train a small UBM with speech from one single French speaker to represent the brief exposure that even a 4-day-old infant may have already encountered.

Once the UBM has been trained, data-specific models representing feature distributions of different utterances can be derived from the UBM by Maximum a Posteriori (MAP) adaptations. Usually, only the component means are shifted during the adaptation. Using factor analysis, the adaptation offset with respect to the UBM can be confined to a low-dimensional subspace, called the Total Variability space. If we denote by $\mathbf{m}$ the stacked vector of UBM component means, the generative subspace model has the form:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{v},$$

where $\mathbf{T}$ is a low-rank matrix (Total Variability matrix) defining the bases for the subspace, and $\mathbf{v}$ is a hidden variable with standard normal prior. As with the UBM, this subspace is typically trained on a large number of speech recordings using EM algorithms (Dehak et al., 2010), but for the purpose of our model it will be trained on the data of a single speaker.

Finally, given an utterance or any other segment of a speech recording, the posterior distribution of the hidden variable can be estimated. The MAP point estimate of this distribution is conventionally called an *i-vector*, and can be used as a low-dimensional fixed-length representation of the speech segment. In other words, any unheard utterance can be approximated as a deviation from the background "native" model. We propose to use this simple representation to model the infant's acoustic perception of previously unheard speech, computing an i-vector for every utterance in our test dataset. The advantage of this vectorial representation of speech is that a measure of distance can be defined between any two utterances.

In LID systems, the typical dimensionality of the subspace is around 400. However, for our experiments, the i-vector dimensionality is set to 200, and we use a UBM with 256 mixture components and diagonal covariance matrices. The reason for such a small model is that the database we propose to use in order to model a brief exposure to French is not large enough to robustly estimate all the parameters of a conventional LID model. Furthermore, since our database contains only a limited amount of variability (UBM trained on one single speaker and language), it is unnecessary to increase the number of dimensions.

We argue that i-vectors are reasonable as models of infants' representation of languages for the following reasons: (1) The entire pipeline (construction of UBM and i-vector extraction) only requires two skills, which have been documented in infants: a good acoustic perception (Eimas, Siqueland, Jusczyk, & Vigorito, 1971), and the ability of performing statistical learning (Saffran, Aslin, & Neport, 1996; Maye, Werker, & Gerken, 2002). (2) The learning algorithm is completely unsupervised, requiring no external information about phonemes or words, nor any information about speaker identity, or number and properties of different languages. The only linguistic hypotheses of this model are that utterances are relevant units for performing language discrimination, that they can be modelled through gaussian mixtures, and that they can be segmented out of continuous speech.

Feature extraction A common representation of the acoustic features of a speech signal used in many speaker and language identification systems are *Mel-Frequency Cepstral Coefficients* (MFCCs), which are based on a transform of the power spectrum on a frequency scale that approximates human auditory perception. For our modeling purposes, these features were calculated using the HTK Speech Recognition Toolkit (Young et al., 2006) in 25 ms windows with a 10 ms shift. We retained the first 7 coefficients (including *C0*, which represents the energy) and added a measure of *F0* (pitch) computed with the Kaldi Toolkit (Povey et al., 2011).

In addition, *Shifted Delta Coefficients* (SDC, a stacked version of delta coefficients calculated across several frames, Torres-Carrasquillo et al., 2002) were included to capture the temporal evolution of the MFCC-F0 features. The SDCs were calculated using the usual 7-1-3-7 configuration, resulting in

an approximation of the contour of the MFCC-F0 features over a span of 200 ms. The resulting 64-dimensional MFCC-F0-SDC vectors contain both spectral and prosodic information presumed available to the human auditory system.

Materials

Training data In order to train the UBM to represent the prior experience of an infant with a brief exposure to French, we used a dataset of casual speech recorded from an adult female French speaker selected from the *Corpus of Interactional Data* (Bertrand et al., 2008). The selected dataset is composed of 602 pre-segmented utterances with a mean length of 2.54 seconds ($min = 0.43$ s, $max = 9.01$ s), giving a total of approximately 25 minutes of clean speech. The original recordings were downsampled to 16kHz.

Test data Similarly to previous experimental studies, to test the discrimination of languages we recorded two proficient bilingual speakers: a male French-English bilingual speaker and a female Spanish-Catalan bilingual speaker. The use of bilingual speakers for the test data aims at reducing any sources of variability not due to the target languages. During each recording session, the speakers read the first two chapters of the book *The Little Prince* in one of their languages, and immediately afterwards they were asked to discuss what they had read. This procedure was then repeated for their second language. All recordings for each speaker were done on a single session.

The audio recordings were semi-automatically segmented into utterances with a 300 ms silence threshold using the speech analysis software *Praat* (Boersma & Weenink, 2014), and subsequently downsampled to 16kHz. The resulting dataset is composed of 319 utterances (French: 65, English: 75, Spanish: 99, Catalan: 80), with a mean length of 3.69 seconds ($min = 2.00$ s, $max = 10.63$ s).

Model of habituation task

Experimental studies of language discrimination in infants use an habituation paradigm (Mehler et al., 1988; Nazzi et al., 1998). In this paradigm, infants are presented with a set of stimuli from one language (L1), and their arousal is measured (in newborns, it is measured with a pacifier connected to a pressure detector). After an initial increase, infants' arousal decreases, indicating habituation. When a threshold has been reached, half of the infants continue with the same class of stimuli, and the other half are switched to a second class (L2). The difference of arousal after the switch in the two groups is used as a measure of discrimination.

Here, we will model this paradigm using an on-line clustering algorithm. In the habituation phase, the system gradually incorporates data from one language (L1) until it reaches a statistical threshold. In the test phase, as for infants, new utterances of L1 (*same* condition) and L2 (*switch* condition) are compared to the habituated model. The input of this model consists of the i-vectors of the test utterances as extracted by our previously trained system. To reduce spurious effects caused by specific subsets of utterances, the habituation task was run 100 times for each language pair using randomly selected subsets in each trial.

Habituation phase The model starts with an initial set of 10 i-vectors $\{v_1, ..., v_{10}\}$ of one language (L1) chosen randomly from our dataset. Firstly, the centroid μ_1 of this initial set (i.e., the mean i-vector) is computed, and the cosine distance of each of the 10 composing vectors to the centroid $d_c(v_i, \mu_1)$ is calculated. Secondly, a new random set of 10 i-vectors $\{v_{11}, ..., v_{20}\}$ of the same language L1 is selected, and their cosine distances to the initial centroid μ_1 are calculated. The distribution of distances of the initial and the second set of vectors are then compared with a t-test.

If $p \leq 0.05$, the two distributions are considered statistically different, that is, the model perceives a difference between the two sets of utterances, and therefore has not yet reached habituation. In this case, the last set of vectors is aggregated to the initial set and the centroid is recalculated, μ_2, as the mean i-vector of the whole set. Following the same procedure, a new group of 10 i-vectors from L1 is selected and their cosine distance to the new centroid $d_c(v_i, \mu_2)$, $\{i = 21, ..., 30\}$, are calculated and compared through a t-test to the distance of the previous 10 vectors to the new centroid $d_c(v_i, \mu_2)$, $\{i = 11, ..., 20\}$. This procedure is repeated as long as $p \leq 0.05$.

When $p > 0.05$ (defined as our saturation threshold), the two distributions are not statistically different and the habituation phase is therefore complete. As a final step, the last group of vectors is aggregated to the previous set and a final centroid is obtained, μ_F. The distance of the last 10 vectors to μ_F is then calculated and retained for the test.

Test phase In this stage, a new set of 10 i-vectors v_i is randomly selected from the same language used in habituation (L1, *same* condition) as well as 10 i-vectors u_j from the second language of the same bilingual speaker (L2, *switch* condition). For each set of vectors, the cosine distance to μ_F is calculated.

We finally perform two t-tests, one per condition, comparing the distribution of distances of the new vectors of L1 or L2 to the distribution of the last 10 habituation vectors. In the *same* condition, as the new utterances belong to the same language as those in habituation, the p-value of the t-test is expected to remain above the saturation threshold, $p > 0.05$. On the other hand, in the *switch* condition, the p-value will depend on the overlap between the distribution of the habituation (L1) and L2: a p-value below the 0.05 threshold would mean that the two distributions are significantly different, indicating discrimination of the two languages, while $p > 0.05$ would indicate a lack of discrimination.

Discriminability measures

To quantify the discriminability of the languages independently of our habituation-dishabituation model, we computed the pairwise ABX discrimination score, a nonparamet-

ric measure of category overlap. It consists in taking all possible ABX triplets of utterances from a language pair, where A corresponds to an utterance of L1, B corresponds to an utterance of L2, and X can be either L1 or L2. For each triplet, X is classified as belonging to L1 or L2 based on whether the cosine distance between X and A is smaller or greater than the distance between X and B. The percentage of correct classifications serves as an index of the discriminability between the two languages. Additionally, we performed a *Principal Component Analysis* (PCA) for each language pair as a way of visualizing the variance and distance of the i-vectors that compose each language.

Results

Habituation task

We ran the habituation model for both language pairs, and within each pair we tested the model with both possible languages in the initial habituation phase. The average amount of steps to reach habituation was similar for all languages (French: 2.1, English: 1.8, Spanish: 1.7, Catalan: 1.7).

As previously observed in infant experiments, the results of 100 trials for each test (presented in Figure 1) show a difference in the pattern of discrimination of the two language pairs. In the case of Spanish-Catalan (bottom panels), the p-values of both the *same* condition and the *switch* condition are significantly above the threshold value of $p = 0.05$, independently of the language presented in habituation (Habituation:Spanish -bottom right panel- *same*: $M = 0.48$, $SD = 0.26$, *switch*: $M = 0.40$, $SD = 0.26$; Habituation:Catalan -bottom left panel- *same*: $M = 0.52$, $SD = 0.28$, *switch*: $M = 0.54$, $SD = 0.27$), suggesting a lack of discrimination of these two languages. On the other hand, the second language pair (French-English, top panels) presented an asymmetry in the responses of the model to the *switch* condition, depending on the language of habituation. When the system is habituated to English as L1 and then switches to French (top left panel), the two languages are discriminated as indicated by a decrease of the p-value below the threshold in the *switch* condition (*same*: $M = 0.49$, $SD = 0.29$, *switch*: $M = 0.012$, $SD = 0.026$). However, if the system is initially habituated to French (top right panel), the switch to English is not detected, with both conditions showing similar p-values (*same*: $M = 0.54$, $SD = 0.29$, *switch*: $M = 0.48$, $SD = 0.25$). While a similar behavior was observed in infant habituation experiments (Mehler et al., 1988), additional analyses are required to understand this asymmetry.

ABX and Principal Component Analysis

To further explore the different response patterns of our model, we performed an ABX task for both language pairs and all possible X categories (ABA, ABB). The results of this test, shown in Table 1, present a similar pattern to the one observed in the habituation task. In the case of Spanish-Catalan, both ABA and ABB trials presented scores slightly above chance level (50%), meaning that nearly half of the Spanish

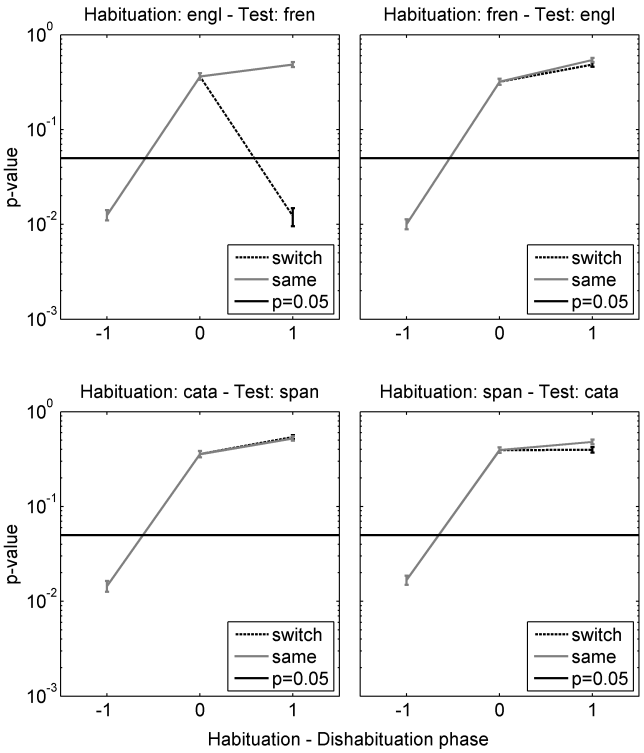


Figure 1: Average p-values over 100 trials of the habituation task for French-English discrimination (top) and Spanish-Catalan discrimination (bottom). The *x* axis represents the steps of the habituation and test phase, where *0* indicates the step where the habituation threshold ($p = 0.05$) was reached. Accordingly, *step -1* represents one step before habituation, and *step 1* represents the test (dishabituation) phase.

utterances were incorrectly categorized as Catalan utterances (and vice-versa). On the other hand, French-English trials presented an asymmetry: a majority of English utterances were correctly classified, while the classification of French utterances remained near chance level. This means that the distance between two given French utterances in the test set is often larger than the distance between a French and an English utterance, pointing at a possible imbalance in the variance of the distributions of their i-vector representations.

Table 1: Summary of ABX results: Percentage of accuracy for the distant language pair (A = English, B = French) and the close language pair (A = Catalan, B = Spanish).

Language Pair	X=A	X=B
English (A) - French (B)	76%	46%
Catalan (A) - Spanish (B)	51%	57%

Finally, we performed a Principal Component Analysis on both language pairs in order to visualize the distribution of the utterances. A representation of the first two dimensions

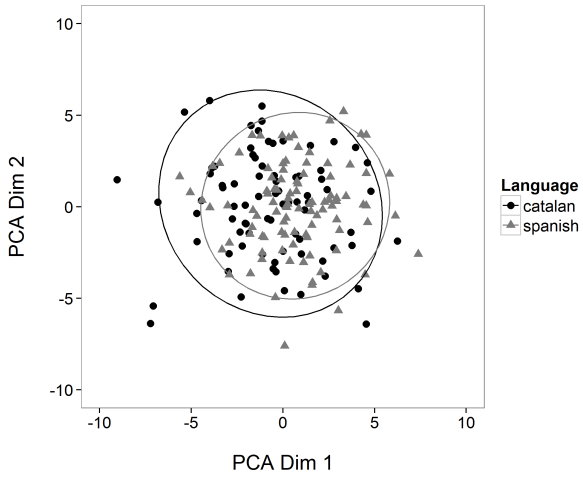


Figure 2: First two dimensions obtained through a Principal Component Analysis of the i-vectors of Spanish and Catalan utterances spoken by a bilingual speaker.

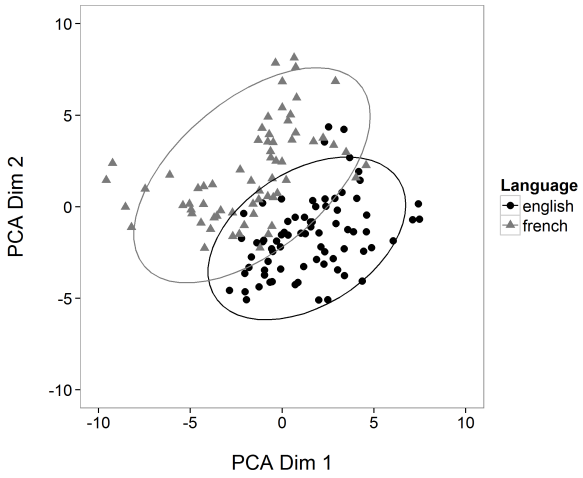


Figure 3: First two dimensions obtained through a Principal Component Analysis of the i-vectors of French and English utterances spoken by a bilingual speaker.

of the PCA for the Spanish-Catalan pair, shown in Figure 2, revealed a high degree of overlap in the distribution of the utterances of these two languages. On the contrary, the first two dimensions of the French-English PCA, presented in Figure 3, show a higher separation between the two languages. However, as suggested by the ABX score, the variance in these two dimensions appears to be larger within French utterances than within English utterances.

Together with the ABX results, this difference in the variance may explain the asymmetry observed in the habituation task: when the model is habituated to English, the variance of the i-vectors that are aggregated during this initial phase remains small, allowing the system to detect a switch to the second language. In other words, the within-language distance distribution is smaller than across-language. However in the inverse case, when the model is initialized with French, the variance of the habituation vectors is relatively large and therefore the switch to English remains unnoticed.

In summary, we found an overall difference in the degree of separation of the i-vectors of both language pairs, which reflected in the behavior of our habituation-dishabituation model. Spanish-Catalan utterances present largely overlapping distributions, causing a lack of discrimination in the habituation task, while French-English utterances have less overlapping yet more asymmetrical distributions, producing an equally asymmetric response of the system.

Discussion

In this paper we introduced a novel application of speech technology tools to model language discrimination in infants. Using a GMM-UBM trained on a small dataset of French utterances, we represented the acoustic space of a monolingual infant after a brief exposure to this language. To test the system's ability to discriminate languages, we modelled the acoustic representation of novel utterances as a pattern of shifts from the means of the UBM. Using this low-dimensional representation, called i-vector, we constructed a model of a habituation task similar to the experimental paradigm often used to test discrimination in infants.

The behavior of our model in the habituation task resembled that observed in previous experiments: our system, pre-exposed to French, was unable to discriminate between two previously unheard languages with highly similar phonologies (Spanish & Catalan), while it discriminated two phonologically distant languages (French & English). Interestingly, just as reported in previous infant studies such as Mehler et al. (1988), the ability to discriminate between French and English depended on the language presented during habituation. When the system was initially habituated to the previously unheard language (English), it detected a switch to the "native" language (French), but it failed at discriminating a switch to English when French was presented in habituation. Further analyses provided a potential explanation for our model's asymmetrical behavior: the variance of the i-vector representations of French utterances is larger than that of English utterances, causing the habituation model to create a broad category for French which hinders the discrimination of English. While in the context of infant studies this asymmetry was regarded as a preference for the native language, our modeling results suggest that the perceived acoustic variability might be responsible for this behavior, providing a new perspective on this issue.

There are three possible explanations for the larger variance of French as compared to English in our test data. First of all, this difference might be a characteristic of the specific bilingual speaker that was recorded for this experiment. To test this hypothesis, it would be necessary to repeat the test with a different French-English speaker. If the same pattern

was observed, it would indicate that the difference does not lie in the speaker but in the language. This could mean that, overall, French speech is acoustically more variable than English. However, and more interestingly, it is also possible that the difference was originated in the training of the Universal Background Model and the Total Variability subspace: as our system was pre-exposed only to French, the model may have developed a larger sensitivity to acoustic differences present in French speech than those in English speech, thus appearing more variable. To discern these two possibilities, the model could be re-trained using English as the background (i.e., "native") language. If the larger variance is due to the sensitivity of the model to its native language, then the asymmetry should be inverted. The results of these future modeling experiments may help us better understand the behavior observed in infants.

In addition, this methodology can be applied to model language discrimination in a variety of other cases. First, the UBM and the TV subspace can be trained with different languages and with varying amounts of data to investigate the impact of language exposure on discrimination (e.g., the model can be trained with a large dataset of Spanish speech and then tested on its ability to discriminate Spanish from Catalan). Second, the system could be trained with a bilingual background to study how multilingualism affects the construction of the acoustic space and consequently its ability to discriminate languages. This bilingual background can be composed of either monolingual speakers of two languages or bilingual speakers, giving further insight into the impact of different bilingual environments on the perceptual system. Third, the acoustic features provided to the model can be adapted (for example, by using filtered speech, or adding additional prosodic information to the feature vectors) to explore the role of different cues in language discrimination. The experimental data available to date provides a means of evaluation for the models, which in turn may generate new testable hypotheses that will help us better understand how young infants achieve this task.

Acknowledgements

We thank Alexander Martin and Laia Fibla for their participation in the recordings, and Hynek Heřmanský and Lukáš Burget for their helpful discussions. This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, the Region Ile de France (DIM Cerveau et Pensée), and an AWS in Education Research Grant award.

References

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, *49*(3), 1–30.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [Computer program].* Retrieved from `http://www.praat.org` (Version 5.3.86)

Bosch, L., & Sebastian-Galles, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological Science*, *21*(3), 343–348.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Martínez, D. G., Plchot, O., Burget, L., Glembek, O., & Matějka, P. (2011). Language recognition in ivectors space. In *Proceedings of interspeech 2011* (Vol. 2011, pp. 861–864). International Speech Communication Association.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, *24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*, 1–19.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Saffran, J. R., Aslin, R. N., & Neport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. (2002). Approaches to language identification using Gaussian Mixture Models and Shifted Delta Cepstral features. In *ICSLP 2002* (pp. 89–92).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., ... others (2006). The HTK book (for HTK version 3.4).

Are Financial Advisors Money Doctors or Charlatans? Evidence on Trust, Advice, and Risk Taking in Delegated Asset Management

Qizhang Sun (Qizhang.Sun@usi.ch)
Institute of Marketing and Communication Management, USI
Lugano-6900, Switzerland

Michael Gibbert (Michael.Gibbert@usi.ch)
Institute of Marketing and Communication Management, USI
Lugano-6900, Switzerland

Thomas T. Hills (T.T.Hills@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry, CV4 7AL, United Kingdom

Eric Nowak (Eric.Nowak@usi.ch)
Institute of Finance, USI
Lugano-6900, Switzerland

Abstract

We test the effects of advice and trust on risk-taking in three online experiments designed to elucidate under what conditions financial advice may increase risk-taking, irrespective of advisor performance. In our study, investors made 100 decisions, selecting between one of two alternatives: risky or conservative. We manipulate the suggestion of an advisor (risky vs. non-risky investments), the fee of the advice, as well as the trustworthiness of the advisor (by increasing the transparency of the advice presented) to test the effect of the advice on risk-taking. The results show that individuals asymmetrically follow the advice they received, with a bias towards following more risky than conservative advice. Moreover, trusted advice was more persuasive irrespective of what the advisor suggested and even the fee is higher.

Keywords: Advice taking; Financial advice; Money doctors; Risk taking; Trust

A central finding of the literature on advice-taking in financial economics is that many investors seek advice even though it often performs poorly relative to market benchmarks (Bergstresser et al., 2009; Chalmers and Reuter, 2012; Del Guercion and Reuter, 2014). In psychology, research also found that individuals keep looking for advice even though they understand that investors perform poorly (Sun et al., 2014). Gennaioli, Shleifer and Vishny (2015) in an analytical paper suggest this is because the trusted advice enables them to be more audacious than they would be otherwise, thus enabling them to take more risk, irrespective of the advisor's actual performance. Our study uses an experimental method as a first empirical test of this hypothesis, thereby bridging the financial and psychological dimension of advice-taking.

In our study, we hypothesize that (1) advisors will enable individuals to take more risks, in line with the core assumption of the model built by Gennaioli et al. (2015), but only when advice favors the risky alternative; (2) Compared to mere advice, trustworthy advice is more effective; moreover (3) trustworthy advice has a larger effect on encouraging individuals to take more risk than justified advice for more conservative investments. (4) Individuals still select the high trust advisor even though it is more expensive.

Money Doctor and Our Study

In a recent analytical paper in financial economics, entitled "Money Doctors", Gennaioli et al. (2015) present a new model of money management, in which investors delegate portfolio management to professional advisors based not on performance but on trust. In the model, the advisor decreases the investor's perception of the riskiness of a given investment, which allows advisors to charge fees to investors who trust them, even though the actual performance of advisors might be poor relative to market benchmarks. However, this still benefits investors, because the advisor encourages them to take higher risks, which they would not dare to bear just by themselves, and so earn superior returns. The central assumption of their work is that advisors charge investors to reduce their anxiety, which then enables investors to take these higher risks. By themselves, investors tend to be reluctant, or even anxious, to choose a risky alternative. However, the risky alternative often comes with a higher pay-off in the long-run. Financial advisors provide an important function in this situation. Specifically, the advisor seems to decrease the investor's perception of the riskiness of a given investment, reducing investor anxiety, enabling the investor to take more risk (and get more reward). To our knowledge, we provide the first study investigating these claims using an experimental method with randomly chosen survey participants. We conduct two different sets of survey experiments. In Experiment 1, we show that trusted advisors enable investors to choose riskier investment options. In Experiment 2, we extend the basic design by

manipulating the trustworthiness of the advisor (via increasing the transparency of the advice presented), and find that non-trustworthy advice attenuates the effect.

In both experiments, financial advisors provide risky as well as conservative advice, and we do model this for two reasons. The first reason is that financial advisors vary – both over time and in the cross-section - in how much risk they advise their clients in real life. Changing economic environments will influence performance criteria. This suggests that even mildly cognizant money managers should change their risk evaluations over time. In some cases, money managers do give conservative advice. Indeed, this is likely to be very common right after a market crash, especially as managers sometimes turn into "noise traders pandering to investors" (Gennaioli et al., 2015). Second, conservative advice is a necessary control to test for asymmetric influences of advice. Conservative advice is necessary in order to test the asymmetry of "anxiety reduction in risk-taking" (Gennaioli et al., 2015). Without conservative advice, we can only establish that people follow advice, which in the laboratory environment may be due simply to participant expectations that this is what the experimenter wants the participant to do.

Repeated Choice Paradigm

Investors rarely do have the true full information set about their investment alternatives. Instead, they need to refer to either their own experience by sampling or other's experiences (e.g., market history). This is classic risk under uncertainty, where the true probabilities of payoffs are unknown and may even be changing. Unlike decisions from description (Kahneman & Tversky, 1979), the decisions from experience paradigm is especially appropriate to decisions under uncertainty, as it requires decision makers to infer the properties of alternatives based on personal experience (Hertwig & Erev, 2009; Hills, Noguchi, & Gibbert, 2013). In particular, we use the decision from experience paradigm based on repeated choice from behavioral economics. In the repeated choice paradigm, participants choose alternatives and receive payoffs after each choice. Prior to decision making, decision makers do not get any information about alternatives.

General Method

We conduct an online survey experiment with Amazon's Mechanical Turk (mTurk). mTurk has recently been advocated to being utilized in large randomized online studies in experimental economics, for example by Kuziemko et al. (2015) to analyses preferences for redistributive policies. Participants were recruited from mTurk. Each participant received a participation fee $0.10 plus a performance-based bonus, which depended on their accumulated investment outcome. The participants were given the opportunity to provide informed consent and then received instructions about the experiment. Subsequently, they were required to complete the experiment task and the socio-demographic questionnaire, including age, gender, nationality, personal investment experience, financial consultant experience, organizational investment experience and professional financial knowledge.

In the experiment task, participants were asked to invest $1 on one of two alternatives, and they were required to choose one of the two alternatives for each of 100 consecutive virtual days. At the end of each day, they were informed about the outcome of the investment, and whether they earned or lost money. The outcome was presented on the selected alternative for one second. The investment alternatives were presented as boxes. Participants were allowed to choose an alternative (one of the presented boxes) by clicking on it. Unbeknownst to the participants, the two alternatives were a risky alternative (e.g., stock) and a safe alternative (e.g., savings account or bond). Every time the participants selected an alternative, they received a random draw from the alternative's underlying payoff distribution.

In the experimental groups, the participants were given an advisor at the very beginning of the experiment. No other information was provided about the advisor. Participants were free to ask (or not to ask) and to take (or not to take) the advice. The participants clicked on a female silhouette to solicit the (free) advice. The advisor only provided advice once, when asked, but the advice stayed on the screen until the very end of the experiment task. In the control group, there was no advisor available to be solicited.

Experiment 1

Method

Participants were 756 individuals randomly recruited from the online experimental platform mTurk. Excluding the missing data, the total number of participants is 721. A 3 (environment type: risky, neutral, or safe choice has the best expected value) X 2 (advice type: risky vs. conservative) X 2 (advice present or absent) between-subject design was used for the experiment. As the control group did not have an advisor to be manipulated, there were nine conditions in total.

One percent of the accumulated investment outcome was paid into the participants' MTurk account. The actual bonus range is from $0.23 to $1.36. The distribution of the alternatives are showed in Table 1. The participants were asked to rate the advisor every five trials on trial 6, 11, 16 and so forth, if they have asked for advice on that trial or the trial before. They were asked "how helpful did you find the advice? Please select from 1 (not at all helpful) to 7 (very helpful)".

Table 1: Distribution of the alternatives in Experiment 1.

	Risky alternative	Conservative alternatives
Risky environment	N ~ (0.75, 1)	always 0.5
Neutral environment	N ~ (0.5, 1)	always 0.5
Bearish environment	N ~ (0.5, 1)	always 0.75

The advice in the risky advice group was "I would recommend 'Alternative 1'. The returns for 'Alternative 1' are, in the long-run, greater than 'Alternative 2.' Although you may occasionally lose money in the short run, 'Alternative 1' offers the best long-run prospect for future returns", while in the conservative advice group the advice is "I would recommend 'Alternative 2'. You may not make huge jumps in earnings as in 'Alternative 1', but you don't lose large amounts of money and it gives you a secure payoff if you keep choosing it". Here 'Alternative 1' and 'Alternative 2' refer to the risky or safe option, respectively, which were randomized in location across participants.

Results and Discussion

Money Doctors. To test the intuition of the Money Doctors model by Gennaioli et al. (2015), we focus on whether (or not) advisors enable the investors to take more risk. To answer this question, we use a generalized linear mixed effects model to test Model 1, where "followed" means followed advice or not (binomial); "consult" stands for before or after advice seeking (binomial); that is, whether or not participants had advice on that trial (day), which aims to add the probability of selecting the suggested alternative before the advice as a baseline; "1 | participants" represent the individuals are used as an intercept random factor.

Followed ~ consult + (1 | participants) *(Model 1)*

The results show that participants are 5 times more likely to select the risky alternative after risky advice as before advice seeking ($B = 1.61$, $SE = 0.09$, $z = 17.50$, $p < .001$). This result supports the assumption in money doctor model, that a financial advisor enables the investor to take more risk. However, not only risky advice is followed, the conservative alternatives are also 2.77 times more likely to be selected after receiving conservative advice as before the advice seeking ($B = 1.02$, $SE = 0.09$, $z = 11.79$, $p < .001$).

A second Model 2 tests if the participants follow risky advice more than conservative advice, where "condition" means if the advisor provide a risky or conservative advice (binomial). The interaction represents the difference in advice following (the difference in the probability of following advice after advice taking) between the two advice conditions (risky advice and conservative advice).

*Followed ~ consult + condition + consult * condition + (1 | participants)* *(Model 2)*

The results show that the participants are 1.79 times more likely to follow the risky advice than the conservative advice ($B = 0.58$, $SE = 0.13$, $z = 4.60$, $p < .001$, Figure 1), which represents the "asymmetric effect" of being more likely to follow risky advice than conservative advice.

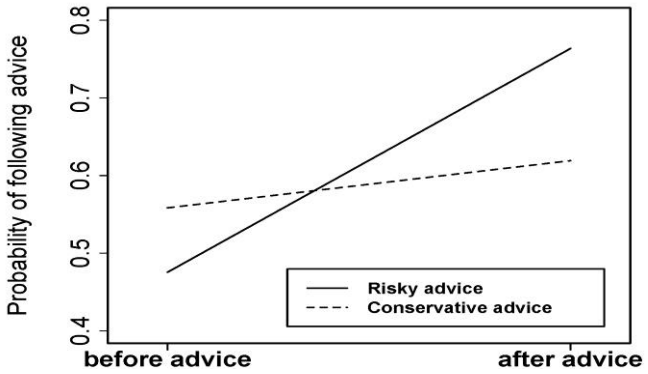


Figure 1. Asymmetric effect (the interaction)

Investment Environment. By comparing the Model 2 and Model 2.1, we are able to test how the macro-economic environment influences advice following. The results show that the three-way interaction, χ^2 (2) = 0.42, $p > .05$, is not significant. Breaking down the three-way interaction, we find that the coefficient of the asymmetric effect in the bullish, neutral and bearish economic environment respectively are 0.63 ($SE = 0.23$, $z = 2.73$, $p < .01$), 0.77 ($SE = 0.23$, $z = 3.36$, $p < .001$), and 0.53 ($SE = 0.20$, $z = 2.62$, $p < .01$).

*Followed ~ consult + consult + envir. + (1 | participants) + consult * condition + consult * envir.+ envir. * condition + consult * condition * envir.* *(Model 2.1)*

The results show that they are more likely to follow the risky advice regardless the macro-economic environments. Is it because people are not able to understand the investment environment? To answer this question, we asked if participants could detect the better alternatives in the absence of advice (i.e., the control group). The results show that in the control group, those who were in the bullish investment environment select the risky alternative more than by chance, $V = 2575$, $p < .05$, 95% confident interval (CI) is from 0.52 to 0.64, and those who were in the bearish environment selected the conservative alternative more than chance, $V = 740$, $p < .001$, 95% CI is from 0.29 to 0.40.

At the end of the experiment, we asked 3 questions- "which alternative do you think is better"," which alternative do you think has a larger value in the long run (larger average value)" and "which alternative do you think has a larger chance to get a larger value outcome"- to the participants to confirm their understanding to the environment. The results show that more than half of the participants in the bullish environment think that the risky alternative is better, χ^2 (1) = 15.68, $p < .001$, has a larger value in the long run, χ^2 (1) = 11.52, $p < .001$ and has a larger chance to get a larger value outcome, χ^2 (1) = 18.00, $p < .001$; more than half of the participants in the bearish environment think that the conservative alternative is better, χ^2 (1) = 21.36, $p < .001$, has a larger value in the long run, χ^2 (1) = 24.2, $p < .001$ yet do not have a larger chance to get a larger value outcome, χ^2 (1) = 2.69, $p > .05$ (only 37.78% of the people select alternative 2; that is conservative alternative, for this question).

These results show that the asymmetric effect is robust in different macro-economic environment. Even though they sense the macro-economic environment, they are still more likely to follow the risk favouring advice. These results suggest that the financial advisor does enable individuals to take more risk (and hope for a larger pay-off), irrespective of the advisor's actual performance— an assumption that is strikingly central to the Money Doctor model of Gennaioli et al. (2015).

Financial Advisors. Are participants able to detect the quality of the advisors? Do they follow the advice simply because they do not sense the quality of the advisor? We test if participants who are in consistent conditions (conservative advice in bearish environments or risky advice in bullish environments) rate the advisor better than those who are in inconsistent conditions. The Wilcoxon rank sum test with continuity correction result is negative, $W = 21143.5$, $p > .05$. In other words, participants are not able to perceive the quality of the advisor. If investors do not sense the quality of the advisor and clear the macro-economic environment in mind, why are they more likely to follow the risky advice? Gennaioli et al. (2015) suggest it is because they trust the advisor, which is to be tested in the Experiment 2.

Financial Experience. We also test the link between financial experience and the asymmetric effect by adding personal information into the model (Model 2). The results show that those who have personal investment experience (three-way interaction: $B = -0.19$, $SE = 0.26$, $z = -0.75$, $p > .05$, the interaction represents the asymmetric effect) and consultant experience (three-way interaction: $B = 0.17$, $SE = 0.32$, $z = 0.53$, $p > .05$) do not show a larger asymmetric effect.

However, those who have the organizational investment experience (three-way interaction: $B = 0.79$, $SE = 0.38$, $z = 2.09$, $p < .05$) and professional knowledge do (three-way interaction: $B = 1.63$, $SE = 0.54$, $z = 3.02$, $p < .01$). These results suggest that those with financial investment training rather than personal experience are less likely to follow the conservative advice. These results are complementary to those of Von Gaudecker (2015) on the effect of financial literacy and financial advice on the diversification of Dutch household portfolios.

Experiment 2

In Experiment 1, we showed that advisors enable people to take more risks irrespective of the quality of the advice. In Experiment 2, we manipulate advisor trust by manipulating the phrasing of the advice.

Method

Participants were 400 individuals randomly recruited from the online experimental platform mTurk. Excluding the missing data, the total number of participants is 387. A 2 (justification: justified vs. non-justified) X 2 (advice type: risky vs. conservative advice) X 2 (advisor presence: present or absent) between-subject design was used. As the control group did not have an advisor to be manipulated, there were five conditions in total. Fifteen percent of the accumulated investment outcome was paid into the participants' MTurk account. The actual bonus range is from \$0 to \$4.57. Regarding the distribution of the alternatives, the conservative alternative always has an outcome of \$0.05, while the risky alternative has a mean of 0.1, and a standard deviation of 1.0. Participants were not asked to rate the advisor.

Advisor trust was manipulated by altering the phrasing of the advice: in the conservative justified group, the advice was "I would recommend 'Alternative 2'. You may not make huge jumps in earnings, but you don't lose anything and it gives you a secure payoff". In the risky justified group, the participants read the advice as "I would recommend 'Alternative 1'. The returns for 'Alternative 1' are, in the long run, greater than 'Alternative 2.' Although you may occasionally lose a large amount of money, you will earn it back if you keep choosing it". In contrast, the participants in the non-justifying groups only saw the advice as "Choose:'Alternative 1'" and "Choose:'Alternative 2'" in the risky and conservative group respectively.

Results and discussion

Trusted Advisor. We test the phrasing of the advice on the probability of selecting a risky alternative using Model 3, where "followed" means followed advice or not (binomial); "consult" stands for before or after advice seeking (binomial); "phrasing" means whether the advice is justified or non-justified; "1 | participants" shows that individuals are used as fixed effects. The interaction represents the difference in advice following between justified advice and non-justified advice.

*Followed ~ consult * Phrasing + (1 | participants)(Model 3)*

The results show that people receiving the justified advice are more likely to follow the advice (interaction: $B = 0.48$, $SE = 0.15$, $z = 3.19$, $p = .001$, Figure 2). This result suggests that the phrasing of the advice plays a role in advice following, which is in line with our hypothesis. As the trust on the advisor can be manipulated by the phrasing of the advice, this result also supports another key finding in the money doctors model: investors follow financial advisors based on trust.

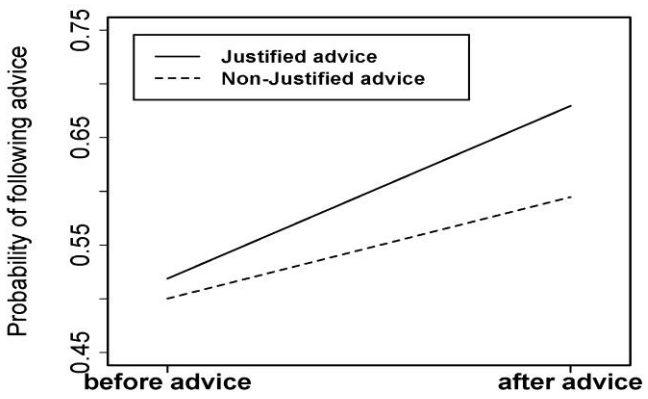


Figure 2. The effect of advice phrasing.

After adding the advice condition (risky advice or conservative advice) into the model (Model 3), the results show that the effect of the advice phrasing is stronger in risky advice (three-way interaction: $B = 1.93$, $SE = 0.37$, $z = 5.27$, $p < .001$). To further break down the interaction, we compare the justified advice condition to the non-justified condition. People were more likely to follow justified than non-justified advice in the risky condition (interaction: $B = 2.33$, $SD = 0.32$, $z = 7.36$, $p < .001$, Figure 3). Further analyses show that people increase the probability of selecting risky alternatives after the advice comparing to before the advice was given in justified condition ($B = 2.35$, $SD = 0.28$, $z = 8.29$, $p < .001$), but not in the non-justified condition. This result means the advisor enables the investors to take more risk only if the investors trust in the advisor. Otherwise, the effect does not kick in. This underlines the importance of trust in terms of boosting risk taking and demonstrates the third assumption of the money doctors model.

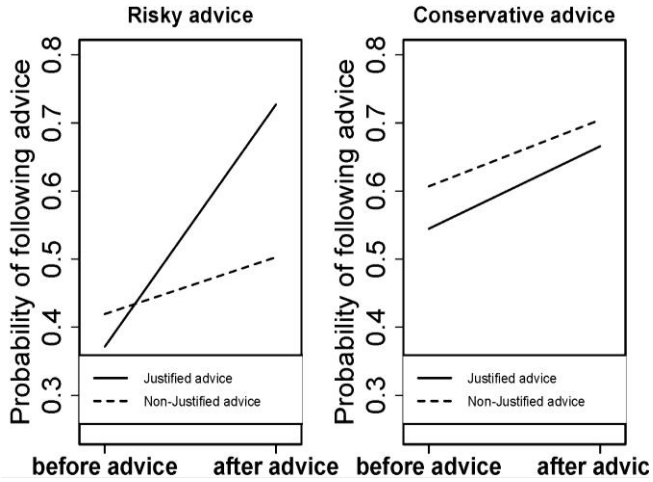


Figure 3. Simple effect of trust in Experiment 2.

In the conservative condition, people were less likely to follow the justified advice (interaction: $B = -0.40$, $SD = 0.18$, $z = -2.18$, $p < .01$, Figure 3). Further analyses show that people increase the probability of selecting conservative alternatives after the advice comparing to before the advice in the justified condition ($B = 0.75$, $SD = 0.13$, $z = 5.72$, $p < .001$), but the effect is also found in the non-justified condition ($B = 1.15$, $SD = 0.13$, $z = 8.94$, $p < .001$). Here, trust in the advisor does not make them more likely to follow the advice compare to those without trust.

Asymmetric effect. Using the justified advice condition and Model 2, we find a strong asymmetric effect; people are more likely to follow risky than conservative advice (interaction: $B = 1.62$, $SD = 0.32$, $z = 5.09$, $p < .001$, Figure 4). To look closer, we separate the risky advice and conservative advice. The results show that people follow both risky advice ($B = 2.35$, $SD = 0.28$, $z = 8.29$, $p < .001$) and conservative advice ($B = 0.75$, $SD = 0.13$, $z = 5.72$, $p < .001$). That means people tend to follow trusted advice, but are still more likely to follow risky than conservative advice.

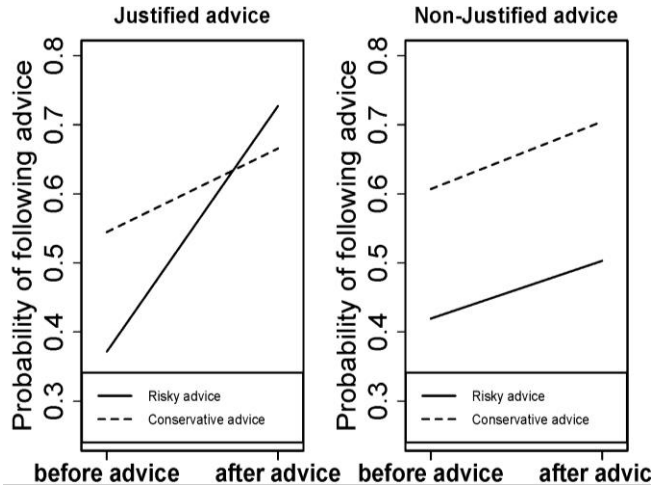


Figure 4. The asymmetric effect in Experiment 2.

We also test the effect in non-justified condition, and the asymmetric effect is reversed (Figure 4), which means people are more likely to follow the conservative advice in this case (interaction: $B = -1.11$, $SD = 0.18$, $z = -6.13$, $p < .001$, Figure 4). Separating the risky and conservative advice condition, we find that under the unjustified condition, people do not follow the risky advice ($B = 0.04$, $SD = 0.13$, $z = 0.30$, $p > .05$), while people follow the conservative advice ($B = 1.15$, $SD = 0.13$, $z = 8.94$, $p < .001$). These results imply that people need to trust the advisor to take more risk, whereas to follow a conservative advice, the advisors do not necessarily need to be trusted. Although people can follow the conservative advice anyway - but only follow a trusted advisor to take risk, people are much more likely to follow the risky advice than the conservative advice once they consider the financial advisor to be trustworthy.

Experiment 3

In Experiment 3, we manipulate the trust on advisors to test the robustness of the trust effect. We also assign different prices to the advisors, in order to prove that an investor prefers a trusted advisor and enables managers to charge a higher fee and still keep them (Gennaioli et al., 2015).

Method

Participants were 118 individuals randomly recruited from the online experimental platform mTurk. Excluding the missing data, the total number of participants is 104. One percent of the accumulated investment outcome was paid into the participants' *MTurk* account. The actual bonus range is from $0.29 to $1.4.

A 2 (high vs. low trust, within subject) X 2 (same price vs. different price, between subject) mixed design was used. There were two phases in the experiment. In both phases, the

participants were forced to make a decision between two alternatives. The first 30 decisions manipulated the participants' trust on the advisors. Then, in the second phase, there were 70 decisions. They were instructed that all the alternatives were different from the previous 30 rounds and the participants still could get a piece of advice from one of our advisors (who were exactly the same advisors as in the first 30 rounds.) However, during the second phase, the advisors charged them a certain fee. They are free to ask (or not to ask), take (or not to take) the advice and choose one of the advisors.

In the first phase, we set one alternative as better than the other in a very obvious manner. The better alternative (alternative 1) followed a normal distribution with a mean 1.5 and standard deviation 0.5, while the worse alternative (alternative 2) followed a normal distribution with a mean 0.2 and standard deviation 1. All participants saw two advisors with their own advice. The advice was directly showed to the participants next to its advisor's profile. To manipulate the trust on the advisor, we made the quality of the advisor easily judgeable by the participants. The high trust advisor suggested the better alternative in a justifying way- "I would recommend 'Alternative 1'. The returns for 'Alternative 1' are greater than 'Alternative 2.' In addition, it is less risk and more secure. Plus, you almost do not lose money if you select 'Alternative 1', whereas you have almost half a chance of losing money if you select 'Alternative 2'"; The low trust advisor simply suggested to the participants who went for the worse one in a non-justifying way- "choose: alternative 2".

In the second phase, the risky alternative (alternative 1) followed a normal distribution with a mean 0.88 and standard deviation 1.5, while the conservative alternative (alternative 2) followed a normal distribution with a mean 0.3 and standard deviation 0.1. Irrespective of the advisor chosen, the participants receive the justifying advice suggesting the risky alternative. In the same price group, both advisors charged 50 cents. In the different price group, high trust advisor charged 100 cents, while the low trust one charged 50 cents.

Results and discussion

Manipulation Check. We asked the participants which advisor they trust more at the end of the first phrase. Ninety-two participants selected the high trust advisor and twelve participants select the other. Chi-squared test result shows that the number of participants selecting the high trust advisor is significantly higher than the number of participants selecting the low trust advisor, χ^2 (1) = 61.54, $p < .001$. This suggests that the first phrase has successfully manipulated participants' trust on the advisors.

Trust. We test whether participants tend to select the high trust advisor in the task phrase. Across the price condition, 47 and 12 participants selected the high trust advisor. Chi-square test results shows that investors are more likely to select the advisor they trust, χ^2 (1) = 20.76, $p < .001$. Breaking down the price conditions, both conditions had more investors buying the high trust advisor. In the same price condition, 20 participants bought the high trust advisor and only 3 bought the low trust advisor, χ^2 (1) = 12.57, $p < .001$. In the different price condition, 27 participants bought the high trust advisor and 9 bought the low trust advisor, χ^2 (1) = 9, $p < .01$. These results show that the investors selected the advisor because of trust, which is in line with our hypotheses and the money doctors model.

Price. After we prove that the investors select advisors because of trust, we move to the next step; that is to prove trust enables managers to charge the investor a higher fee and still keep him. To test this hypotheses, we use a generalized linear model to examine whether the increased price reduces the probability of selecting the high trust advisor. The tested model is with "advisor" as dependent variable and "condition" as independent variable, where "advisor" is which advisor the investors choose (binomial) and "condition" is the price variable (binomial). The result shows a negative effect, $B = -0.80$, $z = -1.10$, $p > .05$. This means that the probability of selecting the high trust advisor does not drop down because of the increased price. Together with the findings about trust in this experiment, all these results prove the arguments put forth by money doctors model.

References

Bergstresser, Daniel, John M. R. Chalmers, and Peter Tufano, 2009, Assessing the Costs and Benefits of Brokers in the Mutual Fund Industry, The Review of Financial Studies 22, 4129-4156.

Chalmers, J., and J. Reuter, (2012) What is the Impact of Financial Advisors on Retirement Portfolio Choices and Outcomes?, NBER Working Paper, 18158.

Guercio, D. D. and Reuter, J. (2014), Mutual Fund Performance and the Incentive to Generate Alpha. The Journal of Finance, 69: 1673–1704.

Gennaioli, N., Shleifer, A., and Vishny, R. (2015). Money doctors. The Journal of Finance, 70(1), 91-114.

Hertwig, R., and Erev, I. (2009). The description–experience gap in risky choice. Trends in Cognitive Sciences, 13(12), 517-523.

Kahneman, D., and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 263-291.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." American Economic Review, 105(4): 1478-1508.

Sniezek, J. A., and Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. Organizational Behavior and Human Decision Processes, 84(2), 288-307.

Sun, Q., Noguchi, T., Hills, T. and Gibbert, M., Paying for Useless Advice to Resolve Information Overload, Working Paper.

Gaudecker, H.-M. V. (2015), How Does Household Portfolio Diversification Vary with Financial Literacy and Financial Advice?. The Journal of Finance, 70: 489–507.

A Recurrent Network Approach to Modeling Linguistic Interaction

Rick Dale (rdale@ucmerced.edu)
Cognitive and Information Sciences, University of California, Merced

Riccardo Fusaroli (fusaroli@dac.au.dk), Kristian Tylén (kristian@dac.au.dk)
Interacting Minds Centre & Center for Semiotics, Aarhus University

Joanna Rączaszek-Leonardi (joanna.leonardi@gmail.com)
Institute of Psychology, Polish Academy of Sciences, Warsaw, Poland

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University
Interacting Minds Centre, Aarhus University

Abstract

What capacities enable linguistic interaction? While several proposals have been advanced, little progress has been made in comparing and articulating them within an integrative framework. In this paper, we take initial steps towards a connectionist framework designed to systematically compare different cognitive models of social interactions. The framework we propose couples two simple-recurrent network systems (Chang, 2002) to explore the computational underpinnings of interaction, and apply this modeling framework to predict the semantic structure derived from transcripts of an experimental joint decision task (Bahrami et al., 2010; Fusaroli et al., 2012). In an exploratory application of this framework, we find (i) that the coupled network approach is capable of learning from noisy naturalistic input but (ii) that integration of production and comprehension does not increase the network performance. We end by discussing the value of looking to traditional parallel distributed processing as flexible models for exploring computational mechanisms of conversation.

Keywords: language; interaction; neural networks; production; comprehension

Introduction

What capacities enable linguistic interaction? There are a large number of extant theoretical proposals. A glance at the literature reveals a host of proposed mechanisms that support conversation and other sorts of interactive tasks. Some of these are specific to social or linguistic cognition, such as mirroring and simulation (Oberman & Ramachandran, 2007), mind or intention reading (Tomasello, Carpenter, Call, Behne, & Moll, 2005), linguistic alignment (Garrod & Pickering, 2004), and use of common ground (Clark, 1996). Others have drawn on domain-general cognitive processes, including memory resonance of social identity (Horton & Gerrig, 2005), perceptuomotor entrainment (Shockley, Richardson, & Dale, 2009), synergies (Fusaroli, Rączaszek-Leonardi, & Tylén, 2014), one-bit information integration (Brennan, Galati, & Kuhlen, 2010), coupled oscillatory systems (Wilson & Wilson, 2005), executive control (Brown-Schmidt, 2009), brain-to-brain coupling (Hasson, Ghazanfar, Galantucci, Garrod, & Keysers, 2012), and situated processes (Bjørndahl, Fusaroli, Østergaard, & Tylén, 2014).

Many of these proposals are individually supported by rigorous experimentation or corpus analysis. However, language happens in the "here-and-now" (Christiansen & Chater, in press) and thus must satisfy a plurality of constraints at the same time: from the perceptuomotor level all the way "up" to social discourse and pragmatics (Abney et al., 2014; Fusaroli et al., 2014; Louwerse, Dale, Bard, & Jeuniaux, 2012). Accordingly, there remains a need to systematically compare and articulate the contributions of the suggested mechanisms in an integrative model of interactional language performance (Dale, Fusaroli, Duran, & Richardson, 2013).

In this paper, we propose a computational framework that enables flexible combination and comparison of different cognitive constraints. We show that coupled simple-recurrent networks are capable of learning sequential structure from latent-semantic analysis (LSA) representation of wordforms in interactive transcripts. As a first case study, we use a traditional neural-network approach to test the role of production-comprehension integration during natural language performance (MacDonald, 2013; Pickering & Garrod, 2014).

Production, Comprehension, and Prediction

We look to production-comprehension integration to illustrate this framework. The relationship between production and comprehension is a key factor in most theories of language processing. In research on conversational or task-based interaction, these two systems are granted considerable and often distinct attention. Does language production vary more as a function of internal constraints of the speaker, or more in response to the needs of his or her listener (for some review, among many, see Brennan & Hanna, 2009; Ferreira & Bock, 2006; Jaeger, 2013)? A prominent recent theory takes these systems to be deeply intertwined. Pickering and Garrod (2014; see also MacDonald, 2013) have argued that an integration of production and comprehension is critical in understanding the mechanistic underpinnings of interaction.

Experimental and neuroimaging work suggests simultaneous involvement of both aspects of language processing during linguistic interactions (e.g, Silbert, Honey, Simony, Poeppel, & Hasson, 2014). However, explicit cognitive modeling can more directly reveal the extent to which, for example, prediction and understanding are improved as a function of

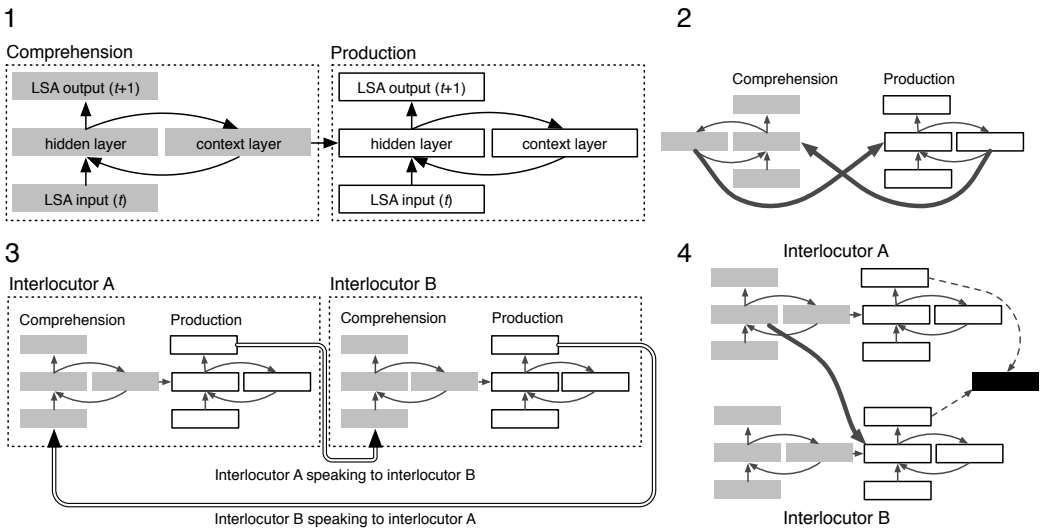


Figure 1: (1) A representation of two coupled simple-recurrent networks (SRN) inspired by Chang (2002). A conversant is modeled as a two-SRN agent. A pair of coupled subnetworks is referred to as an *agent network*. (2) In the original Chang (2002) model, production did not influence comprehension. We model the complete integration of production-comprehension by having these two subnetworks share internal states. (3) A conversation can be modeled as a coupling between two such "nets of nets," leading to a second-order recurrent network. Each agent receives input from the other, and shares the hidden states of its comprehension subnetwork with the input layer of its production when it is its turn. We refer to this second-order network as a *dyad network*. (4) This framework can be parameterized to investigate, for example, the effect of explicitly externally shared information between interlocutors, akin to emerging common ground (black box with dotted lines), or the extent to which one network is facilitated by having access to the "internal state" of another network (thick solid line).

tighter integration. In what follows, we describe one way to model human interaction using parallel distributed processing (PDP). Inspired by a predictive approach to language, we adapt the models of Elman (1990) and Chang (2002) to couple neural networks into two interacting systems, and show that such a model can be parameterized in various ways to test computational claims.

Higher-Order Recurrent Dynamics

We draw inspiration from the successful PDP model of Elman (1990) and adapted by Chang (2002) to investigate sentence processing in a single cognitive agent. The architecture of this simple-recurrent network (SRN) is shown in Fig. 1, panel 1. This network receives input in a comprehension subnetwork. In Chang (2002), this was modeled as a set of input sentence primes. The hidden state of the comprehension network (activation of nodes at the hidden layer) then constrains the production subnetwork, and influences its subsequent performance. Such a network has been shown to effectively model syntactic priming effects (Chang, 2002).[1]

Each person in an interaction can be represented as a pair of SRNs – receiving input and generating output with production and comprehension subnetworks. Modeling conversation then involves coupling these neural network architectures into a "dyad." We couple these nets by taking the output of "speaker" and use it as the input of the "listener," as shown in Fig. 1, panel 3. On a turn-by-turn basis, we can switch who is doing the producing and comprehending. The networks are trained to predict word sequences in this way, in the context of a coupled "conversation." As shown in Fig. 1, panel 3, there are two levels of coupling in this model. These first-order networks (*agent network*) are coupled in their comprehension and production subnetworks in some way. Interaction is modeled as a coupling between two such networks, as a second-order recurrent network (*dyad network*).

This model can be readily adapted to parameterize constraints on processing. In Fig. 1, panel 2, we show that we can "complete the circuit" in the dyads by connecting production to comprehension in the same way. This simple modification inspired two conditions in a preliminary simulation. First, we studied the ability of dyad nets to predict words in interaction under the original formulation, with only comprehension constraining production. We then tested the contribution of full comprehension-production integration by completing the circuit, and compared its performance to the original formulation.

Like any cognitive model, this framework requires an input space that provides structure to the task. Elman (1990) used simulated sentences generated by a simple grammar, and Chang (2002) used hand-coded semantic and syntactic representations in a simplified grammar. To get input vectors for our model, we used transcripts from an interactive task in which two participants communicate to jointly solve a perceptual task (Fusaroli et al., 2012). Taking the word-by-word

[1]Note in Fig. 1 that Chang's original model only included the constraint on production from prior comprehension.

sequences in these transcripts, we created input activations based on latent semantic analysis representations. This reduces the dimensionality and sparsity of the input space and makes the learning problem more tractable for the network. It also tests the framework with complex naturalistic data.

Input Corpus: LSA Word Vectors

The corpus consists of 16 dyads (32 Danish-speaking individuals) totaling more than 1,600 joint decisions, 25,000 word tokens and 1,075 unique word types.[2] Given the sparsity of the lexical space, we transformed the corpus into a latent semantic analysis (LSA) representation (Landauer & Dumais, 1997). This projects words into a lower dimensional feature (vector) space based on how the words occur in the corpus. We define a word's relative cooccurrence to another word by using a simple 1-step window, so that the cooccurrence of word w_i with word w_j is the total number of times they followed each other, $f(i,j) = NP(w_{i,t}, w_{j,t+1})$, where N is the total number of words in sequence, and P the joint probability that words i and j occurred together at times t and $t+1$, respectively. This count serves as an entry in a 1,075 $\times$ 1,075 matrix M, as the entry at the ith row and j^{th} column. This matrix is, of course, quite sparse, because most words do not cooccur with every other word. LSA was employed as a means to overcome such sparsity, providing a lower-dimensional representation of word similarity based on these distributional patterns: $[U,S,V] = \text{SVD}(M)$.

The left eigenvector matrix (U) provides a more compact representation for individual words. Rather than a complete (but sparse) representation across all 1,075 of its column entries, the SVD solution that LSA uses allows us to take a much smaller number of columns of U instead. These columns represent the most prominent sources of variance in the distributional patterns of the word usage.

When inspecting the singular values (S) of the SVD solution in an LSA model, we find that word usage across all transcripts can be captured by about 7 of these columns of U. A schematic of how we use these feature vectors is shown in Fig. 2, which illustrates a pattern of activity across 7 nodes as the input for these networks. This gives us a 7-dimensional representation of words, where activations can be negative or positive, which requires some modification to the training of our SRN subnetworks.

Training with LSA

Because common backpropagation assumes an activation range of $[0,1]$, we had to modify the input and output activation transformations to suit a $[-1,1]$ range. To do this we changed the standard sigmoid function, used as output activation function, to a tanh function that has the desired properties. In order to propagate error back, we differentiate the tanh function at the output nodes. Because derivative $d\tanh = 1 - \tanh^2$, we obtain

$$\delta_{\mathbf{o}} = \mathbf{o} \circ d\mathbf{o} \circ \mathbf{e} = \mathbf{o}(1 - \mathbf{o} \circ \mathbf{o}) \circ \mathbf{e} \quad (1)$$

Where $\mathbf{o}$ is the output vector of the network, $\mathbf{e}$ is the error associated with each node, and $\circ$ represents elementwise multiplication. $\delta_{\mathbf{o}}$ reflects the error assignment to output nodes. Once this is calculated, we can modify the weights $W_{\mathbf{h}\to\mathbf{o}}$ with

$$\Delta W_{\mathbf{h}\to\mathbf{o}} = \alpha \mathbf{h}^T \delta_{\mathbf{o}} \quad (2)$$

α represents the learning rate parameter, and $\mathbf{h}$ the hidden unit activations of a given subnetwork. We used this approach to modify the weights connecting hidden and output layers. All other layers were treated in the common way with the sigmoid function and its derivative, in accordance with traditional iterated backpropagation.

In order to train the networks using LSA vectors as they interact in dyads, we follow the process illustrated in Fig. 2. In a turn-by-turn fashion, the production subnetwork of one agent net would be trained to predict its "spoken" output, while the comprehension subnetwork of the other agent net would receive this output as input and predict it in a word-by-word fashion.

Simulation Procedure

Training and Testing To assess how well the models capture interactional structure of the empirical data, we trained 16 dyad networks in each of two conditions (comprehension to production only vs. full integration). Each network was trained on one pass on the full transcripts of 15 dyads (almost 25,000 word presentations) and then tested on the remaining target dyad. We set α to .01, and the number of hidden units across all subnetworks to 10.[3] We built a baseline control for each test dyad by shuffling its word order, thus disrupting the sequential structure that the networks were expected to learn. The 'A' or 'B' designation of interlocutors was randomly assigned, but used here for convenience of presentation.

Our performance measure was based on the common measure of cosine between the output and target vectors. Cosine is commonly used with the LSA model, since it captures whether word vectors are pointing in the same direction in "semantic space." Cosine varies from $[-1,1]$, with higher values reflecting better predictions by the network.

Predictions First, we expected that the "content" shared between speaker and listener, projected in LSA space, should allow the networks to learn the statistical structure of interaction. Second, we contrasted three hypotheses about production-comprehension integration. (H1) Fully integrated production-comprehension systems would benefit performance, as the networks are able to receive "more in-

[2] Space limitations prevent us from fully describing the construction of this semantic representation, but we note that we also included a "turn end" marker to ensure that words adjacent across turns were not treated as if they were spoken in the same sequence of words by one person.

[3] Space restricts our parameter search, but we found, in general, that hidden layer size did not greatly impact performance in any conditions in our explorations.

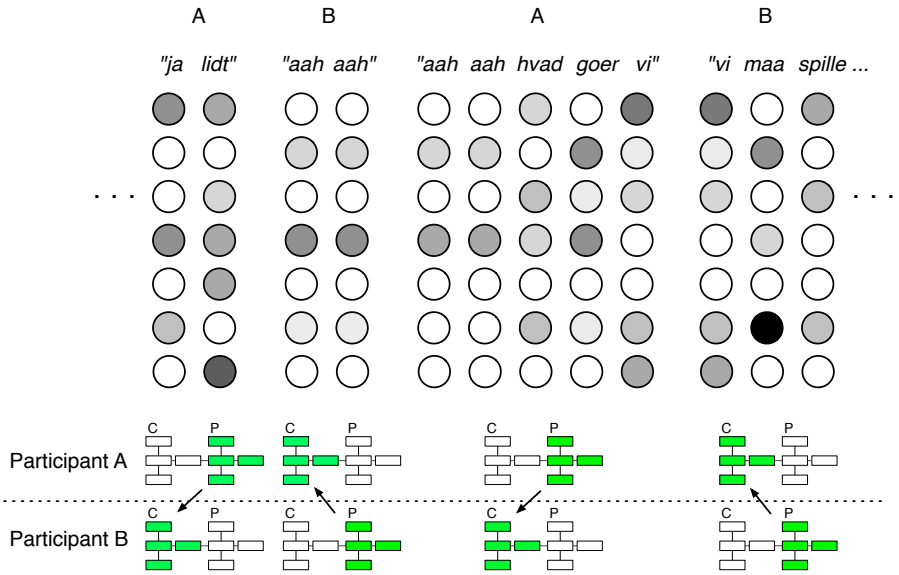


Figure 2: We organized network training by interactive turn. For a given turn, one participant (A or B) is doing the talking. We take the LSA vectors (visualized as a distributed pattern of activation) and have the production network of the speaker on that turn predict its output, and the comprehension network of the other participant predict its input. Within each dyad, the subnetworks of each participant take turns learning to predict the LSA vectors.

formation," in that the comprehension net is now receiving input from production. (H2) Fully integrated production-comprehension systems degrade performance as they introduce noise to the network and an additional set of weights that the network has to learn. (H3) There will be no difference between these networks: Our simplified task has the production and comprehension networks doing very similar things, and so we may not observe any divergence in their performance.

Results

Can Dyad Nets Learn Sequential Structure? When comparing networks in both conditions, it appears that they are very similarly effective at predicting word-by-word LSA vectors in unseen interactions, and that they also show much better performance than the control baseline, in which words are shuffled. This means that networks are processing the order of LSA features, and not simply capturing the activation space in which these LSA features reside. This learning effect is quite large, and is shown in Fig. 3. The appropriate test here is a paired-sample t-test, since each network and its control are trained on matched sets of words with the same network. A t-test across all four subnetworks shows the expected result, for both conditions: t's > 25 and p's $< .000001$.

Does Integration Improve Prediction? The average cosine performance did not differ between the two network conditions, using the same paired-sample t-test across layers, $t(63) = 0.33, p = .74$. This absence of an effect is quite evident in Fig. 3. No reliable difference emerges in direct comparison of any of the subnetworks, either.

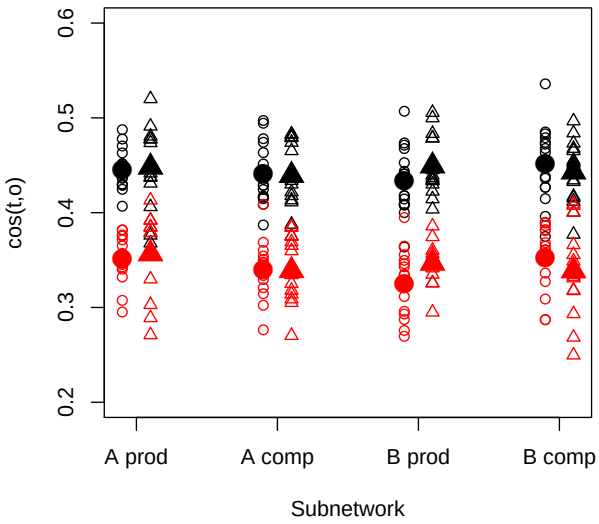


Figure 3: Dyad networks are capable of learning interactional structure. The cosine for agent subnetworks trained on sequential structure show greatly increased scores relative to baseline subnetworks, for which temporal order of the LSA training vectors are shuffled. In general, agent nets with comprehension $\Rightarrow$ production (circle) do not perform differently from agent nets with integration (triangle). They do both show better performance than the control (red). The models are both learning to predict LSA vector sequences. $cos(t,o)$ stands for the cosine of target and observed output vectors.

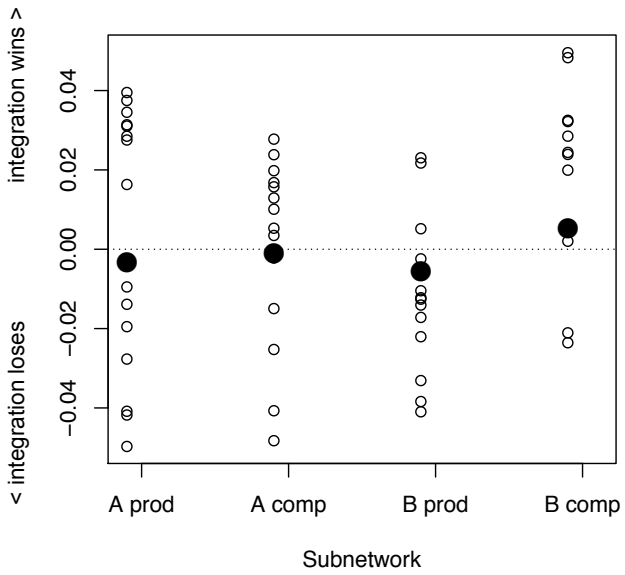


Figure 4: Difference between integrated and unintegrated agent network conditions relative to their respective baselines. It reflects how much more one network can be expected to exceed its baseline relative to the other condition. If integrating production and comprehension improves performance, we expect a positive value on the y-axis.

Does Integration Improve Prediction above Baseline? These results are shown in Fig. 4. In general, as might be expected from the prior analyses, the models are not different from each other in most subnetwork performance. All results are non-significant, with the initial agent net configuration not different from its baseline relative to that of the fully integrated networks, $|t|$'s < 1.

General Discussion

In this paper, we described a flexible computational framework to investigate the cognitive mechanisms underlying linguistic interaction. The first step in this direction is the implementation of coupled neural networks to learn from interaction data. We demonstrated that this adaptation of Chang (2002) is capable of learning the sequential semantic structure in raw, noisy input.

Based on the current debate on interactive alignment, we manipulated their internal cognitive structure to contrast two theoretically motivated models: (i) a model with full comprehension-production integration, and (ii) a model without integration. These alternative coupled networks were then used to model real conversational data in order to investigate hypothesized prediction benefits of full integration. Our results did not reveal an effect of full integration. Put simply, hypothesis (H3) seems to have been supported here: In this computational system, full integration does not bring great gains, if any. Why did we not observe clearer results? To conclude, we outline theoretical and methodological considerations that hint at possible explanations and motivate future implementations of the framework.

First, the results can be interpreted to suggest that 'internal' production-comprehension coupling is in fact not facilitating mutual prediction in this context. This could indicate that recurrent (and thus predictive) structure resides on levels other than the turn-by-turn organization of the conversation. In fact, a recent study (relying on the same corpus) suggests that linguistic patterns critical to performance in the task tend to straddle interlocutors and speech turns making turn-by-turn alignment secondary to recurrent structural patterns at the level of the conversation as a whole (Fusaroli & Tylén, 2016). A future implementation of the model could directly test such ideas (sometimes referred to as the interpersonal synergy model of dialogue: Fusaroli et al., 2014) and compare the performance to other types of conversational interaction that might entail different functional organization.

These results might also be contingent upon a number of methodological limitations that will need to be overcome in future developments. First, the sample size is not impressive and a bigger corpus would possibly enable better training of the networks. Second, in order to deal with the sparse lexical space of real conversations, we reduced the input to LSA vectors. As a consequence both the comprehension and production subnetwork end up dealing with the same kind of data. Integrating comprehension does therefore not add information that is not already contained in the LSA vectors processed in production subnetworks. Thus, the integration is at least partially redundant and cannot be expected to add much to the performance of the model.

There are also more general limitations to overcome. For example, anticipatory dynamics in agent networks should allow overlap at the turn level, as seen in natural interactions. This is a critical feature for modeling the higher-order dynamics of interaction. The PDP approach embraces such computational extensions. For example, networks could be gated, such that off/on states of the production subnetwork will have to be learned by agents. The recurrent property of these networks *should* allow them to predict forthcoming turn switches. The approach offers much in the way of extension, as these networks are, after all, nonlinear function approximators over any arbitrary sets of temporal constraints. For example, we could also develop other input spaces, such as multimodal constraints from nonverbal aspects of interaction, and add them to the verbal components we have explored here.

This flexibility also permits more focused theoretical explorations. The constraints on these networks have theoretical implications that can be readily adapted to further compare and integrate proposed mechanisms, the topic that began this paper. For example, Fig. 1, panel 4 showcases how we might develop the framework to test combinations of other constraints on interaction, such as "common ground." Another example is how internal constraints from one agent network might constrain, and possibly facilitate, the dynamics of the agent to which it is coupled in the dyad network. Theoretically motivated manipulations of this kind would allow more explicit tests of the relationship among these various proposals for the mechanisms of interaction, and compar-

isons to related computational frameworks (e.g., Buschmeier, Bergmann, & Kopp, 2010; Reitter, Keller, & Moore, 2011).

Acknowledgments

Thanks to the Interacting Minds & Center for Semiotics at Aarhus University for its support in bringing the co-authors together for a meeting last year in January, 2015 to discuss this work. Thanks also to Andreas Roepstorff for fun and productive discussions during our visit. The co-authors vibrantly discussed the theoretical status of such a PDP framework, and did not come to a consensus about that status. It did not detract from the fun.

References

Abney, D. H., Dale, R., Yoshimi, J., Kello, C. T., Tylén, K., & Fusaroli, R. (2014). Joint perceptual decision-making: a case study in explanatory pluralism. *Frontiers in Psychology*, *5*.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085.

Bjørndahl, J. S., Fusaroli, R., Østergaard, S., & Tylén, K. (2014). Thinking together with material representations: joint epistemic actions in creative problem solving. *Cognitive Semiotics*, *7*(1), 103–123.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. *Psychology of Learning and Motivation*, *53*, 301–344.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.

Buschmeier, H., Bergmann, K., & Kopp, S. (2010). Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In *Empirical methods in natural language generation* (pp. 85–104). Springer.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651.

Christiansen, M. H., & Chater, N. (in press). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. C. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation*, *59*, 43–95.

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes*, *21*(7-8), 1011–1029.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological Science*, 931–939.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*, 145-171.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11.

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404–1426.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*.

Oberman, L. M., & Ramachandran, V. S. (2007). The simulating social mind: the role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychological Bulletin*, *133*(2), 310.

Pickering, M. J., & Garrod, S. (2014). Neural integration of language production and comprehension. *Proceedings of the National Academy of Sciences*, *111*(43), 15291–15292.

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, *111*(43), E4687–E4696.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(05), 675–691.

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968.

The Aging Lexicon: Differences in the Semantic Networks of Younger and Older Adults

Dirk U. Wulff (Wulff@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Lentzeallee 94
14195 Berlin, Germany

Thomas T. Hills (T.T.Hills@warwick.ac.uk)
Department of Psychology, University Road
Coventry, CV4 7AL, United Kingdom

Margie Lachman (lachman@brandeis.edu)
Department of Psychology, 415 South Street
Waltham, MA 02453, USA

Rui Mata (Mata@unibas.ch)
Department of Psychology, Missionsstrasse 64A
4055 Basel, Switzerland

Abstract

How does the mental lexicon, the network of learned words in our semantic memory, change in old age? To address this question, we employ a new network inference method to infer networks from verbal fluency data of a group of younger and older adults. We find that older adults produce more unique words in verbal fluency tasks than younger adults. In line with recent theorizing, this suggests a larger mental lexicon for older than for younger adults. Moreover, we find that relative to the mental lexicon of younger adults, the mental lexicon of older adults is less small-world-like. Based on several findings linking network clustering to processing speed, this finding suggests that not only the size, but also the structure of the mental lexicon may contribute to apparent cognitive decline in old age.

Keywords: Semantic representation, networks, small world, verbal fluency, aging.

Introduction

Cognitive science commonly depicts semantic memory as a random walk traversing a network of concepts or words, often called the mental lexicon (Abbot, Austerweil, Griffith, 2015; Anderson, 1983; Collins & Loftus, 1975; De Deyne, Verheyen, & Storms, 2014; Hills, Todd, & Jones, 2012; Miller, 1995; Vitevich, 2011). A corollary of this view is that the structure of the network should impact memory performance (Jones, Hills, Todd, 2015; Vitevich, 2008; Borge-Holthoefer & Arenas, 2010). Several studies have attempted to measure the mental lexicon using memory productions from free association or natural language (Ferrer-i-Cancho & Sole, 2001; Steyvers & Tenenbaum, 2005; Morais, Olssen, & Schooler, 2013). They concluded that the macroscopic structure of the mental lexicon follows, similar to networks is other domains, a small-world structure, implying higher clustering and equal average path length than found in random networks (Humphries & Gurney, 2008; Steyvers & Tenenbaum, 2005; Watts & Strogatz, 1998). However, whether this aggregate pattern also characterizes the mental lexica of individuals is far from understood. As a first step toward answering this question, we will apply in this investigation a new network inference method to study and compare the small world-ness of younger and older adults' mental lexica based on verbal fluency data.

The Aging Mental Lexicon

Semantic memory follows a unique developmental trajectory across later age. Whereas episodic memory and fluid abilities, such as working memory capacity, peak in early adulthood, the performance of semantic memory, as measured by vocabulary tests, increases until age 65 to 70 (Hartshorne & Germine, 2015; Keuleers, Stevens, Mandera, & Brysbaert, 2015) or even beyond (Kavé & Halamish, 2015)[1]. Recently, it was argued that the positive trend for semantic memory might actually be responsible for the poor performance of older adults in other cognitive variables (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). Specifically, if older adults have access to an increasing number of words, then their mental lexicon must be larger. As with finding a book in a large as compared to a small library, retrieving information from a large mental lexicon

[1] Moreover, linguistic analyses suggest that common vocabulary tests underestimate the true vocabulary size for older adults (Baayen, 2001).

should take more time and be more error prone, than in a small mental lexicon. The apparent cognitive decline in older adults – usually attributed to a general cognitive slowing for older adults due to neuronal deterioration (Light, 1991) – may thus arise from changes in older adults mental lexica.

As size of the mental lexicon can impact the performance of the memory system, so should its structure (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). Previous research has successfully connected a network's clustering coefficient to processing speed (Nematzadeh, Ferrara, Flammini, & Ahn, 2014; Vitevich, et al., 2011), enhanced priming effects (Nelson & Goodmon, 2002), as well as facilitated recognition and recall (Nelson, Bennett, Gee, Schreiber, & McKinney, 1993; Nelson, Zhang, & McKinley, 2001). Furthermore, the structure of the mental lexicon is naturally related the meaning of words, categorization, and language itself (Borge-Holthofer & Arenas, 2010; De Deyne, Verheyen, & Storms, 2014; Jones & Mewhort, 2007).

To our knowledge, no previous study assessed the effect of aging on the structure of the mental lexicon. Existing research rather focused on the early developmental trajectory of the mental lexicon in children (e.g., Hills, Maouene, Maouene, Sheya, & Smith, 2009; Beckage, Smith, & Hills, 2011). Hills and colleagues (2009) found evidence in favor of preferential acquisition process, in which words are learned as a function the words' degree of connectedness in the learning environment. More research is needed to corroborate this finding and it is unclear whether these trends can be extrapolated to later life. We therefore aim to help investigate the structural development of the aging mental lexicon.

Present Study

In this investigation we provide a first peek into the mental lexicon's adult development by comparing mental lexica of younger and older adults as inferred from verbal fluency data. Doing this we will focus on the average local clustering coefficient, the average shortest path length, and the small world-ness of young and older adults' semantic networks. Out of the many network statistics, the clustering coefficient is most prominent. It is also the only network statistic that has been causally connected to memory performance (Vitevich et al., 2011). The average shortest path length has not been linked to memory performance, however, it carries an intuitive interpretation: Networks with larger average shortest path lengths should lead to slower and less flexible recall performance. Finally, as a composite measure of the clustering coefficient and the average shortest path length, small world-ness speaks to the global structure of a network. Many natural occurring networks exhibit a small world structure, including importantly word occurrences in natural language (Ferrer-i-Cancho & Solé, 2001). Small world networks have been found to "display enhanced signal-propagation speed, computational power, and synchronizability" (Watts & Strogatz, 1998, p. 440).

Method

Data

The data for our analysis was comprised of 60 seconds animal fluency data from a total of 332 participants. Verbal fluency tests ask participants to list within a defined time window as many members of a natural category as they can think of. The dataset was composed of the data from two independent studies, 228 (18.6 animals on average) participants from Hills, Mata, Wilke, and Samanez-Larkin (2013) and 104 participants (21.8 animals on average) from a subsample from the Midlife in the United States study (MIDUS; Lachman, Agrigoroaei, Tun, & Weaver, 2013). The data of the latter was transcribed by us from audio recordings using the Penn TotalRecall[2] software. Following Lerner, Ogrocki, and Thomas (2009), the datasets were subjected to preprocessing, in which variants of the same animal were combined ('kitten' and 'kitty'), but alternate forms of the same species were retained ('cow' and 'calf'). We removed 36 participants for low scores on a dementia screener (value < 26; Folstein, Folstein, & McHugh, 1975) and 16 for producing too few animals (n < 10). A total of 284 participants entered the analysis, with age ranging from 29 to 94 and number of produced animals from 10 to 39. Figure 1 illustrates the sample. For the comparison of older and younger adults we performed a median split on age, resulting in a group of 142 younger adults, aged 29 to 65, producing on average 22 animals, and a group of 142 older adults, aged 66 to 94, producing on average 18.8 animals (see also Table 1).

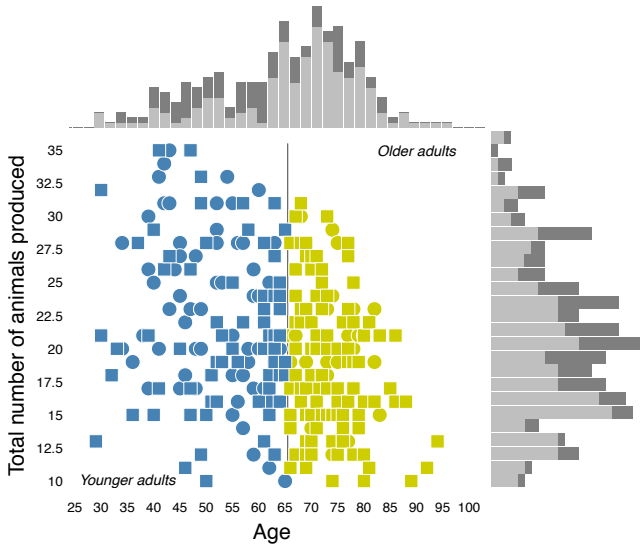


Figure 1: Illustration of the data used to infer the mental lexica of younger and older adults.

Table 1: Description of the raw data after median split.

[2] http://memory.psych.upenn.edu/TotalRecall

Group	N	Age	Mean recalled	Unique recalls N_U	N_U / N_T
Younger Adults	142	19 – 65	22	287	.091
Older Adults	142	66 – 94	18.8	284	.106

Legend: N_U /N_U – Number of unique/total words

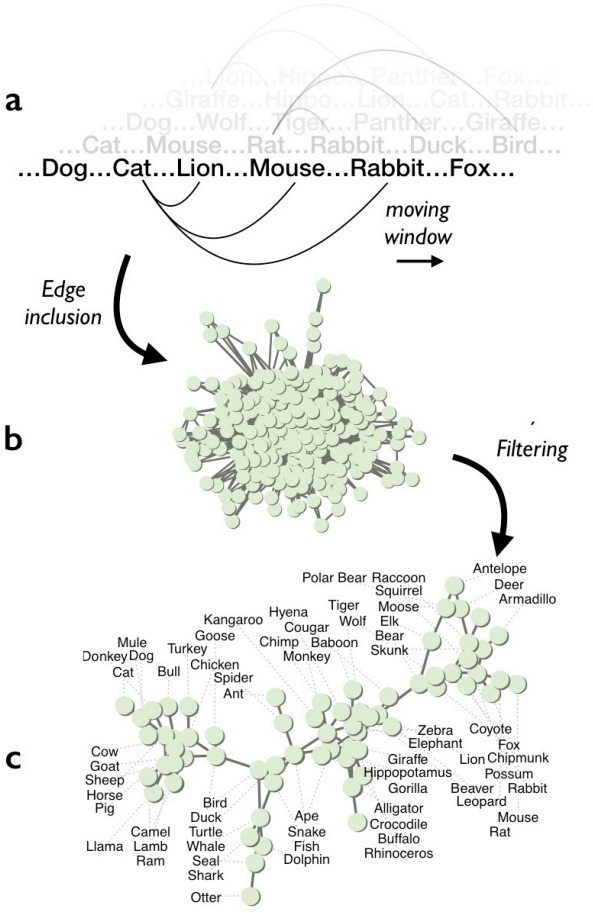


Figure 2. Illustration of the network inference method.

Network Inference Method

To infer the mental lexicon of younger and older adults from verbal fluency data, we borrow from the statistical procedure developed by Goñi et al. (2011). The method assesses for every two words whether they co-occur more frequently than would be expected by chance. Figure 2 illustrates the three steps of the procedure. In the first step (a), a window of size w moves through all verbal fluency sequences of one group and records the number of times two words co-occur within the window. For instance, if w = 3, then all pairs of words with no more than one intervening word entered the next step of the analysis. In the second step (b), a non-weighted, undirected graph is created from the pairs that co-occurred more often than a minimum threshold of m. For example, if m = 2, then all pairs of words that co-occurred at least twice entered the final step of the analysis. Finally, in the third step (c), the recorded frequency of co-occurrence is tested against the random expectation based on the marginal frequencies of words and the lengths of the verbal fluency sequences[3]. Specifically, an edge between a pair of nodes is retained whenever likelihood of the frequency of co-occurrence under the random model surpasses a lower threshold c. As can be seen in the example given, in the bottom panel of Figure 2 the method produces highly intuitive networks. The network shown is based on the older adults data and w = 3, m = 3, and c = .05.

Relative to other methods that are used to construct networks from verbal fluency data, our method has two important advantages. First, it is the only method based on the common contention that related words co-occur within relatively small window sizes, often no more than two or three words apart (Abbot et al., 2015; Troyer et al., 1997; Hills, Jones, & Todd, 2012, Wulff, Hills, & Hertwig, 2013). Second, the flexible parameterization allows to map different parts of the mental lexicon. Specifically, increasing the minimum co-occurrence parameter m means that only the strongly connected core network will be assessed (Baronchelli et al., 2013). A similar case can be made for the window size parameter w.

A critical aspect of investigations into the macroscopic properties of networks is statistical inference. When the endpoint of the investigation is a single network, one cannot rely on standard statistical procedures. Often the only possibility to generate standard errors lies in bootstrap methods that repeatedly construct subnetworks from randomly selected nodes and edges. The reliability of such methods is, however, contested (Sneijders & Borgatti, 1999). To circumvent these issues, we chose to bootstrap participants instead of subnetworks. Specifically, we inferred as many networks as there were participants in a group according to a leave-one-out procedure (Efron & Efron, 1982; Sneijders & Borgatti, 1999). We then compared network measures between the groups based on the pooled standard deviations computed across the 142 leave-one-out-networks for each of the groups.

Network Measures

We focus in our analyses on the two network measures constituting a small world: The average (local) clustering coefficient and the average shortest path length. The clustering coefficient C, sometimes also called transitivity, refers to the proportion of cases in which the neighbors of a node are neighbors themselves. The shortest path length of two nodes L refers to the length of the shortest possible way to traverse from one node to the other. The average clustering coefficient and the average shortest path length

[3] For details see Goñi et al. (2011).

are both computed as the arithmetic mean of their respective measures.

To counter known dependencies of the clustering coefficient and the shortest path length on the size and connectedness of the network (von Wijk, Stam, & Daffertshofer, 2010), we measure both variables relative the expectations of a Erdos-Renyi random graph (Bollobas, 2001). In addition, to prevent influences from the outset of the analysis, we match the data of younger and older adults. Specifically, we implemented the following matching scheme: First, we included older and younger adults with exact matches in samples size of fluency productions. Second, for all participants that could not be matched in the first round, we identified for each older adult the younger counterpart whose number of productions lay closest, cropped the extra productions at the end produced by the young adult, and included the pair. The matched data included 5230 productions, representing about 90% of the original 5802 productions. All of the following results hold for an alternative matching scheme in which young adults' extra productions were cropped at the beginning.

We also include in our analysis the small-world-ness index *S* developed by Humphries and Gurney (2008). This measures combines the clustering coefficient and the average shortest path into a single metric, while controlling for the expectations of a Watts-Strogatz small world network. The measure indicates small-world-ness for values > 1.

Results

Unique Productions

First, we turn to the number of unique productions. Although younger adults produced slightly more unique animals in total, older adults produced more unique items per production, which was both, significant and substantial, according to a bootstrap test ($d = -1.9$, $p < .001$). This finding is consistent with the idea that a larger mental lexicon slows the productions of older adults (Ramscar et al., 2012).

Network Comparison

Figure 3 shows the results of our network inference for the number of nodes in the resulting graph (*N*), the average clustering coefficient relative to random (C/C_{rand}), the average shortest path length relative to random (L/L_{rand}), and the small-world-ness index *S*. Specifically, the figure displays for each measure the median result of the 142

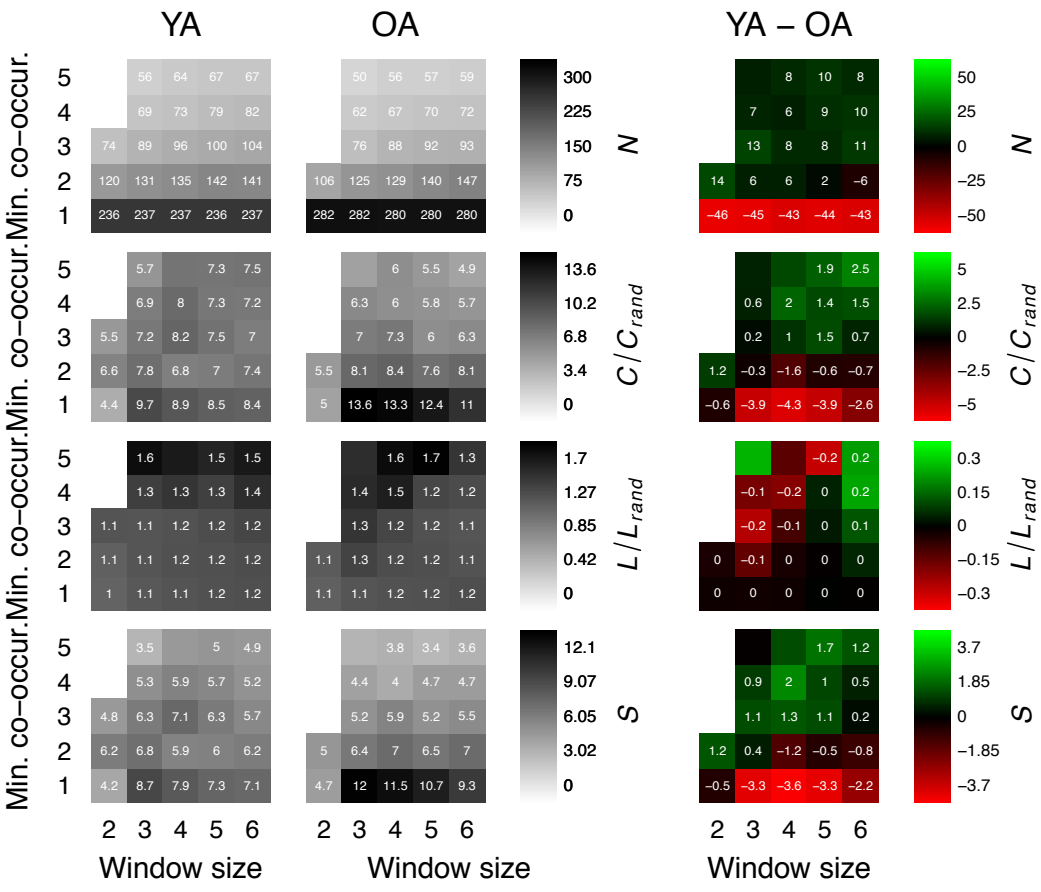


Figure 3. Macroscopic properties of the mental lexicon of older and younger adults as inferred from verbal fluency data.

leave-one-out runs per group. The results are shown for five levels of the minimum co-occurrence parameter *m*, five levels of the window size parameter *w,* and a single criterion value of $c = .05$. The left two columns show the results for the younger (YA) and older adults (OA), while the rightmost column displays their difference. The figure only displays boxes when a giant components comprised of at least 90% of all nodes could be recovered, and numbers, when the statistical test yielded significance at the level of .05.

As can be seen from the figure, the networks and their differences across groups are markedly influenced by the choice of minimum co-occurrence, but not window size. Specifically, the analysis of all data, including pairs of words that only co-occurred once, resulted in about 20% larger networks for the older adults than for the younger adults. Responsible for this pattern is that, consistent with more unique productions for older adults, the older adults frequency distribution of productions has a longer tail of highly infrequent items. Under most circumstances, such infrequent items will likely form unique pairs with the words they are paired with. As it is impossible to discern whether such pairings are the result of a systematic association between the words or mere chance, we followed Goñi et al. (2011) in disregarded co-occurrences that occurred no more than once ($m \geq 2$).

Evaluating the networks for $m \geq 2$ revealed a very clear pattern: The networks of younger adults exhibit more clustering, mostly shorter average path lengths and mostly larger small-world-ness indices. This suggests that the structure of the younger adults mental lexicon resembles more closely a small world structure than that of older adults.

Discussion

To our knowledge our investigation represents the first comprehensive comparison of younger and older adults mental lexica. Consistent with previous research (Ramscar et al., 2014), we have shown that relative to younger adults, older adults produce more unique words in the verbal fluency task, suggesting a larger underlying mental lexicon. Critically, we have also shown that relative to the mental lexicon of younger adults, the mental lexicon of older adults exhibits a less small-world-like structure, implying less clustering. Based on previous findings that associate network clustering with processing speed (e.g., Nematzadeh, Ferrara, Flammini, & Ahn, 2014), this result suggests that the structure of the mental lexicon of older adults might contribute to the apparent cognitive decline in old age.

The inference of a mental lexicon from memory productions is, however, not without caveats. Most importantly, one must consider that memory productions are the process not only of the underlying network, but also a search process operating on the network (Jones, Hills, & Todd, 2015; Raaijmakers & Shiffrin, 1981). This means that one cannot easily attribute differences in inferred networks to the underlying structure. This holds in particular as it has been proposed that the search process is affected by age (Hills et al., 2013). Future studies should aim to disentangle search process and network, possibly by means of simulation. Future studies should also aim to corroborate the present findings by extending the analysis of mental networks to other methods (e.g., similarity ratings) and content (e.g., other natural categories).

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological review, 122*(3), 558.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior, 22*(3), 261-295.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in cognitive sciences, 17*(7), 348-360.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PloS one, 6*(5).

Bollobas, M. (2001). *Random Graphs*, 2nd Edition. Cambridge: Cambridge University Press.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: structure and dynamics. *Entropy, 12*(5), 1264-1302.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology, 68*(8), 1665–1692. http://doi.org/10.1080/17470218.2015.1022560

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review, 82*(6), 407.

De Deyne, S., Verheyen, S., & Storms, G. (2014). Structure and organization of the mental lexicon: a network approach derived from syntactic dependency relations and word associations. Towards a theoretical framework for analyzing complex linguistic networks. Springer.

Dorogovtsev, S. N., & Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B: Biological Sciences, 268*(1485), 2603-2606.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*(1), 36-48.

Ferrer-i-Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences, 268*(1482), 2261-2265.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research, 12*(3), 189-198.

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N. V., Corominas-Murtra, B., ... & Villoslada, P. (2011). The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive processing, 12*(2), 183-196.

Hartshorne, J. K., & Germine, L. T. (2015). When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychological science, 26*(4), 433-443.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review, 119*(2), 431.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal Analysis of Early Semantic Networks Preferential Attachment or Preferential Acquisition? *Psychological Science, 20*(6), 729-739.

Hills, T. T., Mata, R., Wilke, A., & Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental psychology, 49*(12), 2396.

Humphries, M. D., & Gurney, K. (2008). Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PloS one, 3*(4).

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review, 112*(3), 570-574.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review, 114*(1), 1.

Kavé, G., & Halamish, V. (2015). Doubly blessed: Older adults know more vocabulary and know better what they know. *Psychology and aging, 30*(1), 68.

Lachman, M. E., Agrigoroaei, S., Tun, P. A., & Weaver, S. L. (2013). Monitoring cognitive functioning: Psychometric properties of the brief test of adult cognition by telephone. *Assessment, 21*(4), 404-417.

Lerner, A. J., Ogrocki, P. K., & Thomas, P. J. (2009). Network graph analysis of category fluency testing. *Cognitive and Behavioral Neurology, 22*(1), 45-52.

Light, L. L. (1991). Memory and aging: Four hypotheses in search of data. *Annual review of psychology, 42*(1), 333-376.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM, 38*(11), 39-41.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science, 37*(1), 125-145.

Nelson, D. I., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. *Memory & Cognition, 30*(3), 380-398.

Nelson, D. L., Bennett, D. J., Gee, N. R., Schreiber, T. A., & McKinney, V. M. (1993). Implicit memory: Effects of network size and interconnectivity on cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(4), 747.

Nelson, D. L., Zhang, N., & McKinney, V. M. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1147.

Nematzadeh, A., Ferrara, E., Flammini, A., & Ahn, Y. Y. (2014). Optimal network modularity for information diffusion. *Physical review letters, 113*(8), 088701.

Newman, M. (2010). *Networks: an introduction*. Oxford: OUP.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological review, 88*(2), 93.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science, 6*(1), 5-42.

Snijders, T. A., & Borgatti, S. P. (1999). Non-parametric standard errors and tests for network statistics. *Connections, 22*(2), 161-170.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science, 29*(1), 41-78.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138.

Vitevitch, M. S. (2008). What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval? *Journal of Speech, Language, and Hearing Research, 51*(2), 408–422.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: implications for spoken word recognition. *Frontiers in psychology, 2*.

Van Wijk, B. C., Stam, C. J., & Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PloS one, 5*(10).

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature, 393*(6684), 440-442.

Wulff, D. U., Hills, T. T., & Hertwig, R. (2013). Worm holes in memory: Is memory one representation or many?. In *CogSci 2013: 35th Annual Conference of the Cognitive Science Society* (pp. 3817-3822). Cognitive Science Society.

Aiding Preschoolers' word-learning by scaffolding lexical awareness

Sofia Jimenez (sofia.r.jimenez@vanderbilt.edu)

Kaitlin Ryan (kaitlin.f.ryan@gmail.com)

Megan M. Saylor (megan.saylor@vanderbilt.edu)
Psychology and Human Development, 230 Appleton Place #552
Nashville, TN 37203 USA

Abstract

Preschool-aged children develop awareness of the words they do and do not know. Awareness of one's lexicon may encourage word learning if children pay more attention to the definition of unknown words. Here, we tested 3-4-year-old children (N = 91) on a word learning task embedded in an e-book. When a novel word was read, children were either asked if they knew the word, asked a question about the storyline, or asked no question. Then they were given a description without visual input and asked to identify the referent's picture from three choices. Participants who were asked if they knew a word before being provided with the definition identified more referents than children in the other conditions. Children's word learning was predicted by short-term memory.

Keywords: word learning, lexical awareness, preschoolers, object representation, memory

Introduction

Recognizing that a word is unfamiliar might help children learn its meaning. This recognition can prompt the search for a definition, or the use of context to reason about meaning. Although older children and adults are often aware of what they know and what they do not (Klin et al. 1997, Flavell, 1979), preschool-aged children are still developing this ability. The use of the word "know" emerges around a child's third birthday (Bartsch & Wellman, 1995; Montgomery, 1992), along with other words referring to mental states (Johnson & Maratsos, 1977). At the same time, preschoolers are developing their ability to monitor their uncertainty across domains (Lyons & Ghetti, 2011). Since children ages three-to-five are acquiring many words each day, researchers have investigated their ability to recognize when unfamiliar words as well as the factors that influence this recognition.

Merriman and Marazita (2004) argued that there are two components to determining whether one knows a word: recognition that the sound form is unfamiliar (word-cued awareness) or appreciation that one does not know the name for a certain type of thing (meaning-cued awareness). Merriman and Marazita (2004) described preschool children's ability to use these types of cues. They tested word-cued awareness by asking children if made up words and real words were words or not. Rejecting the made up words and accepting the known words shows mastery of this task. Overall there was an increase in awareness of lexical ignorance from three to four. Three-year-olds accepted an average of 51% of the made-up words and 84% of the familiar words. Four-year-olds performance was more robust as they accepted 24% of the novel words and 92% of the familiar words. Their test of meaning-cued awareness mirrored this performance pattern. These results suggest there may be an increase in children's lexical awareness during the preschool years.

Children's awareness of lexical ignorance can be supported by exposure to words, actions, and objects that differ in familiarity. For example, asking children to sort words into piles of known and unknown words with a visual aid leads to greater accuracy on lexical awareness tasks (Merriman & Maritza, 2004), the pairing a novel word with a novel action (Merriman, et al. 1996) and exposure to familiar and novel objects (e.g., Hartin, Stevenson, & Merriman, 2016). This work suggests that lexical awareness in preschoolers may be work best when children are prompted to consider familiarity and novelty. One possibility is that children may benefit from being prompted to consider whether they know the meaning of an unfamiliar word prior to hearing a definition, if they do not deploy their awareness of lexical ignorance spontaneously.

One additional question is how age-related improvements in lexical awareness might be supported by cognitive and linguistic abilities. One possibility is that concomitant developments in cognitive skills may support this understanding (Merriman & Marazita, 2004) such that increased short-term memory capacity or a heightened ability to mentally represent objects may support children's lexical awareness abilities. These cognitive skills might help children recognize when they do not know words to the extent that they support search through their lexicon. Alternatively, the development of these cognitive skills might be separate from the advance of lexical awareness. We investigate these two possibilities in the current study.

Although some children are reluctant to ask adults questions about unfamiliar words, children with large vocabularies are one exception; they are much more likely to ask about the meaning of a word (Jacobson & Saylor, in prep). This suggests that children who already have large vocabularies may have an easier time gaining more vocabulary. This has been referred to as the Matthew effect

(Stanovich, 1986), or the "rich getting richer." However, the direction the relationship is unclear: does having a larger lexicon lead you to knowing when a question should be asked (i.e. more lexical awareness leads to more question asking and vocabulary increases), or do some children have a larger vocabulary because they are predisposed to asking questions. One way to investigate these possibilities is to have adults scaffold lexical awareness by ask preschoolers questions about word knowledge.

Question asking during joint book reading is an established technique called dialogic reading (Whitehurst, et al., 1988) and it aids in word learning (Senechal & Cornell, 1993; Senechal et. al., 1995; Whitehurst et. al., 1988). The benefit of using a dialogic reading-like technique to manipulate lexical awareness is that it children can be asked about the novel vocabulary as well as other elements of the story. Asking if a child knows the meaning of novel items could scaffold awareness of lexical ignorance and lead to word learning. Another possibility is that asking any type of question heightens preschoolers' attention to the text and promotes word learning.

In the current study we explore these possibilities by asking questions that draw attention to lexical ignorance, asking a question unrelated to the target word, or asking no questions. We will also test skills that may influence word learning and the ability to access the lexicon such as short-term memory, object representation and existing language ability. We predict that children with more robust memory, object representation, and language abilities will learn more words during our task. By testing cognitive abilities we hope to more fully explain the developmental trend in lexical awareness.

Method

Participants

Ninety-one 3-4-year olds (range 36-63 months, M = 47 months, 44 Females) were recruited from childcare centers and state birth records in the southeastern United States. Eight additional children were recruited and not included in the analysis for non-compliance (n = 2), biased responding (n = 2) and experimenter error (n = 4). Participants were typically developing, had intact hearing, and heard English in their household 70% of the time or more.

Demographic surveys revealed that 53% of mothers had a post graduate degree or some graduate school, 36% had a college degree or some college, and 6% had a technical/AA degree or a high school degree. Four participants' parents did not respond to this question. 22% of participants' families reported an income of $150,000 or more per year, 25% reported an income between $100,000 to $150,000 per year, 30% reported an income between $100,000 and $50,000 per year, and 12% of families reported an income of less than $50,000 per year. Ten participants' parents did not respond.

Design

All participants were administered five tasks in the same order: the electronic picture book, lexical awareness task, a forward digit span, Test of Early Language Development -3 (TELD-3; Hresko, Reid, & Hammill, 1999) and an object representation task called the Vanderbilt Expertise Task-Kid (VET-Kid). Children tested in the lab completed the tasks during a single visit and most of the participants tested in child-care centers did four of the tasks on one day and the VET-Kid a few days later (Range: 0-21 days, M = 2.55, sd = 3.50).

For the electronic picture book, children were randomly assigned to one of three conditions so that ages were and distribution of males and females across conditions was equal (to be described in more detail below): *lexical awareness (n = 26, 11 females), distractor question (n = 31, 14 females), no question (n = 34, 16 females) See Table 2 for mean ages by condition*.

Materials

Electronic Picture Book. Participants were read an electronic picture book on a laptop that was created for this experiment. It contained six novel target words, which were names for made-up creatures. In the picture book, two people (a brother and sister) participated in various activities (e.g., playing in the garden, eating ice cream) and found novel creatures. Illustrations were presented through Microsoft Power Point on a laptop computer. An experimenter read the text from a separate binder.

For test trials, children were presented each target item with two distractors. The distractors matched the target in color but differed in location and shape (See figure 1). One of the distractors was another novel creature and one was a familiar creature in an unfamiliar color (e.g. a purple pig). The order of the target items and the order of the dimensions were counterbalanced across participants.

Lexical awareness task. A task that assessed participants' lexical awareness was created for the purpose of this experiment. Twelve word pairs consisted of a familiar (i.e. *old*) word paired with a novel (i.e. *new*) word. The familiar words in the pairs increased in difficulty for later items (the words were from the PPVT-IV). A PowerPoint of pictures testing familiar word knowledge was also created.

Object representation task. To determine how well each participant could mentally represent an object, we administered a task that was an adaptation of the Vanderbilt Expertise Task (McGugin, Richler, Herzmann, Speegle, & Gauthier, 2012) originally designed for use with adults. This task was programed on MatLab. Images of three object types, teddy bears, toy cars and kid's shoes were gathered from Google images and shopping websites such Ebay (See *Table 1* for some example pictures).

Procedure

To begin, children were seated at a table across from the experimenter and began the electronic picture book task. There were three conditions that differed according to what

the experimenter said after a novel word was mentioned (e.g., *Grimp* in "They went into the garden to look for a *grimp*."). In the *lexical awareness* condition, the researcher asked children if they knew what the target word meant after a novel word was mentioned (Do you know what that is?"). Children were expected to say "no" as the target words were all novel, but if they said "yes" they were asked to provide a definition. Most participants then admitted that they did not know what it was. However, if they made up a definition, they were told that the word was one they did not know. In the *distractor question* condition, children were asked about an element of the story that was not target-word related ("Do you have a garden?"), and in the *no question* condition the story was read straight through with no interactive questions from the researcher to the participant.

After the target item was mentioned in the electronic picture book and participants were either asked a question, or not, the target was described in three dimensions: color, location and shape. For example, the researcher would tell the participant, "A grimp is orange, has a droopy nose and lives in the trees." The descriptions were given without visual reference. Afterwards participants were asked, "Which creature did they see?" and were presented with three pictures, one target and two distractors. The placement of the target words and distractors were counterbalanced across trials and participants. There were six target words and six different target-word presentation orders for each condition.

Following this, children completed the **lexical awareness task**. Participants were told that they were going to play a game with words where they would hear two words; one was an *old* word that they had heard before and knew what it meant, for example, the word "book." They were told the other word was going to be a *new* word that they had not heard before and didn't know what it meant, like the word "floopydoopy." They were then given two practice trials in which they were given word pairs (e.g. *sock* and *baloota*) and asked to say the *new* word. For each practice trial, if participants responded incorrectly (i.e. by saying the *old* word) they were asked if they knew the definition of the *old* word (e.g. sock) and were guided to the correct answer. Once they provided or were given the correct answers on the two practice items they proceeded to the testing phase.

Participants were reminded about saying the *new* word before each pair was presented. Additionally every four items they were reminded that *new* words were words they had never heard before. The order of familiar and novel words was counterbalanced across items.

After the pairs were administered we tested participants' knowledge of the familiar words by presenting them with a three-option forced task in which they selected the familiar referent (e.g. a picture of a *cat*) among two distractor pictures (e.g. a picture of an oven and a frog). If the participant could not identify the familiar item when prompted then those items were excluded. The number of errors on the task (i.e. saying the known word was the word they did not know) was the measure of performance. If participant failed to identify any particular familiar word the trial was not included as an error, because both words in the pair would have been unknown for the child.

To assess short-term memory children were given a forward digit span task. Then their language abilities were tested with the TELD-3.

The VET-Kid task tested **object representation**. For this task we tested children on three different categories: teddy bears, toy cars and shoes. Children were asked to remember three exemplars of a category (e.g., three teddy bears) in the context of story about a character named Casey who lost his three teddy bears and needed the child to help find them. During training they were shown Casey's three teddy bears. Then they are asked, "Can you find one of Casey's bears" when presented with an array of three teddy bears that included one of Casey's bears.

For the first nine trials participants were asked to identify pictures of the items that were identical to the ones they were trained on (OR-Easy). They received feedback on the first 9 trials, and showing them the target item if they were incorrect. Then children completed twelve transfer trials in which children were asked to identify the target items that were shown from a different angle (OR-Hard). For the transfer trials, children did not receive feedback. Three additional study trials, as well as three catch trials where the target object was paired with unrelated distractors were interspersed to ensure participant's attention and boost confidence. See *Table 1* for the design matrix.

Results

Preliminary analyses revealed no differences in age, lexical awareness, TELD-3, digit span, or object representation between the participants in the three conditions (see Table 2).

A planned comparison revealed that children in the *lexical awareness* condition (M = 3.36, sd = 2.00) learned significantly more words than children in *distractor question* (M = 2.45, sd = 1.67, $p < .05$) and *no question* (M = 2.44, sd = 1.36, $p < .05$) conditions.

To test whether children were able to identify words beyond what would be expected by chance, we conducted tests against chance for each condition. In the two control conditions word learning performance did not differ from chance, for the *distractor question* ($t(30) = 1.51$, $p = .14$) and the *no question* ($t(35) = 1.96$, $p = .06$) conditions, but the *lexical awareness* condition boosted children's word learning above chance levels ($t(25) = 3.40$, $p = .002$).

For the purposes of the rest of the analysis we collapsed across the two control conditions (since responding in these two conditions did not differ). Bivariate correlations revealed that language and memory abilities were correlated with word learning: TELD-3 ($r(87) = .32$, $p = .002$), lexical awareness ($r(86) = .29$, $p = .006$), digit span ($r(91) = .33$, $p = .001$) and object representation ($r(87) = .37$, $p < .001$).

We conducted a multiple-regression with age, condition, TELD-3, lexical awareness, digit span and object representation as predictors of word learning. Our model was significant ($F(6,77) = 4.50$, $R^2_{adjusted} = .21$, $p < .001$) and condition was a significant predictor ($t(77) = 2.58$, $\beta = .26$, $p = .012$), as was short term memory ($t(77) = 2.16$, $\beta = .25$, $p = .03$). Object representation also emerged as a marginal predictor ($t(77) = 1.93$, $\beta = .25$, $p = .06$). See *Table 3* for regression table.

Discussion

We investigated whether adults could scaffold lexical awareness for preschool aged children by asking questions that made children consider whether they know the meaning of a novel word. We also investigated whether cognitive abilities influenced word learning, and if these skills might explain variability in learning. Our results suggest that both asking questions that highlight lexical ignorance and memory abilities affected children's word learning success.

In this study, age, language abilities and a measure of lexical awareness did not independently predict word learning above digit span (short-term memory) or object representation. This was surprising, in part, because previous research has found age and language ability to be important predictors of word learning during shared reading (Sénéchal et al., 1995). One explanation for our results is that the current task loaded heavily on children's memory skills because the descriptions were offered in the absence of the referent objects. Children therefore had to hold information about features in mind rather than immediately mapping a label to a visible referent. The failure of our lexical awareness measure to predict success may indicate that children do not employ their understanding automatically during the preschool period. In other words, it may be that preschoolers sometimes know that they do not know a word, but there are not sure how to act on their ignorance. This is consistent with research suggesting that practice with novel and familiar words and objects leads to greater awareness of lexical ignorance in preschoolers (Marazita & Merriman, 2004; Hartin et al, 2016).

In the current study, the lexical awareness condition may have worked in two ways: it could be drawing attention to the fact that the child does not know the word; alternatively, it could be drawing increased attention to the word form. To test these two processes separately in future work, we plan to draw attention to the phonology of the new word separately from having children think about the meaning of the word. Additionally, in future work we will more directly test word learning by asking children, "Point to the *grimp*," instead of, "Which creature did they see?"

This research provides suggestive evidence that children may use what they do and do not know to propel their vocabulary acquisition and adults' question asking might support preschoolers' metacognitive abilities.

Tables

Table 1: Object Representation Task.

Example Stimuli	Trial type	# of trials
	Study	1
	Identical test	6
See above	Study	1
See above	Identical test	3
	Catch	1
	Transfer test	3
See above	Study	1
See above	Transfer test	3
See above	Catch	1
See above	Transfer test	3
See above	Study	1
See above	Transfer test	3
See above	Catch	1

Table 2: Means and standard deviations of study measures.

Variable	Lexical Awareness	Distractor Question	No Question
Age	48.44 (7.47)	45.42 (10.60)	48.69 (7.18)
Word ID	3.36 (2.00)	2.45 (1.67)	2.44 (1.36)
LA	7.80 (3.33)	8.41 (3.45)	6.72 (3.65)
TELD-3	27.09 (3.94)	26.13 (4.17)	26.03 (4.53)
Digit	6.24 (1.51)	5.90 (1.68)	6.00 (1.26)
Object Rep	13.25 (3.83)	13.15 (3.88)	11.96 (4.43)
OR-Easy	5.52 (1.71)	5.67 (1.69)	5.09 (2.14)
OR-Hard	7.73 (2.25)	7.48 (2.46)	6.86 (3.65)

Table 3: Regression predicting word learning.

Predictors	B	SE B	β
Age	-.01	.02	-.07
Condition	.97	.37	.26*
Digit-span	.28	.13	.25*
TELD-3	.02	.05	.04
Lexical Awareness	.05	.06	.87
Object Representation	.11	.06	.06^

^ = p < .1. * = p < .05.

Figures

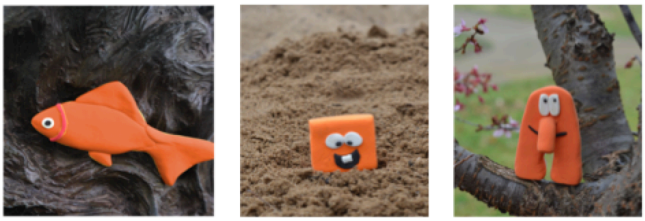


Figure 1: Example test trial from electronic picture book.

Acknowledgments

This research was supported by the Vanderbilt Psychology and Human Development departmental funds as well a National Science Foundation—Graduate Research Fellowship awarded to the first author (1445197). We would like to sincerely thank the members of the Language Development Lab for their help with recruitment, data input and organization as well as the parents and preschools that allowed their children in to participate in this study.

References

Bartsch, K., and Wellman, H. M. (1995). Children talk about the mind. New York: Oxford University Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. doi:10.1037/0003-066x.34.10.906

Hartin, T. L., Stevenson, C. M., & Merriman, W. E. (2016). Preexposure to Objects that Contrast in Familiarity Improves Young Children's Lexical Knowledge Judgment. *Language Learning and Development*, 1-17.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999).The Test of Early Language Development-3(TELD-3). Austin, TX: Pro-Ed.

Jacobson, R. & Saylor, M. M. "What does that mean?" Exploring information-seeking behavior in young children" (in prep)

Johnson, C. N., & Maratsos, M. P. (1977). Early comprehension of mental verbs: Think and know. *Child Development*, 48, 1743-1747.

Klin, C. M., Guzman, A. E., and Levine, W. H. (1997). Knowing that you don't know: Metamemory and discourse processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 1378–1393.

Lyons, K. E., & Ghetti, S. (2011). The development of uncertainty monitoring in early childhood. *Child Development*, *82*(6), 1778–87.

Marazita, J. M., & Merriman, W. E. (2004). Young children's judgment of whether they know names for objects: The metalinguistic ability it reflects and the processes it involves. *Journal of Memory and Language*, *51*(3), 458–472.

McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., & Gauthier, I. (2012). The Vanderbilt Expertise Test reveals domain-general and domain-specific sex effects in object recognition. *Vision Research*, *69*, 10–22.

Merriman, W. E., Evey-Burkey, J. A., Marazita, J. M., and Jarvis, L. H. (1996). Young two-year-olds' tendency to map novel verbs onto novel actions. Journal of Experimental Child Psychology, 63, 466–498.

Merriman, W. E., & Marazita, J. M. (2004). Young children's awareness of their own lexical ignorance: Relations to word mapping, memory processes, and beliefs about change detection. In D. T. Levin (Ed.), *Thinking and Seeing: Visual Metacognition in adults and children.* (57-73). MIT Press.

Merriman, W. E., and Schuster, J. M. (1991). Young children's disambiguation of object name reference. Child Development, 62, 1288–1301.

Sénéchal, M., & Cornell, E. H. (1993). Vocabulary Acquisition through Shared Reading experiences. *Reading research quarterly*, *28*(4), 360–374.

Sénéchal, M., Thomas, E., & Monker, J. (1995). Individual Differences in 4-Year-Old Children's Acquisition of Vocabulary During Storybook Reading. *Journal of Educational Psychology*, *87*(2), 218–229.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 21, 360–406. doi:10.1598=RRQ.21.4.1

Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., Debaryshe, B. D., & Caulfield, M. (1988). Accelerating Language Development Through Picture Book Reading. *Developmental Psychology*, *24*(4), 552–559.

Collective search on rugged landscapes: A cross-environmental analysis

Daniel Barkoczi (barkoczi@mpib-berlin.mpg.de) **Pantelis P. Analytis** (analytis@mpib-berlin.mpg.de)
Charley M. Wu (cwu@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development

Abstract

In groups and organizations, agents use both individual and social learning to solve problems. The balance between these two activities can lead collectives to very different levels of performance. We model collective search as a combination of simple learning strategies to conduct the first large-scale comparative study, across fifteen challenging environments and two different network structures. In line with previous findings in the social learning literature, collectives using a hybrid of individual and social learning perform much better than specialists using only one or the other. Importantly, we find that collective performance varies considerably across different task environments, and that different types of network structures can be superior, depending on the environment. These results suggest that recent contradictions in the social learning literature may be due to methodological differences between two separate research traditions, studying disjoint sets of environments that lead to divergent findings.

Keywords: Social learning; communication networks; collective behavior; search; rugged landscapes.

Introduction

There are two paths to the acquisition of knowledge. Organisms can search for new solutions through trial-and-error, testing various courses of action in isolation from others, or alternatively, they can copy existing solutions by imitating other individuals (March, 1991; Rendell et al., 2010). Imitation is a cognitively simple strategy that can lead to exceptionally good outcomes (Boyd & Richerson, 1985; Rendell et al., 2010; Laland, 2004; Miller & Dollard, 1941), however, it cannot produce new information by itself. Thus, collective learning in groups can be seen as the outcome of both individual learning and the spread of social information through imitation (Tarde, 1903). A different balance of theses two strategies can lead to very different levels of individual and group performance (Rogers, 1988).

In this paper we examine how collectives using individual and/or social learning perform across a wide range of environments embedded in different network structures, governing the communication of social information. Models of social exploration-exploitation dynamics regularly use the analogy of search on a fitness landscape where nearby (similar) solutions can have quite different payoffs, forming a "rugged" multi-peaked landscape (Levinthal, 1997; Lazer & Friedman, 2007; Mason & Watts, 2012). Real-world examples of rugged landscapes may include combinatorial problems, technological innovation, or the fitness of an organism as a function of its genome. A key feature of these environments is that they can be dominated by multiple local maxima, where the best solution in a given neighborhood may be far from the best global solution (see Figure 1 for an illustration).

Behavioral scientists have conducted a handful of experimental studies on collective learning using one or two dimensional functions (Mason, Jones, & Goldstone, 2008; Mesoudi, 2008; Mason & Watts, 2012), or simple combinatorial problems (Wisdom, Song, & Goldstone, 2013). In contrast, management and organization scientists have studied search using simulations and experiments based on the NK model inspired by evolutionary biology (Kauffman & Levin, 1987; Levinthal, 1997; Gavetti & Levinthal, 2000; Rivkin, 2000; Billinger, Stieglitz, & Schumacher, 2013), where the environment is completely determined by two parameters, the N number of components making up a solution, and the K level of interdependence between components. This type of formalism has the advantage that the difficulty of the problem can be directly set by the parameter K, but also the drawback that the problem is specifically tailored to a single type of complexity (i.e., interdependence).

Both communities have investigated the influence of communication networks on collective performance, but have found contradictory results (Derex & Boyd, 2016; Fang, Lee, & Schilling, 2010; Lazer & Friedman, 2007; Mason & Watts, 2012; Mason et al., 2008; Wisdom et al., 2013). Lazer and Friedman (2007) found that less connected (*inefficient*) networks lead to better collective performance in the NK landscape, while Mason and Watts (2012) reached the opposite conclusion, finding support for the superiority of well-connected (*efficient*) networks in a 2-dimensional landscape.

In these studies, the choice of fitness landscape was held constant, with results often derived from a single type of environment. Here we explore a promising explanation for these contradictory results, namely that these studies investigated learning on disjoint sets of environments, differently suited to specific types of collective search behavior. There are also other possible explanations for these contradictory results, such as the use of different social learning strategies (Barkoczi & Galesic, 2016) or the nature of the task that the collective has to perform (Shore, Bernstein, & Lazer, 2015).

In this study, we present simulations of collective search across a wide range of multi-peak environments, which include both two dimensional (Mason & Watts, 2012) and NK environments (Lazer & Friedman, 2007). Our approach uses a cross-environmental analysis to resolve inconsistencies regarding the influence of network structure on collective performance.

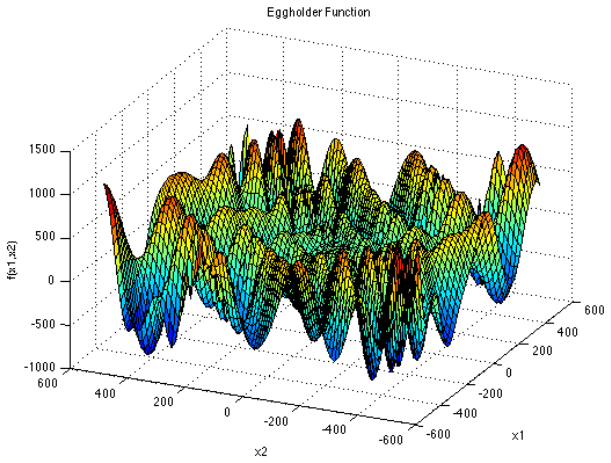


Figure 1: The Eggholder function is a typical example of a bivariate function that generates a rugged payoff landscape. The exact formula can be found at the virtual library of simulation experiments http://www.sfu.ca/~ssurjano/optimization.html.

Methods

Landscapes

In total we studied 15 different multi-peaked landscapes (Table 1). Environments 1-12 are bivariate functions obtained from the virtual library of simulation experiments (Surjanovic & Bingham, n.d.), which are regularly used in operations research and the field of global optimization to study how different optimization algorithms perform (e.g., Hu, Fu, & Marcus, 2007). These environments have been designed and investigated specifically because they pose challenges to adaptive optimization algorithms, covering a wide range of possible environmental structures with regards to the variability of high quality solutions, the ruggedness of the landscape, and the average pay-off. We also study the environment used in Mason and Watts (2012) and two different NK landscapes (K=5 and K=10) of the type used in Lazer and Friedman (2007).

For each environment we normalized the payoff scales to obtain relative payoffs between 0 and 1, with the global maximum was set to 1 and the global minimum set to 0. Following Lazer and Friedman (2007) we re-scaled these normalized payoffs by raising each value to the power of 8. This monotonic transformation creates larger differences in the upper range of payoffs but leaves other features of the landscape unchanged.

Function-based Landscapes. Environments 1-12 are generated from bivariate functions, which translate any two values x and y (evaluated on a specific range of real numbers) into a corresponding payoff between 0 and 1. For our purposes, we first transformed the continuous environments into discrete units by dividing each axis into 1001 equally-spaced regions. Since all the environments were two dimensional, this resulted in a total of $1001 \cdot 1001$ possible locations that an agent could explore. This harmonized the different x-axis and y-axis scales of the individual environments, resulting in the same number of possible solutions across environments.

Most of these environments were initially designed as minimization problems. To be consistent with research on collective behavior in the behavioral and social sciences we inverted the payoffs and turned them into maximization problems.

Mason & Watts and NK Environments. We replicated the environment used in Mason and Watts (2012) and NK environments from Lazer and Friedman (2007). To retain consistency with the function-based environments described above, we enlarged the Mason and Watts (2012) environment to have $1001 \cdot 1001$ possible locations using bicubic interpolation. For the NK environments, we use $N = 20$, resulting in a similar number of possible solutions (2^{20}).

#	Environments	μ	σ	modality
1	Ackley	0.001	0.01	4489
2	Cross-in-Tray	0.14	0.17	64
3	Drop-wave	0.003	0.03	4391
4	Eggholder	0.03	0.07	353
5	Griewank	0.17	0.24	50861
6	Holder table	0.006	0.06	56
7	Langermann	0.02	0.07	2461
8	Rastrigin	0.05	0.1	121
9	Schaffer n.2	0.007	0.03	88457
10	Schaffer n.4	0.001	0.02	87737
11	Schwefel	0.03	0.07	64
12	Shubert	0.005	0.03	761
13	Mason & Watts (2012)	0.001	0.01	1090
14	$N = 20, K = 5$	0.04	0.04	1143
15	$N = 20, K = 10$	0.04	0.04	7131
	Average	0.036	0.07	16612

Table 1: The 15 landscapes that we examined. We report the mean (μ) and variability (σ) of pay-offs, and the modality (i.e., number of peaks) of the landscape. Environments 1-12 are function-based and the environmental statistics remain constant. Statistics for Mason and Watts (2012) were averaged over 10,000 replications of the environment generation function, while the NK environments were averaged over the 100 pre-generated landscapes used in simulations.

Learning processes and behavioral strategies

We considered a population of 100 agents simultaneously learning about the environment using different strategies. In separate conditions, we tested strategies that relied solely on individual learning, solely on social learning, or on a combination of both (McElreath et al., 2008). We first describe the rules associated with each strategy.

1. *Individual learning:* We studied two types of individual learning, where agents evaluated new solutions without the benefit of social information.

   - **Local search.** In our simulations, local search (i.e., hill climbing) was performed by examining neighboring solutions for the largest increase in payoff. In cognitive science, local search has been studied in resource-allocation problems (Busemeyer & Myung,

1987; Rieskamp, Busemeyer, & Laine, 2003), and also corresponds to the gradient descent algorithm used in many learning systems. On the 2D landscapes (Environments 1-13), the local search strategy evaluated the 8 adjacent solutions accessible by modifying either the value of x, y, or both simultaneously by one discrete unit. If any of the explored solutions were better than the current one, it adopted the best among these solutions, otherwise, it kept the current solution. In NK landscapes, local search was implemented by modifying a single randomly chosen digit of the N-dimensional solution (Levinthal, 1997; Lazer & Friedman, 2007; Rivkin, 2000) [1].

- **Random search.** Random search was performed by randomly evaluating a new location in the fitness landscape and moving to that position if it had a higher payoff (i.e., long-jump). Random search relates to the idea of blind variation in psychology (Campbell, 1960), but also corresponds to the random search assumption made by optimal stopping models in economics and statistics (e.g. DeGroot, 2005; Analytis, Stojic, & Moussaïd, 2015).

2. *Social learning:* Social learners sampled the solutions from a number of other individuals (n=k) within the connected population (see Network structures), and adopted the solution with the highest observed payoff if it was better than the agent's existing solution (Lazer & Friedman, 2007). In this paper we report results for n=3. We refer to this this strategy as **imitate-the-best**.

3. *Hybrid social/individual learning:* Hybrid agents used a combination of social and individual learning (Enquist, Eriksson, & Ghirlanda, 2007; McElreath et al., 2008). At each time step, agents followed a sequential strategy selection process (Gigerenzer, 2008), in which they first attempted to find a better solution through social learning (described above), but switched to individual learning, using either local search or random search, if unsuccessful. In this paper we report two variants of this strategy, one with no option to perform random search (rs = 0) and one where 20% of individual learning actions resulted in random search (rs = .2), with the remaining 80% of actions performing local search.

Network structures

To account for different channels of communication, we examined how hybrid agents performed when they were embedded in a fully connected network, allowing communication with all other agents, or a locally connected lattice in which each agent was connected with four other agents from the population (Figure 2). In both networks the agents sampled n=3 individuals at random from their connections. In the locally connected lattice, access to the full population was only available indirectly, over several degrees of separation. In contrast, information flowed more freely in the fully connected network, with the possibility for each agent to imitate each other agent in the population. These two network structures were also studied by Lazer and Friedman (2007).

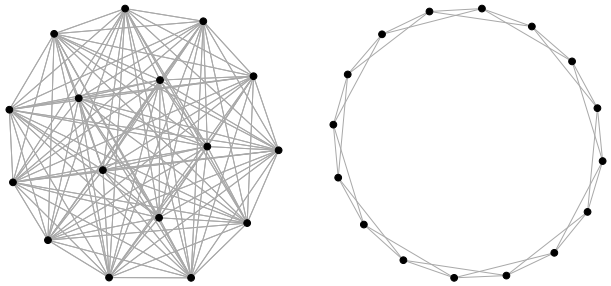


Figure 2: Examples of a fully connected (left) and locally connected lattice (right) networks.

Simulation procedure

We assigned random starting solutions to a population of 100 agents. At each time step, agents used the strategies described above to learn about the environment. Imitation actions were performed simultaneously at the beginning of each time step to avoid sequential effects (Analytis et al., 2015; Bikhchandani, Hirshleifer, & Welch, 1992).

We repeated this process for 100 time steps and 100 replications, and recorded the average payoff achieved in the population. In total we tested 7 different strategy and network combinations; the three pure learning strategies (local search, random search, and imitate-the-best[2]) and four versions of the hybrid strategy (two rs values and two different network structures)[3].

Results

Overall, we found that performance varied across environments, with some environments being more challenging that others (Figure 3). While the rank order remained relatively stable across environments, there were several notable differences. First, random search was always superior to local search in the 2D environments (Environments 1-13), while the opposite was the case in the NK environments (Environments 14-15). Second, imitate-the-best typically outperformed individual search strategies–by driving the population to local peaks faster than individual search–with one exception being the NK environments, where local search performed better. These two results demonstrate that local search is a highly effective strategy in the NK environment, which could be because the peaks of NK landscapes have larger basins of attraction, allowing local search to explore a larger space compared to the 2D landscapes (Kauffman & Levin, 1987).

[1] The differences in the implementation of local search in 2D and high dimensional NK landscapes are due to irreconcilable differences in the number of adjacent solutions as a function of the dimensionality of the problem space. We follow standard implementations from the respective literatures.

[2] Note that imitate-the-best is used as a benchmark, therefore, we only study its performance in a fully connected network.

[3] See: https://github.com/dnlbrkc/collective_search for the simulation code.

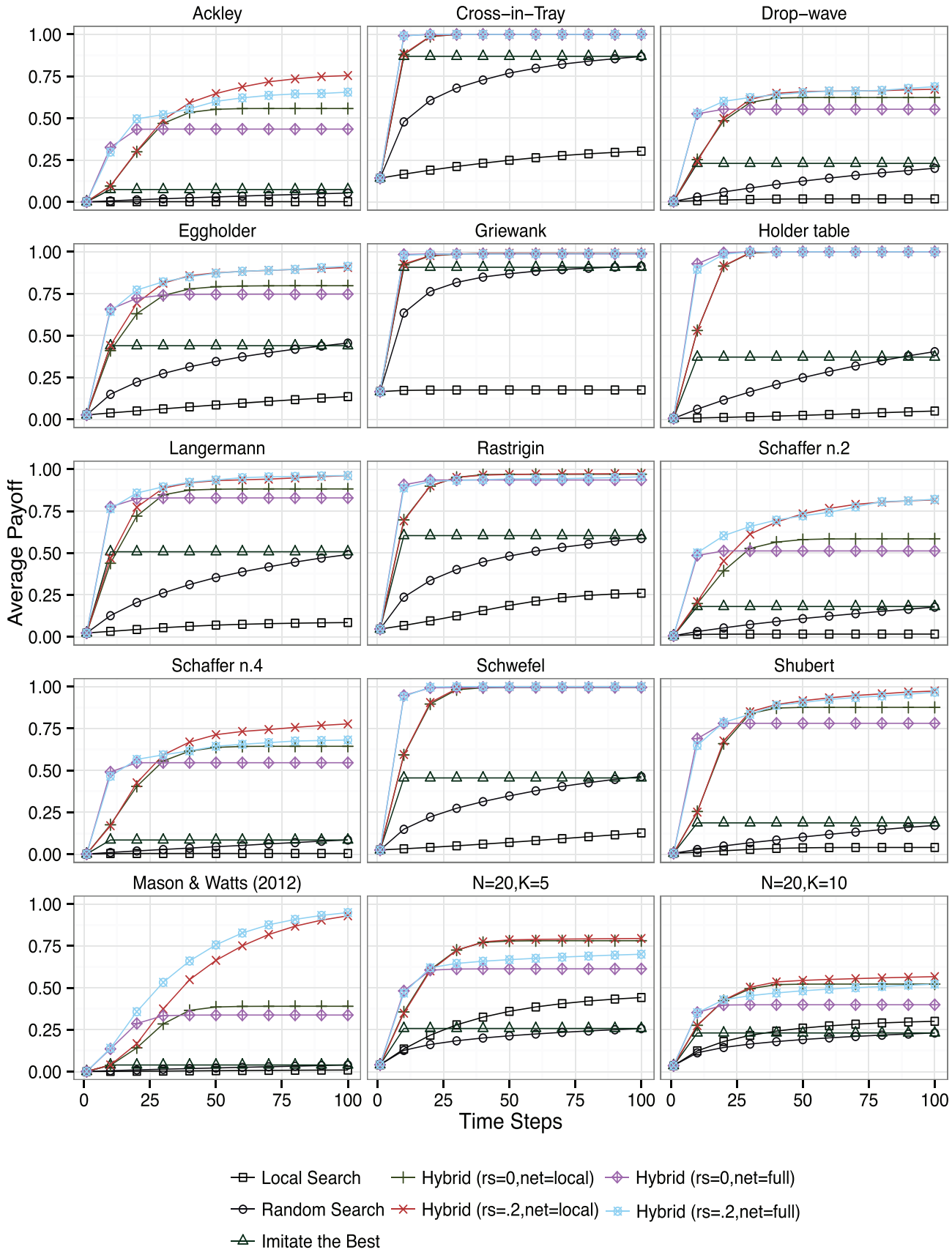


Figure 3: Learning curves for the different models across 15 environments, reporting the change in average payoff (over each population of 100 agents and over 100 replications) for 100 time steps. We studied 7 different models using combinations of simple heuristic rules. Hybrid models utilize both social and individual search strategies, where the random search parameter (*rs*) governs the balance of random search vs. local search, while the network parameter (*net*) specifies either a fully connected or a locally connected lattice network. Hybrid strategies consistently performed best with some random search ($rs = .2$); however, the superiority of different network structures depended on the structure of the environment.

Network structure and performance We replicated the results found in both Mason and Watts (2012) and Lazer and Friedman (2007), and show that the contradictory conclusions can be explained by the choice of environment they studied. Depending on the environment, network structure can lead to three different types of results. First, hybrid agents with a local lattice network can converge on a higher payoff than those with a fully connected network (particularly when $rs = 0$; e.g., Ackley, Schaffer n.4, and NK environments). This replicates the results of Lazer and Friedman (2007), who showed that local connections force the population to explore the landscape more extensively, delaying convergence on local optima. Second, fully connected networks can initially outperform the local lattice network, and reach comparable performance over time (e.g., Eggholder and Mason & Watts environments). This result is consistent with Mason and Watts (2012) and shows that in some environments, slower communication does not lead to higher performance in the long-run. Third, there can very small differences between networks (e.g., Cross-in-Tray and Griewank), which happens when the problem is relatively easy and hybrid agents quickly reach high-value optima.

The role of random exploration Allowing for a mixture of both local exploration (hill-climbing) as well as the ability to occasionally undertake random exploration (long-jumps) greatly benefited hybrid collectives, making it the overall best performing strategy (*rs*=.2) by a large margin. Random search generated new information that gradually diffused to other agents by means of imitation. In fact, heuristics for global optimization problems such as genetic algorithms (Goldberg & Holland, 1988) or simulated annealing (Kirkpatrick, Gelatt, Vecchi, et al., 1983) have explicit strategies to escape from the local optima. The ability to search randomly (via long-jumps) in our simulation played exactly the same role.

Explaining differences across environments. What are the environmental features that influence the performance of different strategies and networks? An obvious candidate is the the number of peaks (i.e., modality) in an environment. Figure 4 shows that modality can explain some of the variability in performance across environments, but not all. Higher modality is correlated with lower payoffs for both the locally connected (top; r=-.53, p=.04) and fully connected (bottom; r=-.55, p=.03) hybrid strategies (where $rs = .2$). Still, a large proportion of the variance remains unexplained. One striking difference is that while the Mason & Watts (2012) and N=20, K=5 environments have similar modality, the rank order of average payoffs are inverted between the two network structures. In the future we aim to identify relevant features of task environments (see Mersmann et al., 2011), that would allow us to make predictions about which cognitive strategies and network structures are best suited to a particular environment.

General discussion

In this paper we investigated how groups of individuals relying on different cognitive strategies performed across 15 challenging search environments. As Herbert Simon illustrated with his scissors analogy, "[h]uman rational behavior (and the rational behavior of all physical symbol systems) is shaped by a [pair of] scissors whose two blades are the structure of task environments and the computational capabilities of the actor" ((1990), p. 7). In a similar vein, to understand collective performance, it is essential to study cognition in tandem with the task environment in which it operates.

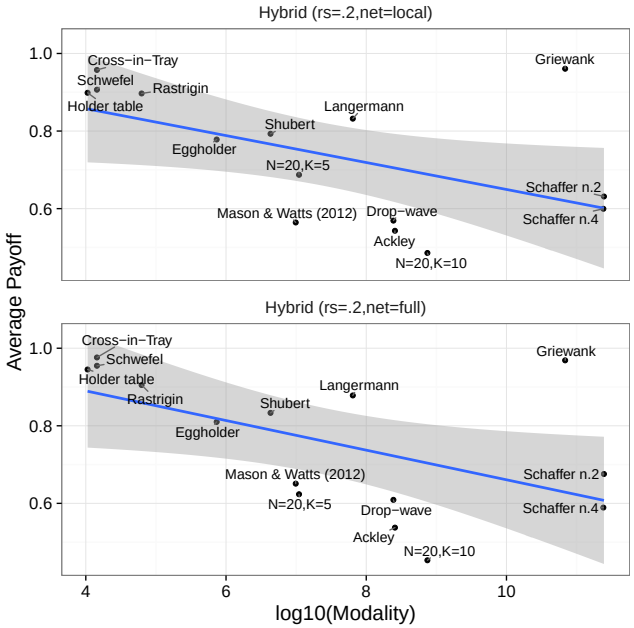


Figure 4: The number of local maxima (modality) explains some of the variability in performance across environments, but not all. Here we show average payoff as a function of the (log_{10}) modality of the environments, for the locally connected (top) and fully connected (bottom) hybrid models (where $rs = .2$). Average payoff was derived by averaging over all individuals and time steps.

A large number of studies have addressed collective problem solving; however, they have focused on only one type of task environment. This makes it (i) hard to compare results across studies, (ii) uncertain whether they generalize beyond a specific environment, and (iii) unknown whether certain strategies lead to superior collective performance when paired with specific environments. To the best of our knowledge, this is the first study in the social and behavioral sciences that analyses collective problem-solving across many different types of environments[4].

Network and environmental structure

An open question in the study of collective learning is the influence of network structure on collective performance in groups and organizations. Derex and Boyd (2016), Lazer and Friedman (2007), Fang et al. (2010) and Mason et al. (2008) found that less efficient networks may led to higher

[4]To this end, we borrowed several environments that have been used as test-beds for search algorithms in Operations Research. Conversely, researchers from the OR community have often turned to collective cognitive systems for inspiration. Algorithms imitating the behavior of ant colonies (Bonabeau, Dorigo, & Theraulaz, 2000) and swarms (Kennedy, 2010) have proved to be particularly powerful in solving combinatorial and global optimization algorithms.

levels of collective performance. In contrast, a recent study by Mason and Watts (2012) came to the opposite conclusion, that more efficient networks are better. A crucial difference between these studies is the type of environment that they investigated, because as we have found, the performance of a type of network structure depends on the underlying task environment. We found that in a number of environments (e.g., Ackley, Schaffer n.4, and NK environments), inefficient networks eventually converged on better payoffs than efficient networks, while in others (including the Mason & Watts environment), efficient networks had an initial advantage, but reached similar performance as inefficient networks over time.

Extensions and limitations

In this study we chose to adhere closely to the designs of Mason and Watts (2012) and Lazer and Friedman (2007) to resolve contradictory results in the literature on the role of network structure in collective performance. We covered a broad range of environmental structures, but we only looked at two popular types of networks. In the future, we intend to extend our analysis to additional networks (see Mason & Watts, 2012), different behavioral rules (see Barkoczi & Galesic, 2016) and types of tasks (see Rahwan, Krasnoshtan, Shariff, & Bonnefon, 2014). However, the key challenge will be to identify features of tasks or environments that favor the usage of certain networks over others.

Acknowledgments

We thank Stojan Davidovic and Mirta Galesic for their useful comments on an earlier version of this manuscript.

References

Analytis, P. P., Stojic, H., & Moussaïd, M. (2015). The collective dynamics of sequential search in markets for cultural products. *Santa Fe Institute Working Paper*.

Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Working paper*.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, *100*(5), 992-1026.

Billinger, S., Stieglitz, N., & Schumacher, T. R. (2013). Search on rugged landscapes: An experimental study. *Organization Science*, *25*(1), 93–108.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (2000). Inspiration for optimization from social insect behaviour. *Nature*, *406*(6791), 39–42.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process.* University of Chicago Press.

Busemeyer, J. R., & Myung, I. J. (1987). Resource allocation decision making in an uncertain environment. *Acta Psychologica*, *66*(1), 1–19.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review*, *67*(6), 380.

DeGroot, M. H. (2005). *Optimal statistical decisions* (Vol. 82). John Wiley & Sons.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 201518798.

Enquist, M., Eriksson, K., & Ghirlanda, S. (2007). Critical social learning: a solution to rogers' paradox of nonadaptive culture. *American Anthropologist*, *109*(4), 727–734.

Fang, C., Lee, J., & Schilling, M. A. (2010). Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, *21*(3), 625–642.

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly*, *45*(1), 113–137.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on psychological science*, *3*(1), 20–29.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning*, *3*(2), 95–99.

Hu, J., Fu, M. C., & Marcus, S. I. (2007). A model reference adaptive search method for global optimization. *Operations Research*, *55*(3), 549–568.

Kauffman, S., & Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *Journal of theoretical Biology*, *128*(1), 11–45.

Kennedy, J. (2010). Particle swarm optimization. In *Encyclopedia of machine learning* (pp. 760–766). Springer.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., et al. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Laland, K. N. (2004). Social learning strategies. *Animal Learning & Behavior*, *32*(1), 4–14.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, *52*(4), 667–694.

Levinthal, D. A. (1997). Adaptation on rugged landscapes. *Management Science*, *43*(7), 934–950.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*(1), 71–87.

Mason, W., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, *137*(3), 422.

Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769.

McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3515–3528.

Mersmann, O., Bischl, B., Trautmann, H., Preuss, M., Weihs, C., & Rudolph, G. (2011). Exploratory landscape analysis. In *Proceedings of the 13th annual conference on genetic and evolutionary computation* (pp. 829–836).

Mesoudi, A. (2008). An experimental simulation of the copy-successful-individuals cultural learning strategy: adaptive landscapes, producer–scrounger dynamics, and informational access costs. *Evolution and Human Behavior*, *29*(5), 350–363.

Miller, N. E., & Dollard, J. (1941). *Social learning and imitation.* Yale University Press.

Rahwan, I., Krasnoshtan, D., Shariff, A., & Bonnefon, J.-F. (2014). Analytical reasoning task reveals limits of social learning in networks. *Journal of The Royal Society Interface*, *11*(93), 20131211.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., ... Laland, K. N. (2010). Why copy others? insights from the social learning strategies tournament. *Science*, *328*(5975), 208–213.

Rieskamp, J., Busemeyer, J. R., & Laine, T. (2003). How do people learn to allocate resources? comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1066.

Rivkin, J. W. (2000). Imitation of complex strategies. *Management Science*, *46*(6), 824–844.

Rogers, A. R. (1988). Does biology constrain culture. *American Anthropologist*, 819–831.

Shore, J., Bernstein, E., & Lazer, D. (2015). Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. *Organization Science*.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Surjanovic, S., & Bingham, D. (n.d.). *Virtual library of simulation experiments: Test functions and datasets.* Retrieved January 26, 2016, from `http://www.sfu.ca/~ssurjano`.

Tarde, G. D. (1903). *The laws of imitation*. H. Holt.

Wisdom, T. N., Song, X., & Goldstone, R. L. (2013). Social learning strategies in networked groups. *Cognitive Science*, *37*(8), 1383–1425.

Is it a nine, or a six? Prosocial and selective perspective taking in four-year-olds

Xuan Zhao (xuan_zhao@brown.edu)[1] Bertram F. Malle (bfmalle@brown.edu)[1]
Hyowon Gweon (gweon@stanford.edu)[2]

[1]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912
[2]Department of Psychology, Stanford University, Stanford, CA 94305

Abstract

To successfully navigate the complex social world, people often need to solve the problem of perspective selection: Between two conflicting viewpoints of the self and the other, whose perspective should one take? In two experiments, we show that four-year-olds use others' knowledge and goals to decide *when* to engage in visual perspective taking. Children were more likely to take a social partner's perspective to describe an ambiguous symbol when she did not know numbers and wanted to learn than when she knew numbers and wanted to teach. These results were shown in children's own responses (Experiment 1) and in their evaluations of others' responses (Experiment 2). By preschool years, children understand when perspective taking is appropriate and necessary and selectively take others' perspectives in social interactions. These results provide novel insights into the nature and the development of perspective taking.

Keywords: social cognition; cognitive development; perspective taking; theory of mind; pedagogy

Introduction

We all see the world from unique perspectives, forming different perceptions and beliefs and pursuing individual goals. Thus, to effectively interact with others, we often have to juggle various viewpoints and even choose between conflicting perspectives: When my dinner guest asks, "could you pass the plate on the left," should I hand over the one on his left or mine? When a friend in a different time zone says, "I'll call you at 6pm", does she mean 6pm for her, or for me? From ambiguity resolution in conversations to large-scale collaborative efforts, *perspective selection*—deciding whose perspective to take—is a central and pervasive problem in navigating the social world.

Previous research has revealed various benefits of taking another's perspective: it allows people to better understand others' mental states, improves interpersonal and intergroup relations, and even promotes generosity and helping behaviors (see Hodges, Clark, & Myers, 2011 for a review). However, perspective taking is not always necessary; our own perspectives are usually sufficient and even more relevant for many daily tasks such as spatial navigation or goal-directed actions. Moreover, deliberately taking on others' perspectives can be cognitively demanding (Hodges & Klein, 2001; Epley, Morewedge, & Keysar, 2004; Surtees & Apperly, 2012) and may even lead to poorer coordination (Galinsky, Ku, & Wang, 2005; Ku, Wang, & Galinsky, 2015). Because indiscriminately taking others' perspectives can be costly and even counterproductive, we must know *when* perspective taking is called for and employ it selectively in proper contexts.

For instance, imagine you see a number "9" on the table, which looks like a "6" to Emma who is sitting across the table. When Emma asks, "What is the number on the table?" would you say that it is a "nine," or a "six"? There is no inherently correct or incorrect answer to this question; the meaning of the symbol is *perspective-dependent*. Depending on the context, however, it may be more appropriate to answer the question from one perspective than the other. For instance, suppose Emma doesn't know numbers at all and wants to learn from you; then reading it as "nine" from your perspective might mislead her to falsely believe that what appears as "6" to her should be read as "nine." However, if Emma already knows numbers and is simply testing your number knowledge, then taking her perspective to say "six" might be unnecessary; it may even be potentially counter-productive if Emma (the teacher) expects you (the learner) to read the number from your own perspective.

This example illustrates a case in which the appropriate perspective is determined by the communicative context. Importantly, others' goals and knowledge can provide critical information about the nature of context. In this study, we investigate young children's ability to use such information to decide when to take others' perspectives. Decades of developmental research have revealed young children's ability to reason about others' mental states, such as their goals, desires, and beliefs—an ability often labeled "theory of mind" (Wellman, Cross, & Watson, 2001). Even toddlers use others' mental states to flexibly modulate their behaviors in social interactions, such as providing the food that others like rather than what they themselves like (Repacholi & Gopnik, 1997; Doan, Denison, Lucas, & Gopnik, 2015), or offering help or information for those in need (Warneken & Tomasello, 2006; Cortes Barragan & Dweck, 2014; Liszkowski, Carpenter, & Tomasello, 2008). Furthermore, children around age 5 readily consider others' prior goals and knowledge to decide what, and how much, information is appropriate in a given communicate context (Gweon, Shafto, & Schulz, 2014; Gweon, Chu, & Schulz, 2014). These studies suggest children's ability to consider others' mental states to modify their social behaviors.

However, little research has examined whether children can also use others' mental states to solve the perspective-selection problem. Developmental research on perspective taking has traditionally focused on whether children at certain ages are capable of reasoning about others' visual experiences (e.g., Piaget & Inhelder, 1956; Clark, 1997; Moll & Tomasello, 2006). By using tasks that explicitly instructed or clearly suggested children to report what others see, researchers have found that although the sensitivity to

others' visual perspectives is present even in infancy (e.g., Luo & Johnson, 2009), the ability to report what others see gradually develops throughout childhood. Importantly, children show varying degrees of success depending on the complexity of the task. For example, around age 2, children can reliably report what object someone else can see (i.e., "Level-1 perspective taking", e.g., Moll et al., 2006), but they cannot decide how another person may see the same object differently until age 3 or even later (i.e., "Level-2 perspective taking", e.g., Moll & Meltzoff, 2011; Flavell et al., 1981). Particularly challenging are tasks that require suppressing one's own visual percepts to deliberately respond from another's conflicting perspective (Moll et al., 2013); in such tasks, even adults are slow and error-prone (e.g., Epley, et al., 2004), as such tasks require a careful deployment of executive resources (Qureshi, Apperly, & Samson, 2010).

The competition between two representations is particularly prominent in the above "9"/"6" example, which provides an ideal setup to test the perspective-selection problem: The perspective-dependent nature of the symbol provides the critical ambiguity, and the relevance of each perspective can only be decided by taking the contexts, such as the communicative partner's mental states, into account. We thus exploit this example[1] to test whether children use others' goals and epistemic states to decide when to take others' perspectives. Importantly, unlike prior studies that explicitly specified the perspective people should take when judging similar stimuli (e.g., Surtees, Samson, & Apperly, 2016), we embedded our ambiguous task in a pedagogical interaction to examine the effect of communicative contexts.

Motivated by recent studies showing children's ability to consider others' mental states to select what information to communicate (e.g., Gweon et al., 2014a; 2014b), we hypothesized that children would consider others' goals and epistemic states to decide whose perspective to take in a social interaction. We targeted four-year-olds because prior research on this age group has found reliable above-chance performance in Level-2 perspective taking tasks that require suppressing one's own representation to report what another person sees (Masangkay et al., 1974; Moll et al, 2013).

We predicted that even four-year-olds would show prosocial and selective perspective taking, spontaneously adopting another person's point of view to describe the ambiguous symbol when it would yield benefits to the other person. In Experiment 1, we looked at children's responses (i.e., reading the number from self-perspective or other-perspective) in a dyadic interaction. In Experiment 2, we examined children's evaluations of others who responded used different perspectives to answer the same question.

[1] We have also considered other stimuli with similar perspective-dependent property (e.g., up/down arrows, "u"/"n", convex/concave curves as mouths on a happy/sad face). The numbers were the most accessible symbols for our target age group (4-year-olds).

Experiment 1

In Experiment 1, we asked children to directly respond to the puppet Emma's question, "What is the number on the table?" Across two conditions, we manipulated Emma's apparent goals and knowledge. We predicted that children would be more likely to report "six" from Emma's perspective when she didn't know numbers and wanted to learn from them (the "Puppet-Learner" condition) compared to when she already knew numbers and was testing the child's knowledge (the "Puppet-Teacher" condition).

Method

Participants Forty 4-year-old children were randomly assigned to the Puppet-Learner (9 boys, 11 girls, *M (SD)* = 4.52 (.35) yrs) or the Puppet-Teacher condition (9 boys, 11 girls, *M (SD)* = 4.48 (.21) yrs). Children were recruited from a university preschool (*N*=22) or a local children's museum (*N*=18). An additional 11 children were excluded for: Not knowing numbers "6" and "9" and failing to learn during the number-pretest session (*N*=4; see Procedure), reading aloud the number before hearing the critical test question (*N*=4), answering from both perspectives (*N*=2), or failing to follow task instructions (*N*=1).

Materials A plastic cutout of number "6" in Arial Narrow font (3 × 5.5 inches) was used for the number pretest (see Procedure). For the practice trial and the test trial, we used two laminated photographs (8.5 × 11 inches): The Practice Photo showed a puppet (Emma) looking at an orange on a table; the Number Photo showed her looking at a number "6" laid flat on the table, which would be a "9" to the participant (Fig. 1). The puppet was gazing at the object in both photos, a gesture shown to trigger a moderate level of visual perspective taking (Zhao, Cusimano & Malle, 2015, 2016).

Procedure The experiment took place in a quiet, separate room. All children sat at a table across from the experimenter. If children's parents or siblings were in the testing room, they were asked to sit outside the children's visual field and not intervene during the procedure.

Number pretest. The experimenter held the number "6" upright and asked the child what number it was. After the child correctly answered "six", she rotated the number to show "9" and asked what number it was. If the child gave an incorrect answer to either question, the experimenter provided the correct answer and repeated the question. Once the child correctly identified both numbers, the experimenter said, "the same thing can be seen as both a 'six' and a 'nine'! How cool is that!"

Practice trial. The experimenter then placed the Practice Photo on the table and introduced the puppet in the photo as "Emma". Then she impersonated Emma using a voice distinct from her own (while avoiding eye contact with the child): "Look at that! It is my favorite fruit! I have a question for you: what is the fruit on the table?" All children answered this question correctly.

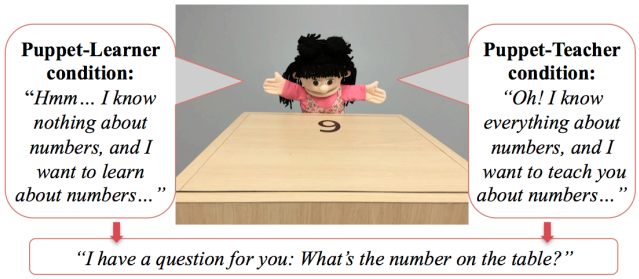


Figure 1. The Number Photo of the test trial in Experiment 1 (in the middle), and Emma's lines in the "Puppet-Learner" and the "Puppet-Teacher" conditions.

Test trial. The experimenter placed the Number Photo on the table and continued to impersonate Emma. In the Puppet-Learner condition, Emma said: *"I know nothing about numbers, and I want to learn about numbers."* In the Puppet-Teacher condition, she said: *"I know everything about numbers, and I want to teach you about numbers."* Then she asked an identical question across conditions: *"I have a question for you. What's the number on the table?"*

After the test trial, all but three children received the questions "What does this number look like from your side?" and "What does this number look like from Emma's side?" The latter question resembles the kind of questions used in previous research to gauge participants' perspective-taking ability upon explicit request (e.g. Flavell et al., 1981).

Results and Discussions

Consistent with our prediction, children were significantly more likely to take Emma's visual perspective and report "six" in the Puppet-Learner condition (14 of 20, 70%, 95% CI: 47.37-89.65%) than in the Puppet-Teacher condition (5 of 20, 25%, 95% CI: 7.41-45.45%; p = .006, Fisher's exact. See Fig. 2). When explicitly requested to take Emma's perspective, the proportions of children who accurately reported "six" were nevertheless comparable (67% and 63% in the Puppet-Learner and the Puppet-Teacher conditions, respectively). Similarly, accuracy rates were also indistinguishable when they were explicitly instructed to respond from their own perspective (61% and 68%, p = .74).

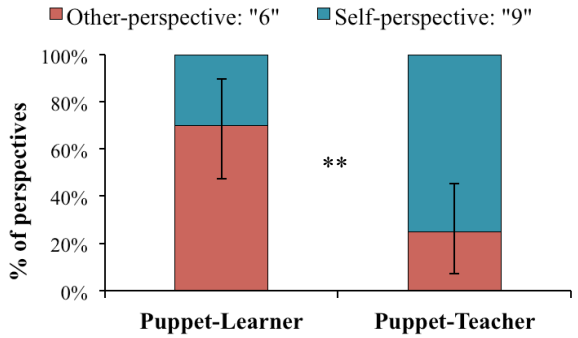


Figure 2. Results from Experiment 1. Vertical black lines show 95% confidence intervals bootstrapped from the data.

These results suggest that children solved the perspective-selection problem by taking into account Emma's goal to teach or to learn and her number knowledge. Specifically, as teachers in the Puppet-Learner condition, children were more likely to describe the number from her perspective. Given that many four-year-olds still struggled with taking specific perspectives even upon explicit instructions, the proportion of children who spontaneously took Emma's perspective in the Puppet-Learner condition is impressive.

In Experiment 2, we sought another way to test our hypothesis. Prior work suggests that children evaluate informants based on the quality of information they provided, preferring those who provide accurate, complete information over those who provide inaccurate, incomplete information (e.g., Koenig & Harris, 2005, Gweon et al., 2014b). Here we asked whether children also evaluate others' perspective-taking decisions using Emma's goals and knowledge. As in Experiment 1, we hypothesized that children would prefer the informant who took Emma's perspective only when she did not know numbers and wanted to learn about numbers.

Experiment 2

Method

Participants Forty 4-year-old children were randomly assigned to the Puppet-Learner condition (6 boys, 14 girls, *M (SD)* = 4.53 (.22) yrs) or the Puppet-Teacher condition (8 boys, 12 girls, *M (SD)* = 4.53 (.30) yrs). They were recruited either at a university preschool (N=8) or at a local children's museum (N=32). An additional 14 children participated but were excluded due to inconsistent responses (N=6; see Procedure), not knowing numbers "6" and "9" and failing to learn them during the number-pretest session (N=3), not completing the procedure (N=3), endorsing both informants as correct (N=1), and sibling interference (N=1).

Materials A cutout of number "6" ("9" if rotated) was used. In addition to the "Emma" puppet, "Bert & Ernie" puppets appeared in the practice trial, and two nearly identical puppets (only differing in their hair colors) were used in the test trial. As in Experiment 1, Emma was standing behind the table and facing the participant; the informants appeared on the other side of the table, facing away from (and thus sharing the same viewpoint with) the participant. We recorded audio tracks of puppets' voices separately and overlaid them on the video images during postproduction. All videos lasted 30 seconds and were played on a Macbook Pro in the experiment. To help children focus on the screen, we placed the laptop inside a "viewing box"—a tightly fitting box around the laptop (approximately 13 ×10 × 10 inches) with a front opening for watching videos.

Procedure The *number pretest trial* was identical to that in Experiment 1 except that the child and the experimenter sat on the same side of the table.

Practice trial. The experimenter invited the child to watch "a show about Emma and her friends". Similar to the Practice Photo in Experiment 1, the Practice Video showed Emma looking at the orange and asking what fruit was on the table. Bert appeared from the lower-left corner of the screen and said, "I know what it is. It's a banana!" and disappeared. Then Ernie appeared from the lower-right corner and said, "I know what it is. It's an orange!" and disappeared. Both puppets then reappeared simultaneously, and the experimenter asked, "Now, who did a better job answering Emma's question? The guy on this side (while tapping on the left side of the box), or the guy on this side (while tapping on the right side of the box)?" If children pointed to the screen to indicate their choice instead of tapping, the experimenter repeated the tapping instruction to avoid any ambiguity in coding their choices. All children answered this question correctly.

Test trial. Similar to the Number Photo in Experiment 1, the Number Video showed Emma looking at the number "9/6" on the table. In both the Puppet-Learner and the Puppet-Teacher conditions, Emma's lines were identical to Experiment 1 (shown in Figure 1), ending with the critical question: "I have a question for you: What number is on the table?" As in the Practice Video, two puppets appeared and disappeared in sequence; one puppet said, "I know what it is. It's a six!", while the other puppet said, "I know what it is. It's a nine!" (order of "six" and "nine" counterbalanced) Both puppets appeared again at the end of the video, and the experimenter asked, "Who did a better job answering Emma's question?" As in the practice trial, children were asked to tap on the side of the box to indicate their choice. If a child could not answer, the experimenter repeated the video and her question. After the children made their choice by tapping, the experimenter tapped on the same side and asked, "What did this guy say?" This question served as a memory check, and those (N=6) who provided inconsistent responses (e.g., tapped to endorse the guy who answered "six" but believed he said "nine") were dropped from analysis.

Similar to Experiment 1, the experimenter also asked "What does this number look like from your side?" and "What does this number look like from Emma's side?" All but four children received these questions.

Results and Discussion

We hypothesized that children in the Puppet-Learner condition would be more likely to prefer the informant who took Emma's perspective to respond (i.e., "it's a six!") than children in the Puppet-Teacher condition. As predicted, their preferences for informants differed significantly across conditions—while 60% of children in the Puppet-Learner condition believed the perspective-taking informant did a better job (12 of 20, 95% CI: 36.86-80.95%), only 15% in the Puppet-Teacher condition thought so (3 of 20, 95% CI: 0-33.33%), p = .004, Fisher's exact test (see Fig. 3).

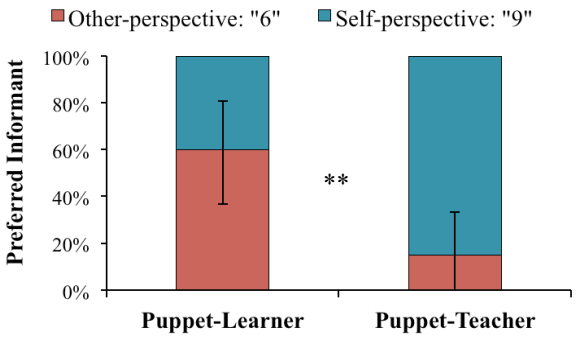


Figure 3. Results from Experiment 2. Vertical black lines show 95% confidence intervals bootstrapped from the data.

In addition, when explicitly asked to report the number from Emma's perspective, children had similar accuracy rates in both conditions (62% and 70%, p = .33). Their accuracy rates were also almost identical when instructed to respond from their own perspective (69% and 75%, p = .48).

Experiment 2 showed that 4-year-olds were able to evaluate the appropriateness of other agents' perspective-taking responses based on the social context; in particular, the requester's mental states were critical for determining which perspective was more relevant and appropriate. Rather than always preferring the informant who engaged in perspective taking, children preferred the informant who answered from Emma's perspective when, and only when, Emma indicated her lack of number knowledge and her intention to learn. Such results are consistent with children's own perspective-selection decisions in Experiment 1.

General Discussion

People perceive the world through their own eyes. However, in everyday interactions with others, they also encounter perspectives that differ from their own. Therefore, people often need to decide which perspective is more relevant in a given social interaction; even though taking the other' perspective is effortful, such effort may be worthwhile when the situation calls for it. Across two studies, we demonstrated that 4-year-olds took into account another's mental states to flexibly and selectively engage in visual perspective taking. In Experiment 1, they responded to their social partner's question by reading an ambiguous symbol either from their own viewpoint or from the partner's viewpoint depending on the partner's goals and epistemic states. In Experiment 2, they showed a similar understanding when evaluating other agents' responses.

This research generates novel insights into the nature and the development of perspective taking and suggests possible directions for future work. To start with, these results reveal previously undocumented perspective-selection ability in preschoolers: While prior research focused primarily on whether, or at what age, children demonstrated perspective-taking competencies upon explicit request, we provide evidence that four-year-olds showed considerable levels of spontaneous perspective taking even without explicit

instructions. Furthermore, their perspective taking was *selective*: Instead of indiscriminately taking others' viewpoint, they did so only when the other "wanted to learn" (i.e., showed a need or interest). Because the proportion of children who took Emma's perspective in the Puppet-Learner condition might have been bounded by the proportion of children who would be able to report the number from Emma's view upon explicit request, it would be worthwhile to compare our results with explicit perspective-taking performance in an independent group of 4-year-olds ("What number does Emma see?"), as well as with results from adults who would have little trouble with perspective taking. Both studies are currently underway.

Our findings pose further questions on *how* children make such astute decisions about when (and when not) to take others' perspectives. Given accumulating evidence on preschooler's impressive inferential capacities in prosocial, pedagogical contexts (e.g., Gweon et al., 2014a), we speculate that children reason about the consequences of their options for both themselves and others to decide whose perspective to take. For example, if Emma is ignorant about numbers, taking her perspective would yield the benefit of correctly informing her, whereas using the egocentric perspective is likely to mislead her. In contrast, when Emma is already knowledgeable about numbers, there is no apparent benefit from taking her perspective; in fact, it is pragmatically more reasonable to answer from their own perspective. Indeed, a handful of children explained their decisions along this line (although we did not solicit verbal justification). However, it is also possible that the majority solved the problem of perspective selection using a simpler heuristic or norm acquired from their own experiences, such as "teachers usually take students' perspectives."

Our research also expands the phenomenon of visual perspective taking by highlighting its social nature. Because researchers have traditionally considered visual perspective taking as a perceptual and spatial ability (e.g. Piaget et al., 1956), relatively few theories or experimental investigations have questioned *why* humans engage in perspective taking at all. Recent work has started to examine how social cues from another person, such as goal-directed behaviors, may influence people's perspective-taking tendency (Tversky & Hard, 2009; Zhao et al., 2015). We extend this line of research in the context of dyadic interactions and demonstrate that even young children consider social information when engaging in visual perspective taking and use it to serve prosocial functions.

Understanding others' perceptions of the physical world is inherently tied to understanding their beliefs, desires, and intentions (Flavell, 2004). Thus, "visual" perspective taking may not depend on a fundamentally different capacity from "social" perspective taking in a broader, abstract sense (c.f. Moll & Kadipasaoglu, 2013). Although our research utilizes a visual perspective-taking task in the sense that children reason about what another sees, success in this task also requires thinking and reasoning from another's perspective. For future work, we hope to further study whether people might show selective and prosocial perspective taking regardless of the apparent visual, mental, or social nature of the task. Finally, note that we used both the social partner's goals (e.g., wanting to learn about numbers) and her epistemic states (e.g., knowing nothing about numbers) to reveal her mental states. In real-world pedagogical contexts, the presence of one mental state (either goal or knowledge) usually implies the presence of the other; nonetheless, teasing apart the roles of each of these mental states (e.g., by embedding them in other social contexts) would be an interesting direction for future research.

Both in everyday situations and in educational practices, perspective taking is generally considered desirable and beneficial for facilitating social interactions. This has led prior researchers to focus primarily on whether, and at what ages, children can accurately take others' perspectives. Our results highlight the importance of investigating when children take others' perspectives, and what information may support such inferences. We hope that our work generates more future research to explore this broad topic. Such research will not only clarify the use of perspective taking in everyday life, but also provide an opportunity to better understand failures of perspective taking in real-world contexts. By identifying situations where people naturally engage in perspective taking, researchers, educators, and practitioners may also find opportunities to create favorable conditions for people to exercise their critical perspective-taking skills.

Acknowledgments

Thanks to Mika Asaba, Sophie Bridgers, Erik Santoro, & Taylor Holubar for help with stimuli; Alyssa Lombardo, Andrew McCabe, Veronica Chu, and Amy Qian for help with data collection. We are grateful to Bing Nursery School, Palo Alto Junior Museum & Zoo, our participants and their families. We also thank Herb Clark and members of Stanford Social Learning Lab for helpful feedback. This work was supported by the Exchange Scholar Program between Brown and Stanford (X.Z.) and Varieties of Understanding Grant from Fordham University and the John Templeton Foundation to H.G.

References

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, *64*, 1–37.

Cortes Barragan, R., & Dweck, C. S. (2014). Rethinking natural altruism: Simple reciprocal interactions trigger children's benevolence: Fig. 1. *Proceedings of the National Academy of Sciences*, *111*(48), 17071–17074.

Doan, T., Denison, S., Lucas, C. G., & Gopnik, A. (2015). Learning to reason about desires: An infant training study. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 578–583.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*(6), 760–768.

Flavell, J. H. (2004). Chapter 1 Development of Knowledge about Vision. In D. T. Levin (Ed.), *Thinking and Seeing: Visual Metacognition in Adults and Children* (pp. 13–36). Cambridge, MA: MIT Press.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual-perception – Further evidence for the level 1–level 2 distinction. *Developmental Psychology*, 17, 99–103.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-Taking and Self-Other Overlap: Fostering Social Bonds and Facilitating Social Coordination. *Group Processes & Intergroup Relations*, *8*(2), 109–124.

Gweon, H., Chu, V., & Schulz, L. E. (2014). To give a fish or to teach how to fish? Children weigh costs and benefits in considering what information to transmit. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 559-564.

Gweon, H., Shafto, P. & Schulz, L. E. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society,* 565-570.

Hodges, S. D., Clark, B, & Myers, M. W. (2011). Better living through perspective taking. In Biswas-Diener, R. (Ed.), *Positive Psychology as a Mechanism for Social Change* (193-218). Dordrecht, The Netherlands: Springer Press.

Hodges, S. D., & Klein, K. J. K. (2001). Regulating the costs of empathy: the price of being human. The Journal of Socio-Economics, 30(5), 437–452.

Koenig, M., & Harris, P. L. (2005). Preschoolers Mistrust Ignorant and Inaccurate Speakers, *76*(6), 1261–1277.

Ku, G., Wang, C. S., & Galinsky, A. D. (2015). The promise and perversity of perspective-taking in organizations. *Research in Organizational Behavior*.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, *108*(3), 732–739.

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, 12(1), 142–149.

Masangkay, Z. S., McCluskey, K. a, McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, *45*(2), 357–366.

Moll, H., & Kadipasaoglu, D. (2013). The primacy of social over visual perspective-taking. *Frontiers in Human Neuroscience*, *7*(September), 558.

Moll, H., & Meltzoff, A. N. (2011). How Does It Look ? Level 2 Perspective-Taking at 36 Months of Age, *82*(2), 661–673.

Moll, H., Meltzoff, A. N., Merzsch, K., & Tomasello, M. (2013). Taking versus confronting visual perspectives in preschool children. *Developmental Psychology*, *49*(4), 646–654.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. British Journal of *Developmental Psychology,* 24(3), 603–613.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space.* London, England: Routledge & Kegan Paul.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236.

Repacholi, B. M., & Gopnik, a. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Sabbagh, M. A., & Baldwin, D. A. (2001). Learning words from knowledgeable versus ignorant speakers: links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054–1070.

Slaughter, V. & Repacholi, B. (2003). Introduction: Individual differences in theory of mind: What are we investigating?. In B. Repacholi and V. Slaughter (Ed.), *Individual Differences in Theory of Mind: Implications for Typical and Atypical Development* (pp. 1-12). New York: Psychology Press.

Surtees, A. D. R., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. Child Development, 83(2), 452–60.

Surtees, A., Samson, D., & Apperly, I. (2016). Unintentional perspective-taking calculates whether something is seen, but not how it is seen. *Cognition*, *148*, 97–105.

Tomasello, M., Dweck, C., Silk, J., Skyrms, B., & Spelke, E. (2009). *Why we cooperate*. Cambridge, MA: MIT Press.

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: spatial perspective-taking. *Cognition*, *110*(1), 124–129.

Warneken, F., & Tomasello, M. (2006). Altruistic Helping in Human Infants and Young Chimpanzees. *Science*, *311*(5765), 1301–1303.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655–84.

Zhao, X., Cusimano, C., & Malle, B.F. (2015). In search of triggering conditions for spontaneous visual perspective taking. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 2811-2816.

Zhao, X., Cusimano, C., and Malle, B. F. (2016). Do people spontaneously take a robot's visual perspective? In *HRI '16: Proceedings of the Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction, Christchurch, New Zealand* (pp. 335-342). Piscataway, NJ: IEEE Press.

Can Monaural Auditory Displays Convey Directional Information to Users?

Takanori Komatsu (tkomat@meiji.ac.jp)
Department of Frontier Media Science (FMS), Meiji University
4-21-4 Nakano, Tokyo 1648525 JAPAN

Seiji Yamada (seiji@nii.ac.jp)
National Institute of Informatics/SOKENDAI/Tokyo Institute of Technology
2-1-2 Hitotsubashi, Tokyo 1018340 JAPAN

Abstract

The purpose of this study is to build a monaural auditory display to convey four pieces of directional information (upward, downward, rightward, and leftward) to users effectively and intuitively without the need for wearing headphones or preparing more than one speaker. We prepared five types of monaural auditory displays consisting of triangle wave sounds and conducted an experiment to investigate which kinds of displays succeeded in conveying the four pieces of information to participants. As a result, we could confirm that one of the prepared monaural auditory displays, designed as a "progress bar" on the basis of the mental-number line and spatial-number association of the response code effect, succeeded in conveying the four pieces of information more effectively compared with the other candidate sets (its average correct rates were about 0.88). This result thus strongly shows that this monaural auditory display was quite useful for conveying primitive spatial information to users

Keywords: Monaural auditory display; Directional information; Mental-number line; Spatial-number association of response code effect.

Introduction

Most user interfaces convey various types of information to users as visual information appearing on visual displays, such as LCDs or LEDs, or as sound information emitted from speakers as speech or non-speech sounds. Currently, visual displays can be used to convey rich information to users, while sound information still has an indispensable role, especially for some users, such as visually impaired persons, those whose eyes are focused on a task at hand, or workers who must move around in their environment. Furthermore, the omnidirectional nature of auditory perception allows listeners to obtain information from sound information, regardless of their location or orientation (Walker & Kramer, 2006a).

In terms of speech and non-speech sounds, it is said that speech sounds are good for conveying precise statements to listeners, but there are also some flaws; that is, speech sounds are limited only to those who can understand the utilized language; paralinguistic information in speech sounds, e.g., gender, inflections, or tones of voices, can cause unexpected interpretations (Nass & Brave, 2005), and understanding such statements places a cognitive load on users, interfering with tasks they are engaged in (Rouben & Terveen, 2007). In comparison, non-speech sounds could contain an immediate and intrinsic relationship between the display dimensions and the information that is conveyed, e.g., in a Geiger counter, wherein radiation measurements are presented to users in the form of a clicking sound, with more clicks representing higher radiation levels (Walker & Kramer, 2006b), so understanding these sounds is quite easy and intuitive for users. Therefore, these sounds have been studied as "auditory displays" (Gregory, 1994; Walker & Kramer, 2006a) and actually used in various kinds of user interfaces for many years (Patterson, 1982; Pollack & Ficks, 1954), especially for conveying warnings or alarms (Edworthy & Stanton, 1999; Edworthy *et al.*, 1991) and for indicating events or objects on graphical user interfaces, such as by using an earcon (Blattner *et al.*, 1989; McGookin & Brewster, 2004), e.g., notifying that a USB device is connected or removed with a brief piece of music consisting of three or four notes, or an auditory icon (Gaver, 1986; Gaver, 1993), e.g., notifying that a file has been closed with the sound of a closing metal cabinet. The sounds have also been used for presenting spatial information to users (Loomis *et al.*, 1998; Walker & Lindsay, 2006).

Auditory displays used as warnings/alarms and earcons/auditory icons are simply emitted from a monaural speaker, while ones used for representing spatial information are emitted from headphones or more than one speaker. These studies use beacon-like sounds, e.g., a beacon sound played in a cyclic on-off pattern that increases in tempo when a user move closer to a target, or stereo sounds to represent left or right directions for differentiating the times sound reaches the two ears (interaural time

difference) or the sound pressure level reaches the two ears (interaural level difference). Therefore, these users must pay attention to beacon-like sounds continuously or wear headphones, or keep the same distance and orientation from the speakers, so this would hinder users engaged in tasks or in information gathering to comprehend their surrounding environment. Furthermore, most studies on auditory displays require a training phase in which users get used to the prepared displays. We regard this training phase as a severe constraint for users.

The purpose of this study is then to build monaural auditory displays to convey directional information to users effectively and intuitively without such physical constraints and a training phase. As specific spatial information, we focus on four primitive directions: upward, downward, rightward, and leftward. These four pieces of directional information could consist of a 2D vertical plane in front of a user, so we regard these pieces of information as fundamental information for spatial recognition. Specifically, we prepared five types of monaural auditory displays consisting of triangle wave sounds and conducted an experiment to investigate which kinds of displays succeeded in conveying the four pieces of directional information to the participants.

We expect that successful auditory displays can convey directional information to users without the several constraints mentioned above, and various auditory displays (including earcon, alarms, and spatial information) can be emitted from monaural speakers. This would make users receive the fruitful benefits of auditory displays regardless of their locations or orientations.

Prepared Auditory Displays

In terms of designing the auditory displays, several studies argued that "data-to-sound mappings that seem intuitive to a sound designer may actually result in less effective performance (Walker & Kramer, 2005)," or "an effective and practical approach to sonification can only result from the determination of actual listener preference and performance (Walker & Lane, 2001)." Therefore, we decided not to propose the best candidate display set on the basis of deep deliberation but to prepare the several candidate sets on the basis of a broad range of possibilities and explore the best one of them.

First, we prepared two auditory displays that indicate upward and downward as continuously increasing and decreasing pitched sounds, respectively [Figure 1] because several studies (Pratt, 1930; Roffler & Butler, 1968) showed strong evidence that "high pitch is consistently mapped to high positions" and this mapping has been used in various applications successfully (Meijer, 1992). Concretely, a triangle wave sound 0.35 s in duration with increasing pitch [onset fundamental frequency (F0): 250 Hz and end F0: 400 Hz] indicates upward, and a triangle wave sound 0.35 s in duration with decreasing pitch indicates downward [onset F0: 250 Hz and end F0: 100 Hz]. Therefore, these upward and downward displays were commonly used among the following five candidate sets for rightward and leftward[1].

- Set 1 [Figure 2]: Pitched sounds increasing and decreasing in a stepwise manner indicate rightward and leftward, respectively. This was designed on the basis of the spatial-musical association of the response code (SMARC) effect (Rusconi *et al.*, 2006); that is, low-pitched tones are associated with the left side of a keyboard like a piano, while high-pitched tones are associated with the right side. Concretely, a triangle wave sound 0.7 s in duration decreasing 25 Hz from an onset F0 of 400 Hz every 0.1 s in a stepwise manner (end F0: 250 Hz) indicates leftward, and a triangle wave sound 0.7 s in duration increasing 25 Hz from an onset F0 of 250 Hz every 0.1 s in a stepwise manner (end F0: 400 Hz) indicates rightward.

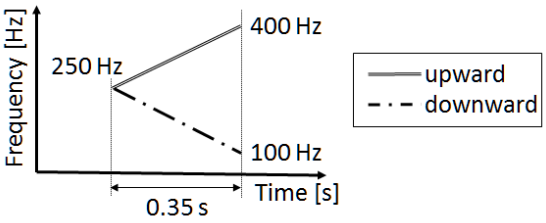


Figure 1: Commonly used upward and downward

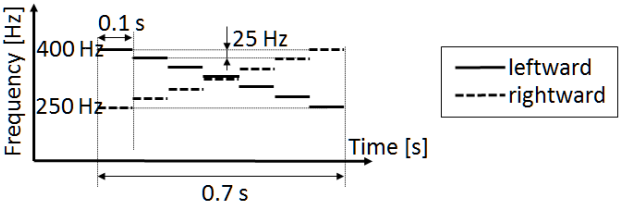


Figure 2: Leftward/rightward of Set 1

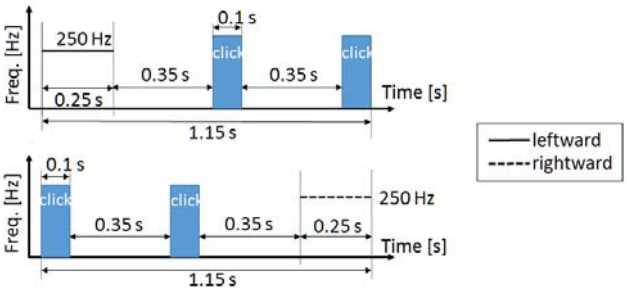


Figure 3: Leftward/rightward of Set 2

[1] URL will be listed here for listening to the prepared auditory displays if this paper is accepted.

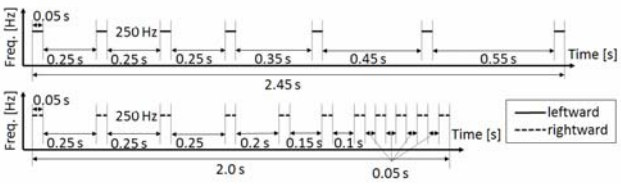


Figure 4: Leftward/rightward of Set 3

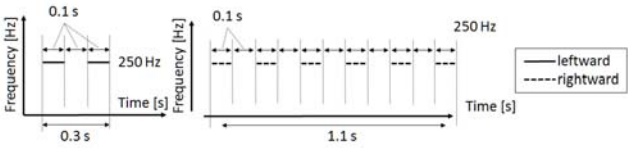


Figure 5: Leftward/rightward of Set 4

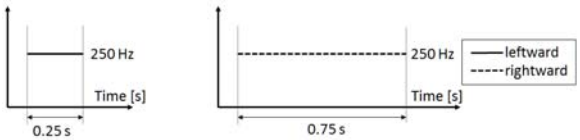


Figure 6: Leftward/rightward of Set 5

- Set 2 [Figure 3]: "Sound-click-click" indicates leftward, and "click-click-sound" indicates rightward. This was designed on the basis of the concept of eye-music (Abboud *et al.*, 2014); that is, the sounds appear to scan a picture from left to right, i.e., elements to the left will be heard first. Concretely, "sound" was a triangle wave sound 0.25 s in duration with a fixed F0 of 250 Hz, and "click" was a recorded sound of a metronome, and its duration was 0.1 s. The interval among these sounds was set at 0.35 s.
- Set 3 [Figure 4]: A sound played 7 times with intervals between sounds becoming gradually wider (after first four sounds; from 0.25 to 0.55 s) indicates leftward, while a sound played 11 times with intervals becoming gradually narrower (from fourth sounds to sixth sounds; from 0.25 to 0.05 s) indicates rightward. This was also designed on the basis of the concept of the above mentioned eye-music, but changing the sound intervals represents deceleration and acceleration, respectively. The sound was a triangle wave sound 0.05 s in duration with a fixed F0 of 250 Hz.
- Set 4 [Figure 5]: A sound played twice indicates leftward, and a sound played six times indicates rightward. This was designed on the basis of the concept of the mental-number line (MNL) (Dehaene *et al.*, 1993) and the spatial-number association of the response code (SNARC) effect (Fias *et al.*, 1996); that is, smaller numbers are associated with the left side, while larger numbers are associated with the right side. The sound was a triangle wave sound 0.1 s in duration with a fixed F0 of 250 Hz, and its interval was also 0.1 s.
- Set 5 [Figure 6]: A shorter sound (duration: 0.25 s) indicates leftward, and a longer one (duration: 0.75 s) indicates rightward. This was also designed on the basis of the concept of the MNL and the SNARC effect, but these concepts were represented as a "progress bar." The F0 of these sounds was fixed at 250 Hz.

Actually, the authors arbitrarily designed the parameters of the sounds, such as the durations (e.g., 0.25 s in Set 2, 0.05 s in Set 3, or 0.1 s in Set 4), numbers (e.g., 7 times for leftward and 11 times for rightward in Set 3 or twice for leftward and 6 times for rightward), or fundamental frequency (most of onset F0 was 250 Hz). The purpose of this experiment was not to find the best parameters of these sounds but to find the best pattern of these sounds as an auditory display.

Experiment

Participants

20 Japanese undergrads (14 men and 6 women; average age was 19.9 years old; all right-handed) participated in this experiment. The mother tongue of all of the participants was Japanese, in which reading and writing is done from left to right, so the SNARC effect was the same as that in Western countries.

Procedure

In an experiment room, each participant was asked to sit at a table in which a USB speaker (JVC Kenwood Cooperation, NC-SP1) was placed in front of her/him. This speaker was connected to an experimenter's laptop PC, and the experimenter, who was placed on the other side of a room partition, played the selected auditory display manually. The sound pressure at the participants' head level was set at about 50 dB (FAST, A). Participants were asked to answer with one of the four pieces of directional information (upward, downward, rightward, or leftward) after hearing one of the prepared auditory displays described in the former section.

The participants experienced all five candidate sets (Set 1 to Set 5) in random order, and in each set, the four pieces of information were presented four times, also in random order. This means that each participant answered 80 times (4 pieces of directional information × 4 times × 5 candidate sets), and this took about 10 minutes. Note that this experiment did not have any training phases in which the participants got used to the prepared auditory displays.

Result

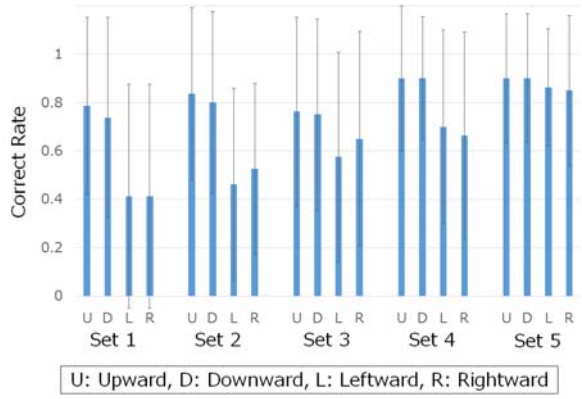


Figure 7: Correct rates in terms of directional information and candidate (error bars represent standard deviations)

Table 1: Average correct rates for each set and each piece of directional information.

	up	down	left	right
Set 1	0.7875 (0.3646)	0.7375 (0.4144)	0.4125 (0.4628)	0.4124 (0.4628)
Set 2	0.8375 (0.356)	0.8 (0.3758)	0.4625 (0.3975)	0.525 (0.3527)
Set 3	0.7625 (0.3911)	0.75 (0.3953)	0.575 (0.4337)	0.65 (0.443)
Set 4	0.9 (0.3)	0.9 (0.225)	0.7 (0.4)	0.6625 (0.4277)
Set 5	0.9 (0.2669)	0.9 (0.2669)	0.8625 (0.2433)	0.85 (0.3102)

To investigate which candidate set succeeded in conveying the directional information to the participants correctly, we calculated the correct rates, indicating how much the presented display was interpreted correctly. Table 1 and Figure 7 show the correct rates in terms of the four pieces of information and five candidate auditory display sets.

The correct rates were analyzed by using a two-way within-participants design ANOVA [independent variable 1: directional information (four levels: upward, downward, leftward, and rightward), independent variable 2: candidate set (five levels: Sets 1 to 5), dependent variable: correct rates]. The results of the ANOVA showed significant differences in the interaction effect [$F(12, 228) = 2.72$, $p < .01$, effect size: $\eta^2 = 0.15$] and in both of the main effects [candidate set: $F(4,76) = 4.2$, $p < .01$, effect size: $\eta^2 = 0.29$, directional information: $F(3,57) = 10.1$, $p < .01$, effect size: $\eta^2 = 0.14$]. The simple main effects of both of the independent variables were analyzed, and the results showed significant differences in correct rates for leftward and rightward among the candidate sets [leftward: $F(4,76) = 5.23$, $p < .01$, rightward: $F(4,76) = 3.72$, $p < .01$] and in the rates for Sets 1, 2, and 4 among the four pieces of information [Set 1: $F(3,57) = 10.92$, $p < .01$, Set 2: $F(3,57) = 10.64$, $p < .01$, Set 4: $F(3,57) = 4.17$, $p < .05$], and the results showed a marginal difference in the rates for Set 3 [$F(3,57) = 2.50$, $p < .10$].

A multiple comparison using an LSD test on the simple main effect of independent variable 2 (candidate set) showed that the rates for leftward for Sets 4 and 5 were significantly higher than those for Sets 1 and 2, the rates for Set 5 were significantly higher than those for Set 3 (MSe = 2.0429, $p < .05$), the rates for rightward for Sets 4 and 5 were significantly higher than those for Set 1, and the rates for Set 5 were significantly higher than those for Set 2 (MSe = 2.3150, $p < .05$). A multiple comparison using an LSD test on the simple main effect of independent variable 1 (directional information) showed that the rates for upward and downward were significantly higher than those for leftward and rightward in Set 1 (MSe = 1.2088, $p < .05$), those for Set 2 (MSe = 1.0851, $p < .05$), those for Set 4 (MSe = 1.2406, $p < .05$), and those for leftward in Set 3 (MSe = 1.0037, $p < .05$).

The results of this statistical analysis can be summed up as follows.

- Sets 4 and 5 showed significantly higher correct rates for leftward and rightward compared with Sets 1 and 2.
- Except for Set 5, the correct rates for upward and downward were higher than those for leftward and rightward in the remaining four candidate sets.

Therefore, we could confirm that the monaural auditory display prepared as Set 5 succeeded in conveying the four pieces of information more effectively compared with the other candidate sets. Moreover, the average correct rates among the four pieces of information for Set 5 were about 0.88 without any training, so these rates are undoubtedly high. This result thus strongly shows that this monaural auditory display was quite useful for conveying the primitive spatial information to users.

Discussion and Conclusions

In this study, we investigated which kinds of our prepared monaural auditory displays can convey four pieces of directional information to users. The information for upward and downward was correctly understood, while that for rightward and leftward in Set 5, designed as a progress bar on the basis of the MNL and the SNARC effect, was correctly understood. Conveying such information to users in an auditory manner is traditionally done as beacon-like or stereo sounds, so our approach (using monaural auditory displays to convey directional information) and the acquired results are completely original and novel. The other noteworthy point of these monaural auditory displays is that a training phase is unnecessary for users, while most auditory studies require these phases. This means that our

proposed auditory display is not only effective but also quite intuitive for users to understand directional information.

As already mentioned, the parameters of the sounds utilized in this experiment, such as durations, numbers, or fundamental frequency, were arbitrarily designed by the authors. Now that we confirmed that the monaural auditory display prepared as Set 5 succeeded in conveying the four pieces of information more effectively, we are planning to find the best parameter sets of these auditory sets, e.g., the appropriate duration for indicating the four pieces of information or appropriate F0 values.

Currently, we have two ideas for concrete applications that use our proposed auditory display. One is to use it independently to indicate the relative direction to users such as in a huge warehouse to find a desired product or in a server machine placed in a huge server rack to convey alarms or warnings to users to indicate which server machine has a problem. The other is to use the proposed display together with the speech sounds used in car navigation to facilitate users' understanding of these sounds. The way of combining the speech sounds and auditory display is almost the same as those of artificial subtle expressions (Komatsu *et al.*, 2010a; Komatsu *et al.*, 2010b); for example, 0.2 s after the speech sound "turn left," a sound indicating a left pattern (0.25 s of sound) is played. We expect that this combination can help new learners of certain languages or people who have right-left confusion to understand directional information from navigation systems smoothly, e.g., an American driving in Japan with Japanese navigation systems. Therefore, this paper is regarded as fundamental research or a first step toward such realistic applications.

Acknowledgments

This study was partially supported by JSPS KAKENHI Grant Numbers 26118005 and 16K00379.

References

Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., & Amedi, A. (2014). EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution, *Restorative neurology and neuroscience, 32*, 247–257.

Blattner, M. M., Sumikawa, D, A., & Greenberg, R. M. (1989). Earcons and Icons: Their Structure and Common Design Principles, *Human-Computer Interaction, 4*, 11–44.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude, *Journal of Experimental Psychology, 122*, 371–396.

Edworthy, J. and Stanton, N. A. (Eds.) (1999). *Human Factors in Auditory Warning*, UK: Ashgate Publishing Limited.

Edworthy, J., Loxley, S., & Dennis, I. (1991). Improving auditory warning design: relationship between warning sound parameters and perceived urgency, *Human Factors, 33 (2)*, 205–231.

Fias, W., Brysbaert, M., Geypens, F., & d'Ydewalle, G. (1996). The Importance of Magnitude Information in Numerical Processing: Evidence from the SNARC Effect, *Mathematical Cognition, 2*, 95–110.

Gaver, W. W. (1986). Auditory icons: Using sound in computer interfaces, *Human-Computer Interaction, 2*, 167–177.

Gaver, W. W. (1993). Synthesizing auditory icons, *Proceedings of INTERCHI'93,* 228–235.

Gregory, K. (Eds.) (1994). *Auditory Display: Sonification, Audification, And Auditory Interfaces*, Addison-Wesley.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., & Nakano, M. (2010a). Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts, *Proceedings of the 28th ACM Conference on Human Factors in Computing Systems*, 1941–1944.

Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., & Nakano, M. (2010b). Proposing Artificial Subtle Expressions as an Intuitive Notification Methodology for Artificial Agents' Internal States, *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 447–452.

Loomis, J. M., Golledge, R. G., & Klatzky, R. L. (1998). Navigation System for the Blind: Auditory Display Modes and Guidance, *Presence, 7 (2)*, 193–203.

McGookin, D. K. & Brewster, S. A. (2004). Understanding Concurrent Earcons: Applying Auditory Scene Analysis Principle to Concurrent Earcon Recognition, *ACM Transaction on Applied Perception, 1 (2)*, 130–155.

Meijer, P. B. L. (1992). An Experimental System for Auditory Image Representations, *IEEE Transactions on Biomedical Engineering, 39 (2)*, 112–121.

Nass, C. and Brave, S. (2005). *Wired for Speech*. Cambridge, MA: The MIT Press.

Patterson, R. D. (1982). *Guideline for auditory warning systems on civil aircraft* (Paper No. 82017.) London, UK: Civil Aviation Authority.

Pratt, C. C. (1930). The spatial character of high and low tones, *Journal of Experimental Psychology, 13 (3)*, 278–285.

Pollack, I. & Ficks, L. (1954). Information of elementary multidimensional auditory displays, *Journal of Acoustical Society of America, 26*, 155–158.

Roffler, S. K. & Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane, *Journal of Acoustic Society of America, 43 (6)*, 1255–1259.

Rouben, A. & Terveen, L. (2007). Speech and Non-Speech Audio: Navigational Information and Cognitive Load, *Proceedings of the 13th International Conference on Auditory Display*, 468–475.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect, *Cognition, 99 (2)*, 113–129.

Walker, B. N. & Kramer, G. (2005). Mapping and Metaphors in Auditory Displays: An Experimental Assessment, *ACM Transaction on Applied Perception, 2(4)*, 407–412.

Walker, B. N. & Kramer, G. (2006a). Auditory Displays, Alarms, and Auditory Interfaces. In W. Karwowski (Eds.), *International Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, New York, NY: CRC Press, 1021–1025.

Walker, B. N. & Kramer, G. (2006b). Sonification. In W. Karwowski (Eds.), *International Encyclopedia of Ergonomics and Human Factors (2nd ed.)*, New York, NY: CRC Press, 1254–1256.

Walker, B. N. & Lane, D. M. (2001). Psychophysical Scaling of Sonification Mappings: A comparison of visually impaired and sighted listeners, *Proceedings of the 7th International Conference on Auditory Display*, 90–94.

Walker, B. N. & Lindsay, J. (2006). Navigation Performance with a Virtual Auditory Display: Effects of Beacon Sound, Capture Radius, and Practice, *Human Factors, 48 (2)*, 265–278.

Spatial Attention to Social Cues is not a Monolithic Process

Samuel M. Harding (hardinsm@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Ty W. Boyer (tboyer@georgiasouthern.edu)
Department of Psychology, P.O. Box 8041
Statesboro, GA 30460 USA

Bennett I. Bertenthal (bbertent@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

Social stimuli are a highly salient source of information, and seem to possess unique qualities that set them apart from other well-known categories. One characteristic is their ability to elicit spatial orienting, whereby directional stimuli like eye-gaze and pointing gestures act as exogenous cues that trigger automatic shifts of attention that are difficult to inhibit. This effect has been extended to non-social stimuli, like arrows, leading to some uncertainty regarding whether spatial orienting is specialized for social cues. Using a standard spatial cueing paradigm, we found evidence that both a pointing hand and arrow are effective cues, but that the hand is encoded more quickly, leading to overall faster responses. We then extended the paradigm to include multiple cues in order to evaluate congruent vs. incongruent cues. Our results indicate that faster encoding of the social cue leads to downstream effects on the allocation of attention resulting in faster orienting.

Keywords: social cues; spatial cueing; selective attention; reflexive orienting; exogenous and endogenous attention

Introduction

At a crowded party surrounded by strangers, sharing a meal with friends, or catching the eye of your server on a café patio, we cannot help but notice and react to the actions of those around us. Through the language of glances and gestures, physical proximity, and facial expressions, we exchange invitations to interact. Interpreting and sending the right social cues is a fundamental part of communicating, building, and maintaining relationships. At the same time, we are not entirely beholden to the social world around us; in a crowded coffee shop, we manage to tune others out to finish our paper draft. The competition between noticing and engaging with others while completing our own personal goals is a fundamental question of what influences our attention and how these endogenous and exogenous processes interact.

Social cues, and in particular, gaze cues, are seen as fundamental to communication (Langton, Watt, & Bruce, 2000). For example, direct gaze can be seen as an invitation to interact, while averted gaze is often used to signal interest in other objects and may communicate an invitation for others to jointly attend to the same thing. Adults and infants are highly sensitive to these gaze cues. Early in life, infants preferentially attend to faces and face-like stimuli (Farroni, Csibra, Simion & Johnson, 2002), while in adulthood, people can rapidly locate faces, even when embedded among numerous non-face distractors (Hershler, & Hochstein, 2005). Conversely, irrelevant faces interfere with locating other non-face objects in a search task (Langton, Law, Burton, & Schweinberger, 2008).

Beyond simply attracting attention, eyes, hands and faces can act as cues to distal locations or objects. Pointing, head turns, and gaze shifts have all been shown to shift attention toward the cued direction, aiding subsequent detection and identification of stimuli located there, with only a few hundred milliseconds' exposure to the cue (Driver, Davis, Ricciardelli, Kuhn, & Benson, 2007; Friesen, & Kingstone, 1998; Langton, & Bruce, 1999; Frischen, Bayliss, & Tipper, 2007; Crostella, Carducci, & Aglioti, 2009; Burton, Bindemann, Langton, Schweinberger, & Jenkins, 2009). In addition, this effect is resistant to interference, and these properties have led some to suggest that social cueing is automatic and stimulus driven. Indeed, this proposal is consistent with specialized processing of social cues in the brain, especially in the superior temporal sulcus (Allison, Puce, & McCarthy, 2000). Nevertheless, this view remains contentious since there is evidence from other neural activation studies showing a large degree of overlap between social and non-social stimuli (Tipper, Handy, Giesbrecht, & Kingstone, 2008), as well as behavioral evidence, including but not limited to spatial cueing tasks showing mixed results in distinguishing between social and non-social cues (Kuhn, & Benson, 2007; Eimer, 1997; Friesen, Ristic, & Kingstone, 2004; Tipples, 2002). Accordingly, exogenous cueing of attention may not be unique to social stimuli, but rather shared by the broader class of directionally-oriented stimuli.

If it is true that social stimuli are no different in their ability to orient attention, what other properties could explain the differences observed in other tasks? Birmingham and Kingstone (2009; Birmingham, Bischof, & Kingstone, 2008) suggest that the crucial difference lies in their

propensity for attracting, rather than their ability to orient attention. This argument, however, is incomplete, in that it fails to explain how the selection process distinguishes between social and non-social objects. One possibility is that encoding of the social stimuli occurs more quickly, allowing those items to compete for attentional resources sooner. In this light, the design of the standard spatial cueing paradigm (Posner, 1980) makes observing a difference less likely, as the presentation of a single, isolated cue at a known location obviates the need for selective attention. Furthermore, even if social cues are detected faster or with greater likelihood, the differences may be small and obscured by the additional time taken to orient towards the periphery, detect the target, and plan the appropriate motor response. Differentiating between the effects of these multiple component processes is necessary for identifying if and how social stimuli are different.

We began this investigation by using a standard spatial cueing paradigm to examine differences in the speed of detecting and responding to a peripheral target after the appearance of a non-predictive pointing hand or arrow cue. It has been argued that hands provide a more salient and accurate cue than eye gaze or head direction, but have received significantly less study (Langton et al., 2000; Ricciardelli, Bricolo, Aglioti, & Chelazzi, 2002). In line with previous research, we expected faster encoding and thus shorter reaction times in response to a pointing hand as compared to an arrow. Moreover, both types of stimuli have been shown to be effective cues, so we do not expect any difference in their cueing strength, measured as a validity effect (invalid – valid RTs).

Experiment 1

Methods

Participants were shown a non-predictive hand or arrow cue and tasked with responding to the location of a subsequently appearing peripheral target stimulus. The cue was visible for either 100 or 600ms before the appearance of the peripheral target. Prior research suggests that shorter delays tap into automatic or reflexive processes, while a greater delay permits more volitional or strategic responding (Friesen et al., 1998). Twenty-two undergraduate psychology students (11 female, between 18-24 years of age) participated as volunteers or for course credit. In all experiments, participants reported normal or corrected to normal vision and were naive to the purpose of the research.

Stimuli and Apparatus The stimuli were comprised of digital images of either a pointing hand or a flesh colored arrow created using Adobe AfterEffects CS5 (Adobe Systems, San Jose, CA), and were presented on a 36.6 x 27cm LCD screen operating at 1024 x 768 pixel resolution. Participants were seated approximately 70cm from the screen and responded using the '1' or '3' key on the computer number pad with their right middle- and index-fingers. Stimuli were presented using E-Prime presentation software (Psychology Software Tools, Pittsburgh, PA).

Procedure Each trial began with a central, white fixation cursor, presented for 400, 500, or 600ms. The fixation disappeared and was replaced with the central cue, which remained visible throughout the rest of the trial. After a delay of either 100 or 600ms, the target appeared in the periphery and remained on screen until the participant made a response, or 5 seconds had elapsed. A blank screen was shown between trials for 600, 700, or 800ms (see Figure 1 for stimulus sizes and arrangement). Subjects were instructed to fixate on the center of the screen, and to respond to the location where the target appeared as quickly and accurately as possible. Additionally, they were informed that the cues were non-predictive, such that targets were equally likely to appear at the cued and uncued location

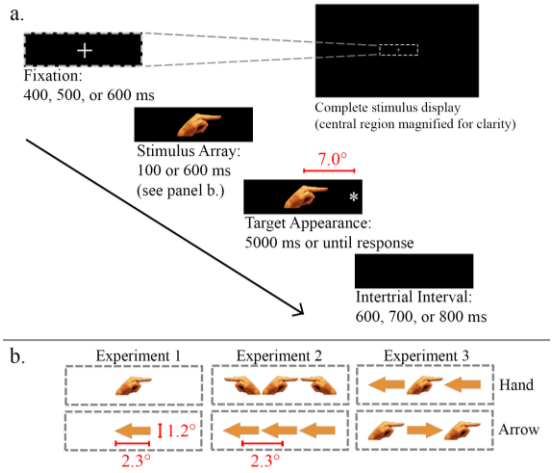


Figure 1. (a**)** Trial Procedure (b) Stimulus arrays used in Experiments 1-3. Sizes and distances visually exaggerated for clarity; true values shown in degrees of visual angle.

Design Three within-subjects factors were manipulated. Stimulus type referred to either a pointing hand or an arrow. The stimulus to target onset asynchrony (SOA) was 100 or 600ms. Validity referred to whether the target would appear at the location cued by the stimulus (valid) or to the opposite side of the screen (invalid).

These factors yielded 8 unique trial types, which were each presented with 20 repeats per block for 3 blocks, yielding 480 total trials per subject. Within each block, the direction of the central cue and the location of the target were counterbalanced for each condition. Before beginning the experiment, participants completed 16 practice trials and were provided feedback on their response time and accuracy. Feedback was not provided during the experimental trials.

The overall proportion correct was 98.4% across all participants. Due to the low amount of errors, we only analyzed correct responses. We also excluded trials with reaction times less than 100ms or greater than 800ms to

exclude anticipations and inattentive responses, which accounted for an additional 1.2% of the total trials. A 2x2x2 repeated-measures ANOVA with stimulus type, SOA, and validity as within-subjects factors was conducted.

Results & Discussion

Consistent with the results from previous studies, mean response times to valid trials (387ms) were faster than invalid trials (401ms), $F(1, 21) = 20.18$, $p < 0.001$, $\eta^2 = 0.490$. This result confirms that even though participants knew that the cues were irrelevant, they could not completely ignore them.

Response times were also affected by the SOAs. The 600ms SOA produced faster overall responses (375ms) than the 100ms delay (413ms), $F(1, 21) = 191.95$, $p < 0.001$, $\eta^2 = 0.901$. This is likely due to participants having additional time to prepare a response (Driver, et al. 1999).

We hypothesized that the social nature of the pointing hand would lead to faster encoding, and thus shorter reaction times than the arrow. Our results were consistent with this prediction in that the mean reaction time was faster to the hand than to the arrow (392 and 397ms respectively), $F(1, 21) = 10.18$, $p < 0.01$, $\eta^2 = 0.327$, though the difference was small. The interaction between stimulus and SOA was not significant, $F(1,21) = 3.70$, $p = 0.068$, suggesting that the additional response preparation benefitted both stimulus types similarly. Plots showing the main effects are shown in Figure 2.

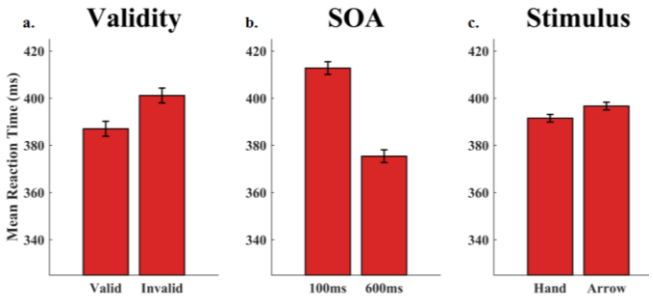


Figure 2. Reaction time means and standard error shown for responses in Experiment 1. Main effects of (a) Validity (b) SOA and (c) Stimulus.

We also tested whether there was a difference between valid and invalid cues. The size of the validity effect was calculated as the mean invalid – valid RT. Response times were analyzed in a 2x2 repeated-measures ANOVA with stimulus and SOA as within-subjects factors. Critically, the stimulus type was not significant, $F(1, 21) = 0.75$, $p = 0.396$, nor was SOA, $F(1, 21) = 0.05$, $p = 0.826$, or the interaction, $F(1, 21) = 0.15$, $p = 0.706$. These results are consistent with our hypothesis that both cues are equally effective at shifting attention toward the target for both automatic (100ms SOA) as well as more intentional (600ms SOA) responses.

The overall pattern of results was consistent with faster encoding of the pointing hand, but the cueing strength for both hand and arrow stimuli (i.e., no stimulus effect for the validity measure) showed no difference. We earlier proposed that this encoding difference is critical in explaining how attention can be selectively oriented towards social stimuli. In order to test this hypothesis, we needed to increase the difficulty of the task to include multiple cues.

Experiment 2

White, Ratcliff, and Starns (2011) studied the effect of surrounding a central cue with flankers on directional judgments, and suggest that multiple cues can simultaneously contribute to a directional decision. In this experiment, we utilized a similar paradigm to explore whether the addition of flanking stimuli facilitate or interfere with processing of the central cue.

Methods

Twenty undergraduate psychology students (14 female, between 18-22 years of age) participated in this experiment as volunteers or for course credit. Participants reported normal or corrected to normal vision and were naive to the purpose of this research.

Stimuli and Apparatus We modified the stimuli from Experiment 1 by pairing the central cue with two additional cues of the same type, i.e. three hands arranged horizontally. The two flankers were always oriented in the same direction as each other, but could point in the opposite direction of the central cue. The total horizontal extent of the stimuli subtended 6.9° visual angle. The distance between the target and the central cue remained the same as Experiment 1.

Design

As in Experiment 1, we manipulated the stimulus type (hand vs. arrow), and the SOA between cue and target (100 vs. 600). With the addition of the two flanking cues, validity was defined as the relationship between the *central* cue and the target location. We term the relationship between the directionality of the flankers and the central stimulus 'congruence.' Trials were 'congruent' when all three stimuli pointed in the same direction, and 'incongruent' when the central cue pointed opposite the two flankers. Thus, on a 'valid, congruent' trial, all three stimuli pointed toward the target, while on an 'invalid, congruent' trial, none of them did. Conversely, on a 'valid, incongruent' trial, the central cue alone pointed towards the target location, while on an 'invalid, incongruent' trial, only the flankers pointed toward the target. Again, we emphasize that none of the cues were predictive of the target location, and subjects were informed about this property.

The full-factorial design yielded 16 conditions, each of which was presented 10 times per block, which was repeated three times for a total of 480 trials. Before beginning the experiment, participants completed 16 practice trials and were provided feedback on their response time and accuracy. We excluded incorrect trials (2.4%) and filtered based on RT (additional 0.6%) of trials.

Results & Discussion

In accordance with Experiment 1, subjects responded faster when the SOA was 600ms (355.8ms) than 100ms (396.5ms), $F(1, 19) = 124.39$, $p < 0.001$, $\eta^2 = 0.867$. Additionally, the central cue seemed to exhibit a stronger overall effect than the flankers, as the main effect of validity was also significant, $F(1, 19) = 6.52$, $p = 0.019$, $\eta^2 = 0.255$, with faster overall RTs for valid trials (374.1ms) than for invalid (378.5ms). In contrast to Experiment 1, there was no main effect of stimulus on response times, $F(1, 19) = 1.69$, $p = 0.209$, with reaction times to the hand (375.3ms) nearly identical to those for the arrow (377.4ms). One explanation is that the redundant cueing of the flankers facilitated encoding for both stimulus types, bringing both close to a floor, or minimum, encoding time.

We expected the flankers to contribute some influence on orienting beyond the central cue. The interaction of congruence and validity was significant, $F(1, 19) = 64.67$, $p < 0.001$, $\eta^2 = 0.773$, indicating that the arrangement of the flankers influenced response times (see Figure 3).

To better understand the role of the flankers on orienting, we tested the validity effect. On congruent trials we observed a mean RT advantage for valid trials of 23ms, while incongruent trials showed an advantage in the opposite direction, on average of 14ms, a reversal of 34ms. These differences were both significantly different from zero, $t(19) = 7.47$, $p < 0.001$, $t(19) = -4.80$, $p < 0.001$ for congruent and incongruent trials respectively. The incongruent flankers thus not only mitigated the cueing effect of the central stimulus, but actually exerted a greater influence and shifted attention to the opposite side. This is a critical finding because it indicates that, while located away from participants' point of gaze, the two flanking stimuli were attended and exerted a greater net influence on orienting than the central cue. Lastly, the interaction of congruence with stimulus type was not significant, $F(1, 19) = 1.84$, $p = 0.191$, suggesting the strength of the cues was the same for both types of stimulus.

The greater cueing strength provided by the redundant cues appears to have created a floor effect and eliminated the stimulus specific effects seen in Experiment 1. To test our critical hypothesis, that encoding speed differences will predict differential allocation of attention towards social vs. non-social stimuli when the two are present simultaneously, we modified the current paradigm by placing hands and arrows together in the same stimulus array.

Experiment 3

In the preceding experiment, flankers modulated the directional influence of the central cue when all three stimuli were from the same class. In this experiment, we test the critical prediction of the selection hypothesis: when presented with both social and non-social stimuli, attention should be biased towards processing the social cues as the result of faster encoding. Specifically, we predict that incongruent flanking arrows will not significantly interfere with a central hand, and thus a validity effect (valid < invalid) will still be observed. By contrast, incongruent flanking hands should interfere with the central arrow resulting in a reverse of the validity effect (invalid < valid).

Methods

The procedure and design for Experiment 3 was identical to Experiment 2 with one exception; rather than showing three of the same stimuli, we included hands and arrows within the same display. The flankers were always pairs of the same class (e.g. two arrows or two hands) and matched with a central cue from the other class (e.g. hands flanking a single arrow or vice versa). Overall error percentages were low (0.5% of trials) and filtering based on reaction times excluded another 1.1% of trials.

Results & Discussion

In line with the previous experiments, we observed a main effect of SOA, $F(1, 19) = 296.79$, $p < 0.001$, $\eta^2 = 0.940$ with longer SOAs showing faster responses, and a main effect of validity, $F(1, 19) = 30.67$, $p < 0.001$, $\eta^2 = 0.617$, with valid responses overall faster than invalid.

Our main prediction was that the flankers would produce different effects depending on the stimulus arrangement. When the arrows flanked the hand, we expected them to have a smaller effect than when the hands flanked the arrows. For congruent stimuli, the addition of the flankers should bolster the validity effect. This was the case: the addition of congruent hand flankers provided a bigger boost to the central arrow (21ms) than the congruent arrows provided the central hand (14ms). When all three stimuli were hands the validity effect was larger (Experiment 2; 22.3ms) than when the flankers were arrows (Experiment 3; 14ms), though a post-hoc t-test showed no significant difference, $t(38) = 1.85$, $p = 0.073$ (two-tailed).

Even more striking are the results from the incongruent condition. When the hands and arrows provided conflicting information, targets in the direction cued by the hand(s) always led to faster response times, regardless of whether they were placed in the center or appeared as flankers. This effect is visible as a 10ms reversal of the validity effect between centrally presented hands and arrows (see the blue bars in Figure 4). Consistent with these results, the interaction between stimulus and congruence was significant, $F(1, 19) = 5.24$, $p < 0.05$, $\eta^2 = 0.216$.

Together, these results suggest that the pointing hand stimuli automatically drew attention at the expense of attention towards the arrows. The asymmetry of the flankers' influence is best understood in terms of faster encoding of the hands as compared to the arrows. This initial difference during encoding resulted in a cascading effect, leading to faster responses to targets cued by the direction of the hands.

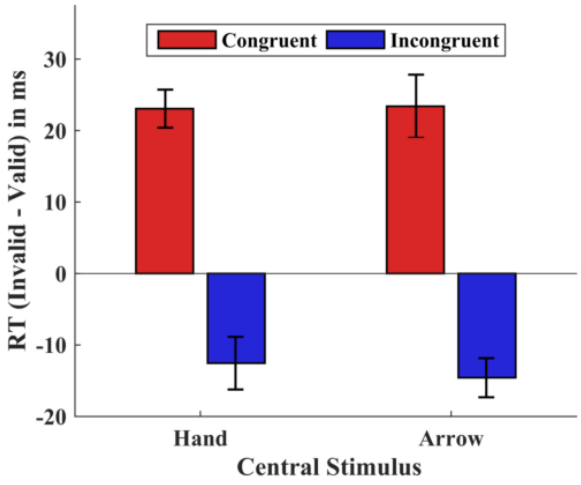


Figure 3. Validity effect (Invalid – Valid reaction times) for response times in Experiment 2. Here we show the main effect of Stimulus and lack of interaction of the Stimulus type with Congruence.

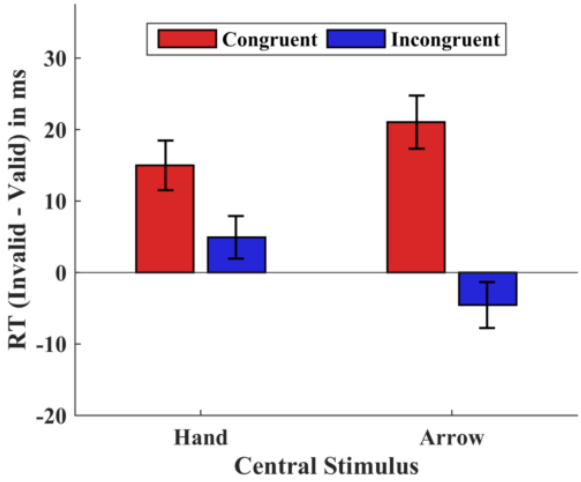


Figure 4. Validity effect for response times in Experiment 3, showing the interaction between Stimulus and Congruence. Note the label on the x-axis refers to the central stimulus which was always paired with flankers from the other class.

One alternative explanation for the differences in Experiment 3 is that the stimuli differed on low-level features like contrast or spatial frequency. In order to address this question, we utilized a computational salience algorithm (Harel, Koch, & Perona, 2006) to compute a feature-based prediction of attention for each stimulus array. To rule out a featural bias towards the flankers, we calculated the ratio of the total salience on the flankers relative to the central cue. These ratios did not correlate with the size of the validity effect, $p = 0.982$, suggesting any low-level changes were insufficient to account for the changes we saw. Flanker to central ratios and graphical depictions of the predicted distributions for each stimulus configuration are shown in Figure 5.

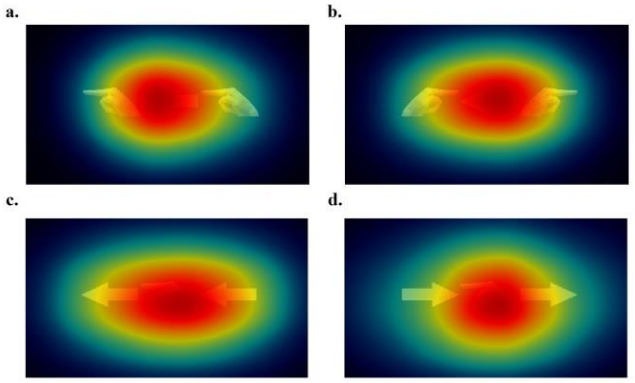


Figure 5. Output from the salience algorithm (Harel at al., 2006) showing the feature-based prediction of the distribution of attention and flanker-to-central ratios for the four stimulus configurations used in Experiment 3. (a) Congruent, Arrow: 1.33 (b) Incongruent, Arrow: 1.42 (c) Congruent, Hand: 1.57 (d) Incongruent, Hand: 1.37.

General Discussion

In Experiment 1, we showed that the pointing hands generated faster overall response times, but saw no difference in the magnitude of the validity effect. These results are in line with faster encoding of the social cue, but no difference in speed of orienting. We modified the basic paradigm in Experiment 2 to include flankers from the same stimulus type as the central cue. The presence of the flankers modulated the effect of the central cue on both congruent and incongruent trials, suggesting simultaneous processing of all three stimuli. Moreover, their presence eliminated the stimulus differences in Experiment 1, which suggests that the redundant cues facilitated encoding of both the hands and arrows. Finally, in Experiment 3, we leveraged participants' ability to process multiple cues to pit pointing hands and arrows against one another. We found evidence suggesting that participants' faster encoding of the hand stimuli resulted in preferential attention towards the hands, and ultimately a greater influence on subsequent orienting.

Previous research has brought to light important differences in the way we attend and process social stimuli, but has struggled to disentangle the multitude of component processes that contribute to the observed results. Our findings are consistent with the position put forth by Birmingham and Kingstone (2009) which proposes that social stimuli preferentially attract attention. Moreover, we suggest this preference emerges as a downstream effect of early encoding differences.

The competing explanation, that the differences between social and non-social stimuli lie in the speed of orienting, was not supported. The validity effect provides a measure of the strength of the stimulus cue for covertly orienting attention. The size of the validity effect was no different

between the hands and arrows in either Experiment 1 or Experiment 2. Only in Experiment 3, when the two stimuli competed for attention, did we see a stronger validity effect in favor of the hands on both congruent and incongruent trials. These results are most parsimoniously explained by faster encoding of the directionality of the pointing hands, without positing the additional effect of orienting differences.

An alternative account of the observed difference between the hand and arrow is via low level features like spatial contrast. We employed one algorithm and found no relationship between the computed salience of the stimuli and participants' relative attention towards the central and flanking cues. However, a number of limitations preclude us from completely ruling out this account. This algorithm was designed to estimate where subjects would fixate during an extended, free viewing paradigm, and our stimuli occupied a relatively small region of the screen subtending only about 7° of visual angle. Subjects were instructed to attend to the preceding fixation cursor and knew the location of the stimulus array, so it is reasonable to assume their point of gaze was directed towards the stimuli. This approach thus only provides an account of how attention may have been distributed across the multiple stimuli during the first fixation. To better distinguish between category and feature-based accounts, future experiments should directly manipulate low level properties as in the work of Sui, Rotshtein, and Humphreys (2013).

References

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS regions. *Trends in Cognitive Sciences*, 4(7), 267–278.

Birmingham, E., Bischof, W.F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition*, *16*(2/3), 341-355.

Birmingham, E., & Kingstone, A. (2009). Human social attention: A new look at past, present, and future investigations. *Annals of the New York Academy of Sciences*, *1156*, 118–40.

Burton, A. M., Bindemann, M., Langton, S. R. H., Schweinberger, S. R., & Jenkins, R. (2009). Gaze perception requires focused attention: evidence from an interference task. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 108–18.

Crostella, F., Carducci, F., & Aglioti, S. M. (2009). Reflexive social attention is mapped according to effector-specific reference systems. *Experimental Brain Research*, *197*(2), 143–51.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, 6(5), 509–540.

Eimer, M. (1997). Uninformative symbolic cues may bias visual-spatial attention: Behavioural and electrophysiological evidence. *Biological Psychiatry*, 46, 67–71.

Farroni, T., Csibra, G., Simion, F., Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9602-9605.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bulletin & Review, 5(3), 490–495.

Friesen, C. K., Ristic, J., & Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *Journal of Experimental Psychology: Human Perception and Performance*, 30, 319–329.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4), 694–724.

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Proceedings of the Neural Information Processing Systems (NIPS)*.

Hershler, O., & Hochstein, S. (2005). At first sight: a high-level pop out effect for faces. *Vision Research*, *45*(13), 1707–24.

Kuhn, G., & Benson, V. (2007). The influence of eye-gaze and arrow pointing distractor cues on voluntary eye movements. *Perception & Psychophysics*, 69(6), 966–971.

Langton, S. R. H., & Bruce, V. (1999). Reflexive Visual Orienting in Response to the Social Attention of Others. *Visual Cognition*, 6(5), 541–567.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50-59.

Langton, S. R. H., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition*, *107*(1), 330–342.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 3–25.

Ricciardelli, P., Bricolo, E., Aglioti, S. M., & Chelazzi, L. (2002). My eyes want to look where your eyes are looking: exploring the tendency to imitate another individual's gaze. *Neuroreport*, 13(17), 2259-2264.

Sui, J., Rotshtein, P., Humphreys, G. W. (2013). Coupling social attention to the self forms a network for personal significant. *Proceedings of the National Academy of Sciences*, 110(19), 7607-7612.

Tipper, C. M., Handy, T. C., Giesbrecht, B., & Kingstone, A. (2008). Brain responses to biological relevance. *Journal of Cognitive Neuroscience*, 20(5), 879–91.

Tipples, J. (2002). Eye gaze is not unique: automatic orienting in response to uninformative arrows. *Psychonomic Bulletin & Review*, *9*(2), 314-318.

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210–38.

When High WMC Promotes Mental Set: A Model of the Water Jar Task

Erin E. Sovansky (esovan2@uic.edu**)**
Stellan Ohlsson (stellan@uic.edu**)**

University of Illinois at Chicago
Department of Psychology (MC 285)
1007 West Harrison Street
Chicago, Illinois, 60607-7137

Abstract

Differences in working memory capacity (WMC) relate to performance on a variety of problem solving tasks. High WMC is beneficial for solving analytical problems, but can hinder performance on insight problems (DeCaro & Beilock, 2010). One suggested reason for WMC-related differences in problem solving performance is differences in strategy selection, in which high WMC individuals tend toward complex algorithmic strategies (Engle, 2002). High WMC might increase the likelihood of non-optimal performance on Luchins' (1942) water jar task because high WMC solvers tend toward longer solutions, not noticing when shorter solutions become available. We present empirical data showing this effect, and a computational model that replicates the findings by choosing among problem solving strategies with different WM demands. The high WMC model used a memory-intensive strategy, which led to long solutions when shorter ones were available. The low WMC model was unable to use that strategy, and switched to shorter solutions.

Keywords: Working memory capacity; problem solving; strategy selection; computational modeling

Background

Problem solving, like cognitive processes generally, is bounded by resource limitations (Simon, 1972). In particular, the capacity of working memory (WMC) has repeatedly been found to be related to problem solving performance. In a major review, Wiley and Jarosz (2012) concluded that, "In analytical problem solving, the superior executive function associated with WMC seem to generally support more successful problem solving." (p. 219).

However, as Wiley and Jarosz point out, this conclusion may only hold for analytical problem solving. In this type of process, problems are solved by extrapolating from prior experience, and the problem solver makes steady, step by step progress towards the goal.

In contrast, creative problem solving is characterized by the need to override prior experience in order to identify solutions that do not confirm to or follow from that experience (Ohlsson, 1992, 2011). Evidence is accumulating that in creative problem solving, the relationship between WMC and performance works differently than the "more is better" relation observed in analytical problem solving. Several studies have documented a reversed relation, in which problem solvers with lower working memory capacity performing better on insight problems than solvers with greater capacity. For example, DeCaro, Van Stockum, and Wieth (2015) found that low WMC participants outperformed high WMC participants on match stick arithmetic problems (Knoblich et al., 1999) and insight word problems (Schooler, et al., 1993, Wieth & Burns, 2006). Similarly, other studies have found that if WMC is reduced through alcohol intoxication (Jarosz, Colflesh, & Wiley, 2012) or solving problems during one's non-optimal time of day (Wieth & Zacks, 2012), insight problem solving improved while analytical problem solving suffered. This outcome is counterintuitive and stands in need of explanation. Why is lower capacity associated with greater probability of reaching an insight solution?

One possible explanation is that WMC influences the types of strategies used during problem solving. Those with high WMC are better able to control attention, giving them an increased ability to suppress distracting information and process more information relevant to the task at hand (Engle, 2002; McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010). Beilock and DeCaro (2007) suggested that individuals with lower WMC may lack the attentional control required to accurately use complex problem solving strategies and instead use associatively based strategies, whereas individuals with high WMC are able to use complex strategies. However, "sometimes high WMC participants may attempt to use complex strategies when simpler, more elegant, or more direct approaches are available" (Wiley & Jarosz, 2012, p. 210). Reliance on complex strategies are often beneficial on analytical problems which require the solver to hold multiple steps in working memory while progressing toward the goal (Jarosz, 2015). However, insight problems often trigger an inappropriate representation of the problem (Ohlsson, 1992), and using a complex strategy based on this representation will lead the solver toward impasse. The solution to an insight problem typically requires that complex yet familiar problem solving strategies are abandoned in favor of searching for a novel solution (Knoblich, Ohlsson, & Raney, 2001).

The Present Study

The purpose of the present paper is to describe a computational model of the classical findings on the water jar task. The latter was introduced into problem solving research by Luchins (1942). In this task, the solver is presented with a set of three jars of specified values, and is instructed to use only these jars to obtain a desired amount of water. The original problem set included a single practice problem to acquaint the solver with the task, followed by a set of ten

problems (see Table 1). The practice problem is solved by filling jar A, then subtracting water with jar B three times (A - 3B). Next, problems 1-5 can be solved by filling jar B, then subtracting water with jar A once and jar C twice (B - A - 2C). Problems 6-7 are solvable using the previous *long formula*, but can also be solved using a *shorter formula* of filling jar A and either adding or subtracting jar C (A +/- C). Problem 8 can only be solved using the shorter formula, and problems 9-10 can once again be solved with either the long or short formula. The main finding of interest is that the participants showed a strong tendency to use the long path on the problems that could also be solved by the shorter and more elegant, single-step path. However, after being given a problem that could only be solved with the shorter solution, they then applied that shorter solution to the two subsequent problems that could be solved with either the short or the long solution. The tendency to continue using the more complicated, but previously successful solution was called the Einstellung effect by Luchins (1942), and is now more commonly referred to as mental set (Smith & Blankenship, 1991).

Because the later problems in the water jar task can be solved using two different solutions, it is a particularly interesting task for examining how strategy use varies with WMC. Switching to the shorter solution is an insight-like process because it requires the solver to change their representation of the problem. In contrast, continuing to use the longer solution suggests that the solver may be using complex algorithmic strategies which tend to be better for analytical problem solving. Based on previous problem solving research, it would then be expected that low WMC individuals may be more likely to switch to the shorter solution than high WMC individuals, and high WMC individuals are more likely to experience mental set.

Two recent studies using the water jar task have found this effect (Beilock & DeCaro, 2007; Van Stockum & DeCaro, 2015). Our model provides a possible explanation why less working memory capacity might be associated with higher probability of finding and using the short solution on water jar problems.

Table 1: Classic water jar problem set (Luchins, 1942).

Problem	Formula(s)	Jar A	Jar B	Jar C	Goal
0	Practice	29	3	0	20
1	Long	21	127	3	100
2	Long	14	163	25	99
3	Long	18	43	10	5
4	Long	9	42	6	21
5	Long	20	59	4	31
6	Long, Short	23	49	3	20
7	Long, Short	15	39	3	18
8	Short	28	76	3	25
9	Long, Short	18	48	4	22
10	Long, Short	14	36	8	6

Problem solving typically relies on prior knowledge and experience, and successful problem solutions and task performances are generated by extrapolating that prior experience and applying it to a new problem or situation. Such analytical problem solving necessarily involves potentially complex interactions between working memory and long-term memory, because the latter has to be engaged in locating the relevant prior experience, retrieve it, and adapt it to the task at hand.

However, insight problems are characterized by the need to override prior experience and engage other processes than memory retrieval in order to achieve a novel solution path. The latter might include so-called *weak methods*, problem solving strategies of high generality (Laird & Newell, 1983). If the cognitive load required to access long-term memory is greater than the cognitive load imposed by those alternative processes, then working memory capacity might be one of the factors that impacts the probability that a problem solver will cease trying to use prior experience and instead engage problem solving processes that are abstract and local and hence might find a different solution than the one implied by prior experience.

It is plausible that the participants store information about the longer path in memory during the initial five problems, and that they draw upon that information when solving each successive problem. However, memory encoding is seldom perfect and similar items, such as steps in the water jar problem, are subject to interference. Applying what is remembered from a previous problem might impose significant cognitive load.

However, people also possess general or abstract problem solving processes. A common type of strategy is to compare the current state of the problem at hand with the desired or designated goal state, and be guided by how or in what respect they differ. Strategies of this sort are commonly referred to as means-ends strategies (Newell & Simon, 1972). The latter type of computation can be performed on information that is visible to the participants (current water levels in the jars and the desired goal state), and so do not impose high cognitive load nor require operations on long-term memory. The participants might prefer to work a problem by extrapolating prior experience, and only switch to an abstract and local difference-reduction strategy when prior experience imposes too high a cognitive load or turns out to be unsuccessful.

A Computational Model

Our model assumes that people have multiple strategies for solving water jar problems that vary with respect to the WMC they require. WMC interacts with problem solving by impacting strategy choice. The current version of the model utilizes three different problem solving strategies. This was not meant to be an exhaustive list of all possible strategies, but to exemplify plausible strategies that a human solver might use for the water jar task. One strategy was *solving from memory*. In this strategy, if there are steps from a previously successful solution path stored in memory, the model can follow the path to check if that will solve the next

problem. This strategy could be considered a type of case-based reasoning (Riebeck & Schank, 1989).

A second strategy is a *difference reduction strategy*, which is a simplified form of a means-ends analysis (Newell & Simon, 1972).. This strategy finds a starting value by evaluating which two operations will put the solver closest to the goal value and will pick the first operation of the two. For each subsequent step, the solver picks a value to add or subtract from the current state that will bring the current state closest to the goal value.

In order to prevent infinite loops of adding and subtracting the same value, the model uses a form of the no-loop heuristic (Atwood & Polson, 1976). If the solver is about to undo a previous step by adding the same value that was subtracted in the previous step or vice versa, it will instead randomly pick a jar.

The third problem solving strategy is *guided random solving*. For this strategy, at each step the solver determines whether a value needs to be added or subtracted in order to get closer to the goal, and then randomly selects a jar to perform the operation.

One potential way in which the model can deviate from human performance is that unless a limit is placed onto the model, the difference reduction and random solving strategies are both capable of performing an infinite number of steps until a solution is reached, whereas a human solver would only be capable of performing a limited number of steps. In order to resolve this problem, the number of steps that could be taken by either of these strategies was limited to a maximum of seven steps. A limit of seven steps was chosen because people on average can hold about seven or fewer items in memory depending on the type of information being stored (Baddeley & Hitch, 1974; Cowan, 2010; Miller, 1956).

The model was implemented in python 3.4 and is approximately 300 lines of code. Working memory was a fixed size storage for which the capacity could be specified. If a problem successfully solved a problem using any strategy, the solution path was saved into the model's long term memory, which then was available to be used by the memory strategy. For the memory strategy, the number of steps that can be saved is limited by WMC. This means that if the solution path was four steps long, but WMC only allows memory for three steps, only the first three steps will be saved. The steps were saved as a list of steps for which each step had an operation (add or subtract), and the jar used to perform the operation (A, B, or C), For example, the path for the longer solution would appear in WM as: ((add, B), (sub, A), (sub, C), (sub, C)). This solution path would remain in WM until a problem is solved using a different solution, and then the new solution would be saved into WM.

The model selects strategies in order from highest to lowest WM demand. It will first attempt the solving from memory strategy, followed by the difference reduction strategy, followed by the guided random strategy. This order was chosen because studies of WMC and strategy choice in problem solving have found that those with higher WMC tend to use more demanding strategies, sometimes instead of using less demanding, but valid strategies. This suggests that if a more demanding strategy can be used, it is more likely that it will be used (Beilock & DeCaro, 2007; Wiley & Jarosz, 2012).

However, because human solvers do not always select strategies in such a deterministic way, there is some noise in the strategy selection so that ten percent of the time it will skip the solve from memory strategy, and if it skips the first strategy, in ten percent of those cases, it will also skip the difference reduction strategy and go straight to the random solving strategy. We have not yet found a way to ground this parameter in the empirical data.

Empirical Study

Participants were 67 undergraduate students who were enrolled in an introductory psychology course and received credit for participation in this study.

Materials and Measures

Working Memory Capacity. WMC was measured using the automated symmetry span task (aSymspan; Redick, et al., 2012), and the automated running span task (aRunspan; Broadway & Engle, 2010). The aSymspan is a computer-based complex span task in which a memory task and processing task are interleaved.

In the *aSymspan*, participants judge whether an image is symmetrical across a vertical axis followed, and are then presented with a red square located in a 4x4 grid. After 2-5 trials, participants are then shown a grid and asked to click on the locations of the red squares in the order they were presented. Participants complete 12 sets of trial, 3 of each length. A participant's score is the number of red squares correctly remembered, and can range from 0-42.

The *aRunspan* is a computer-based simple span task in there is not a separate processing and memory component of the task. Participants are told to remember the last specified amount of letters in a string (3-7). Then participants are shown a string of letters of unknown length one at a time. Participants are then shown a screen with letters and are asked to click on the specified letters in the order they appeared. Score is the number of letters correctly remembered, and can range from 0-75. The WMC measures took approximately 5-10 minutes each to complete.

Problem Solving. Participants completed the water jar problems shown in Table 1 in the order presented. Problems were presented on paper with one problem per page. Participants first received an instruction page which included a completed example problem, and the practice problem (problem 0). Once participants correctly solved the example problem, they were given the rest of the task to complete. This task took approximately 15-20 minutes to complete.

Modeling Results

In order to explore whether the model replicates WMC-related differences in performance on the water jar task, the model was run 20 times, 10 with high WMC and 10 with low

WMC. For high WMC, the model was capable of remembering five steps, and for low WMC, the model was capable of remembering three. The classic Luchins (1942) problem set was used, and problems were performed in the order listed in Table 1.

Overall, the high WMC version of the model had higher accuracy in solving the water jar problems than the low WMC solver. The proportion correct was .97 for high WMC and .80 for low WMC. However, when broken down by problem, it can be seen that the differences in accuracy are driven by a few specific problems (see Table 2). More specifically, the low WMC solver failed to solve problem 5 all ten times, problem 3 nine times, and problem 4 three times. In contrast, the high WMC solver failed problem 5 in two instances, but successfully solved problems 3 and 4 every time. Errors occurred on these problems when the solver used the difference reduction or random strategies.

Table 2. Proportion correct on each water jar problem as a function of WMC.

	Low WMC		High WMC	
Problem	Model	Human	Model	Human
0	1.00	1.00	1.00	1.00
1	1.00	0.91	0.90	0.95
2	1.00	0.91	1.00	1.00
3	0.10	0.91	1.00	0.82
4	0.70	1.00	1.00	1.00
5	0.00	1.00	0.80	0.95
6	1.00	0.95	1.00	0.91
7	1.00	1.00	1.00	1.00
8	1.00	0.68	1.00	0.59
9	1.00	1.00	1.00	0.95
10	1.00	1.00	1.00	0.95
Total	0.80	0.94	0.97	0.92

Strategy use also varied by WMC (see Table 3 for a summary). The low WMC solver was unable to successfully use the solving from memory strategy, and instead used the difference reduction strategy, and rarely the random strategy to successfully solve the problems. The high WMC solver successfully solved from memory on half of the problems, and solved using the difference reduction strategy on just under half of the problems. For the high WMC, when the solver did not solve from memory, the majority of these instances were problems in which the most recent successfully solved problem had a different solution formula, so there was not a relevant path stored in memory.

Table 3. Proportion of strategy use as a function of WMC

Strategy	Low	High
Memory	0.00	0.50
Dif. Red.	0.79	0.47
Random	0.01	0.00
Fail	0.20	0.03

The main question is whether the model replicated the finding that high WMC solvers are more likely than low WMC solvers to use the long solution on problems 6-7 and 9-10. The model behaved in exactly this way (see Figure 1). The low WMC solver always used the short solution formulas instead of the long, whereas the high WMC solver used the long solution formula just under half the time (see Figure 2). When specifically examining problems 6-7, the high WMC solver had an even higher tendency toward using the longer solution, using it 90% of the time. On problems 9-10, because problem 8 was solved using a short solution, the high WMC solver overcame Einstellung and used the short solution.

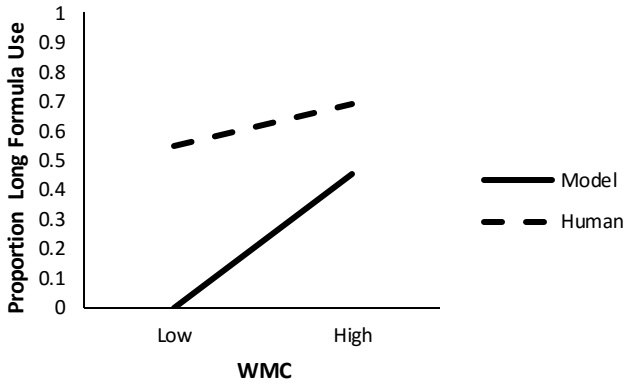


Figure 1. Proportion of long formula use on problems 6-7 and 9-10 as a function of WMC.

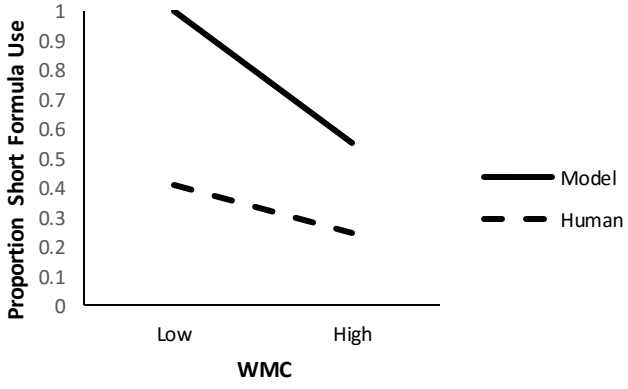


Figure 2. Proportion of short formula use on problems 6-7 and 9-10 as a function of WMC.

Empirical Results

Participants' WMC score was based on a factor score created by calculating the shared variance between aSymspan score and aRunspan score. In order to compare model data to human data, participants were split into low, medium, and high WMC groups. The low and high WMC groups were used to compare to the model.

High and low WMC groups did not differ in overall accuracy, $t(42) = 0.84$, $p = .41$ (Table 2.). Additionally, there were no WMC differences in accuracy on any individual problem. The low WMC human solvers had higher accuracy than the low WMC model, and the high WMC human solvers had slightly lower accuracy than the high WMC model

Although not significant, we found that the high WMC group had a higher rate of using the long solution formula on problems 6-7 and 9-10 compared to the low WMC group, $t(42) = -1.19$, $p = .24$, and a lower rate of using the short solution formula, $t(42) = 1.52$, $p = .14$. When analyzed as a correlation across all participants, there is a negative correlation between WMC and using the short solution formula, $r(65) = -.28$, $p = .02$, and a marginal positive correlation between WMC and using the long solution formula $r(65) = .22$, $p = .08$. Compared to the model, both low and high WMC human solvers showed a higher tendency of using the long solution.

Discussion

This model demonstrated how WMC influences strategy use on the water jar task and how strategy selection in turn affects the likelihood of experiencing mental set. By placing WMC limits on the memory strategy, the model was able to simulate the finding that high WMC solvers are more likely to use the long solution. When WMC was high, like human solvers, the model was more likely to continue using the long solution on problems 6-7, even though the short solution was available. The high WMC solver generally did not switch to the short solution until it failed to solve problem 8 from memory and used difference reduction to search for a new solution. When WMC was low, the solver was not able to store the full four step solution of the long formula, and was incapable of solving from memory using the long formula. The low WMC solver instead used the difference reduction and guided random solving strategies.

Even though the model was able to simulate WMC differences in formula use, compared to human performance, the model under predicted the likelihood of continuing to use the long solution once the short solution becomes available. There are a couple possible explanations for this finding. One possibility is that people resist changing strategies in a way that this model does not account for. Another possibility is that there are more strategies that could lead to using the long solution. Other possible strategies could include the undershoot or overshoot strategies used by Lovett on her model of the building sticks task, which is an isomorph of the water jar problem (Lovett, 1998; Lovett & Anderson, 1996).

One limitation of the model is that the low WMC solver failed to solve certain problems (problems 3-5) at a much higher rate than low WMC human solvers. This may also be because humans were using problem solving strategies not included in the model. One possible future direction would be to perform a think aloud study in order to learn what strategies people are using to solve this problem. Any new strategies that are learned could be incorporated into a future iteration of this model.

Another limitation to this model is that it selects its strategies in a pre-specified order: memory, difference reduction, and then random search. The only variation is that it sometimes skips an earlier strategy. People are not likely to move down a list of strategies in a particular order, especially if a strategy has not proven to be successful on previous problems. Another future direction could be to have the model randomly select a strategy based on the weighted utility of the strategy. The utility could be updated on success or failure of the strategy. If WMC determines which strategies can be used, then it would be expected that high and low WMC versions of the model would give higher utility to different strategies, with high WMC giving higher utility to more memory-intensive strategies.

Our results support the theory that strategy selection in problem solving is influenced by WMC limits. High WMC problem solvers are better able to make use of memory-intensive strategies such as remembering entire solution paths or algorithms. Low WMC problem solvers are less likely to use these strategies because they may require storing more information in working memory than the solver is able, and have to rely on strategies with lower memory demand. On the water jar task, the difference in strategy use meant that the high WMC solver had higher overall accuracy than the low WMC solver, but it was also more likely to use a non-optimal solution when a shorter possible solution was introduced. The model explains the Einstellung effect as a consequence of the interaction between the structure of the task environment and the boundaries on human cognitive capacity. When the task environment supports extrapolating from prior experience and the extrapolation imposes low cognitive load, people will tend to respond on the basis of memory, with Einstellung, ruts, and mindlessness as consequences. But when extrapolation is capacity demanding and the environment allows a strategy that is based on perceptually available information and hence imposes low cognitive load, solutions that go beyond prior experience become possible. Hence, the counterintuitive beneficial effect of low WMC on insightful problem solving.

References

Atwood, M. E., & Polson, P. G. (1976) A process model for water jug problems. *Cognitive Psychology, 8,* 192-216.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The Psychology of Learning and Motivation*. New York, NY: Academic Press.

Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving

under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 983-998.

Cowan, N. The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*, 51-57.

DeCaro, M. S., Van Stockum, C. A., & Wieth, M. B. (2015). When working memory capacity hinders insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19–23.

Jarosz, A. F. (2015). *The relationship between working memory capacity and strategy use on tests of general fluid intelligence* (Doctoral Dissertation). Retrieved from http://indigo.uic.edu/bitstream/handle/10027/19717/Jarosz_Andrew.pdf?sequence=1

Jarosz, A. F., Colflesh, G. J. H., & Wiley, J. (2012). Uncorking the muse: Alcohol intoxication facilitates creative problem solving. *Consciousness and Cognition*, *21*, 487–493.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1534–1555.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*, 1000–1009.

Laird, J., & Newell, A. (1983). A universal weak method. Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Lovett, M. C. (1998). Choice. In J. R. Anderson, & C. Lebiere (Eds.). The atomic components of thought. Mahwah, NJ: Erlbaum.

Lovett, M. C. & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology*, 31(2), 168-217.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*, 1–95.

McCabe, D. P., Roediger, III, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology, 24*, 222–243.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1992). Information processing explanations of insight and related phenomenon. In M. Keane & K. Gilhooly (Eds.), *Advances in the psychology of thinking*. London, UK: Harvester-Wheatsheaf.

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. New York, NY: Cambridge University Press.

Riesbeck, C., & Schank, R. (1989). Inside Case-Based Reasoning.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122*, 166-183.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization, 1*, 161-176.

Smith, S. M., & Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. *The American Journal of Psychology, 104*, 61-87.

Van Stockum, C., & DeCaro, M. S. (2013). The path less taken: When working memory capacity constrains insight. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society. Berlin, Germany: Cognitive Science Society*.

Van Stockum, C., & DeCaro, M. S. (2015). Different working memory components predict flexibility versus mental set. Poster presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL.

Wieth, M., & Burns, B. D. (2006). Incentives improve performance on both incremental and insight problem solving. *The Quarterly Journal of Experimental Psychology, 59*, 1378–1394.

Wieth, M. B., & Zacks, R. T. (2011). Time of day effects on problem solving: When the non-optimal is optimal. *Thinking & Reasoning*, *17*, 387–401.

Wiley, J., & Jarosz, A. (2012). How working memory capacity affects problem solving. *Psychology of Learning and Motivation, 56*, 185–227.

Thermodynamics and Cognition: Towards a Lawful Explanation of the Mind

Larry A. Moralez (Larrymoralez@knights.ucf.edu)
Department of Philosophy, University of Central Florida
4000 Central Florida Blvd., Orlando, FL 32816-1352 USA

Luis H. Favela (Luis.Favela@ucf.edu)
Department of Philosophy, Cognitive Sciences Program, University of Central Florida
4000 Central Florida Blvd., Orlando, FL 32816-1352 USA

Abstract

An argument is developed to show that explanations of biological and physical systems can be unified via the second law of thermodynamics (SLT). The SLT's influence on the evolutionary history of life at the scale of the global Earth system justifies reunifying phenomena—i.e., mind and matter—whose separation dates back to Modern Western philosophy and still influences contemporary scientific investigations. From this perspective it appears that the necessity of ever-increasing entropy in nature may constrain the organization and behavior of living organisms and cognitive processes. Via an example of explaining memory at the scale of the brain-body-environment system, we recommend understanding cognition with respect to its role in increasing entropy in nature. This framework may lead to a fruitful understanding of cognition by appealing to the necessity of physical laws.

Keywords: action selection, cognition, entropy, memory, thermodynamics

Introduction

The classic dualism disjoining scientific descriptions of physical and biological systems rests on a perceived incommensurability. The division is rooted in Modern Western philosophy and has persisted contemporary science. Despite the ubiquitous constraints imposed on nature by the laws of physics, many domains of inquiry are indifferent to the general principles of these laws when investigating living organisms. In the cognitive sciences, dynamic systems theorists have made strides in bridging the two (e.g., Kelso, 1995; Thelen & Smith, 1994). Nevertheless, much of their work stops short of uniting accounts of cognition with a supreme law of physics like the second law of thermodynamics (SLT) (Varpula et al., 2013). If we instead approach cognizing organisms as physical systems, whose laws regulate all physical interaction within such systems, the constraints these laws impose on cognitive processes should be present and observable in organisms. Approaching the study of cognition from this perspective opens the door to promising lines of inquiry that may advance our understanding of the human mind.

The fecundity of bringing together the theoretical frameworks of physics and biology into a common sphere of theoretical and methodological practice is here explored by approaching the study of cognition with respect to the SLT. The case is made as follows: First, we reinforce the claim that the SLT predisposes the organizational and behavioral properties of all biological systems. Second, we argue that cognition, being a biological phenomenon, is likely to be invariably constrained by the SLT, which can be fruitfully understood as the quest for free-energy consumption in the least time. Third, we show that this approach to understanding cognition is readily amenable to the explication of two cognitive phenomena—memory and action selection—that enjoy a unique complementarity in nature in the context of the SLT by acting as an accelerating entropy production system.

The Biophysical Dichotomy

The dichotomy of biological and physical systems (*biophysical* henceforth) is rooted in Modern Western philosophy, particularly Descartes' mind-body dualism and Kant's autonomy of biology from physics (Swenson & Turvey, 1991). Since then, some researchers have attempted to develop domain specific laws of human behavior analogous to Newton's laws governing the heavenly bodies (cf. Stevens, 1957). However, psychology has experienced a steady decline in attempts to establish psychophysical laws that may provide explanations of cognitive phenomena. A study on psychology paper abstracts containing 'law' citations shows a decline from 22 references per 10,000 for entries occurring between 1900 to 1999, to 10 references per 10,000 for entries occurring in the decade to follow. Moreover, the latter works presented fewer attempts to formulate new laws, suggesting an increasing doubt regarding the lawfulness of cognitive processes (Teigen, 2002). Adding to this trend, Davidson (1995) further argued that explanations of the psychological—which understood in terms of propositional attitudes such as intentions, desires, and fears—is rational in nature, whereas the physical domain is chaotic and irrational in nature. Thus, he claims the two domains of description can develop rough correlations but will remain fundamentally incommensurable. It is important to note for present purposes that his argument appeals to minimal anecdotal empirical support, and is very much akin to the Hegelian arguments Chemero (2009) claims plagues immature sciences like cognitive science. An argument is "Hegelian" when apparently empirical propositions are false as a matter of logical necessity, such that certain frameworks for understanding the natural world are ruled out in advance. This is based on the notorious arguments Hegel made,

following Plato, concerning the idea that the number of planets in our solar system is necessarily seven (Craig & Hoskin, 1992; Hegel, 1801/1987). Eventually, though, theories must be justified via appeal to empirical facts.

What is clear is that the biophysical chasm Descartes and Kant constructed is alive and well in the contemporary scientific and philosophical literature. Rowlands (1999) argues that the privilege researchers often grant to the processes going on inside the brain by computational models and neuroimaging are direct descendants of Cartesian internalism, a position that explicitly posits two differing substances—mind and matter. Additionally, some have argued (e.g., Anderson, 2014) that the dichotomy of mind and matter is responsible for subsequent attempts to locate where the mind occurs in the head, resulting in the development of phrenology and its continued influence in cognitive neuroscience (cf. Uttal, 2011). If cognitive scientists truly wish to achieve a scientifically virtuous theory of the human mind, then they may want to alter their trajectories away from the research paradigms established in part by the biophysical dichotomy and instead aim for commensuration. However, this line of thought is itself contingent on empirical evidence that justifies rescinding the biophysical dichotomy. If the empirical data serves to confirm Davidson's claims disjoining biology and physics, then there may be little hope of grounding cognitive processes in the laws of nature in physics. We argue here that by exploring the ramifications of the SLT on biological phenomena, the SLT can be established as a driving force behind the ontogenesis of living organisms at the scale of the global Earth system. This systems-based approach will merit minimizing the historical divisions of biological and physical systems and suggest that new explanations of cognitive process can be developed in cognitive science if it refocuses its inquiries towards the organism at various scales within the context of the fundamental laws of physics—namely, the SLT. If successful, this position will open the door to a promising thermodynamically lawful account of cognitive processes.

The Second Law, Entropy, and the Global Earth System

The SLT is distinct from many other laws of nature in that it can be interpreted not only as a law about laws but also a position on Aristotelian end-directed physics (Swenson & Turvey, 1991). As we shall see, not only does the SLT indicate that there is an end that all physical processes strive towards, but also that living organisms proceed along the same structured path. Understanding cognitive phenomena with respect to this law may lead to promising lines of inquiry and generate hypotheses. Such accounts could provide grounds for revising covering law-like models of explanations of cognitive phenomena (cf. Walmsley, 2008). Knowledge of the prior conditions of a system is combined with a law of nature (e.g., SLT) that, ceteris paribus, enables us to derive an explanation of the corresponding event that details what *had* to happen given the conditions. The promise of this model works symmetrically with prediction as it does explanation: We can combine knowledge of a future set of initial conditions with natural laws to accurately predict the conditions to follow. Explanations and predictions of this kind are grounded in the nomological necessity of laws of nature and not logical or metaphysical necessity. Though attempts to ground cognitive processes in physical laws have recently emerged with varying degrees of success (e.g., Kelso, 1995; Thelen & Smith, 1994), none have successfully appealed to nomic necessity for explanation. If we wish to have a comprehensive explanation of why specific cognitive phenomena exist and how they function, pursuing nomic necessity may be a good place to begin such a pursuit.

Note that we are not claiming that non-covering law explanations will not be sufficient or even ideal for certain phenomena. Rather, being placed within the context of laws of nature can strengthen even non-covering law explanations. We defend a general theoretical framework whereby cognitive phenomena can be understood in a manner similar to physical systems, namely, as continuous with the SLT.

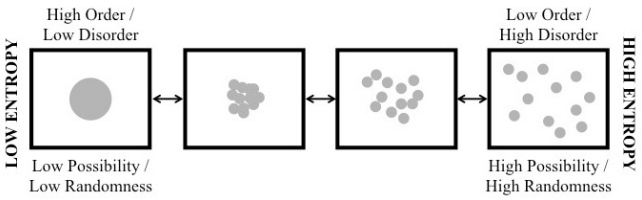


Figure 1: As the entropy of a system increases, it becomes less ordered/increases in disorder. As the entropy of a system decreases, it becomes more ordered/decreases in disorder.

The SLT states that within a closed physical system *entropy* can only increase over time (Boltzmann, 1886/1974; Brissaud, 2005). Entropy is the measure of disorder in a physical system. It is important to understand that when the total amount of order in the system is low, its entropy is high, and vice versa (Figure 1). Physicists generally agree that the only truly closed system known to exist is the universe itself. Thus, the universe is in the process of transitioning from a higher-ordered state to a lower-ordered state. At first glance it would seem the SLT supports the biophysical dichotomy, for it would entail that complex, higher-ordered living systems could not emerge in an open system that is lawfully trending towards disorder without violating the SLT. Dennett's (1995) interpretation takes this line of thought a step further into the domain of biology by defining living systems as those in nature that defy the SLT. However, this approach fails by re-emphasizing that the universe is the only closed system known to exist: It entails that biological systems cannot be closed systems and, consequently, accounts of their description (like Dennett's) cannot presume otherwise. Such interpretations naively disengage the highly ordered biological system from the environment by treating it as

unique with respect to the SLT, only taking into account the order produced explicitly within the system itself.

Contemporary treatments of living systems in isolation from their environments seem to have supplemented beliefs in the biophysical dichotomy that began long ago. However, von Bertalanffy (1950) demonstrated that in open systems such as living organisms, order must arise (decreasing entropy) whenever the opportunity presents itself, and at the same time the total net entropy in nature is increased via some process whereby it is released into the local environment. The SLT equation balances itself out once a multi-scale system perspective is taken into account and higher-ordered biological systems are understood as physically open systems within the closed universe system. If the net entropy produced in nature by the evolution of complex organisms is positive, then the SLT suggests that the proper construal of biological systems as physical systems dynamically situated within their environment must be taken. Evidence for this net increase in entropy at the scale of the global Earth system is suggested by empirical data with respect to the evolutionary history of life on Earth. From this a nascent perspective for understanding biology, and subsequently cognition, emerges.

Schrodinger (1945) argued that for a living thing to exist it must never cease contributing to the total entropy of the universe. Moreover, Swenson and Turvey (1991) established that the global Earth system reveals an increase in biospheric entropy during the development and growth of living matter over geological time. Primal life on Earth was likely to have been largely anaerobic, or without a supply of oxygen and biological compounds to metabolize like present day life. The emergence of the first photosynthetic bacteria linked life directly to the sun. Through an extraordinary development in the history of the planet, the anaerobic organisms were unable to split electrons from Earth's vast water supply for energy. Proto-Cyanobacteria were the first to do otherwise. This was accomplished by making use of the virtually endless supply of photons from the sun to split the electrons within the water molecules. This process resulted in the release of O_2 into the atmosphere. Combined with a vast supply of water and sunlight, this led to the rapid oxidation of the primal Earth atmosphere, a process that equates to a rapid increase in entropy. Since this process began, the global Earth system has generated an exponential increase of entropy production. The aforementioned Pre-Phanerozoic era, where the first photosynthetic organisms employed sunlight to release the O_2 molecules from water, significantly increased the terrestrial levels of entropy. The resulting high concentration of atmospheric O_2 provided a unique opportunity for generating additional entropy. It was then that nascent life forms metabolized the newly formed organic matter and O_2. This resulted in an even greater increase of entropy production along with increasingly higher ordered forms of life. This trend continues to this day. Thus, from the moment life began on Earth the production of entropy by progressively higher ordered biological systems was largely responsible for an exponential increase in net entropy in nature over time. The vantage point that views biological systems as situated thermodynamically within nature advances our claim that the SLT demonstrates an intrinsically homogenous biological and physical universe.

Entropy and the Free Energy Principle

, The SLT has multiple coherent interpretations. One interpretation—the law of maximum entropy production—suggests that entropy maximization and field potential minimization are expressions of the same symmetry (Swenson, 1997; Swenson & Turvey, 1991). When all available energy is evenly distributed throughout a system and there are no remaining local field potentials, then entropy is maximized and the system is at equilibrium. The system will remain at equilibrium (maximum entropy) unless acted upon by an external non-equilibrium system with free energy to consume. The SLT is also referred to as the principle of least action (Annila, 2010, Kaila & Annila, 2015). If the system is at equilibrium and there are no local free energy pools, no action can be taken. But when a non-equilibrium system is within proximity to affect a system at equilibrium—which is always the case for any open system—the equilibrium system reacts efficiently by consuming any readily available free energy with the least amount of action possible. Thus, according to several mutually consistent interpretations of the SLT, a low-entropy system must consume free energy by reducing local field potentials with the least amount of action if it is to remain consistent with the SLT and contribute a net increase of entropy in nature. As noted earlier, this increase occurs somewhere along the system's physical boundaries. This interpretation of the SLT has been referred to as the *free energy principle* (Friston, 2010) and it manifests in many processes in nature.

Rainwater flowing down a dry mountain riverbed readily illustrates the free energy principle (Figure 2A). Water does not merely flow in a straight line down the main body of the riverbed. Rather, the water flow branches off into any available side streams as the water follows the path of least resistance down the riverbed. Multiple paths minimize the constraints otherwise imposed on the water flow if it had been confined only to a straight line down the main body of the riverbed, thereby ensuring field potential minimization and maximum entropy increase in the fastest possible time. The riverbed analogy serves as an easy to grasp instantiation of the free energy principle in nature by demonstrating a familiar physical system's quest to maximize free energy consumption in the least amount of time. Given the ubiquity of the SLT throughout nature, it is unsurprising that the instantiation of the free energy principle is evident in cognitive processes as well (e.g., Friston, 2010; Friston & Stephan, 2007).

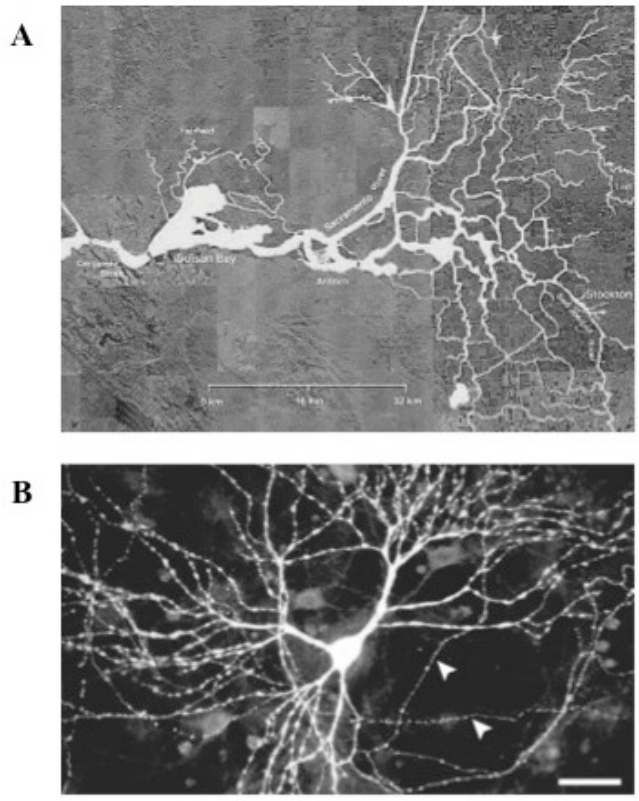


Figure 2: Increased energy flow results in increased speed of flow. (A) River flow of California's Sacramento-San Joaquin River delta (modified from Trump, 2004). (B) Expression of TRPV1, a receptor protein, in cultured hippocampal neurons (Zemelman et al., 2003). The hippocampus and receptor proteins such as TRPV1 are associated with memory formation in the brain (e.g., Gibson et al. 2008).

It may come as no surprise that animals have to move to eat, and that they need to eat to move (Varpula et al., 2013). A living organism's need to consume free energy to move and live entails that these cognitive perception-action cycles at the scale of the organism-environment system (Favela & Chemero, 2016) are constrained by the SLT. At the scale of the brain, the cortical mechanisms for action selection can be understood as a manifestation of this principle because it facilitates what actions an organism selects to stay alive and contributes to the net increase in biospheric entropy. Thus, for the animal, an action may be considered advantageous if it leads to maximized consumption of free energy in the least amount of time. Moreover, action selection is invariably constrained by an organism's memory. Memory is a cognitive phenomenon that can be understood as participating in extending nature's entropy increasing efficacy into the future.

A Multi-Scale Entropy Production System: Affordance Selection and Memory

Neurophysiological data suggests that neural activity within the brain is produced in a manner consistent with the free energy principle (Friston & Stephan, 2007). More specifically, any neuron or neuronal assembly that can change will do so in an effort to consume free energy in the least amount of time. Like rainwater flowing down a mountainside, as an activation signal traverses a neural pathway, it takes the path of least resistance and reduces energy field potentials within the brain as fast as possible (Figure 2). In the case of action selection, neurophysiological data suggests that multiple simultaneous opportunities for action are selected and specified in parallel (Cisek, 2007). This "affordance competition hypothesis" is different from the classical cognitivist perspective where action was thought to be selected first and specified after. Instead, neural activity in the brain produces bodily action once a neurological signal is received from the motor cortex, and the neurological signal that the motor cortex transmits will be the signal that survives the action selection competition carried out in the brain in accordance with the free energy principle. But an organism's free energy consumption is not restricted to the intracranial processes. Rather, in order to survive, the brain needs its free energy consumption process to adapt in a way that can guide the organism through dynamic environments. Unless an organism is situated in a static and threat-less environment with unlimited access to free energy, to survive it will need a memory system that reorganizes the morphology of the brain to enable adaptation to external circumstances. In addition to action selection, memory can be explicated in the context of the SLT.

New memories are made—at least in part—by the brain establishing new or strengthening old neural connections. The mechanism that enables this morphodynamic process in the brain is Hebbian learning. Hebbian learning can be summed up as "neurons that fire together, wire together" (Keysers & Perrett, 2004). The pathways that activated neural signals traverse throughout the nervous system are strengthened, subsequently providing a means of increasing the likelihood and efficiency that whatever environmental stimulus generated the initial neural activation will follow a similar path if reencountered in the future (Figure 2B). Again, the mountain riverbed analogy is helpful here. The more frequently water flow traverses a specific path down the riverbed, the more likely it is to erode and deepen the river. While river erosion and memory formation are wildly different processes, the erosion will serve to increase the likelihood that a similar path will be taken again during the next rainfall, which is akin to how Hebbian learning in the brain facilitates an increased likelihood that neural signals will traverse similar paths in the future. The process of memory formation via Hebbian learning enables an organism to facilitate an increasingly efficient means of dissipating field potentials from local free energy pools with its actions. Thus, combined with a Hebbian learning

process, the free energy that traverses an organism's nervous system simultaneously constructs a feedback process that increases the likelihood that it will select similar field potential minimizing actions in the future. From this perspective, memory can be vaguely understood as a physical feedback mechanism within an organism that enables the past to influence the present. Both implicit and explicit memory can be characterized by this description and is therefore indifferent to the different types of memory often instantiated in living organisms.

Memory is thus unique with respect to the SLT on the grounds that it permits the exploitation of temporal circumstances to increase nature's entropic efficacy. The SLT not only ensures all physical systems—including living organisms—consume free energy pools the moment they are within spatial proximity to be consumed, but if the system is organized with an effective memory system it will enable that system to forgo immediate free energy consumption in order to consume greater amounts into the future. The SLT ensures maximum entropy production in real time with respect to local field potentials. This is evident in cortical mechanisms for action selection. By grounding memory in the SLT we have an exclusive account of nature organizing itself in a way that removes spatial constraints imposed on entropy production. Non-autonomous and memory-less systems are restricted to entropy production only when they encounter free energy within spatial proximity of the system's boundaries. In the case of living physical systems endowed with memory, they are able to forgo immediate field potential minimization that presents itself locally within spatial proximity and instead extends nature's entropy, thereby increasing efficacy forward in time thanks to greater amounts of free energy consumption in the future.

The consequence of this unique cognitive phenomenon is that aspects of human culture—including education, socializing, career and financial planning, and science, all of which require memory to function—allow for sacrifices of immediate field potential consumption for an increased net consumption at a non-local point in time. This suggests that not only are there likely to be additional cognitive processes constrained by the SLT, but that additional phenomena from the domain of the social sciences can potentially be understood from this perspective.

Memory formation is necessary for virtually all living organisms that rely on their personal autonomy to move their bodies through their environment. It enables them to select appropriate action—in the interest of avoiding risk and seeking nourishment—so that they may continue to survive in the wild. In the context of the global Earth system detailed above, the combination of action selection and memory in living organisms compliment one another in a way that results in unique and increasingly viable entropy production system in nature.

Concluding Remarks

Having motivated a rejection of the assumption that biological and physical systems are separate and incommensurable within scientific explanation, we have presented accounts of cognitive processes with respect to the SLT and its symmetry with the free energy principle. This suggests that cognitive, biological, and physical phenomena can be examined and understood in some of the same ways. Further investigations into this matter may ultimately show that memory exists and is organized in a manner that is necessary for nature to continue increasing its entropy within the constraints of the global Earth system. Additional researchers have begun to notice the connection at additional scales and link it to other processes like sleep, perceptual sensation, learning, notions of self, task performance, single cell recordings, and neural information flow in neuroimaging (e.g., Collell & Fauquet, 2015; Dimitrov, Lazar, & Victor, 2011; Varpula, Annila, & Beck, 2013). This framework diminishes the risk of epistemological shortcomings that result from the influence of Modern Western philosophy in contemporary studies of the mind (for discussions of such shortcomings see Anderson, 2014; Assecondi, Bagshaw, & Ostwald, 2014; Chemero, 2009; Clark, 2008; Sporns, 2012; Uttal, 2011).

For a comprehensive and fruitful understanding of the mind, researchers will benefit from not only broadening the scopes of their domains to take into account the brain-body-environment system, but also by examining the constraints imposed on living organisms by the SLT—and perhaps other laws. Moreover, such an approach looks to be a promising line of inquiry for building a *bottom-up* understanding of the mind. "Bottom-up" is used here to refer to a framework that is grounded in the laws of nature that ascend into higher order processes. Motivating this model is an appeal to the conditions for covering law models of explanation, which derives hypotheses from sets of premises that contain at minimum one law of nature. If the premises turn out to be true (e.g., empirical evidence), then there is strong assurance that the explanation will entail nomological necessity, namely, whatever happened *had to* happen.

This approach presupposes the commensurability of our theories and explanations of both biological and physical systems. We defended this presupposition by highlighting the guiding role entropy production has played over the history of the global Earth system. Rescinding the biophysical dichotomy that has held much scientific practice captive requires a non-dualistic approach to understanding the mind. Cognitive systems ought to be examined as biological systems, and biological systems as physical systems that are thermodynamically situated within their environments. From there, we can draw from the copious amounts of knowledge that exists in the domains of cognitive science, physics, biology, neuroscience, and philosophy to continue developing a comprehensive, lawful understanding of the mind.

References

Annila, A. (2010). All in action. *Entropy, 12*, 2333-2358.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain.* Cambridge, MA: MIT Press.

Assecondi, S., Ostwald, D., & Bagshaw, A. P. (2014). Reliability of information-based integration of EEG and fMRI data: A simulation study. *Neural computation, 27*, 281-305.

Boltzmann, L. (1886/1974). The second law of thermodynamics. In D. McGuinness (Ed.), *Theoretical Physics and Philosophical Problems: Selected Writings* (pp. 12-32). Boston, MA: D. Reidel Publishing Company.

Brissaud, J. B. (2005). The meanings of entropy. *Entropy, 7*, 68-96.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society, 362*, 1585-1599.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* New York, NY: Oxford University Press.

Collell, G., & Fauquet, J. (2015). Brain activity and cognition: A connection from thermodynamics and information theory. *Frontiers in Psychology, 6*, 1-11, doi: 10.3389fpsyg.2015.00818.

Craig, E., & Hoskin, M. (1992). Hegel and the seven planets. *Journal for the History of Astronomy*, *23*, 208-2010.

Davidson, D. (1995). Could there be a science of rationality? *International Journal of Philosophical Studies*, *3*, 1-16.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York: Simon and Schuster.

Dimitrov, A. G., Lazar, A. A., & Victor, J. D. (2011). Information theory in neuroscience. *Journal of Computational Neuroscience*, *30*, 1-5.

Favela, L. H., & Chemero, A. (2016). The animal-environment system. In Y. Coelllo & M. H. Fischer (Eds.), *Foundations of Embodied Cognition: Volume 1: Perceptual and Emotional Embodiment* (pp. 59-74). New York, NY: Routledge.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*, 417-458.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*, 127-138.

Gibson, H. E., Edwards, J. G., Page, R. S., Van Hook, M. J., &, Kauer, J. A. (2008). TRPV1 channels mediate long-term depression at synapses on hippocampal interneurons. *Neuron*, *57*, 746–59.

Hegel, G. W. F. (1801/1987). Philosophical dissertation on the orbits of the planets (1801), preceded by the 12 theses defended on August 27, 1801. P. Adler (trans.). *Graduate Faculty Philosophy Journal*, *12*, 269- 309.

Kaila, V. R. I., & Annila, A. (2015). Natural selection for least action. *Proceedings of the Royal Society A*, *464*, 3055-3070. doi: 10.1098/rspa.2008.0178

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: A Hebbian perspective. *Trends in Cognitive Sciences, 8*, 501-507.

Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. Cambridge, MA: Cambridge University Press.

Schrodinger, E. (1945) *What is life?* New York: Macmillan

Sporns, O. (2012). *Discovering the human connectome*. Cambridge, MA: MIT press.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review, 64*, 153-181.

Swenson, R. (1997). Autocatakinetics, evolution, and the law of maximum entropy production: A principled foundation towards the study of human ecology. *Advances in Human Ecology*, *6*, 1-48.

Swenson, R., & Turvey, M. T. (1991). Thermodynamic reasons for perception action cycles. *Ecological Psychology*, *3*, 317-348.

Teigen, K. H. (2002). One hundred years of laws in psychology. *The American Journal of Psychology, 115*, 103-118.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Trump, M. (2004). Sacramento–San Joaquin River Delta. *Wikipedia*, photograph. Retrieved January 29, 2015 from http://en.wikipedia.org/wiki/Image:Wpdms_usgs_photo_sacramento_delta_2.jpg. GNU Free Documentation License.

Varpula, S., Annila, A., & Beck, C. (2013). Thoughts about thinking: Cognition according to the second law of thermodynamics. *Advanced Studies in Biology*, *5*, 135-149.

Uttal, W. R. (2011). *Mind and brain: A critical appraisal of cognitive neuroscience*. Cambridge, MA: MIT Press.

Von Bertalanffy, L. (1950). The theory of open systems in physics and biology. *Science, 111,* 23-29.

Walmsley, J. (2008). Explanation in dynamical cognitive science. *Minds & Machines*, *18*, 331-348.

Zemelman, B. V., Nesnas, N., Lee, G. A., & Miesenböck, G. (2003). Photochemical gating of heterologous ion channels: Remote control over genetically designated populations of neurons. *PNAS*, *100*, 1352-1357. Image retrieved January 29, 2016 from http://www.pnas.org/content/100/3/1352/F2.large.jpg

Benefits for Grounded Feedback over Correctness in a Fraction Addition Tutor

Eliane Stampfer Wiese (eliane.wiese@berkeley.edu)
Graduate School of Education, 4407 Tolman Hall
Berkeley, CA 94720 USA

Rony Patel (rbpatel@andrew.cmu.edu)
Kenneth R. Koedinger (krk@cs.cmu.edu)
Department of Psychology, 342c Baker Hall
Pittsburgh, PA 15213 USA

Abstract

Do students activate conceptual and procedural knowledge simultaneously when learning fraction addition? In grounded feedback, student actions on a target, to-be-learned representation are reflected in a more familiar feedback representation to promote conceptual learning within procedural practice. An experiment with 163 4th and 5th graders shows improved learning with a grounded feedback tutor over a symbols-only control with step-level right/wrong feedback. Learning with grounding also transferred to symbols-only assessment items, providing some support for the simultaneous activation view.

Keywords: fraction addition; simultaneous activation; magnitude representation

Introduction

Math and science are often communicated with abstract symbols. Learning these domains involves fluently using these symbols and correctly applying conceptual principles to them. How might a second representation provide grounding for learning a symbolic representation? *Grounded feedback* is based on the common characteristics of tutor designs that were previously shown to be successful: students manipulate a to-be-learned representation, while a linked representation reflects those inputs in a more accessible form (Mathan & Koedinger, 2005; Nathan, 1998). This feedback aims to take the less-familiar representation that the student is learning and ground it both in another representation and in the student's prior knowledge. We hypothesize that grounded feedback allows the student to apply her prior conceptual knowledge to the more-familiar feedback representation and then decide if her work with the to-be-learned representation is correct. This hypothesis follows from Ohlsson's theory of learning from performance errors: learners identify errors when there is a discrepancy between what the learner expects and what actually happens (Ohlsson, 1996). Grounded feedback provides the context in which the discrepancy can occur. For example, a learner may guess that 1/10 is larger than 1/4 because 10 is bigger than 4. Comparing two equal-sized rectangles, one with 1/10 and one with 1/4 shaded should alert the learner to his error: he expects 1/10 to have more shaded, but sees that it has less. The more accessible rectangle representation serves to disambiguate the meaning of the symbolic representation (Ainsworth, 1999). Importantly, in grounded feedback students do not directly manipulate the more accessible representation. Transfer between symbolic and non-symbolic representations is difficult for students (Uttal et al., 2013), likely because the cognitive demands of working in each type of representation are different (Sarama & Clements, 2009). Therefore, while grounded feedback includes an accessible representation to facilitate sense making and self-evaluation, having students act directly on the to-be-learned representation encourages transfer. Specifically, this grounded feedback tutor teaches fraction addition, and students act directly on the symbolic fractions.

Solving a fraction addition problem correctly may involve at least two steps: rejecting incorrect strategies and using the correct strategy. The most common incorrect strategy for fraction addition problems is the independent whole number strategy (Ni & Zhou, 2005). To execute this strategy, students independently add the numerators and denominators of the addends to get a final answer. For example, when adding 1/2 and 1/3, students incorrectly executing the independent whole number strategy would get the answer 2/5. 2/5 is less than the expected sum. In fact, it's even less than one of the addends (1/2). 34% of fraction addition and fraction subtraction problems resulted in errors due to the use of the independent whole number strategy for 6th and 8th graders (Siegler, Thompson, & Schneider, 2011). Grounded feedback can help students realize why the independent whole number strategy is incorrect (i.e., it results in a magnitude-incongruent answer) and why the correct strategy is correct (i.e., it results in a magnitude-congruent answer).

How might fraction magnitude knowledge and fraction arithmetic knowledge be related? The dynamic view proposes the two are independent and become progressively more so over time (Anderson, 1983). In contrast, the simultaneous activation view argues that arithmetic computation errors are the product of a lack of relevant concepts being simultaneously activated to reduce implausible solutions (in this case, magnitude knowledge; Hiebert, 1987). To test these two views, Byrnes and Wasik (1991) ran two studies to establish temporal precedence between fraction magnitude knowledge and fraction arithmetic knowledge and empirically intervene to teach

fraction arithmetic knowledge with or without magnitude. Results showed that teaching with magnitude did not improve learning above a purely procedural approach. Students made arithmetic errors despite having an abundance of magnitude knowledge, suggesting they had the conceptual knowledge to reject the arithmetic errors, but were not bringing that knowledge to bear, as the simultaneous view would suggest. However, while the magnitude instruction included a *demonstration* of using fraction bars and coordinating between the two representations, students did not actually *practice* this skill themselves. Coordinating the fraction bars and fraction symbols is not trivial for students (Stampfer & Koedinger, 2013), and this coordination may be a pre-requisite skill for the activation of conceptual knowledge in a procedural context. To that end, the grounded feedback condition includes pre-instruction on interpreting the fraction bars, and the grounded feedback tutor showed students the magnitudes of the converted fractions and sums that they were proposing. The dynamic magnitude representations were intended to help students bring their existing magnitude knowledge to bear while practicing the procedure. Our experiment supports this notion, as the grounded feedback students matched the control students on fraction addition gains while outperforming them on more conceptual questions.

A previous experiment comparing grounded feedback to a symbols-only control (the correctness tutor) found similar pretest to posttest gains for both conditions, though the grounded condition had greater pre-test to delayed gains (Wiese, 2015). One explanation for why the grounded condition did not outperform the control was that grounded students often seemed unable to correctly interpret and integrate both representations (Ainsworth, 1999; Wiese, 2015). The current experiment investigates if pre-instruction on the feedback representation and a longer intervention time can lead to greater learning gains relative to a control.

Grounded Feedback for Fraction Addition

Figure 1 shows a screenshot from the grounded feedback tutor, constructed with CTAT (Aleven, McLaren, Sewall, & Koedinger, 2006). Students input numbers at the bottom of the interface, while fraction bars reflect the converted and sum fractions in a more concrete form. The fraction bars aim to ground the symbolic fractions by making their magnitude more salient. In addition, the grounding relies on students' prior knowledge of equivalence: equivalent fractions have the same magnitude, so equivalent fraction bars have the same amount colored in. Grounded feedback allows students to see the consequences of their errors and thus may promote students' evaluation of their own work (e.g., a student may guess that 8/24 + 9/24 = 17/48, but the fraction bars show 17/48 is too small). While the grounded feedback tutor offers on-demand text hints, it does not provide step-level right/wrong feedback, and does not prevent students from erasing correct inputs. A previous experiment compared this tutor to a correctness tutor, which did not include fraction bars but did provide immediate step-level feedback (correct inputs were colored green and incorrect inputs were colored red) (Wiese, 2015). In the correctness tutor, students were not permitted to change correct inputs. With both tutors, students were required to solve the current problem correctly before moving on.

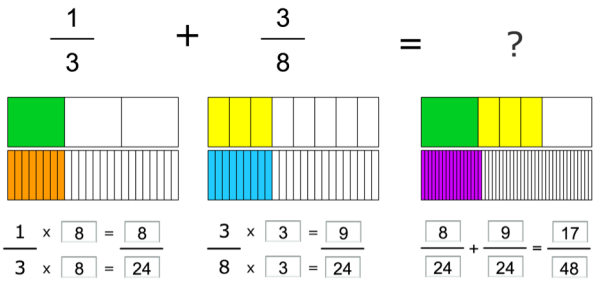


Figure 1: Grounded feedback tutor. Top row of fractions and yellow and green fraction bars are given, second row of bars dynamically shows students' inputs as they are typed in boxes at the bottom. Not shown: the window for on-demand hints, and the "done" button.

Prior Research on the Grounded Feedback Tutor

Prior work found that students learned from the grounded feedback tutor, but also indicated that students found the feedback unclear. Participants in a think-aloud study used the fraction bar feedback to identify and fix mistakes (Stampfer, Long, Aleven, & Koedinger, 2011) and a classroom study with 5th graders found learning benefits (Wiese, 2015). However, those students did not use the fraction bars effectively - they often clicked the "done" button when the fraction bars did not line up (Wiese, 2015). A follow-up study assessed how well 5th graders could evaluate fraction addition equations when fraction bars were provided as scaffolds. Equations were presented in four formats: three included fraction bars, and one was a numbers-only control (Fig. 2) (Stampfer & Koedinger, 2013). Students saw one correct and one incorrect equation in each format, and were asked to indicate if the equation was true or false. Incorrect sums were obtained by adding the numerators and denominators independently. The average of students' scores with the numbers-only format was 21%, far below their performance with the fraction bars. Still, performance with the fraction bars was low: 79% with the pictures-only format, 64% with pictures and numbers, and 46% with half pictures and numbers (Stampfer & Koedinger, 2013). These scores indicate that, while the fraction bars improve performance, they are not enough for students to reliably determine when an equation is correct or not, explaining students' confusion with the tutor.

Pre-Instruction on the Fraction Bar Representation

To help students interpret the fraction bar representations, the current grounded feedback tutor includes up-front instruction on the fraction bars. The instruction consists of multiple-choice problems, beginning with questions on fraction equivalence (expected to be within students' prior knowledge; Stampfer & Koedinger, 2013) and gradually fading in the addition operations and fraction symbols. This progression is based on concreteness fading (Fyfe, McNeil, Son, & Goldstone, 2014). Students were given immediate

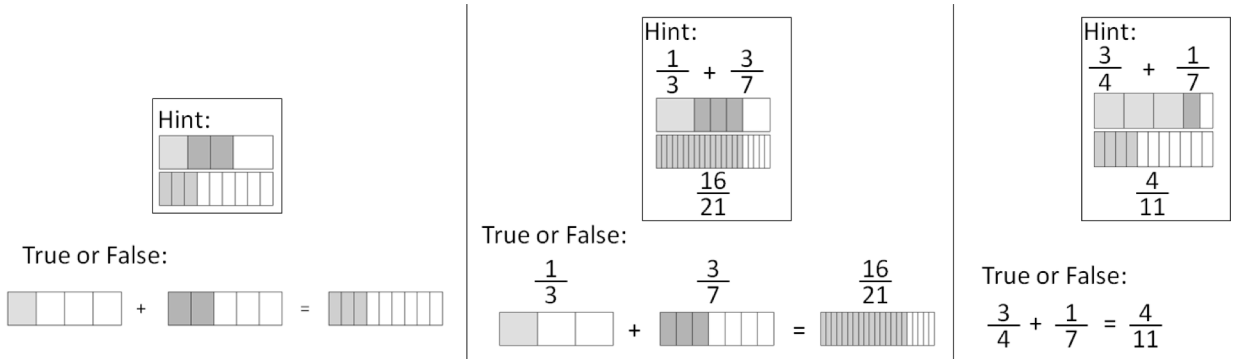


Figure 2: Sample addition questions in the three formats that included fraction bars. From left to right: pictures only, pictures and numbers, half pictures and numbers. The numbers-only format showed the symbolic equation and answer options without any fraction bars.

correctness feedback and on-demand hints. Sample problems are shown in Figs. 3-5.

Experiment: Grounded vs. Correctness

This experiment compared learning with the grounded and correctness feedback tutors, using a pretest-intervention-posttest design. Both tutors included the same brief instruction on using the tutor software and on fraction addition. The grounded feedback tutor included the pre-instruction on fraction bars.

Materials, Participants, and Procedures

The 29-question pre- and posttests included 12 symbolic fraction addition items and 9 evaluation items that proposed a fraction addition equation and asked if the sum was correct, too big, or too small (3 each of pictures only, numbers only, and both pictures and numbers). Answers were scored 1 if correct and 0 otherwise. Two matched tests (same problem types, different numbers) were counterbalanced, question order was determined randomly, and half of the tests were given in reversed question order.

194 students from 9 classes at a local public school participated in the experiment (60 4th graders and 134 5th graders). The school tracked students by achievement, and teachers identified their classes as high (3), average (5), or low (1). 31 students were removed from the sample because they were absent during the pre- or posttest, or they spent less than 45 minutes on their assigned tutor, leaving 163 students (78 grounded, 85 correctness). The experiment took place at the school during class time over four consecutive days. All random assignment was within-class. Students were given a 15-minute pretest, worked with a randomly assigned tutor for up to 80 minutes, and then took a 15-minute posttest the next day. The tests were administered on a computer and students could not return to previously answered questions.

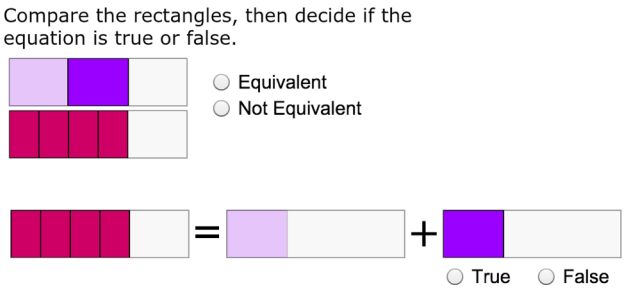


Figure 3: Question 1. 53% of students solved the problem, without hints, on their first try.

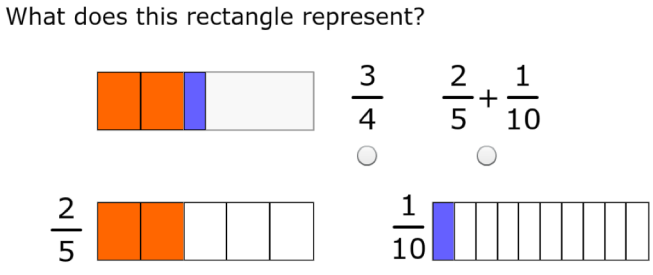


Figure 4: Question 10. 81% of students solved the problem, without hints, on their first try.

Results

Table 1: Average scores (and standard deviations) for overall tests and subtests.

Condition	Test	Total	Addition	Evaluation	Other
Correctness	Pre	.43 (.20)	.32 (.27)	.42 (.26)	.60 (.24)
	Post	.59 (.22)	.49 (.30)	.63 (.26)	.69 (.18)
Grounded	Pre	.42 (.19)	.35 (.26)	.42 (.23)	.57 (.23)
	Post	.63 (.22)	.55 (.32)	.69 (.23)	.71 (.22)

Did the grounded condition learn more than the correctness condition? Overall, yes. Table 1 shows the average scores for the overall pre- and posttests and for the

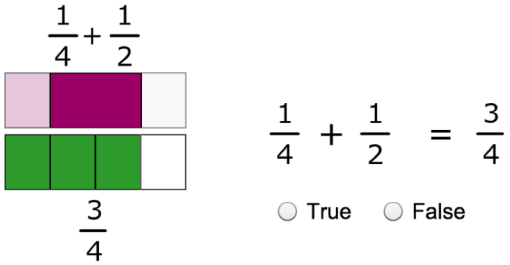


Figure 5: Question 14. 68% of students solved the problem, without hints, on their first try.

three subtests, by condition. To test that pretest differences were not significant, an ANOVA was run on pretest score, with pretest order, pretest form, class tracking level, and condition as fixed factors, and class as a random factor. The first model included all main effects and two-way interactions. After removing non-significant interactions and main effects, the final model included a marginal effect for order ($p = .07$), a marginal order by pretest form interaction ($p = .08$) and a significant class by pretest form interaction ($p = .04$). Condition was not significant ($p = .7$). Paired samples t-tests show all within-condition differences from pre- to posttest are significant ($p < .01$). To test if condition had a significant effect on learning, we re-ran the final model, this time on posttest score, with pretest score as a covariate. The first model included all two-way interactions with pretest score. After removing non-significant interactions and main effects, the final model included class and total pretest score as significant main effects (both $p < .01$) and condition as a marginal main effect ($p = .065$), in favor of grounded feedback. The same tests were repeated on the addition and evaluation subtests – condition was not significant in either case.

How did transfer from the grounded tutor to a symbols-only assessment compare to transfer from the symbols-only tutor to a dual-representation assessment? To determine if there were condition differences for scores on the numbers only and pictures and numbers evaluation items, a MANOVA was run on the posttest scores for each scaffold type, with corresponding pretest scores as covariates and class and condition as fixed factors. The condition by class interaction was not significant in the multivariate test so the model was re-run without it. Multivariate tests showed pretest scores and class were significant ($p < .04$), as was condition ($p = .047$), in favor of grounded feedback. Condition was significant on the posttest score for the pictures and numbers scaffold ($p = .015$, again in favor of grounding), but not for the numbers only scaffold. Figure 6 shows the estimated marginal means for the two scaffold types, by condition.

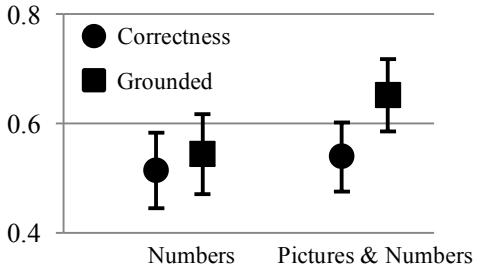


Figure 6: Estimated marginal means for posttest evaluation items that included numbers, with 95% confidence intervals (y-axis is from .4 to .8).

Did Students Learn from the Fraction Bar Pre-Instruction? The fraction bar instruction aimed to help students interpret the grounded feedback. One measure of success is how often students pressed the "done" button when the proposed sum differed from the correct some by more than .1: on average .16 times per problem (.34 on average for the first 20 problems, compared to .99 for the 20 problems in the previous study; Wiese, 2015). Another measure of learning comes from a two-question pre- and posttest bracketing the pre-instruction. Similar to the question shown in Fig. 5, the test questions proposed a fraction addition equation with the fractions represented both symbolically and as fraction bars. Students indicated if the proposed sum was correct, too big, or too small. These pre- and posttests included one true equation and one false equation, where the sum was obtained by adding the numerators and denominators independently. Both before and after instruction, the average score was 63% correct. Errors were categorized as whole number error, other error, or skipped. A whole number error indicates incorrect transfer from whole number addition: answering 'correct' to a sum obtained by adding the numerators and denominators of the addends, and answering 'too big' to the correct sum. Answers that were not correct or whole number errors were coded as other. Table 2 shows the proportion of each error at the fraction bar pre- and posttest (this table includes the 95 students who completed this section, not just the 78 grounded students included in the other analyses).

Table 2: Proportion of correct answers and error types for the fraction bar pre- and posttest

	Correct	Whole Number Error	Other Error	Skipped
Pre	63%	30%	6%	1%
Post	63%	23%	13%	1%

Table 3: Pearson correlations between error types on evaluation items and performance on free-response fraction addition items. *p < .03

Response on Fraction Addition Items	Whole Number Error	Other Error	Correct
Percent Correct	-.42*	.12	.30*
Rate of Whole-Number Error	.31*	-.11	-.21*

After the fraction bar instruction, students had fewer whole number errors. To determine if one type of error indicates better understanding, we examined correlations between each type of error and proficiency at fraction addition problems. The study pretest included two evaluation questions that were isomorphic to those used in the fraction bar pre- and posttest, and 12 free-response symbolic fraction addition problems. For this analysis we include students who saw both of the evaluation questions, and calculated scores and error rates on the addition items based on the questions that students saw (i.e., disregarding questions that students ran out of time for). Table 3 shows the correlations between occurrence of each error type and (1) score on the fraction addition items and (2) rates of student-generated whole-number errors on the addition items. These results show that correct responses on the evaluation items are correlated positively with correct responses on the fraction addition items and correlated *negatively* with whole-number errors on the fraction

addition items; the reverse is true for whole-number errors on the evaluation items.

Case Studies: Using Grounded Feedback

Is grounded feedback easier to work with than correctness feedback? On average, students in the grounded condition solved fewer fraction addition problems (38 vs. 74 for correctness), took longer per problem (65 seconds per problem vs. 40), and requested more hints per problem (1.4 vs. 0.4), indicating that grounded feedback was more difficult.

How did students make use of the grounded feedback? Log data suggests two pathways: responding to the grounded feedback directly to diagnose and correct errors, and using grounded feedback to decide when to ask for a hint. Figures 7-8 illustrate the first strategy for a student converting 3/8 to 24ths. The student is adding 1/3 and 3/8, and got a hint for the denominator of the first fraction that said to multiply 3 by 8. The student correctly chose to multiply 8 by 3 to get the denominator for the second fraction, but then decided to multiply the numerator by 6. Figure 7 shows the student's interface at this point. The grounded feedback shows that 18/24 is bigger than 3/8. Next, the student tries 10 as a numerator (still to big), and then 9 (Fig. 8). After the grounded feedback shows that 9/24 equals 3/8, the student updates the multiplication area to show 3 x 3 = 9. In this case, the student does not seem able to find the equivalent fraction using symbols alone: the student does not begin by multiplying the numerator and denominator by 3. Instead, the student appears to use the grounded feedback to inform a guess-and-check strategy,

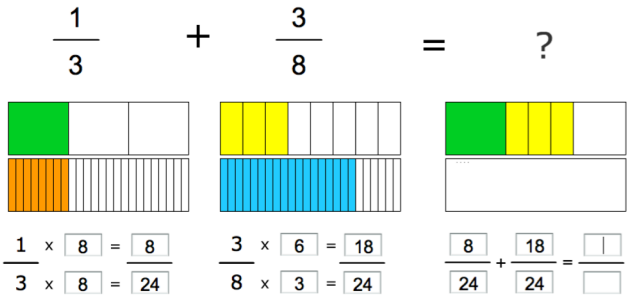


Figure 7: The grounded feedback tutor. The student is converting 3/8 to 24ths

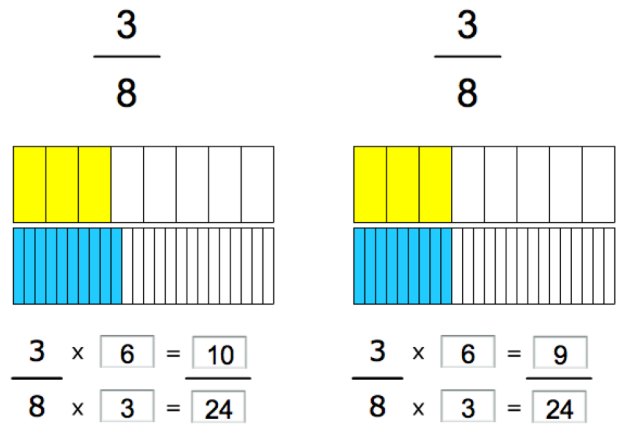


Figure 8: Grounded feedback for each guess-and-check conversion attempt

identifying the direction of the error and correctly deciding when that part of the problem is complete (after converting the second fraction, the student moves on to the sum).

In other cases, the feedback may facilitate learning from hints. In one example, a student adding 4/9 and 1/9 entered 5/18 for the sum (the whole-number error). The student seems to interpret the feedback as showing an error, but appears unsure of how to fix it. Instead of pressing the done butting or guessing, the student asks for hints until the answer is provided. On the next problem, the student converts the addends incorrectly, and then uses the whole-number strategy on the converted fractions, again asking for a hint only after entering the incorrect sum (perhaps the student pays more attention to the addition section of the interface than the converting sections, or the student might not realize that the converted fractions should be equivalent to the addends). This student does not attempt the whole-number strategy on any subsequent problems. Here, the grounded feedback appears to have shaken this student's confidence in that incorrect strategy, perhaps facilitating acceptance of the correct strategy offered in the hints.

Discussion

Correctness feedback is easier to work with than grounded feedback, indicated by students solving many more correctness problems, spending less time per problem, and requesting fewer hints on each problem. How does the additional difficulty of grounded feedback affect learning? The marginal significance in favor of grounded feedback on overall learning and the non-significant difference on the addition subtest indicates that grounded feedback is no worse than correctness. The differences in learning on the evaluation items with pictures and numbers also suggest that the additional difficulties in grounded feedback are desirable. Those items include the same representations present in the grounded tutor. The numbers-only evaluation items only included the symbolic representation present in the correctness tutor. Therefore, the pictures and numbers items can be considered target items for the grounded students while the numbers only items are transfer, and visa versa for the correctness students. With this view, the grounded feedback students were better than the correctness students at transferring their knowledge to the less-familiar format: Grounded students scored just as well on the numbers only problems as the correctness students, while outperforming them on the pictures and numbers items. At the very least, the similar performance of both conditions on the fraction addition items and numbers only evaluation items shows that including the fraction bars during learning did not impede students' performance with numbers on the posttest.

Did students learn from the fraction bar tutorial? Scores on the evaluation items bracketing the pre-instruction did not change. However, students decreased their rates of whole number errors, switching to other errors instead. Whole number errors are negatively correlated with solving symbolic fraction addition problems correctly and are positively correlated with adding both numerators and denominators independently on such problems, while other errors are not correlated with either behavior. Therefore,

whole number errors appear to be more harmful than other errors, and a decrease in whole number errors suggests that students benefitted from the tutorial.

These results indicate that a longer intervention time (80 vs. 40 minutes) and the inclusion of fraction bar pre-instruction addressed the shortcomings of the grounded condition in the previous study (Wiese, 2015). Still, the case studies point to further areas for improvement. Even with the grounded feedback, students do not always seem to recognize when their work is incorrect (e.g., a student may recognize when a proposed sum is incorrect but may not recognize when a converted fraction is incorrect). Including correctness feedback with the grounding may help: Instead of relying on the grounding alone to evaluate the action and diagnose the error, the correctness feedback will evaluate the error, freeing cognitive resources to focus on the diagnosis.

Conclusions

This study shows an advantage for grounded feedback, and certainly no disadvantage, compared to a strong control condition. Students in this study seemed to be better able to interpret the grounded feedback than students in the previous study (Wiese, 2015), although the measures used (rates of incorrectly pressing the "done" button and performance on evaluation items) may be overly coarse. Control students did purely procedural practice with the fraction addition items, and improved on all test sections from pre-test to post-test. Even though students in the grounded condition had their mental resources split between the procedure and the magnitude concepts, they improved just as much on symbolic fraction addition, and outperformed the control on the conceptual evaluation items with symbols and magnitude. The dynamic view would suggest that the grounded condition's improvement on the conceptual items should come at a cost to the procedural ones. That grounded students improved as much as the control on the procedural items offers some support to the simultaneous activation theory.

Acknowledgements

This work is supported in part by the Pittsburgh Science of Learning Center through NSF award SBE-0836012, and by Carnegie Mellon University's Program in Interdisciplinary Education Research (PIER) funded by Grant R305B090023 from the U.S. Department of Education, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C100024 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education*, *33*(2-3), 131–152. doi:10.1016/S0360-1315(99)00029-9

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2006). The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In M. Ikea, K. Ashkely, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 61 – 70). Springer-Verlag Berlin Heidelberg.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Byrnes, J. P., & Wasik, B. A. (1991). Role of Conceptual Knowledge in Mathematical Procedural Learning. *Developmental Psychology*, *27*(5), 777–786.

Fyfe, E. R., McNeil, N., Son, J., & Goldstone, R. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, *26*(1), 9–25.

Hiebert, J. (1987). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, New Jersey: Erlbaum.

Mathan, S., & Koedinger, K. R. (2005). Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. *Educational Psychologist*, *40*(4), 257–265. doi:10.1207/s15326985ep4004_7

Nathan, M. J. (1998). Knowledge and Situational Feedback in a Learning Environment for Algebra Story Problem Solving. *Interactive Learning Environments*, *5*(1), 135–159.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and Learning Fraction and Rational Numbers: The Origins and Implications of Whole Number Bias. *Educational Psychologist*, *40*(1), 27–52.

Ohlsson, S. (1996). Learning from Performance Errors. *Psychological Review*, *103*(2), 241–262.

Sarama, J., & Clements, D. H. (2009). "Concrete" Computer Manipulatives in Mathematics Education. *Child Development Perspectives*, *3*(3), 145–150. doi:10.1111/j.1750-8606.2009.00095.x

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296. doi:10.1016/j.cogpsych.2011.03.001

Stampfer, E., & Koedinger, K. R. (2013). When seeing isn't believing: Influences of prior conceptions and misconceptions. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1384–1389). Berlin, Germany: Cognitive Science Society.

Stampfer, E., Long, Y., Aleven, V., & Koedinger, K. R. (2011). Eliciting Intelligent Novice Behaviors with Grounded Feedback in a Fraction Addition Tutor. In *Proceedings of the 15th annual conference on Artificial Intelligence in Education*.

Uttal, D. H., Amaya, M., Maita, M. del R., Hand, L. L., Cohen, C. A., O'Doherty, K., & Deloache, J. S. (2013). It works both ways: Transfer difficulties between manipulatives and written subtraction solutions. *Child Development Research*, *2013*. doi:10.1155/2013/216367

Wiese, E. S. (2015). *Toward sense making with grounded feedback*. Carnegie Mellon University. Retrieved from http://reports-archive.adm.cs.cmu.edu/anon/hcii/CMU-HCII-15-104.pdf

Our morals really depends on our language: The foreign language effect within participants

Kuninori Nakamura (knaka@seijo.ac.jp)
Faculty of Social Innovation, Seijo University
6-1-20, Seijo, Setagayaku, Tokyo, 155-0081, Japan

Abstract

Recent research has suggested that using a foreign language to present hypothetical moral dilemmas increases the rate of utilitarian judgments about those dilemmas (e.g., Greene et al, 2001) and decreases incoherency between judgments in framing effect tasks (e.g., Tversky & Kahneman, 1981; see Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014; Costa, Foucart, Hayakawa, Aparici, Apesteguia, Heafner, & Keysar, 2014; Keysar, Hayakawa, & An, 2012). However, existing research has mainly investigated this effect using between-participants designs (i.e., different participants in the foreign and native language conditions). Such designs are unable to exclude non-equivalent conditions as a confounding variable. In contrast, this study examined the foreign language effect using a within-subjects design (i.e., all participants responded to moral dilemmas (Greene et al, 2001) and framing effect tasks (Tversky & Kahneman, 1981) in both their native and foreign languages. The "foreign language effect" was replicated, excluding semantic non-equivalence between language conditions as a potential confound. This result supports the hypothesis that the foreign language effect is independent of meaning.

Keywords: foreign language effect; moral dilemmas; framing effect; individual differences

Introduction

Language may affect individuals' manner of thinking. This possibility has attracted many researchers' attention since the famous Sapir-Whorf hypothesis was first advanced (Carroll, Levinson, & Lee, 2012; Sapir, 1921). Languages' effect on thinking has received much empirical study; however, the discussion remains ongoing (for reviews, see Kay & Kempton, 1984; Takano, 1989).

Recent work on the "foreign language effect" (Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014; Costa, Foucart, Hayakawa, Aparici, Apesteguia, Heafner, & Keysar, 2014; Keysar, Hayakawa, & An, 2012) provides interesting data suggesting that languages affect human cognition. In these studies, participants completed various types of reasoning tasks including framing-effect tasks (Tversky & Kahneman, 1984) and moral dilemmas (e.g., Greene et al, 2001) in either their native or foreign languages.

The framing effect provides an initial demonstration of the foreign language effect in reasoning (Costa, Foucart, Arnon, Aparici, & Apesteguia, 2014; Keysar, Hayakawa, & An, 2012). The framing effect causes equivalent descriptions of a decision problem to elicit systematically different decisions. This effect is robust and common; however, it is reduced or disappears in decision tasks not presented in participants' native language. (Tversky & Kahneman, 1981). For example, read the following vignette known as the Asian disease problem (Tversky & Kahneman, 1981);

> Recently, a dangerous new disease has been going around. Without medicine, 600,000 people will die from it. In order to save these people, two types of medicine are being made.
>
> Gain framing:
> If you choose Medicine A, 200,000 people will be saved.
> If you choose Medicine B, there is a 33.3% chance that 600,000 people will be saved and a 66.6% chance that no one will be saved.
>
> Which medicine do you choose?
>
> Loss framing:
> If you choose Medicine A, 400,000 will die.
> If you choose Medicine B, there is a 33.3% chance that there was a 33.3% chance that "no one will die and a 66.6% chance that "600,000 people will die..
>
> Which medicine do you choose?

As you see, the Gain and Loss vignette describe the same contents. However, participants who read the Gain framing tend to choose Medicine A, whereas those who read the Loss framing tend to choose B (Tversky & Kahneman, 1981). This indicates peoples' coherence in risky choice. Keysar,et al. (2012) demonstrate that this coherence in risky choice decrease when people read and answer the framing task in their foreign language.

Costa et al. also explored the foreign language effect in moral thinking (2014). Intuitively, moral judgments about "right" and "wrong" are the result of deep thought and should therefore be consistent and unaffected by factors irrelevant to moral reasoning such as language; however, recent studies (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) indicate that moral judgments are highly context dependent. The most prominent example of this contextual dependency is the difference between the switch and footbridge dilemmas. The switch dilemma assumes that a runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save these people is to activate a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Respondents must decide whether to divert the trolley in order to save five people at the expense of one. Most respondents indicate believing that one should

activate the switch (Greene et al., 2001). In the footbridge dilemma, a trolley threatens to kill five people (as before); respondents imagine themselves standing next to a large stranger on a footbridge that spans the tracks between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge and onto the tracks below. He would die in that case, but his body would stop the trolley from reaching the others. Respondents must thus decide to push the stranger off the bridge or to refrain; most respond that one should refrain. Assuming that the imagined loss of life is morally significant and the means to that loss is insignificant, this discrepancy between the two problems' response tendencies illustrates the contextual dependency of moral reasoning.

Costa et al. (2014) found that this discrepancy varied if the dilemmas were presented in a foreign language: participants solved moral dilemmas, including the switch and footbridge dilemmas, using either their native or a foreign language; across three studies incorporating several different languages, using a foreign language elicited more utilitarian judgments than using one's native language did. This supported the hypotheses that affective processes importantly affect moral reasoning and that using foreign languages decreases affective engagement.

These studies' results indicate systematic differences between cognitive processing in native and foreign languages; specifically, irrational decisions are reduced in framing-effect tasks when choices are presented in a foreign language (see also Costa et al., 2014) and moral dilemmas more frequently elicit utilitarian judgments when dilemmas are presented in a foreign language. Keysar, Hayakawa, and An (2012) used dual process theory to explain the foreign-language effect (e.g., Kahneman, 2003; Sloman, 1996; Stanovich & West, 2000). The dual-process model proposes that human cognition is composed of an analytic, rule-governed, and systematic system that employs many mental resources, and an intuitive, affective, and heuristic system. Keysar et al. proposed that using a foreign language moves people from the immediate affective system to a more deliberate, analytic mode of thinking (2014). Foreign languages are less grounded in speakers' emotions than their native language is (e.g., Pavlenko, 2005), and are typically processed less automatically than speakers' native language; this may lead to more deliberate cognition (Favreau & Segalowitz, 1983). Such deliberate cognition might more frequently elicit rational decisions. Additionally, foreign language is more difficult to process (Alter, Oppenheimer, Epley, & Eyre, 2007), possibly eliciting more analytic decision-making. The foreign language effect aligns with this suggestion (Costa, Foucart, Arnon, et al., 2014; Costa, Foucart, Hayakawa, et al., 2014; Keysar, Hayakawa, & An, 2012; see also Nakamura, 2015).

Previous studies have examined this effect indirectly using between-subjects designs; however, such designs cannot exclude the possibility that their results partly reflect differences between individual participants. Within-subjects designs yield results that do not reflect individual differences, confining comparison to the different languages' effect on each participant. This study therefore aimed to examine the foreign language effect using a within-subjects design.

Renderings of moral dilemmas in different languages may not have equivalent meaning or significance. In Nakamura (2015), Japanese participants responded to various moral dilemmas either in their native language (Japanese) or a foreign language (English) in two experiments. Nakamura used factor analysis of participants' responses to test the dilemmas' semantic equivalence between the two languages, and directly compared responses to the moral dilemmas between the two languages. In both experiments, a foreign language effect resembling that of Costa et al. (2014) was observed in participants' responses; however, factor structures varied between the native and foreign languages, indicating that the moral dilemmas' meaning varied between the two languages. This result implies that between-subject designs may not fully capture the foreign language effect: individual differences in cognition between the foreign and native language may be large enough to change participants' interpretation of the dilemma between the language conditions. Earlier research has consistently used between-subject designs; hence, their results may simply reflect non-equivalent dilemmas between language conditions, rather than language-dependent differences in moral judgment. Given this possibility, demonstrating the foreign language effect requires the ensured preservation of semantic equivalence between moral dilemmas in native and foreign languages.

Differences in factor structure may not reflect individual differences in the dilemmas' interpretation; however, it remains significant that differences between participants might affect responses to the foreign and native language conditions. Experimental design should therefore separate language effects from individual differences to clarify the foreign language effect.

Individual differences also affect interpretation of the foreign language effect in framing-effect tasks; additionally, the foreign language effect is apparent in comparison of risk-averse responses between native and foreign-language conditions: the difference in the risk-averse response rate between gain- and loss-framed conditions in a foreign language condition was smaller than that in a native language condition (Costa et al, 2014). This latter result led Costa et al. to conclude that using foreign languages enhances rational decision-making (2014). Nonetheless, attributing rationality to a participant requires that participant's judgment remains coherent throughout equivalent gain- and loss-framed scenarios; hence, comparison between participants does not directly indicate framing effect-induced irrationality.

Excluding individual differences is thus crucial to the examination of a possible relationship between language and thought (e.g., Kay & Kempton, 1984; Takano, 1989). To the authors' knowledge, previous studies examining this

Table 1 Moral dilemmas used in this study

	Content	Action
Switch	Kill one man to save five workmen	Throw switch to turn the train to the side track
Footbdidge	Kill one heavy man to save five workmen	Throw the man from the bridge
Donor	Kill one young man to save five patient	Transparent young man's organs to five patient
Hospital	Kill one patient to save five	Hit a certain switch, which will cause the fumes to bypass the room containing the three patients
Baby	Kill your baby to save tonwpeople	Smother your child to death
Sculpture	Destroy the sculpture to save one man	Push the sculptures into the valley so that it will roll onto the tracks and block the trolley's passage
Boat	Lie to the guard to save the toursits	Lie to the guard to borrow a nearby speedboat

topic have not adequately determined if foreign and native language differentially affect modes of cognition.

In sum, examining the foreign language effect using a within-subject design would be fruitful for both practice and theory. This study therefore centrally aimed to examine the foreign language effect using a within-subjects design. In this study, Japanese participants responded to moral dilemmas and framing tasks in both foreign and native languages.

Method

Participants

One hundred and thirty-two undergraduates participated; participants were compensated with course credits.

Materials and procedure

Seven moral dilemmas (including the Switch and Footbridge dilemmas) and two types of framing task were used. The framing task included gain and loss framing conditions. Participants thus responded to 22 problems ((7 moral dilemmas + 2 framing tasks (Asian disease and financial crisis) * 2 framing conditions (gain and loss)) * 2 language conditions (native and foreign)). All materials and response scales were presented using booklets; participation was compensated with course credits.

Following Nakamura (2015), moral dilemmas were adopted from Greene et al. (2001). Dilemmas were composed of three moral-personal dilemmas (viz., footbridge, transplant, crying baby) and four moral-impersonal dilemmas (viz., switch, standard fume, sculpture, and speedboat) (Greene et al., 2001; cf. Nakamura, 2013). Table 1 summarizes the dilemmas.

Framing tasks were adopted from Costa et al. (2014). Regarding the Asian disease problem, this study used the problem described in the introduction section. Regarding the financial crisis problem, this study used the following scenario:

> A serious financial crisis has started recently. Without any action, the company you manage will lose 600,000 euro. In order to save this money, two types of actions are possible.

In the gain condition, participants made a choice between the following two options:

> If you choose Action A, 200,000 euros will be saved.
> If you choose Action B, there is a 33.3% chance that 600,000 euros will be saved and a 66.6% chance that no money will be saved.

The loss version was identical, except that regarding Action A, “200,000 euro will be saved.” was exchanged for “400,000 euro will be lost,” and regarding Action B, “600,000 euros will be saved” was exchanged for “400,000 euro will be lost.”

Japanese versions of the moral dilemmas and framing tasks were translated from the above English versions. Regarding the moral dilemmas, participants rated the permissibility of available acts on an eight-point scale (0 = *morally impermissible*, 7 = *morally permissible*.) Regarding the framing tasks, participants chose between the risk-averse and risk-seeking options. Participants were randomly provided with one of six types of booklet to record their choices.

Results and discussion

Moral dilemmas

Figure 1 presents mean estimates of permissibility judgments for the seven moral dilemmas in the foreign and native language conditions. Permissibility judgments in the foreign language condition were higher than in the native language condition for the Switch, Footbridge, and Donor dilemmas. Multivariate t-tests between the languages indicated significant differences between conditions in mean

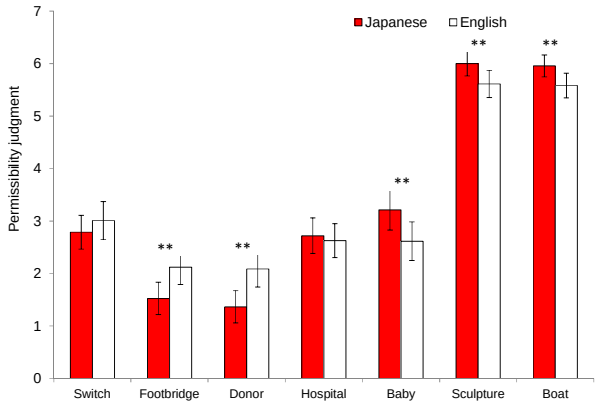


Figure 1. Permissibility judgments in moral dilemmas: error bars indicate 95% confidence interval; *: p<.05, **: p<.01

responses to the dilemmas except Switch and Hospital dilemma (p >.20). Although this study could not found significant difference in Switch dilemma, this trend was also found in Costa et al. (2014). Thus, as a whole, the foreign language effect was replicated in the same way as Costa et al (2014). Notice that directions of the effect of language on moral reasoning are the opposite in Baby, Sculpture, and Boat dilemmas. Costa et al. (2014) proposed that using foreign language would enhance engagement of the rational system. However, these results did not match the prediction by Costa et al (2014). These results suggest that direction of the foreign language effect might depend on moral dilemmas.

The following analysis was subsequently performed to determine if the foreign language effect would persist following control of individual differences. First, factor analysis with promax rotation was performed using maximum likelihood estimation. Table 2 presents eigenvalues, information criteria, and fit indexes for one-, two-, three-, and four-factor solutions. The data best supported the four-factor model; however, that model contained a factor without a significant load, and used a somewhat complex structure. In contrast, the three-factor model used a simple structure (Table 3): the first factor only significantly affected responses in the Switch, Footbridge, and Donor dilemmas; the second factor was only significant in the Baby dilemma; and the third factor was only significant in the Sculpture and Boat dilemmas. Additionally, factor loads were significant for all items. In sum, the four-factor model offered better data fit in the exploratory factor analysis; however, the pattern of factors appeared to support the three-factor model. The three-factor model was therefore adopted.

Two types of confirmatory factor analysis were subsequently performed (Table 4). One model assumed that all dilemmas were affected by only one of the three factors but that factor loads were not equal between the foreign and native languages. This model represents non-equivalence of the dilemmas' meaning between language conditions. The other model constrained values of factor loads to equality between the native and foreign language. This model represents the dilemmas' semantic equivalence between the two languages. The latter model fit the data better than the former, indicating that the moral dilemmas' meaning was equivalent in each language.In sum, the foreign language effect was replicated using moral dilemmas and a within-subjects design; this design excludes individual differences from potentially explaining the language effect; utilitarian judgment was promoted in the foreign language condition.

Table 2. Factor loads in the three-factor model

Dilemma		Factor		
		F1	F2	F3
Switch	Foreign	**0.644***	-0.043	0.069
	Native	**0.646***	-0.068	0.002
Footbridge	Foreign	**0.787***	-0.153	0.01
	Native	**0.866***	0.051	-0.053
Donar	Foreign	**0.460***	0.006	-0.059
	Native	**0.394***	0.018	-0.094
Hospital	Foreign	**0.699***	0.005	-0.012
	Native	**0.614***	0.045	0.014
Baby	Foreign	0.065	**0.636***	-0.01
	Native	-0.002	**0.893***	0.153
Sculpture	Foreign	0.008	0.043	**0.474***
	Native	-0.084	0.03	**0.445***
Boat	Foreign	-0.029	0.073	**0.643***
	Native	0.116	-0.009	**1.013***
	F1	1.000		
	F2	-0.099	1.000	
	F3	0.185	0.103	1.000

Framing tasks

Figure 2 presents results indicating the foreign language effect in the framing-effect task. Differences in risk-averse response rates between the gain- and loss-framed conditions were reduced when participants answered the framing tasks in their foreign language in Financial crisis problem Chi-square tests indicated significant differences between the

Table 3 Fit indexes of the exploratory factor analysis.

	Eigen value	AIC	BIC	Adj. BIC	CFI	RMSEA
1 factor	4.00	6954.16	7074.27	6941.44	0.60	0.15
2 factors	2.49	6831.88	6989.11	6815.21	0.83	0.11
3 factors	1.43	6794.26	**6985.87**	6773.97	0.92	0.09
4 factors	1.00	**6768.51**	6991.58	**6744.89**	**0.98**	**0.05**

Table 4. Fit indexes of the confirmatory factor analysis.

	AIC	BIC	Adj. BIC	CFI	RMSEA
No constraint	6769.633	6898.324	6756.004	0.921	0.069
Constrained	**6765.416**	**6885.528**	**6752.696**	**0.924**	**0.067**

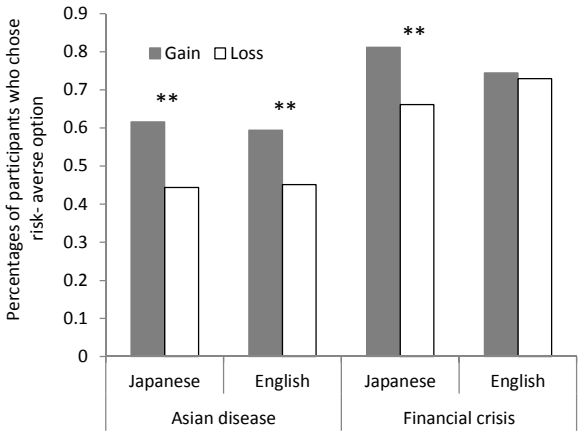


Figure 2. Risk-averse response rates in framing tasks

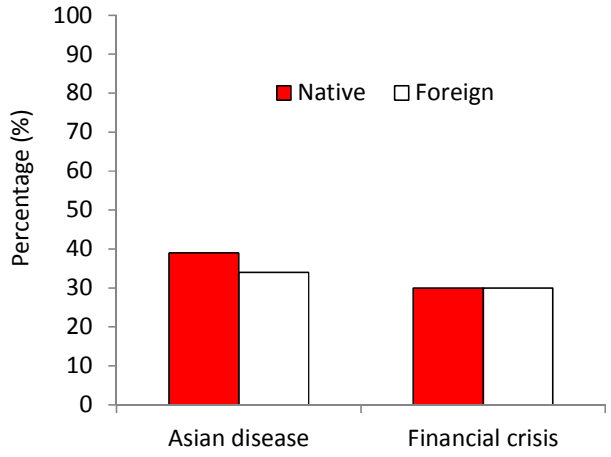


Figure 3. Response coherence in framing tasks

foreign and native language conditions in Financial crisis problem, but not in Asian disease problem. The foreign language effect was thus partly replicated using a within-subjects design.

Analysis also examined differences in coherence between the gain and loss conditions in the foreign and native language conditions (Figure 3). To accomplish this, the percentages of participants who chose the same option throughout the gain and loss conditions were calculated; no significant differences were detected in participants' percentage coherence between the foreign and native language conditions in either the Asian disease or financial crisis problem (Figure 3), indicating that using foreign language do not reduce tendency for incoherence in risky decision making .

Conclusion

The foreign language effect persists in within-subject experimental designs. Existing studies have commonly used between-subject designs, which cannot exclude differences between individuals; in contrast, this study's design excludes the possibility that the replicated foreign effect reflects differences between individuals by controlling for individual differences. Specifically, the results of this study are important because it found the difference between the native and foreign language conditions confirming the equivalence in the moral dilemmas between the two conditions. This study's results are thus more robust of those obtained using between-subjects designs.

Notice that this finding can be positioned as a first example that demonstrated the foreign language effect in its purist form. The foreign language effect indicates, lending words from Costa et al. (2014), that the way of thinking "depends on language." This statement clearly implies that use of language would affect way of thinking in the same person. However, existing studies did not examine this statement directly because of their use of between subject design. Thus, we might say that this study is the first study that showed the "true" foreign language effect.

Additionally, this study's results imply that the foreign language effect contains individual differences within participants between conditions. This implication importantly suggests that the interpretation of earlier results apparently illustrating the foreign language effect may be

seriously confounded (e.g., Costa, Foucart, Arnon, et al., 2014; Costa, Foucart, Hayakawa, et al., 2014; Keysar, Hayakawa, & An, 2012). Further, this study detected no change in participants' response coherence between the language conditions in the framing tasks. This result does not support the proposition that foreign language reduces the framing effect; instead, it suggests that the language effect may itself be dependent on other factors, such as task type or content.

Finally, this study's results indicate that the foreign language effect is unstable between decision-making tasks. The foreign language effect persisted in moral dilemmas following controlling for individual differences by using a within-subjects design; however, the effect's appearance seems to vary between within- and between-subject designs in judgment and decision making tasks, such as the framing task. Additionally, results of the framing tasks indicate that an existence of the foreign language effect depend on how to define the effect. Hence, foreign language's effect on reasoning and decision-making appears to partly depend on task type. Future research should therefore aim to determine the relationship between the foreign language effect and task type and illuminate the mechanism underlying that relationship.

References

Costa, A., Foucart, A., Arnon, I., Aparici, M., Apesteguia, J. (2014). ''Piensa'' twice: On the foreign language effect in decision making. *Cognition, 130*, 236–254.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keyser, B. (2014). Your morals depend on you language. *PLoS ONE, 9*, e94842.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105–2108.

Kahneman, D. (2011). *Thinking: fast and slow*. Farrar, Straus & Giroux

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist, 86*, 65–77

Keysar, B., Hayakawa, S. L., An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science, 23*, 661–668.

Nakamura, K. (2013). A closer look at moral dilemmas: Latent dimensions of morality and the difference between the trolley and footbridge dilemmas. *Thinking and Reasoning, 19*, 178–204.

Nakamura, K. (2015). Effect is true, but explanation is untrue: A closer investigation to the foreign language effect by Japanese participants. *Proceedings of the Thirty-seventh Annual Conference of the Cognitive Science Society*, pp.1697-1702.

Sapir, E. (1921). *Language: An Introduction to the Study of Speech*, New York: Harcourt, Brace.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645-665.

Takano, Y. (1989). Methodological problems in cross-cultural studies of linguistic relativity. *Cognition, 31*, 141–162.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Whorf, B., J. B. Carroll (Ed.) (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Massachusetts: MIT Press.

Learning and making novel predictions about others' preferences

Natalia Vélez[1], Yuan Chang Leong[1], Chelsey Pan, Jamil Zaki, & Hyowon Gweon
{nvelez, ycleong, chelspan, jzaki, gweon}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

We often make decisions on behalf of others, such as picking out gifts or making restaurant recommendations. Yet, without direct access to others' preferences, our choices on behalf of others depend on what we think they like. Across two experiments, we examined *whether* and *how accurately* people are able to infer others' preferences by observing their choices. Our results suggest that people are capable of making reasonably accurate predictions about what others will choose next, given what they have chosen before. These results lay the groundwork to systematically study how people make novel predictions about others' preferences, and when different strategies might be appropriate.

Keywords: preference learning; social cognition; Theory of Mind; decision-making

Introduction

People often make choices to please others, such as buying gifts or making restaurant recommendations. The effortlessness of these mundane, everyday decisions belies their underlying complexity. Others' preferences may differ from our own, and we do not have complete knowledge of what they like. Therefore we must base our decisions on what we *think* the other person likes or wants. These choices are easy when we can simply give others what they have chosen before; even infants can cast aside their own preferences to give others foods that they clearly like (Repacholi & Gopnik, 1997; Doan et al., 2015).

However, choosing for others is rarely so simple. For example, suppose you are recommending a movie to a friend. You might remember some movies your friend has watched and liked before, but recommending exactly those movies would hardly be useful to her. Thus, it is often insufficient, even inappropriate, to simply choose what others have liked before. Instead, you would most likely consider movies that your friend has not seen before and choose the one you think she would like best. How do people make these novel choices on other people's behalf?

In many cases, people's own preferences provide a useful template for reasoning about other people's preferences. For example, you might recommend whatever movie *you* like best, under the assumption that your friend has similar tastes. Indeed, people tend to project their own desires and beliefs to those who are perceived to be similar (Ames, 2004). This is a useful strategy for predicting others' choices, especially when we have sparse, noisy, or ambiguous information about their preferences.

However, it is not a perfect strategy; other people's preferences do not always align perfectly with our own. Observing what others have liked before is another valuable source of information. Humans are capable of drawing powerful generalizations from sparse, noisy data (see Tenenbaum et al., 2011, for a review). Even young children draw systematic inferences about others' goals, preferences, and beliefs in ways that go beyond the observable evidence (Hamlin et al., 2007; Gweon et al., 2010; Kushnir et al., 2010). Previous work has formalized this process as "inverse planning," working backwards from others' observable actions to infer the unobservable mental states that generated them (Baker et al., 2009). Thus, people might use others' past choices to generate an abstract representation of their preferences, abstracting from the specific items that others have chosen to spot qualities that they might also enjoy in novel items. This would be akin to reasoning from a few movies that your friend has liked before—such as *Love Actually, Pretty Woman,* and *Sleepless in Seattle*—that she likes romantic comedies, and that she might enjoy other romantic comedies.

The ultimate goal of the present work is to better understand *how* people generalize from observed previous choices to make predictions about what others would choose next—as a first step, we ask *whether* and *how accurately* people can do this. Experiment 1 validates key features of our approach, and tests whether people can generalize from others' choices in a simplified context. In Experiment 2, participants (henceforth "observers") were faced with a more naturalistic—and much more challenging—task: they observed choices made by a previous participant (the "target") among one set of movies, and then predicted what the other person chose among a completely different set of movies.

Experimental Task: Choosing Novel Movies

The experimental task used here was designed to mirror the everyday task of recommending a movie to a friend (Figure 1). We created posters and plot synopses of novel movies that varied along three features: valence (positive or negative), setting (historical or futuristic), and genre (romantic or action). Each of the three features was varied orthogonally to generate 8 categories of movies; 4 movies were made for each category, resulting in 32 novel movies total. All movies were normed on Amazon Mechanical Turk by an independent group of raters (n = 90; data not shown) to ensure that each movie was categorized according to its intended features. The benefit of using novel movies is

[1] These authors contributed equally to this work.

twofold: first, we mitigated the potentially complex effects of prior knowledge and familiarity, as all movies were novel to all participants; second, we imposed some structure on the features of our movies, thus simplifying the learning problem.

In both experiments, participants watched the choices made by a target among a set of 16 movies, and then predicted what the same target would choose among a novel set of 16 movies. We considered two metrics for human performance. First, we directly compared observers' predictions to targets' real choices; this served as a ground-truth indicator of accuracy. Second, we compared observers' predictions to those of a mixed multinomial logit (MML) model. Model accuracies served as a benchmark for how well observers could be expected to do, given the choices they have seen the target make.

MML models have been used extensively in economics to model consumer choices (e.g., Train, 1980); more recently, they have also been applied to describe the development of preference understanding in young children (Lucas et al., 2014). MML models assume an agent's preferences are: (1) stable over time; and (2) defined over features of objects, and thus generalizable to other objects with similar features. When choosing one option out of a set, an agent's choices are probabilistically related to the utility, or attractiveness, of each option (Luce, 1977). Each option i is represented as a binary vector of its features (x_i) and a vector of weights (β) corresponding to the agent's preferences for individual features. The utility of option i (u_i) is the weighted sum of its features ($\beta \cdot x_i$), scaled by a free parameter T that describes the stochasticity of the agent's choices. (In all cases, T was a free parameter fit with a regularizing prior and the models were fit to maximize the maximum a posteriori estimate.) Taken together, the probability of an individual choosing option i from a pair of options J is defined as:

$$P(c_i \mid X, \beta) = \frac{\exp(\beta \cdot x_i / T)}{\sum_j \exp(\beta \cdot x_j / T)}$$

In the experiments below, the MML model served three important functions. First, the model was used to describe targets' own preferences: in Experiment 1a, we trained the model on one half of participants' own choices to extract the preference weights (β), and then used these learned weights to predict the second half of participants' responses. The model should predict targets' choices to the extent that participants' responses are reliable and that their preferences align well with the features we imposed on the stimuli,

Second, the model was used to evaluate *participants'* inferences about the target's preferences. In Experiment 1b, the model was trained on the first half of the target agent's choices—the same choices that human participants observed—and then tested on the participants' prediction about the target agents' choices among novel options. The model's accuracy reflects whether participants learned the feature weights that describe the other person's choices, and to what extent they used these weights to predict what the other person would choose among a set of novel options.

Finally, the model was used as a descriptive tool to capture participants' strategies when choosing for others. Most notably, in Experiment 2, we extended the model to not only capture participants' inferences about the preferences of a target agent, but also how participants' observations interact with their own preferences when making novel predictions about the target's choices. Critically, throughout the paper, we used the model as a benchmark for human performance, rather than as a formal, computational characterization of the inferential processes involved in preference learning and generalization.

Experiment 1

Experiment 1, we tested two key assumptions of our approach. In Exp. 1a, we asked whether people's preferences are reliable and well defined over the features built into our stimuli; here, participants chose whichever movies they liked best, and we used the MML model to predict their choices. In contrast, in Exp. 1b, we were interested in whether people are able to learn about the preferences of others after observing a series of choices. Here, we simplified the learning problem by having participants learn about a simulated agent whose preferences can be perfectly described in our feature space, and who deterministically chooses the option with highest utility. Taken together, Exp. 1 serves to validate our overall approach and pilot a behavioral paradigm that can be used to study preference learning under noisier, richer conditions.

Experiment 1a: Choosing for self

Participants: 40 adults participated in an online behavioral experiment for pay through Amazon Mechanical Turk. All participants in this and subsequent experiments had U.S. IP addresses and provided informed consent in accordance with the IRB at Stanford University.

Procedure: The study was split into two rounds; in each round, participants saw half (16) of the 32 novel movies. Each round was composed of two tasks: Meet the Movies and Choose for Self (Figure 1a). In Meet the Movies, participants were shown the title, poster, and synopsis of each movie (Figure 1b). To ensure that participants were attending to and forming preferences for the movies, they rated how much they would like to watch each movie using a Likert Scale (1 – Not at all; 7 – Very much). In Choose for Self, participants were shown pairs of movie posters that they had just "met" and were asked to choose which of the two they would rather watch. There were 56 trials in this task (i.e., 56 pairs chosen from 16 movies), spanning all possible permutations of non-identical conditions.

Experiments 1b: Choosing for other

Participants: 50 participants were recruited for an online behavioral experiment through Amazon Mechanical Turk.

Procedure: Participants were told that they would observe and predict the choices of a target agent who had previously participated in Experiment 1a. Unbeknownst to the participants, the responses were generated by simulating the responses of an agent with pre-defined weights who deterministically chose the option with the highest utility.

As before, the study was split into two rounds, and each round began with Meet the Movies (Figure 1a). In the first round, participants observed the target's choices (Observe Other; Figure 1b). Participants were shown pairs of movie posters, and a border appeared around the movie that the target had chosen. To ensure that participants attended to the task, we asked them to imitate the target's choice by selecting the highlighted movie. Instead of using all 56 possible pairs, we excluded 8 pairs in which the artificial agent would be indifferent between the two movies; thus, participants saw 48 choices total.

In the second round (Choose for Other; Figure 1b), participants were again shown pairs of movies, but were instead asked to select the movie that they believed the target had chosen. Participants were not given trial-by-trial feedback, but they were informed that they would receive a bonus based on the number of correct responses in this task.

Note that the MML model had perfect information about each movie's features, while human observers had no prior knowledge of the movies. Even though participants had a chance to "meet" the movies, we reasoned that this brief pre-exposure would be insufficient to eliminate their uncertainty about the dimensions of the feature space as well as the uncertainty about each movie's features. Thus, in Experiment 1b, we provided keywords for each movie (e.g., "positive, historical, romantic") during all tasks, making explicit the features of each movie. This ensured that the task for our human participants was comparable to the task imposed on our model, making the performance comparison more meaningful.

Results and Discussion

Experiment 1a: The model was trained on participants' responses in the first round and tested in the second, and vice versa. Our measure of model accuracy (henceforth cross-validation accuracy) is the model's average accuracy in predicting participants' choices in each iteration; overall, the model accurately predicted participants' choices in the test set (Cross-validation accuracy: M(SD) = 0.54(0.11), tested against 0.50: $t(39) = 2.34$, $p = 0.02$; Fig. 2a).

Experiment 1b: Participants' accuracy was near ceiling in the Observe Other task (M(SD) = 0.98(0.08)), suggesting that participants were alert and attentive during the task. Impressively, people showed fairly high accuracy even in the Choose for Other task, where participants had to use their previous observations to predict the target's responses among a set of new movies, (M(SD) = 0.77(0.22), test against 0.50: $t(49) = 24.27$, $p < 0.001$; Fig. 2b).

Overall, our model performed reasonably well at capturing people's own preferences, suggesting that participants themselves have stable preferences that can be inferred and predicted by our model. Importantly, participants were never told about the three dimensions or the features of each movie in Experiment 1a—nevertheless, people's preferences were well described by the model, suggesting that people's preferences align reasonably well with the feature space we have imposed on the stimuli.

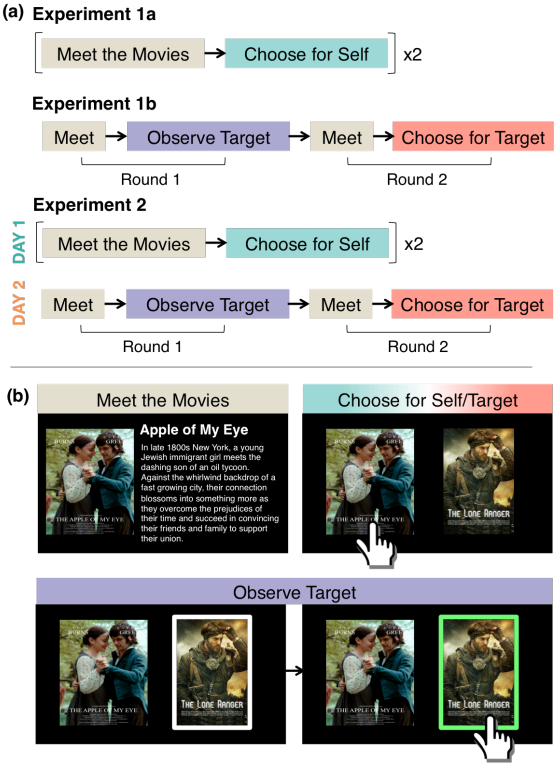


Figure 1: (a) Task order: In Experiment 1a, participants only chose for themselves; in Experiment 1b, participants only chose on behalf of another agent. Experiment 2 combines aspects of both of these into a multi-session experiment. (b) Schematic of tasks and example stimuli.

However, we note that the model is much worse at predicting real people's preferences (Experiment 1a) than people are at predicting an artificial target agent's preferences (Experiment 1b). These results serve complementary functions. On one hand, Exp. 1b provides an approximate upper bound for human performance, in the extreme case where the choice data provided are as clear and consistent as possible. Overall, we find that observers perform admirably when they are provided with good evidence. By contrast, model performance in Exp. 1a provides an estimate of the quality of the evidence available to observers when learning from real people's choices. These results suggest that real people vary wildly in the extent to which their choices are consistent and aligned with the features built into the stimuli.

In Experiment 2, we aimed for a stronger test of people's ability to learn and generalize about other's preferences: having them observe a real person's choices and asking them to predict their real choices. This is a much more

challenging task than learning from an idealized, deterministic target.. However, given that people perform very well when given very clear information, we would still expect people to be able to learn and generalize from real person's choices, at least to the extent that targets' choices are consistent and informative.

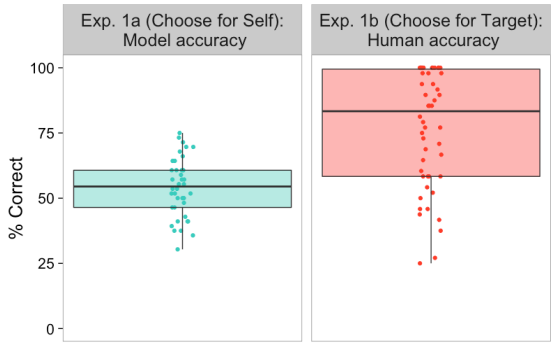


Figure 2: Experiment 1 results. Left: Model cross-validation accuracy in Exp. 1a, indicating how well the model predicted each participants' choices among the test set, based on weights inferred from participants' choices among the training set. Right: Human performance in Choose for Target.

Experiment 2

As described above, Experiment 2 extended the previous experiments by having participants choose for themselves as well as for a target—here, the target was a real participant who participated earlier in the same experiment. This not only allowed us to test people's ability to learn and generalize from noisy, messy choices of real participants, but also allowed us to examine the degree to which participants' own preferences biased their choices on behalf of the target.

Procedure: 51 participants were recruited from the Stanford community for a two-day experiment. On day 1, participants completed the Meet the Movies and Choose for Self tasks, matching the procedure used in Experiment 1a. On day 2, participants first observed a target's choices (Observe Target), then predicted the target's choices among a different set of movies (Choose for Target). This session closely matched the procedure in Experiment 1b, with the critical difference that participants' responses were yoked. That is, participant A's responses during the first round of Choose for Self were presented to participant B during Observe Target, and participant B's predictions in Choose for Target were tested against the actual choices that participant A made during the second round of Choose for Self. In sum, each participant observed and predicted the responses of another participant who had come before. Data from 2 participants were discarded because they did not return for the second day of the experiment.

Results and Discussion

First, the model was trained on participants' binary choices in one round of Choose for Self and tested on the remaining round; as in Experiment 1, the model described participants' own preferences reasonably well (Cross-validation accuracy: M(SD) = 0.61(0.12), test against 0.50: $t(48) = 6.4$, $p < 0.001$). If model accuracy is taken as a proxy for how consistent targets' choices were and how closely they aligned to our stimulus features, then these results suggest that some targets' choices were more informative than others, but that it is possible, on the whole, to generalize from the choices of targets in our sample.

Accordingly, participants' predictions about the target's choices on a novel set of movies also matched the target's actual choices reasonably well (M(SD) = 0.57(0.11), test against 0.50: $t(48) = 4.59$, $p < 0.001$, Fig. 3a) — despite the sparseness and noisiness of the target's choices in both halves of the experiment. Impressively, participants' performance was comparable to that of the model predicting the target's choices based on the same observations (Model performance: M(SD) = 0.59(0.13), paired t-test: $t(48) = -0.92$, $p = 0.361$). When the predictions of the MML model were compared to participants' predictions, we found a reasonable correspondence (M(SD) = 0.63, test against 0.50: $t(48) = 7.1$, $p < 0.001$, Fig. 3a). Most importantly, we found a correlation between human and model accuracy in predicting the target's choices ($r = 0.36$, $p = 0.01$; Fig 3b). Thus, even though 57% might not seem much higher than chance, our results suggest that observers seized the underlying structure in targets' choices when possible.

So far we have considered a completely allocentric version of the MML model, which infers the targets' preference weights from their previous choices and makes predictions based solely on these inferred weights. However, an alternative possibility is that one could make predictions by simply projecting one's own preferences while completely ignoring others' preferences, or that one could use a mix of these two strategies. In an exploratory analysis, we asked whether participants' predictions about the choices of others were biased by their own preferences. We extended our MML model to make predictions based on a weighted average of self and learned target preferences, with the weight determined by free parameter η. Here, the model with $\eta=0$ is completely allocentric (i.e., made solely from others' preferences), whereas the model with $\eta=1$ makes completely egocentric predictions (i.e., made solely from the participant's own preferences).

We examined the distribution of best-fit values of η in our sample (Fig. 4). The distribution of best-fit values of η across participants peaks at $\eta = 0$, suggesting that many of the participants relied on their past observations of the target's choices to make their predictions. There is also a second, smaller peak at $\eta = 1$, indicating that there is a second group of participants who relied solely on their own preferences when making decisions for others. However, there is a wide range of η values in our sample, suggesting that some participants combined self and other preferences to some degree when making their choices. Using a leave-one-trial-out cross-validation procedure, we compared the performance of the combined model to a model that made choices based only on target weights (other model) or only

on observer weights (self model). Notably, for 7 out of the 49 participants, the combined model predicts participants' choices better than both the self and other model alone, suggesting that these participants were using both self and other preferences when making their choices.

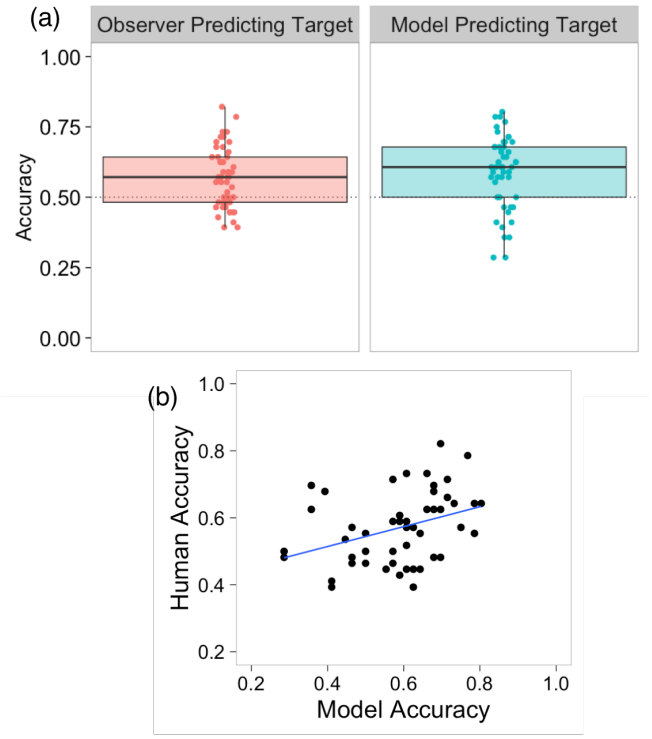


Figure 3: Experiment 2 results. (a) Human (left) and model (right) right accuracy in Choose for Target. The model shown here was trained solely on the target's choices, and thus corresponds to a totally allocentric observer. (b) Correlation between human and model accuracies.

Overall, we find initial evidence that people can learn impressively well from extremely noisy, sparse data to derive reasonable predictions about others' choices, and that their accuracy is commensurate with the quality of the information gleaned from others' past choices. We also find hints that people are not completely free of biases coming from their own preferences.

We note that there are two limitations of the paradigm in its current form, which will be addressed in future work. First, the interpretation of η is confounded when the observer and target have similar preferences. In such cases, the model would predict choice equally well when using self-preferences and when using other-preferences. The best-fit value of η would then depend on small differences in choice histories and are unlikely to reflect differences in participants' strategies. Second, though many targets' choices align closely with our stimulus dimensions and are consistent from one set of movies to the other, other targets' choices were not at all informative, and in these cases our observers were set up to fail. While we randomly paired each observer to a target in the current work, future iterations of this paradigm will screen participants with consistent, learnable preferences, and yoke participants to targets whose preferences are orthogonal to their own.

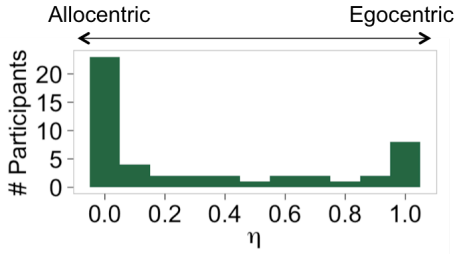


Figure 4: Distribution of η. Smaller η indicate that participants chose allocentrically (i.e., based on the target's choices), and higher η suggest that participants chose egocentrically (i.e., based on their own preferences).

General Discussion

Across two experiments, we found that participants can generalize from the choices a target has made before to accurately predict what the target will choose next. In Experiment 1a, we first verified that the model captures people's own preferences, suggesting that people have stable preferences that can be described by the dimensions we have imposed on the stimuli. Experiment 1b suggested that participants can accurately predict others' choices in an idealized scenario, where they are observing the choices of an artificial target whose actions are deterministic and perfectly aligned with predetermined features of the stimuli.

Experiment 2 provided the main testing ground for our hypothesis: participants not only indicated their own preferences, but also grappled with the much harder task of learning from and predicting the noisy and often inconsistent choices of a real human participant. Despite the task being more challenging, we again found that participants made reasonably accurate predictions. Further, observers' accuracy was commensurate with the consistency of the target's choices, as measured by the accuracy of a simple MML model. In an exploratory analysis, we used this model to probe how strongly participants relied on their own preferences or on previous observations of the target's choices in order to choose on their behalf.

Interestingly, we found that participants' accuracy in making predictions about targets correlated with the accuracy of the MML model. Since the performance of the MML model is an index of consistency of a target's preferences from one set of movies to the next, this result suggests that participants are better able to predict the target if the target provides consistent data for participants to learn about. What happens when the data provided are noisy? One possibility is that participants switch to an egocentric strategy of projecting their own preferences onto the target. Here, we demonstrate that participants use a mixture of egocentric and allocentric strategies; future work needs to be done to examine the factors determining how participants choose their strategy.

When we observe others' choices, we make inferences about the hidden, internal states—such as preferences—that

motivated the choice. Using MML models that have been used in economics and cognitive science (Train, 1980; Lucas et al., 2009), we operationalize this process as learning the weights the other person attaches to underlying features that define the choice. While the model and our feature space are indeed too limited to capture the full richness and complexity of how people represent and learn about preferences, we believe that this simplified space is a good starting point for this investigation.

Given that we have found that people *can* learn and generalize from others' preferences, the natural next step is to explore *how* this is done. In the current work, we made the simplifying assumption that our stimuli varied along three binary dimensions; however, the feature space of people's actual movie preferences is much larger. In fact, for the same type of items, different people could use different sets of features to guide their choices—for example, one person might pay attention to a movie's reviews before choosing to watch it, while another person might decide which movies to watch based on the cast. As such, before learning which features another person values, observers have to first infer what features to learn about. This is a non-trivial problem, and it is similar to the structure learning problem explored in other domains of cognitive science (Gershman & Niv, 2010).

Another exciting extension of this work is to examine the degree to which social closeness affects people's predictions of others' choices. If participants systematically overestimate that the people closest to them are also more similar to them (Savitsky et al., 2011), then they might choose more egocentrically for a friend than for a stranger. However, closeness could very well play the opposite role: because people have had more opportunities to observe the choices of their close friends, they may choose more accurately for their friends than for a stranger. This direction also converges with prior neuroimaging work, which suggests that watching other people receive rewards engages neural systems involved in reward, and that this vicarious reward signal is influenced by perceived similarity between the observer and the target (Mobbs et al., 2009). However, little empirical work has directly tested the neural computations involved in making novel choices for others based on abstract, generalizable preferences, or how social closeness might modulate neural responses associated with preference learning.

The current work explored the deceptively mundane problem of predicting what others will like next based on what they have liked before. This is no small feat — other people's preferences may (or may not) differ substantially from our own, and their choices provide only a sparse and noisy reflection of their preferences. Motivated by both classical theories on egocentric biases (Ross et al., 1976) and more recent computational approaches to understand human learning as rational inductive inferences from sparse, noisy data (Tenenbaum et al., 2011), our experiments provide important empirical groundwork to better understand how this feat might be accomplished.

Acknowledgements

This work was supported by the Varieties of Understanding grant from the John Templeton foundation to HG and an NSF Graduate Research Fellowship to NV. We also thank our reviewers for their invaluable comments and feedback.

References

Ames, D. R. (2004). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. *Journal of personality and social psychology*, *87*(3), 340.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of personality and social psychology*, *87*(3), 327.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. Current opinion in neurobiology, 20(2), 251-256.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*,*21*(8), 1134-1140.

Mobbs, D., Yu, R., Meyer, M., Passamonti, L., Seymour, B., Calder, A. J., ... & Dalgleish, T. (2009). A key role for similarity in vicarious reward. *Science*,*324*(5929).

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, *33*(, 12.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology*, *47*(1), 269-273.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Train, K. (1980). A structured logit model of auto ownership and model choice. *The Review of Economic Studies*, *47*(2), 357-370.

Monitoring the Level of Attention by Posture Measurement and EEG

Ryohei Furutani (s1421604@u.tsukuba.ac.jp) **Yuki Seino (seino@ce.slis.tsukuba.ac.jp)**
Graduate School of Library, Information and Media Studies, University of Tsukuba, 305-8577 Japan

Taro Tezuka (tezuka@slis.tsukuba.ac.jp) **Tetsuji Satoh (satoh@ce.slis.tsukuba.ac.jp)**
Faculty of Library, Information and Media Science, University of Tsukuba, 305-8577 Japan

Abstract

Attention is a factor that affects the performance of various intelligent activities in humans. Up until now, the methods for measuring the level of attention have been mostly based on subjective reports or employing large and costly devices. In this paper, a new method of estimating the level of attention is proposed, based on posture and EEG measurements. These data can be recorded using easily available and less burdensome devices. From the obtained data, the time evolution of attention was explored. Experiments showed that there is negative correlation between posture variance and attention, and also between EEG and attention.

Keywords: attention, posture measurement, EEG

Introduction

Attention is a crucial factor that affects the performance of various intelligent activities in humans. Measuring the level of attention is of a great importance for optimizing the performance. Unless the level is estimated properly, it would not be able to control it. However, most of the ways to monitor the level of attention have been based on participants' reports (Rich & Gureckis, 2015). Although there are methods that uses objective measurement, they required high-cost devices such as fMRI or invasive physiological recordings (Itti & Koch, 2001). In this paper we propose a method of estimating the level of attention using easily available and less burdensome devices. Devices that we used are pressure sensors and a mobile EEG (electroencephalogram) sensor.

Researchers have revealed that there are various types of attention, from a subjective and conscious state of mind to subconscious alert states that can only be measured through psychological tests. In this work, we focus on a subjective and conscious level of attention, which humans can recognize and evaluate by themselves. The goal of this work is to find features in posture and EEG measurement that correlates well with this type of attention.

There are numerous applications realized by monitoring the attentional level. Here we name a few examples in the fields of education, safety management, and cognitive science.

Education: Knowing the level of attention among students is valuable for the lecturer since he can use such feedback for making the lecture more attractive. This is even more important in the case of e-learning, where the lecturer cannot easily see the response of the audience.

Safety management: When an employee has low attention, the monitoring system can warn him or advice the manager for rotation. It is of critical importance since many disastrous incidents were caused by a low level of attention of the worker.

Cognitive science: Monitoring the level of attention is also of interest for cognitive science (Itti & Koch, 2001; Srivastava & Vul, 2016). For example, knowing the time evolution of attention may give a clue to how attention is maintained or generated in the brain.

In these applications, it is preferable to implement a system using easily available devices, in order to make it accessible to a large group of users. The method explored in this work is aimed at realizing such applications.

The underlying assumption of this work is that the posture and EEG of a participant change in correlation to the level of attention. A reason for using posture is that it is a quantity that can be relatively easily measured. Unlike many other physiological quantities, it can also be measured non-invasively. Our preliminary observation showed that the posture of a sitting-down students reflects his/her attention toward the lecture. A focused participant sits fixedly, leaning rather forward. On the other hand, a non-focused participant is likely to change their posture more frequently, often leaning backward. Based on this observation, we propose a hypothesis that the movement of posture reflects the level of attention. This does not mean that a specific posture corresponds to a high or low level of attention. Rather, our model is that indicators such as average or variance of a posture related value correlates with the level of attention. In addition, EEG is known to have correlation with performance in many types of tests (Klimesch, Doppelmayr, Russegger, & Schwaiger, 1998). This indicates that EEG can perform as an indicator for the level of attention. We therefore seek specific indicators in the posture and EEG that are correlated to the level of attention. These indicators are computed from raw signals obtained from sensors.

Related work

Measuring attention

It has been pointed out that attention is limited in time (Nobre & Coull, 2010). Humans cannot sustain a high level of attention for a long period of time. In practice, it is important to know how the level of attention changes over time, and how it can be recovered. This would help making humans more productive. One goal of our work is to uncover such dynamics

of attention.

Itti and Koch proposed a model of visual attention that predicts where the participant focuses at (Itti & Koch, 2001). The model was evaluated on the basis of eye movement and object recognition. Srivastava and Vul proposed a new model for the dynamics of attention when a participant is tracking multiple objects displayed on a screen (Srivastava & Vul, 2016). The level of attention was estimated based on how well the participant can judge whether the color of the moving object changed or not, despite many other distracting objects. Walsh et al. evaluated a simulation model for the level of vigilance using the response time of a participant in certain tasks (Walsh, Gunzelmann, & Dongen, 2014).

Sensor measurement

Sensors similar to the ones used in our work have been proposed for estimating the state of a participant while conducting a certain task. Shen et al. used EEG, blood pressure sensor, and perspiration sensor to monitor participants during e-learning (Shen, Wang, & Shen, 2009). Kanki et al. proposed a system for controlling a wheelchair using an array of pressure sensors embedded to its seat (Kanki, Kuwahara, & Morimoto, 2014). The user could direct the wheelchair by moving his center of gravity.

Method

Participants

We tested the system with 17 participants. They were all graduate and undergraduate students in their 20s. Eleven were male and six were female.

Apparatus and stimuli

In our system, the participant's posture is measured using pressure sensors placed within the chair that the participant is sitting on. Using these sensors, the center of gravity of the participant's upper body is tracked. Hereafter, "the center of gravity" refers to that of the upper body of the participant, ignoring the weight supported by their feet. In addition, EEG is measured using a mobile recording device that the participant wears on his/her head. The system and its constituent parts are described in detail in the rest of this section.

Balance board To measure the posture of the participant based on the movement of the center of gravity, we used Nintendo Wii-U balance board[1], illustrated in Figure 1. The board has pressure sensors embedded at four corners, allowing to measure the weight imposed at each corner. Let A, B, C, and D represent four corners of the board, that is, front-left, front-right, back-left, back-right, respectively, as indicated in Figure 2. Let $w_a(t)$ indicate the weight at corner a at time t, where a is for A, B, C, and D. Let $w_\Sigma(t)$ represent the total weight imposed on the board at time t, and s_x and s_y be the width and the depth of the board, respectively. Using the weights at four corners, the horizontal coordinates for the center of gravity $(x(t), y(t))$ can be obtained as follows.

[1] http://www.nintendo.com/wiiu



Figure 1: A chair with a balance board

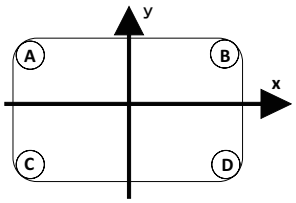


Figure 2: Corners of a balance board

$$w_x(t) = w_B(t) + w_D(t) - w_A(t) - w_C(t) \quad (1)$$

$$w_y(t) = w_A(t) + w_B(t) - w_C(t) - w_D(t) \quad (2)$$

$$w_\Sigma(t) = w_A(t) + w_B(t) + w_C(t) + w_D(t) \quad (3)$$

$$x(t) = \frac{s_x}{2}\frac{w_x(t)}{w_\Sigma(t)} \quad (4)$$

$$y(t) = \frac{s_y}{2}\frac{w_y(t)}{w_\Sigma(t)} \quad (5)$$

The origin of this coordinate system is at the center of the balance board. In other words, $(x(t), y(t)) = (0,0)$ means that at time t the participant's center of gravity has the same horizontal coordinates as the center of the balance board. The size of the balance board used in our experiment was 50 cm by 30 cm, so s_x = 50 cm and s_y = 30 cm. The values of $w_A(t)$, $w_B(t)$, $w_C(t)$ and $w_D(t)$ are recorded from the balance board at 50 Hz, i.e. every 20 milliseconds.

The balance board was placed on a chair whose height was 45.5 cm. The board itself has 5.3 cm height. In total, the height was 50.8 cm, which is close to chairs used in classrooms. Since participants were all graduate and undergraduate students, we considered this total height to be apt. However, the chair had no backrest, which could have made the participant feel less comfortable compared to a regular classroom environment. The height of the table on which the tasks were conducted on was 71.6 cm, which is a little lower than a desk used in a regular classroom.

EEG recording EEG (electroencephalogram) was recorded using Neurosky Mindwave Mobile[2], indicated in Figure 3. Following the recommended usage of the device, electrodes were placed at three sites on the participant's head and forehead, namely Fp1 (frontal pole 1), Fp2 (frontal pole 2), and A1 (auricular 1) recording sites specified by the international 10-20 system. If the participant moves by a large extent during experiments, it can cause some artifacts to EEG recordings. Since such large movements were expected to be prominent for the Reversi task, we placed the game board close to the participant, such that he/she does not need to move his/her arm much.

The power of each frequency band is sampled at 1 Hz. Although various frequency bands are recordable using this de-

[2] http://neurosky.com/



Figure 3: EEG sensing device



Figure 4: Experiment environment

vice, we focused on four bands, namely low α, high α, low β, and high β. Their power at time t is indicated by $\alpha_L(t)$, $\alpha_H(t)$, $\beta_L(t)$, and $\beta_H(t)$, respectively. These are suggested to be most relevant to the level of attention (Klimesch et al., 1998; Yasui, Tian, & Yamauchi, 2008; Yoshida, Sakamoto, Miyaji, & Yamada, 2012). Table 1 shows the frequency for each of these bands, and variables that represent their power.

Table 1: Types of EEG frequency bands

Band	Var.	Frequency	Band	Var.	Frequency
low α	α_L	8∼10 Hz	low β	β_L	12∼20 Hz
high α	α_H	10∼12 Hz	high β	β_H	20∼30 Hz

For estimating the level of attention from the posture of the participant, we propose to use two indicators, namely variances of $x(t)$ and $y(t)$ over time interval I_k. They are based on a hypothesis that when the participant is less focused to the task, they feel more body pain and change the posture more frequently, thus increasing the variance of the center of gravity. For EEG based monitoring, we propose to use the following two variables W and Z that can be calculated from the power of these frequency bands. In the following definitions of $W(t)$ and $Z(t)$, $\alpha_\Sigma(t)$ indicates the total α power and $\beta_\Sigma(t)$ indicates the total β power at time t.

$$\alpha_\Sigma(t) = \alpha_H(t) + \alpha_L(t) \quad (6)$$

$$\beta_\Sigma(t) = \beta_H(t) + \beta_L(t) \quad (7)$$

$$W(t) = \log\left(\frac{\beta_\Sigma(t)}{\alpha_\Sigma(t)}\right) \quad (8)$$

$$Z(t) = \log\left(\frac{\alpha_L(t)}{\alpha_H(t)}\right) \quad (9)$$

W is based on an existing report which states that the strength of the β band is relevant to attention (Yoshida et al., 2012). Z is our own proposal, based on a hypothesis that the composition of the α band is relevant to the level of attention.

Table 2 summarizes the indicators proposed in this paper.

Conditions

Table 3 indicates three types of tasks conducted for each participant. All of the problems were multiple choice problems and were presented by printed documents.

Table 2: Indicators used in the experiments

Indicator	Description
var(x)	The variance of $x(t)$ for all $t \in I_k$.
var(y)	The variance of $y(t)$ for all $t \in I_k$.
mean(W)	The mean of $W(t)$ for all $t \in I_k$.
mean(Z)	The mean of $Z(t)$ for all $t \in I_k$.

Procedure

Figure 4 shows the environment for the experiment. Each participant was asked to sit on top of a balance board placed over a chair. After the outline of the experiment was explained, the participant wears the EEG recording device. The experiment took about one hour for each participant. The participants were divided into two groups. For each group, the three tasks were conducted in the following order.

Group 1 (9 participants): T1 → T3 → T2
Group 2 (8 participants): T2 → T3 → T1

This is to reduce the effect of the participant's level of attention lowering from getting tired from performing many tasks.

We also recorded during resting periods, but since we did not ask the participants to stay still, some of them adjusted the EEG recording device that they were wearing. We are therefore not using data recorded during the resting periods for analysis.

In the Reversi match task, the participants had to play the game under the following constraints.

- For the first three moves, the participant has to wait 10 seconds before making each move.
- After half of all the squares are filled (i.e. 32 squares are filled) with pieces, the sides are abruptly switched, that is, the light side player now plays the dark side and vice versa. This is not told to the participant beforehand.
- Three moves after the switching of the sides, the participant must make each move within 15 seconds after the opponent has made a move.

After these tasks, the participant was requested to fill in a questionnaire after finishing all of the tasks. The questionnaire asks for a subjective description of how his/her level of attention changed as he/she conducted the tasks, scored by an integer from 1 and 5.

Analysis

In the questionnaire, the participants were asked to answer their subjective level of attention by five levels, from "strongly not attentive (1)", "not attentive (2)", "neither attentive nor not attentive (3)", "attentive (4)", and "strongly attentive (5)". Hereafter, we call each of the level as a class.

Table 3: Types of tasks

ID	Task	Description	Time intervals
T1	CAB calculation problems	10 minutes time limit. 27 mental computation problems, 18 arithmetic calculation problems, and 18 geometric problems.	$I_1 \sim I_3$
T2	GAB word problems	10 minutes time limit. 24 problems from 8 documents.	$I_4 \sim I_6$
T3	Reversi match	1 game, which takes about 15 minutes.	I_7, I_8

Tasks T1 and T2 were divided into three equal-sized intervals, and T3 was divided into two equal-sized intervals. These eight time intervals are represented by $I_1, I_2, ..., I_8$, which are defined in Table 3. Each participant had to give five-level grading for his/her level of attention during each of the time intervals. As a result, an 8-dimensional vector $C^{(i)}$ is obtained for each participant i. Its k-th component $C_k^{(i)}$ is the subjective level of attention at interval I_k for participant i , where $k \in (1, 2, ..., 8)$

Let X represent any of four indicators, i.e. var(x), var(y), mean(W), and mean(Z). For X of participant i, we get an 8-dimensional vector $\overline{X}^{(i)}$, whose k-th component $\overline{X}_k^{(i)}$ is obtained by time averaging $X^{(i)}(t)$ over time intervals I_k. $|I_k|$ represents the number of sampling time points in I_k.

$$\overline{X}_k^{(i)} = \frac{1}{|I_k|} \sum_{t \in I_k} X^{(i)}(t) \tag{10}$$

We evaluated the indicators based on how well they correlate with the subjective level of attention. For example, Pearson's correlation coefficient between indicator $\overline{X}^{(i)}$ and the subjective level of attention $C^{(i)}$ is computed for each participant i as follows.

$$R_{(X)}^{(i)} = R(\overline{X}^{(i)}, C^{(i)}) \tag{11}$$

When there are N participants, the average of Pearson's correlation coefficient for indicator X is computed as follows.

$$R_{(X)} = \frac{1}{N} \sum_{i=1}^{N} R_{(X)}^{(i)} \tag{12}$$

Results

Questionnaire

Table 4 summarizes the answers of the participants to the questionnaire. Each row represents a participant, and a column is a time intervals defined in Table 3. Figure 5 shows the distribution of the answers. A box represent the 25th and 75th percentiles, and a red plus is an outlier. The plot shows that for the CAB calculation problem task and the GAB word problem task, there is a tendency that reported levels of attention decreases as the time passes.

Table 4: Participants' answers to the questionnaire

	T1			T2			T3	
participant	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
P1	5	2	2	5	2	2	1	2
P2	4	5	4	5	4	5	1	4
P3	4	3	2	3	1	1	1	2
P4	4	3	3	5	5	5	5	5
P5	5	5	4	5	3	3	5	5
P6	1	2	2	4	4	4	1	4
P7	3	4	5	2	2	3	5	5
P8	2	2	2	4	4	4	1	1
P9	5	4	2	5	5	4	1	4
P10	4	4	5	3	4	5	5	5
P11	5	5	4	5	5	3	5	4
P12	4	5	4	5	4	4	5	5
P13	5	4	4	4	5	4	5	2
P14	4	5	1	2	4	1	5	2
P15	4	3	4	4	4	3	5	5
P16	4	4	2	5	4	4	1	5
P17	5	4	4	5	4	4	5	4

Dynamics

Figures 6, 7, and 8 indicate the center of gravity for 5 different participants (Participants $1 \sim 5$) at different time points, sampled at 1 Hz, during three different types of tasks (T1$\sim$T3). Participants can be distinguished by marker types. Note that the x-axis is more stretched than the y-axis. The plot shows that the distributions of the center of gravity vary among participants. Some participants have wider distribution, indicating they tend to move more during the task, while others are more fixed, having smaller variance. It can also be seen that for the Reversi match task (T3), the participants are more likely to change their center of gravity.

Figures 9, 10, 11, and 12 show the time evolution of the indicators for Participants $1 \sim 4$ during the CAB calculation problem task. Time is indicated in seconds.

Correlation

Table 5 summarizes the correlation coefficients between the subjective level of attention and different indicators. For var(x), may participants show negative correlation with the subjective level of attention, suggesting that when the participant is less attentive, he/she tends to change his/her posture either more frequently or widely to the x direction (sideways), thus increasing the variance. For the column for var(x) and

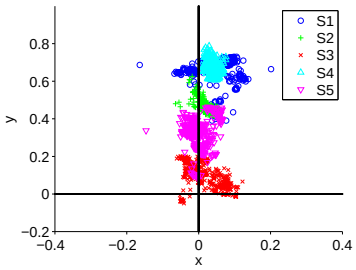


Figure 6: The center of gravity for the calculation task

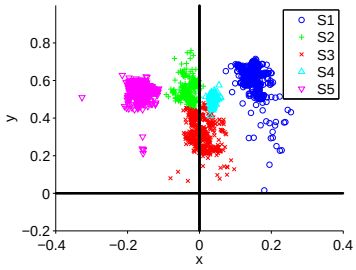


Figure 7: The center of gravity for the word problem task

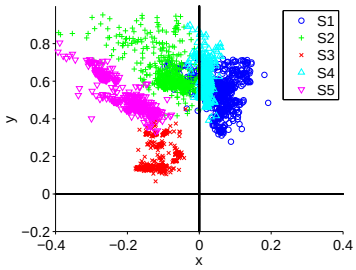


Figure 8: The center of gravity for the Reversi match task

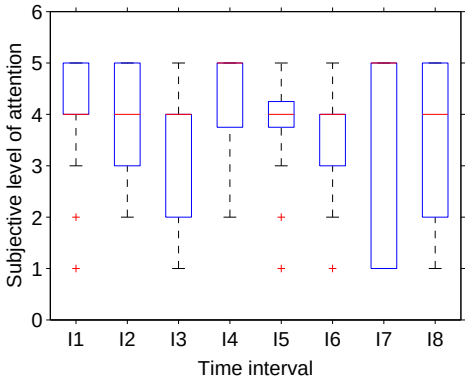


Figure 5: Subjective levels of attention by time intervals

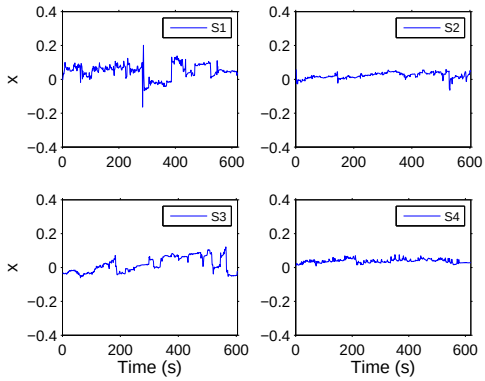


Figure 9: Time evolution of $x(t)$

var(y), cells with negative correlation are emphasized. On the other hand, mean(W) and mean(Z) had positive correlation with the subjective level of attention for many participants, but the difference is not significant. For these columns, cells with positive correlation are emphasized. Since EEG requires intensive preprocessing such as filtering. We are expecting that adding such preprocessing may improve the result. Although the correlation coefficients were of moderate quantity at this moment, we expect that combining these indicators would enable good estimation of the level of attention. Such extension would require determining optimal weight coefficients for combining these indicators. This is left to future work.

Conclusion

A method of estimating the level of attention using posture measurement and EEG was proposed. The posture and EEG of the participants were successfully measured using easily available and less burdensome devices.

The measured attentional level was based on retrospective reports from the participants, which could be inaccurate. We would like to introduce more objective measures of attention which could be measured in real-time. In addition, we plan to measure the lower body of the participants, especially leg positions, using an infrared sensor placed under the table.

In future work, we would like to develop an estimation method for the level of attention by combining the indicators discussed in this paper. Machine learning algorithms can be used to combine the indicators such that it estimates the actual level of attention well. We also plan to search for specific patterns that would appear when the level of attention changes. For example, in Figures 9 and 10, Participant 1 shows abrupt change of $x(t)$ at around 300 seconds (5 minutes) and that of $y(t)$ at around 550 seconds (9 minutes) after the start of the task. This could be due to some disruption of attention. Rather than looking at the average value of the indicator over a time interval ranging a few hundreds of seconds, we plan to look at local features that indicate some change in the level of attention.

Acknowledgments

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI: Grant Numbers 21700121, 25280110, and 25540159.

References

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Review Neuroscience*, *2*, 194–203.

Kanki, Y., Kuwahara, N., & Morimoto, K. (2014). An evaluation of physical strains while driving an electric wheelchair.

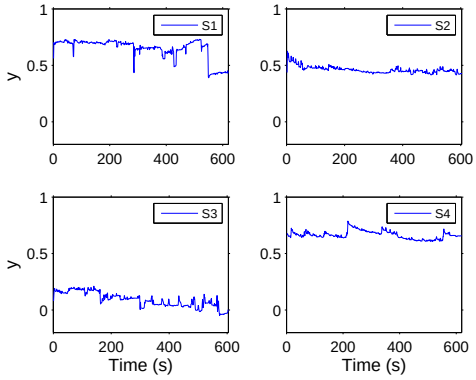


Figure 10: Time evolution of $y(t)$

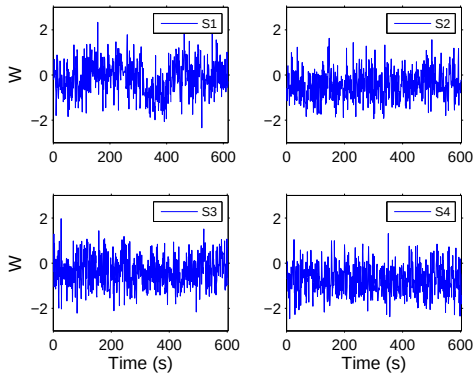


Figure 11: Time evolution of $W(t) = \log\left(\frac{\beta_\Sigma(t)}{\alpha_\Sigma(t)}\right)$

In *Proceedings of the 3rd international conference on advanced applied informatics* (pp. 863–866). Kitakyushu, Japan.

Klimesch, W., Doppelmayr, M., Russegger, H., & Schwaiger, T. P. J. (1998). Induced alpha band power changes in the human eeg and attention. *Neuroscience Letters*, *244*, 73–76.

Nobre, A., & Coull, J. (Eds.). (2010). *Attention and time*. Oxford University Press.

Rich, A., & Gureckis, T. (2015). The attentional learning trap and how to avoid it. In *Proceedings of the 37th annual cognitive science society meeting*. Pasadena, California.

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Educational Technology and Society*, *12*, 176–189.

Srivastava, N., & Vul, E. (2016). Attention dynamics in multiple object tracking. *Topics in Cognitive Science*.

Walsh, M., Gunzelmann, G., & Dongen, H. V. (2014). Comparing accounts of psychomotor vigilance impairment due to sleep loss. In *Proceedings of the 36th annual cognitive science society meeting*. Quebec City, Canada.

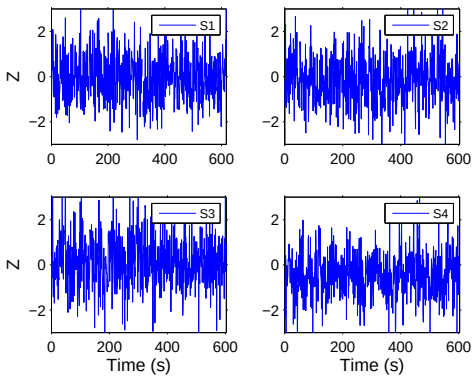


Figure 12: Time evolution of $Z(t) = \log\left(\frac{\alpha_L(t)}{\alpha_H(t)}\right)$

Table 5: Pearson's correlation coefficient between the indicators and the subjective level of attention

Participant	**var**(x)	**var**(y)	**mean**(W)	**mean**(Z)
P1	**-0.3582**	**-0.5319**	**0.4696**	**0.0638**
P2	**-0.8404**	**-0.5287**	-0.3096	-0.0243
P3	**-0.0571**	**-0.4817**	-0.3967	**0.0506**
P4	**-0.0391**	0.3090	**0.5100**	**0.3829**
P5	0.4570	0.6397	-0.7074	-0.2620
P6	**-0.3074**	**-0.4317**	**0.6861**	**0.5125**
P7	0.7891	0.6442	**0.0053**	-0.5354
P8	**-0.7207**	**-0.5835**	**0.5780**	**0.4819**
P9	**-0.5879**	**-0.3808**	**0.3891**	**0.4651**
P10	**-0.1183**	**-0.0600**	-0.1748	-0.1248
P11	**-0.0176**	**-0.3879**	**0.6273**	-0.2381
P12	0.1826	0.4798	-0.5971	-0.3259
P13	**-0.2274**	**-0.5233**	**0.4781**	**0.1991**
P14	**-0.0403**	0.4332	-0.3212	**0.0670**
P15	0.6796	0.8574	**0.2897**	**0.2070**
P16	**-0.3865**	**-0.4959**	**0.3135**	**0.4759**
P17	**-0.0190**	**-0.1357**	-0.3908	-0.1482

Yasui, Y., Tian, Q., & Yamauchi, N. (2008). A data process and wavelet analysis method used for monitoring daily physiological attributes. In *Proceedings of the 30th annual international conference of the ieee engineering in medicine and biology society* (pp. 1447–1450). Vancouver, Canada: IEEE.

Yoshida, K., Sakamoto, Y., Miyaji, I., & Yamada, K. (2012). Analysis comparison of brain waves at the learning status by simple electroencephalography. In *Proceedings of kes conference on knowledge-based intelligent information and engineering systems* (pp. 1817–1826).

Referential choice in identification and route directions

Adriana Baltaretu (a.a.baltaretu@tilburguniversity.edu)

Emiel J. Krahmer (e.j.krahmer@tilburguniveristy.edu)

Alfons A. Maes (maes@tilburguniversity.edu)
Tilburg center for Cognition and Communication, PO Box 90153, 5000 LE, Tilburg, The Netherlands

Abstract

Though communicative goals are an important element in language production, few studies investigate the extent to which these goals might affect the form and content of referring expressions. In this study, we directly contrast two tasks with different goals: identification and instruction giving. Speakers had to refer to a target building nearby or further away, so that their addressee would distinguish it between other buildings (identification) or give route directions and use the same building as a landmark (instructions). Our results showed that irrespective of goals, the referring expressions consisted of the same types of attributes, yet the attribute frequency and formulation differed. In the identification task, references were longer, contained more locative and more post-nominal modifiers. In addition, referential choices were influenced by the visual distance between the speaker and the target: when the speaker observed the target from far, references were longer and contained more often locative modifiers.

Keywords: communicative goals; identification task; route directions; referring expressions.

Suppose you want to point out a building to a tourist, either because that is the hotel he is looking for or because it is part of the route direction you are asked to give. In both cases, you would have to describe the building (the target) in such a way that your addressee can distinguish it from the other buildings (the distractors). Most probably you will have to choose between different attributes (modifiers) that single out the building (e.g., color, location, size, architectural style). Though you need to refer to the same building, the two situations are rather different. In the first case, the goal is to help your addressee distinguish the target from similar objects (e.g., look at X). This is similar to an identification (or discrimination) task where the speaker has to utter a distinguishing description in order to identify a target (Van Deemter, Gatt, van Gompel, & Krahmer, 2012). Compared to an identification task, focused on describing the target, when giving instructions, the distinguishing description is part of an action oriented speech act (e.g., go to X and turn left). The main goal of the latter is helping your addressee to turn on the correct street. The question that arises is to what extent humans tune their referential choices when having different communicative goals (e.g., object identification, route directions)?

This paper focuses on referring expression production in naturalistic scenes, e.g., how a speaker chooses among different attributes and how his choice is influenced by communicative goals. We define route directions as an action oriented discourse composed of descriptive (references to objects) and procedural information (references to actions), aimed at helping a person navigate in an unknown environment (e.g., Allen, 1997; Michon & Denis, 2001). Procedural information refers to a series of actions that one needs to take, while descriptive information typically refers to environmental features (landmarks), such as buildings and their attributes (Allen, 1997). Paired with actions (e.g., turn at X), landmarks ground the direction change that has to be performed in the intersection (Michon & Denis, 2001) and positively affect the quality of the instructions and navigation performance (Tom & Tversky, 2012). By an identification task we refer to a situation in which the speaker has to produce a unique referring expression. The referring expression is required to be a description consisting of a set of properties that singles out a target from similar objects (Van Deemter et al., 2012). The focus of this study is on initial (definite) references whose content is shaped by information available in the visual physical context. In general, these target descriptions consist of a definite article, a head noun, and nominal modifiers.

Communicative goals and referring expressions

The speaker's communicative goals are an important element of language production. This observation has been often acknowledged by qualitative work on dialogue development and some experimental work on reference and spatial language (Di Eugenio, Jordan, Thomason, & Moore, 2000; Daniel & Denis, 2004; Vorwerg & Tenbrink, 2006). Yet, as far as we know, there are no systematic comparisons assessing how much different referring expressions could be. We propose to investigate the extent to which communicative goals affect reference production, in a study that uses naturalistic scenes (depicting buildings in intersections), while taking into account a perceptual factor present in natural settings (the distance from which the target object is being observed). We compare the referring expressions produced when 'identifying' buildings, with the ones produced for the same objects while giving 'route directions', and assess how these references are evaluated by addressees.

There are several studies on reference that emphasize identification as a goal in itself (for a review, see Krahmer & Van Deemter, 2012; Van Deemter et al., 2012). These 'identification studies' start from a single goal and manipulate the attributes needed to identify objects (e.g., the STARS corpus (Paraboni, Galindo, & Iacovelli, 2016); the TUNA corpus (van Deemter, van der Sluis, & Gatt, 2006); the GRE3D3 corpus (Viethen & Dale, 2008, 2011). These studies focus on the properties of the target in contrast to other objects presented in a simple (sometimes grid-like) visual context. The speaker's purpose is to get the addressee pick out the target

object with a particular description. In general, identification is assumed to be part of a larger cooperative interaction (e.g., Brennan & Clark, 1996). However, this interaction is rarely modelled, and this type of studies typically report on a narrow set of experimental tasks. This raises the question to what extent these results would generalize when speakers have different goals. Do speakers produce similar references when describing an object for their addressee and when giving route directions, or do they adapt their references to the communicative context? How would the perceptual complex nature of the scenes influence reference production?

There are reasons to believe that different goals might affect referential choices. For example, when the participants' goal was to negotiate, rather than just identify, the preference for different attributes changed (Di Eugenio et al., 2000). In the COCONUT corpus, players had to buy items on a fixed budget, and in order to carry out this task, they had to describe and negotiate the furniture items that they believed were relevant to the task. In doing so, the item's price became one of the most often used attribute. Moreover, different goals may result in references with different levels of specification (Yoon, Koh, & Brown-Schmidt, 2012). When speakers had to give instructions they avoided scalar adjectives that were unnecessary for identification, and when they had to describe events, they referred more frequently to the objects' size.

Not only communicative goals, but also the importance attached to them might influence the speaker's reference production. When speakers had to identify surgeon' 'buttons on a control panel, compared to click on 'elements for another participant, they were more likely to include detailed descriptions containing redundant attributes and location information (Arts, Maes, Noordman, & Jansen, 2011). These had ultimately helped the addressee fulfil his task faster. Similarly, when participants had the goal of instructing someone how to operate once an alarm clock, compared to teaching someone that needs remember how to do this every night, their referring expressions contained more detailed information (Maes, Arts, & Noordman, 2004).

In everyday language, ones' communicative goals are realized through choices among the expressive means (lexical items, omissions/deletions, syntactic structure, word order, phrases and prepositions, among others; Talmy (2000, p. 346). In this study, we directly compare two tasks with different communicative goals using the same set of visual scenes. It is conceivable that references might have a different level of specification and the types of attributes and their distribution might be different across the two tasks. For example, as speakers focus more on the target in the identification task, they might mention more details about the buildings (such as the number of windows on a façade). On the other hand, route directions have a strong instructional focus and procedural information is arguably more important for the success of the task. Though, referring to a landmark also requires an 'identification' step, it might be more important what the speaker is trying to get the addressee to do with the utterance. As speakers need to convey two types of information, we might expect shorter and more focused references in the route directions than in the identification task.

Apart from communicative goals, a second factor that we explore in relation to reference production and comprehension is related to the distance from which the participants perceive the target. A difference of distance would affect the size of the target, the amount of visual details, and the number of objects in the visual field. In such situations, perception and recognition of object properties might be harder to assess for both speaker and listener; for example, the visibility of some object (or of one of its parts) may not be inferred with complete certainty. Distance differences seem to trigger various strategies when producing references. For example, participants tend to point more often when close to the object and refer less to the target's location, but use more locative phrases when pointing becomes ambiguous (Van Der Sluis & Krahmer, 2004; Bangerter, 2004). We question when pointing is not an available option, how would references change. When the intersection is far, the references might be longer, due to high ambiguity levels, however the opposite might be the case as from close by one has access to more visual details.

Methods

Participants

Eighty native Dutch-speaking students (40 dyads) of Tilburg University (62 women, mean age 21 years) participated in exchange for partial course credits. Participants were randomly assigned to speaker roles (28 women). The study was carried out in accordance with the recommendations of APA guidelines for conducting experiments, and all participants gave written consent for use of their data.

Materials

Experimental materials consisted of 36 target objects (buildings depicted in Google Street View snapshots of intersections). These buildings were pictured from two distance points (36 scenes taken with a camera positioned at 40m away from the target and 36 scenes taken with a camera positioned at 20m away from the target). This resulted in 72 experimental scenes. The target objects were buildings marked with red squares, and placed in the corners of 4-way intersections (see Figure 1). Targets were always placed in the corners with highest visibility, namely on the other side the intersection. The targets' position was counterbalanced, so that in half of the scenes they were placed on the left side of the intersection. In both tasks we used the same scenes. For the route directions task, we added arrows indicating the turning direction (see Figure 2).

In addition, 36 filler scenes were introduced in order to present participants with a range of different navigation scenarios, and avoid participants from relying on fixed responses. The filler scenes depicted buildings in intersections with complex geometric structures.

Procedure

Participants worked in pairs, and completed their task on separate computers. Pointing was discouraged by placing the computer screens in between the participants. The speaker

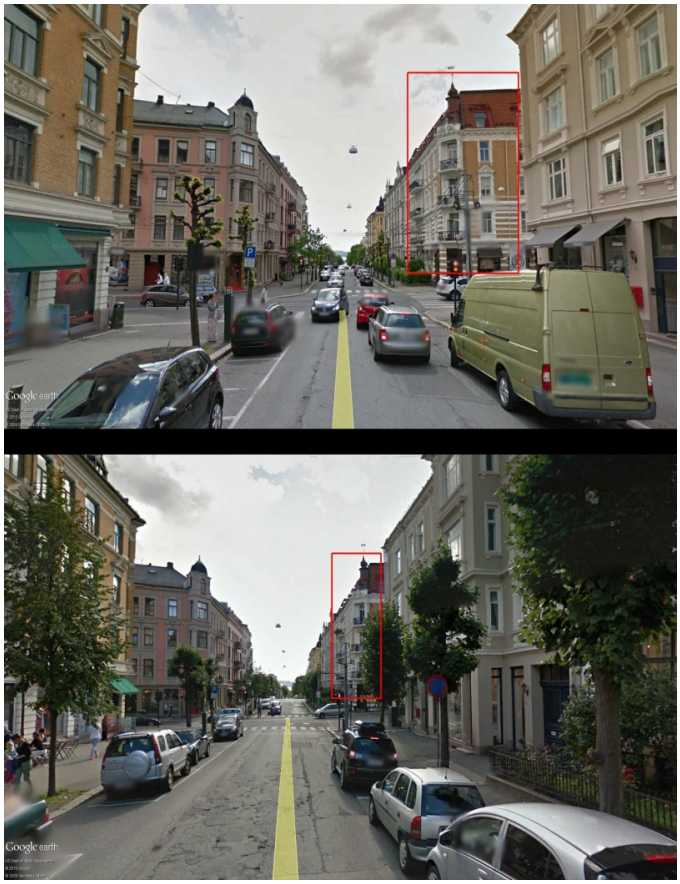


Figure 1: Experimental scenes from the description task depicting a target building near (above) and far (below)



Figure 2: Experimental scenes from the route direction task depicting a target building near (above) and far (below)

received a scene with a marked building (and an arrow indicating the route), while the listener received the same image without any markings.

Participants were randomly assigned to one of the two tasks. In the descriptions task (D-task), speakers were asked to refer to the marked building in such a way that the listener could uniquely identify it. In the route directions task (RD-task), speakers were asked to give route directions, and in doing so make use of the marked building as landmark. In order to elicit uniform responses, speakers were asked to verbally fill in templates (a typical procedure in identification studies, e.g., Dale & Viethen, 2009). They had to fill in the following templates: "click on ... " (descriptions) and "go to ... and turn left / right" (route directions). In both situations, the listeners had to click on the correct building. Listeners were allowed to ask questions if the speakers' instructions were unclear. Experimental scenes were divided in presentation lists, so that each participant would see each target object only once. Participants were randomly assigned to one of the presentation lists. The experiment started with three practice trials, next 72 trials (36 experimental trials and 36 filler trials) were presented in different random orders.

Design

This study had a 2 x 2 design with Task (levels: D-task, RD-task) as between participants factor and Distance (levels: far, near) as within participants factor. For the first analysis, we looked at the length of the references (number of words). Next, we assessed the type of attributes mentioned in relation with the target (e.g., color, location, etc), their frequency and distribution (pre / post nominal modifiers), as well as the length of the remaining phrase after the speaker has mentioned the noun denoting the target building (number of words). The presence / absence of attributes coupled with the target noun was binary coded. A phrase like *the large building on the left, next to the white tower* consists of a target noun (*building*) and the following attributes: size (*large*) and location (*on the left, next to the white building*). In order to analyze the differences regarding the length of the references, data transformations (log data) were applied due to a skewed distribution. For ease of understanding, means and standard deviations reported here represent untransformed data.

In order to test the observed differences, we conducted separate statistical analyses using logit mixed model analysis (Jaeger, 2008), following the recommendations of Barr, Levy, Scheepers, and Tily (2013). We used the mixed logit model

analysis as it can correctly account for random subject and item effects in a one-step analysis. The models were fitted using the LMER function from the LanguageR Package in R (version 2.15.2; CRAN project; The R Foundation for Statistical Computing, 2012). To determine whether the two conditions significantly differed from each other, we started by constructing a maximal model with a full random effect structure. This had Task and Distance as fixed factors; Speakers and Scenes as random factors; intercepts and random slopes for Speakers and Scenes to account for between-subject and between-item variation. In case the dependent variable was binary coded, the factors were centered to avoid collinearity. In case the model did not converge, we only excluded random slopes with the lowest variance until convergence was reached. The results from the first converging model, as well as the structure of the model are reported. The p - values were estimated via parametric bootstrapping over 100 iterations.

Results

In total 1440 references were produced (40 speakers * 36 scenes). Data from two dyads (one from each condition) were discarded on the basis of task misunderstanding. The referring expressions consisted of a noun denoting the target object and all the phrases attached to it. In practice, in the D-task, the references consist of all the phrases after "click on", and in the RD-task, they consist of all the phrases between "Go to" and "and turn". In some cases, participants omitted some part of the template, but kept the overall structure (e.g., *The red building, and turn ...*), and these cases were included in the analysis. In fact, 91 percent of the references had a similar structure (noun and modifiers). The remaining, nine percent of the cases resulted in utterances with different structure (e.g., *The building is ...*), evenly distributed across the two tasks (5% cases in the D-task and 4% in the RD-task). As this could potentially bias aspects such as word counts, we decided to exclude these cases from the analysis.

Length of referring expressions

The first converging model had random intercepts for Speakers and Scenes, R^2 marginal = 0.10, R^2 conditional = 0.56. The length of the references was significantly influenced by the Task ($\beta = -0.369$; $SE = 0.12$; $p < .01$). Referring expressions were longer in the D-task ($M = 16.3$, $SD = 1.3$), than in the RD-task ($M = 11$, $SD = 1.3$). There was a main effect of Distance ($\beta = 0.076$; $SE = 0.03$; $p < .01$). When close to the target, speakers produced slightly shorter referring expressions ($M = 13$, $SD = .88$), than when further away ($M = 14$, $SD = 1.01$). There was no interaction between Distance and Task ($p > .05$).

Type of attributes

Speakers described targets by referring to ten types of attributes (see Table 1). The same types were produced in both tasks. Top three most frequent attributes in both tasks are location, followed by colour and references to structural parts of the target (such as chimney, stairs, doors).

In the D-task, speakers mentioned more often the location and color of the object than in the RD-task. We analyzed statistically these differences. With respect to location, the first converging model had random intercepts for Speakers and Scenes. There was a significant difference between the tasks regarding locative information ($\beta = -2.573$; $SE = 0.81$; $p < .01$). There was a main effect of Distance ($\beta = 0.535$; $SE = 0.22$; $p < 0.01$). When close to the target, speakers referred less often to the position of the object ($M = .77$, $SD = 0.41$), than when further away ($M = .82$, $SD = 0.32$). There was no significant interaction between Task and Distance ($p > .05$).

As for the difference regarding references to the target color, the first converging model had random intercepts for Speakers and Scenes, and there were no main effects ($p > .05$) or interactions ($p > .05$).

Table 1: Type of attributes, examples and attribute frequency split by task

Type	Examples	Frequency	
		Route Directions	Descriptions
location	The building across the intersection, on the left	68%	92%
color	The white building	42%	47%
part of building	The building with (five) balconies / with (red) roof / with (two) windows	35%	36%
decoration	The building with stripes / with flowers pots / with hanging things / with flags	11%	12%
size	The smallest building	8%	7%
shape	The long building	0.6%	0.3%
age	The modern building	1%	0.5%
architectural style	The Italian building	2%	0.3%
materials	The brick building	1%	1%
evaluative	The ugly building	1%	2%

Distribution of modifiers

First there was significant difference regarding the number of words produced after the target noun and the first converging model had random intercepts for Speakers and Scenes ($\beta = -0.508$, $SE = 0.16$, $p < .01$). The number of words produced after the target noun was longer in the D-task ($M = 13.62$, $SD = 1.25$), than in the RD-task ($M = 7.81$, $SD = 1.25$). There was no main effect of Distance ($p > .05$) and no interaction between the factors ($p > .05$).

In general, there were more post-nominal modifiers in the D-task ($N = 1002$) compared to the RD-task ($N = 683$). The reversed pattern was observed for pre-nominal modifiers, which were more frequent in the RD-task ($N = 306$) than in the D-task ($N = 212$). In both tasks, pre-nominal modifiers consisted mostly of location (e.g., *the left building*; *the second building*), color (e.g., *the white building*) and size references (e.g., *the large building*) (see Table 2). Post-nominal modifiers consisted also of location, color and size, but mostly included references to structural parts of the building that syntactically can not be framed otherwise (e.g., *the building with two balconies*).

Table 2: Distribution (number of cases) of pre- and post-nominal modifiers split by task.

	Pre-nominal		Post-nominal	
	Route Directions	Descriptions	Route directions	Descriptions
location	151	116	270	434
colour	128	70	136	221
size	21	17	27	28
other	6	9	250	319

Lexical Fillers

We define lexical fillers (hedges) as words or phrases that are conventionally used for signalling hesitation, marking the reference as provisional (Brennan & Clark, 1996). The references produced in the two tasks included a different amount of lexical fillers (see Figure 3). D-task triggered more lexical fillers, compared to the RD-task.

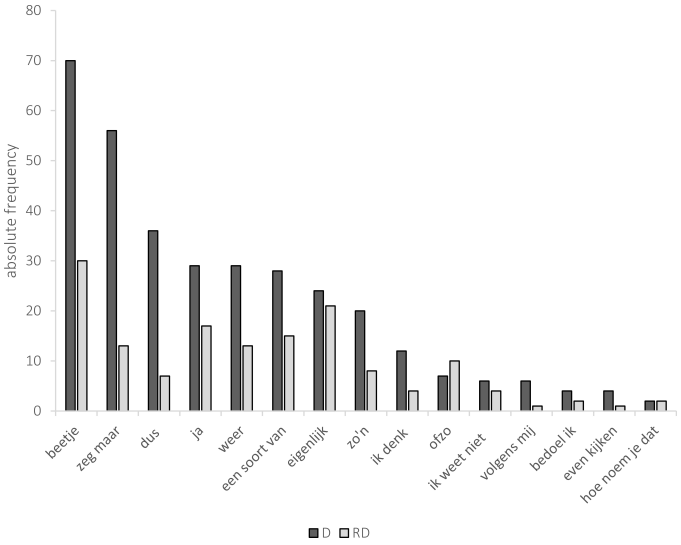


Figure 3: Distribution (number of cases) of lexical fillers split by task

Error rates

There was a small number of cases in which the addressee clicked on wrong buildings (3 cases in RD-task and 6 in the D-task) and relatively few clarification questions (13 questions in the D-task and 26 questions in the RD-task). These questions mostly asked for simple clarifications (e.g., *a big building?*) to which speakers mostly uttered a short confirmation (e.g., *yes*).

Conclusions and Discussion

In this study, we questioned to what extent different communicative goals (object identification, giving route directions) influence reference production, while using complex naturalistic scenes and taking into account a perceptual factor, the visual distance from which a target object is observed. The referring expressions were elicited using real-world scenes, and the design allowed a direct comparison between the references elicited in two communicative settings. Preliminary results showed that identification as opposed to referring to objects in route directions triggered a number of differences regarding reference formulation, and almost no semantic differences.

First, there were no semantic differences, in both tasks speakers used the same types of attributes to refer to buildings. This pattern of results suggests that studies, where identification is the main purpose of the interaction, could generalize to a large extent to other settings. Second, the syntactic formulation of the references was different. Descriptions were longer and contained more post-nominal information and more words following the target noun than references in route directions, which suggests that speakers described the target in more detail. Contrastively, in route directions, referring expressions were shorter and tended to contain more pre-nominal modifiers. Previous research suggests that attributes in pre-nominal position are more efficient for identification (Rubio-Fernndez, 2016), which might suggest that in a larger communicative act speakers 'optimize' their references, in such a way that the addressee can find the target faster and more easily (see also, Clarke, Elsner, & Rohde, 2015). We plan to test how differences in information structure affect addressee performance, when one needs to find his way in an unknown environment. Moreover, in the description task, speakers tended to pay more attention to the location of the object. Locating an entity has been suggested to be a robust and successful strategy in object identification (Paraboni et al., 2016) though it could have contributed to the increased length of the descriptions (see also Vorwerg & Tenbrink, 2006). In addition, and unexpectedly, on the pragmatic level, we found different levels in the use of markers of nuance and hesitation. We question whether the difference in absolute length between tasks was caused by the post-nominal structures, the pragmatic hesitations, or more words to express some attributes as with location? We assume it is a combination of these three.

These differences could be caused by the type of discourse elicited in the tasks. On the one hand, action related information might be more important for route directions. The difference in length might be explained by the fact that speakers want to make sure the addressee finds the correct street, and less so the correct building. Moreover, landmarks are meant to improve route directions, and long complex references would doubtfully be an efficient contribution. On the other hand, descriptions might be a more difficult and less planned task, as suggested by a higher level of post-nominal information and lexical fillers.

In addition, we explored if the distance between the speaker and the target might influence referring expressions. This factor relates to several aspects that increase the uncertainty with respect to what building is referred to. For example when the speaker is further from the target, there are more similar buildings in the scene and the target could be partially occluded. References were longer and contained more location information when the speaker was far from the target.

However the differences between the two conditions, as well as the effect size, were rather small.

In sum, when comparing the effects of different communicative goals on reference production an interesting pattern of results emerges. In line with studies focused exclusively on identification, we noticed a stable preference for the same type of attributes. Yet, different communicative goals seem to shape the information structure of the references and influence the frequency of some of the attributes.

References

Allen, G. L. (1997). From knowledge to words to wayfinding: Issues in the production and comprehension of route directions. In S. C. Hirtle & A. U. Frank (Eds.), *Spatial Information Theory A Theoretical Basis for GIS* (pp. 363–372). Berlin: Springer.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493.

Clarke, A. D. F., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in Psychology*, *6*(1793). doi: 10.3389/fpsyg.2015.01793

Dale, R., & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th european workshop on natural language generation* (pp. 58–65). Stroudsburg, PA, USA: Association for Computational Linguistics.

Daniel, M.-P., & Denis, M. (2004). The production of route directions: Investigating conditions that favour conciseness in spatial discourse. *Applied Cognitive Psychology*, *18*(1), 57–75.

Di Eugenio, B., Jordan, P. W., Thomason, R. H., & Moore, J. (2000). The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, *53*(6), 1017–1076.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, *37*(2), 117–144.

Michon, P.-E., & Denis, M. (2001). When and why are visual landmarks used in giving directions? In D. R. Montello (Ed.), *Spatial information theory* (pp. 292–305). Berlin: Springer.

Paraboni, I., Galindo, M. R., & Iacovelli, D. (2016). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 1–24.

Rubio-Fernndez, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153). doi: 10.3389/fpsyg.2016.00153

Talmy, L. (2000). *Toward a cognitive semantics* (Vol. 1). MIT press.

Tom, A. C., & Tversky, B. (2012). Remembering routes: streets and landmarks. *Applied Cognitive Psychology*, *26*(2), 182–193.

Van Deemter, K., Gatt, A., van Gompel, R. P., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166–183.

van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the fourth international natural language generation conference* (pp. 130–132). Stroudsburg, PA, USA: Association for Computational Linguistics.

Van Der Sluis, I., & Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In S. Kim & D. Youn (Eds.), *Proceedings of international conference on spoken language processing* (p. 1005-1008). Korea: ISCA.

Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. In W. Michael, N. Crystal, & M. David (Eds.), *Proceedings of the fifth international natural language generation conference* (pp. 59–67). Stroudsburg, PA, USA: Association for Computational Linguistics.

Viethen, J., & Dale, R. (2011). Gre3d7: A corpus of distinguishing descriptions for objects in visual scenes. In B. Anja, E. Roger, G. Albert, & S. Kristina (Eds.), *Proceedings of the ucnlg and evaluation: Language generation and evaluation workshop* (pp. 12–22). Stroudsburg, PA, USA: Association for Computational Linguistics.

Vorwerg, C., & Tenbrink, T. (2006). Discourse factors influencing spatial descriptions in english and german. In T. Barkowsky, M. Knauff, G. Ligozat, & D. Montello (Eds.), *Spatial Cognition V Reasoning, Action, Interaction* (pp. 470–488). Berlin Heidelberg: Springer-Verlag.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*(4), 699–707.

Statistical learning creates novel object associations via transitive relations

Yu Luo (yuluo@psych.ubc.ca)
Department of Psychology
University of British Columbia

Jiaying Zhao (jiayingz@psych.ubc.ca)
Department of Psychology
Institute for Resources, Environment and Sustainability
University of British Columbia

Abstract

A remarkable ability of the cognitive system is the creation of new knowledge based on prior experiences. What cognitive mechanisms support such knowledge creation? We propose that statistical learning not only extracts existing relationships between objects, but also generates new associations between objects that have never been directly associated. Participants viewed a continuous color sequence consisting of base pairs (e.g., A-B, B-C), and learned these pairs. Importantly, they also successfully learned a novel pair (A-C) that could only be associated through transitive relations between the base pairs (Exp1). This learning, however, was not successful with three base pairs (e.g., learning A-D from A-B, B-C, C-D), revealing a limit in this transitive process (Exp2). Beyond temporal associations, novel transitive associations can also be formed across categorical hierarchies (Exp3), but with limits (Exp4&5). The current findings suggest that statistical learning provides an efficient scaffold through which new object associations are transitively created.

Keywords: Statistical learning; transitive inference; implicit associations; regularities; categorical hierarchy

Introduction

A remarkable feature of the human cognitive system is that it is able to generate new knowledge based on existing information. For example, upon learning that Starbucks is next to a grocery store which is next to a wine shop, people can automatically infer that Starbucks is close to the wine shop. What cognitive mechanisms support such inference?

One mechanism that detects the relationships among objects in the environment is statistical learning, which involves the extraction of relationships between individual objects in terms of how they co-occur over space or time (Fiser & Aslin, 2001; Saffran, Aslin, & Newport, 1996). Statistical learning operates in multiple sensory modalities and feature dimensions (Conway & Christiansen, 2005; Fiser & Aslin, 2001; Saffran et al., 1996; Turk-Browne, Isola, Scholl, & Treat, 2008), draws attention to the co-occurring objects (Yu & Zhao, 2015; Zhao, Al-Aidroos, & Turk-Browne, 2013), and facilitates information compression (Brady, Konkle, & Alvarez, 2009; Zhao & Yu, 2016).

The critical distinction between statistical learning and other forms of associative learning is that for statistical learning the two objects are related based only on conditional or conjunctive probability, whereas for associative learning the two objects are associated through explicit cues, such as segmentation, grouping, or proximity. Statistical learning occurs incidentally, without conscious intent or explicit awareness, and the nature of the knowledge is implicit, in that observers are not explicitly aware of the co-occurrences between the objects (Turk Browne, Jungé, & Scholl, 2005; Turk-Browne, Scholl, Chun & Johnson, 2009).

Such implicit knowledge is also stimulus specific. That is, exposure to the co-occurrences between individual objects leads to the knowledge that these objects are associated. Given the ability to implicitly extract the co-occurrences, an unexplored question is whether statistical learning forms new associations of objects that have never co-occurred with each other and can only be associated based on prior knowledge.

Initial support for this question comes from studies on transitive inference. For example, previous work has shown that people can successfully learn the hierarchical order of abstract objects (e.g., if A<B and B<C, then A<C) without awareness using trial-by-trial feedback (Greene, Spellman, Dusek, Eichenbaum, & Levy 2001; Kumaran & Ludwig, 2013). A recent study shows that the hippocampus supports the transfer of values across objects that were previously associated, enabling people to make decisions between options that were never directly rewarded (Wimmer & Shohamy, 2012). Given this possibility of reactivating previous connections between objects, we predict that statistical learning can form novel associations among the objects that were never directly associated in the past.

The goal of our current study was to examine whether statistical learning enables the creation of new associations among objects that have never appeared together. Here, we focused on two basic types of transitive relations and asked: (1) given the pairs A-B and B-C, do people automatically learn a new pair A-C (Experiments 1 and 2)? and (2) given object exemplar pairs at one categorical level (e.g., New York-London), do people automatically learn new pairs at the subordinate level (e.g., Central Park-Hyde Park) and the superordinate level (e.g., USA-UK; Experiments 3, 4, and 5)?

Experiment 1

This experiment examined whether new associations could be formed between objects that had never appeared together.

Participants

Thirty undergraduates (21 female; mean age=19.5 years, SD=1.3) from University of British Columbia (UBC) participated for course credit. Participants in all experiments reported normal or corrected-to-normal vision and provided informed consent. All experiments reported here were approved by the UBC Behavioral Research Ethics Board.

Stimuli

The stimuli consisted of nine circles in nine distinct colors (color name=R/G/B values: red=255/0/0; green=0/255/0; blue=0/0/255; yellow=255/255/0; magenta=255/0/255; cyan=0/255/255; gray=185/185/185; brown=103/29/0; black=0/0/0). Each circle subtended 1.6° of visual angle. The colored circles were randomly assigned into six base pairs for each participant and remained constant throughout the experiment. The six base pairs contained three groups of two base pairs. In each group, the second color in the first pair was the same as the first color in the second pair (e.g., A-B, B-C, Fig.1a). This allowed us to test if people could automatically learn a novel pair (A-C) given the two base pairs (A-B and B-C). There were three novel pairs from the six base pairs. Importantly, the two colors in the novel pair were never directly associated with each other. Each base pair was repeated 50 times to form a single continuous temporal sequence of circles in a pseudorandom order with two constraints: no single base pair could repeat back-to-back and no two overlapping base pairs (e.g., A-B, B-C) could occur consecutively.

Procedure

The experiment contained two phases: exposure phase and test phase. During the exposure phase, one colored circle appeared at the center of the screen for 500ms followed by a 500ms inter-stimulus interval (ISI) in each trial. Participants performed a 1-back task where they judged as quickly and accurately as possible whether the current circle had the same color as the previous circle (by pressing the "/" or "z" key for same or different, respectively, key assignment counterbalanced). For the 1-back task, each color had a 20% chance of repeating the previous color, producing a total of 360 trials. This 1-back task served as a cover task irrelevant to learning, in order to conceal the true purpose of the study, ensuring that learning of the color pairs was incidental. Participants were not told anything about the color pairs.

After exposure, participants completed a surprise two-alternative forced choice (2AFC) test phase to examine whether they had successfully learned the base pairs. In each trial, participants viewed two sequences of circles at fixation. Each circle appeared for 500ms followed by a 500ms ISI, and each sequence was separated by a 1000ms pause. Participants judged whether the first or second sequence looked more familiar based on the exposure phase. If they did not respond during the sequence presentation or ISI, the screen remained blank until response. One sequence was a base pair, and the other was a "foil" (e.g., A-E) composed of one color from a base pair (e.g., A-B), and the other from a different base pair (e.g., D-E) while preserving the temporal positions in the pairs. Each base pair was tested against a foil, which was then repeated, resulting in 12 trials in total. It's important to note that each base pair and foil were presented the same number of times at test. Thus, to discriminate the base pair from the foil, participants needed to know which two particular colors followed each other during exposure. The order of the trials was randomized, and whether the base pair or foil appeared first was counterbalanced across trials.

At the test phase, we also examined whether participants learned the novel pair (e.g., A-C from A-B and B-C). The foil was constructed by selecting two colors from two novel pairs while maintaining the temporal positions in the pairs. Each novel pair was tested against a foil. The two colors in the foil came from two different base pairs that never shared a common color. Each novel pair and foil were presented the same number of times at test. If participants chose the novel pair as more familiar, this would suggest that they had formed a new association between two objects that had never directly followed each other but could only be inferred given the knowledge about the two base pairs.

A debriefing session was conducted after test, where participants were asked if they had noticed any colored circles that appeared one after another. For those who responded yes, we further asked them to specify which colors followed each other.

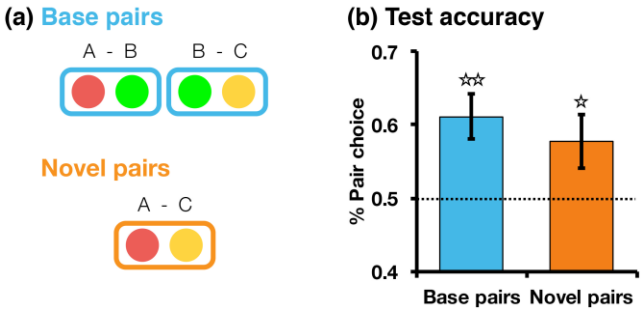


Figure 1: (a) Two example base pairs (A-B, B-C) and a novel pair (A-C). (b) Percent of times the pair was chosen as more familiar than the foil at test (error bars reflect ±1 SEM; *p<.05, **p<.01; dotted line=chance performance of 50%).

Results and Discussion

At the test phase, base pairs were chosen as more familiar than foils for 61.0% (SD=16.9%) of the time, which was reliably above chance (50%) [$t(29)$=3.61, p=.001, d=0.66] (Fig.1b). This indicates that participants have successfully learned the temporal co-occurrences between the two colors in the base pairs.

Moreover, novel pairs were chosen as more familiar than foils for 58% (SD=19.8%) of the time, which was again reliably above chance (50%) [$t(29)$=2.15, p=.04, d=0.39] (Fig.1b), suggesting that participants have also successfully learned the novel pairs, although the two colors in the novel pair never directly followed each other during exposure.

During debriefing, only five participants reported noticing color pairs, but none could correctly report which specific colors followed each other. This suggests that participants had no explicit awareness of the base pairs or the novel pairs.

These findings demonstrate that statistical learning automatically forms implicit novel associations between objects that have never appeared together and can only be associated via transitive relations from prior experiences.

Experiment 2

This experiment aimed to examine the limits of transitive associations by increasing the chain of object associations.

Participants

A new group of 30 undergraduates (20 female, mean age=20.5 years, SD=2.7) from UBC participated in the experiment for course credit.

Stimuli and Procedure

The stimuli and the procedure were identical to those in Experiment 1, except that we added one more base pair such that three base pairs formed a novel pair. As before, six base pairs were created for each participant. For every three base pairs, the second color in the first pair was the same as the first color in the second pair, and the second color in the second pair was the same as the first color in the third pair (e.g., A-B, B-C, C-D, Fig.2a). The novel pair (e.g., A-D) consisted of the first color in the first pair and the second color in the third pair. As before, participants performed a cover 1-back task to ensure incidental encoding of the pairs during exposure. Afterwards, participants completed the surprise test phase where they chose whether the pair or the foil looked more familiar.

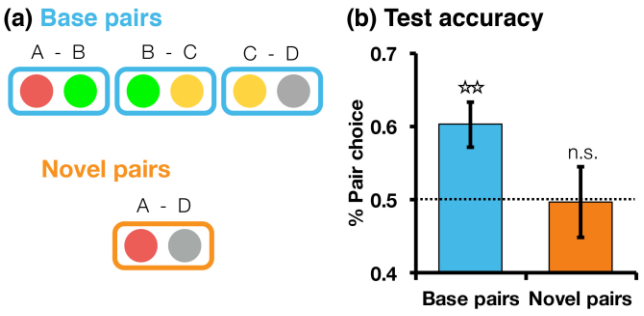


Figure 2: (a) Three example base pairs (A-B, B-C, C-D) and a novel pair (A-D). (b) Percent of times the pair was chosen as more familiar than the foil at test (error bars reflect ±1 SEM; **p<.01; dotted line=chance performance of 50%).

Results and Discussion

At test, base pairs were chosen as more familiar than foils for 60.0% (SD=16.8%) of the time, which was reliably above chance (50%) [t(29)=3.36, p=.002, d=0.61] (Fig.2b), suggesting that participants have successfully learned the co-occurrences between the two colors in the base pairs.

However, novel pairs were chosen as more familiar than foils for 49.7% (SD=26.5%) of the time, which was not reliably above chance (50%) [t(29)=0.06, p=.95, d=0.01] (Fig.2b). This suggests that learning of the novel pairs was not successfully expressed at test.

During debriefing, only one participant reported noticing color pairs, but the participant could not correctly report which specific colors temporally followed each other. This again suggests that participants had no explicit awareness of the base pairs or the novel pairs.

The findings demonstrate that while participants have learned the three base pairs, they could not form the novel association between objects in the first and the third pairs. This reveals a limit in the associations that can be formed transitively. This limit may reflect processing constraints as the number of objects or pairs increases.

Experiment 3

As shown in Experiments 1 and 2, novel associations can be formed transitively between objects that have never co-occurred before, and learning of the transitive association may depend on the length of associations. Here, we aimed to generalize the findings by examining another type of transitivity across the categorical hierarchy. Specifically, given the association between two exemplars at one categorical level, can people automatically learn the same association at the subordinate and the superordinate levels?

Participants

A new group of 80 undergraduates (58 female; mean age =20.6 years, SD=3.2) from UBC participated in the experiment for course credit.

Stimuli

The stimuli consisted of eight city names (New York, London, Vancouver, Paris, Tokyo, Beijing, Barcelona, and Bangkok), the eight corresponding park names (Central Park, Hyde Park, Stanley Park, Champ de Mars Park, Yoyogi Park, Bei Hai Park, Güell Park, and Lumpini Park), and the eight corresponding country names (USA, UK, Canada, France, Japan, China, Spain, and Thailand). The eight cities were randomly grouped into four base pairs for each participant (e.g., New York-London, Fig.3a). The city base pairs produced four park pairs at the subordinate level (e.g., Central Park-Hyde Park), and four country pairs at the superordinate level (e.g., USA-UK). The park pairs and the country pairs served as novel pairs to be tested at the test phase, and were never presented in the exposure phase. Each city base pair was repeated 50 times to form a single continuous sequence of cities in a pseudorandom order with the constraint that no city pair could repeat back-to-back.

Procedure

Since participants may not know which park is in which city, they were first trained on a separate task to associate each park with a given city. In this task, participants viewed a park

and selected which city contained the park (by pressing a key from "1" to "8"), and received feedback on each trial. They had to achieve 100% accuracy on this task before starting the experiment. We did not test city-country association, since we assumed that most participants should know which city is in which country. There was no mention of any country names, city or park pairs, before starting the experiment.

The experiment consisted of an exposure phase and a test phase, as in Experiment 1. During the exposure phase, one city name was presented at fixation for 500ms followed by a 500ms ISI in each trial. As before, participants performed a 1-back task where they judged whether the current city name was the same as the previous name. For the 1-back task, each city name had a 20% chance of repeating the previous name, producing a total of 360 trials. This 1-back task served as a cover task to conceal the true purpose of the study, ensuring that learning of the city pairs was incidental.

After exposure, participants completed the surprise test phase, as in Experiment 1, to see if they had learned the city pairs, and more importantly, to see if they had learned the corresponding park pairs or country pairs which were never presented during exposure. In each trial, participants judged whether the pair or the foil looked more familiar based on what they saw in the exposure phase. There were three blocks of trials. In the first block, each city pair was tested against a foil where two cities never followed each other during exposure. In the second block, each park pair corresponding to its city pair was tested against a foil that contained the two parks corresponding to the two cities in the foil in the first block. Likewise, in the third block, each country pair and the foil corresponded to the city pair and the foil in the first block. The order of the last two blocks was randomized.

A debriefing session was conducted after the test phase, where participants were asked if they had noticed any cities that appeared one after another. For those who responded yes, we further asked them to specify which cities followed each other.

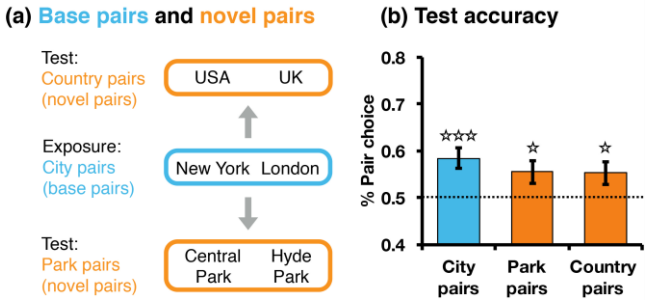


Figure 3: (a) City pairs served as base pairs at one categorical level, and park pairs and country pairs served as novel pairs at the subordinate and the superordinate levels, respectively. (b) Percent of times the pair was chosen as more familiar at test (error bars reflect ±1 SEM; *p<.05, ***p<.001; dotted line=chance performance of 50%).

Results and Discussion

At the test phase, the city base pairs were chosen as more familiar than foils for 58% (SD=19.5%) of the time, which was reliably above chance (50%) [$t(79)=3.82$, $p<.001$, $d=0.43$] (Fig.3b). This indicates that participants have successfully learned the temporal co-occurrences between the two cities in a base pair during exposure.

Moreover, park pairs were chosen as more familiar than foils for 55% (SD=19.6%) of the time, which was again reliably above chance (50%) [$t(79)=2.49$, $p=.01$, $d=0.28$] (Fig.3b), suggesting that participants have successfully learned the park pairs, although no parks were ever presented during exposure. Likewise, country pairs were also chosen as more familiar than foils for 55% (SD=22.3%) of the time, which was again reliably above chance (50%) [$t(79)=2.09$, $p=.04$, $d=0.23$] (Fig.3b), suggesting that participants have also successfully learned the country pairs, although no countries were presented during exposure.

During debriefing, seven participants reported noticing city pairs, but none correctly reported which specific cities followed each other. This suggests that participants had no explicit awareness of the city base pairs.

These results suggest that participants spontaneously formed new associations at both the subordinate and the superordinate levels, based on the regularities extracted at one categorical level. This finding provides further evidence that statistical learning automatically forms implicit novel associations between objects at different levels along the categorical hierarchy, even if these objects are never directly experienced or associated with each other.

Experiment 4

The goal of this experiment was to examine whether the findings in Experiment 3 were specific to city pairs, and whether there were limits in forming the novel transitive associations across the categorical hierarchy.

Participants

A new group of 45 undergraduates (31 female; mean age=20.7 years, SD=2.5) from UBC participated in the experiment for course credit.

Stimuli and Procedure

The stimuli and the procedure were identical to those in Experiment 3, except that park pairs served as base pairs during exposure, and city pairs and country pairs served as novel pairs at test (Fig.4a).

Results and Discussion

At test, the park base pairs were chosen as more familiar than foils for 68% (SD=10.0%) of the time, which was reliably above chance (50%) [$t(44)=12.43$, $p<.001$, $d=1.85$] (Fig.4b), indicating that participants have successfully learned the temporal co-occurrences between the two parks in a base pair during exposure.

Moreover, city pairs were chosen as more familiar than foils for 54.1% (SD=16.3%) of the time, which was again reliably above chance (50%) [$t(44)=1.71$, $p=.05$, $d=0.26$], suggesting that participants have successfully learned the city pairs, even though no cities were ever presented during exposure. However, country pairs were chosen as more familiar than foils for 51.2% (SD=20.4%) of the time, which was not reliably above chance (50%) [$t(44)=0.41$, $p=.34$, $d=0.06$] (Fig.4b).

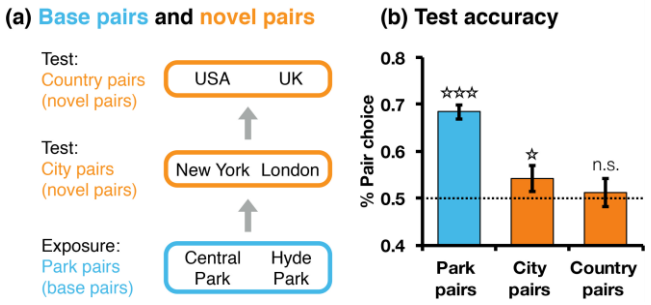


Figure 4: (a) Park pairs were base pairs. City and country pairs were novel pairs. (b) Percent of times the pair was chosen as more familiar (error bars reflect ±1 SEM; *p<.05, ***p<.001; dotted line=chance performance of 50%).

During debriefing, three participants reported noticing park pairs, but none correctly reported which specific parks followed each other. This suggests that participants had no explicit awareness of the park base pairs.

These results suggest that participants could successfully form new associations at one superordinate level above the original subordinate level where regularities were learned, but not at two superordinate levels above. This reveals a limit in how far new associations can be formed beyond the categorical level where objects were originally associated.

Experiment 5

The final experiment aimed to replicate the results from Experiment 4 by exposing regularities at the superordinate level and testing whether new associations can be formed at all subordinate levels.

Participants

A new group of 80 undergraduates (61 female; mean age =20.8 years, SD=5.2) from UBC participated in the experiment for course credit.

Stimuli and Procedure

The stimuli and the procedure were identical to those in Experiment 3, except that country pairs served as base pairs during exposure, and city pairs and park pairs served as novel pairs at test (Fig.5a).

Results and Discussion

At test, the country base pairs were chosen as more familiar than foils for 61% (SD=22.0%) of the time, which was reliably above chance (50%) [$t(79)=4.52$, $p<.001$, $d=0.51$] (Fig.5b), indicating that participants have successfully learned the temporal co-occurrences between the two countries in a base pair during exposure.

Moreover, city pairs were chosen as more familiar than foils for 56% (SD=21.0%) of the time, which was again reliably above chance (50%) [$t(79)=2.46$, $p=.02$, $d=0.27$], suggesting that participants have successfully learned the city pairs, even though no cities were ever presented during exposure. However, park pairs were chosen as more familiar than foils for 54% (SD=20.8%) of the time, which was not reliably above chance (50%) [$t(79)=1.75$, $p=.08$, $d=0.20$].

During debriefing, eight participants reported noticing country pairs, but none correctly reported which specific countries followed each other, suggesting that participants had no explicit awareness of the country base pairs.

These results replicated those in Experiment 4, showing that participants successfully formed new associations at one subordinate level below the original superordinate level where regularities were learned, but not at two subordinate levels below. This again reveals a limit in how far new associations can be formed beyond the categorical level where objects were originally associated.

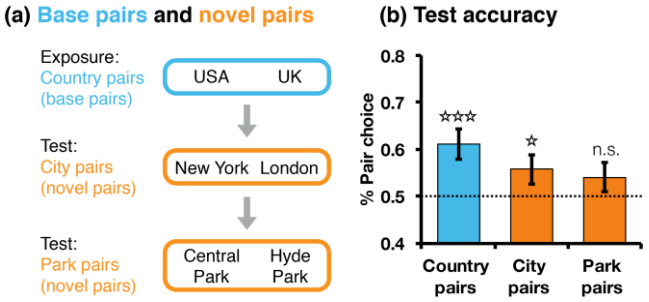


Figure 5: (a) Country pairs were base pairs. City and park pairs were novel pairs. (b) Percent of times the pair was chosen as more familiar (error bars reflect ±1 SEM; *p<.05, ***p<.001; dotted line=chance performance of 50%).

General Discussion

The goal of the current study was to examine whether statistical learning enables novel associations among objects that have never been directly associated before. We found that after experiencing that A predicted B and B predicted C, participants automatically and implicitly inferred that A predicted C, although C never directly followed A (Experiment 1). However, this inference was not successful when there were three pairs (e.g., A-B, B-C, C-D) presented during exposure, revealing a constraint in the extent of the transitive inference (Experiment 2). Extending beyond temporal transitivity, we examined if novel associations could be formed across the categorical hierarchy. We found that after learning a pair of objects at one categorical level (e.g., New York-London), participants automatically and implicitly inferred the same association at the subordinate level (e.g., Central Park-Hyde Park) and superordinate level

(e.g., USA-UK), even if the subordinate or superordinate objects were never presented or associated with each other (Experiment 3). However, this transitive inference seemed to be limited to the subordinate or the superordinate level immediately below or above the original level where the regularities were experienced, and did not extend to two levels beyond the original level (Experiments 4 and 5).

The current findings suggest that statistical learning can produce knowledge above and beyond the mere co-occurrences among objects that are directly observed. The base pairs support automatic generalizations across objects and category levels. This implies that our cognitive system can draw upon prior experiences to automatically form new relationships between objects through transitive inference.

Interestingly, both the knowledge about the co-occurring objects, and the knowledge about the new associations were completely implicit, since no participant could report which objects co-occurred in the experiments. This finding is consistent with the previous work showing that people can infer relational information between objects without awareness (Greene et al., 2001; Munnelly & Dymond, 2014).

While statistical learning allows the creation of new knowledge, it is important to understand the constraints in this process. The failure to learn A-D in Experiment 2 despite having successfully learned the base pairs A-B, B-C, and C-D could be explained by either an increase in the number of objects, the number of pairs, or the number of connections. The failure to infer country pairs despite having successfully learned the park pairs and inferred the city pairs in Experiment 4, and the similar failure in Experiment 5, could be explained by the conceptual distance in the categorical hierarchy. These constraints may be due to the capacity limit of visual short-term memory (Alveraz & Cavanagh, 2004; Luck & Vogel, 1997). Future studies are needed to identify the factors underlying these constraints.

The current study also extends beyond past work showing that people can learn categorical regularities from associations among individual exemplars (Brady & Oliva, 2008). We found that the regularities extracted at one categorical level can be transferred to subordinate and superordinate levels. This suggests that statistical learning not only operates at an abstract conceptual level, but also automatically propagates learned object associations across the categorical hierarchy.

The current study is significant in several ways. We found that statistical learning not only produces knowledge about existing relationships between objects, but also generates new associations between objects that are never directly associated or experienced. These new associations are automatically formed via transitive relations, either through temporal co-occurrences or across the categorical hierarchy, even in the absence of explicit awareness. Importantly, there are constraints in forming these transitive associations. Understanding the scope and the limits of this transitive inference can help reveal how the cognitive system creates new knowledge from prior experiences.

Acknowledgement

This work was supported by NSERC Discovery Grant (RGPIN-2014-05617 to JZ), the Canada Research Chairs program (to JZ), and the Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ).

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*, 106-111.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*, 487-502.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological Science, 19*, 678-685.

Conway, C. M., & Christiansen, M. H. (2005). Modality constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24-39.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.

Greene A.J., Spellman B.A., Dusek J.A., Eichenbaum H.B., & Levy W.B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. *Memory & Cognition, 29,* 893–902.

Kumaran, D., & Ludwig, H. (2013). Transitivity performance, relational hierarchy knowledge and awareness: Results of an instructional framing manipulation. *Hippocampus, 23,* 1259-1268.

Luck, S.J., & Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281.

Munnelly, A., & Dymond, S. (2014). Relational memory generalization and integration in a transitive inference task with and without instructed awareness. *Neurobiology of Learning and Memory, 109*, 169-177.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 399–407.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552–564.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21,* 1934–1945.

Wimmer, G., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science, 338*, 270-273.

Yu, R., & Zhao, J. (2015). The persistence of attentional bias to regularities in a changing environment. *Attention, Perception, & Psychophysics, 77,* 2217-2228.

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological Science, 24*, 667–677.

Zhao, J., & Yu, R. (2016). Statistical regularities reduce perceived numerosity. *Cognition, 146,* 217-222.

Information Search with Depleting and Non-Depleting Resources

Amber Bloomfield (anbloomfi@umd.edu)
J. Isaiah Harbison (isaiah.harbison@gmail.com)
Susan Campbell (susanc@umd.edu)
Petra Bradley (pscheck@umd.edu)
Lelyn Saner (lsaner@casl.umd.edu)
University of Maryland, 7005 52nd Ave
College Park, MD 20742 USA

Abstract

Predictions about information search behavior have been informed by extensive research in food foraging behavior. However, information foraging environments may differ in key ways from food foraging environments, and these differences may impact search behavior. We investigated the effect of patch distribution (depleting or non-depleting) and ability to return to previously searched patches on participants' decision to switch from one patch to another while searching. Whether or not a participant could return after leaving a patch led to fewer samples and fewer relevant items found. Whether or not the patches depleted and whether it was possible to return to a patch influenced stopping rules, indicating that these factors may alter the size of the increment applied through the Incremental Rule.

Keywords: information foraging, marginal value theorem, stopping rules, patch distribution

Introduction

How do information foragers decide where to look and when to look elsewhere? Research on information foraging has benefitted from an analogy to food foraging behavior (Constantino & Daw, 2015; Dennis & Taylor, 2006; Hills, Jones, & Todd, 2012; Hutchinson, Wilke, & Todd, 2008; Jacobs & Hackenberg, 1996; Kalff, Hills, & Wiener, 2010; Sandstrom, 1994; Wilke, Hutchinson, Todd, & Czienskowski, 2009). The Marginal Value Theorem (MVT), introduced by Charnov (1976), describes how animals foraging for food decide when to cease exploiting a current source to move on to another. In a situation of depleting returns (e.g., a berry bush has fewer berries to offer as a forager consumes berries from that bush), the MVT predicts that a forager will exploit a resource until the point at which the resource's rate of return (i.e., richness) has decreased to the mean return value of the environment. The point at which a forager will leave a resource (i.e., the stopping rule) is then determined by the richness of the resource and the cost of moving on to another resource (the switch cost). A strict interpretation of this rule means that foragers should not leave patches that remain richer than the mean return expected from moving on.

While the analogy to food foraging in animals has proven useful, there are a number of potential differences between the environments in which food and information foraging typically take place. First, in food foraging, a given source normally depletes (e.g., picking a berry results in fewer berries remaining), but depending on the information source, the number of items to search may be so large or replenished so frequently that functionally there is no depletion. For example, the number of tweets on a topic could be so numerous that the information forager is unlikely to read all of them, or the number may increase faster than they could be read as additional tweets are created. Search behavior in environments without depletion has not been investigated in previous studies. A related issue concerns search behavior in environments where it is possible to return to previously investigated patches. While returning to a previous patch is usually sub-optimal in food foraging, as that patch was only left because its richness had diminished beneath the expected return for the environment, a previously visited source may offer the highest richness in the environment if the source does not deplete. Search behavior in environments where it is possible to revisit a patch has not been investigated in previous research.

Stopping Rules

Because the MVT stopping rule requires that a forager have access to information about the richness of a patch at each moment as well as knowledge of the average richness for the environment, more computationally feasible rules have been proposed. The rule which has received the largest share of support in previous information search studies is the incremental rule. In applying this rule, the information forager begins with a set threshold of sampling attempts for a patch (e.g., 10 samples before moving on). When a relevant item is found, the threshold is increased by some increment. People tend to use the incremental rule both for internal search (e.g., memory search; Harbison, Dougherty, Davelaar, & Fayyad, 2009) and external search (e.g., searching in a library or on an online database; Wilke et al., 2009).

Application of the Incremental Rule can produce different switching behavior depending on the size of the increment the information forager uses. For instance, the total failures rule is a special case of the Incremental Rule where the threshold is incremented up by one search attempt for every relevant item found; in essence, samples that yield relevant items do not count against the sampling threshold. An

information forager applying the total failures rule will switch patches when the total number of samples not yielding relevant items (i.e., the number of failed samples) exceeds their threshold. Another special case of the Incremental Rule is the total samples rule. In this rule, the size of the increment is zero, such that an information forager will switch patches after a threshold number of samples, no matter how many relevant items are found. Although the total failures rule and the total samples rule are the most common variations of the Incremental Rule, others are possible. For instance, an information forager applying the rule with an increment of size 2 would stay longer in richer patches than a searcher applying either the total failures rule or the total samples rule because each encountered relevant item would increment the sample threshold up by two. These three cases of the Incremental Rule make increasingly similar predictions the poorer the search patch and make the same prediction for patches with zero target items (the case were there would not be an increment).

Predictions

Figure 1 shows the predicted relationship between return rate and number of relevant items found under different increment sizes for the two depletion conditions. These simulated results, like the experiment described below, include three levels of richness: 10%, 30%, and 50% relevant items.

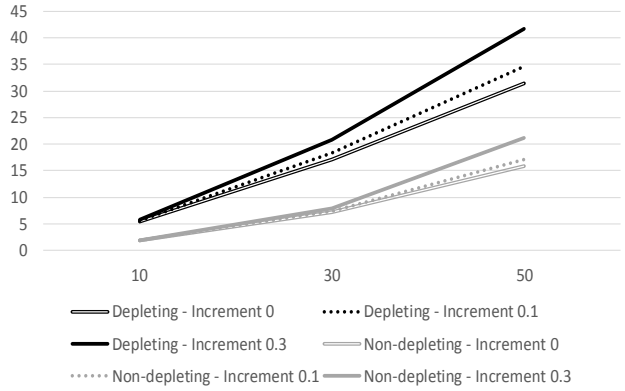


Figure 1: Predicted relationship between return rate and number of relevant items found by depletion condition and increment size.

If participants in our experiment apply the total samples variant of the Incremental Rule (i.e., the size of the increment is zero, represented by the gray and black double lines in Figure 1), they should find more relevant items as patch richness increases, regardless of whether the patches deplete. This is also true for the two other specific cases of the Incremental rule (increments of size 0.1 and 0.3). Therefore, the number of relevant items found does not make a clean distinction between different increment values. However, because of differences in the distribution of relevant items, it is predicted that participants in the Depleting condition will find more relevant items at each level of patch richness than participants in the Non-depleting condition (in Figure 1, the group of black lines are consistently higher than the group of gray lines).

The present experiment also allows us to measure the influence of irrelevant items on search termination in a way that has not previously been explored in the literature. The number of irrelevant items is not accessible to researchers when participants are retrieving from memory because the irrelevant items do not usually elicit a response, and the number of times such irrelevant items are output such that they can be observed is relatively rare (but see Unsworth et al., 2011). In the information search literature, paradigms such as anagram, word search, and fishing provide only a measurement of the time spent not finding relevant words or fish. With the current task, the number of irrelevant items found and their impact on search termination can be measured directly.

The predicted relationship between patch richness and number of irrelevant items found for each depletion condition with three different increment sizes is shown in Figure 2. In this case, there is expected to be a different form of relationship between number of irrelevant items found and return rate dependent on the increment the participant uses. A negative relationship between patch richness and the number of irrelevant items found is predicted if a participant applies the total samples rule (increment of 0). More time is spent sampling relevant items in the richer patches, and since the stopping threshold is not influenced by the number of relevant items found, these relevant items displace the irrelevant items that would be found in poorer patches. Given a larger increment, the relationship should become increasingly positive. Thus, if an information forager applies the total failures rule (increment of 1), a larger number of irrelevant items will be found as the return rate of the patch increases.

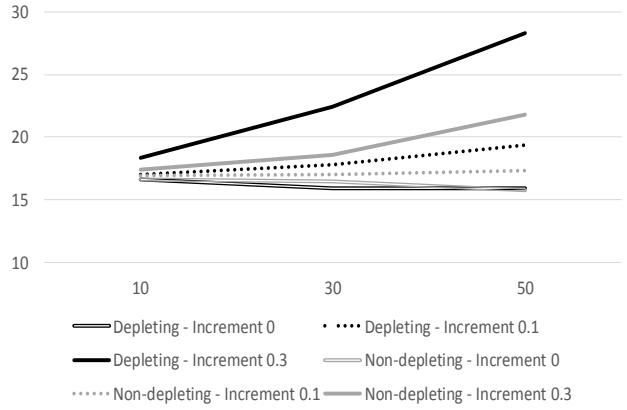


Figure 2: Predicted relationship between return rate and number of irrelevant items found by depletion condition and increment size.

In summary, a different pattern of irrelevant items found in patches of different richness levels will emerge

depending on the size of the increment the information forager uses. The predictions are the same for both depleting and non-depleting patches. However, it is possible that information search under conditions where resources do not deplete will lead participants to apply a different size of increment, and so show a different pattern in the relationship between irrelevant items found and patch richness. It may also be the case that whether or not it is possible to return to a patch will impact the increment used to determine when to stop exploiting a given patch. As noted above, previous research has not explored stopping rules when it is possible to return. One possibility is that information foragers will adopt a more conservative approach to stopping when they know they can return to a patch, potentially entailing a smaller increment size. The analyses will explore these possibilities.

Experiment

Search Task

The participants' search task was to collect information about clients for a matchmaking service wherein the participant would use the information found to judge the compatibility of a pair of clients. On each trial, participants saw the name of a client. During the search task, three different sets of tweets (i.e., brief messages stating a piece of information about a person) were offered for the participant to search. Each patch (of tweets) was described as having been produced by a different algorithm designed to gather tweets about the service's clients. Tweets could be about the client (relevant) or about another person (irrelevant).

The richness of the algorithm patches (i.e., the percentage of relevant tweets), could be 10%, 30%, or 50%. The richness levels available among the three patches in a trial varied across trials, with the restriction that all three patches could not have the same return rate. Thus, the highest level of richness available varied from trial to trial. In the depleting condition, the tweets in each of the three patches were ordered such that the rate of return was highest for the first few selections, but decreased as the number of selections increased; in the non-depleting condition, the rate of return remained constant throughout the patch. Patches in the depleting condition had the highest return rate in the first ten selections, and the rate of return decreased in successive sets of ten selections, with the average richness being 10%, 30%, or 50% if the participant sampled 100 times; patches in the non-depleting condition had a constant return rate of 10%, 30%, or 50% throughout.

For each tweet, participants were asked to indicate whether it was relevant for the target candidate or not relevant by clicking one of two buttons. After determining whether a given tweet was relevant to the target, the participant could either choose to view another tweet from the same algorithm or move to a different patch. The switch cost (i.e., the time required to switch to another patch) was 1 second. Depending on condition, participants were either able to return to a patch they had previously searched (return condition), and all three patches remained available throughout each trial, or a patch disappeared after the participant chose to leave the patch (no return condition).

Relevant tweets correctly identified as relevant by the participant were reflected in a running total at the top of the screen, labeled "Score"; participants were not penalized for incorrect decisions, though relevant tweets deemed irrelevant or irrelevant tweets deemed relevant did not increase the participant's score. For each target, participants had two minutes to search for information. At the end of each search, they moved on to the search task for a new target with a new set of patches.

Matching Task

After 20 search trials, each involving a different target person, participants completed a matching task. This task involved 10 trials; on each trial, the participants evaluated the compatibility of a pair of targets (matched based on the gender of match each target was seeking) using the relevant tweets found during the search task. Only relevant tweets the participant had judged to be relevant were available in this phase of the task. Participants indicated their compatibility judgment using an unlabeled slider (scored as a 0-100 scale) with "Horrible Match" and "Great Match" at the poles. The matching task was used to motivate the search task but data from this task were not analyzed.

Participants

Eighty-five participants were recruited via Amazon Mechanical Turk. "Workers" on Mechanical Turk receive monetary compensation for completing online tasks. Workers who selected this experiment task were informed that they would receive $10 for their participation in a task requiring about one hour of time.

Eight participants' data were removed from the sample due to low judgment accuracy, no switching decisions, performing the task twice, or not completing the task. The total analyzed sample, therefore, was 77 participants.

Procedure

Before beginning the experiment, participants were provided instructions and a practice session that included two search tasks and a single matching task. Participants were able to contact the researchers via email if they had questions or encountered difficulty during the experiment. At the end of the experiment, participants were thanked for their participation and paid via Amazon Mechanical Turk.

Participants were randomly assigned to one of the two search conditions: return and non-return. Participants were also, independently, randomly assigned to one of two patch distribution conditions: depleting or non-depleting. There were 16-20 participants in each of the four possible conditions.

Results

As shown in Table 1, the number of samples participants made varied considerably by patch, as did the number of relevant and irrelevant items found.

Table 1. Summary of visits aggregated over patches visited.

	Mean	SD	Min	Max
Total Visits	1.31	1.03	1	10
Total Samples	23.61	16.34	1	100
Total Relevant	9.81	9.37	0	49
Total Irrelevant	13.67	9.80	0	90

Most participants in the Can Return condition chose to return to a previously exploited patch at least once across the ten search trials (~88% of the 41 participants). As shown in Table 2, people were more likely to return in the Depleting condition than the Non-Depleting condition (return was not possible in the No Return condition).

Table 2. Mean visits per visited patch by condition. Standard deviation in parentheses.

	Can Return	No Return
Depleting	1.65 (1.42)	1.00 (0.00)
Non-depleting	1.47 (1.27)	1.00 (0.00)

The analyses in this section focused solely on the participant's first visit (for participants in the Can Return condition) to the first patch offered in each search set, to investigate how depletion condition and return condition influenced participants' search behavior. This focus on the first patch ensured that participants had ample time to investigate the patch prior to choosing to leave, rather than "stopping" as the result of running out of time in the trial. In addition, each analysis involved a multi-level model which estimated an intercept for each participant to account for any individual variability in search behavior.

The first model fit to the data predicted total samples based on return rate of the patch, depletion and return condition. Total number of samples from a patch was related to whether the participant could return ($B = 6.98$, $p < 0.05$), and the return rate ($B = 7.41$, $p < 0.05$ comparing 30% to 10% and $B = 18.56$, $p < 0.05$ comparing 50% to 10%), but not depletion condition ($B = -1.98$). The means for each condition are shown in Table 3.

Table 3. Mean samples per visited patch by condition. Standard deviation in parentheses.

	Can Return	No Return
Depleting	24.60 (20.89)	28.12 (15.92)
Non-depleting	19.77 (19.1)	28.97 (19.88)

Number of relevant items found was related to all three of the independent variables: return rate ($B = 7.04$, $p < 0.05$ comparing 30% to 10% and $B = 18.20$, $p < 0.05$ comparing 50% to 10%), depletion condition ($B = -4.84$, $p < 0.05$), and return condition ($B = 2.60$, $p < 0.05$). Average number of relevant items found in each condition are shown in Table 4.

Table 4. Mean relevant items found per visited patch by condition. Standard deviation in parentheses.

	Can Return	No Return
Depleting	13.54 (11.10)	14.72 (10.79)
Non-depleting	8.13 (9.51)	10.13 (10.16)

Only the return condition was significantly related to the number of irrelevant items found by participants ($B = 4.16$, $p < 0.05$). Participants in the No Return condition found more irrelevant items than those in the Can Return condition (Table 5).

Table 5. Mean irrelevant items found per visited patch by condition. Standard deviation in parentheses.

	Can Return	No Return
Depleting	12.39 (12.92)	13.57 (7.45)
Non-depleting	13.36 (11.34)	19.31 (11.73)

The relationship between number of relevant items found and depletion condition is expected given differences in the distribution of items in these conditions. Because depleting patches were front-loaded with relevant items and their rate of return diminished with repeated sampling, participants in the depleting condition found more relevant items, on average, than did the participants in the non-depleting condition. The significant relationship between return condition and number of relevant and irrelevant items found is more surprising. Participants consistently found fewer relevant and irrelevant items when they were able to return to a previously exploited patch. Taken together, the results suggest that deciding when to switch to a new patch or a previously visited patch, and performing a switch, cost the participants in the Can Return condition a significant number of relevant items and, because there was no penalty for finding irrelevant items, saved them nothing.

Switching behavior

In any given search trial, participants could either switch from the first patch in the set or continue exploiting that patch for the entirety of the two-minute search time. This "sticking and staying" behavior occurred on 9.8% of search trials. We explored how the return rate of the patch and the depletion and return conditions impacted whether or not participants stayed with a patch without exploring any other options in the search trial. Participants were more likely to stay with the first patch when the return rate of the patch was higher ($B = -2.57$, $p < 0.001$ comparing 30% to 10% and $B = -4.89$, $p < 0.001$ comparing 50% to 10%), but neither depletion nor return condition impacted likelihood of staying with the same patch ($B = -1.36$ and $B = 0.51$, respectively). Participants may have gotten a sense of the

highest available return rate over several trials and stayed with the first patch when they suspected this patch had the highest return rate possible.

Stopping rules

The impact of depletion and return conditions on average relevant and irrelevant items found suggests that the Can Return may have led to a less optimal search strategy than the No Return condition. However, the pattern of number of relevant and irrelevant items found across patches of different return rates can provide insight into the dominant increment used in the stopping rule across participants, and whether the increment changed due to return or depletion condition.

Figure 1 shows how the relationship between return rate and number of relevant items found changes depending on return and depletion condition. Consistent with the overall results regarding number of relevant items found, participants in the Depleting condition found more relevant items across all return rates and those in the No Return condition tended to find more relevant items regardless of return rate. Because all forms of the incremental rule predict a position relationship between return rate and the number of relevant items found, these results are consistent with the use of any size of increment.

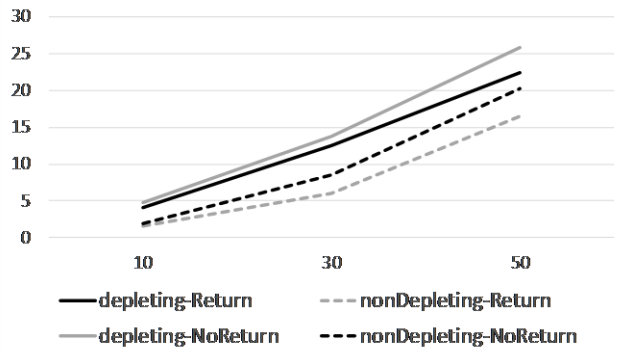


Figure 2. Average relevant items found for each return rate, by depletion and return condition.

Figure 4 shows the average number of irrelevant items found across the three return rates for each depletion and return condition. Across all return rates, participants in the Non-depleting – No Return condition found the most irrelevant items, and there is a slight but positive trend for number of irrelevant items found as return rate increases for this condition and for the Non-depleting – No Return condition. Comparatively, the slope of the Depleting – No Return and Depleting – Can Return appear slightly negative. This difference between the depletion conditions is intriguing because the application of the incremental rule is predicted to result in a different relationship between return rate and number of irrelevant items found depending on the size of the increment the information forager uses.

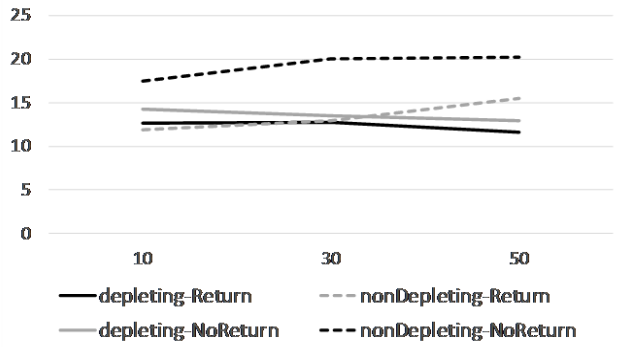


Figure 4. Average irrelevant items found for each return rate, by depletion and return condition.

To explore the dominant stopping rule used by participants in each condition, we fit a linear model predicting number of irrelevant items from patch return rate to each participants' data separately (for all patches, regardless of their position in the search set). Of the 77 participants, there was a non-significant relationship between the number of irrelevant items found and the return rate of the patch for 33. However, 42% of the sample showed a significant negative relationship between number of irrelevant items found and return rate of the patch. Only 14% showed a positive relationship, consistent with using a larger increment. Further, there was variation in the predominant relationship between return rate and irrelevant items found by condition: 63% of participants in the Depletion – No Return condition showed a negative relationship between patch return rate and irrelevant items found; this pattern was found for only 32-40% of participants in the other conditions. Thus, it seems that both whether the patches deplete and whether it is possible to return to previously exploited patches influenced the stopping rule used by participants.

Discussion

The current study explored whether the distribution of relevant items in a patch (depleting or non-depleting) and the ability to return to a previously exploited patch influenced the search behavior of information foragers. Overall, ability to return affected the average number of samples participants made from each patch: those participants who could return made fewer samples, and as a result found fewer relevant and irrelevant items. This suggests that the additional burden of deciding whether to leave a patch to return to a patch previously searched or a patch that has not yet been searched, and the accompanying time taken in switching patches, cut into the amount of time participants in this condition had to exploit any given patch. Because these participants found fewer relevant items compared to participants who could not return, this suggests that performing information search while weighing the ability to return to the current or previously explored patches may lead to suboptimal search. Even in cases where return to a patch is possible, as it is in most real-world

information search tasks, it may be advantageous for information foragers to leave each patch only when they are confident they will not wish to return to it.

Across all participants, the number of relevant items found increased with the richness of the patch. This pattern is to be expected if participants are applying the Incremental Rule while searching. Further, the relationship between number of relevant items found and richness of the patch differs only in degree between uses of an incremental stopping rule with different sizes of increments (see Figure 1). However, our study provided an opportunity to directly observe the relationship between number of irrelevant items found and richness of the patch which has not been possible in previous studies. Increments of different sizes are predicted to yield different relationships when the increment is 0, the relationship is predicted to be negative; as the size of increment increases, the relationship is predicted to become increasingly positive (see Figure 2). Of participants who showed a significant relationship between return rate and number of irrelevant items found, the majority showed a negative relationship. This is consistent with application of the total samples rule, which dictates an increment of 0. There were further differences between the conditions: a negative relationship between return rate and irrelevant items found was most prevalent in the Depleting – No Return condition, more common than any other pattern, while other conditions yielded a greater number of positive or non-significant relationships.

Conclusion

Previous research on information foraging has not explored situations where the proportion of targets or relevant items in a patch does not deplete, or cases where it is possible to return to a previously searched patch, yet both scenarios are possible in real-world search tasks. Our results indicate that these factors influence the stopping rules that information foragers apply. Further, our results suggest that having awareness of the ability to return to previously searched patches may lead to less optimal search behavior.

Acknowledgements

This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

References

Charnov, E. L. (1976). Optimal foraging, the Marginal Value Theorem. *Theoretical Population Biology, 9*, 129-136.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, 1-17.

Dennis, A. R., & Taylor, N. J. (2006). Information foraging on the web: The effects of "acceptable" Internet delays on multi-page information search behavior. *Decision Support Systems, 42*(2), 810-824.

Frank, R. G. (2004). *Behavioral economics and health economics* (No. w10881). National Bureau of Economic Research.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *The Journal of Political Economy*, 279-304.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition, 111*(3), 397-402.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: can a generalist adapt its rules to dispersal of items across patches?. *Animal Behaviour, 75*(4), 1331-1349.

Jacobs, E. A., & Hackenberg, T. D. (1996). Humans' choices in situations of time-based diminishing returns: Effects of fixed-interval duration and progressive-interval step size. *Journal of the Experimental Analysis of Behavior, 65*(1), 5-19.

Kalff, C., Hills, T. T., & Wiener, J. M. (2010). Human foraging behavior: A virtual reality investigation on area restricted search in humans. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 168-173).

Parsons, G. R., & Kealy, M. J. (1995). A demand theory for number of trips in a random utility model of recreation. *Journal of Environmental Economics and Management, 29*(3), 357-367.

Sandstrom, P. E. (1994). An optimal foraging approach to information seeking and use. *The Library Quarterly*, 414-449.

Thorelli, H. B., & Engledow, J. L. (1980). Information seekers and information systems: a policy perspective. *The Journal of Marketing,* 9-24.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Factors that influence search termination decisions in free recall: An examination of response type and confidence. *Acta psychologica, 138*(1), 19-29.

Wilke, A., Hutchinson, J., Todd, P. M., & Czienskowski, U. (2009). Fishing for the right words: Decision rules for human foraging behavior in internal search tasks. *Cognitive Science, 33*(3), 497-529.

Conceptual Expansion During Divergent Thinking

Richard W. Hass (hassr@philau.edu)
College of Science, Health, and the Liberal Arts, Philadelphia University, 4201 Henry Avenue, Philadelphia, PA, 19144, USA

Abstract

Recent research on creative thinking has implicated conceptual expansion as potential cognitive underpinnings. These theories were examined within the context of a laboratory study using two divergent thinking prompts. Participants generated alternative/creative uses for a brick and for a glass bottle (separately) for two minutes and responses were time-stamped using a Matlab GUI. Semantic distances between responses and conceptual representations of the DT prompts were computed using latent semantic analysis. Results showed that semantic distance increased as responding progressed, with significant differences between the two tasks, and intraparticipant variation. Results have implications for theories of creative thinking and represent methodological and analytic advances in the study of divergent thinking.

Keywords: creativity; semantic distance; latent semantic analysis; divergent thinking; conceptual expansion

Conceptual Expansion During Divergent Thinking

Divergent thinking is a problematic topic in the study of creativity for many reasons. One issue is that divergent thinking (DT) refers to both a psychometric construct—thinking in multiple directions—and the set of tasks used to quantify the construct (for a full discussion see Hass, in revision). Perhaps because of DT's psychometric roots, cognitive analysis of creative thinking often omits reference to DT studies (e.g., Finke, Ward, & Smith, 1992; Weisberg, 2006). However, recently there has been a surge of neuroscientific studies using DT as a proxy for creative thinking, citing among other points that DT tests have some predictive validity for real world creative success (Kim, 2006). Despite criticisms of DT as a means to assess creativity (e.g., Weisberg, 2006, Ch. 9), if they are to be used in neuroscientific studies, then cognitive theories must be developed to explain what is transpiring *during* DT. This paper represents one part of a larger project to try to do just that. With processing data in hand, interpretation of neuroscientific studies of DT will become much more straightforward and useful for a cognitive science of creative thinking.

Cognitive explanations for DT performance

Many neuroscientific studies using DT as a proxy have shown that originality on DT tasks is related both to the brain's cognitive control and default mode networks (for a review see Beaty, Benedek, Kaufman, & Silvia, 2015). The main conclusion drawn from these studies is that better control of self-directed thought defines improved performance on DT tasks. However, most studies are correlational in that they do not ask how cognitive control operates during DT, rather, cognitive control is assessed on a separate task and correlated to DT performance. For example, Zabelina, Saporta, and Beeman (2015) showed that DT performance was positively related to how well participants overcame an invalid cue that preceded 20% of trials on the Navon (1977) Local-Global Letter Task. However, there was no relation between DT performance and attention filtering (assessed in terms of a congruency effect on the Letter Sets Task). Moreover, these DT-attention relationships did not match effects relating individual differences in attention to individual differences in real-world creative achievement. The main question asked by the current analysis is whether tracking cognition during DT response generation can shed more light on these kinds of conflicting results.

Only a single study attempting to track cognition during DT exists. Gilhooly, Fioratu, Anthony, and Wynn (2007) took verbal protocols from participants and found that they often invoked distinct strategies during DT. For example, in thinking of alternative uses for a shoe, many participants engaged in self-cuing (repeating the word "shoe"), and reconstructed the problem representation by mentally disassembling the object (i.e., using only the laces of the shoe). So it seems that there may be several levels of cognitive processing operating during DT, and it is imperative that we move toward studies that quantify those processes, rather than rely purely on correlational data. Though the current analysis did not involve verbal protocols, the next section outlines a cognitive framework for the kind of data that were colletted.

Conceptual expansion. Ward (2008) described conceptual expansion as the formation of novel exemplars of a concept during [creative] problem solving. Indeed, Abraham (2014)—in her theoretical examination of the conceptual expansion hypothesis—used a DT task (think of alternative / creative uses for a shoe) as her primary example of conceptual expansion. She argued that envisioning the use of a shoe as a plant pot or as a pencil holder, by definition, expands upon the canonical concept of shoe. Abraham and colleagues (2012) showed evidence of differential brain activity when participants generated common versus unusual responses to DT prompts (see also Chrysikou & Thompson-Schill, 2011). Given Ward's definition of conceptual expansion, they reasoned that this additional activation was evidence of a conceptual expansion processes during "unusual" idea generation.

There are many reasons that conceptual expansion represents a good theoretical framework for creative thinking. Particularly, it allows for research to focus on specific questions regarding the process. For example, does conceptual expansion unfold in a linear fashion? Do people actively monitor the amount of expansion in the responding? Is expansion related to processing speed? Is expansion another way to describe analogical transfer?

Before answering these questions it is important to settle on an operational definition of conceptual expansion. In this analysis, conceptual expansion was operationally defined as the degree of semantic distance between DT responses and the prompt (e.g., think of alternative uses for a brick). Semantic distance was derived from cosine similarity scores obtained via latent semantic analysis (LSA, e.g., Landauer & Dumais, 1997). Though LSA is not a one-to-one mapping of conceptual expansion, it is of interest to examine relationships among semantic distance and response order and inter-response time.

Serial order, response time, and semantic distance

The so-called serial order effect has been described in many studies showing that generally people provide more creative responses to DT prompts later in response array's (e.g., Beaty & Silvia, 2012; Christensen, Guildford, & Wilson, 1957). Beaty and Silvia found that the originality of DT responses (scored with a subjective system) increased as a function of response order, but that participants with higher fluid intelligence scores began with more creative responses during DT than participants with lower fluid intelligence scores, and showed less of an increase. Hass (in revision) replicated the analysis using semantic distance and growth-curve modeling and showed that high fluid intelligence scores related to higher initial semantic distance during DT.

Though the serial order effect seems to be well established, a cognitive explanation is less clear. If it is the case that associative processes spur recall of the concepts that map onto DT responses, then response latencies should be related to the distance between the conceptual content in the response and the conceptual content in the DT prompt (e.g., Kahana, 1996). This hypothesis is directly tested in the current study.

Is LSA a valid means of measuring conceptual expansion?

Before describing the method and results, it is important to discuss the validity of LSA-derived semantic metrics in DT studies. Hass (in revision) provided a discussion of the use of LSA-derived distances in scoring DT responses as opposed to other semantic methods (see also Harbison & Haarmann, 2014). The crux of the argument was that if the distances are culled from comparisons between each response and a fixed conceptual representation of the DT prompt (e.g., *brick*), then the metric has both construct validity and convergent validity with subjective scores. This approach is similar to more traditional DT flexibility scoring[1] (e.g., Madore, Addis, & Schacter, 2015), but with the added benefits of a continuous scale of measurement and the availability of computational models. Both techniques target persisting themes in cognitive theories of creative thinking: remote association (e.g., Mednick, 1962) and conceptual expansion. However, since flexibility scores rely on the creation of ad-hoc categories after data is collected, the system is potentially biased and also provides low-resolution information regarding the graded structure of categories (cf., Gabora, Rosch, & Aerts, 2008).

Also, unlike the current study, prior analyses of DT data with LSA have seemed focused on *replacing* subjective scoring with semantic scoring, which essentially keeps DT tied to the psychometric "summary score" approach. For example, Forster and Dunbar (2009) showed that LSA-derived semantic distance scores from DT data were correlated with originality ratings, and that since distances are objectively calculated, they may be preferred to subjective scoring. Harbison and Haarmann (2014) similarly showed that subjective scores and distances correlated, though they also showed that another natural language processing procedure (point-wise mutual information) was more highly correlated with subjective scores.

Rather than persist with the summary score approach, this analysis used Growth Curve Modeling (e.g., Mirman, Dixon, & Magnuson, 2007) to examine individual differences in the serial-order effect, and to examine relationships between serial order, inter-response time, and semantic distance (as a proxy for conceptual expansion). Variations in semantic distance within individuals were also examined across two oft-used DT prompts: think of creative alternative uses for a brick, and for a glass bottle. It was expected that responses would increase in distance as a function of response order, and also that there would be a linear relationship between distance and IRTs. Prior analyses also revealed differing levels of semantic distance across DT prompts (Hass, in revision), so that analysis was also performed.

Method

Participants

Sixty participants (18 females) were recruited from the participant pool at a large state college in New Jersey. The average age of participants was 19.45 years (SD = 1.46). All participants were given partial course credit for participating. Participants provided informed consent prior to participation. Time-stamp malfunctions led to the elimination of data from three participants.

Materials

All materials were presented on Lenovo ThinkVision monitors. Participants typed responses on computer keyboards. The experiment was automatically administered

[1] Flexibility is defined as the number of category switches in a response array.

using a custom Matlab GUI which provided an editable response field for participants to enter responses. Matlab timestamped both the initial keypress for each response and the final return key. Pressing return cleared the response from the response field and pasted it below to keep a running log of the participants' responses so he or she would be encouraged to continue producing novel responses. Prompts appeared in 50-point font and were visible throughout response generation. Responses appeared in 36-point font.

Procedure

Participants were greeted by an experimenter and filled out a demographic survey while the experimenter initiated Matlab. Instructions appeared on the screen and the experimenter read aloud to the participant the initial instructions regarding using the keyboard to enter responses. Participants then engaged in a short category generation task (30s of naming colors) to grow accustomed to the experimental setting. After that, participants were randomly assigned to two un-related task conditions that lasted 5 minutes[2]. Finally, participants were presented with the instructions for the DT tasks. Participants were told to think of creative uses for common objects that would be presented in text on the screen. They were told that there would be two such tasks and that they would have 2 minutes to complete each task.

The task prompt then appeared above the response field, with the order of the two prompts (brick, glass bottle) randomized by Matlab. The prompts read "Think of uses for a Brick *besides* building a wall" and "Think of uses for a Glass Bottle *besides* holding liquid". These instructions were designed to increase the validity of the semantic analysis using the canonical concept of brick as a building material and bottle as a liquid holder. Participants were instructed to continue responding until time had expired. When the two minutes per task expired, Matlab displayed a message to indicate that the next task was loading. The inter-task time was 10 seconds to allow for a brief break. After completion of the second task, a thank you message appeared on the screen.

Results

Data preparation and semantic analysis

LSA was performed using the tools available at lsa.colorado.edu. Analysis was performed using the data from the TASA corpus, compiled to represent general semantic knowledge gained from primary school through the first year in college. Three hundred factors were used, in keeping with prior analyses that used this tool (e.g., Forster & Dunbar, 2009). Prior to LSA, all responses were spell-checked, and a set of stopwords was removed using functions from the *tm* package (Feinerer, Hornik, & Meyer, 2008) in the R Statistical Programing Environment (R Core Team, 2015).

The "one-to-many" LSA tool was used to compare each DT response from the dataset to a target phrase—a composite description of the DT prompt compiled from Merriam Webster Dictionary entries (see Hass, under review) in document space. The phrase representing the *brick* concept was, "a small, hard block of baked clay that is used to build structures such as houses and sometimes to make streets and paths" (see http://www.merriam-webster.com/dictionary/brick). A similar phrase was used for the glass bottle comparison. LSA represents phrases as the centroid of the word vectors contained in the phrase. The centroid is essentially a vector average, and thus represents a sort of blend of the meanings of the words in each response. This method of representation has been shown to work well for long-passages of text such as student essay responses (e.g., Rehder, et al., 1998).

For each response, the LSA tool computed the cosine of the angle between the vector representing the target (the DT prompt) and the vector representing the response. This represents the *similarity* of two vectors, such that the cosine of the angle between two identical vectors is 1, the cosine of two orthogonal (i.e. unrelated) vectors is 0, and the cosine of two vectors pointing in opposite directions is -1. The cosine similarity values were then transformed into to *distances* by subtracting each from 1 (e.g., Prabhakaran, Green, & Gray, 2013).

Table 1: Descriptive statistics by DT prompt. Inter-response times, and distances were analyzed at the level of response. Fluency was analyzed at the level of participant.

Prompt	Variable	M	SD	s.e(M)
Brick (n_{resp} = 402)	IRT	13.48	12.73	0.64
	Distance	0.88	0.14	0.01
	Fluency	7.04	3.09	0.40
Bottle (n_{resp} = 393)	IRT	13.62	12.31	0.62
	Distance	0.78	0.17	0.01
	Fluency	6.89	2.88	0.38

Inter-response time (IRT) was calculated as the difference in end-of-response time stamps between adjacent pairs of responses. IRT for the first response was defined simply as the time stamp of the first response. Table 1 provides descriptive statistics for IRTs and distances along with average fluency counts for each task.

Statistical Analysis

Inter-response time. Before examining a multilevel model for semantic distance, the relationship between IRTs and response order was examined with a simple correlation. The correlation was small but significant ($r(793) = .18$, $p < .001$). In addition to showing that participants took more time to respond as their 2 minutes on task elapsed, the small magnitude of the correlation means that response order and

[2] Antecedent task condition had no effect on of the results reported in this paper

IRT can be used in a linear model for semantic distance without collinearity issues.

Semantic Distance. A multilevel model for semantic distance was assembled because of the variation in fluency across participants, and to test for possible variation in the relationship between response order and distance across participants. Model testing followed procedures given by Mirman, and colleagues (2007). The significance of predictors and random effects was determined by comparing nested models with a likelihood ratio. For all models, semantic distance was the dependent variable, with response order, and IRT as level-1 predictors. Response order was rescaled with zero as the first response so that the intercept, and IRT was rescaled in grand-mean-deviation form. DT prompt was entered as a level-2 predictor.

Table 2 summarizes the various models compared in terms of model deviance (Mirman, et al., 2007), with significant differences identified as statistically significant likelihood ratios. Model 1 is a baseline linear growth model. The response-order coefficient was significant[3] (β_{11} = 0.01, 95%CI = (0.006, 0.013)), confirming an overall linear serial order effect. Model 2 examined potential nonlinearity in the response order effect. The comparison narrowly missed significance, suggesting that there was an inverted-U trend to the data, but that this did not explain much more of the variance in distances across responses than the linear response-order predictor. The addition of the IRT variable also did not improve the fit (comparison 3). So the best level-1 growth model for semantic distance is defined with a linear response order predictor, and a random intercept per participant.

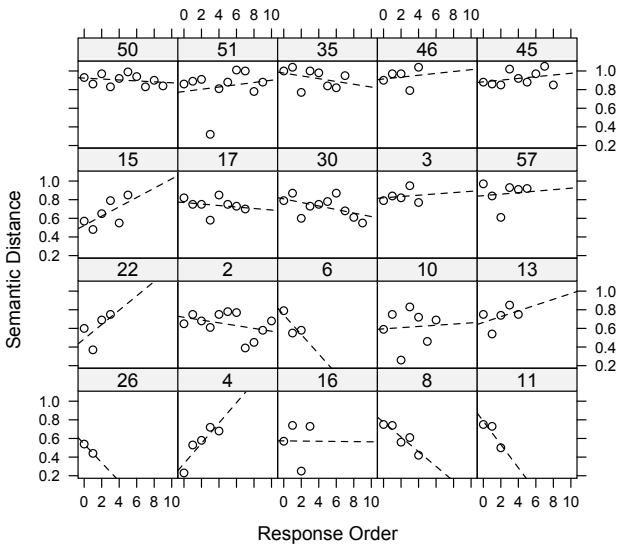
Table 2: Results of Semantic distance model testing. Model 1 is nested in Model 2. Models 3 and 4 are nested in Model 5.

Comparison	logLik	ΔD	*p*
1. Order (linear)	349.13	-	-
2. Order (quadratic) v. 1	350.89	3.53	.06
3. IRT v. 1	350.55	2.84	.09
4. Prompt (intercept) v. 1	405.61	112.96	< .001
5. Prompt (slope) v. 4	455.24	99.26	< .001
6. Prompt x Order v. 5	459.61	8.74	.003

Comparisons 4 through 6 in Table 2 represent tests to determine whether participants' distances varied according to the DT prompt (brick v. glass bottle), and whether the serial order effect varied across prompts and across individuals. These comparisons represent the key contribution of growth-modeling as they allow for examination of variations across individuals.

Comparison 4 shows that there was a significant difference in the average semantic distance of first responses given by participants across prompts. The

a



b

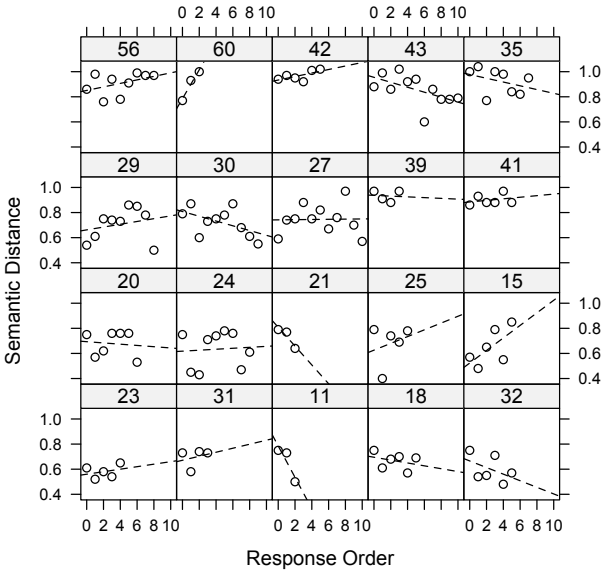


Figure 1: Variations in the serial-order effect across 20 randomly sampled participants. (a) responses to the Alternative Uses for a Brick prompt only, (b) responses to the Alternative uses for a Glass Bottle prompt only. Dotted lines represent OLS regression within participant.

semantic distances of initial responses were significantly lower on the bottle task compared to the brick task (γ_{10} = -0.11, 95%CI = (-0.13, -0.09)). This suggests that semantic distance is context dependent and so conceptual expansion may not be a general ability.

Finally, Figure 1 illustrates the results of comparisons 5 and 6, for 20 randomly sampled participants. There was both significant variation participants' changes in distance across the prompts, and a significant amount of variation in

[3] The use of *p*-values for evaluating coefficients in Multilevel models is controversial so 95% confidence intervals are reported.

participants' serial order effects. Thus, the serial-order effect may not be a universal phenomenon. Rather, semantic distance is a function of the conceptual content in the DT prompt, and though on average, semantic distances between responses and prompts tend to increase, inter-participant variability remains to be explained.

Discussion

The central aim of this paper was to examine conceptual expansion during divergent thinking and relate it to response order and IRT. Conceptual expansion was operationally defined in terms of the semantic distance between the concept represented in the DT prompt and a particular response. Several interesting results emerged, which in turn lead to new questions about DT, creative thinking, conceptual expansion, and response latency.

First, response latency did not directly relate to conceptual expansion. Rather, the degree of conceptual expansion shown by participants was more dependent upon both the concept represented by the DT prompt (brick v. glass bottle) and likely individual differences in semantic memory organization. This is consistent with other evidence that individual differences in semantic memory organization relate to individual differences in creative thinking and creative accomplishments (e.g., Kenett, Anaki, & Faust, 2014; Kenett, Beaty, Silvia, & Anaki, 2016). In those studies, network analysis was applied to category fluency responses from participants, rather than to DT responses, but the interconnectedness and flexibility of participants' semantic networks did indeed correlate with DT performance. Taken together with the verbal protocol analysis performed by Gilhooly and colleagues (2007), it seems that DT performance varies along with variations in participants' semantic processing, and likely according to their retrieval and cuing strategies, though verbal protocols were not taken from these participants.

Interestingly, the serial order effect does not seem to be a cognitive universal, nor does it seem that participants always need more time to come up with more distant responses. This is somewhat inconsistent with the remote-association account of creative thinking forwarded by Mednick (1962, see also Beaty, Silvia, Nusbaum, Jauk, & Benedek, 2014). According to Mednick, more creative people should generate more ideas when prompted, and the associations among ideas should be looser than less creative people (see also Wallach & Kogan, 1965). There are many problematic aspects of the theory for which the current study has implications. First, the variations in distances across prompts and within participants suggests that associative processes vary substantially according to the conceptual context and individual knowledge. That is, we should not assume that people should approach all creative idea generation tasks with the same amount of knowledge, or the same potential to expand on such knowledge.

Second, there may be two conceptual expansion processes operating during DT. In an analogous study, Smith, Vul, and Huber (2013) used LSA-derived semantic similarity to show that adjacent responses in a modified remote associates task (RAT) were semantically dependent. Performance on RAT items are often used to simulate creative insight (e.g., Kounios and Beeman, 2009). Smith and colleagues also argued that the search process is conscious given that responses are sequentially dependent. Though potential local dependencies were not examined in this paper, it is very likely that semantic distances between adjacent responses will illustrate some degree of dependence. If so, it may be evidence of both global and local conceptual expansion processes operating during DT. A global process might monitor the overall conceptual expansion with the DT prompt as the basis for comparison, while a local process might monitor the expansion needed for the next iteration compared with the previous response iteration. These data are amenable to such analysis, and I encourage others to investigate these local-global monitoring phenomena. If they exist, they provide context for the effects described by Zabelina and colleagues (2015) that multiple levels of attention and monitoring are differentially related to creative thinking and creative achievement.

Limitations. It should be noted that the use of LSA-derived distances as a conceptual-expansion metric is limited to the validity of the TASA corpus for representing DT responses. Indeed, there were a few cases in which responses (e.g., smartphone) were not found in the corpus, and responses had to be discarded. Also, though the corpus is able to resolve ambiguity in word meaning through co-occurrence data, there are likely places in which creative wordplay (e.g., use of a brick as "the weight of life") that might yield invalid LSA cosines.

Despite these limitations, this analysis stands as the first step toward understanding how people approach creative thinking tasks like these DT problems from the perspective of cognitive science. Continued examination creative thinking data using semantic distance and other related techniques, couched in growth curve models is highly recommended. Among other issues, this type of analysis is likely to address some of the inconsistencies in creative thinking study results when DT summary scores are correlated with measures of cognitive processing. It is clear that variations in people's semantic knowledge and possibly their ability to monitor progress during creative idea generation is a key factor in explaining how DT unfolds.

Acknowledgments

I would like to thank Matthias Benedek and Emannuel Jauk for providing the basis of the Matlab code used for data collection.

References

Abraham, A. (2014). Creative thinking as orchestrated by semantic processing vs. cognitive control brain networks, *Frontiers In Human Neuroscience, 8,* 1–6.

Abraham, A., Pieritz, K., Thybusch, K., Rutter, B., Kröger, S., Schweckendiek, J., Stark, R., Windman, S., &

Hermann, C. (2012). Creativity and the brain: Uncovering the neural signature of conceptual expansion. *Neuropsychologia*, *50*(8), 1906–1917.

Beaty, R. E., Benedek, M., Kaufman, S. B., & Silvia, P. J. (2015). Default and executive network coupling supports creative idea production. *Scientific Reports*, 1–14. http://doi.org/10.1038/srep10964

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(4), 309–319.

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197.

Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology, 53*, 82–88.

Chrysikou, E. G., & Thompson-Schill, S. L. (2011). Dissociable brain states linked to common and creative object use. *Human Brain Mapping, 32*, 665–675.

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5): 1-54.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications.* Cambridge, MA: MIT Press.

Forster, E. A., & Dunbar, K. N. (2009). Creativity evaluation through latent semantic analysis. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the cognitive science society* (pp. 602–607). Austin, TX: Cognitive Science Society.

Gabora, L., Rosch, E., & Aerts, D. (2008). Toward an Ecological Theory of Concepts. *Ecological Psychology*, *20*(1), 84–116.

Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, *98*, 611–625.

Harbison, J. I., & Haarmann, H. (2014). Automated scoring of originality using semantic representations. In *Proceedings of cogsci 2014*. https://mindmodeling.org/cogsci2014/papers/405/index.html.

Hass, R. W. (in revision). Tracking the dynamics of divergent thinking via semantic distance. *Memory & Cognition.*

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109.

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons, 1–16. http://doi.org/10.3389/fnhum.2014.00407/abstract

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and Flexibility: Investigating the Relation Between the Structure of the Mental Lexicon, Fluid Intelligence, and Creative Achievement. *Psychology of Aesthetics, Creativity, and the Arts*, 1–14.

Kim, K. H. (2006). Can We Trust Creativity Tests? *Creativity Research Journal*, *18*(1), 3–14.

Kounios, J., & Beeman, M. (2009). The Aha! Moment: The Cognitive Neuroscience of Insight. *Current Directions in Psychological Science*, *18*(4), 210–216.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Madore, K. P., Addis, D. R., & Schacter, D. L. (2015). Creativity and Memory: Effects of an Episodic-Specificity Induction on Divergent Thinking. *Psychological Science*, *26*(9), 1461–1468.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*, 220-232.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. http://doi.org/10.1016/j.jml.2007.11.006

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353-383.

Prabhakaran, R., Green, A. E., & Gray, J. R. (2013). Thin slices of creativity: Using single-word utterances to assess creative cognition. *Behavior Research Methods* (online first) http://doi.org/10.3758/s13428-013-0401-7

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*(2-3), 337–354.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75.

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity–intelligence distinction*. New York: Holt, Rinehart, & Winston.

Ward, T. B. (2008). The role of domain knowledge in creative generation. *Learning and Individual Differences*, *18*(4), 363–366.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts.* Hoboken, NJ: John Wiley.

Zabelina, D., Saporta, A., & Beeman, M. (2015). Flexible or leaky attention in creative people? Distinct patterns of attention for different types of creative thinking. *Memory & Cognition*, 1–12. http://doi.org/10.3758/s13421-015-0569-4

Evaluating Causal Hypotheses: The Curious Case of Correlated Cues

Bob Rehder (bob.rehder@nyu.edu)
Zachary Davis (zach.davis@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003 USA

Abstract

Although the causal graphical model framework has achieved considerable success accounting for causal learning data, application of that formalism to multi-cause situations assumes that people are insensitive to the statistical properties of the causes themselves. The present experiment tests this assumption by first instructing subjects on a causal model consisting of two independent and generative causes and then requesting them to make *data likelihood judgments*, that is, to estimate the probability of some data given the model. The correlation between the causes in the data was either positive, zero, or negative. The data was judged as most likely in the positive condition and least likely in the negative condition, a finding that obtained even though all other statistical properties of the data (e.g., causal strengths, outcome density) were controlled. These results pose a problem for current models of causal learning.

Hypothesis testing occupies a central role in learning theory. On this view, learners use observed data to update their beliefs about different possible models of the world. A critical component of this process are learners' judgments regarding how probable, or improbable, it is that the observed data were generated by each of the hypotheses. In this paper we consider what factors affect learner's judgments regarding the likelihood that data was generated a particular *causal* hypothesis.

For example, consider the situation where there are two potential causes (C_1 and C_2) of an effect E. Fig. 1 shows the four hypotheses, or *graphs* (G), formed by crossing the presence/absence of $C_1 \rightarrow E$ with the presence/absence of $C_2 \rightarrow E$. Evaluating the posterior probability of these graphs involves calculating the probability of the observations, or the data (D), were generated by each graph, $p(D|G_i)$, and then applying Bayes' law. Indeed, Carroll, Cheng, and Lu (2013) adopted the hypothesis testing framework shown in Fig. 1 to account for learning data from some traditional associative learning paradigms involving two cues. While Griffiths and Tenenbaum (2005) initially developed the hypothesis testing methodology to account for learning data from simpler situations involving just *one* potential cause (also see Lu et al., 2008; Meder et al., 2014), it has since been extended to multi-cause situations (e.g., three potential causes in Powell et al., 2016).

Our purpose in this article is to highlight what we find to be an interesting property of how these models calculate the probability that data D were generated by a particular causal graph G. To do so we will present a modified version of the notation presented in Carroll et al. (2013). For generative causes, the likelihood of the data D under a particular parameterization of graph G was defined as

$$p(D|\boldsymbol{w}, G) = \prod_{e,\boldsymbol{c}} p(e|\boldsymbol{c}, \boldsymbol{w}, G)^{N(e,\boldsymbol{c})} \quad (1)$$

where $\boldsymbol{c}$ is a vector denoting the presence ($c_i = 1$) or absence ($c_i = 0$) of the cues and $N(e, \boldsymbol{c})$ gives the frequency counts for each combination of the presence/absence of the effect and the cues in D. By denoting the causal strength of the causal relations, the vector $\boldsymbol{w}$ represents the parameterization of G. The probability of the effect was defined as

$$p(e = 1|\mathbf{c}, \boldsymbol{w}, G) = 1 - \prod_{i \in I(G)} (1 - w_i)^{c_i} \quad (2)$$

where $I(G)$ is the set of indices corresponding to the causal links that are present in G (i.e., $I(G_0) = \{\}$, $I(G_1) = \{1\}$, $I(G_2) = \{2\}$, $I(G_3) = \{1,2\}$). Eq. 2 codifies the common assumption that each causal relation operates independently and that multiple causal influences combine according to a noisy-or integration rule (Cheng, 1997). Note that it is also common to include an additional causal strength parameter, w_0, that represents the probability that E might be caused by something other than C_1 and C_2.

Calculating $p(D|G_i)$ from Eqs. 1 and 2 requires integrating over possible causal strengths (the ws), which Carroll et al. assumed were uniformly distributed. Finally, calculating the posterior probability of the graphs, $p(G_i|D)$, requires stipulating their prior probability, which were assumed to be equal.

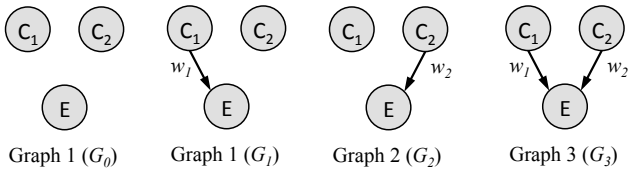


Figure 1

Although Eqs. 1-2 may appear to compute "the likelihood of the data given a model," in fact there is what might be considered an omission—no consideration is given to whether the data is consistent with the *base rate* of the causes stipulated by the models. This omission manifests itself in the absence of a parameter representing the base rates of the causes. Moreover, not only are the base rates of the causes not considered, neither is any *correlation* between C_1 and C_2. Stated more generally, in all these cases the computation of $p(D|G_i)$ is insensitive to any assumptions that learners might have regarding the statistical properties of the causes themselves. Such models are referred to as *conditional Bayesian networks* because they encode the distribution of a subset of variables (in our case, E) given their par-

ents (the Cs) but not the distribution of the parents themselves (Koller & Friedman, 2009).

Of course, a theorist might reasonably assume that when evaluating causal hypotheses it is natural for learners to focus on the manner in which causes co-vary with an effect rather than on the statistical properties of the causes themselves. Here we present data suggesting that this assumption is false. We do this by manipulating, in a two-cue learning scenario like that in Fig. 1, a statistical property of the two causes while holding their relationship with the effect E constant. In particular, we show that people's explicit judgments of $p(D|G_3)$ vary depending on whether the correlation between the two causes is positive, negative, or, zero.

Why Cue Correlations Might Matter

What reason might there be to think that causal learners expect cues to be correlated? Evidence for this view comes from a study of causal *reasoning* (rather than learning) reported by Rehder (2014a). Subjects were taught artificial causal relations intended to be plausible. For example, the causal structure represented by G_3 in Fig. 1 was instantiated in the domain of economics by telling subjects that "low interest rates causes high retirement savings" ($C_1 \rightarrow E$) and that "small trade deficits causes high retirement savings" ($C_2 \rightarrow E$). After learning these materials subjects were presented with 3AFC trials in which they were asked to decide which of two economies were more likely to possess a particular variable (a third "equally likely" alternative was also available). For instance, Fig. 2 depicts two economies in which the subject is being asked to predict which is more likely to have C_2 (in the example, small trade deficits). In the economy on the left of Fig. 2, $C_1 = 1$ (interest rates are low) is given; in the economy on the right, $C_1 = 0$ (interest rates are high). No information about the state of E was provided in either economy. Subjects were significantly more likely to choose the economy on the left, that is, they behaved as if they believed the two causes were positively correlated (also see Perales et al., 2004).

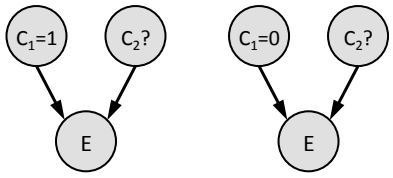


Figure 2

Formally speaking, this result represents a violation of the independence relations entailed by G_3. This is so because although a causal graphical model may have exogenous influences that are not included in the graph, those influences are constrained to be uncorrelated (ruling out, e.g., all unobserved common causes whose values are not constant)—a fact that entails the unconditional independence of C_1 and C_2 (Pearl, 1988; 2000; Spirtes et al., 2000). Of course, one might question whether people had prior beliefs that the causes were correlated. For instance, they may believe that low interest rates are more likely when trade deficits are small than they would be otherwise, despite the instructions during the task. However, this interpretation is defeated by the counterbalancing of the senses of the variables (e.g., the role of C_1 was sometimes played by *high* instead of low interest rates; the role of C_2 was sometimes played by *large* instead of small trade deficits) and by the fact that the same pattern of results obtained in multiple domains (meteorology and sociology in addition to economics).

It has been established that people violate independence relations (a.k.a., commit "Markov violations") with network topologies other than G_3. Rationalizations of these errors all appeal to the possibility that subjects reason with knowledge in addition to that assumed by the experimenters (see Rottman & Hastie, 2013 for a review). Yet none explain the errors found by Rehder (2014a) with G_3. For instance, Park & Sloman (2003) demonstrated that the Markov violations that arose with their *common cause networks* (i.e., $E_1 \leftarrow C \rightarrow E_2$)—namely, subjects incorrectly treated E_1 and E_2 conditioned on C as dependent—were partly due to subjects' beliefs that the two causal links could be disabled by a common factor (also see Burnett, 2004; Lagnado & Sloman, 2004; Mayrhofer & Waldmann, 2015; Walsh & Sloman, 2008). However, this account does not explain the violations that occur with G_3. Relatedly, Rehder & Burnett (2005) explained the large variety of Markov violations they observed by assuming that all variables were related by an underlying common cause (an assumption justified on the basis of the fact that the variables were features of the same category; also see Rehder, 2014b). However, this account also fails to explain the results from Rehder (2014a) because it tested materials that were not features of categories.

The aim of this study is to assess whether the independence violations observed in reasoning tasks generalize to a *learning* context. One explanation for the independence violations observed during reasoning is that they are a side effect of the particular causal reasoning processes subjects invoke to render conditional probability judgments. If this is the case, we would not expect those violations to generalize to a learning task. If, on the other hand, people's understanding of the statistics implied by causal networks really differs from those assumed by formal models, we would expect independence violations to be reflected during both reasoning *and* learning. In particular, if people's beliefs about the statistics of common effect models is such that they think that the causes are positively correlated, then a data sample in which the cues are positively correlated will be viewed as more likely to be generated by a common effect model than samples in which the cues exhibit a zero or negative correlation.

Overview of Experiment

We present an experiment that evaluates the effect of intercue correlations on the evaluation of a causal hypothesis. Note that we adopted a novel experimental paradigm in which we didn't ask subjects to evaluate the relative likelihood of two hypotheses. Instead, we cut out the middle man, so to speak, by (a) presenting a candidate causal theory (one that took the form of G_3), (b) presenting a set of observations, and (c) asking subjects to rate how likely it is that the data was generated by the theory assuming that the theo-

ry is true. That is, we asked subject to *directly* evaluate $p(D|G_3)$.

Our key manipulation is to vary the degree of cue covariation across three levels: negative, zero, and positive. To generate sets of observations (hereafter, "samples"), we first identified a number of target parameters that we wanted the samples to match. Those parameters were, in order of importance, the strength of the causal relationships (.50), the marginal probability of the effect (.50), the log odds ratio (LOD) between the cues (–3, 0, 3 in the negative, zero, and positive conditions), the probability that E was caused by something other than C_1 and C_2 (0), and the marginal probability of each of the causes (.50). Sets of observations that closely instantiates these target parameters are presented in Table 1. The first eight rows of Table 1 correspond to the eight states that can be formed by three binary variables (again, the presence of a variable is denoted by 1 and its absence by 0). Each of these rows presents the number of instances of that type of observation in each of the three conditions. As described in detail later, the presentation of these observations was simultaneous, that is, all observations were visible on the computer screen at the same time.

Round off error induced by the finite-sized samples means that they didn't perfectly match the target parameters and so the bottom portion of Table 1 presents the statistics computed from the actual samples. As can be seen, the LODs between the cues were –3.51, 0, and 3.18, the causal strengths are all very close to .50, the marginal probability of E is always .46, and the probability that E occurs in the absence of C_1 and C_2 is always 0.

How might subjects estimate the likelihood of each of the three data sets given G_3? We first treated G_3 as a conditional Bayesian network in which the probability of the effect conditioned on the causes is computed without making any assumptions about the statistical properties of the causes. We refer to the probability of D given G_3 in this case as $p^{C_1C_2}(D|G_3)$ to emphasize that the network is conditioned on C_1 and C_2. The parameter space of G_3 is thus $\boldsymbol{\theta} = (w_0, w_1, w_2)$ and we sampled (10,000 times) over that space in the manner described in the Appendix. The averaged values of $log(p^{C_1C_2}(D|G_3))$ for each of the three conditions are shown in Table 1, which shows that the sample with a negative inter-cue correlation is most likely to be generated by G_3 (log likelihood of –15.70) followed by the samples with the zero (–16.05) and positive (–16.37) cue correlations. In particular, because this analysis fails to predict a preference for the sample with positively correlated cues, a finding that subjects in fact exhibit such a preference will bolster our claim that the cues of a common effect model are expected to be positively correlated.

We also computed $p(D|G_3)$ treating G_3 as *unconditional* Bayesian network, that is, taking the base rates of C_1 and C_2 into account. In this case the parameter space of G_3 is $\boldsymbol{\theta} = (b_1, b_2, w_0, w_1, w_2)$ where b_1 and b_2 represent the base rates of C_1 and C_2, respectively. Table 1 reveals that for this model, $log(p(D|G_3))$ is highest in the zero correlation condition. These analyses reveal that subjects should not favour the sample with positively correlated cues regardless of whether G_3 is treated as a conditional or unconditional network.

Whereas these analyses sampled over a uniformly distributed parameter space, subjects might think that some parameters values are more likely than others. Using beta distributions we introduced a prior in which $E(b_1) = E(b_2) = .5$, (i.e., C_1 and C_2 are each expected to occur about half the time), $E(w_1) = E(w_2) = .9$, (i.e., $C_1 \rightarrow E$ and $C_2 \rightarrow E$ are expected to be strong causal relations), and $E(w_0) = .1$ (i.e., causes of E other than C_1 and C_2 are expected to be weak). In this prior each expectation was worth 20 "observations." In fact, the introduction of this prior left the qualitative pattern of $p(D|G_3)$ and $p^{C_1C_2}(D|G)$ unchanged.

Table 1.

			Condition		
			Negative	Zero	Positive
C_1	C_2	E	Count	Count	Count
0	0	0	3	6	8
0	0	1	0	0	0
0	1	0	5	3	1
0	1	1	5	3	1
1	0	0	5	3	1
1	0	1	5	3	1
1	1	0	0	1	3
1	1	1	1	5	9
	$LOD(C_1, C_2)$		-3.51	0	3.18
	$p(E=1\|C_i=1, C_j=0)$		0.52	0.53	0.50
	$p(E)$		0.46	0.46	0.46
	$p(E=1\|C_1=0, C_2=0)$		0	0	0
	$p(C_i)$		0.46	0.50	0.58
	$\log(p(D\|G_3))$		-40.98	-35.52	-41.85
	$\log(p^{C_1C_2}(D\|G_3))$		-15.70	-16.05	-16.37

Note that the process via which the samples in Table 1 were generated was chosen to avoid confounds that might complicate the interpretation of the results. One possibility that concerned us is that subjects might interpret the $p(D|G_3)$ query as one requesting the *strengths* of the causal relations that inhered in each sample. This is why all three samples were constrained to have almost equal causal strengths. Moreover, because causal strength judgments themselves are known to be affected by the marginal probability of the effect (the "outcome density bias"), the samples were matched on that factor as well; they were also matched on the probability of E occurring in the absence of C_1 and C_2 (it was 0). Note that one factor on which the samples *weren't* matched is the marginal probability of the causes (which were .46, .50, and .58 in the negative, zero, and positive correlation conditions). Because in G_3 the marginal probability of E will, all else being equal, increase as the cues become negatively correlated and decrease as they become positively correlated, equating $p(E)$ across conditions entails that $p(C_i)$ is smaller in the negative condition and larger in the positive condition. Equating $p(E)$ was deemed

more important because of the documented outcome density bias.

In summary, if subjects' judgments concerning $p(D|G_3)$ reflect an expectation that cues are uncorrelated, then the ratings should be highest in the zero-correlation condition. If it ignores the statistical properties of the cues altogether, then ratings should be highest in the negative-correlation condition. Finally, if it reflects an expectation that cues will be positively correlated, then ratings should be highest in the positive-correlation condition and lowest in the negative-correlation condition. The following experiment tested these predictions.

Method

Materials. Participants were presented with causal hypotheses in three domains: meteorology, sociology, and economics. Each domain had three variables: interest rates, trade deficits, and retirement savings in economics; degree of urbanization, interest in religion, and socio-economic mobility in sociology; and ozone level, air pressure, and humidity in meteorology. Each variable could take on two possible values. One of these values was described as "Normal" and the other was either "High" (or "Large" for some variables) or "Low" (or "Small") according to a ran-

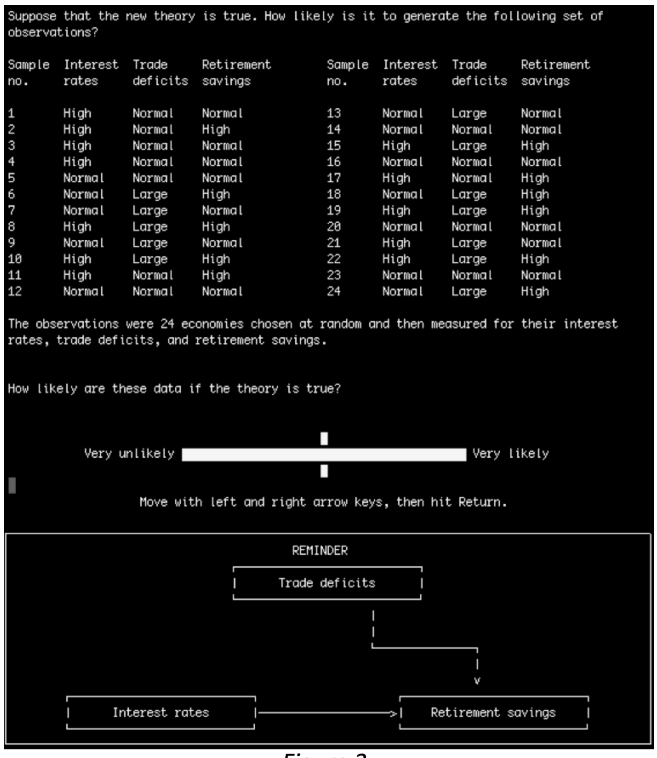


Figure 3

domization scheme described below.

Each hypothesis was of the form G_3 in Fig. 1. The description of each of the two causal relationships consisted of two sentences, one that stated the cause and effect and a second that provided information about the causal mechanism. For example, "High interest rates cause high retirement savings. The high interest rates result in high yields on government bonds, which are especially attractive for retirement savings because they are such a safe investment."

Procedure. Participants first learned a hypothesis in each domain. For each hypothesis, they then judged the degree of fit between it and the three samples in Table 1. Participants thus made a total of nine judgments.

In each domain, participants were first told that researchers (a "research group of economists" in the case of economics) had proposed a "new theory" consisting of two paragraphs describing the two causal relationships. A diagram of those causal relationships analogous to G_3 in Fig. 1 was also presented. Subjects were verbally instructed in two ways that the causal relationships operated independently. Firstly, they were explicitly told in two ways that the causes were independent (e.g., "note that high interest rates and large trade deficits each bring about high retirement savings on its own"; "high interest rates can produce high retirement savings by itself, and large trade deficits can independently produce high retirement savings by itself as well"). Secondly, as described in Materials, they were provided with separate causal mechanisms for each of the causes, providing an account by which each cause could independently bring about the effect.

Participants were then presented with the three samples. Each was presented on one screen that listed the 24 observations (i.e., presentation was simultaneous). See Fig. 3 for an example. Each sample was described as being "chosen at random" and measured on the three variables that made up the hypothesis (interest rates, trade deficits, and retirement savings in the case of economics). Subjects were asked to judge the likelihood of observing this sample assuming that the theory was true. The response scale ranged from "Very unlikely" to "Very likely". So that subjects did not have to rely on their memory of the theory, the diagram of the causal relationships was re-presented at the bottom of this screen (see Figure 3). The next two screens presented the second and third sample. The presentation order of the domains (economies, societies, and weather systems) and the three samples (Negative, Zero, Positive) was randomized as described below.

Design and participants. There were two between-subject factors. A Latin square determined the order of presentation of the three domains (meteorology, sociology, economics). Subjects were randomly assigned to these three cells subject to the constraint that an equal number appeared in each. The senses of the variables' non-normal values were randomized such that they were either all high, all low, or all high in the first, second, and third domain, respectively (HLH) or exhibited the reverse pattern (LHL). Within each domain the presentation order of the three samples was randomized for each subject. 21 New York University undergraduates received course credit for participating.

Results

Initial analyses revealed no effect of domain (economics, sociology, meteorology) or sense (HLH, LHL), so the results presented are collapsed over these factors

As hypothesized, there was a main effect of inter-cue correlation on ratings of ($p(D|G_3)$, $F_{2,32} = 33.82$, MSE = 53.58, $p < .001$). Ratings from 0-100 for each of these correlations

are displayed in Fig. 4. Additionally, participants consistently preferred positive correlation, with positively correlated cues having higher ratings than non-correlated cues ($p < .001$), which in turn had higher ratings than negatively correlated cues ($p < .001$).

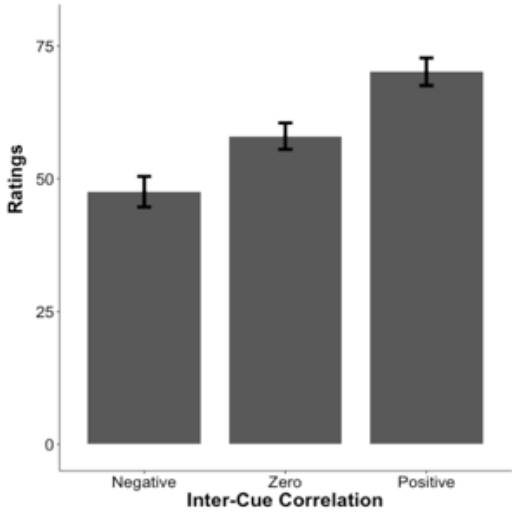


Figure 4. Mean Ratings. Error bars represent standard errors

Discussion

This article presents evidence that people are sensitive to inter-cue correlations in a way not predicted by current causal learning models. In particular, they expect that the two cues in a common effect structure will be positively correlated, resulting in a preference for data that manifests such a correlation. Because the current findings involved a new sort of task, they generalize the violations of the independence constraints stipulated by causal Bayes nets that obtain in reasoning tasks in which subjects render conditional probability judgments. Participants' non-normative expectations of the statistical properties of data generated from causal structure poses problems for models (e.g., Carroll et al., 2013; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014; Powell et al., 2016) that assume that learners estimate $p(D|G)$ in a veridical manner.

There are important methodological differences between the present work and traditional causal learning studies. Whereas we asked participants to judge the probability that a particular causal structure generated some data set (i.e. judge $p(D|G_3)$), participants are usually asked to estimate causal strength (e.g., Cheng, 1997, and many others) or to select the causal structure most consistent with the observed data (e.g., Lu et al. 2008). We have also conducted a straightforward extension of our paradigm by asking subjects to instead estimate the posterior probability of the four hypotheses in Fig. 1—that is to estimate $p(G_i|D)$ for each G_i instead of $p(D|G_3)$. Consistent with the results above, we found that $p(G_3|D)$ was highest in the positively-correlated condition and lowest in the negatively-correlated one. We have also asked subjects to generate hypothetical data from G_3 (having them, in effect, estimate G_3's joint distribution). In fact, that joint distribution reflected their expectation that the two causes were positively correlated.

Our paradigm can reveal the non-normative expectations about statistical structure that people have for other network topologies. For instance, in a common cause structure ($E_1 \leftarrow C \rightarrow E_2$), the fact that E_1 and E_2 should be independent conditioned on their common cause C is often violated (e.g. Rehder, 2014a). It would be straightforward to generate samples of data in which, conditioned on C, the correlation between E_1 and E_2 is either negative, zero, or positive. We predict that the sample with the positive correlations will be treated as more consistent with $E_1 \leftarrow C \rightarrow E_2$.

Most modern models of causal learning fall under the umbrella of conditional Bayesian networks, which are invariant over statistical properties of parent nodes (Carroll et al., 2013; Griffiths & Tenenbaum, 2005; Lu et al., 2008; Meder et al., 2014; Powell et al., 2016). The current study provides evidence that such assumptions contrast with actual causal-based judgments. A fully descriptive formal model of causal learning behaviour, then, must also be able to account for participant expectations of the statistical properties of cues.

References

Burnett, R. C. (2004). Inference from complex causal models (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3156566)

Carroll, C. D., Cheng, P. W., & Lu, H. (2013). Inferential dependencies in causal inference: A comparison of belief-distribution and associative approaches. *Journal of Experimental Psychology: General, 142*, 845-863.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Griffiths, T. L., & Tenebaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Koller, D. & Friedman, L. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lagnado, D.A. & Sloman S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*:856–76

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-984.

Maloney, L. T., Dal Martello, M. F., Sahm, C. & Spillmann, L. (2005), Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences*, **102**, 3164-3169.

Mayrhofer, R. & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*, 65-95.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review, 121*, 277-301.

Park, J. & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186-216.

Pearl, J. (2000). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, *95*(450),

428-431.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
Perales, J., Catena, A., & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. Learning and Motivation, 35, 115–135.
Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology, 86*, 62–86.
Rehder, B. (2014a). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54-107.
Rehder, B. (2014b). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*
Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.
Rottman, B., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. New York: Springer-Verlag.
Walsh, C. R., & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. In G. H. Bower, M. A. Gluck, J. R. Anderson & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon Bower* (pp. 345-358).

Appendix

Computing $p^{C_1C_2}(D|G_3)$ when G_3 is interpreted as a conditional network yields,

$$p^{C_1C_2}(D|G_3) = \int_0^1\int_0^1\int_0^1 p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)\, p(\boldsymbol{\theta}|G_3)\mathrm{d}w_0\mathrm{d}w_1\mathrm{d}w_2 \quad (3)$$

The computation of $p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)$ raises well know issues regarding how people estimate the likelihood of sequences of events. For example, if the likelihood of the effect E in some context is .50, then presumably people will judge a sequence in which E is always present as less probable than one in which E is present about half the time, in the same way that a fair coin is viewed as unlikely to yield a long run of all heads or all tails as compared to a mixed sequence (Kahneman & Tversky, 1972; also see Maloney et al., 2004). This concern is especially relevant in the current experiments in which observations are presented simultaneously rather than sequentially, making non-representative samples easier to detect. Accordingly, here we use the binomial distribution as a first order approximation of $p^{C_1C_2}(D|\boldsymbol{\theta}, G_3)$. In particular,

$$p^{C_1C_2}(D|\boldsymbol{\theta}, G_3) = \prod_{\boldsymbol{c}} B(N(e=1,\boldsymbol{c}); N(\boldsymbol{c}), p(e=1|\boldsymbol{c})) \quad (4)$$

$$p(e=1|\boldsymbol{c}) = 1-(1-w_0)(1-w_1)^{c_1}(1-w_2)^{c_2} \quad (5)$$

where the product ranges over the four distinct settings of C_1 and C_2 in D.[1] B is the binomial distribution which returns the probability of $N(e=1,\boldsymbol{c})$ "successes" (number of times E is present in context $\boldsymbol{c}$) in $N(\boldsymbol{c})$ "trials," where the probability of a success is $p(e=1|\boldsymbol{c})$.

Computing $p(D|G_3)$ when G_3 is interpreted as an unconditional network yields yields

$$p(D|G_3) = \int_0^1\int_0^1\int_0^1\int_0^1\int_0^1 p(D|\boldsymbol{\theta}, G_3)\, p(\boldsymbol{\theta}|G_3)\mathrm{d}b_1\mathrm{d}b_2\mathrm{d}w_0\mathrm{d}w_1\mathrm{d}w_2 \quad (6)$$

and $p(D|G_3)$ is computed using the multinomical distribution M,

$$p(D|\boldsymbol{\theta}, G_3) = M(N(e,\boldsymbol{c});\, p(e,\boldsymbol{c}|\boldsymbol{\theta}, G_3)) \quad (7)$$

where $N(e,\boldsymbol{c})$ is again the vector of counts associated with each unique combination of C_1, C_2 and E and $p(e=1,\boldsymbol{c}|\boldsymbol{\theta}, G_3))$ is the vector of corresponding probabilities, where

$$p(e,\boldsymbol{c}|\boldsymbol{\theta}, G_3)) = p(e|\boldsymbol{c}, \boldsymbol{\theta}, G_3)p(\boldsymbol{c}) = p(e|\boldsymbol{c})b_1b_2 \quad (8)$$

and $p(e|\boldsymbol{c})$ is given by Eq. 5.

[1] Use of a binomial distribution introduces a normalizing constant in the computation of the likelihood, one that cancels out when the relative likelihood of two hypotheses is being computed (such as in Griffiths & Tenenbaum's support model). In contrast, in the current experiments subjects estimate absolute rather than relative likelihoods. Use of a binomial distribution may also be less appropriate in traditional learning experiments in which data is presented sequentially (where a non-representative sample may be less salient) rather than simultaneously.

Do Simple Probability Judgments Rely on Integer Approximation?

Shaun O'Grady (shaun.ogrady@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California,
Berkeley, CA 94720 USA

Abstract

A great deal of research has been conducted on how humans reason about probability, yet it remains unknown what mental computations support this ability. Research on the development of the Approximate Number Sense (ANS) has shown that performance in a magnitude (i.e., estimations of integers) discrimination task is well fit by a psychophysical model (Halberda & Feigenson, 2008). Whether or not estimations of integers plays a role in probability judgments has yet to be investigated. In the present study we use data from two adult experiments as well as results from comparisons of two computational models to investigate the potential relationship between the ANS and probability judgments.

Keywords: Probability; Approximate Number Sense (ANS); estimation

Introduction

Although classic findings in the literature on probabilistic reasoning indicate that adults perform poorly on a wide range of problems involving complex probability (see Kahneman, 2011, for a review), a great deal of evidence suggests that even young children and infants are adept at making correct judgments about simple probability based on proportion (e.g., Denison, Reed, & Xu, 2013; Piaget & Inhelder, 1975). Despite the vast literature on the development of probabilistic reasoning, very little research has been conducted on how these computations are made. In what follows we review findings from research on the development of probabilistic reasoning as well as research on the approximate number sense (ANS) and present findings from two experiments designed to investigate the role of the ANS in adults' judgments about probability based on proportion. We then compare two models proposed to predict probability judgments based on proportion in order to understand the computational process that underlies probability judgments.

Development of Probabilistic Reasoning

Recent evidence suggests that even infants possess powerful statistical reasoning abilities which allow them to predict the outcome of probabilistic events. Researchers have shown that 8-month old infants form expectations about relationships between samples and populations of objects (Xu & Garcia, 2008) and infants as young as 6.5 months expect samples to reflect the ratios of the populations from which those samples were obtained (Denison, Reed & Xu, 2012). These powerful reasoning abilities continue to develop in infancy with older infants making increasingly sophisticated probabilistic judgments about their environment (Denison & Xu, 2009; Gweon, Tenenbaum & Schulz, 2010). As children grow, their judgments about probability become more and more accurate (Piaget & Inhelder, 1975). Recently, Falk, Yudilevich, Assouline & Elstein (2012) found that young children's judgments about simple probability based on proportion are influenced more by the number of favorable events rather than the proportion of favorable to unfavorable events but around 7-8 years of age this changes and children adopt a proportional strategy similar to those used by adults. Based on these findings it seems reasonable to assume that children's improving sense of probability is related to improvements in general number reasoning.

The Approximate Number Sense

Several research teams have shown that both humans and non-human primates have a sophisticated system for making rapid judgments about large and small sets of objects (Whalen, Gallistel & Gelman, 1999; Feigenson, Carey & Hauser 2002, Pica, Lemer, Izard & Dehaene, 2004; Halberda & Feigenson, 2008). This has been termed the Approximate Number Sense (ANS) and has been studied by testing participants' ability to discriminate between two different sets of objects that vary in the magnitude of the difference between the two sets. Several of these studies have found that the acuity of the ANS improves with age and can be characterized by Weber's Law (Halberda & Feigenson, 2008; Pica et al. 2004; Whalen et al. 1999). Accordingly, the ANS is characterized by ratio dependence, meaning that the ability to discriminate between two sets is dependent on the ratio of the magnitudes of those sets.

While there is evidence to suggest that ANS acuity is related to mathematical ability (Halberda, Mazzocco & Feigenson, 2008), some recent evidence has shown that ANS acuity may not influence probabilistic judgments (Patalano, Saltiel, Machlin & Barth, 2015; Winman et al. 2014). Furthermore, some researchers have proposed that rational number reasoning may rely on a set of neural computations that are distinct from that of the ANS (Jacob, Valentin & Neider, 2012). However, it is possible that the computation of rational number comparisons such as proportion and probability, relies on integer estimates. For this reason, it is important to understand the computational process involved in accurate judgments about probability.

Experiment 1: Evaluating Proportions.

Do adults use the approximate number system to reason about the difference between two proportions? To answer this question, we devised a computer-based experiment in which participants are presented with two distributions of red and white marbles and asked to pick the distribution that was most likely to yield a target color marble. If the ANS is recruited when making judgments about probability, then ANS acuity should be negatively correlated with performance on a two-alternative forced-choice task requiring the judgment of probability based on proportion. Additionally, as predicted by Weber's Law, there should be an effect of the distance between the ratios of compared proportions. Finally, if adults are making accurate judgments their responses should be based on the proportion of favorable to non-favorable events rather than the total number of favorable events.

Methods

The methods reported below were originally designed for a study involving young children and for this reason the methods were presented to adults in a child-friendly manner. Only the data from adults is presented below as we are still collecting data from 6- to 12-year-old children.

Participants Forty-eight adult undergraduates (Mean age: 22.62; 37 female) from Psychology classes at UC Berkeley were recruited for participation in this study.

Measures & Stimuli Images were designed using Blender 2.72, 3D animation software (http://www.blender.org/) and consisted of two groups of red and white marbles separated by a black partition. The proportions used are presented in Table 1. All images were presented on a MacBook Pro laptop (OSX; 1280 x 800) using the MatLab programing language with psychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et al, 2007). In order to measure the acuity of participants' approximate number sense, participants played Panamath a game designed to measure the acuity of the ANS (http://panamath.org/) for 10 minutes after the probability task.

(a) (b)

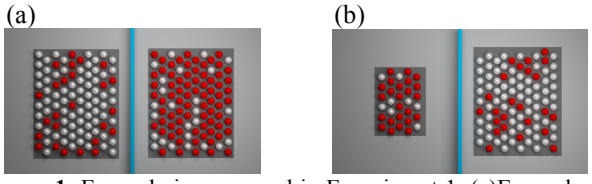


Figure 1: Example images used in Experiment 1. (a)Example of a total equal trial. (b) Example of a target equal trial

Procedure Participants sat about 60 centimeters from the laptop on which the probability game and Panamath task were presented. On the initial screen, participants saw a Big Bird character with two bags on either side and were told that Big Bird really likes red marbles. They were then told that Big Bird is going to close his eyes and take a marble from one of two groups of marbles. Their job was to tell Big Bird which group to take a marble from to help him collect as many of his favorite color marbles as possible. Participants then played four practice trials in which they saw a group of red marbles on one side and a group of white marbles on the other side.

Two types of trials were used for each ratio of ratios listed in Table 1. On 'total equal' trials both groups of marbles contained 100 marbles while on 'target equal' trials the number of Big Bird's favorite color marbles in each group was equal. See Figure 1 for an example image of the two trial types. If participants are using a strategy based on choosing the distribution with the greatest amount of favorable marbles, they should perform at around chance levels on target equal trials.

All of the images were presented on screen for 500ms in order to prevent participants from counting the marbles. Participants played the game for 40 trials presented in one of two orders for both red and white target colors. Trial order and target color were counterbalanced across subjects. Adults participants were told that the game was originally made for young children and therefore used a lot of child friendly language.

Table 1: Ratios of marbles used in Experiment 1.

Ratio of ratios	Bag 1 prop.	Bag 2 prop.	**Ratio of ratios**	Bag 1 prop.	Bag 2 prop.
14.00	0.70	0.05	**2.33**	0.70	0.30
11.00	0.55	0.05	**2.00**	0.60	0.30
10.00	0.50	0.05	**1.83**	0.55	0.30
9.00	0.90	0.10	**1.75**	0.70	0.40
8.00	0.80	0.10	**1.50**	0.60	0.40
6.00	0.90	0.15	**1.45**	0.80	0.55
4.00	0.80	0.20	**1.33**	0.80	0.60
3.50	0.70	0.20	**1.22**	0.55	0.45
3.00	0.75	0.25	**1.17**	0.70	0.60
2.67	0.80	0.30	**1.10**	0.55	0.50

Note: 'Bag 1 prop.' etc, indicates the proportion of favorable marbles.

Results

Analyses of general performance revealed the average percentage correct to be 94.9% (SD = 0.04%) and this was significantly greater than that expected by chance (t = 64.93, df = 47, p < .001). The average Weber Fraction (WF) for the 30[1] adults tested on Panamath was 0.162 (SD = 0.0058). Importantly, WF was not statistically significantly correlated with general performance in the probability game (Pearson's r = -0.32, t = -1.8036, df = 28, p = .082).

Generalized Linear Models with Mixed effects (GLMMs) were used to predict performance based on age, ratio of ratios, and trial type with participant entered as a random effect and found no main effect or interaction effects for gender, trial order or target color. The model predicting performance based on the ratio of ratios and trial type with an interaction ($AIC_{RR*trial}$ = 772.04) outperformed the null model (AIC_{null} = 837.15) (χ^2 = 71.105, df = 3, p < .001), the

[1] The scores of 18 participants were excluded from these analyses because they played the Panamath game for fewer than 8 minutes which is the required amount of time to acquire an accurate measure of number sense acuity.

model predicting performance based on ratio of ratios (AIC_{RR} = 822.81) (χ^2 = 54.7, *df* = 2, *p* < .001), age (AIC_{age} = 839.15) (χ^2 = 71.105, *df* = 2, *p* < .001), trial type (AIC_{trial} = 824.72) (χ^2 = 56.674, *df* = 2, *p* < .001), ratio of ratios and trial type ($AIC_{RR+trial}$ = 810.22) (χ^2 = 40.177, *df* = 1, *p* < .001), ratio of ratios, trial type, and age ($AIC_{RR+trial+age}$ = 812.22) (χ^2 = 40.176, df = 0, *p* < .001). Average performance by trial type and ratio of ratios is plotted in Figure 2.

Analysis of the coefficients revealed that as the ratio of ratios increased the chances of a correct response increased by 4%. The chance of a correct response on total equal trials was 88% less likely than for target equal trials and the effect of ratio of ratios was 4.9 times greater for total equal trials than for target equal trials.

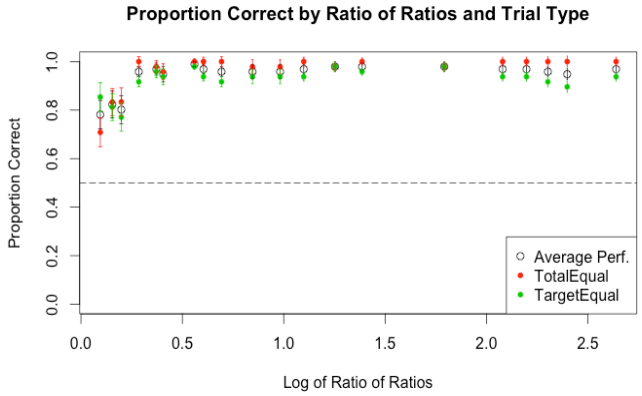


Figure 2: Percentage correct by trial type and ratio of ratios.

Discussion

These results suggest that the ANS may not be used in judgments of probability based on proportion: there was not a statistically significant correlation between participants' Weber fraction and general performance in the probability discrimination task, with a sample size (*n* = 30) where a large effect (*r* > 0.5) would have been detected (Cohen, 1988). Interestingly, there was an effect of ratio of ratios on performance in this task and this effect was in the expected direction: as the ratio of ratios increases so did performance, a finding that mirrors the distance effect reported in the ANS literature.

The GLMMs also revealed that performance improved on trials in which the number or target marbles is equal suggesting that participants' judgments were influenced by the number of favorable marbles. However, it should be noted that performance on both target equal and total equal trials was significantly above chance indicating that participants did not adopt a strategy based solely on the number of favorable marbles but rather that performance was enhanced on trials in which the target number of marbles was equal indicating that some participants may have adopted a strategy of avoiding the choice with the most red on some trials.

Several questions arise in response to these findings. First, and most importantly, it is interesting that there was no significant correlation with ANS acuity but it remains unknown if this is true of children as well. It is possible that adults have learned a non-numerical strategy and therefore no longer use their ANS when making probabilistic judgments. Future research will investigate the influence of ANS acuity on children's probability judgments. Second, the astute reader will notice that the ratio of ratios that were used in Experiment 1 were mostly comparisons of favorable (above 50% chance) distributions to unfavorable (below 50% chance) distributions. It seems likely that different results may be found when comparing favorable distributions to other favorable distributions as well as comparing unfavorable to other unfavorable distributions.

Another issue with the design of Experiment 1 arises from the analysis of trial type in which there was enhanced performance for target equal trials. Using the current design it is impossible to tell if participants were relying solely on approximations of either target or non-target marbles or whether they were reasoning about the proportion of target to non-target marbles. Trials in which the distribution with a lower proportion actually has a higher number of favorable marbles would show such an effect. Finally, it is impossible to tell to what extent participants' judgments are influenced by the proportion of area of the two colors rather than the proportion of the number of marbles, given that all the marbles were the same size. Considering the lack of correlation with ANS acuity in the adult sample, it is possible that participants are making proportional judgments based on area rather than numerosity. Each of these points is addressed by design modifications made in Experiment 2.

Experiment 2: Area, number & chance.

Several changes were made to the design of Experiment 2 in order to address the potential limitations of Experiment 1 and explore the possibility that participants are not recruiting the ANS when making judgments but are instead using other possible alternative strategies. The results we present here constitute pilot data for a more comprehensive study. The full version of the study will include more trials with a wider range of ratio comparisons.

Methods

Participants 19 undergraduates (Mean age: 21 years (SD = 1.94 years); 14 female) enrolled in psychology classes at U.C. Berkeley participated in this experiment.

Measures & Stimuli New ratio of ratios comparisons were chosen to include favorable vs unfavorable (F:U), favorable vs favorable (F:F), favorable vs 50% chance (F:C), 50% chance vs unfavorable (C:U), and unfavorable vs unfavorable (U:U) comparisons similar to Drucker, Rossa and Brannon (2016).

Three trial types were used for each of the ratio of ratios. On 'total equal' trials the total number of marbles in both distributions was equal while on 'number vs proportion' trials there were more favorable marbles in the distribution with the lower proportion. As in Halberda and Feigenson

(2008), 'area anti-correlated' trials included larger favorable marbles in the lower proportion distribution and smaller favorable marbles in the higher proportion distribution such that the proportion of area of favorable color in the lower proportion distribution was equal to the proportion of the number of marbles in the higher proportion distribution and vice versa. See Figure 3 for example images of trial types.

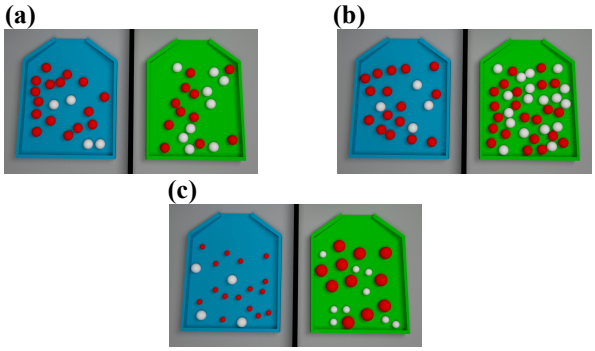


Figure 3: Example images used in Experiment 2. (a) Example of a total equal trial. (b) Example of a number vs proportion trial. (c) Example of an area-anticorrelated trial.

Procedure Images were presented using the same procedure as in Experiment 1 and the instructions were exactly the same except that participants were no longer asked to help Big Bird but instead were asked to choose the box most likely to yield a red marble. During practice trials participants were shown that the trays containing the marbles were dumped into two different boxes which were then shaken up. This was done in order to reduce the use of spatial cues that could influence the participant's selection. box they chose would tip over and dispense a single marble. Participants were also told that the size of the marble did not matter and the large marbles are just as likely to fall out of the box as small marbles. Finally, participants played the Panamath game after the probability task for 10 minutes in order to measure their ANS acuity.

Results

Participants' average performance is presented in figure 4. Mean overall percentage correct was 70.3% (*SD* = 14.9%), and this was significantly better than chance (t = 5.936, *df* = 18, $p < .001$)). The average Weber Fraction (WF) as measured by Panamath was 0.156 (*SD* = 0.029) and participant's WF was not statistically significantly correlated with general performance in the probability game (Pearson's $r = -0.39$, $t = -1.762$, $df = 17$, $p = .096$).

As with Experiment 1, GLMMs were used for the analyses. Model comparisons revealed that the model with the best fit was that which predicted performance based on ratio of ratios, trial type, and the interactions between ratio of ratios and trial type ($AIC_{RR*trial} = 576.33$). This model outperformed the null model ($AIC_{null} = 677.35$) ($\chi^2 = 111.02$, $df = 5$, $p < .001$), as well as models predicting performance based on the ratio of ratios ($AIC_{RR} = 652.97$) ($\chi^2 = 84.64$, $df = 4$, $p < .001$), trial type ($AIC_{trial} = 608.78$) ($\chi^2 = 38.45$, $df = 3$, $p < .001$), and ratio of ratios and trial type ($AIC_{RR+trial} = 580.4$) ($\chi^2 = 8.07$, $df = 2$, $p = .018$). Gender, age, comparison category, trial order and target color were not significant predictors of performance.

Analyses of the coefficients of the superior model revealed that as the ratio of ratios increased participants were more than twice as likely to make a correct choice. Both the number vs proportion and the area anti-correlated trials had a negative effect on the likelihood of a correct response with number vs proportion trials decreasing the chances of a correct response by 93.2% and area anti-correlated trials by 12.7%. Interestingly, the interaction terms indicated that the effect of ratio of ratios increased by 4% for number vs proportion trials and decreased by 52.4% for area anti-correlated trials.

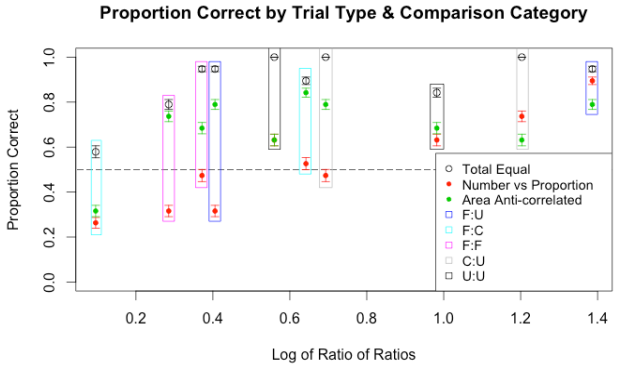


Figure 4: Proportion correct by trial type, ratio of ratios, and comparison category

Discussion

Not surprisingly, performance on this task was much worse than performance in Experiment 1 and this is probably due to the fact that we included many more difficult trials. In particular, the number vs. proportion and area anti-correlated trials were meant to be more difficult than the target equal trials used in Experiment 1. The interaction between ratio of ratios and trial types suggest that participants are using an area calculation as part of their estimations and this is interesting given the lack of an effect of area in the ANS literature (Halberda & Feigenson, 2008). Importantly, the comparison category did not seem to affect participant's judgments which may indicate that adult participants are able to compute an accurate probability estimate regardless of the likelihood of obtaining a target event. However, since there are only two of each comparison category and the comparisons are not evenly spaced throughout the ratios of ratios used in the study, the effect of qualitative distinctions between comparisons must remain open to speculation.

Interestingly, performance was still above chance even with the more difficult trials used in Experiment 2 and there was still no correlation with ANS acuity even though there was still a significant effect of the ratio of ratios. It is still too early to definitively rule out the role of the ANS in probability judgments since participants may be using some computation applied to their estimates of the number of

target and non-target marbles. For example, the participant might estimate the total number of target marbles as well as the total number of marbles for both distributions. Since ANS estimates are inexact, even for those who have with high number sense acuity, an error in any one of these estimates could lead to an incorrect judgment. For this reason it is necessary to understand the computations that lead to accurate performance as well as the strategies that people may adopt when reasoning about the probability of discrete binary outcomes.

Models of Probability Discrimination

Several studies investigating the approximate number sense (Halberda & Feigenson, 2008; Pica et al. 2004) have used computational models based on signal detection theory in the psychophysical literature (Green & Swets, 1966). The psychophysics model assumes that participants' approximations form a Gaussian distribution on a mental number line around a mean close to the actual number of objects being estimated and a standard deviation given by the participants' number sense acuity, or Weber fraction, multiplied by the mean. When presented with two groups of objects and asked which has more the participants' estimates form two Gaussian distributions which overlap based on the distance between their means and the size of their standard deviations. The area of overlap can be thought of as the probability of a subject making an error, which can be modeled using the psychophysical function based on the CDF of the Gaussian distribution:

$$P(R_c) = \Phi\left(\frac{\mu_1-\mu_2}{\sqrt{2w}\sqrt{\mu_1^2+\mu_2^2}}\right) = \frac{1}{2}erfc\left(\frac{\mu_1-\mu_2}{\sqrt{2w}\sqrt{\mu_1^2+\mu_2^2}}\right) \quad (1)$$

where $P(R_c)$ is the probability of a correct response, w is the subjects' Weber fraction, $\mu_1$and μ_2 are the means of the two numbers being estimated, and Φ is the Cumulative Distribution Function (CDF) of the Gaussian distribution.

This model assumes that participants estimate probability as an exact numerical value between 0 and 1. Although Equation 1 seems to be a plausible model for how people are making judgments about probability based on comparisons of two binary distributions, assessment of the likelihood of this model reveals that it is not well fit to the data. Figure 5 provides the model predictions alongside the data from Experiments 1 and 2.

The probability of selecting a target color marble from the distributions used in Experiments 1 and 2 can be represented as rational numbers between 0 and 1. If the subjects in our study are using their ANS to calculate probabilities they would need to approximate the number of favorable objects and divide this by the total number of all objects in the distribution which could be modeled using the distribution of a ratio of two random variables. Although a general form for calculating the probability density function for this distribution exists (Hinkley, 1969), a model using assumption distribution makes similar predictions as the model assuming a Gaussian distribution with the ratio comparisons that are reported for Experiments 1 and 2 above.

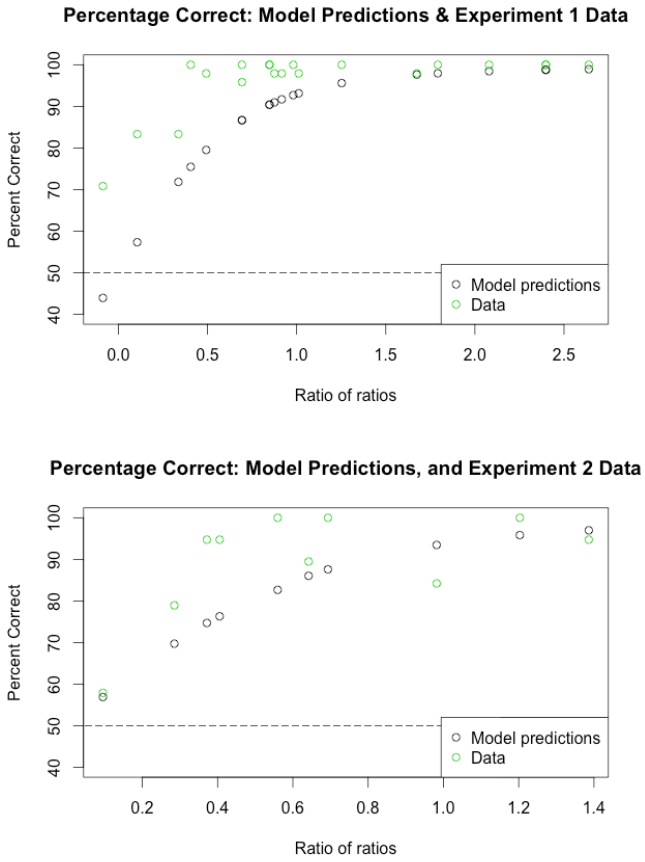


Figure 5: Model predictions for the Gaussian model along with data from Experiments 1 and 2. (a) Predictions and data for Experiment 1. (b) Predictions and data for Experiment 2.

General Discussion

Using a combination of experimental evidence and computational modeling we sought to demonstrate the role of the approximate number sense in adult participants' judgments of probability based on proportion. In two experiments we demonstrated that adults' ANS acuity was not statistically significantly correlated with performance on a probability discrimination task. Since a large correlation would have been detected in these experiments, these results suggest that the ANS is not recruited when people make judgments about probability based on proportion.

It is possible that participants' judgments about probability comparisons are more heavily influenced by qualitative factors such as trial type and comparison category. Future work will include larger numbers of more varied comparisons in order to more thoroughly investigate the factors influencing probability judgments. In addition, we plan to use more complex models that can account for

both quantitative and qualitative factors in judgments of probability based on proportion.

Acknowledgments

We would like to thank Falk Lieder and Dylan Daniels of U.C. Berkeley for their suggestions on this manuscript. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE 1106400)

References

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*:433-436.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Denison, S. & Xu, F. (2010) Twelve- to 14-month-old infants can predict single-event probability with large set sizes.*Developmental Science, 13,* 798-803.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: evidence from 4.5-and 6-month-olds. *Developmental Psychology*, *49*(2), 243.

Drucker, C. B., Rossa, M. A., & Brannon, E. M. (2016). Comparison of discrete ratios by rhesus macaques (Macaca mulatta). *Animal Cognition*,*19*(1), 75-89.

Falk, R., Yudilevich-Assouline, P., & Elstein, A. (2012). Children's concept of probability as inferred from their binary choices—revisited. *Educational Studies in Mathematics*, *81*(2), 207-233.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the" Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457.

Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differ- ences in non-verbal number acuity correlate with mathematics achievement. *Nature*, 455, 665–668.

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, *56*(3), 635-639.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kleiner M, Brainard D, Pelli D, 2007, "What's new in Psychtoolbox-3?" Perception 36 ECVP Abstract Supplement.

Patalano, A. L., Saltiel, J. R., Machlin, L., & Barth, H. (2015). The role of numeracy and approximate number system acuity in predicting value and probability distortion. *Psychonomic Bulletin & Review*, *22*(6), 1820-1829.

Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision 10*:437-442.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children. (Trans L. Leake, P. Burrell & HD Fishbein)*. WW Norton.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*,*10*(2), 130-137.

Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology*, *5*.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old Infants. *Cognition*, *112*(1), 97-104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.

Social Affordance Tracking over Time - A Sensorimotor Account of False-Belief Tasks

Judith Bütepage (butepage@kth.se)
Hedvig Kjellström (hedvig@csc.kth.se)
Danica Kragic (dani@kth.se)
Computer Vision and Active Perception Laboratory,
CSC, KTH, Stockholm, Sweden

Abstract

False-belief task have mainly been associated with the explanatory notion of the theory of mind and the theory-theory. However, it has often been pointed out that this kind of high-level reasoning is computational and time expensive. During the last decades, the idea of embodied intelligence, i.e. complex behavior caused by sensorimotor contingencies, has emerged in both the fields of neuroscience, psychology and artificial intelligence. Viewed from this perspective, the failing in a false-belief test can be the result of the impairment to recognize and track others' sensorimotor contingencies and affordances. Thus, social cognition is explained in terms of low-level signals instead of high-level reasoning. In this work, we present a generative model for optimal action selection which simultaneously can be employed to make predictions of others' actions. As we base the decision making on a hidden state representation of sensorimotor signals, this model is in line with the ideas of embodied intelligence. We demonstrate how the tracking of others' hidden states can give rise to correct false-belief inferences, while a lack thereof leads to failing. With this work, we want to emphasize the importance of sensorimotor contingencies in social cognition, which might be a key to artificial, socially intelligent systems.

Keywords: social cognition, sensorimotor signals, affordances, false-beliefs, theory of mind.

Introduction

Social cognition benefits highly from our ability to infer and predict others' percepts and future actions. Such inferences have been postulated to occur on two levels; low-level sensorimotor contingencies (SMCs) and high-level goal and mental state inferences. While the former phenomenon has been explained with help of the simulation theory (ST) and embodied cognition, the later inferences are commonly addressed with help of the theory-theory (TT) and accounts of the theory of mind (ToM) (Dindo, Donnarumma, Chersi, & Pezzulo, 2015). While the simulation theory is supported by biological concepts such as mirror neurons, the ToM is rooted in psychological approaches to social cognition.

One cornerstone of the theory of mind are "false-belief tests", which, according to the TT approach, necessitate the ability to track and infer others' mental states, beliefs and intentions. Since both young children and autistic children fail false-belief tasks to a great extent, these tests are usually presented as a measure of advanced social intelligence. A recent proposal, however, views false-belief tasks in the light of SMCs, with an emphasis on social affordances and working memory (Brincker, 2014). Instead of tracking and inferring others' mental states, which is computationally expensive, the memory of past affordances, individual and shared, could be involved in false-belief inferences. As social SMCs and affordances belong to low-level cognitive mechanisms, the tracking of these signals comes at a lower cost than mentalizing. While executive functions are of relevance for false-belief tasks, see e.g. (Devine & Hughes, 2014), we want to emphasize the importance of the understanding and incorporation of others' SMCs into action prediction.

In this work, we investigate whether low-level social SMC signals can give rise to false-belief inferences, a phenomenon commonly believed to be caused by high-level cognitive reasoning. To this end, we develop a computational model that demonstrates how the tracking of social affordances allows for false-belief inferences. In order to include the temporal evolution of social interaction and make use of the predictive nature of cognition as proposed by the ST, we make use of a Bayesian generative model which, based on hidden variables and prior knowledge, selects the optimal action towards a given goal. While the presented model is generally applicable and not confined to false-belief tasks, we demonstrate its generative power with help of the well-known Sally-Anne test.

False-Belief Tasks and the Theory of Mind

The Sally-Anne story goes as follows. Sally and Anne are playing with a marble and two boxes, box A and B. Sally puts the marble into box A and leaves the room. In her absence, Anne takes the marble from box A and puts it into box B. Upon Sally's return, the question is: Where will Sally look for the marble? In a clinical or research setting, this story is often either demonstrated with help of a pair of dolls or illustrated with a comic stripe. Participants, asked where Sally will look for the marble, can give two answers. When passing the false-belief test, they successfully infer that Sally can by no means know that Anne moved the marble. Thus, they infer that Sally carries the false-belief that the marble is still in box A, as this is the location where she put it. Failing the false-belief test on the other hand implies that this inference is not accomplished. Instead, the actual current location of the marble, box B, is pointed out to be the goal of Sally's next action.

In an early study, Baron-Cohen et al. (Baron-Cohen, Leslie, & Frith, 1985) showed that healthy children and children with Down syndrome are able to pass the Sally-Anne test in 85-86 % in all cases, while autistic children pass only in around 20 % of all trials. These and other findings have

led to the belief that autistic children lack the ability to infer others' mental states and to develop, if at all, this trait later than their peers. Furthermore, even healthy children pass explicit false-belief tests only from an age of two-five, which is interpreted as a developmental account of a theory of mind (Apperly, 2012).

In the view of the TT and ToM, mind reading is the ability of humans to understand others' beliefs, desires, intentions and mental states by logically reasoning about their behavior with help of mental theories of the human mind. In the Sally-Anne story, Sally's desire to obtain the marble is hindered by her false-belief about the location of the object. The interpretation of ToM with respect to the role of mental states and beliefs differs within the research community (Apperly, 2012). This difficulty is only enhanced due to the fact that different false-belief tasks test varying aspects of ToM and subjects show diverging performances in different tests.

Hence, we claim here that TT can not fully account for the experimental evidence of false-belief tasks. Instead, the understanding of others' SMC signals as well as executive cognitive processes and low-level action constraints, such as spatial and temporal conditions and goal-directedness, have to be considered (Butterfill & Apperly, 2013).

Computational Approaches to False-Belief Tasks

In order to gain more understanding of the underlying dynamics of ToM, computational models can help to identify the essential variables that give rise to correct predictions. We will here focus on three models concerned with explicit (Goodman et al., 2006) and implicit (Berthiaume, Onishi, & Shultz, 2008) false-belief tasks and a Human-Robot Interaction (HRI) setting (Ferreira, Milliez, Lefevre, & Alami, 2015). While ToM has also been addressed with help of inverse reasoning, e.g. inversion of Partially Observable Markov Decision Processes (POMDP) (Baker, Saxe, & Tenenbaum, 2011), these approaches have not been applied to a false-belief setting.

A probabilistic account of ToM has been developed by Goodman et al. (2006). With the help of Bayesian networks, two models, the copy theorist (CT) and the perspective theorist (PT), are compared. Both of these models incorporate variables representing the world state, beliefs and desires of an actor, while only the perspective theorist has access to a variable indicating visual access. Manually defined, discrete probability distributions over mutual influences of these variables allow for the derivation of a posterior distribution over beliefs and desires given the observed events and actions. Additionally, the surprise about an observation can be determined. As the CT is less complex, it is not able to represent a false-belief, resulting in a high level of surprise when Sally is looking into the original box. The PT on the other hand, predicts the false-belief correctly. Due to hand-picked probability distributions and the additional information of visual access, the PT succeeds in the false-belief task. However, the superimposed structure, simplicity of the model and lack of temporal dynamics prohibit reliable conclusions about the applicability of this model.

In contrast to the probabilistic viewpoint Berthiaume et al. (2008) approached the implicit ToM, the idea that humans automatically and implicitly track others' mental states, with a neural network. To train networks with different numbers of hidden unit layers, they presented the models with the state and action data of an implicit Sally-Anne version. The training data was corrupted by adding incorrect samples. While networks with no hidden units were not able to capture false-beliefs, deeper networks could more reliably predict the behavior of the agents. Due to the nature of neural networks, the performance increased with an increasing amount of training samples. Furthermore, the results hint that the networks represented the statistics of the generated training data, with error rates matching the added noise. Although implicit knowledge should be more intuitive than explicit, conscious reasoning, it is dubious that this ability does only depend on correlations of observations. Instead, the internal motivation to predict the actions of other agents and to engage in interaction seems to be an important factor.

The examples introduced above function on the basis of belief-desire inference on the one hand and correlations on the other hand. As stated in the introduction, we propose taking a sensorimotor approach towards false-belief inferences. One recent example of this idea has been introduced by Ferreira et al. (2015) in a HRI setting. Applying two independent POMDP for the robot and human action space respectively, which use estimates of visual and reachable space to determine hidden state information, this system is able to interact with a human in a false-belief setting. By comparing manually designed and learned behavior, the authors conclude that learning results in faster and more reliable predictions. Furthermore, a system that incorporates knowledge about the humans belief space reacts faster towards misunderstandings. Nevertheless, the differences in performance between belief incorporation and its absence are not significant. The advantages in conversation speed might be balanced by the additional computational load during the learning period. Additionally, the focus lies on successful communication while predictive power and a deeper analysis of the system are not presented.

A Sensorimotor Approach to Social Inference

The application oriented approach towards false-belief inference based on visibility and reachability by Ferreira et al. (2015), points towards the explanatory power of sensorimotor signals in a social context. Even Goodman et al. (2006) and Berthiaume et al. (2008) conclude that visual access is an important factor. Without this variable, both belief and desire are not sufficient to account for false-belief inferences.

Thus, we propose that the theoretical considerations concerning ToM in such simple false-belief tasks as the Sally-Anne test need to be revised. Instead of high-level reasoning

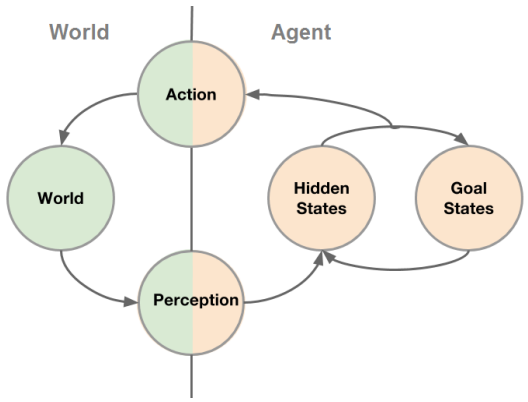


Figure 1: A simplified version of active inference: The agent is connected to the world through sensorimotor channels. Actions are chosen to minimize the distance between approximated hidden state and goal state distributions. The hidden states are thought to represent hidden causes of events in the world.

about mental states we suggest that tracking of SMCs and affordances over time is the primary factor for successful inference.

Inspired by the ideas of predictive coding (Kilner, Friston, & Frith, 2007) and active inference (Friston, Mattout, & Kilner, 2011) in the context of social cognition, we propose a generative model of optimal action selection which can also be employed for optimal action prediction of a co-actor. According to active inference, the human mind is prone to minimize uncertainty about the world state and sees actions as an inferential process to fit internal state distributions to goal distributions. The coupling between an agent and the world is accomplished through sensorimotor channels as shown in Figure 1. Since both actions and hidden causes can give rise to changes in the world, the hidden state distributions are approximating these hidden causes.

Assuming that interaction partners are equipped with a similar probabilistic inference machine, others' actions can be predicted with help of prior knowledge about their social SMC signals and hidden and goal states. Since the consideration of other agents reduces uncertainty about future world states significantly, these automatic mechanisms might account for many aspects of social cognition. Implicit tracking of others' sensory input and affordances is vital for both low-level and high-level interaction. We propose that failing in false-belief task is less caused by the lack of ToM than by the inability to identify and memorize others' social SMC histories. In this case, estimates of the hidden state distributions of the co-actor are impaired and have to be approximated by ones own internal distributions at the current time. As the hidden state distributions differ in the context of false-belief tasks, such an assumption leads to incorrect inferences. In the following, we will introduce a generative model that can imitate both of these behaviors.

The Model

Our generative model is based on a joint distribution over observations and hidden states. Actions based on the current hidden state estimates will have an effect on the actual environment. Optimal action selection is performed in two steps. Firstly, the current observations are incorporated into the distribution over hidden states. Secondly, the optimal action, which minimizes the distance between a given goal distribution and the updated hidden state distribution, is sought. The found optimal action is then executed. In a Bayesian fashion, prior beliefs can be incorporated into the hidden state distribution and affect action selection in a top-down manner. In a further step, this model can be used to predict the actions of an interaction partner by using the same mechanisms but different distributions over hidden states. Thus, instead of inferring beliefs and desires, action prediction is based on approximations of hidden states that represent SMC signals.

Mathematical Notation To clarify the mathematical notation, let $\mathbf{v} \in \mathbb{R}^n$ be a column vector of dimension n and $\mathbf{v}^T$ denote the transpose. Equivalently, let $\mathbf{M} \in \mathbb{R}^{n,m}$ represent a matrix with n rows and m columns. The dimensions of a vector are indexed by v^i for $i \in [1,..n]$, while a matrix entry is indexed by $m^{i,j}$ for $i \in [1,..n]$ and $j \in [1,..m]$. Furthermore, let the index $t \in [1,..T]$ indicate the time step ranging from 1 to T, such that $\mathbf{v}_t$ is a vector at time t. The time index for a single vector in a set of N vectors $\{\mathbf{v}_t\}_N$ is represented by $\mathbf{v}_{t,k}$ for $k \in [1,..N]$.

Our generative model consists of a discrete vector representing the actual world state $\mathbf{w} \in \mathbb{R}^{N_w}$. Due to the noise inherent in perceptual channels, let an observation $\mathbf{o} \in \mathbb{R}^{N_o}$ be a representation of the world state with added noise such that for each dimension i we have

$$o_t^i = w_t^i + \varepsilon^i, \quad \varepsilon^i \sim \mathcal{N}(0, \sigma_i), \tag{1}$$

where $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution with mean μ and variance σ and σ_i is the variance belonging to the ith dimension. Let the hidden states be represented by a set of N_s normal distributions, denoted by $\{\mathbf{s}_t\}_{N_s}$, and each kth hidden state consist of N_f features such that

$$\mathbf{s}_{t,k} \sim \mathcal{N}(\mu_k, \Sigma_k), \tag{2}$$

with mean $\mu_k \in \mathbb{R}^{N_f}$ and covariance matrix $\Sigma_k \in \mathbb{R}^{N_f, N_f}$. Finally, an agent is equipped with a set of N_a actions a_k for $k \in [1,..N_a]$ that produce changes in the environment.

Integration of new observations In this discrete setting, the joint probability distribution over hidden states and observations $P(\mathbf{o}_t, \{\mathbf{s}_t\}_{N_s})$ at time t is defined as a mixture of Gaussians

$$P(\mathbf{o}_t, \{\mathbf{s}_t\}_{N_s}) = P(\mathbf{o}_t | \{\mathbf{s}_t\}_{N_s}) P(\{\mathbf{s}_t\}_{N_s}) \tag{3}$$

$$= \sum_{k=1}^{N_s} P(\mathbf{o}_t | \mathbf{s}_{t,k}) P(\mathbf{s}_{t,k}). \tag{4}$$

Assume that the conditional distribution of the observation given the hidden states is multinomial distributed. Let the parameters of this distribution be a linear combination of the observations with a weight vector $\mathbf{u}_i \in \mathbb{R}^{N_o}$ that depends on the respective hidden state such that

$$P(\mathbf{o}_t|\mathbf{s}_{t,k}) = \frac{\mathbf{u}_k^T \mathbf{o_t}}{\sum_{l=1}^{N_s} \mathbf{u}_l^T \mathbf{o_t}}. \quad (5)$$

If the weight vectors have been determined, through learning or manual design, the update of the joint distribution $P(\mathbf{o}_{t+1}, \{\mathbf{s}_{t+1}\})$ is accomplished by inserting the new observation $\mathbf{o}_{t+1}$ into Eq. 3.

Optimal action selection Action selection is based on the minimization of the distance between a given goal distribution $P_{goal}(\mathbf{o}^*_{t+1}) \sim \mathcal{N}(\mathbf{o}^*_{t+1}, \sigma_{goal}\mathbf{I})$ and the joint probability distribution $P(\mathbf{o}_{t+1}, \{\mathbf{s}_t\}_{N_s})$. Let this distance be defined as the L^2 vector norm of the distance between the mean vectors of both distributions. Since this model is operating in a defined action space, let us assume, that an action a_k produces a discrete, deterministic hidden state $\hat{\mathbf{s}}_{t+1} = Q(\mathbf{s}_{t+1}|\{\mathbf{s}_t\}_{N_s}, a_k)$, where Q denotes a transition function. Then, Eq. 4 at time $t+1$ turns into

$$P(\hat{\mathbf{o}}_{t+1}, \hat{\mathbf{s}}_{t+1}) = \sum_{k=1}^{N_s} P(\hat{\mathbf{o}}_{t+1}|\hat{\mathbf{s}}_{t+1}) \mathcal{N}(\hat{\mathbf{s}}_{t+1} : \mu_k, \Sigma_k), \quad (6)$$

where $\mathcal{N}(\hat{\mathbf{s}}_{t+1} : \mu_k, \Sigma_k)$ denotes the kth Gaussian evaluated at $\hat{\mathbf{s}}_{t+1}$. Define $\mathbf{p} \in \mathbb{R}^{N_s}$ to be a vector consisting of the evaluations of the N_s Gaussian distributions. Then, the approximated observation is given by

$$\hat{\mathbf{o}}_{t+1} = \mathbf{U}^{-T}\mathbf{p}, \quad (7)$$

where the matrix $\mathbf{U}$ consists of the stacked weight vectors $\mathbf{u}$. The optimal action is therefore selected by determining which resulting approximate observation minimizes the L^2 vector norm

$$a^* = \underset{a_k,\, k\in[1,..N_a]}{\operatorname{argmin}} ||(\hat{\mathbf{o}}_{t+1}|a_k) - \mathbf{o}^*_{t+1}||_2^2. \quad (8)$$

When the optimal action is performed, it results in a change of the environment as follows

$$\mathbf{o}_{t+1} = P(\mathbf{o}_{t+1}|\mathbf{s}_{t+1}) Q(\mathbf{s}_{t+1}|\{\mathbf{s}_t\}_{N_s}, a^*). \quad (9)$$

Action prediction Up to this point, the discussion was focused on the optimal action selection for a single agent. In a joint setting, the same mechanism can be used to predict the action of a co-actor. If the observing agent memorizes not only its own past hidden state distributions, but also those of its partner, the observer can use its internal models to make inferences based on this information. As we assume the hidden states to represent SMC signals and affordances, many of the hidden states are either identical or complementary, such that the tracking of the additional states is computationally cheap. Thus, we assume that any agent keeps a memory over its own hidden states up to the current time T, i.e. $\{\mathbf{s}_{1:T}\}^{own}$, and an approximation of its partners hidden states, i.e. $\{\hat{\mathbf{s}}_{1:T}\}^{other}$. This approximation is based on the information made available by the observations and prior assumptions. Since the perceptual channels of the observed agent are inaccessible, the input can only be estimated. But as both agents share the same environment, these estimations can be made under low uncertainty.

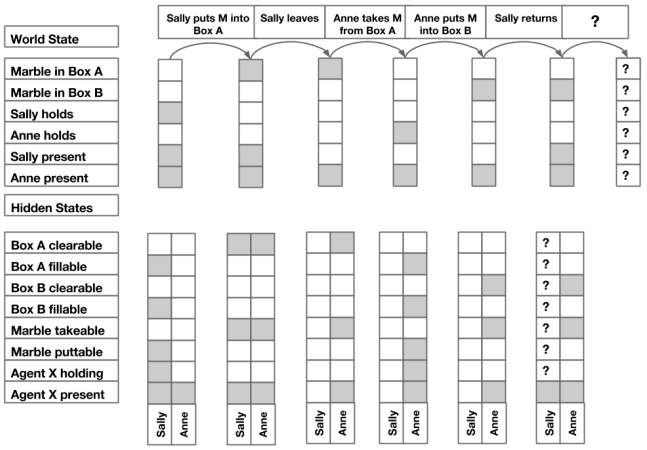


Figure 2: The Sally-Anne story enacted by our generative model. For the purpose of visualization, the noise has been removed. Dark squares indicate a value of 1 and white squares a value of 0. In the upper row, the dynamics of the story in the world state are depicted, while the lower row shows the hidden states of both agents in the view of Anne. Since Sally is gone for several time steps, her hidden states are unknown to Anne.

Experiment

The main goal of this work is to demonstrate that low-level sensorimotor signals can account for inferences in a false-belief task which are usually explained by high-level reasoning based on ToM. As the generative model presented above allows predictions of others' actions based on approximations of hidden states which represent these low-level signals, we can apply it to the Sally-Anne test. Following the idea of past affordance tracking, suggested by Brincker (2014), we aim to show that both passing and failing this false-belief task can be accounted for by prior assumptions about others' past affordances. In this scenario, we need to incorporate the two agents Sally and Anne, the objects box A, box B and marble and finally the room itself.

World Representation In order to test our generative model in a false-belief setting, we need to specify the variables world state and hidden state and the transition function describing effects of applied actions. While the former two variables are defined to be binary, with feature 0 = false and 1 = true, the later represents changes in the hidden states and world state.

Let the world state be defined as a vector with $N_w = 6$,

representing the features *marble in box A*, *marble in box B*, *marble held by Sally*, *marble held by Anne*, *Sally present* and *Anne present*.

Furthermore, let each hidden state consist of $N_f = 9$ features, *box A is clearable*, *box A is fillable*, *box B is clearable*, *box B is fillable*, *marble is takeable*, *marble is puttable*, *room is leavable*, *agent X is holding* and *agent X is present*.

Finally, let the $N_a = 4$ actions be defined as *agent X takes marble from box Y*, *agent X puts marble into box Y*, *agent X leaves room* and *agent X enters room*, where X $\in$ {Sally, Anne} and Y $\in$ {A, B}.

We defined the parameters of the Gaussian distribution in the mixture model and the mapping parameters of the multinomial distribution and the transition function manually. In a more application oriented setting, these parameters could also be learned either with help of learning by demonstration or self-learning.

Sally-Anne test As our generative model is generally applicable, we were able to let the model itself enact the Sally-Anne story instead of predefining the variables manually. By defining the goal observations at each time point t according to the Sally-Anne story, the model determined the optimal action and updated the distributions in accordance with the mechanisms described above. This procedure was iteratively performed up to the point, where Sally returns to the scene. Fig. 2 illustrates the Sally-Anne story in this format.

What will Sally do next?

Upon Sally's return, we ask the question "What will Sally do next?". In order to answer this question, we made use of Anne's inferential model since this agent has knowledge about the current situation and Sally's past hidden states. Given that Sally has been absent, her hidden state representation at time t, i.e. $\{\hat{\mathbf{s}}_t\}^{Sally}$ is unknown to Anne. However, with the aim to predict her next action given that she wants to hold the marble, the hidden states need to be approximated. Anne has two alternatives. Either she replaces the missing information with her own representation at time t

$$\{\hat{\mathbf{s}}_t\}^{Sally} = \{\mathbf{s}_t\}^{Anne} \tag{10}$$

or she recalls Sally's representation at time $\tau < t$ when Sally was last present in the room, i.e.
$\tau = \max_{t'}(t' \in [1,...t] : \{\hat{\mathbf{s}}^9_{t'}\}^{Sally} == 1 | t' < t)$,

$$\{\hat{\mathbf{s}}_t\}^{Sally} = \{\hat{\mathbf{s}}_\tau\}^{Sally}. \tag{11}$$

While this approach is similar to the copy theorist (Eq. 10) and the perspective theorist (Eq. 11) as introduced by Goodman et al. (2006), notice that our approach is not based on the notion of beliefs and desires. Furthermore, we introduce temporal dynamics which make a fluent interaction possible, while Goodman et al. (2006) work in a static environment with predefined variables and distributions. Since our idea is not following the TT approach but the ideas of ST, our agents can not be viewed as theorists. Therefore, we denote the approach of Eq. 10 as *mapping* and the past affordance incorporation in Eq. 11 as *tracking*.

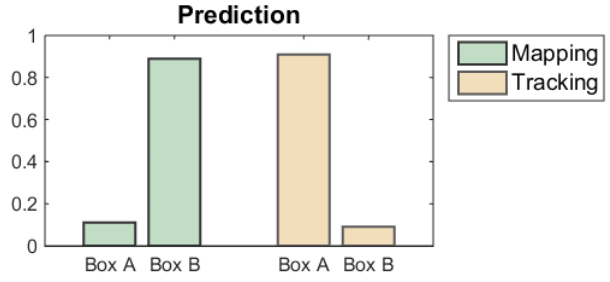


Figure 3: Predictions of the box that Sally would select based on Anne's generative model. To the left, the mapping approach demonstrates that the inference of a false-belief fails, while the tracking approach to the right correctly predicts the box into which Sally had put the marble before leaving.

Results

For the sake of reliability, we performed $N = 1000$ trials of the Sally-Anne test. Due to the induced noise in the mapping from world state to observation and the variance of the Gaussian distributions representing the hidden states, the model did occasionally fail to complete the Sally-Anne story up to the point where Sally returns, i.e. it selected incorrect actions. These trials were omitted and replaced by a newly generated, successful trial. The average of the predicted location where Sally will look for the marble is depicted in Fig. 3 for both the mapping and the tracking approach. While the mapping approach incorrectly predicted that Sally would look at the actual location in 89 % of all trials, the tracking approach correctly inferred Sally's behavior in 91 % of all trials. These results match closely those found in e.g. healthy (tracking) and autistic (mapping) children as reported in Baron-Cohen et al. (1985).

Discussion

In this work, we presented a generative model based on social SMCs which can be employed for both optimal action selection and prediction. Instead of mental state, belief and desire inferences, we hypothesize that SMCs can account for complex social behavior such as the recognition of a false-belief. In this context, the tracking of others' past affordances gives rise to successful inferences, while a failure of these basic sensorimotor functions results in incorrect predictions.

Why would the inference and tracking of others' SMCs be of advantage compared to belief and desire inferences? First of all, the two approaches have not to be seen as contradictory but as complementary. While the social SMC approach might be involved to a great extent in social behavior, high-level reasoning is also a non-negligible part of interaction. Instead, we argue for a shift from pure ToM reasoning towards the integration of essential, sensorimotor functions. As shown in this

work, social SMC signals can account for complex interaction scenarios, while a lack or impairment of these functions leads to the impoverished social capabilities found in infants and autistic children. A focus on the entanglement between low- and high-level cognition in a social context might reveal important information for medical and therapeutic research.

One could argue that the memory of the world state alone would result in the same predictions. Instead of inferring an affordance space, the mere knowledge where the marble had been when Sally was present could suffice. It is important to keep in mind, that the world and agent are two separate entities, coupled through sensorimotor channels. Thus, the agent has no direct access to the location of the marble but only to the hidden state representation of hidden causes in the environment. Without the inference of the co-actor's representation, successful prediction is impaired since the representations of observer and observed are entwined but not identical.

How can such representations of others be acquired? Similar to other computational approaches towards false-belief tests, such as Berthiaume et al. (2008) and Goodman et al. (2006), we defined many parameters manually. However, with a sufficient amount of training data, the model could be learned in an adaptive fashion and be generally applicable in dynamic interaction scenarios. As two agents share a considerable amount of the hidden state space, shared latent variable models are one method that could be applied to this method. While Ferreira et al. (2015) implemented a dynamic, socially interactive system, they did not account for this redundancy in the data. A putative future extension of the presented model is therefore an actively learning system which detects shared and individual latent manifolds in the action space that effectively encode action possibilities. Such an approach is both data efficient and reduces the complexity of high-dimensional interaction spaces. Models based on social SMCs can then be employed in interactive agents that are able to master complex scenarios as e.g. a false-belief setting.

Conclusions

The Theory of Mind has a long tradition to account for complex, social behavior. Mental state inferences and the reasoning about beliefs and desires are viewed as a mile stone in mental development. Advances of the idea of embodied intelligence however have started to give a complementary explanation of social phenomena. In this work, we demonstrated how the tracking of others' SMCs and affordances can be involved in certain false-belief inferences. This low-level approach to high-level cognition can clear the way for artificial agents in which social intelligence emerges naturally through the coupling between action and perception. Furthermore, a deeper insight in the underlying dynamics of social interaction results in valuable information for medical and psychological research and applications. In conclusion, sensorimotor signals are vital for social interaction. Their incorporation into theoretical frameworks of social intelligence is an important step towards an embodied understanding of social communication.

Acknowledgments

References

Apperly, I. A. (2012). What is theory of mind? concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, *65*(5), 825–839.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind?". *Cognition*, *21*(1), 37–46.

Berthiaume, V. G., Onishi, K. H., & Shultz, T. R. (2008). A computational developmental model of the implicit false belief task. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society.* (pp. 825–30).

Brincker, M. (2014). Navigating beyond "here and now" affordances - on sensorimotor maturation and "false belief" performance. *Frontiers in Psychology*, *5*.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637.

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, *85*(5), 1777–1794.

Dindo, H., Donnarumma, F., Chersi, F., & Pezzulo, G. (2015). The intentional stance as structure learning: a computational perspective on mindreading. *Biological Cybernetics*, *109*(4-5), 453–467.

Ferreira, E., Milliez, G., Lefevre, F., & Alami, R. (2015). Users belief awareness in reinforcement learning-based situated human-robot dialogue management. In *7th International Workshop on Spoken Dialogue Systems.*

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*(1-2), 137–160.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1382–1387).

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166.

The Illusion of Explanatory Depth in a Misunderstood Field: The IOED in Mental Disorders

Andrew S. Zeveney (asz214@lehigh.edu)
Department of Psychology, 17 Memorial Drive East
Bethlehem, PA 18015 USA

Jessecae K. Marsh (jessecae.marsh@lehigh.edu)
Department of Psychology, 17 Memorial Drive East
Bethlehem, PA 18015 USA

Abstract

Humans fail to understand the world around them and also fail to recognize this lack of understanding. The illusion of explanatory depth (IOED) exemplifies these failures: people believe they understand the world more deeply than they actually do and only realize that this belief is an illusion when they attempt to explain elements of the world. An unexplored factor of the IOED is how people may become overconfident by confusing their own understanding with others' understanding. In two experiments, we examine the IOED in mental disorders, a domain where society has a limited understanding. In Experiment 1, we demonstrate that people display an IOED for mental disorders as well as devices, but that it is smaller for mental disorders. In Experiment 2, we show that exposing the IOED is specifically linked to generating an explanation, rather than more generally thinking about a phenomenon.

Keywords: illusion of explanatory depth; domain differences; explanation

Introduction

Humans encounter a vast number of phenomena on a daily basis but only possess a shallow level of knowledge about most of these phenomena (Wilson & Keil, 1998). In addition to lacking an understanding of many everyday domains, people fail to recognize the limits of their own understanding, as demonstrated by the illusion of explanatory depth (IOED). First studied by Rozenblit and Keil (2002), the illusion of explanatory depth is the incorrectly held belief that one understands the world on a deeper level than one actually does.

The procedure used by Rozenblit and Keil to elucidate the IOED is simple. Participants were asked to rate their understanding of how devices worked (e.g., how a cylinder lock works). Next, for several of the items, they were asked to generate a detailed explanation of how each worked and to then rerate their understanding of that item. Consistently, ratings were lower after generating an explanation, suggesting that people do not understand the mechanisms of many everyday devices as much as they believe they do and only realize that they lack understanding after attempting to explain. The IOED has been shown in the domains of devices (Lawson, 2006; Mills & Keil, 2004; Rozenblit & Keil, 2002), natural phenomena (Rozenblit & Keil, 2002), and political policy (Alter, Oppenheimer, & Zemla, 2010; Fernbach, Rogers, Fox, & Sloman, 2013).

Why do people demonstrate an IOED? Explanations have typically centered on people's tendencies to overestimate the quality and depth of their mental representations and their failures to accurately judge their ability to provide good explanations (Mills & Keil, 2004; Rozenblit & Keil, 2002). Importantly though, people can and do rely on more than just their own understanding when generating explanations; they often seek out the understanding of others. People recognize that their own understanding has limits (even if they are inaccurate about where these limits lie) and use their understanding of what others know when choosing who to turn to fill in gaps in their own understanding (Keil, 2012; Wegner, 1987; Wilson & Keil, 1998).

Recent work suggests that because people so commonly rely on others for understanding, they may actually overestimate the amount of understanding that they possess when they cannot rely on others. For instance, Fisher, Goddu, and Keil (2015) demonstrated that people had greater confidence in their ability to explain various phenomena when they had first used the Internet to search for explanations for a different set of phenomena. They postulated that people's confidence in their own understanding came from conflating an understanding in the world with an understanding in their own mind. In another line of work by Kominsky and Keil (2014), people's illusory beliefs about their ability to generate differences in meaning between similar pairs of words were predicted by their beliefs that experts would be able to generate more differences. In general, this work suggests that the IOED may in part be confusing what others understand with what is understood in one's own mind.

Linking the IOED to the extent to which others understand a field has interesting implications for thinking about where an IOED might occur. For some domains, we as a collective society have a deep level of understanding of how things in that domain work. For example, there are people who understand devices like cylinder locks very well, even if the average layperson has a limited understanding of how a cylinder lock works. However, there are fields where understanding is fairly limited for everyone.

For instance, in the domain of mental health, there is limited understanding of how disorders should be defined (Barch & Keefe, 2009) or how treatments for mental health issues work (Wampold & Imel, 2015; Zhang & Malhotra, 2013). How would the IOED manifest in such domains where less is collectively understood about how things in this domain can be explained?

In the following set of experiments, we examine how our perceptions of what others understand about a domain influences the IOED for that domain. To explore this question, we examined the domain of mental disorders. Beyond the lack of scientific consensus on how symptoms interrelate in mental disorders to explain the etiology or presentation of a disorder (e.g., Matheson, Shepherd, & Carr, 2014), we have found that laypeople also recognize that understanding of mental disorders is limited. To measure this, we ran a pilot study where we asked a group of pretest participants (N = 50) to rate their beliefs about what we know about mental health, as well as devices, natural phenomena, and medical disorders. For each domain, participants rated their agreement that "we fully understand" the phenomenon (what we will call the society rating) on a scale of 1 (strongly disagree) to 7 (strongly agree). They also rated the difference between what an expert would know about each topic and what the average person knows about that topic (what we will call the gap rating) on a scale of 1 (no difference) to 7 (very great difference). We found that on average people reported higher societal understanding for devices (M = 5.73) than for mental disorders (M = 3.72; p < .001). People also reported a larger gap between lay and expert understanding for mental disorders (M = 5.64) than for devices (M = 3.73; p < .001). These findings provide evidence that laypeople recognize the difficulty we have in understanding mental disorders on a deep level. With these findings in mind, mental disorders serve as a prime domain in which to examine several questions about the illusion of explanatory depth.

Research Questions

Question 1 Do people demonstrate an IOED even in a field where understanding is limited?

Since people recognize that others, particularly laypeople, understand fairly little about mental health, one might anticipate that people will be more accurate about their own understanding. However, the IOED is a robust effect found in multiple domains, so people may still hold an illusion even when they recognize that there is limited understanding overall. The size of this illusion may be smaller in mental health than devices because people endorse a lower level of understanding for mental health at the outset.

Question 2 Do people's individual beliefs about others' understanding predict their own illusion?

If an individual's own beliefs about what others understand about a domain predicts the magnitude of his or her IOED for that domain, it would lend support to the hypothesis that the IOED occurs in part because people believe that understanding exists in the world and mistakenly believe that they have a greater portion of this understanding than they actually do.

Question 3 Is explanation necessary to expose the illusion?

Finally, mental health serves as a domain in which to further test the theory that it is specifically explanation that exposes the illusion (Rozenblit & Keil, 2002). Examining this is particularly important in mental health, as there is plenty of information about things like risk factors and symptoms but limited ability to explain how the interplay of these aspects leads to disorder. Is an attempt to explain necessary to reveal the illusion or will thinking about one's knowledge of disorder be sufficient?

Overview of Experiments

In Experiment 1, we presented participants with a number of phenomena including mental disorders and devices and asked them to rate their understanding before and after attempting to explain each phenomenon. Additionally, we asked participants to rate their beliefs about what others understand. In Experiment 2, we asked participants to rate their understanding of mental disorders but varied whether they were asked to explain the disorder or describe features of each disorder.

Experiment 1

In Experiment 1, we examined the IOED in mental health items using the basic protocol established by Rozenblit and Keil (2002). If people do have an illusion of their own understanding for mental disorders, we would expect to see a drop in ratings after attempting to explain these items. We also tested the IOED in devices in order to have a baseline to which to compare the potential magnitude of the illusion in mental health.

We also examined the correlations between beliefs about the general level of understanding and ratings of personal understanding. If people confuse their own understanding with what others understand, we anticipated that perceptions of the amount of societal understanding would be correlated with higher perceived personal understanding. Additionally, we examined the correlations between beliefs about the gap between lay and expert knowledge and ratings of personal understanding. We anticipated that a perception of a smaller gap between lay and expert knowledge would allow for more confusion between personal understanding and others' understanding, given that understanding is more widely spread across individuals, rather than being concentrated in experts. Thus, we expected that perceptions of a smaller gap would be correlated with higher ratings of personal understanding.

Methods

Participants 150 participants recruited through Amazon's Mechanical Turk participated for payment.

Materials and Procedure Participants were assigned to one of three conditions: devices (n = 50), mental disorders (n = 49), or mental health treatments (n = 51). The mental health treatment condition was included for comparisons not presented here and will not be discussed further. Participants were presented with a five-item stimuli set consisting of items from the given domain (see Table 1 for examples of all stimuli).

Table 1
Example stimuli in each domain

Domain	Stimuli
Devices	How a zipper works
Mental disorders	How the different symptoms of depression develop
Mental treatments	How an antipsychotic works

Participants were instructed on how to rate their understanding using a 1 (low) to 7 (high) understanding scale adapted from Rozenblit and Keil (2002). They were given an example of a high, moderate, and low level of understanding. They then rated the stimuli set on this scale (Time 1 [T1] rating). Following, participants were again presented with the five items, one by one, and prompted to explain each item in as much detail as possible. After each explanation, participants rated their understanding of that phenomenon on the understanding scale (Time 2 [T2] rating).

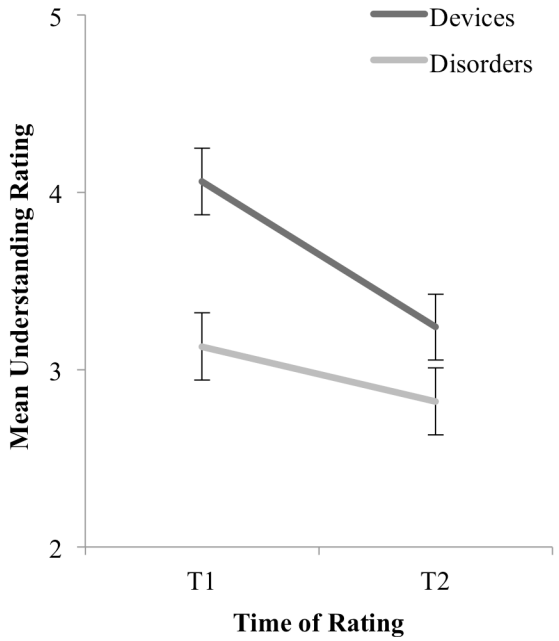
Participants then answered several follow-up questions designed to assess beliefs about others' understanding of each domain. For each of the items in the original set, participants rated their agreement that "we fully understand" the phenomenon ("society rating", rated from 1, strongly disagree, to 7, strongly agree). They also rated the size of the difference between what an expert would know about each topic and what the average person knows about that topic ("gap rating," rated from 1, no difference to 7, very great difference). Additionally, participants answered basic demographic questions.

Results

We examined ratings for devices and mental disorders using a 2 (time: T1 vs. T2; within) x 2 (domain: devices vs. mental disorders; between) mixed ANOVA. We found a significant effect of time, $F(1, 97) = 40.96$, $p < .001$, $\eta_p^2 = .30$, with ratings of understanding decreasing from T1 (M = 3.59, SD = 1.40) to T2 (M = 3.03, SD = 1.33). We also found a significant effect of domain, $F(2, 97) = 7.23$, $p = .008$, $\eta_p^2 = .07$. Overall, participants rated their understanding of devices (M = 3.65, SD = 1.24) as significantly higher than their understanding of disorders (M = 2.98, SD = 1.25).

Finally, there was a significant interaction between time and domain, $F(2, 97) = 3.41$, $p = .036$, $\eta_p^2 = .044$. Sidak-corrected follow-up tests showed the expected illusion in devices, as ratings significantly dropped from T1 to T2, $p < .001$. Additionally, people also demonstrated an illusion in mental disorders, $p = .014$, as ratings dropped significantly from T1 to T2 in this domain as well (see Figure 1). In order to compare the size of the illusion across domains, we examined this interaction another way. At T1, ratings for devices were significantly higher than those for mental disorders, $p = .001$. However, at T2, ratings did not differ between devices and mental disorders, $p = .112$. Given that ratings for devices were greater than mental disorders at T1 but the same at T2, participants showed a greater drop (and thus a larger illusion) for devices than for mental health phenomena.

Figure 1: The IOED in devices and mental disorders



Next, we examined the society and gap ratings. Participants rated society as understanding devices, M = 4.42, SD = 1.57, significantly more than mental disorders, M = 3.55, SD = 1.41, $t(97) = 2.92$, $p = .004$. Participants also rated the gap between expert and lay knowledge as significantly larger for disorders, M = 5.71, SD = 1.15, than for devices, M = 4.84, SD = 1.10, $t(97) = 3.88$, $p < .001$. These findings replicated our pilot data.

We then examined the correlations between society/gap ratings and ratings of personal understanding at T1 and T2 across domains. Society ratings were correlated with personal understanding of disorders at T1 ($r = .310$, $p = .032$) and T2 ($r = .400$, $p = .005$). For devices, society ratings were also correlated with personal understanding at

T1 (r = .290, p = .043) and T2 (r = .412, p = .003). Perceptions of higher societal understanding were related to higher individual understanding both before and after generating an explanation.

For gap ratings, a slightly different pattern emerged. For mental disorders, smaller perceived gaps were correlated with higher personal understanding for mental disorders at T1 (r = -.315, p = .029) and T2 (r = -.491, p < .001). However, gap ratings were not correlated with personal understanding ratings for devices (rs < .036, ps > .810). Thus, the anticipated relationship between smaller gaps and higher personal understanding was found only in mental disorders.

Discussion

As in devices, people demonstrate an illusion of explanatory depth in the domain of mental disorders, even though they endorse less general understanding and in particular, less lay understanding of disorders than devices. This finding serves to validate the robustness of the IOED: even when people admit that a field is not particularly well understood and that understanding is more concentrated within experts, they still overestimate their own understanding.

Importantly, the illusion was larger in devices than in mental disorders, as personal understanding was higher at T1 for devices than for mental disorders but not at T2. Potentially, one explanation is that people confuse their own understanding of a phenomenon with the understanding that others have for that phenomenon. Given that people see mental health as less understood by others, people may initially experience less confusion between others' understanding and their own in mental disorders than in devices. As a result, ratings of personal understanding for mental disorders at T1 are closer to actual levels of understanding, leading to a smaller illusion in mental disorders than devices.

Relatedly, the illusion in mental disorders may also be smaller because people see understanding as more concentrated within experts for disorders than for devices. People may be less likely to confuse their understanding with others' understanding when others' understanding is seen as mostly possessed by experts on a domain, rather than being distributed more evenly among all people. Future research should explore perceptions of both the level of understanding that others possess and the distribution of this understanding across others.

Additionally, this experiment sheds light on the potential relationships between individuals' perceptions of others' understanding and of their own understanding. For mental disorders, perceptions of societal understanding and the lay/expert gap were related to personal understanding at both T1 and T2, and for devices, societal and personal understanding showed the same relationships. Importantly, the same relationships were found both before generating an explanation and after, when ratings of personal understanding are more accurate. Thus, it is not that individuals' beliefs about what society understands and about the differences between expert and lay understanding simply relate to their illusion of understanding: these beliefs are correlated to perceptions of understanding even after the illusion is shattered through explanation. One thing that is important to note is that society and gap ratings were completed only after participants had generated an explanation. Potentially, generating this explanation altered beliefs about what others understand and therefore, obscured the true relationship between beliefs about others' understanding and personal understanding.

Given that the IOED in mental disorders is smaller than in devices, another possible question arises of whether this illusion is similar in nature to the IOED shown in other domains. Prior work on the IOED has suggested that it is only through attempting to explain a phenomenon that one is made aware of one's own limits in understanding (Rozenblit & Keil, 2002). But when considering many phenomena, including mental disorders, people can easily think about these phenomena in ways that do not entail detailed explanations (e.g., thinking about all of the symptoms of a disorder without thinking about how they are interrelated to create an explanation of how the disease came into existence). Is this type of deeper thinking sufficient to expose the illusion or is it actually necessary to generate an explanation? We tested this possibility to determine if the IOED in mental health disorders is reliant on explanations as it is in other domains.

Experiment 2

In Experiment 2, we tested whether explanation is specifically necessary to expose the IOED, or whether simply thinking deeply about a phenomenon more generally is enough. We also tested a wider range of mental disorders, in order to further extend the finding that people do demonstrate an IOED for mental health issues.

Methods

Participants 100 participants recruited through Amazon's Mechanical Turk participated for payment.

Materials and Procedure

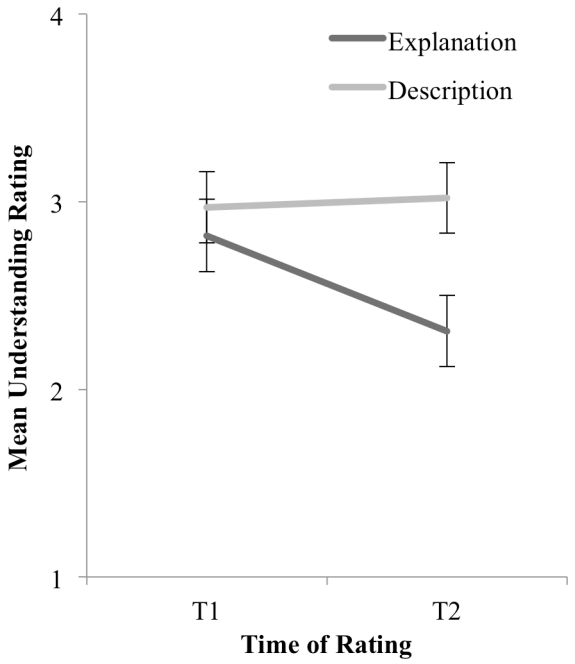
Participants were assigned to either the explanation condition (n = 50) or the description condition (n = 50). All participants were presented with eight mental disorders: depression, OCD, generalized anxiety disorder, schizophrenia, borderline personality disorder, bipolar disorder, ADHD, and anorexia. For a given disorder, participants were presented with the name of a disorder and then were asked to rate their understanding with the same question as in Experiment 1 (T1 rating). Participants in the explanation condition then were asked to explain each disorder as in Experiment 1. Participants in the description condition were asked to list as many characteristics about

the disorder as they could. After completing the explanation or the description task, participants completed the T2 rating from Experiment 1. Participants then completed the society and gap ratings from Experiment 1. Finally, participants rated each disorder in terms of how biologically, psychologically and environmentally caused they were on a scale of 0 (not at all) to 100 (completely). Causation questions were not the focus of the work presented here, so they will not be discussed further.

Results

We examined ratings using a 2 (time: T1 vs. T2) x 2 (condition: explanation vs. description) mixed ANOVA. We found a significant effect of time, $F(1, 98) = 6.79$, $p = .011$, $= .065$, and a marginally significant effect of condition, $F(1, 98) = 2.82$, $p = .011$, $\eta_p^2 = .065$. These main effects should be interpreted in light of a significant interaction between time and condition, $F(1, 98) = 4.12$, $p = .002$, $\eta_p^2 = .097$. Sidak-corrected follow-up tests showed that in the explanation condition, there was a significant drop from T1 ($M = 2.98$, $SD = 1.33$) to T2 ($M = 2.31$, $SD = 1.27$; $p < .001$). However, for the description condition, there was not a significant difference between T1 ($M = 2.97$, $SD = 1.33$) to T2 ($M = 3.02$, $SD = 1.30$; $p = .650$; see Figure 2).

Figure 2: Experiment 2 mean understanding ratings.



Examining this interaction another way, ratings between conditions did not differ at T1, $p = .454$, but are significantly lower at T2 for the explanation condition than the description condition, $p = .002$. Thus, participants in both conditions endorsed the same initial level of understanding but only those who explained their understanding showed a drop at T2. The illusion was only revealed when people attempted to explain mental disorders, and not when they simply described their knowledge.

We also examined the correlations between society/gap ratings and personal understanding. For the description condition, society ratings were correlated with T1, $r = .370$, $p = .008$, and T2 ratings, $r = .408$, $p = .003$, and gap ratings were uncorrelated with T1 ratings, $r = -.171$, $p = .234$, and marginally correlated with T2 ratings, $r = -.269$, $p = .059$. In the explanation condition, participants showed a different pattern. Society ratings were uncorrelated with T1 or T2 ratings, rs > .190, ps > .135, and gap ratings were correlated with T1, $r = -.294$, $p = .042$, and T2 ratings, $r = -.316$, $p = .029$. Higher ratings of societal understanding were related to higher ratings of personal understanding but only in the description condition. Likewise, ratings of a larger lay/expert gap were negatively related to personal understanding only in the explanation condition.

Discussion

As in Experiment 1, participants demonstrated an IOED for mental disorders. However, this illusion was only exposed when people attempted to explain disorders, not when they were simply prompted to think about the aspects of the disorder. This finding provides further evidence that overconfidence in one's understanding can be revealed by attempting to explain that understanding, but by simply reflecting on aspects of the phenomenon that might be related to that understanding.

In this experiment, we also found that the relationships between perceptions of others' understanding and personal understanding differed across conditions. Personal understanding was correlated with gap ratings only in the explanation condition and with social understanding only in the description condition. This finding provides further evidence that generating an explanation changes people's later beliefs about others' understanding. The conditions were only different from one another after T1 ratings were made, so relationships for T1 ratings and others' understanding should be the same if perceptions of others' understanding are not changed by the experiment. Despite the difficulty of interpreting these results, overall our results suggest that ratings of societal understanding and the lay/expert gap are related to individuals' ratings of understanding.

General Discussion

Across two studies we explored whether an illusion of explanatory depth is demonstrated for a domain where people acknowledge that we have limited understanding of the domain. Our pretest study indicated that laypeople do in fact believe that we as a society have less of an understanding of mental disorders than devices and understanding is more concentrated within experts in the

mental health domain. In comparing the IOED across mental health disorders and devices, we saw a smaller IOED in disorders than devices (Experiment 1). This IOED seemed to be based in the same mechanism as previous research (Rozenblit & Keil, 2002) in that it only arose after explanation and not after mere description (Experiment 2). Despite the seeming link between mental disorders being seen as less well understood and a smaller IOED in the mental domain, ratings of others' understanding were correlated not only with initial illusory perception of personal understanding but also with personal understanding after generating an explanation.

Finding that people show a smaller IOED for mental health than devices suggests that though people are overconfident about their understanding in a number of domains, the extent to which they are overconfident may vary. In particular, mental health may be a domain in which people are more accurate about their understanding, and this accuracy may be driven by their perception that this domain is less understood in general. At the same time, there is still a gap between what people believe that they understand before and after attempting to explain.

Our findings suggest avenues for future research related to the IOED. For example, understanding what particular domain differences lead to the difference in size of the IOED across domains would be an important avenue to explore to better understand the underlying mechanism of the IOED. There may be certain domains in which laypeople would be unwilling to endorse any understanding, believing that only an expert can understand that particular subject. How unfamiliar would a domain need to be to laypeople for them to accurately assess their limited understanding? These are questions for future research that deserve exploring.

Furthermore, our research has interesting implications for how people reason about the mental health domain. It is interesting to think about how generating explanations as participants did in this study may alter how people think about their own diagnoses or treatments, and those of the people around them. For example, even if people are aware that they do not completely understand how the disorder of depression comes to be, our results suggest that they would still overestimate how much they understand about the disorder in general. This overconfidence may give people a false sense of understanding what treatment would best address a disorder's symptoms since they feel like they understand how the disorder works at least in some part. It is an interesting avenue for future research to explore how these false senses of understanding influence decision making in health.

Overall, our results are informative as to the extent of people's overconfidence in their own understanding. Even in the domain of mental disorders, where causal links are complex, multifaceted, and invisible, people still endorse a greater understanding than they actually possess. While it seems that in some ways, this illusion is supported by a belief that others in the world understand the domain, the relationship between our own understanding and others' understanding is potentially more complicated and deserves future research.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*, 436-451.

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24*, 939-946.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144*, 674-687.

Kominsky, J. F., & Keil, F. C. (2014). Overestimation of knowledge about word meanings: The "misplaced meaning" effect. *Cognitive Science*, *38*, 1604-1633.

Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, *34*, 1667-1675.

Matheson, S. L., Shepherd, A. M., & Carr, V. J. (2014). How much do we know about schizophrenia and how well do we know it? Evidence from the Schizophrenia Library. *Psychological Medicine*, *44*, 3387-3405.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*, 1-32.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: Research evidence for what works in psychotherapy* (2nd ed*.)*. New York: Routledge.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York, NY: Springer-Verlag.

Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines*, *8*, 137-159.

Zhang, J. P., & Malhotra, A. K. (2013). Pharmacogenetics of antipsychotics: Recent progress and methodological issues. *Expert Opinion on Drug Metabolism & Toxicology*, *9*, 183-191.

Influences of Speaker-Listener Similarity on Shadowing and Comprehension

Lynn K. Perry (lkperry@miami.edu)
Department of Psychology, 5665 Ponce De Leon Blvd
Coral Gables, FL 33124 USA

Emily N. Mech (emech@wisc.edu)
Maryellen C. MacDonald (mcmacdonald@wisc.edu)
Mark S. Seidenberg (seidenberg@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

We routinely encounter speakers with different accents and speaking styles. The speech perception literature offers examples of disruption of comprehension for unfamiliar speech and also of listeners' rapid accommodation to unfamiliar accents. Much of this research uses a single measure and/or focuses on isolated word perception. We investigated listeners' abilities to comprehend and shadow connected speech spoken in a familiar or unfamiliar accent. We found increases in shadowing latencies and comprehension errors in the Dissimilar Speech relative to Similar Speech conditions—especially for relatively informal rather than more academic style speech. Additionally, there was less accommodation over time to Dissimilar than Similar Speech. These results suggest that there are costs both in the immediate timescale of processing speech (necessary for shadowing) and in the longer time scale of listening comprehension when accent and other speech quality is very different from one's own speech.

Keywords: speech perception; accented speech; speech shadowing; listening comprehension

Introduction

In every perceptual domain, familiarity improves perceptual speed and accuracy. In speech, familiarity can take many forms: we can be familiar with the words spoken, with a particular speaker's voice quality, or with more general characteristics of a speech community such as accent and typical speaking rate. An important question to the study of speech perception, then, is how do listeners make sense of unfamiliar speech in order to communicate? In the current experiment, we examined the effect of familiarity of a talker's speaking style on listeners' ability to comprehend rapidly and accurately. We manipulated familiarity with recordings of two speakers reading aloud, one from the same speech community as the participants and one from a different speech community. We will refer to this manipulation as familiarity with the talker's speech style or "accent", recognizing that many factors affect listeners' perception of an individual's speaking style and its similarity to their own speech.

Previous research has demonstrated that non-native accents disrupt intelligibility of speech to native speakers, as do differences in regional accents (Adank, Evans, Stuart-Smith, & Scott, 2009; Floccia, Goslin, Girard, & Konopczynski, 2006; Major, Fitzmaurice, Bunta, & Balasubramanian, 2005). However, it is also clear that both adults and children can accommodate unfamiliar accents with sufficient exposure (Maye, Aslin, & Tanenhaus, 2008; Sumner, 2011; Sumner & Samuel, 2009; White & Aslin, 2011). The ease with which listeners can accommodate to unfamiliar speech patterns is often held as an example of how impressive human language abilities are. This accommodation is often measured via recognition of isolated words (e.g., Sumner, 2011) rather than sentences or longer utterances, and the degree of accommodation to continuous speech is less well understood.

Studies that have investigated perception of connected speech have often used speech shadowing tasks. Shadowing requires participants to repeat aloud the speech they hear as quickly and accurately as possible (e.g., Marslen-Wilson, 1973). Sabatini (2000) examined native Italian speakers trained as professional English-language interpreters' listening comprehension, shadowing, and simultaneous interpretation of nonstandard American English and Indian nonnative English. Consistent with previous findings, the unfamiliar accents were less intelligible to participants in all tasks (Sabatini, 2000). However, listening comprehension was easier than shadowing and simultaneous interpretation. The longer time scale of comprehension tasks is helpful because background knowledge and downstream speech information may clarify the interpretation of earlier input. Although top-down influences occur even on the shorter timescale in which shadowing occurs (e.g., Marslen-Wilson, 1985; Nye & Fowler, 2003), shadowing provides an index of the difficulties listeners encounter in early stages of processing unfamiliar speech. The Sabatini (2000) study clearly revealed that non-native listeners have difficulties processing unfamiliar accents. However, it is unclear whether such effects also occur for native speakers listening to unfamiliar speech, such as markedly different regional accent or voice quality.

We compared shadowing and offline listening comprehension measures to assess effects of speech familiarity at two timescales. Two native speakers of American English from different speech communities each recorded four text passages for use in shadowing and

comprehension tasks. Two passages were more academic in character and two were in a more informal narrative style. This manipulation of Passage Type was included to determine whether the impact of speaker variability depended on type of material. The academic passages contained more low frequency words, which could increase the difficulty of speech comprehension for these passages compared to more informal narratives.

Methods

Participants

59 (36 female) native English speakers from the Midwest United States participated. All participants identified themselves as white native English Speakers. 29 participants were assigned to the Similar Speech Condition (16 female), and 30 to the Dissimilar Speech Condition (20 female). Participants were recruited from introductory psychology classes at the University of Wisconsin-Madison and received course credit for their participation.

Stimuli

Four text passages, two academic and two informal, were used as stimuli. The academic passages were drawn from reading comprehension portions of the Test of English as a Foreign Language (TOEFL) exam. The informal passages were written transcriptions of two stories from the radio program "This American Life." Passages varied between 308 and 350 words in length.

Each passage was recorded by two female native English speakers, both graduate students in their mid-20s. The speaker in the Similar Speech Condition is a white woman from the Midwest, the same region as the participants, while the speaker in the Dissimilar Speech Condition is an African American woman from the southeastern US and identifies herself as a speaker of African American English as well as Mainstream American English. Speakers were informed that the recordings would be used for a study of speech perception and accents. They read through the passages before making the recordings and were instructed to speak as naturally as possible, which led to slight differences in length of each speaker's recordings (Similar Speech: M=134 seconds; Dissimilar Speech: M=133 seconds).

Each speaker's recordings were rated for familiarity by fourteen white participants from the upper Midwest who did not participate in the main experiment. Similar Speech Condition recordings were rated more similar to their own speech, $t(13)=4.00$, $p=.002$; compared to the versions for the Dissimilar Speech Condition. To be sure that the Dissimilar Speech recordings were not rated lower in intelligibility and familiarity for speaker specific—rather than dialect specific—reasons, an additional eleven African American participants, primarily from the Southern U.S., rated speaker familiarity. Critically, these participants rated the Dissimilar (i.e. African American English) Speech recordings as marginally *more similar* to their own speech than the white Midwestern participants had rated these recordings, $t(23)=1.83$, $p=.081$. The African American raters judged the recordings of the white speaker as significantly less similar to their own speech than the white Midwestern participants rated these recordings, $t(18)=3.78$, $p=.001$.

Procedure

Speech similarity (Similar, Dissimilar) was manipulated between participants. Task (shadowing, comprehension), and Passage Type (Academic, Informal) were manipulated within participants, with the order of conditions randomized for each participant. Passages were presented over headphones in a quiet lab room. Each participant heard all four passages once, with assignment of conditions counterbalanced across participants. All experimental tasks were run using E-prime 2.0 software. The experiment took about 20 minutes to complete.

Shadowing Task Participants were instructed to repeat the words in the passages as they were heard as quickly and accurately as possible, speaking into a microphone directly in front of them. One Academic passage and one Informal passage were presented.

Listening Comprehension Task Participants were instructed to listen to each passage and were told they would answer true/false questions afterwards to test their comprehension. After they heard each passage, they answered six true/false questions (presented one at a time) by pressing the T and F keys on the keyboard. One Academic passage and one Informal passage were presented.

Coding

Shadowing Task Two trained research assistants, blind to the experimental hypotheses, coded each participant's speech shadowing for errors and latency.

Errors: Any errors or deviations from the original transcript were coded as omissions, constructive errors, or delivery errors. Omissions were whole words that participants omitted in shadowing. Constructive errors included any added words or changes to words that resulted in a different word or a nonword. Delivery errors included slurred hesitations, stuttering, and unintelligible responses.

Latency: Every tenth word of participants' shadowing was coded for latency relative to the original transcript. Latency was measured at the word onset, which was determined by analysis of the speech spectrogram on Praat software.

Listening Comprehension Listening comprehension accuracy was measured for each participant as the total number of true/false questions answered correctly for each passage.

Analyses

All analyses were conducted using mixed effects regression models. Latency analyses were conducted using linear regression and accuracy analyses were conducted using logistic regression. To determine the best-fit model, we used chi-square tests comparing models with and without the factor of interest. For interactions, we report coefficients and confidence intervals from the full model, and the chi-square test of model fit from the comparison to a model with the interaction removed. For main effects, we report coefficients and confidence intervals from the full model, and the chi-square test of model fit from the comparison to a model with the predictor main effect removed. To determine appropriate random effects, we began with completely specified random effects structures including random slopes for all variables in a given model. Using model comparison, we systematically removed uninformative random effects (Jaeger, 2009). All final models included random intercepts for subjects and items.

Results

We first report performance in each task separately and then examine the relationship between them. Our primary question for each task is whether speech style influenced participants' ability to shadow and comprehend that speech.

Shadowing task

The principal measures concerned the effects of speech condition on speed and accuracy of shadowing responses. We first analyzed the quantity and types of errors that were made. We next examined the speed of participants' speech shadowing by measuring the latency of their productions relative to the stimulus onset.

Shadowing errors Participants' shadowing was highly accurate overall but they made a larger proportion of errors in the Dissimilar Speech Condition, M=.09, than in the Similar Speech Condition, M=.06. Although participants were more likely to make errors for academic speech passages, M=.10, than informal speech passages, M=.06, this difference was larger for those in the Similar Speech Condition (academic: M=.08 of words; informal: M=.04 of words) than in the Dissimilar Speech Condition (academic: M=.10 of words; informal: M=.08 of words). Model comparisons revealed a main effect of speech condition, b=.34, 95% CI[.02, .67]; $X^2(1)$=11.46, p<.001, such that participants in the Similar Speech Condition were more accurate than participants in the Dissimilar Speech Condition, and a main effect of passage type, b=-.91, 95% CI[-1.14, -.68]; $X^2(1)$=8.49, p=.004, such that participants were more accurate at shadowing academic than informal speech. The interaction between passage type (academic v. informal) and speech condition (similar v. dissimilar) was also significant, b=.59, 95% CI[.43, .75]; $X^2(1)$=51.54, p<.00001.

Planned follow-up comparisons revealed that the interaction was driven primarily by participants performance on informal passages, with a highly significant effect of speech condition on informal passages, b=.90, 95% CI[.56, 1.24]; $X^2(1)$=23.35, p<.001, but only a marginally significant effect of speech condition on academic passages, b=.34, 95% CI[-.03, .71]; $X^2(1)$=3.18, p=.07. Additional follow-up comparisons revealed that both the participants in the Similar Speech Condition, b=-.34, 95% CI[-.46, -.22]; $X^2(1)$=6.22, p=.02, and those in the Dissimilar Speech Condition, b=-.92, 95% CI[-1.21, -.62]; $X^2(1)$=8.39, p=.004, demonstrated a significant effect of passage type, such that they made fewer errors on academic than informal passages.

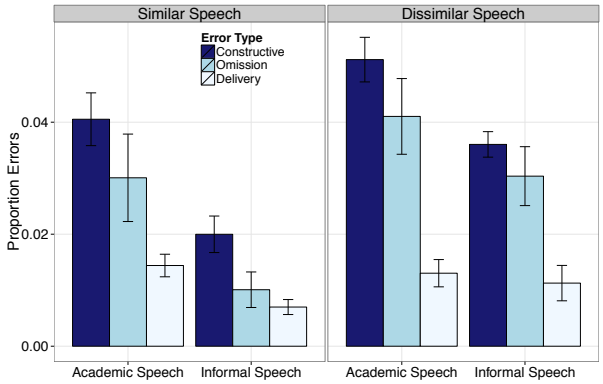


Figure 1. Proportion of errors of each type made in the shadowing task by participants in each speech condition for each passage type. Error bars depict standard error of mean.

We next examined the types of errors that were made. As seen in Figure 1, constructive errors were most frequent, M=.04 of words, followed by omissions, M=.03, and then delivery errors, M=.01. Model comparison revealed a three-way interaction between speech condition, passage type, and error type (constructive, omission, delivery) on participants' accuracy, $X^2(2)$=6.15, p=.05. Follow-up comparisons revealed that this interaction was carried by the interaction between speech condition and passage type, as there was a significant interaction between speech condition and passage type on the number of constructive, $X^2(1)$=10.63, p=.001, omission, $X^2(1)$=37.36, p<.001, and delivery errors, $X^2(1)$=10.19, p=.001. This result suggests that participants show a somewhat similar pattern of errors for each speech condition and passage type.

Finally, we examined whether accuracy changed over time because participants were able to adapt to the speakers' voice and improve in their shadowing performance. Model comparisons revealed that the interaction between speech condition and block was not significant, $X^2(1)$=2.13, p=.14, nor was the three-way interaction between speech condition, block, and passage type, $X^2(1)$=.40, p=.53, suggesting little effect of speech style on accommodation.

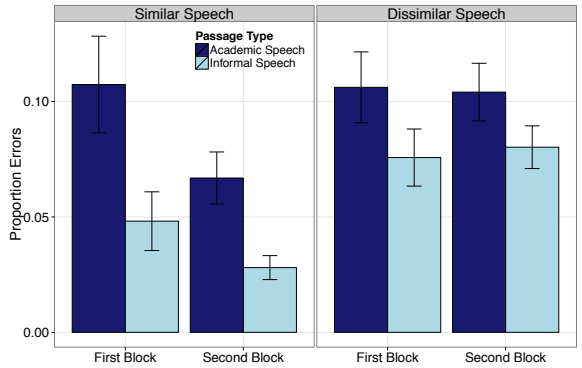


Figure 2. Proportion of errors participants made in each speech condition during shadowing of each passage type depending on whether they did the shadowing task first or second. Error bars depict standard error of the mean.

Shadowing latency In addition to making more errors, participants in the Dissimilar Speech Condition also shadowed more slowly, *M*=1108ms, *SD*=.513; than participants in the Similar Speech Condition, *M*=1032ms, *SD*=480ms (see Figure 3). Similarly, although participants were slower to shadow academic, *M*=1094ms, *SD*=514ms; than informal speech, *M*=1048ms, *SD*=.483; this difference was larger in the Similar Speech Condition (Academic Speech: *M*=1069ms, *SD*=509ms; Informal Speech: *M*=1001ms, *SD*=452ms) than in the Dissimilar Speech Condition (Academic Speech: *M*=1118ms, *SD*=518ms; Informal Speech: *M*=1098ms, *SD*=509ms). Indeed, model comparisons revealed a marginally significant interaction between speech condition and passage type, *b*=.05, 95% CI[-.005, .10]; $X^2(1)=3.27$, *p*=.07. Planned follow-up comparison revealed that this interaction was primarily driven by participants in the Similar Speech Condition being faster to shadow Informal Speech than Academic Speech, *b*=-.07, 95% CI[-.14, -.01]; $X^2(1)=4.34$, *p*=.037. There were no other significant effects.

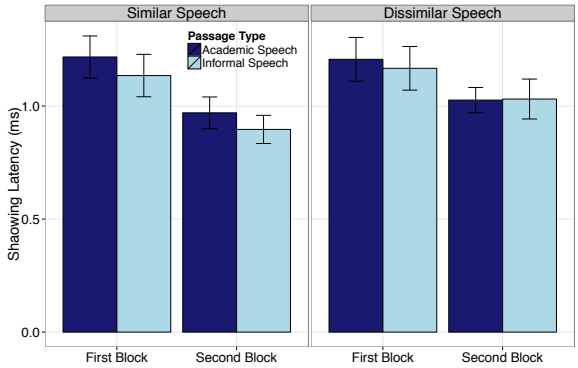


Figure 3. Latency of participants' shadowing of speech of each passage type depending on whether they did the shadowing task first or second. Error bars depict standard error of mean.

We next examined how latencies changed over the course of the task. As can be seen in Figure 3, participants had longer latencies if the shadowing task occurred in the first block, *M*=1169ms, *SD*=557ms; than in the second block, *M*=971ms, *SD*=409ms. This difference was larger for those in the Similar Speech Condition (First Block: *M*=1161ms, *SD*=558ms; Second Block: *M*=930ms, *SD*=378ms) than those in the Dissimilar Speech Condition (First Block: *M*=1176ms, *SD*=556ms; Second Block: *M*=1023ms, *SD*=440ms). However, although there was a main effect of block, *b*=-.24, 95% CI[-.46, -.02]; $X^2(1)=6.43$, *p*=.011, there was no interaction between speech condition and block, $X^2(1)=.30$, *p*=.584. Nevertheless, planned follow-up comparisons revealed that the effect of block was driven by the participants in the Similar Speech Condition, as they showed a significant effect of block, *b*=-.24, 95% CI[-.46, -.03]; $X^2(1)=4.86$, *p*=.028, but those in the Dissimilar Speech Condition did not $X^2(1)=2.19$, *p*=.14. Thus, participants shadowing speech similar to their own showed better accommodation to the speaker's voice than those shadowing dissimilar speech, as evidenced by shorter latencies as the experiment progressed.

Listening Comprehension task

Listening comprehension was higher in the Similar Speech Condition, *M*=.92, than Dissimilar Speech Condition, *M*=.88 (Fig. 4). Comprehension was higher for informal passages, *M*=.92, than academic passages, *M*=.88. As seen in the figure, participants in both speech conditions were more accurate on the Informal passages compared to the Academic ones: Similar Speech Condition (Academic Speech: *M*=.90; Informal Speech: *M*=.95); Dissimilar Speech Condition (Academic Speech: *M*=.87; Informal Speech: *M*=.89). Model comparisons revealed both marginal effects of speech condition, *b*=-.49, 95% CI[-1.04, .06]; $X^2(1)=3.02$, *p*=.082, and of passage type, *b*=-.24, 95% CI[-.05, .96]; $X^2(1)=3.01$, *p*=.083, but no significant interaction between speech condition and passage type, $X^2(1)=.84$, *p*=.36, on participants' listening comprehension.

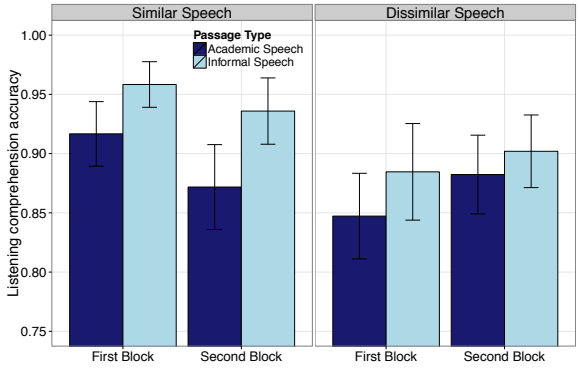


Figure 4. Accuracy in the listening comprehension task for each passage type depending on whether they completed the listening comprehension task first or second. Error bars depict standard error of mean.

We also examined whether listening comprehension changed over the course of the experiment. As can be seen in Figure 4, participants were similar in accuracy regardless of whether they completed the listening comprehension task first or second. Model comparison revealed that there was not a significant interaction between speech condition and

block, $X^2(1)=1.62$, $p=.20$, nor was there a main effect of block on listening comprehension, $X^2(1)=.03$, $p=.86$. Thus, participants' listening comprehension abilities did not improve if they had already heard the speaker in the shadowing task first.

Between task comparisons

Finally, we examined the relationships between our three measures—shadowing accuracy, shadowing latency and listening comprehension. First, as can be seen in Figure 5, participants who shadowed more slowly also tended to make more shadowing errors. Comparison of linear regression models revealed that latency was a significant predictor of shadowing errors, $b=.10$, 95% CI[.07, .14]; $X^2(1)=39.76$, $p<.001$. However, there was no interaction between latency and speech condition in predicting the number of speech errors participants made, $X^2(1)=2.04$, $p=.16$. Importantly, as can be seen in Figure 5, this relationship was not significantly different for participants in the two speech conditions, demonstrating that there was not a speed/accuracy tradeoff in shadowing. Rather, speech latency and errors are highly similar measures of shadowing ability.

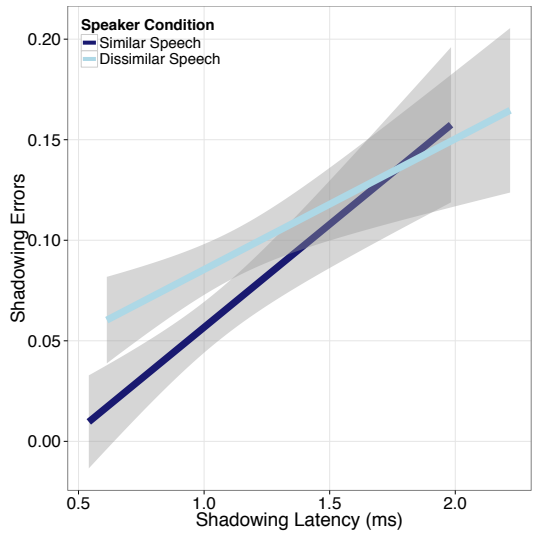


Figure 5. Relationship between shadowing errors and shadowing latency for participants in each speech condition. Error bands depict standard error of mean.

Similarly, as can be seen in Figure 6, participants with better listening comprehension tended to make fewer speech errors in shadowing. Comparison of linear regression models revealed that listening comprehension was a marginally significant predictor of speech errors, $b=-.08$, 95% CI[-.16, .003]; $X^2(1)=3.57$, $p=.06$, but listening comprehension and speech condition did not interact, $X^2(1)=.40$, $p=.53$.

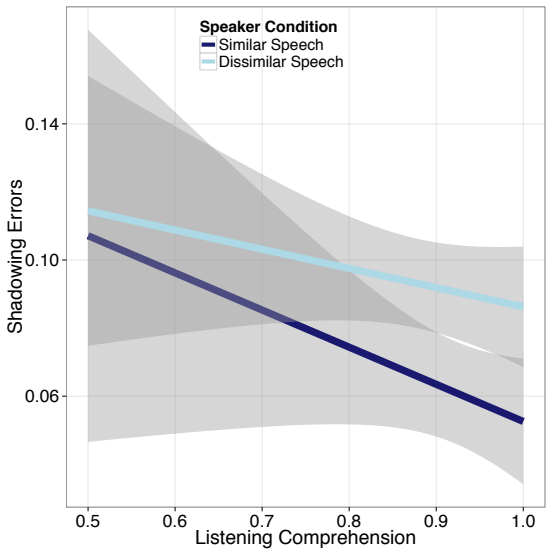


Figure 6. Relationship between shadowing errors and listening comprehension for participants in each speech condition. Error bands depict standard error of mean.

Importantly, there was a main effect of speech condition on shadowing errors, even for when both shadowing latency and listening comprehension were included as covariates, $b=.02$, 95% CI[.007, .04]; $X^2(1)=7.39$, $p=.008$, suggesting that differences in the shadowing task are not driven by individual differences in language ability. Instead, a critical predictor of participants' ability to closely shadow another person's speech, is whether or not that speech is similar to their own speech.

Discussion

In this study we asked whether the familiarity of a talker's speaking style affected listeners' ability to comprehend and closely shadow it. We found that participants were more likely to make more errors and lag further behind during shadowing and to make more comprehension errors for Dissimilar Speech than Similar Speech. The impact was larger for more informal speech than academic speech as informal speech allows for greater variation in speaking style than the more constrained academic prose. Participants make more errors and had longer shadowing latencies for Dissimilar Speech than Similar Speech, and also demonstrated less adaptation over the course of the task. This low level of adaptation is somewhat surprising given that listeners do accommodate slightly to speakers with unfamiliar accents in other studies (e.g., Maye et al., 2008). These effects are consistent with findings regarding another form of accommodation, syntactic alignment, namely the degree to which a listener subsequently uses the same sentence structures as a speaker. Weatherholtz, Campbell-Kibler & Jaeger (2014) found that participants' degree of syntactic alignment to recorded speech varied with the perceived "standardness" of the speaker's accent and perceived similarity to the participant's own speech. In our own task, speech in the Academic passage is arguably more

standard than the Informal passage regardless of speech similarity. This increased standardness could explain why participants' shadowing of Academic speech was relatively unaffected by speech similarity.

Our study is the first to our knowledge to compare listeners' ability to closely shadow and comprehend speech of speakers who both were native speakers of the listeners' language but varied in regional accent and perceived similarity. By comparing listening in these two contexts we are taking an important first step in understanding how listeners process speech that is different from their own.

In the current study, our goal was to examine differences in speech style holding the content of the material constant. The overall differences between the speakers were due to a variety of individual and group factors that were not the focus of the experiment. One area for future research will be to replicate these effects with recordings from additional speakers, in order to distinguish speaker-specific and more regional factors in speech familiarity. Another important direction for future research will be to examine the impact of differences in the speech properties we examined when other factors such as dialect also vary. The comprehension of dialects may also depend on their similarity to the speaker's own. The combination of less familiar dialect and speaking style may create greater difficulties.

Future research should also consider how social information modulates participants' processing of similar and dissimilar speech. Previous research has noted a role for social influence on speech perception (Babel, 2010; Casasanto, 2008; Kinzler, Dupoux, & Spelke, 2007; Weatherholtz et al., 2014). It will be interesting to further explore how social attitudes towards people from different racial groups and regional groups influence individuals' abilities to perceive and comprehend their speech. Given that race and region were both manipulated and correlated with each other in the current study, it is possible that participant attitudes may have been contributing to processing difficulties.

Conclusions

Together our results demonstrate that even when two speakers speak the same language, differences in speaking style can create difficulties in processing at multiple timescales.

Acknowledgments

This study was supported by NICHD R24D075454. We thank Sarah Isaak for help in data collection and coding.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529.

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*(04), 437–456.

Casasanto, L. S. (2008). Does social information influence sentence processing? *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, *BC Love, K McRae, & VM Sloutsky*, 799–804.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276–1293.

Jaeger, T. F. (2009). Random effect: Should I stay or should I go? Retrieved February 8, 2013, from http://hlplab.wordpress.com/2009/05/14/random-effect-structure/

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the Effects of Regional, Ethnic, and International Dialects of English on Listening Comprehension. *Language Learning*, *55*(1), 37–69.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523.

Marslen-Wilson, W. D. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*(1), 55–73.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical Adaptation to a Novel Accent. *Cognitive Science*, *32*(3), 543–562.

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*(1), 63–79.

Sabatini, E. (2000). Listening comprehension, shadowing and simultaneous interpretation of two "non-standard" English speeches. *Interpreting*, *5*(1), 25–48.

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–136.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4), 487–501.

Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change*, *26*(03), 387–420.
White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384.

Hidden Markov Modeling of eye movements with image information leads to better discovery of regions of interest

Stephan Brueggemann (u3003097@hku.hk)
Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong

Antoni B. Chan (abchan@cityu.edu.hk)
Department of Computer Science, Tat Chee Avenue, Kowloon, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract

Hidden Markov models (HMM) can describe the spatial and temporal characteristics of eye-tracking recordings in cognitive tasks. Here, we introduce a new HMM approach. We developed HMMs based on fixation locations and we also used image information as an input feature. We demonstrate the benefits of the newly proposed model in a face recognition study wherein an HMM was developed for every subject. Discovery of regions of interest on facial stimuli is improved as compared with earlier approaches. Moreover, clustering of the newly developed HMMs lead to very distinct groups. The newly developed approach also allows reconstructing image information at each fixation.

Keywords: Eye-tracking; Face Recognition; Hidden Markov Model; Machine Learning;

Introduction

Eye movements provide a direct insight into ongoing cognitive processes. Although mental processes cannot be observed per se, recording eye movements is an unobtrusive measurement of what a person is processing at a particular moment. Thus, eye tracking can be used to study attention, memory, language, problem solving and decision making. There exist different approaches to analyze eye-tracking recordings. Below, we summarize common eye movement analysis techniques in face recognition.

Face recognition is a cornerstone process of meaningful social interactions since it helps us to identify familiar individuals irrespective of the viewpoint, lighting conditions and emotional expression of a face. In previous studies, attempts have been made to better understand the spatial and temporal characteristics of face recognition.

Spatial eye-tracking analyses

One goal of studies on eye-tracking in face recognition is to identify which facial regions people look at when successfully recognizing another person. There exist different approaches for doing so. In a region of interest analysis, the percentage of fixations on a predefined region of interest (ROI) is computed (Henderson et al., 2005). However, there exists a lack of an objective way to identify ROIs. Statistical fixation maps aim to close this gap by constructing ROIs in a data-driven way.

A statistical fixation map can be constructed by plotting all fixations of a subject on an average face, with fixations being subsequently smoothed by convolving Gaussian kernels. Using fixation maps, it was shown that more fixations are placed on the nose area and the eye region compared with other areas (Caldara & Miellet, 2011; Hsiao & Cottrell, 2008).

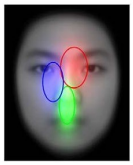
Temporal eye-tracking analyses

Face recognition is a process that can span over several seconds. Therefore it is important to understand the temporal characteristics as well. An early attempt to do so was the scan path theory (Noton & Stark, 1971). According to the theory, fixations are made on facial features in a sequence (scan path). For the same stimulus, the same sequence emerges. Studies have shown that this assumption does not always hold. For example, Walker-Smith and colleagues (1977) showed that the same scan path only emerges about 65% of the time when the same stimulus is presented.

Fixation locations during face recognition do not follow a strict, a priori planned path. More precisely, it has been indicated that saccades are constantly influenced by top-down (Yarbus, 1965) and bottom-up inputs (Mannan, Ruddock, & Wooding, 1997). It can be argued that eye movements can be treated as a Markov stochastic process. In this process, the future state depends only on the current state. Probabilistic time series models are a good fit for understanding eye-movement strategies (Chuk, Chan, & Hsiao, 2014).

Hidden Markov Models

Chuk and colleagues (2014) proposed to use a statistical time series model, the hidden Markov model (HMM) to analyze eye movement recordings in cognitive tasks. The model has several advantages over existing analysis techniques. It combines spatial and temporal analyses, and it produces data-driven ROIs, which can be of different size. Furthermore, the model accounts for individual differences by identifying subject-specific ROIs and transition probabilities.



Prior values	Red	Green	Blue
	0.47	0.35	0.17
Transition probabilities (from\to)	To Red	To Green	To Blue
From Red	0.60	0.21	0.19
From Green	0.42	0.56	0.02
From Blue	0.45	0.06	0.49

Figure 1: The Figure shows a hidden Markov Model with its prior values and transition probabilities. The red, blue and green ellipses on the face represent the hidden states. The table shows the prior values (i.e. the probability that the first fixation is at a given hidden state). The transition probabilities indicate the probability that a subject's fixation moved from one hidden states to another. The Figure is obtained from Chuk and colleagues (2014).

An HMM summarizes a subject's eye-fixation strategies (Figure 1). First, the regions of the face where the most fixations were present are identified (hidden states). In Figure 1, one hidden state is between the eyes, one hidden state is below the right eye and one hidden state covers the lower nose and mouth region. The locations of hidden states are estimated from the fixation locations. The HMM describes the participant's eye fixation strategy by the transition from one hidden state to another hidden state. The HMM specifies the prior probabilities of the initial fixation's hidden states (prior probability vector). The transition matrix describes the probability of moving from one hidden state to another hidden state. More precisely, when a subject's fixation is at a certain hidden state, transition probabilities indicate at which hidden state the next fixation will be.

It was shown that HMMs can successfully model spatial and temporal information and capture individual differences in face recognition strategies (Chuk, Chan, & Hsiao, 2014; Simola, Salojarvi, & Kojo, 2008; Wedel, Pieters, & Liechty, 2003). In Chuk and colleagues (2014), an HMM was developed for each subject to describe the individual's eye-movement patterns. Individual differences were found in both fixation locations and transition probabilities. Moreover, the eye-movement strategies of participants could be classified into one of two groups, namely holistic and analytic, demonstrating individual differences even within the same culture. In addition, it was also shown that correct and incorrect recognitions were associated with distinct HMMs. The main difference between the two groups of HMMs was found in the transition probabilities.

Hidden Markov models with fixation locations and image information

Fixation locations were the input for the HMM model by Chuk and colleagues (2014). 2-D Gaussians were fitted to the fixation locations to identify regions of interest on the face. In the present study, we propose a new way to identify hidden states. We suggest identifying hidden states based on both fixation locations and image information. The newly proposed input features advance the original model in several ways.

First, fixation locations alone are not always the optimal input features because they can be compromised. Recordings with an eye-tracker are not always accurate and over- and undershooting by the eye itself can further introduce noise to the fixation locations. Including image information as an additional input feature ensures that hidden states are identified based on fixation locations and corresponding observed visual stimuli.

Secondly, including image information makes the model more expressive. Fixation location and image information do not always closely correspond. In other words, we are able to better discriminate between fixation locations that are similar regarding their coordinates but correspond to different facial features. Thirdly, face stimuli vary slightly. In Figure 2, it can be seen that a position with a given coordinate can correspond to different facial features. Fixations of a hidden state in the newly proposed HMM will be similar in both fixation location and observed image patch.

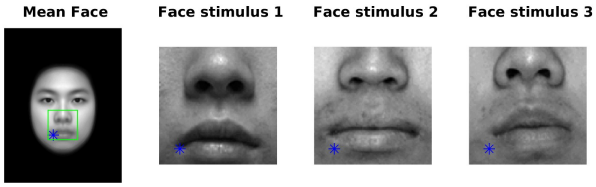


Figure 2: The Figure illustrates the variation in the structure of faces. On the left, the mean average face of all stimuli is shown. The three other images are parts of face stimuli that were used in the experiment. The blue star is located at X=168, Y=363. Although the location is exactly the same, the point belongs to slightly different areas of the face (corner of the mouth vs. area below the mouth).

In short, the newly developed HMM will advance the original model in three ways. First, it allows the identification of regions of interest on the face more accurately. Second, the current approach allows us to read out the image information for every hidden state. We are able to better understand what kind of image information a participant used during face recognition. As a final step, we cluster the newly developed, individual HMMs to investigate group differences.

Methods

Experimental Setup

In the present study, the data by Chuk and colleagues (2014) was re-analyzed. More details can be found in the original paper. In short, participants had to perform a face recognition task. 32 Chinese subjects (16 males) participated in the experiment. The experiment consisted of two blocks. In the training block, participants were instructed to study faces (targets) and in the testing phase they had to recognize the targets among new faces (distractors). In the testing phase, participants indicated whether they recognized a face or not by pressing one of two buttons on a response pad. In order to familiarize participants with the experimental task, they needed to complete a very short practice version of the experiment. The EyeLink 1000 Tower Mount eye tracker was used to record the eye movements from participants. Before the start of the experiment, the standard nine-point calibration procedure was performed. At the beginning of each trial, a drift correction was performed.

In the testing phase, face stimuli were displayed for 5 seconds within which participants were required to respond. The stimulus set consisted of 40 (20 males) gray scale frontal-view Asian faces. The faces had a neutral expression. Stimuli had an inter pupil size of 60 cm and they were all cropped to a size of 320 x 420 pixels. The screen was viewed at a distance of 60 cm. The horizontal visual angle was 6 degrees and the vertical angle was 8 degrees. Faces were aligned by vertical and horizontal eye positions. Participants were not familiar with the face stimuli.

Data Analysis

Input features The fixation location and the corresponding image information were used as input for the model. For every fixation, we extracted a 50x50 pixel image patch around the fixation location on the face stimuli. The image patch was foveated using the formula from Geisler and Perry (1998). More precisely, the spatial resolution of the visual system strongly decreases away from the fixation location. Geisler and Perry (1998) developed a foveated multiresolution pyramid which transforms every image into 5-6 regions of different spatial resolutions (see Figure 3). We performed a Principal Component Analysis (PCA) on all image patches for the purpose of dimensionality reduction. All image patches were converted to vectors. PCA was used to identify its principal components of the set of images. Each image patch was represented with its principal component coefficients, which we used as input features for the HMM.

HMMs with a different number of PCA coefficients can be learned. Firstly, we developed an HMM with a partial representation of image information. 5 coefficients account for 78.43% of variance in the stimuli. For this model, we used the X- and Y-fixation locations and the first 5 PC coefficients as input features. The 5 PC coefficients help to obtain hidden states which tend to be similar in image information.

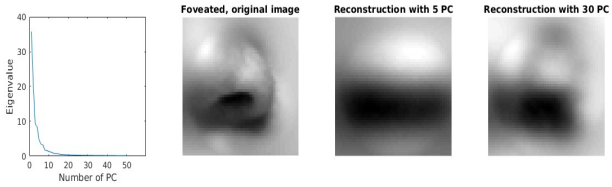
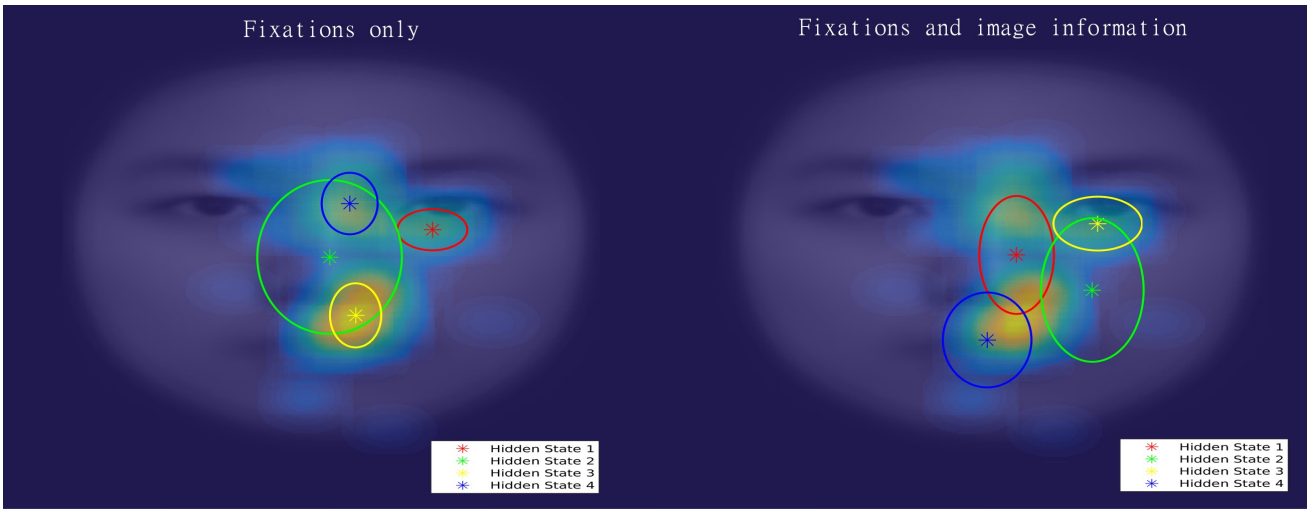


Figure 3: On the left, the eigenvalues for the first 50 eigenvectors are shown. The second image to the left shows an extracted 50 x 50 pixels image patch which was foveated using the formula by Geisler and Perry (1998). The two images on the right show the reconstruction of the original image using 5 and 30 coefficients respectively.

Secondly, we built an HMM with the purpose to obtain improved individual HMMs and to reconstruct the perceived visual stimuli per hidden state. The first 30 components are used as features. They account for 96.62% variance. We matched the number of fixation dimensions with the number of image information dimensions to ensure equal weighting. More precisely, to match the 30-dimensions of the PCA representation, we replicated the X- and Y-coordinates of the fixation 15 times. We did this to balance the influence of the fixations locations and image information on the positioning of the centers of the hidden states. The final feature vector has 60 dimensions. The model with 30 coefficients helps to obtain image patches that are highly similar in image information in a hidden state and furthermore allows reconstructing image information at the fixation.

Hidden Markov Model An HMM was estimated for every participant. Parameters of the HMM were estimated in a two-stage process. Firstly, regions of interest (ROIs) were estimated by learning a Gaussian mixture model (GMM) on the feature vectors. Each Gaussian component in the GMM corresponds to one ROI. The variational Bayesian framework for Gaussian mixture models (VBGMM) was used to estimate the number of GMM components and Gaussian parameters (Bishop, 2006). The VBGMM puts priors on the GMM components and on the GMM parameters, and it tries to find the maximum a posteriori (MAP) estimate. The first step was repeated 2000 times. The model with the highest log likelihood was chosen. Models where a hidden state had a component weight below 0.1 were rejected. In a second step, the transition and prior probabilities of the hidden states are estimated using the forward-backward algorithm (Bishop, 2006).

Clustering HMMs First, we developed HMMs for every participant to model their individual eye-movement strategy during face recognition. In a second step, we investigated if there exist different groups of eye movement patterns among participants. To cluster HMMs, we used the variational hierarchical EM algorithm (VHEM) for HMMs (Coviello, Chan, & Lanckriet, 2012, 2014). The VHEM clusters



Priors

HS 1	HS 2	HS 3	HS 4
0.35	0.52	0.03	0.1

Transition Probabilites

	To HS 1	To HS 2	To HS 3	To HS 4
From HS 1	0.34	0.39	0.24	0.04
From HS 2	0.47	0.08	0.15	0.3
From HS 3	0.87	0	0.13	0
From HS 4	0.9	0	0	0.1

Priors

HS 1	HS 2	HS 3	HS 4
0.43	0.18	0.03	0.35

Transition Probabilites

	To HS 1	To HS 2	To HS 3	To HS 4
From HS 1	0.15	0	0.2	0.65
From HS 2	0.34	0.22	0	0.44
From HS 3	0	0	0.22	0.79
From HS 4	0.22	0.26	0.17	0.35

Figure 4: The Figure shows the fixation heat maps and overlaid HMMs of the first subject. Moreover, prior values and transition probabilities are shown. On the left, an HMM was developed only using fixation locations as input. On the right, we developed an HMM with input features consisting of fixation locations and full image information (30 coefficients). The first HMM has hidden states that closely correspond to the heat map. In contrast, the second HMM has hidden states that center more on facial features.

HMMs into groups and presents common ROIs and transitions for each group. Moreover, heat maps were developed using the fixations of all subjects clustered together. We applied the VHEM algorithm to HMMs with partial image information.

Results

HMM with full image representation

Regions of Interest Figure 4 shows a comparison of the different HMMs that were obtained using the fixation-only input features and the input features consisting of fixation locations and image information for the first subject. The HMM which was developed only using fixation sequences has hidden states which closely correspond to the heat map. Hidden state 2 covers a great part of the central face and therefore makes it difficult to understand where exactly the fixation was present. The hidden states of the HMM with the newly proposed input features correspond to facial features rather than to the heat map. The first hidden state is clearly located on the nose and the second hidden states is on the cheek next to the nose. Moreover, the third hidden state covers the right eye and the fourth hidden state is on the mouth. The HMM with the newly proposed input features seems to better capture the regions of interest on the face stimuli of the first subject.

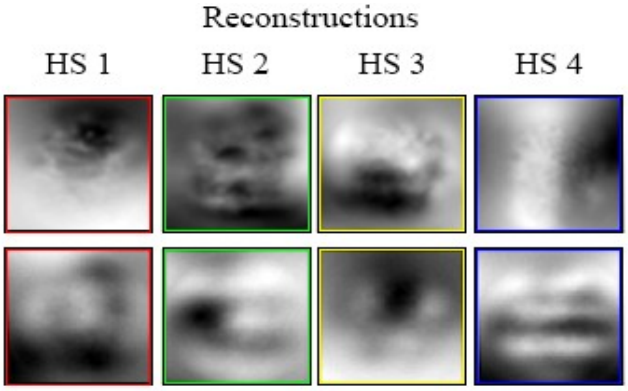


Figure 5: Images show the reconstructions of the image patches per hidden state. The top panel shows the reconstructions of the fixation-only HMM. The bottom panel shows the reconstructions of the image patches of the newly proposed HMM.

Reconstructions of image information Figure 5 shows the reconstructions of the image information for each hidden state of the HMM with the newly proposed input features (bottom panel). It is possible to identify important face regions. The first reconstruction shows the center of the nose and the second reconstruction shows the right side of the

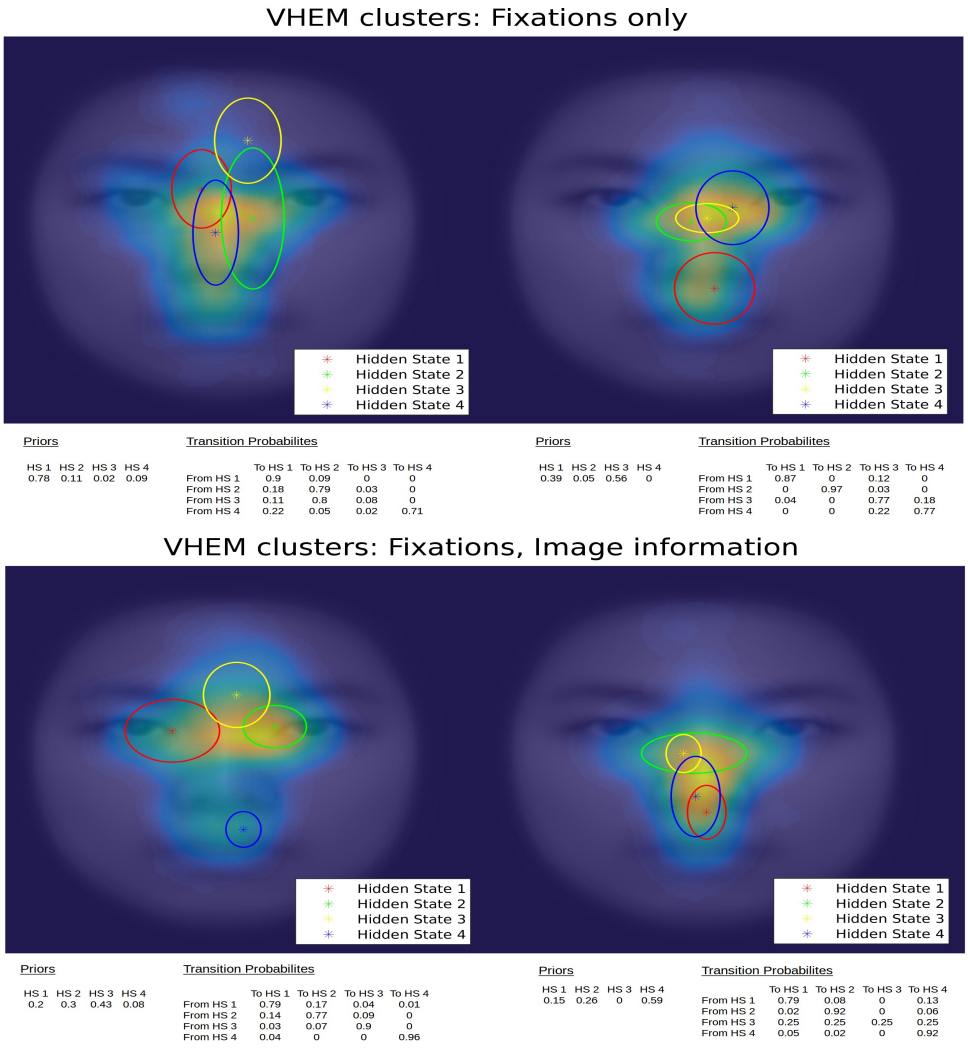


Figure 6: The Figure shows the VHEM clusters for HMMs which were developed based on fixation locations (top panel) and HMMs developed using fixation locations and image information (bottom panel). Heat maps summarize fixation locations of all participants who are part of a given cluster. VHEM clusters and the difference in heat maps of the latter are much more distinct. Prior and transition probabilities are rounded to two decimal places.

nose and parts of the right cheek. Reconstructions from the third and fourth hidden state show the right eye and the mouth respectively. The newly proposed model allows reconstructing the perceived visual stimuli at fixation, although reconstructions are not always very clear. We compared the reconstructions from the newly proposed HMM with reconstructions by the fixation-only HMM. In the top panel, averaged image patches from the hidden states of the fixation-only HMM are shown. They look more washed out and noisy than the reconstructions from the new model.

HMM with partial image representation

Group Differences In Figure 6, results of the HMM clustering are illustrated. After having developed HMMs for every participant, the VHEM algorithm was used to cluster the HMMs into groups to investigate if different fixation strategy groups exist. The top panel shows clustering results of the HMMs which were solely developed using fixation locations. Two different clusters exist. The HMM which is representative of the first cluster (top panel, left HMM) has hidden states which cover a large part of the central, upper area of the face. The hidden states of the second cluster (top panel, right HMM) are located more in the central area of the face.

The VHEM clusters obtained from the HMMs with partial image information are much more distinct. The first group (bottom panel, left) has three hidden states centered on the eye area and one hidden state on the mouth area. The HMM representative of the other VHEM cluster has all its hidden states on the nose area.

Visually, there exists a clearer difference in heat maps for VHEM clusters of HMMs with partial image information (bottom panel). The heat map of the first cluster indicates that many fixations fell on the eye area. The heat map of the second clusters shows that most fixations were on the nose area. The difference in heat maps of the VHEM clusters of HMMs without image information is less pronounced (top panel). In short, it is clear the clustering of HMMs with partial image information resulted in much more distinct clusters than the clustering of HMMs without image information

Discussion

In the present study, a new HMM approach is introduced. We showed that using image information in addition to fixation locations as input features has several benefits. The newly developed HMMs have better ROIs which are based on fixations that are similar in fixation location and image information. Moreover, the newly developed HMMs with full image representations allow to reconstruct the image information that was associated with each ROI. Lastly, clustering the newly developed HMMs resulted in very distinct groups confirming the findings by Chuk and colleagues (2014). The newly introduced HMM approach can be used for different cognitive tasks to investigate spatial and temporal characteristics of eye-movement strategies.

Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project 17402814 to J. H. Hsiao and CityU 110513 to A. B. Chan). We thank Tina Liu for collecting the data for the study and we thank T. Chuk for helpful discussion.

References

Bishop, C. M. (Ed.). (2006). Pattern recognition and machine. New York: Springer.

Caldara, R., & Miellet, S. (2011). imap: A novel method for statistical fixation mapping of eye movement data. Behavior Research Methods, 43, 864 - 878.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden markov models. Journal of Vision, 14, 1 - 14.

Coviello, E., Chan, A. B., & Lanckriet, G. R. G. (2014). Clustering hidden markov models with variational hem. Journal of Machine Learning Research, 15, 697 - 747.

Coviello, E., Lanckriet, G. R., & Chan, A. B. (2012). The variational hierarchical em algorithm for clustering hidden markov models. In P. Barlett (Ed.), Advances in neural information processing systems. New York: Curran Associates, Inc.

Geisler, W. S., & Perry, J. S. (1998). A real-time foveated multi-resolution system for low-bandwidth video communication. Human Vision and Electronic Imaging, 3299, 294 - 305.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. Memory and Cognition, 33, 98 - 106.

Hsiao, J. H., & Cottrell, G. (2008). Two fixations suffice in face recognition. Psychological Science, 19, 998 - 1006.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S.. (1997). Fixation patterns made during brief examination of two dimensional images. Perception, 26, 1059 - 1072.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. Vision Research, 11, 929 – 942.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9, 97
- 113.

Simola, J., Salojarvi, J., & Kojo, I. (2008). Using hidden markov model to uncover processing states from eye movements in information search tasks. Cognitive Systems Re- search, 9, 237 - 251.

Walker-Smith, G. J., & Findlay, A. G. G. J. M. (1977). Eye movement strategies involved in face perception. Percep- tion, 6, 313 326.

Wedel, M., Pieters, R., & Liechty, J. (2003). Evidence for covert attention switching from eye-movements. reply to commentaries on liechty et al., 2003. Psychometrika, http://doi.org/10.1007/BF02295611.

Yarbus, A. L. (Ed.). (1965). Eye movements and vision, translated from Russian by basil haigh. New York: Plenum Press.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. (1985). Matching familiar and unfamiliar faces on internal and external features. Perception, 14, 737 - 746

Structural Alignment in Incidental Word Learning

Ruxue Shao (ruxueshao2018@u.northwestern.edu)
Dedre Gentner (genter@northwestern.edu)

2029 Sheridan Road, Evanston, IL, 60208, USA

Abstract

Young children can sometimes acquire new vocabulary words with only limited, indirect exposure (Carey & Bartlett, 1978). We propose that structural alignment processes lead to fluent detection of commonalities and differences that facilitate incidental word learning. To test this, we adapted the Carey and Bartlett paradigm, varying the alignability of the objects that 4-year-olds saw while hearing the novel word *chromium*. In Experiment 1, children in the High-Alignment condition were significantly better than those in the Low-Alignment condition at identifying chromium objects in a subsequent task. In Experiment 2, we ruled out an alternative account by equalizing the overall amount of information presented to the two groups. We also found that the advantage of high alignment persisted after two-to-four days. These results suggest that structural alignment is a mechanism by which children can learn word meanings even in incidental word learning situations.

Keywords: incidental word learning; structural alignment; alignable differences

Introduction

Children acquire vocabulary words in many ways. Sometimes, a helpful adult points to an object of interest and gives a clear label for it. For example, the teacher might point to a rubber duck and tell the child "Look, a duck!" Yet these instances of explicit word learning are far from the whole picture. Children also learn many words spontaneously through daily conversation, media consumption, joint reading, and so on (e.g. Akhtar, Jipson, & Callanan, 2001; Krcmar, Grela, & Lin, 2007; Senechal & Cornell, 1993).

Carey and Bartlett (1978) conducted a classic study on young children's incidental word learning. They exposed 3.5-year-olds to a new color term (*chromium*) for an unfamiliar color (dark olive green). Children saw two trays that were identical on every dimension except color (see Figure 1), and were asked to "Give me the chromium one, not the red one." Even with such limited input, approximately half of the children were able to demonstrate some knowledge of the word *chromium* at a later time.

Carey and Bartlett's findings have inspired many researchers to further investigate this question from multiple perspectives (e.g. Bion, Borovsky, & Fernald, 2013; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Heibeck & Markman, 1987; Wagner, Dobkins, & Barner, 2013). Our interest here is in what processes enabled children to learn a new word from such minimal, indirect exposure. We propose structural mapping as a mechanism that can lead to rapid spontaneous learning.

According to Structure Mapping Theory, comparison involves a process of *structural alignment* between two representations (Gentner, 1983, 2010). This process not only reveals commonalities, but also highlights *alignable differences*—differences that play the same role in the shared structure. Alignable differences emerge easily and quickly when comparing items that are high on overall similarity (Sagi, Gentner, & Lovett, 2012). Both children and adults find it easier to generate differences for high-similarity object pairs than for low-similarity ones (e.g. Gelman, Raman, Gentner, 2009; Gentner & Gunn, 2001; Markman & Gentner, 1994). Thus, for high-similarity pairs, alignable differences emerge quickly and spontaneously.

From this perspective, the Carey and Bartlett (1978) findings can be interpreted as follows. Children were asked to 'give me the chromium one, not the red one," and their goal (we assume) was simply to comply with this request (which they may have interpreted simply as "give me the one that's not red"). However, when they saw two highly similar trays, a spontaneous alignment process took place, and the alignable difference between the two trays (their color) popped out. Thus, even without overt effort to learn a new word, the children may have associated this newly highlighted color with the new word, *chromium*.



Figure 1: Trays shown in the Carey and Bartlett (1978) study

To test this claim, we adopted a paradigm similar to the one used by Carey and Bartlett (1978), but varied the similarity (and thus the alignability) of the objects pairs presented. The question is whether children will learn the meaning of the term *chromium* from indirect exposure, and whether alignability will be a determining factor.

Experiment 1

We presented children with a pair of objects that were either high in alignment (HA) or low in alignment (LA). Children were asked to "point to the chromium one—the chromium one, not the blue one." After a short delay, we assessed whether they could identify chromium objects in a yes-no task. We predicted that children in the HA condition would show better understanding of the word *chromium* than those in the LA condition.

Methods

Participants Sixty-three 4-year-olds (mean age=54.05 months, SD= 3.28 months, range= 48-59 months) participated in the study. Among them, thirty-two 4-year-olds were assigned to the High-Alignment (HA) condition and the remaining to the Low-Alignment (LA) condition. An additional five children were tested but excluded, one due to showing a yes bias, one for failing to complete the experiment, and three due to experimenter error. All children were recruited from the greater Chicago area through a voluntary participant pool or from local preschools.

Initial Exposure As a warm-up, children first received two simple two-alternative-forced-choice trials (e.g. "point to the cow, not the lion"). Then they received the key Initial Exposure trial. Children in the HA group saw two highly similar (readily alignable) objects (as in the Carey and Bartlett study), while those in the LA group saw a less alignable pair (Figure 2). All figures were made from foam and were approximately the same size. Both groups were asked, "Look at these two! Can you point to the chromium one? The chromium one, not the blue one." Importantly, all children (in both the HA and LA groups) were able to correctly pick out the chromium object during the initial exposure.

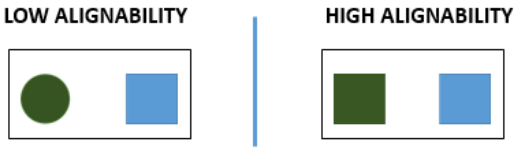


Figure 2: Experiment 1 Initial Exposure objects.

Meaning Assessment Approximately 10 minutes after the initial exposure, children were assessed on whether they had learned the meaning of *chromium*. They were shown eight geometric figures, each with a unique combination of color and shape. Among the figures, three were chromium and five were of other colors (See Figure 3). For each group, there was one previously-seen chromium figure and two new ones. (Children in the HA group had seen the chromium square, while those in the LA group had seen the chromium circle.) In addition, all children had seen the blue square in the Initial Exposure trial and the maroon triangle during the warm-up trials.

Children were shown the objects one at a time, and asked, "Look at this one! Is this a chromium one?" Based on the child's answer, the object was placed in either a Yes Box or a No Box. If the child failed to answer, or indicated that (s)he did not know, the experimenter repeated the question. No participant required more than three requests to answer a question.



Figure 3: Experiment 1Meaning Assessment objects

Results

In the Initial Exposure phase, all children were able to correctly pick out the chromium object on the first attempt. However, as predicted, significant differences were found between the HA and LA groups in the Meaning Assessment task. The HA group was able to correctly identify more chromium objects (more hits) than the LA group (M_{LA} = 1.77, M_{HA}= 2.50, max =3.0; $t(61) = 8.81$, $p = .004$, one-tailed t test). There was no difference in the number of false alarms (identifying non-chromium objects as *chromium*) made by the two groups (M_{LA} = 1.23, M_{HA} = 1.77, max=5.0; $t(61)$ = 2.78, *n.s.* one-tailed t test). Thus the difference in hit rate was not due to an overall inclination of the HA group to respond more positively.

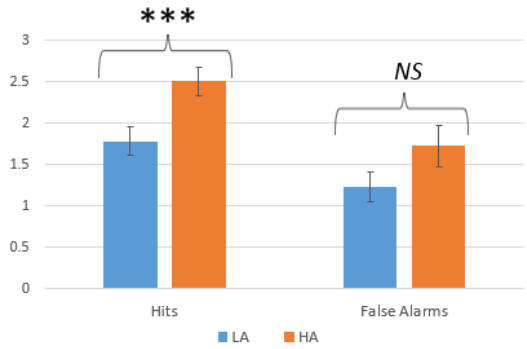


Figure 4: Mean number of hits (max = 3) and false alarms (max = 5) made by the LA and HA groups in Experiment 1.

We also found a difference between the two groups when comparing their performance to chance. Children in the HA condition scored significantly better than chance (1.5) at identifying chromium objects ($t(31)=7.42$, $p<.001$). However, performance in the LA group did not differ from chance ($t(30)=1.33$, *n.s.* one-tailed t test).

Our hypothesis is that for the HA group, a spontaneous, non-deliberate alignment process resulted in children's noticing the alignable difference in color. However, it is also possible that the difference between the two groups lay in differential forgetting. Perhaps children in the LA group had more difficulty remembering the initial chromium object. To test this possibility, we statistically controlled whether children were able to successfully identify the Initial Exposure object as *chromium* during the Meaning Assessment task. When we compared performance on the other objects in the Meaning Assessment Task, we found the same pattern as before, indicating that the performance gap in the Meaning Assessment task cannot be attributed to differential item memory for the initial item.

Discussion

Children in the HA condition were significantly more likely than those in the LA condition to discern the meaning of the word *chromium*. Although both groups were able to "point to the chromium one, not the blue one" when shown the two initial objects side by side, only the HA group showed

evidence of being able to differentiate *chromium* objects from non-*chromium* ones in a subsequent task. These findings are consistent with our predictions based on the structural alignment account.

We believe that the HA group was able to "pick up" the word *chromium* without a deliberate intention to learn a new word. For children in the HA condition, the chromium color was highlighted during the Initial Exposure when it emerged as an alignable difference from the familiar color, blue. This provided participants with a perceptual sensation that could attach to the new word. In other words, we suggest that high-alignment comparison essentially made the mapping of *chromium* to the color 'dark olive green' effortless.

However, before embracing this account, we need to assess an alternative possibility. Children may have been actively trying to learning the meaning of the novel word. In this case, the overall amount of information available to them becomes very important. For the HA group, the two objects shown during the Initial Exposure differed in only one dimension – color. In contrast, the LA group saw objects that differed in both color *and* shape. From a hypothesis-testing point of view, the LA children had to consider two possible hypotheses for the meaning of *chromium*—a color hypothesis (dark olive green) and a shape hypothesis (round). In contrast, the HA children only had to consider one hypothesis: that *chromium* refers to dark olive green. Thus, the advantage shown by the HA group over the LA group could have resulted from the differential amount of information available to the two groups, rather than from the greater ease of alignment.

Experiment 2 was designed to distinguish between these two accounts. We again created two groups—HA and LA—but gave children in both groups two trials, to equate the amount of information they received. In addition, we wanted to assess the strength of the word-meaning link formed here. In Experiment 1, children were tested on the meaning of *chromium* 10 minutes after the Initial Exposure, with a collection of geometric shapes similar to those used in the Initial Exposure. In Experiment 2, we asked whether we would still find evidence of learning if the delay was longer and the set was larger. In the same spirit, we also asked whether children would be able to transfer their learning to a very different set of items.

Experiment 2

The goals of Experiment 2 were (1) to assess the effects of alignability in incidental learning while equating the overall amount of information given; and (2) to explore the robustness of the learning effect. To equate the information given to the two groups, we provided both the HA and LA groups with two exposure trials during the Initial Exposure., such that both groups could logically rule out shape as the meaning of *chromium* and settle on color. We then assessed children's knowledge of the word *chromium* on the day of the exposure (Day 1), and also two-to-four days later (Day 2). Day 2 featured both a retention task and a transfer task.

Participants

Forty 4-year-olds (*M*=53.63 months, SD= 3.56 months, range= 48-59 months) participated in Experiment 2. Nineteen of them were placed in the High-Alignment (HA) condition and twenty-one in the Low-Alignment (LA) condition. Children were recruited through the same methods used in Experiment 1.

Day 1

Materials and Procedure The procedure for Day 1 was highly similar to that of Experiment 1. Children received a warmup trial, and then encountered the word *chromium* and its referent color. However, instead of just one exposure, they were given two chances to form the mapping between word and color, as described below. They then took a short break, and afterwards revisited the word and color in a sorting task.

On each of the Initial Exposure trials, children saw a blue object and a chromium object side by side. The experimenter asked, "Can you point to the chromium one? The chromium one, not the blue one." The Initial Exposure trials were not presented consecutively—a foil trial (similar to the warmup trial) separated the two.

The trials were designed to permit the LA group to rule out the shape hypothesis, leaving them with only the color hypothesis— and thus removing the informational advantage of the HA group. On Initial Exposure Trial 1, the LA group saw two objects differing in color and shape, leaving two possible hypotheses for the meaning of *chromium*:

(1) *chromium* = dark olive green color
(2) *chromium* = cross shape

One Trial 2, the LA group saw two new objects. Crucially, the 'chromium' object (the non-blue object) was now a different shape (an upside-down triangle) but retained the same color. Thus, children in the LA condition could rule out the shape hypothesis and arrive at the correct meaning.

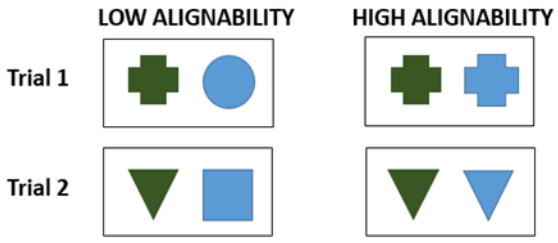


Figure 5: Experiment 2, Day 1: Initial Exposure Objects

Children in the HA condition also received two trials, designed so that the chromium objects were the same as those seen by the LA group. However, to facilitate structural alignment, these objects were paired with another object of the same shape within each trial. The HA group saw a chromium square and a blue square on one trial, and a chromium circle and a blue circle on the other. As in Experiment 1, the objects were identical on every dimension

except for color. Figure 4 shows the Initial Exposure for Experiment 2, Day 1.

A new and extended set of 14 geometric shapes was used in this experiment for the Meaning Assessment task (Figure 6). There were five chromium objects—including two that had been previously seen by both groups—and nine non-chromium objects. The procedures for the Meaning Assessment Task in Experiment 2 Day 1were exactly the same as those for Experiment 1.



Figure 6: Experiment 2: Figures Used in Meaning Assessment (Day 1) and Retention Task (Day 2)

Results As in Experiment 1, all children in the HA and LA groups were able to correctly pick out the chromium object during the Initial Exposure.

In the Meaning Assessment task, children in the HA group were able to correctly identify more chromium objects (more hits) than those in the LA group (M_{LA}= 3.71, M_{HA}= 4.63, max = 5.0; t (38)= 2.18, p=.020, one-tailed t test). Also similar to Experiment 1, there was no difference in the number of false alarms made (M_{LA}= 1.76, M_{HA}= 1.32, max = 9.0; t(38)=0.70, *n.s.*, one-tailed t test).

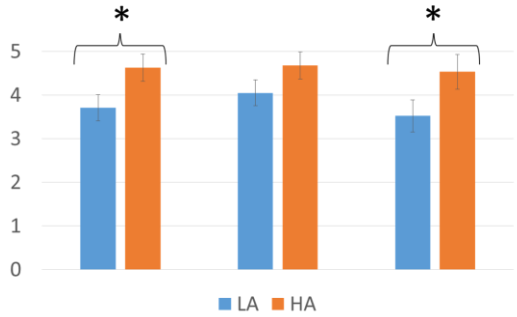


Figure 7: Experiment 2: Mean number of hits (max=9) on Day 1 Meaning Assessment task, Day 2 Retention Task, and Day 2 Transfer Task.

Day 2: Retention and Transfer

Materials and Procedure After two to four days (M= 2.45, SD= 0.75), children returned to the lab for Day 2 of the experiment. Children in the HA and LA groups went through the same procedure: a Retention Task, a Reminder, and a Transfer Task.

The goal of the Retention Task was to assess how well children retained whatever meaning they had assigned to *chromium* on Day 1. Thus, the Retention Task was an exact duplicate of the Meaning Assessment task on Day 1. After the Retention Task, children were presented with a Reminder similar to the Initial Exposure on Day 1. All children saw the same two cards side by side, one depicting a blue fish and the other a chromium fish. The experimenter asked, "Look at these two! Can you point to the chromium one? The chromium one, not the blue one!" All children were successful on this task. After the Reminder, children engaged in the Transfer Task. The Transfer Task used the same format as Day 1's Meaning Assessment task, except that all of the objects being sorted were picture cards of different colored and shaped fish. The fish cards consisted of 14 line drawings of fish. Similar to the set shown in Figure 6, five of the fish were chromium, four were blue, and the rest were of various colors.

Results Although the Retention task's pattern of results suggested that the HA group made more hits than the LA group, the difference between the two groups did not reach significance (see Figure 7). There was no difference between the two groups in the number of false alarms.

When provided with a Reminder exposure to the word *chromium* and the color dark olive green, all children were able to pick out the correct chromium object on the first attempt, consistent with the pattern in the Initial Exposure task in both Experiment 1 and Experiment 2 Day 1.

In the Transfer Task, the HA group again showed an advantage over the LA group: they identified a significantly higher number of dark olive green fish as *chromium* than the LA group (M_{LA}= 3.52, M_{HA}= 4.53, max= 5.0; t(38)=1.84, p=0.037, one-tailed t test). The number of false alarms made by the two groups did not differ, once again mirroring previous results.

Discussion

The findings of Experiment 2 replicated and extended those of Experiment 1. On Day 1, children in both the HA group and LA group received sufficient information to rule out shape and settle on color as the meaning of chromium. Yet we found the same performance pattern as in Experiment 1: although all children were able to correctly pick out the chromium object in the Initial Exposure, the HA group did significantly better than the LA group on a subsequent Meaning Assessment task. These results argue against a difference in information as the explanation for the HA group's advantage. Instead, the greater fluency of structural alignment seems to be the key distinguishing factor here.

On Day 2, children were assessed on the meaning of the word *chromium* using the same objects that they had seen on Day 1 (the Retention Task). Although we did not find significant HA advantage in the Day 2 Retention task, the HA group went on to significantly outperform the LA group in the Transfer task: Children in the HA group demonstrated better understanding of the word *chromium* than those in LA group on the new items shown in the Transfer Task.

General Discussion

The current studies provide evidence supporting the hypothesis that spontaneous structural alignment promotes incidental word learning. In Experiment 1, we manipulated the alignability of the objects that children saw during the Initial Exposure. Although all children in both groups readily

picked out 'the chromium one, not the blue one', the two groups differed dramatically in how much they learned from this experience. Children in the HA condition were able to differentiate between chromium objects and non-chromium objects in a subsequent task, while those in the LA condition performed at chance. In Experiment 2, we equalized the amount of information presented to the two groups during the Initial Exposure and still found an advantage of the HA group over the LA group. This advantage persisted two to four days after the Initial Exposure, as evidenced in the Transfer Task—the HA group (but not the LA group) was able to transfer their understanding to a very different set of objects.

We suggest that the high overall similarity between the initially shown objects benefitted the HA group in two ways. First, it invited spontaneous comparison, initiating a structural alignment process. Second, the high alignability of the two objects made structural alignment fast and essentially effortless, so the alignable difference of color "popped out" and was easily detected. Structural alignment allowed children to quickly and easily "pick up" a perceptual placeholder for the new word *chromium*, even without any conscious intentions of decrypting it.

We suggest that in this task, children were simply engaging in a social interaction with the experimenter, and their goal was to comply with the experimenter's request. Nevertheless, given the fortunate learning situation presented to the HA group, they were able to learn the meaning of a new word *chromium* without any prior intention of doing so.

For the LA group, who saw objects that were dissimilar, structural alignment did not spontaneously occur. Thus, although they were able to correctly pick out the chromium object when given the verbal contrast "the chromium one, not the blue one," there was no evidence that they formed a link between the novel word *chromium* and its corresponding property.

Of course, we note that either group could have focused on the color hypothesis for linguistic reasons. The syntax of the request-- 'the chromium one, not the blue one"—would have allowed an adult to infer that the key dimension of difference was probably color. However, it's not clear that 4-year-olds understand this convention. Indeed, the poor performance of the LA group suggests that they did not benefit from such an insight. We aim to shed light on this issue in further research

These results are consistent with prior work showing that structural alignment processes support children's learning (Childers & Paik, 2009; Ferry, Hespos, & Gentner, 2015; Gentner, Anggoro & Klibanoff, 2011; Waxman & Klibanoff, 2000). Explicit instructions to compare exemplars can promote learning in both adults as well as children (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Loewenstein et al., 1999; Markman & Gentner, 1993). Evidence for such effects has been found across perceptual and conceptual domains, including spatial configurations (Christie & Gentner, 2010), mathematics (Richland, Zur, & Holyaok, 2007; Richland, Holyoak, & Stigler. 2004; Rittle-Johnson & Star, 2007; Thompson & Opfer, 2010), word learning (e.g., Graham et al., 2010), part-learning (Gentner, Loewenstein, & Hung, 2007), and science education (Jee et al., 2010; Kurtz & Gentner, 2013). However, to the best of our knowledge, this is the first study that examines the role of structural alignment in incidental learning situations.

Incidental learning experiences are clearly an important part of children's language learning, yet they are almost invisible in nature. Such experiences can occur at different times and places, with no necessary pedagogical intention on the part of the adult, nor any intention to learn on the part of the child. They depend on the learner noticing the new information for himself or herself. The current studies show that structure mapping is a powerful mechanism for incidental learning. When the target information is highlighted as an alignable difference, it is more likely that spontaneous learning will take place.

Acknowledgments

We thank the children, families, and preschools that participated in the study, Emma Bulzoni and Katherine Cha for their help in conducting the experiments, and members of the Project on Children's Thinking for their insightful suggestions. This work was supported by NSF SLC Grant SBE-0541957 awarded to the Spatial Intelligence and Learning Center (SILC), and by ONR Grant N00014-13-1-0470 to Dedre Gentner.

References

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development, 72*(2), 416-430.

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition, 126*(1), 39-53.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference, 15*, 17-29.

Childers, J. B., & Paik, J. H. (2009). Korean-and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language, 36*(1), 201-224.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*(3), 356-373.

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development*,*86*(5), 1386-1405.

Gelman, S., Raman, L., & Gentner, D. (2009). Effects of language and similarity on comparison processing. Language Learning and Development, 5, 147-171.

Gentner, D., & Boroditsky, L. (200

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science, 7*(2), 155-170.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752-775.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure-mapping and relational language support children's learning of relational categories. *Child Development*, *82*(4). 1173-1188.

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition*, *29*(4), 565-577.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development*, *8*(3), 285-307.

Gentner, D., & Kurtz, K. J. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*(4), 609-642.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152-158.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*(2), 263-297

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487-513.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1-38.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99

Graham, S., A., Namy, L. L., Gentner, D., & Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. *Journal of Experimental Child Psychology*, *107*. 280-290.

Heibeck, T. H., & Markman, E. M. (1987). Word learning in children: An examination of fast mapping. *Child Development*, 1021-1034.

Jee, B. D., Uttal, D. H., Gentner, D., Manduca, C., Shipley, T., Sageman, B., Ormand, C. J., & Tikoff, B. (2010). Analogical thinking in geoscience education. *Journal of Geoscience Education*, *58* (1), 2-13A

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, *10*(1), 41-63.

Kurtz, K. J., & Gentner, D. (2013). Detecting anomalous features in complex stimuli: the role of structured comparison. *Journal of Experimental Psychology: Applied*, *19*(3), 219.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*, 586-597.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431-467.

Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognitiion and Instruction*, *22*, 37-60.

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*, 1128-1129.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*, 561–574.

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019-1050.

Senechal, M., & Cornell, E. H. (1993). Vocabulary acquisition through shared reading experiences. *Reading Research Quarterly*, 360-374.

Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, *81*(6), 1768-1786.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: color word learning as a gradual inductive process. *Cognition*, *127*(3), 307-317.

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*(5), 571.

Reflexive Spatial Attention to Goal-Directed Reaching

Alexis A. Barton (aabarton@indiana.edu)
Bennett I. Bertenthal (bbertent@indiana.edu)
Samuel Harding (hardinsm@indiana.edu)

Department of Psychological and Brain Sciences, Indiana University
1101 E.10th Street, Bloomington, IN 47405 USA

Abstract

Social interaction involves cues such as gaze direction, head orientation, and pointing gestures that serve to automatically orient attention to a specific referent or spatial location. In this paper we demonstrate that an observed reaching action similarly results in a reflexive shift in attention as evidenced by faster responses that are congruent with the direction of the reach, than responses that are incongruent. This facilitation is evident quickly after the onset of the reach action and is due to the rapid prediction of the reach-goal. When the task involves a saccadic response (Experiment 1) this prediction is inhibited and results in a reverse-congruence, faster responses to incongruent than congruent cues, when the cue occurs after the reach is completed. This reverse-congruence is not present when the task involves a key press (Experiment 2) or a mouse movement (Experiment 3). We propose that the inhibition of the predictive saccade is overcome when the eye movements toward the goal are activated to guide the mouse movement. The three experiments together demonstrate that automatic attention distribution and its effects on behavior depend on the response.

Keywords: action perception; reflexive attention; action prediction

Introduction

Social cues that guide another person's attention to a specific location come in various forms such as eye-gaze, head orientation, and pointing gestures (Langton, Watt, & Bruce, 2000). The observer often processes these cues automatically as their attention is quickly deployed in such a way that it is hard to inhibit. For example, when presented with a face that displays averted gaze, participants are faster to detect a target when the gaze is in the same direction as the target's location than when it appears in the opposite location (Driver et al., 1999; Friesen & Kingstone, 1998). Similar compatibility effects are seen following the presentation of a pointing hand (Bertenthal, Boyer, & Harding, 2014; Crostella, Carducci, & Aglioti, 2009). Furthermore, this shift in spatial attention may influence responses differently based on the type of response required. Crostella, et al. (2009) report that observation of a distracting gaze cue interferes with saccadic responses, whereas, a distracting point gesture interferes with a pointing response.

Paradoxically, a grasping hand does not automatically capture attention in the location of the grasped object (Fischer & Szymkowiak, 2004) even though it, too, is a social cue. Furthermore, grasping actions will modulate perspective taking, but not spatial orienting (Mazzarella, Hamilton, Trojano, Mastromauro, & Conson, 2012). Fischer & Syzmkowiak (2004) suggest that a point is a deictic gesture that continues to represent the intentions of the actor, whereas, a grasp is an indication that an action has already been performed and observers are not reflexively drawn to attend to an object that has already been acted upon.

While this claim may apply to the observation of a static grasping hand, there is evidence to suggest that it is less likely to apply to a dynamic reaching and grasping action. Indeed, a static presentation of a pre-shaped grasping hand, unlike the completed grasp, results in automatic spatial attention to a congruently shaped object (Fischer, Prinz, & Lotz, 2008), and even infants as young as 5 months will orient faster to a target that is in the same direction as a static pre-shaped grasping hand (Daum & Gredebäck, 2011).

In fact, recent studies measuring eye movements during the observation of reaches find that observers will predictively shift their gaze to the goal of the reach prior to the completion of the action (Flanagan & Johansson, 2003). These predictive saccades are very similar to the proactive eye movements made by an actor during reaching. It has been hypothesized that this coordination between the observer's gaze and the actor's reach could be attributable to observers mapping perceived actions to a corresponding system of motor representations including the prospective looks to the goal (Flanagan & Johansson, 2003). Alternative explanations for the prediction of observed action goals involve cognitive, rather than motor, mechanisms such as statistical learning or the attraction to goal salience (Eshuis, Coventry & Vulchanova, 2009).

If the prospective eye movements to the reach goal are automatically activated during observation then we expect that covert spatial attention will likewise be automatically shifted in the direction of the reaching action. Furthermore, if spatial attention is captured by a reaching action then observation of a dynamic reach will facilitate fast and accurate responses in the same direction during a spatial cuing task, even if the reach is irrelevant to the task.

The current study aims to investigate the relation between perceiving dynamic, goal-directed actions and spatial attention. We followed the approach of Crostella et al. (2009), and presented a video of a goal-directed reach in conjunction with a spatial cueing task. The reaching action

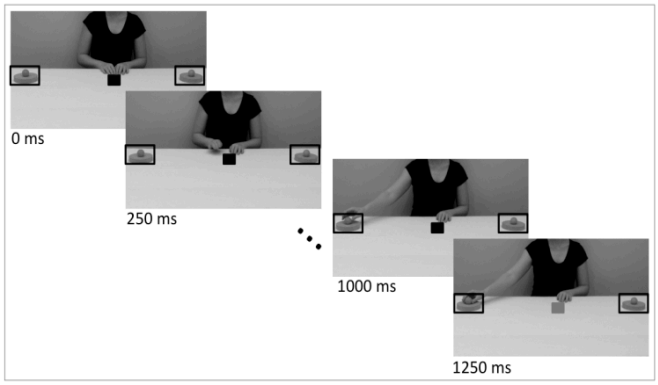


Figure 1: Still images showing the stimulus video of reaching actions with the cue and response targets overlaid. In this trial, the cue to respond occurred after the reach-to-grasp action was completed at 1250 ms.

is non-predictive of target locations and participants were instructed to ignore it. Thus, if we see that responses that are congruent with the direction of the reach are facilitated in comparison to responses that are incongruent, we can assume that observation of the reach automatically shifted the participant's spatial attention toward the reach-goal. Furthermore, we manipulated the timing of the reaching action with regard to the stimulus cue. In this way, we could assess the automatic processing of the reach at multiple time points and evaluate the effects of both ongoing and completed actions.

Experiment 1: Saccadic Response

Methods

Participants Thirty undergraduate students in introductory psychology courses at Indiana University participated in this experiment (Mean age: 19.6 years, SD: 0.8 years; Left handed: n=6; Male: n=9).

Stimuli & Apparatus The stimuli (Figure 1) consisted of videos of an actress, visible from the neck down, reaching to grasp one of two objects resting on the table in front of her. The actress reached with either her right or left hand to the object on her ipsilateral or contralateral side. The full duration of the reach was 1000 ms and the grasp around the object was completed after 1250 ms. Trials were presented in a pseudorandom order with no more than three consecutive trials of the same reach type or same goal location. The stimulus cue began as a filled black square centered on the screen and the two response targets were unfilled black rectangles placed around the location of each potential reach goal.

Stimulus presentation was performed using E-Prime (Psychology Software Tools, Pittsburg, PA). Gaze data was collected with a Tobii TX300 eye tracking system (Stockholm, SE) sampling at 120Hz.

Design and Procedure On each trial the video began with the actress placing her hands on the table in front of her and then reaching for one of the objects. After a variable amount of time (stimulus onset asynchrony, SOA) the stimulus cue changed from black to either blue or red to cue a target. Participants were instructed to fixate the stimulus cue and wait for the color change, then to shift their gaze to the rectangle around the object on the right when the cue was red and to the rectangle on the left when the cue was blue. They were also told that if they looked away from the cue before it changed the trial would be repeated. When gaze was detected within the correct response-target AOI for a minimum of 500ms the color of the response-target rectangle changed to grey to indicate to the participant that a response had been recorded. On 50% of the trials the cued direction was congruent with the direction of the reach and on 50% of the trials the cued direction was incongruent with the direction of the reach. Immediately following the response the video stopped and remained on the screen for a variable inter-trial interval between 300 and 1000ms.

In order to explore the effects of the time course of the observed action on spatial attention, the following SOAs were used in this experiment: -250, 0, 250, 500, 750, 1000, and 1250ms. A negative SOA indicates the stimulus cue changed color before the onset of the reach, whereas a positive SOA indicates the cue occurred after the reach. Participants completed 6 blocks of 60 trials for a total of 360 trials, with congruent and incongruent trials counterbalanced. The number of ipsilateral and contra-lateral trials as well as the number of right and left handed reaches were also counterbalanced.

Results and Discussion

Response Time Figure 2a presents the mean saccadic response times for congruent and incongruent trials at each SOA. Saccadic response times were analyzed in a 2 (congruence) x 7 (SOA) within subjects ANOVA. This analysis revealed a significant main effect of SOA ($F(6,24)=12.65$, $p<.0001$, $\eta^2=.76$) and a significant interaction between SOA and Congruence ($F(6,24)=3.22$, $p=.018$, $\eta^2=0.45$). Paired comparisons revealed significantly faster responses to congruent than incongruent cues at 250ms ($t(29)=3.55$, $p=.001$, $d=0.60$), but the opposite effect, faster responses to incongruent than congruent cues, at 1250ms ($t(29)=2.73$, $p=0.011$, $d=0.26$).

Replication In the preceding study we could not predict a priori which SOAs would show a congruence effect or reverse-congruence effect, which was the reason for a partial replication. We focused on SOAs that previously showed a significant effect (250 and 1250ms) as well as an even longer SOA (1600ms) to determine if this reverse-congruence effect continues beyond the end of the observed reaching action. Analysis of the saccadic response times revealed main effects of SOA ($F(2,23)=5.24$, $p=.01$, $\eta^2=.31$) and congruence ($F(1,24)=7.59$, $p=.01$, $\eta^2=.24$) as well as an interaction between SOA and Congruence ($F(2,23)=6.08$, $p=.008$, $\eta^2=.35$). Paired comparisons revealed congruent responses are faster than incongruent at 250ms ($t(24)=3.67$, $p.=001$, $d=0.48$), and incongruent faster than congruent at both 1250 ms ($t(24)=2.80$, $p=.01$, $d=0.88$) and 1600 ms ($t(24)=3.21$, $p=.004$, $d=0.88$).

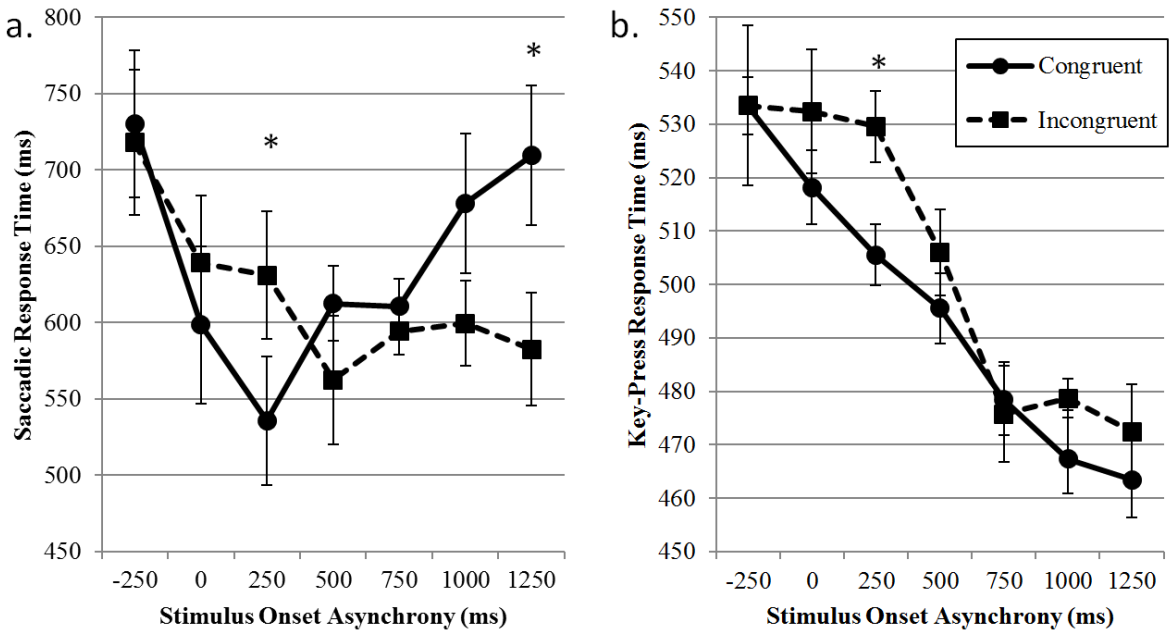


Figure 2. Saccadic response times (a) and Key-Press response times (b) for congruent and incongruent trials in Experiments 1 and 2 respectively. Error bars represent ± standard error of the mean. * p<.05.

Discussion The results from this experiment demonstrate that the observation of a reach reflexively shifts attention in the direction of the goal. Even though participants were instructed to ignore the reaching action, their attention was captured by this stimulus and they were unable to inhibit covertly orienting in the reach direction, which resulted in faster saccadic responses to congruent targets and slower responses to incongruent targets. We speculate that this shift in spatial attention is a consequence of the automatic activation of the predictive saccade to the reach goal due to the direct matching of the observed action to its corresponding motor representation. This congruence effect was strongest at 250 ms SOA, and then began to dissipate as participants had sufficient time to inhibit their reflexive attentional shift and instead plan an intentional, controlled response. It is possible that observing the reaching action nearing its completion also contributed to a reduction in the activation of the eye movement response. This experiment, however, cannot dissociate these two factors because long SOAs and reach completion are confounded.

The reverse-congruence effect seen at longer SOAs was an unexpected result and one that is not typically reported in spatial cueing studies with social stimuli (for an exception see Friesen and Tipper, 2004). This pattern of responses is similar to the well-established phenomenon of inhibition of return (IOR) where saccades toward a previously attended location are slower compared to an unattended location (Posner & Cohen, 1984). In this experiment, however, the time course of the effect is significantly delayed relative to the typical IOR pattern (Klein, 2000), and it has previously been reported that social cues do not result in IOR even at long SOAs (for a review see Frischen, Bayless & Tipper, 2007). A related concept is the 'Social IOR' (Skarratt, Cole, & Kingstone, 2010) which describes the tendency for people to perform slower reaches to a location where their joint-action partner has recently reached. For these reasons, we are hesitant to suggest that the reverse-congruence effect is another example of IOR. Instead, we believe that it is in part due to the task instructions associated with this experiment as we will discuss later on.

The congruence effect in this experiment is consistent with the premotor theory of attention which proposes that a shift in spatial attention necessarily precedes a goal directed action such as an eye movement (Rizzolatti, Riggio, & Sheliga, 1994). This theory would also predict that automatic shifts in attention would be associated with responses in any modality. This is in contrast to the results of Crostella et al. (2009) where there was a one-to-one mapping of the stimulus and the affected response. Those results can be attributed to the automatic activation of a

To test if both the congruence as well as reverse-congruence effects of the current experiment would be replicated with another motor response, we modified the current task by substituting a key-press for a saccadic response.

Experiment 2: Key-Press Response

Methods

Participants Twenty students participated in this experiment (Mean age: 19.7 years, SD: 1.3 years Left handed: n=6 Male: n=5).

Stimuli and Apparatus The stimuli and the SOAs in this experiment were identical to Experiment 1. All stimulus presentation and response collection procedures were

performed using E-Prime (Psychology Software Tools, Pittsburg, PA).

Design and Procedure Participants were instructed to press the 'J' key on a standard American computer keyboard with their Right index finger when the stimulus cue changed to red and the 'F' key with their Left index finger when the cue changed to blue.

Results and Discussion

Response Time Analysis of key-press response times revealed a main effect of SOA ($F(6,14)=4.12$, $p=.02$, $\eta^2=.62$), but no main effect of congruence ($F(1,19)=1.42$, $p=.23$, $\eta^2=.07$) nor interaction between SOA and congruence ($F(6,14)=0.91$, $p=.32$, $\eta^2=.36$). Planned comparison revealed participants were significantly faster to congruent than incongruent cues at 250 ms SOA ($t(19)=2.89$, $p=0.009$, $d=0.15$, Figure 3).

Replication Because there was no evidence of a reverse-congruence effect at the longer SOAs we decided to focus our replication efforts on the range of SOAs where we expected the congruence effect. The analysis of response times confirmed significant main effects for SOA ($F(6,14)=13.77$, $p<.001$, $\eta^2=.86$) and Congruence ($F(1,19)=11.33$, $p=.003$, $\eta^2=.37$) with significantly faster responses to congruent than incongruent trials at 125, 250, 375 and 500 ms SOAs (all $t(19)>2.38$, all $p<.03$).

Discussion The results of this experiment were consistent with our expectation that participants would again respond faster to the congruent target at short SOAs. In contrast to Experiment 1, this experiment showed no evidence of a reverse-congruence effect at long SOAs. This result suggests that the effects of the irrelevant reach and grasp action were different as a function of response mode.

We hypothesize that the reason for a reverse-congruence effect in only Experiment experiment1 is because the saccadic response was influenced by both the irrelevant stimulus facilitating a congruent response as well as the task demands that inhibited a congruent response. As illustrated by Figure 3, each trial begins with inhibition of an eye movement in any direction because participants are instructed to fixate the stimulus cue and not move their eyes. Once the reaching action begins at time 0 ms, there is a covert shift in attention as well as a gradual build-up in the activation of a proactive eye movement. Recall that this attentional bias comes from the observation of the reaching action and the reflexive activation of a predictive eye movement to the reach-goal. Based on our results, this activation peaks at around 250ms and then decreases due to competition from the intentional inhibition of an eye movement. When the reach ends at 1000ms the activation associated with the predictive eye movement ends, but the inhibition remains resulting in an inhibitory rebound.

The results of Experiment 1 are consistent with this explanation. In the 250 ms SOA, congruent responses are facilitated relative to incongruent responses. Then, the saccadic response times in congruent trials consistently

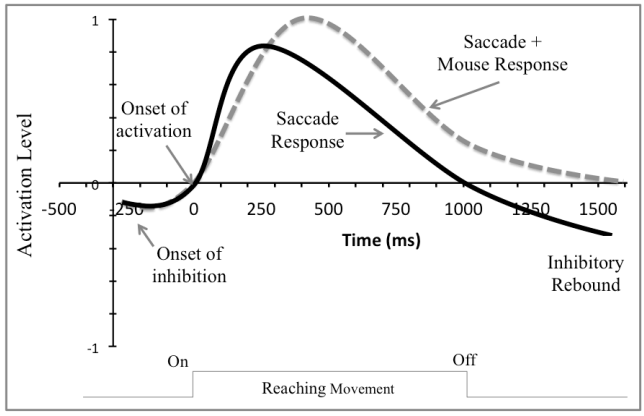


Figure 3. The proposed model of activation and inhibition of eye movements toward the reach-goal that explains the reverse-congruence at long SOAs in saccade response times (E1; solid line) but not saccadic responses in conjunction with mouse movements (E3; dashed line)

increase from 250 ms to 1250 ms SOA while the saccadic response times in incongruent trials remains relatively flat (Figure 2). This is because the predictive eye movement to the reach-goal is being inhibited which slows responses in the same direction. This inhibition does not affect responses in the incongruent trials. Furthermore, this inhibition is not evident in the key-press responses. Instead, the congruent responses are consistently facilitated relative to the incongruent responses. We suggest that this is because the inhibition of the key press response does not increase over the course of the trial and is applied equally to congruent and incongruent responses.

To test our hypothesis about selective inhibition of the predictive saccade, we modified the task to include a mouse movement response and measured how this multi-modal response affected saccadic response times. Note that mouse movements, like goal-directed reaches, are visually guided actions. Therefore, the eyes are activated in this task to both proactively move in the direction of the actor's reach, and to guide the mouse movements to the target in response to the cue. We expected that the additional activation associated with preparing a mouse movement that would not subside until the response was completed would offset the inhibition of the predictive saccade at the completion of the reach by the actor.

Experiment 3: Mouse Movement + Saccade Response

Methods

Participants Twenty-seven students participated in this experiment (Mean age: 19.5 years, SD: 0.9 years; Left handed: n=7; Male: 12), three participants were tested and excluded from analysis due to noncompliance with task instructions.

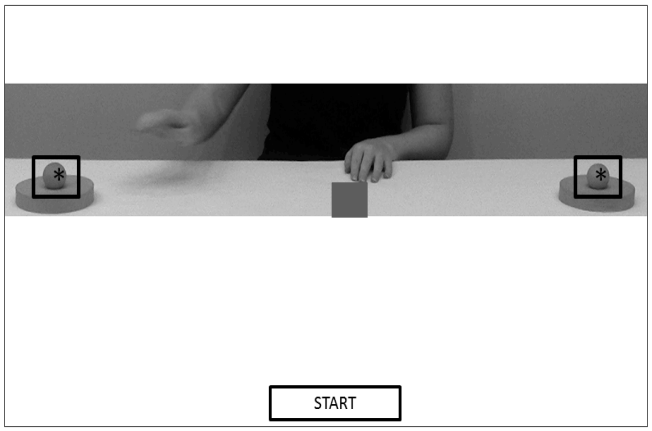


Figure 4. Still frame from Experiment 3 immediately following the color change of the cue in the 500ms SOA condition.

Stimuli and Apparatus Video stimuli from Experiment 1 was cropped to include only the reaching action and presented in the top third of the screen. Additionally, text-box centered at the bottom of the screen with the word 'START' and an asterisk was centered within the response rectangles over each of the objects. The following SOAs were used in this experiment: 250, 500, 750, 1000, 1250, and 1600ms.

Design Participants were instructed to initiate each trial by clicking the mouse on the 'START' box, then remaining on box until the cue changed. They then moved the cursor and clicked on the asterisk in the right box if the cue was red and clicked the asterisk in the left box if the cue was blue.

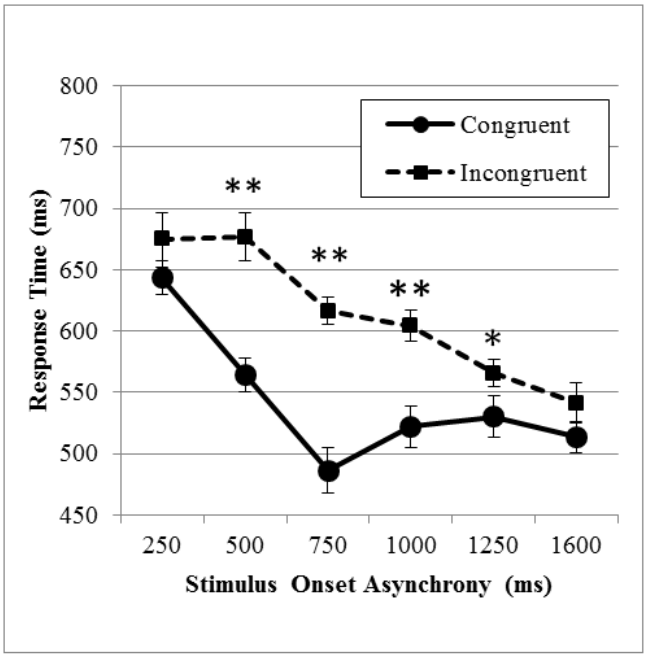


Figure 5. Saccadic response time in Experiment 3. Error bars represent ± standard error of the mean. * p<0.05; ** p<.001.

Results and Discussion

Response Time Analysis of saccadic response times revealed significant main effects of both SOA ($F(5,22)=11.99$, $p<0.001$, $\eta^2=.73$) and congruence ($F(1,26)=29.58$, $p<0.001$, $\eta^2=.53$) as well as a significant interaction between SOA and congruence, ($F(5,22)=3.20$, $p=.025$, $\eta^2=.42$). Planned comparisons revealed significantly faster congruent responses at the 500, 750, 1000, and 1250 ms SOAs (all *t's*(26)> 2.18, all *p's*<.04). Critically, there was no evidence of a reverse-congruence effect at 1250 or 1600ms.

An analysis of the mouse response times revealed only a main effect of SOA ($F(5,22)=4.28$, $p=.007$, $\eta^2=.49$) due to faster response times at the longer SOAs. Neither the main effect of congruence ($F(1,26)=2.64$, $p=.116$, $\eta^2=.09$) nor the interaction between SOA and congruence ($F(5,22)=0.51$, $p=.77$, $\eta^2=.10$) were significant.

Discussion As predicted saccadic response times continued to show an early congruence effect, albeit delayed relative to Experiments 1 and 2, but no reverse-congruence effect as seen in Experiment 1. This is consistent with our hypothesis for the reverse-congruence effect, because, in this case, the eye movement continues to be facilitated as it is needed to guide the mouse movements. Note that there was no evidence of response times in the congruent condition increasing as SOAs increased beyond 250ms. This is because the build-up of an inhibitory response to prevent a premature eye movement was offset by the preparation to move the mouse in the congruent direction. Unlike the activation of a proactive eye movement which was stimulated from observing the actor's reach, the activation of the eye movement associated with the mouse movement continued until the completion of the response.

Interestingly, the mouse response times did not reveal either a congruency or reverse-congruency effect, and, in fact, showed no consistent differences between congruent and incongruent responses. This result underscores the importance of using multiple response measures as the mouse response-time alone may have lead one to conclude that there was no effect of action observation on attention distribution.

General Discussion

Taken together the results of these three experiments demonstrate that attention is automatically captured by the observation of dynamic reaching actions. In all three experiments participant's responses are faster when the response direction is congruent with the direction of the reach, but only when the cue to respond occurs shortly after the onset of the reach stimulus. This only tells part of the story however, because the time course of the reach's influence is not the same across all responses. The eye movement response in Experiment 1 showed a significant reverse-congruence effect, where the responses to cues at long SOAs were slower in the direction congruent with the reach. We suggest that this reverse-congruence reflects a

steady build-up of inhibition in order to ensure that the eye does not move from the fixation square until cued. We tested this hypothesis by adding a visually-guided response in Experiment 3, and the results revealed that the activation associated with preparing a mouse movement overcame the inhibition of the saccadic response..

These results contribute to our understanding of reflexive shifts in attention to observed actions in three ways:. First, we used a dynamic stimulus. By presenting videos of a reaching action, as opposed to still images of a grasp, we were able to examine the role of ongoing actions in a more ecologically valid way. While this opens up the possibility that any dynamic motion stimulus would similarly drive attention, we predict that this would not be the case. Recall that we presented both ipsilateral and contralateral actions. Contralateral actions start on the side opposite from the eventual reach-goal, and would therefore initially facilitate incongruent responses if shifts in attention were driven only by movement. Critically, the congruence effect was consistently present in the 250ms SOA despite the fact that 250ms into a contralateral reach the hand had just reached the midline of the actor (Figure 1) and therefore the perceived direction of the hand movement was still somewhat ambiguous.

Second, presenting a dynamic reaching action helps to clarify an apparent inconsistency in the literature where some studies are reporting no spatial orienting following a grasp (Fischer & Szymkowiak, 2004; Mazzarella et al., 2012) and others are reporting that an incomplete grasp does automatically drive attention (Fischer et al., 2008, Daum & Gredebäck, 2011). By studying the effects of an ongoing reaching action on automatic spatial orienting we observed results that are consistent with both results. Attention is automatically oriented in the direction of the reach while the reach is ongoing. This facilitation ends when the reach is complete, supporting Fischer & Szymkowiak (2004) explanation that completed actions do not drive attention.

Finally, the use of multiple response measures revealed that automatic spatial orienting affected responses differently depending on the response modality. Taken together the results of these experiments support a hypothesis that could not be tested by one experiment alone. The reverse-congruence effect observed in Experiment 1 appears very similar to inhibition of return despite evidence that social cues do not elicit IOR. Increasing the activation of the eye movements in the direction of the reach by adding the mouse movement response eliminated the reverse-congruence effect and allowed us to reject the IOR explanation while providing support for our hypothesis of specific saccade inhibition.

References

Bertenthal, B. I., Boyer, T. W., & Harding, S. (2014). When do infants begin to follow a point? *Developmental Psychology*, *50*(8), 2036–2048.

Crostella, F., Carducci, F., & Aglioti, S. M. (2009). Reflexive social attention is mapped according to effector-specific reference systems. *Experimental Brain Research*, *197*(2), 143–151.

Daum, M. M., & Gredebäck, G. (2011). The development of grasping comprehension in infancy: covert shifts of attention caused by referential actions. *Experimental Brain Research*, *208*(2), 297–307.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, *6*(5), 509–540.

Eshuis, R., Coventry, K. R., & Vulchanova, M. (2009). Predictive eye movements are driven by goals, not by the mirror neuron system. *Psychological Science*, *20*(4), 438-440.

Fischer, M. H., Prinz, J., & Lotz, K. (2008). Grasp cueing shows obligatory attention to action goals. *The Quarterly Journal of Experimental Psychology*, *61*(6), 860–868.

Fischer, M. H., & Szymkowiak, A. (2004). Joint Attention for Pointing but not Grasping Postures. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *40*(1), 168–170.

Flanagan, J. R. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze Cueing of Attention: Visual Attention, Social Cognition, and Individual Differences. *Psychological Bulletin*, *133*(4), 694–724.

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, *4*(4), 138-147.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59.

Mazzarella, E., Hamilton, A., Trojano, L., Mastromauro, B., & Conson, M. (2012). Observation of another's action but not eye gaze triggers allocentric visual perspective. *The Quarterly Journal of Experimental Psychology*, *65*(12), 2447–2460.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X* (pp. 531–556). Hillsdale, NJ: Erlbaum

Rizzolatti, G., Riggio, L., & Sheliga, B. M. (1994). Space and selective attention. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV*, *15*, 231-265.

Skarratt, P. A., Cole, G. G., & Kingstone, A. (2010). Social inhibition of return. *Acta psychologica*, *134*(1), 48-54.

Longitudinal L2 Development of the English Article in Individual Learners

Akira Murakami (am933@cam.ac.uk)
Theodora Alexopoulou (ta259@cam.ac.uk)
Department of Theoretical and Applied Linguistics, 9 West Road,
University of Cambridge, Cambridge, CB3 9DP, United Kingdom

Abstract

We investigate the accuracy development of the English article by learners of English as a second language. The study focuses on individual learners, tracking their learning trajectories through their writings in the EF-Cambridge Open Language Database (EFCAMDAT), an open access learner corpus. We draw from 17,859 writings by 1,280 learners and ask whether article accuracy in individual learners fluctuates randomly or whether learners can be clustered according to their developmental trajectories. In particular, we apply k-means clustering to automatically cluster in a bottom up fashion learners with similar learning curves. We follow learners for a period covering one CEFR level. Given the relatively short learning window, the majority of learners follow a horizontal line. Nevertheless, we also identify groups of learners showing a power-function and U-shaped curve. Crucially, these groups are 'hidden' when the aggregate of learners is considered, a finding highlighting the importance of individual level analysis.

Keywords: learning curve; clustering; individual variation; second language

Introduction

Since the early days of second language acquisition (SLA) research, learner production data have been an important empirical base for developmental research (Selinker, 1972). Early studies revealed the non-linearity of the learning process and, in particular, the existence of U-shaped learning curves (Lightbown, 1983). However, despite important insights regarding learning curves, a number of important issues remain unaddressed. Most studies based on production data tend to be small scale (Lightbown, 1983) and tend to lack longitudinal information for large numbers of individuals. As a result, developmental research has predominately employed cross-sectional designs while it has generally not been possible to investigate the relation between individual learning curves and cross-sectional or aggregate patterns. This has been an important limitation in SLA research in view of evidence from psychology that the averaged pattern can conceal individual developmental trajectories (e.g., Heathcote, Brown, & Mewhort, 2000). These limitations are acknowledged by SLA researchers calling for a stronger focus on individual learners (e.g., Larsen-Freeman, 1997). However, the absence of sufficient amounts of longitudinal data for individuals is an important practical obstacle (Larsen-Freeman & Cameron, 2008).

Recently developed large-scale learner corpora (e.g., Granger, Dagneaux, Meunier, & Paquot, 2009) provide production data rich enough to allow analysis of individuals at various proficiency levels. In the present study, we exploit one such resource, EFCAMDAT, as it enables tracking the productions of sufficient numbers of individual learners. We focus on the accuracy in the use of the English article by learners of English as a second language (L2 learners). Accuracy is an important (though not sole) indicator of acquisition and the article has been central to accuracy investigations from early days (e.g., Dualy & Burt, 1973). In recent work, we show that, despite systematic effects of the native language on the accuracy of the L2 English article, there exists significant individual variation (Murakami, 2016). In the present study, we investigate whether we can identify learning curves in individuals that are absent from the aggregate analysis.

Method

Corpus

The empirical data for the study are drawn from the pre-release version of EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2014) The corpus contains learners' writings submitted to *Englishtown*, the online school of EF Education First. The 16 teaching levels of the EF curriculum range from beginners to advanced proficiency (aligned to A1-C2 in the Common European Framework Reference (CEFR) levels). Writings are on a variety of topics and range from 20-40 words at beginner levels to 150-180 words each at the highest levels. The scripts in the corpus are accompanied by error corrections by teachers. EFCAMDAT is publicly available at http://corpus.mml.cam.ac.uk/efcamdat/.

Each writing includes an anonymous learner ID, national language, the topic of the writing, the date and time of submission, the level and the unit/lesson number the writing was submitted to and error-corrections by teachers (see Geertzen et al, 2014 for further details on EFCAMDAT).

Target Linguistic Feature and L1 Groups

We targeted English articles, indefinite (*a*, *an*) and definite (*the*), because they are highly frequent and therefore provide us with dense longitudinal data while we could calculate their accuracy automatically with high precision.

The effect of mother tongue/First Language/L1 is pervasive in L2 use (Jarvis & Pavlenko, 2007) and specifically shown for the L2 English article (e.g., Murakami & Alexopoulou, 2015; Snape, 2008). Our study targeted 10 L1 groups; Brazilian-Portuguese, Mandarin-Chinese, German, French, Italian, Japanese, Korean,

Russian, Spanish, and Turkish. We used their country of residence as the closest approximation to L1 (Murakami, 2016). L1 Spanish includes two countries of residence; Spain and Mexico. L1 Mandarin-Chinese includes both those living in China and those in Taiwan. L1 Brazilian-Portuguese, German, French, Italian, Korean, Russian, and Turkish learners correspond to those living in Brazil, Germany, France, Italy, Korea, Russia, and Turkey, respectively. This method has yielded reliable L1 effects (Murakami, 2016).

We grouped L1s in two L1-types, PRESENT and ABSENT, depending on whether or not they have an article. The PRESENT group included L1 Brazilian, German, French, Italian, and Spanish. The ABSENT group included L1 Chinese, Japanese, Korean, Russian, and Turkish.

The subcorpus used in the study included nearly 140,000 writings consisting of 10 million words. There were more writings, and thus a larger number of words, at lower levels. The distribution between L1 groups was skewed as well. Over 40% of the data were contributed by L1 Chinese learners of English, and another large portion (14%-23% each) by L1 Brazilian and L1 Russian learners.

Scoring Method

To investigate accuracy in use we need to identify contexts of obligatory use and measure correct suppliance while also capturing erroneous use of the article in contexts where it is not needed (overgeneralization errors). We thus employed the target-like use (TLU) score calculated by the following formula (Pica, 1983);

$$\text{TLU Score} = \frac{\text{number of correct suppliances}}{\text{number of obligatory contexts} + \text{number of overgeneralization errors}}$$

This formula assesses the proportion of correct use and at the same time penalizes the unnecessary use of the article.

Data Extraction

To measure accuracy, we first need to obtain obligatory contexts of article use. We retrieved error-tagged texts and converted them to corrected texts (incorporating the corrections). The number of article occurrences in a corrected text was taken as the number of obligatory contexts in the learner's original writing. For instance, if the phrase *farmer who lived in a small village* was corrected into *a farmer who lived in a small village*, the two instances of *a* in the latter made two obligatory contexts. R scripts were written to automatically retrieve errors using the teacher error tags.

Tracking Development Through Moving Windows

To track individuals over time, we consider consecutive writings and compare TLU scores from one writing to next. However, writings are too short to provide enough obligatory contexts to reliably calculate the TLU score. We thus computed TLU scores over multiple writings constructing windows as close to a single writing with enough obligatory contexts as possible. Our goal was to ensure that the shape of accuracy development is as close as possible to the shape that would be generated if each essay included a large number of obligatory contexts.

We then calculated TLU scores in a moving-window fashion (e.g., van Geert & van Dijk, 2002). Each window included at least 10 obligatory contexts (OCs) and could cover multiple writings. Consider a learner who wrote five writings with OCs as in the following example:

Writing 1; 6 OCs
Writing 2; 6 OCs
Writing 3; 3 OCs
Writing 4; 7 OCs
Writing 5; 2 OCs

Here, the first TLU score would be calculated over Writing 1 and Writing 2 because Writing 1 alone does not include 10 OCs but Writing 1 and Writing 2 combined do. The first window would thus include 12 OCs. In computing the second TLU score, the head of the first window shifts forward by one, and the second window starts from Writing 2. The second window would cover Writing 2 through Writing 4, because Writing 2 alone or Writing 2 and Writing 3 together do not include 10 OCs, but the three writings combined do. The second window would include 16 OCs in total. Similarly, the head of the window now shifts to Writing 3, and the third TLU score would be calculated over Writing 3 and Writing 4. This learner has these three TLU scores in total, as Writing 4 alone, Writing 4 and Writing 5 combined, or Writing 5 alone does not include 10 OCs. Once TLU scores were obtained for all the windows, we analyzed the learners who had 10 or more windows.

Alternatively, it was also possible to construct windows so that no writing overlaps in any window. We chose the overlapping approach because it generally tracks more fine-grained developmental patterns and finer resolution of data is indispensable in the study in order to observe change (Larsen-Freeman & Cameron, 2008). See Murakami (2014) for further discussion.

There were in total 70,879 TLU scores (windows) by 20,394 learners. Out of the 20,394 learners, 1,280 (6.3%) had 10 or more windows. Among those with the minimum of 10 windows, the average number of writings in a window was 2.4 (*SD* = 1.2). The median number of windows was 12. The total number of writings by them was 17,859. The mean number of unique writings over 10 windows was 11.0 (*SD* = 0.9). This is not 10 times as large as the average number of writings in a window because there are overlaps of writings over windows. The average number of Units covered in 10 windows was 25.6 (*SD* = 10.1). This corresponds to just over three EF teaching levels — if the writings always start at levels 1, 3, 7, 10, etc., then it corresponds to one CEFR level. Note that three levels correspond to 24 writings (3 Levels × 8 Units), but the average number of unique writings over 10 windows is less than half of this (11.0). Recall that we selected only error-tagged writings for analysis. As not all writings are error-tagged, consecutive writings in our windows do not

necessarily correspond to adjacent Englishtown tasks and, indeed, span just over three teaching levels on average.

Longitudinal View of Development

There are multiple ways to visualize longitudinal accuracy in use. Figure 1 lists two of them for the 1,280 learners. In both panels, each thin line represents the accuracy development of one learner, and thick lines are locally weighted scatterplot smoothing (LOESS; Singer & Willett, 2003) lines showing the overall trend. Solid, dashed, and dotted lines correspond to the development of the ABSENT, the PRESENT, and the L1 Chinese groups respectively. L1 Chinese was separated because visual inspection (not shown) suggested that they behave differently from the other ABSENT groups, and there are indeed studies claiming that Mandarin-Chinese has linguistic features that play similar roles to English definite (Huang, 1999) and indefinite (Chen, 2004) articles.

The left panel demonstrates the accuracy development of individual learners across Englishtown teaching levels. Englishtown levels 1-16 are shown on the horizontal axis and accuracy is on the vertical axis. Each data point is plotted according to its window proficiency, defined as the weighted average Unit number (1 to 128) of the writings included in the window. Each learner horizontally covers only a short span because a learner typically covers three to four levels. The trend line indicates that accuracy increases as learners' proficiency rises. Individually, however, we can observe that some learners radically go up and down in their accuracy. Since each learner covers a short span in this panel, it is difficult to analyze the development of individual learners here. The trend lines are the cross-sectional trend lines and do not represent longitudinal development.

In order to better understand longitudinal development, the right panel visualizes the development of individual learners so that each learner covers the entire span from left to right. For the sake of comparability, it only targets the first 10 windows of each learner. The horizontal axis here represents the window number. The panel shows that the overall pattern is relatively horizontal, which may appear to contrast with the left panel that shows the overall accuracy increase across Englishtown levels but is in fact consistent, given the short span covered in the longitudinal development in this study. The same panel also demonstrates large individual variation in the development. The accuracy of some learners appears to increase over time, while the accuracy of others seems to fluctuate widely. This suggests that there can further be a difference between the average longitudinal pattern and the learning curves of individual learners.

Clustering Learners According to Their Developmental Shapes

K-Means Clustering and Number of Clusters

Now that we visually grasped the longitudinal development of the morphemes, we will explore the learning curves that characterize the learners we target. To obtain the groupings that reflect the data, we employed a data-driven way of determining developmental shapes, namely k-means clustering.

As the input data, we used the first 10 windows of each learner. Our interest here is the developmental trajectory. However, if k-means clustering is run on the present data as they are, it will take into account the absolute accuracy of each learner and may cluster learners according to their accuracy. To neutralize the effect of absolute accuracy, all the data points were learner-mean-centered: The mean accuracy value of each learner was subtracted from all the data points of the learner within the first 10 windows.

Because the bottom-up approach is open not only to the shape but also number of clusters, determining the shape and number of clusters is two separate issues. The approach we take is clustering with different numbers of clusters and examining how emerging patterns vary across the clustering. For instance, we can establish which k certain patterns, such as U-shaped development, appear at. If a certain pattern consistently appears with varying k's, it is likely that the pattern reflects some kind of mechanism in learners' performance development.

We chose 10 for the maximum number of clusters because there are 10 L1 groups in total and 10 different patterns are expected if each L1 has a distinct developmental pattern. Figure 2 shows the LOESS of the developmental patterns of each cluster when k varies from 2 to 10. Note that although k-means clustering was performed on learner-mean-centered data, we show the result in the original scale (i.e., TLU scores) for the ease of interpretation. The horizontal axis of the figure represents window number and the vertical axis represents TLU scores. Each panel represents the clustering when the k is the value stated above the panel. Each line is the LOESS of the development of the learners in each cluster. Cluster A is always the largest cluster in terms of the number of included learners, followed by Cluster B, which in turn is followed by Cluster C, and so forth.

We can make a few observations here. First, from $k = 4$ onwards, the accuracy of the largest cluster (Cluster A) is relatively unchanged. Second, there is always a cluster showing an upward trend over 10 windows (e.g., Cluster B in $k = 2$) and a cluster showing decreasing accuracy over the period (e.g., Cluster C in $k = 3$). Third, the U-shaped pattern is prevalent. It first appears in $k = 3$ as Cluster A, and can always be observed until $k = 10$ as Clusters C, H, and I. Finally, some clusters are reminiscent of the power and the exponential function. For instance, Cluster B in $k = 3$, Cluster C in $k = 4$, and Cluster D in $k = 5$ are all more or less similar to the concave function in that their accuracy increase slows down as learners progress.

In sum, the patterns of (i) relatively horizontal, (ii) increasing or decreasing accuracy including the power-law and the exponential shape, and (iii) U-shaped curve are robust and can be observed almost irrespective of the number of clusters. On the other hand, increasing the

number of clusters only results in finer splits of the same patterns and hardly leads to the emergence of new patterns. As shown in Figure 2, few new patterns emerge after $k = 3$. This means that having $k = 4$ or more is unlikely to reveal important longitudinal patterns. We thus assume $k = 3$ is optimal for article patterns derived from k-means clustering. The main criterion for decision here is how informative each k is, and whether new information or pattern can be revealed. Note that although we chose $k = 3$ as the number of clusters, we repeated the procedure for $k = 2$ and $k = 4$ and confirmed that our results, including cluster validation discussed later, remain the same. The point here is not to argue that $k = 3$ is the best number of clusters, or subpopulations, but to illustrate that there are indeed subgroups and that they can be hidden from the aggregated data.

Individual Developmental Patterns of Clusters

Figure 3 shows the clusters of article development when $k = 3$. The upper three panels present the clusters situated in the entire Englishtown course, and the lower three panels show the same data over 10 windows. In other words, the upper panel corresponds to the left panel in Figure 1 divided into the three clusters, while the lower panel corresponds to the right panel in Figure 1, again divided into the three clusters. As before, thin lines represent individual learners and thick lines represent LOESS. The learners are approximately equally spread between the three clusters. L1 type does not affect clustering because the three LOESS lines largely overlap. Learners in Cluster A show a smooth U-shaped developmental curve. Their accuracy slowly decreases for the first five or six windows, after which it slowly increases. Their overall accuracy is high, being consistently over 0.8. The accuracy in Cluster B gradually rises until around the seventh window, after which it levels off possibly due to the ceiling effect. The pattern is reminiscent of the power/exponential developmental pattern. Learners in Cluster C show a horizontal development until the fifth window, after which their accuracy decreases. Significant individual differences can be observed in each cluster, some learners radically going down and others rapidly going up.

Cluster Validation

An important question at this point is whether the identified clusters are 'real'. In other words, is it the difference in the true learning curve that the k-means clustering above reflects or is it just the random noise? We empirically tested the cluster validity in the following manner. The overall idea is similar to the usual statistical testing procedure: We compute the null distribution of the metric that measures goodness of clustering based on random data, and if the value of the metric in the observed data falls outside of its 95% range, we consider it as the evidence that the observed clusters are too good to be derived from random data and conclude that our clusters indeed reflect the difference in the learning trajectory. The key decisions we need to make are (i) what metric to use to measure goodness of clustering and (ii) how to conceptualize the null hypothesis and derive the null distribution, both of which are detailed below.

As the metric for goodness of clustering, we used a measure called the *silhouette* (Rousseeuw, 1987). Intuitively, the silhouette value is large if within-cluster dissimilarity is small (i.e., learners within each cluster have similar developmental trajectories) and between-cluster dissimilarity is large (i.e., learners in different clusters have different learning curves). The silhouette is given to each data point, and all the silhouette values are averaged to measure the cluster distinctiveness of a cluster analysis. Any distance can be used to calculate dissimilarity, and squared Euclidian distance was used in the present case. The mean silhouette value in our clustering was 0.151. If this value is higher than the 95% range of the null distribution of the mean silhouette value, we conclude that there are multiple learning curves.

The null hypothesis is that there is no systematic pattern in intra-learner variability. We obtained the distribution of the mean silhouette value under this null hypothesis and tested the significance of our observed value. The idea here is that we practically randomize the order of the writings within individual learners and follow the same procedure as our main analysis. Since the order of writings is random, there should not be any systematic pattern of development observed. The clusters obtained in this manner thus captures noise alone. We calculate the mean silhouette value on the noise-only, random clusters, and obtain its distribution by repeating the whole procedure a large number of times.

More specifically, the following procedure was employed.

1. For each learner,
    (a) We randomly sampled a large number (e.g., 100) of his/her writings. Here, the same writing could be selected multiple times.
    (b) We then calculated the TLU scores of the first 10 windows.
2. With the data obtained in 1, we ran a k-means cluster analysis with k = 3 and calculated its mean silhouette value.
3. 1 and 2 were repeated 1,000 times, resulting in 1,000 mean silhouette values that we consider as the null distribution.
4. We examined whether the 95% range of 3 included 0.151, the observed mean silhouette value in the present study.

The resulting null distribution showed that the upper bound of its 95% range was 0.144. Therefore, our clusters with the mean silhouette value of 0.151 is considered non-random, and what our clusters capture is not only noise.

Discussion

The present study investigated the longitudinal accuracy use of the L2 English article. We observed large individual variation in the developmental pattern. We, therefore, investigated whether we can identify systematic learning curves in a bottom-up manner. A cluster analysis identified

three learning curves that are followed by approximately an equal number of learners.

We demonstrated that there can be differences between average longitudinal development of a group of learners and the individual learning curves that constitute the group. When we aggregated the data and looked at the developmental trajectory of the average accuracy, it was fairly horizontal, (falsely) implying constant accuracy over the course of development. However, clustering demonstrated that there are indeed subgroups hidden in the aggregate pattern. This suggests that the two types of learning curves can differ, and that we cannot necessarily infer the development of individual learners based on the longitudinal data aggregated over multiple learners. Whereas prior research in psychology has shown the mismatch between aggregated and individual learning curves a number of times (Heathcote, et al., 2000), we believe the present study is the first that empirically demonstrated it in the field of SLA.

The obvious question now is why should learners show these different developmental patterns, and, in particular what is the cause of the U-shape pattern. Admittedly, we can offer no insight at this point to this question but we can note at least two potential sources of variation: (i) changes in the internal knowledge of learner related to reanalysis or increased complexity which is known to impact negatively on accuracy accuracy (Skehan & Foster, 1997); (ii) learner individual variation (e.g., working memory, aptitude, etc.). The investigation of these potential sources requires investigation of specific hypotheses (e.g., evidence for increased complexity in the learner language coinciding with drop in accuracy) or access to individual psychological data (which would require investigation beyond the corpus). Though the interpretation of such results is very challenging, we believe that our findings show the need for investigating individual patterns in a more comprehensive way.

Acknowledgments

The authors gratefully acknowledge the sponsorship of EF Education First.

References

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics, 42*(6), 1129–1184.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning, 23*(2), 245–258.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 database: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Millar et al. (Eds.), *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines* (pp. 240–254). Cascadilla Proceedings Project.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*(2), 185–207.

Huang, S. (1999). The emergence of a grammatical category definite article in spoken Chinese. *Journal of Pragmatics, 31*(1), 77–94.

Jarvis, S., & Pavlenko, A. (2007). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Larsen-Freeman, D. E. (1997). Chaos/complexity science and second language acquisition. *Applied Linguistics, 18*(2), 141–165.

Larsen-Freeman, D. E., & Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *Modern Language Journal, 92*(2), 200–213.

Lightbown, P. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. Seliger & M. H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 217–243). Rowley, MA: Newbury House.

Murakami, A. (2014). *Individual variation and the role of L1 in the L2 development of English grammatical morphemes: Insights from learner corpora* (Unpublished PhD thesis). University of Cambridge, Cambridge, United Kingdom.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear L2 development: The case of English grammatical morphemes. *Language Learning*. Advance online publication.

Murakami, A., & Alexopoulou, T. (2015). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*. Advance online publication.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*(1), 69–78.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209-232.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research, 1*(3), 185–211.

Snape, N. (2008). Resetting the nominal mapping parameter in L2 English: Definite article use and the count-mass distinction. *Bilingualism: Language and Cognition, 11*(1), 63–79.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior & Development, 25*(4), 340–374.

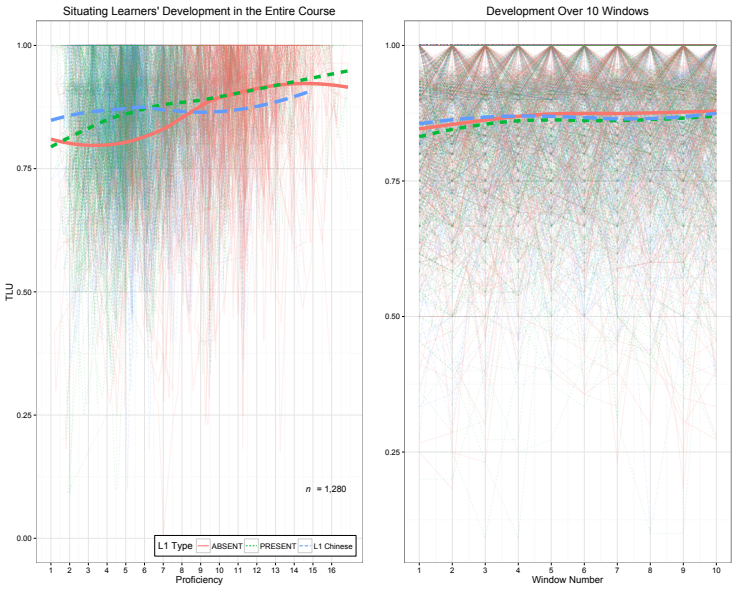


Figure 1: Longitudinal Development of Article Accuracy

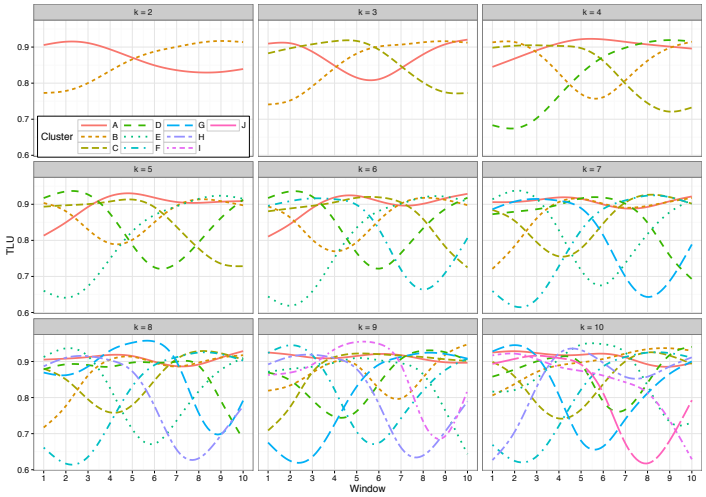


Figure 2: Developmental of Each Cluster in Varying Numbers of Clusters

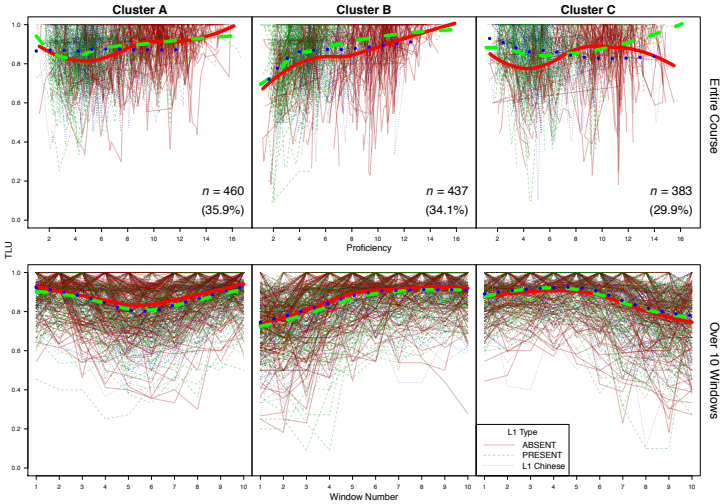


Figure 3: Results of K-Means Clustering

Spatializing emotion: A mapping of valence or magnitude?

Benjamin Pitt[1] **(bpitt@uchicago.edu)**

Daniel Casasanto[1,2] **(casasanto@uchicago.edu)**

[1]Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL 60637 USA
[2]Grossman Institute for Neuroscience, University of Chicago, 5812 S. Ellis Avenue, Chicago, IL 60637 USA

Abstract

People implicitly associate different emotions with different locations in left-right space. Which dimensions of emotion do they spatialize? Across many studies people spatialize *emotional valence*, mapping positive emotions onto their dominant side of space and negative emotions onto their non-dominant side. Yet, other results suggest a contradictory mapping of *emotional intensity* (a.k.a., emotional magnitude), according to which people associate more intense emotions with the right and less intense emotions with the left, regardless of valence. To resolve this apparent contradiction, we first tested whether people implicitly spatialize whichever dimension of emotion they attend to. Results showed the predicted valence mapping, but no intensity mapping. We then tested an alternative explanation of findings previously interpreted as showing an intensity mapping; these data may reflect a left-right mapping of spatial magnitude, not emotion. People implicitly spatialize emotional valence, but there is no clear evidence for an implicit lateral mapping of emotional intensity.

Keywords: conceptual metaphor theory; emotion; magnitude; mental metaphor; valence

Introduction

People implicitly associate different emotions with different locations in left-right space, mapping points along an imaginary continuum of emotions onto an imaginary lateral spatial axis. Which dimensions of emotion do people spatialize? In this study we explore two contrasting mappings between space and emotion that have been proposed, a Valence Mapping and an Intensity Mapping, and suggest a resolution to the apparent contradiction between them.

According to many studies, people implicitly spatialize emotional valence, mapping positive-valence emotions onto their dominant side of space and negative-valence emotions onto their non-dominant side. Right-handers, therefore, tend to associate positive emotions (e.g., happiness) with the right, and negative emotions (e.g., anger) with the left (Casasanto, 2014, for review). This "Valence Mapping" appears to be shaped by the physical experiences we have with our hands: right-handers come to associate "positive" with the side of space on which they can usually act more fluently with their dominant hand, and "negative" with the side on which they act more clumsily, with the non-dominant hand (Casasanto & Chrysikou, 2011; see also Oppenheimer, 2008).

This body-based Valence Mapping has been observed across a variety of populations: in children as well as adults (Casasanto & Henetz, 2012), in members of Western and non-Western cultures (de la Fuente, et al., 2015), and in both neurotypicals and patients with compromised motor systems (Casasanto & Chrysikou, 2011). A "good is right" mapping has also been found using a variety of methods, ranging from questionnaires and diagram tasks (Casasanto, 2009) to memory tasks (Brunyé, Gardony, Mahoney, 2012), motor training tasks (Casasanto & Chrysikou, 2011), and analyses of spontaneous gestures (Casasanto & Jasmin, 2010).

Of particular interest for the present study, the Valence Mapping also produces space-valence congruity effects in reaction time (RT) tasks. Across multiple experiments, right- and left-handers were faster to classify centrally-presented words as positive when responding with their dominant hand, and faster to classify words as negative when responding with their non-dominant hand (de la Vega et al., 2012; de la Vega, et al., 2013; Kong, 2013). A similar pattern was found when people judged positive and negative emotional faces (Kong, 2013). In sum, these results provide strong and generalizable evidence for an implicit left-right Valence Mapping.

Is valence the only dimension of emotion that people implicitly spatialize on the lateral axis? According to one study, people spatialize emotional intensity (which the authors refer to as "emotional magnitude"), associating less intense emotion with the left and more intense emotion with the right (Holmes & Lourenco, 2011; hence H&L). Participants responded to photographs of emotional faces that varied in both valence (e.g., happy vs. angry) and intensity (e.g. happy vs. extremely happy). As indexed by their RTs, participants appeared to associate less intense emotions with the left and more intense emotions with the right – regardless of whether the emotional valence was positive or negative (H&L, Experiment 2).

These data of H&L's are consistent with their proposal that people implicitly associate emotional intensity with left-right space: an "Intensity Mapping." However, these data conflict with the Valence Mapping. Whereas an Intensity Mapping predicts that very negative emotions (e.g. extremely angry) should be associated with the *right* (because they are more intense), the Valence Mapping predicts that these negative emotions should be associated with the *left*, in right-handers, (because they have negative valence).

Here we conducted three experiments to resolve this apparent contradiction. In Experiments 1 and 2, we explored the possibility that people have *both* a Valence Mapping and an Intensity Mapping available in long-term memory, and spatialize emotion according to whichever dimension of emotion is more salient in context (broadly consistent with suggestions of H&L's). To preview our results, we found evidence for the Valence Mapping but no evidence of an

Intensity Mapping. Therefore, in Experiment 3, we tested an alternative explanation for H&L's findings based on the *spatial magnitude* (as opposed to the *emotional magnitude*) of a salient feature of their stimuli.

Experiment 1: Do people spatialize whichever dimension of emotion they attend to?

In Experiment 1, we tested whether orienting participants toward either valence or intensity would cause them to show the corresponding spatial mapping. Participants responded to emotional words and judged them according to either their valence or their intensity, by pressing buttons on the left or right of a keyboard.

Method

Participants Thirty-two right-handed adults from the University of Chicago community who spoke fluent English participated in the main experiment for payment or course credit. Half were randomly assigned to make speeded valence judgments (n=16) and the other half to make speeded intensity judgments (n=16). Twenty-four other participants from the same community completed a questionnaire used to measure normative judgments of the stimulus words' meanings.

Materials Four emotional words were selected, on the basis of a previous experiment in French speakers (Carbé & Gevers, 2013), which varied in their valence and intensity: "horrible" (very negative), "bad" (negative), "good" (positive) and "perfect" (very positive).

In order to quantify differences among these words we asked a group of participants (N=24) who did not participate in the main experiment to rate the four words on intensity and valence. For this norming task, the words appeared in a vertical column along the midline of a printed page. Each participant judged the emotional valence of the words on a scale from -5 (most negative) to 5 (most positive) and, on another page, the emotional intensity of the words on a scale from 0 (least intense) to 10 (most intense). They wrote their ratings on a horizontal line to the right of each word, all using their right hand. The position of the words on the page and the order of valence and intensity judgments was counterbalanced across participants.

Procedure In both the valence and intensity tasks, the four stimulus words appeared one at a time in black text in the center of a computer screen. Participants in the valence task judged whether each word connoted a "positive emotion" or "negative emotion." Participants in the intensity task judged whether each word connoted a "more intense emotion" or "less intense emotion." Participants were instructed to respond as quickly and accurately as possible by pressing one of two keys (the "z" and the "?" keys on the English-US QWERTY keyboard) using the index finger of each hand. If no response was given, trials automatically ended after two seconds. In one block of trials, participants pressed the key on the left to indicate a negative/less intense emotion and the key on the right to indicate a positive/more intense emotion. In a second block, this mapping was reversed, and the order of blocks was counterbalanced across participants. Participants performed 8 practice trials at the start of each block, after which the four stimulus words were presented in random order 24 times each, composing a total of 192 critical trials per participant over two blocks.

After testing, participants were debriefed to determine whether they were aware of the experimental hypothesis, and they then completed a language history questionnaire and the Edinburgh Handedness Inventory (EHI; Oldfield, 1971).

Results & Discussion

Ratings of emotional word stimuli Mean valence ratings ranged from "horrible" (-4.48 +/- .15) to "bad" (-1.89 +/- .43), "good" (2.20 +/- .33), and "perfect" (3.91 +/- .44). Mean intensity ratings ranged from "good," (4.21 +/- .44), the least intense, to "bad" (5.04 +/- .42), "perfect" (7.92 +/- .64), and "horrible" (8.21 +/- .49).

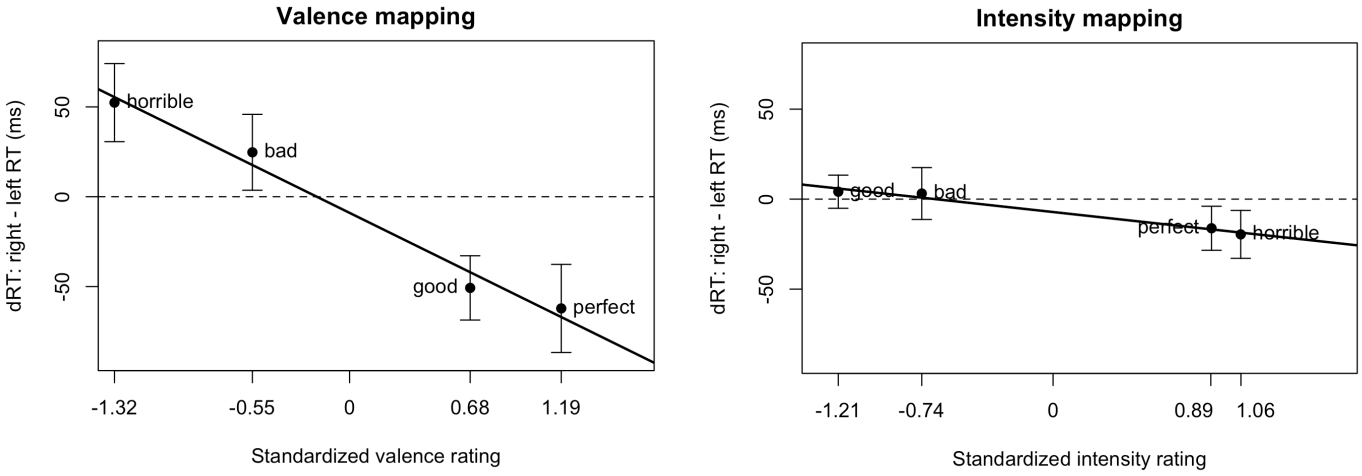


Figure 1. Results of Experiment 1. Left: Significant Valence Mapping in the valence task. Right: Non-significant Intensity Mapping in the intensity task. Error bars show the standard error of the mean.

Accuracy Four subjects failed to follow instructions and were replaced. The error rate was higher in the valence task (5.6%) than in the intensity task (4.3%; $\chi^2(1)=6.08$, p=.01). Inaccurate trials were excluded from the RT analyses.

RTs: Valence Judgments RTs greater than 2.5 standard deviations from subject means were removed, (2.69% of accurate responses). To evaluate whether participants implicitly spatialized the four words according to their valence, we adopted the technique typically used to evaluate the SNARC effect (e.g. Fias, Brysbaert, Geypens & d'Ydewalle, 1996), like H&L. For each participant and word, we calculated the RT difference between hands (dRT=right - left hand RT) and regressed these dRTs over the mean valence rating of each word. The slope of the resulting regression line provides a continuous index of each participant's Valence Mapping. On average, participants in the valence task associated the negatively valenced words with left space and the positively valenced words with right space, as indexed by the negative slope (M=-14.59 ms/unit valence; t(15)=-2.49, p=.02; Figure 1, left), demonstrating a clear Valence Mapping.

RTs: Intensity Judgments RTs greater than 2.5 standard deviations from subject means were removed (2.52% of accurate responses). Analogous to our analysis of the Valence Mapping, we regressed participants' dRTs over the mean intensity rating of each word. Although the relationship between intensity and dRT showed a slight trend in the direction predicted by an Intensity Mapping, this trend did not approach significance (M=-6.13 ms/unit intensity; t(15)=-1.03, p=.32; Figure 1, right). Participants did not show an Intensity Mapping, even though the task required them to make explicit judgments of emotional intensity.

RTs: Comparison of Valence and Intensity Effects To test the difference between the significant Valence Mapping observed in the valence task and the nonsignificant Intensity Mapping observed in the intensity task, we used a linear mixed-effects regression. RTs were predicted by response hand, standardized valence and intensity ratings, and task, with random slopes and intercepts for participants. The 3-way interaction between response hand, ratings, and task was marginally significant ($\chi^2(1)=2.93$, p=.09).

Experiment 2: A second test of Valence and Intensity Mappings

In Experiment 1, we predicted that if a Valence Mapping and an Intensity Mapping are both available to participants in long-term memory, then participants would (i.) implicitly spatialize valence when making speeded valence judgments (consistent with de la Vega, et al., 2012; de la Vega et al., 2013; Kong, 2013), and (ii.) implicitly spatialize emotional intensity when making speeded intensity judgments (consistent with the conclusions of Holmes & Lourenco, 2011). Only the first prediction was upheld. The goal of Experiment 2 was to test the same predictions with greater power, generalizability, and sensitivity.

To increase power, we doubled the number of participants and octupled the number of items; increasing the sample of items from 4 to 32 also increased the generalizability of the results. To increase sensitivity, we collected word ratings from each participant, after the RT task, and predicted dRTs on the basis of these subject-specific word ratings.

Method

Participants Sixty-four right-handed adults from the University of Chicago community participated for payment or course credit. Half were randomly assigned to the valence task (n=32) and the other half to the intensity task (n=32).

Materials The set of 4 emotional words from Experiment 1 was expanded to include 32 words for which valence (positive, negative) was crossed with emotional intensity (high intensity, low intensity): *beloved, brilliant, capable, courteous, determined, devastated, disliked, drowsy, ecstatic, energized, exhausted, good, gorgeous, gregarious, hated, hesitant, hideous, horrible, idiotic, ignorant, insulting, obstinate, perfect, pleased, prepared, rested, unattractive, unhappy*.

Procedure The procedure was identical to that of Experiment 1 except for the following changes. Participants performed 32 practice trials at the start of each block (one for each word), after which these words were presented in random order 5 times, composing 320 critical trials over two blocks. After this speeded RT task, participants rated the valence and intensity of each word. Words appeared on the computer screen one at a time in randomized order and participants spoke aloud their numerical ratings, on the same scales used in Experiment 1. The order of the valence rating and intensity rating tasks was counterbalanced across participants.

Results

Valence judgements Trials in which no response was given (0.55%) were removed as were RTs greater than 2.5 standard deviations from subject means (3.2% of responses).

We regressed each participant's mean dRTs over their standardized valence ratings for each word. As in Experiment 1, participants in the valence task associated the negatively valenced words with left space and the positively valenced words with right space, as indexed by a negative slope (M=-10.92 ms/unit valence; t(31)=2.60, p=.01), demonstrating a clear Valence Mapping.

Intensity judgments Trials in which no response was given were removed (1.30%) as were RTs greater than 2.5 standard deviations from subject means (2.9% of responses). We regressed each participant's mean dRTs over their standardized intensity ratings for each word. We found no systematic association between intensity and side of space (M=5.58 ms/unit intensity; t(31)=.54, p=.59). The mean slope did not differ significantly from a slope of zero, but notably its sign was *positive*, inconsistent with the predicted Intensity Mapping.

Comparison of valence and intensity effects To test the difference between the significant Valence Mapping observed in the valence task and the nonsignificant Intensity Mapping observed in the intensity task, we used a mixed-effects model with random slopes and intercepts for

participants and words. The 3-way interaction between response hand, standardized word ratings, and task was significant, indicating that the Valence Mapping was reliably stronger than the Intensity Mapping ($\chi^2(1)=7.82$, p=.005).

Discussion

The results of Experiment 2 replicated the results of Experiment 1, extending them to a larger sample of participants and a larger set of words: we found a significant Valence Mapping but no significant Intensity Mapping. Experiment 2 also supports an inference that was only weakly supported in Experiment 1. Although the effect of valence on dRT was clearly significant, and the effect of intensity on dRT was clearly nonsignificant, the difference between these two effects was only marginal in Experiment 1. This between-task difference was significant in Experiment 2, indicating that the Valence Mapping was statistically stronger than the (nonsignificant) Intensity Mapping.

In summary, the results of Experiments 1 and 2 showed a highly significant Valence Mapping, consistent with previous results (Casasanto, 2014, for review). They contrast with H&L's results, however, which were interpreted as showing an Intensity Mapping. In Experiment 3, we tested a possible explanation for this apparent contradiction.

Experiment 3: Emotional or spatial magnitude?

Why did H&L find an Intensity Mapping whereas we did not? A potential answer comes from an examination of H&L's stimuli. For the two experiments in which H&L reported an Intensity Mapping (their Experiments 1 and 2), the stimuli were photographs of six actors making facial expressions that varied in their emotional intensity (what H&L called emotional magnitude): neutral, happy, very happy, extremely happy, etc. When classifying the gender of these faces using left- and right-hand response keys, participants showed a pattern of RTs that was consistent with an Intensity Mapping.

However, as acknowledged by H&L, these findings might also be "due to specific facial features that varied across the range of stimuli, rather than the magnitude of emotional expression per se" (p. 318). One such feature was the "size of the mouth; more happy faces had a larger mouth opening than less happy faces" (p. 318).

Although H&L dismiss this confound between emotional intensity and physical space, it is of concern for two reasons. First, across cultures, people tend to focus on the mouths of emotional faces, as much or more than other parts of the face such as the eyes (Jack et al., 2009; Koda & Ruttkay, 2014; cf, Blais et al., 2008), and this tendency is particularly strong in US participants (Jack, Caldara & Shyns, 2012; Yuki, Maddeux & Masuda, 2007). Second, and crucially, previous data show that people implicitly spatialize *area* along a left-to-right continuum. In one experiment, participants were faster to make speeded judgments on smaller circles with the left hand and larger circles with the right hand (Ren, Nicholls, Ma & Chen, 2011), producing a dRT effect much like H&L's.

Given that (a) people implicitly associate smaller spatial areas with the left and larger areas with the right, and (b) H&L's more intense faces appear to have larger mouths, it seems possible that the emotional magnitude effect H&L reported was, in fact, a *spatial magnitude* effect. To test this alternative explanation, here we calculated the mouth areas of the faces in H&L's stimulus set and used them in place of emotional intensity to predict dRTs in H&L's Experiment 2, the experiment that provided the strongest evidence for an Intensity Mapping. (We discuss alternative explanations for the results of H&L's Experiments 1 and 3 in the General Discussion.)

Method

H&L's Experiment 2 used 30 faces from the NimStim database (Tottenham, et al., 2009). The photographs depicted six actors making each of five emotional expressions, which H&L labeled as neutral, happy, angry, extremely happy, and extremely angry. A coder blind to the experimental hypotheses used Adobe Photoshop to select and measure the area of the mouth (including the lips) of each face.

Results and discussion

Analysis of Mouth Areas After the mouth areas were measured, measurements were averaged for each emotional expression (reported in units of 1000 pixels): Neutral (M=3.27; +/- .27); Happy (M=3.64; +/- .40); Angry (M=2.26 +/- .20); Extremely happy (M=10.53; +/- .75); Extremely angry (M=6.75 +/- .85). A first analysis confirmed that the mean mouth area varied significantly across the emotional expressions (F(1, 28)=36.3, p=.000000004; Figure 2).

A second analysis tested whether mouth area covaried with emotional intensity. For each image, the mouth area (from our measurements) was used to predict the intensity rating (i.e., "emotional magnitude" ratings, obtained from H&L). Mouth area reliably predicted emotional intensity ratings (F(1, 28)=17.9 p=.0002).

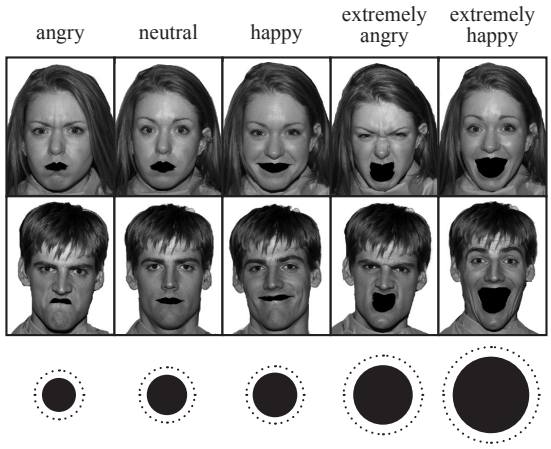


Figure 2. Top: Examples of face stimuli used in H&L Experiment 2, with measured mouth area highlighted in black. Bottom: Relative mouth area (black circles) and standard errors (dotted lines) by emotional expression.

Predicting dRT from Mouth Area The finding of a correlation between mouth area and emotional intensity confirms that spatial magnitude could, in principle, explain an Intensity Mapping. To test whether mouth area can, in fact, explain H&L's effects, we conducted a further analysis, analogous to H&L's, using mouth area instead of emotional magnitude ratings to predict dRTs. For each participant and image, dRTs (obtained from H&L) were regressed over standardized mouth area (from our measurements) to produce a spatial magnitude slope for each participant. The mean slope differed significantly from zero (M=-10.14 ms/unit area; $t(45)=2.72$, $p=.009$), suggesting an implicit mapping of spatial magnitude (i.e. mouth size) onto left-right space (a Spatial Magnitude Mapping). Moreover, when we controlled for mouth area, emotional magnitude slopes no longer differed from zero ($M = -10.48$ ms/unit intensity; $t(45) = 1.40$, $p = .17$); the Intensity Mapping disappeared. When we controlled for emotional magnitude, the effect of mouth area also no longer differed from zero ($M = -3.31$ ms/unit intensity; $t(45)=0.73$, $p=.47$).

These findings show that the effects observed in H&L's Experiment 2 can be explained by the spatial magnitude of a salient aspect of their stimuli. That is, participants may have been spatializing emotional faces either according to their relative emotional intensity or according to the relative size of their mouths; the data cannot distinguish between these two alternatives.

General Discussion

In three experiments, we investigated how people spatialize different aspects of emotion. We sought to reconcile an apparent contradiction between previous studies showing an implicit left-right mapping of emotional valence (Casasanto, 2014, for review) with the results of a study positing an implicit left-right mapping of emotional intensity (Holmes & Lourenco, 2011). In Experiments 1 and 2, we tested whether people can implicitly spatialize emotion according to either valence or intensity, depending on the context. When judging the emotional valence of words, participants reliably showed the Valence Mapping typical of right-handers. By contrast, when judging emotional intensity, participants showed no evidence of an Intensity Mapping, inconsistent with the findings of H&L. Experiment 3 provided a possible explanation for this discrepancy. In H&L's stimuli, emotional intensity was confounded with spatial magnitude: Faces expressing greater emotional intensity also had larger mouths. Mouth area reliably predicted the pattern of RTs reported by H&L, and controlling for mouth area eliminated the effect of emotional intensity. Together, these experiments (a) extend the evidence for an implicit lateral mapping of emotional valence, (b) provide no evidence for the implicit lateral mapping of emotional intensity posited by H&L, and (c) provide an alternative explanation for the RT effects that H&L interpreted as evidence for an Intensity Mapping.

Is there any evidence for an Intensity Mapping?

In light of our findings, what is the evidence that people implicitly spatialize emotional intensity on a left-right continuum? Currently, the evidence for this claim rests on a paper by H&L, whose three experiments we address in turn.

In H&L's Experiment 1, participants classified faces whose emotional expressions varied from neutral to extremely happy. As acknowledged by H&L, this experiment cannot, in principle, distinguish between a Valence Mapping and an Intensity Mapping, since both mappings make the same predictions for positive emotions.

In H&L's Experiment 2, the findings are consistent with an Intensity Mapping and inconsistent with a Valence Mapping. However, here we show that these findings are also consistent with a Spatial Magnitude Mapping, in which smaller area (i.e. mouth size) is associated with the left and larger area is associated with the right (see our Experiment 3). To the extent that participants were spatializing mouth size rather than emotional intensity or valence, the results of H&L's Experiment 2 do not bear on the spatialization of any dimension of emotion.

Although H&L interpreted the results of their third experiment as evidence of an Intensity Mapping, the data do not support this interpretation. Participants in H&L's Experiment 3 classified faces on either happiness (happy/not happy) or anger (angry/not angry). The pattern of RTs depended on the type of judgment participants made: those who judged happiness associated the angriest faces with the left and the happiest faces with the right; those who judged anger showed the opposite mapping, associating the happiest faces with the left and the angriest faces with the right.

Although H&L interpreted these two patterns as evidence of the flexibility of an "emotional magnitude line," both patterns are inconsistent with any linear mapping of emotional magnitude (i.e., intensity) onto left-right space. Whereas an Intensity Mapping predicts that people should associate the angriest and happiest faces with the same side of space, participants in H&L's Experiment 3 mapped extremely happy and extremely angry faces onto opposite sides of space. These results are, therefore, incompatible with the proposed emotional magnitude line.

Instead, these findings may be readily explained by *polarity alignment* (Proctor & Cho, 2006), also called *markedness* (Clark 1969; Dolscheid & Casasanto, 2015; Lakens, 2012; Lynott & Coventry, 2014). Like many other analog continuums in language and mind, *happiness* and *anger* are both bipolar (or marked) continuums. That is, they consist of a +polar (or unmarked) endpoint (*happy*, *angry*) and an opposing –polar (or marked) endpoint (*unhappy*, *not angry*). Lateral space is also a bipolar continuum, where *right* has +polarity and *left* has –polarity, at least for right-handers (see Huber et al., 2014).

In polarity alignment effects, participants show faster RTs when the poles of two continuums align than when they misalign (see Clark, 1969; Proctor & Cho, 2006). This is exactly the effect that H&L observed in their Experiment 3: participants responded faster when the poles of the designated emotion aligned with the poles of the lateralized response (*happy/angry: right; unhappy/not angry: left*) than when they misaligned (*happy/angry: left; unhappy/not*

angry: right). The results of H&L's Experiment 3, therefore, provide no evidence for an Intensity Mapping, but rather show what appears to be a classic polarity alignment effect (Clark, 1969; Proctor & Cho, 2006).

In sum, none of the three experiments presented by H&L provide any clear evidence for an Intensity Mapping (although, in principle, future studies could reveal such a mapping, in some contexts). Instead, previous results may reflect emotional valence (in H&L's Experiment 1), spatial features of the stimuli (i.e., mouth size in H&L's Experiment 2), and polarity alignment (in H&L's Experiment 3).

Conclusions

In two experiments, we show that people implicitly map emotional valence onto left-right space, but we find no spatial mapping of emotional intensity. In a third experiment, we provide an alternative explanation for previous data interpreted as support for an Intensity Mapping. The idea of an Intensity Mapping was motivated by the search for a generalized magnitude system for representing multiple prothetic domains (Holmes & Lourenco, 2011). The Valence Mapping shown here provides no evidence for such a generalized magnitude system because valence is not a prothetic domain (e.g., "happy" is not *more valence* than "sad"). Rather, the present results extend the evidence for the broadly generalized tendency to spatialize our abstract concepts, whether or not they are prothetic, according to the specifics of our physical and social experiences.

Acknowledgments

Research supported by a James S. McDonnell Foundation Scholar Award (#220020236) and NSF grant (BCS #125710) to DC. Thanks to Wim Gevers and Katia Carbé for helpful discussions, to Katia Carbé and Nikolai Maximay for help with data collection, to Geoffrey Brookshire, Laura Staum Casasanto, Tom Gijssels, and Ché Lucero for discussion of the analyses, and to Kevin Holmes and Stella Lourenco for their collegiality and for sharing their data.

References

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS One*, *3*(8).

Brunyé, T. T., Gardony, A., Mahoney, C. R., & Taylor, H. A. (2012). Body-specific representations of spatial location. *Cognition*, *123*(2), 229-239.

Carbè & Gevers (2013). *Investigating the spatial representation of magnitude information and emotional valence concepts.* Paper presented at The 32nd European Workshop on Cognitive Neuropsychology, Bressanone.

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right- and left-handers. *Journal of Experimental Psychology: General, 138*(3), 351-367.

Casasanto, D. (2014). Bodily relativity. In L. Shapiro (Ed.), *Routledge handbook of embodied cognition* (pp. 108-117). New York: Routledge.

Casasanto, D., & Chrysikou, E.G. (2011). When left is "right": Motor fluency shapes abstract concepts. *Psychological Science, 22*(4), 419-422.

Casasanto, D., & Henetz, T. (2012). Handedness shapes children's abstract concepts. *Cognitive Science, 36,* 359–372.

Casasanto, D., & Jasmin, K. (2010). Good and bad in the hands of politicians: Spontaneous gestures during positive and negative speech. *PLoS ONE, 5*(7).

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, *76*(4), 387.

de la Fuente, J., Casasanto, D., Román, A., & Santiago, J. (2015). Can culture influence body-specific associations between space and valence? *Cognitive Science*, *39*, 821-832.

de la Vega, I., De Filippis, M., Lachmair, M., Dudschig, C., & Kaup, B. (2012). Emotional valence and physical space: limits of interaction. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(2), 375.

de la Vega, I., Dudschig, C., De Filippis, M., Lachmair, M., & Kaup, B. (2013). Keep your hands crossed: The valence-by-left/right interaction is related to hand, not side, in an incongruent hand–response key assignment. *Acta Psychologica*, *142*(2), 273-277.

Dolscheid, S., & Casasanto, D. (2015). Spatial congruity effects reveal metaphorical thinking, not polarity correspondence. *Frontiers in Psychology*, *6*(1836), 1-11.

Holmes, K. J., & Lourenco, S. F. (2011). Common spatial organization of number and emotional expression: A mental magnitude line. *Brain and Cognition*, *77*(2), 315-323.

Huber, S., Klein, E., Graf, M., Nuerk, H. C., Moeller, K., & Willmes, K. (2014). Embodied markedness of parity? Examining handedness effects on parity judgments. *Psychological Research*, 1-15.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology*, *19*(18), 1543-1548.

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, *141*(1), 19.

Koda, T., & Ruttkay, Z. (2014). Eloquence of eyes and mouth of virtual agents: cultural study of facial expression perception. *AI & Society*, 1-8.

Kong, F. (2013). Space–valence associations depend on handedness: evidence from a bimanual output task. *Psychological Research*, *77*(6), 773-779.

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 726.

Lynott, D., & Coventry, K. (2014). On the ups and downs of emotion: testing between conceptual-metaphor and polarity accounts of emotional valence–spatial location interactions. *Psychonomic Bull. & Review*, *21*(1), 218-226.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237-241.

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416.

Ren, P., Nicholls, M. E., Ma, Y. Y., & Chen, L. (2011). Size matters: non-numerical magnitude affects the spatial coding of response. *PLoS One*, *6*(8), e23553.

Tottenham, N., et al. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, *168*(3), 242-249.

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, *43*(2), 303-311.

Essays about service-learning events can be mined for program assessment

Anne T. Gilman (gilman@juniata.edu) & Deborah W. Roney (roneyd@juniata.edu)
Juniata College Department of Psychology & Language in Motion, 1700 Moore Street
Huntingdon, PA 16652 USA

Victoria Rehr (vrehr@cscc.edu)
Columbus State Community College Writing Center, 102 Columbus Hall, 550 East Spring Street
Columbus, OH 43215 USA

Helen H. Hu (huh162@psu.edu)
PSU Department of Educational Psychology, Counseling, & Special Education, 125E CEDAR Building
University Park, PA 16802 USA

Mark P. Peterson (mark.phillip.peterson@gmail.com)
Viterbo Universty Department of Biology and Mathematics, 900 Viterbo Drive
La Crosse, WI 54601 USA

Abstract

Psychological applications of human language technology combined with multidisciplinary approaches to similarity calculations and data visualization offer avenues to broaden the use of students' own words in program assessment. We compared multiple analysis approaches on both simple token counts (word roots and character trigrams) and top-down language indicators from 85 student essays about service-learning events. Bioinformatic distance calculations on word root counts provided useable assessment information on attitude change, showing patterns of word use that match the holistic goals of the assignment. Although these patterns were not found in a subsequent batch of 81 essays, the tools we are providing may facilitate other efforts to detect attitude change in student writing about service-learning events.

Keywords: events; semantic similarity; LIWC; text mining

Discovering attitude changes in writing about events

Writing about emotion-laden topics, including events such as starting college, is associated with gains in school success and even physical health (Pennebaker & Chung, 2011). Instructors routinely assess students' content learning and skill development by examining students' assigned writing about events (Reynolds, Livingston, Willson, & Willson, 2010), although questions remain about interpretating these assessment practices (Baker, O'Neil, & Linn, 1993). Institutions of higher learning are facing increasing pressure to provide documentation of changes in knowledge, skill, and attitude in their students (Ewell, 2009). Measuring changes in cultural attitudes is an assessment area with even less consensus than measuring writing skill development. Some measures rely on a few self-reported rating-scale items (Pascarella, Wolniak, Seifert, Cruce, & Blaich, 2005). Other measures that avoid self-report and claim to indicate implicit attitudes can be gamed (Marini, Rubichi, & Sartori, 2012) and should be interpreted with caution (Gawronski, LeBel, & Peters, 2007). Student writing constitutes a greater volume than most professors can assess for learning goals beyond the grading criteria for each assignment. Although this volume remains tiny in comparison to the language data used in computational linguistics research, could some computational techniques help document attitude changes and avoid the need for additional assessment activities? Student privacy protection and variation among assignment types stand in the way of producing a useable large-scale corpus of this type of writing. In this study, we used students' reflections on service-learning events to explore the potential for computational linguistics to supplement other evalutations of student learning.

How computers detect patterns in texts

Enormous growth (Lyman & Varian, 2003) in available, machine-readable text data continues to foster the development of tools for detecting patterns in a much greater volume of writing than a single instructor could feasibly read (Bender & Good, 2010), These tools count written words, phrases, and/or pieces of words in a given text and apply different calcuations using those counts to compare that text to others. For example, "happy birthday to you" is more similar by word count to "you are too happy" than to "many happy returns". A knowledge-added approach might have the phrases "happy birthday" and "happy returns" listed in the same category and thus identify those as more similar to each other than to "too happy". With longer and more numerous texts, the calculations used to measure text similarity–usually expressed as a *distance*—become more complex. The calculations that group documents by topic draw on different text characteristics than those that can detect an author's emotional outlook (Pennebaker, Mehl, & Niederhoffer, 2003), mental health status (Resnik, Garron, & Resnik, 2013), or personal identity (Koppel & Schler, 2003). In this study, we compared several similarity calculations, drawing from bioinformatics as well as computational linguistics, to discover common patterns in essays about service-learning events.

Contrasting computational techniques

A key challenge in finding similar patterns in a group of natural language texts is the sparseness of the distribution of individual words. Among one hundred essays about a service trip

abroad, these terms might each occur three times: *undoubtedly*, *counterrevolutionary*, *twin*. How should similarity calculations treat these terms compared to extremely common ones such as *and* or *the*?

Some approaches divide words into short sequences of letters such as the three-letter trigrams used as part of our calculation; others rely on a dictionary to remove endings and count word roots, so that *revolutionary* and *revolutionaries* count as the same term. Individual words, letter sequences, and word roots are called tokens when used this way. Other approaches add even more knowledge by grouping terms by their part of speech (e.g. noun, verb) and even their emotional connotations (e.g. *glad*, *happy*). Since character trigrams have been employed to allow for errors from faulty optical character recognition in scanned documents (Faber, Hochberg, Kelly, Thomas, & White, 1994), they may aid the processing of documents authored by native and non-native speakers. On the other hand, completely bottom-up approaches often fall short of the mark in natural language processing (Chang & Su, 2004), so we also used top-down semantic and rough part-of-speech information via the Linguistic Inquiry and Word Count (LIWC) tool (Francis & Pennebaker, 1993; Pennebaker et al., 2003), as a comparison facilitated by bioinformatic distance measures.

Applying bioinformatic analysis to student writing

Biology, like computational linguistics, is seeing enormous growth in new tools to find patterns in the rapidly-increasing volumes of data, for example (Buonaccorsi et al., 2014; Sboner, Mu, Greenbaum, Auerbach, & Gerstein, 2011). The focus of bioinformatics in particular on deep analysis of often small numbers of unique samples, makes it a promising source of tools and analogies for processing natural language texts. Many bioinformatic tools converge on the analysis of count data, such as the number of copies of a gene or the number of a bacterial species in a sample. While computational linguistics methods are commonly applied to biomedical research (Ananiadou & McNaught, 2006), applying bioinformatic techniques to language data is still in its infancy. Applying these tools to natural language data stands to increase the reach of natural language processing innovations for text-rich social and behavioral sciences.

Summary

In this paper we present ongoing work to evaluate the utility of low-knowledge and expert-informed computational linguistics tools with and without the application of bioinformatic models for educational program assessment. In addition to new findings, we have developed a small R package to ease the entry of researchers to automated textual analysis. The `wordcountWrapper` package is available at `https://bitbucket.org/petersmp/wordcountwrapper` including source code and a description of the included tools.

Method

Language in Motion (LIM)

As a service-learning program, Language in Motion (LIM) brings college students with extensive knowledge of foreign cultures into rural K-12 classes. For more than a decade, the LIM program has formally documented the professional development gains of participating K-12 teachers and their students. Now, program leaders face growing pressure to label and document the effects on the college students involved: international students, multilingual students, and those returning from study abroad programs. Assessments of LIM's impact for participating K-12 sites do not address program goals such as attitude and perspective changes among presenters.

A bottom-up approach to assessment

The source texts for this project are essays written by students to discuss what they learned from one or more educational outreach events. Each text author (all of whom are college students) made several language and culture presentations to K-12 students, typically featuring a language other than English and stories of the presenter's experiences in another country.

Some program goals, such as practice using the featured language, are comparatively easy to measure independently of these essays; measuring the attainment of goals such as gains in cultural awareness is more difficult. The program administrator (DR), a coauthor and our designated expert, looks for evidence in the essays of progress towards goals such as students learning about themselves and gaining a fresh perspective on their culture of origin. Since expert evaluation of written texts does not map transparently onto assessment criteria that outside reviewers can use, several text-mining techniques were compared to see which most closely approached the expert's holistic essay ratings.

Each presenter submits a narrative evaluation of their experience presenting to younger students, and these essays, averaging 821 words per student, offer the opportunity for automated knowledge extraction. For this study, 85 presenter essays from 2008–2012 were analyzed using computational approaches to increase the depth of evaluation analysis and reporting, including identifying effects not yet articulated in the stated mission of LIM and offering avenues to document qualitative observations.

Data processing

Each essay was converted to an anonymous plain-text document. To remove potentially identifiable (i.e. unique) place names from the corpus and standardize the target culutural information, foreign country names and languages were replaced with *ZZTopia* and *ZZTopian*.

The LIM director rated 22 randomly chosen essays for evidence of learning from their program participation. Our analyses compared those essays rated as showing 'Excellent' success (8) in large program goals with all others (14). Unrated

essays, omitted from clustering analyses, provide useful context for our application of the results found here and will allow for follow-up analysis of the value added by our clustering approaches.

Token counts

Character trigrams and word roots were extracted from each student essay using R (R Core Team, 2013). The functions to read, count, and provide context for these analyses are available from the authors as an R package (`https://bitbucket.org/petersmp/wordcountwrapper`). Punctuation and non-standard characters were discarded, and all numbers were all replaced with a single *7*. Root words were identified using the package `SnowballC` (Bouchet-Valat, 2013). Character trigrams were identified and counted, without spaces between words, using the package `tau` (Buchta, Hornik, Feinerer, & Meyer, 2014). Analysis of whole words was similar to word roots, and trigram analysis including spaces was similar to those without (data not shown).

Standard dimensional reduction

Since word root and trigram data is sparsely distributed, dimensional reduction is needed to make comparisons between texts feasible. To do this, we initially used principal component analysis (PCA), linear discriminant analysis (LDA), and k-means clustering to analyze frequency data for trigrams, word roots, and LIWC categories. Token frequencies within each essay were used rather than raw counts to avoid confounds with differences in essay length, as 'Excellent' essays were approximately twice as long as others (1,269 vs 674 mean words, $t(11.425) = 6.78, p<0.0001$). Each method was then compared to expert ratings to determine their value for future program evaluation.

We compared PCA scores (for any component explaining at least five percent of the variance) between ratings, using a t-test, to determine if any of the dimensions separated our groups. To test the value of the LDA discrimination, we performed a jackknife analysis, serially omitting a single rated sample, and then determining if the LDA test accurately placed it after training on the remaining data. K-means clusters were compared to expert ratings using Cohen's kappa statistic (Cohen, 1960) implemented in the R package `fmsb` (Nakazawa, 2014) to determine if rated groups could be reproduced.

Distance measures from biology

Many methods have been developed in ecology to assess the relative abundance of flora and fauna between divergent sites (Leinster & Cobbold, 2012). One of these measures, the Horn similarity index (Horn, 1966), is now being used in bioinformatics due to its ability to accurately and efficiently handle large numbers of items and samples while accounting for variation in total sampling depth (i.e. length of each essay). We calculated Horn similarities for all pairs of samples using the R package `rnaseqWrapper` (Peterson, Malloy, Buonaccorsi, & Marden, 2015).

Horn similarities were then converted to dissimilarity measures and supplied to the non-metric multidimensional scaling function in the R package `vegan` (Oksanen et al., 2013). Briefly, this calculated the two-dimensional representation of the data that most accurately recreates the pair-wise distances to allow plotting of the spatial relationship between samples.

Differential token usage

To determine which tokens were used differentially by 'Excellent' and other essays, we used the R packages `DESeq` (Anders & Huber, 2010) accessed via `rnaseqWrapper` (Peterson et al., 2015). DESeq uses a negative binomial test on count data to determine whether or not two groups (i.e. 'Excellent' and other essays) have differential representation of a given token (i.e. word root, trigram, or LIWC category).

Designed for gene expression data, DESeq accurately models dispersion across samples by comparing similarly used words to increase the accuracy of the test statistic. However, because it is designed for large scale sequencing projects, the test statistic, particularly its false discovery rate corrected q-value, are likely to be overly conservative for this analysis. Therefore, we report uncorrected p-values but note that these results should be interpreted with caution.

Results

Standard dimensional reduction

Conventional dimensional reduction methods failed to provide any insight into the differences between 'excellent' and other essays for word roots and trigrams, but identified some patterns in LIWC categories. Jackknifed LDA prediction of sample ratings was non-significantly worse than chance for word roots and trigrams, and better than chance for LIWC categories (Fisher's exact test, $p = 0.19$, 0.66, and 0.19).

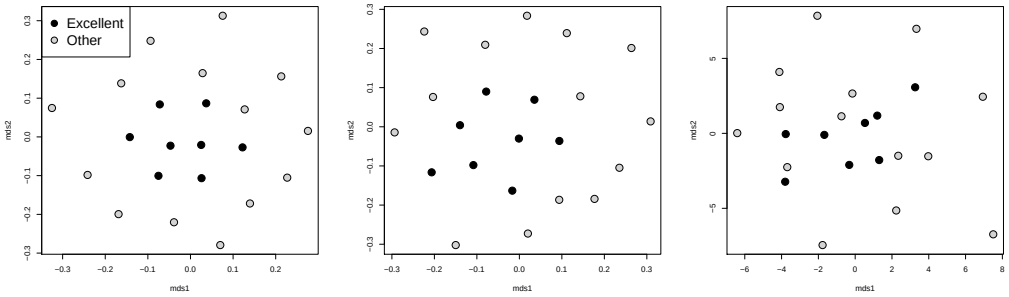
PCA yielded 5 components in word roots, 7 components in trigrams, and 6 components in LIWC categories that each explained at least five percent of variance (for a total of 57%, 55%, and 83% of variance, respectively). None of these differed between 'Excellent' and other ratings (t-test, all $p > 0.05$). For k-means with two clusters, Cohen's kappa was 0.35, 0.05, and 0.21 for word roots, trigrams, and LIWC ($p =$ 0.06, 0.72, and 0.35), respectively.

Distance measures

Horn distance and non-metric dimensional scaling visualization revealed a centralized cluster of essays rated as 'excellent,' with a dispersion of other essays (Figure 1).

Differential token usage

Between 'Excellent' and other essays, 25 trigrams, 5 root words, and 4 LIWC categories significantly differed in usage (Figure 2). Essays rated as excellent used *both* and *look* more frequently than the others, which in turn had higher occurrences of *didn't*, *American*, and words beginning in *cours*. LIWC analysis showed that the top-rated essays used more



(a) Trigrams (b) Word Roots (c) LIWC Categories

Figure 1: Non-metric scaling of pair-wise Horn similarity indices.

negation, in spite of using *didn't* less, and more perception-related terms. Less-successful essays had more longer words and social terms.

Discussion

We developed a toolkit that enables knowledge extraction from collections of free text where traditional clustering approaches fail. This toolkit delineated student accomplishment of stated but hard-to-measure goals in a language outreach program and promises insights for a range of other educational objectives. Of particular note, the addition of two bioinformatic tools, Horn similarities and DESeq, substantially improved these analyses. These tools revealed clustering that were unlikely to be identified by traditional approaches and identified several of the tokens that characterized high-quality essays.

Clustering

Traditional approaches to dimensional reduction and clustering failed to accurately identify groups, but the addition of Horn similarites suggests why this may be the case. PCA, LDA, and k-means clustering attempt to separate groups along a single axis at a time. The distance measures in the non-metric scaling, however, demonstrated that 'Excellent' essays are instead clustering in the middle.

Collecting more essays might reveal separable, dispersed clusters. Cluster dispersion can indicate mention or absence of specific LIM goals, e.g. changing perspective on one's own culture, or simply reflect an extraneous detail such as a specific food name. That is, addition of more sample essays may allow the identification of a single central cluster (the 'Excellent's here) along with multiple edge clusters, each sharing some common set distinguishing it from the high quality central group. The central distribution of 'Excellent' essays, however, impedes discovery of these dispersed clusters.

Furthermore, this centrality may suggest that 'Excellent' essays share a core vocabulary. This could mean that: *a*) Excellents are not including extraneous information, particularly about their assigned country, and/or *b*) Excellents are touching on more of the core concepts from the program. That is, the clusters of other essays may diverge from the central cluster of "Excellent" essays because they are focused on a single culture (and those cultures may cluster) rather than the broader program goals and/or they may focus on a smaller portion of the goals of the program (and cluster by the portion of the program they cover). This type of analysis would, in part, require breaking the anonymity of the analysis and use subject matter expertise to explore those possibilities.

Specific visualization features for large data sets influence what knowledge viewers gain and remember (Ware, Gilman, & Bobrow, 2008). The bioinformatic visualization tools applied here allowed for more meaningful knowledge extraction from a very limited text corpus.

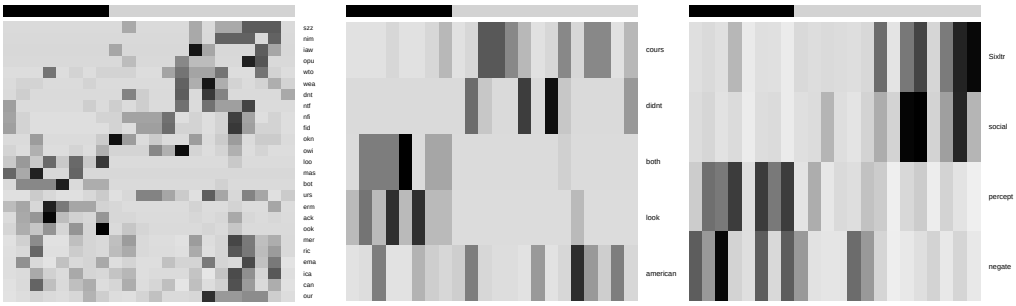
Differential word usage

In addition to this general pattern, we identified specific words and trigrams that are used differently by writers of 'Excellent' versus other essays. In particular, 'Excellent' essays never used the contraction *didn't*, while other ratings used it on average 1.5 times per essay, suggesting that other ratings used less-polished language.

In addition, 'Excellents' used the word root *American* less than other ratings (1.13 vs 3.94 times per essay; similar rates for its component trigrams). This trend continued for focus on the student's country: 'Excellent' essays used the word root *ZZTopia* (an anonymizer for assigned country) less than other ratings (5.61 vs. 8.94 times per essay), though this difference was not statistically significant (p= 0.31). Together, these differences strongly imply that students that are most completely meeting the stated goals of LIM are those that are focusing less on a specific country and more on the cross-cultural goals of the program.

Boundary conditions for LIWC

LIWC analysis did not improve the clustering of essays, though a few categories differed between 'Excellent' and other essays. LIWC can be confounded by departing from the original expressive writing prompt (Hu, Koestler, Stroup, & Gilman, 2013). The present findings confirm that the LIWC tool is distinct from low-knowledge word- and character-based approaches, at the same time demonstrating additional boundary conditions for LIWC.



(a) Trigrams (b) Word Roots (c) LIWC Categories

Figure 2: Heatmap of token usage by rating. For each essay (column), usage of each token (row) significantly differentially used (by DESeq), with usage z-scale transformed by row (darker is higher usage). 'Excellent' essays are to the left in each panel, denoted by the black bars above the heatmap.

No single assessment meets all needs

The educator who provided expert ratings on LIM essays aimed to spend no more than several minutes per essay to arrive at a broad rating of the writer's success in learning from LIM involvement. This rating was specifically not intended to follow from the grade the paper might receive considered only as a piece of writing. Mining assignment text for students' attitude changes should provide different results than expert ratings (Lee, Pincombe, & Welsh, 2005). The tools presented here were applied to a comparison set of 113 essays from British students of anthropology and sociology (Nesi, Gardner, Thompson, & Wickens, 2007), fields where perspective on one's own and others' culture matters. No similar clustering patterns were found according to paper grade or education stage, suggesting that our results are not based on writing quality but on learning from a described event.

Automated essay grading can invite attempts to game such systems (Perelman, 2008). The present study seeks new ways to describe holistic educational outcomes discovered in participants' words—a very different aim than that of grading.

We put 81 LIM essays (23 of which were rated) from 2013-2016 through `wordcountWrapper`. Although the patterns described here were not directly replicated, it is important to note that many of the essays were written by students that went through the program *after* the initial analyses described here were completed. It is possible that changes made in response to insights developed with these tools are impacting the depth and quality of student experiences. In particular, looking at the context of key tokens such as *ZZTopia* revealed that the more recent papers were describing difficult topics about the featured country rather than merely listing topics. The context tools in this package make such comparisons feasible even for hundreds of essays.

Conclusions

Here, we have demonstrated the utility of applying bioinformatic analysis and visualization to text mining for program assessment. Lacking a consensus measure for changes in perspective accompanying service-learning activities, and aware that self-report and implicit measures (including this one) can be gamed, we mined student-authored texts to supplement (not replace) the administrator's assessment data without adding another instrument. Like Paquette, de Carvalho, and Baker (2014), we see this project based on interactions with one expert as a starting point. The R package provides tools to aid novices to NLP analysis the ability to rapidly create substantial count data from a number of formats. The output count tables are intended to be ready to use in any of several analysis and visualization tools, including the ones shown here. In addition, the package has a function to provide context information (i.e., surrounding words) of any token of interest to aid exploratory analyses. Combined, these tools may aid in increasing quantitative reporting to agencies on the effectiveness of difficult-to-assess program objectives across higher education.

Acknowledgments

Portions of this work were supported a summer grant to ATG by the James J. Lakso Endowment for Faculty Excellence and funding to MPP from Penn State University Huck Institues of the Life Sciences and Howard Hughes Medical Institute.

References

Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine*. Boston, MA, USA: Artech House.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106. Retrieved from `http://genomebiology.com/2010/11/10/R106/` doi: 10.1186/gb-2010-11-10-r106

Baker, E. L., O'Neil, H. F., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210.

Bender, E., & Good, J. (2010). A grand challenge for linguistics: Scaling up and integrating models. In *Social, behavioral, & economic sciences 2020.* National Science Foundation (USA). (White Paper)

Bouchet-Valat, M. (2013). Snowballc: Snowball stemmers based on the c libstemmer utf-8 library [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=SnowballC (R package version 0.5)

Buchta, C., Hornik, K., Feinerer, I., & Meyer, D. (2014). tau: Text analysis utilities [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=tau (R package version 0.0-16)

Buonaccorsi, V., Peterson, M., Lamendella, R., Newman, J., Trun, N., Tobin, T., ... Roberts, W. (2014). Vision and change through the genome consortium for active teaching using next-generation sequencing (gcat-seek). *CBE-Life Sciences Education*, *13*(1), 1–2.

Chang, J.-S., & Su, K.-Y. (2004). Pitfalls in applying unsupervised learning to NLP. In *IJCNLP-04*. Hainan Island, China.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *37*.

Ewell, P. T. (2009). *Assessment, accountability, and improvement: Revisiting the tensions*. National Institute for Learning Outcomes Assessment (USA). (Occasional Paper No. 1)

Faber, V., Hochberg, J. G., Kelly, P. M., Thomas, T. R., & White, J. M. (1994). Concept extraction: A data-mining technique. *Los Alamos Science*, *22*, 123–149.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic inquiry and word count* (Tech. Rep.). University of Texas at Austin.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, *2*(2), 181-193. doi: 10.1111/j.1745-6916.2007.00036.x

Horn, H. S. (1966). Measurement of 'overlap' in comparative ecological studies. *American Naturalist*, *100*, 419–424.

Hu, H., Koestler, A., Stroup, S., & Gilman, A. (2013). Comparing text characteristics of expressive and values writing. In *Annual meeting of the Eastern Psychological Association.* New York, USA.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of ijcai'03 workshop on computational approaches to style analysis and synthesis* (Vol. 69, p. 72).

Lee, M. D., Pincombe, B., & Welsh, M. (2005). A comparison of machine measures of text document similarity with human judgments. In *27th annual meeting of the cognitive science society (cogsci2005)* (pp. 1254–1259).

Leinster, T., & Cobbold, C. A. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3), 477–489.

Lyman, P., & Varian, H. R. (2003). How much information? 2003. *University of California, Berkeley, California, USA*. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003

Marini, M., Rubichi, S., & Sartori, G. (2012). The role of self-involvement in shifting IAT effects. *Experimental Psychology*, *59*(6), 348–354. doi: 10.1027/1618-3169/a000163

Nakazawa, M. (2014). fmsb: Functions for medical statistics book with some demographic data [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=fmsb (R package version 0.4.3)

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2007). The British Academic Written English (BAWE) corpus, developed at the Universities of Warwick.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., ... Wagner, H. (2013). vegan: Community ecology package [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=vegan (R package version 2.0-10)

Paquette, L., de Carvalho, A. M., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. In *Proceedings of the 36th Annual Cognitive Science Conference* (pp. 1126–1131).

Pascarella, E. T., Wolniak, G. C., Seifert, T. A., Cruce, T. M., & Blaich, C. F. (2005). Liberal arts colleges and liberal arts education: New evidence on impacts. *ASHE Higher Education Report*, *31*, 1–168.

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. *Oxford handbook of health psychology*, 417–437.

Pennebaker, J. W., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577.

Perelman, L. (2008). Information illiteracy and mass market writing assessments. *College Composition and Communication*, *60*, 128–141.

Peterson, M., Malloy, J., Buonaccorsi, V., & Marden, J. (2015). Teaching rnaseq at undergraduate institutions: A tutorial and r package from the genome consortium for active teaching. *CourceSource*.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (p. 1348-1353). Stroudsburg, PA, USA: Association for Computational Linguistics.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*. Pearson Education International.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology*, *12*(8), 125.

Ware, C., Gilman, A. T., & Bobrow, R. J. (2008). Visual thinking with an interactive diagram. In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrams* (Vol. 5223, p. 118-126). Springer.

Language Informativity: Is starfish more of a fish in English than in Dutch?

Farah M. Djalal (farahmutiasari.djalal@kuleuven.be)
Wouter Voorspoels (wouter.voorspoels@kuleuven.be)
Tom Heyman (tom.heyman@kuleuven.be)
Gert Storms (gert.storms@kuleuven.be)

Faculty of Psychology and Educational Sciences, University of Leuven,
Tiensestraat 102, 3000 Leuven, Belgium

Abstract

Two studies examined how lexical information contained in words affects people's category representations. Some words are lexically suggestive regarding the taxonomic position of their referent (e.g., bumble*bee*, star*fish*). However, this information differs from language to language (e.g., in Dutch the equivalent words hold no taxonomic information: *hommel*, *vlinder*). Three language groups, Dutch, English, and Indonesian speakers, were tested in similarity and typicality judgment tasks. The results show that the lexical information affects only the users of the language (e.g., Dutch speakers rated Dutch-informative items, both in similarity and typicality tasks, higher than English and Indonesian speakers). Results are discussed in light of theories of concept representation and the language relativity hypothesis.

Keywords: lexical; similarity; typicality; cognitive; concepts

Introduction

In concept research, the meaning of a word is generally related to the properties of the (category of) objects it refers to. These properties cover a broad range, such as physical, contextual and taxonomic properties. This information is activated when people have need of the word meaning, for instance when they have to interpret the word in a sentence, or make concept-dependent judgements, such as similarity (how similar are salmon and trout) or categorization decisions (is this novel object a fish?). A number of theories propose that categorization follows a feature-based approach, meaning that people tend to put objects in the same category the more properties they have in common (e.g., Clark, 1973; Nelson, 1974; Mervis, 1987). For instance, an object that grows in soil and has branches and leaves is called a tree. Consequently, all objects that share the same features will be called a tree as well.

Language, however, is generally not considered merely a system of signifiers that map on classes of objects with certain (physical) properties. Indeed, it also entails a particular way of carving up the world and manners of thinking about reality, subtly encoding cultural knowledge, and metaphors that describe ways of viewing natural phenomena (e.g., Lakoff & Johnson, 1980). It is often claimed that, when learning a language, one does not only learn what the different words in a language refer to, but also these more subtle aspects. In other words, learning a particular language may have particular effects on thinking about and perceiving the world (e.g., Casasanto, 2016; Lucy, 2014).

The idea that language shapes thought and perception has some interesting consequences, which become tangible when considering the myriad of languages that can be found in the world. According to language relativity, as two languages are structurally different, their respective speakers should differ in how they think, act and perceive in objectively similar situations. This hypothesis has been examined and confirmed in a number of cognitive domains. The effects on cognition of particular manners of classification have been documented in domains such as color (e.g., Thierry, Athanasopoulos, Wiggett, Dering, & Kuipers, 2009), causation (e.g., Fausey & Boroditsky 2011), and space and motion (e.g., Slobin, 1996). Also, it has been shown that grammatical categories such as gender and tense have effects on cognitive tasks (e.g., Boroditsky, Schmidt, & Phillips, 2003), and language specific metaphors that describe abstract domains such as time, musical pitch and mathematical and number terminology have been shown influential in cognitive tasks in these domains (e.g., Casasanto, 2008; Dolscheid, Shayan, Majid, & Casasanto, 2013).

Considering the evidence available at the moment, the question is not so much whether language can influence thought, but how, in which domains and to what extent this is the case. In the present study, we examine whether the representation of categories in a taxonomically structured domain are under influence of the vocabulary that a language provides for the domains. In some cases, the name that a language provides for a category of objects sometimes holds information regarding its position in the taxonomic structure. For example sunflower is a type of flower, a bumblebee is closely related to bees, a goldfish is a type of fish and a blackbird is a type of bird. Clearly, in English, the names are more than just arbitrary signifiers, containing what we will refer to as "lexical information": Information regarding the referent of a word that follows from the word itself, generally because the name is a combination of constituents of which at least one has a meaning in the language. The obvious question from a language relativistic point of view then is whether the availability of lexical information in a name influences the representation of the category it refers to. To the extent that languages differ in

terms of lexical information, one may find representational differences for everyday categories such as bumblebees or cauliflowers[1]. From the perspective of the dominant theories in concept representation (e.g., exemplar theory, Smith & Medin, 1981; family resemblance, Rosch & Mervis, 1975; prototype theory, Hampton, 1995), we do not expect category labels to have any influence in how the category is represented. Although some theories consider the labels to be features that matter (e.g., the rational model of categorization; Anderson, 1991), the labels are considered arbitrary, and thus their meaning is not taken into account. To examine whether a word's lexical information influences the representation of the object category it refers to, we look at objects for which one language has a head noun that is informative as to the object's position in a broader taxonomy, whereas the equivalent in another language does not entail such information. The word bumblebee, for instance, is informative in English, but not in Dutch, where no reference to bee is made in the word (*hommel*) or in Bahasa (i.e., the Indonesian language; *kumbang*). Conversely, *inktvis* (squid) is suggestive in Dutch (i.e., the word *vis* in Dutch means fish) but not in English (squid) or Bahasa (*cumi-cumi*).

If the representation of object categories is influenced by lexical information, we expect that people's judgments regarding the objects will subtly vary across language groups in a systematic way, particularly when the judgments rely on the representations of the categories. For example, as starfish in English contains a category suggestion, but not in Dutch, we expect English participants to judge a starfish to be more similar to a goldfish than Dutch participants, as starfish in Dutch does not hold the same category suggestion (*zeester*). Moreover, we expect English participants to judge starfish as more typical for the fish category than Dutch participants. Similarity and typicality are fundamental notions in concept representation research, and directly related to how people represent classes of objects. As such, they form a perfect arena to test whether lexical information influences representation.

Study 1: Similarity

In the first study, we examined how lexical information included in object names can influence people's judgments of similarity between two objects that share lexical category information in their name (e.g., the Dutch pair *walvis* and *inktvis*; in English, respectively, whale and squid) and whether the three language groups, English, Dutch, and Indonesian speakers, give higher similarity ratings for pairs of items that are lexically informative in their own language.

Method

Participants Sixty English speakers (44 females and 16 males, mean age: 21 years and 10 months), 74 Dutch speakers (59 females and 15 males, mean age: 18 years and 3 months), and 60 Indonesian speakers (33 females and 27 males, mean age: 24 years and 6 months) were tested. Two Indonesian participants were excluded from the analyses as there was no variability in their answers. The Dutch speakers were students who got credits in exchange for participation, the Indonesian participants were students who live and study in Indonesia, and the English speakers were recruited online using Amazon Mechanical Turk.

Material[2] A list of 70 questions was presented in a web survey. Each question contained a pair of words that shared the same lexical information in one of the three languages. In 16 word pairs the words were informative in Dutch but not in English nor in Bahasa (e.g., *wasbeer* and *beer* [raccoon and bear]; *aardappel* and *sinaasappel* [potato and orange]), 16 word pairs were informative in Bahasa but not in English nor in Dutch (e.g., *ikan hiu* and *ikan pari* [shark and stingrays]; *burung hantu* and *burung kakaktua* [owl and cockatoo]), and 16 word pairs were informative in English (e.g., bumblebee and bee; catfish and jellyfish), but not in Dutch nor Bahasa. As fillers, eight word pairs were informative in all three languages and 14 word pairs were uninformative in all the three languages.

Procedure Dutch and Indonesian participants were given a link to an online survey, whereas the English speakers participated via Mechanical Turk. In the survey, each word pair was presented in the format: "How similar are X and Y?". Participants answered on a 10-point rating scale ranging from 1 (not at all similar) to 10 (extremely similar). The survey took about 5 to 7 minutes. All items were translated into the three languages and all three groups of participants received the survey in their own languages Thus, with the exception of the filler items, items that were informative in one language were uninformative in the other two languages. Each participant was randomly assigned to one of four randomized orders of the items sets.

Results and Discussion

First, the consistency of the similarity judgments within each language group was computed using the split-half method combined with the Spearman-Brown formula. The reliability estimates for English, Dutch, and Indonesian speakers were, respectively, .80, .89, and .90, indicating a high agreement among participants of the same language group in the similarity judgment task. Next, we computed average similarity scores by first calculating by-participant

[1] Obviously, a cauliflower is not a flower. Some names provide useful taxonomic information, but sometimes the information can be misleading. Another example: the word whale in Dutch (*walvis*) wrongly implies that the animal belongs in the fish category. We will consider both cases.

[2] In a pilot study, different participants from the three language groups were asked to judge the familiarity of all the objects' names on a 10-point scale. The results revealed that participants were relatively familiar with all items.

conditional means and then collapsing across participants (see the upper panel of Figure 1).

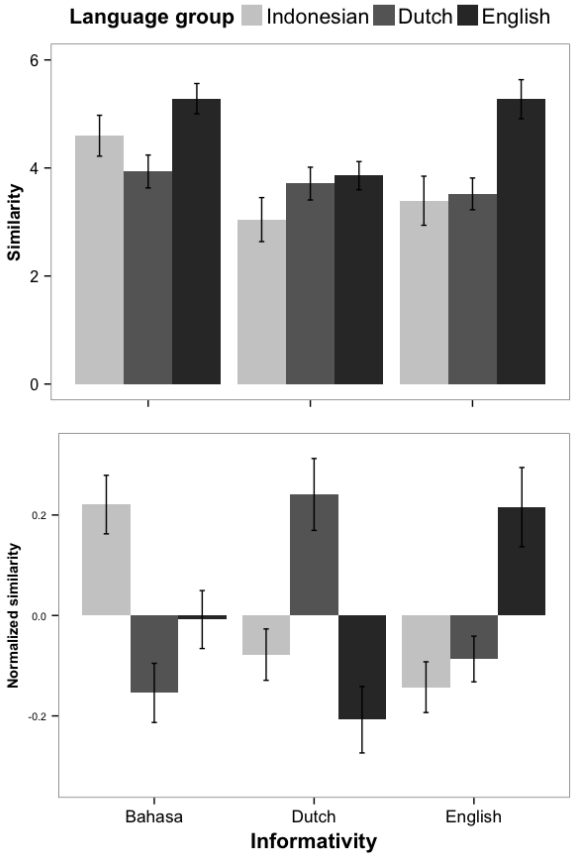


Figure 1: Averaged raw (upper panel) and z-transformed (lower panel) similarity scores in each condition with 95% confidence intervals. The z-transformed values reflect the effect of lexical informativity, controlled for differences between language groups and item groups.

In order to test the hypothesis that people's similarity judgments are influenced by the lexical information in the objects' names, mixed effect analyses were performed on the similarity judgment scores. In the analyses, the language group (English, Dutch, and Indonesian speakers) and the language informativity were included as fixed effects, whereas participants and items were included as random effects such that a maximal random structure was created (Barr, Levy, Scheepers, & Tilly, 2013). The crucial test of the linguistic relativity hypothesis concerns the interaction between language group and language informativity: If lexical information influences the similarity judgments, we expect that two categories are judged more similar by speakers of the language in which the categories' names are informative. To test for this interaction, a model that includes the interaction term was compared with a model that does not. The analyses were carried out in R (version 3.1.2) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). As predicted, the analyses revealed a significant language group × language informativity interaction ($\chi^2(4) = 34.49$, $p < .001$), suggesting that similarity judgments can be influenced by lexical information. However, the interpretation of the interaction is not that straightforward, because baseline similarity scores differed in the three language groups. This is best exemplified by a separate analyses of the filler data, which showed a strong effect of language group ($\chi^2(2) = 26.60$, $p < .001$). Remember, fillers were either uninformative or informative in all three languages, so this finding entails that the three language groups have diverging baselines.

To aid the interpretation of the language group × language informativity interaction, we transformed the similarities into z-scores for every participant separately, after which by-participant conditional means were calculated. Collapsing across participants then yields the average similarity scores shown in the lower panel of Figure 1 (some are negative as a result of the by-participant z-transformation). Figure 1 suggests that items informative in a certain language receive relatively higher similarity ratings from the speakers of that language. Additional contrast analyses performed on the untransformed similarity data confirmed this ($\beta = 0.86$, $SE = 0.26$, $Z = 3.25$, $p = .001$, for Bahasa; $\beta = 1.17$, $SE = 0.25$, $Z = 4.66$, $p < .001$, for Dutch; and $\beta = 1.07$, $SE = 0.23$, $Z = 4.73$, $p < .001$, for English)[3]. These results are consistent with the hypothesis that each language group judges similarity higher for item pairs that are lexically informative in their own language.

In general, it is assumed that people judge similarity based on features that objects have in common. However, this study shows that lexical information may also affect similarity judgments between two objects, even when it concerns objects that are rather perceptual in nature. Note that the participants were most likely not aware of the purpose of the study, as nothing in the instructions hinted at the research question of interest, and the item list contained a relatively large number of fillers. This finding suggests that the particular language, and the implicit cultural knowledge it carries, can influence people's judgments.

Study 2: Typicality

In Study 2, we examine whether lexical information that is included in objects' names could affect people's typicality judgments. Typicality refers to the graded membership structure of concepts, and is considered a crucial variable in natural language concept research (see, e.g., Rosch & Mervis, 1975). Similar to Study 1, items that are informative either in English, Dutch or Bahasa were presented in order to investigate whether language users rate typicality higher for words that are informative in their own language.

[3] The statistical model is the same as in the previous analyses, except that Helmert coding was used to extract the relevant comparisons. Furthermore, to obtain p-values, we treated the *t*-statistic as a *z*-statistic following Barr and colleagues (2013, p. 266).The data and the scripts of the mixed effects analyses can be found on osf.io/tfqhw.

Method

Participants Sixty one English speakers (46 females and 15 males, mean age: 22 years and 4.5 months), 60 Dutch speakers (48 females and 12 males, mean age: 18 years and 5 months), and 67 Indonesian speakers (41 females and 25 males, mean age: 24 years and 8 months) participated in the second study. One participant was excluded since she was originally from Malaysia and spoke Malay as her mother tongue. None of them participated in Study 1, nor were they aware of the purpose of this study. The Dutch speakers were students who got credits in exchange for participation, the Indonesian participants were students who live and study in Indonesia, and the English speakers were recruited online using Mechanical Turk.

Materials A list of 46 questions was presented to the participants in a website survey. Six items were most informative in English (e.g., ladybird, eggplant, jellyfish), six items were most informative in Dutch (e.g., *kikkervisje, inktvis, stinkdier* [tadpole, squid, skunk]), and another six items were most informative in Bahasa (e.g., *ikan hiu, burung hantu, burung kakaktua* [shark, owl, cockatoo]). The remaining items were fillers, consisting of eight items that were informative in all three languages (English, Dutch, and Bahasa), and 20 items that were not informative in any of the three languages.

Procedure All questions were of the following format: "How typical is x for the category of X?". All informative items (x) were paired with the category name (X) that was included in the objects' name. For example: "How typical is a goldfish for the category of fish?". Participants were asked to answer on a 10-point scale rating, ranging for 1 (extremely atypical) to 10 (extremely typical). The survey took about 3 to 5 minutes. All participants were tested in their own language and received one out of three sets of questions, which only differed in the order of the items.

Results and Discussion

The consistency of the typicality judgments in each language group was computed using the same method as in Study 1 (i.e., split-half method combined with the Spearman-Brown formula). The results were, for English, Dutch, and Indonesian speakers, respectively, .91, .91, and .97. These results indicate a very high consistency between subjects in each language group (see the upper panel of Figure 2 for the average typicality scores in each condition). Mixed effects analyses were then performed on the typicality judgment scores in order to investigate whether lexical information included in objects' names could affect people's typicality judgments. The analyses were run in the exact same manner as in Study 1. Again, the effect of interest in this study, the interaction between language group and language informativity, was statistically significant ($\chi^2(4) = 13.08$, $p = .011$).

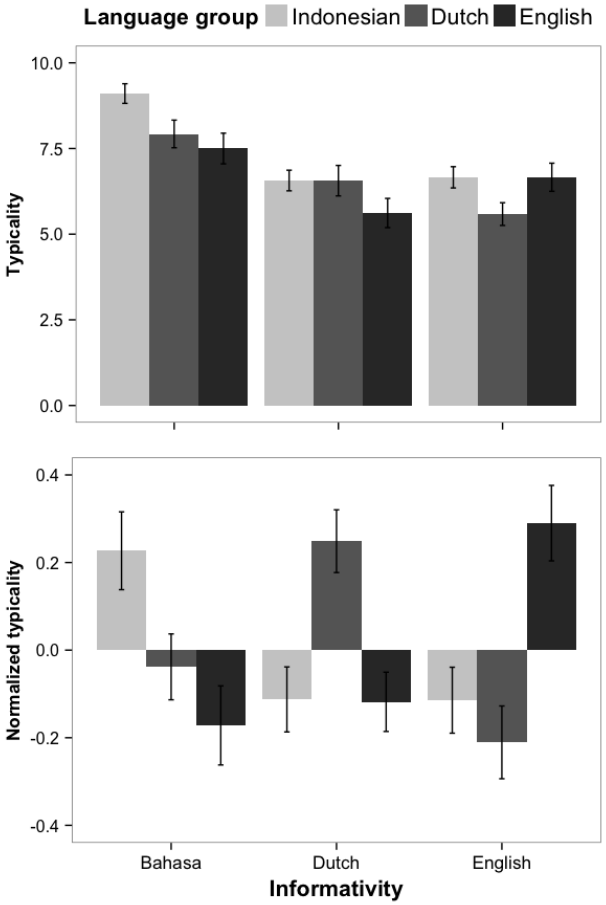


Figure 2: Averaged raw (upper panel) and z-transformed (lower panel) typicality scores in each condition with 95% confidence intervals. The z-transformed values reflect the effect of lexical informativity, controlled for differences between language groups and item groups.

As in Study 1, the interpretation of the interaction gets clouded by cross-cultural differences in baseline (typicality) judgements. More specifically, analyses of the filler data showed a main effect of language group once again ($\chi^2(2) = 7.96$, $p = .019$). Analogous to Study 1, we first transformed the typicality judgments into z-scores per participant, then we calculated by-participant conditional means, and finally we collapsed across participants to obtain the average typicality scores shown in the lower panel of Figure 2. The pattern of results looks very similar to those displayed in Figure 1. That is, items informative in a certain language receive relatively higher typicality ratings from the speakers of that language. Additional contrast analyses performed on the untransformed typicality data are in line with this, although the effect did not reach statistical significance for Bahasa ($\beta = 0.89$, $SE = 0.57$, $Z = 1.54$, $p = .123$, for Bahasa; $\beta = 1.19$, $SE = 0.49$, $Z = 2.42$, $p = .015$, for Dutch; and $\beta = 1.52$, $SE = 0.43$, $Z = 3.57$, $p < .001$, for English).

The results of Study 2 suggest that in all three language groups, participants rated typicality higher for items that were informative in their respective languages. Similar to

Study 1, these results are consistent with the hypothesis that people are influenced by lexical information in giving typicality judgments, even if sometimes the lexical information provided is misleading (e.g., starfish and *inktvis* [squid] rated as more typical members of category fish for, respectively, English and Dutch speakers). Again, it seems that not only non-lexical features are considered when making these judgments – as is the general assumption of models of concept representation – but also suggestive information that is available in the linguistic name of an object.

General Discussion

Two studies were conducted to investigate the influence of lexical information on semantic tasks related to taxonomic concept representation. In Study 1, evidence was found that participants rated similarity higher for item pairs that were informative in their language, as well as in Study 2, where the participants judged typicality higher when the items' name was suggestive of a particular taxonomic position in their mother tongue. Interestingly, these findings suggest that language has some influence on how categories are represented simply by means of the particular names that are used. Not only does this finding have consequences for models of concept representation and categorization, it is also interesting to consider it in light of the discussion concerning the influence language can have on thought.

More than features?

Both the similarity and typicality judgment task showed that even if the information contained in the objects' names was misleading, people are tempted to use this information to categorize the object. While this is an interesting finding as such, the obvious question is which cognitive processes underlie the observed effect of lexical information. Broadly speaking, the effect can be driven either by the representation that is influenced by the label, or the response processes underlying the judgment.

At a representational level, it is possible that some features of the informative constituent of the label are automatically transferred to the representation of the category the label refers to. For example, the representation of 'star fish' in English speakers may automatically include some features of fish, by virtue of the label. When making a judgment of similarity or typicality, these features, although not experientially acquired, will make the categories more similar, or typical: a starfish will be considered more typical of fish and more similar to goldfish because of these transferred features. Alternatively, it may be merely because of the label. According to the rational model of categorization, labels are just another feature, without any special status. Thus, according to this approach, a similarity or typicality judgment also relies on the label, and shared labels or partly shared labels are expected to influence the judgment.

The effect could also reside at the response level. Given that in both experiments, informative trials presented participants with two labels that share a constituent, it is possible that participants were influenced by the mere commonality in the labels. This explanation would imply that whatever the label, independent of it being informative or not, commonality between labels will influence judgments of typicality and similarity (for example, this would imply that English speakers judge a **bee**r and a **bee**hive more similar than Dutch speakers). While this is not our preferred interpretation, it cannot be refuted on the basis of the two experiments presented here.

To examine this hypothesis, an additional ongoing study is conducted using pairs of items in which an item with lexical information is paired with an item that is considered to be a typical member of the category mentioned in the name of the first item (e.g., salmon and jellyfish and salmon). As in the present studies, critical items are only informative in one language. In this way, the idea that people make an inference about the likely relation between the words (e.g., share the ending '*-fish'*) can be controlled.

Importantly, while we cannot conclude to either explanation, our results do not in any way contradict the basic idea that similarity drives categorization, nor that perceptual, contextual and relational features are important. In general, people do categorize objects based on these features and then compare them with the other members or the prototype of the category. However, relying on features – in the traditional sense – may not tell the complete story. As demonstrated in Djalal, Ameel, Heyman, and Storms (2016), there is no clear-cut correspondence between the generated features and the prediction of category membership, and thus it should not come as a surprise that sometimes information of a different nature is relied upon.

Language shapes thought?

The present study provides evidence for language specific characteristics being influential in two fundamental cognitive tasks that have been extensively shown to rely on concept representation. The leap to concluding that the labels in a language can influence meanings should not be made without care, however. For one, as explained in the previous section, the effect may reside at the response level.

Here, we consider another possible explanation that points to a particularly subtle way of how language can shape behaviour in certain tasks. Perhaps the effect of label informativity depends on uncertainty. When people cannot form a sufficiently detailed image of the object a word is referring to, they may rely on other sources of information, one of which is the knowledge encoded in the labels of the language. For example, when people are uncertain as to whether a squid is a fish (a boundary case at least for some people), Dutch speakers may indeed rely on the "cultural knowledge" present in their language (in which the word for squid makes reference to fish) as a source of information, whereas English and Bahasa speakers do not have this knowledge available. While most objects used in the present studies were relatively familiar to participants, this interpretation requires further examination.

While not a mere response effect, relying on language in this way is presumably not what is understood when theorizing about language and though. However, the question is whether the potential relation with an individual's knowledge would make the observed effect less interesting. Perhaps one of the most basic ways in which culture, and language, are influential in a person's behaviour and thought, is the mere fact that she can rely on knowledge that is encoded in the culture, without the need for first-hand experience.

In sum, the present findings in the domain of concept representation are consistent with the (abstract) hypothesis that language shapes thought, but more importantly, they point to a number of hypotheses as to the cognitive processes or representational differences involved in the behavioural effects.

Acknowledgments

Farah M. Djalal is sponsored by a graduate scholarship financed by the 'Bijzonder Onderzoeksfonds' of the University of Leuven. Tom Heyman is a research fellow of the Research Foundation-Flanders (FWO-Vlaanderen). Wouter Voorspoels is a post doc at KU Leuven. The data are stored and can be accessed in Open Science Framework (osf.io/tfqhw). Correspondence should be addressed to Farah M. Djalal, Department of Psychology, University of Leuven, Belgium. Email address: farahmutiasari.djalal@ kuleuven.be.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-429. doi: 10.1037/0033-295X.98.3.409.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi: 10.1016/j.jml.2012.11.001.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7). Retrieved from: http://CRAN.R-project.org/package=lme4.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought* (pp. 61-79). Cambridge, MA: MIT Press.

Casasanto, D. (2008). Who's afraid of the Big Bad Whorf? Cross-linguistic differences in temporal language and thought. *Language Learning*, *58*(1), 63-79. doi: 10.1111/j.1467-9922.2008.00462.x

Casasanto, D. (2016). Linguistic relativity. In N. Riemer (Ed.), *Routledge Handbook of Semantics* (pp. 158-174). New York: Routledge.

Clark, E. V. (1973). What's in a word? On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.). *Cognitive development and the acquisition of language*. New York: Academic Press.

Djalal, F. M., Ameel, E., Heyman, T., & Storms, G. (2016). *On the (mis)match of extension and intension in children's and adults' categories: Features do not tell the whole story*. Manuscript submitted for publication.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science 24*(5), 613-621. doi: 10.1177/0956797612457374

Fausey, C. M. & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review 18*(1), 150-157. doi: 10.3758/s13423-010-0021-5

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*(5), 686-708. doi: 10.1006/jmla.1995.1031.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. Cognitive Science, 4(2), 195-208. doi: 10.1207/s15516709cog0402_4

Lucy, J. A. (2014). Methodological approaches in the study of linguistic relativity. In L. Filipović & M. Pütz (Eds.), *Multilingual Cognition and Language Use: Processing and Typological Perspectives* (pp. 17-44). Amsterdam: John Benjamins.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorization* (pp. 201-233). Cambridge University Press.

Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review, 81*(4), 267-285. doi: 10.1037/h0036592.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605. doi: 10.1016/0010-0285(75)90024-9.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz and S. C. Levinson (Eds.) *Rethinking Linguistic Relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, MA: Harvard University Press.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on pre-attentive color perception. *Proceedings of the National Academy of Sciences*, *106*(11), 4567-4570. doi: 10.1073/pnas.0811155106.

A Developmental Shift in the Relationship Between Sequential Learning, Executive Function, and Language Ability as Revealed by Event-Related Potentials

Joanne A. Deocampo (jdeocampo@gsu.edu)
Department of Psychology, Georgia State University, P. O. Box 5010
Atlanta, GA 30302-5010 USA

Christopher M. Conway (cconway@gsu.edu)
Department of Psychology, Georgia State University, P. O. Box 5010
Atlanta, GA 30302-5010 USA

Abstract

Previous research has shown a link between sequential learning (SL) and language as well as links between executive function (EF) and both language and SL. However, little research has focused on both the development of the relationship between these factors and their neurological underpinnings. Here we report a study of the event-related potential (ERP) correlates of SL and behavioral measures of language and EF in a sample of 7-12-year-old children. Results revealed that both SL and EF had independent associations with language development but that the contribution that both made toward language development shifted dramatically between the ages of 7 to 11-12 years. The results furthermore suggest that this developmental shift may be due in part to the maturation of EF abilities and changes due to neural entrenchment and commitment as a consequence of language acquisition.

Keywords: language development; sequential learning; statistical learning; executive function; event-related potentials (ERP)

Introduction

The ability to encode statistical structure in temporally ordered sequences and make predictions about the world based on that structure is referred to as structured sequence processing or sequential learning (SL). There is a growing body of evidence suggesting that SL is an important mechanism underlying spoken language acquisition (e.g., Conway, Bauernschmidt, Huang, & Pisoni, 2010). For example, infants and adults can both learn artificial spoken languages in the laboratory based only on transitional probabilities (e.g., Romberg & Saffran, 2013, adults; Saffran, Aslin, & Newport, 1996, infants). Auditory statistical learning tasks can also uniquely predict adults' ability to comprehend English sentences (e.g., Misyak & Christiansen, 2012). In addition, several studies have linked individual differences in visual SL (VSL) to individual differences in various aspects of language, including, for example, adults' incidental learning of a grammar in a sequential memory task correlating with their ability to predict words in a spoken sentence (Conway, et al. 2010), individual differences in children's VSL independently predicting their ability to comprehend English syntax (Kidd & Arciuli, 2015), and VSL abilities of children with cochlear implants correlating with their performance on language measures (Conway, Pisoni, Anaya, Karpicke, & Henning, 2011). One recent study found that infants' VSL predicted later vocabulary and gesture comprehension (Shafto, Conway, Field, and Houston, 2012). In addition, there is some neural evidence showing a connection between SL and language. For instance, Christiansen, Conway, and Onnis (2012) found that an event-related potential (ERP) component that is typically considered an index of syntactic processing in natural language was elicited by incongruities in both an SL task and a natural language processing task in adults.

Recent research also suggests strong positive relationships between executive function (EF) and both SL (e.g., Bahlmann, Korb, Gratton, & Friederici, 2012) and language (e.g., January, Trueswell, & Thompson-Schill, 2009). Because SL is the ability to make predictions about the world based on statistical structure in temporally ordered sequences, it makes sense that some EF mechanisms, such as attention, cognitive control, cognitive flexibility, working memory, and inhibition, would contribute to SL ability in the same way that it does to language ability. In addition, as with SL and language, there is some evidence for an overlap in the neural mechanisms underlying SL and EF (Bahlmann et al., 2012).

However, there is very little research that addresses the developmental trajectory of the relationship between SL, EF, and language ability, especially in combination with neural measures. In light of this, we aimed to examine the relationship between SL as measured by ERPs, EF, and spoken language ability throughout a 5-year period in childhood. We chose the age range 7-12 years for two reasons: 1) there has been relatively less research on SL in middle childhood, with the bulk of the research having been done on infancy through preschool and adolescence through adulthood, and 2) as brought up in the Discussion, 7 years and 12 years appear to be ages at which there are major changes in the development of both SL and language (Janacsek, Fiser, & Nemeth, 2012; Johnson & Newport, 1989). We hypothesized that because EF skills appear to contribute to both SL and language abilities and SL is an important component of language development, EF and SL would both directly predict language ability, but that EF would also affect language through the mediator of SL. In addition, we hypothesized that because of the protracted nature of EF's development (e.g., Best & Miller, 2010),

these relationships would differ at various ages, suggesting age as a moderator (see Figure 1).

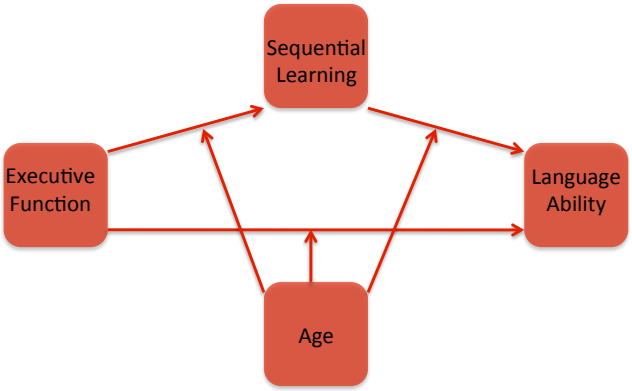


Figure 1: Conceptual model of the proposed relationships between EF, SL, language, and age.

Method

Participants

Thirty-eight typically developing and hearing monolingual English-speaking children between the ages of 7 and 12 years (M = 9.13 years, SD = 1.9 years; 15 female) participated.

Sequential Learning Task

Children were told a story about a magician who tried to make food for his children using his magic hat. Participants were told to "catch" the sporadically presented food by pressing a button. Children then viewed a stream of stimuli consisting of hats of different colors presented one at a time (for each: 500ms stimulus, 500ms black screen; SOA: 1000ms). Occasionally, a target hat with food depicted above it was presented within the stream. Unbeknownst to participants, hats of three different colors each differentially predicted the occurrence of the target hat, which we refer to as high-probability predictors, low-probability predictors, and standards. When the high-probability (HP) predictor was presented, it was immediately followed by the target 90% of the time and the standard 10% of the time. The low-probability (LP) predictor was followed by the target 20% of the time and the standard 80% of the time. In addition, the target was occasionally presented directly after a standard without a preceding predictor (no-predictor, NP). Figure 2 shows a schematic of the sequential learning task.

Based on previous results using a similar task (Jost, Conway, Purdy, Walk, & Hendricks, 2015), it was expected that if children learned the transitional probabilities between the predictors and the target, there should be differences in both response times (RTs) to the targets and ERPs to the predictors based on whether a trial was a HP, LP, or NP trial. These differences would constitute evidence of SL. Although RT findings supported the ERP findings, to conserve space, they will not be discussed in this paper.

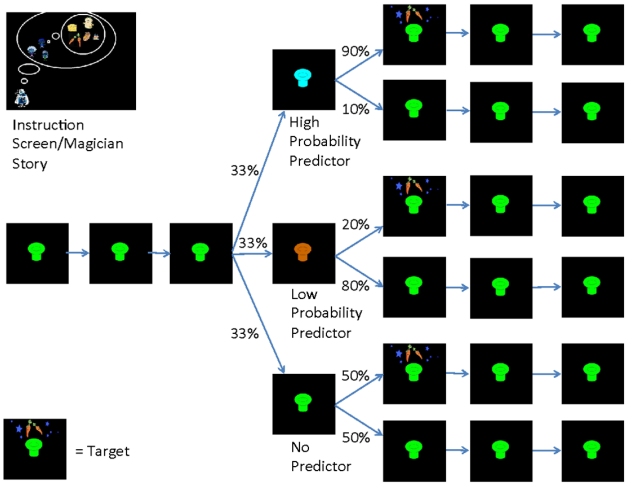


Figure 2: Schematic representation of the sequential learning task. The target followed the high predictor on 90% of HP trials but only on 20% of LP trials. In the NP condition, the target was presented immediately after a standard with no preceding predictor. A random number of standards were presented before each predictor (or NP)..

ERP Recording and Analysis

ERPs were collected using a 32-channel sensor net and preprocessed using Net Station Version 4.3.1 (Electrical Geodesics, Inc.). ERPs were time-locked to the onset of each predictor stimulus or in the case of the NP condition, the standard that preceded the target during the SL task. This resulted in 60 trials for each of the three predictor conditions (HP, LP, and NP).

The remainder of processing was done using custom scripts in MATLAB (version R2012b 8.0.0783; MathWorks) and the EEGLAB Toolbox (version 10.2.2.24a, Delorme & Makeig; 2004) for MATLAB to remove artifacts and replace bad channels. Participants were required to have a minimum of 20 good epochs per condition in each half of the SL task to be included in further analyses.

Executive Function Task

Executive function (EF) was assessed with a version of the Eriksen Flanker Task, which is thought to tap into selective attention, conflict response, and response inhibition mechanisms, all of which fall under the umbrella of EF. Horizontal arrays of arrows were presented on a computer screen, and children were told to respond only to the arrow in the center. They were to indicate whether the arrow was pointing left or right while ignoring the flanking arrows which could be facing the same direction (congruent) or opposite direction (incongruent). Response times to incongruent and congruent trials were recorded separately and average incongruent RTs subtracted from average congruent RTs to give a Flanker score (a higher score indicates higher EF).

Language Assessment

We assessed children's language ability with the Sentence Completion subtest of the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 1999). In this test, sentences are read without their final word, and children are asked to give a semantically and grammatically correct single word ending. We chose this assessment because it parallels a sentence prediction task that was positively associated with adults' SL (Conway et al., 2010).

Results

ERP Analyses

Based on the ERPs to the SL task described in Jost et al. (2015), we chose to analyze the ERP time window 400-700 ms post-stimulus presentation from the medial posterior region of sensor net. This is the time window and region in which Jost et al. (2015) found ERP effects in a very similar paradigm and with children of similar ages. Visual inspection of the grand averaged ERP waveforms (Figure 3) suggests that there was a P300-like positivity in the posterior medial region within a similar window as that found by Jost et al. (2015). This positivity was especially visible for the HP condition in the second half of the task (Figure 3B), consistent with the notion that learning the predictor-target contingencies occurred toward the end of the task, after sufficient exposure to the statistical probabilities (as was also observed in Jost et al., 2015). Furthermore, based on visual inspection, this positivity appears to consist of two different peaks where the predictor conditions appear to be differentiated: 300-600 ms and 600-750 ms post stimulus presentation. Therefore, we chose to analyze each of these time windows separately. Results for the 600-750ms window are presented here, although both windows yielded similar results. This 600-750ms window has the advantage of being within the time window analyzed by Jost et al. (2015) with children as well as within the window where Christiansen et al. (2012) found overlap between ERPs to a language task and an SL task.

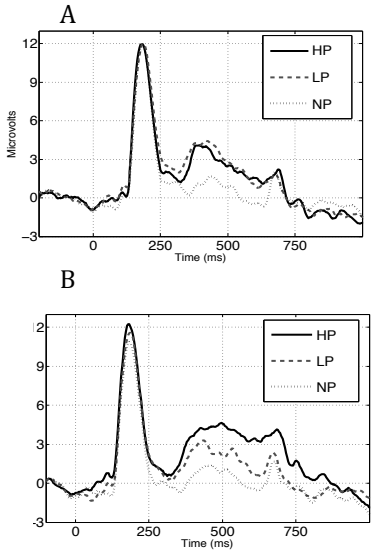


Figure 3: Grand-averaged ERP (n=38) in the medial posterior region to high-probability predictor (HP, solid line), low-probability predictor (LP, dashed line), and no-predictor (NP, dotted line) trials for first block (A) and second block (B) (Positivity upward in microVolts; time in milliseconds).

A 3 (predictor: high probability predictor, low probability predictor, or no predictor) x 2 (block: 1st half SL task or 2nd half SL task) repeated measures ANOVA was done on ERPs in the 600-750ms window to examine the effects of predictor type and block on the mean ERP amplitudes. Results revealed a significant main effect of predictor, $F(2, 74) = 7.67$, $p = .001$, $\eta_p^2 = .17$, and a significant interaction between predictor and block, $F(2, 74) = 5.08$, $p = .009$, $\eta_p^2 = .12$. The main effect of block was not significant. Predictor effect post hoc pairwise comparisons revealed that HP ERP amplitude was significantly greater than both LP ($p = .039$) and NP amplitudes ($p = .004$). For the interaction, post hoc pairwise comparisons showed that only HP increased across blocks ($p = .002$). All posthoc tests were Sidak corrected for multiple comparisons.

Conditional Process Model

To test the conceptual model of the relationship between EF, SL, language, and age depicted in Figure 1 in which EF has both direct and indirect effects (mediated by SL) on language and all relationships with language are moderated by age, we constructed a statistical conditional process model. First, we created composite SL scores to capture the level of learning across the task. First, the ERP amplitude difference of HP minus LP was calculated to capture the amount of differentiation between high and low probability predictors, which indicates a basic level of learning. Then, HP-LP for the 1st block was subtracted from HP-LP for the 2nd block to show change over time during the task, which indicates the amount that learning increased as the task progressed. We refer to this measure as "HP-LP change".

The conditional process model was tested using Hayes's (2013) PROCESS macro Model 59 with bootstrapping in SPSS. Bootstrapping was used to ameliorate the relatively small sample size. One additional participant was removed at this point because she met outlier criteria for the regression. We tested the model with HP-LP changes as the SL variable. For the EF measure, we used the Flanker score (congruent RT minus incongruent RT), and we used a continuous measure of age. Standard Sentence Completion score was the outcome measure.

Both predictive models, that between EF and SL with SL as the outcome and that predicting language (Sentence Completion) as the outcome from EF and SL, were significant (see Tables 1 and 2). SL does mediate the relationship between EF and language, but only with age moderating both the predictive link between EF to SL and the predictive link between SL to language. In addition, the EF measure, Flanker, significantly positively predicted language as measured by Sentence Completion. SL and age also positively predicted language (although SL only did so marginally). Age significantly moderated the relationships between EF and SL and between SL and language and marginally significantly moderated the relationship between EF and language, suggesting that the relationships between all of these constructs change with age. To visualize these changes in relationships that occur with age, we divided

participants into three age groups of approximately equal size, 7-year-olds (n = 12), 8-10-year-olds (n = 13), and 11-12-year-olds (n = 12), and produced scatter plots to show each relationship (see Figure 5).

Table 1: Regression table with SL (HP-LP change) as the outcome

Variable	*B*	*SE B*	*p*
EF (Flanker)	-.006	.004	.181
Age	.044	.426	.917
EF x Age	-.005	.002	.016
R^2	.263		.017

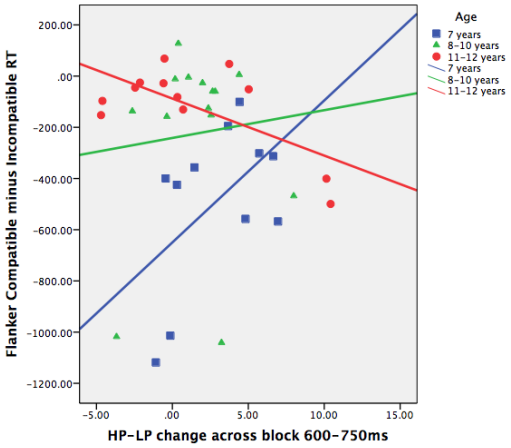
Table 2: Regression table with language (Sentence Completion) as the outcome

Variable	*B*	*SE B*	*p*
SL (HP-LP change)	1.27	.739	.096
EF (Flanker)	.047	.010	.0001
Age	-3.82	1.54	.019
SL x Age	.985	.329	.005
EF x Age	.013	.007	.064
R^2	.371		.0000

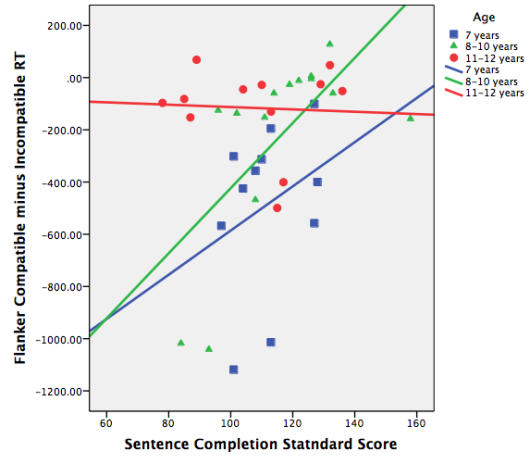
Inspection of the scatter plots reveals that all three relationships are quite different for older (11-12 years) versus younger (7 years) children. Younger children exhibited a positive relationship between EF and SL, which was expected given Bahlmann et al.'s (2012) findings and our prediction that EF would underlie SL abilities. The older children's negative relationship between EF and SL is more puzzling, but seems to be driven by two participants who had high SL scores but low EF scores. For the remaining plots, younger children (and even middle age children) seem to have a relatively strong positive relationship between EF and language skills while there is practically no relationship for older children. The opposite is true for SL and language: older children have a strong positive relationship while younger have little relationship or even a negative relationship. Thus, it appears that EF is a more important contributor to language development for our younger children while SL is more important for our older children.

To further inspect differences between the oldest and youngest children, we created separate ERP waveforms for the two age groups to look for evidence of differences in SL patterns (see Figures 6 and 7). Visual inspection of the waveforms reveals that for the 7-year-olds, there is a P300-like positivity that is most pronounced for the HP condition in the second block of the task. On the other hand, for the

A



B



C

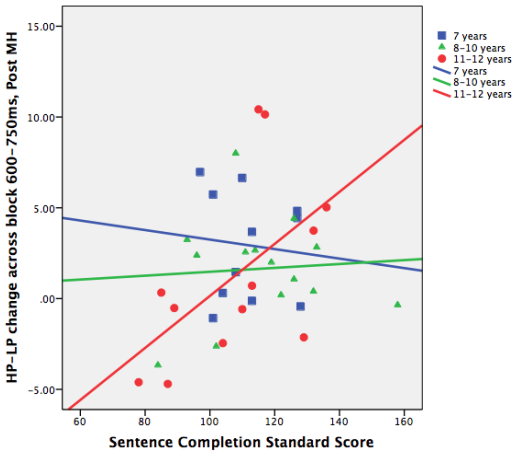


Figure 5: Scatter plots depicting the relationships between A) EF and SL, B) EF and language, and C) SL and language for three age groups: 7 years (blue squares), 8-10 years (green triangles), 11-12 years (red circles).

11-12-year-olds, the HP and LP waveforms are not clearly differentiated; in addition, there does not appear to be much difference between the first and second blocks.

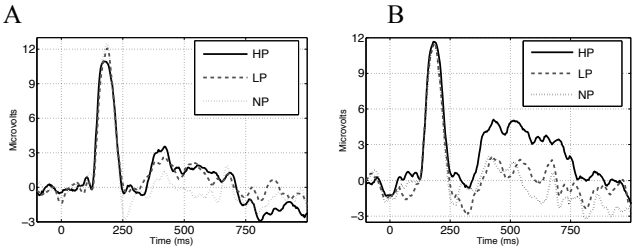


Figure 6: 7-year-old age group's grand-averaged ERP (n=12) in the medial posterior region to HP (solid), LP (dashed), and NP (dotted) trials for first block (A) and second block (B) (Positivity upward in microVolts; time in milliseconds).

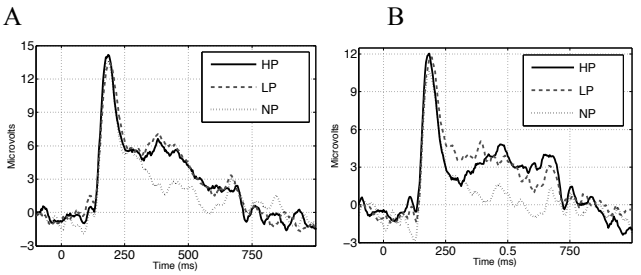


Figure 7: 11-12-year-old age group's grand-averaged ERP (n=12) in the medial posterior region to HP (solid), LP (dashed), and NP (dotted) trials for first block (A) and second block (B) (Positivity upward in microVolts; time in milliseconds).

2(block) x 3(predictor) ANOVAs for each of the two age groups revealed a significant main effect of predictor, $F(2,20) = 3.69$, $p = .043$, $\eta_p^2 = .27$ and significant block x predictor interaction, $F(2, 20) = 8.87$, $p = .002$, $\eta_p^2 = .47$ for 7-year-olds but no significant effects for 11-12-year-olds. This may suggest that while 7-year-olds showed SL over time, 11-12-year-olds did not evidence SL over time. For the 7-year-olds, predictor differences did not hold for post hoc tests, but posthoc pairwise comparisons for the interaction indicated that only HP increased from the first half to the second half ($p = .015$), although there was a trend toward NP decreasing across blocks ($p = .065$).

Discussion

This study sought to examine how sequential learning (SL) and executive function contribute to language across development. In a sample of 7-12 year old children, we used ERPs to measure the neural correlates of SL in a visual statistical-sequential learning task. The ERPs revealed that the youngest children showed evidence of statistical learning whereas the oldest did not (however, see Arcuili & Simpson, 2011, for evidence of increased SL with age using a different SL task). In addition, we used a conditional process model to assess the extent that SL in conjunction with EF was associated with a behavioral measure of language development for the different age groups. The results of this model revealed that both SL and EF had their own independent contributions to language development, but the relationship between both SL and language and EF and language showed dramatic shifts occurring between 7 years and 11-12 years of age.

As SL has been considered to be an implicit process, and implicit learning has been argued to be developmentally invariant (e.g., Reber, 1993; Vintner & Perruchet, 2000; though see Thomas et al., 2004, for an opposing view), the observed finding that there may be a developmental change in SL is particularly important. Further, this development of SL seems to be accompanied by a counterintuitive developmental change to the relationship between SL and language ability.

Although the finding that younger children show better sequential learning than older children and that there may be opposite relationships between SL and language across ages may seem puzzling, some previous findings may provide clarity. For example, in examining SL across the lifespan, Janacsek, Fiser, and Nemeth (2012) found that 4-12-year-olds had the strongest learning effects as measured by RTs with a dramatic decrease in SL ability around 12 years that continued to decline across the lifespan. However, accuracy scores were worst in the children and elderly participants with highest scores at the middle ages. Janacsek et al. (2012) suggested that these findings may be the result of tapping into two separate systems, with accuracy related to voluntary attentional control (an under-developed EF mechanism in early childhood) and RT related to involuntary attention mechanisms. Jost, Conway, Purdy, and Hendricks (2011) also presented findings that younger children may display heightened statistical learning abilities compared to older children and adults. Similarly, McNealy et al. (2011) found that younger children (5-7 years old) showed greater neural activation to weak statistical cues governing a novel stream of nonsense syllables, compared to older children (9-10 and 12-13 year olds) and adults. It may be advantageous to have efficient information processing mechanisms for detecting statistical patterns early in development, which could provide an explanation for why young children are able to learn natural language so effectively.

According to Kuhl (2004), infants and young children who have had relatively little experience are extremely open to learning; however, the more they learn, the more entrenched and neurally committed their brains become to the specific statistical and prosodic patterns of their own language. This may allow children and adults to readily process sequential structure of their native language, but may also have the side effect of making it more difficult to learn other sequential structures. This shift can be seen around 9 months when infants begin to stop using and recognizing sounds from non-native languages but improve in their ability to use and recognize sounds from their native language (Kuhl, 2004); another shift appears to occur

around 7 years and at adolescence, both times at which learning a foreign language fluently becomes much more difficult (e.g., Johnson & Newport, 1989).

Thus the anchor ages that we tested (7 and 12 years) are precisely at points when previous research has found changes in both SL and language that amount to both decreases in performance of unfamiliar items and increases in performance involving familiar items (e.g., native language and its underlying statistical structure). Voluntary attentional control (an EF mechanism improving over this period) and involuntary attention (an already developed mechanism) may also play a role in the relationships between SL, language, and age. We suggest that at around 7 years of age, continued openness to unfamiliar sequence structures is detrimental to language development. By 12 years, SL is lower in general than it is for 7-year-olds, but within that lower level individuals with higher SL combined with overall higher EF provides a benefit to language ability. As suggested by Arciuli and Torkildsen (2012), longitudinal research is necessary to fully flesh out the complicated interrelationships between sequential learning, language, executive function and age. Although these findings are still preliminary, we expect this research to lead to valuable information about the basic nature of the neurocognitive mechanisms of language development.

Acknowledgments

Thank you to our participants and our funding source, The National Institutes of Health (Grant R01DC012037).

References

Arciuli, J. & Simpson, I.C. (2011), Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental Science, 14*(3), 464-473.

Arciuli, J. & Torkildsen, J.K. (2012). Advancing our understanding of the link between statistical learning and language acquisition: The need for longitudinal data. *Frontiers in Psychology*, *3*, 324.

Bahlmann, J., Korb, F.M., Gratton, C., & Friederici, A.D. (2012). Levels of integration in cognitive control and sequence processing in the prefrontal cortex. *PLoS ONE,7*(8), e43774.

Best, J.R. & Miller, P.H. (2010). A Developmental Perspective on Executive Function. *Child Development, 81,* 1641–1660.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language.* Circle Pines, MN: AGS.

Christiansen, M.H., Conway, C.M., Onnis, L. (2012). Similar neural correlates of language and sequential learning: Evidence from event-related potentials. *Language and Cognitive Processes,* 27(2), 231-256.

Conway, C.M., Bauernschmidt, A., Huang, S.S., & Pisoni, D.B. (2010). Implicit statistical learning in language processing: word predictability is the key. *Cognition, 114, 356–371.*

Conway, C.M., Pisoni, D.B., Anaya, E.M., Karpicke, J., & Henning, S.C. (2011). Implicit sequence learning in deaf children with cochlear implants. *Developmental Science*, *14*, 69-82.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: The Guilford Press.

Janacsek, K., Fiser, J., Nemeth, D., 2012. The best time to acquire new skills: age-related differences in implicit sequence learning across the human lifespan. *Developmental Science*, *15*, 496–505.

Johnson, J.S. & Newport, E.L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21,* 60-99.

Jost, E., Conway, C.M., Purdy, J.D, Walk, A.M., & Hendricks, M.A. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. *Brain Research, 1597,* 95-107.

January, D., Trueswell, J.C., Thompson, Schill, S.L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience, 21,* 2434-2444.

Jost, E., Conway, C.M., Purdy, J.D., & Hendricks, M.A. (2011). Neurophysiological correlates of visual statistical learning in adults and children. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2526-2531). Austin, TX: Cognitive Science Society.

Kidd, E. & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development, 87,* 184–193.

Kuhl, P.K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831-843.

McNealy, K., Mazziota, J., & Dapretto, M. (2010). The neural basis of speech parsing in children and adults. *Developmental Science*, *13*(2), 385-406.

Misyak, J.B. & Christiansen, M.H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62,* 302-331.

Reber, A.S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford University Press: New York.

Romberg, A.R. & Saffran, J.R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science, 37,* 1290-1318.

Shafto, C.L., Conway, C.M., Field, S.L., & Houston, D.M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy*, *17,* 247-271.

Thomas, K.M., Hunt, R.H., Vizueta, N., Sommer, T., Durston, S., Yang, Y., & Worden, M.S. (2004). Evidence of developmental differences in implicit sequence learning: an fMRI study of children and adults. *Journal of Cognitive Neuroscience, 16,* 1339–1351.

Vinter, A., & Perruchet, P. (2000). Implicit learning in children is not related to age: evidence from drawing behavior. *Child Development, 71*, 1223–1240.

Ambiguity and Representational Stability: What is the role of embodied experiences?

Ramón D. Castillo (racastillo@utalca.cl)
Centro de Investigación en Ciencias Cognitivas, Universidad de Talca
Avenida Lircay s/n, Talca, CP 3460000, VII Región, Chile

Talia L. Waltzer (twaltzer@ucsc.edu)
Department of Psychology, University of California, Santa Cruz
1156 High Street, Santa Cruz, CA 95064 USA

Heidi Kloos (heidi.kloos@uc.edu)
Center for Cognition, Action, and Perception, Department of Psychology, University of Cincinnati
5130 Edwards One, Cincinnati, OH 45221-0376 USA

Abstract

Embodied cognition is sometimes presented as an alternative to computational approaches, the argument being that cognition is strongly influenced by an agent's body movement. However, the exact nature of this influence is still uncertain. In the current paper, we add to the conversation by analyzing adults' predictions in a high-ambiguity task: Adults had to decide which of two objects would sink faster (or slower) in water. Ambiguity was achieved by pitting object volume and object mass against buoyancy: The winning object of a pair was sometimes the bigger and heavier one, and sometimes it was the smaller and lighter one. The crucial manipulation was whether the stimuli were real-life objects or 2D pictures. All participants were presented with pictures of the objects during a training phase (when they received feedback on their predictions). Real-life objects were either present during the phase prior to the training (jars-first condition), or during the phase after the training (jars-last condition). Findings showed a clear influence of hands-on experiences: When allowed to hold the objects, adults were more likely to demonstrate a simplistic focus on object heaviness. These results call for a more nuanced understanding of the effect of embodied experiences on the stability of representations. While embodiment sometimes can help distinguish relevant from irrelevant information, we show that it can also destabilize representations acquired through visual information.

Keywords: action; knowledge representation; predictions; ambiguity; misconceptions; hands-on explorations

Introduction

What is the source of our thoughts, beliefs, attitudes, and the like? Traditionally, this question has been addressed with models of symbolic activity of the mind: Thoughts might be formed on the basis of combining symbols, which themselves are computed on the basis of simpler symbols, derived from sub-symbolic codes of sensation and perception. Approaches of embodied cognition stand in sharp contrast to the traditional view of computational models. They claim that mental activity, seemingly a bodiless manipulation of symbols, is instead strongly influenced by the very physical non-symbolic movement of our bodies (e.g., Chemero, 2011; Gibbs, 2005; Wilson & Clark, 2009). Rather than suggesting purely symbolic activities of bodiless minds, proponents of embodied cognition make a convincing case that higher-level cognition is constrained by our bodily experience of being in the world (Goldin-Meadow, Cook, & Mitchell, 2009).

In the current paper, we seek to explore the influence of embodied experiences in more detail. Our guiding theoretical framework does not subscribe to a specific representational format or cognitive architecture. Instead, we postulate that the mind makes use of whatever constraints are available in order to perform systematically (e.g., Kloos, Fisher, & Van Orden, 2010). These constraints could come from symbolic content, from bodily experiences, or from constraints outside the mental or bodily activity. The central question, then, pertains to how these different constraints interact. For example, to what extent does embodied experience override, support, or interfere with visual perception?

To explore this question, we analyzed the responses of adults in a high-ambiguity prediction task: Adults had to decide which of two objects would sink faster (or slower) in water. Ambiguity was achieved by pitting object volume and object mass against buoyancy: The winning object of a pair was sometimes the bigger and heavier one, and sometimes it was the smaller and lighter one. Thus, the task could not be solved with a simplistic rule that focuses on one dimension only (i.e., just weight or just size). To be successful, one must integrate both mass and volume by paying attention to the distribution of mass. While this integration can be accomplished, even by children (Kohn, 1093), it is not likely to be an adult's first guess (Castillo & Kloos, 2013). In fact, the initial tendency might be to focus on mass exclusively to make a decision (Castillo, Kloos, Richardson & Waltzer, 2015).

Note that high-ambiguity tasks, while not necessarily common in adults' everyday experiences, have been used extensively to better understand the mind's inner workings. The idea is that a high-ambiguity context reveals internal biases, natural preferences of the mind, so to speak. They are particularly useful to explore the role of embodied experiences: If embodied experience matters, then it should help disambiguate the constraints of the task. An added twist

here is that the specific task we chose – predicting the sinking behavior of objects – yields common misconceptions, namely that a reliance on mass alone could lead to successful performance. Would haptic explorations allow adults to overcome these misconceptions faster? Or would it in fact be more difficult for adults to benefit from such experiences?

Our overall method was as follows: Adults were presented with pairs of transparent objects that differed in size and contained a certain number of weights, clearly visible to participants. There were three phases: a *pre-test*, a *training*, and a *post-test*. Each phase had the same prediction trials, the difference being only in whether participants received feedback (training) or not (pre-test, post-test). The crucial manipulation was, before each prediction, whether adults were presented with real-life objects, or whether they merely saw the objects via 2D pictures. Specifically, one group of participants could explore real-life objects during the pre-test (jars-first condition), and one group of participants could explore real-life objects during the post-test (jars-last condition). During all other phases, stimuli were the 2D pictures of the objects. To what extent does the embodied experience affect performance?

Method

Participants

Participants were 112 adults between 18 and 27 years of age, recruited from a Midwestern university. They were each assigned to one of two conditions: the jars-first condition (17 men, 38 women; M = 19.03 years, SD = 1.69), or the jars-last condition (14 men, 43 women; M = 19.02 years, SD = 1.67). They received partial course credit for participation, following an IRB-approved procedure.

Material and Apparatus

Real-life sinking objects were used in this experiment, dropped into a water tank to create feedback for adults' predictions. The objects were transparent glass jars that differed in their sizes. Round aluminum discs (43g) could be placed inside the jars to manipulate mass. The water tank was 1m tall and had a vertical dividing wall to make it possible for each jar to sink without being affected by the other's turbulences.

The jars were combined into pairs of objects. Figure 1 shows an example for various different trials, which differ in how mass and volume correlated with rate of sinking. The faster sinking object within a pair is marked with a star. In two of the trial types, only one of the features was varied (either mass or volume), and in three of the trial types, both mass and volume were varied. Specifically, in the *small-wins* pair (Fig. 1A), mass was held constant and the size of the jar was varied in such a way that the smaller jar sank faster. In the *heavy-wins* pair (Fig. 1B), volume was held constant and mass was varied in such a way that the heavier jar sank faster. In the *big/heavy-wins* pair (Fig. 1C), the faster sinking object was bigger and heavier than the slower object. In the *small/heavy-wins* pair (Fig. 1D), the faster sinking object was smaller and heavier than the slower object. And finally, in the *small/light-wins* pair (Fig. 1E), the faster sinking object was smaller and lighter than the slower object. There were nine pairs of each trial type, resulting in a total of 45 unique pairs.

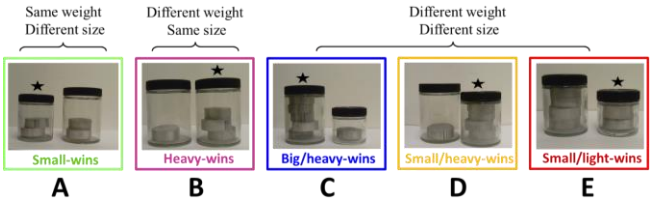


Figure 1: Example trials. Star marks jar that sinks faster. A: Small-wins pair. B: Heavy-wins pair. C: Big/heavy-wins pair. D: Small/heavy-wins pair. E: Small/light-wins pair.

For simplicity, we will only report performance on the big/heavy-wins pairs (Fig. 1C) and the small/light-wins pairs (Fig. 1E). These two types of pairs create the high-ambiguity task context needed for the current purposes. This is because, while mass and volume correlate positively (the bigger of the two objects was also the heavier one), it was sometimes the heavier and sometimes the lighter object that sank fastest. Thus, to perform correctly, it would not be sufficient to pay attention to either mass or volume alone. All other trials had low ambiguity and will be considered fillers (indeed, adults performed largely at ceiling during those trials).

Pairs were presented either as actual jars or as pictures on a screen. Figure 2 shows the picture versions of the stimuli. Each trial included a close-up picture of a pair with discs outside the jar (Fig. 2A) as well as a close-up with discs inside the jar (Fig. 2B). Feedback was always provided as a picture of the jars being dropped in the tank of water (Fig. 2C). A numeric keypad was used to record participants' predictions.

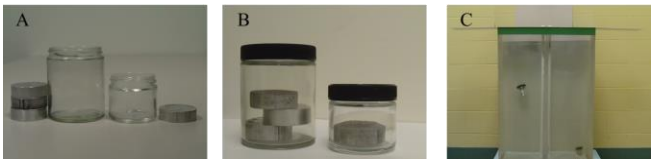


Figure 2: Example pictures. A: Empty jars with weights on either side. B: Jars filled with weights. C: Jars sinking in the water tank.

Procedure

Participants were tested individually in the lab, using DirectRT Precision Timing Software (2012 Version) to administer the experiment on a desktop computer. The experiment consisted of a total of 360 prediction trials, divided into three phases. The first phase was the pre-test (90 trials): Participants made predictions across various jar combinations, without receiving any feedback. The second phase served as training (180 trials): Adults' predictions were followed by corrective feedback. Finally, the last phase was

the post-test (90 trials), featuring prediction trials that were identical to the pre-test (no feedback provided).

Our main manipulation was the timing of the embodied experience. Participants held the real objects either during the pre-test (jars-first condition), or during the post-test (jars-last condition). The training was always carried out with pictures. For generalizability purposes, we also manipulated the type of predictions adults had to make: Participants were either asked to predict which of two jars would sink faster, or which of two jars would sink slower.

During familiarization, participants were shown empty jars of different sizes, as well as several aluminum discs. They were told that all the discs have the same weight. The experimenter then filled the large and small jars with aluminum discs and asked the participant to predict which of them would sink faster in water (or slower). Participants were encouraged to lift the jars before making their predictions. Then they were provided with feedback pictures showing the outcome of the jars after being dropped in water. Finally, participants were shown the keypad and how it works. Prior to the experiment proper, they were informed that the pictures were taken from the real objects.

For predictions with real-life objects (pre-test or post-test, depending on condition), participants sat in front of an opaque box (60 x 25 x 40 cm) that served as a table to hold the objects. It also served as a barrier behind which the researcher kept the 12 jars (see Fig. 3 for a schematic over-head view of this arrangement). Based on a random order determined for each participant prior to the start of the experiment, the jar pairs were presented one at a time. For each pair, participants were asked to make a prediction about which one of the two jars would sink faster (or slower) in water. Participants were encouraged to respond by saying "left" or "right", corresponding to whether the winning (or losing) jar was in their left hand or right hand.

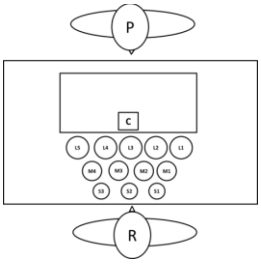


Figure 3: Diagram of the set-up for prediction trials with real jars. *R*: Researcher. *C*: Video camera. *P*: Participant.

For predictions with pictures (pre-test or post-test, depending on condition), participants were first shown an image of two empty jars next to each other, with a stack of discs by each jar. This allowed participants a clear view of the number of discs for each object. After 1.5 seconds, the image was replaced with a picture of the same two jars, but now filled with the discs and closed with a lid. Participants were asked to decide which of the two jars would sink faster (or slower). Figure 4A shows such a trial in schematic form. There was no time restriction for making a prediction. The trial ended when the participant pressed the keypad to provide a prediction. A fifth of the all trials were big/heavy-wins trials, and a fifth of the trials were small/light-wins trials, interspersed with filler trials.

Training was identical to pre- and post-test predictions, the only difference being that a feedback picture was shown for 1.5 seconds, right after the participant made a prediction. On the very first feedback trial, the image was explained. The faster sinking object was pointed out on the computer screen, and participants were provided with explicit feedback (e.g., "Yes, you were right"; "No, look, it was the other one that sank faster"). Training took place between pre- and post-test. Of all the training trials, a fifth were big/heavy-wins trials, and a fifth were small/light-wins trials, interspersed with filler trials.

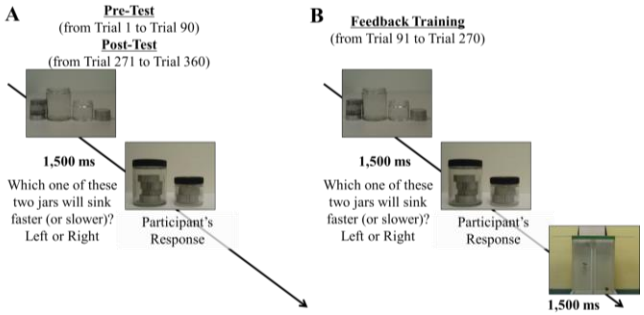


Figure 4: Schematic representation of the prediction trials. A: Example picture trial during pre- or post-test. B: Example picture trial during feedback training.

Results and Discussion

Our dependent variable was the proportion of correct predictions on big/heavy-wins and small/light-wins trials. Figure 5 presents the accuracy data for these two trial types, separated by phase (pre-test, training and post-test), and by the embodiment manipulation (jars-first vs. jars-last).

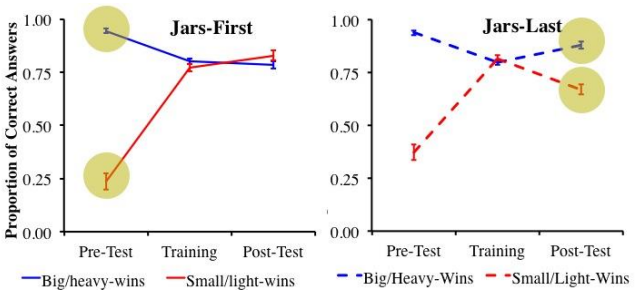


Figure 5: Proportion of correct answers for trial type and phase, separated by condition. Error bars represent standard errors of the mean. Circles highlight when real-life jars were used.

A 2 x 2 x 3 mixed-design ANOVA was carried out, with trial type (big/heavy-wins; small/light-wins) and phase (pre-test; training; post-test) as within-group factors, and with

condition (jar-first; jar-last) as the between-group factor. We found a significant 3-way interaction, $F(2, 220) = 22.47$, $p < .001$, $\eta^2 = .17$, prompting us to look at the results separately by phase (see Table 1 for a summary of the results).

During the pre-test, participants' performance was markedly different for the two types of trials: While performance was at ceiling (or close to) on big/heavy-wins trials ($M_{Pic} = .94$; $M_{Jar} = .94$), participants made systematic mistakes on the small/light-wins trials ($M_{Pic} = .37$; $M_{Jar} = .24$). It appears that participants resolved the ambiguity of the prediction task by focusing on weight exclusively. A 2 x 2 mixed-design ANOVA, with trial type and condition as factors, revealed a main effect of trial type, $F(1,110) > 100$; $p < .001$, a main effect of condition, $F(1,110) = 8.10$; $p < .01$, as well as a significant interaction, $F(1,110) = 4.74$; $p = .032$. The interaction is driven by the fact that participants' mistakes on small/heavy-wins trials were even more pronounced when they handled jars (jars-first condition) than when they viewed pictures (jars-last condition), $F(1,110) = 6.32$; $p = .013$.

During the training, the difference between trial types disappeared, whether participants were in the jars-first condition ($M_{Big/heavy} = .80$, $M_{Small/light} = 0.77$) or in the jars-last condition ($M_{Big/heavy} = .80$; $M_{Small/light} = 0.82$). A 2 x 2 mixed-design ANOVA (trial type by condition) yielded no main effects and no interaction, Fs(1,110) < 2.37; ps > .12. Performance was clearly above chance, $t(111) > 5$; $p < .01$, implying that adults benefited from the training and quickly learned that a focus on mass or volume alone yields mistakes.

To compare performance during training and pre-test, we carried out two 2 x 2 repeated-measure ANOVAs (trial type by phase), one for the jar-first condition, and one for the jar-last condition. For both conditions, the analysis yielded highly significant main effects and interactions, Fs > 80, ps < .001. While performance on big/heavy-wins pairs decreased slightly from pre-test to training in both conditions, ps < .001 it starkly improved for small/light-wins pairs, ps < .001. The results show that training had very similar effects on performance, whether participants had a chance to haptically explore the objects prior to training or not. This confirms that the switch between 3D objects to 2D pictures of the objects did not have a discernable effect on performance.

Finally, during the post-test, the difference between trial types was affected by condition. The trial type by condition mixed-design ANOVA revealed a main effect of trial type, $F(1,110) = 11.21$; $p < .001$, and a marginal main effect of condition, $F(1,110) = 3.55$; $p = .06$, both driven by the highly significant interaction, $F(1,110) = 26.72$; $p < .001$, $\eta^2 = .20$. To be more specific, trial types yielded different performance when participants made their predictions using real jars (jars-last condition: $M_{Big/heavy} = .88$, $M_{Small/light} = 0.67$; $F(1,110) = 36.94$; $p < .001$, $\eta^2 = .25$), but not when they made their predictions using pictures (jars-first condition: $M_{Big/heavy} = .78$; $M_{Small/light} = 0.83$; $F(1,110) = 1.63$; $p = .20$.)

Comparing post-test performance with training performance, we found no significant main effect of trial type in the jars-first condition, $F < 1$. Put differently, when adults were presented with pictures, they retained what they learned during the training and performed well even without feedback. In contrast, in the jars-last condition, when participants were given the opportunity to explore the objects haptically, performance changed from training to post-test. The 2 x 2 repeated-measure ANOVA (trial type by phase) revealed a significant main effect of trial type $F(1,56) = 14.83$; $p < .001$; a significant main effect of phase, $F(1,56) = 5.56$; $p = .02$; and a significant interaction, $F(1,56) = 42.63$; $p < .001$. From training to post-test, performance on big/heavy-wins pairs increased, $p < .001$, while performance on small/light-wins pairs decreased, $p < .001$. Put differently, participants in the jars-last condition reverted back to disambiguating the conflict in making predictions by focusing on the feature of weight.

Table 1. Summary of results.

Pre-test
• Independent of condition, performance was at ceiling when the winning jar was big and heavy.
• Independent of condition, systematic mistakes were made when the winning jar was small and light.
• Systematic mistakes were higher when participants used jars (compared to pictures).
Training
• Independent of condition, performance was equally high on both big/heavy-wins and small/light-wins pairs.
• Participants made some mistakes, but performance was overall above chance.
Post-test
• When participants used pictures, performance remained unchanged (compared to the training).
• When participants used jars, performance increased for the big/heavy-wins pairs, while it decreased for the small/light-wins pairs.

Difference scores. To capture these findings on the level of individual participants, we calculated a difference score for each participant, based on their performance on the big/heavy-wins and small/light-wins trials. Specifically, we subtracted average accuracy scores for the small/light-wins trials from the big/heavy-wins trials. This difference reflects the extent to which participants held a big/heavy bias, choosing the bigger/heavier jar as the winner more often than the smaller/lighter jar. Figure 6 shows obtained results.

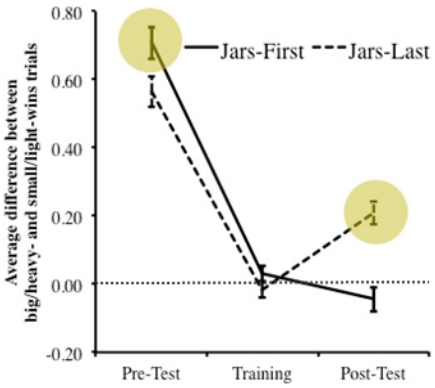


Figure 6: Average difference between big/heavy-wins and small/light-wins trials, per phase and condition. Error bars represent standard errors of the mean. Circles highlight when real-life jars were used.

Results are very much in line with our analysis of trial-based performance: Highest difference scores were obtained during pre-test (M_{Pic} = .56; M_{Jar} = .71), reflecting the naïve heaviness bias. Importantly, the difference score was higher for participants who were given the opportunity to explore the objects haptically than for participants who saw pictures, $F(1,110) = 4.74$; $p = .03$.

Difference scores decreased substantially during training, (M_{Pic} = .02; M_{Jar} = .03), reaching values that were statistically undistinguishable from zero, simple-sample $t < 1$. This suggests that participants no longer based their predictions on mass or volume alone. They quickly discovered a new criterion, yielding the same degree of success (i.e., number of mistakes) in both types of trials.

The central finding is during the post-test, after participants had learned about the shortcomings of a naïve heaviness bias. While participants in the jars-first condition largely retained the difference scores that they had obtained during training (M_{Pic} = .04), participants in the jars-last condition did not (M_{Jar} = .21). Their difference scores shot up, significantly more than the difference scores obtained in the jars-first condition, $F(1,110) = 26.72$; $p < .001$. Participants who were allowed to handle the real objects during the post-test phase seemed to unlearn some of what was learned during training.

Summary and Conclusion

We set out to explore the extent to which embodied experiences override, support, or interfere with experiences that are gained from visual perception. Adults participated in a prediction task about sinking objects, a task that is thought to elicit mistaken beliefs about what makes an object sink faster or slower in water. Feedback during part of the prediction task was expected to change some of those initial misconceptions. Indeed, adults demonstrated a substantial amount of learning during training. At the same time, we succeeded in creating a task that was sufficiently difficult for adults to perform below ceiling, but not so difficult that they would merely make random guesses. This is the kind of regime that is likely to shed light on the constraints on mental activity.

How did embodied experience interface with performance when using 2D pictorial stimuli? Our results are clear: there was no evidence that embodied experience simply *overrode* visual perception. Even though adults were presented with the exact same trials across conditions, when their chance to explore objects haptically took place before training (jars-first condition), performance was decidedly different from when it took place after the training (jars-last condition). This suggests that behavior derived from embodied experiences is not separable from behavior derived from other means of perception. This, of course, is no surprise to a one-mind-one-behavior systems view (e.g., Clark, 2013; Smith, 2005). Visual and embodied perception are likely to be interlinked. Thus, results that argue for a dissociation between visual and embodied experience need to be re-evaluated carefully.

We also found no evidence that embodied experience *supports* visual perception. This is at least the case if support pertains to performing accurately. Whether participants got a chance to haptically explore objects before or after the training, their performance on the small/heavy-wins trials was lower with real jars than when they saw the objects as pictures. This is especially evident after the training, when participants reached equivalent levels of competence. Performance levels stayed the same during the post-test for adults presented with pictures, but critical mistakes arose from adults presented with the real-life objects. These findings undermine blanket claims of the general advantage of hands-on, embodied learning.

Results show that embodied experience *interfered* with visual perception. It did not act separately, and it did not support it, but nevertheless, it interfered with it drastically. This finding far from trivial given the current task, because relevant information, say about object mass and volume, were available to both modalities: participants could count the number of weights and compare the sizes of the objects, whether they were presented in real life or as pictures. If the same information can be obtained in theory, why then did we find differing performance as a function of condition?

Could it be that proprioceptive information simply made the task harder, yielding non-specific mistakes? This is unlikely, given that the differences in performance between the jar-based and picture-based contexts were rather specific, both in the pre-test and the post-test. In fact, there was not a general increase in mistakes for participants exposed to real-life objects: When they explored objects haptically, they performed highly successfully on big/heavy-wins trials, even better than adults who merely saw pictures. Their mistakes increased only on the small/light-wins trials. This pattern of performance, to perform well on big/heavy-wins trials and poorly on small/light-wins trials, is the signature of a heaviness bias, a bias that was more pronounced when participants could hold objects, rather than view them on a computer screen.

One could argue that our set-up was an unfair comparison: Embodied experience might support visual perception, but not in a task in which salience to heaviness yields mistakes. Embodiment might make heaviness salient, due to the inherently salient down-ward force of holding objects. Our results might reflect nothing more but a bias of embodied experience to increase the salience of heaviness, failing to generalize to embodied experience outside of heaviness tasks. While our data do not speak directly to this criticism, it is nevertheless worth questioning. This is because the difference in mass between the two objects in a pair is likely to be far more salient in the picture case than the real-jar case. The weights were too light (only 43g) to create differences that could be readily perceived haptically. It is most likely that adults judged difference in weight on the basis of visual information. Thus, a high salience of heaviness in embodied experiences might not explain our results.

There are several possible reasons for why the embodied experience increased the heaviness bias of adults. One possibility is that, rather than making heaviness more salient, the redundancy of information between visual and tactile information may have prompted the system to revert to a simpler belief (in this case, about heaviness). Without this redundancy of information, adults might have relied on their memory of feedback on specific pairs, and merely guessed on those pairs they could not remember. The haptic information might have disrupted this strategy to some extent. To test this possibility, the study would need to be expanded to include a manipulation of explicit, non-tactile disruption.

Our results support the idea that performance emerges from the interaction of many components that change each other over time driven by the system's own history (Smith, 2005; Smith & Breazeal, 2007). Such interactive processes have a non-linear character, and, beyond a certain size and number of relations among their constituents, they express a complex behavior of self-organization that cannot be explained by the simple features of the elements (Steenbeek & Van Geert, 2008; Van Orden, Holden, & Turvey, 2003). This dynamic pattern is difficult to place in a single component, because it is a product of the coordination of the whole system (Steenbeek & Van Geert, 2008). It is possible that the mind capitalizes on the dynamics of the body when needed (Spencer, Austin, & Schutte, 2012; Turvey, 1990; 2007). However, our results call for a more nuanced understanding of the effect of embodied experiences on the stability of representations.

Acknowledgments

Support for this study was provided by the National Science Foundation (DLS 1313889; Kloos) and Universidad de Talca (VAC 600692; Castillo). This research was completed as part of a dissertation (VAC 600692; FONDECYT 1161533; Castillo). We thank Samantha Linsky, Alexandra Matthews, Catherine Schneider, Presley Benzinger, Tiara Clark, Samantha Hinds, Theresa Grefer, and Allison Stewart for assistance with data collection.

References

Castillo, R. D., & Kloos, H. (2013). Can a flow-network approach shed light on children's problem solving? *Ecological Psychology, 25*, 281-292.

Castillo, R. D., Kloos, H. Richardson, M. J. & Waltzer, T. L. (2015). Beliefs as self-sustaining networks: Drawing parallels between networks of ecosystems and adults' predictions. *Frontiers in Psychology 6(1723)*. doi: 10.3389/fpsyg.2015.01723

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181-204.

Chemero, A. (2011). *Radical embodied cognitive science.* MIT press.

Gibbs Jr, R. W. (2005). *Embodiment and cognitive science.* Cambridge University Press.

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*, 267-272.

Kloos, H., Fisher, A., & Van Orden, G. C. (2010). Situated naïve physics: Task constraints decide what children know about density. *Journal of Experimental Psychology: General, 139*, 625-637.

Kohn, A. S. (1993). Preschoolers' reasoning about density: Will it float? *Child Development, 64*, 1637–1650.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review, 25*, 278-298. doi:10.1016/S1364-6613(03)00156-6

Smith, L. B., & Breazeal, C. (2007). The dynamic lift of developmental process. *Developmental Science*, *10*, 61-68.

Steenbeek, H., & Van Geert, P. (2008). An empirical validation of a dynamic systems model of interaction: Do children of different sociometric statuses differ in their dyadic play? *Developmental science*, *11*, 253-281.

Spencer, J. P., Austin, A., & Schutte, A. R. (2012). Contributions of dynamic systems theory to cognitive development. *Cognitive Development, 27*, 401-418.

Turvey, M. T. (1990). Coordination. *American Psychologist, 45*, 938-953.

Turvey, M. T. (2007). Action and perception at the level of synergies. *Human Movement Science, 26*, 657-697.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*, 331.

Wilson R. A. & Clark, A. (2009). How to situate cognition: Letting nature take its course. In Murat Aydede & P. Robbins (eds.), *The Cambridge handbook of situated cognition*. Cambridge University Press.

Questions in informal teaching: A study of mother-child conversations

Yue Yu[12] (yue.yu@rutgers.edu), Elizabeth Bonawitz[2] (elizabeth.bonawitz@rutgers.edu), Patrick Shafto[1] (patrick.shafto@gmail.com)
Department of [1]Math and Computer Science & [2]Psychology, Rutgers University - Newark
Newark, NJ 07102 USA

Abstract

Questioning is a core component of formal pedagogy. Parents commonly question children, but do they use questions to teach? Research has shown that informal pedagogical situations elicit stronger inferences than the same evidence observed in non-pedagogical situations. Certain questions ("pedagogical questions") have similar features. We investigate the frequency and distribution of pedagogical questions from mother-child conversations documented in the CHILDES database. We show that pedagogical questions are commonplace, are more frequent for middle-class mothers compared to working-class mothers, are more frequent during free play than during daily routines, and are more frequent in mothers who ask more questions. The results serve as a first step towards understanding the role of questions in informal pedagogy.

Keywords: informal pedagogy; mother-child conversation; individual differences; socioeconomic status; CHILDES.

Introduction

To question well is to teach well.
- Henry Barnard, 1860, American Journal of Education

Asking and answering questions has been seen as a core component of teaching and learning at least since the days of Socrates. Research in education suggests that, as a teaching technique, questioning leads to improvements in learning outcomes (Shymansky, Hedges, & Woodworth, 1990; Wise & Okey, 1983), and preliminary research in cognitive development suggests that question asking may support rapid and effective learning (Blewitt, Rump, Shealy, & Cook, 2009). However, not all questions are created equal. We suggest that questions by a knowledgeable informant in the service of teaching ("pedagogical questions") are a special class of questions separate from other information-seeking or rhetorical questions (unlike previous approaches; cf. Olsen-Fulero & Conforti, 1983). We explore whether and how parent-child conversations include pedagogical questions in everyday interactions.

Questioning plays an important role in formal pedagogy: Teachers ask ~400 questions per day in a typical classroom setting (Gall, 1970), and these questions serve various functions from checking class work to motivate thinking (Black, 2001). The effect of questioning on students' learning has been summarized in a meta-analysis demonstrating that questioning has the largest impact on cognitive outcomes (such as general achievement, specific learning outcomes, and problem solving) as compared to the other dozen teaching techniques surveyed (Wise & Okey, 1983).

However, learning begins well before formal classroom experiences. Theory in cognitive development emphasizes the importance of everyday interactions in which adults help children learn—informal pedagogy (Csibra & Gergely, 2009; Tomasello, 1999). Recent advances in developmental psychology have shown that infants and young children are sensitive to adults' cues that suggest an intention to teach, such as joint attention, child-directed speech, name-calling, etc. (Csibra & Gergely, 2009) These "pedagogical cues" lead to strong inferences that the adult is showing important information, and they facilitate learning in various domains (Bonawitz et al., 2011; Butler & Markman, 2014; Sage & Baldwin, 2010; Topál, Gergely, Miklósi, Erdőhegyi, & Csibra, 2008; Vredenburgh, Kushnir, & Casasola, 2014).

One key factor for the effect of informal pedagogy on learning is whether an adult appears to be knowledgeable about what she is doing (Harris & Corriveau, 2011; Koenig, Clément, & Harris, 2004; Shafto, Goodman, & Frank, 2012). In particular, pedagogical demonstrations from knowledgeable others justify drawing stronger inferences than would otherwise be licensed by the data. Given the huge space of possibilities for any problem, being able to harness pedagogical inferences may bolster learning. For instance, Bonawitz, et al (2011) investigated how children explore a novel toy with multiple functions after watching an adult demonstrating one target function. Two conditions, among others, manipulated the knowledge state of the demonstrator: she either appeared to be knowledgeable and helpful when demonstrating the target function (pedagogical condition), or she appeared to be ignorant about the toy and activated the function by accident (accidental condition). Children were more likely to learn the target function in the pedagogical condition than in the accidental condition (similar examples in other domains of learning see Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2014).

If a critical aspect of pedagogical demonstrations is to facilitate learning about the correct hypothesis (or set of hypotheses), then perhaps questions that are asked by one who *already knows the answer* will have similar functions. A learner who infers that the question asker already knows the answer, may thus infer that the *goal of the question is to teach the askee rather than acquire information for the asker*. As in the case of pedagogical demonstration, we suggest the questioner's knowledge state play an important role in learning. Do parents, like teachers, use questions for which they already know the answers to facilitate learning?

As a first step towards a broader goal of understanding the role of pedagogical questions in learning, we explore whether parents use pedagogical questions, and if so, how common they are. We do so by analyzing parent-child conversations from the CHILDES database (MacWhinney & Snow, 1990). We coded mothers' questions to their children into three categories: Pedagogical questions are those for which mothers know the answer; information-seeking questions are those for which mothers do not know the answer; and rhetorical questions are those not intend to be answered verbally. Analyzing parents' questions with regard to these categories will provide information about whether questions are used to guide learning in informal pedagogy. Additionally, we characterize cross-individual, cross-situational, and developmental variation in their use, and review relevant prior work briefly in the following sections, before presenting our results and discussion.

Contextual differences in questioning

Context plays an important role in guiding parent-child interactions. Research has identified general differences in conversations, as well as specific differences in questioning, across different contexts such as book reading, free play, mealtime conversations, caretaking interactions, and household routines (Dunn, Wooding, & Hermann, 1977; Goddard, Durkin, & Rutter, 1985; Snow et al., 1976). The general function and pace of the activities are likely causes of these different interactions by context. Snow, Dubber, and De Blauw (1982) have likewise suggested that contexts that are typically low-stress and non-goal-directed (e.g., free play) are more likely to elicit conversational interactions; whereas contexts that are typically high-stress and goal-directed (e.g., daily routines) are more likely to elicit directive interactions. Intuitively, a parent who is interested in engaging in conversation with their child is more likely to choose the relaxed moments of dyadic free-play, than instances in which daily routines must be quickly accomplished.

We investigate whether types of questions may also differ across these contexts: One possibility is that non-goal-directed activities like free play provide more opportunities for teaching, so that mothers would ask more pedagogical questions. On the other hand, it is also possible that non-goal-directed activities provide more opportunities for conversation, so that mothers would simply ask more questions of all kinds.

Individual differences in questioning

Might questioning differ across mothers and across families from different socioeconomic status (SES)? Research has identified both quantitative and qualitative differences in parent-child conversations across SES, and these differences are predictive of the "achievement gap" in children's learning outcomes (Hoff-Ginsberg, 1991; Hoff, 2013; Snow et al., 1976). Questioning has been shown to be an important part of the difference: One study (Snow et al., 1976) showed that mothers from academic middle-class families posed more wh-questions ("who", "what", "where", "when", "why", "how") and less yes-no questions than mothers from working-class and lower middle-class families. Indeed, one way mother-child conversation may relate directly to learning is through the use of questioning strategies common in formal pedagogy. We investigate this possibility by analyzing the distribution of pedagogical questions for mothers from different SES, as well as for mothers who ask many versus few questions.

Developmental changes in questioning

Developmental changes can be important in revealing the cognitive mechanisms that drive learning. Considerable evidence suggests that the types of questions children are asked changes with development (Levelt, 1975). Infants frequently hear "contingent queries" which ask them to repeat or clarify their utterances (Garvey, 1977; Wilcox & Webster, 1980), whereas toddlers and preschoolers hear more yes/no questions and wh- questions (Snow, et al., 1976).

The types of questions posed to children may also vary with development, especially in terms of the knowledge states of the questioner. Evidence suggests developmental changes in children's understanding and inferences about other people's knowledge states (Wellman & Liu, 2004). Similarly, research has shown that whereas 4-year-olds consistently mistrusted informants who were ignorant or inaccurate, 3-year-olds were less discriminating about who they trust for information (Clement, Koenig, & Harris, 2004). Do the types of questions parents ask change with their children's age? Evidence suggests that older children would be better prepared to correctly interpret pedagogical questions; however, it is possible that parents' use of questions may not reflect the changing cognitive capabilities of their children.

Method

Sample

We searched the CHILDES database (MacWhinney & Snow, 1990) for transcripts that meet the following eight criteria: 1) The transcript was in English; 2) The conversation took place at home; 3) The conversation partners included a mother and a child, and did not include anyone outside the immediate family (interviewer, grandparents, relatives, friends, etc.); 4) The target child was between 3 and 6 years of age; 5) The conversation represented everyday talk, and was not a purposeful conversation such as an interview; 6) The transcripts for mother's and child's speech were not separated; 7) The transcript used punctuation marks, and had at least three question marks in it; 8) If there were multiple transcripts for a same child (such as in longitudinal studies), we only used the first (earliest) transcript that meets all other criteria.

The final sample was 94 transcripts from 14 studies (online supplementary material for a summary of the sample is provided at: http://shaftolab.com/YuCogsci16Sup.docx). Fifty-eight of the 94 transcripts came from one study

Table 1: Coding scheme for questions. Note that the subcategories of pedagogical questions and information-seeking questions are by nature asymmetrical: For example, mothers did not ask children generic questions for which they do not know the answer, so only pedagogical questions contain a "generic" subcategory; On the other hand, mothers always asked about children's needs, opinions, and status as information-seeking questions.

Category / subcategory	Description	Examples	Proportion (*SD*)
Pedagogical	Asker knows the answer of the question	—	23.8% (22.7%)
Generic	Testing child's knowledge about kinds, concepts, rules, or scripts	"What's 'A' stand for?"; "What'd you say [in this situation]?"	2.6% (9.0%)
Specific	Testing child's knowledge about a specific object, event, or person	"Who came for your birthday party?"; "What does this button do?"	21.1% (22.3%)
Information-seeking	Asker does not know the answer of the question	—	61.1% (23.6%)
Specific	Asking about a specific object, event, or person	"What did you do at school?"	28.1% (22.0%)
Check status	Asking about the child's needs, opinions, or physical/emotional/epistemic status	"Are you hungry?"; "Do you remember?"	22.9% (19.4%)
Clarification	Asking the child to repeat what he/she just said	"You what?"; "Huh?"	9.2% (11.2%)
Permission	Asking for permission	"Can I get you changed?"	0.9% (2.8%)
Rhetorical	No verbal answer is expected for the question	—	15.2% (15.9%)
Commands	Giving commands in a question form	"Can you help clean up?"	10.3% (13.0%)
Attention	Raising child's attention with a question	"Well?"; "Jack?"; "Be good, hear?"	4.8% (7.7%)

(Dickinson & Tabors, 2001) which recorded mealtime conversations in low-income, racially diverse families.

Coding Procedures

For each transcript, one coder first recorded the target child's gender, age, and the conversation partners from the heading of the transcript. The coder also recorded the social economic status (SES) of the family and the context of conversation if that information was available. Next the coder derived the total number of the mother's statements and questions in the transcript using tools provided by the database. The frequency of questioning (per 100 statements) was calculated by dividing the number of questions by the total number of statements, and then multiplying by 100.

Questions. For each transcript, two coders blind of the hypotheses independently coded the first 10 questions that 1) ended with a question mark; 2) was asked by the mother; 3) was directed towards the target child; and 4) did not contain missing words ("xxx"). Eight transcripts contained less than 10 such questions (with a minimum of 3 questions), and for those transcripts all questions were coded.

Each question was assigned into one of eight subcategories under one of three major categories (Table 1). The coders first determined whether the question was a pedagogical, information-seeking, or rhetorical question, and then determined their subcategories. Context before and after the question were used to help determine whether the mother knew the answer of the question. Since transcripts do not fully capture the history and details in mother-child interactions, this sometimes required an inference on the part of the coder. However, the inter-rater reliability was acceptable, which shows the majority of questions can be reliably categorized based on our coding scheme. When only the three major categories were considered, Cohen's $\kappa = .80$; When all eight subcategories were considered, Cohen's $\kappa = .78$. Inconsistent codes were reviewed and resolved by a third coder.

Context. After coding all 10 questions and answers for a transcript, one coder determined whether the main context of the conversation was 1) over meal, 2) during free play, or 3) during daily routines (e.g., dressing/bathing the child, preparing to leave). The other coder checked the code and discussed with the first coder if she did not agree.

Data analysis

All data was entered and analyzed in IBM SPSS 22. Two-tailed tests were used for all between-group comparisons, and an α level of .05 was used for all tests.

Results

A total of 907 questions were coded. Table 1 shows the proportion of each type of questions averaged across all transcripts.

Contextual differences in questioning

Seventy-two out of the 94 transcripts were coded as conversation over meal (687 total questions); 17 were during free play (170 total questions); and 5 were during daily routines (50 total questions). Overall *frequency* of questioning did not differ across contexts: average number of questions per 100 statements was 34 for meal, 37 for free play, and 32 for daily routines, $F(2, 91) = 0.50$, $p = .61$, $\eta^2 = .01$. However, the composition of questions did differ (Figure 1). The proportion of pedagogical question was significantly

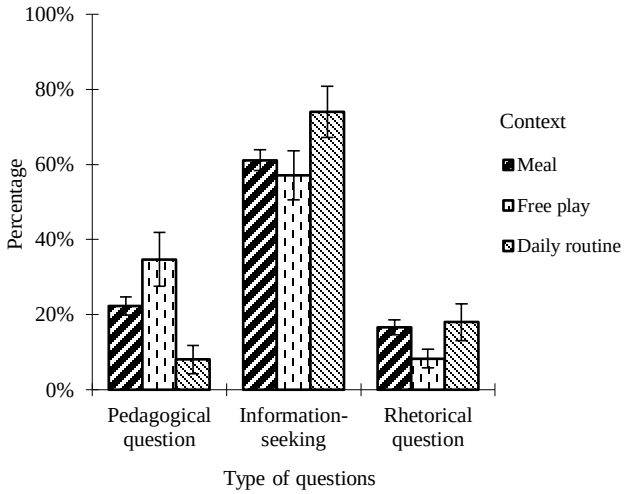


Figure 1. Proportion of pedagogical, information-seeking, and rhetorical questions across contexts. Pedagogical questions were more commonly asked during free play than during daily routine. Information-seeking and rhetorical questions did not differ significantly across contexts. Error bars denote *SE*.

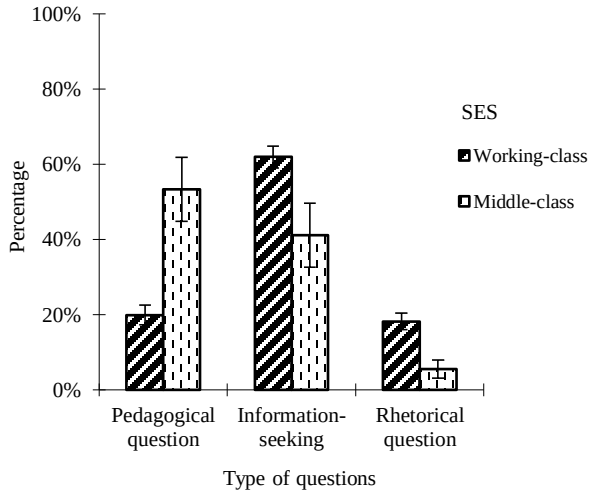


Figure 2. Proportion of pedagogical, information-seeking, and rhetorical questions for working- and middle-class SES. Higher SES asked more pedagogical questions and fewer information-seeking and rhetorical questions. Error bars denote *SE*.

different between contexts, $F(2, 91) = 3.52$, $p = .03$, $\eta^2 = .07$, with Post-hoc Bonferroni corrected test showing that mothers were more likely to ask pedagogical questions during free play (34.7%) than during daily routines (8.0%), $p = .05$, $d = 1.0$. Given the small sample size, we confirmed this difference with a nonparametric Mann-Whitney U Test, and the result was marginally significant: $z = 1.80$, $p = .07$. The proportion of pedagogical questions during meal time (22.2%) was not significantly different from the other two contexts, ps > .2. Unlike pedagogical questions, the proportion of information-seeking and rhetorical questions were not significantly different between contexts, Fs < 2.1, ps > .14.

Individual differences in questioning

Out of the 94 transcripts, 62 were documented as from working-class[1] (592 total questions), 9 were documented as from middle-class families (90 total questions), none were from upper-class families, and 23 were not specified (225 total questions). Overall frequency of questions was not different between working-class and middle-class mothers: average number of questions per 100 statements was 35 for working class and 37 for middle class, $t(69) = 0.56$, $p = 0.58$. However, the composition of questions differed by SES (Figure 2). On average pedagogical questions consisted a smaller proportion of questions in working-class families (20.0%) than in middle-class families (53.3%), $t(69) = 4.41$, $p < .001$, $d = 1.57$. In contrast, the proportion of information-seeking questions was higher in working-class families (62.0%) than in middle-class families (41.1%), $t(69) = 2.52$, $p = .01$, $d = 0.90$. The proportion of rhetorical questions was also higher in working-class families (18.2%) than in middle-class families (5.6%), $t(69) = 2.13$, $p = .04$, $d = 0.76$. These results held with nonparametric Mann-Whitney U Test: Pedagogical questions: $z = 3.45$, $p < .001$; information-seeking questions: $z = 2.18$, $p = .03$; rhetorical questions: $z = 2.39$, $p = .02$.

To separate the effect of context and SES, we performed three stepwise regression analyses, one each for the proportion of pedagogical, information-seeking, and rhetorical questions, with context entered in the first layer and SES entered in the second layer (context was dummy-coded; for SES working-class = 0; middle-class = 1; unspecified cases were excluded). Results showed that SES had an effect in addition to context for pedagogical and information-seeking questions but not for rhetorical questions, pedagogical question: $\Delta F(1, 67) = 10.88$, $p = .002$; information-seeking question: $\Delta F(1, 67) = 6.33$, $p = .01$; rhetorical question: $\Delta F(1, 67) = 0.48$, $p = .5$. After controlling for context, higher SES predicted more pedagogical questions, $\beta = .65$, $p = .002$; less information-seeking questions, $\beta = -.54$, $p = .01$, and did not predict rhetorical questions, $\beta = -.15$, $p = .5$.

The composition of questions was also related to the frequency of questioning. The frequency of questioning was positively correlated with the proportion of pedagogical questions, $r(92) = .26$, $p = .01$.[2] That is, the more questions were asked, the more likely those questions would be

[1] As described in the CHLDES corpus: The majority of working-class samples (58/62) were from a study of low-income families (70% of the sample had 1987 annual household income of < $25,000). The other transcripts came from studies describing their sample as "working-class" and "upper-working-class."

[2] Because the proportion of the three types of questions were not normally distributed, we confirmed these correlations using Kendall rank test. Results were similar: between frequency of questioning and pedagogical questions, $r_\tau = .17$, $p = .02$; information-seeking questions, $r_\tau = -.08$, $p = .27$; rhetorical questions, $r_\tau = -.18$, $p = .02$.

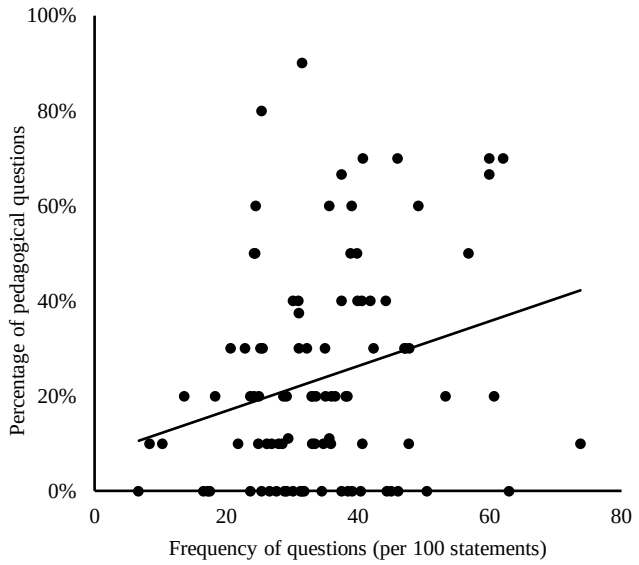


Figure 3. Percentage of pedagogical questions is positively correlated with the overall frequency of questions—pedagogical questions made up a larger proportion for mothers who asked many questions compared to those who asked few.

pedagogical (Figure 3). Frequency was not correlated with the proportion of information-seeking question, $r(92) = -.04$, $p = .7$, and was negatively correlated with the proportion of rhetorical question, $r(92) = -.30$, $p = .003$.[2] The correlation coefficient for pedagogical question was significantly greater than that for both information-seeking and rhetorical questions (pedagogical question vs. information-seeking question: Fisher's $z = 2.03$, $p = .04$; pedagogical question vs. rhetorical question: Fisher's $z = 3.85$, $p < .001$). The same results were found when context was controlled for, with pedagogical questions being more prevalent in proportion as the overall frequency of questions increased $r(91) = .25$, $p = .01$. But, no effect of frequency on proportion of information-seeking questions, $r(91) = -.04$, $p = .7$, and a negative effect of frequency on rhetorical questions, $r(91) = -.30$, $p = .003$. The differences between correlation coefficients also held up when controlling for context (pedagogical question vs. information-seeking question: Fisher's $z = 2.01$, $p = .04$; pedagogical question vs. rhetorical question: Fisher's $z = 3.83$, $p < .001$). These results suggest that the frequency of questioning is differently related to the proportion of each type of question—pedagogical questions made up a greater proportion as the total number of questions increased.

Age changes in questioning

The sample consisted of 42 3-year-olds, 41 4-year-olds, and 11 5-year-olds. The three age groups did not differ on any measurement of questioning behavior, Fs < 1.4, ps > .2. When treated as a continuous variable, age did not correlate with of any measurement of questioning behavior, rs < .16, ps > .13.

Discussion

Educators use questions for which they already know the answer to guide students' learning. Our results suggest that parents do as well. Consistent with research in language learning and academic achievement, we found the composition of mothers' questions differs by context and SES. The proportion of pedagogical questions were higher during free play than during daily routines. Mothers from working-class families asked fewer pedagogical questions than mothers from middle-class families, even though the overall frequency of questioning did not differ. Furthermore, pedagogical questions were more prominent among mothers who asked more questions overall.

We did not observe any developmental changes in the type of questions mothers ask their children. It is possible that this was due to the narrow age range we focused on (3-6 y), and the developmental changes occur either before or after that age range. This would be somewhat surprising given the developmental changes that are going on during this period. In particular, pedagogical questions depend on children's understanding of other people's knowledge states, and 3-5 years of age may be a period of rapid change in children's (explicit) understanding of other's beliefs (Wellman & Liu, 2004). Research on epistemic trust, however, does suggest that children's understanding of ignorance versus knowledge arises somewhat earlier (Nurmsoo & Robinson, 2009), which would be sufficient to differentiate pedagogical from information-seeking questions.

The contextual, individual, and SES differences found in this study may serve as a starting point for investigating the mechanisms behind sociocultural influences on parents' questioning behavior, and their implications on children's learning. The lower proportion of pedagogical questions in low SES settings is surprising. Given that free play was more likely to be associated with pedagogical questions, it is possible that lower SES families do not have as many free play opportunities in their regular daily routines (Snow et al., 1976). However even after controlling for context, we found an effect of SES, suggesting that this difference may be rooted in something deeper than varying degrees of free play in these populations. This points to the importance of investigating the origins and methods of minimizing these differences. Indeed, recent interventions involving parent-child question asking have been shown to significantly influence conversation in low SES families (Ridge, Weisberg, Ilgaz, Hirsh-Pasek, & Golinkoff, 2015). The majority of questions used in Ridge et al's intervention appear to be pedagogical queries, suggesting an important positive route for these kinds of questions in future research.

This study builds upon and extends an accumulating literature on the role of informal pedagogy in children's learning. The results indicate that parents may teach not only through direct instruction, but also through asking questions for which they know the answer. A critical next step, then, is to evaluate children's inferences from these pedagogical questions, and relate that to their learning outcomes. Doing

so will help bridge theories of pedagogy and active learning by considering cases where teaching happens not by giving the evidence itself, but by giving the place to look for evidence.

The sample for our study comes from the CHILDES database, so the scope is limited to information made available, and for the available variables the data were not balanced. Random-assignment experimental work is needed to confirm these results, and to extend them to cover other important factors such as culture. Regardless, these results demonstrate that pedagogical questioning is common in informal pedagogy, and that it varies across individuals and contexts. Thus, we have provided solid evidence that future controlled studies will be investigating questions that are not only of theoretical interest, but are also relevant to understanding children's learning in everyday life.

Acknowledgments

This study is supported by NSF CAREER, DRL-1149116 to P.S. We thank Victoria Golinski & Reham Bader for coding.

References

Black, S. (2001). Ask me a question: How teachers use inquiry in a classroom. *American School Board Journal, 188*(5), 43-45.

Blewitt, P., Rump, K. M., Shealy, S. E., & Cook, S. A. (2009). Shared book reading: When and how questions affect young children's word learning. *Journal of Educational Psychology, 101*(2), 294-304.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*(3), 322-330.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition, 120*(3), 331-340.

Butler, L. P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130*(1), 116-127.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*(4), 148-153.

Dunn, J., Wooding, C., & Hermann, J. (1977). Mothers' speech to young children: variation in context. *Developmental Medicine & Child Neurology, 19*(5), 629-638.

Gall, M. D. (1970). The use of questions in teaching. *Review of educational research*, 707-721.

Garvey, C. (1977). The contingent query: A dependent act in conversation. In M. Lewis & L. A. Rosenblum (Eds.), *Interaction, conversation, and the development of language* (pp. 63-94). New York: Wiley.

Goddard, M., Durkin, K., & Rutter, D. (1985). The semantic focus of maternal speech: A comment on Ninio & Bruner (1978). *Journal of Child Language, 12*(01), 209-213.

Hoff-Ginsberg, E. (1991). Mother-child conversation in different social classes and communicative settings. *Child Development, 62*(4), 782-796.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49*(1), 4.

Levelt, W. (1975). What became of LAD. In W. Abraham (Ed.), *Ut videam: contributions to an understanding of linguistics, for Pieter Verburg on the occasion of his 70th birthday* (pp. 171). Lisse: Peter de Ridder Press.

Nurmsoo, E., & Robinson, E. J. (2009). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. *Child Development, 80*(1), 23-27.

Olsen-Fulero, L., & Conforti, J. (1983). Child responsiveness to mother questions of varying type and presentation. *Journal of Child Language, 10*(03), 495-520.

Ridge, K. E., Weisberg, D. S., Ilgaz, H., Hirsh-Pasek, K. A., & Golinkoff, R. M. (2015). Supermarket Speak: Increasing Talk Among Low‐Socioeconomic Status Families. *Mind, Brain, and Education, 9*(3), 127-135.

Sage, K. D., & Baldwin, D. (2010). Social gating and pedagogy: Mechanisms for learning and implications for robotics. *Neural Networks, 23*(8), 1091-1098.

Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance. *Journal of Research in Science Teaching, 27*(2), 127-144.

Snow, C. E., Arlman-Rupp, A., Hassing, Y., Jobse, J., Joosten, J., & Vorster, J. (1976). Mothers' speech in three social classes. *Journal of Psycholinguistic Research, 5*(1), 1-20.

Snow, C. E., Dubber, C., & De Blauw, A. (1982). Routines in mother-child interaction. In L. Feagans & D. C. Farran (Eds.), *The language of children reared in poverty* (pp. 53-72). New York: Academic Press.

Tomasello, M. (1999). *The cultural origins of human cognition*: Harvard University Press.

Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science, 321*(5897), 1831-1834.

Vredenburgh, C., Kushnir, T., & Casasola, M. (2014). Pedagogical cues encourage toddlers' transmission of recently demonstrated functions to unfamiliar adults. *Developmental Science*.

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development, 75*(2), 523-541.

Wilcox, M. J., & Webster, E. J. (1980). Early discourse behavior: An analysis of children's responses to listener feedback. *Child Development, 51*(4), 1120-1125.

Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching, 20*(5), 419-435.

Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech

Fatemeh Torabi Asr
Cognitive Science Program
Indiana University, Bloomington
fatorabi@indiana.edu

Jon A. Willits
Department of Psychology
University of California, Riverside
jon.willits@ucr.edu

Michael N. Jones
Psychological and Brain Sciences
Indiana University, Bloomington
jonesmn@indiana.edu

Abstract

Distributional Semantic Models have been successful at predicting many semantic behaviors. The aim of this paper is to compare two major classes of these models – co-occurrence-based models, and prediction error-driven models – in learning semantic categories from child-directed speech. Co-occurrence models have gained more attention in cognitive research, while research from computational linguistics on big datasets has found more success with prediction-based models. We explore differences between these types of lexical semantic models (as representatives of Hebbian vs. reinforcement learning mechanisms, respectively) within a more cognitively relevant context: the acquisition of semantic categories (e.g., *apple* and *orange* as fruit vs. *soap* and *shampoo* as bathroom items) from linguistic data available to children. We found that models that perform some form of abstraction outperform those that do not, and that co-occurrence-based abstraction models performed the best. However, different models excel at different categories, providing evidence for complementary learning systems.

Introduction

Distributional models of lexical semantics have had a large impact on cognitive science over the past two decades. In general, these models formalize the distributional hypothesis (Firth, 1957; Harris, 1970), and attempt to learn distributed representations for word meanings from statistical regularities across a large corpus of linguistic input. The resulting representations have been enormously valuable to researchers wanting to select and calibrate word stimuli balanced on semantic dimensions. They have also been successfully used as semantic representations in models of cognitive processes (e.g., word recognition, reading), and in a wide variety of applications ranging from automated tutoring to open question answering.

Due in part to their practical successes, the algorithms that distributional models use to build semantic representations have also been hypothesized to be related to the cognitive mechanisms potentially used by humans to learn semantic representations from regularities in their language input[1]. The various learning mechanisms posited include simple co-occurrence learning, episodic abstraction, reinforcement learning, and probabilistic inference (see Jones, Willits, & Dennis, 2015, for a review).

The majority of distributional models in the cognitive science literature are from the family of *co-occurrence models*. Models of this family tend to apply unsupervised learning mechanisms to a frequency count of how often words co-occur with each other in a context (paragraph, document, or an *n*-word moving window). There are many specific models that differ in theoretically meaningful ways in terms of the learning mechanisms they apply, but all members of this family share the assumption that the learner is basically counting observed co-occurrences of stimuli in the environment. Hence, they are all based on error-less Hebbian-type learning mechanisms.

Perhaps the best-known co-occurrence model in the cognitive literature is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA basically applies a dimensionality reduction mechanism to a sparse word-by-document frequency matrix computed from a large corpus of text. The resulting dense vectors emphasize higher-order statistical relationships among words: Two words that occur in similar contexts across language will have similar representations, even if they never directly co-occurred in a context. LSA vectors have been used to model how words' semantic similarity predicts performance in vocabulary acquisition, categorization, reading, summarization, and a number of other cognitive tasks (see Landauer, 2006, for a review). Other popular models count the co-occurrences of words within a moving window, or how frequently words occur with predefined context words. For a review of the various co-occurrence algorithms and their performance on semantic tasks, we refer the reader to Bullinaria and Levy (2007) or Riordan and Jones (2011).

A second family of *predictive* distributional models has been around for decades, based on principles of predictive encoding and error-driven learning core to theories of reinforcement learning. For example, early recurrent neural networks studied by Elman (1990) and St. John and McClelland (1990) learn distributed representations for a word's co-occurrence history across their hidden layers. Feedforward networks studied by Rogers and McClelland (2004) similarly learn distributed semantic representations, representations that even contain hierarchical taxonomic relations. These model architectures rely on predicting a distributed set of features for each input word, then deriving an error signal from the difference between the prediction

[1] To be fair, the transfer between practical algorithm and human mechanism has been bidirectional: Knowledge of how humans learn was used to inform the design of distributional algorithms, and the subsequent success of certain algorithms on practical tasks has then fed back to help narrow the range of likely cognitive mechanisms that humans use.

and the observed context. Errors are backpropagated through the network to increase the likelihood that the correct output will be predicted given the input in the future.

The mechanisms for predictive learning core to these models have been studied a great deal within reinforcement learning, and there is a considerable literature in cognitive neuroscience exploring how this type of learning is driven at the neural level by dopaminergic error signals, and is moderated by the basal ganglia, cingulate, and hippocampus. In practice, these predictive models have only been applied to small toy datasets to discover distributional structure; their capabilities have not been studied when applied to large, naturalistic corpus-based materials such as those that are used by simple co-occurrence models. However, there has been a recent resurge of interest in predictive models of distributional semantics. Howard et al., (2011) trained a predictive version of the Temporal Context Model (pTCM), a recurrent model of error-driven hippocampal learning, on a large text corpus and demonstrate impressive performance on word association tasks. Similarly, Shaoul et al. (2016) used a Naïve Discrimination Learning (NDL) procedure (a single-layer network trained using the Rescorla-Wagner learning procedure) to show that semantic representations can be learned from a relatively small sample of spoken language. Despite the many appealing properties of predictive distributional models, the bottom line is that they are heavily outperformed by co-occurrence models at accounting for human data on every lexical semantic task that has been tested.

A new type of context-prediction model (a.k.a. "neural embedding model") has emerged in the past few years; as Baroni et al. (2014) put it, these models are "the new kids on the distributional semantics block." The predictive model of Mikolov et al. (2013), referred to in the literature as *word2vec* (W2V), is a feedforward neural network model with a hidden layer that uses error backpropagation to maximize the likelihood of either predicting context given a word, or predicting a word given the context. In this sense, W2V behaves much like the networks studied by Rogers and McClelland (2004), but applied to massive amount of linguistic data, and with some tricks to improve training efficiency. W2V has made a huge stir in the machine learning literature for its ability to outperform every other semantic model on benchmark tasks, and to achieve this impressive performance using an architecture that had been written off in the cognitive and linguistic literatures. Baroni et al. (2014) undertook a careful comparison of state-of-the-art co-occurrence models and W2V, testing them on the same input corpus and with a large battery of different semantic tasks. They concluded that the hype surrounding W2V is warranted: Even under these very well controlled comparisons, W2V outperformed the current top performers studied by Bullinaria and Levy (2007). Since W2V has a similar architecture to many "toy" connectionist models that have been popular in cognitive science, its success on practical tasks is exciting to the field. However, it is important to note that current tests of W2V have been on very large corpora—the tests by Mikolov et al. (2013) were trained on over 1-billion words of text, and the comparisons by Baroni et al. (2014) were trained on a corpus of almost 3-billion words. While estimates of the number of tokens a human will read/hear in a lifetime vary greatly, both of those corpora are orders of magnitude beyond the upper limit of a single human's experience.

Hence, our approach here is to scale the problem way down. We train co-occurrence and predictive models on the real-world speech that children experience from birth to age 5 using the CHILDES corpus (MacWhinney, 1998), a resource used in previous work on computational modeling of early word categorization (e.g., Asr et al., 2013). This linguistic experience is a very different test for the models, and allows us to explore how their learning mechanisms might deal with the noisy data on which children must build their semantic representations. Given the superiority of W2V over co-occurrence models on large data and practical semantic tasks, and the similarity of its learning algorithm to popular error-driven models of development, how do the two families of models compare on their ability to learn complex structure on the same impoverished data that children receive?

Co-occurrence Models

To evaluate the performance of co-occurrence based models, two models were selected that have both performed well in previous evaluations, but that use different learning algorithms (thus providing breadth of coverage of different types of co-occurrence based models). One model uses an abstraction mechanism, whereas the other operates with simple summation of surface-level word co-occurrences.

PCA-Based Vector Model

The first model tested is notable for its use of principle components analysis (PCA) as the primary method of knowledge abstraction. This model computed co-occurrences in a 12-word moving window (12 words in both the forward and backwards directions) for the 10,000 most frequent words in the corpus, resulting in a 10,000-by-10,000 co-occurrence count matrix. These values were then normalized into positive point-wise mutual information values (Bullinaria & Levy, 2007). This matrix was then reduced using PCA, and the first 30 principle components were retained, resulting in 30-element vectors for each of the 10,000 words. This model is composite of several pre-existing models, such as the HAL model (Lund & Burgess, 1996), the COALS model (Rohde et al., 2005) and models by Bullinaria & Levy (2007).

Sparse Random Vector Accumulator

The second co-occurrence model we tested was from the family of Random Vector Accumulators (RVAs; see Jones, et al., 2015, for a review). RVAs are essentially distributed count models. They initialize a unique random vector for each word prior to learning, and the word's memory vector is then updated across learning as the sum of the vectors representing the words with which it has co-occurred.

Hence, RVAs are incremental learners with no abstraction mechanism (compared to PCA). We use a sparse-distributed RVA here (Recchia et al., 2015). To equate with the other models, the RVA was trained on the same 10k x 10k word co-occurrence matrix as was the PCA model. Each word was initially represented by an 8,000-element sparse ternary environment (E) vector. The memory (M) vector was then the frequency-weighted sum of the E vectors for all the words with which the target word co-occurred within a 12-word context window. For example, if *dog* had a co-occurrence frequency of 3, 1, 0, 2, with *cat, shoe, bunny*, and *run*, respectively, then $M_{dog} = 3*E_{cat} + 1*E_{shoe} + 2*E_{run}$.

Predictive Models

To compare to the RVA and PCA co-occurrence models, we tested two error-driven models. The primary objective was to evaluate the performance of the W2V predictive learning algorithm, but we also tested a second model of the error-correction family based on the classic Rescorla-Wagner model of discrimination learning.

W2V

As described above, W2V is a multilayer neural network (with an input layer, an output layer, and one hidden layer) that learns word vectors by iterating through sample contexts. W2V comes with two slightly different architectures of a neural network for learning word embeddings: (1) *cbow* (context bag-of-words) and (2) *skipgram*. In the cbow architecture, a word is predicted as the output from its context input. During training, the input layer produces a weighted sum over the context words within a fixed adjacency window of the target word. Output activations are converted to a probability distribution over the vocabulary (softmax) and the weight matrices are updated through backpropagation of the errors. The network topology is similar for skipgram, except that in this architecture, a target word is used to predict the context in which it appears. In skipgram, several context vectors can be sampled from a certain window of adjacent words (e.g., given the input sentence "she found a cute cat in the garden", the target word "cat" can be used to predict context unigrams "cute", "in", "a" and "the" when a window of size two (from each side) is considered. After training the model on a text corpus, the weight matrix between the hidden layer and the output layer in cbow or the one between the input layer and the hidden layer in skipgram represents the embeddings for the vocabulary words: V * N, where V is the size of vocabulary and N is the dimension of word vectors. In our experiments, we used a python implementation of the W2V model from Gensim (Řehůřek & Sojka, 2010).

Naïve Discrimination Learner

Like W2V, the Naïve Discrimination Learning (NDL) model learns to make predictions about a target word given its lexical context, or vice-versa. Unlike W2V, NDL has only an input layer and an output layer (with no hidden layer), and uses the Rescorla-Wagner learning procedure (Rescorla & Wagner, 1972) to learn a set of weights between target words and its cues (i.e. the other words in the window) that predict it. In addition to a long history of use in the psychology of learning (see Miller et al., 1995, for a review), NDL models are now being explored as models of infant word segmentation and how children learn the meanings of words (Baayen et al., 2015).

Method

Corpus

We used the entire child-directed speech data in the American English subset of the CHILDES corpus (MacWhinney, 1998). This collection includes conversations between children (4 to 60 months of age) and their parents, care-givers and other children. Utterances directed to the target children were combined to create a corpus representative of the linguistic input of children from these ages. The resulting corpus consisted of 4,568 sub-corpora (transcribed documents), containing 36,170 distinct word types and 8,323,266 total word tokens. The corpus is relatively well-distributed across ages and generally forms a decent snapshot of the input children get at various ages. Little pre-processing was done to the corpora beyond simple word tokenization. To equate comparisons across the models, from this corpus, we selected the 10,000 most frequently occurring words and used only those words as inputs into the four models. Words below this rank were excluded due to their low frequency in the corpus (<7).

Evaluation Task

We evaluated the performance of the models based on a word categorization task. For this task, we used 1,244 high frequency nouns from the corpus that unambiguously belonged to a set of 30 categories (like *mammal*, *clothing*, etc.). Thus, each non-identity pair of words either belonged to the same category (as in *dog-cat* for mammals and *shoe-sock* for clothing) or to different categories (as in *dog-shoe*, *dog-sock*, *cat-shoe*, and *cat-sock*).

The category membership of each pair was predicted using each model's similarity score in a signal detection framework. For each word pair, the similarity score was compared to a decision threshold. If the similarity score was above that threshold, the pair was predicted to belong to the same category (classified a "*hit*" if this prediction was correct and a "*false alarm*" if this prediction was wrong). If the similarity score was below the threshold, the pair was predicted to belong to different categories (classified a "*correct rejection*" if this prediction was correct and a "*miss*" if this prediction was wrong). For each model, a single decision threshold was chosen, the threshold that maximized categorization accuracy for that model. Each model's overall performance was assessed by computing balanced accuracy (BA) using the formula below:

$$BA = \frac{1}{2} * \left(\left(\frac{hits}{hits + misses}\right) + \left(\frac{correct\ rejections}{correct\ rejections + false\ alarms}\right)\right)$$

Experiment 1

Our first experiment focused on exploring the parameter space of the W2V model to see how an error-driven

distributional model learns from the specific characteristics of child-directed speech data: short sentences, simple structure, less ambiguity, etc.

W2V Setup

The skipgram architecture is known to perform better on smaller corpora and modeling rare words due to repeated sampling from a fixed window of context words, whereas cbow is trained faster, thus is more suitable for learning word embeddings from big text corpora. We tried both architectures in this experiment to see if this held true in our task. We also manipulated two other parameters in the model: the window size for collecting context words, and the hidden layer size, i.e., dimensionality of the resulting word embeddings. We examined window sizes of 2 and 12 words from each side of the target word. Previous experiments showed that context words in a smaller window convey syntactic information about the target word, whereas, context words in a larger window convey more topical information (Levy & Goldberg, 2014a). Vector dimensionality (size of hidden layer) was set to 30, 50, 100, 200 or 300 nodes. Larger hidden layers provide a finer-grained distributional representation, and thus can be beneficial in learning word similarities to perform a categorization task. In contrast, smaller hidden layer sizes force the network to develop more abstract representations of meaning. Frequency cut-off was set to 1. Other parameters in the W2V training function were held constant across the experiments (for a complete description of the default parameter values please check the Gensim package).

Results

Figure 1 shows the accuracy of the model with different parameter considerations in the categorization task. As expected, the skipgram architecture learned similarity between the words better than the cbow architecture, resulting in superior categorization performance. Interestingly, considering a larger context window size (changing from 2 to 12 context words from each side of the target word) did not enhance the performance of the cbow model significantly. On the other hand, the skipgram model did benefit from more context words. When compared against the cbow results, this suggests that the size of our corpus is small for the connectionist model to learn the similarities between the words in our categorization dataset. That is why sampling smaller *n*-grams from a bigger window of adjacent words (the mechanism used only in the skipgram model) results in a significant boost in the backend performance. This observation is in line with previous experiments on large corpora (Mikolov et al. 2013) where the skipgram model outperformed cbow in semantic association tasks. The cbow model, on the other hand, recognizes syntactic associations (plural/singular) slightly better. This can be due to the way every training instance is used in either model: in skip-gram, unigrams are sampled from the context window and used one-by-one together with the target word in several training steps, whereas cbow uses the entire context frame of a word at once.

The second observation concerned the effect of vector dimensionality. While adding nodes to the hidden layer increased the categorization accuracy in lower scales (i.e., from 30 to 50) the effect disappeared in higher scales (100 to 300). In fact, the best performing model overall is the skipgram with 200 hidden layer nodes (accuracy = 74.9%), not the one with 300. This can be due to the small and relatively less ambiguous vocabulary in the child-directed corpus, for which a smaller hidden layer would suffice and might generalize better to capture similarities between words within the coarse-grained categorization defined in our evaluation task.

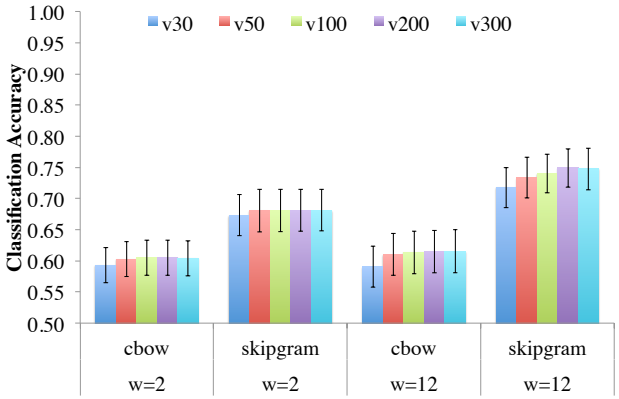


Figure 1. Mean classification accuracy (averaging across the 30 semantic categories, error bars represent 95% CI) of the various W2V models, as a function of the window size (2 vs. 12), the size of the hidden layer vector (30, 50, 100, 200, 300), and architecture (cbow, skipgram).

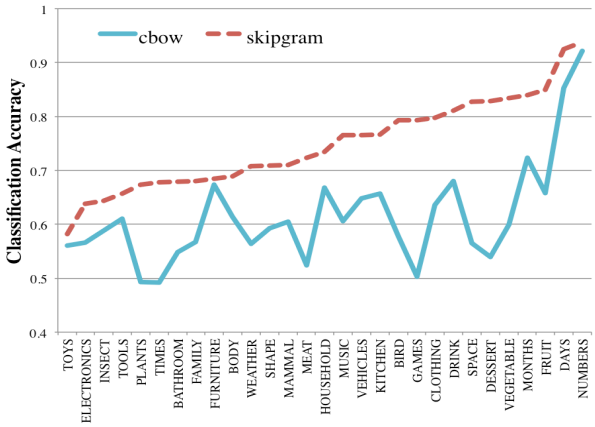


Figure 2. Mean classification accuracy on the 30 different categories, comparing the best performing cbow model to the best performing skipgram model (in both cases, w=12, v=200).

The best performing W2V setup will be used in our next experiment in comparison to other models. Thus before we proceed, it is good to take a closer look at the learning of each word category by either of the architectures. Figure 2, compares the best cbow and skipgram models on every category within the evaluation dataset. First, the two models seem to be learning qualitatively different types of information: cbow captured the within-category similarity of furniture and household items relatively better than it did on dessert, fruit, vegetable and meat items. In contrast,

household and furniture were among the less accurate categories when detected by the skipgram model. Both models performed very well in categorization of numbers, days and months, but had difficulty with other categories such as plants. This suggests that the distributional properties of different semantic categories might be different. A correlation analysis of system performance across categories with *token frequency*, *word types* and *entropy* of the category items revealed that cbow learned categories with larger token frequency better ($p<0.05$), while the performance of skipgram on a category was statistically independent of these factors.

Experiment 2

In the second experiment we compare different distributional models in the word categorization task.

Setup of the Models

We trained each model with its optimal parameter settings. As we found in Experiment 1, for W2V this meant using the skipgram architecture with a window size of 12 and 200 nodes for the hidden layer. For comparison, the optimal performing NDL model had a window size of 3, and like W2V, performed better predicting contexts from words rather than vice-versa (Shaoul et al., 2016). NDL has no hidden layer, and therefore is not comparable in terms of vector size. Since the co-occurrence models do not use error-driven learning, direction (word predicting context vs. context predicting word) was not a feature of these models. However, these models were both trained using a window size of 12 (making them comparable to the better performance of W2V). The PCA model has a parameter equivalent to the number of hidden units in W2V (i.e. the amount of abstraction that is used): the number of principle components retained. The peak performance of this model was obtained with 30 principle components. The RVA model also has a parameter, which is the size of the random environment vectors. The peak performance of the RVA model was obtained with a random vector size of 8000.

Results

Figure 3 compares the performance of each of the four models in the word categorization task. The two co-occurrence models, PCA and RVA performed very differently. The PCA model had the advantage of learning latent relationships in the dataset, which likely helped it perform better on a small dataset such as the CHILDES corpus. Between the two predictive models, W2V was superior to the NDL. As with the PCA model, W2V's hidden layer allows it learn abstract, latent relationships from the co-occurrence data. In machine learning, introducing a hidden layer to a neural network topology is considered as a method for capturing nonlinear correlations between the input and output of the model. In addition to NDL missing the hidden layer, the input to this model was different from that of the W2V skipgram used in this experiment. In fact, like the cbow model tested in our previous experiment, NDL did not benefit from larger context windows (thus its peak performance is that of it being trained on 3 context words from each side).

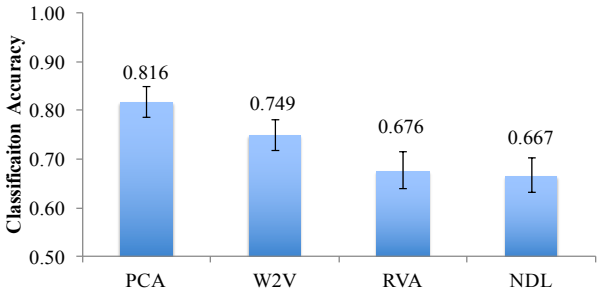


Figure 3. Mean classification accuracy (averaging across the 30 different categories and 95% CI) for the four DSMs.

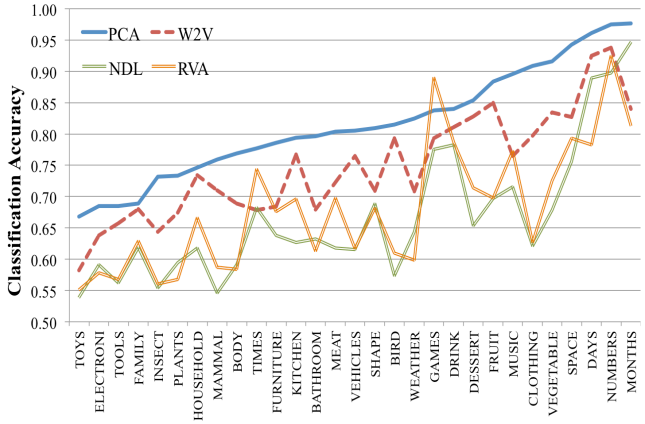


Figure 4. Mean classification accuracy on the 30 different categories for the four DSMs.

Finally, a comparison between the overall performances of the best co-occurrence model (PCA) and the best predictive mode (W2V) in this task revealed the superiority of the co-occurrence model. PCA includes a relatively sophisticated post-processing of the raw co-occurrence vectors, whereas W2V involves an error-based learning. Inputs to the both models were very similar, that is the co-occurrence information was collected from a window size of 12 from each side of a target word. As we mentioned above, the number of principal components in the PCA was a good equivalent to the size of the hidden layer in the W2V. If we compare it against the skipgram with 30 nodes in the hidden layer (in our previous experiment), we find an even bigger difference in the performance of the two models.

How models differ from one another qualitatively? Figure 4 provides an answer to this question by showing performance on each category separately. The models showed some variance in terms of what categories they learned better and what categories they learn worse. The less accurate models (NDL and RVA) tended to align fairly highly, doing relatively well and poorly on the same categories, and showing no significant difference in overall performance ($t(29) = 1.48$, $p = 0.15$). W2V outperformed both NDL and RVA, with a significantly higher overall performance ($t(29) = 4.20$, $p < 0.001$), and beating NDL on 28/30 categories and RVA on 27/30 categories. The PCA model performed the best, having a significantly higher

overall performance than all three models (all p's < .0001), and beating W2V on 30/30 categories, RVA on 29/30 categories, and NDL on 30/30 categories. Accuracy of a category did not correlate with its types or token frequency.

General Discussion

Our analyses compared two models that involved the abstraction of latent representations (PCA and W2V) versus two models that do not perform abstraction (RVA and NDL) when trained on child-directed speech. Both models that perform abstraction strongly outperformed those that do not. Our analyses also compared two co-occurrence models (PCA and RVA) with two predictive models (W2V and NDL). In previous work, W2V has been demonstrated to surpass standard co-occurrence models when trained on large amounts of data (e.g., Baroni et al., 2014; Mikolov, 2013). However, we found an interesting paradox: W2V was actually outperformed by a rather simple PCA-based co-occurrence model when applied to child-directed speech and with a classification of children's concepts. It is important to note that our PCA model was also in the list of models in Baroni et al. that were compared to W2V. Hence, the story isn't as simple as saying one model is "better" than the other. Despite the equivalence of the objective function optimization in W2V to the matrix factorization process of the PCA model (as pointed out in Levy & Goldberg 2014b), W2V involves an incremental error-driven learning process. It operates very well and efficiently with large amounts of data; also might be considered as a better simulation of the cognitive processes. On the other hand, our experiments show that PCA performs better on small and sparse linguistic data that children learn from.

Future work needs to focus on correlating the models' predictions with children's response data. In addition, our target words were all from the same syntactic category (nouns). Different distributional model architectures have been shown to vary in the type of information they learn best (e.g., syntactic vs. semantic association; Mikolov et al., 2013, or different types of semantic relations; Baroni et al., 2014; Jones, Kintsch, & Mewhort, 2006). Our classification task was one that required the models to learn paradigmatic similarity well. Interestingly, in the survey of Baroni et al. (2014), paradigmatic tasks were among the cases where W2V did not outperform the co-occurrence models. Hence, our results may point to the benefit of having complementary learning systems when constructing a semantic representation: An error-driven predictive mechanism and a mechanism applied to direct co-occurrences. There is no reason that the two learning mechanisms need to be mutually exclusive—there is a great deal of evidence that humans use both Hebbian and error-driven learning (e.g., O'Reilly, 1998). It is a reasonable presumption that perhaps the two systems work together to construct semantic memory from episodic experience.

Acknowledgement

This research was funded by grant R305A140382 from the Institute of Education Sciences.

References

Asr, F. T., Fazly, A. and Azimifar, Z. (2013). From cues to categories: a computational study of children's early word categorization. *Cognitive Aspects of Computational Language Acquisition*. (pp. 81-103).

Baayen, R. H., Shaoul, C., Willits, J. and Ramscar, M. (2015). Comprehension without segmentation: A proof of concept with naive discrimination learning. *Language, Cognition, and Neuroscience*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL* (pp. 238-247).

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510-526.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. *In Studies in Linguistic Analysis*, (pp. 1-32).

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics* (pp. 775–794).

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science*, *3*, 48-73.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55, 534-552.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232-254

Landauer, T. K. (2006). Latent semantic analysis. *Encyclopedia of Cognitive Science*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Levy, O. and Goldberg, Y., (2014a). Dependency-Based Word Embeddings. In *ACL (pp. 302-308)*.

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems (pp. 2177-2185)*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

MacWhinney, B. (2000). *The CHILDES project: The database (Vol. 2)*.

Miller, R. R., Barnet, R. C., and Grahame, N. J. (1995). Assessment of the rescorla-wagner model. *Psychological Bulletin*, 117(3). 363-386.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR*.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in cognitive sciences*, *2*, 455-462.

Recchia, G. L., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. *Computational Intelligence & Neuroscience*.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64-99.

Riordan, B., & Jones, M.N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *TopiCS*, *3*, 303-345.

Řehůřek, R., & Sojka, P. (2011). Gensim - Statistical Semantics in Python.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627-633.

Shaoul, C., Willits, J. A., Ramscar, M. Milin, P., & Baayen, R. H. (2016). A discrimination-driven model for the acquisition of lexical knowledge in auditory comprehension. *Manuscript under review*.

St. John, M., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217-257.

Geometric representations of evidence in models of decision-making

Peter D. Kvam (kvampete@msu.edu)
Department of Psychology, Michigan State University
316 Physics Rd, East Lansing, MI 48824 USA

Abstract

Traditionally, models of the decision-making process have focused on the case where a decision-maker must choose between two alternatives. The most successful of these, sequential sampling models, have been extended from the binary case to account for choices and response times between multiple alternatives. In this paper, I present a geometric representation of diffusion and accumulator models of multiple-choice decisions, and show how these can be analyzed as Markov processes on lattices. I then introduce psychological relationships between choice alternatives and show how this impacts the sequential sampling process. I conclude with two examples showing how one can predict distributions of responses on a continuum as well as response times by incorporating psychological representations into a multi-dimensional random walk diffusion process.

Keywords: random walk; decision making; evidence accumulation; multi-alternative; continuous response

Introduction

Traditionally, models of the decision-making process have focused on the case where a decision maker has two options between which to choose. By far the most empirically successful accounts of this process are sequential sampling models, which are able to reproduce decision makers' observed choice proportions and distributions of response times with high fidelity.

Sequential sampling models can be broadly broken down into two categories: diffusion models and accumulator models. The former type of model posits that a person represents the information or evidence they have regarding the decision as a balance between the two choice alternatives. In inferential decisions, this balance can be formally represented as the log odds of one alternative (hypothesis) relative to another, requiring that the sum of the log odds for all hypotheses sum to zero (i.e. so that the probabilities of all hypotheses sum to one). A decision is triggered when the log odds for one hypothesis over the others exceeds a critical threshold value.

Accumulator models, by contrast, represent information or evidence for the alternatives as separate quantities [accumulators]. As a decision maker gathers or receives information, each accumulator is updated, and a decision is triggered once one of the accumulators reaches a critical threshold value. However, each of the accumulators is usually assumed to be independent such that incrementing one's value does not affect another's.

More recently, these models have been extended to account for decisions between multiple alternatives. In the case of models where evidence for alternatives is represented as multiple independent accumulators (one may think of these as 'pure' accumulator models), this is straightforward – one need only add an additional accumulator for each additional choice alternative. A decision is still triggered once any of these accumulators reach a critical threshold, so this addition can be done ad infinitum, or at least until a modeler runs out of computational resources.

For diffusion models, adding alternatives is only slightly more complicated. In a 3-alternative case, rather than decreasing the log odds of B by 1 every time the log odds of A increase by 1, one would have to decrease the log odds of B and C each by 1/2 when the log odds of A increase by 1. This ensures that the log odds across all hypotheses sums to zero / probabilities sum to one. As more alternatives are added, this decrement decreases so that incrementing evidence for a single alternative by 1 also decrements all other alternatives by $\frac{1}{n-1}$. The same decision rule applies – once the (log odds) evidence in favor of one alternative exceeds a critical threshold, a decision is triggered.

Although assuming independence of accumulators or at least a uniform distribution of negative evidence are convenient simplifying assumptions, they are perhaps unrealistic. A substantial body of literature has suggested that there are often strong and unbalanced interactions between different pairs of items in a set. For example, context effects arising from the inclusion of a third alternative – such as decoy, compromise, and similarity effects – substantially alter choices between an original set of two (see e.g. Trueblood et al., 2014). Similarly, in absolute identification tasks, adjacent categories (e.g. 50-60 and 60-70) interact more strongly than non-adjacent ones (50-60 and 80-90) (see Brown et al., 2008).

Models of decision making have been modified in a number of ways to account for these phenomena. For example, decision field theory (Busemeyer & Diederich, 2002) introduced an additional step in the decision process where pairs of items are contrasted against one another before computing accumulator values. The leaky competing accumulator model (Usher & McClelland, 2001, 2004) introduced competition and loss aversion to a similar effect, the multi-attribute linear ballistic accumulator model (Trueblood et al., 2014) includes pairwise comparisons as well as subjective attribute values, and the selective attention, mapping, and ballistic accumulation model (Brown et al., 2008) specifies and utilizes adjacency between categories to define the evidence accumulation process for separate accumulators. Similarly, models of confidence (see Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009) specify adjacency of judgment categories using ordinal states or accumulators.

What all of these models have in common is that they specify *psychological relationships* between alternatives in a choice set. Indeed, Trueblood et al. (2014) note explicitly that decisions are made between the psychological represen-

tations of alternatives rather than physical ones, and that this is the key component allowing the different models to account for context effects. These models each add components in order to avoid making the simplifying assumptions of independence and uniform negative evidence. Instead, they suggest that evidence for alternative A may also be evidence for alternative B and simultaneously be strong evidence against alternative C.

In the next section, I examine how the psychological relationships between alternatives can be thought of as geometric relations in a psychological space.

Geometry of evidence accumulation

In order to introduce a geometric representation of the decision process, I return first to the simple binary cases. Using the binary choice diffusion model, one can establish a rule to construct models relating accumulator and diffusion models, construct geometric models of evidence accumulation among multiple alternatives, and in turn derive a method for modeling evidence accumulation when the number of alternatives is very large or continuous (infinite).

Equal relative evidence models

The basic uni-dimensional diffusion model (Ratcliff & McKoon, 2008) originally described the behavior of particles along a single dimension in space. Its natural geometric analogue is a random walk on a line (Figure 1A). In inference tasks, this single dimension may correspond to the log odds of one hypothesis (H_1) relative to another (H_2).

Put simply, the closer a person's representation of evidence (or preference) is to a boundary corresponding to a choice alternative, the more they currently favor that alternative. Evidence that provides support for an alternative H_1 moves a person's state orthogonally toward the boundary corresponding to H_1 in direction D_1 and away from the boundary H_2, which is in direction D_2. This gives rise to a spatial relationship between the amount of evidence (Ev) a particular piece of evidence supporting one hypothesis H_1 provides for another hypothesis H_2:

$$Ev(H_2) \leftarrow Ev(H_1)cos(\angle D_1 D_2) \quad (1)$$

In the two-alternative diffusion model, the evidence favoring alternative H_1 must move the state in the opposite direction from evidence favoring alternative H_2, so that $cos(\angle D_1 D_2) = -1$.

In order to obtain the case for three alternatives, where evidence for H_1 decreases the log odds of H_2 and H_3 equally, it must be the case that $cos(\angle D_1 D_2) = cos(\angle D_1 D_3) = -\frac{1}{2}$ (and of course it will also be the case with $cos(\angle D_2 D_3)$ to conserve total log odds).[1] This results in an evidence accumulation process that unfolds in a plane, contained within an equilateral triangle (Figure 1B). A decision is triggered when a state crosses one of the sides of the triangle, each of which corresponds to choosing one of the alternatives (see Laming, 1968, for a similar proposal).

Extending this strategy to model decisions between any number of alternatives is relatively straightforward. In order to account for decisions where there are n alternatives, one must create a situation where there are n directions $\{D_1, D_2, ..., D_n\}$ satisfying the property $cos(\angle D_i D_j) = -\frac{1}{n-1}$, $i \neq j$, so evidence for any alternative provides evidence against all others equally. In the case of 4 alternatives, the boundaries corresponding to $H_1 - H_4$ would each be a plane in a 3-dimensional space, together forming a tetrahedron (Figure 1C), and evidence accumulation will unfold in a 3-dimensional space.

In order to accommodate n alternatives, this would naturally be extended to permit evidence accumulation in $(n-1)$ dimensions. The state would exist in the interior of a simplex, with the choice boundaries corresponding to each of its $(n-2)$-dimensional facets.

It is worth a note that the cosine relation in Equation 1 will preserve log odds in any n-dimensional space by virtue of every integral $\int \cdots \int_0^{2\pi} cos = 0$. However, later examples where decision bounds do not form regular figures show that log odds are preserved across the *theoretically possible* space of alternatives but not necessarily across all available ones.

Absolute evidence models

While it is often practically unnecessary to envision accumulator models in a geometric way, doing so illustrates the psychological assumptions that go into these models and allows them to be analyzed as a random walk. In the two alternative case, evidence in favor of H_1 provides no information regarding H_2. Using the relationship defined by equation 1, this means that the directions corresponding to each alternative must be orthogonal, $cos(\angle D_1 D_2) = 0$. The choice boundaries therefore form two sides of a rectangle (Figure 1D).

However, the evidence state does not immediately have a clear log odds interpretation as it did in the diffusion models. One could potentially address this by assuming that there are two theoretical alternatives in directions $-D_1$ and $-D_2$ (if D_1 and D_2 are given by vectors) and anchor log odds to be zero at the initial state. This would allow computation of relative log odds of the hypotheses in the case a person wanted to make a relative judgment of the two alternatives (e.g. preference or confidence). However, doing so is not necessary for predicting choices and response times.

Extending accumulator models to three or more alternatives is relatively straightforward. One need only add additional, orthogonal dimensions to the evidence accumulation space, then set the choice criterion and new direction D_n for each new alternative as orthogonal to existing ones. In the case of three alternatives, this would yield a figure bounded by three choice criteria constituting sides of a cube (Figure 1E). For n alternatives, the orthogonal choice criteria would compose a set of intersecting facets of an n-cube.

[1] Note that this approach offers an alternative solution to the relative-accumulator problem encountered by Nosofsky (1997), where there was more negative evidence than positive evidence added across accumulators if increments and decrements were restricted to values of one.

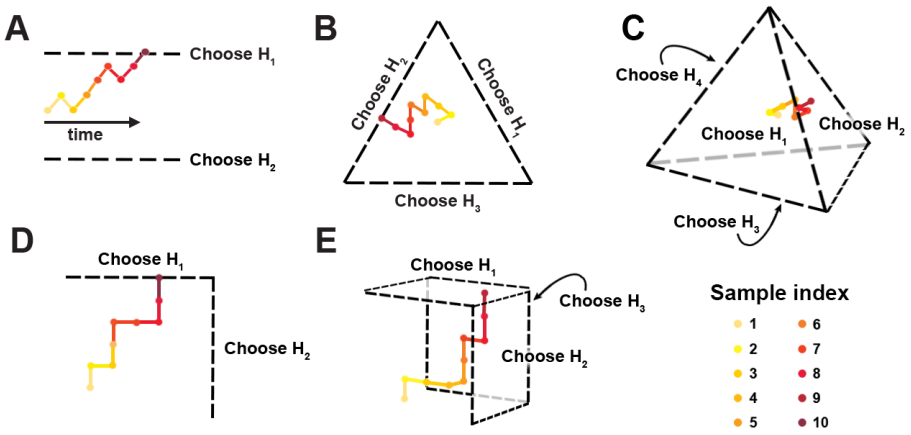


Figure 1: Representation of a person's accumulated evidence and choice criteria for 2-, 3-, and 4-alternative diffusion processes (A, B, C) and 2- and 3-alternative accumulator processes (D, E). An alternative is chosen when a person's represented evidence (yellow / red) crosses the corresponding edge (for models A, B, and D) or face (for models C and E).

Random walks

This geometric framework lends it self to both discrete-time and continuous-time as well as discrete-space and continuous-space random walk representations. A discrete-space structure can be given to the models I have described so far by imposing a lattice structure upon them.

For example, suppose we are interested in using an equal relative evidence diffusion model to describe how a person sorts a color stimulus into three categories: red, blue, or green. This corresponds to the triangular structure shown in Figure 1B. In order to produce a discrete random walk in this space, one can construct a triangular lattice bounded by the three choice criteria.

In this case, a person's internal representation of the stimulus in terms of the log odds of the three hypotheses (red, green, blue) corresponds to their position on the lattice. Initially, they might start out in the middle (unbiased / 0), but as they view the stimulus they should sample pieces of evidence that favor green, red, or blue, causing them to step at angles $\frac{\pi}{6}$, $\frac{5\pi}{6}$, or $\frac{9\pi}{6}$ radians on the lattice. The probability of taking a step in each direction is given by p (for green), q (for red), and $1-p-q$ (for blue, to sum to one). The edges of the lattice defined by choice criteria consist of absorbing states; upon entering one of these states, the person halts the transition process and selects the corresponding alternative.

This state transition process can be represented as a Markov chain much as the 2-alternative diffusion model is (see Diederich & Busemeyer, 2003), with the caveat that each state has three rather than two transition destinations. It can be implemented as a continuous-time random walk by introducing the standard exponentially distributed transition time (requiring one additional sampling rate parameter), allowing prediction of choice probabilities as well as response times for each of the three possible choice alternatives.

Both the relative and absolute evidence models with any number of alternatives can be described in a similar way. The lattice shape will change and the number of transition destinations at each step will grow along with the number of alternatives, but the sampling rate and thresholds operate in a similar way.[2]

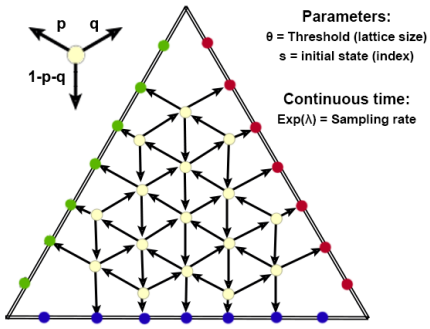


Figure 2: Lattice and state transition structure for a 3-alternative, relative evidence model of color identification.

Note that a person's representation will only be able to step toward available alternatives (not directly away from them) – in Figure 2, this is indicated by unidirectional transitions. Similarly, in accumulator models like the one shown in Figure 1D, the representation will only be able to step to the right or upward. It is perhaps worth noting that the transition probabilities and sampling rate in such a case can be obtained from the accumulation rates of independent Poisson accumulators (as in Smith & Van Zandt, 2000), making it possible to swap between the separate race and single geometric random walk process representations (see, e.g., Ross, 2014).

Such a random walk still requires an initial state. Bias can be introduced by moving the initial state closer to one boundary or another, and starting point variability (required for pre-

[2]Indeed, a fixed sampling rate but a larger number of alternatives will yield slower response times, which may give rise to Hick's Law (Hick, 1952)

dicting fast errors) can be introduced by specifying a mixed state across starting positions.

Psychological spaces

Thus far I have focused mainly on the geometric structure of models where evidence for one alternative has no net effect on the evidence balance between other alternatives. However, this is often an unrealistic assumption. Returning to color categorization, suppose that a participant must match a stimulus to one of 4 categories: red, yellow, green, or blue. One might expect that a stimulus emitting light peaking at a wavelength 610 nm (orange) would provide evidence in favor of both "red" and "yellow" responses, but provide evidence against a "blue" or a "green" response.

In such a case, it makes little sense to treat red, yellow, blue, and green categories as independent or equally related alternatives. Instead, they must be related to one another by constructing a psychological space describing the cognitive representations of the stimulus and choice alternatives.

Doing so requires two generalizations of the framework described in the previous section. First, the directions corresponding to alternatives are permitted to vary. There are several ways to do so. For example, they could be released to vary as free parameters – in 2 dimensions, this could simply be the angle relative to a reference direction, though this would require more parameters when moving to 3 or more dimensions. Alternatively, the directions could be set a priori by the modeler. This could be done by using the physical characteristics of the stimulus, by using existing similarity judgments, or by using existing psychological theory. I give examples of these approaches in the next section.

Second, a person's representation of evidence is modified to reflect the type of information they gather from the stimulus relative to the type of responses they can make. For example, if a person is trying to reproduce the orientation of a stimulus that can vary anywhere from 0 to π radians, they must be able to sample and represent information that favors any direction between 0 and π radians. This will often mean that discrete-state Markov chain representations of the evidence accumulation process are no longer possible, except as approximations or in the rare case that the choice alternatives are arranged in a psychological space so that their orientations allow for a convenient lattice to be superposed upon them.

Random walks in psychological space

Instead of the discrete-state Markov chain, evidence accumulation in a psychological space is enabled by utilizing a multidimensional random walk. In this framework, a person's cognitive state representing evidence they have gathered from a stimulus is described by a point in a multidimensional (e.g. feature-based) space. As a person integrates a new piece information, this state representation is updated by drawing a random variable ϕ that describes a direction in the space and moving one unit of distance in that direction.

The distribution of ϕ is determined by the stimulus and the psychological space in which it is represented. The arrival time of each piece of evidence is again described by an exponential distribution, $Exp(\lambda)$. As before, once the state representation crosses one of the boundaries corresponding to a choice alternative, that alternative is chosen, yielding a choice and response time.

Examples

It is perhaps helpful to visit some examples of how to construct a psychological space and sampling distribution. For each example, we must consider several important factors: what type of response is being elicited (are they discrete / continuous, time-dependent?); how these responses relate to one another in terms of their psychological representation (are they similar, what physical dimensions overlap, how are they processed at a sensory level?); and how a person gathers and represents information about the stimulus relative to the response options (how much evidence does percept X provide for alternative Y?). I do so for two cases: an orientation estimation task where responses can fall anywhere along a continuum and a color identification task where participants must match a color stimulus to a set of categories.

Orientation task

Perceptual tasks often lend themselves to straightforward psychological representations. A common stimulus used to examine how people (as well as other animals) perceive the orientation of objects in the world is the Gabor patch (Figure 3) (see, e.g., Wilson et al., 1983). They may be used in tasks where participants must distinguish one from a set of stimuli, decide between left- or right-leaning orientation, or be combined with color or motion. We examine the simple case where a participant in a task must reproduce the orientation of a Gabor patch stimulus.

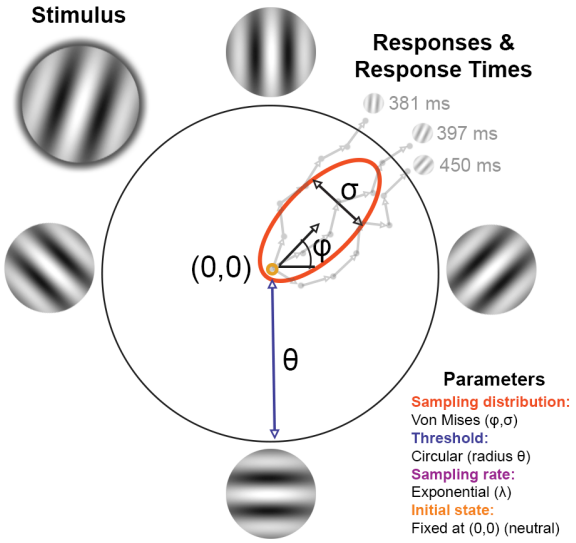


Figure 3: State representation, sampling, threshold, and evidence trajectories for the orientation matching task.

Because these stimuli are symmetric across horizontal and vertical axes, meaning they have no top and bottom, the orientations of these stimuli and the possible responses to them vary from 0 to π radians. For simplicity, assume that stimuli at orthogonal rotations (i.e. $\frac{\pi}{2}$ / $\frac{3\pi}{2}$ or horizontal / vertical) provide evidence against one another. Therefore, they can be arranged in a circle as shown in Figure 3.

In order to complete the task, a participant must gather pieces of information from the stimulus or from memory (if the stimulus has been removed and masked) to construct a cognitive representation that they match to the possible response options. Each piece of information they gather is assumed to be pulled from a von Mises distribution.[3] They sample information until a criterion level of certainty is met, given by the circular threshold, and the point at which the walk crosses the threshold gives the orientation response.

The model still requires additional assumptions regarding the initial representation of the stimulus (before information is gathered) and specification of the sampling rate. For the former, assume a neutral or unbiased initial state, indicating that a participant believes that all orientations are equally likely a priori. Of course, this assumption could be modified if the generating stimuli were not evenly distributed across orientations or if a person was believed to be biased toward (e.g.) a "vertical" response.

The mean of the von Mises sampling distribution (ϕ) can be set directly from the stimulus, given that the true orientation is the most likely to be sampled. Then the error, threshold, and sampling rate are free to vary and can be estimated. In the simple case of diffusion in 2 dimensions with a von Mises sampling distribution and circular decision bounds, first passage times can be derived analytically (see Smith, 2016, for a thorough discussion of this matter). However, in more complex cases with different assumptions regarding non-decision time, functional forms of starting point variability, decision bounds, or trial-to-trial parameter variability, the model may need to be simulated in order to compute the likelihood and/or require approximate fitting criteria such as joint quantiles (as in Heathcote et al., 2002).

Color identification task

I return now to the color identification task, where a participant must classify a fixed-luminosity stimulus as red, green, yellow, or blue. Each category corresponds to a direction D_R, D_G, D_Y, or D_B.

The particular direction corresponding to each color may be fixed or free to vary. One of the benefits of taking the approach presented here is that each direction can in fact be fit by imbuing it with a free parameter or fixed based on a particular theory the modeler has in mind. This allows for model comparison between 2-dimensional representations based purely on hue (Figure 4A), Hering's opponent-process representation (Figure 4B), a multi-dimensional scaled space representation (as in Shepard, 1962)(Figure 4C), or a situation where each of the directions is given by an additional parameter of the model.

Each of these theories makes very different predictions about how stimuli and responses should interact. For example, a stimulus showing light at 430 nm (violet) should result in very long response times in hue or multidimensional scaling space, but not in an opponent-process theory space. Similarly, the multidimensional scaling space predicts far more green-yellow and green-blue confusions than do hue or opponent process constructions. Because of this, the pattern of responses from a decision task can actually inform and distinguish between the theories used to construct the space, or even develop better theories when parameters giving D_R, D_G, D_Y, and D_B are estimated.

As with the orientation task, a neutral initial state, von Mises sampling distribution (or the von Mises-Fisher in a higher-dimensional representation), and exponential arrival times for new samples are reasonable assumptions for how a person's cognitive representation of the stimulus changes over time. These can of course be modified if the base rates of different colors or stimulus information change.

Discussion

One may wonder what the point of modeling different tasks in this way is. The answer is many-fold. One potential use is as a measurement model – the parameters of the model are psychologically meaningful, with sampling rate describing how quickly a person can gather new information, threshold describing the strictness with which they make decisions, and the amount of error in drift or starting point location in the system (e.g., in a von Mises distribution or mixed initial state) allows us to potentially diagnose sources of bias and inaccuracy as with the uni-dimensional diffusion model (Ratcliff & McKoon, 2008; Pleskac & Busemeyer, 2010).

Second, modeling decisions among many alternatives in this way allows for testable predictions and model comparison. For example, without starting point variability, the models presented in the examples predict that response times should be longer for larger errors. This means that one would expect an incidental "red" response to a blue stimulus or a left-leaning response to a right-leaning stimulus to take longer than more similar responses like "green" or vertical-leaning (respectively).

Furthermore, psychological representations of both stimuli and potential responses can be informed by established models of how people represent objects and concepts. Similarity judgments or categorization data can be used to construct a multidimensional feature space that relates available choice alternatives, in turn making a priori predictions about response and response time distributions about decisions among them. Similarly, neural data about activity during decision-making can be connected to the psychological

[3] This is similar to a normal distribution on a circle. Each piece of evidence could be sampled from momentary activation across orientation columns in the visual cortex. This opens up the question of whether activation across the columns mimics a von Mises distribution, which is interesting but beyond the scope of the paper.

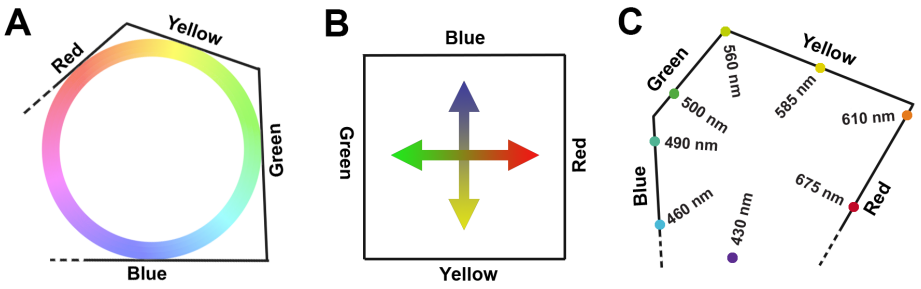


Figure 4: Accumulation space representation for 4-category decision based on hue (A), visual opponent-process theory (B), and Shepard (1962)'s multidimensional scaling construction of color (C).

representations of the stimuli and response alternatives.

When put together, this approach may also provide natural and parsimonious explanations of existing psychological phenomena, such as context effects (Trueblood et al., 2014), Hick's law (Hick, 1952), selection of prices and certainty equivalents (Busemeyer & Diederich, 2002), confidence and confidence response times (Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2009), and bow effects and lateral interactions in absolute identification (Brown et al., 2008).

In short, I hope that the approaches and models presented here are able to inform our understanding of the decision process. At best, they should open up new questions as well as bring together different perspectives and sources of data on how to represent and process information. At worst, they should at least provide a first approach toward modeling how responses and response times arise when decisions are made among many alternatives or along a continuum.

Acknowledgments

The author would like to thank Timothy Pleskac and Andrew Heathcote for helpful discussions that led to the ideas presented in the paper. The author was additionally supported a graduate research fellowship from the National Science Foundation (Grant No. DGE-1424871).

References

Brown, S. D., Marley, A., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*(2), 396–425.

Busemeyer, J. R., & Diederich, A. (2002). Survey of decision field theory. *Mathematical Social Sciences*, *43*(3), 345–370.

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. *Journal of Mathematical Psychology*, *47*(3), 304–322.

Heathcote, A., Brown, S., & Mewhort, D. J. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic bulletin & review*, *9*(2), 394–401.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26.

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* Academic Press.

Nosofsky, R. M. (1997). An exemplar-based random-walk model of speeded categorization and absolute judgment. *Choice, decision, and measurement: Essays in honor of R. Duncan Luce*, 347–365.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59-83.

Ross, S. M. (2014). *Introduction to probability models*. Academic press.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika*, *27*(3), 219–246.

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, *53*(2), 293–315.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological review*, *121*(2), 179–205.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550–592.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological review*, *111*(3), 757–769.

Wilson, H. R., McFarlane, D. K., & Phillips, G. C. (1983). Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision research*, *23*(9), 873–882.

Centering and the Meaning of Conditionals

Nicole Cruz[1, 2] (ncruzd01@mail.bbk.ac.uk), David Over[3] (david.over@durham.ac.uk), Mike Oaksford[2] (m.oaksford@bbk.ac.uk), Jean Baratgin[1] (jean.baratgin@univ-paris8.fr)

[1] Laboratoire CHArt (Université Paris 8 & EPHE), 4-14, rue Ferrus, 75014 Paris, France
[2]Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK
[3]Department of Psychology, Durham University, South Rd, Durham DH1 3LE, UK

Abstract

The centering inference - p & q, therefore if p then q - is important in reasoning research because it is logically valid for some accounts of conditionals (e. g. the material and the probability conditionals), but not for others (e. g. the inferential conditional, according to which a conditional is true if and only if there is an inferential connection between p and q). We tested participants' acceptance of centering compared to valid and invalid inferences not containing conditionals, varying the presence of an inferential connection and of a common topic of discourse between p and q. Participants' acceptance of centering was more similar to valid inferences than to invalid inferences, and there was no reliable effect of a connection between p and q. Acceptance rates were higher when there was a common topic of discourse, independently of the type of inference. The findings support the probability conditional account.

Keywords: validity; uncertainty; conditionals; centering

Introduction

Conditional sentences of the form *if p then q*, like "If it is snowing, then the streets are slippery", are used on a daily basis in ordinary discourse and in science. And yet there is still not a consensus on their meaning. They are central to any account of reasoning, because every inference from *p* to *q* can be rephrased as a conditional. Our understanding of the meaning of conditionals therefore goes hand in hand with our understanding of reasoning processes as a whole (Over & Cruz, in press).

Most reasoning takes place from premises that are not certain, but are held only with higher or lower degrees of belief. This uncertainty affects the inferences people draw, and therefore has to be taken into account in theories of reasoning (Evans & Over, 2004, 2013; Oaksford & Chater, 2007, 2013; Pfeifer & Kleiter, 2009, 2011). Classical logic, with its restriction to the binary values of truth and falsity, cannot represent uncertainty, but this becomes possible with the use of probability theory (Adams, 1998; Coletti & Scozzafava, 2002; de Finetti, 1936/1995; Gilio, 2002; Ramsey, 1926/1994; Jeffrey, 1991).

One inference that has not been studied much in the psychology of reasoning is centering (or conjunctive sufficiency): inferring *if p then q* from *p & q*. Centering is important because it is logically valid for several theories of the conditional (Evans & Over, 2004; Girotto & Johnson-Laird, 2004; Lewis, 1973; Pfeifer & Kleiter, 2010; Stalnaker, 1968), and logically invalid for several others (Douven, 2015a; Cariani & Rips, in press; Kratzer, 2012) allowing a comparison between these theories. In particular, centering is valid for the *probability conditional* (Over, in press; Pfeifer & Kleiter, 2010), but invalid for the *inferential conditional* (Douven, 2015a; Skovgaard-Olsen, Singmann, & Klauer, in press). We will focus on these two theories in the following.

The probability conditional

If people interpret a natural language conditional as a probability conditional, then their degree of belief in the conditional is the conditional probability of its consequent given its antecedent: *P(if p then q) = P(q|p).* This equality is called *The Equation* as a normative philosophical proposal (Edgington, 1995). The psychological hypothesis that people's interpretation of conditionals conforms to the Equation is called the *conditional probability hypothesis* (Evans & Over, 2004). It is proposed that people compute this conditional probability by performing a *Ramsey test*: a mental simulation in which they hypothetically assume *p* to be the case, make any changes to their beliefs that may be necessary to preserve consistency, and assess the probability of *q* on this basis (Evans & Over, 2004; Ramsey, 1929/1994; Stalnaker, 1968). An inferential connection between *p* and *q* may influence people's assessment of the conditional probability, but the effect of such a relation is procedural and pragmatic and not part of the semantics of conditionals.

The inferential conditional

Inferentialism states that a high conditional probability, *P(q|p)* is high, is not sufficient for the conditional to have a high probability, *P(if p then q)* is high. For the conditional to have a high probability, there also has to be an inferential connection between *p* and *q*. Of course, *P(q|p)* can be high because *P(q)* is high and *q* is independent of *p*, but this is not sufficient to make *P(if p then q)* high for inferentialism. In the strongest version of the theory, a conditional is proposed to be true *if and only if* there is a strong enough argument from *p* (plus relevant background knowledge) to *q*, where this argument can be deductive, inductive, or abductive (Douven, 2015a; Krzyżanowska, Wenmackers, & Douven, 2014; c. f. Douven, 2015b). A deductive argument is assumed to be present when *q* is necessarily true given the truth of *p* (because for instance *p* is a subset of *q*). An abductive argument is assumed to be present when *p* is the best explanation for *q* (Douven & Verbrugge, 2010). It is

held that "*q* is an inductive consequence of *p* (given the background premises) iff *q* follows with high statistical probability from *p* (in light of the background premises)" (Krzyżanowska et al., p. 775; c. f. Douven & Verbrugge, 2010, p. 303). If taken literally, this criterion implies that a high conditional probability is sufficient for a conditional to be highly probable. We therefore take the authors to mean that an inductive argument is present not only when the conditional probability of *q* given *p* is statistically high, but when in addition it is higher than the unconditional probability of *q*, that is, when *p* raises the probability of *q* (c. f. Douven, 2008, Oaksford & Chater, 2007, and the use of the delta-p rule in research on causation, as in Cheng, 1997).

This definition of an inferential conditional in terms of the binary values of truth and falsity has not been extended to degrees of belief. But if we make the uncontroversial assumption that degrees of belief are degrees of belief that a statement is true, or true given that it has a truth value, and no further qualification is added to the inferentialist definition, then this definition seems to imply that the degree of belief in a conditional is a direct function of the strength of an inferential connection between *p* and *q*.

A less strong variant of inferentialism integrates the requirement for a connection with the conditional probability hypothesis. It proposes that people assume by default that a conditional has a connection, operationalized as a positive covariation between *p* and *q* when using naturalistic materials. If this assumption is met, then people set their degree of belief in the conditional to that of the conditional probability. If the assumption that there is a connection is not met, then people "penalize" the value of the conditional probability by adjusting it downwards (Skovgaard-Olsen et al., in press; c. f. Douven, 2008). In virtue of its contingent nature, this "penalty" is arguably a pragmatic adjustment rather than a semantic requirement, but its supporters hold that it is semantic.

Previous findings

The Equation has received strong and consistent empirical support (Evans, Handley, & Over, 2003; Fugard, Pfeifer, Mayerhofer, & Kleiter, 2011; Oberauer & Wilhelm, 2003; Politzer, Over, & Baratgin, 2010), but there has been little empirical research on inferentialism. Some studies have found an effect of a covariation between *p* and *q* on people's belief in conditionals (Oaksford & Chater, 2010; Oaksford, Chater, & Larkin, 2000; Skovgaard-Olsen et al., in press), while others have not (Oberauer, Weidenfeld, & Fischer, 2007; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Singmann, Klauer, & Over, 2014). Oberauer et al. (2007) found that the presence of a causal connection affected people's belief in a conditional in the absence of explicit frequency information on the conditional probability. However, when frequency information was given, the effect of a causal connection was negligible. This suggests that information on a connection is relevant for estimating the conditional probability, but unnecessary when this probability is already known. Moreover, inferentialism, unlike the probability conditional hypothesis, is not in accordance with truth table tasks (Baratgin, Over, & Politzer, 2013; Evans et al., 2003; Oberauer & Wilhelm, 2003), where people hold the conditional to be true when *p* and *q* are true, and where they judge the probability of the conditional to be a function of $P(q|p)$. In these tasks, the materials, which would indicate the presence or absence of a connection, remain constant (Over & Cruz, in press).

Proponents of inferentialism have argued that such studies do not undermine their proposal because of the use in them of abstract or pseudo-naturalistic conditionals, and that the scope of their theory is restricted to naturalistic conditionals (Skovgaard-Olsen et al., in press). They have also pointed out that a distinction is made in linguistics between *content* and *inferential* conditionals, and their theory is restricted to the latter. It also does not apply to so-called *Peter-Pan conditionals*, e.g. "If fairies exist, then I am Peter Pan", or *non-interference conditionals*, which are types of conditionals that explicitly lack an inferential connection (Douven & Verbrugge, 2010; Krzyżanowska et al., 2014). One problem for inferentialism is that such restrictions reduce its falsifiability, and expose it to the danger of providing ad hoc explanations for data.

The present experiment

As described above, both variants of inferentialism assume that the role of a connection is semantic and not pragmatic. They conceive of the connection as a necessary component of what makes a linguistic form a conditional and not some other kind of statement. To test this assumption, we draw on a conception of validity from test theory. Note that this notion of (test) "validity" is different from the logical validity of an argument or inference. In test theory, a test is said to be "valid" when it is *sensitive* enough to measure the construct it is designed to measure, and when it is *specific* enough not to measure much else in addition to the construct. In analogy to this, we consider there to be evidence that the effect of a connection is semantic when it is present across conditionals of different types and contents, and when it is specific to conditionals, or at least stronger for conditionals than for other connectives.

Regarding sensitivity, we already mentioned above the restriction in scope of inferentialism, which makes it seem unwarranted to assert it as a theory of the meaning of conditionals in general, without further qualification. Here we focus on specificity: That the effect of a connection is specific or at least higher for conditionals than for other connectives. We therefore compare people's degree of belief in centering with people's degree of belief in inferences not containing conditionals, the logical validity (invalidity) of which is much less controversial.

If an effect of a connection were to be present to a similar degree for inferences with and without conditionals, then inferentialism would have no means to distinguish conditionals from other connectives, unless it reduced, in its

less strong version, to the conditional probability hypothesis.

To illustrate the problem with unconnected conditionals, proponents of inferentialism often use examples like "If Hillary Clinton runs for president in 2016, the earth weighs more than 2 kilograms." (Krzyżanowska et al., 2014). Referring to this example, the authors state "antecedent and consequent have nothing to do with each other, which at least on our hypothesis accounts for the felt falsity of this sentence." The sentence indeed sounds odd, but it not only expresses no inferential connection. As the authors say, antecedent and consequent have nothing to do with each other: they belong to different topics. This raises the question of whether the intuitive oddness of such sentences comes from the absence of an inferential connection, or alternatively from the absence of a common topic of discourse. To assess this question, we compared people's degree of belief in inferences under three conditions: (1) one in which there is a connection between *p* and *q*, (2) one in which there is no connection, but a common topic of discourse, and (3) one in which there is no connection and no common topic of discourse.

Method

Participants

A total of 670 participants from English speaking countries completed the online experiment. Following methodological recommendations for online research (Aust, Diedenhofen, Ullrich, & Musch, 2013; Reips, 2002), we excluded cases with repeated IP address; cases who failed a test question assessing whether they were reading the materials presented to them or just clicking through; cases who indicated at the end of the experiment that they did not take part seriously; cases with one or more trial reaction times of less than 3 seconds; and cases who rated their English language skills as not being high or that of a native speaker. The final sample consisted of 363 participants, reporting a mean age of 42.56 (range: 15-80) and a diverse formal educational background. They received a compensation of $ 0.16 for their participation.

Materials and design

Participants were asked to imagine that they were researchers investigating the birds of the invented island of Liaku, and also that they were following the election results in their far away home country Raimos. On each trial, participants were shown a short context story, followed by a one-premise inference. They were asked to assume that they had a specific degree of belief in the premise of the inference, and were asked to judge how confident they could reasonably be in the conclusion, given this degree of belief in the premise. Participants gave their answer by writing a percentage into a box.

Each participant was shown materials from one of three conditions. In group 1 (the connection condition) *p* and *q* referred to two features of the same bird (e. g. "The next Amri bird you see on Liaku will eat arb seeds" and "The next Amri bird you see on Liaku will build its nests on arb trees"). In group 2 (the no connection – same topic condition) *p* and *q* referred to different features of two different birds (e. g. "The next Amri bird you see on Liaku will eat arb seeds" and "The next Grauk bird you see on Liaku will live in groups"). In group 3 (the no connection – different topic condition) *p* and *q* referred to different topics (e. g. "The next Amri bird you see on Liaku will eat arb seeds" and "The Grauk province of Raimos will increase taxes"). The materials were pretested for the extent to which people perceived there to be a connection between *p* and *q* (median ratings in percent: 78, 19, and 1 for groups 1 to 3, respectively). Participants were randomly allocated to one of the three groups.

Each participant gave ratings to the six inferences displayed in Table 1.

Table 1. The inferences investigated.

	Name	Form
1	and-to-if	*p & q, ∴ if p then q*
2	if-to-and	*if p then q, ∴ p & q*
3	and-to-or	*p & q, ∴ p or q*
4	or-to-and	*p or q, ∴ p & q*
5	and-elim	*p & q, ∴ p*
6	and-intro	*p, ∴ p & q*

Note. "∴" stands for "therefore".

Inference 1 is centering, whose validity differs between theories of the conditional. Inference 2 is its converse, and is more uncontroversially invalid. The remaining inferences do not contain conditionals. Inferences 3 and 5 are more uncontroversially valid, and inferences 4 and 6 are more uncontroversially invalid. We will refer to inferences 2 to 6 simply as valid resp. invalid in what follows.

Each inference was presented three times, showing a premise probability of 1, .8 and .2, respectively. The materials in which the inferences were embedded were different on each trial. The allocation of materials to inferences, as well as the trial order, was varied randomly for each participant.

The experiment was conducted in a single online session using the platform CrowdFlower(c), and took 8.4 min on average to complete.

Results and discussion

As expected, participants' ratings of conclusion probability were lower for lower premise probabilities. However, the pattern of responses for the variables of interest to the hypotheses was very similar across premise probabilities. In particular, there was no qualitative difference in responses to inferences with certain and with uncertain premises. Because we had no specific prediction about effects of premise probability, which we varied merely to increase the generalizability of the findings, subsequent results were aggregated across this variable.

To test whether centering is generally treated as valid, and whether its treatment as valid depends on the presence of a connection and/or of a common topic of discourse between *p* and *q*, we first compared inference 1 (centering) with inference 2 (its converse). We next compared inference 1 with the average ratings for valid inferences 3 and 5. Finally, we compared inference 1 with the average ratings for invalid inferences 4 and 6. The results are displayed in Figure 1.

Centering vs. its converse A mixed ANOVA with group (1, 2, 3) as between subjects factor and inference (1, 2) as within subjects factor led to a significant effect of inference, $F(1, 360) = 18.46$, $p < .001$, $\eta_p^2 = .049$: ratings for inference 1 were higher ($M = 62.90$, $SE = .71$) than ratings for inference 2 ($M = 59.03$, $SE = .86$). There was also a significant effect of group, $F(1, 360) = 3.55$, $p = .03$, $\eta_p^2 = .019$; but no interaction between inference and group ($F < 1$). Follow-up analyses for the effect of group revealed that ratings for group 3 ($M = 58.66$, $SE = 1.14$) were lower than for group 1 ($M = 62.49$, $SE = 1.18$), $F(1, 257) = 5.46$, $p = .02$, $\eta_p^2 = .021$. Ratings for group 3 were marginally lower than for group 2 (M = 61.75, SE = 1.17), $F(1, 236) = 3.95$, $p = .048$, $\eta_p^2 = .016$; but there was no significant difference between ratings for group 1 and for group 2 ($F < 1$).

The finding of higher ratings of conclusion probability for centering than for its converse, regardless of the presence or absence of a connection, is inconsistent with inferentialism. Inferentialism would have predicted an interaction between inference and group, such that ratings are lower for centering than for the other inferences in the groups with no connection. The finding that the effect of group could be traced back to the effect of a common topic of discourse, but not to the presence of a connection, is also not in accordance with inferentialism.

Centering vs. valid inferences 3 and 5 A mixed ANOVA with group (1, 2, 3) as between subjects factor and inference (1, 3&5) as within subjects factor led to a small main effect of inference, $F(1, 360) = 4.73$, $p = .03$, $\eta_p^2 = .013$: ratings for inference 1 ($M = 62.90$, $SE = .71$) were lower than mean ratings for inferences 3 and 5 ($M = 64.57$, $SE = .55$). There was also a main effect of group, $F(1, 360) = 3.82$, $p = .023$, $\eta_p^2 = .021$. Inference and group did not interact ($F < 1$). Follow-up analyses to the effect of group showed that ratings were lower for group 3 ($M = 62.08$, $SE = .88$) than for group 1 ($M = 65.37$, $SE = .91$), $F(1, 257) = 6.79$, $p = .01$, $\eta_p^2 = .026$. There was no difference in ratings between groups 1 and 2, $F(1, 227) = 1.73$, $p = .19$, $\eta_p^2 = .008$; nor between groups 2 and 3, $F(1, 236) = 1.92$, $p = .17$, $\eta_p^2 = .008$.

The absence of an interaction between inference and group is again contrary to the predictions of inferentialism, as is the finding that the effect of group could be traced back to the effect of a common topic of discourse, but not to the effect of a connection. The small effect of inference indicates that people rated the conclusion of centering slightly lower than the conclusion of the two valid inferences. However, the effect was not reliable: it did not reach significance when analyzed individually in any of the groups (for group 1, $t < 1$; for group 2, $t = 1.39$, $p = .17$, 95%CI [-3.7, .65]; for group 3, $t = 1.86$, $p = .07$, 95%CI [-4.9, .15]). There is therefore little evidence for a treatment of centering as invalid.

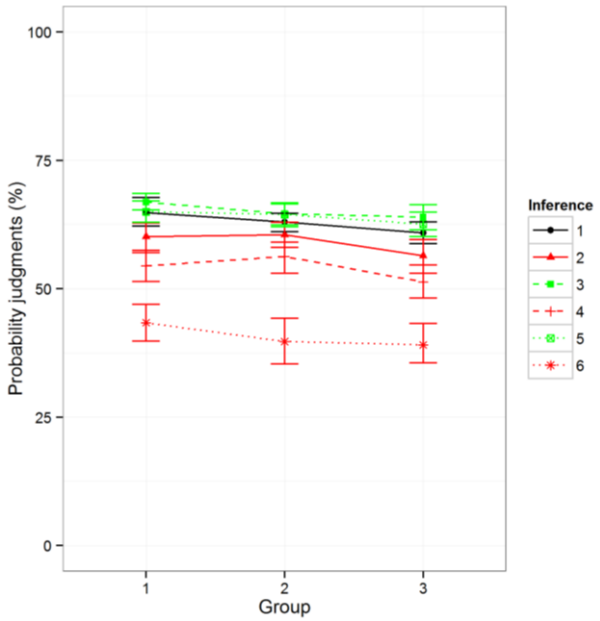


Figure 1. Judgments of conclusion probability for the six inferences investigated, separately for group 1 (connection), group 2 (no connection – same topic) and group 3 (no connection – different topic). The black line represents the target inference of centering. Green lines represent the valid and red lines the invalid comparison inferences. Error bars show 95% CI.

Centering vs. invalid inferences 4 and 6 A mixed ANOVA with group (1, 2, 3) as between subjects factor and inference (1, 4&6) as within subjects factor led to a main effect of inference, $F(1, 360) = 232.88$, $p < .001$, $\eta_p^2 = .393$: mean ratings for inferences 4 and 6 ($M = 47.38$, $SE = .88$) were lower than for inference 1 ($M = 62.90$, $SE = .71$). There was also a main effect of group, $F(1, 360) = 3.61$, $p = .028$, $\eta_p^2 = .020$; but no interaction between inference and group ($F < 1$). Follow-up analyses to the effect of group revealed that ratings for group 3 ($M = 53.04$, $SE = 1.01$) were lower than for group 1 ($M = 56.88$, $SE = 1.04$), $F(1, 257) = 6.39$, $p = .012$, $\eta_p^2 = .024$. There was no significant difference between group 1 and group 2 (for group 2: $M = 55.50$, $SE = 1.14$) ($F < 1$), nor between groups 2 and 3, $F(1, 236) = 2.61$, $p = .108$, $\eta_p^2 = .011$.

The difference between centering and the mean of the invalid inferences 4 and 6 was reliably present (for group 1, $t = 8.23$, $p < .001$, 95%CI [12.1,19.7]; for group 2, $t = 9.08$, $p < .001$, 95%CI [11.7, 18.2]; for group 3, $t = 9.69$, $p < .001$, 95%CI [12.5,18.9]). Ratings for centering were more similar to those for valid inferences (difference: $M = -1.67$,

SE = 0.8) than to those for invalid inferences (difference: *M* = 15.52, *SE* = 1.0), $F(1, 360) = 290.94$, $p < .001$, $\eta_p^2 = .447$, irrespective of group ($F < 1$).

The finding that centering was treated as a logically valid rather than as an invalid inference is at odds with theories of the conditional that reject centering. The consistent absence of an interaction between inference and group, as well as the finding that the effect of group could be traced back to that of a common topic of discourse, but not to that of a connection, is at odds with the predictions of inferentialism more specifically.

We repeated the above analyses using a linear mixed model with a random intercept for participants, finding the same pattern of significant and non-significant results. The model failed to converge when attempting to introduce a random slope for participants, or a random intercept and/or slope for item contents. We report the ANOVA here for reasons of space.

General discussion

We investigated whether people treat centering as a logically valid inference, comparing it to inferences not containing conditionals, and whether people's acceptance of centering depends on the presence of a connection or common topic of discourse between *p* and *q*. We found centering to be treated more like a logically valid inference than like an invalid one, irrespective of the presence of a connection or of a common topic of discourse. Centering was accepted less when there was no common topic of discourse, but was not accepted less when there was no connection. The effect of a common topic of discourse was not specific to centering, but occurred to a similar degree for inferences not containing conditionals.

These findings are not in line with theories implying that centering is logically invalid for the natural language conditional, and not with inferentialism more specifically. They suggest that what matters for the intuition that there is something odd in conditionals like "If Hillary Clinton runs for president in 2016, the earth weighs more than 2 kilograms", is the absence of a common topic of discourse between *p* and *q*, rather than the absence of a specific connection. Further, this intuition does not seem to be specific to conditionals, but appears to apply similarly to inferences with disjunctions and conjunctions, suggesting that the effect is pragmatic and not semantic. If it were semantic, then the stronger version of inferentialism would be unable to distinguish conditionals from other connectives, and the less strong version would reduce to the conditional probability hypothesis.

The suggestion that conditionals with no connection are odd for pragmatic reasons is in accordance with previous findings in support of the conditional probability hypothesis (Evans et al., 2003; Oberauer & Wilhelm, 2003); with findings in support of centering, albeit without varying the presence of a connection (Cruz, Baratgin, Oaksford, & Over, 2015; Politzer & Baratgin, 2015); and with evidence for the role of a connection when using conditionals in the pragmatic context of argumentation (Hahn & Oaksford, 2007).

The present findings are preliminary, however, and require replication and extension before allowing stronger conclusions. One point to consider is that the presence of a connection was varied between subjects in this experiment. This was important because it allowed us to avoid participants reacting to the variation of a connection simply because of its salience in the materials, making them think the experimenter expects them to respond differentially towards it. However, a variation within subjects would make it easier to find an effect of group, and thus also easier to find the interaction predicted by inferentialism. Further, the materials used here were pseudonaturalistic. It would be informative to test whether the findings can be generalized to naturalistic scenarios.

If subsequent research provides further evidence against inferentialism, this would undermine the generality of the inferentialist proposal. However, it would not invalidate it as such. To take the causal domain as an example, there are certainly statements of the form *p causes q*, which are true when *p* causes *q*, and false otherwise. Such statements do not necessarily change their truth conditions when they are rephrased as *if p then q*. Moreover, there is likely to be a continuum in the extent to which a conditional is meant to assert the presence of a causal or other kind of relation.

However, to the extent that a conditional is meant to assert *p causes q*, it can perhaps be considered general as opposed to singular (Cruz & Oberauer, 2014). General conditionals are useful because they allow us to build theories about the world, to help us understand it and make predictions (Chater & Oaksford, 2013). Sometimes we are looking for information that is relevant for refining our theories about the world. But other times we just want to make a prediction of *what* is the case in a particular instance, regardless of *why* it is the case. For example, we might sometimes want to know whether mosquito bites really cause a certain illness. At other times, we might be satisfied with learning simply that the conditional probability of the illness is high given mosquito bites, to take quick and easy measures against the bites. Sometimes a connection may be used as a predictor for the conditional probability, and sometimes the converse may hold. Whether a connection or the conditional probability is more basic may depend on the goals of the reasoner. In this way, a narrower scope for inferentialism in the context of singular indicative conditionals could imply a wider scope for it in the context of generals and counterfactuals.

Acknowledgments

We would like to thank Igor Douven, Karolina Krzyżanowska, and Ulrike Hahn for very helpful discussions on the topic. We are very grateful to the Mexican National Council of Science and Technology (CONACYT), and to the ANR grant Chorus 2011 (project BTAFDOC) for supporting this research.

References

Adams, E. (1998). *A primer of probability logic.* Standford, US: CLSI publications.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527-535.

Baratgin, J., Over, D. E., & Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking & Reasoning, 19*, 308-328.

Cariani, F., & Rips, L. J. (in press). Conditionals, context, and the suppression effect. *Cognitive Science*.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science, 37,* 1171-1191.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367 – 405.

Coletti, G., & Scozzafava, R. (2002). *Probabilistic logic in a coherent setting*. Kluwer, Dordrecht.

Cruz, N., Baratgin, J., Oaksford, M., & Over, D. E. (2015). Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology, 6*, 192.

Cruz, N., & Oberauer, K. (2014). Comparing the meanings of "if " and "all." *Memory &. Cognition, 42*, 1345–1356.

de Finetti, B. (1936/1995). The logic of probability. Translated in R. B. Angell, The logic of probability. *Philosophical Studies, 77,* 181-190.

Douven, I. (2008). The evidential support theory of conditionals. *Synthese, 164*, 19-44.

Douven, I. (2015a). How to account for the oddness of missing-link conditionals. *Synthese.*
DOI 10.1007/s11229-015-0756-7

Douven, I. (2015b). *The Epistemology of Indicative Conditionals: Formal and Empirical Approaches.* Cambridge, UK: Cambridge University Press.

Douven, I. & Verbrugge, S. (2010). The Adams family. *Cognition, 117*, 302-318.

Edgington, D. (1995). On conditionals. *Mind*, 104, 235-329.

Evans, J. St. B. T., Handley, S. J., & Over, D. E. (2003). Conditional and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 321–335.

Evans, J. St. B. T., & Over, D. E. (2004). *If.* Oxford, UK: Oxford University Press.

Evans, J. St. B. T., & Over, D. E. (2013). Reasoning to and from belief: deduction and induction are still distinct. *Thinking & Reasoning, 19*, 267-283.

Fugard, J. B., Pfeifer, N., Mayerhofer, B., & Kleiter, G. D. (2011). How people interpret conditionals: Shifts toward conditional event. *Journal of Experimental Psychology: Learning Memory and Cognition, 37,* 635-648.

Gilio, A. (2002). Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence 34*, 5-34.

Girotto, V., & Johnson-Laird, P. N. (2004). The probability of conditionals. *Psychologia, 47*, 207-225.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*, 646-678.

Jeffrey, R. C. (1991). Matter of fact conditionals. *Aristotelian Society Supplementary Volume, 65*, 161-183.

Kratzer, A. (2012). *Modals and conditionals*. Oxford, UK: Oxford University Press.

Krzyżanowska, K. H., Wenmackers, S., & Douven, I. (2014). Inferential conditionals and evidentiality. *Journal of Logic, Language and Information, 22*, 315-334.

Lewis, D. (1973). *Counterfactuals.* Cambridge, USA: Harvard University Press.

Oaksford, M, & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.

Oaksford, M., & Chater, N. (2010). Causation and conditionals in the cognitive science of human reasoning. *The Open Psychology Journal, 3*, 105-118.

Oaksford M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning, 19*, 346-379.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 883-899.

Oberauer, K., Weidenfeld, A., & Fischer, K. (2007). What makes us believe a conditional? The roles of covariation and causality. *Thinking & Reasoning, 13*(4), 340 – 369.

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models and personal utilities. *Journal of Experimental Psychology: Learning Memory and Cognition, 29*, 680-693.

Over, D. E. (in press). The paradigm shift: The debate. In L. Macchi, M. Bagassi, & R. Viale (Eds.), *Cognitive unconscious and human rationality*. Cambridge, US: MIT Press.

Over, D. E., & Cruz, N. (in press). Probabilistic accounts of conditional reasoning. In Linden J. Ball, & Valerie A. Thompson (Eds.), *International handbook of thinking and reasoning*. Hove, Sussex: Psychology Press.

Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology, 54*, 62-97.

Pfeifer, N., and Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic, 7*, 206–217.

Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. In M. Oaksford & N. Chater (Eds.), *Cognition and conditionals: Probability and logic in human thinking* (pp. 153–173). New York, US: Oxford University Press.

Pfeifer, N., and Kleiter, G. D. (2011). Uncertain deductive reasoning. In K. I. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reason: A Festchrift for Jonathan St. B. T. Evans* (pp. 145-166). Hove: Psychology Press.

Politzer, G., & Baratgin, J. (2015). Deductive schemas with uncertain premises using qualitative probability expressions. *Thinking & Reasoning, 22*, 78-98.

Politzer, G., Over, D. E., & Baratgin, J. (2010). Betting on conditionals. *Thinking & Reasoning, 16*, 172–197.

Ramsey, F. P. (1926/1990). Truth and probability. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 52-94). Cambridge, UK: Cambridge University Press.

Ramsey, F. P. (1929/1990). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers* (pp. 145-163). Cambridge, UK: Cambridge University Press.

Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology, 49*(4), 243-256.

Singmann, H. & Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology, 5*, 316.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition, 150*, 26-36.

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98-112). Oxford, UK: Blackwell.

A rational speech-act model of projective content

Ciyang Qing, **Noah D. Goodman**, **Daniel Lassiter**
{qciyang, ngoodman, danlassiter}@stanford.edu
Stanford University

Abstract

Certain content of a linguistic construction can project when the construction is embedded in entailment-canceling environments. For example, the conclusion that John smoked in the past from the utterance *John stopped smoking* still holds for *John didn't stop smoking*, in which the original utterance is embedded under negation. There are two main approaches to account for projection phenomena. The semantic approach adds restrictions of the common ground to the conventional meaning. The pragmatic approach tries to derive projection from general conversational principles. In this paper we build a probabilistic model of language understanding in which the listener jointly infers the world state and what common ground the speaker has assumed. We take change-of-state verbs as an example and model its projective content under negation. Under certain assumptions, the model predicts the projective behavior and its interaction with the *question under discussion* (QUD), without any special semantic treatment of projective content.

Keywords: Presupposition; projection; Bayesian pragmatics

Introduction

> "How am I to get in?" asked Alice again, in a louder tone. "*Are* you to get in at all?" said the Footman. "That's the first question, you know." — Carroll (1866), ch. VI

Courtroom drama, political misinformation, and ordinary misunderstandings often revolve around *presupposition*, a backgrounded aspect of meaning with a complex logic and communicative function. Famously, presuppositions can be used manipulatively, as in the classic loaded question "Have you stopped beating your wife?", or the sly reporter's "When did you become aware that this policy was a failure?". At the same time, presuppositions serve to streamline conversation by allowing interlocutors to convey multiple pieces of information simultaneously. Alice's question "How am I to get in?" efficiently indicates both Alice's assumption that she will get in, and her wish to know how to enter.

In addition to their important communicative role, presuppositions are interesting because they seem to flout some of the most basic rules of logic. For example, from *John danced* we can infer *John moved*, but we cannot infer this from *John didn't dance*—the inference that *John moved* is canceled by negation. In contrast, from both *John stopped smoking* and *John didn't stop smoking* we are likely to infer that *John used to smoke*—this inference is said to *project* over negation. In natural language semantics and pragmatics, an inference that survives an operator that usually cancels inferences is called *projective content* of the sentence under that operator. Change-of-state verbs can have information about the past as projective content under negation: *John used to smoke* is projective content under negation of "John stopped smoking." There are many other types of projective contents (e.g., the complement of *know*) under different operators (e.g., questions, modals) identified and discussed in the literature. The problem of explaining how certain inferences can survive entailment-canceling operators is called the *projection problem*.

There are two main approaches to the projection problem. According to the *semantic* approach, projective contents are conventional properties of lexical items (e.g., Frege, 1948; Heim & Kratzer, 1998). According to the *pragmatic* approach, projection can be derived from general conversational principles (e.g., Stalnaker, 1974; Simons, 2001, 2006). How could we capture projection patterns using general conversational principles? To illustrate the reasoning, let us expand the example scenario: Alice and Bob are talking about John, an old friend of Alice's who is visiting her. Bob has never met John before so he knows nothing about him. Bob asks Alice, "Does John smoke?" Alice replies, "John did not stop smoking."[1] Taken literally, Alice's utterance seems under-informative: it can be literally true, regardless of whether or not John smokes. If Alice knows whether John smokes and is cooperative, she would not have said something under-informative. So perhaps her answer is informative after all, but how? Maybe she has taken some additional information for granted, assuming that it is in the *common ground* with Bob. Indeed, if Alice took for granted that John smoked in the past, then, together with "John did not stop smoking," this information would mean that John still smokes, which fully answers the question of whether John smokes now. In other words, assuming that Alice took for granted that John smoked in the past best explains Alice's utterance. If Bob further assumes that if Alice took it for granted then it must be true, then he will arrive at the projected content: John used to smoke.

There are different types of projective content (Tonhauser, Beaver, Roberts, & Simons, 2013). It is possible that they project for different reasons. For change-of-state verbs, there are several reasons why one might prefer a pragmatic approach to projection to a semantic one. First, they systematically show projective behavior. Therefore, a generalization would be missing if their projective contents are lexically-encoded properties that could vary arbitrarily from verb to verb. In contrast, a pragmatic approach could in principle explain why a class of verbs with a similar basic meaning would also have the same projection behavior. Second, projection interacts with the contextual question under discussion, as can be seen from the following example (Geurts, 1995). Imagine that Bob asked Alice: "I notice that John keeps chewing on his pencil. Did he recently stop smoking?" In this context Bob is not interested in whether John is currently a smoker, but

[1] Alice's answer is indirect and complex, and hence would be infelicitous without additional contextual justification. This is predicted by our model.

in whether there was a change from smoking to non-smoking which could explain John's odd behavior. As a result, if Alice were to answer "no (it's just an nervous habit)", Bob would not infer that John used to smoke. Similarly, if a customer asks whether an item on sale has been used by anyone before, a reply "it is not refurbished" would imply that it is brand new. Third, projection is sensitive to prosodic focus: "John did not **stop** smoking" seems to suggest he never smoked. Yet a major obstacle to adopting the pragmatic approach has been the difficulty in formalizing the reasoning and showing that it emerges naturally from conversational principles.

In this paper, we build on and formalize previous ideas of the pragmatic approach. We do so within the Rational Speech-Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013): a Bayesian approach to language understanding. We extend the previous models by allowing the listener to reason about the facts that the speaker took to be in common ground (along similar lines to Degen, Tessler, and Goodman (2015)). We find that to account for the example scenario we must also make certain assumptions about which facts are plausible common ground and which question is under discussion. With these assumptions and extensions, the model accounts for projection phenomena of change-of-state verbs under negation. It further predicts an interaction between projective behavior and the *question under discussion* (QUD) (Roberts, 2012), suggesting further experimental research to the growing body of literature (e.g., Cummins, Amaral, & Katsos, 2013; Schwarz, 2015)

A Rational Speech-Act (RSA) model

In this section we introduce an extension to Rational Speech-Act (RSA) model (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) to account for the projection phenomenon of change-of-state verbs under negation, by formalizing the ideas introduced in the previous section. We will continue to use our working example of the conversation between Alice and Bob regarding John's smoking habit.

We consider the following relevant utterances: "John smokes," "John smoked," "John has always smoked," "John stopped smoking," "John started smoking," "John has never smoked," and their negations. In addition, we introduce the null utterance "" (say nothing). The prior probability of an utterance depends on the number of content words (i.e., negation and auxiliaries are excluded) that it has. The shorter an utterance, the higher its prior probability is, as defined in (1).

$$\Pr(u) \propto 2^{-\#\text{content-words}(u)} \tag{1}$$

The meaning/denotation of an utterance is standardly defined as the set of worlds where the utterance is true. We define a world w as a pair. Its first element is whether John smoked in the past and its second element is whether John smokes now. This gives us a set of four possible worlds (the *universe* $U = \{(T,T),(T,F),(F,T),(F,F)\}$). All positive utterances and their denotations are listed in Table 1. In addition, we define that saying nothing is always true, and that the denotation of the negation of an utterance u is $U - [\![u]\!]$.

u	$[\![u]\!]$
"John smokes"	$\{(T,T),(F,T)\}$
"John smoked"	$\{(T,T),(T,F)\}$
"John has always smoked"	$\{(T,T)\}$
"John stopped smoking"	$\{(T,F)\}$
"John started smoking"	$\{(F,T)\}$
"John has never smoked"	$\{(F,F)\}$

Table 1: Positive utterances and their denotations

A *Question Under Discussion* (QUD) is a function Q that takes a possible world as its argument and returns the answer to the question in this world. For example, QUD_{now} is the question "Does John smoke now?" It takes a world and returns its second element, which answers whether John smokes now. Another example is the maximal QUD_{max}, which is the identity function. Intuitively, QUD_{max} is asking which is the actual world. It is maximal in the sense that knowing its answer means knowing the answer to any QUD.

To account for projective behavior, we propose additional components and assumptions to the standard RSA model in the literature (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Lassiter, 2015). To better illustrate why each of them is necessary and how they contribute to the model's prediction, we will present the model incrementally. We will start with the standard RSA model, point out its problems, motivate a modification, explain the problem it addresses, review the remaining issues, motivate another modification, and so on, until we reach the final model.

Standard RSA model

In the standard RSA model (augmented with QUD as in Goodman and Lassiter (2015)), the literal listener, given an utterance and a QUD, randomly samples a world that is consistent with the utterance, and returns the value of the QUD in that world, as in (2). In this paper we always assume that all worlds are equally likely *a priori*, i.e. $\Pr(w) = 1/4$ for each w.

$$L_0(Q(w) \mid u, Q) \propto \sum_{w' \in [\![u]\!]} \delta_{Q(w)=Q(w')} \cdot \Pr(w') \tag{2}$$

Here δ subscripted with a statement is defined to be 1 if the statement is true and 0 otherwise. For example, if Q is QUD_{max}, which asks for a complete specification of the state of the world, after hearing "John didn't stop smoking," the literal listener will rule out world (T,F), and return the remaining 3 worlds with equal probability.

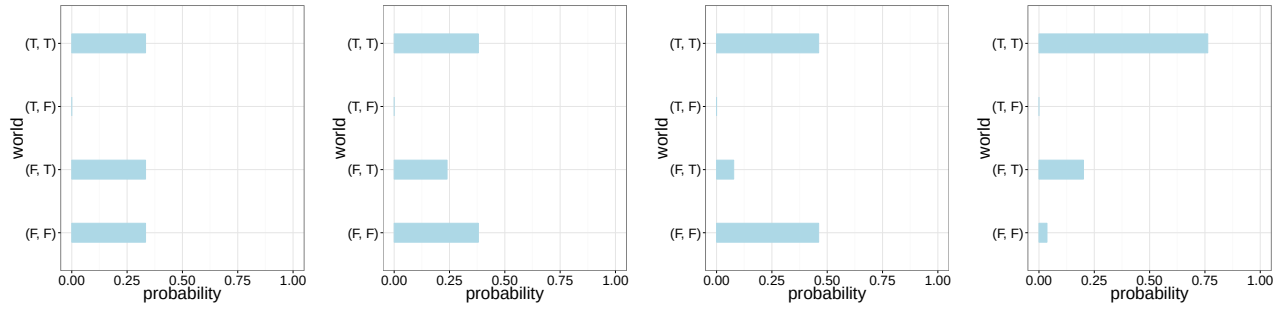
Given the actual world and the QUD, the probability of the speaker's utterance u depends on two factors: the utterance prior and the probability that the utterance will make the literal listener return the correct answer to the QUD, as in (3).

$$S(u|w,Q) \propto \Pr(u) \cdot L_0(Q(w) \mid u, Q)^{\alpha} \tag{3}$$

Here α is a rationality parameter controlling the extent to which the speaker optimizes her utterance to induce the correct answer from the literal listener. When $\alpha \to \infty$, the speaker will always choose utterances that strictly maximize the probability of inducing the right answer. In this paper we set $\alpha = 6$, but the qualitative predictions do not hinge on this specific value.

	Standard (no CG + QUD_{max})	Uniform CS + QUD_{max}	CG prior + QUD_{max}	CG prior + QUD_{now}
literal	$L_0(Q(w) \mid u, Q) \propto \sum_{w' \in [\![u]\!]} \delta_{Q(w)=Q(w')} \cdot \Pr(w')$	$L_0(Q(w) \mid u, C, Q) \propto \sum_{w' \in C \cap [\![u]\!]} \delta_{Q(w)=Q(w')} \cdot \Pr(w')$		
speaker	$S(u \mid w, Q) \propto \Pr(u) \cdot L_0(Q(w) \mid u, Q)^{\alpha}$	$S(u \mid w, C, Q) \propto \Pr(u) \cdot L_0(Q(w) \mid u, C, Q)^{\alpha}$		
listener	$L(w \mid u, Q) \propto \Pr(w) \cdot S(u \mid w, Q)$	$L(w, C \mid u, Q) \propto \Pr(w) \cdot \Pr(C) \cdot S(u \mid w, C, Q)$		
CG prior	–	$\Pr(C) \propto 1$	$\Pr(C) \propto \sum_{CG \subseteq \text{Obs}} P(CG) \cdot \delta_{C=\cap CG}$	
QUD	$\text{QUD}_{\text{max}}(w) = w$	$\text{QUD}_{\text{max}}(w) = w$	$\text{QUD}_{\text{max}}(w) = w$	$\text{QUD}_{\text{now}}((x,y)) = y$

Table 2: Specifications of four RSA models , with $\Pr(w) \propto 1$ and $\Pr(u) \propto 2^{-\#\text{content-words}(u)}$ for all four models



(a) No CG + QUD_{max} (b) Uniform CS + QUD_{max} (c) CG prior + QUD_{max} (d) CG prior + QUD_{now}

Figure 1: Pragmatic listener after hearing "John did not stop smoking" for each model, with $\alpha = 6$

The pragmatic listener, given the QUD, infers the actual world given the speaker's utterance, using Bayes' rule (4).

$$L(w \mid u, Q) \propto \Pr(w) \cdot S(u \mid w, Q) \tag{4}$$

The standard RSA model is summarized in the first column of Table 2, and the predicted pragmatic listener's distribution over worlds is shown in Figure 1(a). As we can see, the standard RSA model predicts a uniform distribution over the three worlds that are consistent with the literal meaning of "John did not stop smoking". It therefore fails to capture the projective content — the inference that John used to smoke.

The reason for this failure is that "John did not stop smoking" is equally under-informative in any of the three worlds compatible with its literal meaning. For example, suppose the actual world is (T,T). Since the literal listener will return this world with probability only $1/3$ after hearing "John did not stop smoking," the speaker is unlikely to choose this utterance. She is more likely to say "John has always smoked" instead, which will always induce the correct answer. The same holds for the other two worlds (F,T) and (F,F) and therefore the pragmatic listener in the standard RSA model will infer that the three worlds are equally likely.

RSA with common ground

We have seen that one important reason that the standard RSA model fails to capture the projective content of "John did not stop smoking" is that its literal meaning is under-informative when considered in the entire universe U. However, as discussed before, there can be information taken for granted by the speaker and the listener, i.e., the common ground, and an utterance that is under-informative when considered in the entire universe U may nevertheless be informative when evaluated in the common ground. To formalize this observation, we now add common ground to the RSA model.

We first define a related notion. A *context set* C is a non-empty subset of the universe (Stalnaker, 1974). Since we have 4 possible worlds, there are $2^4 - 1 = 15$ different context sets. These context sets are intuitively named. For example, `+past` is the context set that contains (T,T) and (T,F), `+past+now` contains only (T,T), $\sim$`+past+now` contains all the worlds except (T,T) ($\sim A$ is defined to be $U - A$, i.e., A's complement), and `change` is the context set that contains (T,F) and (F,T).

A literal listener, given an utterance, the current context set and QUD, randomly samples a world that is consistent with both the utterance and the context set, and returns the value of the QUD in that world, as in (5).

$$L_0(Q(w) \mid u, C, Q) \propto \sum_{w' \in C \cap [\![u]\!]} \delta_{Q(w)=Q(w')} \cdot \Pr(w') \tag{5}$$

For example, given context set `+past` and QUD_{max}, after hearing "John did not stop smoking," the literal listener will rule out (T,F) because of the utterance's literal meaning, and (F,T) and (F,F) because they are incompatible with the context set. Therefore he will always return (T,T). We can see from this example that an utterance that is under-informative when the entire universe is considered can be informative in some other context sets.

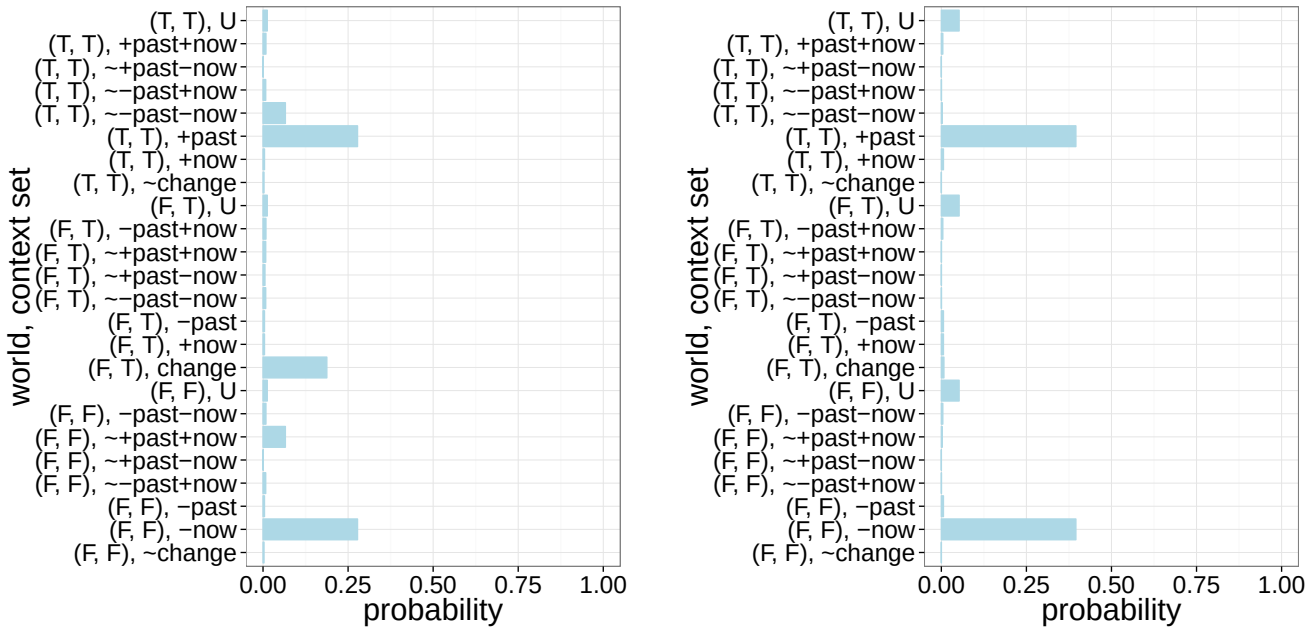
The new speaker model is almost the same as (3), except that it is relativized to the current context set, as in (6).

$$S(u \mid w, C, Q) \propto \Pr(u) \cdot L_0(Q(w) \mid u, C, Q)^{\alpha} \tag{6}$$

Finally, given an utterance and the QUD, the pragmatic listener now jointly infers the real world and the context set the speaker assumes when she produces the utterance.

$$L(w, C \mid u, Q) \propto \Pr(w) \cdot \Pr(C) \cdot S(u \mid w, C, Q) \tag{7}$$

We need to specify a prior distribution $\Pr(C)$ over context sets in (7). We consider two possibilities. First, we consider a uniform distribution over all context sets, i.e., $\Pr(C) \propto 1$.



(a) Uniform CS + QUD_{max}

(b) CG prior + QUD_{max}

Figure 2: Pragmatic listener after hearing "John did not stop smoking," with $\alpha = 6$

Assuming the maximal QUD, the model is summarized in the second column of Table 2 and the pragmatic listener's marginal distribution over worlds is shown in Figure 1(b). We can see that this model predicts that (F,T) is slightly less likely than (T,T) and (F,F), and (T,T) has the same probability as (F,F). This does not capture projection.

The second possibility makes use of the notion of a *common ground* (*CG*) in the pragmatic approach to derive a prior over context sets (Stalnaker, 1974). Intuitively, a common ground represents everything that is taken for granted for conversational purposes. Formally it is a set of propositions (a proposition is a set of worlds), all of which are taken for granted. The context set C, as defined above, can be thought of as the conjunction of all of the propositions that are being taken for granted: $C = \bigcap CG$.

In our example scenario, Alice (the speaker) could reasonably take for granted certain propositions representing plausible observations about John's smoking habits — whether he smoked in the past, and whether he does now. Therefore, assuming that the propositions in the common ground come from observations about John's past and present smoking habits, we can use (8) to naturally define a prior over context sets (henceforth the CG prior).

$$\Pr(C) \propto \sum_{CG \subseteq \text{Obs}} P(CG) \cdot \delta_{C=\cap CG} \tag{8}$$

Concretely, we assume that each of the observations enters the common ground independently, with probability .4 (meaning that the speaker does not tend to take things for granted). In addition, we add a small amount (5%) of noise to (8), so that every non-empty C has a nonzero prior probability. This model assigns low prior probability to those context sets that cannot be built up via conjunctions of natural observations. One example of such a context set is `change`, the rather complex assumption that John has *either* switched from smoking to not, *or* switched from not smoking to smoking — but we do not know which. In contrast, context sets such as `+past` (i.e., taking for granted that John used to smoke) and `-past-now` (i.e., John did not smoke and does not smoke now) receive higher probabilities because they correspond to observations that the speaker could plausibly have made.

Using the CG prior (the model is summarized in the third column of Table 2), the pragmatic listener's marginal distribution over worlds is shown in Figure 1(c). We can see that this model predicts that world (F,T) is very unlikely, and world (T,T) has the same probability as world (F,F). Although it still does not capture projection because (T,T) is predicted to be as likely as (F,F), the model correctly predicts that (F,T) is unlikely. Therefore we have made some progress.

To better understand how the CG prior improves the model and what the remaining problem is, we plot the pragmatic listener's joint distribution of world and context set in Figure 2.

In Figure 2(a), with a uniform prior over context sets, the pragmatic listener has 3 most likely outcomes: world (T,T) with context set `+past`, world (F,F) with context set `-now`, and world (F,T) with context set `change` (this last outcome is slightly less likely than the first two). This is because in these outcomes, "John did not stop smoking" can fully identify the world given the context set, and hence these outcomes best explain the speaker's utterance. As a result, the marginal distribution over worlds is almost uniform over the 3 worlds.

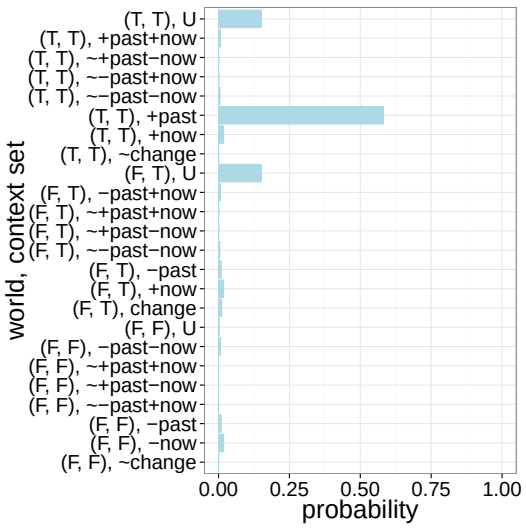


Figure 3: Pragmatic listener after hearing "John did not stop smoking," with $\alpha = 6$, CG prior + QUD_{now}

In contrast, in Figure 2(b), with the CG prior, world (F,T) with context set `change` is no longer a likely outcome, because as noted earlier, the context set `change` is assigned a very low prior. This is why world (F,T) is correctly predicted to be unlikely. Although the CG prior we introduce above is probably over-simplified, the crucial assumption we need is just that not all context sets are equally likely *a priori*, and in particular `change` is a fairly unusual context set and should be assigned a low prior probability, which seems intuitively plausible. As long as this assumption is satisfied, there could be alternative ways to motivate a more realistic prior over context sets without affecting the model's qualitative prediction.

Nevertheless, we can see that world (F,F) with common ground `-now` is still one of the most likely outcomes in Figure 2(b), and hence the marginal probability of (F,F) is the same as (T,T) in Figure 1(c). This is not desirable, but is totally expected from the model: the prior for context set `-now` is the same as for context set `+past`. Therefore, to fully capture projective behavior, we need to further explain why (F,F) with context set `-now` is dispreferred.

Non-maximal QUDs So far, we have been assuming that the QUD is maximal, i.e., the utterance "John did not stop smoking" is chosen to address the question of whether John smoked in the past and whether John smokes now. For this QUD, the RSA model with common ground prior predicts a tie between (T,T) with context set `+past` and (F,F) with context set `-now`.

The maximal QUD is often assumed in applications of RSA models (though see Kao, Bergen, and Goodman (2014)), but in this case there are good reasons to consider non-maximal QUDs. Empirically, as noted in the beginning, projection is sensitive to the QUD. Theoretically, there has been a lot of discussion in the previous literature on the relation between at-issueness and projection (Beaver, 2010; Simons, Tonhauser, Beaver, & Roberts, 2010). Their notion of at-issueness roughly corresponds to QUDs in our model, which may be non-maximal.

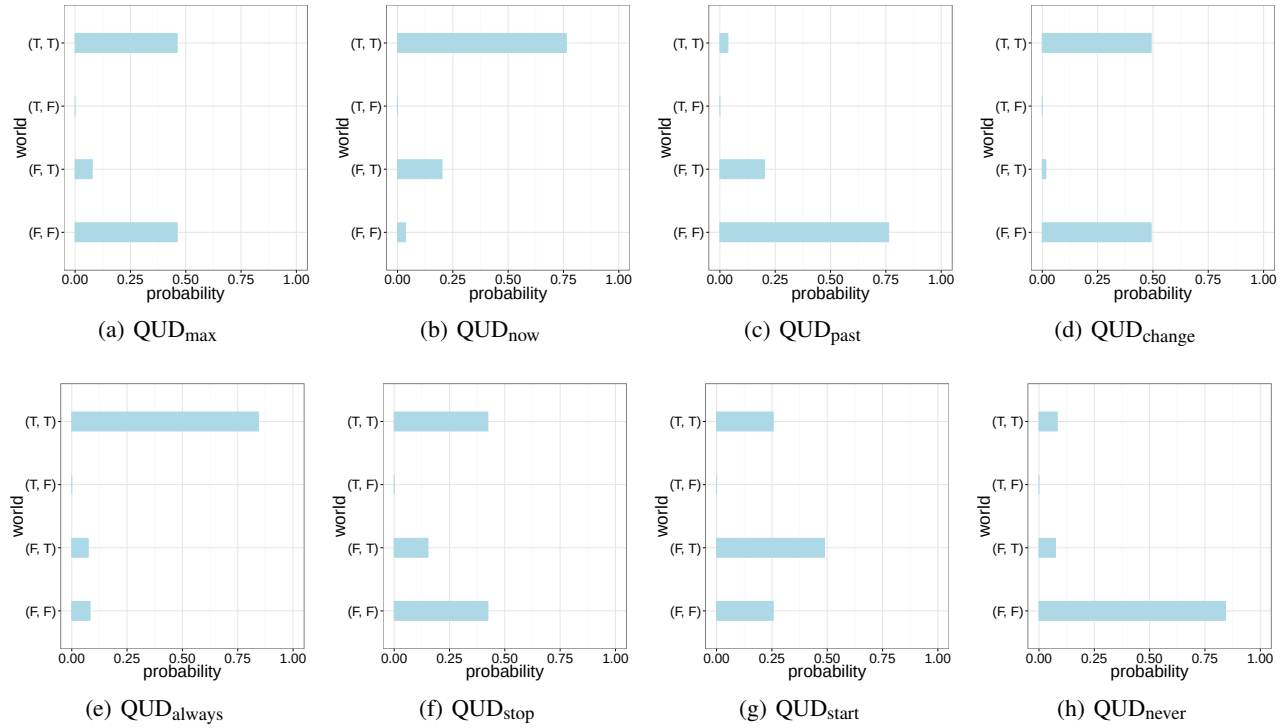
In our running example, Bob explicitly asked about whether John smokes, which means that the QUD is presumably QUD_{now} (i.e., "Does John smoke?"). When we use the previous RSA model with the common ground prior, but replace QUD_{max} with QUD_{now} (summarized in the last column of Table 2), the pragmatic listener's marginal distribution over worlds is shown in Figure 1(d) and the joint distribution of world and context set is in Figure 3. We can see from Figure 3 that (T,T) with context set `+past` is the only most likely outcome, and the world (T,T) is the only most likely world (and its probability increases with a higher α). This is exactly the projection pattern we aim to capture.

To understand why we obtain this result, we note that when the QUD is QUD_{now}, (F,F) with context set `-now` is dispreferred because the context set `-now` already entails the answer to the QUD. That is, it is already known from the context set `-now` that John does not smoke now. This means that the speaker would be maximally informative even if he says nothing. As a result, the speaker would be unlikely to say "John did not stop smoking" when the context set is `-now`, and the pragmatic listener could therefore infer that the context set `-now` is unlikely, which means that (T,T) with context set `+past` is the only winner.

Hence the current model predicts the projective behavior for "John did not stop smoking" in the example scenario, where the QUD is explicitly given by Bob's question. Assuming that people generally care about information about now rather than the past, i.e., the default or most salient QUD is QUD_{now}, the model predicts that the preferred projective content of "John did not stop smoking" without explicit QUD is that John smoked in the past. Note that our speaker model predicts that "John did not stop smoking" is very unlikely to be used to answer QUD_{now}, as there exist simpler utterances "John smokes/does not smoke." This explains the perceived weirdness of Alice's indirect answer to Bob's explicit question.

Other QUDs We have introduced a RSA model with common ground and shown its predictions for QUD_{now} and QUD_{max}. The prediction is sensitive to the QUD—in Figure 4 we show predictions for eight different QUDs. In general, it seems that the model is making plausible predictions. The utterance "John did not stop smoking" implies that John has always smoked if the QUD is whether John has always smoked (Figure 4(e)).[2] It implies that John has never smoked when the QUD is whether John has never smoked (Figure 4(h)). When the QUD is about whether John stopped smoking (Figure 4(f)) or there is a change (Figure 4(d)), the listener will believe that John smoking in the past is about 50% likely (recall Geurts' example described earlier). These results are compatible with Simons et al.'s generalization that only non-at-issue content projects. But further experimental data will be needed to assess whether they are borne out.

[2]Note that QUDs are just partitions and not presuppositional. We use, e.g., *always* to describe the QUDs only as a mnemonic device.



(a) QUD_{max} (b) QUD_{now} (c) QUD_{past} (d) $\text{QUD}_{\text{change}}$

(e) $\text{QUD}_{\text{always}}$ (f) QUD_{stop} (g) $\text{QUD}_{\text{start}}$ (h) $\text{QUD}_{\text{never}}$

Figure 4: Pragmatic listener after hearing "John did not stop smoking" for different QUDs, with $\alpha = 6$ and CG prior

Conclusion and future directions

In this paper, we introduced a probabilistic model in the RSA framework, which analyzes the projective content of change-of-state verbs under negation as the result of the listener using general conversational principles to jointly infer the actual world and the context set that the speaker assumes The model predicts an interaction between projection and the question under discussion, formalizing insights of previous pragmatic approach to projection and providing concrete quantitative predictions that we plan to test experimentally.

Our model is a first step towards a full integration of pragmatic approach to the projection problem into a general probabilistic model of language understanding. We plan to further explore the model and see to what extent it can be generalized to other types of projective content and other entailment-canceling operators, which will help us further understand the division of labor between semantics and pragmatics in the projection phenomena.

References

Beaver, D. (2010). Have you noticed that your belly button lint colour is related to the colour of your clothing. In *Presuppositions and discourse: Essays offered to Hans Kamp* (pp. 65–99). Philadelphia, PA: Elsevier.

Carroll, L. (1866). *Alice's Adventures in Wonderland*. Macmillan.

Cummins, C., Amaral, P., & Katsos, N. (2013). Backgrounding and accommodation of presuppositions: an experimental approach. In E. Chemla, V. Homer, & G. Winterstein (Eds.), *Proceedings of Sinn and Bedeutung 17* (pp. 201–218).

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the thirty-seventh annual conference of the Cognitive Science Society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998. doi: 10.1126/science.1218633

Frege, G. (1948). Sense and reference. *Philosophical Review*, *57*, 209–230. (Originally published in 1892)

Geurts, B. (1995). *Presupposing*. Doctoral dissertation, University of Stuttgart.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory, 2nd edition.* Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184. doi: 10.1111/tops.12007

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell Oxford.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*(6), 1–69. doi: 10.3765/sp.5.6

Schwarz, F. (Ed.). (2015). *Experimental perspectives on presuppositions.* Switzerland: Springer International Publishing.

Simons, M. (2001). On the conversational basis of some presuppositions. In R. Hastings, B. Jackson, & Z. Zvolenszky (Eds.), *Semantics and linguistic theory (SALT) 11* (pp. 431–448). Ithaca, NY: Cornell University.

Simons, M. (2006). Foundational issues in presupposition. *Philosophy Compass*, *1*(4), 357–372.

Simons, M., Tonhauser, J., Beaver, D., & Roberts, C. (2010). What projects and why. In N. Li & D. Lutz (Eds.), *Semantics and linguistic theory (SALT) 20* (pp. 309–327). Ithaca, NY: Cornell University.

Stalnaker, R. C. (1974). Pragmatic presuppositions. In M. K. Munitz & P. K. Unger (Eds.), *Semantics and philosophy* (pp. 197–214). New York: New York University Press.

Tonhauser, J., Beaver, D., Roberts, C., & Simons, M. (2013). Toward a taxonomy of projective content. *Language*, *89*(1), 66–109. doi: 10.1353/lan.2013.0001

Infants' speech and gesture production in Mozambique and the Netherlands

Chiara de Jong (c.dejong_3@tilburguniversity.edu)
Paul Vogt (p.a.vogt@uvt.nl)
Tilburg centre for Cognition and Communication, Tilburg University
P.O. Box 90153, 5000 LE, Tilburg, The Netherlands

Abstract

In this paper, we explore the cultural differences in the production of speech and speech+gesture combinations by infants at the age of 17-18 months in Mozambique and the Netherlands. We found that Dutch infants produce more speech and gestures compared to Mozambican infants. Infants in both communities make most use of content words. The results further show that Dutch infants make more use of proximal pointing than Mozambicans, whereas Mozambicans make more use of the offering gesture. Finally, the amount of semantically coherent speech+gesture combinations of the Mozambican infants is higher than of the Dutch infants.

Keywords: Child language acquisition; culture; infant speech; infant gesture; semantic coherence; speech+gesture combinations.

Introduction

For infants in the early stages of language, gestures can allow them to express certain thoughts at a time they may be unable to express those thoughts fully in the spoken modality (Butcher & Goldin-Meadow, 2000; Goldin-Meadow & Butcher, 2003). Additionally, early gesture use in infants has been found to be a predictor of their later language development (Rowe, Özçalışkan & Goldin-Meadow, 2006, 2008). Despite the importance of gestures for language development, little knowledge is available about the use of (co-speech) gestures by infants in a cross-cultural perspective, especially in non-Western societies. In this paper we investigate the amount of speech and co-speech gestures produced by infants in two cultures. Additionally, we investigate the relation between different kinds of gestures and accompanying speech.

It is well established that there are considerable individual and cross-cultural differences in the language-learning environment of children (Hoff, 2006). For instance, concerning the amount of infant-directed speech (IDS; Lieven, 1994), the communicative intentions of IDS (LeVine et al., 1994; Vogt, Mastin & Schots, 2015) and style of co-speech gestures addressed to infants (Gogate, Maganti, & Bahrick, 2015; Tamis-LeMonda, Song, Leavell, Kahana-Kalman, & Yoshikawa, 2012). Similar results have been found concerning the influence of socio-economic status on the language environment of children (Hart & Risley, 1995; Rowe & Goldin-Meadow, 2009). Moreover, the various differences concerning the input to children have been found to affect children's development of language and other communicative skills (Hart & Risley, 1995; Rowe & Goldin-Meadow, 2009; Tamis-LeMonda et al., 2012). However, while such cultural differences concerning the learning environment of children are well known, they mostly concern the verbal environment or more Westernized cultures. Still little is known about (the role of) gestural communication in children ('s learning environments) from non-Western societies, such as those from Africa.

Studies have shown that the amount of gestural input, just like vocal input, varies across sociocultural settings and is not just dependent of SES or parents educational level (Callaghan et al., 2011; Iverson, Capirci, Volterra, & Goldin-Meadow, 2008; Salomo & Liszkowski, 2013; Vogt & Mastin, 2013). A recent observational study from Mozambique has obtained some novel insights. For instance, it was found that infants in urban Mozambique were stimulated much more frequently compared to infants from rural Mozambique across their whole gesture repertoire and were exposed to three times as much speech and co-speech gestures, which correlated to differences found in their vocabulary development (Vogt & Mastin, 2013). Also the communicative intentions of IDS differ considerably between the Netherlands, and urban and rural Mozambique (Vogt, et al., 2015). In particular, IDS in the Netherlands contained more utterances with a cognitive intention, whereas IDS in urban Mozambique contained more utterances with a socio-emotional intention, and IDS in rural Mozambique contained most of the imperatives. Additionally, the amount of IDS in the Netherlands was almost twice as much as the input in urban Mozambique, and about ten times as much as in rural Mozambique (Vogt et al., 2015).

The observed differences correspond to differences in caregiving practices between the communities, which can be explained based on lifestyles of the different socio-cultural communities (Keller, 2012). In a Western context, parents focus more on the cognitive development of the infant, such as language development. In a non-Western context the focus is more on situation-oriented socialization, with more attention for the development of motoric skills.

It is well established that the input in speech and gesture tends to correlate with children's development of speech and gesture (Hart & Risley, 1995; Iverson et al., 2008; Rowe & Goldin-Meadow, 2009; Salomo & Liszkowski, 2013). We therefore expect that the observed differences in the input to infants from the Netherlands and Mozambique will reflect differences in the *production* of speech and gesture between infants from the Netherlands and Mozambique. However, little is known about how infants in non-Western communities produce gesture and speech in everyday communication.

Salomo and Liszkowski (2013) have shown that the gesture productions of infants from the Netherlands, China

and a Mayan culture varied considerably with Chinese infants producing most gestures, followed by Dutch and Mayan infants. Assessing another three cultures (Canada, Peru and India) in a pointing task, Callaghan et al. (2011) found that significantly more Canadian infants produced declarative pointing gestures compared to Indian infants of the same age, while the number of Peruvian infants who pointed was in between.

The amount of gestures infants produce predicts the early stages of their spoken language development. Rowe et al. (2006; 2008) have shown that parents' use of gestures did not directly influence children's vocabulary size at 42 months. However, parental gesture use did influence the amount of gestures produced by infants at 14 months, which in turn was a significant predictor of vocabulary development at 42 months. Moreover, Rowe and Goldin-Meadow (2009a) found that children from low SES families produced fewer gestures than children from high SES families, which related to their parents' gestural input.

Children's gesture production is of importance since it does not only predict later vocabulary size, but is also involved in other developmental milestones, like the onset of the two-word stage. Children who are unable to express two words within one utterance can make use of a speech+gesture combination to express a two-word idea (Butcher & Goldin-Meadow, 2000; Goldin-Meadow & Butcher, 2003; Rowe & Goldin-Meadow, 2009b). The age at which children start to produce these speech+gesture combinations predicts the onset of two-word utterances (Iverson, et al., 2008; Iverson & Goldin-Meadow, 2005; Rowe & Goldin-Meadow, 2009b). Speech+gesture combinations conveying different information were found to emerge prior to the onset of two-word combinations for all cultures studied so far (Goldin-Meadow, 2009; Iverson et al., 2008; Iverson & Goldin-Meadow, 2005), but these do not include non-Western cultures.

The purpose of this paper is to explore the cultural differences in the speech and gestures infants produce at the age of 17-18 months. In particular, we compare the speech and gesture production of children from middle-class families in the Netherlands with children from low SES families in urban Mozambique. We assess infants' amounts and semantic categories of speech in combination with gesture production, as well as the mean length of utterances and speech+gesture utterances they produce. While investigating the relation of these differences with other developmental milestones, such as vocabulary development or the onset of two-word utterances, would be interesting, this is left for future research.

Methods

Participants and field sites

This study was based on data from earlier research of Mastin and Vogt (2016) and Vogt et al. (2015). The data consists of video recordings of infants in the Netherlands and urban Mozambique at three different ages, when infants were 1;1, 1;6 and 2;1 years old. For the purpose of the present study, only data recorded at 1;6 was used. Since the infants in rural Mozambique did not express many gestures and speech at the age of 1;6, that field site was excluded.

Table 1: Demographic information. *Note:* [a]Maternal education data from one Mozambican participant is missing.

Participant information	Mozambique (N = 11)	Netherlands (N = 11)
Female / male	5/6	5/6
Avg age (SD)	1;5.11 (29)	1;5.10 (20)
Education level mothers		
No education	1[a]	0
Primary education	8	0
Secondary education	1	0
Higher education	0	11

The field site in urban Mozambique was made up of two adjacent residential suburbs in Maputo. The community is relatively poor, with low parental education, and is has a market-based lifestyle. The Dutch sample consists of highly educated, native Dutch families living in the Tilburg region. The sample reflects a typical Western, middle-class, post-modern urban community that has a knowledge-based lifestyle.

The results reported here are from 11 infants from the Dutch community and 11 from urban Mozambique. The average age at the time of recording was 1;5.10 years (Table 1). In Mozambique, local research assistants explained the general purpose and recording procedures of the study to the families in their native language (Portuguese or Changana). In the Netherlands this was done by one of the researchers. Signed informed consent forms were obtained from the infants' mothers. Table 1 shows some basic demographics of our sample. For more details consult Vogt et al. (2015).

Data collection procedures

The families were visited at their home. During those visits naturalistic interactions were recorded of infants' daily, social environment by instructing those people who were present to carry on with what they were doing prior to our visit and to ignore our presence as well as possible. In Mozambique each visit was preceded by an accommodation session a week ahead in order to familiarize infants and their families with the presence of a foreign person who observes them with a video camera. In the Netherlands such an accommodation session was considered unnecessary since the infants have been exposed to video equipment before and the researcher was not foreign to them.

The infants were filmed for 45 to 75 minutes to obtain at least 30 minutes of clear and naturalistic footage for analysis. Video recording was done using a tripod whenever possible, placed approximately 5 meters from where most of the interactions occurred. In smaller confined areas, recording was done by hand-held camera.

Coding procedures

The videos were coded for 30 minutes during which the infant was displaying natural behaviour for a prolonged duration (i.e.: not asleep, off camera and interacting with or disturbed by the experimenters). Prolonged periods of breastfeeding (more than 2 minutes) were excluded too, since this could have introduced a bias toward dyadic interactions. The videos have been annotated for infant engagements (Mastin & Vogt, 2016), infant-directed speech and infants' speech (Vogt et al., 2015), and (co-speech) gestures directed to and produced by infants (Vogt & Mastin, 2013). The coding schemes were developed based on the literature and has been piloted and adapted with videos from an earlier pilot study from Mozambique. In this article, we only present the results of infants' speech and gestures. Interrater agreements are assessed and reported based on 35% cross-coding of data, except for the gestures of which only 15% was cross-coded.

Infant speech All speech produced by the infants during those 30 minutes was transcribed by research assistants who were native speakers of the language. The Mozambican data was subsequently translated to English. For the purpose of the present study, a distinction was made between *meaningful* and *meaningless* utterances. An utterance was considered meaningful when the speech was intelligible, or a clear meaning was conveyed. Only meaningful utterances were analysed in this study.

To investigate what type of words infants produced, we coded the semantic category of each utterance. In case the utterance contained only one word, this is the semantic category of that word; in case it was a multiword utterance, we coded the semantic category of the most prominent word in the utterance. We coded the infants' speech into the following three semantic categories (Cohen's kappa=0.90):

Content words. A word was considered to be a content word if it bears meaning on itself (e.g. 'daddy').
Demonstratives/deictics. A word was considered a demonstrative or deictic if it refers to something else, which can be another time, space, person or object or if it was meant to refer someone's attention to something (e.g. 'look'). These are words that cannot be understood without any additional contextual information (e.g. 'there').
Other. The category that was coded as 'other' consisted of words that did not belong to the previous categories, such as affirmatives or interjections (e.g. 'yes', 'so').

Gestures The gestures produced by the infants were coded during episodes of joint engagement. A gesture is broadly defined as any physical activity with the hand or body that has a clear communicative intent (Zukow-Goldring, 1996). For this research the coding from Vogt and Mastin (2013) and Vogt, Masson-Carro and Mastin (2015) was adopted. The following categories were distinguished (Cohen's kappa=0.58):

Proximal pointing. The infant points at an object in the near vicinity using his index finger or hand, in order to draw the attention of another individual to the target object.
Distal pointing. The infant points at an object placed far away using his index finger or hand, in order to draw the attention of another individual to the target object.
Showing. The infant holds an object, drawing another individual's attention to it, without any direct imperative.
Reaching. The infant moves his hand towards a target object with the intention to obtain it, but does not obtain said object. Also requests for objects by extending the hand were included in this category.
Offering. The infant offers (or gives) an object to another individual with the intention of transferring the object.
Taking. The infant takes an object from the other individual's possession while it has not been offered by the other.
Conventional gestures. This category comprises gestures that are symbolic of nature, such as emblematic gestures, but also gestures that bear an iconic relationship with their referent. For example, waving bye-bye, or indicating the size of the target object with the hands.
Ritualized play. This category accounts for all ritualized interactions or displays that occur between infants and communication partners. For instance, dancing, clapping hands or turn-taking games, such as patty-cake.
Request for attention. This category comprises any gesture that seeks for the attention of the addressee.

Speech gesture combinations In order to determine whether the gestural and vocal modalities are unified into one communication system, gesture and speech combinations were further coded. Unification is characterised by both semantic coherence and temporal synchrony (Butcher & Goldin-Meadow, 2000; Goldin-Meadow & Butcher, 2003). In case an utterance was not accompanied by a gesture, the utterance was coded as *speech only*. When an utterance was accompanied by a gesture, this was coded as a *speech+gesture combination*. A combination was categorized as *semantically coherent* when it was combining gesture with meaningful and related speech (Butcher & Goldin-Meadow, 2000). Following Goldin-Meadow and Butcher (2003), semantic coherence was further divided into two categories (Cohen's kappa=0.66):

Semantically congruent. A speech+gesture combination is semantically congruent when both speech and gesture bear the same meaning. An example of such a combination is a child pointing at a dog and uttering the word 'dog'.
Semantically related. A speech+gesture combination is semantically related when the meaning of the speech is different from the meaning of the gesture, but combined they form a semantically coherent utterance. For instance, a child shows a pair of glasses while uttering 'mommy', meaning that the pair of glasses is their mother's.

Table 2. Average number of utterances per infant, specified by semantic category. *Notes*: +p<.10; *p<.05; **p<.01; ***p<.001.

	Mozambique		Netherlands	
	Mean (SD)	Mean % (SD)	Mean (SD)	Mean % (SD)
All utterances	39.64 (27.22)		73.36 (41.76)+	
Content word	28.18 (22.26)	66.18 (28.34)	36.45 (25.91)	51.32 (23.80)
Deictic/ demonstratives	9.18 (11.27)	28.05 (30.06)	14.82 (14.30)	16.71 (11.74)
Other	2.18 (1.72)	5.78 (6.05)	22.00 (12.47)	31.98 (21.19)***
Speech+gesture	10.00 (6.96)		22.09 (14.90)*	
Content word	5.18 (3.95)	52.23 (27.20)	12.09 (10.26)	56.40 (29.68)
Deictic/ demonstratives	4.55 (4.87)	46.04 (27.87) +	6.55 (7.03)	26.49 (19.77)
Other	0.27 (0.47)	1.72 (3.17)	3.36 (3.72)	17.11 (24.88)*
Speech only	29.73 (21.64)		51.27 (30.14)	
Content word	23.09 (18.67)	81.77 (20.64)**	24.36 (17.61)	48.91 (23.30)
Deictic/ demonstratives	4.64 (7.70)	11.54 (19.81)	8.27 (7.63)	13.26 (9.97)
Other	1.91 (1.51)	6.69 (6.83)	18.64 (10.74)	37.84 (20.10)***

Results

Infant speech

Table 2 shows the average number of utterances and speech+gesture combinations expressed by the infants. In Mozambique the average number of utterances (including both speech only and speech-gesture combinations) per child was M=39.64 (SD=27.22), against M=73.36 (SD=41.76) in the Netherlands. A Mann-Whitney U test showed that this difference tended towards significance (U=31.00, p=.056). Of the total number of utterances, 30.11% was combined with a gesture in the Netherlands (n = 22.09 average per child) and 25.23% in Mozambique (n = 10.00 average per child), which resulted in a significantly larger amount of speech+gesture combinations in the Netherlands compared to Mozambique (U=26.00, p=.023). To analyse the semantic categories of speech between Mozambican and Dutch infants, we compared the mean percentages of each category. In both communities gestures were most often accompanied by content words, 52.23% in Mozambique and 56.40% in the Netherlands. However, in Mozambique, a gesture was more frequently accompanied by a demonstrative or deictic (46.04%) than in the Netherlands (26.49%). This difference is marginally significant (U=34.00, p=.088). In Mozambique, gestures were significantly less frequently accompanied by *other* words (1.72%) than in the Netherlands (17.11%; U=26.00, p=.023).

When infants produced speech without a gesture, Mozambican infants significantly more often produced a content word (81.77%) compared to the Netherlands (48.91%; U=18.00, p=.004). For speech only, the percentages of deictics and demonstratives were similar for both cultures. However, Dutch infants produced significantly more *other* words (37.84%) than in Mozambique (6.69%; U=4.00, p<.001).

Infant gesture

Table 3 shows the mean frequencies with which infants produced the various gestures in speech+gesture combinations. The gesture that was used the most in Mozambique was offering (35.50%), whereas in the Netherlands this was proximal pointing (45.27%). Comparison of gesture use in Mozambique with the Netherlands using the Mann-Whitney U test showed that gesture use only differed significantly for these two gestures: proximal pointing (U=2.00, p<.001) and offering (U=20.00, p=.007). For all the other gestures, no significant difference across the two cultures was observed.

Table 4 shows to what extent the speech+gesture combinations were coherent. While the total number of speech+gesture combinations was higher in the Netherlands, the mean percentage of coherent combinations in Mozambique (M=92.86; SD=14.99) was significantly higher than in the Netherlands (M=80.06; SD=15.23; U=20.00, p=.007). Interestingly, in Mozambique infants produced relatively more semantically congruent combinations (M=39.07; SD=40.84) than in the Netherlands (M=9.70; SD=9.07), while Dutch infants produced relatively more semantically related combinations (M=70.36; SD=15.19) compared to Mozambicans (M=53.79; SD=39.58). A Chi square showed that there was a significant association between country and semantic congruency (χ^2(1, 353)=14.33, p<.001).

Table 3: Average gesture use in numbers and frequencies. *Note:* **p<.01, ***p<.001.

	Mozambique		Netherlands	
	M (SD)	% (SD)	M (SD)	% (SD)
Proximal pointing	0.73 (1.42)	4.53 (8.22)	10 (7.92)	46.64 (24.00)***
Distal pointing	1.18 (1.47)	17.07 (19.63)	2.64 (3.47)	10.05 (9.52)
Showing	0.73 (1.01)	5.66 (8.78)	2.09 (2.81)	10.21 (8.83)

Reaching	1.64 (2.25)	14.75 (19.91)	1.91 (2.74)	7.91 (10.47)
Offering	3.55 (4.74)	32.63 (23.30)**	1.82 (4.24)	6.40 (12.27)
Taking	0.27 (0.47)	2.20 (4.49)	0.82 (2.59)	4.45 (7.84)
Conventional	0.36 (0.67)	2.26 (4.25)	1.91 (2.59)	10.77 (19.52)
Ritualized play	0.91 (1.58)	15.06 (27.52)	0.64 (1.12)	2.67 (5.30)
Attention request	0.64 (1.80)	5.84 (15.26)	0.27 (0.65)	0.91 (2.24)

Table 4: Percentage of semantically coherent speech+gesture combinations conveying congruent or related information. *Note:* **p<.01.

	Mozambique		Netherlands	
	M (SD)	% (SD)	M (SD)	% (SD)
Coherent	9.64 (7.00)	92.86 (14.99)**	18.27 (12.48)	80.06 (15.23)
Congruent	5.18 (4.40)	39.07 (40.84)	16.00 (10.70)	9.70 (9.07)
Related	4.45 (6.33)	53.79 (39.58)	2.27 (2.80)	70.36 (15.19)

Discussion

The objective of this paper is to explore the cultural differences in the speech and gestures production of infants at the age of 17-18 months from the Netherlands and Mozambique. Since the sample size was small and the nature of the research exploratory, generalizability is difficult. Moreover, the two communities do not only differ in culture but also in socioeconomic status. Many of the differences that we find may therefore not only be explained due to cultural differences but also to SES differences. Many of the differences that we found are in line with those found in other studies of the same nature (Hart & Risley, 1995; Rowe & Goldin-Meadow, 2009a).

The results revealed that infants from the Netherlands produced twice as many utterances than infants from Mozambique (Table 2). This is also the case for speech+gesture combinations and for speech only productions, but the latter difference was not significant. These differences reflect the difference observed in the amount of IDS, which was also twice as much in the Netherlands compared to urban Mozambique (Vogt et al., 2015).

Looking at the type of utterances, infants from both countries most often produced content words. For utterances not accompanied by gestures, we found that the percentage of content words is higher in Mozambique compared to the Netherlands. This may partly be caused by the low proportion of *other* words (like affirmatives or interjections) observed in Mozambique compared to the proportion of other words produced in the Netherlands. As a result, the utterances are redistributed over content words and deictics, which may cause a skewed distribution.

The higher proportion of content words in the absence of a gesture may also be related to a relatively frequent use of gestures when producing a deictic or a demonstrative. When combining speech with a gesture, Mozambican infants tend to produce more deictics and demonstratives compared to Dutch infants, though this difference is only marginally significant. This tendency may be explained by the high frequency of the Mozambican deictic '*im*', which means 'look at my gesture'. 'Im' is a high frequency word in Changana and Mozambican Portuguese, especially by and for children, and is practically always accompanied by a gesture, thus boosting the frequency of deictics and demonstratives in combination with a gesture.

The relatively low frequency of the *other* semantic category in Mozambique may reflect the lack of diversity of the IDS, because the more diverse the IDS, the more diverse infants' vocabularies tend to grow (Hart & Risley, 1995). Hart and Risley demonstrated that the diversity of IDS strongly relates to socio-economic status or maternal education. Since the maternal education in Mozambique is relatively low, the diversity of IDS may be low, thus explaining low diversity in the infants' production. Future analyses on the type-token ratio should inspect whether this is indeed the case here.

Looking at the different gestures infants produce (Table 4), we see that while Dutch infants produced more than twice as many gestures, the relative frequencies are highly similar for both cultures except for the amount of proximal pointing and the number of offering gestures observed. Dutch infants produced substantially more proximal pointing gestures than Mozambican infants. Almost half of the gestures Dutch children produce are pointing gestures to an object in close proximity. This is consistent with results from other cross-cultural studies (Callaghan et al., 2011; Salomo & Liszkowski, 2013), and may be explained by the fact that pointing facilitates joint attention and word learning, which are fostered more in Western middle class societies than non-Western middle class societies (Keller, 2012).

The higher frequency of the offering (including giving) gesture in Mozambique might be due to the importance of sharing in non-Western cultures as part of the social responsibility (Keller, 2012). Another explanation can be the high amount of imperatives directed to children in Mozambique (Vogt et al., 2015), resulting in children offering their caregivers the things they asked for.

Finally, we observe that in both countries infants' speech+gesture combinations attained a high level of semantic coherence, and thus in both cultures children have a unified communication system in which gesture and speech are integrated to convey one single message (Butcher & Goldin-Meadow, 2000). Despite the larger number of speech+gesture combinations observed in the Netherlands, Mozambican infants produced proportionally more semantically coherent combinations. Moreover, these combinations often convey the same information in Mozambique and are thus semantically congruent, whereas

in the Netherlands the combinations often convey different information, and are thus semantically related. Given the suggested associations between semantic congruency and temporal synchrony, and between semantic relatedness and the onset of two-word utterances (Goldin-Meadow & Butcher, 2003), these findings could predict related cultural differences in the development of temporal synchrony and two-word utterances.

Acknowledgments

We express our gratitude to Doug Mastin, Ingrid Masson-Carro and various other research assistants for their contributions in collecting and annotating the data. Additionally we offer a special thanks to our participant families for their kind hospitality and collaboration.

References

Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech: when hand and mouth come together. In D. McNeil (Ed.), *Language and gesture*. New York: Cambridge University Press.

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., Tomasello, M., & Collins, W. A. (2011). *Early social cognition in three cultural contexts*. Wiley-Blackwell.

Gogate, L., Maganti, M., & Bahrick, L. E. (2015). Cross-cultural evidence for multimodal motherese: Asian Indian mothers' adaptive use of synchronous words and gestures. *Journal of experimental child psychology*, *129*, 110-126.

Goldin-Meadow, S. (2009). How gesture promotes learning throughout childhood. *Child development perspectives*, *3*, 106-111.

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet*. Mahwah, NJ: Erlbaum Associates.

Hart, B. & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore, MD: Paul H. Brooks.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55-88.

Iverson, J. M., Capirci, O., Volterra, V., & Goldin-Meadow, S. (2008). Learning to talk in a gesture-rich world: Early communication in Italian vs. American children. *First Language*, *28*, 164-181.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, *16*, 367-371.

Keller, H. (2012). Autonomy and relatedness revisited: Cultural manifestations of universal human needs. *Child Development Perspectives*.

LeVine, R. A., Dixon, S., LeVine, S., Richman, A., Keefer, C. H., Leiderman, P. H., & Brazelton, T. B. (1994). *Child care and culture: Lessons from Africa.* Cambridge, UK: Cambridge University Press

Lieven, E. V. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway & B. Richards (Eds.), (1994). *Input and interaction in language acquisition*. New York, US: Cambridge University Press.

Mastin, J. D., & Vogt, P. (2016). Infant engagement and early vocabulary development: A naturalistic observation study of Mozambican infants from 1;1 to 2;1. *Journal of child language*, *43*, 235-264.

Rowe, M. L., & Goldin-Meadow, S. (2009a). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*, 951-953.

Rowe, M. L., & Goldin-Meadow, S. (2009b). Early gesture selectively predicts later language learning. *Developmental science*, *12*, 182-187.

Rowe, M., Özçalıskan, S., & Goldin-Meadow, S. (2006). The added value of gesture in predicting vocabulary growth. *Proceedings of the 30th annual Boston University conference on language development* (pp. 501-512).

Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First language*, *28*,182-199.

Salomo, D., & Liszkowski, U. (2013). Sociocultural settings influence the emergence of prelinguistic deictic gestures. *Child development*, *84*, 1296-1307.

Tamis‐LeMonda, C. S., Song, L., Leavell, A. S., Kahana‐Kalman, R., & Yoshikawa, H. (2012). Ethnic differences in mother–infant language and gestural communications are associated with specific skills in infants. *Developmental Science*, *15*, 384-397.

Vogt, P., & Mastin, J. D. (2013). Rural and urban differences in language socialization and early vocabulary development in Mozambique. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 3787-3792). Austin, TX: The Cognitive Science Society.

Vogt, P., Mastin, J. D., & Schots, D. M. (2015). Communicative intentions of child-directed speech in three different learning environments: Observations from the Netherlands, and rural and urban Mozambique. *First Language, 35*, 341-358.

Zukow-Goldring, P. (1996). Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. *Early development and parenting, 5*, 195-211.

A Neural Model of Context Dependent Decision Making in the Prefrontal Cortex

Sugandha Sharma (s72sharm@uwaterloo.ca)
Brent J. Komer (bjkomer@uwaterloo.ca)
Terrence C. Stewart (tcstewar@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Center for Theoretical Neuroscience, 200 Unversity Ave. W.
Waterloo, ON N2L3G1 Canada

Abstract

In this paper, we present a spiking neural model of context dependent decision making. Prefrontal cortex (PFC) plays a fundamental role in context dependent behaviour. We model the PFC at the level of single spiking neurons, to explore the underlying computations which determine its contextual responses. The model is built using the Neural Engineering Framework and performs input selection and integration as a nonlinear recurrent dynamical process. The results obtained from the model closely match behavioural and neural experimental data obtained from macaque monkeys that are trained to perform a context sensitive perceptual decision task. The close match suggests that the low-dimensional, nonlinear dynamical model we suggest captures central aspects of context dependent decision making in primates.

Keywords: context dependent decision making; decision making; neural engineering framework; neural dynamics; theoretical neuroscience

Introduction

Decision making is a complex process that involves the aggregation of many sensory signals in order to produce a response. What kinds of sensory information are relevant as well as what to do with that information can be highly dependent on the context of the situation. Prefrontal cortex (PFC) contributes to representing and maintaining relevant contextual information, ignoring irrelevant sensory information. Neurons in PFC have been found to be involved in selective attention and selection of an intended action (Tanji & Hoshi, 2008). Thus the relevant sensory evidence can be thought to be integrated in PFC before a decision is made in a contextual situation.

Early selection models have been used in the past to explain the underlying context dependent computations in PFC. These models suggest that the irrelevant inputs are filtered out before the integration stage in PFC thus leading to a larger effect of relevant as compared to the irrelevant sensory information on context dependent behaviour. However, in a recent experiment of context dependent behaviour in monkeys, no evidence of early selection was found (Mante et al., 2013). Instead it was found that the relevant sensory input is selected late by the same PFC circuitry which integrates sensory evidence to make a choice.

In this paper we present a spiking neural model of context dependent decision making based on the above findings. The model is built using the Neural Engineering Framework (NEF; Eliasmith & Anderson (2003)) and is meant to provide insight into the biological mechanisms behind context dependent computations in PFC.

Related Work

To explore context dependent decision making in prefrontal cortex, Mante et al. (2013) designed an experiment in which macaque monkeys were trained to perform two different perceptual discriminations on the same set of visual stimuli based on the given context. On the basis of the results of the experiment, they built a learned recurrent network model to characterize the mechanisms underlying context dependent behaviour.

Experiment

Two monkeys (monkey A and monkey F) were trained on a perceptual discrimination task. On each trial of the task, the monkey was shown a random dot display consisting of dots of two different colours. A fraction of the dots moved coherently in one of the two directions (left or right), while the remaining dots were transiently flashed at random locations. The monkey was provided a contextual cue to instruct it to either discriminate the direction of motion (i.e., the coherent direction of motion of the dots) or the colour (i.e., which of the two colours are on the largest number of dots). The monkey reported its choice with a saccade to one of the two visual targets provided. After each trial the monkey was rewarded for correct responses with a small quantity of juice.

To vary the difficulty of the colour discrimination, the relative number of dots of the two colours was parametrically changed, while keeping the total number of dots constant. The fraction of colour 1 to colour 2 dots i.e., colour coherence (100% coherence: only one colour, 0%coherence: equal number of dots of two colours) was fixed throughout a trial. To vary the difficulty of the motion discrimination, the fraction of dots moving coherently i.e., motion coherence (100% coherence: all dots moving in the same direction, 0% coherence: all dots moving randomly) was parametrically changed. A total of 163,187 trials were performed and the trials varied with respect to the colour coherence and motion coherence of the random dot display, and the contextual cue provided.

Extracellular responses were recorded from the neurons in the frontal eye field (an area of PFC involved in the selection and execution of saccadic eye movements) while the monkeys performed this task. The responses of the population of neurons as a whole were analyzed to understand the neural mechanism underlying context-dependent selective integration of sensory stimuli.

A learned recurrent network model

To identify the mechanism underlying context-dependent behaviour, Mante et al. (2013) trained a network of recurrently connected neurons to solve a task analogous to the one solved by monkeys. Their model reproduced the main features of the data, and they analyzed the trained network to determine the selection mechanism. The results of their analysis indicate that the rich dynamics of PFC responses during selection and integration of inputs can be understood with just two features - a line attractor and a selection vector, which are defined only at the level of the neural population. It's important to note that their results indicate that computations in PFC emerge from collective dynamics of large populations of neurons. Further details about this model can be found in Mante et al. (2013) paper.

However, their model results do not fit the reported data very closely, significantly deviating from the response dynamics reported for the monkey experiments. As well, because the model was learned from data, it is unclear precisely what the nature of the underlying computation is. The identification of two degrees of freedom (i.e., memory and selection) is helpful, but is evident from performing dimensionality reduction on the original data, without a model. Here we propose a specific, low-dimensional, nonlinear computation and demonstrate a superior fit to the original data. Additionally, we show how such computations can be implemented in biologically realistic spiking neurons.

A nonlinear dynamical model

The basic principle of the model we describe is that four different pieces of information must be represented by the same group of neurons. The empirical data indicates that the neurons in the frontal eye field change their activity based on the context (whether the response should be based on colour or motion), colour coherence (the proportion of dots of one colour over another), motion coherence (the proportion of dots moving in one direction over another), and the actual choice being made. We denote this combination of four values as the vector x.

Rather than this information being held separately by different groups of neurons, it is distributed across a single population. This means that each neuron in the population is sensitive to some combination of those four values. Some neurons may be sensitive to colour and motion, while others may respond primarily due to the final choice, and others may be equally sensitive to all four values. In general, we assume that each neuron has some particular combination of values for which it will fire most strongly, consistent with the standard idea of a preferred stimulus or a preferred direction vector (Schwartz et al., 1988). If this vector is e (for "encoder"), then the amount of current flowing into a particular neuron i will be $\alpha_i e_i \cdot x + J_i$, where α and J_i are a gain and bias respectively. This current is used as the input to the standard leaky integrate-and-fire (LIF) neuron model to produce spiking output (it is these spikes which flow through the connections and not the continuous signals). The encoder, gain, and bias are randomly chosen, so different neurons have different background firing rates and different maximum firing rates (set here to be uniformly distributed between 20Hz and 120Hz) along their preferred direction.

The purpose of this model is to combine the three inputs (context, colour coherence, and motion coherence) to produce an appropriate fourth value (the appropriate choice). To provide the input, we directly stimulate the neurons based on the sensory inputs of the given trial (we are not simulating the visual system). From these inputs, the neural population must compute the resulting choice of action. To capture this computation, the neurons are recurrently connected to compute a nonlinear dynamical system that appropriately integrates the inputs, and stores the resulting choice.

We achieve this by writing a mathematical expression for how the choice value should be dynamically computed given the sensory input, and then solving for the neural connectivity that optimally approximates that mathematical expression. This mathematical expression is our hypothesis about how information is integrated in PFC to produce a choice. Specifically, writing the dynamics of the population as representing the 4D state variable, x (where x_4 is the final choice):

$$\dot{x_4} = f(x) \tag{1}$$

we must determine the function f. In a simple PFC memory model, $f(x) = 0$. However, here there are clearly context interactions which must be accounted for. We propose that the nonlinearity in the recurrent connection is given by

$$f(x) = (1 + x_1)x_2 + (1 - x_1)x_3 \tag{2}$$

where x_1 is the task context (1 if the decision should be based on motion, and -1 if it should be based on colour), x_2 is the motion stimulus (between -1 for completely leftward motion, and +1 for completely rightward motion), and x_3 is the colour stimulus (between -1 for all one colour, and +1 for all the other colour). In short, this nonlinearity essentially gates either colour or motion information into a working memory integrator, depending on the context signal. Consequently, the model is a simple hypothesis of multiplicative gating in PFC. Importantly, this multiplicative effect is achieved at the network level using normal synapses between LIF neurons. Mathematically, it basically forms a line attractor, however, this needs to be explored further in future work.

To simulate the model, we used the software package Nengo, a Python-based simulation engine and environment for constructing neural networks that includes facilities for generating NEF-style networks (Bekolay et al., 2014).

A schematic of the model is shown in Figure 1. The central group of neurons consists of 1000 randomly generated LIF neurons with preferred direction vectors e such that each neuron responds differentially to the four x values being represented. These neurons are connected recurrently back to themselves in an all-to-all manner (for a total of 1 million synapses), with the synaptic connection strengths optimized

via the NEF to closely approximate equation 2. We also included a separate group of neurons that just represent the final choice (x_4). The synaptic connection weights from the "pfc" neurons to the "choice" neurons are also found using the NEF, where the target function was simple $choice = x_4$, providing a simple readout of the model's choice.

For each trial, the inputs were set as follows:

- Contextual cue (x_1) - This indicates the context for each trial and takes a value of 1 and -1 for motion and colour contexts respectively. Recall, motion context implies discriminating the direction of motion of the dots (i.e., make a decision based on motion coherence) whereas the colour context implies discriminating the colour of the dots (i.e., make a decision based on colour coherence).
- Colour coherence (x_2) - For each trial one of the six signed colour coherence values: [-0.50, -0.18, -0.06, 0.06, 0.18, 0.50] is randomly provided to the model as an input. The sign of the values indicates different colours i.e., a negative sign corresponds to colour 1 (red) while a positive sign corresponds to colour 2 (green).
- Motion coherence (x_3) - For each trial one of the six signed motion coherence values: [-0.50, -0.15, -0.05, 0.05, 0.15, 0.50] was randomly provided to the model as an input. The sign of the values indicates the direction of motion i.e., a negative sign indicates motion towards left while a positive sign indicates motion towards right.

Different values of motion and colour coherence mentioned above were used to vary the strength of the motion and colour signals between trials. The output of the model is a choice (i.e., the decision made for a particular trial given the context). Each trial in the model is 750ms long (i.e., inputs are provided to the model for 750ms per trial) followed by a delay period of 750ms between consecutive trials. This delay period was added to avoid the influence of the previous trial on the current trial, and can be thought of as analogous to the period when the monkey was being rewarded between trials.

In the experiment, the monkeys were able to quickly switch between behavioural contexts. This implies the presence of rapid modulation or 'gating' mechanisms in the PFC that select relevant sensory information for decision making. In our model, multiplicative gating is used to control the dynamics of "pfc", such that only the relevant evidence in a given context gets integrated. For example in motion context ($contextual_cue = 1$), only the motion coherence values get integrated over time and the colour coherence values are irrelevant to the final decision. This is analogous to the actual experiment because once the monkey knows the context, it pays selective attention to only the relevant evidence and the irrelevant evidence mostly gets ignored. When the integrated relevant evidence passes beyond a threshold, the model is considered to have made a choice.

Note that instead of early selection (i.e., selection of relevant sensory input before the integration stage in PFC), input selection and integration happens within the same PFC circuitry as a part of a single dynamical process.

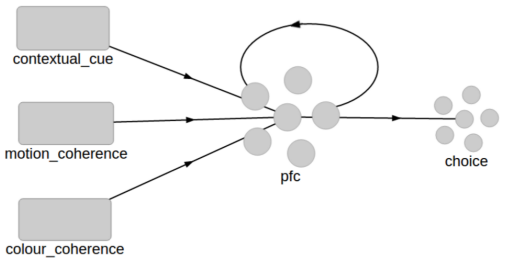


Figure 1: A schematic diagram of the model structure built using Nengo GUI (a graphical user interface to aid in the design and testing of Nengo networks). Here "pfc" contains 1000 neurons and "choice" contains 200 neurons.

Results

The model runs through a series of trials consecutively, and the choice the model makes for each trial is recorded, along with the particular condition variables for that trial (context, motion coherence, colour coherence, and correct choice). Figure 2 shows the spike trains and decoded values from one such trial. The spike trains are shown only for 10 random neurons picked from a population of 1000 neurons in "pfc", and show that different neurons fire at different times according to their preferred stimulus. The spike trains from each individual neuron are recorded for each trial. This allows the model output to be compared to the monkey data on both the behavioural and neural level. Note that we compare our model to the performance of monkey A since monkey A data has primarily been discussed by Mante et al. (2013). Additionally we fix our model parameters and use the data from the same model for behavioural and neural comparisons.

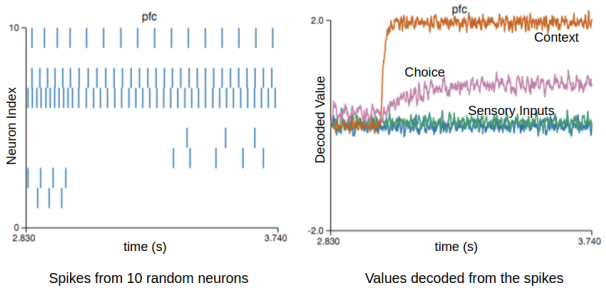


Figure 2: Spike trains and the decoded values from "pfc". The decoded values show the context, choice and the sensory inputs (i.e., colour and motion coherence).

Behavioural comparison

A comparison of the model output to the behavioural data of the monkey is shown in Figure 3. The figure plots the performance of the model as a function of motion and colour coherence in a given behavioural context. The performance is measured in terms of % of choices made to the right, which

implies choosing 'right' direction in motion context and % of choices made to green which implies choosing the 'green' colour (i.e., colour 2) in colour context. The plot on the top left shows that in the motion context, the model makes very few choices towards right for negative values of the motion coherence (recall that negative values simulate dots moving to the left). However the % choices made to the right increases with an increase in the motion coherence, and is highest when the motion coherence takes its maximum positive value (recall that positive values simulate dots moving to the right). The plot on the top right shows that the colour coherence does not have much effect on the choices in motion context. Overall, the two plots show that in the motion context, the choices of the model depend strongly on motion coherence of the dots whereas they depend only weakly on colour coherence. The opposite pattern is observed in the colour context (bottom two plots) in which the choices of the model are largely affected by colour coherence and weakly by motion coherence.

These results show that the model successfully differentiates the relevant sensory evidence in each context and ignores the irrelevant evidence. The result also matches well with the monkey data obtained experimentally as shown in Figure 2. The overall % average error of our model with respect to the experimental data is 7.33%, while that for the learned recurrent network model of Mante et al. (2013) is 7.45%. Importantly, in their model the response is completely independent of the irrelevant input (e.g., the %choices to the right is constant with respect to the motion coherence in colour context). However, in our model, the irrelevant input has a weak effect on the response which is consistent with the experimental data. Our model captures this subtle feature since it is built using biologically realistic LIF neurons. Moreover, our model has only one parameter which is a constant transform (0.45) applied to scale the motion and colour coherence inputs (this parameter was not tuned in the results we present). The synaptic time constant (200ms) was fixed based on biological data (Flint et al., 1997) and the neural firing rates were set to be uniformly distributed between 20Hz to 120Hz. The connection strengths were not fit to the experimental data, but to the equation that we are trying to approximate i.e., Equation 2, using the NEF (Eliasmith & Anderson, 2003). On the other hand, the learned recurrent network model has many parameters which are learned through training.

Analysis of neural spike data

In order to facilitate a quantitative comparison to the monkey data, identical analysis methods as described in the supplementary information from Mante et al. (2013) are used on our spike data. All the analysis is performed on the spike data obtained from the central population of neurons i.e., "pfc" in Figure 1.

First, we determine the condition average population response by grouping the trials based on the unique combination of condition variables. There are five different condition variables: context (motion or colour), motion coherence (range of six values), colour coherence (range of six values), choice (two options), and correctness (whether or not the monkey made the correct choice for the given condition). This amounts to a total of 288 possible unique conditions ($2 \times 6 \times 6 \times 2 \times 2$).

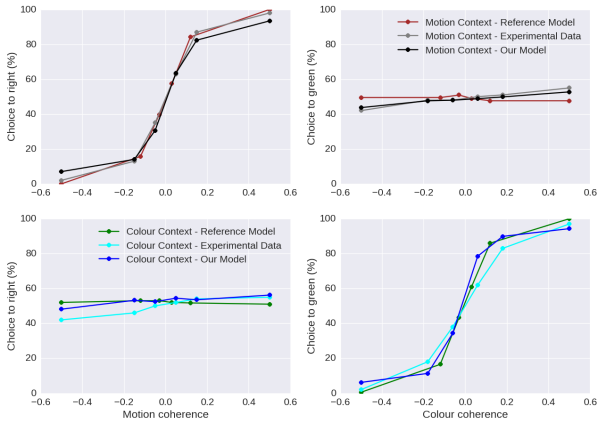


Figure 3: Behavioural output of the model. Results are averaged across 14,688 trials. Reference model and experimental data used for the comparison are taken from Mante et al. (2013) and were averaged across 14,400 and 163,187 trials respectively.

The spike response of each neuron is averaged across all trials having the same unique conditions, and then smoothed by a gaussian filter with $\sigma = 40ms$. The result is then z-scored for each neuron to give the overall response across all times a zero mean and unit standard deviation. This results in a two dimensional data matrix X of size $N_{neuron} \times (T.N_{condition})$ containing the condition averaged population response. Here N_{neuron} is the total number of neurons in the neural population, $N_{condition}$ is the total number of unique conditions and T is the number of time samples. Note that for our model $N_{neuron} = 1000$, $N_{condition} = 288$ and $T = 750$.

Next, we perform principal component analysis (PCA) on the condition averaged response (X) to identify the dimensions in state space that capture the most variance. The PCA is done along the number of neurons and the principle components (PCs) are vectors of length N_{neuros}. We use the first twelve PCs to define a de-noising matrix D of size $N_{neuron} \times N_{neuron}$ using equation 3.

$$D = \sum_{a=1}^{12} v_a v_a^T \tag{3}$$

Next, we perform linear regression to determine the effect of various task variables (i.e., choice, motion, colour and context) on the response of each neuron. In order to perform linear regression, the response of neuron i at time t is represented as a linear combination of task variables as shown in equation 4, where $r_{i,t}(k)$ is the z-scored spike response of neuron i at time t on trial k. The values of the four task variables also correspond to trial k. The regression coefficients $\beta_{i,t}(v)$ for $v = 1$ to 4 describe how the corresponding task variable affects the

firing rate of neuron i at a given time t during the trial.

$$r_{i,t}(k) = \beta_{i,t}(1)choice(k) + \beta_{i,t}(2)motion(k) + \beta_{i,t}(3)colour(k) + \beta_{i,t}(4)context(k) + \beta_{i,t}(5) \quad (4)$$

In order to estimate the regression coefficients, a set of condition matrices is defined, one for each neuron. Each matrix F_i is of the size $N_{coeff} \times N_{trial}$, meaning there is a row for each of the five regression coefficients and a column for each trial. The first four rows of F_i contain the task variables for each trial, while the last row contains only ones, and is used solely to estimate $\beta_{i,t}(5)$. The regression coefficients are then estimated using equation 5.

$$\beta_{i,t} = (F_i F_i^T)^{-1} F_i r_{i,t} \quad (5)$$

The matrix $\beta_{i,t}$ of length N_{coeff} is then rearranged to form $\beta_{v,t}$ of length N_{neuron} which is de-noised by projecting it to the subspace spanned by the first 12 PCs using equation 6.

$$\beta_{v,t}^{pca} = D\beta_{v,t} \quad (6)$$

$\beta_{v,t}^{pca}$ is of length N_{neuron} and contains the de-noised regression coefficients. Next step is to define the time-independent de-noised regression vectors. To do this we determine the time at which the corresponding set of vectors $\beta_{v,t}^{pca}$ has the maximum norm, and select the value of the regression coefficients at that time (equation 7).

$$\beta_v^{max} = \beta_{v,t^{max}}^{pca} \quad (7)$$

β_v^{max} is of length N_{neuron} and is orthogonalized using QR decomposition as shown in equation 8.

$$B^{max} = [\beta_1^{max}\beta_2^{max}\beta_3^{max}\beta_4^{max}] = QR \quad (8)$$

The first four columns of Q become a set of orthogonalized regression vectors $\beta_v^{\perp}$ which define the task-related axes of choice, motion, colour, and context. The condition averaged population response (X) is then reorganized to form a set X_c of $N_{condtion}$ matrices of size $N_{neuron} \times T$. Then we project X_c onto the task related axes (Equation 9) to allow for the visualization of the task related representations in "pfc".

$$p_{v,c} = (\beta_v^{\perp})^T X_c \quad (9)$$

Neural comparison

Figure 4 shows the plots of the neural output of the model along the various task-related axes in two different contexts. The response trajectories start from a point in state space close to the center of the plots, which corresponds to the population response at the beginning. When the sensory input is provided, the response moves away from the center along the axis of choice. Movement of the population response along the choice axis is caused by integration of evidence while the sensory input is being provided to the model. Additionally, the population response moves away from the choice axis, along the axis of motion/colour. The magnitude of the response along the motion/colour axes reflects the strength of the sensory evidence (motion/colour coherence), while its direction (up or down) reflects the sign of the sensory evidence.

Mante et al. (2013) obtained the neural data from two monkeys (monkey A and monkey F) trained to perform the same task. The trajectory for monkey A is shown in Figure 4 while the trajectory for monkey F can be found in the extended data of Mante et al. (2013) paper (Extended Data Figure 7). The response trajectories obtained for both monkeys varied significantly from each other. However the following three key features of population response were shared between the two monkeys:

- In both contexts, movement along the same choice axis corresponds to the integration of relevant evidence (Figure 4, plots a2 and f2).
- The motion evidence evokes response along the axis of motion which is different from the axis of choice in both contexts (Figure 4, plots a2 and d2).
- Motion evidence is strongly represented regardless of whether it is relevant (Figure 4, plot a2) or irrelevant (Figure 4, plot d2).

The response trajectories obtained from our model are very similar to the ones obtained from the experimental data. Importantly, they exhibit the three key features of the population response that were shared among the experimental data from the two monkeys. Plots a1 and f1 in Figure 4 show that the model trajectories show movement along the choice axis which corresponds to the integration of relevant evidence (i.e., motion in plot a1 and colour in plot f1). The movement along the choice axis in plots c1 and d1 is not as much because they plot the irrelevant sensory evidence in their respective contexts. Plots a1 (motion context) and d1 (colour context), both show that the motion evidence evokes response along the axis of motion which is different from the axis of choice. Moreover, these plots also show that motion evidence is strongly represented in both contexts regardless of whether it is relevant or not.

Figure 5 shows a comparison of our model to the learned recurrent network model of Mante et al. (2013). Note that the scaling on the choice axis in the figure is 1.5:1 for our model and the monkey data, and 6:1 for the reference model. The response trajectory obtained by our model is closer to the trajectory obtained from the experimental data. While our model is able to produce a curved trajectory as in the experimental data, the model of Mante et al. (2013) doesn't produce a curved trajectory. This might be because the dynamics of their model doesn't capture the fact that neural response in a given context is weakly dependent on irrelevant sensory input (Figure 3).

Conclusion and Future Direction

The main purpose of this work was to test a hypothesis (Equation 2) regarding the computations underlying context-

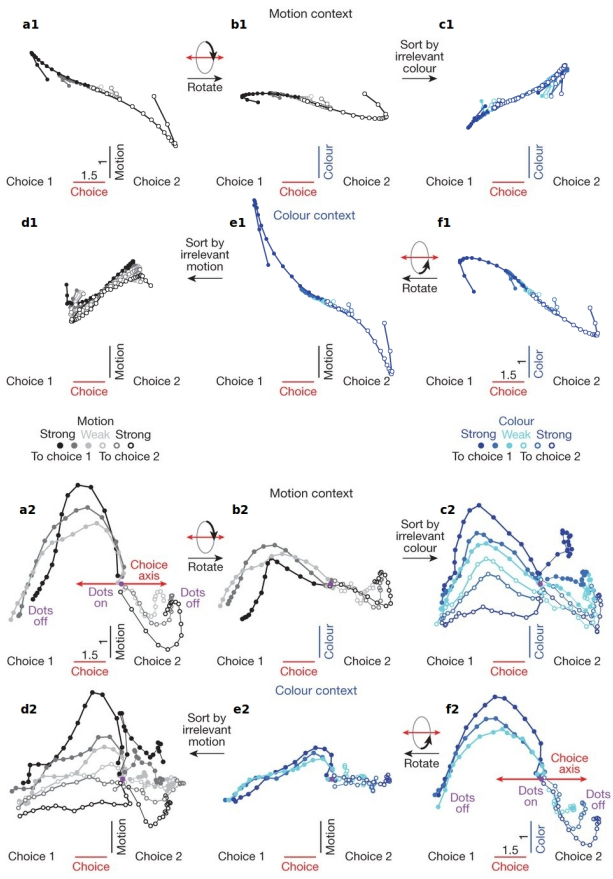


Figure 4: Neural Output of the Spiking Model (plots a1 - f1). Plots a2 - f2 show the neural response trajectories from the experimental data (monkey A). Only the responses from correct trials are plotted. Units are arbitrary. **a1**, **a2**: Effect of relevant motion input and choice in motion context where trials are sorted by the relevant sensory input (motion). **b1**, **b2**: Same data as in a1, a2 rotated by 90°about the choice axis to show the projection on to the colour axis. **c1**, **c2**: Same trials as in a1, a2 but resorted by irrelevant sensory input (colour). **d-f**: Responses plotted in the colour context, analogous to **a-c**. Display format taken from Mante et al. (2013).

dependent behaviour in the frontal eye fields of the prefrontal cortex (PFC). We have presented a spiking neural model to explain how neural and behavioural data can be captured by a simple nonlinear dynamical system. One limitation of the current work is that we directly stimulate the neurons based on the sensory inputs of the task without simulating the visual system required for the task. However, this should not have a major impact on the underlying computations in PFC.

In this model we simulated a task where the sensory information is constant within each trial and the decision is made only at the end of the trial. However, animals are able to change their behaviour in a constantly changing environment. The future work would involve investigation of neural mechanisms of such dynamic choices where the sensory information is constantly changing (even within a given trial) and a

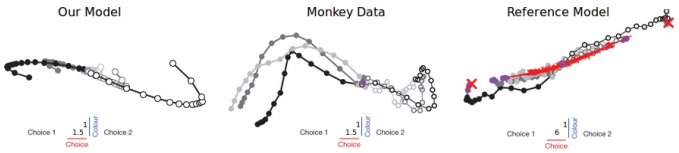


Figure 5: Comparison of our model and the reference model to the experimentally recorded data for the motion context. Data is sorted by motion coherence and plotted on the choice and colour axes. Scaling on the choice axis is 1.5:1 for our model and the monkey data, and 6:1 for the reference model.

decision can be made at any time. One such experiment has been previously performed with monkeys where they were trained to solve the tokens task (Thura & Cisek, 2014). Building a scalable and robust model which can not only perform the task discussed in this paper, but also a more dynamic task like the tokens task can provide insights to more precisely explain the computational mechanisms underlying context dependent decision making in dynamic environments.

Acknowledgments

This work was supported by CFI and OIT infrastructure funding, the Canada Research Chairs program, NSERC Discovery grant 261453, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084 and OGS graduate funding.

References

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., ... Eliasmith, C. (2014). Nengo: A python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*(48). doi: 10.3389/fninf.2013.00048

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Flint, A. C., Maisch, U. S., Weishaupt, J. H., Kriegstein, A. R., & Monyer, H. (1997). Nr2a subunit expression shortens nmda receptor synaptic currents in developing neocortex. *The Journal of Neuroscience*, *17*(7), 2469–2476.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84.

Schwartz, A. B., Kettner, R. E., & Georgopoulos, A. P. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement. *The Journal of Neuroscience*, *8*(8), 2913–2927.

Tanji, J., & Hoshi, E. (2008). Role of the lateral prefrontal cortex in executive behavioral control. *Physiological reviews*, *88*(1), 37–57.

Thura, D., & Cisek, P. (2014). Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. *Neuron*, *81*(6), 1401–1416.

Discourse Analysis as a Solution to Interpretive Problems in Cognitive Development Research

Patrick Byers (Patrick.d.byers@gmail.com)
Department of Psychology, 365 Fifth Ave.
New York, NY 10016 USA

Abstract

Cognitive development researchers have drawn conclusions about young children's developing knowledge of number by studying their behavior, while at the same time acknowledging that behavior is an imperfect index of knowledge, e.g., it may be disputed whether a given behavioral task accurately measures, overestimates, or underestimates children's knowledge. The texts of published research articles from these investigations are the focus of a discourse analysis described in the present article. The results of the discourse analysis suggest that claims about what a person knows are actually generalized descriptions of behavior. Therefore, in studying behavior on tasks to draw conclusions about participants' conceptual knowledge, researchers are merely making behavioral generalizations, not investigating hidden cognitive or epistemic content.

Keywords: conceptual knowledge; discourse; epistemology; performance and competence; conceptual and procedural knowledge

Introduction

Cognitive development researchers have drawn conclusions about young children's developing knowledge of number by studying their behavior, while at the same time acknowledging that behavior is an imperfect index of knowledge. For example, researchers may dispute whether a given behavioral task accurately measures, overestimates, or underestimates children's knowledge, all while tacitly accepting that behavior is what provides proof of knowledge. This ambivalent relationship between knowledge and behavior—knowledge is neither perfectly reflected in behavior, but behavior is the only clue we have as to what children know—has made the distinction between them central to investigations of children's knowledge of number. In fact, many of these investigations are as much concerned with determining valid behavioral indices of that knowledge as with the knowledge itself (Briars and Siegler, 1984; Gelman and Meck, 1983; Greeno, Riley and Gelman, 1984; Rittle-Johnson, Schneider and Star, 2015; Sarnecka and Carey, 2008; Wynn, 1990).

The current paper argues that the distinction between knowledge and behavior as pursued in this research is illusory, and that descriptions of what children know are, in fact, descriptions of behavioral capacities rather than descriptions of some hidden, epistemic content. This conclusion is supported by the results of a qualitative discourse analysis of the ways that claims about what children know are justified, negotiated and contested in research articles on early number concepts.

Knowledge and Behavior in Research on Cardinality

Research on developing number concepts in children has focused extensively on their understanding of the cardinal meanings of number words and counting, or cardinality. Cardinality refers to the way that number words denote entire sets of items, rather than merely the individual members of those sets. Part of the reason for this focus on cardinality is that children are inconsistent in performing the behaviors indicative of understanding cardinality. Children who are otherwise able to use numbers in various ways (e.g., in counting) often fail to demonstrate an understanding of cardinality in certain contexts. For example, Fuson (1988) describes preschool children who, after accurately counting a set of five items, and responding to the question "how many are there" with the last counted word (a behavior ostensibly indicative of understanding cardinality), respond to the request for *the five items* by picking up only the fifth counted item. Other studies report similar findings. For example, when asked to create a set of a requested number of items from a larger pile, preschool-aged children who were otherwise proficient counters responded by grabbing a random (numerically unrelated) number of items (Fuson, 1988**;** Sarnecka & Carey, 2007; Wynn, 1990).

Researchers interpret these types of inconsistent displays of competence by recourse to two acknowledgements about the relation between knowledge and behavior: (1) First, it may be claimed that competence precedes performance, i.e., young children may understand cardinality but have difficulty planning and/or executing actions that reflect this understanding (e.g., Gelman and Gallistel, 1978). (2) Alternately, it is claimed that performance may precede competence, i.e., young children's apparent competence in some contexts may reflect their carrying out rote-learned or imitated behavioral routines that they do not yet fully understand on a conceptual level (e.g., Wynn, 1990).

These interpretations allow researchers to provide stable accounts of children's developing conceptual knowledge of number, despite inconsistent and unpredictable behavioral performances. Yet, as has been explained in more detail elsewhere (Allen & Bickhard, 2013; Byers, 2016), the flexibility that these explanations provide is inherently problematic. If we accept that (1) knowledge can be present despite the absence of corresponding behavior, and (2)

behavioral performances ostensibly indicative of knowledge may be performed in the absence of that knowledge, this means that essentially any behavioral evidence can be consistent with any underlying conceptual knowledge. This is especially problematic because, in the previously cited research, behavior is the only available predictor of knowledge. It is not clear how the additional study of children's behavior could produce more definitive conclusions about children's knowledge of cardinality.

The current paper proposes that the fundamental issue here is not one that can be resolved by further study of children's behavior. Instead, the crucial issue is that the distinction between behavior and knowledge—and specifically the attempt to draw conclusions about knowledge on the basis of behavioral evidence, while simultaneously treating behavior as an ambiguous indicator of that knowledge—undermines the discursive practices through which conclusions about knowledge are normatively reached. These arguments are elaborated in the following sections, which describe a discourse analysis of the texts of research articles that report conclusions about children's knowledge of cardinality.

Discourse Analysis of Knowledge Claims

The discourse analysis focused on the ways that conclusions about what children know are asserted, justified and contested in research articles on children's knowledge of number. The aim of this analysis is to develop a general model of how this occurs that can shed light on what it means to claim that the child *knows* or *understands* something about cardinality or number, and why certain types of evidence provide grounds for asserting or contesting these claims. These issues are important since, insofar as problems in reaching conclusions about children's knowledge of number stem from (or can be understood in terms of) the discursive practices themselves, the most relevant aspect of these practices are the conclusions about knowledge, and their justification.

Sampling of Texts

The discourse analysis was conducted on a sample of ten published research articles,[1] each of which was focused on children's conceptual knowledge of number. In each of these articles, claims are made about what children do/don't know, which are justified with evidence and other arguments. The decision to focus only on articles on numerical knowledge was intended to reduce the scope of analysis for purposes of simplicity. While some of the articles have the same authors, and are therefore not discursively independent of each other, the intention of this study was not to construct a representative sample of all published research in this area, but to understand, in depth, the way that knowledge claims may be asserted/contested/justified in a specific discursive context, with the hope that this may yield findings applicable to other discursive contexts. While it is possible that the current results will not be generalizable due to idiosyncratic features of the sample, this kind of high density sampling allows for a comprehensive analysis that would be unwieldy with a broader sample (in a similar way, a historian might claim that an in-depth study of a particular war (from a few extensive accounts) might yield in-depth knowledge about war that would not be possible from a more brief study of many accounts). Consequently, while the current results must be taken as only potentially generalizable, the hope is that they may illustrate general insights about how conclusions about knowledge are reached.

[1] The selected articles include Briars and Siegler (1984), Brooks, Audet and Barner (2012) Davidson, Eng and Barner (2012), Frye, Braisby, Lowe, Maroudas and Nicholls (1989), Gelman (1972), Gelman and Meck (1983), Nikoloska (2009), Sarnecka and Carey (2007), Sarnecka and Wright (2013), Sophian (1988), and Wynn (1990).

Process of Analysis

As a starting point for the analysis, a preliminary distinction was made between two types of statements to be identified in each of the texts: *knowledge claims*, and the statements that justify them—*justifications* (both are described in detail below in Table 1). These initial categories were necessary as a starting point for the analysis, which after all, cannot come from nowhere. In the remainder of this article, examples of knowledge claims will be italicized and justifications will be underlined.

Knowledge Claims Knowledge claims are assertions of what a person knows. They are not just any claims about knowledge, but rather claims that describe what is known. Knowledge claims most commonly involve the verbs *to know* or *to understand*, but may also use others such as *realize*, *recognize*, or phrases such as the claim that a person *has a concept of something*. A basic example of a knowledge claim is the following: *s/he understands the cardinal meaning of the number six*. Knowledge claims may also be assertions that a person does *not* know something, e.g., *s.he does not understand the cardinal meaning of the number six*.

Justifications Justifications are the most directly given reason used to support or contest a given knowledge claim. Continuing from the example above, the justification would be the reason or proof given for the assertion of the previous statement. For example if it is claimed that *s/he understands the cardinal meaning of the number six* because s/he correctly gives out six items from a larger pile when requested, the latter underlined portion would be a justification. As this description implies, justifications can only be identified relative to a knowledge claim, rather than in terms of the intrinsic characteristics of the justifying statement itself.

Table 1: Composite statements containing *knowledge claims* (italicized) and justifications (underlined).

S/he doesn't understand how counting works because when asked for "four toys" from a larger pile, she responded by grabbing a random number of toys.
The concept of cardinality is measured by *performance on* the Give-N task.
After an experimenter removed an item from a previously counted set, the child correctly responded when asked how many items it now contained, suggesting that *s/he knows the subtraction of an item from a counted set means that the number of items in the set decreases by one*.

Identification of Knowledge Claims and Justifications in the Analyzed Texts

While Table 1 shows knowledge claims and justifications as relatively discrete features in the text, it was recognized that instances of each may only be interpretable as a distributed network of connected statements across a text (e.g. descriptions of a task performance elaborated in multiple separate sections). In addition, the analysis focused on explicitly-, implicitly- and conditionally-asserted knowledge claims/justifications. Finally, care was taken to ensure that knowledge claims/justifications were identified in cases where, due to the flexibility of language, they may not have been apparent as such. For example, although the phrase *children's invariance schemes contain rules for reversing operations* (Gelman, 1972, p. 84) is not immediately apparent as a knowledge claim, it was classified as such because, in context, it clearly meant children *know that addition and subtraction (but not displacement) change the numerosity of a set and they also know which operation is needed to reverse the effects of addition/subtraction to return a set to its original numerosity*.

Refinement of Analytical Categories

The analysis proceeded by first reading each text and identifying paired exemplars of *knowledge claims* and *justifications* therein. These paired exemplars were then juxtaposed, revealing common characteristics of each type of statement, as well as clues as to how and why each member of a pair relates to the other. These features were analyzed with the intention of addressing the following questions: What is it about a knowledge claim and its accompanying justification that allows the latter to provide valid grounds for the assertion of the former? In cases where a justification is brought in to contest an asserted claim, what is it about the justification/claim that makes it possible for this to be done? More generally, under what circumstances can previously-justified knowledge claims be contested, and how is this done? Tentative answers to these questions were scrutinized by assessing the extent to which they were consistent with the broader use of knowledge claims throughout the sampled texts.

Results

The analysis of the selected texts in terms of the above categories (knowledge claim and justification) produced a variety of results. A full presentation of all of these is beyond the scope of the present paper. For the current purposes, the following two results are the basis for the most significant findings of the study.

Similarity of Knowledge Claims and Justifications

The distinction between knowledge claims and justifications was readily apparent in the analyzed texts. However, this distinction appeared to be a relative one, rather than something that could be defined in terms of the intrinsic characteristics of statements exemplifying either category. While this was assumed from the beginning to be true for justifications, the discourse analysis showed it to be true for knowledge claims as well (i.e., a statement is only a knowledge claim relative to the surrounding text).

While certain types of statements were found to be more typical of either knowledge claims or justifications (statements using verbs like *know* or *understand* versus descriptions of behavioral performances), these characteristics did not reliably characterize statements of either type because there was no type of statement that was exclusively found in only one category or the other. So, while assertions that a person *knows X* or *has a concept of X* were more typical of knowledge claims, statements like these were occasionally used as justifications. This is evident in the following quote from Frye et al. (1989), in which a description of what a child knows (first sentence) is presented as the possible justification for two alternative knowledge claims:

> …[the child's knowledge] that the cardinal value reached on the second trick trial ought to be the same as on the first correct trial. That knowledge could be evidence of [an understanding of] cardinality, or of a less specific understanding that if nothing about the array changes, the puppet's response shouldn't change either. (p. 1168)

Similarly, although descriptions of a person's behavior were more typically associated with justifications, these sorts of statements were occasionally used in essentially the same way as knowledge claims, insofar as their assertion was justified by other statements. (Despite not involving words like *know*, *understand*, etc., for reasons that will become clear, these statements can still be considered to be knowledge claims.)

There were also cases in which a statement appeared to simultaneously function as a knowledge claim and justification. This is evident in the following quote from Gelman (1972)

> Together, the reactions to Phase III indicate the children treated subtraction as an operation that was relevant and displacement as one that was irrelevant to number. This

indicates they had the ability to treat number (at least small ones) as invariant. ...

Reactions to the surreptitious changes introduced in Phase III indicate that Ss treated addition as relevant and displacement as irrelevant to number. ...The means of 1.5 and 0.3 reflect the fact that 15 addition and only four displacement Ss were surprised by the changes. All addition Ss noted the change as opposed to only six displacement Ss. (Gelman, 1972, p. 8)

In the quote, the phrase *children treated subtraction as an operation that was relevant and displacement as one that was irrelevant* functions as a justification for the claim that *children had the ability to treat number as invariant*. It also functions as a knowledge claim that is justified by the final two sentences of the quote.

The absence of a clear distinction between the intrinsic characteristics of statements serving as knowledge claims as opposed to justifications becomes even more apparent if we juxtapose a variety of examples of knowledge claims and justifications. In the place of clear boundaries between the two categories, a continuum emerges. As a demonstration of this, consider the following list of quoted knowledge claims and justifications, culled from the Davidson, Eng and Barner (2012, p. 162-162). Within those two pages, the authors entertain whether children…

- …understand how counting represents cardinalities.
- …have the meanings for only a subset of the words in their count list.
- … understand that…when a set beyond the child's counting range has an item removed and replaced with a different item…it retains its cardinality and that the same number word applies.
- …are able to distinguish one and two from each other and from the rest of the numbers in their count list.
- …are more likely to respond correctly when asked if there are now six or seven objects.
- can recite higher numbers (e.g., 5 or 10) [when counting]

The list begins with statements classified as knowledge claims and ends with justifications. While statements on either extreme are clearly distinct, no clear demarcation point between them is apparent. These findings will be further addressed in the interpretation section below.

Researchers' Responses to Inconsistent Evidence

The second relevant finding involved the fact that, at various points in each of the analyzed texts, authors gave reports of inconsistent evidence for a particular knowledge claim. This inconsistency involved findings that the same child would succeed and fail on two tasks that were (ostensibly) measures of the same form of knowledge. One example is summarized by Brooks, Audet and Barner (2012):

> Previous studies report conflicting evidence, and find that children expect numerals to label precise quantities in some tasks but not in others" (**Sarnecka and Gelman, 2004**; Condry and Spelke, 2008). (p. 1066)

In the terminology of knowledge claims and justifications, the above quote can be rephrased as saying 'previous researchers have found the same children's behavior across multiple tasks to constitute justifications for and against the knowledge claim that children *expect numerals to label precise quantities.'* Brooks et al. (2012) resolve this inconsistency by arguing that "some of children's apparent successes are best explained not by domain-specific semantic understanding of number, but instead by language-general pragmatic abilities" (p. 1066). This argument, like others seen throughout the analyzed texts, resolves inconsistent findings by re-describing some of the knowledge claims made on the basis of observed behavior. In this case, a behavioral performance ostensibly indicative of conceptual knowledge of number is re-described as indicative of general pragmatic abilities. Such arguments appear to work by differentiating the conflicting evidence, thereby removing the apparent inconsistency.

Interpretation of Results

The above findings suggest the following interpretations of what knowledge claims and justifications are and how they function:

1. Knowledge claims and justifications are both descriptions of behavioral dispositions.
2. A justification is a more concrete description of some aspect of the behavioral disposition that is described more generally by the knowledge claim.

The ways that these conclusions were derived from the previous two findings is explained in the following sections.

Knowledge Claims and Justifications are Descriptions of Behavioral Dispositions

The idea that knowledge claims and justifications are both descriptions of behavioral dispositions explains the similarity between both types of statements, as well as they ways they were used. This interpretation also makes intuitive sense in the case of most justifications, insofar as these were very often descriptions of behavioral performances. It makes less intuitive sense in the case of knowledge claims, which do not obviously appear to be descriptions of dispositions. Nevertheless, the idea that knowledge claims (including descriptions of very abstract conceptual knowledge, e.g., *understanding the cardinal principle*) are descriptions of behavioral dispositions is strongly supported by a variety of findings. To illustrate this, consider a typical case in which a knowledge claim is justified by a description of some behavioral performance, e.g., *knowledge of the cardinal principle* is justified by the fact that a child was able to accurately create sets of a requested number of items. (Although not all justifications were self-evidently descriptions of behavior, this explanation will account for those other cases as well.) If knowledge claims are themselves descriptions of behavioral dispositions, this explains (1) how researchers determine the particular behavioral performance denoted by the justification as relevant to the knowledge claim, and (2) why

the performance of the behavior described by the justification would be grounds for (or at least consistent with) the assertion of the knowledge claim. For similar reasons, this explanation also accounts for why descriptions of behavioral performances were sometimes used in the same way as knowledge claims, as well as cases in which knowledge claims were justified by other claims about what children know: Since—according to the current interpretation—all knowledge claims and justifications are descriptions of behavioral dispositions, these exceptional cases are not fundamentally unique.

The fact that knowledge claims for conceptual knowledge (e.g., *knowing the cardinal principle*) don't often appear to be descriptions of behavioral dispositions may be a reflection of the generality of the described disposition. While any disposition implies some range of behavior, rather than only one specific behavioral performance, this range may be more broad or narrow in different cases. It is relatively narrow in the case of a disposition such as *knowing how to recite a particular poem*, and much more broad in the case of a disposition like *understanding that poem*. In the former case, there would be clear structural resemblances in the range of behavior instantiating the described disposition. In the latter case, the disposition would be instantiated by a broad variety of behavioral forms whose similarity to each other may only be very abstract.

Knowledge Claims and Justifications as Relative Levels of Abstraction

Descriptions of behavioral dispositions do not describe specific behaviors, but rather a potential for some variety of related forms of behavior. Dispositions can be described at varying levels of abstractness/concreteness, with more concretely described dispositions denoting only highly structurally similar forms of behavior, and more abstract dispositions denoting forms that may share only a very abstract similarity. For instance, the disposition of a person to *count in general* is instantiated by a wider range of behavior than is the disposition to *count a set of five marbles*.

Descriptions of behavioral dispositions (knowledge claims) at different levels of abstraction can be nested hierarchically if the more abstract claim subsumes the disposition described by the more concrete one. Within such a hierarchically nested set of dispositions, the truth of a more general claim can be used to deductively infer the truth of a more concrete claim subsumed under it (e.g., *if a child knows how to count, they should be able to count these marbles*). Likewise, the truth of more concrete claims can be used to inductively infer the truth of a more abstract one (e.g., *the fact that the child can count marbles suggests that they may be able to count in general*).

The possibility of nesting disposition descriptions, and the inferential possibilities that are possible across different levels of abstraction explains how certain statements justify knowledge claims, and how a given knowledge claim provides a basis for expectation of more concrete forms of behavior. When researchers study children's behavior to make conclusions about what they know, they are in fact assessing whether a general disposition can be inductively inferred to be true as a function of whether the specific behaviors (or sub-dispositions) that it entails are in fact observed. Failure to observe these forms of behavior provides grounds for the rejection of the knowledge claim.

The view being developed here also accounts for the problems and solutions that emerged in cases of inconsistent evidence (i.e., finding that a child can perform some but not all of the forms of behavior implicated by a given knowledge claim. If knowledge claims are descriptions of more general dispositions (relative to their justifications), then to assert them is to imply a variety of more concrete dispositions subsumed under the asserted claim. The finding of inconsistent performance of these subsumed behaviors is problematic vis a vis a given knowledge claim, since this permits neither assertion nor rejection of the more general claim. Consequently, it is understandable that, in these instances, researchers tended to re-describe one of the inconsistent performances, since this had the effect of removing it as an exemplar of an inconsistently supported, knowledge claim, and resulting in a now-consistent body of evidence for or against the assertion of the knowledge claim.

This can be shown with the following example. If a child passes all but one task measuring an *understanding of the cardinal principle*, the researcher may claim that the failed task (behavior X) was too difficult due to, e.g., memory demands. In doing this, the researcher is effectively re-describing task X as not only a measure of the cardinal principle, but additionally as a measure of memory ability. The fact that the task has now been claimed to also measure memory ability means that it is no longer a valid concrete instantiation of *understanding the cardinal principle*. Consequently, the child can now be claimed to *understand the cardinal principle* because this no longer implies passing task x, since the latter is now a justification for a claim along the lines of: *demonstrating knowledge of the cardinality principle in a context that involves excessive memory demands.*

Conclusion

The present research analyzed the texts of published research articles in order to explain how conclusions about children's knowledge of number are asserted, justified and contested, as well as how these statements function and what they indicate. The premise for this investigation was the observed difficulty that researchers have faced in drawing definitive conclusions about children's knowledge on the basis of their behavior. The results of the discourse analysis suggest that descriptions of children's knowledge are in fact generalized descriptions of their disposition for certain behavior, and that these knowledge claims are asserted contingent on the observation of behavioral

performances that concretely exemplify certain aspects of the behavioral disposition in question.

These conclusions readily account for why cognitive development researchers have been unable to disentangle children's conceptual knowledge from their performance capabilities. From the perspective of the current findings, such a differentiation is impossible. Conceptual knowledge in some conceptual domain is not some cognitive or epistemic entity that is reflected in—but separate from—certain forms of behavior. Rather, it is a discursive generalization of that behavior. Therefore, the perennial discussion of whether children's conceptual competence in some domain chronologically precedes or follows their behavioral ability in that domain is ultimately incoherent.

The centrality of the competence and performance distinction in cognitive development (e.g., in seminal writings such as Chomsky (1965) and Gelman and Gallistel, (1978)), as well as in the use of related distinctions such as that between procedural and conceptual knowledge (Rittle-Johnson & Schneider & Starr, 2015) likely reflects researchers' desire to distinguish between the hidden potential for certain forms of nervous system functioning and the potential for certain forms of overt behavioral performance. While a distinction along these lines is certainly useful for many purposes, the present results suggest that the nervous system capacities in question cannot usefully be characterized in terms of knowledge claims, since insofar as the latter are descriptions of behavior, they cannot realistically be distinguished from behavioral performances.

Future Questions

It cannot be denied that the behavior of any typical person entails actual, rather than just described, patterned behavioral capacities (these might be described as attractor states, such as the stable limit cycle seen in walking, or the tendency towards walking as a solution to the desire to move). Without a doubt, the quasi-predictability that such patterns provide is a prerequisite for the utility of any generalizations about behavior (e.g., knowledge claims). To the extent that knowledge claims have predictive power, a crucial question for future research is how and why the dispositions described by knowledge claims correspond in any isomorphic way to these actual dispositional capacities of the typical human organism.

Acknowledgments

Thanks to Joseph Glick for intellectual support in the development of this project.

References

Allen, J.W.P. & Bickhard, M. (2013). Stepping off the Pendulum: Why Only an Action-Based Approach Can Transcend the Nativist-Empiricist Debate. *Cognitive Development, 28*(2), 96-133.

Briars, D., & Siegler, R. S. (1984). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology*, *20*(4), 607.

Brooks, N., Audet, J., & Barner, D. (2012). Pragmatic Inference, Not Semantic Competence, Guides 3-Year-Olds' Interpretation of Unknown Number Words. *Developmental Psychology, 49*(6), 1066-75.

Byers, P. (2016). Knowledge claims in cognitive development research: Problems and alternatives. *New Ideas in Psychology, 43*, 16-27. doi:10.1016/j.newideapsych.2016.03.004

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction? *Cognition*, *123*(1), 162–173. http://doi.org/10.1016/j.cognition.2011.12.013

Frye, D., Braisby, N., Lowe, J., Maroudas, C., & Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child Development*, *60*(5), 1158–1171.

Fuson, K. C. (1988). *Children's counting and concepts of number*. Springer-Verlag.

Gelman, R. (1972). Logical Capacity of Very Young Children: Number Invariance Rules. *Child Development*, *43*(1), 75–90. http://doi.org/10.2307/1127873

Gelman, R., & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Cambridge, MA: Harvard University Press.

Gelman, R., & Meck, E. (1983). Preschoolers' Counting: Principles Before Skill. *Cognition*, *13*(3), 343–359.

Greeno, J. G., Riley, M. S., & Gelman, R. (1984). Conceptual competence for children's counting. *Cognitive Psychology, 16,* 94-143.

Nikoloska, A. (2009). Development of the cardinality principle in Macedonian preschool children. *Psihologija*, *42*(4), 459–475. http://doi.org/10.2298/PSI0904459N

Rittle-Johnson, B., Schneider, M., & Star, J. R. (2015). Not a One-Way Street: Bidirectional Relations Between Procedural and Conceptual Knowledge of Mathematics. Educational Psychology Review, *27* (4) 587-597.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674.

Sarnecka, B. W., & Wright, C. E. (2013). The Idea of an Exact Number: Children's Understanding of Cardinality and Equinumerosity. *Cognitive Science*, *37*(8), 1493–1506. http://doi.org/10.1111/cogs.12043

Sophian, C. (1988). Early Developments in Children's Understanding of Number: Inferences about Numerosity and One-to-One Correspondence. *Child Development*, 59(5), 1397–1414. http://doi.org/10.2307/1130502

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193. http://doi.org/10.1016/0010-0277(90)90003-3

Concept Membership vs Typicality in Sentence Verification Tasks

Francesca Zarl (francesca.zarl@phd.units.it)
Psychology Unit "Gaetano Kanizsa", Department of Life Sciences,
University of Trieste, via Weiss 21, I - 34128 Trieste, Italy

Danilo Fum (fum@units.it)
Department of Humanistic Studies, University of Trieste
via Tigor 22, I - 34134 Trieste, Italy

Abstract

In the paper we discuss the relation between fuzzy sets and the graded membership and typicality effects found in the study of concepts. After a short overview of the topic, we present three experiments, carried out using the same method but with different situational contexts, which examine whether graded membership and typicality could be considered as independent factors capable of influencing the performance of human participants involved in sentence verification tasks, or they are somehow interrelated. The paper concludes with a general discussion of the experimental findings and the problems they pose for models of concepts based on the theory fuzzy sets.

Keywords: Concept representation; fuzzy set theory; graded membership; vagueness; typicality; sentence verification, categorization.

Introduction

The study of concepts has constituted a key point for psychological research since its very beginning. In the last decade or so, it has acquired new importance in the light of the Semantic Web endeavor to find a computationally effective way to model the ontologies machines need to perform satisfactorily at cognitive-related tasks.

The so-called classical view (Smith & Medin, 1981) regards concepts as rules or definitions, i.e., as sets of properties which are individually necessary and jointly sufficient to determine whether or not a given entity is an instance of the concept in question. From a formal point of view, concepts can be modeled using sets whose intensions are defined by the rules. According to the law of excluded middle, an element is a member of a set (if it meets all the necessary properties) or it is not (if there is at least one necessary property which is not met). Moreover, since all the elements of a set satisfy the same conditions, there is no difference among them as far as their membership is concerned.

The classical view sounds reasonable and intuitive but it is in fact inadequate to explain several empirical phenomena psychologists have found by running their experiments. A number of empirical studies (e.g., Barsalou, 1989) showed that people, required to define a concept, have difficulty in generating lists of properties that are necessary and sufficient for it. Morevover, in trying to identify the features characterizing many everyday concepts, people disagree with each other, and sometimes they disagree also with themselves, with the same person generating different lists on diverse occasions (Bellezza, 1984).

These results depend on the fact that most of our concepts are vague, and they do not have clearly defined boundaries. If vagueness constitutes sometimes a definite asset, it has the disadvantage that none of our categories will ever fit completely with the world, and there will be always cases in which it will be difficult to discriminate whether an instance belongs to a concept or not.

Hampton (1993) found, for example, that people consider some entities as just barely members of a category and other entities as just merely non-members. Members and non-members form a continuum without a clear distinction between them. While it is easy to classify unambiguous elements, people face serious difficulties in determining whether an in-between element does belong or not to a particular category. McCloskey and Glucksberg (1978) found that people, asked to make membership decisions for exemplar-category pairs in two separate sessions, agreed with each other and were consistent between sessions for items clearly related or unrelated with the category (e.g., *apple-fruit* and *cucumber-furniture*, respectively), while they disagreed and were frequently inconsistent in the case of borderline elements (e.g, *curtains-furniture*).

Another embarrassing result for the classical view is constituted by the fact that people consider certain entities as better exemplars of a concept than others; for instance, *sparrows* and *robins* are somehow considered as better *birds* than *ostrichs* or *penguins* (Rosch, 1973). This finding, known as the typicality effect, represents one of the most common and most robust phenomena found in the study of concepts.

It has been found that typicality affects people's performance in a variety of ways and in wide range of cognitive tasks. Typical exemplars of a category are classified faster and more accurately than atypical ones (Rips, Shoben, & Smith, 1973); they are more likely to be considered as members of the category (Hampton, 1979); they are mentioned more frequently when asked to name the members of the category (Mervis, Catlin, & Rosch, 1976); they are the first to be learned in artificial category learning tasks (Rosch, Simpson, & Miller, 1976); they support better analogical inferences (Rips, 1975), etc.

Being incompatible with the classical view of concepts, graded membership and typicality cannot be modeled according to the set-theoretic account on which such a theory is based. Soon after the discovery of these phenomena, an alternative framework for concepts that relied on the idea of fuzzy sets (Zadeh, 1965) was put forward.

In the paper we discuss the relation between fuzzy sets

and the graded membership and typicality effects found in the study of concepts. After a short overview of the topic, we present three experiments, carried out using the same method but with different situational contexts, which examine whether graded membership and typicality could be considered as independent factors capable of influencing the performance of human participants involved in sentence verification tasks, or they are somehow interrelated. The paper concludes with a general discussion of the experimental findings and the problems they pose for models of concepts based on the theory fuzzy sets.

Graded membership, typicality, and fuzzy sets

Fuzzy sets were introduced by Zadeh (1965) to overcome the limitations of the traditional set-theoretic approach in dealing with classes of objects that are not clearly defined. Because these classes are pervasive in all the processes that somehow involve language, fuzzy sets have been considered since their beginning as a promising formalism to represent conceptual knowledge. Following the publication of a critical paper by Osherson and Smith (1981), the interest for fuzzy sets rapidly declined among cognitive scientists determining their practically disappearance from the literature. Recently, however, a recent reevaluation of the topic by Belohlavek and Klir (2011) showed that many arguments raised against fuzzy set were fallacious, and contributed to restate their relevance for the study of concepts.

Fuzzy sets can be conceived as an extension of classical sets. While the characteristic function of the latter maps its domain into the set $\{0, 1\}$, comprising only two elements, the range of the former's function contains an infinite number of elements: all the real numbers in the interval [0,1], including its limits. In this sense, a classical set is a subset of a fuzzy one. While classical sets consider 0 and 1 symbolically, as Boolean entities, the degrees of membership of fuzzy sets have a true numerical value. One important consequence of this fact is that fuzzy sets can be mathematically manipulated in ways that are not allowed by their classical counterpart.

It is evident that both the graded membership and the typicality effect found in the study of concepts can be modeled through fuzzy sets by introducing a function having as range the interval of real numbers [0,1]. The question to be asked is whether graded membership and typicality should be considered as separate phenomena which however depend on an underlying common factor, or whether they denote distinct dimensions to be captured by different functions.

According to Cai, Yeung, and Leung (2012) graded membership and typicality have different nature and are not necessarily related. An instance of a concept can in fact be assigned a high degree of membership and a low degree of typicality (as it is the case for *ostrichs* as members of the class of *birds*) while, on the other hand, there may be entities that have a membership degree close to zero and non-null typicality. This latter case occurs when people consider some entities as exemplars of a given concept (for example, consider *whales* as *fish* or *tomatoes* as *vegetable*) while, technically, they cannot be considered as such. As stated by the authors, graded membership and typicality play different roles in determining whether an entity is an instance of a concept, and are computed in different ways. To determine the degree of membership, the necessary conditions which define a concept according to the classical view are taken into account. Typicality provides an additional mechanism to rank those instances that meet all the membership requirements (and whose values approximate therefore 1). In this case the non-defining features which are widely shared among the set members are tallied. Sometimes this leads to the assignment of typicality values to entities that, without being members of the set, partake of its characteristic properties.

Kamp and Partee (1995) too deny that a single measure could serve the purposes of quantifying both the degree to which an entity is an instance of a given concept and its typicality, interpreted as the degree of proximity to the best example (or prototype) of the concept. The question they ask is whether and how the two functions could be considered as connected, and in particular for which concepts, if any, do they coincide. As an example of lack of interrelation Kamp and Partee refer to the concept of *male nurse*. In this case, according to the authors, knowing the degree of membership—which depends on the intersection of the classes of *males* and *nurses*—does not help in any way to establish typicality. On the other hand, there are cases (for instance, for the concepts of *red* or *chair*) in which knowing the typicality value of an instance helps to establish its membership degree.

Another interesting point of view on this issue is raised by Hampton (2011) who considers typicality and graded membership as separate functions that are based, however, on the same underlying similarity measure. More particularly, "[t]ypicality is a monotonically rising function of similarity, whereas membership is a nondecreasing function of similarity that starts at 0, starts to rise at a certain point k_1, and then reaches a ceiling of 1 at a further point k_2, where k_1 and k_2 are above the minimum and below the maximum values that similarity can take." (Hampton, 2011, p.219)

To better define the relationship between graded membership and typicality, and their possibile connections with fuzzy sets, we realized the following experiments.

Experiments

All the experiments investigated the role that graded membership and typicality play in determining the behavior of people engaged in sentence verification tasks. The experiments asked participants how much they agreed with a series of sentences claiming that *"...Xs are/are not Ys"*. Beyond asserting or negating the membership of X to Y, the sentences differed in the relationship that connected an instance with its putative category. X could be in fact a typical member of Y, an atypical member, a non-member sharing common features with the members of Y, and a non-member without any evident relation with Y. In each experiment, therefore, member-

ship and typicality were orthogonally varied, as it was varied the polarity (affirmative vs negative) of each sentence.

The experiments differed according to the context in which membership judgments were made. It is reasonable to assume that the degree of agreement with a statement could depend not only on the intrinsic relationship between an entity and a class but also on the particular viewpoint taken by participants, or by the specific circumstances in which the sentences are evaluated. (For a review of studies on concept classification manipulating the context of the task, see: Murphy, 2002, pp. 413-422.)

By adopting different kinds of contexts, we tried to determine the generality and robustness of any influence membership and typicality could have on the concept categorization processes underlying the sentence verification task.

Experiment 1

In Experiment 1 we investigated whether the effect of membership and typicality could be modulated by providing participants with a purposive context for sentence evaluation. Our approach is similar to that employed by Hampton, Dubois, and Yeh (2006) who adopted instructions contrasting a purely pragmatic classification context with a more technical one, and compared them with a no-context condition. Contrary to the authors' expectations, however, none of the dependent measures was influenced by the context. We conjectured that, by emphasizing the context also in the sentence text, we could differently influence the criteria participants used to evaluate the sentences.

Participants. Sixty University of Trieste students (48 females), whose age varied from 18 to 53 years (mean = 23.1; sd = 8.0), participated to the experiment. Participants were randomly assigned to three experimental conditions named *Technical*, *CommonSense*, and *Neutral*, respectively.

Design. A 3x2x2x2 mixed design was adopted having *Context* (*Technical, CommonSense*, and *Neutral*) as a between-subjects factor and *Membership* (*Yes* vs *No*), *Typicality* (*Yes* vs *No*), and *Polarity* (*Positive* vs *Negative*) as factors within.

Materials and procedure. In the experiment 64 Italian sentences – the same used in Zarl and Fum (2014) – each involving a relation between an instance and a category, were used. The sentences were divided into 32 pairs. A sentence in each pair made an affirmative statement while the other negated it. The sentences were constructed by balancing the gender and the number of instances and categories which were both of natural (e.g., *tomato-fruit*) and artificial (e.g., *volleyball-sport*) kind. Eight different types of sentences were constructed by varying the three distinct factors of *Membership, Typicality,* and *Polarity*. Table 1 provides an example for each kind of sentence.

Positive sentences are labeled with P while negative sentences are labeled with N. M means that the instance is a member of the category while while $\overline{\text{M}}$ negates it. Analogously, T means that the instance shares some common features with members of the category while $\overline{\text{T}}$ negates any similarity between the instance and category.

Table 1: A sample of sentences used in the experiment.

Sentence kind	Text
PMT	... *canaries are birds*
PM$\overline{\text{T}}$	... *penguins are birds*
P$\overline{\text{M}}$T	... *bats are birds*
P$\overline{\text{M}}\overline{\text{T}}$	... *toads are birds*
NMT	... *toads are not birds*
NM$\overline{\text{T}}$	... *bats are not birds*
N$\overline{\text{M}}$T	... *penguins are not birds*
N$\overline{\text{M}}\overline{\text{T}}$	... *canaries are not birds*

The interpretation of positive sentences is straighforward. PMT means that the instance is a typical member of the category, PM$\overline{\text{T}}$ that it is an atypical member, etc. The labeling of negative sentences, obtained by negating the labels of the positive ones, is based on the criteria participants would follow in agreeing with the sentence content. So, for example, accepting the N$\overline{\text{M}}$T sentence *"... penguins are not birds"*, which negates the PM$\overline{\text{T}}$ *"... penguins are birds"*, means denying membership while acknowledging typicality as evaluation criterion because *penguins*, even if they are in fact *birds*, lack some features that are typical of this category.

The sentences of each experimental condition were introduced by a different phrase which provided a context for their reading. In the case of the *Technical* group, the sentences began with the expression *"In a technical sense ..."*, while those of the *CommonSense* group were introduced by the words *"According to common sense ..."*. For the *Neutral* condition we borrowed the phrase *"In a sense ..."* by Machery and Seppälä (2011) who used it to allow different interpretations for the concepts of their experiments.

The context according to which participants were asked to evaluate the sentences was varied in the instructions, too. For the *Technical* group, the instructions highlighted the fact that concepts are structured according to a taxonomy based on strict membership rules. Participants assigned to the *CommonSense* condition were said that a looser interpretation of concepts would take into account the similarity that exists between them. Finally, the instructions for the *Neutral* group were quite general and did not provide any specific indication about the setting to be adopted.

All the sentences were gathered in a leaflet whose pages contained eight sentences drawn randomly from the total pool. Next to each sentence was printed a 7 cm line whose extreme points were marked with the labels *Completely disagree* and *Completely agree*, respectively. Participants had to indicate their degree of agreement with the sentence by putting a vertical mark on the line. The position of the mark was measured at the next millimeter and converted into a score in the [0,70] interval of integers.

Results A four-way mixed ANOVA with one variable between-subjects (*Context*) and three variables within-

subjects (*Membership, Typicality, and Polarity)* was performed on the experimental data. All the main effects of the within-subjects factors were significant: Sentences in which the instance was a member of the category obtained higher ratings than sentences in which the instance was a non-member ($F_{(1,57)} = 338.67, p < .001$). Analogously, scores for sentences in which the instance was similar to the category typical members were higher than those of sentences in which there was no similarity between the instance and the category ($F_{(1,57)} = 443.91, p < .001$). Finally, affirmative sentences obtained higher judgments than negative ones ($F_{(1,57)} = 64.10, p < .001$). On the other hand, the different contexts did not have any effect on the participants performance, and did not lead to any significant interaction with the other factors.

Table 2 reports the mean scores for the different sentence kinds. To facilitate the understanding of the data, the contextual conditions have been collapsed in the table.

Table 2: Average scores – contextual conditions collapsed.

	Positive		Negative	
	$\overline{\text{M}}$	M	$\overline{\text{M}}$	M
$\overline{\text{T}}$	13.21	44.27	9.21	33.46
T	36.00	62.80	27.79	54.50

An interesting (and puzzling) three way interaction *Membership x Typicality x Polarity* ($F_{(1,57)} = 4.58, p < .05$) was found among the within-subjects variables. The interaction effect has quite a small magnitude but it constitutes an original and unexpected result that requires an adequate explanation (see: the *General discussion and conclusions* section).

To determine which criterion, between *Membership* and *Typicality*, was more influential in determining the participants judgments, a second analysis of the data was made taking into account only the sentences is which the two criteria were directly contrasted. A two way mixed ANOVA was carried out having *Context* as a variable between, and *Criterion* (i.e., *Membership*, for the PM$\overline{\text{T}}$ and NM$\overline{\text{T}}$ sentences, vs *Typicality*, for the P$\overline{\text{M}}$T and P$\overline{\text{M}}$T ones) as variable within. The ANOVA revealed only the significant main effect ($F_{(1,57)} = 19.13, p < .001$) of *Criterion*, with *Membership* sentences obtaining significantly higher ratings than *Typicality* ones (average scores of 38.86 and 31.90, respectively).

Experiment 2

The experiment aimed to assess whether the influence of *Membership* and *Typicality*, and their possible interaction with *Polarity*, could depend on how the material was presented. In a previous work (Zarl & Fum, 2014), we found that participants were more likely to agree with sentences that contradicted each other (e.g., "*In a sense penguins are birds*" vs "*In a sense penguins are not birds*"), and to assigning therefore high scores to both, when the sentences were submitted separately in random order, while contradiction acceptance was lower when they were presented together for the evaluation. Experiment 2 was therefore a control experiment intended to ascertain the generality of the effects of Experiment 1, and to confirm maybe its findings.

Participants. Forty voluntary participants (31 females) from the Trieste area whose age varied from 20 to 42 years (mean = 26.1, sd = 4.8) were engaged in the experiment. Participants were randomly assigned to two experimental conditions named *Pair* vs *Single*.

Design. A 2x2x2x2 mixed design was adopted with *Context* (*Pair* vs *Single*) as a variable between, and *Membership*, *Typicality* and *Polarity* as variables within subjects.

Materials and Procedure. The materials and the instructions of the *Neutral* condition of Experiment 1 were used, with all the sentences introduced by the phrase "*In a sense* ..." and no particular hint provided for the their interpretation. For participants in the *Single* condition, the procedure too was identical to that of the previous experiment, with the sentences presented in random order. The positive and negative sentences of each pair were instead presented together, one below the other, to participants of the *Pair* condition. For this condition, the pairs presentation order, and the order of presentation of the positive and negative sentence within each pair were determined randomly.

Results. The experiment replicated all the findings of Experiment 1. A four-way ANOVA showed the main effects of *Membership* ($F_{(1,38)} = 181.04, p < .001$), *Typicality* ($F_{(1,38)} = 124.872, p < .001$), and *Polarity* ($F_{(1,38)} = 52.889, p < .001$) only. Table 3 shows the mean scores for the different sentences. Similarly to what was done in Experiment 1, because the presentation *Context* did not proved significant, the data are presented without taking it into account.

Table 3: Average scores – contextual conditions collapsed.

	Positive		Negative	
	$\overline{\text{M}}$	M	$\overline{\text{M}}$	M
$\overline{\text{T}}$	14.49	50.15	7.31	37.38
T	36.12	59.82	25.72	55.26

The ANOVA confirmed also the three-way interaction *Membership* x *Typicality* x *Polarity* ($F_{(1:38)} = 14.21, p < .001$) previously found in Experiment 1 that becomes here more perspicuous: the gain due to *Typicality* is clearly greater for the non-members $\overline{\text{M}}$, in the case of *Positive* sentences, and for members M, when the sentences are *Negative*.

Another interesting interaction *Context* x *Membership* ($F_{(1,38)} = 6.25, p < .05$) found by the ANOVA and absent in Experiment 1 is reported in Table 4.

According to the data, when the sentences are presented simultaneously in pairs, the participants judgments become more polarized (i.e. with higher scores for M sentences and lower scores for the $\overline{\text{M}}$ ones) in comparison with the case in

Table 4: The *Context* x *Membership* interaction

	Member M	Non-Member $\overline{\text{M}}$
Pair	58.88	21.20
Single	51.64	30.25

which the sentences are presented singularly and in a randomized order. This polarization effect obtained in the *Pair* condition confirms what had already been pointed out previously by Zarl and Fum (2014).

Finally, a second ANOVA taking into account the positive and negative sentences of the M$\overline{\text{T}}$ and $\overline{\text{M}}$T kind was performed to determine which criterion, between *Membership* and *Typicality*, was more influential in determining the participants judgments. This mixed two-way ANOVA was carried out having *Context* as between-subjects factor and *Criterion* as factor within. The ANOVA revealed only the significant main effect ($F_{(1,38)} = 31.32, p < .001$) of *Criterion*, with *Membership* sentences (mean = 43.76) obtaining significantly higher ratings than *Typicality* ones (mean = 30.92).

Experiment 3

A final experiment was carried out to check whether the previous findings could somehow depend on the fact that all the sentences were always preceded by an introductory phrase. The idea for this control experiment derived from some occasional remarks made by participants in Experiment 2 and by those assigned to the *Neutral* condition of Experiment 1 who, after reading some sentences introduced by *"In a sense ... "*, asked "In which sense?". Experiment 3 thus contrasted the simple assertion or negation of membership with the more elaborate formulation adopted to induce a neutral context for sentence interpretation.

Participants. The participants were 64 residents in the Friuli Venezia Giulia region in Italy (49 females) whose age varied from 19 to 63 years (mean = 25.74; sd = 8.2).

Design. The same 2x2x2x2 mixed design of Experiment 1 was adopted. In this case the levels of the between-subjects *Context* variable were *Neutral* and *NoContext*, respectively.

Materials and procedure. For participants in the *Neutral* condition, the materials and the procedure were identical to those of the *Single* condition of Experiment 2 (and of the *Neutral* condition of Experiment 1). Participants in the *NoContext* condition read the same sentences with the introductory phrase *"In a sense ... "* removed (e.g. *"Penguins are birds", "Canaries are not birds"* etc.).

Results. The same pattern of results found in the previous experiments was replicated in Experiment 3, too. More particularly, the *Context* factor did not have any statistically significant effect on the participants performance indicating that the previous findings were not influenced by the presence of the introductory phrase *"In a sense ... "*. Table 5 reports the average scores for the different conditions.

Again, a four-way ANOVA revealed the significative main

Table 5: Average scores–contextual conditions collapsed.

	Positive		Negative	
	$\overline{\text{M}}$	M	$\overline{\text{M}}$	M
$\overline{\text{T}}$	11.68	47.62	5.94	32.99
T	37.05	64.05	22.38	54.64

effects of *Membership* ($F_{(1,62)} = 433.67, p < .001$), *Typicality* ($F_{(1,62)} = 585.33, p < .001$), and *Polarity* ($F_{(1,6262)} = 188.44, p < .001$), and their three-way interaction ($F_{(1,62)} = 19.72, p < .001$) indicating the greater advantage due to Typicality that occurs in the *Positive* $\overline{\text{M}}$ sentences.

A mixed two-way ANOVA, using *Context* and *Criterion* as factors was carried out on sentences contrasting directly *Membership* and *Typicality*. It showed a significant ($F_{(1,62)} = 40.48, p < .001$)) main effect of *Criterion*, obtaing the former (mean = 40.31) higher scores than the latter (mean = 29.72).

General discussion and conclusions

In the previous section we have reported three experiments, all leading to concordant results. First of all, the experiments confirmed the graded membership and the typicality effects characterizing concepts, and their importance in determining the degree of agreement with a series of statements asserting or denying the membership of an instance to a class.

Between the two factors, membership played the dominant role. In all the experiments, when the factors were directly contrasted, participants assigned higher scores to sentences in which the instances were atypical members of the class in comparison with those in which the instances simply shared some similarity with the class members.

This is a general result that was not affected by any attempt to modulate it by providing participants with different evaluation contexts. Even when explicitly instructed to take into account the similarity between the instance and the typical exemplars of the class (Experiment 1), participants were guided by the membership criterion. In Experiment 2 the simultaneous comparison between contradictory statements led to a polarization of judgments (with higher scores for sentences asserting membership and lower scores for those denying it, when the instance was indeed a member of the class) but did not change the general pattern of results. The same was true in Experiment 3 when assertions and negations were expressed directly in the sentences or were somehow dampened by the introductory phrase *"In a sense"*.

Membership and typicality are not independent factors but they interact with each other in a subtle way. In case of affirmative sentences, the gain in the score due to typicality is higher for non-members than for the members of a category while it is virtually identical when the sentences are negative. This is the most interesting and original finding deriving from all the experiments and, at least to our knowledge, no existing model of concepts is able to predict and explain it.

The results of our experiments highlight some interesting issues for models of concepts based on the theory of fuzzy sets. Due to space limitations we mention only two, both

related to the interpretation of negative sentences.

From a strictly logical point of view, denying that an instance is a member of a set can be considered as equivalent to stating that it is a member of the set complement: negating that *bats* are *birds* is like claiming that they are *non-birds*. Concepts such as *not a bird*, defined in a negative way, are often cited (Connolly, Fodor, Gleitman, & Gleitman, 2007) as examples of concepts for which no prototype exists. Although it may be relatively easy to establish whether a certain instance is a *non-bird*, in this determination the similarity between the instance and *non-birds* should not play any role. The problem is that in our experiments typicality manifests its effect also in negative sentences: the $\overline{\text{NMT}}$ *"Penguins are not birds"* receives higher scores than the $\overline{\text{NMT}}$, *"Canaries not are birds"*, and the same is true for the sentences whose instance is a member of the category.

It is possibly true, as state by a paper reviewer, that "the fact that typicality is present in negative sentences does not mean that it is the effect of any prototype for the negation of the concept. It just means that participants are still capable of evaluating the distance of a nonmember to the prototype of the positive category in terms of similarity". The fact is that what appears as a strong pragmatic effect of negation cannot be easily adapted into the mathematical theory of fuzzy sets.

A second issue concerns the scores given to a sentence and its negation. The function defining the complement of a fuzzy set is monotonically decreasing, i.e. for every instance a, b of a set A and their respective membership degrees, $\mu(a)$ and $\mu(b)$, if $\mu(a) < \mu(b)$ then $c(\mu(b)) \geq c(\mu(a))$, where c is the function defining the complement of A. While this relation is maintained in the aggregated data, in many cases the individual scores given by participants do not respect this property, a fact that casts some further doubts about the possibility of translating directly the axioms of the mathematical theory of fuzzy sets into predictions about the participants behavior.

References

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 76–121). New York: Cambridge University Press.

Bellezza, F. S. (1984). Reliability of retrieval from semantic memory: Common categories. *Bulletin of the Psychonomic Society*, *22*, 324–326.

Belohlavek, R., & Klir, G. J. (2011). Fallacious perceptions of fuzzy logic in the psychology of concepts. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and fuzzy logi* (pp. 121–148). Cambridge, MA: The MIT Press.

Cai, Y., Yeung, C. M. A., & Leung, H. F. (2012). *Fuzzy computational ontologies in context*. Beijing and Berlin: Higher Education Press and Springer.

Connolly, A., Fodor, J. A., Gleitman, L. R., & Gleitman, H. (2007). Why stereotypes don't even make good defaults. *Cognition*, *103*, 1-22.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *18*, 441–461.

Hampton, J. A. (1993). Prototype models of concept representation. In I. V. Mechelen, J. A. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67–95). London: Academic Press.

Hampton, J. A. (2011). Conceptual combinations and fuzzy logic. In R. Belohlavek & G. J. Klir (Eds.), *Concepts and fuzzy logic* (pp. 209–231). Cambridge, MA: The MIT Press.

Hampton, J. A., Dubois, D., & Yeh, W. (2006). Effects of classification context on categorization in natural categories. *Memory & Cognition*, *34*, 14319–14436.

Kamp, E., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.

Machery, E., & Seppälä, S. (2011). Against hybrid theories of concepts. *Anthropology and Philosophy*, *10*, 97–126.

McCloskey, M., & Glucksberg, S. (1978). Natural categories: Well-defined or fuzzy set? *Memory & Cognition*, *6*, 462–472.

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationship among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*, 283–284.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35–58.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665–681.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1–20.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. M. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111–144). New York: Academic Press.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 491–502.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353.

Zarl, F., & Fum, D. (2014). Theories of concepts and contradiction acceptance. In P. Bernardis, C. Fantoni, & W. Gerbino (Eds.), *TSPC2014: Proceedings of the Trieste Symposium on Perception and Cognition* (pp. 157–161). Trieste, Italy: EUT: Edizioni Università di Trieste.

Better safe than sorry: Risky function exploitation through safe optimization

Eric Schulz[1], **Quentin J.M. Huys**[2], **Dominik R. Bach**[3], **Maarten Speekenbrink**[1] **& Andreas Krause**[4]

[1]Department of Experimental Psychology, University College London, London, WC1H0AP
[2]Translational Neuromodeling Unit, ETH and University of Zürich, Wilfriedstrasse 6, 8032 Zürich
[3]Psychiatric Hospital, University of Zürich, Lenggstrasse 31, 8032 Zürich
[4]Institute for Machine Learning, ETH Zürich, Universitaetstrasse 6, 8092 Zürich

Abstract

Exploration-exploitation of functions, that is learning and optimizing a mapping between inputs and expected outputs, is ubiquitous to many real world situations. These situations sometimes require us to avoid certain outcomes at all cost, for example because they are poisonous, harmful, or otherwise dangerous. We test participants' behavior in scenarios in which they have to find the optimum of a function while at the same time avoid outputs below a certain threshold. In two experiments, we find that Safe-Optimization, a Gaussian Process-based exploration-exploitation algorithm, describes participants' behavior well and that participants seem to care first about whether a point is safe and then try to pick the optimal point from all such safe points. This means that their trade-off between exploration and exploitation indicates intelligent, approximate, and homeostasis-driven behavior.

Keywords: Safe Optimization, Function Learning, Approximate Learning, Gaussian Process, Homeostasis

Introduction

Imagine you are hosting a dinner party. In the afternoon, you open up your fridge and kitchen cupboards to find a plethora of ingredients at your disposal. Aiming to amaze your friends with an unique culinary experience, you decide to prepare something extraordinary not found in recipe books. Considering your options, you generate expectations of how the tastes of different ingredients combine and interact to produce a – hopefully memorable – culinary experience. You have time to try out some options and experience their overall taste, learning about the effects of unusual combinations and methods of preparation. At the same time, however, you need to avoid certain combinations at all costs, for example those that are inedible, poisonous, or otherwise bad.

This scenario is an example of a multi-armed bandit task (Srinivas et al., 2009), where there are a number of actions or 'arms' of the bandit (e.g., the possible dishes) which lead to initially unknown and stochastic outcomes or rewards (e.g., the taste of the dish), which are related to a set of features (e.g., the ingredients, the method of preparation, etc.). Through experience, one can learn the function which maps the features to the rewards and maximize the overall rewards gained over repeated plays of the bandit. A key issue in optimal behavior in such tasks is known as the exploration-exploitation dilemma: should I take an action which I know will lead to a high reward, or try an unknown action to experience its outcome and thereby learn more about the function mapping features to rewards, increasing my chances of gaining higher rewards in the future? In order to avoid certain bad outcomes (e.g., poisonous dishes), one should only explore uncertain options which are likely to be 'safe'. Such restricted exploration-exploitation problems are ubiquitous in daily life, from choosing which restaurant to visit, which car to buy, all the way to whom to befriend. In our previous research on human behavior in contextual multi-armed bandits (Schulz et al., 2015a,b), we found that participants' behavior is well-described by Gaussian Process regression, a non-parametric regression tool that adapts its complexity to the data at hand by the means of Bayesian posterior computation.

The aim of the present study is to assess how people behave when they have to maximize their rewards whilst avoiding outcomes below a given threshold. The task is couched as a function learning task, where participants choose an input and observe and accrue the output of the function. In two experiments with a uni- and bivariate function, we find that participants efficiently adapt their exploration-exploitation behavior to risk-inducing situations. Overall, they are well-described by a Gaussian Process-based safe optimization algorithm that tries to safely expand a set of 'explorers' while simultaneously maximizing outputs within a set of possible 'maximizers; (Sui et al., 2015). Such behavior might be based on the principle of homeostasis maintenance (Korn & Bach, 2015), where organisms need to forage for food while avoiding the probability of starvation. Additionally, we find evidence that participants first assess whether points are safe and then attempt to maximize within this safe subset. This simplification of the task in terms of subgoals resonates well with recent results on approximate planning strategies in complex dynamic tasks (Huys et al., 2015).

Modeling learning and optimization

If the task is to learn and maximize an unknown function, then two ingredients are needed: (a) a model to represent an unknown function, for which we will use Gaussian process regression, and (b) a method to safely

[1]Corresponding author: eric.schulz@cs.ucl.ac.uk

choose the next inputs, for which we will use a safe optimization algorithm.

Learning a function

We assume people represent and learn a function through Gaussian process regression, a universal function learning algorithm which has been supported in previous research (Griffiths et al., 2009; Schulz et al., 2015b).
A Gaussian Process ($\mathcal{GP}$) is a stochastic process of which the marginal distribution of any finite collection of observations is multivariate Gaussian (Rasmussen, 2006). It is a non-parametric Bayesian approach towards regression problems and can be seen as a rational model of function learning as it adapts its complexity to the data encountered. Let $f(\boldsymbol{x})$ be a function mapping an input $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ to an output y. A $\mathcal{GP}$ defines a distribution $p(f)$ over such functions. A $\mathcal{GP}$ is parametrized by a mean function $m(\boldsymbol{x})$ and a covariance (or kernel) function, $k(\boldsymbol{x}, \boldsymbol{x}')$:

$$m(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{x})\right] \tag{1}$$
$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))\right] \tag{2}$$

At time t, we have collected observations $\mathbf{y}_{1:t} = [y_1, y_2, \ldots, y_t]^\top$ at inputs $\boldsymbol{x}_{1:t} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t)$. For each outcome y_t, we assume

$$y_t = f(\boldsymbol{x}_t) + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

Given a $\mathcal{GP}$ prior on the functions

$$f(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right), \tag{4}$$

the posterior over f is also a $\mathcal{GP}$ with

$$m_t(\boldsymbol{x}) = k_{1:t}(\boldsymbol{x})^\top (\boldsymbol{K}_{1:t} + \sigma^2 \boldsymbol{I}_t) \boldsymbol{y}_{1:t} \tag{5}$$
$$k_t(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_{1:t}(\boldsymbol{x})^\top (\boldsymbol{K}_{1:t} + \sigma^2 \boldsymbol{I}_t)^{-1} \boldsymbol{k}_{1:t}(\boldsymbol{x}') \tag{6}$$

where $\boldsymbol{k}_{1:t}(\boldsymbol{x}) = [k(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, k(\boldsymbol{x}_t, \boldsymbol{x})]^\top$, $\boldsymbol{K}_{1:t}$ is the positive definite kernel matrix $[k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\ldots,t}$, and $\boldsymbol{I}_t$ is a t by t identity matrix. This posterior distribution can be used to derive predictions for each possible input $\boldsymbol{x}$ on the next time point, which again follow a Gaussian distribution. A key aspect of a $\mathcal{GP}$ is the covariance or kernel function k. The choice of a kernel function corresponds to assumptions about the kind of functions a learner expects. Here, we will use a squared exponential kernel:

$$k_{\text{sqe}}(\boldsymbol{x}, \boldsymbol{x}') = \theta_1^2 \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}')^2}{2\theta_2^2}\right) \tag{7}$$

This kernel induces a universal function learning engine and has been found to describe human function learning well (Griffiths et al., 2009).

Optimizing a function

Given a learned representation of a function at time t, this knowledge needs to be used to choose a next input at time $t+1$. This is done through an acquisition function that takes the expected output for each input and the associated uncertainty to balance exploration and exploitation (Brochu et al., 2010).
An algorithm that is well-poised to cope with the additional requirement to avoid outcomes below a threshold first separates possible inputs into those that are likely to provide outputs above the threshold (the safe set) and those that are not, and then separates this safe set further into a set of maximizers (inputs that are likely to provide the maximum output) and expanders (inputs that are likely to expand the safe set). Following Berkenkamp et al. (2015), we define upper and a lower bounds of a confidence interval as sum of the current expectation m_{t-1} and its attached uncertainty σ_{t-1}.

$$u_t(\boldsymbol{x}) = m_{t-1}(\boldsymbol{x}) + \beta_t \sigma_{t-1}(\boldsymbol{x}) \tag{8}$$
$$l_t(\boldsymbol{x}) = m_{t-1}(\boldsymbol{x}) - \beta_t \sigma_{t-1}(\boldsymbol{x}). \tag{9}$$

the parameter β_t determines the width of the confidence bound, and we set it to $\beta_t = 3$ to assure high safety a priori (i.e. 99.9%). Using these bounds, we can define the safe set as all the input points in the set $\mathcal{X}$ of available inputs that are likely to lead to output values above the safe threshold, $J_{\min}$

$$\mathcal{S}_t = \{\boldsymbol{x} \in \mathcal{X} | l_t(\boldsymbol{x}) \geq J_{\min}\} \tag{10}$$

The set of potential maximizers contains all safe inputs that are likely to obtain the maximum output value; these are the safe inputs for which the upper confidence bound u_t is above the best lower bound:

$$\mathcal{M}_t = \{\boldsymbol{x} \in \mathcal{S}_t | u_t(\boldsymbol{x}) \geq \max_{\boldsymbol{x}' \in \mathcal{X}} l_t(\boldsymbol{x}')\} \tag{11}$$

To find a set of expanders, we define

$$g_t(\boldsymbol{x}) = |\{\boldsymbol{x}' \in \mathcal{X} \setminus \mathcal{S}_t | l_{t,(\boldsymbol{x}, u_t(\boldsymbol{x}))}(\boldsymbol{x}') \geq J_{\min}\}| \tag{12}$$

where $l_{t,(\boldsymbol{x}, u_t(\boldsymbol{x}))}(\boldsymbol{x}')$ is the lower bound of $\boldsymbol{x}'$ based on past data and a predicted outcome for $\boldsymbol{x}$ which provides a new upper bound $u_t(\boldsymbol{x})$. The function is used to determine how many inputs are added to the safe set after choosing input $\boldsymbol{x}$ and observing the output it provides. This function is positive only if the new data point has a non-negligible chance to expand the safe set. The set of possible expanders is then defined as

$$\mathcal{G}_t = \{\boldsymbol{x} \in \mathcal{S}_t | g_t(\boldsymbol{x}) \geq 0\} \tag{13}$$

Normally, the safe optimization routine picks as the next point a safe point that is within the intersection of expanding and maximizing points, but currently shows the highest uncertainty measured by the difference between

the upper and the lower bound. However, for this first investigation of human behavior within safe exploration scenarios we will focus on how participants choices of input points are influenced by simple membership of the 3 sets and if their behavior can be described by more heuristic, stepwise decision behavior.

Experiment 1: Univariate functions

The first experiment required participants to maximize unknown univariate functions $f : x \to y$. On each trial $t = 1, \ldots, 10$ in a block, they could choose an input value $x \in \{0, 0.5, 1, \ldots, 10\}$ to observe (and accrue) an output $y = f(x) + \epsilon$ with noise term $\epsilon \sim \mathcal{N}(0, 1)$. The underlying functions were sampled from a $\mathcal{GP}$ with a squared exponential kernel (l=1, θ=1). The objective was to maximize the sum of the obtained outputs over all trials in a block. A threshold $J_{\min}$ was introduced and a block was ended abruptly if an output below this threshold was obtained. On average, that threshold was fixed to separate 50% of the points into safe and unsafe points. Before the first trial, an initial safe point above the threshold was provided. A screenshot is shown in Figure .

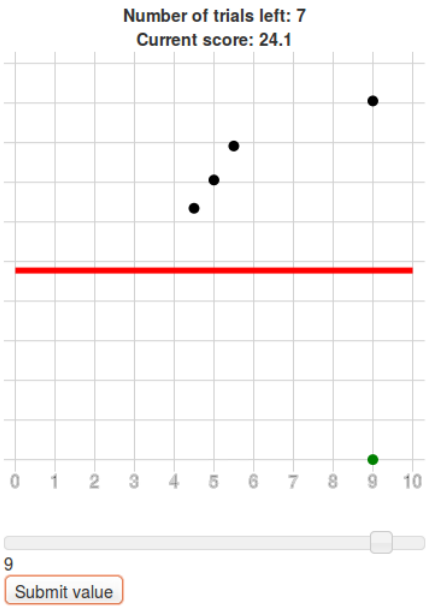


Figure 1: Screenshot of first experiment.

Participants

61 participants (36 female) with an average age of 32.95 (SD = 8.02) were recruited via Amazon Mechanical Turk and received $1 for their participation and a bonus of up to $1, in proportion to their overall score.

Procedure

Participants were told that they had to maximize an unknown function while at the same time trying to avoid sampling below the red line as this would end the current block. After reading the instructions and performing an example task, they had to correctly answer 4 questions to check their understanding, then performed the task, and at the end saw their total score.

Results

As shown in Figure 2, participants obtained outputs higher than expected by chance on the large majority of trials and indeed the average score per participant was significantly higher than chance, $t(60) = 13.311$, $p < 0.01$. In addition, the average number of trials per block

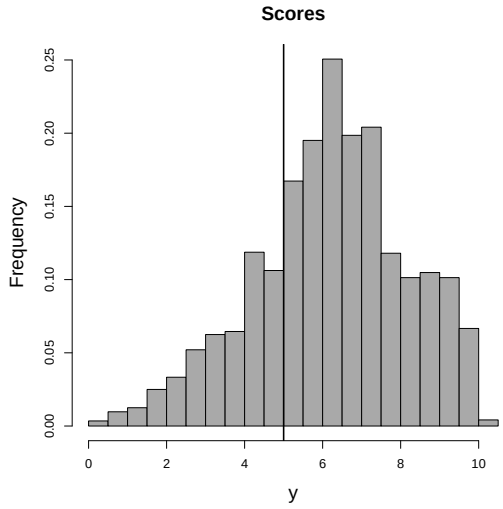


Figure 2: Scores per trial. Black line: chance level.

statistically exceeded what would be expected if participants chose completely at random, $t(548) = 5.1201$, $p < 0.01$ and participants' scores were positively correlated with trials ($r = 0.25$, $p < 0.01$). Taken together, these results indicate that participants learned the task and tended to chose safe inputs.

We used mixed-effects logistic regression analysis to asses which factors influenced participants' choices. The dependent variable was whether each input was chosen or not on each trial for each participant. As predictors, we used indicator variables for membership of an input of the safe, maximization, and expander set. Results indicated that the most plausible model was a model that contains all variables as fixed effects and a participant-specific random intercept, indicating that participants were influenced by set membership in an overall similar fashion. The coefficients of the fixed effects are presented in Table 1 below. Comparing the magnitude of the slopes of the predictors, we can conclude that participants cared about all of the sets, but mostly about

Table 1: Fixed effects estimate. Significant estimates are flagged.

Variable	b	$SE(b)$
Intercept	-4.26^*	0.04
Safe set	1.57^*	0.06
Maximizer set	1.72^*	0.05
Expander set	0.12	0.05

whether or not a point was safe and/or a maximizer. Next, we used a random intercept decision tree analysis (Sela & Simonoff, 2011) to assess whether participants might utilize a simple but effective heuristic strategy that can be implemented as a decision tree, as suggested by Huys et al. (2012). For this, we replaced the indicators of set membership with probability assessments, substituting membership of the maximizer set with the probability of improvement, the safe set with the probability of being above the threshold, and the expander set with the probability of safely expanding the set (assessed through one step ahead forward simulation). The probability of improvement is defined as the probability that an input $\boldsymbol{x}$ produces a higher output than the input $\boldsymbol{x}^+$ that is currently thought to provide the maximum, and can be calculated as

$$PI_t(\boldsymbol{x}) = P\left(f(\boldsymbol{x}) \geq f(\boldsymbol{x}^+)\right) \tag{14}$$

$$= \Phi\left(\frac{m_t(\boldsymbol{x}) - f(\boldsymbol{x}^+)}{\sigma_t(\boldsymbol{x})}\right) \tag{15}$$

where Φ is the cumulative Normal distribution function. Figure 3 depicts the decision tree which best fitted participants' choices. This analysis shows that participants seem to partition the problem into two sub-goals: first they conservatively assess whether or not a point is safe, then they maximize within that safe set. That expanders are not considered within this decision process could be due to the brevity of the task (10 trials) or to risk aversion (i.e., the fear of sampling below the threshold). The non-inclusion of possibly expanding points also

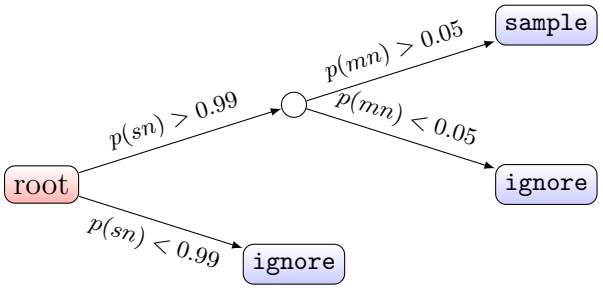


Figure 3: Multi-level decision tree minimizing log-loss.

means that participant only tried to maximize very locally, something that can also be seen when the distance of chosen points to the initially provided input point, $x_{\text{start}} - x_t$ is calculated as shown in Figure 4.

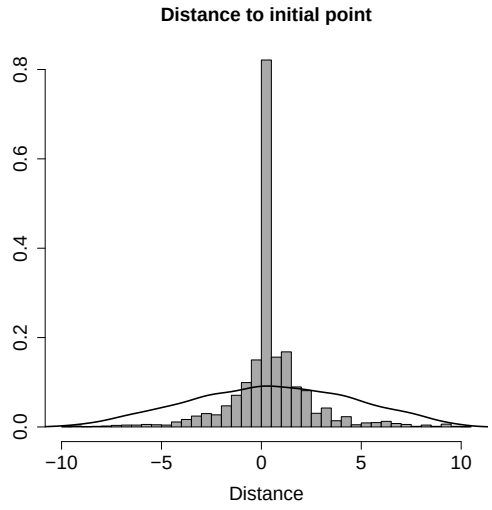


Figure 4: Distance of chosen input points to initially provided point. Black line indicates expected density for sampling at random.

This means that people's behavior in this uni-variate function optimization experiment was based on the attempt of locally maximizing points that they strongly perceived as safe.

Experiment 2: Bivariate functions

In the second experiment, participants were asked to maximize an unknown bivariate function $f : \boldsymbol{x} \rightarrow y$ with $\boldsymbol{x} = (x_1, x_2)^\top$, defined over the grid $x_1, x_2 \in [0, 0.05, 0.1, \ldots, 1]$, with $y = f(\boldsymbol{x}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$. As in Experiment 1, the function f was sampled on each block from a $\mathcal{GP}$ with a squared exponential kernel ($l = 2$,$\theta = 1$). The output values y varied between 0 and 100 and one initial point above 50 was provided. We varied the level of risk within-participants: there were 10 blocks in total out of which 5 were "normal", that is unconstrained maximization tasks without a threshold and 5 were "safe" blocks in which obtaining an output below 50 caused the current block to end abruptly. The blocks were presented in randomly permuted order. A screenshot is shown in Figure 5.

Participants

62 participants (37 male), with an average age of 31.77 years (SD = 8.97) were recruited via Amazon Mechanical Turk and received $1 for their participation and a performance-dependent bonus of up to $1. The average completion time of the whole experiment was 11 minutes.

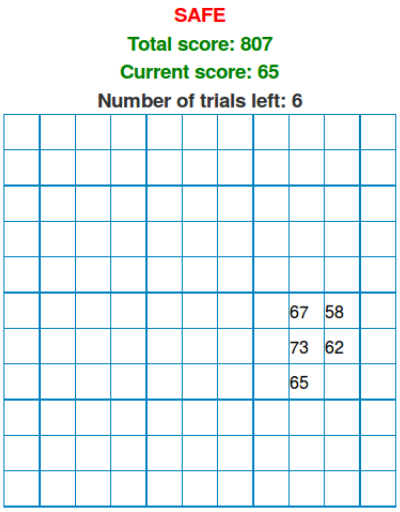


Figure 5: Screenshot of second experiment.

Procedure

The procedure was essentially the same as for Experiment 1, apart from additional detailed instructions regarding the difference between normal unconstrained (without a threshold) and safe (with a threshold) trials.

Results

As shown in Figure 6, participants scored better than expected by chance in both the safe and the normal conditions ($t(\text{normal} = 50) = 24.9$ with $p < 0.01$; $t(\text{safe} = 59) = 9.3$ with $p < 0.01$). The reason why chance level performance is higher in the safe condition is that scores below 50 were not allowed and therefore the output was truncated to be above 50. This time, participants within the safe condition did not complete more trials in a block than expected by randomly choosing inputs on the grid ($t(\text{length} = 5) = -0.32$ with $p = 0.72$). A similar mixed-effects logistic regression analysis as used for Experiment 1 (Table 2) showed that participants seemed to care most about scoring above the threshold in both conditions. As expected, this effect was more pronounced in the safe conditions than in the normal conditions. Still, the presence of this effect in the normal condition is interesting as it did not matter whether or not participants scored below the threshold. If scoring below 50 did not matter, participants should have not cared as much about sampling above this point in the normal condition as they actually did. One explanation for this might be a transfer effect by which participants assume that sampling below 50 is generally bad. The maximizer set only had a small influence on participants' choices that was slightly bigger for the safe condition. Participants showed no tendency to expand the safe set in either of the conditions. This indicates that most chosen inputs were close to the initial safe input and previously chosen inputs. This relatively high risk aversion is understandable, as the bivariate task is more difficult than the univariate one of Experiment 1. Lastly, a random intercept decision tree analysis (Figure 7) showed that in the best fitting model, only the probability of being above the threshold mattered. This indicates that participants only seemed to care about whether or not an input was safe, simplifying the task to a great extent with a strong focus on the probability of losing. Such simplification makes sense in light of the relative complexity of the bivariate task.

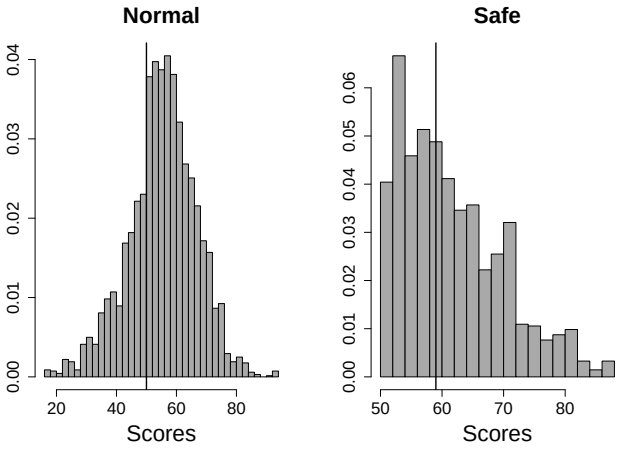


Figure 6: Scores per trial.

Table 2: Fixed effects estimates. Significant estimates are flagged.

Condition	Variable	b	SE(b)
Normal	Intercept	-5.17^*	0.04
	Safe Set	1.35^*	0.04
	Maximizer	0.13^*	0.04
	Expander	0.04	0.05
Safe	Intercept	-5.92^*	0.09
	Safe Set	2.11^*	0.07
	Maximizer	0.23^*	0.09
	Expander	0.03	0.08

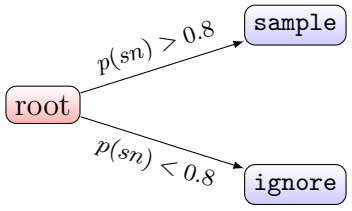


Figure 7: Decision tree minimizing log-loss.

This means that participants again only sampled locally, staying close to the initial point (Figure 8).

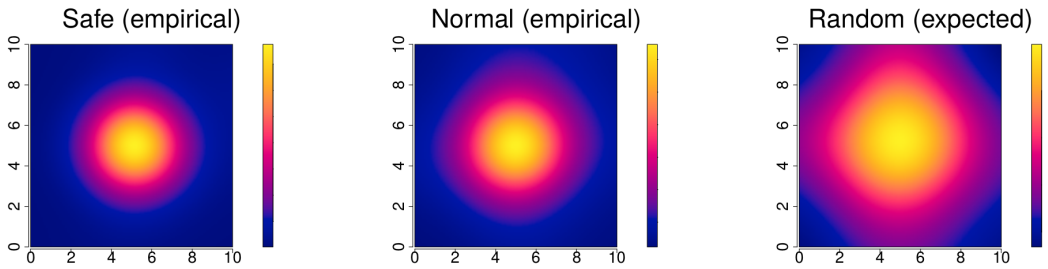


Figure 8: Distance of participants' sample points to initial point for safe condition, normal condition, and the theoretically expected distance for random sampling. Participants stay a little closer to the initial point in the safe condition than in the normal condition. Randomly sampling would theoretically cause much higher distances.

If participants sampled locations are treated as a Poisson process and centred around the initial point, then the posterior density of sampled points shows that participants stayed very close to the initial point in the safe condition, sampled a little further away from the initial point in the normal condition, but never sampled as dispersively as a completely random sampler.

Discussion and Conclusion

Learning unknown functions and exploiting this knowledge to maximise rewards are essential cognitive skills. Such tasks can be formalized as bandit tasks and here we focused on a restricted version thereof where outcomes below a given threshold need to be avoided. We found that participants' behavior was described well by a Gaussian Process safe optimization routine that establishes safe sets and then tries to maximize outputs within these sets. Participants mostly ignored input points that could expand the safe set, shunning risks and maximizing outputs locally, thereby preferring to rather be "safe than sorry".

Participants' behavior was consistent with a sequential heuristic in which they first determined whether inputs were safe and then maximized within this safe set. While this strategy involves only local searches, it can result in truly auspicious behavior, especially when the choices are limited. Participants' focus on avoiding unsafe inputs is consistent with a biological homeostasis maintenance principle that prioritizes not loosing everything over gaining as much as possible. The continued influence of the save threshold on participants' choices in the normal condition, where it had no effect on their potential earnings, might be due to participants generalizing their evaluation of outputs below the threshold as "bad" from the safe conditions.

In future work, we want to focus on what factors drive participants to switch from explorative to safe behavior and in which situations switching constitutes as a normative strategy, for example because it is minimizing costs (Bach, 2015). As we have only focused on functions sampled from a squared exponential kernel here, both for the description of participants' intuitive function learning process and for the actual functions sampled within the task, another direction is to assume different kernel parametrizations of these functions as those lead to diverse theoretical predictions about how fast participants are able to learn (Schulz et al., 2015c). Future work could also extend our approach to active versions of more traditional models such as heuristics and weight-based strategies (Parpart et al., 2015).

Unlike previous work on human behavior in the bandit setting, which has focused on pure optimization primarily, our work explored a relatively novel facet–optimizing risky functions. We expect that this new approach will provide further insights into how people resourcefully optimize outcomes in the real world.

References

Bach, D. R. (2015). Anxiety-like behavioural inhibition is normative under environmental threat-reward correlations. *PLOS Computational Biology*, *11*, e1004646.

Berkenkamp, F., Schoellig, A. P., & Krause, A. (2015). Safe controller optimization for quadrotors with Gaussian Processes. *arXiv preprint arXiv:1509.01066*.

Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian Processes. In *Advances in Neural Information Processing Systems*, (pp. 553–560).

Huys, Q., Eshel, N., ONions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, *8*, e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*, 3098–3103.

Korn, C. W., & Bach, D. R. (2015). Maintaining homeostasis by decision-making. *PLOS Computational Biology*, *11(5):e1004301*.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2015). Active learning as a means to distinguish among prominent decision strategies. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*, (pp. 1829–1834).

Rasmussen, C. E. (2006). Gaussian Processes for machine learning.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015a). Exploration-exploitation in a contextual multi-armed Bandit Task. *Proceedings of the 13th International Conference on Cognitive Modeling. Groningen, NL*.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015b). Learning and decisions in contextual multi-armed bandit tasks. *Proceedings of the 37th annual conference of the cognitive science society*, (pp. 2122–2127).

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. J. (2015c). Assessing the perceived predictability of functions. *Proceedings of the 37th annual conference of the cognitive science society*, (pp. 2116–2121).

Sela, R., & Simonoff, J. (2011). Reemtree: Regression trees with random effects. *R package version 0.90*, *3*, 741–749.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009). Gaussian Process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Sui, Y., Gotovos, A., Burdick, J. W., & Krause, A. (2015). Safe exploration for optimization with Gaussian Processes. In *International Conference on Machine Learning (ICML)*.

Intentionality and the Role of Labels in Categorization

Felix Gervits (Felix.Gervits@tufts.edu)
Tufts University, Human-Robot Interaction Laboratory, 200 Boston Avenue
Medford, MA 02155

Megan Johanson (megan.johanson@gmail.com)
University of Delaware, Department of Psychological and Brain Sciences, 401 Wolf Hall
Newark, DE 19716

Anna Papafragou (APapafragou@psych.udel.edu)
University of Delaware, Department of Psychological and Brain Sciences, 401 Wolf Hall
Newark, DE 19716

Abstract

Language has been shown to influence the ability to form categories. Here we investigate whether linguistic labels are privileged compared to other types of cues (e.g., numbers or symbols), and whether labels exert their effects regardless of whether they are introduced intentionally. In a categorization task, we found that adults were more likely to use labels to determine category boundaries compared to numbers or symbols, and that these effects persisted in all intentionality manipulations. These findings suggest that labels have a powerful effect on categorization compared to other cues; most strikingly, labels (but not other cues) are used during categorization even when people are specifically asked to ignore them. These results provide novel support for the position that labels indicate category membership.

Keywords: categorization; labels; intentionality, category markers

Introduction

The formation of categories is an essential aspect of cognition found to some degree in all animals. Humans, however, are the only species known to incorporate language into their representational categories. Categorization can be defined as the grouping of discriminable properties, objects, or events into classes. It is generally recognized to be a complex process relying on a variety of cues: including perceptual (Oakes & Rakison, 2003), conceptual (Booth & Waxman, 2002) and linguistic (Casasola & Bhagwat, 2007). Though categories can be formed without linguistic labels, it is well-known that the presence of labels facilitates categorization by drawing attention to shared features, relations, or actions (Waxman & Markow, 1995). For instance, when two objects are given the same label, people tend to think of them as more similar to each other compared to when objects do not share a label (Goldstone, 1994). Moreover, people rely more on labels than on perceptual features in making category decisions when the two cues conflict (Deng & Sloutsky, 2012; Sloutsky, Lo, & Fisher, 2001). In cases where perceptual information does not give adequate cues for category formation, a label helps group objects into broad categories based on perceptual features (Johanson & Papafragou, 2016). Even completely redundant labels have been shown to influence the way people learn novel categories (Fisher, 2010; Hoffman & Rehder, 2010; Lupyan, Rakison, & McClelland, 2007).

The studies described above show a powerful effect of labels on the categorization process, but it is not clear whether similar effects will occur in the presence of non-linguistic cues. Some perspective on this issue is provided by Lupyan (2006), who demonstrated that the labeling advantage normally attributed to words can apply to other highly correlated environmental cues as well, and that words are not the only cues that can serve as category markers. Others, however, find support for the idea that labels are indeed privileged as categorization cues. For example, Lupyan et al. (2007) found that adults learned to categorize "aliens" faster after learning novel labels for them compared to seeing behaviors that indicated where the aliens lived. In another study, Lupyan and Thompson-Schill (2012) found that verbal labels (e.g., "dog") activate visual information more quickly and reliably than nonverbal sounds (e.g., a dog barking). Together these data support the view that labels serve as category markers (Gelman & Davidson, 2013). On this position, labels function as an invitation to form categories (Waxman & Markow, 1995): "[e]xactly what makes a dog a dog, or a lamb a lamb, may be unknown [...], but a category label can serve as a placeholder that a reason exists" (Jaswal & Markman, 2007).

Though it has been widely assumed that adults understand the linguistic significance of labels (Markman & Ross, 2003; Yamauchi & Markman, 2000), some argue that labels do not serve to mark categories, but rather are interpreted as highly salient object features (Anderson, 1991; Gliozzi, Mayor, Hu, & Plunkett, 2009; Perfors & Navarro, 2010). In support of this view, Deng and Sloutsky (2012) found that not all adults displayed label-consistent responding when a novel label was pitted against a highly salient feature (moving head) in an induction task. In a categorization task, Perfors and Navarro (2010) found that labels directed attention away from other object features, suggesting that in this way they behave more like salient features rather than cues to category membership. These results suggest that perhaps labels are not always treated as category markers for adults.

At present, however, several issues remain about the potency of labels as categorization cues. Notice that, in virtually all previous studies that have pointed to a strong role for language in categorization, linguistic labels were intro-

duced intentionally by the experimenter. Specifically, labels were almost always presented intentionally alongside a newly-introduced exemplar of a category (e.g., "This is an X") and were explicitly linked to a kind, not to the individual exemplar itself (Deng & Sloutsky, 2012; Lupyan et al., 2007; Yamauchi & Markman, 2000). Since in these studies labels were presented intentionally as part of the task instructions, it is perhaps not surprising that adults took the labels to be relevant to the task. It is possible that participants in a similarity judgment task may develop a strategic bias to respond to the expectation that items that have just been labeled alike should belong to the same category without actually integrating the information from labels with perceptual cues and other information about the categorical structure of the stimuli (Goldstone, Lippa, & Shiffrin, 2001). A more stringent test of the theoretical position that labels are interpreted as category markers would involve dissociating the information carried by the labels from the intentional act of labeling the stimuli during the categorization task. Furthermore, it is important to compare labels to other symbolic (but non-linguistic) cues under the same intentionality conditions to ascertain that labels produce unique and powerful effects on categorization.

To our knowledge, there have not been any studies that addressed this problem in adult categorization. A relevant study from the child literature showed that 4-year-olds were influenced by the communicative intent of an adult speaker in making categorical inferences (Jaswal & Markman, 2007). For example, children were willing to accept that a cat-like animal was a dog only if the speaker made this explicit (e.g., "This is actually a dog"). This suggests that intentionality of verbal cues can have a powerful influence on the formation of categories. Other work has shown that young children learned to associate novel words and even nonverbal sounds to objects, but only when these stimuli were presented in a referential context (Campbell & Namy, 2003; Tomasello & Barton, 1994). These findings are important because they disentangle intentionality from the cues, showing that the manner in which the cues are presented is as important as the linguistic (or non-linguistic) nature of the cue.

Present Study

In this paper, we present the first study to examine the role of intentionality on adults' reliance on labels (and other cues) during categorization. We presented adults with a task in which they had to form categories for novel natural kind exemplars (e.g., flowers or birds) that were perceptually equidistant (perceptual information gave no indication of category membership) from two Standards. We asked whether a shared novel label (noun) might motivate adults to group such perceptually ambiguous stimuli with one of the two Standards. We also compared the usefulness of labels to other cues such as numbers and symbols that do not refer to object kinds. We chose numbers because, unlike nonsense labels, they are familiar and meaningful - yet numbers may lack the referential capacity that labels possess. We chose arbitrary symbols (which were simply the number cues rotated 90 degrees horizontally) because, unlike labels and numbers, they have little or no meaning (even though in terms of low-level perceptual features they were identical to the numbers). A crucial difference from previous studies was that the cues were presented in written, not auditory, format and were never explicitly introduced by the experimenter (e.g., labels never appeared in sentences such as "This is an X"). This design allowed us to manipulate global information about the intentional relevance of the cues given to participants in the beginning of the experiment, such that cues were presented as intentional, accidental or neutral (unspecified).

The present design also allowed us to address specific - but currently untested - hypotheses about the role of language in adult categorization, especially in the absence of convergent perceptual cues about the structure of a category. If human beings treat labels as category markers, adults' use of labels should be sensitive to the intentionality (and hence task-relevance) of the naming contexts; there is evidence that, at least in children, this is the case (Campbell & Namy, 2003). Cues such as numbers or symbols might also function as an invitation to form categories when intentionally introduced alongside novel exemplars of a category. Crucially, however, adults' reliance on such cues should be limited compared to labels across intentionality manipulations: unlike labels, numbers and symbols cannot be referentially linked to the natural kind categories in the experimental stimuli. The potency of numbers or symbols should diminish or disappear altogether when these cues are not accompanied by clear communicative intent.

Methods

Participants

A total of 180 English-speaking adults were used as participants in the study. They were recruited from Introductory Psychology courses at the University of Delaware and were given partial course credit for participating.

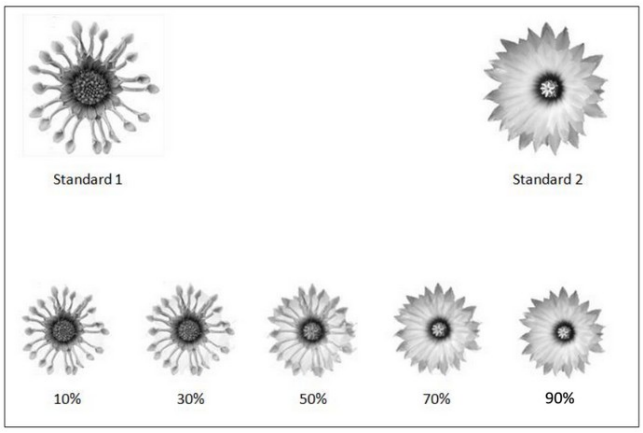


Figure 1: Sample set of Standards and Target

Stimuli

Four sets of black-and-white stimuli were created: one set of birds, two different sets of flowers, and one set of fish. Each of these sets consisted of photographs of two stimuli (Standards) that were morphed through a commercial morphing program into 5 different stimuli (Targets) along a scale of varying similarity to the originals (Johanson & Papafragou, 2016). Each Target was 10%, 30%, 50%, 70% or 90% similar to one of the Standards (see Figure 1). The 10% and 30% Targets were always perceptually closer to Standard 1, whereas the 70% and 90% Targets were always perceptually closer to Standard 2. The 50% Target was perceptually ambiguous, i.e., equidistant from the Standards. To confirm these rankings, a separate group of 13 adults was presented with triads consisting of the two Standards and each of the Targets and was asked which of the Standards each Target went with. People in this group were highly accurate with the 10%, 30%, 70% and 90% (unambiguous) trials (M = .96) but were at chance with the 50% (ambiguous) trials (M = .58).

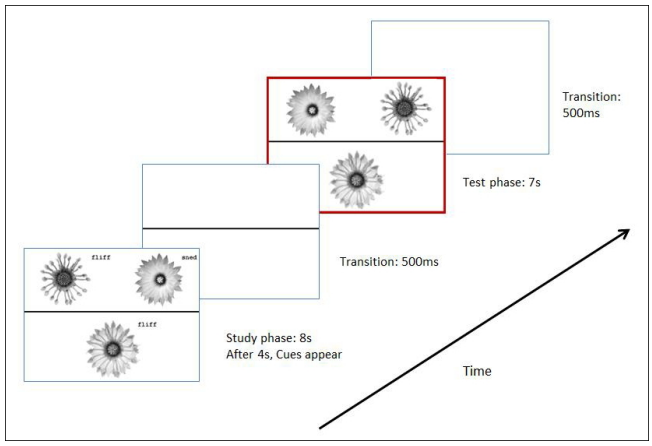


Figure 2: Sample ambiguous trial

Within each of the four stimuli sets, there were five trials for a total of 20 trials. Each trial included a triad display, with the two Standards on top, and a morphed Target image (the 10%, 30%, 50%, 70%, or 90% Target) at the bottom, separated from the standards by a solid line. This triad display was presented for 8 seconds (Study phase). Then the objects disappeared for 500ms but the solid line remained. The objects reappeared with a red border surrounding the display and stayed on the screen for 7 seconds (Test phase). All five trials within a set were presented in block sequence. The fourth trial in each set was always the ambiguous (50%) trial, but the order of the unambiguous trials varied randomly for each set. Blank screens marked transitions between trials within a set (500ms) and between sets (2s). For half of the participants, the left-right position of the Standards was switched in the Test phase so as to prevent any side associations.

Within each set, visual cues appeared 4 seconds (halfway) into the Study phase for the ambiguous trial and for one additional, randomly selected trial that came before the ambiguous one (Cue Trials). Three cues appeared simultaneously within each display, with each one being matched to one of the natural-kind exemplars (Standards and Target) in a triad. Each cue would appear in the top-right corner of each exemplar, blink twice and then stay on for another 2s until the end of the Study phase for that trial. The cue for the Target was identical to the cue for one of the Standards. In the ambiguous Cue trials, the cue for the Target arbitrarily matched that of either the left or right Standard. In the unambiguous Cue Trials, the Target cue always respected the perceptual similarity between Target and Standards. The assignment of cues to Standards and Target was counterbalanced across participants. See Figure 2 for a sample trial progression.

Table 1: Sample cues

Cue	Example
Label	lorp pim
Number	3 6
Symbol	ш ⅂

There were three types of Cues: Labels, Numbers, and Symbols. Participants were randomly assigned to one of three groups depending on Cue type. The labels were "lorp" and "pim" (bird set), "fliff" and "sned" (1st flower set), "blick" and "dax" (2nd flower set), "hep" and "moof" (fish set). The numbers were 6 and 3, 1 and 5, 2 and 7, 8 and 4 for the corresponding sets, presented in a distinctive font. The symbols were the number cues rotated 90 degrees clockwise and hence were identical to numbers in terms of low-level visual properties (Lupyan & Spivey, 2008). See Table 1 for examples of the cues.

Procedure

Participants were tested in small groups in front of a projector screen which displayed the stimuli. Within each Cue group, participants were randomly assigned to one of three Conditions (Neutral, Accidental, Intentional) depending on the instructions given to them in the beginning of the session. Instructions were read aloud to all participants. In the Neutral condition the instructions were as follows: "In this experiment you will be presented with a series of slides, each containing three images. Your task is to determine as best you can which of the two top images the bottom image best goes with. Each slide will appear twice. The first time, you will have 8 seconds to inspect the slide. The slide will briefly disappear and reappear again with a red border around it. Your task then will be to mark down on your answer sheet which of the top two images the bottom image on the slide best goes with. Mark L if the bottom image goes with the top left image or mark R if the bottom image goes with the top right image. Pay attention as the position of the top two images may have

been switched around between the first and second time you see the slide. Please only write your answers when you see the red border around the slide". In the other two conditions, there was additional information at the end of these instructions. People in the Accidental condition were casually told: "We've been having issues with our software this week so occasionally you might see random messages displayed on the screen. These are actually a glitch from another experiment so please disregard them as they are irrelevant to your task". People in the Intentional condition were told: "Pay attention to all information on the screen as it will be helpful in your task". Participants marked their responses sequentially on an answer sheet.

Results

We first examined the percentage of correct responses for unambiguous trials (10%, 30%, 70% and 90%). On these trials, correct responses were those that conformed to perceptual similarity. As expected, performance was highly accurate on the three unambiguous trials per set that were presented without cues in all Cue (M_L = .99, M_N = .98, M_S = .97) and Condition subgroups of participants (M_N = .99, M_A = .98, M_I = .98). Similarly, for the one unambiguous trial per set that were also presented alongside cues, performance was at ceiling across Cue (M_L = .99, M_N = .99, M_S = .97) and Condition sub-groups of participants (M_N = 1.00, M_A = .99, M_I = .97). Thus participants were successful in categorizing the unambiguous stimuli with or without additional cues.

We then looked at ambiguous (50%) trials (see Figure 3). A two-way ANOVA was carried out using the percentage of cue-based responses on these trials as the dependent variable and Cue (Label, Number, Symbol) and Condition (Neutral, Accidental, Intentional) as between-subjects factors. The ANOVA revealed a significant main effect of Cue, $F(2, 171) = 14.898$, $p < .001$; M_L = .80, M_N = .65, M_S = .58. Planned comparisons with Bonferroni corrections showed that Labels led to better categorization performance compared to Numbers ($p < .005$) or Symbols ($p < .001$) but there was no significant difference between Numbers and Symbols ($p = .265$).

The analysis also returned a significant main effect of Condition, $F(2, 171) = 13.779$, $p < .001$; M_I = .79, M_N = .68, M_A = .58. Planned comparisons with Bonferroni corrections showed that performance in the Intentional condition was significantly higher than either the Accidental ($p < .001$) or the Neutral condition ($p < .05$). Moreover, performance in the Neutral condition was higher compared to the Accidental condition ($p = .05$). Interaction effects between Cue and Condition were not statistically significant, $F(4, 171) = .922$, $p = .452$.

As Figure 3 shows, performance with Labels was significantly different from chance in all conditions (Neutral: M_L = .84, $t(19) = 9.0$, $p < .001$; Accidental: M_L = .71, $t(19) = 4.07$, $p < .005$; Intentional: M_L = .86, $t(19) = 10.72$, $p < .001$). By contrast, performance with Numbers and Symbols was at chance-level in both the Neutral (M_N = .60, $t(19) = 1.90$, $p = .072$; M_S = .59, $t(19) = 1.38$, $p = .185$) and the Accidental condition (M_N = .55, $t(19) = .89$, $p = .385$; M_S = .46, $t(19) = -.62$, $p = .545$), and differed from chance only in the Intentional condition (M_N = .81, $t(19) = 7.11$, $p < .001$; M_S = .70, $t(19) = 4.29$, $p < .001$). In sum, all cues were used at above-chance levels when presented intentionally. However, labels were the only cue to be utilized at above-chance levels in all intentionality conditions.

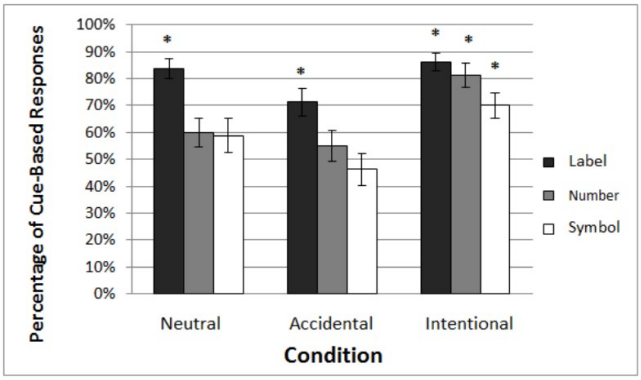


Figure 3: Percentage of cue-based responses for ambiguous trials. Error bars represent standard error of the mean. * reflect significant difference from chance (50 %) performance ($p < .05$)

General Discussion

Summary of Results

In most, if not all, prior demonstrations of the role of labels on adult categorization, the contribution of the representational content of the labels cannot be easily distinguished from the experimenters' intention to direct attention to the labels (and possibly their task relevance) within the categorization task. For the same reason, the potency of labels in adults remains to be compared to that of other, intentionally communicated cues. The present study was the first to compare the role of different levels of intentionality and different cue types (labels, numbers, symbols) in adult categorization.

Overall, we found that labels were particularly potent in influencing category decisions about perceptually indeterminate exemplars of novel natural kinds when compared to other cues such as numbers and symbols. Furthermore, for all cues, intentionality affected peoples' willingness to treat the cues as task-relevant. A particularly interesting finding was that the role of labels persisted throughout various levels of intentionality. People used novel labels to constrain their categorization decisions even when the labels were not tied to a specific communicative intent (Neutral condition). Perhaps most strikingly, people used novel labels even when they were instructed to ignore them (Accidental condition). By contrast, numbers and symbols were used only when their relevance to the task was explicitly highlighted (Intentional condition). These findings break new ground by providing novel support for the position that labels indicate category membership.

Labels as Category Markers

Clarifying the source and boundaries of the effects of labels on category formation is theoretically important, especially for the widely held perspective that labels are inherently category markers and have a privileged connection to kinds because of their representational, or semantic, content (Balaban & Waxman, 1997; Gelman & Davidson, 2013; Gelman & Markman, 1986). This position can naturally accommodate the fact that the use of labels during categorization should depend on the nature of the intentional context in which the labels are presented, and that other cues might shape categorization of novel kinds if accompanied by strong intentional indications. Crucially, however, this position expects that labels (because of their representational nature) should shape category formation even if the experimenter does not direct participants' attention to the labels and/or their potential significance within the task. Furthermore, this view predicts that, across contexts that differ in intentionality, adults should assume that labels are uniquely helpful for categorization compared to other symbolic cues that do not have a privileged representational connection to kinds. All of these predictions of the position that labels are category markers were confirmed in our data.

Taken together, our findings show that the effect of labels during categorization cannot simply be attributed to broad effects of intentional communication that could be achieved with any symbolic representation. Even though communication in general carries a presumption of relevance and is therefore expected to yield "cognitive effects" for the hearer (Sperber & Wilson, 1986), linguistic labels create the expectation of a specific type of cognitive effect - here, kind reference - that is not readily shared by other symbols. By the same token, our findings show that adults do not approach the categorization task through a strategy to use whatever cue they are presented with to base their category decisions (Goldstone et al., 2001). For similar reasons, our results cannot be explained solely by the idea that labels function by directing peoples' attention towards certain kinds of groupings of exemplars over others and encouraging the extraction of similarities (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002), since in principle all cues in our studies could function in that way. The specific and unique effects of labels suggest that adults use sophisticated and rich conceptual reasoning to determine which kinds of information are relevant to kind membership and under what circumstances - exactly as predicted by the position that labels are privileged types of invitations to form categories.

The fact that a newly-presented label was used to guide the categorization of unfamiliar stimuli even when people were given reason to treat the label as accidental might appear surprising, even unexpected, under the theoretical position that humans treat labels as category indicators. We believe that the reason participants did not completely disregard accidental labels was that, despite global information about the lack of intentionality of the experimental cues, the labels appeared to have properties compatible with intentional status: most importantly, they accompanied stimuli that could still be taken as their referents (since there was no mismatch between the labels and the visual evidence for the structure of the novel categories).

This interpretation of our findings is reminiscent of evidence of sensitivity to speaker intent in the developmental literature on categorization which suggests that when the intentional link between a label and a category is drastically severed (e.g., when the source of the label is non-human), even very young learners recognize that labels should not be linked to categories. In one demonstration, 15-month-old infants used labels to form categories only when the labels were presented orally but not when they were presented by a voice recorder (Fulkerson & Haaf, 2003). Similarly, 13- and 18-month-old infants learned to associate novel words and even nonverbal sounds to objects when the objects were embedded in a referential context (e.g., named by the experimenter), but not when they were embedded in a non-referential context (e.g., named through a baby monitor) (Campbell & Namy, 2003).

One alternative interpretation of our results is the possibility that labels were used to such a high degree because they were taken to denote the natural kind depicted by the stimulus. Notice that, unlike past experiments, labels in the present study were presented visually, were not introduced intentionally (Accidental condition), and were not even identified as linguistic stimuli. As a result, we cannot be sure that participants interpreted the novel letter strings we presented them with as labels. Thus it remains possible that the special status of labels could be due to alternative factors that do not have to do with their representational potential (but rather with, e.g., the ease of associating visual stimuli of flowers or birds to word-like forms compared to numbers or symbols, or the greater salience of labels compared to the other cues). Investigating these factors should provide a promising direction for future studies. One other direction for future work is to better disentangle the effects of attention from intentionality. For example, it is possible that people in the Accidental and non-label conditions of our study simply did not pay attention to the cues. However, if attentional mechanisms are responsible for the effects we observed, then it is unclear as to why nonsense labels, but not other cues, were consistently attended. Future studies will need to examine different types of label cues to see if these cues are similarly attended to, and affected by intentionality.

Conclusion

In support of previous findings, we demonstrated that labels have a privileged role in categorization and are used to group objects into broad categories. The present study also included novel manipulations of intentionality that were purposefully removed from actual interactions with an interlocutor. Nevertheless, we found that when participants' attention was directed towards a symbolic cue (a number or an arbitrary sym-

bol) through global linguistic information about the experimenter's intent, participants treated these cues as guides to kind membership at rates above chance; in the absence of intentional information, these cues were not reliably used. Labels were used during categorization in all intentionality manipulations, and even when people were specifically asked to ignore them. These findings lend support to the position that labels serve as category markers for adults, and also confirm the strong role of intentionality (especially in communicative contexts) for learning novel information.

Acknowledgments

This work was partly supported by NSF grant BCS-0641105 to A.P. that supported the graduate research performed by F.G.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child psychology*, *64*(1), 3–26.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948–957.

Campbell, A. L., & Namy, L. L. (2003). The role of social-referential context in verbal and nonverbal symbol learning. *Child Development*, *74*(2), 549–563.

Casasola, M., & Bhagwat, J. (2007). Do Novel Words Facilitate 18-Month-Olds? Spatial Categorization? *Child Development*, *78*(6), 1818–1829.

Deng, W., & Sloutsky, V. (2012). Carrot-eaters and moving heads: Salient features provide greater support for inductive inference than category labels. *Psychological Science*, *23*, 178–186.

Fisher, A. V. (2010). Whats in the name? or how rocks and stones are different from bunnies and rabbits. *Journal of Experimental Child Psychology*, *105*(3), 198–212.

Fulkerson, A. L., & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9-and 15-month-olds' object categorization. *Infancy*, *4*(3), 349–369.

Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology*, *66*(3), 327–353.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319.

Jaswal, V. K., & Markman, E. M. (2007). Looks aren't everything: 24-month-olds' willingness to accept unexpected labels. *Journal of Cognition and Development*, *8*(1), 93–111.

Johanson, M., & Papafragou, A. (2016). The influence of labels and facts on children's and adults' categorization. *Journal of Experimental Child Psychology*, *144*, 130–151.

Lupyan, G. (2006). Labels facilitate learning of novel categories. In *The sixth international conference on the evolution of language* (pp. 190–197).

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.

Lupyan, G., & Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Current Biology*, *18*(10), R410–R412.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–86.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological bulletin*, *129*(4), 592–613.

Oakes, L., & Rakison, D. (2003). Issues in the early development of concepts and categories: An introduction. *Early category and concept development: Making sense of the blooming, buzzing confusion*, 3–23.

Perfors, A., & Navarro, D. J. (2010). How does the presence of a label affect attention to other features? In *Proceedings of the 32nd annual conference of the cognitive science society.*

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695–1709.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Tomasello, M., & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental psychology*, *30*(5), 639.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776.

Constraining the Search Space in Cross-Situational Word Learning: Different Models Make Different Predictions

Giovanni Cassani, Robert Grimm, Steven Gillis, and **Walter Daelemans**
Computational Linguistics and Psycholinguistics (CLiPS) Research Center
Department of Linguistics, University of Antwerp, 13 Prinsstraat
B-2000 Antwerpen, Belgium
{name.surname}@uantwerpen.be

Abstract

We test the predictions of different computational models of cross-situational word learning that have been proposed in the literature by comparing their behavior to that of young children and adults in the word learning task conducted by Ramscar, Dye, and Klein (2013). Our experimental results show that a Hebbian learner and a model that relies on hypothesis testing fail to account for the behavioral data obtained from both populations. Ruling out such accounts might help reducing the search space and better focus on the most relevant aspects of the problem, in order to disentangle the mechanisms used during language acquisition to map words and referents in a highly noisy environment.

Keywords: cross-situational learning; word learning; computational modeling; language acquisition

Ever since the *gavagai* example provided by Quine (1960) to describe the huge amount of referential uncertainty that any language learner has to face while inducing word-object mappings, researchers took an interest in which mechanisms can be exploited to solve this crucial task. In the last twenty years, computational modeling has proven extremely useful in exploring what information encoded in the input children receive might allow them to correctly map referents and words, and which learning mechanisms might best exploit the relevant information (Frank, Goodman, & Tenenbaum, 2009).

Cross-situational learning posits that children keep track of co-occurrences of referents in the world and words uttered to them in several situations to establish unambiguous mappings: while the single situation might be ambiguous, the co-occurrences of words and referents across many different situations help the learner figure out the correct mappings. Starting from the work of Siskind (1996), many different learning mechanisms that exploit this basic principle have been proposed that show comparable performances to many behavioral data from both children (Ramscar, Dye, & Klein, 2013; Smith & Yu, 2008; Suanda, Mugwanya, & Namy, 2014) and adults (Dautriche & Chemla, 2014; Fazly, Alishahi, & Stevenson, 2010; Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013; Yu & Smith, 2007; Yurovsky & Frank, 2015; Yurovsky, Yu, & Smith, 2013), using both corpus studies and laboratory experiments, covering many different conditions. Differences and similarities across models have been explored, with the main goal of showing how apparently different proposals can yield comparable results and make similar predictions when certain components of the learning algorithm are modified (Yu & Smith, 2012; Kachergis, Yu, & Shiffrin, 2016).

In this paper, we compare behavioral evidence to four different models that exploit cross-situational regularities to infer word-referent mappings from the data, to analyze what predictions each model makes and whether they fit with what children and adults do when asked to map a referent to a word. Our aim is to provide evidence about which learning mechanisms proposed in the literature can explain behavioral evidence and which cannot, in order to constrain the search space of possible models to the learning strategies that exploit cross-situational information in the same way humans do. Carefully controlled laboratory settings in which specific features of the word learning task are manipulated can help to achieve this goal, by isolating the information from the input that makes learning possible or impossible and providing valuable data to test computational simulations in a variety of situations (Ramscar, Dye, & Klein, 2013; Kachergis et al., 2016).

While many learning mechanisms can mirror certain behavioral patterns (Yu & Smith, 2012), some may not be able to learn the correct word-referent mappings in specific, controlled paradigms in which subjects do learn such mappings robustly. Identifying these situations and showing why certain mechanisms fail to account for successful learning will help the researchers to constrain the hypothesis space and discard mechanisms that make incorrect predictions.

Dataset

In order to evaluate the predictions of different models of cross-situational learning we make use of the evidence presented in Ramscar, Dye, and Klein (2013). The experiment they reported was conducted with a group of children (mean age 28 months) and two groups of adults, undergraduates and developmental psychologists.

The setting included three objects, [ObjA, ObjB, ObjC], and three labels, $\{Dax, Wug, Pid\}$; during 18 learning trials, each subject saw two objects and then heard one label. Of the three objects, ObjA and ObjC were presented 9 times, never together; ObjB, however, was present in all trials, occurring half of the times with ObjA and half of the times with ObjC. Crucially, ObjA was always presented together with the same label, e.g. *Dax*, and ObjC was always presented with the same label, e.g. *Pid*. Consequently, ObjB occurred half of the times with the label *Dax* and half of the times with the label *Pid*. The third label, *Wug*, never occurred during training.

During testing, the subjects heard one of the three labels

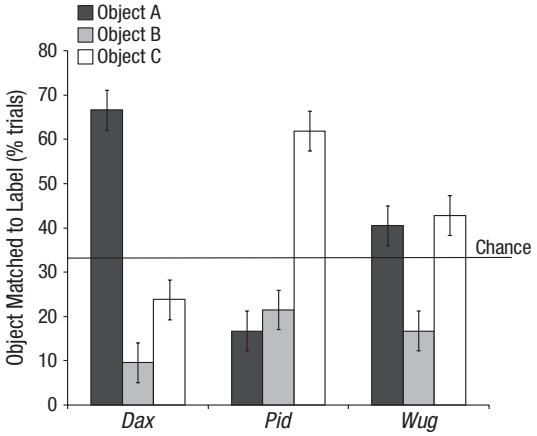


Figure 1: Children learning patterns on the word-learning experiment from Ramscar, Dye, and Klein (2013). The plot is taken from the original paper (Figure 2a).

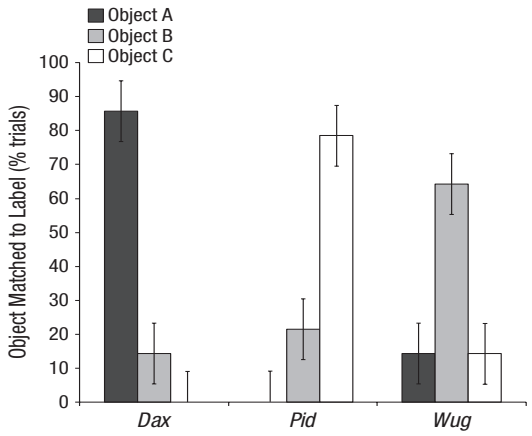


Figure 2: Undergraduates learning patterns on the word-learning experiment from Ramscar, Dye, and Klein (2013). The plot is taken from the original paper (Figure 3a).

and were asked to point to the object to which they thought the label referred to. Two labels occurred during training, one did not - words and objects were counterbalanced and learning trials were randomized across participants.

Behavioral Evidence

Results of the experiment are provided in Figure 1 for children and Figure 2 for undergraduates - the plots show the case in which *Dax* was always presented with ObjA, *Pid* with ObjC, and *Wug* was only showed during testing.

Both groups mapped ObjA and ObjC to the labels that only occurred with each of them. Interestingly, however, undergraduates showed a mutual exclusivity bias and mapped ObjB to *Wug*, which was not presented during training; on the contrary, children picked ObjA and ObjC at comparable rates as referent for the new label. The developmental psychologists were asked to predict the behavior of children but ended up predicting that of undergraduates. The authors of the study conclude that children are more sensitive to the informativity of cues than to logical principles, which on the contrary play a role in adults.

Feature-Label-Order Effects In this experiment, and in many others that address cross-situational word learning, objects are presented before their labels are uttered. Far from being irrelevant to the task, evidence from Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) shows that different learning outcomes arise in behavioral experiments where this order is manipulated. This difference is unfortunately not always considered in cross-situational learning studies: as a consequence, certain models are defined as mapping referents to words and others do the opposite. Moreover, the behavioral data we use were obtained using a paradigm in which the subjects first saw an object and then heard a label. Thus, considering the experimental paradigm and the importance of the order of presentation for learning to take place, when we evaluate a model that was designed to map words to referents, we switch the two layers and make it learn the opposite mapping.

Models of Cross-situational Learning

We compare simple, basic implementations[1] of four different learning mechanisms to highlight what predictions are made by each of them, and whether they match behavioral evidence. We introduce each model separately and briefly discuss its main features; for more detailed explanations, we refer to the cited publications.

Hebbian Learner This model implements the law of contiguity (Warren, 1921), according to which the association between two items becomes stronger when they consistently occur together in the environment. It is usually implemented as a neural network with no hidden layer that incrementally establishes associations between an input and an output layer (Hebb, 1949). An input-to-output association is strengthened by a constant quantity whenever the two co-occur within a learning trial. Associations from inputs that occur in a learning trial and outputs that do not are left unchanged, as are associations from absent inputs to all output units.

The way associations are updated is summarized in equation (1), where t represents a learning trial, c_i indicates an input item, or cue, o_j indicates an output, or outcome, and ΔV_{ij} indicates the value of the update from c_i to o_j after experiencing the learning trial t:

$$\Delta V_{ij} = \begin{cases} k & \text{if } c_i \in t \text{ and } o_j \in t \\ 0 & \text{else} \end{cases} \quad (1)$$

[1]The code of our own re-implementations of each model is available at `https://github.com/GiovanniCassani/cross_situational_learning`, commit n. 2a9dbaa

ΔV_{ij} is then added to the current association from c_i to o_j; k is a strictly positive constant which only affects the absolute value of the associations but not the relations among them, thus changing its value does not affect the learning outcome. This model was showed to successfully model behavioral data in the study by Yu and Smith (2012) and for this reason it is evaluated here. However, the risk exists that every input becomes associated with every output, making it impossible to learn unambiguous input-output mappings (Dawson, 2008).

Naïve Discriminative Learning (NDL) In this model, input-output associations are updated according to the Rescorla-Wagner equations (Rescorla & Wagner, 1972), developed in the context of animal learning and conditioning. This model is often referred to as Naïve Discriminative Learning (NDL, (Baayen, Hendrix, & Ramscar, 2013)) and its relevance to language has been established in different aspects of language learning and processing (Baayen, Milin, Durdević, Hendrix, & Marelli, 2011; Baayen, Shaoul, Willits, & Ramscar, 2015; Ramscar, Dye, & McCauley, 2013; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014).

Its architecture closely resembles the Hebbian learner, as it is a neural network with no hidden layer that incrementally establishes associations between cues and outcomes, where the first constitute the input layer and the latter the output nodes. As for the Hebbian learner, when a cue co-occurs with an outcome, the association between them becomes stronger; moreover, associations from absent cues (in a learning trial) to all outcomes are left unchanged. However, in the NDL model, associations from present cues to absent outcomes are weakened, and can eventually become negative. The model is naïve because every outcome is updated independently of all other outcomes.

The update in associations is summarized in equation (2), where t is a learning trial, ΔV_{ij} is the change in association involving a cue c_i and an outcome o_j.

$$\Delta V_{ij} = \begin{cases} \alpha_i\beta_1(\lambda - \sum_{c\in t} V_c) & \text{if } c_i \in t \text{ and } o_j \in t \\ \alpha_i\beta_2(0 - \sum_{c\in t} V_c) & \text{if } c_i \in t \text{ and } o_j \notin t \\ 0 & \text{if } c_i \notin t \end{cases} \quad (2)$$

α_i is a parameter modifying the salience of an input unit, or cue: while a different value can be set for each cue, this parameter is usually kept constant to remain agnostic with respect to cue importance. β_1 and β_2 specify the importance of positive and negative evidence respectively. These two parameters can again take different values but are usually set to the same quantity to reduce the initial assumptions. λ is the maximum amount of association that each outcome can receive from all inputs and operates as a simple linear scaling factor (Evert & Arppe, 2015). Finally, $\sum_{c\in t} V_c$ is the total association supported by the cues present in the current learning trial: this evidence is used to predict the outcome, and the prediction error is used to update cue-outcome associations. ΔV_{ij} is added to the current association value of cue c_i for each outcome o_j encountered up to trial t. The same happens for all $c_i \in t$. For the reported simulations we selected standard parameter values that allow to make minimal assumptions, setting all $\alpha s = 0.2$; $\beta_1 = \beta_2 = 0.1$; $\lambda = 1$.

Probabilistic learner In its original formulation (Fazly et al., 2010), this model computes a posterior probability distribution over referents for each word, updating the probability mass allocated to each referent in the light of new evidence. A referent r that seldom occurs with a word w but often occurs with many other words will get a small probability for w, while a referent r' that often occurs with word w and rarely with others will have a high probability of being the correct referent for w. The model incrementally updates associations between words and referents and uses them to compute the conditional probability of a referent given a word for all the referents that occurred with the word up to the present learning trial.

More generally, this model can be thought of as computing a posterior distribution over all possible outcomes for each cue. Associations between cues and outcomes are computed as specified in equations (3-5), where t is a learning trial, o is an outcome from the set of outcomes in the learning trial, O_t, c is a cue, from the set of cues in the learning trial, C_t, paired with O_t, and C is the set of cues encountered up to t:

$$a(c|o,O_t,C_t) = \frac{p_{t-1}(o|c)}{\sum_{c'\in C_t} p_{t-1}(o|c')} \quad (3)$$

$$assoc_t(c,o) = assoc_{t-1}(c,o) + a(c|o,O_t,C_t) \quad (4)$$

$$p_t(o|c) = \frac{assoc_t(c,o) + \lambda}{\sum_{o'\in O} assoc_t(c,o') + \beta\cdot\lambda} \quad (5)$$

This model has 3 free parameters. λ is a small smoothing factor; β is the upper bound on the expected lexicon; $p_{t=0}(o|c)$ is the initial value of the probability of an outcome given a cue, before they are encountered in a learning trial. In the simulations reported by Fazly et al. (2010), $\beta = 8.500$, $\lambda = 10^{-5}$, and $p_{t=0}(o|c) = 1/8{,}500$. We kept the same value for λ, set $\beta = 10^4$ and $p_{t=0}(o|c) = 10^{-4}$.

Equation (3) computes the update in association between a cue and an outcome from the current learning trial: this update is proportional to $p(o|c)$ at the previous learning trial and depends on the number of cues in the current trial: more cues cause a lower change, due to higher noise and uncertainty in the current trial. The update computed in (3) is added to the corresponding cue-outcome association, as specified in (4). Associations are not exploited directly but rather used to update a probability distribution over outcomes for each cue. More evidence makes the learner allocate a higher posterior probability to a specific outcome. In (5), the denominator acts as a scaling factor that implements within-trial competition: if a cue c is already associated to one of the previously encountered outcomes, the probability that c maps to another outcome does not receive strong support.

In the original formulation, words were cues and referents were outcomes; however, considering what we discussed in the section about Feature-Label-Order effects (Ramscar et al., 2010), we flipped the encoding so that this algorithm learns a probability distribution for words over referents, coding words as outcomes and referents as cues[2].

Hypothesis-Testing Model (HTM) The HTM model selects, stores and updates a single hypothesis for each learning trial. Initially, it randomly picks a word-referent mapping from the possible ones in the learning trial. When an already encountered word is presented in a subsequent trial, the model looks in memory to retrieve the hypothesized referent for the word and may retrieve it or not. If it does, the hypothesis is strengthened when confirming evidence is found in the current trial and discarded otherwise, in which case a new referent is hypothesized at random for the word being considered. If no hypothesis is recalled, a new referent is hypothesized at random for the word being considered and the old one fades away. As is specified in Medina et al. (2011) and Trueswell et al. (2013), the model depends on one main parameter, α, which models the probability that a formed hypothesis is retrieved from memory. However, Trueswell et al. (2013) argue that the value of this parameter changes when a hypothesis is recalled: the next time the label appears, the hypothesized referent should be retrieved with a higher probability if it was already retrieved. Unfortunately, however, no function was specified to model the change of α after successful retrievals. Therefore, we set the initial and second values for α at 0.6 and 0.81, following the third experiment in Trueswell et al. (2013), were this model was shown to fit behavioral results. Accordingly, the first time a hypothesis can be retrieved with probability equal to 0.6; if it gets confirmed, the next time it will be retrieved with probability equal to 0.81. Since we have many more trials, we set further values, 0.9, 0.95, and 0.99, to model the probability that a hypothesis is retrieved after the third, fourth or fifth time it was retrieved and confirmed. After the fifth time α does not change anymore: we stopped at 0.99 to exclude certainty of recall.

Computational Simulations

In order to closely mimic the learning task that was faced by children in the study by Ramscar, Dye, and Klein (2013), we implemented incremental learners: the connection between a referent and a word is only updated when both have been encountered in a learning trial. This is crucially different from the simulations implemented by Ramscar, Dye, and Klein (2013), where the equilibrium equations (Danks, 2003) of the NDL model (Baayen et al., 2011) were used. In this case, the end state of the model is computed when no more learning trials are available. Equilibrium equations have the advantage of not depending on any free parameter, but all cues and all outcomes are simultaneously available to the learner, differently from the task faced by the subjects. There was no way they could expect a third label to be presented during testing and thus update connections from objects to that label during training.

Here, we focus on the situation where children and undergradutes showed consistent behaviors, i.e. in retrieving an object when presented with a label they encountered during training. If a model fails to account for this aspect of the data, it can be hardly justified as a model of human cross-situational learning, during acquisition as well as in adulthood. Accordingly, we train each simulation using the input presented to the subjects and evaluate the final state of learning. However, since different models learn different things (associations, probabilities, hypotheses), it is hard to directly compare them. We do not assume any linking mechanism that converts internal representations to behavior; we simply look at the learned representations and evaluate whether unambiguous mappings were learned, that could allow subjects to retrieve an object, consistently with learning displayed by human subjects.

In each learning trial, simulated learners were given a set of objects and a word: beside ObjA, ObjB, and ObjC, the set of cues also contained other cues that account for the whole experimental context,[3] for consistency with the original simulation in Ramscar, Dye, and Klein (2013). Table 1 summarizes the input to the computational models.

Table 1: Training trials, as described in Ramscar, Dye, and Klein (2013)

Cues	Outcomes	Freq
ObjA_ObjB_Context1_ExptContext	Dax	9
ObjB_ObjC_Context2_ExptContext	Pid	9

For all models, we ran 200 simulations randomizing the order of presentation of the learning trials: since no model depends on initial random values, the order of the trials is the only potential source of bias. We report referent-word associations at the end of training for the four models in Table 2[4].

Successful learning happens when, for each label, the value corresponding to an object is consistently higher than the values of the other two objects, given that the test procedure consisted of presenting a label and asking for the matching object. In this setting, the Hebbian learner would choose randomly and is not learning much, since, in both the *Dax* and *Pid* columns, two objects have the same association to each label. On the contrary, the NDL model would retrieve the correct object given the two words provided during training, since the ObjA-*Dax* and ObjC-*Pid* associations are higher than any other. Another interesting feature is that it learns that ObjA does not come with the label *Pid*, forming a neg-

[2]Personal communication with one of the authors confirmed that the learning mechanism is not altered by switching the mapping.

[3]This was not the case for the HTM model, in which only ObjA, ObjB, and ObjC were provided as input.

[4]For explanatory purposes we will focus on the three objects, leaving the other cues out.

Table 2: Referent-word associations after 18 training trials (200 simulated learners). For the Probabilistic Learner, conditional probabilities of label given object are showed; for the Hypothesis Testing Model, the proportion of learners that selected each hypothesis is showed.

Model	Cue	Dax	Pid
Hebbian Learner	ObjA	9	.
	ObjB	9	9
	ObjC	.	9
NDL	ObjA	.127 $\pm$.003	-.052 $\pm$.004
	ObjB	.076 $\pm$.003	.076 $\pm$.003
	ObjC	-.051 $\pm$.005	.127 $\pm$.002
Probabilistic Learner	ObjA	.967 $\pm$.003	.
	ObjB	.484 $\pm$.085	.485 $\pm$.085
	ObjC	.	.967 $\pm$.003
HTM	ObjA	.465	.
	ObjB	.535	.53
	ObjC	.	.47

ative association. The Probabilistic Learner makes similar predictions to the NDL model, except for the negative associations. Finally, the HTM performs close to chance, with as many simulated learners mapping *Dax* to ObjA as to ObjB, and *Pid* to ObjB and ObjC, again showing no sign of learning, inconsistently with the behavioral evidence we considered.

Discussion

Our results show that some of the proposed learning mechanisms fail to account for the behavioral data obtained by Ramscar, Dye, and Klein (2013), for both children and adults: specifically, a Hebbian learner (Hebb, 1949) and the HTM (Trueswell et al., 2013) fail to learn robust object-label mappings. Two other models, the Probabilistic Learner (Fazly et al., 2010) and the NDL model (Baayen et al., 2011), show remarkably similar patterns to the behavioral data from both children and adults. The behavioral evidence also makes it clear that it is not necessary for successful cross-situational learning that true word-referent associations are more frequent than spurious associations. As a matter of fact, in the dataset each word-referent pair occurs with the same frequency, defying the very notion of a spurious pairing: ObjA could be paired to *Dax* just as ObjB could, if we only consider frequency of co-occurrence of objects and words. Nonetheless, humans learned consistent mappings, suggesting that simply tracking co-occurrence frequencies is a poor candidate mechanism to explain cross-situational word learning.

As is often the case in attempts to compare models, many decisions need to be taken and different choices can result in different outcomes. For example, we did not equip the HTM with a mutual exclusivity bias, mainly because it is not specified in the paper where the model was proposed and also because we wanted to evaluate basic versions of each model to focus on the proposed learning mechanisms rather than specific features. However, even with such a bias, the HTM would fail to match the behavioral data. Consider the situation in which the model first sees a *Dax* trial and it randomly picks ObjB as a referent. When a *Pid* trial is presented, the learner searches in memory, finds a *Dax*-ObjB hypothesis, decides that *Pid*-ObjB is not legitimate, and maps *Pid* to ObjC. If the HTM starts with a wrong mapping for *Dax*, it will only find the correct mapping for *Pid*, but will keep failing at relating ObjA to *Dax*. The problem lies in the single hypothesis assumption, not in the absence of the mutual exclusivity bias. In order to account for this behavioral evidence, a model should hold in memory the two possible hypotheses. Only then could it appreciate the fact that ObjB occurs with both labels while ObjA and ObjC consistently occur with one. The same problem of failing to appreciate the different background rates of the three objects affects the Hebbian learner, but results from an entirely different architecture, since it only focuses on co-occurrences to update associations. However, the behavioral evidence suggests that subjects do assign importance to missing co-occurrences too, and our simulations show that successful learning is only possible when a model is sensitive to both positive and negative co-occurrences. Taken together, the failures of the HTM and the Hebbian learner point to the importance of storing multiple mappings and being sensitive to both things that co-occur and things that fail to co-occur in the environment (Ramscar, Dye, & McCauley, 2013).

Unlike Trueswell et al. (2013) and Dautriche and Chemla (2014), we only evaluated the end-state of learning and did not consider trial-to-trial patterns, due to the behavioral data we used for comparison. This analysis would have certainly been useful because it allows to follow the learning trajectory. However, if a model does not account for the end state of learning it can hardly explain the mid-states, while a model that fits the final picture might have done so in different ways than the subjects. Thus, the reported evidence appears to be strong enough to make a case against the psychological plausibility of a model, while more evidence is needed about models that fit the behavioral data.

Finally, we did not evaluate any specification of which mechanism can make use of the associations learned during training to actually decide which object to retrieve when presented with a new label. While this is an interesting component of the paradigm in Ramscar, Dye, and Klein (2013) and it is crucial to investigate how learning mechanisms differ between young children and adults, we provided evidence that some learning mechanisms fail to account for behavioral data from both groups even when the much simpler condition of retrieving a referent when presented with a known word is considered. Further analyses are required to identify those mechanisms that can both i) form the correct associations during training and ii) use such associations to retrieve a known referent for an unknown word, in the same way children and adults do, to highlight where their learning mechanisms differ and where they are comparable.

Conclusion

The evidence we provided in this paper complements the study by Yu and Smith (2012) by showing that not every learning mechanism can be instantiated in an algorithm that accounts for behavioral data in cross-situational word learning. A single-hypothesis learning strategy (Medina et al., 2011; Trueswell et al., 2013) and an associative model that only relies on Hebbian learning (Hebb, 1949) fail to fit behavioral data. The jury is still out about the Probabilistic Learner (Fazly et al., 2010) and the Naive Discriminative Learner (NDL, (Baayen et al., 2011)): both models fit the results by Ramscar, Dye, and Klein (2013), but they behave differently, prompting for further research on which mechanisms underpin cross-situational learning in humans.

Acknowledgments

We are grateful to Michael Ramscar and Konstantin Sering for the discussion over the NDL model; to Aida Nematzadeh for her help in better understanding the Probabilistic Learner; to Chen Yu for sharing his latest research with us.
This research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

References

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naïve discriminative learning. *Lang Speech*, *56*(3), 329-347.

Baayen, R. H., Milin, P., Durdević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychol Rev*, *118*(3), 438-481.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2015). Comprehension without segmentation: a proof of concept with naïve discriminative learning. *Lang Cogn Neurosci*, 1-23.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *J Math Psychol*, *47*(2), 109-121.

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *J Exp Psychol Learn Mem Cogn*, *40*(3), 892.

Dawson, M. R. W. (2008). Connectionism and classical conditioning. *Comp Cogn Behav Rev*, *3. Monograph*, 1-115.

Evert, S., & Arppe, A. (2015). Some theoretical and experimental observations on naïve discriminative learning. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics.*

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017-1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychol Sci*, *20*(5), 578-585.

Hebb, D. O. (1949). *The organization of behavior*. New York, NY: John Wiley and Sons.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proc. Natl. Acad. Sci.*, *108*(22), 9014-9019.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychol Sci*, *24*(6), 1017-1023.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, *89*(4), 760-793.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: non-linear dynamics of lifelong learning. *Top Cogn Sci*, *6*(1), 5-42.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, *34*, 909-957.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: current research and theory* (p. 497). New York, NY: Appleton-Century-Crofts.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39-91.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross- situational statistics. *Cognition*, *106*(1), 1558-1568.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *J Exp Child Psychol*, *126*(1), 395–411.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cogn Psychol*, *66*(1), 126-156.

Warren, H. C. (1921). *A history of the association psychology*. New York, NY: Charles Scribner's Sons.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychol Sci*, *18*, 414-420.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychol Rev*, *119*(1), 21-39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53-62.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891-921.

Feature-based Joint Planning and Norm Learning in Collaborative Games

Mark K Ho[1] (mark_ho@brown.edu), James MacGlashan[2] (james_macglashan@brown.edu), Amy Greenwald[2] (amy@cs.brown.edu), Michael L. Littman[2] (mlittman@cs.brown.edu), Elizabeth M. Hilliard[2] (betsy@cs.brown.edu), Carl Trimbach[2] (ctrimbac@cs.brown.edu), Stephen Brawner[2] (sbrawner@cs.brown.edu), Joshua B. Tenenbaum[3] (jbt@mit.edu), Max Kleiman-Weiner[3] (maxkw@mit.edu), Joseph L. Austerweil[1] (joseph_austerweil@brown.edu)

[1]Department of Cognitive, Linguistic, and Psychological Sciences, 190 Thayer St., Providence, RI 02912 USA
[2]Computer Science Department, 115 Waterman St., Providence, RI 02912 USA
[3]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

People often use norms to coordinate behavior and accomplish shared goals. But how do people learn and represent norms? Here, we formalize the process by which collaborating individuals (1) reason about group plans during interaction, and (2) use task features to abstractly represent norms. In Experiment 1, we test the assumptions of our model in a gridworld that requires coordination and contrast it with a "best response" model. In Experiment 2, we use our model to test whether group members' joint planning relies more on state features independent of other agents (landmark-based features) or state features determined by the configuration of agents (agent-relative features).

Keywords: joint intentionality; norms; team reasoning; reinforcement learning; features; computational modeling

Introduction

From driving to running institutions like the U.S. Postal Service, groups need to coordinate their behaviors to accomplish shared goals. Key to this is that agents learn and use *norms* to guide individual and collective behavior. But how do people (or more generally, how can any intelligent agent) represent and learn norms?

One approach is to treat coordination as emerging through the ego-centric behavior of individual agents. For instance, norms can emerge when agents have other-regarding or aligned preferences (Binmore, 2010). Other approaches use off-the-shelf algorithms, like Q-learning, to show how under certain reward structures, "socially-blind" learning mechanisms can produce social norms (Sen & Airiau, 2007; Claus & Boutilier, 1998). More sophisticated approaches allow agents to model others and best respond to the predictions of those models. For example, agents can recursively reason about one another in *cognitive hierarchies* (Camerer et al., 2004; Wunder et al., 2011).

These computational approaches generally make two key assumptions: First, norms are modeled as *emergent* behavioral by-products rather than *intended* outcomes of agents' learning mechanisms. Second, the space of possible norms is generally constrained to a small set of singular actions (e.g. *cooperate* or *defect* in Flood and Dresher's Prisoner's Dilemma). As a result, the *representation of a norm* is never distinguished from the low-level actions that instantiate the norm.

Unfortunately, psychological research and intuition raise doubts about applying these assumptions to people. For instance, people take a group perspective when choosing their actions in coordination games using *focal points* (Schelling, 1960; Bardsley et al., 2010). Similarly, norms like "curb your dog" seem to rely on learned abstract representations that are applied flexibly to new situations. With this in mind, we have formulated a computational model that incorporates two novel properties:

(1) Agents reason *as if* they were part of a single agent with joint mental states like beliefs, desires, and plans. For instance, a postal worker does not simply reason in terms of "I-intentions" (e.g. I will bring these letters to this address), but also in terms of "we-intentions" (e.g. I will deliver these letters so *we* can deliver the mail) (Searle, 1995; Bacharach, 2006).

(2) Norms are represented as *joint planning biases* that reflect instructions to perform (or avoid) actions. Following Biccheri (2006), agents both follow these instructions and expect others to as well. Formally, these are feature-based reward functions for when an agent plans actions. This provides a compact representation for norms that enables generalization.

This model represents a first step towards understanding how norms are learned through joint reasoning and represented abstractly, aspects of human norm learning not captured in previous formulations.

To study how people learn norms, we focus on multi-state, multi-round coordination games. In our tasks, payoffs are always shared and depend on multi-state planning to reach individual goals simultaneously. In two experiments, we examine the extent to which our model captures how people learn norms. Experiment 1 compares people's behavior to our Norm-Learning model and a Best-Response model that plans actions optimally according to a learned model of its partner. Experiment 2 uses the Norm-Learning model to examine how people generalize norms across situations and the extent to which they use landmark-based or agent-relative features.

Computational Models

Norms are instructions that individuals follow and expect others to follow. We formalize this notion and describe how a group of norm-following agents can converge on norms in a decentralized manner.

Multi-Agent Decision Making

Markov Decision Processes (MDPs) and Stochastic Games MDPs model single-agent decision making and are defined by the tuple (S, A, T, R): a set of states in the world, S; a set of actions the agent can take, A; transition dynamics, $T(s'|s,a)$, which assign a probability of transitioning to a state $s' \in S$ after an agent takes action $a \in A$ in state $s \in S$; and a reward function $R(s,a,s')$, which returns a real valued reward when transitioning to state s' after the agent has taken action a in state s'. An agent in an MDP has a policy (a mapping from states to actions) $\pi: S \rightarrow A$. An agent's policy relates directly to the value, or expected future discounted reward, of each state: $V(s_0) = E_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$, where $\gamma \in [0,1]$ is a discount factor specifying the value of immediate rewards relative to temporally distant ones. Here, $\gamma = .95$.

To find an optimal policy, an agent needs to calculate the optimal state ($V^*(s)$) and state-action ($Q^*(s,a)$) value functions. Given these functions, the optimal policy can be derived by taking the action with the highest value: $\pi^*(s) \in \text{argmax}_{a \in A}\, Q(s,a)$ (Sutton & Barto, 1998).

MDPs can be extended to include multiple agents using game theoretic tools (Littman, 1994). A *stochastic game* is defined by the tuple (I, S, A^I, T, R^I), where I is an index set of agents in the environment, S is the set of states; A^I is the set of actions for each of the agents with A^i denoting the action set of agent $i \in I$; $T(s'|s,j)$ defines the task dynamics by specifying transition probabilities given a *joint action*, $j \in \times_i A^i$, of all agents taken in state $s \in S$; and R^I is a set of reward functions for each agent, with $R^i(s,j,s')$ denoting the reward received by agent $i \in I$ when agents in state $s \in S$ take joint action $j \in \times_i A^i$ and the environment transitions to state $s' \in S$.

Because multiple agents with individual reward functions are involved, there is no direct analogue of an 'optimal policy' in stochastic games. Rather, solution concepts can be posited (e.g. Nash equilibria) or a learning mechanism can determine how the multi-agent system converges.

Norm-Learning Model

Norms as reward function biases We assume that norms are instructions that an agent (a) follows, and (b) expects others to follow. More formally, we first represent the instructional content of norms as *reward biases* that cause a group of agents to prefer certain types of actions or states. For example, a norm to "drive on the right" would be represented as a collective preference for states that satisfy that description. This provides a natural, flexible way to represent the content of norms. To simplify the problem, we assume the norm bias is based on a linear combination of state features. Assuming agents have a feature function, Φ, that maps states to feature vectors, the norm bias is represented as:

$$R_{\text{norm}}(s) = \theta^T \Phi(s)$$

where θ^T is the transpose of a feature weighting vector. This allows the model to learn that certain state features (e.g. being on the right) is preferable during joint planning.

Second, we incorporate the motivational influence of norms directly into individual agents' reward functions that are used to calculate a reward-maximizing policy. Formally, for the i-th agent in a community, their total reward function will combine their private reward function and a norm bias:

$$R^i(s) = R^i_{\text{individual}}(s) + R_{\text{norm}}(s).$$

All agents in the community will have the same norm bias, R_{norm}, and know other agents will follow it. Thus, norms are joint reward function biases that agents follow and expect other agents to follow.

Inferring Norms We implement learning norms as *group inverse reinforcement learning* (IRL). In single-agent IRL one observes an agent behaving in an MDP and based on those observations infers the goals or reward function of the agent (Abbeel & Ng, 2004; Baker et al., 2009).

A Norm-Learning agent attempts to infer the norm that a group follows given some history of group interaction. That is, each agent estimates the most likely norm given a history of interaction, $\mathcal{H} = \big((s_0, j_0, s_1), \ldots, (s_{T-1}, j_{T-1}, s_T)\big)$:

$$\hat{R}_{\text{norm}} = \underset{R_{\text{norm}}}{\text{argmax}}\, P(R_{\text{norm}} | \mathcal{H}).$$

Since the norm bias function is a linear weighting of features, this corresponds to finding the most likely weights, $\hat{\theta}$.

Here, we focus on norm learning in collaborative games. That is, we assume that all agents all have the same goal (i.e. have the same $R^i_{\text{individual}}$) but must figure out how to work together to accomplish it. This simplifies inferring the norm bias. Other work should investigate how norm learning interacts in competitive scenarios (e.g. see Kleiman-Weiner et al. in this year's proceedings).

Features for Learning Norms Our representation of norms and implementation of norm learning depends on the features available to individuals in the group. The specific features are important for several reasons. First, to converge on a norm, individuals must have sufficiently similar features available to them to determine which norm the group uses. Second, features must be sufficiently expressive to allow individuals to pin down the norm that they collectively use to solve a task. Third, learning norms in terms of features allows generalization to novel situations. Without a concise, abstract representation of a norm, people would not be able to apply a learned norm to a new context and would need to learn an appropriate norm from scratch. For the tasks in the experiments reported, we describe which types of features are used for constructing norms.

We consider two types of features: *landmark-based features*, such as "Agent X is north of its goal", and *agent-relative features*, such as "Agent X is north of Agent Y". These two types were chosen because the former are an 'asocial' representation, while the latter explicitly involve social others. Moreover, they lead to different predictions in the tasks we use.

Best-Response Model

Best-response agents individually plan using a model of other agents. This means that instead of reasoning about a joint-policy directly, an agent i uses a predictive model of another agent j's policy, $\hat{\pi}_j$, to predict what j will do in a certain state. That is, an agent i will construct a transition function that includes predictions about the behavior of the other agent, $\tilde{T}_i(s'|s,a_i) = T\left(s'\middle|s,\left(a_i,\hat{\pi}(s)\right)\right)$.

Here, we use a level-1 cognitive hierarchy planner as our Best-Response model. It models its partner's behavior directly by counting its partner's actions and decaying past counts by a parameter δ. Additionally, to accelerate learning, the model assigns a pseudocount, α, to joint states in which the partner's location on the grid is the same[1]. Although we could have modeled a higher level of reasoning (e.g. best responding to a level-1 planner) we did not for two reasons. First, previous experimental work has shown that people do not typically reason beyond one or two levels (Camerer et al., 2004). Second, in non-competitive contexts, strategies often converge at higher levels in the cognitive hierarchy, and even level-1 reasoning provides a good estimate of this converged behavior (Bardsley et al., 2010).

Experiment 1: Hallway Task

Task Description

To test whether people best respond or learn norms, we designed the 2-person Hallway task shown in Figure 1. Two agents (circles) start at opposite ends of a 5x3 grid and on each turn simultaneously move up, down, left, right, or wait. The two agents cannot enter the same state or immediately switch positions – if they attempt this, then they collide and remain in the same location as in the previous time-step. Each agent has its own goal tile, indicated by a matching color, and the two agents start the task on one another's goals. Whenever ***either*** agent enters its own goal state, the round ends. However, to succeed in a round (and in the human case win a bonus), the agents must ***simultaneously*** enter their respective goals. This necessitates collaboration.

At the beginning of a round, each agent is exactly 4 tiles away from its goal. But they cannot both take a direct route to their goals without colliding. Rather, they must choose a series of actions that enables them to "break the symmetry"

[1] For example, if the agent is at (1,1) and the partner is at (0,0), then the partner's behavior will be generalized to other joint states where it is at (0,0). In this paper $\delta = 0.5$, $\alpha = 0.5$.

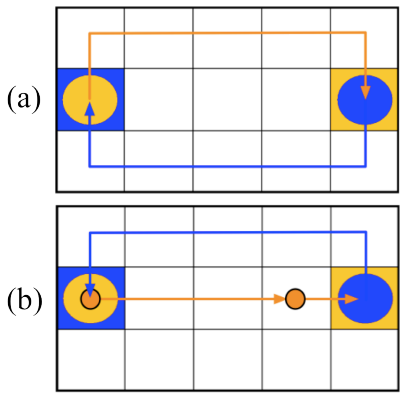


Figure 1: Hallway Task examples where (a) agents pass on top and bottom rows, and (b) they pass on middle and top rows. Smaller circles indicate the agent waited a step.

and pass one another. Critically, at least one of the two agents has to deviate from the center row of the grid and return to the center for the two to successfully complete the task. The other agent will either have to do the same, but on a different row, or wait two time-steps for the other agent. Figure 1 displays two joint plans that successfully complete the task in the minimal number of steps (6). Note though that there are many other possible joint actions that the two agents can take to pass one another.

In a given round, we can consider the row that each player is on when the two pass. Each player can be either on the *top*, *bottom*, or *center* row when attempting to move closer to their own goal. Clearly, successful passing requires that the two agents be on different rows while attempting to pass. Figure 2 visualizes this as an outcome matrix. Executing a successful joint policy, defined as both agents reaching their goal in the minimal number of steps, requires that both agents select different passing rows.

		Player 2's Passing Row		
		Top	Center	Bottom
Player 1's Passing Row	Top	*Fail*	*Success*	*Success*
	Center	*Success*	*Fail*	*Success*
	Bottom	*Success*	*Success*	*Fail*

Figure 2: Matrix representing passing success as a function of each player's row in the gridworld. Note that "Success" means the game was solved optimally.

Model Simulations

Best-Response Suppose two Best-Response agents succeed where player 1 passes through the center and player 2 passes along the top ({*center, top, success*}). Having observed player 2's behavior, player 1's predicts that player 2 will again choose *top*. From player 1's perspective, it is equally optimal to choose *center* as it is to choose *bottom*. However, if player 2 reasons similarly about player 1, then player 2 will treat *top* or *bottom* as equally optimal. If the players choose their respective pairs of equally optimal actions at

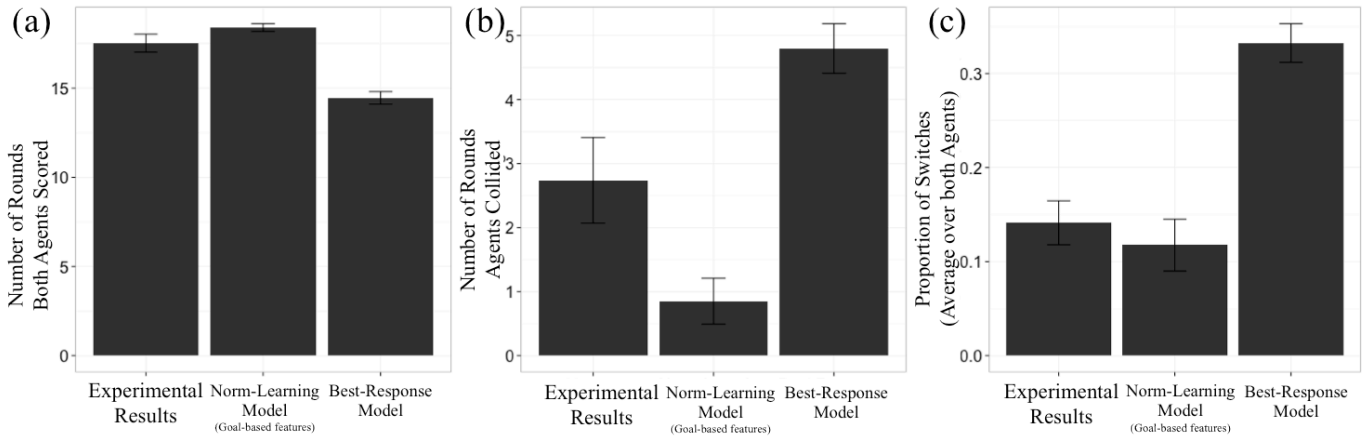


Figure 3: Human (experimental), Norm-Learning (simulation), and Best-Response (simulation) results for (a) number of rounds (out of 20) in which both agents scored, (b) number of rounds in which agents collided at least once, and (c) proportion of rounds in which agents switched their strategy from the previous round (averaged over both agents). These results suggest that human collaboration relies on jointly learned norms rather than best-responding to one's partner.

random (i.e. 0.5 each), then they will stay at {*center, top, success*} with .25 probability, switch to the {*bottom, top, success*} or {*center, bottom, success*} cells with 0.5 probability, or switch to {*bottom, bottom, fail}* with .25 probability. In our implementation, the probabilities differ from this ideal due to the decaying memory. Nonetheless, this illustrates the central prediction of best-response decision making in this collaborative game: that there will be high row-passing switching as well as a moderate amount of collisions from agents simultaneously switching.

Note also that the memoryless, mixed-strategy Nash equilibrium is itself a type of best-response solution concept. In this particular coordination game it is $\langle 1/3, 1/3, 1/3 \rangle$, which leads to an even higher proportion of collisions – $1/3$ – as well as switching – $2/3$.

Norm-Learning In the Norm-Learning model, two agents observe the same history of interaction, and use this to infer the most likely norm that a hypothetical collective agent is using. By using their shared observations and reasoning processes to deduce the most likely norm that they as a group have, they converge on and stay with a particular norm. For the Hallway task, we use a set of *landmark-based features* to define the space of norm reward biases. Specifically, for each of the two agents, we represent which row they are on relative to the row that their goal is on: above (top), on (center), or below their goal's row (bottom). This gives us a total of 6 binary goal-based features.

Unlike the Best-Response model, the Norm-Learning model predicts that people will stick with a combination of rows when performing the task. That is, in dyads that collaborate successfully, participants will not change which row they pass on and there will be few, if any, collisions.

Experiment

Design and Procedure We recruited 50 MTurk participants (25 dyads). They signed a consent form and then completed demo tasks that familiarized them with the grid game interface and task dynamics. They received a $2.00 base payment and an additional $0.10 bonus when they simultaneously reached their goals. Afterwards, each participant completed a post-task survey that included questions about the task and demographics. One dyad was excluded from analysis due to missing data.

For the simulations, agent dyads played 20 rounds and only learned at the end of each round. The Best-Response model updated its model of its partner based on the play the previous round, while the Norm-Learning model updated its distribution over possible norm biases. To simplify inference, we considered the space of feature weights to be $\theta \in \{-1,0,1\}^{f_n}$.

Results and Discussion Participants reported the task being relatively easy where 1 = Very Difficult to 7 = Very Easy (Mean = 5.67; SE = .19). Additionally, participants reported that they were skilled at the task on a scale from 1 = Very Bad to 7 = Very Good (Mean = 5.64; SE = .18).

The dyads were successful at collaborating on the task and winning the bonus. For our analysis, we focused on dyads that scored more than half of the rounds (23 of 24 dyads). These dyads, on average, jointly scored 17.5 out of 20 rounds (SE = 0.50) and jointly scored in the minimum number of steps possible (6) 15.22 out of 20 rounds (SE = 0.85). Human rounds scored did not differ from the Norm-Learning model ($t(35.307) = 1.82$, $p = 0.07$) but did differ from the Best-Response model ($t(38.0) = 4.98$, $p < .001$) (Figure 3a). However, direct comparison by scoring is difficult since the simulations update only once a round completes. This leads the Best-Response model to potentially collide indefinitely and never reach its goal.

Overall, the experimental results resemble the predictions of the Norm-Learning model over the Best-Response model. To compare human behavior in the collaborative Hallway task to the models, we focused on two measures: the number of rounds in which the agents collided at least once,

and the proportion of rounds in which agents switched their strategy from the previous round.

Figure 3b plots the number of rounds in which at least one collision occurred for a dyad. Human collisions fell directly in between the Norm-Learning model and Best-Response model and significantly differed from the numbers in both models (Best-Response: t(34.7) = -2.67, p = 0.01; Norm-Learning: t(33.2) = 2.49, p = 0.02). While this fails to distinguish between the two idealized models presented, it is consistent with a noisy or imperfect norm learning process.

To determine what strategy human and simulated agents used in a given round, we used the following heuristic: the first time a dyad collided in a round, the respective rows of the agents were coded as their strategies (note that agent locations after a collision remain the same as before). If a dyad never collided in a round we looked at the first step in which they were in the same column or switched columns -- i.e. the step at which they passed one another. Their respective rows *after* passing were coded as their strategy.

The amount of trial-to-trial strategy-switching provides a critical contrast between the Norm-Learning model and Best-Response model. Individuals in the human dyads switched their strategies rarely (14% of rounds) (Figure 3c). This closely matches the predictions of the Norm-Learning model ($t(38.8) = 0.66$, $p = .51$) and deviates significantly from the Best-Response model ($t(40.9) = -6.14$, $p < .001$). A related measure of joint strategy diversity tells us about global rather than local variability. To measure this, we calculated the entropy of the frequency distributions of joint strategies. Although the joint strategy entropy for humans significantly differed from both models (Norm-Learning: t(36.6) = 2.81, p < .001; Best-Response: t(36.2) = -5.55, p < .001), human entropy was closer to Norm-Learner entropy (Human: M = 0.68, SE = 0.10; Norm-Learning: M = 0.34, SE = 0.06; Best-Response: M = 1.34, SE = 0.06). Overall then, people displayed behavior more consistent with learning a jointly understood norm rather than best responding to their partner's behavior.[2]

Experiment 2: Feature-Based Norm Generalization

In our model, feature-based norms allow agents to converge on a joint strategy quickly, but they also enable individuals to generalize norms to new contexts. Here, we present the predictions of two sets of landmark-based features and one set of agent-relative features. We show that human norm learning better resembles landmark-based features assuming people learn norms like the Norm-Learning model.

Task Description

We designed the Courtyard tasks shown in Figure 4 to test norm generalization. In Direct Courtyard, agents can immediately move towards their respective goals through entrances. In Indirect Courtyard, agents must first move upwards towards entrances to cross. But both versions require agents to devise a way to pass. In our simulations and experiment, agents first play 10 rounds of Direct Courtyard, followed by 10 rounds of Indirect Courtyard.

If the agents have no means of generalizing learned joint plans, then they will have to find a new one when moving from Direct to Indirect Courtyard. On the other hand, if they learned a norm as a feature-based reward bias, then they can use it to guide their strategy choice on Indirect Courtyard.

Models

We tested three sets of binary feature-based reward biases: goal-relative features – is agent X's current row above, below, or the same as the goal row? (f_n = 2 agents x 3 features = 6 total features) –, entrance-relative features – is agent X's current row above, below, or the same as the entrances? (f_n = 6) –, and agent-relative features – is agent X above, below, or on the same row as agent Y? (f_n = 3). 40 simulations of each model were run.

Experiment

Procedure 90 MTurkers (45 dyads) participated. 3 dyads were excluded due to technical error. The procedure matched Experiment 1 except participants played 2 games.

Results and Discussion To determine whether and how participants generalized from Direct to Indirect Courtyard, we analyzed individual strategies within dominant joint strategies. Each round, we identified the strategy used with the same heuristic as in Experiment 1. For each dyad, we then identified the most frequent strategies in the Direct Courtyard and Indirect Courtyard phases. We then identified each agent's individual strategy in the two phases. Figure 5a shows individual strategy counts in the two phases. Note that if people were forming new joint plans from scratch, the strategies in the two phases would be uncorrelated. This was not the case ($\chi^2(4) = 27.45$, $p < .001$).

To explain the systematicity in how joint plans were generalized to the Indirect Courtyard, we compared this distribution to the simulation results of our three models (Figure 5b, 5c, and 5d). A visual analysis of these tables suggested that the participant results reflect a mixture of goal-relative and entrance-relative features. We confirmed

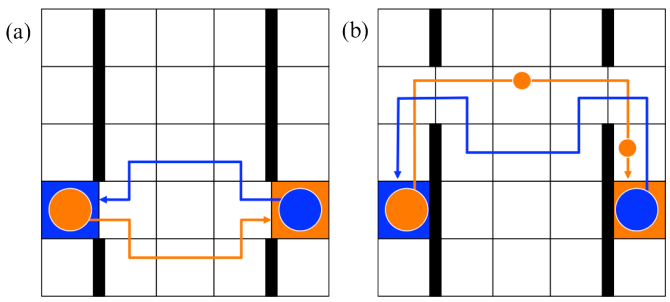


Figure 4: (a) Direct and (b) Indirect Courtyard Tasks with example optimal joint plans. The bold lines are walls that agents cannot pass through.

[2] We assume that best responders do not default to the previous choice if indifferent between options. Future studies will need to rule this out in people.

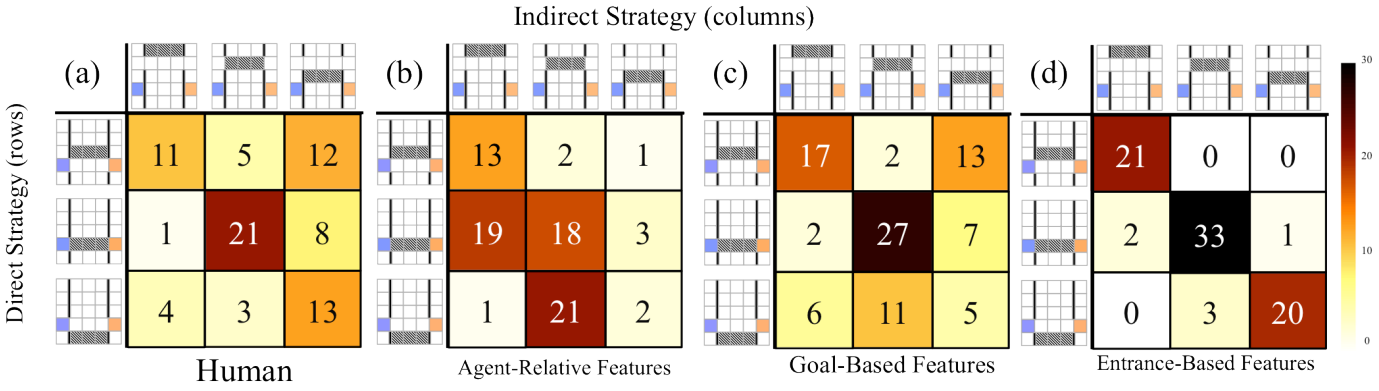


Figure 5: Individual agents' most frequent strategies (counts) by Direct and Indirect Courtyard phases for (a) experimental participants, (b) our Norm-Learning agent with Agent-Relative features, (c) Goal-Based features, and (d) Entrance-Based features. The distribution indicates which row an agent is likely to take in the Indirect Courtyard grid given what row was taken in the Direct Courtyard. Human results are best explained as a mixture of the two landmark-based feature sets (Goal-based and Entrance-based; see text). Grayed out rows in the gridworlds indicate the most frequent individual strategy.

this by calculating maximum likelihood mixture values for the three models: $p_{\text{Entrance}} = 0.21$, $p_{\text{Goal}} = 0.79$, and $p_{\text{Agent}} = 0.0$. Thus, participants tended to generalize norms using landmark-based rather than agent-relative features.

Conclusion

Here, we have presented a novel model of norm learning based on inferring joint reward biases. We compared the predictions of this model to those of a best response model in the Hallway task, and used the same model to show that people use landmark-based rather than agent-relative features to generalize norms across the two Courtyard tasks.

A central aspect of human sociality is engaging in shared intentions and joint plans with others (Searle, 1995). Using the formal tools of multi-agent MDPs, we are able to make quantitative predictions about how collaborating individuals represent and reason about themselves as part of a larger entity. Future work should explore how individuals learn norms over particular types of features, what happens when agents' feature sets differ from one another, and how learned norms interact in competitive scenarios.

Acknowledgments

This material is based upon work supported by the NSF GRF under Grant No. (DGE-1058262).

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In C. Brodley (Ed.), *Proc. of twenty-first international conference on machine learning* (p. 1). New York, NY: ACM.

Bacharach, M., Gold, N., & Sugden, R. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bardsley, N., Mehta, J., Starmer, C., & Sugden, R. (2010). Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning*. *The Economic Journal*, *120*(543), 40–79.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (2010). Social norms or social preferences? *Mind & Society*, *9*(2), 139–157.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Claus, C., & Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In J. Mostow, C. Rich, B. Buchanan (Eds.) *Proc. of the 15th national conf. on artificial intelligence* (pp. 746-752).

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In W. Cohen (Ed.), *Proc. of 11th international conf. on machine learning* (pp. 157–163). San Francisco, CA: Morgan Kaufmann.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard university press.

Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.

Sen, S., & Airiau, S. (2007). Emergence of Norms Through Social Learning. In R. Sangal, H. Mehta, R. K. Bagga (Eds.), *Proc. of the 20th international joint conference on artificial intelligence* (pp. 1507–1512). San Francisco, CA: Morgan Kaufmann.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Wunder, M., Kaisers, M., Yaros, J. R., & Littman, M. (2011). Using Iterated Reasoning to Predict Opponent Strategies. In L. Sonenberg, P. Stone (Eds.), *10th international conference on autonomous agents and multiagent systems* (Vol 2, pp. 593–600). Richland, SC: IFAAMS.

Emotional Implications of Metaphor: Consequences of Metaphor Framing for Mindset about Hardship

Rose K. Hendricks (rhendricks@ucsd.edu)
Lera Boroditsky (lera@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Do metaphors shape people's emotional states and mindsets for dealing with hardship? Natural language metaphors may act as frames that encourage people to reappraise an emotional situation, changing the way they respond to it. Recovery from cancer is one type of adversity that many people face, and it can be mediated by the mindset people adopt. We investigate whether two common metaphors for describing a cancer experience – the battle and the journey – encourage people to make different inferences about the patient's emotional state. After being exposed to the battle metaphor participants inferred that the patient would feel more guilt if he didn't recover, while after being exposed to the journey metaphor participants felt that he had a better chance of making peace with his situation. We discuss implications of this work for investigations of metaphor and emotion, mindsets, and recovery.

Keywords: metaphor; framing; emotion; adversity; cancer; battle; journey; mindset; recovery

Introduction

Linguistic metaphor influences the way we reason about social problems (Thibodeau & Boroditsky, 2011, 2013), physical phenomena (Dolscheid & Casasanto, 2013), and scientific concepts (Gentner & Gentner, 1983). These examples demonstrate that metaphor influences rational and deliberate decision-making. Can metaphor also influence emotion?

Metaphor evokes more emotion-related brain activity than literal terms (Citron & Goldberg, 2014), and emotion-laden metaphorical sentences encourage more mental simulation than less emotional metaphorical sentences (Samur et al., 2015). In addition, people who read about a character in a passage that used metaphors felt a greater sense of intimacy with that character than people who read about the character in the absence of metaphor (Bowes & Katz, 2015).

Most of the work on metaphor and emotion has detailed the ways that emotions are described metaphorically (Fainsilber & Ortony, 1987; Kövecses, 2000). For example, anger is described as seeing red, and participants who thought about anger or became angry perceived the color red more than people who didn't (Fetterman et al., 2011). Anger can also be metaphorically described in terms of heat, and people associate anger-related words and angry faces with heat (Wilkowski et al., 2009). Additional evidence suggests that these metaphorical ways of describing emotions (for example, anger as being contained in a bounded space, as opposed to being a fluid) influence the way people reason about those emotional states (Reali & Arciniegas, 2014). Whether a metaphorical framing can influence people's emotional states remains an open question.

Metaphor is often used in difficult situations. Here, we look at metaphors for overcoming adversity in the context of cancer. We compare the emotional inferences that people make about an individual with cancer after reading about his situation framed as either a battle or a journey, and we find that these metaphors give rise to different emotional inferences related to coping.

Adversity & Coping

The world presents diverse opportunities for dealing with adversity, and the way we think about that adversity matters. In a seminal work, Lipowski (1970) suggested that there are 8 prevalent ways that people attribute meaning to illness. These categories include seeing illness as a challenge, enemy, punishment, weakness, irreparable loss, relief, strategy, or a value. He further suggested that the meaning we attribute to our disease influences our coping abilities, in turn affecting recovery.

One way of coping is by reconceptualizing our experiences. This process of reappraisal allows us to change the way we feel about something by changing the way we think about it. At the neural level, reappraisal involves interactions between the prefrontal control systems that implement reappraisal strategies and emotional appraisal systems like the amygdala that generate an affective response (Ochsner & Gross, 2005). Similar work has shown that physical health improves when health concerns like stress are framed in more positive ways (Crum et al., 2013).

Can we alter our mindset about a difficult situation by invoking different metaphors? Can metaphor facilitate reappraisal of hardship to improve coping and health outcomes?

Language about Cancer

Over 1.6 million people are diagnosed with cancer each year. Upon diagnosis, are they beginning a battle or a journey with the disease? Psychological variables can affect recovery. For example, mental states like chronic stress, depression, and social isolation are linked with the biological pathways involved in cancer's progression, compromising functions like immune response and cancer cell death (Lutgendorf & Sood, 2011). Women with breast cancer who had more optimistic expectations of their

surgeries experienced less pain, nausea, and fatigue one week after surgery than those with negative expectations (Montgomery et al., 2010). Further, breast cancer patients with greater social support and a tendency to minimize the importance of their disease experienced better prognoses, while those demonstrating depression and constraint of emotions had lower levels of survival (Falagas et al., 2007).

Metaphors are pervasive in language about cancer. Casarett and colleagues found that oncologists used metaphors in roughly two thirds of their conversations with patients (2010). Patients also rated physicians who used more metaphors as better communicators and easier to understand than those who used fewer.

The predominant metaphor when discussing cancer is that of a battle (Reisfield & Wilson, 2004; Penson et al., 2004). This metaphor seems so deeply ingrained in our culture that a "fighting spirit" is one of the five categories in the Mental Adjustment to Cancer (MAC) scale, which is frequently used to assess patients' mentalities about their disease (Greer et al., 1989). On this scale, someone with a fighting spirit "fully accepts the diagnosis, uses the word 'cancer', is determined to fight the illness, tries to obtain as much information as possible about it and adopts an optimistic attitude; may see the illness as a challenge" (p. 374). The other four states – helplessness/hopelessness, anxious preoccupation, fatalism, and denial – are all negative. Because the only positive adjustment state involves conceptualizing cancer as a battle, this scale demonstrates that not only do we think of cancer as a battle, but those who designed and use the scale assume that patients *should* think of it that way. On the other hand, in one study women with metastatic cancer tended to associate their disease with the concepts of enemy, punishment, weakness, and irreparable loss more than patients with earlier stage cancer, suggesting that the battle metaphor seems more appropriate to the sickest people (Wallberg et al., 2009).

Our societal belief in the battle metaphor is also evident in interventions like the game Re-Mission, touted on its website (www.re-mission.net) as "the original cancer-fighting game." In the game, players fire treatments at growing tumors, drop chemo bombs on cancer cells, and collect healthy cells to earn points. Cancer patients aged 13-29 who regularly played Re-Mission showed greater adherence to their treatments, self-efficacy, and knowledge about the disease than those who did not (Kato et al., 2008).

Despite the prevalence of the battle metaphor, some researchers, patients, and physicians intuit that a battle, inherently violent, masculine, and power-based, may not be a productive framing for cancer. It may suggest that someone wins (recovers) if they try hard enough, a message that disregards more social and emotional aspects of healing (Reisfield & Wilson, 2004). A battle mindset may encourage suppressing negative feelings and imply the importance of treatment at all costs (Harrington, 2012). Further, it is conceptually inaccurate because cancer does not involve enemy invaders; instead, one's own cells are the enemies (Reisfield & Wilson, 2004). In one study, cancer patients perceived their clinicians as promoting a positive attitude and a "fighting spirit," which they then internalized as encouragement to conceal their emotional distress (Byrne et al., 2002).

One alternative is to describe cancer as a journey. There is no winning, losing, or failing on a journey; the emphasis is instead on a variety of possibilities and a larger process, since life itself is often compared to a journey (Reisfield & Wilson, 2004). One's experience with cancer is therefore just one part of a larger narrative (Harrington, 2012).

Despite the prevailing intuition that battle metaphors may do more harm than good, there is likely no perfect metaphor to talk about cancer. Different metaphors are likely to help different patients experiencing widely different circumstances (Reisfield & Wilson, 2004; Penson et al., 2004). In fact, a corpus analysis of online support forums and blogs written by patients and health professionals demonstrated that the battle was not inherently bad, nor was the journey inherently good (Semino et al., 2015). Both patients and health professionals used each of the metaphors in both empowering and disempowering contexts.

Metaphor & Reasoning

Metaphors often allow us to think about complex ideas by drawing on domains we have experience with. For example, when people used an analogy comparing the flow of electricity in a circuit to that of water through pipes, they made different inferences about the flow of electricity than people who use compared it to crowds of mice running along a track (Gentner & Gentner, 1983). The students made metaphor-consistent inferences, both correct and incorrect, because they mapped elements of the metaphorical domain onto their conceptualization of electrical circuits, when such a mapping was warranted and when it was not.

Can learning a new metaphor shape the way people think about complex concepts? Dutch speakers who learned to talk about musical pitches as thick and thin (as opposed to their conventional way of describing them as low or high) began to think of pitches in the new way, even in a non-linguistic task (Dolscheid & Casasanto, 2013). Similarly, people who learned to talk about the past as heavy judged an older book to be heavier than a new book, while those that learned to talk about the present as heavy judged the newer book as heavier than the older one, even though the book weights were identical (Slepian & Ambady, 2014). English speakers who learned to talk about time in terms of vertical space (with either earlier events either above or below later ones) also showed a corresponding shift in their nonlinguistic mental representations of time, consistent with their newly learned systems of metaphors (Hendricks & Boroditsky, 2015).

Some prior work has examined the effect of linguistic metaphor when more than one metaphor could be used to describe a social problem. In a series of experiments, people who read about crime as either a virus or a beast offered solutions for dealing with the crime that were consistent with the metaphor they had read (Thibodeau & Boroditsky,

2011, 2013). When people read that crime was a beast, they were more likely to offer solutions like increasing law enforcement and punishments for criminals than did virus readers. This is in line with what people would likely do if a literal beast were ravaging the town, which suggests that they reasoned in systematic metaphor-consistent ways.

One question that remains is whether metaphors have consequences for emotion. Can linguistic metaphor shape people's affect and mindset about a difficult situation?

In this paper, we explore the role of metaphor for framing the experience of having cancer. We investigated two common metaphors for an experience with cancer – a battle and a journey. After people were exposed to one of these metaphors in the context of a fictional character's disease, we measured the inferences they made about that person's guilt and ability to make peace with his situation.

According to the intuitions shared by a number of researchers and physicians, if the journey metaphor leads to more productive conceptualizations of a person's experience with the disease, people who read about it as such should think that the character will have better chances of making peace with the disease and will feel less guilty if he does not recover than people who read about his experience as a battle. Our results confirmed these predictions, suggesting that metaphors do influence the emotional mindsets that people adopt toward adversity. This work opens the door for future investigations of the mechanisms underlying a metaphor's influence on emotion.

Methods

Participants

We recruited and paid 528 participants through Amazon's Mechanical Turk. We restricted participants to those whose IP addresses were in the United States, had completed at least 500 tasks, and had an overall approval rate of at least 95%. Data from 25 participants were excluded because they did not finish the survey, resulting in 503 participants for analysis.

This sample included 295 males and 208 females. Most (84%) had completed at least some college. Twelve participants reported having been treated for cancer in the past, and the rest reported never having been diagnosed with cancer. Four of those participants read about the battle, and 8 read about the journey.

Materials & Procedure

We created two passages about an individual's experience with cancer. These passages were identical except in the metaphor used to describe cancer. One used a *battle* framing (n = 249), while the other used a *journey* (n = 254). Each participant read one of the following (boldface added to emphasize differences; no text was bolded in the actual materials):

Joe was just diagnosed with cancer. He knows that for the foreseeable future, every day will be a ***battle against*** *the disease. The* ***battle*** *he has to* ***fight*** *will not always be an easy one. Many people have written about their experiences on the* ***battlefield****, and he can turn to those for consolation. His friends and family want him to know that he will not be alone in his* ***battle****. Even though sometimes he might not feel like talking, other times he may want to share stories of his* ***battle*** *with others, and they will be there for those moments.*

Joe was just diagnosed with cancer. He knows that for the foreseeable future, every day will be a ***journey with*** *the disease. The* ***road*** *he has to* ***travel*** *will not always be an easy one. Many people have written about their experiences on the* ***path****, and he can turn to those for consolation. His friends and family want him to know that he will not be alone on his* ***journey****. Even though sometimes he might not feel like talking, other times he may want to share stories of his* ***journey*** *with others, and they will be there for those moments.*

On the following page, participants rated the extent to which two separate statements described Joe's experience based on the information they had read. Options for agreement with each statement ranged from 1 (not at all) to 7 (very much). The following were presented in a random order:

a) He will feel guilty that he hasn't done enough if he does not recover.

b) He can make peace with his experience.

After responding to the two statements, the next page stated: "Please give any additional information about Joe's experience that you can imagine based on the passage that you read previously." The experiment concluded by asking participants about their personal experience with cancer, gender, highest level of education, and languages they speak.

After reading all of the responses, we noticed that people mentioned Joe's social support and his mindset about his disease often, so we decided to code for these elements. To do so, we read all the responses, blind to participants' metaphor conditions, and noted the presence of comments about social support and positive mindset. We also coded for uses of both battle- and journey-associated language and whether the participant explicitly commented on the metaphor they had read.

Results

The metaphor framing influenced people's emotional inferences about Joe's guilt and his ability to make peace with his situation. Figure 1 shows these main results.

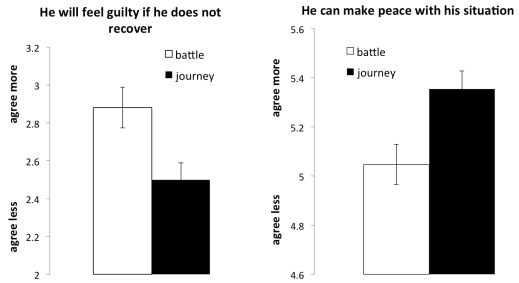


Figure 1: Main results. Agreement scales ranged from 1 (not at all) to 7 (very much).

Guilt

As predicted, people who read about the battle agreed more with the statement that Joe "will feel guilty if he does not recover," (M=2.87, SD=1.69) than did people who read about the journey (M=2.50, SD=1.42). A 2 (metaphor) x 2 (gender) ANOVA confirmed this difference by a main effect of metaphor ($F(1,498) = 7.07$, $p = .008$, $d = .24$). Gender was included because the masculinity inherent in the battle metaphor has been emphasized by previous work (Reisfield & Wilson, 2004; Penson et al., 2004; Harrington, 2012). Men reported that Joe would feel more guilty (M=2.88, SD=1.59) than women did (M=2.40, SD=1.50), confirmed by a main effect of gender ($F(1,498) = 11.1$, $p = .0009$, $d = .31$). The metaphors did not sway men's and women's responses to different extents, as there was no interaction between metaphor and gender ($F(1,498) = 0.16$, $p = .69$). Figure 2 shows the results for the statement about Joe's guilt.

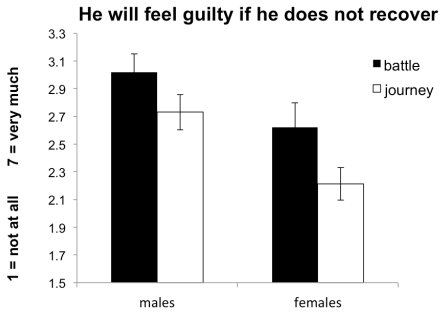


Figure 2: Responses to how guilty Joe will feel.

Making Peace

There was again a main effect of metaphor ($F(1,498) = 7.96$, $p = .005$, $d = .25$), but no main effect of gender ($F(1,498) = 0.33$, $p = .56$) or interaction between metaphor and gender ($F(1,498) = 0.15$, $p = .70$). Figure 3 shows the results for this statement.

People who read the journey metaphor agreed more with the statement that Joe "can make peace with his situation," (M=5.35, SD=1.16) than the group who read the battle metaphor (M=5.05, SD=1.29). There was no main effect of gender ($p = .68$) or interaction between metaphor and gender ($p = .95$). Figure 3 shows the results for this statement.

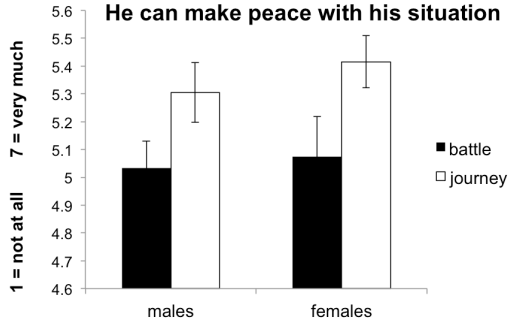


Figure 3: Responses to whether Joe can make peace.

Other Information Imagined

When writing additional information about Joe's experience, people who read about the battle used battle metaphors (42%) more than those who read about the journey (18%; $\chi^2 = 23.04$, $p < .0001$). Similarly, people who read about the journey used journey metaphors (17%) more than those who read about the battle did (5%; $\chi^2 = 18.13$, $p < .0001$). Overall battle metaphors were more common (29%) than journey metaphors (11%; $\chi^2 = 43.46$, $p < .0001$). These results are shown in Figure 4.

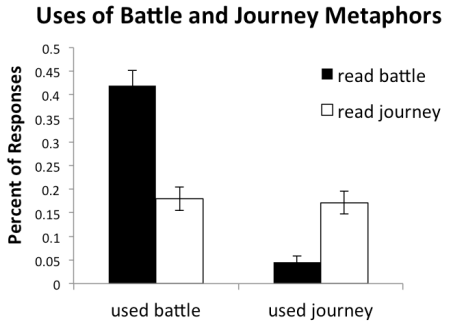


Figure 4: Use of battle and journey metaphors

Many people commented on the mindset they imagined Joe would have after his diagnosis. Examples reflecting the belief that Joe had a positive mindset include:

I imagine that he has a pretty positive mindset, he has people to consult with and people to lean on. I think he would be as positive as he could be because he has support. He's probably scared, but assured that he can make it.

Joe has come to terms with his situation. He knows just how hard the battle he has to face will be, and is prepared for it.

Because we coded for two types of information mentioned (social support and positive mindset), we used Bonferroni adjusted alpha levels of .025 per test. People who read about the journey were numerically, but not significantly, more likely to write about his positive mindset than those who read about it as a battle, $\chi^2 = 3.24$, $p = .07$.

Many participants also mentioned that Joe would receive good social support throughout his experience. These were comments such as:

Joe is lucky to have support from the people that love him. This can help him tremendously.

There was no difference in frequency of mentioning social support between the groups, $\chi^2 = 1.44$, $p = .23$.

There was also no difference in the response length between the two metaphor groups, $t(484.6) = 0.67$, $p = .50$.

Discussion

In this study, we considered the role that metaphor has on emotion about a difficult situation. People made different emotional inferences about recovery when they received different metaphors. Specifically, reading about someone's experience with cancer as a battle encouraged people to infer that the individual would feel more guilt if he did not

recover than reading about the same experience as a journey did. On the other hand, the journey framing encouraged people to believe this person could make peace with his situation more than the battle did.

People were more likely to use metaphors consistent with the one they read than the alternative. This is notable because if metaphors are perpetuated, either in a single person's mind or in conversations, their effects on emotion may also be perpetuated over time. This emotional contagion phenomenon has been observed for social media posts (Kramer et al., 2014).

Social support is important for health outcomes in cancer patients (Lutgendorf & Sood, 2011; Falagas et al., 2007), and there was no difference in the frequency with which this feature of Joe's situation was mentioned between people who read the different metaphors. This suggests that the differences we do observe in inferences about the guilt Joe will feel and his ability to make peace with his situation are not driven by underlying assumptions about Joe's social support. Because both journeys and battles can be experienced with others for support, the comparable frequencies with which the two groups mentioned social support show that not *all* emotional inferences were influenced by the battle and journey metaphors; instead, only aspects that were systematically related to the metaphors were impacted.

Future Work

Do battle and journey metaphors have the same affective influence when you reason about yourself or someone you know well, instead of a hypothetical person?

Are there real-world observable consequences of these metaphors for cancer behaviors or outcomes?

What are the individual differences that affect how people respond to metaphors? Does extensive experience with the metaphorical domain (i.e., battle) make a person more or less susceptible to being swayed by the metaphor than a lack of experience?

Do trait differences mediate a metaphor's emotional influence? For example, people who are competitive, thrive under stress, or are hesitant to ask for help from others may respond to the battle metaphor differently than those who are not. Do people's emotional processing skills mediate the effect that metaphor has on their affective reasoning?

What is the impact of mixing metaphors? A number of participants naturally mixed metaphors in their free responses (for example: *Based on the passage, I can tell that Joe's battle with cancer will be as difficult as those who experienced war on the battlefield. I can imagine that Joe will feel furious and hopeless about his predicament as the war vets did back on the battlefield. I'm glad that he'll have a solid supporting system to rely on throughout his journey.*) When metaphors are combined, do effects on inferences become diminished? Amplified? Are both metaphors' inferences co-activated?

Similarly, does a metaphor's influence on emotion change as it gets repeated and transmitted from one person to the next? Could its effects snowball, gaining power as the metaphor is perpetuated, or might its impact become watered down?

Do the metaphors influence the information that people seek out? Are people conscious of these metaphors' presence or their role in shaping emotion? Do people's preferences for different metaphors mediate the inferences that they make as a result of encountering different metaphors?

Finally, do novel metaphors for the cancer experience encourage different emotional inferences? Since new metaphors do alter conceptual representations of the things they describe (Dolscheid & Casasanto, 2013; Hendricks & Boroditsky, 2015), perhaps novel metaphors can help cancer patients see the disease in a new light, one without as much societal baggage as the current metaphors have.

Limitations

Limitations of this work can be addressed in future iterations. For one, the character that participants read about was always a male. This may have caused males to identify with Joe more than the females, which could at least in part explain males' stronger belief that Joe will feel guilty if he does not recover than females'. In addition, we do not know whether participants have had close friends or family members who have had the disease, which may mediate metaphor's role in shaping their emotions.

Conclusion

Because cancer is a complicated and multi-faceted disease, metaphor will likely always be present in the way we talk about it. This work demonstrates the affective consequences of two common metaphors for discussing cancer – as a battle and as a journey. It paves the way for further consideration of the mechanisms underlying role in shaping our emotions about adversity.

Acknowledgments

We thank Elena Semino and Zsofia Demjen for sharing expertise in cancer metaphors and contributing exploratory analyses. I attribute my zeal for the project to my dad, PH, and cousin, KF, whose cancer experiences convinced me that this topic needed to be studied quantitatively.

References

Bowes, A., & Katz, A. (2015). Metaphor creates intimacy and temporarily enhances theory of mind. *Memory & Cognition*, *43*(6), 953–963.

Byrne, A., Ellershaw, J., Holcombe, C., & Salmon, P. (2002). Patients' experience of cancer: evidence of the role of "fighting"in collusive clinical communication. *Patient Education and Counseling*, *48*(1), 15–21.

Casarett, D., Pickard, A., Fishman, J. M., Alexander, S. C., Arnold, R. M., Pollak, K. I., & Tulsky, J. A. (2010). Can metaphors and analogies improve communication with

seriously ill patients? *Journal of Palliative Medicine*, *13*(3), 255–260.

Citron, F. M. M., & Goldberg, A. E. (2014). Metaphorical Sentences Are More Emotionally Engaging than Their Literal Counterparts. *Journal of Cognitive Neuroscience*, *26*(11), 2585–2595.

Crum, A. J., Salovey, P., & Achor, S. (2013). Rethinking stress: The role of mindsets in determining the stress response. *Journal of Personality and Social Psychology*, *104*(4), 716–733.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The Thickness of Musical Pitch: Psychophysical Evidence for Linguistic Relativity. *Psychological Science*, *24*(5), 613–621.

Fainsilber, L., & Ortony, A. (1987). Metaphorical Uses of Language in the Expression of Emotions. *Metaphor and Symbolic Activity*, *2*(4), 239–250.

Falagas, M. E., Zarkadoulia, E. A., Ioannidou, E. N., Peppas, G., Christodoulou, C., & Rafailidis, P. I. (2007). The effect of psychosocial factors on breast cancer outcome: a systematic review. *Breast Cancer Res*, *9*(4), R44.

Fetterman, A. K., Robinson, M. D., Gordon, R. D., & Elliot, A. J. (2011). Anger as Seeing Red: Perceptual Sources of Evidence. *Social Psychological and Personality Science*, *2*(3), 311–316.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. *Mental Models*, 99–129.

Greer, S., Moorey, S., & Watson, M. (1989). Patients' adjustment to cancer: The mental adjustment to cancer (MAC) scale versus clinical ratings. *J Psychosomatic Research, 33*(3), 373-377.

Harrington, K. J. (2012). The Use of Metaphor in Discourse About Cancer: A Review of the Literature. *Clinical Journal of Oncology Nursing*, *16*(4), 408–412.

Hendricks, R. K., & Boroditsky, L. (2015). New space-time metaphors foster new mental representations of time. *Proceedings of the 37th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

Kato, P. M., Cole, S. W., Bradlyn, A. S., & Pollock, B. H. (2008). A Video Game Improves Behavioral Outcomes in Adolescents and Young Adults With Cancer: A Randomized Trial. *PEDIATRICS*, *122*(2), e305–e317.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, *111*(29), 10779.

Lipowski, Z.J. (1970). Physical illness, the individual, and the coping process. *Int J Psychiatry Med, 1*(2), 91-102.

Lutgendorf, S. K., & Sood, A. K. (2011). Biobehavioral Factors and Cancer Progression: Physiological Pathways and Mechanisms. *Psychosomatic Medicine*, *73*(9), 724–730.

Montgomery, G. H., Schnur, J. B., Erblich, J., Diefenbach, M. A., & Bovbjerg, D. H. (2010). Presurgery Psychological Factors Predict Pain, Nausea, and Fatigue One Week After Breast Cancer Surgery. *Journal of Pain and Symptom Management*, *39*(6), 1043–1052.

Ochsner, K., & Gross, J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, *9*(5), 242–249.

Penson, R. T., Schapira, L., Daniels, K.J., Chabner, B.A. & Lynch, T.J. (2004). Cancer as Metaphor. *The Oncologist*, *9*(6), 708–716.

Reali, F., & Arciniegas, C. (2014). Metaphorical Framing Influences How We Think about Emotions: Some Evidence from Spanish. *Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

Reisfield, G. M. (2004). Use of Metaphor in the Discourse on Cancer. *Journal of Clinical Oncology*, *22*(19), 4024–4027.

Samur, D., Lai, V. T., Hagoort, P., & Willems, R. M. (2015). Emotional context modulates embodied metaphor comprehension. *Neuropsychologia*, *78*, 108–114.

Semino, E., Demjen, Z., Demmen, J., Koller, V., Payne, S., Hardie, A., & Rayson, P. (2015). The online use of "violence"and "journey"metaphors by cancer patients, as compared with health professionals: a mixed methods study. *BMJ Supportive and Palliative Care*.

Slepian, M. L., & Ambady, N. (2014). Simulating sensorimotor metaphors: Novel metaphors influence sensory judgments. *Cognition*, *130*(3), 309–314.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, *6*(2), e16782.

Thibodeau, P. H., & Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PloS One*, *8*(1), e52961.

Wallberg, B., Michelson, H., Nystedt, M., Bolund, C., Degner, L. & Wilking, N. (2009). The Meaning of Breast Cancer. *Acta Oncologica*, *42*(1), 30–35.

Wilkowski, B. M., Meier, B. P., Robinson, M. D., Carter, M. S., & Feltman, R. (2009). "Hot-headed" is more than an expression: The embodied representation of anger in terms of heat. *Emotion*, *9*(4), 464–477.

Trust, Communication, and Inequality

Joanna Bryson (bryson@conjugateprior.org)
Department of Computer Science, University of Bath
Bath, BA2 7AY, United Kingdom
The Center for Information Technology Policy, Princeton University
303 Sherrerd Hall, Princeton, NJ 08544, USA

Paul Rauwolf (paulrauwolf@gmail.com)
The Institute for New Economic Thinking, Oxford University
Eagle House, Walton Well Rd, Oxford OX2 6ED, United Kingdom

Abstract

Inequality in wealth is a pressing concern in many contemporary societies, where it has been show to co-occur with political polarization and policy volatility, however its causes are unclear. Here we demonstrate in a simple model where social behavior spreads through learning that inequality can covary reliably with other cooperative behavior, despite a lack of exogenous cause or deliberate coordination. In the context of simulated cultural evolution selecting for trust and cooperative exchange, we find both cooperation and inequality to be more prevalent in contexts where the same agents play both the roles of the trusting investor and the trusted investee, in contrast to the condition where these roles are divided between disjoint populations. Cooperation is more likely in contexts of high transparency about potential partners and with a high amount of partner choice; while inequality is more likely with high information but no choice in partners for those that want to invest. While not yet a full model of contemporary society, our approach holds promise for examining the causality and social contexts underlying shifts in income inequality.

Keywords: trust; cooperation; inequality; behavioral ecology; agent-based modeling; social learning; cultural evolution

Introduction

Wealth inequality is an issue of considerable social concern. Striking the appropriate balance between egalitarian and meritocratic (for various definitions of *merit*) distributions of wealth and power has been a dominant political concern since at least the time of Marx (1867). In recent decades, the world as a whole has seen wealth inequality decrease, concomitant with a substantial drop in extreme poverty (Olinto et al., 2013). However in the wealthy countries of the OECD, inequality has been increasing. The present increase in inequality has sometimes been attributed to information technology, particularly artificial intelligence (Brynjolfsson & McAfee, 2014; Ford, 2015, for reviews). Although some argue that inequality *per se* is not so much of a concern as the wealth of a society's poorest members, others argue that radical differences in income lead to a lack of shared identity and political disenfranchisement (Plato, 380BC; Atkinson, 2015). While the causality is not yet established, there is good evidence that greater inequality increases political polarization (McCarty & Shor, 2016), leads to political volatility (Dutt & Mitra, 2008), and damages economic growth (Ostry et al., 2014). Establishing causal connections between these correlates is not only of scientific interest, but would have substantial consequences for public policy. However, establishing causality via interventions is both costly and potentially ethically problematic.

Here we propose that our understanding of the causes and consequences of inequality may be advanced if we leverage the more widely explored models of the dynamics of cooperative behavior. Cooperation is now understood to be as fundamental a biological behavior as competition (Marshall, 2015). Cooperation is a strategy by which 'selfish' genes can propagate themselves, as one vector of their transmission may pay a cost to increase the probability that other vectors survive and flourish. In some circumstances, such strategies generate absolute increases in overall advantage for a collection of cooperators; however, in all circumstances at least some individuals must focus sufficiently on their own survival to persist. Consequently, increasing and decreasing social investment may be an appropriate response to changing contexts (MacLean et al., 2010; Bryson et al., 2014); though oscillations in the dominance of social vs. independent strategies have also been shown to emerge without an exogenous environmental cause (Cavaliere et al., 2012).

The work presented here demonstrates that levels of inequality can also vary and emerge with no environmental trigger and no change discrete change in policy. We extend here a previous account of trust and cooperation in human economic transactions. This previous account used an evolutionary model to demonstrate two human-like behaviors: blind trust in unfamiliar strangers, and the costly rejection of unfair but profitable offers. Here we replicate and extend that model in the context of social learning rather than biological evolution. We also show that the results hold in a more biologically plausible spatial context (where interactions are only with neighbors), then examine the dynamics of inequality under both conditions. We find persistent inequality emerging only in limited circumstances, but transient inequality emerging as part of periodic collapses in cooperative behavior. These are transient because the context of trust required fades until the behaviors resulting in inequality are again suppressed.

Background and Previous Work

The Trust Game

In this paper we explore the dynamics of trust, exchange, and inequality in the context of an established model from the behavioral economics literature: the trust game. The trust game has two players playing anonymously, and an experimenter mediating their cooperation. The first player, **the investor**, is

given money by the experimenter, then offered the option of keeping the money, or entrusting it to an anonymous other: **the investee**. Both players know that the experimenter will not simply give the money to the investee, but also multiply it by a factor b. The investee can now keep all the money, or give any amount of it back to investor. In the present paper, we focus on one-shot games, where this is the sum total of all interactions between the two anonymous players. In such a context, there is no obvious motivation for the investee to return any money, but generally they do. Similarly, there is no reason for the investor to have blind faith in this return, but many do make the investment, though often that trust is misplaced (Berg et al., 1995; Güth et al., 1997).

Rauwolf & Bryson (2016) demonstrate that evolutionary pressure can be maintained for trusting in this game in some contexts, as described below. This work extends from that of a number of other theoreticians, who have been exploring whether experimentally-observed levels of trust might be explained if we assumed that investors have either knowledge of the investee's rate of return due to reputation, or at least a population-level expectation based on experience of behavior such as might be captured by evolution (Bear & Rand, 2016). Manapat et al. (2012) have shown that selective pressure for trust can arise when there is sufficient chance of knowing a trustee's return rate. This is true even when information is delayed and inconsistent (Manapat & Rand, 2012).

Tarnita (2015) has similarly shown trust should emerge in this context, however she also exploits the mechanism of a structured environment: that is, that interactions are more likely to occur with some individuals than others. Although sometimes overlooked in mathematical theory (Sober & Wilson, 1998, for a history), in practice it takes time to traverse physical space, so nearly all interactions between organisms in the real world have this property. Further, both because offspring tend to be born near their mothers and bear family resemblance, and also social species tend to learn behavior from each other, those nearer to you are more likely to behave as you do. These properties have been shown to support both the genetic and cultural evolution of cooperation, and their ubiquity may explain the similar ubiquity of cooperation in nature (Hamilton, 1964; Marshall, 2015).

Experimental researchers have also examined the effects of offering investors in the Trust Game information about the trustees' return rate before they decide whether to invest. Investors often reject offers that would be profitable ($r > 1/b$), implying an implicit demand for fairer returns (i.e. $\approx 1/2$ Manapat et al., 2012). Rejecting an offer even when it provides a net gain for the investor again seems irrational and maladaptive, but here again Rauwolf & Bryson (2016) show that it can be advantageous. If there is sufficient competition between potential investees, then a strategy of withholding benefit from those who offer too little return on investment can pay off in the long term.

Previous Work: Methods and Results

Rauwolf & Bryson (2016) use an evolutionary agent-based model to determine whether it might be adaptive to blindly trust someone whose rate of return is unknown, even preferring them to someone who has offered a rate that is beneficial but unfair. We do this in a context of partner choice. We assume that on any given round of the game, a trustee may interact with only a subset of the population. The number of potential partners (k) ranges from 1–5. The other independent variable in the experiment is the probability of knowing (q) the the return rate (r) of any one partner. Return rates are the sole adaptive feature of the population of investees. The population of investors have two adaptive characteristics: their probability of preferring to trust a stranger (t) over keeping their money to themselves when there is no available partner with a known suitable r, and the level of r they demand (d) before entering a game with a player with known r.

Throughout this paper, the benefit of investing b is set to 3, so the investee always receives triple the money invested. We also focus on the binary version of the game: the options available to the investor are only all or nothing (as per Güth et al., 1997). The decision rule of the investor therefore is the following: An investor knows r for j out of k potential partner trustees, where $j/k = q$, the probability of knowing r. An investor invests with trustee i who has return rate r_i with probability:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq d \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x < d \end{cases} \quad (1)$$

An entrusted trustee deterministically returns the investment $\times br$ to the investor, and keeps $b(1-r)$.

A population is initially seeded with random values of r, d and t drawn from the range $[0,1]$. The entire population is born simultaneously, and all live for 500 rounds. There are 500 agents of each type. Each round every investor is offered the opportunity to invest with one of k randomly selected trustees. After 500 rounds, a new generation of investors and trustees are selected, with the probability of a set of variables persisting into the next generation being directly correlated to an agents' wealth. In line with Manapat et al. (2012), each agent is selected for the next generation using the pairwise comparison process. When an attribute is added to the next generation it is slightly mutated over a Gaussian distribution with $\mu = 0$ and $\sigma = 0.01$. After 1000 generations, we observed the populations' averages for the three dependent variables (traits.) We found that the rate of return tended to depend on both information and competition between partners. Nearly all the investment is returned under conditions where all return values are known ($q \approx 1$) and there are many partners ($k \approx 5$), whereas return rates approached 0 as q did, or when $k = 1$. In contrast, trust and demand peaked not at the extreme limits of the parameter space, but at intermediate values. Trust was only selected for in contexts of partial information, and came to very high levels when both $0.2 < q < 0.7$

and $2 \leq k \leq 5$. For higher levels of q, trust drifts neutrally because r is driven high by direct competition. As the number of partners increases, the maximum information rate q for maintaining high t declined further, since effectively more partners provide more knowledge of return rate. The shape of the curve for demand was similar, although demand was always in the range $0.333 < d < 0.6$. The floor is set by b; any return lower than b means investments lose money. Again, where r was driven high by competition (high q), or where there is no information (low q), d drifts unused. Selection for the costly rejection of partners only occurred in the same situations as trust.

Methods: Replication and Extensions

We have produced a new ABM to replicate and extend our previous results. In the present paper, we wished to examine four further questions:

1. *Do the results hold for social learning?* Here we assumed a stable but not particularly cognitive population that learned socially, rather than evolved. Rather than learning from its own experience (which may be useless in a constantly shifting environment) the agents learn from each other. After each round of the game, agents copy the parameters of their potential trading partners if those partners have more money than they do.

2. *Does spatial structure alter the results?* We added location in an x,y grid as a parameter of the agents. In the spatial, local condition, partners were drawn from adjacent neighbors. We also ran a condition without spatial structure to confirm the replication. As reviewed above, space is expected to facilitate cooperation, and also diversity.

3. *What is the impact of investment castes?* Previous models had investors and investees in disjoint populations. We ran conditions assuming the same, and alternative conditions assuming that all individuals could invest, and any individual might be called upon to be an investee. We expected the unified population to show greater, more stable cooperation due to mutualism, as per Estrela et al. (2015).

4. *In what conditions if any will inequality emerge, and will it be stable?* This was our principle motivation for this study. We anticipated that extremely high rates of return (r) or demand (d) might generate inequality.

Simulations were run in NetLogo on a 33x33 grid, so with 1089 locations, each with one either a single agent, or in the situation where investors and investees were different castes, two agents, but these agents did not interact. The environment provided each agent with one unit of currency per round. The agents invested all of their wealth if they chose to invest, so the results could be cumulative; however this rapidly exceeded the arithmetic capacity of NetLogo, so all agents were also taxed at a flat rate every round. For spatial models, the world was assumed to be toroidal, so all agents had 8 neighbors, of which k would be drawn randomly on each round as potential investees.

Results

Our results show the following:

1. The dynamics of learning are very much like the dynamics of evolution—the outcomes are certainly comparable. This is assuming that social learning occurs implicitly or at least automatically. Fig. 1, row 1 replicates Rauwolf & Bryson (2016) most closely in both experimental condition and outcome.

2. Spatial assortment also had surprisingly little impact (Fig. 1, top two rows vs. bottom two).

3. The most striking result of our replication was the difference between a unified population of investors and investees (rows 2 and 4 of Fig. 1) vs. the original disjoint castes (rows 1 and 3). The unified population had a far greater range of values for which the return rate r was high. This result also clarified the contexts affording high levels of demand. Where there is high information (q), r is driven by pure competition and demand (d) drifts sufficiently low that it has not impact. When the populations are separate, for low q, r collapses and d again has no impact. But in the case where the traits are bound in a single population, there is sufficient selective pressure on d to keep costly punishment high even under low information. Interestingly, trust (t) and d seem to complement each other at least when r is sufficiently high to merit trust.

4. The only conditions that afforded substantial inequality were those with the single population (Fig. 2). The top 1% of the population ordinarily had between 1–1.5% of the wealth at least among the investors, although the investee-only caste suffered greater differentiation where there was either high information and no choice or low (but not zero) information and choice. But the greatest inequality, where the top 1% of the population had as much as 50% of the wealth, occurred only when investors and investees were the same individuals, and then only in conditions of high information ($q > 0.5$) and no partner choice ($k = 1$).

5. In the conditions best replicating human performance on the trust game, inequality surges periodically correlated with a drop in return rate. There is no stable equilibrium for cooperation in any of these models, but rather the dynamics are such that cooperation periodically erodes but then recovers. In our model, inequality was a characteristic result of these periods of collapse (Figure 3).

Discussion and Future Work

There is a growing body of evidence has shown that inequality and polarization have tracked each other closely in the USA since at least the 1880s, well before the onset of IT or

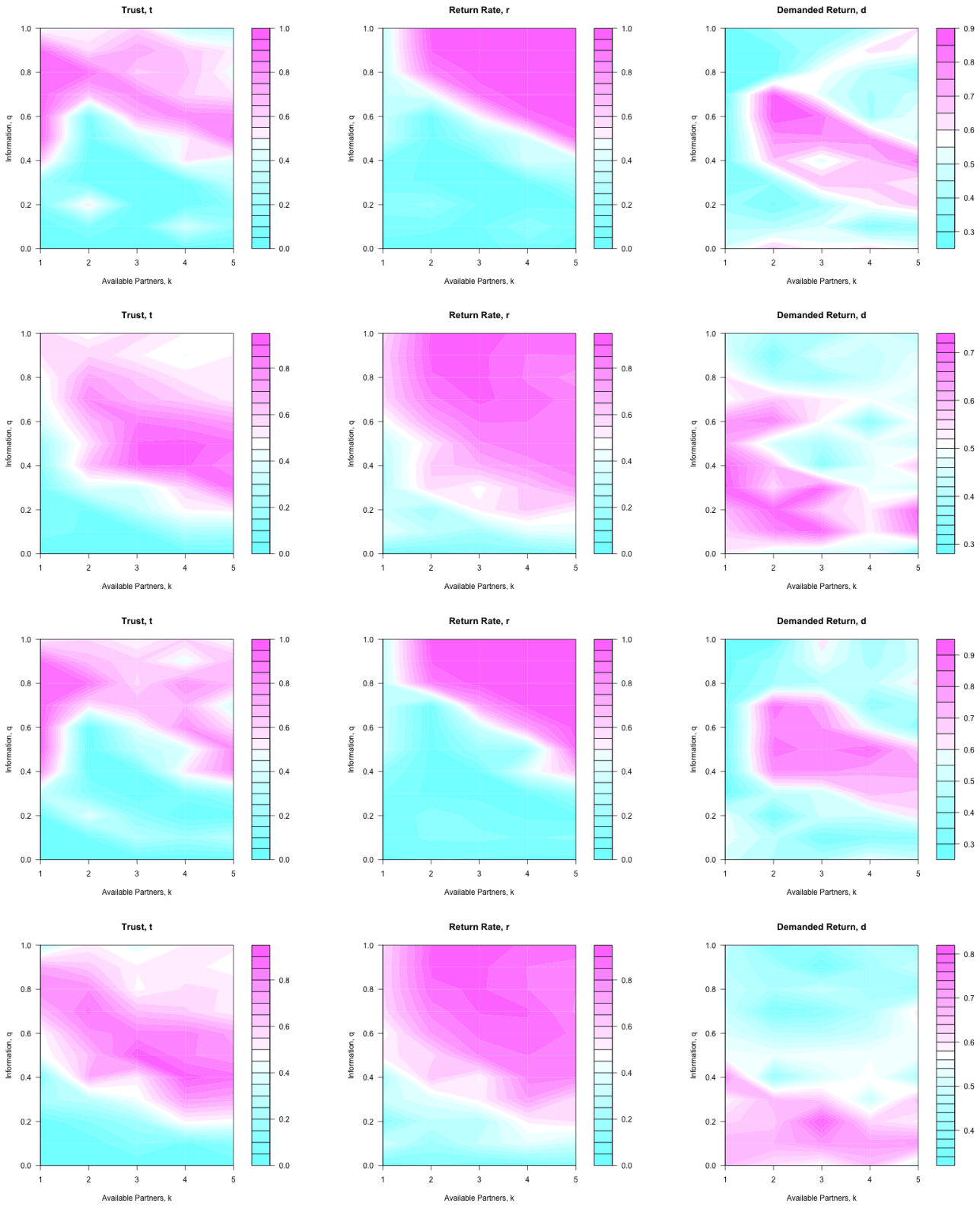


Figure 1: Trust, rate of return, and demand for the four conditions. The top two rows partners are randomly selected from the full population, the bottom two rows they are only selected from adjacent neighbors. The top and third row the populations of investors and investees are disjoint, the second and bottom row all agents play both roles. Values represent averages over $N = 1089$ agents for the final 300 time cycles (of 1200), for each of 5 runs. Note that index colors do *not* indicate consistent ranges, but rather vary by subfigure—see keys for values. See main text for interpretation.

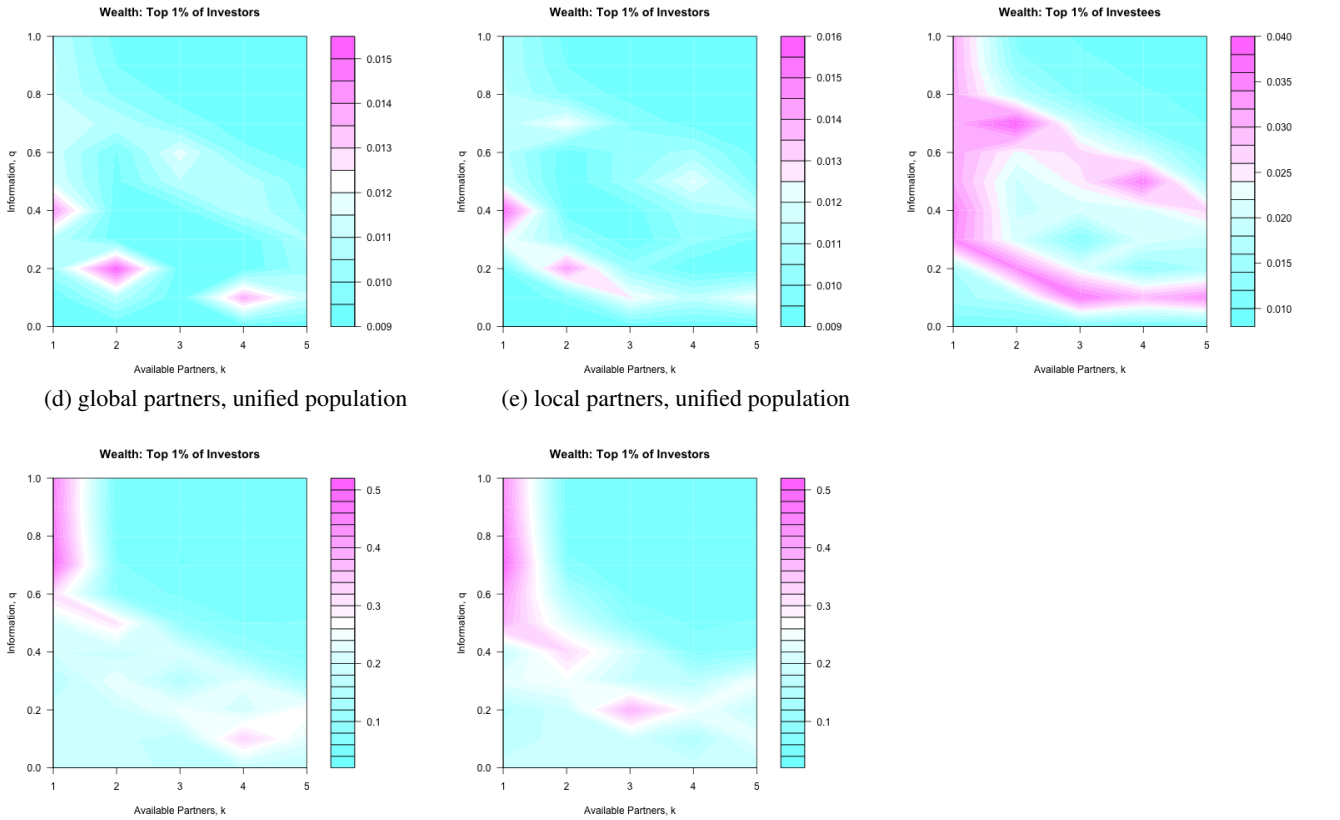


Figure 2: Inequality as measured by the proportion of wealth held by the top 1% of investors for each of the four conditions; values averaged as per Figure 1. Note that index color ranges are *not* consistent across subfigures, see keys. The only condition of extreme inequality occurs when investors and investees are drawn from the same population, there is no partner choice (although agents can still refuse to invest at all), and relatively high levels of information are available about potential investees. Note that localised investment reduces the required level of information to generate inequality.

AI. Both inequality and political polarization peaked immediately before and after World War I, then plummeted during the Great Depression (McCarty & Shor, 2016). Though now viewed as a norm, the long flat trough of both inequality and polarization between WWII and the 1980s may actually be the aberration, as ordinary oscillations in altruism such as were described by Cavaliere et al. (2012) were held in check by financial policy (Bryson & McCarty, 2016).

The results shown here are preliminary, but open the way to time-series analysis which may afford a better understanding of causality in this model, which can then be checked for match to the existing political and economic data. Certainly the dynamics of these models fluctuate greatly: there is no stable equilibrium, but rather tendencies for cooperation which occasionally are compromised for local profit. Creating better correlation to human society will probably also require modeling the stickiness of institutions.

Acknowledgments

Thanks to Will Lowe for assistance with the figures, and Nolan McCarty for discussions on inequality and polarisation. Rauwolf would like to acknowledge partial funding support through a University of Bath fees-waived PhD studentship. Bryson would like to acknowledge the Universities of Bath and Princeton for sabbatical support.

References

Atkinson, A. B. (2015). *Inequality: What can be done?* Harvard University Press.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, *113*(4), 936–941.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

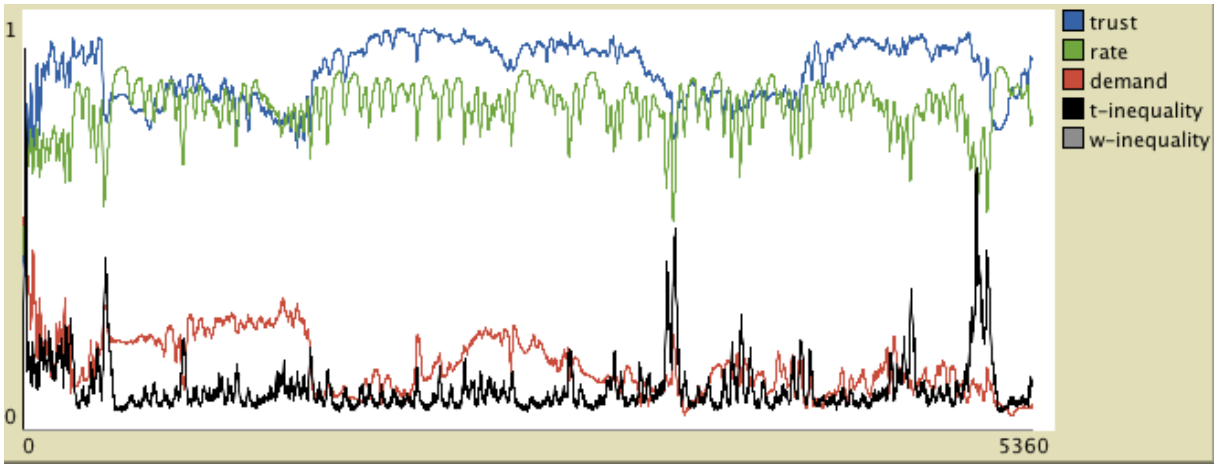


Figure 3: These simulations do not reach stable equilibria, but rather evolve dynamics that persistently converge in ways determined by the socio-economic context. In this exemplar, chosen because of the correspondence to human levels of trust and demand, the context is: the probability $q = 0.5$ of knowing a potential partner's return rate (r), and the number of available partners to choose between per game $k = 3$; investees are drawn from the same population as the investors (no castes); and they are always drawn from the possible 8 immediate neighbors (spatially localized exchanges). The measured values are investee return rate r, investor demand d and trust t—these are population averages ($N = 1089$); inequality is the proportion of overall wealth held by the richest 1% of agents.

Bryson, J. J., & McCarty, N. (2016). *Polarization and inequality: Towards a mechanistic account.* (in prep.)

Bryson, J. J., Mitchell, J., Powers, S. T., & Sylwester, K. (2014). Understanding and addressing cultural variation in costly antisocial punishment. In M. A. Gibson & D. W. Lawson (Eds.), *Applied evolutionary anthropology: Darwinian approaches to contemporary world issues* (pp. 201–222). Heidelberg: Springer.

Cavaliere, M., Sedwards, S., Tarnita, C. E., Nowak, M. A., & Csikász-Nagy, A. (2012). Prosperity is associated with instability in dynamical networks. *Journal of Theoretical Biology*, *299*(0), 126–138. doi: 10.1016/j.jtbi.2011.09.005

Dutt, P., & Mitra, D. (2008). Inequality and the instability of polity and policy. *The Economic Journal*, *118*(531), 1285–1314. doi: 10.1111/j.1468-0297.2008.02170.x

Estrela, S., Morris, J. J., & Kerr, B. (2015). Private benefits and metabolic conflicts shape the emergence of microbial interdependencies. *Environmental Microbiology*, n/a–n/a. doi: 10.1111/1462-2920.13028

Ford, M. (2015). *Rise of the robots: Technology and the threat of a jobless future*. Basic Books.

Güth, W., Ockenfels, P., & Wendel, M. (1997). Cooperation based on trust. an experimental investigation. *Journal of Economic Psychology*, *18*(1), 15–43.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*, 1–52.

MacLean, R. C., Fuentes-Hernandez, A., Greig, D., Hurst, L. D., & Gudelj, I. (2010, 09). A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biol*, *8*(9), e1000486. doi: 10.1371/journal.pbio.1000486

Manapat, M. L., Nowak, M. A., & Rand, D. G. (2012). Information, irrationality, and the evolution of trust. *Journal of Economic Behavior & Organization*.

Manapat, M. L., & Rand, D. G. (2012). Delayed and inconsistent information and the evolution of trust. *Dynamic Games and Applications*, *2*(4), 401–410.

Marshall, J. A. R. (2015). *Social evolution and inclusive fitness theory: An introduction*. Princeton University Press.

Marx, K. (1867). *Das kapital: Kritik der politischen oekonomie* (Vol. I). Hamburg: Otto Meissner.

McCarty, N., & Shor, B. (2016). *Partisan polarization in the united states: Diagnoses and avenues for reform.* (Available at SSRN 2714013)

Olinto, P., Beegle, K., Sobrado, C., & Uematsu, H. (2013). The state of the poor: Where are the poor, where is extreme poverty harder to end, and what is the current profile of the world's poor? *World Bank: Economic Premise*(125), 1–8.

Ostry, J. D., Berg, A., & Tsangarides, C. G. (2014). *Redistribution, inequality, and growth*. International Monetary Fund.

Plato. (380BC). *Republic*. Thrasyllus of Mendes.

Rauwolf, P., & Bryson, J. J. (2016). *Trust mediates costly punishment in environments of partial information and partner choice.* (under review)

Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Tarnita, C. E. (2015). Fairness and trust in structured populations. *Games*, *6*(3), 214. doi: 10.3390/g6030214

What Were They Thinking? Diagnostic Coding of Conceptual Errors in a Mathematics Learning Software Data Archive

Christine M. Massey (massey@seas.upenn.edu)
Jennifer D. Kregor (kregor@sas.upenn.edu)
Laura E. Cosgrove (lacos@sas.upenn.edu)
University of Pennsylvania, Institute for Research in Cognitive Science
3401 Walnut Street, Suite 400A, Philadelphia, PA 19104-6228

Himchan Lee (himchan@gse.upenn.edu)
University of Pennsylvania, Graduate School of Education
3700 Walnut Street, Philadelphia, PA 19104-6216

Abstract

Decades of research have demonstrated that students face critical conceptual challenges in learning mathematics. As new adaptive learning technologies become ubiquitous in education, they bring opportunities both to facilitate conceptual development in more focused ways and to gather data that may yield new insights into students' learning processes. The present study analyzes data archives from a perceptual learning intervention designed to help students master key concepts related to linear measurement and fractions. Using algorithmic data coding on a database of 78,034 errors from a sample of sixth graders, both conceptual errors and other errors were captured and analyzed for change over time. Results indicate that conceptual errors decreased significantly. This approach suggests additional ways that such datasets can be exploited to better understand how the software impacts different students and how next generations of adaptive software may be designed to code and respond to common error patterns in real time.

Keywords: adaptive learning; conceptual development; educational software; learning technology; mathematics cognition; perceptual learning

Introduction

Prior research on students' conceptual development in mathematics has identified a number of areas in which students persistently make characteristic conceptual errors that are often resistant to standard instruction or procedural practice (National Research Council (NRC), 2001; Vosniadou & Verschaffel, 2004). The present study investigates types of student errors and patterns of change in their performance over time using data archives from sixth graders interacting with an educational software module that was explicitly designed to address such conceptual errors in the domain of units of linear measurement on rulers with both integer and fractional subdivisions. These data were generated as part of a large randomized control trial (RCT) by students in 30 classrooms that were randomly assigned to an intervention condition that used four perceptual learning software modules (Kellman & Massey, 2013) focused on fractions and measurement over the course of their sixth grade year (Kellman, Massey & Son, 2010). Data reported here are from the Linear Measurement Perceptual Learning Module (PLM). Separately reported data from this ongoing study indicate that the PLM intervention demonstrates significant learning gains compared to a control condition (Bowden, Massey & Kregor, 2015; Scull, 2015). For the content of interest in the present analysis, HLM analyses from the RCT indicate a significant treatment effect of the PLM condition, replicated across two cohorts, on a test consisting of multiple choice and open-ended items focused on various aspects of linear measurement drawn from large-scale standardized assessments. The PLM is hypothesized to promote students' conceptual understanding by enabling them to recognize the specific properties of units used to measure continuous extents, to apprehend how whole units and fractional parts of units are represented and enumerated on rulers, and to overcome tendencies to inappropriately apply schemes for counting discrete objects to linear measurement. The present study evaluates this mechanism by examining whether students made the types of conceptual errors that would be anticipated based on the existing research literature on conceptual development for linear measurement and fractions, and, if so, whether the software was effective in helping students overcome known error patterns and move to correct responses.

While this analysis shares some general goals with approaches used in educational data mining and learning analytics (Siemens & Baker, 2010), rather than using machine learning techniques and automated algorithms to discover patterns in responses or to model students and predict responses, the study instead uses algorithmic coding to classify error types predicted from the research literature. Whereas prior cognitive studies of conceptual change in mathematics—particularly microgenetic studies examining learning over time—have generally involved intensive study of relatively small numbers of students interacting with a constrained set of tasks or problems, the current study allows us to examine and code detailed records from 716 sixth graders, each of whom completed an average of 215 open-ended interactive problems over the course of multiple weeks. A total of 157,147 completed problems yielded a pool of 78,034 errors for analysis.

Conceptual Challenges in Linear Measurement

Two areas in which U.S. elementary students perform particularly poorly are fractions and measurement (National

Mathematics Advisory Panel, 2008). Research studies indicate that many students do not recognize that units of linear measurement must have continuous extent, and they instead impose discrete counting schemes on ruler measurement, counting numbered hash marks rather than the intervals between marks (Bragg & Outhred, 2004); Mitchell & Horne, 2008). A familiar result of this misunderstanding is that many students are baffled as to why rulers do not begin at "1." Students also demonstrate consistent errors when measuring with "broken" or partial rulers. Other conceptual difficulties include failing to distinguish between position and distance on a ruler or number line, and not understanding how fractions are represented by subdivisions of units (Ball, 1993; Bright, Behr, Post & Wachsmuth, 1988; Lehrer, Jaslow & Curtis, 2003; NRC, 2001). Also challenging are mapping mixed numbers to rulers and reconciling multiple labels for the same point (e.g., 2/4 and 4/8). Students typically learn a standard computational procedure for converting mixed numbers to improper fractions, but they often lack the ability to flexibly regroup fractions and whole numbers, and, in the context of relating mixed numbers to positions and distances on rulers, the computational procedure may not be productive. The Linear Measurement PLM was specifically designed to address these conceptual difficulties, using a perceptual learning approach in which students directly interact with the targeted structures, relations, and representations across a large and varied set of problems with customized animated feedback on every trial.

Perceptual Learning Software

Perceptual learning (PL) refers to a process by which individuals improve their ability to accurately and fluently extract information coming from the environment in some domain (Gibson, 1969; Kellman & Massey, 2013). This efficient pick-up of information characterizes experts, who selectively attend to relevant features, recognize meaningful patterns, extract higher-order relational structure, and ignore irrelevant variation. Typically, PL occurs through repeated experience discriminating and classifying a wide variety of instances as one engages in a given activity. Recent research (Kellman, Massey & Son, 2010) has demonstrated that principles of perceptual learning can be incorporated into learning software and used to accelerate fluent, expert-like information pick-up in academic symbolic domains such as mathematics and chemistry. Although the term "perceptual" may seem to contrast with conceptual understanding (Kellman & Massey, 2013), in fact, the fluent apprehension of fundamental structures and relationships is often a critical foundation for conceptual understanding. In the present work, PL training is aimed at improving learners' understanding of the structure of whole and fractional measurement units and invariant patterns in how they are represented on rulers and on number lines more generally.

The graphic interface for this PLM consists of an interactive display showing a ball on top of a ruler, as illustrated in Figure 1, which provides examples of a simple integer problem and a more complex fraction problem. Information given at the top of the screen identifies the ball's starting point and then gives either the distance the ball should move and asks the student to indicate the endpoint, or gives the endpoint and asks the student to input the distance it would travel. When the student enters a response, the ball carries out the action, followed by animated feedback indicating whether the response was correct, and, if not, showing how the correct answer compares. On each learning trial, the student sees a unique

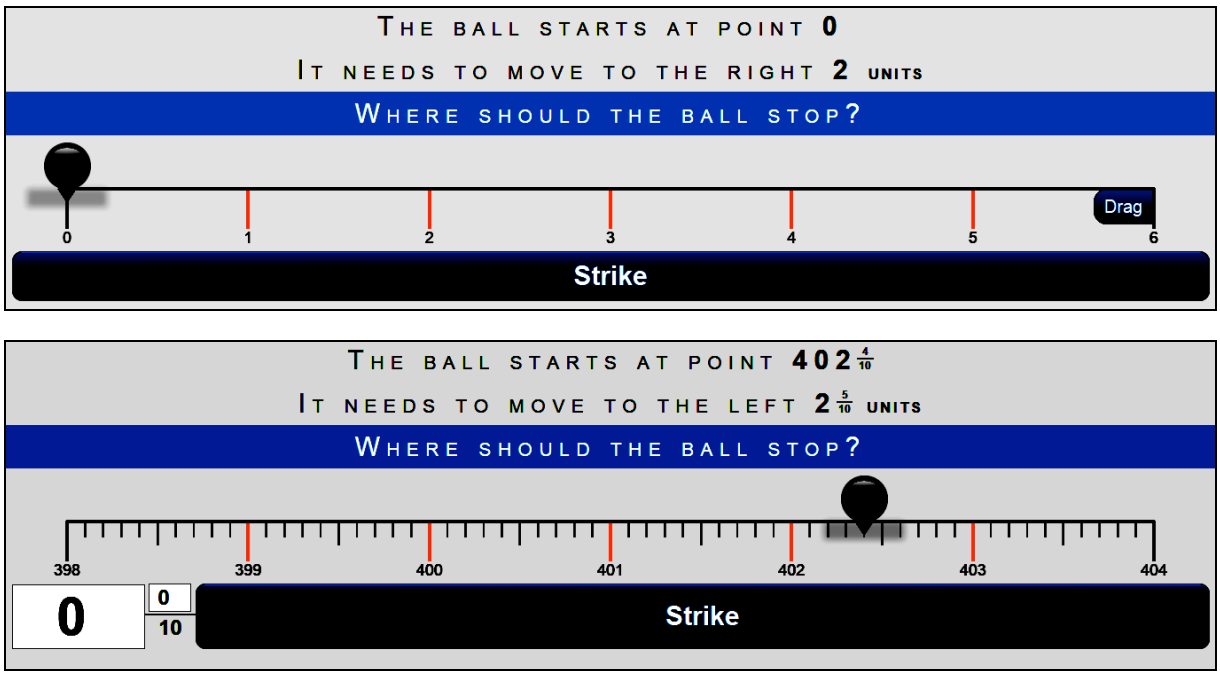


Figure 1: Examples of a simple integer problem (top) and a difficult fraction problem (below).

problem drawn from a very large set of problems organized into eight subtypes, based on whether the problems involve fractions or only integers, whether users enter their responses by moving markers on the ruler or by typing in numerical values, and by whether the unknown in the problem is the final endpoint or the distance traveled on the ruler. Half of the categories consist of simpler integer problems and half are more difficult fraction problems. (Thus one category would be fraction problems on which the distance traveled is given and the user types in the endpoint.) Other problem variations that cut across these 8 categories include whether the ruler is fully or partially labeled; whether the start/end point is 0, 1, or some other point on the ruler (including values in the hundreds); whether the direction of movement is to the right (addition) or to the left (subtraction); and whether the ruler is over-partitioned or congruently partitioned for the units given in the problem.

The software automatically captures time-stamped data, recording every problem seen, the response entered, and the response time (or time-out if no response is entered within 90 seconds). The resulting dataset was analyzed to examine (a) the frequency with which students made the specific conceptual errors anticipated from the research literature on measurement and fractions, (b) what other kinds of errors students made, and (c) whether and how error rates changed as students used the software.

Method

Subjects

Participants in this study were 716 sixth graders in 30 classrooms in schools in a large Northeastern city serving predominantly low-income and minority students. To be included in the analysis, each student had to complete at least 20 problems using the Linear Measurement PLM but did not have to complete the entire PLM. Students used the web-based software during the school day as part of their normal mathematics curriculum.

Procedure

Each unique problem in the software database can be deconstructed into a set of problem parameters. To code errors algorithmically in the large set of student data from participants in this study, we used the problem parameters associated with each problem to create algorithms that operationally define a set of targeted conceptual errors, with a particular focus on miscounting of hash marks and regrouping errors involving fractions. We also developed algorithmic codes for errors related to how students encoded the problems and interacted with the software interface. All codes were built using STATA. Descriptions of the error categories and how they were coded are given below.

Hash Mark Errors Hash Mark errors occur when students focus on discrete hash marks as the unit of measurement and count them starting from one, resulting in answers that are systematically wrong by one. Students can make similar errors with fractional parts of units if they count secondary hash marks on a ruler or number line the same way. When fractions are involved, a student may make the hash mark error only on the integer hash marks, on both the integer hash marks and the fraction hash marks, or only on the fraction hash marks. We designated the first case, along with hash mark errors made on integer problems, as Hash Mark Integer errors, and the latter two cases as Hash Mark Fraction errors.

Regrouping Errors By design, many of the fraction problems in the learning set involve redistributing fractional units from or into integer units across an integer boundary (e.g., the bottom problem in Figure 1.) For right-going Endpoint Unknown problems, this occurs when the sum of the fractional units is greater than "1"; for left-going Endpoint Unknown problems and all Distance Unknown problems, which require subtraction, this occurs when the fraction to be subtracted is greater than the fraction from which it is subtracted. Students often made a characteristic Regrouping error when confronted with boundary-crossing problems. Figure 2 illustrates several examples of such errors on a typical problem. A student would use the correct numerical operation for the integers in the mixed numbers, but use any number of different strategies to deal with the fraction parts: reversing the place of the fractions in order to avoid subtracting the larger from the smaller; answering with either of the given fractions and ignoring the other; or ignoring the fractions entirely. The result for any one of these strategies is nevertheless predictable: an integer answer that is, correctly, the sum or difference of the given integer values but with some incorrect fraction or no fraction appended. This conceptual difficulty is analogous to well-known "buggy algorithms" involving borrowing errors across place value columns in multi-digit arithmetic and in mixed-number subtraction (Brown & Burton, 1978; Fuson, 1990; Scott, 1962). In each case, students fail to process the relational structure of adjacent place values or of fractional and integer units.

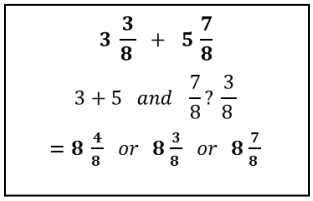


Figure 2: Schematic illustration of possible process for regrouping errors on a problem with a given start point of 3 ⅜ and distance traveled of 5 ⅞.

Problem Encoding Errors In addition to algorithms to capture the targeted conceptual errors described above, algorithms were also developed to capture errors that are specific to the Linear Measurement PLM problem presentation interface. As is often the case in mathematics problem solving, students do not always accurately encode the problem structure in terms of what information is given

and what is to be found. All problems in the PLM involved a triad of a start point, an endpoint, and a distance traveled. The start point was always one of the given values, while endpoint and distance varied between given and unknown roles. One type of problem encoding error occurred when students confused whether distance or endpoint was the unknown. A second type of encoding error occurred when students did not correctly encode the direction of travel (e.g., answered as if the ball moved rightward when the problem specified that it moved to the left).

Given Information Errors Responses were coded as Given Information errors when students entered one of the given values as their answer. Students might do this as a default response when they cannot process the problem structure (similar to "number grabbing" that has been observed when students solve word problems (Bell, Greer, Grimison & Mangan, 1989; NRC, 2001)), or this kind of error may represent a type of disengaged response in which students enter a given value just to enter something.

Unproductive Responses Unproductive Response errors were coded when students pressed "Enter" without giving an answer, timed out without entering any answer, or entered a value that was out of range for the given problem.

Parameter data for each problem were used to create a general code for each error type that would be applicable for all problems or for all problems within a particular subset. For example, a directionality error variable was defined if, for left-going Endpoint Unknown problems, the student's answer for the endpoint is equal to the start point plus the distance. Not all types of errors are applicable for every type of problem, and so the parameter data were used to narrow the test space for particular error codes (e.g. Regrouping errors were only tested on boundary-crossing problems). It is important to note that for nearly all the error codes, the student's answer was flagged only if it corresponded to the answer that would be given if only the named error were made. That is, a student could have concatenated multiple errors—e.g., a Directionality error and a Hash Mark error—and this would not be captured by the error code. Given the risk of miscoding responses unrelated to a complex error combination when operating on a small answer space, however, we chose to avoid concatenating errors.

Results

Out of 78,034 total errors, 38,337 (49.1%) were coded as belonging to a single well-specified error category. An additional 15,753 errors (20.2%) were captured by more than one error code, since the same error could have been made by more than one reasoning process. Since coding of these errors is inherently ambiguous, we removed them for the remainder of the analysis. Errors that were not captured by the algorithmic codes are not considered further in this analysis. Because not all errors can occur on every type of problem, analyses below indicate when the reported frequencies are out of the total of eligible problems rather than all problems.

Table 1 shows the number of students achieving each mastery level. Just over half of the students (52.8%) mastered the entire module, and 60.5% mastered at least 6 of the 8 categories. (To master a category, a student had to complete at least 4 of the most recent 5 problems of that type correctly.)

Table 1: Mastery level by number of students

Mastery Level	N Students	% Students
0	23	3.2%
1	15	2.1%
2	25	3.5%
3	30	4.2%
4	106	14.8%
5	84	11.7%
6	44	6.2%
7	11	1.5%
8	378	52.8%

(Total N of Students = 716)

Table 2 shows the frequency and percentage of each of the captured error types as well as the number of students who made each type of error at least 5 times. As the table indicates, just about half of all errors committed were uniquely captured by the individual codes specified above. Approximately one-fifth of the total errors were the anticipated conceptual errors related to regrouping and to misreading hash marks. Errors captured by multiple codes are not included.

Table 2: Frequency of captured error types

Error	Total Errors Coded	% of Errors All Students	N Students with error 5+ times
Unproductive Responses	11,080	14.2%	479
Given Information	9,588	12.3%	463
Regrouping	8,363	10.7%	464
Hashmark Fraction	5,073	6.5%	382
Problem Encoding	2,533	3.2%	193
Hashmark Integer	1,700	2.2%	111
Total Errors/Total N	78,034		716

Total Problems = 157, 147

To examine changes in the rates of various error types over time, each student's time-ordered sequence of trials was divided into ten phases, from early trials through late trials. Since students completed different numbers of trials, the number of trials falling within each phase is relative to the individual student.

As Figure 3 illustrates, learners typically make steady progress through the PLM, accumulating up to 8 mastery levels as they reach mastery criteria for each category

(typically mastering the easier integer categories first). Figure 3 also shows a distinctive U-shaped curve for average accuracy across time phases. Average accuracy starts at around 68%, as the PLM begins with the simplest integer problems first, and then drops to a low near 50% during the middle of training (phases 5-7), before climbing back up. The steep drop coincides with the appearance of more difficult problems and persists as the easiest problem categories are being retired, which results in students' practice being adaptively focused on more difficult categories. In the last third of training, accuracy again increases as performance improves on harder categories.

Figure 3: Average Mastery Level and Accuracy during Training over Relative Time Phase

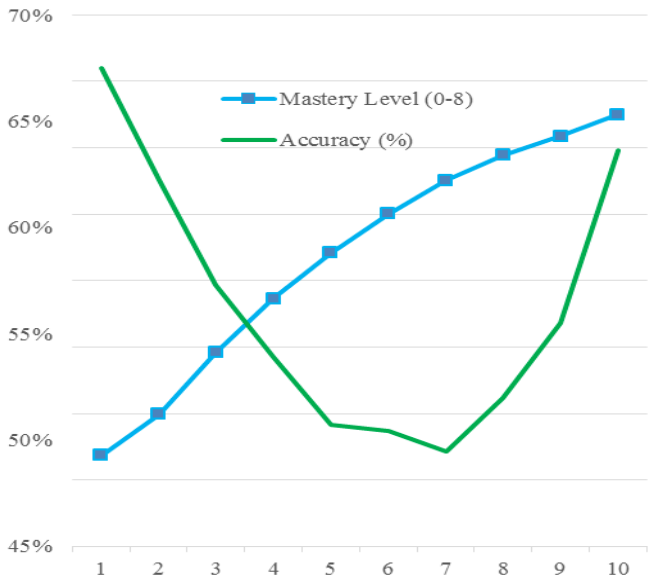


Figure 4: Average Error Rate on Relevant Problems over Relative Time Phase

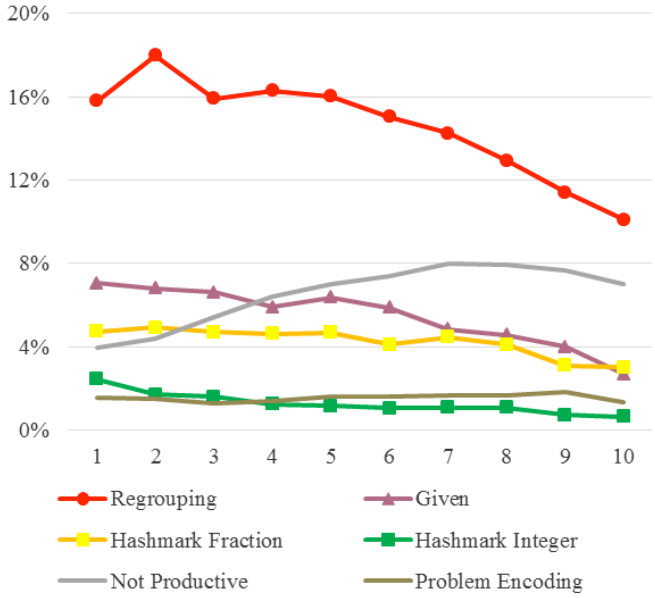


Figure 4 compares the proportion of errors made by each student at each phase of learning, averaged across all students. Regrouping errors showed the highest rate (relative to eligible problems) in all phases, and they decreased steadily in phases 5-10. Notably, the decline in Regrouping errors coincided with an increase in average accuracy and mastery levels across phases 7-10. Both Hash Mark Fraction and Hash Mark Integer errors, which were relatively less common, also decreased across phases. Given Information errors decreased over time, again, with a sharper drop in the later phases. Unproductive Response errors showed a different pattern, initially increasing and then leveling off during the phases in which conceptual errors were declining and correct responses were increasing most rapidly. Problem Encoding errors were relatively uncommon and remained steady across phases.

Repeated Measure ANOVAs were run on each type of error rate to examine mean error rates (averaged across students) across relative time phase. There was a significant effect for nearly all captured errors (using a Huynh-Feldt adjustment for sphericity). Regrouping, Hash Mark Integer, Hash Mark Fraction, and Given Information errors all decreased significantly across phases ($p < .0001$ in all cases). Unproductive Response errors increased significantly across phases ($p < .0001$). Problem Encoding errors did not vary significantly ($p = .07$). Paired t-tests comparing mean error rate at initial and final phases also demonstrated significant results ($p < .0001$) for all error rates except Problem Encoding errors ($p = .62$).

Discussion and Future Directions

The error analyses presented indicate that the Linear Measurement PLM was successful in mitigating several of the specific conceptual errors it was designed to address. Regrouping errors and errors that involved counting hash marks rather than intervals for both integer and fraction rulers declined significantly as students used the software. Most students mastered most or all of the subcategories in the learning set, including fairly difficult fraction problems that required them to be able to flexibly partition and repartition integers and fractions with varying denominators and to add and subtract fractions and mixed numbers. Given that the problems were intentionally varied and required open-ended responses, it is unlikely that students accomplished this formulaically, without gaining some genuine insight into the structure of linear units of measurement and fractions. Indeed, coming to recognize essential structures and relations across novel instances is a hallmark of perceptual learning (Kellman & Massey, 2013).

The targeted conceptual errors showed a distinctive pattern of decline over time, while other types of errors, such as time-outs and unproductive responses, increased during the first half of training. While it may seem paradoxical that some types of errors would increase, this is at least in part a result of the adaptive nature of the software. As students retire certain categories, up until the time all categories are retired, the problems they are seeing come

from not-yet mastered categories and generally become more difficult as students proceed through the module. Thus the pattern of progressive decreases in the targeted conceptual errors indicates that the software was selectively helping learners resolve these conceptual issues.

The methods used in these analyses have significant potential to be extended in ways that would further illuminate students' learning with this kind of adaptive software. Future extensions of this approach will examine the as yet uncaptured errors in the dataset to see if there are more error types that could be coded with well-defined algorithms. Future analyses can also go beyond patterns averaged across students to examine patterns for individual learners or particular subgroups of learners. As the RCT that generated the present dataset proceeds, student-level co-variate data, including demographic data and scores on standardized state tests and on an aligned mathematics test, will become available, which will enable these more detailed explorations. Additional analyses can also investigate at a finer grain how error types interact with specific subtypes or features of problems. While the present analyses have focused particularly on conceptual errors, since that is what the software was primarily designed to address, error data might be analyzed from other points of view. For example, Unproductive Response errors could index disengagement or other motivational or attentional issues for some students. Finally, findings from error analyses of large data archives can be a powerful input to the design process to create new generations of software that are more adaptive in classifying errors in real-time and responding to them in more differentiated ways.

Acknowledgments

We gratefully acknowledge expert assistance from Tim Burke and support from IES, US Department of Education through Grants R305A120288 and R305H06070 to UCLA and the University of Pennsylvania. The opinions expressed are those of the authors and do not represent the views of the US Department of Education.

References

Ball, D. L. (1993). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. In T. P. Carpenter, E. Fennema, & T. A. Ronberg (Eds.). *Rational numbers: An integration of research*. Hillsdale, NJ: Lawrence Erlbaum.

Bell, A., Greer, B., Grimison, L., & Mangan, C. (1989). Children's performance on multiplicative word problems: Elements of a descriptive theory. *Journal for Research in Mathematics Education, 20*(5), 434-449.

Bowden, J., Massey, C. M., Kregor, J. D. (April, 2015). *What predicts successful use and completion of an adaptive mathematics software intervention?* Paper presented at the 2015 Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Bragg, P. & Outhred, L. (2004). A measure of rulers – the importance of units in a measure. *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, 2*, 159-166.

Bright, G. W., Behr, M. J., Post, T. R. & Wachsmuth, I. (1988). Identifying fractions on number lines. *Journal for Research in Mathematics Education, 19*(3), 215-232.

Brown, J. S. & Burton, R. R (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Fuson, K. (1990). Issues in place-value and multidigit addition and subtraction teaching and learning. *Journal for Research in Mathematics Education, 21*(4), 273-280.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* New York: Prentice-Hall.

Kellman, P. J. & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *The Psychology of Learning and Motivation, Vol. 58*, 117-165. New York: Elsevier.

Kellman, P.J., Massey, C.M & Son, J. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, *Special issue on Perceptual Learning, 2(*2), 285-305.

Lehrer, R., Jaslow, L. & Curtis, C. (2003). Developing an understanding of measurement in the elementary grades. In D. H. Clements & G. Bright (Eds.), *Learning and teaching measurement, 2003 yearbook*. Washington, DC: National Council of Teachers of Mathematics.

Mitchell, A. & Horne, M. (2008). Fraction number line tasks and the additivity concept of length measurement. In M. Goos, R. Brown & K. Markar (Eds.). *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 353-360). MERGA.

National Mathematics Advisory Panel (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council, (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, & B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Scott, L. (1962). Children's concept of scale and the subtraction of fractions. *The Arithmetic Teacher, 9*(3), 115-118.

Scull, J. (April, 2015). *Perceptual learning technology in sixth-grade mathematics education.* Poster presented at the 2015 Annual Meeting of the American Educational Research Association, Chicago, Illinois.

Siemens, G. & Baker, R.S.J.d. (2010). Learning analytics and educational data mining: Toward communication and collaboration. *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

Vosniadou, S. & Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. *Learning and Instruction, 14*(5), 445-451.

Generalisable patterns of gesture distinguish semantic categories in communication without language

Gerardo Ortega (gerardo.ortega@mpi.nl) &
Aslı Özyürek (asli.ozyurek@mpi.nl)

Centre for Language Studies, Radboud University and
Max Planck Institute for Psycholinguistics,
Wundtlaan 1, 6525XD Nijmegen, The Netherlands

Abstract

There is a long-standing assumption that gestural forms are geared by a set of modes of representation (acting, representing, drawing, moulding) with each technique expressing speakers' focus of attention on specific aspects of referents (Müller, 2013). Beyond different taxonomies describing the modes of representation, it remains unclear what factors motivate certain depicting techniques over others. Results from a pantomime generation task show that pantomimes are not entirely idiosyncratic but rather follow generalisable patterns constrained by their semantic category. We show that a) specific modes of representations are preferred for certain objects (acting for manipulable objects and drawing for non-manipulable objects); and b) that use and ordering of deictics and modes of representation operate in tandem to distinguish between semantically related concepts (e.g., "to drink" vs "mug"). This study provides yet more evidence that our ability to communicate through silent gesture reveals systematic ways to describe events and objects around us.

Keywords: pantomime, gesture, action/object distinction, modes of representation, iconicity

Introduction

Speakers have at their disposal several strategies to represent a referent with their gestures. If referring to a glass, for example, a speaker may choose to produce a gesture representing how it should be held, may describe its outline or perhaps would depict its cylindrical volume in a three-dimensional space. Despite most research focusing on the relationship between gesture and speech, the mechanisms responsible for the specific form that iconic gestures adopt remains largely unknown. It is unclear for example whether the physical characteristics of the referent may play a role in gestural production and also whether people develop systematic strategies to make distinctions between different semantic categories (i.e., between actions and objects).

Research investigating the mechanisms responsible for gestural production is paramount to better understand the cognitive system that allows efficient communication through the manual channel. A powerful tool to do so is by investigating communication in the absence of speech. By exploring individual's communicative strategies through silent gestures, we open a window into humans' internal representations as well as into our capacity to convey information in the most effective way. Studies have shown that after stripping communication from a conventionalised language, speakers of linguistically diverse languages converge in the same strategies to express information through their silent gestures (Goldin-Meadow, So, Ozyürek, & Mylander, 2008a; Hall, Mayberry, & Ferreira, 2013; Langus & Nespor, 2010; Özçalişkan, Lucero, & Goldin-Meadow, 2016). It remains an empirical question whether individuals develop systematic strategies to make distinctions across different semantic domains when expressing concepts in silent gesture. It is possible that gestures will be highly idiosyncratic and thus will be executed in different ways. Alternatively, it is possible that individuals' knowledge of the world may interact with the available techniques of gestural representations and as a consequence, their silent gestures will display a high degree of systematicity.

Communication without language and what it reveals about the mind

Studies employing descriptive and empirical methods have found that language is a very important factor that shapes many of our cognitive processes (Carrol, 1956; Flecken, Von Stutterheim, & Carroll, 2014; Majid & Burenhult, 2014; Sapir, 1921). Despite the significant differences observed in the behaviours of speakers of linguistically diverse languages, most effects disappear when we look at their silent gestures. To date there are now several studies showing that, regardless of their language, speakers converge in the strategies used to represent events in this mode of manual communication.

One of the most studied effects is the sequencing of events and the agents that perform them (i.e., word order) during manual communication without speech. Languages vary considerably in the way they order the constituents of a sentence. For instance, English favours a Subject-Verb-Object sequence while Turkish prefers a Subject-Object-Verb. One would expect that when expressing an event through silent gesture, English speakers would favour an S-V-O order while Turkish speakers would follow an S-O-V ordering, as they would in their native language. However, this is not the case as speakers of both languages coincide in an S-O-V sequencing of pantomimes. This effect has been proven to be quite robust as has been replicated in multiple occasions with speakers of very diverse languages (Goldin-Meadow et al., 2008a; Hall et al., 2013; Langus & Nespor, 2010).

A similar effect has been observed in the description of motion events. English is a satellite-framed language and as such encodes information of manner (e.g., run) and path (e.g., towards) in a clause. Turkish, in contrast, is a verb-framed language often dropping information about the manner and expressing information about the path only (e.g., enter). When asked to express motion events, the co-speech gestures produced by speakers convey the same information as in their speech (i.e., English speakers express manner and path in their gestures and Turkish speakers express only path) (Kita & Özyürek, 2003; Özyürek, Kita, Allen, Furman, & Brown, 2005). However, the same speakers default to a single strategy when they are expressing the same information in silent gesture. That is, when speakers are in this mode of communication, they no longer align their gestural strategies with the information conveyed in their speech but rather they resort to a strategy that is shared across speakers of different languages (Özçalişkan et al., 2016).

Together these studies demonstrate that while language is an important factor that governs many of our cognitive behaviours, communication through silent gesture overrides any linguistic influence and generates communication with unique properties shared across speakers of different languages. The similarities in patterns observed in pantomimes in different domains (i.e., word order, motion events) have been interpreted as silent gesture being a window onto our internal representations (Özçalişkan et al., 2016) and to our capacity to package information in the most communicatively effective manner (Goldin-Meadow, So, Ozyürek, & Mylander, 2008b; Hall et al., 2013).

Another domain that has the potential to reveal a high degree of systematicity in silent gestures is the representation of concepts across different semantic categories. If silent gestures are also prone to a high degree of systematicity across different individuals, it is possible to expect generalisable patterns in the modes of representation used in specific semantic domains. This possibility has not yet been explored and remains an empirical question.

Manual modes of representation

There is general consensus that gestures may adopt at least four modes of representation. *Acting* (or handling) denotes how an object is manipulated (e.g., supination of a closed fist for 'key'). *Representing* (or instrument) uses the hand to recreate the form of an object (e.g., an extended index finger to represent a 'toothbrush'). *Drawing* (or tracing) describes the outline of a referent (e.g., two index fingers tracing a square to represent a 'window'). *Moulding* depicts the three-dimensional characteristics of an object (e.g., cupped hands describing the shape of a 'vase') (Müller, 2013). Beyond different taxonomies describing the modes of representation that gestures can adopt, it remains unclear what factors motivate certain techniques over others.

Müller (2013) proposes that during their narrations, speakers express in both the spoken and manual channel the relevant aspects of a scene or event. Importantly, the form of the gestures will depend primarily on the speakers' focus of attention or what is considered to be the relevant information. If a speaker, for instance, wants to emphasise the specific way to handle an object he will use a depicting technique that expresses this information. If, in contrast, the main focus of his narrations is the form of an object he will produce a gesture in which the hand configurations represent the shape of the referent. If the focus of his narration is the three-dimensional form of an object he will probably describe the volume of the referent in space. In other words, iconic gestural forms express visual information specifically tailored to describe a unique event.

Recent evidence has shown, however, that people's gestures are not entirely dependent on speakers' focus of attention but rather are constrained by the affordances of the referent. Using a referential paradigm, speakers were asked to describe different objects from a visual prompt. Stimuli were categorised as having high or low affordances (i.e., the degree to which objects allow to be manipulated). The analysis of participants' co-speech gestures showed that objects with high affordances (e.g., wine glass) were often represented through an acting strategy. In contrast, items with low affordances (e.g., sink) were described using a drawing strategy (Masson-Carro, Goudbeek, & Krahmer, 2015).

This study suggests that the form of iconic gestures, at least in co-occurrence of speech, are somewhat constrained by the affordances of the referent. A question that remains unanswered is whether the different representational techniques are also deployed systematically depending of the type of referent (i.e., manipulable and non-manipulable) in silent gesture. Further, it is remains an empirical question whether gesturers will resort to a different strategy to make distinctions between actions and objects, when there is no speech to aid marking this differentiation.

Action-object distinctions in the visual modality

Most of the investigations attempting to understand how the manual channel makes distinctions between actions and objects come from sign language research. The first studies exploring this issue found that in American Sign Language (ASL) pairs like HAMMER and TO-HAMMER are formally marked in the movement of the sign. While actions have a continuous movement (e.g., TO-HAMMER), objects have a restrained, repeated movement (e.g., HAMMER) (Supalla & Newport, 1986). It has also been reported that signs for actions tend to use a larger signing space and are less marked through non-manual features (i.e., mouthings) than signs for objects. These characteristics are not universal given that different sign languages use different formal features to mark these distinctions as has been documented for Australian Sign Language (Auslan) (Johnston, 2001) and Russian Sign Language (RSL) (Kimmelman, 2009). Interestingly, emerging sign languages do not seem to exhibit a clear mechanism to make such distinctions. Al-Sayyid Bedouin Sign Language (ABSL) is an emerging sign language that is gradually developing mechanisms to mark these distinctions more overtly.

More recently, studies have shown that an effective mechanism to make distinctions between actions and objects in the absence of speech is through the representation of the referent with different depicting strategies (Padden et al., 2013). When users of different sign languages were asked to represent vignettes of objects and agents manipulating objects (actions), one can observe that there is systematicity in their patterning of use. ASL and ABSL signers tend to depict actions through acting depictions and objects through representing depictions. This distinction is language-specific because users of an unrelated sign language (New Zealand Sign Language) favour the opposite patterns (i.e., acting for nouns and representing for verbs). Interestingly, this study also revealed that when asked to perform the same task, hearing people always converge in the same strategy to represent the referent. That is, silent gesturers predominantly favour an acting strategy for all their gestural depictions (actions and objects alike) with few instances of representing depiction (Padden et al., 2013; Padden, Hwang, Lepic, & Seegers, 2015). The notion of *patterned iconicity* postulates that sign languages may differ in the strategy used to make action-object distinctions, but they systematically exploit the available depicting possibilities (modes of representation) to make such differentiations. In contrast, gesturers default to the same strategy (acting) for both semantic categories.

These studies show that sign languages alter the phonological structure of the sign or their mode of representation to distinguish actions from objects while pantomime overall defaults to the acting technique. A shortcoming of these studies is that they have limited their observations to the techniques of depiction only within two semantic domains (tools and actions with tools). Given that gestures are holistic units without sub-lexical components (McNeill, 1992) individuals are unlikely to modify their gestures' kinematics to make semantic distinctions in a similar way as signs. It is possible, however, that they may deploy additional strategies and bodily cues such as pointing, showing, eye-gaze, and sequences of gestures to mark such distinctions.

The Present Study

In the present study we turn to the production of silent gesture to investigate whether actions vs. objects and their affordances (i.e., manipulable vs, non-manipulable) modulate the strategy used by speakers to represent a referent manually. More specifically, we ask 1) do individuals use a specific depicting strategy for each semantic category; and 2) what are the additional strategies implemented to make semantic distinctions. To that end, we implemented a pantomime generation task to a group of Dutch speakers and described the gestures produced for a list of words, the strategy they used for each semantic category, and the strategies they implemented to differentiate actions from objects.

Methodology

Participants

Twenty native speakers of Dutch (10 females, age range: 21-46, mean: 27 years) living in the area of Nijmegen, the Netherlands took part in the study.

Procedure

Participants were tested individually in a quiet room with two cameras from different angles recording their gestures. They were told that the task consisted of generating a sign or gesture that conveyed exactly the same meaning as the word on the screen. They were explicitly told two rules: they were not allowed to speak or say the target word; and they could not point at any object present in the room (e.g., pointing at the table or at a wall). They were also told that their videos were going to be shown to another participant who would have to guess the meaning of their gesture.

The stimuli consisted of a total of thirty words from three semantic categories: 10 actions with an object (e.g., to phone, to smoke), 10 manipulable objects (e.g., telephone, lighter), and 10 non-manipulable objects (e.g., pyramid, floor). Words were presented in black font on a white background in a different randomised list for each participant. We decided against presenting the stimulus materials with a visual cue so as to avoid prompting participants. Each trial started with a fixation cross in the middle of the screen for 500 ms and this was followed by the word participants had to represent with their gestures. The target word remained on the screen for 4000 ms during which participants had to come up with their gestural depictions. We limited the allowable time for gestural production so as to force participants to produce their most intuitive responses. Participants' renditions were video recorded and later annotated using the software ELAN (Lausberg & Sloetjes, 2009).

Coding and data analysis

For each target word, participants were observed to produce one gesture or sequences of gestures to depict the referent. Following a strict coding criteria, all gestures produced for each item were annotated. Each gesture or sequences of gestures would consist minimally of a preparation phase, a stroke and a (partial/full) retraction. Once all the gestures were isolated, we classified them according to their mode of representation. Adapting the taxonomy developed by Müller (2013), we categorised each gesture as follows: *Acting* if the gesture represented how the referent is manipulated; *representing* if the hands were used to recreate the form of an object; and *drawing* if participants used their hands to describe the outline or the three-dimensional characteristics of an object (note that we collapsed Muller's drawing and moulding categories into one). Aside from these modes of representation, we also included the category *deictic* which consisted of pointing, showing and/or ostensive eye-gaze to elements of the gesture (see Figure 1 for examples). This is

not a mode of representation *per se* but we decided to include this category in the analysis given the high prevalence of this strategy to make semantic distinctions.

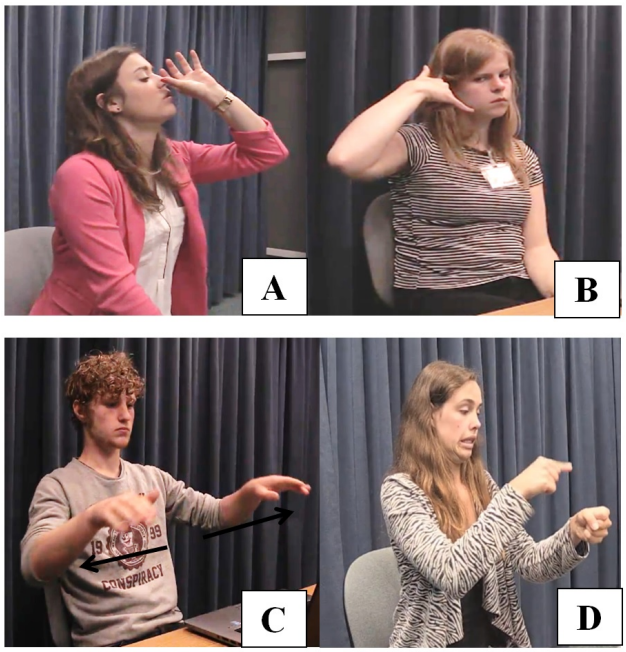


Figure 1: Examples of participants using different modes of representation (Müller, 2013) and deictics. A) The *acting* technique represents how an object is manipulated (e.g., 'to drink'); B) in *representing* the hands recreate the shape and form of an object (e.g., 'telephone'); C) *drawing* traces the outline or three dimensional features of an object (e.g., 'table'), and D) deictics like points, showing or eye-gaze are used to highlight features of a gesture (e.g., pointing at a fist holding an imaginary mug).

After the whole dataset was annotated and categorised according to the gestures' mode of representation, we calculated the number of gestures produced per item per participant across the three semantic categories (actions with an object, manipulable objects, and non-manipulable objects). Then, we calculated the proportion of the different modes of representation per semantic category; and finally, we calculated the proportion of decitics used in the three different categories.

Results

We calculated the number of gestures produced per item per participant across the three semantic domains. We found that actions with objects elicited the least number of gestures (range: 1-2 gestures; mean: 1.1 gestures, *SD* = 0.12), followed by non-manipulable objects (range 1-4 gestures; mean: 1.49 gestures, *SD* = 0.33), and manipulable objects elicited the highest number of gestures (range: 1-4 gestures; mean: 1.77 gestures, *SD* = 0.39). On the arcsine transformed values, a one-way ANOVA revealed a significant difference in the number of gestures produced for each semantic category $F(2,38) = 40.14$, $p < 0.0001$, $\eta^2 = 0.679$. Pairwise comparisons after Bonferroni corrections revealed that the number of gestures produced for each category is significantly different from one another. Actions with objects was significantly lower than manipulable objects [$t(19) = 8.39$, $p < 0.0001$] and non-manipulable objects [$t(19) = 5.24$, $p < 0.0001$]. Non-manipulable objects elicited significantly fewer gestures than manipulable objects [$t(19) = 3.98$, $p < 0.001$].

Table 1 shows the proportion of instances in which participants produced a single vs. multiple gestures to describe a referent across conditions. After removing passes and wrong targets (e.g., the target word 'to sieve' *zeven* often elicited the gesture 'seven' *zeven*) we can see that the vast majority of actions with tools elicited a single gesture, which often depicted how the action is executed (see Figure 2 for the gesture 'to-drink'). In contrast, manipulable objects were predominantly depicted with more than one gesture (see Figure 2 for the gesture 'lighter'). Non-manipulable objects have a split with an almost equal proportion of items being depicted with a single or multiple gestures.

Table 1: Proportion of concepts depicted with a single or multiple gestures across conditions (N=200)

	Action with objects	Manipulable objects	Non-manipulable object
Single	0.91	0.35	0.45
Multiple	0.09	0.61	0.46
Wrong	0.01	0.05	0.10
	1.00	**1.00**	**1.00**

In order to explore whether the gestural forms are restricted by the affordances of the referent, we looked at the mode of representation used across semantic categories (acting, representing or drawing). We focused on the instances in which a single gesture had been elicited to represent a concept. Table 2 shows that actions with objects and manipulable objects use predominantly the acting strategy (i.e., how an object is used). In contrast, non-manipulable objects resort more often to a drawing technique.

Table 2: Proportion of concepts depicted with different modes of representations across conditions (one-gesture depictions only)

	Action w/object (N=181)	Manipulable object (N=70)	Non-manipulable object (N=89)
Acting:	0.86	0.81	0.20
Drawing	0.00	0.01	0.57
Representing	0.14	0.17	0.22
	1.0	1.0	1.0

Finally we looked at the instances in which participants included a deictic (i.e., pointing, showing, ostensive eye-gaze) to refer to a specific feature of their gesture. We observed that out of the whole data set, actions with objects (N=199) and non-manipulable objects (N=180) elicited a small proportion of deictics (0.04 and 0.09 respectively). In contrast, manipulable objects (N=191) elicited significantly more deictics (0.25). For instance, to represent 'lighter' participants would perform the action of lighting a cigarette and then they would point at an imaginary lighter (See Figure 2). Similarly, for 'toothbrush' they would pretend to be brushing their teeth with a handling handshape (i.e., closed fist) and then they would raise it and show it to the camera.

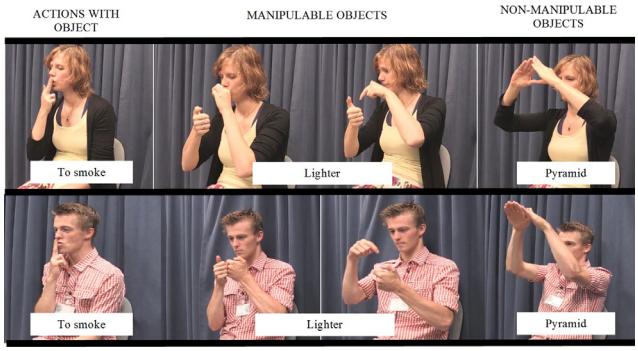


Figure 2: Examples of the modes of representation in pantomimes across different semantic categories. Actions with object were depicted with a single gesture representing how to manipulate an object (e.g., 'to smoke'); manipulable objects were represented with the gesture of an action followed by a deictic (e.g., lighter was depicted with a pantomime of lighting a cigarette and then pointing at an imaginary lighter). Non-manipulable objects were more frequently depicted with drawing depictions (e.g., 'pyramid').

Discussion

In this study we investigated whether certain modes of representations were typically bound to a specific semantic domain and whether there were generalisable patterns observed across different participants. We also looked into the different strategies deployed to distinguish different word types such as actions and objects. The results from a pantomime elicitation task revealed that pantomimes show systematic patterns when speakers are asked to represent a referent in silent gesture. Actions with objects tend to be expressed with a single gesture using an acting mode of representation (i.e., representing how an object is manipulated). Non-manipulable objects in contrast tend to be represented with a drawing strategy and with more than one gesture. Interestingly, manipulable objects elicited significantly more gestures than the other two categories and the most common strategy used was also acting. Crucially, we observed that manipulable objects elicited significantly more deictics than the other two categories. That is, participants would pantomime an event (e.g., eating soup) and then highlight part of this gesture with a deictic (e.g., pointing at an imaginary spoon).

The present data replicates earlier findings that speakers tend to rely on the acting mode of representation in their gestures (Padden et al., 2013, 2015; van Nispen, van de Sandt-Koenderman, Mol, & Krahmer, 2014). However, we also find that this mode of representation falls out of favour when the referent does not allow an effective way of depiction. Gesturers seem to switch from one strategy to the other depending on whether the referent allows for certain modes of representation (i.e., drawing is favoured in the depiction of non-manipulable objects). These findings go in line with recent research showing that the shape of an object and the possible ways to interact with it modulate the form of co-speech gestures (Masson-Carro et al., 2015). These findings also resonate work showing that gesture production relates to simulation of actions (Cook & Tanenhaus, 2009). It is possible that speakers default to an acting strategy in their gestures because they are simulations of their experiences with objects. However, when an object does not lend itself to a clear form of manipulation or the affordances of the object does not permit the use an acting strategy, individuals will turn to an alternative strategy to represent it.

When we look at the different strategies adopted to make distinctions between actions with objects and manipulable objects we see that gesturers do not align different modes of representation to a specific category, as has been shown for established or emerging sign languages (Johnston, 2001; Kimmelman, 2009; Supalla & Newport, 1986). Instead, we see that gesturers complement their acting strategies with deictics to highlight the focus of their gesture. That is, gesturers feel the communicative need to inform the addressee that the intended referent is not the action they are depicting, but the object at hand.

This study adds to our current understanding of gesture production, and the factors that drive their form. However, we should be cautious about the generalisation of these results given that most research in this domain has focused on the gestures produced by Dutch speakers. Future work should investigate whether speakers of different languages adopt the same strategies in silent gesture regardless of their native language. By looking at communication in the absence of speech we see gesturers devise strategies to express complex notions such as action-object distinctions. These strategies operate in tandem with individuals' knowledge of the world and the available strategies to represent a referent. These strategies may be the raw materials of emerging sign languages and the foundations of a conventionalised manual linguistic system.

Acknowledgments

This work was supported by a Veni grant by the Netherlands Organisation for Scientific Research (NWO) awarded to the first author. We would like to thank Deniz

Cetin and Renske Schilte for their collaboration in the annotation of the data.

References

Carrol, J. B. (1956). *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: MIT Press.

Cook, S. W., & Tanenhaus, M. K. (2009). Embodied communication: speakers' gestures affect listeners' actions. *Cognition*, *113*(1), 98–104. doi:10.1016/j.cognition.2009.06.006

Flecken, M., Von Stutterheim, C., & Carroll, M. (2014). Grammatical aspect influences motion event perception: findings from a cross- linguistic non-verbal recognition task. *Language and Cognition*, *6*(01), 45–78. doi:10.1017/langcog.2013.2

Goldin-Meadow, S., So, W. C., Ozyürek, A., & Mylander, C. (2008a). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9163–8. doi:10.1073/pnas.0710060105

Goldin-Meadow, S., So, W. C., Ozyürek, A., & Mylander, C. (2008b). The natural order of events: how speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9163–9168. doi:10.1073/pnas.0710060105

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: evidence from elicited pantomime. *Cognition*, *129*(1), 1–17. doi:10.1016/j.cognition.2013.05.004

Johnston, T. (2001). Nouns and verbs in Australian sign language: An open and shut case? *Journal of Deaf Studies and Deaf Education*, *6*(4), 235–257. doi:10.1093/deafed/6.4.235

Kimmelman, V. (2009). Parts of speech in Russian Sign Language: The role of iconicity and economy. *Sign Language and Linguistics*, *12*(2), 161–186. doi:10.1075/sl

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. doi:10.1016/S0749-596X(02)00505-3

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology*, *60*(4), 291–318. doi:10.1016/j.cogpsych.2010.01.004

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES--ELAN system. *Behavior Research Methods*, *41*(3), 841–9. doi:dx.doi.org/10.3758/brm.41.3.841

Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, *130*(2), 266–270. doi:10.1016/j.cognition.2013.11.004

Masson-Carro, I., Goudbeek, M., & Krahmer, E. (2015). Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition and Neuroscience*, 1–11. doi:10.1080/23273798.2015.1108448

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Müller, C. (2013). Gestural modes of representation as techniques of depcition. In C. Müller, A. Cienki, S. Ladewig, D. McNeill, & J. Bressem (Eds.), *Body - Language - Communication: An International Handbook on Multimodality in Human Interaction* (pp. 1687–1701). Berlin: De Gruyter Mouton.

Özçalişkan, Ş., Lucero, C., & Goldin-Meadow, S. (2016). Does language shape silent gesture? *Cognition*, *148*, 10–18. doi:10.1016/j.cognition.2015.12.001

Özyürek, A., Kita, S., Allen, S. E. M., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures?: Insights from crosslinguistic variations and similarities. *Gesture*, *5*(1), 219–240. doi:10.1075/gest.5.1.15ozy

Padden, C., Hwang, S.-O., Lepic, R., & Seegers, S. (2015). Tools for Language: Patterned Iconicity in Sign Language Nouns and Verbs. *Topics in Cognitive Science*, *7*(1), 81–94. doi:10.1111/tops.12121

Padden, C., Meir, I., Hwang, S.-O., Lepic, R., Seegers, S., & Sampson, T. (2013). Patterned iconicity in sign language lexicons. *Gesture*, *13*(3), 287–305.

Sapir, E. (1921). *Language*. New York: Harcourt, Brace & World.

Supalla, T., & Newport, E. L. (1986). How many sits in a chair? The derivations of nouns and verbs in American Sign Language. In P. Siple (Ed.), *Understanding language through sign language research* (pp. 91–132). New York: Academic Press.

van Nispen, K., van de Sandt-Koenderman, M., Mol, L., & Krahmer, E. (2014). Pantomime Strategies: On Regularities in How People Translate Mental Representations into the Gesture Modality. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)* (pp. 3020–3026). Austin, TX: Cognitive Science Society, Inc.

Singular Interpretations Linger During the Processing of Plural Noun Phrases

Nikole D. Patson (Patson.3@osu.edu)
Department of Psychology, 1465 Mount Vernon Avenue
Marion, OH 43302 USA

Abstract

Plural nouns do not strictly refer to more than one object, which suggests that they are not semantically marked to mean "more than one" and that plurality inferences are made via a scalar implicature. Consistent with that hypothesis, recent evidence using a picture-matching paradigm supports founds that participants were equally fast to respond to a picture of a single object as a picture of multiple objects after reading a sentence containing a plural. This suggests that comprehenders activate both a semantic (i.e., singular) and a pragmatic interpretation (i.e., plural). The current study found that even after a 1500 ms delay, comprehenders still maintain activation of both meanings after reading a sentence containing a plural. This suggests that the activation of the singular meaning may not be due to the processing of a scalar implicature, but rather may be due to the nature of plural conceptual representations.

Keywords: plurals; semantics; pragmatics; scalar implicature; language comprehension; conceptual representations

Introduction

There is a well-known puzzle regarding the interpretation of plural noun phrases. Consider the following set of examples.

1a) Ben fed a shark.
1b) Ben fed sharks.
1c) Ben fed more than one shark.

2a) Ben didn't feed a shark.
2b) Ben didn't feed sharks.
2c) Ben didn't feed more than one shark.

For most people, 1b & 1c are essentially the same in meaning, and distinct from 1a. However, in the negated cases, 2a and 2b are usually judged as equivalent and distinct from 2c (Sauerland et al., 2005; Tieu et al., 2014). This suggests that "more than one" is not always the appropriate interpretation of a plural noun phrase. Because of this pattern of interpretations, linguists have argued that the plural is semantically unmarked, or weakly marked, for number, while the singular is strongly marked for number.

If the plural is semantically unmarked (or weakly marked) for number, then linguistic theory must account for the fact that, in many contexts, most people interpret plural nouns to mean "more than one". One hypothesis is that when comprehending a plural, comprehenders make a kind of pragmatic inference known as a scalar implicature (e.g., Spector, 2007; Tieu et al, 2014). A scalar implicature is a type of inference that arises when a weak expression is used instead of a stronger expression. For example, a sentence like *Zoe ate some of the cookies* is typically interpreted as meaning Zoe ate some but not all of the cookies. However, logically that does not have to be the case. If Zoe ate all of the cookies, it is also true that she ate some of the cookies. Thus, the statement that Zoe ate some of the cookies does not logically rule out the possibility that Zoe in fact ate all of the cookies. Nevertheless, comprehenders seem to assume that speakers use the strongest labels that are compatible with their intended meaning and interpret the fact that a weaker expression was used to indicate that the stronger meaning was not appropriate (Grice, 1975). Thus, they assume that if a speaker intended to indicate that Zoe indeed ate all of the cookies, they would have used the quantifier *all* because that would be the strongest way to communicate that state of affairs. Applying this logic to plural noun phrases, a scalar implicature account assumes that the literal, semantically defined interpretation of the plural is something like "at least one" and an implicature must be made to arrive at the "more than one" interpretation. The logic of the implicature is as follows: a plural can refer to a single entity, but if the speaker intended to refer to only one entity they would have used a stronger form (i.e., the singular) to express that.

Using a truth-value judgment task (e.g., T/F: *Does a dog have tails?),* Tieu et al. (2014), provided evidence for a scalar implicature account of plurality. Tieu et al. found that both adults and children interpreted plural nouns as meaning "more than one" more often in positive contexts (or upward-entailing environments) than negative contexts (or downward-entailing environments). This is consistent with typical scalar implicature patterns (e.g., Chierchia, 2004; Levinson, 2000). Additionally, Tieu et al. found that children were less likely to compute plural inferences than adults, which is also consistent with previous work showing that children are typically less likely to compute scalar implicature inferences than adults (e.g., Noveck, 2001).

Recently, Patson (in press) provided experimental evidence that is consistent with the hypothesis that comprehenders compute a scalar implicature when comprehending a plural. Patson (in press) used a picture-matching paradigm designed to probe the conceptual representation of plural noun phrases (e.g., Patson, George, & Warren, 2014). Using this picture matching paradigm, Patson, George and Warren (2014) had participants read a sentence that contained either a singular noun (as in 3), a plural definite description (as in 4), or a two-quantified plural (as in 5).

3) The parent handed the child the *crayon*.
4) The parent handed the child the *crayons*.

5) The parent handed the child the *two crayons*.

After reading the sentence, participants pressed a button and then they saw a picture of exactly one object, exactly two objects, or multiple (3-6) objects. Participants were instructed to decide whether or not the picture was of an object(s) that was mentioned in the sentence. They were instructed and trained to ignore the number of objects and base their judgment on object identify alone. Patson et al. measured how quickly participants responded affirmatively to the picture. When number was explicit (e.g., singular NP, two-quantified NP), participants were faster to judge a picture when the number matched. For example, after reading a singular noun phrase participants were faster to accurately decide that the picture was of an object that was mentioned in the sentence when there was only one object pictured compared to when multiple objects were pictured. This finding is straightforward: When number information is made explicit comprehenders have a detailed conceptual representation that contains that number information. Interestingly, for the plural definite description conditions, there were no judgment time differences based on picture type. That is, participants did not show a preference for pictures that depicted more than one object compared to pictures that depicted a single object. Patson et al. interpreted this finding as consistent with the theory that plurality is semantically unmarked for number. That is, because number information is not semantically explicit in a plural noun phrase, comprehenders do not explicitly represent number information when building a conceptual representation of a plural noun phrase. Therefore, they showed no preference for pictures of multiple objects over pictures of a single object.

Patson (in press) followed up this work and suggested that the lack of preference for a picture depicting multiple objects compared to a picture depicting a single object is due to the computation of a scalar implicature. In Patson (in press) comprehenders read sentences that contained plural noun phrases. The sentences were written to evoke a particular spatial configuration for the plural set. For example, in the first experiment, the sentential context either described a spatial configuration where the individual items that comprised the plural were spread out (e.g., the wind scattered the leaves) or the items were gathered closely together (e.g., a pile of leaves). After reading the sentence, comprehenders were shown a picture that matched the spatial configuration, mismatched the configuration, or was a single object. Patson found that participants were faster to respond to a picture when it matched the spatial configuration implied in the sentence than when it did not (cf. Stanfield & Zwaan, 2001). Importantly, participants were also faster to respond to a picture of a single item than to a picture that mismatched the spatial configuration implied in the sentence. Patson interpreted the match finding as indicating that comprehenders do not leave number information conceptually unspecified for plural noun phrases. Instead, comprehenders interpret plural noun phrases as meaning "more than one" and create highly detailed conceptual representations that contain information about how the individual entities that make up the plural set are arranged. Furthermore, Patson argued that this interpretation came about via a scalar implicature. This argument was based on the finding that even though comprehenders created highly detailed conceptual representations, they had activated a singular representation as evidenced by the finding that participants were faster to respond to the picture of a single object than the picture that mismatched the sentential context. Patson argued that this pattern of findings could be explained by assuming that during comprehension both the literal, semantic meaning ("at least one") and the pragmatic scalar implicature ("more than one") are computed during the processing of a plural. At the end of the sentence, both the semantic meaning and the pragmatic meaning are still active in memory, so participants are equally fast to respond to a picture that is consistent with the semantic meaning (e.g., a picture of a single object) as well as a picture consistent with the pragmatic meaning (e.g., a picture that matches the sentential context). This interpretation was based on findings presented by Kaup, Lüdtke, and Zwaan (2006). Kaup et al. had comprehenders read negated sentences and then presented them with pictures that either matched the affirmative state of affairs or the negated (actual) state of affairs. For example, participants read sentences like: *The umbrella was not open*. Then they saw a picture of either an open umbrella (affirmative state of affairs) or a closed umbrella (negated, or actual, state of affairs). In their first experiment, Kaup et al. presented the pictures immediately after the sentence. In that experiment participants were equally fast to respond to the affirmative picture as the negated picture. In the second experiment, Kaup et al. inserted a 1500 ms delay between the sentence and the presentation of the picture. In that experiment, participants were faster to respond to the negated picture than to the affirmative picture. Kaup et al. argued that comprehenders create conceptual representations for intermediate stages (e.g., the affirmative state of affairs) of linguistic processing and that those intermediate stages are still active at the end of the sentence. Furthermore, those intermediate stages get suppressed or deactivated over time and the comprehender is left with a conceptual representation that adheres to the actual state of affairs described in the sentence. While Kaup et al. interpreted their findings with respect to mental simulations comprehenders were performing during comprehension, these findings might also hold for scalar implicature processing. Specifically, this account predicts that during the processing of a scalar implicature, comprehenders compute and conceptually represent both the literal, semantic meaning of the scalar term as well as the pragmatic, implicature –derived meaning. Thus, given a delay between the end of the sentence and the presentation of a picture, comprehenders should show a preference for the pragmatic, implicature-derived meaning and not preference for the literal, semantic meaning. The current study tests this prediction.

Whether or not the singular interpretation maintains activation over a delay is an important question for understanding how plural noun phrases are conceptually represented. Very little work has been done to investigate how plural nouns phrases are conceptually represented (e.g., Patson, 2014). Given that number is a fundamental aspect of language processing, it is important to understand how language comprehenders represent it conceptually.

Although little work has been done to explicitly probe the conceptual representation of plural noun phrases, some theorist have speculated about what those conceptual representations might look like. For example, Johnson-Laird (1983) argued that the conceptual representation for large quantities could contain a small set of objects or may be a single token. Under this hypothesis, it is possible that when comprehenders conceptually represent a plural they represent the plural at both the level of the set and also represent a single token of the individuals that make up the plural. Activating a representation of both the set and the individuals that make up the set would be logical given that, depending on the context, the set or the individuals may be more relevant.

The current study was designed to investigate whether the activation of the singular meaning during the comprehension of plural noun phrases (Patson, in press; Patson, George, & Warren, 2014) persists after a 1500 ms delay. If the activation of the singular interpretation is due to the computation of a scalar implicature, then comprehenders should not show a reduced reaction time for a singular picture compared to a plural picture that mismatches the context after a delay (e.g., Kaup et al., 2006). If a reduced reaction time for a singular picture persists after a delay, it suggests that the activation of the singular interpretation was not due to the computation of a scalar implicature, but rather may be due to the nature of plural conceptual representations.

Method

Participants

Fifty four native speakers of American English volunteered to participate. Participants were recruited from the Columbus Center of Science and Industry (COSI).

Design and Stimuli

The experiment had a 2x3 repeated measures design. The first factor was the implied distribution of entities in the plural set. The sentences either implied that the entities within the plural set were spatially gathered (as in 6) or spatially spread out (as in 7).

6) The gardener raked up the leaves.
7) The breeze scattered the leaves.

The sentences used in the current study were a subset of the sentences used in Patson (in press). The sentences were normed to ensure they evoked the correct spatial configuration.

The second factor was the picture type (see Figure 1 for examples). The picture was either spatially gathered (e.g., a pile of leaves), spatially distributed (e.g., spread out leaves), or a single object. The pictures used in the current study were a subset of the pictures used in Patson (in press). In order to verify that all three picture types were similarly easy to identify, Patson (in press) ran a norming study in which participants were shown a label (e.g., leaves) and asked to judge whether a picture was a good match for that label. There were no response time differences based on picture type indicating that any response time differences in the Experiment are not due to differences in how the pictures are visually processed.



Figure 1. Pictures (in greyscale) used in the Experiment.

Forty two experimental items were divided into six lists such that each list contained one condition from each experimental item. Each participant viewed one list. Each list also contained the same set of 42 filler sentences. Filler sentences were structured exactly like the experimental sentences; however, in all of the filler items, the picture that followed the sentence was not of an item mentioned in the sentence. The filler pictures contained a mixture of singular objects, spatially grouped objects, and spatially distributed objects. The correct answer for the sentence-picture matching judgment was "no" for all 42 filler items and "yes" for all 42 experimental items.

Apparatus

The trials were presented using E-Prime v.2 experimental software (Schneider, Eschman, & Zuccolotto, 2002). A Dell P2412H 24-inch monitor (1920 X 1080 pixels) displayed stimuli with a screen refresh rate of 60 Hz. Keyboard presses were used to log responses and record reaction time.

Procedure

Participants were tested individually or in pairs. After they provided informed consent, they were given a verbal introduction to the experiment. Then the computer guided them through example trials followed by four practice trials with feedback. As in Patson et al. (2014) and Patson (in press), the instructions and practice indicated that participants' judgments should be based on object identity. That is, participants were instructed to answer "yes" when they saw a singular picture even when the sentence

contained a plural noun phrase. A left-justified fixation cross signaled the beginning of each trial. When participants pressed the spacebar, the cross was replaced by a sentence. Participants read the sentence at their own pace and pressed the space bar. The sentence disappeared and a blank screen appeared for 1500 ms. After the 1500 ms delay, a picture appeared in the center of the screen. Participants decided whether or not the picture was of an object mentioned in the sentence by pressing 'Y' for yes and 'N' for no on the keyboard. The picture disappeared only when participants made their response. The participants' button presses and response times were recorded.

Results

Overall, accuracy on the task was high (M = 96%, SD = 4%). One participant was dropped from the analysis due to extremely long reaction times (over 3 SDs greater than mean RTs).

The mean reaction times for correct trials are presented in Figure 2.

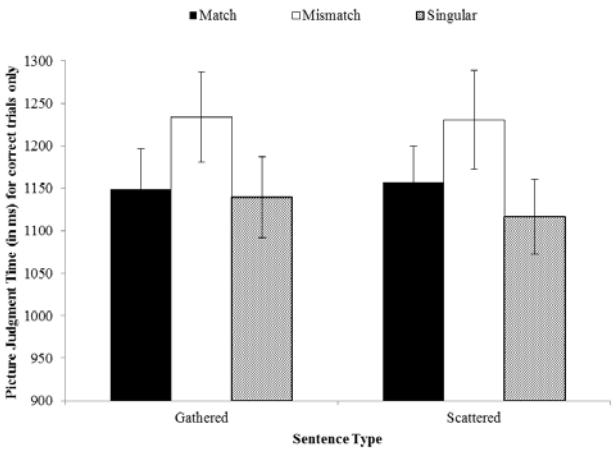


Figure 2. Mean (standard error) picture judgment time by sentence and picture type.

The main effect of picture type was significant, $F1(2, 104) = 13.09$, $MSe = 24556.73$, $p < .001$; $F2(2, 82) = 10.72$, $MSe = 40179.22$, $p < .001$. Neither the main effect of sentence type nor the interaction was significant, all ps > .10.

Planned comparisons indicated that participants were faster to respond accurately to the picture when the picture matched the sentential context than when it did not, $t_1(52) = 3.89$, $p < .001$; $t_2(41) = 3.60$, $p = .001$. Additionally, participants were faster to respond to the singular picture than to the picture that mismatched the sentential context, $t_1(52) = 4.53$, $p < .001$; $t_2(41) = 3.88$, $p < .001$.

Discussion

The results reported here replicate and extend the results reported by Patson (in press). When comprehenders read sentences containing plural noun phrases they are faster to respond accurately to a picture that matches the sentential context than to a picture that does not. Additionally, comprehenders are faster to respond to a picture of a single object than to a picture that does not match the sentential context. What is new in this experiment is that these effects occur even after a 1500 ms delay. This suggests that the singular interpretation that gets activated by the plural is not an intermediate stage of processing, but rather is part of the conceptual representation comprehenders build for plural noun phrases.

Importantly, these data do not undermine the scalar implicature account of plural noun phrases. Indeed, if the singular activation were due to the computation of a scalar implicature, it was predicted that participants in this experiment should not have shown a preference for the singular picture over the picture that mismatched the sentential context. However, the finding that participants did show a preference for the singular picture does not indicate that a scalar implicature was not computed. Instead, it suggests that the activation of the singular meaning was not due to the semantic content of the plural noun phrase, but rather may be due to the nature of how plural nouns are conceptually represented (as I will describe in more detail below). Furthermore, Patson (in press) argued that the semantic meaning of a plural noun phrase is something like "at least one". However, there are linguistic patterns that suggest this may not be the definition of the plural. For example, the sentence *I have zero dogs*, the meaning "at least one dog" would be incompatible with the intended meaning of the sentence. This issue is beyond the scope of this paper, but does suggest that the activation of the singular meaning may not have come from the semantic content of the plural noun phrase.

The data reported here suggest that during the comprehension of a plural noun phrase, comprehenders activate a representation of a single token of the objects that make up the plural and that activation persists over a 1500 ms delay. This suggests that the activation of the single token is an important part of the plural conceptual representation. This is reasonable given the fact that depending on the context the set or the individuals may be more relevant during the comprehension of plural noun phrases (e.g., Patson, 2014; Patson & Warren, 2010). For example, plural noun phrases can have either a collective (as in 8) or a distributed reading (as in 9).

8) Together the girls ate a cookie.

9) Each of the girls ate a cookie.

In the collective case (8), the predicate is applied to the set making the set is the most relevant referent for the plural noun phrase. In the distributed case (9), the predicate is applied to the individuals that make up the set, making the individuals the more relevant referent. Given that either the set or the individuals that make up the set can be more or less relevant (and that this relevance can shift throughout a discourse) it is reasonable that the conceptual representation would contain an explicit representation of both the set and the individuals that make up the set.

There is still an open question about why comprehenders represent a single token rather than multiple distinct entities when representing the individuals that make up the plural set. One possibility is that the single token representation is a way in which the conceptual representation is left indeterminate (e.g., Barsalou, 1999). For example, in discussing how a tiger's stripes may be simulated in a conceptual representation, Barsalou[1] argued that there are two ways in the conceptual image might be indeterminate. First, the stripes may be blurry, such that they cannot be counted. Second, they might be extracted from the representation of the tiger such that they appear in a patch. The data reported here are consistent with Barsalou's second suggestion for indeterminacy. With respect to the plural, the individuals that make up the plural set may be represented by extracting a single token from the plural set to represent each of the individuals. Future work will be necessary to fully investigate how plural noun phrases are conceptually represented.

Additionally, future work should be aimed at investigating how context influences the conceptual representation of plurals. Recently, Zwaan (2014) argued that different contexts may require different levels of conceptual representation. Given that context can influence the relevancy of the individuals that make up the plural set (e.g., Patson, 2014; Patson & Warren, 2010), the strength of the singular representation may be influenced by the context in which the plural appears.

Furthermore, scalar implicatures are more likely to be computed in some contexts compared to other contexts (Degen & Tanenhaus, 2015). Thus, given the linguistic evidence that plurals derive their number information via a scalar implicature (Tieu et al., 2014), future work aimed at investigating how context influences the conceptual representations of plural noun phrases must take into account how likely scalar implicatures are to be computed.

Acknowledgments

This research was conducted at the Buckeye Language Network Language Sciences Lab at the Center of Science and Industry in Columbus, Ohio. I would also like to thank Ciara Corielle and Elaina Hill for assistance collecting data.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond*. Oxford: Oxford University Press.

Degen, J. and Tanenhaus, M.K. (2015). Processsing scalar implicature: a Constraint-Based approach. *Cognitive Science, 39*, 667-710.

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics, 3*, 41-58.

Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed?. *Journal of Pragmatics, 38*, 1033-1050.

Levinson, S. (2000). *Presumptive meanings.* Cambridge, MA: MIT Press.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*, 165-188.

Patson, N. D. (2014). The processing of plural expressions. *Language and Linguistics Compass, 8*, 319-329.

Patson, N.D. (in press). Evidence in support of a scalar implicature account of plurality. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Patson, N. D., George, G.E., and Warren, T. (2014). The conceptual representation of number. *Quarterly Journal of Experimental Psychology*, *67*, 1349-1365.

Patson, N. D., and Warren, T. (2010). Evidence for distributivity effects in comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36,* 782-789.

Sauerland, U. (2003). A new semantics for number. In *Semantics and Linguistic Theory* (pp. 258-275).

Sauerland, U, Anderssen, J. & Yatsushiro, J. (2005). The plural is semantically unmarked. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence*, de Gruyter.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Spector, B. (2007). Aspects of the pragmatics of plural morphology: On higher-order implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presuppositions and Implicature in Compositional Semantics*, Palgrave.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*, 153-156.

Tieu, L., Bill, C., Romoli, J., & Crain, S. (2014). Plurality inferences are scalar implicatures: Evidence from acquisition. *Proceedings of SALT, 24*.

Zwaan, R. A. (2014). Embodiment and language comprehension: reframing the discussion. *Trends in Cognitive Sciences, 18,* 229-234.

[1] Barsalou (1999) and Zwaan (2014) argue for embodied representations. The arguments here are fully consistent with an embodied account of language processing. However, the arguments here do not rely on an assumption that language processing is embodied.

Using determiners as contextual cues in sentence comprehension: A comparison between younger and older adults

Nazbanou Nozari (nozari@jhu.edu)
Department of Neurology, Johns Hopkins University, 1629 Thames Street
Baltimore, MD 21231, USA

Daniel Mirman (dan@danmirman.org)
Department of Psychology, Drexel University, 3141 Chestnut St
Philadelphia, PA 19104, USA

Abstract

Younger adults use both semantic and phonological cues to quickly and efficiently localize the referent during sentence comprehension. While some behavioral studies suggest that older adults use contextual information even more strongly than younger adults, ERP studies have shown that this population, as a group, is less apt at using contextual semantic cues to predict upcoming words. The current study extends the investigation of contextual cue processing in auditory sentence comprehension beyond semantic cue processing, by comparing younger and older adults in their ability to use phonological cues in indefinite articles (a/an) to localize the referent in an eye-tracking visual world paradigm. Our results suggest that both age groups use such phonological information for referent localization, but with different timelines: younger adults use the cues to anticipate an upcoming word, whereas older adults show delayed cue processing after the target word has been spoken. Together with findings from semantic context processing, these results support a model of sentence comprehension in which the use of contextual cues continues with aging, but is no longer as efficient as in the young system for anticipatory word retrieval.

Keywords: aging; sentence comprehension; context; determiners; indefinite articles; eye-tracking.

Introduction

Do older adults process sentences differently from younger adults? The answer to this question is not only interesting from the viewpoint of understanding the language system, but also from the more general perspective of how aging changes the principles of cognitive processing. One such principle is the use of contextual information to facilitate processing of an upcoming stimulus. A large body of research suggests that individuals routinely use contextual information in both reading and listening; for example, upon hearing "The boy will eat the …" listeners quickly look at the edible object among all other objects on the screen (e.g., Altmann & Kamide, 1999). Similarly, upon reading "The day was breezy so the boy went outside to fly ..." participants anticipate the word "kite" and its appropriate determiner "a", showing not only anticipation of the lexical representation but also of its phonological form (Delong et al., 2005, 2012). Do older adults also use contextual information?

Some behavioral studies suggest that older adults use sentential context to process a target word as much as, or even more than, younger adults, especially when the target is presented amid visual or auditory noise (e.g., Madden, 1988; Pichora-Fuller, Schneider, & Daneman, 1995). However, more recent ERP studies have shown that older adults are less likely to use contextual information during sentence comprehension to anticipate the upcoming word (e.g., Federmeier, Kutas, & Schul, 2010; Federmeier, McLennan, Ochoa, & Kutas, 2002; Wlotko, Federmeier, & Kutas, 2012).

Most of these ERP studies have focused on N400, a negative potential that reflects implicit aspects of semantic access, and is larger in magnitude for words that contradict contextually-induced expectations (Kutas & Federmeier, 2000). For example, Federmeier, Van Petten, Schwartz, and Kutas (2003) found that N400 in response to conflict between a word and the sentential context was delayed by over 200 ms in older, compared to younger, adults. Subsequent work endorsed this finding, by showing that N400 reduction in response to strongly-constraining context was smaller and significantly delayed in older adults, especially those with lower reading spans (Federmeier & Kutas, 2005). Federmeier et al. (2010) further showed that the differences in using contextual cues was still visible between younger and older adults, even when the working memory load of the task was minimal, although older adults with higher verbal fluency scores showed ERP patterns that more closely resembled that of younger adults.

The ERP studies reviewed above suggest that older adults, as a group, are less likely to use contextual information to pre-activate potential target representations. To determine whether this age group showed any trace of context use for pre-activating linguistic information, DeLong, Groppe, Urbach, and Kutas, (2012) measured N400 responses to indefinite articles "a" and "an" for upcoming nouns whose cloze probability given the context was parametrically manipulated to be between 0 and 100% (e.g., "Dale was very sorry and knew he owed Mary *a check/an apology* for what he had done."). The main finding was that N400 in response to nouns showed graded sensitivity to nouns' cloze probability in both younger and older adults, but these N400 correlations had a later onset

and lasted longer in the older adults (see also Wlotko et al., 2012). Importantly, unlike younger adults, older adults did not show any effects on the articles, indicating that at that early point they had not yet anticipated the phonological form of the upcoming word. A second finding of the study was a prolonged increased frontal positivity to less likely nouns in young adults and a subset of older adults with high verbal fluency. Together, these findings led the authors to propose that older adults may in fact use contextual cues for pre-processing of the linguistic representations, even though the delayed timeline of this effect might lead one to conclude otherwise.

We test this possibility by comparing younger and older adults in a visual world eye-tracking paradigm. Participants viewed a scene of four objects, while listening to sentences such as "She will see a/an cherry/apple". On the Experimental trials, the indefinite articles "a" or "an" unambiguously cued the target, because the other three objects all had names that started with a phoneme that was incompatible with that article (e.g., for "an" the target would be "apple", and the three distractors would be "baby", "piano", and "duck"). On Control trials, participants heard similar sentences but with the definite article "the", which provides weaker cues to the target identity. If listeners use indefinite articles as cues, they should be able to fixate the target faster on the Experimental, compared to the Control, trials.

Importantly, "a" and "an" are semantically identical, and the disambiguating information they carry is phonological. Past research has suggested that younger adults use phonological cues to anticipate potential targets (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Zwitserlood & Schriefers, 1995). For example, upon hearing the onset /b/ listeners consider all the words that start with /b/ as possible referents (the *cohort effect*; Allopenna et al., 1998). Are phonological cues in the articles also used as cues? Interestingly, articles may not be fully processed in sentence comprehension. Readers often skip over articles in reading (O'Regan, 1979) and children older than age 4 show no disruption in sentence comprehension when an inappropriate article is used (Zangl & Fernald, 2007; McNamara, Carter, McIntosh, & Gerken, 1998). Thus articles, while potentially valuable phonological cues, may be skipped without causing much harm to comprehension. Thus, if older adults are less likely to process contextual cues, articles would be an excellent test bed.

The current experiment investigated whether indefinite articles are used as phonological cues to locate the referent during auditory sentence comprehension in younger and older adults, and whether the two groups process such cues differently.

Methods

Participants

Twelve younger adults (six females, mean age = 19.5, SE = 0.4 years), and twelve older adults (five females, mean age = 63.7, SE = 2.3 years) participated in the study in exchange for course credit or payment. All participants were native speakers of English. Older adults were tested on the Mini-Mental State Exam (Folstein, Folstein, & McHugh, 1975) and all scored within normal range.

Materials

Visual stimuli were presented as 300×300 pixel pictures of black and white line-drawings taken from either the IPNP corpus (Szekely et al., 2004), or from Snodgrass & Vanderwart (1980). Targets were 60 common nouns, half beginning with a vowel, and half with a consonant. There were no significant differences between the items in the two groups in frequency ($t(57) = 1.22$, $p = .23$; reported as frequency per million words, from SUBTLEX (Brysbaert & New, 2009)), number of syllables ($t(58) = -1.37$, $p = .18$), and number of phonemes ($t(58) = .27$, $p = .79$). Each item appeared once as the target in the Experimental condition ("an apple"), once as the target in the Control condition ("the apple"), and six more times as distractor in trials with other target nouns. One hundred and twenty sentences with the structure "She will see [article][target]." were recorded by a native English speaker at 44.1 kHz. All sentences were recorded naturally, without word splicing. This was necessary because the pronunciation of "the" could change depending on whether the following noun starts with a vowel or a consonant. Therefore, "a" and "an" should each have their own proper baseline of "the"; splicing would have removed this natural variability in "the" pronunciation and provided a biased baseline. In the recorded materials, there was no significant difference between the duration of the determiners and their paired "the" controls ($t(29) = .58$, $p = .56$ for "a" vs. "the"; $t(29) = 1.15$, $p = .23$ for "an" vs. "the").

Apparatus

Participants were seated approximately 25 inches away from a 17-inch monitor with the resolution set to 1024×768 dpi. Stimuli were presented using E-Prime Professional, Version 2.0 software (Psychology Software Tools, Inc., www.pstnet.com). A remote Eyelink 1000 eye-tracker recorded participants' monocular gaze position at 250 Hz.

Procedure

Participants were instructed to "listen and look at the pictures" (no response was required). Each trial began with a 1375 ms preview. In the first 1000 ms, the four line-

drawings were presented in the four corners, and in the last 375 ms a shrinking red dot appeared at the center to draw the gaze back to the central location. After the preview, the sentence was presented through speakers at a comfortable listening volume. The position of the four pictures was randomized on every trial. Participants first completed six practice trials, and then moved on to the experiment. Two blocks, each containing 60 intermixed "a", "an" and "the" trials were administered, with a break in between. No picture was repeated as the target within the same block.

Results

Figure 1 shows the proportion of fixation (±SE) on the target for the Experimental (a/an) and Control (the) conditions, separated by article type and age group. Statistical analyses were performed using the Growth Curve Analysis method (GCA; Mirman, 2014), a variant of multilevel regression (or hierarchical linear modeling) that uses orthogonal polynomials to capture the curvilinear pattern of fixation proportions over time. Effects of the variables of interest on the polynomial terms provide a way to quantify and evaluate those effects on statistically independent (i.e., orthogonal) aspects of the fixation proportions trajectory. Data were contrast coded and centered, and unless otherwise specified, the overall target fixation trajectory was modeled with a cubic polynomial. Dependent variables included Age (young vs. old), Condition (definite article "the" vs. indefinite articles "a"/"an") and ArtType (Article Type; type of the indefinite article: "a" vs. "an"). The interaction between these variables and polynomial terms were explored over the intercept, linear and quadratic terms, as higher order interactions are difficult to interpret (see Mirman et al., 2014 for a full discussion). The cubic term was, however, entered in both the fixed the random effect structures to obtain the best fit to the data.

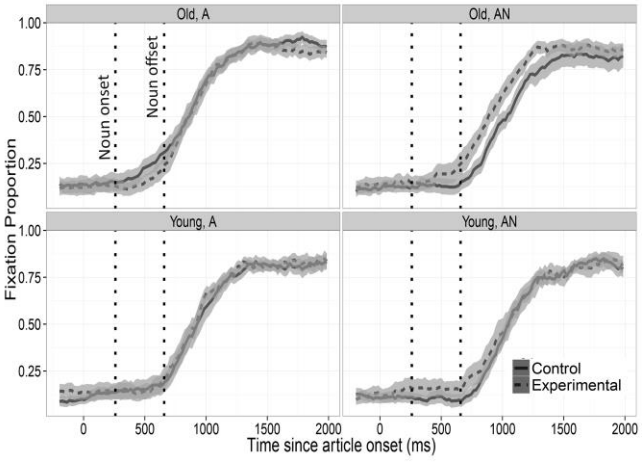


Figure 1: Proportion of fixations (±SE) to the target in Experimental (a/an) vs. control (the) conditions, separately for each article and each age group.

The first analysis addressed four questions: (a) Do phonological cues facilitate target localization? (b) Do both younger and older adults both use these cues? (c) Are there differences in using "a" and "an" cues? And (d) Do the two age groups both show these differences? Table 1 shows the full model results.

Table 1- Results of the GCA first analysis.

Fixed effect structure	Coefficient	SE	t	p-value
Intercept	0.403	0.025	16.012	<0.001
Linear	1.907	0.082	23.264	<0.001
Quadratic	0.405	0.059	6.872	<0.001
Cubic	-0.301	0.032	-9.341	<0.001
Intercept*Age	0.024	0.024	1.018	0.320
linear*Age	0.065	0.066	0.984	0.336
Quadratic*Age	-0.069	0.058	-1.187	0.248
Intercept*Condition	0.031	0.006	5.29	<0.001
Linear*Condition	0.073	0.025	2.878	0.005
Quadratic*Condition	-0.107	0.023	-4.69	<0.001
Intercept*ArtType	-0.013	0.006	-2.265	0.027
Linear*ArtType	-0.023	0.025	-0.909	0.367
Quadratic*ArtType	0.027	0.023	1.188	0.239
Intercept*Condition*Age	-0.002	0.006	-0.309	0.758
Linear*Condition*Age	<0.001	0.025	-0.019	0.985
Quadratic*Condition*Age	0.002	0.023	0.084	0.933
Intercept*Condition*ArtType	0.021	0.006	3.451	0.001
Linear*Condition*ArtType	-0.005	0.025	-0.212	0.832
Quadratic*Condition*ArtType	-0.05	0.023	-2.19	0.032
Intercept*ArtType*Age	-0.001	0.006	-0.091	0.928
Linear*ArtType*Age	-0.051	0.025	-2.024	0.047
Quadratic*ArtType*Age	-0.008	0.023	-0.332	0.740
Intercept*Condition* ArtType*Age	0.011	0.006	1.847	0.069
Linear*Condition*ArtType*Age	0.033	0.025	1.311	0.194
Quadratic*Condition*ArtType*Age	-0.024	0.023	-1.068	0.289
Random effect structure	**Variance**	**Correlation**		
Subject				
Intercept	0.012			
Linear	0.124	-0.17		
Quadratic	0.061	-0.7	0.4	
Cubic	0.013	-0.35	-0.67	0.2
Condition*ArtType|Subject				
Intercept	0.003			
Linear	0.063	0.38		
Quadratic	0.044	-0.43	0.01	
Cubic	0.030	-0.24	-0.47	0.26
Residual	0.002			

The effect of Condition was significant on the intercept, linear and quadratic terms, showing that the phonological cues provided by the indefinite articles reliably facilitated localization of the target compared to the less phonologically-informative "the". This effect was not reliably different between older and younger adults, as evidenced by non-significant Condition*Age interactions on all three polynomial terms, implying that both age groups used these cues. We then asked if facilitation was more pronounced for one type of the definite article over the other. The interaction between Condition and ArtType was significant over two of the three polynomial terms (intercept and quadratic), indicating that the article "an" was more facilitatory than the article "a", when each article was compared to its own baseline "the". Finally, we asked

whether younger and older adults differed in their processing of "a" vs. "an" articles. We found a marginal effect of the three-way interaction (Condition by ArtType by Age) on the intercept, hinting at a possible difference.

In summary, the results of this analysis provided strong evidence that both younger and older adults used phonological cues associated with the definite articles to more efficiently locate the visual target, but also suggested that there might be differences in how the two age groups used such cues. These differences were explored in the next set of analyses.

The second set of analyses focused on separate exploration of facilitation in "an" and "a". The model structure was similar to that of the first analysis, except that ArtType was removed since only subsets of data containing either "an" or "a" were analyzed by each model. Table 2 presents the results of the analysis on the "a" dataset. Neither the effect of Condition, nor its interaction with age, was significant on any of the polynomial terms.

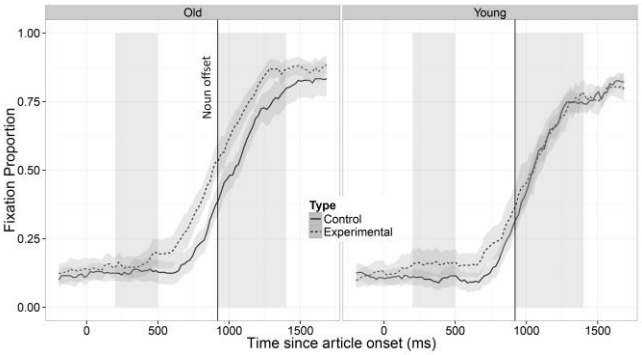


Figure 2: Proportion of fixations (±SE) on the target in Experimental (an) vs. control (the) conditions in each age group. Gray rectangles mark the two analysis time windows.

Table 3 shows the results of the analysis on the "an" subset. The effect of Condition was significant on the intercept and quadratic terms, showing that participants used "an" as a more reliable cue to locate the target compared to "the". Importantly, the interaction between Condition and age was also significant on the linear term. These results indicate that while, as a group, younger and older adults both use "an" as an informative cue for finding the referent, they do so differently. The next set of analyses further explored these differences by focusing on early and late time windows.

Early time window. This analysis focused on a narrow window of 300 ms, starting 200 ms after the presentation of the article (to allow for planning and execution of an eye movement), and ending 240 ms after the onset of the word. This window best captures early use of phonological cues for anticipatory target localization. We found a main effect of Condition on the intercept, such that "an" facilitated target fixation compared to "the" (coefficient = -0.022, SE = 0.009; t = -2.471; p = 0.022). Critically, we also found a reliable interaction between Condition and Age on the quadratic term (coefficient = -0.018; SE = 0.006; t = -3.260; p = 0.003). This interaction suggests that the anticipatory effect of phonological cue was reliably more pronounced in the younger, compared to the older, adults (See Fig. 2).

Late time window. This time window was chosen to reflect the opposite of an anticipatory process, namely to explore if phonological cues were still being used even when the target noun had been completely spoken and there was no longer any ambiguity about the referent. To this end, we chose a time window starting at the end of the noun (920 ms after the article onset) until the point where fixations plateaued (1400 ms after the article onset). This analysis revealed a marginal main effect of Condition on the intercept (coefficient = -0.033, SE = 0.016; t = -2.058; p = 0.052), as well as a significant interaction between Condition and Age on the quadratic term (coefficient = 0.028; SE = 0.011; t = 2.478; p = 0.018) and a marginal effect of this interaction on the intercept (coefficient = -0.028; SE = 0.016; t = -1.782; p = 0.089). Importantly, the direction of this interaction was the opposite of that found in the early time window analysis, showing that in this late time window, older adults continued to rely on the earlier phonological cues from the article to find the referent, considerably more than younger adults (see Fig. 2).

Table 2- Results of the GCA on the "a" subset.

Fixed effect structure	**Coefficient**	**SE**	**t**	**p-value**
Intercept	0.44	0.025	17.467	<0.001
Linear	2.026	0.076	26.493	<0.001
Quadratic	0.276	0.07	3.931	<0.001
Cubic	-0.376	0.035	-0.813	<0.001
Intercept*Age	0.025	0.025	1.021	0.319
Linear*Age	0.079	0.066	1.197	0.245
Quadratic*Age	-0.068	0.069	-0.983	0.337
Intercept*Condition	0.008	0.008	0.948	0.354
Linear*Condition	0.015	0.044	0.354	0.727
Quadratic*Condition	-0.022	0.028	-0.799	0.433
Intercept*Condition*Age	0.011	0.008	1.323	0.2
Linear*Condition*Age	-0.01	0.044	-0.222	0.826
Quadratic*Condition*Age	-0.032	0.028	-1.137	0.268
Random effect structure	**Variance**	**Correlation**		
Subject				
Intercept	0.012			
Linear	0.034	-0.97		
Quadratic	0.085	-0.7	0.7	
Cubic	0.005	-0.24	0.24	0.17
Condition|Subject				
Intercept	0.003			
Linear	0.175	0.4		
Quadratic	0.034	-0.55	-0.12	
Cubic	0.039	-0.32	-0.77	0.32
Residual	0.003			

In summary, the second set of analyses showed that (1) the effect of phonological cue was stronger in "an" than in "a", (2) both younger and older adults used "an" as a cue to localize the visual referent, and (3) the two age groups used

the cue with notably different timelines; early in younger and late in older adults.

Table 3- Results of the GCA on the "an" subset.

Fixed effect structure	Coefficient	SE	t	p-value
Intercept	0.367	0.026	14.299	<0.001
Linear	1.787	0.08	22.423	<0.001
Quadratic	0.534	0.056	9.613	<0.001
Cubic	-0.226	0.031	-7.26	<0.001
Intercept*Age	0.024	0.025	0.96	0.337
Linear*Age	0.072	0.075	0.957	0.338
Quadratic*Age	-0.066	0.053	-1.229	0.219
Intercept*Condition	-0.033	0.01	-3.299	0.001
Linear*Condition	-0.017	0.037	-0.45	0.653
Quadratic*Condition	0.076	0.036	2.095	0.036
Intercept*Condition*Age	-0.006	0.01	-0.608	0.543
Linear*Condition*Age	-0.079	0.037	-2.13	0.033
Quadratic*Condition*Age	0.01	0.036	0.278	0.781
Random effect structure	**Variance**	**Correlation**		
Subject				
Intercept	0.011			
Linear	0.097	-0.14		
Quadratic	0.034	-0.85	0.52	
Cubic	0.007	-0.21	-0.29	0.22
Condition|Subject				
Intercept	0.006			
Linear	0.071	0.62		
Quadratic	0.059	-0.31	0.05	
Cubic	0.024	-0.52	-0.5	0.34
Residual	0.002			

General Discussion

Our results suggest that articles were indeed used as informative contextual cues for localizing the referent during online sentence comprehension regardless of age. This finding suggests that while comprehension does not critically depend on articles, listeners do use these function words as cues. However, this benefit was limited to the less frequent article "an". Note that each article was compared to its corresponding control condition (e.g., "a cherry" is compared to "the cherry" and "an apple" to "the apple"), thus subtle differences in the pronunciation of "the" when followed by a consonant or a vowel were controlled for. The most likely reason for "an" being used as a more prominent cue than "a", is that the subset of words that "an" cues (i.e., words that start with a vowel) is much smaller than the subset of words that are cued by "a". Also in modern American English, the use of "a" along with a noun that starts with a vowel, though infelicitous, is not uncommon, especially in disfluent speech (e.g., "He's a um… artist!"), decreasing the validity of "a" as a unique cue.

Critical for our investigation was that older adults too showed evidence of employing these article cues for locating the referent. This finding complements the literature on contextual cue processing in older adults, which has almost entirely focused on the processing of semantic context. Note that studies that investigate ERP responses on the article (e.g., Delong et al., 2005, 2012), investigate retrieval of the phonological form in response to semantic cues in the sentence, e.g., retrieving the word "kite" or "airplane" (and subsequently the appropriate article a/an), given the context "The day was breezy so the boy went outside to fly ...". Thus these studies follow a different goal from the present study, which tested whether phonological cues associated with the articles themselves are used by listeners. Our results suggest that older adults do use such phonological cues, even on function words that past research suggests are not critical for sentence comprehension (e.g., McNamara et al., 1998; Zangl & Fernald, 2007).

However, our analyses of early and late time windows revealed intriguing differences in the timeline of cue processing between younger and older adults. While younger participants used phonological article cues in an anticipatory fashion, i.e., to locate the target *before* the target name was spoken, older adults showed no evidence of anticipatory cue use. On the other hand, older adults showed continued advantage for the target when it followed "an", compared to "the", in the late time window, *after* the noun had been spoken. In this late time window, younger adults showed no difference in fixating targets that were preceded by either article type, presumably because the noun information was enough for unequivocally localizing the referent in both conditions.

The delay in the processing of phonological article cues in older adults mirrors the reports of delayed N400 effect during semantic context processing (e.g., Delong et al., 2012; Federmeier et al., 2003; 2010; Wlotko & Federmeier, 2012; Wlotko et al., 2012). Moreover, older adults' continued processing of such cues in a late time window is compatible with Delong et al.'s (2012) report of prolonged N400 in this population (see also Wlotko et al., 2012). Our results add to these findings in three ways. First, the delay in processing contextual cues in older adults is not limited to semantic cues. This shows that the issue does not stem from specific dynamics within the lexical-semantic system, and is instead be better explained by models of general slowing in cognitive processing with aging (e.g., Myerson et al., 1990). Importantly, this slowing does not encompass automatic aspects of processing, such as activation of lexical representations through related representations (e.g., Federmeier et al., 2003; see Burk & Shafto, 2008, for a review), but seems to be specific to operations that require active processing and control.

The second way that the current results add to the findings of ERP studies is that it corroborates those findings using a measure that is sensitive enough to capture subtle differences in the timeline of cognitive processing, while still capturing overt behavior. We intentionally used articles which, according to past research, were dispensable to comprehension to show that even such subtle cues can be used in older adults and have measurable behavioral consequences. Finally, the sentences used in the current experiment entailed no anomaly or syntactic complexity, thus providing a straight-forward test of contextual cue use in sentence comprehension.

In summary, these results, together with those of past studies, support a model of sentence comprehension in which the use of semantic and phonological contextual cues to process upcoming words continues in old age, but the efficiency of using such cues for anticipatory word retrieval decreases considerably with aging.

Acknowledgments

The authors to thank Allison Britt and Kristen Graziano for their help with recordings and data collection. This research was funded in part by NIH grant R01DC010805 to DM and by the Moss Rehabilitation Research Institute.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Brysbaert, M., & New, B. (2009). Moving beyond Ku\vcera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Burke, D. M., & Shafto, M. (2008). Language and aging. In F. I. M. Craik & T. A. Salthouse (Eds.). Handbook of aging and cognition (Vol. 3, pp. 373–444). New York: Psychology Press.

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, *121*(3), 226–239.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121.

Federmeier, K. D., & Kutas, M. (2005). Aging in context: Age-related changes in context use during language comprehension. *Psychophysiology*, *42*(2), 133–141.

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161.

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146.

Federmeier, K. D., Van Petten, C., Schwartz, T. J., & Kutas, M. (2003). Sounds, words, sentences: age-related changes across levels of language processing. *Psychology and Aging*, *18*(4), 858.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470.

Madden, D. J. (1988). Adult age-differences in the effects of sentence context and stimulus degradation during visual word recognition. Psychology and Aging, 3(2), 167–172.

McNamara, M., Carter, A., McIntosh, B., & Gerken, L. (1998). Sensitivity to grammatical morphemes in children with specific language impairment.*Journal of Speech, Language, and Hearing Research*, *41*(5), 1147-1157.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. Florida, USA: Chapman & Hall/CRC Press.

Myerson, J., Hale, S., Poon, L. W., Wagstaff, D., & Smith, G. A. (1990). The information-loss model – A mathematical-theory of age-related cognitive slowing. Psychological Review, 97(4), 475–487.

O'Regan, K. (1979). Saccade size control in reading: Evidence for the linguistic control hypothesis. *Perception & Psychophysics*, *25*(6), 501–509.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Stine-Morrow, E. A., Miller, L. M. S., & Nevin, J. A. (1999). The effects of context and feedback on age differences in spoken word recognition. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54*(2), P125–P134.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., … Wicha, N. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250.

Wlotko, E. W., & Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology*, *49*(6), 770-785.

Wlotko, E. W., Federmeier, K. D., & Kutas, M. (2012). To predict or not to predict: Age-related differences in the use of sentential context. *Psychology and aging*, *27*(4), 975.

Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, *3*(3), 199–231.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes*, *10*(2), 121–136.

Scarcity captures attention and induces neglect: Eyetracking and behavioral evidence

Brandon M. Tomm (brandon.tomm@psych.ubc.ca)
Department of Psychology, University of British Columbia
2136 West Mall, Vancouver, BC V6T 1Z4, Canada

Jiaying Zhao (jiayingz@psych.ubc.ca)
Department of Psychology, and Institute for Resources, Environment and Sustainability
University of British Columbia, 2136 West Mall
Vancouver, BC V6T 1Z4, Canada

Abstract

Resource scarcity poses challenging demands on the human cognitive system. Budgeting with limited resources induces an attentional focus on the problem at hand. This focus enhances processing of relevant information, but it also comes with a cost. Specifically, scarcity may cause a failure to notice beneficial information that helps alleviate the condition of scarcity. In three experiments, participants were randomly assigned with a small budget ("the poor") or a large budget ("the rich") to order a meal from a restaurant menu. The poor participants looked longer at the prices of the items and recalled the prices more accurately, compared to the rich participants. Importantly, the poor neglected a useful discount that would save them money. This neglect may arise as a result of attentional narrowing, and help explain a range of counter-productive behaviors of low-income individuals. The current findings have important implications for public policy and services for low-income individuals.

Keywords: Poverty; visual attention; memory; encoding; decision making;

Scarcity is the condition of having insufficient resources to cope with demands, and is an urgent and pervasive problem in the world: Roughly 1.2 billion people live in extreme poverty with less than $1.25 a day, 1.3 billion people live without electricity, and more than 780 million lack access to clean water. This condition presents significant challenges to the human cognitive system. For example, having limited financial resources requires the meticulous calculation of any expenses. Similarly, having limited time requires stringent management of schedules.

The cognitive consequences of scarcity are recently revealed by a number of studies (Mullainathan & Shafir, 2013). For example, scarcity causes myopic behavior which results in the neglect of future events (Shah, Mullainathan, & Shafir, 2012). Specifically, people under scarcity tend to prioritize the task at present and over-borrow resources from the future. Financial scarcity directly impairs cognitive function, reducing fluid intelligence and the ability to exert cognitive control (Mani, Mullainathan, Shafir, & Zhao, 2013). These cognitive and behavioral consequences are particularly problematic because these impairments can lead to suboptimal decision making and behaviors (e.g., poor time management or financial planning skills) that further perpetuate the condition of scarcity.

Currently, it is still unclear what cognitive mechanisms underlie the impairments caused by scarcity. A possible explanation of these impairments is that scarcity presents urgent demands that hijack attentional resources, causing a strong focus on the present task. Such focus can induce a neglect of other potentially important information.

Support for this explanation comes from the previous theoretical and empirical work on the limits of the cognitive system. Specifically, the cognitive system has a finite capacity, and people can only receive and process a limited amount of information at a time (Baddeley, 1992; Luck & Vogel, 1997; Miller, 1956; Pashler, Johnston, & Ruthruff, 2001). Given this limited capacity, engaging in one process consumes cognitive resources needed for another, thus causing interference. For example, studies on inattentional blindness (Simons & Chabris, 1999; Neisser, 1979) show that performing a demanding task (e.g., counting how often the basketball is passed around) results in an inability to notice a salient event (e.g., a man dressed as a gorilla passing by). Basic visual features of unattended stimuli may not even be perceived (Rock & Gutman, 1981). In addition to perception, this interference can cause serious behavioral consequences such as impaired driving (Strayer, Drews, & Johnston, 2003). The limited cognitive resources given competing demands can thus result in attentional trade-offs between focus and neglect.

Here, we propose that scarcity forces attentional trade-offs. Specifically, people operating under scarcity may prioritize urgent tasks at hand, leaving other information unattended. This process can be counter-productive because the attentional neglect can cause the failure to notice useful and beneficial information in the environment that alleviates the condition of scarcity. To investigate the attentional trade-offs under scarcity and the resulting memory performance of task-relevant information, we conducted three experiments in the current study.

Experiment 1

The goal of this experiment was to examine the effects of scarcity on visual attention. We predict that scarcity draws attention to the task-relevant information, but at the same time, also causes the neglect of other useful information in the environment.

Participants

One hundred and ninety undergraduate students (152 female, 35 male, 3 unspecified; mean age = 20.39 years, *SD* = 3.92 years) were recruited from the Human Subject Pool at the Department of Psychology at the University of British Columbia (UBC), and participated in the experiment for course credit. Participants in all experiments reported normal or corrected-to-normal vision and provided informed consent. All experiments reported here were approved by the UBC Behavioral Research Ethics Board.

Stimuli and Procedure

Participants were presented with a restaurant menu which contained 24 food items. For each item, the price and the calories were listed in two columns on the menu (Fig.1). The menu subtended 12.4° of visual angle in width and 16.2° in height. A discount clause was shown on the bottom of the menu ("You may ask for an 18% student discount.").

Participants were randomly assigned with a small budget ($20; the poor condition) or a large budget ($100; the rich condition). Thus, the experiment used a between-subjects design.

Participants were asked to view the items on the menu and think about what they would like to order, as if they were ordering a meal from a restaurant. They were given unlimited time to place the order, and were told not to exceed the assigned budget, but they were not required to spend the entire budget.

The eye gaze of each participant was monitored throughout the experiment using an SMI RED-250 Mobile Eyetracking System (60hz). To examine which part of the menu was prioritized, the menu was divided into four areas of interest: the left column of the food items, the middle column of the price information, the right column of the calories information, and the discount clause.

To measure visual attention, we calculated the dwell time and the number of fixations in each area of interest. Since there was no time limit in the experiment, we used the proportional dwell time (the dwell time spent in each area divided by the total dwell time on the menu) and the proportional fixations (the number of fixations in each area divided by the total number of fixations on the menu) as two measures of visual attention. The heat maps of the average duration of dwell time between the poor and the rich conditions were shown in Figures 1 and 2.

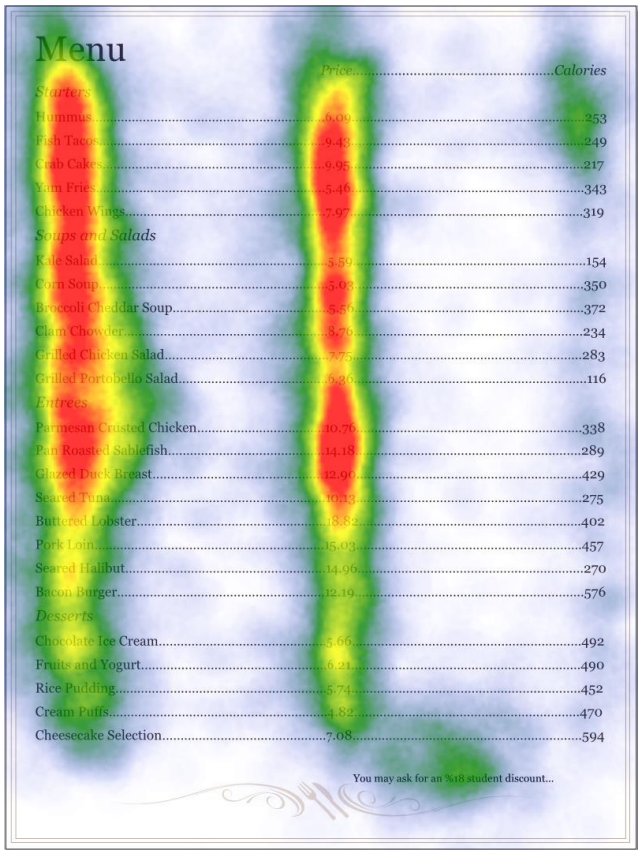


Figure 1. Experiment 1. A heat map of the menu showing the distribution of the average dwell time for the participants in the poor condition (who ordered a meal with $20). Warmer colors represent longer average dwell time.

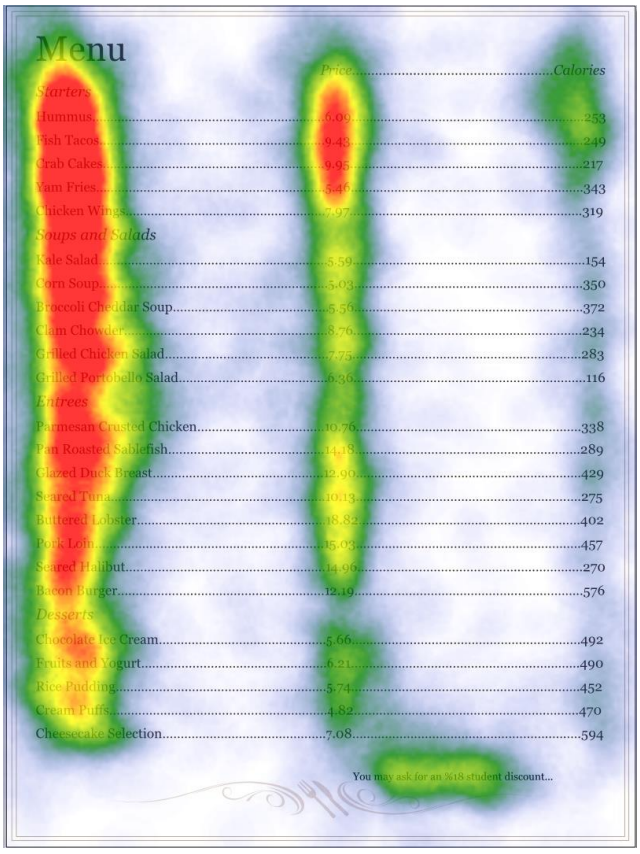


Figure 2. Experiment 1. A heat map of the menu showing the distribution of the average dwell time for the participants in the rich condition (who ordered a meal with $100). Warmer colors represent longer average dwell time.

Results and Discussion

First, we observed that participants in the rich condition took more time to order (M = 89 seconds) than the participants in the poor condition (M = 76 seconds) [$t(188) = 2.24$, $p = .03$, $d = .33$]. This provided motivations for using the proportional dwell time and fixations in the following analyses between the two conditions. Moreover, participants with more than 3 standard deviations away from the mean in each measure were excluded (between 1 and 4 participants in total, depending on the measure).

For the food items (Fig.3), participants in the poor condition spent less dwell time (M = 35.11%) than those in the rich condition (M = 51.98%) [$t(185) = 3.91$, $p < .001$, $d = .57$]. The poor also made few fixations (M = 36.66%) on the food items than the rich (M = 48.50%) [$t(187) = 3.08$, $p = .002$, $d = .45$]. This suggests that the poor participants spent less time considering which food items they would like to order than the rich participants did.

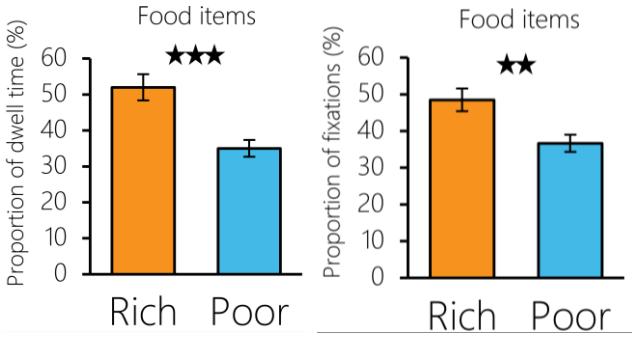


Figure 3. The proportional dwell time and fixations on food items between participants in the poor and the rich conditions (error bars reflect ±1 SEM; **p<.01, ***p<.001).

For prices (Fig.4), participants in the poor condition spent more dwell time (M = 21.08%) than those in the rich condition (M = 15.23%) [$t(185) = 2.16$, $p = .03$, $d = .32$]. Similarly, participants in the poor condition made more fixations on prices (M = 23.07%) than those in the rich condition (M = 15.81%) [$t(185) = 2.91$, $p < .01$, $d = .43$]. This suggests that the poor attended more to prices than the rich participants.

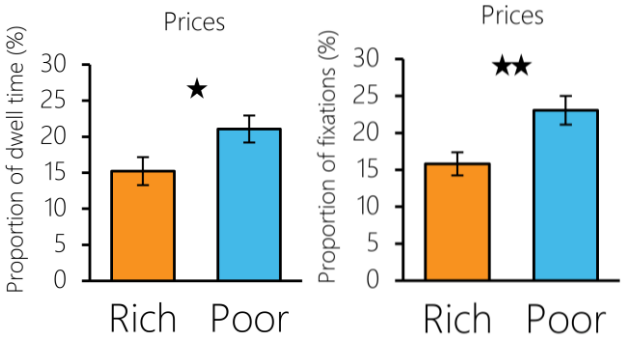


Figure 4. The proportional dwell time and fixations on prices between participants in the poor and the rich conditions (error bars reflect ±1 SEM; *p<.05, **p<.01).

This result could be driven by the possibility that scarcity enhanced attention to all numerical information. Thus, we examined attention to the calorie information (Fig.5). Participants in the poor condition spent less dwell time on calories (M = 2.92%) than those in the rich condition (M = 4.35%) [$t(185) = 2.65$, $p < .01$, $d = .39$]. The poor (M = 3.51%) also made fewer fixations on calories than the rich did (M = 5.09%) [$t(184) = 2.39$, $p = .02$, $d = .35$]. These results indicate that financial scarcity draws attention only to prices and induces a neglect of food items and calories.

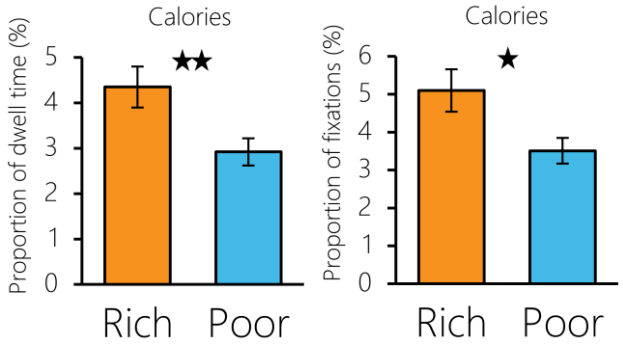


Figure 5. The proportional dwell time and fixations on calories between participants in the poor and the rich conditions (error bars reflect ±1 SEM; *p<.05, **p<.01).

Importantly, for the discount clause (Fig.6), participants in the poor condition spent less dwell time (M = 0.83%) than those in the rich condition (M = 1.83%) [$t(184) = 3.51$, $p < .001$, $d = .52$]. The poor also made fewer fixations on the discount clause (M = 0.85%) than the rich (M = 1.82%) [$t(184) = 3.51$, $p < .001$, $d = .52$]. This suggests that the poor neglected the discount more than the rich participants. However, there was no difference in the number of poor (47%) and rich (55%) participants who looked at the discount [$X^2 = 1.35$, $p = .25$]. This suggests that the poor and the rich are equally likely to look at the discount, but the poor spent less time looking at the discount.

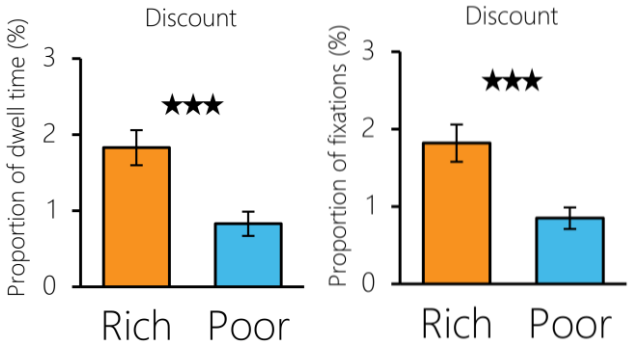


Figure 6. The proportional dwell time and fixations on the discount clause between participants in the poor and the rich conditions (error bars reflect ±1 SEM; ***p<.001).

As another measure of attention to the discount, after placing the order the participants were asked if they had noticed other information on the menu besides the price and

calorie information. While measures of visual attention showed that the poor looked less at the discount than the rich did, there was no reliable difference between the number of poor participants (38%) and rich participants (35%) who explicitly reported noticing the discount during debriefing of the experiment [X^2 = .09, p = .76]. However, less than 5 participants spontaneously requested the discount during their order. Thus, the attentional neglect of the discount was largely implicit.

We also analyzed the absolute proportional dwell time and fixation count for all analyses, which yielded nearly identical results. These absolute measures were not bounded, and therefore were not subject to trade-offs within a limit. This means that although the poor dwelled longer on price information, they did not necessarily dwell less on the discount.

An alternative explanation for the finding that the poor attended less to the discount was that scarcity might result in more efficient processing of task-relevant information. This would suggest that the poor did not need to look at the discount as much as the rich, because they were faster in seeing the discount. Since both the prices and the discount were task-relevant, this explanation would predict that the poor would be efficient in processing both price information and the discount. However, we found that the poor looked more at the prices but less at the discount than the rich did, which could not be explained by the efficiency account.

The discount on the menu could in theory help the poor participants save money and stay within their budget. Despite this usefulness, the poor participants still neglected the discount and focused more on the prices of the food items. This finding is ironic and could help explain why the low-income individuals engage in neglectful behaviors that are counter-productive.

Experiment 2

Experiment 1 demonstrated that financial scarcity prioritizes the processing of price information, at the cost of other useful information. Given the attentional prioritization of prices, we predict that memory encoding of prices will also be enhanced. This prediction is supported by the recent work that suggests that visual working memory can be construed as visual attention preserved internally over time (Chun, 2011; Chun, Golomb, & Turk-Browne, 2011). Feature-based theories of attention also predict selective facilitation in visual processing for task-relevant features (Hayden & Gallant, 2008; Jehee, Brady, & Tong, 2011).

Thus, in Experiment 2 we examined the effects of scarcity on memory encoding, as a result of attentional prioritization. We predict that financial scarcity facilitates memory encoding specifically for price information, and not for other types of information.

Participants

A new group of 60 undergraduate students (43 female, 17 male; mean age = 19.95 years, *SD* = 2.30 years) from UBC participated in the experiment for course credit.

Stimuli and Procedure

To increase the demand for memory encoding, we increased the number of items on the menu. Participants were presented with a menu which now contained 50 food items. Similar to Experiment 1, the menu included the price and calories for each food item.

As in Experiment 1, participants were asked to place a meal order from the menu as if they were ordering from a restaurant. As before, participants were randomly assigned with a small budget ($20; the poor condition) or a large budget ($100; the rich condition). The experiment again used a between-subjects design.

After participants placed their order, they were given a surprise memory test. Participants were asked to recall as many items from the menu as possible. For each item recalled, they were also asked to recall the price and the calorie information of the item as accurately as possible.

Results and Discussion

To measure memory encoding, we calculated the average absolute error between the recalled prices (and calories) and the objective prices (and calories) for each participant (Fig.7). Participants in the poor condition (*Mean error* = $1.32) were reliably more accurate in the price recall than those in the rich condition (*Mean error* = $2.17) [$t(58) = 2.35$, $p = .02$, $d = .61$]. However, there was no reliable difference in the calorie recall between the poor and the rich participants [$t(57) = .80$, $p = .42$, $d = .21$].

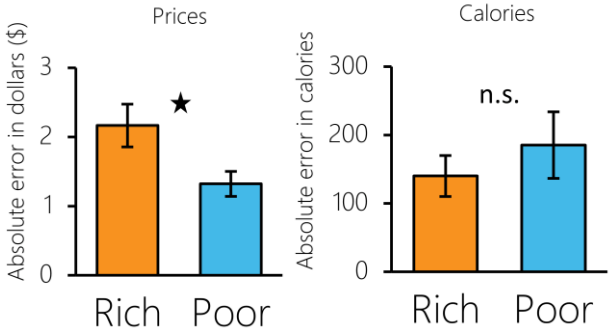


Figure 7. The absolute error in the price recall and the calorie recall between participants in the price poor and the price rich conditions (error bars reflect ±1 SEM; *p<.01).

This enhanced performance in price recall in the poor cannot be explained by the fact that the poor participants ordered fewer items (M = 2.13) than the rich (M = 3.67) [$t(58) = 3.74$, $p < .001$, $d = .98$]. First, there was no reliable difference in the number of recalled times between the poor and the rich [$t(58) = 1.31$, $p = .20$, $d = .34$]. Second, there was no difference in the time taken to place the order between the poor and the rich participants [$t(58) = 1.44$, $p = .16$, $d = .37$]. Third, even if ordering fewer items might improve memory recall, this benefit would be seen in both price and calorie recall, but we found that the poor were more accurate only in price recall, not in calorie recall.

Thus, these findings suggest that financial scarcity improves memory encoding for task-relevant information (i.e., prices), but not for task-irrelevant information (i.e., calories). Scarcity selectively facilitates memory encoding.

Experiment 3

To generalize the findings in Experiment 2 to a different domain, we examined how calorie scarcity affects memory encoding. We predict that calorie scarcity facilitates memory encoding specifically for calorie information, and not for other types of information.

Participants

A new group of 60 undergraduate students (49 female, 11 male; mean age = 20.03 years, *SD* = 2.11 years) from UBC participated in the experiment for course credit.

Stimuli and Procedure

The stimuli and the procedure were identical to those in Experiment 2, except for a critical difference. Participants were randomly assigned with a small calorie budget (500 calories; the poor condition) or a large calorie budget (2000 calories; the rich condition). As before, participants were then asked to place a meal order from the menu as if they were ordering from a restaurant. After the order, participants were given a surprise memory test, where they recalled items from the menu with the price and calorie information.

Results and Discussion

To measure memory encoding, we calculated the average absolute error between the recalled calories (and prices) and the objective calories (and prices) for each participant (Fig.8). Participants in the poor condition (*Mean error* = 48.05 calories) were reliably more accurate in the calorie recall than those in the rich condition (*Mean error* = 71.61) [$t(58) = 2.27$, $p = .03$, $d = .58$]. This suggests that the calorie poor showed better memory encoding of calorie information than the calorie rich.

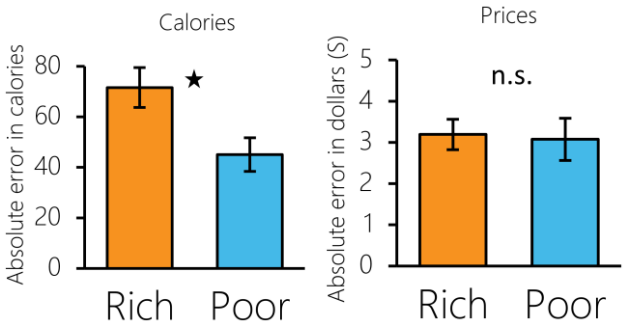


Figure 8. The absolute error in the calorie recall and the price recall between participants in the calorie poor and the calorie rich conditions (error bars reflect ±1 SEM; *p<.01).

A critical test of our prediction was whether this memory facilitation is specific to task-relevant information (i.e., calories). We found that there was no reliable difference in the price recall between the two conditions [$t(58) = 0.19$, $p = .85$, $d = .04$]. Thus, memory encoding was selectively enhanced for the calorie information in the calorie poor participants.

Interestingly, we did not observe worse memory encoding for the neglected task-irrelevant information. That is, the calorie recall was the same for the price poor and the price rich in Experiment 2, and the price recall was the same for the calorie poor and the calorie rich in Experiment 3. Even though Experiment 1 suggested that task-irrelevant information was neglected, here we did not see neglect influenced memory encoding of task-irrelevant information.

General Discussion

The goal of the current study was to examine how scarcity forces attentional trade-offs and influences memory encoding driven by such trade-offs. When operating under a limited financial budget, the poor focused more on the price information, compared to the rich (Experiment 1). This focus came with the neglect of other information in the environment, even if the information could be useful or beneficial to the poor (e.g., the discount). The attentional prioritization of prices also resulted in enhanced memory encoding of price information among the poor participants (Experiment 2). Likewise, the attentional prioritization of calories led to better memory encoding of calorie information among the calorie poor (Experiment 3).

The current findings provide a new perspective on how scarcity shapes the way people perceive and experience the external environment. While the perceptual experiences can be largely characterized by information overload, scarcity selectively orients people's attention to specific aspects of the environment. When operating with financial constraints, people automatically prioritize price-relevant information. Such prioritization facilitates memory encoding of these information, but crucially it comes with a cost, which is the neglect of the beneficial discount on the menu. An alternative explanation is that the poor spent less time looking at the discount because it was highly accessible to them due to high relevance. However, this explanation fails to account for the prioritization of the price information which is also highly relevant to the poor.

The findings also suggest that the rich with abundant resources have the luxury to attend broadly to the environment, compared to those under scarcity. The rich participants in Experiment 1 attended more to the food items, the calorie information of the food, and the discount, than the poor did. This means that the rich are able to consider which food they want to eat, rather than which food they can afford. In addition, they can also consider health-related (calorie) information of the food. Overall, the rich can attend to and process more types of information in the environment to make an informed choice.

The current study also reveals a painful irony of scarcity. People with limited resources were too focused on prices, such that they paid less attention to the beneficial discount that could save money and alleviate the financial burden. This

irony can help explain why low-income individuals sometimes engage in neglectful behaviors that are counter-productive (e.g., missing an appointment for a health checkup, or failure to sign up for benefit programs).

It is worth noting that the current experiments involved an artificial simulation of scarcity in the lab. In fact, just by randomly assigning people to receive a hypothetical small or large budget, we observed a strong effect of scarcity on attention and memory. Moreover, the participants in our experiments were not provided with real money, were not rewarded for frugality, and knowingly were not to receive any food from the menu. In the absence of possible consequences of their decisions, the poor participants still focused on task-relevant information and neglected help to alleviate the condition of scarcity. This raises the possibility that, outside the lab when people operate with scarce resources and can face real consequences of their actions, the effects of scarcity on attention and cognition observed in this study may be amplified.

The current findings can help inform public policy and services targeting low-income populations. Among the OECD countries, enrollment in social assistance and public benefit programs is estimated to range between 40% and 80% (Hernanz, Malherbet, & Pellizzari 2004). Our current study provides a new explanation for the low participation rate. That is, the poor who are eligible for these programs fail to participate because of the attentional trade-offs under scarcity. Low-income individuals may need to focus on their financial challenges and deadlines under scarcity, and either are not aware of these benefit programs and services, or neglect the enrollment procedures. This attentional account is not the only factor that can explain the low participation rate, as there are many other social barriers and stigmas related to enrollment in assistance programs.

Given the attentional constraints under scarcity, we propose that social assistance and public benefit programs should be designed to avoid the attentional neglect in the poor under scarcity. It may be helpful to streamline assistance applications and services to make them more salient, more accessible, and easier to process for the poor. The amount of effort and attention required from the poor should be minimized to increase or maintain participation. Benefit programs and social services can also be made more salient by using prompts and reminders. This could be done through any messaging medium such as text-message or email, and could be effective in catching the attention of those living under scarcity. Based on our current findings, future research can design behavioral interventions to avoid attentional neglect in the poor.

Acknowledgements

We thank Bram Scheerlinck and Desmond Fung for assistance with data collection. We also thank Eldar Shafir, Anuj Shah, Darko Odic, Alan Kingstone, Rebecca Todd, and the members in the Zhao Lab for helpful comments. This work was supported by Sloan Foundation Grant (2014-6-16), the Canada Research Chairs program (to JZ), and the Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ).

References

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation*, *8*, 47-89.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*, 1407-1409.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, *62*, 73-101.

Hernanz, V., Malherbet, F., and Pellizzari, M. (2004). Take-Up of Welfare Benefits in OECD Countries: A Review of the Evidence. *OECD Social, Employment and Migration Working Papers*, *17*, OECD Publishing, Paris.

Hayden, B. Y., & Gallant, J. L. (2009). Combined effects of spatial and feature-based attention on responses of V4 neurons. *Vision research*, *49*, 1182-1187.

Jehee, J. F., Brady, D. K., & Tong, F. (2011). Attention improves encoding of task-relevant features in the human visual cortex. *The Journal of Neuroscience*, *31*, 8210-8219.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279-281.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*, 976-980.

Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much.* Henry Holt and Company, New York.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, *52*, 629-651.

Rock, I., & Gutman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 275.

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*, 682-685.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059-1074.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, *9*, 23-32.

The Influence of Religious Beliefs on False Memory of Fabricated Events

Ellen Searle (eksearle@oakland.edu)
Department of Psychology, 2200 N Squirrel Rd
Rochester, MI 48309

Jennifer Vonk (vonk@oakland.edu)
Department of Psychology, 2200 N Squirrel Rd
Rochester, MI 48309

Brock Brothers (brockrbrothers@gmail.com)
Department of Psychology, 2200 N Squirrel Rd
Rochester, MI 48309

Abstract

Previous research has indicated that memories can be modified in conjunction with one's attitudes, in particular, political beliefs. The current study extended this finding by focusing on the relationship between differing religious beliefs and false memories for news events. We predicted that religious people would be more inclined to remember fabricated news events positively depicting religion and less likely to remember events negatively depicting religion compared to non-religious people. Opposite effects were predicted for events depicting atheism. In contrast, we found that religious people were more likely to falsely remember both events depicting religion positively and negatively compared to non-religious individuals. However, the extent to which individuals felt positively about the events interacted with religious beliefs to predict reported false memories. Religious individuals were more likely to remember events if they felt positively about them whereas atheists were more likely to remember events if they felt negatively about them.

Keywords: religiosity; false memory; beliefs

False memory refers to mistaken recollection of events or aspects of events that never occurred as if they actually happened. Elizabeth Loftus and colleagues have conducted numerous studies on this phenomenon. In one of the first studies to investigate false memory, Loftus and Palmer (1974) showed participants videos of a car accident and then asked them questions regarding the accident. More participants falsely remembered seeing broken glass when the verb "smashed" was used versus when the verb "hit" was used. This seminal study was one of the first to indicate how malleable memory can be, and how easily it is influenced by leading questions – a finding that has contributed to stricter guidelines for witness interrogations (Loftus, 1993).

Further research has identified many possible influences on false memory, such as plausibility of events (Pezdek, Finger, & Hodge, 1997) and prior expectations (Macrae, Scholoerscheidt, Bodenhausen, & Milne, 2002; Vendetti et al., 2014). In addition, group identification can influence pre-existing beliefs and bias memory (Van Bavel and Cunningham, 2012). Beliefs regarding religion and politics can contribute to group belonging and may also lead one to recall events in line with those beliefs.

Frenda et al. (2013) investigated the relationship between political beliefs attitudes toward events, and memory by presenting Democrats and Republicans with images depicting political events. Participants were then asked whether they remembered the event and how they felt about the event. They were not asked how strongly they remembered the event. Not only were participants more likely to recall a fabricated event consistent with their political views, but they also recalled feeling emotions consistent with their attitudes toward the party in question. For example, Democrats were more likely to falsely remember Bush vacationing during Hurricane Katrina and to report feeling more outraged about the event. Republicans, on the other hand were more likely to falsely remember events negatively depicting Democrats, such as Obama shaking hands with the Iranian president. On the other hand, fabricated events depicting Hillary Clinton airing an attack ad linking Obama to Rev. Wright or Joe Lieberman voting guilty during President Clinton's impeachment trial were not remembered because there was low congruence between their political attitudes and how they felt about the event. These findings support previous research suggesting that real memories are influenced by one's positive or negative attitudes toward an event, in addition to being influenced by one's beliefs about events. However, such research reveals mixed results and a recent study showed that people may remember positive and negative life events equally well (Waters et al., 2013). In addition, emotions toward events may promote memory for features of events, but not overall accuracy of the memories (Earles et al., 2016). Less is known about the role of affect on false memories.

In the present study, we were interested in examining the effects of religious affiliation and strength of religious beliefs on false memory for plausible news events some of which had actually occurred, and some of which had not. Although we expected participants to be more likely to falsely remember events consistent with their religious beliefs (or lack thereof), it is possible that, for example, some events portraying religion may be recalled by

nonreligious people if the religious event was similar to their experiences of religion (Pezdek, Finger, & Hodge, 1997).

It was expected that Christians, compared to atheists and agnostics, would be more likely to falsely remember an event that positively described a Christian or an event that negatively depicted nonbelievers, such as atheists. Similarly, nonbelievers were expected to falsely remember events that were consistent with their non-religious views, such as events depicting religion as causing harm rather than doing good. Because the role of religious beliefs and valence in predicting false memories had not previously been examined, our study was designed to investigate whether effects existed, not to reveal underlying mechanisms for such possible associations.

Method

Participants

599 participants (82% female, 84% religious) who received course credit for their participation were recruited from the psychology research pool at a Midwestern University in the U.S. and directed to a secure website (surveymonkey.com) where they completed the study online.

Materials

To assess the strength of participants' religious beliefs, the religious fundamentalism (Altemeyer & Hunsberger, 2004) and true religious belief scales (Zeigler-Hill & Shackelford, unpublished) were administered

The Religious Fundamentalism Scale is a 20 item scale with a Cronbach's alpha typically ranging from .93 to .95. It measures the extent to which one views religion as a single, absolute truth. Like the dogmatism scale, it can be related to authoritarianism (Altemeyer & Hunsberger, 1992). The scale was revised in 2004. An example of an item is, "God has given mankind a complete, unfailing guide to happiness and salvation, which must be totally followed." The items are answered based on a 9-point Likert scale ranging from *very strongly disagree* to *very strongly agree* (α = 95).

The True Religious Beliefs scale is a 59 question compilation of items from several scales including six items from the Shortened Post-Critical Belief Scale (Duriez, Soenens, & Hutsebaut, 2004), ten items from the Beliefs and Values Scale (King et al., 2005), two items from the God Image Inventory (Lawrence, 1997), five items from The God Delusion (Dawkins, 2006, pg207-208), four items from the Literal, Anti-Literal, and Mythological Scales (Hunt, 1972), and additional measures as developed by Zeigler-Hill and Shackelford (Unpublished data). In the current study, we obtained a Cronbach's alpha of .98 for the complete scale.

To assess memory for false events, participants were given a memory task in which they were presented with eight images corresponding to news stories that depicted four actual events (one in which religion was depicted positively and one in which religion was depicted negatively, one in which atheism was depicted positively, one in which atheism was depicted negatively) and four that depicted fabricated events from the same categories.

Procedure

Participants completed several online questionnaires via surveymonkey.com. Participants provided information about their sex, GPA, political, and religious affiliations. Participants who reported being affiliated with Christianity or other religions, or religious without any affiliation were coded as religious. Those who reported being atheist or agnostic were coded as non-religious. They were then administered the scales described above along with two additional scales that were not included in the analyses. After the questionnaires, the participants were presented with the memory task. Images of events were presented in the same pseudorandom order for all participants. For each image, participants responded as to how strongly they recalled the event on a four point Likert scale (not at all to very well) and how they felt about the event on a five point Likert scale (very negatively to very positively) with lack of memory at the middle point (valence).

Results

Data was examined for normality and the memory scores were found to have a positive skew. Memory scores were therefore transformed using a logarithmic transformation in order to correct for this positive skewness. Memory and valence scores were then standardized before being entered into analyses. We conducted bivariate correlations between the predictor variables of religious category, memory for true events, and valence for the false events, and the outcome variable of memory for false events. Correlation results revealed a high positive correlation between false memory for the event negatively depicting religion and valence of this event (r= -.59, p< .001). Religious category was highly correlated with our religiosity measure (described below), r=.62, p< .001.

We then performed hierarchical multiple regression analyses where we regressed each log-transformed and standardized false memory score on to strength of religious beliefs and valence (measure of positive regard) for the corresponding false event. Consistent with Frenda et al. (2013), the transformed memory score for the corresponding true event was entered into the regression at the first step as a control. In addition, religious category was added into this step to see if strength of religious beliefs accounted for any variability above and beyond what was accounted for by religious affiliation.

Correlation analyses between the True Religious Beliefs Scale and Religious Fundamentalism scores suggested a large correlation between the two scores (r = .81). To eliminate the effect of multicollinearity, and because the two scales reduced to a single component using PCA, the scores were standardized and averaged to create a composite religiosity score. This score was entered into the second step of the regression model and allowed to interact with valence in the third step. Results of the regression analyses are displayed in Table 1.

Memory for false positive religious events

Memory for the true event (β=.26, t= 6.59, p < .001) and religious category (β= .08, t= 2.11, p= .04) were positively associated with false memory for the event positively depicting religion. After controlling for religious category and memory for the true event, valence of the false event was positively associated with false memory (β= .46, t= 13.13, p< .001)

A significant interaction was found between strength of religious beliefs and valence of the false event (β =-.08, t= -2.24, p= .03). Predicted values for this interaction are presented in Figure 1. Simple slopes tests were conducted for individuals who were high and low in religiosity. For both those lower in religiosity (β= .55, t= 10.53, p < .001), and those higher in religiosity (β= .40, t= 8.93, p < .001), valence was positively associated with false memory. Participants were more likely to falsely remember events positively depicting religion if they felt positively about them.

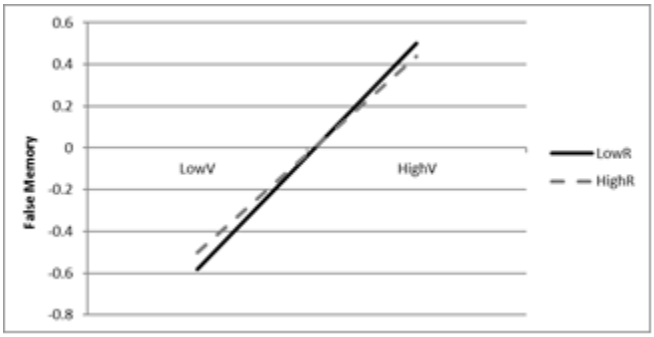


Figure 1.

Predicted values for false memory for positive religious events are presented illustrating the interaction of religiosity and valence at values that are one standard deviation above and below their respective means.

Memory for false, negative religious events

Memory for the true event (β= .21, t= 5.12, p< .001) and religious category (β= .12, t= 3.00, p= .003) were positively associated with memory for the false event. After controlling for religious category and memory for the true event, valence for the false event (β = .28, t = 7.39, p > .001) was positively associated with memory for the false event.

Religiosity interacted with valence to predict false memory (β= .12, t= 2.96, p= .003). Predicted values for this interaction are presented in Figure 2. To further understand this interaction, simple slopes tests were conducted for individuals high and low in religiosity. For both participants high (β= .37, t= 7.75, p < .001) and low in religiosity (β= .16, t= 2.71, p= .007), valence was positively associated with false memory. Participants were more likely to falsely remember events negatively depicting religion if they felt positively about them.

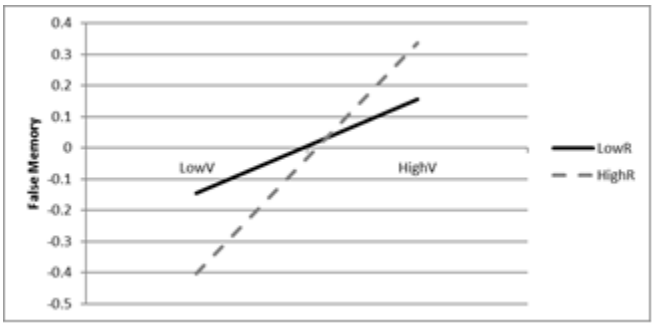


Figure 2.

Predicted values for false memory for negative religious events are presented illustrating the interaction of religiosity and valence at values that are one standard deviation above and below their respective means.

Memory for false, positive atheist events

Memory for the true event (β = .33, t = 8.27, p > .001) was positively associated with memory for the false event. After controlling for religious category and memory for the true event, valence for the false event was negatively associated with false memory for the negative religious event (β = -.19, t = -4.69, p > .001).

Religiosity interacted with valence for the false event to predict memory for the false event (β= -.35, t= -8.89, p < .001). Predicted values for this interaction are presented in Figure 3. To further understand this interaction, simple slopes tests were conducted for individuals high and low in religiosity. For participants high in religiosity, valence was negatively associated with false memory (β= -.55, t= - 9.92, p< .001), but valence did not significantly impact false memory for participants low in religiosity (β= -.04, t = -.87, p= .39). Only participants high in religiosity were more likely to falsely remember events positively depicting atheism if they felt negatively about them.

Table 1: Regression of False Memory Scores on to Memory for the true Event, Religious Category, and Strength of Religious Beliefs and Valence toward the false event.

	Memory False Positive Religious			Memory False Negative Religious			Memory False Positive Atheist			Memory False Negative Atheist		
	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2	β	R^2	ΔR^2	β
Step 1	.08***	.08***		.06***	.06***		.11**	.11***		.07***	.07***	
Memory for real event			.26***			.21***			.33***			.22***
Religious Category			.08*			.12**			-.03			.14***
Step 2	.29***	.21***		.14***	.08***		.14***	.03***		.38***	.31***	
Memory for real event			.22***			.21***			.34***			.17***
Religious Category			.05			.11*			-.07			.08*
Strength of religious beliefs			.04			-.03			.003			-.03
Valence for false event			.46***			.29***			-.19***			-.57***
Step 3	.30*	.006*		.15**	.01**		.25***	.10***		.38	.001	
Memory for real event			.22***			.21***			.32***			.16***
Religious Category			.06			.12*			-.02			.07*
Strength of religious beliefs			.04			-.02			-.01			-.03
Valence for false event			.47***			.26***			-.29***			-.57***
Religious Category * Valence			-.08*			.12**			-.35***			.02

*$p < .05$; **$p < .01$; ***$p < .001$.

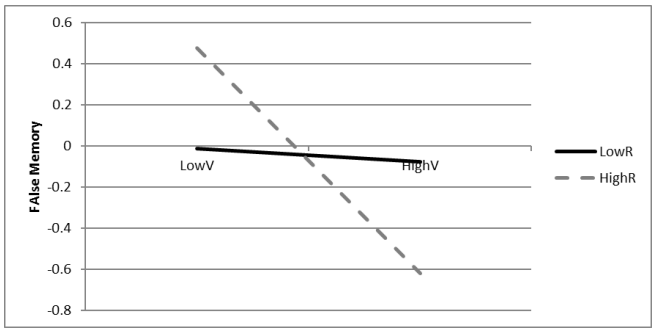


Figure 3.

Predicted values for false memory for negative religious events are presented illustrating the interaction of religiosity and valence at values that are one standard deviation above and below their respective means.

Memory for false, negative atheist events

Memory for the true event (β = .23, t = 5.72, p > .001) and religious category (β= .14, t= 3.58, p< .001) were positively associated with memory for the false event. After controlling for religious category and memory for the true event, valence for the false event was negatively associated with false memory (β = -.57, t = -17.04, p > .001). That is, people who had stronger negative feelings about the event and those who were more religious were more likely to falsely remember an event that portrayed atheists in a negative light. The interaction between strength of religious beliefs and valence for the false event was not significant (β = .02, t= .69, p = .49).

Discussion

The results of our regression analyses support the notion that religious beliefs interact with regard for the events to impact false memory formation. Religious participants and non-religious participants falsely remembered the fabricated events consistent with their religious beliefs, but differently as a function of how they felt about various events. Religious participants were more likely to remember religious events if they felt positively about them and atheist events if they felt negatively about them. Of particular interest is the finding that both religious and nonreligious participants were more likely to remember the event negatively depicting religion if they felt positively about it.

The latter finding could be explained by features of the events themselves. Because only a single event was used for each category of events, it is possible that idiosyncratic features of the individual events had unexpected influence on participants' valence ratings and memories. For instance, the image of the Pope (Figure 5), while telling a negative story, may have been more salient than the caption itself. The event depicts a cover-up of child sexual abuse by Pope Francis (when this was in-fact

done by a previous pope). We intend to address this possibility in a future study.



Figure 5.

The false negative religious event depicted Pope Francis covering up child sexual abuse.

Another possible explanation for the finding is that negative events are in themselves more salient than positive events. Pratto and John (1991) suggested that people automatically attend to negative more than positive stimuli. Other studies, however, also point to the impact of emotion. Laney, Campbell, Heuer, and Reisberg (2004) found that participants remembered emotionally-arousing information better than more neutral information. Kensigner and Corkin (2003) reached a similar conclusion in finding that memory was greater for events that were both negative and emotional. It is possible that both religious individuals, and atheists, who identify as being non-religious (which is defined more by the absence than presence of a particular belief system), feel more strongly about events aligned with the presence rather than the absence of religion. The results of the regression analysis are consistent with this literature given that there was a large effect of valence on memory for the false negative atheist event, which depicted atheists burning a church. This finding may result from participants both having strong negative feelings for the events and strong memories for the events. This event, as well as the true event involving the Westboro Baptist Church, can be considered to be both negative and emotionally-arousing.

Whereas negative events may be salient, it is also possible that religious people tend to evaluate events more positively. Vishkin et al. (2015) found that religious people were more likely to effectively reappraise negative information in more positive light. This finding could explain why religious people generally remembered events depicting religion if they felt positively about them. Perhaps religious people were able to reappraise the negative religious events as more positive.

The finding that false memory is moderated by valence is consistent with Frenda et al. (2013). For example, Republicans in this study were more likely to falsely remember events positively depicting Republicans if they felt positively about the event. Conversely, if such participants felt negatively about an event negatively depicting democrats, they were more likely to remember it. This finding, as well as the findings from our study are consistent with the positive correlation between conservatism and religious attitudes found by Pennycook et al. (2011).

The current study is the first study to directly address the relationship between religiosity and false memory. Further research can focus on investigating the interaction between religious affiliation and attitudes towards the event. In addition, future research could investigate how different religious viewpoints may influence false memory (i.e. Christian events versus Jewish or Muslim events).

Our study is limited by the use of a college sample. Data from college students may not necessarily be generalizable to the overall population particularly with regards to beliefs and memory for public events. In addition, our sample was predominately Christian and female. Because our population included a disproportionate amount of religious and nonreligious participants, the small amount of nonreligious participants may not be truly representative of the atheist population. We are currently addressing this limitation with a second study using a community sample.

Our results suggest that religiosity may influence false memory for religious events. This further adds to the literature suggesting that memory may be biased by prior expectations and attitudes. In addition, our research provides further evidence that prior beliefs may interact with attitudes towards an event to predict false memory. By expanding the existing literature, we have shown that memory can be influence by more ways in which we imagined.

References

Altemeyer, B. & Hunsberger, B. (1992). Authoritarianism, religious fundamentalism, quest, and prejudice. *The International Journal for the Psychology of Religion, 2*, 113-133.

Cooke, N. M., Durso, F. T., & Schvaneveldt, R. W. (1986). Recall and measures of memory organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 538-549.

Duriez, B., Soenens, B., & Hutsebaut, D. (2005). Introducing the shortened post-critical belief scale. *Personality and Individual Differences, 38*, 851-857.

Earles, J. L., Kersten, A. W., Vernon, L. L., & Starkings, R. (2016). Memory for positive, negative and neutral

events in younger and older adults: Does emotion influence binding in event memory? *Cognition and Emotion, 30*(2), 378-388.

Frenda, S. J., Knowles, E. J., Saletan, W., Loftus, E. (2013). False memories of fabricated political events. *Journal of Experimental Social Psychology, 49*, 280-286.

Holyoak, K. J. & Simon, D. (1999). Bidirectional Reasoning in Decision Making by Constraint Satisfaction. *Journal of Experimental Psychology: General, 12*, 3-31.

Hunt, R. A. (1972). Mythological- Symbolic Religious Commitment: The LAM Scales. *Journal for the Scientific Study of Religion, 11*, 42-52.

Kensinger, E.A. & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory and Cognition, 31*, 1169- 1180.

King, M., Jones, L., Barnes, K., Low, J., Walker, C., Wilkinson, S., Mason, C., Sutherland, J., & Tookman, A. (2006). Measuring spiritual belief: development and standardization of a beliefs and values scale. Psychological Medicine, 36, 417-425.

Laney, C., Campbell, H. V., Heuer, F., & Reisberg, D. (2004). Memory for thematically arousing events. *Memory and Cognition, 32*, 1149-1159.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.

Macrae, C. N., Scholoerscheidt, A. M., Bodenhausen, G. V., & Milne, A. B. (2002). Creating memory illusions: Expectancy- based processing and the generation of false memories. *Memory, 10*, 63-80.

Pezdek, K., Finger, K., & Hodge, D. (1997). Planting false childhood memories: the role of event plausibility. *Psychological Science, 8*(6), 437-441.

Pratto, F. & John, O.P (1991) Automatic vigilance: The attention- grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380-391.

Rokeach, M. (1956). Political and religious dogmatism: An alternative to the authoritarian personality. *Psychological Monographs: General and Applied, 70*, 1-43

Van Bavel, J. J. & Cunningham, W. A. (2012). A social identity approach to person memory: group membership, collective identification, and social role shape attention and memory. *Personality and Social Psychology Bulletin, 38*, 1566-1578.

Vendetti, M. S., Wu, A., Rowshanshad, E., Knowton, B. J., & Holyoak, K. J. (2014). When reasoning modifies memory: Schematic assimilation triggered by analogical mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1172-1180.

Vishkin, A., Bigman, Y. E., Porat, R., Solak, N., Halperin, E., & Tamir, M. (2015). God rest our hearts: Religiosity and cognitive reappraisal. *Emotion, 16*, 252-262.

Waters, T. E. A., Bohanek, J. G., Marin, K., & Fivush, R. (2013). Null's the word: A comparison of memory quality for intensely negative and positive events.*Memory, 21*(6), 633-645.

Webster, D. M. & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Personality Processes and Individual Differences, 67*, 1049-1062.

Surprising blindness to conversational incoherence in both instant messaging and face-to-face speech

Gareth Roberts[1,2] **(gareth.roberts@ling.upenn.edu)**

Benjamin Langstein[2,3] **(langstein@haskins.yale.edu)**

Bruno Galantucci[2,3] **(bruno.galantucci@yu.edu)**

[1]Department of Linguistics, University of Pennsylvania
255 South 36th Street, Philadelphia, PA 19104 USA

[2]Department of Psychology, Yeshiva University
2495 Amsterdam Avenue, New York, NY 10033 USA

[3]Haskins Laboratories, 300 George Street
New Haven, CT 06511 USA

Abstract

Language is widely assumed to be a well designed tool for reliably communicating propositional information between people. This suggests that its users should be sensitive to failures of communication, such as utterances that are blatantly incoherent with respect to an ongoing conversation. We present experimental work suggesting that, in fact, people are surprisingly tolerant of conversational incoherence. In two previous studies, participants engaged in instant-messaging conversations that were either repeatedly crossed with other conversations or had lines inserted into them that deliberately contradicted available information. In both cases, a substantial proportion of participants failed to notice. In a new study, confederates inserted unexpected, nonsensical lines into face-to-face conversations. The majority of participants failed to notice. We argue these findings suggest that we should be wary of modeling spontaneous communication in terms of faithful information transmission, or language as a well designed tool for that purpose.

Keywords: communication; miscommunication; language evolution; change blindness

Introduction

A widely held view is that language is primarily a tool for reliably communicating propositional information between individuals, and that it is good at its job. This assumption, which is closely connected to what has been termed the *code model* of human communication (Blackburn, 1999), is not only widespread outside academia, but it also pervades much scholarly discourse. It is central, for instance, to models of communication that follow the work of Shannon and Weaver (1949); it has long been a guiding assumption of prominent adaptive accounts of language evolution (Pinker & Bloom, 1990); and research has been conducted demonstrating the communicative function of apparently maladaptive features of language (e.g., Piantadosi, Tily, & Gibson, 2012). This is not unreasonable, and it could scarcely be denied that a great deal of linguistic interaction involves communicating propositions. But that is not the same as saying that this is the primary purpose of linguistic interaction, or that language is a system finely tuned by evolution to fulfil that purpose. If that is so, then we should expect language users to be sensitive to cases of communicative failure, and for blatant cases of incoherence to be readily identified and rapidly corrected.

In fact, however, there are reasons to be wary of assuming that communication is all that reliable. Humans are poorer than might be expected at faithful information transmission in a variety of non-linguistic perceptual tasks (Rensink, O'Regan, & Clark, 1997; Simons & Levin, 1998; Simons & Chabris, 1999). With respect to language, there is evidence that comprehenders form representations that are merely "good enough" for the task they need to perform (Ferreira, Bailey, & Ferraro, 2002), while work on the "Moses illusion" has shown that comprehenders fail to detect inconsistencies in input where inconsistent material is semantically related to expected material (Park & Reder, n.d.). There is reason to suspect that even more jarring inconsistencies might be missed surprisingly often. One reason for this is that it is unclear whether reliable communication is in fact the primary function of a great deal of linguistic interaction. It has long been pointed that human conversations very often serve what have been termed *phatic* goals (Malinowski, 1923) – that is, they are aimed at creating and supporting social relationships. This observation has several consequences for the reliability of communication. First, it suggests that failures of communication, such as errors of fact or utterances that are incoherent with respect to surrounding discourse, may not be noticed by interlocutors, as speakers may focus on social coherence at the expense of communicational coherence. Second, even if failures of communication are noticed, they may not be queried or remedied – interrupting, correcting, or questioning one's interlocutor detracts from maintaining a pleasant social interaction. Third, whether or not observed instances of miscommunication are resolved, they are likely to be quickly forgotten, because they may not be especially important to the social goals of the interlocutors. In fact, the situation may be worse than has been suggested so far. It might be countered, for example, that even if phatic interactions are

very common, speakers still generally want to convey information, and the language has various features to help them to do this. However, some of these very features may in fact make it easier for speakers to overlook cases of communication failure. For example, the ambiguity that Piantadosi et al. (2012) convincingly argue supports efficient communication also allows cases of apparent incoherence in one's interlocutor to be dismissed. In Gricean models of communication, moreover, a meaning is reconstructed via an inferential process that assumes relevance on the part of the speaker (Grice, 1975; Wilson & Sperber, 2004). While this process is beneficial to successful communication, it is also a means by which incoherent contributions to discourse can be explained away.

The above suggests that humans may be much less sensitive to incoherence in conversation than is typically assumed, but most work on miscommunication to date has relied on qualitative analysis of naturally occurring discourse (e.g., Tzanne, 2000), with very little work devoted to investigating it in controlled environments (see Mills, 2012, for a rare exception). To remedy this, we conducted a series of studies investigating people's sensitivity to conversational incoherence. Two of these studies, which have already been published (Galantucci & Roberts, 2014; Roberts, Langstein, & Galantucci, 2016), involved spontaneous instant-messaging conversations. These two studies will be described below in the Section *Previous studies*. We also report a new third study, which involved face-to-face conversation with a confederate who inserted a nonsensical line. In all three cases participants were asked whether they had noticed the incoherence (Galantucci, Roberts, & Langstein, submitted). Since the three studies are best understood in comparison with one another, we will present the previously published work in some detail.

Previous work: Insensitivity to incoherence in instant-messaging

In each of the two instant-messaging studies, 40 native native-English-speaking students, with no deficits in color vision or communicative ability, participated for $20 each. In the first study (henceforth Study A), participants' conversations were crossed with conversations by other naive participants. Study B was similar, but involved the automatic insertion of lines that were guaranteed to not fit well with information already provided to participants.

Study A: Crossed instant-messaging conversations (Galantucci & Roberts, 2014)

Procedure Participants took part in pairs. Care was taken that the members of a pair did not already know each other and met only at the start of the study. On arriving at the lab, members of a pair were seated in separate rooms and received the same scripted instructions. Each member of a pair sat in front of a computer displaying a cartoon of five famous people and an instant-messaging window (Figure 1). As in widely used instant-messaging programs, the messaging window consisted of two parts. At the bottom was a space where a participant could type messages and relay them to their partner. The rest of the window consisted of a space in which messages would appear. Messages the participant had sent appeared in black and were preceded by "You:"; received messages appeared in red and were preceded by "Partner:". Both members of the pair saw the same exact cartoon except that it was colored differently for each of them. (This was explained to both participants.)

The task given to participants was manipulated. Ten pairs of participants were put in the *Narrowly focused* condition and were asked to identify the color differences by using the instant-messaging software to chat for 15 minutes. The other ten pairs were put in the *Broadly focused* condition and were asked to discuss which of the five famous people they would most and least like to spend a day with.

Although the above account focuses on one pair only, each trial in fact involved two pairs, Pair A and Pair B (both pairs were always in the same condition). The two pairs performed the same task simultaneously, but with different cartoons. Pair A's cartoon contained five different famous people from Pair B's cartoon, set against a very different background (Figure 1). The software ensured that the task began at the same time for both pairs, but no participant was aware of the existence of the other pair. Over the course of each conversation there were four 30-second *crossings* during which each member of Pair A was re-partnered with a random member of Pair B. Participants were not warned before the study that these crossings would occur (indeed, no participant was even aware of the existence of the other pair), and there were no markers to indicate that they were occurring. All messages received, whoever had sent them, continued to appear in red preceded by "Partner:" (Figure 1). Crossings occurred at random points but began at least three minutes into the conversation.

After 15 minutes had elapsed the chat window closed and a new window presented the following questions one by one:

1. How did you find the conversation today?
2. Did the conversation go smoothly?
3. Did you ever feel like you were having trouble communicating with your partner?
4. Did you notice anything unusual in the conversation?
5. Participants in this study are put in one of two groups. 50% of participants are put in the No-Crossing Group. If we put you in the No-Crossing Group then all the messages you received came from your partner. The other 50% of participants are put in the Crossing Group. If we put you in the Crossing Group then some of the messages you received came from a different participant who intended them for someone else and did not know that they would come to you. Which group do you think you were in? Note: If you are correct, you will win $3!

Participants who answered "no" to Question 2 or "yes" to Question 3 or 4 were asked to explain their answers before the next question appeared.

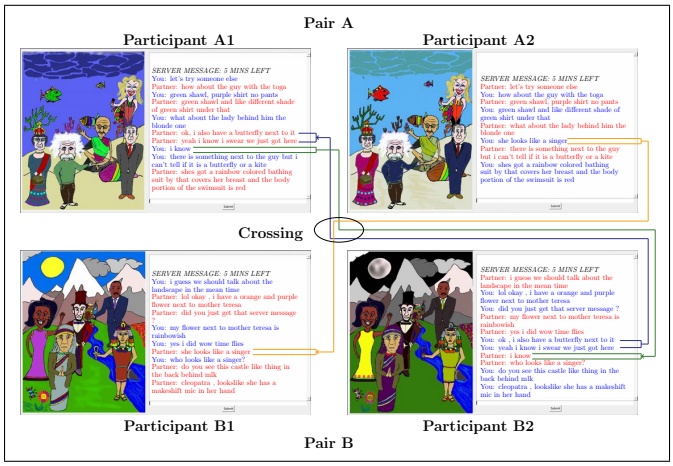


Figure 1: Example conversations from Study 1 (Narrowly focused condition) with crossed messages highlighted.

Results

Questions 2–4 were used to exclude participants whose answers to those questions were inconsistent with their answer to Question 5. In the Narrowly focused condition there was one *Inconsistent detector* (i.e., a participant who guessed "Crossing Group" in answer to Question 5, but otherwise claimed to have noticed nothing out of the ordinary) and no *Inconsistent Non-detectors* (i.e., participants who claimed to have noticed something odd, but answered "No-Crossing Group on Question 5); in the Broadly focused condition there were four Inconsistent detectors and one Inconsistent Non-detector. These patterns suggests the existence of a bias toward suspecting unusual circumstances. To test this, we replicated the study with 20 new participants and no crossings (i.e., each participant chatted with the same partner for the entire session). Consistent with the presence of such a bias, six participants (30%) incorrectly answered that they were in the Crossing Group (four in the Narrowly focused condition and two in the Broadly focused condition).

Crossing detection rates In the Narrowly focused condition, 11 out of 19 participants answered that they were in the Crossing Group, leading to a detection rate of 57.9%. This rate is not significantly different from chance: 95% CI [33.5%, 79.7%] (Figure 3; all CIs presented in this paper were computed using the Clopper–Pearson exact method).

In the Broadly focused condition 10 out of 15 participants answered that they were in the Crossing Group, leading to a detection rate of 66.7%. Again this rate is not significantly different from chance (95% CI [38.4%, 88.2%]; Figure 3) and is rather far from the expectations of the sample reported in the introduction.

Crossing detection rates for blatantly crossed conversations A possible explanation for the low detection rates is that crossing the conversations did not in fact lead to blatant violations of coherence. To check whether this was the case we analyzed the transcripts of the conversations in two ways.

First, we read the transcripts, looking for out-of-the blue references to famous people or salient items not present in the cartoon and which had not been mentioned earlier in the conversation. Such references should have provided blatant indications that the conversation had been crossed with a different one.

The second analysis relied on four naive judges, who were given ten transcripts each for each condition: five from the No-Crossing Group (i.e., from the control study reported above) and five from the Crossing Group. Transcripts were distributed in such a way that (a) every judge received one transcript from each set, and (b) each transcript was given to one judge. We explained the basics of the study to the judges and asked them to decide which condition each transcript came from. Judges were paid $10 an hour or given course credit (in either case the reward was doubled if discrimination was perfect). There was no time-limit for completing the task.

This led us to exclude seven participants who were likely not exposed to blatant indications of crossings. For the Narrowly focused condition, this produced a crossing detection rate of 67%. For the Broadly focused condition this produced a crossing detection rate of 72.7%. In neither case is the rate significantly different from chance (95% CI [38.4%, 88.2%] and [39%, 94%] respectively).

Although the detection rates in the Broadly focused condition were higher than in the Narrowly focused condition, the results are consistent: People are surprisingly insensitive to conversational incoherence.

Study B: Inserted lines in instant-messaging conversations (Roberts et al., 2016)

Study 1 provided intriguing evidence that people are less sensitive to conversational incoherence that might have been expected. However, there are reasons to be wary of the findings. The procedure used led to incoherencies that were inherently uneven, meaning there was no way to ensure that conversations would include noticeable interruptions. This problem was somewhat diminished by including three crossings in every conversation, and (as described above) the resulting conversations were carefully checked in two ways to ensure that participants had in fact been exposed to clear incoherence. Nevertheless, an obvious step was to replicate the study with more controlled interruptions. Study 2 (reported in Roberts et al., 2016) did precisely this by automatically inserting lines that were guaranteed not to be consistent with ongoing conversations. Furthermore, it compared two different kinds of interruption: Inserted lines that conflicted with shared task-relevant information and lines that conflicted with the interlocutor's apparent gender.

Procedure The same procedure was employed for Study 2 as for the Broadly focused condition of Study 1, except in two respects. First, instead of the cartoon used in Study 1, an image composed of six photos of famous historical people (Figure 2) was used to stimulate conversation. Second, there was no crossing of conversations. Instead, two of the messages sent during the conversation (one from each participant) were replaced by fake messages (always at least 270 seconds after the beginning of the conversation and within 120 seconds of the end). All inserted messages all had the same phrasing:

> of these six [FAMOUS PERSON]s kind of an icon for [WOMEN/MEN] like me.

There were two between-subjects conditions (with 10 pairs each). In the *Gender-swap* condition the famous person named was chosen from among Hillary, Einstein, Marilyn, and MLK, all of whom were depicted in the photograph (Figure 2), but the phrase "women like me" or "men like me" was chosen to conflict with the apparent sender's real gender. For example, a line that supposedly came from a female participant might read "of these six Einsteins kind of an icon for men like me". In the *Character-swap* condition the reverse was true: The gender was as expected, but the famous person referred to was not in the picture, but would be chosen from among the names Oprah, Madonna, Gandhi, or Mandela (e.g., a line supposedly from a female participant might read "of these six Oprahs kind of an icon for women like me"). Across both conditions a sentence ending "for women like me" would always refer to a female celebrity and a sentence ending "for men like me" would always refer to a male one. The same two famous people were never used for both inserted lines in the same conversation.[1]

After 15 minutes had elapsed, as in Study 1, a questionnaire was presented. This differed only with respect to Question 5, which referred to "the Normal-Conversation Group" and the "Inserted-Message Group" instead of "No-Crossing Group" and "Crossing Group". Having finished answering Question 5, every participant was told that there had in fact been an inserted message and was shown a transcript of the conversation they had just been engaged in, which included all the messages the participant had actually sent or received. (The inserted message that the participant had received was included; the inserted message that the participant's partner had received was not.) Every message in the transcript was numbered and the participant was asked to scroll through and select the message they thought had been inserted. The distance in number of lines between the participant's guess and the actual inserted message was measured (henceforth *Transcript distance*). The transcripts were also checked for lines from partners which were not themselves inserted, but which referred to inserted lines (e.g., "I don't see Oprah anywhere."). If such a line occurred, the Transcript distance between it and the participant's guess was also measured, and the lower value was used for analysis.



Figure 2: Image used to stimulate conversation in Study 2. All photographs were taken from Wikimedia Commons (https://commons.wikimedia.org/) and are either public domain or available under a creative commons license.

Results

In the interest of being conservative, Inconsistent detectors were excluded only if they also failed to identify the line in the transcript. In the Gender-swap condition there were seven Inconsistent detectors and no Inconsistent Non-detectors. In the Character-swap condition there was one Inconsistent Non-detector and four Inconsistent detectors. These participants were excluded from further analysis.

Detection rates In the Gender-swap condition 66.66% of participants (95% CI = [48.17% 82.04%]) correctly guessed that they were in the Inserted-Message Group. In the Character-swap condition 57.14% of remaining participants (95% CI = [39.35% 73.68%]) correctly guessed that they were in the Inserted-Message Group. In neither case was the detection rate significantly better than chance (Figure 3).

Identification of inserted message The transcript distance for one participant in the Character-swap condition (who guessed that she was in the Normal-Conversation Group) could not be computed because, due to a misunderstanding, she did not perform the line identification task. For each of the remaining participants, if the Transcript distance was greater than three lines this was taken as an indication that the participant had failed to identify the inserted message.

In the Gender-swap condition there were 14 such participants (35%), seven of whom responded that they were in the Normal-Conversation group (63.64%) and seven of whom thought they were in the Inserted-message Group (24.14%),

[1] It should be noted that, on top of the incoherence generated by the informational conflict, the message might introduce other forms of incoherence, such as the absence of an answer to a direct question or an abrupt change of topic. Furthermore, the inserted line might generate messages from the receiver (e.g., "Why bring up Oprah?") which would not make sense to the other participant, who had no knowledge of the inserted message.

suggesting that their answer to Question 5 might have been simply a guess. The difference in line identification between these two groups is significant ($\chi^2 = 5.47, p < .05$).

In the Character-swap condition 11 participants (28.21%) failed to identify the inserted message, seven of whom responded that they were in the Normal-conversation Group (46.67%) and four of whom responded that they were in the Inserted-Message Group (16.67%). The difference in line identification between the two groups is significant ($\chi^2 = 4.39, p < .05$).

NEW STUDY: Inserted lines in face-to-face conversation

Study B found similar results to Study A, with rather more controlled interruptions. However, both studies involved instant messaging rather than face-to-face conversations. While it is important to stress that instant messaging is a very natural communication medium for a great number of people, especially college students like our participants, it remains possible that our results had something to do with the nature of the interaction. We therefore conducted Study 3, in which participants chatted face-to-face with confederates (Galantucci et al., submitted).

Methods

Participants 30 native English-speaking students (16 female, 14 male; mean age 24.73, SD = 6.43), with no deficits in communicative ability, participated for $20 each.

Procedure Each participant engaged in face-to-face conversation with a confederate. In order to elicit a spontaneous and lively conversation, they were asked to discuss five thought-provoking questions such as "Would you rather be color blind, or not be able to understand puns?" or "Would you rather live the rest of your life on a tree or in a cave?". The five questions were written on a whiteboard in front of the individuals, who were told that they had about 15 minutes to order them from funniest to least funny and from hardest to answer to easiest to answer.

At some point during the conversation the confederate would utter the nonsensical sentence "Colorless green ideas sleep furiously". The confederate was instructed to utter the sentence clearly and audibly, at least eight minutes into the conversation, and at a point where the participant was silent. One minute after the sentence was uttered, the conversation was interrupted by an experimenter and the participant and the confederate were escorted to two separate rooms for a follow-up questionnaire. When the 30 participants were asked at the end of the study whether they ever suspected that the person they conversed with was a confederate, only five of them answered positively, and none of them mentioned the nonsensical sentence as the reason for the suspicion.

The questionnaire was as in Studies 1 and 2 apart from Question 5, which referred to "the Control Group" and the "Nonsensical Sentence Group" (but was otherwise identical to Question 5 in Studies 1 and 2). Right after answering Question 5, all participants were asked to estimate how confident they were in their answer, using a scale from 1 to 10 (where 1 corresponded to "Not certain at all" and 10 to "Completely certain"). Then, the participants who thought they were in the Nonsensical Sentence Group were asked to identify the nonsensical sentence uttered by the confederate in a list which included 20 other nonsensical sentences.

Results

Responses to Question 5 In answer to Question 5, 10 participants (33.33%) said they thought they were in the Nonsensical Sentence Group. This detection rate is not significantly different from chance (95% CI [17.3% 52.8%]).

There were no Inconsistent detectors in Study 3. Furthermore, the mean confidence rating for the 20 participants who said they thought they were in the Control Group was 8.8 (95% CI [7.81 9.79]), while the mean confidence rating for the 10 participants who said they thought they were in the Nonsensical Sentence Group was 7.7 (95% CI [6.34 9.06]; Figure 3).

Recognition of the nonsensical sentence Of the 10 participants who said they thought they were in the Nonsensical Sentence Group, only one correctly recognized the sentence "Colorless green ideas sleep furiously" in a list that included 20 other nonsensical sentences. In other words, 90% of the participants who detected the nonsensical sentence forgot it within a few minutes from hearing it. This finding is consistent with the well established fact that recalling linguistic materials that lack a coherent overall meaning is a substantial challenge (Marks & Miller, 1964).

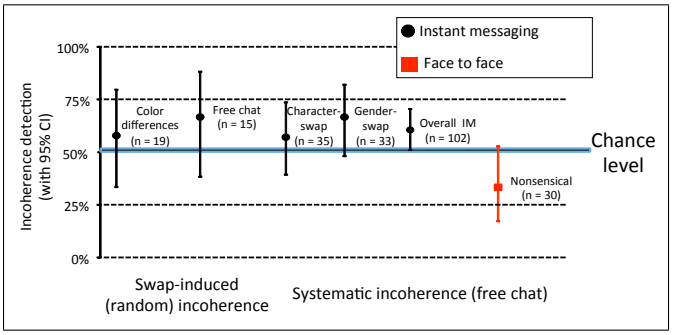


Figure 3: Detection rates across all three studies

Discussion

In each of the three studies described here participants were surprisingly blind to cases of blatant incoherence in spontaneous conversation (Figure 3). Even in the condition with the highest detection rate (the Gender-swap condition of Study 2), more than 25% of participants failed to notice the incoherence. (This rises to a third once Inconsistent detectors are excluded.) The results from Study 3 are particularly striking: In face-to-face conversation a full two thirds of participants failed to notice the introduction of a meaningless line

that could have no relevance to the rest of the conversation, and only one of the participants who did notice the line could identify it a few minutes later. These results sit awkwardly with the view that language is the product of biological evolutionary purposes fine-tuning it for reliable communication. If that were the case, one would expect language users to be reliable communicators, sensitive to instances of communicative failure. Our findings are more consistent with accounts that see language as having evolved to serve social goals alongside communicative ones (Dunbar, 1996), or accounts of language evolution that downplay the role of adaptive biological evolution generally, emphasizing the role of cultural evolution instead (Dediu et al., 2013; Tamariz & Kirby, 2016; Tomasello, 2010).

In either case, it is likely that the social, phatic, dimension of conversation is an important part of the explanation for our findings. This would help to explain why incoherence detection was particularly poor in Study 3, in which participants conversed face-to-face and were thus under a greater social pressure than in the other two studies, where they could not see each other. It might also help to explain why the *highest* detection rate was observed in the Gender-swap condition of Study 2, where the inserted line had a social significance lacking in Study 3 or the Character-swap condition of Study 2 (or, indeed, the majority of crossings in Study 1). However, the difference between conditions in Study 2 was not significant, and other differences make it difficult to compare this condition with the other studies directly.

Identifying precisely what mechanisms lie behind our findings, and how their relative importance varies from context to context, is an important next step in this investigation. In the meantime, however, we have presented compelling evidence that human linguistic communication is not as reliable as tends to be assumed, as well as a paradigm for investigating this phenomenon in a controlled setting.

Acknowledgments

We thank the students in the Experimental Semiotics Lab at Yeshiva University for their help with data collection.

References

Blackburn, P. L. (1999). *The code model of communication: A powerful metaphor in linguistic metatheory*. Unpublished doctoral dissertation. (Retrieved from ProQuest Dissertations & Theses Full Text database – 304574107)

Dediu, D., Cysouw, M., Levinson, S. C., Baronchelli, A., Christiansen, M. H., Croft, W., ... Lieven, E. (2013). Cultural evolution of language. In P. J. Richerson & M. H. Christiansen (Eds.), *Cultural evolution: Society, technology, and religion. strüngmann forum reports* (Vol. 12, pp. 303–332). Cambridge, MA: MIT Press.

Dunbar, R. I. M. (1996). *Grooming, Gossip and the Evolution of Language*. London: Faber and Faber.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Galantucci, B., & Roberts, G. (2014). Do we notice when communication goes awry? an investigation of people's sensitivity to coherence in spontaneous conversation. *PLoS ONE*, *9*(7), e103182.

Galantucci, B., Roberts, G., & Langstein, B. (submitted). Content deafness: When coherent talk just doesn;t matter.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, vol. 3: Speech acts* (pp. 41–58). New York: Academic Press.

Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards (Eds.), *The meaning of meaning* (pp. 146–152). London: Routledge.

Marks, L. E., & Miller, G. A. (1964). The role of semantic and syntactic constraints in the memorization of english sentences. *Journal of Verbal Learning and Verbal Behavior*, *3*(1), 1–5.

Mills, G. J. (2012). The evolution of miscommunication. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The 9th international conference on the evolution of language (evolang ix).* Singapore: World Scientific.

Park, H., & Reder, L. M. (n.d.). Moses illusion: Implication for human cognition. In R. F. Pohl (Ed.), *Cognitive illusions* (pp. 275–291). Hove: Psychology Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*, 707–84.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*, 368–373.

Roberts, G., Langstein, B., & Galantucci, B. (2016). (in)sensitivity to incoherence in human communication. *Language and Communication*, *47*, 15–22.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review*, *5*(4), 644–649.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.

Tomasello, M. (2010). *Origins of human communication*. Cambridge, MA: MIT Press.

Tzanne, A. (2000). *Talking at cross-purposes: The dynamics of miscommunication*. John Benjamins.

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 607–632). Oxford: Blackwell.

Analytical Thinking Predicts Less Teleological Reasoning and Religious Belief

Jeffrey C. Zemla (jzemla@brown.edu)
Samantha M. Steiner (samantha_steiner@brown.edu)
Steven Sloman (steven_sloman@brown.edu)
Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer Street
Providence, RI 02912 USA

Abstract

Individual differences in reflectiveness have been found to predict belief in God. We hypothesize that this association may be due to a broader inclination for intuitive thinkers to endorse teleological explanations. In support of our hypothesis, we find that scientifically unfounded teleological explanations are more likely to be endorsed by intuitive compared to analytical thinkers, and that those who endorse teleological explanations are more likely to have religious beliefs.

Keywords: cognitive style; cognitive reflection test; religious belief; teleological explanations; causal reasoning

Introduction

Religious beliefs have enormous impact on our lives. They affect our moral and political values, our personal relationships, and our life goals. While people are often aware of the overt influences of religion, there may be a deeper connection between religious belief and general cognition (Barrett, 2000). How does cognitive style affect religious belief, and what makes some people more inclined to believe in God than others?

Kelemen (2004) proposed that children are *intuitive theists* who have a natural predisposition to believe in God. The foundation of this claim rests on the finding that children are teleologically promiscuous: Children tend to generate and endorse explanations for phenomena that ascribe a purpose or intention to the effect that is being explained (Kelemen, 1999), even when these explanations are scientifically unsupported. As a result, many children may readily endorse explanations that appeal to God—the ultimate teleological explanation—and this belief may persist through adulthood.

However, childhood religious beliefs do not always indicate what a person will believe as an adult. The Pew Forum on Religion & Public Life (2009) found that roughly half of Americans have changed religious affiliations at least once since childhood. A majority of those who reported no religious affiliation during childhood now belong to a religious affiliation. Conversely, 16% of adult Americans report having no religious affiliation, though only 7% were raised without a religious affiliation. Together, these finding suggest that religious beliefs in the United States are fluid, with some embracing religion only later in life and others abandoning their childhood beliefs. Children may be intuitive theists, but this cannot explain why religious views change over the lifespan.

One of the most commonly cited reasons for becoming affiliated or unaffiliated with a religion is a change in one's belief in God (Pew Forum on Religion & Public Life, 2009). Many of those who left their childhood religion expressed the view that "modern science proves religion is a superstition" and that they "just don't believe in God" anymore. In contrast, many of those who embraced religion later in life indicated that they felt "called by God."

Although it is common to think of belief in God as an unshakeable conviction, there have been many arguments both for and against the existence of God. Christian philosophers have presented teleological arguments in support of God's existence (Aquinas 1274/1938; Paley, 1802), while other scholars have presented scientific arguments against the existence of God (Darwin, 1859; Dawkins, 1986). How susceptible is faith to reasoned argument? If these arguments do affect religious belief, then individuals who are more strongly influenced by teleological arguments should be more likely to believe in God.

Recent research has suggested that individuals who adopt an analytical thinking style are less likely to have strong religious beliefs. For instance, those who neglect base-rates in a statistical problem in favor of qualitative, case-based reasoning were more likely to express a belief in God (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012). Likewise, participants who scored highly on the Cognitive Reflection Test (Frederick, 2005), used to measure the ability to deliberate and suppress intuitive yet incorrect responses to analytical problems, reported lower beliefs in God (Pennycook et al., 2012; Shenhav, Rand, & Greene, 2012). A variety of experimental manipulations support a causal role of analytical thinking in religious belief. Participants who were asked to describe a time when they used intuition to solve a problem were more likely to report a belief in God than those who described a situation that required careful reasoning to solve a problem (Shenhav et al., 2012). Even subtler manipulations, such as visual or linguistic priming of analytical thought, have shown an effect on reported religious beliefs (Gervais & Norenzayan, 2012).

Our hypothesis is that the relation between analytical thinking and religious belief is mediated by a willingness to endorse teleological explanations. Specifically, we hypothesize that intuitive thinkers will be more likely to endorse scientifically unfounded teleological explanations, in part because they are less likely to reflect on causal directionality. In other words, intuitive thinkers may endorse

teleological explanations because they confuse an effect with an intention to cause that effect (i.e., a teleological cause). In contrast, analytical thinkers may be more likely to reject teleological explanations that erroneously reverse cause and effect.

Experiment 1

In Experiment 1, we seek to establish a correlation between teleological reasoning and analytical thinking, as well as replicate two previous findings that show a connection between religious belief and teleological reasoning (Heywood & Bering, 2013; Banerjee & Bloom, 2014), and between religious belief and analytical thinking (Pennycook et al., 2012; Shenhav et al., 2012; Gervais & Norenzayan, 2012).

Methods

Participants. 65 participants (40 male, 25 female, ages 19-66, median age 29) completed the experiment via Amazon Mechanical Turk. All participants were located in the United States.

Materials and Procedure. The experiment consisted of four sections in order: an abbreviated numeracy scale, a set of true/false teleological explanations, an extended cognitive reflection test (CRT), and a religious questionnaire[1].

First, participants completed an abbreviated numeracy scale (Weller et al., 2013). The six-item scale measures a participant's ability to evaluate numerical problems involving percentages, probabilities, and basic mathematical operations. Some evidence suggests that the CRT may be essentially a measure of numeracy (Weller et al., 2013). We included a numeracy scale in order to rule out the possibility that CRT is measuring numeracy alone.

Second, participants completed a test of teleological reasoning consisting of 40 single-sentence explanations that suggest an intentional cause, taken from Kelemen, Rottman, & Seston (2013). Participants read each explanation and judged whether it was true or false. Twenty of the explanations were *test* stimuli. These explanations were scientifically unsupported teleological explanations, such as "*Bees frequent flowers in order to aid pollination."* Although it is true that bees frequent flowers and that they do aid pollination, they do not visit flowers *in order to* aid pollination. That is, the bee's role in pollinating flowers is not intentional.

The remaining twenty explanations were *control* stimuli, half teleological and half causal (non-teleological). Additionally, half of each type were true and half false. For example, two of the control stimuli include "*Houses have doorbells in order to make dogs bark,"* (false teleological) and *"A lightbulb shines because electricity passes through its filaments,"* (true causal). The control stimuli were used to ensure participants had no trouble reasoning about teleological and causal explanations in general. The order of the control and test stimuli were pseudo-randomized.

Third, participants completed the CRT, designed to measure the degree to which one engages in analytical as opposed to intuitive thinking. The original scale (Frederick, 2005) is made up of three questions in which the intuitive response is incorrect. Participants who engage in reflective thinking (analytical reasoning) may override the tendency to respond with their first instinct, and instead arrive at the correct answer. For instance, one question reads, "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?" The intuitive response is 100 minutes, whereas the correct analytical response if 5 minutes. Several variants of the CRT have been proposed that include more than three items. We employed a six-item version that includes the original three questions plus three new questions (Toplak, West, & Stanovich, 2014)[2].

Finally, participants completed a religious questionnaire identical to that used in Shenhav et al. (2012). Participants reported their belief in God on a 7-point Likert scale from "Confident atheist" to "Confident believer" (midpoint "Agnostic/undecided"). Additional Likert questions probed the participant's familial religiosity during childhood, belief in an immortal soul, change in religious belief since childhood, and whether they had a belief that convinced them of God's existence (yes/no).

Results

Following Kelemen et al. (2013), seven participants were excluded from analysis for answering fewer than 80% of the teleological reasoning control questions correct, leaving a total of 58 participants remaining. Although performance on these control stimuli was not related to any of the primary measures of interest (CRT or religious belief, both $p > .14$), there was a significant positive correlation between performance on the control questions and numeracy ($r = .26$, $p = .048$). The control questions were not used in subsequent analyses. Pairwise correlations from Experiment 1 are reported in Table 1.

CRT and Religious Belief. Participants were assigned a CRT score based on the number of correct (analytical) responses (0-6). We found a significant negative correlation between CRT and belief in God ($r = -.37$, $p = .004$). Participants who performed well on the CRT (more analytical responses) reported less of a belief in God than those who scored poorly on the CRT, even though CRT performance does not predict religious beliefs during childhood ($r = -.11$, $p = .41$).

[1] The full experimental stimuli used in both experiments are available at http://research.clps.brown.edu/SlomanLab/cogsci2016/

[2] Toplak, West, & Stanovich, (2014) propose a seven-item version of the CRT, but due to an error we excluded the sixth item.

Table 1: Correlations between variables measured in Experiment 1. * $p < .05$, ** $p < .01$, *** $p < .001$

Variable	Belief in God	Teleo. reasoning	CRT	Numeracy	Familial religiosity	Age	Gender
Belief in God	—						
Teleological reasoning	.24	—					
CRT	-.37**	-.32*	—				
Numeracy	-.14	-.29*	.55***	—			
Familial religiosity	.33*	-.05	-.11	-.03	—		
Age	.20	-.04	-.04	-.31*	.13	—	
Gender (1=M, 2=F)	.29*	-.02	-.22	-.38**	.22	.22	—
Education	.15	-.28*	.17	-.01	-.01	.28*	.23

Numeracy and Religious Belief. Performance on the six-item numeracy scale did not predict belief in God (r= -.14, p = .3). This suggests that the predictive ability of the cognitive reflection test is not due to differences in numerical ability.

Religious Belief and Teleological Belief. To construct a teleological reasoning score for each participant, we counted the number of teleological test questions that were endorsed (0-20). We found that participants who reported a strong belief in God were more likely to endorse teleological explanations, a marginally significant result (r = .24, p = .067). However, a post-hoc analysis controlling for familial religiosity indicated a significant partial correlation between teleological reasoning and belief in God (r = .27, p = .037). This suggests that the effect of teleological reasoning on religious belief may be moderated by religious upbringing.

CRT and Teleological Beliefs. Our original hypothesis was that those who engage in analytical reasoning are less likely to endorse false teleological explanations. Indeed, we found that participants who scored highly on the CRT were less likely to endorse teleological explanations (r = -.32, p = .016). We also found that numeracy predicts endorsement of teleological explanations (r = -.29, p = .029). However, given that the teleological explanations contained no numerical information, it is unlikely that numeracy has direct influence on teleological reasoning.

Discussion

Experiment 1 replicates several findings from prior literature. Specifically, we find that that both analytical thinking and teleological reasoning predict belief in God. We also demonstrate a novel effect: analytical style predicts willingness to endorse teleological explanations. Given these preliminary results, we conducted a follow-up study to examine these effects with a larger sample size. In addition, we propose and test a mechanism by which analytical style influences teleological reasoning.

Experiment 2

Experiment 1 found that participants who engaged in analytical thinking were less likely to endorse scientifically unwarranted teleological explanations. It is hard to imagine how to explain this result and others (e.g., Kelemen & Rosset, 2009) without assuming that teleological explanations have an intuitive appeal that can be overridden by more deliberative processes. In Experiment 2, we test a mechanistic account of why this might be. Specifically, we propose that intuitive thinkers are more likely to neglect causal directionality, leading them to endorse teleological explanations in which the intention to bring about an effect is itself a cause. We test whether participants who attend to causal relations are less likely to endorse teleological explanations.

We suggest this mechanism may lead high CRT to be unwilling to endorse teleological explanations. For example, consider the explanation "Bees frequent flowers in order to aid pollination." An individual who is aware of causal directionality may correctly reason that a bee's flower visit is a *cause* of pollination, but that the reverse is not true: the desire to pollinate a flower is not a *cause* of the bee's visit. That is, bees do not visit flowers *in order to* pollinate them. This ability to identify causal directionality is crucial to successfully evaluate teleological explanations. If intuitive thinkers are more likely to neglect causal directionality, they should be more likely to endorse some false teleological explanations.

Methods

Participants. 188 participants (99 male, 87 female, 1 other, 1 unknown, ages 18-66, median age 26) completed the experiment. 135 participants (72%) were recruited via Amazon Mechanical Turk, while 53 participants (28%) were recruited from the Brown University undergraduate subject pool for course credit. All participants were located in the United States.

Materials and Procedure. Participants completed four tasks: the six-item CRT (identical to Experiment 1), an abbreviated teleological reasoning test, a causal reasoning test, and a religious questionnaire[1]. Questions were inter-

mixed and pseudo-randomized except for those from the religious questionnaire, which were presented last.

To avoid redundancy, we tested an abbreviated version of the teleological reasoning test used in Experiment 1. We used only five test items and three control items. The test items chosen were the most discriminative in Experiment 1, teleological explanations whose responses were highly dependent on CRT score. The three control items all had a high proportion of correct responses in Experiment 1.

The causal reasoning test included six conditional probability problems, including two from Tversky and Kahneman (1977) and four novel problems with the same structure. Each question presented participants with a scenario and asked them to choose the most likely of three options. One question, borrowed from Tversky and Kahneman (1977), read:

Which of the following is more probable?

(a) That a girl has blue eyes if her mother has blue eyes
(b) That a mother has blue eyes if her daughter has blue eyes
(c) Equally likely

The correct answer to this, and to all of our conditional probability problems was (c): Equally likely. Kahneman and Tversky (1977) found that people often exhibit a bias when answering these problems, with a significant portion of participants responding (a) and very few responding (b). This result has been explained as a causal bias due to the strong forward causal relation from a mother's eye color to a daughter's eye color. Though people correctly reason that a daughter's eye color cannot influence her mother's eye color, they fail to identify that this does not influence the relative conditional probabilities of (a) and (b).

Participants with strong causal reasoning skills may be more likely to fall prey to this bias. Conversely, those who neglect causal directionality may be less likely to exhibit this bias and instead arrive at the correct answer. Thus, we hypothesize that participants who engage in deliberate analytical thinking will be better at causal reasoning, and less likely to arrive at the correct answers. As such, we expect participants who score high on the CRT to be *less likely* to answer the causal reasoning questions correctly.

In addition, we included three causal reasoning control questions. These questions were of a similar format to the causal reasoning test questions, except that the conditional probabilities were unbalanced so that the correct answer was not "Equally likely." For instance, one question asked:

Which of the following events is more probable?

(a) That a woman has held a leading part in a musical given that she can sing
(b) That a woman can sing given that she has held a leading part in a musical
(c) Equally likely

The correct answer is (b) as all leads in a musical can sing, though not everyone who can sing has held the lead in a musical. These causal reasoning control questions were used to measure a participant's ability to reason through conditional probability questions more generally.

Finally, we used a religious questionnaire identical to that of Experiment 1, with two additional questions to gauge the participant's belief in God's agency: "To what extent do you believe that a God or Supreme being actively influences events in the world over the course of a day, and over the course of a century?"

Results

Eleven participants were excluded from the analyses for failure to answer all three control teleological questions correctly. The teleological control questions were not analyzed further.

CRT, Teleological Reasoning, and Belief in God. We replicated many of the effects found in Experiment 1. Participants who scored high on the CRT were less likely to endorse teleological explanations ($r = -.32$, $p < .001$, see Figure 1) and less likely to report a belief in God ($r = -.29$, $p < .001$). We also found a robust correlation between teleological reasoning and belief in God ($r = .27$, $p < .001$, see Figure 2). Note that undergraduate students were less likely to answer the CRT questions correctly ($p = .04$) but more likely to express a belief in God ($p = .07$) compared to Mechanical Turk participants.

Our primary hypothesis was that analytical style leads to decreased religious belief because analytical reasoning affects our ability to evaluate teleological arguments. We tested this hypothesis using a mediation analysis, and find that teleological reasoning does mediate the relation between CRT performance and belief in God, Sobel's $z=-2.24$, $p=.025$. This shows that teleological reasoning accounts for a significant portion of the shared variance between CRT and belief in God. However because our design is entirely correlational, we cannot rule out alternative hypotheses. It may be, for instance, that religious beliefs influence teleological reasoning.

Causal Reasoning. Participants were assigned a causal reasoning score based on the number of correctly answered causal reasoning test questions (0-6). We did not find support for our hypothesis that intuitive thinkers are less likely to exhibit a causal reasoning bias. Causal reasoning scores did not correlate significantly with CRT scores, teleological reasoning scores, or belief in God (all $p > .17$).

Table 2. Correlations between variables measured in Experiment 2. * $p < .05$, ** $p < .01$, *** $p < .001$. Shaded cells indicate effects that are significant ($p < .05$) in both Experiments 1 and 2.

Variable	Belief in God	Teleo. reasoning	CRT	Familial religiosity	Causal reason. (test)	Causal reason. (control)	Age	Gender
Belief in God	—							
Teleological reasoning	.27***	—						
CRT	-.29***	-.32***	—					
Familial religiosity	.45***	.05	-.07	—				
Causal reasoning (test)	-.10	-.08	-.01	-.13	—			
Causal reasoning (control)	-.03	-.16*	.26***	.15	-.30***	—		
Age	.11	-.12	.10	.09	-.03	.07	—	
Gender (1=M, 2=F)	.10	.08	-.31***	.03	-.11	.07	.05	—
Education	-.06	-.18*	.27***	.05	-.14	.19**	.26***	-.09

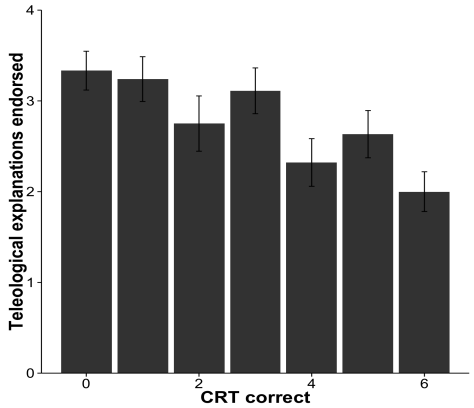


Figure 1. Performance on the CRT negatively predicts willingness to endorse teleological explanations.

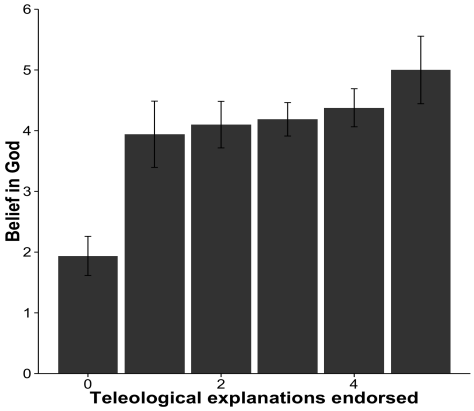


Figure 2. Participants who endorse teleological explanations are more likely to report a belief in God.

We assessed participants' performance on the causal reasoning control questions to ensure that this was not due to a failure to understand conditional probability problems more generally. Performance on these control questions, where the correct answer is consistent with a causal bias, was significantly positively correlated with both CRT scores (r = .26, $p < .001$) and negatively correlated with teleological reasoning (r = -.16, $p = .039$). That is, participants who were better at these control conditional probability problems were more likely to engage in analytical thinking and were less likely to endorse teleological explanations, as expected.

Performance on the causal reasoning control questions was *negatively* correlated with performance on the test questions (r = -.30, $p < .001$). This is consistent with our initial hypothesis: some participants are lured by the forward causal option, leading them to get the control questions correct but the test questions wrong; other participants are not influenced by the causal lure and instead prefer the "Equally likely" option, leading them to get the test questions correct but the control questions wrong. What is surprising is that this behavior is *not* accounted for by analytical style.

Discussion

Our results lend support to our initial hypothesis that teleological reasoning is a mediating factor that affects the relationship between analytical thinking and religious belief. However we cannot rule out the alternative possibility that religious participants were more likely to endorse teleological explanations *because* of their religious beliefs. This may be a reasonable position for some explanations. For instance, a religious thinker who believes in intelligent design may endorse the teleological explanation that "lemurs have adapted in order to avoid extinction." However it seems less likely that those same religious beliefs would cause a participant to endorse non-evolutionary teleological arguments, such as "hurricanes circulate seawater to gather energy," or teleological arguments that appear to superficially endorse evolution, such as "the fittest animals survive so that species can grow stronger."

We found no strong support for our mechanistic hypothesis that intuitive thinkers endorse teleological explanations due to a failure to distinguish between cause and effect, but we did uncover a novel and strange effect: Performance on standard conditional probability questions predicted susceptibility to causal bias on conditional probability problems used by Kahenman and

Tversky (1977). Although we predicted this result, our proposed mediating variable (CRT performance) failed to account for it.

One possibility is that performance on the causal reasoning test questions is influenced by two factors. On the one hand, participants must think analytically. On the other hand, participants must ignore causal directionality. If these two factors are in opposition to each other, it could explain why we observed a null effect for the correlation between causal reasoning test questions and CRT performance.

General Discussion

In two experiments, we replicated previous findings that show religious beliefs are negatively correlated with analytical thinking, but positively correlated with willingness to endorse teleological explanations. In addition, we found that analytical thinking is negatively correlated with willingness to endorse teleological explanations, and that teleological reasoning mediates the relationship between analytical style and belief in God. However, our proposed mechanistic account of why intuitive thinkers endorse teleological explanations was not supported.

One limitation of the present studies is their reliance on correlational data. Future research should attempt to manipulate individual factors, such as teleological reasoning and analytical style, to clarify the causal direction of these effects. Unraveling this complex set of relations will further our understanding of reflectiveness and the sources of religious belief.

Acknowledgments

This publication was made possible through the support of a grant from The Varieties of Understanding Project at Fordham University and The John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of The Varieties of Understanding Project, Fordham University, or The John Templeton Foundation. We would like to thank Amitai Shenhav for helpful discussion and Deborah Kelemen for sharing her experimental stimuli used for the teleological reasoning task. We would also like to thank Daniela Lopez and Cristina Ballarini for their excellent research assistance.

References

Aquinas, T. (1274/1938). *Summa theologica* (Vol. 66, pp. II-II). impressa per Andream de Torresanis.

Banerjee, K., & Bloom, P. (2014). Why did this happen to me? Religious believers' and non-believers' teleological reasoning about life events. *Cognition*, *133*(1), 277-303.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, *4*(1), 29-34.

Darwin, C. (1859). *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. W. F. Bynum (Ed.). AL Burt.

Dawkins, R. (1986). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25-42.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493-496.

Heywood, B. T., & Bering, J. M. (2014). "Meant to be": How religious beliefs and cultural religiosity affect the implicit bias to think teleologically. *Religion, Brain & Behavior*, *4*(3), 183-201.

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, *70*(3), 241-272.

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological Science*, *15*(5), 295-301.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138-143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074-1083.

Paley, W. (1802). Natural Theology; Or Evidences of the Existence and Attributes of the Deity, Collected from the Appearances of Nature by William Paley, Ed. FC and J. Rivington.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335-346.

Pew Forum on Religion & Public Life. "Faith in flux: Changes in religious affiliation in the U.S." Pew Research Center, Washington, D.C. (April, 2009). http://www.pewforum.org/files/2009/04/fullreport.pdf.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 1-6.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168.

Tversky, A., & Kahneman, D. (1977). Causal Schemata in Judgments Under Uncertainty. Eugene, Oregon: Defense Technical Information Center.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J. and Peters, E. (2013), Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach. Journal of Behavioral Decision Making, *26,* 198–212.

Gender Differences in the Effect of Impatience on Men and Women's Timing Decisions

Moojan Ghafurian and David Reitter
moojan@psu.edu reitter@psu.edu
The Pennsylvania State University
University Park, 16802, USA

Abstract

Decisions over the timing of actions are critical in several safety, security and healthcare scenarios. These decisions, similar to discrete decisions, can be influenced by biases and individual traits. In this paper, a bias of impatience is studied in an experiment with 626 participants, with a focus on gender differences. Impatience was moderated with a manipulation of a variable-speed countdown. Men and women differed in how they expressed impatience. While men systematically and irrationally act earlier when become impatient following the slower countdowns, women react by irrationally requesting earlier information about the outcome of each trial, and impulsively pressing an inactive key.

Keywords: Impatience; gender differences; decision-making; timing decisions; women; men

Introduction

When people make decisions, their choices are gravely affected by systematic biases and individual preferences. This set of well-known phenomena has been extensively studied for the case of choices between discrete alternatives (e.g., Tversky & Kahneman, 1974; Gigerenzer, 1996; Wittmann & Paulus, 2008 ;Croskerry, 2002). However, we know much less about the case of decisions over a time continuum, that is, *when* rather than *how* to act. Timing decisions, which are decisions about *when* to take an action, as opposed to *what* to do, are critical in several security, safety, and healthcare scenarios, where acting early and late can both lead to dramatic losses (e.g. when to undergo medical screening, when to change car tires, or when to check smoke detectors in a big company).

In this paper, we study a bias we found on timing decisions: impatience. We specifically examine the hypothesis that men and women react differently to induced impatience. Understanding gender differences, we believe, is essential to accurately modeling the effects of biases and individual differences in timing decision-making.

The existence of gender differences in timing decisions is to be expected, if discrete decision-making serves as a comparison. Risk-taking behavior of men and women has been investigated especially in the context of financial decision making. Powell and Ansic (1997) proposed that women are generally less risk-seeking compared to men, regardless of the costs, ambiguity of the situation, task familiarity, and framing. They also showed that men and women use different strategies in financial decisions, even if this difference did not reliably affect the final outcome.

The relative risk propensity of men and women is strongly affected by framing. For the general case, the classical result is that risk-taking increases in the face of gains compared to potential losses (Daniel Kahneman, 1979). This is true more so for men than for women in surveys and in abstract gambles (Schubert, Brown, Gysler, & Brachinger, 1999): here, women are comparatively more risk-taking towards losses. An explanation for the differences, if they exist, may be that women are more sensitive to punishment (Cross, Copping, & Campbell, 2011). However, Schubert et al. (1999) specifically showed that in financial decisions, the relative risk propensity of men and women is strongly affected by the decision frame. For practical contexts, however, men and women were not reliably different in terms of their propensity to take risks with gains vs. losses. Thus, gender differences are not uncontroversial in discrete decision-making, and warrant careful analysis in other types of decisions, too.

Gender differences do affect how individuals value future rewards in relation to the time gap between the discrete decision made now and the reward obtained later. Dittrich and Leipold (2014) asked participants to choose between two options: receiving $100 in one month, or receiving $100+N in 13 months. Women accept delayed gratification considerably more than men. Based on these results, they concluded that women tend to be more patient compared to men (Dittrich & Leipold, 2014). We consider this to be a result of differences in temporal discount rate. Although impatience can be a factor affecting temporal discounting, we believe that it has a broader definition, which is different from the discount rate. We hypothesize that it can be reflected differently by individuals, and by people of different genders. For lack of a better operationalization, we refer to The Oxford English Dictionary, which defines impatience as "the feeling of being annoyed or irritated by somebody/something, especially because you have to wait for a long time", and "the desire to do something soon or for something to happen soon". This latter definition is the operational one for the present study, and we refer to the earlier definition by checking participants' overall satisfaction in our experiments. We investigate gender differences in the reactions to an impatience-inducing manipulation, which may prompt individuals to act earlier than they would otherwise, or to make costly choices that allow them to obtain information earlier than they would otherwise. These biases apply even when associated with a lower expected reward.

We will present two experiments, in which people decide on timing of actions. In experiment 1, we moderate impatience and study the effect on different genders. In experi-

ment 2, which involves a more straightforward task, we further study differences in how men and women reflect impatience.

Experiment 1

We use experimental games to study participants' choice of timing in situations that are similar to real-world. In the first experiment, we use a game, which is inspired by the *FlipIt* game of "stealth takeover" (Van Dijk, Juels, Oprea, & Rivest, 2013). In our game, participants are required to find a strategy to search for an unknown action made by an opponent. Each round lasts 30 seconds. Each attempt to catch the opponent, and each second of latency in catching her/him have costs. In each round of the game, participants first decide on the timings of checking on the opponent. After setting the timings of checks, participants choose between the two possible ways to start the game: the first choice is to start the game normally. By choosing it, participants get visual updates on whether the opponent has played or not when they reach one of their checks. That is to say, if a participant sets two checks at t_1 and t_2, and the opponent plays at t_O ($t_1 < t_O < t_2$), and he/she starts the game normally, then at t_1 he/she will realize that the opponent has not played yet, and at t_2 he/she will be updated that the opponent has played at t_O and he/she has caught it at t_2 (see Figure 1). The second option is to start it *live*. The live option, which costs the equivalent of two additional checks (200 points), lets participants see when the opponent plays instantly. That is to say, in the aforementioned example, participants see the opponent's action as soon as it plays at t_O, however, they still need to wait until t_2 to be able to catch it (see Figure 1). We hypothesize that impatience can lead to a tendency to receive information faster, and we use the *live* option to study our hypothesis. We consider the live option as an irrational way of starting the game, since participants need to pay to choose it. However, this choice would have no benefits for them: it neither affects the duration of the game, nor their outcome, since even though they see the opponent's action, they would still need to wait for their next check to catch it.

The experiment has four conditions. In all the conditions except for the control condition, we introduce a 15-second countdown (and a sensible explanation "Saving Checks") between choosing to start the round and actually watching the round. We hypothesize that waiting before watching the round increase participants' impatience. We also hypothesize that it is possible to moderate impatience by showing different speeds of countdown, and test it in different conditions of the experiment. Thus, the conditions of the experiment are as follows:

- **NoWait (control)**: Participants experience no waiting and the game starts right away.
- **5CD**: Participants experience a 15-second wait between starting and watching the game, which is accompanied by countdowns from **5 to 1** (each count lasts 3 seconds).

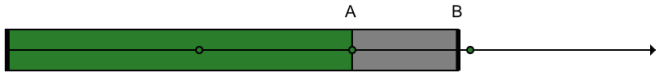


Figure 1: Feedback during a round, which is started normally. The game is current and at point B. Since the latest check time of the participant is set to be at point A, no information after point A is revealed. The dots on the bar represent check times that are set by the participant.

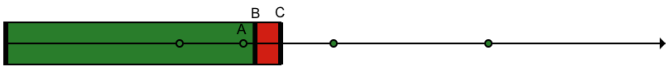


Figure 2: Feedback during a round, which is started live. The game is current and at point C. The latest check time of the participant is set to be at point A, and we can see that the opponent has played at point B. However, the participant needs to wait until the next set check (shown with dots on the game bar) to catch the opponent.

- **10CD**: Participants experience a 15-second wait between starting and watching the game, which is accompanied by countdowns from **10 to 1** (each count lasts 1.5 seconds).
- **15CD**: Participants experience a 15-second wait between starting and watching the game, which is accompanied by countdowns from **15 to 1** (each count lasts 1 second).

After participants are randomly assigned to the experimental conditions, they first complete a survey with four demographic questions, three basic integrity questions, seven risk propensity assessing questions (Meertens & Lion, 2008), and five need for cognition (NFC) assessing questions (Wood & Swait, 2002). Afterwards, they see the game instructions and start the game.

Since we want to make sure that the participants are paying attention to the game, we prevent them from switching to other windows on the screen, and monitor their attention by asking to press a specific key immediately after the unpredictable end of the game.

Cover Story

We add a cover story to the game and call it the *Cookie Monster Game* to make it easier to understand for the participants on Amazon Mechanical Turk. Participants are told that they have invited the Cookie Monster (the opponent) over for dinner. While they are cooking for their guest in the kitchen, the Cookie Monster is waiting in the living room with a big box of cookies. Participants are told that the Cookie Monster will definitely start eating the cookies, and with a constant pace, but no one knows when. They need to find a strategy (set different alarm times) to check on him. Whenever they catch the Cookie Monster, the game ends. However, if they check on the Cookie Monster before he has started eating the cookies, they need to give him some cookies to apologize for not

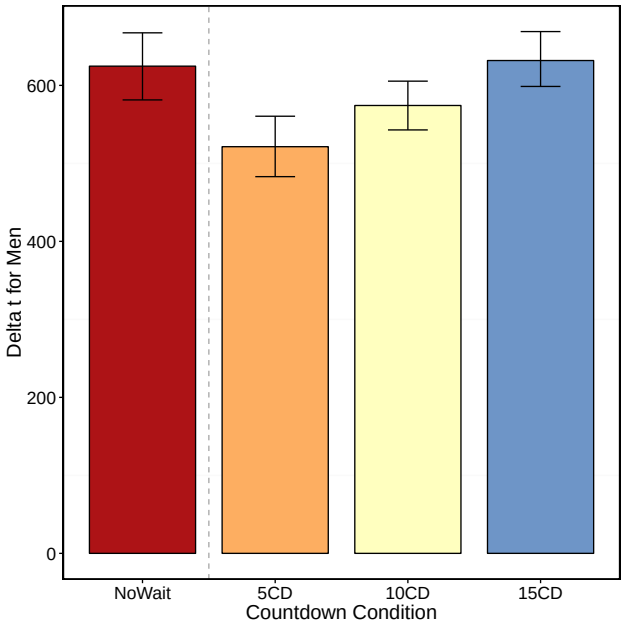


Figure 3: Average time between check times for men in each condition. Lower values reflect earlier (more frequent) checks. Men who see the slower countdown check more frequently.

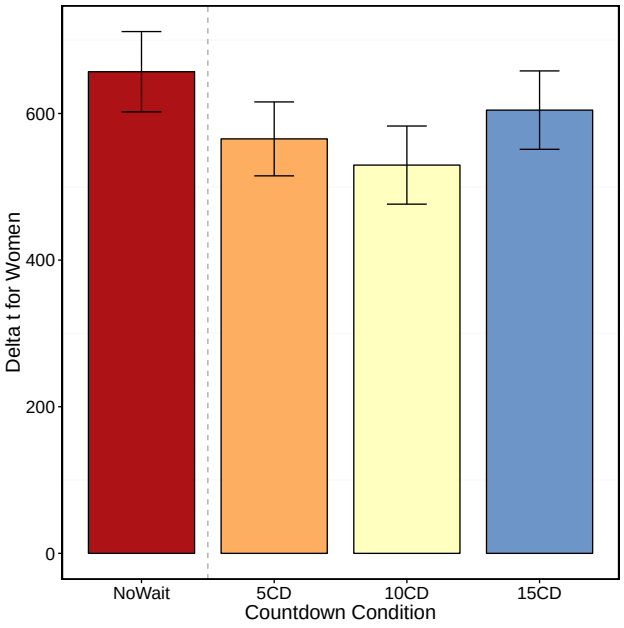


Figure 4: Average time between check times for women in each condition. Lower values reflect earlier (more frequent) checks. Speed of countdown does not significantly affect the frequency of women's checks

trusting him. The number of remaining cookies will translate to the bonus money that they will receive at the end of the experiment.

Method

503 volunteers (211 female and 292 male, age mean 33.9y, range: 18-77) on Amazon Mechanical Turk participated and passed the attention tests. Each participant played 6 rounds of the game. The results of the first round is not included in the analysis: it was labeled "practice round", and the participants first experienced countdowns and learned about the live option in this round.

Results

We calculate an average time difference between the checks, per participant, and per round. We call it Δt. Δt represents the average time that participants decide to wait between their checks, which reflects how frequently they check. We believe that a change of Δt in different conditions can reflect a change in the level of participants' impatience. Δt is calculated for all participants in all experimental conditions, and based on genders. The results for men and women are visualized in Figures 3 and 4.

In a linear mixed-effect model (see Table 1), we investigate the effect of our manipulation on participants' Δt, which we interpret as impatience. The model predicts men's and women's Δt based on round number, need for cognition, and the duration of each count in the countdown (1 for 15CD, 1.5 for 10CD, and 3 for 5CD). Round number reflects participants' proficiency in the game. NFC, which is the tendency to engage in thinking (Olson, Camp, & Fuller, 1984) reflects how much individuals enjoy and are desired to think. We use it in our model as we hypothesize that NFC affects participants' strategy. A random intercept varying by subject is fitted to account for the individual differences.

Table 1: Regression model predicting Δt for different genders. Count duration and round num are centered. The model was not sensitive to age as a predictor.

Covariate	Estimate	SE	t	Pr ($>\|t\|$)
genderF	5.810	0.361	16.114	< 0.0001
genderM	6.257	0.293	21.379	< 0.0001
round num	0.242	0.102	2.368	< 0.05
need for cognition	0.016	0.020	0.767	0.444
genderF:count duration	-0.271	0.538	-0.504	0.615
genderM: count duration	-1.011	0.488	-2.071	$< .05$

Men's Δt is significantly affected by the speed of countdown ($se = 0.488$, $t = -2.071$, $p < .05$). Men seeing the faster countdowns play less frequently than the ones seeing the slower countdowns. Women's Δt, however, is not affected by the countdown manipulation ($se = 0.538$, $t = -0.504$, $p = 0.615$).

In another model we investigate the effect of our manipulation on the choice of live. The model uses NFC, count duration, and round number to predict choice of live. A random intercept varying by subject is fitted to account for the

Table 2: Regression model predicting choice of live. Count duration and round num are centered.

Covariate	Estimate	SE	t	Pr ($>\|t\|$)
genderF	1.135	0.059	19.149	$< .0001$
genderM	1.133	0.060	19.020	$< .001$
NFC	-0.0002	0.002	-0.122	0.903
genderF:count duration	0.112	0.048	2.319	$< .05$
genderM:count duration	0.002	0.044	0.040	0.968

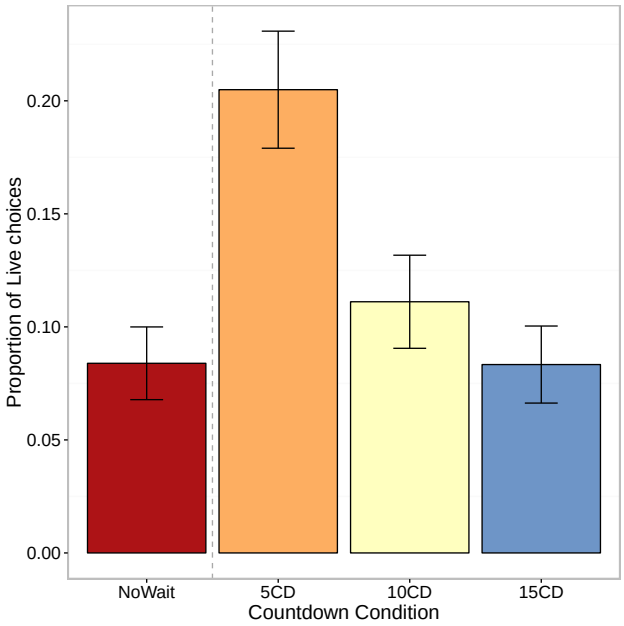


Figure 5: Proportion of live choices for women in each condition

individual differences (see Table 2). Women's choice of live is reliably affected by the speed of countdown ($se = .048$, $t = 2.319$, $p < .05$). Women seeing the slower countdown pay to choose *live* reliably more (see Figure 5). Men's choice of *live*, however, is not affected by the speed of count down ($se = .044$,$t = .040$, $p = .968$).

We report no correlation between Δt and the measured risk propensity.

Overal satisfaction: At the end of the game, participants had a chance to leave us comments. Since it was not mandatory, many participants left without leaving any comments [1]. Positive comments were predicted based on NFC and count duration (see Table 3). The number of men's positive comments is significantly affected by the speed of countdown. Men seeing faster counts leave reliably more positive comments ($se = 0.439$, $z = -2.496$, $p = 0.013$). However, the number of women's positive comments is not affected by the speed of countdown ($se = 0.458$, $z = 0.059$, $p = 0.953$).

[1]Some examples of the positive comments are: "Great taks", "Thanks. It was very fun & unique.", "Wow!!! I loved this game. Please send me another one like this in the future. Thanks". Negative comments pertained to technical issues; they were very low in number.

Table 3: Predicting positive comments based on the game condition.

Covariate	Estimate	SE	z	Pr ($>\|z\|$)
genderF	-1.461	0.312	-4.683	$< .0001$
genderM	-1.039	0.235	-4.428	$< .0001$
need for cognition	0.054	0.019	2.793	$< .01$
genderF:log(countDuration)	0.027	0.458	0.059	0.953
genderM:log(countDuration)	-1.096	0.439	-2.496	0.013

Experiment 2

In order to test more explicitly for impatience, we consider data from a simpler game. In this section, we provide analysis of data reported previously (Ghafurian & Reitter, 2014, Exp.2). 123 participants recruited on MTurk took part for compensation that included an incentive-compatible bonus. The game is similar, but much less complicated than the game in Experiment 1. In this game, the opponent played exactly once during the game, and the participants were asked to either check (play) before the opponents move (early condition), or after it (late condition). Participants were given only one chance to check, and based on the experimental condition, the best strategy was to either play early (at the beginning of the game), or late (at the very end). This experiment allows us to study impatience in a situation where decision-making is dynamic and the rational choice is clear, as opposed to the first experiment, where *checks* were pre-defined and the rational choice was not clear.

Participants in both early and late conditions were also assigned to either of two different trial durations (5 and 15 seconds). To make sure that participants are not motivated to play early (to finish the experiment faster), the game did not end after they made their single check, and continued for the full period of the round. In Ghafurian and Reitter (2014), we discussed how impatience affected the participants' actions. Participants performed reliably better in the early condition, and showed a consistent preference to play early.

With a view of gender differences, we measured the check times as well as how many times participants pressed *check* impulsively, despite the fact that they only had one chance (this was made clear to them). A check was considered *impulsive* when the space key was pressed repeatedly, even though though the button was deactivated and hidden after the first time. The number of times the space bar was pressed (or the area on the screen was clicked) was counted.

We will analyze the timing of each check, as well as the number of impulsive attempts for checking.

Cover Story

To make it easier to understand for the participants on Amazon Mechanical Turk, we add the *Cookie Monster Game*

cover story to this experiment as well. However, in this game, the story is different. The abstract version of the cover story for early and late conditions are as follows:

Late condition: *Imagine you are the Cookie Monster and your neighbor has promised to bring you some cookies. However, she has not told you when. You have only one chance to check very quickly. Your job is to figure out when to check. You know that your neighbor will definitely put the cookies outside for you, but you do not know when. If you check and they are still not there, you have lost your chance. Your job is to decide when to open the door.*

Early condition: The cover story in the early condition is very similar to the late condition. In this condition, the cookies are definitely left for them in the beginning of the game, however, the neighbor might take them at some point.

Number of cookies will translate to a bonus money in both conditions, and in this experiment, the participants either get all the bonus in each round, or none of it.

Method

123 volunteers (43 female, 79 male, 1 unknown; mean age 31y, from 18 to 67) recruited on Amazon Mechanical Turk and each played 10 rounds of the game. Participants were randomly assigned to game lengths of 5 or 15 seconds, and to one of the early or late conditions.

Results

In this experiment, we define Δt as the time delay between the start of the game and participant's check in the early condition, and the time delay between participant's check and the end of the game in the late condition. We also calculate the number of times participants press "check" impulsively, and call it Num_p. We label the 5-second and 15-second games short and long respectively.

We hypothesize that behavior of both men and women are influenced by impatience, but they might reflect it differently. Using linear-mixed effect and models, we investigate Δt and Num_p in early and late conditions, and in the short and long duration games. Random intercept varying by subject is fit in all the models.

Women's Δt is generally affected by the experimental condition. Women in the late condition have a reliably higher Δt, which shows that they fail to wait until the end of the game to make their checks (short game: $se = 0.589$, $t = 2.168$, $p < .05$;long game: $se = 0.486$, $t = 2.197$, $p < .05$). No reliable values for men could be obtained from the data (short game:$p = 0.318$, long game: $p = 0.138$)

In the shorter game, we did not get any reliable effect for Num_p. This is sensible as 5-seconds is too short for pressing a key several times. In the longer game and in the early condition, where people should wait longer for the game to end after they make their check, we see a significant difference between Num_p of men and women (see Table 4): men's Num_p is significantly less than women's: women impulsively press the key reliably more than men ($se = 0.384$, $z = -2.053$, $p < .05$).

Table 4: Regression model predicting Num_p in the long (15-second) game. NFC is centered. Condition by itself (shown in the last two rows of the table) is not meaningful in predicting numbers pressed, since in the early condition, more time is available to press the key after the first check.

Covariate	Estimate	SE	z	Pr ($>\|z\|$)
Intercept	2.631	0.297	8.856	$P < .0001$
NFC	-0.044	0.019	-2.284	$p < .05$
genderM	-0.794	0.385	-2.063	$< .05$
genderF:condLate	-1.571	0.392	-4.007	$p < .0001$
genderM:condLate	-0.520	0.328	-1.588	0.112

In all participants, a higher need-for-cognition trait was reliably associated with a decreased number of repeated button presses ($se = 0.019$, $z = -2.284$, $p < .05$).

We report no correlation between either Δt, or Num_p and the measured risk propensity.

Discussion

In the first experiment, we effectively moderated impatience, using faster countdowns during a 15-second waiting time. This manipulation reliably affected the quality of timing decisions of men and women. We observed differences in how impatience is reflected in behavior by each gender. Impatient men checked more frequently. Impatient women paid to activate a valueless option. Further, men were less satisfied with the game when impatient; men seeing slower counts left fewer positive comments at the end. Women, on the other hand, did not react in this way to the manipulation.

The second experiment, which is a more fine-grained analysis of existing data, showed that women reflect impatience by making their check early in the late condition. Which means that women fail to wait until the end of the game in the late condition, doing which guarantees receiving a bonus. Further, in shorter games and in the early condition, women impulsively ("impulsive" is defined by Carver, 2005 as "at tendency to act spontaneously and without deliberation") pressed a valueless key. We believe that this number represents an impulsive action made because of impatience.[2]

A major difference between the first and the second experiment is the difficulty of finding the best strategy to play. Although the rational solution of the second experiment is immensely straightforward, the best strategy for the first experiment is difficult to find for a participant using deliberative reasoning. The key to solving the game through reasoning is understanding that the hazard rate increases after each unsuccessful check, because we obtain information that the oppo-

[2]We do see a potential confound. If women were less attentive to the technical description of the task, they might have mistakenly believed that once they press the button, the game would end. If this were true, we would see an effect among low numbers of checks, because after the first few, it is evident that the trial does not end when a check is made. The actual number of checks is relatively high (mean: 10.55, range: [1,424]).

nent has not played until now, and because we know that the opponent will definitively play before the round ends.

The results from our experiments are likely compatible with a model that has men perceiving this hazard rate differently. In other words, as men become impatient, the perceived hazard rate might increase (which would explain why they play more frequently in Experiment 1). In contrast, we would model a women's reaction to impatience as an increase in impulsive actions (which would explain why they choose *live* more frequently in Experiment , press the inactive check frequently in Experiment 2, and failed to wait until the end of the game in experiment 2). Further research is required to validate such models.

Finally, men's impatience affected their overal satisfaction of the game, and reflected their impatience by leaving less positive comments. However, women's overal satisfaction of the game was not affected by the level of impatience.

Conclusion

In this paper, we compared men and women's timing decision-making using two games of timing. We found a set of interesting and surprising differences in how genders realize impatience in decision-making. The impatience effects were provoked using a visual countdown run at different speeds in a between-subject design. This manipulation created different changes in behavior in men and women. Men systematically changed the timing of their actions, even if this timing reduced their task performance: the slower the countdown that preceded the round, the more did men advance the timing of their actions. We could not detect such an effect for women. On the other hand, women expended their incentive payments on a costly (but valueless) option that allowed to view the outcome of each trial earlier. This option, however, neither affected the duration of the round, nor the outcome.

In a second, confirmatory experiment, we elicited impatience without a countdown manipulation, and found that, again, men and women behaved differently. Women showed systematically more impatience through their direct game behavior, which affected their outcome. They also showed impulsive actions that could not influence the payoff.

Understanding why people make irrational decisions requires better models of how individuals make those decisions. Rather then reflecting a hypothetical *average* human being, we strive for models that take individual differences into account. Gender is a simple and coarse distinguishing factor, which, as it turns out, is highly relevant. Still, much is not yet understood. However, the empirical comparison of strategies and biases in timing decisions warrants further analysis, not only in terms of the potential biological mechanisms, but also regarding the cognitive architecture that gives rise to such behavioral effects.

Acknowledgements

We thank Jens Grossklags, Frank Ritter, Pezhouhan Tavassoti-Kheiry, Bill McDowell, and Alexander G. Ororbia for their comments.

References

Carver, C. S. (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. *Personality and Social Psychology Review*, *9*(4), 312–333.

Croskerry, P. (2002). Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic Emergency Medicine*, *9*(11), 1184–1204.

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: a meta-analysis. *Psychological Bulletin*, *137*(1), 97.

Daniel Kahneman, A. T. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291.

Dittrich, M., & Leipold, K. (2014). Gender differences in time preferences. *Economics Letters*, *122*(3), 413–415.

Ghafurian, M., & Reitter, D. (2014). Impatience, risk propensity and rationality in timing games. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci), Quebec City, Canada* (p. 2841-2846).

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to kahneman and tversky. *Psychological Review*, *103*, 592–596.

Meertens, R. M., & Lion, R. (2008). Measuring an individual's tendency to take risks: The risk propensity scale. *Journal of Applied Social Psychology*, *38*(6), 1506–1520.

Olson, K., Camp, C., & Fuller, D. (1984). Curiosity and need for cognition. *Psychological Reports*, *54*(1), 71–74.

Powell, M., & Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. *Journal of Economic Psychology*, *18*(6), 605–628.

Schubert, R., Brown, M., Gysler, M., & Brachinger, H. W. (1999). Financial decision-making: are women really more risk-averse? *The American Economic Review*, *89*(2), 381–385.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Van Dijk, M., Juels, A., Oprea, A., & Rivest, R. L. (2013). Flipit: The game of stealthy takeover. *Journal of Cryptology*, *26*(4), 655–713.

Wittmann, M., & Paulus, M. P. (2008). Decision making, impulsivity and time perception. *Trends in Cognitive Sciences*, *12*(1), 7–12.

Wood, S. L., & Swait, J. (2002). Psychological indicators of innovation adoption: Cross-classification based on need for cognition and need for change. *Journal of Consumer Psychology*, *12*(1), 1–13.

Why Sense-Making through Magnitude May Be Harder for Fractions than for Whole Numbers

Eliane Stampfer Wiese (eliane.wiese@berkeley.edu)
Graduate School of Education, 4407 Tolman Hall
Berkeley, CA 94720 USA

Rony Patel (rbpatel@andrew.cmu.edu)
Kenneth R. Koedinger (krk@cs.cmu.edu)
Department of Psychology, 342c Baker Hall
Pittsburgh, PA 15213 USA

Abstract

What is the role of fraction magnitude knowledge in learning fraction addition? An experiment with 71 6th and 7th grade students compared fraction addition instruction and practice with a magnitude representation to a tightly controlled non-magnitude condition. In the magnitude condition, students with better fraction magnitude estimation skills benefitted more from the conceptual instruction and this relationship was moderated by students' knowledge of how magnitude relates to fraction addition and equivalence. However, students with better fraction magnitude estimation skills benefitted *less* from the practice problems with magnitude. In the non-magnitude condition, fraction magnitude estimation was not predictive of learning. This study indicates that students with magnitude knowledge can leverage it to learn fraction addition concepts from magnitude representations, but, for those students, magnitude representations may be a distraction from practicing the procedure.

Keywords: fraction addition; number line estimation; multiple representations

Fractions are Important but Difficult to Learn

Proficiency with fractions is a pivotal skill for students: fraction knowledge in fifth grade predicts math achievement in tenth grade even after statistically controlling for socioeconomic status, general IQ, and whole number arithmetic knowledge (Siegler et al., 2012). Yet, fractions are a difficult topic for students (Siegler et al., 2010) and teachers (Lee, Brown, & Orrill, 2011; Siegler et al., 2010). One obstacle to students' conceptual understanding is difficulty estimating a fraction's magnitude (Behr, Wachsmuth, & Post, 1985). In addition to improving conceptual understanding, fraction magnitude estimation, specifically accuracy on number line estimation, is correlated with accuracy on fraction arithmetic tasks (Siegler, Thompson, & Schneider, 2011).

One potential reason for this correlation is that students who have stronger fraction number line estimation knowledge can avoid incorrect strategies that result in magnitude incongruent answers. Adding 1/2 and 1/3, for example, must result in an answer greater than 1/2 and 1/3. Students who understand that the magnitude of the sum needs to be greater than the magnitude of both positive addends might reject an incorrect strategy that leads to a magnitude incongruent answer. A common incorrect strategy, adding the numerators and denominators, would lead to a sum of 2/5. Since the magnitude of the incorrect sum (2/5) is smaller than that of one of the addends (1/2) the strategy must be incorrect.

Fifth grade students don't always understand that adding positive fractions results in a sum with a magnitude that is larger than each individual addend (Wiese & Koedinger 2014) while sixth grade students almost always understand this concept (Siegler & Lortee-Forgues, 2015). If the correlation between accuracy of fraction number line estimation and accuracy on fraction arithmetic is at least partially driven by the reduction of magnitude incongruent procedural strategies, the correlation between accuracy of fraction number line estimation and accuracy on fraction arithmetic should be greater for those students who know how addition or subtraction affect magnitude versus those students who do not. That is, students cannot logically use their magnitude knowledge to reject magnitude incongruent answers if they do not know the sum is greater than the addends for positive fractions. Therefore, the understanding of how the magnitude of the addends is related to the sum should moderate the relationship between fraction magnitude knowledge and fraction addition knowledge.

Students also need equivalence magnitude knowledge in addition to fraction magnitude knowledge to help select amongst fraction arithmetic strategies. A student cannot use fraction magnitude knowledge to select amongst fraction addition strategies if she or he does not know that equivalent fractions have the same magnitude. For example, when adding 1/2 and 1/4, students need to find a common denominator (e.g., 2/4 plus 1/4). If students do not know that 1/2 and 2/4 are the same magnitude, it makes it difficult to compare the final sum (3/4) to the original addends. This means equivalence magnitude knowledge should moderate the relationship between fraction magnitude knowledge and fraction addition knowledge. The present study tests the two moderators of the relationship between fraction magnitude knowledge and fraction addition knowledge.

Prior Work: Brief Number Line Interventions

Existing literature supports the positive relationship between magnitude knowledge of whole numbers and whole number

arithmetic knowledge (Siegler & Mu, 2008), and the positive, causal influence of whole number magnitude knowledge on whole number arithmetic knowledge with a brief intervention (Booth & Siegler, 2008). Since magnitude-based interventions have been shown to improve students' proficiency whole number arithmetic, they may improve students' proficiency with fraction arithmetic also.

Magnitude Representations May Not Be Sufficient for Fraction Addition Skills

On the other hand, although magnitude training is a promising approach, it may not be sufficient for improving students' arithmetic skills or for their rejection of strategies that ignore magnitude. In one study, 5th grade students were asked to judge if fraction addition equations were correct or not. Half of the equations were correct, and half used the add-both-numerators-and-denominators strategy to yield a sum that was smaller than one of the addends. Even when shown the fraction symbols in the equation and their corresponding magnitudes, students performed around chance (Stampfer & Koedinger, 2013). This result indicates that magnitude was not sufficient for the students to determine the correctness of an addition equation. To reject a strategy that ignores magnitude, students must realize not only that fractions have magnitudes, but also that fractions follow the same rules of addition as whole numbers, namely, that adding two positive fractions yields a sum larger than each addend. A follow-up study confirmed that those 5th grade students had difficulty applying that addend-sum relationship to fractions (Wiese & Koedinger, 2014).

Finally, though choosing an incorrect addition procedure reveals gaps in conceptual knowledge, it also reveals gaps in procedural knowledge. Students may make procedural errors on fraction addition items even though they have high levels of conceptual knowledge. Indeed, Byrnes & Wasik (1991) found, in a an experiment with middle-schoolers, that reinforcing fraction magnitude concepts and explaining the addition procedure did not improve performance above simply executing the procedure. However, Byrnes and Wasik used discrete rectangles instead of a continuous number line. Since there are theoretical reasons a number line specifically might be more conducive to thinking about magnitude (e.g., Siegler et al., 2011) and discrete rectangles might elicit counting behavior as opposed to thinking about magnitude holistically, the present study uses a continuous representation of magnitude.

Study: Fraction Addition Instruction With and Without a Representation of Magnitude

The present study will compare two forms of instruction, one with a magnitude representation and one without. We hypothesize that knowledge and application of the addend-sum relationship (that two positive addends yield a sum larger than each) moderate the relationship between magnitude estimation of fractions and fraction addition skills. Additionally, we propose conflicting hypotheses for the question 'Does a procedure-and-magnitude intervention demonstrate superior learning gains than a procedure-only control?' Magnitude approaches would suggest yes, as considering magnitude should help students reject incorrect answers, while cognitive load theory would suggest no, as the magnitude representations require processing that is extraneous to learning the procedure.

Participants and Assessment Materials

78 6th and 7th graders from two middle schools participated in the study. The analyses include only the 71 students who completed all portions of the study. Within-class random assignment yielded 38 in the procedure-and-magnitude intervention and 33 students in the procedure-only control.

Three assessments were given: before the instruction (pretest), after the instruction (midtest) and after the practice problems (posttest). Three assessment forms were created and students were randomly assigned to one of the six counter-balanced orders. The overall reliabilities (Chronbach's Alpha) for the test forms were .77 or greater.

Fraction Number Line Estimation items presented a target fraction to the left of a number line, which was labeled with a '0' on the left endpoint and a '1' on the right endpoint. Participants saw 10 of these problems, only at pretest.

Students saw four Fraction Addition problems. Two had same denominators while two had unlike denominators. For the addition items, the reliabilities (Chronbach's Alpha) for the three test forms ranged from .72 to .75. The remaining items (described below) assessed how students connected magnitude knowledge to a fraction addition context.

For magnitude knowledge to be beneficial for fraction addition, students will likely need to understand how magnitude relates to addition. Direction of Effects items asked if the addition of two positive fractions is greater than one of those fractions alone (e.g., 1/2 + 1/3 > 1/3, with answer options True, False, and Can't tell from the information given). Addend-Sum items asked if the sum was greater than each of the addends (e.g., showing students 1/2 + 1/3 = 5/6 with a note that the addition is correct, then asking if 5/6 > 1/2 and if 5/6 > 1/3, with the answer options True, False, and Can't tell from the information given).

For linear depictions of magnitude to be beneficial, students will likely need to be able to interpret them correctly, including recognizing that equivalent fractions have the same magnitudes. Fraction Equivalence items presented a number line with a fraction plotted on it. Students were provided with an equivalent fraction (and the statement that the two fractions were equivalent) and asked to plot the equivalent fraction on the same number line. They were then asked if the plotted fraction belonged to the left, right, or in the same spot as the presented fraction with a fill-in-the-blank multiple-choice question. For example, 2/8 belongs ____ 1/4, with the options "to the left of," "in the exact same spot as," and "to the right of."

Instructional Materials

Students were given initial instruction and then practice problems. First, all students saw an example whole number

addition equation, and an explanation that adding whole numbers yields the total of their sizes. A similar example of a fraction addition equation followed, showing that the independent whole number strategy yields an incorrect sum. Accompanying text explained that the correct sum for fraction addition would be the same magnitude as the combined magnitudes of the addends. The instruction included brief multiple-choice and free-response questions to check that students had read the text. This instruction was intended to shake students' confidence in the incorrect whole number strategy, and did not demonstrate the full fraction addition procedure. Students in the procedure-and-magnitude intervention saw depictions of the magnitudes of the addends and sums in the example equations. This depiction was intended to reinforce that a correct sum would equal the combined magnitudes of the addends, while an incorrect sum would not. These depictions were linear magnitude representations, but were not presented on a number line. The procedure-only control did not include depictions of magnitude (see Figure 1). After completing that initial instruction, students did the mid-test.

After the mid-test, students did scaffolded practice problems with correctness feedback for each step, alternating with non-scaffolded problems (Figure 2), which were similar to the fraction addition assessment items. After entering an answer for the non-scaffolded problems, students were given correctness feedback. If the answer was wrong, students were shown a worked example of that problem and were asked to fix their answer. The worked examples in the procedure-and-magnitude intervention included a linear magnitude depiction of the two original addends, the student's incorrect answer, and the correct answer. Students in the procedure-only control did the same tasks, but without the depictions of magnitude.

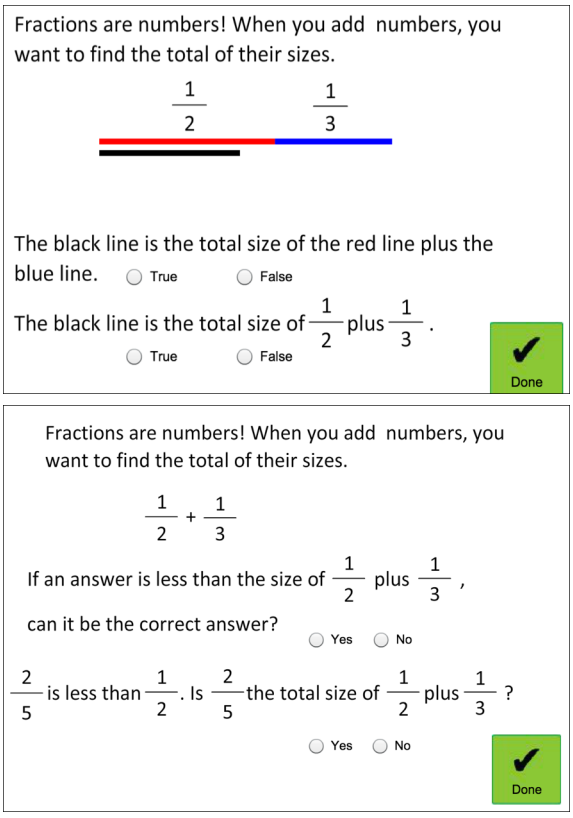


Figure 1: Matched instruction with magnitude (top) and without it (bottom), question four of six. Both formats try to convey that a sum cannot be smaller than the combined magnitudes of its addends.

Method

All study activities took place during normal class time. Students were given 2 minutes for the number-line estimation questions, 5 for the pre-test and post-test, and 25 minutes for the instruction, mid-test, and practice problems, which students did at their own pace. All activities used an online math tutoring system. While the instruction between the pre-test and mid-test was brief, it targeted only one concept: that a correct addition strategy will yield a sum that is equal in magnitude to the two addends together. After the mid-test, students practiced fraction addition with scaffolding, interleaved with worked examples. This study was timed for a standard 40-minute class period. Though students would likely have improved more over a longer study duration, previous work has demonstrated learning with brief interventions that include magnitude (Lomas, Ching, Stampfer, Sandoval, & Koedinger, 2011; Opfer & Thompson, 2008; Siegler & Ramani, 2009).

Results

For the fraction addition items, students rated their answers with a confidence score from 1 to 5. The answer score (correct or incorrect) was combined with the self-reported accuracy to yield a scale from 1 to 10 (from highest confidence in an incorrect answer to highest confidence in a correct answer). This measure is referred to below as 'accuracy' (Note: Similar patterns of results were obtained using accuracy without confidence). Students' responses spanned the entire scale. Table 2 shows means and standard deviations for fraction addition accuracy, and mean scores for direction of effects problems, addend sum problems, and the multiple-choice portion of the fraction equivalence problems. For the latter three items, since there was only one question per student per type of problem, the distributions are binomial and can be expressed in a single parameter.

One striking finding from this study was students' difficulty in progressing through the initial instruction (between the pre-test and mid-test), which was intended to be straightforward. This initial instruction presented six problems, starting with whole number addition and moving to fraction addition, all focused on the idea that the sum of two numbers is the total of their magnitude. In particular, the error rate for question four (Figure 1) reveals gaps in students' reasoning around fraction addition. In the procedure-only control, the majority of students (85%) answered at least one question incorrectly, even though the

Table 2: Assessment Scores

Fraction Addition Accuracy Means (and Std. Deviations)			
Condition	Pretest	Midtest	Posttest
Procedure-only	5.22 (2.81)	5.45 (2.84)	6.48 (2.30)
Procedure & Magnitude	5.14 (2.58)	5.26 (2.61)	6.35 (2.64)
Direction of Effects Mean Scores			
Condition	Pretest	Midtest	Posttest
Procedure-only	45%	61%	55%
Procedure & Magnitude	42%	38%	42%
Addend-Sum Mean Scores			
Condition	Pretest	Midtest	Posttest
Procedure-only	56%	58%	56%
Procedure & Magnitude	55%	54%	57%
Equivalence Mean Scores			
Condition	Pretest	Midtest	Posttest
Procedure-only	21%	21%	13%
Procedure & Magnitude	29%	30%	27%

first question explicitly states that the proposed sum is less than combined sizes of the two addends, and the second question explicitly states that the proposed sum is less than the size of one of the addends alone. While the matched questions were easier in the procedure-and-magnitude intervention (which showed the magnitude of each fraction graphically), students were still far from ceiling (34% had at least one incorrect response). One hypothesis for why students benefit from magnitude knowledge is that it helps them reject magnitude-incongruent answers. That hypothesis suggests that knowledge of magnitude helps students arrive at the correct evaluation of whether or not a sum is incongruent. However, this information was explicitly provided in the procedure-only control, and students' error rates show that version of the question was harder. These results suggest that it is the underlying addition principles that are hard for students to make sense of in a fraction addition context. The hypothesis that magnitude knowledge helps students by making it easier for them to apply principles of addition presupposes that students understand the principles of addition in the first place. Rather, these results suggest the reverse: depictions of magnitude may help students make sense of abstract principles of addition.

Analyses

Across both conditions, a pairwise t-test revealed no difference between pretest fraction addition scores and mid-test fraction addition scores (t = -0.46, df = 68, p = 0.65). In a linear regression, pretest addition scores and condition were predictive of mid-test scores ($F(2,66)$ = 23.29, R^2 = 41.37%, $p < 0.01$). While pretest fraction addition accuracy was predictive of mid-test fraction addition accuracy (B = 0.65, t = 6.814, $p < 0.01$), condition was not predictive (B = -0.20, t = -0.39, p = 0.70).

Within each condition, paired t-tests with pretest fraction addition accuracy and mid-test fraction addition accuracy

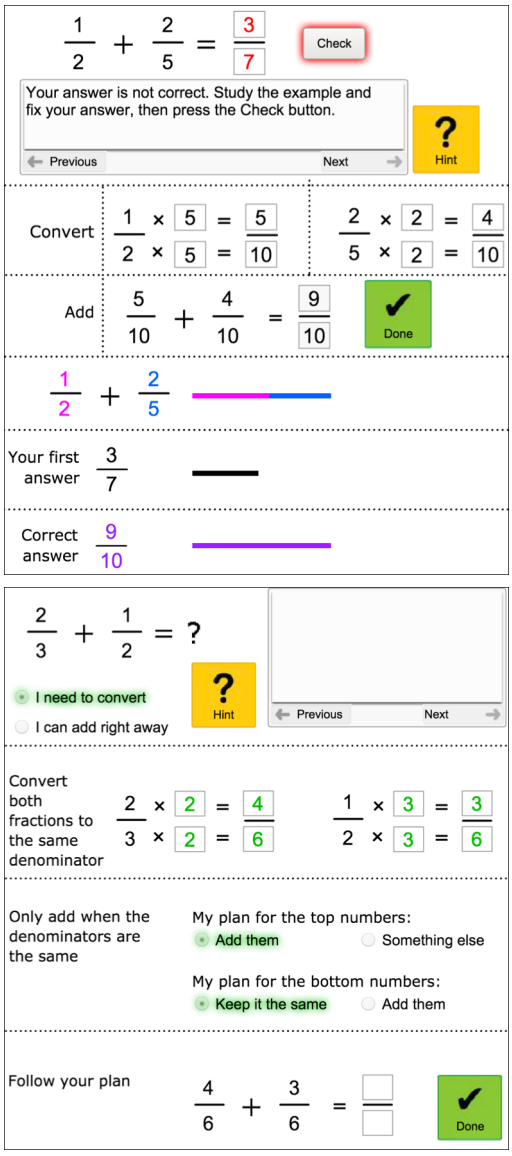


Figure 2: A non-scaffolded problem (top) and scaffolded problem (bottom). Non-scaffolded items initially show only the addition problem. If the student answers incorrectly, the worked example is revealed. Scaffolded items unfold step by step, and are the same in both conditions.

reveal no significant improvement (magnitude: t = -0.07, df = 36, p = 0.94; control: t = -0.75, df = 31, p = 0.46). Within each condition, linear regressions predicted mid-test fraction addition accuracy with pretest fraction addition accuracy and Percent Average Error (PAE) for fraction number line estimation (magnitude: $F(2,34)$ = 12.63, R^2 = 42.62%, $p < 0.01$; control: $F(2,29)$ = 27.68, R^2 = 65.62%, $p < 0.01$). In the procedure-and-magnitude invervention, both pretest fraction addition accuracy (B = 0.29, t = 2.01, p = 0.05) and PAE for fraction number line estimation (B = -11.01, t = -3.47, $p < 0.01$) were significant predictors of mid-test fraction addition accuracy. Note that a negative predictor for PAE means low error (i.e., high accuracy) in number line estimation is associated with higher scores on the fraction

addition accuracy scale at mid-test. Since there was no learning in fraction addition from pre-test to mid-test and PAE was a significant predictor, it seems that students got better at fraction addition if they had strong magnitude knowledge, but got worse if they had weak magnitude knowledge. For the procedure-only control, pretest fraction addition accuracy was a significant predictor of mid-test fraction addition accuracy (B = 0.81, t = 7.33, $p < 0.01$), but PAE was not (B = -1.04, t = -0.45, p = 0.66). The two groups seem to have learned in different ways: students' number line magnitude knowledge influenced their learning for the procedure-and-magnitude intervention, but not for the procedure-only control.

Three separate regressions on the procedure-and-magnitude intervention tested moderators of fraction magnitude knowledge and fraction addition learning from pre-test to mid-test: scores on direction of effects, equivalence, and addend-sum items. Separate regressions were run since the moderators were correlated, violating the independent and identical distribution assumption (i.i.d.) of regression.

Moderators: PAE and Direction of Effects This regression included pretest fraction addition accuracy, PAE fraction number line estimation, pretest direction of effects accuracy, and the interaction between PAE and pretest direction of effects accuracy (F(4,32) = 9.414, R^2 = 54.06%, $p < 0.01$). In this model, pretest fraction addition accuracy is not predictive of mid-test fraction addition accuracy (B = 0.19, t = 1.40, p = 0.17), but PAE (B = -8.82, t = -2.83, $p <$ 0.01), pretest direction of effects score (B = 1.43, t = 2.070, p = 0.05), and the interaction of the two (B = -0.60, t = -1.66, p = 0.10) are predictive. Thus, for fraction addition accuracy, students with low magnitude error and high direction of effect accuracy benefited the most from the procedure-and-magnitude intervention.

Moderators: PAE and Equivalence Knowledge The second regression uses pretest fraction addition accuracy, PAE fraction number line estimation, pretest equivalence knowledge, and the interaction between PAE and pretest equivalence knowledge to predict mid-test fraction addition accuracy (F(4,32) = 8.996, R^2 = 52.93%, $p < 0.01$). Pretest fraction addition accuracy (B = 0.12, t = 0.84, p = 0.41) and pretest equivalence knowledge (B = -1.17, t = -0.91, p = 0.37) were not predictive of mid-test fraction addition accuracy, but PAE (B = -14.81, t = -3.51, $p < 0$ .01) and the interaction between PAE and pretest equivalence knowledge (B = -1.48, t = -2.32, p = 0.03) are predictive of mid-test fraction addition accuracy. Thus, for fraction addition accuracy, students with low magnitude error and high equivalence accuracy benefited most from the intervention.

Moderators: PAE and Addend-Sum The final regression included pretest fraction addition accuracy, PAE fraction number line estimation, addend sum knowledge questions, and the interaction between PAE and addend sum knowledge to predict mid-test fraction addition accuracy (F(4,31) = 6.097, R^2 = 44.03%, $p < 0.01$. Pretest fraction addition accuracy (B = 0.30, t = -3.10, p = 0.06) and PAE (B = -10.49, t = -3.104, $p < 0.01$) were both predictive of mid-test fraction addition accuracy, but addend sum knowledge (B = 0.00, t = 0.00, $p > 0.99$) and the interaction between PAE and addend sum knowledge (B = -0.36, t = -0.97, p = 0.34) were not predictive of mid-test fraction addition accuracy.

Midtest to Posttest Learning Across both conditions, a pairwise t-test revealed improvement in fraction addition scores from mid-test to post-test (t = -3.81, df = 66, $p <$ 0.01; means were 5.35 at mid-test vs 6.41 at post-test). In a linear regression, mid-test addition scores and condition were predictive of post-test scores (F(2,64) = 26.38, R^2 = 45.19%, $p < 0.01$). While mid-test fraction addition accuracy was predictive of post-test fraction addition accuracy (B = 0.62, t = 7.258, $p < 0.01$), condition was not predictive (B = -0.11, t = -0.24, p = 0.81).

Midtest to Posttest Learning and PAE For both conditions individually, paired t-tests on mid-test and post-test fraction addition accuracy show learning (magnitude: t = -2.92, df = 34, $p < 0.01$; control: t = -2.48, df = 31, p = 0.02). In a linear regression with the procedure-and-magnitude invervention only, post-test fraction addition accuracy was predicted using mid-test fraction addition accuracy and PAE for fraction number line estimation (F(2,32) = 24.05, R^2 = 60.05%, $p < 0.01$). Both mid-test fraction addition accuracy (B = 0.92, t = 6.54, $p < 0.01$) and PAE (B = 6.37, t = 2.01, $p < 0.01$) were significant predictors of post-test fraction addition accuracy. Note that a positive predictor for PAE here means low error (or high accuracy) in number line estimation is associated with *lower* scores on the post-test fraction addition accuracy measure. When running the same regression with the procedure-only control (F(2,29) = 8.49, R^2 = 36.92%, $p < 0.01$), mid-test fraction addition accuracy was a significant predictor post-test fraction addition accuracy (B = 0.48, t = 3.95, $p < 0.01$), but PAE was not (B = -1.56, t = -0.62, p = 0.54). Note: We tested the same moderators we used from pretest to mid-test, but none were significant.

Discussion

What is the role of fraction magnitude knowledge in learning fraction arithmetic? Within the procedure-and-magnitude intervention, when students were learning how magnitude knowledge can be used to justify correct and incorrect answers (between pretest and mid-test), students with greater initial magnitude knowledge benefited more than students with lower initial magnitude knowledge. However, when solving the problems and studying the worked examples (between mid-test and post-test), students with greater initial magnitude knowledge benefitted *less* than their counterparts. Perhaps the magnitude knowledge that was initially helpful for learning the general concept of fraction addition became a distractor to learning the

procedural steps. Perhaps the combination of learning the right steps to solving a fraction addition problem *and* seeing how the magnitudes of their answers compared to the magnitude of the actual answer exceeded the desirable difficulty. Trying to learn both pieces of information at once (each of which can be considered complex) may have caused excessive cognitive load and hurt their performance. Those with stronger magnitude knowledge might have been more distracted by the magnitude stimuli if they were attending to it more, while those with weaker magnitude knowledge might have attended to the magnitude stimuli less, resulting in more focus in the steps, and thus, better performance.

Also of note is the lack of differences between the groups in terms of overall learning. While the mechanisms by which the students learned might have been different, both groups overall did not improve from pretest to midtest and learned equally from midtest to posttest. This indicates the benefits of learning by doing: overall, each condition only demonstrated learning from the activities that gave students practice solving the targeted problems. The results from this experiment do not provide evidence supporting the hypothesis that the inclusion of magnitude in a brief fraction addition instruction benefits students' learning. However, the role of magnitude knowledge in the procedure-and-magnitude interention does provide evidence for the theoretical role of magnitude in fraction addition learning: explicit representations of fraction magnitude can help students learn fraction concepts but may detract from learning the procedures.

Finally, the initial instruction (between pre-test and mid-test) posed surprising challenges for the students. Students demonstrated difficulty in making logical inferences involving fraction addition. For example, 85% of students in the procedure-only control thought that an answer that was less than the total size 1/2 and 1/3 could be the correct answer to 1/2 + 1/3. While the correct inference was easier to make in the procedure-and-magnitude intervention, where the magnitudes of the addends and proposed sum were provided, 34% of students in that condition still made errors. Results from students' interactions with the initial instruction indicate that students did not have a solid, context-general foundation with the principles of addition.

It is important to consider how this type of instruction might generalize to other fraction arithmetic operations. While there is evidence for the mechanistic use of magnitude knowledge, in this study there was little practical benefit of the procedure-and-magnitude intervention compared to the control for learning the fraction addition procedure.

Acknowledgements

This work is supported in part by Carnegie Mellon University's Program in Interdisciplinary Education Research (PIER) funded by Grant R305B090023 from the U.S. Department of Education, and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C100024 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Behr, M. J., Wachsmuth, I., & Post. (1985). Construct a Sum: A Measure of Childrens Understanding of Fraction Size. *Journal for Research in Mathematics Education*, *16*(2), 120–131.

Booth, J. L., & Siegler, R. S. (2008). Numerical Magnitude Representations Influence Arithmetic Learning. *Child Development*, *79*(4), 1016–1031.

Byrnes, J. P., & Wasik, B. A. (1991). Role of Conceptual Knowledge in Mathematical Procedural Learning. *Developmental Psychology*, *27*(5), 777–786.

Lee, S. J., Brown, R. E., & Orrill, C. H. (2011). Mathematics Teachers' Reasoning About Fractions and Decimals Using Drawn Representations. *Mathematical Thinking and Learning*, *13*(3), 198–220.

Opfer, J. E., & Thompson, C. a. (2008). The trouble with transfer: insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*(3), 788–804.

Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., … Wray, J. (2010). *Developing Effective Fractions Instruction for Kindergarten through 8th Grade. IES Practice Guide. NCEE 2010-4039. What Works Clearinghouse*. Retrieved from http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED512043

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., … Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, *23*(7), 691–7.

Siegler, R. S., & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science : A Journal of the American Psychological Society / APS*, *19*(8), 759–63.

Siegler, R. S., & Ramani, G. B. (2009). Playing Linear Number Board Games – But Not Circular Ones – Improves Low-Income Preschoolers' Numerical Understanding. *Journal of Educational Psychology*, *101*(3), 545–560.

Siegler, R. S., Thompson, C. a, & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–96.

Stampfer, E., & Koedinger, K. R. (2013). When seeing isn't believing: Influences of prior conceptions and misconceptions. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1384–1389). Berlin, Germany: Cognitive Science Society.

Wiese, E. S., & Koedinger, K. R. (2014). Investigating scaffolds for sense making in fraction addition and comparison. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1515–1520). Quebec City, Canada: Cognitive Science Society.

The Primary and Convergent Retrieval Model of Recall

William J. Hopper (whopper@psych.umass.edu)
Department of Psychological and Brain Sciences, 135 Hicks Way
Amherst, MA 01003 USA

David E. Huber (dehuber@psych.umass.edu)
Department of Psychological and Brain Sciences, 135 Hicks Way
Amherst, MA 01003 USA

Abstract

Memory models typically assume that recall is a two-stage process with learning affecting both processes to the same degree. This equal learning assumption is difficult to reconcile with studies of the 'testing effect', which reveal different forgetting rates following learning from test practice versus learning from restudy. Here we present a new memory model, termed Primary and Convergent Retrieval (PCR) that assumes successful recall leads to a selective enhancement for the second stage of recall (Convergent Retrieval). We applied this model to existing testing effect data. In two new experiments, we confirmed novel predictions of the PCR model for transfer between retrieval cues and for recall latencies. This is the first formally specified model of the testing effect and it has broad implications for the nature of learning and retrieval.

Keywords: Memory Modeling; The Testing Effect, Retrieval Practice

Two-Stage Models of Recall

As briefly reviewed here, memory models typically assume that recall is a two-stage process. For instance, the Search of Associative Memory model (SAM, Raaijmakers & Shiffrin, 1981) contains a *sampling* stage that selects a specific memory trace from a pool of active traces, followed by a *recovery* stage that extracts the details of the sampled memory. Similarly, the MINERVA 2 model (Hintzman, 1984) differentiates between an *intensity* response (measuring overall activation) that is used to weight the contribution of memories to the *echo content*, which is subsequently 'cleaned up' through a recursive process to produce the desired content. Norman and O'Reilly (2003) assumed these separate processes reflect the operations of different brain regions, with parahippocampal cortex providing a *familiarity* response, such as used in recognition, whereas recall requires *pattern completion* that depends on the actions of the hippocampus.

These models, and others, include a stage that isolates relevant memories based on a scalar value for retrieval strength followed by a stage in which sufficient detail is extracted to produce on overt recall response. However, these models also assume that learning is passive and any opportunity to encode a memory will affect both stages in a similar manner. Additionally, the time course of this second stage is not specified by extant memory models. Here, we present a new memory model which specifies the learning processes and time course of this second stage, motivated in large part by studies of retrieval practice effects.

Restudy, Test Practice, and Forgetting Rates

The assumption of passive learning appears at odds with testing effect studies that indicate greater learning from retrieval practice (see Roediger & Karpicke, 2006a for a review). In these studies, participants learn some new material, after which the material is practiced by restudying, or by taking a practice test (either with or without feedback). A retention interval follows this practice, after which participants take a final test. The final test often reveals an advantage for material practiced with a test relative to material practiced by restudying, and this advantage grows with retention interval. In other words, these two types of practice produce different forgetting rates. A striking example of different forgetting rates is found when test practice occurs without feedback. In this case, if the retention interval is short (e.g., 5 minutes), restudying produces higher accuracy than a practice test. However, if the retention interval is longer (e.g., 24 hours), this relationship is reversed and test practice produces better accuracy than restudying (Roediger & Karpicke, 2006b; Toppino & Cohen, 2009; Wheeler, Ewers, & Buonanno, 2003).

This crossover interaction comparing restudy to test practice without feedback is partially explained by realizing that there is no opportunity for additional learning for the items that fail to be retrieved during test practice. Thus, the test practice reflects a bifurcated distribution (Kornell, Bjork, & Garcia, 2011). However, this account still assumes greater learning for the items that were recalled during test practice as compared to the learning from restudy and yet no explanation is provided as to *why* this is the case. Furthermore, the crossover interaction occurs even when considering recallable items, as determined by an initial test prior to subsequent test practice or restudy (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). To explain why the act of recall produces qualitatively different learning than passive restudy, we developed the Primary and Convergent Retrieval (PCR) model of recall.

The Primary and Convergent Retrieval Model

The PCR model retains the two-stage recall architecture of previous memory models, with the first stage termed *Primary Retrieval* (PR) while the second is termed *Convergent Retrieval* (CR). Primary Retrieval describes the initial process of activating all relevant memories based on their

associations with the current retrieval cues (e.g., the current temporal context and any item information such as a word or picture provided as a cue). However, even the most active memory in PR may be incomplete (e.g., some of the features of the memory remain inactive). Convergent Retrieval describes a second stage in which the memory system focuses on the most active memory and attempts to activate any inactive features through intra-item associations between the features of that memory.

General Assumptions of PCR

PCR assumes that all memory traces and retrieval cues are composed of a finite number of features. New unidirectional associations between features are formed according to the temporal sequence of events. For instance, if feature A becomes active at time *t*, and feature B becomes active at time *t*+1, an excitatory connection is formed from feature A to feature B. This form of learning is consistent with spike time dependent plasticity in which potentiation of a synapse only occurs if the pre-synaptic cell fires before the post-synaptic cell (for examples see Dan & Poo, 2006). We assume this learning rule applies in all situations. Thus, context features become associated with item features during study because context is active prior to presentation of the item. More importantly, in terms of explaining the testing effect, initially active features of an item become associated with initially inactive features of the same item, provided that those features are subsequently retrieved. This intra-item learning explains the extra benefit of retrieval practice.

Primary Retrieval

When a retrieval attempt is initiated, all currently active features (e.g., the current temporal context and any retrieval cues) serve as cues to activate features of memory traces. Thus, PR is cue-dependent, meaning that the content and magnitude of the memory system's response depends completely on the features of the retrieval cues. During initial study, the retrieval cues (context and an item presented as a cue) are typically active first, followed by the target item, allowing associations between these retrieval cues and the features of the target item. However, encoding is likely to be incomplete and error prone. Furthermore, the temporal context will naturally change between study and test (e.g., Howard & Kahana, 2002). Thus, the activation of the target features that occurs with PR will be incomplete. Because a naming response requires full retrieval of the item, a pattern completion process is needed for recall success. We term this process 'convergent retrieval'.

Convergent Retrieval

Convergent Retrieval is the process by which initially dormant features that were missed by PR become active, via intra-item excitatory connections between the individual features that define the target item. Even for readily known items (e.g., high frequency words), the retrieval cues may fail to activate enough of the item, such as occurs with 'tip-of-the-tongue'. If CR succeeds in activating all the remaining features, two things occur: 1) the item can then be recalled; and 2) new associations between the features activated by PR and the features subsequently activated by CR can be formed (note that this also holds true when the target item is provided after CR failure, such as occurs in test practice with feedback; this explains why the testing effect is more powerful with feedback).

This second outcome, called intra-item learning, represents a theoretical departure from most memory models, which do not explicitly model the learning between features of an item. This intra-item learning makes it easier to recall the item *regardless of the initial state of activation that occurs with PR*; because the associations between the individual features that compose an item are a property of the item, rather than the association between the item and retrieval cues, the intra-item learning that takes place following successful test practice benefits recall in a cue-independent manner. This intra-item learning reduces the numbers of steps required for CR, resulting in faster recall. Thus, even in situations where intra-item learning fails to increase the probability of recall, it will decrease retrieval latency for the items that are recalled.

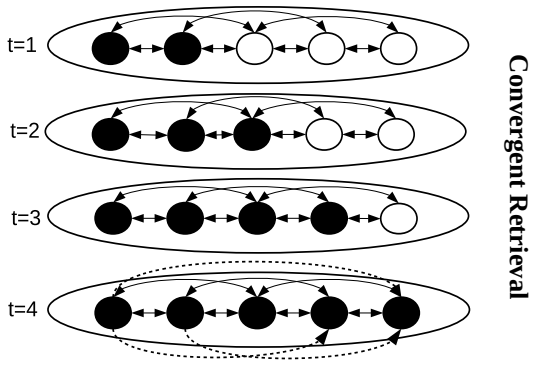


Figure 1: An example of successful Convergent Retrieval. The features initially activated by retrieval cues during Primary Retrieval activate the remaining inactive features according to the associative connections between each feature. Following convergence, additional associations are learned between the initially active features, and the subsequently activated features.

Figure 1 shows an example of the CR process, beginning where PR ends. In this example, retrieval cue features only activate two of the five features of an item in memory (features of an item are shown as circles inside an oval, with currently active features represented by the filled circles; the existing intra-tem associations are indicated by the solid arrows). In this example, if a feature requires two excitatory inputs to become active, all of the initially dormant features will eventually become active across three time steps. New intra-item associations are now formed according to this progression of events (represented by the dashed arrows).

This discussion outlines the guiding principles behind the PCR model. However, a full-scale neural network

implementation would require many auxiliary assumptions (e.g., the exact time function underlying learning, the exact rule for feature activation during CR, the nature of temporal context change, etc.). Next we present a simple abstract mathematical model that approximates these guiding principles, noting that a fully mechanistic instantiation of these principles might deviate somewhat from this simple model.

A Binomial Instantiation of PCR

We assume each item consists of a finite number of features (set to 100 for convenience) and that each item requires a specific number of active features for CR to be successful (features are discretely active or inactive). This captures pre-experimental item differences in which some items are more easily recalled even if encoding is incomplete or context has changed greatly (i.e., even with weak PR). The value of each item's CR 'threshold', θ_i, is sampled from a binomial distribution with probability *t* and 100 counts (one count for each feature). We use the symbol $B(\,)$ to indicate a binomial distribution.

$$\theta_i \sim B(t, 100)$$

We assume that initial encoding is incomplete, or prone to errors such that even an immediate final test (i.e., one for which context has not changed) fails to activate all of the target item's features. As with item differences in threshold, we assume some items are better encoded than others, and the number of features that are encoded for each item, α_i, is also sampled from a binomial distribution, but in this case the probability parameter is *e*.

$$\alpha_i \sim B(e, 100)$$

The parameter *e* may be thought of as an encoding rate parameter, and more time spent studying will result in higher values of *e*.

Restudy and *successful* recall provide another opportunity for encoding item features by associating them with retrieval cues and both forms of practice increase the value of α_i (there is no learning for unsuccessful recall without feedback). This learning during the practice phase of an experiment is again captured with a binomial sample, but in this case the probability parameter is *l*. However, unlike initial study, this learning only applies to features that were not originally encoded and so the number of counts for this binomial distribution is 100 - α_i, resulting in the following expression for change in the number of encoded features.

$$\Delta\alpha_i \sim B(l, 100 - \alpha_i)$$

Forgetting is implemented in the model by reducing the value of α_i. Modeling forgetting by reducing the number of features activated by retrieval cues corresponds to the assumption that temporal context shifts during the retention interval such that PR only activates a subset of the item features that were previously encoded in relation to the context at the time of initial study. The reduction in features activated during PR is captured by a binomial sample with probability *f* as follows.

$$\Delta\alpha_i = -B(f, \alpha_i)$$

As outlined previously, CR success produces new intra-item learning, as the features of an item become associated with each other because they were activated in a progressive manner. We capture intra-item through a reduction in the CR threshold θ_i. This change in threshold is again a binomial sample, but with probability parameter *r* and the number of counts equal to the current threshold.

$$\Delta\theta_i = -B(r, \theta_i)$$

This reduction makes items more easily recalled regardless of how they are cued (i.e., even when PR is weak). In simulations with the model, recall success for each item is discretely determined according to a comparison between α_i and θ_i. If $\alpha_i > \theta_i$, then PR has activated more features than are required for CR success, and thus the entire content of the item is retrieved and available to be named in a recall response. However, as described next, a key component of the recall process is the order in which items are considered for CR and the time that it takes to attempt CR.

In a cued recall test, it is likely that the target memory is the only memory that has any appreciable activation (although this is not true in cued recall experiments that pair the same cue with multiple targets). In the case where the target is the only active memory, the important question is whether that memory is sufficiently active to support CR success and whether CR can be achieved in the allowed time. As outlined in Figure 1, we assume that CR takes time as dormant features progressively activate. We implement this by assuming that the time needed for CR relates to the distance from threshold, α_i - θ_i. More specifically, the Reaction Time (RT) to recall (or fail to recall) target item *i* is a negative exponential function.

$$RT_i = T_{min} + (T_{max} - T_{min})\left(e^{-\lambda|\alpha_i - \theta_i|}\right)$$

Here, T_{max} and T_{min} are upper and lower bounds on possible response times, respectively. Importantly, the term in the exponent uses the absolute magnitude of the difference. This captures the intuition that an item that is on the 'tip-of-the-tongue' is one that will take a long time before it is recalled or before the memory system admits defeat. In contrast, items that are far above threshold are recalled very quickly. Similarly, items that are far below threshold fail to progress in the CR process, and the retrieval attempt is quickly abandoned.

In a free recall test, the order in which memories are considered for CR plays a crucial role. More specifically, a considered item may fail to be recalled owing to CR failure but another possibility is that the item was not recalled because CR was never attempted. We assume that the memory system does not directly know whether CR success is possible for each memory; this knowledge requires the actual engagement of the CR process. However, the amount of feature activation (i.e., the amount of PR) is a good proxy for CR success. In other words, it is more likely that $\alpha_i > \theta_i$ when selecting for items with high α_i. Thus, rather than random sampling as in the SAM model, we assume that CR

is attempted for each item in descending rank order of α_i. In this way, interference occurs in the construction of the rank ordered list; strong memories will be near the top of the list, making it more likely that a test taker runs out of time before considering items farther down the list.

The predicted probability of recall in a free recall experiment is determined as follows. Monte Carlo simulations generate 1,000 hypothetical lists of items in which each item has an initial threshold θ_i, and α_i value owing to initial study, followed by changes to these values with restudy or test practice. Simulating final recall, CR is attempted for each item in the rank ordered list of α_i values for each list, while keeping track of the total time elapsed during the recall session until the allowed time is exceeded or the entire list has been considered for CR.

Applying PCR to Roediger & Karpicke 2006b

Experiment 1 from Roediger and Karpicke 2006b manipulated practice method (test practice without feedback vs. restudy) and retention interval (5 minutes, 2 days, and 1 week), using free recall for both test practice and the final test. Participants were given two prose passages that each contained 30 'idea units'. Participants then took a practice test on one passage, and restudied the other. After one of the 3 retention intervals, participants took a final test on both passages. Figure 2 shows recall accuracy in each condition, as well as mean accuracy on the practice test.

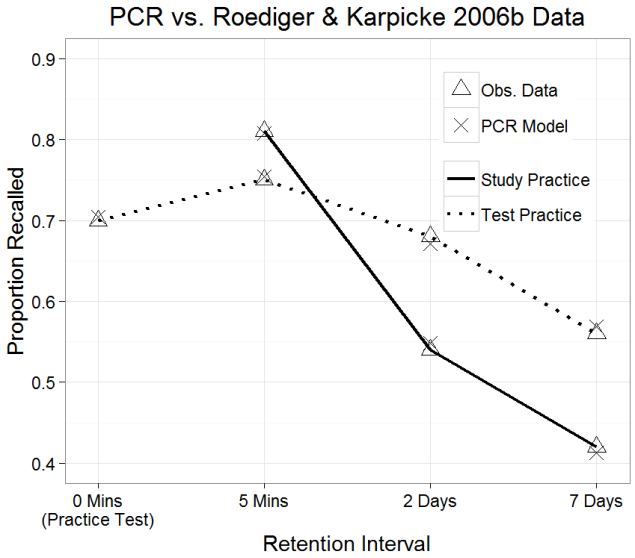


Figure 2: Recall accuracy from Roediger and Karpicke 2006b shown with the accuracy predicted by PCR.

We fit the PCR model to the average free recall accuracies reported by Roediger and Karpicke 2006b (shown in Figure 2). The e, l, r, T_{min}, f_2, and f_7 parameters were allowed to freely vary, while T_{max} was fixed at 60 seconds. Each item's CR threshold θ_i was drawn from a Binomial distribution with probability parameter t fixed at .5. We assumed that no forgetting occurred in the immediate final test condition (i.e., $f = 0$). The likelihood of the data was maximized using the binomial likelihood ratio test, which provides a chi-square goodness-of-fit statistic. The low value of this statistic indicates that the model is not rejected ($\chi^2(1) = .042$, $p = .83$). The predicted accuracy of the best fitting model is shown along with the observed data in Figure 1.

One the one hand, the accuracy of this fit is not surprising considering that 6 free parameters were used to fit 7 conditions. On the other hand, this crossover interaction between type of practice and retention interval is theoretically challenging and these results are problematic for any memory model, regardless of the number of free parameters. Ultimately, the fit to these data can be considered a "proof of concept" that the PCR model is able to explain the pattern of recall accuracy results observed in studies of the testing effect. In the following sections, we test novel predictions of the PCR model in a data set which imposes a much greater amount of constraint.

Hidden Benefits of Retrieval

The PCR model includes the notion of a bifurcated distribution (e.g., learning only applies to items that were recalled during test practice). The notion of a bifurcated distribution suggests that a great deal of learning has occurred for the items that were recalled on the practice test, but does not specify the nature of this extra learning for tested items. The PCR model attributes this extra learning to increased associations between the individual features of a target item. The incremental nature of CR during test practice promotes this type of learning whereas this does not occur with restudy considering that all of the item features are presented simultaneously with restudy. However, in the absence of feedback during test practice, this extra learning is not apparent *when examining accuracy* on an immediate final test because it only occurs for the items that would have been recalled *even if* they hadn't received test practice. Thus, the effects of this additional intra-item learning are masked in the short term and only emerge after a delay period, where this extra learning serves as additional protection against interference and forgetting.

However, it should be possible to reveal the benefit of retrieval practice on an immediate final test by measuring how long it takes for each item to be retrieved, instead of just how many items are ultimately retrieved. The PCR model predicts faster retrieval following test practice even if overall accuracy does not improve on an immediate final tests.

Practice Tests Reduce Retrieval Latency

To see why intra-item learning reduces the latency of subsequent retrievals of the same item, consider the example of CR shown in Figure 1. Even if successful test practice failed to strengthen associations between the retrieval cues and the item features (i.e., even if PR still only activated 2 of the 5 features), CR would be achieved in fewer time steps during subsequent retrieval attempts. The newly formed intra-item associations (shown by the dashed arrows), allow CR to occur in 2 time steps rather than 4 time steps. This is

because the 2 features initially activated during PR are now associated with *all 3* of the initially dormant features, instead of just 1.

Restudy also reduces retrieval latency, but for a different reason. Following restudy, the reduction in latency occurs because more features are active in the first time step (i.e., after PR). In other words, learning from restudy increases the number of features activated by the retrieval cues such that fewer features need to be filled in during CR. It is important to note that this type of learning also occurs with successful test practice. In other words, both restudy and test practice result in more features activated by the retrieval cues. However, successful test practice also boosts intra-item learning, and thus the CR process operates more efficiently (i.e., fewer time steps) to fill in the remaining features. In summary, for an immediate final test, accuracy is predicted to be higher following restudy because restudy boosts PR for all of the items. In contrast, retrieval latency is predicted to be faster following test practice because recalled items will have benefited both from better PR and from more efficient CR.

We tested these predictions in a free recall experiment in which 34 participants studied lists of 15 words, followed by either restudy or a practice test, and then a final test (this procedure was repeated 8 times). On the memory tests, participants were given 90 seconds to recall the 15 words in any order. Participants completed a 30 second math distractor task between both the initial study and the practice phase, and between the practice phase and the final test. This experimental design yields 3 conditions to measure: performance on the practice test itself (i.e., the baseline or "no prior practice" condition), performance on the final test following restudy practice, and performance on the final test following a practice test. In addition to recall accuracy, we analyzed the elapsed time between each item recalled (the Inter-Response Time, or IRT). The accuracy and latency results were in line with the PCR models predictions: restudy produced the highest accuracy (average of 81% correct) and faster IRTs than baseline, while IRTs were the fastest on a final test following test practice, despite no substantial change in accuracy (61% correct on the practice test and 59% correct on the final test). The median IRTs for each possible output position in this experiment are shown in Figure 3. Note that the lack of a "testing effect" in terms of recall accuracy is expected in this situation, as the learning benefits of test practice only apply to items *already* able to be recalled, and thus should only be expected to emerge with a longer retention interval. Demonstrating the generality of these latency effects, a similar speed-up has been found with a final cued recall test following cued recall practice (Broek, Segers, Takashima, & Verhoeven, 2014).

Applying PCR to Free Recall Latencies

The PCR model was simultaneously fit to the observed recall accuracy and IRTs from each item on each list of the free recall experiment, separately for each subject. More specifically, the PCR model predicts the shape of the IRT

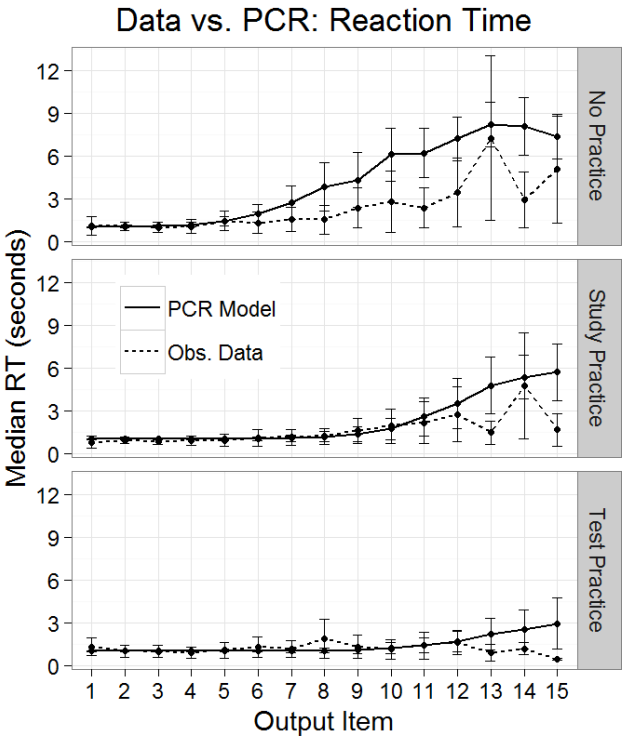


Figure 3: Predicted and observed inter-retrieval times in the no practice, study practice, and test practice conditions. Inter-retrieval times increased as function of output position in the no practice and study practice conditions, but not in the test practice condition.

distributions, and these were used to produce a maximum likelihood fit of the joint probability of producing each observed latency at the observed output order position within the test list (e.g., taking 4.3 seconds after recall of a 4th item to then recall a 5th item). In this manner, the model explained the shape of the separate IRT distributions as a function of output position within the test list, and as a function of prior restudy or prior test practice. The predicted IRTs using the best fitting model parameters for the baseline, restudy, and test practice conditions are shown along with the observed data in Figure 3.

To allow the model to capture the data, the T_{min}, T_{max}, and λ parameters were allowed to freely vary in addition to the *e*, *l*, *r*, and *t* parameters. When using a binomial distribution, the model generated IRT distributions are discrete because there is a finite number of possible RTs representing a finite number of possible values for the difference from threshold (however, this finite number is large when considering that an observed IRT may reflect some number failed CR attempts before a successful CR). A continuous IRT distribution was produced owing to two factors. First, rather than using the discrete Binomial distribution, we used a continuous Beta distribution that approximates the Binomial by having the same mean and variance. Theoretically, this corresponds to consideration of partial feature activation. Second, rather than assuming that each CR attempt was initiated precisely at

the offset of the last CR, we imposed a standard normal (0 mean and 1 second standard deviation) for the start time of each CR. This was implemented with Gaussian kernel estimation of the Monte Carlo data. This procedure produced a family of IRT distributions across the 15 possible output positions (i.e., these 15 distributions integrate to value 1.0), and in this way the model simultaneously explained both the accuracy data and trial-by-trial IRTs.

General Discussion

The PCR model has important theoretical implications for the testing effect literature as well as the broader field of memory research. To this date, there have been no well specified (i.e., mathematical) models of the cognitive processes that underlie the testing effect. Beyond serving this need, the PCR model provides a conceptually novel account of the benefits from retrieval practice. In the testing effect literature, the leading theories include overlearning (i.e., testing simply produces more learning), transfer appropriate processing (i.e., the best way to learn to recall at a later date is to practice recall), and desirable difficulties (i.e., testing requires more effort). Intra-item learning is conceptually different than any of these accounts. It is closest to transfer appropriate processing, but it supposes that the act of recall introduces a king of learning that will benefit future recall attempts for that item under a variety of circumstances (i.e., beyond situations that are identical to those of test practice). This theory provides an explanation of why test practice benefits are largest following recall practice (Carpenter & DeLosh, 2006); because CR is not reliably engaged by tests that do not require recall (e.g., a recognition or some forms of multiple choice tests), these tests do not produce as much intra-item learning.

While PCR incorporates many ideas from existing models of memory, the assumption of different kinds of learning for the two stages that underlie recall is an important departure from prior models. These models concern the association strength between retrieval cues and memories, and then use these associations in different ways to explain the difference between familiarity responses (the first stage) and recall responses (the combination of two stages). By additionally considering intra-item learning, PCR goes beyond passive theories of memory formation to explain why the act of recalling something from memory results in a qualitatively different kind of learning; a kind of learning that is unique to the item, allowing better/faster recall of the item even when retrieval cues change. Nevertheless, in the PCR model, these two types of learning follow from the same learning mechanism that builds associations between features according to the temporal order in which features become active. Thus, the key distinction between passive study and active recall is that passive study is an all-at-once event in terms of items and their features, whereas recall is a gradual unveiling of an item's features.

References

Broek, G. S. E. van den, Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. http://doi.org/10.1080/09658211.2013.831455

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268 – 276.

Dan, Y., & Poo, M.-M. (2006). Spike Timing-Dependent Plasticity: From Synapse to Perception. *Physiological Reviews*, *86*(3), 1033–1048. http://doi.org/10.1152/physrev.00030.2005

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, *16*(2), 96–101. http://doi.org/10.3758/BF03202365

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. http://doi.org/10.1006/jmps.2001.1388

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *The Quarterly Journal of Experimental Psychology*, *65*(5), 962–975. http://doi.org/10.1080/17470218.2011.638079

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. http://doi.org/10.1016/j.jml.2011.04.002

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. http://doi.org/10.1037/0033-295X.88.2.93

Roediger, H. L. I., & Karpicke, J. D. (2006b). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249 – 255.

Roediger, H. L. I., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. http://doi.org/10.1111/j.1745-6916.2006.00012.x

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. http://doi.org/10.1027/1618-3169.56.4.252

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571 – 580.

Modelling the co-development of word learning and perspective-taking

Marieke Woensdregt (m.s.woensdregt@sms.ed.ac.uk)[a]

Simon Kirby [a], Chris Cummins [b], Kenny Smith [a]

[a] Centre for Language Evolution, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

[b] Department of Linguistics and English Language, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

Abstract

Word learning involves mapping observable words to unobservable speaker intentions. The ability to infer referential intentions in turn has been shown to depend in part on access to language. Thus, word learning and intention-reading co-develop. To explore this interaction, we present an agent-based model in which an individual simultaneously learns a lexicon and learns about the speaker's perspective, given a shared context and the speaker's utterances, by performing Bayesian inference. Simulations with this model show that (i) lexicon-learning and perspective-learning are strongly interdependent: learning one is impossible without some knowledge of the other, (ii) lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both even when the learner starts with no knowledge of the lexicon and unhelpful assumptions about the minds of others, and (iii) receiving initial input from a 'helpful' speaker (who adopts the learner's perspective on the world) paves the way for later learning from speakers with perspectives which diverge from the learner's. This approach represents a first attempt to model the hypothesis that language and mindreading co-develop, and a first exploration of the implications for theories of word learning and mindreading development.

Keywords: word learning; perspective-taking; computational model; Bayesian inference;

Introduction

Word learning is a special case of associative learning, as one has to learn a mapping between something observable — a speaker's utterance — and something unobservable — the speaker's meaning. Word learning therefore requires inferring the speaker's referential intention (Waxman & Gelman, 2009), which in turn requires theory of mind (ToM). Learning about words and learning about minds are thus necessarily connected: language learners need to figure out not just the stable mappings between words and concepts (the lexicon) but also a way of inferring speaker intention, which is variable over time and depends on context and speaker-specific features.

In this paper we present evidence that language and ToM development go hand in hand, and we explore the implications of such a co-development by means of an agent-based model. As a test case we look specifically at the interaction between word learning and perspective-taking. Although perspective-taking cannot be equated with ToM, it is an instantiation of the latter and forms a good starting point for formalising the relation between language learning and ToM development.

Learning about words and minds

There is persuasive evidence consistent with the idea that learning about words and learning about minds are interrelated. In a study comparing children with autism (AD) to typically-developing (TD) children, Parish-Morris et al. (2007) showed that although 5-year-old AD children have some ability to use social cues (pointing and eye gaze) to direct their attention in word learning, they perform at chance when learning new words required inferring the speaker's intention, unlike language- and mental-age-matched TD children.

The reverse phenomenon has also been observed, namely that the development of ToM depends in part on having access to language. Deaf children of hearing parents, who lack consistent linguistic input, were shown to have delayed ToM development relative to deaf children of deaf parents, who receive sign language input from birth (Schick et al., 2007). Similarly, a study with TD children showed that simply training children on the use of mental state verbs with sentential complements accelerated their false belief understanding (Lohmann & Tomasello, 2003).

Thirdly, in a study comparing different age-groups of signers of the recently emerged Nicaraguan Sign Language, Pyers and Senghas (2009) showed that the bootstrap effect of language on ToM development continues on into adulthood. Pyers & Senghas found that the first cohort of signers (mean age 27), whose language had very limited mental state vocabulary, were worse at understanding false belief than the second cohort (mean age 17) who had more words for mental states. Moreover, a follow-up study two years later revealed that the first-cohort signers had improved in their false belief understanding and that this either followed or co-occurred with, but never preceded, an expansion of mental state vocabulary.

Finally, recent evidence suggests that mindreading and language skills co-develop. Brooks and Meltzoff (2015) showed that gaze-following in 10.5-month-old infants predicted their production of mental state terms at 2.5-years-old, and that these mental state terms in turn predicted the extent of their false belief understanding at 4.5-years-old, even though gaze-following did not directly predict false belief understanding. Thus, this shows evidence of an indirect relation between early sensitivity to social cues and later mindreading ability, mediated by language.

Models of word learning and perspective-taking

Words are used in complex environments, and each word could label any part of that complex environment. Worse, words can label objects and events which are not currently perceivable to the hearer and/or the speaker (e.g. events which are spatially or temporally distant from the time of speaking). Learners therefore face *referential uncertainty*: every time a word is used, there may be many meanings which a learner could infer as the word's intended meaning.

Computational models of word learning have explored several potential solutions to the problem of referential uncertainty, which could be roughly divided up into three kinds: (i) solutions using learning biases, (ii) social cues solutions, and (iii) intention-reading solutions.

Brute force statistical learning of word-referent associations is impossible if referential uncertainty is unbounded: if all logically possible meanings are equally-plausible candidates for the meaning of any word on any use, then no learner can learn the meaning of any word (an observation commonly attributed to Quine, 1960, in his work on radical translation). Experimental and observational studies have demonstrated that word learners use a number of heuristics to reduce referential uncertainty: learners assume that words refer to whole objects (Macnamara, 1972); they use argument structure and syntactic context to constrain the meaning of new words (Gillette et al., 1999); and they use knowledge of the meaning of other words to constrain hypotheses about the meaning of a new word, for example by assuming that words have mutually exclusive meanings (Markman & Wachtel, 1988). Models of cross-situational statistical learning suggest that brute-force cross-situational learning of large lexicons is possible under surprisingly high levels of referential uncertainty (Blythe, Smith, & Smith, 2010) or even under infinite referential uncertainty if word learners can use their heuristics to rank candidate meanings in terms of their plausibility (Blythe, Smith, & Smith, submitted).

In addition to exploiting linguistic context or their knowledge of likely word meanings, learners can use social cues, which are potentially highly informative in guiding word learning (see Paulus and Fikkert (2014) for eye-gaze and pointing and Yu and Smith (2012) for joint attention). Yu and Ballard (2007) formalised these mechanisms in a model of word learning that integrates the use of statistical regularities and social cues. They provided an associative model with information about which words and objects in a discourse stream were highlighted by social cues (prosody and joint attention), and simply increased the association weight of those items. They then tested the model on how well it could learn a lexicon from transcriptions of two videos of mother-child interactions from the CHILDES corpus. This 'hybrid' model was compared to a 'bare' statistical learning model, and statistical learners who exploited prosody or joint attention, but not both. Best performance was obtained with the model that integrated both types of social cue.

However, there is more to social interaction than just cues that direct attention. The ability to recognise that speech can convey unobservable communicative intentions comes online before children start talking (Vouloumanos, Onishi, & Pogue, 2012) and is used to guide their word learning (Parish-Morris et al., 2007). To formalise the role that inferring speaker intentions plays in word learning, Frank, Goodman, and Tenenbaum (2009) designed a Bayesian model that simultaneously infers word-object mappings and speaker intentions, and tested this model on the same CHILDES videos used by Yu and Ballard (2007). Rather than re-weighting items based on social cues, Frank et al. assume that learners posit an extra unobserved variable mediating between the objects in the physical context and the words that the speaker produces: the speaker's referential intention. The learner then evaluates all possible lexicon hypotheses based on the prior probability of that lexicon and the likelihood of a word given that lexicon and the speaker's referential intention, where the intention hypotheses that are considered by the learner are simply all possible subsets of the objects present in the context, including an 'empty' intention.

This model has two advantages over other associative learning models. Firstly, it can represent the possibility of 'empty intentions', where the word does not refer to any physically present object. Secondly, it can distinguish between words that can be used referentially and words that are used exclusively 'non-referentially', where non-referential (e.g. function) words are simply left out of the lexicon. Frank et al. (2009) show that this model outperforms several alternative statistical learning models (including Yu and Ballard's), both when tested on the lexicon they learned and on the referential intentions they inferred (given their lexicon).

Although these various models constitute important first steps towards modelling the role of intention-reading in word learning, they treat the ability to utilise social cues or infer intentions as a given and fixed capacity, present from the start of word learning. In real-world learning, the ability to learn words and the ability to infer mental states (including referential intentions) improve as a child grows older. As described above, this improvement is partly accounted for by a co-development of language and intention-reading. Below, we will describe a model that takes these considerations into account: rather than modelling word learning as a combination of associative learning with social cues or uninformed intention representations, we provide a model which allows for the co-development of word learning and perspective-taking.

The current model: Integrating development of word learning and perspective-taking

Model description

We model referential intentions as a result of the interaction between a set of attributes of the world — the context — and an attribute of the speaker — the perspective. This perspective can be interpreted in a literal sense, where objects that are spatially or temporally closer to the agent are more salient (see figure 1). Importantly however, it can equally serve as a

model for the sum of an agent's knowledge and beliefs about the world that determine what topics of conversation will be most salient to them in a given situation. The latter is the sort of perspective that requires full-blown ToM to be inferred. All that matters here is that there is a function that maps from the attributes of the world to an agent's saliency distribution over potential topics, and that the agent has a hidden variable (their perspective) that is a parameter in this function.

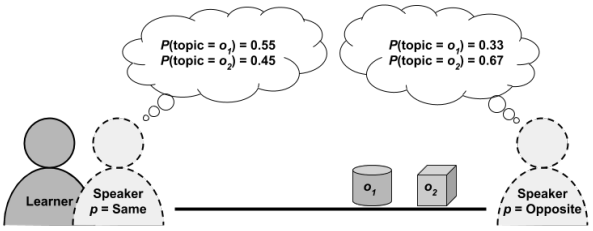


Figure 1: Diagram of how speaker perspective gives rise to referential intention. The speaker on the left is only slightly more likely to choose object 1 (o_1) over object 2 (o_2) as a referent, since both are approximately equidistant. The speaker on the right however is twice as likely to choose object 2 than object 1 since o_2 is twice as close as o_1. The learner has their own perspective on the world and learns with an egocentric bias; assuming that the speaker shares their perspective.

The variables that the learner can observe are the context and the speaker's utterance (see figure 2). The variables that are unobservable are the speaker's perspective, the speaker's referential intention, and the speaker's lexicon. The learner's task is to infer the speaker's perspective and the lexicon based on the same data: the speaker's word use in different contexts.

This model differs from that outlined in Frank et al. (2009) in that it posits an extra unobservable variable: the speaker's perspective, which together with the context determines the speaker's referential intention. Given a specific hypothesis about the speaker's perspective, the learner can compute a prediction of how likely it is that the speaker will refer to a given object in a given context (i.e. how salient the object is for the speaker). Subsequently, given a specific hypothesis about what the lexicon is, the learner can turn this prediction about likely referents into a prediction of likely utterances.

We assume, unlike the models of word learning described above, that all objects that are part of the world are possible referents in every learning context: thus, simple associative cross-situational learning alone will not be able to solve the problem of referential ambiguity. The learner can get around this problem by inferring the speaker's perspective: a hypothesis about this perspective is the only information available that can render the probability distribution over possible referents non-uniform, which in turn allows the learner to infer the most likely word-object mappings. Specifically, this is achieved by incrementing the posterior belief in lexicon hypotheses in proportion to how salient the object that is associated to the utterance in that lexicon is for the speaker, given the perspective hypothesis under consideration.

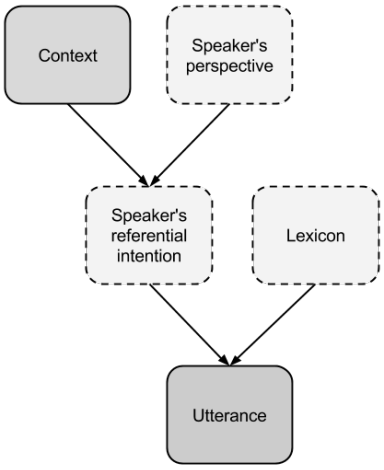


Figure 2: Diagram of the current model. Variables in dark grey and solid lines are observable to the learner, variables in light grey and dashed lines are unobservable. The learner's task is to infer the speaker's perspective and the lexicon based on observations of the speaker's word use in context.

Note that in this model no lexicon hypothesis can be evaluated without simultaneously positing a perspective hypothesis, and vice versa. Thus the complete hypothesis space for the learner consists of all possible combinations of lexicon hypothesis and perspective hypothesis (with the potential of representing different perspectives, and indeed different lexicons, for different speakers). Learning in this model is implemented as Bayesian inference according to the definitions described below.

Posterior The task of the learner in this model[1] is to find the lexicon hypothesis l and perspective hypothesis p that have the highest posterior probability given data D, as shown in equation 1.

$$P(l, p \mid D) \propto P(D \mid l, p)P(l, p) \quad (1)$$

The perspective hypothesis p represents a single parameter in an intention function that maps from the context to the speaker's referential intention. This referential intention is based on the saliency of the objects in the context, which is defined as the inverse of the distance between the speaker's perspective and the object's location (see figure 1). These saliency values are then normalized over all objects in the context, rendering a probability distribution over all objects

[1]We describe the model in terms of a learner who assumes that a single lexicon and a single speaker perspective will account for all of their data: the same model can straightforwardly be extended to model a learner who allows that different speakers might have different lexicons and different perspectives; later we present results for a learner who entertains multi-perspective hypotheses.

that determines how likely the speaker is to choose them as intended referent. This distribution is then used to generate the speakers referential intention.

The learner does not need to infer the intention function itself, only the perspective parameter. This model thus simulates the situation where the learner is 'born' with the ability to represent mental states, but has to learn how to make predictions about the *content* of another agent's mind on the basis of the context. More specifically, the learner is born with a model of how a context will give rise to a speaker's referential intention, given the speaker's perspective, but has to infer from data exactly what the speaker's perspective is.

Likelihood The likelihood of a set of data D is:

$$P(D \mid l, p) = \prod_{d \in D} P(w_d \mid l, p, c_d) \tag{2}$$

where each data point d consists of a context c and a word w that was uttered by the speaker in that particular context. The likelihood of a single word w_d is defined in equation 3.

$$P(w_d \mid l, p, c_d) = \sum_{o \in c_d} P(i_o \mid p, c_d) P(w_d \mid i_o, l) \tag{3}$$

where o stands for object and i_o for the probability that object o will be the intended referent given the perspective hypothesis p.

Thus, the probability of a particular word being uttered in a particular context is equal to the product of the probability of that word being uttered for a given object (according to lexicon hypothesis l) and the probability of that object being the intended referent (according to perspective hypothesis p), summed over all objects.

In the simulations described below all lexicon hypotheses that are considered consist simply of discrete binary mappings between words and objects — in other words, if there are two objects and two possible words, there are nine possible lexicons (object 1 maps to word a or word b or either, and object 2 independently maps to word a or word b or either). Thus, the probability of a given word being uttered for a given intended referent is given by equation 4

$$P(w_d \mid i_o, l) = \begin{cases} \dfrac{1}{|w_o|} & \text{if } w_d \text{ maps to } o \text{ in } l \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $|w_o|$ is the number of words that map to object o in lexicon l.

Prior For all simulations described below, we assume that learners have a neutral prior over lexicons and an egocentric prior over perspectives. That is, the learner starts out assuming that all lexicons are equally probable, and that other agents share their own perspective. Over all combinations of lexicon and perspective prior, the 'composite prior' is simply the product of the two, as shown in equation 5.

$$P(l, p) = P(l)P(p) \tag{5}$$

Simulation results

All simulation results described in this section show what happens in the very simple case where the learner gets input from one or two speakers in a world where there exist only two possible referents (objects) and two words. The set of lexicon hypotheses consists of all functionally distinct ways of mapping two words onto two objects (nine lexicons in total, as described above). The set of perspective hypotheses consists of the two most extreme possibilities: either the speaker's perspective is the same as the learner's own perspective, or it is exactly the opposite. The learner's hypothesis space consists of all possible combinations of lexicon hypothesis and perspective hypothesis.

In a first set of four simulations we explore the influence that perspective-learning and lexicon-learning have on each other. We compare three different cases: (i) the target lexicon is unambiguous (i.e. each object is associated with a distinct word) but the learner is unable to learn that speakers might have a perspective that is different from their own (which we achieve by setting the prior probability of the 'other' perspective to 0); (ii) the learner is initially egocentric yet can learn that speakers can have a perspective that differs from their own (which we achieve by setting the prior probability of the 'other' perspective to 0.1, and the 'own' perspective to 0.9), but the target lexicon is partly ambiguous (e.g. object 1 maps to both word a and word b, while object 2 maps only to word b); (iii) same as in (ii) but with a fully ambiguous lexicon (both objects map to both words); and (iv) the learner can learn that the speakers can have a different perspective from the learner, as in (ii) and (iii), and the target lexicon is unambiguous, as in (i).

Situation (iv) thus simulates a typically-developing child in a normal language environment (under the assumption that words are effectively unambiguous in their linguistic context: Piantadosi, Tily, & Gibson, 2012) — we refer to this as the Typical condition. Situation (i) simulates a word learner with a strongly impaired (or absent) ToM — we refer to this as the No ToM condition. Situation (ii), which we refer to as the Partly Ambiguous Lexicon condition, simulates a typically-developing word learner in an environment where the target lexicon is such that a speaker's utterances are rather uninformative about their referential intentions. This scenario could be compared to the case of deaf children who grow up with hearing parents (i.e. without sign language), since although such parents do exhibit communicative behaviour that could reveal something about their communicative intentions, this is less explicit and more ambiguous than linguistic data (Schick et al., 2007). Finally, situation (iii), which we refer to as the Uninformative Lexicon condition, is an extreme form of this case, where there is a complete absence of behaviour that is informative about the speaker's intentions. This is a case analogous to one in which a reliable language has yet to emerge in a population.

Figure 3 shows the learning results for the four different situations described above. Several interesting learning dynam-

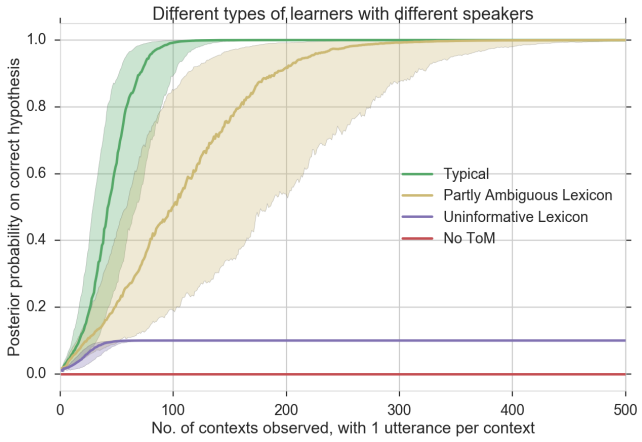


Figure 3: Learning curves for different learners in different learning situations. Learning is measured as the amount of posterior probability assigned to the correct hypothesis, where 1.0 is ceiling. Lines show median over 1000 runs, shaded area shows first and third quartile.

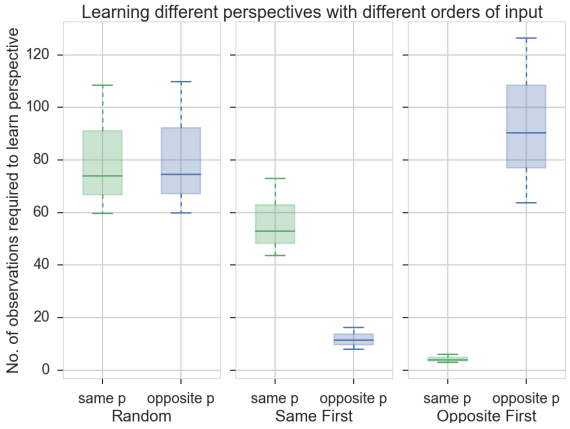


Figure 4: Amount of observations required for learning different speaker perspectives under different input conditions: Random, Same First and Opposite First. Successful learning is defined as > 0.99 posterior probability on correct hypothesis, and the lexicon is learned fully in all conditions before the learner enters the second input phase. Boxes show median, first and third quartile over 100 runs.

ics are apparent. Firstly, inferring the correct lexicon is impossible when the learner cannot infer the correct perspective of the speaker that they get input from (No ToM condition). Secondly, inferring the speaker's perspective becomes more difficult when there is a less direct mapping between their referential intention and their behaviour (Partly Ambiguous Lexicon condition). However, learning in this case is still eventually successful: the ability to infer perspective gives a way into learning the lexicon, thus making it easier to deal with lexical ambiguity. Thirdly, inferring the speaker's perspective becomes impossible when the speaker's behaviour gives no information at all about their intention (Uninformative Lexicon condition). Finally, learning happens most quickly and successfully when the learner is both able to represent different perspectives and the speakers' lexicon is unambiguous (Typical condition).[2]

In a second set of three simulations we present the effect of order of input on lexicon and perspective learning. These simulations are similar to the ones described above, except that the learner receives input from two different speakers who have two different perspectives: one speaker shares the learner's perspective, the other has the opposite perspective. We present the learning results in three different situations: (i) the speaker is randomly picked on each trial, but both speakers get to speak for an equal number of contexts (Random condition); (ii) the learner receives the first half of their input from the speaker that shares their perspective, and the second half from the 'opposite perspective' speaker (Same First condition); and (iii) the learner receives the first half of input from the opposite perspective speaker and the second half from the same perspective speaker (Opposite First condition).

[2]These results are qualitatively similar for learning about larger lexicons of e.g. 3x3 and 4x4 objects and words.

As figure 4 shows, the difference in the amount of observations that is required to learn the opposite perspective is bigger between the two conditions (Same First vs. Opposite First) than the difference in the amount of observations required to learn the same perspective in the two conditions. This means that receiving input from a 'helpful' speaker (a speaker who shares the learner's perspective) first paves the way for later learning about perspectives that are different from the learner's own.[2]

The mediating factor that gives rise to this effect is the lexicon, since the only thing that is different about the learner after having learned the same perspective first is their knowledge of the lexicon. (Which, in all simulations shown in figure 4, is fully learned before the learner enters the second input phase.) This effect relies on the lexicon being shared among members of the population. Language as a convention is what allows the learner to bootstrap knowledge of other's perspectives based on starting with a familiar speaker first.

Discussion

We presented an agent-based model that simulates the co-development of word-learning and perspective-taking through Bayesian inference. This model is different from previous models of word learning in that all objects that are part of the world are considered as potential referents at each learning episode, rendering brute-force cross-situational learning impossible. However, the learner can overcome this referential uncertainty by learning about the speaker's perspective. Both the lexicon and the perspective are learned using the same data (the speaker's word use in context).

This model gives rise to several potentially interesting co-development dynamics. Firstly, lexicon-learning and

perspective-learning are strongly interdependent: learning the one cannot happen without some knowledge of the other. Secondly, lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both variables even when the learner starts out with an inappropriate egocentric bias and no knowledge of the lexicon whatsoever. Finally, the results show that receiving input from a helpful speaker first paves the way for later learning from speakers whose perspective differs from the learner's — the helpful speaker provides data which facilitates learning of the lexicon, which then facilitates learning of the perspective of other less well-aligned speakers (on the assumption that the lexicon is shared among speakers).

To our knowledge, this is the first computational model that does not simply incorporate pragmatic inference as a tool to infer word meaning (Frank et al., 2009), but rather incorporates pragmatic inference as a developing skill that interacts bi-directionally with word learning. Thus, this model is a first step towards formalising the hypothesis that language and mindreading co-develop.

The simulation results of this model described here replicate several empirical findings. Firstly, it mirrors the finding that word-learning depends partly on the inference of mental states (Parish-Morris et al., 2007). Secondly, it mirrors the finding that the development of mindreading depends partly on vocabulary development (Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007). Finally, it generates the developmental prediction that learning from a helpful speaker who shares the child's perspective early on in life will aid vocabulary development, and that this in turn will help the child to learn about alternative perspectives later on.

Several aspects of this model are however very simplistic. Firstly, the learner in this model is 'born' with a ToM. Rather than having to infer the full function that maps from a context to a speaker's referential intention, the learner only has to infer the speaker's perspective. In real life children have to develop not only the ability to infer the content of mental states, but also the underlying ability to represent *that* the content of others' minds is different from that of their own. Future work with this model could incorporate a more realistic model of ToM development that could mimic more closely the stages of ToM development we see in real children.

Secondly, the relation between observations of words and learning about perspectives is very direct. Each word-object mapping that is learned helps with inferring perspective because it allows the learner to evaluate their prediction of referential intent based on their perspective hypothesis. It is not yet clear what the role of language learning is in driving the development of ToM in the real world — this might have to do with access to discourse, explanations or representations of mental states (see e.g. Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007).

Despite these simplifications, this model forms a first exploration into the co-development dynamics of language and ToM.

References

Blythe, R. A., Smith, A. D. M., & Smith, K. (submitted). Word learning under infinite uncertainty.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620–42.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67–78.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–85.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–76.

Lohmann, H., & Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, *74*(4), 1130–44.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*, 121–57.

Parish-Morris, J., Hennon, E. a., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development*, *78*(4), 1265–87.

Paulus, M., & Fikkert, P. (2014). Conflicting Social Cues: Fourteen- and 24-Month-Old Infants' Reliance on Gaze and Pointing Cues in Word Learning. *Journal of Cognition and Development*, *15*(1), 43–59.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280-91.

Pyers, J. E., & Senghas, A. (2009). Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language. *Psychological Science*, *20*, 805–12.

Quine, W. V. O. (1960). *Word and Object*.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: a study of deaf children. *Child Development*, *78*(2), 376–96.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *PNAS*, *109*, 12933–7.

Waxman, S. R., & Gelman, S. a. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–63.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–65.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–62.

Salience versus prior knowledge - how do children learn rules?

Samuel Rivera (rivera.162@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH 43210 USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH 43210 USA

Abstract

Categories are essential for thinking, learning, and communicating. Research has shown that young children and adults treat categories very differently, with young children favoring whole objects while adults focus on the key information in most cases. If so, then how can young children learn categories requiring focused attention to key features? Studies have shown that drawing attention to rules had facilitative effects. We sought to identify whether the effect was driven by instruction about rules or by stimulus-driven factors. Our results suggest that even with instruction, 4-year-olds were not able to attend to key information. Simply making important information more salient, however, allowed them to learn the category and *transfer* to situations when the key feature was no longer salient.

Keywords: category learning; attention; cognitive development

Introduction

Our world is made up of categories and concepts - groups of objects or ideas that share some equivalence. They allow us to generalize, communicate, and make decisions by abstracting away the unnecessary variation in the world. Decades of research have demonstrated that young children have difficulty attending selectively to "category-relevant" information and using this information when learning categories (Smith & Kemler, 1977). Children are thought to be *holistic* learners, with evidence showing a distributed pattern of attention in comparison to the focused attention profile of adults (Best, Yim, & Sloutsky, 2013). This means that children categorize objects based on their overall appearance while adults emphasize key features when categorizing. This phenomenon has been linked to the protracted development of selective attention and executive functioning in children (Smith, 1989; Sloutsky, 2010; Rabi & Minda, 2014). But if young children cannot focus on defining features, how and under what circumstances do they learn rule-defined categories? Answering this question is important for developing a theory of early categorization, but also has immediate practical relevance. Specifically, it informs efforts in teaching young children certain abstract concepts, such as fractions, where understanding a key idea or feature is prerequisite to learning.

Prior work has suggested that the solution to overcoming the holistic predisposition may depend on attention influences, with both instruction (Deng & Sloutsky, 2015) and feature salience (Deng & Sloutsky, 2012; Rabi, Miles, & Minda, 2015) contributing to learning and generalization. Deng and Sloutsky demonstrated this point in their study of the role of labels over development (Deng & Sloutsky, 2012). In their categorization and induction study, they found that when a very salient feature was in conflict with several other feature values, children would respond based on the salient feature value. Rabi and Minda (Rabi et al., 2015) similarly found that salience had a significant effect on what type of information children used to categorize, and their success or failure in learning category rules. Their results suggested that children employed a variety of optimal rule-based, suboptimal rule-based, and similarity based strategies when children were learning categories defined by rules and family resemblance (R+FR). However, children were more successful at learning the rule when the rule-feature was salient. These results show that exogenous forces of the stimulus properties can facilitate category learning in young children.

In a more recent investigation of categorization and induction, Deng and Sloutksy (2005) found that 4-year-olds relied on overall similarity to categorize objects when they were learned by categorization or inference training. In inference training, the objective is to determine a missing feature value instead of the unknown category label as in classification training. When 4-year-olds were trained to categorize and repeatedly given instruction about the rule feature, however, they relied on the rule feature and categorized like 6-year-olds. This work implies a facilitative effect on top-down knowledge in selecting the key information for categorizing. Unfortunately, it is difficult to decouple the influence of instruction that requires effortful top-down control from selection history (Awh, Belopolsky, & Theeuwes, 2012) because instruction included specific exposure to features.

The short survey above demonstrates that instruction as well as feature salience may have effects on young children's ability to learn categories. Our goal is to understand how and when those factors contribute to category learning. One possibility is that young children have difficulty identifying the important features when learning. Thus, the difficulty in focusing on key information during childhood could stem from poor learning strategies in general, which may be remedied with prior knowledge of key information. On the other hand, focusing on the key feature may not be possible early on due to limited selective attention. If that is the case, then an explicit manipulation of feature salience would aid learning. A third possibility is that both instruction and salience contribute to rule-based category learning to different degrees over development.

Thus, in this work we attempt to decouple the effect of both instruction, a form of prior knowledge, and stimulus properties (exogenous influences) on category learning in order to understand how those forces contribute to category and concept learning over development. To address these open questions, we investigated the differential influence of attentional cues on category learning and generalization. We predicted that exogenous cues would have a more potent facilitative effect on children's category learning then endogenous cues due to limited executive functioning. Even with instruction, having to use effortful top-down control may not be as easy for young children. It was unclear whether endogenous instruction would help at all, or to a lesser degree. If young children can direct their attention to important features when learning categories, we would expect no difference between the exogenous versus endogenous cue. In the following we detail our experimental methodology and results. We close with a discussion.

Method

Participants

A total of 273 adults participated in the study for course credit through The Ohio State University research experience program. One additional adult participant who did not complete the study was excluded from the analysis. Adults either participated in the baseline (N = 74), endogenous (N=87), or exogenous (N=112) condition. A total of 87 four-year-olds (M$\pm$SD$= 4.52 \pm 0.27$ years) participated in the study. They completed either the baseline (N=24; M$\pm$SD=4.57 ± 0.22 years in range $[4.04 : 4.87]$), endogenous (N=33; M$\pm$SD=4.45 ± 0.28 years in range $[3.84 : 4.90]$), or exogenous condition (N=30; M$\pm$SD=4.57 ± 0.28 years in the range=$[4.01 : 5.00]$). Children were recruited through local daycares or preschools located in Columbus, Ohio, and public birth records. The majority of child and adult participants were Caucasian. Adults gave informed written consent prior to participation. Child participants gave verbal assent, and a caretaker gave written consent prior to the study.

Stimuli

Stimuli were artificial tree-like visual categories comprised of 6 spatially separated petals on branches extending from a central trunk, as shown in Figure 1. The stimuli spanned approximately 22 by 22 degrees of visual angle on the display. Neighboring petals were separated by approximately 8 degrees of visual angle from their centers, with about 2 degrees of empty space between the petals.

Categories were defined by the color and shape of the features according to a rule plus family resemblance (R+FR) structure. Specifically, the top right petal had a deterministic color and shape that perfectly determined the category. All other features varied probabilistically, such that all but one color and one shape was associated with a contrasting category. Thus, participants could learn the rule, or consider the overall appearance of features to correctly classify the objects

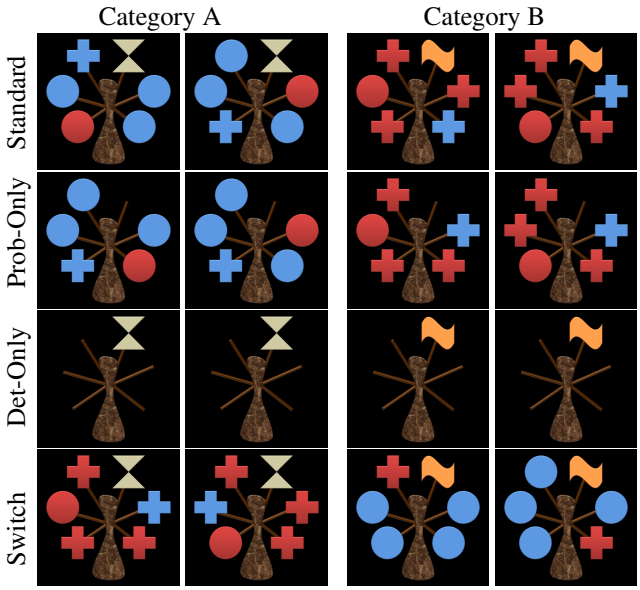


Figure 1: We show example stimuli used in the experiment. The first row shows standard stimuli used for the testing or training portion of the experiment. The bottom 3 rows show stimuli used only during testing. The Prob-Only images only show the probabilistic features or family resemblance features. The Det-Only condition displays only the rule feature. The switch items have the rule and family resemblance features in conflict.

Table 1: Example category structures used in the study. C1 and S1 correspond to color and shape at position one, respectively. Deterministic features are in bold.

	C1	S1	C2	S2	C3	S3	C4	S4	C5	S5	C6	S6
A	**0**	**0**	0	0	0	0	0	0	1	0	0	1
	0	**0**	0	0	0	0	1	0	0	1	0	0
	0	**0**	0	0	1	0	0	1	0	0	0	0
B	**1**	**1**	1	1	1	1	1	1	0	1	1	0
	1	**1**	1	1	1	1	0	1	1	0	1	1
	1	**1**	1	1	0	1	1	0	1	1	1	1

(Rosch & Mervis, 1975). Some example category structures are described symbolically in Table 1, were the feature value 0 corresponds to a value from category A, and a feature value 1 corresponds to category B.

Experiment

We ran a between subjects design where participants learned categories in either a baseline, endogenous cue, or exogenous cue condition. Testing was the same across conditions. The key manipulations are illustrated in Figure 2, with details explained below. Adults completed the task in a quiet room on campus at The Ohio State University. All instructions were displayed on the screen and read by participants at their own pace. Children completed the study one at a time with a trained experimenter reading the instructions and giving encouragement throughout.



(a) Baseline (b) Endogenous (c) Exogenous

Figure 2: We show examples of stimuli during the training trials of the different experiment conditions. In the endogenous condition, the arrow remained during the entire trial but was not shown during test trials. Participants were told that the arrow points to an important feature that will help them categorize. In the exogenous condition, the yellow circle blinked during the training trials, but not during testing. Participants were told that some things may blink, but were not told that the feature is important.

Baseline In the baseline and all other conditions, participants learned to discriminate between two different categories. Participants were told that they would be playing a game where they would help creatures who live in either the desert or ponds to find their homes. Stimuli were shown centrally, with a small picture of a desert on the left, and a pond on the right. Adults responded at their own pace by pressing the left or right arrow which immediately prompted the feedback display. On the feedback screen the test exemplar was shown near the correct target (pond/desert), with either a central smiling or neutral face to indicate correct or wrong answers.

Participants first completed a series of 6 practice trials where they categorized frogs and camels. Each practice trial repeated until it was completed correctly. After practice, participants were told that they would fly away to a magical place with new types of creates. They completed 30 training trials with feedback and an equal number of exemplars from each category. The left versus right position of the target was counterbalanced across participants. After the training trials, participants were asked to complete test trials without feedback. They had 8 standard test trials, 8 Switch trials, 8 probabilistic only (Prob-Only) trials, and 8 deterministic only (Det-Only) trials presented in a fixed order and without breaks between. Examples of stimuli from the different test trials are shown in Figure 1.

Child participants completed the same study, with a few age appropriate adjustments. Namely, children made their response by touching the pond or desert. In addition, only smiling faces were shown for correct answers and no face for incorrect answers. Corrective feedback was still given on every training trial in the form of the exemplar shown near the correct target on the feedback screen. The entire procedure took children about 10-12 minutes.

Exogenous The exogenous condition was like the baseline except for the key manipulation. Namely, during the instructions at the beginning of the training trials after practice, participants were told that "Some things may blink but that is okay. Help them find their homes just like we did before." These instructions were given so that participants wouldn't be confused by the introduction of blinking after the initial practice trials. Then during the training trials, a yellow circle blinked around the deterministic feature. Nothing blinked on test trials.

Endogenous The endogenous condition was just like the baseline except for two key changes. During the post-practice instructions, adults were told that they "will see an arrow that will point to an important part of the creatures that will help you make your choice. Remember, the arrow points to something important!". Child instructions were more detailed. They were told that "There is a clue to this game! The clue will help you get all the smiley faces. You will see an arrow that will point to an important part of the creatures that will help you make your choice. Just like people who have different parts like arms and legs, the new creatures will have different parts. An arrow will point to a part that will help you know where it goes! Remember, the clue is that the arrow points to an important part that will help you know where the creature goes! Can you tell me the clue?" Experimenters repeated the instructions as necessary until children could mention the arrow clue and understood. About halfway through data collection, we added a memory check where we asked children if they remembered the clue. About 75% of polled children mentioned the arrow on the memory test. There was no effect of memory test result (pass/fail) on the training accuracy (p=0.80).

Results

We were interested in the differential effects of attention cues on learning and transfer to novel category exemplars. The different test trials probed different types of learning and response strategies. The standard test trials were an indicator of learning in general. Above chance (50%) accuracy on the standard test trials indicates successful category learning by either overall similarity or rule. The switch trials tell us how participants bias their decision. If the switch accuracy drops slightly from test but remains above chance, then participants may use both types of information but rely mostly on the deterministic feature. If the switch accuracy drops to below chance, then participants favor the probabilistic features in their decision. The Prob-Only trials probe whether participants had learned to use probabilistic features at all. If accuracy remained as high as on test, then we would infer that participants were using only this information. If it dropped slightly from test but remained above chance, then participants were using information from all features. If it dropped to chance on these trials, then we would infer that participants were only using the deterministic feature. On the other hand, the Det-Only trials tell us whether participants learned to use the rule feature. If accuracy dropped to chance, then we would infer that participants were using the probabilistic features to make their decision.

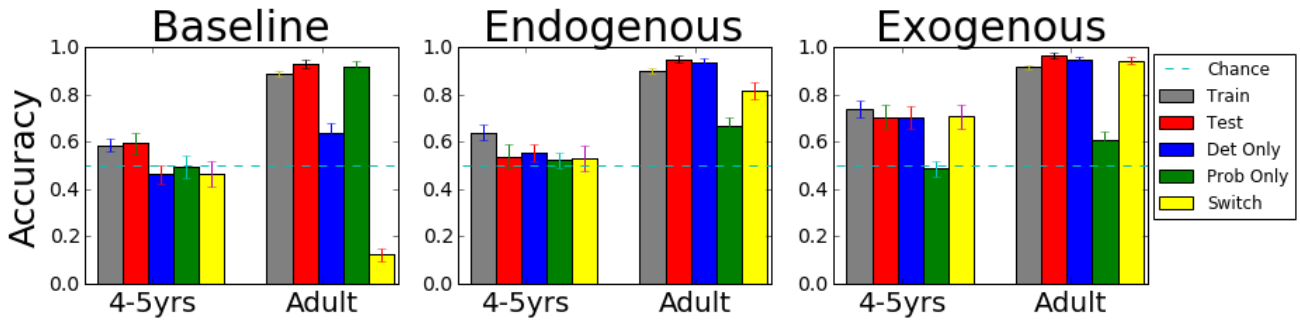


Figure 3: We plot mean accuracy across experiment conditions and age groups. Error bars denote standard error.

Our predictions were that adults would use rule features across all training conditions, with either cue providing equal support for learning that rule. That is because adults have been shown to optimize attention such that they only attend to the rule after learning (Shepard, Hovland, & Jenkins, 1961; Rehder & Hoffman, 2005; Hoffman & Rehder, 2010). We predicted that children, however, would receive more benefit from the exogenous versus endogenous cue, but would rely on the more holistic exemplar appearance when making their judgements. We expected their memory for the holistic appearance to be reduced in the exogenous condition versus the endogenous condition, from prior results showing the effects of very salient features on learning (Deng & Sloutsky, 2012).

The main analyses were concerned with the effects of the different training cues on accuracy for the different types of test trials, and changes in these effects over development. Therefore, we focused on participant accuracy within the different test trials. The accuracies are shown in Figure 3. We fit a 2x3x4 mixed ANOVA model with between subject factors of age (adult vs. child) and condition (baseline vs. endogenous vs. exogenous) and within subject factor of test type (Standard vs. Det-Only vs. Prob-Only vs. Switch) on accuracy. Analyses revealed a significant 3 way interaction, $F(4.66, 822.83) = 16.994, p < .001, \eta_p^2 = 0.088$ (Greenhouse-Geisser corrected for sphericity). We broke down by age group to further understand the effects.

For adults, we found a significant interaction between test type and condition, $F(4.26, 574.79) = 100.90, p < .001, \eta_p^2 = 0.43$ (Greenhouse-Geisser corrected for sphericity). We further broke down the interaction by condition. Standard test blocks across conditions were all above chance, indicating that adults successfully learned the categories in all conditions (baseline: $t(73) = 22.90, p < 0.001$; endogenous: $t(86) = 30.68, p < 0.001$; exogenous: $t(111) = 45.73, p < 0.001$; adjusted $\alpha = 0.002$ for 2x3x4 test type comparisons).

Pairwise comparisons within the adult baseline condition revealed that standard test accuracy and Prob-Only accuracy were not statistically different. Standard test accuracy was greater than Det-Only and Switch accuracy ($p < .001$), suggesting that adults were using probabilistic features in their categorization decisions. Surprisingly, we found that the reliance on the probabilistic features was overwhelming in the switch condition, with accuracy in Switch block being significantly lower than all other test blocks (pairwise comparison, $p < .001$). Contrary to our predictions, a one sample t-test revealed that Switch accuracy was significantly below chance level (M=0.12, one sample $t(73) = -12.93, p < 0.0001$). Thus, although participants could discriminate the categories above chance level with only the deterministic feature (Det-Only: M= 0.64, one sample t(73)=3.41, p=0.001), in the presence of conflict they relied on the overall probabilistic appearance. We will revisit this novel and surprising finding in the Discussion.

Within the Endogenous condition, there was a different pattern. Standard test accuracy was not statistically different from the Det-Only, but was significantly better than the Prob-Only and Switch ($p < .001$). This suggests that adults in the endogenous condition relied on the deterministic feature overall, since they did as well in the Det-Only condition as in the Standard test block, but not as well when only probabilistic features were visible (Prob-Only). However, they did learn to rely on the probabilistic cue to a slight degree because accuracy was reduced from the standard test to switch trials - when deterministic and probabilistic features were in conflict. Unlike the baseline, however, the switch accuracy remained above chance indicating that the deterministic feature dominated in the decision.

Within the exogenous condition, the Standard test accuracy was greater than the Prob-Only condition ($p < .001$). However, Standard test accuracy was not significantly different from the Det-Only or Switch trials. This suggests that like the endogenous condition, adults relied on the deterministic feature. However, unlike the endogenous condition, there was no reduction in accuracy in the Switch, suggesting that probabilistic features were not used in the decision at all. Thus, the effect of salience was more pronounced that that of prior information in causing adults to learn and use the rule information.

We next investigated differences in the child age group. For children, we found a significant interaction between test type and condition, $F(5.76, 238.94) = 2.546, p = 0.023, \eta_p^2 = 0.058$ (Greenhouse-Geisser corrected for sphericity). We

further broke down the interaction by condition. Within the baseline and endogenous conditions, no test conditions were significantly different from each other ($ps > 0.18$). Pairwise comparisons confirmed that accuracy in the baseline and endogenous conditions were not significantly different from each other ($p > 0.99$). A one sample t-tests of standard test trials within the baseline and endogenous conditions revealed that accuracy was not significantly different from chance level. The baseline was approaching significance ($t(23) = 2.13p = 0.044$), but not when adjusted for multiple comparisons (adjusted $\alpha = 0.002$). Thus, children did not successfully learn to discriminate the categories in the endogenous or baseline conditions.

Within the exogenous condition, however, a one sample t-test of accuracy in the standard test block confirmed that children successfully learned the categories ($t(29) = 4.058, p < 0.001$). Pairwise comparisons showed that only the Prob-Only accuracy was significantly less than all other testing conditions ($ps < .01$). One sample t-test revealed that accuracy in Prob-Only was not different from chance ($M = 0.4875, t(29) = -0.37, p = 0.71$). This suggests that children learned the category in the exogenous condition, and used the deterministic feature consistently in making their categorization judgements. When only probabilistic features were available, their accuracy dropped to chance performance.

These results are consistent with our prediction that children would do better learning the categories when salience draws their attention to the rule. As in studies mentioned above and similarly to adults in this study, the salience was overwhelming such that the probabilistic features were ignored in the decision. We should also emphasize here that the salient cue was not shown during the testing trials. Therefore, the information that was conveyed through a salient cue manipulation during training transferred to the non-cued test trials.

Discussion

We have presented a study to understand the conditions under which children can learn to use key information for categorizing objects. Decades of research has suggested that young children prefer holistic object information or overall similarity when categorizing objects, with a few more recent studies demonstrating facilitative effects of attentional manipulations on learning. We sought to disentangle the role of instruction and salience in these demonstrated effects. We make two important novel contributions.

First, we demonstrated a case where adults who learned the R+FR categories relied on the family resemblance features when categorizing. This demonstration is quite novel and surprising because decades of research has repeatedly shown that adults will use the rule features and ignore probabilistic features when learning categories with feedback. Furthermore, the R+FR category structure we employ is nearly identical to those used previously (See Table 1 vs. Deng & Sloutsky, 2012, Table 1). We suspect that the result stems from an important property that is obscured by the symbolic category structure representations. Namely, whereas the dimensions in other studies have distinct feature values, the family resemblance dimensions in our category structure denote highly redundant features. In other studies an exemplar with symbolic structure (000) corresponds to an object with distinct hands, feet, and body shapes that are all consistent with category A. In our study, however, a (000) corresponds to an exemplar having many blue features.

We suspect that adults in the baseline condition are sensitive to the redundancy, and learn to use *mostly blue circles* or *mostly red plus signs* as a type of higher order feature. There is evidence that adults do learn higher level features from primitives when learning categories and concepts (Schyns, Goldstone, & Thibaut, 1998). Furthermore, children are less consistent in their creation of these higher order features - possibly explaining their lesser reliance on this simple strategy. Interestingly, either attention cue made adults revert to the typical pattern of rule-based responding. Further research will be necessary to understand what strategy adults use in the baseline condition, and why they rely on the probabilistic features.

Our second main result is that instruction may not be sufficient for teaching young children to rely on key bits of information while simple salience is very powerful in that regard. We made a great effort in the endogenous condition to explain that the information necessary to the task would be in a location denoted by the arrow. We used analogies of body parts to convey the idea, and repeated the instructions as many times as necessary. However, children were not able to learn the categories given all those hints and guidance. In contrast, in the exogenous conditions we gave no hints or guidance. We only told participants that things would blink so that they would not be confused. We gave no instruction about the importance of blinking locations, but children learned that categories successfully and even transferred their knowledge to test trials where the blinking cue was not shown. The point that the knowledge transfers to the non-blinking case is critical because it implies that salience can be used to bootstrap concept learning for application in novel situations. Furthermore, detailed instruction may not be sufficient to overcome the limitations of limited executive function. Thus, our finding offers important insight into answering the question of how we can teaching young children concepts that require them to override their holistic predisposition.

Acknowledgments

This research was supported by NSF grant (BCS-1323963) to Vladimir Sloutsky.

References

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. doi: 10.1016/j.tics.2012.06.010

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*, 105–199.

Deng, S., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, *23*(2), 178–186.

Deng, S., & Sloutsky, V. M. (2015). The Development of Categorization: Effects of Classification and Inference Training on Category Representation. *Developmental Psychology*, *51*(3), 392–405.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of experimental psychology. General*, *139*(2), 319–340. doi: 10.1037/a0019042

Rabi, R., Miles, S. J., & Minda, J. P. (2015). Learning categories via rules and similarity: Comparing adults and children. *Journal of Experimental Child Psychology*, *131*, 149–169.

Rabi, R., & Minda, J. P. (2014). Rule-based category learning in children: The role of age and executive functioning. *PLoS ONE*, *9*(1).

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. doi: 10.1016/j.cogpsych.2004.11.001

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. doi: 10.1016/0010-0285(75)90024-9

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998, feb). The development of features in object concepts. *The Behavioral and brain sciences*, *21*(1), 1–54.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). *Learning and memorization of classifications.* (Vol. 75) (No. 13). doi: 10.1037/h0093825

Sloutsky, V. M. (2010). From Perceptual Categories to Concepts: What Develops? *Cognitive Science*, *34*(7), 1244–1286. doi: 10.1111/j.1551-6709.2010.01129.x

Smith, L. B. (1989). A model of perceptual classification in children and adults. *Psychological review*, *96*(1), 125–144. doi: 10.1037/0033-295X.96.1.125

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, *24*(2), 279–298.

Creative Interaction with Blocks and Robots

Daniel Smithwick (Djs2@Mit.Edu)
MIT-SUTD International Design Center
Cambridge, MA 02139 USA

Larry Sass (Lsass@Mit.Edu)
Department of Architecture, 77 Massachusetts Ave
Cambridge, MA 02139, USA

David Kirsh (Kirsh@Ucsd.Edu)
Cognitive Science
La Jolla, CA 92093-0515 USA

Abstract

In order to creatively interact with robots we need to understand how creative thinkers work with objects to explore new ideas physically. Our approach involves comparing the model-making strategies of architects with students to expose the creative extras architects bring to working with physical models. To study this we coded students and architects performing a design task. Architects differed from students along three dimensions. First, architects were more selective; they used fewer blocks overall and fewer variations. Second, architects appear to think more about spatial relationships and material constraints. Lastly, architects more often experiment with re-orientations: they position a block one way to see its relations to its neighbors; they reposition it another way to see how that changes how things look and feel. These findings suggest that designers interact with the material more effectively than students. This embodied know-how is something next generation robots can support and possibly enhance.

Keywords: design thinking; interaction; robotics

Introduction

Much of what has been said of design thinking is about manual sketching activity: it is the means by which designers have 'reflective conversations' with their design (Schon, 1992); it is how designers 'see as' and 'see that' (Goldschmidt, 1991); it situates designers and enables them to 'think on the fly' (Suwa, Purcell, and Gero 1998); and it enhances a designer's ability to perceive visual-spatial features and conceive multiple design ideas (Bilda and Demirkan, 2003). Indeed, much of design thinking initially takes form as exploratory sketching activity.

However, sketching is not always necessary in design activity (Bilda, Gero, and Purcell, 2006) and it is not the only means of exploring a design manually. Architects also make physical models, including: sketch models, diagram models, concept models, massing models, presentation models, and more (Mills, 2011).

Unlike sketching, which involves marking a two-dimensional paper surface, physical model making takes place in three dimensional space and involves different forms of material interaction. For instance, different modeling materials afford different actions: chipboard can be layered, paper can be folded, wood can be milled, concrete can be cast, and so forth.

If sketching is thought of as visual design thinking (Goldschmidt, 1994), model making, with its emphasis on building, assembly and manipulation, ought to be considered physical design thinking – a more tangible, interactionist way of exploring designs.

Framing certain forms of action as enactive thinking is central to theories of embodied and situated cognition (Anderson, 2003). Gesture can facilitate thinking in calculation (Martin, 2005) and problem solving (Goldin-Meadow and Beilock (2010). Body movement enables dancers to probe movement structures in ways inaccessible through observation alone (Kirsh, 2011).

Our objective here is to extend the notion of design thinking to physical model making. What sketching is to visual thinking model making is to physical thinking. To explore this idea we created a simple design world – a blocks world – where model making is abstracted to picking, manipulating, and placing blocks in a configuration on a site. We devise a simple coding scheme that tracks the key material interactions over time. By video recording the design sessions of designers and non-designers we are able to compare their interactions and by using the method of voice aloud protocol analysis we can relate these to what is said during activity. Assuming designers do behave differently than non-designers, our central concern is to elucidate a framework that will reveal these differences and explain them.

Our ultimate motivation is to provide a theoretical and empirical foundation for interactive robotic design tools that enhance the material-based nature of exploratory model making activity. As digital technologies become more physical – more interactive and integrated into the design environment – theories of physical design thinking are needed to guide the development of supportive tools.

Background: Coding Design Activity

Protocol analysis methods have often been used to codify actions in design activity and correlate them with a designer's thoughts.

Coding schemes are used to define and count discrete actions within design activity. Many studies are based on Suwa and Tversky's coding scheme (1997) which identified four information categories that architects see and think about while sketching: emergent properties, spatial relationships, functional relationships and background knowledge. Suwa et al (1998) adapted these categories to correspond to the flow of cognitive processes involved in

human cognition: distinguishing physical, perceptual, functional, and conceptual actions. Physical actions include drawing, looking, and gesturing. Perceptual actions involve attending to features, relationships, and making comparisons. Functional actions include relating non-visual information with spatial features e.g., circulation of people through rooms. Conceptual actions involve setting up goals and making value judgments based on domain knowledge.

By codifying the content of a designer's actions researchers have not only been able to speculate on the flow and structure of design thinking but also to compare novice and expert behavior. The differences found suggest there are learned designerly behaviors. Kavakli and Gero (2002) found structural differences in the behaviors of experts and novices. They observed that in terms of the overall number of actions performed during sketching, an expert is more active and productive than a novice. Furthermore, experts increased their rate of action over time while novices decreased theirs. They found that an expert is three times more selective in how many actions they deal with at once, suggesting their design process is more controlled and efficient.

In the current work, we are interested particularly in how physical action in model making drives design thinking and in what ways this may differ between experts and novices. Looking more closely at the kinds of physical actions defined by Suwa et al (1998), however, we find their action types ill-suited for describing material interactions in model making. They distinguish these seven:

- Revise the shape, size or texture of a depiction
- Create a new depiction
- Trace over a depiction on same/new sheet
- Write sentences or words
- Depict a symbol
- Look at previous depiction
- Move a pencil/depiction

In physical model making one may perform these actions in the course of making a model, e.g. to depict a shape to be cut out of paper, but this would be a sketching action plus some other physical action. Strictly speaking, model making means working with material in hand. The actions possible depend largely on the type of material, opening up a wide range of interactions: folding, twisting, laminating, stacking, sorting, cutting, milling, pouring, and so on. In addition, model making commonly involves the application of skill and technique with tools as diverse as knives, drill presses, laser cutters, and 3D printers. This makes model making as a general activity cumbersome to analyze in terms of a small set of simple discrete physical actions.

This complexity may explain why little research has been conducted on model making as a kind of physical thinking process. Studies that do examine model making are highly constrained. For example, to explain the benefit of hands-on model making in engineering design, Lemons et al (2010) had participants construct models with Lego bricks. They focused on their subjects' accounts of what they were thinking, using verbal protocols exclusively and did not code the different physical actions performed such as joining, disjoining, rotating or sorting bricks.

Methodology

Experiment Setup

In our experiment 9 participants were given 15 minutes to build a physical model of their dream house by arranging blocks on a wooden site model (Figure 1). Participants were supplied with 44 3D printed parallelepiped-shaped blocks and free to use as many or as few as they wished. The parallelepiped shape enabled fairly complex assemblies with varied spatial relationships. No other tools or medium (e.g. no pencil or paper) was allowed. Prior to the start of the experiment subjects were given a minute or two to familiarize themselves with the look and feel of the blocks and told that during the experiment it would be helpful if they voiced aloud their thoughts concerning their design or their process as they manipulated blocks. Video/audio was captured looking down on the site model and photographs were taken throughout each session.

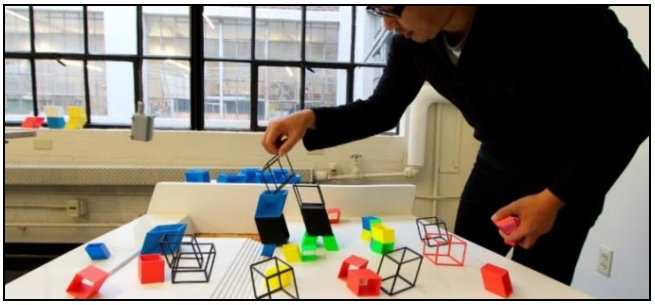


Figure 1: Experiment site model and 3D printed blocks.

Participants

Of the nine participants three were architects with four to eight years of professional practice as well as teaching experience at the graduate level. The six student participants were spilt into two groups. Three students were undergraduates majoring in architectural design, and three were undergraduates from non-design related departments.

Table 1: Study Participants

Experienced architects	3	3M, 0F
Students in architecture	3	1M, 2F
Students in non-design fields	3	2M, 1F
Total	**9**	**6M, 3F**

Hypothesis

Our hypothesis was that in an exploratory design task requiring physical manipulation, there will be significant differences between the way architects and students interact with their materials. The nature of these differences should be discernable through a coding scheme and connect in some way with the strategies different skill levels rely on.

Coding Scheme

Our coding scheme treats interaction with a block to have three parts: a picking up step, a manipulating step and a placement step. Within this broad activity structure or 'framework' we distinguish four types of actions and then seven distinct interaction sequences.

Primitive Interaction Types There are four interaction types: *adding* blocks to a configuration, *subtracting* blocks from a configuration, *modifying* blocks within a configuration, and *relocating* blocks across the site. For our purposes we abstract from details of how a block was picked up, manipulated and placed and focus instead on where it came from and where it ends up. Each of these four types is defined by the locations of the picking and placing actions as identified in Table 2. For each participant we counted each occurrence of the action type and measured its duration.

Interaction Sequences An interaction sequence is an ordered set of primitive interactions over time. There are two types of interaction sequences:

- *Linked interactions* occur when the same block is acted on two or more times in a row. For example, when a subject adds a block to a configuration and without really putting it down, does a modify action (turns it) we say the two actions are linked. Or when a block is put down then taken away, even if there is a delay, two actions are linked.
- *Non-linked interactions* occur when subjects pick up different blocks. When the actions are different the sequence is a simple non-linked sequence. When the same action is performed three or more times on different blocks – e.g. Add-Add-Add – we call the sequence repetitive.

See Table 3 for all interaction sequence definitions.

Table 2: Definition of interaction types

Interaction Type	Pick Location	Place Location	Description
ADD	Site	Configuration	Participant adds block from the site to the configuration
SUBTRACT	Configuration	Site	Participant removes block from config-uration and places on site
MODIFY	Configuration	Configuration	Participant adjusts blocks within con-figuration only
RELOCATE	Site	Site	Participant moves block across the site

Table 3: Definition of Interaction Sequence Types

Interaction Sequence	Name	Interactions Involved
Non-linked Repetitive	Manage	Relocate-relocate-relocate
	Disassemble	Subtract-subtract-subtract
	Assemble	Add-add-add
	Explore	Modify-modify-modify
Linked	Test:Reject	Add-modify-subtract
	Test:Accept	Add-modify
	Test:Eject	Modify-subtract

Results

Despite our small sample size there were important differences between the three participant groups. At a structural level there were clear differences in the block count, block variation, and block arrangement of the final model. At a protocol level there were suggestive differences but variance in the quantity of talk aloud results prevented us from finding anything more than anecdotal differences. At an interaction level, however, there were clearer differences in the primitive interaction counts as well as trends in the interaction sequences across the three groups.

Dream House Models

Figure 2 shows renderings of the participants' final models. Each rendering is an orthographic projection of the model.

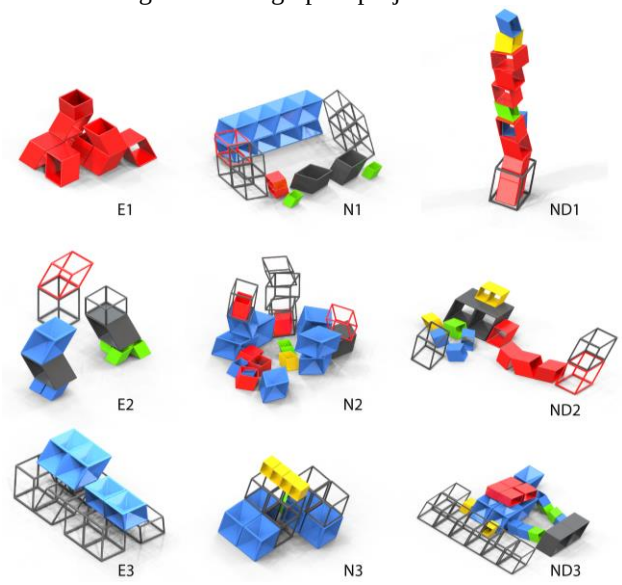


Figure 2: Renderings of models by architects, left; novices, middle; and non-design students, right.

Block Count Architects tended to use fewer blocks in their completed models than the other two groups. Mean number of blocks used: architects = 11.0, novices = 17.3 non-designers = 19.0. Novices and non-designers tended to keep adding blocks to their configuration until the supply ran out. Experts seemed to identify a limited set of preferred blocks, e.g. all red blocks, and then worked with those alone.

Block Variations Similarly, architects used fewer types of blocks. Of the ten distinct types of blocks: architects used 3.3 types (mean) in their final model, novice student designers used 6.3 (mean); and non-design students used 7.3 (mean).

Verbal Protocol Analysis

To explain the differences in behavior we looked at the content of what was said to see if there were differences in design approaches of the different groups. The key difference we found is that architects take a more abstract approach, thinking in broader architectural concepts than students (as one would expect), whereas students think less about the spatial and structural properties of shapes and more concretely about the function of a placement in an emerging structure. Non-architects seem to explore less and in a less abstract way.

For example, one novice design student (see subject N1 in Figure 4) described her process as one of using blocks to represent typical household features – specific functional units:

> "I'm using this [block] as the entrance…and I'm going to use these [blocks] as the grass and flowers…and these [wireframe blocks] will be windows"

A non-design student (ND3 in Figure 4). voiced a similar approach, expressing practical concerns about the house, commenting that:

> "It'd be kind of cool to have a wall of windows along the side of the house…maybe the south facing side where there's a lot of sun"

Furthermore, ND3 was thinking about particular rooms and how to connect them:

> "This is a wide open living room, here's an open hallway on the side of the building…the little blocks I'm treating as hallways and maybe stairs, this is the kitchen, here is a studio…there's bedrooms on the second floor."

Architects had a different approach. They commented on the how moving shapes around facilitated their exploration of the spatial relationships between blocks. E2 started off his design session by saying:

> "Using the blocks, I am creating face-matching walls."

He was giving himself a constraint – joining the blocks face-to-face – and seeing what relationships became apparent.

E1 remarked very early on in the session:

> "I like that these [the red blocks] create an 'outdoor/indoor'…a kind of exterior spatial definition and an interior spatial definition…it's super strong…and you don't have to do much to let them do that…which is nice" .

He was thinking sculpturally of the external shape of his model in relation to the site in which it is situated, and simultaneously thinking about how the blocks seem to define an interior space that is confronted differently by someone on the inside. The same architect emphasized that he conceptualized the inside too in a more abstract manner,

> "I can't go through it and say what's my bedroom, what's my living room…but as a plain figure I like it and can imagine it occupied in many ways."

Architects seem to use the blocks to visually and kinesthetically fish for abstract structural ideas and relations. They don't have a clear idea of what they might be making first and then assemble blocks to see if it works. Rather, they use the blocks to throw up ideas that they can evaluate as interesting or not.

Interaction Results

To determine whether expertise was revealed in other quantitative measures we first coded each participant's video session in terms of primitive interactions (Add, Subtract, Modify, or Relocate), looked for differences in summary statistics, and then analyzed sequences. In Table 4 we show the number of times each action was performed by a subject and the percentage of all actions that action represents. E.g. E2 performed 27 Add actions; these accounted for 24.5% of all of his interactions.

There was huge variance. Total number of actions for the nine subjects range from 44 to 162. These counts vary across and within each group and depend largely on the amount of time spent in the task. Total time spent acting on blocks for each participant ranges from 113 to 429 seconds.

We did a Chi-Square analysis to test the null hypothesis that that there is no significant relationship between participant group and interaction count. The result is $x^2(6, n=9) = 32.22$ $p < .01$. This low probability is sufficient to reject the null hypothesis. But what is the relationship?

Owing to the small sample size and significant individual differences results were suggestive but inconclusive at the summary level. Figure 4 reveals that both architects and novice architectural students perform Modify actions more often than Relocate actions. For instance, they will position a block one way in relation to its neighbors and see what it looks like; reposition it another way and see what that looks like, and so forth. This has the effect of displaying more block relationships. Non-design students, on the other hand, interact with blocks more often on an individual basis, isolated from a developing configuration. Non-designers seem to experiment less.

Interaction Sequences

Differences between architects and all other subjects show up more clearly when we look at sequences of action. Overall, architects spend a greater percentage of their block activity performing linked interactions than both novices and non-designers, and suggestively novice designers likewise perform more linked actions than non-designers: architects 60.7%, novices 50.8%, and non-designers 43.8% (Figure 5). That is, action spent thinking and manipulating the same block varies with expertise, the more expert the more actions on the same block. This is consistent with the tendency to experiment. Looking Figure 6, there is a clear trend correlating expertise and the tendency to test and reject a placement, and test and eject a placement.

Table 4: Interaction primitive results per participant by group

Interaction Type	Architects			Novices			Non-designers		
	E1	E2	E3	N1	N2	N3	ND1	ND2	ND3
# of ADD actions	11	27	16	42	36	26	21	22	33
% of total	25.0	24.5	28.1	28.6	22.2	24.5	39.6	23.9	31.1
Time (s)/action	1.0	1.0	1.0	1	1.1	1	1.2	1.1	2.5
# of SUBTRACT actions	4	12	4	9	13	9	6	5	7
% of total	9.1	10.9	7.0	6.1	8.0	8.5	11.3	5.4	6.6
Time (s)/action	1.0	1.0	1.0	1.0	1.5	1	1	1	1
# of MODIFY actions	19	39	20	59	90	35	13	31	17
% of total	43.2	35.5	35.1	40.1	55.6	33.0	24.5	33.7	16.0
Time (s)/action	5.5	6.6	3.5	5.0	2.6	4.5	4.9	4.2	2.8
# of RELOCATE actions	10	32	17	37	23	36	13	34	49
% of total	22.7	29.1	29.8	25.2	14.2	34.0	24.5	37.0	46.2
Time (s)/action	1.1	3.5	1.4	1.9	2.0	1.4	2.4	4.7	1.8
TOTAL ACTION TIME (s)	**130**	**410**	**113**	**429**	**340**	**243**	**126**	**335**	**227**
TOTAL ACTIONS	**44**	**110**	**57**	**147**	**162**	**106**	**53**	**92**	**106**

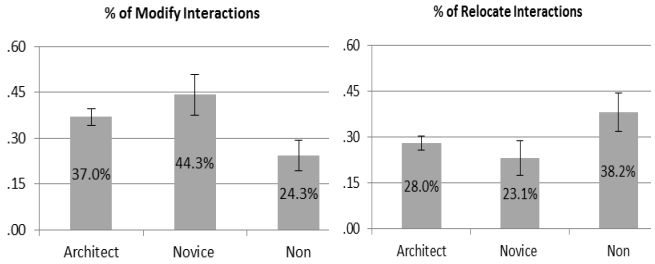


Figure 4: Percentage of Modify and Relocate

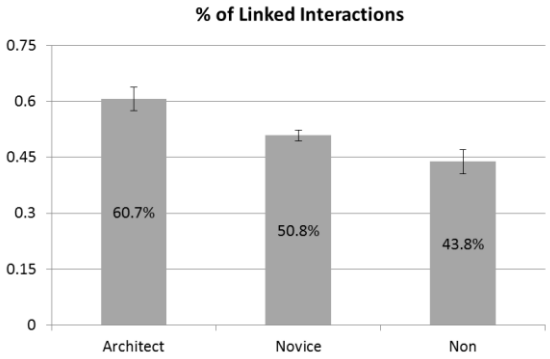


Figure 5: Percentage of Linked Interactions

Architects perform over two times as many Reject interaction sequences and over three times as many Eject sequences compared to non-design students. It seems the architects are more selective when choosing which blocks to keep in the model.

This also correlates with the tendency to devote less activity to managing and assembling blocks. Non-design students perform Assemble interaction sequences seven times as much as the architects. On the whole, architects, more so than the other groups, employ interaction strategies which link longer chains of related interactions.

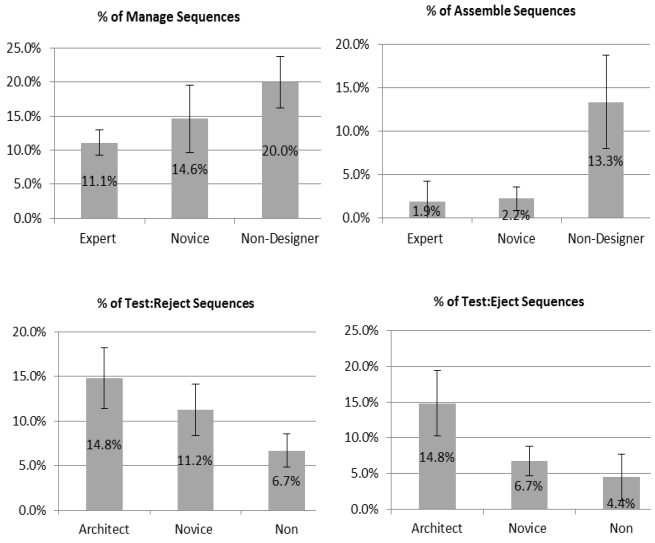


Figure 6: Percentage of Non-linked (Above) and Linked (below) sequence types by group.

Discussion

Robots are becoming increasingly available to architects and designers. So far their role is to help build parts and assemble models. Why not integrate them in the early exploratory stage of design (Figure 7)?

The challenge in bringing robots in as cognitive collaborators and not just as hired hands, is that joint action normally requires collaborators to know what each other is doing. That is still a long way off. But as our study has shown, architectural designers interact with physical material in characteristic ways – certainly in ways that are distinguishable from students. These can be coded in terms of primitives and sequences and as more regularities are discovered a theory of 'thinking through material interaction' may become the framework through which designers may creatively work with robots.

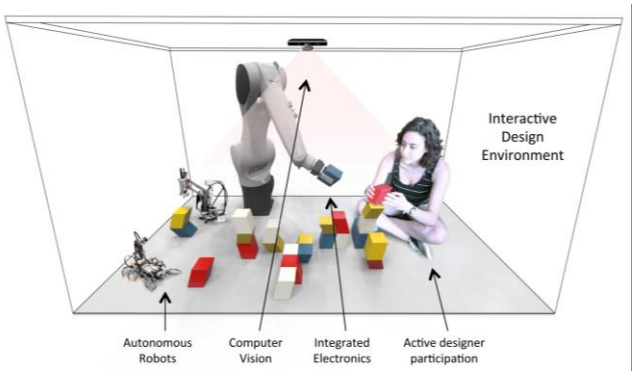


Figure 7: Vision for an interactive robotic arm

Equipped with vision sensors and object recognition functionality, robotic tools could observe a designer's material interactions and track the emerging structure of his or her model. At first the value of this tracking would be to support low cost undoing or duplication. A robot could quickly duplicate interactions or perform iterations and variations of the designer's moves, enabling rapid prototyping of variations or construction of symmetric configurations. But additionally, a robot can be used to compute complex arrangements that are hard to imagine and difficult to manually configure, such as assembling block arrays or reflections. This will be especially beneficial as materials themselves become more designed and complex. A robotic partnership could expand the visualization and manipulation capacities of architects and give room for new thoughts.

Conclusion

In order to understand how architects work with physical objects to visually and kinesthetically explore design possibilities, we gave architects, architecture students and non-design students a task to build a model of their dream house. Their design environment was a 4' square site with contours and a set of forty-four different shaped, sized and colored blocks. We devised a coding scheme to analyze the material interactions of designers and this world based on three basic actions: picking, manipulating, and placing blocks on the site. Using this scheme we were able to distinguish the interactions exhibited by architects from novice students and non-design students.

Our main observation is that architects more than student architects seem to materially explore their design in an abstract manner at first, looking for interesting structural relations between the site and collections of blocks. They also consider other architectural concepts such as negative space. Student architects and non-design students are far more functional and pragmatic, placing blocks down to serve as living rooms or kitchens, and other familiar parts of houses. This difference in concern leads architects to experiment with space differently.

Our second observation is that the amount of activity spent experimenting and fishing for new ideas, seems to correlate with design experience. Architects more than novices and more than non-design students use more time exploring through manipulating and reflecting on the possibilities of each block, than on placing a block down and rushing off to get another block.

A blocks world is inevitably a limited design world. We believe our results are suggestive enough, however, to show that subjects use blocks to think with. If our goal is to develop robotic cognitive assistants frameworks such as ours will need to be further developed and elaborated.

Acknowledgments

The authors would like to thank the MIT-SUTD International Design Center for funding this research.

References

Anderson M (2003). Embodied Cognition: A field guide. *Artificial Intelligence*:1-40.

Bilda Z, Demirkan H (2003). An insight on designers' sketching activities in traditional versus digital media. *Design Studies* 24:27-50.

Bilda Z, Gero J, Purcell T (2006). to sketch or not to sketch? That is the question. *Design Studies* 27:587-613.

Goldschmidt G (1991). The dialectics of sketching. *Creativity Research Journal* Vol 4 No 2:123-143.

Goldschmidt G (1994). On visual design thinking: the vis kids of architecture. *Design Studies* Vol 15 No 4:158-174.

Goldin-Meadow S, Beilock S (2010) Action's Influence on Thought: The Case of Gesture. *Perspectives on Psychological Science* 5(6):664-674.

Kavakli , Gero J (2002). The structure of concurrent cognitive actions: a case study on novice and expert designers. *Design Studies* 23:25-40.

Kirsh D (2011). Thinking with the body. *Proceedings from the 30th Annual Cognitive Science Society*.

Lemons, G (2010). The benefits of model building in teach engineering design. *Design Studies* 31:288-309.

Martin T, Schwartz D (2005). Physically Distributed Learning: Adapting and Reinterpreting Physical Environments in the Development of Fraction Concepts. *Cognitive Science* 29:587-625.

Mills C (2011). *Designing with Models*. Wiley New York.

Schon D (1992) Designing as reflective conversation with the materials of a design situation. *Behavior Research* Vol 5 no 1:3-14.

Suwa M, Tversky B (1997). What do architects and students perceive in their design sketches? A protocol analysis. *Design Studies* 18:385-403.

Suwa M, Purcell T, Gero J (1998). Macroscopic analysis of design processes based on a scheme for coding designer's cognitive actions. *Design Studies* 19:455-483.

A Hierarchical Probabilistic Language-of-Thought Model of Human Visual Concept Learning

Matthew C. Overlan (moverlan@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Steven T. Piantadosi (spiantadosi@bcs.rochester.edu)
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY USA

Abstract

How do people rapidly learn rich, structured concepts from sparse input? Recent approaches to concept learning have found success by integrating rules and statistics. We describe a hierarchical model in this spirit in which the rules are stochastic, generative processes, and the rules themselves arise from a higher-level stochastic, generative process. We evaluate this *probabilistic language-of-thought* model with data from an abstract rule learning experiment carried out with adults. In this experiment, we find novel generalization effects, and we show that the model gives a qualitatively good account of the experimental data. We then discuss the role of this kind of model in the larger context of concept learning.

Keywords: Probabilistic language of thought, Bayesian inference, abstract rule learning, computational model, induction, generalization, behavioral experiment

Introduction

A foundational question about human cognition is how we learn as much as we do from input that is often extremely limited. From a few or even just one example, we can make powerful and accurate generalizations. In language learning, for example, there have long been debates about how we learn the complex grammatical structures that we do, given input that is relatively sparse. The ability to learn complex mental representations on the basis of small data sets is also at work in other cognitive domains such as visual perception (Marr & Nishihara, 1978) and causal reasoning (Gopnik & Sobel, 2000) . These mental representations are important because they abstract and summarize the regularities in our environment. By generalizing knowledge gained in specific settings, they allow us to act in novel settings. For the field of cognitive science, a key question is, since there are infinitely many generalizations that are consistent with a finite input, why and how do we generalize in the ways that we do?

To understand people's generalizations, cognitive scientists often debate the merits of statistical versus rule-based approaches. Statistical approaches are appealing because the statistical regularities of data can often be learned relatively easily. Advocates of rule-based approaches argue however, that statistical regularities are inadequate and that, without explicit rules, it is difficult to explain the full richness and complexities of people's generalizations (Marcus, 1999).

As pointed out by others, statistical and rule-based approaches are not mutually exclusive, but rather can be profitably combined (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Aslin & Newport, 2012). For example, in the field of artificial intelligence, the natural language processing community has long used models that use statistics to infer structured, arguably rule-based representations of syntax (Manning & Schütze, 1999). Within the cognitive sciences, the Rational Rules model (Goodman, Tenenbaum, Feldman, & Griffiths, 2008) showed how we can account for human performance by considering rule learning as Bayesian statistical inference over a structured rule space. Hybrid statistical/rule-based models are sometimes referred to as "probabilistic language of thought" models.

Work in developmental psychology has strongly suggested that even infants generalize in ways that go beyond simple statistical co-occurrences. Marcus, Vijayan, Bandi Rao, & Vishton (1999) showed that seven-month-old infants can learn to recognize sequences of syllables that follow an *ABA* pattern like "ga ti ga" and "wo fe wo", where the first and third syllables are the same but differ from the middle syllable. Gerken (2006) later showed that when infants are exposed to stimuli that are consistent with both a broader (e.g., *ABA*) and a narrower generalization (e.g., *AxA* where x is a specific syllable, not a class or set of syllables), the infants tend to prefer the narrower generalization. This ability to seemingly learn abstract rules from small data sets has even been glimpsed in non-human animals (van Heijningen, Chen, van Laatum, van der Hulst, & ten Cate, 2013).

To account for these types of experimental findings, Frank & Tenenbaum (2011) modeled rule-like patterns with strings that have a symbol for each token in the pattern. The symbols indicate whether the token in a given position is a particular token x, (is_x), the same as another token at position n ($=_n$), or a wildcard (i.e., the symbol could be any token). They modeled learning as Bayesian inference over these structures: $P(h|D) \propto P(D|h)\ P(h)$, where h is a hypothesis representing a specific choice of symbols and D is the observed data. The likelihood $P(D|h)$ is the probability of obtaining the exemplars in D via independent random draws from the set of all strings consistent with h. Importantly, this likelihood follows the "size principle" (Tenenbaum & Griffiths, 2001). If h is a broad hypothesis that is consistent with many possible sets of strings, then obtaining the specific set D from h is small (i.e., $P(D|h)$ is small). In contrast, if h is a focused hypothesis that is consistent with relatively few possible sets of strings, then obtaining D from h is large (i.e., $P(D|h)$ is large). Consequently, likelihood functions that follow the size principle fa-

vor focused hypotheses over broad hypotheses. In Frank and Tenenbaum's model, hypotheses were *a priori* equally likely (i.e., $P(h)$ was a uniform distribution).

Their model gives an impressive account of findings in the literature for abstract rule learning across several domains. Although this work is an important early step in developing a probabilistic language-of-thought account of human generalization, it leaves open many important questions. Their model is limited in the sense that it only includes the bare machinery necessary to account for the specific findings that they consider. To what extent can their model, or rather their general theoretical framework, serve as a foundation for a richer and more broadly-applicable model providing a more comprehensive account of generalization? What is the full range of rules that people might learn and that cognitive models will need to account for? As noted by Frank and Tenenbaum, people have built-in biases for certain hypotheses over others. What are those biases, and how can they be included in a cognitive model?

In light of these outstanding questions, we developed a new probabilistic language-of-thought model for rule learning. This model uses a two-level generative process for explaining items in a data set. At the top level is a stochastic generative process for generating rules. As explained below, the generative process in our model is a probabilistic context-free grammar, and this grammar generates rules. At the bottom level, each rule is a stochastic generative process for generating data items. An innovative aspect of our model is that rules are themselves stochastic generative processes. Because data items are generated stochastically from rules, and because rules are generated stochastically from a probabilistic grammar, the overall generative process forms a hierarchy.

To our knowledge, only one other computational model of concept learning employs such a structure. Lake, Salakhutdinov, & Tenenbaum (2015) created a model of handwritten character recognition that employed a two-level generative model. The top level defined a distribution over the abstract, symbolic representation of a character (the *type*), and then given that specification, the bottom layer defined a distribution over concrete instances of that character as visual strokes (the *token*). Our work relates to and extends this work by casting a hierarchical model in a more general context. The Lake et al. model is highly customized for its domain, so it is unclear how to apply insights from that work to other domains, except at the broadest conceptual level. Our model, however, is built upon the more general Language of Thought framework. Since this framework has already been successfully applied to other domains, and since models in this framework only require the specification of very general primitives, our work is much more readily adaptable to other domains. This also makes for a much more plausible cognitive explanation, as the learning system requires far less manual engineering.

The model is a natural evolution of prior work on rule learning. Previously, theoretical progress was made by incorporating stochasticity into the rule-learning process, and here we incorporate stochasticity into the rules themselves. The remainder of the paper is an exploration of this idea in an abstract rule learning task in a visual domain.

Experiment

We conducted an experiment with adults to test their ability to learn rule-based visual concepts from a small number of examples. Our visual stimuli were part-based 3D objects where the parts act as tokens in an abstract rule (see Figure 1). There were two groups of subjects (30 subjects per group). For one group of subjects, the experiment used the rule *ABA* (as in the experiment by Marcus et al. discussed above). For the other group, the experiment used the rule *xBB* (as in the experiment by Gerken discussed above; *x* is a specific token that is identical in all exemplars).

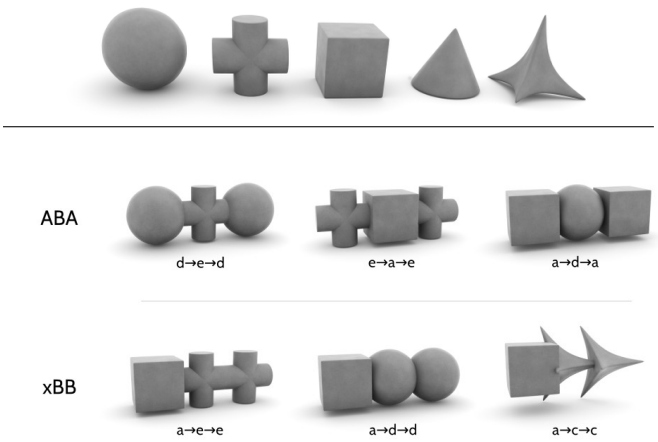


Figure 1: On top, the five parts used in the experiment. Participants viewed these during the instructions phase. On the bottom are the training exemplars for *ABA* and *xBB* conditions.

The experiment was web-based, carried out on Amazon Mechanical Turk. All subjects were US residents over the age of 18. To eliminate the possibility of order and experience effects, each subject participated in only a single condition. The experiment consisted of three stages: an instruction stage, a training stage, and a testing stage. As part of the instruction stage, participants were shown all five possible part shapes. Following the instruction stage, subjects participated in a training stage where they were shown three exemplars from a category. Each subject was allowed to view the exemplars for as long as he or she wished. Training was followed by testing. During testing, subjects were shown an array of 24 test items. Test items had the same general structure as training exemplars (three parts arranged linearly), but differed in which parts occupied each position in an item. Participants chose 'yes' or 'no' for each test item to indicate whether it belonged to the same category as the training exemplars. The exemplars remained available for viewing at the top of the

web page for the duration of the test stage. At least one exemplar was present in the test items. If a participant answered 'no' to that item, his or her results were excluded from the analysis (1 and 2 subjects were excluded in the *ABA* and *xBB* conditions, respectively).

Cognitive Model

In our implementation, concepts or hypotheses begin as lambda calculus expressions. Lambda calculus is a form of logic that is a universal model for computation (i.e., it is equivalent in power to a Turing machine). It characterizes computation using function abstraction and application via variable binding and substitution. For ease of readability, we present these expressions here as procedural "computer programs". These programs construct objects by sampling parts from a fixed alphabet and then combining those parts in a spatial order. They do so using simple set operations, such as removing an element from a set. They can also abstract over parts by assigning them to variables which can be reused. An example program is the following:

```
let x1 = sample(A)
let x2 = sample(A − x1)
output x1 → x2 → x1
```

This program generates the set of objects following an *ABA* pattern. The object is constructed by first randomly sampling part x_1 from the full alphabet, then randomly sampling x_2 from the set consisting of all parts except x_1, and finally combining those parts in the order $x_1 \rightarrow x_2 \rightarrow x_1$. The arrows specify the spatial relationship between parts. Although repetition detection is not explicitly built in, it arises as a natural consequence of the ability to form variable abstractions.

We model learning by assuming that people select the most probable rule or hypothesis h given the set of training exemplars D: $\text{argmax}_h P(h|D)$. The posterior distribution over h can be calculated using Bayes' Rule: $P(h|D) \propto P(D|h)P(h)$. This expression has a natural interpretation in our framework, with the two probabilities corresponding to the two levels of the hierarchy.

The likelihood $P(D|h)$ is the probability that hypothesis h generated the training exemplars in data set D. Assuming the exemplars are drawn independently with replacement (known as the "strong sampling hypothesis" (Tenenbaum & Griffiths, 2001)),

$$P(D|h) = \prod_{d \in D} P(d|h)$$

where d is an individual exemplar. Note that each h, as in the example above, is itself a stochastic generative model. Therefore it naturally defines a distribution over its outputs. Consider "running" the example program repeatedly. Each run produces an independent output that depends on the randomly sampled tokens. $P(d|h)$ is the limiting distribution over those outputs.

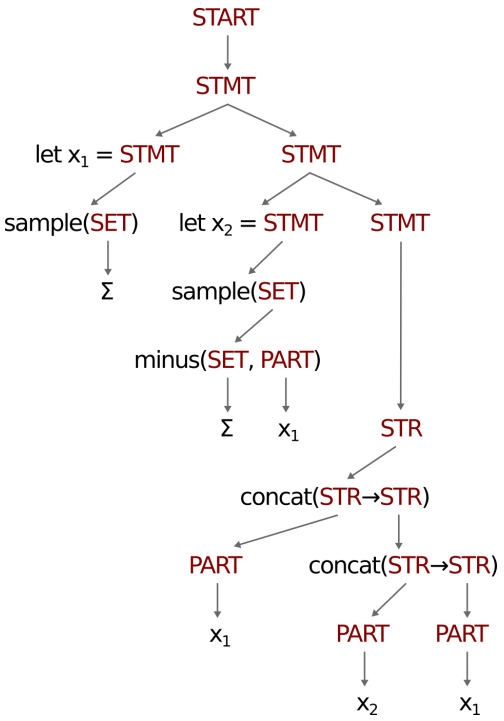


Figure 2: Parse tree for the example program discussed in the main text.

The prior distribution $P(h)$ is the prior probability of hypothesis h. By employing the language-of-thought framework, it too has a natural interpretation. Consistent with earlier language-of-thought models, our model implements the idea that hypotheses are language-like in that they are compositional. Just as sentences are structures built out of words, our model's hypotheses are structures built out of primitives.

This structure is specified by a probabilistic context-free grammar G which defines the syntax for how primitives can be combined. For this experiment, we provided primitive functions *sample(SET)* which samples uniformly from a set, *set_minus(SET, PART)* which removes a part from a set, and *concatenate(STR, STR, ...)* which concatenates strings (which are in turn made up of parts). These primitives are in addition to variable abstraction, which is an inherent property of hypotheses by virtue of their lambda calculus core.

Because the grammar is probabilistic, it defines a distribution over the structures it generates. Each non-terminal in the grammar has an associated distribution that specifies the probability that a production rule will be used to expand that non-terminal. The prior probability of hypothesis h is the product of the probabilities for each of the production rules used in constructing h. In other words, if $\mathcal{T}$ is the parse tree for h and r_i is a rule in this tree, then

$$P(h) = \prod_{r_i \in \mathcal{T}} P(r_i|G).$$

Note that this prior distribution implements a form of Occam's Razor. Since each probability $P(r_i|G)$ is less than one,

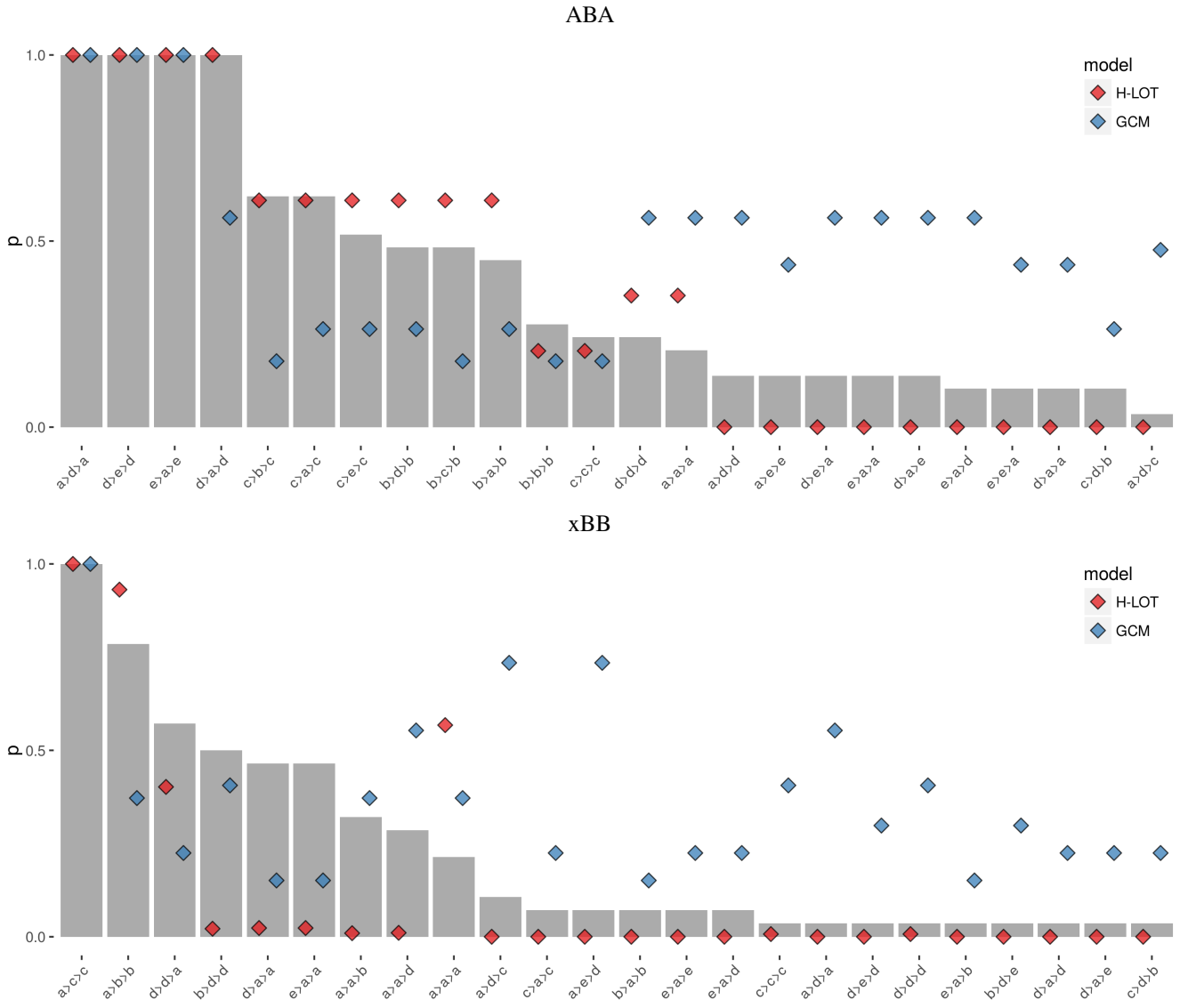


Figure 3: For each test item, the graphs show the probability that subjects judged the item to be from the same category as the training exemplars (plotted by the gray bars in each graph) compared to the probability predicted by our model (labeled H-LOT) and the GCM model. The top and bottom graphs are for the *ABA* and *xBB* conditions, respectively.

hypotheses with short derivations tend to have higher probability than those with long derivations. The parse tree for the example program discussed above is shown in Figure 2.

Finally, each hypothesis has an associated vector of variables θ which allow us to model additional factors that influence generalization. For this experiment, we used two variables. One indicated that the hypothesis should be *orientation invariant*, so that an expression that produced the object $a \rightarrow d \rightarrow d$ would also produce $d \rightarrow d \rightarrow a$. The other indicated that the alphabet should only contain parts that have been seen in the training exemplars rather than the full alphabet of parts. Each of these parameters has an associated prior probability, making the full posterior

$$P(h, \theta | D) \propto P(D|h, \theta) P(h) P(\theta).$$

We estimated this distribution in two steps. First we fixed $P(\theta)$ and sampled the discrete variables h and θ using a Metropolis-Hastings sampling algorithm (a type of Markov chain Monte Carlo algorithm). Because we are sampling in a discrete space, we can approximate the full distribution by saving unique samples and then normalizing. In our Metropolis-Hastings algorithm, we used a slightly modified version of the standard tree regeneration proposal distribution by Goodman et al. (2008). Next, since we have no *a priori* information or theory to indicate how strong the prior tendency to generalize to novel parts or to show invariance to orientation should be, we fit $P(\theta)$ via gradient descent. Since

our search space is tractable enough for the initial sampling step to obtain an approximately complete set of samples, we do not need to do any more sampling after fitting θ.

Results

The results are shown in Figure 3. For each test item, these graphs show the probability that subjects judged the item to be from the same category as the training exemplars (plotted by the gray bars in each graph) as well as the probabilities predicted by two different models. For our model, its prediction for a given test item was calculated by summing the posterior probabilities for all hypotheses that produce that item:

$$P(t|D) = \sum_{h,\theta} P(h,\theta|D) I_{ext(h_\theta)}(t)$$

where I is the indicator function and $ext(h_\theta)$ is the extension of (the set of objects generated by) hypothesis h given parameters θ.

We also show results for the Generalized Context Model (GCM) (Nosofsky, 1986), an exemplar model of category learning that is commonplace in the cognitive science literature. The GCM is a similarity-based model; it determines the category membership of a test item based on its similarity (or inverse of distance) to the exemplars. For this domain, a natural distance function would be one that assigns low distances (and thus high similarities) to pairs of objects that have many parts in common, and high distances to those that have parts that differ. This distance function would serve as a useful comparison because, unlike our model, its representation is relatively unstructured, and it does not have internal variables. We chose the Levenshtein string edit distance, which gives the minimum number of insertions, deletions, or substitutions needed to transform one string into another. We gave this distance function a string representation of the objects. For test item t, the predicted proportion of responses is given by

$$P(t|D) = \sum_{d \in D} e^{-c \cdot Lev(t,d)}$$

where c is a scaling parameter that we fit via gradient descent.

Subjects' responses in our experiment showed large variability, as illustrated by the fact that many subject probabilities (see gray bars in Figure 3) are not near 0 or 1. Despite this variability, our model provides a reasonably good account of subjects' responses, particularly in the *ABA* condition. The GCM model performs poorly; people's generalizations in this task reflect the latent structure present in the objects, thereby going beyond simple similarities.

The tables in Figure 4 show the three hypotheses with the highest probabilities according to our model. These results suggest that the model correctly infers the target rules (*ABA* in the top table of Figure 4 and *xBB* in the bottom table). However, people often deviate from the exact patterns given by these rules, sometimes in interesting ways. For example, the model and subjects may or may not generalize to test items containing parts beyond those used by the training exemplars.

ABA

p	hypothesis	extension set size
.4	let x_1 = sample(A) let x_2 = sample($A - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_1$	20
.24	let x_1 = sample(A_R) let x_2 = sample($A_R - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_1$	6
.15	let x_1 = sample(A) let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_1$	25

xBB

p	hypothesis	extension set size
.39	let x_1 = "a" let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_2$	5
.21	let x_1 = "a" let x_2 = sample($A - x_1$) output $x_1 \rightarrow x_2 \rightarrow x_2$	4
.18	let x_1 = "a" let x_2 = sample(A) output $x_1 \rightarrow x_2 \rightarrow x_2$ or $x_2 \rightarrow x_2 \rightarrow x_1$	5

Figure 4: The three top scoring hypotheses for each condition as given by the model, along with their posterior probability (p) and the number of objects each hypothesis generates. A is the set of all parts, and A_R is the set of parts that are present in any of the exemplar objects. The model predicts that learned hypotheses should not strictly follow the size principle, as shown by the non-monotonic set sizes.

In addition, the model and subjects may or may not generalize to the linear reversal of patterns (e.g., *BBx* instead of *xBB*). These creative generalizations suggest that people's concept space may be rich in ways that have only rarely been explored by computational models in the cognitive science literature.

An interesting finding is that people often generalize in ways that violate the "size principle" discussed above. For instance, subjects in the *xBB* condition seem to often infer the rule *ABB*. That is, despite the fact that the same token appeared in the leftmost position of all training exemplars, subjects seemed to infer that the leftmost token can be any part so long as it differs from the tokens appearing in the other po-

sitions. In general, the set sizes that people learned are highly variable. This tells us that people are being strongly influenced by the biases and structure of their hypothesis space, which is captured reasonably well by the biases and structure of the language-of-thought prior.

Discussion

We have shown how our theory of rule-based concepts as structured, generative models provides a framework which can profitably model concept learning in the domain of abstract rules. Our approach has several features that make it attractive as a general paradigm for theorizing in this domain. First, it operates as an *ideal observer* (Geisler, 2003)—that is, it defines optimal behavior under a set of assumptions. The language-of-thought framework has the attractive property that the modeler is forced to make those assumptions explicit; they must be encoded directly in the rules, choices of primitives, and probabilities of the grammar. For instance, the model in this paper assumed two computational primitives: a uniform sample operation and an operation to remove an element from a set. Furthermore, these assumptions are psychologically interpretable. The modeling choices afforded by the framework, such as the production probability of a rule, can typically be mapped directly onto psychological phenomena. For instance, we saw that in our experiment, subjects only mildly penalized the complexity added by the use of set operations. Because the framework allows us to decompose the structure of concepts in these ways, we can identify the relevant dimensions along which to aim further work.

As an example of how the framework gives us a lens through which we can frame analyses, we identify several avenues for further investigation both in the domain of abstract rule learning and in wider concept learning. The data hinted that people may be incorporating primitives other than those that we included in our model, perhaps ones that arbitrarily permute or shuffle tokens, or ones that invert them, swapping parts A and B. Would the incorporation of such primitives improve model performance in this domain, and would those primitives be relevant in other domains? There are several dimensions of variation that may influence generalization—the number of training exemplars, the number of tokens in an object, the number of unique tokens across all exemplars, etc. The probabilistic basis of the model allows it to make predictions along all of these dimensions, but further empirical data is needed to test those predictions. More broadly, as the framework allows us to identify potential representational biases, we can then ask why people have those biases? Are they the result of some deeper computational principle? And do they need to be innate or can those biases be learned?

The work presented here is a proof of concept that a two-level hierarchy of generative models can be a powerful framework for modeling and interpreting human rule-learning phenomena. Thinking of concepts as structured, stochastic rules has promising potential to be a general theoretical tool for investigating concept learning in many contexts and domains.

Acknowledgments

This work was supported by AFOSR (FA9550-12-1-0303) and NSF (BCS-1400784) research grants.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current directions in psychological science*, *21*(3), 170–176.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371.

Geisler, W. S. (2003). Ideal Observer Analysis. *The visual neurosciences*, *10*(7), 12–12.

Gerken, L. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child development*, *71*(5), 1205–1222.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Marcus, G. F. (1999). Do Infants Learn Grammar with Algebra or Statistics? *Science*, *284*(5413), 436–437.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77–80.

Marr, D., & Nishihara, H. K. (1978). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society B: Biological Sciences*, *200*(1140), 269–294.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, *115*(1), 39–61.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *The Behavioral and brain sciences*, *24*, 629–640; discussion 652–791.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science (New York, N.Y.)*, *331*(6022), 1279–85.

van Heijningen, C. a. a., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: which rule? *Animal cognition*, *16*(2), 165–75.

Perceiving Fully Occluded Objects via Physical Simulation

Ilker Yildirim*[1] (ilkery@mit.edu)
BCS, MIT and The Laboratory for Neural Systems, The Rockefeller University

Max H. Siegel* (maxs@mit.edu)

Joshua B. Tenenbaum (jbt@mit.edu)
BCS, MIT

Abstract

Conventional theories of visual object recognition treat objects effectively as abstract, arbitrary patterns of image features. They do not explicitly represent objects as physical entities in the world, with physical properties such as three-dimensional shape, mass, stiffness, elasticity, surface friction, and so on. However, for many purposes, an object's physical existence is central to our ability to recognize it and think about it. This is certainly true for recognition via haptic perception, i.e., perceiving objects by touch, but even in the visual domain an object's physical properties may directly determine how it looks and thereby how we recognize it. Here we show how a physical object representation can allow the solution of visual problems, like perceiving an object under a cloth, that are otherwise difficult to accomplish without extensive experience, and we provide behavioral and computational evidence that people can use such a representation.

Keywords: physical object representations; analysis-by-synthesis; object perception; occlusion; psychophysics

> *The common and almost despairing feeling . . . was that practically anything could happen in an image and furthermore that practically everything did.*
>
> David Marr, *Vision*

Introduction

Object perception is notoriously difficult, in part because the appearance of an object can vary in almost any way. The problem has been studied in neuroscience, cognitive psychology, and artificial intelligence, leading to a loose consensus that object perception can be solved by the brain (or a computer) learning to "untangle" or become "invariant to" sources of variation in the image (DiCarlo, Zoccolan, & Rust, 2012; LeCun, Bengio, & Hinton, 2015). On this account, sensory input is repeatedly transformed, ideally leading to a (biological or artificial) neural code that is diagnostic for a particular object regardless of variation in the image (Riesenhuber & Poggio, 1999).

We study an alternative solution to the object perception problem, which is enabled by a different representation for objects and a different attitude towards variation. The basic idea is to model the causal processes that lead to an observed image, explaining and reproducing image variation rather than attempting to ignore it. More specifically, we take an object to be represented (at least in part) by a set of physical attributes necessary for supporting physical interaction and image generation. We posit that objects are represented, at a minimum, by attributes including three-dimensional geometry; rigidity; mechanical material properties; and optical material properties.

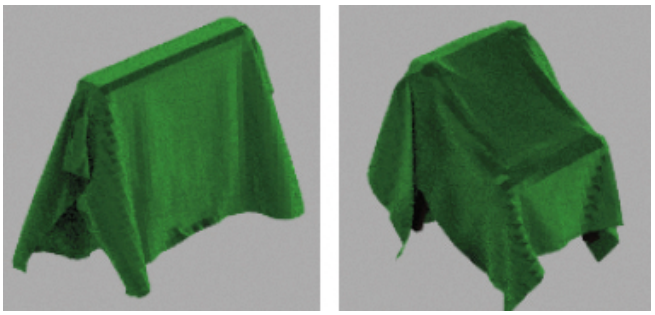


Figure 1: Two objects occluded by cloths.

Consider the covered objects displayed in Figure 1. Both objects are completely occluded, but it is easy to say which of them might be a chair. These images were generated by using a physics simulator to drop a simulated cloth onto two separate 3D models (one a chair), then using a rendering engine to produce images from the resulting scene.

This describes a process for generating the image, but the same process can also be used to interpret an image. When asked which image has a chair in it, we can simulate dropping cloths onto a chair mesh, and compare the rendered results with each candidate (this is an intuitive sketch of a procedure; it can be made precise with Bayes' theorem, which specifies how to turn a forward model into an inverse model. See also (Battaglia, Hamrick, & Tenenbaum, 2013)).

There are several notable differences between the latter approach (which we will call analysis-by-synthesis) and the "consensus" approach (which we will call the invariant features approach). First and most notably, invariant features approaches must learn each kind of scene transformation independently. We explained above how knowledge about chairs and cloths can be combined, in the analysis-by-synthesis approach, to recognize a chair underneath a cloth. By contrast, invariant features approaches cannot directly leverage existing knowledge to recognize the compound object. They must be trained, separately, to discount the cloth in order to recog-

[1] indicates equal contribution.

nize the chair[2]. Second, while invariant features approaches are mostly agnostic to the kinds of image transformations that might exist, analysis-by-synthesis implicitly handles a wide variety of transformations without being explicitly trained or taught to do so.

Analysis-by-synthesis in vision has a long history (Yuille & Kersten, 2006; Tu, Chen, Yuille, & Zhu, 2005), and has recently seen increased attention (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015; Erdogan, Yildirim, & Jacobs, 2015). Our work focuses primarily on two less-studied aspects: First, while most work in object perception studies unoccluded or partially occluded objects, we are interested in objects that are fully occluded, so that the only perceptible effect of the object is on its occluder. Said another way, perceiving objects through cloths requires an observer to do without most or all of the visual information that one normally uses. Second, unlike most previous work, we are interested primarily in how the object representation enables this kind of perception. If objects are represented geometrically in a way that can interact with physics, then the procedure outlined above shows how to solve the cloth task without more training.

The object-under-cloth task is most interesting in the case of novel cloth-object pairs (such as an airplane under a cloth). We predict that both people and analysis-by-synthesis models will perform well on this task, but invariant features models will not without further training. We ran a large scale study to assess how well people can perform the object-under-cloth task. After describing the task and the results of the study, we will evaluate the performance of each of the candidate models, and discuss the implications of our findings.

Experiments

We performed two experiments where participants needed to generalize from a single view of 3D object shown at a canonical view to either a novel view of that object (Experiment 1) or to a fully occluded image of that object again at a novel view (Experiment 2).

Participants

58 participants were recruited from Amazon's crowdsourcing web-service Mechanical Turk. The experiment took about 20 minutes to complete. Each participant was paid $1.50.

Stimuli

The stimuli were generated using a subset of the meshes from the ShapeNet (Chang et al., 2015) database using Blender (Blender Online Community, 2015), a 3D modeling and rendering program. The meshes we used represented objects from five different categories: guns, cars/buses, bicycles, laptops and pillows.

We rendered each mesh in three ways: (1) unoccluded, with texture, and from a canonical viewpoint, (2) unoccluded, without texture, and at a random viewing angle, and (3) fully occluded with a cloth draped over it, and with the mesh randomly rotated.

For (2) and (3), the rotation was sampled from the full sphere with a slight preference around the canonical viewpoint – viewpoint of (1). More specifically, a rotation angle of $\pm 35^\circ$ of the canonical viewpoint on all three axes was 1.5 more likely than the rest of the sphere. For (3), we simulated a cotton-like cloth draped over the rotated mesh for a total of 100 simulation steps, and obtained a rendering of the very last step of the simulation.

We used a total of 197 meshes to generate 100 five-tuples of one study item of unoccluded object rendered at a canonical viewpoint with texture, and four test items consisting of two unoccluded objects without texture each rendered after randomly rotating the meshes, and two objects rendered after randomly rotating each and then occluding with a cloth. The unoccluded study items were never seen twice, but the test items were repeated multiple times, each at a different rotation or viewing angle. On 57 of the 100 tuples, the distractors were of the same category, and 43 of the 100 tuples, the distractors were of different category. Example pairs of unoccluded study items and occluded test items are shown in Figure 2a.

Procedure

Both experiments were match-to-sample tasks where both the study and the test items were presented simultaneously and all stayed on the screen until the participants responded. In Experiment 1 ($N = 27$), the study item was an unoccluded image of an object from a canonical viewpoint; the test items were images of two unoccluded objects after randomly rotating each. Participants had to choose which of the test items contained the study item (Figure 2b).

In Experiment 2 ($N = 31$), the study item was also an unoccluded image of an object; but the test items were images of two objects rotated randomly and occluded with a cloth. Again, participants had to choose which of the test images contained the study item (Figure 2b).

In each of the experiments, participants completed 10 practice trials before moving onto 90 experimental trials. Participants were provided with their running average performance at every 5th trial throughout the experiment.

Results

Results of the experiments are shown in Figure 3a. Participants performance on Experiment 1 (with the unoccluded test items) were high overall (93%).

Participants performance was surprisingly high in Experiment 2 with the occluded test items at 80% (Figure 3a), and as expected, their performance was lower when compared to Experiment 1.

The type of the distractor (whether it is of the same category as the study item or different category) introduced a much stronger decline in performance in Experiment 2 than in Experiment 1 (Figure 3b).

[2]Alternatively, models might be "retrained" or similar; these methods also require more training.

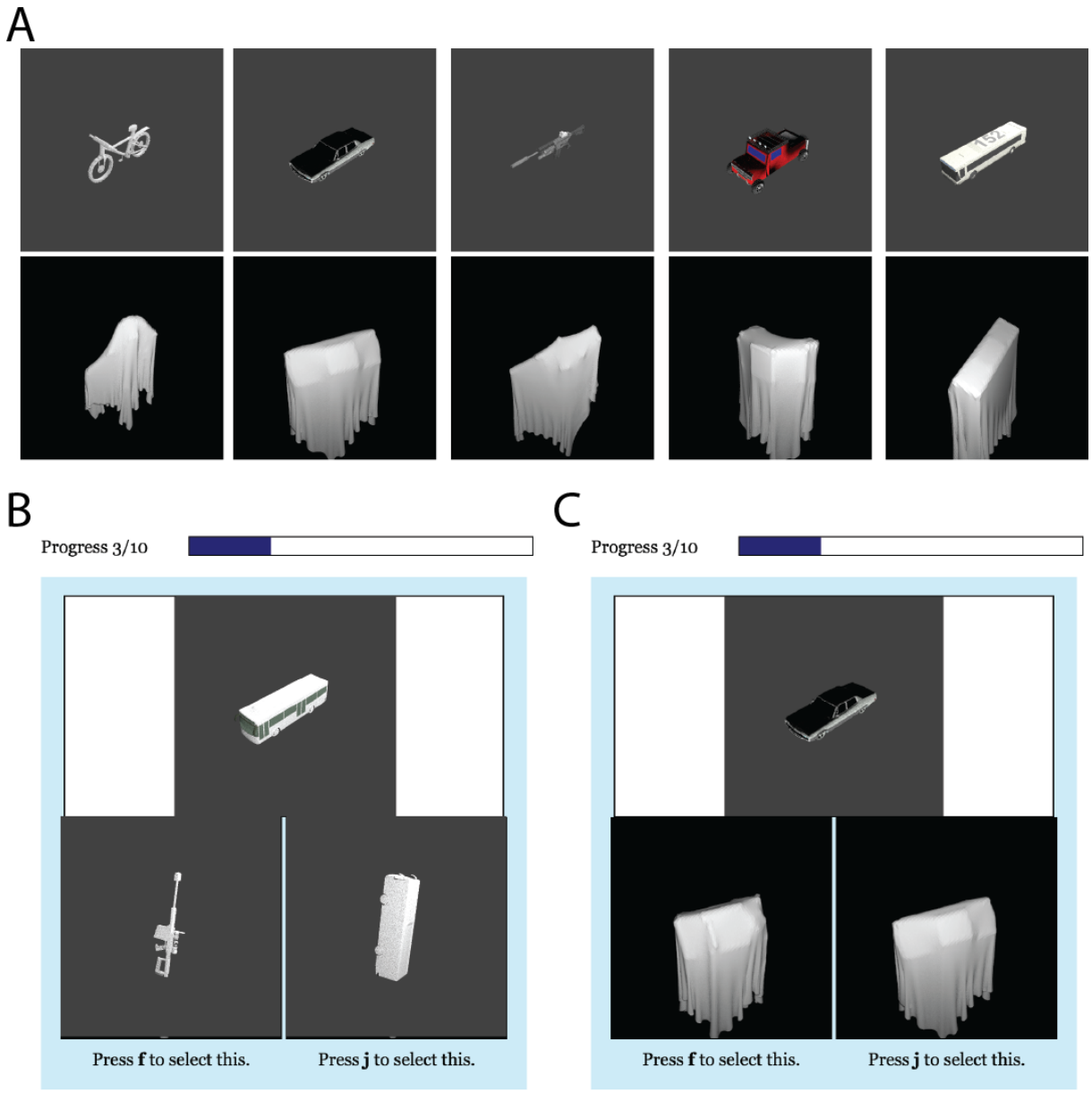


Figure 2: (A) Pairs of images of meshes and simulation results after rotating the mesh randomly and draping a cloth over it. (B) Screenshot of an example trial in the unoccluded experiment. (C) Screenshot of an example trial in the occluded experiment.

Models

We considered two models as potential accounts of our subject's performance: a physics-based analysis-by-synthesis model, and a model derived from features learned by a deep convolutional neural network trained to classify millions of *unoccluded* images.

Physics-based analysis-by-synthesis

We developed a Bayesian computational model that uses knowledge of the causal processes underlying image formation to interpret new images. Aspects of (synthetic) image formation may be divided into two categories: physics-based object factors (e.g., 3D shape, mass, friction, soft-body dynamics, soft-body and rigid-body interaction); and graphics (e.g. rotation and lighting direction). When each of these factors are specified, we end up with a likelihood function that gives the probability of an image given latent parameters, $P(I|\Psi)$.

The model maps input images to the underlying scene parameters using Bayesian inference. Bayes' rule enables us to use the (forward model) likelihood along with a prior distribution (here taken to be uniform) to get the *posterior* distribution of the parameters given the image. The posterior, which includes beliefs about the underlying mesh, is the object of

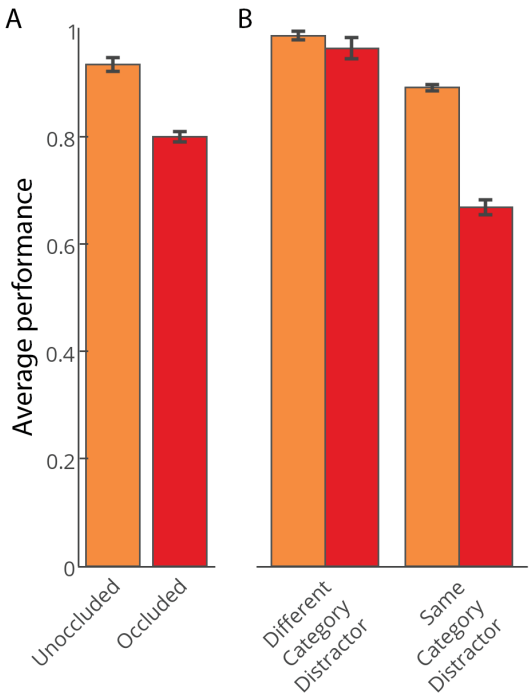


Figure 3: (A) Average performace of the participants in the two experiments. (B) Performance of the subjects divided by whether the distractor is of the same or different category as the study item. Error bars indicate standard error of the mean.

interest for inference tasks.

Bayes' rule states that the posterior is proportional to the likelihood times the prior, that is,

$$P(\Psi|I) \propto P(I|I_s,\Psi)\delta_{p(\cdot)}\delta_{g(\cdot)} \quad (1)$$

where Ψ are the latent variables (e.g. 3D shape, mass, friction, soft-body dynamics, rotation, lighting) that drive the physics and graphics engines; $\delta_{p(\cdot)}$ denotes a physics engine and $\delta_{g(\cdot)}$ denotes a graphics engine (here, we used Blender for both engines); $P(\Psi|I)$ denotes the posterior distribution over physical object representations; I_s denotes the generated image given $\Psi, g(\cdot)$, and $p(\cdot)$; and $P(I|I_s,\Psi)$ denotes the image log-likelihood for a given set of physical object parameters (we assumed a Gaussian likelihood function with $sigma = 0.01$; $N(I|I_s,\sigma)$.

In our implementation, we deterministically assigned the cloth properties such as stiffness, mass and friction to their true values, but assumed uncertainty for the exact 3d shape, S, of the study item and the rotation, R, of the test items. We approximated shape uncertainty as a uniform distribution over the most similar five shapes given a study item, I and its underlying shape, S_I, from the ShapeNet dataset. Similarity between a pair of meshes from the ShapeNet was determined by correlating the concatenated images of each mesh without any texture at four view points – the canonical viewpoint where each mesh is already aligned at in the dataset, and the three orthogonal viewpoints (top, right, front).

The uncertainty about the rotation or pose, R (a vector of length three of Euler angles), was taken to be higher for occluded test scenes than for unoccluded test scenes. For unoccluded test images, we modeled rotation as $R \sim N(R_{true}, 0.025)$, whereas for occluded test images, we modeled it as $R \sim N(R_{true}, 0.1)$, where R_{true} is the true rotation of the test items. We approximated the rotation uncertainty using five randomly chosen samples from the normal distribution.

We took a sampling-based approach to simulate participants from our experiments. For a given simulated participant and trial number, we took one joint sample of shape and rotation from a pool of 25 samples (5 possible shapes $\times$ 5 rotations). Also using the true cloth parameters for occluded trials (i.e., rest of the parameters in Ψ), the model generated a sampled image, I_s, compared it to each of the test images using correlation as its similarity metric, and returned the test image that was more similar to I_s. We simulated 40 subjects in each of the occluded and unoccluded experiments. We report the average performance of these 40 simulated participants.

Deep convolutional network

We also evaluated a pre-trained deep convolutional neural network, which has been shown to succeed at a number of challenging visual recognition tasks (Jia et al., 2014). The network, like all convolutional neural networks, is a feed-forward hierarchical model that performs a series of convolutions and nonlinearities; such models can contain (as does the model we evaluate) millions of learnable parameters. The model was trained to classify objects using the Imagenet (Deng et al., 2009) dataset; it is representative of a wide class of neural network models that find application in computer vision.

Results

Figure 4 show the performance of the two models. The pre-trained network performs worse than human participants (Figure 4, left). Furthermore, unlike the human participants, the decline in the performance from by the type of the distractor category is similar for both the occluded and unoccluded stimuli sets. The network performs at chance in the most difficult condition of occluded and same-category distractor trials.

The physics-based analysis-by-synthesis model captures the average performance of the participants across the two experiments accurately (Figure 4, right). Moreover, our model captures the details of the subjects performance when broken down by the distractor type.

Discussion

Can humans recognize objects even when they are fully occluded? The behavioral study presented here indicates that the answer is yes, at least when the candidates are known (we suspect that people can often do so otherwise, as in the chair example in the introduction). What underlies our participants' performance? We think that this ability reflects the

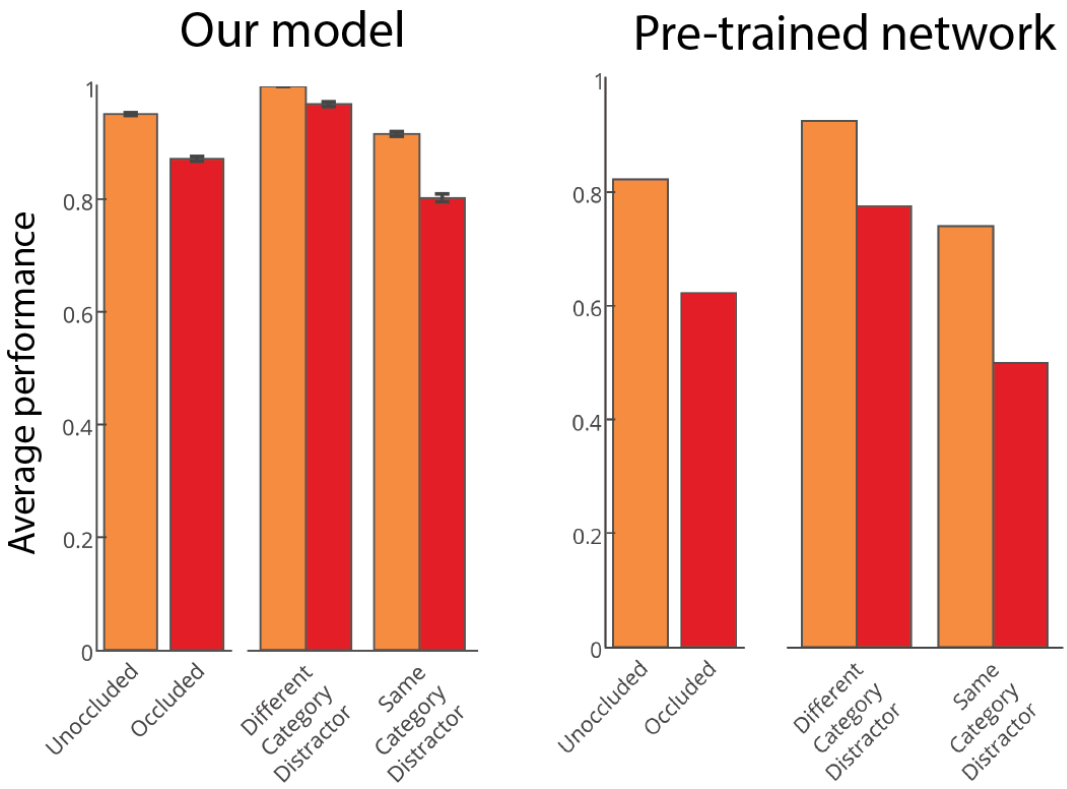


Figure 4: (Left) Average performance of the physics-based analysis-by-synthesis model on the unoccluded and occluded stimuli sets, and the breakdown of its performance by the tyep of the distractor category. Error bars indicate standard error of the mean. (Right) Average performance of the pre-trained network on the unoccluded and occluded stimuli sets, and the breakdown of its performance by the type of the distractor category. Dashed line shows the chance-level performance

presence of *causal theories* in human perception - in this case, theories about how soft and rigid objects bodies interact, and about how the resultant shape gets rendered into a visual percept. This approach is different from the standard model of visual perception. Most accounts of visual object recognition assert that the brain maps images to identity or class labels via increasingly more abstract feature hierarchies (DiCarlo et al., 2012; Krizhevsky, Sutskever, & Hinton, 2012; Riesenhuber & Poggio, 1999). In this approach, objects are not explicitly represented – they do not have physical properties such as mass and friction or 3D shape. The cost of not having explicit and causal representations of the scenes is an inability of composition – or in other words, requirement for data (lots of data) in the face of every new scene setting such as the setting presented here: objects fully occluded under cloths.

These issues are not just theoretical: we found that a state-of-the-art neural network (albeit trained only on unoccluded images) had trouble with our behavioral task. In particular, it could not approach human performance. We believe that any similar architecture (supervised learning) will have similar problems, because such approaches do not attempt to model the causal structure that gives rise to percepts.

We suggested an alternative solution. Instead of relying on the environment to teach us about variation, we proposed using basic knowledge about the world to derive and understand how the world can influence our percepts. We considered the problem of perceiving objects under cloths. Invariant features theories of perception face a problem in this domain, because cloth-draped objects differ dramatically in appearance from their unoccluded states. Treating perception as an inverse compositional process provides a solution.

There are a number of future directions that we wish to explore. First, our match-to-sample task is only one way of getting at people's abilities to perceive objects under heavy occlusion and in physical settings. We plan to build upon this paradigm for future experiments. In the Introduction we posed a question without answering - "which of these occluded objects is a chair?". This is perhaps the most interesting future direction to pursue; it requires accessing and manipulating the concept of a chair, instead of imagining object transformations as we studied here. Second, we plan to build hybrid architectures that involve fast and feedforward neural network pipelines for broad stroke comprehension of the scene and top-down physics-based architectures for in depth physical interpretation (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

Acknowledgments

This research was supported by the Center for Brains Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and by ONR grant N00014-13-1-0333.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Blender Online Community. (2015). Blender - a 3d modelling and rendering package [Computer software manual]. Blender Institute, Amsterdam. Retrieved from `http://www.blender.org`

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... Yu, F. (2015). *ShapeNet: An Information-Rich 3D Model Repository* (Tech. Rep. No. arXiv:1512.03012 [cs.GR]). Stanford University — Princeton University — Toyota Technological Institute at Chicago.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, ieee conference on* (pp. 248–255).

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach. *PLoS Comput Biol*, *11*(11), e1004610.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the acm international conference on multimedia* (pp. 675–678).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4390–4399).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, *63*(2), 113–140.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Thirty-seventh annual conference of the cognitive science society.*

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, *10*(7), 301–308.

“This problem has no solution”: when closing one of two doors results in failure to access any.

Hippolyte Gros (hippolyte.gros@cri-paris.org) (1,2,3)
(1) University Paris Descartes, Sorbonne Paris Cité,
12 Rue de l’École de Médecine, 75006 Paris, France.

Emmanuel Sander (emmanuel.sander@univ-paris8.fr)
(2) Paragraphe Lab, EA 349, University Paris 8, Department of Psychology,
2 Rue de la Liberté, 93526 Saint-Denis Cedex 02, France.

Jean-Pierre Thibaut (jean-pierre.thibaut@u-bourgogne.fr)
(3) LEAD-CNRS, UMR 5022, UB, Univ. Bourgogne Franche-Comté,
Pôle AAFE – Esplanade Erasme, 21065 Dijon. France.

Abstract

We investigated what happens when the spontaneous encoding of a problem is incongruent with its solving strategy. We created word problems from which two distinct semantic representations could be abstracted. Only one of these representations was consistent with the solving strategy. We tested whether participants could recode a semantically incongruent representation in order to access another, less salient, solving strategy. In experiment 1, participants had to solve arithmetic problems and to indicate which problems were unsolvable. In experiment 2, participants received solved problems and had to decide whether the solution was appropriate or not. In both experiments, participants had more difficulties acknowledging that problems inducing an incongruent representation could be solved than they had for problems inducing a congruent representation. This was confirmed by response times. These results highlight how semantic aspects can lead even adults to fail or succeed in the solving of arithmetic problems requiring basic mathematical knowledge.

Keywords: arithmetic problem solving; semantic structures; semantic encoding; strategy choice.

Introduction

The view that solvers build an abstract problem schema that guides a solving strategy free of contextual effects such as the semantic content of the problem has been challenged by recent investigators (e.g. Kotovsky & Simon, 1985; Bassok & Olseth, 1995). They have argued that the representation of the situation (objects, relations between objects) of a problem influences the solving procedure by reference to mental models (Reusser, 1990) that are mental constructs analogous to the situation depicted in the problem. Further work has shown that elements that are irrelevant from a mathematical point view (e.g. the type of objects referred to in the problems) were taken as cues while constructing an interpreted structure that solvers were using as a frame during the solving process (Bassok, Wu & Olseth, 1995). Recent studies have highlighted the importance of such a framework (Bassok, DeWolf & Holyoak, 2015; Lee, DeWolf, Bassok, Holyoak, 2016). The existence of an abstract interpreted structure raises the question of its congruence with the algorithm relevant for the solution. For example it has been shown that spontaneously, additions are associated with objects belonging to the same level of categorization, such as oranges and apples, and not to functionally connected objects such as oranges and baskets, even though there is no mathematical reason why one would not add these two types of objects (Bassok, 2001). To what extent can the algorithm be used when the interpreted structure is not congruent with the mathematical one? To date, it remains an open issue to show whether an incongruence between the representation abstracted from the wording and the mathematical structure will lead the solver to a dead end, even if the solving algorithm is one which only requires mathematical skills already mastered by the solver.

In order to investigate the representational issues, recent studies focused on problems that were solvable by alternative strategies, each one relying on a specific representation (Thevenot & Oakhill, 2005, Coquin-Viennot & Moreau, 2003, Gros, Thibaut & Sander, 2015). They showed that the strategy that was used to solve the problem was strongly constrained by the representation solvers were relying on. The present study is based on this type of problems. The main question addressed regards what participants do when the solving strategy associated with the representation spontaneously evoked by the problem is prohibited due to the lack of some necessary information. In other words, what happens when the “door” the participants usually use is blocked? Will solvers switch to an alternative representation congruent with the other strategy, which is still practicable, or will they remain stuck to their initial representation and think that the problem has no solution? In the latter case, it would mean that their initial representation of the problem is so constraining that they became blind to the alternative one and to the strategy associated with it. The solvers would thus be incapable of noticing and using the other “door”, yet wide open and leading to the solution.

Content effects

The characteristics of the situation depicted in a word problem have been shown to influence the solver’s ability to

find the correct solution of a problem. Isomorphic problems sharing the same mathematical structure can thus have different levels of difficulty depending on their wording (De Corte, Verschaffel, De Win, 1985).

More specifically, the influence of a wide range of abstract semantic properties on the encoding of word problems has been highlighted in different studies: continuous versus discrete property of a constant change (Bassok & Olseth, 1995), continuous versus discrete property of the entities of a problem (DeWolf, Bassok & Holyoak, 2015), symmetrical versus asymmetrical property of a relation (Bassok, Wu & Olseth, 1995), categorical versus functional property of an association between elements (Bassok, Chase & Martin, 1998) have all been shown to influence the solving strategies used by the solvers.

Overall, the crucial role of such abstract semantic features on the representations involved in problem solving demonstrates that any solving task necessarily depends on the solver knowledge regarding the elements included in the problem. This world knowledge can ease the solving process when the semantic information is congruent with it, or it can inversely impair the solving of a problem if the available information leads to a representation that is not congruent with the solving strategy.

Multiple-strategy word problems

In order to address this issue, Gamo, Sander & Richard (2010) studied multiple-strategy word problems in which the solution could be found in two different ways: either a 3-steps strategy requiring to perform a complementation inference (where the value of the difference between a set and a whole is calculated), or a 1-step strategy based on a comparison inference (requiring the calculation of the difference between two homologous quantities). According to Bosc-Miné and Sander (2007), it is easier to make a complementation inference when the context highlights the part/whole relation between the elements, as is the case in problems featuring cardinal values. On the other hand, an ordinal representation of the values usually highlights the comparison relations between the featured values, and thus promotes comparison inferences. This effect is fairly intuitive when we consider prototypes of these two categories of problems:

- Problem A: "John bought a 5€ exercise book and scissors. He paid 14€. A pen costs 3€ less than the exercise book. Paul bought scissors and a pen. How much did he pay?"
- Problem B: "Antoine took painting classes for 5 years, and stopped when he was 14 years old. Jean began at the same age as Antoine and took the course during 3 fewer years. At what age did Jean stop?"

Both types of problems can be solved using two identical strategies, namely the complementation-difference strategy (14-5=9 then 5-3=2 then 9+2=11, in which 11 is the answer) and the comparison-difference strategy (14-3=11, in which 11 is the answer). Yet, data collected by Gamo et al. (2010) showed that the comparison strategy is only scarcely used (4% of the situations) on problems similar to the problem A, which evoke a cardinal part/whole representation, whilst this strategy is widely used (64% of the problems) to solve the problems comparable to the problem B, in which the difference between the duration of the two classes is directly encoded in the ordinal representation of the problems. On the other hand, using the comparison on problem A requires to identify that the difference between the total amount paid by Paul and by John is equal to the price difference between the pen and the exercise book. The price of the scissors doesn't need to be calculated to solve the problem.

In order to gain a better understanding of these two types of induced representations, it is possible to draw a schema of their structure. Regarding the problem A, Gamo et al. (2010) suggested that the interpreted representation is cardinal. We can thus represent the different elements of the problem as depicted in figure 1.a.

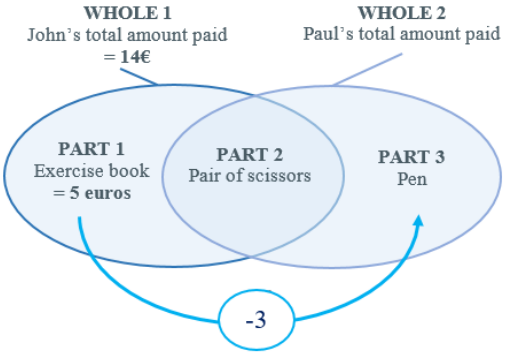


Figure 1.a: Cardinal representation of headcount problems such as those found in Gamo, Sander & Richard (2010).

It is then clear that the representation promotes the calculation of the intersection (Part 2) between Whole1 and Whole2 in order to get the value of Whole2, thus favoring the complementation strategy. The corresponding calculations are as follows: *Whole1 – Part1 = Part2*, then *Part1 – Difference = Part3*, then *Part2 + Part3 = Whole2*.

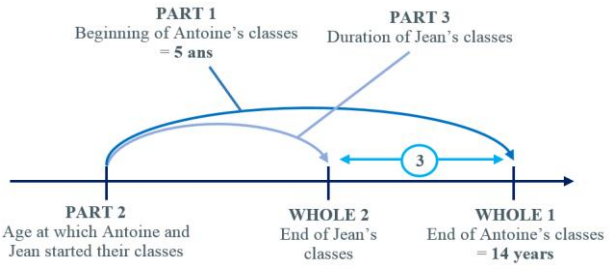


Figure 1.b: Ordinal representation of age problems such as those found in Gamo, Sander & Richard (2010).

Now regarding problem B, the expected representation has an ordinal nature, and can be drawn as on Figure 1.b. The ordinal representation of problem B values puts forward the fact that the difference between Whole1 and Whole2 is equal to the difference between Part1 and Part3. It is straightforward to deduct that the quickest solving strategy

is the 1-step comparison strategy: *Whole1 – Difference = Whole2*.

The solving strategy thus is directly encoded in the representation built, despite the fact that these two problems are isomorphic. Only an expert able to have a more exhaustive comprehension of the mathematical relations between the elements of these problems could access an abstract representation subsuming the other two, as detailed in Figure 1.c.

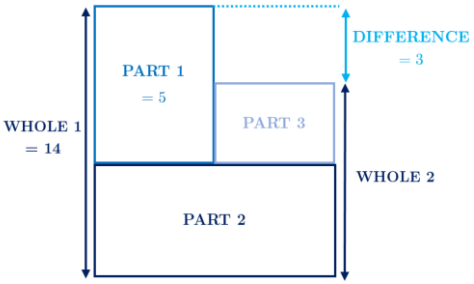


Figure 1.c: Abstract representation of the mathematical structure of the problems found in Gamo et al. (2010).

This structure can equally depict both problems, and makes it possible to use both solving strategies. The variations observed in the choice of one strategy over the other clearly indicate that the participants only abstracted intermediate, context-specific representations (with either ordinal or cardinal structure) instead of the abstract mathematical structure of the problems. It therefore seems that the semantic nature of the quantities used has a major impact on the type of representation being evoked, and subsequently on the type of solving strategy being used.

In what follows, we used the two types of problems described above. The above analysis shows that the quantities grounding each solving strategy differ. We gave participants problems in which one quantity was missing. As a result, they could not be solved according to one strategy (the 3-steps strategy) while remaining solvable according to the other. The issue was whether participants would use the other solving strategy (the 1-step strategy) in all cases or, by contrast, whether they would more often fail to solve them when their spontaneous strategy depended on knowing the value of the missing data than when the missing data was not necessary for making use of their spontaneous solving strategy.

Experiment 1

Materials

We created a set of problems which shared the same mathematical structure as described in figure 1.c, with one notable exception: we removed one of the numerical values so that participants could not access the 3-steps complementation strategy anymore, and solely had to rely on the 1-step matching strategy in order to solve the problems. Instead of asking participants to solve the problems with the fewest possible steps, we tested whether the preference for the semantically congruent strategy could lead to failure to solve the problems when only one of the two strategies was available.

We thus created two types of problems: (i) *solvable problems*, in which we removed the "part 1" value, making it impossible to use the 3-steps complementation strategy (see left part of figure 2), and *unsolvable problems*, where the value of Whole1 was not given, making it impossible to solve the problems using any strategy (see right part of figure 2). The latter were used as distractors.

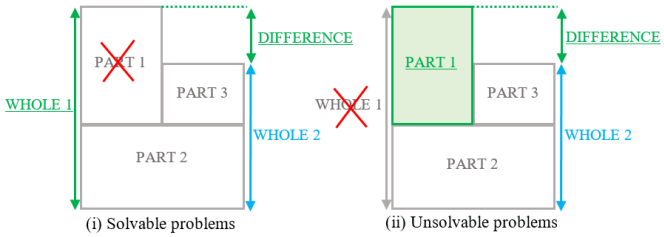


Figure 2: Structure of both solvable and unsolvable problems. Underlined values are those given in the problem. The goal was in both cases to find the value of Whole 2.

We implemented those two types of problems with different quantities inducing either an ordinal representation or a cardinal representation. Six different quantities were used in total; three ordinal quantities (duration, height, and number of floors) and three cardinal quantities (weight, price, number of elements). For each quantity, we created two different problems so as to make sure that the effects measured could not be attributed to one specific context. Those 12 problems were present in both forms: solvable problems and unsolvable problems, making a total of 24 problems in the pool. Finally, the problems were all written in the same way: the same number of sentences was used, and the numerical values (each below 15 so as to limit the calculus induced difficulties) were presented in the same order.

Procedure and experimental design

Each participant was presented with 12 different problems: 2 for each quantity, a solvable problem and an unsolvable one. The 6 unsolvable problems acted as distractors since we tested our hypotheses with the solvable problems. The order in which the problems were presented was randomized for each participant, and so were the two different versions of each problem.

The experiment took place online, on the survey platform Qualtrics. Participants who agreed to participate in the experiment were asked to complete the entire survey. On the first page, the following instructions were given: "You will find an arithmetic problem on each page of this survey. Please take the time to read and understand each of these problems, as this is not a speed test. Your task is to identify which problems can be solved and to indicate for each of them the operation you used to solve them as well as the solution you found. Be careful: some of the problems cannot be solved with the available information."

On each page of the survey, a problem was displayed with the following question below it "Given the data provided, is it possible to find the solution?" with two buttons "Yes" and "No". When the participants pressed "Yes", two new questions appeared, asking them to indicate both the operation needed to solve the problem and the result of the operation. Participants used the keyboard to write down their answers. After participants answered all 12 problems, a new page was displayed asking them for their gender, date of birth, and whether they took any break during the completion of the experiment.

Participants

A total of 89 adults completed our experiment. We removed from the analyses 15 participants who either mentioned taking a break during the test or answered at least one of the questions in less than 5 seconds or more than 5 minutes, which gave 74 participants (44 females, M=33.8 years, SD=13.4 years). Subjects were all fluent French speakers and were recruited through social networks and emails.

Hypotheses

Our first hypothesis regarded the ability to correctly solve the problems that induced a representation which was semantically incongruent with the solving strategy. We hypothesized that participants would have a better rate of success on solvable problems with ordinal quantities compared to solvable problems with cardinal quantities, because participants whose first spontaneous representation was associated with a 3-step strategy would fail to switch to a representation of the problem associated with the 1-step strategy.

Our second hypothesis regarded the situations in which participants overcame the difficulty related to an incongruent representation and managed to find the correct solution of a cardinal problem. We hypothesized that higher response times would be recorded on successfully solved cardinal problems compared to successfully solved ordinal problems.

Results

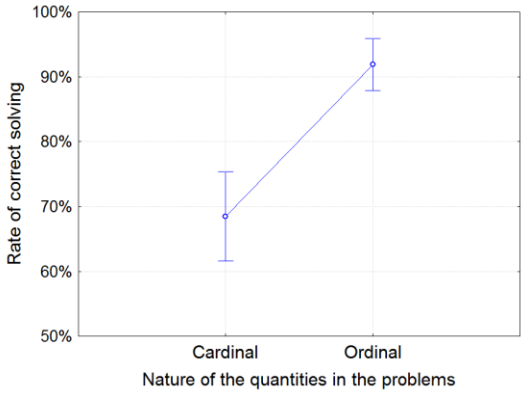


Figure 3: Rate of correct solving on solvable problems. Vertical bars denote 0.95 confidence intervals.

In order to test our first hypothesis, we analyzed the ratio of correct answers on solvable problems, depending on the cardinal or ordinal nature of the quantities used. Ordinal solvable problems were successfully solved in 91.9% of the trials, and cardinal solvable problems in only 68.5% of the trials (see Figure 3). A paired t-test was performed on participants' mean rate of success for cardinal and ordinal problems and showed that the difference was statistically significant ($t(73)=6.38$, $p<0.001$), therefore confirming our first hypothesis.

We then studied the response times on correctly solved problems, depending on the nature of the quantities in the problems. We wanted to see whether accessing the correct 1-step matching strategy on problems inducing a semantically incongruent cardinal representation would require a higher time than needed for problems inducing an ordinal representation. We removed from the analysis 4 participants who did not manage to correctly solve at least 1 cardinal problem and 1 ordinal problem. On average, participants took 68.7 seconds to successfully solve cardinal problems, against 49.8 seconds for ordinal problems (see Figure 4). A linear mixed model applied to the response times of successfully solved problems confirmed that the difference was significant ($F(1,7369)=20.38$, $p<0.001$), thus verifying our second hypothesis.

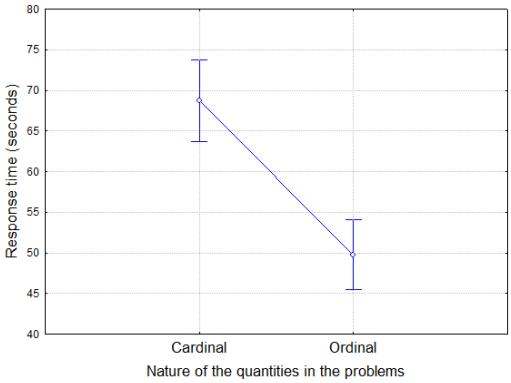


Figure 4: Mean response time on correctly solved problems. Vertical bars denote 0.95 confidence intervals.

Discussion

The results obtained confirmed that depending on the semantic nature of the quantities present in arithmetic word problems, participants induced a semantic representation which was either semantically congruent (ordinal) or semantically incongruent (cardinal) to the only solving strategy available. The fact that the rate of success was significantly different between these two types of problems showed that the semantic congruence does not just promote one strategy over the other (see also Gamo, Sander & Richard, 2009; Gamo, Taabane & Sander, 2011), but can actually lead participants to failure when only one strategy is available. Therefore, a semantically incongruent representation can significantly impair the solver's ability to find the solution of a problem if no semantic recoding takes place.

The study of the response times gave us insights into the need of such a recoding process. The longer time needed to solve cardinal problems suggested the existence of an extra reasoning step which allowed the solver to access a semantically incongruent solving strategy. The fact that the response rates were different for cardinal and ordinal problems suggested that such a process does not systematically happen and the longer RTs showed that this recoding step was a costly one.

In order to better understand these processes, in a second experiment, we explicitly gave the answer to the participants, and asked them to evaluate its validity. The goal was to go beyond the first experiment by evaluating whether the difficulty lied in the ability to discover the solution, or in the capacity to recognize it as a valid one. We made the hypothesis that, even when presented with the complete solution, participants would tend to refuse it more often if the representation they inferred from the problem was semantically incongruent with the solution. They would essentially refuse to use the door presented in front of them, because of their semantically biased representation of the situation.

Experiment 2

Procedure

The only aspect in which this experiment differed from the previous one was the nature of the task proposed to the participants. Instead of telling them to solve the problems themselves, we proposed a potential solution for each problem, and asked them to identify whether the solution was a valid one, or whether the problem could not be solved. For every problem, the question "Given the data provided, is it possible to find the solution?" was displayed, and below two choices were available: "No, we do not have enough information to solve this problem." and "Yes: *numerical value 1 – numerical value 2 = answer. Sentence presenting the answer*". For example, on one of the "number of floors" problems, the following solution was given: "Yes: 11-2=9. Karine arrives at the 9th floor."

Participants

A total of 223 adults completed our experiment. 27 of them were removed from the analysis according to the same criteria as in the previous experiment, and the analyses were performed on the remaining 196 participants (135 females, M=34.5 years, SD=14.8 years). Subjects were native French speakers recruited through social networks and emails.

Hypotheses

We expected to find the same results as in the previous experiment. Even though the solution was directly given to the participants, we hypothesized that they would still refuse it more often for problems with cardinal quantities than for ordinal problems. Similarly, we hypothesized that the correct identification of a solution would require more time for cardinal problems than for ordinal ones.

Results

As in the previous experiment, we first analyzed the ratio of correct answers on solvable problems depending on the semantic nature of the quantities used. The results detailed in Figure 5 showed that the solvable cardinal problems did indeed lead to a lower rate of success (63.6%) than the solvable ordinal problems (88.4%).

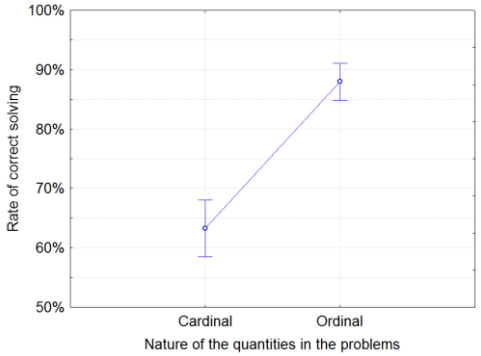


Figure 5: Rate of correct answers on solvable problems. Vertical bars denote 0.95 confidence intervals.

A paired t-test performed on the participants' mean rate of success confirmed that this difference was significant ($t(195)=9.26$, $p<0.001$).

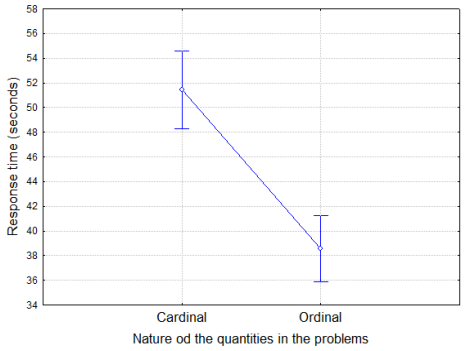


Figure 6: Mean response time on problems where the solution was correctly identified. Vertical bars denote 0.95 confidence intervals.

In order to assess the validity of our second hypothesis, we then studied the response time associated with the correct answers on solvable problems, thanks to a linear mixed model. We removed from the analysis the 26 participants who did not manage to correctly solve at least 1 cardinal problem and 1 ordinal problem. Figure 6 shows that getting a correct answer required a shorter reasoning time for ordinal (38.6 seconds) than for cardinal problems (51.4 seconds). This difference was statistically significant ($F(1,169)=30.28$, $p<0.001$), thus confirming our hypothesis.

Discussion

This task of solution validity assessment confirmed the effects observed on the first experiment, which involved a task of solution discovery. Even when participants were given the correct answer as well as the operation used to calculate it, they tended in more than 35% of the cases to

reject it if the problem induced a representation that was semantically incongruent with it.

Similarly, even when participants managed to successfully identify the correct answers of cardinal problems, they needed more time than for ordinal problems, which, we believe, indicates semantic recoding.

These results suggest that the semantic congruence effects that have been previously highlighted are not restricted to the spontaneous access to a solving strategy, but also to the evaluation of a given solving strategy.

Conclusion

These two experiments confirmed that the process to overcome a semantic incongruence between an interpreted representation and a solving algorithm is a costly one. Both the success rates and the response times were influenced by the incongruence, therefore showing the prevalence of semantic aspects over mathematical ones when faced with a situation from which we do not spontaneously abstract a solving-relevant representation.

We believe that these results pave the way for the emergence of a new approach to problem solving, integrating the three main components of any problem solving activity: the procedural aspects (the algorithms), the representations inferred from the solver's knowledge about the world, and the mathematical structure of the problems. We think that characterizing the way these three levels of description interact would deeply contribute to the understanding of the underlying cognitive processes. In this study, we tried to describe how a representation inferred from the world semantics can prevent the solver from accessing the relevant solving algorithm. Even when this algorithm only consists in a mere subtraction.

This underlines the importance to conduct extensive work on mathematical reasoning from an educational perspective in order to help the learners overcome semantic congruence effects. The robustness of these processes among adults suggest that they cannot spontaneously disappear, and instead need to be the target of a substantial and focused teaching so as to be reduced. As it has already been highlighted (Bassok, Chase & Martin, 1998; Lee, DeWolf, Bassok, & Holyoak, 2016), mathematics textbooks tend to preferably use semantically congruent examples when teaching a specific arithmetic concept. Yet, our study seems to indicate that people face greater difficulties when trying to solve incongruent problems compared to congruent ones. Would it not be preferable, then, to teach students how to specifically address these situations, and how to resort to a semantic recoding of their initial representations? Maybe then, the remaining door would be available to them.

References

Bassok, M. (2001). Semantic alignments in mathematical word problems. *The analogical mind: Perspectives from cognitive science*, 401-433.

Bassok, M, Chase, V. M, & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology*, 35, 99-134

Bassok, M., & Olseth, K., (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1522-1538.

Bassok, M., Wu, L., & Olseth, L. K. (1995). Judging a book by its cover: interpretative effects of content on problem-solving transfer. *Memory and Cognition*, 23, 354 e 367.

Bosc-Miné, C., & Sander, E. (2007). Effets du contexte sur la mise en œuvre de l'inférence de complément. *L'Année psychologique*, *107*(03), 397–425.

Coquin-Viennot, D., & Moreau, S. (2003). Highlighting the role of episodic model in the solving of arithmetical problems. *European Journal of Psychology and Education*, 18, 267-279.

DeCorte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solution. *Journal of Educational Psychology*, 77, 460-470.

DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. *Journal of Experimental Psychology: General*, *144*(1), 127-150.

Gamo, S., Sander, E., & Richard, J.-F. (2010). Transfer of strategy use by semantic recoding in arithmetic problem solving. *Learning and Instruction*, *20*(5), 400–410.

Gros, H., Thibaut, J.-P., & Sander, E. (2015). Robustness of semantic encoding effects in a transfer task for multiple-strategy arithmetic problems. *Proceedings of the 37th Annual Conference of the Cognitive Science Society, Los Angeles, USA*.

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive psychology*, *17*(2), 248–294.

Lee, H. S., DeWolf, M., Bassok, M., & Holyoak, K. J. (2016). Conceptual and procedural distinctions between fractions and decimals: A cross-national comparison. *Cognition*, *147*, 57-69.

Rapp, M., Bassok, M., DeWolf, M., & Holyoak, K. J. (2015). Modeling discrete and continuous entities with fractions and decimals. *Journal of Experimental Psychology: Applied*, *21*(1), 47.

Reusser, K. (1990). From text to situation to equation: Cognitive simulation of understanding and solving mathematical word problems. in H. Mandle. De Corte, N. Bennett, & H. F. Friedrich (Eds), *Learning and instruction in an international context* (Vol. II). New York: Pergamon Press.

Thevenot, C., & Oakhill, J. (2005). The strategic use of alternative representations in arithmetic word problem solving. *The Quarterly Journal of Experimental Psychology Section A*, 58(7), 1311-1323.

Analogical Generalization and Retrieval for Denominal Verb Interpretation

Clifton J. McFate (c-mcfate@northwestern.edu)
Qualitative Reasoning Group, EECS Department, 2133 Sheridan Road
Evanston, IL 60208 USA

Kenneth D. Forbus (forbus@northwestern.edu)
Qualitative Reasoning Group, EECS Department, 2133 Sheridan Road
Evanston, IL 60208 USA

Abstract

The creativity of natural language poses a significant theoretical problem. One example of this is denominal verbs (those derived from nouns) such as *spoon* in "She spooned me some sugar". Traditional generative approaches typically posit a unique entry in the lexicon for this usage, though this approach has difficulty scaling. Construction Grammar has evolved as a competing theory which instead allows the syntactic form of the sentence itself to contribute semantic meaning. However, how people learn syntactic constructions remains an open question. One suggestion has been that they are learned through analogical generalization. We evaluate this hypothesis using a computational model of analogical generalization to simulate Kaschak and Glenberg's (2000) study regarding interpretation of denominal verbs.

Keywords: Analogy; Construction Grammar; Linguistics; Analogical Generalization

Introduction

Linguistic creativity poses a theoretical problem for models of language learning and representation. One example of this is the use of denominal verbs (Clark & Clark, 1979) e.g.

1) a. She *spooned* me some sugar.
   b. The neighbor *hosed* the patio clean.

English speakers can easily interpret 1a as a transfer event involving sugar (with the spoon as instrument). Similarly 1b describes using the hose to wash a patio. How do we reach these interpretations? In many conventional theories the semantics of the sentence would be driven by the main verb. However, as Clark and Clark's (1979) analysis shows, the semantics of denominal verbs are highly context sensitive, and it certainly seems odd that *spoon* would independently have a transfer meaning in our lexicon. Construction grammar approaches present an alternative solution to this problem—that the interpretations are grounded in the argument structure itself (Goldberg, 2003).

Construction grammar proposes that the building blocks of language are *constructions*, pairings of form and meaning which are hierarchical and vary in both complexity and degree of fixedness. A morpheme is a kind of construction, as is the double-object (NP-V-NP-NP) pattern in 1a or the partially filled idiom "Jog X's memory" (Goldberg 2003).

An *expression* is a combination of non-contradictory constructions. In 1a, each of the NPs '*she*', '*me*', and '*some sugar*' is an NP construction. They combine with the double-object construction (exemplified in 1) that specifies each NP's location and role. A construction can also specify general semantic meaning. For example, the double-object construction is associated with transfer events (Goldberg, 2003). With construction grammar, we don't need *spoon* to have a new lexical entry. Instead, the double-object construction imposes its meaning on the denominal verb.

Kaschak and Glenberg (2000) investigated how people interpret novel denominal verbs and found that, consistent with a constructionist approach, participants were sensitive to argument structure differences when interpreting textual entailment and selecting a meaning for the novel verbs. Specifically, participants were more likely to attribute a transfer interpretation to a double-object construction (like that in 1a). However, a significant unanswered question is how humans acquire and represent such constructions.

One proposal is that children use analogy processes to generalize abstract constructions from item-based examples (Tomasello, 2003). Under this account, language learners would map the structural form and communicative function of an utterance to other examples, thus gradually abstracting away individual features. The generalization associates a semantic function with the linguistic form and could be used to interpret novel verbs appearing in the same context.

Our goal is to provide computational evidence for the idea that analogical generalization and retrieval could result in constructions accounting for Kaschak and Glenberg's denominal interpretations. We evaluate this using SAGE (Mclure et al, 2015), a computational model of analogical generalization based on Gentner's (1983) structure-mapping theory. Using a computational model further allows us to investigate precisely what kinds of representations are needed for successful generalization.

We train SAGE over a small corpus of annotated sentences and use MAC/FAC, a model of analogical retrieval, to retrieve and apply the learned constructions to Kaschack & Glenberg's stimuli. For our semantic representation we use Fillmore *et al's* (2001) FrameNet, a frame-semantic resource for English. The result is a semantic role assignment that we can evaluate as consistent with either a transfer or transitive interpretation.

Background

Construction Grammar

Recently, several different constructionist approaches have emerged (e.g. Boas and Saag, 2012; Steels, 2011).

However, as Goldberg (2003) summarizes they generally have a few core principles in common.

One principle is that construction grammar allows argument structure to contribute semantic meaning. Also unlike generative approaches (e.g. Chomsky, 1981), differing syntactic expressions are not derived from underlying deep structures; rather each surface form is its own instantiation of a set of constructions.

Constructionist approaches have been especially useful in explaining more idiomatic constructions such as the comparative-correlative ('The X-er, The Y-er) as in "The bigger they are, the harder they fall" (Cullicover & Jackendoff, 1999). Here, the structure of the sentence provides extra semantic meaning beyond that provided by the head verb. It is enriched but not derived from the verb.

Perhaps most importantly, constructionist approaches view constructions as learnable using general cognitive mechanisms on the basis of input from the environment (Goldberg, 2003). There is no *universal grammar* (e.g. Chomsky, 1981; Lidz *et al*, 2003) Thus, a constructionist account naturally predicts piece-meal acquisition of language such that a child's early utterances are those that occur in its linguistic environment. There is considerable evidence for this pattern. For example, Rowland and Pine (2000) found that subject-auxiliary inversion errors in a child's speech were mostly specific to individual wh-word auxiliary pairs, and that the pairs the child produced correctly were those that occurred more frequently in the input. Similarly, Theakston *et al* (2001) found that children's usage of a verb in a transitive or intransitive construction was predictable based on how frequently the verb appeared in that construction in the mother's speech.

More evidence of item-specific learning comes from Akhtar and Tomasello (1997) who found that older children (3;8) were able to use a transitive construction to interpret *agent* and *patient* roles for a novel verb while younger children (2;9) generally could not. However, both groups demonstrated understanding of transitive commands with known verbs. The older children had generalized a productive transitive construction while the younger children's understanding was verb specific.

However, if children's early language centers around item-based constructions, then an important question is how they develop more abstract syntactic knowledge.

Analogical Generalization

Analogy has been proposed as a mechanism for language learning and more specifically for generalizing from item-based constructions (Gentner & Namy 2006; Tomasello, 2003). There is abundant evidence that comparison facilitates language learning and relational extraction. Christie & Gentner (2010) found that comparison significantly improved relational abstraction in children when extending novel spatial labels to new situations. Further, Namy & Gentner (2002) found that common labels can invite comparison, facilitating the formation of categories based on relational rather than perceptual similarities.

Gentner's (1983) structure-mapping theory proposes that the process of comparison consists of aligning structured relational representations. The process is guided by constraints which prohibit many-to-one matches between entities and require matching parent relationships to have matching children. Further, structure-mapping proposes a bias towards systematicity, i.e. that people will prefer mappings that preserve higher-order (e.g. causal) structure. This preference has been demonstrated in humans with both match selection and prediction tasks (Clement & Gentner, 1991). Structure-mapping forms the theoretical basis for our model of analogical generalization. Forbus *et al*'s (2016) structure-mapping engine (SME) is a computational implementation of structure-mapping.

Our computational model of analogical generalization, SAGE (Mclure *et al*, 2015), is built on top of SME. SAGE operates over sets of predicate calculus representations called cases. Given a new (probe) case and a library of known cases, SAGE compares the case to ones in the library using a two-stage retrieval model, MAC/FAC (Forbus *et al*, 1995). The first stage compares flat feature-vector representations of the probe and case-library cases. While fast, this stage doesn't take into account structural similarity. The second stage operates over cases selected in the first phase and compares them using SME.

If the structural similarity score provided by SME is above a defined threshold[1], the new example will be combined with the retrieved case to create a generalization with a probability distribution governing features of the matched entities. Future examples can be added to the generalization or combine with other singletons to form a new generalization. If dissimilar to all, they join the case library as an ungeneralized example. Figure 1 shows the generalization process using representations from our experiment.

SAGE has been used in tasks such as learning concepts from maps (Mclure et al, 2015). Additionally, SAGE models a phenomenon from the analogy literature called *progressive alignment*, whereby highly-alignable cases pave the way for understanding less similar pairs, (Kotovsky and Gentner 1996; Kandaswamy *et al*, 2014).

There is evidence that progressive alignment contributes to language development. For example, Goldwater *et al* (2011) proposed a structure-mapping account of children's construction learning, as assessed by their syntactic priming (that is, by whether and to what extent they can re-use the syntactic form of a just-heard utterance when framing a new utterance). Goldwater et al. (2011) found that both 4- and 5-yr-old children showed syntactic priming, but (in keeping with our earlier discussion) 4-yr-olds were more concrete—the target had to be highly similar to the prime in order for them to show priming. Goldwater and Echols (2014) further

[1] An assimilation threshold of 1.0 means perfect identity would be required to merge, whereas a threshold of 0.0 means any two descriptions will be merged.

found that 4-yr-olds primed with high-similar sentences could go on to show priming from less similar sentences—evidence for progressive alignment in learning constructions (See also Childers & Tomasello, 2001).

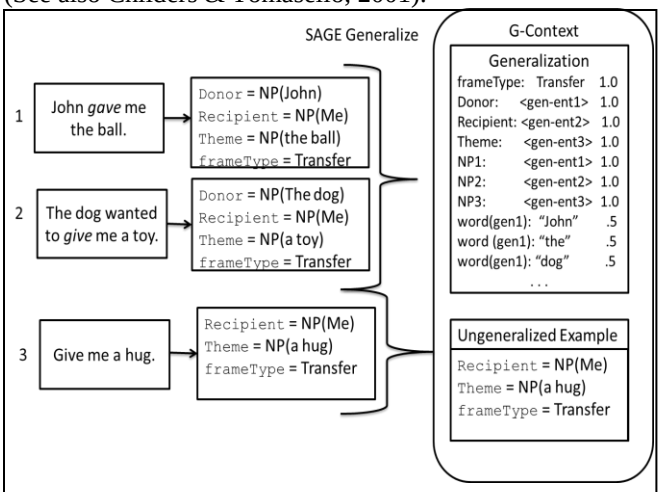


Figure 1: SAGE Generalization

Double-object sentences 1 and 2 are compared using SME. Since their similarity score is above threshold, their shared structure (semantic roles and phrase structure) is abstracted into a generalization with a probability distribution governing their variable features. Sentence 3, an imperative, is compared using MAC/FAC, but lacks a `Donor` FE and its `Theme` FE is in object position. It becomes an ungeneralized example.

Experiment

The first goal of the experiment is to evaluate whether analogical generalization can result in argument-structure constructions. The second is to evaluate whether these constructions can be used to interpret denominal verbs through analogical retrieval.

Simulation Target

Our target for simulation is the first experiment from Kaschak and Glenberg (2000). In Experiment 1, participants were presented with pairs of sentences (e.g. Figure 2) headed by either a conventional verb or a novel denominal verb. One was a double-object construction while the other was a transitive construction. Each also had an additional `Purpose` argument. This ensured both examples had the same number of entities, so readers must rely on structure.

Participants were evaluated on one of two tasks. In Task 1, participants were presented with one of two inferences. One was consistent with a transfer interpretation and the other a transitive action e.g. "Tom got the apple" vs "Lyn acted on the apple". They were then asked which example most strongly implied the inference. If they were influenced by the argument structure, then they should choose the double-object sentence for the transfer inference. In the inference task, participants overwhelmingly chose the double-object construction for transfer inferences (92% for conventional verbs and 80% for novel denominals).

1. Lyn crutched Tom her apple so he wouldn't starve. (double-object form)
2. Lyn crutched her apple so Tom wouldn't starve. (transitive form)

Figure 2: Example from Kaschak and Glenberg (2000)

In a second task, participants were shown each sentence independently and asked which of a pair of meanings e.g. "to act on using a crutch" or "to transfer using a crutch" best fit the denominal verb. If they were influenced by the argument structure, then they should select the transfer meaning for the double-object construction.

Participants in the meaning-choice task showed a weaker though still significant preference for the transfer definition when presented with the double-object construction (61% following double-object vs 42% following a transitive).

Our simulation models the semantic classification of transfer vs transitive action. We trained SAGE on example sentences annotated with semantic frames and roles. For each double-object example from Kaschak and Glenberg (2000), the simulation should assign semantic roles consistent with a transfer frame. For each transitive example, it should label it with a transitive frame. Kaschak & Glenberg further claim that the full semantic meaning of the sentence is based on the affordances of the noun the denominal came from, a claim we do not examine.

Representations and Materials

We use Fillmore *et al's* (2001) FrameNet as our semantic representation. FrameNet defines frames that relate an evoking lexical item and its dependent structures to roles in a semantic description. For example, the word *Give* evokes the `Giving` frame which includes required an optional roles (frame elements) such as a `Donor` and `Recipient`. The realization of frame elements in a construction is called a valence pattern. Consider example (2) below:

2) I saw John *give* the teacher the apple.

With *give* as the target, the NP "John" would be annotated as the `Donor` of a giving frame. The NP, "the teacher", would be the `Recipient`, and "the apple" would be the `Theme`. This annotation format only identifies the arguments to the target, ignoring other aspects of the sentence.

The training set was created by manually annotating 21 sentences from a 6th grade reading comprehension workbook (Spectrum, 2007). Each training case consisted of a sentence and predicate calculus representations of its target verb, the arguments to that verb, their positions, their roles, and the words in each argument. For sentence 2, the first NP would be represented as follows:

```
(isa NP1 NP)
(FE-Donor "give" NP1)
(wordMemberOf NP1 "John")
(loc1 sentence1 NP1)
```

The test set was created by automatically chunking the stimuli from Kaschak and Glenberg (2005) into its valence pattern phrases using regular expressions. Each phrase was

labeled with its phrase type as above, but no information about the evoked frame or the frame element was given. All 20 double-object examples from their experiment were used. We discarded one transitive example because its additional argument was not the same form as the rest of the examples. Thus, our total test set consisted of 39 examples.

Procedure

In the training phase, SAGE was given each of the training stimuli as an individual case for generalization (see Figure 1). The generalization threshold was set to .8. The training set consisted of nine sentences that evoked `Giving`. Seven examples were inheritors of the `Transitive_action` frame (e.g. `Cause_motion`) and four were distractor `Motion` frames. The training examples covered a wide range of constructions including examples with several prepositional attachments. Of the nine `Transfer` sentences, only two illustrated the double-object in isolation. Six examples included an additional argument.

A hallmark of SAGE is its sensitivity to the order of the training stimuli, demonstrating progressive alignment. We first evaluated the model with a hand-made progressive alignment training order (PA-ordered) that clustered similar constructions and then with randomly generated orders.

In the evaluation phase, each example from our test set was used as a probe for analogical retrieval using MAC/FAC over the learned case-library. The top-scoring response is compared to the probe using SME. Structures in the base that are missing in the target are hypothesized as candidate inferences. These candidate inferences are the model's interpretation of the incoming sentence. This is shown in Figure 3.

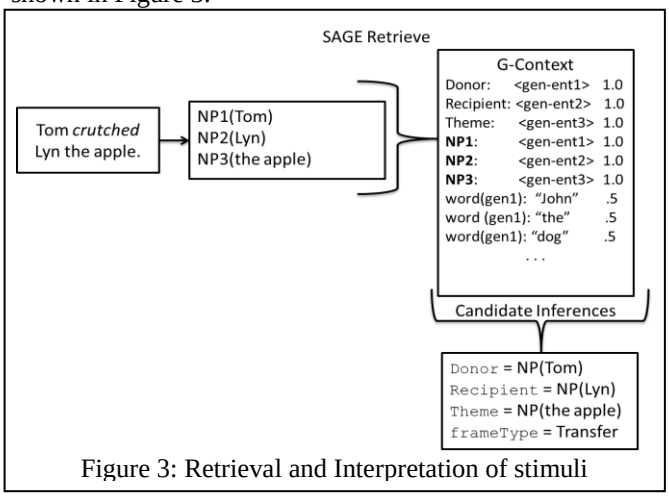


Figure 3: Retrieval and Interpretation of stimuli

Results

The results were evaluated based on the candidate inferences generated by the retrieval process. For the 20 double-object examples, a response was counted correct if it both correctly identified the frame-type of the verb (e.g. *crutched* = `Giving`) and if it correctly labeled the 3 NPs as `Donor`, `Recipient`, and `Theme`. This corresponds to a double-object response in Kaschak and Glenberg (2000). An interpretation was judged incorrect if it over-assigned the three FEs which could happen if the double-object and prepositional ditransitives formed a single generalization.

For the transitive stimuli, we counted the interpretation correct if it chose an inheritor of `transitive_action` or a generalization containing only those inheritors. Generic non-transitive motion, for example, was incorrect. We did not evaluate the specific frame elements because a generic transitive action such as *"Lyn acted on the apple"* could be consistent with many different types of transitive action and the FEs vary across FN frames.

We compared to two baselines. The first was a random baseline. For each example, we assume a random baseline system that could label it *Transfer*, *Transitive*, or *other* (corresponding to the `Motion` distractors). Thus it would have a 1/3 chance of classifying the frame-type correctly. For transfer classifications, it would further have to assign each FE correctly (1/6 at random). Thus for the 19 transitives a random system would have a 1/3 chance of being correct. For the 20 transfer examples, random guessing would have a 1/18 chance of being correct. This gives a random baseline an overall mean of 7.4 (19%). The second was a baseline of guessing *Transitive* for each example. Under our evaluation measure, this would result in 19 correct answers (49%). We call these the *random* and *choose-transitive* baseline.

PA-Ordered Training Set:

On 19 out of 20 examples, the model correctly interpreted the verb as evoking a `Giving` frame and correctly assigned all frame elements (`Donor`, `Recipient`, `Theme`).

On the transitive examples, the model correctly interpreted the event as non-transfer in 18 out of 19 examples. This gives us a total of 37 out of 39 (95%).

With a single manually determined training order, the model is deterministic—there is no variance. Our model significantly out-performs the expected performance of the random baseline ($P < .05$). It is larger than the static *choose-transitive* baseline and thus out-performs that as well.

Random Training Orders:

We also evaluated our model using randomized training orders. Across 25 trials using random orders, the mean total correct dropped to 26.88 (68.9%). The double-object stimuli were most affected, dropping to a mean accuracy of 9.96. The max total accuracy across all 25 trials was 37 (95%) correct with 19 correct double-object classifications. The minimum accuracy was 19 (49%) correct classifications, with only 4 correct double-object classifications.

A one-sample t-test demonstrated that the random order mean was a significant improvement over the *random* baseline $t(24) = 16.05$ $P < .05$. Its performance was also a significant improvement compared to the *choose-transitive* baseline $t(24) = 6.49$ $P < .05$.

Generalizations:

We manually inspected the generalizations produced by the model trained on hand-ordered stimuli. SAGE produced four generalizations. One was the simple double-object construction. Another contained two of the double-object

constructions with an additional modifier. The next contained two simple `cause_motion` transitive sentences and the final generalization contained two examples of transitive actions with an additional argument.

Without phrase labels:

We also evaluated performance without phrase type labels. Instead, the probe included the arguments to the verbs as unlabeled 'chunks'. The model was trained with the hand-ordering. In this condition the model correctly identified the `Giving` frame but consistently inferred a double-PP pattern, thus misaligning FEs. Without labels, it essentially only operates over the number of chunks and their size which isn't enough to distinguish between examples.

Discussion

After training, the model successfully applied the transfer and transitive semantics to the novel denominal verbs. The results of this model therefore support the claim that analogical generalization could be a mechanism for generalization and application of linguistic constructions.

As predicted, receiving the examples in progressive alignment order led to the best results. This is consistent with the progressive alignment phenomenon seen in human learning. Indeed, the linguistic environment itself may facilitate this kind of learning, as Cameron-Faulkner *et al* (2003) found that 51% of child-directed speech began with one of 52 constructions, and 45% began with one of seventeen words. Future work demonstrating progressive alignment in language learning would support our model. We may predict that sequential comparison of canonical verbs would improve performance on the novel denominals. Kaschak & Glenberg (2005) do not evaluate this.

The model makes several representational assumptions. First, we assume that humans are able to chunk sentences into arguments for a target verb. We don't claim that these representations are part of a particular larger parse structure nor do we endorse a specific account of how humans form these arguments. Going from raw input to preliminary structures will be a significant focus of future work.

Our results also suggest the importance of both chunking a sentence into arguments and classifying sentence chunks. Without phrase-labels, performance dramatically decreased. One option is to include classification as a part of the chunking process. A two-phase model could use analogy to infer chunks from low-level features such as word order or dependency parse information.

Related Work

Semantic role labeling has been an active area of research in computational linguistics, though open-domain semantic parsing remains elusive. An example is Das *et al's* (2014) statistical SEMAFOR parser. Ovchinnikova's (2012) statistical abduction system additionally uses a knowledge-base extracted from WordNet and FrameNet. However, none of these systems focus on constructions. As far as we know, our model is the first use of a cognitive simulation to evaluate the analogical account of construction learning.

Connor *et al*'s (2008) Baby SRL system classified transitive agent and patient arguments using a linear classifier trained over a corpus of child directed speech. They investigated the theory that children use the number of nouns as a heuristic for classification, and replicated child performance including over-generalization errors. Adding a feature for verb position greatly improved performance.

Baby SRL targets a simpler construction and uses a word-level representation, but they propose that the approach could be extended to use phrases. A fundamental difference between our approaches is that SAGE produces classifications for each phrase jointly (candidate inference), while theirs labels independently. This should affect error patterns, with independent labeling more likely to assign the same roles to multiple arguments. Furthermore, a linear classifier would not generally model progressive alignment.

Solan *et al* (2004) propose an incremental unsupervised algorithm which represents sentences as paths through a word-graph and identifies classes of equivalent words and patterns. Their model trained over a corpus of child directed speech and demonstrated intermediate performance on a 9th-grade ESL proficiency test. They did not include semantic roles as is necessary for our task. A benefit of their approach is that constructions are learned hierarchically, where we currently learn a generalization for each valence pattern. One direction for future work could use hierarchical models of analogical generalization (Liang & Forbus, 2014).

Bergen, Chang, & Narayan (2000) propose Embodied Construction Grammar (ECG) in which constructions link linguistic form to conceptual schemas used to specify parameters for simulation. The simulation generates inference and prompts response. Their conceptual schemas are similar to our frames in specifying a predicate-argument mapping, though constructions in ECG also contribute additional simulation parameters such as perspective. While we don't incorporate simulation into our model of semantics, analogical generalization could support learning of ECG's conceptual-schemas.

Finally, Steels' (2011) Fluid Construction Grammar is a construction grammar formalism that was designed for work on human-robot interaction, though not necessarily as a cognitive model.

Conclusion & Future Work

Humans must adapt to an ever-shifting linguistic landscape. Constructionist accounts of language facilitate this by pairing semantic meaning and syntactic surface forms. Furthermore, constructionist theories view language as learnable using general cognitive abilities. We examine whether analogy could be a mechanism used in both the generalization of constructions and their application. We simulated experiment 1 of Kaschak and Glenberg's (2000) denominal verb study using a computational model of analogical generalization and retrieval, SAGE, which allowed for inspections of and concrete hypotheses about the representations used in construction building. Our model produced correct transitive and ditransitive semantics,

supporting the role of analogy in language learning.

Much remains to be done. We explored only a handful of English constructions. Future work will focus on modeling other linguistic studies as well as applying these techniques to larger-scale problems. Finally, our system currently treats phrase-level chunking as a black box. Future work will involve building these representations from the ground up.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. Developmental psychology, 33(6), 952.

Bergen, B., Chang, N., & Narayan, S. (2004). Simulated action in an embodied construction grammar. *In Proc. 26th Conference of the Cognitive Science Society*.

Boas, H. & Saag, I, (Eds.) 2012. *Sign Based Construction Grammar*. Stanford, CA: CSLI Publications.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, 27(6), 843-873.

Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2008). Baby SRL: Modeling early language acquisition. In *Proc. 12th Conference on Computational Natural Language Learning.*

Culicover, P.W. & Jackendoff, R. (1999) The view from the periphery: the English comparative correlative. *Linguist. Inq.* 30, 543–571

Childers, J., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, 37, 730–748.

Chomsky, Noam. 1981. Lectures on Government and Binding. Dordrecht:Foris.

Christie, S. & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11 (3). 356-373.

Clark, E. V., & Clark, H. H. (1979). When nouns surface as verbs. *Language*, 767-811.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89-132.

Das, D., Chen, D., Martins, A. F., Schneider, N., & Smith, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1), 9-56.

Forbus, K., Ferguson, R., Lovett, A., & Gentner, D. (2016). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*.

Fillmore, C. J., Wooters, C., & Baker, C. F. (2001). Building a Large Lexical Databank Which Provides Deep Semantics. *Proceedings of PACLIC-2001*.

Forbus, K., Gentner, D. and Law, K. (April-June, 1995). MAC/FAC: A model of Similarity-based Retrieval. *Cognitive Science*, 19(2), 141-205.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2), 155-170.

Gentner, D. & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6), 297-301.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5), 219-224.

Goldwater, M.B., & Echols, C. H. (2014) Structural priming as analogical mapping. *Presented at the 13th ICSCL, Amsterdam.*

Goldwater, M., Tomlinson, M., Love, B., Echols, C. (2011). Structural priming as structure-mapping: Young children use analogies from previous utterances to guide sentence production. *Cognitive Science*, 35(1), 156-170.

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of memory and language*, 43(3), 508-529.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.

Liang, C. and Forbus, K. (2014). Constructing Hierarchical Concepts via Analogical Generalization. *In Proc. 36th Conference of the Cognitive Science Society*.

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, 87(3), 151-178.

McLure, M.D., Friedman S.E. and Forbus, K.D. (2015). Extending Analogical Generalization with Near-Misses. *In Proc. 29th AAAI Conference on Artificial Intelligence*.

Namy, L. L,. & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, 131, 5-15.

Ovchinnikova, E. 2012. Integration of World Knowledge for Natural Language Understanding. Atlantis Press, Springer.

Rowland, C. F., & Pine, J. M. (2000). Subject–auxiliary inversion errors and wh-question acquisition: 'What children do know?'. *J. Child Language*, 27(01), 157-181.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2003). Unsupervised context sensitive language acquisition from a large corpus. *In Advances in Neural Information Processing Systems*,

Spectrum. (2007) *Spectrum Reading Grade 6*. School Specialty Publishing. Columbus, OH.

Steels, L. (Ed.). (2011). *Design patterns in fluid construction grammar* (Vol. 11). John Benjamins Publishing.

Taylor, J. L. M., Friedman, S. E., Forbus, K. D., Goldwater, M. and Gentner, D. (2011). Modeling structural priming in sentence production via analogical processes. *In Proc. 33rd Annual Conference of the Cognitive Science Society*.

Theakston, A., Lieven, E., Pine, J., and Rowland, C. 2001. The role of performance limitations in the acquisition of verb argument structure. *J.Child Language* 28, 127–152.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Mindfulness meditation as attention control training: A dual-blind investigation

Alexa R. Romberg[1,2] and Henk J. Haarmann[1]
{aromberg, hhaarmann}@casl.umd.edu
[1]Center for Advanced Study of Language, 7005 52nd Ave
College Park, MD 20742 USA
[2]Department of Human Development and Quantitative Methodology, 3304 Benjamin Building
College Park, MD 20742 USA

Abstract

Mindfulness meditation is a form of secular meditation that emphasizes non-judgmental awareness of the present moment. Research into mindfulness has greatly expanded in recent years (Davidson & Kasniak, 2015) and a growing literature has documented effects of mindfulness training on cognition. However, the specific aspects of mindfulness meditation training for novice practitioners that might influence cognition remain unexplored. The present study used a rigorous, dual-blind design to investigate whether the attention-monitoring component of mindfulness meditation reduces mind-wandering and improves performance during reading comprehension and sustained attention tasks. When compared with relaxation meditation, mindfulness training improved recall of specific details from a text but did not reduce mind-wandering or affect sustained attention. The results are discussed with respect to design considerations when studying a meditation intervention.

Keywords: mindfulness; meditation; mind-wandering; text comprehension; sustained attention

Introduction

In recent years, there has been a dramatic increase in the application of mindfulness meditation training (MMT) across many sectors of our society. MMT is a form of secular meditation emphasizing non-judgmental awareness of thoughts, feelings, and sensations in the present moment (e.g., Kabat-Zinn, 1994). In some sense, mindfulness can be thought of as the inverse of mind-wandering (MW; Mrazek, Smallwood, & Schooler, 2012), with mindfulness arising from full engagement with the present moment and MW arising from attention moving away from the present moment.

Several studies have found positive effects of MMT on cognition, including fundamental functions such as attention control and working memory (e.g., Jha, Morrison, Dainer-Best, Parker, Rostrup, & Stanley, 2015; MacLean, Ferrer, Aichele, Bridwell, et al., 2010; Mrazek, Franklin, Phillips, Baird & Schooler, 2013; Zeidan, Johnson, Diamond, David, & Goolkasian, 2010; but see Banks, Welhaf, & Srour, 2015) and higher-order abilities, such as reading comprehension (e.g., Mrazek et al., 2013). Indeed, positive benefits of MMT on cognition have been found after brief inductions lasting 6-8 minutes (e.g., Mrazek et al., 2012) as well as training programs of between 4 and 8 hours of practice (e.g., Mrazek et al., 2013; Zeidan et al., 2010) and highly intensive interventions (e.g., MacLean et al., 2010). A challenge to interpreting the findings from these studies is that they employed comparison groups that were engaged in very different kinds of activities than MMT, such as wait-list controls (e.g., Jha et al., 2015; MacLean et al., 2010), attending a nutrition class (Mrazek et al., 2013) or listening to an audiobook (Zeidan et al., 2010). The many differences between treatment and control groups leaves open the possibility that differences between groups were due to nonspecific factors, such as differing expectancy effects across groups. Just as importantly, the studies leave unexplored *what components* of MMT contribute to cognitive effects.

MMT has many components, some of which may be more strongly related to specific aspects of cognition than others. The present study sought to closely examine one of those components, attention monitoring, while controlling for a second, relaxation. Specifically, we tested the hypothesis that the *attention-monitoring component* of MMT provides a form of attention training for the novice practitioner that reduces MW in non-meditation contexts. In testing our hypothesis we adhered to many recommendations suggested by Davidson & Kazniak (2015) in their recent critical review of research on mindfulness and meditation.

First, in designing our comparison group, we used the "dismantling strategy" advocated by Davidson & Kazniak (2015) in which the control intervention contains all of the elements of the treatment intervention except for the component under investigation. In order to isolate potential effects of the attention-monitoring component of MMT above and beyond the general benefits that one might expect from beginning a meditation practice, we trained two groups of participants in meditations inspired by MMT. Both forms involved directing attention inward and focusing on the breath but the two groups differed in the explicit goal that was set for them in the meditation and the specific guidance given during meditation. The Relaxation Meditation Training group (RMT) were told that the meditation training they were doing was a form of "relaxation training" and that their goal was to relax as much as possible. They were guided to relax by observing their breath. The Mindfulness Meditation Training group (MMT) were told that the meditation training they were doing was a form of "attention training" and that their goal was to improve their mental focus. They were guided to engage in a relaxation task of counting their breaths and were given additional instructions specifically related to MW.

Second, our study procedure followed a "dual-blind" model and included the 4 desiderata outlined by Davidson

& Kazniak (2015): Comparison groups were matched for 1) structural dosage, 2) daily practice and 3) instructor characteristics, and 4) participants were blind as to which was the "experimental" intervention. In addition, experimenters were blind to the group assignment of participants during data collection.

In order to test for effects of the attention-monitoring component of MMT on MW and cognitive performance, we used outcome measures that have been shown by previous work to be sensitive to the effects of MW and that are relevant to both basic and applied theories of cognition: text comprehension and sustained attention. Clearly, MW reduces time-on-task, which may in and of itself impair performance. Further, MW may reflect a failure of attention control that could have repercussions across the cognitive system (McVay & Kane, 2009).

In the present study, meditation-naïve participants made 4 visits to the lab over 8 calendar days. On the first visit they completed demographic questionnaires, the pre-test for text comprehension and sustained attention and finally were introduced to the meditation and given 20 minutes to practice. Visits 2 and 3 consisted of 45 minutes of meditation practice. Finally, on visit 4, participants began with 10 minutes of meditation practice, followed by the post-tests in text comprehension and sustained attention. Thus, our study design can be thought of as testing state effects of MMT relative to RMT, augmented by prior practice. Our analyses focus on 1) replicating the expected effects of MW on objective performance markers in text comprehension and sustained attention and 2) testing the Meditation Group X Session interaction for both the measures of MW and objective performance markers in the text comprehension and sustained attention tasks.

Method

Participants

Participants were young adults with no significant prior meditation experience. They were selected on a first-come-first serve basis from respondents to an online pre-screening questionnaire. All reported (1) normal or corrected-to-normal hearing and vision, (2) being a native speaker of English, (3) being between 18 and 40 years of age, (4) no more than minimal prior meditation experience, and no experience with mindfulness meditation, (5) no intensive long-term yoga experience, and (6) no prior diagnosis of a neurological or psychiatric condition. Participants were paid $90 for the approximately 6 hours of total study time.

Data was collected from a total of 86 participants. The sample consisted of university students (62 undergraduate and 20 graduate students, 3 non-students and 1 who declined to report) and was ethnically diverse (34% European American, 33% African American, 18% Asian American, 5% Hispanic, 9% other). Nine participants were excluded from all analyses for indicating in the first session that they failed to meet one or more of the inclusion criteria (6), consistently failing to follow the instructions (2), attrition after the first session (1). Participants in the final sample (N = 77) were randomly assigned to two experimental groups: Mindfulness Meditation Training (MMT; N=37; 27 females; M age=20.4 years, SD=2.0) or Relaxation Meditation Training (RMT; N=40; 28 females; M age=22.4 years, SD=4.3).

Materials

Meditation Training In each meditation training session, participants were seated at a computer, reviewed written instructions and then listened to pre-recorded audio over headphones. Instructions were tailored to the meditation condition and session, emphasizing either relaxation meditation training (RMT) or the attention monitoring meditation training (MMT). Participants were not informed that different participants were doing different types of meditation or the specific hypotheses of the study.

Participants were instructed to sit comfortably, with feet on the floor, hands in their lap and chin up, with their eyes closed or looking ahead with a dull stare. The first session included one 20-minute auditory guided meditation, sessions 2 and 3 consisted of 40 minutes of auditory guided meditation and session 4 included one 10-minute auditory guided meditation and a 5-minute non-auditory meditation reminder. The guided meditations used in training were created by CASL researchers after referencing Jon Kabat-Zinn meditations. They consisted of instructions alternating with silence to allow for the participant to practice on their own. The transcripts were recorded by a male talker with experience with meditation.

The guided meditations for the two conditions contained the same general pattern of instructions and silences but the content of the instructions varied between groups. Instructions for MMT focused on counting breaths. The auditory and written instructions also explicitly asked participants to (a) notice when MW occurred, without judging the content or frequency of the experience, and (b) bring attention back to the breath-counting task when they became aware of their MW. The focus on counting breaths was intended to give participants a concrete marker by which to notice MW when it occurred. In contrast, instructions for RMT were designed to follow the same general pattern of instruction (i.e., they also had a task to do during meditation) but to maximally encourage relaxation. The guided meditation asked participants to relax by focusing on their breath (without counting) and did not explicitly mention MW at all. After completing the Text Comprehension task and before completing the SART, participants engaged in a 5-minute "reminder" meditation. They read instructions on the computer asking them to use the techniques they had been learning during their training. At the end of the 5 minutes a tone sounded to end the meditation.

Self-evaluation of meditation training and relaxation Immediately prior to each auditory meditation, participants indicated their level of relaxation at that moment using a

visual analog scale. In Sessions 2 to 4, participants were also asked whether they had used the meditation techniques they were learning on their own since the last session.

After each meditation training session participants again rated their relaxation. They also completed a retrospective attention report, using a slider to indicate the percentage of time they were 1) completely engaged with the task, 2) thinking about the task more generally, and 3) thinking about things unrelated to the task (i.e., task unrelated thoughts or TUTs).

Text Comprehension The text comprehension task was modeled on the self-paced paradigm used by Feng, D'Mello and Graesser (2013). Participants read passages one sentence at a time, pressing the space bar to advance to the next sentence. Each passage was followed by a set of multiple-choice questions with four possible responses. The questions fell into one of two categories: Sentence-Linked or General. Each Sentence-Linked question assessed comprehension of a specific detail from the passage (e.g., What did the poet Homer do in old age?), while General questions assessed comprehension of the passage as a whole (e.g., One can conclude Momaday knows about ……?).

Attention probes were inserted between sentences at several points in each passage, fixed across participants. Each probe asked "Were you mind wandering when you read the previous sentence? Press 1 for Yes and 2 for No." There were two categories of attention-probes: Each *Question-Linked* attention probe (66% of total) immediately followed a sentence whose comprehension would be assessed on a Sentence-Linked question. *Filler* attention probes were placed in pseudorandom positions in the passage and were not linked to specific questions.

The passages used were from the reading comprehension section of Forms G and H of the Nelson-Denny Reading Test™, a standardized assessment of reading ability appropriate for high school and four-year college students (www.riversidepublishing.com). Each Form consisted of 7 passages: one long passage (approximately 600 words) with 6 attention probes (4 Question-Linked and 2 Filler) and 7 associated questions and 6 shorter passages (approximately 225 words each) with 3 total probes (2 Question-Linked and 1 Filler) and 4-5 associated questions. Participants completed one Form (i.e., G or H) at Session 1 and the other Form at Session 4, randomly assigned and counterbalanced across subjects.

The text comprehension task began with detailed instructions presented on the computer. These instructions defined MW as when "attention drifts away from the task" and provided examples of MW. Participants were told to report NO mind-wandering if they were "completely engaged in the task of reading and understanding the passage" and to report YES if they were either a) thinking about what they were reading AND something else or b) thinking only about something else. After the instructions, participants completed a short practice passage with attention probes and then advanced through the 7 passages in a fixed order.

Sustained Attention to Response Task (SART) The SART is "go/no-go" task in which participants make a response to "non-targets" and withhold that response to "targets." Each trial in the SART consisted of a # symbol (42 pt) presented for 1200 ms followed by a single uppercase letter (36 pt) presented for 800 ms. All stimuli were presented in the center of the computer monitor in black Arial Monospaced font on a grey background. Participants were told to make no response to the target (the letter "X") and to respond by pressing the letter "m" with the index finger of their dominant hand for all non-targets (all non-X letters of the alphabet).

Participants completed 4 blocks of 150 trials each. Blocks alternated between Low Frequency Target (12.5% of total trials) and High Frequency Target (50% of total trials) conditions. The exact sequence of trials within each block was randomly determined at the beginning of each block for each participant. Before the 4 blocks of test trials, participants completed a block of 15 practice trials with feedback in order to make sure they understood when to respond. The Low Frequency condition was expected to be significantly more challenging than the High Frequency condition. Block order was counterbalanced across participants and fixed across sessions.

Our dependent measures were target sensitivity (d_L) and response time variability (calculated as RT-CV, or the standard deviation of each participant's response latencies divided by their mean). Target sensitivity provides a measure of target accuracy while controlling for overall response bias (see Equation 1, in which H and FA refer to the proportions of hits and false alarms, respectively). RT-CV is generally thought to capture MW, with participants who are less attentive to the task showing greater variability in response latencies (McVay & Kane, 2009). RT-CV is typically negatively correlated with d_L.

(1) $d_L = \ln \{[H(1 - FA)]/[1 - H)FA]\}$

As a subjective measure of MW, participants were asked to complete the same retrospective attention report after the SART as the post-meditation report described above.

Procedure

While participants worked at individual computers, there were between 1 and 12 participants present in the testing room for each session. Session proctors were blind to the participants' meditation condition, as (i) presentation of written and spoken meditation materials was done automatically through E-Prime computer scripts, (ii) the audio instructions were delivered through closed headphones, and (iii) single sessions typically included participants in both meditation conditions. This ensured that proctors could not influence participants' expectations based on meditation group. Proctors actively monitored participants to ensure they were completing the expected tasks during their session. All participants completed all 4

sessions within a minimum of 7 and a maximum of 9 calendar days (days between Sessions 1 and 4: MMT: *M*=7.1 , *SD*=0.5; RMT: *M*=7.3, *SD*=0.6).

Results

Meditation Training Manipulation

Participants' self-reports were analyzed to examine whether the general effects of the meditation experience (level of relaxation and overall engagement) varied between groups (see Table 1). Participants in both the MMT and RMT groups reported feeling substantially more relaxed after each meditation session than they did before each session. A linear mixed-effect model was fit to the reported change in relaxation with fixed effects of Session (1 to 4) and Meditation Group (MMT or RMT) and random effects of Subject. The model revealed a significant Session X Meditation Group interaction (*b*=3.06, *SE*=1.47, *t*= 2.08). The RMT group, but not the MMT group, reported increasing changes in relaxation across the 4 sessions. This difference was largely driven by the fact that the RMT group reported a smaller change in relaxation after the first meditation session than the MMT group. After Session 1, the meditation was equally relaxing for both groups.

The same linear mixed-effect model as described was fit to percentage of On-Task time. The model revealed a significant effect of Session (*b*=4.20, *SE*=0.77, *t*=5.49) but no effect of Group or interaction. The effect of Session is likely due to the increase in On-Task reported for Session 4 (which was only 10 minutes in duration). Participants in both groups were equally likely to report having used the meditation techniques at home since the last session (59% of the MMT group and 65% of the RMT group, $\chi^2(1)<1.0$)) Thus, participants in both groups were equally engaged in the meditations.

Table 1. Mean (SD) self-reported changes in relaxation (scale 0-100) and percent of time on task for each meditation session. Note that sessions varied in duration.

	Session	1	2	3	4
Relax Change	MMT	23.0 (18.6)	20.2 (14.6)	20.3 (14.7)	20.9 (17.0)
	RMT	14.8 (18.7)	23.6 (17.3)	24.8 (19.0)	22.6 (20.6)
% OnTask	MMT	49.3 (21.8)	44.7 (22.1)	47.6 (23.9)	61.1 (21.6)
	RMT	45.2 (22.6)	45.8 (20.8)	48.4 (21.5)	59.5 (24.6)

Text Comprehension

Descriptive statistics for accuracy on Sentence-Linked and General questions and amount of MW are provided for both groups in Table 2.

MMT facilitates retention of details A multi-level logistic model was fit to Sentence-Linked question accuracy with fixed effects of Group (MMT or RMT) and Session (1 or 4) and random effects of Subject, Item and Subject X Session interaction. Group and Session were dummy coded with MMT and Session 1 as the reference groups. There was a main effect of Session (*b*=0.38, *SE*=0.16, *z*= 2.34, *p*=0.02) and a significant Group X Session interaction (*b*=-0.47, *SE*=0.22, *z*= -2.12 , *p*=0.03) and no main effect of Group (*p*=0.87). Thus, Sentence-Linked accuracy increased significantly from Session 1 to 4 for participants in the MMT group but not participants in the RMT group. When the same model was fit to the General question accuracy and to the amount of reported MW, no main effects or interactions were significant for either measure.

Table 2. Means (SD) for test accuracy and reported MW (percentage of "yes" responses) for text comprehension.

	MMT		RMT	
Session	1	4	1	4
Sentence-Linked	0.75 (0.1)	0.80 (0.1)	0.75 (0.2)	0.73 (0.2)
General	0.73 (0.1)	0.73 (0.1)	0.73 (0.1)	0.75 (0.1)
MW	0.28 (0.16)	0.28 (0.23)	0.34 (0.20)	0.30 (0.21)

Effects of MW equivalent across groups and sessions Participants' mean accuracy for Sentence-Linked questions are illustrated in Figure 1, split by whether participants reported mind-wandering during the corresponding sentence. As is clear from the figure, participants had significantly better comprehension of sentences read with their full attention than sentences on which they reported MW. A multi-level logistic model was fit to the raw Sentence-Linked question accuracy with fixed effects of Meditation Group, Session, reported MW (Yes or No) and their interactions and random effects of Subject, Item and Subject X Session interaction. The model revealed a significant main effect of MW (*b*= -0.69, *SE*=0.12, *z*= -5.9, *p*<0.001) and no other significant effects (all *p*>0.3).

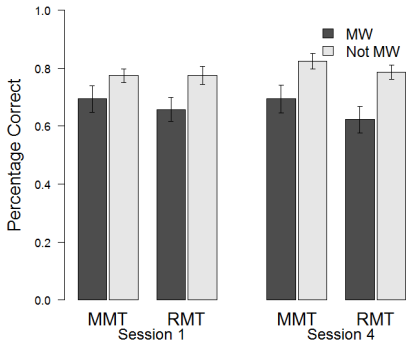


Figure 1. Accuracy on Sentence-Linked questions split by participants' responses to subsequent MW probes.

In addition to MW affecting comprehension of what is read *during* a MW episode, MW may also negatively influence the reader's ability to integrate information about

the passage as a whole. We tested the relationship between total amount of MW reported during a passage and participants' accuracy on the General questions for that passage. Participants' mean accuracy for General questions for each passage were fit with a linear mixed-effects model with fixed effects of Meditation Group (MMT or RMT), Session (1 or 4) and reported MW (Yes or No), and random factors of Subject and Passage. Interactions were tested, but in no case were interactions between the fixed effects significant. Participants who reported more MW during a passage were significantly less accurate on the General questions for that passage (b= -0.11, SE=0.03, t= -4.0) and the MW effect was equivalent across Session and Group.

Sustained Attention

Target detection and MW were equivalent across groups

The data reveal the general pattern of effects expected for the SART frequency manipulation: Participants had higher d_L and lower RT-CV (MW) on high-frequency target than low-frequency target blocks (see Table 3). Greater RT-CV was also associated with lower d_L for both conditions, for both Groups and Sessions (r(75)=-0.64 to -0.75, all p<0.001).

Table 3. Means (SD) for SART dependent measures.

	MMT		RMT	
Session	1	4	1	4
d_L Low Freq	5.39 (1.86)	6.04 (1.57)	5.75 (2.08)	5.99 (2.15)
d_L High Freq	7.0 (1.88)	7.37 (1.78)	7.23 (1.72)	7.38 (1.93)
RT-CV Low Freq	0.26 (0.08)	0.23 (0.06)	0.24 (0.08)	0.23 (0.09)
RT-CV High Freq	0.22 (0.06)	0.23 (0.06)	0.21 (0.07)	0.22 (0.07)
%TUT self-report	29.6 (21.0)	21.5 (19.3)	28.3 (22.6)	24.3 (24.0)

There was no evidence for a differential effect of MMT over RMT on target sensitivity, though d_L scores improved overall across sessions: A multi-level linear model was fit to the target sensitivity scores with fixed effects of Meditation Group, Session and Target Frequency (High or Low) and random effects of Subject, Subject X Session and Subject X Target Frequency interactions. The model revealed a very strong and significant main effect of Target Frequency (b=1.44, SE=0.102, t=14.31) and a smaller main effect of Session (b=0.36, SE=0.172, t=2.07), but neither of these factors interacted with Group, nor was there a main effect of Group. The same interaction model as above was fit to RT-CV and revealed only a significant effect of Target Frequency (b=-0.021, SE=0.004, t=-5.42), with no interactions.

While RT-CV did not decrease after training, participants in both groups reported greater task engagement during SART in Session 4 relative to Session 1. A multi-level linear model was fit to reported TUTs with fixed effects of Meditation Group and Session and random effects of Subject. The model revealed a significant effect of Session (b=-5.92, SE=2.53, t=-2.34) but no main effect of or interaction with Meditation Group[1].

Discussion

Our data reveal at most mixed results with respect to the hypothesized effect of MMT influencing cognitive performance through improvements in attention control. We found that novice practitioners who completed MMT improved more at retrieving specific details from a passage than those who completed RMT. However, this change in performance was not found for general comprehension and did not appear to be facilitated by decreased MW, as we hypothesized, as the two groups reported similar amounts of MW during reading at post-test. The sustained attention task revealed no significant group differences across sessions.

A limitation of the current design is the reliance on self-report of MW during text comprehension. It is possible that the attention-monitoring instructions in the MMT made participants more *aware* of their MW without giving participants enough practice to better control it (Davidson & Kazniak, 2015). Such increased awareness might result in participants reporting MW at equal or higher rates even if objectively they were MW less. This is particularly plausible since MW is a graded phenomenon, and participants might initially consider themselves "on-task" if they were even partially engaged in the task but after MMT might consider themselves "off-task" at the same level of engagement.

We did find a consistent detrimental effect of MW on reading comprehension. Both recall of specific details and synthesis of the passage as a whole were negatively affected by MW (see also Feng et al., 2013; Smallwood et al., 2008). We also found that increased MW, as measured by both an objective measure (response time variability) and retrospective self-report, was detrimental to target detection. These results are important for establishing key paradigms that produce reliable effects of MW and can therefore be used to test the effects of attention training.

Given the sensitivity of the outcome measures to MW, the fact that we did not find that MMT was more beneficial than RMT for reducing MW or for improving most objective performance markers is revealing. This pattern stands in contrast to that found by Mrazek and colleagues (2013), who found increases in reading comprehension accuracy were mediated by decreases in MW after MMT. Our study differs from that of Mrazek in several ways that might account for this divergence. First, our training was about a

[1] Individual models fit to the MW measures for the sustained attention task do find significant effects of session for the MMT group for both the objective measure, RT-CV (b=-0.03, SE=0.01, t=-2.47) and the subjective measure, reported TUTs (b=-8.05, SE=3.89, t=-2.07). The RMT group did not have significant effects of session for either RT-CV (b=0, SE=0.01, t<1) or TUTs (b=-3.95, SE=3.29, t=-1.20). This is at most weak evidence of an effect of MMT on MW during the sustained attention task.

quarter of the duration of that employed by Mrazek et al. and therefore may have not been long enough to influence MW (though note that benefits of MMT have been found after interventions as short as 8 minutes (Mrazek et al., 2012)). Second, the structure of our reading task prevented rereading and this may have affected the relationship between MMT and participants' comprehension. One possibility is that MMT facilitated partial engagement during MW and this partial engagement was sufficient to support encoding of some additional details but not to support the abstract schema formation required for synthesis of the passage as a whole. Further study is required to determine how MMT, MW and reading comprehension may be related under different task demands.

In the present study, participants in both groups showed improvements in sustained attention at post-test. Because we did not have a non-meditation control group, we cannot say whether this change in performance is due to non-specific benefits of meditation or are practice effects. One recent study did find group differences in cognition (working memory) when comparing MMT specifically to RMT (Banks et al., 2015). However, the role of MMT appeared to be one of protecting cognition from the effects of stress rather than an overall benefit to cognition and the relationship between meditation, MW and cognition was not tested. One possibility for future research would be to focus more specifically on the effects of stress on performance, rather than on MW in the absence of stress.

Our experimental manipulation, comparing meditation training with and without instructions to monitor attention, was necessarily subtle. Both groups reported similar changes in relaxation and engagement with the meditation. It's further possible that asking about MW during pretest prompted the RMT group to closely monitor their attention even though we didn't instruct them that attention control was the objective of their training. The current study is an initial attempt at the dual-blind design advocated by Davison and Kasniak (2015). However, the present study necessarily samples a very specific set of the many possible parameters for this type of intervention. It is certainly possible that changing the duration of training, its mode of delivery or other factors could result in different patterns between the MMT and RMT groups. However, it is precisely this type of comparison that will allow us to understand how *and why* different meditation practices may influence cognition.

Acknowledgments

This material is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. We wish to thank Nick Davey and Patrick Cushen for their role in developing the meditation materials, Ashlyn Testut, Esther Robinson and Leon Li for data collection and Stefanie Kuchinsky and Nick Pandza for helpful discussion.

References

Banks, J. B., Welhaf, M. S., Srour, A. (2015). The protective effects of brief mindfulness meditation training. *Consciousness and Cognition, 33*, 277–285.

Braver, T. S., Barch, D. M., Gray, J. R., Molfese, D. L., Snyder, A. (2001). Anterior cingulate cortex and response conflict: Effects of frequency inhibition and errors. *Cerebral Cortex. 11*(9): 825-836.

Davidson, R. J., & Kaszniak, A. W. (2015). Conceptual and methodological issues in research on mindfulness and meditation. *American Psychologist*, *70*, 581-592.

Feng, S., D'Mello, S., & Graesser, A. C. (2013). Mind-wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review, 20*, 586–592.

Jha, A. P., Morrison, A. B., Dainer-Best, J., Parker, S., Rostrup, N., & Stanley, E. A. (2015). Minds at attention: Mindfulness training curbs attentional lapses in military cohorts. *PloS one*, *10*(2), e0116889.

Jha, A. P., Stanley, E. A., Kiyonaga, A., Wong, L., Gelfand, L. (2010). Examining the protective effects of mindfulness training on working memory capacity and affective experience. *Emotion 10*: 54-64.

Kabat-Zinn, J. (1994). *Wherever you go, there you are: Mindfulness meditation in everyday life*. New York: Hyperion.

MacLean, K. A., Ferrer, E., Aichele, S. R., Bridwell, D. A., Zanesco, A. P., Jacobs, T. L., King, B. G., Rosenberg, E. L., Sahdra, B. K., Shaver, P. R., Wallace, B. A., Mangun, G. R., Saron, C. D. (2010). Intensive meditation training improves perceptual discrimination and sustained attention. *Psychological Science, 21*(6), 829–839.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind-wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 196–204.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and GRE performance while reducing mind-wandering. *Psychological Science*, *24*(5), 776–781.

Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2012). Mindfulness and mind-wandering: Finding convergence through opposing constructs. *Emotion, 12*(3), 442–448.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory & Cognition, 36*(6), 1144–1150.

Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., & Goolkasian, P. (2010). Mindfulness meditation improves cognition: Evidence of brief mental training. *Consciousness and Cognition, 19*(2), 597–605.

Which Learning Algorithms Can Generalize Identity-Based Rules to Novel Inputs?

Paul Tupper (pft3@sfu.ca)
Department of Mathematics, Simon Fraser University
Burnaby, BC, V5A 1S6, Canada

Bobak Shahriari (bshahr@cs.ubc.ca)
Department of Computer Science, University of British Columbia
Vancouver, BC, V6T 1Z4, Canada.

Abstract

We propose a novel framework for the analysis of learning algorithms that allows us to say when such algorithms can and cannot generalize certain patterns from training data to test data. In particular we focus on situations where the rule that must be learned concerns two components of a stimulus being identical. We call such a basis for discrimination an *identity-based* rule. Identity-based rules have proven to be difficult or impossible for certain types of learning algorithms to acquire from limited datasets. This is in contrast to human behaviour on similar tasks. Here we provide a framework for rigorously establishing which learning algorithms will fail at generalizing identity-based rules to novel stimuli. We use this framework to show that such algorithms are unable to generalize identity-based rules to novel inputs unless trained on virtually all possible inputs. We demonstrate these results computationally with a multilayer feedforward neural network.

Keywords: phonology; learning algorithms; symmetries; connectionism

Introduction

Suppose a subject is asked to learn an artificial language in which all words consist of two letters. They are told that CC, AA, HH, EE, and RR are all examples of valid words in the language but that GA, EH, RA, ER, MG are not valid words. Now suppose that the learner is asked whether YY and YZ could be valid words in the language. Presumably they will say that YY could be a valid word in the language whereas YZ could not be. The obvious feature that all the valid words have in common is that they consist of two identical letters. This feature is not shared by the invalid words. We say in this case that the learners have learned an identity-based rule, and are able to generalize the rule to novel inputs.

We do not know if this exact experiment has ever been performed, but there have been analogous tests in the phonological domain (Berent, Marcus, Shimron, & Gafos, 2002; Gallagher, 2013). In artificial language learning tasks, human subjects are sensitive to identity relations between segments, and are able to generalize them to novel inputs. This kind of effect is not specific to language though: consider a task where subjects are presented with pictures of pairs of socks, and are asked to say whether they form a matching pair.

Surprisingly, given how obvious the above pattern is to human learners, many computer models of learning are not able to learn identity-based rules like those implicit in the data above, without being presented with nearly all possible inputs. These computational learners may give the same rating to both the forms YY and YZ, since neither of them have any similarity to the training words in a manner that is deemed relevant by the algorithms. Important classes of such algorithms include basic connectionist algorithms (Rumelhart & McClelland, 1988) and the "Plain" (Baseline version) of the UCLA Phonotactic Learner (Hayes & Wilson, 2008). There are ways to modify these algorithms to perform better on such tasks, for example, by introducing copying (Colavin, Levy, & Rose, 2010), special representations of identical segments in the input (Gallagher, 2013), or weight sharing across connections as is done in convolutional neural networks (LeCun & Bengio, 1995).

There are many informal arguments given for why the basic versions of these algorithms cannot learn identity-based rules. Such algorithms are unable to generalize "outside the training space" (Marcus, 2003), or "do not instantiate variables" (Berent, 2012). Though these terms describe a genuine limitation of such algorithms, they suffer the drawback of not being defined formally. Even though computational learners themselves are clearly defined, whether a particular algorithm is able to learn identity relations or instantiate variables is impossible to determine precisely since the criterion for these conditions is not formalized. Our present goal is to provide a rigorous framework for these informal statements about algorithms, and to provide criteria for when an algorithm cannot generalize identity-based rules to novel inputs.

In the following we define learning *algorithms*, *symmetries* of sets of words, and what it means for an algorithm to be *invariant* under a symmetry. In our main result we show that if an algorithm is invariant under some symmetry, and the training data is invariant under the same symmetry, then the algorithm cannot learn a grammar that is not invariant under that symmetry. As an application, we demonstrate a symmetry that identity-based rules are not invariant under, and then show that a wide class of algorithms are invariant under it. This means that such algorithms cannot learn identity-based grammars with invariant training data, in contrast to human performance on analogous tasks. We then demonstrate how feed-forward neural networks suffer from these limitations, independent of the number of hidden layers in the network.

Formal Definitions

We consider a set W, which we call the set of words, containing all well-formed inputs. We stress that in the linguis-

tic case W consists of both words that are good (grammatical) and words that are bad (ungrammatical). In what follows we will consider words to be strings of letters, but individual words can be anything, such as strings of segments or feature bundles.

To fix ideas, in what follows we will often consider a particular example of a set of words: we let $\widetilde{W}$ be the set of all two letter words, where the letters are capitals taken from the English alphabet, such as AA or MG.

We define the training data D to be a collection of word-rating pairs $\langle w, r\rangle$ where w is a word in W and r is a number interpreted as a *rating* of how "good" w is. For example, using the word set $\widetilde{W}$, a dataset D might consist of

$$\langle CC, 1\rangle,\ \langle AA, 1\rangle,\ \langle EE, 1\rangle,\ \langle GA, 0\rangle,\ \langle EH, 0\rangle,\ \langle RA, 0\rangle. \quad (1)$$

This dataset says that CC, AA, and EE have rating 1 (and thus are good words) and that GA, EH, and RA have rating 0 (and thus are bad words). Alternatively, in a training task where only good words are given to the learner, D might consist only of good words paired with the rating 1. But there are other possibilities: words could be paired with a rating given by their prevalence in a corpus, for example.

To formally define a learning algorithm, consider what a learning algorithm such as the UCLA Phonotactic Learner (Hayes & Wilson, 2008) does. First, a collection of data D is input to the algorithm and used to choose a set of parameters p in a model. We can formalize this as $p = \mathscr{A}(D)$. Once we have p, given any new input w the algorithm outputs a *score*, which we can formalize as $f(p, w)$. Typically, the computation of p from D is computationally intensive whereas once we have p, the score $f(p, w)$ is cheap to evaluate. This matches our experience of human behaviour where learning a language occurs over long periods of time, whereas judgements of the well-formedness of novel words are readily produced by adult speakers.

Here we will abstract away issues of parameter setting and computational effort and just view an algorithm as a map that takes a set of training data D and an input w and outputs a rating. We consider the map $\mathscr{L}$ given by

$$\mathscr{L}(D, w) = f(\mathscr{A}(D), w).$$

Specifically, a learning algorithm $\mathscr{L}$ is a map that takes training data D and word w and outputs a score $\mathscr{L}(D, w)$. The interpretation is that this score is what you would get if you used the data D to train the algorithm and then used the resulting computational model to evaluate the word w.

We note that we interpret both the ratings coupled with words in the training data D and the scores output by the algorithm $\mathscr{L}$ as measures of the goodness of a word. This is natural, since we expect the algorithm to give good scores to words that have high ratings in the training data. However, ratings and scores are distinct in general; for example, ratings in D could be how common a word is in a corpus and scores from $\mathscr{L}$ could be intended to model how well-formed a word is on a scale from 0 to 1.

We define a symmetry σ to be a bijective map from the set of words W to itself; in other words, a map such that $\sigma(w)$ is in W for all w in W, and for all v in W there is an u in W such that $\sigma(u) = v$. As an example of a symmetry, let $\tilde{\sigma}$ be the map from $\widetilde{W}$ to itself given by

$$\tilde{\sigma}(\mathbb{X}\mathbb{Y}) = \mathbb{Y}\mathbb{X}, \quad (2)$$

for any letters $\mathbb{X}$, $\mathbb{Y}$. Thus the symmetry $\tilde{\sigma}$ reverses the order of letters in two-letter words.

We introduce symmetries in order to analyze algorithms: we are not claiming that they have any psychological or linguistic reality. Indeed, as far as we know all maps that are naturally occurring phonological processes are not symmetries. For one thing, most phonological maps satisfy $\sigma(\sigma(x)) = \sigma(x)$ for all x (also known as idempotency (Magri, 2015)). But this can only happen with a symmetry if $\sigma(x) = x$ for all x, meaning that σ does nothing.

A word w is invariant under a symmetry σ if $\sigma(w) = w$. To apply a symmetry to a set of training data, we say that σ just acts on each word in every word-rating pair in the data set, but does not change the rating of that word. So if the word-rating pair $\langle w, r\rangle$ is in D, then the pair $\langle \sigma(w), r\rangle$ is in $\sigma(D)$. For example, if we applied $\tilde{\sigma}$ (as defined in (2)) to the dataset in (1) we would get the dataset

$$\langle CC, 1\rangle,\ \langle AA, 1\rangle,\ \langle EE, 1\rangle,\ \langle AG, 0\rangle,\ \langle HE, 0\rangle,\ \langle AR, 0\rangle.$$

We say that a dataset D is invariant under a symmetry σ if $\sigma(D)$ has precisely the same word-rating pairs as D. The simplest way for data D to be invariant under a symmetry σ is if each word in each word-rating pair in D is invariant under σ. But there are other ways. For example, the symmetry $\tilde{\sigma}$ leaves the data

$$\langle BB, 1\rangle,\quad \langle GG, 2\rangle,\quad \langle EE, 0\rangle$$

invariant because the words BB, GG, EE are all invariant under $\tilde{\sigma}$. On the other hand the data

$$\langle BG, 1\rangle,\quad \langle GB, 1\rangle,\quad \langle EA, 2\rangle,\quad \langle AE, 2\rangle$$

is also invariant under $\tilde{\sigma}$, but in this case the individual words are not invariant, it is just that w and $\sigma(w)$ always have the same rating in this data set.

We say an algorithm $\mathscr{L}$ is invariant under σ if $\mathscr{L}(\sigma(D), \sigma(w)) = \mathscr{L}(D, w)$ for all D and w. In words, the rating that the algorithm gives to w when trained on D is the same that the algorithm gives to $\sigma(w)$ when trained on $\sigma(D)$.

Our main result is a simple consequence of these definitions.

Theorem 1 *If algorithm $\mathscr{L}$ and training data D are invariant under symmetry σ then*

$$\mathscr{L}(D, w) = \mathscr{L}(D, \sigma(w)),$$

for all w in W. In other words, the algorithm $\mathscr{L}$ gives the same rating to w and $\sigma(w)$ when trained on D.

Proof. We have

$$\mathscr{L}(D,w) = \mathscr{L}(\sigma(D),\sigma(w)) = \mathscr{L}(D,\sigma(w))$$

where the first equality follows from the invariance of $\mathscr{L}$ under σ, and the second inequality follows from the invariance of D under σ. □

Example: Consider a language containing 10 letters, each letter having a sonority value between 1 and 5 according to the following table. (Sonority is an abstract phonological variable, roughly corresponding to how close a segment is to a vowel.) Words in the language consist of only two letters. Suppose that all words in the language have increasing or constant sonority. So, BA, MO, ZW, BD could all be words in the language, but AD, AN, and WV could not be. Consider the letter reversing symmetry $\tilde{\sigma}$ given in (2). If you apply $\tilde{\sigma}$ to an ungrammatical word (e.g. AB) you get a grammatical word (BA). If you apply $\tilde{\sigma}$ to a grammatical word with increasing sonority you get an ungrammatical word. Words with two letters of the same sonority give you back another word with letters of the same sonority.

Table 1: Segments in a Hypothetical Language

segments	sonority
A O	5
W Y	4
M N	3
V Z	2
B D	1

Now suppose you have a learning algorithm $\mathscr{L}$ that is invariant under $\tilde{\sigma}$. This means that if you take a data set D, train the algorithm on it, and then use the algorithm to evaluate word w, you will get the same result if you train the algorithm on $\tilde{\sigma}(D)$ (in which all the words are reversed) and then use the algorithm to evaluate $\tilde{\sigma}(w)$, which is just the reversal of w.

Suppose we give the algorithm data D that is invariant under $\tilde{\sigma}$. For simplicity we assume that D consists only of grammatical words each assigned the rating 1. In this case, the only way D can be invariant under $\tilde{\sigma}$ is if all the words in D have constant sonority, and for every such word $\mathbb{X}\mathbb{Y}$ in D, $\mathbb{Y}\mathbb{X}$ is also in D. Can the algorithm correctly learn the generalization that words in the language must have increasing or level sonority from this data set?

Theorem 1 shows that it cannot, as follows. According to the theorem, $\mathscr{L}(D,w) = \mathscr{L}(D,\tilde{\sigma}(w))$. All we need to do is let w equal a word of increasing sonority, such as BA, to see that the algorithm with training data D gives the same score to BA and AB. Since the first is grammatical and the second is ungrammatical, the algorithm clearly has not learned the correct rule governing grammaticality in the language. This is pretty commonsensical: one way to think of it is that there is nothing in the algorithm or the training data to make the algorithm prefer AB to BA, since both the algorithm and the training data are invariant under $\tilde{\sigma}$, and $BA = \tilde{\sigma}(AB)$. Of course, this is not necessarily a defect of the algorithm $\mathscr{L}$; if some words with increasing or decreasing sonority were included in D, then D would not be invariant under $\tilde{\sigma}$, and $\mathscr{L}$ could learn the grammar.

In the next section we will give a less straightforward example, allowing us to formalize the idea of identity-based rules for learning algorithms.

Identity-Based Rules

We now use the above result to show that certain algorithms cannot learn identity-based rules unless trained on words containing virtually all letters in the alphabet. That is, the algorithm cannot extend the identity-based rules to words containing letters that it has not explicitly been trained on. This is in sharp contrast to human learners who are able to generalize identify-based rules (in the phonological context, for example) to segments they have not encountered before(Berent et al., 2002).

We return to the example at the beginning of the paper: $\widetilde{W}$ is the set of all words consisting of two letters. We stipulate that grammatical words are those consisting of two identical letters and all other words are ungrammatical. Suppose we want the algorithm to learn this grammar, but train it on data omitting any words containing the letters Y and Z. What algorithms will not be able to learn the correct grammar under these conditions?

Define the symmetry σ of W by the following:

$$\sigma(\mathbb{X}_1 Y) = \mathbb{X}_1 Z, \quad \sigma(\mathbb{X}_1 Z) = \mathbb{X}_1 Y, \quad \sigma(\mathbb{X}_1\mathbb{X}_2) = \mathbb{X}_1\mathbb{X}_2,$$

for all letters $\mathbb{X}_1, \mathbb{X}_2$, with $\mathbb{X}_2$ not equal to Y or Z. In other words, if the second segment is Y, σ changes it to Z, if the second segment is Z, σ changes it to Y, and if the second segment is neither, then the word is unchanged.

Now suppose our training data D contains no words with either segments Y or Z as the second segment. D may contain both grammatical words (e.g. CC) with rating 1 and ungrammatical words (e.g. CE) with rating 0. Then D is invariant under σ. Theorem 1 shows that if the algorithm $\mathscr{L}$ is also invariant under σ then it will give the same rating to w and $\sigma(w)$ for any word W when trained on D. In that case we would have that it gives the same rating to the words YY and YZ, showing that it cannot learn the identity based grammar.

Below we provide an example of an algorithm invariant under this symmetry. But in general we informally argue that any algorithm that does not in some way explicitly check for identity between the letters, or somehow enforce a similar treatment of those two letters in processing, cannot correctly learn that YY is a more well formed word than YZ, if it is never given words with a second letter Y or Z as training data.

Randomized Algorithms

Many algorithms for learning use randomness at some point in their operation. It may either be in the computation that

takes the input data to the parameters p (for example, by which order the input words are used) or in the map from the parameters and a new input word to a word score s. In the former case $p = \mathscr{A}(D)$ is a random function of D; in the latter $s = f(p,w)$ is a random function of p and w. In either case, this leads to $\mathscr{L}(D,w)$ being random for any fixed D and w.

Under these conditions, it is unlikely that invariance of the form described above will hold. Instead we now define invariance of $\mathscr{L}$ under σ to be

$$\mathbb{E}\mathscr{L}(\sigma(D),\sigma(w)) = \mathbb{E}\mathscr{L}(D,w),$$

where $\mathbb{E}$ denotes expectation. (If X is a random variable, $\mathbb{E}X$ is approximately what we would get if we took the average of a large number of samples of X.)

We now get the same result as before. This is a strictly stronger result than Theorem 1, since a deterministic algorithm is just a special case of a randomized algorithm.

Theorem 2 *If random algorithm $\mathscr{L}$ and training data D are invariant under symmetry σ then*

$$\mathbb{E}\mathscr{L}(D,w) = \mathbb{E}\mathscr{L}(D,\sigma(w)),$$

for all w in W. In other words, the algorithm $\mathscr{L}$ gives on average the same rating to w and $\sigma(w)$ when trained on D.

Proof. We have

$$\mathbb{E}\mathscr{L}(D,w) = \mathbb{E}\mathscr{L}(\sigma(D),\sigma(w)) = \mathbb{E}\mathscr{L}(D,\sigma(w))$$

where the first equality follows from the invariance of $\mathscr{L}$ under σ, and the second inequality follows from the invariance of D under σ. □

Experiments

We demonstrate the consequences of our theorems in a computational experiment where we use a deep neural network to learn the grammar described in our introduction. The networks are trained using data in which two-letter words with two identical letters are good, and two-letter words with two different letters are bad. The network is then asked to assess novel words containing segments it has not seen in the training set. Randomness enters into the training of these networks in various places and so Theorem 2 is the relevant result in this case. Consequently, we do not compare individual trainings of the network on the novel stimuli. For each novel stimulus we train the network numerous times and take the average score over all the trainings. It is these scores that are compared between stimuli.

Task and Dataset

Before discussing the neural network learners that were tested, we describe the dataset and task that was required of them. As before, our set of words W consisted of all two letter words with letters running from A to Z. The training set consisted of the 24 words AA, BB, …, XX paired with rating 1, along with 48 randomly generated words with mismatched segments taken from the list A, …, X, each paired with rating 0.

To assess the ability of the learner to generalize to novel inputs, after training we tested it on the words

$$\text{YY},\ \text{ZZ},\ \mathbb{X}\text{Y},\ \text{YZ},\ \mathbb{X}\text{Z},\ \text{ZY},$$

where $\mathbb{X} \in \{\text{A},\text{B},\ldots,\text{X}\}$ were randomly selected. For each learner, the experiment was independently repeated 40 times with different random seeds.

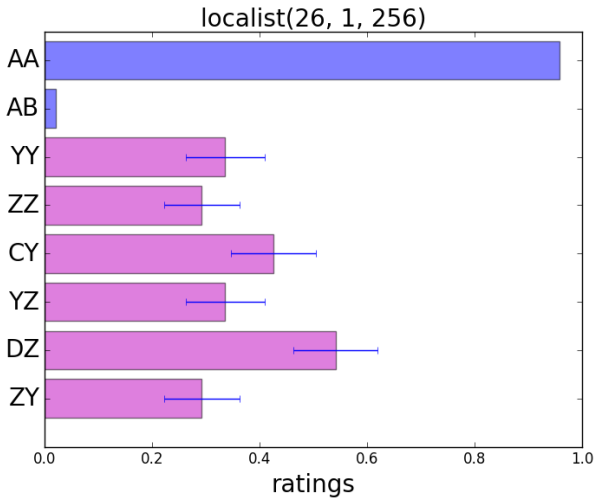
Encodings. We distinguish two different representations for the segments A to Z, namely the localist and distributed encodings. Both of these representations use a fixed length bit string. However, while localist codes (also known as 1-of-k encoding) are constrained to include a single non-zero bit, distributed codes can be any arbitrary combination of k bits, for some fixed k. Distributed encodings are a much more compact representation of data; indeed, for the same string-length k, we can represent an exponentially large number of segments 2^k. The experiment was run on both types of encoding with $k = 26$. When distributed encoding was used, codes for each letter were randomly generated each repetition, so the exact encoding of the segment X, for instance, is almost certainly different between two repetitions of a given run.

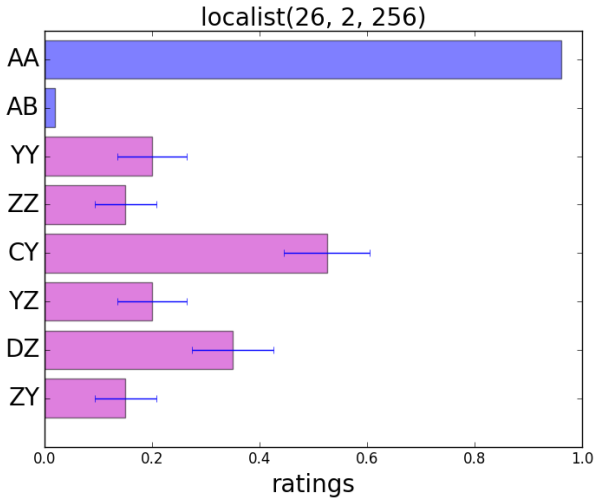
Neural Network Learners

We tested our theoretical findings on the most popular model in the machine learning literature today: the artificial neural network. The words were fed into the neural network by simply concatenating the two 26-bit codes of their letters. We experimented with many different architectures, ranging from one to three hidden layers, and from 256 to 1024 units per layer, with tanh nonlinearities for all hidden units. We trained the models via backpropagation using an iterative quasi-Newton gradient descent algorithm called the limited memory Broyden-Fletcher-Goldfarb-Shanno method (L-BFGS), with a maximum of 100 iterations. Both the neural network and its optimization are implemented in `torch` (Collobert, Bengio, & Mariéthoz, 2002).

Results

We present results for the case of each hidden layer having 256 units, as the results are similar for more units per hidden layer. In Figure 1, for the localist encoding, we plot the average score output by the neural network for each of the test words above, for 1, 2, and 3 hidden layers. In addition, the averaged training scores are reported in the top two bars of each panel.





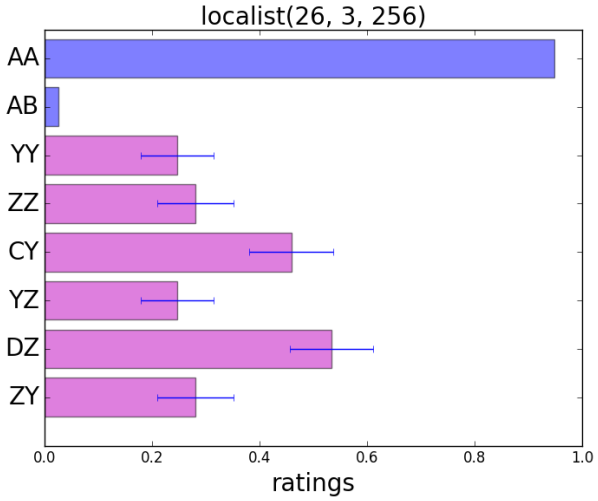
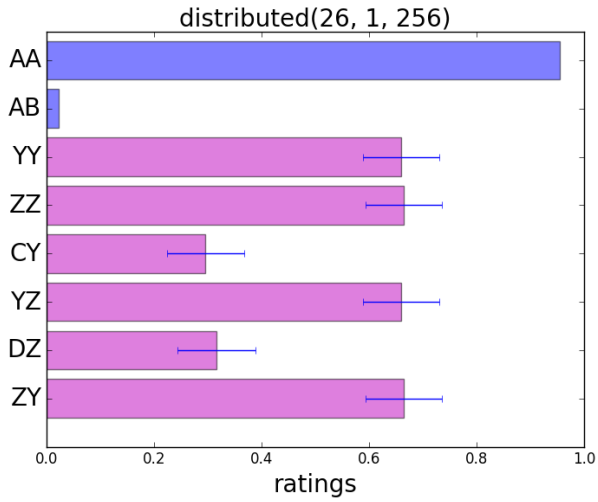
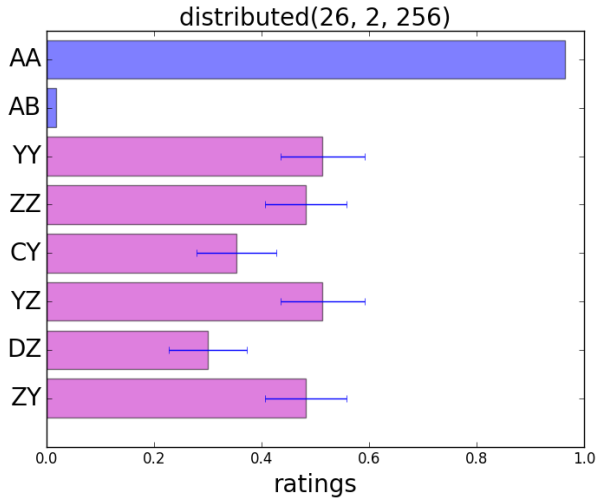


Figure 1: Scores for various words for the network with localist encoding for 1, 2, and 3 hidden layers.





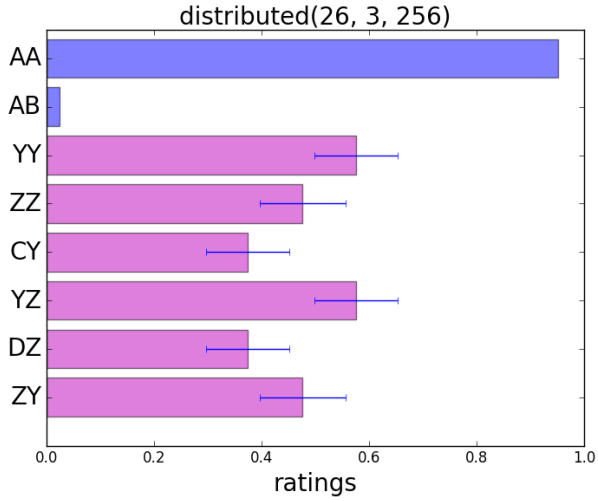


Figure 2: Scores for various words for the network with distributed encoding for 1, 2, and 3 hidden layers.

Looking at the top plot in the figure, showing the results for one hidden layer, the words YY and ZZ get scores of around 0.3 in contrast to the score of near 1 given for the well-formed input AA. The networks are unable to determine that YY and ZZ are grammatical. Likewise, the other test words with differing segments and containing the segments Y or Z have scores ranging from approximately 0.3 to 0.5. The networks are not able to distinguish between grammatical and ungrammatical words in this case.

The ability of the networks to generalize to novel inputs is not improved by adding further hidden layers. The second and third plots in Figure 1, corresponding to two and three hidden layers, show very similar results to the first. To within statistical accuracy, the scores for YY, ZZ, YZ, and ZY are all the same. The networks are not able to discriminate between grammatical and ungrammatical words when the words included the novel segments Y and Z.

This poor performance is perhaps not surprising for the localist encoding, as observed by Marcus (Marcus, 2003): in the localist encoding, introducing new segments correspond to activating new input units that were never active during training, and therefore whose weights never changed from their random initializations. However, in Figure 2 we show that the poor performance remains true in the case of distributed representations. In the first plot, we show the results for a single hidden layer. The networks give a rating higher than 0.5 for both YY and ZZ, which is higher than the score given by the localist networks, but the same high rating is given to the words YZ and ZY. A similar pattern is repeated for the two and three-layer case. The networks are not able to discriminate between grammatical and ungrammatical words containing novel segments, even when distributed representations are used.

Discussion

That connectionist networks are unable to generalize what are sometimes called "algebraic" rules to novel inputs is not a new observation (Marcus, 2003; Berent, 2012). Our contribution has been to give a formalized description and proof of this phenomenon. Furthermore, our results and computer experiments reinforce that Deep Learning, in the form of the ability to train connectionist networks with multiple hidden layers, does not alone overcome these limitations.

Acknowledgments

The authors thank Nilima Nigam for comments on an earlier draft of this manuscript. PT was supported by an NSERC Discovery Grant, a Research Accelerator Supplement, and held a Tier II Canada Research Chair. BS was supported by an NSERC Discovery Grant.

References

Berent, I. (2012). *The phonological mind*. Cambridge University Press.

Berent, I., Marcus, G. F., Shimron, J., & Gafos, A. I. (2002). The scope of linguistic generalizations: evidence from Hebrew word formation. *Cognition*, *83*(2), 113 - 139.

Colavin, R., Levy, R., & Rose, S. (2010). Modeling OCP-place in Amharic with the Maximum Entropy phonotactic learner. In *Proceedings volume of the 46th meeting of the Chicago Linguistics Society.*

Collobert, R., Bengio, S., & Mariéthoz, J. (2002). Torch: a modular machine learning software library. *Technical Report IDIAP-RR 02-46*.

Gallagher, G. (2013, 8). Learning the identity effect as an artificial language: bias and generalisation. *Phonology*, *30*, 253–295. doi: 10.1017/S0952675713000134

Hayes, B., & Wilson, C. (2008, 2016/01/09). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10).

Magri, G. (2015). Idempotency and chain shifts. In K.-m. Kim (Ed.), *Proceedings of WCCFL 33: the 33rd annual West Coast Conference in Formal Linguistics.* Cascadilla Proceedings Project.

Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing* (Vol. 1). IEEE.

Pragmatic relativity: Gender and context affect the use of personal pronouns in discourse differentially across languages

Zeynep Azar (zeynep.azar@mpi.nl)
Radboud University, Center for Language Studies, Wundtlaan 1
6525 XD Nijmegen, The Netherlands

Ad Backus (A.M.Backus@uvt.nl)
Tilburg University, Department of Culture Studies, Warendelaan
5037 AB Tilburg, The Netherlands

Aslı Özyürek (asli.ozyurek@mpi.nl)
Radboud University, Center for Language Studies, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Wundtlaan 1
6525 XD Nijmegen, The Netherlands

Abstract

Speakers need to use a variety of referring expressions (REs) (e.g. full noun phrases, pronouns or null forms) in pragmatically appropriate ways to produce coherent narratives. Languages, however, differ from each other in terms of a) whether REs as arguments can be dropped or not and b) whether personal pronouns encode gender or not. Here we examine two languages that differ from each other in these two aspects and ask whether the co-reference context (i.e., referents are maintained or re-introduced) and the gender encoding options affect the use of REs differentially. We elicited narratives from Dutch and Turkish speakers about two types of three-person events, one including people of the same and the other of mixed-gender. Speakers of both languages followed a general principle of using full forms such as noun phrases (NPs) while re-introducing a previously mentioned referent into the discourse and reduced forms (overt or null pronoun) while maintaining the same referent; a language independent strategy in discourse production. Turkish speakers, unlike Dutch speakers, used pronouns mainly to mark emphasis. Furthermore, Dutch but not Turkish speakers used pronouns differentially across the two videos. Thus, we argue that linguistic possibilities available in typologically different languages might tune speakers into taking different principles into account to establish coherence in narratives in pragmatically coherent ways.

Keywords: referring expressions; gender encoding; pronouns; cross-linguistic comparison; discourse production

Introduction

Throughout discourse, speakers often refer to the same entities, but they do not always use the same referring expression (RE). It has been suggested that speakers are sensitive to the information status of the referents: They use fuller forms (e.g. a full noun phrase) to (re-)introduce a new, less accessible referent, and prefer reduced forms (e.g. overt or null pronouns) when maintaining an already given, more accessible referent (Ariel, 1990; Chafe, 1976). A general conclusion of these studies is that speakers encode only as much information as their addressee needs in order to uniquely identify the intended referent. Since an NP encodes more semantic information, it is preferred for less accessible referents while pronouns, which do not encode much semantic information, are used for maintained that is for more accessible referents. Therefore, the accessibility of a referent and the amount of information encoded in the referring expression that refers to it show an inverse relation. Speakers of different languages seem to follow the same strategy even though languages differ in the forms they make use of, such as whether dropping arguments is allowed or not (Azar and Özyürek, 2015; Berman and Slobin 1994; Perniss and Özyürek, 2015; Yoshioko, 2008). The difference between languages which do allow argument drop (pro-drop languages) and those that do not (non-pro-drop languages) surfaces mostly in contexts where referents are maintained. The default form in non-pro-drop languages is the overt pronoun while for pro-drop languages it is the null pronoun (Carminati, 2002).

This picture raises the question of what function personal pronouns has in tracking reference in different contexts, especially in pro-drop languages. To our knowledge, there are only a few studies comparing pro-drop to non-pro-drop languages directly in a discourse elicitation task (e.g., Aksu-Koç & Nicolopoulou, 2015; Perniss and Özyürek, 2015; Yoshioko, 2008). However, Perniss and Özyürek (2015) collapsed NPs and overt pronouns into one category, *overt* and Aksu-Koç and Nicolopoulou collapsed overt and null pronouns into one category, *pronoun*. Thus there are not enough studies showing how a specific category of RE, especially the overt pronoun, is used across typologically different languages in the same discourse context.

Languages also differ from each other in whether personal pronouns encode gender, and how this affects the organization of extended discourse across languages is not known. It is possible that in addition to general principles of accessibility, there might be other principles that are differentially taken into account during the choice of a specific RE in discourse.

To investigate these questions, we compare adult speakers of two typologically different languages, Turkish (pro-drop) and Dutch (non-pro-drop), in an elicited discourse study. While Turkish has a single third person pronoun, Dutch

third person singular pronouns are gender specific. We manipulated the gender of the characters in the events in our experimental stimuli to see whether Turkish and Dutch speakers organize their discourse and their use of REs differently, especially in the case of pronouns. We also examine the specific functions of personal pronouns in discourse narratives of Turkish speakers.

Linguistic characteristics of Turkish and Dutch

One of the differences between Turkish and Dutch that is relevant for reference tracking is that they differ in which specific context null pronouns are allowed. Overt realization of arguments is normally required for grammatical productions in Dutch. However, subject arguments may be left unmentioned when consecutive clauses refer to consecutive actions performed by the same subject. Null subjects have been reported to be relatively low in Dutch discourse narratives (e.g., 23.38% in Flecken, 2011; 12% in Gullberg, 2006). In Turkish, however, in maintained contexts the null pronoun is the default form. It is argued that an overt pronoun is used only when the referent has an emphatic or contrastive function (Enç, 1986). Some of the studies on Turkish, however, are based on native intuitions of the authors and not on empirical data (Enç, 1986; Erguvanli-Taylan; 1986; Özsoy, 1987) or on written discourse (Kerslake, 1987; Turan, 1995). A few studies with discourse production data from adults in an experimental setting either focused on only one RE form (e.g. Küntay, 2002, NPs only) or collapsed overt and null pronoun into one category (Aksu-Koç & Nicolopoulou, 2015) or they had a relatively small sample size (Azar & Özyürek, 2015). Thus, the function of overt pronouns in relation to null pronouns in tracking referents in adult discourse in Turkish and especially in marking maintained and re-introduced referents is still not very clear.

The other relevant difference between the Turkish and Dutch is (+/-) gender encoding of the personal pronouns. Turkish third-person pronoun (*o* for singular and *onlar* for plural) does not encode gender and is the equivalent of *he*, *she*, *it* in English. On the other hand, Dutch singular third person pronouns *hij* (he) and *ze*/ *zij* (she) are gender specific.

Predictions

Concerning the general reference tracking strategies, we expect both Turkish and Dutch speakers to re-introduce referents dominantly with NPs, thus with fuller forms. We expect both groups of speakers to prefer maintaining referents with reduced forms, null pronoun for Turkish (pro-drop) speakers and overt pronoun for Dutch (non-pro-drop) speakers. As for gender, we expect Dutch speakers to use more overt pronouns while narrating the mixed-gender video compared to the same-gender video, specifically in the re-introduction context. In that context, speakers can refer back to a previously mentioned referent with a personal pronoun encoding gender without causing an ambiguous interpretation of the pronoun. We hypothesize that a personal pronoun might be more helpful for the addressee in an event narration with mixed-gender characters as it uniquely identifies a particular character. This advantage is absent with same-gender characters. Compare (1a) to (1b); personal pronouns in (1a) are ambiguous while in (1b) the hearer can easily identify who the personal pronouns *she* and *he* refer to. Since the third person pronoun encodes gender only in Dutch, we expect an effect of gender on pronoun use only for Dutch speakers but not for Turkish speakers

(1a)
Suzan and Ellen went to college together.
She was studying math while she was studying literature.

(1b)
Suzan and Robert went to college together.
She was studying math while he was studying literature.

Method

Participants

Twenty pairs of native speakers of Dutch studying in Nijmegen (14 female; mean age 21.5) and twenty pairs of native speakers of Turkish studying in Istanbul (17 female; mean age 22.2) participated in our study in return for payment or course credits. They had normal or corrected-to-normal vision and no history of language impairment.

Stimuli

Our stimuli consisted of two short silent videos. Figure 1 illustrates stills depicting different actions from each video. In one video three women (same-gender condition) were engaged in cooking activities (Perniss and Özyürek, 2015). The two women who are seen to be cutting vegetables in the stills below had a more prominent role compared to the woman who is standing, and both were involved in a similar number of actions (N=10; 11). In the other video two women and one man (mixed-gender condition) were engaged in office activities. The woman and the man seen to be sitting in the first still were more prominent compared to the woman sitting behind a computer, and again each was involved in a similar number of actions (N=15; 16). Both videos included actions mainly performed by a single character although the overall activity depicted in each video (cooking and office activities) could be seen as joint.

Procedure

Participants were invited to a quiet room in pairs and randomly assigned the role of either the speaker or the addressee (they were not made aware of the role division). The addressees were not confederates and were naïve to the stimulus materials. Participants were instructed by the experimenter in the relevant language. Speakers were instructed to watch the videos one by one and to tell the addressee what they had seen after each video.



Figure 1: Still frames from the same-gender (all female) and the mixed-gender (two females; one male) video stimuli

The stimuli were presented on a computer screen. The addressees did not see the videos. Addressees were instructed that after each narrative, they could ask clarification questions. They were also informed that they would be given two short written questions about each narrative. The purpose of this was to ensure that the speakers included enough details in their narratives and that the addressees pay attention to the narratives. Once the instructions were given, the experimenter left the room and came back with the questions for the addressee after each narrative. The order of the two videos was counter-balanced. Each session was video recorded.

Data coding

We coded and analyzed speech only from the speakers of each tested pair. We transcribed the video narratives using the standard orthography of each language with the frame-by-frame video annotation software ELAN[1]. Each narrative was divided into main clauses, defined as utterances with a single subject argument and a single predicate. The subject argument of a main clause itself could express an event or an activity as in the case of nouns modified by a relative clause (e.g., the woman who is cooking). We coded only main clauses with an animate grammatical subject. Each subject referent was coded for one of the following referring expressions: *noun phrase* (*NP*) (bare noun, noun with a simple modifier or relative clause modifier, etc.), *overt pronoun* (personal pronoun, demonstrative pronoun, indefinite pronoun, etc.) and *null pronoun*. In addition, each main clause was coded for subject-to-subject local co-reference (cf. Hickmann and Hendricks, 1999). A *Maintenance* context implies that the subject referent of the current main clause is identical to that of the immediately preceding main clause. A *Re-introduction* context implies that the subject referent of the current main clause is not identical to the subject in the immediately preceding main clause but has been previously mentioned in the discourse (cf. Gullberg, 2006). (2) contains four successive clauses extracted from a single discourse narrative in our Turkish dataset as an example of these coding categories. Subject referents are italicized and letters donate co-referentiality.

(2)

Kız$_i$ ordan kalkıyo.	NP/re-introduction
Ø$_i$ kitaplığın yanına gidiyo.	null pronoun/ maintenance
Çocuk$_j$ kalkıyo masadan.	NP/re-introduction
O$_j$ da gidiyo kitaplığa.	overt pronoun/ maintenance

Girl$_i$ is standing up.
(She)$_i$ is going to the bookshelf.
The boy$_j$ is standing up off the table.
He$_j$ is also going to the bookshelf.

Results

We analyzed a total of 1.046 Turkish main clauses (426 re-introduced and 620 maintained) and 792 Dutch main clauses (360 re-introduced and 432 maintained) for the proportion of subject referents encoded with an NP, overt pronoun and null pronoun. We did not attest a systematic reason why Turkish speakers produced more clauses than Dutch speakers. Turkish speakers seemed to provide more details regarding the events in the stimulus videos. We will first present how co-reference context affects the choice of a specific RE cross-linguistically and later whether this effect is modulated by the gender of the characters mentioned in the discourse.

Effect of context

We calculated the mean proportion of NPs, overt and null pronouns out of all animate subject referents in narratives. Arcsine transformation was performed on the means before any analyses were carried out. We report the untransformed means.

We performed a mixed-effect analysis of variance with the RE type (NP, overt pronoun, null pronoun) and co-reference context (re-introduction, maintenance) as repeated measures and language (Turkish, Dutch) as independent measure. Bonferroni corrections were applied for multiple comparisons to all analyses and Greenhouse-Geisser correction was applied where the assumption of sphericity was violated. We report the corrected degrees of freedom. The analysis showed a significant main effect of RE type $F(2) = 16.390$, $p<.001$, $\eta p^2 = .301$. It also showed a significant interaction of RE type and language $F(2) = 68.986$, $p<.001$, $\eta p^2 = .645$, RE type and co-reference context $F(2) = 247.436$, $p<.001$, $\eta p^2 = .867$ and RE type, co-reference context and language $F(2) = 31.891$, $p<.001$, $\eta p^2 = .456$.

Next, we performed a mixed-effect analysis of variance separately for each co-reference context with RE type (NP, overt pronoun, null pronoun) as repeated measure and language (Turkish, Dutch) as independent measure. The analysis for the re-introduction context showed a significant

[1] ELAN is an annotation tool developed at the Max Planck Institute for Psycholinguistics, Nijmegen at The Language Archive department (cf. Lausberg, H., & Sloetjes, H., 2009).

main effect of RE type $F(1.240) = 184.903$, $p<.001$, $\eta p^2 = .830$ and a significant interaction of RE type and language $F(1.240) = 7.391$, $p=.006$, $\eta p^2 = .163$. Separate repeated measures of analysis of variance for each language showed a significant main effect of RE type for both Turkish $F(1.278) = 109.752$, $p<.001$, $\eta p^2 = .852$ and Dutch $F(1.013) = 87.661$, $p<.001$, $\eta p^2 = .996$. Both groups of speakers re-introduced subject referents mostly with NPs (*M*=.737, *SE*=.026 for Turkish; *M*=.767, *SE*=.036 for Dutch). Turkish speakers used more null pronouns (*M*=.185, *SE*=.024) than overt pronouns (*M*=.078, *SE*=.015). Dutch speakers showed the reverse pattern. They used overt pronouns more (*M*=.230, *SE*=.036) in comparison to null pronouns (*M*=.003, *SE*=.002). Compared to Dutch speakers, Turkish speakers used overt pronouns less $t(38) = -3.875$, $p<.001$ and null pronouns more $t(38) = 7.452$, $p<.001$.

The analysis for the maintenance context showed a significant main effect of RE type $F(1.194) = 52.922$, $p<.001$, $\eta p^2 = .940$ and a significant interaction of RE type and language $F(1.194) = 95.725$, $p<.001$, $\eta p^2 = .736$. Separate repeated measures of analysis of variance for each language showed an effect of RE type both for Turkish $F(2) = 352.838$, $p<.001$, $\eta p^2 = .960$ and Dutch $F(1.109) = 39.875$, $p<.001$, $\eta p^2 = .677$. Turkish speakers maintained subject referents mostly with null pronouns (*M*=.747; *SE*=.016), $p<.001$ and they used NPs and overt pronouns equally often (*M*=.126, *SE*=.013; *M*=.128, *SE*=.014). Dutch speakers on the other hand maintained subject referents mostly with overt pronouns (*M*=.699; *SE*=.049) $p<.001$, then null pronouns (*M*=.236, *SE*=.046) and the least with NPs (*M*=.065, *SE*=.014), $p<.007$). The frequency of null subjects in maintenance contexts in Dutch is similar to the numbers reported in previous literature on Dutch narratives; see (3) for an example where the speaker introduces a referent into discourse and leaves the subject referent unexpressed in the consecutive clauses. Independent sample t-tests showed that compared to Dutch speakers, Turkish speakers used fewer overt pronouns $t(38) = -9.135$, $p<.001$ but more null pronouns $t(38) = 10.788$, $p<.001$ and NPs $t(38) = 3.158$, $p=.003$.

(3)

Die andere vrouw$_i$ staat op.	NP/re-introduction
Ø$_i$ zet bureaustoel ook weer weg.	null pronoun/maintenance
Ø$_i$ loopt naar 'n boekenkast.	null pronoun/maintenance
Ø$_i$ gaat daar staan.	null pronoun/maintenance

That other woman$_i$ stands up.
(She)$_i$ put the office chair back.
(She)$_i$ walks to the bookshelf.
(She)$_i$ stands there.

As these analyses show (Figure 2), the overt pronoun is not the preferred option for any of the contexts in Turkish, unlike in Dutch. Further paired sample t-test showed that Turkish speakers used more pronouns in maintenance contexts $t(19) = 2.869$, $p<.010$ than in re-introduction contexts. Additional analysis showed that when overt pronouns were used in Turkish, 82% percent of the occurrences were accompanied by the clitic *-de/-da* (meaning *also*) which modifies the noun preceding it (see 4a). This clitic has been suggested to be a marker of focus and emphasis (Bican, 2000). In (4a) the speaker emphasizes that the woman is helping the man, she is also participating in the action the man has been performing. In Dutch on the other hand overt pronouns were used as the default form to maintain referents and were not accompanied by an extra emphasis marker (see 4b).

(4a)

Ondan sonra *Ø*$_i$ oturuyor yanına.	maintenance/null pronoun
O$_i$ da yardım ediyor.	maintenance/overt pronoun

Then *(she)*$_i$ is sitting next to (him).
She$_i$ is also helping.

(4b)

En *Ø*$_k$ gaat naast die jongen zitten.	maintenance/null pronoun
En *ze*$_k$ helpt mee.	maintenance/overt pronoun

And *(she)*$_k$ is sitting next to that boy.
En she$_k$ is helping along.

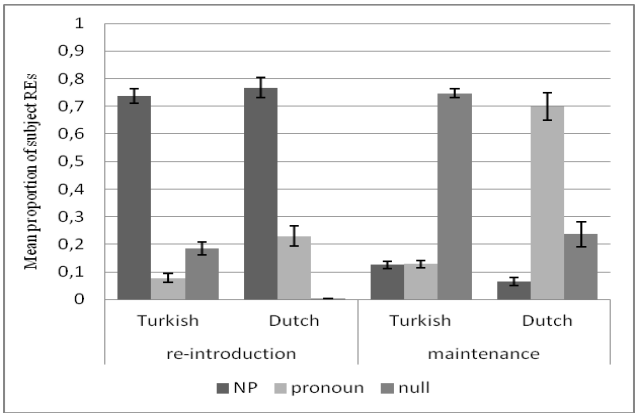


Figure 2: The distribution of RE types across the two co-reference contexts in Turkish and Dutch. Error bars represent standard errors of the mean.

Effect of gender

Since we found an interaction of RE type and co-reference context, we analyzed the two contexts separately in looking for effects of gender in the videos. We conducted a mixed effect analysis of variance separately for each co-reference context, re-introduction and maintenance, with RE type (NP, overt pronoun, null pronoun) and video type (same-gender, mixed-gender) as repeated measures and language (Turkish, Dutch) as independent measure. The Greenhouse-Geisser correction was applied in all analyses. We report the corrected degrees of freedom.

The analysis for the re-introduction context did not show a significant main effect of video type or an interaction of RE

type and video type. It only showed a significant main effect of RE type $F(1.210)= 186.847$, $p<.001$, $\eta p^2 = .831$ and a significant interaction of RE type and language $F(1.210)= 7.234$, $p=.007$, $\eta p^2 = .160$ Therefore, contrary to our prediction, neither Turkish nor Dutch speakers re-introduced subject referents differently in the narratives of the two types of videos.

The analysis for the maintenance context on the other hand did show a significant main effect of RE type $F(1.227)=47.899$, $p<.001$, $\eta p^2 = .558$, and language $F(1)=5.563$, $p=.024$, $\eta p^2 = .128$. It also showed a significant interaction of RE type and language $F(2)=91.540$, $p<.001$, $\eta p^2 = .707$, video type and language $F(1)=8.163$, $p=.007$, $\eta p^2 = .177$ and RE type and video type $F(1.652)=11.938$, $p<.001$, $\eta p^2 = .239$. Separate repeated measures of analysis of variance for each language showed an interaction of RE type and video type only for Dutch $F(1.208)=9.325$, $p= .004$, $\eta p^2 = .329$. Therefore further analyses were conducted only in Dutch. Due to the interaction of RE type and video type, we performed repeated measures of analysis of variance for each video type separately. The analysis showed a significant main effect of RE type for both the same-gender F (1.164) =61.429, $p<.001$, $\eta p^2 = .723$ and the mixed-gender videos F (1.196) =24.235, $p<.001$, $\eta p^2 = .525$. Pairwise comparisons of RE type within each video showed that in the same-gender narratives, Dutch speakers used overt pronouns ($M=.780$, $SE=.050$) significantly more often than NPs ($M=.047$, $SE=.016$) and null pronouns ($M=.172$, $SE=.046$), $p<.001$, but did not differentiate between the last two $p=.057$. While narrating the mixed-gender videos, Dutch speakers again used the overt pronoun as the most preferred RE type ($M=.63$, $SE=.056$). However, this time they used null pronouns ($M=.280$, $SE=.054$) more frequently than NPs ($M=.09$, $SE=.020$). Further paired sample t-tests showed that Dutch speakers used more pronouns while maintaining referents during the narratives of the same-gender video compared to the mixed-gender video $t(19)=3.163$, $p=.005$. Figure 3 depicts the preferred expressions for each type of videos in maintenance context for both languages.

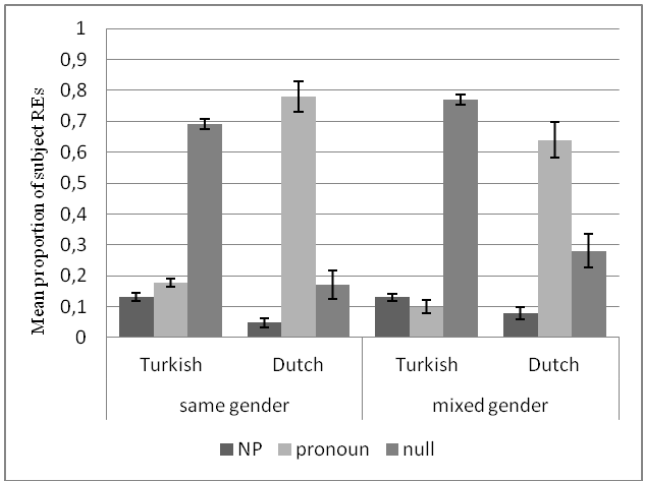


Figure 3: The distribution of RE types in maintenance contexts across the narratives of the two types of videos in Turkish and Dutch. Error bars represent standard errors of the mean.

Conclusion and Discussion

We have shown that speakers of typologically different languages in general use fuller forms while re-introducing a previously mentioned referent into the discourse and reduced forms while maintaining the same referent. This provides additional support to the previous theoretical and empirical work on reference tracking that has identified this as a language independent strategy in discourse production.

We have also provided additional cross-linguistic evidence by directly comparing adult speakers of two typologically different languages. This study also investigated which reduced form (overt or null pronoun) is preferred in each co-reference context and for what functions. Pronouns are used differently in creating coherent discourse in different languages.

Our findings from adult data support previous claims that Turkish speakers use overt pronouns mainly to mark emphasis in extended discourse and mostly in maintenance contexts. They use null pronouns as the default form in maintenance contexts. Dutch speakers on the other hand do not necessarily use pronouns to mark emphasis but rather as a default form in maintenance contexts. This current analysis directly comparing two languages then specifies the differential functions that pronouns might have in typologically different languages.

A novel contribution to the existing literature on spoken discourse production and reference tracking is our focus on gender. By manipulating the gender of the referents to be mentioned throughout the extended discourse, we were able to show that whether the genders of the people mentioned in the story are the same or not influences the reference tracking strategies of Dutch speakers but not of Turkish speakers. Pronoun use in Turkish is limited to pragmatic purposes and additionally pronouns do not encode gender. Thus using pronouns when the referents are of different gender is not an additional discourse strategy that Turkish speakers can use to create unambiguous references. Differences in gender encoding, however, manifested in different ways than we originally predicted.

We had predicted that speakers of Dutch, which encodes gender in third person singular pronouns, will use more overt pronouns while re-introducing referents during the mixed-gender narratives. Although we found an effect of gender manipulation for Dutch, we found it in the maintenance context rather than in the re-introduction context. We tentatively argue that it could be cognitively more challenging for Dutch speakers to use pronouns with mixed genders in the maintenance contexts since in order to use the correct pronoun they need to keep track of the gender of the character they just mentioned, on top of the actions they are engaged in. Tracking both actions and the

gender of the referents may be cognitively more demanding and therefore the speakers might use fewer pronouns the mixed-gender narratives compared to the same-gender narratives. While narrating the same-gender video, on the other hand, speakers do not have to keep track of the gender of their referents in maintenance contexts since all the characters in that video are female. Speakers do not necessarily need to activate the gender of the referents; using 'zij/ze' (*she*) is always grammatically correct and therefore a "safe" strategy. The reason why we do not find a similar effect in re-introduction context may be because speakers re-introduce referents mainly with an NP in narratives of both videos, almost all the time and thus the use of pronoun does not seem to be a common strategy in this context anyway.

The comparisons we provide in this study open new ways for the "pragmatic relativity", namely revealing differential attention (at least as measured by the use/non-use of certain forms) to those aspects of events and contrasts across speakers of typologically different languages to create pragmatically appropriate and coherent narratives. Further research needs to systematically study the interplay between different factors and discourse contexts, drawing extended discourse data from typologically different languages to understand the whole mechanism underpinning how speakers of different languages track referents and their significance for non-linguistic cognition.

Acknowledgements

This research is funded by the Center for Language Studies, Radboud University Nijmegen, the Netherlands and partially by the European Council Erasmus Staff training grant awarded to the first author by the International Office Radboud University. We thank Dr. Pamela Perniss for the same-gender stimulus video. We also thank Dr. Ayşe Caner and Dr. Nihan Ketrez for providing the location for the data collection in Istanbul, Turkey and their help for recruiting participants.

References

Aksu-Koç, A. & Nicolopoulou, A. (2015) Character Reference in Young Children's Narratives: Cross-linguistic Comparison of English, Greek, and Turkish. *Lingua* 155, 62-84.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Azar, Z. & Özyürek, A. (2015). Discourse management: Reference tracking in speech and gesture in Turkish narratives. *Dutch Journal of Applied Linguistics*, 4(2), 222-240.

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bican, F. Y. (2000). A computational analysis of information structure in Turkish. In A. Sumru Özsoy, Didar Akar, Mine Nakipoğlu-Demiralp, E. Eser Erguvanlı-Taylan and Ayhan Aksu-Koç (Eds.) *Studies in Turkish Linguistics*, (pp. 151-163).

Carminati, M. N. (2002). *The processing of Italian subject pronouns*. PhD Thesis, University of Massachusetts Amherst.

Chafe,W. L. (1976). Giveness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.) *Subject and topic* (pp. 25–55). New York: Academic Press.

Enç, Mürvet. (1986). Topic switching and pronominal subjects in Turkish. In D.I. Slobin and K. Zimmer (Eds.) *Studies in Turkish linguistics*, (pp. 195-209). Amsterdam: John Benjamins.

Erguvanlı-Taylan, E. (1986). Pronominal versus Zero Representation of Anaphora in Turkish. In Slobin, Dan I. and Karl Zimmer (Eds.), *Studies in Turkish linguistics*, (pp. 206-233), Amsterdam: John Benjamins.

Flecken, M. (2011). Assessing bilingual attainment: macrostructural planning in narratives. *International Journal of Bilingualism*, 15(2), 164-186.

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language Learning*, 56(1), 155-96.

Hickmann, M. & Hendriks, H. (1999). Cohesion and form in children's narratives: a comparison of English, French, German, and Mandarin Chinese. *Journal of Child Language*, 26, 419-452.

Kerslake, C. (1987). In E. H. Boeschoten and L.Th. Verhoeven (Eds.), S*tudies on modern Turkish: proceedings of the third conference on Turkish linguistics* (pp. 91-104). Tilburg: Tilburg University Press.

Küntay, A. (2002). Development of the Expression of Indefiniteness: Presenting New referents in Turkish Picture-series Stories. *Discourse Processes*, 33, 77-101.

Lambrecht, K. (1994). *Information structure and sentence form*. Cambridge: Cambridge University Press.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, 41(3), 841-849.

Özsoy, A. S. 1987. Null Subject Parameter and Turkish. In Hendrik E. Boeschoten and Ludo T. Verhoeven (Eds.) *Studies on modern Turkish: Proceedings of the third conference on Turkish linguistics*, Tilburg University Press, pp. 82-90.

Perniss, P. M., & Özyürek, A. (2015). Visible cohesion: A comparison of reference tracking in sign, speech, and co-speech gesture. *Topics in Cognitive Science*, 7(1), 36-60.

Turan, D. (1995). *Null vs. overt subjects in Turkish discourse: A centering analysis*. PhD Dissertation, University of Pennsylvania.

Yoshioka, K. (2008). Gesture and information structure in first and second language. *Gesture*, 8(2), 236-55.

Investigating Rational Analogy in the Spirit of John Stuart Mill: Bayesian Analysis of Confidence about Inferences across Aligned Simple Systems

Brad Rogers (bwrogers@indiana.edu)
David Landy (dlandy@indiana.edu)
Department of Psychological and Brain Sciences, 1101 East 10th Street
Bloomington, IN 47405 USA

Abstract

What does it mean for analogy to be rational? John Stuart Mill described a probabilistic underpinning for analogical inference based on the the odds of observing systemic pairwise correspondence across otherwise independent systems by mere chance. Although proponents and critics have debated its validity, Mill's approach has yet to be implemented computationally or studied psychologically. In this paper we examine Mill's approach and show how it can be instantiated using Bayes theorem. Then we describe two experiments that present subjects with partially-revealed, aligned binary strings with varying degrees of intra- and inter-string regularity. Experimental results are compared to a formal rational analysis of the stimuli revealing conditions whereby participants exhibit confidence patterns consistent and inconsistent with Mill's rational basis of analogy.

Keywords: Analogy, Bayes, Confidence, J.S. Mill, Rationality

Introduction

Analogy has been suggested to play central roles in human reasoning activities such as learning facts and concepts (Rumelhart & Norman, 1981), discovering scientific principles (Hesse, 1966), and perceptually interpreting our daily situations (Chalmers, French, & Hofstadter, 1991; Goldstone, Landy, & Brunel, 2011). Previous research has focused on identifying mechanisms of analogy making by comparing people's analogical preferences to predictions made by computational models (Falkenhainer, Forbus, & Gentner, 1989; Goldstone, 1994; Hofstadter, 1995; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Lu, Chen, & Holyoak, 2012). These models propose solutions to the problem of finding the best mapping between well described systems based on respective theories of analogy. Inferences have been studied in this paradigm as extensions of possible mappings, but the rationality of such extensions has not been investigated. Although these computational achievements have advanced our understanding of analogy construction, treating analogies as operations over well-defined situations has left an important context under-explored—making predictions and choices in novel and uncertain circumstances where the relevant facts may not be known in advance (Falkenhainer, 1990).

Mill's Rational Basis of Analogy

The idea that analogy has a rational underpinning is not a recent development. In *A System of Logic, Ratiocinative and Inductive*, Mill argued that the probabilistic basis of rational induction does not distinguish relational similarity from property similarity (Mill, 1882). Still he does discuss the nature of their difference:

> *"In the strictest induction, equally with the faintest analogy, we conclude because A resembles B in one or more properties, that it does so in a certain other property. The difference is, that in the case of a complete induction it has been previously shown, by due comparison of instances, that there is an invariable conjunction between the former property or properties and the latter property; but in what is called analogical reasoning, no such conjunction has been made out. There have been no opportunities of putting in practice the Method of Difference, or even the Method of Agreement; but we conclude (and that is all which the argument of analogy amounts to) that a fact m, known to be true of A, is more likely to be true of B if B agrees with A in some of its properties (even though no connection is known to exist between m and those properties), than if no resemblance at all could be traced between B and any other thing known to possess the attribute m."*

Mill describes a key distinction between "complete" property induction and analogical property induction. In the former, a statistic is developed for that specific property by observing it in a sample of many systems thereby establishing a probabilistic connection between instances of that property across those systems generally. In the latter, a statistic is developed according to the pairwise correspondence of a sample of unique properties between two systems thereby establishing a general probabilistic connection between their properties within the two systems or kinds of systems. Mill describes the pairwise correspondence as a proportion:

> *"where the resemblance is very great, the ascertained difference very small, and our knowledge of the subject-matter tolerably extensive, the argument from analogy may approach in strength very near to a valid induction. If, after much observation of B, we find that it agrees with A in nine out of ten of its known properties, we may conclude with a probability of nine to one, that it will possess any given derivative property of A."*

Although Mill's extrapolation of the proportion statistic may be mathematically simplistic, his case for a sampling postulate is clear. It would be terribly unlucky to observe a systemic correspondence in a large sample of pairwise properties and relations across two systems if in fact the true correspondences occur only at a rate of perfect chance. To the extent that we have observed greater correspondence, we would be more surprised to discover that an inferred alignment is untrue.

Mill's approach has drawn some criticism. For example, missing in his argument is the distinction between property and relation in as much as they influence the *independence* of the sample of pairwise correspondences. Along these lines, it seems likely that analogies that occur to us may follow paths toward increasingly higher match rates, a decidedly biased sampling scheme that could lead to systematic over-confidence in analogical inferences (Bartha, 2010). But an argument about biased sampling does not preclude the possibility of an unbiased rational sample, rather it necessitates the existence of such a normative ideal. Another criticism by Bartha of the sampling postulate of analogy is directed not toward its rational basis, but rather the difficulty and bias in representing real systems such that the sampling postulate might be applied. Although analogy in situ undoubtedly relies upon such representations, proponents of this criticism should then be satisfied that a rational solution might exist for any given representation and that this criticism is not about the fundamental analogical form.

Minimal Probabilistic Model of Systemic Pairwise Correspondence

The essence of Mill's proposal is that a proportion can be developed that describes the rate of pairwise matches between two systems, and that this proportion conveys predictive value to inferences made in virtue of that sampled correspondence. Harrod advanced Mill's simplistic calculation by applying the distributional property of systems generated randomly according to a known proportion—if p is the proportion of observable properties of S that are actually shared with T, and n of these properties are observed, then the probability that exactly k of them will be shared with T is given by the binomial distribution (Harrod, 1956):

$$Pr(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Harrod's formulation is a step forward in applying a probabilistic range of a known proportion to an unobserved sample in a system. But it does not fully capture the description of the analogy problem, that of estimating an unknown proportion from an observed sample. This is a problem of inverse probability whereby we seek to understand the probability of different possibly true pairwise match proportions given observation. Since we know the distribution of samples from a hypothesized model, we can apply Bayes theorem to calculate the distribution of hypothetical models (Bayes & Price, 1763; Laplace, 1825):

$$P(H|E) = P(E|H)P(H)/P(E)$$

Applying the binomial distribution for the probability of the sample given the proportion, it has been shown that the solution is beta distributed:

$$p(k) \sim \frac{x^k(1-x)^{n-k}}{B(k+1, n-k+1)}$$

The application of Bayes has brought us to an acceptable model of pairwise correspondence. However, actual reasoning about uncertain situations also involves reasoning about the resulting system. Mill has proposed a basis for reasoning across systems, but any viable approach must also account for this intra-system reasoning. In the analogy literature a classic example of this intra-system reasoning is that an intelligent person who is swayed by the analogy between the solar system and the atom does not then suspect that the nucleus is yellow, large, or causes plants to grow on Earth. Instead, the analogy is integrated with what one knows about each system on its own. Approaches that deal with these concerns have typically separated these two forms of reasoning (Falkenhainer, 1990; Lee & Holyoak, 2008). For the purpose of this initial investigation, it is less important which kinds of relational systems are explored and more important that the systems have structure in some way. In this study, we will analyze systems whose elements also exhibit a proportion. This is admittedly a minimally structured system, but it is adequate to demonstrate Mill's approach.

We apply a Bayesian mixture model that simultaneously evaluates the inferential power of the inter-system relations and the intra-system pairwise proportion (Wasserman, 2000). For computational efficiency, we can implement this using a Dirichlet distribution (or an equivalent hierarchical model) where each category comprises a unique possible combination of paired system properties (i.e., 00,01,10,11):

$$p(k_{i=1}, ..., k_{i=\kappa}) \sim \prod_{i=1}^{\kappa} p_i^{k_i} \frac{1}{B(1+k_{i=1}, ..., 1+k_{i=\kappa})}$$

**Uniform priors were used for purposes of simplicity*

Approach

We investigate Mill's proposed basis of rational analogy in the context of the simplest relational systems and their pairwise correspondences using randomly generated binary strings. Although such systems lack the context and richness often found in human analogy making, they also minimize difficulties and potential sources of biases. This petri dish approach to analogy is a starting place toward evaluating probabilistic aspects of analogical reasoning in more complex situations.

In this study we present two experiments that employ different designs but using stimuli of a common format.

Subjects are presented pairs of randomly generated strings that are only partially revealed and that possess varying degrees of intra- and inter-string regularities. Subjects are given a narrative about the general meaning of the strings, but not the meaning of individual characters. They are asked to evaluate the string pairs, to make a prediction about a target missing character, and to estimate the probability that their prediction would turn out to be true.

Stimuli

Experiments 1 and 2 presented stimuli of a consistent format—pairs of partially revealed binary strings—but employed different designs regarding the generative procedure of the strings. Participants were presented a narrative that we called "Digital Matchmaker" that cast them in the role of a data analyst at a leading online matchmaking website. Their task is to evaluate pairs of incomplete binary strings representing the responses to a questionnaire of two single people who have been matched together for a date.

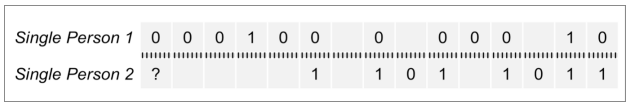


Figure 1: Example "Digital Matchmaker" Stimulus

They are told that matched couples have been made by an algorithm that can look for strong same alignment ("birds of a feather") and for strong opposite alignment ("opposites attract") to responses in the question. But matched couples may also have been made by a professional matchmaker who uses intuitions about photographs rather than survey responses to propose matches. Participants were also told about a study that asked people to fully complete their questionnaires. The study claimed to find that that regularities initially present within a single's incomplete responses and across matched singles' incomplete responses tend to mostly continue in about the same way when fully completed.

Each stimulus contains a target missing survey response that is deemed "crucial to match success." Participants must: 1) select a radio button indicating a prediction of a 0 or a 1 for the target, and 2) use a slider with numeric feedback estimating the probability that their prediction is correct.

Experiment 1

The purpose of Experiment 1 is to evaluate the relationship between people's guesses (and self-reported confidence) and the predictions of the rational analysis. The strings were generated to roughly evenly sample the range of rational predictive strength, from very weak evidence to very strong evidence.

Design

A large pool of stimuli were created using a script in R that creates and combines pseudo-random strings to meet the constraints of target sample sizes, base rates for each string, and match rates across strings. The script then randomly selects the precise positions of elements and matches, and assigns '0' and '1' symbols for a particular stimulus. The large pool of strings was then sampled, and the stimuli were selected whose rational evidence strength was closest to the target. Targets were selected at each centile on the range from 0.50 to 0.92. In this manner, a continuum of stimuli was randomly generated. The set of items was then divided into two lists, by putting odd items in one list and even items in the other.

Participants

We recruited N = 45 participants from Amazon's Mechanical Turk. Participants were paid $2.50 for an average of under 20 minutes of work. Participants were randomly assigned to one of two groups which determined which stimuli they judged.

Procedure

Participants responded to 22 stimuli with varying strengths of regularities. Each participant was randomly assigned one of two sets of stimuli (the even list or the odd list) with roughly equivalent rational probabilities. Participants responded to each randomly ordered stimulus from the list by predicting the target and estimating the probability that their prediction was correct.

Results & Discussion

Figure 2 shows the relationship between the proportion of unique participant predictions and the rational model. Overall, the correlation between choice proportions and the rational approach was fairly strong, but systematically non-linear: as the rational probability for the stimuli increases from 0.50 the mean choice probability quickly increases. After that, median confidence increases slowly or not at all.

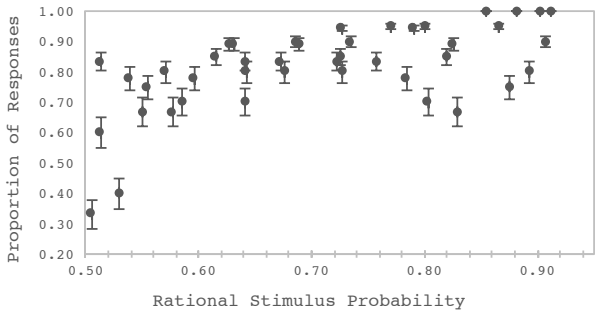


Figure 2: Proportion of Rationally Consistent Responses

**All error bars represent +/- one standard error*

People did systematically deviate from the rational model predictions for some particular stimuli. Upon examination, it appears that the regularities in those stimuli may be more spatially disperse with more interruptions dispersed rather than clustered. These effects may be investigated in future studies.

Figure 3 presents the probability estimates in light grey made by each participant for each stimulus. The median responses are shown by darkened black circles. The dashed line represents the rational model predictions.

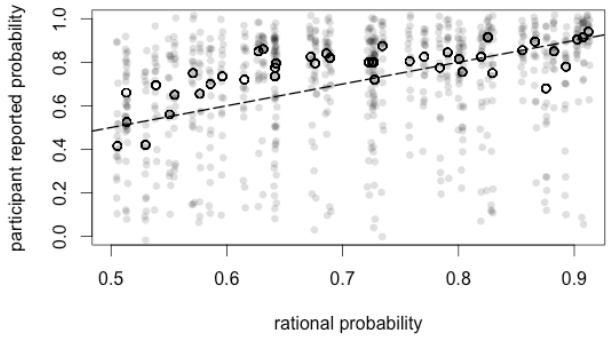


Figure 3: Response Probability vs. Rational Probability

Participants confidence increased with the rational model's probability estimate. The median participant responses by stimulus correlate to the rational probabilities by a factor of 0.71 (S.E.=0.03). The correlation is strongest at the extremes of 0.50 to 0.55 and 0.75 to 0.90, while participants' probability estimates systemically exceed the rational results on the interval of 0.55 to 0.75. This suggests a non-linear relationship with the rational model, where people may behave with overconfidence on this range.

Experiment 2

The purpose of Experiment 2 is to understand how people's predictions are influenced by intra-string regularities versus inter-string regularities.

Design

Experiment 2 used a 2x2x2 factorial design resulting in 8 groups of stimuli that are rationally distinct across groups and rationally similar within groups. The three factors examined were: 1) observed proportion of intra-string regularities (low/high), 2) observed proportion of inter-string regularities (low/high), and 3) agreement or conflict between the rational predictions corresponding to the observed intra- and inter-string regularities. The first two factors vary the degree of evidence coming from within the string, and the degree evidence coming from the analogy or relational correspondence with the other string. The strongest evidence holds when those two sources of evidence point in the same direction (*agree* stimuli); when they conflict rational evidential strength is reduced.

Participants

We recruited N = 52 participants from Amazon's Mechanical Turk. Participants were paid $3.00 for an average of under 30 minutes of work. Participants were randomly assigned to one of six groups, which determined which stimuli they judged.

Procedure

Participants responded to three stimuli from each group for a total of twenty-four responses per participant. Six sets of twenty-four stimuli were created and randomly ordered to minimize the possibility that anomalies in the stimuli or anomalies in the order of presentation would bias results. Participants were randomly assigned one of these six sets. Participants responded to each stimulus by predicting the target and estimating the probability that their prediction is correct.

Results & Discussion

We present results in for Experiment 2 that distinguish between stimuli in which intra- and inter-string regularities lead to predictions that "Agree" versus "Conflict." Agree stimuli elicited responses that are consistent with the presented regularities indicating that participants were mostly able to find at least some basis for their predictions, even when regularities occurred at a low rate. For agree stimuli, participant responses are designated as either rationally consistent or rationally inconsistent since both intra- and inter-string regularities compel the same prediction. Figure 4 shows the proportion of predictions for Agree stimuli that are rationally consistent for each other design factor.

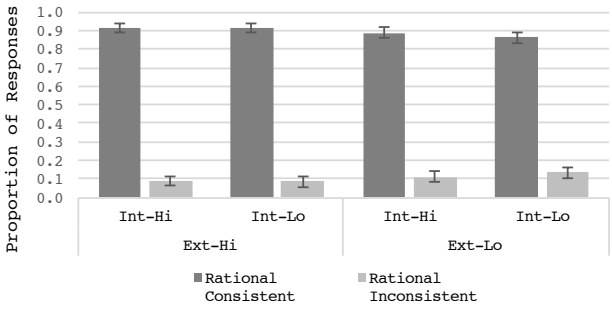


Figure 4: Proportion of Predictions in Agree Stimuli

Participant predictions were generally consistent with the rational predictions with a rate of 0.90 for the highest strength regularities and a rate of 0.85 for the lowest.

Rational probability estimates for the Agree stimuli are shown in Figure 5 below in the plots on the left and right respectively. Participants' self-reported probabilities were systematically higher than the rational model to an extent roughly consistent with the results from Experiment 1 over this rational range. More importantly, though, the main effects between intra-string and inter-string regularities are

statistically significant and follow a relative pattern consistent with the rational model predictions.

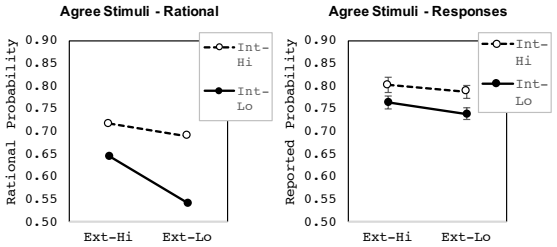


Figure 5: Rational & Response Probabilities in Agree Stimuli

Figure 6 shows the proportion of predictions for Conflict stimuli that are either consistent with the intra-string regularity (internal consistent) or the inter-string regularity (external consistent) for each other design factor.

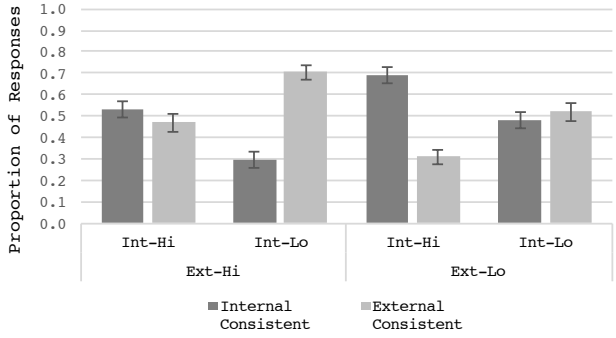


Figure 6: Proportion of Predictions in Conflict Stimuli

Participant predictions are much less consistent with the rational choices when the intra- and inter-string regularities lead to conflicting predictions. This is consistent with the idea that the relative strengths of intra- and inter-string regularities influence the chance that a participant would make a choice consistent with the rational choice. Participant predictions exhibited quantitatively symmetric preference patterns across these two conditions indicating that there does not appear to be a preference for intra- versus inter-string regularities on whole.

Participant probability estimates for Conflict stimuli demonstrate a distinct break from the rational results. Figure 7 shows the rational and participant probability estimates for the Conflict stimuli. Conflict stimuli results are further distinguished by whether the participant predictions were internal consistent or external consistent, respectively.

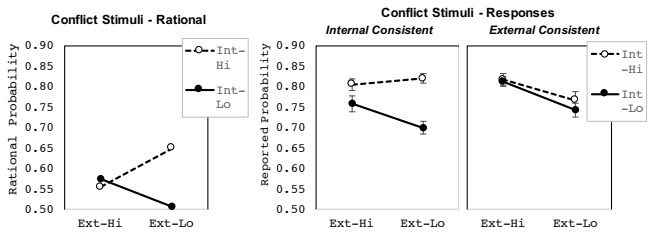


Figure 7: Rational & Response Probabilities in Conflict Stimuli

The rational analysis produces a substantial drop in probability for the hi-hi case, whereby the competing regularities are largely offsetting. Although this effect is borne out in participant predictions, participant confidence does not drop. Rather people indicate probabilities consistent with only integrating the evidence for their chosen prediction, neglecting the evidence for the other prediction. Although some participants were consistent across the three hi-hi stimuli choosing only inter-string or intra-string predictions, just as many people switched the basis of their prediction leading us to conclude that a possible individual preference for either intra- or inter-string regularities could not explain the effect.

General Discussion

This paper describes an initial investigation into the rationality of analogy that relies on a probabilistic basis. First described by Mill in 1882, this approach holds that the odds of sampling pairwise correspondence in otherwise independent systems is greater when a greater degree of pairwise correspondence between those systems truly exists. We developed a plausible minimal Bayesian model that instantiates Mill's proposal and applied this model to a rational analysis of pairs of aligned simple strings.

We compared the results of the rational analysis to the results of two psychological experiments. The first showed that people integrate observed patterns probabilistically and in a manner that correlates strongly with results of the rational analysis. People appear to over-estimate probability as the rational results increase from 0.50, but their estimates converge with the rational model around 0.75 and above.

The second experiment found that when intra- and inter-string regularities lead to the same prediction, people integrate varying combinations of strengths of these regularities in strong correspondence to the patterns of the rational analysis. But when the regularities lead to conflicting predictions, people fail to integrate across the regularities in a manner consistent with the rational analysis, instead only estimating probability according to the regularity that is consistent with their chosen prediction. Moreover, there was no statistically discernable pattern in the choice of the prediction including differences in individual preferences.

Implications for Mill's Theory

Our goal was to construct the simplest possible situation that would allow us to explore the relationship between Mill's theory and human behavior. What we found was that Mill's theory provides a meaningful guide, but discrepancies also point to complexities of the human reasoning process. The positive results for these simple tasks suggest that Mill's theory may provide a useful apparatus to understand peoples' reasoning. Still, the stimuli may be so simple that many different strategies would yield Mill-Like response patterns.

The results of Experiment 1 and the results of Experiment 2 Agree stimuli show that participants' behavior is correlated with predictions based on Mill's theory of rational analogy. These results lend credibility to using a Millian model to gain insight to how and why people use analogies.

The Conflict stimuli in Experiment 2, however, identified an inconsistency between the rational model and observed behavior. If people behave rationally, then equal evidence for competing predictions made on the basis of opposing regularities should effectively "cancel out," lowering the overall probability of a correct response. Instead we observe that people appear to dismiss the evidence for one of the competing theories and make their predictions and confidence statements based largely on the other. However, this behavior may be a more general phenomenon about how people make inferences that is not unique to analogy. First, we did not find any significant bias in predictions consistent with intra- or inter-string predictions in conflict stimuli suggesting that people are able to consider pairwise regularities across strings in the same way they consider regularities within strings. Second, previous research has shown a similar non-rational behavior in a categorization inference task where participants expressed confidence in their categorizations consistent only with the evidence for their chosen categories, neglecting evidence for fits in other categories (Murphy & Ross, 1994).

Two criticisms have been made about Mill's theory: the "counting problem" and the "problem of bias"(Bartha, 2013). The counting problem is about the seemingly arbitrary choices in defining the representation or schema over which analogies can be made. The problem of bias is one about the search and discovery of analogical mappings. Neither criticism is about analogy fundamentally, but rather about other process (i.e., representation and search) that are necessarily implicated in analogical cognitive process. These processes are independent from the basis of Mill's theory. Further, we do not promote Mill's theory as description of the analogical cognitive process. Rather, a rational basis could help answer why analogical processes operate as they do. Cognitive mechanisms need not have a rational basis, and even when they do, we should not mistake that basis as necessary but rather well adapted for the specified task.

Future Directions

We consider Mill's theory akin to the modern computational theories of analogy, but oriented toward the problem of induction with incomplete information rather than analogical discovery in well described domains. It is not hard to envision, for example, how pairwise probabilistic analysis in a more complex structure might produce qualitatively similar patterns to SMT analysis (Gentner, 1983). Binary strings are a small first step toward integrating probability into pairwise systems, but they minimize the core characteristic of analogy—relational similarity. The next step is to instantiate Mill's theory within more general relational configurations and compare results to people and other models.

Acknowledgments

This material is based upon support of BR by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1342962.

References

Bartha, P. (2010). *By Parallel Reasoning*. New York: Oxford Uniersity Press.

Bartha, P. (2013). Analogy and Analogical Reasoning. In *The Stanford Encyclopedia of Philosophy*.

Bayes, T., & Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1991). High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *Journal of Experimental and Theoretical Artificial Intelligence*, (March).

Falkenhainer, B. (1990). A unified approach to explanation and theory formation. In J. Shrager & P. Langley (Eds.), *Computational Models of Scientific Discovery and Theory Formation* (pp. 157–196). Morgan Kaufmann Publishers, Inc.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). Structre Mapping Engine. *Artificial Intelligence*, *41*, 1–63.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science*, *7*(2), 155–170.

Goldstone, R. L. (1994). Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 3–28.

Goldstone, R. L., Landy, D., & Brunel, L. C. (2011). Improving perception to make distant connections closer. *Frontiers in Psychology*, *2*(December), 385.

Harrod, R. (1956). *Foundations of Inductive Logic*. London: Macmillan.

Hesse, M. (1966). *Models and Analogies in Science*. Notre Dame: Notre Dame Press.

Hofstadter, D. R. (1995). *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science: A Multidisciplinary Journal*, *13*, 295–355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Laplace, P.-S. (1825). *Philosophical Essay on Probabilities:* Springer Science & Business Media.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1111–1122.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*(October 2015), 617–648.

Mill, J. S. (1882). *Chapter XX. Of Analogy*. A System of Logic: Ratiocinative and Inductive. *1882*.

Murphy, G. L., & Ross, B. H. (1994). Predictions From Uncertain Categorizations. *Cognitive Psychology*.

Rumelhart, D. E., & Norman, D. A. (1981). Analogical processes in learning. In *Cognitive Skills and Their Acquisition* (335–359).

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107.

Translating testimonial claims into evidence for category-based induction

Tamar Kushnir (tk397@cornell.edu)
Cornell University, Department of Human Development, Martha Van Rensselaer Hall
Ithaca, NY 14853 USA

Susan A. Gelman (gelman@umich.edu)
University of Michigan, Department of Psychology, 530 Church Street
Ann Arbor, MI 48109 USA

Abstract

Inductive generalizations about the properties of *kinds* are based on evidence. But evidence can come either from our observations, or from the testimony of knowledgeable informants. The current study explores how we combine information from these two sources to make inductive inferences. Participants learned about a novel object category, and observed the property occur with some frequency in a sample of category members. Different groups of participants also heard an informant making either Generic, Quantified, or Specific claims about the prevalence of the property. Participants who heard generic claims were more resistant to a straightforward use of statistical evidence in their generalizations. Moreover, participants who rated the informant as more knowledgeable (across conditions) gave higher prevalence estimates. The results suggest two pathways through which testimony translates into evidence for category learning, and raise questions on how to best combine evidence from these different sources into a common representational form.

Keywords: category-based induction; probabilistic reasoning; generics; rational models; testimony; epistemic trust

Introduction

It's nearly impossible to learn something new without in some way generalizing what you learned. This core feature of cognition – that our experiences teach us about *kinds* of things and not just individual instances – supports the inferences, predictions, and explanations that help us make sense of the world around us. Much empirical research – guided by formal models based on Bayesian inference – suggests that these generalizations are principled and rational (Gopnik & Wellman, 2012; Oaksford & Chater, 2007; Schulz, 2012; Tenenbaum & Griffiths, 2001); that is, they are both based on our current conceptual knowledge and appropriately responsive to new evidence. This process of rational (i.e. Bayesian) updating is how we are able to get so much inductive power from so little new information.

The rational models approach has empirical support from concept learning studies in infants (Denison, Trikutam & Xu, 2014; Teglas et al, 2011), children (Kushnir & Gopnik, 2007; Schulz, Bonawitz & Griffiths, 2001), and adults (Griffiths et al, 2011; Griffiths & Tenenbaum, 2009). In all of these studies, evidence presented to participants is statistical (associations that are observed to occur with some frequency) and as such can be represented as probability distributions. But new information comes to us in many forms, and, as human learners, the process of translating evidence into probabilistic representations may not always be as straightforward as these studies would imply.

Consider the following example. My best friend just visited an exotic island and brings back with her two Ylang-Ylang seeds. One seed grows into fragrant flowers, and the other grows into flowers with no smell at all. A reasonable generalization, based on my observations and no other information, might be to assume that these plants (in their natural habitat) are fragrant about half of the time.

Consider, instead, what would happen if she brought home stories from her travels instead of seeds, and told me about the Ylang-Ylang flowers she saw with an amazing fragrance. What can I conclude from these stories? Based on her general description, and the fact that I trust my friend is telling the truth, I'd probably assume that all (or most) of the plants had a lovely smell.

The example illustrates two very different ways of translating information about a new category and its properties into probabilities: One is by sampling from the category and observing the prevalence of that property in the sample (i.e., the statistical likelihood of a member of the sample having said property) then treating the sample as representative of the population. Both adults and children do this (Gweon, Tenenbaum & Schulz, 2010; Xu & Tenenbaum, 2007).[1]

Another is by believing the testimony of trusted sources. Especially when we are young, but also when we are older, we rely on others to impart generalities by making kind-based claims (Cimpian & Markman, 2009; Gelman et al, 1998; Koenig et al., 2015). Arguably, much of our conceptual knowledge is acquired in this way, rather than through direct sampling, because many of the deep, non-obvious, and essential properties of categories are not directly observable (Gelman, 2003). Thus, it remains a

[1] We leave aside for now, specifics on the way the sample was generated. In our example, it could have been a random sample (she happened to pick some flowers while on vacation), a sample chosen intentionally (she liked those flowers specifically, so she picked them) or a sample chosen pedagogically (she wanted to share information about their fragrant properties, so she picked an representative set to show me). For details on how rational models account for the sample-generating process in inductive generalization, see Shafto, Goodman & Frank, 2012; Shafto, Goodman & Griffiths, 2013.

critical question: does it make sense to straightforwardly translate new information about kinds garnered through testimony into probability estimates?

There is some evidence that generic claims ("Ylang-Ylang flowers have a lovely smell") can be used as the basis for probability estimates. Novel generic claims lead people to make high estimates of the prevalence of a property in the category, and to generalize broadly to new category members (Brandone, Gelman & Hedglen, 2015, Cimpian, Brandone & Gelman, 2010; Rhodes, Leslie & Tworek, 2010). However, there are plenty of cases in which generic claims could lead to estimates of *low* probability (as in the claim "Mosquitoes carry West Nile virus"). Morover, both children (Brandone, Cimpian, Leslie, & Gelman, 2012) and adults (Cimpian, Gelman, & Brandone, 2010; Prasada, Khemlani, Leslie & Glucksberg, 2013) view generic claims as distinct from claims about quantities. Thus, with the addition of even a little prior knowledge (such as the fact that West Nile is a rare disease) it is easy to see how a generic claim could be recast as a probability estimate.

Are more precise, quantified ("some/most/all Ylang-Ylang flowers have a lovely smell") statements about kinds easier to translate into probabilities? Perhaps. But logically, absolute claims ("all Ylang-Ylang flowers have a lovely smell") fail to be true after even *one* exception, and, a discovery of that exception may call into question the credibility of the person making the claims. This would present its own complication: how do we create probability estimates from *unreliable* testimony?

The current study represents a first attempt at addressing these less-than-straightforward cases. We use a simple design to teach participants about a novel object category and a property (a causal property, discoverable, but not immediately obvious) of a sample of category members. Participants observed the property occur with some frequency in the sample. Two groups also heard an informant making category-based claims – either Generic or Quantified (i.e. "all") – about the property. We compared these two groups with a group of participants who heard the informant make a Specific (and accurate) claim about the property of only one of the objects in the sample.

Our central aim was to investigate how participants would integrate the statistical data with category-based testimonial claims into estimates of property prevalence. Most straightforwardly, participants should take the observed frequency of the property in the sample as representative of the prevalence of the property in the category as a whole. However, based on prior work, we might expect two differences unique to generic testimonial claims. On one hand, given the flexibility of generic claims (that they allow for exceptions, and can be true for low-prevalence properties), participants who hear generic claims may not rely as much on observed frequency in their prevalence estimates. On the other hand, given that generic claims are often taken as indicating high prevalence, participants who hear generic claims may over-estimate the prevalence of the property across the range of observed frequencies.

Our second aim was to investigate how knowledge attributions are influenced by the nature of the testimonial claims made, and in turn how these attributions influence inferences about the category and its properties. In our study, observed frequencies afford participants the opportunity to *verify* testimonial claims. Thus, most straightforwardly, specific (and verifiably accurate) claims made about a single object should lead to high knowledge attributions. On the other hand, category-based claims that are not well-matched by observed frequencies (e.g. a general claim about the property, but a low observed frequency in the sample) may lead to low knowledge attributions. This might be particularly true when the observations logically contradict the claim (as in the absolute, quantified "all" claims). But, it may also be true of generic claims (as they suggest high prevalence). On the other hand, given the flexibility of generic claims to varying interpretations, knowledge attributions following such claims may remain high. These knowledge attributions may be an additional pathway through which category-based claims influence prevalence estimates; perhaps leading people to assign greater weight to testimony that is seen as coming from a knowledgeable source.

Methods

Participants

Nine-hundred-and-thirty-three adults (512 female, 412 male, age range 16yrs – 86yrs, mean age = 37.37, SD = 12.16) participated in the study through Amazon Mechanical Turk in exchange for monetary compensation. Data was collected on US participants only. Participants were majority white (79%) and non-Latino/a (92.9%). A majority had attended college (35% with a 4-year degree, 12.9% with a 2-year degree, 27.5% some college).

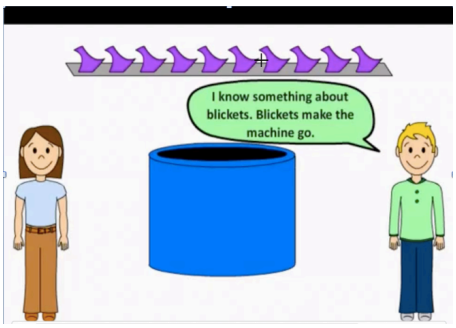


Figure 1: A sample of the video seen by participants on trial 1 in the *Generic Language* condition. In this condition, the cartoon figure on the right, a character named "Zorg," makes a claim about the objects before they enter the "special music machine."

Procedure

A random number generator assigned each participant to one of 30 conditions, for a total of 30-33 participants per condition. Each condition combined one of 3 types of testimonial claims (Specific, Generic, Quantified) and one

of 10 observed frequencies of the novel causal property (ranging from 1/10 to 10/10 inclusive).

The procedure consisted of two trials, each with the same characters, observed frequency, and testimony, but with different objects (Trial 1: "blickets" and Trial 2: "midos") and a differently-colored machine. On each trial, participants first watched a short video then answered a set of questions. A screen shot of the first video is shown in Figure 1. The video began with a girl who introduced the "special music machine," her friend Zorg, and an array of objects on a shelf that she labeled (e.g. "Here are some blickets" in Trial 1). Zorg then made one of the following testimonial claims, depending on condition:

Specific claim: *"I know something about this blicket. This blicket makes the machine go."*[2]

Generic claim: *"I know something about blickets. Blickets make the machine go."*

Quantified ("All") claim: *"I know something about all blickets. All blickets make the machine go."*

After hearing Zorg's claim, the video showed the 10 objects passing through the machine (they went into the top and came out of the bottom) one at a time. When the machine "activated" it changed color and made a loud trumpeting sound. The first object always had the property of activating machine insuring that, in the Specific condition, Zorg's claim was accurate. After the first object, a random subset of the remaining objects activated the machine depending on the observed frequency assigned.

Following the video, participants made two knowledge attribution ratings: A **Category Knowledge Attribution** (*How much do you think Zorg knows about blickets?)* and a **Machine Knowledge Attribution** (*How much do you think Zorg know about the machine?).* Answer choices for both were on a three-point scale: 0 – Nothing, 1 – A little bit, 2 - A lot. Participants also answered a **Property Prevalence Estimate**: *"Imagine there were more blickets here. What percentage of these blickets would make the machine go?* Participants were allowed to enter numbers ranging from from 0-100. Following the questions for Trial 1, participants began Trial two by watching the second video and answering the same set of questions about "midos."

Results

Property Prevalence Estimates: We examined the influence of Trial (within participants), Observed Frequency (between participants), Testimony (between participants), and the interaction of the two manipulated variables on Prevalence Estimates using a linear mixed effects model.[3] Parameter estimates for the model are shown in Table 1. There was no significant effect of Trial ($F(1,1836)=.02$, *ns*). There was a significant main effect of Observed Frequency ($F(1,1836)=2034.6$, $p<.001$) such that estimates increased as frequency of activation increased, regardless of the type of testimony heard. There was also a main effect of Testimony ($F(2,1836)=96.4$, $p<.001$). Prevalence estimates were on average highest after hearing Quantified (i.e. "All") claims (Mean=58%, SD=31%), and on average 3.3% above than the observed frequencies in each condition (SE=1.16, $t(297) = 2.84$, $p<.01$). Estimates were next highest after hearing Generic claims (Mean=54%, SD=30%) and were not different on average from observed frequencies in each condition (M=-.12, SE=1.33, $t(311) = -0.09$, *ns*). Estimates were lowest after hearing Specific claims (Mean=52%, SD=30%) and were on average 4.18% lower than the observed frequencies in each condition (SE=1.06, $t(302) = -3.94$, $p<.001$).

Finally, there was a significant interaction between Testimony and Frequency ($F(2,1836)=4.8$, $p<.01$). An illustration of this interaction can be seen in the graph of the model predicted values in Figure 2. In the Generic conditions, the slope of prevalence estimates was significantly flatter than in the other two conditions (see parameter estimate in bold in Table 1): they were on average higher for low-frequency properties than in the Specific condition and lower for high-frequency properties than in the Quantified condition.

Table 1: Parameter Estimates of the mixed model predicting Prevalence Estimates by condition

Parameter	Estimate	SE	*t*
Intercept	5.01	1.91	2.62**
Slope of Specific condition	.84	.03	27.38***
Intercept difference between Generic and Specific	10.13	2.67	3.79***
Intercept difference between Quantified and Specific	7.08	2.70	2.62**
Slope difference between Generic and Specific	**-.12**	**.04**	**-2.69****
Slope difference between Quantified and Specific	.00	.04	.02

*$p<.05$; **$p<.01$, ***$p<.001$

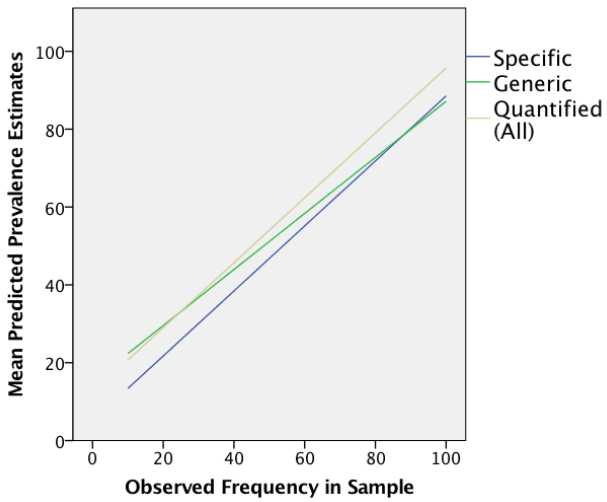


Figure 2: An illustration of the main effects of Frequency, Testimony, and interaction between the two on participants' estimates of property prevalence.

[2] In the Specific claim condition, the blicket Zorg refers to is always the one closest to him, and as he speaks he "points" to it with a dashed line connecting his hand to the object.

[3] A Q-Q plot of the prevalence estimates showed that they were approximately normally distributed.

Object Knowledge Attributions: We then examined the influence of Trial, Observed Frequency, Testimony, and their interaction on attributions of Category Knowledge (e.g. "How much does Zorg know about blickets?") using an ordinal logistic GEE. There was no significant effect of Trial (Wald χ^2 (1) = 2.5, *ns*). There was a significant main effect of Observed Frequency (Wald χ^2 (1) = 59.4, p<.001), such that attributions of knowledge increased as the frequency in the sample increased. There was also significant main effect of Testimony (Wald χ^2 (2) = 38.2, p<.001): attributions of knowledge were highest in the Specific condition (Mean = 1.34, SD= .55), next highest in the Generic condition (Mean=1.25 SD=.54) and lowest in the Quantified condition (Mean=1.09 SD=.58).

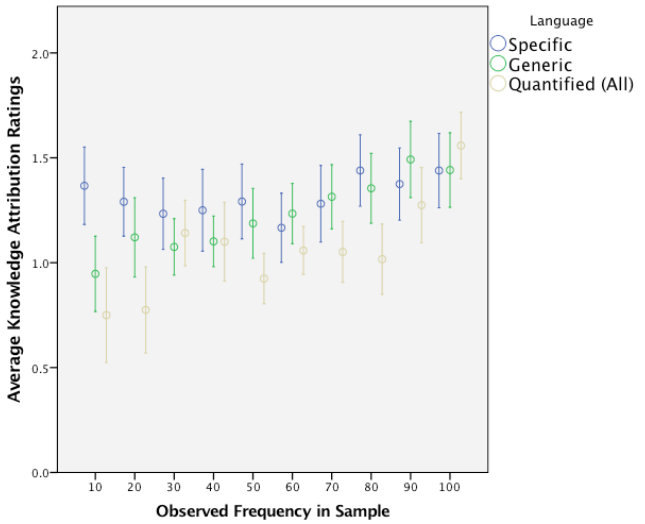


Figure 3: Average knowledge attributions (Category and Machine Knowledge Questions, Trials 1 and 2) by Observed Frequency and type of Testimony. Error bars represent 95% confidence intervals of the mean.

Critically, there was a significant interaction between Observed Frequency and Testimony (Wald χ^2 (2) = 17.2, *p*<.001). Parameter estimates reveal the nature of the interaction: while the slope in the Specific condition was not significantly different from 0 (Wald χ^2 (1) = 2.1, *ns*) the slopes in the Generic and Quantified conditions were significantly different from the Specific condition (Generic vs Specific: Wald χ^2 (1) = 15.3, *p*<.001; Quantified vs Specific: Wald χ^2 (1) = 9.6, *p*<.01) and significantly positively correlated with Observed Frequency.

Object Knowledge Attributions: We ran the same analysis on attributions of Machine Knowledge ("How much does Zorg know about the Machine?"). There was no significant effect of Trial (Wald χ^2 (1) = .2, *ns*). There was again significant main effect of Observed Frequency (Wald χ^2 (1) = 46.1, p<.001), a significant main effect of Testimony (Wald χ^2 (2) = 26.8, p<.001), and a significant interaction (Wald χ^2 (2) = 12.2, p<.01). The interaction followed the same pattern as above: the slope in the Specific condition was not significantly different from 0 (Wald χ^2 (1) = 1.8, *ns*) the slopes in the Generic and Quantified conditions were significantly different from the Specific condition (Generic vs Specific: Wald χ^2 (1) = 9.3, *p*<.01; Quantified vs Specific: Wald χ^2 (1) = 8.8, *p*<.01) and significantly positively correlated with Observed Frequency.

Together these results suggest that, when Zorg made claims about a specific object, participants' knowledge attributions did not depend on observed frequency. However, when Zorg made category-based claims about the properties of the objects, participants' knowledge attribution ratings increased as observed frequency of the property in the sample increased. This interaction is illustrated in a graph of the average knowledge attributions across both types of questions (Figure 3).

Knowledge Attribution Influences on Prevalence Estimates: Our final question concerned the influence of knowledge attributions on inductive generalizations. We addressed this question by adding knowledge attributions to our mixed effects model from Table 1 to see if either type of knowledge attribution (Category Knowledge or Machine Knowledge) predicted prevalence estimates above and beyond the main effects and interactions shown. Only the main effect of Category Knowledge attributions was significant ($F(2,1832)= 23.16$, p<.001). An illustration of the knowledge attribution effect is shown in Figure 4. Participants who rated Zorg as less knowledgeable about blickets/midos gave lower than average prevalence estimates. Participants who rated Zorg as more knowledgeable about blickets/midos gave higher than average prevalence estimates. Partial correlations, controlling for observed frequencies, show that the relationship was strongest in the Quantified (*r*=.45, *p*<.001) and Generics conditions (*r*=.40, *p*<.001), and still significantly positive but less strong in the Specific condition (*r*=.14, *p*<.05).

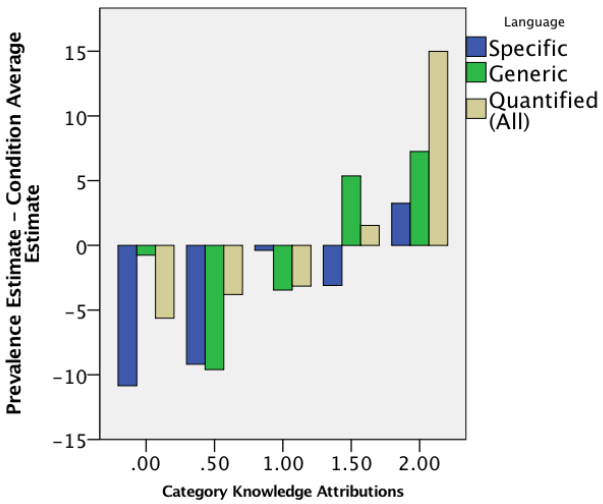


Figure 4: An illustration of the effect of Category Knowledge Attributions on prevalence estimates. Y axis shows the difference between actual estimates and the condition average (to account for condition differences found in prior analyses).

General Discussion

New information comes to us in different ways; we may make our own observations, or we can believe what others tell us. How do we integrate evidence from these different sources into a common representational form? The present study begins to address this broad question. We designed a category learning study in which participants heard a testimonial claim about a category and one its properties. They then observed the property with some frequency in a sample of category members. Participants were asked to translate this evidence into an estimate of the prevalence of the property in the category as a whole, and also asked to attribute knowledge to the informant who made the testimonial claim.

To begin with, it should be noted that participants in our study responded rationally to the statistical evidence from the sample they observed. Low observed frequencies led to low prevalence estimates, and high observed frequencies led to high prevalence estimates. However, testimony influenced this basic and well-replicated result.

Generic claims ("*blickets* make the machine go") had a unique influence on prevalence estimates compared to the other types of testimony. After hearing generic claims about the category and its novel property, participants' estimates of low prevalence properties were higher, and their estimates of high prevalence properties were lower. This explains why the estimates were, on average, well-matched to the observed frequencies. But it also suggests that, at the extremes, these claims would be less well matched. Thus, we offer one intriguing interpretation of this "flattened" relationship: that generics have the effect of making people *resistant* to straightforward use of statistical evidence in their inductive inferences about categories.

Generic claims also led to some degree of skepticism about the knowledge of the informant, in particular when the observed frequency of the property was low. This is consistent with prior work suggesting that generic claims are often interpreted as indicating high prevalence (Cimpian et al., 2010). So, observing high frequencies helps verify the credibility of these claims, and observing low frequencies calls their credibility into question.

Quantified, absolute claims ("*all blickets* make the machine go") led to prevalence estimates that were on average higher than straightforwardly predicted by the sample. Absolute claims also led to the greatest degree of skepticism about the knowledge of the informant. But this skepticism did not seem to be based on a straightforward logical relation between the sample and the claim; if it were, then any frequency of less than 100% would have overwhelmingly led to responses that Zorg knows "nothing" or "little" about the category (which it did not). Rather, the relation between knowledge attributions and frequency paralleled the results in the Generic condition, suggesting that perhaps a documented tendency to interpret quantified claims as generic (Leslie & Gelman, 2012) played a role. This possibility, and other potential influences on how we use frequency evidence to verify the reliability of logical claims, would be interesting to explore in future work.

Non category-based claims about a specific object ("*this blicket* makes the machine go") also influenced prevalence estimates; they led to estimates lower than straightforwardly predicted by the observed frequency. Knowledge attributions in this condition tended to be high, and were independent of the observed frequencies. This suggests that participants based their knowledge attributions on the accuracy of the specific claim, disregarding evidence (i.e. the properties of the remaining objects) irrelevant to verifying them.

The final result was that, across all types of testimony and all observed frequencies, knowledge attributions positively influenced prevalence estimates. Thus, all else equal, participants who were skeptical gave lower prevalence estimates relative to others in the same condition, and participants who were convinced gave higher relative prevalence estimates.

All together, the results suggest two pathways through which testimony translates into evidence for category learning. Claims about categories and their properties lead to adjustments to the way observed frequencies translate into general category knowledge. And, the credibility of claims about categories and their properties themselves can be verified through observation, and subsequent skepticism or belief in these claims leads to further adjustments to inductive generalizations.

The influence of generic testimony in our study has particularly important implications for a principled, rational account of category learning, inductive inference, and belief revision. Generics may, as stated, make us resistant to straightforward interpretations of probabilistic data. They may also, by extension, lead to a greater resistance to counter-evidence in belief revision. Generic *speakers* also have the advantage of being judged as knowledgeable despite counter-evidence. In support of this idea, other work has shown that generic claims lead to high estimates of knowledge even when they are *not* verifiable (Koenig et al., 2016). Thus, what may make generics so influential in category learning is that they are resistant to statistical evidence by both pathways.

Do these findings pose a problem for the rational view? Not necessarily. Our preliminary findings suggest that evidence from testimony doesn't completely discount our own observations, but rather can be integrated with them. The details of the integration may depend in part on the type of testimony (Generic, Specific, and Quantified among others). It may also be expected to interact with domain-specific knowledge, and additional types of evidence. Many of these details could be worked out in future empirical and computational work.

Our data suggests that perhaps not all of the interesting nuances in human learning and inductive inference are explained by differences in prior knowledge, or by assumptions made about the evidence-generating mechanism (e.g. strong vs. weak sampling, teacher vs.

learner driven, intervention vs. observation). Instead we suggest that evidence from *what people say* is not straightforwardly reducible to these distinctions. When we relax our assumption that evidence comes in only one form, we can come closer to understanding the potential of human learning.

Acknowledgments

We'd like to thank Alexandra Was and Natalie Davidson for help creating stimuli and with data collection. Thanks to three anonymous reviewers, and to Felix Thommes for helpful comments on the revised version. This research was supported by funding from NSF (DLS #1023179) to T.K. and NICHD grant HD-36043 to S.G.

References

Brandone, A. C., Cimpian, A., Leslie, S., & Gelman, S. A. (2012). Do lions have manes? For children, generics are about kinds rather than quantities. *Child Development,* 83(2), 423-433.

Brandone, A. C., Gelman, S. A., & Hedglen, J. (2015). Children's developing intuitions about the truth conditions and implications of novel generics versus quantified statements. *Cognitive Science*, *39*(4), 711-738.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science,* 34(8), 1452-1482.

Cimpian, A., Gelman, S. A., & Brandone, A. C. (2010). Theory-based considerations influence the interpretation of generic sentences. *Language and Cognitive Processes,* 25(2), 261-276.

Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition*, *113*(1), 14-25.

Denison, S., Trikutam, P., & Xu, F. (2014) Probability versus representativeness in infancy: Can infants use naïve physics to adjust population base rates in probabilistic inference? *Developmental Psychology*. 50(8):2009-19.

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., Pappas, A., & Keil, F. C. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development*, i-157.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York, NY: Oxford University Press.

Gopnik A, & Wellman H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085-1108.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*, 1407-1455.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*, 661-716.

Gweon, H., Tenenbaum, J., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences, 107*(20), 9066-9071.

Koenig, M. A., Cole, C. A., Meyer, M., Ridge, K. E., Kushnir, T., & Gelman, S. A. (2015). Reasoning about knowledge: Children's evaluations of generality and verifiability. *Cognitive Psychology, 83*, 22-39.

Leslie, S., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive Psychology*, *64*(3), 186-214.

Oaksford, M., & Chater, N. (2007) *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Prasada, S., Khemlani, S., Leslie, S., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition,* 126(3), 405-422.

Rhodes, M., Leslie, S., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences, 109*(34), 13526-13531.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*, 382-389.

Schulz, L.E., Bonawitz, E. B., & Griffiths, T. (2007). Can being scared make your tummyache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology, 43*(5), 1124-1139.

Shafto, P., Goodman, N. D., & Griffiths, T. L (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology, 71*, 55-89.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences, 104*, 19156–19159.

Xu, F., & Tenenbaum, J.B. (2007) Word learning as Bayesian inference. *Psychological Review, 114*, 245-272.

Simple Search Algorithms on Semantic Networks Learned from Language Use

Aida Nematzadeh, **Filip Miscevic**, and **Suzanne Stevenson**
Department of Computer Science
University of Toronto
{aida,miscevic,suzanne}@cs.toronto.edu

Abstract

Recent empirical and modeling research has focused on the semantic fluency task because it is informative about semantic memory. An interesting interplay arises between the richness of representations in semantic memory and the complexity of algorithms required to process it. It has remained an open question whether representations of words and their relations *learned from language use* can enable a simple search algorithm to mimic the observed behavior in the fluency task. Here we show that it is plausible to learn rich representations from naturalistic data for which a very simple search algorithm (a random walk) can replicate the human patterns. We suggest that explicitly structuring knowledge about words into a semantic network plays a crucial role in modeling human behavior in memory search and retrieval; moreover, this is the case across a range of semantic information sources.

Keywords: semantic networks; semantic search; semantic memory; computational modeling

Introduction

Semantic memory plays a significant role in cognition because it is the locus of storage for concepts and their relations. There are a number of competing hypotheses for the representation of semantic memory, such as semantic networks (*e.g.*, Collins & Loftus, 1975; Steyvers & Tenenbaum, 2005), vector space models (*e.g.*, Landauer & Dumais, 1997), and topic models (Griffiths, Steyvers, & Tenenbaum, 2007). The content and structure of semantic memory is of great interest because it impacts how effectively people can store, search for, and retrieve information.

Recent work in computational modeling has illustrated in an interesting way the trade-off between the representation of semantic memory and the nature of the algorithms required to process it (Hills, Jones, & Todd, 2012; Thompson & Kello, 2014; Abbott, Austerweil, & Griffiths, 2015). The models in question focused on the semantic fluency task, in which people name as many members of a cue category as they can in a certain amount of time. This task is informative about representation and processing of semantic memory because it requires people to access semantically-related words. Here we focus on the empirical data of Hills et al. (2012), who argue that people follow an optimal foraging pattern that is similar to animals searching for food: a semantic patch is *exploited* until the rate of word retrieval is less than the long-term average rate of retrieval, and then a new patch of related words is *explored*.

Hills et al. (2012) and Abbott et al. (2015) suggest that very different computational approaches are required to model this empirical data. Hills et al. (2012) adopted a vector space representation of semantic memory – one that encodes word–word co-occurrence patterns. Using this representation, they showed that the best match to human behavior required a two-stage algorithm with an explicit strategy to switch from exploiting the current semantic patch to exploring a new patch. In contrast, Abbott et al. (2015) showed that a simple random walk that operates uniformly was sufficient to model the pattern of behavior. To achieve this, their model used a semantic network representation that encoded relations among words from free association norms. These results clearly demonstrate the interplay of representation and algorithm in replicating the same empirical data on semantic memory.

Having a semantic memory that is appropriately structured to support efficient real-time access might constitute a good balance in the representation/process trade-off. But people must learn such a structure. Creating a semantic network by directly encoding human association norms, as Abbott et al. (2015) do, avoids the statistical learning problem that people face (Jones, Hills, & Todd, 2015). It has thus remained an open question whether representations of words and their relations *learned from language use* can enable a simple search algorithm to mimic the observed behavior in the fluency task.

Our first contribution here is to show that this is indeed possible: we create a semantic network using learned meanings of words from a cognitively plausible computational model, and show that a simple, uniform random walk exhibits the observed foraging pattern of search. Moreover, we also show that if an explicit semantic network is created from the vector-space semantic information of Hills et al. (2012), the same random walk algorithm on that network shows the desired match with human behavior. We thus conclude that explicitly structuring knowledge about words into a semantic network plays a crucial role in modeling observed behavior in memory search and retrieval; moreover, this is the case across a range of semantic information sources (not solely in the case of free association data). We also perform structural analyses of the networks to consider the relation between their connectivity properties and their behavior.

Related Semantic Fluency Data and Models

Hills et al. (2012) argue that search through semantic memory is guided by the same strategy as that used by animals foraging for food. In support of this view, they found that participant responses in a semantic fluency task (i.e., 'name as many animals as you can in 3 minutes') came in bursts of semantically related "patches" (animal categories as defined by Troyer, Moscovitch, and Winocur (1997), such as 'pets' or 'farm animals'). Moreover, the timing of these responses was consistent with the marginal value theorem of optimal foraging in physical space (Charnov, 1976). Specifically, the time

it took for participants to retrieve the next novel item relative to the last one – referred to as the inter-item retrieval time (IRT) – increased with each item within a patch. When the IRT exceeded the participant's average IRT across the entire trial, a switch into a different patch of semantically-related words occurred, and the IRT then decreased. This pattern can be seen in Figure 1a in the Results section.

Hills et al. (2012) investigated the ability of different search algorithms to model this empirical data, using semantic representations of words learned by a vector space model, BEAGLE, on the Wikipedia corpus (Jones & Mewhort, 2007). They show that a two-stage algorithm best replicates the data, using local cues (word–word similarity) to find the next item within a patch, along with an explicit strategy to switch to a global cue (word frequency) to guide exploration of a new patch. Moreover, they showed that a simpler search algorithm – a random walk that used only the word–word similarities – could not capture the observed foraging pattern.

In contrast, Abbott et al. (2015) showed that a simple random walk on a semantic network *could* replicate human IRT patterns just as well as the two-stage algorithm of Hills et al. (2012). However, their semantic representation was created using human association norms (Nelson, McEvoy, & Schreiber, 1998). Jones et al. (2015) raised the issue that this semantic representation implicitly encodes the structure of a search process similar to the fluency task, thereby making it possible for a search algorithm simpler than that used by Hills et al. to replicate the empirical data.

In the remainder of the paper, we explore whether a structured representation that results in a simpler search and retrieval algorithm can be learned from the kind of data that people are naturally exposed to. Similarly to Abbott et al. (2015) and in contrast to Hills et al. (2012), we construct a semantic network to explicitly encode the appropriate relations among words. However, unlike Abbott et al., our model learns these relations from a language corpus rather than simply encoding human association norms. Moreover, unlike Hills et al., we use a corpus of child-directed speech to reflect more naturalistic language input, and use a semantic representation that explicitly draws on conceptual knowledge.

Our Semantic Representation

We briefly review our computational word learner, then describe the process for constructing semantic networks. Since we model the empirical data from Hills et al. (2012) examining semantic fluency of animal words, we focus on the subset of words in our training data that also occur in their dataset.

The Word Learner

We use an incremental and probabilistic cross-situational learner shown to mimic a range of child and adult behaviors in vocabulary learning (Fazly, Alishahi, & Stevenson, 2010). The model takes as input a sequence of utterance–scene pairs, U–S, where U represents the linguistic input to a child, and S represents the non-linguistic data a child perceives in language learning. The input is highly ambiguous, as the mapping between individual words in U and the relevant semantics in S is not explicitly indicated – U is represented as a set of words, and S as a set of semantic features:

U: {*crocodile*, *float*, *in*, *the*, *river*}
S: { ..., REPTILE, VERTEBRATE, ..., BODY-OF-WATER, ... }

From such input, the model uses an incremental version of expectation-maximization to learn a probability distribution $P(.|w)$ for each word w over all observed features.

The utterances in the input are taken from a corpus of child-directed speech. To create the associated scene representations, each word in the corpus is entered into a gold-standard lexicon. (This lexicon is never seen by the model.) Each word in the lexicon has a set of semantic features representing its gold-standard meaning. The features for each animal word (and nouns in general) are the names of each ancestor node (hypernym) of the word's first sense in WordNet[1]. Each noun is thus represented by definitional features that reflect conceptual knowledge: general features such as OBJECT, which appear with many words, and more specific features such as REPTILE, which appear with fewer words. For example:

crocodile: { CROCODILIAN REPTILE, DIAPSID, REPTILE, VERTEBRATE, ⋯, WHOLE, OBJECT, ⋯ }

Scene S for utterance U is formed by taking the union of the gold-standard semantic features for all words in U. Thus the semantic input to the model represents naturalistic features that are distributed realistically across related entities, and reflect a conceptual hierarchy intended to approximate the type of conceptual categories children are forming.

An interesting property of the learner is that the learned meaning probabilities for a word w, $P(f|w)$ for observed features f, reflects not only the co-occurrences of w with its gold-standard features: The probabilities importantly capture the influence of contextual features in the input as well. For example, *crocodile* and *hippopotamus* will be distinguished by high probabilities for the definitional features $P(\text{REPTILE}|\textit{crocodile})$ and $P(\text{MAMMAL}|\textit{hippopotamus})$, but are both likely to have a higher than chance value for the feature BODY-OF-WATER since both animals live in rivers. Thus the learned semantic representation in the model captures both definitional and contextual similarities of words.

Constructing a Semantic Network

Other recent research has used free-association norms or conceptual hierarchies like WordNet as the basis for a semantic network; two words are connected by an edge in the network if there is a direct connection between them in the representation (Steyvers & Tenenbaum, 2005; Abbott et al., 2015). By contrast, for our meaning representation (as for BEAGLE data), the appropriate network connections among the words must be determined by considering *how related* any pair of words is in that representation (since all words are implicitly more or less related). We follow Nematzadeh, Fazly, and Stevenson (2014b) in their approach to creating a

[1] http://wordnet.princeton.edu

semantic network over our model's learned meaning representations. Each such word is represented as a node in the network, and pairs of nodes are connected if the cosine similarity of their associated meaning probability vectors exceeds a certain threshold τ. The meaning similarity serves as the weight on an inserted edge.

We experiment with various values for the edge-threshold τ, and at higher values, the resulting network becomes somewhat disconnected: groupings of very similar words form sets of connected components, usually animals of a similar subcategory (e.g. 'farm animals' or 'pets'). This reflects the fine-grained differences in word meaning that the learner has acquired. Because these learned representations do not completely capture taxonomic knowledge – i.e., that 'animal' is a subsuming category of those groupings naturally occurring in the network – we treat the word *animal* differently in deciding on its network connections.[2] Specifically, we use a lower threshold, τ_a, to determine when to add edges including the node for *animal*. This ensures that *animal* is connected to a number of the groupings of animals, and increases the connectivity of the network.

The resulting graph may not be fully connected. Since the fluency task starts with the cue word *animal* and can only reach nodes that are directly or indirectly connected to it, we take the semantic network for our purposes to be the connected component of the graph that includes *animal*. The number of nodes in the semantic network may thus be smaller than the number of observed animal words.

Experimental Methods

The Semantic Networks

The child-directed speech that forms the basis for the input to our word learner is the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) of CHILDES (MacWhinney, 2000). Of the 518 unique animals classified by Hills et al. (2012) using the categories described by Troyer et al. (1997), 111 of these are present in the full corpus and thus in our gold standard lexicon. However, only 93 of these appear in the 481K-word corpus (120K utterances) we use for training. Thus, a semantic network of learned meanings – called a Learner network – will have a maximum of 94 nodes (93 words from the animal subcategories plus *animal* itself).

Recall that the learned representations from our model reflect both definitional and contextual aspects of word meaning; this contextualization of meaning has been shown to influence the structure of resulting semantic networks (Nematzadeh et al., 2014b). For comparison, we create "gold-standard" semantic networks, called Gold, whose edge connections are determined using the gold-standard (definitional) meanings rather than the learned meanings. These networks enable us to see the impact of having the hierarchical semantics from WordNet without the contextually learned aspects of meaning. The Gold networks have a maximum of 112 nodes (111 animal terms+*animal*).

Finally, we used the same method to create BEAGLE semantic networks using the data reported by Hills et al. (2012). BEAGLE consists of word co-occurrence data that encodes contextualized meanings; however, some hierarchical conceptual knowledge is reflected in the 400M-word Wikipedia corpus it was trained on. The BEAGLE data contains 364 animal words that appear in Hills et al. (2012), and thus these networks have a maximum of 365 nodes.

For all semantic networks, we use cosine similarity as the (potential) edge weights, and consider various levels of the thresholds τ_a (for edges that include the word *animal*) and τ (for all other edges) on these weights for inclusion of edges.[3]

Simulating Behavior with Random Walks

Our goal is to see whether the structure of our semantic networks is sufficient to obtain the observed foraging behavior using a simple, uniform search algorithm. To that end, we perform random walks with variations as discussed by Abbott et al. (2015). Each random walk begins at the word *animal* to simulate the fact that *animal* is the cue for the fluency task (*i.e.*, "name as many animals as you can"). Each step in a random walk – *i.e.*, the move from the current node n_c to the next node n_n – is determined by a probabilistic selection over the edges incident on n_c. The selection process may choose the edge to follow in proportion to the edge weights (a weighted walk), or use a uniform distribution over all edges connected to n_c (an unweighted walk). (A further variation in which there is a probability p of jumping back to the word *animal* after any step in the random walk had no appreciable impact on our results, so we do not report that method here.) Due to the probabilistic nature of the algorithm (in selecting edges to traverse), we report results averaged over 282 random walks for each network under parameter settings of interest.

To reflect the time limit in the semantic fluency task, Abbott et al. (2015) fix the number of steps in the random walks to produce approximately the same number of words as human participants. Because this walk length is dependent on properties of the graph being traversed, Abbott et al. set this for each network, using walk lengths of 45 with the BEAGLE data and 2000 on their own semantic network. We take an alternative approach: Instead of picking one walk length to produce a certain number of words, we explore the interaction of different walk lengths with parameters of the networks to see which combinations lead to an appropriate number of words produced. We aim for a range of number of words produced around that of people – i.e., 37 ± 5.

Evaluating IRTs and Patch Switches

In assessing the fit of the random walks to human data, we use the same mapping of steps in the walk to the IRT as used by Abbott et al. (2015). Only the first visit to a node counts

[2]The calculation of model probabilities $P(f|w)$ entails that general features like ANIMAL (that many words share) have lower probability than specific features that distinguish the words.

[3]Our code and data are available at `https://github.com/FilipMiscevic/random_walk.git`.

as producing a word (just as repeats of words are not counted in the human task); the IRT is thus counted between such first visits: i.e., the IRT is the number of steps in the walk between a node n_i the first time it is visited and the next node n_j in the walk that has not been previously visited. Any nodes revisited between such an n_i and n_j increase the IRT between them.

Patch switches occur in the fluency task when participants switch from listing animals in one subcategory (such as 'farm animals') to another (such as 'pets'). Motivated by findings in Hills, Todd, and Jones (2009), we use a "fluid patch model" with the Troyer et al. (1997) categories of animals in analyzing our results. This approach takes into account that animals may belong to multiple categories: a patch switch is considered to have occurred whenever the current novel word and the next novel word in the walk do not have *some* category in common. Patch switches are used to determine the patch entry position in analyzing the match of the random walks to human data (e.g., "1" represents the first item in a patch; "-1" is the last word before a patch switch; see Figure 1).

To assess whether our networks match the human IRT pattern, we consider specific thresholds for the ratio of the IRT at certain points to the overall mean IRT of the random walk. For the patch entry point (1), this ratio for the human data is around 1.2 (cf. Figure 1a); we consider a minimum threshold 1.1 as achieving a fit, with a stricter ratio of 1.2 indicating a better match to human data.[4] For the IRT at position 2, where there is a decrease following the patch switch, we similarly set a maximum threshold of 0.80 of the mean IRT over all. For all other positions, the ratio of IRT to mean IRT must be less than or equal to 1.0. We report walks as matching human data when they meet all these thresholds (and note when the stricter of the patch entry thresholds is met).

Experimental Results

Parameter Search and Selection

Several parameters influence both the number of words produced in a random walk on our networks, and the precise pattern of IRTs and patch switches. The thresholds τ and τ_a used in determining the edges to include in the networks (for non-*animal* and *animal* nodes, respectively) affect both how connected the network is and the actual pattern of connectivity (e.g., all over loosely connected, or a disjoint set of connected components). For example, having fewer edges does not necessarily lead to less connectivity, but might increase the path length between a given pair of words.

Similarly, the number of steps the random walk is allowed – the "walk length" L – clearly influences the number of words produced, but it affects the patterning as well. For example, longer walks have more opportunity to explore more subcategories of words, which can affect the patch switching. Also, a longer walk does not necessarily mean that more words are produced – it might simply raise the IRTs by spending more time revisiting nodes.

[4] Hills et al. (2012) note that to mimic foraging the value simply needs to be higher than the average IRT.

	Weighted			Unweighted		
Network	N	IRT	IRT+	N	IRT	IRT+
Learner	81	44	22	92	**64**	11
Gold	56	14	14	56	**33**	19
BEAGLE	25	**19**	0	25	11	0

Table 1: The *percentage* of 36 walks (weighted and unweighted), varying $\tau \times L$, that match people with respect to (a) the number of words produced (N); (b) N and the IRT pattern (IRT); (c) IRT at the stricter threshold on patch entries (IRT+).

Given that the structure of the network and the random walk length interact to produce both a certain number of words and a given IRT pattern, we perform a parameter search over pairs of reasonable values of the edge threshold τ and the walk length L. (We fix τ_a for connecting *animal* at 0.40, which we found to give good results across all networks.) We vary τ in increments of 0.05, with the range chosen for each type of network based on preliminary experimentation. We vary L from 35 (the approximate number of words produced by people) to 135 (within which all networks showed human-like behavior for some value of τ).

We search over $\tau \times L$ to find the combinations that yield walks over Learner, Gold, and BEAGLE that match the human pattern of responses. We looked for parameters that: (i) produce a range of number of words similar to that of people, and (ii) produce an IRT pattern that matches that of people (as detailed in Methods). Instead of simply finding one parameter setting that achieves these goals, we find a number of such walks across a range of settings, indicating the robustness of the approach to semantic representation.

Overall Patterns Observed

Generally τ and L work together to produce the desired output patterns – i.e., the higher τ's need higher L's to produce the right number of words. We select a range of four τ values (Learner [.70–.85], Gold [.75–.90], BEAGLE [.40–.55]) and nine settings of L (60–100) that exhibit the best performance in showing human behavior (as in (i) and (ii) above). This yields a set of 36 walks in each of the weighted and unweighted settings to analyze; see Table 1.

Overall, BEAGLE performs somewhat better with weighted walks and our networks somewhat better with unweighted walks. The high τ in our networks means edge weights have a small range and are thus very similar – i.e., they are not much more informative than picking uniformly. Also, BEAGLE is trained on a corpus over 800 times the size of ours, so our Learner weights may simply be noisier.

We find that the best performance for BEAGLE (weighted) only matches the target human pattern for 19% of the walks; the best for Gold does so for 33% and the Learner for 64% (both unweighted). Even with weighted walks, our Learner achieves the pattern in 44% of the walks, over twice the number of BEAGLE. We believe that our learned representations, which encode both conceptual knowledge from WordNet coupled with contextual influences from corpus co-occurrences, more robustly reflect the nature of the similarity relations among words for this task. Thus, Learner performs

	Weighted			Unweighted		
Network	N	IRT	IRT+	N	IRT	IRT+
Learner	52	**38**	**38**	38	33	14
Gold	33	**29**	**29**	33	**29**	**29**
BEAGLE	33	**29**	0	33	14	0

Table 2: The *percentage* of 21 walks (weighted and unweighted), for best τ per network, that match people with respect to (a) the number of words produced (N); (b) N and the IRT pattern (IRT); (c) IRT at the stricter threshold on patch entries (IRT+).

better than both Gold and BEAGLE that each only (primarily) capture one of these types of knowledge.

Interestingly, we get these patterns with walk lengths in the range of 60–100, where Abbott et al. (2015) used lengths of 2000 to produce words at the rate of people. Perhaps word co-occurrence data more directly captures relations amongst a wide variety of words compared to the association norms of their data. Future analysis of their network compared to ours may reveal why their walks apparently revisit nodes much more frequently.

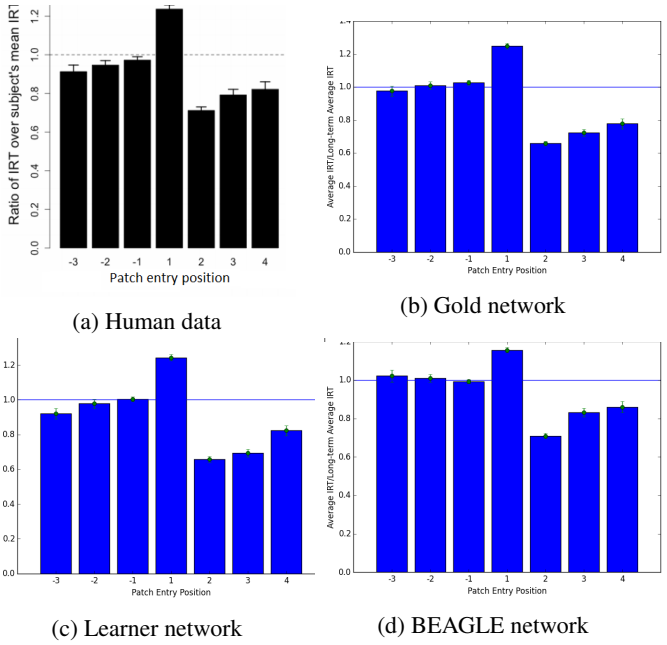
Comparing Best Results

To look more closely at specific patterns, we compared the networks under the best τ parameter for each (Learner: 0.80, Gold: 0.85, BEAGLE: 0.50), with the full range of $L = 35 - 135$; see Table 2. For these settings, we found all networks did the same or slightly better using a weighted walk compared to unweighted. All networks perform very similarly, with the primary difference that the Learner network matches the human target behavior in more walks. Moreover, both Gold and Learner meet the stricter IRT ratio of 1.2 in most cases of weighted walks, while BEAGLE only meets the less strict ratio of 1.1. See Figure 1 for the results of a sample walk ($L = 95$ [Learner], 85 [Gold], 80 [BEAGLE]).

In summary, human-like IRT patterns were observed for random walks on each of the three networks. Importantly, this includes random walks using the BEAGLE data, which Hills et al. (2012) previously showed could not produce such a pattern when used directly. This demonstrates that creating a semantic network from the BEAGLE representation imposes important structure on the raw co-occurrence data, helping the network to focus on meaningful word–word connections. Moreover, the fact that our Learner network shows a very good match to human behavior demonstrates that appropriate representations for a semantic network can be acquired by a cognitively-plausible word learner.

Analyzing the Structure and Semantics of Networks

Previous research suggests that a *small-world* network – a sparse graph with highly-connected sub-networks organized around "hubs" – enables efficient access to semantic information (Steyvers & Tenenbaum, 2005). The idea is similar to foraging: first the hubs are explored, and then a new sub-network connecting to the matched hub is exploited. Indeed, a semantic network created from the association norms used by Abbott et al. (2015) has been shown to have a small-world structure (Steyvers & Tenenbaum, 2005). As in Nematzadeh



(a) Human data (b) Gold network

(c) Learner network (d) BEAGLE network

Figure 1: (a) Human IRTs reproduced from Hills et al. (2012). (b–d) Modeling IRTs in weighted random walks using the parameters described in Comparing Best Results.

	Structure			Semantics		
Network	σ	Nodes	Edges	P	R	F-score
Learner	24	88	205	0.75	0.49	0.59
Gold	24	112	302	0.72	0.50	0.59
BEAGLE	7	136	304	-	-	-

Table 3: The small-world and clustering results for best networks. σ is small-worldness; P and R are average precision and recall.

et al. (2014b), we use well-known graph metrics to calculate a "small-worldness" score (σ) for each of our best networks (see Table 3); when $\sigma > 1$, the network conforms to a small-world structure.[5] We find that all networks exhibiting the target IRT pattern have a small-world structure; in other words, a small-world structure may be *necessary* in producing the human pattern. However, having a small-world structure is not *sufficient*: most of the networks under the wide range of parameter settings we examined have small-world structure, but not all exhibit the foraging behavior.

We observe that an appropriate *graph structure* on its own cannot guarantee efficient search and retrieval: For that, the content of the sub-networks need to appropriately link *semantically-related* words. Indeed, Abbott et al. (2015) also find that their network captures appropriate groupings of animals. We considered whether our networks also reflect the structure of animal subcategories. For the Learner and Gold networks, we can do this by removing the *animal* node and its edges (which we added as the cue word for the random walk), and then labeling each connected component of the network with the most frequently occurring category from Troyer et al. (1997). We take a mean of precision and recall for each

[5]We calculate σ on the largest connected component of the network, which is the part accessible during the random walk; when calculated on the entire network, we obtain similar values.

such cluster, weighted by its size, and compute the F-score (see Table 3). Although not all subcategories of animals are connected to each other (lower recall), the sub-networks have mostly animals from the same subcategory (high average precision), supporting the observed human-like patch switching.

Unfortunately, the networks from BEAGLE do not form such connected components, making this approach to clustering analysis inappropriate. We note here that Abbott et al. (2015) claim the BEAGLE data shows only a "weak signature of animal clusters". We also observe that the small-worldness value is overall larger in our networks than that of BEAGLE; these properties of BEAGLE networks may explain why they do not perform as robustly as our networks in replicating the behavioral data.

Discussion and Future Work

There is an interesting interplay between the richness of representations in semantic memory and the complexity of algorithms required to process it. We show that it is plausible to learn rich representations from naturalistic data for which a very simple search algorithm (a random walk) is enough to replicate the patterns observed in people. Two key factors play a role in the success of our approach: (1) Our learned representations capture the hierarchical relations among words as well as their contextual similarities. (2) We explicitly impose a structure onto our learned representations by creating a semantic network in which words are connected only if their similarity exceeds a certain threshold.

Our work builds on recent research by Hills et al. (2012) and Abbott et al. (2015) in which different representation–algorithm pairs (vectors of co-occurrence statistics and strategic search vs. association norms and random search) replicate the same behavioral data from a fluency task: people name animal words from a subcategory (*e.g.*, pets) until their rate of retrieval is less than the long-term average rate of retrieval, and then they switch to a new subcategory (*e.g.*, farm animals). Importantly, our approach has the advantage that our representations are learned from naturalistic language learning data. Although here we created the semantic networks using the final learned representations of the model, these networks can also be acquired incrementally during word learning (Nematzadeh et al., 2014a).

We further demonstrate that a random walk on a semantic network created from the vector representations of Hills et al. (2012) can produce the observed human pattern. This shows that the co-occurrence statistics learned from a large corpus encodes the required semantic information; however, the explicit structure of a semantic network is needed to simplify the search process. Moreover, our analysis reveals that to replicate the behavioral data all semantic networks (using the various representations) need to have certain connectivity properties – *i.e.*, they consist of highly-connected components, and most nodes are reachable from other nodes via relatively short paths. In future work we will examine other distributional properties of semantic networks that might interact with the search process, such as the scale-free nature of networks, investigated in Thompson and Kello (2014).

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psyc. Rev.*, *122*(3).

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psyc. Rev.*, *82*(6), 407.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cog. Sci.*, *34*(6), 1017–1063.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psyc. Rev.*, *114*(2), 211.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psyc. Rev.*, *119*(2), 431.

Hills, T. T., Todd, P. M., & Jones, M. N. (2009). Optimal foraging in semantic memory. In *CogSci Proceedings.*

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psyc. Rev.*, *122*(3).

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psyc. Rev.*, *114*(1), 1.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psyc. Rev.*, *104*(2), 211.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). Erlbaum.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida free association, rhyme, and word fragment norms.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2014a). A cognitive model of semantic network learning. In *Proceed. Conf. on Empirical Methods in Natural Lang. Processing.*

Nematzadeh, A., Fazly, A., & Stevenson, S. (2014b). Structural differences in the semantic networks of simulated word learners. In *CogSci Proceedings* (pp. 1072–1077).

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cog. Sci.*, *29*(1), 41–78.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Thompson, G. W., & Kello, C. T. (2014). Walking across wikipedia: a scale-free network model of semantic memory retrieval. *Frontiers in psychology*, *5*.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138-146.

Racial Essentialism is Associated with Prejudice Towards Blacks in 5- and 6-Year-Old White Children

Tara Mandalaywala (tmandalaywala@nyu.edu)
Department of Psychology, New York University

Marjorie Rhodes (marjorie.rhodes@nyu.edu)
Department of Psychology, New York University

Abstract

Psychological essentialism is a cognitive bias that leads people to view members of a category as sharing a deep, underlying, inherent nature that causes them to be fundamentally similar to one another in non-obvious ways. Although essentialist beliefs can be beneficial, allowing people to view the social world as stable and predictable, essentialist beliefs about social categories such as race or ethnicity are also thought to underlie the development of stereotyping and prejudice. Whereas recent studies in adults have found that racial essentialism is associated with increased prejudice, the development of this relationship has rarely been examined. The present research examined the implications of essentialism for prejudice in a population of white five- and six-year old children in the United States, and revealed that essentialist beliefs about race are associated with increased implicit and explicit prejudice towards members of a minority racial group.

Keywords: essentialism; social; race; prejudice; cognitive development

Introduction

Humans form categories based on a wide range of features (e.g., color, size, shape), yet in some conceptual domains—particularly for biological and social thought—certain categories take on particularly central roles in cognition and behavior because they are embedded in essentialist beliefs about the structure of the world (Gelman, 2003). Psychological essentialism is a pervasive cognitive bias that leads people to view members of a category as sharing a deep, underlying, inherent nature (a category "essence"), which causes them to be fundamentally similar to one another in non-obvious ways (Medin & Ortony, 1989). In humans, essentialist beliefs about social categories allow people to view the social world as stable and predictable, yet these beliefs might also carry with them deleterious consequences. By emphasizing the distinctiveness of group boundaries, leading people to see social group membership as discrete and immutable, and reifying the idea that social groups are objectively constructed and naturally occurring, essentialist beliefs about social categories have been theorized as a basic cognitive bias that underlies the development of stereotyping and prejudice (e.g., Haslam, Rothschild, & Ernst, 2002; Yzerbyt, Corneille, & Estrada, 2001; Allport, 1954).

Indeed, Mandalaywala, Amodio, and Rhodes (in prep) found that essentialist beliefs about race were associated with increased implicit and explicit prejudice in a sample of white American adults, suggesting that essentialist beliefs provide the cognitive basis for prejudicial attitudes. These findings were consistent with studies that documented a link between the belief that race has a biological basis and increased prejudice towards black individuals (Williams & Eberhardt, 2008; Jayaratne et al., 2006). However, the association between essentialism and prejudice has not been found across the board (e.g., Bastian, Loughnan & Koval, 2011; Bastian & Haslam, 2006; Haslam, Rothschild & Ernst, 2002). These inconsistencies in previous work with adult populations suggest the need for a more precise framework and better understanding of how and why essentialist beliefs relate to prejudiced attitudes. One of the primary points of clarification necessary is whether essentialist beliefs underlie the development of prejudice, or whether essentialist beliefs are invoked as justification for prejudiced attitudes that people might already hold.

There are several mechanisms by which essentialist beliefs could foster the development of prejudice. As essentialism leads individuals to believe that group differences arise from stable and inherent factors, essentialism could lead individuals to view all members of an out-group as more different from themselves than all members of the in-group (No et al., 2008). Similarly, essentialism could lead individuals to believe that social groups demarcate fundamentally distinct *kinds* of entities, perhaps even leading them to dehumanize members of essentialized out-groups. Both of the above processes have been shown to contribute to prejudice in adults (Haslam, Bastian, Bain, & Kashima, 2006; Brewer, 1999). Additionally, essentialism could lead to prejudice by influencing how people respond to negative information about individual group members. For example, because essentialism promotes category-wide generalizations and emphasizes within group similarity, essentialism could make people more likely to draw conclusions about the nature of an entire group based on the negative actions of single individuals (Andreychik & Gill, 2014; Prentice & Miller, 2007).

In light of the mixed patterns of data found in adult populations, however, some theorists have proposed that essentialist beliefs do not lead to the development of prejudice. Instead, these theorists have suggested that essentialism and prejudice are sometimes related to one

another because essentialist beliefs are purposefully adopted in order to justify already held social attitudes or practices. For example, Morton, Hornsey, and Postmes (2009) found that endorsement of certain social essentialist beliefs depended on whether essentialism was being used as a justification for excluding an in-group member or an out-group member, and Mahalingam (2003) found that higher caste Indians were more likely to endorse essentialist interpretations of caste than were lower caste Indians. These findings suggest that people might use essentialism in a post-hoc manner to justify their already held beliefs and, in certain circumstances, to justify their advantaged position in the social hierarchy.

One way to clarify the way in which essentialism is related to prejudice is to consider how this relationship unfolds in early development. Children begin to show race-based social preferences and in-group biases in the preschool years (e.g., Shutts, 2015; Baron & Banaji, 2006), but no prior work has tested whether such prejudiced beliefs and behaviors are related to essentialist beliefs in early childhood. If essentialism underlies the development of prejudice—by accentuating group boundaries, for example—then we might expect that essentialism and racial prejudice will be related to one another as soon as both of these phenomena begin to emerge. Alternately, if essentialist beliefs are adopted strategically to justify already held prejudices and practices, then we might expect essentialist beliefs and racial attitudes to remain distinct in early childhood, and only begin to relate to each other after children gain more experience with the structure of their social world. Thus, examining the relation of essentialism to prejudice in early childhood can provide an initial way to begin to distinguish how these critical cognitive and social phenomena relate to one another.

Like adults, children are prolific essentializers, holding essentialist beliefs about biological and social categories (Taylor, Rhodes, & Gelman, 2009; Waxman, Medin, & Ross, 2007; Gelman, 2003). Social essentialist beliefs emerge in early childhood (by at least age four), but with substantial individual and cultural variation in both the ages at which social essentialist beliefs are acquired (Pauker, Xu, Williams, & Biddle, in press; Kinzler & Dautel, 2012; Rhodes & Gelman, 2009) and in the precise social categories that are essentialized (Diesendruck et al., 2013; Waxman, 2012; Deeb et al., 2011). In the United States, where race is a salient category, children between four- and nine- years of age develop essentialist beliefs about race—coming to view racial categories as stable and determined by birth (Hirschfeld, 1995), as marking similarities within members and differences between groups (Waxman, 2010), and as an objectively accurate way of classifying people (Rhodes & Gelman, 2009), with these beliefs typically emerging at the higher end of this age range in white children (Rhodes & Gelman, 2009).

The implications of essentialist beliefs for other aspects of children's cognitive and social development have rarely been examined. In a notable exception (Pauker, Ambady, & Apfelbaum, 2010), children who believed that skin color is constant over time (e.g., "when you grow up, will you be a white man/woman or a black man/woman?") were also more likely to endorse stereotypes about out-group members (e.g., when asked which of two children would be good at playing basketball, these children were more likely to select the picture of the black child than the picture of the white child). Similarly, at older ages, children who believed that traits are stable over time were more likely to form stereotypes about novel groups (Levy & Dweck, 1999).

The present research goes beyond these previous studies by: (1) directly measuring category essentialism (instead of beliefs about the stability and constancy of individual traits), and (2) examining the relation of essentialism to implicit and explicit prejudicial *attitudes* (i.e. social partner preferences and feelings of warmth towards majority and minority race individuals). Although stereotypes and prejudicial attitudes often co-occur, they are distinct in that prejudicial attitudes are evaluative and rely less strongly on cognitive constructs and explicit knowledge than stereotypes do. Thus, the current study examines the relationship between essentialist beliefs and implicit and explicit prejudice towards blacks in a sample of five- and six-year-old white children who are in the process of developing essentialist beliefs about race.

Examination of the ontological timeline of the emergence of essentialist beliefs and prejudice will shed light onto whether essentialism underlies the development of prejudice, or whether it is invoked as a justification for prejudice after these attitudes are already formed. If essentialism is necessary for the formation of prejudice, then we should find evidence of a relationship in children who have recently acquired both essentialist beliefs and prejudice towards out-group members. However, if essentialism is applied as a justification or rationalization of prejudice, then we might not see a relationship early in development, with a relationship only emerging once more sophisticated knowledge and experience about the structure of the social world is acquired.

Methods

Participants

Participants were 72 five- and six-year-old white children (30 male, M_{age} = 5.88 years, range = 5.00-6.92 years). Child ethnicity was determined by parental report. Participants were recruited from and tested at the Children's Museum of Manhattan. Written parental consent was obtained for all participants and children provided oral assent. All study procedures were approved by the Institutional Review Board of New York University.

Procedure

Participants completed multiple tasks to assess the extent of their essentialist beliefs about race, as well as both their implicit and explicit racial prejudice towards blacks during a single 20 minute testing session. All participants completed the tasks in the same order.

Participant's levels of essentialist beliefs about race were assessed with the visitor task, a commonly used measure of

category essentialism (e.g., Rhodes & Gelman, 2009). This measure of essentialism goes beyond racial constancy, tapping into the beliefs that there are objectively correct ways of classifying people, and that members of the same group are more similar to each other and more dissimilar to members of other groups. In the visitor task, participants were shown a pair of images, consisting of photographs of two children, and were asked to evaluate whether it could be acceptable for a "visitor from some place far away" to group that pair together as the "same kind". If participants view race as an *objective and natural* way to classify people, consistent with an essentialist view of race, they should reject the pairing of a white and black child together and accept only those pairings that follow their expected racial boundaries (i.e. white with white, or black with black). Each child saw four "mismatched" pairs and the total number of times they rejected these pairs was counted and divided by four to obtain an essentialism score in which a value closer to one indicates greater racial essentialism and a value closer to zero indicates less racial essentialism.

To measure participants implicit prejudice towards blacks, we used an evaluative version of the Implicit Association Test (IAT: Greenwald, McGhee & Schwartz, 1998) which is a dual categorization, reaction time test that measures the relative strength of association between compatible pairings (i.e. white faces with positive words and black faces with negative words) as compared to incompatible pairings (i.e. white faces with negative words and black faces with positive words). As not all participants were able to read at the time of testing, we utilized a child appropriate version of the IAT (chIAT) that replaces positive and negative words with smiley and frowny cartoon faces, respectively (Baron & Banaji, 2006).

The chIAT consisted of 5 blocks of trials, during which stimuli were presented in the middle of the computer screen in a randomized order. In Block 1 (10 trials), participants were shown a computer screen on which a cartoon smiley face was presented on the far left side of the screen in front of a yellow background, and a cartoon frowny face was presented on the far right side of the screen in front of a blue background. The participant was first oriented to the picture of the smiley face and told to "press the yellow button anytime you see a smiley face in the middle of the screen", and was then oriented to the picture of the frowny face and was told to "press the blue button anytime you see a frowny face in the middle of the screen". The participant was told to press the buttons as quickly as possible, but without making too many mistakes. After checking for comprehension, the child was randomly and sequentially shown the faces of 10 cartoon smiley or frowny faces (5 smiley and 5 frowny) while pressing the yellow or blue button. Across all Blocks, incorrect responses (e.g., pressing the incorrect button for the face being presented) were marked with a large red "X" and children had to press the correct button before the next face would be presented. In order to avoid excessive memory demands, the reference cartoon smiley and frowny faces in front of the yellow or blue background, respectively, remained on the screen at all times. Block 2 (10 trials) utilized the same training and testing as in Block 1, but with the faces of white and black children in place of smiley and frowny faces, respectively. Racial labels (i.e. black or white) were never used in the instructions, or at any point during the study. Block 3 (30 trials) was the first set of critical trials in which participants viewed the "compatible" pairing (white faces/smiley faces; black faces/frowny faces), Block 4 (20 trials) was used to train the participant on the black and white faces new, switched, locations, and Block 5 (30 trials) was the critical trial for the "incompatible" pairing (white faces/frowny faces; black faces/smiley faces). The block order in which the compatible and incompatible pairings were presented was counterbalanced across participants, with half receiving the compatible pairing first, and half the incompatible pairing first. An IAT D score was calculated following the scoring procedure in Greenwald, Nosek & Banaji (2003), excluding any participant who made incorrect responses on more than 20% of trials (excluded $n = 11$). A positive IAT D score indicates a stronger association for compatible pairings, and thus greater implicit prejudice towards blacks, whereas a negative score indicates greater implicit prejudice towards whites, and a score of zero indicates no implicit prejudice.

Finally, to assess explicit prejudice towards blacks, participants completed two tasks of their racial attitudes: a social preference task and a feelings thermometer task. In the social preference task, participants were shown three sets of pictures, each consisting of one white child and one black child, and for each set they were asked who they would prefer to associate with (e.g., "Here are two kids. Who would you like to invite to your birthday party/to the zoo/to the movies?"). The number of times each participant selected the white child was divided by three (the total number of questions), so that a higher percentage score indicates greater social preference for whites.

In the feelings thermometer task, adapted from the adult version (Amodio & Devine, 2006), each participant was shown six children (3 white and 3 black) one at a time, and asked about each child, "Do you like this kid, or do you not like this kid?" and "How much do you like/not like this kid? Do you sort of like/not like them, like/not like them, or really like/not like them?" Each participant then received an average score for the warmth of their feelings towards all three white children (white composite score) and all three black children (black composite score) separately, to determine the warmth of their feelings towards white and black individuals. As the feelings thermometer measure was incorporated after data collection for this study had already begun, there are fewer participants who completed this measure ($n = 22$).

Data were analyzed in SPSS 22.0. There was no effect of whether the participant received the compatible or incompatible pairing first on the chIAT, so Block order was not taken into account in subsequent analyses. We used Pearson's correlations to examine the relationships between essentialist beliefs and implicit and explicit prejudice.

Results

Overall, participants gave more essentialist responses than would be expected by chance (M = .84, CI = .77-.90; $t(72)$ = 9.92, p < .001), and there was no change in the magnitude of essentialist beliefs about race across development (r (72) = .06, p = .59).

Across all participants, there was an overall implicit bias towards whites (M = .21, CI = .07-.35; $t(61)$ = 3.09, p =.003), with participants demonstrating a stronger association between white/good and black/bad, than between white/bad and black/good. In line with the hypothesis that essentialism leads to the development of prejudice, essentialist beliefs about race were associated with increased implicit prejudice, as measured by the chIAT (r (61) = .31, p = .01: Fig. 1).

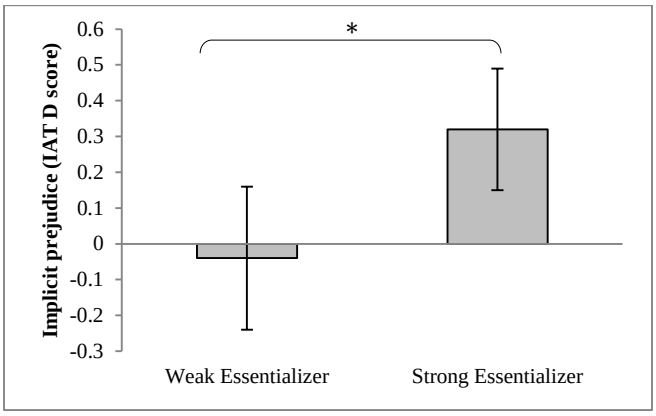


Figure 1: Relationship between essentialist beliefs about race and greater implicit prejudice as measured by the chIAT as shown for weak (M = -.04, CI = -.24-.17, n = 19) and strong (M = .32, CI = .15-.49, n = 42) essentializers. Strong essentializers were significantly more likely to exhibit implicit prejudice towards blacks than weak essentializers ($t(59)$ = -2.55, p =.013).

Essentialist beliefs about race were also marginally associated with greater feelings of warmth towards white individuals, as measured by the white composite score on the feelings thermometer (r (22) = .40, p = .06). We found no relationship between essentialist beliefs about race and feelings of warmth towards blacks (r (22) = -.13, p = .56). There was also no relation between essentialism and responses on the social preference task (r (63) = .08, p = .56).

Discussion

We found that essentialist beliefs about race were related to white children's implicit prejudice towards members of a socially disadvantaged, racial minority out-group, and explicit feelings of warmth towards members of their own socially advantaged, racial majority in-group. Echoing the results of studies done in adults, we found that by the age of five, white children show a clear relationship between essentialism and increased implicit prejudice towards blacks, an out-group, and are beginning to show a relationship between essentialism and greater liking of whites, their in-group. Co-occurrence of essentialist beliefs and prejudice early in development and during the pre-/early-school years when exposure and knowledge of social structure and hierarchy is less fully developed supports the hypothesis that essentialism contributes to the early formation of prejudiced attitudes.

There are several remaining questions that will be important to address in future work. Although this study found that the link between essentialism and prejudice is not dependent upon extensive social experience and knowledge of stereotypes, the precise mechanism by which essentialism might affect the formation of prejudice is still an open question. There are multiple, non-mutually exclusive ways by which essentialism could lead to increased prejudice, and these mechanisms could vary depending on the social group being essentialized, the social group status of the individual who holds the essentialist beliefs, or developmental stage. For example, as members of a socially dominant group, essentialism in white individuals might be associated with greater prejudice towards socially disadvantaged racial groups by emphasizing the boundaries and distinctiveness between racial groups, or by leading them to view out-group members as a distinctly different *kind* of individual (e.g., No et al., 2008). In either manner, these views are likely to increase their in-group bias and positive feelings towards their own racial group, leading them to feel more warmly towards their own group members. White adults and children both demonstrate similar relationships between essentialism and greater in-group liking, in support of the idea that essentialism in white individuals is associated with prejudice towards out-groups by reinforcing group boundaries and increasing in-group preferences.

Without data on black children, it is difficult to determine whether essentialism contributes to the formation of prejudice in the same way across all social groups. As essentialism is related to increased prejudice towards in-group members in black adults (Mandalaywala, Amodio, & Rhodes, in prep), it is possible that essentialism in black children or adults might be associated with prejudice through more of a post-hoc justification of a socially disadvantaged position than was found in white children or adults. By utilizing a more diverse sample and studying these same relationships in black children, we can better understand whether the mechanistic link between essentialism and prejudice is similar across children (i.e. similar relationship regardless of child characteristics), varies based on specific characteristics of both the individual and the particular category being essentialized (i.e. different relationships depending on child characteristics), and/or is developmentally variant (i.e. relationship changes between childhood and adulthood as a consequence of social experience).

Children as young as four- to six-years-old show increased liking and prefer to affiliate with those higher in social status and wealth, and at these ages they also begin to equate race and social status (Shutts, 2015). Therefore, even

though the data we present here did not support the hypothesis that essentialist beliefs are applied to justify prejudice, it is possible that different results would be obtained in minority race children, opening up the possibility that some social groups might use essentialist reasoning in a post-hoc manner, while for others essentialism is a necessary component for the formation of prejudiced attitudes

However, as we did not manipulate children's essentialist beliefs, we are unable in this study to assess a causal relationship between essentialism and prejudicial attitudes. Future research should experimentally manipulate essentialist beliefs about race and assess the consequences for racial attitudes and prejudice to determine whether essentialism is simply related to prejudice or actually leads to it. Findings from recent experimental research in adults suggest that there is a causal relationship in which essentialism directly influences the degree of prejudice reported towards blacks (Mandalaywala, Amodio, & Rhodes, in prep; Williams & Eberhardt, 2008), but the present study does now allow us to support that claim in children.

This study supports the proposal that certain key features of intergroup cognition are continuous across development (e.g., Dunham, Baron, & Banaji, 2008), and that essentialist beliefs about race are more likely to contribute to the formation of prejudiced attitudes than to be a consequence of prejudiced attitudes, at least in socially advantaged white children. Whereas additional research incorporating a more diverse sample of children will continue to clarify the underlying relationship between essentialism and prejudice and the direction of this relationship, this research lays the basis for future studies to investigate new ways to shape early developmental processes to reduce intergroup biases and prejudice towards vulnerable groups in society.

Acknowledgments

This research would not have been possible without the time and efforts of our undergraduate research assistants or without the families who participated in this research at the Children's Museum of Manhattan. We also thank Yarrow Dunham for providing stimuli and scripts for the chIAT, and David Amodio for advisement on measures of prejudice. This work was supported by NSF grant BC-1226942 to Rhodes.

References

Aboud, F. E. (2003). The formation of in-group favoritism and out-group prejudice in young children: Are they distinct attitudes? Developmental Psychology, Vol 39, 48-60.

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Amodio, D.M., & Devine, P.G. (2006). Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. Journal of Personality and Social Psychology, 91, 652-661.

Andreychik, M.R., & Gill, M.J. (2014). Do natural kind beliefs about social groups contribute to prejudice? Distinguishing bio-somatic essentialism from bio-behavioral essentialism, and both of these from entiativity. Group Processes & Intergroup Relations, 18, 454-474.

Baron, A. S., & Banaji, M. R. (2006). The development of implicit attitudes: Evidence of race evaluations from ages 6, 10, and adulthood. Psychological Science, 17, 53–58.

Bastian, B., & Haslam, N. (2006). Psychological essentialism and stereotyping endorsement. Journal of Experimental Social Psychology, 42, 228-235.

Bastian, B., Loughnan, S., & Koval, P. (2011). Essentialist beliefs predict automatic motor-responses to social categories. Group Processes & Intergroup Relations, 14, 559-567.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love or outgroup hate. Journal of Social Issues, 55, 429–444.

Deeb, I., Segall, G., Birnbaum, D., Ben-Eliyahu, A., & Diesendruck, G. (2011). Seeing isn't believing: The effect of intergroup exposure on children's essentialist beliefs about ethnic categories. Journal of Personality and Social Psychology, 101, 1139-1156.

Diesendruck, G., Goldfein-Elbaz, R., Rhodes, M., Gelman, S., & Neumark, N. (2013). Cross-cultural differences in children's beliefs about the objectivity of social categories. Child Development, 84, 1906-1917.

Dunham, Y., Baron, A.S., & Banaji, M.R. (2008). The development of implicit intergroup cognition. Trends in Cognitive Sciences, 12, 248-253.

Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. Oxford University Press.

Greenwald, A.G., McGhee, D.E., & Schwartz, J.K.L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. Journal of Personality and Social Psychology, 74, 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. Journal of Personality and Social Psychology, 85, 197-216.

Haslam, N., Bastian, B., Bain, P., & Kashima, Y. (2006). Psychological essentialism, implicit theories, and intergroup relations. Group Processes and Intergroup Relations, 9, 63–76.

Haslam, N., & Levy, S. R. (2006). Essentialist beliefs about homosexuality: Structure and implications for prejudice. Personality and Social Psychology Bulletin, 32, 471-485.

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? British Journal of Social Psychology, 41, 87-100.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. British Journal of Social Psychology, 39, 113-127.

Hirschfeld, L. A. (1995). Race in the making: Cognition, culture, and the child's construction of human kinds. MIT Press.

Jayaratne, T. E., Ybarra, O., Sheldon, J. P., Brown, T. N., Feldbaum, M., Pfeffer, C. A., & Petty, E. M. (2006).

White Americans' genetic lay theories of race differences and sexual orientation: Their relationship with prejudice toward Blacks, and gay men and lesbians. Group Processes & Intergroup Relations, 9, 77-94.
Kinzler, K.D., & Dautel, J.B. (2012). Children's essentialist reasoning about language and race. Developmental Science, 15, 131-138.
Levy, S.R., & Dweck, C.S. (1999). The impact of children's static versus dynamic conceptions of people on stereotype formation. Child Development, 70, 1163-1180.
Mahalingam, R. (2003). Essentialism, culture, and power: Representations of social class. Journal of Social Issues, 59, 733-749.
Mandalaywala, T.M., Amodio, D.M. & Rhodes, M. (in prep). Essentialism promotes racial prejudice by increasing acceptance of the status quo.
Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning, pp. 179-195, Cambridge.
Morton, T.A., Hornsey, M.J., & Postmes, T. (2009). Shifting ground: The variable use of essentialism in contexts of inclusion and exclusion. British Journal of Social Psychology, 48, 35-59.
No, S., Hong, Y., Liao, H., Lee, K., Wood, D., & Chao, M. M. (2008). Lay theory of race affects and moderates Asian Americans' responses toward American culture. Journal of Personality and Social Psychology, 95, 991-1004.
Pauker, K., Ambady, N., & Apfelbaum, E. P. (2010). Race Salience and Essentialist Thinking in Racial Stereotype Development. Child Development, 81, 1799-1813.
Pauker, K., Xu, Y., Williams, A., & Biddle, A. M. (in press). Race essentialism and contextual differences in children's racial stereotyping. Child Development.
Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. Current Directions in Psychological Science, 16, 202-206.
Rhodes, M., & Gelman, S. A. (2009). A developmental examination of the conceptual structure of animal, artifact, and human social categories across two cultural contexts. Cognitive Psychology, 59, 244-274.
Rhodes, M., Leslie, S.J., & Tworek, C. (2012). Cultural transmission of social essentialism. Proceedings of the National Academy of Sciences, 109, 13526-13531.
Shutts, K. (2015). Young children's preferences: Gender, race, and social status. Child Development Perspectives, 9, 262-266.
Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. Child Development, 80, 461-481.
Williams, M. J., & Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. Journal of Personality and Social Psychology, 94, 1033-1047.
Waxman, S.R. (2010). Names will never hurt me? Naming and the development of racial and gender categories in preschool-aged children. European Journal of Social Psychology, 40, 593-610.
Waxman, S.R. (2012). Social categories are shaped by social experience. Trends in Cognitive Sciences, 16, 531-532.
Waxman, S.R., Medin, D., & Ross, N. (2007). Folkbiological reasoning from a cross-cultural developmental perspective: early essentialist notions are shaped by cultural beliefs. Developmental Psychology, 43, 294-308.
Yzerbyt, V., Corneille, O., & Estrada, C. (2001). The interplay of subjective essentialism and entitativity in the formation of stereotypes. Personality and Social Psychology Review, 5, 141-155.

The Combinatorial Power of Experience

Brendan T. Johns,[1] Randall K. Jamieson,[2] Matthew J. C. Crump,[3] Michael N. Jones, [4] & Douglas J. K. Mewhort[5]
[1] btjohns@buffalo.edu, Department of Communicative Disorders and Sciences, University at Buffalo, Buffalo, NY
[2]Department of Psychology, University of Manitoba, Winnipeg, MB
[3]Department of Psychology, Brooklyn College, Brooklyn, NY
[4]Department of Psychological and Brain Sciences, Indiana University, IN
[5]Department of Psychology, Queen's University, Kingston, ON

Abstract

Recent research in the artificial grammar literature has found that a simple exemplar model of memory can account for a wide variety of artificial grammar results (Jamieson & Mewhort, 2009, 2010, 2011). This classic type of model has also been extended to account for natural language sentence processing effects (Johns & Jones, 2015). The current article extends this work to account for sentence production, and demonstrates that the structure of language itself provides sufficient power to generate syntactically correct sentences, even with no higher-level information about language provided to the model.

Keywords: Language production; Computational models of language; Corpus-based models.

Introduction

Human languages are both productive and regular. They are productive in that an infinite number of utterances are possible for any language, and regular in that the utterances produced by speakers are systematic. In order to explain these aspects of language, it has been proposed that it is necessary to have a formal grammar underlying language performance (Chomsky, 1988). A grammar of sufficient complexity can construct utterances of any length while maintaining consistency in utterance construction.

The need and evidence for formal grammars in explaining language performance has been argued in many different places (e.g. Christiansen & Chater, 2008; Evans & Levinson, 2009). The goal of this article is not a rehashing of this debate but instead to examine the interaction between regularity and productivity, cornerstones of any theory of language, with the ultimate goal being to understand the constraints that the structure of language itself provides. Specifically, given that languages are highly regular and consistent, the question that will be examined here is what power this regularity provides to the production of grammatically correct utterances.

Historically, to make the study of language more manageable, researchers have studied how people learn the grammar of small artificial languages. Classic accounts of artificial grammar results propose that as subjects are exposed to strings randomly generated from a pre-defined grammar, they are capable of using this experience to internally generate a representation of this grammar, and in turn use this abstraction to discriminate between grammatical and ungrammatical strings (Reber, 1967). However, people are unable to verbally describe the rules they used to accomplish this discrimination (i.e., it is implicit knowledge). Recent results by Jamieson and Mewhort (2009, 2010) call this explanation into question. Jamieson and Mewhort demonstrated, across a wide variety of tasks and manipulations that a simple explicit memory model (Minerva 2; Hintzman, 1986) can account for most of the relevant findings in the artificial grammar literature.

The reason why a simple recording of the environment could explain artificial language results is clear: the stimuli contained in an artificial grammar task provide a lot of information about the structure of the underlying grammar. That is, like natural language, the artificial languages that are constructed from artificial grammars are *regular*. This regularity causes the underlying structure of a grammar to emerge across the exemplars that are displayed in an artificial grammar experiment, even with no higher-level abstraction taking place during learning.

To formalize this, Jamieson & Mewhort (2011) have implemented the latter position of a string using a model of memory that combines the storage and retrieval operations from Hintzman's (1986) MINERVA 2 model of episodic memory with the holographic reduced representations from Jones and Mewhort's (2007) BEAGLE model of word meaning. According to the model, each studied grammatical sentence is stored to memory. When a probe is presented at test, it retrieves all of the stored sentences in parallel. If the information retrieved from memory is consistent with the probe, the probe is judged to be grammatical; else, it is judged to be ungrammatical.

Despite the model's simplicity, it predicts a surprising number of results in the artificial-grammar task including (a) the linear relationship between mean judgement accuracy and the redundancy of the generative grammar, (b) judgements of grammaticality for individual test items, (c) grammatical string completion, and (d) variation in peoples' judgements depending on how they represent strings in memory (Chubala & Jamieson, 2013; Jamieson & Mewhort, 2009, 2010, 2011). But, why?

The power of this model comes from the natural correlation between the form and amount of structure in an exemplar produced with a grammar and the grammar that was used to produce it (Jamieson & Mewhort, 2005). It follows, then, that each studied grammatical exemplar provides information about the underlying grammar. It also follows that a collection of grammatical exemplars will almost always provide a sum of information greater than that provided by one exemplar alone. That is, it is the combination of exemplars that gives the model its power. The question, then, is not how information about the grammar can be stored in memory, but how is that information harnessed at retrieval?

In artificial grammar experiments, parallel retrieval of grammatical exemplars is sufficient (Jamieson & Mewhort, 2011). But, can that class of explanation scale from handling small artificial grammars to handling how people learn grammar in their natural language?

The goal of this article is to examine the power that the productive and regular nature of language provides in allowing an exemplar model of natural language to produce grammatically correct utterances. Exemplar models are attractive models to examine this question because the complexity of the model lies in the complexity of the information that is being encoded. By storing natural language sentences within an exemplar memory, and determining how effective this stored information is at generating grammatically correct utterances, an examination of the power of the structure of language is provided.

The Exemplar Production Model (EPM)

Our model combines the storage and retrieval scheme from Hintzman's (1986) MINERVA 2 model of episodic memory with the reduced holographic representation scheme from Jones and Mewhort's (2007) BEAGLE model of semantics, and is similar to an exemplar approach to language comprehension recently explored by Johns & Jones (2015).

Representation

In the model, each word is represented by its own unique random vector, **w**, of dimensionality *N*, where each dimension takes a randomly sampled value from a normal distribution with mean zero and standard deviation $1/\sqrt{N}$. In the simulations that follow, $N = 1024$.

Any given sentence is represented by two vectors, both of which are constructed from the word representations. The first vector, **c**, is called the sentence's context vector and is computed as,

$$\mathbf{c} = \sum_{i=1}^{n} \mathbf{w}_i$$

where **c** is the context vector, *n* is the number of words in the sentence, and $\mathbf{w}_i$ is the vector that represents the word in serial position *i* of the sentence. As shown, the context vector sums the information from all of the words that appear in the sentence, but it does not include any information about the order in which the words occurred. For example, the context vector that encodes "eat your dinner" is equal to the context vector that encodes "dinner your eat".

The second vector, **o**, is called the sentence's order vector and is computed as,

$$\mathbf{o} = \sum_{i=1}^{n} \mathbf{w}_i \circledast \mathbf{l}_i + \sum_{i=2}^{n} \mathbf{w}_{i-1} \circledast \mathbf{w}_i + \sum_{i=3}^{n} \mathbf{w}_{i-2} \circledast \mathbf{w}_{i-1} \circledast \mathbf{w}_i$$

where **o** is the order vector, *n* is the number of words in the sentence, w_i is the word in serial position *i*, w_{i-1} is the word in serial position $i - 1$, w_{i-2} is the word in serial position $i - 2$, l_i is a vector that represents serial position *i*, and $\circledast$ denotes directional circular convolution (see Jones & Mewhort, 2007; Plate, 1995). As shown, the order vector sums information about (a) what word appears in each serial position in the sentence (i.e., serial position information), (b) which pairs of words follow one another from left to right in the sentence (i.e., bigram information), and which triplets of words follow one another from left to right in the sentence (i.e., trigram information). Given the inclusion of trigram information, the formula cannot be applied to a sentence with fewer than three words.

Finally, a sentence's vector representation, **s**, is a $2N$ dimensional vector formed by concatenating the N dimensional context vector and the N dimensional order vector such that dimensions $1 \ldots N$ in **s** store the context vector and dimensions $N+1 \ldots 2N$ in **s** store the sentence's order vector. Thus a sentence is represented as a vector **s** that is equal to **c** // **o**, where // represents concatenation.

Storage of language experience

To represent experience with language, we store *m* sentences to a $m \times 2N$ matrix, where rows represent sentences and columns represent features that encode the information in the sentence. Thus, memory for 1000 sentences is represented in a 1000×2048 matrix whereas memory for 125,000 sentences is represented by a $125{,}000 \times 2048$ matrix.

Retrieval

Retrieval in the model is probe-specific, similarity- driven, and parallel. When a probe is presented to memory, it interacts with the information in the stored traces to construct the memory of a previously experienced event. Decision follows from the construction. Since retrieval is similarity-driven, a probe retrieves traces that are similar to it. Because a probe retrieves whole traces from memory and these whole traces record both context and order information in a sentence, a probe that includes just the context information will also retrieve the order information that it has co-occurred with in the past. This is how the model simulates cued-recall, and it is the mechanism that the model uses to retrieve a sentence (i.e., word order) given a context vector (i.e., an unordered list of words).

The echo, **e**, is computed as,

$$\mathbf{e} = \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{N} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{N} p_j^2} \sqrt{\sum_{j=1}^{N} M_{ij}^2}} \right)^9 \times M_i$$

where **p** is the context vector that encodes an unordered list of words (i.e., includes information in serial positions $1 \ldots N$ with serial positions $N+1 \ldots 2N$ set to zero), **M** is the memory matrix that stores the model's sentence knowledge, **e** is the echo, and *m* is the number of sentences stored in memory. As

with a sentence representation, features 1…*N* in **e** represent the context vector retrieved from memory and features *N*+1…2*N* in **e** represents the order vector retrieved from memory. Although it is not explicitly stated in the formula, if the similarity between a probe and memory trace is less than zero, the similarity is rewritten as equal to 0 (i.e., equivalent to excluding the trace from the calculation of the echo). This same procedure was used in Johns and Jones (2015) and Kwantes (2005), and is useful because it allows for the amount of noise in the echo to be reduced.

Decision

Our goal is to measure the model's ability to produce a syntactic sentence composed of words presented in an unordered words list. For example, given the words eat, dinner, and your, we would like the model to produce "eat your dinner" rather than "dinner eat your".

To accomplish the transformation from unordered word list to syntactic production, the model compares the order vector in the echo to each of the *n!* order vectors corresponding to the *n!* ways or ordering the words in the unordered list. For example, given the list eat, your, and dinner the model retrieves an order vector based on the context vector, $\mathbf{c} = \mathbf{w}_{eat} + \mathbf{w}_{your} + \mathbf{w}_{dinner}$, and then compares the retrieved order vector against all 3! = 6 sentences that can be constructed from the three words: "eat your dinner", "eat dinner your", "your eat dinner", "your dinner eat", "dinner eat your", and "dinner your eat". The order vector that is most similar to the information in the echo is selected as the best alternative. Because all other orders bear some similarity to the order information in the echo, the operation can also be used to rank order the model's preference over all possible *n!* sentences from first (i.e., most similar) to last (i.e., least similar).

Summary

In summary, the model builds and stores a representation of each sentence in a text corpus. When presented with an unordered word list, the model retrieves a corresponding order vector and produces a word order that corresponds to the order vector that is most similar to the order vector retrieved from memory.

Simulations

The simulations that follow apply the model to a sentence production task. Each simulation involved two major steps. First, we constructed a record of language experience by storing *m* sentences of length *n* to memory; the sentences were sampled randomly from a corpus. Second, we computed the model's ability to translate each of 200 unordered word lists of length *n* into ordered sentences of length *n*.

We expect the model will re-write unordered word lists as syntactic sentences. If true, our simulations would demonstrate that parallel retrieval from a record of language is sufficient to produce at least one hallmark of syntactic behavior. This would reinforce the power of exemplar models of language, and would add to the growing literature on the importance of individual experience with language (Abbot-Smith & Tomasello, 2006; Jamieson & Mewhort, 2010, 2011; Johns & Jones, 2015).

Sentences

Given our goal is to conduct an analysis of natural language, it was critical that we get a fair sample of natural language use. To meet that demand and to model a wide range of language experience, we assembled a pool of 6,000,000 sentences from a number of sources including Wikipedia articles, Amazon product descriptions (attained from McAuley & Leskovec, 2013), 1000 fiction books, 1050 non-fiction books, and 1500 young adult books. Once collated, we edited the list to exclude repetitions and, then, we organized the total list into sub-lists of sentences composed of 3, 4, 5, 6, and 7 words. Finally, we used the sentences in the final pool to construct a list of 200 three word test sentences, 200 four word test sentences, 200 five word test sentences, 200 six word test sentences, and 200 seven word test sentences. All sentences simple in construction, and use mostly high frequency word, but given the complexity of the task provide a useful assessment of the model's performance (all test sentences can be found at http://btjohns.com/experience_sents.zip). No general syntactic construction was used, but the majority consist of single phrase structures. Examples of sentences used for each sentence size are:

3 - *Eat your dinner*.
4 - *The children were flourishing*.
5 - *He walked into the bedroom*.
6 - *She held something in one hand*.

Simulation parameters

We conducted simulations as a function of two key parameters: sentence length (i.e., *n*) and language experience (i.e., *m*). Analysis of sentence length was accomplished by conducting separate simulations for sentences of length *n* = 3, 4, 5, 6, and 7. Analysis of language experience was accomplished by conducting separate simulations given *m* = 1000, 2000, 3000 … 125,000 sentences stored in memory. To ensure that results were not conditional on a particular record of language experience, each simulation was conducted using a different random sample of *m* words. Crossing both factors produces a 5 (sentence length) x 125 (language experience) factorial design.

Two measurements of performance

We measured sentence completion performance two ways. The first method tallied the percentage of tests in which the model most preferred the word order that corresponded to the original sentence.

The second method ranked the model's decisions for all possible word orders from first (i.e., most similar to the order vector in the echo) to last (i.e., least similar to the order vector in the echo) and, then, recording the rank at which the original input sentence appeared. For example, if the model was given "eat your dinner" it would produce a rank order of all six possible sentences composed of the three input words. If "eat

your dinner" was the third preferred word order, the trial would be scored as a rank 3 decision.

In summary, the first percent correct measure was an absolute index of model performance, as if the model (like an experimental subject) provided a single response for each test sentence. The second rank based measure offers a more nuanced assessment that measures how close the model was to making the right decision whether its first choice matched or did not match the exact word order in the test sentence.

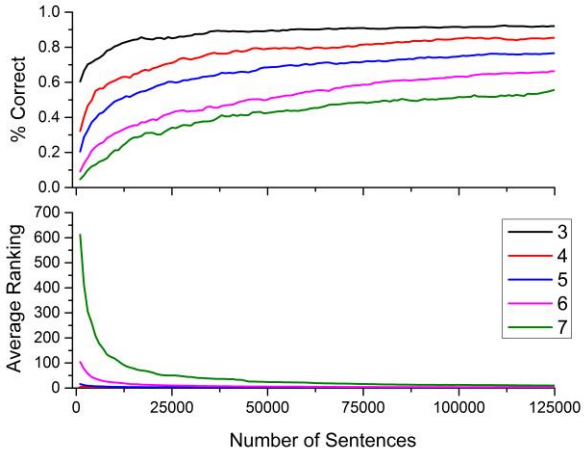


Figure 1. Syntactic completion rates. Top panel shows the percentage of items that the model reproduced the exact word order in the test sentence. Bottom panel shows the mean rank order of the model's preference for the exact word order in the test sentence where 1 = best.

Results

Figure 1 shows results over the complete 5 × 125 factorial design. The top panel in Figure 1 shows the percent correct production rate (e.g., the model returned "eat your dinner" when presented with "dinner your eat"). The bottom panel in Figure 1 shows the mean rank of the model's decisions. In both cases, performance for 3, 4, 5, 6, and 7 words test sentences are presented as different lines with language experience measured in number of sentences ordered along the abscissa.

As shown in the top panel of Figure 1, the model's ability to reproduce word order in the presented test sentence varies as a monotone function of sentence length, being best for short sentences and worse for long sentences. However, one must exercise caution in making that comparison. Chance performance changes dramatically as a function of sentence length. At the extremes, when n = 3 the probability of guessing the correct word order is equal to 1 in 6 whereas chance is equal to 1 in 5040 when n = 7. Thus, performance at m = 1000, the final score of 90% correct in the three word condition might be considered as or even less impressive than the corresponding but lower score of 52% correct in the seven word condition.

Also shown, the model's performance improves as a function of language experience, with the improvement being fastest early on in the addition of sentences to memory and slowing considerably after memory includes a record of approximately 50,000 sentences.

The bottom panel in Figure 1 presents a more nuanced picture of results. As shown, when language experience is modest, m = 1000, the model commits far more misses (i.e., large mean rank scores). But as language experience increases, mean rank scores drop considerably to nearly 1 for most sentence lengths, and under 10 for all. Combined, these figures show that even though the model does not always choose the particular word order in the input sentence, it nevertheless has a strong preference for that specific word order. So, what does a near miss mean?

Although we have scored near misses as wrong, they may occasionally be correct in the broader sense. For example, consider that the model preferred "they quietly went down the stairs" when tested on "they went quietly down the stairs". Although the model failed to produce the input sentence, it nevertheless generated a syntactic alternative.

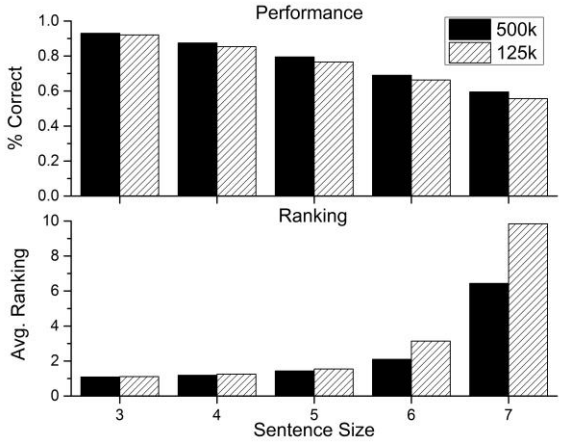


Figure 2. Performance from Figure 1 where m = 125,000 and results from a new simulation were m = 500,000.

Figure 2 shows results when m = 125,000 sentences (hatched bars; previous data from Figure 1) and when m = 500,000 sentences (closed bars). The top panel shows results with the percent correct measure; the bottom panel shows results with the average rank measure. Although the results in the hatched bars in Figure 2 are already presented in Figure 1, the re-presentation shows differences that cannot be seen in in the ranking data in Figure 1.

As shown, the mean rank was less than 2 for the three, four, and five word sentences, was less than 4 for the six word sentences, and was less than 10 for the seven word sentences. Given the average expected ranks for guessing with 3, 4, 5, 6, and 7 word sentences are equal to 3, 12, 60, 360, 2520, and 5040 respectively, the results are very impressive indeed. Plus, as we have already mentioned, whereas a rank greater

than one indicates that the model failed to reproduce the exact order of words from the input sentence, the model still might have chosen an alternative syntactic completion—a possibility that increases with the number of words in the sentence. We plan to assess this possibility in the future.

Finally, the solid bars in Figure 2 show results for simulations where *m* was increased from 125,000 to 500,000—a larger sample of language experience that is more in line with work in other corpus based model analyses (e.g., Landauer & Dumais, 1997).

As shown, increasing the number of sentences in memory from 125,000 to 500,000 produced a modest 2-3% overall improvement of performance but with most of that improvement in judgments about longer sentences. This result reinforces our previous conclusion that after an initial 50,000 sentences are stored in memory (see Figure 1), each additional sentence has a diminished impact on the model's performance.

In summary, Figures 1 and 2 represent a very high level of performance, even at large sentence sizes. For smaller sentence sizes of 3, 4, and 5, the model operated at greater than 75% correct, and was greater than 50% correct even at seven word sentences. Interestingly, much of the improvement in the model's performance was made with a relatively small number of sentences, with small improvements after 50,000 sentences. The reduction in performance across sentence sizes is linear, suggesting that as more sentence types are possible, the introduction of noise reduces model performance by a constant. In fact, the final ranking across the different sentence sizes was almost entirely due to the number of alternatives in the construction process, with a Pearson correlation equal to 0.99 between final ranking and number of permutations. As already discussed, this may also be a function of there being more possible syntactic constructions for words being greater with a higher number of words.

In conclusion, the model demonstrates a simple point: a raw record of language experience combined with parallel retrieval can provide a powerful mechanism to account for how people order words in sentences. The analysis also suggests that the body of linguistic experience need not even be overwhelmingly large and that a relative few (i.e., 50,000 exemplars) can go a long way to helping a person produce syntactic word orders in their natural language. Finally, the analysis also demonstrates that an appreciation of syntax can emerge from the application of a simple parallel retrieval mechanism applied to a realistically scaled representation of language experience.

General Discussion

Natural languages are defined by productivity and regularity. They are capable of producing an infinite number of different utterances, with all the utterances having a consistent structure. To account for these differing aspects of language it has been proposed that a formal grammar is necessary.

A formal grammar is a top-down approach, which seeks to understand language processing from the connections between abstract categories. The approach taken here with the Exemplar Production Model is the opposite of this: use a model that only knows the structure of past utterances, and use that experience to construct a future utterance. That is, it is a bottom-up approach, where past experience controls future behavior. The EPM was designed to exploit the productivity and regularity of natural language, in order to determine the power of experience in producing grammatical utterances.

The EPM proposes that it is not an analysis or encoding of a single utterance that provides the model power. Instead it is the overlap in the usage of language: even though no two utterances may be identical (productivity; this was ensured in the above simulations by removing repeat sentences), the structure of a language emerges as a function of recorded exemplars. That is, it is the combinatorial statistics that emerges from the parallel retrieval of a relatively large number of sentence exemplars that provides the model the ability to construct grammatically correct sentences. What this suggests is that the consistent, but not identical, structure of studied utterances allows for the development of grammar-like behavior, albeit without an actual grammar.

The EPM is a simple model that encodes pure location and linear n-gram information to encode an exemplar of a sentence. A classic retrieval operation, based off of MINERVA 2, is used to construct the likely ordering of a sentence. Every possible ordering of a sentence is tested, with the ordering that is most similar to the expected structure being produced. There is no higher-level processing integrated into the model, and so the behavior of the model is entirely experience-dependent. In that sense, the theory is perfectly continuous with our previous efforts to build an exemplar-based model of language learning and comprehension using the same mechanisms and ideas (see Johns & Jones, 2015). However, there are some differences in the details of the current and previous models that need to be resolved before a complete integration of the two is realized. We take the problem of that integration as a challenge that would move toward the kind of model needed to generate a complete picture of how an exemplar-based model of memory can serve as a valuable competitor in the discipline's pursuit of a theory of language and language use.

The model was shown to be able to construct the correct ordering of simple sentences of sizes 3 to 7 to a high degree, with a linear drop in performance as sentence size increases. This demonstrates that past experience with language provides a large amount of power in producing grammatically correct utterances.

However, the really interesting part of the model's behavior is the performance of the EPM as a function of the number of exemplars it has studied. Performance rapidly increases with the first 20-25,000 sentences studied, with small improvements subsequently (the learning function most resembles a power law). Even when the total number of exemplars studied is quadrupled from 125k to 500k sentences, only a small improvement in performance is seen. However, it does provide a look into what the regular nature

of language provides productivity: even with a small amount of linguistic experience, the correct structure of language emerges, due to the highly structured nature of natural language. Language is far from random and this consistency provides a simple model the ability to construct grammatically correct sentences, without any higher-level processing. As more exemplars are stored, the overlap in structure of the sentences emerges (due to the productivity of language), which allows for the model to exploit the combinatorial nature of language usage.

This is not to say that this approach does not have any challenges. The main one being that the model potentially operates at the wrong level of analysis – phrases may be the right unit of exemplars rather than whole sentences, as is the typical case in generative linguistics. Sentences then can be constructed by determining the correct order of phrases, integrating higher-level information into the exemplar construction process. This would also allow for the model to operate with lower number of words, which would be advantageous due to the model becoming computationally burdensome at a high number of words. It would also allow for the model to be tested on longer sentence sizes.

Another issue with the model concerns its encoding scheme: if it is generating the structure of a sentence of size *n*, it studies only exemplars of the same size. More research is required to determine the best mechanism to encode location in a relative fashion, where exemplars of differing length can be included in the same retrieval process.

However, these problems arise because of the simplicity of the approach, which is also its most promising feature. There is very little built into the machinery of the model and it still operates at a high level of performance. It provides a promising framework to examine language production and comprehension from a bottom-up point of view and allows for an examination into the power of experience in explaining linguistic behavior.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23, 275-290.

Chomsky, N. (1988). Language and the problems of knowledge. *The Managua Lectures*. Cambridge, MA: MIT Press.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489-558.

Chubala, C. M., & Jamieson, R. K. (2013). Recoding and representation in artificial grammar learning. *Behavior Research Methods*, 45, 470-479.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32, 429-492.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411-428.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 9–23.

Jamieson, R. K., & Mewhort, D. J. K. (2009). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, 62, 550-575.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String-completion and performance for individual items. *Quarterly Journal of Experimental Psychology*, 63, 1014-1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to Kinder (2010). *Quarterly Journal of Experimental Psychology*, 64, 209-216.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, 69, 233-251.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, 12, 703-710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

McAuley, J., and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (RecSys), 165–172.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6, 623–641.

Reber, A. S. (1967). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 2, 88-94.

Discriminability of sound contrasts in the face of speaker variation quantified

Christina Bergmann (chbergma@gmail.com)
Alejandrina Cristia (alecristia@gmail.com)
Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS; 29, rue d'Ulm
75005 Paris, France

Abstract

How does a naive language learner deal with speaker variation irrelevant to distinguishing word meanings? Experimental data is contradictory, and incompatible models have been proposed. Here, we examine basic assumptions regarding the acoustic signal the learner deals with: Is speaker variability a hurdle in discriminating sounds or can it easily be ignored? To this end, we summarize existing infant data. We then present machine-based discriminability scores of sound pairs obtained without any language knowledge. Our results show that speaker variability decreases sound contrast discriminability, and that some contrasts are affected more than others. However, chance performance is rare; most contrasts remain discriminable in the face of speaker variation. We take our results to mean that speaker variation is not a uniform hurdle to discriminating sound contrasts, and careful examination is necessary when planning and interpreting studies testing whether and to what extent infants (and adults) are sensitive to speaker differences.

Keywords: language acquisition; speech; acoustics; machine classification

Background

One important problem that infants are confronted with during language development is speaker variability. The same word might be implemented differently at the phonological level (as popularized in the song stating "I say tom[e]to, you say tom[ɑ]to"), acoustic targets can vary (Cristia, 2011), and speakers may differ in the shape of their vocal tract. As a result, the same word spoken by different people varies in its acoustic realization, even within the same accent. This problem is so potent that it requires drastic measures within speech recognition technologies to ensure proper handling of speaker variation: Typically, systems implement speaker normalization components which are trained with hundreds of speakers to attain a reasonable, but not perfect, performance on unknown speakers. Infants face the same problem: they have to recognize words as identical across speakers while maintaining the ability to distinguish different words (e.g., *take*$_1$ and *take*$_2$, versus *bake*$_1$). Alas, infants cannot use the same mechanisms as automatic speech recognition systems, while managing to learn their native language successfully and become adults who can generally deal with speaker variation (while being nonetheless affected by it). It is unclear which abilities infants bring to this task and which computational mechanisms they use.

Psycholinguistic models of speech and speaker processing range between two extremes. In the *abstractionist* view, infants are born with or rapidly acquire a system which separates speaker dependent and linguistically relevant phonetic information. For example, following a study with neonates, Dehaene-Lambertz and Pena (2001) state that "normalization is present from birth and is not the consequence of the establishment of phonetic prototypes following extensive exposure to speech." This means sounds and words are represented as abstract entities, invariant across speakers. The other end of the theoretical spectrum is the *episodic* view, where phonetic and speaker-specific information is not separated (Goldinger, 1998). In *hybrid* models, both invariant and speaker dependent formats of representation are combined to varying degrees (for a recent review see Kleinschmidt & Jaeger, 2015).

All models make implicit assumptions about the acoustic signal, which forms the basis for processing, representation, and learning. If a normalization and abstraction mechanism is innate, it seems necessary that, for a learner who starts from acoustic representations, linguistically contrastive information can be a priori separated from speaker-specific 'noise'. However, it is unlikely that such a separation is universally possible, since linguistic and speaker-specific variation has been stated to be intertwined and difficult to separate in several cases (Kleinschmidt & Jaeger, 2015). If, on the other hand, a normalization and abstraction mechanisms is learned, then previous to learning, abstractionist and exemplarist models are indistinguishable. In brief, before we can separate abstractionist from exemplarist models, it is important to establish which cases of speaker variability can be dealt with on the acoustic level and which require learning. As a preliminary of this, we need a method, which can, in a case by case fashion, quantify the amount of acoustic variation introduced by speaker variation.

Here, we propose a systematic measurement of the discriminability of phonetic categories of English (e.g., /a/ vs /o/, /a/, vs /i/, etc) as a function of speaker variability. We use a computerized ABX discrimibility score, whereby the acoustic distance between tokens that belong to the same category is compared to the distance between tokens from different categories. We systematically vary whether all tokens are spoken by the same or different speakers to quantify the impact of speaker variation[1].

In this paper, we first review some of the findings on the effect of speaker variability in infants' discrimination abilities. We then compute the acoustic discriminability scores of English vowels and consonants with a focus on the contrasts that were tested empirically.

[1]This method is not a model of infants or adults performing a particular discrimination paradigm. Instead, it is intended to evaluate how well the phonetic categories of a language are intrinsically separated in an acoustic/auditory representation.

Table 1: Studies on infants' sound discrimination abilities in the face of speaker variation. *Age* is reported in months; *Support* indicates how the results were interpreted; *Discrimination abilities* are divided into tasks where the speaker does not change (*within*) and where speaker variation was present (*across*). *Differences* are split into numerical ("$>$" signifies higher performance within-speaker, $\neq$ means there is a difference but a direction cannot be established) and statistical difference, based on tests of the interaction (sound type and task; significant: *, non-significant: ^{ns}.)

Contrast	Age	Reference	Support for	Discrimination results		
				Within	Across	Difference
/p/-/t/	0	Dehaene-Lambertz and Pena (2001)	Abstract	Yes	No	$>^{ns}$
/b/-/d/	2	Jusczyk, Pisoni, and Mullennix (1992)	Episodic	Yes	Yes/No	$>$
/a/-/i/	2,3,6	Marean, Werner, and Kuhl (1992)	Abstract	Yes	Yes	NA
/a/-/i/	5	Polka, Masapollo, and Ménard (2014)	Hybrid	NA	Yes	NA
/a/-/i/	6	Kuhl (1979)	Abstract	Yes	Yes	NA
/a/-/ɔ/	6	Kuhl (1983)	Abstract	Yes	Yes	NA
/b/-/p/ (/n/-/ŋ/)	7.5	Clough (2015)	Abstract	Yes	Yes	$\neq^{ns}$
/ɛ/-/ɪ/	12	Escudero, Bonn, Aslin, and Mulak (2015)	Episodic	Yes	NA	NA

Infants' Sound Discrimination Abilities in the Face of Speaker Variability

Infants' ability to deal with speaker variation has been investigated using a range of tasks, including word segmentation (Houston & Jusczyk, 2000), word learning (Rost & McMurray, 2009), learning of phonotactic rules (Seidl, Onishi, & Cristia, 2014), and sound discrimination. The last is most relevant for the present study, an overview can be found in Table 1. For reasons of space, we present in detail a representative selection of this line of work.

The logic of infant discrimination studies is as follows: infants first hear sequences of isolated syllables serving as background, followed by deviating stimuli. If a significant difference arises between (new tokens of) background and deviating stimuli, this is taken as evidence that infants can discriminate the two stimulus classes. Most studies analyze overt behavior, such as looking to an unrelated visual display, as indicator of infants' processing. We summarize each study in order of infant age. While infants' general discrimination ability matures over the course of the covered age range (Tsuji & Cristia, 2014), we cannot discern a clear developmental trend in the existing experimental literature pertaining to their ability to deal with speaker variation.

Dehaene-Lambertz and Pena (2001) measured electrophysiological responses to a deviating syllable compared to a background syllable, which was either spoken by one or multiple speakers, in a group of sleeping neonates. The authors took a main effect of condition (deviant versus background) and the lack of an interaction with speaker (within versus across) as evidence of infants' ability to ignore irrelevant information by normalizing over speakers. However, post-hoc tests within the infant group who heard multiple speakers revealed that discrimination in the "across" condition was not significant. Further, the analyzed electrodes (and thus presumably underlying regions) differed for the within- versus across-speaker condition. Despite these differences, the study by Dehaene-Lambertz and Pena (2001) is taken to support an abstractionist viewpoint, whereas the findings by Jusczyk et al. (1992), frequently cited to support an episodic view, are actually comparable, as we show next.

Jusczyk et al. (1992) tested two-month-olds in a high-amplitude sucking habituation-dishabituation paradigm. In this paradigm, infants are first exposed to the background stimulus contingently with their high-amplitude sucking. In this phase, sucking always results in repetition of the same stimulus list, so, usually, the sucking rate wanes over time. The measure of interest is infants' sucking rate when hearing test tokens, which are either the same as before for controls or new tokens for the experimental group. A difference in sucking rate in the latter compared to the former group indicates that infants dishabituated due to detecting a difference in the stimuli. In Jusczyk et al. (1992), infants in the single-talker condition were habituated with the word "bug" spoken by one voice, and tested with "dug" in the same voice; or they first heard "bug" spoken by 6 different talkers (3 male, 3 female), and were tested with "dug" in the same 6 voices. Infants detected changes regardless of condition. In a follow-up experiment, the authors introduced a 2-minute delay between habituation and test. In this setting, only infants in the single-talker condition detected the phoneme change, whereas infants in the multi-talker condition failed. This failure was replicated when the 6 habituation talkers were all drawn from one gender. Through additional experiments, Jusczyk and colleagues demonstrated that infants detected a phonemic change even in the face of within-talker variation (i.e., using multiple different tokens from the same talker), leading them to conclude that talker variability can be disruptive. However, considering only the first experiment, infants succeeded in the multi-talker condition, and additionally there were no direct statistical comparisons of within- versus across talker-conditions.

Kuhl (1979) trained six-month-olds over several days to react to a change in stimuli with a head-turn, by initially presenting a salient deviating stimulus and letting a toy appear on one side at the same time. Thus, turning the head to the side indicates infants' detection of phonemic changes, which the authors then used to test infants with vowels in the face

of variation of vowel pitch, talker identity, and both, in that order. Only when infants succeeded in detecting changes at a given performance level and within a set number of trials were they exposed to subsequent stages with more variability. All tested infants completed the experiment, including the most variable stage, which included both talker and pitch variation. This finding is frequently taken as evidence for abstraction. A follow-up study by Kuhl (1983) extended these results to different vowels, and Marean et al. (1992) obtained similar results with younger babies. Due to the study design it is not possible to compare within- and across-talker performance directly.

The most recent study on the topic measured dishabituation to a psycho-acoustically salient contrast versus a more subtle one (/p/-/b/ versus /n/-/ŋ/), either in the presence of, or without, talker variability (Clough, 2015). Infants showed a significant difference between habituation and novel stimuli, although in opposite directions across the two phonetic contrasts. As is typical in this paradigm, infants looked longer when hearing the novel stimulus in the /p/-/b/ condition, for the more subtle contrast they looked longer when hearing the habituation stimulus. The second pattern, a so-called familiarity preference, is difficult to interpret, but might indicate greater processing demands (Hunter & Ames, 1988). A direct test of discrimination in within- and across-speaker conditions revealed no difference in performance.

In summary, the evidence of infants' ability to discriminate sound contrasts in the face of speaker variation and change is scarce and has been used to support incompatible standpoints: Either infants abstract over speaker-specific characteristics or not. It becomes crucial to establish how speaker differences impact the acoustic signal to determine how infants will be affected by them in an episodic framework or to better understand which problem they have to solve to achieve abstraction. In this paper, we assess the impact of speaker variation on the acoustic signal, both overall and focusing on those contrasts previously tested on infants.

Experiment

We test whether speaker variation impacts the discriminability of speech sounds in the absence of lexical or phonological knowledge (Schatz et al., 2013, 2014; Martin et al., 2015). Importantly, we do not implement automatized speaker normalization procedures and provide no language-specific information on any level.

Supplementary information, including lists of experimental tokens, raw data, figures, scripts, and results is available at `https://osf.io/mvnjy/`.

Speech Material

The Articulation Index Corpus (Schatz et al., 2015) contains noise-free recordings of all American English phones in diphone pairs (e.g., /ba/, /la/, ...), pronounced by 20 speakers (8 women). The contrast /n/-/ŋ/ cannot be tested, as /ŋ/ does not naturally appear at the onset of syllables in English (Clough, 2015); it thus was not included in our study. Otherwise, our corpus choice is similar to the stimuli typically used in infant studies, as both are recorded under noise-free conditions and based on prompted and isolated instead of spontaneous, connected productions.

Acoustic Representations

We use two common acoustic representations of the speech signal: Mel filterbanks and Mel Frequency Cepstral Coefficients (MFCCs). These representations encode the spectral properties of thin slices of the speech signal, and have been argued to be similar to the first stages of human auditory processing (Gold & Morgan, 2000). In the subsequent reports we focus on Mel filterbanks since the results do not differ substantially for MFCCs.

Discriminability Scores

To quantify how discriminable a contrast is within and across speakers, we compute an ABX discriminability score (Schatz et al., 2013, 2014). This score quantifies how often a test sound X, e.g., ba_1, is correctly identified as member of the same category as A (ba_2) and not B (da_1).

The machine-based discriminability score was used previously by Martin et al. (2015) to assess whether mothers speak more or less clearly to their infants compared to adults, systematically testing the so-called hyperarticulation hypothesis. The scores can be computed automatically over large data sets and allow us to quantify discriminability in a non-parametric way over various sound contrasts using a single metric.

We use the ABXpy package (Schatz, Thiolliere, Synnaeve, & Dupoux, 2016) and follow essentially the same procedure as Martin et al. (2015).[2] To calculate the score the following steps are taken: (1) Encode all available tokens in terms of their acoustic properties (here: Mel filterbanks or MFCCs); (2) Align pairs (either from the same category, e.g., ba_1 and ba_2 or crossing categories, e.g., ba_1 and da_1) via dynamic time warping (using the implementation by Synnaeve, 2016) and compute their distance; here, we use the mean of the inverse cosine deviation from the diagonal (signifying identity); (3) Identify all possible combinations of tokens which can function as A, B, and X, respectively; (4) Compare distances of $A-X$ and $B-X$, where A, X versus B represent a given contrast (e.g., /b/-/d/), and count a success when X was attributed to the correct category because the pairwise distance was smaller for the pair from the same category; (5) Count the successfully categorized triplets and divide by the total number of triplets to get the normalized final score, which ranges between 0 and 1; chance level is .5.

We compute ABX scores in two conditions: Either all three tokens in a triplet stem from the same speaker, or the test token X is sampled from a different speaker than both A and B. To directly tap into the difference of within- versus across-speaker discriminability, we compare absolute

[2] We use mean and not sum difference, which leads to overall more robust performance. Details can be found at `https://osf.io/mvnjy/`.

scores for the two conditions. We further compute a difference score by subtracting across-speaker scores from within-speaker scores. The absolute scores take into account the ease of discriminating each contrast while the difference scores show how much each contrast is affected by speaker variability. To obtain distributional information, we randomly subsample speakers for each contrast 1000 times.

Results

The mean discriminability scores, both within- and across-speaker, for every available contrast in the corpus (a total of 358) are shown in Figure 1. Figure 2 shows the difference scores for the subset of contrasts that we selected on the basis of infant research on this topic (results using both male and female speakers). Table 2 contains the discriminability scores that form the basis for the depicted difference scores. Together, the two figures and the table illustrate the following:

1. There is always an advantage (a higher score,) for within- compared to between- speaker discriminability. The mean difference score for all contrasts is .115 (95% CI: [.087, .142]; score range: .003-.278).
2. A ceiling effect leads to the smallest observed differences (points closest to the diagonal), as exemplified by /a/-/i/.
3. Only very few contrasts drop to chance level (see Figure 1).
4. The speaker variability effect seems to be lower when only including women (see the subsets in Figure 1), but the difference between these groups is reliable (mean across-speaker score .827; mean across-women score .853; mean within-speaker score for both speaker sets .941; 95% CI for paired difference scores on all contrasts: [-.020, .083]).

These general observations hold when changing the acoustic representation to MFCCs, which leads to overall lower absolute scores, leaving the difference score largely unaffected (mean scores within-speaker .876 (vs .941); across-speaker .760 (vs .827); difference score .116 (vs .115)).

Table 2: Summary of results for contrasts tested in infant studies, presented as means and 95% CIs.

	Within-Score	Across-Score	Difference
a-ɔ	.734 [.682, .788]	.553 [.523, .593]	.181 [.143, .218]
p-t	.858 [.828, .888]	.686 [.658, .717]	.173 [.140, .207]
b-p	.813 [.777, .848]	.642 [.623, .664]	.171 [.142, .202]
ɛ-ɪ	.871 [.837, .906]	.710 [.675, 746]	.161 [.125, .194]
b-d	.813 [.781, .848]	.654 [.629, .682]	.159 [.124, .194]
a-i	.999 [.998, 1]	.990 [.984, .996]	.009 [.004, .014]

Discussion

We set out to quantify the impact of speaker variability on sound discriminability. Our results show that there is an overall negative impact of speaker variation on sound discriminability. Considering psycholinguistic models, our findings have implications for both abstractionist and episodic models. For the former, this study specify the set of contrasts that are especially non-invariant, and for which a putative innate

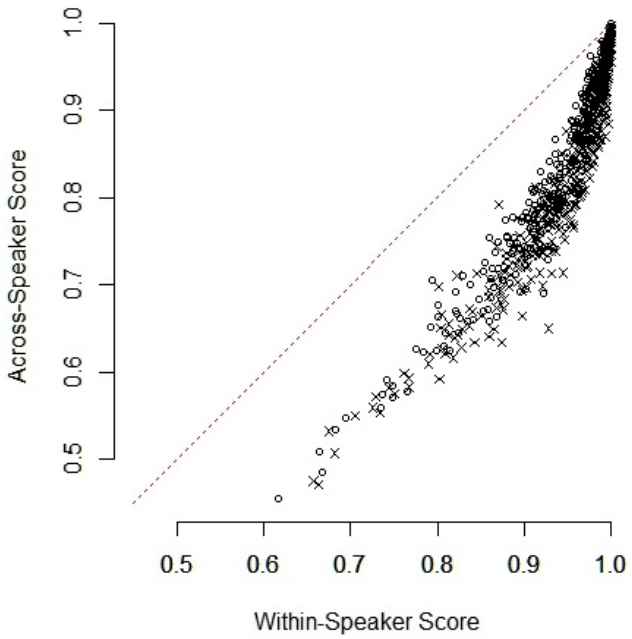


Figure 1: Mean discriminability scores for all possible contrasts in the corpus, showing within- versus across-speaker scores. The **x** mark comparisons across all available speakers, the **o** the results for a subset containing only female speakers. The diagonal (no difference) is indicated by a dotted line.

normalization module has to be specified. For episodic and hybrid models, we make quantitative predictions on the deterioration expected when speaker variability is introduced.

How do our data compare to studies testing whether infants are able to normalize across speakers? The adverse effect of speaker variation we observe is not catastrophic, and can even be seen as modest, as the vast majority of sound contrasts remain discriminable acoustically (better than chance) in the face of talker variation. Furthermore, the impact varies greatly across contrasts, and we observe a ceiling effect. It is important to reflect that this could potentially undermine the possibility of empirically testing episodic models, as follows. A frequent prediction, at least in infant literature, seems to be that episodic models should yield a complete absence of discrimination abilities as soon as talker variation is introduced (Dehaene-Lambertz & Pena, 2001). However, our results show an overall adverse effect of speaker variation. It should be noted that infants' abilities might not be reflected in their performance in the indirect measures described in detail above, although the two are frequently equated (e.g., Apfelbaum & McMurray, 2011; Bergmann, ten Bosch, Fikkert, & Boves, 2013).

In addition, as the size of the adverse effect depends strongly on the chosen contrast, it is possible to observe no difference at all. In fact, the first study supporting early normalization in infants (Kuhl, 1979) chose /a/-/i/, a contrast that is, put simply, always easy. The follow-up study with

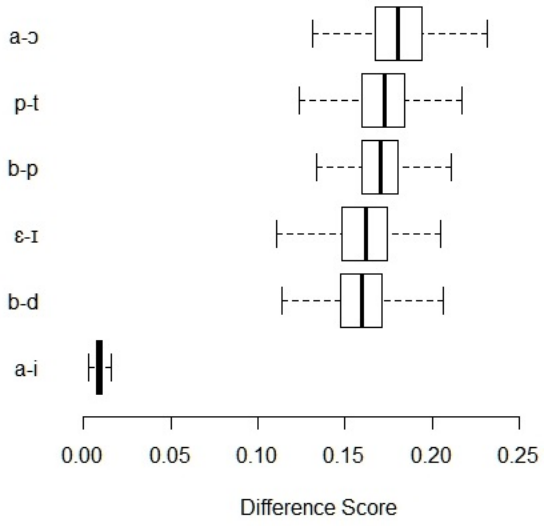


Figure 2: Differences of mean discriminability scores within and across speaker, for the six contrasts tested on infants. A positive difference score indicates that the performance is better within than across speaker.

/a/-/ɔ/ (Kuhl, 1983), the only contrast in the infant literature that approaches (but is above) chance level, trained infants over several days in increasingly variable and difficult conditions. This training might explain continued success with added speaker variability, an advantage our evaluation did not implement.

Clough (2015) tested /b/-/p/ and found a numerical difference of within- versus across-speaker performance, which can be taken as consistent with our finding that this contrast is strongly affected by speaker variation. There was, however, no statistically significant effect, which the author herself attributes to insufficient statistical power. We believe that the same low power may also explain the lack of a significant interaction in the study by Dehaene-Lambertz and Pena (2001). Although this is difficult to determine (Gelman & Stern, 2006), we repeat that the discrimination response was significant in the within-speaker condition, and not in the across-speaker condition.

The observation that non-significant results might be grounded in a lack of power, especially when testing interactions, makes it necessary to discuss the sensitivity of infant measures. As a recent meta-analysis showed, there is substantial variability (I^2=76.87%) in infants' native vowel discrimination performance and only a medium-sized overall effect (Cohen's d=.6, SE=.05; data retrieved from `metalab.stanford.edu`, consulting the dataset of Tsuji & Cristia, 2014). This leads to frequently underpowered studies on the topic, a problem only worsened when testing an interaction with a small effect (within- or across-speaker discrimination). In other words, since speaker variation has a consistent, but moderate effect (we cannot expect a complete breakdown of discrimination abilities across speakers), the impact on infant behavior might be difficult to measure and/or require a(n unfeasibly) high number of participants.

We suggest, nonetheless, to test infants' developing discrimination abilities in the face of speaker variability by comparing two contrasts that are matched on within-speaker discriminability, but that differ maximally in the across-speaker task, as measured by the difference score. The difference between these two scores would provide a measure of how much infants are sensitive to specific difficulties introduced by speaker change, and would therefore provide us with the required quantitative evaluation of episodic versus abstractionist models.

Extrapolating from our data to language acquisition outside the laboratory, our results suggest that infants' input becomes more difficult to learn from when talker variability is present. This holds from both theoretical viewpoints, as long as abstraction is not innate, and has to be (at least partially) acquired. Given that infants tune into their native language based on the acoustic signal, being exposed to higher input variability in the form of more speakers leads to a more difficult learning problem.

The present results were based on a corpus that was both maximally exhaustive and recorded under ideal conditions, much like the stimuli infants are typically confronted with in the lab. Follow-up experiments will address how our results generalize to corpora of infant-directed speech (IDS), which are often not available in sufficiently high quality. It remains an open question, and one orthogonal to the issue of hypo- or hyperarticulation in IDS (Martin et al., 2015), whether the acoustic markers, such as higher, but also more variable, pitch, emphasize or lessen speaker differences. We may thus in an extension of the present work quantify the learning problem infants face in real life when confronted with multiple talkers on a daily basis.

The present work has implications for adult models of speech processing, as it maps out the extent of the problem of introducing speaker variation, a task adults need to solve as well, albeit with more knowledge and experience. Here, too, predictions from abstractionist models have to be carefully examined. To better test whether or not listeners generate a speaker invariant representation it is not correct to expect chance performance when introducing multiple speakers; for some contrasts we predict no adverse effects for a system without any normalization and language knowledge. Thus, paradigms sensitive to decreased performances and a contrast chosen to maximize the predicted difference can better tap into the question of how adults process speaker variance.

In sum, we have shown that speaker variation poses a hurdle for sound discrimination but does not necessarily lead to confusion, even for a naive learner. In quantifying the effect, we can derive more precise predictions when testing competing psycholinguistic models.

Acknowledgements

We thank T. Schatz, X.-N. Cao, R. Thiolliere, M. Bernard, and G. Synnaeve for their help. The present work was supported by the European Horizon 2020 programme (Marie Skłodowska-Curie grant No 660911), the European Research Council (E-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-14-CE30-0003 MechELex, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC), and the Fondation de France.

References

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138.

Bergmann, C., ten Bosch, L., Fikkert, P., & Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Frontiers in Psychology*, *4*(676).

Clough, L. T. (2015). *Does variability impact infants' sound discrimination?* (Unpublished Thesis, accessed via `http://hdl.handle.net/10150/578957`)

Cristia, A. (2011). Fine-grained variation in caregivers' /s/ predicts their infants' /s/ category. *The Journal of the Acoustical Society of America*, *129*(5), 3271–3280.

Dehaene-Lambertz, G., & Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport*, *12*(14), 3155–3158.

Escudero, P., Bonn, C. D., Aslin, R. N., & Mulak, K. E. (2015). Indexical and linguistic processing in infancy: Discrimination of speaker, accent and vowel differences. In *Proceedings of the International Congress of Phonetic Sciences.*

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Gold, B., & Morgan, N. (2000). Chapter 14: Ear Physiology. In *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (pp. 189–203). New York: J Wiley.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, 69–95.

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. (1992). Some consequences of stimulus variability on speech processing by 2-month-old infants. *Cognition*, *43*(3), 253–291.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*(6), 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, *6*(2), 263–285.

Marean, G. C., Werner, L. A., & Kuhl, P. K. (1992). Vowel categorization by very young infants. *Developmental Psychology*, *28*(3), 396–405.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347.

Polka, L., Masapollo, M., & Ménard, L. (2014). Whos talking now? infants perception of vowels with infant vocal properties. *Psychological Science*, *25*(7), 1448–1456.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Schatz, T., Cao, X.-N., Kolesnikova, A., Bergvelt, T., Wright, J., & Dupoux, E. (2015). *Articulation Index lscp* (Speech Corpus No. LDC2015S12). Linguistic Data Consortium.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *14th Annual Conference of the International Speech Communication Association.*

Schatz, T., Peddinti, V., Cao, X.-N., Bach, F., Hermansky, H., & Dupoux, E. (2014). Evaluating speech features with the minimal-pair ABX task (II): Resistance to noise. In *15th Annual Conference of the International Speech Communication Association.*

Schatz, T., Thiolliere, R., Synnaeve, G., & Dupoux, E. (2016). *ABXpy v0.2.* Retrieved from `http://dx.doi.org/10.5281/zenodo.45268`

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker variation aids young infants phonotactic learning. *Language Learning and Development*, *10*(4), 297–307.

Synnaeve, G. (2016). *DTW_Cython.* Retrieved from `http://dx.doi.org/10.5281/zenodo.45257`

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.

A Twist On Event Processing: Reorganizing Attention to Cope with Novelty in Dynamic Activity Sequences

Jessica E. Kosie (jkosie@uoregon.edu)
Department of Psychology
University of Oregon, Eugene, OR 97403, USA

Dare A. Baldwin (baldwin@uoregon.edu)
Department of Psychology
University of Oregon, Eugene, OR 97403, USA

Abstract

Fluent event processing appears to critically involve selectively attending to information-rich junctures within continuously unfolding sensory streams (e.g., Newtson, 1973). What counts as information-rich would seem to depend on a variety of factors, however, including the novelty/familiarity of such events, as well as local opportunity for repeated viewings. Using Hard, Recchia, & Tversky's "Dwell-time Paradigm," we investigated the extent to which viewers' attention to unfolding activity streams is affected by novelty/familiarity and a second viewing. Viewers' dwell times were recorded as they advanced twice each through three slideshows varying in familiarity but equated on other dimensions. Dwell time patterns revealed reorganization on a number of fronts: a) familiarity elicited decreased dwelling overall, b) dwell-time patterns changed systematically on second viewing, and c) familiarity modulated the specific nature of change associated with repeated viewing. These findings illuminate reorganization in attention as action information is first encountered and then quickly incorporated to guide event processing.

Keywords: action segmentation; event processing

Skill at making sense of human action is essential to normal adult functioning. Imagine, for example, an everyday event such as cooking together with someone else. Among other things, this seemingly mundane activity requires fluently interpreting one's partner's ongoing activity and the causal effect it has on the world, seamlessly integrating all of this with an unfolding interpretation of any linguistic contributions that co-occur, and coordinating one's own activity stream to mesh with an understanding of what the other is dynamically enacting or plans to enact. Although the complexity of the processing involved is immense, action processing at the adult level appears to be so fluent that errors are rare and processing occurs largely outside of conscious awareness.

Much remains mysterious about the mechanisms subserving everyday event processing. In recent years, however, considerable progress has been made in relation to one particular component of event processing: segmentation. Event segmentation involves transforming continuously unfolding sensory information into discrete units that can be remembered, categorized, and described with language.

In principle, there are infinite possible ways to segment any sensory stream; it is therefore striking that observers asked to identify event boundaries within continuously flowing activity display considerable agreement, and this agreement holds at multiple levels of event representation. The event boundaries viewers nominate typically coincide with goal transitions, and viewers readily scale their segmentation up or down, as the task demands, in terms of the grain at which they identify event boundaries (e.g., Newtson, 1973; Zacks, Tversky, & Iyer, 2001). For example, at a coarse-grain level, viewers tend to nominate transitions between higher-level tasks (e.g., in a food preparation sequence, the transition between an actor closing an oven door and turning to wash some dirty dishes). At a more fine-grain level, viewers identify smaller-scale events within these tasks (e.g., lifting a cookie pan, placing the cookie pan in the oven, grasping the oven door, closing the oven door). Viewers' judgments at these different levels of generality display alignment, such that coarse-gain boundaries tend to align with boundaries at the fine-grain level, suggesting that viewers organize event segments hierarchically in their processing (Zacks & Swallow, 2007). Adults' segmentation is relatively robust to at least minor novelty, as well.

Several sources of evidence indicate that segmentation occurs spontaneously as viewers process unfolding sensory streams, and may be largely accomplished via implicit mechanisms. For one, magnetic resonance recordings conducted during passive event viewing display neurophysiological activity that selectively correlates with segment boundary judgments participants provided on a subsequent, second viewing of the behavior stream (e.g., Zacks et al., 2001). As well, action boundaries implicitly "intrude" on participants' processing as they carry out unrelated tasks, such as detection of an unrelated stimulus (e.g., Huff, Papenmier, & Zacks, 2012). For example, reaction times reveal that participants process the segmental structure of unfolding behavior while they engage in a change detection task.

Hard, Recchia, and Tversky (2011) recently introduced another valuable implicit measure of action segmentation: the "dwell-time paradigm." In this task, slideshows are constructed by selecting frames at a regular increment from

digitized event videos (e.g., one frame every 500 msec). Viewers advance through slideshows at their own pace by the click of a mouse; latencies between mouse clicks index their "dwell time" to each slide. When dwell times are referenced to event boundary judgments (subsequently provided either by slide show viewers themselves, or by other viewers), systematic dwell-time patterns emerge: a) viewers tend to dwell longer on slides coinciding with event boundaries relative to points occurring within event units (henceforth termed a "boundary advantage"), and b) dwelling is longest to slides coinciding with coarse-grain level boundaries relative to more fine-grain level boundaries. Importantly, dwell-times arising when adults view the same set of slides in scrambled order do not reveal such patterns to the same degree. Thus dwell-time patterns appear to simultaneously reflect both the extraction of segmental structure as viewers process an unfolding activity sequence, and the hierarchical organization of the segments they extract. Also noteworthy is that viewers' dwell times to coarse-grain level slides positively predicted their subsequent recall score for the activity sequences they viewed. The fact that dwell times make contact with event memory helps to validate dwell times as a psychologically meaningful measure of action processing.

One might question why dwell times tend to increase at event boundaries. Zacks and colleagues' (Kurby & Zacks, 2008; Zacks, Kurby, Eisenberg, & Haroutunian, 2011) Event Segmentation Theory provides a natural account for the phenomenon. They suggest that event processing is fundamentally a process of predictability monitoring, with information value being highest where predictability is low. Event boundaries appear to represent just such information-rich predictability "troughs" within activity streams. As one event comes to a close, uncertainty abounds regarding what will occur next. Once the subsequent event is underway, predictability surges, only to plummet again as that event approaches completion. Imagine, for example, one detects another initiating a reach for a coffee mug. Motion in the midst of this reach is highly predictable (the hand predictably configures to prepare to grasp the mug as it approaches the mug), but once the reach is completed, predictability suddenly plummets: many subsequent alternatives for action arise. The actor might lift the mug, push it away, spin it around, or simply hang on to it for a time. At such boundaries, predictability is low and viewers gain highly relevant information to interpret the motion stream. Thus selectively attending to event boundaries – regions of "predictable unpredictability" – conceivably facilitates event processing.

Viewers appear to capitalize on a broad range of clues to identify segment boundaries whether explicitly reporting on the location of event boundaries or using more implicit methods such as the dwell-time paradigm. These clues include expectations/inferences about goals and intentions, as well as physical motion parameters (e.g., motion change, velocity change, motion acceleration/deceleration, and change in motion directory) (e.g., Zacks, 2004; Zacks, Kumar, Abrams, & Mehta, 2009). Statistical regularities also inform judgments about segment boundaries (Baldwin, Andersson, Saffran, & Meyer, 2008). The sensory stream possesses an inherent predictability structure; some phenomena tend to succeed one another with greater regularity than others. For example, within everyday intentional activity, the act of chopping a vegetable would be more likely to follow the act of grasping a knife than would the act of opening a refrigerator door. When predictability is high, small-scale acts cohere into larger event units. Conversely, junctures at which predictability is low reflect event boundaries. Put another way, growing familiarity with predictability structure produces increasing access to the segmental structure of activity streams. One implication of this account is that viewers' processing will change as familiarity with the predictability structure of the sensory stream grows.

Meyer, Hard, and Baldwin (in preparation) recently documented that dwell-times are sensitive to reorganization in attentional patterns that accrue as familiarity with segment-related predictability structure arises during statistical learning. One group of adults (*knowledgeable* viewers) viewed a digitized video depicting a novel activity sequence with underlying statistical regularities (i.e., a ten-minute "exposure corpus" with four randomly combined "actions" – each composed of a fixed triad of small-motion elements, such as *feel-blow-look*), after which they advanced at their own pace through slides extracted once every 500 msec from the same exposure corpus. A second group (*naïve* viewers) provided dwell times to the same slideshow without having had any prior opportunity to acquire the statistical regularities through viewing the exposure corpus video. After slideshow viewing, both groups provided discrimination judgments for pairs of short videos, with each pair depicting one "action" (a statistically regular triad from the exposure corpus) and one "part-action" (a sequence actually viewed, but spanning a low-predictability transition from the exposure corpus).

Several noteworthy findings emerged. First, discrimination judgments revealed better-than-chance discrimination accuracy for "action" segments relative to "part-action" foils, but only in the *knowledgeable* condition, in which viewers had watched the digitized video of the exposure corpus prior to viewing the dwell-time slideshow. Thus, as expected, only those who received extended exposure to the statistical regularities within the novel activity sequence appeared to have achieved a solid knowledge of the segmental structure of the activity stream overall. Second, dwell-time patterns differed for those in the *knowledgeable* versus *naïve* conditions. In particular, *knowledgeable* viewers displayed the predicted action-level boundary advantage in dwell times indicative of sensitivity to the higher-level segmental structure defined by the statistical regularities inherent in the exposure corpus. In contrast, no comparable boundary advantage emerged for dwell-times in *naïve* viewers, who

had lacked prior opportunity to acquire the statistically-defined segmental structure of the exposure corpus. Lastly, only those in the *knowledgeable* condition who displayed strong discrimination performance revealed the segment-related boundary advantage in dwell-times.

All in all, dwell-time patterns in this experiment clearly revealed learning-based reorganization in the way that viewers modulated their attention during the unfolding activity stream. These findings provided the first evidence to date documenting patterns of attentional reorganization specifically attributable to newly-acquired familiarity with segment-related predictability structure. In particular, viewers who had the opportunity to learn the predictability structure of the sequence displayed longer dwell times at junctures within the unfolding activity sequence where they came to be able to predict that predictability was low (i.e., between the end of one statistically coherent unit of action and the beginning of another). This experiment also showcased a striking degree of alignment across discrimination and dwell-time methodologies; that is, viewers who displayed newly-acquired knowledge of segmental structure in the explicit discrimination judgment task also displayed sensitivity to that same structure in their implicit dwell-time patterns. These findings provide a strong basis for confidence in the dwell-time paradigm as a valid and sensitive index of familiarity-driven reorganization in action processing.

The statistical learning paradigm utilized in the research just described has the advantage of offering complete control of the predictability structure of the activity stream. At the same time, of course, the activity streams utilized in such research by necessity are somewhat unnatural, undercutting generalization of findings to event processing in the world at large. Thus, it is important to extend this research by investigating how familiarity alters action processing in more naturalistic activity sequences. We have identified one such sequence -- tying shoe laces -- that offers considerable potential for exploring these issues while maintaining necessary methodological controls. Skill at tying shoelaces is challenging to acquire; children between 5 and 8 years of age typically require many months to grasp its fundamentals and achieve an outcome that is at least moderately effective. This may be in part because the necessary motion elements for causal success are difficult for novices to extract from the motion stream through observation alone. Although shoelace tying comes to seem trivial to the expert (i.e., the average adult), producing successful shoelace tying in fact requires not only manual dexterity, but also knowledge of causal mechanics and fluent sequencing of coordinated motions necessary to effect successful tying. Shoelace tying can be accomplished via alternative methods (in regard to which viewers differ in their familiarity). This fact presents a unique opportunity to investigate the effects of viewer familiarity on the processing of the complex motion stream that shoe-tying generates.

In the current study, we investigated differences in adults' processing of shoelace tying sequences that may arise in conjunction with increasing familiarity (as when those who understand the causal mechanics are introduced to a new method). There are at least three methods for tying shoe laces that achieve the same intentional outcome: *loop*, in which a loop is created, another lace is wrapped around the loop and a second loop is pulled through; *ears*, in which two loops are created and tied together; and *twist*, in which both laces are simultaneously manipulated to capture loops from opposing fingers. In North America, adults tend to be familiar with both the *loop* and *ears* methods, but prefer one of these methods in their own tying (thus they have highest familiarity with one particular method). In contrast, most North American adults have little to no familiarity with the *twist* method.

We predicted that, overall, methods that were less familiar would elicit overall longer dwell times, and that this would hold regardless of the specific details of the tying method involved. We also expected to replicate the "boundary advantage effect" found in previous dwell-time research; specifically, that dwell times to boundary slides would be significantly longer than to those for slides depicting within-unit content. However, we predicted that familiarity might influence observers' ability to detect and increase attention to event boundaries. Thus, a boundary advantage might differ systematically for more familiar versus more novel event sequences.

An additional advantage of shoelace tying sequences is that they allow for naturalistic repetition within the unfolding event stream. That is, we were able to explore dwell time patterns to the unfolding event once on a first shoe and again for a second shoe. We predicted that dwell-time patterns would differ from the first viewing (first shoe) to the second viewing (second shoe) as participants gained experience with a given shoelace-tying method. Thus, the above-mentioned patterns were predicted to vary across first and second viewings.

Method

Stimuli

Videos of the three methods of shoe tying were filmed on a camera at a rate of 30fps. Videos of the *loop, ears,* and *twist* methods were equal in length (each 139 seconds). Only the actor's shoes, socks, pant legs, and hands were visible in the videos. To increase consistency across the slide shows derived from the videos of the three methods of shoe tying, the lead-in and completion portions of all three slide shows were in fact identical. The slide shows differed only in their middle portion, which contained the actions that were distinctive to each method. Consistent with prior research, slideshows were created by extracting one frame every second from each of the three videos (e.g., Hard, Recchia, and Tversky, 2011; Meyer, Baldwin, & Sage, 2011). The

three resulting slideshows thus contained an identical number of frames (N = 139). Sections of slides were classified as causally distinctive (e.g., differing across the three slideshows and unique to the method being demonstrated) versus non-distinctive (e.g., depicting actions that did not differ across the three methods, such as tying a double knot). Two expert coders further classified slides as depicting boundaries or within-unit action. Examples of images depicting boundary and within-unit slides from each of the three slideshows are depicted in Figure 1.

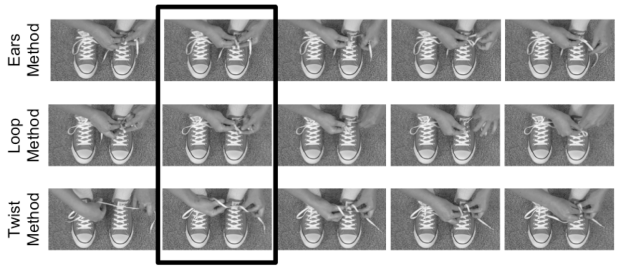


Figure 1: Example boundary and within-unit slides from slideshows depicting the *ears*, *loop*, and *twist* methods. Outline indicates boundary slides in all three slideshows. All other slides depicted were categorized as within-unit.

Participants and Procedure

142 undergraduates (61% female, M_{age} = 20 years) participated for course credit; data from three participants were excluded from analyses due to experimenter error (2) and illness. After a brief training phase to familiarize participants with the self-paced slideshow format, participants advanced at their own pace through three slide shows (the three different shoelace tying methods, in counterbalanced order across participants). Participants' dwell times, or latency between mouse clicks from one slide to the next, were recorded using Psychtoolbox (Brainard, 1997). Finally, participants were asked to rate their familiarity with the three methods of shoe tying and to demonstrate the method that they used every day.

Results

Dwell times greater than three standard deviations from the overall group mean were considered outliers. Participants' data were excluded from analyses when more than 10% of their dwell times were outliers, resulting in the exclusion of eight participants. Data from the remaining 131 participants were positively skewed, as is typical for looking time data; a log10 transformation was utilized to normalize the distribution prior to analyses.

The shoelace tying method that participants used every day (assessed by the shoe-tying demonstration task) was classified as their most familiar method. For 75.9% of participants this was the *loop* method and for 21.2% participants this was the *ears* method. Only 2.9% of participants identified the *twist* method as most familiar; their data were excluded from further analysis due to the small sample size. For purposes of analysis, participant's other highly-rated method (*loop* or *ears* depending on participants' most familiar method) was classified as *moderate familiarity* for each participant. For all participants included in analyses, the *twist* method was classified as *low familiarity*. Thus, even though participants' *high familiarity* method was the method most familiar to them, the details of the specific method differed across subjects (i.e., for some the method of highest familiarity was *loop*, and for others, it was *ears*).

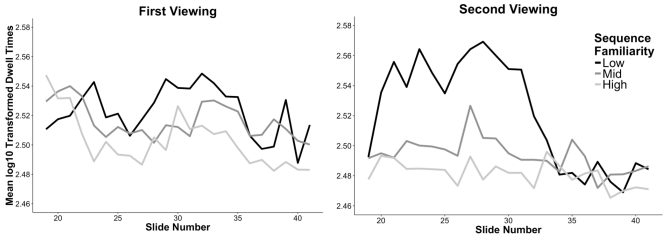
Our first analysis examined whether viewers dwelt longer to the degree that events were less familiar, and whether this pattern held up regardless of their own preferred method of tying (and thus the details of the specific method that was high- versus mid-range in familiarity). Log10-transformed dwell times (henceforth, simply "dwell times") were examined via a 3 (familiarity: high, mid, low) X 2 (preferred method: *loop* vs. *ears*) mixed between-within ANOVA with familiarity as the within-subjects variable and preferred method as a between-subjects variable. We found a significant main effect of familiarity, $F(2, 258) = 5.17$, $p = .006$, $\eta_p^2 = .04$, which, as predicted, displayed a significant linear trend, $F(1, 129) = 9.65$, $p = .002$, $\eta_p^2 = .04$ ($M_{HighFam}$ = 2.50, $SEM = .02$; $M_{MidFam} = 2.52$, $SEM = .02$; $M_{LowFam} = 2.54$, $SEM = .02$). The main effect of preferred method did not emerge as significant, $F(1, 129) = 0.64$, $p = .426$, $\eta_p^2 = .005$, nor did it interact with familiarity, $F(2, 258) = 1.12$, $p = .327$, $\eta_p^2 = .009$. That is, the linear trend (lowest dwell times to participants' most familiar method and highest to their least familiar method) held regardless of participants' preferred method for tying shoes and was robust to differences in the details of which method was classified as low versus mid familiarity. This linear trend of familiarity, combined with a lack of interaction involving familiarity and preferred method, confirms that dwell-times are responsive to familiarity, regardless of the specifics of the particular shoelace tying method at issue, which helps to further validate dwell times as a meaningful index of viewers' processing that is at least to some degree independent of specific motion patterns.

In question with the second analysis was the extent to which viewers displayed increased dwelling to slides depicting event boundaries, relative to slides displaying within-segment content, a basic question regarding replication of prior findings. Also of interest in this analysis was whether a "boundary advantage" was affected by a) the familiarity of shoe-tying methods, and/or b) first versus second viewing of a given method. A 2 (slide type: boundary vs. within) X 3 (familiarity: high, mid, low) X 2 (viewing: first vs. second) repeated-measures ANOVA was conducted to examine these effects. The main effect of slide type was significant, $F(1, 130) = 94.56$, $p < .001$, $\eta_p^2 = .42$, with dwelling significantly longer to boundary slides ($M = 2.52$, $SEM = .016$) than within-unit slides ($M = 2.51$, $SEM = .016$),

replicating prior findings using the dwell time methodology.

The main effect of familiarity was also significant, $F(2, 260) = 5.85$, $p = .003$, $\eta_p^2 = .04$, with a significant linear trend, $F(1, 130) = 11.20$, $p = .001$, $\eta_p^2 = .04$, reflecting increased dwelling for methods that were less familiar, further confirming what was reported in the first analysis. Participants additionally displayed longer dwell times on first ($M = 2.53$, $SEM = .016$) versus second viewing ($M = 2.50$, $SEM = .016$), $F(1, 130) = 40.82$, $p < .001$, $\eta_p^2 = .24$, also indicative of dwelling reducing with increasing familiarity. We found significant two-way interactions between familiarity and slide type and between viewing and slide type, $ps < .05$. However, we will leave these aside as they seemed best interpreted in the context of a nearly significant three-way interaction between slide type, familiarity, and viewing, $F(2, 260) = 2.90$, $p = .06$, $\eta_p^2 = .02$. Simple effects tests revealed that, on first viewing, dwell times to boundary slides were significantly greater than dwell times to within-unit slides across all levels of familiarity, all $ps < .001$. On the second viewing, however, dwell times to boundary slides were greater than dwell times to within-unit slides only for the least familiar method, $p < .001$. For slideshows depicting shoe tying methods at low and moderate levels of familiarity, there was no significant difference in dwell times to boundary versus within-unit slides on the second viewing, $ps > 0.18$. These findings point to the boundary advantage being responsive to the familiarity of the event sequence depicted.

A final exploratory analysis focused specifically on reorganization in dwell times for the least familiar *twist* method across viewings. As Figures 2 and 3 depict, a notable surge in dwell times seemed to occur specifically for the second viewing of the *twist* method at the most causally distinctive portion of the shoe-tying sequence. Viewers seemed to target the causally distinctive portion of the *twist* slide show with sustained enhanced attention during their second viewing. This analysis examined the degree to which this was a systematic attentional increase from first to second viewing, and whether viewing in relation to slide type (boundary versus within slides) was affected equivalently by this attentional enhancement. A 2 (slide type: boundary vs. within) X 2 (distinctiveness: causally distinctive vs. not causally distinctive) X 2 (viewing: first vs. second) repeated-measures ANOVA examined these effects for only the lowest familiarity *twist* method of shoe tying. As expected, dwell times were significantly longer to boundary slides ($M = 2.54$, $SEM = .02$) than within-unit slides ($M = 2.52$, $SEM = .02$), $F(1, 130) = 52.63$, $p < .001$, $\eta_p^2 = .29$, and to distinctive portions of the slideshow ($M = 2.54$, $SEM = .02$) than non-distinctive portions ($M = 2.52$, $SEM = .02$), $F(1, 130) = 5.45$, $p = .02$, $\eta_p^2 = .04$. Dwell times were also marginally significantly longer on first ($M = 2.54$, $SEM = .02$) than second viewing ($M = 2.52$, $SEM = .02$), $F(1, 130) = 3.25$, $p = .07$, $\eta_p^2 = .02$. Additionally, all two-way interactions were significant or marginally significant: the slide type X distinctiveness and distinctiveness X viewing interactions were both significant ($ps < .001$) and the slide type X viewing interaction was marginally significant ($p = .08$). However, these effects must be interpreted in the context of a significant three-way interaction between slide type, distinctiveness, and viewing, $F(1, 130) = 39.76$, $p < .001$, $\eta_p^2 = .23$. This significant three-way interaction suggested that slide types were affected differentially by attentional enhancement to the causally distinctive portion of the *twist* slideshow. A boundary advantage was observed only for non-distinctive portions of the slideshow on first viewing ($p < .001$). However, by the second viewing of the least familiar *twist* sequence, participants' dwell times were higher to boundary slides than within-unit slides for the distinctive as well as the non-distinctive portions of the slideshow ($ps < .001$).



Figures 2 and 3: Dwell-time patterns across low, mid, and high familiarity slideshows on first (Figure 2) and second (Figure 3) viewings.

Discussion

To recap briefly, we investigated the extent to which adults' processing of shoelace tying sequences reorganized with respect to familiarity with a given method of tying shoe laces, and over repeated viewings of that method. We also examined the extent to which enhanced attention to event boundaries might be preserved over the course of such attentional reorganization. Overall, less familiar methods of shoe tying elicited longer average dwell times per slide, and this was so regardless of the specific details of the tying method involved. Replicating prior research, viewers displayed enhanced attention to boundary slides relative to slides depicting within-segment content. However, while this boundary advantage was present on first viewing across all three methods it disappeared on the second viewing for all but the least familiar, *twist*, method. Moreover, analyses focusing specifically on the least familiar *twist* method revealed further striking evidence for attentional reorganization. For the *twist* method, a boundary advantage was observed for non-distinctive portions of the activity stream (depicting activity familiar to viewers and similar across all methods of shoe tying) on both viewings. In contrast, for the causally distinctive portion of the shoe tying sequence (depicting activity unique to the *twist* method), a boundary advantage emerged only on the second viewing. Thus viewers' identification of boundaries within the *twist*

activity stream underwent reorganization after just a single exposure to this unfamiliar shoe tying method.

Of particular interest was the impact of repeated viewing on observers' enhanced attention to events. For shoe tying sequences at high and moderate levels of familiarity, a boundary advantage was present on first viewing and markedly absent on second viewing. By the second viewing, these two relatively familiar action streams no longer contained predictability troughs -- regions of "predictable unpredictability" -- and the boundary advantage effect was reduced. For the least familiar method, in contrast, participants continued to display selective attention to event boundaries on second viewing, suggesting that just one viewing hadn't yet rendered the activity stream familiar and predictable. Further examination revealed that, though the overall boundary advantage remained, attention to event boundaries underwent reorganization across repeated viewings within this unfamiliar method.

Our final analysis, in particular, implicated enhanced targeting of predictable unpredictability on second viewing within the causally distinctive portion of the highly unfamiliar *twist* method The "predictable unpredictability" hypothesis suggests that observers must be able to identify (predict) when unpredictability will occur in order to selectively attend to these junctures. While viewers did not selectively target unpredictability (event boundaries) with increased attention in the causally distinctive portion of *twist* tying on first viewing, just the that single first viewing provided enough exposure to allow them to identify these maximally informative regions as early as the second viewing of this highly unfamiliar event. These findings showcase how rapidly viewers reorganize attention as new action information is first encountered and then incorporated to guide event processing on subsequent viewings.

Our findings also offer incentive to extend related questions to research with children, as shoelace tying is a skill that is typically acquired at approximately five to eight years of age. Additionally, the dwell-time paradigm has been validated for use with preschoolers (Meyer, Baldwin, & Sage, 2011), suggesting that the current study might be amenable to use with children who are at varying levels of shoe-tying ability. Questions of interest include: How might reorganization of streaming sensory information differ for those who are at the cusp of acquiring this new skill? Would variation in children's executive function affect their ability to effectively modulate attention to the unfolding event? How might direct motor experience reshape children's processing of shoe-tying sequences as they are in the midst of acquiring this basic motor skill? These and similar questions can inform research on the acquisition of event processing fluency in childhood.

The findings we report here further validate the usefulness of the dwell-time paradigm for examining how viewers deploy attention as they process dynamically unfolding sensory streams. More specifically, these findings showcase how attention readily reorganizes to cope with novelty as dynamic activity sequences unfold across time. It will be important to extend this methodology to other naturalistic event sequences varying in familiarity; this effort is underway. All in all, the current findings provide altogether new information about the timing and nature of changes to processing as novel event information is encountered.

Acknowledgements

The authors would like to thank Alan L. Baldwin, Rose Maier, Jenny Mendoza, Robbie Ross, and Jason Wallin for valuable conceptual input regarding this research.

References

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition, 106,* 1382-1407.

Brainard, D.H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433-436.

Hard, B., Recchia, G, & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General, 140,* 586-604.

Huff, M., Papenmeier, F., & Zacks, J. M. (2012). Visual target detection is impaired at event boundaries. *Visual Cognition,* 1-17.

Kurby, C. A., & Zacks, J. M.(2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12,* 72–79.

Meyer, M., Hard, B., & Baldwin, D. (in prep). *Attention reorganizes as structure is detected in dynamic human action.* Unpublished manuscript, University of Oregon.

Meyer, M., Baldwin, D. A., & Sage, K. (2011). Assessing young children's hierarchical action segmentation. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology, 28,* 28–38.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28,* 979-1008.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movements and intentions to understand human activity. *Cognition, 112,* 201–216.

Zacks, J. M., Kurby, C. A., Eisenberg, M. I., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience, 23 (12)*, 4057-4066.

Zacks, J. M., & Swallow, K. M. (2007). Event segmentation, *Current Directions in Psychological Science*, *16(2),* 80-84.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130,* 29–58.

Recursion in Nicaraguan Sign Language

Annemarie Kocab (kocab@fas.harvard.edu)[1], Ann Senghas (asenghas@barnard.edu)[2], Jesse Snedeker (snedeker@wjh.harvard.edu)[1]

[1]Department of Psychology, 33 Kirkland Street, Cambridge, MA 02138 USA
[2]Department of Psychology, 1009 Broadway, New York, NY 10027 USA

Abstract

Syntactic recursion is argued to be a key property of natural languages, allowing us to create an infinite number of utterances from a finite number of words and rules. Some have argued that recursion is *uniquely* human. There are at least two possibilities for the origins of recursion: 1) Recursion is a property of the language faculty. 2) Recursion is an historical accomplishment and is culturally constructed over millennia. Here we ask whether an emerging sign language, Nicaraguan Sign Language (NSL), exhibits syntactic recursion by comparing the language of the first three age cohorts of signers. Signers (n=27) watched and described vignettes designed to elicit relative clauses. Results suggest that signers from all three cohorts have strategies to fulfill the discourse function of relative clauses, picking out an individual from a set. The grammatical form of the utterances differs across cohorts, with signers from later cohorts clearly producing embedded structures.

Keywords: language emergence; syntactic recursion

Introduction

Recursion, defined as "a procedure that calls itself, or… a constituent that contains a constituent of the same kind" (Pinker & Jackendoff, 2005), is argued to be a key property of natural languages, allowing us to create an infinite number of utterances from a finite number of words and rules[1]. For instance, we can generate sentences like "Sally said that the boy cried" from "The boy cried," and "Billy thinks that Sally said that the boy cried" from "Sally said that the boy cried," and so on. Some theorists have argued that recursion is *uniquely* human and might even be the sole property that sets human language apart from animal communication systems (Hauser, Chomsky, Fitch, 2002). However, at least one language, Pirahã, has been reported to lack syntactic recursion (Everett, 2005), raising doubts about this claim. It remains an open question as to whether recursion is indeed a hallmark of human language and if so, where the capacity to implement rules within rules comes from.[2]

A central goal in cognitive science has been the identification of what uniquely human properties account for our language system. One possibility is that language arises from our mental architecture, and the capacity to create language is present is every human mind. The language faculty may be a part of our genetic endowment (e.g., Chomsky, 1968; 2000; Pinker, 1994) or language may be a product of general changes in our conceptual resources and computational abilities (e.g., Christiansen & Charter, 2008). The properties and organization of language then reflect the structure of human cognition (Chomsky, 1975; Pinker, 2007). A second possibility is that language developed gradually over historical time, rather than phylogenetic time, through a process of "cumulative cultural evolution" (Tomasello, 2011; Tomasello, Kruger, & Ratner, 1993). Here, language is a side effect of the human capacity for social learning and cultural transmission (e.g., Tomasello, 2008).

Recently, the study of emerging sign languages, such as Nicaraguan Sign Language (NSL), has offered the opportunity to better understand how languages are created and the roles of the individual learner and the community of users. NSL is a new language created by a deaf community in Managua, Nicaragua. Before the 1970s, there was little opportunity for deaf people to gather together and interact, and consequently Nicaragua had no standardized sign language. In the mid-1970s and early 1980s, the government opened a new primary school for special education followed by a vocational school. For the first time, deaf Nicaraguan children and adolescents were able to gather together in large numbers. While lessons were in spoken Spanish and instruction primarily focused on lip-reading and speaking Spanish (with limited success), students communicated through gestures, and a new sign language emerged and continues to develop (Polich, 2005). Each successive group of children who enters the community introduces linguistic complexity into the language that adults are unable to acquire (Senghas, 1995; Senghas & Coppola, 2001). This pattern gives rise to a unique community, where the older signers, the initial creators of the language, represent earlier stages of the language than do younger signers (Senghas, Kita, & Özyürek, 2004). Today, NSL has multiple co-existing age cohorts of users, from the creators of the language to the young children now learning and changing the language. In this project, we examine the language of signers from the *first cohort* (children who entered the community in the 1970s and early 1980s, the *second cohort* (entered the community in the mid-1980s to early 1990s),

[1] Note that recursion is not unique to the language system. Recursion is a property of generative systems and is found in other cognitive domains, such as visual cognition.

[2] There is a great deal of controversy about recursion, its definition, and its relevance to language and types of output generated by recursive functions (e.g., Tomalin, 2006; Watumull et al., 2014). In this paper, we adopt one definition, used by Pinker & Jackendoff (2005) and others: the embedding of a constituent in another constituent of the same type.

and the *third cohort* (entered the community in the mid-1990s to early 2000s).

By comparing the language of different cohorts of children who entered the community at different times, research on NSL to date has revealed a rich and nuanced picture of how historical and cognitive processes interact (e.g., Flaherty & Senghas, 2011; Pyers et al., 2010; Senghas, 2003). Many features of language do not develop immediately from the structure of a single human mind, nor do they require long periods of historical evolution. Instead, these elements emerge over the span of a few generations, indicating that convergence on these forms does not require prolonged historical evolution, but may require a community of users, transmission, and in some domains, sequential age cohorts of learners.

The present study addresses a fundamental unanswered question about a key property of language and the human mind. Specifically, we explore when recursion emerges in the creation of a new language.

The capacity for, and importance of, recursion in the language system can be seen in children's acquisition of syntactic recursion. Languages vary in which structures are recursive. For instance, possessives are recursive in English, but not in German (Roeper, 2011). This cross-linguistic variability in syntactic recursion has been argued to pose a learning challenge for children acquiring their first language (Roeper, 2011). Nonetheless, children can and do acquire syntactic recursion. Preschool-aged children are initially less likely to assign a recursive interpretation to embedded structures, parsing them as conjoined instead. For instance, when presented with an array of five balls in the order of red, green, blue, orange, and green, and asked to show the experimenter "the second green ball," 3- and 4-year-olds chose the first green ball in the array rather than the second (Matthei, 1982). Similarly, other work demonstrates that young children initially prefer conjoined interpretation for possessives in English and Japanese and then acquire the embedded structure (Fujimuri, 2010 as cited in Roeper, 2011; Limbach & Adone, 2010).

Everett (2005) argues that Pirahã lacked syntactic recursion. Further experimental work was not successful in eliciting embedded possessives or relative clauses in elicitation tasks (Piantadosi et al., 2012). In addition, a corpus analysis based on 15 stories (approximately 10,000 sentences) revealed no strong evidence of recursion for relative clauses, complement clauses, possessives, or conjunctions and disjunctions, but there was possible evidence suggesting recursive use of topics or repeated arguments (Piantadosi et al., 2012). The question of whether or not Pirahã lacks syntactic recursion remains open (e.g., Nevins, Pesetsky, & Rodrigues, 2009).

In the current study, we turn to an emerging language and ask whether NSL exhibits syntactic recursion and if the cohorts of signers show differences in the types of structures they produce. There are at least three broad possibilities for the origins of recursion. The first is that recursion is a property of the language faculty and thus should emerge early in a new language, like NSL (and Pirahã is anomalous, but see Piantadosi et al., 2012). The second is that recursion is an historical accomplishment (like mathematics) and is culturally constructed over time (e.g., Deutscher, 2005; Tomasello, 1999; 2008; 2011). On this view, recursion should not emerge in just three generations of NSL signers. A third possibility is that recursion is a property of the human mind, but a new language takes time to converge on (syntactic) forms for its expression, perhaps because of the need for a community of users.

Certain linguistic structures have been identified as typical examples of recursion, such as adjectives (*the second green ball*), relative clauses (*the boy who kicked the ball fell down*), complements (*John thinks that Sally is not coming today*), possessives (*my mother's brother's daughter's dog*), and prepositional phrases (*the cat in the hat on the table*) (Everett, 2010; Roeper, 2011). We investigate one structure that can take recursive interpretations: relative clauses. Comparing the language of three age cohorts of NSL signers, we ask how early we might observe evidence for syntactic recursion to better understand where the capacity for recursion stems from and when it emerges in a new language.

Experiment 1

Relative clauses serve to pick out an individual or subset from a set of referents (e.g., *the boy who was typing* from a set of boys performing different actions). Syntactically, relative clauses are full clauses that modify a noun phrase. In many languages, they contain a gap: a missing argument that is co-referential with the noun that they modify. In other languages, or syntactic contexts, this gap is filled, for example, with a resumptive pronoun as in "There are guests who I am curious about what *they* are going to say" (Prince, 1990). Semantically, they contain an element that is related to the noun or noun phrase. In the experiment, we examine whether NSL signers have strategies to fulfill the discourse function of relative clauses, namely picking out an individual from a set and whether there are form-based distinctions in the syntax.

Method

Participants in this study were 27 deaf Nicaraguan signers who were exposed to the emerging sign language by 6 years of age. Ten of the participants entered the signing community before 1983 (mean age of entry: 3.99, range: 3.0-5.7), and are referred to as the *first cohort* (mean age: 41.69, range: 37.2-46.8). Ten participants entered the community between 1986 and 1990 (mean age of entry: 4.01, range: 3.1-5.3), and are referred to as the *second cohort* (mean age: 30.84, range: 28.2-33.4). Seven participants entered the community between 1993 and 1998 (mean age of entry: 3.77, range: 3.0-4.6), and are referred to the *third cohort* (mean age = 22.06, range: 20.7-24.9).

Stimuli and procedure Two versions of six events were created. One version was designed to elicit relative clauses

and the other conjoined clauses. The relative clause versions of the events each depicted three characters (see Figure 1). The three characters were similar in appearance (same gender, similar physical features and coloring, similar clothing) but performed distinct identifying actions (e.g., one writing, one typing on a laptop, and one reading). Signers first saw individual clips depicting each character separately. Then signers saw a clip with all three characters together performing their respective actions, with a fourth individual watching the scene (see Figure 2). After a short period of time, the fourth individual leaves, and one of the three characters then engages in a new action (e.g., falls off a chair). Signers are asked to tell the fourth individual who had left the scene prior to observing the new action what happened.

The conjoined clause version differed from the embedded clause version in that the event portrayed only one character, instead of three characters (e.g., one character typing on a laptop then falling off a chair, see Figure 3). Here, the observing individual leaves prior to seeing any of the actions.

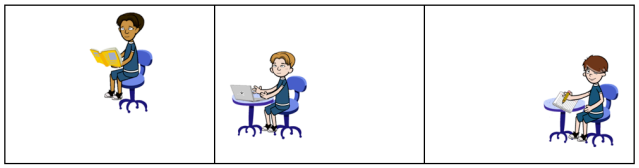


Figure 1. Example of the three characters in the relative clause stimuli. Each individual performs a different action (e.g., typing).

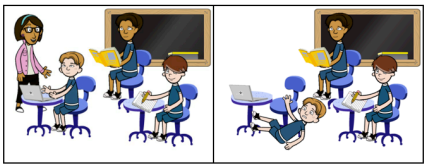


Figure 2. Example of new action in the relative clause stimuli. The left panel depicts all three characters with a fourth individual observing the scene. The right panel depicts a unique individual engaged in a particular action (e.g., falling).

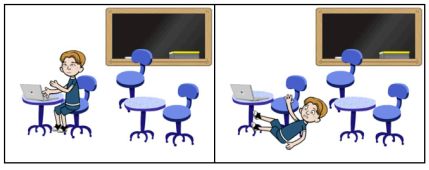


Figure 3. Example of conjoined clause stimuli.

The stimuli were normed on Amazon Mechanical Turk with monolingual English speakers. They were found to reliably elicit relative clauses (*the boy who was typing fell*) or reduced relatives (*the boy typing fell*) in the relative clause version, and conjoined clauses (*the boy typed and fell*, *the boy typed then fell*) in the conjoined clause version.

For the relative clause versions, signers first saw movies of the individual characters and were asked to describe each clip to ascertain that they interpreted the actions portrayed in the way we intended. Signers then saw the clip with all three characters and were asked to describe the event. For the conjoined clause versions, signers saw a single movie and were asked to describe the event. We filmed their descriptions for later analysis.

Coding Signed descriptions were coded for by the first author, a fluent signer of ASL (American Sign Language) with 8 years of research experience with NSL. Our coding centered on three levels of representation: *semantics* (differences in the discourse function of the descriptions), *syntax* (differences in the form), and *prosody*.

Semantics: We coded whether signers from all cohorts were able to fulfill the discourse function of relative clauses, picking out an individual from a set. In the relative clause version, the fourth individual in the scene sees the different actions of the characters (e.g., typing, reading, writing), but not the new action (e.g., falling). Because the fourth individual is aware of the different actions of individuals in the set, but not the new action of one individual, we expect that signers will identify the target individual when describing what happened, producing a description like *the boy who was typing fell*, contrasting the target with the other boys who were performing different actions (reading and writing). The establishment of a set and then the selection of an individual of that set is the semantic function that relative clauses serve, where the interpretation is one like: "There is a boy typing, a boy reading, and a boy writing. The boy who was typing fell" as opposed to "There is a boy typing, a boy reading, and a boy writing. The boy is typing and falls."

In contrast, in the conjoined clause version, there is a single character performing both actions. The observing individual leaves before seeing any of the actions. We expect that signers will produce what amounts to a conjoined clause, describing the scene as one where *a boy is typing and then falls*. Here, there is only a single character in the scene and the observer does not see the initial action, thus there is no identifying action that picks out a specific individual (e.g., the one who is typing). As such, a description of this event should semantically differ from a description in the relative clause version, where no set has been established.

Syntax: We coded one possible marker of form difference: the repetition and reduction in length of the verb that picks out the unique individual. Sign languages possess a rich system of verbal morphology, where signs can be inflected for person, number, location, and manner (see Padden, 1983; 1990). In ASL, verbs can be inflected through changes in movement to indicate aspectual information or in location to indicate subject-object agreement (Klima & Bellugi, 1979; Padden, 1983). Previous observation of NSL has noted use of a reduction in the

length of verbs as anaphor, suggesting this device may be a possible candidate for relative clause embedding.

In the relative clause version of our events, there is an opportunity to mention the individuating verb twice: the first when establishing the set (one boy *typing*, one boy reading, and one boy writing) and the second when describing the new action (the boy who was *typing* fell). Previous observations of NSL suggest that one possible marker of a relative clause is a reduction in the length of the verb the second time it is mentioned. The second, reduced mention of typing refers back to the boy who is typing, established at the beginning of the description.

In the conjoined clause case, there is no set of individuals and thus no action that individuates. As such, typing is only mentioned once, and does not serve as an identifier, but rather as a description of the first of two actions observed in the scene.

Prosody: We coded for differences in prosody. A relative clause (the boy who was typing fell) has a different prosodic structure than a conjoined clause (the boy was typing then fell). This difference may be marked with a longer pause or additional descriptions between mention of typing and mention of falling in the conjoined clause case compared to the relative clause case.

Results

The data were submitted to logistic mixed effects regression models with item and subject as random effects. For the analyses, we constructed separate logistic mixed effects models, with the trial coded as 1 when the variable of interest was present and 0 when it was absent. Cohort was coded using two dummy variables, with the first cohort acting as the baseline (the intercept). We compared two models: one model with cohort as a predictor variable and one model without cohort as a predictor variable.

First, we looked at whether signers from all three cohorts mentioned the identifying action of the target individual (e.g., typing) *and* the new action (e.g., falling). The model with and without cohort showed no significant difference, $F(3, 5) = 2.169$, $p = .338$ (Cohort 1: 97%; Cohort 2: 98%; Cohort 3: 100%). Then we examined whether signers established the set of referents from which the unique individual is picked out, specifically if they mentioned all three characters in the event. Again, we found no significant difference between the model with cohort and the model without cohort, $F(3, 5) = 4.042$, $p = .133$ (Cohort 1: 60%; Cohort 2: 73%; Cohort 3: 77%). Signers from all three cohorts established the set of referents and described the identifying action and new action of the target individual, providing robust evidence that all three cohorts have strategies to fulfill the discourse function of relative clauses.

The third analysis looked at whether signers repeated the individuating verb, as we would expect if they were producing a relative clause. If this device is developing in the language, we expect to see a difference across the cohorts. This was observed in the ANOVA comparison of the model with cohort and the model without cohort, $F(3, 5) = 5.314$, $p = .070$. We then conduced pairwise comparisons to determine where the differences among cohorts lay (see Figure 3). The difference between the first cohort and the second cohort was significant ($p < .01$), as was the difference between the first cohort and the third cohort ($p < .01$). The difference between the second and third cohorts was not significant ($p = .960$).

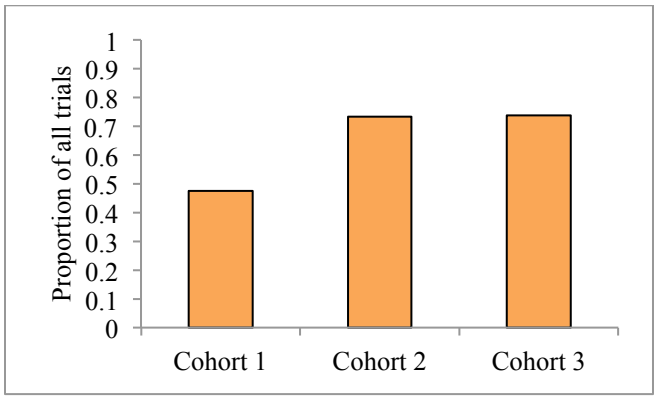


Figure 3. Proportion of trials in which signers repeated the individuating verb.

Finally, we looked at whether there was a reduction in the length of the second mention of the individuating verb. Here, we constructed linear mixed effects models for the comparison, computing the timing difference using a ratio of the length in milliseconds of the second mention of the individuating verb to the length of the first mention of the verb. The ANOVA analysis revealed that the model with cohort as a predictor performed significantly better, $F(4, 6) = 6.470$, $p = .039$ (see Figure 4). Follow-up comparisons revealed that the difference between the first and second cohorts was marginally significant ($p = .077$), and the difference between the first and third cohorts was significant ($p = .014$). The difference between the second and third cohorts was not significant ($p = .354$).

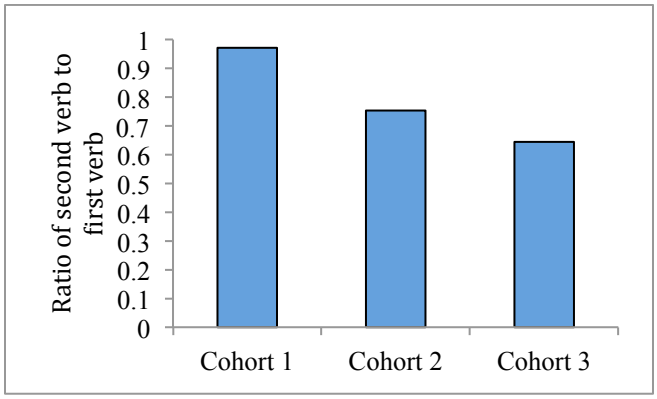


Figure 4. Ratio of second-mentioned verb to first-mentioned verb out of trials in which signers repeated the individuating verb.

Future work will analyze differences in descriptions of the relative clause stimuli and the conjoined clause stimuli.

Preliminary coding suggests that there are some prosodic differences as well as a greater number of words separating the two verbs in the conjoined clause descriptions compared to the two verbs in the relative clause descriptions, suggesting that descriptions with a relative clause differ prosodically from descriptions with a conjoined clause. We plan to do additional coding as well as perform a closer examination of the type of words that separate the two verbs in the two types of clauses (e.g., some words are related to the action encoded by the second verb).

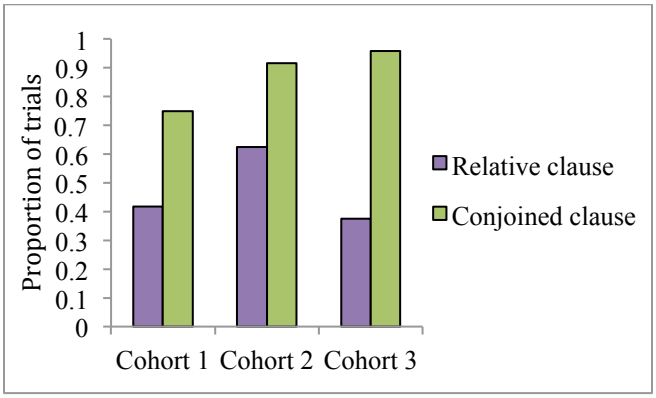


Figure 5. Proportion of trials in which words separated the two verbs in relative clause descriptions and conjoined clause descriptions.

Discussion

In the current study, we investigated whether the capacity to implement linguistic rules within rules (syntactic recursion) emerges early in a new language or if it must be constructed over historical time.

Our results suggest that signers from all three cohorts have strategies to fulfill the discourse function of relative clauses, namely picking out an individual from a set, producing semantically different utterances when describing a relative clause version of an event and a conjoined clause version. However, the grammatical form of the utterances differs across cohorts, with signers from later cohorts producing sentences with relative clauses. Specifically, third-cohort signers may be producing a reduced (shorter) form of the verb in the relative clause (the boy who was *typing* fell) compared to the first (longer) mention of the relevant individual and his action (there is a boy typing).

Follow-up studies will attempt to answer open questions from this work, ascertaining whether the prosodic differences observed between relative clauses and conjoined clause are indeed due to differences in form rather than differences in timing of the events. Future work will also test an alternative possible explanation for the reduction of verbs that we observed, namely that any subsequent mention of the same verb (rather than just the verb in a relative clause) is accompanying by a reduction in length.

The recent emergence of a sign language in Nicaragua offers us the opportunity to capture the creation and development of a new language. Here, we investigate whether linguistic rules that may allow for infinite expression are present in each individual human mind or if the formation and expression of these rules depend on social context and transmission across individuals. Early findings from this work suggest that syntactic recursion, while not appearing immediately in a new language, may nonetheless be an early-emerging property. Importantly, while we found suggestive evidence for syntactic recursion in the language of only the third cohort signers, signers from all three cohorts were able to express semantically different utterances, indicating that even if a language may not have syntactic recursion, human thought is recursive.

Acknowledgments

We thank the Nicaraguan participants for their enthusiastic participation. This research was supported by an NSF Graduate Research Fellowship to A.K., and a NIH/NIDCD grant R01 DC05407 to A.S.

References

Chomsky, N. (1968). *Language and Mind*. New York: Harcourt, Brace & World.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Christiansen, M. H., & Charter, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489-558.

Deutscher, G. (2000). *Syntactic Change in Akkadian: The Evolution of Sentential Complementation*. New York: Oxford University Press.

Deutscher, G. (2005). *The Unfolding of Language*. New York: Holt Paperbacks.

Emmorey, K. (2002). *Language, Cognition, and the Brain: Insights From Sign Language Research*. Mahwah, NJ: Lawrence Erlbaum.

Everett, D.L. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology*, *46*, 621-646.

Everett, D. L. (2010). You drink. You drive. You go to jail. Where's recursion? In M. Speas & T. Roeper (Eds.), *Recursion*. New York: Oxford University Press.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in nonhuman primates. *Science*, *303*, 377–380.

Flaherty, M., & Senghas, A. (2011). Numerosity and number signs in deaf Nicaraguan adults. *Cognition*, *121*, 427-436.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*, 1569-1579.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language*, *85*, 355–404.

Padden, C. (1983). Interaction of Morphology and Syntax in American Sign Language. Doctoral dissertation, Linguistics Department, University of California, San Diego.

Pinker, S. (1994). *The Language Instinct*. New York: Harper Perennial Modern Classics.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, *95*, 201-236.

Prince, Ellen. 1990. Syntax and discourse: A look at resumptive pronouns. In Proceedings of the sixteenth annual meeting of the Berkeley Linguistics Society: General session and parasession on the legacy of Grice, 482–497. Department of Linguistics, University of California, Berkeley.

Pyers, J., Shusterman, A., Senghas, A., Spelke, E., & Emmorey, K. (2010). Evidence from an emerging sign language that language supports spatial cognition. *Proceedings of the National Academy of Sciences*, *107*, 12116-12120.

Richie, R., Yang, C., & Coppola, M. (2013). Modeling the emergence of lexicons in homesign systems. In Proceedings of the Thirty-fifth Annual Conference of the Cognitive Science Society, 1223-1228.

Roeper, T. (2011). The acquisition of recursion: How formalism articulates the child's path. *Biolinguistics*, *5*, 57–86.

Senghas, A. (1995). Children's Contribution to the Birth of Nicaraguan Sign Language. Doctoral dissertation, Department of Brain and Cognitive Sciences, MIT.

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development*, *18*, 511-531.

Senghas, A., & Coppola, M. (2001). Children creating language: how Nicaraguan sign language acquired a spatial grammar. *Psychological Science*, *12*, 323-328.

Senghas, A., Kita, S., & Özyürek A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*, 1779-1782.

Tomasello, M. (2011). Human culture in evolutionary perspective. In M. Gelfand, C. Chiu, & Y.Hong (Eds.), *Advances in Culture and Psychology*. Oxford: Oxford University Press.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello, T., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*, 495-552.

Comparing competing views of analogy making using eye-tracking technology

Yannick Glady, Robert M. French, & Jean-Pierre Thibaut
yannick.glady@gmail.com, {jean-pierre.thibaut, robert.french}@u-bourgogne.fr
LEAD-CNRS, UMR 5022, Univ. Bourgogne Franche-Comté, Pôle AAFE
21065 DIJON. FRANCE

Abstract

We used eye-tracking to study the time course of analogical reasoning in adults. We considered proportions of looking times and saccades. The main question was whether or not adults would follow the same search strategies for different types of analogical problems (Scene Analogies vs. Classical A:B:C:D vs a Scene version of A:B::C:D). We then compared these results to the predictions of various models of analogical reasoning. Results revealed a picture of common search patterns with local adaptations to the specifics of each paradigm in both looking-time duration and the number and types of saccades. These results are discussed in terms of conceptions of analogical reasoning.

Keywords: Analogical reasoning; eye tracking; analogy tasks; strategies

Introduction

Analogical reasoning is a central feature of human cognition (Gentner & Holyoak, 1997; Holyoak, 2012; Hofstadter & Sander, 2013) and involves the transfer of relations from a source domain to a target domain. Analogical reasoning has been extensively studied from adult experimental, developmental and modeling perspectives and several general models have been proposed in order to characterize this form of reasoning in both children and adults (see French 2002; Gentner & Forbus, 2011, Holyoak, 2012).

The present work uses eye tracking data to study the temporal organization of the search for a solution in adults.,. Specifically, we compare different types of analogical reasoning tasks (Scene analogies, a standard ABCD task, a Scene version of the ABCD task -- see Figures 1a, 1b, and 1c, respectively), and the search profiles resulting from participants' analysis of the stimuli for each of them. In eye tracking studies, it has been shown that there is a correlation between the amount of attention paid to a particular item and the gaze-fixation (Deubel & Schneider, 1996; He & Kowler, 1992), and between the fixation time associated with a given item and its informativeness (Nodine, Carmody, & Kundel, 1978). All of this argues in favor of using eye tracking technology to study analogy-making strategies. Indeed, any analogical reasoning task involves analyzing the stimuli, comparing them, and mapping in order to find the relations shared by the two compared domains, this according to a temporal sequence.

Existing models of analogy make different predictions regarding *how and when* participants focus on and compare stimuli. Gentner and Forbus (2011) distinguish "align-first" models from "projection-first" models. Markman and Gentner (1993) propose an "alignment-first" conception which emphasizes the alignment of the stimuli that compose the base and the target domains. At some point during the interpretation, the comparisons of local elements, of local and global structures from both sides, one derives which elements should be put into correspondence (e.g., Falkenhaimer, Forbus, & Gentner, 1989). In the A:B::C:D paradigm, one would systematically tend to align A with C and look for a D (or Ds) to be aligned with B. In a scene analogy task (see Markman & Gentner, 1993; see below Figure 1), one would predict numerous saccades between the two pictures.

According to "projection-first" models (e.g., LISA, Hummel & Holyoak, 1997) participants first search for a relation unifying the base pair (i.e., the A:B pair in the A:B::C:D paradigm) and, then, try to find matches corresponding to this relation in the target pair (i.e., the C:Ds). This model would, therefore, predict more initial attention to the A:B pair, meaning more A-B transitions and more gaze time in the first image in the Scene analogy task. Later in the trial, one expects more gazes towards C and Target and more C-Target transitions in both the A:B:C:D task or the scene analogy task.

By contrast with the above models, the "relational priming" model (Leech et al., 2008) gives no role to mapping. According to this view, children first study the A-B pair. Then, the relation found between A and B (conceived of as the retrieval of a relevant *transformation* between A and B, e.g., "cuts" for knife and bread) *automatically primes* the retrieval of a relationship between C and the target item into which it is transformed. With respect to the time course of comprehension, this model predicts that participants should study the A-B pair first, turning, later, to C and the solution set. However, participants should not actively compare C with the semantic distractors or the semantic distractor with the Target because the C-Target pair would be *automatically primed* by the A-B pair relation. The solution is found through *priming* without systematic comparisons within and between the pairs, thus no mapping.

A limited number of eye tracking studies involve analogy-making (e.g., Gordon & Moser, 2007; Thibaut et al., 2011; Thibaut & French, 2016). Bethell-Fox et al. (1984) used an A:B::C:D task, with easy and difficult *geometrical* analogies. They found that participants with less fluid intelligence relied more on the elimination of implausible

answers, i.e. had more transitions within pictures in the solution set. Interestingly, difficult items elicited more saccades back to A and B, i.e., more time spent on the A:B pair of the problem before looking to alternatives. Participants also looked at the alternatives more often than in simple trials. The authors also found that when participants first looked at the correct answer, they later tended to look at a lower number of alternatives than when an incorrect answer was first looked at. If participants looked less at other options when their first look was for the correct answer, this suggests that they had already constructed a solution for the A:B pair which allowed them to recognize that the correct solution was correct.

In developmental studies Thibaut et al. (2011) and Thibaut & French (2016) with A:B::C:D semantic analogies found key differences between adults and children. First, adults focused on the A:B pair at the beginning of the trial, paying little or no attention to C and to stimuli in the solution set. At the end of the trial, the Target was their sole focus of attention. By contrast, children organized their search around C which they actively focused on during the entire trial. At the beginning of the trial they paid more attention to C and B but also looked at the Target and the semantic distractor earlier than adults. Results also showed that children paid less attention to A and B. Overall, adults behaved in a projection-first way whereas children followed neither a projection-first (or constructive) strategy nor an alignment-first strategy but, rather, a strategy organized around C (Glady, Thibaut, French, 2012; Glady et al., 2014; Thibaut & French, 2016).

Gordon and Moser (2007) used scene analogies (see Markman, & Gentner, 1993) in which participants had to point which item in a scene had the same role as an item pointed to by the experimenter in the other scene (e.g. pointing to a boy chasing a girl, if the experimenter pointed to a dog chasing a cat, see below Figure 1). Adults initially focused on the "actor-patient" pair in the source image (i.e., a dog chasing a cat, which is analogous to our A and B terms) and then looked for the solution in the target image (a second actor-patient pair, e.g., a girl chasing a boy, analogous to our C and D terms). This is consistent with the projection-first conceptions (study A:B then analyze C:D). They also showed that there were longer fixations on the stimuli with an arrow pointing to it in the source scene and its match in the target scene.

Goals and Rationale

In the present paper, we compared adults' visual strategies in three different analogical tasks, the scene analogy task used by Gordon and Moser (2007) (see Figure 1a), the standard A:B::C:D task (hereafter "ABCD task", see Figure 1b) and a modified "scene" version of the A:B::C:D task in which the stimuli were inserted in scenes (hereafter, the Scene ABCD task, Figure 1c). The resulting stimuli were similar to the Scene analogy task. In the Scene Analogy task, participants have to find what plays the same role in the target scene (e.g. the boy who is chasing a girl, see Figure 1) as the element that is pointed to in the source scene (e.g., a cat chasing a mouse). In this task, the stimuli are semantically connected in meaningful scenes. By contrast, in the A:B::C:? task, stimuli are presented separately and the resulting display has no global meaning. Given the instructions, participants have to compare separated pictures in order to find the common relation. The idea behind the Scene ABCD task was to insert the stimuli in simple but meaningful scene (e.g. a bird flying close to a nest). In this case the instructions remain the same as in the standard ABCD task (find the one that goes with a designated stimulus [e.g., a dog], in the same way as the bird goes with the nest). Recent evidence suggests that grouping objects in meaningful scenes might positively affect attention toward these objects and their relations (see Humphreys et al., 2010 for a review).

The main difference between the two "scene" tasks is that in the Scene analogy task, the focus is on the identification of the role of the designated object in the relation between A and B. Once this is done, one must identify stimuli connected by the same relation in the second picture and the stimulus which plays the analogous role as the one designated in the first pair. Thus, the emphasis is on the role of the designated item within an easily identified relation between the items. In the Scene-ABCD, when one is gazing at the AB scene, he/she has to find the relation connecting them and then orient the search in the second image around the designated C ("what goes with C") and find the object that fulfills the same role in the other objects. Here, the emphasis is on finding a common relation, something which might not be straightforward. In the standard A:B::C:D case, finding the direction of the relation might be simpler since location cues make it more salient.

Are the search patterns equivalent in the three tasks? Is the Scene-ABCD task more similar to the Scene task or to the ABCD task? In terms of the above models, the question was whether a unified profile of answers would appear for all tasks. For example, would there be a projection-first pattern (A:B then C:D) in all cases or not? Would we observe more alignments (AC and BT saccades),for Scene analogy tasks than in the two other tasks, this being because participants had to identify equivalent roles in both scenes? Since scenes are semantically integrated, they might elicit more intra-domain saccades than in ABCD task. In order to test Leech et al. (2008), we will focus on the way participants analyze the second scene or pair, especially the presence or absence of Target-Semantic distractor comparisons.

In what follows we will attempt to test these hypotheses experimentally. Crucially, we divided the time course of each trial into three equal time slices, which allowed us to study the evolving dynamics of problem resolution. Specifically, we used a Task (Scene Analogy, Standard A:B::C:?, Scene ABCD) x Time slice (1st, 2nd and 3rd slices) x Stimulus Type (A, B, etc.) design as within-subject factors. Response accuracy, reaction times and eye movements were recorded.

Methods

Participants

Subjects were 20 adults (14 females, 6 males; mean age=20.4 years; SD=2.21; range: from 17 to 27 years), Univ. Bourgogne Franche-Comté students. They received course credit for their participation.

Materials

Three tasks were used, each composed of three training trials and four experimental trials. The first task was a Scene Analogy task (Figure 1a), the second a Standard ABCD task (Figure 1b) and the third a Scene-ABCD task, i.e., a standard A:B::C:? task with the items composing the problems inserted in scenes (Figure 1c). In each task, the stimuli were composed of 7 black and white line drawings. The tasks were presented sequentially and appropriately counterbalanced. The order of the trials within each task was random.

In the Scene Analogy task (Fig. 1a), the pictures were based on materials in which the distractor was chosen to be semantically related to one member of the relation in the bottom picture (see Richland et al., 2006; Gordon & Moser, 2007), for example, a ball in the "boy chasing a girl" case.

In both the Scene Analogy task (Figure 1a) and the Scene A:B::C:D task, trials were constituted of two scenes (501x376 pixels each) containing 7 black and white line drawings (corresponding to A, B, C, T, semantic distractor and two unrelated distractors) framed by a black rectangle. The top scene contained the A and B pictures; the remaining pictures were in the bottom scene. The only difference in the general presentation of the two tasks was that an arrow pointed to the B stimulus in the Scene Analogy task, and to the C stimulus in the Scene ABCD task.

In the Standard ABCD trials, the A, B, C drawings were presented above the picture. On the top right, a black empty square was the solution location. The four remaining pictures (Target, Semantic Distractor and two Unrelated distractors) were presented at the bottom of the screen (Figure 1). In the analyses presented below, the potential-answer items were designated as follows: T (correct Target stimulus), SemDis (semantic distractor), and UnDis (distractor semantically unrelated to either A, B or C). The size of each picture was 200x195 pixels. Pictures were not aligned in two rows in order to minimize saccades of participant's eyes involving stimuli "on the way" to another stimulus.

The tasks were displayed on a Tobii T120 eye-tracker with a 1024x768 screen resolution using an E-Prime© software (version 2.8.0.22) embedded in a Tobii Studio (version 2.1.12) procedure to record participants' scanpaths.

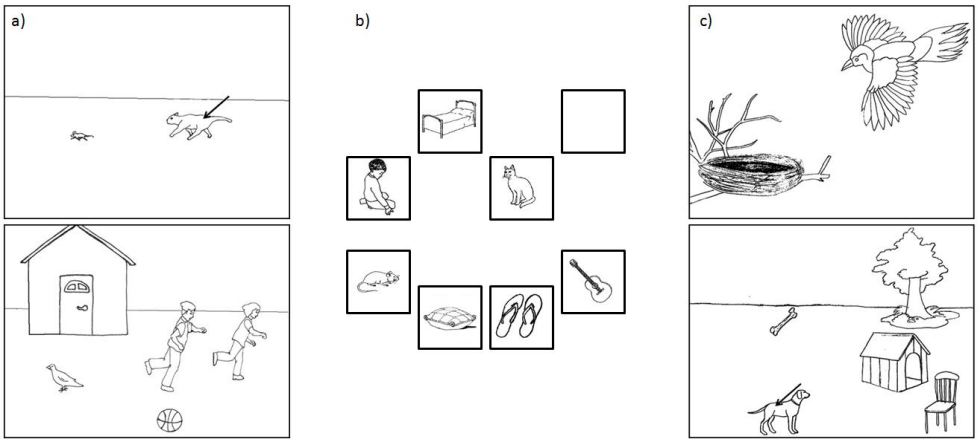


Figure 1: The three tasks used in this experiment: a) scene analogy task, b) standard ABCD task, c) Scene ABCD task.

Procedure

Participants were tested individually in a quiet room at the University. Following calibration, participants were shown a training trial and were given the instructions corresponding to each condition: In the Scene Analogy task, they were shown a stimulus in the above frame and were asked to point to the stimulus that played the same role in the below picture as the one played by the object that was pointed to in the first picture. In the Scene-ABCD task they were told "There are two pictures [pointing to A and B]. A:B go together well. Can you see why [A and B] go together?" Once the participant had given a relation linking A and B, the experimenter gave feedback. "OK! Now, do you see this one [pointing to C]? What you have to do is to find in this picture [pointing to the below scene] the item that goes with this one [C] in the same way as this one [B] goes with this one [A]." In the standard ABCD task, participants were asked to find in the solution set the stimulus that goes with C in the same way as A goes with B. No further instruction and feedback were given during test trials. Eye tracking data were recorded when the presentation of the problem started and stopped when an answer was given.

Results

Trials were excluded of the eye-tracking analysis when more than 50% of the eye movement data missing. No participant was excluded of the statistical analyses.

Participants' performance was above 95% correct in the three tasks. These analogies were simple because they were used for comparison with children in another experiment.

An ANOVA on reaction times as a function of Task Type revealed no main effect of task, p >.2)

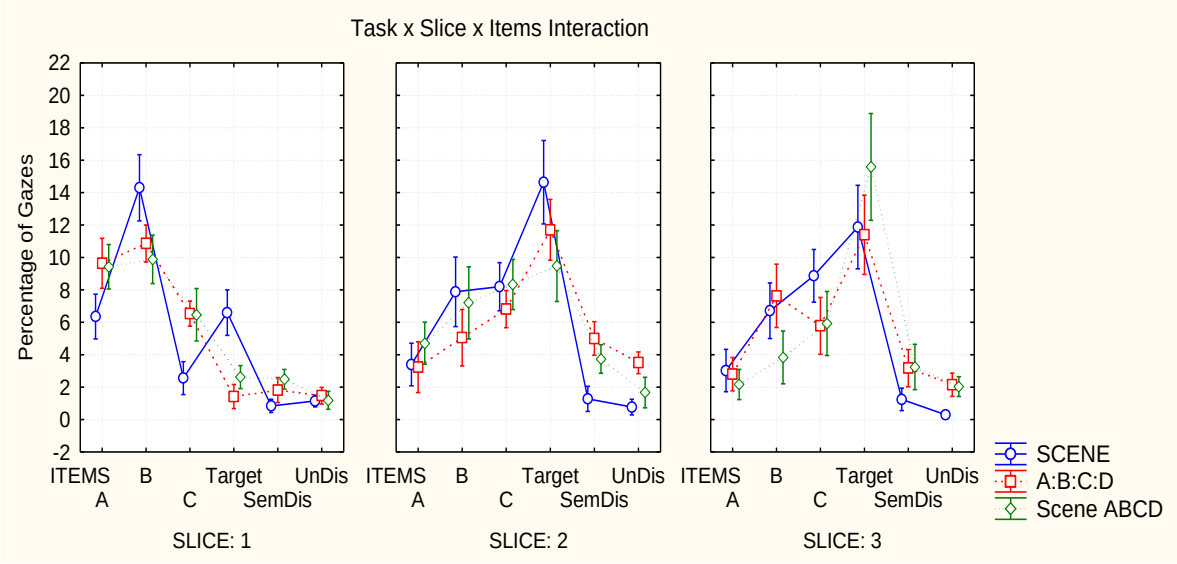


Figure 2: Interaction between Task(Scene, Standard ABCD, and Scene ABCD), Time Slice, and Stimulus Type for the percentage of gazes. Patterns of gazes were partially different in Scene condition and the two versions of the ABCD task (see text).

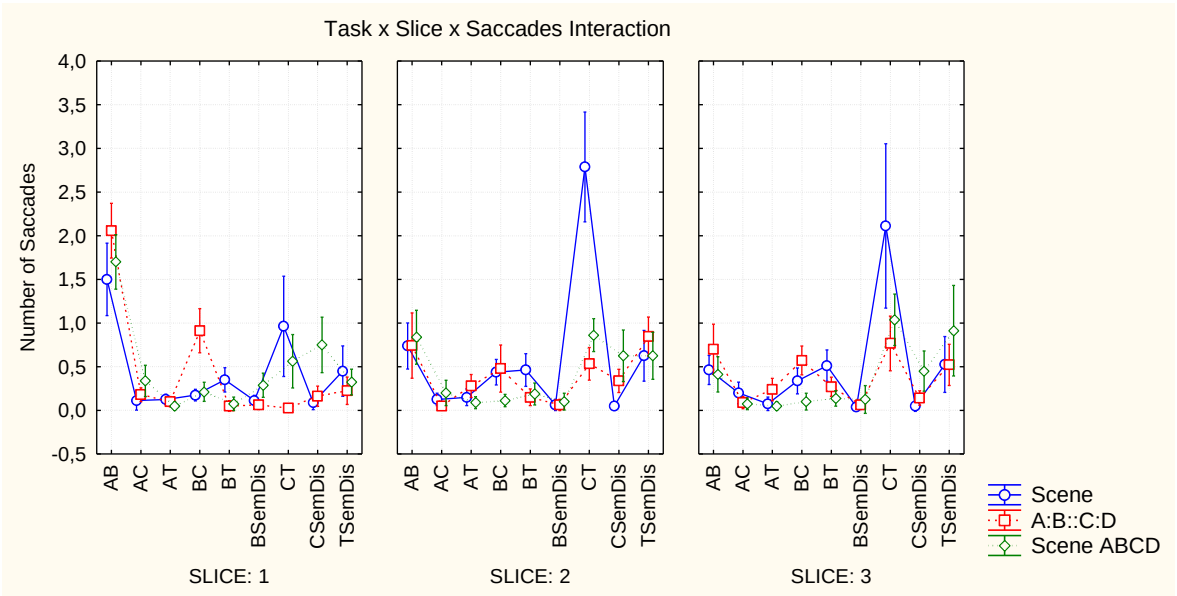


Figure 3: Interaction between Task (Scene, Standard ABCD, and Scene ABCD), Time Slice, and Stimulus Type for the percentage of gazes. Patterns of switches were partially different in Scene condition and the two versions of the ABCD task (see text).

Eyetracking data

Visual Strategies

We analyzed the proportions of gazes towards each type of stimuli with a three-way ANOVA with Type of Stimulus (A, B, C, T, SemDis, UnDis), Task Type (Scene Analogy, ABCD, Scene ABCD) and Slice (1st, 2nd, 3rd) as within-subject factors (Figure 1). There was a main effect of Type of stimulus, $F(8,328)=17.1$; $p<.001$; $\eta^2_p=.294$). There was a significant interaction between Task and Type of Stimulus, $F(10,190)=9.96$; $p<.0001$; $\eta^2_p=.34$) between Slice and Type of Stimulus $F(10,190)=3.57$; $p<.0001$; $\eta^2_p=.65$), and a significant interaction between the three factors, $F(20,380)=6.34$; $p<.0001$; $\eta^2_p=.25$) which was the main result to be considered here.

Figure 2 shows that the three tasks started with gazes on the AB pair, which is in favor of the projection-first strategy. The Standard ABCD and the Scene ABCD tasks

were quite similar in the first slice of the trial, with A and B dominant, and far fewer gazes to the potential solutions (Target, semantic distractor, and unrelated distractor). By contrast, in the Scene condition, B (i.e., the stimulus pointed to by the experimenter) was dominant at the beginning of the trial, whereas A and Target received approximately the same number of gazes, which meant that the Target received more gazes than in the two other conditions. In the second slice, C and the Target dominated in all three tasks, showing that participants were making their decision while comparing these two stimuli (see the "Saccades" section below). During this second slice, they continued to look at B (recall that in the scene analogy task, B corresponds to the object that is designated by the arrow). The third slice is similar. In the two ABCD tasks, the second slices were progressively organized around the Target. The Semantic distractor and the unrelated distractor received more gazes than in the Scene condition. This suggests that, in the ABCD tasks, participants look at the entire space of stimuli while making their decision. Note that the pattern of gazes for the ABCD task was similar to the one obtained by Thibaut and French (2016).

Saccades We also analyzed the saccades between stimuli. They tell us which stimuli are compared and when. We focused on a subset of 9 transitions, the most relevant here (see Thibaut et al. 2011; Thibaut & French, 2016). The subset was composed of A-B, C-T(arget), A-C, B-T and also A-T, B-C, B-SemDis, C-SemDis, and T-SemDis.

The first four of the above transitions are crucial to determining whether participants follow strict projection-first strategies (A-B then C-Target), or have many alignments (i.e., A-C and B-T and between sets comparisons) or a combination of both, depending on the moment of the trial. The other transitions refer to other comparisons between C and the solution set (or within the bottom scene). We ran a three-way repeated-measures ANOVA with Transitions (A-B, C-T, A-C, B-T, A-T, B-C, C-SemDis, T-SemDis), Slice (1st, 2nd, 3rd), and Task (Scene, ABCD, and Scene ABCD) as within-subject factors. The most important result was the significant interaction between Transition Type, Task and Slice, $F(32,608)=3.40$; $p<.0001$; $\eta^2_p=.15$ (see Figure 3). The main effects and the two-way interactions were also significant.

In *the first slice*, the main common result is the presence of A-B transitions and the absence of A-C and B-T transitions. The main difference between the three tasks was on B-C which appeared only in the Standard ABCD task. One interpretation of this result is that the B and C are spatially close. Once they have received the instructions (i.e. "what goes with C") they looked at it, and moved to A and B which they compared as witnessed by the number of A-B saccades. In the Scene analogy task, in the first slice, A-B and C-T saccades were present, suggesting that participants analyzed A and B and C-T in parallel in order to find the equivalent stimuli in the two scenes. By contrast, the search is more sequential, A-B then C-T, in the ABCD task with an overwhelming prevalence of A-B saccades at the beginning. The Scene-ABCD shows the same prevalence of A-B saccades. However, one clearly sees the construction of a semantic space in the second scene with the presence of C-T, C-SemDis and T-SemDis saccades. In the second slice, C-T saccades are overwhelmingly dominant in the Scene condition which suggests that participants have identified the stimuli that are involved in the target relation and are comparing them to attribute a semantic role to them. In the Scene-ABCD task, C-T, C-SemDis and T-SemDis are equally distributed. There is a slight dominance of T-SemDis in the ABCD task suggesting that comparisons between semantically-related-to-C stimuli are important at this point (confirming Thibaut & French 2016). These patterns suggest that in both the ABCD and Scene-ABCD tasks participants compare C, T and SemDis in order to find the correct solution. By contrast, evidence for A-C and B-T saccades remains scarce in the three tasks, suggesting that *aligning the images is not a primary objective in any of them*: the three are organized around the A-B, C-T pattern. The pattern of results here is consistent with the projection-first hypothesis, rather than with the hypothesis of alignments taking place at some point. There was also strong evidence of comparison of stimuli in the solution set (T, SemDis) which is *not* consistent with the relational priming view.

Discussion

The present paper extends the results of previous eye-tracking studies by comparing three different analogy tasks by means of eye tracking. Our first purpose was to assess whether and when participants tended to use projection-first or alignment-first strategies (see Introduction). Data confirmed previous studies by Thibaut and colleagues showing higher rates of A-B and C-T saccades in participants' patterns of visual search than of A-C and B-T saccades at any point during the trial, and a prevalence of gazes towards A and B in early steps of the trial. Thus, the first main result is that *participants tend to construct or interpret an analogy by comparing the relation constructed in the AB domain before looking applying it in the target domain*.

A priori, it could have been argued that scene analogy trials would require more alignments (A-C and B-T transitions) (Markman & Gentner, 1993) than ABCD tasks. Indeed, in order to establish role equivalence, participants should first identify which objects might be compared in both scenes (A,B and C, T), which means that they should saccade between these stimuli in the two pairs. However, in reality, this rarely occurred (few A-C, B-T, B-C and A-T saccades). One can observe a small number of B-T transitions in the Scene task, but only in the second and third time slices. However, their number remains low compared to A-B or C-T, transitions. When they do appear, they are not accompanied by the A-C transitions that should appear simultaneously if one wants to establish the existence of parallel search of equivalent objects between the two pairs.

Moreover, there was no evidence of relational priming in any of the three conditions. Participants first constructed a solution for the first Scene or pair. Then, they systematically, explored all the related-to-C options especially in the two A:B::C::D tasks.

Regarding differences between the tasks, participants had fewer gazes towards distractors and fewer saccades involving the semantic distractor in the Scene task than in the other tasks. Overall the two scenes tasks (i.e., Scene and Scene-ABCD) differed in the way the solution was constructed. The Scene task was organized around T from the start (T gazes and C-T saccades) and fewer of the others whereas the proportion of gazes towards C, T and SemDis and transitions involving them was initially more balanced in the Scene-ABCD condition, suggesting that participants were constructing the solution from comparisons between these stimuli. Note that other factors might also play a role, such as item difficulty (not manipulated here). Indeed, Glady, French & Thibaut (2014) showed that item difficulty was associated with more gazes towards distractors, which was interpreted as an evidence for the necessity of a deeper analysis of the stimuli in order to find the solution.

In conclusion, our results revealed general patterns in the search of a solution, together with local adaptations to the specifics of the tasks.

Acknowledgements

This research has been supported by French ANR Grant 10-BLAN-1908-01, ANAFONEX to the third author, a joint ANR-ESRC grant 10-056 GETPIMA to the second author, and a FABER grant from the Conseil Regional de Bourgogne to the first and the third authors.

References

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*, 205-238.

Deubel, H., & Schneider, W. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, 36, 1827–1837.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1-63.

French, R. M. (2002). The computational modeling of analogy-making. *Trends in Cognitive Science, 6(5)*, 200-205.

Gentner, D., & Smith, L. (2012). Analogical reasoning. *Encyclopedia of human behavior*, Vol. 1. Elsevier Inc.

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist*, *52*(1), 32.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science, 2(3)*, 266-276.

Glady, Y., Thibaut, J.P., and French, R. M. (2014). Adults' Eye Tracking Search Profiles and Analogy Difficulty. *Proceedings of the Thirty-fourth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 535-540.

Glady, Y., Thibaut, J.P.and French, R. M. (2012). Explaining children's failure in analogy making tasks: A problem of focus of attention? *Proceedings of the Thirty-fourth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 384-389.

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition*, *15*(1), 20–35.

He, P., & Kowler, E. (1992). The role of saccades in the perception of texture patterns. *Vision Research*, *32*(11), 2151–2163.

Hofstadter, D. R. and Sander, E. (2013). *Surfaces and Essences: Analogy as the fuel and fire of thinking.* NYC, NY: Basic Books.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning*. New York, NY: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427.

Humphreys, G. W., Yoon, E. Y., Kumar, S., Lestou, V., Kitadono, K., Roberts, K. L., & Riddoch, M. J. (2010). The interaction of attention and action: From seeing action to acting on perception. *British Journal of Psychology*, *101*(2), 185–206.

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *Behavioral and Brain Sciences*, *31*(04), 357-378.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431-467.

Nodine, C. E., Carmody, D. P., & Kundel, H. L. (1978). Searching for Nina. In J. Senders, D. F. Fisher, & R. Monty (Eds.), *Eye movements and the higher psychological functions*. Hillsdale, NJ: Erlbaum.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-273.

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development*, *38*, 10-26.

Thibaut, J.-P., French, R. M., Missault, A., Gérard, Y., & Glady, Y. (2011). In the eyes of the beholder: What eye-tracking reveals about analogy-making strategies in children and adults. *Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society* (pp. 453–458).

Memory for the Random: A Simulation of Computer Program Recall

Fernand Gobet (fgobet@liv.ac.uk)
Department of Psychological Sciences, University of Liverpool,
Bedford Street South, Liverpool, L69 3BX, UK
Iain Oliver (iso@psyc.nott.ac.uk)
School of Psychology
University Park, University of Nottingham
Nottingham NG7 2RD, UK

Abstract

Contrary to a widely held belief, experts recall random material better than non-experts. This phenomenon, predicted by the CHREST computational model, was first established with chess players. Recently, it has been shown through a meta-analysis that it generalises to nearly all domains where the effect has been tested. In this paper, we carry out computer simulations to test whether the mechanism postulated with chess experts – the acquisition and use of a large number of chunks – also applies to computer programming experts. The results show that a simplified version of CHREST (without the learning and use of high-level schemata known as templates) broadly captures the skill effect with scrambled programs. However, it fails to account for the differences found in humans between different types of randomisation. To account for these differences, additional mechanisms are necessary that use semantic processing.

Keywords: chunk; computer programming; expertise; memory recall; random material

Introduction

Computer programming involves a variety of skills: the ability to understand programs written by others, to design, write and debug one's own programs, and to use problem-solving strategies to turn a set of constraints and desiderata into a correct and running program. Several general theories have been proposed to account for these abilities. Some authors (e.g., Adelson, 1981; McKeithen, Reitman, Rueter & Hirtle, 1981; Ye & Salvendy, 1994) have proposed that semantic knowledge plays an essential role. Others have proposed that expertise in programming, like in other domains, stems from the acquisition of a large number of perceptual chunks, which are the building blocks on which later semantic and procedural knowledge is constructed (e.g., Chase & Simon, 1973; Simon & Gobet, 2000).

Using a variety of chess-related tasks including a recall task, Chase and Simon (1973) gathered good evidence for the psychological reality of perceptual chunks. In addition, they found that there was a massive skill effect for the recall of positions taken from Masters' games, but that this effect disappeared with random positions. It was later shown that chess Masters keep a small, but reliable, superiority with random positions (Gobet & Simon, 1996). This result was actually predicted by CHREST (Chunk Hierarchy and REtrieval STructures), a computer model based on the idea of chunking (De Groot & Gobet, 1996; Gobet, 1993). The reason is simple: an expert, who has acquired more chunks than a weak player, is more likely to recognise a few chunks in a given position fortuitously, and thus obtains a better recall. Crucially, Sala and Gobet (2016) have recently demonstrated in a meta-analysis that this effect is present in nearly every domain of expertise reviewed. The overall correlation between expertise and recall of random material was moderate but statistically significant ($r = .42$, $p < .001$).

Chase and Simon's (1973) chunking theory spawned a large number of experimental studies. Several of these studies have been carried out in the domain of computer programming, and the importance of chunking in programming is generally accepted (e.g., Adelson, 1981; Barfield, 1986; McKeithen et al., 1981). In addition, Schmidt (1986) found that recall of computer programs correlates with their comprehension. Since high comprehension is a distinguishing feature of expertise in programming, this correlation suggests that chunks, as measured by the recall task, may play a causal role.

To our knowledge, no computational model has been developed so far to simulate the empirical data about memory for computer programs. The goal of the present paper is to fill in this gap, using the CHREST architecture as a modelling environment. Given Sala and Gobet's (2016) recent finding that the skill effect with random material generalises to many domains of expertise, the focus will be on the recall of randomised programs and the role played by perceptual chunks.

The paper is organised as follows. First, we briefly review research on memory for computer programs. Second, we describe a computer simulation using CHREST. Third, we compare the results of the simulations with those obtained with humans. Finally, we reflect on the impact of our results upon research into expertise.

Memory for Programs

Several studies have been carried out to investigate memory for computer code by individuals of different levels of expertise. While some of the studies were also interested in cognitive processing differences, this review will focus upon the studies where the recall task has been used—that is, a brief presentation of material taken from the domain of expertise, and a subsequent test of memory. In selecting the studies, we have also used the criterion that the experiment

should compare memory recall ability between expert and novice programmers, and that some measure of performance (e.g., percentage of lines correctly recalled) was provided. These criteria resulted in the selection of four studies: Adelson (1981), Barfield (1986), Bateson, Alexander and Murphy (1987), and Guerin and Matthews (1990). Important features in these experiments include the participants' skill level and the type of stimulus given to them.

Assessing Programming Ability

Unlike similar endeavours into chess, there is no standard rating scale measuring a computer programmer's level of expertise. For the four experiments to be reviewed, each participant's level of expertise with a particular language was determined by their experience with it.[1] In general, the participants in the novice group had some experience with the programming language used in the experiment. The expert group usually consisted of programmers that had completed, or lectured on, courses in the language. Some programmers were rated as experts because they had experience in more languages than the target one in the experiment.

Experimental Material

For all experiments, the materials were examples of real computer code. Each experiment only used one programming language to draw its examples from, even if some of the participants knew more than one language. But unfortunately, there are no two experiments using the same language so as to allow direct comparison of results. Indeed, differences in the languages, experimental designs, and scoring methods make detailed comparisons awkward. Some authors (e.g., Guerin & Matthews, 1990) even criticised other experimenters for their choice of target programming language.

Summaries of Experiments

Adelson (1981). This experiment addressed the question of how experts represent and use programming concepts. It tried to show that experts use a hierarchy to organise their information and base its structure upon functional aspects of programming. This is opposed to novices who organise information based on the program syntax.

The experiment used the Polymorphic Programming Language (PPL). PPL is a variant of PL/I, which is a combination of FORTRAN, ALGOL and COBOL (see Schmidt, 1986). The novices were five undergraduates who had completed a course in PPL, and the experts were five lecturers in that language. Sixteen lines of PPL code taken from three separate, complete programs were used as stimuli. Each line of code was presented separately on a screen. Lines were presented in a random order and each line was visible for 20 seconds. After all lines had been presented, the participants had 8 minutes to recall the code. This procedure was repeated for nine trials.

Experts recalled more than the novices (see Figure 1). Adelson suggests that the discrepancy between Chase and Simon's (1973) chess data and her data comes from the fact that the code consists of lines taken from three complete programs and not of lines randomly selected from 16 different programs.

Adelson also looked at the size of chunks used in recall, defining a chunk as a sequence of items recalled in succession with less than a 10-second pause between them. Experts' chunk size was greater than novices' (on average 3.5 and 2.4 items, respectively). Based on these and additional results, Adelson concluded that experts organise information using functional principles, while novices categorise on a more syntactic (surface) basis.

Barfield (1986). Barfield was interested in being able to distinguish novice from expert problem solving behaviour and knowledge acquisition, and concentrated on chunking as the main process that discriminates individuals of different skill levels. He suggested that programmers take in the complex stimuli as meaningful chunks before they are processed.

Four levels of expertise were used. Naïve participants (n = 42) had not completed any programming courses. Novices (n = 80) had completed just one course in BASIC. Intermediates (n = 73) had completed a minimum of one BASIC course plus two or three courses in other languages. Experts (n = 26) were graduates in computer science as well as having at least one course in BASIC.

The material consisted of one 25-line program written in BASIC. The experimenters identified likely modules within the program, but no visible boundaries were marked (i.e. no spaces between lines). The experiment had three conditions: the stimulus could be presented either (a) in executable order, (b) with the order of lines randomised, or (c) with the order of modules randomised (in that case, the lines within a module preserved their order). The participants were allowed three minutes to study the stimulus and four minutes to recall it.

The results are summarised in Figure 1. Naïve and novice participants obtained the same level of performance regardless of stimulus type, indicating that little, if any, of the chunk knowledge possessed by experts is present with novices. According to the results, intermediates can chunk together lines of code as long as they are in executable order. As expected, randomising the lines did negatively affect the performance of the experts, although they still did better than Novices and Naïve participants. The randomising of modules did not affect the performance of experts and this was taken as support for Barfield's chunking explanation. However, Guerin and Matthews (1990) argue that Barfield is measuring recall and not comprehension, so even though the semantic structure of the program is tampered with, it will not affect the results because BASIC programmers are not as sensitive to the semantic complexity as programmers using other languages. They also criticise

[1]This is far from being a foolproof method. From research into other expertise domains, it is known that experience correlates only imperfectly with expertise (Gobet, 2016).

Barfield for not randomising lines within modules to complete his experimental design.

Bateson, Alexander and Murphy (1987). This paper aimed to expose expert-novice differences in syntactic and semantic memory, along with tactical and strategic skill. The gist of Bateson's experiment is to demonstrate the importance of semantics in gauging programmers' differences over tasks that use measures of memory and chunk size in syntactic recall. We discuss only the first of Bateson et al.'s battery of tasks, the syntactic memory task.

Two groups were used for this task: novices (n = 20), who had completed no more than three programming courses, and experts (n = 30), who had completed more than three courses. All participants had completed 12 weeks in an introductory FORTRAN class. The material used for the experiment was four short programs of equal length written in FORTRAN. Two of the four programs had lines randomised.

Participants were given only one normal and one random program, and were given three minutes to study a program and then allowed four minutes of free recall to write down what they remembered. The means for the proportion of total program recall are shown in Figure 1. As with Adelson (1981), there is a skill effect even when the order of the lines is randomised.

Guerin and Matthews (1990). This study aims to demonstrate the role of semantic knowledge in expert programmer ability. Only the first of their three experiments is described here. Guerin and Matthews used COBOL as their target language, and a genuine 116-line COBOL program was used as material. Two groups were used; the novices (n = 52) had an average of 0.5 years of programming experience and the experts (n = 52) had an average of 4.7 years. The participants were given 10 minutes to study the stimulus and then 8 minutes of free recall. There were four conditions: (a) Normal program; (b) Random lines within program modules; (c) Random modules; and (d) Random lines within program modules and random modules. A module is described by Guerin and Matthews as being a chunk of a program; however, they do not go into detail as to whether they are describing a functional section of a program or an amount of information thought capable of being memorised in one go. Nor do they specify how large these modules are.

Guerin and Matthews used a unique method to score recall trials. Instead of counting correctly recalled items, a system of points based on positional and lexical accuracy was devised in order to better describe recall performance. For each line, one point was earned if more than half of the components of the line were recalled correctly; an additional point was added if it was recalled in the correct sequence in the program, and a final point was added if it was recalled exactly as the original. Thus, a maximum of three points could be earned for each line.

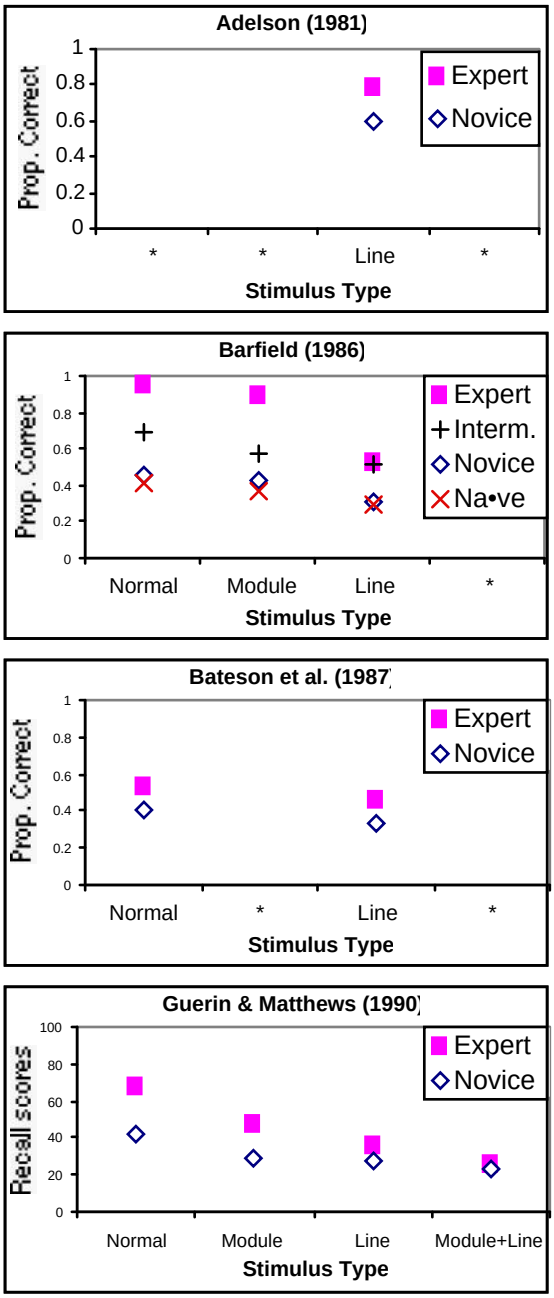


Figure 1. Memory for computer code as a function of level of expertise and type of randomisation. Adelson (1981) and Barfield (1986) used the number of lines correct, and Bateson et al. (1987) used the number of items correct (to facilitate comparison, we have converted these absolute numbers into proportion correct). Guerin and Matthews (1990) devised their own scoring method.

In all four conditions, experts were better than novices. The ordering of the conditions was as follows: Normal > Random modules > Random lines within program modules > Random Lines and Modules (see Figure 1). Finally, Guerin and Matthews found that recall correlated highly

with comprehension, which was measured by having participants write a summary of the purpose of the program and how this purpose was achieved. The authors conclude that the superior semantic knowledge and comprehension skills of experts that allowed them to obtain better performance than novices could not be used when the lines of the program were randomised.

Summary. Several phenomena clearly stand out from these experiments. First, as expected, experts always do better than non-experts with executable code. Second, there is a robust correlation between recall and comprehension. Third, some evidence has been uncovered that experts have larger chunks than novices. Finally, the expertise effect is also present when the code has been scrambled in various ways. In the remainder of this article, we describe simulations with CHREST showing that this cognitive architecture captures the skill effect with randomised computer programs.

The CHREST Architecture

As mentioned above, the chunking theory has often been proposed, in its informal form, as an explanation of the skill effect found in memory tasks for computer programs. In the remainder of the paper, we wish to explore to what extent a computational implementation of the theory, which emphasises perceptual chunking, can account for the empirical data we have just reviewed.

CHREST (Chunk Hierarchy and REtrieval STructures; De Groot & Gobet, 1996; Gobet, 1993; Gobet & Lane, 2005; Gobet & Simon, 2000) is an expansion of the EPAM (Elementary Perceiver And Memorizer) cognitive architecture (Feigenbaum & Simon, 1984; Richman, Staszewski & Simon 1995). At the core of EPAM and CHREST lie mechanisms for encoding chunks into long-term memory (LTM) through the construction of a discrimination net and mechanisms for handling information in short-term memory (STM). Together, EPAM and CHREST have been used to account for domains such as verbal behaviour, chess memory, expert digit-span memory, use of multiple representations in physics, letter perception, spelling and acquisition of language (see Gobet et al., 2001, Gobet & Lane, 2005, for reviews).

CHREST consists of the following components: discrimination network, semantic LTM, and STM. STM, which consists of at most four chunks, is mostly a queue (first-in, first out). However, the largest chunk met at any point in time (the *hypothesis)*, is kept in STM until a larger chunk is met or constructed (see Gobet & Simon, 2000).

The net is grown by two EPAM-like learning mechanisms, *familiarisation* and *discrimination*. When a new object is presented to the model, it is sorted through the discrimination net. When a node is reached, the object is compared with the *image* of the node, which is its internal representation. If the image under-represents the object, new features are added to the image (familiarisation). If the information in the image and the object differ on some feature or some sub-element, a new node is created (discrimination).

Table 1 shows the key time parameters used with CHREST. These parameters are taken from previous work (Feigenbaum & Simon, 1984; De Groot & Gobet, 1996; Gobet & Simon, 2000) and are important in that they impose stringent constraints on how much information processing can be performed both during the training phase and during the presentation of the stimulus in the test phase. Note that creating an LTM chunk, adding a new link to a chunk, or familiarising a chunk occurs in parallel with the other operations.

CHREST incorporates mechanisms for incrementally creating schemas (known as templates in the theory), allowing information to be rapidly encoded in slots (Gobet & Simon, 2000). In this paper, we are primarily interested in how far perceptual chunks can account for skill effect in the recall of scrambled programs. Therefore, we did not use templates in the simulations.

Table 1: Main time parameters used in CHREST

Cognitive operation	Duration
creating an LTM chunk	8 s
familiarising an LTM chunk	2 s
placing a chunk into STM	50 ms
comparing two chunks	50 ms
carrying out a test in the discrimination net	10 ms

Simulations

For training and testing the model, a large collection of data was gathered from a variety of sources. Using the internet and some reference books, a corpus of about one hundred different FORTRAN programs was built.

Training Phase

During training, CHREST is given programs in FORTRAN—a naturalistic material—as input so that the vocabulary of the language as well as some sequences of items can be learnt. The lines of code had a mean length of 7 words. Elements (e.g. numbers, punctuation and other special characters) are recognised as distinct, individual items by the model. This type of input allows the model to build a discrimination net that encodes both the primitive items and legal strings from the computer language.

The same basic model is used to simulate different levels of ability; that is, only the amount of input is varied, and no other mechanisms or parameters are altered. This study will focus on the difference between *novices* and *experts*, who are simulated by passing CHREST either one program or a corpus of eighty-eight programs during training (these numbers were chosen arbitrarily).

As each item is passed to CHREST, the model constructs its discrimination net. The net initially starts with an empty root node. Primitive items are usually the first to be added to the net. Then, after a period of learning, the images at the nodes will come to represent sequences of items.

Test phase

Once the appropriate training had been undertaken (study of one program for novices and eighty-eight for experts), the novice and expert level models were tested using twelve new FORTRAN programs that were not included in the training set. Test programs were selected that did not include too many "print" statements and so that they were all of roughly the same number of lines and words. The simulations were run like an experiment with human participants, with a presentation time of five seconds per line. Various levels of randomisation were applied to the test programs before they were passed to the models. In addition to the conditions used in the studies reviewed above, we also thought it interesting to use a condition in which all elements of a program were randomised. There were therefore five conditions in total:

1. **Normal.** The sequence of the program is unaltered.
2. **Random Modules.** Segments of the program are randomised, but the line order within a segment is retained.
3. **Random Lines**. The lines of the program are randomised. Information within a line is unaltered.
4. **Modules and Lines**. Both modules and lines within modules are randomised.
5. **All Random**. All elements within a program are randomised, yielding a total randomisation.

To create the Random Modules, lines of code that acted as a meaningful unit of instruction were grouped. For example, lines belonging to declaration statements would be retained together as a module. For the All Random condition, the maximum and minimum line lengths of the original program were first noted; then, all the elements in the program were randomised and lines of random length were constructed, with the condition that the values fell between the original lengths.

Each model received two "practice" problems which were always the same. The first practice program was a Normal type and the second was an All Random type. The test programs were then presented. In addition to the practice problems, each model received two examples of each condition, thus making a total of 12 test programs. To control for random variation due to the order of programs in the learning set, CHREST was run with 40 simulations per skill level. The random order in which the programs appeared, the random order of the conditions and the randomisation of programs were all reset for each simulation.

For each program, CHREST read the program line by line, storing recognised chunks into STM, and, when applicable, using the following learning mechanisms. First, as described before, CHREST can add a chunk as a test to another chunk. It takes 8 seconds to carry out this discrimination operation. Typically, a new test is added to the hypothesis. Second, for chunks that have been in STM for at least 8 seconds, a new branch is added to access them by a novel path; this essentially means that episodic cues that permit access to this node are added to the discrimination net (Gobet & Simon, 2000). Such nodes can be recalled during the reconstruction phase even if they are no longer in STM.

During the recall phase, CHREST could output the information held in STM and in the nodes that had been created or for which new access links had been created. Recall was scored in the following way. A list of items that CHREST had recalled, in the order they were retrieved, was collected from the model. This recall list was matched alongside the original stimulus that was presented to CHREST. The first line in the stimulus was compared to the recall list and if the first items matched, then the lines were compared to find out how many of the items were recalled correctly before a mismatch occurred; both lines were then discarded and the next stimulus line compared to the recall. If the stimulus line did not match the recall line, then the line was discarded and the next one matched against the recall list, until all the stimulus lines were used. Not only does this method show how many items were recalled correctly, but it shows how many errors of commission the model made.

Results

Figure 2 illustrates the results of the simulations. We can see that for all cases, CHREST predicts a skill effect. These predictions are borne out by the data, with the exception of Guerin and Matthews' (1990) Module+Line condition, where no effect was found with the human participants. None of the studies reviewed incorporated the All condition, where the order of all elements of the code is randomised. Although the difference between the Novice and Expert models is small for the All condition, it is statistically reliable ($t(78) = 3.71$, $p < .001$). To test this counter-intuitive prediction of the model, we collected data with C programmers (n = 9) and novices (n = 9); the results have supported the prediction. Given that we found a skill effect with full randomisation, it is unclear why no such effect was found with Guerin and Matthews' Module+Line condition, which destroys less structure than our method.

While the simulations show a differential effect for the type of randomisation, this effect is limited to the All conditions vs. the other conditions. The model fails to capture the differential recall shown by humans from the Normal condition to the Module+Line conditions. It is likely that humans pick up semantic information from the modules, as proposed for example by Adelson (1981), which are beyond the essentially perceptual knowledge that this simplified version of CHREST can store.

Conclusion

The experimental studies reviewed in this paper clearly show that randomisation of aspects of computer code affects recall, while still preserving a skill effect in most cases. In

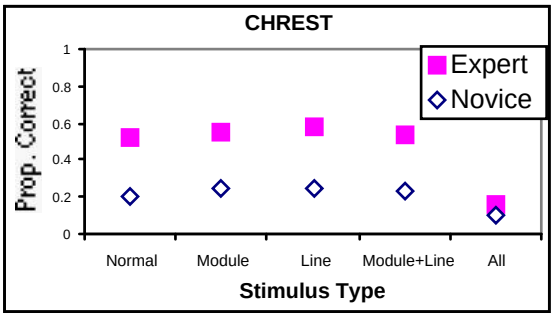


Figure 2. CHREST's memory for computer code as a function of level of expertise and type of randomisation.

order to account for these data, we have described a simplified version of the CHREST simulation model, where emphasis was given to perceptual chunking. We have shown that the model accounts for the skill differences with scrambled programs, which supports Simon and Gobet's (2000) contention that theories based on chunking mechanisms can account for skill effects in memory for computer programs. However, the model did not show differences between the randomisation conditions, as humans did. In this respect, the results differ to those obtained with chess, where it has been shown that CHREST is able to successfully capture recall differences with chess positions that were randomised in different ways (Gobet & Waters, 2003). A difference between the simulations in the two domains is that CHREST used templates with chess, but not with computer programs.

The presence of a skill effect even with randomised material has been demonstrated not only in chess and computer programming, but also in nearly all domains of expertise where this has been studied (Sala & Gobet, 2016). This finding strongly suggests that theories of expertise cannot only propose high-level and holistic mechanisms, but must also include some low-level mechanisms such as chunking to account for the empirical data. In this respect, CHREST is obviously on the right track. Further work will establish whether the presence of templates (Gobet & Simon, 2000) can help the model capture the differential recall of different types of scrambled programs, which is often claimed to tap into differences in high-level, semantic knowledge (e.g., Adelson, 1981).

References

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory and Cognition, 9*, 422-433.

Barfield, W. (1986). Expert-novice differences for software: Implications for problem-solving and knowledge acquisition. *Behaviour and Information Technology, 5*, 15-29.

Bateson, A. G., Alexander, R. A., & Murphy, M. D. (1987). Cognitive processing differences between novice and expert computer programmers. *International Journal of Man-Machine Studies, 26*, 649-660.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess.* Assen: Van Gorcum.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305-336.

Gobet, F. (1993). A computer model of chess memory. *Proceedings of 15th Annual Meeting of the Cognitive Science Society*, (pp. 463-468). Hillsdale, NJ: Erlbaum.

Gobet, F. (2016). *Understanding expertise: A multi-disciplinary approach*. London: Palgrave.

Gobet, F., & Lane, P. C. R. (2005). The CHREST architecture of cognition: Listening to empirical data. In D. Davis (Ed.), *Visions of mind* (pp. 204-224). Hershey, PA: IPS.

Gobet, F., Lane, P.C.R., Croker, S., Cheng, P.C-H., Jones, G., Oliver, I., & Pine, J. (2001). Chunking mechanisms in human learning. *TRENDS in Cognitive Sciences, 5*, 236-243

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1-40.

Gobet, F. & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24*, 651-682.

Gobet, F., & Waters, A. J. (2003). The role of constraints in expert memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 29*, 1082-1094.

Guerin, B., & Matthews, A. (1990). The effects of semantic complexity on expert and novice computer program recall and comprehension. *The Journal of General Psychology, 117*, 379-389.

McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organisation and skill differences in computer programs. *Cognitive Psychology, 13*, 307-325.

Richman, H. B., Staszewski, J., & Simon, H. A. (1995). Simulation of expert memory with EPAM IV. *Psychological Review, 102*, 305-330.

Sala, G., & Gobet, F. (2016). Experts' memory superiority for domain-specific random material generalises across fields of expertise: A meta-analysis. Manuscript submitted for publication.

Schmidt, A. L. (1986). Effects of experience and comprehension on reading time and memory for computer programs. *International Journal of Man-Machine Studies, 25*, 399-409.

Simon, H. A. & Gobet, F. (2000). Expertise effects in memory recall: A reply to Vicente and Wang. *Psychological Review, 107*, 593-600.

Ye, N., & Salvendy, G. (1994). Quantitative and qualitative differences between experts and novices in chunking computer software knowledge. *International Journal of Human-Computer Interaction, 6*, 105-118.

Feature Overlap in Action Sequence

Alexandra M. Stubblefield (a.stubblefield@wsu.edu)
Department of Psychology, PO Box 644820, Johnson Tower 233
Pullman, WA 99164 USA

Lisa R. Fournier (lfournier@wsu.edu)
Department of Psychology, PO Box 644820, Johnson Tower 233
Pullman, WA 99164 USA

Abstract

This study determined if features of an action plan held in working memory are activated to the same extent (consistent with serial memory theories) or in a gradient (consistent with theories that assume serial order is imposed prior to response selection). Two visual events (A and B) occurred in a sequence. Participants planned an action (3-finger, key sequence) to the first event (Action A) and maintained this action in working memory while executing a speeded response (1-finger key-press) to the second event (Action B). Afterwards, participants executed Action A. We manipulated whether Action B overlapped with the first, second or final feature of Action A, and examined the pattern of correct, Action B RTs at the different overlap locations by finger (index, middle, ring), as well as the error rates of both Action A and Action B. Results indicate that 3-finger sequences were not activated equally or in a gradient. Instead, feature activation reflected a serial position curve or a reverse serial position curve dependent on finger.

Keywords: feature overlap, cognitive interference, action planning, partial repetition costs

Introduction

Everyday actions such as reaching for a water glass or preparing a meal require action planning; one must decide what to do and when to do it (Keele, 1968; Lashley, 1951). In some cases action plans can be comprised of complicated sequences that must be planned, like preparing a meal (Rhodes, Bullock, Verwey, Averbeck, & Page, 2004). When carrying out these action sequences, some elements of the sequence need to be held in memory, while others are being executed (Logan, 2004; Schneider & Logan, 2006). Various theories provide different ideas as to how these action elements are represented in memory. There are theories of serial memory, which posit that all action features within an action plan are activated equally and simultaneously (Crump & Logan, 2010; Hartley & Houghton, 1996; Lashley, 1951; Rosenbaum, Inhoff, & Gordon, 1984). These theories assume that serial order may be imposed later, during response selection and execution. On the other hand, there are theories that postulate that earlier features within the action plan are more active than later ones, creating a gradient of activation in working memory (WM) that preserves serial order (Averbeck, Chafee, Crowe, & Georgopoulos, 2002; Page & Norris, 1998; Dell, Burger, & Svec, 1997; Rhodes et al., 2004). How the features of an action plan are represented in WM was investigated in the current study.

Research consistent with theories of serial memory has shown that highly skilled, hierarchically controlled tasks, such as speaking or typing, appear to activate all action features within an action plan in parallel (Rosenbaum, Inhoff, & Gordon, 1984; Crump & Logan, 2010; Logan, Miller, & Strayer, 2011). Crump and Logan (2010) gave skilled typists words five to seven letters long as a prime, followed by either a repeated exposure to the same word that typists had to re-type or a single-letter prompt they had to type. If the single-letter prompt was given, it was the first, middle, or last letter of the original word, or it was a completely unrelated letter. Results showed that the single-letter prompt was typed faster if the letter appeared in the original word, and that priming was greater for the first letter than for the middle or last letters - with no difference in priming between the middle and last letters. This suggests that the original word primed all the letters in the word sequence. The authors assumed that activation of the first letter was greater due to priming, likely because it needed to be typed first. They stated that since the middle and last letters were activated to a similar extent, that their results support the idea of equivalent parallel activation of features. They argue that their findings are consistent with models that suggest equivalent, parallel activation with serial order determined at response selection (Crump & Logan, 2010; Hartley & Houghton, 1996; Lashley, 1951; Rosenbaum, Inhoff, & Gordon, 1984).

In contrast, some models assume that serial order is determined by a gradient of parallel feature activation (Dell et al., 1997; Page & Norris, 1998; Rhodes et al., 2004). These models assume that tasks that are less practiced may produce a primacy gradient of parallel feature activation in which each successive feature in the sequence is activated less than the preceding one (Rhodes et al., 2004; Dell et al., 1997; Page & Norris, 1998). Neurophysiological evidence for this model is provided by Averbeck, Chafee, Crowe, & Georgopoulos (2002), who trained rhesus macaque monkeys to draw geometric shapes (e.g. triangle, rhombus) using a specific line segment order. The researchers first identified patterns of neural activity in the prefrontal cortex that corresponded to the drawing of each segment, and then recorded activity from these neurons just prior to and during the time in which the monkeys drew a shape. They found that prior to executing the segments, the neurons associated

with the different segments of the shape were activated in parallel, and the strength of activation was highest for neurons corresponding to the first segment drawn, followed by a decrease in neuronal activation corresponding to each subsequent segment drawn, revealing a gradient of activation of features within the action plans.

Other evidence supporting the idea that feature activation within an action plan may follow a gradient is provided by Fournier, Gallimore, Feiszli, and Logan (2014). They used a partial repetition paradigm (Stoet & Hommel, 1999). In this paradigm, two stimuli were presented in a sequence. Participants were instructed to plan an action to the first stimulus and hold it in memory while they perceived and responded to a second stimulus. After they responded to the second stimulus, they executed the response of the action plan maintained in memory. Previous research showed that the execution of an action plan can be delayed if there are features of that plan that partially overlap with features of another action plan currently maintained in WM, relative to when there is no overlap between action plans (e.g., Stoet & Hommel, 1999; Wiediger & Fournier, 2008). This delay in responding is referred to as partial repetition costs. Partial repetition costs are assumed to occur when a feature code from the current action plan reactivates (primes) the action plan maintained in WM (Hommel 2004; Fournier et al., 2014). Reactivating the action plan retained in WM leads to temporary confusion as to which action plan is relevant for the current task: the current plan or the plan maintained in WM (Hommel 2004; 2005; Mattson & Fournier 2008, Fournier et al., 2014). Fournier et al. (2014) found partial repetition costs when the current action overlapped with the first feature of the action sequence maintained in WM, but no significant costs when the overlap was on the last feature. Partial repetition costs were used to measure activation strength of each feature in the action feature sequence within an action plan. Larger partial repetition costs (or greater reaction times [RTs]) when overlap is on a particular feature, are indicative of higher activation of that feature compared to others in the action plan; lower costs (lower RTs) are indicative of less activation of that feature. The authors concluded that the features of an action plan are not activated to the same degree in WM. Rather, the first feature of the action plan is activated to a greater extent than the last feature, consistent with a gradient of activation.

However, a limiting factor of the Fournier et al. (2014) study is that there were only two action features representing the action plan maintained in WM (e.g., move joystick "left" then "up"). As a result, it is not clear whether this pattern of activation truly follows a gradient of activation (Dell et al., 1997; Page & Norris, 1998; Rhodes et al., 2004) or if there is just something special about the first feature (Dell et al., 1997; Crump & Logan, 2010). The present study investigated whether action plans are represented by a gradient of feature activation in memory by improving upon Fournier et al.'s design. Participants in this study maintained action plans that consisted of three features, as opposed to two features. The advantage of using a partial repetition paradigm to understand how action plans are represented in working memory is that it allows insight into the representation of an action plan prior to execution.

A modification of the partial repetition paradigm (Stoet & Hommel, 1999) was employed to determine if features of an action plan maintained in WM are represented by a gradient of activation or whether the first feature in the action sequence has a higher level of activation than other features in the sequence which do not differ in activation strength (no gradient). Participants planned an action to the first stimulus event (Event A) and maintained this action plan in WM while waiting for the presentation of a second visual event (Event B) that required an immediate speeded response. After participants executed the response to Event B, they executed the planned action for Event A. The main manipulation was whether response to Event B overlapped with the first, second, or third response in the action sequence planned for Event A. If features of an action plan maintained in WM are activated in a gradient, then response RTs to Event B (Action B) should differ based on whether Action B overlaps with the first, second, or third feature of the action plan maintained in WM (Action A). Specifically, Action B RTs should be slowest when it overlaps with the first feature, intermediate when it overlaps with the second feature, and fastest when it overlaps with last feature of Action A. If instead, there is not a gradient of activation but rather just something special about the first feature, there should be higher activation for the first feature relative to the second and third features, and equivalent activation between the second and third features.

Methods

Participants

Forty-four undergraduates from Washington State University participated for optional extra credit in psychology courses. This study was approved by the Washington State University Institutional Review Board, and informed consent was obtained. All participants were right-handed and had at least 20/40 uncorrected or corrected visual acuity assessed using a Snellen eye chart.

Apparatus

Instructions and stimuli were presented on a computer screen 61 cm from the participant. Key-press responses were collected using a standard keyboard. The keys 7, 8, and 9 on the keyboard were covered with white stickers, and were used to record responses from the index, middle, and ring finger of the right hand respectively. The space bar on the keyboard was pressed by the thumb of the right hand to initiate each trial and the pinky finger rested on the P key. The participants' right hand placement on the keyboard and space bar were consistent throughout the task.

Event A was a sequence of three different boxes lighting up. The three boxes were arranged in a horizontal array centered one half centimeter above a white fixation cross at the center of the screen. The boxes were each two

centimeter squares, spaced a half centimeter apart from one another. The sequence in which the three boxes lit up varied, and each box lit up once in the sequence (i.e. "left box–right box–middle box). Event A required a three key-press response using the index, middle and ring fingers of the right hand. Finger mappings for this response were always spatially compatible where the first box corresponded to the index finger, the middle box corresponded to the middle finger, and the last box corresponded to the ring finger. Participants planned and maintained an action for Event A in the reverse order in which the Event A stimuli appeared. For example, if the light stimulus appeared first in the right box, then in the middle box, followed by the left box, then the participant planned to press the key corresponding to the left box first (index finger), then the middle box (middle finger), followed by the right box (ring finger).[1]

Event B was a left arrowhead (<), a circle (O), or a right arrowhead (>) centered one half centimeter below the fixation cross. Each symbol measured two centimeters in height. Event B required a speeded double key-press response with the index, middle, or ring finger of the right hand, dependent on the particular symbol presented. Participants responded with a double key-press with the index finger to the right arrowhead, a double key-press with the middle finger to the circle, and a double key-press with the ring finger to the left arrowhead. These finger mappings were consistent across all participants.

Procedure

All stimuli appeared on a black background. The sequence of events for each trial is presented in Figure 1. A white fixation cross occurred in the center of the screen and remained throughout the trials. When initiating a trial by pressing the space bar, the fixation cross appeared for 1000ms followed by a horizontal array of three boxes above the fixation cross (Frame A_0). After this, Event A (Frames A_1, A_2, and A_3) appeared for 1200ms above the fixation cross, followed by the fixation cross alone for 1500ms. During this inter-stimulus interval, participants were instructed to plan their response (Action A) to the lighted box sequence (Event A). Event B then appeared for 50ms below the fixation cross followed by a blank screen which appeared until a response to Event B (Action B) was detected, or for 1300ms in the absence of a response.

1 The box sequence was chosen as a stimulus after many pilot studies were run that included stimuli with response associations that proved to be too complex for participants to learn. These included, but are not limited to, colored shapes and number sequences. It would be possible to train participants to provide sufficient practice for them to lean more difficult tasks; however it was important to the study that responses not be highly automatic. Additionally, participants responded to the reverse box order in attempt to make sure participants' working memory was being sufficiently engaged by the task, as pilot studies indicated that responding to the same order was too easy of a task to elicit significant interference effects.

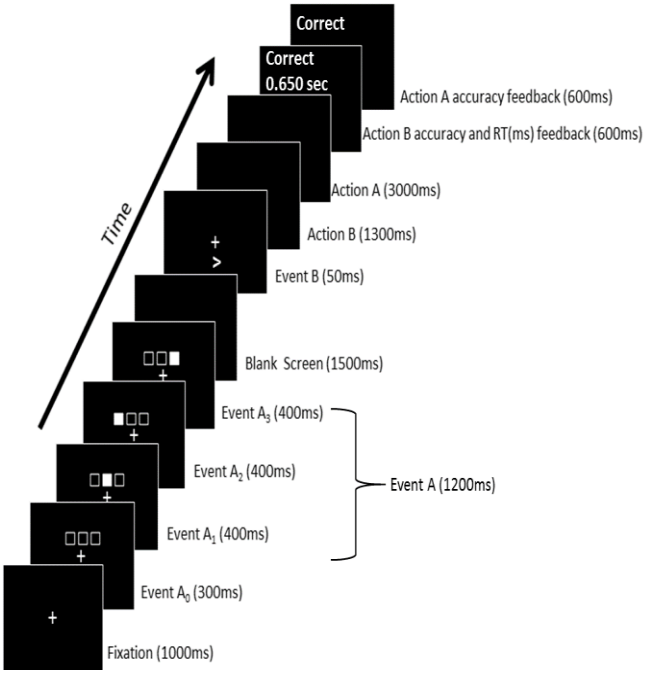


Figure 1: The sequence of events for each trial.

Following Action B, a blank screen appeared until Action A was executed, or up to 3000ms. Participants were instructed to respond to Event B (Action B) as quickly and accurately as possible. They were instructed to then execute their response to Event A (Action A) maintained in memory as accurately as possible. Following the execution of Action A, reaction time (RT) and accuracy feedback for Action B was presented (600ms), followed by accuracy feedback for Action A (600ms). The initiation screen, with the instructions "Press Space Bar to continue," then appeared until the next trial was initiated. The next trial began when participants pressed the space bar on the keyboard with their right thumb.

Both RT and accuracy for Action B were measured. Action B RT was measured from the onset of Event B until the first key press response was detected. Only accuracy was measured for Action A. Participants were instructed to not execute Action A until after executing Action B. Participants were also instructed not to use any finger movements or other external cues to help them remember Action A, but instead to hold this action plan in memory. Those who used body movements or other external cues (identified via post-task questionnaire) were eliminated from data analyses (eight participants), as they did not follow instructions. Additionally, any participants with an overall task accuracy below 80% were excluded (two participants). Data were analyzed for the remaining thirty-four participants.

Feature overlap between Action B and Action A was manipulated within participants. Action B, executed with the index, middle, or ring finger overlapped with the first, second, or third response feature of Action A. All possible

stimuli for Action A and Action B were equally paired together, and feature overlap (first, second, or third feature) occurred with equal probability in a random order in each block of trials. Participants completed 24 practice trials followed by ten blocks of 24 experimental trials.

Results

A 3 X 3 repeated measures analysis of variance (ANOVA) with the within-subjects factors of feature overlap (first, second, or third) and finger (index, middle, or ring) was conducted separately on the correct RTs for Action B, percentage of errors for Action B, and the percentage of errors for Action A. Analysis of mean RTs for Action B was restricted to trials in which Action B and Action A were both executed accurately. Analysis of Action B error rate was restricted to trials in which Action A was executed accurately. All comparisons between means were conducted using Bonferroni Pairwise Comparisons (p<.05). Figure 2 shows the mean correct RTs and correct error rates for Action B. As is evident in the figure, RTs based on feature overlap varied, and this RT pattern differed depending on which finger was used to respond.

Action B There was a significant main effect of feature overlap [F(2,32)=8.51, p<.01, ηp^2=.21], and of finger [F(2,32)=42.65, p<.01, ηp^2=.56], and a significant interaction between these two factors [F(4,30)=8.63, p<.01, ηp^2=.21] for RT. For the middle finger, the pattern of responses across feature overlap conditions (see Figure 2) resembled a serial position curve, while the pattern for the index and ring finger responses resembled a reverse serial position curve. Pairwise comparisons conducted on the main effect of finger showed that RT for the index and ring fingers were overall slower compared to the middle finger (ps<.01), while index and ring fingers did not differ from one another (p=.56). Moreover, pairwise comparisons conducted on the interaction showed the following. For the middle finger, RTs were significantly slower when Action B overlapped with Action A on the third feature (M=652ms) compared to the second feature (M=624ms; p<.01), and RT for the first feature overlap condition (M=640ms) was not significantly different than the second or third feature overlap conditions (p=.09, p=.19 respectively). For the index finger, RTs were significantly slower when Action B overlapped with Action A on the second feature (M=718ms) compared to the first feature (M=685ms; p<.01), and RTs for third feature overlap condition (M=703ms) were not significantly different than first (p=.17) or second (p=.13) feature overlap conditions. For the ring finger, RTs were significantly slower when Action B overlapped with Action A on the second (M=726ms) and third feature (M=705ms) compared to the first feature (M=686ms; p<.01; p=.05, respectively), and RTs were not significantly different between the second and third feature overlap conditions. These results suggest that the amount of feature activation within the planned action sequence was not equal across the first, second, and third feature overlap conditions. Also, the

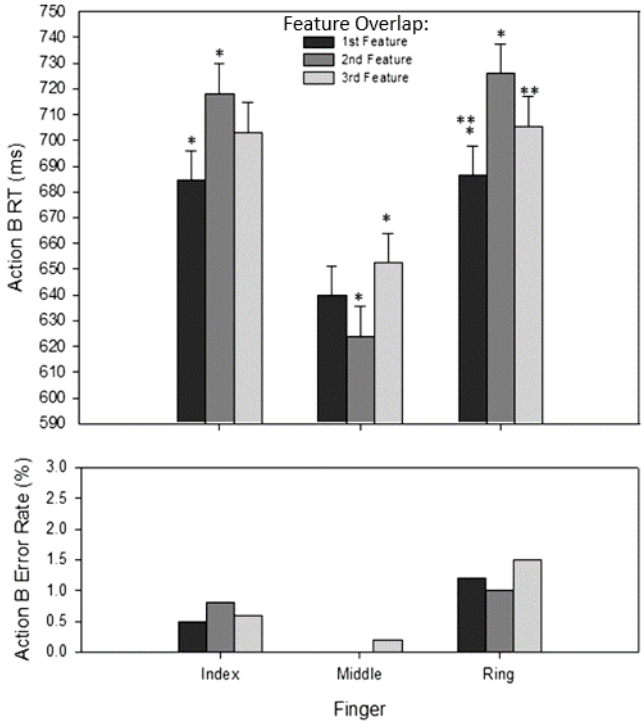


Figure 2: Feature Overlap. Action B correct reaction times (RTs) and percent error rates for each Action B finger response (index, middle, and ring) when Action B overlapped with Action A on the first, second, or third feature.

finger used to respond to Action B influenced the speed of Action B responses and the pattern of Action B responses found across feature overlap conditions.

Mean error rate was small, only 1%. There was a significant effect of finger [F(2,32)=6.31, p<.05, ηp^2=.16] for error rate. However, the main effect of feature overlap (F<1) and the interaction between feature overlap and finger (F<1) were not significant. Pairwise comparisons conducted on the main effect of finger showed that error rate for the middle finger was significantly smaller than that for the index finger (p=.04) and the ring finger (p=.01). These results suggest that the Action B RT interpretations above are not due to a speed-accuracy tradeoff.

Action A Mean error rate was 12%. There was a significant main effect of feature overlap [F(2,32)=5.72, p<.01, ηp^2=.15] and a significant interaction between feature overlap and finger [F(4,30)=2.52, p=.05, ηp^2=.07] for error rate. There was no main effect of finger [F(2,32)=0.20, p=.82, ηp^2=.006] on error rate. Pairwise comparisons conducted on the interaction showed the following. For the ring finger, error rates were significantly greater when Action B overlapped with Action A on the first feature (M=13%) and second feature (M=17%) compared to the third feature (M=9%; p=.05; p<.01, respectively). For the index finger, error rates did not significantly differ across the first (M=12%), second (M=15%), or third (M=10%)

feature overlap conditions (*p*s>.08). For the middle finger, error rates also did not significantly differ across the first (*M*=13%), second (*M*=11%), or third (*M*=12%) feature overlap conditions, (all *p*s=1.0). The error rate in recalling Action A across the feature overlap conditions resembles a serial position curve for the middle finger and a reverse serial position curve for the index and ring fingers – although this pattern was not significant. This error rate pattern was similar to the pattern of RTs found for Action B. This suggests that both responding to Action B and recalling Action A were similarly affected by feature overlap.

Discussion

In summary, results revealed differences in Action B performance when Action B overlapped with Action A on the first, second, or third feature in the Action A sequence. Also, the differences in Action B performance across the action overlap conditions varied depending on the finger involved in executing Action B. The results suggest that features of the action sequence maintained in WM have different levels of activation. However, the predicted gradient of activation was not found. Instead the pattern of activation resembled a serial position curve for actions executed with the middle finger, and a reverse serial position curve for actions executed with either the index finger or ring finger. Importantly, variable activation across action features within an action plan is not consistent with serial memory models that assume equal activation of features within an action plan (Crump & Logan, 2010; Hartley & Houghton, 1996; Lashley, 1951; Rosenbaum, Inhoff, & Gordon, 1984), or those models that assume features are activated in a gradient (Averbeck, Chafee, Crowe, & Georgopoulos, 2002; Page & Norris, 1998; Dell, Burger, & Svec, 1997; Rhodes et al., 2004).

The current findings are consistent with Fournier et al. (2014) in that they also showed unequal activation between conditions of feature overlap on the first feature and second feature. More specifically, they found higher activation of the first feature of an action plan than the second, which is consistent with the results of the middle finger in the present study. However, the pattern for the index and ring fingers showed the opposite pattern. It is unclear, given the design of their study, if the response pattern they got would have represented a serial position curve if they had included a third feature overlap condition in their study.

There were some differences in design between the current study and Fournier et al. (2014) that may account for the different feature activation patterns. In their study, the design used a more fluid joystick response to execute the action being maintained in working memory, compared to the more distinct key-press movements of this study. For example, actions requiring a left and upward response were carried out by moving a joystick with the right hand to the left and then upward. In the current study, fingers were mapped to particular keys, and required three very distinct key-presses using different digits on the right hand. It is possible that fluid movements within an action plan are represented differently than discrete movements within an action plan. For example, action features representing fluid movements may be more strongly associated within a single action plan (e.g., Fournier & Gallimore, 2013). Also, in the Fournier et al. study, participants were taught to associate certain stimuli with directional movements (e.g. left-up), and those directional movements were spatially congruent with the physical movement they executed with the joystick. This was also true in the current study, however, most participants (87%) subjectively reported (via questionnaire) representing the sequence of finger movements in memory as numbers (index as 1, middle as 2, ring as 3). It is possible that differences in how the action plan is represented could have contributed to differences in response patterns.

The different Action B response patterns found across feature overlap conditions dependent on response finger may be simply due to the faster Action B responses for the middle finger versus the index and ring fingers. This response speed difference may be attributed to differences in discriminability among the stimuli mapped to the middle versus the index and ring fingers. It may also be due to the differences in the speed at which one can execute a motor response with the middle finger versus the index and ring fingers. Why the different response patterns occurred across feature overlap conditions based on the overall speed of the Action B response is an interesting question for future research. In a follow-up experiment (not reported here), Action B was reduced to two responses (index and ring finger based on direction of arrowheads), and the same feature overlap patterns for the index and middle fingers were found. Other follow-up experiments investigating the different feature overlap patterns across fingers (i.e., whether due to the actual finger or the spatial representation of finger response) are in progress.

One important limitation of the current study is that we could not be certain that participants represented the Action A sequence (the reverse box order of Event A) in WM or the Event A sequence (the forward box order of Event A) in WM prior to executing Action B. This is currently being addressed in follow-up experiments via post-task survey. Also, we do not know if the activation patterns found in this study are limited to the stimulus and responses used in this study or whether they will generalize to other types of stimuli and responses.

The serial-position curve pattern of activation found across the different feature overlap conditions for middle finger responses is similar to that found in a comparable study by Mushiake, Saito, Sakamoto, Itoyama, & Tanji (2006), though the patterns found across the index and ring fingers both show the opposite pattern. Importantly, the features within the action plan maintained in WM showed variable activation, indicating that these action features maintained in WM were not equally activated.

In summary, this study shows variable patterns of activation across multiple features within an action plan maintained in working memory. This study also provides additional evidence that serial order is represented prior to

response execution in that the different patterns of activation across features were not equal. This contradicts both serial order models that predict equal activation across all features, and those that predict a gradient of activation across features. Specific patterns of feature activation might be dictated by task goals, the response requirement of the task, or how one is representing the task goals themselves.

References

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences, 99*(20), 13172-13177.

Crump, M. J., & Logan, G. D. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1369-1380.

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*(1), 123-147.

Duncan, J., & Humpheys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433-458.

Fournier, L. R., & Gallimore, J. M (2013). What makes an event: Temporal integration of stimuli or actions? *Attention, Perception and Psychophysic, 75(6),* 1293-1305.

Fournier, L. R., Gallimore, J. M., Feiszli, K. R., & Logan, G. D. (2014). On the importance of being first: Serial order effects in the interaction between action plans and ongoing actions. *Psychonomic Bulletin & Review, 21*(1), 163-169.

Hartley, T., & Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. *Journal of Memory and Language, 35*(1), 1-31.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*(11), 494-500.

Hommel, B. (2005). How much attention does an event file need? *Journal of Experimental Psychology: Human Perception and Performance, 31*(5), 1067-1082.

Keele, S. W. (1968). Movement control in skilled motor performance. *Psychological Bulletin, 70*(6p1), 387-403.

Lashley, K. (1951). The problem of serial order in behavior. In L.A. Jeffress (ed.), *Cerebral mechanisms in behavior*. Ney York: Wiley.

Logan, G. D. (2004). Working memory, task switching, and executive control in the task span procedure. *Journal of Experimental Psychology: General, 133*(2), 218-236.

Logan, G. D., Miller, A. E., & Strayer, D. L. (2011). Electrophysiological evidence for parallel response selection in skilled typists. *Psychological Science, 22*(1), 54-56.

Mattson, P. S., & Fournier, L. R. (2008). An action sequence held in memory can interfere with response selection of a target stimulus, but does not interfere with response activation of noise stimuli. *Memory & Cognition, 36*(7), 1236-1247.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., & Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron, 50*, 631-641.

Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review, 105*(4), 761-781.

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science, 23*(5), 699-746.

Rosenbaum, D. A., Inhoff, A. W., & Gordon, A. M. (1984). Choosing between movement sequences: A hierarchical editor model. *Journal of Experimental Psychology: General, 113*(3), 372-393.

Schneider, D. W., & Logan, G. D. (2006). Hierarchical control of cognitive processes: Switching tasks in sequences. *Journal of Experimental Psychology: General, 135*(4), 623-640.

Stoet, G., & Hommel, B. (1999). Action planning and the temporal binding of response codes. *Journal of Experimental Psychology: Human Perception and Performance, 25*(6), 1625-1640.

Wiediger, M. D., & Fournier, L. R. (2008). An action sequence withheld in memory can delay execution of visually guided actions: The generalization of response compatibility interference. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1136-1149.

Modeling Adaptation to a Novel Accent

Kasia Hitczenko[1] (khit@umd.edu)
Naomi H. Feldman[1,2] (nhf@umd.edu)
[1]Department of Linguistics and [2]Institute for Advanced Computer Studies
1401 Marie Mount Hall, University of Maryland, College Park, MD 20742

Abstract

Listeners quickly adapt to novel accents. There are three main hypotheses for how they do so. Some suggest that listeners expand their phonetic categories, allowing more variability in how a sound is pronounced. Others argue that listeners shift their categories instead, only accepting deviations consistent with the accent. A third hypothesis is that listeners both shift and expand their categories. Most work has supported the category expansion hypotheses, with the key exception of Maye et al. (2008) who argued for a shifting strategy. Here, we apply the ideal adaptor model from Kleinschmidt & Jaeger (2015) to reexamine what conclusions can be drawn from their data. We compare adaptation models in which categories are shifted, expanded, or both shifted and expanded. We show that models involving expansion can explain the data as well as, if not better than, the shift model, in contrast to what has been previously concluded from these data.

Keywords: accent adaptation; speech perception

Speech is highly variable. A given sound or word can be pronounced differently depending on the context, the speaker's accent or gender, and the discourse situation, among many other factors. In order for successful communication to take place, people's speech recognition system must adapt to this variability. Previous work has shown that listeners adapt to unfamiliar speech quickly. Although unfamiliar speech results in an initial processing cost, this cost generally disappears after a few minutes of exposure (Clarke & Garrett, 2004).

How do listeners adjust their sound categories in response to newly introduced variability? To address this question, we consider one example of adaptation to variable speech: adult perception of accented speech. In accented speech perception, the variability stems from the fact that the speaker and listener have different accents, so the speaker's pronunciation can differ, often substantially, from what the listener expects.

There are three main hypotheses about how individuals adapt their sound categories in response to accented speech. The *Expand* hypothesis suggests that individuals relax their categories, allowing more flexibility in how a particular category is produced. The *Shift* hypothesis suggests that individuals shift their categories, only allowing more flexibility in the direction of the accent they heard. Finally, the *Shift and Expand* hypothesis suggests that individuals do both: they allow more flexibility in how a given sound is produced, but do so around a shifted mean. Most work has found evidence that category expansion is involved in adaptation, either with a shift (*Shift and Expand*) or without (*Expand*).

Evidence for the *Expand* hypothesis comes primarily from experiments in which adults were exposed to an unfamiliar accent and, at test, had adapted to changes that were not present during the initial exposure (Schmale et al., 2012). In other words, they accepted variability that they had no evidence for. For example, listeners who were exposed to an accent in which back vowels were lowered accepted items that were raised versions of English words as words in a lexical decision task, despite not being familiarized with instances of this kind (Weatherholtz, 2015). In another study, listeners accepted large deviations in how particular vowels were pronounced after only being exposed to small deviations in pronunciation (Witteman et al., 2010). In addition, after being presented with an accent in which stops were devoiced syllable-finally, listeners accepted mispronunciations where stops were devoiced syllable-initially, even if they were never devoiced in this position during the exposure period (Eisner et al., 2010). Finally, toddlers exposed to an accent where [a] was produced [æ] also accepted [ɛ] for [a] (White & Aslin, 2011). In all of these cases, adults and children alike seem to be allowing more flexibility in how a sound is produced, even if it does not align with what they previously heard in this accent.

Although the *Shift and Expand* hypothesis has not been directly studied in the accent adaptation literature, it is consistent with many of the findings that support the *Expand* hypothesis. These findings mostly provide evidence that categories expand during adaptation, without considering whether they also shift. In addition, recent work suggests listeners use this strategy in other related instances of category adaptation, including phonetic recalibration and selective adaptation, so perhaps they are also using it in accent adaptation (Kleinschmidt & Jaeger, 2015). When adults are played sounds that are ambiguous between two categories (e.g. /b/ and /d/) and then shown visual stimuli that disambiguate them, they adapt their categories in accordance with the visual stimuli (phonetic recalibration). On the other extreme, when people are repeatedly played prototypical instances of a category (e.g. /b/), they adjust their categories in a way that makes them less accepting of non-prototypical instances of this category (selective adaptation). A model in which categories are both shifted and expanded in response to speech is able to explain these results, suggesting that adaptation involves both shifting and expanding.

Most work on accent adaptation has found evidence for category expansion, but a key finding by Maye et al. (2008) stands as an exception. In their experiment, participants listened to passages of the Wizard of Oz, with all of the front vowels lowered (e.g. 'witch' was pronounced 'wetch'). After just twenty minutes of exposure, participants were more likely to accept items that were mispronounced with lowered vowels (e.g. 'wetch') as being words, but were no more likely to judge items mispronounced with raised vowels (e.g. 'weech') as words. From this result, Maye et al. (2008) intuit that people

shift their categories, rather than expand them.

In this work, we model Maye et al.'s data to see whether their conclusion is warranted. We apply the ideal adaptor model from Kleinschmidt & Jaeger (2015) to compare three adaptation models: one in which categories are shifted, one in which they are expanded, and one in which they are both shifted and expanded. We compare these to a fourth model that directly learns the vowel mappings (e.g. that [i] is produced [ɪ] in this dialect). We show that the data from Maye et al. (2008), which have typically been used to argue against expansion hypotheses, can be captured by all four hypotheses. This suggests that the Maye et al. (2008) results may not provide strong evidence for the *Shift* hypothesis after all.

We begin with a description of the Maye et al. (2008) experiment, before describing our simulations.

Maye et al. (2008)

Maye et al. (2008) experimentally investigated how adults adapt their sound categories in response to accented speech, asking whether people shift or expand their vowel categories. The experiment was conducted in two sessions that were spaced a few days apart. In Session 1, participants listened to twenty minutes of the Wizard of Oz spoken in their accent by a synthesized male voice. The participants then took part in a lexical decision task, described below, in which they decided whether various test items they heard were words or not.

In Session 2, the same participants listened to the same passage; however, in this session, all of the front vowels were lowered. The vowel [i] was pronounced as [ɪ], [ɪ] was pronounced as [ɛ], [ɛ] was pronounced as [æ], [æ] was pronounced as [a] and [a] was unaltered, resulting in a merger between [æ] and [a]. None of the back vowels were lowered. After twenty minutes of exposure to this accented speech, in which participants heard fragments like 'the weckud wetch of the wast,' instead of 'the wicked witch of the west,' participants took part in the same lexical decision task from Session 1. Participants' responses between the two sessions were compared to see how exposure to accented speech affected their judgments.

Six types of test items were included in the lexical decision task. *Witch* items are words under the participants' accent, but are not words under the lowered accent ('witch' is a lowered version of 'weech', which is not a word in standard English). *Wetch* items are words in the lowered accent, but are not words in the standard American accent ('wetch' is a lowered version of 'witch'). *Weech* items are words in standard American English mispronounced with raised front vowels ('weech' is a raised version of 'witch'). The remaining three test item types did not contain any front vowels, so these would be pronounced identically under both accents. *Girl* items are words under both accents. *Loke* items are non-words under both accents, but are lowered versions of a word in standard English ('loke' is a lowered version of 'look'). *Tuke* items are non-words under both accents, but are raised versions of a word in standard English ('tuke' is a raised version of 'took'). Participants were presented with test items of these types - half of them occurred in the story and half of them did not - and were asked whether they were a word or not.

Maye et al. (2008) argue that if people are generally more flexible with how categories are produced (*Expand* hypothesis), then the participants should show an increase in endorsement rates for both *Weech* (raised) items and *Wetch* (lowered) items. On the other hand, they suggest that if listeners shift their categories in the direction of the accent they hear (*Shift* hypothesis), then participants should only show an increase in endorsement rates for *Wetch* items, not *Weech* items.

The results from their experiment are shown as part of Figure 2. Participants show an increase in endorsement rates for *Wetch* items, but not *Weech* items. Based on this finding, Maye et al. (2008) conclude that people are shifting their categories, not expanding them. In this paper, we model the data from this experiment and find that their results do not reliably support the *Shift* hypothesis over either expansion hypothesis (*Shift and Expand* or *Expand*). In the next section, we outline the modeling framework - the ideal adaptor framework - that we use to test these three hypotheses.

The Ideal Adaptor Model

We use the ideal adaptor model from Kleinschmidt & Jaeger (2015) to instantiate the *Shift*, *Expand*, and *Shift and Expand* hypotheses. We generalize the model to two dimensions, representing vowels as two-dimensional Gaussians. The vowels are defined by their first and second format values, which are continuous measures that are standardly used to distinguish between vowels. At any given point, the model has a set of categories, each of which is defined by a mean and covariance. With exposure to new speech, these categories change (the mean and covariance get updated) by taking a weighted average of the previous category information and the newly encountered data. We formalize shifting a category as changing its distribution's mean based on new evidence. We formalize expanding a category as updating the covariance of its distribution based on new evidence.

Broadly speaking, if the newly encountered data point lies near the category that generated it, then the category mean and covariance will be minimally changed, but if the newly encountered data point lies far from the category that generated it, then the mean and covariance will be more substantially changed to account for that data point. Crucially, the properties of this model are such that we can manipulate how much confidence we place on the previous mean and covariance, to manipulate how much the mean and covariance get updated. If we have high confidence in the previous mean (or covariance), then it will be less affected by new data, whereas if we have low confidence in the previous mean (or covariance), then the new data will affect the parameters more. We take advantage of this property to implement three of the hypotheses we would like to test. For the *Shift* model, we place high confidence in the previous covariance and low confidence in the previous mean. For the *Expand* model, we place low confidence in the previous covariance and high confidence in the previous

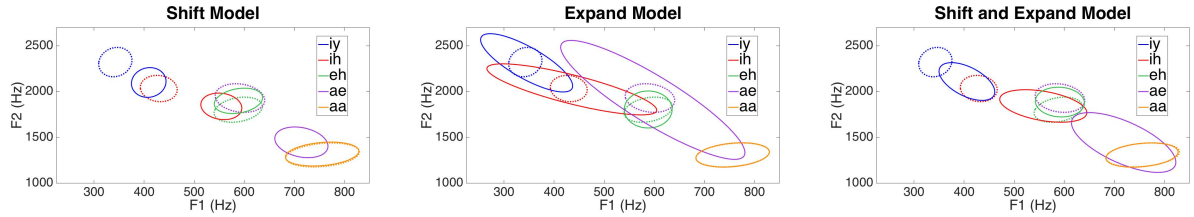


Figure 1: Original categories (dotted lines) and updated categories (solid lines) for each model.

mean. Finally, for the *Shift and Expand* model, we place low confidence in both the previous mean and covariance.

We assume a Normal Inverse Wishart prior over the category means and covariances based on speaker productions and assume that these means and covariances are updated as follows. The updated mean of a category after seeing N new acoustic values, μ_n, is the sum of the previous mean of the category, μ_0, weighted by the confidence in the previous mean, K, and the mean of the new acoustic values, $\bar{x}$, weighted by how many of them there are,

$$\mu_n = \frac{K}{K+N}\mu_0 + \frac{N}{K+N}\bar{x} \tag{1}$$

As confidence in the previous mean, K, increases, more weight will be placed on the previous mean than the mean of the new instances and the category mean will not change very much.

Similarly, the updated covariance of a category after seeing N new acoustic values, Σ_n, is the sum of the previous covariance, Σ_0, weighted by the confidence in the previous covariance, V, and the covariance contributed by the new acoustic values, weighted by the number of new values observed, N. The covariance contributed by the new acoustic values is a sum of the covariance of the new acoustic values, s^2, and a weighted term that accounts for the deviation of the mean of the new acoustic values from the previous mean,

$$\Sigma_n = \frac{V}{V+N}\Sigma_0 + \frac{N}{V+N}\left(s^2 + \frac{K}{K+N}(\mu_o - \bar{x})(\mu_o - \bar{x})^\top\right) \tag{2}$$

As confidence in the previous covariance, V, increases, the previous covariance will be weighted more heavily and the category covariance will not be updated as much. We use this ideal adaptor framework to implement our three adaptation models (see Kleinschmidt & Jaeger (2015) for more details of the generative model). The next section explains how we simulate the experiment performed by Maye et al. (2008).

Simulation of Accent Adaptation Experiment

To simulate the experiment, we first initialize the models with vowel categories based on standard American English, then update their categories based on the acoustic values participants would have heard in both experimental sessions, and, based on each model's updated categories, simulate the lexical decision task participants took part in. In the following sections, we outline each of these components.

Category Initialization

Each model was initialized with a category set that corresponded to a standard American English speaker. The standard accent categories were estimated based on data reported by Hillenbrand et al. (1995), in which participants produced vowels in a "h_d" context (e.g. 'head,' 'had,' 'heed'). Because the speech in the experiment was spoken by a male voice, we limited our estimates to vowels produced by male speakers. For each vowel type, there were 45 productions, each produced by a different speaker. For each vowel category, we took the 45 productions in F1-F2 space and calculated the mean and covariance of these 45 data points, which we used as the mean and covariance of the Gaussian distribution representing that category. These were the categories the models started with (the categories corresponding to the standard accent), shown in dotted lines in Figure 1.

Familiarization Data

We approximated what the participants in the experiment heard in both of the experimental sessions. Based on average reading times, the participants would have heard about 2500 words spoken during the 20-minute story excerpt (Olive et al., 1993). To get the actual items, we automatically transcribed all of the words in the Wizard of Oz, using the CMU Pronunciation Dictionary (Weide, 1998), and then manually verified that the results were accurate. We sampled 2500 words in proportion to how frequently they occur in the story. The same words were presented to the model in Session 1 and Session 2. For each vowel in the presented words, a particular formant-value pair was sampled. The distribution from which the vowel token was sampled depended on the session and the frontness of the vowel. For both front and back vowels in Session 1, a token was sampled from the distribution corresponding to that vowel in standard American English. For Session 2, front vowels were sampled from the standard American English distribution corresponding to the lowered vowel (for example, [i] tokens were sampled from the [ɪ] category) and for back vowels, the same formant values were repeated from Session 1. Throughout the experiment, the test item vowels were sampled from the same underlying distributions, which represent the (unchanged) sound categories of the standard American dialect. The listeners' categories were updated based on this exposure.

Category Adaptation

For the *Shift*, *Expand*, and *Shift and Expand* hypotheses, we update the models' categories according to (1) and (2). We also implement a fourth model, in which we update each category by setting its mean and covariance to the mean and covariance of its lowered version (*Vowel Mapping* model). For example, the updated [i] category will have the mean and covariance of

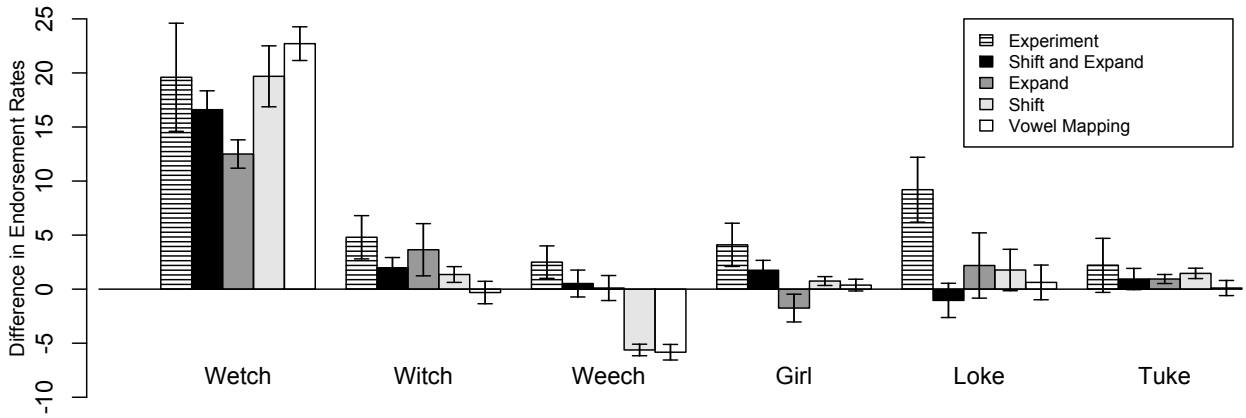


Figure 2: Difference in endorsement rates between Session 1 and Session 2 from Maye et al. (2008) and models.

the original [ɪ] category, the updated [ɪ] category will have the mean and covariance of the original [ɛ] category, and so forth.

There is some evidence reviewed in Kleinschmidt & Jaeger (2015) that listeners build different sets of categories for different groups of speakers. For example, people could learn that adult men speak differently than toddlers and instead of representing speech by both groups as identical, they could build two sets of categories to capture their speech differences. They might have one /i/ category corresponding to adult males and another /i/ category corresponding to toddlers.

Based on these results, when the model encounters the accented speech from Session 2, a new set of categories is built to represent the accented speech. These categories are initialized where the previous categories were, but only these newly-formed categories are updated in response to this new type of speech. At the end of the experiment, the model has one set of categories for speech spoken in a standard accent and another set for speech spoken in a lowered accent.[1]

As discussed above, we manipulate the speed of the adaptation by choosing particular values of K and V (confidence in previous mean and variance). Based on parameter settings found to capture experimental data in Kleinschmidt & Jaeger (2015), we set both K and V to 100 for the *Shift and Expand* model, we set K to 100, but V to 10000 for the *Shift* model, and we set K to 10000 and V to 100 for the *Expand* model.

Lexical Decision Task

After the categories are updated in Session 1 and Session 2, a lexical decision task is simulated. We adapt the model of lexical decision developed by Norris (2006) for written word recognition tasks to the problem of spoken word recognition tasks. The goal in a lexical decision task is to determine whether a test item is a word or not, not to identify the particular word. As a result, inference consists of inferring the word status (w) of a test item given its form. To do so, we apply Bayes' rule to calculate the relative probability that a test item is a word, given the acoustic input, x,

$$P(w = word\,|\,x) = \frac{P(x\,|\,w = word)P(w = word)}{\sum_w P(x\,|\,w)P(w)} \quad (3)$$

That is, the posterior probability that the test item is a word is proportional to the product of the probability that the particular acoustic value in question was generated from a category that would make the test item a word (likelihood) and the prior probability that a test item is a word.

Because there are two category sets at the end of the Session 2 (one set corresponding to unaccented speech and another corresponding to the accented speech), we need to sum over both of these category sets (c=standard and c=lowered),

$$P(w = word\,|\,x) = \frac{\sum_c P(x\,|\,w = word, c)P(w = word)P(c)}{\sum_{w,c} P(x\,|\,w, c)P(w)P(c)} \quad (4)$$

This is because we need to incorporate all possible ways that the item could have been generated. For example, 'wetch' could be an unaccented pronunciation of the non-word 'wetch' or it could be an accented pronunciation of the word 'witch' and both of these possibilities need to be considered.

In order to calculate the probability in (4), we need the values of $P(x\,|\,w = word, c)$, $P(w = word)$, and $P(c = standard)$. We consider each of these in turn, beginning with $P(x\,|\,w = word, c)$. In theory, we need to sum over all possible words, L, that the test item could be an instance of to get this probability. For example, if we want to calculate the posterior probability of 'wetch' being a word, we need to sum the possibility that this was a mispronunciation of the word 'watch' under standard English, a mispronunciation of 'witch' under standard English, a mispronunciation of 'watch' under accented English, etc. Because most of the elements in this sum are close to zero, we make the simplifying assumption that participants are only considering the possibility that one specific word or one specific non-word generated the input (5). All of the test items are either a word, or based on a word (e.g. 'wetch' is a lowered version of 'witch'), so we make use of the most relevant word, l^*, and non-word in our calculations. For example, if participants hear 'witch,' they infer whether this came from

[1] We also ran a version of each model where we directly updated the initial categories, without building a new set of category to represent the accented speech. These fail to fit the data because they predict a substantial decrease in the endorsement of *Witch* items, which is not observed in the experiment.

Model	P(c = standard)	P(w = word)
Shift	0.75	0.9
Expand	0.9	0.9
Shift and Expand	0.9	0.9
Vowel Mapping	0.95	0.9

Table 1: Parameter setting that produced results that best matched Maye et al. (2008) for each model.

Model	Log-Likelihood
Shift	-394.82
Expand	-388.72
Shift and Expand	-388.04
Vowel Mapping	-395.97

Table 2: Log-likelihood of models given Maye et al. data. Less negative values indicate a better match with the data.

a word or a non-word, by considering the probability that it was generated from the word 'witch' or the non-word 'weech.'

$$P(x \mid w = word, c) = \sum_{l \in L} P(x \mid l, c) \approx P(x \mid l^*, c) \quad (5)$$

To get the probability of the phonetic form of the word given the particular lexical item and category set, $P(x \mid l^*, c)$, we take the vowel token in the word in F1-F2 space and calculate the value of the Gaussian probability density function of the relevant vowel category at this point. This probability is highest near the participant's category means.

The remaining two terms are free parameters: one corresponds to the probability of a test item being a word, $p(w = word)$, and the other corresponds to the probability of using the standard English set of categories, $p(c = standard)$. To deal with these two free parameters, each version of the model was run on all combinations of these two parameters, ranging from 0.05 to 0.95 by increments of 0.05, for a total of 200 combinations. Because there is quite a bit of randomness involved throughout the process, the model was run twenty times on each parameter combination and the results were averaged across these twenty runs.

Putting this into (4), we get the relative probability of the input, x, being a word versus not. The model classifies an item as a word or non-word by sampling from its posterior distribution, $P(w \mid x)$.

To summarize, we update the model's categories in one of three ways: by shifting them, expanding them, or shifting and expanding them. In addition, we implemented a fourth model that sets categories to their lowered version. We compare the four models' results to human performance to see which can capture the experimental findings.

Simulation Results

The front categories before and after they are updated are shown in Figure 1. The next two sections compare the models' performance to the Maye et al. (2008) results.

Difference in Endorsement Rates

Figure 2 shows the difference in endorsement rates between Session 1 and Session 2 from the experiment and for each model. The figure shows each model's results from the parameter settings that had the best match with the original data (shown in Table 1). This was determined by calculating the log-likelihood of the model (i.e. the probability of the differences in endorsement rates observed in the experiment given the model). These results are shown in Table 2 - less negative log-likelihoods indicate better matches. Because all four models have the same number of free parameters, model comparison methods like the Bayesian information criterion (BIC) reduce to simply comparing the model log-likelihoods.

Overall, comparing the log-likelihoods, all three adaptation hypotheses outperform the *Vowel Mapping* model. Qualitatively, all four models are able to capture the increase in endorsement rates for *Wetch* items between sessions, though the *Expand* model underestimates this increase. The *Expand* model is, however, also able to capture the slight increase in endorsement rates observed for *Weech* items.

The *Shift* and *Vowel Mapping* models have the worst log-likelihoods, mainly because they both incorrectly predict a slight decrease in endorsement rates for *Witch* items between the two sessions. This happens because the free parameter settings that result in an increase in endorsement rates for *Wetch* items (i.e. high reliance on the new category set) also result in a decrease in endorsement rates for *Witch* items. Although this tension exists for all of the models, the *Expand* and *Shift and Expand* models overcome this because of their high-variance updated categories that continue to overlap with the original category. The *Shift* and *Vowel Mapping* models share the property that their updated categories do not overlap with the original categories, which explains their similar predictions.

It is worth noting that none of these models capture the *Loke, Tuke,* or *Girl* results very well. The participants seem to be generalizing the lowering of vowels to back vowels to a certain extent, which our models cannot capture because back vowels are left completely unaltered.

Overall, despite the conclusions that have been drawn from these results, all three adaptation processes (*Shift*, *Expand*, and *Shift and Expand*) can explain them to approximately the same degree, with the adaptation processes involving category expansion performing slightly better.

Absolute Endorsement Rates

Figure 3 shows the absolute endorsement rates for each of the test item types before and after Session 2 for the participants and the *Shift and Expand* model. The model underestimates people's absolute endorsement rates for all of the items, except *Girl* and *Witch* items. In particular, before they are even exposed to the new speech, participants accept *Wetch* items as words 39% of the time, whereas the model barely reaches this endorsement rate after updating its categories with the unfamiliar accent. Similarly, the model accepts *Weech* items around 35% of the time, but participants accept them as words between 60-70% of the time. Participants accept *Loke* and *Tuke* items about 30% of the time, but the model accepts them

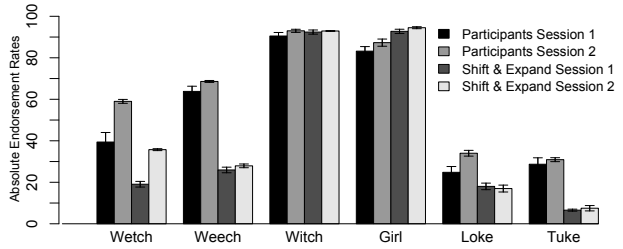


Figure 3: Absolute endorsement percentages by participants in Maye et al. (2008) and *Shift and Expand* model.

far less often. On the other hand, the model predicts that participants accept *Girl* item types over 90% of the time, despite the fact that participants are only accepting these words around 80% of the time. Therefore, there are aspects of the experimental data that the model does not seem to be capturing.

One possible explanation for why we observe this difference between the participants and our models is that unlike our models, people have had exposure to many different types of accents before. If they have heard accents that affect vowels in similar ways as the accent in this experiment, then perhaps they already have a representation for this type of speech and accept it more willingly. The models have only had exposure to a standard American accent. Incorporating some knowledge of other accents would be likely to increase the models' acceptance of these test item types.

Discussion

This paper explores how people adapt their sound categories in response to novel accents. We simulated the Maye et al. (2008) experiment using four different models, each of which adapted categories differently. In one version, categories were adapted by being shifted, but not expanded (*Shift* hypothesis). In a second version, categories were expanded, but not shifted (*Expand* hypothesis). In the third model, categories were shifted and expanded (*Shift and Expand* hypothesis). The final *Vowel Mapping* model directly learned the category mapping between accents. All models captured the data reasonably well, though the expansion models seemed to fare slightly better than the *Shift* and *Vowel Mapping* models. Crucially, this was the case on data that have been used to argue that people use the *Shift* strategy, rather than the *Expand* strategy.

How is *Expansion* able to capture the data? Maye et al. (2008) argue that only a *Shift* strategy can explain why participants show an increase in endorsement rates for *Wetch* items, but not *Weech* items, yet we found that expansion hypotheses were able to capture this asymmetry. Their assumption of symmetry is true if the affected vowels are all equidistantly spaced along a line. In any space where this is not true, such as F1-F2 space, their argument does not hold. On top of that, their test item groups differed in the number of items involving each vowel pair. For example, there were fifteen *Weech* items involving [ɪ] raising to [i], but only six *Wetch* items involving [i] lowering to [ɪ]. Because some vowels are closer together than others, such differences could produce endorsement rate asymmetries between different test item types that are not reflective of listeners' adaptation strategy. Contrary to their assumption, an *Expand* model can lead to asymmetric responses in lexical decision tasks, as they observed in their data.

We only tested one instantiation of each of the *Shift*, *Expand*, and *Shift and Expand* models; therefore, the conclusions we can draw about which of these models best capture the experimental findings are limited. There may, for example, be other instantiations of the *Shift* hypothesis (i.e. with different mean and covariance confidence parameters) that outperform the models discussed here. Overall, however, all three adaptation models are able to capture, to a better or worse degree, the main qualitative findings from Maye et al. (2008).

Most work on accented speech perception has suggested that category expansion is a component of accent adaptation. A key exception comes from Maye et al. (2008), who argue that people shift their categories, without expanding them. Although their experimental results may intuitively support a *Shift* strategy, building a model of the results shows that this argument is not conclusive. In fact, the *Shift & Expand* and *Expand* strategies explain these findings equally well to, or potentially better than the *Shift* hypothesis. Taken together with the rest of the accent adaptation literature, this work suggests that people may be relying on expansion strategies in accent adaptation.

Acknowledgments We thank the reviewers, the UMD Probabilistic Modeling group, and the audience at NECPhon 9 for helpful feedback. This work was supported by NSF grant IIS-1421695.

References

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.

Eisner, F., Weber, A., & Melinger, A. (2010). Generalization of learning in pre-lexical adjustments to word-final devoicing. *The Journal of the Acoustical Society of America*, *128*(4), 2323–2323.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562.

Norris, D. (2006). The bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological review*, *113*(2), 327.

Olive, J. P., Greenwood, A., & Coleman, J. (1993). *Acoustics of american english speech: a dynamic approach*. Springer Science & Business Media.

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, *15*(6), 732–738.

Weatherholtz, K. (2015). *Perceptual learning of systemic cross-category vowel variation*. Doctoral dissertation, The Ohio State University.

Weide, R. (1998). *The cmu pronunciation dictionary, release 0.6*. Carnegie Mellon University.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384.

Witteman, M. J., Weber, A., & McQueen, J. M. (2010). Rapid and long-lasting adaptation to foreign-accented speech. *The Journal of the Acoustical Society of America*, *128*(4), 2486–2486.

Gibson's Reasons for Realism and Gibsonian Reasons for Anti-Realism: An Ecological Approach to Model-Based Reasoning in Science

Guilherme Sanches de Oliveira (sanchege@mail.uc.edu)
Department of Philosophy, 2700 Campus Way
Cincinnati, OH 45221 USA

Abstract

Representational views of the mind traditionally face a skeptical challenge on perceptual knowledge: if our experience of the world is mediated by representations built upon perceptual inputs, how can we be certain that our representations are accurate and our perceptual apparatus reliable? J. J. Gibson's ecological approach provides an alternative framework, according to which direct perception of affordances does away with the need to posit internal mental representations as intermediary steps between perceptual input and behavioral output. Gibson accordingly spoke of his framework as providing "reasons for realism." In this paper I suggest that, granting Gibson his reasons for perceptual realism, the Gibsonian framework motivates anti-realism when it comes to scientific theorizing and modeling. If scientists are Gibsonian perceivers, then it makes sense to take their use of models in indirect investigations of real-world phenomena not as representations of the phenomena, but rather as autonomous tools with their own affordances.

Keywords: perception; ecological psychology; affordances; representation; philosophy of science; scientific modeling

Introduction: Representationalism and Skepticism

Representationalism concerns, roughly speaking, the thesis that the mind operates by generating, storing and processing representations of the world. In this view, as traditionally understood, perceptual input is transformed into internal representations or "information-bearing structures" (Pitt 2013) which in turn are utilized to create behavioral output. Recent representationalists differ in how they define 'representation', distinguishing between internalism and externalism about both the "content" of representations and the "vehicles" that carry representational content (see, e.g., Clark and Toribio 1994, for a rejection of the representationalism/anti-representationalism dichotomy and the suggestion that representation comes in different degrees and types; along similar lines, Hurley 2001 provides a comprehensive taxonomy of externalist views, defending a distinction between what- and how-externalism). Representational theories of the mind, in particular the internalist varieties, are intimately connected to the threat of skepticism: if our knowledge of the "external world" is indirect and mediated by "internal reconstructions" of our surroundings—if, that is, we only have direct experience of our mental representations, not of the world itself—then it is reasonable to wonder about the accuracy of such representations and, consequently, about the reliability of our knowledge.

Epistemological consideration of the problem of perceptual knowledge dates back at least to Plato, who likened input from the senses to the shadows on the dark cave wall, imperfect reflections of the outside world (cf. Newman 2014). A classical example of skeptical doubts resulting from a more explicitly representational approach to the mind is found in the work of René Descartes, who took there to be "reasons for which we may, generally speaking, doubt about all things and especially about material things" (Descartes, Weissman, & Bluhm 1996, p. 55). In his *Meditations* Descartes discusses skeptical threats involving dreams and evil demons. We sometimes have dreams that seem so realistic that we have the impression that the dream is indeed real. Based on this common experience, Descartes asks: how can we be sure that we are not currently dreaming? Further, how can we be certain that our entire lives were not a dream? A second skeptical doubt for Descartes concerns the possibility that an evil, deceitful demon or deity gives us misleading sensory inputs such that we believe in the existence of an external world when in reality none exists. Different versions of these skeptical threats have persisted since Descartes and influenced philosophical theorizing. A variation of the evil demon argument that has been influential in recent analytical philosophy, for example, concerns the idea that if one were a disembodied brain in a vat and were fed representations of the external world, one would be unable to know that those were not real (Harman 1973, Putnam 1981).

The thesis that the mind operates by generating, storing and processing mental representations—particularly internal structures that encode information about the world—has been labeled "the central hypothesis of cognitive science" (Thagard 2014). But that is not to say that there are no alternatives. Major examples of anti-representationalist accounts are found in dynamical systems theory (van Gelder 1998), situated robotics (Brooks 1991; Webb 1995, 2005), and radical enactivism (Hutto and Myin 2013). Ecological psychology has informed and motivated much of this anti-representationalist work (see, e.g., van Gelder and Port 1995 on the its influence on the dynamical systems approach), and will be the focus of the present paper.

Ecological Psychology Revisited

Ecological psychology, as articulated by J. J. Gibson in his 1979 book *The Ecological Approach to Visual Perception*,

was founded on three precepts. First, Gibson suggested that perception is *direct*: rather than having to reconstruct the world by generating representations built upon sensory input, we have unmediated perceptual access to what is present in our surroundings. Second, for Gibson perception is inherently *action-oriented*, and we always perceive in order to do something. Gibson claims that "perceiving is an act, not a response, an act of attention, not a triggered impression, an achievement, not a reflex" (1979, p. 149). In his view, as active explorers of the world, we are in constant engagement with our environment, such that perception is in principle and in practice indissociable from our acting and preparing for action. Third, Gibson proposed a radically novel account of what the objects of perception are. Rejecting the idea that we perceive discrete properties such as an object's size, shape and composition and then have to analyze those properties in order to determine how to relate to the object, Gibson proposed that perception is of "*affordances*" or "action possibilities." Ordinary examples include the possibility to sit on (in the case of a chair), to pass in between (in the case of an aperture), or to cut with (in the case of knives). Gibson's proposal is that the descriptions favored by mainstream psychology and cognitive science are mistaken when they take perception to be the first stage in a process that involves the generation and internal processing of sensory inputs in order to produce external behavioral outputs (similar to what Hurley 2001 calls the "sandwich view" of the mind). Instead, for Gibson we simply and directly perceive *how* we can engage with objects and *what* we can *do* in the environment, such that perception-cognition-action become an indissociable unit. In light of this, the title of Gibson's 1979 book and its emphasis on visual perception can seem misleadingly simplistic: in reality, by treating perception as necessarily action-guided and of action possibilities, Gibsonian ecological psychology collapses perception, cognition and action, at once providing a unitary account of the three.

Gibson's concept of "affordance" has generated considerable controversy. His own description was that affordances are what the environment "offers the animal, what it provides or furnishes, either for good or ill" (1979, p. 127). This somewhat obscure formulation was later interpreted as referring to the *dispositional properties* of the objects themselves (Scarantino 2003). Another interpretation, one that seems to be more faithful to Gibson's proposal, frames affordances in *relational* terms as the interplay of subjective skills and environmental properties, or "relations between the abilities of organisms and features of the environment" (Chemero 2003, p. 189). A more recent adaptation of this relational interpretation attempts to account for social and cultural differences by defining affordances as "relations between aspects of a material environment and abilities available in a form of life" (Rietveld and Kiverstein 2014). Regardless of the particular variety, similar relational interpretations have in the past couple of decades influenced expansions of the Gibsonian framework, which, paired with the explanatory tools of dynamical systems theory (Kugler, Kelso, & Turvey, 1980), have led to the use of non-linear modeling to investigate systems that exhibit "interaction-dominant dynamics" (Van Orden et al. 2003). These additions and applications have secured to the Gibsonian framework a place as a successful approach in contemporary experimental psychology and cognitive science. As for the concept of 'affordances', it has taken its own life in the domain of design research (see, e.g., Norman 1990) as well as in neuroscience, where it has been framed in terms of neural representations of object-related motor intention (Jeannerod 1994, Craighero et al 1998, Tucker and Ellis 2001) and linked to the mirror systems believed to be associated with social cognition (Bach et al 2011, Brincker 2015).

Gibson's Reasons for Realism

Framed in relational terms, as seen above, affordances can be interpreted as undermining the distinction between subject and object: action possibilities exist objectively regardless of subjective awareness, and still they are subjective in that they change according to the individual organism and its characteristics. Gibson himself seemed unbothered by the dichotomy, claiming that "an affordance is neither an objective property nor a subjective property; or it is both if you like" (Gibson, 1979, p. 129). An anti-dualist in many respects, Gibson rejected also the existence of any difference between the natural and the artificial, and, accordingly, between "affordances in general" and the affordances of human artifacts. Psychologist Alan Costall has questioned this point, proposing that artifacts have what he calls "canonical affordances," that is, conventional and normative action possibilities: even if other uses can be found, chairs, for example, are *for sitting on*. Costall considers this expansion of the conceptual framework to complement Gibson's own interests by taking analysis beyond the subject-object dyad: "The object needs to be understood within a network of relations not only among different people, but also a "constellation" of other objects drawn into a shared practice" (Costall 2012, p. 92). Whether this is a reasonable suggestion and particularly whether it provides additional insights into scientific modeling qua affordance-based practice is something to be explored elsewhere. For now, with the basic outline of the Gibsonian approach in place, let us consider what conclusions it motivates for philosophical discussions about realism.

In a paper published a decade before his explicit articulation of the ecological approach in the 1979 book, Gibson argued that the view of perception he was developing provided the basis for a kind of naïve realism, one that justified the confidence we ordinarily have in (the reliability of) our epistemic access to the world around us. He summarized his account at the time saying that it was based on: "(1) the existence of stimulus information, (2) the fact of invariance over time, (3) the process of extracting invariants over time, and (4) the continuity of perception with memory and thought" (1967, p. 167-168). Despite the difference in terminology, the spirit of the theory seems

similar enough to the 1979 book, particularly with respect to the assumption that information is available perceptually, as well as to the merging of perception and cognition. Again, Gibson's view, even in this earlier stage, is that perception is of publicly available information; this is a direct approach, with no intermediary steps between perception and action that involve the generation and manipulation of mental representations. If there existed such representations, there would be a risk that they are inaccurate, that they *misrepresent*. But Gibson does away with the need to posit representations—he claims: "I am quite certain that there is no such thing as a phonograph record in the ear and no such thing as a picture in the eye—no reproduction of an external event or object that the organ transmits to the brain" (1967, p. 169). Accordingly, Gibson concludes that his ecological approach provides answers to traditional epistemological concerns by substantiating a kind of direct naïve realism: taking the senses as perceptual systems, he says, offers "sophisticated support for the naive belief in the world of objects and events, and for the simple-minded conviction that our senses give knowledge of it" (1967, p. 168)—that is, since perception is a process of actively and directly picking up ecological (publicly available) information, skeptical threats do not even get off the ground and we can rest assured—as non-philosophers tend to—that we stand on solid epistemic foundations.

Scientific Models and Representation

Assuming for the sake of argument that Gibson is right both in the basic ecological framework and in the interpretation that direct perception provides the basis for a realist outlook, this paper is concerned with whether Gibson's reasons for *perceptual realism* translate into reasons for *scientific realism*—i.e, realism about the unobservable entities postulated in scientific theorizing and about the relationship between scientific products and the "real world." Taking the recent debate about scientific modeling as a starting point for this investigation, and assuming that scientists are Gibsonian perceivers, the question here concerns what conclusion Gibson's reasons for perceptual realism motivate with regards to the epistemic import of the use of models in scientific research.

In recent years, philosophers have increasingly acknowledged the role played in contemporary science by *mediated* or *indirect* forms of investigation through the construction and manipulation of models like mathematical equations, computer simulations, and robotic replicas. A widespread view in this literature is that scientific modeling is a *representational* activity (Giere 2004, 2010; Morrison and Morgan 1999; Morrison 2015; Suarez 2003, 2004; van Fraassen 2008; Weisberg 2013). The *representational view of models* amounts to the assumption that models are *representations* of the target systems or phenomena they simulate, and that the epistemic value of model-based scientific research is grounded on the supposedly *representational* relationship between a model and its target. Even while acknowledging that models often neglect some details from the target phenomena and include known distortions and simplifications, the representational view of models takes it that models *represent* the target—that is, that despite containing purposeful deviations from "veridical representation[s] of real-world phenomena," a model is a representation of the phenomena (Weisberg 2013, p. 98). And in addition to taking the relationship between models and their targets to be *representational*, the representational view of models assumes further that this representational relationship is what secures the epistemic value of modeling, or what allows scientists to learn about a target through indirect experimentation with a model: models can teach us about the world "precisely because (…) they fulfill a representative function" (Morrison & Morgan 1999, p. 24-25). Recent contributions to this literature have put forward different accounts of scientific representation, providing distinct views of how a model represents its target: some have framed representation in terms of *isomorphism* or one-to-one structural correspondence (van Fraassen 2008) or in terms of *resemblance* or *similarity* (Giere 2004, 2010); other accounts suggest that, beyond the properties of the model-target dyad, scientific representation involves drawing *inferences* (Suarez 2003, 2004) and active *interpretation* (Contessa 2007), or that it is best understood through the lens of *semiotics* (Knuuttila 2010). Despite the variety of takes on how scientific representation works, the representational view of models presupposes that scientists build and use models to learn about some target phenomenon because models represent their target.

An Ecological Approach to Scientific Modeling

What would an alternative to this representational view of model-based scientific research look like? More specifically, what would a *Gibsonian* alternative to representationalism about scientific modeling look like? The account I outline here builds upon the ideas of treating models as "autonomous agents" and "instruments of investigation" (Morrison & Morgan 1999) and as "epistemic tools" (Knuuttila 2011). Morrison and Morgan emphasize how models are partially independent both from theory (the scientific conceptual framework within which models are built) and from the phenomena they are intended to simulate, such that experimenting with a model can shed light both on the relevant scientific theory and on the target of investigation. Like Morrison and Morgan, Knuuttila highlights the manipulability of models as central to their epistemic contribution in science—researchers need to be able to manipulate, intervene on, and otherwise experiment with a model in order to learn something from it. But Knuuttila further adds that the manipulability of models implies their concreteness or "materiality," based on which she proposes that models can play the role of an external scaffolding that, in the sense suggested by Hutchins (1995) and Sterelny (2004), *extends* the scientists' cognitive processing. This last, stronger claim about cognitive extension, while interesting and worthy of further

exploration, actually complicates the issue at hand more than is necessary: the notion of cognitive extension is not one Gibson included in his framework, and can be set aside for present purposes. Accordingly, the working hypothesis to be considered here is simply that the creation and manipulation of models in the context of scientific research can be understood in comparison with how we create and engage with instruments, artifacts and tools more generally.

From a Gibsonian perspective, the ordinary objects we encounter in daily life afford a variety of uses, depending on various properties such as shape, size, and composition. For example, "a pointed elongated object affords piercing" while "a rigid object with a sharp dihedral angle, an edge, affords cutting and scraping" (Gibson 1979, p. 133). Indeed, we manufacture objects with those properties (such as spears and needles in the first case and knives in the second) *because* they afford such uses. Importantly, however, affordances are not reducible to the properties of objects or the environment. In the *relational* framework reviewed above, action possibilities are relations between organism and environment. Thus, in any situation, affordances will vary according to not only what the environment is like, but also *who* is engaging with it: the actions that are possible or not, are possible or not *for* some entity. The characteristics that make a surface climbable for a fit adult of average height are not the same that make it climbable for a child, for a dog, or for a climbing plant, and the differences are not simply a matter of body scale but of the general abilities and skills that the organism has.

Applying this framework to scientific modeling allows us to recognize that models are designed and built by modelers and scientists *so as to* afford some use(s) of interest. Whether it is a robotic model that is built to display some particular behavior, or a computer simulation that is designed to allow interventions of some sort, or a mathematical model that is specified in order to generate predictions based on some data input, scientific models are indissociable from the research context in which they are built. The guiding questions of a model-based project thus determine what the features of the model need to be so that it can be useful in that context. But usefulness is not an absolute value, which means that, from the early stages, model-building is a matter of constructing a tool that will provide relevant action possibilities *for someone*. This specificity imposed by research context and intended uses/users makes it so that no model is ever a perfect representation of its target: in fact, rather than accurate or inaccurate *descriptions* of some target phenomenon, the suggestion here is that models are, at best, the most useful tool available for a given *task*. The affordances of a model, then, are the action possibilities the model offers or makes available to model-builders and intended users.

An interesting consequence of Gibsonian ecological psychology is that it blurs the distinction between the subjective and the objective: affordances are subjective, or "relative to the animal" and "unique for that animal" (Gibson 1979, p. 127), but they are also "objective, real, and physical" (p. 129) in the sense that an "opportunity for action" exists for an animal even if that animal does not act upon it. First, which action possibilities an object affords an organism is a fact irrespective of that organism's awareness of it, that is, "*independent of the individual's ability to perceive this possibility*" (McGrenere & Ho 2000, p. 179). And additionally, the particular affordances of any object can vary across individuals at a given time as well as for the same individual over time: "knee-high for a child is not the same as knee-high for an adult, so the affordance is relative to the size of the individual" (Gibson 1979, p. 128), varying also, as seen above, according to what kind of organism it is —whether it is a human, a dog, a plant, and so on.

In this view, the affordances of a scientific model are objective in that they exist independently of their being perceived by an individual scientist, but they are also subjective in that the same model may not afford the same action possibilities to different individuals. This is significant because it illuminates both the differences between scientists and non-scientists, as well as, within science, the different practices that characterize distinct scientific disciplines. On the one hand, the variability of affordances resonates with the fact that in order to become a scientist one has to go through years of training and arduous work—it requires learning, or in Gibson's terminology, the "education of attention" (1979, p. 254); it is only through serious intentional effort that the novice can learn to perceive certain affordances that are already present and also develop new ones, as is evident in the case of highly complex mathematical models, which are gibberish to the uninitiated. On the other hand, the variability of affordances also sheds light on the differences that exist between scientific disciplines: training with the methods and tools of a particular discipline attunes an individual to the particular approach favored in that discipline, making more salient specific features of the world (both the phenomena of investigation and the models that are built and used to investigate them). The effect of specialization over time is thus present also collectively, shaping how disciplines develop differently from one another, so that, for example, the modeling techniques which appear essential to a behavioral ecologist may seem dispensable to a molecular biologist and entirely useless to a physicist, and vice versa. Members of the same species surely share many of the same affordances (as in the example of cutting objects); yet, oftentimes a lot of effort is required for individuals to become aware of certain affordances or even to develop new ones (as in the cases of training to become a scientist, and of intra- and interdisciplinary methodological differences).

Gibsonian Reasons for Anti-Realism

As seen above, Gibson took his direct-perception approach to substantiate a realist stance: our perception of the world is reliable because it is not based on the production of "internal representations" that could be wrong about the "external world," but rather perception is the very process by which we actively and directly pick up ecological

information. Now even if in this Gibsonian perspective we can be realists about our perception of the world, the ecological view of models sketched above motivates a more nuanced conclusion for realism about scientific knowledge generated from model-based research. Applying to scientific methodology the insight that perception is direct and of action possibilities, we conclude that the models and simulations used in science are *tools* to whose affordances we also have direct perceptual access. Just like any organism perceives in its surroundings opportunities to act in particular ways, some scientists perceive in aggregates of silicon and synthetic compounds, or in simulation softwares, or in mathematical equations the action possibilities that enable them to study the phenomena they are interested in. The point, however, is that this ecological approach to modeling also articulates the model-target relationship in anti-representational terms. That means that rather than taking the resulting robotic, computational or mathematical model to be an accurate or inaccurate representation of some target, the model is an independent *tool* that is *useful* insofar as it offers the manipulability that is considered relevant in the given research context.

In the 1979 book, Gibson briefly speaks of a kind of "second-hand" perceptual knowledge in which ecological information is somehow mediated. This is the case, for example, in the use of instruments like telescopes and microscopes, and in pictures and verbal descriptions. He claims: "The reality-testing that accompanies unmediated perceiving and that is partly retained in perceiving with instruments is obviously lost in the kind of perceiving that is mediated by pictures. Nevertheless, pictures give us a kind of grasp on the rich complexities of the natural environment that words could never do" (Gibson 1979, p. 263). While the (partial) persistence of ecological information in these forms of mediated perception allows Gibson to remain a perceptual realist, the same does not apply to knowledge generated through model-based scientific research because models are not instruments that mediate perception, but rather autonomous tools and artifacts built for a range of action-oriented purposes.

Consider projects in situated robotics. Some robotic models are explicitly intended to simulate real-world phenomena, such as phonotaxis in crickets (Webb 1995; Reeve et al 2005) and thigmotaxis in rat pups (Schank et al 2004; May et al 2006), while others merely set out to design autonomous intelligent robotic agents for exploratory purposes (Brooks 1991; Beer and Williams 2015). It seems intuitive to say that the latter do not *represent* any particular target because the models are not designed to replicate the performance or architecture of particular types of real-world creatures. Yet, the same conclusion applies, within the present framework, to the former cases: robots do not *represent* their targets (real crickets and rat pups), but are autonomous tools whose building and manipulation lead to insights that can be applied to the targets—in other words, the models are not descriptions of *what* the target organisms *are like*, but rather they are tools that are *like and unlike* the target organisms in contextually-relevant ways. And the same idea applies more broadly to other kinds of models (whether concrete, virtual, mathematical, etc.) in different disciplines. In the Gibsonian perspective being proposed here, it is not a supposedly representational relationship between model and target that makes scientific modeling epistemically valuable—rather, it is the agent-specific action possibilities scientists exploit in building and manipulating models that enable an indirect way of learning. In some cases modeling involves the generation of predictions that are subject to empirical confirmation, but in other cases it involves simply the generation of some behavior that is importantly similar and different from what is observed in the target. In these and other cases, however, models do not yield representations of the target that we can be realists about—no description, that is, that can be *universally valid*, *independent from particular perspectives* (Chakravartty 2011) or *assumption-free* and *accurate* (Gaukroger 2012).

In the ecological view of human cognition-action-perception, the usefulness of model-based science is explained in terms of the action possibilities afforded by models and by targets. Those are necessarily action possibilities *for* someone; and in most if not all cases, they are action possibilities that are limited to trained experts in the right research context, material environment, and conceptual problem-space. The fact that this results in Gibson's *reasons for (perceptual) realism* motivating a kind of *anti-realist* stance with regards to the epistemic products of scientific modeling should not be seen as negative for either the original Gibsonian framework nor for the proposed ecological view of modeling; rather, this proves that Gibson's insights remain relevant for empirical and theoretical work alike, contributing to interdisciplinary debates where philosophical and psychological tools combined can provide a better understanding of scientific practice and methodology.

References

Bach, P., Bayliss, A. P., Tipper, S. P. (2011). The predictive mirror: Interactions of mirror and affordance processes during action observation. *Psychonomic Bulletin & Review*, 18(1), 171-6.

Beer, R. D., and Williams, P. L. (2015). Information Processing and Dynamics in Minimally Cognitive Agents. *Cognitive Science* 39, 1-38.

Brincker, M. (2015). Beyond sensorimotor segregation: On mirror neurons and social affordance space tracking. *Cognitive Systems Research* 34-35, pp. 18-34.

Brooks, R. A. (1991) Intelligence without representation. *Artificial Intelligence* 47, 139-159.

Chakravartty, A. (2011). Scientific Realism. *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2015/entries/scientific-realism/>.

Chemero, A. (2003). An Outline of a Theory of Affordances. *Ecological Psychology*, 15(2), 181–195.

Clark, A., and Toribio, J. (1994). Doing Without Representing? *Synthese* 101: 401-431.

Contessa, G. (2007). Scientific Representation, Interpretation, and Surrogative Reasoning. *Philosophy of Science*, 74, pp. 48–68.

Costall, A. (2012). Canonical affordances in context. *AVANT* Volume III, Number 2/2012. ISSN: 2082-6710.

Craighero, L., Fadiga, L., Rizzolatti, G., and Umilta, C. (1998). Visuomotor Priming. *Visual Cognition* 5 (1/2), 109-125.

Descartes, R., Weissman, D., & Bluhm, W. T. (1996). *Discourse on the Method: And, Meditations on First Philosophy*. Yale University Press.

Gaukroger, S. (2012). *Objectivity: A Very Short Introduction*. Oxford University Press.

Gibson, J. J. (1967). New Reasons for Realism. *Synthese* 17, 2.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin Co. Boston.

Giere, R. (2004). How Models Are Used to Represent Reality. *Philosophy of Science*, 71, pp. 742–752.

Giere, R. (2010). An agent-based conception of models and scientific representation. *Synthese* 172:269–281.

Harman, G. (1973). *Thought*. Princeton University Press.

Hurley, S. (2010). The Varieties of Externalism. In R. Menary (Ed.), *The Extended Mind*. MIT Press.

Hurley, S. (2001). Perception and Action: Alternative Views. *Synthese* 129: 3-40.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

Hutto, D. D., and Myin, E. (2013). *Radicalizing Enactivism. Basic Minds without Content*. MIT Press.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, 17, pp 187- 202.

Knuuttila, T. (2010). Not Just Underlying Structures: Towards a Semiotic Approach to Scientific Representation and Modeling. In Bergman, M., et al. (Eds.), *Ideas in Action: Proceedings of the Applying Peirce Conference* (pp. 163–172).

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science* 42, 262–271.

Kugler, P. N., Kelso, J. A. S., and Turvey, M. T. (1980). Coordinative structures as dissipative structures I. Theoretical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 3–47).

May, C. J., Schank, J. C., Joshi, S., Tran, J., Taylor, R. J., and Scott, I. (2006). Rat Pups and Random Robots Generate Similar Self-Organized and Intentional Behavior. *Complexity* 12, 53–66.

McGrenere, J., & Ho, W. (2000). Affordances: Clarifying and evolving a concept. *Proceedings of Graphics Interface* 2000, 179-186.

Morrison, M., & Morgan, M. (1999). *Models as Mediators*. Cambridge University Press.

Morrison, M. (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press.

Newman, Lex, "Descartes' Epistemology", *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2014/entries/descartes-epistemology/>.

Pitt, D. (2013). Mental Representation. *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2013/entries/mental representation/>.

Putnam, H. (1981). *Reason, Truth, and History*. Cambridge University Press.

Reeve, R., Webb, B., Horchlerb, A., Indiveri, G., Quinn, R. (2005). New technologies for testing a model of cricket phonotaxis on an outdoor robot. *Robotics and Autonomous Systems* 51, 41–54.

Schank, J. C., May, C. J., Tran, J. T., and Joshi, S. S. (2004). A Biorobotic Investigation of Norway Rat Pups (Rattus norvegicus) in an Arena. *Adaptive Behavior*, Vol 12, no 3–4, 161–173.

Sterelny, Kim. (2004). Externalism, epistemic artefacts and the extended mind. In R. Schantz (Ed.), *The externalist challenge. New studies on cognition and intentionality*. de Gruyter.

Suarez, M. (2003). Scientific representation: against similarity and isomorphism. *International Studies in the Philosophy of Science*, 17 (3). 225-244.

Suarez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science*, 71 (5). 767-779.

Thagard, P. (2014). Cognitive Science. *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2014/entries/cognitivescience/>.

Tucker, M., and Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, 8 (6), 769-800.

van Fraassen, B. (2008). *Scientific Representation*. Oxford University Press.

van Gelder, T. (1998) The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences*, 21, 615-665.

van Gelder, T., and Port, R. F. (1995). It's about time: An overview of the dynamical approach to cognition. In R. F. Port and T. van Gelder (1995) (Eds.). *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press.

Van Orden, G., Holden, J., and Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132, 331-351.

Webb, B. (1995). Using robots to model animals: a cricket test. *Robotics and Autonomous Systems* 16, 117-134.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Measuring the Causal Dynamics of Facial Interaction with Convergent Cross Mapping

Eric Postma (e.o.postma@uvt.nl)
Tilburg center for Cognition and Communication
Tilburg, The Netherlands

Marie Postma-Nilsenová (m.nilsenova@uvt.nl)
Tilburg center for Cognition and Communication
Tilburg, The Netherlands

Abstract

The nature of the dynamics of nonverbal interactions is of considerable interest to the study of human communication and future human-computer interaction. Facial expressions constitute an important source of nonverbal social signals. Whereas most studies have focused on the facial expressions of isolated individuals, the aim of this study is to explore the coupling dynamics of facial expressions in social dyadic interactions. Using a special experimental set-up, the frontal facial dynamics of pairs of socially interacting persons were measured and analyzed simultaneously. We introduce the use of convergent cross mapping, a method originating from dynamical systems theory, to assess the causal coupling of the dyadic facial-expression dynamics. The results reveal the presence of bidirectional causal couplings of the facial dynamics. We conclude that convergent cross mapping yields encouraging results in establishing evidence for causal behavioral interactions.
Keywords: mimicry; convergent cross mapping; facial expressions

Introduction

In recent years, the development of automatic coding algorithms boosted the study of facial expressions (Littlewort et al., 2011). By automatically tracking the facial dynamics in response to experimental conditions, the nonverbal facial responses to a large variety of situations can be assessed. Human facial expressions constitute a complex dynamical system formed by the facial muscles that are controlled by brain dynamics which, in turn, are governed by endogenous and exogenous sources. Facial expressions have an important signalling function. By means of expressions, nonverbal information is transmitted to social partners and provides an important context to the verbal information (Smith, Cottrell, Gosselin, & Schyns, 2005). In successful dyadic interactions, the facial expressions of interlocutors may be expected to be coupled. For instance, in mother-child interactions, the smile of the mother may invoke a smiling response from the child (Cohn, 2010), which, in turn, reinforces the smile of the mother. In a behavioral context, such a transient positive feedback loop is associated with behavioral synchrony or mimicry, the tendency of interacting humans to synchronize or mimic their behaviors (Louwerse, Dale, Bard, & Jeuniaux, 2012). In a mathematical context, such feedback loops can be modeled in terms of a coupled dynamical system.

The aim of this paper is to study behavioral mimicry by means of dynamical system theory. More specifically, the goal is to determine if dynamical system theory can be used successfully to measure the presence and direction of causality in the interaction dynamics of dyadic facial expressions. Detecting synchrony or mimicry in behavioral interactions is a challenge. Simply measuring the correlation between two dynamic time-series representing non-verbal behaviors of the two interlocutors (e.g., their smiling intensities) falls short for two reasons: (i) there may be causation without correlation, and (ii) there may be correlation without causation. An example of causation without correlation is a case in which the causing effect equally often induces a positive result and a negative result. An example illustrating correlation without causation is the presence of an external effect that induces synchronous behavior. For example, a sudden flash of light induces the synchronized closing of the eyes of both conversation partners.

Many methods have been proposed to measure synchrony and mimicry in behavioral interactions, for instance: windowed cross correlation (Boker, Xu, Rotondo, & King, 2002), frame differencing co-occurrence (Paxton & Dale, 2012), Granger causality (Kalimeri, Lepri, Kim, Pianesi, & Pentland, 2011), and cross recurrence quantification (Shockley, 2005). For a more complete overview, the interested reader is referred to a relatively recent survey of methods (Delaherche et al., 2012). The main limitation of these methods is that they fail to measure causality in an appropriate manner. Windowed cross correlation suffers from the aforementioned limitations of correlation. Frame differencing co-occurrence only considers the simultaneity of time-specific derivatives of pixel intensity associated with the visually discernible behaviors of two persons. Granger causality has its merits for linear dynamics, but fails to account for the prevalent nonlinear dynamics of human behavior. Finally, cross recurrence quantification has a strong foundation in dynamic systems theory and provides many measures for characterizing the interaction dynamics. However, in itself it is not suitable to establish a causal coupling of temporal sequences. The only way to determine causality is to manipulate the alleged causing stimulus and measure its effect (Richardson & Dale, 2005).

A relatively recent method called Convergent Cross Mapping (CCM) (Sugihara et al., 2012) originates from the domain of complex system dynamics and allows for measuring causal couplings. In contrast to cross recurrence quantification, it provides means for the detection of causal relations

between interacting nonverbal signals without suffering from the two shortcomings of straightforward correlation.

Detecting Causality

In the context of complex dynamical systems, the dynamics of two interacting systems is represented by an *attractor manifold*: i.e. the admissible states within the state space spanned by the two (possibly multidimensional) variables. In what follows, we briefly explain Convergent Cross Mapping in terms of attractor manifolds.

Convergent Cross Mapping

When N interacting dynamical variables describe a single attractor manifold in N-dimensional phase space (i.e., a coupled system of N differential equations), the temporal development of each individual variable contains information about the influence of the other $N-1$ variables. This may be compared to a musical ensemble in which the contributions of each of the constituent members depend on those of all the other members. The realization that each variable in a coupled dynamical system contains information about all other variables, led Flores Takens (Takens, 1981) to propose the reconstruction of an N-dimensional attractor by mapping a single variable against one or more time-delayed versions of itself. He showed that the resulting "shadow attractor" is highly similar ("diffeomorphic") to the original N-dimensional attractor manifold. We illustrate these concepts by means of the standard example: the $N = 3$ dimensional Lorenz attractor defined by a system of three coupled differential equations:

$$\begin{aligned} dX/dt &= 10(Y-X), \\ dY/dt &= X(28-Z)-Y, \\ dZ/dt &= XY-8Z/3. \end{aligned} \tag{1}$$

The three variables X, Y, and Z, describe the time-varying system state, with t representing time. As can be seen in these equations, the dynamics of X depends on Y (and on X itself), those of Y on X and Z, and those of Z on X and Y. The top part of Figure 1 shows the attractor manifold M of the famous Lorenz system. The middle and the bottom part illustrates the shadow attractors M_X and M_Y formed by plotting $X(t)$ against its time-delayed versions $X(t+\tau)$ and $X(t+2\tau)$ and $Y(t)$ against its time-delayed versions $Y(t+\tau)$ and $Y(t+2\tau)$. By comparing the two shadow attractors, their similarities become apparent. These similarities reflect the fact that the two variables X and Y (and Z) are (by definition) causally related.

Convergent Cross Mapping exploits the same principle in the opposite direction. If two dynamical variables A and B are coupled, than if A has a causal relation to B, it is possible to predict the shadow attractor of A from the shadow attractor of B, and vice versa. In practice, the CCM algorithm measures the causal coupling between variables X and Y by selecting a point of shadow attractor M_X together with a number of its nearest neighbors. These are used to predict

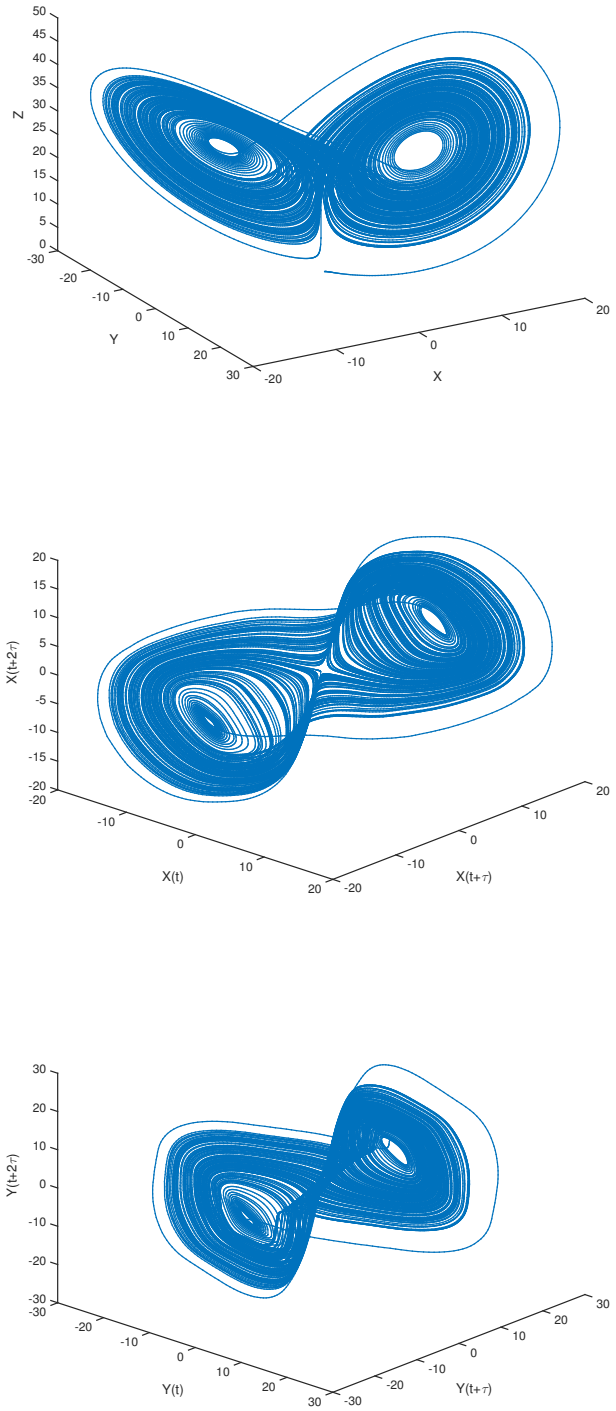


Figure 1: Top: Illustration of the Lorenz attractor manifold M. Middle: Shadow attractor manifold M_X obtained by plotting $X(t)$ against a time-delayed versions of itself: $X(t+\tau)$ and $X(t+2\tau)$ ($\tau = 30$). Bottom: Shadow attractor M_Y obtained for Y. The shadow attractors have a similar shape.

Table 1: Statements rated by the participants on the five-point scale "completely disagree", "disagree", "neutral", "agree", and "completely agree" . Only the ratings of statement 1 are used in the present study.

#	Statement
1	The conversation was good
2	I liked my conversation partner
3	My conversation partner seemed to like me
4	I listened carefully to my conversation partner
5	My conversation partner listened carefully to me

the temporally coupled points on shadow attractor M_Y. If M_X and M_Y are causally coupled, as is the case in the Lorenz attractor, increasing the number of nearest neighboring points in M_X leads to improved predictions of the points in M_Y. The improvement in prediction quality as a function of the number of points is a sign of an unidirectional causal coupling. The prediction quality is measured by means of a correlation value ρ_{CCM} of the actual and predicted values. Hence, the admissible values of ρ_{CCM} are confined to the unit interval.

Experimental Method

We applied CCM to measure causal facial interaction dynamics. The data was collected in an experiment studying mimicry and synchrony in a dyadic conversation. Twelve participants (10 female) were instructed to tell each other about the best and worst events they experienced in their lives. Throughout the session, one participant was assigned the role of the storyteller, whereas the other participant was assigned the role of the listener. After the session, each participant evaluated the quality of the interaction by filling in a brief survey. Participants were asked if they knew their conversation partner, and if so, how often they met. In addition, they rated five statements on a five-point scale ("completely disagree", "disagree", "neutral", "agree", "completely agree"). The five statements are listed in Table 1. The motivation for including these statements was to evaluate the relation between the perceived quality of the conversation and the CCM measurements. In the present study, only the ratings for statement 1 of the survey were used for a post-hoc comparison of the perceived quality of the interaction and the presence and strength of the causal coupling as determined by CCM.

Data Collection

Pairs of participants were each seated in front of an Eye Catcher system in separate rooms. An Eye Catcher system (Ex'ovision) is a recording and display device that by means of an internal see-through mirror, allows for the frontal recording of the face of a person watching the displayed face of his or her conversation partner. Two Eye Catcher systems were coupled in such a way that the pairs of participants could hear and see each other. Simultaneously, both participants were recorded during the session using digital recording software running on two Macbooks (OSX). All videos had a resolution of 640×400 pixels and a frame rate of 30 frames per second. For the current experiment, 6 pairs of sequences were selected for causal analysis. Four pairs involved female-female interactions and the other two mixed gender interactions. For each segment, two fragments of 3000 frames (100 seconds) were selected during which either participant spoke about a positive experience.

Table 2: Action units (AUs) extracted from the video sequences.

AU	Facial Expression
AU1	Inner Brow Raise
AU2	Outer Brow Raise
AU9	Nose Wrinkle
AU12	Lip Corner Pull
AU20	Lip Stretch

Facial Expressions Extraction

The resulting video sequences were processed using the Computer Expression Recognition Toolbox (CERT; Version 5.1, build 1208::867:869M) (Littlewort et al., 2011). CERT outputs for each frame of a facial video sequence estimates of action unit (AU) intensities. Action units are the building blocks of facial expressions as defined by Ekman and Friesen (1976). Each action unit describes the appearance or movement of a local region of the face. An example of an action unit is the "Inner Brow Raise", which is represented by its shorthand "AU1" (action unit 1). The action unit intensity estimates generated by CERT can be negative (evidence for absence) or positive (evidence for presence). The magnitude is proportional to the visual absence or presence of the action unit. Table 2 lists the five action units for which we extracted estimates for further analysis, cf. (Cohn, 2010). Two action units measure facial expressions in the upper facial region (AU1 and AU2), one in the central region (AU9), and two cover the lower facial region (AU12 and AU20).

Dynamical Systems Analysis

Convergent Cross Mapping was applied to paired sequences of the same AU intensities. (We did not examine causal couplings between different pairs of AUs.) We used the R implementation of Convergent Cross Mapping (*rEDM*_0.2.6) combined with Matlab routines for file input and output. CCM has two parameters: the embedding dimension and the time lag. The embedding dimension parameter specifies the number of time-delayed versions of a variable to reconstruct the shadow manifold. For instance, the shadow attractors displayed in Figure 1b and c were reconstructed with an embedding dimension of 3, corresponding to the original and two time-delayed version of X and Y, respectively (i.e., $X(t)$, $X(t+\tau)$, $X(t+2\tau)$ and $Y(t)$, $Y(t+\tau)$, $Y(t+2\tau)$). Its optimal value is highly dependent on the task. We determined the optimal value by varying the embedding dimension from 2 to 10. The time lag parameter defines the time lag separating the predict-

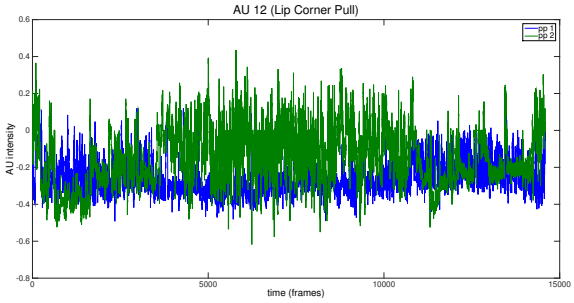


Figure 2: Example of paired time-series of action unit 12 estimates. The two series represent the two interacting participants of dyad 6, pp11 and pp12. The horizontal axis represents time in frames ($\frac{1}{30}$th seconds), the vertical axis represents the AU intensity.

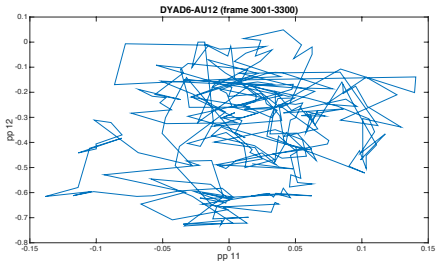


Figure 3: Visualization of the "attractor manifold" obtained by plotting the AU12 intensities of pp11 against those of pp12 for a sample of 300 frames (10 seconds). Adjacent frames are connected by lines.

ing and predicted AU intensity. Because we were interested in multiple time lags ranging from synchrony (time lag = 0) to mimicry (time lag > 0), we examined time lags ranging from 0 (synchrony) to 30 (mimicry) in steps of one. Each time step corresponds to $1/30$-th of a second. Hence, the time lags cover the interval from 0 to 1 second.

Results

Figure 2 shows an example of the AU12 estimates for the two interacting participants of dyad 6 (pp11 and pp12). Whenever a participant is listening, his or her AU12 intensity tends to be larger than that of the speaker. In the figure, participant pp12 is listening to pp11 in the middle time segment.

As an illustration of the interaction dynamics, Figure 3 shows the "attractor manifold" of a 10-second fragment of the AU12 signature of a dyadic interaction between pp11 and pp12. A circular structure represents a period coupling. Clearly, the interaction dynamics of pp11 and pp12 are more noisy and complicated than those of the Lorenz system. This is due to the noisy nature of the CERT action unit estimates and the complicated dynamics of behavioral interactions.

The results of applying the CCM algorithm to the paired time-series of the five action units revealed that only AU12 gave rise to causal coupling. For this action unit, we found that the prediction quality ρ_{CCM} increased as a function of the number of samples L for all dyads and in both directions. This indicates that all dyads exhibited a bidirectional causal coupling of their lip corner pull expression dynamics. The optimal strengths were obtained at time lags varying from 0 (synchrony) to 0.8 seconds (mimicry). The values of the prediction quality $\rho_{CCM}(L_{max})$ at the maximum number of samples examined L_{max} provide an indication of the coupling strengths.

Table 3 lists the strongest causal relations (the maximum values across the five action units examined) found for the 6 dyads examined. For each dyad, the Table lists the composition of each dyad, female-female (F-F), male-female (M-F), or female-male (F-M), the rating of statement 1 of the post-hoc evaluation of the conversation given by the participants ranging from strongly disagree (0) to strongly agree (5), the ratings by the two participants are separated by a /), and a specification of predictive quality $\rho_{CCM}(L_{max})$ obtained for the largest value of L examined, $L_{max} = 300$ in both directions: from the first participant to the second ($\rho_{CCM\rightarrow}$) and vice versa ($\rho_{CCM\leftarrow}$). As listed in the last column, the strongest causal relations were obtained for AU12 (Lip Corner Pull) which is a measure of (subtle) smiling. The results did not depend critically on the embedding dimension, although the best predictive qualities were obtained for the maximal value of the embedding dimension examined (10).

Table 3: CCM results for AU12. The first column specifies the dyad number. The second column indicates the gender composition of the dyad, where the first and second member of the dyad are separated by a hyphen (F = female, M = male). The third column lists the ratings of statement 1 in the survey ("The conversation was good") as a measure of the quality of the conversation (rating 1 represent "strongly disagree" and 5 "strongly agree"), via neutral (3) to very positive (5), the ratings by the two participants are separated by a /). The fourth and fifth columns list the predictive qualities $\rho_{CCM\rightarrow}$) and $\rho_{CCM\leftarrow}$) in both directions.

dyad	gender	evaluation	$\rho_{CCM\rightarrow}$	$\rho_{CCM\leftarrow}$
1	F-F	2/1	0.68	0.43
2	F-F	1/1	0.35	0.42
3	F-F	2/1	0.47	0.59
4	F-F	3/3	0.39	0.54
5	F-M	3/3	0.64	0.74
6	M-F	4/3	0.46	0.56

Figure 4 shows the results of applying Convergent Cross Mapping to paired time series of AU12, see also Table 3. The six plots depict the outcomes for dyads 1 to 6 (top to bot-

tom). Each of the plots shows the prediction qualities ρ_{ccm} as a function of the number of samples (Library Size L) for the two directions of interaction between the two dyad members A and B. Increasing the number of samples L from the shadow manifold M_A for predicting the shadow manifold M_B leads to an increase in the correlation ρ_{ccm} for A $\rightarrow$ B (green curve). Similarly, increasing the number of samples from M_B to predict M_A, increases the correlation ρ_{ccm} for A $\leftarrow$ B (blue curve). The observation that for all dyads both values of ρ exhibit a steady increase with the number of samples, indicates a bidirectional coupling for all dyads. Importantly, such increases are not seen for arbitrary sequences taken from different dyads, i.e., from two persons that did not interact. When pairing the same AU sequences of two persons of different dyads, a small increase is observed up to a maximal value of $\rho_{ccm} \approx 0.2$.

The coupling strengths show some variation both within and between dyads. For our small sample of dyads, there does not seem to be clear relation between the composition of the dyad or the evaluation of the conversation and the prediction strengths (neither for the other survey statements).

Discussion and Conclusions

As stated in the introduction, the goal of the present study was to determine if dynamical system theory can be used to measure the presence and direction of causality in the interaction dynamics of dyadic facial expressions. We employed the method of Convergent Cross Mapping for determining a causal relation between nonverbal behavior in conversation partners. The fact that the method was able to predict the facial expressions of the interacting partners confirms that the presence of a causal coupling in nonverbal behavior can be measured with CCM. Importantly, the prediction improves with an increasing number of samples per conversation partner in a given dyad. In the CCM method, this is considered a true sign of causality.

Previous studies indicated an increased tendency of females as compared to males to exhibit nonverbal mimicry (Dimberg, 1990; Briton & Hall, 1995). LaFrance and colleagues (LaFrance, Hecht, & Paluck, 2003) found a clear gender difference in smiling especially for female dyads as compared to male dyads. Although we did not find an effect of the evaluation of the conversation on the coupling strength, evidence for such an effect has been reported (Chartrand & van Baaren, 2009). Future work aims at increasing the number of dyads submitted to CCM analysis to establish if the effect can be confirmed. Finally, the range of lags at which the evidence for causal coupling is strongest agrees with the range of values reported in the literature (Dimberg & Thunberg, 1998; Dimberg, Thunberg, & Elmehed, 2000; Sato & Yoshikawa, 2007; Achaibou, Pourtois, Schwartz, & Vuilleumier, 2008).

To the best of our knowledge, this is the first time that CCM has been applied to measure the presence, strength, and direction of causal coupling in nonverbal behavioral dynamics. Although our results are encouraging, additional study is re-

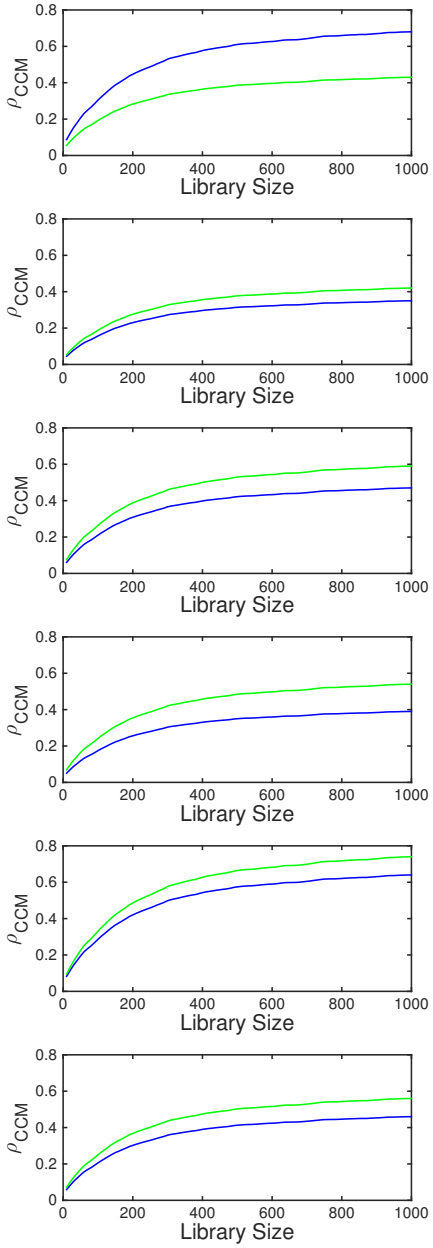


Figure 4: Illustration of the output of Convergent Cross Mapping applied to the paired AU12 time series of dyads 1 to 6 (top to bottom). The curves show the increase of the prediction quality ρ_{CCM} as a function of the number of samples (Library Size L). The increase is a qualitative measure of causal coupling. The green curve represents the causal effect of the first dyad member on the second, and the blue curve the causal effect in the reverse direction.

quired to establish the weaknesses and strengths of applying CCM in cognitive science. An important issue concerns the question to what extent behavioral coupling dynamics correspond to the complex system dynamics for which CCM and related methods were formulated. As a case in point, the "attractor manifold" depicted in Figure 3 appears quite chaotic when compared to the smooth attractor manifold of the Lorenz system in Figure 1. Although a difference in appearance is to be expected, given that the Lorenz system is a deterministic mathematical model and the dyadic AU12 dynamics reflect real-world behavioral measurements, a further analysis of coupled facial expressions is needed to establish the precise nature of the dynamics in relation to established mathematical formulations. Still, since CCM has been applied successfully to other noisy real-world time-series, such as the causal relation between galactic cosmic rays and annual variations in global temperature (Tsonis et al., 2015), we are confident about its validity as a measurement tool for establishing behavioral causality.

Notwithstanding our encouraging results, the relative small number of samples (6 dyads) does not warrant any far-reaching conclusions. In our future studies we will determine if the results generalize to a larger set of dyads, to other action units, and to other nonverbal behaviors. The main conclusion that we draw from our exploration is that Convergent Cross Mapping appears to provide a viable and fruitful alternative to existing measures of synchrony and mimicry.

References

Achaibou, A., Pourtois, G., Schwartz, S., & Vuilleumier, P. . (2008). Simultaneous recording of eeg and facial muscle reactions during spontaneous emotional mimicry. *Neuropsychologia*, *46*(4), 1104-1113.

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002, September). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338–55.

Briton, N., & Hall, J. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, *32*(1-2), 79-90.

Chartrand, T., & van Baaren, R. (2009). Human mimicry. *Advances in Experimental Social Psychology*, *41*, 219-274.

Cohn, J. F. (2010, November). Advances in behavioral science using automated facial image analysis and synthesis. *IEEE Signal Processing Magazine*, 128-133.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012, July). Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing*, *3*(3), 349–365. doi: 10.1109/T-AFFC.2012.12

Dimberg, U. (1990, September). Facial electromyography and emotional reactions. *Psychophysiology*, *27*(5), 481-494.

Dimberg, U., & Thunberg, M. (1998). Rapid facial reactions to emotional facial expressions. *Scandinavian Journal of Psychology*, *39*(1), 39-45.

Dimberg, U., Thunberg, M., & Elmehed, K. (2000). Unconscious facial reactions to emotional facial expressions. *Psychological Science*, *11*(1), 86-89.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56-75.

Kalimeri, K., Lepri, B., Kim, T., Pianesi, F., & Pentland, A. S. (2011). Automatic modeling of dominance effects using granger causality. In *Human behavior understanding* (p. 124-133). Springer Berlin Heidelberg. I.

LaFrance, M., Hecht, M. A., & Paluck, E. L. (2003). The contingent smile: a meta-analysis of sex differences in smiling. *Psychological Bulletin*, *192*(2), 305-334.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011, March). The computer expression recognition toolbox (CERT). *Face and Gesture 2011*, 298–305. doi: 10.1109/FG.2011.5771414

Louwerse, M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior Matching in Multimodal Communication is Synchronized. *Cognitive Science*, *36*(8), 1–48.

Paxton, A., & Dale, R. (2012, October). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior research methods*. doi: 10.3758/s13428-012-0249-2

Richardson, D., & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension. *Cognitive Science*, *29*(6), 1045–1060.

Sato, W., & Yoshikawa, S. (2007). Spontaneous facial mimicry in response to dynamic facial expressions. *Cognition*, *104*(1), 1-18.

Shockley, K. (2005). Cross Recurrence Quantification of Interpersonal Postural Activity. In *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 143–177).

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005, March). Transmitting and decoding facial expressions. *Psychological Science*, *16*(3), 184–189. doi: 10.1111/j.0956-7976.2005.00801.x

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., & Munch, S. (2012, October). Detecting causality in complex ecosystems. *Science (New York, N.Y.)*, *338*(6106), 496–500. doi: 10.1126/science.1227079

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence* (Vol. 898, pp. 366–381). Springer Verlag.

Tsonis, A. A., Deyle, E. R., May, R. M., Sugihara, G., Swansona, K., Verbeten, J. D., & Wang, G. (2015). Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3253-3256.

Linguistic niches emerge from pressures at multiple timescales

Molly Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

What accounts for the vast diversity in the world's languages? We explore one possibility: languages adapt to their linguistic environment (*Linguistic Niche Hypothesis*; Lupyan & Dale, 2010). Recent studies have found support for this hypothesis through correlations between aspects of the environment and linguistic structure. We synthesize this previous work and find that languages spoken in cold, small regions tend to be more complex across a range of linguistic features. We also test a novel prediction of the Linguistic Niche Hypothesis by examining the learnability of languages for first-language, child learners.

Keywords: Linguistic Niche Hypothesis; language evolution

Introduction

What factors shape language? Psychologists have made significant progress understanding this question in the domains of communicative interaction and children's developmental trajectories. In both cases, accounts rely on positing two pressures on the cognitive system—one internal and one external. In the case of communication, theorists argue that speakers are influenced by cognitive constraints (minimize effort) and by the needs of the communicative partner (be understandable; Horn, 1984). In the case of acquisition, there are internal maturational constraints, as well as external pressures from the quality and quantity of linguistic input (Hart & Risley, 1995). In the present paper, we explore the possibility that the same two pressures—system internal and external—may also shape *language systems* over the course of language evolution.

Central to this hypothesis is the notion of a timescale: there are different units of time over which processes operate, and processes at shorter timescales influence those at longer timescales (Blythe, 2015, see also Fig. 1). At the shortest timescale are individual utterances in communicative interactions (pragmatics). At a longer timescale is language acquisition. Both experimental and modeling work suggest that communicative interactions at the pragmatic timescale influence processes like word learning at the acquisition timescale (e.g., Baldwin, 1991; McMurray, Horst, & Samuelson, 2012; Frank, Goodman, & Tenenbaum, 2009; Frank & Goodman, 2014).

In addition to pragmatics and acquisition, a third relevant timescale is language evolution: the timescale over which entire language systems change. As for acquisition, there is evidence that language systems may be the product of processes at the pragmatic timescale. For example, languages universally structure semantic space to reflect optimal equilibria between communicative pressures (e.g., Kemp & Regier, 2012; Regier, Kay, & Khetarpal, 2007; Baddeley & Attewell, 2009).

However, the presence of communicative pressures at the pragmatic timescale is unable to explain cross-linguistic variability in linguistic structure. That is, why does Polish have rich morphology but English relatively sparse? A growing body of work argues that this variability may be due to cognitive constraints internal to the language learner (Chater & Christiansen, 2010) as well as properties of the environmental context (Nettle, 2012). This hypothesis, termed the *Linguistic Niche Hypothesis* (Lupyan & Dale, 2010; Wray & Grace, 2007), suggests that language systems adapt to the internal and external pressures of the linguistic environment.

A number of recent studies provide correlational support for this proposal. At the lowest level of the linguistic hierarchy, languages with larger populations are claimed to have larger phonemic inventories (Atkinson, 2011; Hay & Bauer, 2007), but shorter words (Wichmann, Rama, & Holman, 2011). Speakers with more second language learners have also been suggested to have fewer lexical items (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015). At the level of morphology, speakers with larger populations tend to have simpler morphology (Lupyan & Dale, 2010; Bentz & Winter, 2013). Finally, there is also evidence that population size may influence the mappings between form and meaning. In particular, this work suggests that languages tend to map longer words to more complex meanings (Lewis, Sugarman, & Frank, 2014), but that this bias is smaller for languages with larger populations (Lewis & Frank, 2016).

The plausibility of the Linguistic Niche Hypothesis depends largely on the presence of a possible mechanism linking environmental features to aspects of language systems. A range of proposals have been suggested (Nettle, 2012). For

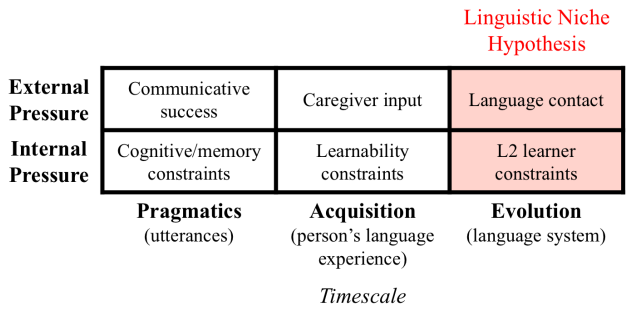


Figure 1: Pressures on language, internal and external to the cognitive system at three different timescales. The Linguistic Niche Hypothesis suggests that language evolution is influenced by the internal and external pressures in the particular environmental context in which a language is spoken.

example, one possibility is that children (L1) and adult (L2) language-learners differ in their learning constraints. In particular, children may be better at acquiring complex morphology than adults, and so languages with mostly children learners may tend to have more complex morphology. A second possibility is that speakers in less dense social networks have less variable linguistic input, and this leads the language system to have more complex morphology.

Providing evidence for these mechanisms is empirically challenging, however. Because there are many factors that shape a linguistic system, large datasets are needed to detect a correlation with environmental factors. In addition, there is non-independence across languages due to genetic relationships and language contact, and so data from a wide range of languages are needed to control for these moderators (Jaeger, Graff, Croft, & Pontillo, 2011). Third, the hypothesized mechanisms linking languages to their environments are somewhat underspecified. Finally, the large scale of this hypothesis makes it difficult to directly intervene on, and so we must rely primarily on correlational data to make inferences about mechanism.

In this work, we try to address some of these challenges by clarifying the empirical landscape. We do this by aggregating across datasets that find covariation between environmental variables and linguistic structure. This serves two purposes. First, it allows us to examine the relationship between the same set of environmental predictors across a range of linguistic features. And, second, it allows for the same analytical techniques and areal controls to be used across datasets. By addressing these inconsistencies, we are better able to compare directly relationships between environmental and linguistic features. A more coherent picture of the empirical landscape may in turn provide insight into the mechanism linking language systems to their environments.

We also explore a novel aspect of the Linguistic Niche Hypothesis: the relationship between L1 and L2 learnability. We ask whether the same languages that are more easily learnable by second language learners are also easier to learn for first language learners, or whether there is some tradeoff in learnability. As a proxy a language's learnability for child-learners, we use the mean age of acquisition of children's first words in a language.

In what follows, we first present a study examining the relationship between environmental and linguistic features using the same analytical techniques across all variables (Study 1). In Study 2, we examine the relationship between first language learnability and environmental and linguistic features.

Study 1: Environmental pressures on language

The Linguistic Niche Hypothesis suggests that languages are shaped by their environment, but the exact nature of these effects has varied across the literature—both in terms of the variables considered and the direction of the effect. To explore this variation, we combined data from five existing datasets that included environmental or linguistic data. The datasets were selected for being publicly available and containing a large sample of languages. Below we describe each of these datasets, followed by our analytical methods, and results.

Datasets

Lupyan and Dale (2010). This dataset contains grammatical information from WALS (Dryer & Haspelmath, 2013), and demographic and geographic information from Ethnologue and the Global Mapping Institute (Gordon, 2005). The demographic and geographic variables included total population of speakers, number of neighboring languages, area of region in which the language is spoken (km^2), mean and standard deviation temperature (*celsius*), and mean and standard deviation precipitation (*cm*). We used these data to create a metric of morphosyntactic complexity calculated from 27 of the 28 morphosyntactic variables analyzed in the original paper.[1] For each variable, we coded the strategy as simple if it relied on a lexical strategy or few grammatical distinctions (e.g., 0-3 noun cases), and complex if it relied on a morphological strategy or many grammatical distinctions (e.g., more than 3 noun cases). We summed the number of complex strategies to derive a measure of morphosyntactic complexity for each language, including only languages with data for all 27 variables. [n = 1991 languages]

Bentz et al. (2015). Two variables were used from this dataset: ratio of L2 to L1 speakers and number of word forms. Estimates of number of word forms were taken from translations of the *Universal Declaration of Human Rights*. Number of word forms was calculated as the number of unique words divided by the number of total words (type-token ratio). Higher type-token ratio indicates more word types in that language. Speaker population data were taken from a variety of sources, where L2 speakers were restricted to adult non-native speakers only. [n = 81]

Moran, McCloy and Wright (2012). Estimates of number of consonants and vowels in each language were used from this dataset. [n = 969]

Lewis and Frank (2014). This work finds that languages tend to map more complex meanings (measured via semantic norms) to longer words. The bias is estimated as the correlation (Pearson's r) between word length and complexity ratings for a set of 499 words translated via Google Translate. We used estimates of the correlation that partialed out the effect of spoken frequency. [n = 79]

Wichmann, Rama, and Holman (2014). This database contains translations for 40-lexical items across many languages. Word length was calculated as the mean number of characters in the ASJPcode transcription system across words in each language. [n = 4421]

Aggregating across datasets, we analyzed 8 environmental variables in total: L2-L1 population ratio, total population size, number of neighbors, area of spoken region, mean and standard deviation temperature, and mean and standard de-

[1] WALS variable 59 was missing from the dataset.

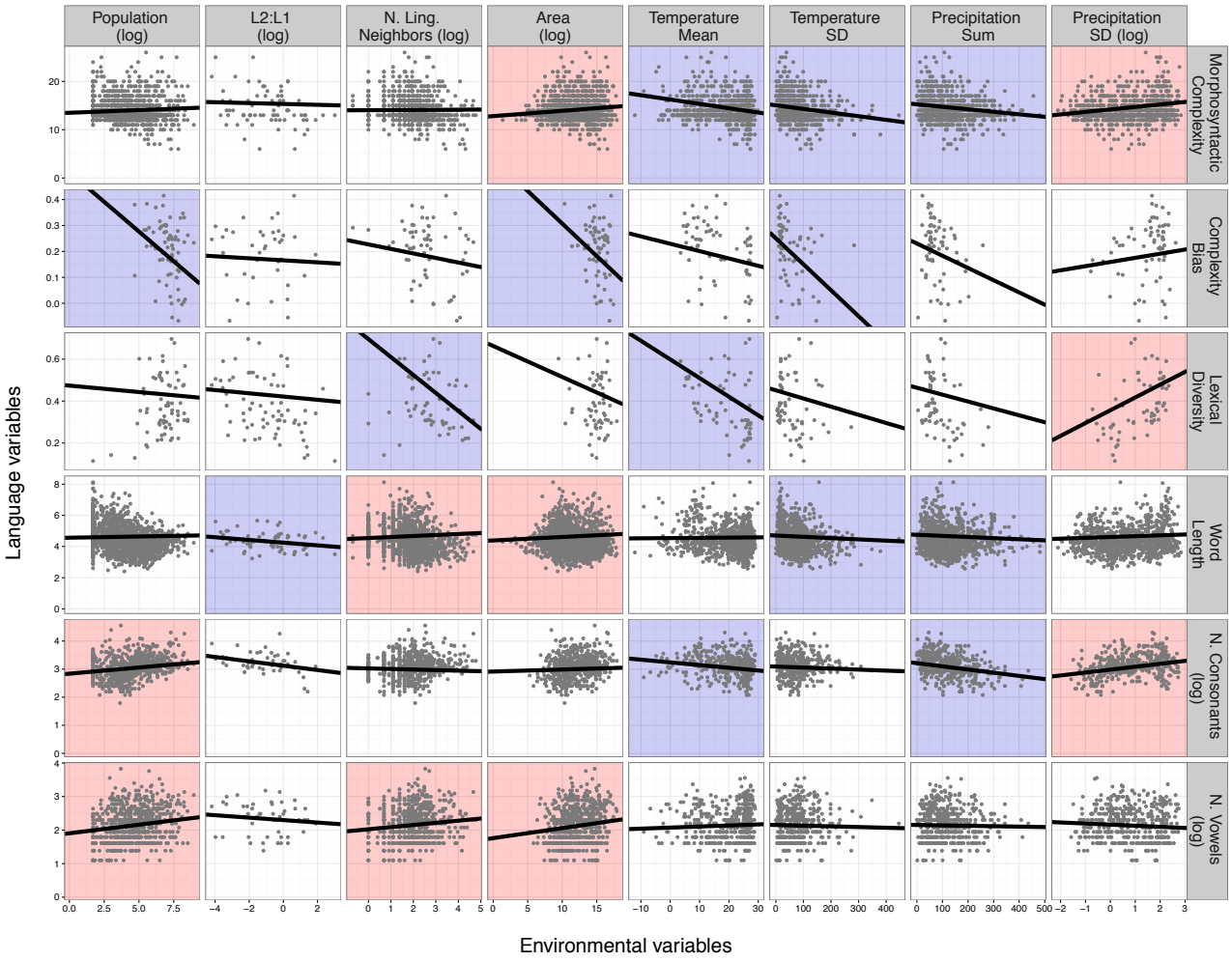


Figure 2: Relationship between environmental and linguistic variables, which each point represents a language. Red (positive) and blue (negative) indicate models where the environmental variable is a significant predictor of the linguistic variable. Lines show the fixed effect estimate (slope) and intercept of the mixed effect model. Number of languages varies across plots due to variation in the number of overlapping languages across datasets.

viation precipitation. These variables were selected from a larger set because they were not highly correlated with each other ($r < .8$). We analyzed 6 total linguistic variables: number of vowels, number of consonants, word length, type-token ratio, complexity bias, and morphosyntactic complexity.

Method

Datasets were merged using common ISO-639 codes when available. Five variables were log-transformed to better approximate a normal distribution (population, L2 to L1 ratio, number of neighbors, area, number of consonants, number of vowels).[2]

Main analysis We tested for a linear relationship between each environmental and language variable. A significant challenge in making inferences about language data is non-independence. This non-independence can come from at least two sources: genetic relatedness and language contact. Following Jaeger et al. (2011), we control for these factors statistically by using linear mixed-effects regression. We control for genetic non-independence by including a random intercept and slope by language family. We control for language contact by including country of origin as a random intercept (models with random slopes failed to converge).[3] We selected country of origin as a proxy for linguistic community because it was available for all languages in our dataset. Both control variables were taken from the WALS dataset. We considered a predictor significant if the test statistic on the fixed effect coefficient exceeded 1.96.

Principal component analysis This first analysis provides a uniform analysis of the many environmental and linguistic

[2]All code and data for the paper are available at `http://github.com/mllewis/langLearnVar`

[3]The model specification was as follows: `language.variable ~ environmental.variable + (environmental.variable | language.family) + (1 | origin.country)`.

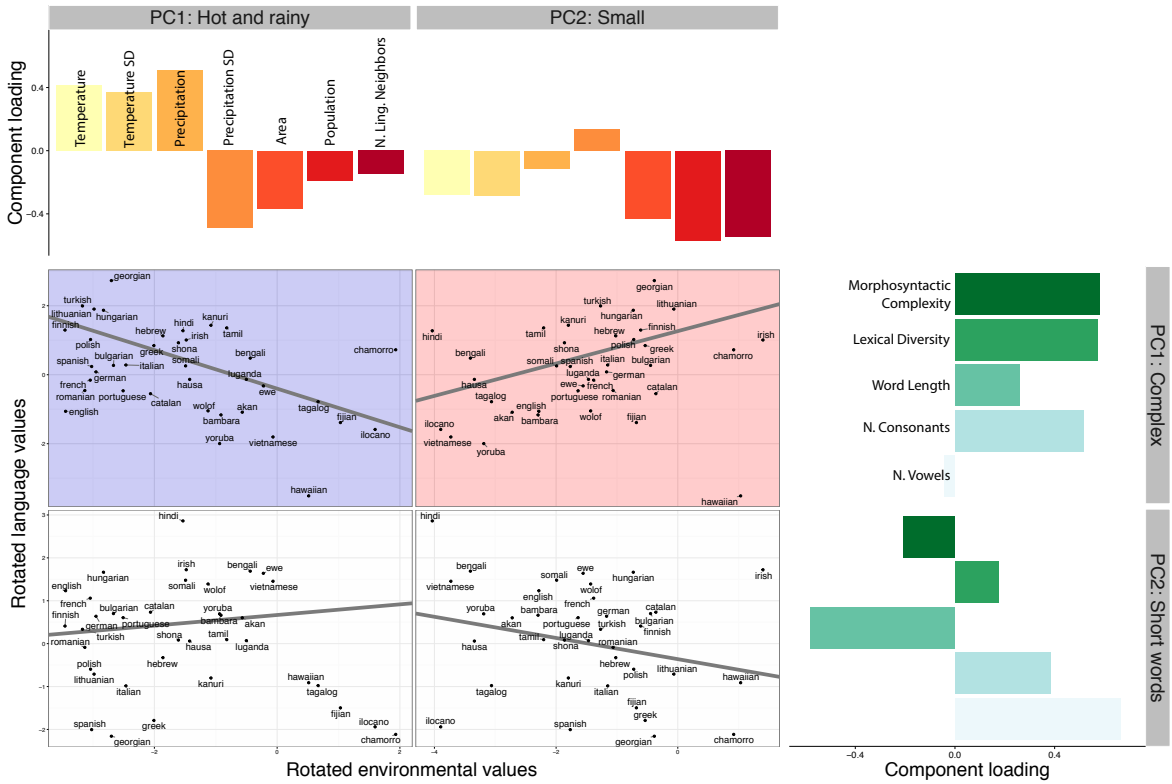


Figure 3: Languages spoken in cold, small regions tend to be more complex. The bar plots show the loadings on the first two principal components for the environmental variables (n = 7; orange) and language variables (n = 5; green). The scatter plots show the relationship between the first two principal components for both sets of variables. Each point corresponds to a language, and lines show the linear fit from the mixed effect model. Significance and direction of a linear relationship are indicated by the coloring of the scatterplot (blue: significant and negative; red: significant and positive).

variables that have been used to test the Linguistic Niche Hypothesis. However, the large number of variables makes it difficult to distill a coherent picture from these data. Given that many of these variables are partially correlated with each other, we used a technique for reducing the dimensionality of the dataset—principal component analysis. We found the principal components associated with the variance for the environmental variables and the linguistic variables, and then fit the same model as in the primary analysis using the rotated values. Complexity bias was excluded because it was only available for a small subset of languages. All variables were scaled.

Results

In the main analysis, we fit mixed effect models predicting each language variable with each environmental variable using areal controls. The results are presented in Figure 2. For each language variable, there was at least one environmental variable that reliably covaried, though some previously-reported effects were not significant in this analysis. We return to this in the discussion. Data can be explored interactively here: `https://mlewis.shinyapps.io/lhnn/`.

The principal component analysis revealed two primary components of variance for both the environmental and linguistic variables. For the environmental variables, the first two principal components accounted for .69 of the total variance (PC1: .39; PC2: .30). The weights on these variables across the two components can be seen in the upper panel of Fig. 3. The first component loads most heavily on variables related to the climate. It can be thought of as corresponding to hot and rainy regions. The second component loads most heavily on variables related to the size of the region a language is spoken in, both in terms of number of speakers and physical size. This principal component can be roughly interpreted as the 'smallness' of a linguistic community.

For the linguistic variables, the first two components also accounted for most of the variance, .70 (PC1: .39; PC2: .31; right panel of Fig. 3). The first component loads positively on all variables, except number of vowels. In particular, this component is associated with more consonants, longer words, more word types, and greater morphosyntactic complexity. Broadly, this component is related to the amount of cognitive difficulty associated with learning a language. The second component is associated with having short words, but large phonemic inventories.

Figure 3 shows the relationship between the principal components. Both environmental principal components were reliable predictors of the first linguistic principal component

(PC1: $\beta = -0.56$, $t = -3.52$; PC2: $\beta = 0.47$, $t = 2.08$). This suggests that languages that tend to be spoken in cold and small regions are more likely to be more complex. Neither of the environmental principal components were reliable predictors of the second linguistic principal component.

Discussion

These two analyses suggest that more complex languages are spoken in cold, small regions. Importantly, we find this relationship across a range of linguistic features—morphosyntactic complexity, linguistic diversity, word length, and consonant inventory—using the same analytic technique across all measures.

This finding is broadly consistent with previous work that finds relationships between individual metrics of complexity and various demographic variables. Nevertheless, we find null effects for several reported relationships in the literature. For example, the relationship between population size and morphosyntactic complexity (Lupyan & Dale, 2010) is not reliable in our model with areal controls, though the correlation is significant (r =.08; p <.001) and we replicate their finding in a binned analysis (Fig. 3 of Lupyan & Dale, 2010). There are many possible reasons for these differences (e.g., different measure of complexity, different areal controls), highlighting the need for a common analytical approach across datasets.

Why might languages in small, cold regions have more complex languages? One possible mechanism is that languages spoken in larger places have more L2 learners, and that L2 learners are less skilled than L1 learners at acquiring complex language. As a result, these languages adapt by simplifying. The relationship between climate and linguistic complexity is less clear, but one possibility is that speakers in colder regions are less itinerant, and therefore have less contact with adult speakers of other languages.

Study 2: Variability in L1 learning

The proposed mechanism in Study 1 makes an important assumption: L2 learners, but not L1 learners, are poor learners of linguistic complexity. Lupyan and Dale (2015) have argued that morphological complexity in fact *facilitates* learning for L1 learners by providing redundancy in the linguistic signal. A straightforward prediction of this hypothesis is that languages that are more easily learnable by L2 learners will be less learnable by L1 learners.

In Study 2, we explore this prediction. As a proxy for language learnability for L1 learners, we use the mean age of acquisition (AoA) of words in a language by L1 learners (children). If there is a tradeoff between learnability for L1 and L2 learners, languages that are less complex should be harder for children to learn, and thus have later AoAs.

Method

We use subjective measures of AoA from the the Łuniewska et al. (2015) norms. These AoAs were collected from adult participants for the translation equivalents of 299 words in 25 languages. To evaluate the validity of this measure, we compared these ratings to more objective measures of AoA collected from parent-report using the CDI (Wordbank; Frank, Braginsky, Yurovsky, & Marchman, in press). We fit a model predicting the objective ratings with the subjective ratings for the small sample of common languages ($n = 7$). We included language as a random by-intercept and by-slope effect. Subjective ratings were a strong predictor of objective ratings ($\beta = 1.00$, $t = 5.45$), suggesting that the Łuniewska et al. (2015) norms were a reasonable proxy for cross-linguistic AoA.

We averaged across words in the Łuniewska et al. (2015) database to get a mean AoA for each language. We then used the same mixed-effect model as in Study 1 to predict AoAs with each of the linguistic and environmental variables analyzed in Study 1.

Results

Number of consonants positively predicted AoA, suggesting that languages with more consonants tend to have later AoAs ($\beta = 1.04$, $t = 1.97$). In addition, temperature positively predicted AoA ($\beta = .13$, $t = 2.57$) and variability in precipitation negatively predicted AoA ($\beta = -1.6$, $t = -2.83$). No other variables were significant predictors of AoA.

Discussion

Study 2 explores a prediction about a mechanism for the relationship between population size and linguistic complexity: L1 learning is facilitated by complexity in the linguistic signal (via redundancy), but L2 learning is hindered. We find only limited support for this proposal. Of the factors that loaded on "complexity" in the principal component analysis in Study 1, number of consonants was the only reliable predictor of AoA.

Nevertheless, we do find several surprising correlates of AoA—number of consonants, temperature, and precipitation variability—even in this very small sample of languages. The mechanism underlying these relationships is not clear. It could be for example that L1 learners in colder regions have more language input, and therefore earlier AoAs. Or, if we assume that temperature and variability in precipitation are proxies from L2 pressure (as suggested in Study 1), it could be that languages with more L2 pressure have later AoAs, and therefore are harder for L1 learners to acquire. A larger sample of languages will be needed to address these questions.

Conclusion

Languages vary in many ways across multiple timescales of analysis. Here we suggest that this variability can be accounted for by considering the relationship between these timescales and two types of pressures, those internal and external to the cognitive system. In the present work, we have explored a hypothesis at the language evolution timescale—the Linguistic Niche Hypothesis—which suggests that cross-linguistic variability is the result of different cognitive constraints of learners and environmental pressures.

We contribute to the empirical findings related to this hypothesis by synthesizing previous correlational evidence using common analytical techniques across datasets. Across a range of linguistic and environmental metrics, we find that more complex languages tend to be spoken in smaller, colder regions. We also find evidence that the learnability of a language for L1 learners may be related to aspects of the language (number of consonants) and the environment (temperature and variability in precipitation). Understanding how *both* child and adult learners shape language systems is an important question for future work.

Accounting for variability at the timescale of language evolution is an empirically challenging enterprise. Moving forward, we suggest that a fruitful avenue for progress is holistic descriptions of the empirical landscape, and appeals to processes at multiple timescales of analysis.

Acknowledgments

We would like to thank Gary Lupyan and Rick Dale for sharing their data with us.

References

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*, 346–349.

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment information theory shows why we have only three lightness terms. *Psychological Science*, *20*, 1100–1107.

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*, 874–890.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers have fewer word forms.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*, 1–27.

Blythe, R. A. (2015). Hierarchy of scales in language dynamics.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*, 1131–1157.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `http://wals.info/`

Frank, M., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578.

Frank, M., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Gordon, R. (2005). Ethnologue: Languages of the world. SIL International.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Hay, J., & Bauer, L. (2007). Phoneme inventory size and population size. *Language*, *83*, 388–400.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context*, *42*.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*, 281–320.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Lewis, M., & Frank, M. C. (2016). Learnability pressures influence the encoding of information density in the lexicon learm. In *The Evolution of Language: Proceedings of the 11th International Conference.*

Lewis, M., Sugarman, E., & Frank, M. C. (2014). The structure of the lexicon reflects principles of communication. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowsk, B., Southwood, F., Pomiechowska, A., ... others (2015). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, *5*.

Lupyan, G., & Dale, R. (2015). The role of adaptation in understanding linguistic diversity. *The Shaping of Language: The Relationship between the Structures of Languages and their Social, Cultural, Historical, and Natural Environments*.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831.

Moran, S., McCloy, D., & Wright, R. (2012). Revisiting population size vs. phoneme inventory size. *Language*, *88*, 877–893.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1829–1836.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436–1441.

Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology*, *15*, 177–197.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543–578.

Know Your Adversary: Insights for a Better Adversarial Behavioral Model

Yasaman D. Abbasi[1], Noam Ben-Asher[3], Cleotilde Gonzalez[2],
Debarun Kar[1], Don Morrison[2], Nicole Sintov[1], Milind Tambe[1]
[1]{ydehghan,dkar,sintov,tambe}@usc.edu; [2]{coty, dfm2}@cmu.edu; [3]nbenash@us.ibm.com
[1]941 Bloomwalk, SAL 300, University of Southern California, SAL (300), Los Angeles, CA 90089, USA
[2] Social and Decision Sciences, 5000 Forbes Avenue, BP 208, Carnegie Melon University, Pittsburg, PA 15213, USA
[3]US Army Research Labs & IBM T.J.Watson Research Center, 1101 route 134 Kitchawan Rd, Yorktown Heights, NY 10598

Abstract

Given the global challenges of security, both in physical and cyber worlds, security agencies must optimize the use of their limited resources. To that end, many security agencies have begun to use "security game" algorithms, which optimally plan defender allocations, using models of adversary behavior that have originated in behavioral game theory. To advance our understanding of adversary behavior, this paper presents results from a study involving an opportunistic crime security game (OSG), where human participants play as opportunistic adversaries against an algorithm that optimizes defender allocations. In contrast with previous work which often assumes homogeneous adversarial behavior, our work demonstrates that participants are naturally grouped into multiple distinct categories that share similar behaviors. We capture the observed adversarial behaviors in a set of diverse models from different research traditions, behavioral game theory, and Cognitive Science, illustrating the need for heterogeneity in adversarial models.

Keywords: Human Behavioral Modeling, Opportunistic Security Game, Cognitive Models, Heterogonous Adversaries

Introduction

Given the global challenges of security, optimizing the use of limited resources to protect a set of targets from an adversary has become a crucial challenge. In terms of physical security, the challenges include optimizing security resources for patrolling major airports or ports, screening passengers and cargo, scheduling police patrols to counter urban crime (Tambe 2011; Pita et al., 2008; Shieh et al., 2012). The challenge of security resource optimization carries over to cybersecurity (Gonzalez, Ben-Asher, Oltramari & Lebiere, 2015), where it is important to assist human administrators in defending networks from attacks.

In order to build effective defense strategies, we need to understand and model adversary behavior and defender-adversary interactions. For this purpose, researchers have relied on the insights from Stackelberg Security Games (SSGs) to provide ways to optimize defense strategies (Korzyk, Conitzer, & Parr, 2010; Tambe, 2011). SSGs model the interaction between a defender and an adversary as a leader-follower game (Tambe 2011). A defender plays a particular defense strategy (e.g., randomized patrolling of airport terminals) and then the adversary takes an action after having observed the defender's strategy. Past SSG research often assumed a perfectly rational adversary in computing the optimal defense (mixed or randomized) strategy. Realizing the limitation of this assumption, recent SSG work has focused on bounded rationality models from behavioral game theory, such as the Quantal Response behavior model (McFadden 1976, Camerer 2003), but typically a homogeneous adversary population is assumed, and a single adversary behavior model is prescribed (Kar et al., 2015).

In contrast to this previous work which often assumes a homogeneous adversary population with a single behavioral model, this paper focuses on the heterogeneity in adversary behavior. Our results are based on the study conducted in Opportunistic Security Games (OSGs). In that experiment, (Abbasi et al., 2015) evaluated behavioral game theory models assuming a homogeneous adversary population. However, our results show that adversaries can be naturally categorized into distinct groups based on their attack patterns. For instance, while one group of participants (about 20% of the population) is seen to be highly rational and taking reward maximizing action, another group (nearly 50%) is seen to act in a completely random fashion. We show through experiments that considering distinct groups of adversaries leads to interesting insights about their behavioral model, including the defender strategies being generated based on the learned model.

There are two strands of previous work related to this paper. First, in behavioral game theory models in security games, mostly homogenous adversary models have been studied, but some recent research has considered the heterogeneity of human adversarial behavior. They have achieved it by either assuming a smooth distribution of the model parameters for the entire adversary population (Yang et al., 2014), such as a normal distribution or by utilizing a single behavioral model for each adversary (Haskell et al., 2014; Yang et al., 2014). However, they have not categorized the adversaries into distinct groups based on their attack patterns. In this paper, we show that adversaries can be categorized into multiple distinct groups, and each such group can be represented by distinct degrees of rationality.

The second strand of related work is with respect to the exploration of available options, which is an important aspect of decision making in many naturalistic situations (Pirolli & Card, 1999; Todd, Penke, Fasolo, & Lenton, 2007; Gonzalez & Dutt, 2011). In line with previous work (Hills & Hertwig, 2010; Gonzalez & Dutt, 2012), in this paper, we show that there is a negative relationship between exploration behavior and maximization of rewards. However, in their work, they did not contrast behavioral models with cognitive models and did not provide insights for behavioral game theory models which we provide. In particular, we study the relationship between exploration and human reward maximization behavior by parameters of bounded rationality models of

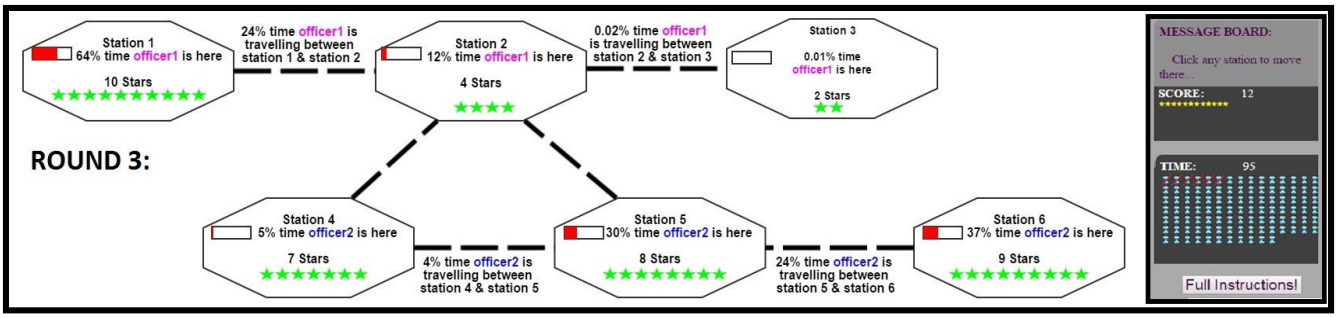


Figure 1. Game Interface

human adversaries. Our observations are also with respect to the security games domain where this kind of relationship between exploration behavior and maximization of rewards has not been studied before. Furthermore, in our work participants were shown all relevant information, such as rewards about all the alternative choices, while in earlier work participants had to explore and collect information about various alternatives.

To model the different categories of human behavior, we provide a family of behavioral game theory and cognitive models. In behavioral game theory models, we have explored models such as the popular Quantal Response (McKelvey & Palfrey 1995) and the Subjective Utility Quantal Response models (Nguyen et al., 2013). These models have been shown to successfully capture human rationality in decision making in the security games domain (Tambe, 2011). In addition, based on the tradition of Cognitive Science, we use a model derived from a well-known cognitive theory, the Instance-Based Learning Theory (IBLT) (Gonzalez, Lerch, & Lebiere, 2003), developed to explain human decision making behavior in dynamic tasks and used to detect adversarial behaviors (Ben-Asher, Oltramari, Erbacher & Gonzalez, 2015). This is the first such use of cognitive models in security games. In summary, in this paper we build on the existent literature of security games and adversary behavior modeling by: (i) investigating the heterogeneity of adversarial behavior in an experimental study designed for OSGs, by categorizing adversaries into groups based on their exploration patterns; (ii) comparing computational models and showing the impact of heterogeneity on future behavior prediction; and (iii) showing the impact of considering heterogeneity on the defender strategies generated.

A behavioral study in an OSG

To collect data regarding adversarial behavior from playing an OSG repeatedly, we used data collected from experiments by (Abbasi et al., 2015) using a simulation of urban crime in a metro transportation system with six stations (Figure 1).

Methods

Game Design. The players' goal is to maximize their score by collecting rewards (represented by stars in Figure 1) while avoiding officers on patrol. Each player can travel to any station, including the current one, by train as represented by the dashed lines in Figure 1.

There are two moving officers, each protecting three stations. The probability of their presence at each station or route, i.e. patrolling strategy, is determined beforehand using an optimization algorithm similar to the one presented in (Zhang et al., 2014). The algorithm optimizes defender strategies given an opportunistic adversary behavior model.

The stationary coverage probabilities for each station and trains are revealed to the players. This means that players can see the percentage of the time that officers spend on average at each station and on the train, so they can determine the chance of encountering an officer at a station. However, during the game, the players cannot observe where officers are actually located unless they encounter the officer at a station.

The game can finish either if the player uses up all the 100 units of available time in each game, or the game is randomly terminated after a station visit, which may happen with a 10% probability after each station visit. The random termination encourages players to choose each action carefully, as there is a chance the game may terminate after each visit.

The player's objective is to maximize his total reward in limited time. Players must carefully choose which stations to visit, considering the available information about rewards, officers' coverage distribution on stations and time to visit the station.

Procedures. Each participant played eight games in total; starting with two practice rounds to become familiar with the game, followed by two validation rounds (two simple versions of the main games), in which the participants were presented with obvious choices to confirm they understood the rules and game's objective, and finally, four main games from which we collect and use the data for our analyses. To ensure that the collected data is independent of the graph structure, the four main games were played on four different graphs, presented in a random order to the participants. Each graph had six stations with a different route structure and patrolling strategy.

Participants. The participants were recruited from Amazon Mechanical Turk. They were eligible if they had previously played more than 500 games and had an acceptance rate of minimum 95%. To motivate the subjects to participate in the study, they were rewarded based on their total score ($0.01 for each gained point) in addition to a base compensation ($1). In total, 70 participants took part in the

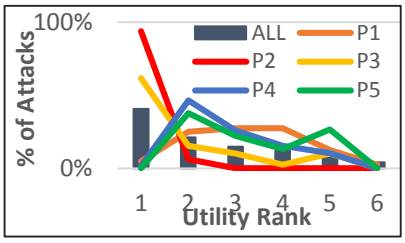


Figure 2: % of Attacks on utility rank

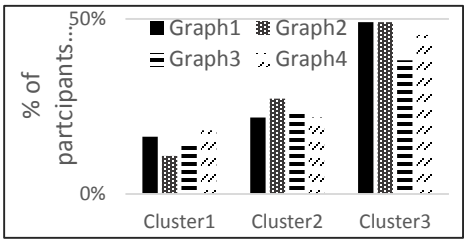


Figure 3: Clustering Distribution

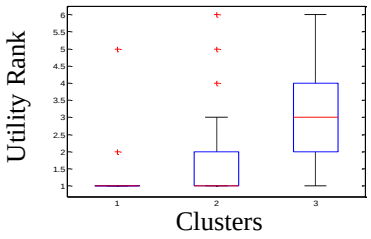


Figure 4: Utility Rank by Cluster

game and went through a validation test. Data from 15 participants who did not pass validation were excluded.

Human Adversarial Behavior

Using data from all the main games, Figure 2 illustrates the distribution of attacks (i.e., moves) from all participants (black bars) on stations ranked by the participants' expected utility (average earning per time)[1], as well as attacks of five randomly selected individuals (P1 to P5). To normalize the utility scores among graphs, we have used the ranking of stations' utility (utility rank) instead of its absolute value (the highest utility in a graph is ranked 1). The graph illustrates significant heterogeneity behavior among individuals (line charts), and comparison to the average behavior (bar chart).

Given this heterogeneous behavior, we have applied the Hierarchical Clustering algorithm (Johnson, 1967) on different features related to an individuals' exploration behavior and found that *mobility score* was the best feature to cluster the participants. The *mobility score* is a measure of exploration: it is a ratio of the number of movements between stations over the number of trials (total number of movements) by a participant in the game. Figure 5: % of participants based on the Mobility Score

shows the distribution of participants based on their mobility score for each graph. The mobility score varied widely (0% to 100%) with a significant proportion of participants at the two extremes. Informally, the exploration behavior seems to fall into three categories: (i) those who did no exploration; (ii) those who always explored and (iii) those who engaged in a middling level of exploration. Indeed, the clustering algorithm resulted in three groups of participants: participants whose mobility score is less than 10% belong to Cluster1, participants with 10% to 80% mobility score belong to Cluster2, and participants whose mobility score is greater than 80% belong to Cluster3.

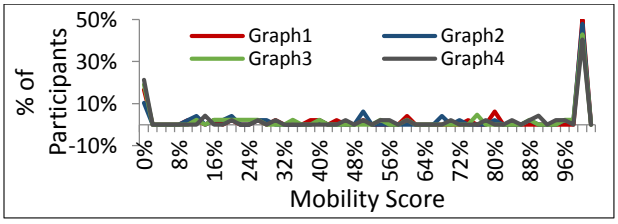


Figure 5: % of participants based on the Mobility Score

[1] $EU = (1 - stationary\ coverage) * reward/\ time$

Figure 3 shows the percentage of participants belonging to each cluster for four different graphs (Graph 1 to Graph 4). The percentage of participants belonging to each cluster is nearly consistent across all graphs: approximately, 20% in Cluster1, 30% in Cluster2 and 50% in Cluster3.

In Figure 4, using the data from all the graphs per cluster, we show the distribution of utility ranks for each of the three clusters. Interestingly, mobility scores were highly correlated with the utility ranks of the attacked stations ($R^2 = .85$ & $p < .01$). We observe that participants in Cluster1 (the lowest mobility scores), attacked stations with the highest utility (average utility rank of 1.04). In contrast, participants in Cluster3 (the highest mobility score), attacked stations that varied more widely in the utility rank (average utility rank of 3.3). Participants in Cluster2 also attacked a variety of stations but were leaning (on average) towards higher utility rank stations (average utility rank of 1.7). These observations provide interesting insights for building defender strategies, as illustrated in Section Model Results.

Models of Adversarial Behavior in OSG

In what follows, we present a series of models that have been proposed recently to represent adversarial behavior.

Quantal Response Model (QR)

Quantal Response models the bounded rationality of a human player by capturing the uncertainty in the decisions made by the player (McKelvey & Palfrey 1995; McFadden 1976). Instead of maximizing the expected utility, QR posits that the decision-making agent chooses an action that gives high expected utility, with probability higher than another action which gives a lower expected utility. In the context of OSG, given the defender's strategy s (e.g., stationary coverage probability at station i (s_i) shown in Figure 1), the probability of the adversary choosing to attack target i when he is in target j and when the defender's coverage is s, $q_{i,j}(s)$, is given by the following equation:

$$q_{i,j}(s) = \frac{e^{\lambda * EU_{i,j}(s)}}{\sum_{1 \le k \le 6} e^{\lambda * EU_{(k,j)}(s)}}$$

where λ is his degree of rationality and $EU_{i,j}(s)$ is the expected utility of the adversary as given by:

$$EU_{i,j}(s) = \frac{r_i}{time\ (i,j)} * (1 - s_i)$$

where r_i is the number of stars at station i, $time$ (i,j) refers to time taken to attack station i when player is in station j

Subjective Utility Quantal Response (SUQR)

The SUQR model combines two key notions of decision making: Subjective Expected Utility, SEU, (Fischhoff et al., 1981) and Quantal Response; it essentially replaces the expected utility function in QR with the SEU function (Nguyen et al., 2013). In this model, the probability that the adversary chooses station i when he is at station j, when the defender's coverage is s, is given by $q_{i,j}(s)$. $SEU_{i,j}$ (s) is a linear combination of three key factors. The key factors are (a) r_i, (b) s_i, and (c) $time_{i,j}$, $w = <w_r, w_{sta}, w_{time}>$ denotes the weights for each decision making feature:

$$q_{i,j}(s) = \frac{e^{SEU_{i,j}(s)}}{\sum_{t' \in T} e^{SEU_{i,t}(S)}} \text{ where}$$

$$SEU_{i,j}(s) = w_r . r_i + w_{sta} . s_i + w_{time} . time_{i,j}$$

Instance-Based Learning Model

The IBL model of an adversary in the OSG makes a choice about the station to go to, by first applying a randomization rule at each time step:

If draw form U(0,1) >= Satisficing threshold
Make a random choice
Else;
Make a choice with the highest Blended value.

This rule aims at separating highly exploratory choices from those made by the satisficing mechanism of the IBL, the Blended Value. *Satisficing* is a parameter of this model. The Blended value *V* represents value of attacking each station (option *j*):

$$V_j = \sum_{i=1}^{n} p_{ij} x_{ij}$$

where x_{ij} refers to the value (payoff) of each station (the number of stars divided by time taken) stored in memory as instance *i* for the station *j*, and p_{ij} is the probability of retrieving that instance for blending from memory (Gonzalez & Dutt, 2011; Lejarraga et al., 2012) defined as:

$$p_{ij} = e^{\frac{A_i}{\tau}} \Big/ \sum_l e^{\frac{A_l}{\tau}}$$

where l refers to the total number of payoffs observed for station *j* up to the last trial, and τ is a noise value defined as $\sigma \cdot \sqrt{2}$. The σ variable is a free noise parameter. The activation of instance *i* represents how readily available the information is in memory:

$$A_i = ln \sum_{\substack{t_p \\ \epsilon \text{ observed}}} (t - t_p)^{-d} + \sum_{\substack{Attribute \\ \epsilon \text{ Situation}}} P(M_{Attribute} - 1) + \sigma \ln(\frac{1 - \gamma_{i,t}}{\gamma_{i,t}})$$

Please refer to (Anderson & Lebiere, 1998) for a detailed explanation of the different components of this equation. The Activation is higher when instances are observed frequently and more recently. For example, if an unguarded nearby station with many starts (reward) is observed many times, the activation of this instance will increase, and the probability of selecting that station in the next round will be higher. However, if this instance is not observed often, the memory of such station will decay with the passage of time (the parameter *d*, the decay, is a non-negative free parameter that defines the rate of forgetting). The noise component σ is a free parameter that reflects noisy memory retrieval.

Model Results

We aggregated the human data and divided the data set into two groups: training and test datasets. The data from the first three graphs played by the participants were used for training and the last graph played was used for testing the models. This resulted in 1322 instances in the training set and 500 instances in the test data set.

For comparison of different models, we use Root Mean Squared Error (RMSE) and Akaike Information Criterion (AIC) metrics. RMSE represents the deviation between model's predicted probability of adversary's attack ($\hat{p}$) and the actual proportion of attacks of participants from each station to others (p).

$$RMSE(\hat{p}) = \sqrt{MSE}(\hat{p}) \text{ where MSE } (\hat{p}) = \frac{1}{n}\sum(\hat{p} - p)^2$$

AIC provides a measure of the relative quality of statistical models; the lower the value, the better the model. The metric rewards goodness of fit (as assessed by the likelihood function), and penalizes overfitting (based on a number of estimation parameters).

$$AIC = 2 * \#\, model's\ pramters - 2 * \ln(likelihood)$$

Table 1 shows the results on the full data set. The model parameters obtained from the training data set were used to make predictions on the test dataset. The prediction errors from all the models are relatively similar, even though they provide different perspectives. QR and SUQR predict the stable state transition probabilities of the attacker while the IBL is a process model that captures learning and decision dynamics over time. We also examine the parameter values and performance of the models for each cluster (Table 2).

Table 1: Metrics and Parameter on the full data set

Model	*Parameters*	*RMSE*[2]	*AIC*
QR	0.4188	0.25	3962
SUQR	<3.97,-2.51,-2.55>[3]	0.23	3685
IBL	<1.4, 3.2,0.3>[4]	0.24	4359

The value of λ (higher value of λ corresponds to higher rationality level) in the QR model decreases significantly from Cluster1 (high value of λ=1.81) to Cluster3 (λ=0). These findings are consistent with our observation of the utility

[2] the average is over 288 entries, representing moves from any of 6 stations to any other station, in four graphs, and for two cases where the player observes the officer or not

[3] $w = <w_{re}, w_{sta}, w_{time}>$

[4] <noise, decay, Satisficing threshold >

ranks of targets chosen by adversaries in each cluster, as shown in Figure 5. This is significant because past research has assumed that all participants either behave based on an average value of λ or that each individual's value of λ can be sampled from a smooth distribution. In this study, however, we show that a significant number of participants (70%: 20% in Cluster1 plus 50% in Cluster3) have values of λ which fall at two extreme ends of the spectrum, thus modeling perfectly rational and completely random adversaries respectively.

Moreover, considering the fact that SUQR weights indicate the importance of each attribute to the decision maker, the results of SUQR parameter extraction for different clusters reveal some interesting points. First, the fact that Cluster1 has the largest weights for all attributes (in the absolute terms) implies that Cluster1 participants are very attracted to the stations with high rewards and highly repelled by high defender coverage; which conforms with the observed behavior of Cluster1 participants in maximizing the expected utility. Second, although SUQR outperforms QR overall and in Cluster2 and 3, QR has lower prediction error (statistically significant for paired t-test at $t(288) = 02.34$, $p<0.01$) on data for Cluster1. This is intuitive if participants are utility maximizers, this would be captured better when in the QR model. On the other hand, a model like SUQR, which reasons based on different features of the game capture better the propensity of the participants to switch between stations, and hence perform better on Clusters 2 and three where participants do not have a clear movement pattern. Therefore, identifying different groups of adversaries gives us valuable insight into the types of behavioral models that can be used in different scenarios to generate accurate future predictions

Table 2: Metrics and Parameters on each Cluster

	Model	*Parameters*	*RMSE*	*AIC*
Cluster 1	QR	1.81	0.01	52
	SUQR	<7.16,-4.53, -13.43>[3]	0.06	67
	IBL	<2.3, 0.9, 0.9>[4]	0.27	238
Cluster2	QR	0.6582	0.28	1023
	SUQR	<5.63,-3.14 ,-4.16>[3]	0.27	927
	IBL	<0.9, 1.4, 0.8>[4]	0.30	1821
Cluster 3	QR	0	0.26	2188
	SUQR	< 1.9,-1.1,0.13>[3]	0.23	2007
	IBL	<0.01, 1.8, 0.1>[4]	0.27	2529

The results from the IBL model suggest that the categories of adversaries found in this study do not emerge naturally from the learning process. Indeed, in this study participants had little opportunities to learn. Instead, it appears that participants either use the information readily available to them in the OSG and attempt to maximize their gains, or they explore the choices randomly which may lead them to less optimal decisions. These two modes of behavior were captured in the IBL model by a meta-rule with a Satisficing parameter. This meta-rule is not part of the IBL model, but it helps to overpass the natural choice by Blending (similar to the Inertia meta-rule used in Gonzalez & Dutt, 2011). This meta-rule was added to explicitly account for random exploratory behavior observed in the OSG. Therefore, the Satisficing parameter helps in selecting between the two modes of behavior to form the different clusters. The Satisficing parameter is highest in Cluster1, lower in Cluster2, and lowest in Cluster3. Cluster1 results from most choices being made by the IBL's Blending while Cluster3 results from a random choice. However, this parameter interacts with the IBL model's decay and noise parameters. For example, in Cluster1, most decisions are made for the station with highest Blended value, and there is a need for a high noise value to introduce the variability found in human behavior. In contrast, choices in Cluster3 are mostly done randomly, but in the rare occasions when the model makes choices based on the highest Blended value, it attempts to benefit from recent past experiences (i.e., low decay) and with low noise to the decision processes. Therefore, identifying such meta-rules for accounting for explicit descriptive information in addition to the IBL model's learning mechanisms is an important aspect of capturing adversary behavior in security games.

It is interesting to observe that the behavioral game theory models provide a significantly better fit in Cluster1, compared to the IBL cognitive model, while the values of behavioral game theory models are comparable to those of the IBL model in Clusters 2 and 3. The IBL model, being a learning model, is poor at making highly accurate decisions with little or no experience as in the OSG study.

Finally, to demonstrate the impact of considering distinct heterogeneous groups of adversaries, we consider one of the most recent works (Kar et al., 2015) which advocated the use of a homogeneous adversary model. We show on data collected from their domain that there is a significant difference between the defender strategies generated by a homogeneous (SUQR) and a heterogeneous model which considers three distinct clusters (Bayesian SUQR). The bar charts in Figure 6 shows the percentage of change in defender strategy, for example, for target 16, the change in coverage probability from defender strategy generated against a homogeneous to that against a heterogeneous model is 110%.

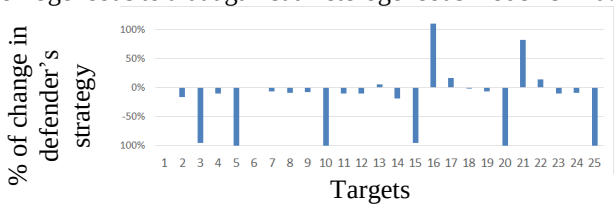


Figure 6: Strategy against homogenous & heterogeneous

Conclusions

Significant research has been conducted towards understanding adversary behavior in security games, which has led to several deployed real-world applications (Tambe 2011), such as PROTECT for the protection of major ports in

the US (Shieh et al. 2012) and ARMOR for scheduling of police patrols at major airports such as LAX (Pita et al. 2008). Although researchers in security games have relied on modeling adversaries via a single homogeneous model, or a heterogeneous model with a smooth distribution over model parameters, in this paper, we showed the heterogeneity in adversary behavior by clustering adversaries into distinct groups based on their exploration patterns. Three clusters emerged based on the adversaries' exploration patterns, two of which fall at two extreme ends of the parameter spectrum, capturing perfectly rational and completely random behavior. We also observed that in our OSG domain, exploration is negatively correlated with utility maximization.

We demonstrate that accounting for the diversity of adversary behavior leads to different model parameters and can provide more accurate predictions of future behavior. Specifically, we show on data collected based on an Opportunistic Security Game that: (i) QR captures the behavior of utility maximizing adversaries much better than SUQR or IBL based models; (ii) the behavioral and cognitive models have similar prediction performance for adversaries who do not act in a perfectly rational fashion. Furthermore, we show that considering the heterogeneity in adversary behavior leads to different defender strategies being generated. The effectiveness of such strategies is an important area of future work.

Acknowledgments

This research was partly supported by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA) to Cleotilde Gonzalez. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. This research is also supported by MURI grant W911NF-11-1-0332, and award no. 004525-00001 by US-Naval Research laboratory.

References

Abbasi, Y. D., Short, M., Sinha, A., Sintov, N., Zhang, Ch., Tambe, M. (2015). Human Adversaries in Opportunistic Crime Security Games: Evaluating Competing Bounded Rationality Models. *Advances in Cognitive Systems.*

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Lawrence Erlbaum Associates.

Ben-Asher, N., Oltramari, A, Erbacher, R.F., and Gonzalez, C. (2015). Ontology-based Adaptive Systems of Cyber Defense. *(STIDS).*

Camerer, C.F. (2003) *Behavioral game theory, Experiments in strategic interaction*. Princeton University Press

Fischhoff, B., Goitein, B., and Shapira, Z. (1981). Subjective utility function: A model of decision-making. *American Society of Information Science*.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review*.

Gonzalez, C., Ben-Asher, N., Martin, J. & Dutt, V. (2015). A cognitive model of dynamic cooperation with varied interdependency information. *Cognitive Science.*

Gonzalez, C., Ben-Asher, N., Oltramari, A., & Lebiere, C. (2015). Cognition and Technology. *Cyber defense and situational awareness*.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*.

Haskell, W., Kar, D., Fang, F., Tambe, M., Cheung, S., & Denicola, L. E. (2014). Robust protection of fisheries with compass. *IAAI*

Hills, T. T., & Hertwig, R. (2010). Information Search in Decisions From Experience Do Our Patterns of Sampling Foreshadow Our Decisions, *Psychological Science*

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, Chicago

Kar, D., Fang, F., Delle Fave, F., Sintov, N., Tambe, M. (2015) "A Game of Thrones": When Human Behavior Models Compete in Repeated Stackelberg Security Games. (AAMAS).

Korzhyk, D., Conitzer, V., & Parr, R. (2010, July). Complexity of Computing Optimal Stackelberg Strategies in Security Resource Allocation Games. InAAAI.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*.

McFadden, D. L. (1976). Quantal choice analaysis: A survey. In *Annals of Economic and Social Measurement*.

McKelvey, R.D., Palfrey, T.R. (1995) Quantal response equilibria for normal form games. *Games and Economic Behavior*.

Nguyen, T.M., Yang R., Azaria A., Kraus S., Tambe M. (2013). Analyzing the Effectiveness of Adversary Modeling in Security Games, In *AAAI*.

Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M., Western, C., and Kraus, S. (2008). ARMOR Security for Los Angeles International Airport. In AAAI

Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., ... & Meyer, G. (2012). Protect: A deployed game theoretic system to protect the ports of the United States. *AAMAS.*

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*.

Tambe, M. (2011). Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned. *Cambridge University Press.*

Yang, R., Ford, B., Tambe, M., & Lemieux, A. (2014). Adaptive resource allocation for wildlife protection against illegal poachers. *AAMAS*.

Zhang, C., Jiang, A.X., Short, M.B., Brantingham, J.P. and Tambe, M. (2014). Defending Against Opportunistic Criminals: New Game-Theoretic Frameworks and Algorithms *Gamesec*.

Semantic, Lexical, and Geographic Cues are used in Geographic Fluency

Janelle Szary (jszary@indiana.edu)
Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

Semantic fluency tasks have increasingly been used to probe the structure of human memory, adopting methodologies from the ecological foraging literature to describe memory as a trajectory through semantic space. Clusters of semantically related items are often produced together, and the transitions between these clusters of semantically related items are consistent with theories of optimal foraging, where the search process exhibits a balance between exploration and exploitation behaviors (Hills, Jones, & Todd, 2012). Here, we use a semantic fluency memory task in which subjects recall geographic locations. For each pairwise transition, we measure temporal, geographic, semantic, lexical, and phonetic distances. In general, the dimensions are loosely but reliably correlated with each other. Segmentation of the retrieval sequence into patches supports the notion that subjects strategically leave patches as within-patch resources diminish, but also suggests that subjects may shift their attention between different sources of information, perhaps reflecting dynamically changing patch definitions.

Keywords: Memory search; Semantic fluency; Optimal foraging; Spatial search

Introduction

In general, effective search strategies require a balance of both exploration and exploitation. The searcher, whether it be a bumblebee searching for nectar, or a child searching for hidden Easter eggs, needs to know when to stop looking in one place and start looking in another. Too much exploration, in the later example, would be a situation in which the child runs wildly around the yard, but never stops to looks behind any leaves (where the Easter eggs would be hidden); Too much exploitation, on the other hand, might be a situation in which the child, finding one egg hidden in the rose bushes, spends the afternoon meticulously sorting through the roses for more eggs, and fails to search the rest of the yard. Although this example is exaggerated, it illustrates the importance of finding an appropriate balance between the two strategies: At either extreme, the child will do very poorly in an Easter egg hunt. A successful search process relies on appropriately modulating between exploration and exploitation.

The same principle applies to search in the cognitive domain: At one end of the extreme, a lack of focus would make it hard to retrieve relevant information, and, at the other end of the extreme, perseveration on one piece of information would also be ineffective. The parallel between search in physical space and cognitive space is not coincidental. Significant evidence suggests that the neural and molecular processes that evolved to govern search in physical space have been exapted to control goal-directed behaviors in other, cognitive modalities (reviewed in Hills, 2006). For example, the dopaminergic pathways of the basal ganglia control both cognitive attention and movement, and increases in dopamine have been associated with exploitation of resources and highly focused behaviors, while reductions in dopamine have been associated with exploratory or inattentive behaviors (see Hills, 2006). The exaptation hypothesis has led to the idea that there are general cognitive search mechanisms that control goal-directed search behaviors in both internal and external spaces (Hills, Todd, & Goldstone, 2008). As evidence for this claim, Hills et al. (2008) demonstrated that priming certain physical search strategies can influence cognitive search behavior.

Further support for the exaptation hypothesis comes from the successful application of models from the animal foraging literature to cognitive search behavior. For example, in the foraging literature, a search process is considered optimal when it follows the marginal value theorem (MVT; Charnov, 1976), in which a forager should continue exploiting a patch of resources while it continues to provide relative rewards, but should leave and switch to a new patch once the rate of rewards (for the given patch) drops below a long-term average. Hills et al. (2012) used the semantic fluency task, in which subjects are asked to name as many items from a given category as possible, to show that cognitive search is well described by MVT, where patches are clusters of semantically related items. As a subject's relative success in a given semantic cluster decreases, the subject is more likely to switch to a new semantic cluster. Models of optimal foraging in semantic memory based on MVT have been successful in matching human performance, but they rely on the construction of appropriate semantic spaces and patches, which can be done in different ways with differing results (see discussion in Hills et al., 2015; Abbott, Austerweil, & Griffiths, 2015).

Anecdotally, the similarity between physical and cognitive search is supported by the fact that we often describe our internal, cognitive representations of information as networks or maps (Steyvers & Tenenbaum, 2005; Tolman, 1948). Further, there is growing support for the embodied cognition perspective that cognitive knowledge is grounded in physical space. Montez, Thompson, and Kello (2015) asked participants to spatially organize a set of items produced during a previous semantic fluency task, and found that the spatial distances correlated with the previously observed temporal distances: Subjects spatially organized items the same way they had previously organized them during semantic retrieval. Louwerse and colleagues (Recchia & Louwerse, 2014; Louwerse & Zwaan, 2009) have shown that from the statistics of language we can extract a great deal of information, such as the locations and sizes of cities.

Using an extended semantic fluency task where subjects spent twenty minutes retrieving the names of cities and towns from their home state of California, Szary, Kello, and Dale (2015) showed that the temporal structure of retrieval sequences captured the geographic structure of cities in physical space. Although the task is known as a 'semantic fluency' task, previously we could only infer semantic information from the timing of the retrieval process. Here, we build upon the study presented in Szary et al. (2015) by computing semantic, lexical, and phonetic measures to compare to the existing geographic and temporal measures. The new analyses allow us to formulate new research questions. Consider the hypothesis that internal search is exapted from external search, and relies on a balance between exploration and exploitation. Because physical space has a finite number of dimensions, it is quite easy to measure exploration versus exploitation behaviors in external search. By comparison, it is less straightforward how one should characterize exploration or exploitation behaviors in the high-dimensional cognitive space of an internal search. Here, we re-investigate semantic fluency data along a number of dimensions: temporal, geographic, semantic, lexical, and phonetic. In this exploratory study, we show that retrieval sequences do indeed contain structure along each of these dimensions. That is, when remembering geographic information, people do, indeed, use the embodied physical structure of the information, but we can also measure their use of semantic, lexical, and even phonetic information. With regards to the optimal foraging perspective, these metrics may reflect different dimensions in which patches exist, highlighting a number of open questions for future work, as discussed in the conclusion.

Methods

Experimental Procedure

Participants Participants were recruited from a subject pool of University of California, Merced undergraduate students who participated for course credit (4 male, 8 female; mean age = 19.92 years, *SD* = 1.08 years), and reported being native or proficient English speakers. All but two participants reported living in California for their whole lives, while the other two reported living in California for the majority of their lives (15 of 19 years, and 16 of 18 years, respectively). Subjects were comfortably seated by themselves at a table in small experiment room, and wore Shure microphone headsets. Speech was collected using an M-Audio MobilePre recording interface and Audacity software.

Task Subjects completed two recall tasks, presented in counterbalanced order, each of which lasted for twenty minutes. In one task subjects recalled items from the category of cities and towns in California, and in the other they recalled from the category of all animals. For the purposes of the present paper, we discuss only results from the category of cities and towns in California. Subjects were given the following instructions: "Your goal [is] to think of as many items from [the] category as you can. When you think of an item, just say it out loud. You can be as specific or as general as you wish. For example, if the category were *Food* you could say 'Fruit', and you could also say 'Orange' or 'Mandarin Orange'. But keep in mind that your goal is to recall as many different items as possible. If you are unsure if an item does or does not belong to the category, just say it anyhow, don't spend time worrying about whether something counts or not," (adapted from Rhodes & Turvey, 2007).

Audio Transcription The speech recordings were loaded into Praat audio analysis software for annotation, and were transcribed as in Szary et al. (2015). That is, each recalled item was transcribed, and onset times of items were marked. Repeated items were removed. Incorrect items ("Reno", which is in Nevada, not California), geographic landmarks ("Monterey Bay" bay, "Sierra Nevadas" mountains), and non-specific areas ("Bay Area", which refers to several locations around the San Francisco Bay) were removed. Pronunciation errors ("Rancho Cucamongo" instead of the correct "Rancho Cucamonga") and common abbreviations ("L.A." instead of the official "Los Angeles") were corrected. Districts, neighborhoods, planned communities and census-designated areas with names recognized by the U.S. Geological Survey (e.g. "Hollywood", "Downieville"; *United States Board on Geographic Names*, 2016) were retained.

Measures

For each dataset, inter-retrieval intervals (*IRIs*) are measured as the amount of time (in milliseconds) between consecutive recall events $city_i$ and $city_{i+1}$, and represent the temporal distances we observed. For each transition ($city_i$ to $city_{i+1}$), we compute additional distances using geographic, semantic, phonetic, and lexical measures.

Geographic Distance The latitudinal and longitudinal coordinates for most cities were provided by the world.cities dataset, in the maps package for the R programming language (Becker, Wilks, Brownrigg, Minka, & Deckmyn, 2016). Missing values were added by hand, and were retrieved from Wikimedia's GeoHack tool (*Wikimedia Tool Labs GeoHack*, 2016). Geographic distances (*GDs*) are measured as the number of miles between consecutively recalled cities. GDs are calculated using the Haversine formula, which gives the great-circle distance between two points on an sphere (Sinnott, 1984).

Semantic Distance Semantic similarity was calculated using vector representations obtained by a Word2Vec model from the gensim package for Python (Řehůřek & Sojka, 2010), which was trained on the full English Wikipedia dataset, available from *dumps.wikimedia.org/enwiki*, 2016. The cosine between the vectors representing two city names is taken as a measure of the semantic distance between those cities.

Lexical Distance The standard Levenshtein distance, which measures the number of edits required to transform one

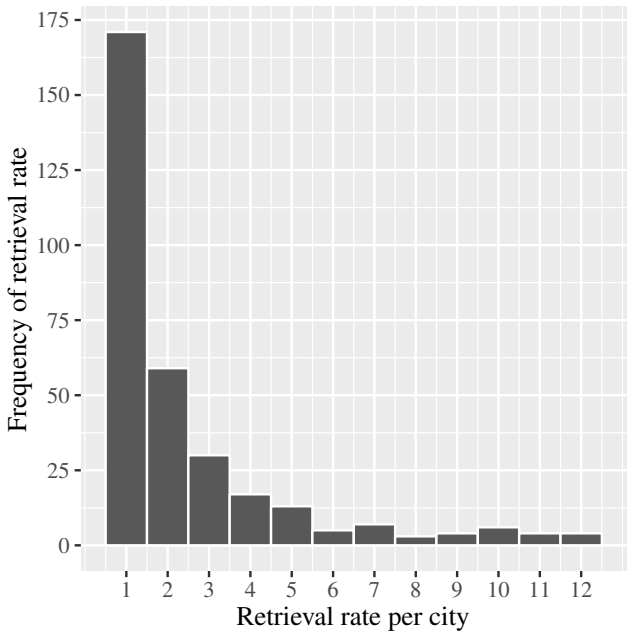


Figure 1: The number of cities that has each retrieval rate.

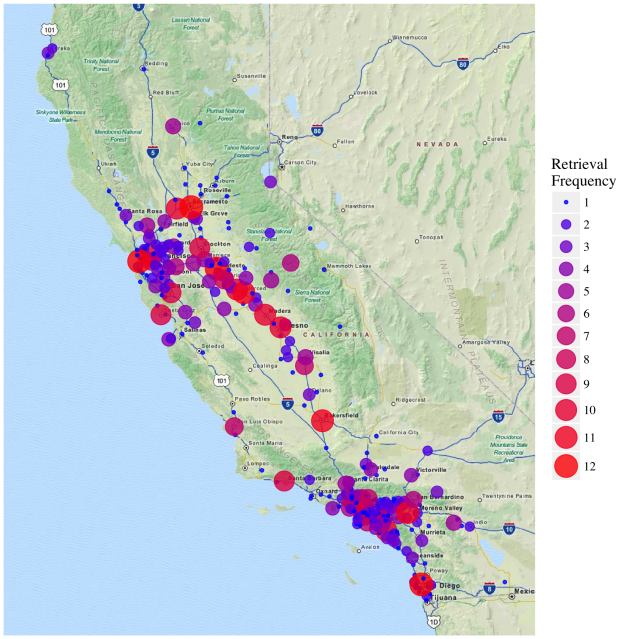


Figure 2: City locations and their retrieval rates.

string into another, is used as a measure of lexical distance.

Phonological Distance Finally, as an exploratory measure of phonological difference, we converted city names into their four-digit soundex codes, which are approximate representations of how words sound. We then compute the Levenshtein distance between these codes, which serves as our measure of phonological distance.

Results

Across all participants, a total of 323 unique city names were retrieved. On average, each participant retrieved 67 cities ($SD = 28$). Many of the cities ($n = 171$) were retrieved only once, and only 4 cities were retrieved by all 12 subjects. Figure 1 shows the number of cities with each retrieval rate. To visualize the structure in these highly variable retrieval rates, Figure 2 shows the locations of retrieved cities along with their retrieval frequencies.

Distance Measures

Across all transitions in our dataset, our distance measures tended to be subtly correlated. Replicating the findings from Szary et al., 2015, geographic similarity was positively correlated with temporal proximity across all observed pairwise transitions, as measured by $-IRI$s, $r(787) = 0.16$, $p < 0.001$. In addition, we found that geographic similarity captured semantic similarity, $r(787) = 0.13$, $p < 0.001$. Figure 3 shows the positive correlations between geographic distance measures and both semantic and temporal distance measures. As evident in this graph, semantic and temporal distances are also closely related, with $r(787) = 0.17$, $p < 0.001$. While temporal distance has slight positive correlations to both lexical and phonological distance, $r(787) = 0.12$, $p < 0.001$ and $r(787) = 0.08$, $p = 0.02$, respectively, neither lexical nor phonological distances are correlated with geographic distance. Finally, although lexical and phonological distances are correlated, $r(787) = 0.21$, $p < 0.001$, they have different relationships to the semantic distance measure. Semantic and phonological distances are positively correlated, $r(787) = 0.18$, $p < 0.001$, but semantic and lexical distances actually show a slight negative correlation, $r(787) = -0.09$, $p < 0.01$.

Observed Versus Randomized Distances To make sure that the distance measures differ from those that would be observed by chance, we shuffled each subject's retrieval sequence 100 times, and compared this to simulated data. For each metric, distances were greater for the shuffled data as compared to the original data, as seen in Figure 4. Geographic distances were greater in the shuffled data ($M = 172.7$ miles) than the observed data ($M = 115.0$ miles), $t(813) = -12.76$, $p < 0.001$. Semantic distances were also greater in the shuffled data ($M = 0.71$) than the observed data ($M = 0.63$), $t(798) = -12.9$, $p < -.001$. Lexical distances were greater in the shuffled data ($M = 8.15$) as compared to the observed data ($M = 8.39$), $t(836) = -2.81$, $p < 0.01$. Finally, phonetic distances were greater in the shuffled data ($M = 3.26$) as compared to the observed data ($M = 3.35$), $t(804) = -2.94$, $p < 0.01$.

Patch Transitions

Kernel density estimation was used to determine a segmentation threshold (at the first local minimum) for each participant's IRI sequence ($M = 28$, $SD = 10$ seconds). This thresh-

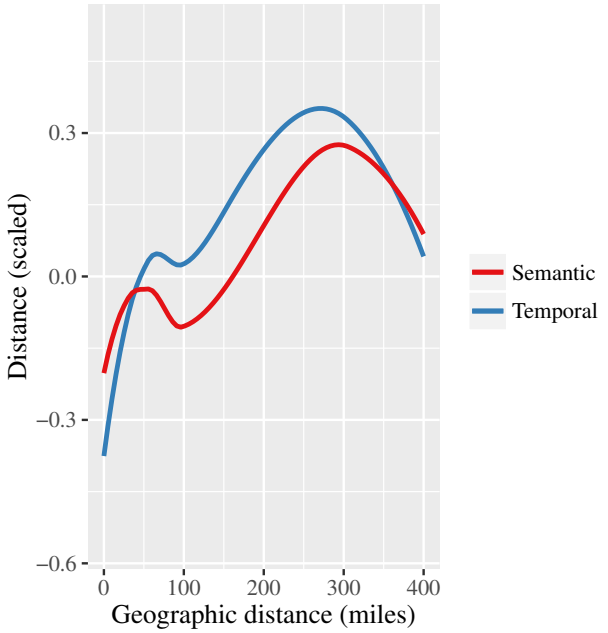


Figure 3: Semantic and temporal distances as a function of geographic distance. For visibility (but not in the analyses), semantic and temporal ranges are scaled and normalized, and geographic extremes are removed. A loess smoothing curve is fit to remaining datapoints to show trends.

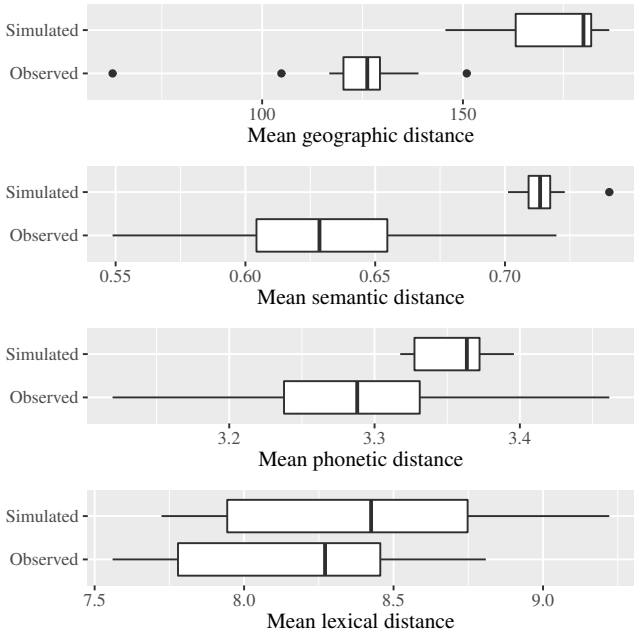


Figure 4: Boxplots showing the differences in mean distances between the simulated and observed datasets.

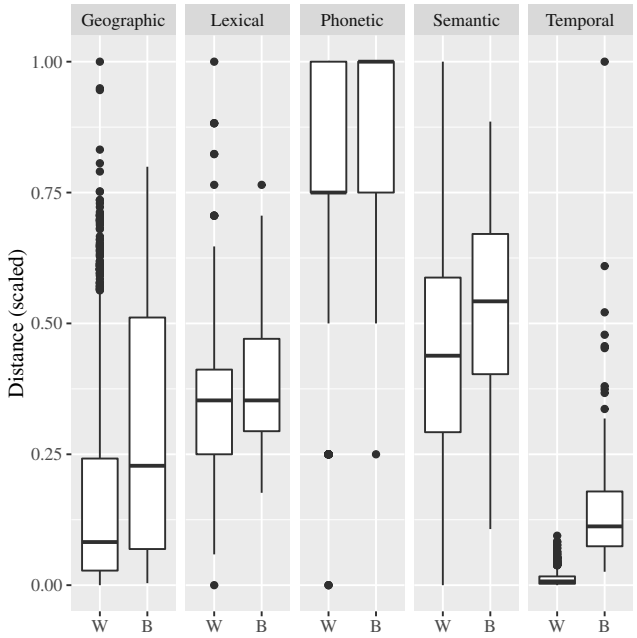


Figure 5: Boxplot of group means comparing pairwise transitions that occurred within patches ("W") to pairwise transitions that occurred between patches ("B"). Distance measures are scaled to share a common axis for illustration purposes only.

old was used to segment retrieved items into patches. Our datasets had an average of 11.4 ($SD = 4.7$) patches. While most patches were small (median size = 2), some were quite large (maximum size = 51; $M = 5.8$, $SD = 9.2$).

Figure 5 compares pairwise transitions that occurred within patches to the pairwise transitions that occurred at patch boundaries (switches). Mean semantic and geographic distances were both significantly higher between patches (at the transition points) as compared to within patches. For semantic distances, $M_{between} = 0.71$ and $M_{within} = 0.62$, $t(211) = 5.80$, $p < 0.001$; For geographic distances $M_{between} = 162.2$ miles and $M_{within} = 105.5$ miles, $t(181) = 4.56$, $p < 0.001$. Lexical distance was slightly but significantly higher between patches ($M = 8.68$) as compared to within patches ($M = 8.05$), $t(194) = 2.94$, $p < 0.005$. Finally, phonetic distance was also slightly higher between patches ($M = 3.44$) as compared to within patches ($M = 3.23$), $t(235) = 3.10$, $p < 0.005$.

Figure 6 shows this same data, but instead of averaging across all within-patch distances, considers the transitions immediately preceding patch switches. Specifically, it shows averages for the switches (index = 0) and the 9 transitions preceding switches. While patch switches have higher average distances on each metric, Figure 6 suggests that there may be patterns leading up to patch switches, perhaps reflecting the depletion of resources available through a given metric.

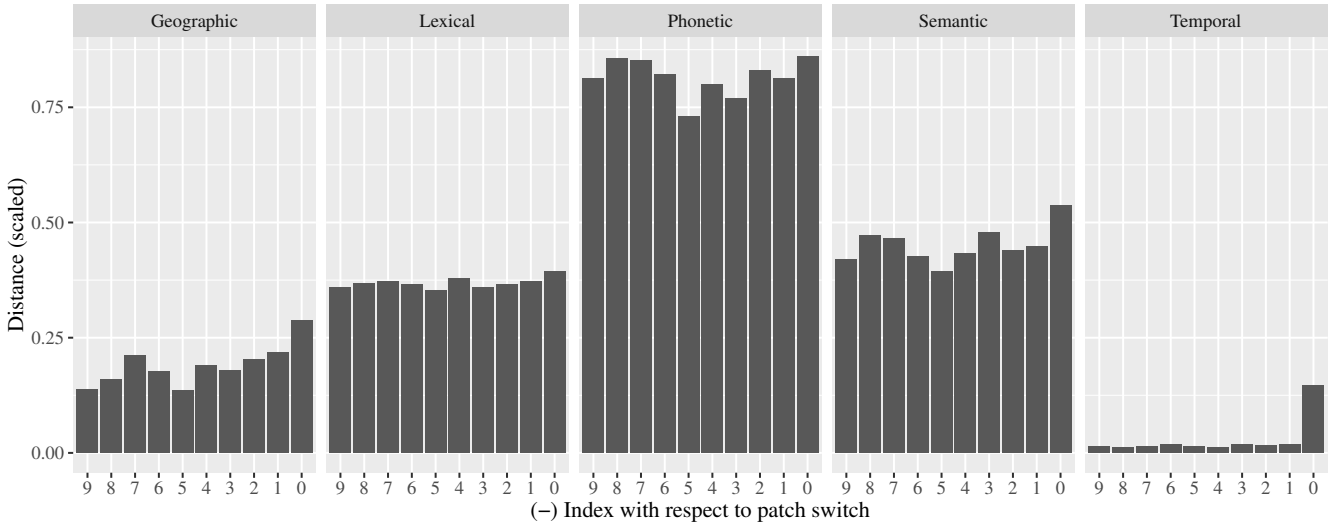


Figure 6: Bar heights show scaled averages for each distance metric at the 9 transitions preceding a patch switch, and the switch itself (index = 0).

Shifting Strategies

Although we focus primarily on summary statistics in this paper, our larger goal is to explore how different cues (geographic, semantic, lexical, phonetic) interact dynamically during retrieval processes. Although a detailed quantitative analysis is beyond the scope of the present paper, Figure shows the evolution of all 4 distance metrics over the course of the trial for each subject, in service of our larger goal.

While there are some similarities across participants (such as geographic distance being relatively lower than other measures), there are also striking differences in the way the measures unfold over time for the different participants. This observation is consistent with the idea that distances may reflect the cueing strategies underlying the search process, where geographic cues would lead a subject to retrieve geographically proximal items, resulting in low geographic distances. Consider, as a simple example, the phonetic and semantic distances for Subject 12. Over the course of the task, semantic distances show a relative increase, while phonetic distances show a relative decrease. From the perspective that our distance metrics (in some loose way) reflect the cueing strategies at work, we see that Subject 12 initially uses a semantic cueing strategy, thinking of items that are semantically related to one another (perhaps listing affluent cities with colleges: "La Jolla... Berkeley... Stanford"). But, as semantically related resources are depleted, the subject switches to a phonetic strategy (perhaps "Stanford... Stanton... Winton").

General Discussion

Overall, most of our distance metrics were positively correlated with each other. Temporal distances were correlated with geographic, semantic, lexical, and phonetic distances. Additionally, geographic distances were correlated with semantic distances, but not with lexical or phonetic distances. Semantic distances were correlated positively with phonetic distances, but negatively with lexical distances. Still, lexical and phonetic distances were positively correlated.

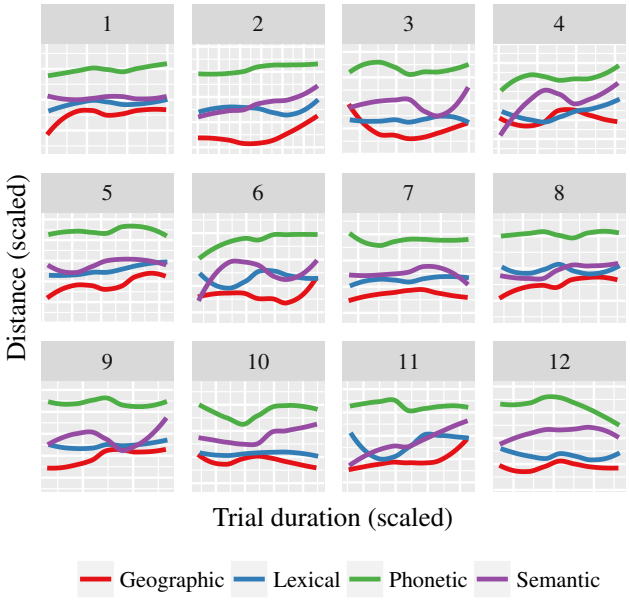


Figure 7: Loess curve fits to each metric (scaled) for each subject.

To test whether there is geographic, semantic, phonetic, or lexical structure in the sequence of retrieved items beyond what would be observed by chance, we compared distances from the observed sequences to distances from sequences of the same items in shuffled orders, as in Szary et al. (2015).

For each of our distance measures, simulated datasets with randomized orders showed greater distances. That is, there is structure in the sequence of retrieved items beyond what would be expected for a random sampling of items, or if a given distance metric were unrelated. This suggests that subjects use each source of information to cue their retrieval processes. In the actual datasets, but not shuffled datasets, pairwise transitions include items that are closely related (on one or more metric).

We also find that each distance metric spikes at the transitions between patches, and tends to increase in the transitions leading up to patch transitions (seeFigure 6). While a quantitative analysis is beyond the scope of the current paper, this finding is consistent with the notion that semantic search utilizes an optimal foraging policy. A hypothesis consistent with optimal foraging would predict that, as a forager retrieves the resources within a given semantic patch, the availability of additional resources becomes depleted, which may be reflected by increases in the the amount of time to find additional items in the transitions preceding patch switches. Here, we use an automatic clustering algorithm applied to the temporal sequence to define patches. While temporal distances are used to define the patch switches, it is difficult to draw conclusions about how they may effect those switches. Visual inspection of the other metrics, however, is consistent with the idea that patches are becoming increasingly sparse (across multiple dimensions) in the time leading up to the decision to switch.

For the argument that search is a generalized cognitive process which relies on the appropriate balance between exploration and exploitation, the word *appropriate* is key. What makes the balance appropriate is known to change depending on search context, such as whether you're searching for randomly distributed or highly clustered resources, but it can also change dynamically over the course of a search process, such as when the resources in a given patch have been depleted. An open question, then, is how the different dimensions of relatedness (geographic, semantic...) interact in what we think of as 'patches'. In order to accurately characterize the line between exploration and exploitation in cognitive space, future work must address how these information sources combine in patches. For example, how does a forager incorporate multidimensional information? Are the patches themselves multidimensional? Or do we move, orthogonally, through cue-strategies and information patches, respectively? As increasing evidence suggests that search strategies are a domain-general process, across a number of cognitive contexts (Hills, Todd, & Goldstone, 2010; Rhodes & Turvey, 2007), a complete understanding of how we leverage different cues to sift through the information in semantic memory becomes increasingly important.

Acknowledgments

Thank you to Fatemeh Torabi Asr and Rick Dale for assistance with data collection and analysis.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558-569.

Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., & Deckmyn, A. (2016). *maps: Draw geographical maps* (R package version 3.1.0 ed.). Retrieved 2016, from https://CRAN.R-project.org/package=maps

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129-136.

dumps.wikimedia.org/enwiki. (2016). Retrieved from https://dumps.wikimedia.org/enwiki/latest/

Hills, T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*, 3-41.

Hills, T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431-440.

Hills, T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, *19*(8), 802-808.

Hills, T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Group, T. C. S. R. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46-54.

Louwerse, M. M., & Zwaan, R. A. (2009). Language encodes geographical information. *Cognitive Science*, *33*, 51-73.

Montez, P., Thompson, G., & Kello, C. T. (2015). The role of semantic clustering in optimal memory foraging. *Cognitive Science*, *39*(8), 1925-1939.

Recchia, G., & Louwerse, M. (2014). Grounding the ungrounded: Estimating locations of unknown place names from linguistic associations and grounded representations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (p. 1270-1275). Austin, TX: Cognitive Science Society.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA. (http://is.muni.cz/publication/884893/en)

Rhodes, T., & Turvey, M. T. (2007). Human memory retrieval as lévy foraging. *Physica A: Satistical Mechanics and its Applications*, *385*(1), 255-260.

Sinnott, R. W. (1984). Virtues of the haversine. *Sky and Telescope*, *68*(2), 159.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41-78.

Szary, J., Kello, C., & Dale, R. (2015). Memory foraging in a spatial domain. In *Proceedings of the 37th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189-208.

United states board on geographic names. (2016). Retrieved from geonames.usgs.gov

Wikimedia tool labs geohack. (2016). Retrieved from https://www.mediawiki.org/wiki/GeoHack

Sentire Decision-Making in a Mixed-Motive Game

Joana Campos (joana.campos@gaips.inesc-id.pt)

Ana Paiva (ana.paiva@inesc-id.pt)
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa

Abstract

The complexity of situations makes individuals use emotions to make sense of their environment and interdependent others. In this paper, we build on the idea that physiological reactions give emotional information about the subject and we focus on Electrodermal Activity (EDA), an index of arousal, to inspect *deep* processes of a dyadic interaction in a mixed-motive game. Our interest lies on how conflict episodes unfold, to design intelligent agents that are more socially aware and thus able to express and recognise dyadic forms of conflict. A qualitative analysis of the data allowed us to identify moments where players made choices to cope with ongoing conflict or prospects of it in the future.

Keywords: Conflict; Electrodermal Activity; Skin Conductance Responses ; Agent Modelling;

Introduction

Emotions are inherently informative. They are central to guide people's behaviour, to support the interaction with others and they carry information about how one is feeling, her motives, preferences, motivations and goals. When interacting with others, emotional expressions help the interlocutor to decode one's internal state, feelings, worries and satisfaction. Emotions (or instead the *appraisal* of the environment) are also transmitted through bodily changes, such as action tendencies and physiological responses (Lazarus, 1994).

In the work described in this paper, we build on the idea that physiological reactions give information about the subject and we focus on Electrodermal Activity (EDA). EDA results from the interaction of local electrical processes of the skin with the sympathetic nervous system of an individual (Boucsein, 2012). It can, thus, inform us about a myriad of psychophysiological processes. Our interest lies on how conflict episodes unfold, to design intelligent agents that are more socially aware and thus able to express and recognise dyadic forms of conflict. We believe that investigating *deep*[1] details of the interaction, such as emotional signalling, is critical for designing more natural experiences between agents and humans. In the future, systems that are able to respond to small details of the interaction (social signals) will become more effective and efficient in their design purpose (Vinciarelli et al., 2009). Focusing on EDA, it is a signal that might be potentially useful in helping a therapist to evaluate which strategies have a more beneficial impact in an autistic child (Chaspari et al., 2014). Furthermore, jointly with heart rate, in can also give insights about how immediate emotions can inform bidder's behaviour in an auction game (Astor et al., 2013).

[1]By *deep* data we mean, rich data representative of one's perspectives, emotions and motivations. We believe that these elements establish the context for many dyadic interactions, which are no more than a form of relating.

In this paper, we use EDA as a *tracing methodology* to study how emotional reactions to events in a bargaining game can change the dynamics of the negotiation process. We attempt to establish links with phases of conflict in a bargaining scenario, by analysing patterns of skin conductance responses (SCRs). For this purpose, we analyse the EDA of 22 children (aged 10 to 12 years-old) in a mixed-motive game under incomplete information. We created a real-life setting reduced to a mixed-motive game, in which the child's previous experiences and relation with the interactional partner play a relevant role in the interaction. We have found that, although there was not overt manifestations of conflict in the interactions, we could identify moments of intense anxiety (high number of SCRs), due to extreme initial offers or during more intense negotiation moments, when the parties are forcing the same deals without making concessions. We conclude the paper discussing our achievements, limitations and whether EDA is actually useful in this type of scenario.

Related Work

Electrodermal Activity

Electrodermal Activity (EDA) results from the interaction of local electrical processes of the skin with the sympathetic nervous system of an individual (Boucsein, 2012). It was verified that the skin becomes a better conductor in response to external stimuli (physical and/or emotional) having influence in myriad of psychophysiological processes. Over the years, EDA has been studied in several domains ranging from attention/engagement, information processing and emotion to clinical research (Dawson, Schell, & Filion, 2007). Researchers found that EDA is an adaptive mechanism to mobilise a 'flight-or-fight' reflex, and is related to anticipation, anxiety and distress (Andreassi, 2000). In the overall, is linked to cognitive and emotional processes.

The EDA complex[2] includes both a *tonic level* (SCL [3]) and a *phasic response* (SCR [4]). The former is the tonic level of conductivity of the skin, a slowly habituating measure of arousal. It slowly decreases when the subject is at rest and rapidly increases when a novel stimulus is introduced, and then gradually decreases (Dawson et al., 2007). The latter is

[2]Contemporary research has focused on measuring the skin resistance response which relies on external current for its observation (*exosomatic method*). Its reciprocal – the *skin conductance response* – has been the unit of analysis used widely by researchers. Within other reasons, *skin conductance* is easier to measure and increases with arousal hence, it provides a more intuitive mechanism for data interpretation.

[3]Skin Conductance Level

[4]Skin Conductance Response

a small fraction of the tonic level and a rapid changing component (often referred to as peak in the signal), which is sensitive to stimulus novelty, intensity and significance (Dawson et al., 2007). Stimuli that elicit a SCR are referred to as "event-specific responses", and on the opposite, if a SCR occurs in the absence of an identifiable stimulus, it is called a "non-specific response" (NS-SCR).

EDA and Decision Making

This interest in affect and emotion in judgement and decision making (JDM) research has renewed the interest in skin conductance and its potential to support research on how emotion impacts decisions in general (Figner & Murphy, 2011). For instance, Wout et al. (2006) studied the role of emotion in the Ultimatum Game. They focused on how skin conductance responses varied when the proposer offered a fair or an unfair division. Furthermore, they also assessed if emotional reactions to the type of offer would be the same if the proposer were a human or a computer. They observed a relationship between human unfair offers and higher skin conductance associated to rejection of those offers. In other words, SCR amplitude measured before response was a better predictor of acceptance or rejection than the offer itself, but the same was not verified when the proposer was a computer.

Another example is the work by Krosch et al. (2012), who explored how people make decisions in morally challenging situations, using self-reported and skin conductance measures. Moral decisions lack a clear right or wrong solution and the authors studied which choices minimise post-decisional consequences[5] for the decision maker. In their experiment, subjects participated in a hypothetical choice scenario, with no interaction with a counterpart. Nevertheless, the results suggest that people reflected increased physiological arousal when facing a conflicting choice, which may be related to post-decisional worry.

Constant and rapid decision making was studied by Palomki et al. (2013) in the context of a poker game variant – No Limit Texas Hold'em (NLHE). This scenario provides a rich context where the participant has to make rapid investment decisions, under risk and uncertainty. The authors use anticipatory arousal of various actions as a predictor of what action is likely to be taken by the decision maker. The authors have related the emotional reactivity to expected and experienced outcomes.

As it is possible to infer from the above research directions, skin conductance is a phenomenon that does not reflect a single psychological process, even within JDM research. Nonetheless, is a mechanism that can provide information about otherwise hidden processes that reflect how people make decisions and form judgements (Figner & Murphy, 2011). In our work, we are specially interested in skin conductance as a *process tracing* methodology to track cognitive states or stages during a decision making task. For that reason,and similarly to what have been done before, we focus on skin conductance responses (SCRs) in anticipation and decision making.

[5]Decision difficulty, post-decisional worry, and negative arousal.

EDA in Mixed-Motive Game: *Game of Nines*

In mixed-motive negotiations, conflicts are bound to emerge because participants have opposing preferences, and each attempts to maximise his or her own gains. An experimental setting in which the potential for conflict exists, acts as a model of a social interaction that is object of study. For this study, we used a variation of the "Game of Nines".

The "Game of Nines"

The "Game of Nines" is a mixed-motive bargaining game and it was firstly used by Kelley (Kelley, Beckman, & Fischer, 1967). This bargaining game was selected because it creates a setting where two negotiators face dilemmas concerning their goals and forms of communication. For a full description of this game refer to (Campos et al., 2015).

Briefly, the game is played with ordinary playing cards and involves three parties: the two players and the house/bank (which is not played by anyone). Each player holds eight cards from one (Ace) to eight, available every round. During each bargaining round, the players had to jointly agree on a possible contract. Each contract corresponded to a card that would be played by player A and one that would be played by player B so that their sum would not exceed 9. In each round, a *minimum necessary share* (MNS) is assigned to each player, privately, by the house[6]. This MNS value is only known by the person to whom it was assigned. for a profitable agreement, the negotiator has to bargain for a value above his MNS (e.g., if a player has a MNS equal to 4 and he plays a 6, he will get 2 as a reward), without knowing the extent of the concessions the other can make. If the participants do not reach an agreement in a limited amount of time both get zero. The player who makes the most profit wins a prize at the end.

What is *conflict* in this setting?

In this created setting, the structure of the task determines the degree of interdependence between the dyads. That is, it establishes the fuel for potential conflict at the beginning of each round. The task structure is determined by the MNS values that are discovered by each participant when each opens a new envelope. The rationale behind using fixed MNS values (see Table 1) throughout the game is to create the same situation each time a new round begins (with the exception of round 4). In our view of conflict (see (Campos, Martinho, & Paiva, 2013; Campos et al., 2015)) at least one party must perceive the incompatibility between the two and attribute the interference in achieving the desired goal to the other party.

Method

Twenty-two children (13 girls and 9 boys, age range 10-12) participated in dyadic sessions of the game. All dyads, with

[6]In our experiment the players took a card from an envelop,which was identified with the number of the correspondent round.

Table 1: Minimum necessary share per round

	Round				
	1	**2**	**3**	**4**	**5**
Player 1	2	1	3	4	2
Player 2	2	3	1	4	2

one exception, were same-sex participants. The mixed-sex dyad was excluded from the analysis described in this paper.

Opt-out consent forms were provided to all of the children's parents or guardians and all games were video and audio recorded, as well as, their skin conductance.

Each pair had to play 5 rounds. In each new round, each participant took a MNS (a numbered card) from an envelope. The child was instructed not to show the card during the trial and to never agree to a value below that number. They were told to negotiate as they saw fit. At the end of the round, each participant had to show the other his or her MNS value. The player with more total points in the end wins the game. The points were not converted to money, but the winner chose a prize between two choices.

Procedure

Before the game sessions, the children completed a sociometric questionnaire and were administered a personality test. The former was applied mainly to ensure that children on opposite poles (neglected vs. popular) or children who did not like each other were not paired together, given the sensitive nature of this experiment[7], to avoid undesirable effects on the participants. The paired children were from the same class, hence they knew each other and shared a history.

For the experiment, each dyad was collected from their classrooms and bracelets to measure their electrodermal activity were immediately attached to their wrists. Then, the children were conducted to a room that had been made available for the purpose. The participants were sat face-to-face at the opposite ends of a table that had a card board to assist them throughout the game. After explaining the rule, the participants were "walked through" two rounds of the game to learn its mechanics (The pre-game sessions took, on average, 15 minutes). Next, they were left alone to play the game.

To motivate the participants to do well, we told the players that the person who accumulated more points during the game would win a prize. In the end, both children won prizes, but the winner was able to choose between two options (one item was better than the other).

Equipment and Recordings

Skin conductance was continuously recorded using the Affectiva Q-Sensor, which has been proven to be highly correlated with data from traditional skin conductance monitoring systems (Poh et al., 2010), with a 8Hz sampling rate.

[7]The results from both questionnaires are beyond of the scope of this paper and are not going to be discussed here.

Sensors were placed in the child's non-dominant wrist right after they were collected from their classroom. After that, they had to walk to the experiment room (2 minutes walk, including climbing stairs). Once they got to the room where the experiment took place, there was a 15 minutes (on average[8]) warm-up period, during which the rules were explained and the participants were "walked through" two rounds of the game to learn its mechanics. Next, children were left alone to play the game and the start of the interaction was "marked" by pressing the sensor's button.

Data was collected during Winter, in January and February.

Detection and quantisation of SCRs

Although measuring EDA is relatively easy nowadays, the analysis of stimulus-related activity remains a challenge, mainly in situations where inter-stimulus intervals (ISIs) are short and peaks may overlap. To address this issue Bach and colleagues (Bach, Flandin, Friston, & Dolan, 2010) and Benedek and Kaernbach (Benedek & Kaernbach, 2010) have proposed two different approaches, based on signal deconvolution, which promise a reliable measure of SCRs. Both works resulted in two different tools for the detection of SCRs and nsSCRs (*SCRalyse*[9] and *Ledalab*[10], respectively). All the results we present in this report were obtained using Ledalab, which we found to be easier to use.

Using *Ledalab*, EDA signals were de-noised with a low-pass Butterworth filter of 1sec length. Afterwards the signal was analysed using continuous decomposition analysis (CDA), which decomposes SC data into continuous tonic and phasic activity (Benedek & Kaernbach, 2010). Furthermore, we used a response window of 1 to 3 sec after each marker, with an amplitude threshold of 0.03 μS to determine the SCR. All nsSCRs were extracted with same minimum amplitude.

Data Analysis

Making decisions consists in evaluate possible options and think about eventual future consequences, take an action and then possibly re-evaluate the decision made. Therefore, both experienced and anticipated emotions influence the decision making process, mainly in quite uncertain situations (Barry & Oliver, 1996). Hence, in this analysis we look for variations in SCRs in three stages (Figure 1): *pre-decision* (anticipation of utility), *decision* and *post-decision* (experienced utility). Pre-decision corresponds to the 3s before the event onset, post-decision which refers to the 3s after the established response window for the event/decision itself. We call to these three stages a *decision block*.

The values presented in this analysis were normalised using a logarithmic transformation (SCRamp = log(1 + SCRamp)). Furthermore individuals that showed no electrodermal activity were excluded from analysis (Dawson, Schell, & Courtney, 2011).

[8]This warm-up period was no less than 10 minutes, minimum required to obtain good results from the Q-sensor.

[9]http://scralyze.sourceforge.net/

[10]http://www.ledalab.de/

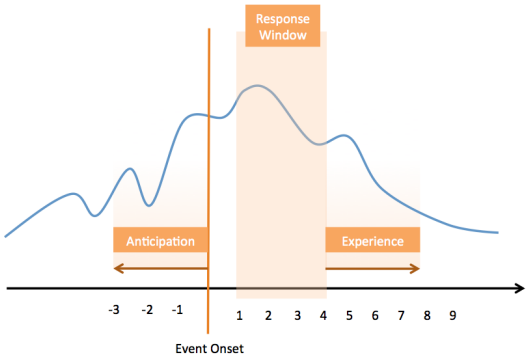


Figure 1: Analysis of the SCRs in three stages: *pre-decision* (anticipation of utility), *decision* and *post-decision* (experienced utility). We call to the set composed by the three stages *decision block*

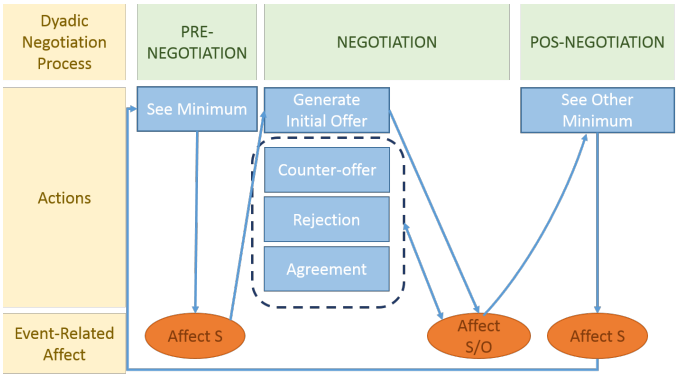


Figure 2: Flow of the interaction in each round of the *Game of Nines* and the affect experienced by the actor of the action (S) and the other (O) due to the game events. The dotted box represents a negotiation cycle.

The analysis of EDA, in particular the skin conductance responses (SCRs), was focused on the actions in the game: *a*) see own minimums; *b*) make an offer; *c*) decline or reject an offer; and *d*) see each other's minimums. Moreover, it is the combination of these actions that may allow us to know more about the emotional dynamics throughout the decisions made within the game. In the next section, we analyse them more closely.

Discussing Results

In our approach, a prototypical negotiation is embedded in the design of the experiment itself. We consider that every round is composed by a *pre-negotiation phase*, a *negotiation* and a *post-negotiation phase* (Figure 2).

In this created setting, children bring their expectations about their partner, their past experiences and the relationship they have built with their partner. These elements, along with the structure of the task, determine the degree of interdependence between them. That is, it establishes the fuel for potential conflict at the beginning of each round. Then, they engage in a negotiation cycle that represents attempts to reconcile conflicting interests and it may be a process marked by intense emotions, which depends on the individual's motivation (Van Kleef & Sinaceur, 2013). Building on the idea that physiological reactions give information about the internal state of the subject, we looked into the EDA of participants, with particular focus on event-related *affect* [11]. A summary of the % of SCRs elicited in each event is in Figure 3.

In ***pre-negotiation***, players see their minimums and formulate their expectations, aspirations and limits, which may vary throughout the interaction. The number of SCRs decrease in *anticipation* of that event, although there is not a significant difference between rounds ($\chi^2(4) = 1.143, p = 0.887$)[12]. This trend may be caused by a decrease in the novelty effect, one of the sources of SCRs (Boucsein, 2012; Dawson et al., 2011). The *experience* of the same event does not vary much over time ($\chi^2(4) = 4.154, p = 0.386$), but we verify a peak in ***Round 4***, which reflects the child's awareness of the few options available for negotiation and might be related to the expectancy of possible negative outcomes (Dawson et al., 2011), We did not find, however, a relationship between the *first offer* and the SCRs elicited by the event *See Minimum* in each round. Independently of the round the *first offers* were high, except in one case:

Case 1 A player demonstrated a feel of guilt in one round after obtaining a very large profit on the previous one. That feel of guilt resulted in cooperative behaviour, that is, she bid low in her first offer (with arousal). This type of behaviour had been identified before by Ketelaar and Tung Au (2003), who experimentally showed that the feeling of guilt fuels cooperative behaviour.

Nevertheless, the increasing number of SCRs exhibited by the *See Minimum* event itself, as the game progresses, may be related to the internal desire of achievement, guided by self-interest only (Barry & Oliver, 1996).

During the ***negotiation phase***, players make offers, counteroffers and generate responses to them. Those actions have effects on the actor of the action (intrapersonal effects), but they certainly have interpersonal effects as well. In Figure 2 we call those effects Affect S/O, i.e. affect experienced by self and by the other. Regarding the affect on the actor of the action (S), the % of SCRs was higher in Rounds 1, 2, and 5 and occurred when players ask for a large slice of the profit. Furthermore, SCRs were elicited over intense moments of the negotiation, that is, slices of the interaction wherein players are making none or very small concessions or even applying

[11] The term affect is used here as a way of describing an affective reactions to a stimulus (event, object or person) without demarcating it as positive or negative (Slovic, Peters, Finucane, & MacGregor, 2005)

[12] Cochran Q test

ROUND	See Min			Bid			Accept/Reject			See Min Other		
	B	D	A	B	D	A	B	D	A	B	D	A
1	40	46.7	13.3	17.6	29.4	0	0	25	7.1	25	56.3	31.3
2	38.5	38.5	7.7	16.7	29.6	3.7	4.3	23.4	4.3	25	37.5	18.8
3	35.7	42.9	28.6	10.8	21.6	8.1	3.1	25	3.1	31.3	50	18.8
4	23.1	23.1	38.5	8.7	18.8	2.9	5.7	24.5	3.8	6.7	46.7	20
5	28.6	50	28.6	12.5	30.4	0	2.6	13.2	2.6	25	41.7	8.3

80-100
50-80
40-50
30-40
20-30
5-20
0-5

Figure 3: Percentage of SCRs elicted by each action by round. **B** - Before/Anticipation; **D** - Event; **A** - After/Experience

avoidance strategies. In ***Round 4*** disputes are over smaller gains that originate some or no profit.

Furthermore, we studied more closely what was a party's reaction to the other's offer since it may be related to the action one takes next. SCRs were elicited in the opponent in two situations: *threat* and *opportunity*. *Threat* corresponds to situations in which the offer generates low or negative profit. The immediate response is to reject in opposition to prompt a counteroffer. *Opportunity* is linked to offers that reflect what the opponent wants to accept, resulting in acceptance of the deal by the opponent. We verify a significant difference in the mean amplitude of the SCRs elicited in the opponent ($F(1) = 4,141, p = 0.046$) when the bid represents a *Threat* ($\overline{amp} = 0.23$) or an *Opportunity*($\overline{amp} = 0.40$)[13].

Accepting and rejecting offers are two moves part of the "*negotiation dance*" and throughout the interaction there are more rejections than acceptances (13 out of 49). Rejection of offers that elicit an SCR are related to high bids made by the proposer that are not the first bid. The next action is usually to replicate the other's offer or wait for the other's proposal. But in particular cases the players used the *give up* card or threatened to use it, this occurred in situations where the opponent did not want to change his offer or was continuously offering bad deals to the player. On the other hand, agreements that elicit a SCR are linked to bad decisions or to concessions that still provide a good profit, but it is not possible to distinguish between the two.

Moving to the ***post-negotiation phase***, this a phase characterised by the revelation of the minimums. The premise is that the behaviour in subsequent rounds will suffer from previous decisions (unfair deals, exploitation or deception). Barry et al. (Barry & Oliver, 1996) alert for situations of asymmetry, wherein a negotiator experiences positive affect and the other negative. Based on previous findings, this may lead to a certain level of "retained hostility" that is exhibited in avoidance or retribution in the following encounters. This state of "retained hostility" was observed twice in our data and are somehow related to the post-negotiation phase:

Case 2 A negotiator simply engaged in avoidance tactics, using irony to make counteroffers. She did not use the card *give up*, but she had threaten to do it a few times.

[13]This data should be taken with a grain of salt though. A spike linked to an opportunity only occurred 7 out of 68 times, although we verified a significant ($p < 0.05$) difference in amplitudes by running ANOVA.

Case 3 A negotiator A got the worse deal in comparison to his opponent B (in rounds 1,2 and 3), except in Round 4 when B played a card below his minimum. In round 5 they engaged again in negotiation with few concessions, player A used the give up card without saying anything before, just to punish the adversary.

These two cases shift our attention to the levels of affect post-settlement and its consequences, when the cause for failure is attributed to the opponent (Weiner, 1985), although they are not directly visible in EDA data. That is, we are able to identify intense moments of the interaction, but more information is needed to make these inferences.

Conclusions

The use of EDA as an index of emotional arousal in judgement research is a common approach and has gained momentum in more recent years (Figner & Murphy, 2011). Among other reasons it a cheap and easy to use method that promises access to hidden processes and information that may reflect how people make decisions, unobtrusively. Research on EDA and decision-making processes has focused essentially in games that involve a risky decisions (Palomki et al., 2013; Slovic et al., 2005; Bechara & Damasio, 2005) to understand people's moves and to study the physiological signals associated with it. With some exceptions (Krosch et al., 2012), there is not much research exploring other types of decision making nor the dynamic process in more natural settings, as the one presented in this report. To approach such complex experimental design we chose to examine the impact of affect in specific moments of the negotiation process: *pre-negotiation*, *negotiation* and *post-negotiation* phase. From that point onwards we continuously broke down the components in smaller pieces. In the *pre-negotiation*, for instance, we analysed the *see minimum* event, which in turn was analysed around a decision block that includes: *pre-decision* (anticipation of utility), *decision* and *post-decision* (experienced utility). This method allowed us to apply between-subjects analysis of each of the actions in the game.

A limitation of our study is that most of the data obtained come from pairs 5, 6, 8 and 10, exhibiting different type of relationships, a pair of best friends and pairs in which the subjects are indifferent to each other. All those were engaged in longer and more disputed negotiations, resulting in the 3 manifestations of conflict (cases 1, 2, 3). Nevertheless, those were

the subjects who kept showing arousal in repeated events suggesting that SCRs are more related to stimulus significance than to stimulus novelty. Furthermore, the decision process was not always fully captured due to the short ISIs between actions, which did not allowed us to have complete *decision blocks*. EDA readings are not useful in such cases (Dawson et al., 2007) and it is debatable if can be used in more natural experimental settings, unless the experimental paradigm is better controlled. Moreover, EDA alone may not help us to automatically make sense of an episode. SCRs only become interpretable by taking into account time and sequence of actions and stimulus conditions, which not makes SCRs alone good predictors of behavioural actions.

All these small details of the interaction, which we call *deep data*, are critical for designing more natural experiences between agents and humans. This study shifted our attention to the role of emotion in the manifestation of conflict and probably the most important finding is the absence of no active behaviours. Dawson et al. (2007) underlines that EDA is linked to the *inhibition* system and thus, involved in processes such as responding to punishment, passive avoidance or to frustrative reward. Hence, it is most responsive when the subject does not take an action. This is an important implication for the design of agent system that deal with conflict: emotions are *retained* and sometimes subjects choose to act upon them and that is visible in cases 2 and 3 described above.

Acknowledgements This work was partially supported by national funds through FCT - Fundação para a Ciência e Tecnologia with reference UID/CEC/50021/2013, a scholarship (SFRH/BD/75342/2010) granted by FCT and was partially supported by the EU FP7 ICT-317923 project EMOTE.

References

Andreassi, J. L. (2000). *Psychophysiology: Human Behavior & Physiological Response*. Psychology Press.

Astor, P. J., Adam, M. T., Jhnig, C., & Seifert, S. (2013). The joy of winning and the frustration of losing: A psychophysiological analysis of emotions in first-price sealed-bid auctions. *Journal of Neuroscience, Psychology, and Economics*, *6*(1), 14.

Bach, D. R., Flandin, G., Friston, K. J., & Dolan, R. J. (2010). Modelling event-related skin conductance responses. *International Journal of Psychophysiology*, *75*(3), 349–356.

Barry, B., & Oliver, R. L. (1996). Affect in dyadic negotiation: A model and propositions. *Organizational Behavior and Human Decision Processes*, *67*(2), 127–143.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, *52*(2), 336–372.

Benedek, M., & Kaernbach, C. (2010, June). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91.

Boucsein, W. (2012). *Electrodermal Activity*. Springer.

Campos, J., Alves-Oliveira, P., & Paiva, A. (2015). Looking for Conflict: Gaze Dynamics in a Dyadic Mixed-Motive Game. *Autonomous Agents and Multi-Agent Systems*.

Campos, J., Martinho, C., & Paiva, A. (2013). Conflict inside out: a theoretical approach to conflict from an agent point of view. In *Proceedings of the 2013 AAMAS*.

Chaspari, T., Goodwin, M., Wilder-Smith, O., Gulsrud, A., Mucchetti, C., Kasari, C., & Narayanan, S. (2014). A non-homogeneous poisson process model of Skin Conductance Responses integrated with observed regulatory behaviors for Autism intervention. In *2014 IEEE ICASSP*.

Dawson, M. E., Schell, A. M., & Courtney, C. G. (2011). The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics*, *4*(2), 111.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The Electrodermal System. In *Handbook of psychophysiology*.

Figner, B., & Murphy, R. O. (2011). Using skin conductance in judgment and decision making research. *A handbook of process tracing methods for decision research*, 163–184.

Kelley, H. H., Beckman, L. L., & Fischer, C. S. (1967, October). Negotiating the division of a reward under incomplete information. *Journal of Experimental Social Psychology*, *3*(4), 361–398.

Ketelaar, T., & Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition & Emotion*, *17*(3), 429–453.

Krosch, A., Figner, B., & Weber, E. U. (2012). Choice processes and their post-decisional consequences in morally conflicting decisions. *Judgment and Decision Making*, *7*(3), 224–234.

Lazarus, R. S. (1994). *Emotion and Adaptation*. Oxford University Press.

Palomki, J., Kosunen, I., Kuikkaniemi, K., Yamabe, T., & Ravaja, N. (2013). Anticipatory electrodermal activity and decision making in a computer poker-game. *Journal of Neuroscience, Psychology, and Economics*, *6*(1), 55.

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on*, *57*(5), 1243–1252.

Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. *Health psychology*, *24*(4S), S35.

Van Kleef, G. A., & Sinaceur, M. (2013). The Demise of the" Rational" Negotiator: Emotional Forces in Conflict and Negotiation. *Handbook of Research on Negotiation*.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009, November). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12).

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, *92*(4), 548.

Wout, M. V., Kahn, R., Sanjey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental brain research*, *169*(4), 564–568.

Modeling the Contribution of Central Versus Peripheral Vision in Scene, Object, and Face Recognition

Panqu Wang (pawang@ucsd.edu)
Department of Electrical and Engineering, University of California San Diego
9500 Gilman Dr 0407, La Jolla, CA 92093 USA

Garrison W. Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

Abstract

It is commonly believed that the central visual field (fovea and parafovea) is important for recognizing objects and faces, and the peripheral region is useful for scene recognition. However, the relative importance of central versus peripheral information for object, scene, and face recognition is unclear. Larson and Loschky (2009) investigated this question in the context of scene processing using experimental conditions where a circular region only reveals the central visual field and blocks peripheral information ("Window"), and in a "Scotoma" condition, where only the peripheral region is available. They measured the scene recognition accuracy as a function of visual angle, and demonstrated that peripheral vision was indeed more useful in recognizing scenes than central vision in terms of achieving maximum recognition accuracy. In this work, we modeled and replicated the result of Larson and Loschky (2009), using deep convolutional neural networks (CNNs). Having fit the data for scenes, we used the model to predict future data for large-scale scene recognition as well as for objects and faces. Our results suggest that the relative order of importance of using central visual field information is face recognition>object recognition>scene recognition, and vice-versa for peripheral information. Furthermore, our results predict that central information is more efficient than peripheral information on a per-pixel basis across all categories, which is consistent with Larson and Loschky's data.

Keywords: face recognition; object recognition; scene recognition; central and peripheral vision; deep neural networks

Introduction

Viewing a real-world scene occupies the entire visual field, but the visual resolution across the visual field varies. The fovea, a small region in the center of the visual field that subtends approximately 1° of visual angle (Polyak, 1941), perceives the highest visual resolution of 20 to 45 cycles/degree (cpd) (Loschky, McConkie, Yang, & Miller, 2005). The parafovea has a slightly lower visual resolution and extends to about 4-5° eccentricity, where the highest density of rods is found (Wandell, 1995). Beyond the parafovea is generally considered to be peripheral vision (Holmes, Cohen, Haith, & Morrison, 1977), which receives the lowest visual resolution. Due to the high density and small receptive field of retinal receptors,the central (foveal and parafoveal) vision encodes information of higher spatial frequency and more detail; peripheral vision, on the contrary, encodes coarser and lower spatial frequency information.

This retinotopic representation of the visual field is mapped to visual cortical areas through a log-polar representation. Recent studies have shown that orderly central and peripheral representations can be found not only in low-level to mid-level visual areas (V1-V4), but also in higher-level regions, where perception and recognition for faces or scenes is engaged (Malach, Levy, & Hasson, 2002; Grill-Spector & Malach, 2004). More specifically, Malach et al. (2002) proposed that the need for visual resolution is a crucial factor in organizing object areas in higher-level visual cortex: object recognition that depends more on fine detail is associated with central-biased representations, such as faces and words; object recognition that depends more on large-scale integration is associated with peripheral-biased representations, such as buildings and scenes. This hypothesis is supported by fMRI evidence, which shows that the brain areas that are more activated for faces (FFA; Kanwisher, McDermott, and Chun (1997)) and words (VWFA; McCandliss, Cohen, and Dehaene (2003)) sit in the eccentricity band expanded by central visual-field bias, whereas buildings and scenes (PPA; Epstein, Harris, Stanley, and Kanwisher (1999)) are associated with peripheral bias. More recent studies even suggest that the central-biased pathway for recognizing faces and peripheral-biased pathway for recognizing scenes are segregated by mid-fusiform sulcus (MFS) to enable fast parallel processing (Gomez et al., 2015).

In the domain of behavioral research, studies have shown that object perception performance is the best around 1°-2° of fixation point and drops rapidly as eccentricity increases (Henderson & Hollingworth, 1999; Nelson & Loftus, 1980). For scene recognition, Larson and Loschky (2009) used a "Window" and "Scotoma" design (see Figure 1), to test the contributions of central versus peripheral vision to scene recognition. The Window condition (top rows of the right-hand columns of Figure 1) presents central information at various visual angles to the subjects, while the Scotoma condition (second row on the right) blocks it. Using images from 10 categories, subjects were required to verify the category in each condition. The recognition accuracy as a function of visual angle is shown in Figure 2. They found that foveal vision is not accurate for scene perception, while peripheral vision is, despite its much lower resolution. However, they also found that central vision is more efficient, in the sense that less area is needed to achieve equal accuracy. The visual area is equal at 10.8°, and the crossover point, where central vision starts to perform better than peripheral, is to the left of that point.

Despite the common belief that central vision is important for face and object recognition, and peripheral vision is important for scene perception shown in studies above, a careful examination of the contribution of central versus peripheral vision in object, scene, and face recognition is needed. In this work, we modeled the experiment of Larson and Loschky (2009) using deep convolutional neural networks. Furthermore, we extended the modeling work to a greater range of stimuli, and answer the following questions: How does the model perform as the number of scene categories is scaled up? Besides scenes, can the model predict the importance of central vision versus peripheral information in object and face recognition? What is the result compared to scenes?

In the following, we show that our modeling results match the observations of Larson and Loschky (2009), and that it scales up to over 200 scene categories. By running a similar analysis for large-scale object and face recognition, our model predicts that central vision is very important for face recognition, important for object recognition, and less important for scene recognition. Peripheral vision, however, serves an important role for scene recognition, but is less important for recognizing objects and faces. Furthermore, across all conditions we tried, central vision is more efficient than peripheral vision on a per-pixel basis (when equal areas are presented), which is consistent with the result of Larson and Loschky (2009).

Method

Image Preprocessing

To create foveated images, we preprocessed the images using the Space Variant Imaging System[1]. To mimic human vision, we set the parameter that specifies the eccentricity at which resolution drops to half of the fovea to 2.3°. Example images and their preprocessed retinal versions are shown in the first and second columns of Figure 1.

As in the experiments of Larson and Loschky (2009), we used the Window and Scotoma paradigms as specified by van Diepen, Wampers, and dYdewalle (1998) to process the input stimulus. The idea of both paradigms is to evaluate the value of missing information - if the missing information is needed, then the perception process may be disrupted and recognition performance may drop; if the missing information is not necessary, then the processing remains normal.

Input images in our experiments are 256×256 pixels, and we assume that corresponds to $27° \times 27°$ of visual angle, the number in (Larson & Loschky, 2009). In (Larson & Loschky, 2009), they used four sets of radius conditions for Windows and Scotomas: 1° represents the presence or absence of foveal vision; 5° represents the presence or absence of central vision; 10.8° presents equal viewable area inside the Windows or outside the Scotomas; 13.6° presents more viewable area in the Windows than the Scotomas. In order make the prediction of the model more accurate, we added five additional radius conditions in all of our experiments: $3°, 7°, 9°, 12°$, and $16°$. The example Window and Scotoma images are shown in Figure 1.

[1] http://svi.cps.utexas.edu/software.shtml

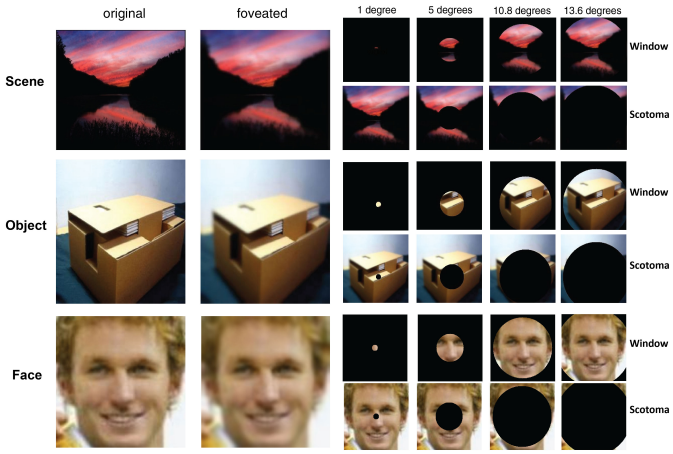


Figure 1: Examples of images used in our experiment. First column: original images. Second column: foveated images. Third to last column: images processed through "Window" and "Scotoma" conditions with different radii in degrees of visual angle.

Deep Convolutional Neural Networks (CNNs)

Deep CNNs are neural networks with many layers that stack computations in a hierarchical way, repeatedly performing: 1) 2-dimensional convolutions over the stimulus generated from previous layers using learned filters, which are connected locally to a small subregion of the visual field; 2) a pooling operation on local regions of the feature maps obtained from convolution operation, which is used to reduce the dimensionality and gain translational invariance; 3) nonlinearities to the upstream response, which is used to generate more discriminative features useful for the task. As layers go higher, the receptive fields of filters are generally larger, and the learned features go from low-level (edges, contours) to high-level object-related representations (object parts and shapes) (Zeiler & Fergus, 2014). Several fully-connected layers are usually added on top of these computations to learn more abstract and task-related features.

We used deep CNNs in our experiments for two reasons. First, deep CNNs are the best models in computer vision: they achieve the state-of-the-art performance on many large-scale computer vision tasks, such as image classification (Krizhevsky, Sutskever, & Hinton, 2012; He, Zhang, Ren, & Sun, 2015), object detection (Ren, He, Girshick, & Sun, 2015), and scene recognition (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014). Thus, the models should achieve decent performance in our experiments. Smaller networks or other algorithms are not competent for our tasks. Second, deep CNNs have been shown to be the best models of the visual cortex: they are able to explain a variety of neural data in human and monkey IT (Yamins et al., 2014; Güçlü & van Gerven, 2015; Wang, Malave, & Cipollini, 2015). As a result,

it is natural to use them in our work modeling a behavioral study related to human vision.

Experiments

In this section, we first describe our model of the behavioral study of Larson and Loschky (2009). We then introduce the experiment for measuring the contribution of central versus peripheral vision for large-scale scene, object, and face recognition tasks.

Modeling Larson and Loschky (2009)

In Larson and Loschky (2009), scene recognition accuracy was measured across 100 human subjects on 10 categories: Beach, Desert, Forest, Mountain, River, Farm, Home, Market, Pool, and Street. For each trial in the Windows and Scotomas conditions, subjects were first presented a scene image, and then were asked to press "yes" or "no" for the cue (category name) presented on the screen. Their experimental result is summarized in Figure 2. They showed that central vision (5° window condition) performs less well than peripheral vision in terms of getting maximum recognition performance. They further demonstrated the peripheral advantage is due to more viewing areas in the Scotomas conditions, and central vision is more privileged when given equal viewable areas (10.8°).

We obtained the stimuli of the above 10 categories from the Places205 database (Zhou et al., 2014), which contains 205 scene categories and 2.5 million images. All input stimuli were preprocessed using the retina model described in the above section. As 10 categories is small and can easily lead to overfitting problems in training deep CNNs, we trained our recognition model by performing fine-tuning (or transfer learning) based on pretrained models. The model pretrained on the Places205 database can be treated as a mature scene recognition pathway, and fine-tuning can be thought as additional training for the task. To investigate whether different network architectures, especially depth, have different impact on the modeling result, we applied three different pre-trained models, namely:

1. AlexNet (Krizhevsky et al., 2012): A network with 5 convolutional layers and 3 fully connected layers, about 60 million trainable parameters. Achieved 81.10% top-5 accuracy on the Places205 validation set.

2. VGG-16 (Simonyan & Zisserman, 2014): A network with 13 convolutional layers and 3 fully connected layers, about 138 million trainable parameters. Achieved 85.41% top-5 accuracy on the Places205 validation set.

3. GoogLeNet (Szegedy et al., 2015): A network with 21 convolutional layers and 1 fully connected layer, about 6.8 million trainable parameters. Achieved 87.70% top-5 accuracy on the Places205 validation set.

For all models, the fine-tuning process starts by keeping the weights except for the last fully connected layer intact, and initializing the weights of the last layer to be random with zero mean and unit variance. To be compatible with the "yes" or "no" condition in the behavioral experiment, we replaced the last layer in the networks with a single logistic unit, and trained the networks for each of the 10 object categories separately, using half of the training images from the target category and half randomly selected from all other 9 categories. As the last layer needs more learning, we set the learning rate of the last layer to 0.001, and all previous layers to $1e^{-4}$. The training set of the 10 scene categories contains a total number of 129,210 full resolution images, and we trained all networks using minibatch stochastic gradient descent with batch size from 32 to 256, using the Caffe deep learning framework (Jia et al., 2014) on NVIDIA Titan Black 6GB GPUs. All networks were trained for a maximum number of 24,000 iterations to ensure convergence. Each test set contains 200 images (100 from target category and 100 from all other categories), and the label distribution is the same as the training set.. Test images were preprocessed to meet each of the Windows and Scotomas condition. We tested the performance of the fine-tuned models on all conditions by reporting the mean classification accuracy, which is shown in Figure 2.

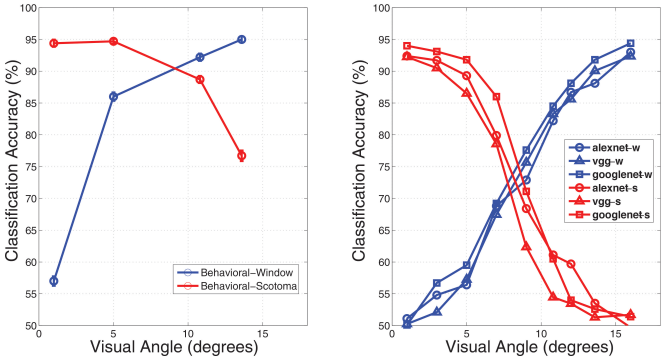


Figure 2: Results for scene recognition accuracy as a function of viewing condition (Windows (w) and Scotomas (s)) and visual angle. Left: result of Larson and Loschky (2009). Right: our modeling result.

From Figure 2, we can clearly see our result for all three models qualitatively matches the result of Larson and Loschky (2009). First, for Window and Scotoma conditions, an increasing radius of visual angle (x axis) yields a monotonic increase or decrease in classification accuracy (y axis). The sharper increase from 1° to 5° in the behavioral study may be due to the higher efficiency of human central vision. Second, we replicated the fact that central vision (less than 5°) is less useful than peripheral vision in terms achieving the best scene recognition performance. Third, however, when using equal viewable areas (10.8°), central vision performs better than peripheral, exhibiting higher efficiency. Fourth, the critical radius (the crossover point where the two conditions produce equal performance, see Figure 2b) is 8.26° (averaged across all models), which is within the 8.22°-9.24° range reported by Larson and Loschky (2009). This suggests our models are quite plausible.

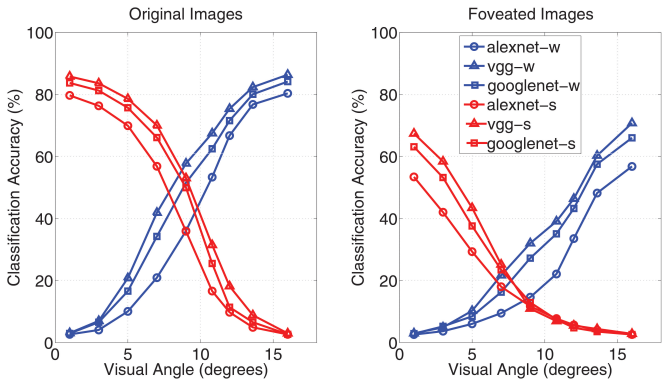


Figure 3: Results for large-scale scene recognition accuracy as a function of viewing condition (Windows (w) and Scotomas (s)) and visual angle. Softmax output is used instead of logistic unit, so chance is 0.005. Left: experiment using original images. Right: experiment using foveated images.

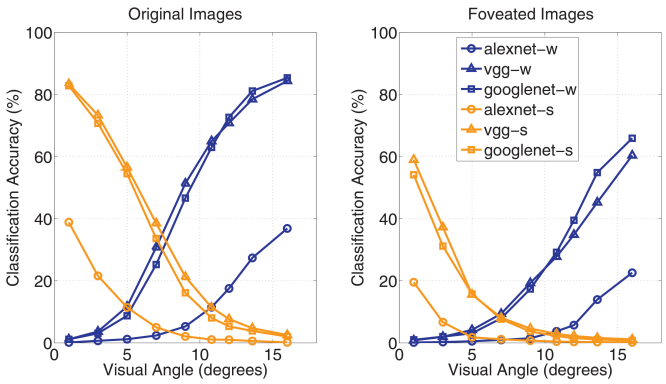


Figure 4: Results for large-scale object recognition accuracy as a function of viewing condition (Windows (w) and Scotomas (s)) and visual angle. Softmax output is used instead of logistic unit, so chance is 0.001. Left: experiment with original images. Right: experiment with foveated images.

When comparing the performance across the three models we use, we cannot find a notable difference in terms of performance, though GoogLeNet usually performs slightly better, indicating that depth of processing might be the key factor in obtaining better performance.

Large-Scale Scene, Object, and Face Recognition

The above modeling work is based on a scene recognition task using 10 categories. In real life, however, there are a much larger number of scene categories. Beyond scenes, general object recognition and face recognition are the two most important recognition tasks that are performed regularly. The relative importance of central versus peripheral vision among the three categories needs to be examined carefully. Using a similar modeling approach, we describe our findings in large-scale scene, object, and face recognition in the sections below.

Scene Recognition We used all 205 categories in the Places205 dataset. The trained models of AlexNet, VGG-16, and GoogLeNet are deployed to examine the recognition accuracy on the Place205 validation set, which contains 20,500 images, in all Windows and Scotoma conditions. In addition, we tested the models using images both processed and unprocessed by the retina model to examine the generalization power of the learned features. The result is shown in Figure 3.

From Figure 3, we can see the general trend that we observed in Figure 2 still holds: peripheral vision is more important than central vision, but central vision is more efficient. All models behave similarly. However, we can see the performance on images preprocessed through the retina model is inferior. Apparently, since there are many more categories in this experiment, the foveation has more of an effect. Recall that the models are trained using images with full resolution; missing the peripheral information may the cause learned features to imperfectly generalize.

Object Recognition We ran our object recognition experiment on the ILSVRC 2012 dataset (Russakovsky et al., 2015), which contains 1000 object categories and over 1.2 million training images. We used the pretrained models of AlexNet, VGG-16, and GoogLeNet, which achieve top-5 accuracy of 80.13%, 88.44%, and 89.00%, respectively, on the ILSVRC 2012 validation set. Similar to scene recognition, we tested all models under all Windows and Scotoma conditions, using original and foveated images. The results are shown in Figure 4.

At the first glance of looking at Figure 4, we may draw the conclusion that the result is the same as scene recognition: central vision is still more important than peripheral vision. However, when we compare the scene and object recognition results (shown in Figure 5), we can clearly see that central information in object recognition is more important than that in scene recognition: the accuracy of the Scotoma conditions drops much faster for object recognition than scene recognition as visual angle increases from 1° to 7°, suggesting that losing central vision causes a greater impairment for object recognition performance. This is consistent with our knowledge that central vision plays a more important role in object recognition than scenes, as there are more high spatial frequency details in objects than scenes. Another finding from this experiment is that AlexNet (8 layers) performs much worse than VGG-16 (16 layers) and GoogLeNet (23 layers), suggesting that depth is important to produce good performance.

Face Recognition We performed the face recognition experiment on the Labeled Faces in the Wild (LFW) dataset (Huang, Ramesh, Berg, & Learned-Miller, 2007), which contains 13,233 labeled images from 5,749 individuals. As there is only 1 image for some identities, researchers usually pretrain their network on larger datasets (not publicly available) and test their models on the LFW dataset. In this experiment, we tested three pretrained models, namely Lighten-A (10 layers; (Wu, He, & Sun, 2015)), Lighten-B (16 layers), and VGG-Face (16 layers;(Parkhi, Vedaldi, & Zisserman, 2015)),

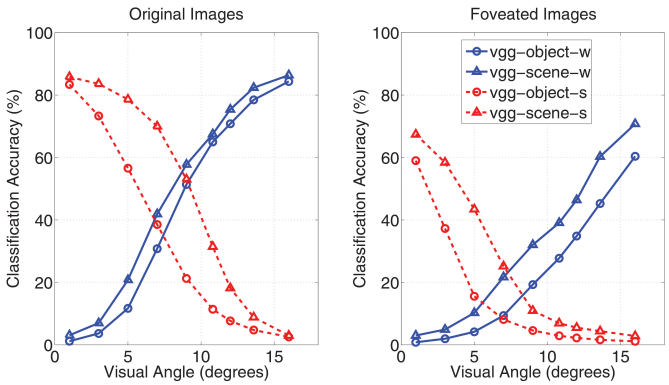


Figure 5: Comparison results for scene and object recognition using the VGG-16 model. Losing central vision decreases performance for object recognition more quickly than scene recognition. Left: original images. Right: foveated images.

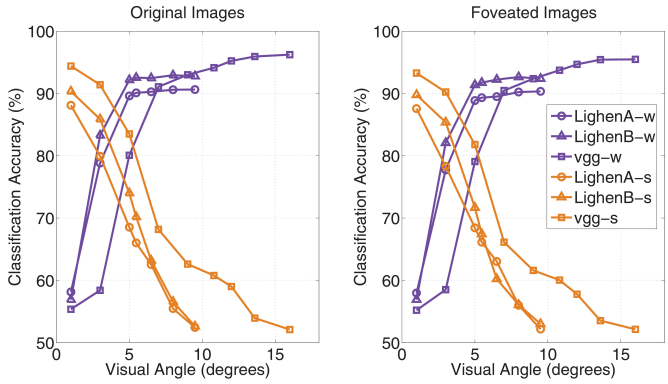


Figure 6: Results for large-scale face recognition accuracy as a function of viewing condition (Windows (w) and Scotomas (s)) and visual angle. Left: experiment with original images. Right: experiment with foveated images. For Lighten-A and Lighten-B models, the visual angle only expands to 9.5°, as the input image is smaller (144×144) than for the VGG model (256×256). The accuracy for face verification task is measured as the true positive rate at Equal Error Rate (EER) point on the ROC curve. Chance is 0.5.

on the face verification task for the LFW dataset, where they achieve accuracy of 90.33%, 92.37%, and 96.23%, respectively. Face images were preprocessed so that they occupy the entire visual field (Figure 1). Same as the previous experiments, we tested all models using Windows and Scotoma conditions, with original and foveated images. Results are shown in Figure 6.

We see very different performance in Figure 6 compared to object and scene recognition. First, central information is obviously much more important than peripheral information for face recognition, given the accuracy at 5° is much higher for the Window condition than the Scotoma condition for Lighten models, and very similar with each other for the VGG model. This is consistent with our intuition that face recognition is a fine-grained discrimination process. Second, the Window performance grows much more slowly after 7°, suggesting the more peripheral region provides little additional information for recognizing faces, unlike objects and scenes, which needs lots of peripheral information to obtain the maximal accuracy. Third, the foveated images produce nearly identical results as the original image, demonstrating that face recognition only involves central vision, and the blurred peripheral vision is not needed.

Finally, as central vision appears to be more efficient (on a per-pixel basis) than peripheral vision in all experiments we tried, we tested the relative efficiency of the central vision over peripheral vision by measuring the recognition accuracy as a function of viewable area. The result is shown in Figure 7.

From Figure 7, we can clearly see that the recognition accuracy of central vision is always superior than peripheral vision for all tasks. However, central vision is even more efficient when recognizing faces than recognizing objects or scenes, as viewable areas over 50% of the whole image can only provide a limited boost for face recognition, while significantly improving the accuracy of object and scene recognition. Contrarywise, peripheral information provides little to no help for face recognition, unless over 90% of the image is presented, but the accuracy still suffers due to the loss of central vision. However, peripheral information is important for object and scene recognition (and more important for scene recognition, as shown in Figure 5).

These large-scale scene, object, and face recognition modeling results suggest there is an order of relative importance of central versus peripheral vision in those tasks: peripheral vision is most important for scene recognition, less important for object recognition, and basically not helpful for face recognition. Central vision, however, plays a crucial role in face recognition, is important for object recognition, and is less important for scene recognition.

Conclusion

In this paper, we modeled the contribution of central versus peripheral visual information for scene, object, and face recognition, using deep CNNs. We first modeled the behavioral study of Larson and Loschky (2009), and replicated their findings of the importance of peripheral vision in scene recognition. In addition, by running a large-scale scene, object, and face recognition simulation, our models make testable predictions for the relative order of importance for central versus peripheral vision for those tasks.

Acknowledgments

This work was supported by NSF grants IIS-1219252 and SMA 1041755 to GWC. PW was supported by a fellowship from Hewlett-Packard.

References

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, *23*(1), 115–125.

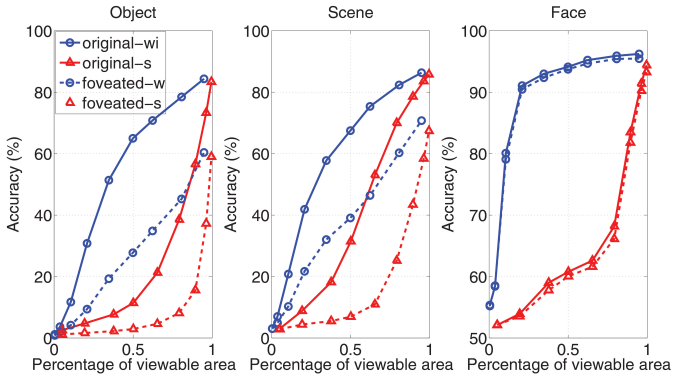


Figure 7: Accuracy for object (left), scene (middle) and face (right) recognition as a function of the percentage of viewable area presented under Window (blue) and Scotoma (red) conditions, using original (solid line) and foveated images (dashed line).

Gomez, J., Pestilli, F., Witthoft, N., Golarai, G., Liberman, A., Poltoratski, S., ... Grill-Spector, K. (2015). Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. *Neuron*, *85*(1), 216–227.

Grill-Spector, K., & Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.*, *27*, 649–677.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005–10014.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, *10*(5), 438–443.

Holmes, D. L., Cohen, K. M., Haith, M. M., & Morrison, F. J. (1977). Peripheral visual processing. *Perception & Psychophysics*, *22*(6), 571–577.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). *Labeled faces in the wild: A database for studying face recognition in unconstrained environments* (Tech. Rep. No. 07-49). University of Massachusetts, Amherst.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302–4311.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Larson, A. M., & Loschky, L. C. (2009). The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, *9*(10), 6.

Loschky, L., McConkie, G., Yang, J., & Miller, M. (2005). The limits of visual resolution in natural scene viewing. *Visual Cognition*, *12*(6), 1057–1092.

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas. *Trends in cognitive sciences*, *6*(4), 176–184.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, *7*(7), 293–299.

Nelson, W. W., & Loftus, G. R. (1980). The functional visual field during picture viewing. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(4), 391.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British machine vision conference.*

Polyak, S. L. (1941). The retina.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015, April). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 1-42. doi: 10.1007/s11263-015-0816-y

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015, June). Going deeper with convolutions..

van Diepen, P. M., Wampers, M., & dYdewalle, G. (1998). Functional division of the visual field: Moving masks and moving windows. *Eye guidance in reading and scene perception*, 337–355.

Wandell, B. A. (1995). *Foundations of vision.* Sinauer Associates.

Wang, P., Malave, V., & Cipollini, B. (2015). Encoding voxels with deep learning. *The Journal of Neuroscience*, *35*(48), 15769–15771.

Wu, X., He, R., & Sun, Z. (2015). A lightened cnn for deep face representation. *arXiv preprint arXiv:1511.02683*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–eccv 2014* (pp. 818–833). Springer.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487–495).

The Charon Model of Moral Judgment

Deirdre Kelly (deirdrekrkelly@gmail.com)

Jim Davies (jim@jimdavies.org)

Institute of Cognitive Science, Carleton University
Ottawa, Ontario, K1S 5B6 Canada

Abstract

We present a model of moral judgment, Charon, which adds to previous models several factors that have been shown to influence moral judgment: 1) a more sophisticated account of prior mental state, 2) imagination, 3) empathy, 4) the feedback process between emotion and reason, 5) self-interest, and 6) self-control. We discuss previous classes of models and demonstrate Charon's extended explanatory power with a focus on psychopathy and autism.

Keywords: morals, morality, ethics, ethical reasoning, moral reasoning, moral judgment, philosophy, deontological, utilitarian, dual-process, modeling, psychopathy, autism, reasoning, emotion, emotions, self-interest, empathy, compassion, willpower, self-control, imagination.

Introduction

Traditional psychological models of moral judgment mainly focus on the types of reasons that people have to do the right thing (e.g., Rest, 1986; Gaudine & Thorne, 2001; Thorne & Saunders, 2002). More recently Haidt (2001) introduced a social-intuitionist model of moral judgment in which judgments are emotional, quick, and intuitive. The role of reason in Haidt's account is entirely post-hoc and does not play a role in generating judgments.

The role of emotion in moral judgment has also been investigated in dual-process models. Greene et al. (2001) distinguish between two types of moral processes – deontological and utilitarian. They argue that the former are an intuitive, emotional process and that the latter is reason-based.

This paper examines the state of psychological moral judgment modelling. We present a new model that better accommodates a breadth of empirical moral judgment evidence and, further, can explain the differences in moral judgment between abnormal populations, such as psychopaths and people with autism.

Reason-based Models

Reason-based models, such as Rest's (1986), involve recognition of moral issues, and then application of moral reasons, rules, or frameworks (such as utilitarianism) to arrive at a moral judgement. Others have added to this basic model to account for the influence of situational factors (e.g., Trevino, 1986). Models of moral judgment have similarly almost exclusively focused on reason-based decision making for the past thirty years.

One major problem with reason-based models is that they have trouble accounting for moral judgements that seem to have nothing to with good moral reasoning, such as when somebody thinks it's immoral to burn a flag. Much of moral judgment happens quickly, with no deliberation, suggesting the importance of emotion in moral judgement.

According to Kohlberg (1973), as people become more proficient at reasoning and have a more fully developed concept of themselves as connected to the largest sphere of relations, they will develop an abstract, universalized form of morality. At its highest stages, these moralities take the form of utilitarianism, and better still Kantianism. Kohlberg's focus was on reasoning alone. The way people are moral is by becoming ever more proficient reasoners. As people improve their understanding of the world, they are better able to make moral decisions.

Haidt's Social Intuitionist model

Haidt (2001) introduced a social intuitionist model, later called Moral Foundations Theory (2012), that holds that moral judgements are based on application of several moral foundations (five in earlier versions, six in its 2012 version) that work through emotion. For Haidt, the role of moral reasoning is post hoc. Once we have already made a snap judgment, we use moral reasoning to justify or confabulate the decision we have already made.

Haidt's empirical support for the post-hoc role of reasoning in moral judgment comes from his moral dumbfounding experiments, in which he presented subjects with various moral scenarios and found that most would have an emotional judgment of things and hold that judgment even after their reasons failed to support it (Haidt, 2001; Haidt et al., 2000).

Research has highlighted the important role that disgust plays in moral judgment (Haidt, 2012; Pizarro et al., 2011). Haidt (2012, 2001) argues that disgust reactions, such as that which is felt when faced with cases of incest, or the idea of your neighbour eating his dead dog, are in themselves moral judgments of the wrongness of the acts. Pizarro takes this a step further in his research. He and his colleagues argue that even the presence of disgust is enough for someone to frame the situation they later view as being moral. In other words, disgust can trigger a moral mindset.

We argue that Moral Foundations Theory underestimates the use of reason in moral judgement. Examination of

Haidt's results indicates that depending on the scenario people sometime change their minds from their original intuition based on their their post-hoc reasoning. For example, when prompted to consider various ways in which the incest case is not actually "morally wrong"; i.e. no one is getting hurt, there will not be an infant conceived, etc., people will sometimes change their minds and arrive at a new moral judgment. 10-23% of people changed their judgments when presented with counter-evidence to their original intuitions (Haidt 2001; Haidt et al., 2000). Though not a majority, this is a sizable percentage and a model of moral judgement must accommodate it.

Dual-Process Models

Dual-process models of moral judgment are a relatively new area of investigation across disciplines. As a result, there are few existent models that take this approach. While it has been gaining traction in psychology, neuroscience, and other more applied disciplines, very little to date has been done from a philosophical standpoint.

Dual-process models of moral decision-making claim that emotion/intuition and reasoning are both integral to human morality.[1] Joshua Greene and his colleagues in the early 2000s were the first team to apply this framework, and other dual-process accounts resemble theirs.

Greene and colleagues used Philippa Foot's (1978) trolley/footbridge cases to examine how people make moral judgments. Trolley cases involve a person having to make a decision to act (such as by pulling a switch) where their choice to act will save a greater number of lives. For example, the classic trolley case involves a train coming down the tracks, if you choose to do nothing then the train will hit and kill 5 people, but if you pull a switch then the train will be diverted and will only kill 1 person. Greene et al. (2001) found that the large majority of people are willing to pull the switch in the trolley case. They argue that this is because when choosing whether or not to pull the switch people do some kind of utilitarian calculus. In this study, people will often explain their actions in the following way: "It's better to save five people, and to let one person die, because five lives saved is better than one." This reflects a utilitarian type of reasoning. Greene et al. were interested in explaining why it is that despite willingness to pull the switch in the trolley case, people will not act to save the five in the footbridge case where it is required that you physically push another person onto the tracks to save the five people's lives.

To explain this finding, Greene et al. invoke a dual-process model of moral judgment. They claim that people are more averse to doing things that involve personal harm and that footbridge cases involve an entirely different type of process than in the trolley case. They explain that when faced with footbridge cases, people invoke (what Greene considers) a deontological approach to morality. Greene and colleagues explain that deontological approaches are emotional and that the idea of personally harming another harming another person causes an emotional interference that impedes doing a utilitarian calculus.

Greene sees complex moral decisions as being the result of an internal struggle between reason-based and emotion-based approaches happening in the brain.

Messervey, Nelson, and Peach (2016) and Messervey (2013) also present a dual-process model equipped for the fact that depending on situation and context people can appeal to two different processes for moral judgment. Their focus is on stress' influence on ethical judgment. They explain that in situations where you are under time constraints, are uncertain, or stressed, deliberative processing is difficult. This predicts that when a person is stressed they are more likely to rely on intuitive moral judgments.

Criticism of Previous Models

We have presented three broad classes of models that attempt to account for the range of moral judgments' influences. However there are empirical findings that have not yet been accommodated by any existing model.

First, they do not account for the way in which imagination affects moral decision-making. Visual imagery has been shown to have an effect on the ways in which people reason. For example, being asked to visualize a situation before making a judgment about it tends to make people less utilitarian in their rule-application. They tend to reject the concept that the ends justify the means compared to those who do not visualize the situation (Amit & Greene, 2012).

The second is that they do not account for weakness of will and self-interest. This is surprising, considering that many consider morality as the opposite of being selfish. But in the classes of models we have looked at, there is no piece that explicitly accounts for differences in how selfish people are (either in general or in the moment).

Similarly, someone with a weak willpower (or with a willpower weakened by something like ego depletion), might make different moral judgments.

Third, these models do not make explicit the role of empathy, which has been shown to affect moral judgment (Batson, et al., 1997).

In response to these ignored issues, we present our own model of moral reasoning.

The Charon Model

The Charon modal is a dual-process model of moral judgment built on previous models, such as those of

[1] This approach aligns moral decision-making much closer to current theories of decision making more broadly. Daniel Kahneman's (2013) work on decision-making has shown how both reason and emotion play separate and integral roles in decision-making. They can cause separate judgments. They influence each other. They are both necessary for decision-making to work well.

Messervey (2013) and Reynolds (2006).

Important ways in which our model (see Fig. 1) deviates from previous models is its introduction of 1) a more sophisticated account of prior mental state, 2) imagination, 3) empathy, 4) the feedback process from the emotion to reason and reason to emotion, 5) self-interest, and 6) self-control.

The structure of the model, with arrows indicating influence, can be seen in Fig. 1.

Top-Down Mental States

There are many ways in which top-down processing can affect decision-making. Long-held beliefs and desires, for example, will have an impact on the way in which we perceive incoming information as well as constrain our imaginative processing. Our phenomenological experiences are determined by the interaction of mental states with either sensory input or imagination.

The types of mental states that can influence moral judgment are broad. However, we want to focus on three ways in which mental states affect moral decision-making.

The first, introduced by Messervey, Nelson, and Peach (2016), is the role of stress. Being in a stressed mental state has consequences from the onset of decision-making. The major consequence of stress is that it makes it difficult to reason effectively. In other words, if someone is already stressed then deliberative processing will be impaired. As predicted by dual-process models, those under cognitive load suffer impairments in utilitarian but not deontological reasoning, and removing time pressure increases utilitarian judgment (Greene, 2012, 127).

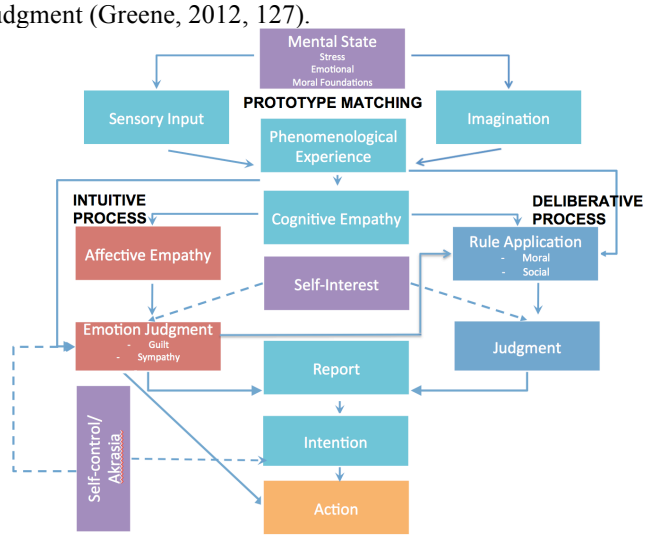


Figure 1. The Charon model of moral judgment. Arrows indicate influence. Dotted lines are mediating.

General affect can alter moral behaviour. Isen and Levin (1972) ran a study where they found that getting an extra dime out of a pay phone increased the likelihood of that person helping another by 22 times. This supports the hypothesis that a positive emotional state will result in more ethical behaviour. We assume that moral judgment was also affected, but we know of no study that differentiates moral action and judgment for the "dime effect."

Haidt's (2012) moral foundations are also stored in the "mental state" box in our model. Given the different reactions that people have to Kohlberg's (1976) moral dilemmas, leaving a broad category for rule application helps to explain the differences. Additionally, by building in the moral foundations as part of the mental states, it will also help in predicting how they will reason. For example, someone who is high on the fairness consideration may lean more towards not stealing.

Imagination

Amit and Greene (2012) examined the role visual imagination plays in moral judgments. They found that those with more visual cognitive styles compared to those with verbal styles were significantly more likely to prefer deontological judgments. They tended to prefer the rights of the individual over the collective good. Further, impairing visualization ability by overloading visual capacity made all participants less deontological in their judgments.

This lends empirical support to the introduction of the role of imagination in moral judgments. Presumably, imagination generates more visceral reactions than abstract thought, preferentially activating the emotional system, which, as we have seen, results in deontological judgments.

Empathy

Empathy, sometimes characterized as that which creates an emotional understanding of the other (Lamm, Batson & Decety, 2007) has long been linked to morality, dating as far back as Hume. More recently, there are those, such as Slote (2007), who have argued that morality is founded on empathy. While we hold that empathy is neither sufficient nor necessary for moral judgment, the ability to empathize with another person can change the ways in which people reason as well as their emotion judgments. As such, it is important that it be included as part of a moral judgment model. There are those who see empathy as that which creates an emotional understanding of the other. Our model characterizes empathy as involving having two distinctive processes. These are cognitive and affective empathy.[2]

Cognitive empathy includes the perspective taking necessary to identify the emotional state of another person and the imagination required to put oneself in that other person's shoes. In other words, it allows a person to imagine how she would feel if she were in similar circumstances to that other person. Batson et al. (1997) has shown that asking participants to take on others' perspectives makes us more likely to offer them help.

Affective empathy can be understood as actually feeling,

[2] While there are possibilities of further subdividing empathy, it would unnecessarily over complicate our model at this point.

as opposed to simply perceiving, the emotional state that another is in. It is a state of emotion matching to the other person. Affective empathy is believed to be very important for altruistic behaviour and personal relationships, as has been shown to increase the likelihood of helpful behaviours (Mullins-Nelson, Salekin, & Leistico, 2006).

Cognitive empathy appears to have the ability to change the way in which we reason when we have a better understanding of what the other person is going through. However, cognitive empathy without affective empathy may result in increases of unethical behaviour. For example, psychopaths are noted for their manipulative behaviours and are known to exploit the information that they have on others (Hare, 2003). Their cognitive empathy is intact, but affective empathy is not. Perspective taking without appropriate emotional response may result in a higher probability of unethical behaviour.

Rule Application

Our model includes a box for rule application to account for the influence of explicitly represented moral principles (such as "I don't eat meat"), as well as the use of rote retrieval of previous cases (or prototypes) of moral judgments (Reynolds, 2006).

This box also contains "scripts" that can be used in moral judgment. To demonstrate how scripts work, we will briefly discuss moral reasoning in those with autism, who do not use emotional cues (Brewer et al., 2015). People with autism are solely reliant on deliberative, reason-based approaches for making moral decision. One of the most effective strategies for teaching children with autism is the use of social scripts. These scripts include everything from everyday tasks such as proper hygiene and setting the dinner table to more complex, such as how to interact with someone in distress. Kelly and Maibom (2012) argued that autistic morality is founded in these social scripts. This may offer insight into how, despite having emotional deficits (Blair, 2005) that impair emotional processing, they are able to make ethical judgments and act accordingly. We assume that this process occurs in those without autism as well.

As described earlier in the case of moral dumbfounding, people change their minds in 10-23% of the cases of trolley/footbridge problems, based on subsequent reasoning. Charon allows for this because, like other dual-process models, reason and emotion run in parallel, both contributing "opinions" that are weighed to eventuate in a final judgment.

Self-Interest

People's self-interest clearly affects their moral *action*. Sometimes people will do something they know is wrong because it will help them. But self-interest can also affect moral *judgment*. In a magazine survey, 85% of people agreed that "If someone sues you and you win the case, should he pay your legal costs?" But only 44% agreed with "If you sue someone and lose the case, should you pay his costs?" (cited in Greene, 2012, 83).

Thus, in Charon, self-control affects moral judgment as well as moral intention (Fig. 1).

Self-Control / Akrasia

A final distinction of our model, the inclusion of self-control, will be made by appealing to psychopathy. As asserted earlier, a person can enter a situation with depleted self-control; however self-control can also fail throughout a morally-charged event. A place in where it seems to fail in psychopaths is at the level of intention.

Lack of self-control is part of the psychological construct for psychopathy. For example, M. Sib Ansari and colleagues (2010) found that psychopaths are hypersensitive to rewards, such as money and drugs. As was earlier established, psychopaths can avail themselves of utilitarian reasoning, but even when they make a judgment that something is wrong they seem unable to act accordingly (Cima et al., 2010). These findings suggest that this might be a failure of self-control: they might know right from wrong, but when they have something to gain, that goes out the window.[3]

Even for non-psychopaths, people with impulsivity issues might fail to act ethically because of a failure of self-control, even after producing the right moral judgment.

Another study found that training in self-control resulted in decreased anger (and retaliation) when facing aggression from another person (Denson, Capper, Oaten, Friese, & Schofield, 2011). The fact that anger was decreased is suggestive that moral judgment was affected, because anger has been shown to result in moral judgments (Haidt, 2012). But future studies should tease the effects of action and judgment out more carefully.

Applying Charon to Successful Psychopathy

Hare describes the psychopath in the following way: "A social predator who charms, manipulates and ruthlessly plows their way through life...completely lacking in feelings for others, they selfishly take what they want and do as they please, violating social norms and expectations without the slightest sense of guilt or regret (Hare, 2003, xi)." Though this combination of traits most often results in criminal behaviour, psychopaths make up between 1-2% of the general population (Hare, 2003)—far more than are imprisoned.

Those who meet the diagnostic criteria for psychopathy, but maintain successful and productive lifestyles are known as "successful" psychopaths. The main features which distinguish a successful psychopath is their ability to abstain

[3] Another interpretation is that even when psychopaths know right from wrong, and have self-control, they are simply not sufficiently motivated to care.

from criminal behaviour and that they succeed in their professional domain (Lykken, 1995).

Surveys sent to to people in three professions – attorneys, psychologists, and professors—asked people to report on psychopaths they worked with. It was found that successful psychopaths were described as dishonest, exploitative, low in remorse, minimizing of self-blame, arrogant, and shallow (Mullins-Sweatt, Glover, Derefinko, Miller, & Widiger, 2010).

Unsuccessful psychopaths tend to have lower scores in the facets dutifulness, self-discipline and deliberation (Lynam & Widiger, 2007). An important difference between successful and unsuccessful psychopaths is that the former is higher in conscientiousness (Mullins-Sweatt et al., 2010).

An examination of the successful psychopath suggests four ways in which the underling cognitive features of their moral decision-making system contribute to their success.

The first process is their cognitive empathy. As Babiak and Hare (2006) note, being able to read people can easily contribute to being able to manipulate them, understand their weaknesses, and use these to one's advantage. While this doesn't necessarily lead to a "moral" course of action, it can and does lead to success in many professional areas.

A third way in which the successful psychopath fits well within the Charon model of moral decision-making is that they have higher conscientiousness (Lynam & Widiger, 2007; Mullins-Sweatt et al.). In trying to explain why it is that they remain unincarcerated, one needs to refer back to the self-control/akrasia box of the model. This is the sequentially last cognitive control that we have over morality prior to action. At the very end, it determines whether we act or refrain from action both in the case of good and bad action. The successful psychopath has better inhibition than does the regular psychopath. Even if they were to arrive at a bad choice, they have the ability to refrain from action in the case of having made an immoral judgment because they have more self-control.

Finally, while there has been very little empirical work done investigating the specific moral deficits of successful psychopaths, there is no evidence to suggest that they are different from other psychopaths when it comes to their ability to use utilitarian reasoning to arrive at judgments. They would have the capacity to follow through on these judgments because of their higher self-control. While the use of utilitarian reasoning unchecked by emotional processes can lead to a cold morality, there are times when this is the necessary course of action. For example, for politicians who need to make decisions regarding who and when to send people to war when it means that they may die requires some utilitarian calculations. It is potentially the case that those with only utilitarian reasoning to use for these purposes may be well-suited to making these kinds of decisions as they are less likely to allow emotions to impede their decision-making process.

To summarize, while the Charon model can differentiate between the successful and unsuccessful psychopath and explain why it is that the former are more successful, it cannot conclude that the successful psychopath are necessarily more morally successful. They are better at not doing illegal things because they have higher self-control, but many of the ways in which they move ahead professionally are of dubious moral methods. They are manipulative, cold, hurtful, and will leave a metaphorical line of bodies behind them to get to where they want to be. (Babiak & Hare, 2007). The severity of their moral lapses may not be as severe as those of the unsuccessful psychopath, but it is a difference in kind. It would not be fair to conclude that one is moral and the other is not. Both make bad moral decisions in general, just one is more inhibited about it and stays on the right side of the law.

Conclusions

The Charon model builds on past research in ethical decision-making to produce a unique approach that accounts for a broader range of evidence than previous models. Its use of empirical evidence from psychopathy and autism results in a more robust concept of emotion judgment. It is intended to account not only for quick moral judgments, but also judgments that are arrived at after considerable reflection.

Where previous models contribute valuable concepts, such as reason, emotion, moral foundations, and dual-processing, we have shown evidence of the influence of many more factors in the complex process of moral judgment, including a more sophisticated account of prior mental state, imagination, empathy, the feedback process between emotion and reason, self-interest, and self-control.

We expect that future research will show that even more aspects of mind affect moral judgment. We encourage modelers to incorporate all of these findings into their models.

References

Amit, E., Greene, J. (2012). You see the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23(8), 861—868.*

Ansari, M. Sib, et al. (2010). Mesolimbic dopamine reward system hypersensitivity in individuals with psychopathic traits. *Nature Neuroscience, 13*, 419-423.

Babiak, P., & Hare, R. (2006). *Snakes in suits: When psychopaths go to work*. US: Harper Business.

Batson, C. D. et al. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group?. *Journal of Personality and Social Psychology, 72*, 105-118.

Blair, R. J. (2005). Responding to the emotions of others: Disassociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition, 14*, 698-718.

Brewer. R., Biotti, F., Catmur, C., Press, C.,Happé, F., Cook, R., & Bird, G. (2015). Can neurotypicals read autistic facial expressions? Atypical production of emotional facial expressions in Autism Spectrum Disorder. Autism Research, 00, 1-10.

Cima, M., Tonnaer, F., & Hauser, M. D. (2010). Psychopaths know right from wrong but don't care. *Social Cognitive and Affective Neuroscience, 5*, 59–67.

Denson, T. F., Capper, M. M., Oaten, M., Friese, M., & Schofield, T. P. (2011). Self-control training decreases aggression in response to provocation in aggressive individuals. Journal of Research in Personality, 42, 252–256.

Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. CA: University of California Press.

Gaudine, A., & Thorne, L. (2001). Emotion and ethical decision-making, *Journal of Business Ethics, 31*, 175-187.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). Emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. New York: Pantheon Books.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814-834

Haidt, Bjorkland, & Murphy. (2000). Moral dumbfounding: When intuition finds no reason. *Lund psychological reports.*

Hare, R. D. (2003). *Manual for the hare psychopathy checklist-revised (2nd ed.)*. Toronto: Multi-Health Systems.

Isen, A., & Levin, P. (1972). Effects of feeling good on helping: Cookies and kindness. Journal of Personality and Social Psychology, 21, 384-388.

Kahneman, D. (2013). *Thinking, fast and slow*. Canada: Anchor Canada.

Kelly, D. K., & Maibom, H. (2012). "But that's your role": Social models and autistic reasoning. International Conference on Thinking, London, England. *Moral education*. Toronto: University of Toronto Press.

Kelly, D., Messervey, D. L., & Nelson. E. (In press). *Exploration of the personality types and situational factors in a qualitative analysis of Israeli soldiers during the First Intifada*. (Director General Military Personnel Research and Analysis Scientific Report). Ottawa, ON: Defence Research and Development Canada.

Kohlberg, L. (1976). Moral stages and moralization: : The cognitive development developmental approach. In C. Beck and E. Sullivan (Eds.), *Moral Education*. Toronto: University of Toronto Press.

Lamm, C., Batson, C. D., & Decety, J. (2007). The neural stubstrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience, 19*, 42-58.

Lykken, D. T. (1995). *The antisocial personalities.* Mahwah, NJ: Erlbaum.

Lynam, D., & Widiger, T. (2007) Using a general model of personality to understand sex differences in the personality disorders. *Journal of Personality Disorders, 21*, 583-602.

Messervey, D. L. (2013). *What drives moral attitudes and behaviour*? Director General Military Personnel Research and Analysis Technical Report 2013-003. Ottawa, ON: Defence Research and Development Canada.

Messervey, D. L., Dean, W. H., Nelson, E., & Peach, J. (2016). *The Defence Ethical Decision-Making Model*. Manuscript in preparation.

Mullins-Nelson, J., Salekin, R., Leistico, A-M. (2006). Psychopathy, empathy, and perspective-taking in a community sample: Implications for the successful psychopathy concept. *International Journal of Forensic Mental Health, 5*, 133-149.

Mullins-Sweatt, S. N., Glover, N. G., Derefinko, K. J., Miller, J. D., & Widiger, T. A. (2010). The search for the successful psychopath. *Journal of Research in Personality, 44*, 554-558.

Pizarro, D.A., Inbar, Y., & Helion, C. (2011). On disgust and moral judgment. *Emotion Review, 3,* 267-268.

Rest, J. (1986). *Moral development: Advances in research and theory*. New York: Praeger.

Reynolds, S. J. (2006). A neurocognitive model of the ethical decision-making process: Implications for study and practice. *Journal of Applied Psychology, 91*, 737-748.

Slote, M. (2007). *Ethics of Care and Empathy*. London: Routledge.

Thorne, L., & Saunders, S. (2002). The socio-cultural embeddedness of individuals' ethical reasoning in organizations (cross-cultural ethics). *Journal of Business Ethics, 35*, 1-14.

Trevino, L. (1986). Ethical decision making in organizations: A person-situation interactionist model. *The Academy of Management Review, 11*, 601-617.

Investigating Semantic Conflict between General Knowledge and Novel Information in Real-Time Sentence Processing

Angele Yazbec (yazbec@psy.fsu.edu)
Department of Psychology, 1107 W. Call Street
Tallahassee, FL 32306

Michael Kaschak (kaschak@psy.fsu.edu)
Department of Psychology, 1107 W. Call Street
Tallahassee, FL 32306

Arielle Borovsky (borovsky@psy.fsu.edu)
Department of Psychology, 1107 W. Call Street
Tallahassee, FL 32306

Abstract

There is extensive evidence that listeners use general knowledge to predict upcoming sentence endings; however, less is known about how novel information is integrated when there is disagreement between general knowledge and novel information. The present studies use the visual world paradigm to study the semantic competition between new information and general knowledge. Experiment 1 demonstrates that listeners learn to use limited exposure to new information and their general knowledge to anticipate sentence endings that align with the action of the sentence. Experiment 2 demonstrates participants learn to use combinatorial information from stories to elicit anticipatory eye movements to the target over the general knowledge distractor. Evidence from these experiments indicates even in the presence of semantic conflict with general knowledge, listeners rapidly increase the weight of novel information rather than general knowledge.

Keywords: visual world paradigm; sentence processing; general knowledge

Comprehending Novel Events in Real Time

Listeners actively interpret spoken language about familiar events by rapidly integrating information from multiple sources to incrementally generate expectations about upcoming input (Huettig, Rommers, & Meyer, 2011 for a review). However, not all spoken events convey highly expected information. How do listeners interpret this unexpected information in real time? Some research suggests that contextual support can prompt listeners to adapt their semantic expectations in potentially anomalous contexts. For instance, Nieuwland and Van Berkum (2006) measured N400 semantic mismatch responses to written sentences conveying normally anomalous events (e.g. a peanut in love). When these sentences were situated as plausible within a larger discourse (e.g. a peanut falling in love with an almond), participants did not show a N400 semantic mismatch effect to sentences about a peanut in love by the end of the story. Similarly, other studies find that comprehenders are faster to read pragmatic anomalies (e.g. "The mouse picked up the dynamite") when they are presented in cartoon settings that support these otherwise infelicitous events (e.g. the cartoon show *Tom and Jerry*; Filik, 2008; Filik & Leuthold, 2008). This prior research indicates that, with sufficient contextual support, it is possible to interpret otherwise unlikely events as semantically plausible. However, these findings do not answer whether participants were generating specific predictions based on the new information, or simply matching the information with the current discourse.

Recent studies suggest that listeners can use recently encountered novel events (e.g. a monkey riding in a bus) to generate predictions during spoken sentence processing (Amato & MacDonald, 2010; Borovsky, Sweeney, Elman, & Fernald, 2014). Thus, this new information can be used to support anticipatory language comprehension. Additional evidence from Kleinschmidt and Jaeger (2015) suggests that in speech perception, listeners compare statistics of a novel context to prior beliefs of how speech should sound. Subsequently, listeners rapidly adapt listening behavior by weighing newly acquired information more than prior beliefs. Thus, listeners quickly learn which source of information to rely on during comprehension.

It is still unclear, however, whether and how listeners adjudicate between cases where long established general knowledge and new information directly conflict. Nieuwland and Van Berkum (2006) provide some clues: When participants read a story containing a novel situation (e.g. a smitten peanut), by the end of the story, sentences that conveyed general knowledge about the event (e.g. the peanut was salted) elicited a strong N400 mismatch effect. This suggested that, during an extended discourse, listeners temporarily "suspended" their general knowledge about peanuts in favor of the newly relevant information. The current research seeks to disentangle these questions by using a visual world paradigm (VWP) approach to explore incremental interpretation of sentences that violate general expectations, (e.g. a pilot who flies a kite, rather than an airplane). Experiment 1 explored how listeners comprehend

isolated sentences that conflict with general knowledge without any other supporting context. Experiment 2 investigated how listeners' comprehension of these same events when proceeded by stories describing those events.

Experiment 1

This experiment explores how listeners resolve semantic competition between general knowledge and new information when isolated sentences (containing an agent, action and thematic object) are the only source of conflict. It is important to note that the thematic object is always something unexpected: participants should not anticipate this object because it does not align with their general knowledge. For example, when hearing, "The pilot flies the kite," participants should make anticipatory fixations on AIRPLANE, the expected ending based on general knowledge. Only after "kite" is spoken should participants primarily fixate on KITE. This presents direct semantic competition between novel information and general knowledge. At the beginning of the experiment, it is hypothesized that this limited information will not support anticipatory interpretation of sentence, as measured by fixations towards an image of the (to be spoken) thematic object before it is spoken. Instead, listeners should rely on general knowledge for comprehension and fixate on the object that coheres most strongly with that possibility.

After hearing a series of sentences that violate general knowledge, however, it is expected that participants should adapt their comprehension strategies such that they learn to anticipate objects related to the action of the sentence, but not the general knowledge distractor. Thus when hearing, "The pilot flies the kite," in the latter half of the experiment, participants should fixate on the flyable objects of KITE and SPACESHIP until the sentential object ("kite") is spoken.

Method

Participants

Fifty-four adults participated in the study and received course credit (M= 19.11 years, male = 14). Inclusionary criteria included: normal or corrected-to-normal vision, normal hearing, no history of diagnosis or treatment of cognitive, speech, language, or attentional issues, and monolingual English speaker.

Materials

Design

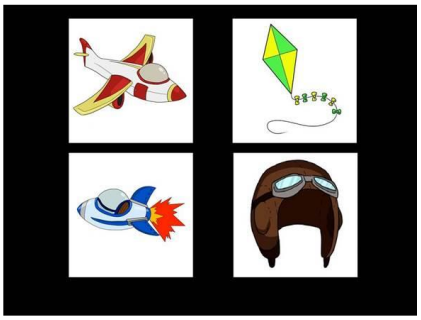
Hand-drawn, cartoon-styled pictures illustrated agents, objects, and agents acting on objects and were adjusted to a 400 X 400 pixel image on a white background. Sentences were pre-recorded by a female, native American-English speaker and sampled at a 44,150 Hz intensity level and normalized offline to 70 dB hearing level.

Agents and objects were paired such that the agent would be acting on an object that is unexpected based on general knowledge. Thus, the generally-expected object for one agent served as the target for a different agent. Each agent-object pairing was checked against Latent Semantic Analysis and the Edinburgh Associative Thesaurus to ensure the objects selected as general knowledge distractors were expected items in each pairing. Based on the combined data from these sources and previously normed relationships (Borovsky, Elman, & Fernald, 2012), 30 sentences were created. Each sentence has the following standardized construction: article 1, agent, action, article 2, object.

Sentence Comprehension Stimuli

Participants completed 15, four-alternative, forced-choice VWP tasks to assess sentence comprehension. Across all versions, image/sentence combinations were counterbalanced such that each of the 30 novel relationships was equally likely to be tested, and the locations of images of the target, verb-related distractor, agent-related distractor, and general knowledge distractor were equally likely to appear in all quadrants on the screen.

Because we were interested in the timing of fixations as the sentence unfolds, the durations of each word were controlled offline using Praat audio editing software (Boersma & Weenink, 2012) following the procedure outlined in Borovsky et al., 2014. Sentences were carefully normed such that they were all the same length and each word in each sentence started at the same time. Each trial started with 2000 ms viewing period during which the pictures appeared on the screen without auditory stimuli. After 2000 ms, the sentence was spoken and participants clicked on the object that corresponded with the sentence.



Spoken sentence:					Story Target	Agent-Related	Verb-Related	General Knowledge
The	*pilot*	*flies*	*the*	*kite.*	KITE	HELMET	SPACESHIP	AIRPLANE
The	*pilot*	*wears*	*the*	*spacesuit.*	SPACESUIT	AIRPLANE	T-SHIRT	HELMET
The	*astronaut*	*flies*	*the*	*airplane.*	AIRPLANE	SPACESUIT	KITE	SPACESHIP
The	*astronaut*	*wears*	*the*	*t-shirt.*	T-SHIRT	SPACESHIP	HELMET	SPACESUIT
The	*boy*	*flies*	*the*	*spaceship.*	SPACESHIP	T-SHIRT	AIRPLANE	KITE
The	*boy*	*wears*	*the*	*helmet.*	HELMET	KITE	SPACESUIT	T-SHIRT
Art 1	**Agent**	**Action**	**Art 2**	**Object**				

Figure 1. Illustration of the stimuli used for sentence comprehension tasks in Exp 1 and 2.

Procedure

Experimental Task

Participants sat in a stationary chair in front of a computer with a 17-inch LCD display. A five-point calibration

procedure with a black and white 20-pixel bull's-eye image ensured proper set-up and tracking of the participant's right eye. The computer running the Eyelink Experiment Builder software (SR Research, Mississauga, Ontario, Canada) presented stimuli to the participants. The instructions were to listen carefully to the sentences and then use the mouse to click on the picture that goes with the spoken sentences. Before presenting the stimuli, the 20-pixel bull's-eye image would appear on the center of the screen for drift correction. After fixating on it, the sentence comprehension trial began.

Eye Movement Recording

Eye-movements were recorded using an EyeLink 1000 remote eye-tracker with a remote arm configuration and sampled at 500 Hz. The eye-tracking camera was attached to an LCD display and adjusted so that it was 580-620 mm away from the participant's right eye. Participants wore a target sticker over the right eye to accommodate for head and eye movements relative to the camera. Eye movements, classified as saccades, fixations, and blinks by the software, were measured during the sentence comprehension task starting from the moment objects were presented on the screen until participants selected a picture. Eye movements were binned into 20 ms intervals offline for analysis.

Results

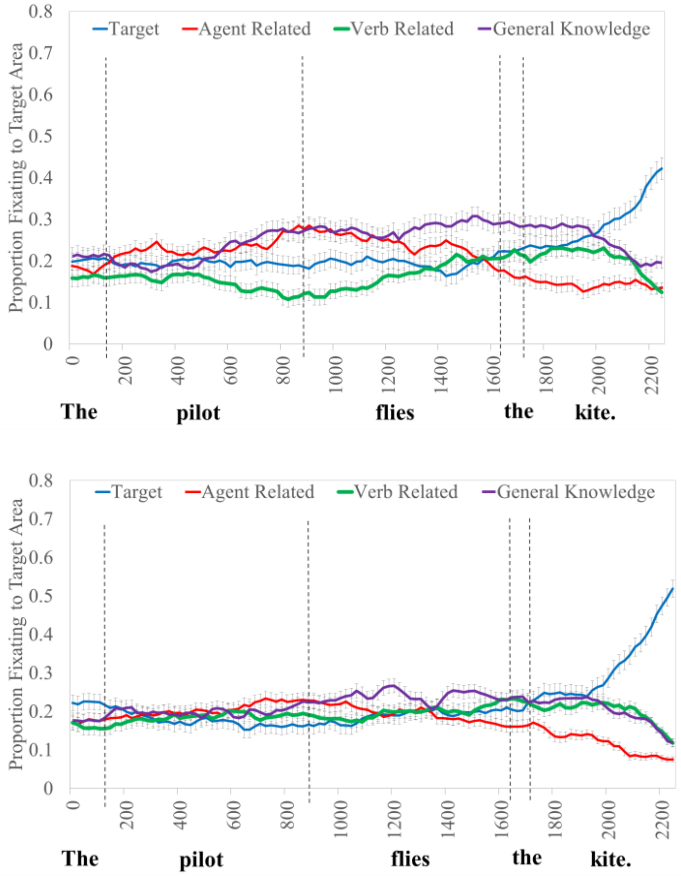
Behavioral Accuracy & Eye Movement Analysis

All analyses mentioned below are conducted only with accurate trials (accuracy= 97.9%). Many analyses were conducted to explore anticipatory fixations to the target and general knowledge distractor and whether there were any strategy changes throughout the experiment. Anticipatory fixations are defined as fixations on one object over the other objects during the action window (860-1599 ms after sentence onset) and/or second article window (1620-1720 ms after sentence onset).

Time-course Visualization

Time-course fixations across the sentence were calculated by mean proportion of time spent fixating on Target, Agent-Related, Action-Related, and General-Knowledge Distractor items in 20 ms bins, averaged across all participants.

During the first seven trials, listeners launched anticipatory fixations to the general knowledge distractor rather than the target and verb-related distractor. For example, when hearing, "The pilot flies the…" listeners anticipate AIRPLANE because according to general knowledge, airplane relates to pilots and what they fly. For the last eight trials of the experiment, there is evidence that listeners increased the weight of other possible endings; they anticipated any object that fits in with the action as a potential sentence ending, including the general knowledge distractor. For example, upon hearing the sentence, "The pilot flies the kite" they anticipate the KITE, AIRPLANE, and SPACESHIP since they relate to "flies."



Figures 2 & 3. Time-course plot of fixations to all interest areas in 20 ms bins for the first 7 and last 8 trials (*N*=54).

Analysis of Anticipatory Fixations

We measured anticipatory fixations by computing fixation proportions to the target versus the general knowledge distractor in 20 ms time bins (see Borovsky et al., 2014 for similar approach). The following main log-gaze proportions ratio were calculated: Target vs. General-Knowledge, log (P(Target/P(General)), Target vs. Verb-Related, log (P(Target/P(Verb)), Target vs. Noun-Related, log (P(Target/P(Noun))General-Knowledge vs. Target, log (P(General/P(Target)), General-Knowledge vs. Verb-Related, log (P(General/P(Verb)), and General-Knowledge vs. Noun-Related, log (P(General/P(Noun)).

Transforming the proportion of looks to each image avoids violations of linearity and homogeneity of variance and allows for an investigation in which looks to the target are relatively biased against fixations to the other objects, but not necessarily meaning fewer looks to other items. Scores of zero indicate equal number of looks to the target and distractors, positive scores show that looks to the target exceeded looks to distractors, and negative scores indicate that looks to distractors exceeded looks to the target.

A nonparametric cluster analysis was performed to determine when fixations to the target and each distractor significantly differed from each other during the sentence

(see Groppe, Urbach, & Kutas, 2011 and Maris & Oostenveld, 2007 for more detail). Results from the first seven trials reveal that listeners anticipated the general knowledge distractor over the target during the agent, action, and object windows, 720-1680 ms after sentence onset (t=68.24, p<0.01), and the verb-related distractor in the same windows, 540-1680 ms after sentence onset (t=205.34, p<0.01). Fixations to the target do not exceed those to the general knowledge distractor until the object window, 2080 ms after sentence onset (t=46.97, p<0.05). Listeners also ruled out the noun related distractor as a potential thematic object starting at the action window, 1480 ms after sentence onset (t=119.26, p<0.001). In line with our hypothesis, listeners rely heavily on general knowledge early in the experiment to anticipate sentence endings.

The last eight trials, however, reveal that fixations to the general knowledge distractor do not exceed those to either the target (t=28.76, p=0.11) or verb-related distractor (t=10.51, p=0.42) at any point during the sentence. Despite this accommodation to entertain other potential sentence endings, the noun related distractor is still not considered a possible candidate. Listeners anticipate the general knowledge distractor over the noun related distractor starting at the action window, 1420 ms after sentence onset (t=125.40, p<0.001) and the target over the noun related distractor starting at the second article window, 1700 ms after sentence onset (t=202.03, p<0.001).

While we see that objects related to the sentence action are anticipated as potential sentence endings, fixations to the target do not exceed those to the general knowledge distractor until the second article window, 2000 ms after sentence onset (t=99.14, p<0.01). This suggests that participants rapidly learned anything relating to the sentence action may be an appropriate sentence ending. Contrary to our hypothesis, however, they did not rule out the general knowledge distractor as a possible sentence ending.

Experiment 2

Findings from Experiment 1 indicated that listeners expanded their expectations of sentence endings by using their knowledge of sentence themes in addition to general knowledge to anticipate any object associated with the action of the sentence. Despite this change, the general knowledge distractor was not uniquely eliminated as a potential sentence ending by the end of the study. In Experiment 2 we ask whether the addition of stories supporting an option that conflicts with general knowledge would shift this effect more strongly with time. A possible outcome is at the beginning of the task, listeners will rely heavily on general knowledge for comprehension. About halfway through the experiment, information from the stories will override general knowledge, which would be evident in anticipatory fixations to the story target over the general-knowledge distractor. This would suggest listeners rapidly integrate new information to launch anticipatory fixations to the story target. Alternatively, there could be no anticipatory fixations to the target, but equivalent looks to the target and general knowledge distractor until listeners start to hear the object of the sentence. This would indicate that even with additional contextual support, a few exposures to novel events are not enough to override general knowledge and elicit anticipatory looks to the target.

Method

Participants

Forty-eight college aged participants participated in the study and received course credit (M= 19.07 years, male = 16). Inclusionary criteria were the same as Experiment 1.

Materials

Story Design

To familiarize participants with novel information, they listened to stories set in a cartoon world before completing a comprehension task. These stories introduced novel relationships of familiar agents, actions, and objects. In each story, three agents performed the same two actions on different objects. The general-knowledge distractor for one agent would appear as the story-target for a different agent. First, the agent would appear on the screen. Afterwards, the agent would act on an object. Subsequently, the second and third agents would be introduced, completing the same action as the first agent, but acting on different objects.

Look! There's a pilot.	*What's that? He's flying a kite.*	*What's he doing now?*	*He's wearing a spacesuit.*
		**blank screen**	
Look! There's an astronaut.	*What's that? He's flying an airplane.*	*What's he doing now?*	*He's wearing a t-shirt.*
		**blank screen**	
Look! There's a boy.	*What's that? He's flying a spaceship.*	*What's he doing now?*	*He's wearing a helmet.*
		**blank screen**	

Figure 4. Illustration of a single story block used in Exp 2. Each image was presented once with an accompanying sentence (written in italics).

Procedure

Experimental Task

The experiment was administered using the same eye tracking system and eye movement recording procedure as Experiment 1. Instructions were to listen carefully to the

stories and then use the mouse to click on the picture that goes with the spoken sentences following the stories.

Participants heard stories about novel, semantically conflicting events accompanied by depictions of the agent and the agent acting on an object. Afterwards, participants completed three sentence comprehension trials corresponding with the three agents in the story. On the screen, four of the objects from the story were presented. Before the presentation of the sentence comprehension stimuli, the 20-pixel bull's-eye image would appear on the center of the screen to serve as drift correction. After fixating it, the sentence comprehension trial began.

Results

Behavioral Accuracy & Eye Movement Analysis

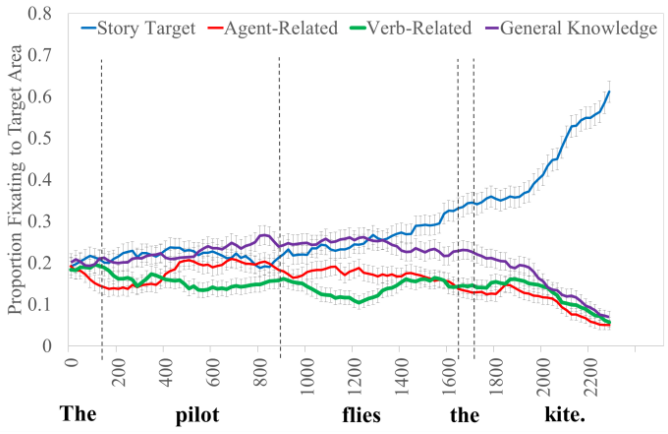
All analyses mentioned below include only accurate trials (accuracy = 98.6%). Analyses were conducted to explore anticipatory fixations to the story target over the general-knowledge distractor and whether comprehension strategy changes occurred throughout the experiment.

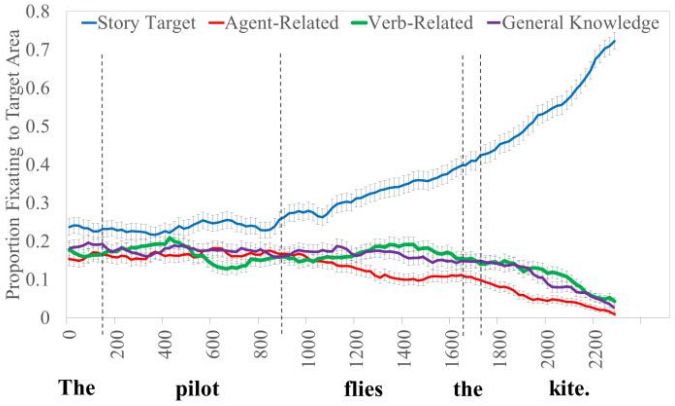
Time-course Visualization

Time-course fixations across the sentence were calculated by mean proportion of time spent fixating on the Story-Target, Agent-Related, Action-Related, and General-Knowledge Distractor items in 20 ms bins, averaged across all participants.

Visual inspection of the first seven trials shows nearly equivalent fixations to the story-target and general-knowledge distractor until after the action of the sentence is spoken. Participants are not uniquely anticipating either the story target or the general knowledge distractor as a potential sentence ending.

The remaining eight trials, however, show that listeners are able to anticipate the story target much earlier in the sentence; thus it appears that they are rapidly learning to weigh recently learned new information more than general knowledge, thus leading to more efficient comprehension.





Figures 5 & 6. Time-course plot of fixations to all interest areas in 20 ms bins for the first 7 and last 8 trials (N=48).

Analysis of Anticipatory Fixations

Due to the novel nature of the relationships in the stories, we measured whether anticipatory fixations occurred by measuring fixation proportions to the story target versus the general knowledge distractor in 20 ms time bins. This log-gaze proportion ratio was calculated: Story-Target vs. General-Knowledge, log (P(Target/P(General)).

Nonparametric cluster analyses determined when fixations to the story target exceed those of the general-knowledge distractor. Results reveal during the first seven trials, fixations to the story target do not significantly exceed those to the general-knowledge distractor until the object window, 1820 ms after sentence onset (t= 103.27, p < 0.001). Thus, there were no anticipatory fixations to the story target early in the experiment.

The remaining 8 trials, however, reveal that fixations to the story target exceed fixations to the general knowledge distractor starting at the action window, 1520 ms after sentence onset (t=213.58, p< 0.001). Thus, listeners begin to anticipate the story target over the general knowledge distractor.

For a finer grain analysis of changes in fixation patterns over time, an ANOVA was conducted to test for differences in mean fixations (averaged across all participants) to the story target and general knowledge distractor in the combined verb and second article windows (the anticipatory windows immediately before the object window) across the five experiment blocks. Significant differences in fixations to the story target and general knowledge distractor were present (F(4, 229) = 4.24, p< 0.01). Tukey's post-hoc tests revealed increased fixations to the story target over the general knowledge distractor in blocks 2 and 5 (p= 0.045) and 2 and 4 (p=0.049). There were marginally significant findings for increased fixations to the story target in blocks 1 and 5 (p= 0.072), 3 and 5 (p=0.073), 1 and 4 (p=0.078), 3 and 4 (p=0.079). Remaining comparisons were not significant. This shows that over the course of the experiment, listeners weigh recently learned information more than general knowledge. This increased weight of new information facilitated more efficient comprehension.

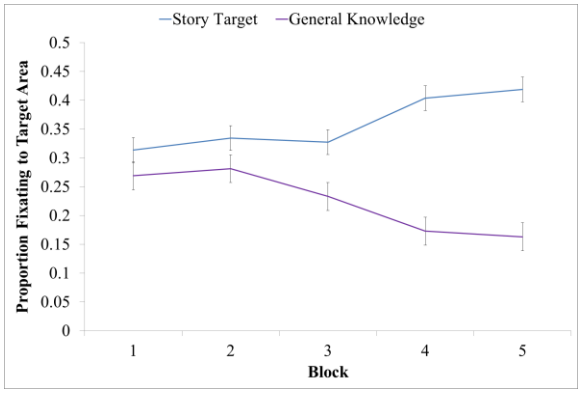


Figure 7. Mean fixations to the target and general knowledge distractor across experiment blocks (*N*=48).

General Discussion

Experiment 1 demonstrates that participants quickly considered alternative sentence endings when encountering repeated sentences that specifically conflicted with general event knowledge. Rather than generating fixations towards a general knowledge expected ending, as seen in sentence comprehension studies where sentences contain generally-expected information (e.g. Kamide, Altmann, Haywood, 2003), participants did not generate specific expectations for any particular ending. General knowledge is the primary source for comprehension (Nieuwland & Van Berkum, 2006); however, listeners quickly learned from the task that objects aligning with the action of the sentence could also be possible sentence endings. Despite increased consideration of these alternative possibilities, listeners never completely discounted their general expectations.

Experiment 2 reveals that additional discourse context via a narrated cartoon story could shift how listeners' weigh different sources of information for comprehension. During initial trials, listeners appear to weigh general knowledge and new information equally. Later trials, however, reveal that listeners weigh new information more than their general knowledge, and use that information to elicit anticipatory fixations to the story target. This suggests when general knowledge and novel information conflict, listeners rapidly learn to weigh new information more than general knowledge to understand the ongoing situation. This ability to override general knowledge is consistent with previous findings in discourse processing (e.g. Filik, 2008; Filik & Leuthold, 2008; Nieuwland & Van Berkum, 2006).

Both experiments reveal in the presence of semantic conflict between general knowledge and new information listeners rapidly increase their reliance on new information, as demonstrated by their willingness to entertain alternative sentence endings. Rapid changes in weighing new information have been demonstrated in previous studies (e.g. Kleinschmidt & Jaeger, 2015). Thus, while the utility of general knowledge never disappears, new information can outweigh general knowledge when given enough support to demonstrate the utility of the novel information.

Acknowledgements

The authors would like to thank Julie Carranza and Sam Stinchcomb for providing the vocal and artistic talents behind the stimuli used in these experiments.

References

Amato, M.S. & MacDonald, M.C. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition, 116* (1), 143-148.

Boersma, P. & Weenink, D. (2012). Praat: Doing phonetics by computer (Version 5.3.22).

Borovsky, A., Elman, J.L, & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112,* 417-436.

Borovsky, A., Sweeney, K., Elman, J.L., & Fernald, A. (2014). Real-time interpretation of novel events across childhood. *Journal of Memory and Language, 73,* 1- 14.

Filik, R. (2008). Contextual override of pragmatic anomalies: Evidence from eye movements. *Cognition, 106,* 1038-1046.

Filik, R. & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology, 45,* 554-558.

Groppe, D.M, Urbach, T.P., & Kutas, M. (2011). Mass univariate analysis of event-related potentials/fields: A critical tutorial review. *Psychophysiology, 48,* 1711-1725.

Huettig, F., Rommers, J., & Meyer, A.S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151-171.

Kamide, Y., Altmann, G.T.M., & Haywood, S.L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.

Kiss, G.R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley & N. Hamilton-Smith (Eds.), *The computer and literary studies.* Edinburgh University Press.

Kleinschmidt, D.F., & Jaeger, T.F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, adapt to the novel. *Psychological Review, 122* (2), 148-203.

Landauer, T.K. & Dumais, S.T. (1997). *Psychological Review, 104,* 211-240.

Maris, E. & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG- data. *Journal of Neuroscience Methods, 164,* 177-190.

Nieuwland, M.S. & Van Berkum, J.J.A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience, 18* (7), 1098-1111.

Extracting Human Face Similarity Judgments: Pairs or Triplets?

Linjie Li[1], Vicente Malave[2], Amanda Song[2], and Angela J. Yu[2]
(lil121@ucsd.edu, vmalave@cogsci.ucsd.edu, mas065@ad.ucsd.edu, ajyu@ucsd.edu)
[1]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA
[2]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA

Abstract

Two experimental protocols, pairwise rating and triplet ranking, have been commonly used for eliciting perceptual similarity judgments for faces and other objects. However, there has been little systematic comparison of the two methods. Pairwise rating has the advantage of greater precision, but triplet ranking is potentially a cognitive less taxing task, thus resulting in less noisy responses. Here, we introduce several information-theoretic measures of how useful responses from the two protocols are for the purpose of response prediction and parameter estimation. Using face similarity data collected on Amazon Mechanical Turk, we demonstrate that triplet ranking is significantly better for extracting subject-specific preferences, while the two are comparable when pooling across subjects. While the specific conclusions should be interpreted cautiously, due to the particularly simple Bayesian model for response generation utilized here, the work provides a information-theoretic framework for quantifying how repetitions within and across subjects can help to combat noise in human responses, as well as giving some insight into the nature of similarity representation and response noise in humans. More generally, this work demonstrates that substantial noise and inconsistency corrupt similarity judgments, both within- and across-subjects, with consequent implications for experimental design and data interpretation.
Keywords: similarity judgment, triplet ranking, pairwise rating, information theory, Bayesian modeling

Introduction

Several protocols have been developed in recent years to collect expensive and time-consuming human perceptual similarity judgments, such as among face images. Similarity is a pairwise numeric relationship between a pair of objects, where a higher value of similarity indicates that the objects are perceived to be more similar. For cognitive science, this is useful for predicting future judgments on unseen stimulus pairs, inferring a low-dimensional internal representation of the object space, identifying individual and group differences, and so on. For artificial intelligence, this type of data is often used as "ground truth" to label or categorize data, train or evaluate machine learning algorithms, predict future preferences in consumer marketing, etc.

There are two common ways to collect similarity ratings. Pairwise rating typically ask subjects to rate the perceived similarity of stimulus pairs using numbers on a specified numerical scale (such as a Likert scale). Algorithms such as classical multidimensional scaling (W. Torgerson, 1952; W. S. Torgerson, 1958) and modern variants (Borg & Groenen, 2005) make use of numeric, pairwise ratings. Another type of experiment has instead asked subjects to make ordinal judgments. One such algorithm, triplet ranking, consists of asking subjects to choose which pair of stimuli among three presented is the most similar. Relative comparisons were discussed early in the multidimensional scaling literature (W. S. Torgerson, 1958). Sometimes they are converted directly to numeric values and then used with an algorithm designed for pairs. More recently, algorithms have appeared in machine learning which learn directly from ordinal information Shepard (1962), or triplets, with no intermediate step: Generalized Non-metric Multidimensional Scaling (GNMDS) (Agarwal et al., 2007) , the Crowd Kernel algorithm (Tamuz et al., 2011) and Stochastic triplet Embedding (STE) (van der Maaten & Weinberger, 2012).

While both pairwise rating and triplet ranking have been used extensively in the literature, there has scant acknowledgement of the types of noise that can corrupt the two kinds of responses, and thus little systematic comparison of the informational utility of the two. In information terms, pairwise rating has the advantage of having more precision, and thus more capability of transmitting more information about human preferences. However, this greater precision could potentially be offset by the greater cognitive difficulty for humans to come up with numerical ratings, rather than making relative judgments. Moreover, different human subjects may interpret the numerical scale slightly differently, contributing to even more inter-subject noise. These factors can potentially translate into greater response noise or self-inconsistency, thus largely or even completely negating the precision advantage of pairwise rating over triplet ranking. A recent paper comparing several methods of collecting similarity data, (Demiralp et al. (2014)), compared pairwise rating in relative judgements terms of efficiency and consistency, and found that relative judgments can be more consistent.

More generally, in terms of the design of experiments involving extracting human similarity judgments, there has been little exploration of how many times a particular stimulus display (resulting in a judgment) should be repeated within subject or across subjects. Indeed, most algorithms simply ignore the fact there may be noise within- and across-subjects, treating the data is noise-free. In terms of experimental design, there is an obvious need to quantify and characterize the noise in order to chose the number of trials within and across subjects. In terms of cognitive science, a better understanding of the noise corrupting similarity judgments can yield insight into the nature of similarity representation in the brain.

In this paper, we utilize several different information-theoretic and probabilistic measures to quantify the information utility of pairwise rating and triplet ranking for extracting facial similarity judgments. Based on a simple Bayesian model, we compute the posterior distribution over the param-

eters of the distribution, as well as a marginal predictive distribution of the response for the next subject or the same subject on the next trial. We can then compute the information gained (entropy reduced) relative to both of these distributions, as well as a prediction error measure, for both both data collection methods. We collected face similarity judgment data, in both pairwise and triplet forms, on Amazon Mechanical Turk, and then used the various measures to quantify the informational utility of the two methods. We demonstrate that when predicting future responses from a subject's own data, triplet is better; however, when predicting a subject's response from others' data, the two methods are comparable.

The rest of the paper is organized as follows: in the Methods section, we explain the experimental design and the data modeling/analysis methods. In the Results section, we show how triplets are better than pairs within subject and vice versa across subjects. In the Discussion section, we discuss the broader impacts of our results, as well as fruitful directions of future work.

Methods

Experimental Design

We collected human similarity judgments on face images through Amazon Mechanical Turk, relying on two types of similarity judgments: a pairwise rating task, and a triplet ranking task. In the pairwise task, subjects were sequentially presented pairs of faces and asked to rate the similarity of each pair on a 9-point Likert scale (Figure 1). In the triplet task, we asked subjects to decide which pair of faces out of the three were the most similar to each other (Figure 2). The order within trials (which pairs were presented on left and right, which triplet appeared in which order), as well as the order of trials, was randomized for each subject. To ensure data quality, we also used catch trials, asking for ratings of identical stimuli (both in the pairwise and triplet cases) to identify non-compliant subjects.

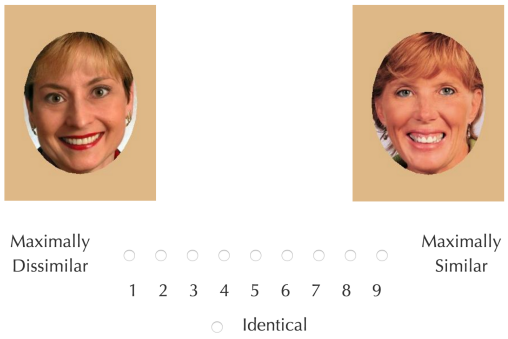


Figure 1: Sample questionnaire of pairwise rating.

In the experiment, we present 7 faces to 90 subjects in the two formats, exhaustively covering all possible pairs/triplets, for a total of 35 sets of triplets and 21 pairs. To aid the assessment of self-consistency, we present each pair and each triplet four times to each subject. Subjects carried out the

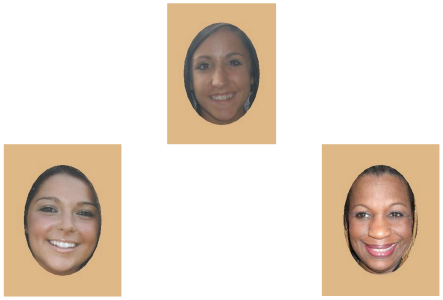


Figure 2: Sample questionnaire of triplet rankings.

experiments in five steps. Subjects are first presented a description of the task with an option of accepting it. Once the task was accepted, subjects complete a short training session, using an interface identical to the actual task interface. After the training session, subjects are prompted with the full set of faces and asked to think about the most similar and dissimilar stimuli in the set – this is to ensure subjects are aware of the full range of possible extent of similarity/dissimilarity, so as to reduce inconsistency on pairwise ratings. Afterward, subject complete the experimental task. In the last step, they provide information about themselves and submit their results. Two out of 90 subjects were thrown out due to being non-compliant.

The face images were taken from the 10k US Adult Faces database provided by Aude Oliva's group at MIT (Bainbridge et al., 2013) and then cropped for uniformity in presentation.

Data Conversion

Since subjects have 9 options in the pairwise setting but only 3 choices in the triple setting, direct comparison between these two types of data is difficult. Since the pairwise data can be easily converted to triplet data, while the reverse, or converting both to a common format, is not possible without making many assumptions about the underlying response generation process, we choose to convert the pairwise data into equivalent triplet data (based on which of the three pairs receives the highest similarity rating), and then use identical measures to compare the two in the remainder of the paper.

Correlation Analysis

As a first analysis of within-subject and across-subject consistency, we perform a correlation analysis. 35 stimuli (triplets) were rated by 88 subjects, each stimulus repeated 4 times, and there are 3 possible responses per trial.

We compute the average across-subject correlation as follows: for each stimulus, we compute the Pearson correlation coefficient between the empirical distributions (across the 3 possible responses) of two subjects, and average across all stimuli and all possible pairings of subjects.

We compute the average within-subject correlation as follows: for each stimulus in a subject, count the number of

pairwise responses that agree and then normalize by 6, the total number of pairwise comparisons across four trials. The score therefore has a value of 1/6 for really noisy data (e.g. (1, 1, 2)) and 1 for really consistent data (e. g. (3, 0, 0)). This value is then averaged across stimuli for each subject.

Statistical Modeling

Given a triplet composed of three faces {A,B,C}, the subject chooses which of the three pairs, {AB,AC,BC}, is most similar. For simplicity, we model a subject's responses to one stimulus as a multinomial distribution, $P(\mathbf{d}_x^i|vecr_x) = \text{Mult}(4,\mathbf{r}_x)$, where $d_x(l)$ is the number of times the lth pair in a triplet x is chosen ($l = 1,2,3$), and $r_x l$ is the probability of choosing the lth pair. In the across-subject analysis, we assume that all the subjects share the same preference vector $\mathbf{r}_x$ and thus generate responses from the same distribution.

We assume a conjugate prior, i.e. a Dirichlet prior distribution $p_0(\mathbf{r}_x^i;\theta)$, where $\theta = [1,1,1]$. The posterior distribution is thus also Dirichlet, and its mean is the predictive prior distribution for the next response/subject.

Information Gain Given the Bayesian response generation model, we can compare how much information is provided by pairwise or triplet data. We define *information gain* as the reduction in entropy, and we calculate the information gain relative to both the posterior distribution and the predictive distribution for the two methods, both within-subject and across subjects.

Prediction Error Use the Bayesian model, we can compute a predictive prior distribution over the next response/subject based on previous responses/subjects on the stimulus. We use MAP estimation (mode of the distribution) to predict the next response, and can therefore compute a predictive accuracy measure.

Results

As a first analysis of within-subject and across-subject consistency, we perform a correlation analysis in order to measure how consistent the subjects were with their own responses on previous trials, or with the responses of other subjects. Figure 3 shows that cross-correlation of responses across subjects for the same triplets, suggesting distinct subgroups of individuals based on overall similarity judgments. The clusters indicate that there are some consensus among the subjects that might be using the same criteria to judge facial similarity. As shown in Figure 4, we find that self-consistency is higher for triplet data than the pairwise data, while inter-subject consistency is higher for pairwise data than triplet data. The difference of across subject correlation between triplets and pairs is significant, while two methods have comparable within subject correlations.

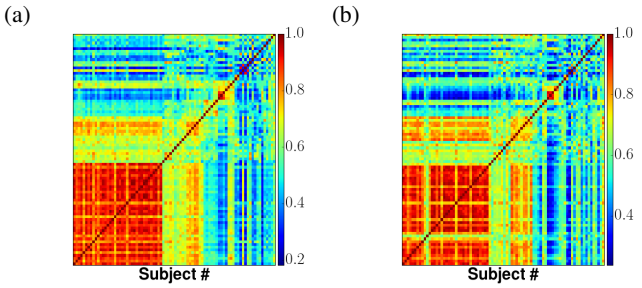


Figure 3: Cross correlation matrix of (a) triplet comparison and (b) pairwise rating. These heat maps indicate how subjects are correlated with each other. Subjects are ordered using hierarchical clustering.

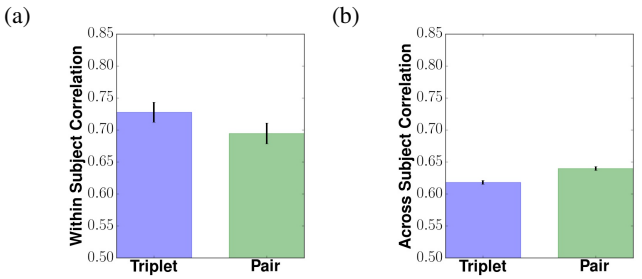


Figure 4: (a) Average within-subject correlation and (b) across-subject correlation for triplet comparison and pairwise rating. Error bars denote standard errors of the mean.

Correlation (consistency) is a very coarse measure of the utility of data, as a subject giving the same response (e.g. 1) to all stimuli on all trials would achieve maximal correlation but actually yield minimal information about any true preferences. We therefore need measures that quantify not only consistency but also diversity of responses, and that brings us naturally to information-theory. We therefore utilize a simple Bayesian generative model to capture how noisy responses arise from true (hidden) similarity percepts. We then use this model to compute both entropy-reduction related to the model parameters, based on the posterior distribution, as well as to the subject's future responses, based on the predictive prior distribution (see Methods).

We find that triplets are more informative than pairs within subject, while they are comparable across subjects. Figure 5 presents the information gain related to the posterior distribution and the predictive distribution, as we see more and more data from one subject. After all four repetitions, the information provided by triplet comparison is more than that gained from pairwise ranking, and the difference is significant. Even though the incremental information gain is decreasing as the number of repetitions increases (as we would expect), there is still significant information gained through the fourth repetition.

The result of the across-subjects analysis (Figure 6) indicates that the two methods are similar in efficacy. The incremental information gain converges to 0 as number of subjects increases. However, the first 15 subjects provide around 80% of total predictive information gain and approximately 60% of total posterior information gain.

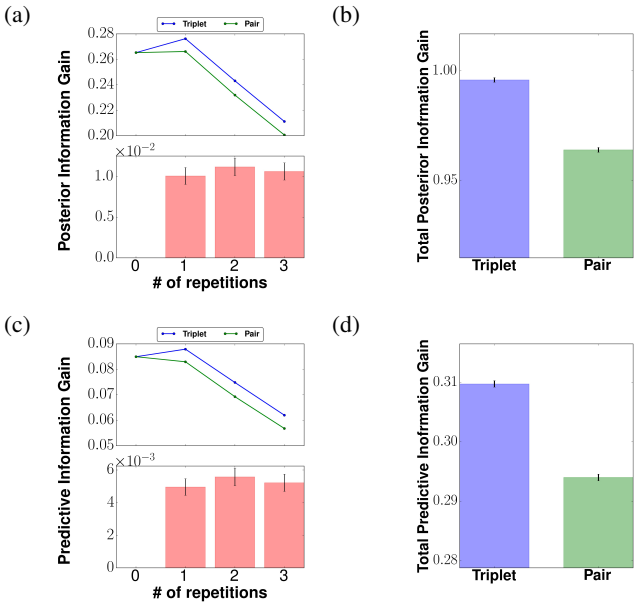


Figure 5: Within-subject information gain of (a) posterior distribution and (c) predictive distribution; total within-subject information gain of (b) posterior distribution and (d) predictive distribution. The lower bar graphs in (a) and (c) plot the point-wise differences of the upper line plots (triplet - pairwise).

While information gain measures how confident we are about the estimated model parameters or about the model predictions of future responses, it does not tell us how much better the model is getting at predicting future responses. In particular, while model precision can improve, its accuracy may saturate or even decline. One can think of the entropy-related measures as quantifying *variance* in predictive accuracy, while the prediction error measure quantifes *bias*. We therefore also use the Dirichlet-multinomial response model to make predictions (using MAP estimate) and compute an average accuracy measure by comparing to human responses.

As illustrated in Figure 7, as more and more data are collected, both within- and across-subject prediction error and entropy decrease. Prediction error decreases sharply before the 5th subjects across subjects and the second repetition within subjects. Notably, predictive error rate stops decreasing sooner than predictive entropy, indicating the model is probably somewhat misspecified: as predictive uncertainty (variance) decreases, total accuracy is already saturated (bias persists). In addition, prediction error converges to approximately 0.23 within subject, compared to 0.32 across subjects.

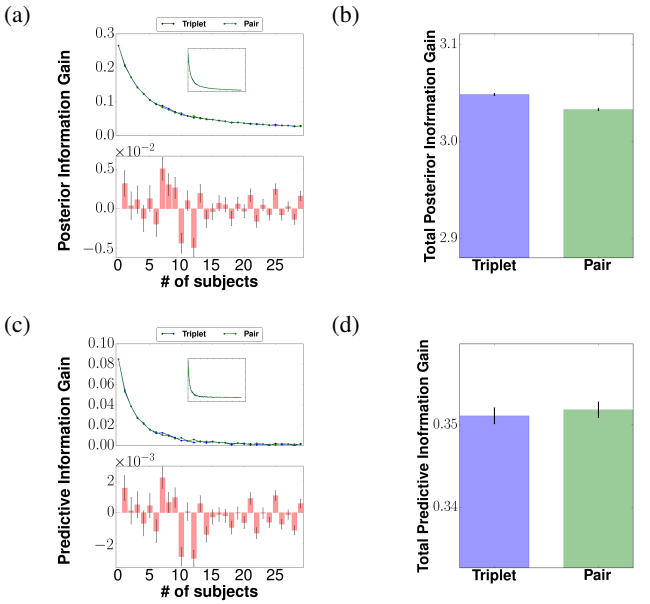


Figure 6: Across-subjects information gain of (a) posterior distribution and (c) predictive distribution; cumulative across subjects information gain of (b) posterior distribution and (d) predictive distribution. The upper line plots in (a) and (c) show the results from the first 30 subjects; the insets show the results from all subjects. The lower bar graphs in (a) and (c) show the point-wise differences of the upper line plots (triplet - pairwise).

This indicates that variability across subjects is not only due to noise but also due to systematic individual differences in preferences.

Similar to other measures, triplet data yield a higher predictive accuracy within-subject than pairwise data, but are similar when predicting across subjects.

Discussion

Similarity learning has been a well-studied topic in cognitive science research. In this area, the study of facial similarity has been particularly prominent, both due to the important role facial processing plays in human interactions, as well as due to the extremely high dimensionality of the face image space and lack of an obvious low-dimensional featural representation. While various experimental methodologies have been utilized to elicit facial similarity judgments, there has been little systematic comparison of their efficacy. More troublingly, most algorithms have assumed human responses to be free of noise and to be completely interchangeable from one subject's response to another's. To tackle some of these issues, we introduce a suite of statistical and information theoretic measures for investigating the amount and type of noise within- and across-subjects. We applied these methods, along with a simple Dirichlet-multinomial Bayesian model for response generation, to a novel crowdsourced dataset. We found that triplet ranking is more informative and predictive

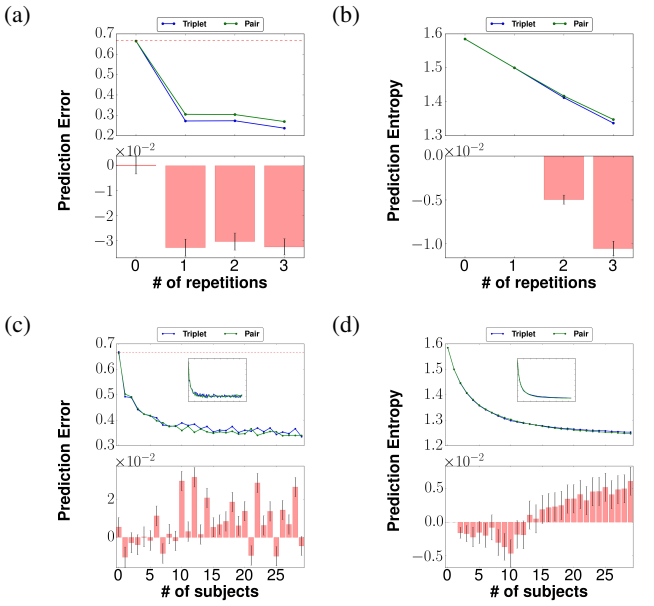


Figure 7: Within-subject (a) prediction error and (b) prediction entropy; across-subject (c) prediction error and (d) prediction entropy. Lower bar graphs indicates the point-wise differences of the upper line plots (triplet - pairwise).

for eliciting facial similarity judgments within a particular subject. It is consistent with the hypothesis that humans are better at making relative similarity judgments. Although pairwise rating has greater granularity, it has often been observed that humans give more self-consistent responses when reporting relative preferences than assigning numeric values to individual items, especially in complex judgments that involve high-dimensional input. Forcing humans to assign numerical values to complex judgments can not only fail to add information but can additionally corrupt the information available in simpler relative ranking responses. In contrast to within-subject analysis, we found triplet ranking and pairwise rating to be comparable in across-subject modeling. Why this is the case is an interesting topic for future investigations.

We covered two of the possible experimental designs used to collect similarity data, but other exist, such as spatial arrangement (Demiralp et al. (2014)), which could also benefit from our analysis. Our model has its limitations as well. We convert pairwise ratings into equivalent triplet rankings for all our analysis: this step may induce a loss of information in the pairwise data. Converting pairs to triplets not only allows an apple to apple comparison, but also minimizes the number of assumptions we need to make. One may argue that measuring mutual information between subjects' responses is a more intuitive and reasonable model free comparison. However, with 9 possible choices in the pairwise setting, we need much more data to compute mutual information of the pairwise data than the triplet data, which make it infeasible.

Our results have broader implications. Our analysis is relevant to the general problem of crowdsourcing similarity models. Depending on the goal of a similarity experiment, different methods should be chosen. If the experiment aims at modeling personal preferences, triplet comparison appear to provide higher quality data. When the goal is to find a population-level model of similarity judgements, without worrying about individual differences, then pairwise data compare well to triplet data. More generally, our work speaks to how many repetitions within and across subjects should be employed. According to the information gain analysis, our data reveal that most of the information is provided by the first 15 subjects. However, as any given model is probably not perfect in capturing human responses, greater model precision may not translate to greater ability to predict future human responses. This is consistent with our observation that, in terms of predicting future responses, two repetitions within subject and five repetitions across subjects exhaust information gain. Moreover, we provided a framework to compute how good a data collection scheme is.

The simple Dirichlet-multinomial model we introduced provides a baseline for the comparison between triplet rankings and pairwise ratings. If a more complicated model is to be proposed in the future, with our framework, all the analysis can be performed to quantify the efficacy of a technique, or compare across various techniques. An obvious next step for a better model is to integrate the relationship among different stimuli. The response model we utilized here simply assumes all triplets to be independent from each other. Similarity data are usually used to fit a latent variable model (such as multidimensional scaling), where faces are shared among triplets, and therefore one can use one set of triplet responses to potentially predict responses about another triplet with partially overlapping faces, or even completely new data. One valuable direction of future research would be to find a low-dimensional embedding of face images, in which we model similarity responses as arising from the perceptual distance between faces. The analyses proposed in this paper easily extend to latent variable models, and would be the focus of our future work.

Acknowledgments

This research is in part funded by a UCSD FISP awarded to AJY. VM is supported by a San Diego Fellowship.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D. J., & Belongie, S. (2007). Generalized non-metric multidimensional scaling. In *International conference on artificial intelligence and statistics* (pp. 11–18).

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Cao, C., & Ai, H.-Z. (2015). Facial similarity learning with humans in the loop. *Journal of Computer Science and Technology*, *30*(3), 499–510.

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Navarro, D. J., & Griffiths, T. L. (2006). A nonparametric bayesian method for inferring features from similarity judgments. In *Advances in neural information processing systems* (pp. 1033–1040).

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.

Torgerson, W. (1952). Multidimensional Scaling: I. Theory and Method*. *Psychometrika*, *17*(4), 401–419.

Torgerson, W. S. (1958). Theory and methods of scaling.

van der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding.

Decision contamination in the wild: Sequential dependencies in Yelp review ratings

David W. Vinson, Rick Dale
Cognitive and Information Sciences
University of California, Merced
[dvinson][rdale]@ucmerced.edu

Michael N. Jones
Departments of Psychology, Cognitive Science, and Informatics
Indiana University, Bloomington
jonesmn@indiana.edu

Abstract

Current judgments are systematically biased by prior judgments. Such biases occur in ways that seem to reflect the cognitive system's ability to adapt to the statistical regularities within the environment. These cognitive *sequential dependencies* have been shown to occur under carefully controlled laboratory settings as well as more recent studies designed to determine if such effects occur in real world scenarios. In this study we use these well-known findings to guide our analysis of over 2.2 million business review ratings. We explore how both within-reviewer and within-business (between reviewer) ratings are influenced by previous ratings. Our findings, albeit exploratory, suggest that current ratings are influenced in systematic ways by prior ratings. This work is couched within a broader program that aims to determine the validity of laboratory findings using large naturally occurring behavioral data.

Keywords: Sequential dependency; Online reviews; Large natural data; Decision making

Introduction

Humans are surprisingly bad at rating the absolute magnitude of their internal cognitive states. Regardless of the task, judgments of the *absolute* magnitude of a stimulus, experience, or feeling, are inherently contaminated by *relative* information from the sequence of judgments prior to the current one. Although we tend to believe that our judgment reflects the absolute value of the current experience, a good deal of the judgment is in fact determined by the relative difference between the current experience and experiences from previous trials (Laming, 1984; Stewart, Brown, & Chater, 2005). This pattern is complicated by the fact that decisions are also influenced by other factors, such as stimulus, response, and feedback (see Donkin, Rae, Heathcote, & Brown, 2015, for a review).

These cognitive *sequential dependencies* (SDs) occur whenever behavior on a trial is influenced by behavior on preceding trials. Far from rare, SDs are ubiquitous in cognition, contaminating absolute judgments from low-level perception all the way up to high-level moral judgments. We see the effect of previous trials on RT, accuracy, the type of errors produced, and interpretation of ambiguous stimuli. SDs seem to affect all levels of the cognitive system, including motor control (Dixon, McAnsh, & Read, 2012), spatial memory (Freyd & Fink, 1984), face perception (Hsu & Yang, 2013; Liberman, Fischer & Whitney, 2014), selective attention (Kristjansson, 2006), decision making (Jesteadt, Luce, & Green, 1977), and language processing (Bock & Griffin, 2000).

SDs have primarily been studied in the laboratory or at least with well-controlled experimental stimuli. They are more difficult to study in real-world scenarios because of the very large number of trials that would be required to identify their effects. In stimulus identification, for example, the immediately preceding (n-1) and non-adjacently preceding (n-2…7) items exert opposing forces on identification of the stimulus presented on trial n (Lockhead, 2004). To observe this pattern in a reasonable amount of time in the lab, carefully designed stimulus sequences are needed.

In this paper, we explore SDs in a real-world situation by mining a large natural database of online review ratings from Yelp, Inc. This is one of many freely available structured databases that can be explored. Here we use the dataset to determine if current review ratings are contaminated by previous reported experiences. In what follows we first review SD trends observed in standard laboratory tasks.

SDs in the Laboratory

Assimilation occurs whenever the judgment of stimulus n moves closer on the measurement scale to the judgment of stimulus n-k than it otherwise would have been. *Contrast* is the opposite effect, when the judgment of stimulus n moves further away on the measurement scale from the judgment of stimulus n-k. In this sense, assimilation can be thought of as a pulling force from the preceding stimulus, while contrast can be thought of as a pushing force (Zotov, Jones, & Mewhort, 2011).

Much of the early work on SDs was psychophysical in nature and involved rating unidimensional stimuli such as the loudness of a tone or length of a line (Garner, 1953; Holland & Lockhead, 1968). Identifying the *absolute magnitude* of these stimuli (e.g., line length) has been well studied: Errors when identifying stimulus n assimilate

towards the stimulus on trial *n*-1[1]. Participants are more likely to estimate the absolute value of some stimuli more similar to their most proximal previous estimate. Oddly, *categorization* of the same stimuli shows the opposite effect from the most recent response—when placing stimuli into categories, classification of stimulus *n* shows contrast from stimulus *n*-1 (Stewart, Brown, & Chater, 2002; Ward & Lockhead, 1971).

The contrast effect (push) of trial *n*-1 on the category rating of trial *n* is not limited to low-level perception, but is seen across levels of cognition. As a striking high-level demonstration, consider Parducci's (1968) example of classifying the event of "poisoning a neighbor's barking dog that was bothering you" on a moral judgment scale from 1-10 scale (where 10 is "extremely evil"). This terrible statement was rated as more evil by subjects if it was preceded by a mild judgment ("stealing a towel from a hotel") than if it was preceded by a nastier judgment ("using guns on striking workers")—a contrast effect when classifying moral judgments. Similar patterns of SDs have been seen in a variety of laboratory tasks designed to tap real-world scenarios, including brake initiation latencies in driving behavior (Doshi, Tran, Wilder, Mozer, & Trivedi, 2012), jury evidence interpretation (Furnham, 1986), and clinical assessments (Mumma & Wilson, 2006). In addition, SDs seem to be immune to practice—they are seen even in overlearned and expert behaviors.

At first glance, SDs appear to be an irrational bias in decision making (or perhaps in event memory), and have been traditionally viewed as the natural by-product of low-level brain dynamics such as residual neural activation. However, more recent theoretical perspectives suggest that SDs may be a rational property of any cognitive system. These accounts characterize SDs in terms of an individual's adaptation to the statistical regularities of a nonstationary environment with related stimulus bundles (Qian & Aslin, 2014; Wilder et al., 2010; Yu & Cohen, 2009).

Our interest is to mine Yelp, guided by knowledge from laboratory studies, to look for these naturally occurring contaminations that may affect how a business is currently rated and can expect to be rated in the future. Future business demand is largely influenced by online reviews (Cantallops, Silva, 2014; Mudambi & Schuff, 2010) affecting a business's revenue between 5-9% with this number increasing by 50% for businesses with more than 50 reviews (Luca, 2011). Computational models that explain how SDs emerge from the decision making process are now being developed, at least for low-level perceptual tasks (e.g., Mozer, et. al, 2010). These models have great promise in that they may be reversed and then applied to rating data to "decontaminate" the SD pollution in the rating, essentially producing a more accurate estimation of the individual's absolute experience of a business by removing the pollution from the relative information. This has an obvious benefit to both the service quality Yelp aims to provide, as well as a more accurate assessment of the business in question.

In Yelp, reviewers rate their experience with a business on a scale of 1 to 5 stars. Because both the rating and rating scale are most similar to categorization tasks studied in the laboratory (i.e., what is the best label to classify the exemplar, experience with the business, on a scale of 1-5 stars), our predictions are loosely drawn from SDs in categorization. Businesses typically specialize in specific services (such as a restaurants that serve American cuisine) while reviewers typically do not provide more than one review per business—similar to many individuals rating the same moral statement, opposed to one individual rating different statements. In particular, we expect that within reviewers we will see a *contrast* effect from ratings across businesses: The rating of a business will be artificially inflated if previous ratings from this reviewer were lower than if they were higher. Secondly, and more tentatively, we expect that businesses may act like categories themselves—a rating of a business is likely to *assimilate* towards preceding ratings. In addition, while little known work investigates the effects of SDs on increasing temporal distances, we anticipate that the effects of stimulus distance will be similar to temporal distance (cf. Ward, 1973). In this sense, our predictions of Yelp review ratings are a simple extension of both the perceptual work of Zotov et al. (2011), and the moral judgments of Parducci (1968).

Natural datasets are wrought with noise. Yet, where they lack structure they make for in sheer size. We do not anticipate that SDs will play such a substantial role as to alter the usefulness of user or business ratings on its face. Instead we expect to fine *echoes* of cognitive influence detectible in large datasets of naturally occurring behavior. We consider this work a guided exploration, in an effort to bridge laboratory findings with relevant and functional natural behavior.

Method

We used the most recent release of the Yelp Inc., dataset, part of Yelp's Dataset Challenge[2]. The dataset consists of just over 2.2 million reviews spanning 12 years from 2004-2016, with ratings between one (negative) and five (positive) stars, from approximately 552,000 reviewers on roughly 77,000 businesses. Reviews were provided from nine cities across four different countries (United States, Canada, Scotland and Germany). Interestingly, star ratings follow a J-shaped distribution (**fig. 1**, *top*) with mostly four and five star ratings a dip in two star ratings and roughly and equal number of one and three star ratings. In addition, the number of reviews increased steadily over Yelp's lifetime (**Fig. 1**, *center*).

[1] Interestingly, the same absolute judgment that assimilates to the most proximal past judgment *contrasts* from stimuli *n*-2…5.

[2] Further information on how to access the dataset for free as part of Yelp's dataset challenge can be found here at http://www.yelp.com/dataset_challenge

[2] Further information on how to access the dataset for free as part of Yelp's dataset challenge can be found here at http://www.yelp.com/dataset_challenge

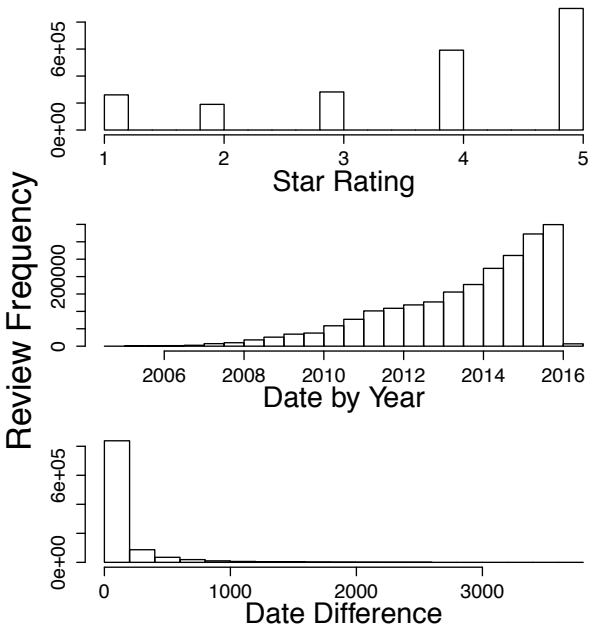


Figure 1: Frequency of reviews by star rating (*top*), frequency of reviews by year (*center*), frequency of reviews at different temporal distances in days (*bottom*).

We tested whether previous reviews influence the current review both within reviewers and within businesses. For this reason, we predict reviews from the same reviewer will be pushed *away* from previous reviews showing a contrast effect (cf. Zotov et al., 2011). Alternatively, there may be an assimilation effect for reviews within businesses. When successive stimuli are presented from the same category, the representation of that category is pulled toward the exemplar of the previous trial. Similarly, successive reviews on the same businesses, a type of category, may be pulled toward ratings from previous reviews. To be sure, this prediction is more exploratory as the nature of most laboratory studies on SDs have not focused on the influence of previous judgments from other individuals on the assessment of the same category. We anticipate that these effects will dissipate the farther away the previous review is from the current review.

Measures We first determined how far the current review rating was from its mean:

$$R_x - M(R_{T\text{-}x})$$

Where R_x is the current rating and $M(R_{T\text{-}x})$ is the average rating by reviewer or business with the current value x removed to account for possible inflation within our statistical models. This allows us to determine whether the current review is systematically biased away from the average response relative to the value of the preceding *n-k* review(s). To assess how distance is related to this deviation measure, we use *Review Distance* (*k*), and *Date Difference*. *Review Distance* is a lag measure of the number of reviews (*k*) between the current review and previous review, while *Date Difference* is a time measure of the number of days between reviews. Reviews that are farther displaced both in time and in the number of reviews may show dependence on previous review ratings.

Results

We first determined whether one's current review was related to one's previous review rating at *k*-distances. **Fig. 2** presents the mean and standard error bars for deviation of the current review rating from the mean (*y*-axis) by the previous star ratings (*x*-axis) at seven different Review Distances (*k*) within reviewers. The figure reveals a *contrast* effect that dissipates the farther away the previous review is from the current review. At *n*-1 (the immediately preceding review) for example, a 1-star rating resulted in an artificial increase in the subsequent rating from the overall mean rating. The opposite would be the case if the *n*-1 rating was 5 stars—the subsequent rating would be a lower star rating on average than it otherwise should have been. In this sense, the data are very much consistent with Parducci's (1968) "dog poisoning" example in that the current rating is systematically biased in the opposite direction from the previous rating.

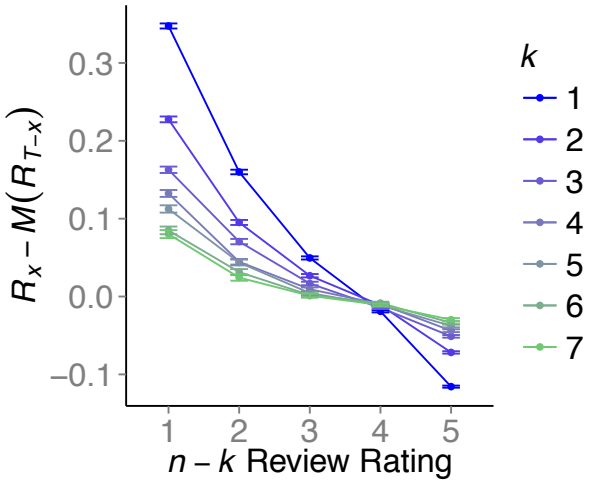


Figure 2: Within-reviewer contrast between previous and current review ratings at *k* Review Distances

To assess the visual impression quantitatively, we use a linear model to predict current review ratings by *n-k* ratings for each of seven different values of *k*. That is, we treated each value of *k* as distinct. The results, presented in **Table 1,** reveal that as the value of *k* increases, or the current review is farther displaced from the previous review, the contrast effect dissipates. However, due to size of our dataset, all results show a significant negative relationship, accounting for ~2% of variance at the closest Review Distance ($k = 1$).

Table 1: Regression model by Reviewer

k	99.9% (CI_b)	F (df)		R^2_{adj}
1	(-.11, -.10)	$2.8x10^4$	(1, $1.7x10^6$)	.016
2	(-.07, -.06)	8482	(1, $1.4x10^6$)	.01
3	(-.05, -.05)	3681	(1, $1.4x10^6$)	.003
4	(-.04, -.03)	1997	(1, $1.1x10^6$)	.002
5	(-.04, -.03)	1333	(1, $1.0x10^6$)	.001
6	(-.03, -.02)	756	(1, $9.4x10^5$)	.001
7	(-.03, -.02)	562	(1, $8.7x10^5$)	.001

Turning now toward within-business reviews, **Fig. 3** presents the mean and standard error bars for the deviation of the current review rating from the mean (y-axis) by the previous star rating (x-axis) at different Review Distances (k). This figure suggests an *assimilation* effect that dissipates the farther away the previous review is from the current review.

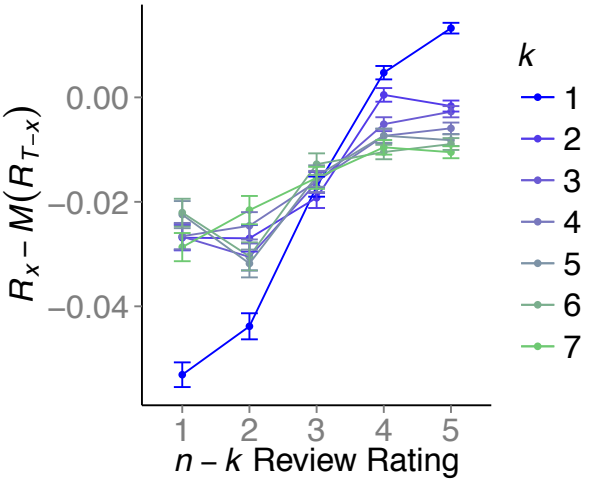


Figure 3: Within-business assimilation between previous and current review ratings at k Review Distances

The results of seven linear regression analyses on the within-business data are presented in **Table 2**. The linear regression model shows a significant negative relationship between previous and current star ratings. As the value of k increases the model accounts for less of the variance in current star ratings. All results show a significant negative relationship, though accounting for less than .1% of variance at the closest Review Distance (k = 1). Hence, this within-business assimilation effect is considerably weaker than the within-reviewer contrast.

Table 2: Regression model by Business

k	99.9% (CI_b)	F (df)		R^2_{adj}
1	(.02, .02)	1179	(1, $2.1x10^6$)	<.001
2	(.01, .01)	203	(1, $2.1x10^6$)	<.001
3	(.01, .01)	175	(1, $2.0x10^6$)	<.001
4	(.004, .007)	121	(1, $1.9x10^6$)	<.001
5	(.003, .007)	75	(1, $1.9x10^6$)	<.001
6	(.002, .006)	56	(1, $1.8x10^6$)	<.001
7	(.002, .006)	54	(1, $1.8x10^6$)	<.001

Next, we briefly explore whether there is a similar effect for reviewers and businesses by *date*. Yelp provides the date (in hours) for each review. However, reviewers occasionally provide multiple reviews within the same time frame (e.g., hours, days). To clearly differentiate between previous and current reviews, we first took the average review rating per day, by reviewer and business, and rounded this to the nearest star rating (1-5). The distribution of the number of days occurring between successive reviews is shown in **Fig. 1** (*bottom*). Temporal distances between successive reviews for both within-reviewer and –business were log-normally distributed and thus log transformed for all subsequent analyses (i.e., there were significantly more reviews that occurred closer to one another in time, than across time). We call this *Date Difference*—the number of days between reviews—and use it in subsequent analyses below.

Using a simple linear regression model, we first centered and squared Date Difference. There was a significant interaction between Date Difference and lagged star rating (t = -63), such that as the time between a reviewer's previous star rating increased, an observed contrast effect became more extreme, $F(3,1.0x10^6) = 1.65x10^4$, $R^2_{adj} = .047$.

Turning to within-business effects of temporal distance on current review ratings, a linear regression analysis revealed a significant interaction between Date Difference and lag star rating (t = -53), $F(3,1.9x10^6) = 1061$, $R^2_{adj} = .002$. This within business effect, albeit weak, is also a between reviewer effect—a study yet to be tested in a controlled laboratory environment. That is, businesses are often not reviewed sequentially by the same reviewer, if ever. However, the effects of this interaction are less clear, showing a slight assimilation effect, that reverses to contrast only at the longest temporal intervals, requiring a more sophisticated series of analysis, discussed below, prior to further interpretation.

Discussion

This study was a guided exploration into the influence that previous business ratings might have on current ratings. We tested the presence of sequential dependencies in business reviews both within reviewer and business, finding that there are significant, albeit subtle, sequential patterns.

Prior research guides interpretation of these findings. Past work shows that individuals are likely to provide contrasting evaluations when asked to rate different stimuli on a similar

rating scale (Parducci, 1968). In addition, with longer temporal intervals the effects of previous responses on current ones tends to dissipate (Doshi et al., 2014). Our predictions are loosely drawn from prior work on SDs in categorization tasks (Zotov, et al., 2011) as well as moral judgments (Parducci, 1968), such that within-reviewer ratings may contrast with previous ratings while within-business ratings may show effects of assimilation, exploring the effects of longer stimulus intervals (n = 2, 3, etc.) on current evaluations is a relatively new approach. In this respect, the current study stands as an initial exploration into the effects of SDs *in the wild.*

We found that a reviewer's current rating deviates from their mean rating in contrast with previous ratigns. If a reviewer's previous rating was positive, their current rating is more likely to be less positive than average. This effect dissipates the farther the previous review is from the current review, an effect that replicates previous findings that show contrast when making sequential moral judgments on the same scale (Parducci, 1968). In addition, this effect was observed across time, such that successive ratings that were displaced across different temporal distances were more likely to contrast with previous ratings. Findings from this analysis suggest that the observed contrast effect may be stable across time. This warrants further exploration

We found that ratings given to the same business were more likely to assimilate to the previous rating, an effect that flattened at greater review distances. The weakness of this effect is most likely due to the nature of the dataset, such that many reviewers provide reviews to a single business, so that any effect is naturally between reviewers. Just as a participant's report of an experience is not independent of their prior experience, in social circumstances there may be a sequential dependence *across* persons. We speculate (very tentatively) that such an effect, if true, would have interesting implications for how we ought to conceptualize our own judgments as entirely independent of others' judgments.

In addition, we found a very slight assimilation effect within businesses between previous and current review ratings over time. However, this effect reversed at longer temporal distances. While initially this effect appears to be consistent Review Distance findings, the reversal is puzzling. Such an effect suggests, perhaps, that evaluations of our seemingly independent experiences are represented relative to others' previous ratings. As such, further exploration as well as experimentation is necessary to more fully understand how sequential dependencies influence natural behavior.

One possible aim for future studies is to control for the business's current rating—normalizing the reviewer's rating using the average rating for the business around the time at which the reviewer's rating is made. This would provide a more absolute difference measure, adjusted to the business's average which could be used to determine if observed SD effects can be explained by the business's current rating. In addition, one could generate an artificial baseline dataset with a fixed number of ratings per reviewer. However the type of distribution we assume when generating such data may artificially inflate our findings if it does not reflect a natural distribution we see here.

Note that the Yelp distribution is not normal, exhibiting a J-shape or bimodal distribution at 1-star and 4/5-stars with a mean of 3.75. Recent studies suggest that a J-shape bimodal distribution, unique to review data, may be the result of an underreporting bias (Hu, Zhang, & Pavlou, 2009), such that reviewers are more likely not to provide reviews when the average business rating is similar to their own experience. Determining an appropriate baseline measure will be dependent, in this case, on how we interpret the cause of the J-shape distribution. For instance, this may be dependent on the type of reviewer that is considered. Critics are more likely to have a unimodal distribution whereas non-critic reviewers tend to produce a J-shape distribution (Dellarocas & Narayan, 2006). One speculative hypothesis is that when one has a choice to write a review (e.g., non-critics), they are susceptible to influences of SDs of other's reviews, resulting in a bimodal shape compared to those who may have less of a choice to write a review (e.g., critics).

Computational models that explain how sequential dependencies emerge from the decision making process can help decontaminate current evaluations so as to obtain a more accurate measure of one's experience (e.g., Mozer, et. al, 2010). Such models, though currently only developed for low-level perceptual tasks, might be fruitfully applied to areas such as online rating systems shown to impact a business's future success (Luca, 2011). Our current work is a first step toward uncovering contamination effects that may be a rational property of the cognitive system (Qian & Aslin, 2014; Wilder et al., 2010; Yu & Cohen, 2009) within naturally occurring behavior. Developing tools that can adjust for such effects might help to provide ratings that reflect the consumer's true experience.

Conclusion

Data sets such as the Yelp, Inc. dataset are incredibly noisy. Reviewers sometimes don't review for various reasons and businesses change their names and their products adapting in real time to the demands of consumer behavior. In SD experiments the stimulus is often held constant, but this may not be the case in the real world. Restaurants go out of business, while other change drastically over time. Moreover, Yelp reviews occur over a much larger time course than sequential dependency experiments, and Yelp reviewers do not see their previous review ratings at the time when they make a new rating. If there are trends in business quality or reviewer performance over time, or adjustments to Yelp's user interface (Yelp, too, must adapt to its customers), the ability to discover echoes of cognitive effects *in the wild* may be affected.

The current work targets a broader goal of validating well-known findings from carefully controlled laboratory studies in large, unconstrained, natural and noisy behavioral data. Our aim was to determine if we can use well-known

cognitive findings from controlled laboratory experiments, to sift through that noise in an effort to understand one's true experience and how that experience is affected by cognitive biases. To this end, our exploratory analysis found that current judgments, such as business review ratings, are in some way dependent on previous judgments. Reviewer's current ratings tend to be displaced from their average rating in a direction that contrasts with their previous ratings. While a business's current review rating tends to assimilate with its previous rating. Our findings, at times unpredicted and surprising, provide new avenues for future research while validating the efficacy of previous well-established laboratory findings *in the wild*.

Acknowledgments

This work was in part funded by an IBM PhD fellowship awarded to David W. Vinson for the 2015-16 academic year.

References

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning?. *Journal of Experimental Psychology: General*, *129*(2), 177.

Cantallops, S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41-51.

Dellarocas, C., & Narayan, R. (2006). A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statistical Science*, *21*(2), 277-285.

Dixon, P, McAnsh, S., & Read, L. (2012). Repetition effects in grasping. *Canadian Journal of Experimental Psychology,* 66.

Donkin, C., Heathcote, B. R. A., & Brown, S. D. (2015). Why is accurately labelling simple magnitudes so hard? A past, present and future look at simple perceptual judgment. *The Oxford Handbook of Computational and Mathematical Psychology*.

Doshi, A., Tran, C., Wilder, M. H., Mozer, M. C., & Trivedi, M. M. (2012). Sequential dependencies in driving. *Cognitive science, 36*(5), 948-963.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *Journal of General Psychology, 113,* 351-357.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 126-132.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology*, *46*(5), 373.

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics*, *3*(6), 409-414.

Hsu, S. M., & Yang, L. X. (2013). Sequential effects in facial expression categorization. *Emotion*, *13*(3), 573.

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM*, *52*(10), 144-147.

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152-183.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology, 24*(21), 2569-2574.

Lockhead, G. R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. *Review of General Psychology*, *8*(4), 265.

Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp.com. *Harvard Business School*

Kristjánsson, Á. (2006). Simultaneous priming along multiple feature dimensions in a visual search task. *Vision research*, *46*(16), 2554-2570.

Mozer, M. C., Pashler, H., Wilder, M., Lindsey, R. V., Jones, M. C., & Jones, M. N. (2010). Decontaminating human judgments by removing sequential dependencies. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culota (Eds.), *In Advances in Neural Information Processing Systems 23* (pp. 1705-1713). La Jolla, CA: NIPS Foundation.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly*, *34*(1), 185-200.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*.

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*(40), 14400-14405.

Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological review*, *112*(4), 881.

Ward, L. M. (1973). Repeated magnitude estimations with a variable standard: Sequential effects and other properties. *Perception & Psychophysics*, *13*(2), 193-200.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics*, *9*(1), 73-78.

Wilder, M., Jones, M., & Mozer, M. C. (2010). Sequential effects reflect parallel learning of multiple environmental regularities. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 2053-2061). La Jolla, CA: NIPS Foundation.

Yu, A. J., & Cohen, J. D. (2009). Sequential effects: superstition or rational behavior?. In *Advances in neural information processing systems* (pp. 1873-1880).

Zotov, V., Jones, M. N., & Mewhort, D. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics 73*, 621-639.

Using Motor Dynamics to Explore Real-time Competition in Cross-situational Word Learning: Evidence From Two Novel Paradigms

John P. Bunce (jbunce@ucmerced.edu)
Psychological Sciences, University of California Merced

Drew H. Abney (dabney@ucmerced.edu)
Chelsea L. Gordon (cgordon7@ucmerced.edu)
Michael J. Spivey (spivey@ucmerced.edu)
Cognitive and Information Sciences, University of California Merced

Rose M. Scott (rscott@ucmerced.edu)
Psychological Sciences, University of California Merced

Abstract

Children and adults can use cross-situational information to identify words' referents. What do learners retain about the potential referents that occur with a word: do they encode multiple referents or a single guess? We tested this question using novel mouse tracking and finger tracking paradigms. Adults were exposed to novel words in a series of ambiguous training trials and then tested on the words' referents. In some test trials, participants saw the target and three referents that had never occurred with the word; other test trials included a high-probability competitor that had repeatedly occurred with the word. Participants' mouse movements were slower, less accurate, and took a more complex path to the selected referent when the competitor was present, indicating that participants were aware that both the target and competitor had previously occurred with the word. This suggests that learners can accrue information about multiple potential referents for a word, and that mouse tracking provides a promising way of assessing this knowledge. However, this knowledge was not evident in participants' finger movements, suggesting that the dynamics of finger movements might not capture real-time competition between referents.

Keywords: cross-situational learning; language acquisition; mouse tracking

Introduction

For any given utterance of a word, the referential scene offers many possible interpretations. Researchers have long assumed that learners cope with such referential ambiguity in part by considering additional referential contexts in which the same word occurs (e.g., Fisher, Hall, Rakowitz, & Gleitman, 1994; Pinker, 1984; Siskind, 1996; Yu & Smith, 2007). Across situations, scene elements that are relevant to a word's meaning should occur more consistently than those that are not relevant. If learners could identify the elements that consistently co-occurred with a word across uses, then this would help them determine the word's likely referent.

Recent evidence suggests that children and adults can use cross-situational information to identify a word's referent (e.g., Scott & Fisher, 2012; Smith & Yu, 2008; Yu & Smith, 2007). For instance, Yu and Smith (2007) presented adults with a series of training trials in which they saw four novel objects and heard four made-up words. Across trials, each novel label consistently co-occurred with only one object. Following training, participants received a series of test trials in which they heard one novel label and saw its target referent and three distracters. Participants selected the target referent significantly more often than expected by chance.

These findings have raised many questions about the mechanism by which learners exploit cross-situational information in word learning. In particular, how much information do learners retain about the potential referents that occur with a word? Some researchers have proposed that learners simultaneously accrue information about an entire set of potential referents for a word (Fazly, Alishahi, & Stevenson, 2010; Yurovsky, Fricker, Yu, & Smith, 2014). On their first encounter with a word, learners encode whatever referents co-occur with that word. The next time they encounter the word, learners compare the current set of potential referents to the set previously stored in memory, adding new possibilities and updating the co-occurrence probabilities for previously encountered referents.

Other researchers, however, have argued that when learners first encounter a word, they make a guess or conjecture about the word's meaning (e.g., Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013). Learners retain this hypothesis and discard information about alternative referents. The next time learners encounter the word, they retrieve and evaluate their conjecture. If the hypothesized referent is present, then they strengthen and retain the hypothesis. Otherwise, learners discard the hypothesis and generate a new one based on the current referential scene.

Empirical attempts to test these accounts have yielded mixed results. While some studies suggest that learners accumulate knowledge about multiple competing referents for a word (e.g., Dautriche & Chemla, 2014; Smith, Smith, & Blythe, 2011; Yurovsky & Frank, 2015; Yurovsky et al., 2014), others suggest that participants retain only a single potential referent for a word across observations (Medina et al., 2011; Trueswell et al., 2013). These conflicting findings

are difficult to reconcile because the experiments have differed along many dimensions, including the number of potential referents that occurred on each observation, whether those referents were presented in isolation or in a natural scene, and the interval between observations for a word (see Yurovsky & Frank, 2015; Yurovsky et al., 2014).

Here, however, we focus on a feature that all of these prior studies share: participants' knowledge about the potential referents for a word was inferred from their patterns of explicit guesses across trials. Although the referent that a participant selects provides one index of their knowledge, this measure might nevertheless fail to capture valuable information about the process by which that selection was made. A participant might select the correct referent for a word because that participant had previously guessed that referent and thus confidently selects it again without considering other referents. Alternatively, the participant might consider how often each of the available referents had occurred with the word in the past and select the correct referent because it had the highest co-occurrence probability. In order to distinguish between these two possibilities, one would need to examine the participant's decision-making process as it unfolded in real time.

Mouse tracking provides one way of capturing this decision-making process (e.g., Dale, Kehoe, & Spivey, 2007; Spivey, Grosjean, & Knoblich, 2005). For instance, Spivey et al. (2005) asked participants to click on one of two objects on a computer screen. When the objects were phonological competitors (e.g., *pickle, picture*), participants took longer to select the target, achieved maximum velocity later, and exhibited more deviation toward the distracter than they did when the two objects' names were dissimilar. Thus, the velocity, duration, and shape of participants' mouse trajectories revealed real-time competition between alternative referents as they made their selection.

When participants select a referent for a word in a cross-situational learning task, do they experience real-time competition between referents that previously co-occurred with the word? To test this question, we devised novel mouse-tracking (Experiment 1) and finger-tracking (Experiment 2) versions of Yu and Smith's (2007) paradigm. Participants viewed training trials in which multiple novel labels occurred with multiple referents, followed by test trials in which a single label occurred with four objects. On each test trial, participants selected the object that they thought the word referred to, and we tracked their mouse/finger movements as they did so. In half of the test trials (competitor-absent trials), participants saw the target referent and three objects that had not previously occurred with the word. In the remaining test trials (competitor-present trials), one of the three non-target objects had occurred with the word in 50% of the training trials (high-probability competitor). If participants retain co-occurrence information for the set of potential referents for a word, then in the competitor-present trials they should experience online competition between the high-probability competitor and the target as they make their selection. This competition should impact their response trajectories in the competitor-present trials, resulting in differing patterns of motor dynamics across the two trial types. If, however, the participants track a single conjecture, then the frequency with which the available referents had previously occurred with the word should have no influence and response trajectories should not differ across trial types.

Experiment 1

Method

Participants 208 undergraduates (139 females) completed the experiment for course credit. All the participants used their right hand to perform the task.

Stimuli Referents were high-resolution photos of 18 common objects; each was paired with a 1- or 2-syllable nonsense word. Words were phonotactically probable in English and recorded by a female native English speaker.

Design Participants received 27 training trials and 18 test trials. On each training trial, participants saw four objects, one in each corner of the screen, and heard four labels played over the computer speaker (Figure 1). The objects for each trial were randomly selected with the constraint that each word occurred six times with its target referent, three times with a high-probability competitor referent, and less than three times with all other objects. We randomly generated two unique sets of word-object pairs.

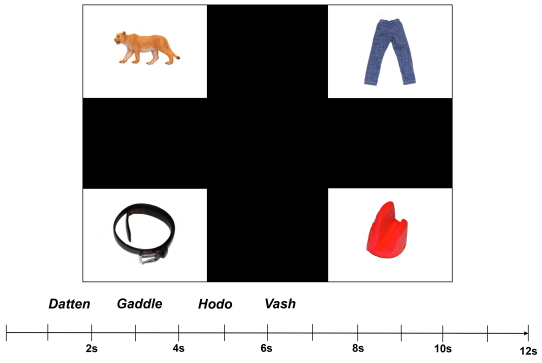


Figure 1: Sample of a single learning trial.

In each test trial, participants saw four objects, one in each corner of the screen, and heard a single label. On competitor-absent trials, objects consisted of the target and three objects that had appeared in training but had never co-occurred with the word. On competitor-present trials, the objects consisted of the target, the high-probability competitor, and two objects that had appeared in training but had never co-occurred with the word. Participants saw one of two randomized test orders. The onscreen positions of the objects were randomly generated with the constraint that on competitor-present trials the target and the high-probability competitor could not be diagonally adjacent.

This ensured a consistent angle between the target and the high-probability competitor relative to the starting position.

Procedure Participants were instructed that they would see a series of objects and hear words and afterwards they would be tested on which word went with which object. Participants then viewed the training trials on a 65 cm by 45 cm computer screen. On each trial, participants saw four objects and heard four consecutively presented audio labels. The first label occurred 1s after the onset of the trial; each subsequent label occurred 1s after the previous label. Each trial lasted 12s; trials were separated by 1s of black screen.

Following training, participants moved to a second identical computer in an adjoining room. Participants were told that they would see sets of objects accompanied by a single word and that after hearing each word, they should drag the green dot that appeared in the center of the screen to the object that they thought matched the word. Participants were told to make their decision as quickly and accurately as possible. At the start of each trial, the objects and the green dot appeared on screen; after 1s, a single audio label was delivered. The green dot was initially locked in place and unlocked at the offset of the audio label. This prevented the participants from making a selection prior to hearing the word. Once the participants released the green dot over one of the objects, the trial ended. Trials were separated by 1s of black screen. During each trial, we recorded the streaming *x, y* coordinates of the computer mouse as participants dragged the dot from the start position to their chosen referent object (sample rate ≈ 71 Hz).

Data Preprocessing On each trial, participants' final *x, y* coordinates were taken as their referent selection. To examine participants' real-time decision making, trajectories were remapped to orient the target location to the top-right corner by inverting the trajectories along the x-axis and y-axis. All trajectories were lined up to a common *x*, *y* starting position (0, 0), then individually normalized by resampling trajectories at 101 equally time-spaced values and computing, by means of linear interpolation, the corresponding mouse-coordinate values (separately for the *x* and *y* coordinate vectors).

All data analyses were conducted with R 3.1.2 (2014) and the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). All of the subsequent analysis of variance (ANOVA) models include participant as a random effect.

Results and Discussion

Participants selected the target significantly more often than expected by chance (.25) on both competitor-absent trials (M = .50, SD = .24), $t(207) = 14.88$, $p < .001$, $d = 2.07$, and competitor-present trials (M = .41, SD = .25), $t(207) = 9.22$, $p < .001$, $d = 1.28$. However, participants selected the target significantly more often on competitor-absent trials than on competitor-present trials, $t(207) = 5.93$, $p < .001$, $d = .37$.

To determine whether participants experienced real-time competition between potential referents, we examined participants' mouse trajectories. We identified trials where participants' selected either the target or the high-probability competitor (we did not analyze trials in which participants selected other objects because the angle between the start position and the object varied based on the object selected). We then separated the trajectories into three trajectory types: competitor-absent (795 trajectories), competitor-present correct (target selected; 454 trajectories), and competitor-present incorrect (high-probability competitor selected; 275 trajectories).

We next examined the participants' reaction times (from label offset to mouse-click release). A one-way ANOVA revealed a significant main effect of trajectory type, $F(2, 325) = 6.06$, $p = .003$. Planned comparisons revealed significantly faster reaction times for competitor-absent trajectories (M = 1618 ms, SD = 947) than for competitor-present correct trajectories (M = 1767 ms, SD = 1030), $z = -2.74$, $p = .015$, and competitor-present incorrect trajectories (M = 1802 ms, SD = 942), $z = -2.86$, $p = .01$. The speed of competitor-present correct and competitor-present incorrect trajectories did not differ, $z < 1$. The fact that participants were slower on competitor-present trials suggests that they experienced real-time competition between the target and the high-probability competitor.

To further examine this competition, for each trajectory we computed the maximum deviation (MD): the largest positive x-coordinate deviation from a straight line between the starting position and the selected object. For each participant, we calculated average MD values for each trajectory type. A one-way ANOVA on participants' MD revealed a significant main effect of trajectory type, $F(2, 325) = 5.44$, $p = .005$. Planned comparisons revealed significantly smaller MD values for competitor-absent trajectories (M = 65.60, SD = 68.07) than competitor-present correct trajectories (M = 76.23, SD = 77.66), $z = -2.38$, $p = .05$. Competitor-absent trajectories exhibited significantly smaller MD values than competitor-present incorrect trajectories (M = 82.77, SD = 83.89), $z = -2.91$, $p = .01$. The MD values of competitor-present correct and competitor-present incorrect trajectories did not differ, $z < 1$.

Finally, angle information and sample entropy were computed using the mousetrack R package (Coco & Duran, 2015). Angle information has been used in previous mouse-tracking studies (e.g., Dale et al., 2007) to investigate how initial movements deviated from the point of origin. Angle trajectory of mouse movements is computed as the angle relative to the y-axis for each sample in a trajectory. This provides a single measure that integrates information about x-axis and y-axis movements. A trajectory starting at the origin and moving directly to the participant's final selection would have a constant angle trajectory. If participants experienced competition between referents, then this should result in more complex angle trajectories.

To measure the complexity of angle trajectories, we submitted angle trajectory to an analysis of sample entropy for each trial (Richman & Mooreman, 2000). Sample entropy measures the complexity of a given time series. It is

robust for small time series (Yentes et al., 2013) and has been used to measure the complexity or "disorder" of mouse movement trajectories (Dale et al., 2007; McKinstry, Dale, & Spivey, 2008). Sample entropy is computed for the angle trajectories by counting the number of similar sequences, m and m+1 (up to m=5), within a similarity tolerance parameter, $0.2*SD_{\text{angle trajectory}}$ and then taking the negative logarithm of the ratio of similar sequence pair across m and m + 1, $-\ln(m/m+1)$. A time series of similar distances between data points across sequence lengths will result in lower sample entropy values. Larger sample entropy values are considered to have higher complexity.

We interpret higher values of sample entropy of angle trajectories as exhibiting competition effects through more disordered movements toward the selected object, whereas lower values of sample entropy indicate more ordered, regular movements toward the selected object. A one-way ANOVA on sample entropy revealed a significant main effect of trajectory type, $F(2, 325) = 6.67$, $p = .002$. Planned comparisons revealed that trajectories were significantly less complex for competitor-absent trajectories ($M = .13$, $SD = .06$) than for competitor-present correct trajectories ($M = .14$, $SD = .07$), $z = -2.63$, $p = .023$, and competitor-present incorrect trajectories ($M = .15$, $SD = .08$), $z = -3.23$, $p = .003$. Within the competitor-present trajectories, the complexity of the trajectories did not differ, $z < 1$.

Participants were slower and their trajectories exhibited greater deflection and complexity when the competitor was present than when it was absent. This suggests that on competitor-present trials, the target and high-probability competitor were partially active as potential response alternatives as participants were making their selection.

These results are inconsistent with what one would expect if learners retained only a single conjecture about a word's meaning. If participants only recalled their prior guess for a given word, then when that hypothesized referent was present in the test trial, they should have selected it. When that conjecture was absent, participants should have selected a referent at random from the available choices. In either case, their decision-making process should not have been affected by how frequently the available referents had previously co-occurred with the word. Contrary to this prediction, the accuracy, speed, and shape of participants' response trajectories differed across trial types, suggesting that participants were sensitive to the fact that both high-probability competitor and target previously co-occurred with the word. Thus, our results suggest that under at least some circumstances, learners can accrue information about multiple potential referents for a word.

Experiment 2

The results of Experiment 1 suggest that mouse tracking has the potential to capture learners' underlying knowledge about alternative referents for a word during cross-situational learning. In Experiment 2, we explored whether we would obtain similar results if we tracked participants' fingers as they performed our task with a touchscreen device. If so, this would provide a portable way of assessing cross-situational learning outside of the laboratory. It would also facilitate the assessment of real-time decision-making in young children, who have difficulty interacting with a mouse (e.g., Agudo, Sanchez, & Rico, 2010).

Method

Participants 79 undergraduates (60 females) completed the experiment for course credit (none participated in the previous experiment). All participants used their right hand to perform the task.

Stimuli, Design, and Procedure The stimuli, design, and procedure were identical to Experiment 1 with one exception: participants completed the test phase on a 24 cm by 19 cm touchscreen tablet and we recorded the streaming x, y coordinates of the participants' finger position as they dragged the dot from the start position to their chosen referent object (sampling rate ≈ 143 Hz). Participants were instructed to not lift their finger as they dragged the dot.

Results and Discussion

Participants selected the target significantly more often than expected by chance (.25) on both competitor-absent trials ($M = .48$, $SD = .28$), $t(78) = 7.16$, $p < .001$, $d = 1.62$, and competitor-present trials ($M = .47$, $SD = .27$), $t(78) = 7.38$, $p < .001$, $d = 1.67$. Unlike in Experiment 1, participants' accuracy did not differ across trial types, $t < 1$.

We next examined participants' finger trajectories, separated into three trajectory types: competitor-absent (293 trajectories), competitor-present correct (211 trajectories), and competitor-present incorrect (59 trajectories).

A one-way ANOVA on participants' reaction times revealed a marginally significant main effect of trajectory type, $F(2, 99) = 2.93$, $p = .058$. Planned comparisons revealed that competitor-present incorrect trajectories were significantly slower ($M = 540$ ms, $SD = 408$) than competitor-absent trajectories ($M = 395$ ms, $SD = 298$), $z = 2.38$, $p = .04$, and marginally slower than competitor-present correct trajectories ($M = 392$ ms, $SD = 331$), $z = 2.15$, $p = .08$. However, competitor-absent and competitor-present correct trajectories did not differ in speed, $z < 1$.

A one-way ANOVA on participants' MD also revealed a significant main effect of trajectory type, $F(2, 99) = 6.08$, $p = .003$. Competitor-present incorrect trajectories ($M = 38.70$, $SD = 50.08$) exhibited significantly larger MD values than did competitor-absent trajectories ($M = 22.48$, $SD = 29.35$), $z = 3.25$, $p = .003$, or competitor-present correct trajectories ($M = 21.45$, $SD = 28.91$), $z = 3.31$, $p = .003$. MD values for competitor-absent and competitor-present correct trajectories did not differ, $z < 1$.

Finally, a one-way ANOVA on sample entropy revealed a significant main effect of trajectory type, $F(2, 99) = 4.22$, $p = .02$. Competitor-present incorrect trajectories ($M = .14$, $SD = .11$) were significantly more complex than competitor-absent trajectories ($M = .11$, $SD = .10$), $z = 2.76$, $p = .02$, or competitor-present correct trajectories ($M = .11$, $SD = .09$), z

= 2.78, p = .01. Competitor-absent and competitor-present correct trajectories did not differ in complexity, $z < 1$.

The presence of the high-probability competitor did not impact the accuracy, speed, or shape of participants' finger trajectories. In contrast to the findings of Experiment 1, these results provide no indication that participants' considered the high-probability competitor as a potential response alternative as they made their selection.

General Discussion

Recent studies suggest that adults and children are able to use cross-situational information to identify the referents of novel words under at least some circumstances (e.g., Yu & Smith, 2007). However, it remains unclear how much information learners retain about the potential referents for a given word. The present study attempted to shed light on this question using novel mouse- and finger-tracking paradigms. Adult participants were exposed to novel words in a series of ambiguous learning trials and then tested on their knowledge of the words' referents. In some test trials, participants saw the word's target referent and three alternative referents that had never co-occurred with the word before, while in other trials the target referent was accompanied by a high-probability competitor that had repeatedly occurred with the word during training. In Experiment 1, participants were faster and more accurate when the high-probability competitor was absent than they were when it was present, and their mouse trajectories revealed differing patterns of motor dynamics across the two types of test trials: when the high-probability competitor was present, participants deviated more from a straight line and followed a more complex path to the selected referent. In Experiment 2, however, we observed no differences in accuracy or motor dynamics across trial types.

On the one hand, the results of Experiment 1 demonstrate that continuous measures can provide information about learners' knowledge of potential referents that is not evident in their discrete guesses. For instance, participants' in Experiment 1 were more accurate on competitor-absent than on competitor-present trials. This could reflect the fact that participants were tracking multiple potential referents for each word and the resulting competition increased the difficulty of competitor-present trials. However, unlike competitor-absent trials, competitor-present trials included two referents that had previously co-occurred with the word. These trials therefore afforded the opportunity to confirm an incorrect conjecture: if participants had previously guessed that the word referred to the high-probability competitor, they would select it if present, resulting in lower accuracy on competitor-present trials. Examining participants' mouse trajectories as they made their guesses allowed us to tease apart these two possibilities: the differing patterns of motor dynamics across the two trial types indicated that participants experienced competition between the high-probability competitor and the target. Even when participants ultimately selected the target, the way in which they did so differed when the high-probability competitor was present. These results thus suggest that assessing the decision-making process in real-time can reveal information not captured by forced-choice measures.

Converging evidence for this conclusion comes from Trueswell et al. (2013), who eye-tracked participants as they performed a cross-situational learning task. Adults viewed a series of trials in which a novel label occurred with two or five everyday objects. On each trial, participants selected the object that they thought the word referred to. Examination of participants' trial-by-trial guesses revealed that when they incorrectly guessed which referent went with a word, they performed at chance on the next encounter with that word. This suggested that participants only remembered their previous conjecture and if that guess was disconfirmed on the next trial, they were unable to remember which alternative referents were present the last time they heard the word. In contrast to their forced-choice responses, however, participants' eye movements suggested that under some conditions, they retained knowledge of multiple referents. Specifically, when participants saw only two referents on each trial, they looked significantly longer at the target than the competitor referent, regardless of whether they had guessed correctly on their previous encounter with a word. Together with the findings of Experiment 1, these results suggest that continuous measures have the potential to capture fine-grained information that learners retain about alternative referents, even when this information does not appear to impact their overt guesses.

More generally, the results of Experiment 1 suggest that mouse tracking offers a promising avenue for exploring the mechanisms behind cross-situational word learning. Incorporating mouse tracking into cross-situational paradigms in which participants select a referent on each exposure to a word (e.g., Smith et al., 2011; Trueswell et al., 2013) could provide new insight into the amount of information participants retain on a given exposure as well as how this information changes across observations. Recent work also suggests that when learners receive similar cross-situational evidence for *two* potential referents for a word, this can disrupt cross-situational learning (e.g., Bunce & Scott, in press; Yurovsky, Yu, & Smith, 2013). Mouse tracking could be used to examine the influence of carefully controlled co-occurring distracters in order to better understand when and how competition between referents leads to breakdowns in cross-situational word learning.

On the other hand, the negative results of Experiment 2 suggest that finger tracking, at least as implemented here, might not capture real-time competition between potential referents. This failure to replicate the results of Experiment 1 could be due in part to the smaller sample size in Experiment 2. However, a power analysis indicated that this sample size should have been more than adequate to detect the difference in accuracy between competitor-absent and competitor-present that we observed in Experiment 1. The fact that we nevertheless failed to observe a difference in accuracy across trial types suggests that the results of Experiment 2 were not merely a product of sample size.

Why then was there no impact of the high-probability competitor on participants' performance in Experiment 2? One possibility is that the shift in context (computer to tablet) interfered with participants' memory of the potential referents for each word, especially those that had occurred with the word less frequently. If so, then perhaps we did not observe real-time competition between the referents because participants' memory for the non-target referents was degraded and thus there was no competition to detect.

Alternatively, it may be that the participants in Experiment 2 did retain knowledge of multiple competing referents, but our finger tracking measure failed to capture that knowledge due to differences in how participants interacted with the touchscreen. Specifically, participants tested on the computer tended to keep their hand on the mouse throughout the test phase, whereas those tested on the tablet only touched the screen while making their selection. As a result, participants tested on the computer may have been more likely to initiate the movement of the dot *during* the decision-making process. Those tested on the tablet may instead have waited to touch the dot until after they had decided which referent they intended to select. Consistent with this possibility, the interval between the audio label and when participants' engaged the dot was longer on average in Experiment 2 than in Experiment 1. This suggests that perhaps we did not detect competition in Experiment 2 because our finger-tracking measure did not capture participants' online decision-making process. Future work will examine whether requiring participants to touch the dot throughout each trial will allow us to detect real-time competition in finger movements.

References

Agudo, J. E., Sánchez, H., & Rico, M. (2010). Playing games on the screen: Adapting mouse interaction at early ages. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT),* 493-497.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1-48.

Bunce, J. P. & Scott, R. M. (in press). Finding meaning in a noisy world: Exploring the effects of referential ambiguity and competition on 2.5-year-olds' cross-situational word learning. *Journal of Child Language.*

Coco, M. I. & Duran, N. D. (2015). mousetrack: Mouse-tracking measures from trajectory data. R package version 1.0.0.

Dale, R., Kehoe, C. E., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*, 15–28.

Dautriche, I. & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 892-903.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017-1063.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. R. (1994). When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua, 92,* 333-375.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science*, *19*, 22-24.

Medina, T. N., Trueswell, J. C., Snedeker, J., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *PNAS*, *108*, 9014-9019.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

R Core Team. (2014*). R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology: Heart and Circulatory Physiology*, 278, H2039–H2049.

Scott, R. M. & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition, 122*, 163-180.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61,* 39-91.

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: an experimental study of word-learning mechanisms. *Cognitive Science, 35,* 480-498.

Smith, L. B. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558-1568.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *PNAS*, *102*, 10393–10398.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology, 66,* 126-156.

Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., & Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Annals of Biomedical Engineering*, *41*, 349-365.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18,* 414-420.

Yurovsky, D. & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition, 145,* 53-62.

Yurovsky, D., Fricker, D., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review, 21,* 1-22.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science, 37,* 891-921.

Modeling N400 amplitude using vector space models of word representation

Allyson Ettinger[1] **(aetting@umd.edu)**
Naomi H. Feldman[1,2] **(nhf@umd.edu)**
Philip Resnik[1,2] **(resnik@umd.edu)**
Colin Phillips[1] **(colin@umd.edu)**
[1]Department of Linguistics, [2]Institute for Advanced Computer Studies
1401 Marie Mount Hall, University of Maryland, College Park, MD 20742 USA

Abstract

We use a vector space model (VSM) to simulate semantic relatedness effects in sentence processing, and use this connection to predict N400 amplitude in an ERP study by Federmeier and Kutas (1999). We find that the VSM-based model is able to capture key elements of the authors' manipulations and results, accounting for aspects of the results that are unexplained by cloze probability. This demonstration provides a proof of concept for use of VSMs in modeling the particular context representations and corresponding facilitation processes that seem to influence non-cloze-like behavior in the N400.

Keywords: N400, vector space models, semantic relatedness

Humans process words in relation to a context. During sentence comprehension, incoming words are processed with respect to mental states elicited by preceding words. How words and their preceding contexts are represented, and how these representations interact during processing, is not yet understood—but questions underlying this puzzle are relevant both to cognitive neuroscience and to computer science. In the present paper we bring together lines of research from both domains to enable explicit modeling of a particular class of context representation and corresponding influence on word processing: the representations that seem to underlie facilitation by lexical semantic relatedness.

In cognitive neuroscience, a measure often used to study the effects of context on word processing is the N400 component, a brain response detectable by the event-related potential (ERP) technique. The N400 is elicited by every word of a sentence, occurring approximately 400 milliseconds after the word is encountered. Its amplitude appears to be modulated by the relation of the word to its context: the worse the fit to context, the larger the N400 amplitude. However, the exact nature of the relation reflected by the N400 is complex and likely varies based on particular circumstances. A widespread generalization is that the N400 amplitude tracks "cloze" probability (Kutas & Hillyard, 1984; Kutas, Lindamood, & Hillyard, 1984), a measure based on the proportion of people who choose a word in a given context during an untimed fill-in-the-blank task. To the extent that N400 amplitudes track cloze probability, we may assume that all information used in an untimed fill-in-the-blank task can exert processing influence very shortly after the context is presented (400 milliseconds after arrival of the next word). This plausibly includes a rich and syntactically-composed representation of the context.

In other cases, however, N400 amplitude does not track cloze probability, suggesting that it does not straightforwardly reflect fit between the incoming word and a fully composed representation of the context. For instance, adding negation to a context should dramatically change likely continuations, but negation has little effect on the N400 (Fischler, Childers, Achariyapaopan, & Perry, 1985) unless additional contextual support is provided (Nieuwland & Kuperberg, 2008). Similarly, in the absence of extended processing time, the N400 appears to be less sensitive to structural information about the agent and recipient of an event (Chow, Smith, Lau, & Phillips, 2015). Specifically, in such cases the N400 amplitude can fail to reflect the low cloze probability of completions such as "A robin is not a bird" (negation violation) or "He forgot which waitress the customer had served" (agent/recipient swap). In these cases, the N400 seems to reflect fit to a less structured context representation: more like general lexical fit, or collective facilitation by semantically related words. Evidence has long indicated that the N400 is sensitive to semantic relatedness facilitation, both by sentence contexts (Kutas & Hillyard, 1980, 1984) and by single-word contexts (e.g. Bentin, McCarthy, & Wood, 1985; Kutas & Hillyard, 1989; Holcomb, 1988; Brown & Hagoort, 1993). Semantic relatedness effects on the N400 are broadly accepted and frequently cited, but to our knowledge no explicit models generating quantitative predictions of these effects have yet been proposed.

Meanwhile, in computer science, explicit models have emerged that allow straightforward computation of relations between words—and by extension, of relations between individual words and groups of words. Vector space models (VSMs), now widely used for natural language processing in computer science, use distributional characteristics of words in text (that is, the types of contexts that words tend to occur in) to form representations for individual words in the form of numeric vectors. A prominent early example of this concept in cognitive science is that of latent semantic analysis (LSA), discussed extensively by Landauer and Dumais (1997). Since the development of LSA, much continued progress has been made in building and optimizing such VSMs.

Once a word is represented by a vector, we can think of this representation as being situated in a multi-dimensional space, and we can compute relations between different word representations based on their orientations or locations in that space. Most common is the use of cosine similarity, a measure based on the angle between vectors. VSMs also allow flexible computation of the relation between a word and a group of words—for instance, using a very simple combination function such as an average, one can generate a single vector represen-

tation reflecting characteristics of multiple words, without additional information (such as agent/recipient roles) that would be encoded in a representation arrived at through full syntactic composition. Such a representation could be used to simulate the types of representations hypothesized to underlie non-cloze-like behavior in the N400, when amplitude seems to reflect sensitivity to general semantic relatedness and not to other information present in fully composed representations.

There is one study to our knowledge that finds a correspondence between VSM relations and N400 amplitude: Parviz, Johnson, Johnson, and Brock (2011) assess a suite of variables as possible predictors of N400 amplitude in a non-linear regression model, including cosine similarity from an LSA-type VSM, which they find to be significant. Several studies have also shown correspondence between semantic priming and VSM measures (Mandera, Keuleers, & Brysbaert, 2016; Lapesa & Evert, 2013; Jones, Kintsch, & Mewhort, 2006; Padó & Lapata, 2007; Herdağdelen, Erk, & Baroni, 2009; McDonald & Brew, 2004), suggesting that VSMs can predict cognitive semantic relatedness effects more generally.

In the present paper we build on this foundation to implement a VSM-based model of aspects of sentence processing, intended to test whether a model with simple, averaging-based context representations can successfully simulate non-cloze-like N400 results, as the above reasoning would predict. We choose to model the results of the Federmeier and Kutas (1999) N400 study. This study is one in which certain results deviate from predictions of cloze probability, making it a valuable testing ground for a model intended to capture semantic relatedness effects believed to underlie non-cloze-like N400 behavior. The study is also ideal because it explicitly manipulates relations between target words and their contexts, but bases assumptions about these relations on measures such as cloze probability and plausibility ratings. We use this opportunity to test whether relations computed based simply on collective effects of context words will generate better predictions of the observed results. We show that this VSM-based model captures many major characteristics of the study's N400 results, including the key result that deviates from cloze predictions. The model's relation computations also largely align with the assumptions made by the authors about their stimulus manipulations—with one main exception, which is also the key factor accounting for the deviation from cloze. This suggests that the model has successfully captured aspects of semantic relatedness-based processes exerting influence on the N400, and that access to predictions based on such a model can lend useful perspective in interpreting N400 results.

Federmeier and Kutas (1999)

Federmeier and Kutas (1999) investigated N400 effects in contexts that predict a particular completion word, and are then completed by words with varying levels of similarity to that predicted word. The authors found that unpredicted (zero-cloze) words elicit larger N400s, as expected. However, when the unpredicted item is similar to the predicted word in strongly predictive contexts, the N400 amplitude is reduced.

Table 1: Sample stimuli.

Stimulus (**expected/within/between**)
He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of **football/baseball/monopoly**.
The day before the wedding, the kitchen was just covered with frosting. Annette's sister was responsible for making the **cake/cookies/toast**.
He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that **lipstick/mascara/earring**.

To accomplish this, Federmeier and Kutas constructed two-sentence contexts with three possible ending types: "expected", "within-category", and "between-category". Expected targets are predicted by the context, with high cloze probability. Within-category and between-category targets are both unexpected in the context—cloze probability of approximately zero—but within-category targets share a category with the expected target.[1] If N400 amplitude were to track cloze probability, then we would see reduced N400 amplitude for the expected target condition, and roughly identical, unreduced N400 amplitude for the two unexpected target types, regardless of category relationship to the expected target.

The stimuli were furthermore binned into two conditions based on the extent to which the context constrained toward the expected word: stimuli were classified as either "high-constraint" and "low-constraint", according to a median split on cloze probability of the expected target.

Figure 1 shows the results of Federmeier and Kutas's study. Negative voltages are plotted upward, with higher N400 amplitude (corresponding to a word that is less expected or facilitated) represented by a greater negativity. In both constraint conditions, we see that the expected target has extremely low N400 amplitude, compatible with strong facilitation. Additionally, in both constraint conditions, between-category targets show very high N400 amplitude, compatible with lack of facilitation. There is also a main effect of constraint level, but the key difference emerges for within-category targets: in high-constraint contexts only, within-category targets show reduced N400 amplitude—despite the fact that within-category targets (like between-category targets) have roughly zero cloze probability. Federmeier and Kutas interpret this result as evidence of a mediating influence of the expected target in high-constraint contexts: strong prediction of the expected target in these contexts causes features of that target to be pre-activated, and because of semantic overlap between expected and within-category targets, the latter targets are facilitated as well.

Federmeier and Kutas's interpretation operates on the as-

[1] Federmeier and Kutas explain that "Categories were chosen to be those at the lowest level of inclusion for which the average undergraduate student could be expected to readily differentiate several exemplars." See Table 1 for examples.

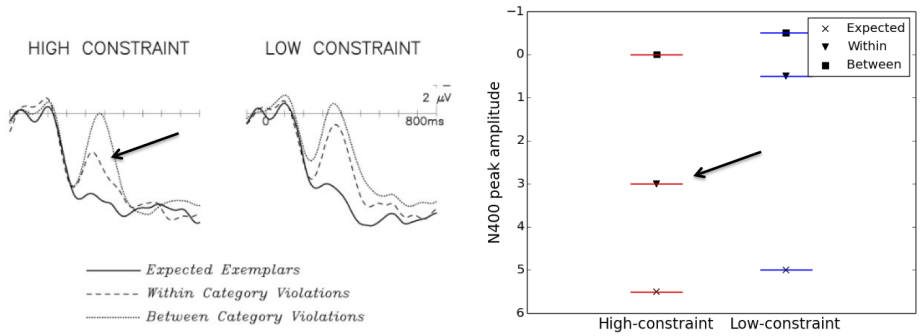


Figure 1: Federmeier and Kutas (1999) N400 results. Left: original results as reported by the authors. Right: Results re-plotted as points representing peak N400 amplitude, for greater ease of comparison to simulation results below. Arrows indicate key facilitation in high-constraint within-category condition.

sumption that high- and low-constraint contexts are equally related to the within-category targets, such that the observed facilitation of within-category targets in high-constraint contexts must be explained by some additional factor. This need motivates their hypothesized mediation by pre-activation of expected target features in high-constraint contexts. An alternative explanation would offer itself if high-constraint contexts were directly more facilitative of within-category targets than are low-constraint contexts. Federmeier and Kutas assume that this is not the case, based on cloze probability measures and plausibility ratings. However, there are other ways that we might conceive of relation to context—in particular, we should consider relations based simply on the collective effect of context words (as opposed to the fully structured and compositional context representations likely to be informing untimed cloze and plausibility decisions). In the next section, with the help of VSMs, we explore whether a notion of fit to context based on collective semantic relatedness can explain the facilitation where cloze and plausibility do not.

Federmeier and Kutas make available a sample of 40 of their experimental stimuli; we run our simulation on that sample.

Model

For testing assumptions and modeling the results of this study, we choose a VSM generated by the word2vec model (Mikolov, Chen, Corrado, & Dean, 2013). Unlike LSA, word2vec uses a neural network to optimize word vectors based on their ability to predict nearby words. In systematic comparisons of VSM performance on various semantic tasks, this model has shown consistently strong and often superior performance (Baroni, Dinu, & Kruszewski, 2014; Levy, Goldberg, & Dagan, 2015). For this reason, we select word2vec as a state-of-the-art VSM of word representations. We train the model on approximately 2 billion words of semantically diverse web data from the ukWaC corpus (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008), training vectors of 100 dimensions using the skip-gram architecture, which maximizes the probability of surrounding words given the current word.

Once we have trained this VSM, each word of the vocabulary is represented as a point within the resulting vector space. For a sentence context, we will refer to vectors for the expected target, within-category target, and between-category target as vectors E, W, and B, respectively.

We model the mental state induced by preceding context words through a simple averaging procedure: vectors for selected context words are averaged to obtain a single context vector C. This representation reflects the collective effect of the included words, without many of the additional structural cues that might inform a cloze decision. In selecting context words, we attempt to isolate the most informative words, which we hypothesize will have the strongest influence upon the context representation. We try two selection methods: *anchored* and *agnostic*.

In the anchored setting, we use relation to the expected target as a proxy for informativeness: using the expected target as an anchor, we select the four context words with highest cosine similarity to that expected target.[2] We employ a minimum cosine similarity of 0.2 (chosen by examination of context word cosine similarities in a small subset of stimuli) to further filter words bearing little relation to the target.

In the agnostic setting, we take the top four words based on negative log frequency (that is, the least frequent words), excluding person names (e.g., *Annette*). This is equivalent to choosing words based on maximum surprisal (information content) as determined by a unigram probability model.

The modeling results suggest *prima facie* that the anchored setting is more successful in isolating the most significant words of the context. If so, this is likely due to the fact that the frequency metric underlying the agnostic setting, while reasonable, is a rather blunt tool for assessing informativeness.[3]

Within these settings, we test two types of average: unweighted, and weighted inversely by linear distance. The latter average aims to instantiate the hypothesis that the effect of a context word would decay over time, with earlier words having less influence than later words.

[2] One target. *polar bear*, is made up of two words; this is represented as the average of the two separate word vectors.

[3] As we caution below, however, at the current stage we should not be overzealous in making fine-grained modeling decisions based on the linear fit of only six datapoints.

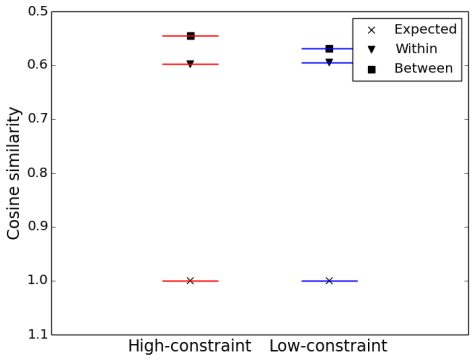


Figure 2: Cosine similarity to expected target

Once we have obtained this context vector C, it can be represented as a point within the space that contains vectors E, W, and B, and its relation to these vectors can be computed. For every stimulus, we take the cosine similarity between C and each of E, W, and B, and we average these cosine similarity values across stimuli within each condition, in order to simulate average N400 amplitude.

We also compute cosine similarity between E and W and between E and B. This allows us to assess the model's representation of the relations between different completion words.

Simulation Results

Figure 2 shows the results of the comparison between target types E, W, and B—this test simply serves as a control, to compare the model's relation computations against those assumed by Federmeier and Kutas, and to check for confounds. In Figure 2 and those that follow, cosine similarity is plotted on the y-axis with the negative direction upward, to facilitate comparison to N400 plots in Figure 1: higher cosine similarity predicts lower N400 amplitude. Note in Figure 2 that the expected word vector E is at cosine similarity of 1, as this is a comparison of a vector to itself. As for the other two comparisons, we see that the model predicts on average a nearly identical level of relation between expected words and within-category words in both constraint conditions. We see a slightly greater distance between the expected word E and the between-category word B in the high- than the low-constraint condition. In both cases the model's relations are roughly consistent with the categorical relations assumed by the experimental manipulation: within-category items are indeed represented as being closer to the expected targets than are the between-category items. The lack of any discernible difference in the expected/within-category target relation between constraint conditions also rules out the possible confound of differing relation strengths between the targets themselves.

Figure 3 shows the full simulations under the anchored and agnostic settings, respectively. (The right hand side of Figure 1 presents Federmeier and Kutas's results in the same plotting format, for ease of comparison.) In these figures we see several things. First, we see a main effect of constraint consistently captured across settings: for each ending type, average cosine similarity to context is higher in the high-constraint condition, corresponding to greater facilitation (lower N400 amplitude). This is consistent with the main effect observed in Federmeier and Kutas's N400 results.[4]

In addition, we see that for the most part, looking independently at the high- and low-constraint conditions, the three ending types pattern as the experimental paradigm predicts: expected targets are most facilitated by the context, while within- and between-category targets are less facilitated. We also see that under all settings, in the high-constraint condition the within-category target falls at an intermediate position between the other two target types. In the low-constraint condition, however, three of the four settings have within- and between-category conditions in reversed or roughly identical positions. The fact that between-category targets in the low-constraint condition fail to fall farthest from the context, often switching with within-category targets, may reflect similar factors to those that lead to within- and between-category conditions having statistically indistinguishable N400 amplitudes in Federmeier and Kutas's results.

Returning to our main effect of constraint: recall Federmeier and Kutas's assumption that facilitation of high-constraint within-category targets cannot be explained by direct relation to context. We see in Figure 3 that the VSM-based representation of context—under both anchored and agnostic word selection settings—does predict greater facilitation of within-category targets in the high-constraint as compared to the low-constraint condition, suggesting that direct relation to context could offer a valid explanation for this deviation from cloze probability. This result both lends support for the explanatory power of our simple non-syntactically-composed context representations, and gives us reason to consider direct facilitation by contextual semantic relatedness as an alternative account for Federmeier and Kutas's results.

Discussion

In this study, we used a vector space model to predict N400 amplitudes observed in Federmeier and Kutas (1999). We find that by representing words in a vector space, averaging vectors of informative context words, and taking cosine similarity measures between the averaged context vector and each of its possible completions, we are able to simulate key aspects of Federmeier and Kutas's N400 results: the basic patterning of expected, within-category, and between-category items within constraint conditions, as well as the main effect of constraint. Our model accounts for the deviation from predictions of cloze probability in the high-constraint condition, and in doing so calls into question the assumption that the key result of this study cannot be explained by a direct facilitation between context words and the within-category targets.

[4]Having access only to 40 items of the original Federmeier and Kutas study, we are not making claims of statistical significance for this pattern of results in the models. This simulation is intended instead as an exploratory proof of concept.

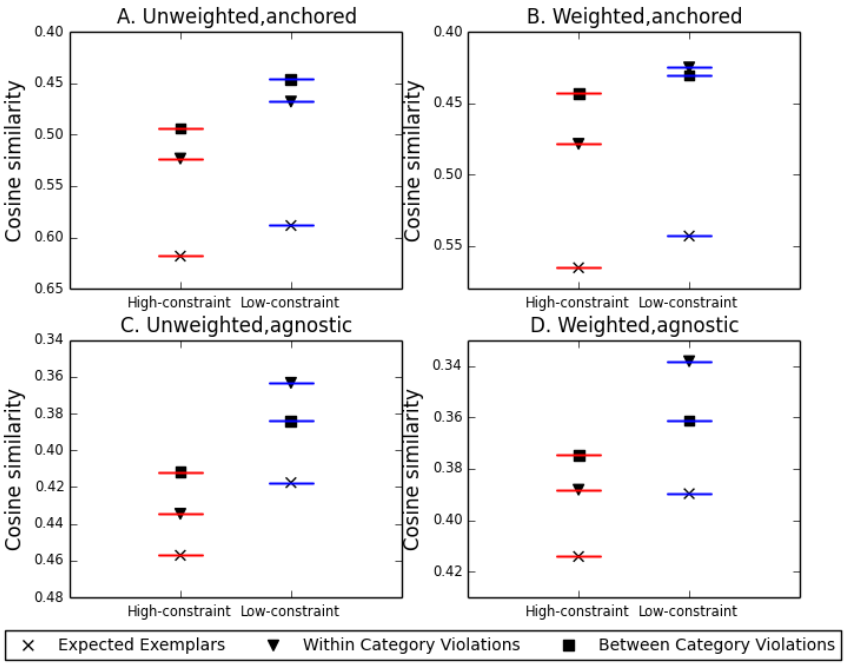


Figure 3: Simulations in four settings. A) Context average unweighted by linear distance and words selected with expected target as anchor. B) Context average weighted by linear distance and words selected with expected target as anchor. C) Context average unweighted by linear distance and words selected by low frequency. D) Context average weighted by linear distance and words selected by low frequency.

At face value, if we assume a linear relation between cosine similarity and N400 amplitude, then Figure 3B is the most faithful simulation of the N400 results. We might take this as evidence in favor of a cognitive model in which activation spreads from informative words (with relation to expected target being a better proxy for informativeness), and in which a word's influence decays over time. However, we caution against drawing strong cognitive conclusions from this single set of simulations. First, we are modeling only six datapoints, without claims of statistical significance. Second, we are for the moment assuming a linear relation between cosine similarity and N400 amplitude, which is very likely an oversimplification. Consider ceiling and floor effects, which are understood to influence N400 amplitude. Floor effects, at least, are likely a factor in Federmeier and Kutas's results, as the study finds no significant effect of constraint on N400 to expected targets, despite the fact that high- and low-constraint contexts are defined precisely by how predictive they are of the expected target. The fact that our cosine similarity measure does reflect an effect of constraint on expected targets suggests that we are capturing important aspects of the context-to-target relation with this measure. However, it also suggests that we will need a nonlinear linking hypothesis to predict the N400 with more precision. This means that we should not be quick to dismiss the other settings in Figure 3, as they could ultimately prove to be the more accurate simulations once we identify the proper linking hypothesis.

We see these simulations as a valuable proof of concept. To understand what the N400 can tell us about contextual representations and their influence on incoming words, we need to be able to tease apart the contributing factors at play when N400 amplitude tracks untimed measures such as cloze, versus the factors at play when it deviates from such measures. VSMs allow for explicit modeling of collective word relations, and as a result they are a promising tool for generating quantitative predictions from a range of hypotheses regarding the semantic relatedness-based processes that may underlie deviations of N400 amplitude from cloze. In the above simulations, we indeed find support for the ability of these models to use averaging-based context representation and simple relation computations to capture aspects of N400 behavior that deviate from the predictions of cloze probability.

It should be noted that our results need not be in direct conflict with Federmeier and Kutas's general framework. Since cosine similarity is computed by dimension-wise comparison of one vector to another, one could think of higher cosine similarity between context and target as representing greater pre-activation of target features as a result of context. As for Federmeier and Kutas's hypothesis of mediating influence by pre-activation of the expected target: in our anchored setting, one might argue that selecting context words based on the expected target instantiates a version of Federmeier and Kutas's mediation account. However, though the expected target does have an increased role in this setting, it is still the context words, and not the expected target, that have the relevant relations to incoming target words in our simulations.

So although this demonstration does not discredit the validity of Federmeier and Kutas's account, it does illustrate a genuine alternative.

It is also important to clarify that our claim is not that our averaging procedure—and the representation that it produces—is an appropriate reflection of the full extent of language processing. We are, however, positing that less structured representations of this kind are likely to underlie the N400 under some circumstances. As discussed above, many aspects of language processing that we know, a priori, will be overlooked by this averaging process, are also aspects of language processing that we have seen the N400 at times to be insensitive to: for instance, this averaging process will not encode agent/recipient information, and it will also fail to capture effects of negation. Such selective insensitivities are in line with N400 studies cited above. It seems not unreasonable, therefore, to suppose that this simple averaging procedure may be approximating a real representational stage tapped into by the N400.

Using the N400 as a probe into online language processing, our results suggest that VSMs are well positioned to capture elements of language interpretation that are driven by lexical semantic relatedness effects. A question that arises now is whether VSMs can also help us to model the more structured compositional processes that seem to underlie the N400 when it does track cloze probability. Structured semantic composition with VSMs is an active area of current research (e.g., Mitchell & Lapata, 2008; Socher, Huval, Manning, & Ng, 2012; Fyshe, Wehbe, Talukdar, Murphy, & Mitchell, 2015), and as progress continues in this area, it will be interesting to investigate whether the influences of more structured context representations can also be captured through VSMs. Other interesting questions will include whether these results extend to other types of VSMs, or to different approaches to semantic similarity, such as manual feature generation.

Acknowledgments This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1322106. This work benefited from helpful comments of three anonymous reviewers, and discussions with Tal Linzen as well as audiences and visitors across the University of Maryland language science community.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 238–247).

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*(4), 343–355.

Brown, C., & Hagoort, P. (1993). The processing nature of the n400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44.

Chow, W.-Y., Smith, C., Lau, E., & Phillips, C. (2015). A 'bag-of-arguments' mechanism for initial verb predictions. *Language, Cognition, and Neuroscience*.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (wac-4) Can we beat Google* (pp. 47–54).

Fischler, I., Childers, D. G., Achariyapaopan, T., & Perry, N. W. (1985). Brain potentials during sentence verification: Automatic aspects of comprehension. *Biological Psychology*, *21*(2), 83–105.

Fyshe, A., Wehbe, L., Talukdar, P. P., Murphy, B., & Mitchell, T. M. (2015). A compositional and interpretable semantic space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Herdağdelen, A., Erk, K., & Baroni, M. (2009). Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 50–53).

Holcomb, P. J. (1988). Automatic and attentional processing: An event-related brain potential analysis of semantic priming. *Brain and Language*, *35*(1), 66–85.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Kutas, M., & Hillyard, S. (1989). An electrophysiological probe of incidental semantic association. *Cognitive Neuroscience, Journal of*, *1*(1), 38–49.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*.

Kutas, M., Lindamood, T. E., & Hillyard, S. A. (1984). Word expectancy and event-related brain potentials during sentence processing. In S. Kornblum & J. Requin (Eds.), *Preparatory states and processes* (pp. 217–237).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Lapesa, G., & Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013)* (pp. 66–74).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Mandera, P., Keuleers, E., & Brysbaert, M. (2016). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (p. 17).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Acl* (pp. 236–244).

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Parviz, M., Johnson, M., Johnson, B., & Brock, J. (2011). Using language models and latent semantic analysis to characterise the N400m neural response. In *Proceedings of the Australasian Language Technology Association Workshop 2011* (pp. 38–46).

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1201–1211).

Grammatical gender affects odor cognition

Laura J. Speed (l.speed@let.ru.nl)[1] **Asifa Majid (asifa.majid@let.ru.nl)[1,2,3]**
[1]Centre for Language Studies, Radboud University, Nijmegen, NL
[2]Max Planck Institute for Psycholinguistics, Nijmegen, NL
[3]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL

Abstract

Language interacts with olfaction in exceptional ways. Olfaction is believed to be weakly linked with language, as demonstrated by our poor odor naming ability, yet olfaction seems to be particularly susceptible to linguistic descriptions. We tested the boundaries of the influence of language on olfaction by focusing on a non-lexical aspect of language (grammatical gender). We manipulated the grammatical gender of fragrance descriptions to test whether the congruence with fragrance gender would affect the way fragrances were perceived and remembered. Native French and German speakers read descriptions of fragrances containing ingredients with feminine or masculine grammatical gender, and then smelled masculine or feminine fragrances and rated them on a number of dimensions (e.g., pleasantness). Participants then completed an odor recognition test. Fragrances were remembered better when presented with descriptions whose grammatical gender matched the gender of the fragrance. Overall, results suggest grammatical manipulations of odor descriptions can affect odor cognition.

Keywords: olfaction; odor memory; grammatical gender; linguistic relativity; French; German

Introduction

"*…a warm passionate fragrance that combines the uniquely liberating notes of crispy apple and white florals with a vanilla and sandalwood heart. They fuse to create a warm, inviting and free-spirited scent, that you can wear with passion*."

Boss Orange Woman – The Perfume Shop

Choosing a personal fragrance can be a difficult undertaking. But how much of what we perceive about a fragrance is based on smell alone? Does the language used in advertisements, product descriptions, and magazine reviews influence us? With over five hundred new fragrance launches every year, sniffing fragrances individually would be an arduous and tiring task. Instead, reading fragrance descriptions can be the most efficient means by which to make a decision. But what do we imagine when we read something like: "*white florals with a vanilla and sandalwood heart*"?

Research has shown we are, in fact, poor at imagining odors (Crowder & Schab, 1995). Moreover, we are bad at identifying and naming odors. It is estimated that we can correctly name only around 50% of common odors, such as coffee or peanut butter (e.g., Cain, 1979; de Wijk, Schab, & Cain, 1995).

There are a number of possible explanations for these facts. One theory claims the olfactory cortex is poorly linked with semantic and linguistic information in the brain (Olofsson & Gottfried, 2015). Olfactory information is processed by fewer channels than other sensory domains, and is argued to be linked with lexical information more directly than modalities such as vision (Olofsson & Gottfried, 2015). This may mean olfactory information is less elaborated by the time lexical retrieval occurs, making it hard to link an odor with a name. Another possibility is the difficulties with olfaction are a product of cultural experience (Majid, 2015). In the West, we lack experience attending to and talking about smells (e.g., San Roque et al., 2015). However, speakers of some languages, such as the Jahai in the Malay Peninsula, are, in fact, just as good talking about smells as they are talking about colors (Majid & Burenhult, 2014). For people in such cultures, odor is an integral part of their daily lives, featuring in their cultural practices and belief system (Burenhult & Majid, 2011).

Research has shown that because of the limitations in thinking and talking about odors, verbal labels and descriptions can easily influence how odors are perceived (Herz, 2003; 2005). This is comparable to the proposal that language is more powerful at influencing thought for more abstract domains, such as time (e.g., Boroditsky, Schmidt, & Phillips, 2003). That is, similar to time, odor is a domain that can be difficult to conceptualize and verbalize, and so is a modality in which language has a strong influence. In fact, Herz (2003) suggests olfaction should be influenced by language more than other perceptual modalities because we cannot see odors, we cannot easily spatially locate them, nor can we easily identify them. So, instead we search for any other contextual information (such as language) to inform odor perception. In sum, it has been suggested that because conceptual representations of odors are weak, they can more easily be shaped by other sources of information, such as words.

Hedonic ratings of odors, for example, differ when odors are given verbal labels compared to when they are presented alone (Herz, 2003). De Araujo, Rolls, Velazsco, Margot, and Cayeuk (2005), and Herz and Clef (2001) found odors were rated as more pleasant when they were labeled with positive (e.g., *cheese*) instead of negative terms (e.g., *body odor*). Verbal labels modulated regions of the brain thought to be activated by odor pleasantness, suggesting the labels affected perception of the pleasantness of the odor rather than simply biasing pleasantness ratings (de Araujo et al., 2005). Similarly, Zellner, McGarry, Mattern-McClory, and Abreu (2008) found explicitly labeling unisex fragrances as

male (or *female*) made participants perceive the fragrance as more masculine (or feminine) (measured by colors matched to fragrances). These effects have been described as "olfactory illusions" (Herz, 2003; 2005).

Previous studies examining effects of language on olfaction have used explicit labels, so it is possible participants in these studies were strategically using the linguistic information they were given. So, although de Araujo et al., (2005) find effects of labels on olfactory pleasantness in the brain, the effects could still be the result of top-down integration of explicit semantic information with an ambiguous olfactory percept. The current study aims to investigate the influence of language on odor cognition by moving away from explicit semantic information to implicit semantic cues provided through grammatical gender. By focusing on grammatical cues we can address, in a novel manner, the extent to which language can affect odor cognition. Specifically, we test whether the grammatical gender of descriptions of fragrances affects how a fragrance is perceived and remembered.

Grammatical gender divides nouns into classes according to the behavior of associated words (e.g., articles, adjectives; cf. Corbett, 2006). In some languages, nouns possess a gender based on natural gender, or "sex", i.e., masculine, feminine. Grammatical gender is typically semantically arbitrary for objects without a natural gender (that is, there is nothing inherently masculine or feminine about the objects to which grammatical gender is assigned). Moreover, grammatical gender of nouns can vary across languages (for example *apple* is masculine in German, *der Apfel*, and feminine in French, *la pomme*). Despite this apparent arbitrariness of gender assignment to nouns, grammatical gender has been shown to affect how speakers of such languages think about objects. For example, Spanish and German speakers are more likely to ascribe masculine qualities to grammatically male objects and feminine qualities to grammatically female objects: German speakers described a *key*, which has masculine grammatical gender in German, using terms such as "*hard, heavy, jagged, metal, serrated and useful*", whereas Spanish speakers, for which the grammatical gender is female, instead used terms such as "*golden intricate, little, lovely, shiny and tiny*" (Boroditsky et al., 2003).

Effects of grammatical gender have been found in tasks that do not explicitly promote the use of grammatical categories, suggesting grammatical gender information is accessed automatically and implicitly (Boutonnet, Athanasopoulos, & Thierry, 2012). Using ERPs during a semantic categorization task of pictures, grammatical gender consistency affected LAN amplitude, an ERP marker of morphosyntactic processing, in Spanish-English bilinguals, but not English monolinguals (Boutonnet et al., 2012). This effect was found within an all English context, suggesting grammatical gender information can be accessed automatically and implicitly (although see, e.g., Vigliocco, Vinson, Paganelli, & Dworzynski, 2005).

This study builds on previous work in two fundamental ways. First, we test a non-lexical verbal manipulation on odor: i.e., grammatical gender. Second, we test the effects of grammatical gender on thought in a new way. Instead of explicitly judging the referent objects of nouns (e.g., Boroditsky et al., 2003), here participants judged odors associated with nouns possessing male or female grammatical gender.

We gave French and German speakers the same perfumes and the same descriptors, differing only in their grammatical gender: if the descriptors were masculine in one group, they were feminine in the other. Participants read the descriptions of odors (with masculine vs. feminine nouns), then smelled masculine and feminine fragrances, and rated the fragrance on a number of dimensions. After that, participants completed a recognition test for the fragrances they had smelled. Note, participants were never explicitly told whether the perfumes were masculine or feminine (cf., Zellner et al., 2008). This information was implicitly conveyed through the nouns by virtue of their grammatical class. We predicted the perception and memory of the fragrances would be affected by the congruence between grammatical gender of the nouns used to describe the fragrance and the gender of the fragrance.

Method

Participants

30 native German speakers (21 female; average age 26.9, SD = 9.9) and 31 native French speakers (20 female; average age 31.2, SD = 12.8) participated in the experiment.

Material

Eight fragrances were used, four marketed as masculine scents and four marketed as feminine scents. Fragrances were selected according to online "bestseller" lists in Germany and France (see Table 1). In addition, a further eight fragrances (four masculine, four feminine) were selected to be used as distractors in the recognition test at the end of the experiment. To present each fragrance, plastic pellets were sprayed with a small amount of a fragrance and then placed inside a squeezy bottle.

Eight fragrance descriptions were used (four female grammatical gender, four with male grammatical gender) (see Table 2). Each description contained three nouns of the same grammatical gender. Nouns were selected so that their grammatical genders were different in German and French. For example, one set of ingredients was *pumpkin*, *sage*, *marjoram*, with all ingredients masculine nouns in German (*Kürbis, Salbei, Marjoran*), but feminine nouns in French (*citrouille, sauge, marjolaine*).

Each fragrance was paired once with a grammatically female description and once with a grammatically male description, distributed across two experimental lists.

Table 1. Male and female fragrances

Fragrance	Gender
Hugo Boss - Boss Orange	*f*
Armani - Si	*f*
Calvin Klein - Eternity	*f*
Dior - J'adore	*f*
Chanel - Bleu de Chanel	*m*
Joop! Homme	*m*
Davidoff- Cool Water	*m*
Hugo Boss - Boss Bottled	*m*

Table 2. Masculine (*m*) and feminine (*f*) fragrance descriptions.

German		French	
Kürbis, Salbei, Marjoran	*m*	citrouille, sauge, marjolaine	*f*
Apfel, Rhabarber, Kardamom	*m*	pomme, rhubarbe, cardamom	*f*
Zitrone, Sonneblume, Melone	*f*	citron, tournesol, melon	*m*
Iris, Ringelblume, Makrone	*f*	iris, souci, macaron	*m*
Muskat, Farn, Lehm	*m*	muscade, fougère, argile	*f*
Schiefer, Lavendel, Zimt	*m*	ardoise, lavande, canelle	*f*
Gewürznelke, Kiefer, Seife	*f*	girofle, pin, savon	*m*
Eiche, Magnolie, Zeder	*f*	chêne, magnolia, cèdre	*m*

Procedure

E-Prime (Version 2.0) was used to present written fragrance descriptions and collect participants' responses. Participants were instructed to carefully read the description of each fragrance that included key ingredients. To make sure participants paid attention to the descriptions, they were told they would need to remember the fragrances and descriptions at the end of the experiment. When they had read a description, the experimenter placed the squeezy bottle beneath the participants' nose and instructed them to smell as it was squeezed. The bottle was squeezed three times with a gap of around four seconds between each squeeze. Order of fragrance presentation was randomized. After smelling the fragrance, participants pressed the space bar on the keyboard to continue to ratings of the fragrance. Participants were then asked to rate the fragrance in terms of pleasantness, intensity, how likely they would be to buy the fragrance for their father or brother, their mother or sister, how much they would be willing to pay for the fragrance, and how clearly they could smell the ingredients in the fragrance. Ratings were made on a scale of 0 to 100, and participants responded by moving a mouse along a scale and clicking. After completing the five ratings, a 2000ms blank screen was presented before the next trial.

After all fragrances had been rated, participants were told they must complete a recognition test. They smelled sixteen fragrances, half they had smelled before, and half were new. As before, the squeezy bottles were placed beneath the participants' nose and they were asked to smell when the bottle was squeezed by the experimenter. If the fragrance was new, participants were told to click in a box labeled "new", but if the fragrance had been smelled previously they were to click in a box labeled "old". Each box turned red when a response had been made.

Results

Linear mixed effect models in R (R Core Team, 2013), using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014), were conducted on the rating scores for each rating question separately, and for accuracy in the recognition test (proportion of fragrances correctly recognized as "old"). Fragrance gender (male vs. female), grammatical gender (male vs. female), language (German vs. French) and the interaction were entered as fixed factors and fragrance item and participant were random factors.[1] We predicted ratings and memory of the fragrances would be different when the gender of the fragrance matched the grammatical gender of the descriptors compared to when they did not match. For brevity sake, we only report significant effects.

As would be expected, participants indicated they were more likely to buy a female fragrance than a male fragrance for their mother or sister ($t = 2.62$, $p < .01$), and vice versa more likely to buy a male fragrance for their father or brother ($t = 5.27$, $p < .001$). Overall, male fragrances were rated as more pleasant than female fragrances ($t = 1.95$, $p = .05$).

For ratings of ingredient clarity there was a significant interaction between fragrance gender and language ($t = 2.8$, $p < .01$) such that ratings were higher for French descriptions than German descriptions for female, but not male fragrances. There was also a significant three-way interaction between grammatical gender, fragrance and language ($t = 2.64$, $p < .01$). This three-way interaction reflects a significant interaction between grammatical gender and language for male fragrances ($t = 4.29$, $p < .001$), but not for female fragrances ($t = .97$, $p = .33$). Participants indicated they could perceive the ingredients in male fragrances more clearly with masculine descriptions in French, but with feminine description in German (see Figure 1). There were no other effects in the ratings of fragrances.

[1] A separate set of participants smelled each fragrance and judged whether they thought it was for a man or woman. Based on these ratings Joop was classified as a female fragrance instead of a male fragrance.

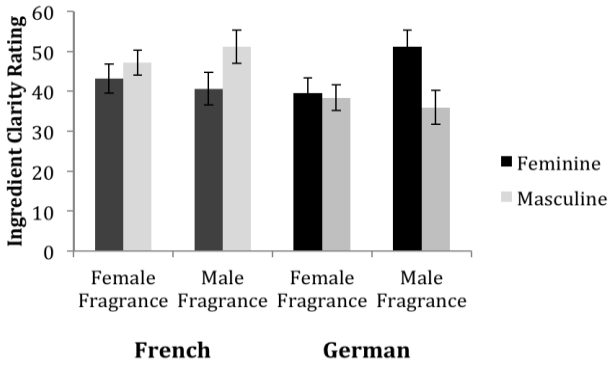


Figure 1: Fragrance ratings for question "How clearly can you perceive the ingredients in the fragrance?"

Importantly, and in line with our predictions, we found an interaction between fragrance gender and grammatical gender in memory performance (t = 2.0, p = .05). Participants were more likely to correctly recognize a fragrance when the gender of the fragrance matched the grammatical gender of the description than when it did not match (see Figure 2). There were no further significant effects in memory.

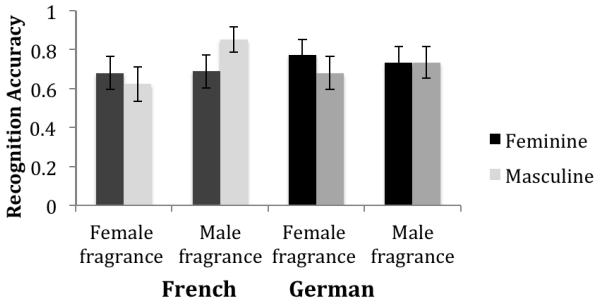


Figure 2: Recognition accuracy for male and female fragrances described with masculine and feminine nouns in French and German

Discussion

We found memory for male and female fragrances was greater when the fragrance was described using nouns with grammatical gender matching the fragrance gender. This finding parallels previous work examining gender information during sentence processing, showing processing is enhanced for gender congruent information compared to gender incongruent or neutral information (Friederici & Jacobsen 1999; Guillelmon & Grosjean, 2001). Similarly, Boroditsky and Schmidt (2000) found memory for the sex of proper names given to objects (e.g., Erica) was higher when the sex matched the grammatical gender of the object.

What could explain these effects? There are, at least, three possible mechanisms to consider. Grammatical gender information from the descriptions could be combined with gender information from the fragrance, so that the more similar the information, the stronger the memory trace formed. Alternatively, grammatical gender information could have primed a specific gender construal, making it easier to subsequently process information matching in gender – meaning fragrances of the same gender were more easily processed, and hence remembered better. Similar findings have, however, been explained as inhibition in processing when there is incongruent gender information (e.g., Jakubowicz & Faussart, 1998). That is, information is more difficult to process, or the memory trace is interfered with, when gender information is mismatching.

Odors are difficult to name and identify: we cannot see them and we have difficulty localizing them in space (Engen, 1982). We therefore rely more heavily on external context to extract meaning (Herz, 2003). Herz (2003) proposes there may be a dual-coding system for olfaction (similar to Paivo's (1971) original dual-coding theory) in which olfactory perception is sensitive to both verbal context and sensory experience. Thus, verbal labels attributed to odors can be a crucial factor in odor interpretation. Previous studies have concluded that odor memory is improved with the addition of verbal labels in general (e.g., Rabin & Cain, 1984; Lyman & McDaniel, 1986) because they provide extra retrieval paths (Lyman & McDaniel, 1990), for example. Here we show the type of label is important for memory, with memory being enhanced only when features of the verbal label match odor features.

We used fragrances in this study, which contain a mixture of scents. Odor perception is a configural process (Thomas-Danguin et al., 2014), with little access to constituent parts of an odor. Perceiving all individual ingredients within a fragrance is thus almost impossible (our ability to perceive odor constituents in a mixture is limited; Laing & Francis, 1989). Fragrances can be considered complex, and so we have tested odor cognition at its most vulnerable. Whether or not grammatical gender information could influence perception and memory for simpler or more familiar odors with more clearly identifiable sources is an open question.

Evidence from other perceptual domains suggests language is more likely to influence perception when perception is difficult. Results from Pavan, Skujevskis, and Baggio (2013) support the view that semantic information is more likely to interact with perception when the sensory signal is reduced or the task is more difficult. In a direction discrimination task, listening to direction verbs affected perceptual sensitivity when the visual stimuli were presented at threshold, but not when presented at suprathreshold. Similar findings are found in speech perception: watching a speaker's lip movements enhances speech comprehension, particularly in noisy environments (e.g., Ma, Zhou, Ross, Foxe, & Parra, 2009). Again, since odor cognition is more difficult to conceptualize than the other senses, we could expect language to have its maximal influence here (cf., Herz, 2003), and even more so for complex mixtures of odors.

We found an interaction between fragrance gender, grammatical gender and language in ratings of ingredient

clarity such that participants perceived the ingredients in male fragrances more clearly with masculine descriptions in French, but with feminine descriptions in German. We primarily included this rating question to encourage people to carefully read the fragrance description, as it would be possible to complete the other ratings based on the smell of the fragrance alone. Since this is the only rating specifically requiring participants to combine descriptions with the odor, it is not surprising this is where we see effects of language. However, the results are puzzling. Why would grammatical gender behave differently in French and German?

One possibility is the effect is not driven by grammatical gender, but another aspect of the words used, such as conceptual gender. That is, people may have masculine or feminine associations to the objects themselves. It is known that people "genderize", or assign conceptual gender to, objects (Yorkston & de Mello, 2005). Certain objects are associated more with maleness and potency, and others femaleness and beauty (Foundalis, 2002). For example, Sera, Berge, and Pintado (1994) found English speakers (for whom there is no grammatical gender) consistently judged natural objects as female, and artificial objects as male. In the present experiment, it is possible *slate*, for example, had more masculine associations than *magnolia*. If one set of words in our study had systematic masculine/feminine conceptual associations, then their congruence with the gender of the fragrance could affect responses. It would therefore be important in future work to carefully control stimuli such that grammatical and conceptual gender are manipulated orthogonally.

An alternative explanation for the difference between French and German is related to the transparency of the gender systems (e.g., Sera et al., 2002; Vigliocco et al., 2005). French has only two grammatical genders: masculine and feminine, but German has three: masculine, feminine and neuter. Famously, the mapping of natural gender to grammatical gender in German is not clear-cut. So although *woman* is feminine, as would be expected, *girl* and *wife* are neuter grammatical gender (Twain, 1880). In French grammatical gender information is widespread in every utterance: articles, nouns, and adjectives all carry morphological information about gender. But in German marking of grammatical gender is more haphazard, so not all indefinite articles and adjectives carry gender information. In addition, case interacts with gender in complex ways in German, but not French. If grammatical gender is learned by noticing the relationship between natural gender and grammatical gender, then an inconsistent relationship between natural and grammatical gender within a language would lead to weak effects of grammatical gender. This idea is supported by the fact that systematic and robust grammatical gender effects have been found in Romance languages (e.g., French and Spanish), but not in German (Sera et al., 2002; Vigliocco et al., 2005).

We found participants preferred female fragrances for female relatives and male fragrances for male relatives. This suggests the gender of a fragrance comes to mind readily when smelling it. Fragrances are typically marketed as male or female, thus over time associations are learned between certain types of odors and natural gender. So, although correctly identifying the source of an odor is difficult (e.g., Cain, 1979; de Wijk et al. 1995), gender may be one of the dimensions along which odors can be successfully described, along with pleasantness – at least for fragrances (cf., Yeshurun & Sobel, 2010).

Our results have further implications for marketing. We have shown information from language can be combined with information from odors, subsequently enhancing encoding and recognition. This is important from a marketing perspective: it is no good a fragrance being pleasantly perceived if the product itself cannot be remembered. When odor identification is weak, it becomes particularly vulnerable to contextual information, making it the ideal venue to use interesting marketing devices and ploys.

In sum, we have shown grammatical information can affect how odors are perceived and remembered. Thus, not only is odor cognition vulnerable to explicit semantic labels, it can also be affected by more subtle linguistic manipulations, such as grammatical gender.

Acknowledgments

This work was funded by The Netherlands Organization for Scientific Research: NWO VICI grant "Human olfaction at the intersection of language, culture and biology". Thanks to Annetta Kopecka, Miklos Dorsche, Nina Krijnen, Alice Reinhartz, Patricia Manko and Julia Misersky for help with stimuli creation, translations and testing.

References

de Araujo, I. E., Rolls, E. T., Velazco, M. I., Margot, C., & Cayeux, I. (2005). Cognitive Modulation of Olfactory Processing. *Neuron*, *46*(4), 671–679.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7).

Boroditsky, L., & Schmidt, L. A., (2000). Sex, Syntax, and Semantics. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. Philadelphia, PA.

Boroditsky, L., Schmidt, L., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.) *Language in mind: Advances in the study of language and thought* (pp. 61–80). Cambridge, MA: MIT Press.

Boutonnet, B., Athanasopoulos, P., & Thierry, G. (2012). Unconscious effects of grammatical gender during object categorisation. *Brain Research*, *1479*, 72–79.

Burenhult, N., & Majid, A. (2011). Olfaction in Aslian ideology and language. *The Senses & Society*, *6*(1), 19–29.

Cain, W. S. (1979). To know with the nose: Keys to odor identification. *Science*, *203* (4379) 467–470.

Corbett, G. G. (2006). Gender, grammatical. *The Encyclopedia of language and linguistics.* (pp.749–756.) 2nd Edition. Oxford: Elsevier

Crowder, R. G., & Schab, F. R. (1995). Imagery for odors. In R. G., Crowder, & F. R. Schab (Eds.) *Memory for odors* (pp. 93-108). Hillsdale, NJ: Erlbaum

Foundalis, H. E. (2002). Evolution of gender in Indo-European languages. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Fairfax, VA

Friederici, A. D., & Jacobsen, T. (1999). Processing grammatical gender during language comprehension. *Journal of Psycholinguistic Research*, *28*(5), 467–484.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition*, *29*(3), 503–511.

Herz, R. S. (2003). The effect of verbal context on olfactory perception. *Journal of Experimental Psychology: General*, *132*(4), 595–606.

Herz, R. S. (2005). The unique interaction between language and olfactory perception and cognition. In Rosen, D. T. (Ed.) *Trends in experimental research* (pp 91–99). London: Nova Publishing.

Herz, R. S., & Clef, J. von. (2001). The influence of verbal labeling on the perception of odors: Evidence for olfactory illusions? *Perception*, *30*(3), 381 – 391.

Jakubowicz, C., & Faussart, C. (1998). Gender agreement in the processing of spoken French. *Journal of Psycholinguistic Research*, *27*(6), 597–617.

Laing, D. G., & Francis, G. W. (1989). The capacity of humans to identify odors in mixtures. *Physiology & Behavior*, *46*(5), 809–814.

Lyman, B. J., & McDaniel, M. A. (1986). Effects of encoding strategy on long-term memory for odours. *The Quarterly Journal of Experimental Psychology*, *38*(4), 753–765.

Lyman, B. J., & McDaniel, M. A. (1990). Memory for odors and odor names: Modalities of elaboration and imagery. *Journal of Experimental Psychology*, *16*(4), 656–664.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using highdimensional feature space. *PloS One*, *4(3)*, e4638.

Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, *130*(2), 266–270.

Majid, A. (2015). Cultural factors shape olfactory language. *Trends in Cognitive Sciences*, *19*(11), 629–630.

Olofsson, J. K., & Gottfried, J. A. (2015). The muted sense: neurocognitive limitations of olfactory language. *Trends in Cognitive Sciences*, *19* (6), 314-321

Paivio, A (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Pavan, A., Skujevskis, M., & Baggio, G. (2013). Motion words selectively modulate direction discrimination sensitivity for threshold motion. *Frontiers in Human Neuroscience*, *7*, (134).

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rabin, M. D., & Cain, W. S. (1984). Odor recognition: Familiarity, identifiability, and encoding consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 316–325.

San Roque, L., Kendrick, K. H., Norcliffe, E. J., Brown, P., Defina, R., Dingemanse, M., … Majid, A. (2015). Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistics*, *26*(1), 31–60.

Sera, M. D., Berge, C. A. H., & Pintado, J. C. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development*, *9*(3), 261–292.

Thomas-Danguin, T., Sinding, C., Romagny, S., El Mountassir, F., Atanasova, B., Le Berre, E., Le Bon, A., Coureaud, G. (2014). The perception of odor objects in everyday life: a review on the processing of odor mixtures. *Frontiers in Psychology*, *5*.

Twain, M. (1880). *A tramp abroad*. Leipzig: Bernhard Tauchnitz.

Vigliocco, G., Vinson, D. P., Paganelli, F., & Dworzynski, K. (2005). Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General*, *134*(4), 501–520.

de Wijk, R.A., Schab, F.R. & Cain, W.S. (1995) Odor identification. In R. G., Crowder, & F. R., Schab (Eds). *Memory for odors.* Hillsdale, NJ: Erlbaum 21-37.

Yeshurun, Y., & Sobel, N. (2010). An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. *Annual Review of Psychology*, *61*, 219–241.

Yorkston, E., & de Mello, G. E. (2005). Linguistic gender marking and categorization. *Journal of Consumer Research*, *32*(2), 224–234.

Zellner, D. A., McGarry, A., Mattern-McClory, R., & Abreu, D. (2007). Masculinity/femininity of fine fragrances affects color-odor correspondences: A case for cognitions influencing cross-modal correspondences. *Chemical Senses*, *33*(2), 211–222.

How Different Frames of Reference Interact: A Neural Network Model

Weizhi Nan (nanwz@psych.ac.cn)[a],
Yanlong Sun (ysun@tamhsc.edu)[b],
Xun Liu (liux@psych.ac.cn)[a],
Hongbin Wang (hwang@tamhsc.edu)[b]
[a] Key Laboratory of Behavioral Science, Institute of Psychology,
Chinese Academy of Sciences, 16 Lincui Road, Chaoyang District, Beijing 100101, China
[b] Center for Biomedical Informatics, Texas A&M University Health Science Center,
2121 West Holcombe Blvd, Houston, TX 77030, USA

Abstract

It has been argued that people use multiple frames of reference (FORs) for representing and updating spatial relationships between objects in a complex environment. When there are conflicts among representations of multiple FORs, they compete to determine behavior. "Frame of Reference-based Map of Salience" theory (FORMS) suggests that FORs with high salience may be processed in priority. Here, we report a computational neural network model for a two-cannon task, which naturally involves multiple FORs with different levels of salience: intrinsic frame of reference (IFOR) and egocentric frame of reference (EFOR). The goal is to investigate the computational neural mechanisms underlying human spatial performance. Our simulation results fit earlier behavioral results well. The model suggests although multiple FORs may be initially represented independently, they interfere with each other by the inhibitory competition of neurons in the later process (in hidden layer) for conflict resolution. Moreover, salience may modulate the competition by prioritizing FORs with high salience levels. These results represent a connectionist support for the FORMS theory.

Keywords: frame of reference; inhibitory competition; salience; neural network model

Introduction

People adopt multiple frames of reference (FORs) to represent the spatial relationship of objects in a complex environment (Klatzky, 1998; Mou & McNamara, 2002; Piaget & Inhelder, 1956; Sun & Wang, 2014; Tamborello, Sun, & Wang, 2012; Wang, Johnson, & Zhang, 2001; Zacks & Michelon, 2005). Based on the relationship with the observer, FORs can be classified into three types, egocentric FOR (EFOR), intrinsic FOR (IFOR) and allocentric FOR (AFOR) (Klatzky, 1998; Mou & McNamara, 2002; Tamborello et al., 2012; Wang & Spelke, 2002). An EFOR-based representation is anchored to the observer, which needs to be updated following the movement of the observer's eye, head, body coordinates (Wang et al., 2001). In an IFOR-based representation, an object or an object group in the viewing environment but exogenous to the observer is used as the reference point. For example, a table is used as an IFOR anchor in the description "the cup is on the table". IFORs remain stable with the observer's movement but have to be updated when the reference object moves. In an AFOR-based representation, the entire environment, such as a room or a city, is taken as the reference point. For a comprehensive review, see (Mou, Fan, McNamara, & Owen, 2008; Mou & McNamara, 2002; Sun & Wang, 2010, 2014; Tamborello et al., 2012; Yamamoto & Philbeck, 2013).

Particularly of our interests is how multiple FOR representations develop and interact with each other. Mathematically, all FORs are equivalent. Depending on specific situations, however, some FORs are more useful or convenient. In comparison, EFOR is probably more automatic and almost effortless, AFOR is quite stable but computationally demanding, and IFOR is a balance between flexibility and stability. "Frame of Reference-based Map of Salience" theory (FORMS) states the human brain represents spatial information simultaneously using multiple FORs, each being a spatial map of salience (i.e. only salient objects or locations are represented on each map), and that human performance is determined by the interaction of all relevant FOR-based representations (Sun & Wang, 2010, 2014; Tamborello et al., 2012; Wang et al., 2001; Wang, Sun, Johnson, & Yuan, 2005; Yamamoto & Philbeck, 2013).

To understand the interaction among multiple FORs and the effect of salience, Tamborello et al. (2012) designed a two-cannon task (a modified and simplified version is shown in Figure 1), in which two cannons (a red and a blue) were surrounded by 8 pellets in red or blue color. The salience of the cannons was determined by the pellet color ratio such that the cannon with more same-color pellets was more salient. In the task, when one of the pellets was randomly chosen as the target pellet (flashing), the cannon in the same color became the target cannon. Participants were asked to rotate the target cannon to the location of the flashing target pellet in the shortest way by pressing either the left-arrow or the right-arrow button on the keyboard. An analysis (Figure 3A and 3B) showed that reaction time (RT) became longer when there was a conflict among different FOR-based representations. RT in cannon angle 180° condition was longer than RT in cannon angle 0° condition (Cannon Angle Effect). In the cannon angle 180° condition, RT for trials where the target pellet appeared in the minority color group was longer than RT for those in the majority color group (Salience Effect). RT in target cannon point-down condition was longer than RT in target cannon point-up condition (Target Cannon Orientation Effect). And the target cannon orientation effect was larger in cannon angle 180° condition than that in cannon angle 0° condition,

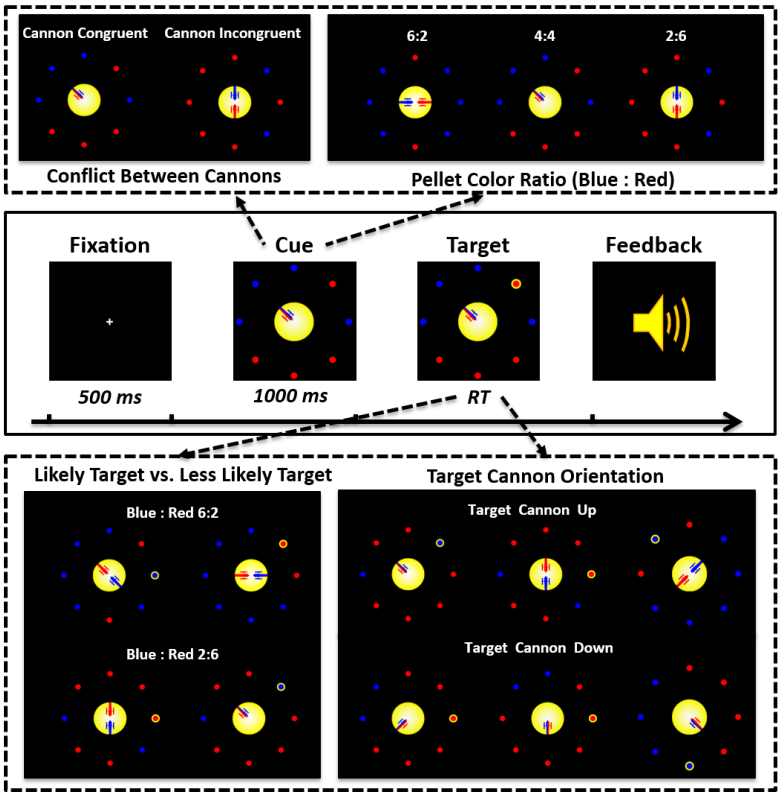


Figure 1. A schematic illustration of the two-cannon task. Middle: trial procedure. At the beginning of each trial, a fixation cross was presented at the center of the screen for 500 ms, and subjects were asked to fixate on it, then two cannons (one red and one blue) and eight pellets (in either red or blue) were presented together on the computer screen. After a one-second pause, a randomly selected pellet would flash as the target pellet. The participants' task was to use the arrow keys to rotate the cannon in the same color of the target pellet toward the target pellet as quickly as possible. Auditory feedback was played for either the correct (a shot sound) or incorrect (an alarm sound) response; top left: cannon congruent (0°) and cannon incongruent (180°); top right: pellet color ratio (blue : red = 6:2, 4:4, and 2:6); bottom left: likely target pellet (target occurs in the majority color pellets) and less likely target pellet (target occurs in the minority color pellets); bottom right: target cannon orientation up (12 o'clock is 0°, clockwise rotation, set 315°, 0°, and 45° as the target cannon up condition) and down (set 135°, 180°, and 225° as the target cannon down condition).

Table 1. The training data for the model. Each row represents a group of trials that covers different FORs with different orientations and salience, different target pellet color. The first three columns ("FOR number", "Cannon Angle" and "Pellet Ratio B:R") represent different conditions; Column "Weights" represents relative frequency; Column "Types" represents how many trial types are contained in each group; the last ten columns represent specific input information, with salience level ranging from 0.2 to 0.8, and cannon orientation or target location as up-right (UR), up (U), down-left (DL), or down-right (DR); the last column is the response (right or left key).

FOR Number	Cannon Angle	Pellet Raito B:R	Weights	Types	Blue Cannon Orientation	Blue Cannon Salience	Red Cannon Orientation	Red Cannon Salience	EFOR Orientation	EFOR Salience	Blue Target Pellet	Target Pellet Location	Red Target Pellet	Output
3FORs	0°	2:6	1	36	UR	0.2	UR	0.8	U	0.5		DR	1	Right
3FORs	0°	2:6	1	36	UR	0.2	UR	0.8	U	0.5	1	DR		Right
3FORs	0°	4:4	1	36	UR	0.5	UR	0.5	U	0.5		DR	1	Right
3FORs	0°	4:4	1	36	UR	0.5	UR	0.5	U	0.5	1	DR		Right
3FORs	0°	6:2	1	36	UR	0.8	UR	0.2	U	0.5		DR	1	Right
3FORs	0°	6:2	1	36	UR	0.8	UR	0.2	U	0.5	1	DR		Right
3FORs	180°	2:6	3	36	UR	0.2	DL	0.8	U	0.5		DR	1	Left
3FORs	180°	2:6	1	36	UR	0.2	DL	0.8	U	0.5	1	DR		Right
3FORs	180°	4:4	1	36	UR	0.5	DL	0.5	U	0.5		DR	1	Left
3FORs	180°	4:4	1	36	UR	0.5	DL	0.5	U	0.5	1	DR		Right
3FORs	180°	6:2	1	36	UR	0.8	DL	0.2	U	0.5		DR	1	Left
3FORs	180°	6:2	3	36	UR	0.8	DL	0.2	U	0.5	1	DR		Right
EFOR			50	6					U			DR		Right

which means there was an interaction between cannon angle effect and target cannon orientation effect. These results indicate that people encounter difficulties when they have to process different conflicting FOR representations and that they seem to prioritize processing of each FOR by salience.

How the conflicts occur raises an issue. In particular, how are the conflicts among FORs represented and what is the mechanism of conflict resolution? Previously, we have hinted that spatiotemporal association and predictive learning play a major role in such tasks (Sun & Wang, 2014). In the current study, we evaluate the hypothesis by developing a neural network model of the two-cannon task. The model is implemented in Leabra (local, error-driven and associative, biologically realistic algorithm), a biologically based computational modeling framework (O'Reilly, 1998; O'Reilly, Munakata, Frank, & Hazy, 2012). Among all the features, Leabra incorporates complex network dynamics using bidirectional excitatory connections and fast pooled inhibition, which makes it an ideal candidate for exploring the interference among conflicting cognitive process.

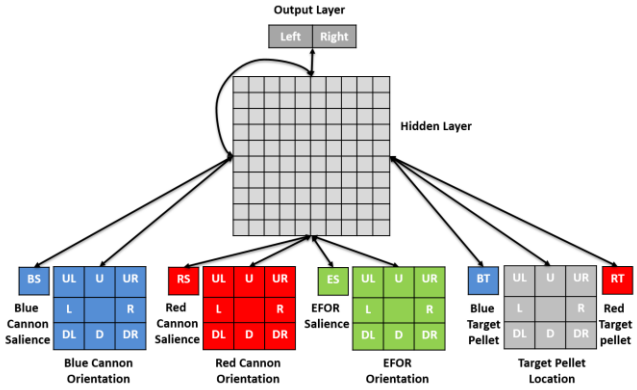


Figure 2. The model of the two-cannon task. Three 3x3 FOR orientation layers (excluding the center neuron, the other eight neurons corresponding to eight orientations: U-up, UR-up-right, R-right, DR-down-right, D-down, DL-down-left, L-left, UL-up-left) with three 1x1 salience layer; one 3x3 target pellet location layer (excluding the center neuron, the other neurons corresponding to eight locations) with two 1x1 target pellet color layer (representing the color of the target pellet); one 10x10 hidden layer; one 2x1 output layer (Left, Right).

The Neural Network Model

Model Architecture and Connectivity

The model (Figure 2) consists of three levels of layers (input, hidden and output). At the input level, there are nine input layers. Each layer encodes one piece of stimulus information independently. Three 3x3 layers are used to encode the orientations of two IFORs and one EFOR (Blue Cannon Orientation, Red Cannon Orientation and EFOR Orientation in Figure 2). Three 1x1 layers aside the three FOR input layers are used to encode their salience (Blue Cannon Salience, Red Cannon Salience, EFOR Salience). One 3x3 layer filled in gray color is used to encode the locations of target pellet (Target Pellet Location). Two 1x1 layers aside the target location input layer are used to encode the color of target pellet (Blue Target Pellet, Red Target Pellet). At the hidden level, one 10x10 layer is used to represent and process all the vision input information. At the output level, one 2x1 layer is used to encode the response.

The connections among the layers are all bi-directional. The hidden layer has a self-recurrent excitatory connection to itself. The weight of forward connection is stronger than the weight of the feedback and self-recurrent excitatory connection with a ratio of 2:1. The k-winner-take-all parameter is used in the hidden layer (the percentage, k= 25%) and the output layer (the number of neurons, k= 1), which controls the activation level of the two layers without using explicit inhibitory neurons. All other parameters used in the model take the default values of Leabra (O'Reilly et al., 2012).

Training and Testing

The training data set consists of 13 groups of trials (Table 1). The first 12 groups are used to train the three FORs-coexisting conditions and the last group is for the single EROR condition. The first 12 groups are generated based on three independent variables: cannon angle (0°, 180°), the pellet color ratio (blue: red = 2:6, 4:4, and 6:2), and target pellet color (Blue, Red). The orientations of each cannon (shown in "Blue Cannon Orientation" column and "Red Cannon Orientation" column, same orientations for cannon angle 0° condition, opposite orientations for cannon angle 180° condition) control the cannon angle (shown in "Cannon Angle" column). The pellet color ratio (shown in "Pellet Ratio B: R" column) will change the salience of each cannon (shown in "Blue Cannon Salience" column and "Red Cannon Salience" column. e.g. if blue: red= 2:6, we set 0.2 for blue salience and 0.8 for red salience). The target pellet color (shown in "Blue Target Pellet" column and "Red Target Pellet" column) controls the target pellet color of each trial and the blue/red target pellet also indicates the correct response should be made by the blue/red cannon. In each of the 12 groups, there are 36 trial types separated by the cannon orientation and target pellet location (shown in "Types" column. 8 cannon orientations and different target pellet location excluding locations same with or opposite to cannon and EFOR orientations). In order to increase the robustness of salience effect, we set the salience groups 3 times more than the other groups (shown in "Weights" column).

EFOR is always automatic activated as the point-up orientation because in behavioral studies participants ("observers") were always facing with the computer screen. When the target cannon points above the horizontal, the target cannon's IFOR matches the participant's EFOR; when it points below the horizontal, IFOR does not match EFOR, causing a conflict between IFOR and EFOR (Target Cannon Orientation Effect). Thus, we set the last group for the EFOR, and specially train the single EFOR with an only point-up orientation group. In EFOR group, there are 6 trial types (1 EFOR orientation and 6 target pellet location excluding locations same with or opposite to EFOR orientation). In order to increase the robustness of the effect, we set the weight of the EFOR group as 50. The total trial number of training data is (1*10+3*2) *36+50*6=876. And the testing data contains only three FORs coexisting groups, and each group only repeats by one time, so the total trial number of testing data is 12*36=432.

We set the maximum cycles of each trial as 200. If the cycle reaches 200, the current training trial would stop and switch to next trial. The stop criterion of the whole training is the average sum square error (SSE) reaching 0 by four times continuously. In the testing phase, when the activation of output layer reaches a set threshold (0.85) a response is said to have been made. The number of cycles is taken as the measure of model RT.

Thirty "simulated subjects", each with randomly initialized weights were trained and tested to get thirty independent simulation results.

Simulation Results

Analysis

The dependent variable is the number of computational cycles of each trial. The independent variables are cannon angle (0°, 180°), target cannon orientation (Up, Down), pellets color ratio (Blue: Red = 2:6, 4:4, 6:2), target pellet color (Red, Blue). One 2 (cannon angle) × 3 (pellet color ratio) × 2 (target pellet color) repeated-measures ANOVA was performed on the cycles to search for the main effects of cannon angle effect and the salience effect (Figure 3C. and Table 2). A 2 (cannon angle × 2 (target cannon orientation) repeated-measures ANOVA was performed on the cycles to search for the main effects of target cannon orientation effect, cannon angle effect, and their interaction (Figure 3D and Table 3). A correlation analysis was performed on the simulation results and the behavioral results (Figure 4). According to the FORMS theory and the earlier behavioral experimental results, we expected our model could learn from the training data, and the simulation results would have a positive correlation with earlier behavioral experimental results. The effects of interest are summarized below.

Cannon Angle Effect and Salience Effect

In Figure 3C and Table 2, there was a significant main effect of cannon angle effect, F (1, 29) = 719.96, p < .001, ηp2 = .96, indicating that the cycles in the cannon angle 0° condition (23.51 ± 0.07) were fewer than the cycles in the cannon angle 180° condition (26.28 ± 0.11). There was a significant interaction of pellet color ratio and target pellet color, F (2, 58) = 9.10, p < .001, ηp2 = .24. Post hoc analysis suggested that the pattern as below: the cycles of the red target pellet (24.81 ± 0.12) were marginally fewer than that of the blue target pellet (25.09 ± 0.13) in the blue: red = 2:6 condition, p = .09. In contrast, cycles of the red target pellet (25.01 ± 0.13) had a trend to be larger than that of the blue target pellet (24.77 ± 0.11) in the blue: red = 6:2 condition, p > .05. The cycles of the two target pellet colors (red target pellet: 24.89 ± 0.11, blue target pellet: 24.82 ± 0.11) showed no difference in the blue: red = 4:4 condition, p > .05. This interaction pattern only appeared significantly in the cannon angle 180° condition and not in the cannon angle 0° condition, as revealed by the significant three-way interaction among cannon angle, target pellet color, and pellet color ratio, F (2,58) = 47.84, p < .001, ηp2 = .62. Post hoc analysis in cannon angle 180° condition suggested that the pattern as below: the cycles of the red target pellet (25.96 ± 0.99) were fewer than that of the blue target pellet (26.82 ± 1.15) in the blue: red = 2:6 condition, p < .01. In contrast, cycles of the red target pellet (26.63 ± 1.06) were larger than that of the blue target pellet (25.90 ± 0.88) in the blue: red = 6:2 condition, p < .01. The cycles of the two target pellet colors (red target pellet: 26.20 ± 0.79, blue target pellet: 26.18 ± 0.91) showed no difference in the blue: red = 4:4 condition, p > .05. No other significant effects were obtained, ps > .05.

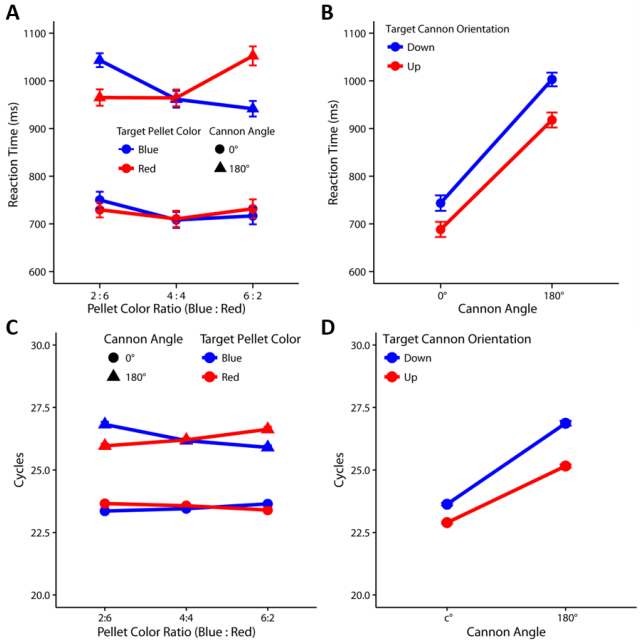


Figure 3 A. RTs for target pellet color, pellet color ratio and cannon angle in behavioral results; B. RTs for target cannon orientation and cannon angle in behavioral results; C. Cycles for target pellet color, pellet color ratio and cannon angle in simulation results; D. Cycles for target cannon orientation and cannon angle in simulation results.

Table 2: Cycles of target pellet color, pellet color ratio, and cannon angle

Cannon angle	Target pellet color	Pellet color ratio blue: red		
		2:6	4:4	6:2
180°	blue	26.82±1.15	26.18±0.91	25.90±0.88
	red	25.96±0.99	26.20±0.79	26.63±1.06
0°	blue	23.36±0.52	23.45±0.52	23.64±0.53
	red	23.66±0.50	23.57±0.60	23.40±0.51

Table 3: Cycles of target cannon orientation and cannon angle

Target cannon orientation	Cannon angle	
	0°	180°
Down	23.63±0.52	26.87±1.03
Up	22.89±0.48	25.16±0.65

Target Cannon Orientation Effect and Cannon Angle Effect

The cycles results (Figure 3D and Table 3) showed significant main effects for target cannon orientation effect

and cannon angle effect [F (1, 29) = 75.82, p < .001, ηp2 = .72; F (1, 29) = 717.79, p < .001, ηp2 = .96]. The interaction of the two factors was also significant, F (1, 29) = 25.03, p < .001, ηp2 = .46. Post hoc analysis suggested that target cannon orientation effect in the cannon angle 180° condition (1.71 ± 0.20) was larger than the effect in the cannon angle 0° condition (0.74 ± 0.13).

Correlation

A correlation analysis between the simulation results and behavioral results was conducted. For cannon angle effect and salience effect, we combined all the 12 conditions (target color × cannon angle × pellet color ratio) together to analyze the relation among these conditions. It is evident results (Figure 4A) that there was a significantly positive correlation, r = 0.53, p < .01, df = 358. For the cannon angle effect and target cannon orientation effect, we combined the 4 conditions (target cannon orientation × cannon angle) together to analyze the relation among these conditions. Results (Figure 4B) showed that there was also a significantly positive correlation, r =.56, p < .01, df = 118.

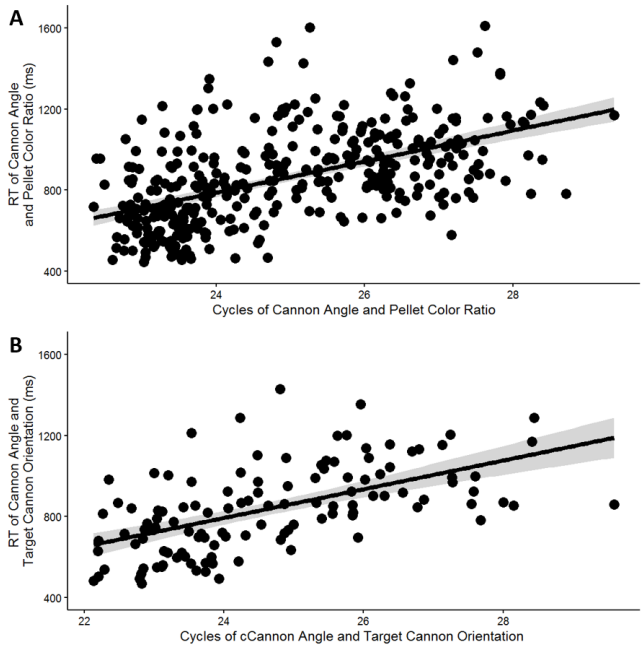


Figure 4 A. Correlation in target color, cannon angle and pellet color ratio; B. Correlation in target cannon orientation and cannon angle.

Discussion

The results from the neural network model of the two-cannon task are consistent with our earlier interpretations of the behavioral results based on the FORMS theory. The model shows stable performance and replicates all major effects with a close match to the behavioral results. Importantly, these results provide a neural basis to our theory where we can characterize the competition between different FOR-based representations by inhibitory competition among groups of neurons and predictive learning.

According to the FORMS theory, multiple representations with respect to different FORs may co-exist in a segmented but competitive fashion. The salience of any particular representation is driven by not only the perception of a static scene (e.g., the ratio between blue and red pellets) but also the prediction of the changing environment (e.g., a red cannon would be more likely to be the target cannon because there are more red pellets). This means that in a complex environment with multiple spatial relationships, representations anchored to different FORs have to be constantly maintained and updated. The simulation results from the neural network model suggest that the process of maintenance and updating can take place distributive within the same group of neurons (hidden layer), and therefore afford the possibility of interference among different FOR-based representations at the neural level.

The model showed the cannon angle effect, target cannon orientation effect and the interaction between them. The cannon angle effect was demonstrated by the fewer cycles in the cannon angle 0° condition (the orientation of the potential IFOR was the same as the orientation of the target IFOR) than in the cannon angle 180° condition (the orientation of the potential IFOR was opposite to the orientation of target IFOR). And the target cannon orientation effect was revealed by the fact that the cycles of the target cannon point-up condition (the orientation of target IFOR was the same as the orientation of EFOR) were fewer than that of the target cannon point-down condition (the orientation of target IFOR was opposite to the orientation of EFOR). These two effects suggest that in the hidden layer, there may be the same group of neurons responsible for computing the output based on the different FORs (blue cannon anchored to blue IFOR, red cannon anchored to red IFOR and observer anchored to EFOR). Therefore, when the orientation of the target IFOR was same as the orientation of the potential IFOR or EFOR, the response became faster. On the contrary, when their orientations were different, inhibitory competition occurred, and neurons took more time to focus on the target IFOR, leading to slower responses. In addition, the interaction between the two effects was revealed by the larger target cannon orientation effect in cannon angle 180° condition than that in cannon angle 0° condition. According to the adding factor theory, when two processes occur in the same phase, they will compete with each other for the cognitive resource, resulting in longer RT (Liu, Banich, Jacobson, & Tanabe, 2004; Liu, Park, Gu, & Fan, 2010; Sternberg, 1969). In this study, it is likely that some shared neurons for FOR information processing in the hidden layer give rise to these effects. The salience effect was demonstrated by fewer cycles in the likely target pellet condition than that in the less likely target pellet condition. We hypothesize that neurons in the hidden layer learned to predict the cannon with the same color of majority pellets as the target cannon.

If the prediction was right, the response would be fast. If the prediction was incorrect, it would take more time to update the target cannon, so the response would be slow. The salience level associated with each FOR might enhance the competition and lead to a prioritization of the corresponding FOR. However, it was the real-time prediction and inhibitory competition that contributed to the extra computational time for resolving potential conflicts (Sun & Wang, 2014).

The above analysis suggests that the interaction among neuron groups in the hidden layer may be responsible for the modeling results. The next step is to perform further data analysis, such as PCA and cluster analysis to evaluate these predictions. Another step is to add the cue data to the testing data set to make the model process the input information in temporal sequences. Then the predictive learning and the target decision making could be more clearly separated in the model.

Conclusion

Our neural network model replicates the behavioral results well, supporting the claim that representations with multiple FORs co-exist and compete to determine performance. Importantly, it suggests a plausible neural mechanism underlying the FORMS theory. Depending on whether there are conflicts among different FOR representations and whether the actual outcome is consistent with the expectation, competition takes place at different levels and results in the engagement and disengagement of different FOR-based representations. According to the salience effect, the internal spatial representation of the environment is always dynamically constructed and updated toward the anticipated outcomes, rather than just representing static associations of the current spatial configuration. These results support our interpretation of spatiotemporal association and predictive learning.

Acknowledgments

This work was partially supported by a Natural Science Foundation of China grant (31328013), an Air Force Office of Scientific Research (AFOSR) grant (FA9550-12-1-0457), and an Office of Naval Research (ONR) grant (N00014-16-1-2111).

References

Klatzky, R. L. (1998). *Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections.* Paper presented at the Spatial cognition.

Liu, X., Banich, M. T., Jacobson, B. L., & Tanabe, J. L. (2004). Common and distinct neural substrates of attentional control in an integrated Simon and spatial Stroop task as assessed by event-related fMRI. *Neuroimage, 22*(3), 1097-1106.

Liu, X., Park, Y., Gu, X., & Fan, J. (2010). Dimensional overlap accounts for independence and integration of stimulus-response compatibility effects. *Atten Percept Psychophys, 72*(6), 1710-1720.

Mou, W., Fan, Y., McNamara, T. P., & Owen, C. B. (2008). Intrinsic frames of reference and egocentric viewpoints in scene recognition. *Cognition, 106*(2), 750-769.

Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *J Exp Psychol Learn Mem Cogn, 28*(1), 162-170.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences, 2*(11), 455-462.

O'Reilly, R. C., Munakata, Y., Frank, M., & Hazy, T. (2012). *Computational cognitive neuroscience*: PediaPress.

Piaget, J., & Inhelder, B. (1956). The child's concept of space. *New York: Humanities Pr*.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276-315.

Sun, Y., & Wang, H. (2010). Perception of space by multiple intrinsic frames of reference. *Plos One, 5*(5), e10442.

Sun, Y., & Wang, H. (2014). Insight into others' minds: spatio-temporal representations by intrinsic frame of reference. *Frontiers in Human Neuroscience, 8(58)*, 1-11.

Tamborello, F. P., 2nd, Sun, Y., & Wang, H. (2012). Spatial reasoning with multiple intrinsic frames of reference. *Experimental Psychology, 59*(1), 3-10.

Wang, H., Johnson, T. R., & Zhang, J. (2001). *The mind's views of space.* Paper presented at the Proceedings of the Third International Conference of Cognitive Science.

Wang, H., Sun, Y., Johnson, T. R., & Yuan, Y. (2005). Prioritized spatial updating in the intrinsic frame of reference. *Spatial Cognition and Computation, 5*(1), 89-113.

Wang, R., & Spelke, E. (2002). Human spatial representation: insights from animals. *Trends in Cognitive Sciences, 6*(9), 376.

Yamamoto, N., & Philbeck, J. W. (2013). Intrinsic frames of reference in haptic spatial learning. *Cognition, 129*(2), 447-456.

Zacks, J. M., & Michelon, P. (2005). Transformations of visuospatial images. *Behav Cogn Neurosci Rev, 4*(2), 96-118.

Syntax Accommodation in Social Media Conversations

Reihane Boghrati[1], **Joe Hoover**[2], **Kate M. Johnson**[2], **Justin Garten**[1], **Morteza Dehghani**[1,2]
boghrati, jehoover,katejohn, jgarten, mdehghan@usc.edu
[1]Department of Computer Science
[2]Department of Psychology
University of Southern California
Los Angeles, CA 90089 USA

Abstract

The psycholinguistic theory of Communication Accommodation proposes that people modify communication dynamics (e.g. vocal patterns, gesture, word choice, syntax, etc.) to minimize (or maximize) their social differences. Research on communication accommodation has shown that people who want social approval will modify their linguistic style to match that of their interactant; however, most studies have been conducted on small-scale datasets and in laboratory situations. In this work, we investigate the relationship between linguistic syntactic usage and conversation participation in a more naturalistic conversational setting: social media conversations on Reddit.com. We introduce a novel approach for calculating document-level syntactic similarity by relying on natural language processing methods (parse tree generators) and graph theory techniques (minimum weight perfect matching on complete bipartite graphs). Using the proposed method, we present the results of two experiments which demonstrate that users who comment on a post tend to use syntax similar to that of the original post. Specifically, we provide evidence that comments on a post are more likely to follow the syntactic structure of the original post, compared to both random comments and also posts by the author of the comment.

Keywords: Communication Accommodation Theory; Social Media; Syntax Similarity; Linguistic Style Convergence

Introduction

Communication Accommodation Theory (CAT) is a prominent theory in psycholinguistics that targets the effects of language on peoples behavior (Giles, 2008). It proposes that people adjust features of their communication dynamics, such as vocal patterns and gestures, while interacting with others in order to maximize or minimize their social differences (Shepard, Giles, & Le Poire, 2001). This speech syntax accommodation convergence or divergence can be conscious or unconscious and is associated with a speaker's involvement in group dynamic processes, such as social integration and group identification or disidentification (Giles, Coupland, & Coupland, 1991). To some extent, CAT shares similarities with a related psycholinguistic theory, Syntax Priming Theory (Bock, 1986), which suggests that speakers tend to unconsciously replicate specific syntactical patterns in their own speech through mere exposure. For example Gries (2005) analyzed two specific syntactic structures, dative alternation and particle replacement of transitive phrasal verbs, and showed syntactic priming in a corpus-based study. In a different study, Reitter, Moore, and Keller (2010) focused on spoken-language and provided evidence for priming of syntactic rules. However, an important difference between Communication Accommodation Theory and Syntax Priming Theory is that the former posits that this process of dynamic mimicry is sensitive to social context and serves a functional social purpose.

Notably, a considerable body of research has investigated the relationship between word usage and psychologically relevant characteristics. There have been several studies that focused on function words as an indicator of linguistic style, personality, and communication (Niederhoffer & Pennebaker, 2002). For example, Danescu-Niculescu-Mizil, Gamon, and Dumais (2011) identified 14 categories of function words and demonstrated that users who are in the same conversation in Twitter, tend to use words from the same category in their tweets compared to those who are not [1]. Pennebaker (2011) emphasized the relationship between people's use of function words (i.e. style words) and a range of psychological characteristics. For instance, they found that first person pronoun use (vs "we" pronouns) is positively related with honesty. In another study, Boyd and Pennebaker (2015) showed that function words can help to identify the author of a piece of writing. These studies provide important insight into the relationship between linguistic dynamics and sociopsychological constructs.

Research has also provided evidence for the positive social outcomes of CAT when people match their word use to that of their conversation partners (Jacob, Guéguen, Martin, & Boulbry, 2011; Guéguen, 2009). For instance, Van Baaren, Holland, Steenaert, and van Knippenberg (2003) found that waitresses who repeat customers orders back to them tend to receive higher tips than waitresses who do not. Another similar study on a product-marketing scenario suggests that when an interviewer mirrors a participant's verbal and non-verbal gestures, the participant will give the product being discussed a higher rating (Tanner, Ferraro, Chartrand, Bettman, & Van Baaren, 2008). Convergence can occur when people seek social approval from their interactant (Giles & Powesland, 1975). Further, research shows that people who are in low power positions tend to adapt their language to the language of their superiors (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012). However, syntax divergence often appears when a member of a group tries to distinguish him/herself from an out-group and signal social identity to an in-group (Samter, 2007).

[1]Function words, or style words, are often contrasted with content words, which generally show what people say, whereas style words indicate how people communicate. Style words categories include pronouns, prepositions, articles, conjunctions, auxiliary verbs, etc (Tausczik & Pennebaker, 2010).

Similarity Attraction Theory (a component theory of CAT) (Byrne, 1971) suggests that people who exhibit overlapping behaviors and beliefs are more likely to be attracted to one another. Extrapolating from CAT, one way to become similar to others and seek social approval is to induce language style convergence. When people feel close to one another, they affect each other more. Similarly, reducing users perceived social differences by adjusting language style matching could help us communicate more effectively with users. Giles, Taylor, and Bourhis (1973) conducted a study on bilinguals from two ethnolinguistic groups and showed that listeners of one ethnic group were more accepting and thought more highly of speakers from the second group when they used the first group's native language to explain a picture.

In the current research, we investigate the relationship between language style and discussion participation through the lens of CAT. More specifically, we examine the hypothesis that language use among conversation participants tends toward syntactic convergence by analyzing the syntax and structure of their sentences throughout a document (in our study, post or comment). To the best of our knowledge, this is the first project which studies CAT by focusing on syntax and structure of language, not just word usage. In other words, we focus on how people put words together instead of what words they use.

The conversation-structure of the social network on Reddit.com makes it a good fit for our analysis (Weninger, Zhu, & Han, 2013). Reddit posts are organized in topical forums called *subreddits*. Users express their thoughts or post their questions on subreddits to receive comments from other users. Both the natural setting and large size of this dataset make it unique from the previous research in this area using lab-based experiments.

In what follows, we first introduce our novel approach for calculating syntactic similarity of two given documents. We also provide details about the data we used throughout our project. Next, we discuss the two experiments which we conducted to examine our hypothesis about within-subreddit and within-person variations in writing style. Each experiment's procedure is followed by its results. Finally, we discuss the conclusion of our work and future directions.

Method

In this section we introduce the approach and the dataset we used to study syntax accommodation in social media conversations.

Syntax Similarity

To the best of our knowledge, this research is the first to study syntax accommodation in social media conversations through the lens of CAT. To achieve this goal, we employed a novel approach for calculating the syntactic similarity of two documents, which we describe in detail in this section. In the following section, we report results generated by applying this method to test our hypotheses about syntax accommodation among Reddit users.

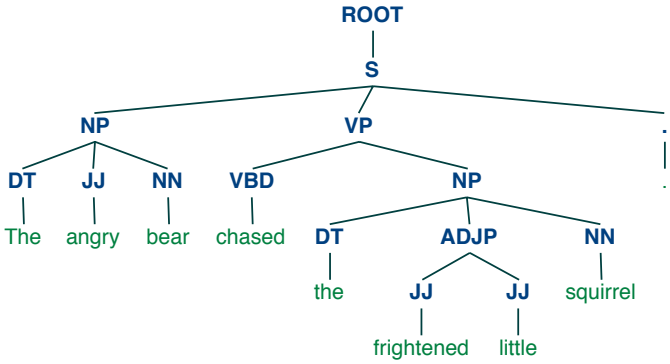


Figure 1: An example of a parse tree for the sentence "The angry bear chased the frightened little squirrel."

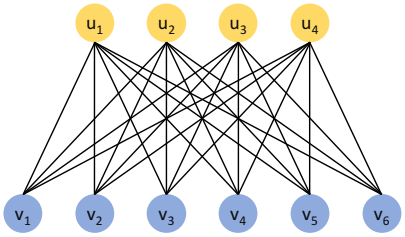


Figure 2: An example of a complete bipartite graph. The blue vertices are considered as set U and the yellow vertices are considered as set V.

Our method for calculating the syntactic similarity between two documents involves three major steps.

First, we build a parse tree for each sentence in the two documents being compared (De Marneffe, MacCartney, Manning, et al., 2006). A parse tree is a tree-shaped data structure that represents the syntactic structure of a sentence (e.g. Figure 1). Second, we calculate the between-document pairwise similarity for each sentence by comparing the parse trees for all sentences in one document to the parse trees of all sentences in the other document.

We use the Edit Distance algorithm to calculate this similarity between the two parse trees (Navarro, 2001). This algorithm counts the minimum number of operations needed to transform one tree into the other one. Operations may include deleting, adding or updating a node in a tree. We then apply the Edit Distance algorithm to every pair of parse trees between the two documents. For example, if one document has two sentences and the other document has three sentences, the edit distance algorithm is performed six times.

We then normalize these results by dividing the output of the edit distance algorithm by the number of nodes in the parse trees. This division mitigates potential biases in our measurement of syntactic similarity due to increases in the number of operations as the number of nodes increase. The normalized output represents the syntactic similarity between the two input sentences parse trees, with a separate measure

of syntactic similarity for each sentence pair between the two documents. Thus, as the metric approaches zero it indicates greater similarity between two parse trees.

Third, we use a complete bipartite graph structure and an algorithm called the Hungarian algorithm (Kuhn, 1955) to pair the sentences with the greatest syntactic similarity between two documents. We chose this algorithm to avoid introducing unnecessary noise into the measurement of document-level syntactic similarity. In the field of graph theory, a bipartite graph is defined as a graph which is composed of two independent sets of vertices or nodes, U and V. That is, no two vertices within the same set are connected by an edge (e.g. path), a property that is referred to in the field as being adjacent, but each vertex in one set share an edge with every vertex in the other set (Figure 2).

Alternatively, we could have estimated the syntactic similarity between documents by taking the average of each set of sentence-level similarities to yield a point estimate of document-level syntactic similarity. However, it seems likely that simply averaging the similarities of all sentence pairs would risk introducing considerable noise to the overall model. For example, consider a case in which two documents are being compared and each document contains three sentences. Further, imagine that within both documents the sentences that constitute those documents have substantially different syntactical structures, but across documents each sentence has a pair with a nearly identical syntactical structure. We would argue that because these two documents contained sentences with close syntactic pairs, they should be considered as having strong syntactic similarity. However, simply averaging the syntactic similarity between all sentence pairs would wash out this similarity.

In our method, to compare two documents U and V, we construct a complete bipartite graph with one set's vertices representing document U's parse trees and the other representing document V's parse trees. Further, each edge between a given pair of vertices from each set is weighted by the edit distance between the parse trees represented by those vertices calculated in step two.

Once the complete bipartite graph is constructed, we use the Hungarian algorithm (Kuhn, 1955) to find the pattern of vertex pairings that minimizes the weights of all edges, an optimization process referred to as minimum weight perfect matching. Perfect matching refers to a set of pairwise non-adjacent edges, in which every vertex should appear in exactly one matching. Minimum weight perfect matching is a special case of perfect matching which attempts to choose edges with lower weights. Given vertices $i \in U$ and $j \in V$, the weight function $w(i, j)$ refers to the weight of the edge between two vertices i and j. The goal in minimum weight perfect matching problem is to minimize the following equation:

$$\sum_{i \in U \text{ and } j \in V} w(i, j) \tag{1}$$

The minimum perfect matching of the bipartite graph of sentences pairs the two most similar sentences from the two given documents (i.e. two sets of U and V). The output of the Hungarian algorithm is a perfect matching on the weighted complete bipartite graph of sentences, which gives us an optimized measure of similarity between two documents. Note that the smaller the output is, the more similar the two documents are. By using the Hungarian algorithm to optimize the pairing of sentences to maximize similarity, our method sidesteps the problem of signal distortion that would be caused by a simple averaging algorithm. Accordingly, documents that have both high syntactic similarity but also high within-document syntactic variation are still scored as having high syntactic similarity.

Figure 3 is an overview of the approach which was described.

We conducted two experiments to test the hypothesis of syntax accommodation in social media conversations. We hypothesize that comments on a social media post tend to follow the syntactic structure of the original post. Before we discuss the details of the experiments, we will first discuss how we compiled our dataset.

Data

We collected data from Reddit.com to examine CAT in social media conversations. Reddit is a social network in which users create content (texts or links) and other users may comment on the created content. Reddit is organized into different topical areas which are called subreddits. The posts and comments on each subreddit are moderated by users for off topic, making posts on this social media platform cleaner and more topic-specific than posts found on other forms of social media.

We used the Text Analysis, Crawling and Interpretation Tool (TACIT) (Dehghani et al., 2015) to collect data from two subreddits: */r/liberal* and */r/conservative*.

In almost all of the collected subreddit posts, users expressed their beliefs and opinions about a particular topic. This aspect of real conversations and debates makes these two subreddits suitable for testing for the presence of CAT and syntax accommodation in social media conversations.

To assemble our corpus, we gathered all the posts in the /r/liberal and /r/conservative subreddits and their top-level comments, that is the comments are written directly in response to the post, not to other comment.

We also collected the historical data for the users who had commented on the /r/liberal and /r/conservative subreddits through TACIT. Note that some of the users didn't have any historical data.

To facilitate syntax comparison between a comment's text and the original post's text, we removed the posts which had links to other webpages or images but no text content (referred to as selftext). We also removed the posts with no selftext in the users' historical dataset.

At the end of the data gathering phase, we had a corpus of 167 and 146 posts from /r/liberal and /r/conservative subreddits (with the total of 7256 comments). The average length

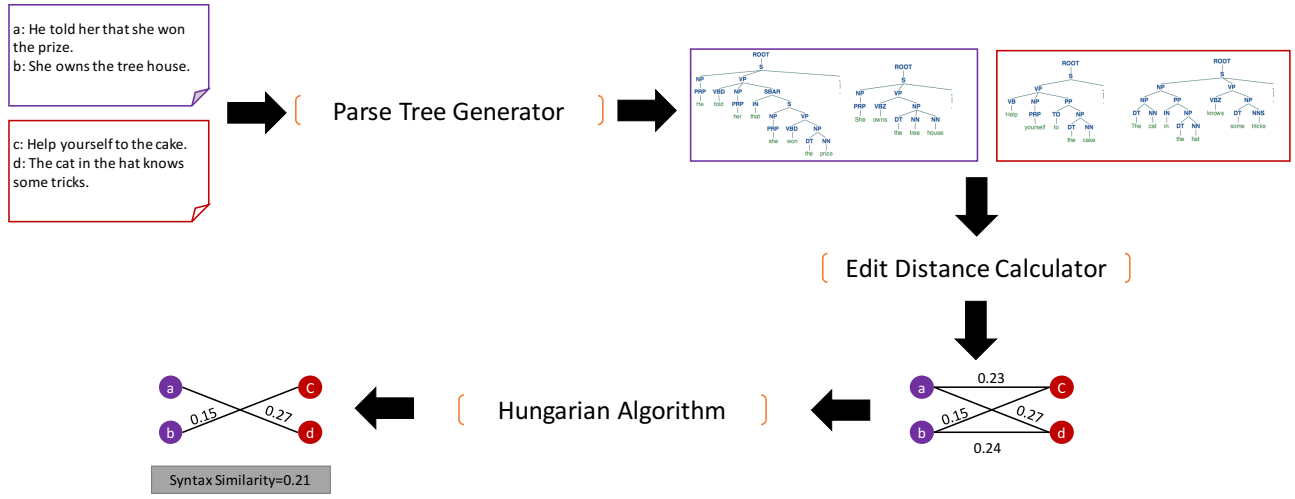


Figure 3: Syntax Similarity Calculation Process. The edit distance calculator module calculates the similarity of each pair of parse trees which are outputted by the parse tree generator module. In the last step, the Hungarian algorithm module finds the minimum weight perfect matching of the graph of sentences' parse trees. The bold edges are the ones that are selected by the Hungarian algorithm. The overall syntax similarity of the two document is the summation of the selected edges' weights divided by the number of edges.

of a post or comment in /r/conservative /r/liberal subreddits is 53 and 87 words respectively. Also the corpus of historical data included posts from 2902 users.

Experiment One

The first experiment aims to examine differences in syntax among comments and posts in a given subreddit. We hypothesize that comments written in response to a post are more syntactically similar to the original post compared to a random comment from a random post. In order to exclude the effects of homophily in syntax accommodation, we selected the random comparison comment from the same subreddit as the original comment. Therefore, the original post, comment, and random comment are all written within the same community.

Approach

In this experiment, we use the entire corpus of posts and top-level comments from the two subreddits, /r/liberal and /r/conservative.

Suppose comment C_0 is written on post P_0 , and post P_0 is posted in subreddit S_0. Also, suppose that C_1 is a random comment which is written on a random post, P_1, which is in the same subreddit as the post P_0 (i.e. S_0). Using the method proposed in the previous section, we calculate the syntactic similarity of C_0 and P_0 and the syntactic similarity of C_1 and P_0, which is a randomly chosen comment from the same subreddit community. We run the syntactic similarity for each of the 7256 comments in the subreddit S_0. If the syntactic similarity of C_0 and P_0 is significantly higher than the syntactic similarity of C_1 and P_0, we may conclude that comments on a post are more likely to follow similar syntactic structure to the original post compared to random comments.

$$Syntax_Similarity(C_0, P_0) > Syntax_Similarity(C_1, P_0) \quad (2)$$

In other words, if the left-hand side of the equation 2 is higher than the right-hand side, it provides evidence that the syntax of a post affects the syntax of top-level comments that follow it. Because all predictions were a priori and directional, they were verified with one-tailed tests (Rosenthal, Rosnow, & Rubin, 2000).

Results

To conduct the first experiment, each comment in the subreddit S_0 was considered as C_0 in the equation 2. We calculated the syntactic similarity of each comment to its original post, P_0 for all comments in /r/liberal subreddit and in /r/conservative subbredit. We also calculated the syntactic similarity of P_0 and a random comment C_0 in each run. As mentioned in the method section, smaller output numbers indicate higher syntactically similarity between the comment and the post.

The results support our hypothesis that a comment, C_0 and its original post, P_0 are more syntactically similar than a random comment C_1 and that post P_0 $(t(14354) = 3.5967, p < 0.01)$.

Experiment Two

The second experiment was deigned to focus on users' historical data in order to look at within-person variations in writing style.

We hypothesize that users' writing styles in comments are influenced by the original post's syntax, and that their com-

ments are more syntactically similar to the post's writing style than to their syntax in previous posts. If this hypothesis holds, it provides strong evidence for syntax accommodation in social media conversations.

Approach

In this experiment, we assessed users' previous posts. This dataset contains the historical data of 2902 users who have commented on the two subreddits, /r/liberal and /r/conservative.

Suppose comment C_0 is written on post P_0 from subreddit S_0, and comment C_0 is written by user U_0. Also suppose that P_1 is a random post which is written by user U_0 in another subreddit S_1. We measure the syntactic similarity of C_0 and P_0 and also the syntactic similarity of C_0 and the randomly picked post, P_1. We run this experiment for the dataset of historical data of all the users who have commented on /r/liberal or /r/conservative subreddits. If the syntactic similarity of C_0 and P_0 is significantly higher than the syntactic similarity of C_0 and P_1, we may infer that the syntactic structure of users' comments is more affected by the original post's syntax, compared to the user's syntax in previous posts.

The left-hand side of equation 3 represents the syntactic similarity of the comment and the original post, and the right-hand side shows the syntactic similarity of the comment and a random post from the user who has written the comment. If the left-hand side is significantly higher than the right, our hypothesis will have been supported.

$$Syntax_Similarity(C_0, P_0) > Syntax_Similarity(C_0, P_1) \quad (3)$$

Results

The entire dataset of users who had commented on /r/liberal or /r/conservative subreddits is used for this experiment. For all of the comments in our dataset of /r/liberal and /r/conservative subreddits, if the comment was written by a user in the pool of users with historical data, the syntactic similarity of C_0 and the original post P_0, is calculated. We also calculate the syntactic similarity of C_0 and a random post from the user's historical data, P_1. We hypothesize that a comment is more syntactically similar to its original post compared to user's previous post. Because the random post and the comment was written by the same user, we used a dependent t-test to test our hypothesis and our results show that C_0 is more syntactically similar to P_0 compared to P_1 ($t(4772) = 1.7943, p = 0.036$).

Discussion

We applied a novel approach for calculating the syntactic similarity of two documents in order to investigate the Communication Accommodation Theory hypothesis at large scale. This approach consists of three main steps. First, we generate parse trees for all of the sentences in the two given documents. Next, we measure the difference between each pair of sentences from the two documents using the Edit Distance algorithm, which computes the number of operations needed to transform one tree in to another. Finally, we create a complete weighted bipartite graph in which nodes represent sentences and the weights of edges represent edit distance between two paired sentences and then by finding the minimum weight perfect matching, we compute the syntactic similarity of the two documents.

Using this method, we provided evidence for the Communication Accommodation Theory in social media conversations through two experiments using a Reddit dataset of two subreddits consisting of 313 posts, 7256 top-level comments, and 2902 users' historical data. We showed that users tend to follow the language syntax of their conversation partner. In the first experiment, we tested whether comments which are written on a post are affected by the original post's syntax. In the second experiment, we examined the similarity of a comment to the original post compared to a random post written by the writer of the comment.

As discussed in the previous section, our results demonstrate that comments that were written on a post are more syntactically similar to the original post compared to a random comment. This finding supports our first hypothesis that a post's syntax affects the syntax of the comments that follow it. Furthermore, a user's comment is syntactically more similar to the original post compared to not only a random comment, but also his or her previous posts. This finding provides strong evidence for the existence of syntax accommodation in the social media conversations. When users write comments on a post, they tend to use a similar syntactic structure as the post's syntax, rather than using their own previous writing style.

Conclusion and Future Works

The two studies discussed in this paper provide evidence for the importance of syntactic similarity for linguistic style accommodation in social media conversations. Our results suggest that a post's syntactic structure influences the comments' syntax, providing an evidence for syntax accommodation in social media conversations. Future research should continue to explore how language style affects people's behaviors or beliefs and whether we can encourage interpersonal liking or behaviors through syntax matching.

Previous studies have mostly emphasized semantics or word usage in language while our results provide evidence for the importance of syntax as a lens to social cognition. In this work we demonstrated the effect of posts' syntax on comments' syntax regardless of the order of sentences. However, further experiments should be performed to study syntax accommodation at finer-grained levels. Additionally, some syntactic structures might trigger syntax accommodation more than the other structures, while in our work, we considered that all have the same effect.

As mentioned earlier, CAT posits that people modify their communication dynamics to minimize or maximize their social differences by either converging or diverging their language style. Building on Similarity Attraction Theory, we

intend to explore whether using syntactically similar language would influence users' behavior. Furthermore, we aim to identify features that promote constructive debates and whether a good debate is predictable based on the convergence of users language styles.

Acknowledgment

This research was supported in part by NSF IBSS Grant #1520031.

References

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, *18*(3), 355–387.

Boyd, R. L., & Pennebaker, J. W. (2015). Did shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 0956797614566658.

Byrne, D. E. (1971). The attraction paradigm (vol. 11). *Academic Pr*.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on world wide web* (pp. 745–754).

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on world wide web* (pp. 699–708).

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., ... Parmar, N. J. (2015). Tacit: An open-source text analysis, crawling and interpretation tool. *Crawling and Interpretation Tool (September 15, 2015)*.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of lrec* (Vol. 6, pp. 449–454).

Giles, H. (2008). *Communication accommodation theory.* Sage Publications, Inc.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation.* Academic Press.

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society*, *2*(02), 177–192.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, *34*(4), 365–399.

Guéguen, N. (2009). Mimicry and seduction: An evaluation in a courtship context. *Social Influence*, *4*(4), 249–255.

Jacob, C., Guéguen, N., Martin, A., & Boulbry, G. (2011). Retail salespeople's mimicry of customers: Effects on consumer behavior. *Journal of Retailing and Consumer Services*, *18*(5), 381–388.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, *33*(1), 31–88.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42–45.

Reitter, D., Moore, J. D., & Keller, F. (2010). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Samter, W. (2007). *Explaining communication: Contemporary theories and exemplars*. Psychology Press.

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. *The new handbook of language and social psychology*(1.2), 33–56.

Tanner, R. J., Ferraro, R., Chartrand, T. L., Bettman, J. R., & Van Baaren, R. (2008). Of chameleons and consumption: The impact of mimicry on choice and preferences. *Journal of Consumer Research*, *34*(6), 754–766.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393–398.

Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in social networks analysis and mining (asonam), 2013 ieee/acm international conference on* (pp. 579–583).

The Developmental Trajectory of Children's Statistical Learning Abilities

Limor Raviv[1] (limor.raviv@mail.huji.ac.il) & Inbal Arnon[2] (inbal.arnon@gmail.com)
[1]Department of Cognitive Science, [2]Department of Psychology, Hebrew University, Jerusalem

Abstract

Infants, children and adults are capable of implicitly extracting regularities from their environment through statistical learning (SL). SL is present from early infancy and found across tasks and modalities, raising questions about the domain generality of SL. However, little is known about its' developmental trajectory: Is SL fully developed capacity in infancy, or does it improve with age, like other cognitive skills? While SL is well established in infants and adults, only few studies have looked at SL across development with conflicting results: some find age-related improvements while others do not. Importantly, despite its postulated role in language learning, no study has examined the developmental trajectory of auditory SL throughout childhood. Here, we conduct a large-scale study of children's auditory SL across a wide age-range (5-12y, N=115). Results show that auditory SL does not change much across development. We discuss implications for modality-based differences in SL and for its role in language acquisition.

Keywords: statistical learning; developmental differences

Introduction

One of the deepest questions in cognitive science is how children learn about the structure of their environment. A fruitful line of work on this question examines the ability to extract knowledge about the world via statistical learning. Statistical learning (SL) refers to the ability to implicitly detect recurring patterns in sensory input based on statistical properties, and use them to learn higher order structure, like that found in language (Romberg & Saffran, 2010; Thiessen & Erickson, 2015). The term "SL" was originally coined in a speech segmentation study showing 8-month-old infants can use transitional probabilities between syllables as a cue to word boundaries (Saffran, Aslin, & Newport, 1996).

In the past two decades, numerous studies have shown that SL is present from very early infancy (Bulf, Johnson, & Valenza, 2011, Kuhl, 2004), in a variety of modalities (auditory, visual and tactile, see Conway & Christiansen, 2005), and can facilitate learning across a range of linguistic domains (Saffran, 2003) – from phonemic inventory (Maye, Werker, & Gerken, 2002), through word order preferences (Gervain, Nespor, Mazuka, Horie, & Mehler, 2008), to syntax (Gomez & Gerken, 1999). This body of literature illustrates learners' ability to extract structure by attending to distributional regularities in their environment.

However, while SL has been studied extensively with infants and adults, much less work has looked at how these abilities develop from infancy to adulthood, even though such findings are crucial for understanding the role and nature of SL abilities. The paucity of research leaves an important question unanswered: What is the developmental trajectory of SL across childhood? Is SL an early-maturing capacity that is stable in an individual across development, or does it improve with age? On the one hand, SL is already present in very young infants and postulated to play a role in language acquisition, suggesting it is an early-maturing capacity. On the other hand, most other cognitive abilities do develop with age.

Only few studies have looked at how SL abilities change during development and they show a mixed pattern of results (see detailed review in the next section): while some argue SL is age-invariant (e.g., Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), others report an improvement with age (e.g., Arciuli & Simpson, 2011). Because these studies examined SL in different modalities (auditory vs. visual), it could be that the effect of age differs across domains. Importantly, although SL is found in multiple modalities and with various sensory inputs (e.g., Conway & Christiansen, 2005), there is growing evidence for modality-based differences in adults' SL abilities (Frost, Armstrong, Siegelman & Christiansen, 2015; Krogh, Vlach & Johnson, 2012). If SL is a unitary capacity, we may expect it to develop similarly across modalities. In contrast, a modality-specific mechanism may show different developmental trajectories in different modalities. However, there is little data that can be brought to bear on these questions.

SL Across Development

In theory, there are several possible predictions on the developmental trajectory of SL. The first is that SL improves with age, just like many other cognitive abilities (e.g. working memory, see Gathercole, Pickering, Ambridge & Wearing, 2004). This prediction is also motivated by recent findings from the field of implicit learning. SL is often seen as a type of implicit learning, occurring without explicit intent and/or overt awareness (Perruchet & Pacton, 2006). Traditionally, implicit learning mechanisms were considered to be age-invariant: they were seen as early-maturing and automatic capacities that do not improve with age (Reber, 1993). Yet this view has been challenged in recent years (Lukács & Kemény, 2014). While some studies support age-invariance (Vinter & Perruchet, 2000; Amso & Davidow, 2012), there is growing evidence that implicit learning does improve with age (Vaidya, Huger, Howard & Howard Jr, 2007; Janacsek, Fiser, & Nemeth, 2012; Lukács & Kemény, 2014). Since SL involves implicit learning, we may expect it to show a similar developmental trajectory and improve with age across modalities (Misyak, Goldstein & Christiansen, 2012).

A different prediction can be made when we consider the role of SL in language acquisition. Since infancy and early childhood are considered to be the prime-time for language learning (Birdsong, 1999), SL skills may be fully developed

in infancy and not improve with age, a claim supported by the presence of SL in newborns (Bulf, Johnson, & Valenza, 2011). Such age invariance would be in line with Reber's claim that some implicit learning mechanisms are early-maturing. A recent fMRI study suggests that auditory SL may even become worse with age. McNealy, Mazziotta & Dapretto (2011) found an age-related decrease in sensitivity to weak statistical cues: younger children showed better sensitivity to low transitional probabilities compared to older children and adults. They suggest that this age-related decrease may help explain adults' worse language learning skills. Under this account, auditory SL skills may even show a negative age effect and deteriorate with age.

A third, more nuanced, prediction on the effect of age on SL takes into account modality-based differences. The fact that SL is found in multiple sensory modalities suggests it is domain general mechanism that works similarly on different kinds of input, linguistic and nonlinguistic (Saffran & Thiessen, 2007; Kirkham, Slemmer & Johnson, 2002; Saffran, Pollak, Seibel & Shkolnik, 2007). This idea receives support from the correlation between visual SL and auditory linguistic measures (Shafto et al., 2012). However, there is growing evidence of differences in learning between the auditory and visual domains that are more consistent with a modality-sensitive model of SL (Frost et al., 2015).

On an individual level, performance on auditory and visual SL tasks is not correlated, indicating they may tap onto different abilities (Siegelman & Frost, 2015). There also seem to be qualitative differences in learning across modalities, though studies differ in which modality shows better learning: Siegelman & Frost (2015) found that adults were better in the visual domain, while other studies found that adults showed better learning in the auditory domain (Conway & Christiansen, 2005; Saffran, 2002). This difference may also be modulated by constraints involving presentation rate: increasing stimuli presentation rate led to better learning the auditory domain, but worse learning in the visual domain (Emberson, Conway & Christiansen, 2011). One way of reconciling these findings is by characterizing SL as a domain general mechanism that applies similar computational principles to all input types but that is nevertheless modality-specific in that it reflects the particular constraints and perceptual biases imposed by different sensory input (Frost et al. 2015).

From a developmental perspective, age may affect learning differently in the visual and auditory domains. In particular, given its role in language acquisition, auditory SL may change less with age (or not at all) compared to visual SL. Such a finding would provide support for the modality-sensitive nature of SL. Yet there is currently very little empirical evidence to support any of these predictions or distinguish between them. To date, only few studies have examined the effect of age on SL, and they display a mixed pattern of results. Interestingly, some of the differences could be modality related.

In the visual domain, two studies that compared children and adults on the same task found no effect of age. Bertels, Boursain, Destrebecqz, & Gaillard (2015) report that children (aged 9 to 12) performed similarly to adults on a visual SL task. Jost, Conway, Purdy, Walk, & Hendricks (2015) found similar ERP patterns during a visual SL task among younger children (8-year-olds), older children (11-year-olds) and adults. Infants' visual SL also does not seem to change between 2- and 8-months (Kirkham et al., 2002). Overall, these findings seem to point to age-invariance in visual SL. However, they are based on quite small samples, examine a relatively narrow slice of development and compare groups of children in given ages rather than examining age as a continuous factor, all of which may mask the effect of age on performance (Arciuli & van Kos Torkisden, 2012; Bertals et al., 2015). Accordingly, a more comprehensive study of visual SL across development did find clear age-related improvement: Arciuli & Simpson (2011) examined visual SL in 183 children between the ages of 5 and 12 and found that older children showed significantly better learning. This single comprehensive study suggests that visual SL does in fact improve with age.

Crucially, even less work examined SL across development in the auditory domain, despite its postulated role in language acquisition. Only one study compared children and adults on the same auditory task and reported age-invariance: Saffran et al. (1997) found no difference in auditory SL between 6-year-olds and undergraduate students, with both age groups showing similar learning. This is somewhat surprising given that auditory grammar learning (AGL) improves with age, with adults showing better learning than six and nine-year-olds (Saffran, 2001). Studies on auditory SL during infancy also suggest age-related changes in early development. Specifically, infants' auditory learning biases change between 8- and 10-months (Emberson, Misyak, Schwade, Christiansen & Goldstein, 2015), and they show improved auditory SL abilities between 12- and 15-months (Gómez & Maye, 2005).

In sum, while Arciuili & Simpson's (2011) findings suggest that visual SL improves throughout development, the only study to compare auditory SL in children and adults found no difference between them. One possibility is that age does affect SL in the auditory modality as well, but that this pattern is not detected when using relatively small samples and comparing only one age group to adults (as in Saffran et al., 1997). Alternatively, it is possible that there are fundamental differences between visual and auditory SL that are also reflected in different developmental trajectories across modalities. In particular, SL may improve less in the auditory domain, or even show age invariance across childhood. These possibilities are hard to evaluate given the existing literature, as no study has examined auditory SL across development in a comprehensive way as was done in the visual domain by Arciuli & Simpson (2011).

The Current Study

To test the different predictions on the effect of age on auditory SL, we conducted the first large-scale, cross-sectional study of children's performance on an auditory SL

task (ASL) across a wide age-range (5-12y). We ask whether performance is affected by age, and if so - then how: will auditory SL abilities improve across development as found in the visual domain, or will they show a different developmental trajectory?

Method

Our task was closely modeled on the classic segmentation task from Saffran et al. (1996). The task was completely computerized, with a human experimenter sitting by the children and providing them with verbal instructions.

Participants

115 children (age range: 5-12y, mean age: 8:3y, 63 boys and 52 girls). All children were visitors at the Bloomfield Science Museum in Jerusalem and were recruited for this study as part of their visit to the Israeli Living Lab. All participants received a small educational reward in return for participation. Parental consent was obtained for all children. All children were native Hebrew speakers, and none of them had known language or learning disabilities.

Materials

The auditory stimuli consisted of a synthesized "alien" language, containing 5 unique tri-syllabic words (made up of 15 different syllables): "*dukame*", "*gedino*", "*kimuga*", "*nalobi*" and "*tobelu*". All words were synthesized using the PRAAT software in order to control for syllable duration and frequency. Average word length was 850ms. The words were concatenated together in a semi-randomized order (with the constraint that no word will appear twice in a row) to create an auditory familiarization stream. In this stream, the transitional probabilities (TPs) between syllables within a word was always 1, while the TPs between words were 0.25 (because syllables were not repeated across words). The exposure phase lasted 2;20 minutes, with each word repeated 32 times. Importantly, there were no breaks between words and no prosodic or co-articulation cues in the stream to indicate word boundaries.

The test phrase included 25 two alternative forced-choice trials (2AFC), in which participants heard two possible "words" (separated by 500ms), and had to choose which one sounds more like the language they just heard. On each trial, participants heard a real word (like "*dukame*") either followed or preceded by a foil word. Foil words were constructed by taking the first syllable from one word, followed by the second syllable form another word, and the third syllable from a third word. Thus, each syllable in the foil words appeared in a similar position in real words, but with different surrounding syllables (for example, "*kilome*" or "*dubega*"). This created a difference in the statistical properties of the words and foils: while the TPs between every two adjacent syllables within a word are 1, the TPs between every two syllables in a foil test item are 0, as participants never heard these syllables one after the other during familiarization. If participants learn the statistical properties of the syllables in the stream, they should be able to distinguish between words and foils. The possible score on this task ranged from 0% accuracy (0/25 trails correct) to 100% (25/25 trails correct).

Procedure

Children were told they are about to hear an alien language, and were then exposed to the familiarization stream using isolating headphones. A picture of an alien in a spaceship appeared on the screen during the entire duration of the familiarization stream. Following exposure, children were told that they are about to hear an alien who is not a good speaker of the alien language, and that they must help him by telling him which of the two words he will say sounds more like the alien language they just heard.

The 25 test trails were presented to children in random order (with the constraint that the same word/foil will not appear in two consecutive trails). The order of words and foils on each trial was also randomized, so that half the trials were word-first trials (the word was heard before the foil) and half were foil-first trials (the foil was heard before the word). After hearing both possibilities, children were asked to press either '1' or '2' according to whether they thought the correct word was the first or the second they heard. In case children felt they didn't know the answer, the experimenter encouraged them to try and guess what sounds more familiar according to the alien language they heard. At the end of the task, the experimenter thanked the child for helping the alien learn the language, and showed them to the prize basket to pick their reward for participation.

Results

As a group, children showed learning in the auditory task with a mean accuracy score of 55%, which is significantly above chance ($t(114)=4.74$, $SE=1$, $p<0.0001$). Figure 1 shows children's mean performance on the task as a function of age (in half years): age seems to affect performance with older children showing better accuracy. To test for significance, we used a mixed-effect logistic regression model. Our dependent binominal variable was success in a single ASL test trial (see Table 1a). The model included fixed effects for age (in half years) and trial number as centered continuous factors, order of appearance in the test (word-first trials vs. general mean, deviation coding) and gender (females vs. general mean, deviation coding). Following Barr et al. (2013), the model had the maximal random effects structure that would converge, including random intercepts for participants and items, by-participant slopes for trial number and order of appearance, and by-items slopes for age and order of appearance.

The model showed that age had a significant positive effect on performance, with older children displaying better accuracy ($\beta=0.05$, $SE=0.02$, $p<0.05$). The effect of gender was not significant ($\beta=-0.05$, $SE=0.04$, $p>0.1$). Trial number was also not a significant predictor of accuracy ($\beta=-0.004$, $SE=0.005$, $p>0.1$), confirming that no learning (or unlearning) was happening during the test phase itself, despite the repetitions of both foils and words.

Interestingly, order of appearance in the test highly affected performance, with better accuracy on word-first trials (β=0.15, SE=0.03, p<0.001). Thus, children were better in trials where the real word was heard before the foil. Since the order of presentation was counter-balanced this could not reflect a preference for pressing 1 or 2. The advantage for word-first trials could reflect the "interval bias" found in 2AFC tasks in which the first word is used as a baseline for assessing the "wordness" of the next item (Einhom & Hogarth, 1981; García-Pérez & Alcalá-Quintana, 2011).

These results suggest that auditory SL improves with age. However, a somewhat different pattern emerged when we examined participants' performance in different age bins[1] (matched in number of participants, see Table 2).

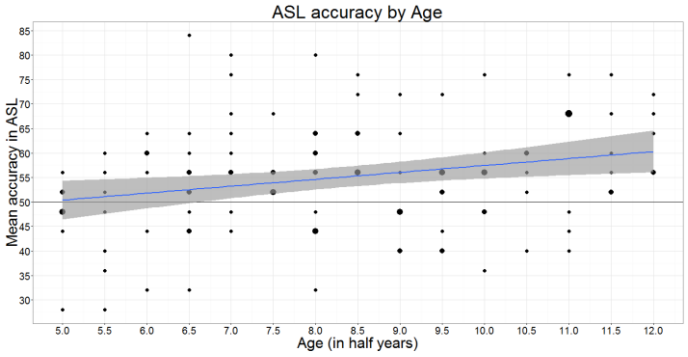


Figure 1: Accuracy in the task by age (in half years).
Each dot represents an accuracy score (ranging from 0% to 100%) shown by one (or more) participants in the relevant age range (size corresponds to the number of participants with this score). For example, some 8-year-olds were 60% accurate while others were 80% accurate. The plotted blue line represents the linear regression line, with the standard confidence interval appearing in gray. The black line represents the 50% chance level.

Table 1a: regression model for ASL (ages 5-12y)

	Estimate	Std. Error	z value	*p*-value
(Intercept)	0.198826	0.043712	4.548545	**< .001 *****
Age	0.056382	0.02222	2.537491	**< .05 ***
Trial Number	-0.00402	0.005519	-0.72823	> .1
Order of Appearance	0.156133	0.039387	3.964068	**< .001 *****
Gender	-0.05032	0.042881	-1.17336	> .1

Table 2: ASL Accuracy by Age Bins

	N	Mean VSL score	p-value
Age Group 5 to 6	21	**48%**	> .1
Age Group 6.5 to 8	38	**56%**	**< .01 ****
Age Group 8.5 to 10	32	**55%**	**< .05 ***
Age Group 10.5 to 12	24	**59.5%**	**< .001 *****
All Children	115	**55%**	**< .001 *****

[1] We applied age bins only for means of presenting the data, yet used age as a continuous factor (in half years) in all our analyses.

We found that children in the youngest age group (ages five to six) did not show significant learning: unlike children in all other age groups, their performance did not differ from chance (M=48%, t(20)=-0.786, SE=2, p=0.44). This is also reflected in Figure 1, which shows that the majority of children aged 6 and below are performing at chance level (we will address this issue in the discussion).

We therefore conducted a second analysis to see if the effect of age on performance was driven by the inclusion of the youngest age group that showed no learning. We ran an additional model with a similar effects structure using the data obtained only from children older than 6, excluding the 21 children in the youngest age bin (Table 1b).

As suspected, the effect of age disappeared in the new model: without the youngest group of children, who showed chance-level performance, age was no longer predicative of accuracy (β=0.02, SE=0.02, p>0.1). That is, auditory SL did *not* show an improvement between the ages of 6.5 and 12, a significant developmental window. In line with previous reports (Saffran et al. 1997), this finding suggests that auditory SL is a rather stable, age-invariant capacity across childhood, at least after age 6.

Table 1b: regression model for ASL (ages 6.5-12y)

	Estimate	Std. Error	z value	*p*-value
(Intercept)	0.266438	0.051252	5.198571	**< .001 *****
Age	0.022181	0.029971	0.740099	> .1
Trial Number	-0.00586	0.006286	-0.93166	> .1
Order of Appearance	0.159104	0.04499	3.536454	**< .001 *****
Gender	-0.04966	0.047842	-1.03799	> .1

Discussion

Our study shows that auditory SL did not improve with age after excluding the youngest age group of children (aged 6 and below), who performed at chance. Even though we initially found an age-related improvement in the task, our second analysis showed that this effect was driven by the youngest group of children, who simply did not show learning. After excluding this group, we found no significant change in children's auditory SL skills across development (between the ages of 6.5 to 12y). This result differs from the findings of Arciuli & Simpson (2011), who reported an improvement with age in the visual domain, and is in line with Saffran et al. (1997), who found age-invariance in the auditory domain when comparing six-year-olds to adults on a similar auditory segmentation task. That is, auditory SL, unlike visual SL, seems to be age invariant.

The finding that auditory SL is age-invariant across development is supportive of Reber's claim that some implicit learning mechanisms, presumably those that are used early in life may not be affected by age. The age-invariance seen in this study is consistent with the postulated role of auditory SL in language acquisition (e.g., Saffran et al., 1996), and the fact that language learning skills are at their prime during infancy and early childhood (Birdsong, 1999).

Our findings also points to modality-based differences in the developmental trajectory of SL abilities: unlike visual SL (Arciuli & Simpson, 2011), auditory SL did not change during development. This finding also strengthens the claim that the improvement reported in the visual domain by Arciuili & Simpson (2011) reflects true changes in SL, and is not merely the result of general improvement in other cognitive abilities related to task performance (e.g., attention, working memory). If the improvement in visual SL was caused by the maturation of other cognitive skills, we would expect a parallel pattern across domains, which is not the case. Taken together, our study supports modality-sensitive models of SL where the type of stimuli has an effect on learning (Frost et al., 2015).

Importantly, not enough is known about the nature of the difference between modalities as no direct comparison was made between auditory and visual SL in children using tasks with similar input statistics. Our task differed from that of Arciuli & Simpson (2011) in having five words compared to four visual triplets: while it is possible this affected the results, it is unlikely to have caused the effect of age to disappear. If anything, the boundaries between words are more marked in our study (because of the larger number of words). Nevertheless, to comprehensively assess possible modality-based differences in the developmental trajectory of SL skills, a direct comparison should be made between auditory and visual SL skills using similar populations, similar designs, similar statistical properties of the input, similar exposure and preferably the same children performing both tasks to detect correlations in learning across domains.

Our results are limited in an important respect: we found that children below 6 did not show any explicit learning in our task. Since infants do exhibit significant learning in ASL when using more implicit tasks (e.g, Saffran et al., 1996), it is possible the youngest children are capable of learning the auditory regularities but fail in manifesting this knowledge using more explicit tasks. Children under six may be too young to exhibit their implicit knowledge and/or to understand the verbal instructions given to them. Indeed, studies show that it is significantly harder for young children to learn an artificial language in laboratory settings, resulting in inferior performance overall (Ferman & Karni, 2010; Folia et al., 2010). Consequently, more implicit tasks should be used in future studies with younger children (before age 6).

Such findings – from early childhood – are crucial in understanding the developmental trajectory of auditory SL. This time period is the one in which children's language skills develop the most, and in which many other cognitive changes occur. It could be that auditory SL skills only improve until a certain period in childhood and remain steady from that point on. In other words, auditory SL may be fully developed by the first year of life, or may develop only during early childhood.

Conclusions

The results of this study suggest that auditory statistical learning does not improve with age during childhood, but is at adult capacity by age six. This finding suggests that auditory SL is an early-maturing and stable capacity in an individual. However, more research is required to (a) examine whether SL abilities in different modalities are affected differently by age, and (b) examine the developmental trajectory or SL from infancy to early childhood. Future work should compare the developmental trajectories of auditory and visual learning across development using similar tasks, and studies with younger children should employ more implicit tests.

References

Amso, D., & Davidow, J. (2012). The development of implicit learning from infancy to adulthood: item frequencies, relations, and cognitive flexibility. Developmental Psychobiology, 54(6), 664-673.

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental Science*, 14(3), 464-473.

Arciuli, J., & von Koss Torkildsen, J. (2012). Advancing our understanding of the link between statistical learning and language acquisition: The need for longitudinal data. *Frontiers in Psychology*, *3*(AUG), 1–9.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language, 68(3), 255-278.

Bertels, J., Boursain, E., Destrebecqz, A., & Gaillard, V. (2015). Visual statistical learning in children and young adults: how implicit? *Frontiers in Psychology*, *5*(January), 1–11.

Birdsong, D. (Ed.). (1999). *Second language acquisition and the critical period hypothesis*. Routledge.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127-132.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24-39.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Journal of Accounting Research*, 1-31.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. The Quarterly Journal of Experimental Psychology, 64(5), 1021-1040.

Emberson, L., Misyak, J. B., Schwade, J., Christiansen, M. H., & Goldstein, M. H. (2015). How abstract is Statistical learning? Comparing learning across visual and auditory perceptual modalities in infancy. Presented in *interdisciplinary advances in statistical learning*, Spain.

Ferman, S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. PLoS One, 5(10), e13648.

Folia, V., Uddén, J., De Vries, M., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. Language learning, 60(s2), 188-220.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity : the paradox of statistical learning. *Trends in Cognitive Sciences*, 1–9.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*(7), 2332-2352.

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, *57*(1), 56–74.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Gómez, R., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy*, *7*(2), 183–206.

Janacsek, K., Fiser, J., & Nemeth, D. (2012). The best time to acquire new skills: Age-related differences in implicit sequence learning across the human lifespan. *Developmental Science*, *15*(4), 496–505.

Jost, E., Conway, C. M., Purdy, J. D., Walk, A. M., & Hendricks, M. a. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. *Brain Research*, *1597*, 95–107.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35-B42.

Krogh, L., Vlach, H. A., & Johnson, S. P. (2012). Statistical Learning Across Development: Flexible Yet Constrained. Frontiers in Psychology, 3, Article 598.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. Nature reviews neuroscience, 5(11), 831-843.

Lukács, Á., & Kemény, F. (2014). Development of Different Forms of Skill Learning Throughout the Lifespan. *Cognitive Science*, *39*, 383–404.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), 101–111.

McNealy, K., Mazziotta, J. C., & Dapretto, M. (2011). Age and experience shape developmental changes in the neural basis of language-related learning. Developmental science, 14(6), 1261-1282.

Misyak, J. B., Goldstein, M. H., & Christiansen, M. H. (2012). Statistical-sequential learning in development. Statistical learning and language acquisition, 13-54.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*(5), 233–8.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford: Oxford University Press.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906-914.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. Journal of Memory and Language, 44(4), 493-515.

Saffran, J. R. (2002). Constraints on statistical language learning. Journal of Memory and Language, 47(1), 172-196.

Saffran, J. R. (2003). Statistical Language Learning : Mechanisms and Constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. a., & Barrueco, S. (1997). Incidental Language Learning: Listening (and Learning) out of the Comer of Your Ear. *Psychological Science*, *8*(2), 101–105.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition, 105(3), 669-680

Saffran, J. R. & Thiessen, E. D. (2007). Domain-General Learning Capacities, in *Blackwell Handbook of Language Development* (eds E. Hoff and M. Shatz), Blackwell Publishing Ltd, Oxford, UK.

Shafto, C. L., Conway, C. M., Field, S. L., & Houston, D. M. (2012). Visual Sequence Learning in Infancy: Domain-General and Domain-Specific Associations with Language. Infancy, 17(3), 247-271.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120

Thiessen, E., & Erickson, L. (2015). Perceptual Development and Statistical Learning. The Handbook of Language Emergence, 87, 396-414.

Vaidya, C. J., Huger, M., Howard, D. V., & Howard Jr, J. H. (2007). Developmental differences in implicit learning of spatial context. Neuropsychology, 21(4), 497-506.

Vinter, A., & Perruchet, P. (2000). Implicit learning in children is not related to age: evidence from drawing behavior. *Child Development*, *71*(5), 1223–1240.

Under Pressure: How Time-Limited Cognition Explains Statistical Learning by 8-Month Old Infants

Martyn Lloyd-Kelly (martynlk@liverpool.ac.uk)
Fernand Gobet (fernand.gobet@liverpool.ac.uk)
Department of Psychological Sciences, University of Liverpool,
Bedford Street South, Liverpool, L69 7ZA, UK

Peter C. R. Lane (peter.lane@bcs.org.uk)
School of Computer Science, University of Hertfordshire,
College Lane, Hatfield, AL10 9AB, UK

Abstract

In a classic experiment, Saffran, Aslin, and Newport (1996) used a headturn preference procedure to show that infants can discriminate between familiar syllable sequences ("words") and new syllable sequences ("non-words" and "part-words"). While several computational models have simulated aspects of their data and proposed that the learning of transitional probabilities could be mediated by neural-net or chunking mechanisms, none have simulated the absolute values of infants' listening times in the different experimental conditions. In this paper, we used CHREST, a model based on chunking, to simulate these listening times. The model simulated the fact that infants listened longer to novel words (non-words and part-words) than familiar words. While the times observed with the model were longer than those observed with infants, we make a novel finding with regard to phonological store trace decay. We also propose how to modify CHREST to produce data that fits closer to the human data.

Keywords: chunking; CHREST; headturn preference procedure; transitional probability; word segmentation

Introduction

Much of learning occurs implicitly, without conscious intention or awareness of the knowledge learned (Reber, 1993). Examples include learning in naturalistic environments (development of a first language and the acquisition of expertise) and controlled laboratory experiments (learning of artificial grammars and serial reaction time tasks). Particularly strong evidence is provided by first language acquisition: children become proficient in vocabulary, syntax, semantics and pragmatics without explicit instruction.

A standard example of implicit learning is word segmentation. To master language, infants must segment speech into words yet, word boundaries are unclear. For example, short silences in speech provide unreliable cues since they sometimes occur not only between words, but also within words. It is thought that children use a combination of several cue types to segment speech, including prosody, allophonic variation, phonotactic regularities, transitional probabilities, semantics, and words occasionally occurring in isolation (Ambridge & Lieven, 2011; Rowland, 2014; Gobet, 2015).

Of these cue types, we are particularly interested in forward transitional probabilities; predictions of syllable occurrence in a word based on the final syllable of the preceding word. These probabilities provide useful information for segmenting the speech stream, even if they are learnt approximately. The importance of forward transitional probabilities for word segmentation was demonstrated in a classic experiment by Saffran, Aslin and Newport (1996).[1] The paper proposes that infants use mechanisms which allow them to compute statistical properties of their language rapidly and such mechanisms are demonstrably active in 8 month-old infants.

SAN studied two groups of 24 8-month-old infants in two experiments. Each experiment was composed of a *learning* and *test* phase. During learning, each participant was played a continuous stream of four nonsense words, composed of three syllables (e.g. "bidaku") and randomly assorted so that no word occurred in succession in the stream, for 2 minutes. Words were uttered at a rate of 270 syllables per minute and the stream was designed so no information about prosody or cues to word boundaries existed, except for the transitional probabilities between words (unknown to the participants). Each group of infants were assigned to one of two experiment conditions, A or B, which varied the words used in the learning stream. During testing, a blinking light appeared on the wall ahead of the infant to direct their attention. Following fixation on this, a blinking light randomly appeared above a speaker on one of two side walls. Following fixation on this light, a test word from a set of 4 was played until the participant either turned their head 30° from the speaker or until the participant had fixated on the speaker for 15s (the "headturn preference procedure" (Jusczyk & Aslin, 1995)). Two words in the set were present in the learning stream (*familiar* words), the other two words (*novel* words) were either not present in the learning stream (experiment 1) or were *part-words*, i.e. composed of the last syllable of a familiar word and the first two syllables of another familiar word.

SAN's results demonstrated that infants listened to familiar words for less time than novel words in both experiments: 7.97s (SE = 0.41) vs. 8.85s (SE = 0.45) in experiment 1 ($P < 0.04$) and 6.77s (SE = 0.44) vs. 7.60 (SE = 0.42) in experiment 2 ($P < 0.03$). Thus, after an exposure of only 2 minutes, infants were able to learn enough information about the order of syllables to discriminate between familiar and novel words. Experiment 2 directly tested the hypothesis that infants can learn forward transitional probabilities; SAN took their results as evidence that they had.

Several computational models have simulated aspects of

[1] "SAN" hereafter.

SAN's study. These include simple recurrent networks (Elman, 1991; French, Addyman, & Mareschal, 2011), connectionist autoassociators (French et al., 2011), Kohonen networks (Anderson, 1999), and PARSER (Perruchet & Vinter, 1998), a symbolic model. In these models, transition probabilities are approximated by learning mechanisms based on connectionist algorithms or the creation of chunks. However, to our knowledge, no model has been used to simulate the ability of infants in SAN's study to discriminate between words, non-words and part-words and to replicate the times recorded by SAN when the headturn preference procedure is used to assess familiarity with words.

In this paper we aim to account for SAN's results using chunking mechanisms that indirectly implement the learning of forward transitional probabilities. Most importantly, we aim to simulate SAN's exact timing data, as opposed to simply demonstrating notable discrimination between the words used by SAN. Doing this adds confidence that a model captures the actual cognitive processes used by the infants rather than an abstract measure of their discrimination ability.

Participant Modelling

Each participant is simulated using an instance of the computational CHREST (**C**hunking **H**ierarchy and **R**etrieval **ST**ructures) model; a symbolic cognitive architecture (Gobet & Lane, 2010). In this study, CHREST is augmented with a phonological loop (Baddeley & Hitch, 1974) that is subject to time-sensitive trace decay (Baddeley, 2007).[2] Since CHREST's cognitive operations are also time-sensitive, we show how these structures can interact to segment speech and attempt to replicate the times recorded in SAN's study.

CHREST

CHREST has been used to account for data in several domains characterized by the acquisition of implicit learning. These include the acquisition of expertise in board games (de Groot & Gobet, 1996), implicit learning tasks (Lane & Gobet, 2012), children's acquisition of vocabulary (Jones, Gobet, & Pine, 2000) and syntax (Freudenthal, Pine, & Gobet, 2009).

A CHREST model consists of three major components: long-term memory (LTM), short- term memory (STM) and an input interface (the phonological loop, in this case). CHREST uses *patterns*, ϕ, generated from the environment using its input interface to create, modify and retrieve LTM nodes, i.e. to learn and recognise. These nodes are organised in a discrimination network: nodes are first organised according to their modality (action, visual or verbal) then according to the information that differentiates them from other nodes. STM is a fixed-size, first-in, first-out list structure that stores retrieved LTM nodes (whereupon they are called "chunks") to facilitate recognition. Note that there exists a STM structure for each modality.

[2] A simplified phonological loop has been used in earlier CHREST models, see (Jones, Gobet, & Pine, 2005; Lane & Gobet, 2012) and references therein.

The version of CHREST used in this paper simply learns and recognises verbal information when requested. To learn, content from the phonological loop is added to a ϕ in the order it was inserted into the phonological loop (oldest information added first) and ϕ is then sorted through the model's LTM. The LTM node reached, θ, is simultaneously placed into STM and its *image* (the pattern contained in θ) is analysed to determine if ϕ contains any new information. If there is new information in ϕ, CHREST will attempt to learn it either either by *discriminating* or *familiarising*. When recognising, CHREST simply checks if ϕ is present in a verbal STM chunk retrieved after sorting ϕ through LTM.

Discrimination increases the number of nodes in a CHREST model's LTM and occurs either when θ is a modality root node (ϕ is not recognised at all), or θ is not a modality root node but its image either mismatches ϕ or is *finished*, i.e. no new information can be added to it. In the first case, a new node is created and connected to the relevant modality root node (see Figure 1(b)); the new node is technically called a *primitive*. In the second case, a new node is created and connected to θ (see Figure 1(c)). The connection created is a *test-link* that contains the first unit of information that is present in ϕ but not θ's image.

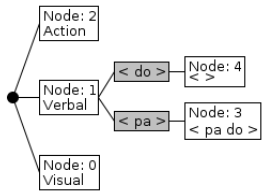
Familiarisation (see Figure 1(d)) adds new information to images of existing LTM nodes and occurs when ϕ contains a sub-pattern, ρ, that exists as a primitive; ρ is not present in θ's image; information preceding ρ in θ's image and ϕ are the same; θ's image is not finished.

Learning consumes a CHREST model's cognitive resource for a specified period of time. Consider a CHREST model, *c* that takes 10000ms to discriminate, 10ms to traverse a link in LTM and 50ms to update its STM. If *c*'s LTM is in the state depicted in Figure 1(a) and it is asked to learn the verbal pattern < pa go > at time t (resulting in the production of the LTM state seen in Figure 1(b)), *c* will take 20ms to sort < pa go > through LTM (10ms to sort its modality and 10ms to traverse the link to node 3). When node 3 is retrieved, it will be placed into verbal STM at time $t + 20\text{ms} + 50\text{ms}$ and node 5 will be created at time $t + 20\text{ms} + 10000\text{ms}$. Any requests to learn another pattern between t and $t + 20\text{ms} + 10000\text{ms}$ will be blocked.

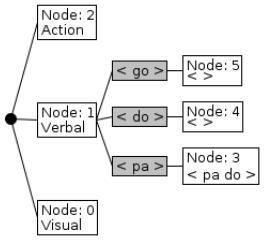
Phonological Loop

The phonological loop implemented adheres closely to the architecture proposed by Baddeley and Hitch (1974). However, we only implement the *phonological store* and omit an *articulatory rehearsal mechanism* since sub-vocal rehearsal of phonological loop content is not believed to emerge until around 7 years of age (Baddeley, Gathercole, & Papagno, 1998; Cowan & Kail, 1996; Gathercole & Hitch, 1993).

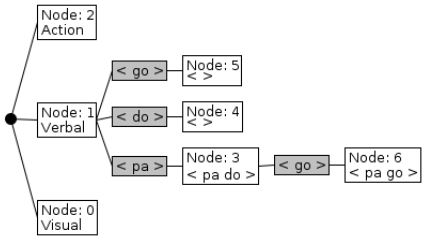
The phonological store is a first-in, first-out list structure of arbitrary length whose spaces can be occupied by, at most, one syllable (the smallest unit of phonological input specified by SAN). The debate regarding how phonological store



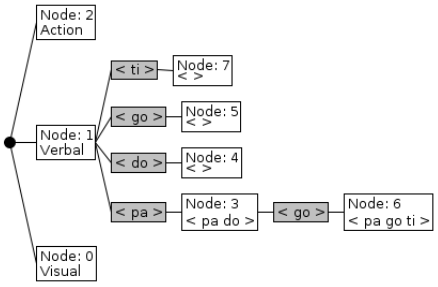
(a) Before pattern presentation.



(b) After first presentation of pattern (go unrecognised; discrimination).



(c) After second presentation of pattern (mismatch on chunk retrieved, < pa do >; discrimination).



(d) After fourth presentation of pattern (< ti > learned as primitive during third presentation but not present in chunk retrieved after fourth presentation, < pa go >; familiarisation).

Figure 1: LTM states when learning verbal pattern < pa go ti >. Tests on links are indicated by grey rectangles.

capacity is limited is noted.[3] Since this debate is unresolved and Baddeley appears to favour the trace decay explanation of phonological store capacity limitation (Baddeley, 2007), we implement this mechanism. Consequently, when syllables are added to a phonological store, they will only be retained for the trace decay value specified before they are removed.

We set the default trace decay value to 600ms according to data reported by Glass, Sachse, and Suchodoletz (2008) who estimate the duration of sensory auditory memory to be between 1 and 2 seconds in 2- and 3-year-old children; extrapolating infant trace decay to 600ms seems reasonable.

[3]The phonological store's capacity constraint may be caused by a bound on the number of items that it can store at any time or an artefact of content trace decay. See Baddeley (2007) for a review.

Experiment Simulations

A complete simulation run consists of repeating SAN's experiment battery (experiments 1 and 2), with 24 simulated participants in each experiment, 50 times. This set-up ensures that the sample sizes in our simulations are equal to the sample size of SAN, making statistical comparisons between the simulated and human data more valid.

Each experiment consisted of a 2 minute learning phase followed by a test phase. There are some notable assumptions we made when considering how to implement the experiments computationally that should be noted:

- SAN do not indicate a participant's prior familiarity with the syllables used so every simulated participant is initialised with a LTM containing nodes for each syllable used in their experiment. The alternative is to initialise simulated participants with empty LTMs; this is implausible since SAN's participants must have learned some of the syllables used in the study before they took part, since the syllables are taken from the infant's native language.

- The time taken by infants in SAN's study to initially fixate on the lights in the test phase is not specified. So, before the presentation of each word in the simulated testing phase, the simulated participant's S/LTM operations were allowed to complete and both their phonological store and STM were cleared.

During the learning phase, a learning string (randomly generated for each participant according to SAN's conditions) was used to populate the phonological store of each simulated participant. SAN report that 270 syllables are uttered every minute and a learning phase lasts for 120000ms. So, if every syllable took an equal amount of time to utter, a syllable should be uttered every 222ms: $120000 \div (270 \times 2) = 222$. Since every word is composed of 3 syllables, a word should be uttered every 666ms ($222 \times 3 = 666$). Consequently, according to our best estimate, a syllable is placed into a simulated participant's phonological store every 222ms and simulated participants constantly tried to learn the contents of their phonological store.[4]

During the test phase, the test words used by SAN in their study were presented to each simulated participant. Again, their phonological store was populated according to the timings specified in the previous paragraph. However, like SAN's study, after a whole word had been uttered, a 500ms break occurred before the next test word was presented. Simulated participants attempted to recognise the test word constantly; if recognition failed then the participant would attempt to learn the contents of its phonological store.

To ensure that the only variability in recognition times would be produced by the ordering of words in the learning phase (rather than the order of learning in the testing phase

[4]If a phonological store was empty, no attempts to learn would occur thus the simulated participant's attentional and cognitive resources were not consumed.

too), the order of test word presentation for each simulated participant was not varied over the entire simulation. We assume this consideration was also afforded by SAN; however, there is no information regarding this.

We approximated the headturn preference procedure as follows: as in SAN's study, test word presentation occurred continuously. However, unlike SAN's study, presentation ended if the simulated participant either recognised the test word or the test word was not recognised after 15 seconds from first presentation; the presentation time length was then recorded. This strategy allows for direct testing of the headturn preference procedure's proposal that novel (unrecognised) stimuli causes perceptual fixation. It also allows us to investigate whether there is evidence to posit that the participant's fixation on a speaker in SAN's study is an indicator of the participant's engagement in discrimination or familiarisation. Since learning in our simulations is analogous to a participant fixating on a speaker in SAN's study, if the presentation times recorded in our simulations are a good fit to the looking times recorded in SAN's study, this hypothesis could be accepted.

We first report results obtained with default CHREST values for discrimination (10000ms) and familiarisation (2000ms); trace decay in the phonological store is set to 600ms (as justified earlier). We then report the results of a grid search aimed at finding the optimal values for these three parameters. In 27 simulations, each repeated 50 times, simulated participants were set-up with a unique combination of values for the three parameters mentioned:

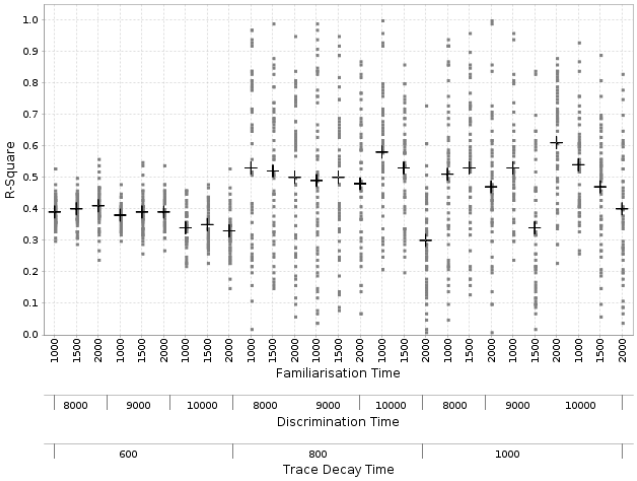
- Phonological store trace decay: 600ms, 800ms, 1000ms
- Discrimination: 8000ms, 9000ms, 10000ms
- Familiarisation: 1000ms, 1500ms, 2000ms

The mean familiar word presentation times and mean novel word presentation times for each simulated participant type repeat were used together with SAN's mean listening times for familiar and novel words in both experiments to calculate r^2 and root mean square error (RMSE) values. This data was then used to answer the following questions:

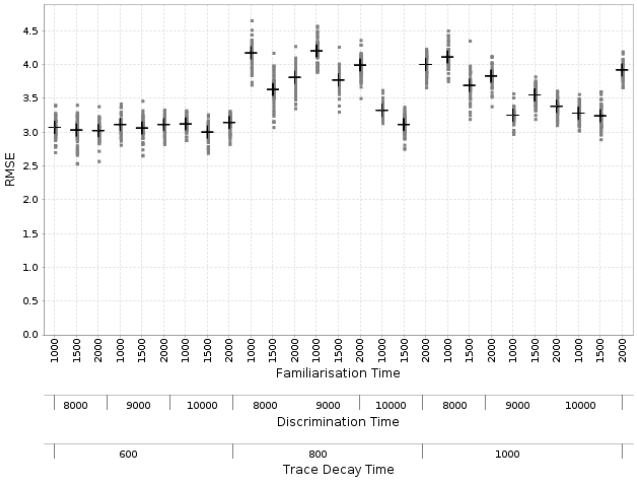
1. Do simulated participants with default times for trace decay, discrimination and familiarisation produce presentation times that are a good fit to SAN's data?
2. Do a different combination of values for these parameters offer a better fit to SAN's data?

Results

Figure 2 displays the r^2 and RMSE values calculated for each repeat of the 27 distinct combinations of the trace decay, discrimination time and familiarisation time parameters along with their means. Note that, whilst higher r^2 values indicate a better fit of the simulation data to the human data, higher RMSE values indicate a worse fit.



(a) r^2 values.

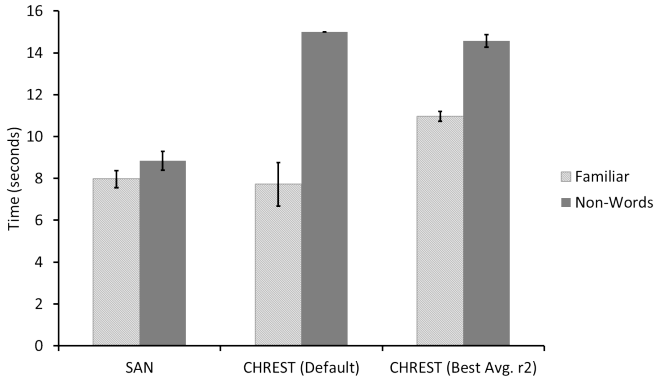


(b) Root mean squared error (RMSE) values.

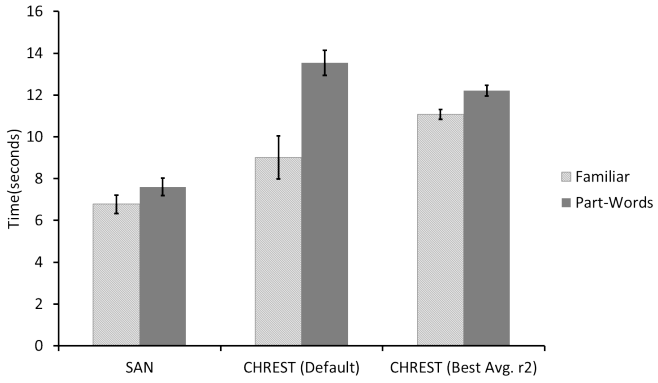
Figure 2: Model fit data for each simulated participant type's repeat experiment battery (mean values denoted by '+').

Default Parameter Model Fit

CHREST clearly discriminates between familiar words and non-words (experiment 1) and familiar words and part-words (experiment 2), as shown in Figure 3. Since 50 replications were carried out, we used Fisher's method to aggregate the p values achieved by the model with default parameter values. The resulting chi squares were highly significant: $\chi^2(100) = 1,792.8$, $p < 4.09 \times 10^{-308}$, for experiment 1, and $\chi^2(100) = 681.6$, $p < 2.35 \times 10^{-87}$, for experiment 2. However, the fit with human data was poor: r^2 = 0.33, RMSE = 3.14. The problem is that CHREST magnifies the effect with novel stimuli, compared to the human data; infants appear to become bored much more quickly than the model. Interestingly, French et al. (2011) report the same effect with the SRN model after measuring the proportion of familiar words better recognised than novel words.



(a) Experiment 1



(b) Experiment 2

Figure 3: Human listening times vs. CHREST presentation times.

Better Model Fits

When trace decay = 600ms, average r^2 values are poor and no clear pattern of r^2 improvement or deterioration emerges as trace decay values increase. The same is true when one considers how r^2 values change as discrimination and familiarisation times increase. The best r^2 on average (0.61) was obtained when trace decay = 1000ms, discrimination = 9000ms and familiarisation = 2000ms.

Conversely, with regard to RMSE values, the best fits are obtained when trace decay = 600ms. Whilst RMSE values generally deteriorate as trace decay is incremented to 800ms and 1000ms, pockets of high RMSE values are still observed. The best RMSE value on average (3.00 seconds) was observed when trace decay = 600ms, discrimination = 10000ms and familiarisation = 1500ms.

Figure 3 illustrates how the average presentation times for the simulated participants that achieve the best r^2 values on average compare to the human data. It can be seen that, whilst the general trend is well captured, the absolute values obtained by the model are still incongruent with SAN's data.

Conclusions

In this paper, we have used CHREST to simulate the listening times displayed by the infants in SAN's highly influential experiment. The key assumptions were that syllables were maintained in a phonological store and that listening time was directly mapped to the time taken to learn and recognize sequences of these syllables. Based on the previous literature, we speculated that the likely value for the trace decay time of the phonological store would be 600ms.

Simulated participants with default values for the trace decay, CHREST discrimination time and CHREST familiarisation time parameters succeeded in capturing the result that infants can significantly discriminate between familiar words and novel words (both non-words and part-words). In fact, the times achieved by simulated participants showed that differentiation between the two types of words was sharper than that observed with infants. Conversely, the sensitivity analysis provided inconsistent results: simulated participants whose trace decay is set to 600ms achieve both the best RMSE on average along with one of the worst r^2 on average. The best r^2 on average was obtained with simulated participants whose trace decay = 1000ms, discrimination = 9000ms and familiarisation time = 2000ms.

Interestingly, a trace decay time of 600ms means that a simulated participant is able to retain only three syllables or, a word using SAN's set-up, in its phonological store at any time. At the beginning of the learning and testing phase, a simulated participant will "hear" the first syllable of the first word at 222ms. This syllable will therefore decay at time 822ms, 156ms after the third syllable is heard at 666ms but 66ms before the fourth syllable (the first syllable of a new word) is heard. This buttresses statements presented by Gathercole and Adams (1993): "...in a 3-year longitudinal study of children, we found that 10% of children aged between 2 years 10 months and 3 years 1 month could already achieve a digit span of four, whereas 36% of the same cohort did not reach this level until 2 years later". If we accept that phonological store capacity is mutable and that the digit span task is a good indicator of phonological store capacity, the data obtained in this paper would lend credence to the proposal that the trace decay time of the phonological store is around 600ms for very young infants.

However, we acknowledge that the model's fit overall was low. The low fit with regard to r^2 measurements is in part explained by the fact that SAN only published 4 data points making goodness of fit comparisons difficult. In addition, it is generally accepted by developmental researchers that the headturn preference procedure is inherently noisy (Bergmann, ten Bosch, Fikkert, & Boves, 2013).

The high RMSEs observed may be explained by one of our design choices. We approximated the headturn preference procedure by assuming that presentation of the test word continued until CHREST had fully recognised it. The fact that the simulated times are too long suggests that this condition might be too strong. Another approach would be to assume that, rather than needing to have fully familiarized the LTM node image with a word in order for the word to be "recognised", it is sufficient to recognise *most* of the word

(2 syllables); this would certainly reduce the times recorded. Whether this change would produce times that fit SAN's data better can only be established by running more simulations.

Word segmentation is an essential aspect of (implicit) language acquisition, and this study has illuminated how time-limited cognitive processes, together with chunking mechanisms, mediate statistical learning, showing that transitional probability learning is not necessary. The paper's novelty was to simulate how listening times, as opposed to just the degree of recognition between familiar and novel nonsense words constructed using considerations of forward transition probabilities, differed as a function of key timing parameters associated with the phonological store and cognitive processes.

While the results of the simulations were encouraging, there is room for improvement. Future research should apply the model to more experimental results on word segmentation obtained with the headturn preference procedure, so that goodness of fit can be calculated with more data points resulting in more reliable conclusions. In particular, more accurate simulation of "head-turning" will hopefully provide data whose accuracy is much improved. This could be achieved using CHREST's simulated eye movements. We also intend to establish if a trace decay value of 600ms for the phonological store generalises to similar word segmentation studies of infants around the age of 8-months old. This is perhaps the most interesting and exciting result obtained in this paper.

References

Ambridge, B., & Lieven, E. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Anderson, B. (1999). Kohonen neural networks and language. *Brain and Language*, *70*(1), 86–94.

Baddeley, A. (2007). Working memory, thought, and action. In (pp. 35–62). Oxford University Press.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), (Vol. 8, pp. 47–89). Academic Press.

Bergmann, C., ten Bosch, L., Fikkert, P., & Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Frontiers in Psychology*, *4*(676).

Cowan, N., & Kail, R. (1996). Models of short-term memory. In S. E. Gathercole (Ed.), (pp. 29–50). Erlbaum Associates, Ltd.

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess: Heuristics of the professional eye*. Assen: Van Gorcum.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614–636.

Freudenthal, D., Pine, J. M., & Gobet, F. (2009). Simulating the referential properties of Dutch, German and English root infinitives in MOSAIC. *Language Learning and Development*, *15*, 1–29.

Gathercole, S. E., & Adams, A.-M. (1993). Phonological working memory in very young children. *Developmental Psychology*, *29*(4), 770–778.

Gathercole, S. E., & Hitch, G. J. (1993). Developmental changes in short-term memory: A revised working memory perspective. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories of memory* (pp. 189–210). Erlbaum.

Glass, E., Sachse, S., & Suchodoletz, W. (2008). Development of auditory sensory memory from 2 to 6 years: an MMN study. *Journal of Neural Transmission*, *115*(8), 1221–1229.

Gobet, F. (2015). Vocabulary acquisition. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (pp. 226–231). Oxford: Elsevier.

Gobet, F., & Lane, P. C. R. (2010). The CHREST architecture of cognition: The role of perception in general intelligence. In E. Baum, M. Hutter, & E. Kitzelmann (Eds.), *Proceedings of the 3rd conference on artificial general intelligence* (Vol. 10, pp. 7–12).

Jones, G. A., Gobet, F., & Pine, J. M. (2000). A process model of children's early verb use. In L. R. Gleitman & A. J. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Meeting of the Cognitive Science Society* (pp. 723–728). Mahwah, NJ: Lawrence Erlbaum.

Jones, G. A., Gobet, F., & Pine, J. M. (2005). Modelling vocabulary acquisition: An explanation of the link between the phonological loop and long-term memory. *Journal of Artificial Intelligence and Simulation of Behaviour*, *1*, 509–522.

Jusczyk, P., & Aslin, R. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23.

Lane, P. C. R., & Gobet, F. (2012). CHREST models of implicit learning and board game interpetation. In J. Bach, B. Goertzel, & M. Ikle (Eds.), *Proceedings of the 5th conference on artificial general intelligence* (Vol. LNAI 7716, pp. 148–157). Berlin, Heidelberg: Springer-Verlag.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*(2), 246–263.

Reber, A. S. (1993). *Implicit learning and tacit knowledge*. Oxford University Press.

Rowland, C. (2014). *Understanding child language acquisition*. London: Routedge.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Exploring the Neural Mechanisms Supporting Structured Sequence Processing and Language Using Event-Related Potentials: Some Preliminary Findings

Gretchen N.L. Smith (gsmith50@student.gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010
Atlanta, GA 30302-5010 USA

Gerardo E. Valdez (gvaldez2@student.gsu.edu),
Department of Psychology, Georgia State University, P.O. Box 5010
Atlanta, GA 30302-5010 USA

Anne M. Walk (amcclur3@illinois.edu)
Department of Kinesiology and Community Health, University of Illinois, 405 North Mathews Avenue
Urbana, Illinois 61801 USA

John D. Purdy (jdprdy@gmail.com)
Department of Psychology, Saint Louis University, 3700 Lindell Blvd.
St. Louis, MO 63108

Christopher M. Conway (cconway@gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010
Atlanta, GA 30302-5010 USA

Abstract

Structured sequence processing (SSP) refers to the neurocognitive mechanisms used to learn sequential patterns in the environment. SSP ability seems to be important for language (Conway, Bauernschmidt, Huang, & Pisoni, 2010); however, there are few neural studies showing an empirical connection between SSP and language. The purpose of this study was to investigate the association between SSP and language processing by comparing the underlying neural components elicited during each type of task. Healthy adult subjects completed a visual, non-linguistic SSP task incorporating an artificial grammar and a visual morpho-syntactic language task. Both tasks were designed to cause violations in expectations of items occurring in a series. Event-related potentials (ERPs) were used to examine the underlying neural mechanisms associated with these expectancy violations. The results indicated the P3a component elicited by the SSP task and the P600 component elicited by the language task shared similarities in their topographic distribution. These preliminary analyses suggest that the P3a and P600 may reflect processes involving detection of sequential violations in non-language and language domains, which is consistent with the idea that language processing relies on general-purpose SSP mechanisms.

Keywords: Structured Sequence Processing; Sequence Learning; Statistical Learning; Artificial Grammar Learning; Language Processing; Syntax; Event Related Potentials; P3a; P600

Introduction

Structured sequence processing (SSP), also termed sequential learning or statistical learning, is a core cognitive mechanism used to learn patterns of information from the environment over time. SSP emerges early in development (Aslin, Saffran, & Newport, 1998) and is largely automatic and implicit (Cleeremans, Destrebecqz, & Boyer, 1998), though explicit processes likely occur in parallel (Sun, Slusarz, & Terry, 2005). SSP involves learning complex embedded patterns in which each item that occurs next is determined probabilistically based on what item occurred previously (see Conway & Christiansen, 2001 for a more detailed discussion on different types of sequential learning).

A key facet of SSP is that it serves as a tool for making predictions about which elements will occur next in a sequence (Christiansen, Conway, & Onnis, 2012). When sequential patterns are learned, this information can be used not only to generate expectancies about upcoming stimuli in the sequence, but also to detect when stimuli deviate from expectation (Ferdinand, Mecklinger, & Kray, 2008). Bar (2007) suggests that a "circular mechanism" occurs in which the brain limits processing of stimuli that are predictable, while allotting cognitive resources to stimuli that are novel and/or unexpected. These "predictive processing" operations are generally beneficial to many aspects of cognition, including perception, movement, decision-making (Bubic, von Cramon, & Schubotz, 2010) and language (Federmeier, 2007).

SSP appears to be especially important in the domain of language (Conway, Bauernschmidt, Huang, & Pisoni, 2010). In particular, SSP may support knowledge and use of grammatical language, such as word order (Conway et al., 2010), phonology (Saffran, 2003), morphology and syntax

(Ullman, 2004). However, the association between SSP and language has largely been assumed. Only recently have behavioral studies demonstrated empirical links between SSP and natural language processing (e.g., Conway et al., 2010). Additionally, even fewer studies have empirically compared these two mechanisms at a neural level (e.g., Patel, Gibson, Ratner, Besson, & Holcomb, 1998). In a neural-based investigation using a within-subject design, Christiansen et al. (2012) examined the electrophysiological responses elicited during a visual SSP task and a visual syntactic natural language processing task. The findings indicated both the SSP task and the natural language task elicited a late positive-going deflection in voltage potential—a P600 component—that has been linked with the processing of syntactic violations (Lelekov, Dominey, & Garcia-Larrea, 2000). Furthermore, topographic maps of the P600 effects showed similar distribution between conditions (Christiansen et al., 2012). Overall, these results provided some of the earliest direct, within-subject empirical evidence that the same neural mechanisms may be used for SSP and syntactic natural language processing (Christiansen et al., 2012).

However, more direct neural evidence of a link between SSP and natural language processing is needed, using different types of SSP and language tasks. Therefore, the purpose of this study was to investigate the relation between SSP and natural language processing by comparing the electrophysiological profiles elicited during each type of task, using a within-subject design. The SSP task was designed to resemble an artificial grammar learning (AGL) paradigm (Reber, 1967), in which complex statistical regularities are embedded in the sequences. One key aspect of the SSP task used in this study is that it is more purely non-linguistic in nature than previous SSP tasks using language-like stimuli (e.g., Christiansen et al., 2012) or stimuli that are readily verbalizable and easily mapped onto vocalizations (e.g., Patel et al., 1998). Evidence showing that similar neural responses are elicited during a more fundamentally non-linguistic SSP task and a natural language task would provide additional—and possibly more compelling—support that language processing is based in part on mechanisms utilized to extract and encode structured sequential information in a general-purpose manner.

It is possible our visual non-linguistic SSP task and our morpho-syntactic visual natural language processing task would both elicit a P600, similar to the Christiansen et al., (2012) findings. The authors of that study hypothesized that the P600 might reflect processing broadly involved with making predictions about upcoming items in a series, which is not confined solely to language (Christiansen et al., 2012). It is also possible our SSP task might elicit ERP components that have been associated with extraction and encoding of non-linguistic structured sequential patterns. Previous studies have suggested that the N200 [negative-going deflection, occurring approximately 200 milliseconds (ms) after stimulus onset] and P3b (positive-going deflection occurring approximately 300 ms after stimulus onset) components are elicited in sequential learning paradigms and may reflect the processing of expectancy violations (e.g, Carrión & Bly, 2007). Additionally, the P3a (positive-going deflection occurring approximately 250 ms after stimulus onset) has been evoked from "novel" stimulus paradigms (Courchesne, Hillyard, & Galambos, 1975), has been linked with the recognition of grammatical violations in a second language (Jakoby, Goldstein, & Faust, 2011), and has been associated with focused attention (Comerchero & Polich, 1999).

Given the present study was exploratory, we expected to observe any of the ERP components mentioned above. Consequently, the central hypothesis was simply that violations in a non-linguistic SSP task and violations in a morpho-syntactic natural language task would elicit similar electrophysiological response profiles.

Method

Subjects

Forty-three subjects (ages 18-22; 25 female) participated. All subjects were recruited from Saint Louis University, were native speakers of English, with normal to corrected-to-normal vision and who, at time of testing, reported no history of hearing loss, difficulty with speech, or history of cognitive, perceptual, or motor disorder.

Experimental Paradigm

Measures of SSP and language were administered separately in a single test session. All subjects performed the measure of SSP first and the measure of language second.

Measure of SSP The measure of SSP was similar to the "Simon" visual-spatial SSP task used in previously published work (see Conway et al., 2010 for details). In this measure, subjects viewed sequences of 4 black squares appearing one at a time on a white background in 1 of 4 possible quadrants (upper left, upper right, lower left, lower right) (See Figure 1 below). The task was to reproduce each sequence immediately following presentation by touching the squares in the correct order on a touchscreen. Unknown to subjects, the measure of SSP consisted of two parts: a learning phase and a test phase, which differed in the types of sequences presented.

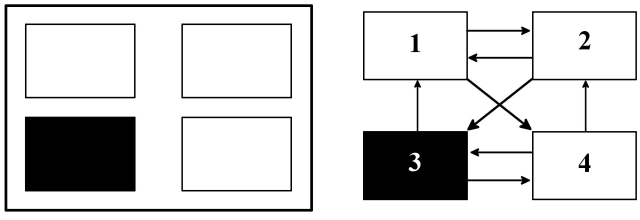


Figure 1: SSP Task (on left) with rule structure (on right).

In the learning phase, sequence elements were generated according to an underlying artificial grammar that specified the probability of a particular element in a sequence occurring given the preceding element (see Figure 1 above). For each sequence, the starting element (1-4) was randomly determined, and then the grammar was used to determine each subsequent element, until the full sequence length was reached. For example, given the starting element 3, the element 2 had a zero probability of occurring next, while the 1 and 4 elements had an equal (50%) chance of occurring. No element could follow itself in the sequence. The mapping of the rules to the locations was randomly determined for each subject; however, for each subject, the mapping remained consistent across all trials. All sequences were 5 elements long.

The learning phase consisted of 40 grammatical sequences. The phase began with a blank screen that appeared for 1 second. In the first part of the learning phase (20 sequences), each element in the sequence was displayed for 600 ms, followed by a 200 ms pause in which nothing was displayed on the screen. The final element in the sequence was followed by a 200 ms pause before the whole 2x2 grid of squares was displayed with a "Done" button in the middle. Using the touch screen, the subject then reproduced the sequence just presented, followed by pressing the word "Done". Immediate feedback was given as to the correctness of each response. The second part of the learning phase (20 sequences) was the same as the first part, except each element in the sequence was displayed for 400ms, followed by a 200 ms pause.

The test phase consisted of 64 sequences. One fourth of the sequences were "grammatical-trained" (i.e., the sequences were identical to grammatical sequences presented in the learning phase), one fourth were "grammatical-untrained" (i.e, new sequences sharing the same underlying structure as the other grammatical sequences), and one half were "ungrammatical" (i.e., the sequences violated the grammar). For the ungrammatical sequences, the starting element (1-4) was randomly determined, then any element could occur next in the sequence except repeating elements were not allowed. The timing of the test phase was identical to the timing used in in the second half of the learning phase. The subjects were not told that this was a test phase or that there were different types of sequences (grammatical-trained, grammatical-untrained, and ungrammatical). From the perspective of the subject, the test phase was the same reproduction task they had been doing all along.

Scoring for the Measure of SSP For the test phase, a sequence was scored as correct if a subject correctly reproduced it. A score of 5 was given for each correctly reproduced sequence and was based on the length of each sequence. As in previous studies (e.g., Conway et al., 2010) a learning score was then obtained by subtracting the total score for the ungrammatical sequences in the test phase from the grammatical sequences in the test phase. A higher learning score indicates better performance on structured sequences compared to ones that violate the structure, suggesting that successful SSP occurred.

Measure of Language Sixty-four sentences were presented in the measure of language. These sentences varied according to whether or not they contained grammar violations. Thirty-two sentences were grammatically correct, and 32 sentences contained morpho-syntactic violations pertaining to verb agreement with the subject. For example, "The famous singer walks onto the stage." was a grammatically correct sentence used in the task and "The famous singer walk onto the stage." was a sentence with a violation. Each sentence began with a white fixation point (+) and was presented 1 word at a time in white text on a black screen. Each element (word or fixation point) appeared for 400 ms and was followed by a blank screen for 400 ms. Thirty-two 7-word sentences and 32 8-word sentences were presented. The 32 8-word sentences each contained an auxiliary verb. The target words (grammatical or violation) always occurred at the 4th word in the 7-word sentences and at the 5th work in the 8-word sentences. Target words were always verbs. The ratio of grammatically correct sentences to sentences with morpho-syntactic violations was 1:1, for each sentence length.

The phrase "Was that a good or a bad sentence? Press 1 for good, press 2 for bad." appeared on the screen immediately following presentation of the final word in the sentence. The task was to make a keypad response on button 1 if the sentence was "good" and to respond on button 2 if the sentence was "bad." Subjects were not given explicit instruction as to what "good" or "bad" meant (i.e., that grammatical sentences were "good" and sentences with violations were "bad), nor were they told that some sentences were grammatical and some had violations. No feedback was given. A 1-second pause was given between a response and the presentation of the next sentence.

Scoring for the Measure of Language A sentence was scored correct if the subject made the correct grammaticality judgment for that sentence (i.e., a button press on "1"/ "good" for grammatical sentences; a button press on "2"/ "bad" for sentences with morpho-syntactic violations).

Expectancy Violations for Both the Measure of SSP and the Measure of Language

Both measures were designed to cause violations in expectations of items occurring in a series (i.e., a violation of the learned sequence in the SL task and a violation of grammar in the language task). Event-related potentials (ERPs)—portions of ongoing electroencephalogram (EEG) time-locked to cognitive events of interest—were used to associated with these expectancy violations.

EEG/ERP Data Acquisition and Preprocessing

EEG was recorded during the test phase of the measure of SSP and throughout the measure of language using a 128-channel high-density sensor net with vertex recording reference (Electrical Geodesics, Eugene OR). Standard sensor net application techniques were followed. Recordings were made using NetStation acquisition software (Electrical Geodesics, Inc.), with a 0.1–100Hz bandpass filter and digitized at 250 Hz. Electrode impedances were kept below 50 kiloohms. Rest breaks were given as needed.

ERP for the measure of SSP was time-locked to the presentation of a stimulus that violated the artificial grammar and was compared to a stimulus in a similar position in a sequence that was grammatical. ERP for the measure of language was time-locked to the presentation of a word in the sentence that violated the morpho-syntactic grammar and was compared to a word in a similar position in a sentence that was grammatical.

Data was preprocessed using Netstation (Electrical Geodesics, Inc.). The continuous raw EEG recording was filtered through a 0.1 Hz high pass filter and a 30 Hz low pass filter. Channels were marked bad for a given trial if blinks or eye movements were detected, if amplitudes >150 μV, if the channel was flat (had zero variance), or if manual inspection suggested noise specific to that channel. Channels marked bad were interpolated in the raw EEG from data measured at nearby electrodes. After exclusion of artifacts, the continuous EEG was segmented into epochs in the interval -200 msec to +1000 msec with respect to the onset of the target stimulus (i.e., violation of grammar, for both the measure of SSP and the measure of language). Data were not re-referenced from the vertex channel.

Data from 3 subjects was excluded from analysis due to bad EEG channels that were either too high in number or too clustered together. Data from 8 subjects was excluded due to poor data quality or missing data. Therefore, data from a total of 32 subjects was analyzed.

Regions of Interest

Nine regions of interest (ROI) were defined for data analysis, with each containing 9 channels: frontal (FRz), central (CNz), posterior (POz), left anterior (LAn), left central (LCn), left posterior (LPo), right anterior (RAn), right central (RCn), and right posterior (RPo) (see Figure 2 below).

Results

Behavioral Average Task Performance for the Measure of SSP and the Measure of Language

Average accuracy given in percentage correct for reproduction of the three sequence types presented in the measure of SSP was as follows: 73% (grammatical-trained), 71% (grammatical-untrained), and 70% (ungrammatical). These three scores were not significantly different from one another (p=.999). Average accuracy given in percentage correct for the grammaticality judgment of the sentences presented in the measure of language was 93%.

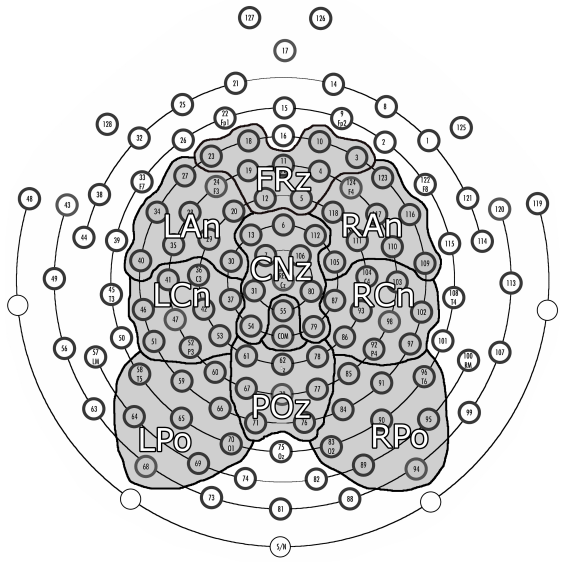


Figure 2: 2-D layout of the 128-channel sensor net (top is front). For data analysis, the channels were grouped into 9 regions of interest (outlined above), each consisting of 9 channels.

Electrophysiological Response Elicited by the Measure of SSP

Visual inspection indicated a P3a component for ungrammatical sequences relative to both types of grammatical sequences in several ROI. Paired samples t-tests were conducted on the grand-averaged mean amplitude waveforms associated with the P3a component 270-330ms after the sequence violations, with significant effects in the central (CNz) [t(31)=3.968, $p<.001$], frontal (Frz) [t(31)=3.321, $p=.002$], and right anterior (RAn) [t(31)=2.303, $p=.028$] regions [See Figure 3 below for an example of the P3a effect in the frontal (FRz) region].

Correlations were computed between the grand-averaged mean amplitude waveforms associated with the significant P3a effects for ungrammatical sequences relative to both types of grammatical sequences and the averaged learning score on the measure of SSP (grammatical-ungrammatical). Results showed that learning score was significantly negatively correlated with the P3a effect in the frontal (FRz) region [r(31)= -.300, p=.05].

Electrophysiological Response Elicited by the Measure of Language

Visual inspection indicated a P600 component for ungrammatical sentences relative to grammatical sentence in several ROI. Paired samples t-tests were conducted on the mean amplitude waveforms associated with the P600

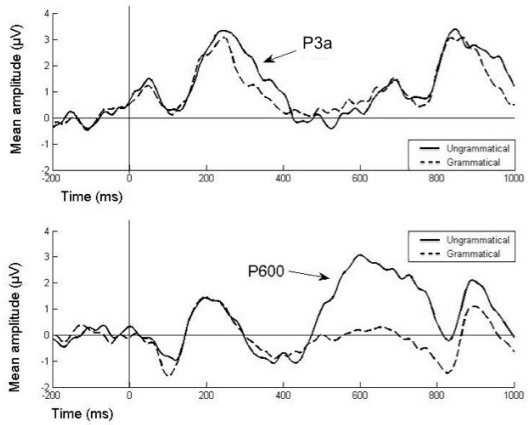


Figure 3: P3a component in the frontal region and P600 component in the central region.

component 525-925ms after the syntactic violations, with significant effects in the central (CNz) [$t(31)=6.616$, $p<.001$], right posterior (RPo) [$t(31)=2.880$, $p=.007$], left posterior (LPo) [$t(31)=5.360$, $p<.001$], right anterior (RAn) [$t(31)=7.610$, $p<.001$], left anterior (LAn) [$t(31)=2.902$, $p=.007$], and left central (LCn) [$t(31)=3.433$, $p=.002$] regions [See Figure 3 above for an example of the P600 effect in the central (CNz) region].

Comparison Between the Electrophysiological Responses Underlying SSP and Language

Topographical maps were created from ungrammatical minus grammatical difference waves associated with the P3a and P600 described above, to compare the scalp distribution of electrical activity for both components in the two tasks (See Figure 4 below). Visual inspection showed some similarities in the topographic profile for the two components; specifically, the early phase of the P600 (525-580ms) resembles the full P3a component. A correlation was computed between the difference waves (ungrammatical-grammatical) between the early phase of the P600 and the full P3a component. The results showed a positive correlation in the right anterior (RAn) region between the P3a (270-330ms) and the early P600 (525-580ms) that approached significance [r(32)=.291, p=.106].

Discussion

The present study provided some initial analyses comparing the electrophysiological responses elicited in a visual, non-linguistic SSP task and a visual morpho-syntactic natural language processing task. Following exposure to sequences that followed an embedded artificial grammar, subjects showed a P3a to violations of the grammar, while showing a P600 to morpho-syntactic violations in a visual natural language processing task. The P600 elicited in the language task is consistent with findings from paradigms that involve the processing of syntactic violations. The P3a elicited in

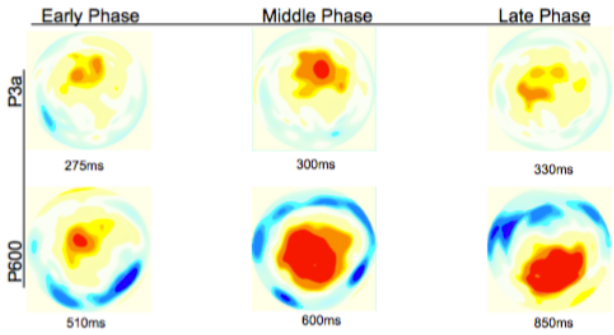


Figure 4: Topographic maps showing distribution of the P3a and P600 components, measured at the scalp.

the SSP task is consistent with previous AGL ERP studies showing a P300 component in response to violations of a grammar (e.g., Opitz, Ferdinand, & Mecklinger, 2011). The presence of the P3a could be due to violations of the artificial grammar drawing attentional resources, corresponding with Bar's (2007) "circular mechanism" of predictive processing previously described. The negative correlation between the SSP learning score—an indication of better performance on grammatical sequences—and the P3a effect for ungrammatical sequences relative to grammatical sequences supports this notion.

Although averaged behavior performance on the measure of SSP showed no learning effect as a group, the correlation between the SSP learning score and the ERP effect for the ungrammatical sequences suggests that even though an overall group learning effect was not observed, there is a distribution of learning scores that seem meaningful, with some individuals showing learning and others not showing learning. Future work will investigate whether individual differences in cognitive processes such as attention and working modulate learning of structured sequences.

One interpretation of the topomaps findings is that the P3a and P600 are distinct components, yet both reflect processes involving the detection of sequential violations in artificial and natural grammar processing tasks, respectively. It has been previously suggested that P300 and P600 components may both reflect the processing of incongruent information in different types of tasks (e.g., Christiansen et al., 2012). A similar role for both types of components, therefore, suggests some degree of overlap between SSP and language processing mechanisms.

The early phase of the P600 showed a distribution similar to the full P3a elicited in the SSP task, suggesting that they may have a common neural origin. Although the distribution of the full P3a is more confined than the full P600, this could be due to the relative unfamiliarity and limited exposure to the SSP task, whereas the wider scalp distribution of the P600 might be the result of the extensive and prolonged exposure humans have had with language. With more exposure to non-linguistic SSP tasks, the P3a elicited to violations in the artificial grammar might show

robustness similar to the P600 elicited to violations in natural language grammar. On the other hand, the differences in the ERP correlates observed for the two tasks could instead be due to differences in the types of violations inherent to each (something akin to word order violations in the AGL task but subject-verb agreement violations in the natural language task).

Coulson, King, and Kutas (1998) noted similarities between the P600 and P3b elicited by manipulations to probability and saliency in an oddball paradigm. They concluded the P600 might be a part of the P300 family of components (Coulson et al., 1998). Still, similarity in distribution may not reflect similarity in neural generators [see Osterhout & Hagoort, 1999 (response to Coulson et al., 1998) for a more detailed discussion on cautions of attempting to determine similarity of neural mechanisms from EEG recorded at the scalp]. To help address this limitation, we are using source localization analyses to examine whether the components elicited in our two tasks share a common neural origin. Although preliminary, these findings suggest the possibility that the neurocognitive mechanisms involved in detecting sequential violations in a non-linguistic AGL task are similar to those involved in detecting morpho-syntactic violations in natural language, with the P3a possibly being an earlier version of the P600 or a reduced variant of it.

Acknowledgements

N. Buchholz, J. Daltrozzo, J. Deocampo, G. Frishkoff, E. Hilvert, S. Pardasani, B. Rickles, K. Ross, G. Signiski, S. Singh, and R. Town. Funding: NIH R01DC012037.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321 324.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*, 280 289.

Bubic, A., von Cramon, D. Y. & Schubotz, R. I. (2010) Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 1-15.

Carrión, R.E. & Bly, B.M. (2007). Event-related potential markers of expectation violation in an artificial grammar learning task. *NeuroReport, 18,* 191-195.

Christiansen, M.H., Conway, C.M**.**, & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes*, *27*, 231-256.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*, 406 416.

Comerchero, M.D. & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology 110*(1), 24–30.

Conway, C.M**.**, Bauernschmidt, A., Huang, S.S., & Pisoni, D.B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356-371.

Conway, C.M**.**, & Christiansen, M.H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*, 529-546.

Coulson, S., King, J.W., Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*(1), 21-58.

Courchesne, E., Hillyard, S.A., Galambos, R. (1975). Stimulus novelty, task relevance and the visual evoked potential in man. *Electroencephalography and Clinical Neurophysiology*, *39*(2), 131-43.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491-505.

Ferdinand, N. K., Mecklinger, A., & Kray, J. (2008). Error deviance processing in implicit and explicit sequence learning. *Journal of Cognitive Neuroscience*, *20*, 629 642.

Jakoby, H., Goldstein, A., & Faust, M. (2011). Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment. *Psychophysiology*, *48*(11), 1517-1531.

Lelekov, T., Dominey, P. F., & Garcia-Larrea, L. (2000). Dissociable ERP profiles for processing rules vs instances in a cognitive sequencing task. *NeuroReport*, *11*(5), 1129-1132.

Opitz, B., Ferdinand, N. K., & Mecklinger, A. (2011). Timing matters: the impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience, 5*(8), 1-9.

Osterhout, L., & Hagoort, P. (1999). A superficial resemblance does not necessarily mean you are part of the family: Counterarguments to Coulson, King, and Kutas (1998) in the P600/SPS-P300 debate. *Language and Cognitive Processes*, *14*, 1-14.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, *10*, 717-733.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Behavior*, *6*, 855-863.

Saffran, J. R. (2003). Statistical learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110-114.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual process approach. *Psychological Review*, *112*(1), 159-192.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231-270.

Bifurcation analysis of a Gradient Symbolic Computation model of incremental processing

Pyeong Whan Cho (pcho4@jhu.edu)

Paul Smolensky (smolensky@jhu.edu)
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218 USA

Abstract

Language is ordered in time and an incremental processing system encounters temporary ambiguity in the middle of sentence comprehension. An optimal incremental processing system must solve two computational problems: On the one hand, it has to keep multiple possible interpretations without choosing one over the others. On the other hand, it must reject interpretations inconsistent with context. We propose a recurrent neural network model of incremental processing that does stochastic optimization of a set of soft, local constraints to build a globally coherent structure successfully. Bifurcation analysis of the model makes clear when and why the model parses a sentence successfully and when and why it does not—the garden path and local coherence effects are discussed. Our model provides neurally plausible solutions of the computational problems arising in incremental processing

Keywords: Harmonic Grammar; Gradient Symbolic Computation; neural networks; incremental processing; parsing; dynamical systems theory; bifurcations

Introduction

Sentence comprehension requires building a hierarchical constituent structure from sequentially presented words. Psycholinguists agree that human language comprehension is incremental, which means the processing system builds up an interpretation with partial input before a whole sentence is presented. An optimal[1] incremental processing system must solve two closely related computational problems: (a) it must be able to keep multiple possibilities consistent with the information that has been processed, say context. (b) At the same time, it has to reject the interpretations inconsistent with context which may be consistent with bottom-up input.

For clarity, consider a finite language consisting of four sentences {"cats sleep", "cats yawn", "dogs sleep", "dogs yawn"}. A first word (e.g., "cats") of a sentence creates temporary ambiguity which can be resolved after processing a second word (e.g., "sleep") of the sentence. An optimal system must be able to consider both grammatical continuations without choosing one over the other after processing the first word. Simultaneously, the system must be able to reject impossible interpretations before reading a second word (e.g., "sleep") such that it does not take a wrong interpretation (e.g., "dogs sleep").

In this study, we propose a neural network model of incremental processing and show (a) how the proposed model can solve two computational problems in a principled way and (b) when and why the model fails to build a target structure. It will turn out that the model's failure looks like either the garden path effect (Frazier, 1987; Frazier & Rayner, 1982) or the local coherence effect (Tabor, Galantucci, & Richardson, 2004) both of which are well established in the psycholinguistics literature.

The remainder of the paper is organized as follows. First, we briefly describe a phrase structure grammar of our interest. Second, we present an overview of the model; summarize the results without technical details; and discuss the implication of the model. The technical details of the analysis are presented in "Bifurcation analysis of the GSC model", which is followed by a conclusion.

A Grammar of Interest

In the present study, we investigate the Gradient Symbolic Computation (GSC) model (Smolensky, Goldrick, & Mathis, 2014) that implements a phrase structure grammar G, consisting of four rewrite rules: G = {S1 $\rightarrow$ A B, S2 $\rightarrow$ A C, S3 $\rightarrow$ D B, S4 $\rightarrow$ D C} where S1, S2, S3, S4 are the starting symbols. Grammar G generates a finite language presented in the introduction. We designed the simple language for the following reasons: (a) the language shares the same computational problems with more complex languages. (b) The GSC model implementing a more complex grammar is more difficult to analyze. The simplicity of the grammar makes it possible to investigate the model thoroughly. (c) Parsing a sentence of the language looks trivial but can be difficult for continuous-time recurrent neural network models that utilize highly local constraints and do not explicitly monitor complete structures.

Overview of the Model and Results

The GSC model is a brain-like, continuous-time continuous-state dynamical systems model that can gradually build discrete symbolic structures. There is a transparent mapping between the symbolic structures and the activation patterns (vectors in a continuous vector space) (Smolensky, 1990) so it allows us to study the nature of the intermediate states in the middle of sentence comprehension.[2] The GSC model does stochastic optimization of constraints implementing the

[1] Human language processing may not be optimal due to performance issues. In this study, we investigate the competence of the proposed model.

[2] A popular connectionist alternative is the Simple Recurrent Network (Elman, 1990; Frank, 2009). Although it has been successful in modeling human sentence comprehension performance, the model works in discrete time and develops a rather hard-to-understand internal representation.

grammar and those implementing a force *quantization* pushing towards those neural states that encode fully discrete symbolic trees (for details, see Smolensky et al., 2014). The model maximizes Total Harmony, a measure of goodness of an activation state: $H(\mathbf{a},\gamma) = H_G(\mathbf{a}) + H_B(\mathbf{a}) + \gamma H_Q(\mathbf{a})$ where $\mathbf{a}$ is an activation state and γ (≥ 0) is quantization strength. Grammar Harmony H_G measures how much a state satisfies grammatical constraints and is defined by principles (Hale & Smolensky, 2006). Bowl Harmony H_B measures how close a state is to a baseline activation state. Quantization Harmony H_Q, roughly speaking, measures how close a state is to discrete symbolic states.[3]

In incremental processing, we assume: (a) the model reads the first word of a sentence at 0 of γ. (b) γ increases monotonically in time without assuming any specific function. (c) The model reads the second word at a certain value of $\gamma\,(=\gamma_c)$ and then γ continues to increase. The topic of the paper is a formal understanding of how γ's temporal trajectory results in correct parsing, garden path, or local coherence errors.

We performed numerical bifurcation analysis of the model (for a brief introduction, see Meijer, Dercole, & Oldeman, 2009)—the details are presented in the next section. The results suggest that there is a range of γ_c values ($\gamma_1 < \gamma_c < \gamma_2$) that guarantees accurate parsing (case 1). If the model reads the first word (e.g., 'A') of a sentence (e.g., 'A B') too long such that $\gamma_c > \gamma_2$, the model randomly chooses one structure (e.g., [$_{S2}$ A C]) over the other (e.g., [$_{S2}$ A B]) before it reads the second word, althouth both structures are consistent with the first word. It is a garden path error (case 2). If the model reads the second word too quickly ($\gamma_c < \gamma_1$), the model does not reject the structures (e.g., [$_{S3}$ D B] or [$_{S4}$ D C]) inconsistent with the first word (e.g., 'A'). Thus, when the model reads the second word at $\gamma_c (< \gamma_1)$, sometimes the model fails to build the target structure because a non-target structure (e.g., [$_{S3}$ D B]) is consistent with the second word ('B') and it is still considered by the model as a possible grammatical structure. We argue it is an local coherence error (case 3).

We argue our model has a potential as a processing model of incremental structure building. Our model provides neural/mathematical solutions of the computational problems arising in incremental processing; we know when and why the model parses a sentence correctly and when and why it does not. Unlike other constraint satisfaction models (Spivey & Tanenhaus, 1998; Tabor & Hutchins, 2004; Vosse & Kempen, 2000), the GSC model is contructed in a principled way using Harmonic Grammar (Prince & Smolensky, 1997) and tensor product representation (Smolensky, 1990). Its dynamics can be understood thoroughly not with simulations but based on principles of dynamics (see the next section)—the GSC model is not a blackbox. Unlike probabilistic symbolic models of sentence comprehension (Hale, 2001; Levy, 2008), the GSC model describes how the state changes in continuous time and proposes the solutions of computational problems at the algorithmic level. More importantly, the GSC model is not an implementation of structural probabilistic models where the structural hypothesis space is discrete. A blend state in the GSC model—an intermediate state that is located between the states encoding fully discrete structures—is not the representation of a probability distribution across discrete symbolic structures.

The GSC model proposes an interesting way of keeping past and predicting the future. In the GSC model, the present (*blend* state) contains past. Unlike the memory-based model proposed by Lewis, Vasishth, and Van Dyke (2006), there is no need of retrieval processes.[4] Given a word input, the present contains the future as well. In the next section, it will be shown that the GSC model travels along a one-dimensional manifold (a subspace of the full representation space) which can evolve to multiple grammatical structures consistent with both top-down context and bottom-up word input.

The proposed GSC model is not complete. Processing difficulty in sentence comprehension is typically measured in reading times rather than in parsing accuracy in behavioral experiments but the model does not make a prediction on word reading times directly.[5] First, we point out that psycholinguistic models such as the garden path model (Frazier, 1987) and the unrestricted race model (Traxler, Pickering, & Clifton Jr., 1998) propose longer reading times indicate the revision of the initial parse that turns out to be wrong. At this point, the GSC model does not have the ability to revise its interpretation. Parsing failures (cases 2 and 3) overviewed in this section (see also paths [2] and [3] in Figure 4) should be interpreted as the *initial* parsing failure rather than the final product. If we assume the reanalysis requires more time, there is a close relationship between the accuracy of *initial* parsing and the reading times. One way to implement the revision of the initial interpretation is to allow the model to reduce γ when it detects the present blend state is inconsistent with the bottom-up input—probably by detecting a sudden decrease in Grammar Harmony. Second, slow reading times may not necessarily suggest the revision of the initial parse. For example, in the constraint-satisfaction models (MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey & Tanenhaus, 1998; Tabor & Hutchins, 2004), processing delay is observed when multiple interpretations compete with each other without parsing failure. In the GSC model, there may

[3] $H_G(\mathbf{a}) = 0.5\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a} + \mathbf{b}^{\mathrm{T}}\mathbf{a} + \mathbf{ext}^{\mathrm{T}}\mathbf{a}$; $H_B(\mathbf{a}) = -0.5\beta||\mathbf{a}-\mathbf{z}||^2$; $H_Q(\mathbf{a}) = -0.5\sum_{r\in R}(\sum_{f\in F} a_{f,r}^2 - 1)^2 - 0.5\sum_{f\in F, r\in R} a_{f,r}^2(a_{f,r}-1)^2$ where $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and the bias vector which are constructed by Harmonic Grammar. $\mathbf{a}$ is an activation vector, $\mathbf{ext}$ is an external input vector, $a_{f,r}$ is an acitvation value of a filler/role binding (Smolensky, 1990), F and R are the index sets of fillers and roles, $\mathbf{z}$ (= 0.5 $\mathbf{1}$) is the baseline activation vector, β is the strengh to pull a state to $\mathbf{z}$ and was set to 10. H_Q implements a constraint that only one filler must occupy a role.

[4] We do not argue retrival processes do not involve in human sentence comprehension. The GSC model suggests retrieval may not be the core process in incremental structure building.

[5] The model has a potential to make reading time predictions but at this point, we do not know with what information the model can make a decision when to read a next word. The model may be able to make such decisions by monitoring γ.

be a manifold which takes longer to travel along than other manifolds.

We conclude the GSC model can provide a different approach to incremental processing and may improve our understanding of the process.

Bifurcation analysis of the GSC model

Background

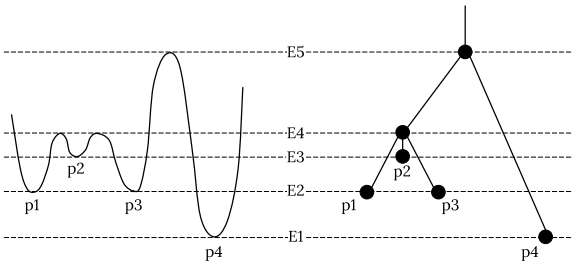


Figure 1: An arbitrary energy landscape (energy = -Total Harmony) and a disconnectivity graph constructured from the energy landscape. p1-p4 indicate the local minima and E1-E5 (E1 $< \cdots <$ E5) indicate the energy levels.

Consider an arbitray energy function $E(x,p)$ in which x is a state variable and p is a parameter. The left side of Figure 1 presents an arbitrary energy landscape (energy = -Total Harmony) when p is set to a certain value. Now consider a system that minimizes the energy function. When no noise is assumed, the system rolls down the hills of the energy landscape (i.e., gradient descent): $dx/dt = -dE(x,p)/dx = f(x,p)$. At a local minimum or a local maximum, the system will not change anymore because the gradient vanishes there. Those states x^* where $f(x^*,p) = 0$ are called *equilibrium points* (or *fixed points*). We can check what happens if a system is slightly displaced from an equilibrium point x^*. If the system approaches x^*, the state (e.g., the local minima p1-p4) is asymptotically stable and often called an *attractor*. If the system moves away from x^*, the equilibrium state (e.g., the local maxima) is unstable.

If the system is high dimensional, it is not easy to visualize the energy landscape. In that situation, we can focus on the topology of the energy landscape by constructing a disconnectivity graph (Wales, Miller, & Walsh, 1998; Becker & Karplus, 1997). The terminal nodes of the disconnectivity graph (the right side of Figure 1) correspond to the attractors (local minima of the energy landscape, or equivalently local maxima of the harmony surface). The nonterminal nodes indicate the height of a ridge (called *energy barrier*) between two local minima. The magnitude of noise required to cross over an energy barrier is determined by the barrier's height (Chiang, Hwang, & Sheu, 1987). For example, the height of the energy barrier from p1 to p2 (= E4-E2) is greater than the height of the energy barrier from p2 to p1 (= E4-E3), suggesting there is a certain level of noise at which the p2-to-p1 transition is possible while the p1-to-p2 transition is not.

The equilibrium points may be created, destroyed, or change their stability when a parameter's value changes. Such qualitative change in the dynamics of the system is called *bifurcation* (Strogatz, 1994) and the parameter values at which bifurcations happen are called *bifurcation points*. When a bifurcation occurs, the topology of the disconnectivity graph changes. In this stuy, we use the continuation method (Meijer et al., 2009) to discover the equilibrium points at different parameter values and detect bifurcations. At the same time, we construct a disconnectivity graph from the set of equilibrium points detected at a particular parameter value to visualize the toplogy of the energy landsacpe. The combination of the two techniques allows us to understand the dynamics of the GSC model thoroughly.

Method

The GSC model maximizes Total Harmony $H(\mathbf{a};\gamma,\alpha)$, or equivalently minimizes energy $E(\mathbf{a};\gamma,\alpha) = -H(\mathbf{a};\gamma,\alpha)$ in which $\mathbf{a}$ is a 27-dimensional activation state vector, γ (≥ 0) is the quantization strength parameter, and α is a variable indicating whether the model is reading the first ($\alpha = 0$) or the second ($\alpha = 1$) word of a sentence 'A B'.[6]

First, we used the continuation method to discover the one-dimensional manifold of equilibrium points in the 28 dimensional vector space (27 state variables + 1 parameter γ) when the first word 'A' was presented (i.e., $\alpha = 0$). To use the continuation method, we should know the equilibrium points at a particular γ value. In the GSC model, it is not difficult. When $\gamma = 0$, we can ignore nonlinear quantization dynamics and at the setting, the GSC model is a linear dynamical system which was constructed to have a single global optimum. The equilibrium point was discovered by the Newton's method. Alternatively, the global attractor can be discovered by integrating the differential equations $\nabla_{\mathbf{a}} H(\mathbf{a},\gamma,\alpha)$ numerically. Once an equilibrium state was discovered, we used the continuation method to discover the one-dimensional manifold of equilibrum points.

We iterated the process with every attractor discovered when $\gamma = 200$. When γ is very large, the equilibrium points are mostly determined by Quantization Harmony. The function was designed to have the attractors at a subset of the vertices of the unit hypercube $[0,1]^{27}$. The actual attractors must be close to those vertices. By applying the Newton's method, we could find all the attractors when $\gamma = 200$. Besides the attractors, there are many unstable equilibrium points at the γ value and finding those points is much more difficult. Instead of trying to find those unstable fixed points at the paramter setting, we used the continuation method to discover the unstable equilibrium points, assuming that the unstable equilibrium points are connected to the stable equilibrium points on the manifold of the equilibrium points when γ is allowed to

[6] Given that all four sentences of the language are symmetric, we focus on the case in which the model is reading a sentence 'A B' and investigate when and how the model can build the target structure [$_{S1}$ A B] successfully.

change. If we start to follow the manifold from a stable equilibrium point, we will reach an unstable equilibrium point.

The numerical bifurcation analysis with the continuation method was performed by using a MATLAB package, the Continuation Core and Toolboxes (COCO)[7]. After performing bifurcation analysis, we chose specific γ values (γ = 20, 25, 35, 70) and constructed a disconnectivity graph at each γ value to investigate the topology of the energy landscape.

We did not perform bifurcation analysis and construct the disconnectivity graphs when the model reads the second word 'B' ($\alpha = 1$). Instead, the equilibrium points and the topology of the harmony landscape were inferred, based on symmetry, from those discovered from numerical bifurcation analysis when the model reads the first word 'A'.

Result

The continuation method discovered 795 branches of the equilibrium points including 1589 equilibrium points at 200 of γ when $\alpha = 0$.[8] Most branches are not relevant in our discussion of the model dynamics so we focus on a small number of important branches (see Figure 2). First, look at the branches (1), (2), and (5). A saddle-node bifurcation occurs at $\gamma_1 \approx 21.3$ and a subcritical pitchfork bifurcation occurs at $\gamma_2 \approx 59.6$ (for introduction, see Strogatz, 1994). Second, look at the branches (3) and (4) which evolve to the states representing [$_{S3}$ D B] and [$_{S4}$ D C] both of which are inconsistent with the first word 'A'. Those branches are disconnected from the other branches. We suspect the branches evolving to the states representing grammatical/ungrammatical structures inconsistent with the input word are disconnected from the major branch evolving to the states representing the grammatical structures consistent with the input.

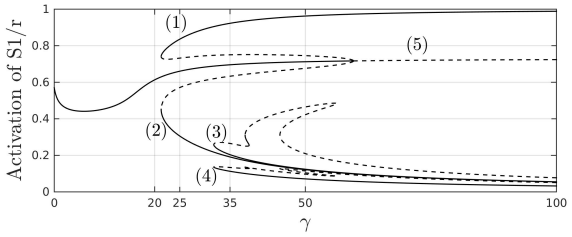


Figure 2: The bifurcation digram. The solid and dashed lines indicate stable and unstable equilibria, respectively. Only the branches evolving to (1) [$_{S1}$ A B], (2) [$_{S2}$ A C], (3) [$_{S3}$ D B], (4) [$_{S4}$ D C], and (5) a blend of S1 and S2 are presented. The vertical lines indicate the γ values at each of which a disconnectivity graph was constructued (see Figure 3).

[7]http://sourceforge.net/projects/cocotools/

[8]The COCO could not finish continuation successfully along 96 of 891 branches. All those cases occurred when the software switched to a new branch at a branch point (where multiple branches intersect each other) and tried to follow a differnet branch; the branch before switching was connected to an ungrammatical structure (e.g., [$_{C}$ D _]) at 200 of γ. Those cases need further investigation but those branches are not important in the discussion of the model.

From Figure 2, we can predict the state change while the model is reading the first word 'A'. Recall that γ is assumed to increase in time. The system has a single global attractor in the region where $\gamma < \gamma_1$. Thus, regardless of the initial state, the system will reach the global equilibrium point before γ passes γ_1. At γ_1, two more attractors emerge from a saddle-node bifurcations but they are separated from the equilibrium point along branch (5) where the system is. The system keeps following the major branch until γ passes γ_2 at which a subcritical pitchfork bifurcation occurs and the equilibrium point loses its stability. Thereafter, even with a very small noise, the system moves to either branch (1) or (2).

Figure 3 shows the topology of the energy landscape at four different γ values (see also Figure 2). When $\gamma = 20$ (Figure 3a), there is a single attractor. When $\gamma = 25$ (Figure 3b), there are three local optima on branches (1), (2), and (5). If the system has followed the major branch, the system will be at the attractor on branch (5) represented by the leftmost terminal node. The attractor is separated from other attractors on branches (1) and (2) by energy barriers. When $\gamma = 35$ (Figure 3c), the energy landscape has more attractors but all newly emerged attractors are separated from the three attractors on branches (1), (2), and (5) by high energy barriers. The bottommost part of the graph inside the square has the same local structure as Figure 3b. The blend state on branch (5) is still stable so the system will be at the equilibrium point. Figure 3d presents the disconnectivity graph when $\gamma = 70$. The graph is very complex and not easy to read. The squred region at the bottom is magnified in the third panel of the top row in Figure 4. At that time, the equilibrium point on branch (5) is not stable any more. The system moves to the branches of stable equilibrium points, either (1) or (2).

Recall that each word in the language is consistent with two grammatical structures and inconsistent with the other two grammatical structures. Given the symmetry, the equilibrium states when the model reads the second word 'B' ($\alpha = 1$) can be inferred from the equilibrium states when the model reads the first word 'A'. More specifically, when the model reads the second word 'B', the system forms a different set of branches of equilibrium points which has the same structure as shown in Figure 2. At this time, the equilibrium points along branch (2) evolves into the state representing [$_{S3}$ D B] and the equilibrium points along branch (5) evolves into the blend of the blend of [$_{S1}$ A B] and [$_{S3}$ D B]. Equilibrium points on branch (1) still evolves into the state representing [$_{S1}$ A B].

Now consider what will happen if the model reads the second word 'B' at γ_c. Figure 4 shows three qualitatively different cases in incremantal processing.

Case 1 (see path [1] in Figure 4): The model reads 'B' when $\gamma_1 < \gamma_c < \gamma_2$; γ_c = 35 in the current example. While γ increases from 0 to γ_c, the state quickly approaches a global attractor and follows the major branch (see branch (5) in Figure 2). The state change is indicated by the arrow (annotated as [1]/[2]) from panel (a) to panel (b) in Figure 4. When the model reads 'B' at γ_c = 35, the harmony landscape changes

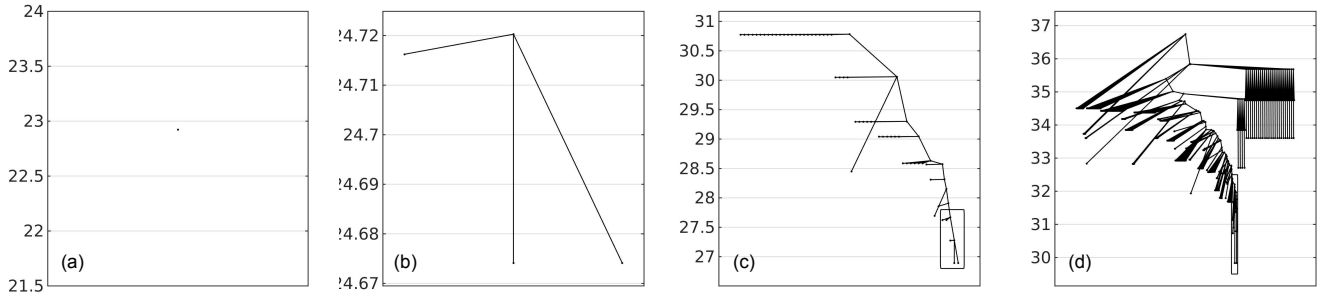


Figure 3: Topological change in the harmony landscape (given a first word 'A') as γ increases (γ = 20, 25, 35, 70)

abruptly. The previously stable equilbrium state becomes unstable so the state moves to a local optimum when $\alpha = 1$. It turns out the previous equilibrium point is located inside the basin of attraction of an equilibrium point along a branch corresponding branch (1) in Figure 2. The state change is indicated by the arrow (annotated as [1]) from panel (b) to panel (e). As γ increases further, the new equilibrium point evolves to the state representing the target structure [$_{S1}$ A B]; see the arrow (annotated as [1]) from panel (e) to panel (f). When small noise is assumed, the model builds the target structure with certainty.

Case 2 (garden path; see path [2] in Figure 4): The model reads 'B' when $\gamma_c > \gamma_2$; $\gamma_c = 70$ in the example. As in Case 1, the system reaches the global attractor and follows the manifold of the global attractors in the beginning. However, when γ passes $\gamma_2 (\approx 59.6)$, the equilibrium point on branch (5) loses its stability. With a small noise, the system moves to a new equilibrium point, either one on branch (1) or one on branch (2); see the arrows from panel (b) to panel (c). As suggested in Figure 3, the system cannot move to other attractors because they are separted by high energy barriers. When the model reads the second word at $\gamma_c = 70$, the harmony landscape changes abruptly. The attractors (1) and (2) in the top right panel are located in the basins of attraction of the new attractors (1) and (2) in the bottom right panel, respectively. The state will be at either (1) or (2) in the bottom right panel depending on where the system was (see the arrows from panel (c) to panel (f). It suggests if the system was following branch (2) while processing the first word, the system will build a non-target structure [$_{S2}$ A C] with the second word input, although it is inconsistent with the bottom-up input. This case is like the garden path effect in that the system chose one possibility over the others when it encounters ambiguity.

Case 3 (local coherence; see path [3] in Figure 4): The model reads 'B' when $\gamma_c < \gamma_1$; $\gamma_c = 20$ in the example. As in Case 1, the system reaches the global attractor and follows the manifold of the global attractors in the beginning. When the model reads 'B' at $\gamma_c = 20$, the state moves to a new global optimum; see the arrow [3] from panel (a) to panel (d). As γ increases, the system follows the major branch of equilibrium points (see the arrow from panel (d) to panel (e)) until γ passes γ_2 when the system moves to either a branch leading to the target structure or a branch leading to a non-target structure [$_{S3}$ D B] (see the arrows [3] from panel (e) to panel (f)). The latter case is an extreme version of the local coherence effect; the system failed to use top-down information (context provided by the first word 'A') and relied only on the bottom-up information to choose the locally coherent but globally incoherent structure.

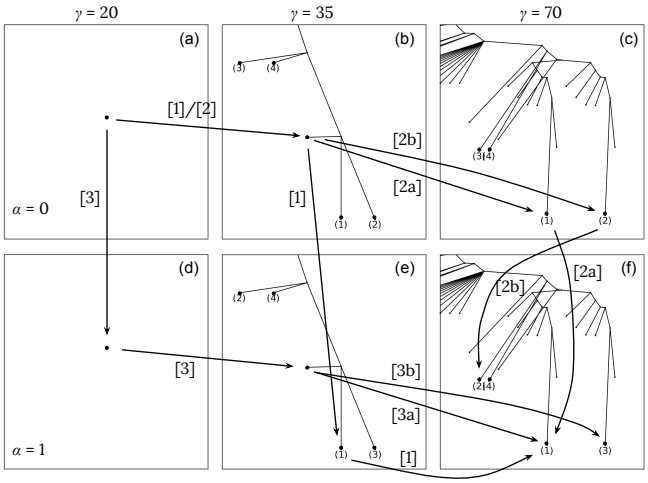


Figure 4: Parsing accuracy depends on γ_c. The arrows indicate the state change expected while the model processes a sentence 'A B' ([1] $\gamma_c = 35$, [2] $\gamma_c = 70$, [3] $\gamma_c = 20$). The panels in the second and third columns correspond to the regions included in the panels (c) and (d) in Figure 3.

Conclusion

In this study, we investigated how the GSC model handles temporary ambiguity to parse a sentence successfully in incremental processing. The model considers all structures consistent with context (the words that the model has processed) without choosing one over the others by being at a *stable blend* state which can evolve into the possible grammatical structures, suggesting the usefulness of the continuous representation space. We point out that the blend state does not represent a disjunctive set of those structures so the GSC model is not an implementation of the probabilistic models. On the other hand, the model rejects the structures

inconsistent with the input by developing high energy barriers between attractors.

More importantly, the model explains when and why the model fails to parse a sentence correctly. When γ increases too quickly, the model chooses one interpretation over the other possible ones. When γ increases too slowly, the model fails to separate globally incohrerent interpretations from globally coherent ones. The present study suggests that accurate parsing in the GSC model requires an optimal control of quantization strength γ.

Acknowledgments

We thank Geraldine Legendre, Akira Omaki, Kyle Rawlins, Ben Van Durme, Colin Wilson, Laurel Brehm, Nick Becker, Belinda Adam and especially Matthew Goldrick for their contributions to this work, and acknowledge the support of NSF INSPIRE grant BCS-1344269.

References

Becker, O. M., & Karplus, M. (1997). The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, *106*(4), 1495–1517.

Chiang, T., Hwang, C., & Sheu, S. (1987). Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, *25*(3), 737–753.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1139–1144).

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and Performance XII: The Psychology of Reading* (pp. 559–586). Lawrence Erlbaum Associates.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics.

Hale, J., & Smolensky, P. (2006). Harmonic Grammars and harmonic parsers for formal languages. In P. Smolensky & G. Legendre (Eds.), *The harmonic mind: From neural computation to optimality-theoretic grammar. Volume I: Cognitive architecture* (pp. 393–416). The MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703.

Meijer, H., Dercole, F., & Oldeman, B. (2009). Numerical bifurcation analysis. In R. A. M. Ph.D (Ed.), *Encyclopedia of Complexity and Systems Science* (pp. 6329–6352). Springer New York. (DOI: 10.1007/978-0-387-30440-3_373)

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, *275*(5306), 1604–1610.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159–216.

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: a framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102–1138.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521–1543.

Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Westview Press.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 431–450.

Traxler, M. J., Pickering, M. J., & Clifton Jr., C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558–592.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*(2), 105–143.

Wales, D. J., Miller, M. A., & Walsh, T. R. (1998). Archetypal energy landscapes. *Nature*, *394*(6695), 758–760.

Coalescing the Vapors of Human Experience into a Viable and Meaningful Comprehension

Tomer D. Ullman (tomeru@mit.edu)
Department of Brain and Cognitive Science, MIT

Max Siegel (maxs@mit.edu)
Department of Brain and Cognitive Science, MIT

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Science, MIT

Samuel J. Gershman (gershman@fas.harvard.edu)
Department of Psychology, Harvard University

Abstract

Models of concept learning and theory acquisition often invoke a stochastic search process, in which learners generate hypotheses through some structured random process and then evaluate them on some data measuring their quality or value. To be successful within a reasonable time-frame, these models need ways of generating good candidate hypotheses even before the data are considered. Schulz (2012a) has proposed that studying the origins of new ideas in more everyday contexts, such as how we think up new names for things, can provide insight into the cognitive processes that generate good hypotheses for learning. We propose a simple generative model for how people might draw on their experience to propose new names in everyday domains such as pub names or action movies, and show that it captures surprisingly well the names that people actually imagine. We discuss the role for an analogous hypothesis-generation mechanism in enabling and constraining causal theory learning.

Clerk: Occupation?
Comicus: Stand-up philosopher. I coalesce the vapors of human experience into a viable and meaningful comprehension.
Clerk: Oh, a bullshit artist!
(*History of the World, Part 1*. Dir. Mel Brooks)

Introduction

How do people come up with new concepts, causal models or theories, to make sense of their world? Whether it is scientists trying to explain the natural world with formal theories of physics or chemistry, or children and adults trying to explain their experience with common-sense theories such as folk biology or folk psychology, the question of how fundamentally new ways of thinking unfold over time and change in response to evidence has long been of interest to cognitive science (Carey, 2009; Gopnik & Wellman, 2012; Schulz, 2012b).

This paper is about the more everyday aspects of this process: New thoughts of on-the-fly explanations and causal models to make sense of everyday problems and puzzles. The difference between radical conceptual change and prosaic concept generation is one of degree, like the difference between coming up with a general explanation for how objects balance and fall – a difficult process that may take months or years (Baillargeon, 2008) – and coming up with an off-the-cuff explanation for why soda tastes fizzy, or a plausible name for a new action movie.

While everyday creativity is not categorically different from scientific creativty, the specific challenge of explaining creativity in everyday idea generation has been pointed out by Schulz (Magid, Sheskin, & Schulz, 2015; Schulz, 2012a). Schulz called for both empirical and theoretical research on the issue. On the theoretical side, everyday creative thought poses a problem for computational models that have been developed to capture more radical theory change (Goodman, Ullman, & Tenenbaum, 2011; Ullman, Goodman, & Tenenbaum, 2012). Briefly put, these models see humans as reasoning over a hypothesis space ("theory space") that contains explanations and concepts. People do not know in advance what the right concepts and theories are, so they must stochastically search through these large (potentially infinite) spaces, adjusting and discarding their concepts as they go. If a newly proposed concept or theory better fits the data, and is more likely under a general prior favoring such things as 'simplicity', then that newly proposed concept will be accepted.

Such a process might make sense for theory-change that takes years to unfold, but surely (Schulz argues) it is underconstrained when it comes to everyday thought. The search spaces are too large, and the process does not take into account the specific constraint of a task. For example, suppose a friend asks you to come up with a name for their hip new Thai restaurant, located near a cluster of start-ups. You may never have been faced with such a problem before, but after some thought you might come up with "Thai-Tech". It is not a particularly good name, but it is at least a relevant one. It is better than "Mummified Ragdoll Fifteen" or "Croatian delight", or any of the other infinite combinations of words that language affords. Such proposals would not just be rejected if proposed, they would not be thought of in the first place.

That people do not waste time with nonsense (in the sense of proposing gibberish answers used above) seems almost trivial, and suggesting a concept-production algorithm that runs through all combinations in the English language for any given task seems like a clear non-starter. And yet, many stochastic search algorithms for hypothesis generation suffer from exactly this problem. How can such algorithms be amended to not consider 'obviously wrong' proposals without actually proposing them first?

The challenge of everyday thinking is the challenge of reducing hypothesis spaces quickly and on-the-fly, in response to a task that might never have been considered before. The reduced hypothesis spaces might still be large, and they might still contain 'bad' ideas and concepts. But these concepts and ideas would at least be relevant to the task. They would be capable of being wrong, in the sense that a human would recognize them as a bad or good response to the task, as opposed

to completely unconnected.

In this paper we take up the computational challenge of everyday reasoning, and propose a structured approach for narrowing hypothesis spaces by inferring a generative model over relevant examples taken from memory. The resulting "Bounded-Space" model (or, BS model, for short) can generate reasonable proposals for learning or problem solving within a domain, but it does not exempt a thinker from stochastically searching within the reduced space and evaluating the proposals. Ideas generated in the reduced space may be bad ones. They may even be, for lack of a more polite term, mostly bullshit. But by quickly cutting down the space of possible thoughts from 'mostly nonsense' to, at worst, 'mostly bullshit', everyday thinking and learning can be powerfully and usefully constrained.



Figure 1: A space of possible names for things. Only a small subregion contains names relevant for a given task (dark area, in this case pub names). A stochastic search over the entire space would result in many examples that are not even bad.

We develop and test our BS model using the relatively simple task of coming up with new names for items, specifically movies of different genres and pubs (this is similar to a task proposed in Magid et al., 2015). We compare its proposals to those of people, and to real world data. We find that the basic BS model fares worse than people, but is already within a reasonable range. A cognitively plausible extension to the BS model that evaluates a small number of samples before selecting one is surprisingly consistent with the range of people's novel name generation behavior. The BS model provides an initial proposal for capturing everyday thought, but extending it beyond the (relatively) simple task of proper names to tasks such full-blown explanations is not trivial, a point that we take up in the discussion.

Modeling on-the-fly thought

Generating new thoughts and concepts can be seen as a stochastic search through a hypothesis space, guided by some metric of success or quality (Ullman et al., 2012). Ideally, the hypothesis space should be large enough to encompass any possibly correct or useful theory or concept. Infinite hypothesis spaces that fit this bill can be easily specified through grammars over conceptual primitives (Piantadosi, Tenenbaum, & Goodman, 2012), and in principle it is possible to stochastically search through such spaces by proposing and rejecting amendments to the current hypothesis. But the problem of everyday thinking suggests that these proposals must be strongly constrained by abstract domain knoweldge, such that most of the proposals that can potentially be considered (a-priori of any data or constraint) will never be considered, and our actual proposals focused efficiently on candidates that have some hope of being useful.

To see the problem more clearly, try to come up with a name for a new pub. Perhaps you never faced such a task, but presumably you can do it with some degree of success. Such a thought process *could* be implemented by a stochastic search through all the possible phrases in the English language, but this would lead to a ridiculous proportion of not only bad proposals, but completely irrelevant proposals. In Figure 1 we consider a fictional space of all possible names for things. Only a tiny portion of that space can even be called 'Pub Space'.

Assuming a thinker (such as yourself) was never asked to come up with a new pub name before, how can they quickly reduce the space of all possible new names to just ones relevant for pubs? Presumably you thought of a better pub name than "Towards a Better Designed Belgian Office", and if someone posed that as a suggestion you could evaluate it as a bad proposal. [1] But "Towards a Better Designed Belgian Office" is a potentially good answer to a different question. Again, the issue is that *before* any particular question, puzzle or problem is posed, we wish to have a large search space that can generate many possible concepts, thoughts and solutions. But *after* a particular task is set, we wish to restrict the space to only the relevant solutions. How can we know ahead of time what counts as a 'relevant' solution, without first proposing it and evaluating it?

We propose a method (Bounded-Space, or BS) for constructing new, relevant concept spaces on-the-fly, illustrated through the pub example in Figure 2. For any particular task the thinker first draws from memory several examples that match the desired concept (Figure 2.1). For example, if asked to come up with a new pub name, the thinker might first draw some known pub names from memory. We assume that relevant examples are available in memory, even if they are relevant in a broad sense. "Broad sense" here means that the examples match the general structure of the task. For example, if the task is coming up with an explanation for why the Roman Empire fell, relevant examples can include causal explanations in general ("State changes in X can be caused by an outside Y"), rather than particular reasons why the Roman Empire fell ("Crises of legitimacy"). Without any relevant examples that come to mind the question itself is a non-starter, e.g. "Can you ganoosh a new Floop?". The initial retrieval of relevant examples might rely on associative memory, but is a problem outside the scope of this paper.

The thinker then uses inverse inference, conditioned on the

[1] Or rather, as "not even bad', just ridiculous.

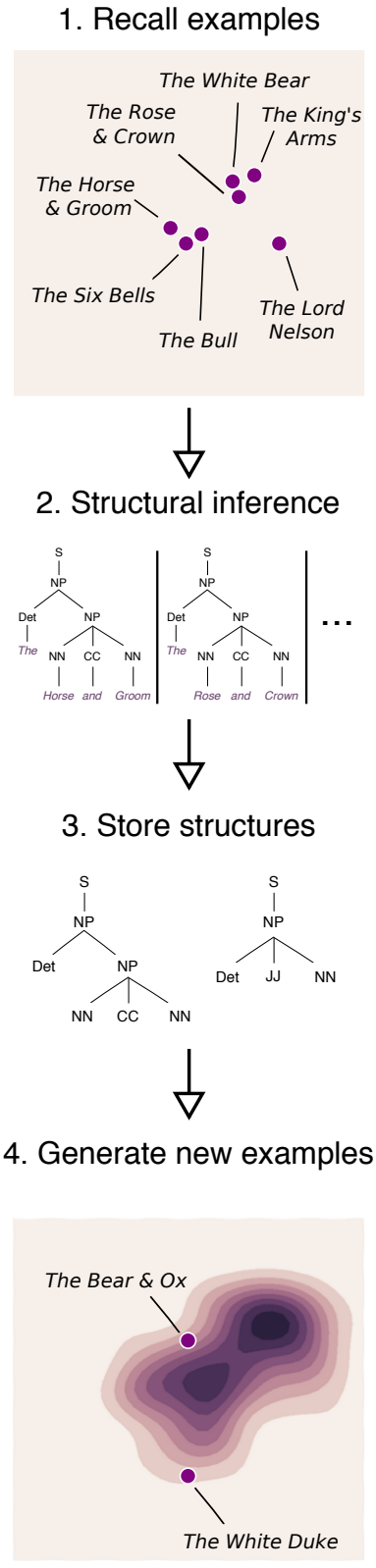


Figure 2: An illustration of the BS model applied to the domain of pub names.

examples, to construct a generative model that can produce these examples. Such a thinker might focus on the form of the examples, noticing for instance that all the examples share a certain abstract causal structure. In our particular case study, which is restricted to new names for pubs and movies, we consider a grammatical analysis of the examples from memory (Figure 2.1), although many other structural constraints are possible in the general case. For example, the thinker would notice that "The Rose and Crown" share the same syntactic structure as "The Horse and Groom", namely "The [Noun] and [Noun]". The thinker then stores the relevant underlying structure without the particular terminals (Figure 2.3). This database of structures is a proxy for a generative model over the space containing the examples.

In order to come up with a new example (Figure 2.4), the thinker chooses a structure in proportion to the number of times it was used in any of the examples from memory. For instance, the thinker might generate from "The [Noun] and [Noun]". When reaching any particular terminal (e.g. [*Noun*]), the speaker either reuses an appropriate part from a stored example (e.g. any of the nouns in the examples) or a semantically similar part ("Ox" rather than "Horse"). In this paper we use high-dimensional word embeddings via Global Vectors (Pennington, Socher, & Manning, 2014) to implement a similarity space. Word-vectors that are close together in this space (in a Euclidean or cosine distance sense) tend to be semantically similar in psychological tasks. At the end of this process, the thinker might for instance come up with "The Bear and Ox" as a plausible pub name.

It is important to stress that this model only implements the proposal stage of a conceptual search process. An additional evaluation step – not modeled here – is necessary to accept or reject a particular proposal. There may be additional constraints that one considers in the evaluation, such as the catchiness of the name. "The Bear and Ox" might be a terrible name for a pub for any number of reasons under additional evaluation, but at least it appears reasonable, as opposed to "A Floral and Tasty Essence Wrapped in Chamomile"[2]. The role of the BS proposal model is to get the thinker within a space where good and bad proposals can be evaluated, rather than having to spend the majority of the time with non-starters.

Under the assumption that an evaluation function exists, it is possible to extend the model so that the thinker draws a number of examples at a time, and only reports one of them. This is natural if the thinker must provide an example or series of examples, and cannot choose to simply refuse to provide an example even if it is evaluated as poor. We consider such an extension by having the model draw k samples (BS-k), and choosing among them in proportion to their quality (provided by the assumed evaluation function), relative to the total quality of the sample. If $k = 1$, we recover the original BS model. The k parameter has the psychological interpretation of 'the number of examples people draw and evaluate internally before reporting a single answer'. In order to compare this model to people, we consider the task of coming up

[2]Which is better as an example response for 'make up a wine review'.

with new names for things, specifically movies of different genres and pubs.

Experiment

Participants, materials and methods

Two groups of participants, Producers ($N = 40$, 18 female, median age 29) and Raters ($N = 50$, 16 female, median age 33), were recruited through Amazon's Mechanical Turk service and paid a small monetary sum for their participation.

The Producers were asked to come up with 5 new names for 4 different movie genres (action, horror, comedy and romance give 20 movie names in total per participant), and 5 new names for pub names. Producers entered their responses using a free-form text field. The Raters were asked to rate names for different categories on a 1-5 scale ("Very Bad" to "Very Good"). Names were category specific, meaning a particular question might be "How good is the title *Parade of Bullets* as a name for a action movie?"

In order to construct different names, we used an equal mixture of names from the Production experiment, names from real instances of the category, and names from the BS model. To keep the task manageable for Raters, we used 25 names from each source (Production, Real, BS) per category, creating 375 names in total. Each Rater saw half of these names, such that each name was rated by 25 raters.

The Production names were selected by randomly choosing 5 Producers for each category (we chose this method, rather than randomly selecting from all the production data, in order to asses whether Producers come up with better or worse names as their guesses progress). The real names were selected by randomly choosing among all Wikipedia entries for that category (for movies) and from a list of popular pub names (for pubs). The model names were selected by first choosing examples at random from the Wikipedia entries for movie genres, and a list of popular pub names. There were on the order of 1,000 examples for each movie genre, and 260 pub names.[3] These examples were syntactically parsed using a variant of the NLTK package in Python (Bird, Klein, & Loper, 2009), and a library of syntax trees was built for each genre. Syntax trees were then chosen in proportion to how often they appear in the examples, and terminals in each tree were chosen from the appropriate part of speech as it appears in the examples, or randomly replaced with a semantically similar part of speech using GloVe (Pennington et al., 2014) with probability $P_{\text{replacement}} = 0.2$. Once examples are available, generating a new BS example takes less than a second.

Results

Participants rated both the Production names (those made up by people) and the Real names (those taken from Wikipedia) similarly: the mean ratings were 3.12 for Production and 3.07 for Real. People rated the BS model names lower on average (mean rating was 2.57, difference is significant at $p < 0.001$). The BS-k model with $k = 5$ achieves a mean rating of 2.99, which is still statistically different from the average rating for Production and Real names, but the difference is now much smaller (we expand on why we chose $k = 5$ below).

Figure 3 shows the comparison in more detail, as a distribution over the ratings from 1 to 5. A χ^2 test shows the distributions for the Production names and the Real names are not distinguishable, while the BS and BS-5 models are highly distinguishable from both and from each other ($p < 0.0001$). Table 1 illustrates examples of high quality, low quality and average names for different genres.

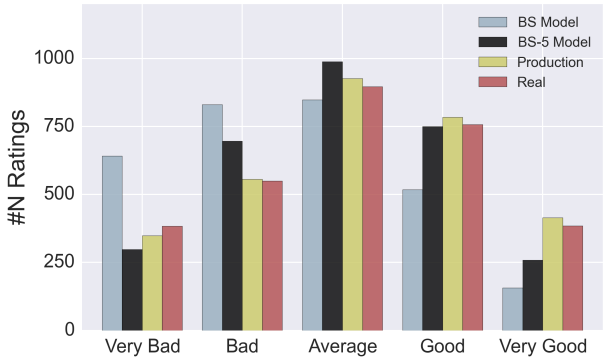


Figure 3: The distribution of ratings for different names, by the source of the name.

The ordinary BS model is not as successful at producing new examples as people, but it is not far. As mentioned, the idea of the BS model is to construct a reasonable space to sample from, as a first step towards in a propose-evaluate cycle. It is likely that people are also considering bad names but not reporting all of them, and this evaluation step is not captured by the BS model. The BS-k extention of the model uses the rating of the model names as a proxy for an evaluation step, considering k samples at a time and choosing one in proportion to its ratings relative to the total rating in the k samples. If $k = 1$, we recover the original model. But as k grows larger, we come closer to the rating distribution of the Production and Real names (Figure 4), where 'closer' means a lower Kullback-Leibler divergence (KL), between the distribution of quality scores for the names produced by the BS-k model and the Production or Real names. As Figure 4 indicates, considering only around 3 to 7 samples at a time can make a marked improvement to the BS model. Larger values of k provide diminishing returns, and so for the analyses here we considered $k = 5$. Note that the limiting behavior of increasing k is *not* to produce higher and higher rated names, but rather to converge on a steady-state distribution reflecting a two-stage generate-and-evaluate loop for new names. It is striking how quickly this distribution converges to that of the

[3] A more psychologically plausible version of example recall would use likelihood sampling over the space of movies, and would require asking a different group of subjects to recall actual movies in response to genre prompts.

names made up by people (or the Real names), suggesting that the BS-k model represents at least a plausible first guess for how people come to produce the new names they do.

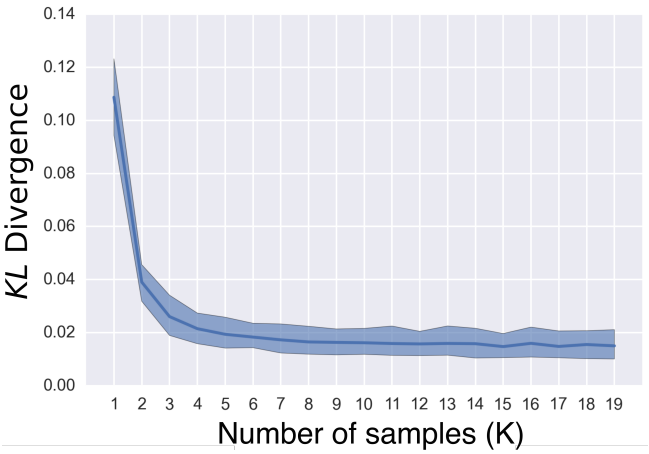


Figure 4: Kullback-Leibler divergence (*KL*) between the distribution of ratings over model names and Production names, for different values of *k* in the BS-k model. Error bars show 95% confidence intervals.

Discussion

There is something magical about everyday thought, and something odd about how people can respond quickly and reasonably to new questions they never heard with answers no one else has heard. A child could take years to come up with an intuitive biology that unites trees and animals into one concept (Carey, 2009), but a 5-year-old can answer 'What makes the wind?' at the speed of thought. And even if their answer is wrong it might be amusing ("The trees waving their arms make the wind"), rather than absurd ("The wind blows because the moon is bigger than a bag of Doritos").

Such magical stuff deserves hard-nosed experimental scrutiny (Magid et al., 2015), and a better computational answer than "People search randomly through all possible concepts and explanations and evaluate each candidate by how well it explains the data" (Ullman et al., 2012). Here we considered one particular proposal for constructing a reasonable hypothesis space on the fly by using a structural analysis of examples that match the task at hand. This BS model presupposes that thinkers have a way of carrying out such a structural analysis (in our particular case, we assumed a thinker can use grammar to recognize the structural similarity between instances, using "Saving Private Ryan" and "Chasing Amy" to construct a general "VERBing PROPER-NAME" movie schema).

The BS model serves as just one part in a propose-evaluate search, where the evaluation step is difficult to capture and can involve a large amount of world knowledge. Still, for the limited task considered it appears to produce reasonable-sounding new examples for a given category. Within the language domain, the model can potentially be extended beyond short titles to include such things as wine reviews ("The finishing notes are not supported by the light body"). Outside language, a similar approach to idea generation can take advantage of other structures used to organize thoughts. For example, a causal-explanation BS model could analyze the underlying Bayes-net structure of examples from memory, and use those to propose new explanations from a common structure (say, $X \wedge Y -> Z$, the Roman Empire fell because of a combination of tribal invasions and Jewish thought[4]). Applying this approach to generate good hypotheses for causal learning and intuitive theory formation is, we hope, a promising next step – perhaps wrong, but at least not ridiculous, as an account of where learners' new concepts come from.

Acknowledgments We are grateful to Laura Schulz for stimulating discussions. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Baillargeon, R. (2008). Innate ideas revisited: For a principle of persistence in infants' physical reasoning. *Perspectives on Psychological Science*, *3*, 2–13.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. " O'Reilly Media, Inc.".

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological review*, *118*(1), 110.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*, 1085.

Magid, R. W., Sheskin, M., & Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development*, *34*, 99–110.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *EMNLP*, *12*, 1532–1543.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*, 199–217. doi: 10.1016/j.cognition.2011.11.005

Schulz, L. (2012a). Finding new facts; thinking new thoughts. *Rational constructivism in cognitive development. Advances in child development and behavior*, *43*, 269–294.

Schulz, L. (2012b). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*, 382–389.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455–480.

[4] http://courses.washington.edu/rome250/gallery/ROME%20250/210%20Reasons.htm

	Rating	BS Model	Production	Real
ACTION	High	The Raging V (3.7) Phoenix First (3.4)	War Hawks (4.1) Retribution (3.9)	Bloodfist II (3.9) Marked for Death (3.8)
	Average	Drop (3.0) Conspiracy Mars (2.7)	Stunt Man (3.1) Run Faster (3.1)	Blue Steel (3.2) Bullet in the Head (3.1)
	Low	Of Art (1.9) Teenage Chinese (1.5)	The State of Iraq (2.6) Eat It (2.1)	I Come in Peace (2.3) Curry and Pepper (1.5)
HORROR	High	The Dawn (3.5) The Prophecy on One (3.4)	The House of Death (4.2) Cabin of the Dead (4.0)	The Dunwich Horror (4.1) Murders in the Rue Morgue (4.0)
	Average	Sharks (2.7) Seed 5 (2.4)	Silence (3.1) Blood Draws (3.0)	The Headless Eyes (3.3) The Wizard of Gore (3.0)
	Low	Space Vegas (1.9) Family (1.5)	Paid Maidens 3000 (2.1) Cat Napping (1.7)	Count Yorga, Vampire (2.7) Mephisto Waltz (2.2)
COMEDY	High	Love Punch Drunk (3.0) All American Rita (2.8)	Hot Mess (3.5) Tame Your Own Shrew (3.4)	Above the Limit (3.0) Mr. Flip (3.0)
	Average	Puddles of Max (2.6) America Dave (2.6)	Come on Man (2.8) Match This (2.7)	The Enchanted Drawing (2.6) Those Awful Hats (2.6)
	Low	Princess Year (2.2) West (2.1)	Jane 2 (2.1) Ha Ha Ha (1.8)	New Pillow Fight (2.3) Clowns Spinning Hats (2.0)
ROMANCE	High	Love in a Separation (3.5) Private Woman (3.4)	When You Least Expect It (3.7) Daydreaming in New York (3.7)	At First Sight (4.0) Bed of Roses (3.7)
	Average	The Pearl Rollercoaster (2.7) Walls of Sky (2.5)	Take it Slow (3.3) Cool Happiness in New York (3.0)	Bitter Moon (3.2) Ballistic Kiss (2.9)
	Low	A Speckled McKee (2.2) Death (1.4)	Cheeky (2.5) Red Beans and Rice (1.9)	1871 (2.3) The 5th Monkey (2.0)
PUBS	High	The Hound's Head (3.5) The Royal Hood (3.1)	The Old Barrel (3.8) The Rusty Spur (3.8)	The Rose & Crown (3.8) The White Lion (3.7)
	Average	The Green Bay (2.6) The Garter (2.4)	The Dark Forest (3.0) The Dog's Ear (3.0)	The Castle (3.2) The Cross Keys (3.1)
	Low	The Cow (2.2) The Pear (1.9)	The Cat's Meow (2.6) The Paper Cut (2.3)	The White Hart (2.5) The New (1.8)

Table 1: Examples of different names, organized by source and average rating, with their rating indicated in parenthesis. The names were assigned as Low, Average and High based on their relative position compared to the median rating for that source and genre. 90 names are shown out of the total 375.

Music Reading Expertise Modulates Visual Spans in both Music Note and English Letter Reading

Tze Kwan, Li (saralis@hku.hk)
Department of Psychology, University of Hong Kong
627 Jockey Club Tower, Pokfulam Road, Hong Kong SAR

Susana T. L. Chung (s.chung@berkeley.edu)
School of Optometry, University of California
Berkeley, CA, USA

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, University of Hong Kong
627 Jockey Club Tower, Pokfulam Road, Hong Kong SAR

Abstract

Here we investigated how music reading experience modulates visual spans in language reading. Participants were asked to identify music notes, English letters, Chinese characters, and novel symbols (Tibetan letters) presented at random locations on the screen while maintaining central fixation. We found that for music note reading, musicians outperformed non-musicians at some peripheral positions in both visual fields, and for English letter reading, musicians outperformed non-musicians at some peripheral positions in the RVF but not in the LVF. In contrast, in both Chinese character and novel symbol reading, musicians and non-musicians did not differ in their performance at peripheral positions. Since both music and English reading involve a left-to-right reading direction and a RVF/LH advantage, these results suggest that the modulation of music reading experience on visual spans in language reading depends on the similarities in the cognitive processes involved.

Keywords: Music reading expertise; visual span; English reading; Chinese reading; symbol reading

Introduction

Recent research has shown that experts have superior perceptual cognitive abilities than novices in meaningful tasks related to their expertise, such as in sports (Mann, Williams, Ward & Janelle, 2007) and in chess playing (Reingold, Charness, Pomplun & Stampe, 2001). From these studies, experts typically demonstrated superior response accuracy (ACC), faster response time (RT), fewer eye fixations of longer duration, or a larger visual span in the tasks. Visual span has been defined as the region around the fixation point within which visual stimuli can be recognized (O'Regan, Lévy-Schoen & Jacobs, 1983). The size of one's visual span can be enlarged through expertise training as in perceptual learning, and this can consequently improve reading/processing speed (Chung, Legge, & Cheung, 2004). For example, chess experts were found to have a larger visual span, fewer fixations, and faster RT in structured chess configuration detection than novices (Reingold et al., 2001), suggesting they are able to extract more information from one fixation in the field of expertise than novices. Similarly, musicians with extensive music reading experiences may have a larger reading span than poor readers of music notation. Indeed, one study found that well-trained musicians had a span of 6.5 notes, whereas the poorest reader only read 3.5 notes at a time (Sloboda, 1974).

Music reading involves mapping a set of spatially distributed notes and chords in a staff to a horizontal melodic line (Stewart, 2005). In the horizontal direction, musicians read further ahead from left to right and focus on areas between notes rather than a single note to ensure in-time playing (Goolsby, 1994). In the vertical direction, musicians read musical markings below staff without making an eye fixation, or attempt to read two staffs at the same time (such as in piano playing). These suggest that musicians may attempt to perceive as much information as possible through both horizontal and vertical peripheral vision (Goolsby, 1994), and consequently develop a larger visual span in both directions in music reading tasks.

Not only does music reading expertise influence visual span in music reading, but it may also modulate visual span in language reading due to similarities in the cognitive processes involved in both tasks. Both music and English are read from left-to-right, and thus music notes and letters are recognized in the right visual field (RVF) more often (Brysbaert & Nazir, 2005; Wong & Hsiao, 2012). This perceptual learning results in processing advantages in the RVF/left hemisphere (LH) (Brysbaert & Nazir, 2005). Indeed, previous studies have shown an RVF/LH advantage in both music note (Segalowitz, Bebout & Lederman, 1979) and English word processing (Brysbaert & d'Ydewalle, 1990). In addition, music notation reading involves note-to-sound mapping, and similarly English word reading involves grapheme–phoneme correspondence (Brown, Martinez & Parsons, 2006; Hsiao & Lam, 2013). Both types of mapping involve decomposing visual stimuli into components for mapping to sound components, and this kind of analytic encoding process is shown to be dominant in the LH (e.g., Bradshaw & Nettleton, 1981; Hébert & Cuddy, 2006; Hsiao & Lam, 2013; Segalowitz et al., 1979). Thus, music notation and English word reading may share similar underlying neural mechanisms. In support of this notion,

patients with music reading deficiencies due to brain lesions in the LH also showed word reading difficulties in alphabetic languages such as English (Hébert & Cuddy, 2006). In short, music and English reading may share similar cognitive processes as evidenced by their similar processing advantages in the RVF/LH.

In contrast to left-to-right music and English reading, Chinese can be read in all directions (left to right, right to left, or vertically). Moreover, due to its unique logographic orthography, each Chinese character is regarded as a morpheme and corresponds to a syllable in the pronunciation. Since there is no grapheme-phoneme correspondence in Chinese, decomposition of a character into components is not required. This may account for a left visual field (LVF)/right hemisphere (RH) advantage typically observed in Chinese orthographic processing (e.g., Tzeng, Hung, Cotton & Wang, 1979; Tan et al., 2001; Hsiao & Lam, 2013). Consistent with this speculation, brain imaging studies typically showed more bilateral or right-lateralized activation in the visual area in Chinese character processing as compared with English word reading (e.g., Tan, Laird, Li & Fox, 2005). These findings suggest that music-reading expertise may have limited influence on Chinese reading due to the different hemispheric lateralization and cognitive processes involved.

In this study, we examine how music-reading expertise modulates visual span in music reading, as well as in English and Chinese reading according to their similar/different cognitive processes involved. We also investigate whether music-reading expertise modulates visual span in reading novel symbols (i.e., Tibetan letters) as a transfer effect. We hypothesize that musicians would have a larger visual span than non-musicians in music reading due to an expertise effect, and that for English reading, musicians may have a larger visual span than non-musicians in the RVF due to the similar cognitive processes inherent in both music and English reading. More specifically, musicians who are also expert English readers may have received more training in the RVF/LH through extensive music reading, which facilitates perceptual processing in the RVF/LH, and may thus further benefit English reading in the RVF. In contrast, for Chinese character reading, musicians and non-musicians may not display differences in visual span due to the different cognitive processes involved in music and Chinese reading. More specifically, musicians' RVF/LH processing advantage in music reading may not translate to an increased visual span in Chinese reading due to the reliance on LVF/RH processing in Chinese reading. Here we also examine if a larger visual span of musicians in music note reading can be found in novel symbol (i.e., Tibetan letter) reading due to a possible transfer effect. To test these hypotheses, here we conducted a visual word/symbol identity matching task, in which participants were asked to attend to a briefly presented screen filled with music notes, English letters, Chinese characters, or Tibetan letters, and match a target stimulus presented at a given location afterwards. We examined how musicians and non-musicians differ in their performance.

Methods

Participants

Participants consisted of 64 Cantonese (L1)-English (L2) bilinguals from Hong Kong, whose ages ranged from 18 to 29 (M = 22, SD = 2.9). They had similar linguistic and college education backgrounds. They were categorized as musicians (n = 32) and non-musicians (n = 32), with 16 males and 16 females in each group. Musicians were well-trained pianists, who started music training at age 3-10 (M = 4.9, SD = 1.8). All of them were either piano teachers, music undergraduate/postgraduate students, or frequent piano players. They had attained grade 8 or above in the graded piano examinations of the Associated Board of The Royal Schools of Music (ABRSM), with 8-25 years experience in piano playing (M = 16.3, SD = 4.2) and regular music reading hours per week (M = 9.3, SD = 11.5). In contrast, non-musicians did not receive any music training.

Aside from their music training background, musicians and non-musicians were closely matched in other aspects detailed as follows. All participants were right-handed, which was assessed using the Edinburgh Handedness Inventory (Oldfield, 1971; musicians: M = 75.78, 5th right decile; non-musicians: M = 70.00, 4rd right decile, $t(62)$ = 1.295, *n.s.*). Both musicians and non-musicians had normal or corrected to normal vision (20/20) as shown in the Freiburg Visual Acuity and Contrast Test (FrACT; Bach, 2006; musicians: M = 1.27; non-musicians: M = 1.33, $t(62)$ = -1.043, *n.s.*). Both groups' verbal and spatial working memory performance were matched in an N-back task (Lau, Ip, Lee, Yeung & Eskes, 2013; Verbal ACC: musicians: M = 87.5%; non-musicians: M = 79.6%, $t(62)$ = 1.863, *n.s.*; Spatial ACC: musicians: M = 81.7%; non-musicians: M = 74.0%, $t(62)$ = 1.746, *n.s.).* All participants started learning English as a second language at age 3 (M = 3.4, SD = 1.57), and no participants had any experience with the Tibetan language.

Materials

Materials consisted of four types of stimuli: English letters, Chinese characters, music notes and Tibetan letters. English lower-case letters (a-z, n = 26) were included. Chinese characters stimuli (n = 4805) were selected from the *List of Graphemes of Commonly-used Chinese Characters* (Chinese Language Education Section, HKSAR, 2012), comprising Chinese characters ranging from low to high number of strokes and frequency according to Ho's (1998) database. As for the music notes (n = 11), crotchets (1 beat) ranging from D4 to G5[1] were selected. Tibetan letters (n = 46) were included as novel symbols.

Design

Participants completed a visual word/symbol identity matching task with English letters, Chinese characters, mu-

[1] D4 to G5 ranges across one octave from the D note below the first line to the G note above the fifth line.

sic notes, and Tibetan letters in four separate progressive blocks with 36 positions around a central fixation (Fig. 1).

In the visual word/symbol identity matching task, progressive testing started from level 1 to level 3 according to participants' ACC in the 36 testing positions of stimulus identification (Fig. 1). The progression threshold for each level was set at 50% ACC. For example, in level 1, participants proceeded to the next adjacent positions (position 8,9,14) in level 2 only if their performance in position 15 was above chance level (50%); otherwise they did not proceed to the adjacent positions at level 2 after finishing all trials at position 15. Similar progression rules applied to level 2 positions when advancing to level 3. The experiment was terminated if a participant was not able to proceed to any adjacent position.

We measured visual span in two different ways: the number of musicians and non-musicians who reached the testing positions under this progressive testing paradigm, and participants' ACC in the task. To ensure the readability of stimuli, we doubled the size of stimuli in the present study from their usual size for expert readers in daily life. For music notes, we doubled the size of a crotchet found in Grieg's (1888) Anitra's Dance from Peer Gynt Suite No.1, Op.46 in *Piano Pieces the Whole World Plays*, which is a piano piece for post-intermediate piano players. For English letters, Chinese characters and Tibetan letters, we doubled the size of an alphabet/a character/a letter found in the text of English/Chinese/Tibetan newspapers respectively.

Participants' viewing distance was fixed at 61 cm in our task. Under this viewing distance, a music note (crotchet) with its corresponding five-line staff subtended a horizontal and vertical visual angle of 1.19° x 2.25°. The 6 (horizontal) x 6 (vertical) testing positions of music notes subtended a visual angle of 7.10° x 13.41°. Here we included crotchets as a single unit comparable to the other stimuli in the visual word/symbol identity matching task (i.e., English letter, Chinese character, Tibetan letter). English letters were displayed in Courier – a serif font with fixed width – to ensure constant center-to-center spacing between letters. The lowercase letter 'x', as a reference letter without ascenders or descenders such as letter 'h' and 'g' respectively, subtended 0.47° of visual angle horizontally and 0.50° vertically on a rectangular background that subtended a visual angle of 0.47° x 0.76°, with 1 x standard letter spacing. The 6 (horizontal) x 6 (vertical) testing positions of English letters subtended a visual angle of 2.81° x 4.55°.

Chinese characters were presented in Microsoft DF-Hei font, a serif font with fixed width, to ensure constant center-to-center spacing between characters. Each character subtended a horizontal and vertical visual angle of 0.72° x 0.76° with 1x standard center-to-center character spacing. The 6 (horizontal) x 6 (vertical) testing positions of Chinese characters subtended a visual angle of 4.30° x 4.55°.

Tibetan letters were displayed in Himalaya font. The letter '', as a reference letter without ascenders or descenders, subtended a horizontal and vertical visual angle of 0.69° x 0.73° on a rectangular background that subtended a visual angle of 0.72° x 1.25°, with 1x standard letter spacing. The 6 (horizontal) x 6 (vertical) testing positions of Tibetan letters subtended a visual angle of 4.30° x 7.52°.

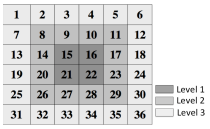


Figure 1. A sample of 36 testing positions of Chinese characters in a 6 x 6 layout around central fixation

The average luminance of stimuli was adjusted to 3.63 cd/m^2. With 73.8 cd/m^2 background luminance, the Weber contrast of the stimuli was -0.95. Experiments were conducted using SR Experiment Builder with an EyeLink 1000 eye tracker (SR Research Ltd., Canada) to ensure participants' central fixation. A chinrest was used to reduce head movement. Calibration and validation was performed before the start of each block. Block order was counterbalanced and trials were randomized across participants.

Procedure

Each trial started with a drift correction to ensure accurate central fixation. After detecting central fixation, a screen filled with stimuli was presented for 200 ms as the time constraint, to allow only one fixation without eye movement in letter recognition (Legge, Mansfield & Chung, 2001). After a 500ms blank screen, a target stimulus was then presented at one of the 36 testing positions around central fixation at the designated level (Fig. 1); the screen remained unchanged until participants responded (Fig. 2). Participants had to judge whether the target stimulus was identical to the stimulus presented earlier at the same position on the screen filled with stimuli, as quickly and accurately as possible, without shifting their gaze away from the central fixation (+). Each position consisted of 10 'yes' and 10 'no' trials that were randomly created without repetitions. Participants responded by pressing buttons on a response box with both hands. ACCs were recorded.

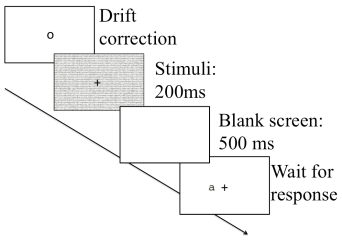


Figure 2. Procedure of the visual word/symbol identity matching task

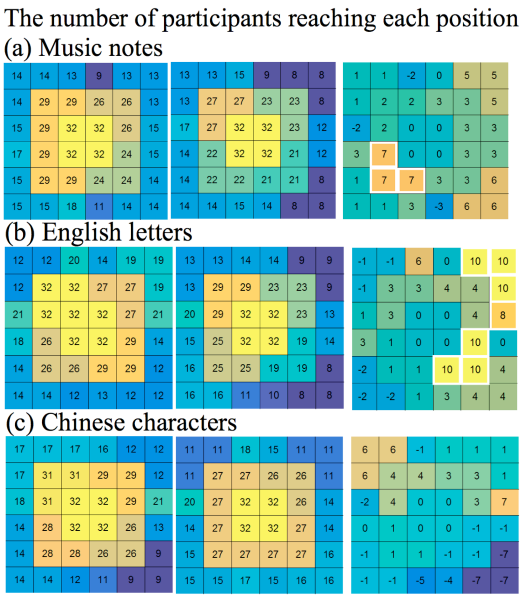
Prior to the visual word/symbol identity matching task, a demographic and music background questionnaire, Freiburg Visual Acuity and Contrast Test (FrACT; Bach, 2006), Edinburgh Handedness Inventory (Oldfield, 1971) and an N-back task (Lau et al., 2013) were conducted to assess partic-

ipants' language and music learning background, visual acuity, handedness and working memory capability.

Results

We first measured visual spans according to the number of musicians and non-musicians who reached the testing positions under our progressive testing paradigm. As shown in Figure 3, more musicians reached the 36 testing positions than non-musicians in music note and English letter reading, but not in Chinese character and Tibetan letter reading. For each testing position, we used Chi-square tests to examine whether there was a significant difference in the number of musicians and non-musicians reaching each position under progressive testing. For music notes, there were significantly more musicians reaching 3 level-2 positions in the lower LVF than non-musicians, including position 20 ($\chi^2(1) = 4.730$, $p = .030$, $\varphi = 0.27$), 26 ($\chi^2(1) = 4.730$, $p = .030$, $\varphi = 0.27$), and 27 ($\chi^2(1) = 4.730$, $p = .030$, $\varphi = 0.27$; Fig. 3a). No significant differences in the number of musicians and non-musicians reaching other positions were observed.

For English letters, there was a significant difference in the number of musicians and non-musicians reaching 7 positions in the RVF, including position 5 ($\chi^2(1) = 6.349$, $p = .012$, $\varphi = 0.32$), 6 ($\chi^2(1) = 8.349$, $p = .012$, $\varphi = 0.36$), 12 ($\chi^2(1) = 8.349$, $p = .012$, $\varphi = 0.36$), 18 ($\chi^2(1) = 4.016$, $p = .045$, $\varphi = 0.25$), 23 ($\chi^2(1) = 8.333$, $p = .004$, $\varphi = 0.36$), 28 ($\chi^2(1) = 8.333$, $p = .004$, $\varphi = 0.36$), and 29 ($\chi^2(1) = 8.333$, $p = .004$, $\varphi = 0.36$). More musicians reached positions at level 2 in the lower RVF and positions at level 3 in the upper RVF than non-musicians in English reading (Fig. 3b). No significant differences in the number of musicians and non-musicians reaching other testing positions were observed. For Chinese characters and Tibetan letters, no significant differences were found in the number of musicians and non-musicians reaching the 36 testing positions (Fig. 3c, 3d).



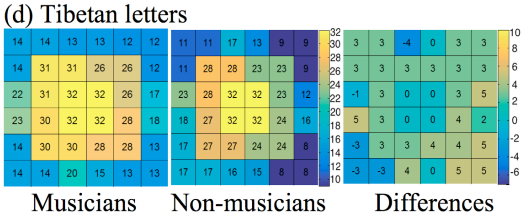
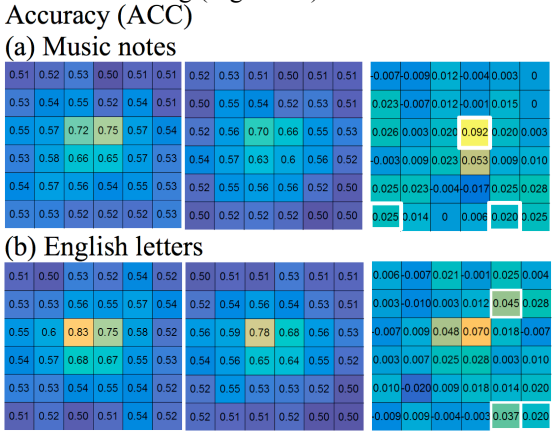


Figure 3. The number of musicians and non-musicians reaching each testing position, and the differences between the two groups (musicians - non-musicians) in reading (a) music notes, (b) English letters, (c) Chinese characters, and (d) Tibetan letters under progressive testing paradigm. (White borders: $p < .05$)

To examine participants' identification ACC, *t*-tests were used to compare musicians' and non-musicians' ACCs in the 6 (horizontal) x 6 (vertical) testing positions in the visual word/symbol identity matching task. In our analysis, we substituted the ACCs below the chance level (0.5) and missing data (i.e., for participants who did not reach a position due to their ACCs being below 0.5 in the corresponding position at the previous level) with the chance level performance (0.5). For music note ACC, a significant difference was found in position 16 ($t(62) = 2.469$, $p = .016$, $d = 0.63$), 31 ($t(62) = 2.072$, $p = .042$, $d = 0.53$), and 35 ($t(62) = 2.056$, $p = .044$, $d = 0.52$; Figure 4a). Musicians performed better than non-musicians in music note reading at the central (level 1) position in the upper RVF and two positions at level 3 in the lower LVF and RVF. For English letter ACC, a significant difference was found in position 11 ($t(62) = 2.590$, $p = .012$, $d = 0.66$), 35 ($t(62) = 2.729$, $p = .008$, $d = 0.69$), and 36 ($t(62) = 2.425$, $p = .018$, $d = 0.62$; Figure 4b). Musicians performed better than non-musicians when English letters were presented at one position at level 2 in the upper RVF and two positions at level 3 in the lower RVF. For Chinese character ACC, a significant difference was found in position 15 ($t(62) = 2.069$, $p = .043$, $d = 0.53$; Figure 4c). Musicians performed better than non-musicians when Chinese characters were presented at the central position in the upper LVF. For Tibetan letter reading, no significant differences were found between musicians and non-musicians, suggesting that musicians did not hold a significant advantage over non-musicians in Tibetan letter reading (Figure 4d).



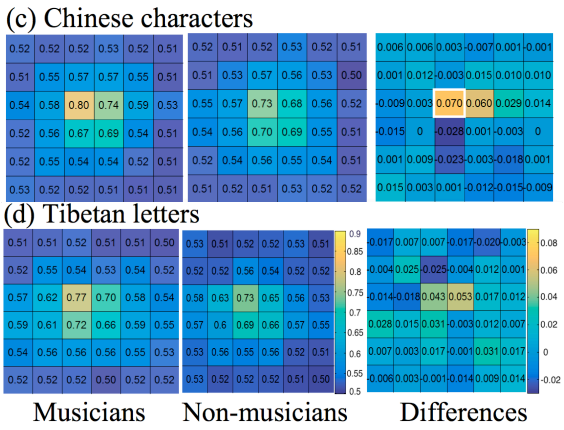


Figure 4. The ACCs of musicians, non-musicians, and the differences between the two groups (musicians - non-musicians) in reading (a) music notes, (b) English letters, (c) Chinese characters and (d) Tibetan letters. (White borders: $p < .05$)

Discussion

Here we examined how music reading expertise influences visual span in music note, English letter, Chinese character, and novel symbol (i.e., Tibetan) reading. For music note reading, musicians outperformed non-musicians at the central position in the upper RVF and some peripheral positions in both the lower LVF and RVF, suggesting that they have a larger visual span than non-musicians. This result is consistent with previous findings that experts may develop a larger visual span than novices in meaningful tasks related to their expertise, as shown in a chess configuration detection task (Reingold et al., 2001) and a music note playing task (Sloboda, 1974). The finding that musicians' advantage could be found in both visual fields was consistent with Proverbio, Manfredi, Zani and Adorni's (2012) finding that visual processing of music notes involves bilateral activations in the fusiform (BA37) and inferior occipital gyri (BA18). Note that this result is in contrast to Segalowitz et al.'s (1979) study, which showed a RVF advantage in a chord-playing task. This difference may be due to the involvement of left-lateralized motor planning in music playing tasks. Putting together, these results suggest that lateralization of music processing may depend on the task requirements.

For English letter reading, consistent with our hypothesis, musicians outperformed non-musicians at several peripheral positions in the RVF. This finding suggests that music-reading expertise may modulate visual span in English reading due to their similarities in the cognitive processes involved. Both music notations and English words are read from left-to-right, and thus music notes and English letters are recognized in the RVF more often (Brysbaert & Nazir, 2005; Wong & Hsiao, 2012). With extensive music reading experience, musicians hold a processing advantage in the RVF/ LH through perpetual learning, and this may translate to better English reading performance in the RVF. Moreover, both the note-to-sound mapping in music reading and the grapheme–phoneme correspondence in English reading (e.g., Brown, Martinez & Parsons, 2006) may involve more LH analytic processing than RH processing (Bradshaw & Nettleton, 1981; Hébert & Cuddy, 2006; Hsiao & Lam, 2013; Segalowitz et al., 1979). Due to these similarities in cognitive processing between music and English reading, music-reading expertise may facilitate perceptual processing in the RVF/LH, which further benefits English reading in the RVF.

For Chinese character reading, musicians performed better than non-musicians at the central position in the upper LVF, but not in any peripheral locations. This result is consistent with our hypothesis that music-reading expertise has less influence on the visual span for Chinese character reading than English letter reading. This effect may be due to the different cognitive processes involved in music and Chinese reading. Contrary to left-to-right music reading, Chinese can be read in all directions (left to right, right to left, or vertically), possibly resulting in different perceptual learning. Moreover, Chinese has no grapheme-phoneme correspondence, and thus decomposition of a character into components is not required. As such, Chinese reading tends to be right-lateralized (Tzeng et al., 1979) or bilateral (Tan et al., 2001) in its orthographic processing due to its unique logographic orthography. Taken together, the different processing advantages between music (RVF/LH) and Chinese (LVF/RH) reading may reduce the facilitation of music reading expertise on the visual span of Chinese reading. However, musicians' advantage in the central upper LVF suggests the possibility of modulation of music reading experience in Chinese character processing. Future work will examine this possibility.

For novel symbol (i.e., Tibetan letter) reading, no performance differences were found in any of the positions between musicians and non-musicians. This result suggests that the two groups have similar visual spans in novel symbol processing. The absence of a transfer effect to novel symbol processing suggests that the modulation of music reading experiences on visual spans in reading is limited to stimuli of expertise, and cannot be generalized to novel symbols.

Note that in the current study, we measured visual span as participants' identification performance when the target location within the stimulus was unknown beforehand. In other words, the visual span was measured in a distributed attention condition, in which participants had to pay attention to the whole stimulus without orienting their attention to a specific location beforehand. This is in contrast to some previous studies of visual span, in which a cue was provided prior to the presentation of the stimulus (e.g., Legge et al., 2001) to allow participants' orientation of attention to the targeted location beforehand. Future work will examine whether similar modulation effects can be observed when a different measure of visual span is used.

To conclude, this study examined how music reading expertise influences visual spans in reading music notes, English letters, Chinese characters, and novel symbols reading.

As an expertise effect, musicians outperformed non-musicians in music reading at some central and peripheral positions in both visual fields. Interestingly, for English letter reading, musicians also outperformed non-musicians in some peripheral positions in the RVF but not in the LVF. In contrast, in both Chinese character and novel symbol reading, musicians and non-musicians did not differ in their performance in any peripheral positions. These results can be explained by both music and English processing advantages in the RVF/LH, whereas Chinese character processing is more right-lateralized or bilateral. Thus, the modulation of music-reading expertise on visual spans in language reading depends on the similarities in the cognitive processes involved.

Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (ECS scheme project # HKU 758412H to J.H. Hsiao). We thank Professor Gail Eskes for providing the n-back program and MindSea Development Inc. for writing the software. We also thank Dr. Sing Hang Cheung for providing useful comments and a prototype of Chinese stimuli.

References

Bach, M. (2006). The Freiburg Visual Acuity Test-variability unchanged by post-hoc re-analysis. *Graefes Arch. Clin. Exp. Ophthalmol., 245*, 965-971.

Bradshaw, J. L., & Nettleton, N. C. (1981). The nature of hemispheric specialization in man. *Behav. Brain Sci., 4*, 51-63.

Brown, S., Martinez, M. J., & Parsons, L. M. (2006). Music and language side by side in the brain: a PET study of the generation of melodies and sentences. *Eur. J. Neurosci., 23*, 2791-2803.

Brysbaert, M., & d'Ydewalle, G. (1990). Tachistoscopic presentation of verbal stimuli for assessing cerebral dominance: Reliability data and some practical recommendation. *Neuropsychologia, 28*, 443- 455.

Brysbaert, M., & Nazir, T. (2005). Visual constraints in written word recognition: evidence from the optimal viewing-position effect. *J. Res. Read., 28*, 216- 228.

Chinese Language Education Section, Curriculum Development Institute, Education Bureau, Hong Kong SAR (2012). *List of Graphemes of Commonly-used Chinese Characters.* Hong Kong: Education Bureau, HKSAR.

Chung, S. T., Legge, G. E., & Cheung, S. H. (2004). Letter-recognition and reading speed in peripheral vision benefit from perceptual learning. *Vision Res., 44*, 695-709.

Goolsby, T. W. (1994). Profiles of processing: Eye movements during sightreading. *Music Percept.,12,* 97-123.

Grieg, E. (1888). Anitra's Dance from Peer Gynt Suite No.1, Op.46. In A. E. Wier *Piano Pieces the Whole World Plays* (pp. 72-74). Beijing: People's Music Publishing House.

Hébert, S., & Cuddy, L. L. (2006). Music-reading deficiencies and the brain. *Adv. Cogn. Psychol., 2*, 199-206.

Ho, H. H. (1998). *Hong Kong, mainland China & Taiwan: Chinese character frequency–A trans-regional, diachronic survey*. Retrieved January 27, 2015 from The Chinese University of Hong Kong, Website: http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/

Hsiao, J. H., & Lam, S. M. (2013). The Modulation of Visual and Task Characteristics of a Writing System on Hemispheric Lateralization in Visual Word Recognition—A Computational Exploration. *Cognitive Sci., 37*, 861-890.

Lau, E. Y. Y., Ip, M. S. M., Lee, T. M. C., Yeung, A. W. M., & Eskes, G. A. (2013). Neurocognitive and Psychosocial Outcomes of Obstructive Sleep Apnea in Hong Kong Chinese: Similar to or Different from Western Populations? *J. Sleep Disor: Treat Care*, 2: 3.

Legge, G. E., Mansfield, J. S., & Chung, S. T. (2001). Psychophysics of reading: XX. Linking letter recognition to reading speed in central and peripheral vision. *Vision Res.,* 41, 725-743.

Mann, D. T., Williams, A. M., Ward, P., & Janelle, C. M. (2007). Perceptual-cognitive expertise in sport: A meta-analysis. *J. S. E. P., 29*, 457.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*, 97-113

O'Regan, J. K., Lévy-Schoen, A., & Jacobs, A. M. (1983). The effect of visibility on eye-movement parameters in reading. *Percept. & Psychophys., 34*, 457-464.

Proverbio, A. M., Manfredi, M., Zani, A., & Adorni, R. (2013). Musical expertise affects neural bases of letter recognition. *Neuropsychologia, 51*, 538-549.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychol. Sci., 12*, 48-55.

Segalowitz, S. J., Bebout, L. J., & Lederman, S. J. (1979). Lateralization for reading musical chords: Disentangling symbolic, analytic, and phonological aspects of reading. *Brain Lang., 8*, 315-323.

Sloboda, J. (1974). The Eye-Hand Span--An Approach to the Study of Sight Reading. *Psychol. Music, 2*, 4-10.

Stewart, L. (2005). A neurocognitive approach to music reading. *Ann. N. Y. Acad. Sci.*, 1060, 377-386.

Tan, L. H., Laird, A. R., Li, K., & Fox, P. T. (2005). Neuroanatomical correlates of phonological processing of Chinese characters and alphabetic words: A metanalysis. *Hum. Brain Mapp., 25,* 83-91.

Tan, L. H., Liu, H. L., Perfetti, C. A., Spinks, J. A., Fox, P. T., & Gao, J. H. (2001). The neural system underlying Chinese logograph reading. *Neuroimage, 13*, 836-846.

Tzeng, O. J. L., Hung, D. L., Cotton, B., & Wang, S. Y. (1979). Visual lateralization effect in reading Chinese characters. *Nature (London), 282*, 499-501.

Wong, Y. K., & Hsiao, J. H. W. (2012). Reading direction is sufficient to account for the optimal viewing position in reading: The case of music reading. *CogSci 2012 proceedings,* 2540-2545.

Design from Zeroth Principles

Jordan W. Suchow (suchow@berkeley.edu)
Michael D. Pacer (mpacer@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology,
University of California, Berkeley, USA

Abstract

A successful design accounts for the structure of the problem it is aimed at solving. When it is a human-directed design, this includes the expectations of its users. How do we arrive at such a design? One approach starts from first principles (e.g., simplicity, unity, symmetry, balance) to evaluate the quality of proposed designs. Here, we introduce *design from zeroth principles*, a form of human-in-the-loop computation that synthesizes a design that conforms to its users' expectations. The technique begins by constructing a transmission chain seeded with a random design. Each user in the chain is exposed to the design and then recreates it, passing along their recreation to the next user, who does the same. Through this iterative process, the users' perceptual, inductive, and reconstructive biases directly transform the initial design into one that is better fit to human cognition. Such designs are easier to learn and harder to forget. We evaluated the approach in three domains — stimulus–response mappings, vanity phone numbers, and letter placement in typeset words — and show that it produces a good design in each.

Keywords: design, cognitive ergonomics, inductive bias, transmission chain, user interface

Introduction

Successful design in the rationalist tradition begins by evaluating the problem that a designed object or system aims to solve: the goals, any constraints imposed by the environment or by human factors, and the surrounding context, broadly construed (Simon, 1996). Another tradition appeals to principles that are purported to be universal — simplicity, balance, unity, order, liveliness — rather than to direct considerations of function (Lidwell, Holden, & Butler, 2010; White, 2002). Following these principles will, in theory, lead to successful designs.

But what makes a design successful? Certainly, in cases where the design acts as the interface between a user and a system, the success of the design hinges in part on the user's experience in working with the design. Users have expectations about how to interact with the world to accomplish their goals, and a good design conforms to those expectations — i.e., when humans are the users, good designs fit the human mind. Practices have developed around ensuring this, including techniques to compare variants of a design through statistical hypothesis testing (e.g., A/B testing), to measure performance under user-focused metrics (e.g., usability research), and to elicit feedback from potential users (e.g., focus groups).

In cognitive science, a person's expectations can be described as *perceptual*, *inductive,* and *reconstructive* biases, as they pertain to perception, learning, and memory, respectively. Bias in this sense means merely that the distribution of expected designs is not uniform — some match expectations better than others. Tools and techniques from the domain of cognitive science that identify, extract, and amplify these biases can thus aid designers in their quest to find cognitively fit forms.

Transmission chains are one such technique for extracting and amplifying biases in memory and learning. Originating in early experiments by Frederic Bartlett, transmission chains pass information from one person to the next (Bartlett, 1932). At each step of the chain, the transmitted information is transformed. So long as a few technical conditions hold, repeated application of a transformation leads to erasure of the information contained in the input, leaving behind a signature of the transformation process itself.

In this paper, we introduce a method that uses transmission chains to synthesize a design. Our technique begins by constructing a transmission chain seeded with a random design. Each user in the chain is exposed to the design and then recreates it, passing along their recreation to the next user, who does the same. Through this iterative process, the users' inductive and reconstructive biases directly transform the initial design into one that is better fit to human cognition. No formal design principles are assumed. Thus, we call this process *design from zeroth principles*.

The plan for the paper is as follows. We begin with a technical description of the transmission chain technique and its ability to amplify biases in perception, learning, and memory. Next, we apply the technique in three domains: stimulus–response mappings, vanity numbers, and letter placement in typeset words. We conclude with a discussion of how our method can be extended and elaborated, some criteria for design problems that might benefit from the technique, and possible modifications to the transmission-based scheme that relate to other forms of human-in-the-loop computation.

Transmission chains reveal biases

In a transmission chain, information is passed from one person to the next. In the children's game *Telephone*, for example, a child invents a sentence and whispers it to the next child in line, who then does the same. By the time the sen-

tence reaches the end of the chain, it has changed. Hilarity ensues.

This kind of system can be formally modeled as a Markov chain, a stochastic process in which transformations are defined by a transition matrix specifying the probability of going from one given state to any other state (Kalish, Griffiths, & Lewandowsky, 2007). In *Telephone*, for example, the transition matrix defines the probability that a given sentence will transform into a given another (e.g., that "laid him on the green" becomes "Lady Mondegreen").

If a Markov chain obeys certain requirements,[1] it can be proven that it will eventually converge to a *stationary distribution*, a distribution of states unchanged by the transformation. This means that even if we seed a chain with a random state, after enough steps, the information contained in the input will be lost, leaving behind a signature of the transformation process itself. Identifying the stationary distribution of a Markov chain requires having a model for the probability with which it transitions from one state to another. In the case of transmission chains, one such model is provided by assuming that perception, learning, and memory follow the principles of Bayesian inference. Given an observed stimulus d, people consider hypotheses h about its nature, and then produce a reconstruction d'. The Bayesian analysis of transmission chains assumes hypotheses are sampled from the posterior distribution $p(h|d) \propto p(d|h)p(h)$, where $p(d|h)$ gives the probability of seeing d if it were generated from h (known as the *likelihood*) and $p(h)$ is the *prior* distribution over hypotheses and encodes people's expectations about the prevalence of different hypotheses. If d' is generated by sampling from the likelihood, then the stationary distribution of this Markov chain is the *prior predictive distribution* $p(d) = \Sigma_h\, p(d|h)p(h)$. Consequently, we should expect the outcome of transmission to reflect people's expectations, as expressed in the form of this prior predictive distribution.

Cognitive scientists have used transmission chains in studies of *serial reproduction* and *iterated learning* to study reconstructive biases in memory and inductive biases in learning. For example, in Kalish, Griffiths & Lewandowsky (2007), participants first learned a functional relationship between two magnitudes (the length of a rectangular bar and the width of another) by observing pairs. Notably, participants were then tested on some examples that they had never directly observed. Responding to these novel stimuli requires generalization beyond what they have observed. The authors found that the functional form passed in these transmission chains gradually reverts to a linear relationship regardless of the data that seeds the chain. From this, they concluded that this functional form is what people expect.

[1] The condition that needs to be met is that the Markov chain must be *ergodic*, which is to say that starting from any state, one can eventually reach any another state; that the expected number of steps needed to reach each other state is finite in expectation; and that returning to any one state does not occur only as a multiple of some $k > 1$).

If a good design is one that fits the expectations of its users, then any difficulty in perceiving, learning, or remembering a design indicates that it may be inconsistent with the user's cognitive biases. By passing the design through a transmission chain, the users' perceptual, inductive, and reconstructive biases will transform the initial design into one that is better fit to human cognition. In this way, it becomes possible to improve a design without appealing to first principles — thus we call it *design from zeroth principles*.

Experiment 1: Stimulus–response mappings

In which direction should a screw be turned in order to drive it further into wood? Which light switch should be flipped to turn off the patio light? And which knob should be turned to light the front left stove burner? Assigning these mappings are design decisions, and some mappings are better than others. Designs with *stimulus–response compatibility* offer a simple and clear mapping between an action and a response, leading to shorter reaction times and lower rates of error (Fitts & Seeger, 1953; Proctor & Reeve, 1989; Kornblum, Hasbroucq, & Osman, 1990).

In Experiment 1A, we applied design from zeroth principles to stimulus–response mappings between light switches and lights. Experiment 1B evaluated the resulting mappings.

Methods, 1A

Experiment 1A constructed a transmission chain where participants passed along a mapping between light switches and lights. The chain was seeded with a random mapping.

Participants. We recruited 100 participants on Amazon Mechanical Turk, an online crowdsourcing platform. Each participant was paid $0.25 for a few minutes of work.

Stimulus. The stimulus was a depiction of a set of six light switches and six lights (Fig. 1). When pressed, the switch turned on one of the lights for 1000 ms.



Figure 1. A set of six light switches and lights. Which switch maps to which light?

Procedure. Procedure. First, the participant learned the mapping. On each trial, one of the switches was highlighted in green. The participant was instructed to press the switch and observe what happened. Each switch was highlighted once over the course of training, such that the participant observed the entire mapping. Then the participant was tested on what they had learned. On each of six trials, one of lights was highlighted with a bounding box. The participant was instructed to turn on the highlighted light by pressing the corresponding switch. Once pressed, the switch became dis-

abled and could not be reselected. There was no feedback. The order in which the pairings were learned and tested was random. Together, the participant's six responses define a new mapping that was then passed along to the next participant in the chain. There were 10 chains of 10 participants.

Results & Discussion, 1A

Over time, mappings in the chain became more regular, by the tenth generation coming to resemble the solution where all the switches are mapped to the light directly above them (average Kendall rank correlation coefficient $\tau = 0.96$ between light position and switch position; Fig. 2). In fact, 6 of 10 chains converged to exactly this solution.

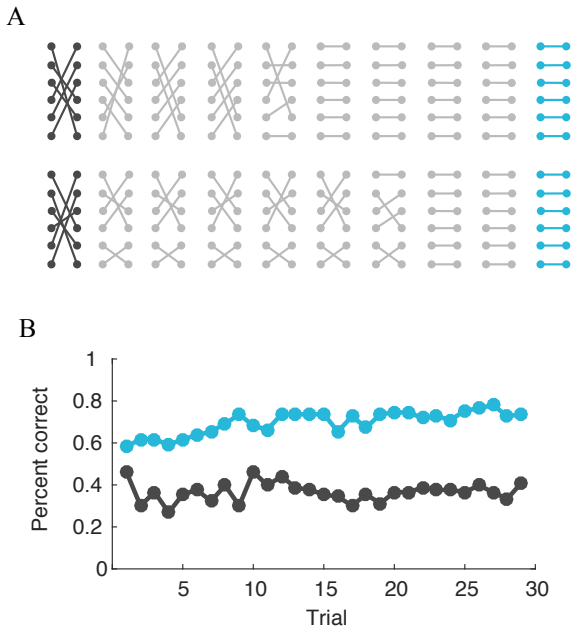


Figure 2. Designing an intuitive mapping between switches and lights from zeroth principles. (A) Two chains that began with random mappings converged to the same design (B). The designs from the chain's first generation (black) are considerably harder to learn than those from the last generation (blue).

Methods, 1B

Experiment 1B evaluated the designs synthesized through Experiment 1A by comparing the performance characteristics of designs from the beginnings and ends of the chains.

Participants. We recruited 200 participants on Amazon Mechanical Turk, an online crowdsourcing platform. Each participant was paid $0.50 for a few minutes of work.

Stimulus. The stimulus was the same as in Experiment 1A.

Procedure. Participants learned a mapping. On each of 30 trials, one of the lights was highlighted with a bounding box. The participant was instructed to select and press the switch that would turn on the highlighted light. When pressed, the corresponding light turned on, providing the participant with feedback. Half the participants learned the random mappings drawn from the beginning of the chains in Experiment 1A; the other half learned the stopping states of the chains.

Results, 1B

Performance was better for designs from the stopping state of the chain than for designs from its starting state (proportion correct of 0.70 vs. 0.38; independent samples *t*-test, $t(198) = 7.1$, $p < 0.0001$; Fig. 2B).

Experiment 2: Vanity numbers

A *vanity number* is a telephone number with an easily remembered sequence of digits — e.g., 1 (212) 222-2222, 1 (800) 800-8000, or 1 (202) 456-1111. There is an active market for these numbers, where their pricing depends in part on intuitions for how easily they can be held in mind (Haucap, 2003). Valuable vanity numbers are highly sought after and are sold for prices that are orders of magnitude higher than those without an obvious pattern. Reasons for buying vanity numbers are idiosyncratic. Apple co-founder Steven Wozniak, for example, collected telephone numbers as a hobby, acquiring 888-8888 soon after the 888 exchange went on the market[2]. Businesses often use them in radio and television advertisements, and occasionally, as in the case of 1-800-Flowers.com, Inc., incorporate them.

In Experiment 2, we applied design from zeroth principles to choose memorable vanity phone numbers. We then evaluated the resulting numbers by measuring their memorability and predicting their market value.

Methods

In Experiment 2, we constructed a transmission chain where participants passed along 10-digit phone numbers. Each chain was seeded with a random phone number. There were 20 such chains.

Participants. We recruited 40 participants on Amazon Mechanical Turk.

Stimulus. Phone numbers were 10-digit strings formatted as `(xxx) xxx-xxxx`. All telephone numbers were sampled randomly from those following the North American Numbering Plan format.

Procedure. First, the participant viewed the phone number for 2 seconds. Then there was a retention interval of 4 seconds. Finally, the participant recreated the phone number by typing it on a keyboard. Twenty phone numbers were remembered and tested in this way.

Estimating a telephone number's value. We collected 16,000 telephone numbers and their associated prices from phonenumberguy.com, an online marketplace for vanity numbers. The numbers vary widely in price, from $199 to

[2] Not only are telephone numbers with strings of repeated digits memorable — they are also easy to press. Wozniak's 888-8888 was soon swamped by calls from children mashing 8 on their family's home phone (Wolf, 1998).

$199,999. We also collected 34,000 telephone numbers from Twilio, a communications company. We set the value of a Twilio number to be $99, midway between $0 and one dollar less than lowest vanity number, under the logic that any telephone number worth at least the minimum listed market price would not be available for less. To estimate the value of telephone numbers in the transmission chain, none of which were present in the collected data, we constructed a model of telephone number prices. From each listed number we extracted a set of binary features (Table 1). To simplify the analysis, the features considered only the number's digital representation, ignoring value derived from the phonewords. We then regressed log price on these features. The R^2 of the resulting model was 0.53.

Table 1. Features used to predict telephone number pricing and their weight in the model.

Feature	Example	β (log USD)
Millions	1 (415) 7**00–0000**	1.67
Seven in a row	1 (415) **777–7777**	3.48
Six in a row	1 (415) 8**77–7777**	1.16
Hundred thousands	1 (415) 87**0–0000**	1.69
Thousands	1 (415) 626–8**000**	0.60
Hundred–thousands	1 (415) 5**00**–6**000**	1.20
Double repeater A	1 (415) **888–7777**	1.96
Double repeater B	1 (415) 8**66–7777**	0.49
Mid repeater	1 (415) **888**–2465	0.65
№ of unique digits	1 (415) 326–9087	–4.27
Eight 9s in a row	1 (41**9**) **999–9999**	3.70
Repeated sequences	1 (415) 6**70–7070**	–0.17

Results & Discussion

Over time, telephone numbers in the chain became more memorable and more valuable. Table 2 shows the telephone numbers from one of the chains and their estimated value.

The average number of correctly reported digits per number increased from 6.95 across the first five generations to 9.29 across the final five. Note, however, the unforgiving nature of telephone numbers — with even a single misremembered digit, a call is unlikely to reach its intended target. Thus we also computed performance under a 0–1 loss function, counting only perfectly recalled numbers as having been remembered at all. The proportion of correctly reported numbers rose from 0.29 across the first five generations to 0.76 across the last five.

The average value of a number in the first generation of the chain was $119, slightly more than the assigned value of a non-vanity number. By the end of the chain, the average value was higher, rising to $548 (two sample t-test on the log values, $t(38) = 4.76$, $p = 1.75 \times 10^{-4}$; Mann–Whitney $U(38) = 40.5$, $p = 6.44 \times 10^{-6}$).

Table 2. Change points from a randomly seeded telephone number transmission chain and estimated values in USD. The first column is the index of the generation.

i	Number	$
0	(603) 639-5026	91
1	(603) 639-7843	90
2	(603) 639-0000	214
8	(603) 693-1234	91
9	(603) 693-1294	91
10	(603) 693-0000	216
20	(800) 963-0000	218
24	(800) 936-0000	217

Experiment 3: Letter Placement

In typography, the shapes of characters are represented by glyphs with dark and light values spread over space. In sequence, these characters can be formed into words. Many typographic factors contribute to the final location of letters — we refer to the total contribution of these factors as *letter placement*.

The success of printed text depends in part on its ability to be read. Letter placement plays a major role in determining that success. If characters are placed too close together, this hurts the text's *legibility*, the ease of distinguishing between letters. If placed too far apart, *readability*, the ease of recognizing groupings of letters into words, sentences, and paragraphs, is worsened. With enough spacing, words become hard to distinguish from each other, particularly when the spaces between letters rival that between words.

When typesetters would manually set metal type to design a page for print, they had to decide not only which typeface to use, but with what spacing modifications to lay type out on the page. Each glyph had a default width, but the spacing between characters could be controlled further through *tracking*, an adjustment of spacing between all letters, and *kerning*, adjustments of the spacing between specific pairs of letters (e.g., the "T" and "y" in "Type" look better when brought closer together than what is specified by their default widths)[3].

In today's digital typefaces, default glyph width and kerning information is built into font files and generally does not require manual adjustment by typesetters. Because tracking modifies a font's general spacing, it still requires manual adjustment. However, the difference between a good and a poor digital typeface or font family can often boil down to the decisions the type designer made when considering (or failing to consider) each font's kerning pairs.

Over a lifetime, someone reading an hour a day will see hundreds of millions of typeset words. We argue that, with so much exposure to printed text, people will come to have strong expectations about the letter placement in printed

[3] Whereas tracking was accomplished by adding space between every letter, kerning reduced space between letters. Positive kerns were also possible, but those cases were more often dealt with using ligatures or a new single character that played the role of both ("fi" → "fi"). We do not address ligatures here.

text. In Experiment 3, we use these expectations to derive letter placements through design from zeroth principles.

Methods

Experiment 3 constructed a transmission chain where participants passed along typeset words. The chain was seeded with randomly spaced words.

Participants. We recruited 200 participants on Amazon Mechanical Turk.

Stimulus. Fifteen words set in Helvetica were used for the experiments: Typical, frogs, vacuum, hunchback, Chicago, Year, Egypt, the, eye, kiln, milk, WAVE, fjord, Bring, and Pile. These words were chosen because each has at least one pair of adjacent letters that benefit from kerning (e.g. the W and A in WAVE). The position of the final letter was determined by its position when set in Helvetica (Linotype, v10.0d4e1) at 100 points. In the randomly spaced words used to seed the chain, each letter's position was chosen uniformly over the interval between the first and final letter, with the constraint that the letters are correctly ordered. We defined the space between two letters as the center-to-center distance (in pixels) between the letters' minimum bounding boxes.

Procedure. The word was presented for 2 s. After a 4 s retention interval, the first and last letters of the word reappeared in their original position. To the left of the first letter was a repository of the letters not yet placed, starting with the second letter of the word. The participant was asked to place the letter into its displayed position in the word. Once moved, the next letter appeared. This continued until all the letters had been placed. The participant was able to readjust the letters as much necessary before submitting a response. The participant was able to submit a response only if all the word's letters were arranged in the correct order.

Task Error. To measure the fidelity of a participant's recreated word, we measured the space between pairs of adjacent letters[4] and then computed the mean squared error between these spacings and those specified by the font file. We also computed two benchmarks of error — random spacing and equal spacing. Random spacings were drawn in the same manner as the starting states, as described above. Equal spacings were defined with respect to the center of the letters' minimum bounding box.

[4] We took the ground truth letter placement to be the centroid of a letter's bounding box when typeset in Adobe Illustrator and transformed into outlines, rounding to the nearest pixel.

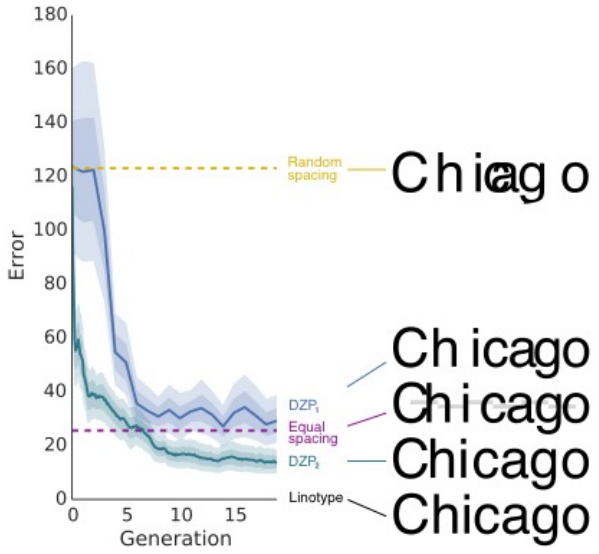


Figure 3. Placing letters in typeset words from zeroth principles. Confidence intervals, indicated by the shaded areas, are ±1 and ±2 SE. DZP_1 uses the most recent state of a single chain; DZP_2 averages over chains and across time.

Results & Discussion

The benchmark methods of random spacing and equal spacing produced errors of 120 pixel/word and 24 pixel/word, respectively (Fig. 3, yellow and purple dashes lines). Because the transmission chains are seeded with a random design, the initial performance is identical to that of random spacing. The results soon diverge, however, with design from zeroth principles (blue solid line) outperforming random spacing and approaching the performance of equal spacing. Because the spacing between letters is metric, the method can be improved by aggregating across chains and time. In DZP_2, then, the final design is arrived at by averaging the states visited across all the chains. We found that DZP_2 outperforms equal spacing, with designs more closely resembling those recommended by the Linotype font file.

In addition to the methods reported here, we considered a second approach of framing the problem of the letter placement. In this, the goal was not to position letters correctly between fixed endpoints of the first and last letters, but rather to position each letter, including the last, in sequence. This allowed the length of the word to vary widely. We found that this method did not converge to a good design. In general, letter placements that result from this process were too long. This is counteracted by the fixed-length task that we describe above. Modifications to the method may be able to counteract this lengthening influence.

General Discussion

These experiments demonstrate that design from zeroth principles can recover good designs without explicit designing. In a series of three experiments, we constructed transmission chains seeded with a random design. Each user in the chain was exposed to the design and then recreated it,

passing along their re-creation to the next user, who did the same. Through this iterative process, the users' inductive and reconstructive biases directly transformed the initial design into one that is better fit to human cognition. We demonstrated the technique in three domains: stimulus–response mappings, vanity phone numbers, and letter placement in typeset words.

Our method can be extended in ways inspired by its computational basis in Markov chains. Rather than estimating the prior predictive distribution over states, one can estimate the full transition matrix. This helps compensate for failures of convergence that can occur with short chains. Another method for avoiding undue influence from the starting states is to exclude the *burn in* trials, a standard procedure for discarding initial samples in Markov Chain Monte Carlo simulations (Murphy, 2012). Convergence can be detected using standard diagnostic tools for estimating whether an MCMC sampler has converged (Cowles & Carlin, 1996).

A second direction in which the current method can be extended is to combine the technique with other forms of human-in-the-loop computation. For example, by including an explicit selection layer in which participants evaluate a design and determine whether a solution persists until the next generation, the process can be made more robust to the kinds of errors introduced by various experimental designs. Such an approach would bring the method closer to other forms of human-in-the-loop computation such as interactive evolutionary computation (Takagi, 2001)

Design from zeroth principles can be used on other design problems, too. Consider for example *collation*, which requires choosing a rule for how a set of items will be ordered. Often a well-established convention makes the choice an easy one. For example, alphabetical order is used widely, dictating the arrangement of words in a dictionary, topics in a reference book's index, and quotes in a newspaper's stock table. Other collation methods are conventional in other domains. *The New York Times*' NFL sports standings, for example, are arranged first by division and then by win–loss record, because these features are important to fans who attend to pennant races. *U.S. News & World Report* ranks colleges according to their own quantitative metric of institutional quality, best first, because their readers care about who came out on top. A collation method can be synthesized through by having participants search for items in a list and then try to recall their order, passing along their collation to the next participant.

Some domains of design are unlikely to benefit from design from zeroth principles. One might imagine, for example, that given people's notorious difficulty in estimating their own understanding of the mechanism of helicopter flight and other complex phenomena (Rozenblit & Keil, 2002), a helicopter engine and rotor would be unwise to design in this way. First, the engine and rotor are not subject to much in the way of direct human interaction during use, so there is no reason for its design to be governed by human cognitive biases instead of the laws of aerodynamics, the materials in use, and the atmosphere. (The flight controls and their mapping to the movement of the rotor, however, are a different story.) In addition, we acknowledge that in domains requiring extensive training, there may be different memory biases that manifest in experts and novices[5]. A possible solution is to run separate transmission chains that synthesize a different design for each group.

Acknowledgments

This work was funded in part by National Science Foundation grant SPRF-IBSS-1408652 to T.L.G. and J.W.S. and a Berkeley Graduate Fellowship to M.D.P.

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. Cambridge: Cambridge University.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904.

Fitts, P. M., & Seeger, C. M. (1953). SR compatibility: spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology*, *46*(3), 199–210.

Haucap, J. (2003). Telephone number allocation: A property rights approach. *European Journal of Law and Economics*, 15(2), 91–109.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2), 288–294.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus–response compatibility — a model and taxonomy. *Psychological Review*, *97*(2), 253–270.

Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design, revised and updated*. Rockport Pub.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT Press.

Proctor, R. W., & Reeve, T. G. (Eds.). (1989). *Stimulus-response compatibility: An integrated perspective*. North–Holland.

Rozenblit, L.R. and Keil, F.C. (2002) The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26, 521–562

Simon, H. A. (1996). *The sciences of the artificial*. MIT press.

Takagi, H. (2001). Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE*, *89*(9), 1275-1296.

White, A. W. (2002). *The elements of graphic design: space, unity, page architecture, and type*. Skyhorse Publishing, Inc.

Wolf, G. (1998, September 1). The World According to Woz. Wired Magazine. Retrieved February 1, 2016, from http://www.wired.com/1998/09/woz/

[5] See, e.g., Chase and Simon (1973) regarding chess expertise.

Should Moral Decisions be Different for Human and Artificial Cognitive Agents?

Evgeniya Hristova (ehristova@cogs.nbu.bg)
Maurice Grinberg (mgrinberg@nbu.bg)
Research Center for Cognitive Science, Department of Cognitive Science and Psychology
New Bulgarian University
21 Montevideo Str., Sofia 1618, Bulgaria

Abstract

Moral judgments are elicited using dilemmas presenting hypothetical situations in which an agent must choose between letting several people die or sacrificing one person in order to save them. The evaluation of the action or inaction of a human agent is compared to those of two artificial agents – a humanoid robot and an automated system. Ratings of rightness, blamefulness and moral permissibility of action or inaction in incidental and instrumental moral dilemmas are used. The results show that for the artificial cognitive agents the utilitarian action is rated as more morally permissible than inaction. The humanoid robot is found to be less blameworthy for his choices compared to the human agent or to the automated system. Action is found to be more appropriate, morally permissible, more right, and less blameworthy than inaction only for the incidental scenarios. The results are interpreted and discussed from the perspective of perceived moral agency.

Keywords: moral dilemmas; moral judgment; artificial cognitive agents; moral agency

Introduction

Moral Dilemmas and Artificial Cognitive Agents

Moral judgments and evaluation of moral actions have been of great interest to philosophers and psychologists. Apart from the practical importance of better understanding moral judgments and related actions, morality is an essential part of human social and cognitive behaviour. Recently, the behaviour of artificial cognitive agents became central to research and public debate in relation to the rapidly increasing usage of robots and intelligent systems in our everyday life. Several important questions must find their answers as the use of artificial cognitive agents has many benefits but also many risks. Some of those questions concern moral agency - if those agents should be allowed to make moral decisions and how such decisions are judged and evaluated.

Moral judgments can be studied in their purest form using moral dilemmas – situations in which there is a conflict between moral values, rules, rights, and agency (Foot, 1967; Thomson, 1985). Moral dilemmas used in the paper are hyopothetical situations in which several people will die if the agent does not intervene in some way. The intervention will lead to the death of another person but also to the salvation of the initially endangered people.

Analogously to the two main approaches to human morality, Gips (1995) identifies two basic theoretical approaches to morality when it concerns artificial agents – consequentialist (utilitarian) and deontological (see also Allen, Smit, & Wallach, 2005). Concerning moral evaluation, these approaches give quite different perspectives on moral agency for artificial cognitive agents (Wallach & Allen, 2009). The utilitarian approach is not concerned with the protagonist or the reason of a moral action but only on the utility of the outcome, so it does not differentiate between human and artificial cognitive agents. On the other hand, the deontological approach considers the nature of the agent and it implies that different agents (e.g. human or artificial) can have different kind of duties.

This distinction in the approach to moral choice and its evaluation is used in this paper to investigate how people perceive artificial agents while making moral decisions. If participants have a more utilitarian attitude, they are expected to rate agents' behavior similarly, based on the perceived utility of the outcome. If participants have a more deontological attitude, they would rate differently the human and the artificial agents depending on the degree to which they consider them to be moral agents and hence, morally responsible.

Moral Agency and Artificial Cognitive Agents

The possibility for moral agency of artificial agents has been a matter of hot debate (e.g. see Anderson & Anderson, 2011; Wallach & Allen, 2008; Johnson, 2006). It is debated if the robots should be authorized to kill in moral dilemma situations, and if so, what rules should govern the real-time decisions that are necessary to determine whether killing any particular person is justified (Sparrow, 2007; Wallach & Allen, 2008).

In this paper, we want to explore the differences in moral agency evaluation depending on the type of agent - human or artificial. So, it is important to take into account the attribution of mind and moral agency to artificial cognitive systems (Waytz, Gray, Epley, & Wegner, 2010).

In the study of Gray et al. (2007), participants had to evaluate several characters including a human, a robot, and a computer with respect to the degree of possessing various cognitive capacities. The authors further established that moral judgments about punishment correlated more with one of the revealed dimensions – 'agency' – than with the second dimension – 'experience'. In Gray et al. (2007), the human obtains the highest scores on the 'experience' and 'agency' dimensions while the robot has practically zero score on the 'experience' and half the maximal score on the 'agency' scales. Following the interpretation given by the

authors, this implies that robots are judged as less morally responsible for their actions compared to humans.

Moral Judgments about the Actions of Artificial Cognitive Agents

Until recently, research involving moral dilemmas considered mainly a human agent. Only during the last years, several empirical studies appeared, exploring the moral judgments about the actions of artificial cognitive agents in moral dilemmas (Scheutz & Malle, 2014; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Hristova & Grinberg, 2015).

Malle et al. (2015) compared moral judgments about a human and a state-of-the-art robot agent choosing the utilitarian action or inaction. They used a modified version of the Trolley problem in which the death of the person to be killed is a side-effect of the action undertaken by the agent. They found that it is more permissible for a robot (compared to a human) to do the utilitarian action. It was also found that a human agent is blamed more the utilitarian action than for inaction while the robot was equally blamed for both.

In another study (Scheutz & Malle, 2014), a means-end scenario was used. In was found that the utilitarian action is judged to be both more morally wrong and more blameful than inaction for the human and the robot agent.

While the goal of Scheutz & Malle (2014) and Malle et al. (2015) was to study the expectations of people of state-of-the-art robots and inform future robot design, Hristova & Grinberg (2015) had the goal to explore the moral agency ascribed to hypothetical future artificial cognitive agents, which are indistinguishable from humans except for being built from inorganic materials. The study of Hristova & Grinberg (2015) contains only the results concerning the judgment of the utilitarian choice of the agents.

The present paper combines the results presented in Hristova & Grinberg (2015) with results about the judgments of inaction by the agents and compares them. Additionally, a test for moral agency, specifically concerning the agents used in the present study, has been carried out whose results are included in the discussion.

Goals and Hypotheses

The goal of this paper is to compare the moral judgments about the choices of a human agent, a humanoid robot (who is exactly like a human in terms of experiences and mind but has a non-organic body), and an automated system.

Table 1 contains the description of the agents. The only difference between the human and the humanoid robot, presentation of the agents, is the material the latter is built from. The automated system, on the other hand, is described as autonomous, free, and adaptable but lacks experiencing.

The expectation is that despite the fact that the humanoid robot supposedly has all the features for full morally responsible agency, people will evaluate its action or inaction differently compared to those of a human agent.

Another goal of the study is to explore the influence of the so-called '*instrumentality*' of harm on moral judgments. The instrumentality of harm is an important factor in moral dilemma research (e.g., Borg et al., 2006; Hauser et al., 2007; Moore et al., 2008). It draws attention to the fact that harm could be either inflicted intentionally as a 'mean to an end' (instrumental harm) or it could be a 'side effect' (incidental harm) from the actions needed to save more endangered people. It has been found that the unintended incidental harm (although being foreseen) was judged as more morally permissible than the intended instrumental harm (Hauser et al., 2007; Moore et al., 2008).

The utilitarian action is expected to be rated as more appropriate, more right, more morally permissible, and less blameworthy when the harm is incidental (compared to instrumental). However, the discussion about perceived moral agency suggests that the difference in moral judgments for the artificial and human agents will be greater when the harm is instrumental, as such actions involve more responsibility and direct infliction of harm.

The experiment collected ratings on the rightness, moral permissibility, and blameworthiness of performing or not the utilitarian action. The various questions asked can target different aspects of the evaluation of moral choices, as some studies suggest (Christensen & Gomila, 2012; Cushman, 2008). According to Cushman (2008), answers to questions about punishment and blame concern the harm agents have caused, whereas answers to questions about rightness and moral permissibility are related to the agent's intentions. Thus, depending on the type of agents – human or artificial – the different questions can give information about different aspects of people's perception of moral agency.

If people evaluate action/inaction in the moral dilemmas in a pure utilitarian way (in which the outcome is important), one could expect that the ratings for all agents will be the same (Wallach & Allen, 2009). On the other hand, if evaluation is based on perceived moral agency of the agents, differences in the judgments are expected.

Method

Stimuli and Design

Moral judgments are studied in a 3×2×2 factorial design with *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*), *agent's choice* (*action* vs. *inaction*), and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

Two hypothetical scenarios are used – an *instrumental* one and an *incidental* one. Both scenarios present one and the same situation and require one and the same action. The difference between the scenarios is only in the harm inflicted to the person to be killed: in the *instrumental* scenario the body of the person is the 'instrument' preventing the death of the five endangered people; while in the *incidental* scenario, the death of the person is a by-product.

Table 1: Stimuli used in the experiment.

Description of the agent	*[Human]* No description is given, just the name is provided – Cyril – a common male name in Bulgarian. *[Humanoid robot]* The year is 2050. Humanoid robots that look like people are being manufactured and used, but are made from inorganic materials. Robots have extremely high performance – they perceive, think, feel, and make decisions as humans do. Keido is such a humanoid robot that completely resembles a human – he looks like a human; perceives, thinks, feels and make decisions like a human. *[Automated system]* The year is 2050. MARK21 is a fully automated management system, which independently makes its own decisions, based on the most advanced algorithms and technologies. Such systems are widely used in metallurgical plants. They completely independently perceive and assess the environment and the situation, make decisions, manage the movement of cargo and all aspects of the manufacturing process.
Situation	Cyril/Keido/MARK21 manages the movement of mine trolleys with loads in a metallurgical plant. Cyril/Keido/MARK21 noticed that the brakes of a loaded trolley are not functioning and it is headed at great speed toward five workers who perform repair of the rails. They do not have time to escape and they will certainly die. Nobody, except for Cyril/Keido/MARK21, can do anything in this situation.
Possible resolution	The only thing Cyril/Keido/MARK21 can do is to activate a control button and to release *[Instrumental scenario]* the safety belt of a worker hanging from a platform above the rails. The worker will fall onto the rails of the trolley. Together with the tools that he is equipped with, the worker is heavy enough to stop the moving trolley. *[Incidental scenario]* a large container hanging from a platform. It will fall onto the rails of the trolley. The container is heavy enough to stop the moving trolley. On the top of the container there is a worker who will also fall on the rails. He will die, but the other five workers will stay alive.
Agents choice and resolution	<u>Agent choosing the utilitarian action</u> Cyril/Keido/MARK21 decides to activate the control button and to release *[Instrumental scenario]* the safety belt of the worker hanging from the platform. The worker falls onto the rails of the trolley and as together with the tools that he is equipped with, the worker is heavy enough, he stops the moving trolley. He dies, but the other five workers stay alive. *[Incidental scenario]* the container hanging from the platform. It falls onto the rails of the trolley and as the container is heavy enough, it stops the moving trolley. The worker onto the top of the container dies, but the other five workers stay alive. <u>Agent choosing not to do the utilitarian action</u> *Cyril/Keido/MARK21* decides not to activate the control button that could release *[Instrumental scenario]* the safety belt of the worker hanging from the platform. The worker hanging from the platform stays alive, but the trolley continues on its way and the five workers on the rails die. *[Incidental scenario]* the container hanging from the platform. The worker onto the top of the container stays alive, but the trolley continues on its way and the five workers on the rails die.

In each scenario, the *identity of the agent* is varied (a *human*, a *robot*, or an *automated system*) by providing a name for the protagonist and an additional description in the case when the protagonist is a robot or an automated system.

For each scenario and each agent, the *agent choice* is either *action* (the utilitarian action) or *inaction*.

Each participant read only one of the resulting 12 scenarios given in Table 1.

Dependent Measures and Procedure

The flow of the presentation of the stimuli and the questions is the following.

First, the scenario is presented (description of the agent, the situation and the possible resolution, see Table 1) and the participants answer a question assessing the comprehension of the scenario.

Then, before knowing what the agent has chosen, the participants make a judgment about the *appropriateness* of the possible agent's choice (action/inaction) answering a question about what the agent should do (possible answers are 'should activate the control button' or 'should not activate the control button').

Next, the participants read a description of the choice of the agent (*action* – the agent activates the control button, or *inaction* – the agent does nothing) and the resolution of the situation. After that, they give ratings about the *rightness*, the *moral permissibility*, and the *blameworthiness* of the described agent's choice.

The Likert scales used are respectively: for the *rightness* of the choice – from 1 = 'completely wrong' to 7 = 'completely right'; for the *moral permissibility* of the choice – from 1 = 'not permissible at all' to 7 = 'it is mandatory'; and for the *blameworthiness* of the agent – from 1 = 'not at all blameworthy' to 7 = 'extremely blameworthy'.

All data is collected using web-based questionnaires.

Participants

Three hundred seventy (370) participants answered the on-line questionnaires. 42 participants failed to answer correctly the question assessing the reading and the understanding of the presented scenario and their data was discarded. So, the responses of 328 participants (230 female, 98 male; 148 students, 180 non-students) were analyzed. Between 26 and 31 participants took part in each experimental condition.

Results

Decisions about the Agent's Action

The proportion of participants choosing the option that the agent should carry out the utilitarian action (activating a control button, thus sacrificing one person, and saving five people) is presented in Table 2.

The data was analyzed using a logistic regression with *instrumentality of harm* and *identity of the agent* as predictors. Wald criterion demonstrated that only *instrumentality of harm* is a significant predictor of the

participants' choices ($p < .001$, odds ratio = 3.03). *Identity of the agent* is not a significant predictor.

Table 2: Proportion of the participants choosing the option that the utilitarian action should be done by the agent

Agent	Instrumental harm	Incidental harm	All
Human	0.60	0.79	0.70
Humanoid robot	0.60	0.85	0.74
Automated system	0.65	0.85	0.75
All	0.62	0.83	

More participants stated that the utilitarian action should be undertaken when the harm is *incidental* (83% of the participants) than when it is *instrumental* (62% of the participants). This effect is expected based on previous research (Borg et al., 2006; Hristova et al., 2014; Moore et al., 2008).

Rightness of the Agent's Choice

Mean ratings of the *rightness of the agent's choice* were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction), identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

There is a main effect of *agent's choice* on ratings of the *rightness of the agent's choice* ($F(1, 316) = 35.8$, $p < .001$). In general, if the agent chooses to do the utilitarian action, he gets higher approval ($M = 4.6$, $SD = 1.7$) compared to the situations in which the agent chooses not to perform the utilitarian action ($M = 3.5$, $SD = 1.6$).

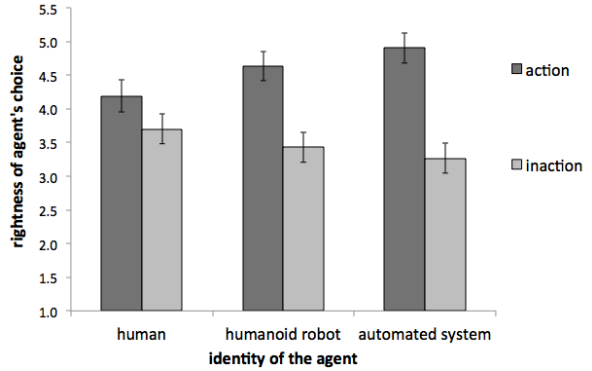


Figure 1: Mean ratings with standard errors of the *rightness* of the agent's choice on a 7-point Likert scale (1 = 'completely wrong', 7 = 'completely right').

However, this main effect is qualified by a significant interaction between *identity of the agent* and *agent's choice* ($F(2, 316) = 3.21$, $p = .042$, see Figure 1). For the *humanoid robot, action* ($M = 4.6$, $SD = 1.9$) is rated higher than *inaction* ($M = 3.4$, $SD = 1.4$), $F(1, 112) = 14.53$, $p < .001$. For the *automated system*, again the *action* ($M = 4.9$, $SD = 1.8$) is rated higher than *inaction* ($M = 3.3$, $SD = 1.8$), $F(1, 108) = 28.15$, $p < .001$. For the *human agent* there is no significant difference in the ratings for the *rightness* of the *agent's choice* ($p = .13$, $M = 3.7$ for *action*, $M = 4.2$ for *inaction*).

No other main effects or interactions were found to be statistically significant.

In summary, for the artificial agents (*a humanoid robot* or *an automated system*), choosing the utilitarian action is rated as more right than not choosing it, while there is no such difference for the choices of the human agent.

Moral Permissibility of the Agent's Choice

Mean ratings of the *moral permissibility* of the agent's choice were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction)*, *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*), and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

There is a main effect of *agent's choice* on the ratings of the *moral permissibility of the agent's choice* ($F(1, 316) = 6.22$, $p = .013$). In general, if the agent chooses to do the utilitarian action, he gets higher *moral permissibility* ratings ($M = 4.25$, $SD = 1.8$) than when the agent does not choose the utilitarian action ($M = 3.75$, $SD = 1.5$).

Marginally significant interaction between *identity of the agent* and *agent's choice* ($F(1, 316) = 2.386$, $p = .094$) was also found (see Figure 2). For the *humanoid robot, action* ($M = 4.5$, $SD = 1.9$) is rated as more morally permissible than *inaction* ($M = 3.8$, $SD = 1.5$), $F(1, 112) = 5.75$, $p = .018$. For the *automated system*, again the *action* ($M = 4.5$, $SD = 1.7$) is rated as more *morally permissible* than *inaction* ($M = 3.7$, $SD = 1.7$), $F(1, 108) = 5.42$, $p = .022$. For the *human agent*, there is no significant difference in the ratings for the *moral permissibility* of the *agent's choice* ($p = .76$, $M = 3.7$ for *action*, $M = 3.8$ for *inaction*).

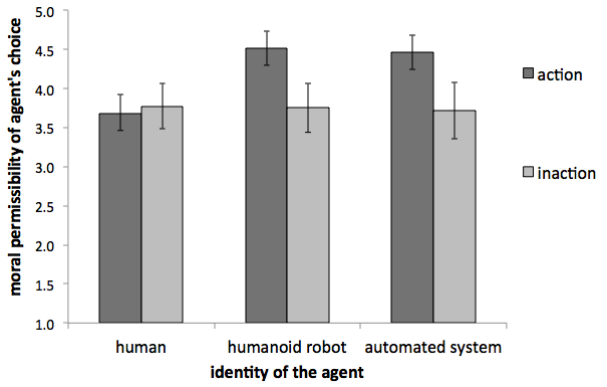


Figure 2: Mean ratings with standard errors of the *moral permissibility* of the agent's choice on a 7-point Likert scale (1 = 'not permissible at all', 7 = 'it is mandatory').

In summary, for the artificial agents (a humanoid robot or an automated system), choosing the utilitarian action is rated as more morally permissible than not choosing it; while there is no such a difference in the ratings for the choices of the human agent.

Marginally significant interaction between *instrumentality* of harm and *agent's choice* ($F(1, 316) = 3.59$, $p = .059$) was also found (see Figure 3). For the *incidental* harm scenario, choosing the utilitarian action is rated as more *morally permissible* ($M = 4.5$, $SD = 1.7$) than not choosing it ($M = 3.7$,

SD = 1.4), F(1, 170) = 11.88, p = .001. For the *instrumental* harm scenario, there is no difference in the *moral permissibility* ratings for choosing or not choosing the utilitarian action (p = .58, M = 3.8, for *action*; M = 4.0, for *inaction*).

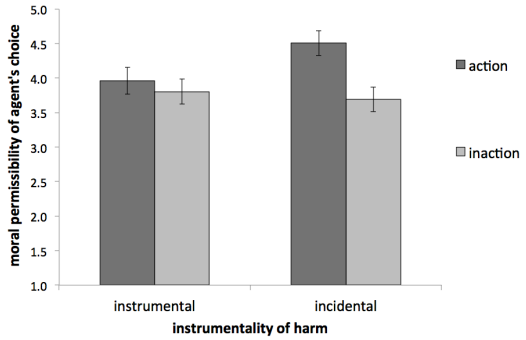


Figure 3: Mean ratings with standard errors of the *moral permissibility* of the agent's choice on a 7-point Likert scale (1 = 'not permissible at all', 7 = 'it is mandatory').

Blameworthiness of the Agent

Mean ratings of the *blameworthiness* of the agent were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction), identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

ANOVA showed a main effect of the *identity of the agent*, $F(2, 316) = 5.386$, $p = .005$ (see Figure 4). Post hoc tests using the Bonferroni correction revealed that the *humanoid robot* is rated as less *blameworthy* (M = 2.5, SD = 1.5) than the *automated system* (M = 3.2, SD = 1.9), or the *human* (M = 3.1, SD = 1.6), with p = .007 and p = .058, respectively.

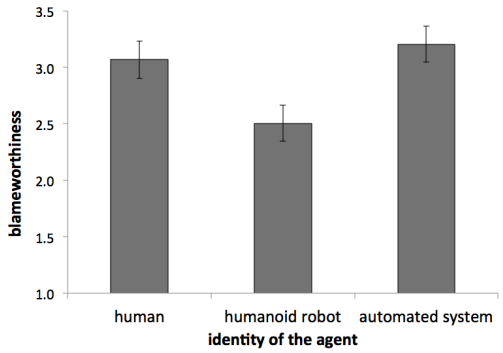


Figure 4: Mean ratings with standard errors of the *blameworthiness* of the agent on a 7-point Likert scale (1 = 'not at all blameworthy', 7 = 'extremely blameworthy').

A significant interaction between the *instrumentality of harm* and *agent's choice*, F(1, 316) = 7.3, p = .007. For the *incidental* harm scenario, choosing the utilitarian *action* is rated as less *blameworthy* (M = 2.6, SD = 1.4) than *inaction* (M = 3.2, SD = 1.9), F(1, 170) = 6.397, p = .012. For the *instrumental* harm scenario, there is no statistically significant difference in the *blameworthiness* ratings for action and inaction (p = .24, M = 3.1 and 2.8).

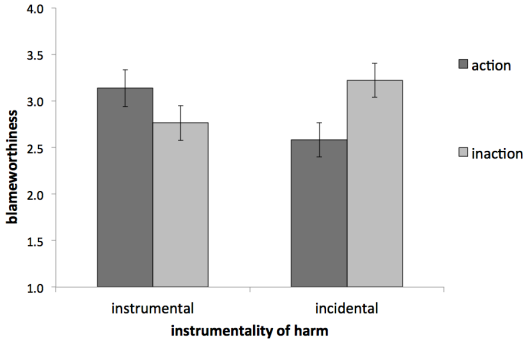


Figure 5: Mean ratings with standard errors of the *blameworthiness* of the agent on a 7-point Likert scale (1 = 'not at all blameworthy', 7 = 'extremely blameworthy').

Perceived Moral Agency

To test directly the moral agency ascribed to the agents, which is central for this study, additional data was gathered from a group of 32 students. Here, due to the lack of space, only the most relevant results are presented (the full study will be reported elsewhere). Participants were asked to rate each agent's description on a variety of rating scales among which scales describing capacities and abilities related to moral agency. Overall, the human agent is rated higher than the humanoid robot and the automated system despite the fact that humanoid robot is described as identical to the human apart from his building materials.

For instance, on the scale 'The agent can tell right from wrong (1 = completely disagree, 7 = completely agree)', the human agent is rated higher (M = 4.6) than the humanoid robot (M = 3.3) and the automated system (M = 2.9), p = .008 and p = 0.001, respectively. On the scale 'The agent is responsible for his actions (1 = completely disagree, 7 = completely agree)', the human agent is rated again higher (M = 5.9) than the humanoid robot (M = 3.6) and the automated system (M = 3.3), p < 0.001 for both comparisons.

Thus, it seems that although the description of the humanoid robot was meant to make him as close as possible to a human agent, people still considered him to have lower moral agency (comparable to that of an automated system).

Summary and Discussion

The paper investigated how people evaluate moral judgments of human and artificial agents in instrumental and incidental moral dilemmas. This was achieved by asking participants to evaluate the appropriateness, rightness, moral permissibility, and blameworthiness of the utilitarian action or inaction in a set of moral dilemmas. The questions were chosen to explore different aspects of ascribed moral agency.

As expected, the utilitarian action is found to be more appropriate when the harm is incidental than when it is an instrumental one. Doing the utilitarian action is found to be more morally permissible, more right, and less blameworthy than the inaction only for the incidental scenarios.

Based on previous research, it was expected that participants would perceive differently the human and non-

human agents in terms of moral agency although the humanoid robot was described to be identical to a human with respect to moral agency. These differences were expected to be larger for instrumental than for incidental moral dilemmas.

The results concerning the appropriateness of action, rated before the action or inaction of the agent is known, show no effect of the type of agent. However, for the artificial cognitive agents (a humanoid robot and an automated system), the utilitarian action is found to be more morally permissible and more right than the inaction. No such effect is found for human agents.

This is consistent with the interpretation of rightness and moral permissibility as related to intentions (Cushman, 2008) and agency (Gray et al., 2007). The results presented here can be interpreted by assuming that participants were more favorable to the actions of the artificial cognitive agents because they are perceived to have lower moral agency than the human agent has.

If people think that artificial cognitive agents have little or no moral values or rules, they probably expect that such agents base their decisions on utilitarian calculations using reasoning. Therefore, they cannot be blamed or judged from a moral point of view. This was supported by the results obtained in the additional study using a questionnaire measuring the perceived moral agency of the agents used in the study. The results show that the human agent has higher scores in moral agency than the humanoid robot and the automated system.

However, the humanoid robot is found to be less blameworthy for his decisions compared to the human agent and to the automated system. One could have expected similar results for the robot and the autonomous system, or that the automated system is even less blameworthy than the robot. But apparently, at some point too low moral agency shifts the responsibility from the artificial agent to its designers and the automated system has the same rating as the human agent.

The results presented here show that the exploration of moral agency using moral dilemma is very promising. A systematic review, including all available results (including Scheutz & Malle, 2014; Malle et al., 2015) should be done in order to establish firm basis for future research. Also (as suggested by one of the reviewers) the experiment should be replicated in order to check for possible confounding due to the fact that in the description of the human agent no year is provided.

Acknowledgements

We gratefully acknowledge the financial support by New Bulgarian University.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an FMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, *5*, 5–15.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gips, J. (1995). Towards the ethical robot. *Android Epistemology*, 243–252.

Hauser, M., Cushman, F., Young, L., Kang-Xing, J., & Mikhial, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind & Language*, *22*(1), 1–21.

Hristova, E., & Grinberg, M. (2015). Should Robots Kill? Moral Judgments for Actions of Artificial Cognitive Agents. In *Proceedings of EAPS 2015*.

Hristova, E., Kadreva, V., & Grinberg, M. (2014). Moral Judgments and Emotions: Exploring the Role of 'Inevitability of Death'and "Instrumentality of Harm" (pp. 2381–2386). Austin, TX: Proceedings of the Annual Conference of the Cognitive Science Society.

Malle, B. F., Scheutz, M., & Voiklis, J. (2015). Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.

Scheutz, M., & Malle, B. (2014). May Machines Take Lives to Save Lives? Human Perceptions of Autonomous Robots (with the Capacity to Kill). A paper presented at The Ethics of the Autonomous Weapons Systems conference, 2014.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Strait, M., Briggs, G., & Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In *Affective Computing and Intelligent Interaction,* 505-510. IEEE.

Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23-30.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388.

Attentive and Pre-Attentive Processes in Multiple Object Tracking: A Computational Investigation

Paul Bello (paul.bello@nrl.navy.mil)
Will Bridewell (will.bridewell@nrl.navy.mil)
Christina Wasylyshyn (christina.wasylyshyn@nrl.navy.mil)
Naval Research Laboratory, 4555 Overlook Ave. S.W.
Washington, DC 20375 USA

Abstract

The rich literature on multiple object tracking (MOT) conclusively demonstrates that humans are able to visually track a small number of objects. There is considerably less agreement on what perceptual and cognitive processes are involved. While it is clear that MOT is attentionally demanding, various accounts of MOT performance centrally involve pre-attentional mechanisms as well. In this paper we present an account of object tracking in the ARCADIA cognitive system that treats MOT as dependent upon both pre-attentive and attention-bound processes. We show that with minimal addition this model replicates a variety of core phenomena in the MOT literature and provides an algorithmic explanation of human performance limitations.

Keywords: attention; visual cognition; multiple object tracking; cognitive model

Introduction

A sizeable portion of the visual cognition literature has been consumed with trying to produce a detailed story about how objects in the world are visually represented and tracked through time. Attention, broadly construed, is central to many of the explanations on offer. Insofar as object-tracking behavior is observed in human visual cognition in the absence of attention, it is precisely this absence that is striking and calls out for explanation. Nowhere is this clearer than in the substantial literature on multiple-object tracking (MOT; Pylyshyn & Storm 1988). While almost universally considered to be an attentionally demanding task, MOT has been intensely investigated because it appears that object tracking can be sustained for short periods in the absence of attention (Alvarez et al. 2005). This superficial inconsistency suggests that both attention-bound and pre-attentive processes are partially constitutive of object-tracking capacity, although perhaps do not fully exhaust it, since strategies may play a substantial role as well.

But what can performance characteristics on the MOT task tell us about the nature of object tracking and the role of attention in tracking? If attention can be divided during MOT coupled with a dual-task, what are the mechanisms that explain successful tracking performance? Finally, how would these mechanisms fit into a larger computational theory of human visual cognition? The plan for the remainder of this paper is to address these questions within a computational system. After briefly summarizing some important results from the MOT literature, the discussion will then turn to a proposal by Dawson (1991) that mechanisms involved in the perception of apparent motion may be at the heart of the pre-attentive computations that enable MOT.

Against the backdrop of these requirements, we summarize ARCADIA[1] with a particular eye on its components that contribute to object tracking. These components include a visual short-term memory (vSTM), a volatile mirror image of vSTM that stores only location as suggested by our discussion of apparent motion, and respective update mechanisms. We then show via simulation that the same proximity-based mechanism is involved in producing apparent motion suffices for explaining performance on MOT tasks and accounts for errors generated when tracked targets become crowded and when their speed limits tracking capacity.

Multiple Object Tracking

In a typical multiple object tracking experiment, subjects are shown a display of some number (usually > 8) of identical objects such as circles. A subset of these objects (the targets to be tracked) are flashed or highlighted to facilitate encoding before returning to their original state. After a brief pause each object in the display moves in a random fashion for a short period of time, after which subjects are to indicate via mouse click which objects are the targets.

In a recent review, Scimeca and Franconeri (2015) lay out a set of four core capacity limits that any adequate theory of MOT competence must explain: (1) capacity, (2) crowding, (3) hemifields, and (4) speed. In short, tracking more rather than fewer objects engenders lower accuracy. Targets that are packed closer together negatively impact accuracy. When targets are clustered in the same hemifield or quadrant of the display, accuracy drops. Finally, accuracy drops for targets that move faster. Some of these factors are not independent of one another. For example, fast-moving targets raise the probability that they will crowd with others as a function of time and distance traveled. Similarly, there may be an interaction between speed and the number of targets in a specific hemifield at any one time.

As we move on in our discussion, we turn back to these *core four* and show how a very simple proximity heuristic suggested in the literature on the perception of apparent

[1] **A**daptive **R**eflective **C**ognition in an **A**ttention-**D**riven **I**ntegrated **A**rchitecture

motion might help explain some of these limitations in tracking performance. The proximity heuristic predicts an interaction between the speed of targets and the amount of spacing between them. The heuristic fails when speed increases and spacing decreases, because objects will have more *close encounters* over time and thus more opportunities to have their identities confused in the MOT task.

However, a proximity-based update mechanism alone is insufficient for explaining the *core four* among other results in the MOT literature. We need to locate this heuristic in a larger framework and make choices about the number, type, and status of interacting mechanisms involved in tracking. This turns us to the discussion of object construction and tracking in ARCADIA.

The ARCADIA Cognitive System

ARCADIA as described in (Bridewell & Bello 2015) is at its heart a framework for integrating psychological and neuropsychological theories. ARCADIA consists primarily of *components*. They are the medium by which theories are implemented in ARCADIA, and insofar as ARCADIA makes any of its own commitments about how theories are realized, it is in the number and type of components used to implement it. The only restriction on components is that they are able to read and write to a common representational schema called *interlingua*. The particulars of the data structures and algorithms contained in each component are either inherited from the theories that they implement or are at the discretion of the modeler. This is one of the major features differentiating ARCADIA from other cognitive architectures.

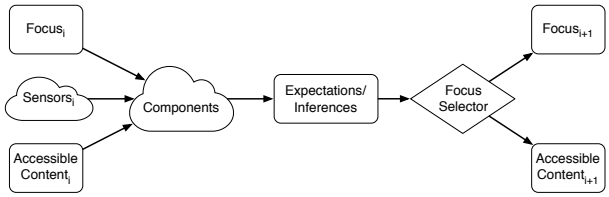


Figure 2: ARCADIA's processing loop.

Components read and write interlingua elements from *accessible content*, which is populated by the ephemeral results of system-wide behavior every cycle. Accessible content is flushed and re-populated on every cycle, making attention, both exogenous and endogenous, a critical enabler for encoding and active maintenance of mental representations (i.e., interlingua elements) across contiguous cycles.

As shown in Figure 1, on each cycle, a privileged item is selected from accessible content and broadcast system-wide to all components. The selected item serves as the *focus of attention* for that cycle, and once broadcast, all focus-responsive components take the focus and the current set of accessible content and compute their results, which are then added to the subsequent set of accessible content. The focus

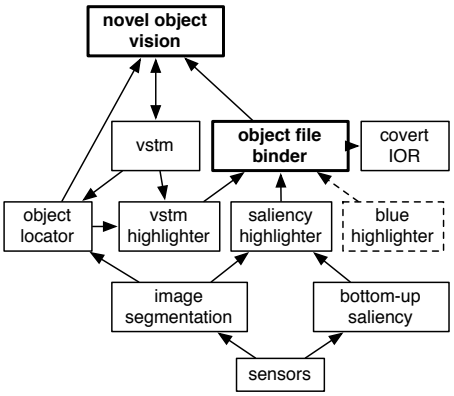


Figure 1: Informational exchange between ARCADIA components during the basic MOT task. Bolding of text indicates that the respective component is responsive to the focus of attention, and therefore attention-bound. Components having dashed borders are task-specific.

of attention is chosen by ARCADIA's focus selector, which is loaded with an *attentional strategy* for whatever task is currently being performed. Attentional strategies in ARCADIA are control knowledge, and establish selectional preferences over items in accessible content for what to focus on during each cycle. A detailed explication of ARCADIA's interlingua, processing loop, and focus selection was given by Bridewell & Bello (2015).

Modeling Object Construction and Tracking

Visual processing in ARCADIA is divisible into pre-attentive and attentive computations that can occur simultaneously. The set of computations underlying object construction can be seen in Figure 2, which may be a helpful roadmap for navigating the subsequent description. Pre-attentively, ARCADIA employs components that compute salience maps via methods described by Itti and colleagues (1998) and proto-object representations via image segmentation. This latter component produces interlingua elements that encode basic color histogram and region information wherever closed contours are found in the image. This serves as a rough and ready approximation to a high-speed, high-capacity iconic memory. The image segmenter also provides proto-object regions to the object locator component, which will be described in detail shortly and also works pre-attentively.

The next set of components in ARCADIA's visual system is responsible for pre-attentively producing requests for orientation. As shown in Figure 2, ARCADIA's saliency highlighter looks at accessible content for interlingua elements having saliency maps and others containing proto-objects. Each region containing a proto-object is co-registered back onto the saliency map and checked for salience value. The saliency highlighter outputs the N regions (with N <= 4; see Xu & Chun 2006) containing salient proto-objects, which become candidates for orientation. The vSTM highlighter produces top-down requests for orientation on objects that have been encoded into visual short-term memory.

The next layer of the visual system is responsible for object-construction, maintenance and tracking. Once orientation requests have been generated, whichever attentional strategy is currently loaded into ARCADIA's focus selector will be used to select from available orientation requests based on their relative prioritization in the strategy. Because this stage of vision is attentive, the selected region is broadcast system-wide, and any components that detect property information, such as shape or color, and can produce inferences or judgments about the content of the region now do so. The resultant judgments are passed by the property detectors up to accessible content as interlingua elements. A binding component then binds all the features that are detected in the region into an object file, which corresponds to a fully formed visual-object and a list of its properties. If focused on, new objects are tested against object representations in vSTM by the novel object vision component to determine whether they are actually new objects or should be treated as an update to a sufficiently similar object encoded in vSTM. ARCADIA assumes a fixed-slot four-element capacity for vSTM with a queue structure, so that when at capacity, new objects encoded in vSTM displace the oldest object in memory. On each cycle, vSTM pushes a list of its elements into accessible content, which are used both by the vSTM highlighter component and the object locator component.

Pre-Attentive Location Update: Object Locator

So far, we have described the normal course of processing for object construction, encoding and attention-dependent vSTM update. But what vision components are unique to MOT performance? Surprisingly, on our account there are none. Instead, the tracking mechanism most often implicated in MOT performance is motivated by other concerns.

Given two temporally contiguous visual snapshots, the human visual system is faced with the problem of re-identifying objects residing within the first frame with objects residing in the second frame, sometimes called *the correspondence problem*. Moreover, this problem is made difficult if the objects in question are in motion and change locations between frames, as would be the case for both the objects in tracking tasks. Dawson (1991) identified and computationally explored a potential solution as a corollary to his work on characterizing the mechanisms underwriting the phenomena of apparent motion.

In summary, Dawson finds that the correspondence problem is solved in the human visual system through the mutual satisfaction of three soft constraints, only two of which we will concern ourselves with in this work. The first of these constraints ensures a one-to-one mapping of each object in the first frame to a corresponding object in the second frame. The second constraint embodies a proximity-based principle such that each object in the first frame is assigned to the nearest object in the second frame in terms of Euclidean distance. The solution to the correspondence problem (1) is insensitive to object features other than spatial location and relative velocity and (2) operates on timescales well beneath those required to solve the problem attentively or deliberatively.

In ARCADIA, the object locator component serves as a bank of visual indices (Pylyshyn & Storm 1988, Alvarez & Franconeri 2007) that reference object locations. Object locator mirrors the internal structure of vSTM and stores a set of locations associated with each object encoded in it. There are two critical differences between vSTM and object locator. The first is that while vSTM stores conceptual representations of objects as collections of properties, the representations in object locator only store object location. Secondly, while vSTM requires attention to update location information for the objects it contains, object locator performs updates pre-attentively on each cycle, using Dawson's nearest neighbor proximity-based heuristic.

Object locator uses proto-object information deposited in accessible content by the image segmenter along with the contents of vSTM in updating its location information. To this end, object locator computes the N nearest proto-object neighbors in terms of Euclidean distance for each element of vSTM, and updates its location values (which correspond to their respective counterparts in vSTM) with new location information from their nearest proto-object neighbor. This basic computation immediately entails that there will be interactions between speed, crowding, and performance in MOT, since fast-moving objects will generate more instances of crowding over time and generate more opportunities for object locator to incorrectly identify vSTM elements with the wrong proto-object.

MOT: Task Specific Components and Strategy

One of the defining features of our account of MOT is just how little must be added to our model of object construction and tracking in order to simulate the MOT task. We only add a single task-specific component: a "blue highlighter." Our MOT simulation highlights the initial target set in blue before changing them back to their initial color. Blue highlighter detects proto-objects from the image segmenter with blue color histograms, and produces fixation requests on those regions of the image. We assume that the experimental instructions given to ARCADIA *qua* human subject indicate that targets will initially flash blue so that ARCADIA's attentional strategy reflects prioritization of blue-directed orientation requests over any others.

While not mentioned in the last section, ARCADIA's attentional strategy for object construction and tracking is given below:

1. If a new object file is available, make it the focus of attention so that it can be compared to and/or encoded in vSTM.

2. Otherwise, if one of the highlighters requests moving covert visual attention and there are no inhibitors

preventing the movement,[2] attend to the specified proto-object.

3. If neither of these options exists, attend to an arbitrarily selected interlingua element.

For MOT, the only change we make to the strategy above is to induce a preference ordering over highlighters such that requests from blue highlighter are prioritized over requests from vSTM highlighter, which are prioritized over requests from saliency highlighter. This has the effect of ARCADIA encoding blue targets when they first flash and serially revisiting each vSTM-encoded object over the course of tracking. Since object locator updates a mirror image of whatever is encoded in vSTM, it is unaffected by the attentional strategy above.

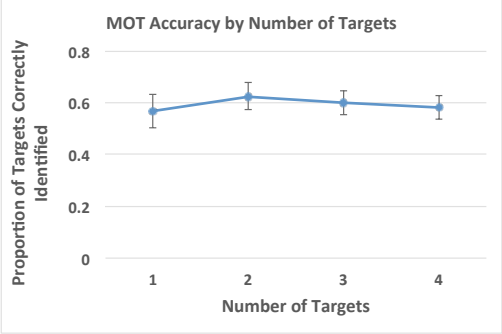


Figure 3: Results when varying the number of targets.

Computational Simulation

To test the set of predictions we have made thus far, we ran the model of ARCADIA shown in Figure 2 in a standard MOT task with sixteen total objects for five seconds per trial. We varied speed, spacing, and the number of initial targets to be tracked. Spacing between objects (circles of diameter D) varied at three levels: 0, 0.5D, and D. Effectively, these values amount to either allowing collision or requiring one or two full diameters of space between objects as they moved. Speed was normed by determining the value at which ARCADIA consistently failed to correctly identify any targets, even at D spacing. We divided this value by four and determined slow, regular, medium, and fast speed levels for the targets. Finally either 1, 2, 3, or 4 targets could be tracked, leading to a 4 x 4 x 3 configuration. We ran each configuration five times for a total of 240 system runs. Each run began with randomly selected targets in random locations with randomly selected initial trajectories. After each run, we computed the proportion of targets that ARCADIA successfully tracked.

[2] We do not discuss inhibition here since it plays no role in MOT, but other ARCADIA models utilize both task-related inhibition and covert inhibition of return (Bridewell & Bello, 2016)

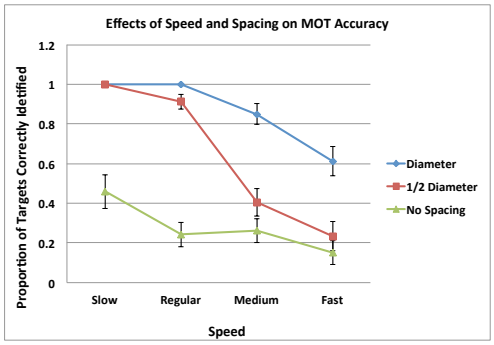


Figure 4: The interaction between speed and spacing in ARCADIA's performance on the MOT task.

Simulation Results

A 4 (number of targets) x 4 (speed) x 3 (spacing) repeated measures ANOVA was conducted. The proportion of targets correctly identified served as the dependent variable. There was no main effect for the number of targets tracked, $F(3, 12) = 0.68$, $p = 0.58$, indicating that up to 4 targets can be tracked robustly. Means and standard errors for number of targets tracked can be seen in Figure 3. There was a main effect of speed, $F(3, 12) = 44.52$, $p < 0.001$, with the proportion of targets correctly identified decreasing as speed increased ($M_{slow} = 0.82$, $M_{regular} = 0.72$, $M_{medium} = 0.51$, $M_{fast} = 0.33$). There was a main effect of spacing, $F(2, 8) = 158.81$, $p < 0.001$, with the proportion of correctly identified targets increasing as spacing decreased ($M_{0\ spacing} = 0.28$, $M_{.5D\ spacing} = 0.64$, $M_{D\ spacing} = 0.87$). These two significant main effects were qualified by a significant speed by spacing interaction, $F(6, 24) = 8.70$, $p < 0.001$. This interaction can be seen in Figure 4.

General Discussion

In general, the results of the simulation study are consistent with the vast majority of literature on limitations in MOT performance. The lack of a main effect of target reflects robust tracking of one to four objects via visual indices. The main effects of speed and spacing found here are in accordance with previous findings and correspond to two elements of Scimeca and Franconeri's *core four* signature performance limits on MOT. Crucially, we also found a significant interaction between speed and spacing, which is both predicted by any nearest-neighbor type model. While not conclusive with respect to the prediction made by Franconeri and colleagues (2010) that takes spacing to be the only theoretical boundary on MOT performance, our results do provide evidence that spacing plays a considerable role in offsetting the performance-reducing effects of high-speed object movement. To more fully pursue Franconeri's hypothesis, we would need to decrease target size and increase the number of possible spacing conditions, but we would expect to find performance dropping off much more slowly as a function of speed, given extra space.

We do not report results here related to hemifield and quadrant effects nor do we report capacity effects. The former, while interesting, are contentious (see Hudson et al. 2012 for details). Capacity effects for tracking loads greater than four are not reported here since ARCADIA's vSTM is only four objects deep. Explanations for how subjects manage to track more than four objects are open to many different interpretations that invoke strategies and mechanisms beyond the simple pre-attentive updating mechanism described here.

So-called "flexible resource theories" have been proposed to explain human tracking of more than four targets (Alvarez & Franconeri 2007). One of the nagging problems about resource theories is that it remains unclear what a "resource" could be (Franconeri et al. 2010). Similarly, many resource-theory explanations fail to invoke the distinction between the attentive and pre-attentive components of tracking, even though the weight of evidence points to the existence of a pre-attentive basis for MOT performance. Finally, others have argued that tracking perceptual groups of objects can give the appearance of tracking more than four targets because one or more of the four are actually sets of targets rather than individuals (Yantis 1992). This is just one example of a potential strategy that could be employed to explain tracking performance past four objects.

These factors and open possibilities have persuaded us to be methodologically conservative in the work we report here. We assume that the storehouse for visual indices in our approach mirrors the structure of vSTM, which we are conservatively assuming is four objects in capacity. One of the features of our theory is that it is insensitive to the internal structure of vSTM. If, for example, evidence persuades us to implement a resource-based account of vSTM that allows for a larger number of objects to be represented at coarser resolution, our pre-attentive update mechanism will mirror this structure and behave accordingly.

Comparison to Other Computational Models

The approach we have taken with implementing a simple nearest-neighbor pre-attentive update within a larger vision framework in ARCADIA has several explanatory advantages. The most important feature of our proposal is to link visual index updating to a known psychological process involved in other parts of visual cognition. Alternative computational models of MOT use Kalman filters and thus share an important similarity: location updating is a function of prediction (Vul et al. 2009, Zhong et al. 2014, Srivastava & Vul. 2015). However, a number of studies have demonstrated that human subjects seem to not rely on extrapolation of object trajectories during MOT, which may appear to rule out Kalman filters as lacking face validity for modeling human tracking performance (Keane & Pylyshyn 2006, Franconeri et al. 2012). A small number of studies demonstrate trajectory extrapolation in MOT under highly circumscribed conditions (Fencsik 2007, Howe & Holcombe 2012). In these latter two studies, extrapolation was only observed for tasks having a tracking load of two or less. Somewhat more disturbingly, Howard and colleagues (2011) find that some of the aforementioned results indicating trajectory extrapolation involved verbal instructions to subjects that may have unintentionally primed subjects to use explicitly extrapolative strategies.

The sharp limitations on tracking load in studies that implicate extrapolation in MOT are suggestive of different mechanisms at play, or perhaps some difference at the tracking-strategy level having to do with allocation of attentional resources. Even if extrapolation is happening for loads of two or less, it very well may be that this process is purely attention-bound and not reflective of the pre-attentive location updating mechanism under discussion in this paper. In any case, ambiguity of this sort compels us to take care as modelers to distinguish between pre-attentive and attention-bound processes, a distinction that is central in our own work.

Future Work

Perhaps the lowest hanging fruit involves modeling results that show subjects are capable of MOT with occlusions using a proximity heuristic such as the one implemented in ARCADIA's object locator (Franconeri et al. 2012). Because object property information other than location is updated attentively in ARCADIA's vSTM, identifying information about targets would be lost over the course of tracking, which is primarily served by pre-attentive mechanisms (Pylyshyn 2004). Pylyshyn (2006) has suggested that distractor inhibition plays a role in explaining why target information is recalled poorly. ARCADIA's attentional strategy for MOT prefers fixating on objects encoded in vSTM over anything driven by salience or other bottom-up processes. In this way, inhibition is built in as a function of being task focused.

Because ARCADIA is driven by attentional strategies it can be used to capture a variety of plausible strategy-driven features of MOT. Recent results are suggestive of better tracking performance for targets in crowded parts of the MOT display (Srinistava & Vul 2015) due to strategic deployment of attention to minimize uncertainty. Attentional strategies along with other components to detect relations could also be used to encode targets as the vertices of a polygon to be tracked (Yantis 1992). ARCADIA would need to be outfitted with a theory of overtly deployed visual attention and an accompanying model of eye movements to begin attacking any of the above in earnest. Work on these additions to ARCADIA has begun, allowing for a much richer and fuller exploration of the range of human object tracking capacity.

Acknowledgments

The authors would like to acknowledge generous support from the Office of Naval Research under grants N0001414WX20179 and N0001415WX01339. The views expressed in this paper are solely the authors and should not

be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Alvarez, G., Horowitz, T., Arsenio, H., DiMase, J., & Wolfe, J. (2005). Do multielement visual tracking and visual search draw continuously on the same visual attention resources? *Journal of Experimental Psychology: Human Perception and Performance, 31*, 643–667.

Alvarez, G., & Franconeri, S. (2007). How many objects can you attentively track?: Evidence for a resource-limited tracking mechanism. *Journal of Vision*, *7*, 1–10.

Bridewell, W., & Bello, P. (2015). Incremental object perception in an attention-driven cognitive architecture. In *Proceedings of the thirty-seventh annual conference of the cognitive science society* (pp. 279-284). Austin, TX: Cognitive Science Society.

Bridewell, W., & Bello, P. (2016). Inattentional blindness in a coupled perceptual-cognitive system. In *Proceedings of the thirty-eighth annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Dawson, M. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review, 98*, 561–603.

Fencsik D., Klieger S., & Horowitz T. (2007). The role of location and motion information in the tracking and recovery of moving objects. *Perception & Psycho-physics , 69,* 567–577.

Franconeri S., Jonathan S., & Scimeca J. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science, 21*, 920–925.

Franconeri S., Pylyshyn Z., & Scholl B. (2012). A simple proximity heuristic allows tracking of multiple objects through occlusion. *Attention, Perception and Psychophysics*, *74*, 691–702.

Howard, C., Masom, D., & Holcombe, A. (2011). Position representations lag behind targets in multiple object tracking. *Vision Research*, *51*, 1907-1919.

Howe P., & Holcombe A. (2012). Motion information is sometimes used as an aid to the visual tracking of objects. *Journal of Vision*, *12*, 1–10.

Hudson C., Howe P., & Little D. (2012). Hemifield effects in multiple identity tracking. *PLoS One 7*: e43796.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Keane, B., & Pylyshyn, Z. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive Psychology*, *52*, 346–368.

Pylyshyn, Z., & Storm, R. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial index model. *Cognition*, *32*, 65–97.

Pylyshyn, Z. (2004). Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition*, *11*, 801–822.

Pylyshyn, Z. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition*, *14*, 175–198.

Scimeca, J., & Franconeri, S. (2015). Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*. Advance online publication.

Srivastava, N., & Vul, E. (2015). Attention dynamics in multiple object tracking. *Proceedings of the thirty-seventh annual conference of the cognitive science society* (pp. 2266-2271). Austin, TX: Cognitive Science Society.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems*, *22*, (pp. 1955-1963).

Xu, Y., & Chun, M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Science, 13*, 167–174.

Yantis, S. (1992). Multielement visual tracking: attention and perceptual organization. *Cognitive Psychology, 24,* 295–340.

Zhong, S., Ma, Z., Wilson, C., Liu, Y., & Flombaum, J. (2014). Why do people appear not to extrapolate trajectories during multiple object tracking? A computational investigation. *Journal of Vision*, *14*, 1–30.

How do Distributions of Item Sizes Affect the Precision and Bias in Representing Summary Statistics?

Midori Tokita (tokita@mejiro.ac.jp)
Faculty of Health Sciences, Mejiro University,
Ukiya, Iwatsukiku, Saitama 339-8501, Japan
Akira Ishiguchi (ishiguchi.akira@Ocha.ac.jp)
Faculty of Core Research, Ochanomizu University,
Otsuka, Tokyo 112-8610 Japan

Abstract

Many studies have shown that observers can accurately perceive and evaluate the statistical summary of presented objects' attribute values, such as the average, without attending to each object. However, it remains controversial how the visual system integrates the attribute values (e.g., information on size) of multiple items and computes the average value. In this study, we tested how distributions of item sizes affect the precision and bias in judging average values. We predicted that if observers utilize all of the available size information equally, the distribution would have no effect, and vice versa. Our results showed that, with novice observers, judgement precision differed among size distributions and that the observers overestimated the size of the average value compared to the actual size under all conditions. These results imply that observations of some items in a set could be weighted more easily than others, with the possibility that this process is easier for larger items than smaller ones. However, this was not the case for experienced observers, who showed no effects of distribution type on average assessment performance. Our findings imply that the process of representing the average value may not be explained by a single definitive mechanism and, is rather mediated by a mixture of multiple cognitive processes.

Keywords: average size; statistical summary representation; size distribution

Introduction

It has been shown that observers are able to quickly and accurately extract average values over a range of visual properties, including size (Chong & Treisman, 2005 ; Oriet & Brand, 2013), brightness (Bauer, 2009), orientation (Dakin and Watt, 1997; Parkes, Liend, Angelucci, Solomon & Morgan, 2001), emotional expression (Haberman & Whitney, 2009, 2011). This ability is not limited to static and simultaneous events; it has been observed in sequentially presented events (Albrecht, Scholl, & Chun, 2012; Corbett & Oriet, 2011; Hubert-Wallander & Boynton, 2015） and dynamic objects, such as expanding and contracting disks (Albrecht & Scholl, 2010). Moreover, the ability to represent statistical properties is not limited to visual cues but is also observed in perceptions from auditory input, such as extracting frequency information from sequences of sounds (Piazza, Sweeny, Wessel, Silver, & Whitney, 2013) and temporal details of sounds (McDermatt, Schemisch, & Simoncelli, 2013).

These representations of statistical summary representations (SSRs), have been proposed to assist our judgment and behavior and more efficiently than attending to each objects and/or events individually (e.g., Alvarez, 2011; Ariely, 2001, 2008; Chong & Treisman, 2003; Robitaille & Harris, 2011). Although there is a general understanding that human observers can accurately represent sets of features, the mechanism by which people extract summary statistics is yet to be fully understood.

One of the debates over the mechanism of SSRs is whether the average value is computed using the entire information on display or using a subset of items in a set; ideas on this have been classified into three types based on findings from and discussions in previous studies.

The first idea is based on the claim that SSRs are computed without computing individual items. Many studies have provided the evidence that people can estimate the average size of a set of items without relying on focused attention on individual items when attention is distributed across a set of similar items (e.g., Attarha, Moore, & Vecera, 2014; Chong & Treisman, 2003, 2005; Oriet & Brand, 2013; Oriet & Hozempa, 2016; Tokita, Ueda, & Ishiguchi, 2016).

In the second idea, the assertion is that all items in a set are processed but not all items contribute equally to the mean. This idea suggests that if some measures are very reliable and others are not, observers may give the more reliable measures more weight when combining them (Alvarez, 2011). For example, Haberman and Whitney (2010) tested how the deviant emotional expression was utilized in averaging emotion shown on multiple faces and suggested that people implicitly and unintentionally discount the emotional outliers, thereby computing a summary representation that involves the majority of the information present. Hubert-Wallander and Boynton (2015) tested how SSRs were computed when stimuli were sequentially presented across time and found that they do not incorporate all items equally.

The basis of the third idea is that average size is computed using a limited sampling strategy which does not necessitate an ensemble representation computed in parallel across all items on display (Myczek & Simons, 2008; Fockert & Marchant, 2008; Marchant, Simons, & Fockert, 2013). For example, Fockert and Marchant (2008) claimed that observers do not always accurately average together the

entire set and that the average is either biased by features of the attended item or based on the practical strategy of extracting the mean of a smaller subset. In line with this argument, Marchant, Simons, and Fockert (2013) tested how the regularity of the item sizes affects the performance of average size perception and demonstrated that judgments of average size become less accurate with increases in the set size and heterogeneity of the item sizes.

In this study, we further explored the mechanism of estimating the average size of items in a set by manipulating the frequency distribution of items. This is a useful approach for determining whether observers utilize information on all items equally, they weight the information of some items and less of others, or they use a limited number of items in a set (Chong & Treisman, 2003; Duffy, Huttenlocher, Hedges & Crawford, 2010; Marchant et al., 2013). For example, Duffy et al. (2010) used asymmetric (skewed) distributions in which there was more than one possible central value and demonstrated that observers adjusted estimates toward the category's running mean.

We tested how frequency distributions of item sizes affect the precision and bias in representing average size. Four types of the distributions were used: uniform, negatively skewed (i.e., Asym1), positively skewed (i.e., Asym2), and pseudo-normal distributions. Figure 1 shows the sizes and numbers of items in the stimulus set for each distribution. We measured the Weber fraction of the discrimination task to assess precision and also measured the point of subjective equality (PSE) to test the accuracy of the estimation.

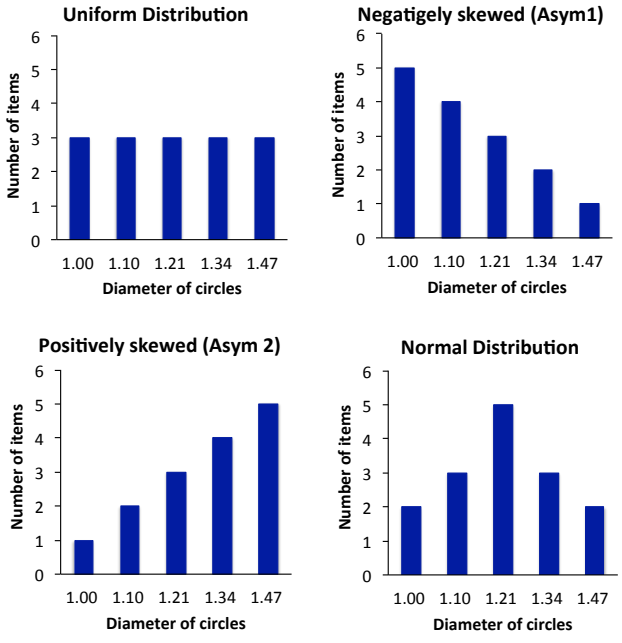


Figure 1: Four types of distributions were introduced: uniform, negatively skewed (i.e., Asym1), positively skewed (i.e., Asym2), and pseudo-normal distributions.

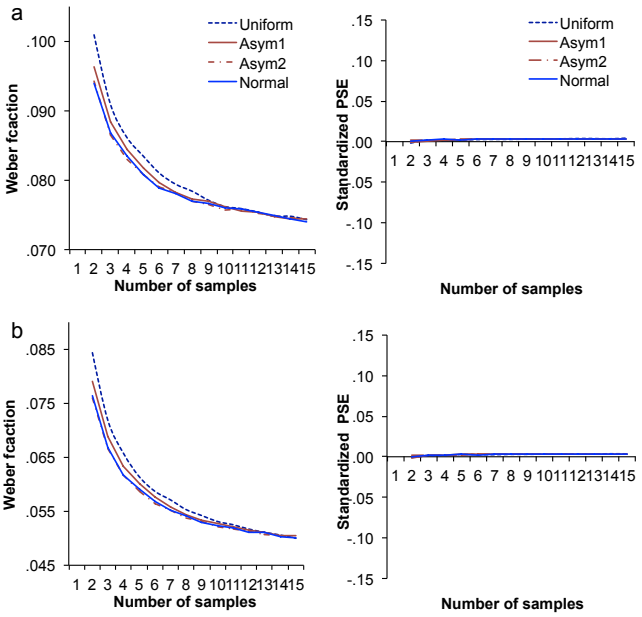


Figure 2: Simulated performances (Weber fraction and standardize PSE) for each distribution expressed as a function of the number of sampled items under different noise conditions: a. High internal noise, b. Low internal noise

To predict the performance of size average estimation, we used a computer simulation for calculating Weber fractions and the PSEs for each distribution expressed as a function of the number of sampled items, as shown in Figure 2. Two levels of noise were used in the simulation, with the values of the noise in each level being obtained from previous research (Tokita et al., 2016). As shown in the left-hand-side figures, when the average value was computed in parallel across all items in the display, performance was unaffected by the distribution of item size. When a limited number of items in the set were used to compute the average value, the performance was influenced by item distribution. Under Asym2 and normal distributions, precision was found to be better than that under uniform and Asym1 distributions. Precision under Asym1 was higher than under uniform. As shown in the right hand side of Figure 2, the PSE results show that observer accuracy was unaffected by distribution.

In addition, we conducted experiments with to two types of observer: novice and experienced. Since some studies have suggested that there are considerable individual differences in simultaneous visual tasks (Tokita & Ishiguchi, 2010; Herbert & Whitney, 2015), we considered it important to test how the effect of distribution differs between novice and experienced observers. Note that "the observers" refer to novice observers in this paper.

Experiment

We tested how the distributions of item sizes affect performance and bias in representing the average sizes. To examine the performance, we obtained Weber fractions that would indicate the precision in the participants' estimation of average sizes. To examine bias, we obtained PSEs, which indicate the constant error in estimation. In deriving the Weber fractions and PSEs, we used the method of constant stimuli, in which the observers in each trial decided which size of stimuli—the average size of a displayed item set (i.e., standard stimuli) or a single item size (i.e., comparison stimuli)—had larger.

Method

Participant As for novice observers, fourteen undergraduate volunteers from Mejiro University participated in exchange of course credit. All observers were naïve as to the purpose of the study. Three experienced observers including one of the authors, who have participated in many psychophysical experiments such as average estimation experiments were added. Two of the observers did not know the purpose of the study. All had normal or corrected-to-normal vision.

Apparatus Stimuli were displayed on the iMac desktop computer monitor (21-inch) controlled by a Macintosh computer (Mac OS X). Stimuli were generated using the Psychophysics Toolbox Version 3 (Brainard, 1997; Pelli, 1997) for MATLAB (Version 8.4, Mathworks, MA). Participants viewed the screen with both eyes and seated approximately 65 cm from the screen.

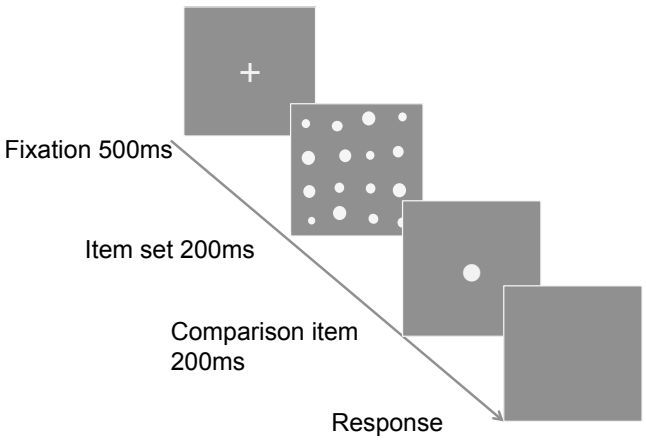


Figure 3: A schematic view of the stimulus presentation. Each trial started with a fixation cross for 500ms. The items in a set were presented first for 200 ms. The comparison item was presented for 200ms after blank for 500 ms, and then a blank screen until response.

Design Distributions of item sizes were manipulated. There are four distribution types; uniform distribution, negatively skewed distribution (Asym1), positively skewed distribution (Asym2), and pseudo-normal distribution. A set of items (i.e., standard stimuli) was presented in the first interval and a comparison item was presented in the second interval.

Stimuli The standard stimuli consisted of fifteen of filled light gray disks of various seizes, which were presented on dark gray background. The disk sizes were equally spaced on a log scale separated by a factor of 1.25. The comparison stimuli was a single disk with a given levels. There were five comparison levels, -0.14, -0.07, 0, 0.07, and 0.14 diameter differences on the power function scale. Three of participants needed wider range of stepwise levels (i.e., -0.16, -0.08, 0, 0.08, and 0.16) due to the low accuracy.

The items were arranged on the array. The array was divided into 4 × 4 matrix. Each item was displayed at the center of each cell with a position jitter.

In each trial, all of the disks shown were randomly scaled by a small multiplicative factor to discourage the participants from basing heir judgments on previously seen items. Three multiplicative factors (0.9, 1, 1.1) were used and the same factor scaled all items in any one trial.

Procedure A schematic view of the stimulus presentation is shown in Figure 3. Observers completed one 65-min session that consisted of a practice block of 24 trials, followed by five experiment blocks of 100 trials each (4 distribution types × 5 comparison level × 5 repetitions). There are 500 trials in total. The distribution types and the comparison level and the order of trials were all randomly mixed.

Each trial started with a fixation cross for 500ms. The items in a set were presented first for 200 ms. The comparison item was presented for 200ms after blank for 500 ms, and then a blank screen until response. The next trial automatically began 500ms after the response.

Observers' task was to decide whether the comparison item was larger or smaller than the average size of item in a set. A two-alternative (larger or smaller) forced choice procedure was used. When they thought that the comparison item is smaller than the average size of items in a set, they pressed '1', otherwise, they pressed '3'. No feedback about the correctness of responses was provided.

Analysis The PSEs and Weber fractions were measured using the method of constant stimuli. First, the relative sizes for the comparison item were plotted on the x-axis, and the proportion of greater responses for each comparison stimulus was plotted on the y-axis, and fits were done for individual data. The plotted data points constructed the psychometric function approximated by a cumulative Gaussian function for individual data.

This discrimination threshold was defined as the smallest amount of the stimuli number change, for which a correct response rate of 75% was achieved. The PSEs were obtained as the values of the locations on the psychometric

function at which the standard and comparative choice probabilities were equal to 50%.

Results

The fits of the data points to the psychometric functions were generally good, and the Pearson product–moment correlation coefficient exceeded .9 in most cases. The data of three novice observers were not satisfactory fitted to the psychometric functions, thus, we excluded their data form further analysis.

Figure 4a shows the results of Weber fractions in each distribution types. The precision of representing average size appeared to be partly affected by the size distribution. Figure 4b shows the standardized PSEs in each distribution condition. The average size of the set of items seems to be overestimated as compared with the comparison item in all distribution conditions with the novice observers.

Novice observers To test whether and how the size distributions affect the precision of representing average values, one way (4 distribution conditions) repeated measures analysis of variance was conducted on the individual Weber fraction. This yielded significant main effect of distribution, $F(3,11)= 2.94$, $p<.05$. Bonfferoni Post-hoc analysis revealed that the judgment in Asym2 condition (positively skewed distribution) gave higher precision than uniform, $p< .05$, and Asym1 (negatively skewed distribution) $p< .05$ conditions.

In a similar way, to test how the distribution affected the accuracy of representing average sizes, one way (4 distribution conditions) repeated measures analysis of variance was conducted on the individual PSE. This yields no main effect of distribution, $F(3,11)= .554$, $p>.1$. This suggests that the accuracy of representing average size was not affected by the size distribution. As average of PSE seems larger than the 0, we conducted a one-sample t test to compare the mean standardized PSEs of each condition with a PSE of 0. The analysis revealed that the mean of the PSE was significantly larger than 0 at the standard stimuli in the uniform, $t(10) = 4.99$, $p < .01$, Asym1, $t(10) = 2.63$, $p < .05$, Asym2, $t(10) = 4.40$, $p< .01$, and normal distribution, $t(10) = 3.30$, $p < .01$. This suggests that average size were overestimated as compared with the actual size in all distribution condition.

Experienced observers To test whether and how the size distributions affect the precision of representing average values, one way (4 distribution conditions) repeated measures analysis of variance was conducted on the individual Weber fraction of the experienced observers. This analysis revealed no significant effect of distribution, $F(3, 2)= .523$, $p >.1$.

To test how the distribution affected the accuracy of representing average sizes, one way (4 distribution conditions) repeated measures analysis of variance was conducted on the individual PSE. This yields no main effect of distribution, $F(3, 2)= 1.342$, $p >.1$. This suggests that the accuracy of representing average size was not affected by the size distribution. This suggests that in experienced observers there is no sign of bias average size were overestimated as compared with the actual size in all distribution condition.

Discussion

We tested how the distribution of item size would affect the precision and bias in the judgement of average size of items in a set by observers. Four types of distributions, namely uniform, negatively skewed, positively skewed, and pseudo-normal, were examined. We predicted that, if all items in a set were processed equally, the performance of the observers in judging the average would be unaffected by item distribution conditions. Our results demonstrated three significant findings. First, precision was significantly higher under Asym2 (i.e., positively skewed distribution) compared to the others, among which there were no significant differences.

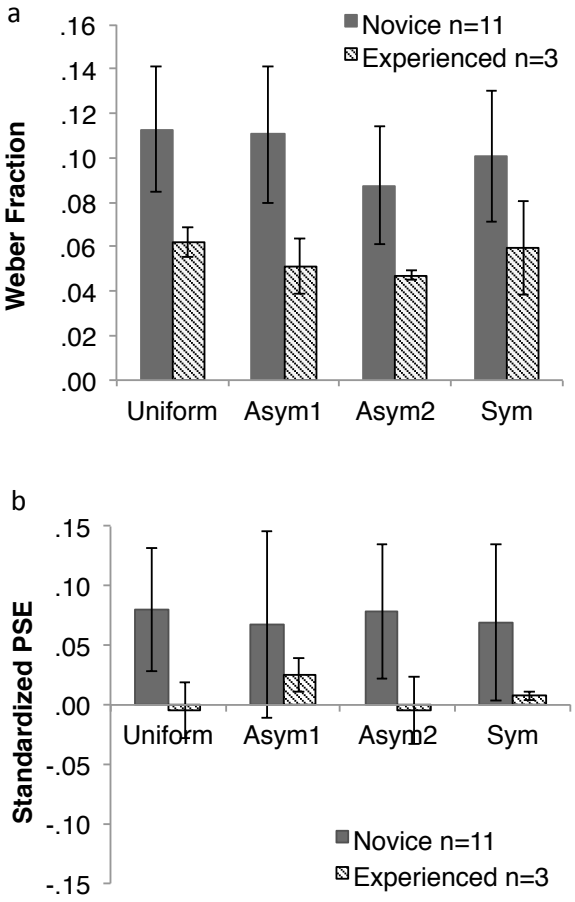


Figure 4: a Mean Weber fraction for each distribution condition. Error bars represent standard deviations. b. Mean PSEs for each distribution condition. Error bars represent standard deviation.

Second, the estimated average values of the items in a set were overestimated relative to the actual values under all item distributions. Third, the accuracy and precision of the experienced observers was unaffected by distribution type when carrying out the averaging task.

The first set of results implies that observers may not use all of the available information equally when assessing the average value of the items in a set, which could be because of two alternative processes. One possibility is that a limited number of elements could be being used to calculate average values. Another possibility is that all of the information on size in a display may not be weighted equally when judging averages. Instead, some of the items could be being weighted more than others. The results partially support those of Marchant et al. (2013), who indicated that the difference in the component of the item sizes affects the performance of perceiving the average.

The finding that overestimation of the average size relative to the actual size under all distributions is intriguing. Bias in the judgment of the average has rarely been investigated, and no consistent pattern of bias has yet been observed. There are two possible causes for the bias. One is the observer's perceptual saliency of large stimuli; the observer may process the large items more attentively than the small ones, and thus, it is possible that observers utilize the information of larger items relatively more frequently in computing the average size. Another possibility is that observers may automatically give more weight to the larger items than to the smaller ones, irrespective of their saliency.

The fact that experienced observers were unaffected by distribution type suggests that they may process all the items in a set equally when elucidating their average value. These results support the idea that people can extract the average size of a set of items without relying on focused attention to individual items in the set. Taken together with the results for novice observers, it is implied that how we represent the average value may not be accounted for just from a single process but from a variety of processes that depend on individual cognitive characteristics. The implication is somewhat consistent with the findings of Haberman, Brady, and Alvarez (2015). They pointed out that the mechanism of representation of summary statistics may involve various levels of processes, and individual differences may reveal those levels. As our data shows, with a wide variety of item distributions, it is important to explore the basis of those differences.

In conclusion, our results demonstrate that average judgement precision differs with distribution and the size of the average value was overestimated by the observers under all conditions. However, the results for experienced observers showed that the performance of their judgement of the average was unaffected by distribution type. Our findings imply that the process of assessing the average value may not be explained by a single definitive mechanism but is rather mediated by a mixture of multiple cognitive processes.

Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research (C) (40571112) and Scientific Research(B) (15H03462) provided by the Japan Society for the Promotion of Science to the authors.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting summary statistics representations over time. *Psychological Science, 21,* 560-567.

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: computing summary statistics from multimodal stimuli. *Atten Percept Psychophys, 74*, 810-815. doi: 10.3758/s13414-012-0293-0

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhance visual cognition. *Trends in Cognitive Science, 15,* 122-131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychol Sci, 19*, 392-398.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157-162.

Attarha, M., Moore, C. M., & Vecera, S. P. (2014). Summary Statistics of Size: Fixed processing capacity for multiple ensembles but unlimited processing capacity for single ensample. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1440-1449.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. Vision Research, 43, 393-404.

Chong, S. C., & Treisman, A. (2005). Statistical processing: computing the average size in perceptual groups. *Vision Research, 45*, 891-900.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: perceptual averaging in the absence of individual item representation. *Acta Psychologia, 138*, 289-301.

Dakin, S. C., & Watt, R. J. (1997). The computation of orientation statistics from visual texture. *Vision Res. 37*, 3181–3192. doi: 10.1016/S0042-6989(97)00133-8

Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Percept Psychophys, 70*, 789-794.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 718-734.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Atten Percept Psychophys, 72*, 1825-1838. doi: 10.3758/APP.72.7.1825

Haberman, J., & Whitney, D. (2011). Efficient summary statistical representation when change localization fails.

Psychon Bull Rev, 18, 855-859. doi: 10.3758/s13423-011-0125-6.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in Ensemble perception Reveal Multiple, Independent Levels of Ensemble Representation, *Journal of Experimental Psychology: General*, *144*, 432-446.

Hubert-Wallander, B., Boynton, G. M. (2015). Not all summary statistics are equal: Evidence from extracting summaries across time, *Journal of Vision*, 15:5, doi: 10.1167/15.4.5

Marchant, Alexander P., Simons, Daniel J., & de Fockert, Jan W. (2013). Ensemble representations: Effects of set size and item heterogeneity on average size perception. *Acta Psychologica, 142*, 245-250.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nat Neurosci, 16*, 493-498. doi: 10.1038/nn.3347nn.3347 [pii]

Myczek, K., & Simons, D. J. (2008). Better than average: alternatives to summary statistics representations for rapid judgments of average size. *Percept Psychophys, 70*, 772-788.

Oriet, C., & Brand, J. (2013). Size averaging of irrelevant stimuli cannot be prevented. Vision Research, 79, 8-16.

Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision*, 16(3), 3. doi:10.1167/16.3.3

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science, 24*, 1389-1397.

Robitaille, N., & Harris, I. M. (2011). When more is less: extraction of summary statistics benefits from larger sets. *Journal of Vision*, 11(12). doi: 10.1167/11.12.18 S0001-6918(11)00142-9 [pii]

Simons, d. J., & Myczek, K. (2008). Average size perception and the allure of a new mechanism. *Perception & Psychophysics, 70*, 1335-1336.

Solomon, J. A., Morgan, M., & Chubb, C. (2011). Efficiencies for the statistics of size discrimination. *Journal of Vision*, 11, 13.

Tokita, M., & Ishiguchi, A. (2010). How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? *Atten Percept Psychophys, 72*, 1839-1853.doi:10.3758/app.72.7.1839

Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a global sampling process in extraction of summary statistics of item sizes in a set. *Frontiers in Psychology, 7*. doi: 10.3389/fpsyg.2016.00711

Who should I tell?
Young children correct and maintain others' beliefs about the self

Mika Asaba (masaba@stanford.edu) **Hyowon Gweon** (hyo@stanford.edu)
Department of Psychology, Stanford University, Stanford, CA 94305

Abstract

We care tremendously about what other people think of us. Motivated by two lines of prior work – children's inferential and communicative capacities and strategic reputation management – we examine how children infer what others think of them given others' observations of their performance, and how they influence these beliefs through disclosing their performance. In Experiment 1, 3-5 year-olds played a luck-based game; one confederates watched the child win and another confederate watched the child lose. We asked the child to disclose an additional, unobserved win to one of the two confederates. We find that younger children overwhelmingly choose the person who previously saw them win. However, as age increased, children were more likely to choose to disclose to someone who previously saw them lose. In Experiment 2, adults played a similar third person version and selectively chose the person who saw the main character previously lose.

Keywords: Theory of Mind; social cognition; cognitive development; communication; reputation management

Introduction

We are deeply curious about the minds of other people. From merely observing others' behaviors to probing them with direct questions, we frequently engage in learning about others' beliefs, and even attempt to change these beliefs by communicating with others. Among many kinds of unobservable contents of other people's minds, there is one suite of beliefs that we care extraordinarily about: others' beliefs about *us*.

Decades of research on Theory of Mind (ToM) – our capacity to understand and reason about others' unobservable mental state – has revealed much about both its early emergence (Baillargeon, Scott, & He, 2010) and its developmental trajectory (Wellman, Cross, & Watson, 2001). A host of prior work has examined various domains in which ToM plays a central role in our everyday social interactions, such as helping others (Buttelmann, Carpenter, & Tomasello, 2009), morally evaluating others (Cushman, Sheketoff, Wharton, & Carey, 2013), and teaching and communicating with others (Gweon, Pelton, Konopka, & Schulz, 2014; Strauss & Ziv, 2012). In this study, we highlight a yet another important role of ToM in our social lives: it allows us to reason about others' beliefs and evaluations of *us*, and help us change or maintain these beliefs by informing others about ourselves.

Imagine the following scenario: your friend Tom observes you win many amazing tennis matches, but later your friend Sam watches you lose several games in a row. Based on what they have seen, you might reasonably infer that your friends would hold different beliefs about your skill level; Tom would think of you as a competent tennis player, while Sam might think quite the opposite. Based on these inferences, you may even attempt to correct Sam's belief (e.g., by telling him that you recently won many matches) while maintaining Tom's belief (e.g., by omitting your losses). We propose that these seemingly simple intuitions emerge from a sophisticated ability to: (1) infer what others believe about us given some evidence about the self, (2) understand that such beliefs can be revised and updated given new information, and (3) decide how to selectively communicate information about the self to cultivate or maintain others' positive beliefs about us.

Previous developmental work has provided evidence for young children's competence in each of these components outside of the domain of beliefs about the self. First, children draw inferences about others' beliefs, epistemic states, and even moral dispositions based on their experiences and behaviors (Hamlin, 2013; Koenig & Harris, 2007). Second, children understand that others' beliefs can change with additional data (Song, Onishi, Baillargeon, & Fisher, 2008), and that these revisions are sensitive to the amount and strength of evidence (Schulz, 2012; Gweon, Shafto, & Schulz, 2014). Third, preschool-aged children also understand that they can actively cause such revisions to others' beliefs by selecting and communicating information for others (Ding, Wellman, Wang, Fu, & Lee, 2015; Gweon, Chu, & Schulz, 2014; Gweon, Shafto, & Schulz, 2014; Rhodes, Bonawitz, Shafto, Chen, & Caglar, 2015). For instance, children differentially select evidence to show others, depending on whether the goal is to teach or deceive (Rhodes et al., 2015) and the learner's prior beliefs (Gweon, Shafto, & Schulz, 2014).

However, in most of these studies, the contents of others' beliefs involve states of the external world with a ground truth (e.g., locations or names of objects, causal mechanisms). For instance, in the Sally-Anne task or other variations of false-belief tasks (Wimmer & Perner, 1983; Gopnik & Astington, 1988), children are questioned about the contents of others' beliefs that are either accurate or inaccurate with respect to the current states of the world. However, in people's beliefs about the self (or people's *beliefs about people* more generally), there is not always a clear ground truth; these beliefs are formed via various sources of information that reflect a person's unobservable, internal qualities (e.g., competence, niceness, knowledgeability, etc.). Therefore, the accuracy of such beliefs can only be assessed with respect to the representativeness of such observations. Therefore, it remains an open question whether children's competence in reasoning about others' factual beliefs about the world also naturally extend to reasoning about others' beliefs about the self.

In reasoning about others' beliefs about us, we not only care about the accuracy of others' beliefes but also about the *desirability* of the beliefs; that is, we care tremendously about whether others think we are competent, nice, and knowl-

edgeable, independent of its ground truth. Indeed, prior research on children's impression management suggests that even young children are motivated to create a positive reputation with others (Shaw, Li, & Olson, 2013; Lee, 2013), using appropriate communicative (e.g., lying) and behavioral (e.g., sharing, helping) means to do so. For instance, even 3-year-old children lie about peeking into a box when an experimenter told them not to do so (Chandler, Fritz, & Hala, 1989; Polak & Harris, 1999). Five-year-olds help more and cheat less in the presence of others (Engelmann, Herrmann, & Tomasello, 2012; Leimgruber, Shaw, Santos, & Olson, 2012), and even when they have been told that other children think they are good (Fu, Heyman, Qian, Guo, & Lee, 2015). By the end of the preschool years, children show both a motivation for others to think well of them as well as a sensitivity to particular kinds of information that might lead others to form positive impressions of them in prosocial or moral contexts.

While these studies have demonstrated children's desire to maintain a positive reputation with others, most of these studies looked at children's behaviors with others with whom they had no prior experience; that is, previous work has mainly shown children's attempts to create a good "first impression". However, many of our daily interactions involve people we have already interacted with before, and thus have already formed some beliefs or impressions of us. If we want these individuals to hold positive beliefs about ourselves, we must employ our Theory of Mind to manage these beliefs: we need to infer what others believe about us, and when such beliefs are discrepant from what we want others to believe, we must provide additional information to change these beliefs.

As adults, we intuitively understand that one can be selective about *to whom* we tell different pieces of information. Importantly, such selective and targeted communication of information is driven by our desire for others to know only certain (and often desirable) aspects of ourselves. We also recognize that the same piece of information (e.g., winning a game) can differentially impact what others think of us. In the tennis example, disclosing to Sam (who thinks you are not very skilled) that you won a game might greatly improve his beliefs about your skills, whereas telling Tom (who thinks you are good) might simply reaffirm what he thinks of you.

In our study, we asked whether young children prefer to provide new positive information about themselves to someone who has previously seen them succeed over someone who has seen them fail. In Experiment 1, we examine children's choices about whom to disclose information about their performance in a simple, luck-based game. In Experiment 2, we asked how adults perform in similar third-person tasks.

Experiment 1: 3-5 year-olds

In Experiment 1, children played four trials of a simple luck-based game. They lost the game once in front of one confederate and won once in front of another, and then were asked to disclose another final win to either one of the confederates. We targeted 3-5 year-olds, given prior work showing that children in this age range are strongly motivated to have others think positively of them (Shaw et al., 2013), while also showing remarkable developmental changes in their ability to reason about and track others' beliefs (Wellman et al., 2001).

In light of this work, we can consider a few possible patterns of data. First, children might be strongly biased to disclose an additional win to someone who previously saw them win (and thus holds a positive belief about them). Second, children might prefer to disclose to someone who has seen them lose (and holds a negative belief about them), recognizing that they can correct such negative beliefs. Finally, we might see a developmental change in children's choices; younger children might have a preference for someone who already holds a positive belief about them, whereas older children might understand the value of disclosing a win to someone who holds a less positive belief about their performance.

Methods

Subjects Seventy-three preschool-aged children (M_{Age}(SD): 4.56(.71), range: 3.23 - 5.98) were recruited from an on-campus preschool. An additional eleven children were tested and dropped from analysis due to failure to complete the study (N = 7) or experimenter error (N = 4).

Materials Children played a simple game with the experimenter using laminated cards (2"x 3"). The winning cards were red on the front with a large yellow star in the middle, while the losing cards were blue on the front with no star; all cards were yellow on the back. Two headshot photos (3"x 5") of each confederate wearing neutral colored clothing. Stickers were used as rewards. See Figure 1B for stimuli.

Procedure Participants were tested in a quiet room inside of the preschool. First, the experimenter introduced the game to the participant: in each trial, the experimenter would lay out several cards face down on the table, and the child could choose one of the cards to flip over. Then, the experimenter showed the two types of possible cards: a red card with a star on it meant that the child won (i.e., win a sticker) ; a blue card with no star meant the child lost (i.e., receive nothing). The experimenter asked the child what happens if each type of card was picked; all children were able to correctly report the outcome of each card. The experimenter showed the child the two pictures of the confederates, provided their names (e.g., Anne and Sally), and said, "Sometimes, my friends will come in and watch the game. Does that sound okay?" All children agreed for the confederates to come in during the game.

Children played four trials of the game; the cards were stacked such that the experimenter could control the outcome for each trial. If the child chose a winning card, the experimenter said, "Oh wow, you got a star! That means you get a sticker!" If the child chose a losing card, the experimenter said, "Oh no, that means you do not get anything." The child always won on the first trial. Before each of the second and third trials, the experimenter told the child, "Okay, I think my friend Anne (Sally) wants to watch now!" And then she

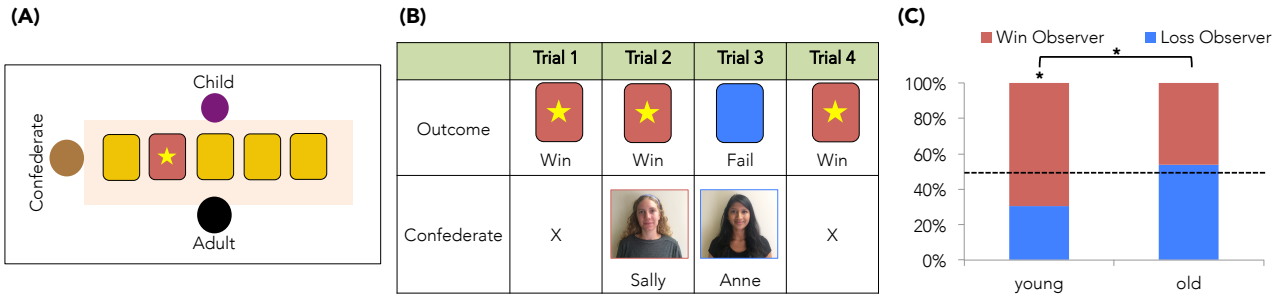


Figure 1: Experiment 1 game setup (A), design and order of trials (B), and results (C), $*p < .05$.

looked at the door (behind the child) and and said, "Hey Anne (Sally), is that you?" Anne (Sally) then entered the room, smiled politely, and sat in a chair between the child and experimenter (see Figure 1 for schematics of the room setup). The confederate maintained a neutral to mildly positive facial expression, made no comments while the game was playing, and did not respond to the outcome of the game. The child won in front of one confederate and lost in front of another (order and confederate counterbalanced). After the outcome was revealed, the confederate left the room.

On the final, fourth trial, no confederate was present and the child always won. Then, the experimenter brought out the pictures of the confederates and told the child that now he or she can tell Anne or Sally about the final win. The child indicated his or her response by saying a name or pointing to one of the photos; if the child did not respond or said both, the experimenter asked the child to only choose one confederate. The experimenter called the chosen confederate into the room. After the child told the confederate what happened, the confederate left the room and the child then received three stickers for the three winning trials.

Results and Discussion

Three- to five-year-olds were asked to disclose the final win to either someone who previously saw them win ("Win" observer) or someone who previously saw them lose ("Loss" observer). Our main interest was which confederate children chose, and whether children's choices change with age. We found no effects of gender, the order of the win and lose in the second and third trials, or the identity of the confederates.

Collapsing across all participants, we found that 31 of 73 participants (42.5%) chose to disclose the final win to the "Loss" observer (p = .242). Next, we used the median age (4.64) to split participants into a younger (N = 36; M_{Age}(SD): 3.96(.43)) and older group (N = 37; M_{Age}(SD): 5.14(.33)). We found that only a small proportion of the younger group, 11 of 36 (30.5%) chose to disclose the final win the "Loss" observer (p = .029, by binomial test). However, the older group showed a different pattern; although the number of children who chose the "Loss" observer was not significantly above chance (p = .743, by binomial test), they were significantly more likely to disclose to the "Loss" observer compared to the children in the younger group (20 of 37 or 54.1%, younger vs. older: p = .059, by fisher's exact test). Indeed, in a logistic regression model (1 = Chose "Loss", 0 = Chose "Win"), Age was a significant predictor in children's choice of confederates (β = .828, p = .025), with older children were more likely to disclose to the "Loss" observer.

In sum, we found that although younger preschool-aged children preferred to disclose the final win to the "Win" observer, children gradually become more likely to disclose to the "Loss" observer. Interestingly, the "Win" observer already had a positive belief about the child, and thus disclosure of an additional win would do little to change her beliefs. While older children were significantly more likely to disclose to the "Loss" observer, the proportion of children who did so was only slightly above chance. Moreover, this result shows young children's striking ability to distinguish between the two observers after only single trials.

Given the results from the logistic regression, one question to further explore is *when* we might preferentially select the "Loss" observer over the "Win" observer. In the next study, we asked whether even adults would choose the "Loss" observer. It is also interesting to consider how the nature of the game (e.g., luck-based vs. skill-based games) might differentially affect motivation to correct others' beliefs. Here, children played a simple, luck-based game but if the game's outcome was determined by effort or skill, there might be a greater concern for reputation and thus a stronger tendency to maintain or change others' beliefs in the desired direction. In Experiment 2, we first investigate adults' choices in an online third-person version of the task (2A), and directly compare participant's responses in a game of chance and a game that involves more skill (2B).

Experiment 2A: Adults (Luck-Based)

In Experiment 2A, we first asked about adults' preferences to correct others' negative impressions. In order to make the scenario plausible for an online study, we asked people to predict another agent's choice to disclose, rather than their own choices. Adults observed an agent (Bill) succeed or lose in a card game task as two confederates observed the agent at different times. We asked the participants to choose who Bill should tell about his final successful performance: the friend

who watched him win, or the friend who watched him lose.

Methods

Subjects Adults (N=100) were recruited from Amazon's Mechanical Turk (AMT), M_{Age}(SD): 34.2(8.9), range: 22 - 60. Fifteen additional participants were tested but excluded from analysis due to failing a planned attention-check question (see Procedure).

Materials Participants observed several cartoon pictures in the online game. The game consisted of a vertical board with five cards on it (see Figure 2); the dealer is depicted in green. A card with a star on it meant that Bill won; a card with an X on it meant that Bill lost. Only the characters backs were shown during the observation pictures.

Procedure

Participants first learned how the game worked and what each card meant; only those who correctly reported that the winning card has a star were included in analyses. We ran two versions of the study in which we varied the amount of evidence. In one version (N=50), Bill won once by himself, then won once in front of one friend ("Win" observer) and lost once in front of the other ("Loss" observer). Each win or loss trial was represented in its own image on a page. For the fourth and last trial, Bill won once all by himself. Then, participants were asked who Bill will tell about his final win: the friend who previously saw Bill win once or the friend who previously saw Bill lose once. Participants were asked to explain their choice afterwards.

In another version (N=50), Bill won or lost *three* trials in a row in front of each confederate, and participants were asked to choose to whom Bill should disclose his final three wins. For both versions, we counterbalanced the order of the observed outcomes (W L vs. L W; W W W L L L vs. L L L W W W) as well as which of the two friends observed during these trials.

Results and Discussion

We found no effect of the order of the observed outcomes (W first vs. L first), the order of the friend (Anne or Sally), or the number of trials (1 or 3 trials). Thus the data were collapsed for further analysis. Participants were asked to choose which one of the two friends Bill should tell about his final win: the friend who previously saw him win lose before versus the friend who previously saw him lose before. 61% (61 of 100) of participants chose the person who previously saw Bill lose (binomial test, $p = .03$). Thus adults were sensitive to the kind of outcome each friend observed, and predicted Bill to selectively disclose to the friend who previously saw him lose.

Experiment 2B: Adults (Skill-Based)

In Experiment 2B, we asked how manipulating the nature of the game (luck-based vs. skill-based) might affect adults' reasoning about selective disclosure. When the game involves skill rather than luck and is thus more informative about the player's underlying competence, more participants might predict that the agent would communicate his positive performance to someone who holds a negative belief about his competence than someone who holds a positive belief.

Methods

Subjects Participants (N=200) were recruited from AMT, M_{Age}(SD): 39.1(10.1), range: 24 - 59. Twelve participants were excluded from analysis due to failing the two planned attention-check questions (see Procedure).

Materials As in Experiment 2A, participants observed a cartoon scenario where Bill played a game as one of his friends (Anne or Sally) was present. The game featured a dart board divided into 6 alternating blue and red slices.

Procedure

The structure of the task was nearly identical to Experiment 2A, and the only difference was that Bill played darts, rather than a card game. This allowed us to easily manipulate the nature of the game (luck-based vs skill-based) using nearly identical stimuli. Participants were randomly assigned to the Skill and Random conditions (N = 100 in each condition). In the Random condition, Bill simply pressed a button to activate a dart-throwing machine that would place the dart at a random location on the board. In the Skill condition, Bill aimed at the board to throw the dart, but no information was provided about his competence. In all conditions, participants were told that Bill would "win" if the dart landed on one of the blue slices on the board, and "Loss" if it landed on red.

As in Experiment 2A, we ran two versions of this task. In one version, one of Bill's friends observed him win once and the other friend watched him lose once. Then, Bill won once in the last weekend when no one was watching. We asked participants who Bill will tell, the friend who observed him win before or the friend who observed him lose before. In another version, Bill won or lost *three* times (instead of once) in front of each confederate and would disclose three wins to one of them. We counterbalanced the order of the outcomes and which of the two friends observed during these trials.

Results and Discussion

There was no effect of the order of the observed outcomes (W vs. L), the order of the friend (Anne or Sally), or the number of observation trials (one or three); thus the data were collapsed for further analysis. As in previous experiments, participants were asked to which friend Bill should tell about his win.In the Random condition, 68% (68 of 100) of participants chose the person who previously saw Bill lose (binomial test, $p < 0.001$), replicating the results from Experiment 2A. In the Skill condition, 71% (71 of 100) of participants chose this person ($p < 0.001$, by binom. test). Contrary to our prediction, there was no difference between the two conditions.

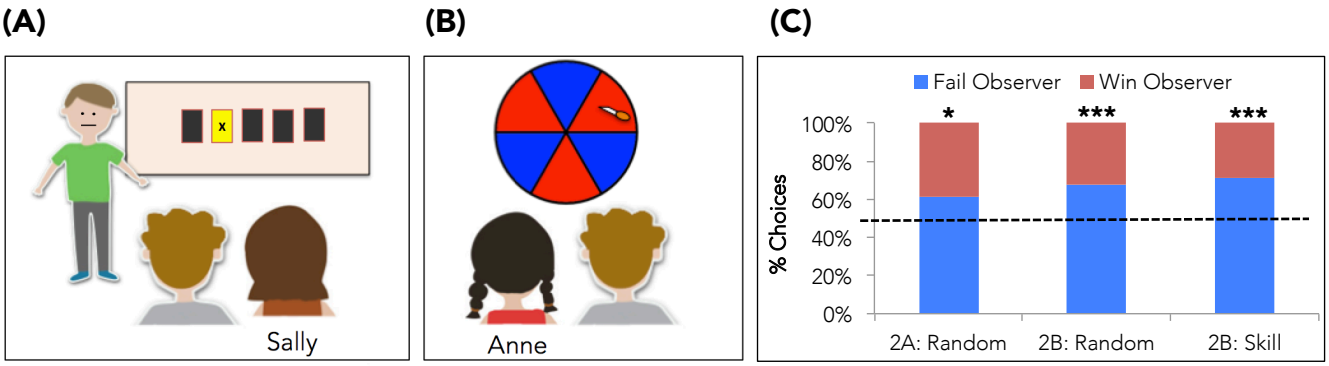


Figure 2: Experiment 2A "win" trial (A), Exp. 2B "loss" trial (B), and Exp. 2A and 2B results (C), $^{***}p < 0.001$, $^{*}p < 0.05$.

General Discussion

In this study, we explored how people communicate information about the self to others. More specifically, we asked whether children as well as adults would choose to disclose positive information about the self (e.g., winning a game) to someone who previously observed a negative outcome or to someone who observed a positive outcome about the self. If people understand that: (1) others' prior observations of the self can result in positive or negative beliefs about one's competence, and (2) one can revise such beliefs by communicating information about the self, they would be sensitive to others' prior observations and preferentially choose to communicate to the person who observed a negative outcome.

In Experiment 1, we tested our prediction with children between ages 3 - 5. We found that while younger children preferred to disclose their positive performance to someone who has previously seen them win rather than lose, older children were more likely to choose the person who has previously seen them lose. While older children's choice was not significantly above chance, children's age predicted their choice. These results suggest that even though young children might be biased to prefer someone who already holds a positive belief about them, children gradually understand the benefit of correcting others' negative beliefs. Under this interpretation, the chance-level choice in older children might suggest that they were torn between their preference for one observer (i.e., "Win") and their desire to communicate their winning outcome to the other observer (i.e., "Loss" observer).

One alternative explanation is that young children's preference for someone who holds a positive belief about them gradually disappears; in this case, older children's choice might reflect a genuine absence of preference between the two observers. However, intuitively even adults prefer people who hold positive impressions of them, and it is unlikely that such preference is absent in the older half of the children in our study. Future studies are needed to provide stronger support for the idea that our results reflects children's developing understanding that others' beliefs about the self can be revised and maintained via selective disclosure of information.

In Experiment 2, we tested adults' intuitions in analogous tasks and further explored the role of luck vs. skill in these inferences. Adults showed a systematic preference for the person who observed a negative outcome of the game, suggesting that they understand that one can change others' negative beliefs about the self by communicating positive information about the self. While we failed to find a difference between the luck-based and skill-based games, it is possible that the positive glow from a lucky win was just as powerful as a skill-based outcome for adults, or that the skill-based game also increased the desire to maintain people's positive beliefs about an underlying competence. We also did not find a difference in selectivity for the "Lose" observer when we varied the amount of evidence (one vs. three instances), but this may have also been due to an increase in desire to affiliate with the person who previously viewed multiple wins.

Future studies will further explore how the nature of these games and amount of observed evidence might differentially affect others' impressions of oneself. For instance, in the face of someone who holds a less positive belief about the child, we could ask whether the child would like to disclose a luck-based positive outcome or a skill-based positive outcome. Further, we could compare how children differentially disclose information to others who have weak positive (e.g., viewed one win) or strong positive (e.g., viewed multiple wins) evidence of the child's performance.

Previous work in children's understanding of disclosure have utilized relatively complicated vignettes and asked children to either endorse or reject others' disclosure about another agent to a third party agent (Shaw & Olson, 2015) or predict whether or not others will disclose information about themselves (e.g., Hicks, Liu, & Heyman, 2015). Our novel and simple paradigm provides a rich and naturalistic environment from which we can better understand how children infer others' beliefs of them based on what others have observed, and how they communicate information about themselves to change or maintain these beliefs.

Last, we hope that this work will inspire future research integrating work in social cognitive development and impression management behaviors. While reputation management behaviors have largely concerned valence-based impressions

(e.g., "good" or "bad"), there are many other kinds of beliefs that others can have about us. For instance, even our preferences or a set of beliefs can provide rich information about who we are, and selectively communicating these to others can be an effective way to manage the representations of ourselves that others hold.

Inferring others' beliefs about the self and knowing how to change and maintain them might be particularly important for young children who are just beginning to develop a coherent self-representation and forming initial relationships with others. This capacity to reason about others' beliefs of us is deeply important for navigating our daily communicative interactions, building meaningful, lasting bonds with others.

Acknowledgments

We thank Sumudu Rathnayake, Sara Altman, Chelsey Pan, and Amy Qian for help as confederates. We also thank Bing Nursery School and the children and families who participated. This work was supported by a Varieties of Understanding grant from the John Templeton Foundation to HG.

References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342.

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-scale deceit: deception as a marker of two-, three-, and four-year-olds' early theories of mind. *Child development*, *60*(6), 1263–1277.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6 - 21.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-Mind Training Causes Honest Young Children to Lie. *Psychological Science*, *26*(11), 1812–1821.

Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-year olds, but not chimpanzees, attempt to manage their reputations. *PloS One*, *7*(10).

Fu, G., Heyman, G. D., Qian, M., Guo, T., & Lee, K. (2015). Young children with a positive reputation to maintain are less likely to cheat. *Developmental Science*, *19*(2), 275-283.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37.

Gweon, H., Chu, V., & Schulz, L. E. (2014). To give a fish or to teach how to fish? Children weigh costs and benefits in considering what information to transmit. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 559–564.

Gweon, H., Pelton, H., Konopka, J. a., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341.

Gweon, H., Shafto, P., & Schulz, L. E. (2014). Children consider prior knowledge and the cost of information both in learning from and teaching others. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 565–570.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186-193.

Hicks, C. M., Liu, D., & Heyman, G. D. (2015). Young children's beliefs about self-disclosure of performance failure and success. *British Journal of Developmental Psychology*, *33*(1), 123–135.

Koenig, M. A., & Harris, P. L. (2007, 10). The basis of epistemic trust: Reliable testimony or reliable sources? *Episteme*, *4*, 264–284.

Lee, K. (2013). Little liars: Development of verbal deception in children. *Child Development Perspectives*, *7*(2), 91–96.

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young Children Are More Generous when Others Are Aware of Their Actions. *PLoS ONE*, *7*(10).

Polak, A., & Harris, P. (1999). Deception by young children following noncompliance. *Developmental Psychology*, *35*(2), 561–568.

Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of evidence to teach and deceive others. *Frontiers in Psychology*, *6*.

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382-389.

Shaw, A., Li, V., & Olson, K. (2013). Reputation is everything. *Navigating the social world*.

Shaw, A., & Olson, K. (2015). Whose idea is it anyway? The importance of reputation in acknowledgement. *Developmental Science*, *18*(3), 502-509.

Song, H.-J., Onishi, K., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition*, *109*(3), 295?315.

Strauss, S., & Ziv, M. (2012). Teaching is a natural cognitive ability for humans. *Mind, Brain, and Education*, *6*(4), 186–196.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, *72*(3), 655–684.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103 - 128.

Extended Metaphors are Very Persuasive

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
Matias Berretta (matias.berretta@oberlin.edu)
Peace Iyiewuare (peace.iyiewuare@oberlin.edu)
Department of Psychology, Oberlin College
120 W. Lorain Street; Oberlin, OH 44074 USA

Abstract

Metaphors pervade discussions of critical issues and influence how people reason about these domains. For instance, when crime is a *beast*, people suggest enforcement-oriented approaches to crime-reduction (e.g., by augmenting the police force); when crime is a *virus*, on the other hand, people suggest systemic reforms for the affected community. In the current study, we find that extending metaphoric language into the descriptions of policy interventions bolsters the persuasive influence of metaphoric frames for an array of important issues. When crime is a *beast*, people are even more likely to endorse "attacking" the problem with harsh enforcement tactics; when crime is a *virus* people are even more likely to to endorse "treating" the problem through social reform.

Keywords: Metaphor, framing, analogy, persuasion, political psychology, reasoning

Introduction

An economic system entails the production, distribution, and consumption of scarce resources. In natural language, though, economic systems are often described metaphorically, as *gardens* (e.g., "The *seeds* of economic *growth* were planted years ago. Today, they are just starting to *bear fruit*, and soon we will *reap the rewards*"), *vehicles* (e.g., "The economy is *off track* or *broken down*"), *bodies* (e.g., "The economy used to be *healthy* but is now *suffering*"), and *machines* (e.g., "The economy is *broken* and needs to be *fixed*").

Conventional metaphors make up as much as 10-20% of natural discourse (Steen et al., 2010) and can be especially persuasive (e.g., Sopory & Dillard, 2002). Metaphors highlight particular relationships in the domains they describe and, as a result, can encourage systematic patterns of inference (Lakoff & Johnson, 2008). For instance, if the economy is a *vehicle* and it is *broken down*, then getting it moving again might require a financial *jumpstart*. On the other hand, if the economy is a *stunted plant*, giving a momentary *jolt of nutrients, sunlight, and water* is unlikely to be an effective long-term solution for the health of the *plant*. Instead, the economy might be better served by consistent *sunlight, water, nutrients* and a *supportive environment* (e.g., investments in education and job training to provide a strong workforce).

Recent work has found that these implicit structural entailments of metaphors affect how people reason about important issues (e.g., Hauser & Schwarz, 2014; Landau, Sullivan, & Greenberg, 2009; Thibodeau & Boroditsky, 2011, 2013, 2015). For instance, a *war* metaphor for cancer makes for an excellent slogan and may facilitate fund-raising efforts at a societal level, but it also seems to downplay the role of relatively mundane behavior change in cancer prevention at an individual level (e.g., smoking less; Hauser & Schwarz, 2014); priming people to think of immigration as a type of *bodily contamination* leads people to adopt anti-immigration attitudes (Landauer & Dumais, 1997); and framing crime as a *virus* (rather than a *beast*) leads people to favor social reforms as a tool for crime-reduction over harsh enforcement and punishment (Thibodeau & Boroditsky, 2011, 2013, 2015).

Here we ask whether extending metaphoric language into the description of candidate responses can facilitate (or inhibit) the persuasive influence of a previously instantiated metaphoric frame. Will people be even more likely to support a proposal to reduce crime by focusing on the educational system when such a program is described as a "treatment" for a crime *virus*? Can metaphor framing effects be negated (or even reversed) when an extended metaphor is used to describe a proposal that would otherwise be incongruent with the metaphor frame, as in "The city should treat a crime [*virus*] by increasing the police force"?

Consistent extended metaphors may facilitate metaphor framing effects by (a) re-instantiating the initial frame and, in turn, further highlight the ways in which the congruent response maps on to the entailments of the initial frame (Nayak & Gibbs, 1990) or (b) by providing a lexical cue that links the frame to a response (Graesser & Bower, 1990). For instance, describing education reform as a "treatment" for a crime *virus* may emphasize the ways in which such a policy is similar to a treatment program for a disease (e.g., by focusing on the root cause of the problem). This would suggest that metaphors play an active role in shaping representations of complex problems and that this role can be facilitated by actively situating a response in relation to the conceptual entailments of the frame.

Alternatively, people may be drawn to "treating" a problem that is framed as a *virus* because the *virus* frame serves as a lexical prime for other *virus*-related language. Although such an account may be less compelling on theoretical grounds, if one could use an extended metaphor to persuade people to choose particular policy interventions, even through lexical priming, this would be an important and interesting result. It would suggest, for instance, that describing crime as a *virus* could promote a variety of policy interventions: both those that are conceptually congruent with the entailments of the metaphor and those that are not, simply by extending language that is consistent with the metaphor frame into the description of the response (e.g., "*Treat* the crime [*virus*] by increasing street patrols!").

Study 1: Conceptual Congruence

Methods

Participants Data from 99 valid participants were collected through Mechanical Turk.

Materials and Procedure Stimuli consisted of 10 stories, including scenarios about cheating, crime, education, ecology, housing, income inequality, a medical mistake, partisan politics, science, and sports. Each story was paired with two metaphor frames and two candidate responses that were designed to reflect realistic judgments relating to policy interventions, risk management, or blame attribution (see, e.g., Thibodeau & Boroditsky, 2011).

In Study 1, after reading about a non-metaphorically framed description of a target issue, participants were asked to match metaphor frames to candidate responses. For instance, after reading a description of a crime problem, participants were told: "Two of the city's officials are debating how to solve this problem; they tend to talk about the problem in different ways. One argues that crime is a *virus*; the other argues that crime is a *beast*. If you had to guess, which of the crime-reducing approaches listed below do you think is supported by each of the officials?:"

1. Increase street patrols that look for criminals.
2. Reform educational practices and create after school programs.

For each issue we identified candidate responses that seemed to map onto the conceptual entailments of different metaphor frames. For instance, a proposal to reform a city's educational system seemed more consistent with a crime *virus* – in which a city is a body that can be brought back to health by treating the root cause of the problem – whereas a proposal to increase a city's police force seemed more consistent with a crime *beast*.

Study 1 served as a manipulation check of these intuitions. The extent to which the sample displays similar patterns of matching behavior is taken to reflect the degree to which the mappings between the metaphor frames and response options are conceptually congruent. Highly consistent matching at a group level would indicate clear conceptual relationships between the metaphor frames and response options. Less consistent matching would indicate weaker conceptual relationships between the metaphor frames and response options.

Each participant was asked to match responses to frames for each of the 10 issues. The order of the issues, frames, and candidate responses was randomized across participants.

Results

The results of this task revealed that there were clear mappings between the responses and issue frames, consistent with the design of the materials – 81.1% consistent overall (95%CI: [.786, .834]). Analyzed separately (i.e., with 10 separate chi-square tests of independence), we found a significant difference in how the responses were matched to the frames for each issue, $\chi^2[1, N = 99]s > 13, ps < .001$, confirming that participants matched the response options to the frames at a rate much higher than chance (50%).

This suggests that our intuitions as researchers about the conceptual relationship between the metaphoric frames and responses options were consistent with the population from which the sample in the experiment will be drawn.

Study 2: Lexical Congruence

In the experiment described in the following section (Study 3), we implemented a 3 frame (*metaphor A, metaphor B, no metaphor*) by 3 extended metaphor (*consistent, inconsistent, no extended metaphor*) design, thereby creating nine versions of each stimulus item. In the *consistent* condition, extended metaphors were paired with their conceptually congruent response (e.g., "treat" with "education reform" and "attack" with "increasing police"). In the *inconsistent* condition, we paired extended metaphors with their conceptually incongruent response (e.g., "treat" with "increasing police" and "attack" with "education reform"; see Table 1).

Table 1: Examples of consistent and inconsistent uses of extended metaphor in the description of response options to an issue that described a crime problem. Italics added to highlight the extended metaphoric language.

Consistent
a. *Treat* the problem by reforming educational practices and creating after school programs. b. *Attack* the problem by increasing street patrols that look for criminals.
Inconsistent
a. *Attack* the problem by reforming educational practices and creating after school programs. b. *Treat* the problem by increasing street patrols that look for criminals.

In this section we confirm that the extended metaphors did, in fact, extend the initial frames *consistently* and *inconsistently* by using latent semantic analysis (LSA; Landauer & Dumais, 1997). LSA is a tool that measures the similarity of words, phrases, and texts as a function of their contextual co-occurrence and has been shown to, among other things, reliably predict response times in for lexically primed target words in a lexical decision task (Hutchison, Balota, Cortese, & Watson, 2008).

To conduct this analysis, we identified the words that were used to instantiate the initial metaphor frames for each issue (e.g., "virus plaguing" and "beast preying") and the words used to insatiate the extended metaphors in the description of the responses (e.g., "treat" and "attack"). Then we entered these word pairings into the LSA database, which yielded four similarity scores per issue (see Table 2). In every case two were designed (expected) to be consistent (more similar) and two were designed (expected) to be inconsistent (less

similar).

Table 2: Similarity scores between metaphor frames and extended metaphors as measured by LSA. For this issue (about crime), "virus"-"treat" and "beast"-"attack" were designed to be consistent whereas "virus"-"attack" and "beast"-"treat" were designed to be inconsistent.

	Treat	Attack
Virus	.28	.24
Beast	.18	.26

For each issue, we averaged ratings of similarity between the two *consistent* and the two *inconsistent* pairings. For instance, in this case the average cosine (LSA's metric of similarity) between the *consistent* pairings was .27 and the average cosine between the *inconsistent* pairings was .21. A paired t-test revealed that the consistent matches ($M = .261, SD = .139$) were more similar than the inconsistent matches ($M = .167, SD = .134$), $t[9] = 5.018, p < .001$, confirming the design of the materials.

The average similarity between the consistent matches was greater than the average similarity between the inconsistent matches for all of the ten issues except one. The one issue that yielded anomalous results described a medical clinic that had made a mistake in filling prescription medication. One metaphor framed the clinic as an *ecosystem* and was extended with the word "interactive"; the other framed the clinic as an *assembly line* and was extended with the word "station." LSA revealed that the relationship between the *consistent* pairings was the same as the relationship between the *inconsistent* pairings in this case.

The lack of a difference between the *consistent* and *inconsistent* extended metaphors for this issue raises a larger issue about metaphor frames that has so far been overlooked: sometimes a metaphor frame can be extended in multiple ways. For instance, in the crime context, the word "treat" was found to be more similar to the *virus* than *beast* frame; the word "attack" was found to be more similar to *beast* than *virus* frame. However, *viruses* can be "attacked" and *beasts* can be "treated." For this reason we divided the issues into two groups: those that were more ambiguous (less distinct or specific) with respect to the similarity between the initial frames and extended metaphors (i.e., cases where the initial frames were similar to both extended metaphors) and those that were less ambiguous (more distinct or specific) with respect to the similarity between the initial frames and extended metaphors. We will consider this dichotomy in analyzing the results of the experiment presented below. It may be the case that the extended metaphors are especially likely to facilitate (or inhibit) the effect of the initial frame when the similarity between the frame and extended metaphoric language is less ambiguous (more specific).

The difference in similarity between the initial frames and metaphor extensions was significantly higher for the five items that were identified as less ambiguous (more distinct and specific; $M = .143, SD = .037$) than for the five items that were identified as more ambiguous ($M = .046, SD = .027$), $t[8] = 4.733, p = .001$.

Together, the two norming studies validate the experimental design. For each issue, there were systematic conceptual and lexical relationships between the metaphor frames and candidate responses. A group of nave participants in Study 1 confirmed that the two response options for each issue mapped on to the conceptual entailments of different metaphor frames. Study 2 used LSA to confirm that the *consistent* extended metaphors were more similar to the initial frames than the *inconsistent* extended metaphors. In the following experiment, we will test whether people are sensitive to these relationships when metaphor frames are embedded in the description of a target issue and when extended metaphors are used to describe the response options.

Study 3: Framing Experiment

Methods

Participants Data from 988 valid participants were collected through Mechanical Turk.

Materials and Design Nine versions of each of 10 stimulus items were created by crossing the three framing conditions (*metaphor A, metaphor B, none*) with the three extended metaphor conditions (*consistent, inconsistent, no extended metaphor*). Participants were presented with one version of each of the 10 issues. Their task was to answer a follow-up question for each issue, for which there were two candidate responses. The follow-up question was designed to reveal whether people were sensitive to implicit conceptual entailments of the metaphor frames. In some cases, the follow-up question asked the participant to choose between policy responses (e.g., as in the case of crime); in other cases the follow-up question asked participants to attribute blame for an outcome or to speculate on some other aspect of the target domain (e.g., for the issue that described a mistake at a medical clinic, participants were asked whether the "nurse who administered the medicine" or the "computer system that relayed messages between the nurse and pharmacist" was more responsible for the mistake).

In the *consistent* and *inconsistent* conditions, metaphoric language was used to describe both of the response options, thereby affording careful control over the experimental manipulation (see Table 1). Each participant was exposed to one of the nine versions of each stimulus item. The order and version of the stimuli was randomized across participants.

Congruence To analyze data from the conditions that included a metaphoric frame, we coded responses as "congruent" or "incongruent" with the associated frame. For instance, in the context of the crime example, "Increase street patrols" was coded as congruent with the *beast* frame and incongruent with the *virus* frame, regardless of the presence/absence of extended metaphoric language.

This approach captures the joint effects of the pairs of metaphor frames and is consistent with prior work on metaphor framing (e.g., Robins & Mayer, 2000; Thibodeau & Boroditsky, 2011, 2013, 2015). One advantage of such a coding scheme is that it provides a clear metric for interpreting the degree to which metaphors influence judgments. Metaphor frames that do not systematically influence the way people think about an issue will yield a congruence score close to .5; metaphor frames that influence people to choose the response option that is consistent with the frame's entailments will yield a congruence score above .5; and metaphors that influence people to choose the response option that is inconsistent with the frame's entailments will yield a congruence score less than .5.

For instance, the results of the current experiment found that people were more likely to think that a city should focus on education reform in response to a crime *virus* (59.0%; no extended metaphor condition) than *beast* (51.9%; no extended metaphor condition). Since the shift is consistent with the predicted effects of the metaphor, the congruence score for this item is greater than .5 ($.590 + (1 - .519) = 1.071; 1.071/2 = .536$). In other words, 53.6% of responses to this item were congruent with the metaphor frame (i.e., 7.2% more participants chose the response option that was congruent with the virus metaphor when crime was framed as a *virus* compared to when crime was framed as a *beast*).

Note that if everyone, in both framing conditions, thought that the city should "increase the police force" in response to the crime problem, the congruence score for the item would be .5: all of the responses in the *beast* condition would be coded as congruent with the frame but none of the responses in the *virus* condition would be coded as congruent with the frame ($1 + 0 = 1; 1/2 = .5$). In this way, transforming participants' judgments into a congruence score allows for clear comparison across issues with different metaphor frames and response options.

Coding responses as congruent or incongruent with the metaphor frame requires that we collapse over the two metaphor frames for a given issue. As a result, congruence cannot be computed when the issue is presented without an initial metaphor (i.e. for the *no metaphor* frame trials).

Results

We found that participants chose the congruent response option 54.9%, 51.3% and 50.1% of the time in the *consistent*, *no extended metaphor*, and *inconsistent* extended metaphor conditions, respectively. A mixed ANOVA on the mean congruence ratings for the 10 issues with extended metaphor condition treated as a factor and issue as a repeated measure revealed a statistically significant difference in the degree to which participants chose the congruent response by extended metaphor condition, $F[2,18] = 4.011, p = .036$.

Due to the relatively small number of items, no pair-wise differences between extended metaphor conditions were significant in post-hoc testing. However, the pattern of results suggested that extending consistent language into the description of the response options made people more likely to choose the congruent response whereas extending inconsistent language into the description of the response options made people less likely to choose the congruent response. We present additional analyses that help to distinguish between the conditions below.

Did the specificity of the extended metaphor matter? In order to test whether the quality of the relationship between the initial frame and the extended metaphor affected this pattern of behavior, we added an additional factor into this analysis: whether LSA scored the relationships between the initial frames and extended metaphors as more ambiguous (i.e., both frames were similar to both extended metaphors) or less ambiguous (i.e., the frames were much more similar to the matching *consistent* extended metaphor and less similar to the *inconsistent* extended metaphor).

The result was a statistically significant interaction between this factor and the extended metaphor condition, $F[2,16] = 3.872, p = .043$. As shown in Figure 1, differences in congruence between the *no extended metaphor*, *consistent* extended metaphor, and *inconsistent* extended metaphor conditions were greater for items that were less ambiguous, $F[2,8] = 7.137, p = .017$, than for items that were more ambiguous, $F[2,8] = .134, p = .877$.

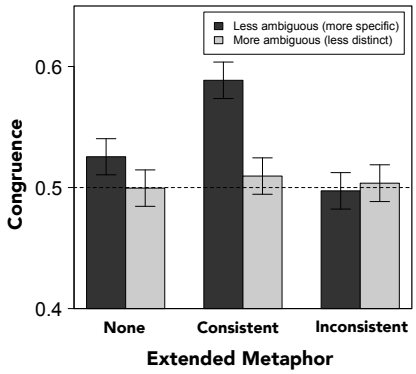


Figure 1: Proportion of congruent responses by extended metaphor condition for items with a less ambiguous relationship between the initial frame and extended metaphor (more specific to the consistent extended metaphor) and for items with a more ambiguous relationship between the initial frame and extended metaphor (indicating that the extended metaphor may be somewhat appropriate in the context of both frames).

This analysis suggests that the degree to which extended metaphors facilitate the persuasive influence of a frame depends, in part, on the specificity of the extended metaphor. For instance, although the word "treat" fits a *virus* frame better than it fits a *beast* frame (and the word "attack" fits a *beast* frame better than a *virus* frame), "treat" can also be used in reference to a *beast* (similarly, the word "attack" can be used in reference to a *virus*). For this issue, people chose

the congruent option more in the *consistent* than *inconsistent* extended metaphor conditions (55% compared to 51%); however, this difference is relatively small when compared to items in which the initial frames were specifically more related to the *consistent*, and not the *inconsistent*, extended metaphor. For instance, another issue described a research scientist as either *climbing a mountain* (extended with "gains ground") or *solving a puzzle* (extended with "looks for connections"). For this issue, 72% of responses were congruent in the *consistent* condition compared to 50% in the *inconsistent* condition). Not only was the relationship between a given frame and the matched extended metaphor relevant to participants' judgments, but the relationship between the frame and the alternative extended metaphor mattered as well.

How strongly did the metaphors affect judgments? So far, we have shown that people were more likely to choose the conceptually congruent response when it was described using *consistent* extended metaphoric language and that the specificity of the relationships between the initial frames and extended metaphors moderated this effect. However, we have not tested whether people chose the conceptually congruent response more often than one would expect by chance in any of the three conditions.

Here, we use mixed-effect logistic regression to test this important question and to make further comparisons between the conditions. On this approach, analyses are conducted at the level of the individual trial, rather than by averaging data over items or participants (i.e., by fitting a single model to participants' binary judgment for each issue; Bates, Maechler, Bolker, & Walker, 2013; Jaeger, 2008). This allows us to take advantage of the statistical power afforded by the relatively large sample and increases the reliability of the results (by reducing the probability of a Type 1 or Type 2 error; Jaeger, 2008). In the model, we included random effects for participant and issue to simultaneously account for error variance associated with these factors (i.e., participant and issue were treated as repeated measures; cf. Clark, 1973). An additional advantage of this approach is that it allows us to compare across all three levels of the framing manipulation (*metaphor frame A, metaphor frame B, and no frame*) as well as all three levels of the extended metaphor manipulation (*consistent, inconsistent, and no extended metaphor*).

We first confirmed the results presented above by testing for an interaction between the framing and extended metaphor manipulations. We compared two models[1]: one that included predictors for interactions between these factors and one that did not. We found that including predictors for the interactions significantly improved the fit of the model, $\chi^2[4] = 11.378, p = .023$. Post-hoc testing (Bonferroni-corrected $\alpha = .017$) revealed that participants were more likely to choose the congruent response in the *consistent* extended metaphor condition than in the *inconsistent* metaphor condition, $\chi^2[1] = 10.410, p = .001$, or the *no extended* metaphor condition, $\chi^2[1] = 7.608, p = .006$. There was no difference between the *inconsistent* and *no extended* metaphor conditions, $\chi^2[1] = .646, p = .422$.

We then tested whether people were significantly more likely than one would expect by chance to choose the congruent response option for each of the three extended metaphor conditions. We found that the metaphor frames affected participants' judgments when the candidate responses were described with *consistent* extended metaphors, $\chi^2[2] = 25.812, p < .001$, and when the candidate responses were described with *no extended* metaphors, $\chi^2[2] = 13.203, p = .001$. However, the metaphor frames did not affect participants' judgments when the candidate responses were described with *inconsistent* extended metaphors, $\chi^2[2] = 3.963, p = .138$.

In sum, these analyses confirmed the omnibus difference in participants' likelihood of choosing a congruent response by extended metaphor condition, and additionally revealed that people were significantly more likely to choose the congruent response when extended metaphors were used *consistently* to describe the candidate responses, relative to the *inconsistent* or *no extended* metaphor conditions. It also revealed that people were more likely to choose the congruent response than one would expect by chance in the *consistent* and *no extended* metaphor conditions but not in the *inconsistent* extended metaphor condition.

In other words, these analyses revealed evidence of a simple metaphor framing effect when the candidate responses were described without extended metaphors, which was amplified (facilitated) when the candidate responses were described with *consistent* extended metaphors. This effect was not negated when the candidate responses were described with *inconsistent* extended metaphors, as responses in this condition were no different from what one might expect by chance (i.e., this condition did not show a metaphor framing effect).

General Discussion

The results of this study indicate that metaphor framing is especially persuasive when *consistent* metaphoric language is extended to descriptions of candidate responses. Although people are more likely to endorse approaches to crime-reduction that emphasize social reform when crime is framed as a *virus*, they are even more likely to do so when the reform is described as a "treatment" (i.e., in the context of a re-instantiation of the initial metaphor frame).

We found differences of about 10 (54.9% congruent) and 3 (51.3% congruent) percentage points in the *consistent* and *no extended* metaphor conditions that were attributable to the metaphor frame. Of note, we also found that the specificity of the extended metaphor mattered. Extended metaphors that were specifically related to the matching frame, and distinct from the non-matching frame, were most likely to facilitate

[1]The deviance between the models (i.e., difference in likelihood ratios) is reported as an index of model fit: model deviance approximates a chi-square distribution with the number of added parameters as its degrees of freedom (Menard, 2002).

persuasion (58.9%; a shift of about 18 percentage points). For instance, the facilitative effect of the extended metaphor was less pronounced in the context of crime because *viruses* and *beasts* can both be "treated" and "attacked."

For comparison, in prior work, we have found shifts on the order of 15 percentage points in a free response task and 8 percentage points in a forced choice task, when response options were described without extended metaphors (Thibodeau & Boroditsky, 2011, 2013). One notable difference between the present work and prior studies is the use of multiple target domains and pairs of metaphor frames. Thus, we are in a position to make a more general claim about how metaphor frames influence reasoning, as well as to explore some of the boundary conditions of metaphor framing effects in future work (Steen, Reijnierse, & Burgers, 2014).

As noted in the introduction, there are at least two reasons that extended metaphors may facilitate persuasion: either because they re-instantiate the conceptual structure of the initial frame or because they provide a more associative (lexical) link to the initial frame. The present work suggests that the conceptual entailments and lexical associations of metaphors are mutually beneficial sources of information (see, e.g., Patterson, 2014), as people were most likely to be influenced by the metaphor frame when the conceptually related response option was described with a matching extended metaphor.

Although it is valuable to think about how these distinct sources of information contribute, separately and in combination, to long-term and on-line representations, it may not be possible to dissociate them completely. Further, at a process level, both of these sources of information can be can be modeled as a spreading activation in an associative network (Flusberg, Thibodeau, Sternberg, & Glick, 2010; Rogers & McClelland, 2008; John, 1992).

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using eigen and s4. *R package version*, *1*(4).

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., & Glick, J. J. (2010). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology*, *1*(197), 1-11.

Graesser, A. C., & Bower, G. H. (1990). *Inferences and text comprehension.* (Vol. XIX; G. H. B. Arthur C. Graesser, Ed.). Academic Press.

Hauser, D. J., & Schwarz, N. (2014). The war on prevention bellicose cancer metaphors hurt (some) prevention intentions. *Personality and Social Psychology Bulletin*, 0146167214557006.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1036–1066.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

John, M. F. S. (1992). The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*(2), 271–306.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Landau, M. J., Sullivan, D., & Greenberg, J. (2009). Evidence that self-relevant motives and metaphoric framing interact to influence political and social attitudes. *Psychological Science*, *20*(11), 1421–1427.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Nayak, N. P., & Gibbs, R. W. (1990). Conceptual knowledge in the interpretation of idioms. *Journal of Experimental Psychology: General*, *119*(3), 315.

Patterson, K. J. (2014). The analysis of metaphor: To what extent can the theory of lexical priming help our understanding of metaphor usage and comprehension? *Journal of psycholinguistic research*, 1–22.

Robins, S., & Mayer, R. E. (2000). The metaphor framing effect: Metaphorical reasoning about text-based dilemmas. *Discourse Processes*, *30*(1), 57–86.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(06), 689–714.

Sopory, P., & Dillard, J. P. (2002). The persuasive effects of metaphor: A meta-analysis. *Human Communication Research*, *28*(3), 382–419.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From mip to mipvu* (Vol. 14). John Benjamins Publishing.

Steen, G. J., Reijnierse, W. G., & Burgers, C. (2014). When do natural language metaphors influence reasoning? a follow-up study to thibodeau and boroditsky (2013). *PloS one*, *9*(12), e113536.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, *6*(2), e16782.

Thibodeau, P. H., & Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PloS one*, *8*(1), e52961.

Thibodeau, P. H., & Boroditsky, L. (2015). Measuring effects of metaphor in a dynamic opinion landscape. *PloS one*, *10*(7), e0133939.

How Does Generic Language Elicit Essentialist Beliefs?

Emily Foster-Hanson (emily.fosterhanson@nyu.edu)
Department of Psychology, New York University

Sarah-Jane Leslie (sjleslie@princeton.edu)
Department of Philosophy, Princeton University

Marjorie Rhodes (marjorie.rhodes@nyu.edu)
Department of Psychology, New York University

Abstract

Generic language (e.g., "tigers have stripes," "girls hate math") is a powerful vehicle for communicating essentialist beliefs. One way generic language likely communicates these beliefs is by leading children to generate kind-based explanations about particular properties; e.g., if a child hears "girls hate math," he may infer that there must be an inherent causal basis for the generalization, which in turn supports essentialist beliefs. However, it is also possible that simply hearing a category described with generics elicits the belief that the category is an appropriate kind to generalize about. On this account, even if the generic is negated ("girls don't hate math"), the generic language might nonetheless lead children to essentialize the category. The current study supports the latter possibility, suggesting that even hearing negated generics ("girls don't hate math") may still foster social essentialism.

Keywords: generic language; essentialism; conceptual development

Introduction

The ability to organize entities into categories is a key element of human cognition. Focusing on the similarities that unite category members allows even young children to use information about one member of a category to make inferences about other members (e.g., if one dog is friendly then other dogs are probably friendly too; Gelman & Markman, 1987). People tend to believe that certain categories, such as animal species, mark objective divisions in nature, and that members of these categories share a deep, underlying essence which gives rise to both observable and unobservable properties. These beliefs, referred to as psychological essentialism (Medin & Ortony, 1989), reflect pervasive cognitive biases that appear early in development (Gelman, 2003; Hirshfield, 1996). Psychological essentialism is an important element of conceptual development, helping to scaffold knowledge acquisition by guiding children to focus on similarities between category members (Gelman & Coley, 1990).

Children hold essentialist beliefs about only a fraction of the categories with which they are familiar. For instance, children and adults generally view animal categories in more essentialist terms than artifact categories (Diesendruck, 2003; Gelman, 2003; Gelman & Coley, 1990; Rhodes & Gelman, 2009; Rhodes, Gelman, & Karuza, 2014). Likewise, children develop essentialist beliefs about certain social categories but not others; the extent to which children view a particular social category in essentialist terms depends on the cultural input they receive throughout development (Diesendruck & Haber, 2009; Mahalingham & Rodriguez, 2006; Rhodes & Gelman, 2009). For example, European American 7- to 10-year-old children growing up in more politically conservative communities hold more essentialist beliefs about racial categories than those in more politically liberal communities (Rhodes & Gelman, 2009). For this reason, social categories provide an excellent window into how cultural input shapes essentialist beliefs.

An important form of cultural input that guides conceptual development is language. For instance, labeling objects with words leads preverbal infants to focus on similarities and facilitates categorization (Fulkerson & Waxman, 2007). Generic language – which characterizes kinds rather than individuals, e.g., "tigers have stripes," "birds lay eggs" – is a particularly powerful linguistic tool in guiding children's category reasoning. Children understand generics by preschool age (Cimpian & Markman, 2008; Cimpian, Meltzer, & Markman, 2011; Gelman & Raman, 2003) and they are common in child-directed speech (Gelman, Taylor, & Nguyen, 2004; Gelman, Waxman, & Kleinberg, 2008), particularly in pedagogical contexts (Gelman, Ware, Manczak, Graham, 2013).

Generic language is importantly related to psychological essentialism. For example, parents produce more generics for animal categories than for categories like artifacts, which are not generally viewed in essentialist terms (e.g., Brandone & Gelman, 2009). Further, limited exposure to generics leads to increased essentialist beliefs about novel animal (Gelman, Ware, & Kleinberg, 2010) and social (Rhodes, Leslie, & Tworek, 2012) categories. For instance, Rhodes, Leslie, and Tworek (2012) had children and adults read a storybook about a novel social category called "Zarpies," consisting of either generic statements (e.g., "Zarpies sleep in tall trees"), or matched specific statements about one member of the category (e.g., "This Zarpie sleeps in tall trees"). Subjects who heard generic statements came to view the category in more essentialist terms than those who heard specific statements. In addition, inducing essentialist beliefs about Zarpies led parents to *produce* more generic language about the kind.

Generic language is a powerful means of communicating essentialist beliefs, yet how generics accomplish this has not

been explained. One likely mechanism through which generic statements increase essentialism is by eliciting kind-based explanations as to why the particular property is generally shared among category members. Prior work has found that children are more likely to expect properties predicated in generic statements to arise from deep, kind-based causal structures, whereas they expect the same properties predicated in specific statements to arise from more idiosyncratic causes (Cimpian & Erickson 2012; Cimpian & Markman, 2009; 2011). For example, when children heard a novel ability attributed to a gender category using generic language (e.g., "boys are really good at *leeming*") they were more likely to explain the ability in essentialist terms, attributing it to inherent traits of the category (e.g., "because boys are tougher than girls"). In contrast, when the same novel property was introduced with a non-generic statement (e.g., "this boy is good at *leeming*") they offered explanations based on individual effort (e.g., "because he got teached;" Cimpian & Markman, 2011; Cimpian & Erickson, 2012). Thus, perhaps explaining properties as arising from kind-based causal mechanisms may lead children to develop a belief in a place-holder "essence" or common causal mechanism for the category members' shared properties. For instance, if children hear "Zarpies buzz when they are sad," they might reason that this property is due to an inherent shared nature, e.g., because Zarpies are not able to cry. Such reasoning may lead them to intuit that Zarpies generally share a common underlying nature that could potentially give rise to other shared properties as well.

An alternative (or perhaps additional) mechanism through which generic statements could elicit essentialism, however, is by more generally signaling that a particular category is appropriate to make generalizations about. By uttering a generic statement, a speaker could simply be communicating the belief that it is appropriate to generalize about the category in question. As children depend on cultural cues to determine which categories support rich inductive generalizations (Diesendruck & Haber, 2009; Mahalingham & Rodriguez, 2006; Rhodes & Gelman, 2009), generics may serve this function by signaling to a child that a given category provides a strong basis for generalizing. Notably, on this hypothesis, the specific nature of the properties being generalized is not the primary factor in explaining essentialist beliefs. What matters is that the kind is being flagged as an appropriate locus of generalization; the specific link between the kind and the predicated property does not play an important role.

Crucially, these two processes need not be mutually exclusive. It is quite possible that essentialist beliefs about a category might be fostered in tandem both by cues that the category is appropriate to generalize about and through causal reasoning about particular properties. Importantly, however, the relative weights of these two processes in eliciting essentialist beliefs imply different recommendations for how a parent or educator might most effectively correct the generic statements that children hear in their daily lives (e.g., "girls hate math") in order to lessen their effects on children's essentialist beliefs. For instance, if generics elicit increased essentialism primarily by leading children to construct a kind-based causal framework for particular properties, then negating the properties would be sufficient to lessen the generic statement's effect (e.g., "No, girls don't hate math). Such a correction clarifies that the property is not generally associated with the kind, presumably halting the need to generate a kind-based causal explanation for the generalization. If, however, generics elicit essentialism primarily by signaling that a particular kind is appropriate to generalize about, then such corrections may not be effective in reducing the tendency to essentialize. That is, as the statement "girls don't hate math" is still itself a generic, such a correction may still leave intact the implication that *girls* constitutes a rich, inductive basis for generalizing. If this hypothesis were supported, then correcting these statements to apply to a particular individual would be a more effective strategy (e.g., "No, only this one girl hates math").

The current study tests whether negating generic statements, and thereby hindering the causal reasoning process, nonetheless elicits essentialist beliefs about a novel social category. Is it necessary for children to believe that a set of positive properties is associated with a novel category in order to develop essentialist beliefs? Or is simply hearing the kind treated as an appropriate locus of a generic generalization sufficient? If the input does not present clear properties about which to construct a kind-based explanatory framework, do children nonetheless develop essentialist beliefs when hearing generic language? The present study focuses on social categories, as they are particularly sensitive to cultural input. Children heard generic statements about a novel social category called "Zarpies," based on materials used by Rhodes et al. (2012). Children heard generic statements (e.g., "Zarpies have striped hair") uttered by one puppet, which were immediately denied by a second puppet. Puppet 2's corrections were of three different types, depending on condition. For a third of participants, these corrections were in the form of direct negations (Generic Negation condition, e.g., "No, no, no! Zarpies don't have striped hair.") Another third of participants heard Puppet 2 offer an alternative property (Generic Replacement condition, e.g., "No, no, no! Zarpies have spotted hair.") The remaining subjects heard Puppet 2 correct only Puppet 1's generalization to the kind (Specific condition, e.g., "No, no, no! *This* Zarpie has striped hair.")

If generics communicate essentialist beliefs mainly by signaling that a kind is appropriate to generalize about, then both generic conditions (Generic Negation and Generic Replacement) should elicit increased essentialism compared with the Specific condition. In both generic conditions, Puppet 2 does not challenge the implication that it is appropriate to generalize across the category, only that the particular property given by Puppet 1 is a correct generalization. In contrast, in the Specific condition, Puppet

2 challenges the generalization's scope directly (by saying that the property applies to just this particular Zarpie).

However, if constructing kind-based causal explanations for particular properties is necessary for generics to elicit essentialist beliefs, then only the Generic Replacement condition should elicit increased essentialism relative to the Specific condition. In the Generic Replacement condition, children receive a set of properties to reason about (i.e., the properties described by Puppet 2), whereas in the Generic Negation condition, they do not know what particular properties are held by Zarpies (as Puppet 2 negates all of the generics stated by Puppet 1).

Method

Participants

Participants included 51 four- and five-year-old children (25 male, M_{age} = 4.91 years, range = 3.99-5.89 years; approximately 41% European American, 9% African American, 5% Asian American, 13% Hispanic, 14% Multiethnic, and 18% unreported). Participants were recruited from and tested at the Children's Museum of Manhattan. Written parental consent was obtained for all participants; children provided oral assent. All study procedures were approved by the Institutional Review Board of New York University.

Procedure

The procedure was divided into three phases. In the first phase, children were introduced to a novel social category, called "Zarpies." The experimenter showed participants an illustration of four Zarpies, who varied by race and gender, and said, "Here is a picture of some Zarpies." The experimenter then pointed to each Zarpie one at a time and said, "This is a Zarpie." Children were asked to repeat the category name to ensure they had understood the introduction before moving on to the experimental phase.

The second phase of the study contained the manipulation by condition in the form of a short puppet show. Subjects were first randomly assigned to one of three conditions. They then watched a puppet show with two animal puppets operated by the experimenter. The experimenter introduced the two puppets as "Sally" (Puppet 1) and "Jenny" (Puppet 2) and told participants that they would be playing a game with Puppet 1 where she looked inside a special box and told the child what she saw. However, they were told, Puppet 1 could not see very well and had forgotten her glasses, so Puppet 2 was going to help her by telling her if she got it wrong. To ensure children understood the procedure, they completed a warm-up round in which Puppet 1 mistakenly said a small blue bear toy in the box was a yellow ball and was corrected by Puppet 2.

After the warm-up, children were presented with a series of 16 generic statements about Zarpies from Puppet 1 and corrections from Puppet 2. The statements were modified from those used by Rhodes et al. (2012), and each expressed a unique physical or behavioral property of Zarpies. Children were first told that there were pictures of Zarpies in the box and that Puppet 1 was going to look at them one at a time and tell them what she saw. Puppet 1 then looked inside the box and uttered a generic statement about Zarpies (e.g. "Zarpies have striped hair"). Puppet 2 then also looked inside the box and disagreed with Puppet 1 in one of three ways, depending on condition. In the Generic Negation condition, Puppet 2 corrected Puppet 1 by simply negating the statement (e.g., "No, no, no! Zarpies don't have striped hair"). In the Generic Replacement condition, Puppet 2 instead provided an alternative property (e.g., "No, no, no! Zarpies have spotted hair"). In the Specific condition, Puppet 2 corrected the statement as that it only applied to the depicted individual (e.g., "No, no, no! *This* Zarpie has striped hair"). After each correction, the experimenter moved on to the next statement by Puppet 1, followed by another correction by Puppet 2, and so on. Subjects heard all 16 items in a row.

The third phase, following the puppet show, consisted of questions that probed the child's essentialist beliefs about the novel social category Zarpies. Test items included two measures of essentialist beliefs, modeled on those used in Rhodes et al., (2012): three explanation questions, and three inheritance questions (using a switched at birth task). For the explanation items, children were asked to explain a property that they had not heard about during the puppet show (e.g., "Why is this Zarpie chasing a shadow?") Explanation content and scope were coded according to a schema used in previous research (Cimpian & Erickson, 2012; Cimpian & Markman, 2009; Gelman et al., 2010; Rhodes et al., 2012). Explanation content was coded with intrinsic causes ("he loves to chase shadows") as 1, and with incidental causes ("he wants to catch it") as 0. Explanation scope was coded with category-based explanations ("because Zarpies like to…") as 1, and with individual explanations ("because he likes to…") as 0.

For inheritance items, children were told that a baby was born to a Zarpie mother but raised by a non-Zarpie mother. They were asked to predict whether the baby would display three novel properties of the Zarpie mother, or alternative properties of the adoptive mother, when it grew up. For instance, children were told that the Zarpie mother loves to eat flowers, but the other mother loves to eat crackers, and were asked whether the baby would love to eat flowers or crackers. Inheritance items were also scored for essentialism using a schema from prior work, with properties of the Zarpie mother coded as 1 and those of the adoptive mother coded as 0 (Gelman et al., 2010; Rhodes et al., 2012; Springer & Keil, 1989; Waxman, Medin, & Ross, 2007). After each property, children were asked whether their judgments were flexible, i.e., if the baby might also display the unselected property. Inflexible "no" responses were coded as 1, and flexible "yes" responses were coded as 0 (Rhodes et al., 2012; Taylor, Rhodes, & Gelman, 2009).

Participants' responses were recorded on an answer form by the experimenter, and an independent coder who was blind to the study design coded explanation responses.

Sessions were also videotaped, and a secondary coder transcribed participants' responses verbatim from video and coded all videos for children's responses. Initial agreement between the coders was 87%, with disagreements resolved by the first author.

A set of memory questions was included at the end of the study. The questions differed across conditions and were included in order to check that subjects remembered the corrections from Puppet 2 rather than the original generic statements uttered by Puppet 1. Participant responses were scored such that higher scores in each condition represented endorsement of the properties expressed by Puppet 2.

Results

Experimental Results

Responses consisted of a series of binary responses in which the dependent variables were the number of times participants gave essentialist responses. Data were analyzed with binomial generalized regression models using the Generalized Linear Model (GLM) procedure in SPSS 2.0. Our primary analysis consisted of a total essentialism score -- the total number of essentialist responses given across the explanation items and the inheritance questions, presented as probabilities of essentialist responses across items.

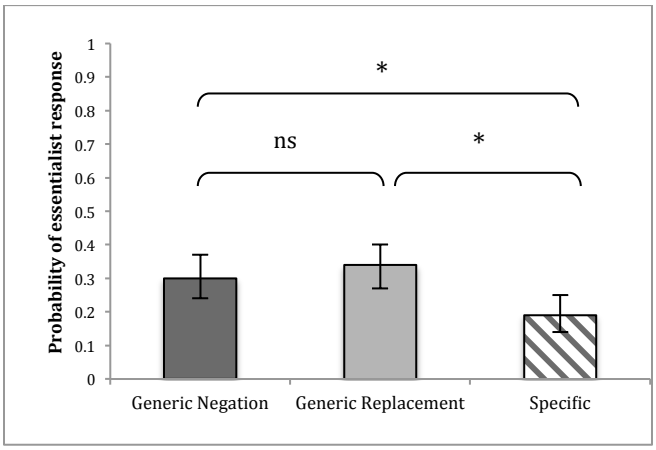


Figure 1: Probabilities of giving an essentialist response by condition: Generic Negation = "Zarpies don't have striped hair;" Generic Replacement = "Zarpies have spotted hair;" Specific = "This Zarpie has striped hair;" (* $p < 0.05$).

Children's essentialist responses varied by condition (Wald X^2 (2) = 12.72, $p < .01$). Children gave more essentialist responses in the Generic Negation condition (e.g., "Zarpies don't have striped hair," $M = .30$, $CI = .24$-$.37$) and the Generic Replacement condition (e.g., "Zarpies have spotted hair," $M = .34$, $CI = .27$-$.40$) than in the Specific condition (e.g., "This Zarpie has striped hair," $M = .19$, $CI = .14$-$.25$; $ps < .05$), and responses in the two Generic conditions did not differ from one another. Relative to the Specific condition, the Generic Negation condition increased the odds of essentialist responses by 1.81, CI = 1.14-2.86, and the Generic Replacement condition increased the odds of essentialist responses by 2.13, CI = 1.35-3.36.

Memory Check Results

Performance on memory check questions showed that memory for corrections from Puppet 2 overall was very good, and exceeded chance overall ($M = .86$, $CI = .81$-$.92$) and within each condition (all $ps < .01$). Memory performance was not correlated with essentialist responses (r (51) = .15, $p = .31$).

Discussion

Hearing generic statements about a novel social category led children to develop the same degree of essentialist beliefs, regardless of whether the generics predicated positive properties of the kind or were simply negative, denying that a given property is shared among the members of the kind. These results are a preliminary indication that generic statements instill essentialist beliefs by signaling that a given category is an appropriate basis for generalizing.

This interpretation does not preclude a role for kind-based explanations in the development of essentialist beliefs. Hearing particular properties expressed in generic language may lead children to view those *properties* in essentialist terms, even if it is not the primary means through which generics increase essentialist beliefs about the kind in general. For example, hearing a novel ability, e.g., *leeming*, associated with a gender category may lead children to expect that *leeming* ability is due to stable, fixed causes linked to category membership. To take a real-world example, hearing athletic ability associated with a racial category might lead people to expect that an individual's athletic ability can be explained and predicted by his race. Thus, generating kind-based explanations may play an important role in the development of children's (and adult's) essentialist beliefs that particular properties are linked to categories (e.g., that athletic ability is linked to race), even if generating such explanations is not critical to the development of essentialist beliefs about kinds more generally (e.g., that race is an essential kind).

One limitation of the current study is that children could be interpreting the negated statements in the Generic Negation condition (e.g., "Zarpies don't have striped hair") as generic statements about negative properties (i.e., the property of not having striped hair.) Future research is investigating this possibility by more fully disassociating properties from the category. In an ongoing follow-up study, Puppet 2 completely rejects Puppet 1's utterance: "No, no, no! That's not what this picture shows. That's not right about Zarpies!" If children are in fact interpreting the negated generic statements in the current study as expressing alternative (negative) properties, then children should not develop increased essentialist beliefs about Zarpies in this new condition. In contrast, if simply hearing a generalization about the category is sufficient to elicit increased essentialism, as the results of the present study

indicate, then children should still develop increased essentialist beliefs in this new condition.

These results also do not provide an exhaustive account of how generic language communicates essentialist beliefs. For instance, it remains unknown how much generic input children need to develop these beliefs, and whether the effects are cumulative across time. The generic input children heard about the novel social category Zarpies in the current design is very limited; children most likely hear significantly more input about the actual categories they come to view in essentialist terms in their daily lives. Future research might investigate if and how these factors modulate the effect of generic language on essentialism.

One interesting question left open by the current findings is the extent to which a listener's existing conceptual structure interacts with the minimal category information offered in the current study. For instance, Gelman et al. (2010) found that hearing a series of generic statements about a novel animal category lead to increased essentialist beliefs about the category in both children and adults. Children and adults generally hold high levels of essentialist beliefs about animal kinds and these beliefs are less dependent on cultural input than social categories (Rhodes & Gelman, 2009). Would minimally informative generics like those used in the current study be even more powerful in eliciting essentialist beliefs about these categories given children's existing essentialist expectations?

In a similar vein, would adults – who hold more established conceptual structure – be equally sensitive to the minimal generics used in the current study? Prior research has shown adults to be susceptible to generic language in developing essentialist beliefs about novel categories, both biological (Gelman et al., 2010) and social (Rhodes et al., 2012). Future research might investigate whether adults, like children, develop these beliefs even if the specific properties expressed in the generic statements are negated. Rhodes et al. (2012) found that adults gave more essentialist responses than children after hearing generic language about Zarpies. Thus, it is possible that adults require even less input than children to develop essentialist beliefs, perhaps due to a more robust expectation that the social world consists of essential kinds.

A more detailed understanding of the ways in which subtle linguistic cues foster the development of essentialist beliefs is of great social importance. Though psychological essentialism can be a powerful tool for learning about regularities in one's environment, viewing certain types of categories, like race and gender, in essentialist terms can have pernicious consequences. Essentialist beliefs about social categories can facilitate social stereotyping and prejudice (Gelman, Heyman, & Legare, 2007; Haslam, Rothschild, & Ernst, 2002; Hirschfield, 1996; Keller, 2005; Leslie, 2008, 2013; Prentice & Miller, 2007). For example, children who hold essentialist beliefs about gender might believe that if one girl is bad at math, then girls in general are probably bad at math (Haslam et al., 2002; Leslie, 2013). Understanding how corrections to generic statements lessen their effect on essentialism implies possible strategies for counteracting stereotyping and prejudice. Based on the current findings, simply negating a generic statement may not be very effective, while correcting the scope of the statement from generic to specific may be more potent.

One reason generic language may have such a powerful effect on the development of stereotypes is that simply negating them is not sufficient to undermine their influence. Thus, hearing a negated generic statement ("girls don't hate math") can nonetheless lead children to conceptualize gender in more essentialist terms. By uttering even a negated generic statement about girls a speaker is communicating the view that gender is a valid category for generalizations. The present results indicate that this alone is sufficient to elicit increased essentialist beliefs.

Acknowledgments

We thank our undergraduate research assistants for help with recruitment and data collection. We are also grateful to the parents and children who participated at the Children's Museum of Manhattan. This work was supported by NSF grant BC-1226942 to Rhodes and Leslie.

References

Brandone, A.C., & Gelman, S.A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts: A window onto a conceptual divide. Cognition 110(1), 1–22. http://dx.doi.org/10.1016/j.cognition.2008.08.005

Cimpian, A., & Erickson, L. C. (2012). The effect of generic statements on children's causal attributions: Questions of mechanism. Developmental Psychology, 48(1), 159–170. http://dx.doi.org/10.1037/a0025274

Cimpian, A., & Markman, E. M. (2008). Preschool children's use of cues to generic meaning. Cognition, 107(1), 19–53. http://dx.doi.org/10.1016/j.cognition.2007.07.008

Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. Cognition 113(1), 14–25. http://dx.doi.org/10.1016/j.cognition.2009.07.004

Cimpian, A., Markman, E.M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. Child Development, 471-492. http://dx.doi.org/10.1111/j.1467-8624.2010.01525.x

Cimpian, A., Meltzer, T. J., & Markman, E. M. (2011). Preschoolers' use of morphosyntactic cues to identify generic sentences: Indefinite singular noun phrases, tense, and aspect. Child Development, 82(5), 1561–1578. http://dx.doi.org/10.1111/j.1467-8624.2011.01615.x

Diesendruck, G. (2003). Categories for names or names for categories? The interplay between domain-specific conceptual structure and language. Language and Cognitive Processes, 18(5-6), 759-787. http://dx.doi.org/10.1080/01690960344000116

Diesendruck, G., Haber, L. (2009). God's categories: The effect of religiosity on children's teleological and essentialist beliefs about categories. Cognition 110(1), 100–114. http://dx.doi.org/10.1016/j.cognition.2008.11.001

Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not Tones) Facilitate Object Categorization: Evidence From 6- and 12-Month-Olds. Cognition,105(1), 218–228. http://dx.doi.org/10.1016/j.cognition.2006.09.005

Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195154061.001.0001

Gelman, S. A., & Coley, J. D. (Sep 1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. Developmental Psychology, 26(5), 796-804. http://dx.doi.org/10.1037/0012-1649.26.5.796

Gelman, S. A., Heyman, G. D., & Legare, C. H. (May-Jun 2007). Developmental changes in the coherence of essentialist beliefs about psychological characteristics. Child Development, 78(3), 757-774. http://dx.doi.org/10.1111/j.1467-8624.2007.01031.x

Gelman, S. A., & Markman, E. M. (Dec 1987). Young children's inductions from natural kinds: The role of categories and appearances. Child Development, 58(6), 1532-1541. http://dx.doi.org/10.2307/1130693

Gelman, S. A., & Raman, L. (2003). Preschool children use linguistic form class and pragmatic cues to interpret generics. Child Development, 24, 308-325. http://dx.doi.org/10.1111/1467-8624.00537

Gelman, S. A., Taylor, M. G., & Nguyen, S. P. (May 2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. Monographs of the Society for Research in Child Development, 69(1), 93-118. http://dx.doi.org/10.1111/j.0037-976X.2004.00279.x

Gelman, S. A., Ware, E. A., & Kleinberg, F. (Nov 2010). Effects of generic language on category content and structure. Cognitive Psychology, 61(3), 273-301. http://dx.doi.org/10.1016/j.cogpsych.2010.06.001

Gelman, S. A., Ware, E. A., Manczak, E. M., & Graham, S. A. (2013). Children's sensitivity to the knowledge expressed in pedagogical and nonpedagogical contexts. Developmental psychology, 49(3), 491. http://dx.doi.org/10.1037/a0027901

Gelman, S. A., Waxman, S. R., & Kleinberg, F. (Apr-Jun 2008). The role of representational status and item complexity in parent-child conversations about pictures and objects. Cognitive Development, 23(2), 313-323. http://dx.doi.org/10.1016/j.cogdev.2008.03.001

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? British Journal of Social Psychology, 41(1), 87-100. http://dx.doi.org/10.1348/014466602165072

Hirschfeld, L. A. (1996). Race in the making: Cognition, culture, and the child's construction of human kinds. MIT Press.

Keller, J. (2005). In genes we trust: the biological component of psychological essentialism and its relationship to mechanisms of motivated social cognition. Journal of personality and social psychology, 88(4), 686. http://dx.doi.org/10.1037/0022-3514.88.4.686

Leslie, S. J. (2008). Generics: Cognition and acquisition. Philosophical Review,117(1), 1-47. http://dx.doi.org/10.1215/00318108-2007-023

Leslie, S. J. (2013). Essence and natural kinds: When science meets preschooler intuition. Oxford studies in epistemology, 4, 108-166. http://dx.doi.org/10.1093/acprof:oso/9780199672707.003.0005

Mahalingam, R., & Rodriguez, J. (2006). Culture, brain transplants and implicit theories of identity. Journal of Cognition and Culture, 6(3), 453-462. http://dx.doi.org/10.1163/156853706778554968

Medin, D. L., & Ortony, A. (1989). Chapter: Psychological essentialism. 179-195.

Prentice, D. A., & Miller, D. T. (2007). Psychological essentialism of human categories. Current directions in psychological science, 16(4), 202-206. http://dx.doi.org/10.1111/j.1467-8721.2007.00504.x

Rhodes, M., & Gelman, S. A. (Oct 2009). Five-year-olds' beliefs about the discreteness of category boundaries for animals and artifacts. Psychonomic Bulletin & Review, 16(5), 920-924. http://dx.doi.org/10.3758/PBR.16.5.920

Rhodes, M., Gelman, S. A., & Karuza, J. C. (Jan 2014). Preschool ontology: The role of beliefs about category boundaries in early categorization. Journal of Cognition and Development, 15(1), 78-93. http://dx.doi.org/10.1080/15248372.2012.713875

Rhodes, M., Leslie, S., & Tworek, C. M. (2012). Cultural transmission of social essentialism. PNAS Proceedings of the National Academy of Sciences of the United States of America, 109(34), 13526-13531. http://dx.doi.org/10.1073/pnas.1208951109

Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. Child Development, 60, 637-648. http://dx.doi.org/10.2307/1130729

Taylor, M. G., Rhodes, M., & Gelman, S. A. (Mar-Apr 2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. Child Development, 80(2), 461-481. http://dx.doi.org/10.1111/j.1467-8624.2009.01272.x

Waxman, S., Medin, D., & Ross, N. (2007). Folkbiological reasoning from a cross-cultural developmental perspective: early essentialist notions are shaped by cultural beliefs. Developmental Psychology, 43(2), 294. http://dx.doi.org/10.1037/0012-1649.43.2.294

Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures

Sebastian Musslick[1,*], Biswadip Dey[2,*], Kayhan Özcimder[1,2,*],
Md. Mostofa Ali Patwary[3], Theodore L. Willke[3], and Jonathan D. Cohen[1]
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.
[2]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.
[3]Parallel Computing Lab, Intel Corporation, Santa Clara, CA 95054, USA.
[*]Equal Contribution, Corresponding Author: *musslick@princeton.edu*

Abstract

The limited ability to simultaneously perform multiple tasks is one of the most salient features of human performance and a defining characteristic of controlled processing. Based on the assumption that multitasking constraints arise from shared representations between individual tasks, we describe a graph-theoretic approach to analyze these constraints. Our results are consistent with previous numerical work (Feng, Schwemmer, Gershman, & Cohen, 2014), showing that even modest amounts of shared representation induce dramatic constraints on the parallel processing capability of a network architecture. We further illustrate how this analysis method can be applied to specific neural networks to efficiently characterize the full profile of their parallel processing capabilities. We present simulation results that validate theoretical predictions, and discuss how these methods can be applied to empirical studies of controlled vs. and automatic processing and multitasking performance in humans.

Keywords: multitasking; cognitive control; capacity constraint

Introduction

The human ability to carry out multiple tasks concurrently – a longstanding focus of cognitive research – presents an interesting puzzle. In some domains, humans can fluidly execute a large number of behaviors concurrently (e.g., locomote, navigate, talk, and bimanually gesticulate). However, in other domains, this capacity is strikingly limited (e.g., conduct mental arithmetic while constructing a grocery list). In addition to their obvious practical importance, constraints on multitasking are also of theoretical significance. Any general theory of cognition must address how choices are made among the limited set of behaviors that can be carried out at a given time (Anderson, 2013; Lieder & Griffiths, 2015; Kurzban, Duckworth, Kable, & Myers, 2013; Shenhav, Botvinick, & Cohen, 2013), and thus the sources of such limitations occupy a central role in cognitive theory.

Whether a set of tasks can or cannot be carried out concurrently has often been attributed to a fundamental distinction between automatic and controlled processing, with the former relying on parallel processing mechanisms and the latter on a limited capacity, serial processing (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). However, this begs a fundamental question: why are control-dependent processes capacity limited? Early theories, as well as some of the most successful unified theories of cognition (e.g., ACT-R) have assumed that this constraint reflects an intrinsic, structural property of the control system itself (e.g., limited capacity of working memory). However, alternative accounts have suggested that limitations in multitasking capacity reflect local properties of the mechanisms used for task execution, rather than an intrinsic property of the control system itself. According to such accounts, constraints on multitasking arise when two tasks call upon the same local resources (e.g., representations specific to the tasks) for different purposes (Allport, 1980; Meyer & Kieras, 1997; Navon & Gopher, 1979; Salvucci & Taatgen, 2008) and thus cannot be performed at the same time.[1]

Building on this idea, it has been proposed that a fundamental purpose of control mechanisms is to prevent crosstalk, by limiting the engagement of representations used by multiple processes ("mulituse representations") to a single purpose (e.g., task) at any given time (e.g. Cohen, Dunbar, & McClelland, 1990; Botvinick, Braver, Barch, Carter, & Cohen, 2001). From this perspective, constraints on multitasking of control-demanding processes reflect the *purpose* of control, rather than an intrinsic limit in control mechanisms. To the extent that the processing pathways required to perform different tasks rely on shared (i.e., multiuse) representations, not only do they become increasingly reliant on control (to specify the current intended use, and avoid crosstalk from competing uses), but the multitasking capacity of the network becomes limited (i.e., driven toward serial processing). In other words, control mechanisms are guilty by association, rather than themselves the source of constraints on multitasking.

One question that might be asked is: how does the constraint on multitasking imposed by pathway overlap scale with network size? A naive assumption might be that, in large networks (such as the brain), the constraint is relatively weak, and thus is inadequate to explain the prohibitive constraints apparent in human control-dependent processing. We have addressed this question in previous work, by examining the effects of pathway overlap (multiuse of representations) in

[1]Multitasking (and apparent parallelism) can, in some situations, be achieved by rapid switching between serial processes (as is common in computers). Here, we focus on forms of multitasking that reflect truly concurrent processing, sometimes referred to as perfect timesharing or pure parallelism. In the General Discussion, we consider how our findings concerning the conditions for such parallelism relate to the capability for rapid serial processing.

two types of networks of varying size (Feng et al., 2014). We found that even modest degrees of pathway overlap produced a strikingly strong constraint on parallel processing that was nearly scale invariant. This supported the idea that constraints in human multitasking may reflect representational multiuse, rather than a limitation intrinsic to control mechanisms themselves. However, while this work was suggestive, it relied on numerical simulations that were restricted to a limited range of parameters. It also failed to provide a clear path from these theoretical ideas to empirical validation.

Here, we conduct an exhaustive analysis of the relationship between pathway overlap and parallel processing in single-layered, feed-forward, non-linear networks. Our findings validate and extend those of Feng et al. (2014), identifying additional factors that influence the relationship between pathway overlap and parallel processing capability. We also show how these analysis methods can be used to fully specify the multitasking capabilities of a network, and validate derived theoretical predictions in simulated neural networks. Critically, we suggest how this method could also be applied to empirical data to determine the multitasking capabilities of natural agents in realistically large task spaces. Finally, we discuss related results using these methods to examine the interaction between pathway overlap, learning and generalization.

Graph-Theoretic Approach to Parallel vs. Serial Processing Capability

Following Feng et al. (2014), we consider single-layered, feedforward networks with N input and N output layer components. Each component represents an input or output dimension (vector subspace), and the connection from an input to an output component constitutes the processing pathway for a given task (defined as a unique mapping from all possible vectors in input subspace to corresponding vectors in the output subspace, that is independent of the mappings for all other combinations of input and output components in the network). The network can be represented as a directed bipartite graph $\mathcal{G}_B = (\mathcal{V}, \mathcal{E})$, in which the node set $\mathcal{V}$ can be partitioned into two disjoint sets of nodes $\mathcal{V}_{in}$ and $\mathcal{V}_{out}$, representing the input and output layer components respectively. Moreover, an edge $(i, j) \in \mathcal{E} \subseteq \mathcal{V}_{in} \times \mathcal{V}_{out}$ represents a directed pathway from the input layer to the output layer in the network (i.e., a task). We introduce the matrix $A = [a_{ij}] \in \{0,1\}^{N \times N}$ to represent the network structure and define its elements such that $a_{ij} = 1$ when, $(i, j) \in \mathcal{E}$, $i \in \mathcal{V}_{in}$, $j \in \mathcal{V}_{out}$ and $a_{ij} = 0$ otherwise.

Pathway Overlap and Interference

The matrix A, extracted from the adjacency matrix of the bipartite graph, captures the overall network structure, since by definition the graph is directed and has no self-loops. In particular, it represents the degree to which pathways overlap (i.e., share representations): the sum of each row of matrix A reflects the multiuse of input representations (out-degree of input nodes), the sum of each column reflects the same for output representations (in-degree of output nodes), and together these indicate the degree of pathway overlap in the network. We assume that such overlap produces interference, prohibiting performance of the tasks involved. We formalize three types of interference, as shown in Fig 1. Convergent interference (shown in green) occurs when two sources of input compete to determine a common output. In addition, we consider divergence (shown in red) as a form of interference in our analysis. Although this does not pose an impediment to performance (i.e., it is possible to generate two distinct responses to the same input), it represents a restriction on the number of independent sources of input (and therefore number of tasks) that the system can process at once, and thus can be treated formally as a type of interference in our analysis of multitasking capability. Finally, we consider a third, indirect form of interference that supervenes on the first two (shown in blue). In this case, the two tasks in question do not directly interfere with one another. However, their simultaneous engagement would necessarily engage a third task (shown in purple) that would produce interference; accordingly the two tasks shown in blue can not be performed simultaneously. It is important to note that not only the *amount* of interference (of the forms just described), but also how it is distributed over the network impacts multitasking performance. Here, for simplicity, we assume a uniform distribution of pathways among the input and output components[2], which means the pathway overlap P is equal to the in-degree and out-degree of each component in the network.

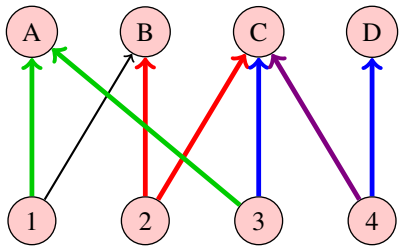
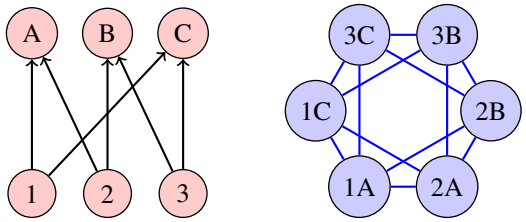


Figure 1: Illustration of the three types of interference considered in our analysis (see *text*).

To quantify multitasking capability, we begin by constructing an interference graph $\mathcal{G}_I$ associated with the original bipartite graph $\mathcal{G}_B$. By assigning each edge of the original graph $\mathcal{G}_B$ to a node in $\mathcal{G}_I$, each node in the $\mathcal{G}_I$ is used to represent a task. Interference between tasks is then represented by assigning edges to pairs of nodes in $\mathcal{G}_I$ if the tasks represented by those nodes are subject to any of the three forms of interference defined above (formally, this corresponds to the square of the line graph of $\mathcal{G}_B$). The adjacency relationships between nodes in $\mathcal{G}_I$ thus describe which tasks in the original network can be executed concurrently (i.e., in parallel). This, in turn, can be used to identify the maximum multitasking capability of the network, as discussed in the next section.

[2]Such a uniform distribution is also reflective of a relatively broad range of distributions in constraining multitasking.



(*a*)Bipartite Graph - $\mathcal{G}_B$ (*b*)Interference Graph - $\mathcal{G}_I$

Figure 2: The first subfigure (2a) illustrates a network of size $N = 3$ and pathway overlap $P = 2$ (in-degree and out-degree of each component is 2). The second subfigure (2b) shows the associated interference graph with 6 nodes, and each of these nodes has 4 neighbors due to interferences.

Maximum Independent Set (MIS) as a Measure of Multitasking Capability

Identifying the multitasking (maximum parallel processing) capability in the original network can be cast as the problem of finding the largest set of nodes in the interference graph wherein no two nodes are adjacent. This is formally known as the *maximum independent set* (MIS). Finding the MIS of a graph is an NP-hard problem, that has been studied extensively in the graph theory literature (Tarjan & Trojanowski, 1977). Figure 3 summarizes the effect of pathway overlap and network size on the MIS for networks with uniform pathway distribution (comparable to Feng et al., 2014), confirming that parallel processing capability is severely constrained by pathway overlap in a manner that is virtually scale invariant for network size (source code available at `github.com/musslick/CogSci-2016`). In the sections that follow, we show how these analysis tools can be used to infer the particular parallel processing capabilities of specific networks, validate predictions made based on extracted interference graphs in simulations of network multitasking performance, and describe how these tools could be used to infer similar information regarding human performance from neuroimaging data.

Application to Neural Network Models

In the previous section we introduced graph-theoretic analyses to investigate factors affecting the parallel processing capability in simplified network structures. Here, we examine the extent to which these analyses can be applied to more complex models (such as artificial neural networks) and empirical data (e.g. neuroimaging analyses). We will describe how neural representations of tasks can be used to generate predictions about how many and which combinations of tasks a network (or person) can perform in parallel (a space of possibilities that grows combinatorially with the number of tasks, and thus quickly becomes intractable to direct empirical inquiry), based on measurements of single task performance (that grows only linearly in the number of tasks). These analyses may provide useful diagnostic tools for exhaustively assessing multitasking capabilities based on amounts of data that are practical to acquire.

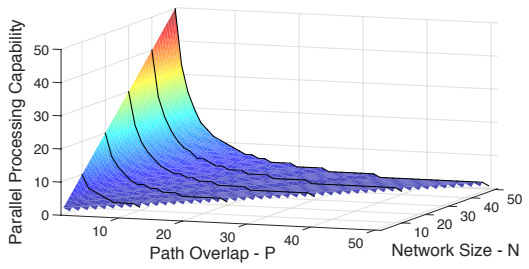


Figure 3: Network parallel processing capability as a function of pathway overlap (P) and network size (N) for networks with uniform pathway distributions.

Network Architecture and Processing

We focus on a network architecture that has been used to simulate a wide array of empirical findings concerning human performance (e.g. Cohen et al., 1990; Cohen, Servan-Schreiber, & McClelland, 1992; Botvinick et al., 2001). Such networks typically consist of four layers (see Figure 4): an input layer with two partitions, one of which represents the current stimulus and projects to an associative layer, and another that encodes the current task and projects to both the associative and output layers; an associative (hidden) layer that projects to the output layer; and an output layer that represents the network's response. Input units are clamped to either 0 or 1 to represent the current stimulus and task. These values are multiplied by the matrix of connection weights from the input layers to the associative layer, and then passed through a logistic function to determine the pattern of activity over the units in the associative layer. This pattern is then used (together with projections from the task units in the input layer) to determine the pattern of activity over the output layer. The latter provides a response pattern that is evaluated by computing its mean squared error (MSE) with respect to the correct (task-determined) output pattern.

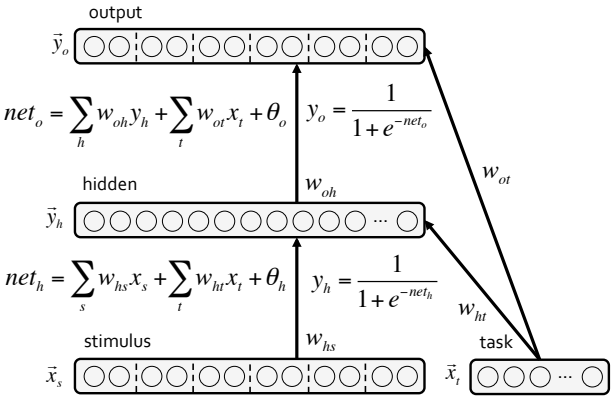


Figure 4: Feedforward neural network used in simulations. The input layer is composed of stimulus vector $\overrightarrow{x_s}$ and task vector $\overrightarrow{x_t}$. The activity of each element in the associative layer $y_h \in \overrightarrow{y_h}$ is determined by all elements x_s and x_t and their respective weights w_{hs} and w_{ht} to y_h. Similarly, the activity of each output unit $y_o \in \overrightarrow{y_o}$ is determined by all elements y_h and x_t and their respective weights w_{oh} and w_{ot} to y_o. A bias of $\theta = -2$ is added to the net input of all units y_h and y_o.

Stimulus input units are structured according to dimensions (subvectors of the stimulus pattern), each of which is comprised of a set of feature units with only one feature unit activated per dimension. Similarly, output units are organized into response dimensions, with only one response unit permitted to be active per dimension. Each task is represented by a single task input unit that is associated with a set of unique, one-to-one mappings between the input units in one stimulus dimension and the output units in one response dimension, and that is independent of the mappings for all other tasks. Here, we focus on networks ($N = 6$) in which there are six input dimensions comprised of two features each, and six output dimensions comprised of two responses each. Such networks support a total of $6 * 6 = 36$ possible tasks; and, since each stimulus input dimension consists of two features, $2^6 = 64$ possible input patterns per task (including both task-relevant and task-irrelevant features). The number of hidden layer units in each network is set to 200 to avoid constraining the network to low-dimensional mappings of the input space. Networks are initialized with a set of small random weights and then trained using the backpropagation algorithm (David E. Rumelhart & Williams, 1986) to produce the task-specified response for all stimuli in each task. That is, the network is trained to generate the response for the corresponding stimulus in the task-relevant dimension, while suppressing responses in all other response dimensions.

Extracting Directed Bipartite Graph from Task Representations

Our analysis focuses on the representations (patterns of activity) over the associative and output units, insofar as these reflect the computations carried out by the network required to perform each task. In particular, we are interested in the characteristics of these representations for each task, how they compare across tasks, and how these factors correspond to multitasking performance. The representations associated with each task can be characterized by calculating, for each unit in the associative and output layers, the mean of its activity over all of the stimuli for a given task; this mean pattern of activity can then be used as a representation of the task[3]. Correlating these patterns of activity across tasks yields a task similarity matrix that can be examined separately for the associative and output layers of the network. This can then be used to assess the extent to which different tasks rely on similar or different representation within each layer of the network. Figure 5 provides an example of such similarity matrices (thresholded for similarity correlations above $r > 0.8$). Tasks that have similar representations over the associative layer can be inferred to rely on the same input dimension – that is, they share an input component in the bipartite graph representation of the network – and tasks that are similar at the output layer can be inferred to share an output component.

[3]A formally equivalent analysis could be carried out using the weight matrix of the network. Here we focus on patterns of activity, as these may serve as useful predictors for patterns of activity observed in empirical data, such as fMRI and/or unit recordings.

Accordingly, a bipartite graph (of the type shown in Figure 3) can be constructed by measuring the patterns of activity observed in the network while it performs each individual task. This can then be analyzed, using the graph-theoretic methods described above, to examine the full multitasking profile of the network – that is, both the maximum concurrency (parallel processing) capability of the network and, perhaps more interestingly, the exact profile of which combinations of tasks can and cannot be performed concurrently (see Figure 6). This procedure is substantially more efficient, and scales more gracefully (linearly with size of the network) than determining the multitasking profile by simulating and examining performance of the network for all combinations of tasks (which scales factorially).

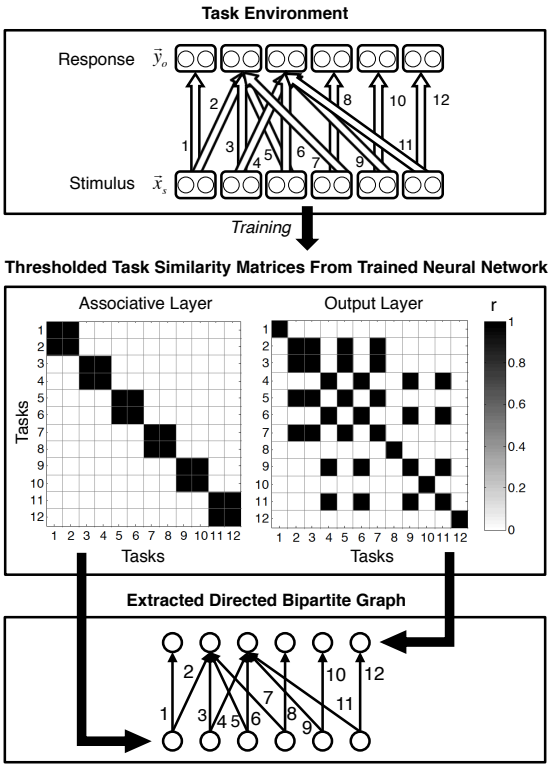


Figure 5: A task environment consists of 12 possible tasks represented as input-output mappings. The bipartite graph can be extracted from thresholded task similarity matrices that are obtained from task activity correlations at the associative layer and output layer of a trained network.

Simulation Experiment

To validate the methods described above, we compared simulated network performance with analysis predictions for 100 networks of size $N = 6$, each trained on a different subset of 12 randomly sampled tasks (source code available at `github.com/musslick/CogSci-2016`). Tasks were chosen subject to the constraint that each stimulus dimension was associated with two tasks. For each network we extracted a bipartite graph from the task similarity matrices as outlined above. Figure 5 shows the results for an example network, from which a bipartite graph was generated that recovered

the exact task structure imposed during training. That is, the network learned to use similar associative layer representations for tasks that involved the same stimulus dimension (e.g. tasks 1 & 2 in Figure 5), and learned similar output representations for tasks involving the same response dimensions (e.g. tasks 2, 3, 5 & 7 in Figure 5).

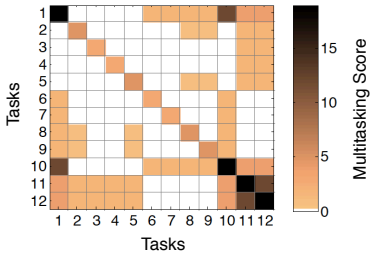


Figure 6: Extracted adjacency matrix of interference graph. Off-diagonal colored entries indicate all tasks that can be paired with a given task. Colors of off-diagonal elements indicate the number of all possible multitasking conditions in which the corresponding task-pairing can occur. Colors of diagonal elements indicate the number of all multitasking combinations in which the corresponding task can occur.

For each network, a bipartite graph can be constructed and used to extract a corresponding interference graph. The adjacency matrix of the latter indicates which tasks can be paired. This can be used, in turn, to identify all combinations of tasks that can be performed concurrently (see example in Figure 6), as well as the MIS which specifies the greatest number of tasks that can be performed concurrently. For each of the 100 trained networks, we extracted its interference graph and computed the multitasking performance (MSE) for all combinations of tasks that belonged to an independent set. We compared this to a random sample (of identical size) composed of combinations of tasks in which two or more of the tasks did not belong to an independent set. Figure 7 shows that these analyses yield accurate predictions about which tasks can be performed concurrently and which cannot. As the network was never trained on multitasking, concurrent performance of tasks from independent sets can still produce some error. However, the performance for those tasks is markedly and reliably better than multitasking performance for combinations of tasks in which not all belong to the same independent set, $t(98) = 232.34, p < .001$ (2 tasks); $t(98) = 132.03, p < .001$ (3 tasks); $t(79) = 29.64, p < .001$ (4 tasks).

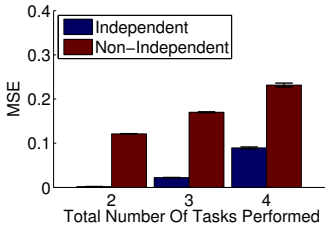


Figure 7: Multitasking performance for sets of identified independent and non-independent tasks. Error bars indicate the standard error of the mean for multitasking conditions of networks trained in different task environments.

General Discussion and Conclusion

We have introduced a graph-theoretic approach to compute the multitasking (parallel processing) capability of feedforward, single-layer non-linear networks. This was achieved by generating an interference graph from the directed bipartite graph representation of the network, which provides a compact representation of its multitasking capabilities. Identifying the MIS in the interference graph reveals the maximum number of concurrent tasks that can be executed without performance loss. The interference graph can also be used to identify all combinations of tasks that can be performed in parallel. We have shown that, consistent with previous work (Feng et al., 2014), introducing even modest amounts of pathway overlap induces dramatic constraints on multitasking capability. We then illustrated how the graph-theoretic analysis can be applied to specific networks, using the patterns of activity associated with individual tasks to characterize the full profile of the network's multitasking capabilities.

At a practical level, these methods suggest possibilities for empirical research. For example, if patterns of neural activity (measured using direct neuronal recordings and/or fMRI) can be identified for a set of individual tasks, then the analyses described above can be used to predict multitasking performance for all possible combinations of tasks in the set. The measurements required to carry out this analysis grow linearly with the number of tasks in the set, whereas the number of measurements required to characterize the interactions among them from behavior would grow factorially with set size. That is, these analyses may be particularly useful in situations in which exhaustively assessing the entire space of task combinations is empirically impractical.

Theoretically, the approach provides a formal framework for studying the relationship between learned task representations and controlled (serial) vs. automatic (parallel) processing. Specifically, it permits quantitative analysis of how task environment and learning impact the tradeoff between compactness of representation (associated with serial, control-dependent processing) and multitasking capability (associated with parallel, automatic processing). In related work others (e.g. Caruana, 1997; Bengio, Courville, & Vincent, 2013; Saxe, McClelland, & Ganguli, 2013) have shown that compact, "multiuse" representations not only make more efficient use of network resources (e.g. fewer associative units) but also are likely to arise most quickly (especially in hierarchically structured environments), and support greater generalization during learning. However, the present work illustrates the costs this incurs with regard to parallelism of processing: as pathway overlap (mulituse of representations) increases, processing in the network is rapidly driven to be serial, and becomes reliant on control mechanisms to avoid cross-talk. We suggest that this tension underlies the tradeoff between controlled and automatic processing observed in human performance, and that constraints on the capacity for human multitasking reflect this tension.

While we have focused on forms of multitasking arising

from concurrent parallelism, our findings are likely to have implications even when multitasking is achieved by rapidly switching between tasks. One of the most robust findings in the cognitive literature is the cost associated with switching between tasks, reflecting at least in part carry-over effects that the representations of one task have on the next (Yeung, Nystrom, Aronson, & Cohen, 2006, for a review see Kiesel et al., 2010). To the extent that such carry-over effects reflect interference from shared representations, then this may determine the speed and/or accuracy with which sequential switching can be achieved in a manner similar to its impact on pure parallelism. That is, multiuse of representations may define a continuum from pure sequential processing, though rapid task switching, to pure parallelism, and the methods we describe may provide a way of analyzing networks to determine where they lie along this continuum.

There are a variety of network parameters that can impact the extent to which multiuse representations arise during training (including weight initialization, regularization constraints, number of hidden units, etc.). Here, we have focused on a simple network parameterization (two-layered, feedforward, with random weight initialization and no regularization) as a first assessment of the usefulness of the analytic approach. Another simplification in our treatment was the construction of binary bipartite and interference graphs, by thresholding the real-valued correlation matrix of network representations. Additional simulation results (not reported) suggest that the analysis methods we report are robust across a wide range of thresholds and learned task representations. However, a generalization of the method to address graded interference effects (e.g., using weighted graphs) is an important avenue for future research. More generally, it will be important to explore the extent to which the methods and analyses we describe can be extended to networks with more complex and realistic architectures (e.g., multi-layered and/or recurrent, with varying pathway distributions and graded degrees of interference). We hope that the work described here will encourage a proliferation of efforts along these lines.

References

Allport, D. A. (1980). Attention and performance. *Cognitive psychology: New directions*, *1*, 12–153.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(8), 1798–1828.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological review*, *97*(3), 332.

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *The American journal of psychology*, 239–269.

David E. Rumelhart, G. E. H., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking vs. Multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 129-146.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switchinga review. *Psychological bulletin*, *136*(5), 849.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *The Behavioral and brain sciences*, *36*(6), 661–679.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In *37th cognitive science society conference, tx.*

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms. *Psychological review*, *104*(1), 3.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, *86*(3), 214.

Posner, M., & Snyder, C. (1975). attention and cognitive control,. In *Information processing and cognition: The loyola symposium* (pp. 55–85).

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychological review*, *115*(1), 101.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. In *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 1271–1276).

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127.

Tarjan, R. E., & Trojanowski, A. E. (1977). Finding a Maximum Independent Set. *SIAM Journal on Computing*, *6*(3), 537-546. doi: 10.1137/0206038

Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-task competition and cognitive control in task switching. *The Journal of Neuroscience*, *26*(5), 1429–1438.

Fractal Scaling and Implicit Bias: A Conceptual Replication of Correll (2008)

Mary Jean Amon (amonmj@mail.uc.edu)
University of Cincinnati, Department of Psychology
PO Box 210376, Cincinnati, OH 45221-0376

John G. Holden (john.holden@uc.edu)
University of Cincinnati, Department of Psychology
PO Box 210376, Cincinnati, OH 45221-0376

Abstract

A racial priming article claimed that, relative to a control condition, an exotic variety of variability, called 1/*f* noise, is altered when stereotypes impact participants' judgments in an implicit prejudice task (Correll, 2008). However, Madurski and LeBel (2014) recently described two powerful, faithfully cloned, and apparently decisive studies that each failed to return a successful *literal* replication of Correll's report. Madurski and LeBel outlined and subsequently eliminated several potential extraneous reasons for their replication failures, such as different participant demographics, participant non-compliance, poor psychometrics, and hardware discrepancies. By contrast, this article reports a successful *conceptual* replication of the pattern reported by Correll (cf. Schmidt, 2009). Notably, this conceptual replication required adjustments to Correll's original method and statistical analyses. All the changes were dictated by a systems theory of 1/*f* noise that was largely in place prior to Correll's report (Kello, Beltz, Holden, & Van Orden, 2007; Van Orden, Holden, & Turvey, 2003; 2005). Implications for the replication debate are discussed, with emphasis on contextualizing implicit cues.

Keywords: 1/*f* noise; prejudice; response time; replication; complexity science

A Conceptual Replication of Correll (2008)

Correll's (2008) implicit prejudice paradigm was modeled after a previously published weapon-identification task (Payne, 2001). It used racial priming to contrast automatic and controlled cognitive processes associated with stereotype activation. The task first presented participants with a racial prime (a photograph of a Black or White face) and then replaced the prime with either a stereotype-relevant or -irrelevant target (a gun or a tool). The task recorded response accuracy and response time as participants attempted to quickly identify targets by pressing either a "gun" or a "tool" button on successive trials.

Presumably, common racial biases might account for a participant's tendency to mistakenly identify a tool as a gun, subsequent to the presentation of a Black face as a racial prime. If so, participants' responses are thought to reflect either automaticity—perhaps driven by racial stereotypes—or the deliberate avoidance or invocation of stereotypes (Payne, 2001). That being said, issues persist regarding the generality and reliability of stereotype automaticity effects (e.g., Müller & Rothermund, 2014; Cesario & Jonas, 2014).

Correll (2008) extended Payne's statistical analysis by including a test for 1/*f* noise on each individual participant's successive series of response times—a trial-series—by analogy to a time-series. 1/*f* noise, also known as "pink noise," is a distinctive pattern of long-range correlation in successive measurements, taken more or less contiguously in time. The surprising aspect of pink noise is that it is a statistical fractal. It is comprised of proportionally nested, statistically self-similar patterns of fluctuation. This phenomenon is in stark conflict with the common Gaussian statistical intuition that uncontrolled variability lacks systematic structure.

Uncertain or variable task demands can perturb scaling, as participants attempt to accommodate competing performance goals. As such, task uncertainty forces unanticipated adjustments in the coordinative activity supporting responses in individual trials. Since they are unsystematic—a source of random variability, or *white* noise—their impact is to *whiten*, or weaken an observed scaling relation, relative to a baseline condition that presents more predictable events (Holden, Choi, Amazeen, & Van Orden, 2011; Kello, et al., 2007; Van Orden et al, 2003; Van Orden, Kello, & Holden, 2010; Van Orden, 2009).

Task uncertainty also *obscures* scaling. Fourier decompositions assume completely regular temporal sampling intervals. Unsystematic shifts in the trial-by-trial pace of the experiment introduce apparent but spurious shifts in the frequencies, amplitudes, and phases of the set of ideal sinusoidal functions that are used to decompose the complex empirical waveform. Other things being equal, a response time spectral scaling analysis that relies on unsystematic sampling intervals will be whitened approximately in proportion to the amplitude of the variability in the sampling rate (Holden et al., 2011).

Scaling in computer controlled response time studies arises at the interface of *extrinsic* cognitive and physiological variability and *intrinsic* sources of contextual variability introduced by the event cycle and demands of the task. One implication of this fundamentally adaptive and multi-scale reciprocal coordination is that the effects of laboratory manipulations are causally intertwined with this ongoing coordinative activity (Holden et al., 2011; Van Orden et al., 2010).

Correll's key manipulation gave differential instructions to participants to introduce competing task requirements, relative to a baseline condition. Participants were instructed

to complete a weapon identification task while either deliberately using or explicitly avoiding use of racial stereotypes relevant to the priming photographs. Target identities and the depicted race primes were manipulated orthogonally. Thus, attending to race provides a potential to decrement performance. Attempting to explicitly ignore race also invokes a racial frame of reference, which also may decrement performance in a "*don't think of an elephant*" manner (e.g., Lakoff, 2005; Wegner, 1989). Instructing participants to attend to a functionally irrelevant task dimension makes the task just a bit more difficult than it would be otherwise. As such, by the uncertainty hypothesis, invoking and avoiding race is predicted to yield weaker scaling, more similar to white noise than a baseline condition.

Our review of the methodology used in both the original and replication studies suggested several potential methodological artifacts might obscure the impact of the racial priming manipulation on $1/f$ scaling. These artifacts could render the task less sensitive to perturbations induced by the manipulation itself. From the perspective of a widely agreed upon definition of a literal replication—do the same thing twice, and get the same result—the two replication failures of Madurski & LeBel (2014) were clear and decisive. Yet, as we explained, the impact of task uncertainty suggests the pattern of change predicted and reported by Correll is nevertheless plausible. We considered several potentially confounding methodological issues:

1. Both the original and replication studies presented only 200 trials. This is not enough observations to establish compelling scaling relations in either study. Generally, one seeks to establish $1/f$ scaling across at least two orders of magnitude of frequency. Technically, 200 trials appear to meet this criterion (i.e., 2, 20, 200 trials). However, to both comply with this rule of thumb, and to accommodate several major statistical pitfalls of spectral and scaling analyses, collecting at least 1,024 valid trials is strongly recommended (Eke, Herman, Kocsis, & Kozak, 2002; Holden, 2005). *Thus, we presented 1100 trials to participants in every condition of both our literal and conceptual replication attempts*.

2. The response boxes used by Correll were accurate to the nearest millisecond (ms). The response keyboards used to collect the replication data sets were accurate to ±7.5 ms. This difference in precision is inconsequential in studies that pursue differences in mean response time. However, it represents a relatively large amplitude source of unsystematic variability across trials. This added variability is capable of obscuring scaling differences (Holden et al., 2011). By itself, this issue *cannot* explain the replication failure. It is notable, however, that none of the replication control conditions yielded spectral exponents as large as those depicted in Correll's baseline condition. *Both our literal and conceptual replications used ms accurate keyboards, and adopted a symmetric response layout, in which the left-hand 'z' and right-hand '/' keys were mapped to the tool and gun responses, respectively*.

3. The procedure used to compute the power spectra in the original and replication studies is inconsistent with those commonly used in the $1/f$ scaling literature. The response times were log transformed, presumably to obviate the need for outlier censoring. A log transform has little discernable impact on the power spectrum of compact and symmetric Gaussian variables. However, response time distributions entail a potent positive skew. Informal contrasts of untransformed versus log transforms of comparable previously published data sets indicated that a log transform shrank scaling exponents by about 12%. This is problematic for standard subtractive statistical contrasts that use scaling exponents as dependent variables. As one approaches the floor of zero, scaling differences diminish. *We used previously established spectral techniques for all our analyses*.

4. The weapon identification response times are typically very fast (≈300 to 400 ms). As such, the relative time-course of the 1-second inter-trial interval is too long to reveal $1/f$ noise that is closely tied to the trials. The long inter-trial interval effectively shifts the minimum period of sampled change to about 2 sec. Moreover, a perceptibly long downtime between trials hampers the emergence of the close coordination between participant and task that reveals scaling. This uncertainty, in turn, becomes a source of unsystematic variability that both perturbs and obscures scaling—it impairs a task's ability to detect differences in scaling across contrasted cells. *We used a constant 500 ms inter-trial interval in our conceptual replication condition*.

As implemented, the Correll task is unlikely to be particularly sensitive to scaling changes. In light of the exact literal replication failure, we pursued both a literal and an optimized conceptual replication of the weapon identification task. We used four separate experimental cells. We increased the number of trials in all four cells from 200 to 1,100, and we used both ms accurate keyboards, along with a symmetric response-button layout to collect response time data in all cells.

Except for the response-button layout and increase in the number of trials, two cells used a method and event-timing identical to that described by Correll. We sought to optimize two additional experimental cells to detect a scaling difference with changes to ancillary methodological and procedural aspects of the original study. The goal was to use the uncertainty principle to guide improvements in the task's sensitivity to changes in $1/f$ scaling. The optimized cells reduced the inter-trial intervals from 1 sec to 500 ms and increased the presentation duration of the racial primes from 200 to 300 ms. In addition, an error message and beep that followed incorrect responses was removed in the optimized conditions, as these intermittent alerts perturb trial pacing and risk interrupting the coordination between the participant and task. The same experimenter explained the task to each participant in the optimized cells to reinforce the written instructions. Finally, we adopted spectral techniques modeled after those used previously to identify scaling differences response time studies.

Method

Participants A total of 128 undergraduate psychology students participated in exchange for course credit. They were recruited into a study described as investigating vigilance during social tasks. Participants ranged in age from 18 to 41, with a median age of 19. Sixty-five percent of participants were female, 34% were male, and 1% did not specify their gender. Eighty-three percent of participants identified as White, 7% as Black, 5% as Multiracial, 4% as Asian, and 1% as Hispanic. A White female graduate student greeted and consented participants individually. Once informed consent was obtained, participants were seated in front of a computer monitor. All research was carried out in accordance with the protocol approved by the University of Cincinnati's Institutional Review Board.

Design And Procedure Participants completed a two-option, forced choice response time task. Each trial began with the appearance of a *face prime*, drawn from a bank of five Black and five White male pictures. Next, either a hand-tool or handgun was shown and quickly masked by a series of black and white rectangles. The visual mask remained on the screen until the participant selected either the 'z' or the '/' key to identify the object as a tool or a weapon, respectively. All the pictures were presented in a black-and-white format. The task randomly interleaved 1,100 trials that balanced equal numbers of four trial identities: *Black Prime-Tool, Black Prime-Gun, White Prime-Tool,* and *White Prime-Gun*.

Participants were randomly assigned to one of four replication conditions. The first two cells mirrored Correll's control and avoid bias conditions, and maintained the use of a 1 sec inter-stimulus interval, a 200 ms prime presentation, and an error beep-and-message. In line with Correll's procedure, participants in the literal control and avoid bias conditions read instructions for the weapon-identification task off a computer screen.

The two remaining conditions implemented conceptually optimized versions of the control and avoid bias conditions. The inter-trial interval was reduced to 500 ms, and the prime duration was increased to 300 ms. Participants in the optimized control and avoid bias cells received both verbal and screen-printed instructions, in an attempt to maximize their potency. Participants in both versions of the avoid bias cells were informed that some people tend to respond more quickly and accurately to guns after a Black face than after a White face. They were directed to try to avoid racial bias when identifying objects. By contrast, instructions in both versions of the control cells did not raise the topic of race. Immediately following instructions, a single multiple-choice item was administered that asked participants in the optimized cells to indicate their primary goal while completing the computer task. Following their response, participants received feedback reiterating the instructions. The participant's multiple-choice responses were not recorded because of a programming error in the data collection script.

All participants completed 25 practice trials. During the practice trials, if no response occurred by 1 sec, a message to respond faster was displayed. The same warning was displayed after 1.5 sec in the two conceptual replication practice trials to allow participants to learn the task at their own pace. Participants required between 30 and 40 minutes to complete the literal replication cells and 20 to 30 minutes to complete the conceptual replication cells. Two self-report questions and a demographic form were administered once participant's completed the weapon-identification task. The self-report questions asked participants to rate task difficulty and their effort to avoid racial bias on a seven-point scale. Participants were then debriefed and thanked for their time and effort.

Results

The data sets for all consented participants were included in our subsequent statistical analyses. We adapted our statistical analysis for one participant that yielded an idiosyncratic response pattern. The participant produced an apparent 94% error rate (*d*-prime = -3.14), likely a consequence of reversing the keyboard response-mapping. We included this data, but assumed it represented a *d*-prime of 3.14 and a 6% error rate.

We adopted 1-tailed p equal to or less than .05 significance levels in all our statistical contrasts. We did this to accommodate the following facts: First, our uncertainty hypothesis makes a clear *a priori* directional prediction that scaling will be reduced, relative to controls, in the avoid bias cells. Increased scaling in the avoid bias condition would contradict the previously established theoretical narrative, the original Correll result, and the Madurski and LeBel (2014) replication failure. Thus, while it was a potential outcome, it would amount to a replication failure because it cannot be interpreted from any established perspective on the manipulation. Second, the previous studies suggest that if an effect is present, it is likely very weak. We used 128 participants, the integer nearest to the 126 participants recruited by Madurski and LeBel that is evenly divisible by 4. Our prediction was that the replication failure would hold in the literal conditions and be overturned in the optimized cells. As such, our key planned statistical contrast, between the two optimized cells, entailed a sample that was about 50% of the size required to yield a power level of .80 according to Madurski and LeBel (2014).

Scaling Analyses The data censoring procedures used to prepare each individual's response time trial-series for spectral analysis were modeled after the steps described in Holden (2005). Each 1,100 trial data set was sorted into the sequential order in which the trials were presented in the experiment. A two-stage procedure was then used to censor extreme observations from the resulting trial-series. First, responses less than 10 ms or greater than 2,000 ms were removed. Second, the series mean and standard deviation were computed, and observations that fell beyond ± 4 standard deviations from an individual's mean response

time were also censored. A common ± 3 standard deviation criterion was too restrictive, eliminating more than 75 observations from four trial-series, and at least 50 observations from five additional series. Censorship procedures insure that extreme values do not overwhelm the statistical procedures. However, one must use the most conservative and inclusive thresholds possible, since omitting too many observations dilutes the sequential patterns the fractal analyses seek to reveal (Holden, 2005). Errors were included in the analysis to preserve trial order, (e.g., Gilden, 1997). If more than 1,024 observations remained following the censoring procedures, enough initial trials were deleted to yield a trial-series that contained 1,024 observations, and the series was then transformed into normalized Z-scores. If less than 1,024 observations remained after censoring, the trial-series was normalized into Z-scores, and padded with zeros until it contained 1,024 observations. Zero padding does not impact the scaling exponent. A 127-frequency power spectrum was computed using the procedures described by Holden (2005). A spectral scaling exponent was computed for each participant's trials series. Spectral exponents were derived from the slope of a regression line, computed from the 50 lowest log-log frequency-power data-pairs (Wijnants, Cox, Hasselman, Bosman, & Van Orden, 2012). The analysis was implemented on MATLAB software.

The scaling exponents were subsequently used as dependent variables for two planned contrasts. A between subjects ANOVA contrasted the scaling exponents in the optimized control and optimized avoid bias conditions, testing the hypothesis that instructions to avoid racial bias resulted smaller average scaling exponents—spectral whitening. We observed a barely reliable difference in the predicted direction, $r^2 = .06$, $F(1, 62) = 3.88$, $p = .05$, opt. control $M = .39$ ($SD = .15$), opt. avoid $M = .31$ ($SD = .14$). As predicted, scaling exponents in the optimized avoid bias cell were on average slightly whitened relative to that of the optimized control cell.

The outcome is consistent with the uncertainty hypothesis, but the difference is hardly compelling when considered in isolation. However, the "effect" was sussed from the ashes of a convincing and powerful replication failure, and the core of the original manipulation was entirely preserved. All our methodological changes were designed to make both optimized cells more sensitive to $1/f$ scaling, resulting in a contrast that was more sensitive to scaling changes. In every case, our methodological and statistical adjustments were either derived from established practice in the scaling literature, or dictated by the emergent coordination and sampling theories that motivated the uncertainty hypothesis. As expected, an identical contrast of the literal replication control and avoid bias conditions failed to reveal a reliable differences in the spectral scaling exponents, $p > .05$.

Figure 1 depicts the power spectra of the optimized control and optimized avoid bias conditions. They illustrate the quadratic shape typical of a power spectrum representative of a mixture of pink and white noise (Holden et al., 2011; Gilden, 1997; Thornton & Gilden, 2005). The optimized avoid bias spectrum diverges from the other spectra as a function decreasing frequency. However, this raises a question: *Is the sole basis of the observed effect an idiosyncratic difference in the lowest-frequency bands*?

This question is best addressed by an alternative statistical analysis called standardized dispersion analysis (SDA analysis, Bassingthwaighte, Liebovitch, & West, 1994; Holden, 2005). SDA is derived from the central limit theorem and is more sensitive to scaling differences in the higher-frequency range, where artifactual whitening introduces a *quadratic* trend that biases spectrum-based scaling exponents. (Thornton & Gilden, 2005; Holden et al., 2011). However, SDA itself is easily biased by low-frequency trends in a trial-series (Van Orden et al., 2003; 2005). For this reason, low-frequency trends are simply removed from each trial-series in advance of the analysis with least-squares linear de-trending.

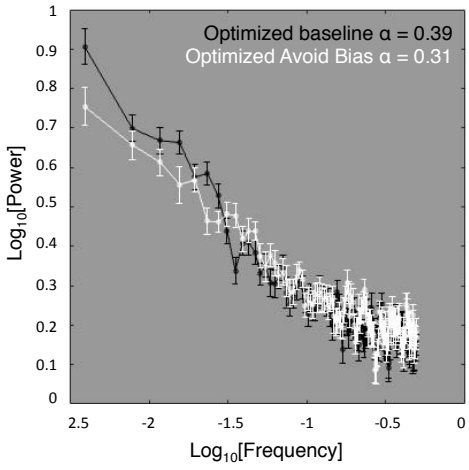


Figure 1: Power spectra of the optimized control and optimized avoid bias conditions averaged across participants. The black markers and lines correspond to the optimized baseline condition, the white markers and lines correspond to the optimized avoid bias condition. The whiskers indicate 1 SEM for each spectral frequency coefficient. The power spectrum of the optimized avoid bias is reliably whitened, relative to the baseline spectrum. The same contrast of the literal replication cells did not reveal a reliable difference.

The SDA analysis yields a fractal dimension statistic (*FD*) that is analogous to a spectral scaling exponent, but derived from a different statistical framework. An *FD* value that is statistically equivalent to 1.5 indicates white noise, values less than 1.5 but not greater than about 1.2 are symptomatic of pink noise. It is notable that, when viewed through the lens of SDA, the control and avoid bias conditions mirror each other in the literal and optimized versions of the replication (see Figure 2). SDA analysis does not rely on sinusoidal functions as a mathematical basis, and it is less susceptible to distortions resulting from irregular

sampling in time. Neither of the two-cell literal and conceptual replication contrasts yielded a reliable difference. A 2 (task version) × 2 (instruction) ANOVA analyzing FD values based on task version and instruction set revealed a reliable instruction effect, $r^2 = .03$, $F(1, 124) = 3.99$ $p = .05$. The FD measurements distinguished the control and avoid bias instructions. The main effect of task version and the task × instruction interaction were non-significant, $p > .05$. Once again, the difference teeters on the margin of traditional statistical significance. We did not anticipate completing the SDA analysis, or this specific contrast, prior to conducting our replication attempt. We implemented it as an *ad-hoc* check for evidence of scaling differences that were not exclusively contingent on the low-frequency differences established with spectral methods. Our overall conclusion is the race manipulation is relatively weak, but the observed scaling differences are just sufficient to reasonably claim a successful conceptual replication.

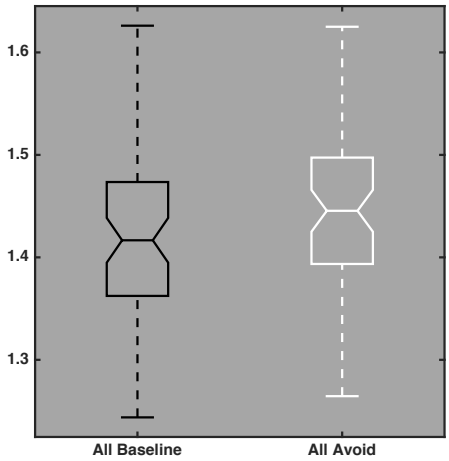


Figure 2: Boxplots of the FD statistics, used to detect differences in scaling. The baseline and avoid bias conditions were collapsed across replication type. SDA is not as sensitive to artifacts in the high frequency range as spectral analysis. A fractal dimension that is statistically equivalent to 1.5 indicates white noise. Fractal dimensions less than 1.5 but no less than about 1.2 indicate pink noise. The FD statistic was reliably larger in the avoid bias condition indicating whitened response time trial-series, relative the baseline condition.

The overall error rate for the different trial types was relatively low, 6.97% ($SD = 6.32$), a pattern consistent with Payne (2001), control $M = 5.20\%$ ($SD = 3.48$), opt. control $M = 9.61\%$ ($SD = 6.30$), avoid $M = 3.69\%$ ($SD = 2.44$), opt. avoid $M = 9.37\%$ ($SD = 8.15$)]. A *d*-prime statistic measures decision sensitivity. In the weapon identification task, this translates to the ability to both accurately identify weapons and tools, and to refrain from falsely identifying a tool for as a weapon and vise versa. A 2 (Type) × 2 (Instructions) between-subjects ANOVA compared *d*-prime as a function of replication type and bias instructions. There was a main effect of replication type, $r^2 = .19$, $F(1, 124) = 29.30$, $p < .05$, $M_{\text{Literal}} = 3.64$ ($SD = .74$), $M_{\text{Conceptual}} = 2.87$, ($SD = .87$), and no interaction or main effect of bias instructions. The cells in the literal replication yielded larger average *d*-prime values than the conceptual cells. This outcome is unsurprising since the literal replication cells included error feedback and the conceptual replication cells did not. Tests for a Race × Object interaction were non-significant for reaction times and error rates, $p > .05$.

Discussion

We managed to replicate the basic pattern of scaling changes predicted by Correll (2008). To do so, we changed a number of methodological details of the task protocol that were not closely linked to the issue of racial bias. Overall, the impact of the racial primes on participant's decision performance was somewhat weak. In fact, other than scaling changes, we found little compelling evidence for other effects that are typically associated with implicit racial priming in the weapon identification task, such those previously reported in mean response time and error rates. Of course, the task protocol was optimized to detect scaling changes, not differences in error rates or mean effects. In this case, the uncertainty hypothesis dictated methodological changes that overturned two compelling replication failures. It illustrates a strongly counterintuitive success of the emergent coordination framework.

Apparently, methodological details of the task events played a crucial role in the outcome of the study. These details were largely ancillary to the issue of racial priming. On one hand, the trial-series analyses tracked the principal instructional manipulation to avoid racial bias. On the other hand, conventional analyses failed to corroborate a role for implicit racial priming in errors or mean response times. As such, the basis for the difference is difficult to pin down. One possibility is that the scaling changes simply tracked the difference between verbal and written instructions, but that would not explain how the SDA analysis detected the same scaling difference when both the literal and optimized data were aggregated, nor does it explain the difference observed in the optimized cells. The only methodological detail that tracks the observed scaling patterns is the presence or absence of race in the instructions.

More generally, a counterintuitive implication of success of the optimized cells is the possibility that almost no factors are truly benign in experiments designed to isolate the influence of specific factors by holding all others constant. In fact, we take seriously the prospect that all psychological effects are context dependent. Similar paradoxes illustrate why some scientists take lessons learned from disciplines of quantum physics and nonlinear dynamics seriously (e.g., Atmanspacher, Römer, & Walach, 2002; Flach, Dekker, & Stappers, 2007; Gabora & Aerts, 2002; Wang, Solloway, Shiffrin, & Busemeyer, 2014). If context dependency is the norm, it has important implications for the replication crisis and the discipline at large.

Notably, a successful conceptual replication provides more and not fewer potential scientific stances on the effect of implicit bias. One might reasonably conclude the effect is so weak, and likely unreliable, that it is not worthy of inclusion in scientific discourse. On the other hand, while its impact is weak, it is theoretically aligned with other more powerful uncertainty-based scaling manipulations, and it lends credibility to a more general theoretical framework that serves the goals of a scientific enterprise.

References

Atmanspacher, H., Römer, H. & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics, 32,* 379-406. doi: 10.1063/1.2158709

Bassingthwaighte, J. B., Liebovitch, L. & West, B. J. (1994). *Fractal physiology*. New York: Oxford University Press.

Cesario, J., & Jonas, K. J. (2014). Replicability and models of priming: What a resource computation framework can tell us about expectations of replicability. *Social Cognition, 32,* 124-136.

Correll, J. (2008). 1/*f* noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology, 94,* 48-59. doi: 10.1037/0022-3514.94.1.48

Eke, A., Herman, P., Kocsis, L., & Kozak, L. R. (2002). Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement, 23*, R1. doi: 10.1088/0967-3334/23/1/201

Flach, J. M., Dekker, S., & Stappers, P. J. (2007). Playing twenty questions with nature (the surprise version): Reflections on the dynamics of experience. *Theoretical Issues in Ergonomic Science, 9*, 125–154.

Gabora, L., & Aerts, D. (2002). Contextualizing concepts using a mathematical generalization of the Quantum formalism. *Journal of Experimental and Theoretical Artificial Intelligence, 14*, 327-358. doi: 10.1080/09528130210162253

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science, 8,* 296-301. doi: 10.1111/j.1467-9280.1997.tb00441.x

Holden, J. G., (2005). Gauging the fractal dimension of cognitive performance. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. Retrieved from http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

Holden, J. G., Choi, I., Amazeen, P. G., & Van Orden, G. (2011). Fractal 1*/f* dynamics suggest entanglement of measurement and human performance. *Journal of Experimental Psychology: Human Perception & Performance, 37*, 935-948. doi: 10.1037/a0020991

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General, 136,* 551-568. doi: 10.1037/0096-3445.136.4.551

Lakoff, G. (2005). *Don't think of an elephant! Know your values and frame the debate.* White River Junction, VT: Chelsea Green Publishing.

Madurski, C., & LeBel, E. P. (2014). Making sense of the noise: Replication difficulties of Correll's (2008) modulation of 1/*f* noise in a racial bias task. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-014-0757-4

Müller, F., & Rothermund, K. (2014). What does it take to activate stereotypes? Simple primes don't seem enough: A replication of stereotype activation (Banaji & Hardin, 1996; Blair & Banaji, 1996). *Social Psychology, 45*, 187-193. doi:10.1027/1864-9335/a000183

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181-192. doi: 10.1037//0022-3514.81.2.181

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90-100. doi: 10.1037/a0015108

Thornton, T.L. & Gilden, D.L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review, 12*, 409-441. doi: 10.3758/BF03193785

Van Orden, G. C. (2009). Voluntary performance. *Medicina, 46,* 581-594.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*, 331-350. doi: 10.1037/0096-3445.132.3.331

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/*f* scaling. *Journal of Experimental Psychology: General, 134*, 117-123. doi: 10.1037/0096-3445.134.1.117

Van Orden, G. C., Kello, C. T., & Holden, J. G. (2010). Situated behavior and the place of measurement in psychological theory, *Ecological Psychology, 22,* 24-43. doi: 10.1080/10407410903493145

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 9431-9436. doi: 10.1073/pnas.1407756111

Wegner, D. M. (1989). *White bears and other unwanted thoughts: Suppression, obsession, and the psychology of mental control*. New York: Viking/Penguin.

Wijnants, M. L., Cox, R. F. A., Hasselman, F., Bosman, A. M. T., & Van Orden, G. (2012). Does sample rate introduce an artifact in spectral analysis of continuous processes? *Frontiers in physiology*, *3*. doi: 10.3389/fphys.2012.00495

The Relationship Between the Numerical Distance Effect and Approximate Number System Acuity is Non-Linear

Dana Chesney (chesneyd@stjohns.edu)
Department of Psychology, St. John's University, 8000 Utopia Pkwy.,
Jamaica, NY 11439 USA

Abstract

People can estimate numerical quantities, like the number of grapes in a bunch, using the Approximate Number System (ANS). Individual differences in this ability (ANS acuity) are emerging as an important predictor in research areas ranging from math skills to judgment and decision making. One commonly used ANS acuity metric is the size of the Numerical Distance Effect (NDE): the amount of savings in RT or errors when distinguishing stimuli values as the numerical distance between them increases. However, the validity of this metric has recently been questioned. Here, we model the relationship between the NDE-size and ANS acuity. We demonstrate that the relationship between NDE-size and ANS acuity should not be linear, but rather should resemble an inverted J-shaped distribution, with the largest NDE-sizes typically being found for near average ANS acuities.

Keywords: Numerical Distance effect; Estimation; Approximate Number system

Introduction

People can evaluate non-symbolic numerical magnitudes (i.e., which pack has more wolves) without counting (Taves, 1941). Indeed, people can respond based on these perceived magnitudes without necessarily linking these values to symbolic numbers (Kaufman, Lord, Reese, & Volkmann, 1949). The system that makes these non-symbolic numerical magnitude evaluations is herein referred to as the Approximate Number System (ANS). Essentially, the ANS allows us to perceive numerical magnitudes from the world in an analog fashion, similarly to how we perceive other magnitudes, like size (Kaufman et al.). The facility of the ANS to make numerical magnitude judgements (ANS 'acuity') is thought to vary among individuals, such that some individuals can make faster and more accurate judgments with smaller magnitude ratios than other individuals (Halberda & Feigenson, 2008). Better ANS acuity has been linked to better math skills in elementary schools and better performance on standardized tests (Gilmore, McCarthy, & Spelke, 2010; Halberda et al., 2008). Recent work has even suggested that the ability to understand the magnitude of numbers, a skill at least partially grounded in ANS acuity, influences judgment and decision making in adults (Peters, et al., 2008; Schley & Peters, 2014). As such, it is important that researchers use valid and reliable measures to assess ANS acuity. Unfortunately, while assessments of mathematical skill are well understood, being similar to math tests one might take at school (e.g. Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Lipkus, Samsa, & Rimer, 2001; Weller, Dieckmann, Tusler, Mertz, Burns & Peters, 2013), metrics of individual differences in ANS acuity are less well investigated.

One widely used metric of ANS acuity is the 'size' of the Numerical Distance Effect (NDE): The difference in reaction time or accuracy between distinguishing values that are close to each other and distinguishing values that are more distant from each other. In a practice which seems to have which originated with Sekuler and Mierkiews (1977), researchers will measure the speed and accuracy of numerical comparisons (e.g. "Which is larger?") at smaller (harder) and larger (easier) distances. Bigger differences in speed and accuracy between these easier and harder trials are interpreted as evidence of poorer ANS acuity (see Price, Palmre, Battista, & Ansari, 2012 for discussion). Recently, NDE-size has come under fire, with several studies questioning both its reliability, and its ability to distinguish individual differences in ANS acuity (Gilmore, Attride, & Inglis, 2011; Holloway & Ansari, 2009; Inglis & Gilmore, 2014; Maloney, Risko, Preston, Ansari, & Fugelsang, 2010; also see Sasanguie, Defever, Van den Bussche, & Reynvoet, 2011). Given the extent of the use of NDE-size in past literature as a metric of ANS acuity and the growing collection of recent publications that questions its validity, we believe it critical to examine the theoretical support as to whether NDE-size has the potential to serve as a metric of ANS acuity. Here, we model the theoretical relationship between ANS acuity and NDE-size. To foreshadow our conclusions, we do not find evidence that NDE-size can serve as a general measure of ANS acuity.

Individual Differences, ANS Acuity, and the NDE

Since Halberda et al.'s (2008) seminal work retrospectively predicting elementary school math ability from ANS acuity in fourteen-year-olds, a flurry of studies have linked ANS acuity to mathematical and academic success in children, adolescents, and adults (see Chen & Li, 2014, for meta-analysis and review). This connection between ANS acuity and other numerical skills can be explained by a mapping between non-symbolic ANS magnitudes and symbolic numbers. Humans, unlike other animals, have resources beyond the ANS to help them evaluate numbers. People can represent numbers verbally (e.g., "three thousand") and with other symbols (e.g., 3000) and do so with much greater precision than the ANS can achieve: The symbolic number 3000 is distinguished from 2999, but one cannot perceive the magnitude difference between 3000 and 2999 grains of

rice. However, there is strong evidence that people invoke ANS-based analog magnitudes when considering symbolic numbers (Dehaene, Bossini, & Pascal, 1993; Moyer & Landaur, 1967). In Moyer and Landaur's (1967) seminal study, it was demonstrated that people show *distance effects* when making judgments about the quantities referenced by symbolic magnitudes. For example, people are faster at determining 6 is smaller than 9 than that 7 is smaller than 8. Such effects would result from neither use of symbolic look-up tables nor sequential-count based comparison processes, but (as discussed below) are a classic pattern in analog magnitude comparisons (Moyer & Landaur).

If symbolic number processing involves the ANS, one could therefore predict that performance on tasks involving symbolic numbers may be influenced by individual differences in ANS acuity. Moreover, as higher order mathematical skills build off one's understanding of symbolic numbers, one could further posit that these mathematical skills may be predicted by one's ANS acuity. Following this logic, some tasks used to assess individual differences in the ANS use symbolic numbers rather than non-symbolic numerical magnitudes (e.g., Holloway & Ansari, 2009; Seilger & Opfer, 2003; Sekuler & Mierkiews, 1977), even though symbolic numbers access ANS magnitudes only indirectly. Indeed, it appears that Moyer and Landaur's now classic approach of using the presence of distance effects to demonstrate that the ANS is invoked in symbolic magnitude comparisons inspired the later use of NDE-size as an ANS acuity metric (Sekuler & Mierkiews).

ANS Theory and NDE-Size

The exact nature of the ANS has yet to be completely determined, but it is well established that the perception of non-symbolic numerical magnitudes obeys Weber's law (Cordes, Gelman, Gallistel, & Whalen, 2001; Dehaene, Izard, Spelke, & Pica, 2008; Mechner, 1958; Meck & Church, 1983; Whalen, Gallistel, & Gelman, 1999). As is typically the case for magnitude perception (see Kingdom & Prins, 2009), numerical magnitudes are not perceived exactly, but rather percepts are normally or quasi-normally distributed around a mean value (which may itself be biased). The ability to distinguish between two quantities is dependent on the amount of overlap between their perceived magnitude distributions. Importantly, the overlap in the distributions of any two values - and thus the ease with which two values can be distinguished - is dependent upon their ratio, rather than upon the absolute distance between them. As a result, one can observe both *size effects* and *distance effects* in magnitude discriminations. 'Distance effects' refers to the observation that, within a given range, it is easier to distinguish numerical quantities that are more distant from each other (6 dots [::::] vs. 12 dots [::::::]) than those that are closer together (8 dots [::::] vs. 10 dots [::::::]). 'Size effects' refers to the observation that it is easier to distinguish numerical quantities at the same distance in smaller magnitude ranges (6 dots [::::] vs. 8 dots [::::]) than in larger magnitude ranges (14 dots [::::::::] vs. 16 dots [::::::::]). ANS magnitude comparisons yield standard psychophysical functions: The likelihood that an individual will successfully discriminate between two magnitudes will increase curvilinearly from chance to asymptote at or near 100% accuracy, as the ratio of the larger to the smaller value increases. RTs similarly decrease with the comparison ratio (Whalen et al., 1999, see Kingdom & Prins, 2009, for a discussion of psychophysical functions).

To put it simply, Weber's law implies that a) the standard deviation (SD) of the distribution around an estimated magnitude is proportional to that magnitude's mean (M) and b) this proportion is constant. This constant proportion is, by definition, the Weber Fraction (*w*) of the perceiver's ANS. Thus, ANS acuity is defined by an individual's 'Weber Fraction' (Cordes et al., 2001; Deheane, et al., 2008; Halberda, Mazzocco, & Feigenson, 2008, Seigler & Opfer, 2003; Whalen et al., 1999). After accounting for other biases, this *w* determines the variability in the representation of a particular magnitude, which in turn determines the amount of overlap between any two magnitudes, which finally determines how likely it is that an individual will be able to tell two non-symbolic magnitudes apart (ANS acuity). The smaller an individual's *w*, the better an individual will be at discriminating between non-symbolic numerical magnitudes, because there is less overlap in their numerical magnitude perceptions. Also, following Weber's Law, there is greater overlap in the magnitude distributions perceived from stimuli at smaller ratios (::::/:::, 1.33) than at larger ratios (::::::/:::, 2).

It follows that the ANS' contribution to an individual's NDE-size should be a function of the specific magnitudes being compared and the individual's ANS acuity (their Weber fraction, *w*). Thus, we can model the relative size of the ANS' contribution to NDE-size for any specific task and any given *w* by calculating the proportion of overlap of the distributions for any set of magnitude comparison pairs, and then calculating the savings in overlap for longer distances. For example, we can find the predicted ANS magnitude distribution overlap at short distances (4 vs. 5, 5 vs. 6) and at long distances (1 vs. 5, 5 vs. 9) for a given acuity level (*w*) and then find the differences between these overlaps. As long as judgments are based on ANS distributions, error rates and RTs should be functionally related to the amount of overlap in these distributions.

Model

As previously discussed, the ANS obeys Weber's Law. We follow the Linear Model of these phenomena here, which claims magnitudes perceived from stimuli are normally distributed, such that the means of perceived magnitude distributions increase linearly with the size of the stimuli and that the standard deviations of these distributions are proportional to their means (Cordes et al., 2001). Thus, the Coefficient of Variation (CV = SD/M) is constant for a given individual on a given task. This constant CV is the *w* of the individual's ANS: their ANS acuity. We model this here by representing perceived magnitudes as Gaussian

distributions about unbiased means equal to the 'stimulus' value (M), where SD = *w* * M. As a result, the overlap in modeled ANS distributions is a function of the stimulus *ratio* and the Weber fraction (*w*) of the ANS: As SD is proportional to M, the overlap of the distributions derived from 8 and 10 is the same as that derived from 80 and 100. (Note that while the model presented in this manuscript treats magnitude representations as linearly spaced, with scalar variability, they might alternatively be modeled as logarithmically spaced with constant variability. However, a Logarithmic model would yield similar findings.)

Here, we calculated the proportion of overlap between modeled magnitude distributions. Additionally, following the method used by Halberda et al. (2008), we used the *erfc* (the complementary error function) to determine the rate at which a given pair of magnitudes will not be distinguished. Assuming there are no other sources of error (such as bias or inattention), this erfc should be equal to twice the error rate of ANS-based magnitude judgments, as the observer would be presumed to choose the correct answer by chance on half of such trials. The equations used are given in the Appendix in the same Matlab code format we used for our calculations.

We seek to establish the theoretically maximal relationship between ANS and NDE-size, assuming no interfering factors or noise. Thus, we model theoretically ideal NDE-size at a given *w* for a given comparison pair as simply the differences between the overlaps or between the erfcs of different comparison ratios given that *w*. Real-world data would involve other sources of RT and error (attention to task, non-decision time, etc.) and make this relationship less clear. However, as these factors are separate from the ANS, they are excluded from this ideal model.

Results

The Relationship Between Overlap, ERFC, and *W*

In Figure 1, we illustrate the modeled ideal overlap of ANS magnitudes at various ratios (greater/lesser) as well as the calculated erfc. Recall, when controlling for total magnitude, smaller ratios map to smaller distances, and larger ratios map to larger distances. Smaller ratios have greater overlaps and greater error rates than larger ratios. The drop in both overlaps and erfcs is initially steep, but then "turns the corner" to asymptote to 0. Smaller *w*s (better acuity) and larger *w*s (worse acuity) both yield this same pattern, but the initial drop is steeper and the asymptote is reached faster for smaller *w*s. Using these values, the theoretical maximum contribution of ANS acuity to NDE-size can be found for any *w* on any particular NDE task.

The Relationship Between NDE-Size and *W*

As discussed above, an individual's NDE-size (i.e., savings in error rate and RT when comparing numerical magnitudes at large vs. small distances) should be functionally related to the difference in the amount of overlap between the comparison values' distributions, which is in turn related to the comparison ratios involved and the individual's *w*. We thus model NDE-size for a given *w* and pair of stimulus ratios as the difference in the model overlaps or erfcs:

NDE-size calculation for:
Overlap: Overlap(small ratio) – Overlap(large ratio)
erfc: erfc(small ratio) – erfc(large ratio)

We first model the ideal relationship between NDE-size and *w* for a comparison-pair set based on the distance effect paradigm originally developed by Moyer and Landauer (1967), and later expanded upon by researchers intending to measure NDE-size (see Sekuler & Mierkiews, 1977; Peters et al., 2008). In Moyer and Landaur's original paradigm, participants saw all possible pairs of non-equal integers between 1 and 9, and indicated which was larger. Although, it was Moyer and Landauer's intent simply to demonstrate that symbolic number comparisons yield distance effects, researchers often use modified forms of this procedure when measuring NDE-size (see Peters et al., 2008). A participant is shown a stimulus value and asked to indicate whether that value is greater or less than a central comparison value. The

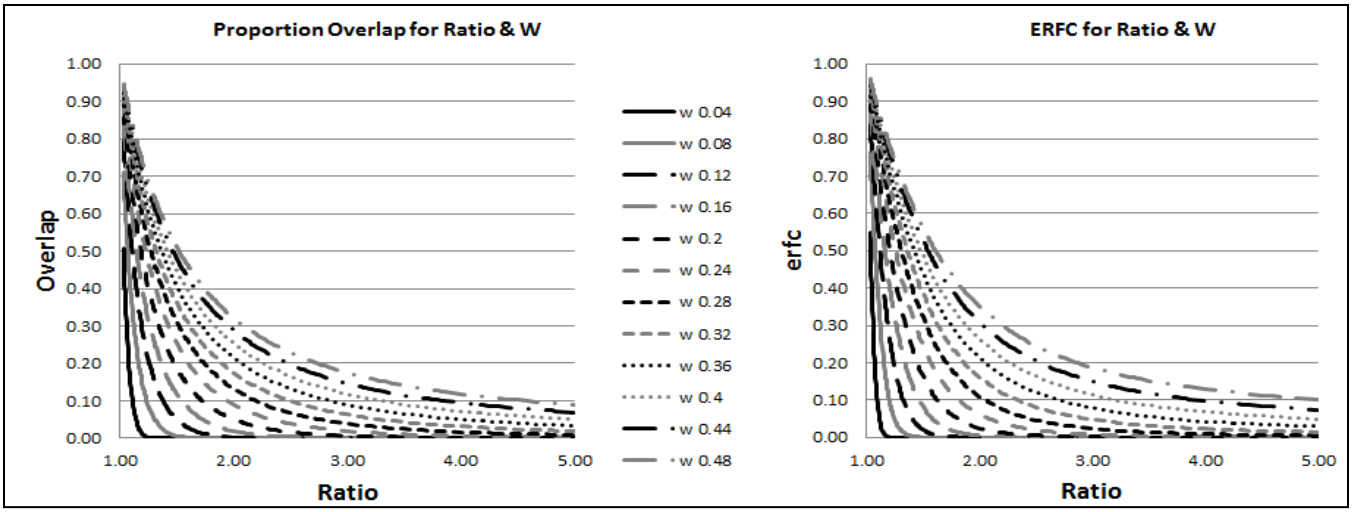


Figure 1: The modeled ANS magnitude distribution overlaps (left) and erfcs (right) for *w*s ranging from .04 (excellent acuity) to .48 (poorer acuity), at High/Low comparison value ratios ranging from 1 (equal values) to 5 (e.g., 50 vs. 10)

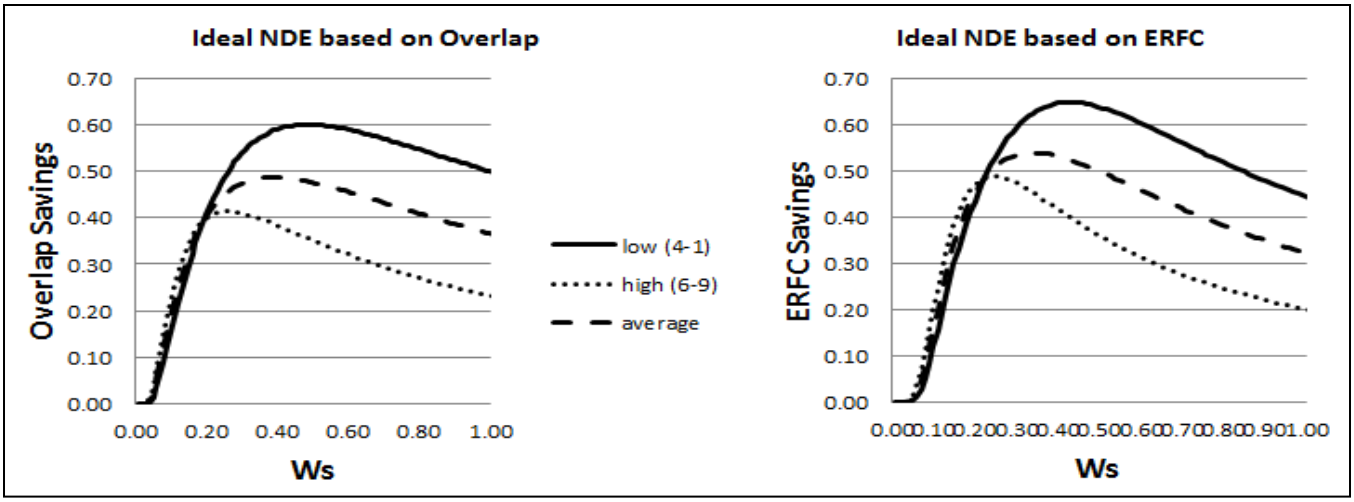


Figure 2. The modeled NDE-sizes for overlaps (left) and erfcs (right), based on values calculated for close distances (5 vs. 4 or 6) minus those calculated for far distances (5 vs. 1 or 9) distances for low ranges ([5 vs. 4] – [5 vs. 1]), high ranges ([6 vs. 5] – [9 vs. 5]), and their average.

stimuli follow a 2x2 design, varying both distance and direction from the central comparison value (e.g., 5). Half of the values are less (e.g., 1, 4) and half are greater (e.g., 6, 9) than the central value. Also, half are close (e.g., 4, 6) and half are far (e.g., 1, 9) from the central value. An individual's NDE-size is operationalized as the difference in accuracy and/or RT on close vs. far trials.

Recall, the overlap of ANS magnitude distributions (and thus the projected savings in RT and error rate) are dependent on the *ratio* of the values being compared, not their absolute distances. Thus, in this paradigm, although the absolute distances of the stimuli are symmetrical around the central comparison value, the ratios are asymmetrical. For the stimuli greater than the central value, the ratios are 6/5 (1.2) for the close value and 9/5 (1.8) for the far value. For the stimuli less than the central value, the ratios are 5/4(1.25) for the close values and 5/1 (5) for the far values. The spread of near and far ratios is thus much greater for the stimuli below than above the comparison value. However, the standard analyses used in the literature classify stimuli merely as near and far, and thus collapse across these ratio differences. We note that our model would yield the same NDE-sizes for any task using these same stimuli ratios and calculation methods regardless of the overall stimulus magnitudes invoked: The same pattern would be predicted for comparing 10, 40, 60, and 90 to a central value of 50.

The modeled NDE-sizes are presented in Figure 2. As can be seen, the relationship between *w* and NDE-size is not linear. Rather, it follows an inverted J shaped curve. NDE-size initially increases with *w*, slows to a peak, and then decreases with *w*. A strong positive linear relationship between *w* and NDE-size only exists for *w*s ranging between ~.05 and ~.20, quickly rising from near 0 savings to a 40% overlap savings, and 45% erfc savings. Both overlap and erfc then are near flat between *w*s of .20 and .60, with overlap savings peaking at 49% for *w*s of .39 and efrc savings peaking at 54% for *w*s of .34. Savings in overlap and erfc decline slowly with *w*s past these peaks. Clearly, the general presumption that larger NDE-sizes correlate with larger *w*s does not always hold. One could only expect to find a positive correlation with NDE-size and *w* if the target population's *w*s were located between .05 and .20. Indeed, depending on the population's *w* distribution, one could predict a positive, negative, or non-existent correlation between NDE-size and *w*.

An alternative method of gauging NDE-size is to look at the slope of the regression of RTs or error rates on ratio or distance, treating distance in a continuous fashion, rather than dichotomizing it to 'close' and 'far' (see Sekuler & Mierkiewicz, 1977). Negative slopes indicate the presence of a distance effect as larger comparison ratios (higher/lower) and – within a given magnitude range – larger distances would typically yield faster RTs and fewer errors than smaller ratios and distances. "Larger" (i.e., more strongly negative) absolute slopes are treated as indicating larger *w*s, and thus poorer ANS acuity.

Here we model the theoretically ideal NDE-slopes one would predict based on both the ratios and the absolute distances used for the dot-array comparison task developed by Chesney, Bjälkebring, and Peters (2015). This task used ratios between 1 and 2.6, with the total number of dots in an array ranging between 10 and 30. As illustrated in Figure 3, the same pattern emerges here as with the dichotomized analysis: a J-shaped curve. A strong linear relationship between *w* and slope only holds between *w*s of .08 and .2. Moreover, this difference in overlap begins to decrease with *w* as *w* exceeds .32. Although the exact location of these inflection points will vary depending on the stimuli used in a particular study, those we show in Figure 3 align very closely to those we modeled for paradigms finding the "slope" of the NDE, where RTs (here, overlaps) are regressed on the distances between compared values for all possible unequal parings of the values 1-9 (see Sekuler & Mierkiewicz, 1977). We also demonstrate regressions using ratio as the IV yield a stronger relationship between NDE-slope and *w*, but in either case this relationship is non-linear.

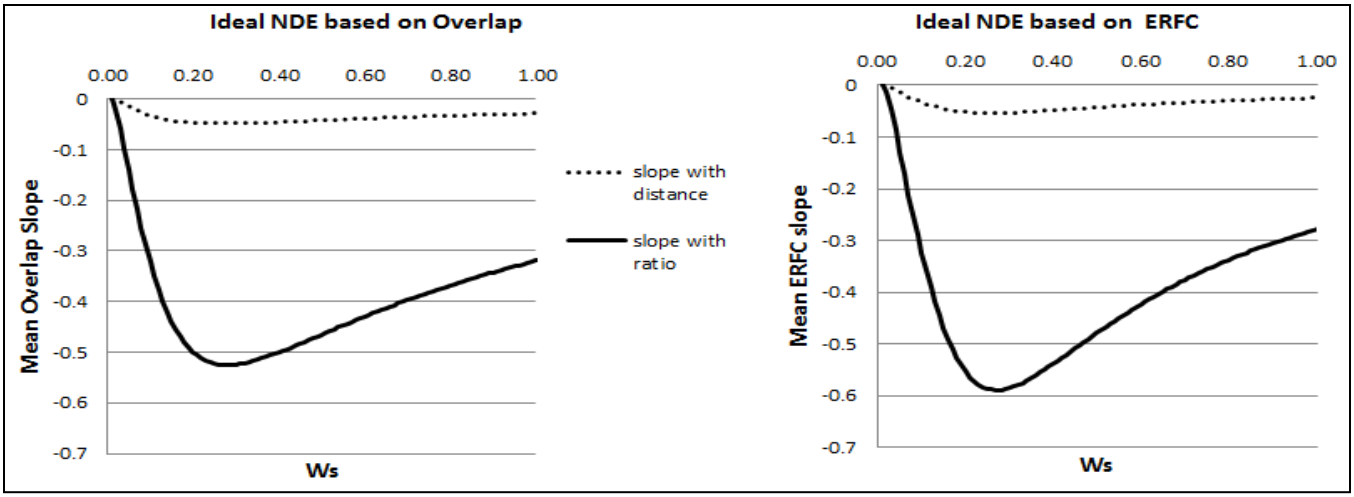


Figure 3: The modeled NDE-slopes for overlaps (left) and erfcs (right) for given *w*s based on both distances and ratios

Discussion

Despite the fact that the presence of numerical distance effects can indicate the involvement of the ANS in a task, it is becoming apparent that NDE tasks have limited utility for measuring individual differences in ANS acuity. With this model, we provide a novel theoretical exploration of the reasons why this is the case: Even under ideal conditions, one cannot expect NDE-size and ANS acuity to be linearly related. Rather, NDE-size initially increases, but then decreases with ANS acuity, with small NDE-sizes expected both for individuals with particularly small and particularly large *w*s. Our model demonstrates that the relationship between NDE-slope or NDE-size and *w* is dependent on both the ratio range of the stimuli and the *w* range in the population. This can be illustrated by considering the modeled NDE values in Figure 1. For example, if the target population's *w*s range between .04 and .20 and stimuli-pair ratios range between 1.25 and 2, one could indeed predict large NDE-sizes for larger *w*s (worse acuity), because the smaller *w*s would have already neared asymptote for this range of ratios, and would thus have small slopes. However, if the target population's *w*s ranged from .2-.4 and the stimuli ratios ranged from 1.05-1.2, the opposite pattern would emerge. These stimuli ratios are sufficiently small that all of these *w*s would be in the initial drop-off range, with smaller *w*s dropping faster: Smaller *w*s, (better acuity) would yield larger slopes.

For typical NDE-tasks, our model shows peak NDE-size is approached at *w*s of ~.2-.3. The location of this peak is a real concern, given the distribution of ANS *w*s in the normal population. Several studies with educated adult participants have found that their *w*s typically center around ~.22 (e.g., Cordes et. al, 2001; Whalen et al., 1999). Moreover, other studies of educated adults have found mean *w*s of .11 (Dehaene et al., 2008), and studies with infants have found *w*s of 1.0 (Xu & Spelke, 2000). Thus, one cannot expect that the range of *w*s for the population under test will coincide with the range of *w*s for which the relationship of *w* to NDE-size is quasi-linear on a standard task. Neither can researchers *a priori* assume a particular *w* distribution in a novel population, so as to be able to adjust their NDE tasks to yield a theoretically supported prediction of a linear relationship between NDE-size and *w*. This is problematic to the literature as a whole, and particularly for research attempting to draw conclusions about the nature of ANS acuity's involvement in other cognitive tasks.

Conclusions

Individual differences in ANS acuity have increasingly come to be considered an important predictor of human cognition and behavior. As such, it is important that the measures used to assess individual differences in ANS acuity are both reliable and valid. We recommend that future researchers assess ANS acuity via tasks whose reliability has been established. We specifically recommend against using NDE tasks to assess ANS acuity, as the validity of such measures is in doubt.

References

Chen, Q. & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: a meta-analysis. *Acta Psychologica, 148,* 163-172. doi:10.1016/j.actpsy.2014.01.016.

Chesney, D. L, Bjälkebring, P., & Peters, E. (2015). How to estimate how well people estimate: Evaluating measures of individual differences in the approximate number system. *Attention, Perception, & Psychophysics.* doi: 10.3758/s13414-015-0974-6

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7,* 25- 47.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*, 698–707. doi:10.3758/BF03196206

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science, 320,* 1217-

1220. doi:10.1126/science.1156540
Dehaene, S., Bossini, S. & Pascal, G. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General. 122*, 371-396.
Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *The Quarterly Journal of Experimental Psychology, 64*, 2009-2109. doi:10.1080/17470218.2011.574710
Gilmore, C.K., McCarthy, S.E., & Spelke, E.S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition, 115*, 394–406.
Halberda, J. & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, 6-year-olds and adults. *Developmental Psychology, 44*, 1457-1465. doi:10.1037/a0012682
Halberda, J., Mazzocco, M. & Feigenson, L. (2008). Individual differences in nonverbal number acuity predict maths achievement. *Nature, 455*, 665-668. doi:10.1038/nature07246
Holloway, I.D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103*, 17-29. doi:10.1016/j.jecp.2008.04.001.
Ingis, M. & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica, 145*, 147-155. doi:10.1016/j.actpsy.2013.11.009
Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37–44. doi:10.1177/0272989X0102100105
Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *American Journal of Psychology, 62*, 498-535.
Kingdom, F.A.A. & Prins, N. (2010). *Psychophysics: A Practical Introduction*. Academic Press: an imprint of Elsevier, London.
Maloney, E.A., Risko, E.F., Preston, F., Ansari, D., & Fugelsang, J.A. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica 134*, 154-161. doi:10.1016/j.actpsy.2010.01.006
Mechner, F. (1958). Probability relations within response sequence maintained under ratio reinforcement. *Journal of the Experimental Analysis of Behavior, 1*, 109-121. doi:10.1901/jeab.1958.1-109
Meck,W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes*, 9, 320–334. doi:10.1037/0097-7403.9.3.320
Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*, 1519–1520. doi:10.1038/2151519a0
Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). Intuitive numbers guide decisions. Judgment and Decision Making, 3, 619-635. Available at SSRN: http://ssrn.com/abstract=1321907
Price, G. R., Palmer, D., Battista, S., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica, 140*, 50-57.
Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (2011). The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. *Acta Psychologica, 136*, 73–80.
Schley, D.R., & Peters, E. (2014). Assessing "economic value": Symbolic number mappings predict risky and riskless valuations. *Psychological Science, 25*, 753-761. doi:10.1177/0956797613515485
Sekular, R., & Mierkiewiez, D. (1977). Children's Judgments of Numerical Inequality. *Child Development, 48*, 630-633. doi:10.2307/1128664
Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-243. doi:10.1111/1467-9280.02438
Taves, E. H. (1941). Two mechanisms for the perception of visual numerousness. *Archives of Psychology, 37*(265).
Weller, J., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch Analysis approach. *Journal of Behavioral Decision Making, 26*, 198-212. doi: 10.1002/bdm.1751
Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Non-verbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*, 130-137. doi:10.1111/1467-9280.00120
Xu, F., & Spelke, E.S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, B1-B11.

Acknowledgments

This research was supported by NSF grant SES–1155924. We thank Ellen Peters for her help and support.

Appendix

The Matlab code used to calculate the proportion of overlap in the ANS distributions of any two values, given *w:*

```
invCdf = 1- normcdf((High - Low)/(w*(High+Low)),0,1)
overlap = (2*invCdf)/(2-(2*invCdf))
```

The Matlab code used to find the erfc – the value of the complimentary error function – which is the rate at which the ANS values will not be distinguished and thus double the ideal error rate:

```
erfc(abs(High-Low)/(sqrt((High^2)+(Low^2))*sqrt(2)*w))
```

"High" referred to the higher value in a comparison pair (e.g., 6), while "Low" referred to the lower value in the pair (e.g., 5). "w" referred to ANS acuity (Weber fraction, *w*).

Strategic search in semantic memory

Evgenii Nikitin (E.Nikitin@warwick.ac.uk)
University of Warwick

Thomas Hills (T.T.Hills@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry, United Kingdom, CV4 7AL

Abstract

We search for various things every day – food, information on the Internet or someone's name in memory. Despite the different nature of these tasks, they all have a common feature – a final goal with an unknown location in a complex environment. This property of the search raises a problem of trade-off between exploration of new opportunities and exploitation of the known information. We used the data from the semantic fluency task experiment to investigate how humans switch between exploration and exploitation strategies when they search in memory and whether they do it optimally. On comparing four different search models, the one that assumes that humans switch search strategies according to the semantic quality of the current neighbourhood best fits the data. Moreover, participants who set higher thresholds for the words with better quality of the neighbourhood tend to retrieve more words from memory. We also used regression analysis to find out which factors affect efficiency of both search strategies.

Keywords: Semantic memory; Memory search; Exploration and exploitation

Introduction

Search tasks can be classified into three broad categories (Hills & Dukas, 2012): (a) external physical search (for example, animals which are foraging for food), (b) external information search (e.g., visual search), and (c) internal information search (such as search in memory). At first glance, these domains do not have much in common. However, it turns out that characteristics of these environments and strategies that individuals use are quite similar. One of the crucial properties that these strategies should have is the ability to successfully deal with the exploitation-exploration trade-off. Too much exploration can prevent a searching subject from maximizing the short-term reward, whereas redundant exploitation can lead to decreasing long-term reward.

The marginal value theorem (MVT) is formal description of optimal patch leaving rules for animals foraging in environments with well-defined patches (Charnov, 1976). It states that in a patchy system an animal should leave the patch when its food intake drops below the average long-term level expectation. The elements of this theory have been applied generally across a number of domains with patchy environmental structure, such as the search for digital information (Pirolli & Card, 1999) and, what we focus on here, search in semantic memory (Hills, Jones, & Todd, 2012).

In one test of semantic memory, the semantic fluency task (SFT), participants are given a fixed amount of time to produce as many words from a category (e.g. animals) as they can. Hills et al. (2012) suggested that semantic memory is organized as a spatial environment, where similar words are grouped in patches. To measure the distance between words in this space, different measures of semantic similarity have been implemented. Hills et al. (2012) used the BEAGLE model (Jones & Mewhort, 2007): Here a word is represented by two vectors – a random initial vector and a memory vector. As the text corpus is being processed, every time a particular word is encountered in a corpus, its memory vector is updated by adding the sum of the random vectors of the other words that appear with it. Final similarity between two words is calculated as the dot product between their resulting memory vectors. One of the advantages of this model is that two words might be recognized as semantically similar even if they rarely occur in context with one another, but nonetheless share similar contexts (e.g., bee and wasp).

The SFT is often analysed by looking at the time between successful retrievals, namely interresponse time or IRT (Davelaar & Raaijmakers, 2012). Using BEAGLE memory representation and hand-coded categories of animals, Hills et al. (2012) showed that IRTs increase up to the long-term average IRT, when participants then leave patches. This suggests that cognition is responding strategically to the structure of semantic memory.

Another approach to search in semantic memory finds that the same retrieval patterns can be generated by a random walk in memory (Abbott, Austerweil, & Griffiths, 2015). However, this study potentially confounds the search process with the representation (Jones, Hills, & Todd, 2015), and it remains unclear whether or not people search in memory strategically or how they might respond strategically to local memory structure. People may adapt their memory search strategies according to the properties of the environment (Hills, Kalff, & Wiener, 2013). The aim of the current study is to expand research in this area by creating a comprehensive model of search that incorporates the potential influence of local memory structure.

This study does this by incorporating the concept of *proximity*, which measures how semantically similar words are to the most recent retrieval. This measure roughly describes the quality of the local neighbourhood of a word. In particular, we find support for the cognitive influence of a maximally local measure of proximity, which we call *maximum proximity*; that is, how similarity of the nearest word.

The paper is organized as follows. In Section 2, we give a full description of several competing computational models of semantic memory search. In Section 3, we describe the details of the experimental study used for analysis. Section

4 includes the results of the model fitting. In Section 5, we discuss these results and give our conclusion in Section 6.

Model description

We begin by defining a formal model of the retrieval process. We assume that there are n words in semantic space that form the set $W = \{w_1, w_2, ..., w_n\}$, and each transition from the word w_a to the word w_b can be written as $w_a \rightarrow w_b$. There are also $1, ..., i, ..., M$ participants, and each of them has retrieved N_i words. The retrieval sequence of each participant can be described by vector $R_i = (r^i_1, ..., r^i_j, ..., r^i_{N_i})$, where each element is an element of W. For example, vector $R_i = (w_5, w_4, w_8)$ would mean that participant i retrieved three words – word 5, then word 4, and word 8. Finally, there are vectors $T_i = (t^i_1, ..., t^i_j, ...t^i_{N_i})$ for each participant that contain retrieval times of each word.

Dynamic retrieval model

To describe a stochastic process of retrieving items from memory, we use a dynamic Luce choice rule (Luce, 1959), similar to the SAM model (Raaijmaikers & Shiffrin, 1981). Given a set of cues, this model allows us to calculate the probability of retrieving a particular word from the set of the possible alternatives. We used the dynamic model that implies switching between exploitation (local search, denoted by L) and exploration (global search, G). In this context, exploitation is an attempt to find semantically similar words in the local neighbourhood, whereas exploration is associated with "jumps" across memory that do not rely on similarity between words but on individual word frequency. Probability of retrieval of item $j+1$ after item j of vector R_i is calculated as:

$$P^i(r^i_j \rightarrow r^i_{j+1}|L) = \frac{F^{\alpha_i}_{r^i_{j+1}} S^{\beta_i}_{r^i_j, r^i_{j+1}}}{\sum F^{\alpha_i}_{w_k} S^{\beta_i}_{r^i_j, w_k}}, \qquad w_k \notin \{r^i_1, ...r^i_j\} \tag{1}$$

$$P^i(r^i_j \rightarrow r^i_{j+1}|G) = \frac{F^{\gamma_i}_{r^i_{j+1}}}{\sum F^{\gamma_i}_{w_k}}, \qquad w_k \notin \{r^i_1, ...r^i_j\} \tag{2}$$

for the local and global search cases respectively. The numerator represents retrieval strength of the word r_{j+1}, and cue intensities of all the not yet retrieved words are added up in the denominator. α_i, β_i, and γ_i represent the attention weights assigned to the given cue; $F_{r^i_{j+1}}$ stands for log-frequency of the word r^i_{j+1}; $S_{r^i_j, r^i_{j+1}}$ is semantic similarity between words r^i_j and r^i_{j+1}.

The likelihood function is as follows:

$$Lik_i(\alpha_i, \beta_i, \gamma_i) = \prod_{j=1}^{N_i} [P^i_j(L) P(r^i_j \rightarrow r^i_{j+1}|L, \alpha_i, \beta_i) + \\ + P^i_j(G) P(r^i_j \rightarrow r^i_{j+1}|G, \gamma_i)] \tag{3}$$

Unfortunately, we are not able to directly observe which type of strategy (local or global) was used for retrieval of each word. However, we can use the idea of the search thresholds, which is described in the next subsection, to obtain $P^i_j(L)$ and $P^i_j(G)$.

Search threshold modelling

Humans do not have to use only one of the strategies (local or global) within one retrieval attempt and do not have to make the decision about which strategy to use just after retrieval of the new word. In this study, we assume that the subject always starts by browsing in the local neighbourhood in an attempt to find a similar word in semantic space (local search). After a certain amount of time, if this attempt fails, the subject switches to the global strategy. Analogously, foraging animals do not leave a patch immediately after finding a new piece of food, but rather do it when the rate of resource intake from this patch falls below a long-term average level (or at another chosen moment).

The maximum amount of time the subject is willing to spend on the local search we call the *search threshold*. According to Marginal Value Theorem, this threshold should be set to the moment when intake from the current patch becomes equal to the long-term average retrieval rate. However, some assumptions of MVT are often not satisfied in real-world environments (McNamara & Houston, 1987). For example, MVT assumes that individuals experience a smooth, continuous flow of rewards, whereas in reality rewards are often discrete. In addition, recent work showed that switching between local and global search increases over time and this may indicate that patch quality reduces as individuals retrieve from successive patches (Hills, Todd, & Jones, 2015). Moreover, MVT ignores the idea that a forager might be able to estimate the quality of the current patch and use this information to improve his strategy. With this in mind, we compare several competing threshold models:

1. *Static (infinite threshold) model.* This model assumes that there is no switching between strategies, and both cues (similarity and frequency) are always used. Therefore, the threshold is
$$X^i_j = \infty \tag{4}$$

2. *Fixed threshold model.* According to this model, participants use constant thresholds. Participant i will switch to the global search strategy if he fails to find a similar word within a_i seconds.
$$X^i_j = a_i \tag{5}$$

3. *MVT threshold model.* Marginal Value Theorem predicts that the subject leaves the patch when the retrieval rate in the current patch drops to the average long-term rate of retrieval in the memory space as a whole. The instantaneous retrieval rate from the patch is calculated as the average cumulative reward obtained in the patch (McNamara & Houston, 1987). We can then calculate the threshold by

equalling these two rates:

$$\frac{N_i}{\sum_{k=1}^{N_i} t_k^i} = \frac{NP}{TP + X_j^i} \tag{6}$$

$$X_j^i = \frac{NP \sum_{k=1}^{N_i} t_k^i}{N_i} - TP, \tag{7}$$

where NP is the number of words retrieved from the current patch, TP is the time spent in the current patch, X_j^i is the search threshold, $\frac{N_i}{\sum_{k=1}^{N_i} t_k^i}$ is the final long-term retrieval rate[1].

4. *Maximum Proximity threshold model.* This model is based on the hypothesis that people will switch to the global strategy faster if they sense that there are no semantically similar words nearby. For instance, the words *orangutan* and *octopus* have roughly the same frequency in language. However, humans might be able to almost instantaneously understand that *orangutan* has very similar words in the semantic space (for example, other monkeys), whereas *octopus* does not. Hence, they may switch to global search faster for *octopus*. To model this hypothesis, we find the not yet retrieved word that has the maximum similarity with the current word, and use this measure (*maximum proximity*) as a factor in the model. Furthermore, we use an exponential transformation of the similarity space (Shepard, 1987).

$$X_j^i = e^{a_i + b_i \max_k S_{r_j^i, w_k}},\ w_k \notin \{r_1^i, ... r_j^i\} \tag{8}$$

Where a and b are parameters fitting the exponential and $max_k S$ represents the maximum proximity. We also checked how the fit of the model changed if we used other measures of the quality of the local neighbourhood. In particular, we fit the models where we use the total similarity of 10, 20, 30, etc. nearest words as a factor in the threshold model.

Using the above models (Equations 4-8), we can then calculate $P_j^i(L)$ and $P_j^i(G)$ using the following rule:

$$P_j^i(L) = \begin{cases} 1, & \text{if } X_j^i \geq t_j^i \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

$$P_j^i(G) = \begin{cases} 1, & \text{if } X_j^i < t_j^i \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

That is, if the subject did not manage to retrieve a word within the threshold time X_j^i, he or she switches to the global strategy (G). We can now rewrite our likelihood function (Equation 3). For example, for the Maximum Proximity threshold model, it will look like this:

$$Lik_i(\alpha_i, \beta_i, \gamma_i, a_i, b_i) = \prod_{j=1}^{N_i} [P_j^i(L|a_i, b_i) P(r_j^i \to r_{j+1}^i | L, \alpha_i, \beta_i) + \\ + P_j^i(G|a_i, b_i) P(r_j^i \to r_{j+1}^i | G, \gamma_i)] \tag{11}$$

We can now find the values of parameters that maximize this likelihood function and compare the fit of the different models.

Search times analysis

To investigate which factors affect efficiency of local and global strategies, we used regression modelling. Each threshold model was used to classify all retrievals into one of two groups, L (local) or G (global). After categorising all retrievals, we calculated the time of local retrieval(TL_j^i) and the time of global retrieval (TG_j^i) for each observation:

$$TL_j^i = \begin{cases} t_j^i, & \text{if } X_j^i \geq t_j^i \\ X_j^i, & \text{otherwise} \end{cases} \tag{12}$$

$$TG_j^i = \begin{cases} t_j^i - X_j^i, & \text{if } X_j^i < t_j^i \\ NA, & \text{otherwise} \end{cases} \tag{13}$$

We then used regression modelling to find the relationships between TL_j^i and TG_j^i and different predictors, such as maximum proximity of the current word, time elapsed from the beginning of the task, etc. However, we need to remember that observations where $X_j^i < t_j^i$, are censored in the local case. We cannot calculate the actual time of retrieval because the strategy switch occurred before the local search was successful. For the further analysis we removed these observations from the dataset, although other approaches are also possible[2]. The dependent variable is positive and has a skewed distribution. Gamma GLM is often used for modelling such cases.

The following set of potential predictors was used:

- Maximum proximity – it should be easier to retrieve the word if MP of the word is higher.

$$MP_j^i = \max_k S_{r_j^i, w_k},\ w_k \notin \{r_1^i, ... r_j^i\} \tag{14}$$

- Current elapsed time – this variable was used to control for the different time effects, such as fatigue or time pressure. We also used a squared term of this variable as its effect might be non-monotonic.

$$TT_j^i = \sum_{k=1}^{j} t_k^i \tag{15}$$

[1] We also fit models using constantly updating long-term retrieval rates, but this produced inferior models to those presented here.

[2] For example, Cox proportional hazards model (Cox & Oakes, 1984) could be used as it can account for the censored observations. However, this model requires proportional hazard assumption to hold, and it is in general less robust than gamma GLM (Basu, Manning, & Mullahy, 2004).

- Total remaining frequency – retrievals should become slower after more frequent words are retrieved from memory

$$RF_j^i = \sum_{k=1}^{n} F_{w_k},\ w_k \notin \{r_1^i, ...r_j^i\} \tag{16}$$

- Average interresponse time (IRT) – we use this measure as a proxy of each participant's overall retrieval speed instead of adding a dummy variable for each participant to the model.

$$IRT_i = \frac{\sum_{k=1}^{N_i} t_k^i}{N_i} \tag{17}$$

- Global – a binary variable that indicates whether this retrieval was a result of the local (0) or global (1) search. We also added interactions of this variable with other predictors to the model.

Estimation results are described in the Results section.

Method

The dataset used in this study was initially collected for Hills et al. (2012) and is available online. Participants were 141 undergraduate students at Indiana University, who participated for course credit. Participants used computers for completing the task. They were asked to type in as many animals as they could remember in three minutes. One participant was eliminated from the current analysis because he produced a word that was not encountered in the Wikipedia text corpus used for training the BEAGLE model. Moreover, participants were allowed to type the last word after three minutes had passed, and we have removed the words retrieved after the deadline from consideration. This leaves us with $n = 354$ unique animals and a 354x354 semantic similarity matrix. This matrix gives us a spatial representation of the semantic memory. More detailed description of the experiment can be found in Hills et al. (2012).

Analysis

Evaluation of models

First, using the maximum likelihood method, we estimated parameters for the four threshold models (Equations 4-8). We then calculated individual AICs (Akiake Information Criterion) for each participant. Table 1 includes AIC summed across the subjects, the number of participants for whom each model achieves the minimum AIC, and average number of retrieved words of these participants. As we can see, the maximum proximity model is the best fit, and the MVT model is the worst. Furthermore, performance of participants who are better described by the MP models is on average higher than performance of the other participants. Additionally, we included Bayesian Information Criterion (BIC) values in this table, and they support our findings.

We now consider only the Maximum Proximity model that best fit the data. This model is not completely identifiable, as

Table 1: Comparison of the threshold models

Model	AIC	Δ_i	Number of people with min AIC	Number of retrieved words	BIC
Exp. maximum proximity threshold	46453.96	0.00	105	35.25	47499.21
Fixed threshold	46965.22	511.26	26	33.81	47801.42
Static model	47739.86	1285.90	9	34.22	48157.96
Marginal Value Theorem threshold	48894.71	2440.75	0	–	49521.86

for given values $\alpha_i, \beta_i, \gamma_i$ there are many combinations of parameters a_i and b_i that lead to the same value of the likelihood function. For the analysis of the effect of the coefficients, we used median value of b_i for each participant as a point estimator. When calculating thresholds below, we used an average value of the thresholds predicted by all equally likely models.

84 participants have positive median values of the coefficient b_i, which means that they to a global strategy more quickly when the maximum proximity of the word is lower. These participants produced on average 36.18 words, whereas participants with negative median b_i retrieved on average 33.20 words, and a Wilcoxon rank sum test indicated that this difference was statistically significant ($z = 2.22, p < .05$). After that, we split participants in five roughly equal groups based on the value of b_i. Figure 1 demonstrates that average performance grows as a function of b_i, reaches its maximum in $b_i = [3;5]$ interval and then drops again. This suggests that setting higher thresholds for words with high MP is beneficial for task performance, but only up to a certain point. This is consistent with U-shaped performance functions proposed for many tasks that rely on attentional control (Hills & Hertwig, 2011).

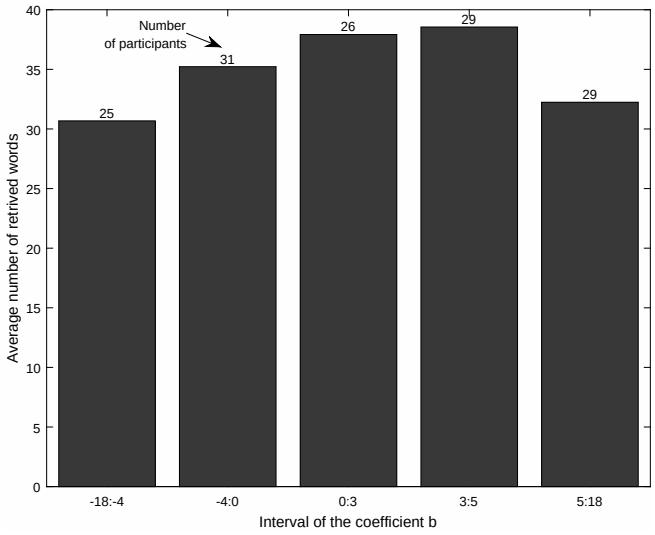


Figure 1: Relationship between number of retrieved words and the coefficient b_i

Next, we analysed how the fit of this threshold model depends on the measure that is used to estimate the quality of the local neighbourhood. We varied the number of nearest items that we included in the local neighbourhood and refit the model. As Figure 2 demonstrates, the goodness of fit is

decreasing as a function of the number of the words, and the maximum proximity model that used only the nearest word fits the data the best. Sensitivity of the model to the choice of this quality measure also indirectly demonstrates that the model outperforms its competitors not just due to the greater number of the free parameters.

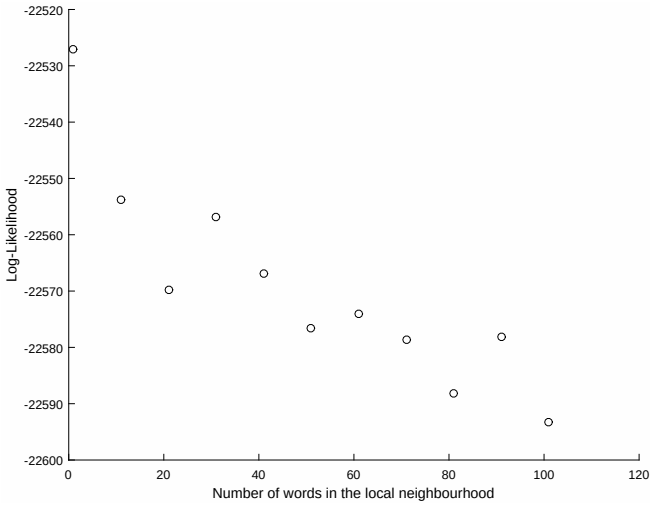


Figure 2: Connection between the quality of fit and the number of words in the local neighbourhood

Results of the GLM model estimation[3] are presented in Table 2. The results suggest that higher maximum proximity on average leads to faster retrievals. However, the effect of remaining frequency is important only for the global search, indicating a potential separation of frequency between local and global search. Interaction of the global search and remaining frequency variables is significant, and the magnitude of this effect is much larger than for the other variables. The effect of elapsed time is non-monotonic; we suppose that this form of the relationship might be caused by various time effects, such as fatigue and time pressure (e.g., participants might start making more effort closer to the end of the task).

Discussion

Overall, the results of analysis support the idea that search in memory is not purely random, bur rather sensitive to the structure of environment. Humans seem to be adapting their search strategy to the current state of the semantic space. We provided the following evidence to support this claim.

First, the retrieval process was best described by models with a finite threshold, which implies switching between local and global strategies. The infinite threshold model had the lowest AIC for only 6% of participants. Furthermore, the best model assumes that the time of this switching depends on the quality of the local neighbourhood, which can be approximated by a maximally local measure of proximity, MP. The MP model has the lowest AIC for 75% of participants. This

[3]Variables were standardized prior to the estimation.

Table 2: Regression modelling

	Standardized coefficients:
Intercept	1.30^{***}
	(0.02)
Maximum proximity	-0.05^{**}
	(0.02)
Time elapsed	0.18^{***}
	(0.05)
(Time elapsed)^2	-0.13^{***}
	(0.01)
Remaining frequency	-0.04
	(0.05)
Average IRT	0.18^{***}
	(0.02)
Global search	0.005
	(0.03)
Global search × Maximum proximity	-0.02
	(0.02)
Global search × Remaining frequency	-0.36^{***}
	(0.08)
Global search × Average IRT	0.23^{***}
	(0.03)
Global search × Time elapsed	-0.18^{*}
	(0.07)
Observations	4,758
Nagelkerke R^2	0.18
Log Likelihood	−10,181.250
Standard errors are in parentheses	*p<0.1; **p<0.05; ***p<0.01

finding supports the idea that humans use additional information about the current state of the environment to improve their retrieval process.

The poor fit of the MVT suggests that memory foraging strategies people use may involve more (or different) information than that described for the MVT. Going back to animal foraging, Iwasa, Higashi, and Yamamura (1981) noted that if additional information about the quality of the current patch (neighbourhood) is available, then simple strategies, such as MVT or constant threshold, will no longer be optimal. In our case, people might be able to estimate remaining richness of the neighbourhood in order to improve their search strategy. Similar to animals, they would be more prone to exploration in poor neighbourhoods and to exploitation in rich environments.

Next, participants who set higher thresholds for the words with greater MP tend to retrieve more words from memory. However, setting thresholds that are too high seems to hurt the performance. These results suggest that at least some participants were capable of quickly estimating quality of the semantic neighbourhood without directly accessing the words in it. The mechanism and efficiency of such estimation are both of great interest. One of the possible explanations is that people store "pre-harvest information" about quality of the different neighbourhoods and are capable of quickly accessing this information during the retrieval process (Valone, 1991).

Finally, regression analysis showed that remaining frequency of semantic space has strong effect only on the speed of the global search, whereas local search is affected only by the quality of the local neighbourhood. This finding supports the idea that these are two distinctive search strategies, and that they have the different nature.

Conclusion

Undoubtedly, search in memory involves a fair amount of randomness. However, calling this process purely random seems to be an overstatement. In this study, we attempted to show that humans might use certain methods to improve their performance in situations when they are required to retrieve words from a semantic category, such as names, countries or animals. In particular, we might be capable of saying that some words have more similar items in memory than other words, and of using this information for optimal strategy switching. Further research is necessary to understand how accurate these estimations are and what they are based on.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*, 558–569.

Basu, A., Manning, W. G., & Mullahy, J. (2004). Comparing alternative models: log vs cox proportional hazard? *Health Economics*, *13*, 749–65.

Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, *9*(2), 129–136.

Cox, D., & Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman & Hall.

Davelaar, E., & Raaijmakers, J. (2012). Human Memory Search. In P. Todd, T. T. Hills, & T. Robbins (Eds.), *Cognitive Search: Evolution, Algorithms, and the Brain* (chap. 11). Cambridge, MA: MIT Press.

Hills, T. T., & Dukas, R. (2012). The Evolution of Cognitive Search. In P. Todd, T. T. Hills, & T. Robbins (Eds.), *Cognitive Search: Evolution, Algorithms, and the Brain* (pp. 11–24). Cambridge, MA: MIT Press.

Hills, T. T., & Hertwig, R. (2011). Why Aren't We Smarter Already: Evolutionary Trade-Offs and Cognitive Enhancements. *Current Directions in Psychological Science*, *20*(6), 373–377.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–40.

Hills, T. T., Kalff, C., & Wiener, J. M. (2013). Adaptive Levy Processes and Area-Restricted Search in Human Foraging. *PLoS ONE*, *8*(4).

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: how we search through memory. *Topics in Cognitive Science*, *7*(3), 513–534.

Iwasa, Y., Higashi, M., & Yamamura, N. (1981). Prey Distribution as a Factor Determining the Choice of Optimal Foraging Strategy. *The American Naturalist*, *117*(5), 710–723.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden Processes in Structural Representations: A Reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*, 570–574.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. doi: 10.1037/0033-295X.114.1.1

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: John Wiley and Sons.

McNamara, J., & Houston, A. (1987). Foraging in patches: there's more to life than the Marginal Value Theorem. In S. S. Commons M.L. Kacelnik A. (Ed.), *Quantitative analysis of Behavior, Volume IV* (pp. 23–39). Lawrence Erlbaum Associates, Inc.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4).

Raaijmaikers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Valone, T. J. (1991). Bayesian and prescient assessment: foraging with pre-harvest information. *Animal Behaviour*, *41*, 569–77.

Context, but not proficiency, moderates the effects of metaphor framing: A case study in India

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
Daye Lee (daye.lee@oberlin.edu)
Department of Psychology, Oberlin College
120 W. Lorain Street; Oberlin, OH 44074 USA

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
Department of Psychology, Purchase College, SUNY
735 Anderson Hill Road; Purchase, NY 10577 USA

Abstract

Metaphors suffuse language and affect how people think. A meta-analysis of metaphor framing studies conducted between 1983 and 2000 concluded that metaphors are about 6% more persuasive than literal language (Sopory & Dillard, 2002). However, each of these studies was conducted in English with samples drawn from populations of native English speakers. Here, we test whether and how language proficiency moderates the influence of metaphor frames. Sampling from a population of non-native, but generally proficient English speakers from India, we found that metaphor frames systematically affected people who reported using English primarily in informal contexts (i.e., among friends and family and through the media) but not those who reported using English primarily in formal contexts (i.e., for school or work). We discuss the implications of this finding for countries like the US, where English is increasingly a non-native language for its residents, and for theories of language processing more generally.

Keywords: Metaphor, framing, analogy, persuasion, political psychology, reasoning

Introduction

Does the context in which people use a language moderate the influence of metaphor on thought? Are fluent speakers of a language more likely to use metaphors to reason about complex problems than people who are less fluent in the language?

The results of several studies suggest that metaphor frames *can* influence how people think about important socio-political issues (Hauser & Schwarz, 2014; Jia & Smith, 2013; Landau, Keefer, & Rothschild, 2014; McGuire, 2000; Ottati, Rhoads, & Graesser, 1999; Robins & Mayer, 2000; Sopory & Dillard, 2002; Thibodeau & Boroditsky, 2011, 2013, 2015). A meta-analysis of metaphor framing studies conducted between 1983 and 2000 concluded that metaphors are about 6% more persuasive than literal language (Sopory & Dillard, 2002). More recent experiments have found that metaphors can influence how people think about everything from cancer prevention (Hauser & Schwarz, 2014) to immigration (Landau, Sullivan, & Greenberg, 2009) and crime (Thibodeau & Boroditsky, 2011, 2013, 2015).

However, each of these studies was conducted in English with samples drawn from populations of native English speakers. Given the prevalence of metaphor in discussions of socio-political issues and the changing landscape of language demographics in countries like the United States (where the percentage of people who report speaking a language other than English at home increased 158.2% between 1980 and 2010; Ryan, 2013), there are important practical motivations for investigating whether and how how non-native (second or foreign language) speakers are influenced by metaphor frames.

Considering the role of language fluency in metaphoric thinking may also inform how cognitive scientists think about the role of language in reasoning more generally. One possibility is that people think more "systematically" when using a second language (Costa et al., 2014; Keysar, Hayakawa, & An, 2012). Second languages may be less valenced and arousing than first languages (e.g., Ayçiçeği & Harris, 2004; Pavlenko, 2007) and processed less automatically (Favreau & Segalowitz, 1983). If being influenced by a metaphor frame is similar to well-studied decision-making biases (e.g., loss aversion in gain/loss framing), then second language speakers of English may be less likely to be influenced by metaphorical descriptions of social dilemmas.

On the other hand, if metaphor is a core feature of language or if there is nuanced variability in peoples' facility with second languages, then one may expect to find systematic effects of metaphor on decision-making among (at least a subset) of a non-native speaking population (e.g, Harris, 2004).

In the present study, we identify a population of non-native, but generally proficient English speakers from India. We distinguish between individuals who tend to use the language in more formal environments like school and work from individuals who tend to use the language in more informal environments like with friends and family or through media like television and radio. We hypothesized that people who used English with friends and family (i.e., in informal contexts) – or simply more frequently – would be more likely to be influenced by the metaphor frames (Freed, Segalowitz, & Dewey, 2004). Using English in informal settings (or more frequently) may imbue the language with more emotional valence and lead this subgroup to process metaphors more like native English speakers.

Experiment

Population and Participants

We recruited participants from India, a country with a diverse population and a rich history of multilingualism (Annamalai, 2004). Although there are more than 150 recognized languages in India and over 500 dialects (Ahmad, 2009), two of the most popular are the two official languages of the country: Hindi and English. As a result, many people who live in India speak both English and Hindi. Importantly though, there is variability in the setting in which people use these languages, the age at which they begin learning these languages, and their level of proficiency with each language.

In addition to providing interesting language demographics, the population of India represents an understudied group in cognitive science (Henrich, Heine, & Norenzayan, 2010). With the increased use of internet-based methodologies (e.g. Amazon's Mechanical Turk; Buhrmester, Kwang, & Gosling, 2011), however, this population is now much more accessible. A recent survey has found that there are more people on Mechanical Turk in India than any other country (Pavlick, Post, Irvine, Kachaev, & Callison-Burch, 2014).

We asked participants in our sample whether they used English in primarily formal (e.g., school and work) or informal contexts (e.g., at home, among friends, via media), as well as a variety of other questions about their familiarity with and use of English. As expected (see Table 1), participants who reported using English more often in informal settings tended to report learning English at an earlier age ($p = .059$) and using English more frequently overall – in every setting that we asked about (e.g., at home, school/work, and among friends and family); these participants also reported that they understood and could speak English better.

As shown in Table 1, participants who reported using English more often in informal settings tended to be more educated: 41.6% of these participants reported having an advanced degree (Master's or Doctorate) compared to 24.7% of participants who reported using English primarily in more formal settings. Thus, the difference between groups may be related to differences in education or motivation or intelligence between groups – but not in socioeconomic status, political ideology, or current age. We consider this possibility by including participants' educational history as a covariate in our analyses.

Table 1: Patterns of English use as a function of where participants most commonly use the language. Asterisks indicate statistically significant differences between groups at the $^{*}p < .05$ and $^{***}p < .001$ levels.

	Informal	Formal
N	209	160
Age of acquisition	8.25	9.49
Frequency*** (1-5)	3.50	3.11
Home*** (1-4)	3.00	2.24
Media*** (1-4)	3.26	2.90
Friends*** (1-4)	3.08	2.71
Family*** (1-4)	2.83	2.11
School*** (1-4)	3.73	3.27
Work*** (1-4)	3.59	3.29
Other*** (1-4)	2.85	2.44
Understand*** (1-5)	4.42	4.07
Speak*** (1-5)	4.31	3.83
Age	31.95	33.44
Education***(1-7)	6.30	5.92
Ideology(0-100)	49.16	47.73
SES (1-4)	2.31	2.26
Gender: Male*	54%	66%

Stimuli

Descriptions of 10 socio-political dilemmas were created. For each issue, two metaphor frames were designed to promote different ways of thinking about the dilemma (see Table 2 for domains and frames). For instance, a crime problem was either described as a VIRUS *plaguing* or BEAST *preying on* a community; in a discussion of education, schools were either described as GARDENS *nurturing* or FACTORIES *molding* young minds.

The metaphor frames and candidate responses were adapted from news reports and other descriptions of the target issues; they were designed to be interpretable by a broad audience (none were about specific issues in the United States, for instance) with common conventional metaphors. Related versions of the stimuli have been used to investigate how metaphors influence thought among native English speakers, thereby providing a point of comparison for the present study (Thibodeau & Gehring, 2015; Thibodeau, in press).

Table 2: Issues and Frames

Issue	Frame A	Frame B
Crime	Virus	Beast
Income Inequality	Destabilize	Split
Education	Gardens	Factories
Environment	Backbone	Gem
Partisan Politics	Theater	Battle
Scientific Research	Puzzle	Summit
Housing	Failing Organ	Blemish
Medicine	Ecosystem	Assembly Line
Cheating	Boxer	Goalkeeper
Sports	Detective	Sniper

A single follow-up question for each issue asked the participant to suggest a policy intervention, attribute blame, or make an inference about the dilemma described. In every case, participants were asked to select between two candidate responses. For instance, after reading about a crime problem, participants were asked which of two policy interventions they thought would be most likely to reduce crime.

For half of the items, the response options were aligned with the metaphor frames; for half of the items, they were not. Consider the following three ways of addressing a crime problem:

(a) Increase street patrols that look for criminals.

(b) Reform educational practices and create after school programs.

(c) Expand economic welfare programs and create jobs.

Prior work has found that "increasing street patrols" is more consistent with a crime BEAST while "reforming educational practices" and "expanding economic programs" are more consistent with a crime VIRUS (Thibodeau & Boroditsky, 2011, 2013). Thus, when a participant is asked to choose between options (a) and (b) or between options (a) and (c), one should expect to find a systematic influence of the frames. Reading about a crime BEAST should make people more likely to suggest "increasing street patrols"; reading about a crime VIRUS should make people less likely to suggest "increasing street patrols" in favor of "education" or "economic" reform. And indeed that is what previous work has found. However, if a participant is asked to choose between options (b) and (c), one should not expect to find a systematic influence of the frames (i.e., because both options are either congruent or incongruent with the metaphor frame).

The present experiment was designed so that each participant would encounter five issues for which the response options mapped clearly onto the metaphor frames (e.g., option *a* contrasted against option *b*) and five issues for which the response options did not map onto different frames (e.g., option *b* contrasted against option *c*).

Since there were only two metaphor frames for each issue, the distinction between *well-aligned* and *non-aligned* refers to a manipulation of the response options. For instance, response options (a) "increase patrols" and (b) "education reform" were considered *well-aligned* to the metaphor frames for the crime dilemma, while response options (b) "education reform" and (c) "economic reform" were considered *non-aligned* to the metaphor frames for the crime dilemma, regardless of whether crime was framed as a BEAST or VIRUS. Each participant read one of the four versions (one of the two metaphor frames; one of the two pairs of response options) of each of the 10 dilemmas.

This distinction between *well-aligned* and *non-aligned* items was validated in a previous study. Thibodeau and Gehring (2015) operationalized this distinction by presenting participants with two metaphor frames and two response options for the 10 issues listed in Table 2 (N = 100 per item). Participants were instructed to match the frames to response options. The degree to which the sample matched response options to frames consistently was taken to reflect the clarity or congruence of conceptual mappings.

For instance, one group of participants was told that two politicians were using different metaphors (BEAST and VIRUS) to argue for different approaches to a city's crime problem: one was encouraging the city to "Increase street patrols that look for criminals" while the other was encouraging the city to "Reform educational practices and create after school programs." For this item, Thibodeau and Gehring (2015) found that 77% of participants matched "patrols" to BEAST and "education" to VIRUS, suggesting that these two response options were *well-aligned* with the frames. Another group of participants was asked to match the same frames with different approaches to the city's crime problem: "Reforming educational practices and create after school programs" and "Expanding economic welfare programs and create jobs." For this item, participants showed less agreement: 59% matched "education" to BEAST and "economy" to VIRUS, suggesting that these two response options were *not well-aligned* with the frames.

Items designated *well-aligned* yielded patterns of matching that were significantly more consistent ($M = .778$, $SD = .086$) than items designated *not well-aligned* ($M = .575$, $SD = .056$), $t(9) = 11.605$, $p < .001$, $d = 3.670$ (Thibodeau & Gehring, 2015).

Including this manipulation yielded 40 items: 10 issues, each with two metaphor frames, paired either with responses that mapped clearly onto the frames (*well-aligned*) or with responses that did not map clearly onto the frames (*non-aligned*).

Procedure

Data from 400 participants were recruited from Amazon Mechanical Turk. Participants were required to live in India and have a good performance record on previous tasks. Participants were instructed to participate only if they were at least 18 years old.

The task involved reading and answering questions about 10 metaphorically framed dilemmas – of which five included response options that were *well-aligned* with the frames and five that included response options that were *not well-aligned*. An embedded timer recorded how long participants spent reading and responding to the dilemmas. Data from participants who spent less than 5 seconds on average or more than a minute on average reading and responding to the dilemmas were excluded ($n = 31$), leaving data from 369 participants for analysis.

After reading and answering questions about the 10 dilemmas, participants were asked about their familiarity with English (i.e., "How frequently do you use English?" "How would you rate your English speaking ability?" "How would you rate your English comprehension ability?"; all rated along 5-point scales), about the context in which they primarily used the language (*formal*: school, work; or *informal*: among friends and family, through media like TV and radio), and about the age at which they began learning English. Participants were also asked about the frequency with which they used English overall and in specific settings (i.e., with friends, family, at school, at work, at home, in media consumption, and in other contexts). They responded to these questions

using a 4-point scale that ranged from "never' to "very frequently."

Finally, participants were asked a set of more general demographic questions, including their age, educational history, socio-economic status, political ideology, and gender. Summary statistics related to the language history and demographics of the sample are shown in Table 1.

Analysis

Responses were coded as "congruent" or "incongruent" with the frame they were presented with. For instance, in the context of the crime example, "Increase street patrols" was coded as congruent with the BEAST frame and incongruent with the VIRUS frame for the *well-aligned* version (i.e., when the response options contrasted "increasing street patrols" with "education reform"); for the *non-aligned* version (i.e., when "education reform" was contrasted against "economic reform"), "economic reform" was coded as congruent with a crime VIRUS and incongruent with a crime BEAST.[1] This approach captures the joint effects of the pairs of metaphor frames and is consistent with prior work on metaphor framing (e.g., Robins & Mayer, 2000; Thibodeau & Boroditsky, 2011, 2013).

One advantage of such a coding scheme is that it provides a clear metric for interpreting the degree to which metaphors influence judgments. Metaphor frames that do not systematically influence the way people think about an issue will yield a congruence score close to .5; metaphor frames that influence people to choose the response option that is consistent with the frame's entailments will yield a congruence score above .5; and metaphors that influence people to choose the response option that is inconsistent with the frame's entailments will yield a congruence score less than .5.

Results

Separate averages were computed by participant for the proportion of congruent responses to the *well-aligned* items and for the *non-aligned* items. Then a mixed-ANOVA was fit to these scores with two predictors: a within-subjects factor for item type (*aligned* or *not*) and a between-subjects factor for the context in which participants' reported they used English predominantly (2 levels: in *formal* settings like school or *informal* settings like at home). The model revealed a statistically significant interaction between item type and context, $F(1,367) = 4.406$, $p = .037$, $\eta^2 = .012$. There were no main effects of item type or English usage context, Fs < 2.1, ps $> .15$.

As illustrated in Figure 1, participants who reported using English in informal settings were systematically influenced by the metaphor frames when the response options were *well-aligned* with the frames ($M = .544$, $95\%CI = [.503, .585]$). Responses were no different from chance in the other three cases – including for participants who reported using English in formal settings for items that were *well-aligned* (see Figure 1).

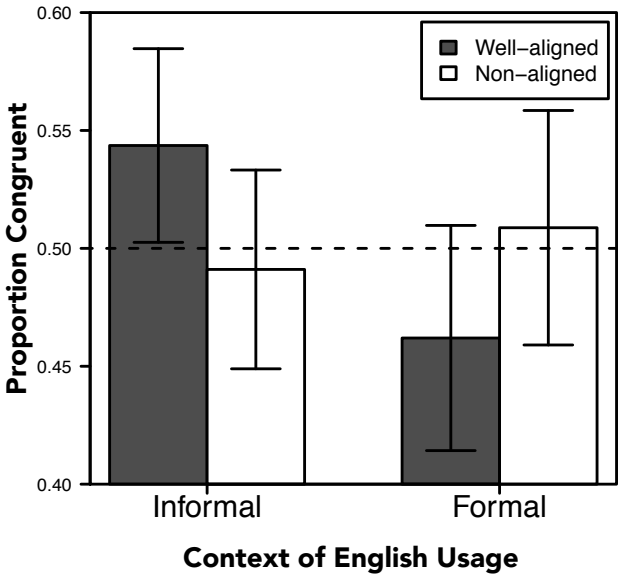


Figure 1: Congruent responding by item type (*aligned* or *not*) and where the participant reported using English. Error bars denote 95% confidence intervals.

The percentage of congruent responses by issue for *well-aligned* items are shown in Table 3. One indication that the metaphors affected people who used English in an informal context more than for people who used English in a formal context can be seen in the variability in congruence for the two groups. For participants who reported using English in an informal context, the proportion of congruent responses was more consistent across items ($SD = .055$, $95\%CI_{SD} = [.038, .100]$) than for participants who reported using English in a formal context ($SD = .082$, $95\%CI_{SD} = [.056, .150]$).

Table 3: Percentage of congruent responses by issue for *well-aligned* items by English usage context.

Issue	Informal	Formal
Crime	52	41
Income Inequality	55	36
Education	51	48
Environment	55	44
Partisan Politics	64	62
Scientific Research	59	51
Housing	46	57
Medicine	57	38
Cheating	47	53
Sports	54	49

As noted earlier, participants who reported using English

[1] Although patterns of matching were less systematic for *non-aligned* items, they were different from chance, $M = .575$, $95\%CI = [.54, .61]$; range $= [.52, .68]$, allowing us to code responses as congruent or incongruent with the frame.

primarily in informal settings tended to be better educated than participants who reported using English primarily in formal settings. To test whether this difference could account for the effect of usage context, we added participants' educational history to the ANOVA. In this analysis, the interaction between usage context and item type was consistent with that of the previous model, $F(1,365) = 4.445$, $p = .036$. The model did not reveal a main effect of education or interactions (either 2- or 3-way) between education, item type, and usage context, Fs < 3, ps $> .08$. However, we did find a marginally significant interaction between education and item type, $F(1,365) = 2.969$, $p = .086$. Controlling for where participants reported using English, the model suggested that more educated participants were somewhat less likely to give a response that was congruent with how the *well-aligned* dilemmas were framed, $B = -.017$, $SE = .012$ (for comparison, $B = .025$, $SE = .012$, for the interaction between item type and usage context).[2]

We also tested whether language usage frequency, a scalar measure, rather than the setting in which participants reported using English most often, moderated the influence of the metaphor frames. A two-way mixed ANOVA revealed no interaction between item type (*aligned* or *not*) and English usage frequency, $F(1,367) = 2.115$, $p = .147$, suggesting that the setting in which people use language impacts how people process metaphor frames—and not just how frequently they use the language.[3]

Discussion

The present study was designed to test whether language proficiency moderates the effects of metaphor framing on reasoning. Consistent with our hypothesis, we found that Indian participants who reported using English in informal settings – at home, with friends and family, and through the media – as opposed to in formal settings like school and work were most likely to show a systematic influence of the metaphor frames. As predicted (and consistent with prior research), this effect was restricted to items for which there was a clear alignment between the frames and response options.

There are several important implications of this finding. First, it represents an initial step toward understanding how non-native speakers process and use metaphors that commonly suffuse discussions of socio-political issues. We found a shift of about 9 percentage points toward the congruent response among participants who reported using English in informal settings, for *well-aligned* items. The magnitude of this effect is similar to what was found with these stimuli among native English speakers: a shift of about 6 percentage points.

This finding suggests that the context in which we learn and use a language can influence how we interpret linguistic information and how the language affects decision-making. Rather than simply distinguishing between native and non-native speakers of a language or between people who are more or less proficient in the language, we found that usage context affected behavior. On this view, "foreign language effects" (Costa et al., 2014; Keysar et al., 2012) may result from differences in where and how multilinguals use the languages they speak.

Studies of where and how people use language support this interpretation (Marian, Blumenfeld, & Kaushanskaya, 2007). First languages tend to be learned primarily through interactions with friends and family and through media (in relatively informal contexts), whereas second languages are often learned in classroom environments (in comparatively formal contexts).

Of course, it is difficult to distinguish the general influence of language proficiency from the more nuanced influence of contextual factors, as these variables are highly related. People who tend to use a language in familial or social settings tend to consider themselves more proficient in the language. However, an analysis of our data, in which a measure of language fluency was substituted for the measure of usage context, revealed no statistically significant effects, suggesting that the context in which people reported using English mattered more than their self-reported fluency with the language. That said, future work should seek to obtain more objective measures of language proficiency in order to more carefully examine these relationships.

The present work also complements ongoing cross-linguistic investigations of metaphor in conceptual representation. Cognitive linguists have identified remarkable similarity in the kinds of metaphors used for abstract domains like emotions, mental states, and time in different languages (Yu, 1998; Ferreira, 2008). For instance, HEAT is a common metaphor for ANGER in English, as well as in Chinese and Portuguese.

To our knowledge, there has not been systematic investigation of the kinds of conventional metaphors commonly used to discuss socio-political issues across languages. Crime, for instance, may or may not be commonly discussed as a VIRUS or BEAST in places other than the US and in languages other than English. However, our findings suggest that people who use English informally — to talk with friends and family and through the media — are affected by metaphor frames in ways that are quantitatively similar to native English speakers.

[2] Substituting the measure of education for English usage context revealed no significant main effects or interactions: $F(1,367) = 1.637$, $p = .202$ for the interaction between item type and education in this model.

[3] A model with all three predictors revealed a significant interaction between item type and language context, $F(1,365) = 4.437$, $p = .036$, consistent with the initial analysis reported, and no other significant main effects or interactions.

References

Ahmad, A. (2009). *Geography of the south asian subcontinent: A critical approach*. New Delhi: Concept Publishing Company.

Annamalai, E. (2004). Nativization of english in india and its effect on multilingualism. *Journal of Language and Poli-*

tics, *3*(1), 151–162.

Ayçiçeği, A., & Harris, C. (2004). Bilinguals' recall and recognition of emotion words. *Cognition & Emotion*, *18*, 977-987.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3–5.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS ONE*, *9*(4), e94842.

Favreau, M., & Segalowitz, N. S. (1983). Automatic and controlled processes in the first-and second-language reading of fluent bilinguals. *Memory & Cognition*, *11*(6), 565–574.

Ferreira, L. C. (2008). A psycholinguistic study on metaphor comprehension in a foreign language. *ReVEL*, *6*(11), 1-23.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second langauge fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*, 275–301.

Harris, C. L. (2004). Bilingual speakers in the lab: Psychophysiological measures of emotional reactivity. *Journal of multilingual and multicultural development*, *25*(2-3), 223–247.

Hauser, D. J., & Schwarz, N. (2014). The war on prevention bellicose cancer metaphors hurt (some) prevention intentions. *Personality and Social Psychology Bulletin*, *41*(1), 66-77.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–83.

Jia, L., & Smith, E. R. (2013). Distance makes the metaphor grow stronger: A psychological distance model of metaphor use. *Journal of Experimental Social Psychology*, *49*(3), 492–497.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*(6), 661–668.

Landau, M. J., Keefer, L. A., & Rothschild, Z. K. (2014). Epistemic motives moderate the effect of metaphoric framing on attitudes. *Journal of Experimental Social Psychology*, *53*, 125–138.

Landau, M. J., Sullivan, D., & Greenberg, J. (2009). Evidence that self-relevant motives and metaphoric framing interact to influence political and social attitudes. *Psychological Science*, *20*(11), 1421–1427.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967.

McGuire, W. J. (2000). Standing on the shoulders of ancients: Consumer research, persuasion, and figurative language. *Journal of consumer research*, *27*(1), 109–114.

Ottati, V., Rhoads, S., & Graesser, A. C. (1999). The effect of metaphor on processing style in a persuasion task: A motivational resonance model. *Journal of Personality and Social Psychology*, *77*(4), 688.

Pavlenko, A. (2007). *Emotions and multilingualism*. Cambridge University Press.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92.

Robins, S., & Mayer, R. E. (2000). The metaphor framing effect: Metaphorical reasoning about text-based dilemmas. *Discourse Processes*, *30*(1), 57–86.

Ryan, C. (2013). Language use in the united states: 2011. *US Census Bureau: American Community Survey Reports*, *3*, 1-16.

Sopory, P., & Dillard, J. P. (2002). The persuasive effects of metaphor: A meta-analysis. *Human Communication Research*, *28*(3), 382–419.

Thibodeau, P. H. (in press). Extended metaphors are the home runs of persuasion: Don't fumble the phrase. *Metaphor and Symbol*.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, *6*(2), e16782.

Thibodeau, P. H., & Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PloS one*, *8*(1), e52961.

Thibodeau, P. H., & Boroditsky, L. (2015). Measuring effects of metaphor in a dynamic opinion landscape. *PloS one*, *10*(7), e0133939.

Thibodeau, P. H., & Gehring, K. (2015). Comparing metaphors reveals their persuasive capacity. In D. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Yu, N. (1998). *The contemporary theory of metaphor: A perspective from chinese*. Philadelphia: John Benjamins.

The Determinants of Knowability

Samuel G. B. Johnson[1], Kristen J. Kim[2], & Frank C. Keil[1]
(samuel.johnson@yale.edu, kjkim@princeton.edu, frank.keil@yale.edu)
[1]Department of Psychology, Yale University, New Haven, CT 06520 USA
[2]Department of Psychology, Princeton University, Princeton, NJ 08544 USA

Abstract

Many propositions are not known to be true or false, and many phenomena are not understood. What determines what propositions and phenomena are perceived as *knowable* or *unknowable*? We tested whether factors related to scientific methodology (a proposition's *reducibility* and *falsifiability*), its intrinsic metaphysics (the *materiality* of the phenomena and its *scope* of applicability), and its relation to other knowledge (its *centrality* to one's other *beliefs* and *values*) influence knowability. Across a wide range of naturalistic scientific and pseudoscientific phenomena (Studies 1 and 2), as well as artificial stimuli (Study 3), we found that reducibility and falsifiability have strong direct effects on knowability, that materiality and scope have strong indirect effects (via reducibility and falsifiability), and that belief and value centrality have inconsistent and weak effects on knowability. We conclude that people evaluate the knowability of propositions consistently with principles proposed by epistemologists and practicing scientists.

Keywords: Lay epistemology; folk science; experimental philosophy; psychology of religion.

Introduction

Facts and falsehoods abound in both lay and scientific discourse. There are facts about the relationship between electricity and magnetism, and the possibility of multiple universes; about the mechanisms of cell division, and the origins of life; about the relationship between supply and demand, and the causes of crime. Beyond science, ordinary humans debate, both with one another and with themselves, questions about the existence of God, the afterlife, and the beginning of the universe.

Science aims to separate the true from the false, as does everyday cognition. Yet, sometimes it is unclear how to determine what is true: Cognitive science, for example, is much closer to understanding early vision than the nature of conscious experience. Indeed, some view the latter question as essentially *unknowable* (e.g., Chalmers, 1996; cf. Dennett, 1991).

The question of what is knowable and what is unknowable has important implications for individuals and for society. Humans have a strong drive to understand the world and explain their lives (e.g., Gopnik, 1998; McAdams, 1993), filling in gaps in their understanding using whatever means are available (Johnson, Rajeev-Kumar, & Keil, 2015). Even though much is unknown, we have developed various strategies for resolving this ignorance. We allocate science funding not only on the basis of what discoveries would be important, but also what discoveries are feasible. And when a topic is not amenable to empirical inquiry, it is seen as a more appropriate domain for faith (Preston & Epley, 2009). We spend our days pursuing the knowable, even as the unknowable keeps us up at night.

Here, we ask what leads people to perceive a phenomenon as knowable. Our strategy is to examine those factors that practicing scientists and philosophers have considered relevant to assessing knowability, anticipating that laypeople may use similar reasoning.

Much evidence has accumulated that adults, and even children, come to understand the world much as scientists do (e.g., Carey, 1985; Gopnik, Meltzoff, & Kuhl, 1999; Keil, 2003). This is not necessarily surprising: To the extent that evolution has endowed us with reasoning mechanisms that track reality, and to the extent that culture has invented scientific methods that track reality, one would expect those methods to converge. Thus, laypeople and scientists potentially share idealized methods for generating and assessing knowledge claims.

This approach led us to consider six factors, divisible into *methodological*, *intrinsic*, and *relational* factors.

Methodological factors. The first two factors concern the applicability of the (scientific or folk-scientific) methods by which propositions are assessed:

Reducibility. According to classic definitions (e.g., Putnam, 1973), a higher-level proposition is *reducible* to lower-level propositions (such as the facts of fundamental physics) if the higher-level can be deduced from the lower-level or, especially, if the higher-level can be *explained* in terms of the lower level. If all phenomena can be reduced to fundamental physics, this reduces the burden of science to understanding the physics and its implications. Arguably, then, to the extent that a proposition is reducible to propositions about fundamental physics, it should be more knowable.

Although the doctrine that all sciences can be reduced to fundamental physics is controversial among philosophers of science (e.g., Putnam, 1973; Strevens, 2009), it is plausible that laypeople see phenomena as knowable to the degree they are reducible. This is in part because laypeople are themselves reductionists: They prefer lower-level explanations of higher-level phenomena, within limits (Burke, Johnson, & Keil, 2016).

Falsifiability. The most well-known doctrine of scientific knowability is certainly Popper's theory of *falsificationism* (Popper, 1959/1934). The idea is that a proposition counts as scientific only if it can be proven false by an empirical test. This proposal has likely received such widespread influence among scientists because it resonates with intuitive notions of how

evidence relates to hypothesis (in philosophy, see Mill, 1967/1843; in psychology, see White, 1990).

Intrinsic factors. The next two factors concern the intrinsic nature or content of the phenomenon itself:

Materiality. Some (alleged) phenomena manifest in the realm of material things; that is, they affect entities composed of physical particles. For instance, electricity is a material phenomenon because it concerns the motions of electrons, whereas the soul (if it existed) would be immaterial by definition. Debates over the knowability of how conscious experiences arise, for example, revolve largely around the question of whether the mind is material or immaterial (Chalmers, 1996).

Scope. Whereas phenomena in physics seem highly knowable because they are highly material and reducible, they also seem *unknowable* because they have an extremely wide scope of application. For example, consider the gravitational force. On the one hand, a phenomenon such as gravity that impinges on everything in the universe will present no difficulty to scientists in terms of finding material to test. But on the other hand, the scientific method is intrinsically *comparative* (e.g., Mill, 1967/1843). One cannot, for example, compare those bits of matter that are subject to gravity from those that are not. Perhaps in part for this reason, gravity is still poorly understood among physicists.

Relational factors. Finally, we consider two factors that concern the ways that propositions relate to each other in one's broader system of beliefs:

Belief Centrality. Some propositions are central to our conceptions of the world; if we no longer believed that 2+3=5, for instance, the world would have to be constituted in a wholly different way (if this is even conceivable). To the extent that a proposition is foundational to our other knowledge, that other knowledge can act as evidence in favor of that proposition. To some degree, this mirrors the approach to epistemology taken by rationalist philosophers such as Descartes. In the *Meditations*, Descartes is able to justify knowledge of the external world only on the basis of God's existence (Descartes, 2013/1641).

Although laypeople probably would not endorse rationalist approaches to knowledge justification, they may nonetheless see epistemically central propositions as more certain than less central propositions. For this reason, such propositions may be seen as less falsifiable and, perhaps counterintuitively, as *less* knowable. Put differently, some beliefs are axiomatic or *basic*: They are not susceptible to contradiction from other knowledge sources because those other sources are themselves dependent on that knowledge (a view known as *foundationalism*; e.g., Plantinga, 1981). Yet, if people are both foundationalists *and* falsificationists, then basic beliefs may actually be seen as fundamentally unknowable because they are unfalsifiable.

Value Centrality. Finally, some propositions are not central to one's other *beliefs*, but instead to one's *values*. For some (e.g., Kant, 2002/1785; Lewis, 1952), our moral sense is the most compelling evidence for God's existence. In addition to reasons of epistemic justification, people may also have motivational reasons to believe in propositions that justify their values: "If there is no God," said Jean-Paul Sartre, "then everything is permitted." Whereas Sartre was an atheist, and hence accepted the nihilist conclusion of this conditional, many people who accept the conditional are more likely to infer the contrapositive: Since some things are *not* permitted, God must exist. Such motivated reasoning may relegate highly value-laden beliefs to the realm of unknowability.

The Current Work. On the basis of the idea that laypeople and scientists may share idealized standards for generating and assessing knowledge, we predicted that reducibility and falsifiability would be especially important, as these criteria govern the methodological basis on which knowledge is achieved. The psychology literature makes less clear predictions about the intrinsic or relational factors. One possibility we explore is that the intrinsic factors are relevant insofar as they are relevant to the methodological factors (e.g., material things lend themselves more to reduction and falsification.)

We test these six potential factors in three studies. First, we measure judgments of each factor for various scientific items (e.g., consciousness, economic cycles) in Study 1, and various pseudoscience items (e.g., clairvoyance, karma) in Study 2. Then, Study 3 adopts an experimental approach, attempting to replicate the patterns from Studies 1 and 2 using controlled stimuli.

Studies 1 and 2

In our initial studies, we tested the relationships among the six methodological, intrinsic, and relational factors for a set of 40 scientific and 30 pseudoscientific phenomena.

Method

Participants. We recruited 350 participants from Amazon Mechanical Turk, who were randomly assigned to complete the science items of Study 1 (N = 176) or the pseudoscience items of Study 2 (N = 174).

Items. We developed 40 science items for Study 1 and 30 pseudoscience items for Study 2. For each item, participants were asked to "consider the topic below." The topic was presented in a box, together with a brief parenthetical description. For example, one of the items in Study 1 read "Determinism (i.e., the idea that all events are the results of prior conditions, such as the movement of atoms or previous physical events)" and another read "The determinants of economic cycles (i.e., the causes behind economic recessions, depressions, etc.)." In Study 2, one item read "Clairvoyance (i.e., being able to know things about objects, people, or events without using the senses)" and another read "Karma (i.e., the moral sum of a person's actions that determines his or her future fate)."

Measures. Between-subjects, participants completed a measure of either *knowability*, *reducibility*, *falsifiability*,

materiality, *scope*, *belief centrality*, or *value centrality* (*N* = 50 per condition) for either the 40 science or the 30 pseudoscience items. The scales were phrased as follows:

Knowability. "To what extent do you think it would be possible for scientists to attain a complete understanding of this topic?" from 0 ("Not at all") to 10 ("Completely").

Reducibility. "To what extent do you think a complete understanding of this phenomenon would invoke only facts about the movement of particles and purely physical components?" from 0 ("Not at all material") to 10 ("Very material").

Falsifiability. "Suppose one scientist made a claim about the above topic, and another scientist wanted to design an experiment to find out whether the claim is true or false. To what extent do you think it is possible to design such an experiment?" from 0 ("Not at all possible") to 10 ("Completely possible").

Materiality. "How much do you think the above topic is within the material realm?" from 0 ("Not at all material") to 10 ("Very material").

Scope. "Consider everything in the universe. How many of these things does the above topic apply to?" from 0 ("Very few") to 10 ("Very many").

Belief Centrality. "Consider all the things you could know about the topic above. If all of this knowledge turned out to be wrong, how many other things that you know would also turn out to be wrong?" from 0 ("Very few") to 10 ("Very many").

Value Centrality. After reading each topic description, participants were asked: "How important is the above topic to your moral beliefs or values?" from 0 ("Not at all important") to 10 ("Very important").

	K	R	F	M	S	BC
R	.44***	—				
F	.84***	.10	—			
M	.75***	.47***	.61***	—		
S	–.07	.63***	–.34**	.01	—	
BC	.22º	.22º	–.02	.19	.52***	—
VC	.39***	.27*	.12	.38**	.47***	.79***
C	.33**	.26*	.06	.31**	.52***	—

º < .10 * < .05 ** < .01 *** < .001

Note. Entries correspond to knowability (K), reducibility (R), falsifiability (F), materiality (M), scope (S), belief centrality (BC), value centrality (VC), and a composite measure averaging belief and value centrality (C).

Table 1: First-order correlations (combined dataset).

Results

For each item, we calculated the mean score on each judgment, across all participants making that judgment for the item. Analyses are all at the item level. (However, Study 3 will rely on a subject-level analysis.)

Correlations. The first-order correlations among the measures are given in Table 1.

Both methodological factors were associated with knowability. There was a moderately strong relationship between knowability and reducibility, r = .44, such that phenomena were seen as more knowable when reducible to fundamental physics. There was also a very strong relationship between knowability and falsifiability, r = .84, such that phenomena were seen as more knowable to the extent that disagreements could be resolved in terms of empirical tests. These correlations are both consistent with popular views among scientists and philosophers of science. Interestingly, falsifiability and reducibility were almost completely uncorrelated, r = .10, suggesting that people conceptualize these as independent dimensions.

The intrinsic factors also seem to have a relationship with knowability and with the methodological factors. Materiality was a consistent predictor not only of knowability, r = .75, but also of reducibility, r = .47, and of falsifiability, r = .61. One possible interpretation of this pattern is that more material things are seen as more knowable *because* they are more reducible and more subject to empirical falsification. We test this possibility using regression analyses below.

Although scope was not associated with knowability, r = –.07, it *was* associated with both falsifiability and reducibility: Phenomena that applied to very many things were seen as *more* reducible, r = .63 (perhaps because the most general phenomena tend to be lower on the reductionist hierarchy), but as *less* falsifiable, r = –.34 (perhaps because there are few opportunities to observe the absence of something that is ubiquitous). Thus, scope may have an *indirect* relationship with knowability, via these two opposite pathways (see regressions below).

The results for the relational factors are somewhat less clear. Because these measures were highly correlated with each other, r = .79, we averaged them to form a composite *Centrality* measure. This measure was a significant predictor of knowability, r = .33, although the reason is less clear from the data. We had hypothesized that highly central beliefs might be seen as less falsifiable (and hence *less* knowable), but this was not borne out by the data. We further analyze and discuss this effect below.

Regression Models. Table 2 shows regression coefficients for a model predicting knowability judgments from just the methodological factors (reducibility and falsifiability) in Step 1, and from all factors in Step 2. (The composite centrality measure was used to avoid multicollinearity.) This reveals that the methodological factors (reducibility and falsifiability) were strong predictors of knowability (b = 0.58 and 0.66, respectively, in Step 2), even after adjusting for the other predictors. Although reducibility and falsifiability seem to completely screen off materiality (b = 0.09, *ns*), there does seem to be a residual effect of scope (b = –0.27), such that wider scope phenomena were seen as less knowable, even after accounting for the other factors.

Finally, centrality seems to be associated with knowability, such that more phenomena more central to beliefs and values are seen as more knowable (b = 0.45).

	DV: Knowability	
	Step 1	**Step 2**
R	.45 (.06)***	.58 (.09)***
F	.83 (.05)***	.66 (.06)***
M		.09 (.10)
S		–.27 (.07)***
C		.45 (.08)***
R^2	.827	.898

Table 2: Unstandardized *b*s and SEs (combined dataset)

One concern about these analyses, however, is that they collapse across the science and pseudoscience items, which may differ in a variety of ways. Indeed, even at a gross level, the means significantly differed between these item sets on all measures (Table 3).

	Study 1 (Science)	**Study 2 (Pseudoscience)**
K* **	5.90 (1.16)	2.91 (1.16)
R* **	5.10 (1.50)	3.41 (0.82)
F* **	5.97 (1.70)	4.22 (1.46)
M* **	5.22 (1.01)	3.59 (0.73)
S*	4.95 (2.39)	3.95 (1.43)
BC* **	5.54 (1.19)	4.43 (0.86)
VC* **	5.51 (1.21)	3.44 (1.19)

Table 3: Descriptive statistics (Studies 1 and 2)

Thus, correlations between knowability and the other measures could potentially be driven by gross differences between item sets rather than meaningful differences among items. To address this concern, we repeated the regression of knowability judgments on the predictors, separately for Studies 1 and 2 (Table 4).

	DV: Knowability	
	Study 1 (Science)	**Study 2 (Pseudoscience)**
R	.45 (.13)**	.47 (.18)*
F	.68 (.07)***	.35 (.10)**
M	–.23 (.14)º	.46 (.27)
S	–.25 (.09)**	–.03 (.16)
C	.16 (.10)	–.34 (.26)
R^2	.823	.848

Table 4: Unstandardized *b*s and SEs (Studies 1 and 2)

This analysis revealed that even *within* each item set, the relationship between reducibility and falsifiability was robust. There was again no consistent relationship with materiality once the other factors are accounted for, and the relationship with scope held up only for the science items. The most substantial divergence between the combined and more fine-grained analyses was for centrality: Whereas centrality was a significant positive predictor on the combined dataset, it does not have any consistent relationship with knowability within either dataset. This suggests that this effect was spurious, caused by differences across (rather than within) datasets.

We can also use these data to look at what predicts the methodological factors themselves, using regressions to predict reducibility (Table 5) and falsifiability (Table 6).

	DV: Reducibility	
	Step 1	**Step 2**
F		.07 (.08)
M	.57 (.10)***	.63 (.12)***
S	.45 (.06)***	.59 (.07)***
C		–.38 (.10)***
R^2	.606	.678

Table 5: Model predicting *reducibility* (Studies 1 and 2).

	DV: Falsifiability	
	Step 1	**Step 2**
R		.16 (.18)
M	.92 (.13)***	.78 (.19)***
S	–.30 (.08)***	–.43 (.14)**
C		.16 (.17)
R^2	.496	.505

Table 6: Model predicting *falsifiability* (Studies 1 and 2).

The intrinsic factors both had large and consistent influences on the methodological factors. Materiality had a strong, positive effect on both reducibility and falsifiability judgments (b = 0.63 and 0.78, respectively). This indicates that the strong correlation between materiality and knowability is due to the effects of materiality on reducibility and falsifiability, rather than any direct effect on knowability.

Consistent with the first-order correlations, scope had a strongly positive effect on reducibility but a strongly negative effect on falsifiability (b = 0.59 and –0.43, respectively), indicating *opposite* indirect effects on knowability. That is, to the extent that a phenomenon applies to many things in the universe, it is seen as more reducible to physics but *less* falsifiable. These two pathways seem to largely cancel out, leading to a weak relationship between scope and knowability overall.

Centrality appears to have a negative association with reducibility (b = –0.38), after adjusting for the other factors, such that phenomena more central to one's beliefs and values tend to be less reducible to physics. However, centrality seems more likely to be an *effect* of reducibility than a *cause*, particularly because centrality has no clear relationship with knowability, whereas reducibility does. Centrality also has no relationship with falsifiability (b = 0.16, *ns*), consistent with the idea that centrality is not a key driver of knowability.

Discussion

Overall, these results paint a consistent picture of what determines knowability. Across many ways of looking at the data, the methodological factors of reducibility and falsifiability play a large role, accounting for the majority of the variance in knowability. The intrinsic factors (scope and materiality) appear to have indirect effects on knowability via reducibility and falsifiability, but little or no direct effect. Finally, centrality appears to be associated with reducibility, but is more likely an effect than a cause. These folk-scientific methods appear to closely mirror the way that professional scientists and philosophers think about what is knowable.

Despite these consistent findings across very different sets of items, and despite the good model fits (accounting for 80–90% of the variance in knowability, and 50–70% of the variance in reducibility and falsifiability), these results are correlational. Thus, conjectures about the direct and indirect causal relationships are difficult to assess. We therefore adopted an experimental approach for converging evidence.

Study 3

In Study 3, we used artificial stimuli to look more directly at the relationships of the six factors with knowability. We would predict that the methodological factors should have strong relationships with knowability, consistent Studies 1 and 2. We would also expect a relationship with at least materiality (given its positive effects on both reducibility and falsifiability) and perhaps scope (given its positive effect on reducibility and negative effect on falsifiability). We would not predict a relationship with belief or value centrality, since these factors did not seem to play a causal role in Studies 1 and 2.

Method

Participants. We recruited 100 participants from Amazon Mechanical Turk; 1 was excluded from analysis due to missing data.

Items. Participants completed 12 novel items (4 concerning physical phenomena [e.g., nerium force], 4 concerning biological phenomena [e.g., zenilan synthesis], and 4 concerning psychological phenomena [e.g., perception cavelation]). For each item, participants were told that "There are many unknown phenomena in our universe that scientists are trying to understand" and that (for example) "Currently, the biological phenomenon of *zenilan synthesis* is not understood." After a brief description of the item (varying across conditions; see below), participants were asked to rate "To what extent do you agree with the claim that someday in the future, scientists will understand everything there is to know about zenilan synthesis?" on a scale from 0 ("Strongly Disagree") to 10 ("Strongly Agree").

Manipulations. There were 12 within-subjects conditions in a 6 (factor: Reducibility, Falsifiability, Materiality, Scope, Belief Centrality, Value Centrality) x 2 (High, Low) design. (We refer to these conditions by the initial letter of the factor, with a '+' or '–' sign; e.g., high materiality is denoted 'M+'.) Only the target factor was described for each item. Each description began with "Although zenilan synthesis is not understood, it is widely believed that…" followed by phrasing varying across conditions. In the R+/R– conditions:

> …a complete understanding would invoke [only facts about / facts that cannot be reduced to] the movement of particles and purely physical components.

In the F+ and F– conditions:

> …if a scientist wanted to make a claim about it, it would [*not*] be possible for another scientist to design an experiment to find out whether the claim is true or false.

In the M+ and M– conditions:

> …it is [*not*] within the material realm.

In the S+ and S– conditions:

> …it [applies to many / does not apply to most] things in the universe.

In the BC+ and BC– conditions:

> …if everything you could know about zenilan synthesis was wrong, very [many / few] other things that you know would also be wrong.

In the VC+ and VC– conditions:

> …it is [very / not at all] important to many moral beliefs and values.

The 12 conditions were balanced with the 12 items for each participant. The '+' and '–' versions of each factor were always adjacent in the presentation order and matched for domain (physical, biological, or psychological). The order of the 6 factor conditions was randomized, as was the order of the '+' and '–' versions.

Results and Discussion

The results were consistent with the patterns of influence uncovered in Studies 1 and 2 (see Table 7). Propositions high in reducibility and falsifiability were rated more knowable than propositions low on these factors [$t(98)$ = 4.42, $p < .001$ and $t(98)$ = 7.34, $p < .001$], consistent with the direct effects of these factors in Studies 1 and 2. In addition, propositions corresponding to phenomena that were highly material and wide in scope were rated more knowable than propositions low on these factors [$t(98)$ = 8.01, $p < .001$ and $t(98)$ = 5.58, $p < .001$], consistent with the indirect effects of these factors uncovered in Studies 1

and 2, via reducibility and falsifiability. However, centrality to beliefs and values did not influence knowability judgments [$t(98) = –0.52$, $p = .61$ and $t(98) = 1.32$, $p = .19$], again consistent with Studies 1 and 2, where these factors appeared to be correlated with, but not causally antecedent to, knowability.

	DV: Knowability	
	+	–
R ***	6.05 (2.45)	5.19 (2.53)
F ***	6.54 (2.52)	4.30 (2.94)
M ***	6.45 (2.20)	4.16 (2.81)
S ***	6.30 (2.48)	4.88 (2.69)
BC	5.28 (2.94)	5.37 (2.56)
VC	5.82 (2.68)	5.50 (2.67)

Table 7: Means (SDs) in Study 3.

These results are consistent with the picture that knowability is determined directly by methodological factors, and indirectly by intrinsic factors. One concern is that we cannot fully rule out the possibility that relational factors (belief or value centrality) also have an influence on perceived knowability, because merely *describing* a proposition as central would not manipulate that proposition's *actual* centrality relative to one's other beliefs (especially given that these stimuli were artificial). That said, our manipulations *were* strong enough to produce effects of the other factors. In conjunction with the naturalistic (but correlational) results of Experiments 1 and 2, we think it unlikely that centrality plays a key role in perceived knowability for most phenomena.

General Discussion

When is a proposition knowable (even if unknown), or a phenomenon understandable (even if not understood)? Consistent with the idea that people act as intuitive scientists, people adopt some of the same epistemological principles as philosophers and practitioners of science: Propositions are knowable to the extent that they can be reduced to more fundamental facts and to the extent that they are subject to empirical falsification. These methodological factors depend in turn on facts about the intrinsic physics or metaphysics of the underlying phenomena: They are seen as more reducible to the extent that they are material and wider in scope, and as more falsifiable to the extent that they are material and narrower in scope. Although relational factors (centrality to one's beliefs and values) were sometimes associated with knowability (and also with some of the other factors), these associations do not appear to be causal. We nonetheless regard the issue of belief and value centrality (both their causes and their consequences) as an intriguing direction for future work.

We are currently building on these findings in several ways. First, we are expanding on Studies 1 and 2 to test other factors, such as domain (an intrinsic factor), that might be associated with knowability and might also help to explain some of the variance in the other intrinsic and methodological factors. Second, we are expanding on Study 3 to further test the relationships among factors (e.g., the indirect effects via falsifiability and reducibility). Third, we are looking at individual differences that might predict perceptions of knowability (e.g., religiosity, trust in science, and need for cognitive closure). Finally, we are testing real-world applications of these findings, such as ways to frame real phenomena so as to influence their perceived knowability and potential downstream consequences.

References

Burke, N., Johnson, S.G.B., & Keil, F.C. (2016). *Folk reductionism*. Manuscript in preparation.

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

Chalmers, D. (1996). *The conscious mind*. Oxford University Press.

Dennett, D.C. (1991). *Consciousness explained*. Boston, MA: Little Brown.

Descartes, R. (2013/1641). *Meditations on First Philosophy*. Cambridge University Press.

Gopnik, A. (1998). Explanation as orgasm. *Minds & Machines*, *8*, 101–118.

Gopnik, A., Meltzoff, A.N., & Kuhl, P. (1999). *The scientist in the crib.* New York, NY: HarperCollins.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015). *Sense-making under ignorance*. Manuscript under review.

Kant, I. (2002/1785). *Groundwork for the metaphysics of morals*. Yale University Press.

Keil, F.C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*, 368–373.

Lewis, C.S. (1952). *Mere Christianity*. New York, NY: HarperCollins.

McAdams, D.P. (1993). *The stories we live by*. New York, NY: Guildford Press.

Mill, J.S. (1967/1843). *A system of logic ratiocinative and inductive*. London, UK: Longmans.

Plantinga, A. (1981). Is belief in God properly basic? *Noûs*, *15*, 41–51.

Preston, J., & Epley, N. (2009). Science and God: An automatic opposition between ultimate explanations. *Journal of Experimental Social Psychology*, *45*, 238–241.

Popper, K. (1959/1934). *The logic of scientific discovery*. London, UK: Routledge.

Putnam, H. (1973). Reductionism and the nature of psychology. *Cognition*, *2*, 131–146.

Strevens, M. (2008). *Depth*. Harvard University Press.

White, P.A. (2009). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, *108*, 3–18.

A scaleable spiking neural model of action planning

Peter Blouw (pblouw@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Bryan P. Tripp (bptripp@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada N2L 3G1

Abstract

Past research on action planning has shed light on the neural mechanisms underlying the selection of simple motor actions, along with the cognitive mechanisms underlying the planning of action sequences in constrained problem solving domains. We extend this research by describing a neural model that rapidly plans action sequences in relatively unconstrained domains by manipulating structured representations of objects and the actions they typically afford. We provide an analysis that indicates our model is able to reliably accomplish goals that require correctly performing a sequence of up to 5 actions in a simulated environment. We also provide an analysis of the scaling properties of our model with respect to the number of objects and affordances that constitute its knowledge of the environment. Using simplified simulations we find that our model is likely to function effectively while picking from 10,000 actions related to 25,000 objects.

Keywords: planning; affordances; spiking neurons; neural engineering framework; semantic pointer architecture

Introduction

Action selection is a topic of long-standing interest for understanding human behavior (Shallice, 1982; Daw et al., 2005). Recent results in neurophysiology have clarified some of the underlying neural mechanisms. In particular, moment-to-moment decisions about motor actions such as reaching, grasping, saccades, etc. appear to arise from parallel competitions among representations in multiple frontal and parietal areas (Andersen & Gui, 2009; Cisek, 2012).

Accordingly, action decisions are influenced by many factors (Wolpert & Landy, 2012). Often, a dominant factor is a larger goal that may be several steps removed from any immediately feasible action (for example, one may have to open a laptop and start an application before one can type something). Thus, rapid and frequent multi-action planning is an important part of life. Planning of action sequences has been studied extensively in the context of the Tower of Hanoi and Tower of London tasks, which appear to particularly involve frontal areas (Goel & Grafman, 1995), but (in contrast with decisions about immediate actions) the neural mechanisms are unclear.

Furthermore, compared to more naturalistic tasks (such as making a sandwich) these tasks are arguably relatively deliberative and constrained. In contrast, preparing a meal, cleaning up after it, fixing a loose hinge discovered on the cupboard in the process, etc. require variable (sometimes novel) action sequences that are assembled with little effort, using sophisticated knowledge of the many objects involved and the actions they afford. Human behavior in such contexts is only loosely related to artificial-intelligence planning methods (Russell & Norvig, 2003), which produce complex and optimal plans within narrow domains.

Our goal in this study is to better understand how the knowledge needed for rapid action-sequence planning might be stored and processed in the brain. To this end, we develop a spiking-neuron model that plans action sequences by chaining together action preconditions and effects (Fikes & Nilsson, 1971) while interacting with a simulated environment. Each planning step selects from actions that are related to available objects, in order to constrain each decision and allow planning to proceed quickly (about 100ms of simulated time per step).

While this kind of planning process could in principal be supported by a variety of action representations and model architectures (Krüger et al., 2011; Fincham et al., 2002), some architectures and representations require precision than is unavailable from noisy, spiking neurons, and/or require very large numbers of neurons. A working spiking neural model may therefore be a source of insight into constraints on the brain's solution to this problem.

In what follows, we present our model and analyze its performance on a naturalistic planning challenge (namely, boiling water in a simulated kitchen environment). We contend that the model satisfies two important constraints on the processes underlying action selection. First, the model gives a neurally plausible account of the kinds of representations and processes that underlie planning in cognitive systems. And second, the model can scale to accomplish goals that require performing fairly complex sequences of actions in domains that require an understanding of large numbers of objects.

The Semantic Pointer Architecture (SPA)

To implement our model, we use the Semantic Pointer Architecture (Eliasmith, 2013), a recently developed framework for constructing neurally plausible models of cognitive phenomena. Previously, the SPA has been used to build Spaun, a large-scale simulation of the brain that performs a wide range of cognitive functions (Eliasmith et al., 2012). In what follows, we provide a condensed description of the SPA drawn from material found in Stewart et al. 2014.

A typical SPA model defines a set of subsystems corresponding to particular brain regions. Each subsystem is implemented as a collection of simulated spiking neurons (we use the leaky-integrate-and-fire (LIF) model in this case). Synaptic connections between the neurons in distinct subsystems are then optimized to compute some desired function on a latent vector space that is represented by the neurons'

spiking activities.

For example, a common subsystem is a working memory. Formally, working memory can be described as a differential equation. The neurons in a working memory subsystem can be characterized as representing a vector (or scalar) x, and the input to these neurons can be taken to represent another vector, u. Assuming that x remains constant when u is zero (i.e. a memory is stored), and that x changes proportionally to value of u when u is non-zero, the representational state of the subsystem can be written as $dx/dt = u$.

Helpfully, an arbitrary differential equation of this kind can be approximated with an ensemble of spiking neurons using the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). This approximation is achieved by randomly assigning a 'tuning curve' to each neuron that specifies its spike rate in relation to the represented value x. For instance, a given neuron might fire rapidly when the value of x is zero, but fire much more slowly when the value of x is positive. For a given neuron in a SPA model, this tuning curve is assigned randomly using a distribution of firing patterns that is consistent with available empirical evidence.

Once synaptic connections are introduced between two populations of neurons, it is possible to use a local optimization technique to ensure that the tuning curves in each population are appropriately related to the variables they are hypothesized to represent. For example, if the tuning curves in the first population are defined in relation to a variable x, while the tuning curves in the second population are defined in relation to a variable y, then it is possible to find a set of connection weights that approximate the computation $y = f(x)$. In the presence of recurrent connections, this technique can be used to approximate any function of the form $dx/dt = f(x,u)$. The quality of the approximation depends on both the number of neurons being used and the degree of non-linearity present in the function.

In general, the SPA suggests that the representations being manipulated by the brain are semantic pointers (SPs), which are compressed neural representations that can be identified with vector variables such as x and y above. In the context of action planning, however, we need to manipulate symbol-like, structured information using SPs. Consequently, the forms of compression that are most relevant are identified by Vector Symbolic Architectures (VSAs; Gayler, 2004). VSAs are a set of mathematical formalisms that enable structured collections of symbols to be represented as high-dimensional vectors. For example, a symbol corresponding to the concept of $KETTLE$ could be defined in a VSA as a 500-dimensional vector (i.e. a distributed representation). These vectors can be randomly chosen (as they are here), or they can be chosen such that similar terms ($KETTLE$ and POT, for example) correspond to similar vectors.

To encode structured combinations of vectors, VSAs introduce a compressive binding operation. Different VSAs choose different binding operators, and for our work we use circular convolution, written as $\otimes$, meaning that this particular VSA uses holographic reduced representations (HRRs; Plate, 2003).

To give an example of how structured information is encoded using this operator, suppose we want to represent the knowledge that kettles tend to be used to boil water and tend to be located in kitchens and staff lounges. We might represent this as: $KETTLE = LOCATION \otimes (KITCHEN + STAFF_LOUNGE) + GOALS \otimes WATER_BOILED$. Importantly, VSAs also define an inverse operation: given an element of the structure, we can determine the associated representations by computing, e.g., $KETTLE \otimes LOCATIONS^{-1}$, which is approximately equal to $KITCHEN + STAFF_LOUNGE$.

These VSA operations can be computed within the SPA to create structured semantics pointers. Moreover, the SPA allows such semantic pointers to be routed between different subsystems and manipulated in various ways. For instance, sensory areas can transform stimuli into appropriate conceptual SPs, and motor areas can take SPs representing actions and transform them into a sequence of muscle movements (Eliasmith, 2013). In order to implement these kinds of transformations, the SPA includes a model of the cortex-basalganglia-thalamus loop that performs (cognitive) action selection. Connections between cortex and the basal ganglia compute utilities over a set of possible actions. The basal ganglia identify the highest utility value, and pass this information on to the thalamus, wherein all of the neurons corresponding to actions with lower utility values are suppressed.

Models constructed using the SPA therefore define a set of cognitive actions that can be performed (note that these are distinct from the physical actions discussed throughout the rest of this paper). Each action is defined in terms of a set of effects E_i (e.g. the routing of an SP from one subsystem to another), and a utility function U_i that indicates when the action ought to be performed. For example, the following action results in the contents of a subsystem labeled *ultimate_goal* being routed to a subsystem labeled *immediate_goal* when a subsystem labeled *signal* represents an SP *PLAN*. This action will be selected by the basal ganglia if it has the highest utility of all actions.

$$U_i : signal \cdot PLAN$$
$$E_i : immediate_goal \leftarrow ultimate_goal$$

The NEF provides an efficient method for defining these actions, with each one requiring roughly 300 basal ganglia neurons to implement (Stewart et al., 2014).

A Neural Action Planning Model

Given that the set of actions needed to accomplish a goal depends on the state of the world, and that such actions modify the state of the world once performed, we define a simple simulated environment for our model to interact with. This environment consists of an arbitrary number of entities that have particular states and locations. For example, a 'ket-

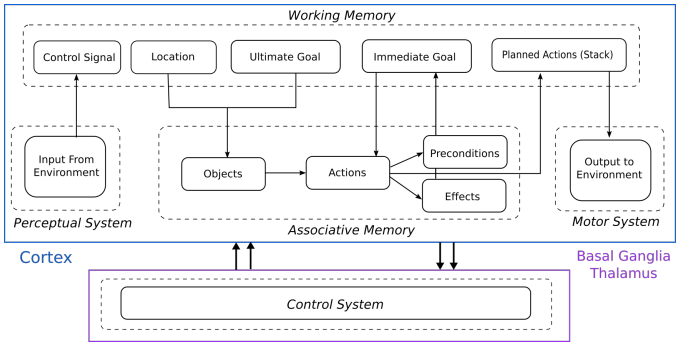


Figure 1: Functional architecture of the model. Regions surrounded by the dashed lines correspond to broadly individuated components responsible for perception, action, memory and cognitive control. Regions surrounded by solid black lines correspond to potentially simpler subsystems of these components. Arrows between subsystems indicate the main pathways through which the control system is able to manipulate the flow of information.

tle' entity might be located on a counter, and might have the state of being plugged in. At an implementational level, the environment is a program that the planner interacts with by sending commands that instruct the program to either return the state of an entity or execute an action. Importantly, actions can only successfully modify the state of the environment in particular circumstances. For instance, the action *BOIL_KETTLE* can only be successfully executed if the kettle contains water and is plugged in.

The planner itself comprises five main components: a perceptual system that monitors the the environment, a motor system that manipulates the environment, a working memory system that stores goals and planned sequences of actions, an associative memory system that matches certain locations and goals to certain actions and objects, and a control system that controls how all of these components interact with one another. Each component is implemented using the SPA, as described above. Figure 1 provides a high-level depiction of the model's functional architecture.

Functionally, the model is provided with input in the form of SPs representing a location and an "ultimate goal" (or "prior intention", Jeannerod, 2006), along with an input that signals the model to start planning. The location and goal representations are mapped by an associative memory to an SP representing a set of objects that are relevant to accomplishing the given goal in the given location. For example, given the location *KITCHEN* and the goal *WATER_BOILED*, the model will represent the objects *TAP* and *KETTLE* as being relevant.

Next, this object representation and the goal representation are mapped by an associative memory to a representation of an appropriate action, which is then stored in working memory as the final item in the action plan being constructed. Here, the action would be *BOIL_KETTLE*, which, simplifying somewhat, would be added to the plan via the following cognitive action:

$$U_i : control_signal \cdot GET_ACTION$$
$$E_i : stack \leftarrow stack \otimes PUSH + action$$

where *PUSH* is a random SP that is used to bind action SPs to particular positions in a structured representation of an action sequence.

Since the action can only be executed in specific circumstances, an associative memory is used to map it to a representation of a set of preconditions that the environment must satisfy. The immediate goal of planning is then updated by adding in these preconditions and subtracting out the effects of the planned action. This operation is performed by a cognitive action of the form:

$$U_i : control_signal \cdot SET_GOAL$$
$$E_i : i_goal \leftarrow i_goal - effects + precons$$

where *i_goal* and *precons* abbreviate the subsystems labeled 'Immediate Goal' and 'Preconditions', respectively.

This whole process of finding an appropriate action and updating the immediate goal of planning is repeated until an action whose preconditions are satisfied by the environment is found. At this point, perceptual feedback from the environment prompts the model to begin executing the actions in the planned sequence. First, the most recently added action in the planned sequence is routed to the planner's motor system, where it is subsequently presented as a command to the simulated environment. Then, the action just executed is removed from the planned sequence and the process repeats (i.e. the next action in the sequence is routed to the motor system etc.):

$$U_i : control_signal \cdot POP_STACK$$
$$E_i : stack \leftarrow stack \otimes PUSH^{-1} - motor$$

The representation of the action sequence in working memory is the VSA equivalent of a stack, and the the process of executing actions amounts to popping items off of this stack and routing them through the motor system as commands to the environment. However, because the SPA makes use of a compressive binding operator, the stack is not perfect (Plate, 2003). As more actions are added to a planned sequence, the likelihood of recovering all of the actions drops considerably. For this reason, our model is designed to execute all of the actions that can be recovered from the planned sequence, after which it begins to re-plan in light of the changed environment.

The fact that the model is able to switch from acting to planning in this manner is important for enabling it to recover from errors. For example, if the planner chooses an incorrect action representation at a given point in the planning process, or fails to execute items in the correct order, it will then re-plan and correct its mistake. The process of re-planning in this way is typically successful because the model relies on

the perceptual feedback from the environment when deciding whether or not the preconditions of a particular action are satisfied.

We emphasize that the model does not directly implement the above logic or symbolic variables. The model consists of 341380 spiking neurons, with synaptic weights that are optimized to approximately perform this information processing. An implementation of the model is available online at https://github.com/pblouw/action-planning/

Results

In what follows, we report results concerning the consistency with which our model is able to successfully perform a task over numerous simulations involving different neuron parameters. We also report results concerning the scalability of the model to situations involving large numbers of possible objects and actions.

First, however, we discuss the results of an example simulation to provide greater insight into the behavior of the model. As shown in Figure 2, the model plans by chaining backwards through a sequence of actions and continually updating its immediate goal. Once an action with preconditions satisfied by the environment is added to the plan, the model ceases planning and begins acting (this can be observed in the change to the Control Signal representation in Figure 2). Once no more actions can be extracted from the current plan, the model stops acting and proceeds to re-plan until it achieves its ultimate goal.

Goal Completion Analysis

We first test the model by assessing whether or not it is able to consistently accomplish its ultimate goal when allowed to interact with the environment. Each row of Table 1 reports the results of an experiment in which the model's behavior is simulated for up to 4 seconds over 50 trials. Each experiment involves setting the environment to an initial state that requires the planner to perform an increasing number of actions. For example, in the first experiment, only two actions need to be performed to accomplish the ultimate goal of boiling water. Each trial involves a unique instance of the model, in that a new random seed is used to generate all of the LIF neuron parameters (each instance of the model is accordingly analogous to a unique experimental subject). A trial is deemed successful if the model achieves the ultimate goal within the simulation time. We report the percentage of successful trials, along with the average time needed per trial to achieve the goal. The time needed to achieve the goal typically varies due to the differences in the number of actions the model is able to retrieve from a planned sequence. Occasionally, the model also makes an error that forces it to re-plan so as to accommodate an unexpected environmental state. Note that the environment model changes instantaneously, so these time scales correspond only to the neural model's internal processing time. Overall, our results indicate that the model is robust to changes in the environment that increase the number of actions required to achieve the goal.

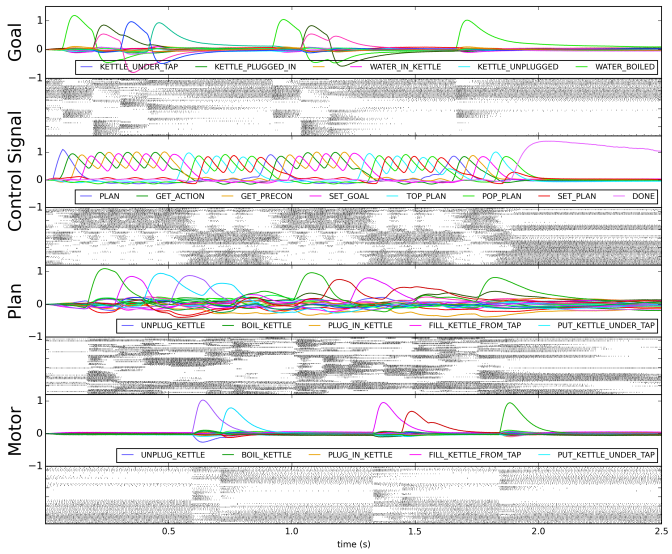


Figure 2: A sample run of the model interacting with the environment. Each colored plot depicts the similarity between the representational state in the indicated component of the model and a set of known representational states, shown under each plot. Underneath each representational plot is the spiking activity from a set of randomly sampled neurons from the labeled model component.

Table 1: Performance Analysis of Planner over 50 Trials

Sequence Length	Success Rate (%)	Average Time (s)
2	94	0.48 (SD 0.17)
3	98	0.90 (SD 0.20)
4	94	1.40 (SD 0.51)
5	94	1.92 (SD 0.64)

Scaling Analysis

Given these successful results, we decided to test whether the action-selection mechanism of this model would work with human-scale knowledge bases. Our model selects an object or action from an associative memory if its SP has a high inner product with an SP that represents the current context (e.g. location; goal). An unrelated item is unlikely to be accidentally selected, because random SPs in high-dimensional spaces tend to be nearly orthogonal. However, if the memory contains a very large number of entries, the total probability of an incorrect selection may become problematic. We explored this issue in simplified HRR models with no neurons.

Figure 3 illustrates how such confusion can arise. An associative memory was created with 25000 500-dimensional object HRRs, each of which contained lists of goal and location vectors of varying length (drawn from a Poisson distribution with a mean of 2). The associative memory was then queried with random goal/location combinations. The inner products

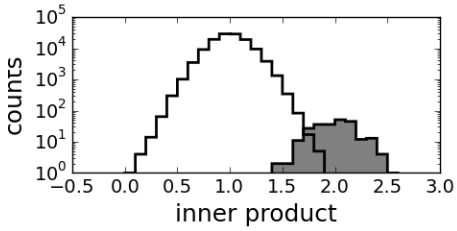


Figure 3: Inner products of 500 random goal/location context HRRs with 25000 object HRRs (each with two goals and locations on average). Histograms of inner products of queries with partially matching (open), and matching (filled) object vectors. There is some overlap between partial and full matches due to the large number of partial matches (note the log scale). The non-matches (not shown) have a mean of zero and do not overlap the matches.

of these queries with the associative memory entries was distributed around two for correct matches, i.e. for entries that actually contained both the specified goal and the specified location (bottom panel), and around zero for entries that did not contain either the goal or location (top panel), with no overlap. However, there was some overlap between matches and partial matches (which contained either the goal or location but not both). Because there are a large number of partial matches, substantial confusion could arise if a small fraction of them overlap.

Figure 4 estimates the effect of these matching properties on our model. The top panel shows average numbers of objects (out of 25000) that match random combinations of locations and goals (as a function of the mean numbers of goals and locations per object, which were set equal). Our model would not perform well if this number rose above about ten, because the inner product of an HRR x with a sum that includes x becomes noisy if the sum is large (Plate, 2003). In these simulations, locations were drawn from a list of 250 and goals from a list of 1000. The numbers of objects, goals, and locations were meant to correspond roughly to the numbers of these things that are familiar to a person. No explicit category structure is imposed on the object representations, though the fact that these representations draw on a shared stock of locations and goals yields varying degrees of representational similarity.

The center panel shows the precision, i.e. (true positives) / (true positives + false positives), of object selection with various HRR dimensions, over 500 random queries in each condition, with the threshold set below at least 90% of the true positives. Higher HRR dimensions improve precision.

Finally, the bottom panel shows the precision of action selection (with 10000 possible actions), using queries that combined desired effects (from 2500 possibilities) with sums of object vectors (including false positives) from the object queries. The precision is fairly high with higher-dimensional vectors. The consequence of an error would be to plan ei-

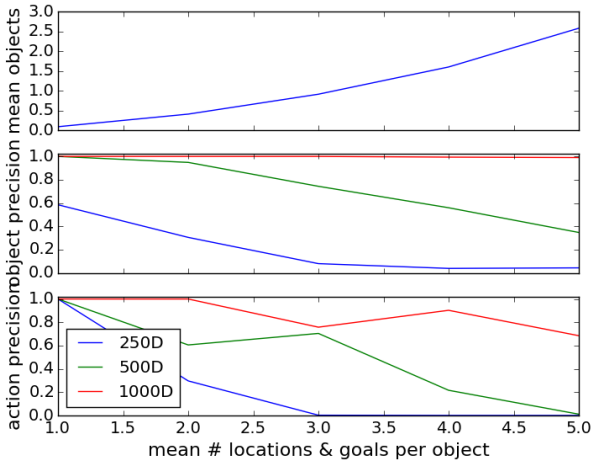


Figure 4: The scaling effects of the average numbers of locations and goals per object concept. Top: Average numbers of objects matching goal/location queries. Middle: Precision of object selection for HRRs of 250, 500, and 1000 dimensions. Bottom: Precision of action selection based on single effects and sums of object vectors from the queries in the middle panel (including false positives). Large HRR vectors are required for high precision with large numbers of object, goal, etc. concepts.

ther an irrelevant action or an action that requires an object that isn't available. One likely reason precision is lower for action selection than object selection is that multiple objects are selected in the first step, and their sum has a somewhat noisy inner product with the individual-object components of action vectors in the second step.

Overall, these results demonstrate that the neural model is likely to function effectively while picking from 10,000 actions related to 25,000 objects. Preliminary simulations indicate the neural model is able to scale to at least 5000 objects with no loss of performance in the kettle boiling scenario. However, additional neural simulations are needed to fully confirm the scaling analysis provided here.

We also found that somewhat greater precision could be obtained by restructuring action queries to include effects and locations, rather than effects and objects. To obtain these results, we built a memory of 10000 actions, each with one effect and a Poisson random number of locations. In this case we performed each query by randomly selecting an action with one or more locations, and querying with its effect and its first location (so there was always at least one correct match). The precision of these queries was somewhat better than that of the affordance-based queries.

This simple analysis does not account for a number of factors, including (for example) correlations between goals and locations. However, it suggests that the action-selection mechanism of our spiking model is likely to scale to relatively

complex environments.

Discussion

The main contribution of our work is to present a neurally plausible model of relatively domain-general action planning. The model is able to plan both quickly and effectively, and it is robust to various changes to the planning environment. The representational properties of the model, moreover, indicate that it is capable of scaling to naturalistic planning environments in which there are vast numbers of potential actions that could be relevant to accomplishing a particular goal. This is an important feature given the degree to which many existing models in action planning literature are restricted to highly specific problem domains (e.g. the Tower of Hanoi puzzle). Interestingly, the scope of our model is closely related to deficits in ideational apraxia (Zadikoff & Lang, 2005).

One potential concern about our model is that it is only able to produce short sequences of 1-3 actions before re-planning (i.e. the stack that stores planned sequences of actions degrades easily). This seems unrealistic in the context of a task such as water-boiling, which most people do with minimal conscious reflection. Some kind of sequence consolidation for routine actions is clearly needed (Taatgen & Lee, 2003; Cooper et al., 2014). However, we are not arguing that the planning performed by our model is reflective of conscious deliberation, in which case the time course over which planning occurs is not implausible.

Finally, given some of the limitations of our model, an important direction for future work concerns grounding it in a richer environment. One possibility would involve integrating the model into a robot that operates in a real kitchen. Among the many practical challenges this would entail, a key problem concerns recognizing object states using a model of the visual system.

Acknowledgments

This work was supported by a NSERC Discovery Grant and an Ontario Graduate Scholarship. We thank Terry Stewart, Ashley Kleinhans, Renaud Detry, Benjamin Rosman, Nasim Mortazavi, and Serge Thill for helpful discussions.

References

Andersen, R., & Gui, H. (2009). Intention, action-planning, and decision making in parietal-frontal circuits. *Neuron*, *63*(5), 568-583.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology*, *22*(6), 927-936.

Cooper, R., Ruh, N., & Mareschal, D. (2014). The goal circuit model: A hierarchical multi-route model of the acquisition and control of routine sequential action in humans. *Cognitive Science*, *38*, 244–274.

Daw, N., Niv, T., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Eliasmith, C. (2013). *How to build a brain: An architecture for neurobiological cognition*. Oxford University Press.

Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press.

Eliasmith, C., Stewart, T., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202- 1205.

Fikes, R. E., & Nilsson, N. J. (1971). STRIPS : A New Approach to the Application of . Theorem Proving to Problem Solving '. *Artificial Intellegence*, *2*, 189 – 208.

Fincham, J., Carter, C., van Veen, V., Stenger, A., & Anderson, J. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(5), 3346–3351.

Gayler, R. W. (2004). Vector Symbolic Architectures Answer Jackendoffs Challenges for Cognitive Neuroscience. *arXiv preprint cs/0412059*.

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia*, *33*(5), 623–642.

Jeannerod, M. (2006). *Motor Cognition: What actions tell the self*. Oxford University Press.

Krüger, N., Geib, C., Piater, J., Petrick, R., Steedman, M., Wörgötter, F., ... Dillmann, R. (2011). Objectaction complexes: Grounded abstractions of sensorymotor processes. *Robotics and Autonomous Systems*, *59*(10), 740–757.

Plate, T. (2003). *Holographic reduced representations*. CSLI Publications.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Pearson Education.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *298*(1089), 199–209.

Stewart, T., Choo, F.-X., & Eliasmith, C. (2014). Sentence processing in spiking neurons: a biologically plausible left-corner parser. In *Proceedings of the 36th annual conference of the cognitive science society* (p. 1533-1538). Cognitive Science Society.

Taatgen, N., & Lee, F. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Human Factors*, *45*(1), 61–76.

Wolpert, D., & Landy, M. (2012). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*(6), 996-1003.

Zadikoff, C., & Lang, A. E. (2005). Apraxia in movement disorders. *Brain*, *128*(7), 1480–1497. doi: 10.1093/brain/awh560

Systems Factorial Analysis of Item and Associative Retrieval

Gregory E. Cox, Amy H. Criss
(gecox100, acriss)@syr.edu
Department of Psychology, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244-2340

Abstract

Using hierarchical Bayesian estimation of RT distributions, we present a novel application of Systems Factorial Technology (Townsend & Nozawa, 1995) to the retrieval of item and associative information from episodic memory. We find that item and associative information are retrieved concurrently, with positive memory evidence arising from a holistic match between the test pair and the contents of memory, in which both item and associative matches are pooled together into a single source. This retrieval architecture is inconsistent with both strictly serial processing and independence of item and associative information. Pooling of item and associative matches implies that while item and associative information may be separable, they are not qualitatively different, nor are qualitatively different processes (e.g., familiarity vs. recollection) used to retrieve these kinds of information.

Keywords: Memory models; associative recognition; systems factorial technology; Bayesian statistics.

Introduction

Associations have served as important theoretical constructs throughout the study of cognition. The present work is concerned with the distinction made in memory between memory for isolated events—which we refer to as "items"—and memory for combinations of events—which we refer to as "associations". Given that myriad phenomena in human learning depend on being able to store and retrieve both of these kinds of information and a dissociation between item and associative memory is important for a variety of neurological diagnoses, it is crucial to understand the representations and processes behind memory for items and associations. In this paper, we use qualitative properties of response dynamics to determine what kind of processing is involved in item and associative retrieval and how the two could be represented in order to support such processing.

Various theories posit that item and associative information are qualitatively different and are retrieved using independent processes. Linearities in receiver-operating characteristic (ROC) functions obtained in associative recognition have been argued to reflect the operation of an "all-or-none" recall process for associative information that is independent of the continuous-valued familiarity process used to assess item information (Yonelinas, 1997). Unfortunately, beyond the difficulty of statistically determining whether a ROC curve is linear or curvilinear, such functional forms are not, in fact, diagnostic of the type of processing involved: Curvilinear ROCs are consistent with discrete-state processing (Province & Rouder, 2012) and linear ROCs are consistent with continuous-valued processes (Wixted, 2007). Process-dissociation procedures (Jacoby, 1991) have also been used to assess whether associative information must be retrieved using a distinct recall process, but such procedures do not lead to accurate estimates of the contribution of different processes even with simulated data (Ratcliff, Van Zandt, & McKoon, 1995), nor has any strong evidence been found for the operation of multiple retrieval processes in item recognition (Dunn, 2004; Pratte & Rouder, 2012).

While there is little evidence for a qualitative distinction between item and associative information, there is good evidence for a *quantitative* difference. Item recognition accuracy decreases faster over time than associative recognition accuracy (Hockley, 1991, 1992), suggesting that associative recognition gives less weight to context (Murdock, 1997) or that it begins from a baseline that is defined by item information (Humphreys, Bain, & Pike, 1989). A focus on studying items impairs associative memory, but a focus on associative study has no negative impact on item memory (Hockley & Cristi, 1996), and associative recognition depends on the strength of *both* items and pairs (Buchler, Light, & Reder, 2008). Studies of retrieval dynamics find that the ability to discriminate between studied and unstudied items begins earlier than the ability to distinguish intact from rearranged pairs (Gronlund & Ratcliff, 1989; Rotello & Heit, 2000; Nobel & Shiffrin, 2001). While this delay has been attributed to a qualitatively different "recall-to-reject" process (Rotello & Heit, 2000; Malmberg, 2008), such a process cannot explain why stronger associations are not retrieved more quickly (Wickelgren & Corbett, 1977; Dosher, 1984; Nobel & Shiffrin, 2001) nor why participants do not use partial cues to aid associative recognition (i.e., using a singly presented word to retrieve its studied associate; Gronlund & Ratcliff, 1989). Instead, such dynamics are more consistent with the need to augment an item-only retrieval cue with compound associative information (Cox & Shiffrin, 2015).

Thus, associative information appears to be at least partially dependent on item information at both storage and retrieval, consistent with the fact that associative interference occurs only among pairs comprised of the same types of items (Criss & Shiffrin, 2004; Aue, Criss, & Fischetti, 2012). At the same time, item information is also affected by the presence of an association such that recognition of intact item groups is superior to rearranged item groups even when associative information is irrelevant (Tulving & Thompson, 1973; Clark & Shiffrin, 1987). Modeling these and other data required an assumption that presenting a pair/triplet intact at test led to an enhancement for memory for the component items, and that a holistic pair/triplet cue was used during tasks requiring retrieval of associative information (Clark & Shiffrin, 1992). Priming studies suggest that this holistic cue may be configural in nature, in that it is only effective when *all* of the component items are present and there is no effect of a partial

overlap between study and test (Dosher & Rosedale, 1989, 1997).

It is clear that, despite attempts to dissociate item and associative memory, the two appear to be closely related. Unfortunately, extant empirical and theoretical work is not sufficient to fully characterize their relationship nor whether the processes used to retrieve item and associative information are in fact separable. The present work is a novel application of Systems Factorial Technology (SFT; Townsend & Nozawa, 1995) to associative recognition in which we use qualitative properties of response dynamics to determine what kinds of processing and interactions are present during the retrieval of item and associative information from event memory.

Experiment

SFT is a set of experimental and analytical techniques designed to identify the class of model to which an information processing system belongs (for an overview, see Houpt, Blaha, McIntire, Havig, & Townsend, 2014). It does so by comparing how processing dynamics differ as the presence and strength of different information sources is varied. To use SFT to describe item and associative retrieval, we employ an associative recognition paradigm in which participants study a list of pairs. After study, participants are presented with a series of test pairs and must give a positive response only to test pairs that exactly match one of the studied pairs. We separately manipulate the strength of the item and associative information in each test pair, implementing a double factorial paradigm. This paradigm allows for the computation of several key SFT statistics needed to characterize the processes underlying item and associative retrieval.

Two broad classes of processing architecture are parallel models, in which both item and associative information are retrieved simultaneously, and serial models, in which one type of information is not retrieved until the first retrieval process is finished. For example, many dual-process theories assume that associative retrieval does not begin until item retrieval has completed, an example of serial processing (Diller, Nobel, & Shiffrin, 2001; Malmberg, 2008). If either item or associative information can be used to make a response, the system is called "self-terminating". If, however, both types of information must be retrieved to generate a response, the system is "exhaustive".

Item and associative retrieval can also interact. For example, if positive evidence from both item and associative retrieval is pooled into a single "match" accumulator, we say that positive responses are facilitated whereas negative responses are inhibited, and vice versa. In the extreme case, both match and mismatch information might be pooled from both sources, termed "coactive" processing; in this case, item and associative information are effectively inseparable, since it is impossible to tell whence a particular bit of information arose. Just as interactions blur the line between item and associative retrieval, they make it impossible to make a meaningful distinction between serial and parallel processing

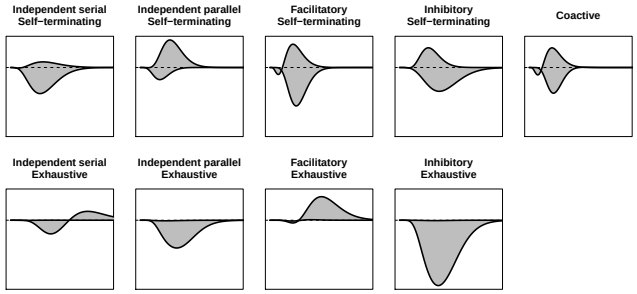


Figure 1: Allowed forms of the survivor interaction contrast (SIC) function for each processing architecture with time along the x-axis and dashed line at zero.

(Townsend, 1976); interactive serial processes are functionally equivalent to parallel cascade processing (McClelland, 1979). We refer to interactive processes as parallel, however, since the defining feature of parallel processing is that the two processes are able to operate concurrently, which is required by our notion of interaction.

Two SFT statistics can help us diagnose the architecture and interactions present in item and associative retrieval. First is the Survivor Interaction Contrast (SIC) function:

$$SIC(t) = [S_{LL}(t) - S_{LH}(t)] - [S_{HL}(t) - S_{HH}(t)]$$

where $S_{LL}(t)$ is the survivor function of the response time distribution for correct responses to test pairs with low (L) item strength and low (L) associative strength, and the other terms are defined similarly. This statistic represents how the processing dynamics of one factor (item retrieval) change as the other factor (associative retrieval) changes, taking particular forms depending on the architecture of the system. In situations where the two factors are selectively influenced by experimental manipulations and accuracy is high, the form of the SIC function can be found analytically (Townsend & Nozawa, 1995). However, when item and associative retrieval may interact—possibly leading to violations of selective influence—and/or demonstrate low to moderate accuracy, the predicted forms of the SIC function can vary. To that end, we followed Eidels, Houpt, Altieri, Pei, and Townsend (2011) and Fific, Nosofsky, and Townsend (2008) and obtained SIC predictions for various architectures by simulating accumulator models with a wide variety of parameter values. These simulations yielded an allowed set of SIC functions for each architecture, depicted in Figure 1, which allows models to be ruled out if the observed SIC is not among the allowed set.

The other statistic we use to characterize item and associative retrieval is the capacity function, specifically, the capacity function adapted to low-accuracy settings given by Equation 2 in Townsend and Altieri (2012). This function compares processing dynamics when two sources of information are present (i.e., in an intact pair, there is both an item match and an associative match) to what would be expected if a parallel unlimited capacity system processed the two sources separately. As shown in Figure 2, this function, termed the "assessment" function, also takes characteristic forms depend-

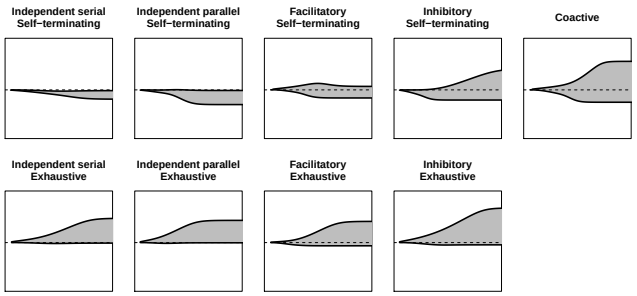


Figure 2: Allowed forms of the OR correct-and-fast assessment function for each processing architecture with time along the x-axis and dashed line at one.

Table 1: Design of studied pairs. Numerals denote either the left or right half (at random) of a particular image.

Image half A	Image half B	Item strength	Associative strength
1	2	H	H
3	4	H	L
5	6	H	H
7	8	H	H
9	10	H	L
11	12	H	L
13	14	L	H
15	16	L	L
17	18	L	H
19	20	L	H
21	22	L	L
23	24	L	L
25	26	H	H
27	28	H	L
29	30	L	H
31	32	L	L

ing on the type of processing architecture. Generally speaking, values of the assessment function greater than one reflect facilitation while values less than one indicate inhibition. Jointly, the SIC and assessment functions characterize how information is combined between item and associative retrieval and the manner in which it is processed dynamically.

Method

Participants 135 Syracuse University students took part in this experiment in exchange for course credit.

Materials Stimuli consisted of indoor and outdoor scene images, 256×256 pixels in size, chiefly derived from image sets used by Konkle, Brady, Alvarez, and Oliva (2010). The images were first screened to remove any legible writing (to preclude this as a strategy to remember particular images) as well as people (since these were particularly salient relative to other scene content). While space constraints prevent describing the procedure in detail, we used the color histograms of the images to select a pool of 512 to use in the experiment. Our criteria were designed to simultaneously maximize within-image symmetry (so the left and right halves would be similar to each other) and between-image dissimilarity (to reduce interference from other items).

Design Each study/test list was comprised of 16 unique pairs of image halves, divided evenly into four strength conditions, summarized in Table 1: High associative strength/high item strength (HH), high associative strength/low item strength (HL), low associative strength/high item strength (LH), and low associative strength/low item strength (LL).

Table 2: Design of test lists. Numerals refer to the same images in the study design in Table 1, with apostrophes denoting the unstudied half of the image labeled by the numeral.

Item strength	Associative strength: H	L	N_H	N_L
H	1, 2	3, 4	5, 8	9, 12
	Quadrant 1 (Intact pairs)		*Quadrant 2 (Rearranged pairs)*	
L	13, 14	15, 16	17, 20	21, 24
N_H	25', 26'	27', 28'	7', 6'	11', 10'
	Quadrant 3 (Novel pairs)		*Quadrant 4 (Novel rearranged pairs)*	
N_L	29', 30'	31', 32'	19', 18'	23', 22'

High associative strength pairs were presented 3 times during the study list while low associative strength pairs were presented only once during study. Item strength was manipulated by presenting the two image halves comprising the pair separately, paired with themselves. The image halves in a low item strength pair appeared *only* as part of the study pair. The image halves in a high item strength pair were shown during study paired with themselves twice. Thus, in total, the study list comprised 64 pair presentations: 8 low strength pair presentations, 24 high strength pair presentations (8 pairs repeated 3 times), and 32 self-pairings of an image half (2 presentations each of the 2 halves of 8 high item strength pairs).

Test lists were comprised of 16 pairs, summarized in Table 2. The 4 pairs in the upper left quadrant are presented intact and should be given a positive response. The remaining 12 pairs in the other 3 quadrants are foils: Rearranged pairs (quadrant 2) are created by swapping the left and right image halves of two pairs, as in normal associative recognition studies. Intact novel pairs (quadrant 3) are created by replacing each image half in an intact pair with the unstudied half of the respective images. Rearranged novel pairs (quadrant 4) apply both of these transformations: A pair is rearranged and then each of its component image halves is replaced with the unstudied image half.

Procedure Participants engaged in 10 study/test blocks. Prior to each study block, participants were told that they would be shown sets of image pairs and that they should try to remember which images appeared together. Study pairs were then presented on a white background for 2 seconds at a time, with a 0.5 second blank between each pair presentation. Presentation order was randomized under the constraint that two successive pairs did not contain any of the same image halves. The two image halves subtended approximately 3 degrees of visual angle, with approximately 0.5 degrees of blank space between them. Although the image pairs were centered horizontally on the screen, the left and right halves were independently offset from vertical center by random values sampled uniformly from [-0.25, 0.25] degrees of visual angle. This offset served two purposes: first, to emphasize that the two image halves were not meant to be treated as part of the same image; second, to avoid confusion between successive pair presentations.

After presentation of all study pairs, participants were informed that they would be shown another set of image pairs and should give a positive response only to pairs of images

Table 3: Observed mean proportion of positive responses and mean of median correct RTs (in seconds, in parentheses) for each pair type.

Item strength		Associative strength: H	L	N_H	N_L
	H	0.75 (*1.071*)	0.54 (*1.162*)	0.37 (*1.216*)	0.34 (*1.194*)
	L	0.72 (*1.095*)	0.46 (*1.197*)	0.31 (*1.209*)	0.22 (*1.105*)
	N_H	0.26 (*1.157*)	0.21 (*1.111*)	0.15 (*1.094*)	0.16 (*1.123*)
	N_L	0.27 (*1.151*)	0.22 (*1.082*)	0.14 (*1.101*)	0.12 (*1.060*)

that had appeared together at the same time on the most recent list (they were told that each test would only be for the most recent list). Positive and negative responses were randomly mapped to the "F" or "J" keys for each participant. Participants were instructed to respond as quickly and accurately as possible. Test instructions appeared on screen for a minimum of 15 seconds, after which participants could press "enter" to proceed to the test list. Each test trial began with a fixation cross centered on the screen for 0.5 seconds followed by presentation of the test pair (which followed the same sizing and random vertical offset procedure used during study). The test pair remained on screen until a response was made, after which feedback was given. Participants were told whether their response was "correct" or "incorrect", with font color green if correct and red if incorrect. Regardless of correctness, if the response was made in under 300 ms, feedback included a statement to "please take more time to respond" and if the response was made in over 4 seconds, feedback included a statement to "please try to respond more quickly". Feedback appeared for a minimum of 1 second, an additional 0.5 seconds if the participant was incorrect, and an additional 3 seconds if the response was too fast. A random time sampled at uniform between 0.25 and 0.75 seconds preceded the onset of the next test trial.

Results

Prior to analysis, we excluded trials in which the response time was shorter than 200 ms (47 out of 21,369 trials) or longer than 5 s (127 trials). Based on the remaining trials, we excluded participants who failed to give a higher rate of positive responses to intact pairs than to foils or who did not give any correct responses in at least one condition. All subsequent analyses were carried out on the remaining 18,760 trials from 118 participants.

Mean proportion of positive responses and median correct RT for each combination of item and associative strength are given in Table 3. A 4 (item strength) $\times$ 4 (associative strength) within-subjects ANOVA on the proportion of positive responses finds main effects of item strength ($F(3,351) = 298.4$, $p \sim 0$) and associative strength ($F(3,351) = 238.3$, $p \sim 0$) as well as a significant interaction ($F(9,1053) = 42.4$, $p \sim 0$). Using the same 4×4 ANOVA to analyze median correct RT, we again find main effects of item strength ($F(3,351) = 10.44$, $p \sim 0$) and associative strength ($F(3,351) = 3.01$, $p \approx 0.03$) as well as a significant interaction ($F(9,1053) = 8.99$, $p \sim 0$).

Systems Factorial Analysis

Precision terms: $\tau_{a:b} \sim \Gamma(1.01, 0.01)$

Intercept: $m_0 \sim \mathcal{N}(0, 0.001)$

For each level of item strength i, associative strength j, response r, subject s...

First-order main effects

$m_i \sim \mathcal{N}(0, \tau_{1:1})$, $m_j \sim \mathcal{N}(0, \tau_{1:2})$, $m_r \sim \mathcal{N}(0, \tau_{1:3})$, $m_s \sim \mathcal{N}(0, \tau_{1:4})$

Second-order interactions

$m_{i:j} \sim \mathcal{N}(0, \tau_{2:1})$, $m_{i:r} \sim \mathcal{N}(0, \tau_{2:2})$, $m_{i:s} \sim \mathcal{N}(0, \tau_{2:3})$,

$m_{j:r} \sim \mathcal{N}(0, \tau_{2:4})$, $m_{j:s} \sim \mathcal{N}(0, \tau_{2:5})$, $m_{r:s} \sim \mathcal{N}(0, \tau_{2:6})$

Third-order interactions

$m_{i:j:r} \sim \mathcal{N}(0, \tau_{3:1})$, $m_{i:j:s} \sim \mathcal{N}(0, \tau_{3:2})$, $m_{i:r:s} \sim \mathcal{N}(0, \tau_{3:3})$, $m_{j:r:s} \sim \mathcal{N}(0, \tau_{3:4})$

Fourth-order interaction

$m_{i:j:r:s} \sim \mathcal{N}(0, \tau_4)$

Linear combination of predictors:

$$m_{i,j,r,s} = m_0 + m_i + m_j + m_r + m_s + m_{i:j} + m_{i:r} + m_{i:s} + m_{j:r} + m_{j:s} + m_{r:s} + m_{i:j:r} + m_{i:j:s} + m_{i:r:s} + m_{j:r:s} + m_{i:j:r:s}$$

Likelihood, given item strength $I[n]$, associative strength $A[n]$, response $R[n]$, and subject $S[n]$ on trial n:

$$\text{RT}_n \sim \text{Ex-Gaussian}\left(m_{I[n],A[n],R[n],S[n]}, s_{I[n],A[n],R[n],S[n]}, t_{I[n],A[n],R[n],S[n]}\right)$$

Figure 3: Hierarchical model structure for RT distributions; although the model structure is depicted for only the m parameter of the Ex-Gaussian, the same structure was applied to $\log s$ and $\log t$, the logarithmic transformation required because the Ex-Gaussian is defined only for $s, t > 0$. Group-level RT distributions are obtained by marginalizing each parameter over subjects, which amounts to excluding terms involving the subject factor s (all of which have mean 0), e.g., $\bar{m}_{i,j,r} = m_0 + m_i + m_j + m_r + m_{i:j} + m_{i:r} + m_{j:r} + m_{i:j:r}$.

Hierarchical parametric RT distribution estimation In contrast to most applications of SFT, we have a large number of participants but few observations per participant. We therefore estimate RT distributions using a hierarchical Bayesian model (Rouder, Lu, Speckman, Sun, & Jiang, 2005), which also enables robust statistical tests. We assume that RT distributions have an Ex-Gaussian form, with likelihood

$$f(x) = \frac{1}{t\sqrt{2\pi}} \exp\left(\frac{s^2}{2t^2} - \frac{x-m}{t}\right) \int_{-\infty}^{\frac{x-m}{s} - \frac{s}{t}} \exp\left(-\frac{y^2}{2}\right) dy,$$

where x is the response time and m, s, and t are the parameters of the distribution. The Ex-Gaussian has been found to provide a good description of RT distributions in recognition memory (Ratcliff & Murdock, 1976) and can accommodate a variety of distribution shapes. For each participant, there are a total of $4 \times 4 \times 2 = 32$ RT distributions defined by the factorial combination of item strength, associative strength, and positive/negative response. We estimate these parameters for each participant according to the multilevel model described in Figure 3. We report SFT statistics at the group level, based on the Ex-Gaussian distributions obtained by marginalizing the three parameters (m, $\log s$, and $\log t$) for each of the 32 distributions over participants (see caption of Figure 3). The model was implemented in JAGS (Plummer, 2013), which was used to obtain 10,000 posterior samples split over 10 parallel chains after 2000 iterations of "burn-in" each.

Correct acceptance architecture The SIC function for correct acceptance of intact pairs (Figure 4, upper left) demonstrates a credible positive peak (95% credible interval [CI] of the maximum is [0.024, 0.142]) but no evidence for any negative deflections (95% CI of the minimum is [-0.013, 0]). We can thus rule out all forms of exhaustive process-

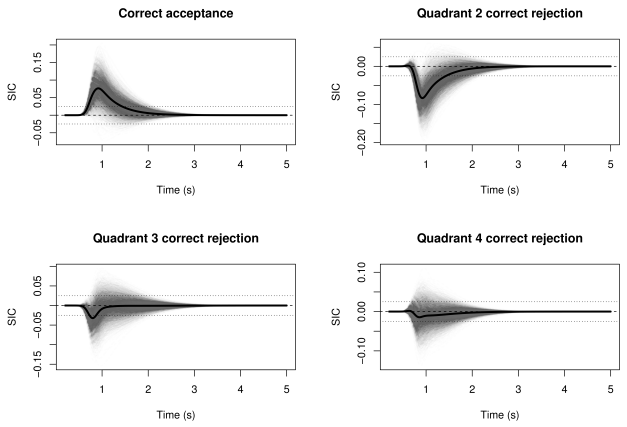


Figure 4: Posterior distributions over SIC functions for correct responses in each quadrant (see Table 2). Solid lines are posterior means while faded lines are posterior samples.

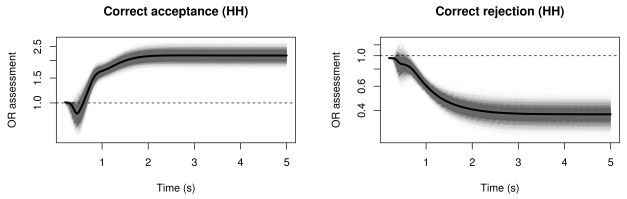


Figure 5: Posterior distributions over capacity assessment functions for correct responses. Solid lines are posterior means while faded lines are posterior samples.

ing except facilitatory, since each of these requires either a flat SIC or one with a credible negative deflection. In addition, coactive and facilitatory self-terminating processing, while they allow for SIC's that are majority positive, require at least a small early negative deflection; nonetheless, we withhold ruling these out since this negative deflection may be too small to detect. The assessment function (Figure 5, left) is everywhere greater than or equal to one, which is impossible under independent serial or parallel self-terminating processing and implies facilitation of positive responses. In sum, the SIC and capacity functions for correct acceptance are most consistent with facilitatory exhaustive and inhibitory self-terminating processing, but may allow for coactive or facilitatory self-terminating processing.

Correct rejection architecture While participants should give a positive response to pairs in the upper left quadrant of Table 2, they should give negative responses to pairs in the remaining three quadrants. Within each of these quadrants, item and associative strength is still varied but must be viewed in reverse: a high strength is actually *low* evidence in favor of the correct (negative) response. For example, the SIC for Quadrant 2 is given by $[S_{HN_H}(t) - S_{HN_L}(t)] - [S_{LN_H}(t) - S_{LN_L}(t)]$. Similarly, whenever positive responses are self-terminating, negative responses must be exhaustive; and facilitation of positive responses is inhibition for negative responses. Therefore, based on the remaining possible correct acceptance architectures, correct rejections could arise from inhibitory self-terminating (reverse of facilitatory exhaustive) or facilitatory exhaustive (reverse of inhibitory self-terminating) processing, as well as possibly inhibitory exhaustive or coactive processing.

The SIC's for the three types of correct rejection (Figure 4) differ quantitatively but are qualitatively similar. None demonstrate any credible positive deflections (95% CI of the maximum for quadrant 2 is [0, 0.017], for quadrant 3 is [0, 0.044], and for quadrant 4 is [0, 0.042]), and the SIC for quadrant 2—correct rejection of rearranged pairs—demonstrates a credible *negative* deflection (95% CI of the minimum is [-0.146, -0.030]). This pattern rules out facilitatory exhaustive processing, since this does not allow for any negative deflections. The assessment function for correct rejections (Figure 5, right) is never greater than one, consistent with the remaining possible architectures and implying inhibition of negative responses.

Complete picture We have so far examined the SIC and assessment functions without regard to the task demands, namely, that correct acceptances by definition require exhaustive processing of both item and associative information—only correct rejections can be made on the basis of a single source. After thus ruling out self-terminating processing for correct acceptances—particularly as facilitatory self-terminating processing entails a negative SIC deflection that we did not observe—we are left with two remaining architectures that can account for the complete qualitative pattern of response dynamics: Parallel facilitatory processing where positive responses are exhaustive; and fully coactive processing. In both architectures, positive match information is pooled between item and associative sources, but in the case of facilitatory processing, mismatch (negative) information arises separately from item and associative retrieval.

Discussion

Based on measures from Systems Factorial Technology (Townsend & Nozawa, 1995; Townsend & Altieri, 2012), we found that item and associative information are retrieved concurrently, with match information pooled between both sources. It is also possible that mismatch information is pooled, resulting in fully coactive processing, making item and associative information effectively indistinguishable. While our results allow for the possibility that associative information may be retrieved more slowly than item information (Gronlund & Ratcliff, 1989), our results are inconsistent with strictly serial processing, such as the proposal that item familiarity is assessed prior to the retrieval of associative information (e.g., Rotello & Heit, 2000; Malmberg, 2008). We can also rule out independence of item and associative retrieval. The assumption of independence lies at the core of both the process dissociation procedure (Jacoby, 1991) and mixture analysis of ROC curves (Yonelinas, 1997). Because this assumption is violated, any conclusions drawn within those frameworks are invalid with respect to item and associative retrieval (for additional arguments along these lines,

see Curran & Hintzman, 1995; Hillstrom & Logan, 1997).

Match information—and possibly mismatch information in the case of coactive processing—was found to be pooled between item and associative sources, suggesting that positive memory evidence derives from a holistic match between the pair and the contents of memory. This is consistent with priming studies that find that multiple memory cues are combined in an interactive fashion (Ratcliff & McKoon, 1988; Dosher & Rosedale, 1989, 1997). The ability for mismatch information to arise separately from multiple sources has previously been implicated in short-term memory search (Mewhort & Johns, 2000), and may well apply in the context of long-term associative recognition, particularly as some of the critical mismatch information in Mewhort and Johns (2000) was relational in nature. Taken together, separate processing of mismatch information and pooling of match information suggests that while item and associative information may be separable, they cannot be qualitatively different, otherwise it would not be possible to combine them into a single holistic match. While this representational schema is not entailed by any extant memory theory, TODAM's allowance for separate item and associative traces (Murdock, 1982) allows that theory the flexibility to construct such a representation.

Our conclusions are based on model-independent qualitative properties of response dynamics, demonstrating how SFT can be applied in situations in which the number of observations per individual is limited and responses are more error-prone. By using Bayesian hierarchical estimation of RT distributions, we can accurately characterize group-level response dynamics. We are limited, however, in what we can say about the degree to which individuals may vary in their retrieval architectures; we are currently employing quantitative individual RT modeling (e.g., Ratcliff, 1978; Brown & Heathcote, 2008) to address this issue. Nonetheless, we have shown how tools from SFT can be used to answer critical questions in novel domains.

References

Aue, W. R., Criss, A. H., & Fischetti, N. W. (2012). Associative information in memory: Evidence from cued recall. *Journal of Memory and Language*, *66*, 109–122.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Buchler, N. G., Light, L. L., & Reder, L. M. (2008). Memory for items and associations: Distinct representations and processes in associative recognition. *Journal of Memory and Language*, *59*, 183–199.

Clark, S. E., & Shiffrin, R. M. (1987). Recognition of multiple-item probes. *Memory & Cognition*, *15*(5), 367–378.

Clark, S. E., & Shiffrin, R. M. (1992). Cuing effects and associative information in recognition memory. *Memory & Cognition*, *20*(5), 580–598.

Cox, G. E., & Shiffrin, R. M. (2015). A dynamic approach to secondary processes in associative recognition. In D. Noelle et al. (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Criss, A. H., & Shiffrin, R. M. (2004). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, *32*(8), 1284–1297.

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning. Memory, and Cognition*, *21*(3), 531–547.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 414–435.

Dosher, B. A. (1984). Degree of learning and retrieval speed: Study time and multiple exposures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 541–574.

Dosher, B. A., & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General*, *118*(2), 191–211.

Dosher, B. A., & Rosedale, G. (1997). Configural processing in memory retrieval: Multiple cues and ensemble representations. *Cognitive Psychology*, *33*, 209–265.

Dunn, J. C. (2004). Rememberknow: A matter of confidence. *Psychological Review*, *111*(2), 524–542.

Eidels, A., Houpt, J. W., Altieri, N., Pei, L., & Townsend, J. T. (2011). Nice guys finish fast and bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology*, *55*, 176–190.

Fific, M., Nosofsky, R. M., & Townsend, J. T. (2008). Information-processing architectures in multidimensional classification: A validation test of the systems factorial technology. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 356–375.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 846–858.

Hillstrom, A. P., & Logan, G. D. (1997). Process dissociation, cognitive architecture, and response time: Comments on Lindsay and Jacoby (1994). *Journal of Experimental Psychology: Human Perception and Performance*, *23*(5), 1561–1578.

Hockley, W. E. (1991). Recognition memory for item and associative information: A comparison of forgetting rates. In *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 227–248).

Hockley, W. E. (1992). Item versus associative information: Further comparisons of forgetting rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(6), 1321–1330.

Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*(2), 202–216.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with R. *Behavior Research Methods*, *46*, 307–330.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208–233.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287–330.

Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental Psychology: General*, *129*(2), 262–284.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(3), 609–626.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*(4), 839–862.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 384–413.

Plummer, M. (2013). *JAGS: Just another Gibbs sampler.* Retrieved from `http://mcmc-jags.sourceforge.net/`

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1591–1607.

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences*, *109*(36), 14357–14362.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385–408.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General*, *124*(4), 352–374.

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, *28*(6), 907–922.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.

Townsend, J. T. (1976). Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology*, *14*, 219–238.

Townsend, J. T., & Altieri, N. (2012). An accuracy-response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review*, *119*(3), 500-516.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–359.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Wickelgren, W. A., & Corbett, A. T. (1977). Associative interference and retrieval dynamics in yes-no recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(2), 189–202.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*(6), 747–763.

Children learn non-exact number word meanings first

Junyi Chu
junyichu@ucsd.edu
Department of Psychology
UC, San Diego

Katie Wagner
kgwagner@ucsd.edu
Department of Psychology
UC, San Diego

David Barner
barner@ucsd.edu
Department of Psychology
UC, San Diego

Abstract

Children acquire exact meanings for number words in distinct stages. First, they learn *one,* then *two,* and then *three* and sometimes *four.* Finally, children learn to apply the counting procedure to their entire count list. Although these stages are ubiquitous and well documented, the foundation of these meanings remains highly contested. Here we ask whether children assign preliminary meanings to number words before learning their exact meanings by examining their responses on the Give-a-Number task to numbers for which they do not yet have exact meanings. While several research groups have approached this question before, we argue that because these data do not usually conform to a normal distribution, typical methods of analysis likely underestimate their knowledge. Using non-parametric analyses, we show that children acquire non-exact meanings for small number words like *one, two, three, four* and possibly for higher numbers well before they acquire the exact meanings.

Keywords: Number word learning; Counting; Cognitive Development; Language acquisition

Introduction

How do children acquire the meanings of large exact numbers like "47"? Although humans are able to represent quantity in absence of language, attested systems of number representation are either limited to representing small arrays of objects precisely, or to representing larger quantities approximately (Feigenson, Dehaene, & Spelke, 2004). No attested non-verbal system is able to represent large numbers precisely (Laurence & Margolis, 2005). Here, we investigated how children acquire labels for exact numbers by exploring the very first meanings they exhibit. To do this, we analyzed previously collected data in a simple but novel fashion and found that, before children learn exact meanings for words like *one, two, three, four,* and perhaps larger numbers, they first learn non-exact meanings.

In the US, children learning English generally begin number word learning around age 2, by learning to blindly recite a small subset of number words in sequence (*one, two, three,* etc.), much like they learn the alphabet, though in tandem with procedures that involve pointing at objects in sequence (Gelman & Gallistel, 1986; Fuson, 1988; Wynn, 1990, 1992). Not long after, they learn an exact meaning for the word *one*, and at this time can give one object when asked for *one*, but cannot reliably give larger numbers when asked. These children are often referred to as "1-knowers". After a long delay – around 6 to 9 months later – children learn an exact meaning for *two,* and are called "2-knowers". Next, they learn *three* ("3-knowers"), and sometimes *four* ("4-knowers"), in sequence. After this, they appear to discover something more general about numbers: That the counting procedure which they began with can be used to label and generate larger sets. Having noticed this, children become able to give any set within their count list and map number words to large precise cardinalities (Wynn, 1990, 1992; see also Schaeffer, Eggleston, & Scott, 1974). These children are often referred to as Cardinal Principle knowers or "CP-knowers" (for discussion see Davidson, Eng, & Barner, 2012; Lipton & Spelke, 2005; Wagner, Kimura, Cheung, & Barner, 2015).

Several recent studies have explored the role of different perceptual systems in this learning trajectory. According to one class of accounts (e.g. Carey, 2009; Klahr & Wallace, 1976; Le Corre & Carey, 2007), the fact that children are limited to learning labels up to *three* or *four* in absence of counting suggests that these meanings are learned from a capacity limited system of object representations. Consistent with this, studies using several distinct experimental paradigms have found that preverbal infants can keep track of a maximum of 3 objects in parallel, a capacity sometimes dubbed "parallel individuation" (Feigenson & Carey, 2003, 2005). Other researchers, however, have argued that number words might be constructed instead from representations in the approximate number system, or ANS (e.g. Cantlon, 2012; Spelke & Tsivkin, 2001; Wagner & Johnson, 2011). The ANS allows humans to represent sets approximately and compare them according to their ratio, such that 20 vs. 40 and 200 vs. 400 are equally discriminable, whereas 200 vs. 220 is not. These representations appear to be available from birth (e.g. Brannon, 2002; Xu & Spelke, 2000), and become sensitive to increasingly finer numerical differences well into elementary school (Halberda & Feigenson, 2008).

Currently, the relative roles of these two candidate systems in number word learning is uncertain. First, it is unclear in principle how either system *could* supply the meanings of number words. As argued by Laurence and Margolis (2005), parallel individuation is a system for representing the properties and locations of individual objects, and not the properties of sets – like cardinality. Further, parallel individuation, though at best useful for learning small numbers, could not define meanings for numbers beyond three or four. Meanwhile, the ANS, though not limited to representing small numbers, is non-exact, and cannot be used to reliably distinguish large quantities like 817 and 820. Perhaps more important, neither system supplies the type of logical representations that characterize counting and the foundations of arithmetic, such as exact equality and the successor function (i.e., that the successor of any natural number n is equal to $n+1$).

In recent years, many researchers have examined children's number word knowledge to determine if children

have preliminary meanings prior to exact ones, and if so, whether these meanings show signatures of parallel individuation and/or the ANS. If preliminary meanings are rooted in parallel individuation, such meanings would be limited to *four* since higher numbers are beyond the capacity limit of this system. In contrast, if children have preliminary meanings rooted in the ANS, these meanings could extend beyond *four* and should show a distinct characteristic of the ANS. Specifically, the standard deviation of children's errors should increase as the target numerosities increase (e.g., Le Corre & Carey, 2007; Wagner & Johnson, 2011).

Although several research groups have pursued these questions, differences in procedures and analysis have led to oftentimes conflicting conclusions. Early studies examined whether children who can count have mappings between the ANS and large number words (up to 100) using dot-array estimation tasks. These studies found that many CP-knowers (Le Corre & Carey, 2007) and less-skilled counters (Lipton & Spelke, 2005) lack such mappings.

More recent studies have probed the role of the ANS by asking whether, before learning to count, children already show evidence of mappings between number words and approximate magnitudes for at least a limited set of numbers beyond their knower level. For instance, Sarnecka and Lee (2009) argued that if children map number words to the ANS, then errors on the Give-a-Number (Give-N) should be normally distributed around the requested numerosities, with higher means for larger requests. Contrary to this prediction, they found that the means of incorrect responses beyond knower level were not higher for larger requests. However, as we show in the present study, excluding correct responses and analyzing only errors introduced systematic bias that underestimates knowledge. Assuming that children's responses are random, we expect the mean of children's responses (both correct and incorrect) to be the same across all requested numbers, as assumed by Sarnecka and Lee in their analysis. Critically, however, removing correct responses alters this expectation, such that a negative slope is expected. This is because the size of responses removed from analyses necessarily increases with the size of the requested number (e.g., correct responses to requests for *ten* are larger than for *one*). This makes it challenging to draw conclusions from incorrect responses only.

A more common approach is to examine the means of all responses beyond children's knower levels, both correct and incorrect (e.g., Barner & Bachrach, 2010; Gunderson, Spaepen, & Levine, 2015; Wagner & Johnson, 2011). Using data from Give-N and a set labeling task called "What's on this Card?", studies by Barner and Bachrach (2010) and Gunderson et al. (2015) found that children have non-exact meanings for the number just beyond their knower level (KL+1). However, responses for larger numbers (KL+2 and above) produced a flat slope, suggesting undifferentiated meanings (see Gunderson et al., 2015, for evidence that Wagner & Johnson's data support the same conclusion).

Whereas some studies (Barner & Bachrach, 2010; Le Corre & Carey, 2007) find that different tasks generate identical patterns of data, others find differences across tasks. For example, Gunderson et al. only find evidence of non-exact meanings using the What's on this Card task, which requires labeling a presented set. Shaeffer, Eggleston, and Scott (1974) find the opposite, with better performance generating sets than labeling them, a result replicated by Odic, Le Corre, and Halberda (2015). However, in the cases of Shaeffer et al., Odic et al., and Wagner and Johnson, comparisons across studies are nearly impossible, since no common measure was used across studies (e.g., Wynn's Give-N criteria for knower level assignments). In the one case where knower levels were assessed (Odic et al.), this was done with a counting What's on this Card task, which might have underestimated or overestimated children's knower levels relative to the standard Give-N task.

A further barrier to understanding these past studies is that, while most analyze all data, including correct responses, they focus on children's mean response as the dependent variable. This is problematic because although means provide an excellent characterization of central tendency for normally distributed data, children's responses on the Give-N task violate normality assumptions in profound ways, as shown in Figure 1, which depicts the responses of 2-knowers to requests for *four* and *five.*

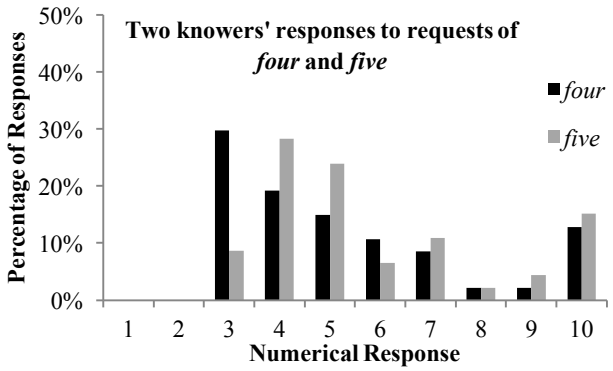


Figure 1: 2-knowers' responses for requests for *four* and *five.*

Give-N data predictably deviate from the normal distribution for two reasons. First, most obviously, normality is violated by the fact that children frequently give all available items (e.g., 10 out of 10) when they are uncertain of how to respond (see Figure 1). Second, children's responses are both upper- and lower-bounded. They are upper-bounded by the number of objects the experimenter chooses to present to the child: If the child is offered 10 objects, then only 10 can be given, even if the child might normally wish to give 12 or 15. And they are lower bounded by the child's knower level: If a child is a 2-knower then responses to *three* necessarily have a lower bound of three items. This is because the criteria for 2-knower classification require that, 2/3 of the time, the child gives two objects when asked for *two,* and further, provides two objects only when asked for *two* but not when asked for other numbers. Thus, the lower bound is given by how we choose to classify children's knowledge of numbers. Though some studies did not attempt a knower level

assignment (e.g., Schaeffer et al., 1974; Wagner & Johnson, 2011) children's responses are still expected to be lower bounded because there is ample evidence that children do acquire exact meanings for number words in stages (e.g., Wynn, 1990; 1992), thereby constraining responses to requests for higher numbers.

These considerations lead us to conclude that all past studies may underestimate children's non-exact knowledge of numbers beyond their knower level, and that differences in data analysis may explain conflicting results and asymmetries between different tasks (which vary according to whether they are subject to problems of upper and lower bounds). To explore this conjecture, we analyzed data from a non-titrated variation of the Give-N task and replicate the analyses of both Sarnecka and Lee (2009) and Gunderson et al. (2015). In this variation, each child received three trials for each number tested in a pseudo-randomized order, regardless of their performance. This is important because it allowed us to examine children's responses for inquiries well beyond their knower level. We compared these analyses to non-parametric analyses, and show that children acquire non-exact meanings for small number words like *one, two, three, four* and possibly higher numbers well before they acquire their exact meanings.

Methods

Participants

Our data set combined data from a study of English monolinguals (Almoammer et al., 2013) and a study of bilinguals (Wagner et al, 2015). Only subset knowers were included from each study. Monolinguals were 44 children aged 2;0 to 4;3 years (M = 3;1). There were 29 bilinguals including 14 who spoke both English and French (M=3;4, range = 2;4 – 4;0) and 15 who spoke both English and Spanish (M=3;8, range = 2;10 – 4;11)[1]. Only English dominant bilinguals (i.e. those with higher knower-level when tested in English) were included, to ensure that stronger number knowledge in a second knowledge did not affect performance in English. Analyses below confirm no difference between monolingual and bilingual children.

Procedures

Give-N Task. This task was adapted from Wynn (1990) and was used to assess children's knowledge of number word meanings in each language the child spoke. The experimenter began by presenting the child with a plate and ten similar objects. For each trial, the experimenter asked the child to place a quantity on the plate, avoiding singular and plural marking by asking, "Can you put *n* on the plate? Put *n* on the plate and tell me when you're all done." Once the child responded, the experimenter asked, "Is that *n*? Can you count and make sure?" and encouraged the child to count. If the child recognized an error, the experimenter allowed the child to change his/her response.

Children completed up to twenty-one quasi-randomized trials, consisting of three trials for each of the seven numbers tested (i.e., 1, 2, 3, 4, 5, 8, and 10). Children were defined as an *n*-knower (e.g., 3-knower) if they correctly provided *n* (e.g., 3 fish) on at least two out of the three trials that *n* was requested and, of those times that the child provided *n*, two-thirds of the times the child did so it was in response to a request for *n*. If *n* was five or higher, the child was classified as a CP-knower.

Results

Knower Level Classification

In our dataset, there were 28 non-knowers, 15 1-knowers, 16 2-knowers, 11 3-knowers, and 3 4-knowers. Because the number of 4-knowers was small (n = 3) and because prior evidence suggests CP-knowers are likely to be misclassified as 4-knowers (Wagner et. al, 2015), we took a conservative approach and excluded 4-knowers.

Further division of non-knowers. While it is likely that a large subset of non-knowers know nothing about the meanings of number words, it is also possible that some have begun developing meanings for small numbers like *one* and *two.* To assess this, we divided non-knowers into two groups. The first group (N=16) was comprised of non-knowers who gave the same number of objects in response to every request, while children in the second group (N=12) provided at least two unique responses. We excluded the first group from our analyses; for the remainder of the paper, the term *non-knower* refers only to this second group.

Replication of Sarnecka & Lee (2009)

Replicating Sarnecka and Lee's (2009) results, we found that the means of children's incorrect responses bore no positive relationship to the target value (see Table 1 and Figure 2a). The mean incorrect response was significantly different from the correct response for all but one target number, *five,* as the means for all requested numbers were close to 5. This was true for all knower levels.

Replication of Gunderson et al. (2015)

Rather than rely on incorrect responses only, Gunderson et al. (2015) analyzed the means of both correct and incorrect responses, and asked whether the slope of the means for larger numbers was positive (see Table 2). We applied this method to our data set (see Figure 2b). By comparing Figure 2a and Figure 2b, we see that a positive slope emerges once correct responses are included. In Gunderson et al. (2015)'s first experiment, they found no statistically significant positive slope from 5 to 9 amongst subset knowers but they did find a statistically significant positive slope from KL+1 to 9, though this effect was driven primarily by 3-knowers. Like Gunderson et al. we found no positive slope for subset knowers from 5 to 10 (0.11, p = 0.12, d = 0.25). However,

[1] Preliminary analyses comparing bilingual and monolingual data revealed no visual differences in the distribution of responses. For brevity, we combined both samples in the current report.

the slope from KL+1 to 10 reached statistical significance (0.13, $p = 0.0031$, $d = 0.48$). We also included an additional analysis of slope, looking at KL+2 to 10 and found that this slope also reached statistical significance for subset knowers (0.11, $p = 0.037$, $d = 0.33$). Non-knowers did not show a positive slope for KL+1 to 10, KL+2 to 10, or 5 to 10.

Table 1: Means of Incorrect Responses

N	# Trials	# Incorrect responses	Mean Incorrect	Comparison to Target		
				df	*t*	*p*
1	166	17	6.44	13	6.3	<0.01
2	167	34	6.73	18	6.5	<0.01
3	164	86	5.10	39	4.3	<0.01
4	167	129	5.20	49	3.5	0.01
5	165	139	5.50	52	0.49	0.63
8	164	153	4.93	53	8.7	<0.01
10	165	126	4.34	46	17	<0.01

Analyses of Median Responses

Having shown that we replicate previous findings in the literature using previously deployed analyses, we next sought to ask whether these studies underestimated knowledge. Already we have shown that removing correct responses from an analysis of mean responses systematically underestimates knowledge of numbers from *one* to *three.* Here we asked whether a focus on means also results in underestimation of knowledge.

To assess this, our first step was to summarize data by median response (including both incorrect and correct responses) as in Figure 2c. In place of an analysis of slope, we conducted sign tests comparing children's median responses on KL+1 to 10, KL+2 to 10 and 5 to 10 (see Table 3). None of these comparisons were statistically significant for non-knowers. However, for subset knowers, sign tests comparing both KL+1 to 10 and the KL+2 to 10 reached statistical significance, both ps < 0.01. Although subset knowers also showed a positive slope from 5 to 10, these differences were not significant, $p = 0.11$.

A second approach to addressing this question compared children's responses for requests beyond their knower-level to chance. To do this, we created a null distribution of responses which accounted for the fact that children give certain quantities more frequently than others. For example, 2-knowers gave ten objects on 0.17 of trials for requests above two. Therefore, if 2-knowers responses are independent of the requested number, we should expect that on 17% of requests for *ten,* 2-knowers will give 10 objects. We used binomial tests to compare children's actual responses (e.g., 2-knowers provided ten objects on 0.31 of *ten* trials) to chance (e.g., 0.17) to ask whether they were more likely to provide correct responses to requests for numbers from KL+1 to 4, and separately from 5 to 10. We did not include analyses for KL+2 to 4 since such an analysis would be limited to responses of *three* and *four* by 1- knowers and responses to *four* by 2-knowers.

The responses of non-knowers did not differ from chance for requests between KL+1 and 4 or between 5 to 10. In contrast, subset knowers were more likely than chance to respond with a correct amount for both requests from KL+1 to 4 from 5 to 10 (see Table 4). To eliminate the possibility that correct responses for 5 to 10 were driven by children giving all items (which happened to be 10), we also separately analyzed requests for 5 and 8 only and found that subset knowers were still correct more frequently than chance (0.12; chance = 0.085, $p = 0.04$).

Discussion

Well before children acquire the exact semantics of number words, they first acquire non-exact meanings both for numbers below five, and larger numbers too. To show this, we analyzed data from the Give-N task, and conducted analyses that included both correct and incorrect responses, that focused on median responses rather than means, and that defined chance responding in a way that respected the constraints of the Give-N tasks (adjusting for children's base rates of responding). Specifically, we replicated the findings of past studies using their analyses, and then showed that these studies systematically underestimate knowledge. Consequently, we provide the first evidence that standard tasks like Give-N, which have been used for over 40 years, can detect children's non-exact meanings for number words well beyond their knower level.

Our findings are compatible with the idea that children's preliminary number meanings are rooted in the approximate number system. However, evidence from other studies suggests that even much older children – e.g., 7 year olds – lack strong associative mappings between the ANS and number words beyond 5 (Sullivan & Barner, 2014). To the extent that these older children are able to make accurate estimates, these are driven by a global structure mapping wherein the relation between number words (e.g., *ten*) and particular magnitudes (e.g., 10) are highly malleable across contexts, unlike the meanings of the positive integers, which remain fixed independent of context.

An alternative explanation consistent with both the current data and the notion that estimates above 5 are driven more by structure mappings than mappings between number words and the ANS is that children's responses are guided by principled knowledge that numbers later in the count list denote greater quantities (the "later-greater" principle). Contrary to this hypothesis, however, several previous studies have found that no knowledge of this principle in subset knowers and even in many CP-knowers (Le Corre, 2014; Condry & Spelke, 2008; Davidson et al., 2012).

These studies in older children point to the possibility that the knowledge driving children's inexact meanings (whether ANS mappings or the later-greater principle) may not extend to their entire count list. Consistent with this idea is data showing that other types of numerical knowledge, specifically the knowledge of numerical successors, is first acquired incrementally long before children abstract the

Table 2: Slopes of the means of all responses from KL+1 to 10 and 5 to 10

		KL+1 to 10				KL+2 to 10				5 to 10			
	N	mean	SD	*p*	*d*	mean	SD	*p*	*d*	mean	SD	*p*	*d*
Non-knowers	9	-0.006	0.39	0.84	-0.016	-0.067	0.35	0.58	-0.19	-0.120	0.66	0.60	-0.18
All Subset Knowers	**42**	**0.128**	**0.26**	**0.0031*** **	**0.48**	**0.107**	**0.32**	**0.037***	**0.33**	**0.108**	**0.44**	**0.12**	**0.25**
1-knowers	15	0.030	0.21	0.59	0.14	0.030	0.26	0.66	0.11	0.070	0.46	0.57	0.15
2-knowers	16	0.192	0.21	0.0022***	0.92	0.172	0.25	0.015**	0.67	0.138	0.42	0.21	0.33
3-knowers	11	0.167	0.37	0.161	0.46	0.117	0.47	0.43	0.25	0.117	0.47	0.43	0.25

Table 3: Slope and Sign Tests of Median Responses

		KL+1 to 10		**KL+2 to 10**		**5 to 10**	
	N	Slope	Sign test *p*	Slope	Sign test *p*	Slope	Sign test *p*
Non-knowers	9	0.056	0.34	-0.125	0.96	0	0.89
All Subset Knowers	**42**		**<0.01**		**<0.01**		**0.11**
1-knowers	15	0.25	0.36	0.14	0.05	0.2	0.36
2-knowers	16	0.43	<0.01	0.33	<0.01	0.4	0.27
3-knowers	11	0.17	0.34	0.2	0.34	0.2	0.34

Table 4: Correct Responses for KL+1 to 10 compared to chance

	KL+1 to 4				**5, 8, 10**			
	# Trials	Proportion Correct	Chance	*p*	# Trials	Proportion Correct	Chance	*p*
Non-knowers	88	0.25	0.19	0.12	65	0.035	0.069	0.91
All Subset Knowers	**281**	**0.31****	**0.24**	**<0.01**	**402**	**0.16****	**0.11**	**<0.01**
1-knowers	130	0.31**	0.22	<0.01	129	0.10	0.084	0.28
2-knowers	118	0.26	0.22	0.23	175	0.19**	0.12	<0.01
3-knowers	33	0.45	0.36	0.16	98	0.21*	0.14	0.04

successor principle and can apply it to their entire count list (Wagner et al, 2015; Davidson et al, 2012).

Determining the source of children's inexact meanings for number words will inform what information children have at their disposal when learning the counting principles. This knowledge could be a useful stepping stone to acquiring the counting principles, but on it's own, it cannot be sufficient, given that children are forming these inexact meanings months before learning the counting principles.

Acknowledgments

Teachers, children and families at VIP Village Preschool in Imperial Beach, CA. NSF (#1420249) awarded to DB.

References

Almoammer, A., Sullivan, J., Donlan, C., Marusic, F., Zaucer, R., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *PNAS, 110*(46), 18448-18453.

Barner, D. & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology, 60,* 40-62.

Brannon, EM.(2002). The development of ordinal numerical knowledge in infancy. *Cognition, 83*(3), 223-240.

Cantlon, J. F. (2012). Math, monkeys, and the developing brain. *PNAS, 109,* 10725-10732.

Carey, S. (2009). Where our number concepts come from. *Journal of Philosophy, 106* (4), 220-254.

Condry, K.F. & Spelke, E.S. (2008). The development of language and abstract concepts: The case of abstract number. *J Exp Psychol Gen, 137, 22-38*

Davidson, K., Eng, K. & Barner, D. (2012). Does learning to count involve a semantic induction? *Cognition, 123,* 162-173.

Feigenson, L. & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants' manual search. *Developmental Science, 6,* 568-584.

Feigenson, L. & Carey, S. (2005). On the limits of infants' quantification of small object arrays. *Cognition, 97*, 295-313.

Feigenson, L., Dehaene, S. & Spelke, E. (2004) Core systems of number. *Trends Cogn Sci, 8*(7), 307-314.

Fuson, K. C. (1988). *Children's counting and concepts of number*. New York: Springer-Verlag.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Gunderson, E., Spaepen, E. & Levine, S. (2015). Approximate number word knowledge before the cardinal principle. *J Exp Child Psychol, 130,* 35-55.

Halberda, J. & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, 6-year-olds and adults. *Developmental Psychology, 44(5)*, 1457-1465.

Klahr, D. & Wallace, J. G. (1976). *Cognitive development: An information processing view.* Hillsdale, NJ: Erlbaum.

Laurence, S. & Margolis, E. (2005). Number and natural language. In P. Carruthers, S. Laurence, & S. Stich (eds), *The Innate Mind: Structure and Contents* (pp. 216-235). Oxford University Press.

Le Corre, M. & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*, 395-438.

Le Corre, M. (2014). Children acquire the later-greater principle after the cardinal principle. *Brit J Dev Psychol, 32(2), 163-177.*

Lipton, J. S. & Spelke, E. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child Development, 76*(5), 978-988.

Odic, D., Le Corre, M. & Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition, 138,* 102-121.

Sarnecka, B. & Lee, M. (2009). Levels of number knowledge during early childhood. *J Exp Child Psychol, 103,* 325-337.

Schaeffer, B., Eggleston, V. H., & Scott, J. L. (1974). Number development in young children. *Cognitive Psychology, 6*, 357-379.

Spelke, E. S., & Tsivkin, S. (2001). Initial knowledge and conceptual change: Space and number. *Language Acquisition and Conceptual Development*, 70-97.

Sullivan, J. & Barner, D. (2014). The development of structural analogy in number-line estimation. *Journal of Experimental Child Psychology,* 171-189.

Wagner, J. B., & Johnson, S. C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition*, *119*(1), 10-22.

Wagner, K., Kimura, K., Cheung, P. & Barner, D (2015). Why is number word learning hard? Evidence from bilingual learners. *Cognitive Psychology*, *83,* 1-21.

Wynn, K. (1990). Children's understanding of counting. *Cognition, 36,* 155-193.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, 24, 220-251.

Xu, F. & Spelke, E.S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*(1). B1-B11.

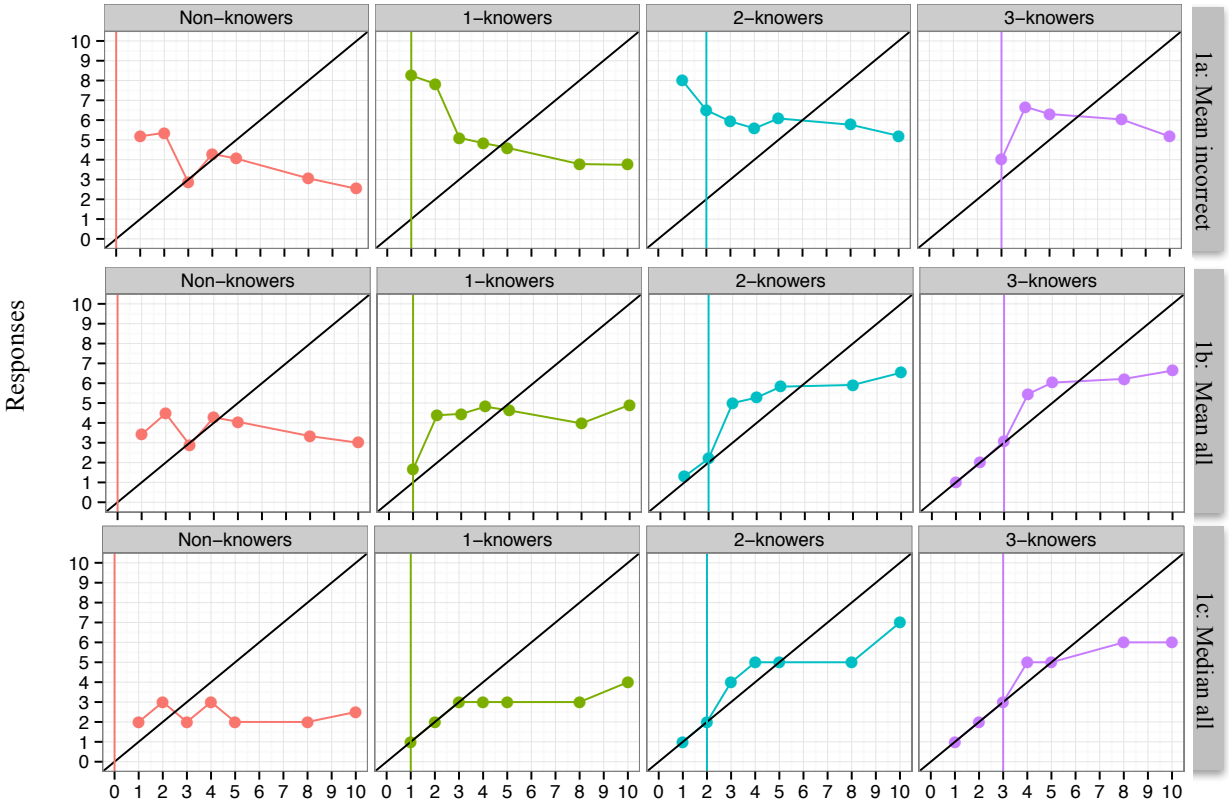


Figure 2: Responses for each knower level, using different summary statistics. (a) Mean incorrect responses, (b) Mean responses, and (c) Median responses.

A Dream Model: Reactivation and Re-encoding Mechanisms for Sleep-dependent Memory Consolidation

George Kachergis[1] **(george.kachergis@nyu.edu)**
Roy de Kleijn[2] **(kleijnrde@fsw.leidenuniv.nl)**
Bernhard Hommel[2] **(hommel@fsw.leidenuniv.nl)**
[1]Psychology Department, New York University, USA
[2]Institute of Psychology / Leiden Institute for Brain & Cognition, Leiden University, the Netherlands

Abstract

We humans spend almost a third of our lives asleep, and there is mounting evidence that sleep not only maintains, but actually improves many of our cognitive functions. Memory consolidation–the process of crystallizing and integrating memories into knowledge and skills–is particularly benefitted by sleep. We survey the evidence that sleep aids memory consolidation in various declarative and implicit tasks and review the basic neurophysiological structure of sleep with a focus on understanding what neural systems are involved. Drawing on machine learning research, we discuss why it might be useful for humans–and robots, perhaps–to have such an offline period for processing, even though humans are clearly capable of learning incrementally, online. Finally, we propose and simulate two mechanisms for use in computational memory models to accomplish sleep-based consolidation via either or both 1) re-encoding knowledge representations and 2) reactivating and strengthening recent memories.

Keywords: memory consolidation; sleep; dreaming; hippocampal replay; memory model

Introduction

Researchers have long been confused why people–and many other animals–spend around a third of their lives asleep. Why has nature burdened us with needing to be in such a vulnerable, unproductive state for so long? Hypotheses abound about why we have evolved to sleep, ranging from it allows us to conserve energy (during a night full of terrors, even), to it simply being a necessary restorative for tired muscles and minds (Siegel, 2013). Both of these example hypotheses are both reasonable and hard to refute (especially the former). Evidence in favor of the latter has been mounting: in a wide variety of cognitive and motoric tasks, performance drops when sleep patterns are interrupted.

Not only does sleep deprivation cause performance deficits, but there is now substantial empirical evidence that both declarative (i.e., facts and events–'what', 'where', 'when') and procedural (i.e., skills–'how') memory benefit from even short periods of sleep. Memory is typically described as three processes: 1) encoding: forming new traces from experience, 2) consolidation: integrating memories with prior knowledge and strengthening/crystallizing the trace, and 3) retrieval: task-dependent extraction of overall familiarity or recall of particular traces. Sleep is generally accepted to aid in consolidation, but under what particular circumstances it helps is not fully understood (Diekelmann, Wilhelm, & Born, 2009, for a review), nor by which mechanisms it works.

It is our goal to hypothesize several mechanisms by which sleep could serve to improve consolidation, and to evaluate which of these are suggested by existing empirical evidence. We will begin by considering the computational advantages and disadvantages of online (i.e., incremental; awake) learning versus offline (i.e., batch; asleep) learning, with intuitions imported from the machine learning literature. By considering the often-competing computational needs of the mind (e.g., to quickly and accurately store new information, but also to integrate such episodes with existing knowledge), we hope to better organize existing findings for rational analysis. Thus, we then proceed to review the physiological characteristics of sleep, along with some of the diverse memory effects that have been found. Typical studies often have a similar design (but different task) to the first one that found a benefit for sleep: Jenkins and Dallenbach (1924) found improved retention of nonsense syllables after a night of sleep as compared to the same amount of time spent awake. However, more recent studies have looked at shorter periods of sleep and also measured the time spent in different sleep stages, which exhibit different types of neural activity. After synthesizing the empirical results, we offer the beginnings of a computational approach to modeling memory consolidation. Despite the large amount of interest in developing computational models of episodic memory (Hintzman, 1984; Shiffrin & Steyvers, 1997, e.g.) and semantic memory (Jones & Mewhort, 2007, e.g.), these models do not yet incorporate sleep-based consolidation mechanisms. We offer several suggestions for concrete changes to an existing model of episodic memory.

Computational Issues

Machine learning algorithms can be classified as either incremental–allowing data to be added to the model instance by instance–or batch, requiring a (sometimes large) set of training instances before before the model produces useful predictions. Before we dive into a high-level discussion of costs and benefits to these two approaches, let us consider a motivating example.

A Tale of Two Robots

Imagine two cooking robots running two different versions of the same underlying learning algorithm during a week of daily training to become pizza chefs: Bob runs a batch version, and Ingrid runs an incremental version. Each day, Bob and Ingrid each watch 10 pizzas being made. Bob watches closely, recording every move for processing later that night, whereas Ingrid learns while watching–but is sometimes over-

whelmed by too many simultaneous things to pay attention to. On the other hand, even after seeing the first pizza made, Ingrid is able to make predictions about the next, and can evaluate and update her knowledge (and attention to observing particular steps) based on prediction failures. However, by storing all of the input over the day (and perhaps forever), Bob can often make a model that better approximates the real experience than what Ingrid's incremental adjustments add up to. In fact, there may even be particular sequences of experience that can lay a false foundation: for example, consider if Ingrid's first 'pizza' experience is a dessert pizza, or 'pizza chicken'–not very stereotypical pizzas. For an incremental learner who may hastily build a foundation from the first experience, such a starting point may cause difficulties for focusing on and updating the correct features over the next few recipes experienced. In contrast, Bob's batch algorithm has the advantage of being able to downweight outliers at the end of the day before building his model. Bob's batch and Ingrid's incremental algorithms each have their advantages and disadvantages, and some generalizations are made below.

Storage versus Processing

Incremental or online algorithms (e.g., naïve Bayes) clearly offer the advantage of being able to work (however poorly) with very little data, and can learn immediately when new data are acquired. Moreover, since instances are processed immediately, they do not need to be stored for later updating. One disadvantage is that online updating may require significant computational resources, perhaps at an inconvenient time. In contrast, batch (i.e., offline; e.g., support vector machines, decision trees) learning algorithms may need a large store of data and quite some time to build an initial useful model, and adding a single training instance may require iterating over the entire (and increasing) data store to update the model. A survey of learning algorithms will reveal the classic algorithmic tradeoff: one can store more, and process less upfront (but retrieval can be costly), or process more upfront and store less.

Another problem with many incremental algorithms is the potential to arrive at different learning outcomes based on the order the instances are encountered in. In many cases, such order effects are undesirable, but on the other hand humans and animals show a variety of order effects (e.g., in associative learning: Kachergis (2012). Could sleep be a chance to mitigate the order effects brought on during online learning? A few batch-update models have been found to have roughly-equivalent incremental versions. For example, latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990, LSA) learns semantic similarities of words via the singular value decomposition (SVD)–an expensive matrix operation–of a large word $\times$ document co-occurrence matrix. This large matrix–adults know over 70,000 unique words, and have read thousands of documents–must be kept in memory to be updated when a new document is read. Updating the model requires performing the SVD again, so it would be quite expensive to update knowledge every time a new document is read. It is more sensible to read a batch of documents–although, of course, this means that any new knowledge is not available in the model until the latest batch is incorporated. Fortunately, an incremental SVD algorithm has been proposed that not only is less computationally expensive, but also does not require storing the full word $\times$ document (Sarwar, Karypis, Konstan, & Riedl, 2002).

Models that use batch updating require storing all of the instances in long-term memory, allowing the model to iterate over all episodes–even multiple times–to extract higher-level features (e.g., correlations of multiple features). On the other hand, incremental updating can reduce the need to store so much information, much of which may be redundant or already over-learned. We conclude that sleep might be a way to get the best of both worlds: incremental learning based on salient features for immediate use, in addition to storage of daily episodes–especially exciting or confusing memories–that can be replayed during sleep to make more thorough, careful updates to knowledge representations before further compressing the memories.

Effects of Sleep

After a brief summary of the neurophysiological characteristics of sleep, we survey sleep effects found in various declarative and implicit memory tasks.

Structure of Sleep

Sleep in mammals and birds consists of cycles of four stages, proceeding from non-rapid eye movement (NREM) stages 1, 2, and 3 (also called slow-wave sleep), to rapid eye movement (REM) sleep. Human adults typically go through four or five cycles each night, reaching REM sleep every 90 minutes or so. More slow-wave sleep (SWS; NREM3) occurs early in the night, whereas more REM sleep occurs in the last few hours of a night's sleep. Each stage is characterized by particular muscle behaviors and brain activity (Schulz, 2008).

NREM1, between wakefulness and sleep, is recognized by somewhat active muscles (e.g., fluttering eyelids and rolling eyes) and alpha waves (7.5-12.5 Hz)–as in an awake state–transitioning to a theta rhythm (6-10 Hz).

NREM2 exhibits lower muscle activity and no awareness of surroundings, as well as a theta rhythm (6-10 Hz), with periodic sleep spindles (11-16 Hz) and K-complexes. It is thought that sleep spindles reflect the brain inhibiting processing, thereby keeping the sleeper asleep, and occur as a result of information flow between thalamic and cortical areas. Sleep spindles often occur together with K-complexes, which are basically delta waves lasting around one second. Their function is thought to be stimulus suppression as well as memory consolidation support.

Sleepers in SWS (NREM3) show very little reactivity to environmental stimuli, and predominantly delta wave activity ($<$ 3.5 Hz, high amplitude). SWS (i.e. deep sleep) is also when parasomnias such as sleepwalking and night terrors occur. This is the hardest part of sleep to wake up from, and sleep interruption during this stage often lead to feeling

groggy and sleepy. There is evidence to suggest that this is an essential stage of sleep: after sleep deprivation, the amount of SWS increases drastically.

Additionally, REM sleep is characterized by rapid, random eye movements as well as most memorable dreaming. At the same time, muscle ationa prevents dreamers from acting out their dreams. Similar to SWS, deprivation of REM sleep increases the amount of REM sleep in a recovery period, indicating that REM is crucial for normal functioning. From neural recordings of rats, it appears that memory replay during non-REM sleep occurs at a 10x speedup, whereas REM replay is roughly at the speed of the behavioral episode (Bendor & Wilson, 2012). During non-REM sleep, replay happens during short periods of increased activity in cortex and hippocampus related to cortical 'up-states' (i.e., frames). Both cortex and hippocampus replay similar content, but replay is initiated in the hippocampus, whereas cortical frames start roughly 50ms before hippocampal frame, leading researchers to conclude that the replay of sequential event memories may be driven by hippocampus, though the memories may be selected by cortex (Lee & Wilson, 2002).

One of the neurotransmitters that is often studied in relation to memory (especially Alzheimer's disease) and sleep is acetylcholine (ACh). ACh levels are lower during sleep than during awake, except during REM sleep when it reaches higher than awake levels. Hasselmo (1999) postulates that ACh inhibits feedback loops within the hippocampus and from it to the neocortex). This means that high ACh levels during waking hours (and REM) supports encoding new declarative memories, whereas low ACh during SWS allows replay of hippocampal memories, which are then stored more permanently in the neocortex.

Implicit Memory

Implicit or non-declarative memories are memories we do not have conscious access to, such as motor skills and procedural and perceptual memory. A positive influence of sleep on a finger tapping task has been demonstrated by Walker, Brakefield, Morgan, Hobson, and Stickgold (2002), which found that subjects performed 20% better after a night of sleep compared to subjects spending an equivalent amount of time awake. Furthermore, a correlation was found between the amount of stage 2 NREM sleep and performance improvement. The spindles that are characteristic of stage 2 sleep are thought to trigger specific intracellular mechanisms required for neural plasticity (Walker & Stickgold, 2006). More evidence for this specific influence of stage 2 NREM sleep comes from selective sleep deprivation studies. Smith and MacNeill (1994) found worse retention of a visuomotor adaptation task after stage 2 NREM sleep deprivation compared to REM sleep deprivation.

Similar spindles are visible in the somatosensory cortex of newborn rats and humans following spontaneous twitches during REM sleep, suggesting that learning is taking place in the brain following these movements (Khazipov et al., 2004). As infants spend relatively more of their sleep time in the REM stage compared to adults, muscle twitching during REM sleep may be a way for the body to build (especially in infants) and reinforce (in adults) specific neural pathways from brain to muscles that are frequently used. This mechanism may simply be a low-level version of a similar mechanism known as motor babbling, which allows infants to build higher-level action-effect associations.

Although most sleep-based consolidation studies have used implicit memory tasks, several have been done with declarative memory tasks.

Declarative Memory

The type of memory most people are familiar with is declarative or explicit memory. This type of memory provides us with knowledge of facts, such as Athens being the capital of Greece (semantic knowledge), as well as knowledge of personal events, like what you had for dinner last night (episodic knowledge).

Declarative memory is thought to be largely dependent on the hippocampus. Evidence supporting this theory consist of patient studies as well as various lesion studies in rats. One famous case includes memory disorder patient H.M., who was unable to form episodic memories after surgery removing a large part of his hippocampus. In rats, Eichenbaum (1990) showed that rats with hippocampal system lesions are severely impaired in the flexible use of previously learned information–normally attributed to episodic memory.

But what is exactly the mechanism by which the hippocampus performs this task? The prevailing hypothesis is that the hippocampus enables the consolidation of memory–i.e., converting an initial memory trace to a stable representation. Sleep is thought to play an important role in memory consolidation, for both declarative and implicit memory. During SWS, the episodic information that is stored in the hippocampus is replayed and projected to brain regions in the neocortex, which stores stable, permanent memories. The direction of this information flow reverses during the REM sleep that occurs later in time. It has been proposed that this allows the hippocampus to remove the unstable, short-term memories in order to make room for new memories to be stored there (Wamsley & Stickgold, 2011).

Indeed, there are many studies showing that sleep improves declarative memory using several paradigms. Retention of nonsense syllables has been shown to improve with sleep since (Jenkins & Dallenbach, 1924), and more recent research has shown improvement on a paired associates word list after SWS-rich sleep (Gais & Born, 2004).

In fact, it seems that not only sleep, but dreaming specifically can have beneficial effects on declarative memory. Wamsley, Tucker, Payne, Benavides, and Stickgold (2010) showed that after training on a virtual navigation task, improved performance at retest was correlated with relevant dream imagery during an afternoon nap. Similar thoughts during wakefulness, however, were not correlated with improved performance. It seems that reactivation of recently formed memories facilitates memory consolidation, with

dream imagery reflecting this process. Although recall shows more consistent sleep-related benefits than recognition, declarative memory often shows benefits even after only one to two hours of sleep (Diekelmann et al., 2009). Consolidation of declarative memories has been found to depend more on early hours of sleep, perhaps due to the dominance of SWS, and nondeclarative memory was more aided by the late, REM-dominated hours of sleep (Plihal & Born, 1999). Lee and Wilson (2002) ran rats repeatedly through a sequence of spatial receptive fields of hippocampal CA1 place cells, and during slow wave sleep (SWS) later ran through the same order of activation at a 20-fold temporal compression of the behavioral sequence, showing that the hippocampus is important for spatial learning–and likely the formation of long-term temporally-extended episodic memories in humans.

In general, SWS seems to aid declarative memory whereas REM enhances procedural and emotional memory in many cases (Mednick & Alaynick, 2010). However, in a study using a sequence learning task (the serial response time task), in which subjects were either explicitly told to learn a (repeating) sequence of button presses or learned it implicitly, the subjects learning explicitly improved only in the sleep condition, and to an extent that was correlated with the amount of NREM sleep (Robertson, Pascual-Leone, & Press, 2004). More intriguing was the fact that implicit learners all improved–regardless of sleep–over a twelve hour period, but not after only 15 minutes. Thus, it may be that for procedural memories, intentionally encoded memories are consolidated during sleep, whereas implicit memory consolidation may simply require time.

Theories of Sleep and Dreaming

An early attempt to explain why we dream was the activation-synthesis hypothesis (Hobson & McCarley, 1977), which posits that dreams come from the cortex trying to make sense of the random noise produced in the brainstem during REM sleep. We now know that dreaming also occurs during non-REM sleep, and even in deep, 'slow-wave' sleep.

Walker, Stickgold, Alsop, Gaab, and Schlaug (2005) proposes that there is a two-phase process by which memories are consolidated during sleep. The hippocampus, which stores recent episodic memories replays events during slow-wave sleep for the neocortex, where long-term memories reside. The communication between the two brain areas at this time is one way, from the hippocampus to the neocortex. During the REM dreaming that follows, though, the flow of information flips, from the neocortex back to the hippocampus. Stickgold suggested that once the neocortex connects the new memories to others in storage, it sends a message back to the hippocampus to erase them.

For declarative memory, there are two basic theories of how memory consolidation is improved during sleep: the active hypothesis states that consolidation depends on sleep, whereas the permissive hypothesis views consolidation as a time-dependent, interference-sensitive process that uses periods of low hippocampus input to process prior information (Mednick & Alaynick, 2010). Procedural memory is just generally thought to be 'enhanced' by sleep, but this idea is not universally accepted (Mednick & Alaynick, 2010). An early neural model found that offline replay helped maintain declarative memories using hippocampal-neocortical interplay (Kali & Dayan, 2004). We will focus on proposing specific computational mechanisms for improving declarative memory, since the current models are more readily adapted to this task, and the empirical evidence indicating the necessity for this is strong.

Model

Here we will propose and test two mechanisms of sleep-based consolidation that can be added to modern computational models of episodic memory. Our modifications will be specified in terms of the REM[1] (Retrieving Effectively from Memory) model Shiffrin and Steyvers (1997), which is a multitrace memory model that makes optimal recognition decisions assuming that memory is subject to noise. In multitrace memory models, a memory trace is represented by a large vector of feature values representing the context and content of the event. Some features may be abstract and learned, whereas others are assumed to be simple and sensory-based (i.e., primitive). REM has both episodic traces as well as lexical-semantic traces, the latter of which are decontextualized accumulations of the various episodic traces, updated across a lifetime. In REM, individual traces are assigned random feature values, each drawn from a geometric distribution with parameter g. That is, the probability that a feature has value v is

$$P(v) = g(1-g)^{v-1} \tag{1}$$

The geometric distribution makes small feature values more probable (and thus frequent) than large values. REM uses these varying base rates in calculating evidence for a memory match: matching a low (common) feature value is not as strong evidence as matching a high value. Conscious experience activates the lexical-semantic (LS) trace for the attended stimuli and updates it with the current context features. An episodic memory trace is formed by copying context and LS features with probability u per time unit t. When a feature is stored, it is only copied correctly with probability c; otherwise, a random value is drawn according to Equation 1. Missing features have a value of 0 (uninformative). Our first proposed modification is to update lexical-semantic (LS) features (i.e., neocortical representations) during a sleep period, when episodic traces since the last sleep period are (randomly, although it may be more accurate to prioritize surprising or emotionally-charged traces) reactivated. The detailed lexical re-encoding mechanism is described in detail below. Updating LS features during sleep leaves the hippocampal episodic traces available for retrieval and recognition throughout the day. REM assumes that when the same stimuli appear multiple times in similar contexts, the old trace may be updated by

[1] In this section, REM refers only to the model.

filling in missing features from LS traces, instead of making a new trace (this differentiation process is how it accounts for the word frequency mirror effect and null list strength effect).

Retrieval in REM uses context features–reinstated by the probe, whatever its source (internal or external)–to activate a subset of long-term memory (e.g., to the studied list of items). This is in part done for simplicity of simulation–so that a number of extra-experimental parameters and processes need not be assumed, but is also a reasonable thing to expect from memory. For recognition, REM computes a likelihood ratio indicating how well test cue j (from the LS traces) match each episodic trace i in the activated subset being considered. This likelihood ratio, in which g_{sys} is the base rate in the long-term, n_q and n_m are the number of non-zero mismatching and matching features, respectively, is defined

$$\lambda_{i,j} = (1-c)^{n_q} \prod_{v=1}^{\infty} \left[\frac{c + (1-c) g_{sys} \cdot (1-g_{sys})^{v-1}}{g_{sys}(1-g_{sys})^{v-1}} \right]^{n_m} \quad (2)$$

Thus, the decision depends on not only the number of matching features, but also on how diagnostic the features are: large values are more rare in the geometric distribution, and are thus more informative. Therefore, prior knowledge about the base rate of different values (g_{sys}) may affect recognition. REM decides that cue j was presented on the list if the average of the likelihood ratios is above a criterion; else it is rejected. Since small feature values will tend to be quite common and thus undiagnostic, whereas the more useful large feature values are rarely encountered, a potential mechanism for improving memory would be to redistribute feature values. By choosing one or more unique, diagnostic features for each trace (or set of highly-related traces), memory will be improved. Although an exhaustive cross-check to make sure features are unique would be quite computationally expensive (which is why it should be offline), but a simple, greedy version might work well, with high probability. We now describe and test two consolidation methods that we have hinted at.

Simulation

Using REM as the basis for storing and retrieving memories, we define and test two sleep-based consolidation mechanisms in a simulation of studying 100 items 1, 3, or 5 times[2]. The *reactivation method* simply reactivates n (here, $n = 50$) random memory traces from the day, and then for each of these memories m finds the best matching LS trace t and copies in missing (0-valued) features from t to m. In essence, the reactivation mechanism assumes that people dream about recent memories and fill in missing details in the traces using background knowledge. Although these traces are strengthened, the underlying knowledge representations (the LS traces) remain static. In contrast, the *re-encoding method* randomly selects r LS traces (here, $r = 50$) for re-encoding, with probability proportional to their familiarity in recent memory (here, the day). For each of these r LS traces, f features (here, $f = 6$) are randomly chosen, and those feature values are incremented by 1, thus making those features more diagnostic, and the item easier to remember in the future. However, now that the LS (i.e., neocortical) representation has changed, memories that were encoded with the old representation will be more difficult to retrieve. This can be largely remedied by probing memory for copies of the old representation while re-encoding, and updating those items in memory, as well: simply increment whatever is stored for those f features. Figure 1 shows the mean d' ($z(H) - z(FA)$) of 50 simulated subjects after studying 100 items one, three, or five times, using REM alone or either of the proposed consolidation mechanisms, both of which aid memory. Note that these two methods are complementary, and could be used together with greater effect. The effect of reactivation lessens with more item repetitions, while the re-encoding effect can be quite strong, regardless of repetitions.

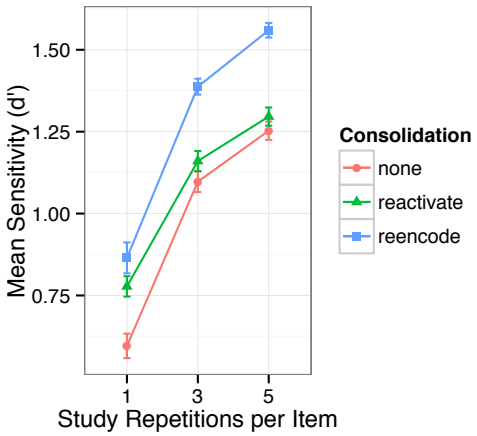


Figure 1: Mean d' for REM simulations using no consolidation, or one of reactivation- or re-encoding-based consolidation after studying 100 items 1, 3, or 5 times. Error bars show +/-1SE for 50 simulated subjects.

Variations of these two methods are quite plausible for modeling sleep-based consolidation in current memory models. However, at least one important aspect is clearly missing: structured sequences of events must be linked in some way, but are not in the REM framework (what defines an 'event' is not yet known). Future models should address how event boundaries are marked. It may be possible–and would be desirable, for compression–to identify routine sequences (e.g., eating your typical breakfast, or driving home from work via your usual route) and to replace the specific feature values in traces containing these sequences with placeholders pointing to the appropriate LS representation, thus simplifying future updating. Modeling procedural sleep-consolidation effects may require distributed neural models such as Kachergis, Wyatt, O'Reilly, Kleijn, and Hommel (2014).

Discussion

We have proposed two computational mechanisms that could beneficially operate on memories during sleep, along with a rationale for why humans may find it advantageous to be

[2] https://github.com/kachergis/REM_consolidation

capable of learning by both incremental and batch updating. The proposed mechanisms are all fairly computationally expensive–involving many comparisons and updates to long-term lexical-semantic traces that are presumably stored in neocortex, making them suitable for conducting during sleep. Note that while a complex version of redistributing diagnostic feature values would have to be done in batch, the simple greedy version used here (choosing random features to increment) is more batch-incremental. Although we specified these mechanisms in terms of the REM (Shiffrin & Steyvers, 1997) model, the same mechanisms could be used in related multitrace modeling frameworks such as SARKAE (Nelson & Shiffrin, 2013) or MINERVA2 (Hintzman, 1984). We proposed the reactivation and re-encoding mechanisms based on our survey of sleep-related effects on a variety of declarative and procedural tasks, which we found to convincingly implicate sleep as beneficial to memory consolidation. Our modeling confirms that the proposed methods benefit memory, and offers an avenue for both predicting and matching detailed results from sleep studies.

Acknowledgments

The preparation of this work was supported by the European Commission (EU Cognitive Systems project ROBOHOW.COG; FP7-ICT-2011).

References

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature Neuroscience*, *15*(10), 1439–1444.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Jour. of the American Soc. for Info. Sci.*, *41*(6).

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*(5), 309–321.

Gais, S., & Born, J. (2004). Declarative memory consolidation: Mechanisms acting during human sleep. *Learning & Memory*, *11*(6), 679–685.

Hasselmo, M. E. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in CogSci*, *3*, 351–359.

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods: Instrument and Computers*, *76*, 96–101.

Hobson, J. A., & McCarley, R. W. (1977). The brain as a dream state generator: An activation-synthesis hypothesis of the dream process. *Am. J. of Psych.*, *134*(12), 1335–48.

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *American Journal of Psychology*, *35*(4), 605–612.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Kachergis, G. (2012). Learning nouns with domain-general associative learning mechanisms. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proc. of cogsci 34* (p. 533-538). Austin, TX: Cog. Sci. Society.

Kachergis, G., Wyatt, D., O'Reilly, R. C., Kleijn, R. de, & Hommel, B. (2014). A continuous time neural model for sequential action. *Phil. Trans. of the Royal Soc: B*, *369*.

Kali, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neurosci.*, *7*(286–294).

Khazipov, R., Sirota, A., Leinekugel, X., Holmes, G. L., Ben-Ari, Y., & Buzsaki, G. (2004). Early motor activity drives spindle bursts in the developing somatosensory cortex. *Nature*, *432*(7018), 758–761.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194.

Mednick, S. C., & Alaynick, W. A. (2010). Comparing models of sleep-dependent memory consolidation. *Journal of Experimental Clinical Medicine*, *2*(4), 156–164.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.

Plihal, W., & Born, J. (1999). Effects of early and late nocturnal sleep on priming and spatial memory. *Psychophysiology*, *36*(5), 571–582.

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. *Current Biology*, *14*(3), 208–212.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2002). Incremental singular value decomposition algorithms for highly scalable recommender systems. *5th International Conference on Computer and Information Technology*.

Schulz, H. (2008). Rethinking sleep analysis. *Journal of Clinical Sleep Medicine*, *4*(2), 99–103.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM–retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*(2), 145–166.

Siegel, J. (2013). The evolution of sleep. In *Encyclopedia of sleep* (Vol. 1). Waltham, MA: Elsevier.

Smith, C., & MacNeill, C. (1994). Impaired motor memory for a pursuit rotor task following stage 2 sleep loss in college students. *Journal of Sleep Research*, *3*(4), 206–213.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron*, *35*, 205–211.

Walker, M. P., & Stickgold, R. (2006). Sleep, memory, and plasticity. *Annual Review of Psychology*, *57*, 139–166.

Walker, M. P., Stickgold, R., Alsop, D., Gaab, N., & Schlaug, G. (2005). Sleep-dependent motor memory plasticity in the human brain. *Neuroscience*, *133*(4), 911–917.

Wamsley, E. J., & Stickgold, R. (2011). Memory, sleep, and dreaming: Experiencing consolidation. *Sleep Medicine Clinics*, *6*(1), 97–108.

Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A., & Stickgold, R. (2010). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology*, *20*, 850—855.

Solving the knowledge-behavior gap: Numerical cognition explains age-related changes in fairness

Nadia Chernyak (chernyak@bu.edu)[1,4], Beth Sandham (bs418@bath.ac.uk)[2], Paul L. Harris (paul_harris@gse.harvard.edu)[3], & Sara Cordes (cordess@bc.edu)[4]

[1]Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215
[2]Department of Psychology, University of Bath, Bath, UK BA2 7AY
[3]Graduate School of Education, Harvard University, Cambridge, MA 02138
[4]Department of Psychology, Boston College, Chestnut Hill, MA 02467

Abstract

Young children share fairly and expect others to do the same. Yet little is known about the underlying cognitive mechanisms that support fairness. Across two experiments, we investigated whether children's numerical competencies are linked with their sharing behavior. Preschoolers (aged 2.5-5.5) participated in either third-party (Experiment 1) or first-party (Experiment 2) resource allocation tasks. Children's numerical competence was then assessed using the Give-N-Task (Sarnecka & Carey, 2008; Wynn, 1990). Numerical competence – specifically knowledge of the cardinal principle explained age-related changes in fair sharing in both the third- and first-party contexts. These results suggest that an understanding of the cardinal principle serves as an important mechanism for fair sharing behavior.

Keywords: fairness; numerical cognition; preschoolers; knowledge-behavior gap

Introduction

By the preschool age, young children show a remarkable concern for fairness. This concern appears in their explicit endorsements of fairness as a social norm (Damon, 1977; Smith, Blake, & Harris, 2013), in their affective responses towards inequalities (LoBue, Nishida, Chiong, DeLoache, & Haidt, 2011), and in their own distribution of resources (Olson & Spelke, 2008). By middle childhood, children go to great lengths to be fair and to appear fair to others, even in third-party contexts, in which they presumably have no stake: they discard resources in order to avoid inequalities (Shaw & Olson, 2012) and they spontaneously correct an adult's unequal distribution of resources (Paulus, Gillis, Li, & Moore, 2013). Thus, children show sufficient motivation to share at an early age – they recognize what fairness is and understand why it is important.

In spite of their interest in fairness, it is remarkable that many children do not typically behave fairly themselves (Blake, McAuliffe, & Warneken, 2014; Sheskin, Chevallier, Lambert, & Baumard, 2014). Prior work has consistently documented that fair sharing emerges late in the preschool period and into middle childhood (e.g., Fehr, Ernst, & Rockenbach, 2008). As such, researchers have recently begun to become interested in what is now termed the "knowledge-behavior gap": although young children recognize situations that are fair or unfair (Schmidt & Sommerville, 2012; Sloane, Baillargeon, & Premack, 2011), they do not always behave fairly themselves (see Smith et al., 2013). Why children show a knowledge-behavior gap has been of recent theoretical interest (e.g., Blake et al., 2014). In our work, we explore one possibility: the extent to which children's understanding of number underpins their abilities to share fairly with others.

Sharing is inherently a number-based problem. In order to understand how six candies should be shared between three people, one should understand that 6 divided by 3 results in 2 candies each. Similarly, a division is unfair if it does not show cardinal equivalence – if, for example, one person had 3 candies, and another received 1. Such simple numerical calculations underlie our understanding of higher-order concepts such as fairness, equality, and generosity. In spite of the clear connection between sharing and numerical cognition, few studies have charted how number knowledge might relate to children's resource distribution.

Several earlier findings suggest that numerical cognition and sharing are, in fact, related: First, older work has found that in middle childhood, children come to understand concepts of division through the action schema of sharing (Correa, Nunes, & Bryant, 1998; Desforges & Desforges, 1980; Frydman & Bryant, 1988; Squire & Bryant, 2002a; 2002b). That is, children find division problems easier when they are presented with the end goal of sharing resources fairly (e.g., "How many candies are in each box?") than when the end goal is to figure out the number of recipients present during fair sharing (e.g., "How many boxes did we use?"). Second, as compared with older children, younger children, with presumably limited numeracy skills, have a harder time splitting resources fairly and fail to recognize the connection between sharing and cardinal equivalence (e.g., don't recall correctly that each recipient has the same amount; Frydman & Bryant, 1988; see also Pepper & Hunting, 1998).

The converging evidence suggests that numerical competencies might be one cognitive prerequisite to fair sharing. Thus, although children may be able to recognize fair sharing from a young age, they do not yet possess the requisite numerical skills to produce fair shares. In our studies, we looked at the relationship between children's sharing behavior and numerical cognition.

In our first experiment, we began by looking at young children's sharing behavior in a third-party context (i.e., sharing between two recipients when there is no cost to the self). In the second, we looked at a more stringent test of fair sharing: sharing in a first-party context, in which children shared between themselves and another recipient. Across both experiments, we also assessed children's numerical understanding using a version of the Give-N task (Sarnecka & Carey, 2008; Wynn, 1992), used to classify young children according to whether they know and understand the cardinal principle (i.e., that the purpose of counting is to determine to cardinality of a set).

Experiment 1

Participants

Seventy-three children (28 male, 45 female) were tested at a local children's museum or during a laboratory visit (Mean age=3y;8m, Range = 2y;6m–5y;6m). Nineteen additional children were excluded due to experimenter error (n = 5), parental/sibling interference (n = 7), failure to complete the task (n = 3), or due to no video (either recording error or lack of parental consent to video record; n = 5).

Procedure

Materials In the resource distribution tasks, we used four plush animals (hedgehog, panda, dog, and elephant), four wooden boxes – one for each puppet with pictures of the puppet on the tops and insides, and 10 small dinosaur toys. Materials for the Give-N task were a set of small yellow rubber ducks and a blue basket used to symbolize a pond.

Resource Distribution Tasks Children were presented with two trials: one in which they were presented with Four Resources and one in which they were presented with Six Resources (order counterbalanced). In each trial, the child was introduced to two puppets in succession (e.g., "Doggie" and "Ellie") and a set of dinosaur toys (either four or six). The researcher arranged the toys linearly between the two puppets and pointed to each toy in turn without any verbal counting. The researcher then placed two wooden boxes in front of the puppets and told the child that s/he would get to split the toys between the puppets: "You get to decide which toys to give to [Recipient 1] and which ones to give to [Recipient 2]. So whichever toys you want to give to [Recipient 1], you can put right here [point to Recipient 1's box], and whichever ones you want to give to [Recipient 2], you can put right here [point to Recipient 2's box]". Children who left any toys on the table were re-prompted ("And what do you want to do with these?") until all toys were placed into the boxes. Children were asked additional follow-up questions regarding their memory of the number of resources they shared and sharing justifications (not further discussed or analyzed here).

Give-N Task Following the two resource sharing tasks, each child completed a version of the Give-N task (Sarnecka & Carey, 2008; Wynn, 1990, 1992) to determine whether the child understood the cardinal principle (that the purpose of counting is to determine the number of items in a set). Children were given a set (approximately 10 ducks) and a blue basket (a "pond"). They were then asked to place N ducks "into the pond" (basket), in which N varied from 1-6. On every trial, after responding, children were asked "Is that N ducks?" and were allowed to change their response if they wanted. The experimenter first asked for 1 duck and continued to 3 if the child correctly placed 1 duck into the pond. The experimenter then continued to ask for N+1 if the child successfully placed N ducks into the pond or for N-1 if the child failed to place N ducks into the pond. The experiment concluded after the child either (a) had two successes on N and two failures on N+1, or (b) succeeded twice on N = 6 ducks. Following prior procedures (LeCorre & Carey, 2007), children were classified as either Subset Knowers (children who could count properly up to a specific subset of under 6 items but no more, and thus failed to understand that the purpose of counting is to determine cardinality; n = 42) or Cardinal Principle (CP) Knowers (proficient counters; n = 31).

Results

Our first question concerned whether there were age-related differences in fair sharing (defined as an equal split between the two animals). We ran a binary logistic regression using age, gender, and trial type (four vs. six; entered as a within-subjects effect) as the predictors and likelihood of fair sharing as the response (Model 1; see Table 1). There was a significant effect of Age, B = 0.92, SE(B) = 0.30, 95%CI = 0.34–1.50, Wald(1) = 9.60, p = .002, with older children being more likely to share fairly. Additionally, there was also a significant effect of Gender (with females sharing more fairly than males), B = 1.27, SE(B) = 0.44, 95%CI = 0.40–2.13, Wald(1) = 8.18, p = .004, and no effect of Trial Type, p = 0.83. Therefore, as children aged, they became more likely to share fairly.

Table 1: Parameter Estimates (and Standard Errors) for Models Used in Experiment 1

	Model 1	Model 2	Model 3
Gender (1=Female)	**1.27 (0.44)****	**1.13 (0.44)****	**1.26 (0.44)****
Age	**0.92 (0.30)****	--	0.51 (0.35)
Trial Type (1=6 Resources)	-0.07 (0.34)	-0.07 (0.34)	-0.07 (0.35)
CP Knowledge (1=CP Knower)	--	**1.57 (0.48)****	**1.10 (0.56)***

Note: Response = likelihood of sharing fairly. Significant effects are in bold. *$p \leq 0.05$; **$p \leq 0.01$.

Our next – and focal – research question was whether children's acquisition of cardinal principle (CP knowledge) might explain age-related differences in fair sharing. We first tested whether CP knowledge related to fair sharing. To investigate whether CP knowledge predicted fair sharing, we re-ran Model 1, but used CP knowledge (CP knowers vs. subset knowers, coded as 1 and 0 respectively) instead of age as a predictor (Model 2). There was a significant effect of CP Knowledge, B = 1.57, SE(B) = 0.48, 95%CI = 0.64 – 2.50 Wald(1) = 10.92, p = .001. As with the first Model, there was a significant effect of Gender, B = 1.13, SE(B) = 0.44, 95%CI = 0.27 – 1.98, Wald(1) = 6.62, p = .01, and no effect of Trial Type, p = 0.83. Therefore, understanding of the cardinal principle predicted children's abilities to share fairly.

We then confirmed that Age was significantly associated with CP Knowledge, B = 2.29, SE(B) = 0.54, Wald(1) = 18.15, p = 0.00002.

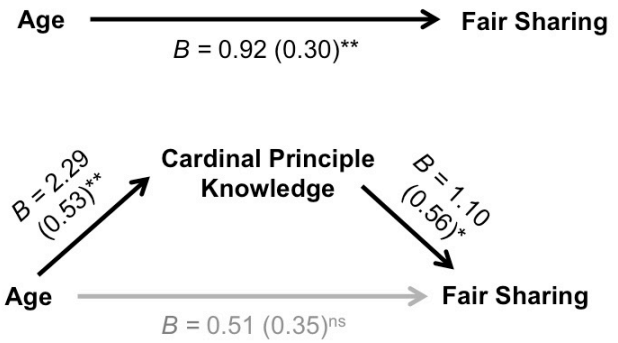


Figure 1: Mediation analysis in Experiment 1

As a last step, we tested whether CP knowledge mediated the effect of age on fair sharing. Here we combined Models 1 and 2 and used both age and CP knowledge as predictors (Model 3). With CP Knowledge added to the model, there was a significant effect of CP Knowledge (such that CP knowers were more likely to share fairly than subset knowers), B = 1.10, SE(B) = 0.56, 95%CI = -0.007 – 2.20, Wald(1) = 3.79, p = .05, and no longer any significant effect of Age, p = 0.15. As with the previous models, there was also a significant effect of Gender, B = 1.26, SE(B) = 0.44, 95%CI = 0.39 – 2.13, Wald(1) = 8.00, p = .005 and no significant effect of Trial Type, p = 0.83. A formal mediation test suggested that CP Knowledge mediated age-related differences in fair sharing, Sobel test z = 1.79, p = 0.07 (see Figure 1). Therefore, knowledge of the cardinal principle, and not age, predicted children's abilities to share fairly across both trials.

Discussion

Our results show that understanding the cardinal principle explained age-related changes in third-party sharing, suggesting that numerical cognition serves as an important mechanism for fairness as children age. As children acquire cardinality, they become more adept at social skills such as dividing resources equally between two recipients.

Because third-party sharing involves a situation in which there is no cost to the self and because third-party sharing is something that even infants expect others to do (Schmidt & Sommerville, 2012), our task was able to look at sharing ability in a context unconfounded with potential motivational concerns. That is, our results suggest that even in a context in which prior has found that children are likely motivated to share equally (e.g., Olson & Spelke, 2008), only children who had acquired cardinality had the requisite skills are able to do so.

In Experiment 2, we looked at a more stringent test of fair sharing: first-party sharing. Prior work has consistently documented age-related differences in first-party fairness (e.g., Fehr, Bernard, Rockenbach, 2008; Smith et al., 2013), but as with third-party fairness, the mechanism driving this increase is not yet clear. Moreover, unlike third-party fairness, first-party fairness requires additional cognitive skills (e.g., inhibitory control), which develop with age during the preschool period. We were interested in whether numerical cognition would continue to predict fair sharing behavior, even in a context in which fairness is costly to the child.

Experiment 2

Participants

Ninety-two children (37 male, 55 female) were tested at a local children's museum or during a laboratory visit (Mean age = 3y;10m, Range = 2y;6m – 5y;4m)[1]. Eighteen additional children were excluded due to either protocol error (n = 2), failure to complete the task (n = 9), missing video (either equipment failure or lack of parental consent to video record (n = 4)), or being outside the a priori targeted age range (n = 3).

Procedure

Materials Materials were the same as Experiment 1, with the following modifications: (a) only one stuffed animal recipient was used in each resource distribution trial (the second recipient was the child), and (b) stickers instead of dinosaur toys were used as resources.

Resource Distribution Tasks Because first-party sharing with anonymous recipients has been shown to be particularly difficult for preschool-aged children, we contextualized the situation and used a modified procedure that has been shown to successfully induce sharing in preschoolers (Chernyak & Kushnir, 2013). Children were introduced to a puppet that was described as feeling "very sad", and told that they could give him/her some stickers to make the puppet feel better. The distribution phase

[1] Three children's parents did not provide a date of birth but identified them as being within the proper age range for the study. Their ages are not recorded in final calculations. They are by necessity excluded from any analyses with age.

proceeded in exactly the same manner as in Experiment 1. Children were then shown a set of linearly arranged stickers (either four in the Four Resource Trial or six in the Six Resource Trial; order counterbalanced) arranged linearly between two boxes (one labeled as the puppet's box which had pictures of the puppet on the top and inside; and the other labeled as the child's box which had no pictures), and told they could split the stickers however they wished between themselves and the puppet.

Give-N Task After the two resource distribution tasks, children completed a Give-N task, which proceeded in exactly the same manner as in Experiment 1. Children were classified as either Subset Knowers (48 children) or CP knowers (44 children).

Results

Overall, the rates of fair sharing were lower in Experiment 2 (first-party sharing) than in Experiment 1 (third-party sharing). A direct comparison of the proportion of fair sharing trials across the two experiments revealed that children were more likely to be fair in Experiment 1 (n = 99 of 146) than in Experiment 2 (n = 103 of 184), Fisher's exact test, p = 0.03. Therefore, children were less likely to be fair when sharing in a first-party context. We then asked whether CP knowledge nonetheless continued to predict children's fair sharing.

As with Experiment 1, our first question concerned whether there were age-related differences in fair sharing. We ran a binary logistic regression using age, gender, and trial type (four vs. six; entered as a within-subjects effect) as the predictors and likelihood of fair sharing as the response (Model 1; see Table 2). There was a significant effect of Age, B = 0.84, SE(B) = 0.25, 95%CI = 0.35 – 1.34, Wald(1) = 11.07, p = .001, a significant effect of Trial Type (with children being more likely to share fairly in the Four Resource trial), B = -0.88, SE(B) = 0.27, 95%CI = -1.41 - -0.35, Wald(1) = 10.72, p = .001, and no effect of Gender, p = .26. Therefore, as children aged, they became more likely to share fairly in a first-party context.

Table 2: Parameter Estimates (and Standard Errors) for Models Used in Experiment 2

	Model 1	Model 2	Model 3
Gender (1=Female)	0.44 (0.39)	0.74 (0.39)	0.51 (.40)
Age	**0.84 (0.25)****	--	0.19 (.32)
Trial Type (1=6 Resources)	**-0.88 (0.27)****	**-0.88 (0.27)****	**-0.95 (.29)****
CP Knowledge (1=CP Knower)	--	**1.54 (0.39)****	**1.58 (0.52)****

Note: Response = likelihood of sharing fairly. Significant effects are in bold. *$p \leq 0.05$; **$p \leq 0.01$.

Our next – and focal – research question was whether children's CP knowledge might explain age-related differences in fair sharing. To investigate whether CP knowledge predicted fair sharing, we re-ran Model 1, but used CP knowledge instead of Age as a predictor (Model 2). Confirming the idea that acquiring cardinality predicts first-party fairness, there was a significant effect of CP Knowledge, B = 1.54, SE(B) = 0.39, 95%CI = 0.77 – 2.30, Wald(1) = 15.42, p = .00008. As with the first Model, there was a significant effect of Trial Type (with children being more likely to share fairly in the Four Resource trial), B = -0.88, SE(B) = 0.27, 95%CI = 0.35 – 1.40, Wald(1) = 10.75, p = .001, and a marginally significant effect of Gender (with females being more likely to share fairly than males), p = 0.06. There were no interactions with CP knowledge or trial type (p = .64). Therefore, understanding the cardinal principle predicted children's abilities to share fairly even in a first-party context.

We then confirmed that as in Experiment 1, age was significantly associated with CP Knowledge, B = 2.51, SE(B) = 0.50, Wald(1) = 25.62, p = 0.0000004.

As a last step, we tested whether CP knowledge mediated the effect of age on fair sharing. As with Experiment 1, we combined Models 1 and 2 and used both age and CP knowledge as predictors (Model 3). There was a continued significant effect of CP knowledge (such that CP knowers were more likely to share fairly than subset knowers), B = 1.58, SE(B) = 0.52, 95%CI = 0.57 – 2.60, Wald(1) = 9.36, p = .002, and no longer any significant effect of Age, p = 0.55. As with the previous models, there was also a significant effect of Trial Type, B = -0.95, SE(B) = 0.29, 95%CI = -1.53 – -0.38, Wald(1) = 10.55, p = .001 and no significant effect of Gender, p = 0.20. A formal mediation test confirmed that CP knowledge fully mediated age-related differences in fair sharing, Sobel test z = 2.60, p = 0.009 (see Figure 2). Therefore, knowledge of the cardinal principle, and not age, predicted children's abilities to share fairly in our first-party context.

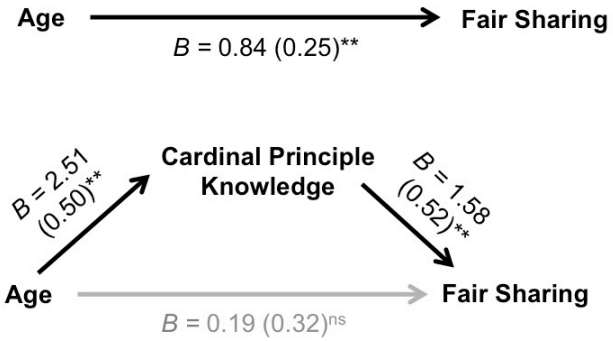


Figure 2: Mediation analysis in Experiment 2

We were also interested in the types of errors children make when they failed to share fairly. We looked at trials during which children had *not* shared fairly (n = 81 of a total of 184 trials). One possibility is that young children tend to default to selfishness. If this is the case, we might expect

that younger children will consistently behave selfishly and keep most resources for themselves. Another possibility, however, is that children are motivated to share fairly, but do not always have the cognitive skills (i.e., numerical cognition) in order to do so. If this is the case, children should be just as likely to make either fair sharing error (either generous sharing or selfish sharing).

Confirming the latter possibility, children did not strategically default to selfishness: Of the trials during which they did not share fairly, approximately half were generous (n = 42 of 81; 52%) and the other half were selfish (n = 39, 48%).

We also tested whether CP knowledge predicted how close children's errors were to fair sharing behavior. We computed a "Difference from Fairness" score (DFS) reflecting the magnitude of the deviation of the child's sharing behavior from perfect fair sharing behavior. The DFS reflected the absolute value of the difference between the number of resources the child gave to the puppet (child's sharing behavior) and the number reflecting fair sharing (i.e., 3 in the six resource trial and 2 in the four resource trial). Thus, children who made "greater" errors and deviated more from fair sharing received higher DFS scores.

We ran a Poisson model using DFS score as the response and CP knowledge, Age, Trial Type (entered as a within-subjects effect), and Gender as the predictors. The results revealed a significant effect of CP knowledge, B = -.33, SE(B) = .17, 95%CI = -0.66 - -0.002, Wald(1) = 3.69, p = .05, and no other significant effects, all p's > .25. Therefore, CP knowers had lower DFS scores, suggesting that children with limited numerical cognition also deviated to a greater extent from fair sharing behavior.[2]

To make sure that this result could not be explained by children simply becoming more generous as they aged (and thus becoming more fair in the selfish subsample), we also ran this model using number of stickers donated to the puppet as a response. There were no significant effects in either the subsample of kids who had not shared fairly, or in the full dataset. This rules out the possibility that as children age and gain numeracy skills, they simply become less interested in the resources or more generous with them. Instead, this supports the idea that as children acquire cardinality, their behavior approaches fairness.

Discussion

As in Experiment 1, numerical cognition predicted and explained age-related differences in children's abilities to share fairly, even in a context in which children may be particularly motivated to share unfairly (be selfish). Children who did *not* share fairly did not strategically default to selfishness – instead, they were just as likely to be generous as they were to be selfish. CP knowers made errors that were smaller in magnitude (closer to fairness) than subset knowers. These results suggest that even in a context in which children may have been motivated to be unfair, numerical cognition continued to predict fair sharing behavior.

General Discussion

A body of recent work has documented the early development of our concern for others' welfare. We add to this work by showing that between the ages of 2.5 and 5.5, children develop the capacity to distribute resources fairly.. Importantly, our results highlight a key cognitive mechanism that enables such sharing behavior. Numerical cognition predicted young children's abilities to share fairly and their sharing strategies. Strikingly, numerical cognition also mediated age-related changes in fair sharing.

Recent work has found that young children recognize fairness before age two (e.g., Schmidt & Sommerville, 2012), but do not necessarily act fairly themselves (e.g., Fehr et al., 2008; Posid, Fazio, & Cordes, 2015; Smith et al., 2013). Our findings have two important implications for fairness. First, we find that sufficient motivation is *not* enough to enable fairness in young children. Even in a third-party context, in which children are motivated to share equally (Olson & Spelke, 2008), children were not always able to do so. This suggests that fairness involves not only sufficient motivation, but also the coordination of later-developing socio-cognitive abilities.

Second, we find that nonsymbolic numerical abilities are also not enough to enable fair distribution behavior. One possibility could have been that sharing 4 stickers would require only the ability to discriminate between 1 vs. 3 items, or between a 2/2 distribution and a 1/3 distribution - a problem that even infants are capable of solving (Feigenson & Carey, 2003; Schmidt & Sommerville, 2012). Instead, however, our results found that knowledge of cardinality and explicit understanding of counting was critical to children's abilities to act fairly. This suggests that while nonsymbolic numerical abilities may be implied in children's passive and implicit understanding of fairness, active manipulation of resources requires the later-developing ability of counting proficiency.

Future work might focus on the mechanism by which the acquisition of cardinality scaffolds sharing decisions. One possibility is that children already have a rudimentary understanding of equality as a social norm (Schmidt & Sommerville, 2011; Geraci & Surian, 2011; Sloan, Baillargeon, & Premack, 2012), and that acquiring the cardinal principle simply helps children realize that norm in their own behavior.

Our work focused on how knowledge of the cardinal principle is related to third-party and first-party resource distribution. However, both fairness and numerical cognition involve a host of other sub-component competencies that follow distinct developmental pathways. For example, fairness involves the ability to engage in third-party moral evaluation of others (thought to be an early-developing capacity; Schmidt & Sommerville, 2012) as well as the ability to engage in altruistic sharing (e.g., Svetlova,

[2] We did not find any relationship between DFS scores and CP knowledge in Experiment 1.

Nichols, & Brownell, 2010). Similarly, numerical cognition includes the approximate number system (ANS; e.g., Xu & Spelke, 2000), knowledge of the cardinal principle (e.g., Wynn, 1990), the ability to map symbols onto their respective magnitudes (e.g., Mundy & Gilmore, 2009), and other related competencies (e.g., school-based arithmetic). Future work is warranted to explore the specific numerical competencies (and cognitive systems) that underlie each type of social behavior.

More generally, our work points to important links between social and cognitive development in early childhood. This link is important to consider from the perspective of developing young children's number knowledge and their sharing behavior: For example, it is important to keep in mind the child's individual cognitive competencies (e.g., number knowledge) when studying social behavior. Similarly, giving children experiences with sharing may help their numerical understanding. More generally, bridging social and cognitive development may help us gain better insights into the developmental processes of each.

Acknowledgments

We would like to thank members of the Boston College Infant and Child Cognition Lab, especially Emma Lazaroff, Chloe Ang, Becky Krol, Jade Dunstan, Monica Lee, Olivia Noe, and Karina Hamamouche for assistance with data collection and coding; the Boston Children's Museum and the *Living Laboratory* at the Museum of Science, Boston; and participating families. This work was funded by Grant #56348 from the John Templeton Foundation.

References

Blake, P. R., McAuliffe, K., & Warneken, F. (2014). The developmental origins of fairness: the knowledge–behavior gap. *Trends in Cognitive Sciences*, *18*, 559-561.

Chernyak, N., & Kushnir, T. (2013). Giving preschoolers choice increases sharing behavior. *Psychological Science*, 24, 1971-1979.

Correa, J., Nunes, T., & Bryant, P. (1998). Young children's understanding of division: The relationship between division terms in a noncomputational task. *Journal of Educational Psychology*, *90*, 321-329.

Desforges, A., & Desforges, C. (1980). Number‐based strategies of sharing in young children. *Educational Studies*, *6*, 97-109.

Damon, W. (1979). The Social World of the Child. San Francisco: Jossey-Bass.

Fehr, E., Bernhard, H., & Rockenbach, B. (2008). Egalitarianism in young children. *Nature*, *454*, 1079-1083.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object‐files: evidence from infants' manual search. *Developmental Science, 6,* 568-584.

Frydman, O., & Bryant, P. (1988). Sharing and the understanding of number equivalence by young children. *Cognitive Development*, *3*, 323-339.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*, 395-438.

LoBue, V., Nishida, T., Chiong, C., DeLoache, J. S., & Haidt, J. (2011). When getting something good is bad: Even three‐year‐olds react to inequality. *Social Development*, *20*, 154-170.

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, *103*, 490-502.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*, 222-231.

Paulus, M., Gillis, S., Li, J., & Moore, C. (2013). Preschool children involve a third party in a dyadic sharing situation based on fairness. *Journal of Experimental Child Psychology, 116,* 78-85.

Pepper, K. L., & Hunting, R. P. (1998). Preschoolers' counting and sharing. *Journal for Research in Mathematics Education*, *29,* 164-183.

Posid, T., Fazio, A., & Cordes, S. (2015). Being sticker rich: Numerical context influences children's sharing behavior. *PLoS ONE*, e0138928.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*, 662-674.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, *141*, 382.

Schmidt, M. F., & Sommerville, J. A. (2012). Fairness expectations and altruistic sharing in 15-month-old human infants. *PLoS One*, *6*, e23223.

Sheskin, M., Chevallier, C., Lambert, S., & Baumard, N. (2014). Life-history theory explains childhood moral development. *Trends in Cognitive Sciences*, *18*, 613-615.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, *23,* 196-204.

Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: why young children endorse norms of fair sharing but do not follow them. *PLoS One*, *8*, e59510.

Squire, S., & Bryant, P. (2002a). The influence of sharing on children's initial concept of division. *Journal of Experimental Child Psychology*, *81*, 1-43.

Squire, S., & Bryant, P. (2002b). From sharing to dividing: young children's understanding of division. *Developmental Science*, *5*, 452-466.

Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. *Child Development*, *81*, 1814-1827.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*, 155-193.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*, B1-B11.

The Pragmatics of Spatial Language

Tomer D. Ullman (tomeru@mit.edu)
Department of Brain and Cognitive Science, MIT

Yang Xu (yang_xu_ch@berkeley.edu)
Department of Linguistics, Cognitive Science Program, UC Berkeley

Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, Stanford University

Abstract

How do people understand the pragmatics of spatial language? We propose a rational-speech act model for spatial reasoning, and apply it to the terms 'in' and 'near'. We examine people's fine-grain spatial reasoning in this domain by having them locate where an event occurred, given an utterance. Our pragmatic listener model provides a quantitative and qualitative fit to people's inferences. **Keywords:** Pragmatics, Implicature, Spatial Language

Introduction

Space is continuous, language is discrete, and spatial language is coarse and limited—built from a restricted and closed class of spatial prepositions (Talmy, 1983, 2000; Landau & Jackendoff, 1993) such as "in," "on," and "near." How can we communicate accurately about spatial relations with an impoverished, discrete spatial vocabulary? A partial solution lies in the pragmatics of spatial language. Pragmatic enrichment allows coarse fixed meanings to gain useful context-specific refinements (Grice, 1975; Horn, 1984). This is especially useful when the conveyed states are much finer-grained than the literal vocabulary. Moreover, the spatial domain provides a useful test of pragmatic theory: there is a great deal of room for enrichment in a fine-grained domain.

To illustrate the potential pragmatics of spatial language, consider the top half of Figure 1: this is a map of a small city with two quarters (represented by the red and blue rectangles) and a plaza that is located inside one of the quarters (represented by the dashed circle). Suppose you were told that "a gold lily grew in the red quarter." Where would you think the flower had grown? Taking "in" at face value (i.e. the literal meaning) would mean uniform uncertainty over the red region. But a pragmatic listener could arrive at a more specific interpretation: The speaker did not say the lily was in the plaza, nor near the plaza, nor on the edge of the red quarter.... From this a listener could infer that the lily was probably in none of these locations, and derive a much more specific guess as to where the lily was. In many ways this is a standard scalar implicature (Horn, 1984), such as "some" implying "not all," but the interpretation space is much richer and the effect of context is easily manipulated—if the plaza were not inside the red quarter, or placed differently inside it, a pragmatic listener's interpretation should change.

Because of the fine-grained space of interpretations, spatial language provides a particularly good domain to explore quantitative models of natural language pragmatics, such as the recently successful Rational Speech Acts (RSA) framework (Frank & Goodman, 2012; N. D. Goodman & Stuhlmüller, 2013). Previous empirical work has discussed the role of pragmatics in spatial locative expressions (Herskovits, 1985, 1987). However, formal work in this domain is scarce and recent computational studies have emphasized production over comprehension (Carstensen, Kon, & Regier, 2014; Golland, Liang, & Klein, 2010). Thus, no studies have looked at the quantitative effects of implicature in the spatial domain. We aim to bridge this gap with a formal RSA-based model, that quantitatively predicts how a listener interprets spatial language in a 2D map domain.

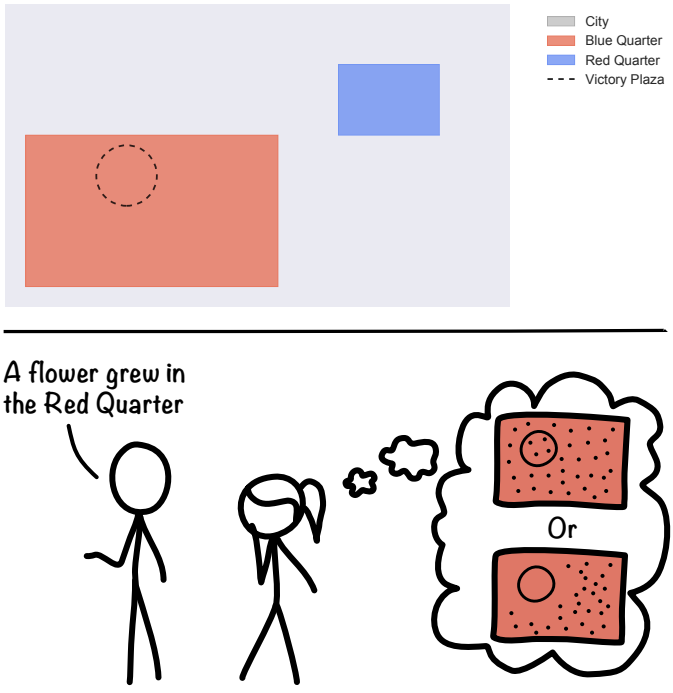


Figure 1: Illustration of a simple spatial reasoning situation. How will the listener interpret the speaker's words?

Modeling spatial implicature

We model spatial language understanding within the Rational Speech Acts (RSA) framework of Frank and Goodman (2012). This framework has previously been applied to domains such as object properties, with terms such as quantifiers (e.g. "some of the apples are red" implying that only a subset of the apples are red). We adapt this framework to spatial language by taking locations as the domain of interest, and

building a lexicon for describing spatial relations.

The RSA framework specifies a pragmatic listener reasoning about the intention of an informative speaker, who in turn reasons about a literal listener; the literal listener updates her beliefs by conditioning on the truth of an utterance. Altogether, writing the location of interest (e.g. of the gold lily) as s and the utterance (e.g."in the red quarter") as u:

$$P_{L_0}(s|u) \propto \delta_{[\![u]\!](s)} P(s) \quad (1)$$

$$P_{S_1}(u|s) \propto \sum_{s' \text{ s.t. } |s-s'|<\varepsilon} P_{L_0}(s'|u)P(u) \quad (2)$$

$$P_{L_1}(s|u) \propto P_{S_1}(u|s)P(s) \quad (3)$$

Beginning with the simplest listener model, L_0, Equation 1 specifies a listener who interprets the utterance via its truth-functional denotation $[\![u]\!]$ (described in more detail below), and simply restricts the prior distribution over locations to those where this denotation is true. The speaker, S_1, in Equation 2 aims for the literal listener to guess the correct location, balanced against the prior probability of the utterance. However, because it is vanishingly unlikely for L_0 to guess exactly the right location, our speaker only cares if the true and gusseded locations are within a small distance ε; this corresponds to an *approximate question under discussion* as used by Kao, Wu, Bergen, and Goodman (2014). Finally, the pragmatic listener, L_1, in Equation 3 updates her beliefs in accord with Bayes' rule, under the assumption that the S_1 speaker would have chosen the utterance she heard.

The denotation of an utterance u is a function from location to Boolean: $[\![u]\!](s) \in \text{Bool}$. We will assume that the spatial language relevant to our scenarios depends on a context set of regions in $\mathbb{R}^2$, as in Figure 1: The City, The Red Quarter, The Blue Quarter, The Victory Plaza.[1] We considered three types of spatial utterances, each of which combines a preposition with a region name:

- "In" utterances (e.g. "In the Red Quarter") are true within the region: $[\![\text{“in R”}]\!](s) := s \in R$.
- "Near edge" utterances (e.g. "Near the edge of the Red Quarter") are true within distance d of a region boundary: $[\![\text{“near edge R”}]\!](s) := \mu(s, \beta(R)) < d$. (Where μ measures distance to a set, and β returns the boundary of a region.)
- "Near" utterances (e.g. "Near the Red Quarter") are true near the edge but not inside: $[\![\text{“near R”}]\!](s) := [\![\text{“near edge R”}]\!](s) \wedge \neg\, [\![\text{“in R”}]\!](s)$.

People no doubt have access to many more spatial utterances, and their combinations. We restrict ourselves to a small set of utterances for simplicity, and to these specific utterances because they were studied in previous research (Herskovits, 1985), and because a separate production task (not reported here) shows these are common spatial terms. We set a uniform prior over the utterances, $P(u)$, and location, $P(s)$.

[1] Both the denotation and Equations 1–3 depend on the context of named regions. We leave this dependence implicit to avoid clutter.

The model as stated leaves open the question of what distance counts as "Near" or "Near edge" in the shared lexicon (what d to use), as well as what counts as "approximately equal" between the true state and the inferred state (what ε to use). We set ε to 20 units (pixels) based on initial exploration. Because the flexibility of "Near" seems to depend on the item and context ("near your coffee mug" is not the same as "near your coffee house", see also Burigo & Coventry, 2010), we allow the literal listener to establish the most useful d for each utterance. We extend Equation 1 to include uncertainty about the tolerances for each "near" utterance:

$$P_{L_0}(s|u) \propto \sum_{\vec{d}} \delta_{[\![u]\!]^{\vec{d}}(s)} P(s)P(\vec{d}), \quad (4)$$

where $\vec{d}$ is a real number, the tolerance. We assume $P(\vec{d})$ is uniform over 10-30 units (pixels) for each utterance[2].

We implement this model in the probabilistic programming language Church (N. Goodman, Mansinghka, Roy, Bonawitz, & Tarlow, 2012). A full implementation can be found at `http://forestdb.org/models/spatialImplicature.html`. We used this implementation to generate model predictions for each spatial configuration ("city map") and utterance used in the experiment below.

Experiment

We examined people's spatial inferences and the predictions of our model for the spatial terms 'in' and 'near', by putting participants in the role of a listener and asking them to guess where an event happened on a map. Participants made their guess in response to the utterance of a speaker with access to the location of the event.

Participants, materials and methods

Participants ($N = 49$, 13 female, median age 29) were recruited through Amazon's Mechanical Turk service.

We created 4 city maps like the one shown in Figure 1, each containing 2 "Quarters" of different size and color, and a circle labeled "Victory Plaza". The location of the Plaza varied between the cities, while the location of the Quarters remained the same. To broadly control for color and position, we created another set of 4 maps by flipping the original maps along the vertical and horizontal axis, and changing the color of the quarters, making 8 maps in total. Participants were randomly assigned to one of the two map groups.

Participants were told that they would see a series of city maps, that a special flower called a 'Gold Lily' can grow anywhere in the city they're shown, and that their task is to find the Gold Lilies.

Participants were further told that for each map a person will tell them where a Gold Lily grew, and that this person can say the Gold Lily grew *in* a location or *near* a location (this narrows people's lexicon to one more directly comparable with the model). This person was said to be reasonable,

[2] For comparison, the maps used are 500x300 units (pixels).

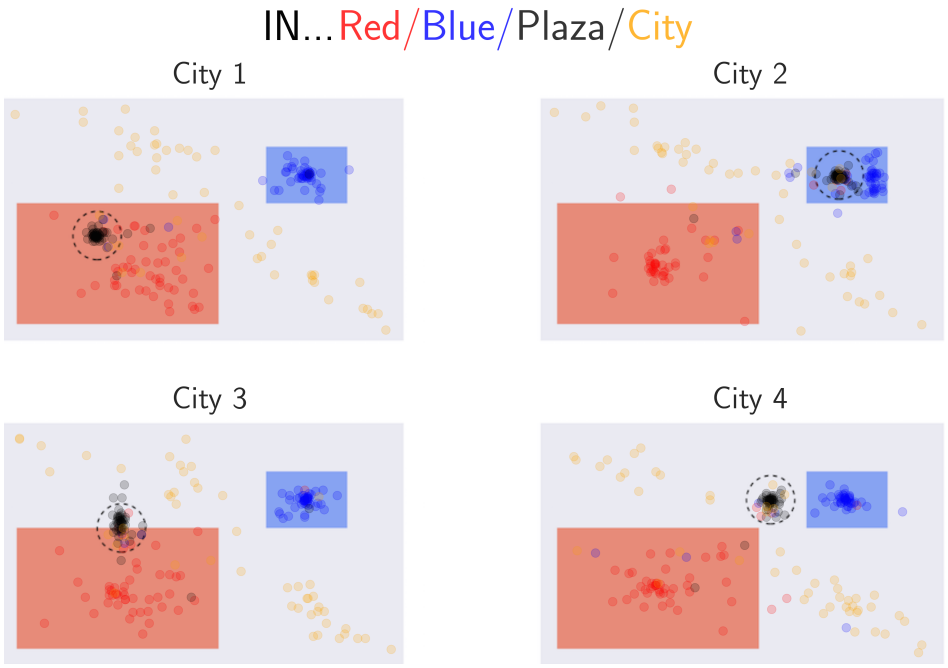


Figure 2: Experiment results pooled across participants for 'In' utterances. Each dot is a participant guess for where a Gold Lily grew, color coded by the region that 'In' refers to (e.g. red dots are different responses to "In the Red Square").

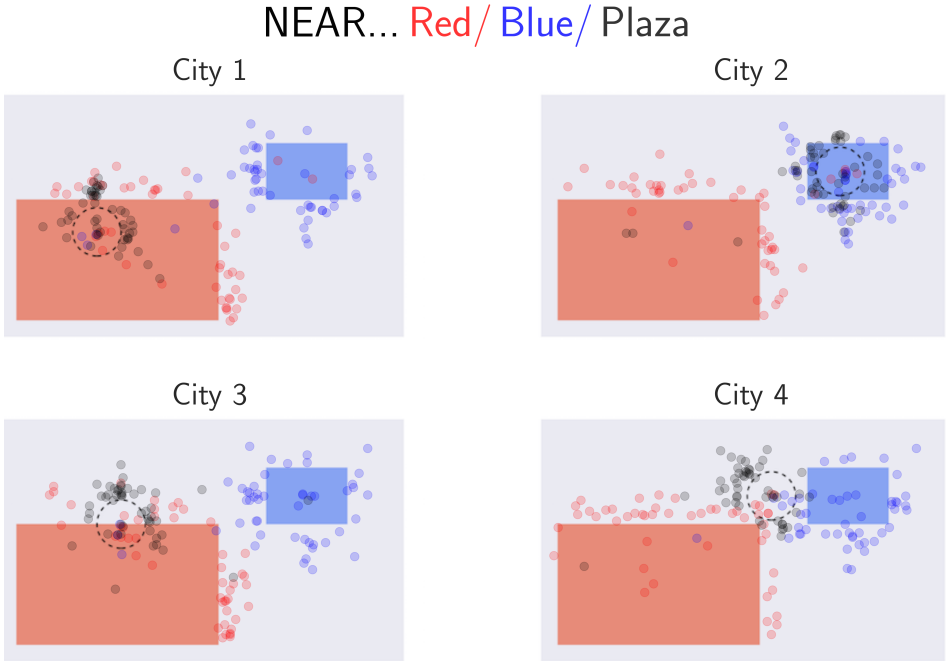


Figure 3: Experiment results pooled across participants for 'Near' utterances, details as in Fig. 2.

and honest. As an example, a participant might read the sentence 'A person tells you: A Gold Lily grew in the Red Quarter'. Participants made their guess by clicking directly on the maps they were shown.

For each of the 4 maps in their group, participants were prompted with a sentence made of $word \times location$ combinations, where $word \in [In,\ Near]$ and $location \in [Red\ Quarter,\ Blue\ Quarter,\ Victory\ Plaza,\ City]$. The combination "Near the City" was not used, as this area was not in the scope of the image and likely to create confusion. In total, each participant was prompted with 28 sentences in random order. Each map had a legend to its right.

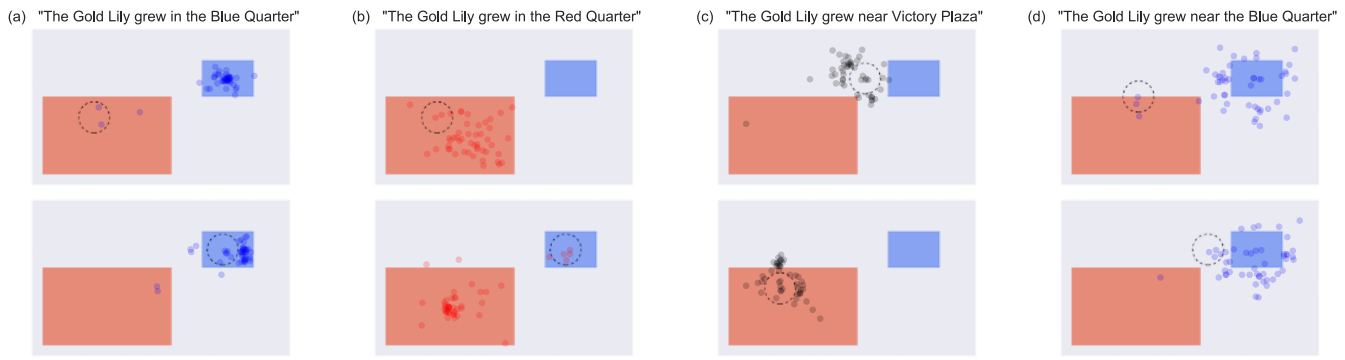


Figure 4: Examples highlighting trends from Figures 2 and 3. The response to the same utterance (column) changes by the different relative configuration of the regions.

Qualitative Results

Figures 2 and 3 show responses for "in" and "near"sentences, collapsing across the rotated-and-inverted-color cities and color-coded by the region referred to. For example, in the top-left of Figure 2, the blue dots correspond to all participant guesses for where the flower grew in City 1, when prompted with "In the Blue Quarter".

To highlight qualitative effects, we focus on the cases highlighted in Figure 4. The reader is encouraged to examine the remaining twenty cases, which show similar effects.

'In X' implies 'In X but not in Y' As shown for example in Figure 4(a): Both the top and bottom of (a) show guesses for lily location when told it grew in the Blue Quarter. When people hear 'In Blue' in City 1 (top) they infer 'In center of Blue'. But in City 2 (bottom) they infer 'In Blue, but not in the Plaza'. Specifically, in City 1 (top) a tic-tac-toe-like division of the Quarter shows that the grid center, accounting for 11% of the area, captures 59% of the guesses. In City 2 (bottom) the same grid center contains just 8% of the guesses, as people shift guesses to the right of the Plaza. Such a pattern of results holds for the other regions as well (Figure 2).

Edge avoidance When there is no direct intersection between regions (except the City, which all regions intersect with), people do not guess uniformly in the region, but rather favor the center. For example, in Figure 4(b), bottom: People told 'In Red' place most of their guesses in the center of the region. A tic-tac-toe grid-analysis shows the red-grid center accounts for 11% of the area but 65% of the guesses. In 10,000 simulations of 49 participants that guess uniformly over the Red Quarter, *none* have the center account for more than ∼30% of the responses. This pattern of results holds for the other regions as well, as shown in Figure 2, though it is disrupted when the Plaza is placed inside the region as in the top panel of Figure 4(b).

'Near' is non-uniform on edges As shown in Figure 4(c) and (d), when told 'Near Plaza' or 'Near Blue' people do not seem to be guessing uniformly, rather their guesses depend on the context of other regions. In the top panel of (c) people place their guesses for 'Near Plaza' to the top-left of the Plaza, while in the bottom panel of (c) people place their guesses for 'Near Plaza' to the top and right of the plaza. In the top panel of (d) the guesses for 'Near Blue' are to the left and bottom of the Blue Quarter, while in bottom of (d) they shift away from the left of the Blue Quarter (now partly occupied by the Plaza). Figure 5 shows the percentage of participant guesses by quadrant, indicating the by region under consideration. A χ^2 test on binned responses near each region (as in Figure 5) shows 11 of the 12 response-patterns are significantly different from uniform ($p < 0.05$), with the Blue Quarter in City 3 being the only exception.

This qualitative pattern of results make it clear that context affects participants' interpretations of spatial language, as expected under a pragmatic account. In the next section, we turn to a quantitative analysis, comparing results with the different listener models described above.

Comparison to Model(s)

To quantitatively compare people to the literal and pragmatic listener models (L_0 and L_1), we converted participants' responses for each condition into smooth two-dimensional dis-

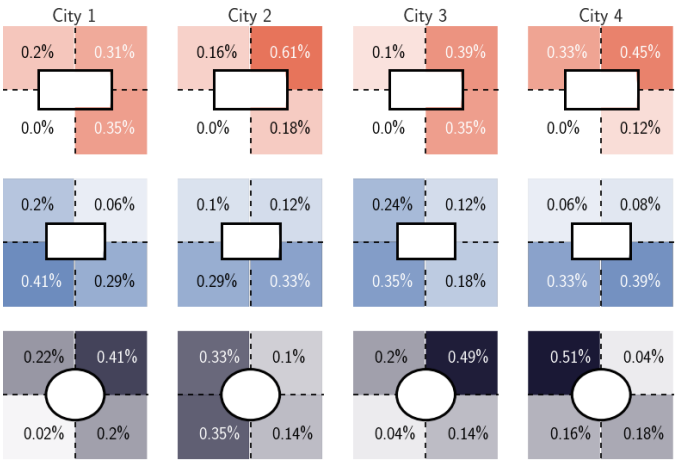


Figure 5: Distribution of participant 'Near' guesses, by region (color-coded) and city.

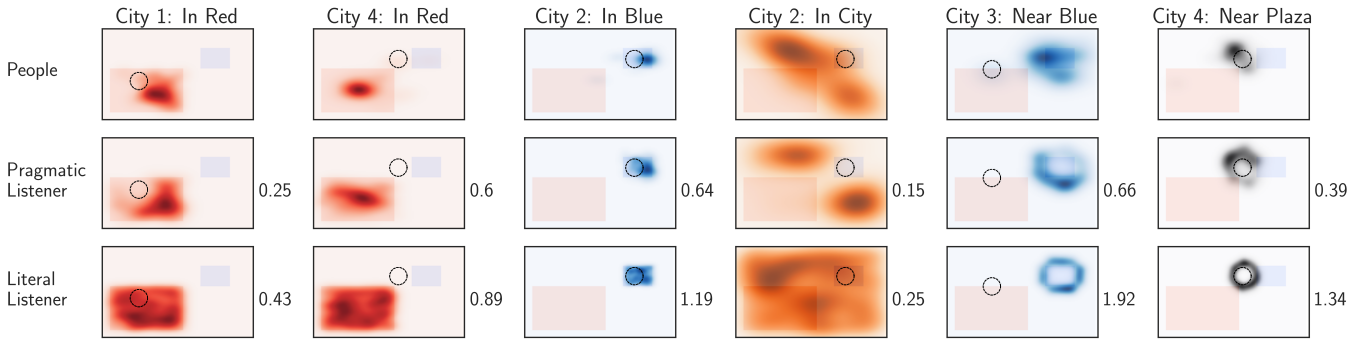


Figure 6: Example comparisons between people and the two listener models. Colored patches indicate probability distributions inferred from people's responses and model samples. Numbers to the right of each listener subplot indicate the *KL* distance between the distributions of people and the model samples, lower numbers indicate a better match with people.

tributions. We used non-parametric, multivariate kernel density estimation to infer these distributions (Epanechnikov, 1969). The same estimation was applied to samples drawn from the probabilistic programs that represent the L_0 and L_1 models. The model distributions are 'sharpened' by raising them to a power α and renormalizing. Such a sharpening parameter controls how much of the model's distribution is smeared or centered around the maximum likelihood, and is functionally similar to an 'optimality' parameter used in other implicature studies (N. D. Goodman & Stuhlmüller, 2013; Kao et al., 2014). We set $\alpha = 1.5$, although the overall findings are robust within an explored range $\alpha \in [1,2]$.

Example comparisons between the models and people are shown in Figure 6. When hearing 'In Red' in City 1 (left-most column) The literal listener L_0 places a uniform posterior distribution on the Red Quarter, while L_1 and people avoid Victory Plaza and lean to the right of the Red Quarter. This correspondence can be measured in terms of the Kullback-Leibler (*KL*) distance between the distributions. In this particular example, $KL(\text{people}|L_0) = 0.43$, while $KL(\text{people}|L_1) = 0.25$.

As Figure 6 shows, the L_1 model captures many of the interesting spatial-implicature interpretations shown by people. Indeed, the L_1 listener shows the same general qualitative patterns discussed in the previous section ("in X" implies "in X but not Y", edge avoidance, non-uniformity on edges). A full comparison figure exceeds the space limits of this paper, but can be found at `http://www.mit.edu/~tomeru/spatialPragmatics/allComparisons.pdf`. Figure 7 shows the *KL* distances for both models across all questions and cities. The pragmatic listener is closer to people than the literal listener for 26 of the 28 (93%) comparisons.

While the pragmatic listener is able to account for the pattern of results found with people, it is interesting to highlight some particular cases where it does not do so well. The left column of Figure 8 shows one of the two cases where the *KL* distance between people and the literal listener is shorter than between people and the pragmatic listener. While the implicature pattern of the pragmatic listener seems generally correct – it places its probability in the Red Quarter while avoiding the Plaza – the model is over-avoiding the area near the plaza relative to people. The right column of Figure 8 shows an example where even though the pragmatic listener is closer to people than the literal listener, it seems generally mistaken in that it places much of its probability distribution within the Red Quarter, which people avoid.

Overall, the results suggest that the pragmatic listener model can account for the quantitative and qualitative pattern of pragmatic spatial inferences of people, within our domain.

Discussion

Space is fine grained, more so than language, leaving a great deal of room for the influence of context and pragmatic enrichment. As a result, spatial language is a promising avenue for testing quantitative models of pragmatic reasoning. We investigated people's fine-grained pragmatic reasoning in a spatial domain, and found support for a model within the rational speech act (RSA) framework.

The model's lexicon was more restricted than natural spatial language. Even within our domain, utterances might include 'in the middle of the City', 'in between the Red and Blue Quarters', 'above Victory Plaza', 'to the left of the Blue Quarter, near the Plaza'.... In future work, a parallel production task (in which people convey a map location using language) should be used to inform and expand the lexicon.

There are many other extensions of the paradigm. For instance, when reasoning pragmatically, people take into account the knowledge of their partner (N. D. Goodman & Stuhlmüller, 2013). In a spatial context, a speaker might adjust the landmarks and utterances they use depending on the listener: 'next to Jacques' café' for those in the know vs. 'next to the Louvre' for out-of-towners. The listener could correspondingly adjust their inferences depending on what they believe the speaker knows.

Within the model, an important directions for exploration is the treatment of *pragmatic slack* in phrases like "near the Plaza." We modeled the tolerance of the lexical items as uncertain, letting the literal listener infer an appropriate value. This literal-slack model allows the slack to adjust in order to increase the odds of a speaker's utterance being true; it neglects other potentially important pressures, such as making the speaker's utterance *informative*. One option is *lifting* the

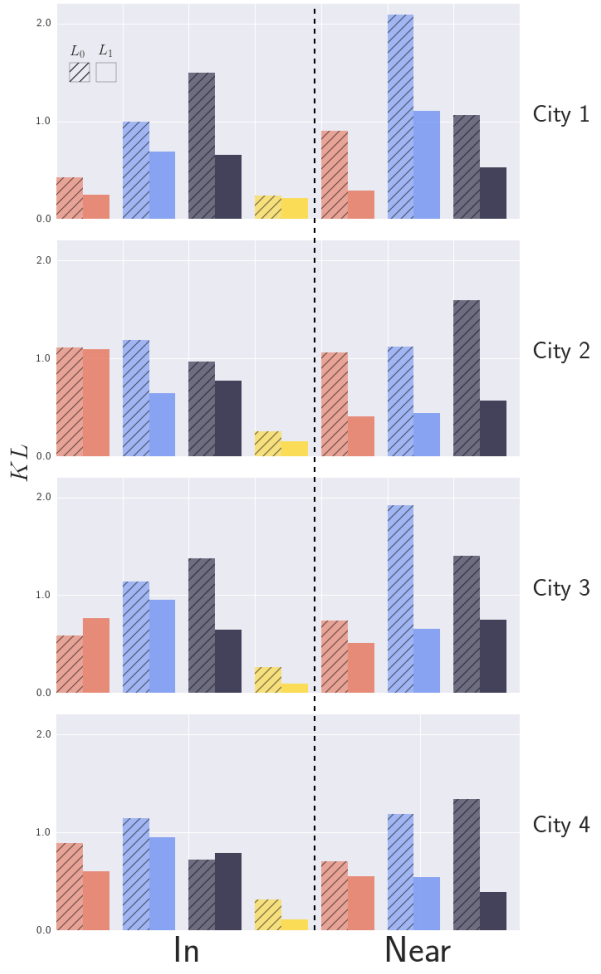


Figure 7: KL distances between people and the listener models, by question and city. Hatched bars are literal listener L_0 and plain bars are pragmatic listener L_1. Bar color indicates region. Lower bars indicate greater agreement with people.

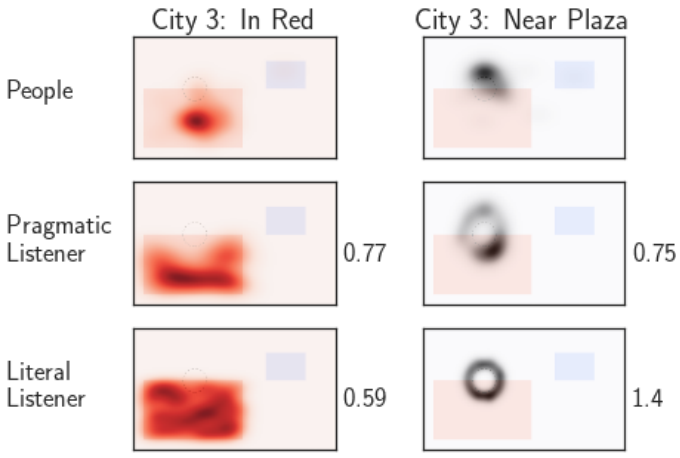


Figure 8: Mismatches between pragmatic listener and people, with literal listener for comparison. Details similar to Fig. 6.

slack variable into the pragmatic listener (as in N. D. Goodman & Lassiter, 2014). While this model is simple to construct, future work is needed to find cases which distinguish it from the model used here, possibly in a communicative paradigm with participant as both speaker and listener.

Acknowledgments This work was supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and the Center for Brains, Minds and Machines (CBMM), NSF STC award CCF-1231216.

References

Burigo, M., & Coventry, K. (2010). Context affects scale selection for proximity terms. *Spatial Cognition & Computation*, *10*(4), 292–312.

Carstensen, A., Kon, E., & Regier, T. (2014). Testing a rational account of pragmatic reasoning: The case of spatial language. , 2009–2013.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 410–419).

Goodman, N., Mansinghka, V., Roy, D., Bonawitz, K., & Tarlow, D. (2012). Church: a language for generative models. *arXiv preprint arXiv:1206.3255*.

Goodman, N. D., & Lassiter, D. (2014). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *Handbook of Contemporary Semantic Theory. Wiley-Blackwell*.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, vol.3* (pp. 41–58). Academic Press.

Herskovits, A. (1985). Semantics and pragmatics of locative expressions*. *Cognitive Science*, *9*(3), 341–378.

Herskovits, A. (1987). *Language and spatial cognition*. Cambridge University Press.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications*, 11–42.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences*, *16*(02), 255–265.

Talmy, L. (1983). *How language structures space*. Springer.

Talmy, L. (2000). *Towards a cognitive semantics, Volume 1: Concept Structuring Systems*. MIT Press.

Measuring Interest in Science: The Science Curiosity Scale

Asheley R. Landrum[1] (ALandrum@asc.upenn.edu)

Joseph Hilgard[1] (JHilgard@asc.upenn.edu)

Heather Akin[1] (HAkin@asc.upenn.edu)

Nan Li[1] (NLi@asc.upenn.edu)

Dan M. Kahan[1,2] (Dan.Kahan@yale.edu)

[1]Annenberg Public Policy Center
University of Pennsylvania
202 S. 36th Street, Philadelphia, PA 19104

[2]Yale Law School
Yale University
127 Wall Street, New Haven, CT 06511

Abstract

In the current study, we present the methods for creating and validating a science curiosity scale. We find that the scale presented here is unidimensional and highly reliable. Moreover, it predicts engagement with a science documentary clip more accurately than do measures of science intelligence or education. Although more steps are needed, this provides initial evidence for the utility of our measure of science curiosity.

Keywords: curiosity; science curiosity; scale; psychometrics; Item Response Theory

Introduction

Do people differ in their desire to seek out and consume science information for personal satisfaction? Determining the answer to such a question requires having a genuine measurement of science curiosity. Although many scales purport to measure such a construct, performance assessments of these scales reveal that most (if not all) of such attempts are psychometrically weak and often not genuinely predictive of what they are supposed to be assessing (e.g., Blalock et al., 2008; Osborne, Simon, & Collins, 2003). The aim of the current study was to take the first steps toward developing an original, valid, and reliable science curiosity scale.

One problem with existing scales that purport to measure science interest or science curiosity is heavy reliance on self-reported measures. Although asking people directly is often a good way of gaining information, there are potential problems. For instance, asking participants to what extent they agree with statements such as "I'm curious about the world in which we live," and "I find it boring to hear about new ideas," (Fraser, 1978) is likely to provoke socially desirable responding.

Public opinion research, for example, has used numerous "science attitudes" batteries that purport to measure science interest by literally asking people if they "like" science. The National Science Foundation Indicators (2014), for instance, feature an array of "public attitudes toward science" items. These items consistently find that members of the public hold overwhelmingly pro-science attitudes (e.g., 4 out of 5 Americans say they are interested in new scientific discoveries, National Science Board, 2014). As these items are subject to the same problems related to self-report measures generally, it is unclear whether the American public actually *does* hold overwhelmingly pro-science attitudes or if this positivity is a direct result of socially-desirable responding. Therefore, when using self-report measures, it is important to determine whether the scales are measuring the underlying disposition or trait of interest and not simply the motivation for others to perceive them as possessing it.

Current Study

Our strategy for conducting a valid science curiosity scale was to combine a number of self-report measures with behavioral and performance ones. First, we used the behavioral and performance items to validate the self-report items—that is, to confirm that the variance in the self-report items could be treated as originating in difference in science curiosity rather than because of some other reason. Then, we combined the self-report items with the behavior and performance items to form a scale that would reliably discriminate among study participants with varying levels of science curiosity.

To counteract the problem of socially-desirable responding, we disguised our objectives by presenting our scale as a marketing survey. That is, we embedded individual self-report items relating to science interest in modules consisting of multiple items reflecting an array of interests (e.g., sports, entertainment, business, and politics). Thus, there was no particular reason for participants to suspect that we were specifically interested in capturing their motivation to learn about science, and we could avoid inadvertently encouraging participants to express pro-science sentiments or to engage in the form of overstatement that pervades many self-report scales.

In addition to the self-report interest and behavior items, we included an objective performance item. Near the middle of the survey, participants were told that we wanted to get their reactions to a news story "of interest to them." In order to ensure that participants were presented with a news story that matched their interests, they were provided with a discrete list of news story sets (See Figure 1) and asked to choose the set that would be of most interest to them. One set consisted of science stories, while the others consisted of popular entertainment, sports, and financial news. Given that reading an article and answering questions is more

cognitively taxing than simply answering questions about one's self, we perceived that a participant's purposeful selection of a science story over the others as one valid indicator of genuine science interest.

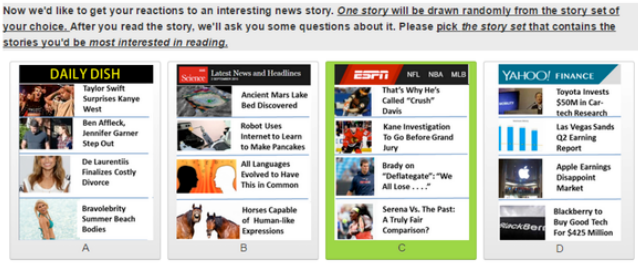


Figure 1. News Story-Set Selection Task. Subjects were instructed to select one set of the four, from which a story would be selected "at random" for them to read and answer two questions on. The task was conceived of as a performance-based measure of interest in science.

With the collected data, we used item response theory (IRT) to combine the items into a composite scale. IRT, unlike simpler alternatives to scale creation and validation, does not assume each item in a questionnaire is equivalent to another item. Rather, it is expected that a pro-curiosity response to some items may indicate higher levels of curiosity than pro-curiosity responses on other items (also known as differences in item "difficulty"). These differences are taken into account by IRT when calculating the scores on the latent variable (Embretson & Reise, 2000).

Moreover, IRT allows one to examine how informative the proposed scale is for each level of the latent variable. While some scales may have high inter-item reliability (e.g., Cronbach's alpha), they may be informative for only a portion of the scores. For example, the cognitive reflection test (i.e., CRT, Frederick, 2005) is very discriminating among people who score above the mean, but provides little to no information among those who score below (Kahan, 2014).

After determining reliability, we were able to behaviorally validate the scale by determining to what extent it predicted engagement with a clip from a science documentary. Indeed, these data are part of a larger study that examines how science curiosity measures might be useful tools for science filmmakers to better target their documentaries to more diverse audiences.

Method

Participants

Participants were 2,267 adults (54.3% female) who are members of a nationally-representative panel of participants from YouGov[1]. Of the participants, 76% reported being white, 9.3% reported being black or African American, 7.8% reported being Hispanic, 2.1% reported being Asian, 0.6% reported being Native American, 2.6% reported being of mixed race, and 1.4% reported being of another race. The participants ranged in age from 18 to 90 years old (Mean = 46.7, Median = 48, SD = 16.23).

Procedure

As previously stated, participants completed the science curiosity items as part of a larger experimental survey examining science curiosity and engagement with science documentary films. Participants completed the survey over the internet, either on computers or mobile devices.

The first part of the survey included 7 modules, in which our science curiosity items (items related to scientific research or discoveries and new technologies) were embedded with items related to other issues, such as crime, education, government or politics, sports, religion, international affairs, business or finance.

The first module was *News Interest*. In this module, participants were asked to rate how closely they follow the news related to each topic: not at all (1), a little, but not closely (2), closely but not very closely (3) or very closely (4).

The second module was *Leisure Activity*. In this module, participants were asked to indicate how many times in the past year they had engaged in several activities including visiting a science or technology museum, attending a live sporting event, visiting an art museum, attending a musical performance or concert, going to a zoo or aquarium, going to a public library, going to a gun show, visiting a theme park or amusement park, attending a political rally or political event, attending a public lecture (on history, science and technology, public affairs or politics, religion, economics, or other).

The third module was *Books*. In this module, participants indicated whether they had read a book in the past year on each of several topics. Topics included crime, science fiction, mystery, education, government or politics, sports, religion (other than Holy Scripture text), international affairs, business or finance, scientific research or discoveries, history.

The fourth module was *Conversation*. In this module, participants indicated what types of topics they discuss with their friends, family members or co-workers. For each topic, participants were asked to say whether they discussed it never, rarely, more than rarely but not often, or often. Topics included crime, education, government or politics, sports, religion, international affairs, business or finance, scientific research or discoveries, new technology, entertainment or celebrities.

The fifth module was *Social Media*. In this module, participants indicated whether (and if so, how often) they share news stories on social media. Participants who said that they did so were asked to rank the topics in order of how likely they were to share. Topics included crime, education, government or politics, sports, religion, international affairs, business or finance, scientific research

[1] For information about YouGov's panel of participants see: https://today.yougov.com/about/about-the-yougov-panel/

or discoveries, new technologies, entertainment or celebrities.

The sixth module was the *Reading Selection Task*. As previously stated, participants were told that we wanted to get their reactions to an interesting news story drawn from a story set of his or her choice. Thus, participants were asked to pick the story set that contained the stories that they would be most interested in reading from entertainment, science, sports, and business or finance (see Figure 1). Following their selection, participants were shown one story from the set and were asked to read it and answer two factual questions about that story.

The seventh module was *Self-Reported Interests*. In this module, participants were told that they would see several topics and for each topic they would be asked to indicate how interested they were in that topic: not at all interested (1), slightly interested (2), more than slightly—but not very—interested (3), or very interested (4). Topics included government or politics, sports, religion, foreign travel, scientific research or discoveries, new technologies, entertainment or celebrities, nature, and music.

Following these modules, participants watched a clip from a science documentary about the evolution of color vision and were asked questions about their interest in the clip, factual questions about the clip, and agreement questions (e.g., whether they thought the documentary supplied convincing evidence of how color vision came about). Moreover, several behavioral variables were collected such as how long participants watched the clip before turning it off. In addition, participants answered a battery of items related to their beliefs about policies and risks, and their cultural worldviews (Kahan, Braman, Slovic, Gastil, & Cohen, 2008) and a questionnaire measuring ordinary science intelligence (e.g., Kahan, in press).

Results

Self-Report Science Interest

First, because we aimed to use the Reading Selection Task and the behavior items as validators of the self-report items (News Interest, Conversation, and Self-Reported Interests modules), we started by forming a scale that aggregated these self-report interest items. The resulting self-report science interest (SRSI) scale displayed a high degree of measurement precision ($\alpha = 0.85$). More importantly, the scale's properties suggest valid measurement of science curiosity.

First, SRSI was positively correlated with the subjects' science comprehension as measured by the Ordinary Science Intelligence assessment (i.e., Kahan, in press; Kahan et al., 2012), $r = .26$, $p < .001$, and education, $r = .21$, $p < .001$. See Figure 2. Although science curiosity and science comprehension are not the same constructs, one would suppose that people who are proficient in science comprehension would also be more likely to like science and that those who were genuinely interested in science would at least be modestly more proficient in comprehending it.

Second, SSRI predicted variance in the responses to the behavioral measures. For instance, it predicted which subjects would select the science set in the Reading Selection Task, $X^2(1) = 181.17$, $p < .001$, which subjects attended a science lecture in the last year $X^2(1) = 230.86$, $p < .001$, and which subjects had read a book about scientific research and discoveries $X^2(1) = 1197.56$, $p <. 001$.

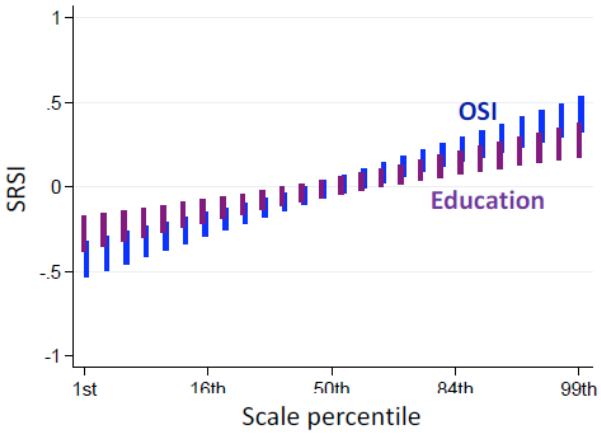


Figure 2. Relationship of Self-Report Science Interest (SRSI) to science comprehension (OSI) and education. Results based on multivariate linear regression (OSI: b=0.21, p<0.01; Education: b=0.14, p<0.01; R^2=0.08). Colored bars are 95% confidence intervals. Scales are normalized with the mean equal to zero and units expressed in standard deviations.

Using Item Response Theory to Create a Science Curiosity Scale

The power of individual items to contribute to measurement precision at different levels of a latent variable can be incorporated into a scale using Item Response Theory modeling (Embretson & Reise, 2000). We used IRT to form a composite scale that combines responses to the self-report interest items (SRSI), the self-report behavioral items (book item and public lecture item), and the Story Selection Task item. The ranking scores for the science news stories in the social media module were also included for participants who indicated that they did indeed share material via social media platforms (social media module).

The resulting scale—the "Science Curiosity Scale" (SCS)—displayed desirable psychometric properties. It was unidimensional (Figure 3, image C), supporting the inference that it measured a single, unitary latent disposition. Moreover, the scale reflected a high level of reliability—at or above $\alpha = 0.80$— across the entire range of latent science interest disposition (Figure 3, image D).

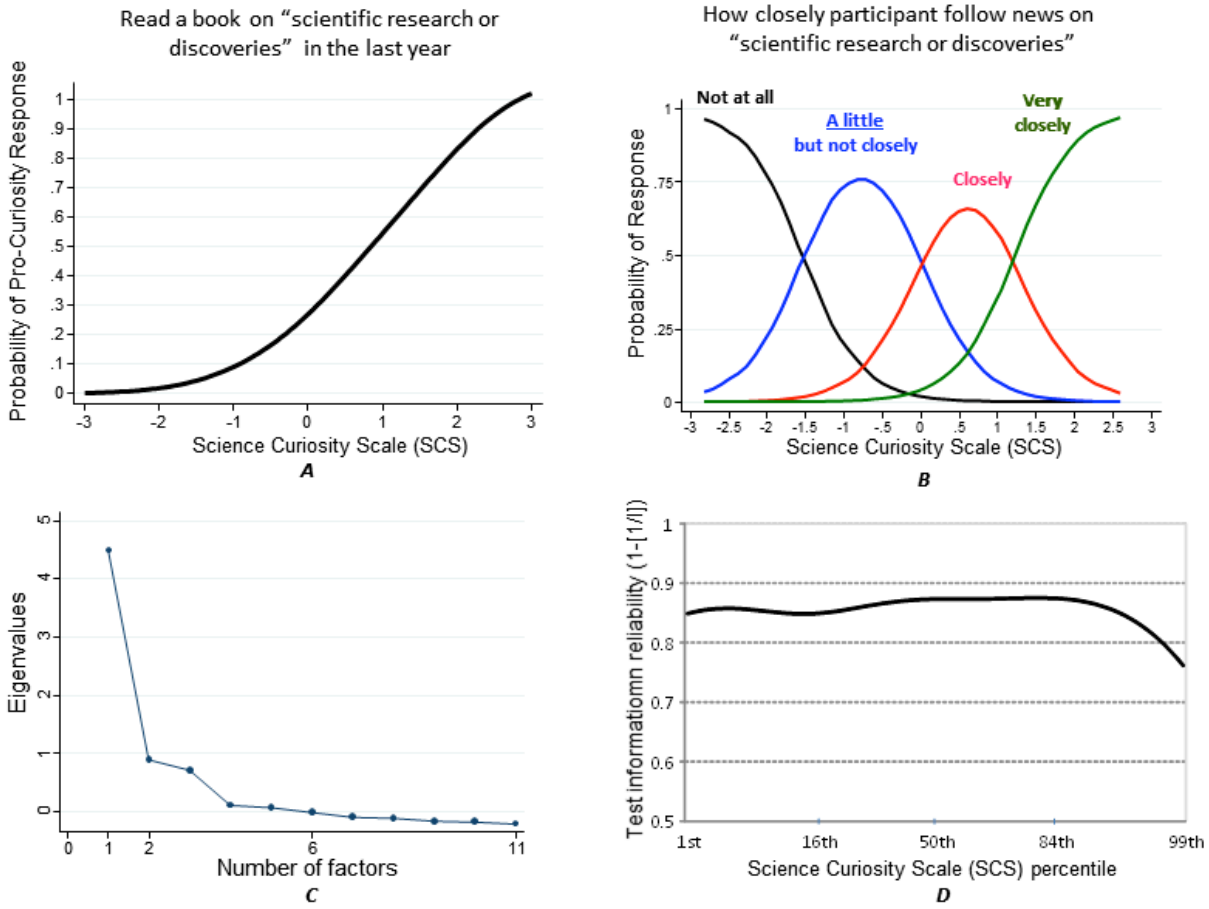


Figure 3. The Science Curiosity Scale (SCS) based on a 2pl Item Response Theory Model. SCS scores are standardized with the mean centered at 0 and units measured in standard deviations. Images A and B reflect representative "item response profiles": the relative probability of the indicated response conditional on a specified level of the latent science curiosity disposition, which is used to estimate subjects' SCS scores. Image C reflects the unidimensionality of the scale. Image D illustrates the measurement precision (test information reliability, similar to Cronbach's alpha) at various levels of science curiosity.

Science Curiosity Predicts Engagement with a Science Documentary Clip

In our study, we aimed to externally validate SCS with engagement with a documentary clip focused on the evolution of color vision. To measure engagement with the video clip we had a combination of self-report measures and behavioral measures that we combined in a manner similar to SCS using IRT. One item asked participants how interesting they found the documentary (M = 3.22, SD = 1.69, Range = 0 to 5). Another item was part of an experimental condition in which we provided half of the sample the ability to turn off the clip whenever they felt they watched enough. This allowed us to measure the number of minutes of the clip watched (out of 10 minutes; M = 6.42 minutes, SD = 4.12). We hypothesized that people who were more science curious would watch the clip for a longer period of time than people who were less science curious. We also offered the sample the option of requesting the full episode of the documentary. If participants were interested in watching the full episode (and selected "Yes", 51%), we would email them a link to the full episode (no payment required). We had hypothesized that participants who were more science curious would be more likely to request the full episode of the documentary. Moreover, we combined these items using IRT to create an index of engagement with the science documentary clip using IRT. See Figure 4.

Indeed, subjects' SCS scores were a strong predictor of their level of engagement with the documentary clip. The practical significance of the predictive power of SCS can be gauged by examining its relationship to various components of the engagement index. Subjects who scored one standard deviation (84th percentile) or above on the SCS were disproportionately likely to rate the clip as "very interesting" and to watch the entire clip. Subjects who scored +1 or higher on SCS were also far more likely than others to request access to the full episode from the documentary.

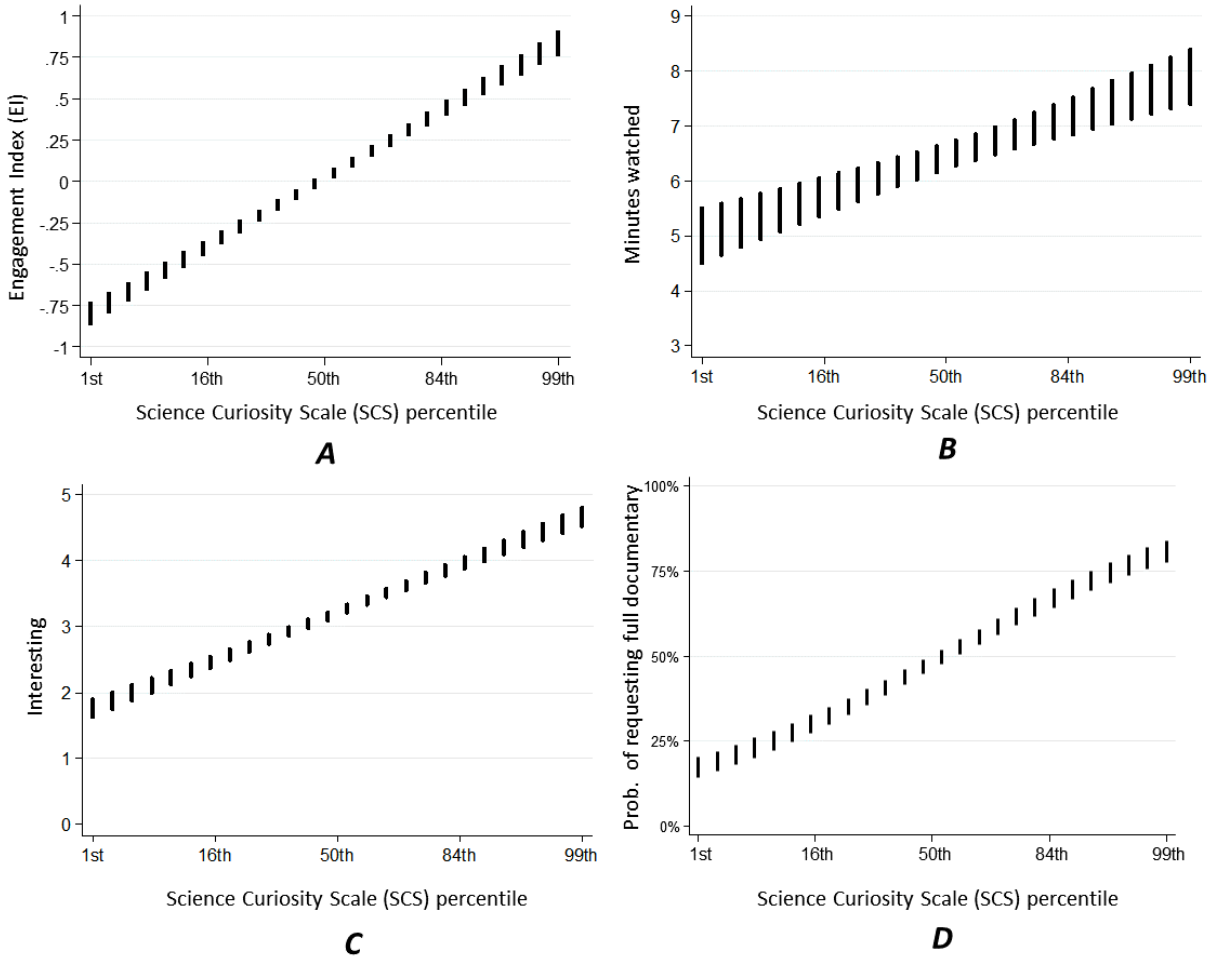


Figure 4. Engagement with the documentary clip as a function of science curiosity. N=2500 for images A, C, & D, and N=1250 for image B. Images A, B, and C are based on linear regression analyses and Image D is based on logistic regression. Bars represent 95% confidence intervals.

SCS also appeared to be a stronger predictor of subjects' engagement with the clip than did their ordinary science intelligence (OSI) scores. See Figure 5. One would expect science comprehension to predict engagement with a science documentary. However, because taking pleasure in contemplating scientific discovery and the capacity to recognize and make use of scientific evidence are distinct dispositions, one would also expect a valid science interest measure to be more discerning of engagement.

Although it is uncommon for researchers to present evidence behaviorally validating curiosity scales, investigations of such scales typically find that the disposition being measured reduces to reasoning proficiency (Loewenstein, 1994). The power of SCS to predict engagement with the clip independent of, and more powerfully than, Ordinary Science Intelligence is thus a highly desirable property of the Science Curiosity Scale.

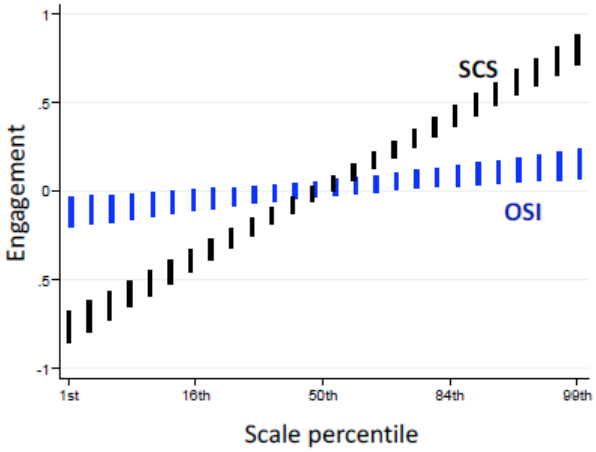


Figure 5. Relative impact of OSI and SCS on engagement with the clip. Results based on multivariate linerar regression (including SCS, OSI, and cross-product interaction predictors. Bars are 0.95 CIs.

Our study design, of course, demanded that a science curiosity measure be developed and validated independently of engagement with the documentary segment itself. Nevertheless, the power of SCS to predict how interesting subjects found the show, how much of the segment they chose to view, and how likely they were to request access to the full episode supplies additional reason to be confident that SCS does indeed measure a general science-interest disposition.

Discussion

The current study demonstrates the feasibility of constructing a valid science curiosity measure that can be used, for example, to evaluate how well science films engage those individuals most interested in contemplating the insights of scientific discovery. However, more work needs to be done, including follow-up studies that aim to continue to validate the Science Curiosity Scale.

One limitation of the scale as it is currently constructed is the amount of items required. In order to avoid some of the problems commonly associated with self-report items, we had to bury indicators of science interest in an entire battery of distractor items. Not all researchers will be able to spare the expense for this measurement. While we anticipate simplifying the measure in future studies, we also will aim to figure out how to put the distractor items to best use—for example, using items that negatively correlate with science curiosity as part of the scale.

All in all, this study demonstrates that by combining appropriately subtle self-report items with behavioral and performance items, it is possible to construct a scale that measures individuals' desire to seek out and consume scientific information for personal satisfaction. Such a measure would likely provide many contributions to the advancement of knowledge. For instance, a science curiosity measure may help improve science education by facilitating investigation of the forms of pedagogy most likely to promote learning (Blalock et al., 2008). In addition, those who study the science of science communication (Fischhoff & Scheufele, 2013; Kahan, 2015) could also use a science curiosity measure to deepen their understanding of how public interest in science shapes the responsiveness of democratically accountable institutions to policy-relevant evidence.

Acknowledgments

The authors would like to thank HHMI and Tangled Bank Studios for their assistance in developing the project as well as for allowing us to use clips from their documentary *Your Inner Fish*. They would also like to thank Kathleen Hall Jamieson and the Annenberg Public Policy Center for their support of the project.

References

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. *International Journal of Science Education, 30*(7), 961-977.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fischhoff, B., & Scheufele, D. A. (2013). The science of science communication. *Proceedings of the National Academy of Sciences, 110*(Supplement_3), 14031. doi:10.1073/pnas.1312080110

Fraser, B. J. (1978). Development of a test of science-related attitudes. *Science Education, 62*(4), 509-515.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25-42.

Kahan, D. M. (2014). We need a crt 2.0! And irt shoudl be used to develop it.

Kahan, D. M. (2015). What is the 'science of science communication'? *Journal of Science Communication, 14*(3), 1-10. doi:10.2139/ssrn.2562025

Kahan, D. M. (in press). 'Ordinary science intelligence': A science comprehension measure for use in the study of risk perception and science communication. *Journal of Risk Research*. doi:10.2139/ssrn.2466715

Kahan, D. M., Braman, D., Slovic, P., Gastil, J., & Cohen, G. L. (2008). Cultural cognition of the risks and benefits of nanotechnology. *Nature Nanotechnology, 4*(2), 87.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. N. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732-735. doi:10.1038/NCLIMATE1547

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116*(1), 75.

National Science Board. (2014). *Science and technology: Public attitudes and understanding*. Arlington, VA: National Sceince Foundation.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education, 25*(9), 1049-1079.

Conflict-based regulation of control in language production

Michael Freund (freund@jhu.edu)
Department of Neurology, Johns Hopkins University,
1629 Thames Street, Suite 350, Baltimore, MD 21231, USA

Barry Gordon (bgordon@jhmi.edu)
Department of Neurology, Department of Cognitive Science, Johns Hopkins University,
1629 Thames Street, Suite 350, Baltimore, MD 21231, USA

Nazbanou Nozari (nozari@jhu.edu)
Department of Neurology, Department of Cognitive Science, Johns Hopkins University,
1629 Thames Street, Suite 350, Baltimore, MD 21231, USA

Abstract

Is language production dynamically regulated by cognitive control? If so, how domain-general is this process? In two experiments, we studied conflict adaptation, or conflict-driven adjustments of control, in two paradigms: Picture-Word Interference (PWI), which induces linguistic conflict, and Prime-Probe (PP), which induces visuospatial conflict. Exp. 1 tested within-task conflict adaptation separately in PWI and PP. Exp. 2 tested cross-task adaptation by alternating the two tasks in a task-switching paradigm. We found reliable within-task conflict adaptation in both PWI and PP, but neither an analysis of individual differences (Exp. 1), nor a direct manipulation of between-task conflict (Exp. 2) revealed cross-task adaptation. We further report a robust 2-back within-task adaptation in Exp. 2 to refute alternative accounts of null cross-task adaptation. These findings support models of dynamic, top-down control in language production that posit at least some degree of domain-specificity.

Keywords: language production; cognitive control; domain-generality; conflict adaptation; picture-word interference

Introduction

Much research has shed light on the nature and levels of representation in language production, but less has explored how language production is controlled. A recent theory posits that language production is monitored via mechanisms similar to those that monitor other cognitive operations (Nozari, Dell, and Schwartz, 2011). Electrophysiological and neuroimaging studies support this claim by showing a similar ERP negativity (the Error-related Negativity, ERN) as well as common cortical regions involved in monitoring of linguistic and non-linguistic tasks (e.g. Gauvin, De Baene, Brass, & Hartsuiker, 2016; Piai, Roelofs, Acheson, & Takashima, 2013; Riès, et al., 2011). More specifically, this account proposes that the amount of conflict generated between target and competing representations in the production system signals the need for increased control (which, when following an error, often manifests as error detection), and the subsequently-recruited control helps resolve this conflict (see also Yeung, Botvinick, & Cohen, 2004).

While Nozari et al.'s (2011) conflict-based account is domain-general in the sense that it proposes conflict as a signal that is monitored in both linguistic and non-linguistic systems, it does not commit to a unitary neural or cognitive system that monitors for conflict regardless of domain. In fact, the authors provided evidence from computational modeling and individuals with brain damage, showing that the consequences of conflict detection (e.g. detecting errors) are specific to the source of conflict: the amount of conflict between lexical representations (e.g. cat and dog) only predicted the ability to detect semantic errors, while the amount of conflict between phonological representations (e.g. /k/ and /d/) only predicted detection of phonological errors. Importantly, increased conflict at the lexical level did not lead to better detection of phonological errors and vice versa. This specificity arises because each layer of the production system generates conflict independently of other layers and presumably of other cognitive systems, and it is the internal dynamics of these conflict generators that determine the strength of the conflict signal. Thus, the model poses a domain-specific component to the monitoring process. This notion of domain-specificity has been supported by ERP studies showing that while detection of both linguistic and non-linguistic errors lead to ERNs, their magnitudes are not correlated between the two error types at the level of individuals (e.g. Acheson & Hagoort, 2014).

Ultimately, detection of conflict serves a purpose beyond signaling for errors—it helps regulate cognitive control to resolve future conflict and optimize task performance. This regulatory loop has been tested using the "conflict adaptation paradigm" (e.g. Weissman, Egner, Hawks, & Link, 2015). In a typical conflict adaptation paradigm, performance on the current trial is evaluated as a function of the level of conflict on the previous trial. The canonical finding is an interaction between the amount of conflict on the current and previous trials, such that performance is better on current incongruent (high conflict) trials if they are preceded by an incongruent compared to a congruent trial. For current congruent (low conflict) trials, performance is better when preceded by another low conflict rather than high conflict trial. The effect is explained in terms of online regulation of cognitive control: when facing an incongruent trial, the control system increases

engagement by biasing processing away from the distractor. Thus, the next incongruent trial will be responded to more easily. This biasing works against congruent trials, where distractors aid performance.

A conflict adaptation paradigm is ideal for testing whether a cognitive system such as language production is subject to online regulatory control via top-down processes. This was tested as the first question of the current study. More importantly, a task-switching version of the conflict adaptation paradigm has been used to test whether increased conflict in one domain helps with recruitment of control that resolves conflict in another domain (e.g. Egner, 2008; Kan et al., 2013). For example, Kan et al. (2013) interleaved trials from the color-word Stroop task with those of a task in which participants passively viewed a Necker cube (a perceptually bistable figure), which can induce visuospatial conflict. These authors showed that a high-conflict Necker cube trial improved performance on subsequent incongruent button-press Stroop, and interpreted this finding as evidence for a domain-general control system that encompasses verbal and visuospatial domains. Other studies, however, have found no evidence of cross-domain adaptation (Egner, Delano, & Hirsch, 2007; Forster & Cho, 2014; Wühr, Duthoo, & Notebaert, 2015).

Three methodological issues make reconciliation of these contradicting findings difficult. First, some of the studies employed factorial combinations of the two tasks (e.g. Stroop and Simon), where each stimulus is simultaneously congruent or incongruent with respect to each task. Participants always perform Stroop, but the position of the button could be congruent or incongruent with the location of stimulus presentation (Simon). While this design avoids a switch cost, it potentially dilutes the effect of conflict, as an incongruent Simon is paired with a congruent Stroop in a single trial. This dilution could lead to a weaker recruitment of control and thus a weaker adaptation effect (Kan et al., 2013). Second, some of these studies did not control for low-level learning and memory confounds that can obscure the adaptation effect (see Schmidt, 2013 for a review). Finally, all but one of these studies used non-verbal Stroop tasks with arbitrary response mappings, which poses additional demands on working memory. The one study that did use verbal Stroop (Wühr et al., 2015) was designed for a different purpose and had an imbalanced design (i.e. a ratio of 8:2 of task A to task B) that is known to obscure the adaptation effect (Freitas & Clark, 2015). Our design addresses these issues.

The current study was designed to answer two questions: (1) is language production subject to dynamic, top-down regulation? (2) Does conflict in language production serve to regulate performance in a non-linguistic task and vice versa? Conflict adaptation in language production was tested in the well-established Picture-Word Interference (PWI) paradigm (e.g. Schriefers, Meyer, & Levelt, 1990), which avoids arbitrary stimulus-response associations and allows for much more stimulus variability than Stroop. Conflict adaptation in a non-linguistic domain was tested using the visuospatial Prime-Probe task (PP), which uses an optimal design to avoid low-level learning and memory confounds and employs meaningful response mappings (see Weissman et al., 2015). Exp. 1 tested whether each task in isolation showed evidence of adaptation. Finding adaptation in PWI would support models that claim language production is regulated online via top-down control.

The second question was tested in two ways. In Exp. 1, we conducted an analysis of individual differences that investigated the correlation between the size of conflict adaptation in PWI and PP tasks. A positive and reliable correlation would support a domain-general control process. This test, however, is subject to limitations of a correlational analysis (e.g. Redick et al., 2013). Thus, in Exp. 2 we directly tested whether conflict in one task regulated control in the other task, by using a task-switching adaptation paradigm. We alternated PWI and PP trials and assessed the response to conflict in one task as a function of conflict in the other. If PWI shows conflict adaptation as a function of PP conflict (and vice versa), we can conclude a domain-general process of control regulation. The absence of cross-task adaptation would support some level of domain-specificity in the control system.

Finally, Exp. 2 was designed to allow us to test conflict adaptation as a function of conflict on two trials prior (2-back; e.g. Forster & Cho, 2014). Given the alternating nature of the design, the 2-back conflict adaptation provides a second test of within-task conflict adaptation. Importantly, this test also ensured that any potential null effects in the cross-task adaptation were not due to other confounds. Collectively, the results of the two experiments inform us about whether language production is subject to dynamic regulation and whether a fully domain-general model or a partially domain-specific model (e.g. Nozari et al., 2011) is better supported by the evidence.

Experiment 1

Methods

Participants. Forty-eight native English speakers recruited from the Johns Hopkins University community (32 women; mean age = 21.2 years) participated for payment. All participants gave informed consent under a protocol approved by the Institutional Review Board of Johns Hopkins Neurology.

Materials. For the PWI task, a list of 120 target-distractor word pairs was compiled to form the incongruent PWI stimuli. Target and distractor words were semantically related (e.g. target = *bus;* distractor = *car*; as suggested by norms on Mechanical Turk), were matched in length and frequency, and had minimal phonological overlap. Then, 120 300x300 pixel black-and-white line drawings corresponding to each target word were selected from Google images. The word (the distractor for incongruent and the target for congruent stimuli) was overlaid in the center of each image

in black uppercase 36 point Helvetica, to create 120 congruent and 120 incongruent stimuli. Four experimental conditions (*cC, iC, iI, cI;* N = 60 in each) were constructed, where lowercase "*c*" and "*i*" denote the congruent and incongruent status of the *previous* trial respectively, and capital "*C*" and "*I*", the congruent and incongruent status of the *current* trial. To avoid cumulative semantic interference (e.g. Schnur, 2014) semantically-related pictures were spaced by at least 12 unrelated items. Each stimulus appeared in all four conditions across participants. Final stimuli comprised two blocks of 120 trials, prepared in four different orders to avoid systematic order effects.

For the PP task, the materials were identical to Weissman et al. (2015), and consisted of the outlines (in black, on white background) of large arrows (primes) and small arrows (probes, or targets; 75% smaller than the large arrows) pointing in the four cardinal directions. Similar to PWI, four conditions (*cC, iC, iI, cI;* N = 96 in each) were constructed, with congruent trials having primes and probes pointing in the same directions, and incongruent trials in opposite directions. Final stimuli comprised four blocks of 96 trials, prepared in four different orders to avoid systematic order effects.

Procedures. The experiment was run in E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). Stimuli were displayed at the center of a 15 x 12 inch Dell monitor approximately 25 inches in front of the participants. Response times (RTs) for PWI were registered using an Audio-Technica microphone connected to the E-Prime's SRBOX. Responses were also recorded digitally and transcribed offline for the identification of errors. RTs for the PP were registered using a Dell keyboard.

First, participants silently reviewed a slideshow containing labeled images of all PWI targets in the experiment. Next, they completed a 10-item practice block of the PWI paradigm, followed by the two experimental blocks in counterbalanced order. Each trial began with a centrally presented fixation for 900 ms followed by a 50 ms blank screen. Stimuli were then presented for 3000 ms or until a response was registered. The next trial started after a 50 ms blank screen.

The PP task was conducted after PWI. Following a 48-item practice block, participants completed four blocks of 96 items in counterbalanced order. PP trials consisted of the presentation of a prime of either horizontal or vertical orientation, followed by a probe oriented along the same dimension. Participants indicated the direction of the target by pressing one of the four arrows on the keyboard corresponding to the correct direction, with index and middle fingers of left and right hands.

Results

Figure 1 shows the conflict adaptation pattern in RTs and errors for PWI and PP tasks. As can be seen, the pattern of RTs and errors in both tasks is compatible with conflict adaptation. Due to the small number of errors, statistical analyses focused on the RT data.

PWI. Erroneous responses (5%), microphone problems (2%), and trials following these errors (to avoid post-error slowing effects; 7%) were excluded from the RT analysis. RTs were log-transformed prior to analysis in order to better approximate a Gaussian distribution and outliers (<1%) were removed using QQ Plots (e.g. Schmidt & Weissman, 2015). Data were analyzed using linear mixed-effect models in R v3.2.3 with the lmerTest package (e.g. Schmidt & Weissman, 2015). Fixed-effect structures included the main effects of previous-trial congruency, current-trial congruency, and their interaction (i.e. the canonical test of conflict adaptation). For random effect grouping factors of subject and item, structures included random intercepts as well as random slopes of previous-trial congruency, current-trial congruency, and their interaction.

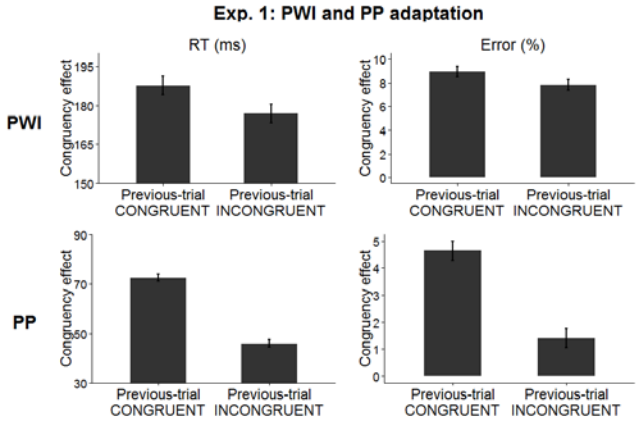


Figure 1: Conflict adaptation in RT and errors in PWI (upper panel) and PP (lower panel). Each bar reflects the subtraction of current congruent from current incongruent (*I−C;* means of participant means ± SE). Conflict adaptation predicts that this difference should be larger for the previous-congruent (left bars) than the previous-incongruent (right bars), a pattern reflected in all four diagrams.

Analysis of log-transformed RTs revealed significant conflict adaptation through an interaction between previous-trial congruency and current-trial congruency ($\beta = -0.005$, $t = -2.0$, $p = 0.049$). Post-hoc tests revealed a reliable effect on current congruent trials: RT was slower on *iC* trials (M = 760, SE = 4.34 ms) relative to *cC* (M = 743, SE = 5.21; $\beta = 0.02$, $t = 4.35$, $p < 0.001$), but *iI* (M = 937, SE = 5.9 ms) and *cI* (M = 930, SE = 4.89 ms) trials were not significantly different ($\beta = 0.004$, $t = 0.53$, $p = 0.60$).

PP. Response errors and subsequent trials each accounted for 3% of the data and were excluded. Mixed-level models were specified using procedures analogous to those used in PWI analysis. Analysis of log-transformed RT revealed a robust effect of conflict adaptation ($\beta = -0.09$, $t = -9.98$, $p < 0.001$), replicating previous results (Weissman et al., 2015). Pairwise

comparisons indicated that this adaptation effect emerged in both current congruent trials, as a slowing of reaction time in *iC* trials (M = 301, SE = 1.91 ms) relative to *cC* (M = 283, SE = 2.1 ms; β = 0.06, t = 9.27, p < 0.001) and in current incongruent trials, as a speeding of reaction time in *iI* trials (M = 347, SE = 2.08 ms) relative to *cI* (M = 356, SE = 2.04 ms; β = −0.02, t = −4.54, p = 0.01).

Cross-task Correlation of Adaptation. To probe whether the size of conflict adaptation in one task was predictable from the size of adaptation in the other task, we conducted an analysis of individual differences. Each participant's adaptation effect was calculated as (cI − cC) − (iI − iC), once for RTs and once for errors. Adaptation size in one task was not predictive of adaptation size in the other[1] (R = 0.06, p = 0.7).

Discussion

When tested in isolation, conflict adaptation was found in both PWI and PP, demonstrating that, similar to spatial tasks, language production is subject to dynamic, top-down regulation. While both congruent and incongruent trials showed the effect in PP, the effect was only reliable on the congruent trials in PWI. This pattern is not uncommon in adaptation studies (Duthoo et al., 2014; Kan et al., 2013; Weissman et al., 2015), but could also reflect a lack of power in detecting a reliable effect in post-hoc tests that use only half of the materials. To address this issue, we doubled the number of trials in Exp. 2.

Additionally, we found no correlation between the size of conflict adaptation in PWI and PP at the level of individuals. This could imply that the two tasks indeed use different regulatory mechanisms, but could also reflect the problems associated with using the correlational method. Internal consistency is known to be low for effects calculated as subtractions (e.g. Redick et al., 2013), and when internal consistency of measures is low (as was the case here), correlations between measures are unreliable. Exp. 2 addressed this problem by directly manipulating cross-task conflict adaptation by interleaving the two tasks.

Experiment 2

Methods

Participants. Thirty-two native English speakers recruited from the Johns Hopkins University community (24 women; mean age = 24.8 years) participated for payment. None had participated in Exp. 1.

Materials. The same materials as Exp. 1 were used with a minor addition: we created eight new PWI stimuli to bring the number of PWI trials to a multiple of 16 (as was necessitated by the PP design, in which four targets were balanced across four conditions). The design of Exp. 1 was changed in two ways: (1) we interleaved PWI and PP trials within each experimental block in an alternating ABAB pattern to minimize switch costs. (2) We duplicated the number of blocks, so that the same PWI target appeared in all four conditions within each participant in counter-balanced order. This change was made to increase the power to detect conflict adaptation in PWI, as well as in a potential cross-task adaptation effect. Each PWI target occurred once before each PP condition for each participant, and each PP condition occurred equally often before each PWI target for each participant. This design ensured that any item-specific effects of PWI on PP or vice versa would be balanced between our conditions of interest. This also led to a balanced design for analyzing the 2-back conflict-adaptation. The final materials consisted of four blocks, each containing 256 experimental trials prepared in four different orders to avoid systematic order effects.

Procedures. Procedures were similar to those used in Exp. 1. Following the familiarization of PWI stimuli, participants completed three practice blocks. The first was a 10-trial PWI block, the second was a 48-trial PP block, and the third was a 20-trial task-switching block. They then completed the four experimental blocks in counterbalanced order.

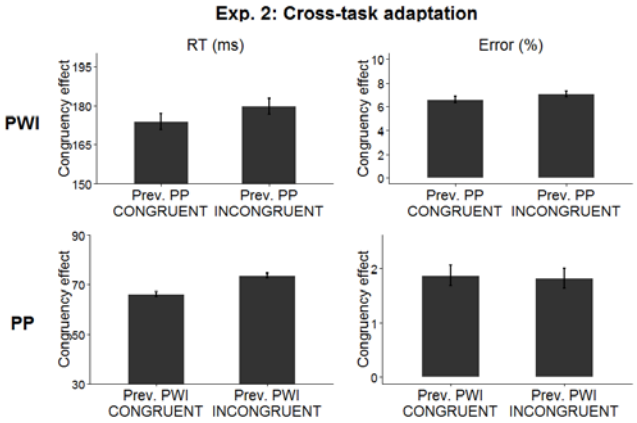


Figure 2: Performance in PWI (upper) and in PP (lower) as a function of previous task congruency. Graphs depict means of participant means for RTs and error rates (± SE). No evidence for conflict adaptation was found. PP showed a reliable anti-adaptation effect.

Results and Discussion

Figure 2 shows the results of cross-task conflict adaptation for PWI (as a function of PP conflict; upper panel) and PP (as a function of PWI; lower panel) on RTs and errors. Neither pattern is compatible with cross-task conflict adaptation. The effects on RTs were tested using similar mixed-level models

[1] To ensure that outliers did not influence these results, we also calculated non-parametric correlations, which returned similar results to the parametric test (Spearman's rho = 0.06; P = 0.7).

as Exp. 1 with maximally specified random-effects. Fifteen and 11% of responses were excluded from the PWI and PP respectively, due to errors, post-error status, microphone malfunction, or outlier status.

Cross-task Adaptation. In PWI, the congruency of the previous PP trial did not interact with the congruency of PWI ($\beta = 0.006$, $t = 0.89$, $p = 0.38$). In PP, the congruency of the previous PWI trial interacted significantly with the congruency of PP ($\beta = 0.02$, $t = 2.8$, $p = 0.03$), but the direction of this interaction is the opposite of what conflict adaptation predicts, that is, congruent PWI trials (compared to incongruent ones) caused slightly *longer* RTs on subsequent congruent PP trials (*cC:* M = 316, SE = 3.24; *iC:* M = 309, SE = 2.72, while incongruent PP RTs were similar regardless of previous-trial conflict (*iI:* M = 383, SE = 2.96; *cI:* M = 382, SE = 2.75).

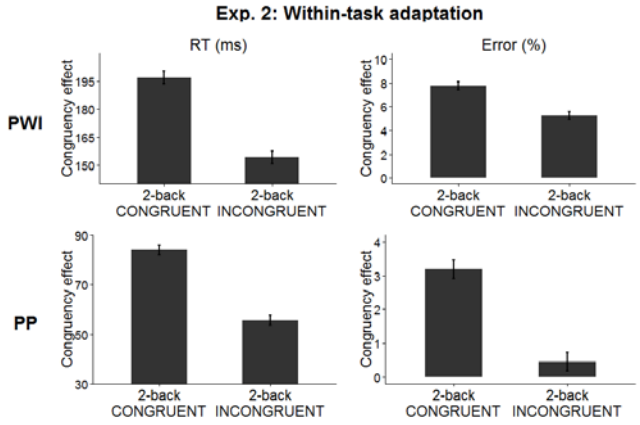


Figure 3: Performance in PWI (upper) and PP (lower) as a function of the n-2 trial congruency. Graphs depict means of participant means for RTs and error rates (± SE). Robust evidence for within-task conflict adaptation was found.

These results suggest that the ability to resolve PWI conflict was not improved by recent experience of PP conflict (and vice versa), in line with predictions of domain-specificity. But, two alternative hypotheses must first be refuted: (1) the null effect may simply reflect the lack of statistical power to detect a significant effect. (2) More theoretically-interesting, is the possibility that interference between task representations in working memory prevented sustained activation of those representations from one trial to the next (Braem, Abrahamse, Duthoo, & Notebaert, 2014). The latter could mask the operation of a domain-general regulatory loop due to excessive task demands. To address these concerns, we examined higher-order sequence effects in our data, namely adaptation as a function of conflict occurring *two* trials before. Due to the ABAB task-switch pattern, the 2-back trial would allow for assessment of within-task conflict adaptation. If robust adaptation is found in PWI and PP as a function of conflict on two trials before, we can reject these alternative hypotheses.

2-back Within-task Adaptation. Within-task adaptation was assessed with the interaction between 2-back congruency and current-trial congruency. This interaction was significant for both PWI ($\beta = -0.01$, $t = -5.77$, $p < 0.001$), and for PP ($\beta = -0.09$, $t = -5.96$, $p < 0.001$). Moreover, post-hoc tests in PWI revealed significant effects on both current congruent trials (*iC:* M = 742, SE = 4.28; *cC:* M = 722, SE = 3.96; $\beta = 0.04$, $t = 5.69$, $p < 0.001$) and current incongruent trials (*iI*: M = 896, SE = 4.68; *cI*: M = 919, SE = 4.61; $\beta = -0.02$, $t = -2.41$, $p = 0.02$). The same was true for PP: both current congruent (*iC:* M = 320, SE = 2.58; *cC:* M = 307, SE = 3.53; $\beta = 0.05$, $t = 4.84$, $p = 0.001$) and current incongruent trials (*iI:* M = 375, SE = 2.52; *cI:* M = 391, SE = 3.49; $\beta = -0.04$, $t = -3.69$, $p = 0.02$) showed a reliable adaptation effect in the predicted direction. These analyses refuted low power and limitations of working memory as alternative explanations for the absence of cross-task adaptation between PWI and PP.

General Discussion

Our first question was whether language production is regulated online through similar mechanisms that regulate non-linguistic tasks. The sequence effects observed in PWI in both experiments took the classic pattern of adaptation, demonstrating dynamic, top-down regulation of language production. While often discussed as "conflict adaptation," alternative accounts have been proposed (e.g. Lamers & Roelofs, 2011). Importantly though, all accounts agree that the pattern reflects online adjustments of top-down control based on the bottom-up requirements of the task.

We then asked whether the conflict-monitoring and control loop is shared between language production and a task that does not involve linguistic representations (Exp. 2). The most general view of conflict monitoring would predict that an increase in conflict in any domain would lead to increased control in other domains. To this end, we first replicated conflict adaptation using the spatial PP task (Weissman et al., 2015), then tested whether interleaving trials from PWI and this task would lead to cross-task conflict adaptation. We found no evidence in support of adaptation in one task as a function of conflict in the other, in either analysis of individual differences (Exp. 1) or in cross-task adaptation (Exp. 2). Demonstration of robust 2-back (within-task) adaptation in both PWI and PP allowed us to reject low statistical power and limitations of working memory as alternative explanations for the null cross-task adaptation effect. Thus, the current results convincingly refute a fully domain-general control system in which control is insensitive to the nature of conflict.

Our results are in agreement with several other studies also concluding at least some specificity in the process of control regulation (Egner et al., 2007; Forster & Cho, 2014; Wühr et al., 2015). One prominent exception is Kan et al. (2013), which found evidence for cross-task adaptation in button-press Stroop and passive viewing of the Necker cube. One difference between Kan et al.'s study and the current study is response modality (button press vs. oral), but since Stroop and PWI conflict most likely occur at a similar level (Piai,

Roelofs, & Schriefers, 2014), this difference is unlikely to have caused the discrepancy.

A more salient difference is the level of forced engagement of control in the spatial tasks. In the PP task, a speeded response is required on each trial, forcing the spatial control system to engage in all participants. But, viewing a Necker cube does not force engagement of control. Incidentally, cross-task adaptation in Kan et al. (2013) was only found in a subset of participants (N = 14) with a large number of switches between the two percepts of the cube, while no effect was found in other participants with fewer switches. It is difficult to determine what characteristic of the participants (e.g. motivation, effort, etc.) led to cross-task adaptation, but the effect was limited to a select group.

The current findings, along with prior evidence for domain-specificity in the monitoring-control system (Egner et al., 2007; Forster & Cho, 2014; Wühr et al., 2015), invite caution in interpreting the results of Kan et al. (2013) as strong support for a fully domain-general control system. Instead, our findings favor accounts that posit at least some level of domain-specificity, even when allowing for contribution of domain-general processes (e.g. Nozari et al., 2011; Gauvin et al., 2016).

References

Acheson, D., & Hagoort, P. (2014). Twisting tongues to test for conflict-monitoring in speech production. *Frontiers in Human Neuroscience*, *8*.

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Braem, S., Abrahamse, E., Duthoo, W., & Notebaert, W. (2014). What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis. *Frontiers in psychology*, 5.

Duthoo, W., Abrahamse, E., Braem, S., Boehler, C., & Notebaert, W. (2014). The Congruency Sequence Effect 3.0: A Critical Test of Conflict Adaptation: e110462. *PLoS One*, *9*(10), e110462.

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*(10), 374–380.

Egner, T., Delano, M., & Hirsch, J. (2007). Separate conflict-specific cognitive control mechanisms in the human brain. *NeuroImage*, *35*(2), 940–948.

Freitas, A., & Clark, S. (2015). Generality and specificity in cognitive control: conflict adaptation within and across selective-attention tasks but not across selective-attention and Simon tasks. *Psychological research*, 79(1), 143-162.

Forster, S., & Cho, R. (2014). Context Specificity of Post-Error and Post-Conflict Cognitive Control Adjustments. *PLoS ONE*, *9*(3).

Gauvin, H., De Baene, W., Brass, M., & Hartsuiker, R. (2016). Conflict monitoring in speech processing: An fMRI study of error detection in speech production and perception. *NeuroImage*, *126*, 96–105.

Kan, I., Teubner-Rhodes, S., Drummey, A., Nutile, L., Krupa, L., & Novick, J. (2013). To adapt or not to adapt: The question of domain-general cognitive control. *Cognition*, *129*(3), 637–651.

Lamers, M., & Roelofs, A. (2011). Attentional control adjustments in Eriksen and Stroop task performance can be independent of response conflict. *The Quarterly Journal of Experimental Psychology*, 64(6), 1056-1081.

Nozari, N., Dell, G., & Schwartz, M. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, *63*(1), 1–33.

Piai, V., Roelofs, A., Acheson, D., & Takashima, A. (2013). Attention for speaking: domain-general control from the anterior cingulate cortex in spoken word production. *Frontiers in Human Neuroscience*, *7*.

Piai, V., Roelofs, A., & Schriefers, H. (2014). Locus of semantic interference in picture naming: Evidence from dual-task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(1), 147.

Redick, T., Shipstead, Z., Harrison, T., Hicks, K., Fried, D., Hambrick, D., ... & Engle, R. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, 142(2), 359.

Riès, S., Janssen, N., Dufau, S., Alario, F., & Burle, B. (2011). General-purpose monitoring during speech production. *Journal of Cognitive Neuroscience*, *23*(6), 1419–1436.

Schmidt, J. (2013). Questioning conflict adaptation: proportion congruent and Gratton effects reconsidered. *Psychonomic Bulletin & Review*, *20*(4), 615–630.

Schmidt, J., & Weissman, D. (2015). Congruency sequence effects and previous response times: conflict adaptation or temporal learning? *Psychological Research*.

Schnur, T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language*, *75*, 27–44.

Schriefers, H., Meyer, A., & Levelt, W. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, 29(1), 86-102.

Weissman, D., Egner, T., Hawks, Z., & Link, J. (2015). The congruency sequence effect emerges when the distracter precedes the target. *Acta Psychologica*, *156*, 8-21.

Wühr, P., Duthoo, W., & Notebaert, W. (2015). Generalizing attentional control across dimensions and tasks: Evidence from transfer of proportion-congruent effects. *The Quarterly Journal of Experimental Psychology*, *68*(4), 779–801.

Yeung, N., Botvinick, M., & Cohen, J. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, 111(4), 931.

Active Viewing in Toddlers Facilitates Visual Object Learning: An Egocentric Vision Approach

Sven Bambach, David J. Crandall, Linda B. Smith†, Chen Yu†
{sbambach, djcran, smith4, chenyu}@indiana.edu
School of Informatics and Computing, Indiana University
†Department of Psychological and Brain Sciences, Indiana University
Bloomington, IN, 47405 USA

Abstract

Early visual object recognition in a world full of cluttered visual information is a complicated task at which toddlers are incredibly efficient. In their everyday lives, toddlers constantly create learning experiences by actively manipulating objects and thus self-selecting object views for visual learning. The work in this paper is based on the hypothesis that active viewing and exploration of toddlers actually creates high-quality training data for object recognition. We tested this idea by collecting egocentric video data of free toy play between toddler-parent dyads, and used it to train state-of-the-art machine learning models (Convolutional Neural Networks, or CNNs). Our results show that the data collected by parents and toddlers have different visual properties and that CNNs can take advantage of these differences to learn toddler-based object models that outperform their parent counterparts in a series of controlled simulations.

Keywords: vision, visual object learning, convolutional neural networks, head-mounted cameras

Introduction

Visual object recognition is of fundamental importance to humans and most animals, whose everyday lives rely on identifying a large variety of visual objects. Because of its importance, even human infants possess sophisticated perceptual and learning processes to form categorical representations of visual stimuli (Quinn & Eimas, 1996). Even as toddlers, they already seem to be able to easily recognize everyday objects. A vexing question for cognitive scientists is how young learners achieve this ability in a visually noisy and dynamic world where objects are often encountered under seemingly suboptimal conditions, including in unusual orientations, varying lighting conditions, or partial occlusions (Johnson & Aslin, 1995; Casasola, Cohen, & Chiarello, 2003). Despite recent progress in object recognition in the computer vision community (Krizhevsky, Sutskever, & Hinton, 2012), even the most powerful computational algorithms trained with large amounts of data are arguably not yet able to learn as efficiently as toddlers do.

Many previous studies on early visual object recognition focus on examining exactly what visual information is extracted from the retinal image to construct invariant descriptors of objects. For this purpose, many experimental paradigms have been invented that repeatedly expose young visual learners to stimuli displayed on a computer screen (familiarization phase), and then measure looking times towards familiar and novel stimuli (test phase). These paradigms are powerful, allowing us to examine, in a rigorously controlled way, which visual features are extracted, how they are stored in memory, and how they are activated to recognize new instances. However, we also know that these experimental paradigms are very different from young children's everyday learning experiences: active toddlers do not just passively perceive visual information but instead generate manual actions to objects, thereby creating self-selection of object views (Yu et al., 2009). Indeed, recent work shows that infants who have more experience in manual object exploration have more robust expectations about unseen views of novel objects (Soska, Adolph, & Johnson, 2010). Another study using head-mounted cameras to record toddlers fields of view found a preference towards planar views of objects: toddlers dwelled longer on these views while manually exploring held 3-d objects than would be expected if the objects were rotated randomly. This bias substantially increased between the ages of 12-36 months (Pereira et al., 2010).



Figure 1: All instances of a toy as seen by cameras mounted on heads of toddlers (left) and parents (right) during joint play between 10 toddler-parent dyads, showing greater diversity in toddler views. Instances are shown to scale and colored boxes depict the field of view size.

Visual object recognition depends on the specific views of objects experienced by the learner. In everyday contexts such as toy play, toddlers actively create many different views of the same object. In light of this, the overall hypothesis in the present study is that active viewing may create high-quality training data for visual object recognition. To test this hypothesis, we used head-mounted cameras to collect first-person video data from a naturalistic environment in which

parents and children were asked to jointly play with a set of toy objects. Figure 1 shows examples of different views of the same toy car, collected from toddlers' view (left) and parents' view (right) during the same play sessions. Clearly, toddlers created more diverse views in terms of relative object size, orientation, and occlusion compared to their parents. In the present study, we first quantify the differences in visual properties of objects between toddler and parent views, finding a higher variation among visual instances for the child. A learning system could take advantage of such variation by building more generalizable representations for recognizing unseen instances, thus better facilitating visual object recognition.

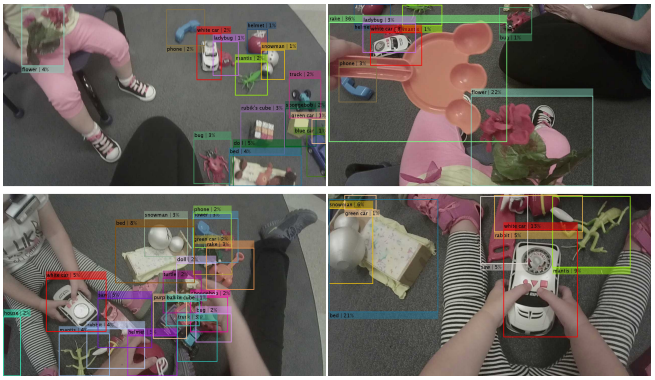
To test this idea, the main focus of the study was to train machine learning models based on Convolutional Neural Networks (CNNs), which are currently considered the most powerful visual learning models in the computer vision community (Krizhevsky et al., 2012), with data from the two different views, and to examine the extent to which these models take advantage of visual information created and perceived by toddlers. The results show that the CNNs perform better on object recognition in multiple simulation conditions when trained with the toddlers' data than with the parents' data. To the best of our knowledge, this is the first study to collect and use egocentric video in everyday contexts and demonstrate a working learning system taking advantage of object view self-selection by active toddlers for visual object recognition.

Data Collection

To test our hypotheses and models, we collected two types of image data, one for training our CNN models and one for testing them. For the training data, we used head-mounted cameras to capture first-person video of toddlers and parents as they jointly played with a set of toys in a naturalistic, unconstrained setting. For the test data, we collected a controlled dataset in which we photographed the same set of objects, but against a clean background and from a systematic set of canonical viewpoints. We now describe each dataset in detail.

Training Data

The training data was collected in a small (~$15m^2$) room with a soft carpet to facilitate sitting on the floor. This "toy room" had an adult-sized chair and a toddler-sized chair, but otherwise no large objects or other distractions. Figure 2 gives an impression of the setting. Our data was collected from 10 child-parent dyads (9 mothers and 1 father; 6 girls and 4 boys, mean child age 22.6 months and SD = 2.1 months). Before entering the toy room, both the parent and toddler were equipped with a head-mounted camera. Both cameras were small (4.8cm×4.8cm×1.5cm), lightweight (22g) *Looxcie 3* cameras with a 100° diagonal field of view. Video data was recorded directly onto a microSD card. Cameras were attached (with velcro) to an adjustable headband to ensure a tight but comfortable fit on the center of each participant's forehead. Next, we randomly arranged 24 toys (Figure 3)



(a) Parent view (b) Child view

Figure 2: First-person video examples that were captured during joint child-parent play in our toy room, contrasting parent view (a) and child view (b). Each row shows one dyad. Also shown are bounding boxes and toy sizes (as % of FOV).



Figure 3: The 24 toys that were used in all of our experiments.

in the center of the floor and encouraged the dyad to play together as they pleased. Once they were engaged with the toys, we left the room and did not give further instructions. Most parents sat on the floor, while toddlers switched between sitting on the floor and walking or crawling around to pick up new toys. Two toddlers briefly sat in the small chair.

For each child-parent dyad, we extracted the greater of 10 minutes of video or the longest period of continuous toy play (uninterrupted by the child taking off the camera or losing interest), yielding at least 3 minutes 35 seconds and an average of 7 minutes 58 seconds of video per dyad. All videos were captured with a resolution of 720×1280 pixels at 30 frames per second, and each video pair between toddler and parent was synchronized.

The location of each of the 24 toy objects within the captured first-person video data was manually annotated. To do this, we subsampled the video stream at one frame every five seconds, and then manually drew bounding boxes around each toy in each frame. Figure 2 shows four annotated example frames. Since toys were often occluded by other objects or truncated at the frame boundaries, we used the following guideline: if only part of a toy was visible, we drew a box around the part if it was visually identifiable as the right toy; if multiple parts of an identifiable toy were visible, we drew a box that included all visible parts of the toy.

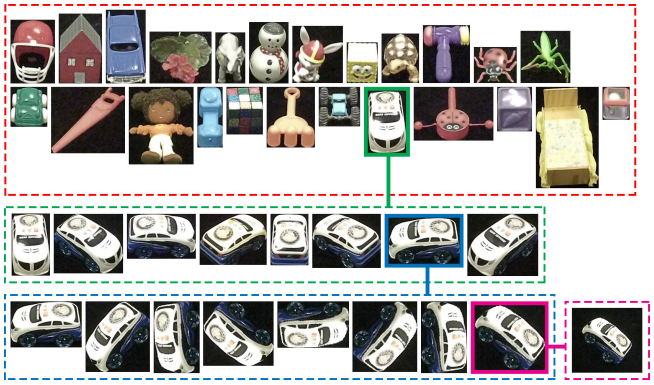
Overall, we captured 9,646 toy instances from the toddler views and 11,313 from the parent views, for an average of 401

instances per class across all toddlers and 471 across parents. There were no large outliers for any of the toys in terms of appearance frequency; the least frequent toy appeared 307 and 341 times for toddlers and parents, respectively, while the most frequent appeared 559 and 600 times.

Testing Data

We also created a separate test set of the same 24 toy objects. The goal of this test data was to have a large variety of clean, systematically-collected, unobstructed third-person views for each toy, to serve as a view-independent and therefore objective way to evaluate the performance of visual object recognition. We again used the *Looxcie 3* camera but this time captured static photos (at the same resolution as the video). The toy room floor was covered in a black cloth to obscure background clutter, and the camera was mounted onto a tripod, pointing towards the floor at a 45° angle. Each toy was put on the floor at a distance of 50cm from the tripod center. The height of the camera was 45cm, creating a distance from lens to toy center of around 67cm, which approximately centered each toy in the camera frame.

We captured 8 photos from each toy, one from each 45° angle rotation around its vertical axis. Sample images from every toy are shown inside the red box in Figure 4, while the green box shows one of the toys from all 8 viewpoints. To create even more diversity, we additionally rotated each image around the optical center of the camera in 45° increments (blue box in Figure 4). Images were then cropped to a bounding box around the object. To add scale variation, we padded and rescaled images to simulate zooming out by a factor of two (cyan box in Figure 4). In total, our test data consisted of $8 \times 8 \times 2 = 128$ images for each toy and 3,072 images total.



(a) Controlled toy images



(b) Toy image with added occlusion

Figure 4: (a) Samples from the controlled test data. Each of the 24 toys (red) was photographed from 8 viewpoints (green), and each resulting image was further rotated 8 times (blue). To add scale variation, each image was also cropped at a lower zoom level (cyan). (b) Sample test images with synthetic occlusion.

Study 1: Quantifying and Comparing Object Properties in Egocentric Views

During joint play, toddlers and parents generate many instances of visual objects within their self-selected fields of view. Our first study quantified and compared properties of object appearance across the two views.

Object Appearance in the Field of View

We begin by studying how many toys are present within the field of view, as well as the perceptual size of those toys.

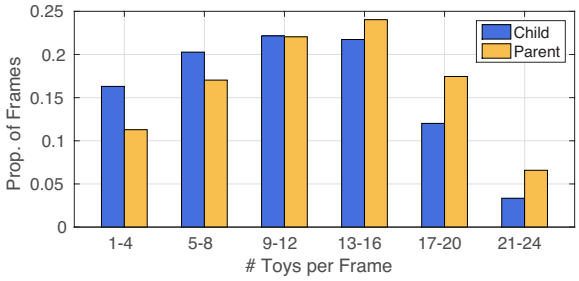
Number of Visual Objects Figure 5(a) presents histograms showing the number of visual objects that appear simultaneously in the field of view. Toddlers have a larger fraction of frames (16.3%) with only 1-4 objects compared to parents (11.3%). Conversely, parents are more likely to have most objects in view at once, with 24.0% of parent frames containing more than 17 objects versus only 15.4% of toddler frames.

Visual Object Sizes Next, we investigate the size of visual objects within the fields of view. We approximate the actual size of an object with the area of its bounding box, and measure the fraction of the field of view that is occupied by this box. Figure 5(b) shows that 42.3% of object instances occupy $\leq 2\%$ of parents' field of view, while only 3.5% of all objects appear dominantly in view ($> 8\%$ of FOV). On the other hand, toddlers exhibit a more spread-out distribution: only 28.3% of objects appear small ($\leq 2\%$ FOV) while 11.9% of objects occupy more than 10% of the view. For perspective, the white car (red bounding box) in the bottom row of Figure 2 occupies 5% of the parent view (a) and 13% in the child view (b). These results are consistent with findings from previous head-camera studies (Yu et al., 2009).

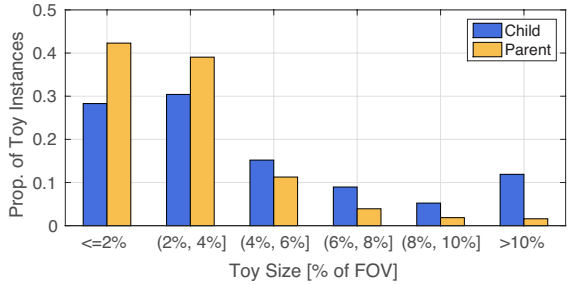
Variation in Visual Object Appearance

Finally, we aim to quantify the visual diversity across the views. We resize each toy image to a canonical size (10×10 pixels), and, for each subject, compute the pixelwise mean squared error (*MSE*) between all instances of the same object. In other words, we take each $10 \times 10 \times 3$ color image, represent it as a 300-dimensional vector, and then compute the mean *MSE* distance between all possible pairs of instances for each object and subject. This score should be low for a subject who sees many visually similar instances of an object, and high for one that sees much variation. We compute scores on a per subject basis to control for inter-subject differences in object appearance.

Figure 6 compares the visual diversity between views generated by toddlers and parents. For each object, we subtract the average MSE score across all toddlers from the corresponding score across parents, so that positive values indicate higher diversity in toddler views while negative values indicate higher diversity in parent views. Using this metric, we



(a) Number of toy objects visible per frame



(b) Size of toy objects within the field of view

Figure 5: Comparison of how objects appear in the fields of view of toddlers and parents, in terms of (a) number of objects appearing simultaneously, and (b) size of objects in view.

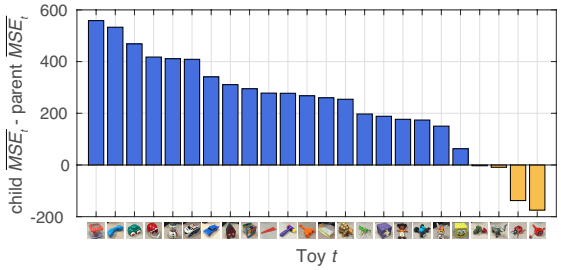


Figure 6: Difference between toddlers and parents in the visual diversity for each of the toys. Positive values indicate higher diversity for toddlers. See text for details.

find that toddlers on average generated more diverse views for 20 out of the 24 toys. We also experimented with other image representations such as grayscale image vectors and GIST features (Oliva & Torralba, 2001) (which capture shape and texture information), and found similar tendencies, with 21 and 15 toys being considered as more diverse respectively.

Discussion

Our study showed large differences in the views of objects that toddlers and adults interact with, even when jointly interacting with them at the same time. While parents are more likely to have "overview" views including many objects, toddlers are more likely to pick out and inspect single objects up close, resulting in fewer, larger objects in view. Additionally, this object selection process seems to create more diverse viewpoints for the children than for the parents.

Study 2: Visual Object Recognition Based on Deep Learning Models

Deep learning using Convolutional Neural Networks (CNNs) has recently shown impressive success in computer vision, improving the state-of-the-art for visual recognition by a large margin (Krizhevsky et al., 2012). We investigate how well a CNN trained with real-world toy instances (as captured during our joint play experiments) recognizes the same 24 visual objects in a separate, controlled testing environment. We do not claim that a CNN constitutes the perfect model to emulate visual object learning in toddlers (or humans in general). Instead, we are interested in CNNs as ideal learners. We assume that the network will learn to use whichever visual features are sufficient to distinguish the 24 objects. Given the differences in captured visual object views of parents and toddlers, two separate networks, one trained with toddler data and the other trained with parent data, might learn different (better) strategies. More directly, we hypothesize that the toddler data captures a richer representation of each object, leading to better classification performance on the controlled test data.

We first describe CNN implementation details and verify that the networks can learn visual object appearance based on first-person data. We then test the networks in a series of experiments based on our controlled views of each object.

Convolutional Neural Networks

CNNs are a special type of multi-layer, feed-forward neural networks, consisting of multiple convolutional layers followed by multiple fully-connected layers. Neurons between the convolutional layers are connected sparsely and with shared weights, effectively implementing a set of filters. Filter responses are passed to a non-linear activation function as well as a local pooling function before serving as input to the next layer. Intuitively, the convolutional layers learn filters (from low-level in early layers to high-level in deep layers) that extract image features, while the fully-connected layers act as a classifier. Please see (Krizhevsky et al., 2012) for more details on CNNs.

Implementation For all our experiments, we used the well established AlexNet CNN architecture (Krizhevsky et al., 2012), consisting of five convolutional layers and three fully-connected layers. The input layer of the network has a fixed size of 224×224×3 neurons, which means the network expects input images to be resized to 224×224 pixels. Instead of training the network from scratch, we follow the common protocol of beginning with a network pre-trained on the ImageNet dataset (Deng et al., 2009), which consists of millions of images. We adjust the final layer to have 24 neurons to accommodate our 24-way object classification task, and then use the parameters learned from ImageNet as initialization for training on our data. Each network is trained via back propagation with a softmax loss function, using batch-wise stochastic gradient descent with a learning rate of 0.001, momentum of 0.9, and batch size of 256 images.

Simulation 1: CNNs Learn from the Training Data

Before we experiment with controlled test images, we need to ensure that CNNs are indeed able to learn visual object models from our first-person data. As detailed in Data Collection, our training data includes 24 different visual objects with 11,313 parent views and 9,646 toddler views. Figure 1 illustrates how some of these images look. We now consider these images as two different datasets (toddler data and parent data), and perform a 6-fold cross validation split on both. This means that for each dataset we train six different CNN networks that each use a randomly-selected one-sixth of the data for testing and the remaining five-sixths for training.

For both parent and toddler data, we found that the back-propagation converged after about 10 epochs, i.e. after observing the training data around 10 times. The average test accuracy across splits was 89.9% for the toddler views and 93.1% for the parent views. To put this into perspective, random guessing achieves $1/24 \approx 4.2\%$, while guessing the majority class achieves 5.8% for toddlers and 5.3% for parents.

We investigated failure cases by inspecting the ten images that each network was least confident about (i.e., having the lowest predicted probability of the true class). Both parent and toddler networks showed similar patterns, where most mistakes were caused by either two or more objects overlapping each other, strong motion blur, or a combination of both. From this we conclude that the failures are reasonable and that CNNs are indeed able to learn from the first person data.

Simulation 2: Using Testing Data from a Third-Person View

Now we investigate how well learned concepts from the first-person training data transfer to the clean testing data. We trained a CNN on the first-person toddler training data, and a separate CNN on the first-person adult training data, and then tested both with the same controlled testing data described above (3,072 images). To avoid learning frequency biases, since some objects have more training instances than others, we uniformly sampled the training data from each class. Given that CNN training is non-deterministic due to random training subset sampling and parameter initialization, we repeated the full training and test procedure over 10 independent trials. We stopped training each network after convergence (around 12 epochs).

As shown in Figure 7(a), the networks trained on toddler data achieve higher recognition accuracy by 6.3 percentage points compared to the networks trained on parent data, reaching 79.6% as opposed to 73.3%. Figure 7(a) also compares the distribution of mean accuracies for each object. Overall, the child networks achieve the same or better results for 16 out of the 24 toys, indicating that the differences in overall accuracy are not caused by some minority of classes.

Simulation 3: Recognizing Occluded Objects

Another interesting question is how well the toddler and parent views allow the trained networks to deal with occlusion.

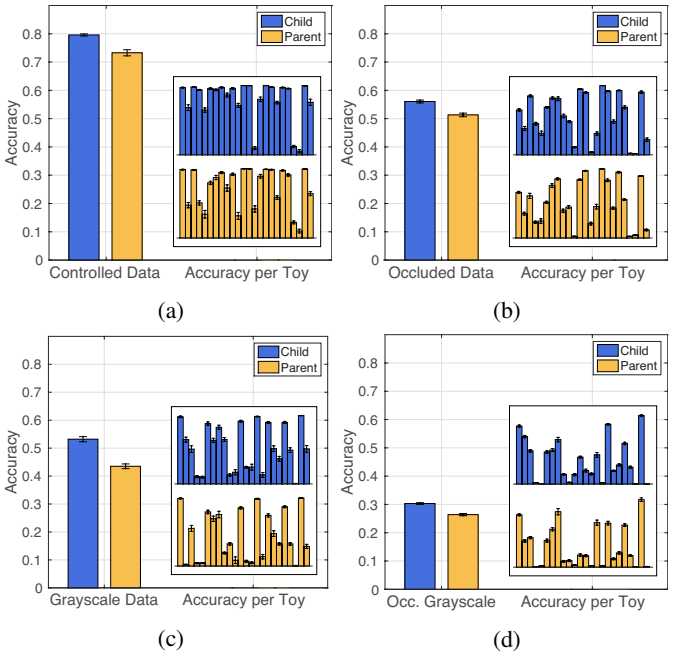


Figure 7: Classification accuracies of CNNs trained with first-person image data from toddlers (blue) and parents (orange), when tested on controlled image data of the same objects. Bars show standard errors across 10 trained networks.

To test this, we systematically added occlusion to each testing image, by splitting the image into quadrants and then occluding each possible combination of one to three quadrants with gray boxes. This resulted in 14 occlusions per image, as shown in Figure 4(b). The occluded testing data thus consisted of $14 \times 3{,}072 = 43{,}008$ images.

Figure 7(b) presents results of testing the same 2×10 networks from Simulation 2 on the occluded data. Toddler networks retain better overall mean accuracy compared to the parent networks (56.1% vs. 51.3%). The relative performances when compared to the non-occluded data drop by ~30% for both parent and toddler networks, which suggests that both are affected similarly by occlusion.

Simulation 4: The Effect of Color Information

The differences in performance on the controlled object images might be because one set of networks relies more on color information to learn object models based on the first-person training data. To examine this idea, we repeat all experiments with grayscale images.

Cross-validation First, we investigate if the absence of color information increases the difficulty to learn from the two first-person datasets. We repeat the same 6-fold cross validation experiments as described above in Simulation 1, but this time train the networks with grayscale images. The average test accuracy across splits decreased to 76.9% (from 89.9%) for the toddler networks and to 83.1% (from 93.1%) for the parent networks. Thus, the absence of color accounts for a rather small drop in learnability of the data.

Controlled Testing We repeat our Simulation 2 and 3 experiments and train two sets of 10 networks, one with the grayscale toddler images and the other with grayscale parent images, and test them on grayscale versions of the testing dataset images. Figure 7(c-d) summarizes the results. The toddler networks again significantly outperform parent networks in terms of overall mean accuracy, both for the non-occluded and the occluded testing data, with mean accuracies of 52.3% over 43.5%, and 30.3% over 26.4%, respectively. The relative performance drops compared to the color experiments are in favor of the toddler networks on the non-occluded data (-34% vs. -40%), but slightly in favor of the parent networks on the occluded data (-51% vs. -54%).

Discussion

Our results suggest that naturalistic first-person images of unconstrained toy play can be used to train CNN-based object models that generalize to recognizing the same objects in a different context. It appears that toddlers generate high quality object views that facilitate learning, as networks trained on toddler data consistently outperformed parent networks.

Summary and General Discussion

In the present paper, we collected egocentric video data of free toy play between toddler-parent dyads, and used it to train state-of-the-art machine learning models (CNNs). Our results showed that (1) CNNs were indeed able to learn object models of the toys in this first-person data and (2) that these models could generalize and recognize the same toys in a different context with different viewpoints. Finally, we showed that (3) the visual data collected by toddlers seems to be of particularly high quality as models trained with toddler data consistently outperformed those trained with parent data in multiple simulation conditions.

In the real world, toddlers spend hours every day playing with toys, actively manipulating objects and, as a result, create learning experiences by self-selecting object views for visual learning. It may not sound surprising that better data leads to better learning. Nonetheless, if active viewing by toddlers creates high-quality training data for object recognition, as evident in the present study, then the potential long-term impact of day-in and day-out object play that repeatedly and incrementally provides such data may be the key to why toddlers are incredibly efficient in visual object learning. If so, future research should focus not only on studying particular learning mechanisms in experimental tasks but also on how high-quality data is created by learners themselves. A better understanding of human learning systems would require a better understanding of both learning algorithms and the data fed into them. This paper represents a first step towards this direction by linking high-density video data collected in naturalistic contexts with state-of-the-art machine learning.

Our future work will focus on further understanding the factors that may account for the observed performance differences (e.g. different spatial scales and resolutions of object images versus innate differences in the viewpoints). Another future direction within our framework is to further examine how CNNs take advantage of visual instances captured from diverse viewpoints. Our results here seem to support view-based theories of visual object recognition, stating that human learners store viewpoint-dependent surface and/or shape information (Tarr & Vuong, 2002), and recognize objects by their similarity to stored views (Bülthoff & Edelman, 1992). By diagnosing the network directly and visualizing learned high-level filters and the image regions they are likely to fire on (Zeiler & Fergus, 2014), we may be able to provide more direct evidence on the mechanisms of object recognition.

Acknowledgments

This work was supported by the National Science Foundation (CAREER IIS-1253549, CNS-0521433, BCS-15233982), the National Institutes of Health (R01 HD074601, R21 EY017843), and the IU OVPR through the IUCRG and FRSP programs. It used compute facilities provided by NVidia, the Lilly Endowment through its support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship. We thank Sam Dong, Steven Elmlinger, Seth Foster, and Charlene Tay for helping with the data collection.

References

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *PNAS*, *89*(1), 60–64.

Casasola, M., Cohen, L. B., & Chiarello, E. (2003). Six-month-old infants' categorization of containment spatial relations. *Child development*, *74*(3), 679–693.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*.

Johnson, S. P., & Aslin, R. N. (1995). Perception of object unity in 2-month-old infants. *Dev Psychol*, *31*(5), 739.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Proc. NIPS* (pp. 1097–1105).

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, *42*(3), 145–175.

Pereira, A., James, K., Jones, S., & Smith, L. (2010). Early biases and developmental changes in self-generated object views. *Journal of vision*, *10*(11), 22.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *J Exp Child Psychology*, *63*(1), 189–211.

Soska, K. C., Adolph, K. E., & Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Dev Psychol*, *46*(1), 129.

Tarr, M. J., & Vuong, Q. C. (2002). Visual object recognition. In *Stevens' handbook of experimental psychology*. Wiley.

Yu, C., Smith, L., Shen, H., Pereira, A., & Smith, T. (2009). Active information selection: Visual attention through the hands. *IEEE Trans Auton Ment Dev*, *1*(2), 141–151.

Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proc. ECCV* (pp. 818–833).

Processing Consequences of Onomatopoeic Iconicity in Spoken Language Comprehension

David Peeters (david.peeters@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract

Iconicity is a fundamental feature of human language. However its processing consequences at the behavioral and neural level in spoken word comprehension are not well understood. The current paper presents the behavioral and electrophysiological outcome of an auditory lexical decision task in which native speakers of Dutch listened to onomatopoeic words and matched control words while their electroencephalogram was recorded. Behaviorally, onomatopoeic words were processed as quickly and accurately as words with an arbitrary mapping between form and meaning. Event-related potentials time-locked to word onset revealed a significant decrease in negative amplitude in the N2 and N400 components and a late positivity for onomatopoeic words in comparison to the control words. These findings advance our understanding of the temporal dynamics of iconic form-meaning mapping in spoken word comprehension and suggest interplay between the neural representations of real-world sounds and spoken words.

Keywords: Iconicity; Onomatopoeia; Sound Symbolism; Language Processing; Event-Related Potentials

Introduction

Language comprehension involves the mapping of forms onto meaning representations. In spoken Indo-European languages, the relation between word form and meaning is often opaque and arbitrary. For instance, the French word form *arbre* and the English word form *tree* do not resemble the actual referent that these words denote (De Saussure, 1916). However, word forms can also be *iconic* of their meaning and thereby render the form-meaning mapping largely non-arbitrary. One example is the existence of onomatopoeic words, such as *beep, click,* and *slurp* in English, in which the phonological word form resembles the real-world sound it refers to. Other examples of iconic manners of signification in language can be found in ideophones and mimetics (Dingemanse, 2012; Kita, 1997), in the iconic hand gestures people produce in temporal and semantic alignment with their speech (Kendon, 2004; McNeill, 1992), and in many lexical items in sign languages (Taub, 2001; Thompson, 2011). Current theorizing therefore considers not only arbitrariness but also iconicity a fundamental feature of language (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Imai & Kita, 2014; Perniss, Thompson, & Vigliocco, 2010; Perniss & Vigliocco, 2014).

In line with its appraisal as a fundamental property of language, several studies have started investigating the processing consequences of iconicity. The majority of studies investigating the role of iconicity in the processing of existing lexical items focuses on sign language (see Perniss et al., 2010; Vinson, Thompson, Skinner, & Vigliocco, 2015). Results are mixed. In lexical decision tasks native signers do not recognize iconic signs at a different speed or accuracy compared to signs with an arbitrary form-meaning mapping (Bosworth & Emmorey, 2010; Klann, Kastrau, & Huber, 2005), which suggests that iconicity does not convey a lexical processing advantage (Bosworth & Emmorey, 2010). In contrast, a facilitatory role of iconicity has been found in picture-sign matching tasks and in picture naming (Grote & Linz, 2003; Ormel, Hermans, Knoors, & Verhoeven, 2009; Thompson, Vinson, & Vigliocco, 2009; Vinson et al., 2015), which suggests that a close mapping between the form and meaning of a lexical item can aid lexical retrieval (Thompson et al. 2009).

The processing consequences of iconicity at the behavioral and neural level in *spoken* language remain largely unclear and much of the work done has used non-words as stimuli (see Lockwood & Dingemanse, 2015). At the behavioral level, studies have followed up on and replicated Köhler's (1929) observation that people associate non-words containing voiceless stops (e.g., *takete* or *kiki*) with spiky shapes and non-words without voiceless stops or containing only continuant consonants (e.g., *maluma* or *bouba*) with curvy shapes (Köhler, 1929; Ramachandran & Hubbard, 2001; Westbury, 2005). At the neurophysiological level, it was found that the brain is sensitive to the sound symbolic congruency between an auditorily presented non-word label and a subsequently presented visual object as early as 160 ms after onset of the visual object, reflected by a significant increase in negative amplitude of an event-related potential (ERP) component in the N2 time-window for congruent compared to incongruent sound symbolic mappings (Kovic, Plunkett, & Westermann, 2010). However, such consistent associations between the acoustic and phonological characteristics of non-words and the shape of objects are interesting, but do not necessarily imply that people process existing iconic lexical items such as onomatopoeic words differently than words with an arbitrary form-meaning mapping.

Neuroimaging studies suggest that the brain represents existing iconic words such as ideophones and onomatopoeic words qualitatively differently compared to non-iconic words. Listening to onomatopoeic words recruits both brain areas involved in the processing of verbal sounds and areas processing non-verbal (animal) sounds (Hashimoto, Usui, Taira, Nose, Haji, et al., 2006). More specifically, additional activation in the right posterior superior temporal sulcus (pSTS) for comprehending iconic words compared to non-iconic words may reflect that iconic words function as both

linguistic and non-linguistic symbols (Kanero, Imai, Okuda, Okada, & Matsuda, 2014). These findings thus suggest differences in representation for iconic words compared to words with an arbitrary form-meaning mapping due to the former's connection to sensory information and everyday sensory experience. A performance benefit in auditory lexical decision for onomatopoeic words compared to non-iconic control words in patients with aphasia indeed suggests that iconic word forms may activate their meaning partially via non-linguistic pathways (Meteyard, Stoppard, Snudden, Cappa, & Vigliocco, 2015).

The aim of the current study is to further investigate the processing consequences of iconicity at the behavioral and neurophysiological level in spoken word comprehension, by focusing on onomatopoeia as the textbook example of iconicity in spoken Indo-European languages. Onomatopoeic words recruit a different set of brain regions compared to non-iconic words (Hashimoto et al., 2006), but the behavioral consequences of iconicity in the processing of spoken words remain unclear. Furthermore, the temporal dynamics of these differential patterns of neuronal activation remain unknown. Therefore, in the current study both reaction times and the electroencephalogram (EEG) were recorded while participants performed an auditory lexical decision task that contained onomatopoeic words and matched control words. If iconicity aids in lexical retrieval (cf. Thompson et al., 2009), this may be reflected by faster response times (RTs), higher accuracy, and a less negative deflection in the N400 component for onomatopoeic words compared to control words. Conversely, if iconicity does not convey a lexical processing advantage (cf. Bosworth & Emmorey, 2010) no such effects should be observed. If onomatopoeic words are more effortful to process compared to control words (similarly to ideophones, cf. Lockwood & Tuomainen, 2015), this may lead to slower RTs, lower accuracy, and a late positive effect in the ERPs (Lockwood & Tuomainen, 2015). Finally, if the findings by Kovic et al. (2010) generalize to onomatopoeic iconicity, a difference in amplitude of the N2 component may be expected for onomatopoeic words versus control words.

Method

Participants

Thirty-four native speakers of Dutch (28 female; 18-29 years of age, mean age = 22.2) participated in the combined RT and EEG study for monetary reward. They were all right-handed as assessed by a Dutch translation of the Edinburgh Inventory for hand dominance (Oldfield, 1971), had normal or corrected-to-normal vision, and no history of neurological insult or language disability. Data from three participants was excluded from behavioral and EEG analyses due to an error percentage on the behavioral task that exceeded 20%.

Stimulus Materials

Both RTs and ERPs were recorded to 320 trials that consisted of 160 words (40 Dutch onomatopoeic words, 40 Dutch control words, and 80 Dutch fillers) and 160 Dutch-like nonwords. All 160 words were selected from the SUBTLEX-NL database via the online interface (Keuleers, Brysbaert, & New, 2010; http://crr.ugent.be/isubtlex/). Before recording the stimuli, the onomatopoeic words were matched as carefully and strictly as possible with the control words on a large range of lexical characteristics including word length (i.e., number of graphemes, phonemes, and syllables), frequency (word, lemma, and bigram), dominant lexical class (all taken from SUBTLEX-NL), orthographic and phonological neighborhood density (Coltheart N as taken from Brysbaert, Stevens, Mandera, & Keuleers, in press), age of acquisition, and concreteness (taken from Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014). Separate *t*-tests for each lexical characteristic confirmed that onomatopoeic words and control words did not differ on any of the lexical variables. The 80 fillers were added to keep the proportion of onomatopoeic words presented in the experiment low and to keep the number of nouns and verbs in the experiment similar. As confirmed by the results of a post-experimental questionnaire, this kept the purpose of the experiment unclear to all participants. The 160 Dutch-like nonword stimuli were derived from unused Dutch nouns and verbs that were also taken from SUBTLEX-NL. Nonwords were constructed by Wuggy (Keuleers & Brysbaert, 2010) by changing one to four letters of these existing Dutch words, in line with Dutch orthotactics. Appendix A shows the critical stimuli (onomatopoeic words and control words).

The 320 stimuli were spoken at a normal rate by a female native speaker of Dutch, recorded in a sound proof booth, and digitized at a sample frequency of 44.1 kHz. They were equalized in maximum amplitude using *Praat* software (version 5.2.46; Boersma, 2001). Average stimulus duration was 816 ms (*SD* = 148). There was no difference in duration between onomatopoeic words (*M* = 721 ms, *SD* = 98) and control words (*M* = 723 ms, *SD* = 96).

Rating for Iconicity

To check whether the onomatopoeic words were indeed judged to be more iconic than the control words, 12 native speakers of Dutch (11 female, 18-24 years of age, mean age = 21.1), who did not participate in the main experiment, were asked to rate the 160 word stimuli (i.e. the onomatopoeic words, control words, and fillers) for iconicity. They were seated in a soundproof booth in front of a computer screen. Words were presented to them one at a time via headphones in a randomized way. On the screen they saw a Likert-scale from 1 to 7 where 1 indicated "not iconic at all" (*helemaal niet iconisch* in Dutch) and 7 indicated "very iconic" (*heel erg iconisch* in Dutch). Prior to the start of the rating procedure, instructions explaining the notion *iconicity* (similar to Meteyard et al., 2015) were presented on the screen. Onomatopoeic words (*M* = 4.83,

SD = 0.53) were rated as significantly more iconic than control words *(M* = 2.15, *SD* = 0.77), *t* (11) = 11.04, *p* < .001.

Procedure

After completing informed consent, participants in the main experiment were seated in a comfortable chair at a distance of 100 cm in front of a computer monitor in a shielded, dimly illuminated room. The experiment was programmed using *Presentation* software (Neurobehavioral Systems). Spoken stimuli were presented through EEG-compatible headphones. Before the start of the experiment, written instructions were presented on the screen. Participants were instructed that they would hear spoken stimuli. They were asked to carefully listen to the stimuli and to indicate by pressing a button with the left or right index finger whether the presented item was an existing Dutch word (right finger) or not (left finger). They were asked to make their decisions as quickly and as accurately as possible. Also, they were asked to blink their eyes only when a specific symbol was presented on the screen.

A trial consisted of a fixation cross (200 ms), followed by the spoken stimulus (*M* = 816 ms) paired with a blank screen (2000 ms in total), followed by a symbol (- -) during which participants could blink their eyes (2000 ms). The response deadline was set to 2000 ms. Responses given after this deadline were considered as errors. The experiment consisted of two blocks of 160 stimuli. Between the blocks participants could have a pause for as long as they wanted. Ten test-items with the same characteristics as the stimuli preceded the main experiment as a practice set.

EEG recording and analysis

The electroencephalogram (EEG) was recorded continuously from 59 active electrodes held in place on the scalp by an elastic cap. In addition to the 59 scalp sites, three external electrodes were attached to record EOG, one below the left eye (to monitor for vertical eye movement/blinks), and two on the lateral canthi next to the left and right eye (to monitor for horizontal eye movements/saccades). Finally, one electrode was placed over the left mastoid bone and one over the right mastoid bone. All electrode impedances were kept below 20 KΩ. The continuous EEG was recorded with a sampling rate of 500 Hz, a low cut-off filter of 0.01 Hz and a high cut-off filter of 200 Hz. EEG was filtered offline (high-pass at 0.01 Hz and low-pass at 40 Hz). All electrode sites were online referenced to the electrode placed over the left mastoid and re-referenced offline to the average of the right and left mastoids.

ERPs were calculated by averaging the EEG time-locked to a point 100 ms pre-stimulus onset and lasting until 800 ms after the onset of the stimulus. The 100 ms pre-stimulus period was used as a baseline. Trials defined as errors or outliers in the behavioral analyses, and trials containing ocular or muscular artifacts, were not taken into consideration in the averaging process. After removal of trials containing errors, outliers, and artifacts, 87.1% of trials entered the ERP analyses (86.9% of onomatopoeic trials; 87.2% of control word trials). Separate ERPs were computed for the onomatopoeic words and the control words. An approach to data analysis was applied in which the head surface is divided into four quadrants and a vertical midline column of 9 electrodes each, yielding five regions (left anterior, LA; right anterior, RA; left posterior, LP; and right posterior, RP; and the midline column, Midline; see Peeters, Hagoort, & Özyürek, 2015).

The mean amplitudes of the ERP waveforms for each condition per subject were entered into repeated measures ANOVAs with factors Iconicity (2: onomatopoeic words, control words) and Region (5: LA, RA, LP, RP, Midline). Based on the considerations outlined in the Introduction, by-participant analyses were performed on the N2 component (150-200 ms after word onset), the N400 component (350-550 ms after word onset), and a late positive time-window (600-800 ms after word onset).

Results

Behavioral (RT and error) analyses

Of the total raw dataset, 5.32 % were errors (wrong responses and responses given after the RT deadline) and therefore removed. In addition, RTs outside the range of 2.5 standard deviations above the participant's mean were considered outliers and were excluded from the analyses (2.26 % of all data). Analyses of variance were performed on the mean RTs and mean error rate per experimental condition in the participant analyses (F_1), and on the mean RT and mean error rate per item in the item analyses (F_2). Iconicity (onomatopoeic words vs. control words) was a within-participant factor in the participant analyses and a between-item variable in the item analyses.

Table 1: Mean RTs (in ms) and Mean Error Rates (in percentages) per condition in the experiment. Standard deviations are indicated between brackets.

Condition	Mean RT	Mean Error Rate
Onomatopoeia	1006 (125)	4.68 (6.48)
Control words	1017 (122)	3.95 (4.17)
Fillers	1058 (111)	7.42 (6.67)
Nonwords	1193 (127)	5.28 (3.15)

Overall, onomatopoeic words (*M* = 1006 ms) were responded to numerically faster than control words (*M* = 1017 ms). However, the analysis on the mean RTs showed no significant main effect of Iconicity, F_1 (1, 30) = 2.39, *p* = .132; F_2 (1, 78) < 1. Overall, participants made numerically more errors to onomatopoeic words (*M* = 4.68 %) than to control words (*M* = 3.95 %). However, also in the error analyses no significant main effect of Iconicity was found,

F_1 (1, 30) < 1; F_2 (1, 78) < 1. Thus, no behavioral differences were found between onomatopoeic words and control words in auditory lexical decision. Table 1 shows the mean RT and the mean error rate for each condition.

Electrophysiological analyses

As can be seen in Figure 1, the onomatopoeic words yielded a decrease in amplitude of the N2 component, a less negative-going N400 component, and a late positive deflection compared to the control words.

N2 time-window (150-200 ms). The repeated measures ANOVA in the N2 time-window with factors Iconicity (2) and Region (5) showed a significant main effect of Iconicity, F (1, 30) = 6.51, p = .016, η_p^2 = .178. This effect was not modulated by Region (F < 1).

N400 time-window (350-550 ms). The repeated measures ANOVA showed a significant main effect of Iconicity, F (1, 30) = 5.46, p = .026, η_p^2 = .154. This effect was not modulated by Region (F < 1).

Late time-window (600-800 ms). The repeated measures ANOVA showed a significant main effect of Iconicity, F (1, 30) = 4.43, p = .044, η_p^2 = .129. This effect was not modulated by Region (F < 1).

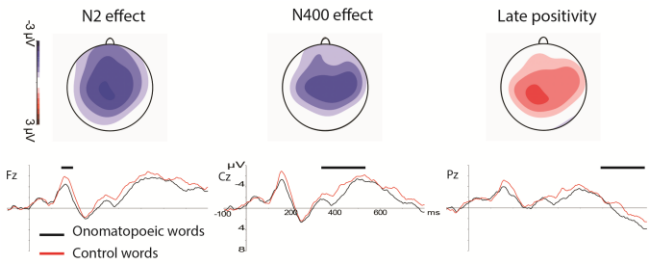


Figure 2: Grand average ERP waveforms time-locked to spoken word onset for a frontal (Fz), a central (Cz), and a parietal (Pz) midline electrode site, reflecting the averaged electrophysiological response to onomatopoeic words (black line) and control words (red line). The three topographic plots show the voltage differences between the two conditions over the scalp for the N2 effect (150-200 ms), the N400 effect (350-550 ms), and the late positivity (600-800 ms).

Discussion

The current study investigated the processing consequences of iconicity at the behavioral and neurophysiological level in spoken word comprehension. It was found that onomatopoeic words were processed as quickly and accurately as words with an arbitrary mapping between form and meaning. Event-related potentials showed a significant decrease in negative amplitude in the N2 and N400 components and a late positivity for onomatopoeic words in comparison to the control words.

The behavioral results are in line with findings from lexical decision tasks in sign language research, which have shown that native signers do not recognize iconic signs at a different speed or accuracy compared to signs with an arbitrary form-meaning mapping (Bosworth & Emmorey, 2010; Klann et al., 2005). Such findings suggested that iconicity does not convey a lexical processing advantage (Bosworth & Emmorey, 2010). However, the time it takes to make a lexical decision by pressing one of two buttons reflects more than lexical processing alone, as it also involves a decision component and the preparation and execution of a motor response. A more direct measure of processing ease may be the amplitude of the N400 component, particularly when words are presented in isolation. Similar to effects of word frequency on the amplitude of the N400 (Allen, Badecker, & Osterhout, 2003; Van Petten & Kutas, 1990), the decrease in negative amplitude for onomatopoeic words compared to control words suggests facilitated access to the lexicon (Lau, Phillips, & Poeppel, 2008) in case of an iconic mapping between form and meaning (cf. Thompson et al., 2009; Vinson et al., 2015).

This presumed ease of mapping form to meaning for onomatopoeic words as reflected in decreased negative amplitude of the N400 component may even be related to the facilitatory effect of higher word frequency in spoken word comprehension, if one takes into account people's everyday sensory experience of perceiving linguistic and non-linguistic (real-world) sounds. In everyday life people not only encounter onomatopoeic words like *beep, click,* and *slurp* in speech, but they also hear the beeps their alarm clock makes, the clicking of their pen, and the slurping sounds of their partner eating soup - sounds that acoustically resemble their onomatopoeic verbal counterpart (Assaneo, Nichols, & Trevisan, 2011). This is not the case for non-iconic words like *throw* and *write* - the word *write* does not acoustically resemble the sounds produced by the process of writing. People's experiences with hearing real-world non-linguistic sounds are (understandably) not taken into account by measures of lexical frequency in psycholinguistic databases, thereby underestimating the frequency with which onomatopoeic sounds (i.e. the sum of verbal and non-verbal experiences) are encountered in everyday life.

But why then did onomatopoeic words not yield faster reaction times than control words? One possible explanation is again related to the fact that onomatopoeic words are so strongly linked to real-world sounds. In an fMRI study, Hashimoto et al. (2006) found that listening to onomatopoeic words recruits brain areas involved in the processing of both verbal sounds and non-verbal sounds. Along similar lines, Kanero et al. (2014) suggested that additional activation in the right posterior superior temporal sulcus (pSTS) for comprehending iconic words compared to non-iconic words reflects that iconic words function as both

linguistic and non-linguistic symbols (see Thierry, Giraud, & Price, 2003). The dual nature of onomatopoeic words, activating both linguistic representations and non-linguistic real-world sound representations, may lead to a conflict when a lexical decision has to be performed on these stimuli. Non-iconic words are arguably more unambiguously identifiable as lexical items than iconic words that have a strong non-lexical component. At the behavioral level, this may cancel out the earlier benefit for onomatopoeic words as indicated by the N400 difference, yielding statistically similar response times for onomatopoeic and control words in lexical decision. Because of its timing and directionality, it could be that the late positivity in the ERPs reflects this enhanced difficulty in making a post-lexical meta-linguistic decision for onomatopoeic words (see e.g. Holcomb, Grainger, & O'Rourke, 2002).

In addition to the N400 effect and the late positivity, an early difference between onomatopoeic words and control words was found in the amplitude of the N2 component. Notably the directionality of this effect was opposite to the directionality of the effect reported by Kovic et al. (2010) in which congruent (compared to incongruent) sound symbolic mappings elicited larger amplitude in the N2 time-window. This difference in directionality of the effect may be due to the difference in stimuli used across the two studies. Kovic et al. (2010) focused on *cross-modal* mappings between an auditorily presented *non*-word label and a subsequently presented visual object. The current study used existing onomatopoeic words, i.e. verbal sounds that refer to non-verbal sounds - all within the auditory modality.

As outlined above, fMRI evidence suggests that hearing an onomatopoeic word activates not only brain areas involved in the processing of verbal sounds, but also (right-hemisphere) areas such as right pSTS involved in processing real-world sounds, outside of the canonical language network (Hashimoto et al., 2006; Kanero et al., 2014). The N2 effect in the current study not only confirms that onomatopoeic words are indeed processed qualitatively differently from non-iconic control words, but also suggests that the patterns of activation for iconic and non-iconic words start to diverge already during early stages of spoken word comprehension. Similar early effects of iconicity have been observed in the visual domain for ideophones (Lockwood & Tuomainen, 2015). These early processing differences may even be enhanced in everyday communication for words that have marked iconic properties such as reduplication and lengthening of syllables through which they stand out from other words (Dingemanse, 2012; Lockwood & Tuomainen, 2015) as naturally occurring auditory oddball stimuli.

To sum up, the current study found that at the behavioral level onomatopoeic words were processed as quickly and accurately as words with an arbitrary mapping between form and meaning. Complementing neuroimaging studies, decreased amplitude of the N2 component of the ERP for onomatopoeic words suggested activation differences between onomatopoeic words and control words during early stages of spoken word comprehension. The link between linguistic and sensory experience in the case of onomatopoeia may lead to facilitated form-meaning mapping, but render lexical decision more difficult in healthy participants. These findings are consistent with the view that the neuronal architecture supporting language comprehension is not an encapsulated entity but interacts with the neuronal infrastructure engaged in auditory perception more broadly.

References

Allen, M., Badecker, W., & Osterhout, L. (2003). Morphological analysis in sentence processing: An ERP study. *Language and Cognitive Processes*, *18*(4), 405-430.

Assaneo, M. F., Nichols, J. I., & Trevisan, M. A. (2011). The anatomy of onomatopoeia. *PloS one*, *6*(12), e28317.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International 5:9/10*, 341-345.

Bosworth, R. G., & Emmorey, K. (2010). Effects of iconicity and semantic relatedness on lexical access in American Sign Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1573-1581.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta psychologica*, *150*, 80-84.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (in press). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. Journal of Experimental Psychology: Human Perception and Performance.

De Saussure, F. (1916). *Course in general linguistics*. New York, USA: McGraw Hill.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, *6*(10), 654-672.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in cognitive sciences*, *19*(10), 603-615.

Grote, K., & Linz, E. (2003). The influence of sign language iconicity on semantic conceptualization. In W. G. Muller & O. Fischer (Eds.), *From sign to signing*. Amsterdam, The Netherlands: John Benjamins.

Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., & Kojima, S. (2006). The neural mechanism associated with the processing of onomatopoeic sounds. *Neuroimage*, *31*(4), 1762-1770.

Hockett, C. D. (1960). The origin of speech. *Scientific American, 203(3)*, 88-96.

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience, 14*(6), 938-950.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical transactions of the Royal Society B: Biological sciences*, *369*, 20130298.
Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*, 54-65.
Kanero, J., Imai, M., Okuda, J., Okada, H., & Matsuda, T. (2014). How sound symbolism is processed in the brain: a study on Japanese mimetic words. *PLoS one, 9*(5), e97905.
Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.
Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*(3), 627-633.
Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior research methods*, *42*(3), 643-650.
Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics, 35,* 379-415.
Klann, J., Kastrau, F., & Huber, W. (2005). Lexical decision with no iconicity effect in German Sign Language: An fMRI-study. *Brain and Language*, *95*(1), 167-169.
Köhler, W. (1929). *Gestalt Psychology*. NY: Liveright.
Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, *114*(1), 19-28.
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920-933.
Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, *6:1246*.
Lockwood, G., & Tuomainen, J. (2015). Ideophones in Japanese modulate the P2 and late positive complex responses. *Frontiers in psychology*, *6:933*.
McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought.* Chicago: University of Chicago Press.
Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*.
Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.
Ormel, E., Hermans, D., Knoors, H., & Verhoeven, L. (2009). The role of sign phonology and iconicity during sign processing: the case of deaf children. *Journal of Deaf Studies and Deaf Education*, *14*(4), 436-448.
Peeters, D., Hagoort, P., & Özyürek, A. (2015). Electrophysiological evidence for the role of shared space in online comprehension of spatial demonstratives. *Cognition*, *136*, 64-84.
Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, *1*:*227*, 1-15.
Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130300.
Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia--a window into perception, thought and language. *Journal of consciousness studies*, *8*(12), 3-34.
Taub, S. F. (2001). *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge: Cambridge University Press.
Thierry, G., Giraud, A. L., & Price, C. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron*, *38*(3), 499-506.
Thompson, R. L. (2011). Iconicity in language processing and acquisition: What signed languages reveal. *Language and Linguistics Compass*, *5*(9), 603-616.
Thompson, R. L., Vinson, D. P., & Vigliocco, G. (2009). The link between form and meaning in American Sign Language: lexical processing effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 550-557.
Thompson, R. L., Vinson, D. P., Woll, B., & Vigliocco, G. (2012). The road to language learning is iconic evidence from british sign language. *Psychological science*, *23*(12), 1443-1448.
Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380-393.
Vinson, D., Thompson, R. L., Skinner, R., & Vigliocco, G. (2015). A faster path between meaning and form? Iconicity facilitates sign recognition and production in British Sign Language. *Journal of Memory and Language*, *82*, 56-85.
Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language*, *93*, 10-19.

Appendix A

Critical stimuli used in the experiment

Onomatopoeic words: babbelen, bonken, brommen, brullen, fluisteren, galmen, giechelen, gorgelen, grommen, hakken, hik, joelen, kievit, klateren, kletteren, klikken, klotsen, knarsen, knetteren, knisperen, knorren, koekoek, kraken, kreukelen, kuchen, kwaken, kwekken, loeien, miauwen, murmelen, piepen, plof, plonzen, rits, ritselen, sissen, slurpen, spatten, tikken, tjilpen.

Control words: afknippen, afprijzen, beheksen, bewonen, blesseren, blinken, broeden, haag, harsen, heg, hinken, inzepen, kneden, knikkeren, kroelen, kruimelen, mand, metselen, omzomen, onweren, optillen, pluizen, prutsen, rafelen, refrein, reiken, slijmen, slijpen, speuren, spieken, staren, tieren, tutten, uitlenen, vastzitten, vijlen, werpen, wiegen, witlof, zuchten.

Language Evolution in the Lab: The Case of Child Learners

Limor Raviv[1] (limor.raviv@mail.huji.ac.il) & Inbal Arnon[2] (inbal.arnon@gmail.com)
[1]Department of Cognitive Science, [2]Department of Psychology, Hebrew University, Jerusalem

Abstract

Recent work suggests that cultural transmission can lead to the emergence of linguistic structure as speakers' weak individual biases become amplified through iterated learning. However, to date, no published study has demonstrated a similar emergence of linguistic structure in children. This gap is problematic given that languages are mainly learned by children and that adults may bring existing linguistic biases to the task. Here, we conduct a large-scale study of iterated language learning in both children and adults, using a novel, child-friendly paradigm. The results show that while children make more mistakes overall, their languages become more learnable and show learnability biases similar to those of adults. Child languages did not show a significant increase in linguistic structure over time, but consistent mappings between meanings and signals did emerge on many occasions, as found with adults. This provides the first demonstration that cultural transmission affects the languages children and adults produce similarly.

Keywords: language evolution; cultural transmission; iterated learning; developmental differences

Introduction

Usage-based theories suggest that the kinds of structures we observe in natural languages arise from general biases and constraints on human capacities, such as learning, memory and processing (Tomasello, 2009). Importantly, these weak individual tendencies can become amplified and fixated over time through the process of *cultural transmission*: the transmission of language over generations through a repeated cycle of use, imitation, observation and induction (Kirby, Griffiths & Smith, 2014). Iterated learning models (ILMs), which simulate the process of cultural transmission of a given behavior over multiple generations, support this theory by showing how the iterative nature of cultural transmission can lead to the creation of strong linguistic universals without the need to assume strong innate biases.

Mathematical and computational simulations of iterated learning show that cultural transmission amplifies weak cognitive biases over time by shaping structural properties to fit agents' existing tendencies and predispositions (Reali & Griffiths, 2009). Moreover, findings from non-linguistic ILM studies with adult participants that examine the transmission of various behaviors (e.g., drawings, whistles, gestures, visual patterns) show that the learned behaviors become significantly more predictable and more structured over generations (Cornish, Smith & Kirby 2013; Verhoef, Kirby & de Boer 2014). In particular, systems become easier to learn, with participants in later generations making considerably fewer mistakes. Notably, systems also become less random and more structured over time, often displaying compositional structure and consistent reuse of smaller building blocks.

Linguistic ILM studies, which are the most relevant to the question of language evolution, show that language learnability increases thanks to an increase in linguistic structure (Kirby, Cornish & Smith, 2008). The languages produced in these studies develop consistent mappings between meanings and signals over time, with similar meanings expressed using similar strings. This is achieved either by the creation of homonyms that mark a shared dimension (e.g., color) but lead to under-specification, or by developing morphological structure, in which different affixes are used to encode different semantic dimensions.

While these findings support the role of cultural transmission in language evolution, they are limited to adult learners. Only one study has used ILM to compare children to adults on a non-linguistic task (Kempe, Gauvrit & Forsyth, 2015), and no published study has looked at the emergence of linguistic structure over time in children. Yet such findings are crucial for making inferences on how learning biases may affect language structure. In particular, adult participants may rely on their extensive and explicit linguistic knowledge when learning an unfamiliar artificial language (Cornish, Tamariz & Kirby, 2009). Consequently, they may have a prior bias in favor of linguistic structure, which (consciously or not) influences their performance, causing structure to emerge. In other words, the structure observed in adult studies may not reflect a cognitive bias *responsible* for the evolution of language over time, but rather a bias that is the *result* of it. If this is true, we cannot draw strong conclusions from these findings on the processes underlying the emergence of linguistic structure in the first place. This criticism can be avoided by looking at children, who have less extensive experience with language (Ramscar & Gitcho, 2007). The lack of evidence from child learners is also problematic because children are the most frequent learners in the actual process of linguistic cultural transmission. Their performance is a test case for the verification of the hypothesis that structure can emerge over time through cultural transmission.

Interestingly, several different predictions can be made regarding children's possible performance in this paradigm: On the one hand, children may be expected to perform like adults, or even better, given their superior language learning abilities in real-world settings (Birdsong, 1999). This prediction is consistent with the claim that children have a special role in the formation of linguistic structure (Bickerton, 1984). Studies with deaf children who are born to hearing parents (and were not exposed to a formal sign language) suggest that children have unique abilities in imposing structure and introducing regularities like word

order that are not found in the gestures of their mothers (Goldin-Meadow & Mylander, 1998). Moreover, research on the developing Nicaraguan Sign Language (NSL) suggests that not only do younger learners reach better fluency, but that NSL has evolved to be more learnable and more grammatically structured in the second generation of child learners (Senghas & Coppola, 2001). This prediction is also supported by the single ILM study that compares children to adults on the same non-linguistic task: Kempe et al. (2015) found that in a visual recall task, children created more identifiable and less complex visual patterns in comparison to adults. They conclude that structure (or less random patterns) emerged more readily in child chains, with children reducing complexity to a level that allowed them to reproduce the patterns as successfully as adults – despite having inferior visual working memory. Taken together, these findings suggest that children may impose more structure compared to adults.

On the other hand, children have less mature cognitive resources (e.g., working memory) and are generally worse in artificial language learning tasks in laboratory settings (Ferman & Karni, 2010; Perry, Axelsson & Horst, 2015), suggesting they might show inferior performance overall. There are also reasons to believe that children differ from adults in their learning and processing skills, which may lead to different biases, different preferences and different trends of learning across development (Arcuili & Simpson, 2011; Hudson-Kam & Newport, 2005). Supporting this claim, artificial language learning studies show clear age-related differences in both learning and generalization: children are more conservative in learning a new structures in comparison to adults (Boyd & Goldberg, 2012), and overgeneralize more than adults (Wonnacott, Brown & Nation, 2013). Importantly, if children are guided by biases that are quantitatively and/or qualitatively different than adults', like overgeneralization and eliminating variation, they may differ in their sensitivity to cultural transmission effects and may exhibit different patterns entirely.

The Current Study

Our goal is to contrast these two predictions by conducting the first large-scale study of iterated language learning in children and adults. We use a novel, child-friendly paradigm that closely resembles previous work with adults. Importantly, we will use the same task with both age groups to enable the comparison between them, as was done in Kempe et al. (2015). We examine the changes in the structure and learnability[1] of the languages produced by children and adults over time, with two questions in minds: (1) Is there an overall difference in performance between children and adults? (2) Will children, like adults, produce more learnable and more structured languages over time? Given that skills like statistical learning, explicit learning, attention and working memory all improve with age, we predict some degree of difference in the overall performance between children and adults. However, we ask whether cultural transmission affects both age groups in a similar way, resulting in similar trends and rate of change. As children may have different biases for regularization, generalization and systemization, linguistic structure may emerge *faster* in one of the age groups.

Importantly, we will directly compare the performance of children and adults by using a more sensitive statistical analysis than used in previous studies. Unlike Kirby et al. (2008), in which the increase in structure and learnability was demonstrated by examining the differences between the first and last generation only, we will examine the course of change over all 10 generations by using regression models. Using regression models has several advantages: we can see changes in linguistic parameters across the course of all generations, rather than just the first and the last; we can examine interactions between our effects of interest, like age group and the effect of time; and we can control for several factors (e.g., gender) and for individual differences using random slopes and intercepts.

Method

The experiment utilizes a diffusion chain paradigm, the most common technique in ILM studies, in which all learners (apart from those in the initial generation) are trained on the output produced by previous learners in the chain. Diffusion chains in this experiment consisted of 10 generations of single participants. Our design is based on Experiment 1 in Kirby et al. (2008) with the following modifications: we used alien figures instead of geometric shapes; the number of items was reduced by half; a human experimenter interactively accompanied the learning; and we used a "syllable bank" instead of free typing. All other conditions were matched to Kirby et al. (2008), including a learning bottleneck, three dimensions of meaning, varying length of words and multiple diffusion chains. In this experiment, there were no limitations on the number of repeated words participants can produce. In Kirby et al. (2008), this design led to a decrease in the number of unique words (improving learnability but creating ambiguities) and to the emergence of linguistic structure, with homonyms assigned according to some semantic dimension.

Participants

90 children (age range: 6.5-12y, mean age: 8:8y, 41 boys and 49 girls) and 40 adults (age range: 21-68y, mean age 33y, 10 men and 30 women), comprising a total of 4 distinct adult chains and 9 distinct child chains. All child participants were visitors at the Bloomfield Science Museum in Jerusalem and were recruited for this study as part of their visit in the Israeli Living Lab in exchange for a small reward. Of the adult participants, half were family members visiting the Living Lab and half were undergraduate students at the Hebrew University, recruited

[1] Thanks to Smith & Kirby, who kindly provided us with their code for these analyses, we were able to use the same algorithms to compute structure and learnability as used in the original paper.

for this study for credit or a small payment. All participants were native literate speakers of Hebrew.

Materials

At the beginning of each diffusion chain, 12 words were randomly drawn from a closed set of 16 nonsense words[2], all of which did not contain or resemble any existing words in Hebrew (as judged by a separate sample of native speakers). These 12 words were then randomly assigned as labels to 12 different items, creating the initial language on which the first participant was trained. We used different types of alien figures, appearing in different colors, either alone or in a group. Thus, items varied along three semantic dimensions: alien type (A, B or C), color (blue or red) and plurality (single or plural). Stimuli included all possible combinations of these three semantic dimensions. Figure 1 below shows the meaning-space structure used in this task, with an example on either side:

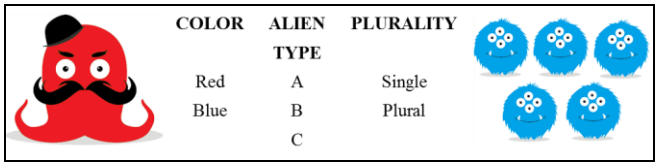


Figure 1: The three semantic dimensions of items in the task

Procedure

Participants were told they are about to learn an alien language that describes many different types of aliens, and that they should try and learn it as best they can. The experiment had three stages: initial exposure, practice and test. Participants were always exposed to a random subset of the target language (SEEN words) during initial exposure and practice, simulating a learning bottleneck. Specifically, participants were trained on just 9 out of 12 words in the language, yet were tested on all items, including UNSEEN words. Note that while adult participants in Kirby et al. (2008) were trained for over 45 minutes, such long sessions are impractical with children. We therefore settled on two sets of exposure to the SEEN words, once during initial exposure (including active verbal production) and once during practice (including active written reconstruction).

The initial exposure phase consists of a random sequence of items from the SEEN subset appearing on the screen together with their label. The experimenter read the label out loud several times and encouraged participants to remember this pairing. Both children and adults were required to reproduce the label aloud before moving on to the next item. During the following practice phase, participants were exposed again to all SEEN items and then had to recreate the labels using a pre-given "syllable bank". Then, participants completed a test phase: they were presented with a series of items without labels, and were required to provide the correct labels according to what they've learned so far, using the same "syllable bank".

Importantly, transmission was implemented in the following way: for each participant, we took the 12 labels produced by him/her during the test and used them as the input language for the next participant in the chain. Specifically, while the first participant was trained on the random initial language drawn by the computer, the second participant learned the language produced by the first participant, and so on for 10 generations of participants.

Results

We examine the performance of children and adults on the same task by looking at two parameters: (1) language learnability, measured by transmission error (normalized Levinstein distances between input and output strings); and (2) linguistic structure, measured with the z-scores produced by a Monte-Carlo algorithm with 1,000 iterations, in respect to the degree of similarities between signals and meanings in a given language (Pearson correlation between form and meaning distances). For a detailed explanation regarding the coding of these parameters, see Kirby et al. (2008).

The transmission error reflects participants' accuracy in reproducing the language, with easier languages eliciting fewer mistakes. Thus, an increase in learnability should be accompanied by a decrease in transmission error. As for structure, the z-score (or structure score) of a given language indicates how likely it is that its structure is created by chance. The higher the z-score is, the smaller the chances that the mapping between words and meanings in this language is random. Thus, an increase in linguistic structure should result in a significant increase in structure score. If the structure score for a given language is higher than 1.96, the language has significantly consistent mappings between words and meanings which is less than 5% likely to have been created by chance.

Language Learnability

Figure 2 shows the changes in mean transmission error as a function of time in both child and adult chains. A decrease in error over generations indicates an increase in language learnability. As can be seen, transmission error generally decreased over time. Following Kirby et al. (2008), we examined the difference in error between the first and final generations. This analysis confirmed a significant reduction in error for both children (mean error at generation 1 =0.75, mean error at generation 10 =0.43, $t(8.2)=3$, $p<0.05$) and adults (mean error at generation 1 =0.69, mean error at generation 10 =0.15, $t(5.1)=9.91$, $p<0.01$).

We used a mixed-effect linear regression model to predict mean transmission error in each generation (Table 1). The fixed effects were gender, generation number, age group and the interaction between the latter two. The model had the maximal random effects structure justified by the data that would converge, including random intercepts for different chains.

[2] All words in the initial and later languages were limited to the same 8 syllables, chosen based on Hebrew phonology: "*šu*", "*gu*", "*di*", "*ki*", "*so*", "*mo*", "*bal*" and "*taz*". We included CVC syllables with open vowels, which are common used in Hebrew.

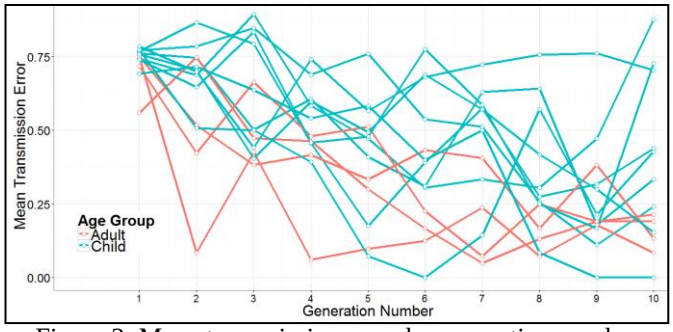


Figure 2: Mean transmission error by generation number and age group.

Table 1: Learnability model

	Estimate	Std. Error	z value	*p*-value
(Intercept)	0.35943	0.0416	8.63893	< .001 ***
Generation Number	**-0.05099**	**0.00933**	**-5.46028**	**< .005 ****
Age Group (Children vs. Adults)	**0.19643**	**0.04973**	**3.94921**	**< .005 ****
Gender (Male vs. Female)	-0.02982	0.03213	-0.92797	> .1
Age Group X Generation Number	0.00549	0.01136	0.48344	> .1

The model showed that generation number has a strong negative effect on transmission error, with errors significantly decreasing over generations (β=-0.05, SE=0.009, t=-5.46, p<0.005). That is, the languages of both children and adults become easier to learn over time. We found a significant difference between children and adults, with the mean transmission error being significantly higher in children (β=0.19, SE=0.04, t=3.94, p<0.005). This suggests that children make more mistakes than adults overall. Crucially, the interaction between age group and generation number was not significant (β=0.005, SE=0.01, t=0.48, p>0.1), so though children are inferior learners, the effect of time on learnability is similar across age groups.

But why are languages becoming more learnable over time? Similar to Kirby et al (2008), the languages of both children and adults in this experiment were characterized by a rapid decrease in the number of distinct words. Here, the number of unique words dropped to as few as only two words in certain chains. Because there were less unique words to memorize overall as chains progress, participants in later generations had smaller chances to make a mistake, which naturally increases learnability. Confirming this claim, lower rates of transmission error were strongly associated with a smaller number of distinct words for both age groups (t(128)=11.2, *r*=0.7, p<0.01)

When participants use this strategy of underspecification, multiple semantic dimensions are encoded using holistic labels, losing much of their informativity. Using just a hand-full of words is functionally useless and indeed varies from natural language, which tend to be expressive. Yet while it is underspecified, such ambiguous languages can still be structured and systematically encode some dimensions of meaning. We now turn to examine language structure.

Language Structure

Figure 3 below shows the changes in structure score as a function of time in both child and adult chains. An increase in structure score over generations reflects an increase in linguistic structure. Dots that fall above the black line represent languages that have consistent and non-random signal-to-meaning mapping at p<0.5.

Importantly, the emergence of consistent mappings between meanings and signals was accomplished by both children and adults multiple times during this experiment, with approximately 20% of languages having more structure than randomly structured languages. Most of these languages were produced in later generations, supporting the role of cultural transmission in the emergence of linguistic structure. Looking at these significantly structured languages confirmed that homonyms were not assigned at random: children and adults created languages in which homonymity was structured along some sematic dimension, closely resembling the result of Kirby et al. (2008).

For example, in one child chain the final language converged to 3 distinct words representing each alien type regardless of color and quantity: all aliens of type A were called "*didi*", all aliens of type B were called "*balgu*" and all aliens of type C were called "*šuki*" (Figure 4). Similar structure emerged in adults' chains. In a different child chain, systematic structure emerged already in generation 8, and was transmitted flawlessly to the last two participants. This language converged to just 2 distinct words representing alien color, regardless of type and quantity: "*ditaz*" for all red aliens and "*balšu*" for all blue aliens.

Like Kirby et al. (2008), we found that adult languages did show a significant difference in structure score between first and final generation (mean structure at generation 0 =0.6, mean score at generation 10 =1.89, t(5.3)=-3.14, p<0.05). Interestingly, children's languages did not show such a change (mean structure at generation 0 =0.54, mean structure at generation 10 =1.05, t(10.7)=-0.78, p=0.45).

We used a mixed-effect linear regression model to predict the structure score in each generation (Table 2), with similar fixed and random effects structure as in the previous model. The model showed a significant difference between children and adults (β=-0.6, SE=0.2, t=-2.3, p<0.05), with adult languages being significantly more structured than those of children. However, though an increase in generation number was associated with higher structure scores, this positive effect was unfortunately not strong enough to reach significance[3] (β=0.1, SE=0.08, t=1.22, p>0.1). There was also no significant interaction between age group and generation number (β=0.02, SE=0.1, t=0.28, p>0.1), indicating that though adults' languages were more structured overall, time affects structure similarly across age groups: both children and adults showed the same non-significant trend of increase in structure over generations.

[3] We believe the source of this null-effect is the Monte-Carlo algorithm's poor approximation in case the input sample is uniformly distributed and/or has little variation, as in the case of languages with a small number of homonyms.

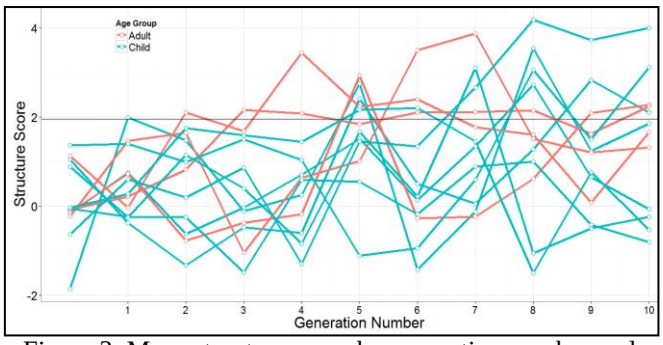


Figure 3: Mean structure score by generation number and age group.

Table 2: Structure model

	Estimate	Std. Error	*t* value	*p* value
(Intercept)	1.261367	0.229747	5.490242	<.001 ***
Generation Number	0.106558	0.08679	1.227778	> .1
Age Group (Children vs. Adults)	**-0.65355**	**0.273864**	**-2.3864**	**< .05 ***
Gender (Male vs. Female)	0.325006	0.220094	1.47667	> .1
Age Group X Generation Number	0.029634	0.105019	0.28218	> .1

	Alien A	Alien B	Alien C	
Red	didi	balgu	šuki	Single
	didi	balgu	šuki	Plural
Blue	didi	balgu	šuki	Single
	didi	balgu	šuki	Plural

Figure 4: A significantly structured child language in generation 10

	Alien A	Alien B	Alien C	
Red	ditaz	ditaz	ditaz	Single
	ditaz	ditaz	ditaz	Plural
Blue	balšu	balšu	balšu	Single
	balšu	balšu	balšu	Plural

Figure 5: A significantly structured child language in generations 8 through 10

Discussion

We found that when underspecification was possible, children and adults behaved similarly: their languages became easier to learn over time at a similar pace and by using the same strategy of reduced words and structured homonymity. Significantly structured languages with non-random signal-to-meanings mapping emerged in child and adult chains on many occasions. However, there was no evidence of significant increase in structure over time for children on either analysis. Importantly, despite adults' overall superiority in this experiment (making less mistakes and creating more structured languages), the effects of cultural transmission on languages' structure and learnability were similar for both age groups. Taken together, this study includes several novel findings: (a) Adults significantly outperform children in this paradigm; (b) Children's languages become significantly more learnable over time in the same manner and pace as adult languages; (c) Children, like adults, can create significantly structured languages. In light of these findings, we can now discuss the questions introduced in the beginning of this paper.

We found that adults were overall better than children on both measured parameters (question 1): they were better learners in general (reflected by lower transmission errors overall) and created more structured languages from the very beginning. This result is in line with previous artificial language learning studies showing that children are inferior learners in laboratory settings despite their optimal acquisition of natural language (Ferman & Karni, 2010). This finding may be driven by developmental differences in key cognitive functions: children have a more limited memory capacity, more problems in sustaining attention, less mature problem-solving strategies and more difficulties in making object-label associations in artificial language tasks, all of which are relevant skills in this paradigm.

With regard to the emergence of more learnable and structured languages (question 2), children and adults showed a somewhat different trend. Language learnability significantly increased over time for both children and adults in a similar fashion, with the same reduction in error as chains progress. Importantly, despite making more mistakes in general, children developed easier and more learnable languages in the same pace as adults and by using the same strategies (i.e., introducing under-specification), suggesting that both age groups are basically guided by the same learnability biases. Similarly, Kempe et al. (2015) reported no differences in the learnability biases of children and adults. Nevertheless, we found no evidence for a significant increase in structure over time: generation number had the same positive (yet *not* significant) effect on structure for both children and adults using mixed-effects models. In other words, while we predicted that children's languages would become more learnable and more structured over repeated iterations, our results support this prediction only partially. Note that if we examine adults' performance using the less-subtle comparison between initial and final generations, our results mirror Kirby et al. (2008): adults did show a significant increase in structure between the first and last generations, while children did not. This discrepancy raises the question of the reliability of comparing only the first and last generations: a different pattern may be seen in previous studies when including information from all 10 generations. Interestingly, both analyses do not align with the predictions drawn from NSL studies and Kempe et al. (2015): children did not have a stronger bias for linguistic structure in comparison to adults.

While a number of child languages did have significant systematic structure with non-random signal-to-meaning mapping, such languages were only a small subset of child languages. There are several possible interpretations of this result: it could indicate that children are less likely to introduce structure during cultural transmission, a finding that is in line with accounts that view adults as the major

agents of linguistic change (Labov, 2007). Alternatively, it could reflect children's difficulty with the additional cognitive demands posed by artificial language learning tasks. In our study, the discovery of the semantic features of items (such as color, type and plurality) was crucial for the emergence of linguistic structure. If children failed to remember all the features of the aliens they saw, this would dramatically reduce their chances of creating non-random languages with consistent mappings between meanings and signals. Since children have trouble in remembering all the features of novel items presented to them (Perry et al., 2015), this could explain why we didn't find an increase in linguistic structure for child chains. Interestingly, the regression analyses revealed parallels between children and adults, with the effects of cultural transmission being similar in pace and magnitude for both age groups.

Finally, the languages that emerged in our study were degenerated in terms of expressivity (as in Kirby et al., 2008). In their paper, a second experiment was conducted where homonyms were filtered out of the language before transmission to the next participant. Under this condition, morphology-like structure emerged in adults. Future work with children should also include a similar experiment in which ambiguities are not allowed to examine whether children also create compositional structure under such conditions. Another option is to introduce a communicative pressure, which serves as the natural equivalent to disfavoring underspecification (Kirby, Tamariz, Cornish & Smith, 2015). Such work is important for evaluating children's ability to create compositional linguistic structure, a crucial feature of natural languages.

Conclusions

In sum, the results of this study suggest that iterated learning models operate in the same way on both children and adults. This finding strengthens the claim that cultural transmission can truly shape languages to be more compatible with learners' limitations and needs. Yet, since children failed to show an increase in structure, more work is required in order to verify this theory.

References

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. Developmental Science, 14(3), 464-473.

Bickerton, D. (1984). The language bioprogram hypothesis. Behavioral and brain sciences, 7(02), 173-188.

Birdsong, D. (Ed.). (1999). Second language acquisition and the critical period hypothesis. Routledge.

Boyd, J., & Goldberg, A. (2012). Young children fail to fully generalize a novel argument structure construction when exposed to the same input as older learners. Journal of Child Language, 39, 457–481.

Cornish, H., Smith, K., & Kirby, S. (2013). Systems from sequences: An iterated learning account of the emergence of systematic structure in a non-linguistic task. In Proc. 35th Annual Conference of the Cognitive Science Society (pp. 340-345).

Cornish, H., Tamariz, M., & Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: What experiments can tell us. Language Learning, 59(s1), 187-205.

Ferman, S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. PloS one, 5(10).

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. Nature, 391(6664), 279-281.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development, 1(2), 151-195.

Kempe, V., Gauvrit, N., & Forsyth, D. (2015). Structure emerges faster during cultural transmission in children than in adults. *Cognition*, *136*, 247-254.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31), 10681-10686.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. Current opinion in neurobiology, 28, 108-114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. Cognition, 141, 87-102.

Labov, W. (2007). Transmission and diffusion. *Language*, 344-387.

Perry, L. K., Axelsson, E. L., & Horst, J. S. (2015). Learning What to Remember: Vocabulary Knowledge and Children's Memory for Object Names and Features. Infant and Child Development.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. Trends in cognitive sciences, 11(7), 274-279.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition, 111(3), 317-328.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. Psychological Science, 12(4), 323-328.

Tomasello, M. (2009). Constructing a language: A usage-based theory of language acquisition. Harvard University Press.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. Journal of Phonetics, 43, 57-68.

Wonnacott, E., Brown, H. E., & Nation, K. (2013). Comparing generalization in children and adults learning an artificial language. Poster from Child language seminar. Manchester, UK.

Outcome or Strategy? A Bayesian Model of Intelligence Attribution

Marta Kryven (mkryven@uwaterloo.ca)
Department of Computer Science, University of Waterloo

Tomer Ullman (tomeru@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

William Cowan (wmcowan@cgl.uwaterloo.ca)
Department of Computer Science
University of Waterloo

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

People have a common-sense notion of intelligence and use it to evaluate decisions and decision-makers. One can attribute intelligence by evaluating the strategy or the outcome of a goal-directed agent. We propose a model of intelligence attribution, based on inverse planning in Partially Observable Markov Decision Processes (POMDPs) in a probabilistic environment, inferring the most likely planning parameters given observed actions. The model explains the agent's decisions by a combination of probabilistic planning, a *softmax* decision noise, prior knowledge about the world and forgetting, estimating the agent's intelligence by a proxy measure of efficiently optimising costs and rewards. Behavioural evidence from two experiments shows that people cluster into those who attribute intelligence to the strategy and those who attribute intelligence to the outcome of the observed actions. People in the strategy cluster attribute more intelligence to decisions that minimise the agent's overall cost, even if the outcome is unlucky. People in the outcome cluster attribute intelligence to the outcome, judging low-cost outcomes as a sign of intelligence even if the outcome is accidental and make neutral judgements before they observe the result. Our model explains human intelligence judgements better than perceptual cues such as the number of revisits or moves.

Keywords: Theory of mind; Intelligence attribution; Social cognition; Bayesian inference; Partially Observable Markov Decision Processes; Inverse planning

Introduction

In everyday life people make fast, intuitive judgements of intelligence. For example, it is more intelligent to rush to catch an infrequent bus than to run towards a subway that departs every 3 minutes, and solving a puzzle from scratch is more intelligent than looking up the answer in a key. Existing qualitative accounts of intelligence explain in general terms why people might describe actions as intelligent or unintelligent. According to Dennett's *rationality principle*, people expect intentional agents to act efficiently in order to achieve their desires, given their beliefs about the world (Dennett, 1989). People may describe behaviour as intelligent if it agrees with their expectations of what the agent should do, and as stupid if the behavior can be explained by a failure of attention, overconfidence or loss of control (Aczel, Palfi, & Kekecs, 2015).

One way to attribute intelligence is to judge the agent's efficiency in achieving a goal, echoing the rationality principle. Efficiency means achieving a rewarding goal at a minimal cost. So, an observer attributing intelligence needs to understand the costs and rewards of a situation and how to plan under uncertainty to maximise the likelihood of success. Evaluating the specific action of an agent seeking a goal requires the observer to ask 'Is this the best goal for that agent?' and 'Is this the best way to achieve that goal?'. The particular end-goal might even be negative, but the behavior could still be conceded as intelligent and thus preferable. Children as young as two understand and value competence, preferring agents who can perform an action on the first attempt, even if the agent is mean (Jara-Ettinger, Tenenbaum, & Schulz, 2015).

However, fully evaluating an agent's planning procedure can be computationally hard, and observing an agent's outcome can provide a shortcut to evaluating its intelligence. In effect, the observer might think 'If they achieved their goal, they must be smart', even though the goal was achieved by sheer luck or prior knowledge. Adults under conditions of cognitive load (Olson et al., 2013) and 7-year old children perceive lucky persons as more likable (Olson, Banaji, Dweck, & Spelke, 2006; Olson, Dunham, Dweck, Spelke, & Banaji, 2008), which suggests that lucky others may also be seen as more intelligent.

In this paper we develop a formal model of intelligence attribution based on Bayesian inverse planning (Baker, Saxe, & Tenenbaum, 2011). We describe two behavioral experiments that validate the model, considering the role of rational expectations and mental short-cuts. The next section gives an informal sketch of the model, followed by formal modelling and behavioral experiments to evaluate the model.

Computational Framework

To describe intelligent behaviour formally, consider a simple context in which it can be observed: a 2-D animation of geometric shapes. Most adults describe the actions of agents in such displays by constructing a narrative about the actors' mental states (Heider & Simmel, 1944). Such 2-D animations were previously used to inform computational models of goal attribution (Baker et al., 2011), preference attribution (Jara-Ettinger, Baker, & Tenenbaum, 2012) and social behaviour (Ullman et al., 2009). Building on these studies we propose a model of intelligence attribution in the context of a rational agent looking for an object in a 2-D world.

Consider an agent (for concreteness, a mouse) looking for a goal (a treat) in a maze. The mouse is familiar with the layout of the maze: it knows which rooms are big and which are small, it knows which rooms are close and easily accessible and which are far away. The mouse knows there is a treat in the maze, but does not know where it is. The treat is equally likely to be anywhere. What should an intelligent mouse do

to find the treat? It should plan efficiently. And when seeing the mouse running through the maze, how does one know if it is intelligent? By comparing its behavior to the behaviour expected of an agent that plans efficiently.

We formalise a rational agent's decision-making process as probabilistic planning in a Partially Observable Markov Decision Processes (POMDP). POMDP assumes that the agent acts sequentially to maximise the reward and minimise the cost of each action, given its beliefs about the world (Figure 1). After each action the agent updates its beliefs based on observations caused by the previously chosen action.

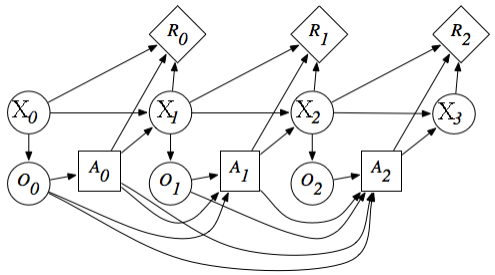


Figure 1: (a) POMDP is a sequential process described by beliefs X_t, observations O_t, rewards R_t and actions A_t at time step t. Arrows indicate a causal relationship.

A viewer observes decisions made by an agent and judges the agent's intelligence by inferring the parameters guiding the agent's planning. The observer estimates the agent's intelligence by a measure of agent efficiency, so that the more optimal agents are judged as more intelligent. We define the agent optimality rank (Table 1) based on how well the corresponding planing strategies optimise the long-term costs and rewards over randomly-generated mazes with a randomly placed goal. Thus, this optimality rank is defined for the model problem and not as a universal cost-minimising strategy. Prior knowledge indicates a non-uniformly distributed prior belief about the goal's location. Since the agent supposedly does not know where the goal is, we use such prior knowledge (with correct information) to represent luck. Decision noise is captured by a *softmax* parameter, set such that actions are chosen with a probability proportional to their reward. To model the observer, we use Bayesian inverse inference on the POMDP by integrating the likelihood of the observed actions with the prior over a set of possible POMDP parameters (Baker et al., 2011).

In principle, this model can describe an evaluation of any behaviour encoded in discrete time and space and can admit a variety of cost functions, reward functions and discount rate models. For concreteness, we limit the planning part of the model to three possible reasons for deviating from an optimal search strategy: prior knowledge about the world, decision noise, and forgetting. While this model is not meant to fully capture the complexities of human intelligence attribution, it provides a computational ideal-observer benchmark to test against experimental data and perceptual-based metrics.

Table 1: Agent optimality rank

Rank	Description
6	Optimal
5	Suboptimal with decision noise or Suboptimal or with prior knowledge
4	Suboptimal with decision noise and prior knowledge
3	Suboptimal with decision noise and forgetting
2	Suboptimal with prior knowledge, forgetting, and decision noise
1	Suboptimal and seemingly random

The Gridworld and POMDP in Detail

Formally, a model world (a maze) is described by discrete time, $0 \leq t \leq T$, and a grid of cells, $W = \{w(i,j)\}$, $w(i,j) \in \{wall, empty, goal\}$. A single cell contains a goal: $\exists(i_g, j_g) \rightarrow w(i_g, j_g) = goal$. The agent acts based on its belief about the world, which is a set of probabilities $X_t = \{P(\mathbf{x_s})_t\}$. Here $X = \{\mathbf{x_s}\}$ is the set of all possible world states such that $\mathbf{x_s}$ represents a world with the goal in cell s. X_0 encodes the set of the agent's initial beliefs.

The agent knows its own location at time t, L_t and sees 180^0 of visual field. The results are observation probabilities $O_t = P(W|X_t, L_t)$. A cell s is visible if the four rays cast from each of the corners of L_t to the corresponding corners of cell s do not intersect walls. If s is visible then $O_t(s) = 1$, otherwise $O_t(s) = 0$. Unseen cells are portrayed as dark, and cells previously seen are patterned (Figure 2). The agent moves one grid cell at a time, choosing among available deterministic actions $A(L_t) = \{a_i\} \in \{N, S, W, E\}$. An action is available if it does not lead into a wall. The agent updates its beliefs contingent on observations using standard Bayesian updating: $X_{t+1} = P(W|A_t, X_t) \propto P(O_t|A_t, X_t)P(A_t|X_t)P(W)$

Calculating Rewards

At time t, the agent calculates the value of each action $Q(a, L_t, X_t)$ and the reward function $R : X_t \times A_t \mapsto \mathbb{R}$ and chooses an action A_t with a likelihood proportional to its reward. Strictly optimal planning in POMDPs is computationally intractable, and in our example the solution is approximated by one step lookahead, or by direct design, on an assumption that the space of beliefs remains unchanged after taking one action (Hauskrecht, 2000):

$$Q(a, L_t, X_t) = \sum_{\mathbf{x_s}} P(\mathbf{x_s})_t \rho(\mathbf{x_s}, L_t + a) + \gamma \max_{a_i \in A(L_t + a)} \{Q(a_i, L_t + a, X_t)\}, \quad (1)$$

where γ is a discount factor, and ρ is proportional to the square of the distance traveled to reach each cell: $\rho(\mathbf{x_s}, L_t + a) \propto \frac{1}{||(L_t+a)-(i,j)||^2}$.

Thus, $\rho(\mathbf{x_s}, L_t + a)$ represents the value of being in a location $L_t + a$ in the world $\mathbf{x_s}$ and $\sum_{\mathbf{x_s}} P(\mathbf{x_s})_t \rho(\mathbf{x_s}, L_t + a)$ is the value of an action a given current beliefs. The second term

in equation (1) describes the discounted value of subsequent actions, assuming actions are chosen optimally.

Finally, the reward function is defined as:

$$R(Q_t(a_i)) = \frac{exp(Q_t(a_i)/\tau)}{\sum_j exp(Q_t(a_j)/\tau)}, \quad (2)$$

where τ is a $softmax$ decision noise. As $\tau \to 0$ the agent deterministically chooses the action with the highest value, and as $\tau \to \infty$ the agent acts at random. For our experiments we either set $\tau = 0$ (optimal) or used a low level of $\tau = 0.01$ such that actions were chosen with a probability proportional to $Q(a_i, L_t, X_t)$ but with most actions optimal (decision noise conditions). A forgetting parameter $0 \leq f \leq 1$ regresses the agent's beliefs toward the mean after each step so that the agent gradually forgets whether a previously observed cell is empty and must re-check already visited locations. Furthermore, prior knowledge about the world is used to to simulate luck: an agent with correct prior knowledge finds the goal sooner.

Experiments

Experiment 1

The first experiment tests whether people attribute intelligence differently to optimal and to suboptimal actions.

Participants. 12 participants recruited from the University of Waterloo, 4 females and 8 males, median age 27.5. Both experiments received ethics clearance from a University of Waterloo Research Ethics Committee and from an MIT Ethics Review Board.

Stimuli. 30 animated movies of a mouse looking for food in a maze were shown in two blocks on a computer screen using Psychtoolbox (Brainard, 1997). The movies were computer-generated by solving a POMDP on one of 9 mazes with two levels of forgetting (on, off), prior knowledge (a correct prior, or a uniformly distributed prior belief) and decision noise (on, off). We varied the appearance of the mouse, the layout of the maze, the textures of the maze and the POMDP parameters on each trial. In every movie the mouse found the treat. The location of the food was counterbalanced so that in one half of the mazes the mouse could find it equally quickly by optimal planning or by luck. In another half of the mazes, the goal was placed so that an optimal agent had to search exhaustively, while a lucky agent with prior knowledge could get finish in fewer steps.

Each movie was labeled according to the most likely POMDP settings estimated by the inverse-planning inference. A movie was labelled *optimal-lucky* if the optimal planner and the prior knowledge planner were equally likely and *optimal-unlucky* if the optimal planner was most likely. Labels *prior knowledge*, *decision noise*, *noise-forgetting*, *noise-forgetting-prior* and *noise-prior* accordingly represented combinations of parameters. A control condition *suboptimal-control* showed an inefficient, highly forgetful and random agent. Each of the conditions occurred four times and the control condition occurred twice. The full set of stimuli used in this and the following experiment can be downloaded from `http://cgl.uwaterloo.ca/~mkryven/`

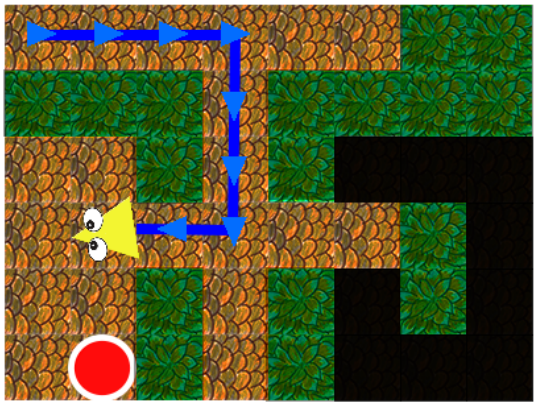


Figure 2: An example of an *optimal-lucky* condition. The agent makes an (optimal) decision to go into the room on the left (reader's perspective). Incidentally, it finds the goal quickly. Dark squares indicate that the agent has not yet seen the area.

Method. Each participant read the instructions on the computer screen and viewed four familiarisation examples followed by the 30 stimuli in two blocks. After viewing each movie the participant recorded his or her rating into a provided Likert answer sheet on a scale from 1 (less intelligent) to 5 (more intelligent).

Results. There was no main effect of block. A two factor ANOVA of rating for *participant* $\times$ *condition* shows a main effect of participant (mean rating 3.44, standard deviation 0.46, $p < .0001$) and of condition ($p < .0001$, Table 2). There was no significant difference between the *optimal-lucky* and *optimal-unlucky* conditions ($p = 0.99$) or between the *optimal-lucky, optimal-unlucky* and *decision noise* conditions ($p = .08$) indicating that participants did not penalise occasional inefficient moves. However, the difference between the *optimal-lucky, optimal-unlucky* and the *prior knowledge* conditions ($p = .001$, difference = 0.8) indicates that participants attributed intelligence to the agent's strategy more than to the agent's outcome.

According to a Scheffe Post-Hoc test, suboptimal conditions were rated differently depending on the cause inferred by the model. Thus, the *decision noise* condition was rated higher than *noise-forgetting* ($p < .0001$, difference = 1) or *noise-forgetting-prior* ($p < .0001$, difference = 2), and *prior knowledge* condition was rated higher than *noise-forgetting* ($p = .0002$, difference = 0.9) or *noise-forgetting-prior* ($p < .0001$, difference = 1.8). The *suboptimal-control* was rated lower than all other conditions ($p < .0001$). Pearson correlations between the agent optimality rank and human rating is .73 and regression of optimality and rating ($r^2 = 53.7\%$),

showing a good fit of our model to the data.

Table 2: Mean ratings and Std. err. of the mean by condition

Condition	Rating	SEM
optimal-lucky	4.7	0.11
optimal-unlucky	4.6	0.11
decision noise	4.1	0.11
prior knowledge	3.9	0.11
noise-prior	3.5	0.11
noise-forgetting	3	0.11
noise-forgetting-prior	2.1	0.11
suboptimal-control	1.6	0.15

The agent's strategy can thus help explain human attributions of intelligence. But can perceptual cues, the path length or the number of cells revisits, explain it equally well? Assuming that a move is scored either for moving from one cell to another or turning, ANOVA for $participant \times moves \times revisits$, shows significant effects of participant ($p < .0001$), moves ($p = .001$) and revisits ($p < .0001$). Table 3 shows mean ratings for different averaged levels of moves and of revisits, where move levels were obtained by splitting the trials into four bins of equal size and calculating the average number of moves per bin. Pearson correlations between the number of moves and ratings, $-.42$, number of revisits and ratings, $-.49$, and multiple regression of moves, revisits and rating ($r^2 = 31.5\%$), calculated over individual trials, show that our model provides a closer fit to the data than the perceptual metrics alone.

In summary, participants did not attribute more intelligence to lucky agents, and judged agents with an efficient decision-making strategy to be more intelligent. The participants attributed the highest intelligence to the approximately rational agents, which either chose optimally or deviated from the optimal decision within a margin explainable by a *softmax* decision noise. Moreover, the inferred cause of sub-optimality matters: random agents are judged least intelligent of all, and forgetful agents as less intelligent than agents with decision noise.

Table 3: Mean ratings and Std. err. of the mean for number of moves and revisits

Moves	Rating	SEM
14	4	0.76
19	3.2	0.14
23	3.1	0.38
31	3.7	0.37

Revisits	Rating	SEM
0	3.9	0.55
2	4.2	0.27
4	3.6	0.18
9	2.2	0.25

Experiment 2

In the second experiment we investigated how do people form a judgement of intelligence. May people decide after observing just one decision, or do they accumulate evidence over time? We used the same mouse in a maze scenario as in Experiment 1, except that on half of the trials the movie stopped after the mouse chose one of the rooms but before the treat was found. In the other half of the trials the movie played until the mouse found the treat.

Participants. 32 participants were recruited via Amazon Mechanical Turk, 2 were discarded for failing to answer questions. The analysis thus included 30 participants (11 females, median age 34).

Stimuli. 40 animated movies were generated by solving POMDP on 8 different mazes. We varied the appearance of the mouse, the layout, textures and orientation of the maze, the location of the goal, and the POMDP settings on each trial. There were 19 complete trials, with each condition (*optimal-unlucky, prior knowledge, decision noise* and *decision noise with prior knowledge*) occurring 4 times and *optimal-lucky* occurring three times. Another 19 trials were incomplete, 12 showing a suboptimal decision (*suboptimal-incomplete* condition) and 7 showing an optimal decision (*optimal-incomplete* condition). Two control trials showed highly forgetful, inefficient mice.

Method. Participants read the instructions on a computer screen in a web browser and viewed 4 familiarisation examples followed by the 40 animated movies shown in two blocks. After viewing each movie, the participant selected a rating from a Likert scale between 1 (less intelligent) to 5 (more intelligent). At the end of the Web survey participants were asked: *How did you make your decision?*

Table 4: Mean ratings and Std. err. of the mean by condition

Condition	Rating	SEM
optimal-lucky	4.5	0.08
prior knowledge	4.3	0.07
optimal-unlucky	4.1	0.07
noise-prior	3.9	0.07
decision noise	3.8	0.07
optimal-incomplete	3.5	0.06
suboptimal-incomplete	3.3	0.05
control	1.2	0.1

Results. There was no main effect of block. A two-factor ANOVA of rating for $participant \times condition$ shows a main effect of participant ($p < .0001$, mean 3.63, standard deviation 0.45) and a main effect of condition ($p < .0001$, Table 4). Main effects of age ($p = .02$) and gender ($p = .0006$) indicate that older participants (difference between groups split by the median age 0.24; median ages 26 and 43) and females (difference 0.3) were more generous. ANOVA of rating over incomplete conditions for $participant \times condition$ shows a small difference between intelligence attributed to optimal and suboptimal decisions (difference = 0.17, $p = .009$).

In agreement with Experiment 1, the *control* condition was

judged as least intelligent ($p < .0001$). However, there was no difference between the *optimal-lucky* ($p = .73$), *optimal-unlucky* ($p = .19$) and the *prior knowledge* conditions. Moreover, *optimal-lucky* and *optimal-unlucky* agents were rated as different ($p = .0.004$, difference 0.47) indicating that online participants judged lucky and efficient agents equally, and optimal agents as more intelligent when they were lucky. Pearson correlation between the agent optimality rank and ratings is .55, regression of optimality and rating $r^2 = 34.6\%$.

The relationship between ratings and perceptual cues was analysed only on full trials, as incomplete trials are always shorter. ANOVA of rating for *participant* $\times$ *moves* $\times$ *revisits* shows significant main effects of participants ($p < .0001$), moves ($p = .009$) and revisits ($p < .0001$). The mean ratings for number of moves and revisits are shown in Table 5. Pearson correlations between the number of moves and ratings, $-.31$, and between number of revisits and ratings, $-.59$. Multiple regression of moves, revisits and rating over individual trials ($r^2 = 36.4\%$), suggests that the ratings in Experiment 2 can be explained by the perceptual cues as well as by the agent's efficiency.

Table 5: Mean ratings and Std. err. of the mean for number of moves and revisits

Moves	Rating	SEM
10	4.4	0.11
20	3.6	0.11
27	3.5	0.11

Revisits	Rating	SEM
0	4.1	0.13
1.5	4.2	0.1
3.5	3.8	0.07
7	2.1	0.09

Did the online participants in Experiment 2 prefer lucky agents? To test the hypothesis that people may use more than one way of attributing intelligence we used a *k-means* (Lloyd, 1982) and a Gaussian Mixture Model with a Bayesian Information Criterion over a 3-dimensional space of correlations of individual ratings with moves, revisits and optimality rank. Two clusters were preferred over three or one. People in the bigger cluster (19 participants) based their rating on the agent's strategy, and people in the smaller cluster relied on perceptual cues.

ANOVA of rating for *participant* $\times$ *condition* over each cluster shows that our model fits the judgements of people in the strategy cluster, but not in the outcome cluster (Table 6). Participants in the Outcome cluster rated *optimal-lucky* agents higher than to the *optimal-unlucky* ones ($p < .0001$, difference = 1.54) and were indifferent between optimal and suboptimal decisions on incomplete trials ($p = .5$). In contrast, people in the Strategy cluster were indifferent between *optimal-lucky* and *optimal-unlucky* agents ($p = .92$) and on incomplete trials rated *optimal-incomplete* agents higher than the *suboptimal-incomplete*($p = .00002$, difference 0.32).

The discrepancy between model and data reveals two different styles of attributing intelligence: attributing intelligence to efficient strategies, or to shorter paths (Figure 3 and

Table 6: Correlations of ratings with perceptual features and with optimality over invividual trials

Metric	Strategy Cluster	Outcome Cluster
Pearson, number of revisits	-.50	-.73
Pearson, number of moves	-.12	-.62
Pearson, Optimality	.65	.40
r^2, revisits, moves	35%	55%
r^2, optimality	44.3%	21.8%

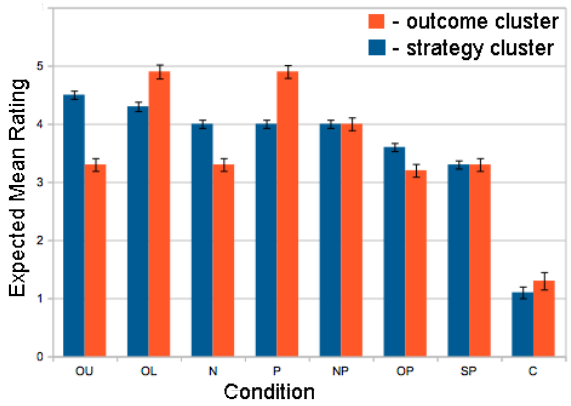


Figure 3: Comparing ratings between the two clusters *OL-optimal-lucky*, *OU-optimal-unlucky*, *N-decision noise*, *P-prior knowledge*, *NP-noise-prior*, *OP-optimal-incomplete*, *SP-suboptimal-incomplete*, *C-control*.

Table 7). Indeed, the participants' answers to the verbal question support this conclusion. We divided the participant answers to *'How did you make your decision?'* into two groups: **outcome** and **strategy**. For example, we coded a response as **strategy** if it said: *'Based on if the mouse checked every nook and cranny.'* and as **outcome** if they said *'Based on how long it took for the rat to find the treat'*. Two independent raters agreed on 27 out of 30 participants, coding 8 of them as **outcome** and 19 as **strategy**. The remaining 3 were randomly assigned to either group. Importantly, the 8 participants independently agreed on as **outcome** by both raters (using verbal measures) were also the 8 participants identified as belonging to the outcome cluster using K-Means clustering.

Discussion

Our model proposes a formal account of intelligence attribution and an apparatus for generating quantitative predictions. The model predicts that people perceive some suboptimal decisions as more intelligent than others, depending on the inferred causes of suboptimal planning, in agreement with behavioural data.

Participants were clustered into two groups: those in the Strategy group attribute intelligence based on partial results, while those in the Outcome group do not decide until they

Table 7: Mean ratings of conditions and Std. err. of the mean of people in the Strategy (S) and the Outcome (O) clusters

Condition	S	SEM	O	SEM
optimal-unlucky	4.5	0.08	3.3	0.11
optimal-lucky	4.3	0.07	4.9	0.12
decision noise	4	0.07	3.3	0.11
prior knowledge	4	0.07	4.9	0.11
noise-prior	4	0.07	4	0.11
optimal-incomplete	3.6	0.04	3.2	0.006
suboptimal-incomplete	3.3	0.04	3.3	0.006
control	1.1	0.1	1.3	0.14

have seen the consequences[1]. Both groups, however, see random actions as less intelligent than those that can be explained by causal inference.

Why did people differ in intelligence attribution? Our experiments used simple mazes that most adults can easily solve. While people in the Strategy cluster expected the agent to do what a human would do, people in the Outcome cluster did not. According to Gardner's theory of multiple intelligences (Gardner, 2011) as skills in different domains people in the Outcome cluster may attribute the agent's luck to an invisible skill (a sense of smell). Alternatively, people may assume that lucky agents must be intelligent based on a belief in a just world (Lerner, 1980). The former implies that people themselves should act rationally when solving a maze, and the latter that people should decide randomly. We plan to address this question in future work.

Another interesting avenue for future work is to investigate how children attribute intelligence. Although children recognise and value competence (Jara-Ettinger et al., 2015), children also prefer lucky agents (Olson et al., 2006, 2008). To our knowledge there are no studies evaluating children's attribution of intelligence, and at what age the abilities to reason about outcome vs. strategy emerge.

One possible criticism of our specific implementation is that it encodes space as a grid of squares and makes a new decision each time the agent moves into a new square, which may not be an accurate representation of how people navigate. An alternative – encoding the maze as a weighted graph where each room is represented by a vertex – may be a better approximation of how people represent space. In addition, more fine-grained rationalistic explanations in terms of variable costs and reward functions may be a better causal model for actions currently attributed to decision noise.

The goal of current research in Artificial Intelligence (AI) is to create intelligent computer applications, such as self-driving cars, automatic trading and intelligent energy-saving appliances designed to take over routine human decision-making. Such applications must be not only be algorithmically correct, but also must interact with people in a way humans understand as intelligent. Thus, to make better AI applications we need to measure intelligence in technical terms, and our model takes a step in that direction.

[1] As Solon advised to Croesus, *'Count no man happy until he is dead'*.

Acknowledgements

We thank the reviewers for valuable comments and feedback, which helped us make this manuscript better. This work was supported by the Center for Minds, Brains and Machines (CBMM), under NSF STC award CCF-1231216.

References

Aczel, B., Palfi, B., & Kekecs, Z. (2015). What is stupid? people's conception of unintelligent behavior. *Intelligence*, *53*, 51–58.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the thirty-second annual conference of the cognitive science society* (pp. 2469–2474).

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Gardner, H. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.

Hauskrecht, M. (2000). Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 33–94.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 243–259.

Jara-Ettinger, J., Baker, C. L., & Tenenbaum, J. B. (2012). Learning what is where from social observations. In *Proceedings of the thirty-fourth annual conference of the cognitive science society* (pp. 515–520).

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent. toddlers inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640.

Lerner, M. (1980). *The belief in a just world; a fundamental delusion*. New york: Plenum Press.

Lloyd, S. P. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, *28*(2), 129–137.

Olson, K. R., Banaji, M. R., Dweck, C. S., & Spelke, E. S. (2006). Children's biased evaluations of lucky versus unlucky people and their social groups. *Psychological Science*, *17*(10), 845–846.

Olson, K. R., Dunham, Y., Dweck, C. S., Spelke, E. S., & Banaji, M. R. (2008). Judgments of the lucky across development and culture. *Journal of Personality and Social Psychology*, *94*(5), 757.

Olson, K. R., Heberlein, A. S., Kensinger, E., Burrows, C., Dweck, C. S., Spelke, E. S., & Banaji, M. R. (2013). The role of forgetting in undermining good intentions. *PLoS ONE*.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

Communicating generalizations about events

Michael Henry Tessler (mtessler@stanford.edu) and **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University

Abstract

Habitual sentences (e.g. *Bill smokes.*) generalize an event over time, but how do you know when a habitual sentence is true? We develop a computational model and use this to guide experiments into the truth conditions of habitual language. In Expts. 1 & 2, we measure participants' prior expectations about the frequency with which an event occurs and validate the predictions of the model for when a habitual sentence is acceptable. In Expt. 3, we show that habituals are sensitive to top-down moderators of expected frequency: It is the expectation of future tendency that matters for habitual language. This work provides the mathematical glue between our intuitive theories' of others and events and the language we use to talk about them. **Keywords:** events; generics; pragmatics; Bayesian data analysis; Bayesian cognitive model

Figuring out that a person or thing tends to exhibit a behavior can be crucial and hard-won knowledge. It is not surprising then that language has ways to convey this information about propensity: *Bill smokes*, *That man steals from children*, *My car doesn't start*. These are called *habitual sentences*. Like many other linguistic means of conveying generalizations, the truth-conditions of habituals are extremely flexible: If Bill smoked three cigarettes last month, can you say that *Bill smokes*? If he smoked a pack last week, but just swore a solemn oath to quit? In this paper we present the first empirical data on the felicity of habitual sentences and describe a formal model of habitual interpretation, based on recent Bayesian models of the pragmatics of vague language.

Linguists have pointed out the parallel between habituals like *Bill smokes* and generic sentences like *Swans are white* (Carlson, 1977, 2005; Cohen, 1999). Both convey generalizations (about events and categories, respectively), and both exhibit dramatic flexibility in their truth conditions: *Swans are white* even though there are black swans, and it may the case that *Bill smokes* even if he goes without a cigarette for an entire family vacation. Indeed, cases like *Mosquitos carry malaria* (wherein only a tiny percentage have the property) seem to parallel habitual sentences of rare actions like *Susan writes novels*. Susan may only have written 3 novels in her life, but still this seems like a valid generalization to convey.

In this paper, we take the analogy between generic and habitual language seriously by elaborating a recent computational theory of generic language to derive predictions for the truth conditions of habitual sentences. The theory of Tessler and Goodman (under review) posits that the semantics of a generic statement is an uncertain threshold on the degree of property prevalence (i.e. how many instances of the kind have the feature) and derives context-sensitive meanings through pragmatic inference given the distribution of property prevalence (i.e. in general, what prevalences are likely across different categories). We adapt this theory to explain habituals by adjusting the underlying degree to be the *propensity to take part in an event* (e.g. how often does a person tend to do an action) and derive predictions for felicity judgments of a range of habituals. In Expt. 1, we measure *a priori* beliefs about how frequently people do a diverse set of actions. In Expt. 2, we use those priors and the pragmatic theory to make predictions about the truth conditions of habitual sentences under different frequencies of action.

If habituals (and generics) are truly language for conveying generalizations, it would be useful for them to reflect expectations, not merely observations. Is the underlying dimension for habituals the actual, past frequency of the event (e.g. how often Bill has actually smoked in the past week) or the predictive probability that this event will occur in the near future (e.g. how likely it is that Bill will smoke next week)? In Expts. 3a & 3b, we show that the object of communication is a speaker's prediction about the future frequency of action, and not past frequency. This has important implications about the relationship between linguistic generalizations and intuitive theories, which we explore briefly in the discussion.

Computational model

A habitual sentence expresses a generalization about the tendency of an individual to participate in a kind of event. For a given individual (e.g. BILL) and an event or behavior (e.g. SMOKING), we refer to the rate at which the individual participates in the event for a given time window as the *propensity*, denoted λ. A natural denotation for the habitual is a simple threshold on propensity $\lambda > \tau$ (c.f. Cohen, 1999), yet no fixed value for τ would lead to the observed flexibility of truth conditions (e.g. *Bill smokes* vs. *writes novels*). Building on Lassiter and Goodman (2013, 2015), we posit that this threshold is not a fixed property of the language, but is established by pragmatic inference.

We model a speaker S_2 who reasons about a pragmatic listener L_1; this listener is considering the propensity of a certain behavior of an individual. The listener L_1 has uncertainty about the appropriate threshold for the habitual in this context ($\tau \sim$ Uniform over possible frequencies), and reasons about what an informative speaker S_1 would be likely to say. The hypothetical speaker S_1 in turn reasons about an idealized literal listener L_0, who has access to the threshold τ (i.e. S_1 believes L_0 will interpret him in exactly the way he means). Writing the propensity as λ, this leads to a set of equations:

$$P_{S_2}(u \mid \lambda) \propto \exp\left(\alpha_2 \cdot \ln \int_\tau P_{L_1}(\lambda, \tau \mid u)\, d\tau\right) \quad (1)$$

$$P_{L_1}(\lambda, \tau \mid u) \propto P_{S_1}(u \mid \lambda, \tau) \cdot P(\lambda) \cdot P(\tau) \quad (2)$$

$$P_{S_1}(u \mid \lambda, \tau) \propto \exp\left(\alpha_1 \cdot \ln P_{L_0}(\lambda \mid u, \tau)\right) \quad (3)$$

$$P_{L_0}(\lambda \mid u, \tau) \propto \delta_{[\![u]\!](\lambda,\tau)} P(\lambda). \quad (4)$$

We take the speakers S_1 and S_2 to consider two utterances: the habitual, with $[\![u]\!](\lambda,\tau) := \lambda > \tau$, or nothing (staying silent), with $[\![u]\!](\lambda,\tau) :=$ True, and to select utterances softmax optimally, with a degree of optimality governed by α_1 and α_2, respectively. Equation 1 can then be interpreted as a model of felicity or truth judgments (Degen & Goodman, 2014; Tessler & Goodman, under review). The speaker will choose to produce the habitual when the true propensity λ is more likely under L_1's posterior given the habitual than under her prior (implied by S_2 "staying silent"). A fully implemented version of the model can be found at `http://forestdb.org/models/habituals-cogsci2016.html`.

The prior, $P(\lambda)$ in Eqs. 2 and 4, specifies prior beliefs about the propensity of a specific event or behavior (e.g. SMOKING) across a set of different individuals. *A priori*, it is unclear how far to extend this *contrast set*: Does it include beliefs about *all* people, or just individuals within a predefined subclass e.g. individuals of the same gender or the same age? This is important because the meaning of the habitual (the threshold τ) is derived with respect to the prior. If the priors differed by gender (e.g., in the propensity to WEAR A BRA) and if language interpreters took the prior to be with respect to a particular gender, the model would predict differences in the truth conditions by gender (e.g., in *Susan wears a bra.* vs. *Bill wears a bra.*). We explore this issue in Expts. 1 & 2.

Experiment 1: Prior elicitation

In this experiment we elicit the prior $P(\lambda)$ for different events in order to generate model predictions for corresponding habituals. Given that some individuals rarely or never engage in an event, while others do quite frequently, we would expect the prior to be a mixture distribution between (at least) these two possibilities, similar in spirit to Zero-inflated or Hurdle Models of epidemiological data (Rose, Martin, Wannemuehler, & Plikaytis, 2006). Indeed, there may be more than these two possibilities, corresponding to individuals with different traits or demographics (e.g., different expected frequencies depending on age or gender).

Methods

Participants We recruited 40 participants from Amazon's Mechanical Turk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 12 minutes and participants were compensated $1.25 for their work.

Materials We created thirty-one events organized into pairs or triplets from 5 different conceptual categories: food and drug (e.g. *eats caviar*, *eats peanut butter*), work (e.g. *sells things on eBay*, *sells companies*), clothing (e.g. *wears a suit*, *wears a bra*), entertainment (e.g. *watches professional football*, *watches space launches*) and hobbies (e.g. *runs*, *hikes*). Items were chosen to intuitively cover a range of likely frequencies of action, as well as to provide a minimal comparison to another item by having a common superordinate action (e.g. *eating* caviar vs. peanut butter).

Procedure For each event, participants were asked two questions, with associated dependent measures:

1. "How many {men, women} have DONE ACTION before?"
   Participants responded "N out of every J." by entering a number for N and choosing J from a drop-down menu (options: {1000 - 10 million}; default: 1000).
2. "For a typical {man, woman} who has DONE ACTION before, how frequently does he or she DO ACTION?"
   Participants responded "M times in K." by entering a number for M and choosing K from a drop-down menu (options: {week, month, year, 5 years}; default: year).

For example, one set read: "How many men have smoked cigarettes before?"; "For a typical man who has smoked cigarettes before, how frequently does he smoke cigarettes?"

We anticipated there might be different beliefs about the frequency of events depending on whether the actor is male or female, so we asked about both genders. Participants answered both questions for each gender on each slide (4 questions total per slide, order of male / female randomized between-subjects), and every participant completed all 31 items in random order. The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/priors/priors-2.html`.

Data analysis and results

We built a Bayesian data analysis model for this prior elicitation task. Question 1 elicits the proportion of people who have done an action before. We model this data as coming from a Beta distribution: $d_1 \sim \text{Beta}(\gamma_1, \xi_1)$. Question 2 elicits the rate, or relative frequency, with which a person does the action. This was modeled by a log-normal distribution: $\ln d_2 \sim \text{Gaussian}(\mu_2, \sigma_2)$. Each item was modeled independently for each gender. We implemented this model using the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014), and found the credible values of the parameters by running MCMC for 100,000 iterations, discarding the first 50,000 for burnin.

The priors elicited cover a range of possible parameter values as intended (Figure 1, scatter), resulting in parametrized distributions of dramatically different shapes (insets). We observe a correlation in our items between the mean % of Americans who have DONE ACTION before (Question 1) and the mean log-frequency of action (Question 2) ($r_{1,2} = 0.74$). Items that tend to be more popular actions also tend to be more frequent actions (e.g. *wears socks*) and visa-versa (e.g. *steals cars*), though there are notable exceptions (e.g. *plays the banjo* is not popular but done frequently when done at all, as is *smokes cigarettes*; *goes to the movies* is a popular activity though not done very often). This diversity is relevant because the speaker model (Eq. 1) will produce habitual sentences (e.g. *Sam goes to the movies vs. the ballet.*) contingent on the shape of the prior distribution.

From the inferred parameters and assumed functional forms, we get an inferred $P(\lambda)$ modeled as a mixture of individuals with the possibility of carrying out the action and those without the possibility of doing it. That is, $P(\lambda)$ was

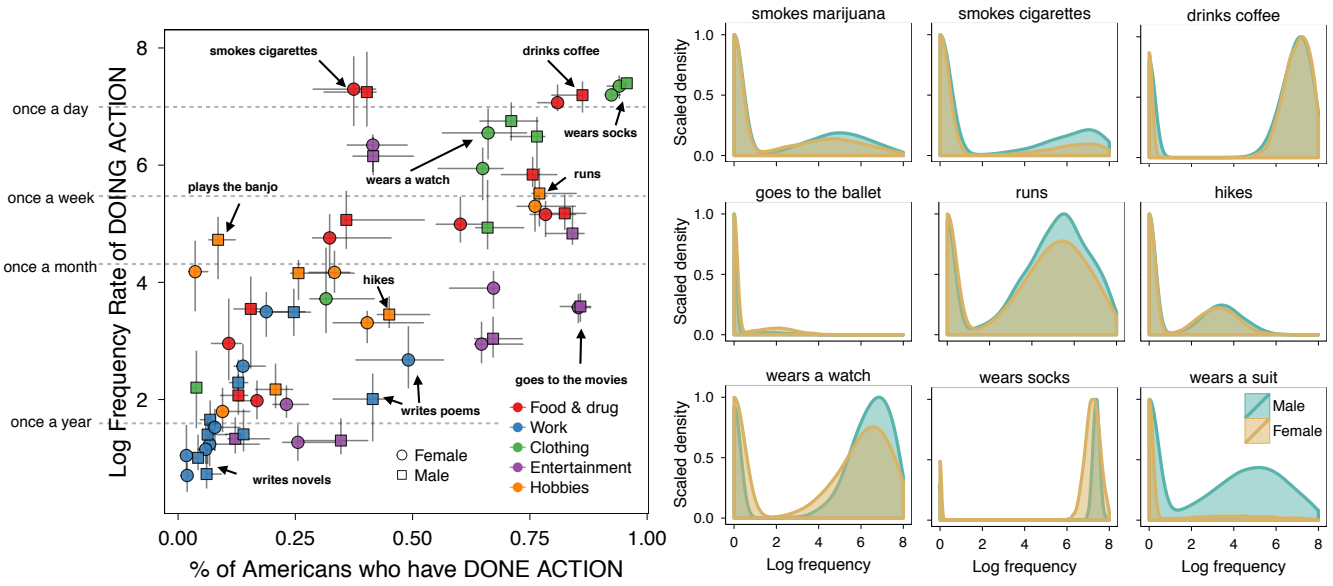


Figure 1: Frequency prior distributions empirically elicited for thirty-one events for both male and female genders. Left: Prior distributions are summarized by θ — the proportion of people who have done the action before — and γ — the mean log frequency of doing the action (for people who had done it before). Error bars denote 95% Bayesian credible intervals. Right: Density plots display posterior predictive distributions on frequency using the structured Bayesian model in Eq. 5. Log frequency scale is on the order of 5 years. Approximate rates of once per: {year $\sim$ 1.5; month $\sim$ 4; week $\sim$ 5.5; day $\sim$ 7}.

constructed by sampling λ as follows:

$$\theta \sim \text{Beta}(\gamma_1, \xi_1)$$
$$\ln\lambda \sim \begin{cases} \text{Gaussian}(\mu_2, \sigma_2) & \text{if Bernoulli}(\theta) = \text{T} \\ \delta_{\lambda=-\infty} & \text{if Bernoulli}(\theta) = \text{F} \end{cases} \quad (5)$$

In addition to specifying the correct way to combine our two prior-elicitation questions, using this inferred prior in our language model resolves two technical difficulties: (1) It smooths effects that are clearly results of the response format[1] and (2) it better captures the tails of the prior distribution which have relatively little data and need to be regularized by the analysis. Figure 1 (right) shows example inferred priors.

Some items show substantial differences between the genders (e.g. *wears a bra*) and some show subtle differences (e.g. *watches professional football*). We will explore the possibility of different truth conditions for habituals of different gendered characters in Experiment 2, for select items with priors that differ substantially by gender.

Experiment 2: Felicity judgments

A present-tense habitual sentence is of the form SINGULAR NOUN PHRASE + PRESENT TENSE SIMPLE VERB PHRASE (e.g. *Bill smokes cigarettes.*). We next explore the endorsements of habituals of this form made from the items whose propensity priors were measured in Experiment 1.

[1]For example, a very common rating is *1 time per year*. Presumably participants would be just as happy reporting *approximately* 1 time per year; the raw data does not reflect this due to demands of the dependent measure.

Methods

Participants We recruited 150 participants from MTurk. To arrive at this number, we performed a Bayesian precision analysis to determine the minimum sample size necessary to reliably ensure 95% posterior credible intervals no larger than 0.3 for a parameter whose true value is 0.5 and for which the data is a 2 alternative forced choice. This analysis revealed a minimum sample size of 50 per item; since participants only completed about one third of the items, we recruited 150 participants. The experiment took 4 minutes on average and participants were compensated $0.55 for their work.

Procedure and materials On each trial, participants were presented with a *past frequency statement* for a given event of the form: "In the past M {weeks, months, years}, PERSON DID X 3 times". For example, *In the past month, Bill smoked cigarettes 3 times*. The particular intervals used (number M and window {weeks, months, years}) were selected after examining the predictions of the speaker model (Eq. 1), for each item independently, to yield a variety of predicted endorsement rates. The items were the same as in Expt. 1.

Participants were asked whether they agreed or disagreed with the corresponding habitual sentence: "PERSON DOES X" (e.g. *Bill smokes cigarettes*). Participants saw 25 out of the 31 items paired randomly with a male or female character name; the other 6 trials were presented with both male and female names (on separate trials; 37 trials total) to explore the nature of the contrast class (see Model section). The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/truth-judgments/tj-2.html`.

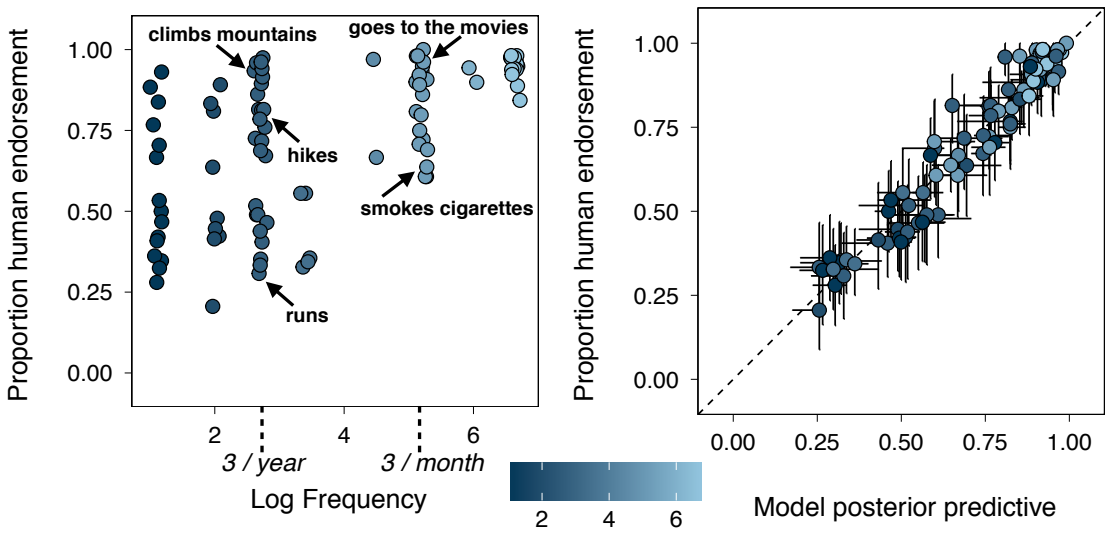


Figure 2: Human acceptability judgments as a function of the log frequency of action (left) and speaker S_2 model predictions (right) for ninety-three unique items (event x frequency). Color denotes target-individual frequency of action (log scale), with lighter colors indicating more frequent actions. Actual frequency noted on x-axis for examples (left). Error bars correspond to 95% bootstrapped confidence intervals for the participant data and 95% Bayesian credible intervals for the model predictions. Error bars suppressed and points jittered on left facet for visual clarity.

Results

Behavioral results On each trial of the experiment, the participant was told a person did a particular action 3 times during some time window. Figure 2 (left) shows the correspondence between the frequency of the event (normalizing to a 5-year time scale and taking the logarithm) and the felicity of the corresponding habitual sentence. It is clear that a habitual sentence can receive strong agreement even when the actions are very infrequent (log frequency $\sim$ 1; 3 times in a 5-year interval; e.g. *writes novels*, *steals cars*). We also see even when actions are done relatively frequently (e.g. 3 times in a one month interval; log frequency $\sim$ 5), there are habitual sentences participants are reluctant to endorse completely (e.g. *wears socks*, *drinks coffee*). In our data, actions completed with a high frequency (3 times in a one week interval; log frequency $\sim$ 6.5) receive at least 75% endorsement, though there is still variability among them (e.g. between 10-25% of people disagree with *wears a watch* and *wears a bra*). Overall, frequency of action predicts only a fraction of the variability in responses ($r^2(93) = 0.33$). For actions that are done on the time scale of years or longer (lower median of frequency), frequency itself no longer explains the endorsements ($r^2(50) = 0.07$)

We further examined the six items for which we observed gender differences in the prior elicitation task (Expt. 1). We find no differences between endorsements of the habitual of characters with male and female names, and overall, the mean endorsements by gender are strongly correlated $r(93) = 0.91$. This may be because the felicity of habitual sentences depends on a comparison to individuals of both genders (i.e, the *contrast class* is other people; not just other men or women). Less interestingly, the lack of a difference may be the result of gender being not very salient in our paradigm, perhaps because the names used were not sufficiently gendered.

Model fit and results We used the pragmatic speaker model S_2 (Eq. 1) with the priors elicited above (Expt. 1) to predict felicity judgments in Expt. 2, assuming the target propensity (to be conveyed by S_2) is the provided frequency. Because we observe no difference between the felicity judgments for habituals of male and female characters, we use a 50% mixture of the inferred priors for each gender to construct a single frequency distribution $P(\lambda)$ across individuals. The model has two free parameters—the speaker optimality parameters, α_i, in Eqs. 1 & 3. We use Bayesian data analytic techniques to integrate over these parameters (Lee & Wagenmakers, 2014), comparing the posterior predictive distribution to the empirical data in Expt. 2. To construct the posterior predictive distribution over responses, we collected 2 MCMC chains of 100,000 iterations, discarding the first 50,000 iterations for burn in. The Maximum A-Posteriori value and 95% highest probability density interval for α_1 is 19.3 [14.9,19.9] and α_2 is 1.5 [1.4,1.6].

As shown in Figure 2, right, the probabilistic pragmatics model does a good job of accounting for the variability in responses ($r^2(93) = 0.94$), including actions done on the time scale of years or more ($r^2(50) = 0.92$). The model decides when the habitual is a useful way to describe the person's behavior, assuming that what the person did in the past is representative. This raises an interesting question: Does the propensity communicated by the habitual indicate an objective, past frequency or a subjective, future expectation?

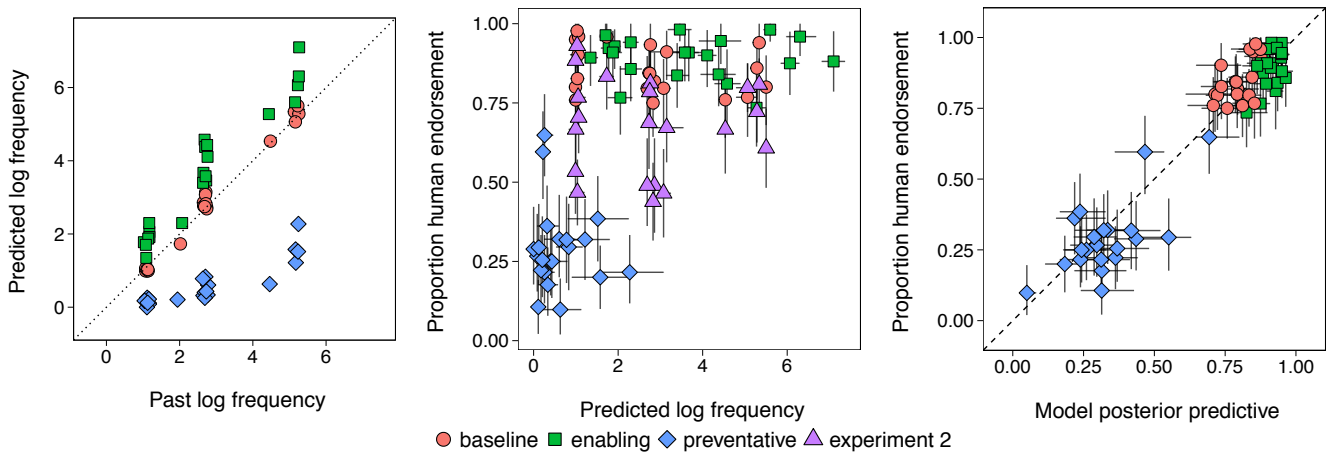


Figure 3: Left: Predicted log frequency as a function of past log frequency given to the participant (Expt. 3a; CIs suppressed and jitter added for visual clarity). Middle: Human endorsements of habitual sentences (Expt. 3b) vs. Predicted log frequency (Expt. 3a), with data for corresponding items from Expt. 2 (assumed to have the same predictive log frequency as baseline). Right: Endorsements (Expt. 3b) vs. Speaker S_2 model predictions using empirically elicited predictive frequencies (Expt. 3a).

Experiment 3: Objective frequency versus subjective expectation

While past frequency is often a good indicator of future tendency, people can change abruptly due to a variety of decisions and outside events. Does habitual language communicate propensity in terms of past frequency or future expectations? In this set of experiments, we address this by introducing causal factors that enable or prevent future actions (e.g. buying cigarettes; developing an allergy). In Expt. 3a, we measure *predictive frequency* when past frequency alone is observed and when these causal factors are introduced. In Expt. 3b, we examine felicity judgments of the habitual sentence (e.g. *John smokes cigarettes.*; *John eats peanut butter.*) in the presence of these causal modifiers. This will allow us to test whether habituals are best explained by a speaker S_2 who communicates the known past or expected future frequency.

Experiment 3a: Prediction elicitation

Methods We recruited 120 participants from MTurk, using the same criterion as Expt. 2. The experiment took 4 minutes on average and participants were compensated $0.40.

The procedure was identical to Expt. 2 except for the inclusion of a second sentence on a subset of trials and the use of a different dependent measure. On all trials, participants were presented with a *past frequency sentence* (see Expt. 2). Additionally, on one third of the trials, participants were presented with a **preventative sentence** (e.g. *Yesterday, Bill quit smoking.*). On one third of the trials, participants were presented with an **enabling sentence** (*Yesterday, Bill bought a pack of cigarettes.*) The final third of trials had no additional evidence and were identical to Expt. 2.

Only twenty-one of the original thirty-one items were used in order to shorten the experiment. To increase expected variability, participants saw only the frequencies that led to most intermediate endorsement of the habitual in Expt. 2. In addition, we did not include separate trials for both male and female names for the select items we did in Expt. 2, since we saw no differences in their endorsements of the habitual.

Participants were asked "In the next TIME WINDOW, how many times do you think PERSON does EVENT?", where the TIME WINDOW was the same as given in the *past frequency statement*. The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/priors/predictive-1.html`.

Behavioral results

Figure 3 (left) shows the predicted future frequency as a function of the past frequency given to the participant and the type of causal information given. We observe in the baseline condition that future frequency perfectly tracks past frequency (e.g. participants believe if a person smoked cigarettes 3 times last month, they will smoke cigarettes 3 times next month). This means that our model makes identical predictions for Expt. 2 whether the target is past frequency or expected future frequency (indicating, as expected, that we must look to the new data to distinguish these models). Critically, we observe the preventative information appreciably decreasing and the enabling information slightly increasing predicted frequency (Figure 3 left; blue and green dots).

Experiment 3b: Felicity judgments

Methods We recruited 150 participants from MTurk, using the same criterion as Expt. 2. The experiment took 4 minutes on average on participants were compensated $0.40 for their work. None of the participants had completed Expt. 3a. The only difference from Expt. 3a is the dependent measure. On each trial, participants were asked if they agreed or disagreed with the corresponding habitual sentence (as in Expt. 2). The experiment in full can be viewed at `http://stanford.edu/~mtessler/habituals/experiments/truth-judgments/tj-3-preventative.html`.

Results

There is a clear and consistent negative effect of preventative information on endorsements for the habitual sentence (Figure 3, middle; blue points). When collapsing across items and subjecting the data to a generalized mixed-effects model with random by-participant effects of intercept and random by-item effects of intercept and conditions, we find evidence for a small effect of *enabling* conditions on endorsements (M = 0.89; 95% Bootstrapped CI [0.88, 0.91]) as compared to baseline (M = 0.85 [0.83, 0.87]) [$\beta = 0.42; SE = 0.15; z = 2.8$], and a large effect of *preventative* conditions on endorsements (M = 0.29 [0.26, 0.31]) [$\beta = -3.22; SE = 0.21; z = -15.2$].

We use the mean predicted log frequency from Expt. 3a as the input to the speaker S_2 model to predict the felicity judgments measured in Expt. 3b. We infer the two model parameters using the same analysis approach in Expt. 2. The model matches the data well ($r^2(63) = 0.91$; Figure 3, right). The same model using the past frequency as the object of communication does not match the data at all ($r^2(63) = 0.02$). These results suggest that the felicity of habituals is based on an underlying scale of predicted future propensity, not merely the observed frequency in the past.

Interestingly, we observe endorsements in this experiment that are appreciably higher than in Expt. 2 for the same items (Figure 3, middle; red vs. purple points). This may be due, in part, to an effect of the experimental context on participants: in this experiment the overall population of frequencies is much lower (both because we selected moderate frequencies from Expt. 2 and because of the preventative information) and participants may infer that the experimenter believes this to be a representative range and adjust judgments accordingly. Future investigation into this issue is warranted.

Discussion

We presented a computational model for communicating generalizations about events. The model decides if a habitual sentence is a pragmatically useful way to describe a person's behavior, taking into account the listener's prior beliefs about the action—how common it is and the likely frequency (measured in Expt. 1). We validated this model by eliciting felicity judgments for habitual sentences covering diverse activities with a wide range of experimentally manipulated frequencies of action (Expt. 2). We further investigated the nature of the underlying "propensity" scale by introducing enabling and disabling evidence, measuring the predicted future frequency (Expt. 3a) and using that, with the model, to predict the felicity of habitual sentences (Expt. 3b). To our knowledge, the experiments presented here are first empirical investigations into the truth conditions of habitual sentences and the first test of a formal model of habitual language.

The model we present here is almost identical to a model we have used to describe generic language (Tessler & Goodman, under review). The only difference is in the underlying scale: for generics, it is the *prevalence* of the property; for habituals, the *propensity* of the action. This provides a formal bridge between generalizations about categories (i.e. *generics*) and generalizations about events (i.e. *habituals*), a connection often noted in the linguistics literature (Carlson, 1977, 2005; Cohen, 1999). Generics often use a bare plural (e.g. *Bears like to eat ants.*) and don't lay claim to any well-defined set of individuals (many bears may not like to eat ants). Habituals use the simple present tense (e.g. *John smokes*) without any well-defined period of time (John may go many days without smoking). In both cases, a pragmatically inferred threshold on prevalence or propensity, respectively, explains the varying truth conditions of these kinds of sentences. Scales, and scalar representations, provide a simple and general quantitative way to express truth conditions.

Accurately predicting the environment is critical for survival and development. Habituals convey generalizations about events and are helpful for future predictions about events. For example, knowing that an event generally happens may be a useful abstraction for causal inference (e.g. Griffiths & Tenenbaum, 2005). Generalizations about people's behavior in particular—as we've investigated in this article—are important to understand, as they likely facilitate trait induction and essentialist beliefs (Gelman & Heyman, 1999). The computational model presented here provides a mathematical bridge between the way we talk about people's behavior and our intuitive theories of others and events.

Acknowledgments

This work was supported by NSF Graduate Research Fellowship DGE-114747 to MHT and a James S. McDonnell Foundation Scholar Award and ONR Grant N00014-13-1-0788 to NDG.

References

Carlson, G. N. (1977). *Reference to kinds in english*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Carlson, G. N. (2005). Generics, Habituals and Iteratives. In *The encyclopedia of language and linguistics* (2nd ed.).

Cohen, A. (1999). Generics, Frequency Adverbs, and Probability. *Linguistics and Philosophy*, *22*.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the 36th annual conference of the Cognitive Science Society.*

Gelman, S. A., & Heyman, G. D. (1999). Carrot-Eaters and Creature-Believers: The Effects of Lexicalization on Children's Inferences About Social Categories. *Psychological Science*(6).

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2015-12-30)

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334–84.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23* (pp. 587–610).

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge Univ. Press.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J. Biopharmaceutical Statistics*, *16*(4), 463–481.

Tessler, M. H., & Goodman, N. D. (under review). *A pragmatic theory of generic language.*

Learning Non-Adjacent Dependencies in Continuous Presentation of an Artificial Language

Felix Hao Wang (wang970@usc.edu)[a], Jason Zevin (zevin@usc.edu)[a,b,c], Toby Mintz (tmintz@usc.edu)[a,b,c]

[a]Department of Psychology, University of Southern California, 3620 McClintock Ave, Los Angeles, CA, 90089
[b]Department of Linguistics, University of Southern California, Los Angeles, CA, 90089
[c]Program in Neuroscience, University of Southern California, Los Angeles, CA, 90089

Abstract

Many grammatical dependencies in natural language involve elements that are not adjacent, such as between the subject and verb in *the child always runs*. To date, most experiments showing evidence of learning non-adjacent dependencies have used artificial languages in which the to-be-learned dependencies are presented in isolation by presenting the minimal sequences that contain the dependent elements. However, dependencies in natural language are not typically isolated in this way. In this study we exposed learners to non-adjacent dependencies in long sequences of words. We accelerated the speed of presentation and learners showed evidence for learning of non-adjacent dependencies. The previous pause-based positional mechanisms for learning of non-adjacent dependency are challenged.

Keywords: implicit learning; non-adjacent dependencies

Introduction

Sentences in natural languages contain grammatical dependencies, such as those that arise from agreement marking between the sentence subject and the verb. Sometimes these dependencies hold between adjacent words (or morphemes), and sometimes the dependencies are non-adjacent. For example, the dependency between the singular subject *child* and the agreeing inflected verb *runs* in *the child runs* is an adjacent dependency, whereas in, *the child always runs* it is non-adjacent. These dependencies are expressed by hierarchical syntactic structures in formal syntactic grammars. However, there has been considerable interest in investigating learning mechanisms that could detect these dependencies in linear sequences within spoken utterances. Such mechanisms could be useful for discovering syntactic structure in children acquiring a language, and could also aid proficient language users in building syntactic parses. For example, there have been a number of studies using artificial and natural languages that have investigated how language learners acquire non-adjacent dependencies (e.g., Gómez, 2002; Newport & Aslin, 2004; Peña, Bonatti, Nespor & Mehler, 2002; Romberg & Saffran, 2013; Pacton & Perruchet 2008), and how early in the acquisition process such dependencies are detected (Gómez, 2002; Gómez & Maye, 2005; Santelmann & Jusczyk, 1998).

While most studies on adjacent dependency learning report success, the same cannot be said for learning of non-adjacent dependencies. The studies to date have found evidence of non-adjacent dependency learning only in limited situations, with some studies reporting success in learning and others reporting failure. Interestingly, a characteristic of experiments that showed successful learning is that the minimal sequences that contained a dependency were presented as discrete chunks. In other words, the chunks were surrounded by silences, and the edges of such a chunk consisted of the (non-adjacent) dependent elements. For example, studies that have probed non-adjacent dependency learning between words in artificial languages typically have used trigrams in which the dependent words were at the trigram edges, and subjects were presented the trigrams one at a time, with silence intervening between presentations (Gómez, 2002; Gómez & Maye, 2005; Gómez, Bootzin & Nadel 2006; Romberg & Saffran, 2013). With the one trigram at a time design, the words immediately before and after the silences are salient for learning the dependencies given that they make up the dependency. Similarly, in experiments investigating non-adjacent dependencies between syllables in syllable sequences, learning occurred only when brief pauses were introduced before (and after) each syllable trigram (Peña et al.,2002). When syllables were concatenated continuously, participants showed no learning (see also Newport & Aslin, 2004). In the studies just discussed, the fact that subjects' success in learning non-adjacent dependencies was correlated with whether the trigrams containing the dependency were pre-segmented suggests that the chunked presentation might have played an important role in learning. One reason in which pre-segmenting the material in this way could be helpful is that it places one or both dependent elements in an edge position. Indeed, Endress, Nespor & Mehler (2009) argued that edges are privileged in the kind of position-related computations they afford, and placement at edges could be an important constraint for learning non-adjacent dependencies.

However, non-adjacent dependencies in natural language are not restricted to edge positions, and are often embedded in longer sequences. Thus, learning the dependency relations of a natural language may require learning non-adjacent dependencies of items that may not always occur at boundaries marked by silences. Given the apparent difficulty in detecting non-adjacent dependencies of continuous sequences of syllables (Newport & Aslin, 2004;

Peña et al., 2002; Gebhart, Newport, & Aslin 2009), the experiments presented here were designed to assess how detection and learning of word-level non-adjacent dependencies fairs when the critical sequences are embedded in longer sequences, such that the dependent items are not at edges.

In this study, we present non-adjacent dependencies with words concatenated together without pauses. Similar previous attempts without pauses with syllables (Newport & Aslin, 2004; Peña et al, 2002) have all reported failure to learn. Instead of CV syllables in previous studies, our study used recorded monosyllabic words with a presentation rate close to the normal speech rate (3Hz). This temporal characteristic is significantly different from previous experiments with words, when individual words were presented every 0.75 seconds (Gomez, 2002; Romberg & Saffran, 2013). We believe that this faster rate may facilitate learning in a number of ways, for a number of reasons. For one, previous theories suggested that speech processing generally occurs at the theta rate (for a review, see Kiebel, Daunizeau & Friston, 2008). For another, faster presentation may expand short memory capacity (Frensch & Miner, 1994). Moreover, it has been suggested that presenting auditory material rapidly may aid auditory statistical learning (Emberson, Conway & Christiansen 2011). Thus, presenting auditory materials rapidly arguably presents the best chance for people to learn non-adjacent dependencies in speech.

We recently described the effect of presenting English sentences for entraining grammatical boundaries to aid learning non-adjacent dependencies (Wang, Zevin & Mintz, under review). We found that non-adjacent dependency is learnable with English bracketing the boundaries of the dependency. However, whether non-adjacent dependency is learnable without English is unknown, especially under the current learning conditions, where no pauses are inserted to indicate where the dependency boundaries are. The variability at the intermediate position of the dependency has also been theorized to influence the learnability of non-adjacent dependency, where dependency with low variability is generally hard to learn (Gomez, 2002). In the current paper, we employed low variability (n=3) in the intermediate position. To summarize, we used no pauses to indicate dependency boundaries, and low variability in intermediate position of the dependency, both of which has been theorized to exacerbate the learning problem.

However, we found that the fast presentation rate is enough to yield learning in all three experiments, and that makes this the first demonstration of learning of non-adjacent dependencies at the syllable/word level. Given all of the failures to learn in the literature, we present the first success demonstration of learning word level non-adjacent dependencies (Experiment 1). We consequently replicated the finding with similarly designs (Experiment 2 & 3).

Experiment 1

Methods

Participants. Thirty-eight USC undergraduates were recruited. Half of them participated in each counterbalancing condition.

Stimuli. We recorded speech from a native English speaker and digitized the recording at a rate of 44.1 kHz. We recorded 9 novel words to form the non-adjacent dependency: 3 at position 1 (rud, swech, voy), 3 at position 2 (dap, wesh, tood) and 3 at position 3 (tiv, ghire, jub).

After all the words were recorded in list intonation, we spliced the words from the recording. Each word by itself from the recording lasts between 300ms to 737ms, and we used the lengthen function in Praat (Boersma, 2001) to shorten all the words into approximately 250ms. An additional 83ms of silence was added to the end of each word to increase its intelligibility. Thus, words occurred at a rate of 3Hz.

Design and Procedure. The experiment consisted of three blocks, each with a training phase and a number of testing trials. Each learning trial consisted of listening to materials passively. Each learning trial contained 144 non-adjacent dependency triplets. Given the word presentation rate of 3Hz, each sentence lasted 1 second, and each learning trial lasted 2.4 minutes. There were no extra pauses between any novel words of artificial language. The testing section consisted of a set of 18 question trials. Each question trial involved participants giving a familiarity rating after hearing a three word sequence. Half of the 18 questions play a triplet from the language with the correct dependency, and the other half from the counterbalancing condition, with order of presentation randomized each time.

Each novel sentence was a concatenation of 3 novel words, 1 each from choices of 3 for each position, as specified in the Stimuli section. We denote the words making up the dependencies with A and B, and the words filling up other positions with X (and Y). The pattern we tested in Experiment 1 can thus be represented as AXB. All the possible combinations occurred for A_iXB_i where the first position word predicted the third position word. As such, there were 3 AB pairs and 3 X words, which made 9 possible different artificial sentences. A counterbalancing condition was created such that the ungrammatical strings that occurred in the test are grammatical in the training sequence in the counterbalancing condition, similar to Gomez (2002). That is to say, where A_iXB_i is grammatical in one condition and A_iXB_j is ungrammatical, the reverse is true for the other condition, and both conditions use the same test items.

Training phase. At the start of the experiment subjects heard the following instructions:
"In this study, you will be presented with rapid succession of made-up words. Press Space to start listening."

Participants listened to the sound stream passively while the screen was blank during the training phase.

Testing phase. Immediately after a training phase, we showed the instructions for the testing section on the screen. The instruction made it clear that participants would hear word sequences and make judgments about them. There were a total of 18 test trials during each testing section, half of which were from the correct dependency, and the other half from the counterbalancing condition (i.e., ungrammatical). The sequence of presenting the test trials was randomized for each participant.

Participants initiated each test trial. Per trial, participants clicked on a button to play an artificial language sentence, and a question followed asking the participant to indicate whether some sequences are from the previous section that they have heard. A scale with radio buttons showed up after playing the sentence and participants were asked to answer the question "Do you think that you heard this sequence in the previous section?" There were five possible items to choose from, "Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not". Participants could click on any of the radio buttons to make their choice, and this trial ended and the next trial began.

Results

For each question in the testing section, participants rated their familiarity for a given test sequence. We coded the scale of "Definitely", "Maybe", "Not Sure", "Maybe Not" and "Definitely Not" into numeric values of 1 through 5 (Definitely = 1). This allowed us to compared ratings for grammatical items vs. the ungrammatical items.

Next, we examined the means and standard error of the ratings. We show the rating information of grammatical and ungrammatical items by block in Figure 1. To compare ratings statistically, we ran mixed effect linear regressions with the data. In the regression, ratings were compared with test item (correct vs. incorrect) as the fixed effect, and subject as the random effect. We found that participants were able to learn the non-adjacent dependency in general (β= -0.160, p<0.001).

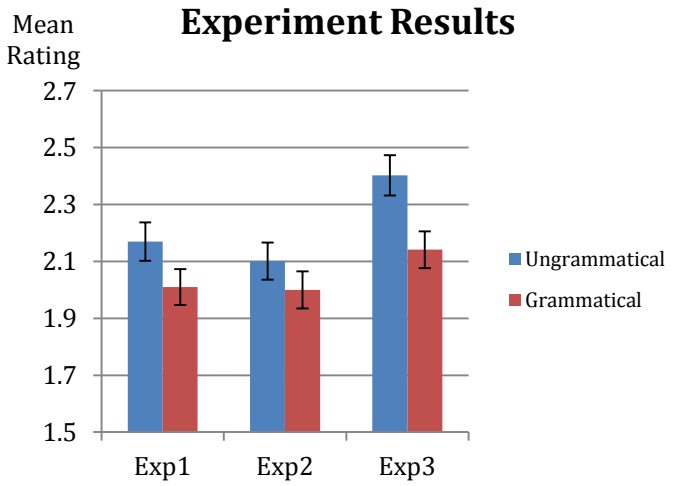


Figure 1. Rating data from all three experiments. The mean and the 95% confidence interval were plotted for each item type (grammatical/ungrammatical). The bar graph indicated that the grammatical items were judged to be more likely to be in the language than the ungrammatical items. All three experiments showed significant learning.

Discussion

Experiment 1 demonstrated that the non-adjacent dependency with one intermediate item can be learned efficiently as long as the words are presented in quick succession. However, natural languages rarely have non-adjacent dependencies stack one after other. In Experiment 2, we explore learning when there is a word between the dependencies, with the pattern of YAXB. Success in Experiment 2 should also be considered as a conceptual replication of Experiment 1.

Experiment 2

Experiment 2 tests non-adjacent dependency learning with the pattern is YAXB, where A & B words formed the dependency. Whereas Experiment 1 presented triplets with the dependency continuously, Experiment 2 has the triplet portion (AXB) separated from the next dependency by a word (Y). The choice of Y words is random with respect to other parts of the artificial language.

Methods

Participants. Thirty-eight USC undergraduates participated in Experiment 2. These participants have not participated in other experiments reported here. Half of the participants were in each counterbalancing condition.

Stimuli. We used the stimuli in Experiment 1. We used 12 novel words to form the non-adjacent dependency: 3 at position 1 (blit, pel, tink), 3 at position 2 (rud, swech, voy), 3 at position 3 (dap, wesh, tood) and 3 at position 4 (tiv, ghire, jub). There are four positions in Experiment 2 because the pattern is YAXB where A & B formed the dependency.

Again, all the words were approximately 250ms long with an additional 83ms of silence was added to the end of each word. When words are concatenated in a continuous stream, they would occur at a rate of 3Hz.

Design and procedure. The experiment consisted of three blocks, each with a training period followed by a sub-block of 18 question trials. Each learning period consisted of listening to materials passively. Each learning trial contained 144 non-adjacent dependency triplets. Given that words were at 3Hz and each sentence contained 4 words, each sentence took a second and a third. Each learning trial lasted 3.2 minutes. During the testing, each testing section contains 18 questions, half from the language and half from

the counterbalancing condition, with order of presentation randomized each time.

Training phase. At the start of the experiment subjects heard the following instructions:
"In this study, you will be presented with rapid succession of made-up words. Press Space to start listening."
Participants listened to the sound stream passively while the screen was blank during the training phase.

Testing phase. Immediately after a training phase, we showed the instructions for the testing section on the screen. The instruction made it clear that participants would hear word sequences and make judgment about the sequences. There were a total of 18 test trials during each testing section, half of which were from the correct dependency, and the other half from the counterbalancing condition. The sequence of presenting the test trials was randomized for each participant.

Participants initiated each test trial. Per trial, participants clicked on a button to play an artificial language sentence, and a question followed asking the participant to indicate whether some sequences were from the previous section that they have heard. A scale with radio buttons showed up after playing the sentence and participants were asked to answer the question "Do you think that you heard this sequence in the previous section?" There were five possible items to choose from, "Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not". Participants could click on any of the radio buttons to make their choice, and this trial ended and the next trial began.

Results

We examined the means and standard error of the ratings (Figure 1). To compare ratings statistically, we ran mixed effect linear regressions with the data. In the regression, ratings were compared with test item (correct vs. incorrect) as the fixed effect, and subject as the random effect. We found that participants were able to learn the non-adjacent dependency in general (β= -0.101, p=0.021).

Experiment 1 & 2 demonstrate that the non-adjacent dependency with one intermediate item can be learned efficiently as long as the words are presented in quick succession.

Experiment 3

Natural languages are not restricted to have only one word in between the dependency (e.g., *the child very rarely runs*). In cases where there is more than one item in between the items that form the dependency, it has been suggested that learning becomes more difficult (Santelmann & Jusczyk, 1998). In Experiment 3, we explore learning when there are two intermediate items between the dependencies, with the pattern of AXYB.

Methods

Participants. Thirty-eight USC undergraduates participated in Experiment 2. These participants have not participated in other experiments reported here. Half of the participants were in each counterbalancing condition.

Stimuli. In Experiment 3, we explore learning when there are two intermediate items between the dependencies, with the pattern of AXYB. We used the stimuli in Experiment 1. We used 12 novel words to form the non-adjacent dependency: 3 at position 1 (rud, swech, voy), 3 at position 2 (blit, pel, tink), 3 at position 3 (dap, wesh, tood) and 3 at position 4 (tiv, ghire, jub).

Again, all the words were approximately 250ms long with an additional 83ms of silence was added to the end of each word. When words are concatenated in a continuous stream, they would occur at a rate of 3Hz.

The experiment consisted of three blocks, each with a training period followed by a sub-block of 18 question trials. Each learning phase consisted of listening to materials passively. Each learning trial contained 216 non-adjacent dependency triplets. Given that words were at 3Hz and each sentence contained 3 words, each sentence took a second. Each trial lasted 3.6 minutes. There were no extra pauses between any novel words of artificial language. During the testing, each testing section contains 18 questions, half from the language and half from the counterbalancing condition, with order of presentation randomized each time.

Training phase. At the start of the experiment subjects heard the following instructions:
"In this study, you will be presented with rapid succession of made-up words. Press Space to start listening."
Participants listened to the sound stream passively while the screen was blank during the training phase.

Testing phase. Immediately after a training phase, we showed the instructions for the testing section on the screen. The instruction made it clear that participants would hear sound sequences and make judgment about the sequences. There were a total of 18 test trials during each testing section, half of which were from the correct dependency, and the other half from the counterbalancing condition. The sequence of presenting the test trials was randomized for each participant.

Participants initiated each test trial. Per trial, participants clicked on a button to play an artificial language sentence, and a question followed asking the participant to indicate whether some sequences are from the previous section that they have heard. A scale with radio buttons showed up after playing the sentence and participants were asked to answer the question "Do you think that you heard this sequence in the previous section?" There were five possible items to choose from, "Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not". Participants could click on any of the radio buttons to make their choice, and this trial ended and the next trial began.

Results

We examined the means and standard error of the ratings (Figure 1). To compare ratings statistically, we ran mixed effect linear regressions with the data. In the regression, ratings were compared with test item (correct vs. incorrect) as the fixed effect, and subject as the random effect. We found that participants were able to learn the non-adjacent dependency in general ($\beta = -0.261$, $p < 0.001$).

In sum, we found that robust learning is present even when there are 2 items between the words forming the non-adjacent dependency.

Discussion

As we mentioned, learning non-adjacent dependencies in the lab has been demonstrated in very restricted situations. There are a variety of reasons for this. For the most part, past literature suggested (Newport & Aslin, 2004; Peña et al., 2002) that pauses are critical to the learning of non-adjacent dependencies. Our design does not contain pauses, which makes our study the first we know that showed success of learning non-adjacent dependencies without resorting to pauses. There have been studies of non-adjacent dependencies with auditory artificial language where the non-adjacent dependency is embedded in which dependent items sometimes occur at edges enables the detection of non-adjacent patterns (Mintz et al., 2014; Reeder, Newport & Aslin, 2013; Wang & Mintz, under review). In the cases where successful learning of non-adjacent dependencies has been reported, at least one edge (beginning or ending) is marked with pauses (Mintz et al., 2014; Reeder, Newport & Aslin, 2013). When both edges are not marked with pauses, learning failed (Wang & Mintz, under review). It is possible that having exposure to elements at edge positions facilitated, detecting non-adjacent dependencies at least initially. However, natural languages contain non-adjacent dependencies at non-edge positions, thus making it difficult to evaluate learning theories that requires the presence of pauses.

Why would the presentation rate make a difference? There are a number of possibilities. The auditory system for speech perception may be tuned towards a particular frequency (Kiebel, Daunizeau & Friston 2008), so efficient speech processing may play a role. This line of explanation is along the lines of modality-specific statistical learning theories (Emberson, Conway, & Christiansen 2011). They argued for the central role of modality-specific processing by contrasting the opposite influence of changing presentation rate in visual and auditory statistical learning. These theories hold promising directions for understanding the modality-specific statistical learning mechanisms, but they are also vague regarding why specific kind of statistical learning (in this case, non-adjacent dependency learning) would benefit from fast presentation. We leave these questions for future research.

Success in learning non-adjacent dependencies without pauses point to the possibility the non-adjacent dependency learning mechanisms with spoken language do not critically require the presence of pauses as a prosodic cue, contrary to previous theories (see Peña et al. 2002, for a discussion). Peña et al. 2002 argued that successful learning requires a prosodic analysis whereby boundaries and positional information is obtained before non-adjacent dependencies are learnable. Across all the previous studies that report success on non-adjacent dependency learning with spoken artificial language (Peña et al. 2002; Newport & Aslin, 2004), this has been the case. The current work suggests that pauses are not a necessary condition for learning non-adjacent dependencies. In the absence of explicit pauses, we speculate that the learning mechanism may still have access to virtual boundaries that arise via a distributional analysis that detects simpler repeated patterns. There may be some kind of distributional or syntactic analyses that can make uses of these positional boundaries which in turn may reduce the computational load for calculating dependency relations between non-adjacent items, and induce the detection of higher order dependencies, such as non-adjacent dependency in the current paper. Future work should examine these possibilities. In sum, the position based accounts (Endress et al., 2009) may still apply to the current findings, except that positional information may not come from prosodic processes, but it may be obtained from distributional analysis as well.

One methodological note is regarding the measure we took, which is a rating scale of confidence. This is different from how artificial language is assessed in the past literature, which involves a variant of this question, "Have you heard this sentence in the language before?", "Is this sentence in the language?", etc, requiring a yes/no answer from participants. There are a number of problems with this approach, most of which involves the interpretation of the phrase "in the language". What does it mean to a naïve participant that a novel sentence is in a novel language? Does it mean that the sentence literally heard? Or does it mean that it follows some kind of a rule? Regardless, given any interpretation of "in the language", participants also need to decide on the criterion when a phrase is "close enough" to be in the language. Many artificial language studies from our labs suggest that participants may simply answer yes to all questions, because it is not clear to them what the experimenter is asking (for similar results, see Gómez, 2002). In light of these findings, we used a rating scale instead of collecting yes/no responses. Rating that making subjects make explicit judgments, we asked participants to report their confidence level that a phrase has been heard. This measure, degrees of certainty, does not require any commitment to any type of meta-linguistic knowledge of knowing what it means to be "in a language", but rather, assesses familiarity with a phrase. Making use of this measure has yielded much success with multiple artificial language/statistical learning studies from our lab already.

Getting back to our study, we wish to emphasize the role of the timing. There are other word level non-adjacent dependency studies (Gómez, 2002; Romberg & Saffran,

2013), but the timing profile is different. In those experiments, utterances were concatenated words, such that there was around 0.8 s between word onsets. This is a relatively slow rate of speech, which can be considered unnatural as far as speech perception is concerned in terms of its timing characteristics. It is conceivable that this mode of presentation makes detecting patterns of non-adjacent elements more difficult because they are not temporally close.

Lastly, existing theories (Gomez, 2002, among others) suggest that the dependency is hard to detect without highly variable middle elements. This is different from our design in important ways. In our design, the variability of the middle elements (n=3) is very low according to Gomez 2002, making the dependency hard to learn. We show that this hard problem of learning of non-adjacent dependency can be solved when the non-adjacent dependency is presented at a typical speech rate. It remains possible that the variability issue is important when the presentation rate of speech is slow, but at least with fast presentation rate, low variability does not seem to lead to failure to learn. Future work is needed to examine whether increase variability will make learning more robust.

In sum, we have shown that temporally controlled word-level non-adjacent dependency is learnable without pauses. We propose that learning about distributional analysis may be best obtained the learning material is presented at the optimal rate is critical, and the importance of the speech rate may outweigh constraints previously proposed, such as presence of pauses and variability in the middle element.

References

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* (2001): 341-345.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, 12(3), 396-406.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *The Quarterly Journal of Experimental Psychology*, 64(5), 1021-1040.

Endress, A.D., Nespor, M. & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, 13(8), 348-353.

Frensch, P. A., & Miner, C. S. (1994). Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory & Cognition*, 22(1), 95-110.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing Structures in Midstream: Learning Along the Statistical Garden Path. *Cognitive Science*, *33*(6), 1087–1116. http://doi.org/10.1111/j.1551-6709.2009.01041.x

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431-436.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7, 183-206.

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science*, 17(8), 670-674.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology*, 4(11).

Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. *Cognitive psychology*, 75, 1-27.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 80.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604-607.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2), 127-162.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, 66(1), 30-54.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, *37*(7), 1290-1320.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, 69(2), 105-134.

Shi, R., & Melançon, A. (2010). Syntactic Categorization in French-Learning Infants. *Infancy*, 15(5), 517-533.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552-564.

Van den Bos, E., & Christiansen, M.H. (2009). Sensitivity to non-adjacent dependencies embedded in sequences of symbols. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

Vuong, L.C., Meyer, A.S. & Christiansen, M.H. (in press). Concurrent statistical learning of adjacent and non-adjacent dependencies. *Language Learning*.

Wang, H., Höhle, B., Ketrez, N. F., Küntay, A. C., & Mintz, T. H. (2011). Cross-linguistic Distributional Analyses with Frequent Frames: The Cases of German and Turkish. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th annual Boston University Conference on Language Development* (pp. 628-640). Somerville, MA: Cascadilla Press.

Wang, F. H., & Mintz, T. H. (under review). Learning Non-Adjacent Dependencies Embedded in Sentences of an Artificial Language: When Learning Breaks Down.

Process Modeling of Qualitative Decision Under Uncertainty

David A. Broniatowski (broniatowski@gwu.edu)
Department of Engineering Management and Systems Engineering, 800 22nd Street NW #2700
Washington, DC 20052 USA

Valerie F. Reyna (vr53@cornell.edu)
Human Neuroscience Institute
Cornell University, Ithaca, NY 14853, USA

Abstract

Fuzzy-trace theory assumes that decision-makers process qualitative "gist" representations and quantitative "verbatim" representations in parallel. Here, we develop a formal model of fuzzy-trace theory that explains both processes. The model also integrates effects of individual differences in numeracy, metacognitive monitoring and editing, and sensation seeking. Parameters of the model varied in theoretically meaningful ways with differences in numeracy, monitoring, and sensation seeking, accounting for risk preferences at multiple levels. Relations to current theories and potential extensions are discussed.

Keywords: Decision making; need for cognition; risky choice; framing effect; Allais paradox

Fuzzy-Trace Theory

Risk preferences are fundamental to psychological and economic theory, and to decision neuroscience. We propose a model of risk preferences that integrates theoretical principles relevant to mental representations with individual differences in metacognitive monitoring and risk-taking propensity. Our model is based on fuzzy-trace theory, an account of decision-making under risk, which posits that decision-makers use qualitative, categorical, "gist" representations of the meaning of decision information, in parallel with precise, eidetic or metric, "verbatim" representations of the exact words and numbers in that information (for an overview, see Reyna, 2012). By "mental representation," we mean the manner in which a stimulus is encoded into a subject's memory. Decision makers operate on these representations of the stimulus rather than on the stimulus itself. Specifically, a gist representation captures the basic *meaning*, or "essence," of a stimulus. Furthermore, fuzzy-trace theory posits a hierarchy of gist that is, in the domain of numbers, analogous to scales of measurement (Reyna, 2008; Stevens, 1946). We approximate this hierarchy with three levels of representation: categorical, ordinal, and interval, described below.

The outline of this paper is as follows: first we present a motivating example. Next, we describe our novel formalization, accounting for factors that vary with individual-differences. Finally, we test our model's parsimony and predictions.

Example

Fuzzy-trace theory is motivated by the insight that one's representation of a decision problem can drive decision outcomes. For example, consider a choice between:

1. winning $180 for sure; versus
2. a .90 chance of winning $250 and .10 chance of no money.

One might represent this decision as a simple choice between the following two options:

1. Possibility of some money
2. Possibility of some money and possibility of no money.

Given this representation, most decision makers would favor option 1 because it promises some money without the chance of no money. Alternatively, one could instead represent the choice as follows:

1. More chance of winning less money
2. Less chance of winning more money and possibility of winning no money.

This representation, although more precise, does not allow for a clear decision to be made because most people would prefer winning more money to winning less money, but they would also prefer more chance of winning to less chance of winning. Finally, one may choose a precise representation of the problem whereby one calculates the expected value of each option by multiplying its respective outcomes by their probabilities, as follows:

1. Expected value of $180 (i.e., $180 * 1)
2. Expected value of $225 (i.e., $250 * 0.90 + $0 * 0.10)

This representation seems to favor option 2, since it promises more money on average. Overall, this example illustrates our approach to categorical gist representation.

The Decision Space

We represent the complements in these options as points in a 2-dimensional space, representing all possible combinations of amounts of money (or, generally, some outcome) and probability that a decision-maker could encounter:

Categorical representation of decision space.

The gist representation of the choice above is between:

1. Possibility of some money
2. Possibility of some money and possibility of no money.

These gists are represented in a 2-dimensional space. All points in this space are interpreted according to the part of the diagram in which they are located. These gist representations can overlap. For example, a point that falls into the part of the space marked as "possibility of no money" also falls into the part marked as "possibility of some money." Thus, multiple gist representations are possible for some points.

The extended fuzzy-processing preference.

We model the relationship between categories as a hierarchy in which higher elements are preferred interpretations when compared to lower elements (Broniatowski & Reyna, 2014). Each decision complement is a point in our space. If we determine each complement's gist representation, and then locate that gist in the associated hierarchy, the model stipulates how that decision complement is interpreted.

Mapping problem information to categorical mental representation. We assume that the categorical distinction between "some" of a quantity and "none" of a quantity is primitive. Thus, our model assumes that all points are mapped to one of these two categories. Throughout this paper, we assume that stimuli map to either "some" or "none" in both the domains of outcomes chance for scientific parsimony.

Formalizing the extended fuzzy-processing preference. We introduced the extended fuzzy-processing preference to enable us to differentiate between overlapping gists. According to the extended fuzzy-processing preference, subjects will prefer to interpret a decision option as within the subcategory containing the fewest points– i.e., the highest element within the associated hierarchy. We formalize the extended version of the fuzzy processing preference by specifying that one always prefers the interpretation associated with the category in the decision space with the lowest dimension (similar to Feldman's, 1997, "maximum codimension rule"). In other words, given a point in our space and a set of possible gist interpretations for that point, a subject will always prefer the interpretation that is highest in the associated hierarchy.

Importantly, a preferred interpretation, or mental representation, is not always the same as a preferred outcome. For example, a decision-maker may prefer to interpret of ".10 chance of no money" as "possibility of no money" when compared to "possibility of some money." Given a choice between two options, interpreted respectively as "possibility that of some money" and "possibility of no money," most decision-makers would choose the former option. Thus, once mental representations are chosen, we must define a preference ordering over the decision options with these interpretations.

Values map mental representations to preference. Decision makers choose between options based on which has the high-valued affect. The affect assigned to a given option is a function of how that option is represented. For example, "no money is possible" is a preferred interpretation for the point ($0, .10, a .10 chance of no money) when compared to "some money is possible." However, a prospect that is interpreted as "some money is possible" has a high valence when compared to one that is interpreted as "no money is possible." Thus, a decision-maker would choose the option with the high valence. In order to formalize this prediction, we again use a partial order – i.e., every pair of elements within the category hierarchy may be less than, greater than, equal to, or unrelated to one another in the domain of values. Full mathematical details of this partial order are presented in the paper by Broniatowski & Reyna (2014).

Mapping problem information to ordinal mental representations. Fuzzy-trace theory predicts that decision-makers use ordinal (e.g., "more" vs. "less") in parallel with categorical and interval representations. When mapping problem information to ordinal mental representation, "more" is always in the direction away from zero and "less" is always in the direction toward zero. Importantly, points may only be compared at the ordinal level if they exist within a common category. For example, one may compare "no money with .10 chance" to "$180 for sure" because both may be represented as "possibility if some money" (even if this is not the preferred interpretation for either option, it is an admissible interpretation for both). Since 0 is less than 180 and .10 is less than certainty, the corresponding ordinal representations are "less_1 money with less_2 chance" and "more_1 money with more_2 chance."

Mapping ordinal mental representations to preference. Ordinal decision-making assumes that each dimension in the decision space has a preferred direction. When comparing two decision options, if the ordinal representation of one option is preferred along all dimensions of the decision space, and is strictly preferred along at least one dimension, then that decision option is preferred overall. For example, "more money with more chance" is preferred when compared to "less money with less chance." Otherwise, a decision cannot be made and the ordinal representation is indifferent, such as when "more_1 money with less_2 chance" is compared to "less_1 money with more_2 chance".

Formalizing How Each Representation Chooses Among Decision Options

A "gist hierarchy" is a set of mental representations ranging in precision from a categorical gist representation to an interval verbatim representation, and sets of rules for making decisions that are unique to each of these representations. At the categorical level, each point is represented according to the extended fuzzy-processing preference. At the ordinal level, a point is chosen if it is weakly preferred along all dimensions and strongly preferred along at least one such dimension. Points in disjoint categories cannot be compared. At the interval level, decisions options are evaluated according to their expected values (i.e., the sum of the value of each outcome multiplied by its probability) – the simplest interval representation (i.e., it assumes no decision weights).

Combining Information Across Representations

We address conflicts between representations in our model by assuming that each representation casts a "vote" for its preferred decision options. For example, given a choice between two decision options, each of the categorical, ordinal, and interval representations "votes" (-1 for the first option, +1 for the second option, or 0 if indifferent) for a preferred option according to its own particular representational logic. A sum across these votes (i.e., a weighted sum, as explained below) determines the final decision. We chose summation because it is the simplest combination rule for this sort of aggregation.

An Error Theory for Risky Decision Problems

We represent error using a standard multinomial logistic distribution. For decisions with two options, effect size typically follows a standard logistic distribution. (In principle, our model could be extended to decisions with multiple options using a multinomial logistic distribution.) For our specific application, we model the probability, P, that a subject will choose a given decision outcome in a risky choice gamble by:

$$P(\vec{x}) = \frac{1}{1 + e^{-(\vec{a}\cdot\vec{x}+b)}}.$$

Here, $\vec{x}$ is a three-element vector containing an entry for each representation (categorical, ordinal, and interval), $\vec{a}$ is a three-element vector containing an entry for each corresponding decision weight. We also introduce a factor, b, capturing the risk-taking propensity of a given set of subjects. Thus, we account for conflict between representations by adding weighted votes from each representation.

Factors affecting the decision weight vector. In the domain of decision making, two major individual difference factors associated with metacognitive monitoring and editing have been proposed – numeracy (e.g., Peters et al., 2006; Liberali et al., 2012) and Need for Cognition (NFC; Cacioppo, et al.,1996; Stanovich & West, 2008).

People who are higher in numeracy and/or NFC are more likely to spontaneously convert and compare alternative "framings" of a problem, reducing cognitive biases.

Numeracy. Peters and colleagues (2006) defined numeracy as "the ability to process basic probability and numerical concepts" and found that more numerate subjects were less susceptible to attribute framing effects. In addition, Schley and Peters (2014) found that more numerate individuals treated numbers as more linear when making a risky decision, suggesting that they rely less on categorical gist and more on interval (linear) representations of probabilities and outcomes.

Table 1: Evaluation of Model Fit.

Model	Like-lihood	AIC	BIC
Null	-7491	14982	14987
Saturated	-6570	13491	14049
Analytic Categories	-6672	13409	13510
Single average value for a and b	-6826	13676	13715

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion

Need for cognition. Prior work suggests that subjects reconcile answers to oppositely framed versions of the same problem when both frames are presented within-subjects, when subjects respond to multiple presentations of the same problem, or when they are exposed to obviously factorial design manipulations within subjects. Kahneman and Frederick (2002) have argued that such designs can lead subjects to focus on the variables that are being manipulated, and to compare different versions of the same underlying problem instead of treating each independently. Thus, the magnitude of framing effects varies systematically with experimental design (e.g., Stanovich & West, 2008).

The tendency to reconcile responses to different versions (or related problems) when they are presented within-subjects is greater for those higher in NFC. Subjects with high NFC tend to edit their choices more than those with low NFC, presumably because they are more likely to notice the common structures underlying these problems (i.e., high NFC subjects display "analytic override;" e.g., LeBoeuf & Shafir, 2003; Stanovich & West, 2008). Furthermore, numeracy and NFC are separate sources of individual differences that are not correlated (Peters et al., 2006; Liberali et al., 2012).

We model the effects of numeracy and NFC using the decision weight vector $\vec{a}$. Furthermore, if we make the simplifying assumption that all of these decision weights are equal, we may replace $\vec{a}$ by a scalar factor, a, which captures the "strength" of a given set of votes. When a is high, subjects will strongly favor one option over another, unless different representations conflict with one another. When a is low, subjects will tend towards indifference.

Risk-taking propensity. In addition, our model incorporates personality differences associated with risk-taking (e.g., Caspi et al., 1997), including factors related to cross-cultural differences (e.g., Du et al., 2002) and sensation seeking or reward-related approach (e.g., Zuckerman, 2007). We represent this in our model by a linear additive risk preference, b, which, when positive, is used to indicate a predisposition toward the higher rewards available in the more risky option in a gamble. The linear additive nature of this factor is based on evidence presented by Reyna and colleagues (2011) who found evidence supporting additive effects (i.e., additive beyond verbatim and gist processing) of subjects' sensation seeking on risk taking.

A worked example. Consider the decision between a certain gain of $180 and a 0.90 chance of winning $250 and a 0.10 chance of no money discussed above. Recall that the categorical representation prefers the certain option (-1), the ordinal representation is indifferent (0), and the interval representation prefers the risky gamble (1).

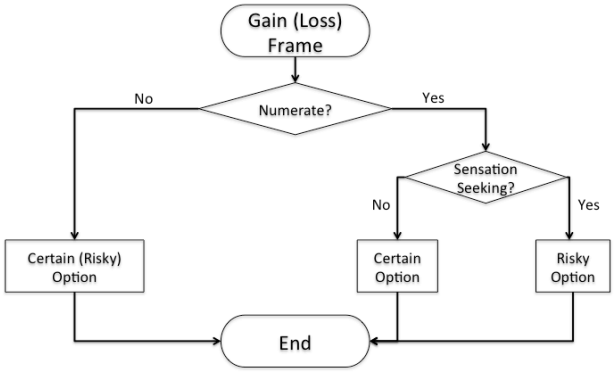


Figure 1: Process Flowchart for Between-Subjects Condition, No Replications, Individual Representation

Thus, $\vec{x}$ = [-1,0,1] in our model. For an experiment that is conducted with frame manipulated between subjects, $\vec{a}$ = [1,1,1]. Thus, $\vec{a} \cdot \vec{x}$ = -1 + 0 + 1 = 0 (indicating that the categorical and interval representations compete). Finally, suppose we estimate our sample's risk propensity from prior data to be b = 0.25, indicating a slight preference for risk. Then, the probability that a randomly chosen subject from our sample will choose the risky gamble option is P= $1/(1+e^{-0.25})$ = 56%, and the probability that a randomly chosen subject will choose the certain option is 44%.

Assessing The Goodness of Fit of our Model

To demonstrate the scientific parsimony of this model, we have adapted a technique used by Busemeyer, Wang, & Shiffrin (2015), wherein we compare our model's fit to a "null" model (in which each decision option is equally likely), a "saturated" model (in which maximum-likelihood parameter values are separately estimated for each replication in our sample), and a model that estimates parameters based upon theoretically-motivated categories: namely mathematical ability (PISA scores) and experimental design (for the a parameter), and stimulus type, nationality, and age (for the b parameter), with parameter values estimated separately within each category using a jackknife estimator. Specifically, given a model, y, the log likelihood function for each experimental replication in our sample of 88 is:

$\ln[L(y_i)] = n_{1,1}\ln(p_{1,1}) + n_{1,2}\ln(p_{1,2}) + n_{2,1}\ln(p_{2,1}) + n_{2,2}\ln(p_{2,2})$

where $n_{1,1}$ is the number of people choosing the first decision option in the first problem, $p_{1,1}$ is the predicted proportion of subjects choosing the first decision option in the first problem, given model y, etc., and

$$\ln[L(y)] = \sum_i \ln[L(y_i)]$$

is the total log-likelihood of model y. Given ln[L(y)], we may calculate the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as follows:

$$AIC = 2k - 2\ln[L(y)]$$

where k is the total number of distinct values for a, b, and x in a given model (e.g., the total number of analytic categories). Similarly,

$$BIC = k\ln(n) - 2\ln[L(y)]$$

where n is the total number of data points in our sample (i.e., 176 data points for 88 pairs of problems, listed in the paper by Broniatowski & Reyna, 2015). Table 1 shows the log-likelihood, AIC, and BIC values for the models that we tested (we tested other models as well, but none surpassed the model with analytic categories). The model containing analytic categories for a and b has the lowest values of both AIC and BIC, and only the saturated (i.e., overfit) model has a higher log likelihood.

Individual Differences Analysis

Estes and Maddox (2005) point out that aggregate data analyses may lead to different conclusions from those reached through individual-level analyses. This concern does not apply to data derived from between-subjects designs, on the other hand, we may use the results of within-subjects designs to perform process level tests of our model. Specifically, we examined three replications of a framing problem for which framing was manipulated within-subjects and individual level frequency data were reported (Frisch, 1993; Stanovich & West, 1998; LeBoeuf & Shafir, 2003). We extracted the number of subjects that were consistent between frames (either choosing the certain gamble or the risk option in both frames), the number of subjects who displayed a framing effect (choosing the certain option in the gain frame and the risky gamble in the loss frame), and the number of subjects who displayed reverse framing (the risky gamble in the gain frame and the certain option in the loss frame. These data were used to test the following process-level accounts of our model.

Between-subjects framing. Figure 1 shows a process-level account of a framing problem for which frame was manipulated between subjects with a single presentation. Here, a subject is exposed to a stimulus containing a given frame. If they are not numerate, they choose the decision option consistent with that frame. If they are numerate, and they are not sensation seeking, they choose the certain option. Otherwise, they choose the risky option. Thus, an individual who chooses the risky option in the gain frame would be numerate and sensation seeking according to our model. Furthermore, we may perform aggregate-level analysis of between-subjects data, which enables us to replace individual choices with probabilities. Specifically, the probability that a subject randomly chosen from our sample would pick the risky option in the gain frame is $p_{gain} = p_{numerate} * p_{ss}$. Similarly, the probability that a subject would pick the certain option in the loss frame is $p_{loss} = (1 - p_{numerate}) + p_{numerate} * p_{ss}$. Furthermore, our model stipulates that $p_{gain} = \frac{1}{1 + e^{-(-a+b)}}$ and

Table 2: 3 Sampled Experimental Replications of Individuals' Decisions When Frame is Manipulated Within-Subjects

Reference	% High NFC			Empirical			Predicted				
	N	MLE	JK	C	F	RF	C	F	RF	χ^2	p
				$\vec{x}$ = [±1,0,0], a=0.89, b=0.22							
Frisch (1993)	99	48	59	63	29	7	71	23	6	2	0.29
Stanovich & West (1998)	292	56	57	202	73	17	205	69	18	0.3	0.87
LeBoeuf & Shafir (2003)	287	61	54	206	60	21	194	74	19	4	0.16

Note. C = Consistent – same decision in both frames; F = Framing – certain option in gain frame & risky gamble in loss frame; RF = Reverse Framing – certain option in loss frame & risky gamble in gain frame; MLE = Maximum Likelihood Estimate; JK = Jackknife estimator generated by averaging across all MLE values except the one corresponding to a given problem

$p_{loss} = \frac{1}{1+e^{-(a+b)}}$. Finally, we may estimate the values of a and b using the by averaging across the maximum likelihood estimates of a and b for all problems with the exception of the problem whose parameters we are trying to estimate. Thus, we have a system of two equations with two unknowns, such that

$$p_{numerate} = 1 + \frac{1}{1+e^{-(-a+b)}} - \frac{1}{1+e^{-(a+b)}} \quad \&$$

$$p_{ss} = p_{numerate}{}^{*}(1+e^{b-a})$$

Within-subjects framing. Figure 2 shows a process-level account of a framing problem for which frame was manipulated within subjects.

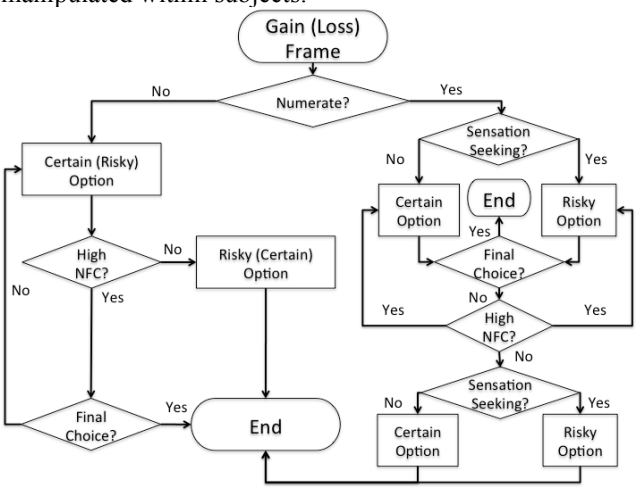


Figure 2: Process Flowchart for Within-Subjects Condition, Individual Representation

Here, subjects are exposed to a stimulus containing a given frame in a manner analogous to a between-subjects framing problem: If they are not numerate, they choose the decision option consistent with that frame. If they are numerate, and they are not sensation seeking, they choose the certain option. Otherwise, they choose the risky option. The individual then makes a second choice on an oppositely-framed version of the same framing problem (if they first saw the loss frame, they now see the gain frame and vice versa). Individuals with high NFC will recognize the similarity of the two problems and remain consistent across frames (LeBoeuf & Shafir, 2003) whereas individuals with low NFC will treat the second problem as if it were independent of the first. As in the between-subjects case, we may replace individual choices with probabilities. This formulation enables us to estimate the proportions of subjects that will be consistent, exhibit framing, or exhibit reverse framing for a given sample where frame is manipulated within-subjects. Specifically, the proportion of subjects in a given sample that is consistent across frames is given by $p_{NFC} + {p_{ss}}^2(1-p_{NFC})p_{numerate}$, the proportion of subjects that exhibiting framing behavior is given by

$$(1-p_{NFC})(1-p_{numerate}) + (1-p_{ss})p_{ss}(1-p_{NFC})p_{numerate}$$

and the proportion of subjects exhibiting reverse framing is:

$$\boldsymbol{(1-p_{ss})p_{ss}(1-p_{NFC})p_{numerate}}$$

Process-level tests of our model. We compare our model's predictions to individual level behavior (i.e., sequential choices) by estimating values of $p_{numeracy}$, p_{NFC} and p_{ss} for a given experimental sample using the equations above. Specifically, we compare our model's predictions to three replications of a framing problem presented within-subjects that are reported in the literature (Frisch, 1993; LeBoeuf & Shafir, 2003; Stanovich & West, 1998). Specifically, we estimated the baseline proportion of subjects who are numerate from the problems in our sample for which framing was manipulated between-subjects. Specifically, we estimated the value of $\hat{a}$ as the average of maximum-likelihood a values for low-PISA between-subjects framing problems (since all within-subjects problems are also from low-PISA countries) weighted by the total number of subjects (recall that the first choice made in a within-subjects design is analogous to a between-subjects framing problem), excluding the data from Stanovich and West (1998) and LeBoeuf and Shafir (2003) from our average in order to avoid *post hoc* estimation (Busemeyer & Wang, 2000), yielding a value of $\hat{a}$ = 0.89. Similarly, we estimate the value of $\hat{b}$ as the weighted average of all MLE b values of the ADP when presented to comparable (i.e., non-Chinese, who are known to be more

risk-taking; Du et al, 2002) college student samples, again excluding the data from Stanovich and West (1998) and LeBoeuf and Shafir (2003), yielding a value of $\hat{b}$=0.22. By extension, $p_{numeracy}$=59% and p_{ss}=58%. Finally, we calculated p_{NFC} to compare the number of subjects choosing consistency, framing, or reverse framing with our model's predictions. The value of p_{NFC} was chosen to minimize the sum of squares between the predicted and actual numbers of subjects within each category. Finally, we averaged across all values of p_{NFC} excluding the value associated with that particular problem (i.e., using a jackknife estimator). Table 2 shows that the data do not differ significantly from our model's predictions in any of these three experiments.

Conclusions

This model is the first to explicitly formalize the key concepts of gist, the gist hierarchy, and qualitative decision-making. Previously, (Broniatowski & Reyna, 2014; 2015) we introduced the mathematical underpinning and error theories underlying our model. Here, we demonstrated that this model is both scientifically parsimonious and robust. The structure of our model also enables it to outperform leading theoretical alternatives, such as Cumulative Prospect Theory (CPT; Tversky & Kahneman, 1992), which make modal, and not precise, predictions. Thus, our formalized theory explains a wide variety of phenomena, integrating known effects and novel predictions.

References

Broniatowski, D. A., & Reyna, V. F. (2014). A Mathematical Formalization of Fuzzy Trace Theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society

Broniatowski, David A., & Reyna, V. F. (2015). Formalizing Risky Choice with a Logistic Model of Fuzzy Trace Theory. In Noelle, Dave, Dale, Rick, Warlaumont, Anne, Yoshimi, Jeff, Matlock, Teenie, Jennings, Carolyn, & Maglio, Paul (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Busemeyer, J. R., & Wang, Y. M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*(1), 171–189.

Busemeyer, J. R., Wang, Z., & Shiffrin, R. M. (2015). Bayesian model comparison favors quantum over standard decision theory account of dynamic inconsistency. *Decision*, *2*(1), 1-12.

Cacioppo, J. T., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197-253.

Caspi, A., Begg, D., Dickson, N., Harrington, H., Langley, J., Moffitt, T. E., & Silva, P. A. (1997). Personality differences predict health-risk behaviors in young adulthood: evidence from a longitudinal study. *Journal of Personality and Social Psychology*, *73*(5), 1052-1063.

Du, W., Green, L., & Myerson, J. (2002). Cross-cultural comparisons of discounting delayed and probabilistic rewards. *Psychological Record*, *52*(4), 479-492.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403-408.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, *54*(3), 399-429.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, *41*(2), 145–170.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment*, 49–81.

LeBoeuf, R. A., & Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, *16*(2), 77–92.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407–413.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making*, *7*(3), 332–359.

Reyna, V. F., Estrada, S. M., DeMarinis, J. A., Myers, R. M., Stanisz, J. M., & Mills, B. A. (2011). Neurobiological and memory models of risky decision making in adolescents versus young adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1125-1142.

Schley, D. R., & Peters, E. (2014). Assessing "Economic Value": Symbolic-Number Mappings Predict Risky and Riskless Valuations. *Psychological Science*, *25*(3), 753–761.

Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking & Reasoning*, *4*(4), 289–317.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672-695.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Zuckerman, M. (2007). Sensation Seeking and Risk. In M. Zuckerman, *Sensation seeking and risky behavior*. (pp. 51–72). Washington: American Psychological Association

A Nonlinear Dynamical Systems Theory Perspective on Dual-Processing Accounts of Decision-Making under Uncertainty

Marieke M. J. W. van Rooij (m.vanrooij@bsi.ru.nl)
Behavioral Science Institute,
Radboud University Nijmegen, the Netherlands

Luis H. Favela (luis.favela@ucf.edu)
Department of Philosophy, Cognitive Sciences Program
University of Central Florida, USA

Abstract

Dual-processing accounts of reasoning have gained renewed attention in the past decade, particularly in the fields of social judgment, learning, and decision-making under uncertainty. Although the various accounts differ, the common thread is the distinction between two qualitatively different types of reasoning: explicit/implicit, rational/affective, fast/slow, etc. Consequently, much research has focused on characterizing the two different processes. Less extensive are the attempts to find mediators that influence which process is used. In this paper, we argue that the missing perspective on these dual-processing theories is rooted in dynamical systems theory. By shifting the perspective to the dynamic interaction and transitions between different types of reasoning, we provide a theoretical framework for dual-processing with an emphasis on phase transitions. As a special case, we focus on dual-processing in decision-making and judgment under uncertainty for which we will propose suggestions for future experimental evaluation.

Keywords: decision-making under uncertainty; dual-processing; nonlinear dynamical systems theory; phase transitions

Dual-Process Theories of Reasoning

Dual-processing accounts of reasoning go back to one of the oldest ideas in psychology, namely, that the human mind is not driven by a single, unified process but by two. Exemplars of such ideas include Plato, Hermann von Helmholtz, William James, and Sigmund Freud (Frankish & Evans, 2009).

The history of psychological frameworks for conceptualizing capacities of the mind—especially in terms of decision-making and reasoning—culminated in the 1970s with *dual-process* theories. Dual-process theories have taken on various forms (e.g., see Evans, 2008 for a review). Nevertheless, there are some common features. First, these theories tend to explain the working of the human mind in terms of two qualitatively distinct cognitive systems, and are referred to as type 1/type 2 (Goodwin & Wason, 1972), System 1/System 2 (Stanovich, 1999), or intuitive/deliberative (Kahneman, 2003). Moreover, these two kinds of cognitive systems tend to be differentiated along the following dichotomies: unconscious/conscious, fast/slow, automatic/controlled, emotional/rational, intuitive/rule-based, etc. There have been, and continues to be, broad applications of these dualistic frameworks. These include, but are not limited to, social judgment (Doherty & Kurtz, 1996), stereotyping (Bodenhausen, Macrae, & Sherman, 1999; Wilson, Lindsey, & Schooler, 2000), learning and problem-solving (Sun, Slusarz, & Terry, 2005), creative thinking (Allen & Thomas, 2011), and perhaps most famously, decision-making and judgment under uncertainty, which we will discuss below as a special case.

The aforementioned phenomena have been investigated and explained in terms of two distinct cognitive systems or processes. However, although they are all "dual-process" theories, many of the models utilize slightly different characterizations of what is being "dual-processed." For the purposes of the current work, we will utilize "dual-processing" as referring to *System 1*, which is fast-working, implicit, and affect-related, and *System 2*, which is slow-working, explicit, and, analytic.

Criticism

Dual-processing theories of reasoning are not without opposition (for a recent discussion, see Evans & Stanovich, 2013). Here we highlight three points of criticism we think are most prominent and crosscutting. We refer to these as *conceptual vagueness*, *evidence,* and *transition*.

Perhaps the most fundamental challenge facing dual-process theories is a lack of conceptual clarity. As noted above, although for

current purposes we refer to dual-processing phenomena in terms of System 1 and System *2*, there are a number of other labels that are readily used, e.g., *Type 1* and *Type 2, or intuitive vs. deliberate processing*. In addition, although the features that fall under each category can be delineated (e.g., Evans & Stanovich, 2013; Frankish & Evans, 2009), it is unclear whether the literature is consistent in the use of such categorizations and if the features are agreed upon. For example, the term *unconscious* is readily used as a feature describing System 1, and *conscious* as a feature describing System 2. Though, what does conscious and unconscious even mean? Theoretical terms that refer to the domain of the "mental" are particularly burdened with the need to be conceptually clear precisely because they are inaccessible directly to third person, objective scientific practice. The lack of such clearly defined and operationalized terms is a particular weakness of dual-process theories that rely on one process being slow and conscious, whereas the other is fast and unconscious.

Another critique of dual-process theories stems from a lack of sufficient empirical evidence. This point is interweaved with the previous critique stemming from conceptual vagueness. It is difficult to have evidence for something if it is unclear what that something is. If we do not know what it is to be conscious or unconscious, it is tough to say that data is evidence of it. Along these lines, without clear conceptual distinctions between System 1 and System 2 features, the question will remain whether there is sufficient empirical evidence to support the strong dichotomy (e.g., Kruglanski & Gigerenzer, 2011; Osman, 2004).

The final critique is the unwarranted strong focus on static rather than dynamic properties of reasoning. Scarcely any attention has been directed to the *transition* from one form of processing to another. The current understanding of factors that influence dual-processing strictly focus on static and binary relationships such that the factors necessary to cause one system to activate do not cause the other system to activate. For example, *meta-cognitive* difficulty has been found to activate System 2 (Alter, Oppenheimer, Epley, & Eyre, 2007), but it is unknown how or exactly when the activation occurs.

Note that it is not our aim to merely reiterate the existing criticisms to dual-processing models. Our starting point is the phenomenon. For example, the difference between calculating two plus two and 17 times 24 is that for most of us, the latter likely involves a combination of mental gymnastics, finger counting, and putting pen to paper whereas the answer to the first immediately comes to mind (cf. Winerman, 2012). Consequently, instead of discussing questions about the validity of dual-processing as a model for cognitive performance, we propose an entirely different take. We propose treating the transition between different types of cognitive processes in terms of nonlinear dynamic change, which, as we will show, makes the discussion about the strict dichotomy of dual-processes theories obsolete.

Our Proposal: A Nonlinear Dynamical Systems Theoretical Approach to Dual-Processing

Although there are several theoretical and methodological challenges facing dual-process theories of reasoning, there have been strides made towards more nuanced and sophisticated accounts. Evans and Stanovich (2013) put forward a theoretical approach to dual-processing in terms of default responses that can be intervened upon by higher-order reasoning processes. This account provides a potential response to the transition critique discussed above and avoids the pitfalls associated with models that place features of cognitive processing necessarily within a framework of binary categories. Instead, there are *types* of processing that share System 1 or System *2* characteristics. Cognitive processing can occur in both kinds of properties—for example, conscious and unconscious—such that it is not the binary nature of properties that distinguish the kinds of thinking so much as the processes underlying the thinking that differentiates them.

Additionally, Wastell (2014) has attempted to address the issue of conceptual vagueness by appealing to the concepts of *complexity theory*, particularly the notion of *emergence*. Wastell's is a theory of human reasoning that tries to bring together emergence and modularity theory. He asserts that reasoning, "is a product of our interaction with the environment and our innate ability to create, store, and utilize virtual reasoning modules" that "emerge" from module-environment information fit (2014, p. 354-355).

As Evans and Stanovich prudently note, there is a lot of work still to be done, as the development of dual-process theories—not to mention human reasoning in general—is an evolving project that is likely to continue to develop (2013). In the spirit of contributing to

this evolving project, we present a nonlinear dynamical systems theory perspective on qualitative transitions in human reasoning, with emphasis on the case of decision-making under uncertainty.

Nonlinear Dynamical Systems Theory

A system can be thought of as a set of parts that have relationships to one another, especially when the integrated elements accomplish a shared or defined task (Norman & Kuras, 2006). Almost all systems are *dynamic*, i.e., their behavior changes over time. Dynamical systems theory refers to the mathematical understanding of change and stability in dynamic systems. Dynamical systems theory provides the tools to describe, model, and evaluate interactions and transitions between qualitatively different behaviors without demanding the definition of qualitatively different systems. Dynamic systems behaviour can be *nonlinear*, i.e., the output is disproportionally related to the input due to multiplicative interactions between its components (Carello & Moreno, 2005; van Rooij, Nash, Rajaraman, & Holden, 2013). This can result in sudden unexpected qualitative transitions from one stable behavior to the next. These kinds of transitions are particularly relevant for a different understanding of dual-processing phenomena in reasoning.

To illustrate how nonlinear dynamical systems theory can shed new light on dual-processing phenomena in reasoning, let us take a look at the transition of ice to liquid water. At certain critical values of pressure and temperature (e.g., 0° Celsius at sea-level), the tiniest change in either can cause the ice to rapidly become liquid whereas at other values a much larger change may not result. This change between two qualitatively different states of water is called a *phase transition*. Phase transitions can be anticipated or accompanied by universal dynamical patterns, so-called *catastrophe flags* (Gilmore, 1993; Isnard & Zeeman, 1976; Thom, 1972). For example, although common understanding is that the transition of liquid back into ice occurs at 0° Celsius, in reality water does not freeze until it is -4° Celsius. This means that between 0° and -4° Celsius, water is *multistable*, and the transition of ice into water—and vice versa—does not necessarily occur at the same temperature value. Moreover, a so-called *hysteresis* effect is observed: After ice turns into water due to increasing temperature, a significantly larger decrease in temperature is needed to re-establish the previous solid state.

Sudden change, multistability, and hysteresis are three of the eight known catastrophe flags (Gilmore, 1981). All eight catastrophe flags are behavioral properties of nonlinear dynamical systems *at the level of their dynamics*, although some (e.g., hysteresis and multistability) can also be observed at the level of the system behavior itself. Observing catastrophe flags around qualitative change is a strong indicator that the system of interest is a nonlinear dynamical system, and that the observed change is indicative of a phase transition. Catastrophe flags have been observed in relation to a broad range of human behavioral changes. The most extensive and successful applications of nonlinear dynamical systems theory to human behavior are to human development (e.g., Thelen & Smith, 1998; van Geert, 1994), movement, action, and perception (e.g., Fajen & Warren, 2003; Haken, Kelso, & Bunz, 1985; Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007; Warren, 1984).

Applications to human higher-order cognition are less numerous for at least two reasons: First, the dominant computational-representational paradigm in cognitive science resulted in a strong focus on stable states rather than change (cf. Stephen & Van Orden, 2012). Second, in addition to issues related to localizing cognitive functions in the brain (cf. Anderson, 2014), defining precise boundaries around cognitive systems (e.g., Chemero, 2009; Favela & Chemero, 2016) and properly accounting for relevant degrees of freedom in cognitive systems can be very challenging (cf. Kelso, 1995). There have a number of successes though. One particular empirically compelling demonstration of a higher-order cognitive phase transition is by Van Orden and colleagues (1999), who identified homophones as a bifurcation phenomenon.

In the following section, we present a case study demonstrating the application of nonlinear dynamical systems theory with an emphasis on phase transitions to inform a novel approach to dual-processing in decision-making. It is important to note that the concept of a "phase transition" is not merely a metaphor for physical change; it refers to *real* change in a system. A phase transition is nature's mechanism of qualitative change. It provides a way to explain the existence of qualitatively different behaviors within a continuum.

Case Study: Dual-Processing as Phase Transition in Decision-Making and Judgment under Uncertainty

As noted above, that humans have multiple decision-making capabilities is an idea that has reoccurred over the history of psychological theorizing. One commonality among such theories (e.g., Plato, Helmholtz, Kahneman, etc.) is that these multiple kinds are distinct, modular, and cognitively encapsulated processes. Such a treatment is faced with serious theoretical and methodological challenges stemming from issues related to conceptual vagueness and lack of empirical evidence. One particular criticism, which we think has not received enough attention, stems from understanding the transitions among these various capabilities. Our main claim is that such transitions are best understood as *phase transitions*.

Suggestions for Experimental Evaluation

Taking a nonlinear dynamical systems theoretic view on dual-processing phenomena in decision-making under uncertainty means we need to study the transitions in cognitive processing along the continuum of reasoning from ultra-quick memory retrieval to slow explicit processing. This requires finding the transition itself, which includes identifying order and control parameters that characterize the underlying system.

Finding the Transition. In a clever series of experiments, Alter et al., (2007) devised ways in which they could activate their version of System 2. For example, they found that participants presented with information that was difficult to read were less prone to reasoning biases typically associated with System 1 and concluded that incidental experiences of difficulty or "disfluency" activated System 2. Unfortunately, in each of the experiments, only two conditions were used. The obvious extension to this type of work would be to present participants with a gradual change in legibility of the presented information in order to find the actual transition, and see whether this is indeed discontinuous (as dual-processing theories suggest) or continuous.

Finding the Order and Control Parameters. *Order parameters* capture macroscopic states of a system (Haken, 1988/2006). These order parameters determine the behavior of individual parts of the system. Order parameters can also be thought of as the dependent variable of a system, that is, the target of measurement in an experiment. *Control parameters* are variables that guide a system's dynamics. Control parameters can also be thought of as independent variables of a system. Note however that although it can be helpful to think of control parameters as independent variables and order parameters as dependent variables, it is important to keep in mind that the relationship between order and control parameters is not exactly the same as that of dependent and independent variables. The control variable does not *cause* the system's behavior. This is partially due to a "slaving" relationship between order and control parameters (Haken, 1988/2006). Although control parameters are intertwined with the dynamics exhibited by the order parameter, the relationship is nonlinear.

Identifying the order and control parameters allows for the development of models that capture the full range of a system's dynamics. In the same way that the interaction between pressure and temperature results in phase transitions of water, we must therefore discover variables that lead the cognitive system through its possible states in decision-making. Following Thelen and Smith (1998), such a task can be summed up as follows: First, identify the order parameter of interest. Second, characterize the qualitatively different behaviours corresponding to different values of the order parameter. Third, describe the dynamic trajectory of the order parameter. Fourth, identify points of transition. Finally, identify potential control parameters and manipulate putative control parameters to experimentally generate transitions.

Recently, we utilized these steps to demonstrate catastrophe flags in the behavioral transition between risk-seeking and risk-averse strategies in a simple risky choice task (van Rooij, Favela, Malone & Richardson, 2013a, 2013b) and similarly in the preference between immediate and delayed rewards (van Rooij & Richardson, 2014). The success of these studies bolsters the case for taking nonlinear dynamical systems theoretic approach to dual-processing.

Another factor hypothesized to affect decision making in terms of dual-processing is *cognitive load*. Cognitive load refers to the demand that cognitive activities place on working memory (e.g., Miller, 1956). When System 1 is under heavy cognitive load, its ability to correct System 2 decreases, which can lead to more unexpected decision behavior. For example, increased cognitive load results in higher than

expected levels of risk aversion (Benjamin, Brown, & Shapiro, 2013). We hypothesize that cognitive load may act as a so-called *splitting factor*, i.e., an order parameter influencing the moment of occurrence of the phase transition.

Discussion & Future Directions

Attempts to characterize the workings of the human mind in terms of two cognitive systems or processing, as well as debates about the conceptual strength of these characterizations, distract from a more fundamental issue: Starting with the phenomenological distinction between slow/deliberate and fast/intuitive reasoning, it is the interactions between these processes as well as the dynamic control that should be the focus of scientific inquiries into the subject. Accordingly, this requires an approach rooted in nonlinear dynamical systems theory.

Nonlinear dynamical systems are also referred to as *complex systems*, and are formed from (many) parts such that the behavior of the system is irreducible or deducible from the properties of the parts (Norman & Kuras, 2006). In other words, the behavior of complex systems is dominated by the dynamic interactions between parts rather than the properties of the parts themselves (Van Orden, Holden, & Turvey, 2003, 2005).

Most natural systems are complex in the ways just described. A clear example is the human brain. Not only is brain activity constantly interacting and controlled by behavioral and environmental factors (e.g., Van Orden, Hollis, & Wallot, 2012), but emerging, global-level properties such as consciousness are not easily captured in one-to-one relations between single neurons or even single areas of the (Anderson, 2014). A growing body of literature indeed places human cognition firmly within a framework of nonlinear dynamical systems or complex systems theory. We believe dual-processing theories could also be understood within this framework and outlined possibilities for experimental evaluation.

References

Allen, A. P., & Thomas, K. E. (2011). A dual process account of creative thinking. *Creativity Research Journal*, *23*(2), 109-118.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General 136*(4), 569.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economic Association*, *11*(6), 1231-1255.

Bodenhausen, G. V., Macrae, C. N., & Sherman, J. S. (1999). On the dialectics of discrimination: Dual processes in social stereotyping. In S. Chaiken & Y. Trope (Eds.). *Dual Process Theories in Social Psychology (*pp. 271–90). New York: Guilford.

Carello, C. & Moreno, M. A. (2005). Why nonlinear methods? In M. A, Riley & G. C. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences* (pp. 1-25). Retrieved March 10, 2012 from www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT press.

Doherty, M. E., & Kurz, E. M. (1996). Social judgement theory. *Thinking & Reasoning*, *2*(2-3), 109-140.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255-278.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.

Fajen, B. R., & Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 343.

Favela, L. H., & Chemero, A. (2016). The animal-environment system. In Y. Coelllo & M. H. Fischer (Eds.), *Foundations of Embodied Cognition: Volume 1* (pp. 59-74). New York, NY: Routledge.

Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: an historical perspective. In *In Two Minds: Dual Processes and Beyond* (pp. 1-29). Oxford: Oxford University Press.

Gilmore, R. (1993). *Catastrophe theory for scientists and engineers*. Courier Corporation.

Goodwin, R. Q., & Wason, P. C. (1972). Degrees of insight. *British Journal of Psychology*, *63*, 205-212.

Haken, H. (1988/2006). *Information and self-organization: A macroscopic approach to complex systems, third edition*. New York, NY: Springer.

Haken, H., Kelso, J. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological cybernetics*, *51*, 347-356.

Isnard, C. A., & Zeeman, E. (1976). Some models from catastrophe theory in the social sciences. *The Use of Models in the Social Sciences. Tavistock, London, UK*.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697-720.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*, 97-109.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Norman, D. O., & Kuras, M. L. (2006). Engineering complex systems. In D. Braha, A. A., Minai, & Y. Bar-Yam (Eds.), *Complex Engineered Systems: Science Meets Technology* (pp. 206-245). Berlin: Springer.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review 11*(6), 988-1010.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science*, *26*, 867-891.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, N.J: Lawrence Erlbaum Associates.

Stephen, D. G., & Van Orden, G. (2012). Searching for general principles in cognitive performance: Reply to commentators. *Topics in Cognitive Science*, *4*, 94-102.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: a dual-process approach. *Psychological review*, *112*, 159-192.

Thelen, E., & Smith, L. B. (1998). Dynamic systems theories. In W. Damon (Ed.), *Handbook of Child Psychology: Theoretical models of human development* (Vol. 1, 5th ed., pp. 563–634). New York, NY: Wiley.

van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos*. Harvester Wheatsheaf.

Van Orden, G. C., Holden, J. G., Podgornik, M. N., & Aitchison, C. S. (1999). What swimming says about reading: Coordination, context, and homophone errors. *Ecological Psychology*, *11*, 45-79.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331-350.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/*f* scaling. *Journal of Experimental Psychology: General*, *134*(1), 117-123.

Van Orden, G., Hollis, G., & Wallot, S. (2012). The blue-collar brain. *Frontiers in Physiology*, *3*, 1-12.

van Rooij, M. M. J. W., Favela, L. H., Malone, M. L., & Richardson, M. J. (2013a). A Dynamical Model of Risky Choice. In M. Knauff, M. Pauen, N. Sebanz, & I. Waschsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1510-1515). Austin, TX: Cognitive Science Society.

van Rooij, M. M., Favela, L. H., Malone, M., & Richardson, M. J. (2013b). Modeling the dynamics of risky choice. *Ecological Psychology*, *25*(3), 293-303.

van Rooij, M. J. W., Nash, B., Rajaraman, S., & Holden, J. G. (2013). A fractal approach to distribution analysis and dynamic inference. *Frontiers in Physiology*, *4*. doi: 10.3389/fphys.2013.00001

van Rooij, M. M. J. W, & Richardson, M. J. (2014). *Delay Discounting Explained with a Two-attractor Nonlinear Model*. 6th International Nonlinear Science Conference 2014, 20-22 March 2014, Nijmegen, the Netherlands.

Warren, W. H. (1984). Perceiving affordances: visual guidance of stair climbing. *Journal of experimental psychology: Human perception and performance*, *10*(5), 683.

Wastell, C. A. (2014). An emergence solution to the reasoning dual processes interaction problem. *Theory & Psychology*, *24*, 339-358.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*, 101-126.

Winerman, L. (2012). A machine for jumping to conclusions. *Monitor on Psychology*, *43*(2), 24.

Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction

Max Kleiman-Weiner[1] (maxkw@mit.edu), Mark K. Ho (mark_ho@brown.edu)[2], Joseph L. Austerweil[2] (joseph_austerweil@brown.edu), Michael L. Littman[3] (mlittman@cs.brown.edu), Joshua B. Tenenbaum[1] (jbt@mit.edu)

[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[2]Cognitive, Linguistic, and Psychological Sciences / [3]Computer Science, Brown University, Providence, RI 02912

Abstract

Successfully navigating the social world requires reasoning about both high-level strategic goals, such as whether to cooperate or compete, as well as the low-level actions needed to achieve those goals. While previous work in experimental game theory has examined the former and work on multi-agent systems has examined the later, there has been little work investigating behavior in environments that require simultaneous planning and inference across both levels. We develop a hierarchical model of social agency that infers the intentions of other agents, strategically decides whether to cooperate or compete with them, and then executes either a cooperative or competitive planning program. Learning occurs across both high-level strategic decisions and low-level actions leading to the emergence of social norms. We test predictions of this model in multi-agent behavioral experiments using rich video-game like environments. By grounding strategic behavior in a formal model of planning, we develop abstract notions of both cooperation and competition and shed light on the computational nature of joint intentionality.

Keywords: joint intention, cooperation, coordination, reinforcement learning, teams

Introduction

Our most important relationships involve understanding when to cooperate and when to compete. From siblings to coworkers, humans rely on both planning and context to know which situations they should cooperate in and which they should compete in (Galinsky & Schweitzer, 2015; Rand & Nowak, 2013). And yet in real life, unlike a behavior economics experiment, cooperation and competition are abstract with respect to a given situation. A cooperative or competitive interaction unfolds over time – there isn't a single moment where competition or cooperation "happens". Even if the decision to cooperate or compete has been made, efficiently implementing those strategies can be difficult. A person determined to cooperate and knowing what the other person wants will have to develop a detailed plan of action to realize that cooperative intention. Likewise for a person intent on competing. In this work we aim to bridge high-level strategic decision making over abstract social goals such as cooperation and competition with low-level planning over actions to actually realize those goals.

The ability to form these hierarchical joint intentions is a key component of social behavior. The motivated instinct to both infer and evaluate complex social plans emerges in early childhood (Warneken & Tomasello, 2006; Hamann, Warneken, Greenberg, & Tomasello, 2011). Young children not only rapidly infer the goals of other agents, but spontaneously execute complex plans to cooperate with others. For instance, a cooperative intention might generalize to include not just the low-level details of a joint task but also tell how to share the spoils. The ability to infer the intentions of oth-

Matrix-Form Games

		Yellow: Cooperate	Yellow: Compete
Blue	Cooperate	7,7	-1,8
	Compete	8, -1	4,4

Figure 1: A social dilemma written as a normal-form game. The numbers in each square specify the payoff in terms of utility to the blue and yellow player respectively for choosing the action corresponding to that square's row and column. If both agents choose cooperate they will collectively be better well off than if they both choose compete. However in any single interaction, either agent would be materially better off by choosing to compete.

Stochastic Games

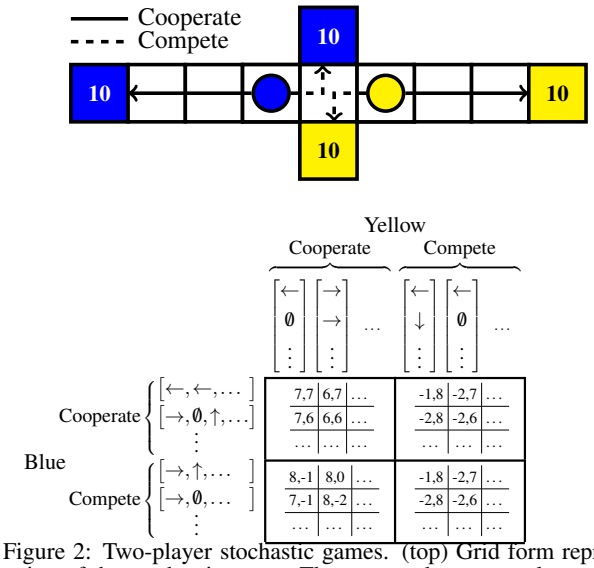


Figure 2: Two-player stochastic games. (top) Grid form representation of the stochastic game. The arrows show example strategies that can be used to realize both cooperative and competitive outcomes. (bottom) Matrix representation of the strategy space, with low-level strategies sorted by a high-level goal. The arrows correspond to moving in a specific direction and the $\emptyset$ corresponds to waiting. Note that the action space is effectively unbounded but the strategies naturally cluster into a small number of high-level goals. If both agents go to the sides then they will both score the reward but if they fight for the middle in hopes of using less moves they will collide and only one will get any reward.

ers and participate in a dynamic joint endeavor (sometimes called the "we-mode") is thought to be a key building block of large scale collaborative culture (Tomasello, Carpenter, Call, Behne, & Moll, 2005).

Naturalistic Games

Game-theoretic investigations of social behavior often represent strategic interactions as *matrix-form* games like the one shown in Figure 1. In these games, the rows and columns correspond to the action space of the two players and the cells describe the payoffs to each agent that would result from those actions. While useful as a succinct representation of a social decision, these games lack the ecological validity of real social decisions which require planning across space and time. When presented to participants, it can be difficult to extract the right information and even after significant training, many people don't even look at the payoffs most relevant for strategic reasoning (Costa-Gomes, Crawford, & Broseta, 2001). When the number of decisions grows beyond two decisions per agent, these problems are exacerbated.

Instead we use a paradigm commonly deployed in multi-agent systems research which has not been explored behaviorally (De Cote & Littman, 2008). In this paradigm, strategic interactions are represented as naturalistic spatial environments that people play intuitively like video-games. Figure 2 shows an example of one of these multi-agent planning environments that is conceptually related to the social dilemma shown in Figure 1. Unlike the matrix-form game, these environments also require low-level planning over spatial actions to realize a strategic goal. The action space of these games is much larger than those typically studied in matrix-form games but the strategies are still intuitive.

Each player controls the movement of one of the colored circles. On each turn players choose to either move their circle into an adjacent square (not including diagonal moves) or to remain in the same position. Attempting to move is costly resulted in the loss of one point. Choosing to remain in the same position did not incur any cost. Both players select an action during the same turn and their positions are updated simultaneously. Each square can only be occupied by one player at a time so if both players try to move to the same square, one of the players chosen by chance will enter the contested square while the other remains in place. However both pay the cost for attempting to move. If one player stays in the same position and the other player tries to move into their square, no movement occurs. Finally, players cannot move through each other and switch places.

The colored squares are the goals. When either player reaches a square with the same color as their avatar, that player receives ten points and the round ends. Thus the only way for both players to receive points is if they both enter squares that match their avatar's color on the same turn. These dynamics were chosen to be identical to those in (De Cote & Littman, 2008) so that our data can also be compared to the models of that work. Because each interaction generates data about both the action plan and the payouts, we can use these games to start to investigate the mechanisms people use to coordinate on cooperative and competitive outcomes. Furthermore, they allow us to study how humans innovate to find these strategies out of such a large possible space of action plans.

Model

Hierarchical Social Planning

We develop a hierarchical model of strategic planning that unifies low-level action planning with high-level strategic reasoning and allows for learning across both levels. In brief, agents have two "modes" of low-level planning: a cooperative mode and a competitive mode. These two modes are connected through a high-level strategic planner that determines which mode should be deployed based on previous interactions. After each round, agents use Bayesian theory-of-mind to determine whether or not the other agent's low-level actions are consistent with the cooperative planning mode vs. the competitive planning mode. The agent can then condition their own next actions on the inferred high-level intentions of the other agent realizing a sophisticated strategic response.

Both modes include forms of model-based learning which allows for learning to generalize across environments as well as model-free reinforcement of actions. In this work we focus specifically on the high-level goals of cooperation and competition but other high-level goals such as teaching, punishing or communication are also relevant in these games and will be investigated in future work. The challenge of hierarchical planning is to link these high-level goals to a lower-level plan of action.

Our work builds on and is inspired by classical formalisms of intention and joint planning from the AI literature (Levesque, Cohen, & Nunes, 1990; Grosz & Kraus, 1996) as well as more modern formulations for planning under uncertainty such as DEC-POMDPs and I-POMDPs (Gmytrasiewicz & Doshi, 2005; Gal & Pfeffer, 2008; De Cote & Littman, 2008). However the earlier models do not handle uncertainty in a probabilistic way and hence struggle with quantitative predictions about behavior while the later are often intractable over long planning horizons and don't explicitly represent abstract social goals.

We briefly introduce stochastic games following the notation of De Cote and Littman (2008) and then discuss repeated stochastic games. A two-player stochastic game is: $\langle S, s_0, A_1, A_2, T, U_1, U_2, \gamma \rangle$ where S is the set of all possible states with $s_0 \in S$ the starting state. Each agent can choose from a set of actions A_1 and A_2 which together form a joint action space $A_1 \times A_2$. The state-transition function, $T(s, a_1, a_2) = P(s'|s, a_1, a_2)$ maps a state and joint action to a distribution over new states. The utility functions of the two agents $U(s', s, a_1, a_2) = R$ describe the agent's goals in terms of quantitative costs and rewards. Finally $0 \leq \gamma_{\texttt{game}} \leq 1$ is the discount rate of reward. In repeated stochastic games, a series of stochastic games are played one after another in succession between the same pair of players. We now discuss the

cooperative and competitive modes of planning in detail.

Cooperative Planning

Since there is no specific action that corresponds to cooperation in these stochastic games (all actions are spatial movements), we develop an abstract notion of cooperation which generalizes across contexts. We postulate that a cooperative action is one that is good for the group i.e., efficiently maximizes the utility of all agents. Since under this assumption, the goal of cooperation is to rationally achieve a group goal, we consider a *group-agent* that optimizes a utility function composed of the utility of all agents (Sugden, 1993, 2003; De Cote & Littman, 2008).

Computationally, we represent this group utility function as a linear weighting of the utility of the two agents: $U^G = (w)U_1 + (1-w)U_2$ where $w \in [0,1]$ controls how the two agents are relatively valued by the group-agent. For example when $w = 0.5$ the group-agent impartially weighs the utility of both agents equally. We are not implying that this group-agent actually exists but rather that each player can simulate the same group-agent by taking an objective view of the planning environment outside and separate of their own personal goals (Nagel, 1986). We note that this utility function can include other social preference such as inequality aversion or merit based allocations.

Since the group-agent can directly control the actions of both players (like a "we" agent), it can treat the stochastic game as a single-agent MDP. Rational planning over joint actions (a_1, a_2) is achieved through value-iteration:

$$P(a_1,a_2|s) = \pi^G(s,a_1,a_2) \propto e^{(\beta Q^G(s,a_1,a_2))}$$

$$Q^G(s,a_1,a_2) = \sum_{s'} P(s'|s,a_1,a_2)[U^G(s',s,a_1,a_2) + \gamma \max_{(a_1',a_2')} Q^G(s',a_1',a_2')]$$

where the group-agent policy, $\pi^G(s)$, is to choose actions with probability proportional to their future expected utility. A high value of β means that the group-agent is more likely to select the action with the highest Q-value and a low value of β means that the group-agent is more likely to select suboptimal-actions. In all experiments we used a relatively high value of $\beta = 4$. We note that π^G is not only a policy for action, but also includes the future-oriented intentions of what the two agents *should* do once they get to a new state. These intentions include how to recover from failed coordination attempts. We used a discount rate of $\gamma = 0.9$ in all the models presented here.

Although each agent might consider the policy of the group-agent, the individual agents can only control their own actions. To transform this group-agent policy into an individual policy, individual agents marginalize out the actions of the other player from the joint policy: $\pi_1^G(s,a_1) = \sum_{a_2} \pi^G(s,a_1,a_2)$ and $\pi_2^G(s,a_2) = \sum_{a_1} \pi^G(s,a_1,a_2)$. These policies contain intertwined intentions, not only an *intention to* take a specific action but also the *intention that* the other agent reach certain states. This "meshing" of plans between the two agents has been called a key component of joint and shared intentionality (Bratman, 1993, 2014). Unlike social preference based accounts of cooperative behavior where each agent individually plans to maximize joint utility, in this account, cooperation is a built in cognitive feature of planning itself – agents *plan together*.

When there is a single unambiguous action for both players that maximizes joint utility, coordination is readily achieved. However in the environments we investigate, there are often multiple actions that can generate optimal rewards for the group-agent. We now discuss two mechanisms for learning social norms that can break these symmetries and lead to robust coordination on a single jointly optimal plan.

We first consider the case where two different actions are equally good from the perspective of a group-agent that weighs the utility of the two agents equally but the rewards will be allocated unequally. For example, consider game (C) in Figure 3 where one agent needs to go around the other. Because moving costs 1 point, the agent who goes around the other will only earn 7 points while the agent who waits will earn 9 points. From the perspective of the group-agent with $w = 0.5$, it doesn't matter who goes around since the joint utility is equal. However if one agent was favored over the other ($w \neq 0.5$) this symmetry would be broken and the disfavored agent would take the long route. Thus prior knowledge about asymmetries in how the group should operate can lead to more robust coordination although potentially at the cost of less fair cooperation.

The two agents may start with a different prior on the value of w and thus when simulating the group-agent will fail to coordinate. Consider the case where both agents think they should be valued more than the other and hence expect the other player to go around them. We propose a mechanism based on "virtual bargaining" accounts of social choice that lead to each agent's w to converge over time to the same value without any explicit communication (Binmore, 1998; Misyak, Melkonyan, Zeitoun, & Chater, 2014). After each interaction, agents can infer the w that best explains the joint behavior of their previous interaction: $P(w|H) \propto P(H|w)P(w)$ where H are the data from previous interactions and the likelihood of those interactions is defined by the marginalized joint policies generated from planning with a specific w: π_1^G and π_2^G. In our analysis, each agent starts out with a prior of $w = 0.5$ and updates it after each round based on the inferred w of the previous interaction. Thus over time w will converge and as predicted by the theory of virtual bargaining, more patient agents who insist on the advantage will gain a greater share of the joint reward in future coordinated interactions where an equitable split isn't possible. For example, if in a previous interaction agent 1 took a more costly route, then in the next round agent 1 will be more likely to take the costly route again generating a social norm for cooperative coordination. Since w is an input to the planning process itself, it allows for generalizing these norms to new environments.

Finally, in some environments, there are multiple plans that are equally good for both agents, creating a different type of symmetry which cannot be broken by w. For example, the decision to go clockwise or counterclockwise in game (A) of Figure 3 is equally good for both players as long as they both go in the same direction. To capture the intuition that once agents successfully coordinate, they should continue to coordinate in that way e.g., after luckily choosing to go clockwise in game (A), they will go clockwise again on the next round, agents learn a function $N^G(s,a_1,a_2)$ based on the frequency of previous joint actions which is added to the state-action Q^G-value used by the group-agent. This norm based reinforcement affects the policies of the individual agents through marginalization. The norms reinforced by this mechanism do not generalize across environments although feature based norms can generalize when there are features in common between two environments e.g., see Ho et al. in this years proceedings.

Competitive Planning

As before, in these stochastic games there is no action that directly corresponds to "compete". Instead, we ground competitive planning as each agent attempting to maximize their individual utility under the assumption that the other agent is doing the same. To tractably realize this game-theoretic best-response, we extend the cognitive hierarchy / level-K formalism used in behavioral game theory to temporally extended polices instead of just actions (Camerer, Ho, & Chong, 2004). In brief, a level-K agent best responds to a level-$(K-1)$ agent which grounds out in the level-0 agent. Specification of the level-0 agent is sufficient to specify the full hierarchy.

In this work we use a level-0 agent that doesn't consider the existence of the other player and tries to efficiently reach her goal without taking any strategic consideration of how the other player might affect her progress. This level-0 agent is more naturalistic than randomly acting agents which are commonly used in behavioral modeling (Camerer et al., 2004; Yoshida, Dolan, & Friston, 2008). A level-0 agent of this type only makes sense in these naturalistic environments since one can easily imagine acting alone unlike in matrix-form games. The level-0 agent for player i is:

$$P(a_i|s,k=0) = \pi_i^0(s) \propto e^{\beta Q_i^0(s,a_i)}$$
$$Q_i^0(s,a_i) = \sum_{s'} P(s'|s,a_i)(U_i(s,a_i,s') + \gamma \max_{a_i'} Q_i^0(s',a_i'))$$

where $P(s'|s,a_i)$ represents transition dynamics that do not depend on the other player. Having defined the level-0 player we can recursively define all of the other levels in the hierarchy in terms of lower levels:

$$P(a_i|s,k) = \pi_i^k(s) \propto e^{\beta Q_i^k(s,a_i)}$$
$$Q_i^k(s,a_i) = \sum_{s'} P(s'|s,a_i)(U(s,a_i,s') + \gamma \max_{a_i'} Q_i^k(s',a_i')))$$

Since the other agent is treated as a knowable stochastic part of the environment, the dynamics of the other player are encapsulated in $P(s'|s,a_i)$ which are marginalized out using the $k-1$ player: $P(s'|s,a_i) = \sum_{a_{-i}} P(s'|s,a_i,a_{-i})P(a_{-i}|s,k=k-1)$ where $-i$ is a shorthand to refer to the "other" player. Because of the maximization operator, a level-K agent implements a best response to a level-$K-1$ agent. Thus zeroth-order agents have their own goals but ignore the other player, first-order agents act on their own goals but assume that the other agent is ignoring their existence and so on. In our experiments we used $K=1$ although results were similar with higher values of K.

Even when competitively planning, agents can still improve their behavior through learning and can even develop certain conventions when they serve mutual self-interest such as symmetry breaking in coordination games. Again we consider two mechanisms. The first mechanism improves agent i's model of agent $-i$ by using the frequency of i's previously successful behavior to modify the state-action Q-values of $-i$ such that previously successful action are more likely to occur again. This model-based mechanism, improves agent i's policy since she will best-respond to a more accurate model of agent $-i$. The second mechanism is model-free reinforcement of player i's state-action Q-values when player i herself successfully reaches a goal. Neither of these norms trivially generalize across different planning environments that don't share states.

Coordinating Cooperation and Competition

Finally, we describe how agents can use both the cooperative and competitive modes of planning to decide whether to cooperate or compete. Since these modes of planning abstract away the details of cooperation and competition, high-level strategic planning can use these low-level planners without considering their details. Agents first use these planning modes to infer the high-level intention I of the other player (i.e., their planning mode) using Bayesian theory-of-mind: $P(I|D) \propto P(D|I)P(I)$ where $P(D|I)$ are just the cooperative or competitive policies. This probabilistic approach is justified because intentions can be ambiguous. For instance, when both agents reach the goal in a coordination game it could just be because of luck so the behavior isn't very diagnostic of the intention. Yet in social dilemma only the cooperative intention is consistent with behavior where both reach the goal. Using these inferred strategic intentions, a high-level planner can take a simple and intuitive form such as reciprocal cooperation (e.g., tit-for-tat) or reinforcement learning at the level of strategy rather than actions (Fudenberg & Levine, 1998).

Behavioral Experiments

We developed client/server software that allows for real-time interactions between two participants randomly matched through mTurk. All participants went through a short single player tutorial that familiarized them with the controls of the games, the dynamics of the game environment, the costs of movement and value of the goals. After the tutorial, pairs of

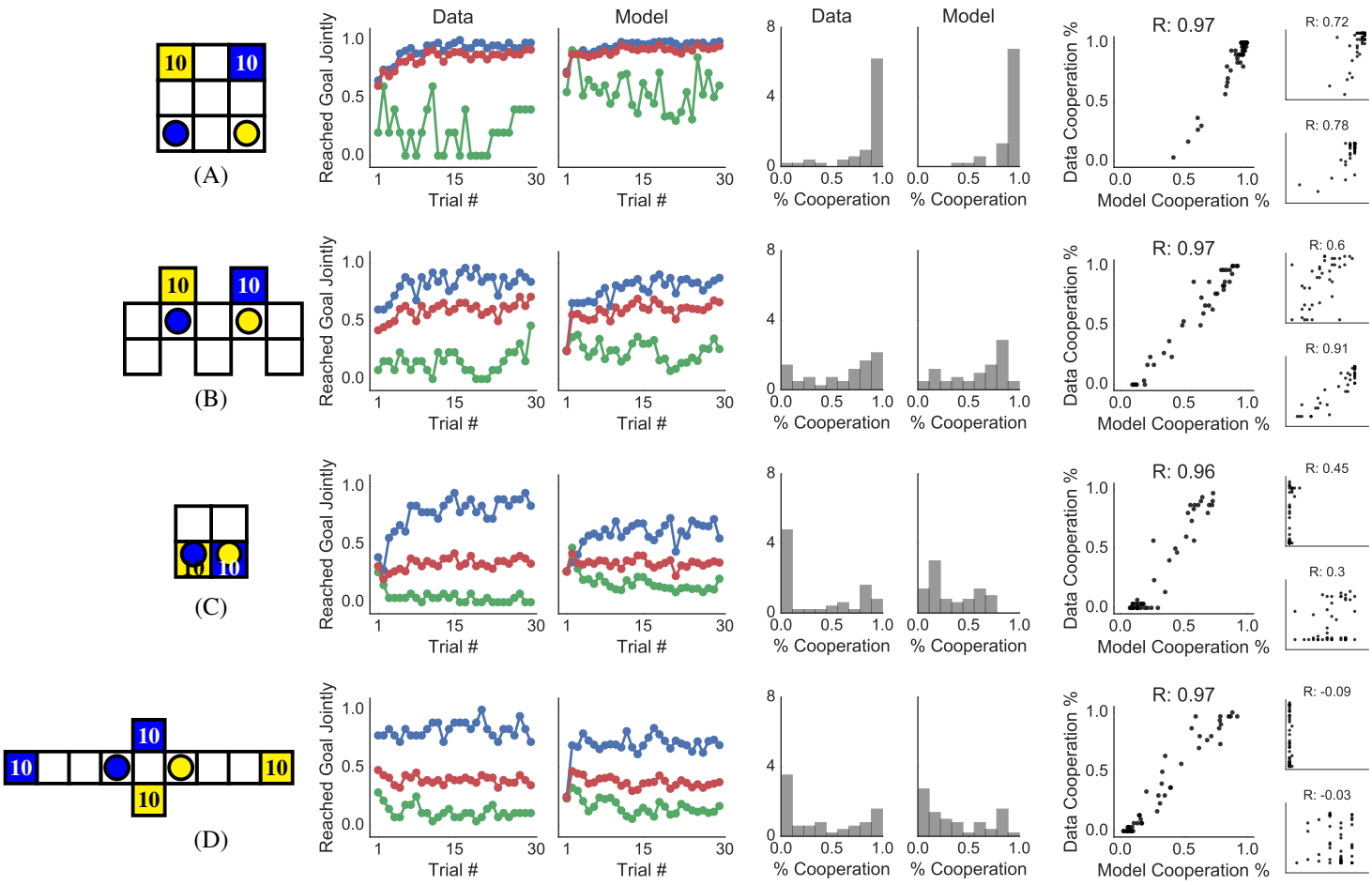


Figure 3: Participant data and model predictions for four environments. Each row shows data and model predictions for the environment in column 1 which was repeated 30 times. Rows 1 and 2 are coordination games and rows 3 and 4 are social dilemmas. Column 2 shows the average rate of cooperation for each round of play averaged over the high-cooperating cluster of participants (blue), low-cooperating cluster of participants (green) and all participants (red). Column 3 are histograms of the proportion of cooperation for all pairs of participants. Column 4 quantifies the model predictions where each point represents the frequency of cooperation for a given dyad observed in the data and as predicted by the model. The inset shows correlations of the two lesioned models with the same human data: (top) only compete (bottom) only cooperate.

participants were matched together and played 30 rounds of the same game with the same partner. Subjects were not told the exact number of rounds they would play together in order to prevent horizon effects from backward induction. Once both participants submitted moves, the game state and score were updated and the process continued until the end of the round. Participants had 30 second for each move and the game ended if a participant exceeded their 30 second time bank two moves in a row. We only analyzed data from complete interactions where the pair of participants completed all 30 rounds of the game together. All experiments were incentivized with bonuses proportional to the number of points accumulated.

To compare model predictions with human behavior, we first focused on analyzing whether or not both players reached a goal on a given round, a behavioral signature of cooperation in these games. For each pair of participants, the model observes the interaction in the previous rounds, performs inference on the latent high-level goal and social norms, and samples a prediction for the behavior of the pair in the next round. We compare this sampled prediction with actual human behavior to assess model performance. The same model parameters were used for all pairs of participants.

Figure 3 shows the results of the behavioral experiments and the model predictions for four environments ($\approx$ 50 participant pairs per environment), two coordination games and two social dilemma. Since model predictions were made at the level of each pair of participants, averaging the behavior and model predictions across dyads obscures individual differences in the dynamics of cooperative and competitive learning. To investigate the model predictions in a more fine-grained way, we used unsupervised clustering to split the pairs of participants into two group. In short, for each pair of participants we construct a 30-dimensional binary vector where each dimension corresponds to one of the 30 rounds. Each element is set to one if both participants reached a goal in the round corresponding to that dimension and set to zero otherwise. We ran K-means clustering with $K = 2$ which split

the data into a high-cooperating cluster and a low-cooperating cluster allowing for better visualization of the data and model prediction and gave some rough indication about the model ability to handle individual differences.

In all four environments, some of the pairs converged on a cooperative plan but the incentive structure of the game i.e., whether or not the game was a coordination game or social dilemma affected the likelihood that both participants jointly reached a goal. Overall, participants jointly reached the goal more frequently in coordination games than in the social dilemma. As shown in Figure 3 the model qualitatively captures the rate of cooperation and competition in both the high-cooperating cluster and the low-cooperating cluster as well as the average over all participants. Another coarse measure of behavior in these games is the distribution of the frequency of cooperative behavior across pairs of participants. In coordination games, the distribution was left-skewed and in social dilemma the distribution was right-skewed. These distributions were captured both qualitatively and quantitatively across these games by the model.

We compared the full model which included both modes of planning and strategic reasoning over those two modes with two lesioned models which just used one of the two planning modes. One lesioned model always used the competitive planning mode and the other lesioned model always used the cooperative planning mode. Overall, neither lesioned model could capture the rates of cooperation between the two clusters and qualitatively failed to explain the distribution of cooperative behavior in each game. Both lesioned models failed to predict the dynamics of strategic reasoning between cooperation and competition in social dilemma and had weaker correlation with participants' behavior in the coordination games.

Discussion

In this work we developed a hierarchical model of social planning to understand how humans coordinate their low-level action plans to realize high-level strategic goals such as cooperation and competition. We formalize cooperation and competition as abstract planning procedures over low-level actions. Both model-based and model-free learning can create social norms which facilitate robust and stable coordination. One of our main contributions is formalizing how cooperative norms can make cooperation more robust across environments, a key step for long-lasting collaborative endeavors. While we only had space to show a subset of our full results, we are currently looking at how agents use these planning programs and the norms that they learn to generalize cooperation to completely new environments with the same partner. We will also use these models to study how observers attribute cooperative and competitive intentions to other agents.

One interesting feature of the model is how an asymmetric w in the cooperative planner can break symmetries making successful coordination more likely. In future work we'd like to explore how priors on this parameter in social hierarchies might enable more effective teamwork e.g., boss-employee relations (Galinsky & Schweitzer, 2015). Finally, in our current paradigm, the desires of all agents are common knowledge. Investigating environments that require jointly inferring the goals of others and the plan needed to help realize a cooperative outcome will be examined in future work. By grounding strategic social reasoning in a theory of planning we can begin to investigate the mechanisms of joint intentionality and how these joint intentions enable the scale and scope of human cooperative behavior (Tomasello, 2014).

Acknowledgement This work was supported by a Hertz Foundation Fellowship, NSF-GRFP, the Center for Brains, Minds and Machines (CBMM), NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. We thank Alejandro Vientós, Banti Gheneti, and Paul Masterson.

References

Binmore, K. G. (1998). *Game theory and the social contract: just playing* (Vol. 2). Mit Press.

Bratman, M. E. (1993). Shared intention. *Ethics*, 97–113.

Bratman, M. E. (2014). *Shared agency: A planning theory of acting together*. Oxford University Press.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 861–898.

Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 1193–1235.

De Cote, E. M., & Littman, M. L. (2008). A polynomial-time nash equilibrium algorithm for repeated stochastic games. In *24th conference on uncertainty in artificial intelligence.*

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games* (Vol. 2). MIT press.

Gal, Y., & Pfeffer, A. (2008). Networks of influence diagrams: A formalism for representing agents' beliefs and decision-making processes. *Journal of Artificial Intelligence Research*, *33*(1), 109–147.

Galinsky, A., & Schweitzer, M. (2015). *Friend and foe: When to cooperate, when to compete, and how to succeed at both*. Random House.

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *J. Artif. Intell. Res.(JAIR)*, *24*, 49–79.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, *476*(7360), 328–331.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. (1990). On acting together. In *Aaai* (Vol. 90, pp. 94–99).

Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. *Trends in cognitive sciences*.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in cognitive sciences*, *17*(8), 413.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social philosophy and policy*, *10*(01), 69–89.

Sugden, R. (2003). The logic of team reasoning. *Philosophical explorations*, *6*(3), 165–181.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, *28*(05), 675–691.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12).

Effects of Gesture on Analogical Problem Solving: When the Hands Lead You Astray

Autumn B. Hostetter (Autumn.Hostetter@kzoo.edu)
Department of Psychology, 1200 Academy Street
Kalamazoo, MI 49006 USA

Mareike B. Wieth (mwieth@albion.edu) and Katlyn D. Foster
Department of Psychological Science,
Albion, MI USA

Keith Moreno and Jeffery Washington
Department of Psychology, 1200 Academy Street
Kalamazoo, MI 49006 USA

Abstract

We investigated the role of speech-accompanying gestures in analogical problem solving. Participants attempted to solve Duncker's (1945) *Radiation Problem* after reading and retelling a story that described an analogous solution in a different domain. Participants were instructed to gesture, instructed not to gesture, or given no instructions regarding gesture as they retold the story. Participants who were instructed to gesture as they retold the analogous story were more likely to mention perceptual details in their description and less likely to apply the analogous solution to the problem than participants who were instructed not to gesture. These results suggest that gestures can be detrimental to analogous problem solving when the perceptual elements of a story are irrelevant to its schematic similarity with a problem.

Keywords: gesture; analogical reasoning; problem solving

Introduction

Analogical problem solving involves using an analogous source, such as a military story about a general overtaking a fortress by dividing his troops, to solve a target problem, such as Duncker's (1945) Radiation problem in which rays are divided to bombard a tumor (e.g., Gick & Holyoak, 1980). One of the most difficult steps in analogical problem solving is noticing that there is a connection between the source story and target problem. Noticing the connection is particularly difficult when the source and the target are from different domains (e.g., a military story and a medical problem) and involve very different objects (e.g., armies and fortresses vs rays and tumors) (see Anolli, Antonietti, Crisafulli, & Cantola, 2001). In order to notice the connection despite such differences, problem solvers must schematitize the goals and relations between objects in each (Gentner & Smith, 2013). A military story can be applied more easily to a medical problem if both are being considered as instances where there are similar goals (e.g., overcoming a centrally located target) and possible means (e.g., dividing).

Such schema formation is more likely the less problem solvers focus on the objects involved in any particular source. Encountering multiple source stories, for example, increases the likelihood that a schema will be formed (Gick & Holyoak, 1983) as does seeing an informative diagram depicting the important features of the schema (Beveridge & Parkins, 1987; Chen, 1995). Moreover, pictures depicting a solution to a target problem are only helpful to problem solvers when they contain the same objects as those involved in the problem (Chen, 1995), perhaps because they cue participants that there is a similarity and encourage the formation of a schema that can apply to both target and source. In contrast, when the objects in the source are different from those in the target problem, focusing on those objects deters from noticing the connection with the target (Chen, 1995; Keane, 1987) and prevents participants from forming a schema that is applicable to both target and source.

While focusing on the objects in a source can be detrimental to problem success in a different domain, activating kinesthetic information that embodies the relevant schema (e.g., dividing, converging) can be beneficial. Catrambone, Craig, and Neressian (2006) found that acting out the military story with wooden blocks increased the likelihood of noticing the connection to the radiation problem. They argue that acting out the story with blocks encouraged the formation of a schema that included kinesthetic information about converging forces, and was therefore more general than a direct representation of the objects involved in the military story. Further, kinesthetic information can encourage schema formation even when there is no source story at all (Thomas & Lleras, 2007; 2009). Thomas and Lleras (2007) found that participants were more likely to solve the radiation problem when they produced eye movements in an unrelated task that mirrored the problem's solution (crossing in and out from the skin to the tumor many times, as multiple smaller rays would do) than when they produced unrelated eye movements.

It appears then that body movements can affect whether a problem solver forms a schema that can be applied to the radiation problem. In the present study, we explore the

effects of a very specific kind of body movement in analogical problem solving – representational gestures. Representational gestures, hereafter simply gestures, are movements of the hand and body that coincide with speech. Gestures are actions, in that they involve the movement of the body; however, as has been argued by Goldin-Meadow and Beilock (2010), they are more than actions, too, because they represent ideas in a way that actions often do not. Gestures have been described as outward manifestations of mental representations that are deeply rooted in the sensorimotor system (Hostetter & Alibali, 2008). When speakers produce gestures, they are using their motor system to represent an idea that is already drawing on the perceptual system (Hostetter & Skirving, 2011). Thus, producing gestures is a natural consequence of the high involvement of the sensorimotor system in concept representation (e.g., Glenberg, 2010).

Further, there is increasing evidence that producing a gesture is not only epiphenomenal, but can actually strengthen underlying perceptual representations in the speaker's mind (e.g., Cook, Yip, & Godin-Meadow, 2010) and make it more likely that gesturers will focus on perceptual information as they solve problems. For example, children who gesture as they solve Piagetian conservation problems are more likely to describe perceptually present aspects of the stimuli (e.g., the height of the objects) than when they do not gesture (Alibali & Kita, 2010). In addition, Alibali, Spencer, Knox, and Kita (2011) found that adults who gestured while solving problems about gear movements were more likely to perseverate on perceptual and motor strategies than speakers who did not gesture.

Because gestures strengthen underlying perceptual representations in the speaker's mind, they have the potential to help or hurt schema formation, just as focusing on objects more generally can either help or hurt (Chen, 1995). When the objects being gestured about are relevant to a problem, gestures may strengthen problem solvers' focus on those objects, thereby making it more likely that they will notice a connection with the problem and form a schema that can be applied to both. On the other hand, when the objects being gestured about are irrelevant to a problem, strengthening the perceptual representations of those objects may make it harder for participants to look past the differences between the source and the target in order to form a schema that can be applied to both (i.e., a fortress and a tumor).

In support of this, Beilock and Goldin-Meadow (2010) found that participants who gestured about the weight of the discs during an explanation of their solution to the Tower of Hanoi had more trouble solving a new version of the problem when the disc weights had been changed than participants who did not gesture. Although the solution process is the same regardless of the weight of the discs involved, by gesturing about the disc weights, problem solvers strengthened the representation of weight in their mind, thereby making it more difficult to generate a schema about how to solve the Tower of Hanoi that was independent of this perceptual element. Similarly, Cooperrider and Goldin-Meadow (2014) found that speakers who gestured during an explanation of a source story were less likely to solve an analogous problem than those who did not gesture. However, they did not directly manipulate gesture in their study, making it difficult to know if the presence of gesture was a side effect of already rich perceptual representations or actually a causal mechanism that focused speakers' representations on perceptual elements.

In the present experiment, we examine whether gesturing about a source story affects the likelihood that a problem solver will notice the connection to a target problem in a different domain. If gestures are much like other kinds of actions, then they should encourage a representation of the source story that is schematic, thereby improving the likelihood of noticing the connection between the source and the problem. Much like the eye movements studied by Thomas and Lleras (2007) and the acting-on-blocks movements studied by Catrambone et al. (2006), producing speech accompanying gestures about the military story could encourage speakers to form a schematic representation of converging forces that in turn makes it easier to notice the similarity with a medical problem. On the other hand, if gestures are different from actions because of their tight connection to perceptual ideas (e.g., Goldin-Meadow & Beilock, 2010; Hostetter & Alibali, 2008), gesturing about the military story could encourage a representation that is less schematic and more focused on the specific objects present in the source story. Under this view, gesturing may make it less likely that speakers will schematitize the source story, and thus less likely that they will notice its similarity with the problem.

To examine these two competing hypotheses, we gave participants multiple attempts to solve Duncker's (1945) *Radiation problem*. In between attempts, participants read and retold an analagous military story, under the ruse of participating in an unrelated study. Participants were either instructed to gesture, instructed not to gesture, or given no instructions regarding gesture during their retells.

Method

Participants

The final sample consisted of 72 undergraduates (44 female) at two small colleges who volunteered to participate in exchange for extra credit. An additional 20 participants were not included because the participant solved the radiation problem during the pre-check (before having a chance to tell the story), the experimenter made a mistake in the protocol, or the camera was not positioned well enough to capture the participants' gestures during the retells.

Stimuli

The military story. The story, taken from Gick and Holyoak (1980), describes a general who raises an army to overtake a fortress located in the center of a town. The

general learns that there are mines on all of the roads approaching the fortress. The mines will detonate if any large group of people crosses, but small groups of people can pass safely. The general splits his army into smaller groups that approach the fortress via different roads. The groups are not large enough to detonate the mines, but together, they are large enough to overtake the fortress when they converge in the center. Because speakers are more likely to gesture when they have seen an image of what they are describing (e.g., Hostetter & Skirving, 2011), the story was printed with a picture underneath the text of the fortress surrounded by roads.

The medical problem. Each participant was presented with Duncker's (1945) *radiation problem*. Briefly, the problem describes an inoperable tumor that can be destroyed with a certain kind of ray. However, at sufficient intensities to destroy the tumor, the ray will also destroy the healthy tissue around the tumor. Participants are asked to think of possible ways to destroy the tumor with the ray without affecting the healthy tissue. Although there are several possible solutions to the problem, we were only interested in the *convergence solution* that is suggested by the military story. Specifically, the ray can be split into several smaller rays that can pass safely through the healthy tissue, but still converge on the tumor with enough intensity to destroy it.

Procedure

Participants arrived individually and were told that the study was about the effects of taking breaks on problem solving. Participants first had 3 minutes to attempt to solve the medical problem as a pre-check to make sure that they did not already know the solution, either because others had told them about it or because they had encountered it in a course or in another study. Participants who generated the convergence solution during the pre-check were not included in the study.

Following the pre-check, those participants who had not generated the convergence solution were told that they would be given a break before attempting the problem again. During the break, they would be involved in another study about how people remember and communicate information. Participants were told that they would retell a story in their own words to a video camera. They then spent two minutes reading the military story.

Participants were then randomly assigned to one of three conditions. In the *Instructed Gesture* condition, participants were told to include hand gestures to depict important aspects of the story. In the *Restricted Gesture* condition, participants were told that, while they might be tempted to include hand gestures to depict important aspects of the story, they should not do so. The experimenter suggested that participants sit on their hands to remind themselves not to gesture. In the *No Instruction* condition, the experimenter did not mention gesture in the instructions and simply told participants to include important aspects of the story as they retold it. In all three conditions, the experimenter produced the same scripted hand gesture when delivering the instructions.

In all conditions, participants took as long as they wanted to retell the story. When they were finished, they were given another 3 minutes to generate solutions to the medical problem. The experimenter left the room during this time. Upon returning, the experimenter asked the participants to explain their solutions to the problem. If participants described the convergence solution, their experimental session was ended and they were debriefed.

If they did not describe the convergence solution, they were told that they would have one more opportunity to solve the problem after taking another short break, during which they would retell the military story a second time to the video camera. Participants were reminded of their gesture instructions before beginning. Participants were then given three more minutes to generate solutions to the medical problem. Although this procedure differs from the original protocol pioneered by Gick and Holyoak (1980), giving participants multiple opportunities to retell the story and work on the problem is similar to the procedure of other studies that have found an effect of movement on problem solving (e.g., Thomas & Lleras, 2009).

Results

All participants who were instructed to gesture as they retold the military story produced at least one representational gesture (n = 25). In addition, 20 of the 22 participants who were given no instructions regarding gesture spontaneously gestured as they retold the story, and 3 of the 25 participants who were told not to gesture produced at least one representational gesture despite having been asked not to. For all analyses reported here, the three individuals who did not follow the gesture restriction instructions have been excluded. We had hoped that more participants in the no instruction condition would spontaneously choose not to gesture, thereby allowing a comparison of those who gesture

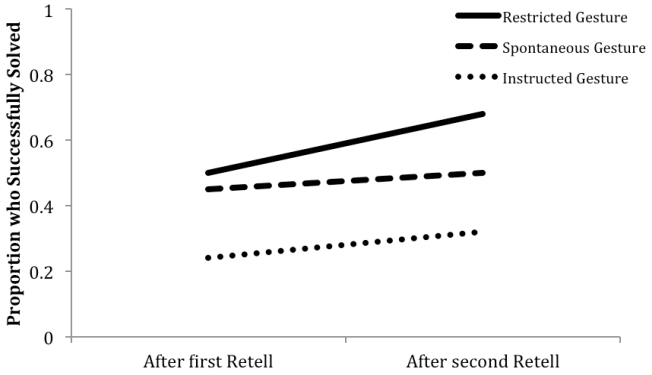


Figure 1. The proportion of participants in each condition who successfully generated the convergence solution to the Radiation problem after each retelling of the General story.

Table 1. Percentage of participants in each condition who mentioned each spatial element in their retelling of the story

Element	Gesture Instructed	Spontaneous Gesture	Gesture Restricted	χ^2
Fortress is centrally located	68	45	32	6.34*
Roads radiate outward	52	40	54	1.00
Army is large	52	15	27	7.36*
Army split into smaller groups	84	85	82	0.08
Groups come from different directions	68	70	32	8.29*
Groups meet in the middle	11	0	14	2.84

Note. * $p < .05$

spontaneously and those who do not. However, given that all but two participants in the no instruction condition gestured, such an analysis is not possible. We have therefore eliminated the two individuals who did not gesture in the no instruction condition from further analysis. Thus, the comparisons reported here are between participants who did not gesture after being instructed not to (n = 22), participants who gestured after being instructed to (n = 25), and participants who gestured spontaneously with no gesture instructions (n = 20).

Figure 1 shows the cumulative proportion of participants in each condition who generated the convergence solution to the medical problem following each retell. We analyzed the data in a mixed logistic regression model that included condition (*Instructed Gesture, Restricted Gesture, No Instruction*) as a fixed factor and college as a random intercept. Because gesture condition was manipulated between subjects and each participant only solved one problem, subject is not included as a random effect. We included a random intercept for college but not slope because the model failed to converge when slope was included.

For solution success after the first retell, there was no significant effect of condition, $\chi^2(2, N = 67) = 4.78$, $p = .09$. However, there was a significant effect of condition when considering success rate overall, $\chi^2(2, N = 67) = 6.75$, $p = .03$. Participants who were instructed not to gesture were twice as likely (68%) to generate the convergence solution to the medical problem by the end of the experiment than participants who were instructed to gesture (32%), $\beta = 1.64$, $SE = 0.65$, $z = -2.51$, $p = .01$. Participants who gestured spontaneously had a solution rate (50%) intermediate to and not significantly different from the other two groups.

Our hypothesis is that participants who gesture are less likely to solve the problem because of an increased focus on spatial and perceptual aspects of *The General Story*. To test this hypothesis, we coded each participant's retells for whether they mentioned six specific perceptual details regarding the spatial layout of the story (see Table 1). A one-way analysis of variance revealed significant differences in the total number of spatial details mentioned by participants during their retells across the three conditions, $F(2, 66) = 3.89$, $p = .025$. Participants who were instructed to gesture included significantly more spatial details in their retells than participants who were instructed not to gesture or participants who gestured spontaneously. As can be seen in Table 1, participants who were instructed to gesture were more likely than other participants to mention that the fortress was centrally located, that the army was large, and that the small armies approached the fortress from different roads. As predicted, instructing participants to gesture appears to increase the attention they pay to spatial details of the story as they are retelling it. Interestingly, however, participants who chose to gesture spontaneously, without specifically being told to gesture, did not focus on spatial details any more than participants who were told not to gesture.

Given that participants who were instructed to gesture were also the most likely to describe spatial details in their story, could their decline in success on the problem be due to their spatially rich descriptions, rather than to the gestures that accompanied them? In particular, two of the spatial details are analogous to the key transformations involved in the dispersion solution to *The Radiation Problem*: the army is split into multiple smaller armies and the smaller armies approach the fortress from different roads. To examine the possibility that it is mentioning these details, rather than gesturing about them per se, that affects problem success, we ran mixed logistic regression models predicting problem success based on whether each detail was mentioned. There was no effect for either detail (split: $\chi^2(1, N = 67) = 0.53$, $p = .47$; approach: $\chi^2(1, N = 67) = 0.06$, $p = .80$). It appears that mentioning these details is not predictive of problem success.

Finally, we more closely examined the gestures produced by participants to see if gesturing about particular concepts affected problem success. We coded the gestures produced by participants for whether they represented one of two concepts that are important to the dispersion solution: multiplicity and convergence. A gesture was coded as representing multiplicity when it indicated multiple places in space either by using two hands simultaneously or alternately or by using one hand sequentially to indicate three or more points in space. Such gestures typically occurred in the context of describing the multiple roads around the fortress or the many small army groups. A gesture was coded as representing convergence if it indicated moving towards the center from multiple radial locations, either by moving two hands inward simultaneously, by moving one hand from radial to center multiple times, or by drawing the fingers of one hand inward. These gestures occurred most often with speech about the armies moving towards the fortress along the different roads.

We considered whether each participant produced a gesture about multiplicity or about convergence in separate mixed logistic regression models. Although not significant by conventional standards, participants who gestured about the concept of multiplicity were less likely to solve the problem than participants who did not, $\chi^2(1, N = 67) = 3.68$, $p = .06$. Specifically, participants who gestured about multiplicity as a result of being told to gesture had worse performance than participants who were told not to gesture, $\chi^2(1, N = 47) = 4.92$, $p = .03$. In contrast, participants who gestured about multiplicity spontaneously did not differ from those who were told not to gesture, $\chi^2(1, N = 42) = 1.32$, $p = .25$. The same pattern emerged when convergence gestures were considered; participants who gestured about the concept of convergence were significantly less likely to solve the problem than participants who did not, $\chi^2(1, N = 67) = 4.32$, $p = .04$, but the effect is specific to individuals who were told to gesture compared to those who were told not to gesture, $\chi^2(1, N = 47) = 7.23$, $p = .007$. There is no difference in solution rate between those individuals who spontaneously gestured about convergence and those who were told not to gesture, $\chi^2(1, N = 42) = 0.72$, $p = .40$.

Discussion

Being instructed to gesture about the military story reduced the likelihood that participants would generate the convergence solution to the medical problem. We posit that this effect occurred because speakers who were told to gesture were particularly likely to focus on the objects involved in the military story, a focus that made it more difficult to form a schema that could also be applied to the medical problem. Indeed, speakers who were told to gesture were more likely to mention specific spatial details about the objects in their description than other participants.

The difference between spontaneously producing a gesture and producing a gesture after being instructed to do so is noteworthy. Previous research showing the effects of gesture on problem solving has not specifically compared the effects of gestures that were instructed with those that were spontaneous, though in many of these studies, participants were specifically instructed to use their hands during the explanation task (e.g., Beilock & Goldin-Meadow, 2010). It appears that, at least in the case of analogical problem solving studied here, being instructed to gesture produces or exaggerates the effect. There are several reasons why this could occur.

First, being told to gesture likely makes participants more consciously aware of their gestures. Having the conscious intention to produce a gesture could strengthen the underlying mental simulation that gives rise to gesture (e.g., Hostetter & Alibali, 2008). As a result, speakers who intentionally gesture may have richer mental representations that are more strongly grounded in perceptual and motor simulations than speakers who avoid gesturing or who gesture without thinking about it. Alternatively, there may be differences in the cognitive effects after the gesture is produced. Goldin-Meadow and Beilock (2010) argue that gestures bring perceptual information to the forefront of speakers' thinking; this may only occur when speakers are consciously aware of their gestures, as when they have specifically been told to produce them.

Regardless of why being told to gesture hurts, the present results suggest that speech-accompanying gestures do not have the same effect on problem solving as the movements studied by Thomas and Lleras (2007, 2009) or the actions studied by Catrambone et al. (2006). Specifically, Thomas and Lleras found that eye movements, movements that did not represent any particular concept, aided problem solving in their studies. However, speech-accompanying gestures are different from these movements because gestures are meaningful to the speaker; in order to be produced, they must represent something. When a speaker produces a gesture about the concept of multiple roads radiating outward from the fortress, the gesture corresponds to the specific mental image the speaker has in mind. Thus, as Goldin-Meadow and Beilock (2010) argue, gestures are a special kind of action. By virtue of representing images, gestures act to strengthen those images in the speaker's mind, and can make it more difficult to form a general schema that transcends the objects represented in those particular images.

Our results are also different from the findings of Catrambone et al. (2006), who found that acting out the military story using blocks aided analogical problem solving over simply retelling the story. In contrast, acting out the story using gestures in the present study impaired analogical problem solving. There are of course many methodological differences between the two studies that could account for the difference in findings, but there may also be a meaningful difference between the gestures studied here and the block movements studied by Catrambone et al. Using blocks to act out the story requires that the storytellers think abstractly about how the story elements can be represented with blocks, which may encourage thinking schematically about the relational elements of the story rather than the objects. In contrast, producing gestures about the story elements may not require the storyteller to engage in an abstract mapping between story and hand because gestures are used so ubiquitously and automatically to express meaning (e.g., Hostetter & Alibali, 2008).

Finally, it is possible that rather than gestures being detrimental to problem success, the act of not gesturing was beneficial. That is, perhaps being told not to gesture freed up cognitive resources that could then be used to think schematically about the story (e.g., Waltz, Lau, Grewal, & Holyoak, 2000). It is difficult for us to rule this possibility out definitively. However, we find the explanation untenable for two reasons. First, when given no instructions regarding gesture, most participants spontaneously produced them, suggesting that gesturing while retelling the story was natural. Thus, refraining from gesturing was likely more difficult than gesturing. Second, previous research has shown that producing gestures reduces cognitive load in

other tasks (e.g., Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001). Thus, there is no evidence to support the idea that gesture production utilizes more resources than speech alone. Rather than utilizing cognitive resources, we argue that producing gestures focused speakers' attention on the concrete perceptual and motor elements of the story, thereby reducing their likelihood of forming an abstract schema that could be applied to the problem.

This explanation suggests several possible avenues for future research. First, whether gestures about the source help or hurt should depend on the similarity of objects and perceptual features between the source and target. Focusing on the objects involved in the source is not always detrimental; when there are similar objects involved in the source and target, thinking about the objects can actually help problem solvers notice the connection (e.g., Chen, 1995). Thus, gesturing about a source story that involves rays might actually improve the likelihood of noticing the connection with a target problem that is also about rays. We are currently testing this prediction.

Second, the participants in the present study read only a single source story. Previous research has shown that being exposed to multiple source stories increases the chances that problem solvers form a schema that can be applied more generally to a problem (e.g., Gick & Holyoak, 1983). Perhaps if problem solvers gestured about multiple source stories, they would notice the similarity in their gestures across stories and be more likely to form the abstract schema of converging forces. Cooperrider and Goldin-Meadow (2014) tested this prediction, but found no evidence to support it. In fact, they found that speakers who gestured about multiple source stories were less likely to generate the convergence solution than speakers who did not gesture, paralleling the effect we found here.

In conclusion, gestures have been shown to play a role in solving problems of various types (e.g., Alibali et al., 2011; Beilock & Goldin-Meadow, 2010). We have extended this work to examine the role of gesture in analogical problem solving and provided evidence that gestures make it *more difficult* to notice a similarity between a source and target that are in different domains. These findings suggest that when it is more important to focus on schematic relationships rather than perceptual properties, problem solvers may be best off keeping their hands still. For example, in a Physics problem involving physical forces on particular objects, gesturing about the objects in the problem may focus participants' attention on those particular objects, rather than on the more general forces at work that are essential to understanding how to solve the problem.

References

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10*, 3-28.

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science, 22*, 1138-1144.

Anolli, L., Antonietti, A., Crisafulli, L., & Cantoia, M. (2001). Accessing source information in analogical problem-solving. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 54,* 237-261.

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological Science*.

Beveridge, M., & Parkins, E. (1987). Visual representation in analogical problem solving. *Memory & Cognition, 15,* 230-237.

Catrambone, R. C., Craig, D. L., & Nersessian, N. J. (2006). The role of perceptually represented structure in analogical problem solving. *Memory & Cognition, 34,* 1126-1132.

Chen, Z. (1995). Analogical transfer: From schematic pictures to problem solving. *Memory & Cognition, 23,* 255-269.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2010). Gesturing makes memories that last. *Journal of Memory and Language, 63*, 465-475.

Cooperrider, K., & Goldin-Meadow, S. (2014). The role of gesture in analogical problem solving. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society (pp. 2068-2072). Austin, TX: Cognitive Science Society.

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58*, 270.

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In D. Reisberg (Ed.) *The Oxford Handbook of Cognitive Psychology,* (pp. 668-681). New York, NY: Oxford University Press.

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 586-596.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306-355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on Psychological Science, 5*, 664-674.

Goldin-Meadow, S., Nusbaum, H., Kelly, S., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science, 12,* 516-522.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15*, 495-514.

Hostetter, A. B., & Skirving, C. J. (2011). The effect of visual vs. verbal stimuli on gesture production. *Journal of Nonverbal Behavior, 35*, 205-223.

Keane, M. (1987). On retrieving analogues when solving problems. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 39,* 29-41.

Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review, 14*, 663-668.

Thomas, L. E., & Lleras, A. (2009). Swinging into thought: Directed movement guides insight in problem solving. *Psychonomic Bulletin & Review, 16*, 719-723.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition, 28,* 1205-1212.

From *uh-oh* to *tomorrow*
Predicting age of acquisition for early words across languages

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, and Michael C. Frank
{mikabr, yurovsky, marchman, mcfrank}@stanford.edu
Department of Psychology, Stanford University

Abstract

Why do children learn some words earlier than others? Regularities and differences in the age of acquisition for words across languages yield insights regarding the mechanisms guiding word learning. In a large-scale corpus analysis, we estimate the ages at which 9,200 children learn 300-400 words in seven languages, predicting them on the basis of independently-derived linguistic, environmental, and conceptual factors. Predictors were surprisingly consistent across languages, but varied across development and as a function of lexical category (e.g., concreteness predicted nouns while linguistic structure predicted function words). By leveraging data at a significantly larger scale than previous work, our analyses highlight the power that emerges from unifying previously disparate theories, but also reveal the amount of reliable variation that still remains unexplained.

Keywords: language acquisition; word learning; development

Introduction

Word learning is one of the central challenges of language acquisition. Learners must integrate multiple information sources to map the word forms they hear onto representations of their meanings. Across many laboratory experiments and small-scale models, a number of strategies have emerged as plausible components of word learning, including tracking co-occurrence statistics between words and referents to deduce word meaning across situations; attending to social cues like pointing and eye gaze; relying on biases, such as a basic level category bias; and drawing on knowledge of relations between words to use known meanings to learn new ones.

Each of these strategies has been reliably demonstrated in the constrained learning context of the laboratory, indicating that they are possible parts of the word learning process. However, small-scale experimental studies typically do not tell us whether these strategies operate uniformly across children, ages, and languages. It is also difficult to explore how strategies interact to create the longer-term dynamics of vocabulary acquisition. How do the various strategies differ in their relative contributions? And how does their influence change over the course of development?

Our approach to addressing these questions is to use large-scale vocabulary development data to examine these interactions. By aggregating across a large number of children, we can look past individual differences in acquisition to investigate not only which words are relatively easy or hard to learn, but also what features affect their acquisition. For example, distributional learning strategies rely critically on frequency. Thus, to make a first assessment of the contribution of distributional learning, we can examine the relationship between the age at which words are typically acquired and word frequency in child-directed speech.

Such an approach has revealed that in English, within a lexical category, words that are more frequent in speech to children are likely to be learned earlier (Goodman, Dale, & Li, 2008). And further studies have found evidence for semantic networks (Hills, Maouene, Maouene, Sheya, & Smith, 2009), neighborhood density (Stokes, 2010), iconicity (Perry, Perlman, & Lupyan, 2015), and linguistic distinctiveness (B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015) as additional predictors of age of acquisition (AoA), suggesting that they are likely contributors to vocabulary development. But these exciting findings are nevertheless limited in their generality because they used different datasets, focused on different predictors, and almost exclusively analyzed English data. It is thus impossible to compare the relative importance of the many relevant factors under consideration and to draw robust conclusions.

To remedy this issue, we present analyses based on data from Wordbank (wordbank.stanford.edu), an open repository of cross-linguistic language development data (Frank, Braginsky, Yurovsky, & Marchman, in press). By aggregating administrations of the MacArthur-Bates Communicative Development Inventory (CDI; Fenson, 2007), a family of parent-report vocabulary checklists, Wordbank provides large-scale vocabulary data based on analogous instruments from more than 40,000 children in 14 different language communities. Wordbank presents a novel resource for richer and more powerful analyses of vocabulary learning over development and across languages.

We integrate AoA estimates from Wordbank with characterizations of the word learning environment from the CHILDES database (MacWhinney, 2000) and elsewhere, a multiple data source methodology originated by Goodman et al. (2008). Building on this work, we examine interactions between a variety of linguistic, environmental, and conceptual factors. Using a similar approach on a high-density longitudinal corpus for a single English-acquiring child, Roy et al. found that the length, usage frequency, and mean length of the utterances in which it occurred were all predictive of a word's AoA. But due to the nature of the dataset, this analysis used production-based AoA estimates and was further limited by relying on data from only one child acquiring a single language.

Our work provides a complimentary analysis by using CDI comprehension data available in Wordbank to look at the earliest words that children learn across several different languages. We estimate AoA for approximately 400 words from CDIs in each of seven languages. We also estimate each

word's frequency and mean length of utterance (MLU) based on the set of utterances in CHILDES containing the word. Additionally, we obtain ratings of each word's concreteness, valence, arousal, and relevance to babies from previously collected norms. We use these measures to predict words' AoA, assessing the relative contributions of each, as well as how they change over development and interact with lexical category. Each of these analyses has the potential to advance our understanding of the theoretical underpinnings of word learning.

A first theoretically-motivated question is which lexical categories are most influenced by input-related factors, like frequency and utterance length, compared with conceptual factors like concreteness and valence. For example, the "division of dominance" theory suggests that nouns might be more sensitive to cognitive factors, while predicates and closed-class words might be more sensitive to linguistic factors (Gentner & Boroditsky, 2001). On the other hand, on syntactic bootstrapping theories (Gleitman, 1990), nouns are argued to be learned via frequent co-occurrence (operationalized by frequency) while verbs might be more sensitive to syntactic factors (operationalized here by utterance length), and neither would be particularly sensitive to conceptual complexity (Snedeker, Geren, & Shafto, 2007).

A second question of interest is the extent to which there is variability across languages in the relative importance of predictors. For example, are there differences in the importance of grammar-related factors in morphologically more complex languages like Russian and Turkish, compared with simpler ones like English? Differences of this type might be revealing of the degree to which learners face different challenges in different language environments. Alternatively, consistency may suggest the operation of similar learning mechanisms and strategies that are not as dependent on the complexities of phonology, morphology, and syntax in a particular language.

By incorporating a variety of theoretically-important factors, basing our analysis on a large sample of words and children, and building towards more cross-linguistic coverage, our study presents a more thorough investigation of the question of what properties determine words' learnability.

Data

We use CDI data from Wordbank to estimate the age of acquisition for words across seven languages: English, Italian, Norwegian, Russian, Spanish, Swedish, Turkish. We then ask what factors are most important for predicting this age of acquisition. Table 1 gives an overview of our data sources.

Estimating Age of Acquisition

To estimate the age at which words are acquired, we used vocabulary data collected using the MacArthur-Bates Communicative Development Inventory, specifically the Words & Gestures (infant) form for 8- to 18-month-olds. When filling out a CDI form, parents are asked to indicate whether their child understands and/or says each of around 400 words. From these data, for each word on the CDI, we computed the proportion of children at each age who were reported to understand the word. We then fit a logistic curve to these proportions using a robust generalized linear model (using the `robustbase` package in `R`) and determined when the curve crosses 0.5, i.e. at what age at least 50% of children are reported to understand the word. Following Goodman et al. (2008), we take this point to be each word's age of acquisition.

Language	CDI Items	CDI Admins	CHILDES Tokens
English	386	2,452	7,858,051
Italian	351	649	328,168
Norwegian	338	2,922	204,406
Russian	337	768	32,398
Spanish	333	778	1,458,327
Swedish	311	467	698,515
Turkish	327	1,115	44,347

Table 1: Dataset statistics

Predictors

Each of our predictors is derived from independent sources. For each word that appears on the CDI Word & Gestures form in each of our seven languages, we obtained an estimate of its frequency in child-directed speech, the mean length of utterances in which it appears in child-directed speech, its length in characters, and ratings of its concreteness, valence, arousal, and relevance to babies. Items such as *child's own name* were excluded. Example words for these predictors in English are shown in Table 2.

Frequency and MLU are measured relative to the word's language. But since existing datasets for conceptual ratings are primarily available for English, we mapped all words onto translation equivalents across CDI forms, allowing us to use the ratings for English words across languages. While neces-

Measure	Value	Words
aoa	min	mommy, bottle, peekaboo
	max	babysitter, teacher, naughty
frequency	min	living room, cockadoodledoo, grrr
	max	you, it, that
babiness	min	donkey, penny, jeans
	max	baby, bib, bottle
concreteness	min	how, now, that
	max	apple, ball, banana
mlu	min	cockadoodledoo, peekaboo, uh oh
	max	babysitter, when (question), day
arousal	min	shh, asleep, blanket
	max	naughty, money, scared
valence	min	sick, owie, ouch
	max	happy, hug, love
num	min	i, in, it
characters	max	cockadoodledoo, refrigerator, living room

Table 2: Examples of words with the lowest and highest values for age of acquisition and each predictor.

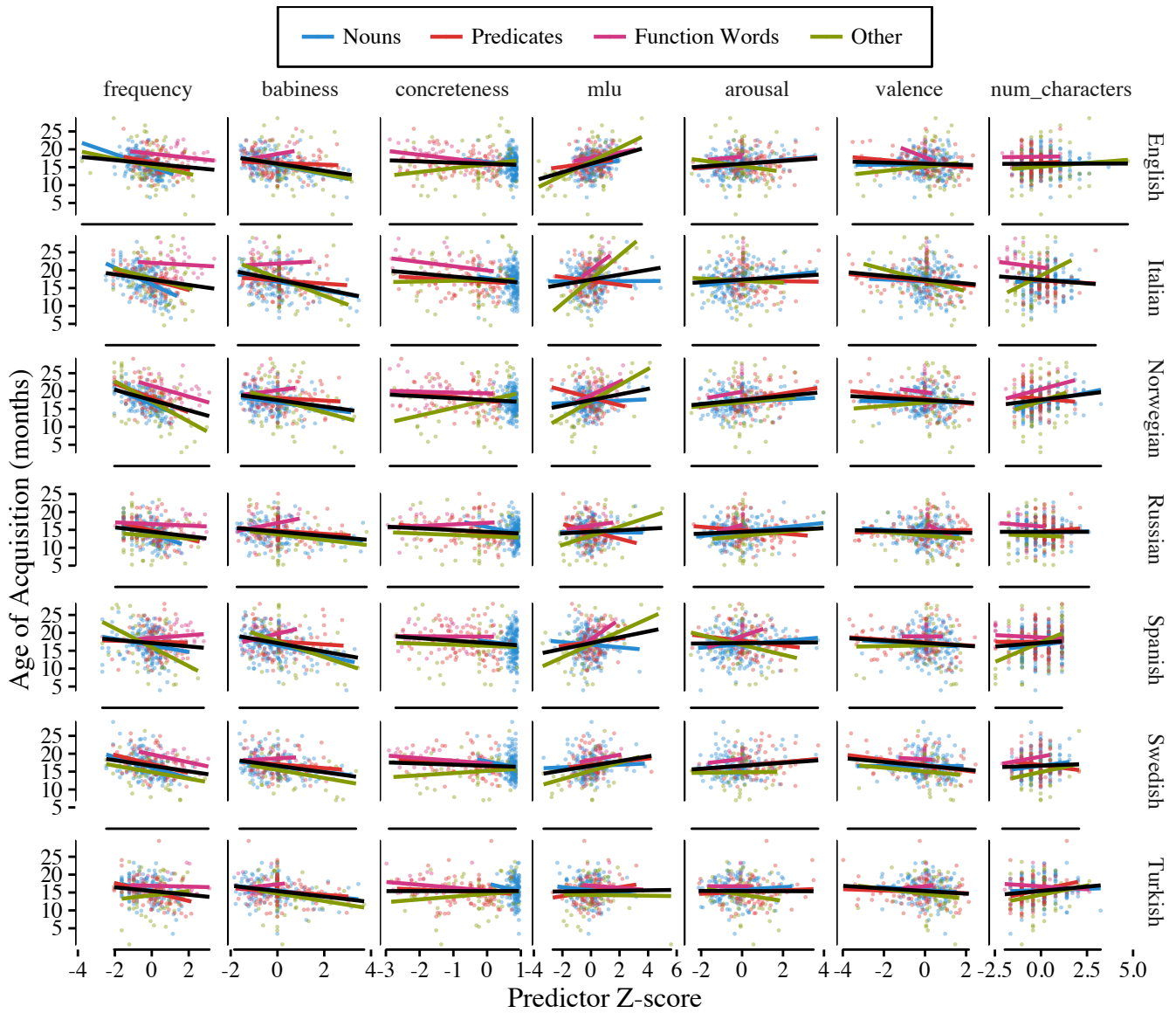


Figure 1: Relationship between predictors and AoA for each lexical category in each language. Each point represents a word, with lines indicating linear model fits for each lexical category (in colors) and overall (in black).

sarily imperfect, this method allows us to examine languages for which limited resources exist. Translation equivalents are available in the Wordbank database using the `wordbankr` package in `R` (Frank et al., in press).

Each numeric predictor was centered and scaled so that all predictors would have comparable units. Lexical category was determined on the basis of the conceptual categories presented on the CDI form (e.g., "Animals"), such that the Nouns category contains common nouns, Predicates contains verbs and adjectives, Function Words contains closed-class words, and Other contains the remaining items (following Bates et al., 1994).

Frequency For each language, we estimated word frequency from unigram counts based on all corpora in CHILDES for that language. Each word's count includes the counts of words that share the same stem (so that *dogs* counts as *dog*) or are synonymous (so that *father* counts as *daddy*). For polysemous word pairs (e.g., *orange* as in color or fruit), occurrences of the word in the corpus were split uniformly between the senses on the CDI. Counts were normalized to the length of each corpus and then log transformed.

MLU For each language, we estimated each word's MLU by calculating the mean length in words of the utterances in which that word appeared, for all corpora in CHILDES for that language. Words that only occurred in one utterance were excluded.

Length We computed the number of characters in each word in each language. While imperfect, this metric of length is highly correlated with number of phonemes and syllables (Lewis & Frank, under review).

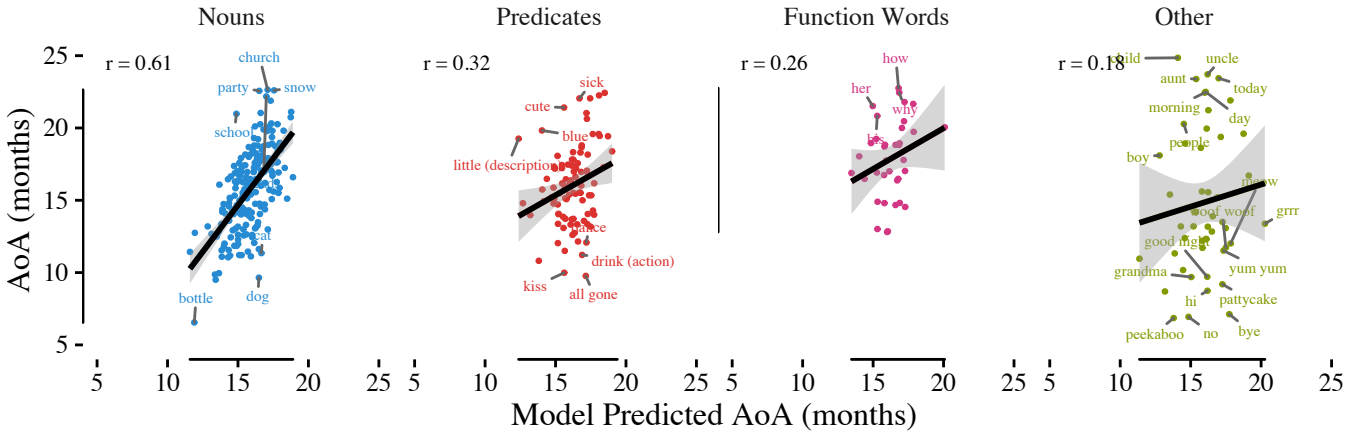


Figure 2: Comparison between the model-predicted and actual ages of acquisition for words in English. Words with an absolute error above 5 months are labelled for reference.

Concreteness We used previously collected norms for concreteness (Brysbaert, Warriner, & Kuperman, 2014), which were gathered by asking adult participants to rate how concrete the meaning of each word is on a 5-point scale from abstract to concrete. For the 120 CDI words that were not part of the collected norms, we imputed ratings from the mean of all CDI words' ratings.

Valence and Arousal We also used previously collected norms for valence and arousal (Warriner, Kuperman, & Brysbaert, 2013), for which adult participants were asked to rate words on a 1-9 happy-unhappy scale (valence) and 1-9 excited-calm scale (arousal). For the 119 CDI words that were not part of the collected norms (mostly function words), we imputed ratings from the mean of all CDI words' ratings.

Babiness Lastly, we used previously collected norms of "babiness," a measure of association with infancy (Perry et al., 2015) for which adult participants were asked to judge a word's relevance to babies.

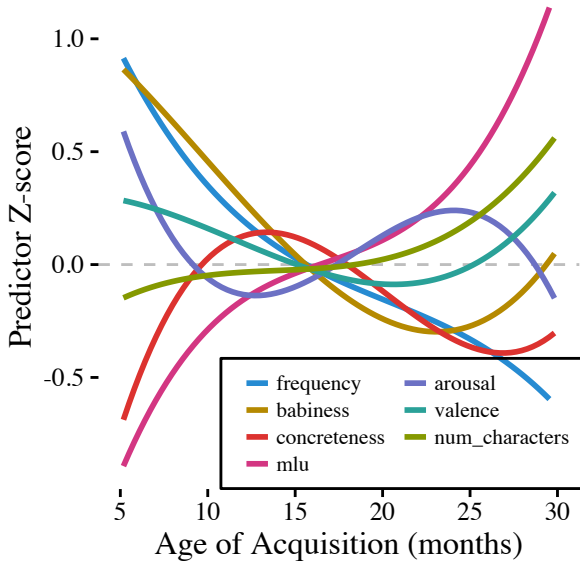


Figure 3: Trends in predictor values across development (for all items in all languages). Curves show best-fitting cubic regression.

Analysis

An overview of our entire dataset can be seen in Figure 1, which shows each word's estimated age of acquisition against its predictor values, separated by language and lexical category. We present three analyses of these data: 1) how predictor values change over development, 2) their relative contributions to predicting AoA, and 3) their interaction with lexical category.

Developmental Trajectories

To assess developmental trends, we examine how the values of each predictor change as a function of estimated AoA. Figure 3 shows these trajectories, with a cubic curve smoothing over all words. Words that are learned earlier are more frequent, higher in babiness, and appear in shorter utterances. Concreteness exhibits a U-shaped trajectory, with the earliest learned words actually being relatively abstract (e.g., social routines and animal sounds).

Predicting AoA

We fit a linear regression for each language's data, as well as a linear mixed-effects model with language as a random effect for all the data pooled across languages. For illustrative purposes, Figure 2 shows the predictions of the English model plotted against the empirical AoA estimates.

Figure 4 shows the coefficient estimate for each predictor in each language and for all languages combined. We find that frequency, babiness, concreteness, and MLU are relatively stronger predictors of age of acquisition, across languages and in the full, cross-linguistic model. Overall there is considerable consistency in how the predictors pattern in various languages, although with some interesting differences. For example, MLU in English appears to be unusually strong, while frequency in Spanish looks unusually weak. There is also variability in the overall fit of the models to the data, with some languages (e.g., Norwegian), having much more of the variance explained than others (e.g., Turkish).

A potential concern for comparing these coefficient estimates is predictor collinearity. Fortunately, in every language, the only high correlations were between frequency

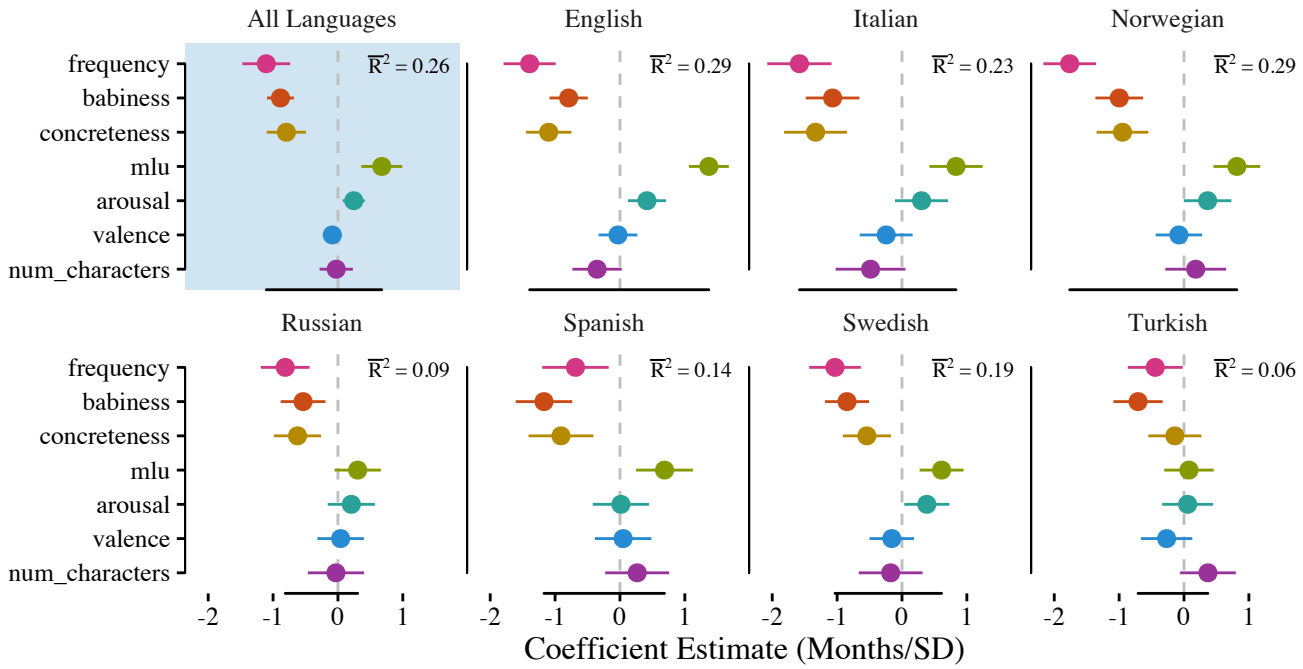


Figure 4: Estimates of predictor coefficients by language and for the all language model. Values above 0 indicate a positive relationship (i.e. words with higher MLU tend to have a higher AoA), while values below 0 indicate a negative relationship (i.e. words with higher frequency tend to have a lower AoA. Ranges indicate 95% confidence intervals.

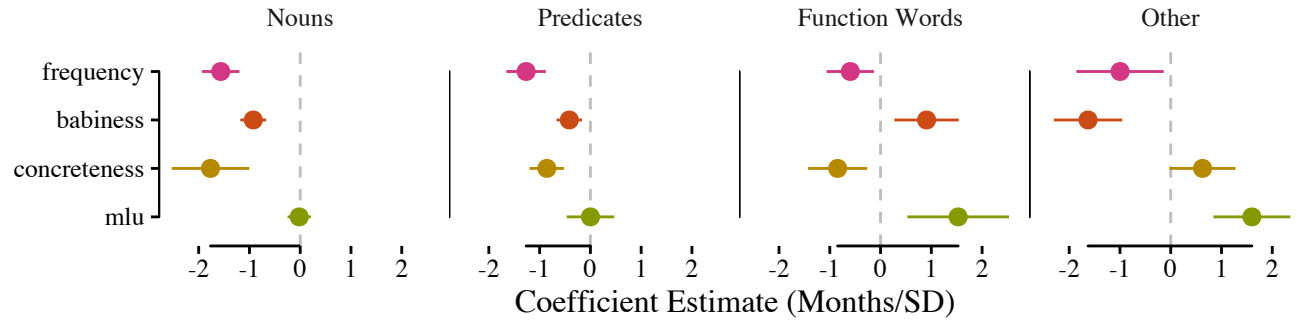


Figure 5: Estimates of predictor coefficients by lexical category, without any separation by language and omitting the three weaker predictors. Ranges indicate 95% confidence intervals.

and number of characters, a reflection of Zipf's Law (Zipf, 1935), and between frequency and concreteness, probably as a consequence of the complexity bias (Lewis & Frank, under review).

Lexical Category

Previous work gives reason to believe that predictors' relationship with age of acquisition differs among various lexical categories (Goodman et al., 2008). To investigate these effects, we separated our data by lexical category and fit separate linear mixed-effects models for each, limiting the predictors to the four that were significantly predictive overall. Figure 5 shows the resulting coefficient estimates. Frequency matters most for nouns and comparatively little for function words, while MLU is irrelevant for both nouns and predicates, but highly informative for function words and other items.

Discussion

What makes words easier or harder for young children to learn? Previous experimental work has largely addressed this question using small-scale experiments. While such experiments can identify sources of variation, they typically do not allow for different sources to be compared in detail. In contrast, observational studies allow the effects of individual factors (with frequency being the most common) to be measured across ages and lexical categories (e.g., Goodman et al., 2008). Scale comes at a cost in terms of detail, however, since the availability of both predictors and outcome data has been quite limited.

By including seven languages and as many predictors, our current work expands the scope of previous observational studies of age of acquisition. Our data show a number of patterns that confirm and expand previous reports. First, predictors changed in relative importance across development. For example, certain concepts that were more strongly associated with babies appeared to be learned early for children

across languages (as in Tardif et al., 2008).

Second, we found general consistency in predictor coefficients across languages (even as overall model fit varied, at least in part due to the amount and quality of data for different languages). This consistency supports the idea that differences in culture or language structure do not lead to fundamentally different acquisition strategies, at least at the level of detail we were able to examine.

Lastly, the predictors varied in strength across lexical categories. Frequent, concrete nouns were learned earlier, consistent with theories that emphasize the importance of early referential speech (e.g., Baldwin, 1995). But for predicates, concreteness was somewhat less important, and for function words, MLU was most predictive. Overall these findings are consistent with theories that emphasize the role of linguistic structure over conceptual complexity in the acquisition of other lexical categories beyond nouns (Gentner & Boroditsky, 2001; Snedeker et al., 2007).

Despite its larger scope, our work shares a number of important limitations with previous studies. First and foremost, our approach is to predict one set of individuals with data about the experience of a completely different set and ratings of concepts gathered from yet others. In contrast to dense-data analyses (B. C. Roy et al., 2015), this approach fundamentally limits the amount of variability we will be able to capture. In addition, the granularity of the predictors that can be extracted from corpus data and applied to every word is necessarily quite coarse. Ideally, predictors could be targeted more specifically at particular theoretical constructs of interest (for example, the patterns of use for specific predicates).

Finally, our work underscores the incompleteness of the current understanding of vocabulary development. Even for English, the language in which our model captures the most variance ($r^2 = 0.29$), much still remains unexplained. Furthermore, this variance is highly reliable—cross-validation using half of the English-speaking children to predict ages of acquisition for the other half yields $r^2 = 0.98$. This gap highlights an important theoretical challenge in the study of early language: linking individual datapoints to the broader patterns of acquisition. We have strong theories of how individual learning situations proceed, but must unify these theories to make progress on understanding language learning at scale.

All data and code for these analyses are available at `https://github.com/mikabr/aoa-prediction`

Acknowledgements

Thank you to Wordbank contributors and the MacArthur-Bates CDI Advisory Board. This work supported by NSF BCS #1528526.

References

Baldwin, D. A. (1995). Understanding the link between joint attention and language. *Joint Attention: Its Origins and Role in Development*, 131–158.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(01), 85–123.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Fenson, L. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual*. Paul H. Brookes Publishing Company.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In *Language acquisition and conceptual development*. Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

Lewis, M. L., & Frank, M. C. (under review). The length of words reflects their conceptual complexity.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PloS One*, *10*(9), e0137147.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. *Psychological Science*, *18*(1), 79–87.

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 670–683.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Zipf, G. K. (1935). The psycho-biology of language.

Lexical Complexity of Child-Directed and Overheard Speech: Implications for Learning

Ruthe Foushee (foushee@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Mahesh Srinivasan (srinivasan@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Abstract

Although previous studies have found a link between the quantity and quality of child-directed speech learners receive and their vocabulary development, no previous studies have found a parallel link between overheard speech measured at a very young age and vocabulary development (Shneidman & Goldin-Meadow, 2012; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013; Weisleder & Fernald, 2013). This is despite the fact that children are able to learn words from overheard speech in laboratory settings (Shneidman & Woodward, 2015). Drawing on the idea that children preferentially attend to stimuli that are at a manageable level of complexity (Kidd, Piantadosi, & Aslin, 2012, 2014), the present research explores the possibility that children do not initially tune into overheard speech because it is initially too complex for their stage of lexical development (i.e., contains too great a proportion of unfamiliar words). Using transcripts from CHILDES and the Santa Barbara Corpus, and estimates of vocabulary by age from the MB-CDI, we find that child-directed speech is significantly less complex than overheard speech through at least 30 months. If attention based on complexity at least partially accounts for the statistical independence of overheard speech and vocabulary development in early childhood, then children might only begin learning from more complex, overheard speech sometime after 30 months.

Keywords: lexical development; attention; corpus analysis

Introduction

In every study designed to investigate what constitutes effective input for language-learning, researchers find that child-directed speech reliably predicts children's vocabularies months later, while overheard speech does not (Shneidman & Goldin-Meadow, 2012; Shneidman et al., 2013; Weisleder & Fernald, 2013). Studies like these record a period of children's everyday linguistic input at 1.5–2.5 years, and relate the quantity (number of words) and quality (number of word types, or lexical diversity) of speech coded as child-directed or overheard to vocabulary assessed at 2.5–3.5 years. Surprisingly, the independence of overheard input and later vocabulary development persists even in contexts where the majority of children's input comes from overheard speech (Shneidman & Goldin-Meadow, 2012), which characterizes many cultural communities (Correa-Chávez & Rogoff, 2009; Ochs, 1982; Schieffelin, 1990; Ward, 1971). Also adding to this puzzle are indications that children *can* learn from overheard speech: preschoolers pick up profanity, and can learn new words in laboratory tasks by as early as 18 months (Akhtar, Jipson, & Callanan, 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe, Liebal, & Tomasello, 2012; Martínez-Sussman, Akhtar, Diesendruck, & Markson, 2011; Shneidman, Buresh, Shimpi, Knight-Schwarz, & Woodward, 2009).

Outside of child-directed speech, recent work suggests that infants may selectively attend to input that is in an optimal, intermediate zone of complexity: neither too simple nor too complex (Kidd et al., 2012, 2014). In looking time studies with 7–8-month-olds, children's probability of looking away from a visual display is lowest when the visual or auditory events are within an intermediate range of predictability based on the preceding sequence of events. Infants are highly likely to look away both when the event is too predictable (e.g., identical to what had happened on the previous several trials), and when it is too unpredictable (i.e., a completely unexpected event). Such studies indicate that children may be implicitly tracking the complexity of different inputs and budgeting their attentional resources, attending only to those stimuli which are in a learnable range. From this work, it seems one of the many ways to capture children's attention is by providing stimuli that are neither too simple nor too complex, according to their knowledge state.

In this paper, we explore the hypothesis that children's implicit attention management might provide an explanation for the limited impact of overheard speech on language acquisition. Might overheard speech initially fail to predict vocabulary development because it is excessively complex, causing children to ignore it until they can learn from it? If so, the lack of a relationship between overheard speech and vocabulary outcomes might be due to how early previous studies took measures of each input type.

To explore this hypothesis, we examine whether we can account for findings about effective input for word-learning by defining an optimal window of complexity for an aspect of language acquisition. In particular, we focus on defining complexity for lexical development, since vocabulary was the outcome measure used in previous studies of overhearing.

While language input varies in complexity along many other dimensions (e.g., morphological, syntactic), for the present purposes, we will be focusing on the familiarity of words in the speech stream. Even so constrained, complexity is a moving target: as the abilities of the learner change, so does what she would consider unpredictable, or engaging. Our measure of lexical complexity therefore takes into account the expanding body of words in the child's vocabulary.

Our analysis uses tools from information theory and estimates of the current vocabularies of children at different ages to compute the complexity of the content of child-directed and overheard speech. Our results show that child-directed speech is consistently less complex than overheard speech through at least 30 months of age. These results provide support for the notion that children might be adaptively selecting their input by allocating attention to speech that falls into an intermediate zone of complexity given their current stage of lexical development.

Goals in Relating Attention & Linguistic Input

Before presenting the methods and results of our analysis in detail, it is worth considering the goals of this investigation. There are at least two ways in which the speech adults direct to children might elicit more attention than speech they direct to other adults. First, across many cultures and contexts, speech to very young children takes a distinctive, exaggerated acoustic form (see Soderstrom, 2007, for a review). Current work is still elucidating the degree to which typical child-directed speech is tailored to children's learning at all levels of linguistic analysis (Eaves, Feldman, Griffiths, & Shafto, in press; Graf Estes & Hurley, 2013; Rafferty & Griffiths, 2012, *inter alia*). However, studies of listening preferences indicate that it at least captures children's attention (Werker, Pegg, & McLeod, 1994), a prerequisite for acquiring new word-to-meaning mappings (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011; Singh, Nestor, Parikh, & Yull, 2009). Second, adults may selectively use more words that are familiar to the child in child-directed speech than they would in adult-directed speech. This greater proportion of known words, and therefore decreased lexical complexity, may help capture the child's attention, and facilitate their learning of the words they don't yet know. The early disparity between the style and complexity of child-directed versus overheard input might then explain why previous studies have found an asymmetric relation between them and young children's vocabulary outcomes.

By examining the complexity of child-directed versus overheard speech as the child's word knowledge increases, we can begin to understand the ways in which the findings of studies relating input and vocabulary might be different (1) if the children were older, or (2) if they were exposed to overheard speech of lesser complexity. The hypothesis that young children's apparent lack of vocabulary learning from overheard speech is due at least in part to its greater complexity suggests that we might see a correlation between overheard input and vocabulary development in both these cases. To this end, we will investigate how input type (Child-Directed or Overheard), the child's own state of vocabulary knowledge, and learnability-driven attention might interact across lexical development. Taking into account the words familiar to the child, we approximate the relative lexical complexity of child-directed and overheard speech when the child is 12–30 months old.

In addition to validating the greater complexity of overheard speech, our results suggest an important role for the exaggerated style of child-directed speech in getting lexical development off the ground, at least in cultures where it is available. Children's earliest input is likely too full of unknown words to attract any attention based on manageable lexical complexity, but the acoustic profile of typical child-directed speech may maintain the attention necessary for their learning.

This investigation is preliminary, but the idea is that for a given child, with a given vocabulary, we might be able to estimate when she will begin to learn reliably from overhearing. That is, identify the point after which measures of *both* overheard and child-directed speech should predict vocabulary development. In reality, this will vary at least based on the vocabulary size of the individual child, and the typical complexity of her caretakers' adult-versus child-directed speech. Here, we first determine whether child-directed and overheard speech are generally differentiated in terms of lexical complexity across early development, using a complexity measure based on the proportion of words familiar to the average child. If overheard speech is more complex, and if adults calibrate the complexity of their child-directed speech as the target child matures, then when the complexity of the two input types are comparable, we would expect children to attend to overheard speech, and subsequently be able to learn from it. Estimating the beginning of this developmental window is our second goal.

Method

Child's Lexicon

The Macarthur Bates Communicative Development Inventory (MB-CDI) is a family of parental report vocabulary instruments for children ages 0–30, many of whose administrations are publicly accessible online (Frank, Braginsky, Yurovsky, & Marchman, under revision). To obtain a lexicon for a given age, we pulled the list of words reportedly comprehended or produced on 50% or more of the administrations archived on wordbank.com for children of that age, using the `wordbankr` package. We did this for children of each month is age from 12–30 months, the period across which we know from previous studies that child-directed, but not overheard, speech predicts later vocabulary development.

Child-Directed Speech

We obtained corpora of speech directed to children of each of our 19 ages from the Child Language Data Exchange System

(CHILDES) Database (MacWhinney, 2000). The corpora were comprised of transcripts representing English-language, naturalistic (not activity-oriented) parent-child interactions at which no other adult or child was present. All files in the database that met these specifications were included for analysis, so the number of children and files for each age varied, from 869–91701 tokens directed to 2–27 target children of different households.

Adult-Directed Speech

While it would be ideal to obtain samples of overheard speech from the same CHILDES transcripts, they contained little verifiably inter-adult speech. Therefore, transcripts meant to reflect the speech that children might overhear came from the Santa Barbara Corpus (Du Bois et al., 2000–2005), a database of transcribed audio recordings of American English conversations from diverse contexts and regions. All files which documented informal, in-person and monolingual conversations between young to middle-aged adults were included. This resulted in a set of 19 transcripts, and a total of 87,496 words.

Complexity Measures

All morphological roots and their frequencies were obtained for each set of transcripts using the *freq* command in CLAN. For each age and input type, 1,000 words were randomly sampled from the total set, using its frequency distribution. Following Kidd, Piantadosi, and Aslin (2012), a complexity measure for each sample of speech was calculated by taking the negative log probability of the words in the child's lexicon in that sample. For an example, say 600 of the 1,000 words in a sample of 18-month-old-directed speech were among those 310 words reported as known by at least half the 18-month-olds on the MB-CDI. The proportion of known words would be $(600/1000)$, or 0.60. The complexity measure for that sample for that age would be $(-\log(0.60))$, or 0.51. This measure is intended to reflect the density of novel words via its complement: the density of known words. Therefore, as the proportion of words children of a given age *already* know in a sample increases, complexity decreases.

Each child- and adult-directed corpus was sampled 100 times for each age, and the complexity of each sample relative to the average child's lexicon at that age was calculated. We expect the complexity of both child-directed and overheard speech to decrease with age as children's vocabularies increase. Beyond that, we can get a sense of whether parents modify their speech to be less complex for children, which may or may not occur across our age span.

To know this, we would want to compare the mean relative complexity of the child- and adult-directed speech samples at each age, with the prediction that child-directed speech should always be significantly lower. We are also interested in the rates at which the two input types decrease, and whether there is an age at which their complexity scores become comparable.

Results & Discussion

Complexity of Child-Directed vs. Overheard Speech

As predicted, mean complexity of adult-directed speech (ADS) was significantly greater than that of child-directed speech (CDS) at all ages except one, 21 months (paired t-tests, all significant at $p < 0.001$ Bonferroni-corrected for multiple comparisons). As can be seen in Figure 1, the complexity of child-directed speech is significantly *greater* than adult-directed speech at 21 months ($t(99) = 41.86$, $p < 0.001$), which given the similarity in complexity of 21-month "overheard" speech, we expect to be an artifact of the specific CHILDES samples used. The child-directed speech at this age came from 23 transcripts spanning 10 children in a variety of situations, which may have contributed to the greater number of unknown words, though this is the case at many of our ages. Another curious trend is the apparent short-term increase in complexity of both input types after 18 months. As the general relationship between overheard and child-directed speech is maintained during this period, we speculate that the cause may be related to our vocabulary measure, though this and the 21-month-old data will be points for future investigation.

These abnormalities also serve to highlight how remarkable it is that child- and adult-directed speech were consistently different in this analysis, given the variability in the transcripts from which the speech samples were pulled, and the general sparsity of the data at each age. The persistent difference between the two input types suggests adults do in fact adjust their vocabulary based on whether they are speaking to a child or fellow adult. Whether they do so appropriately, that is, whether they continuously calibrate their speech as the child ages, we will explore in future analyses.

Of course the words on the MB-CDI do not represent all the words children know, and the complexity range, particularly for child-directed speech, is therefore likely systematically overestimated. The undoubtedly frequent occurrence of the child's own, known name, along with other household-specific lexical items in speech likely directed more often to her than to another adult, were completely ignored.

Modeling Complexity Trajectories

The ages for which the above analysis was possible were constrained by the vocabulary measure we used, so the dataset will necessarily be expanded in future work. For each input type, we fit an exponential function with the complexity score as our dependent variable, and age as our independent variable. The function was selected on a primarily theoretical basis, based on the shape we expected the relationship between complexity and age to take. Choosing this form necessarily means the two models will converge to zero, however, so we speculate on their relationship past the age of 30 months with caution. We were interested in predicting the age at which the two input complexities would cease to be significantly different, with the hypothesis that this might be when we would expect children to regularly learn from overheard

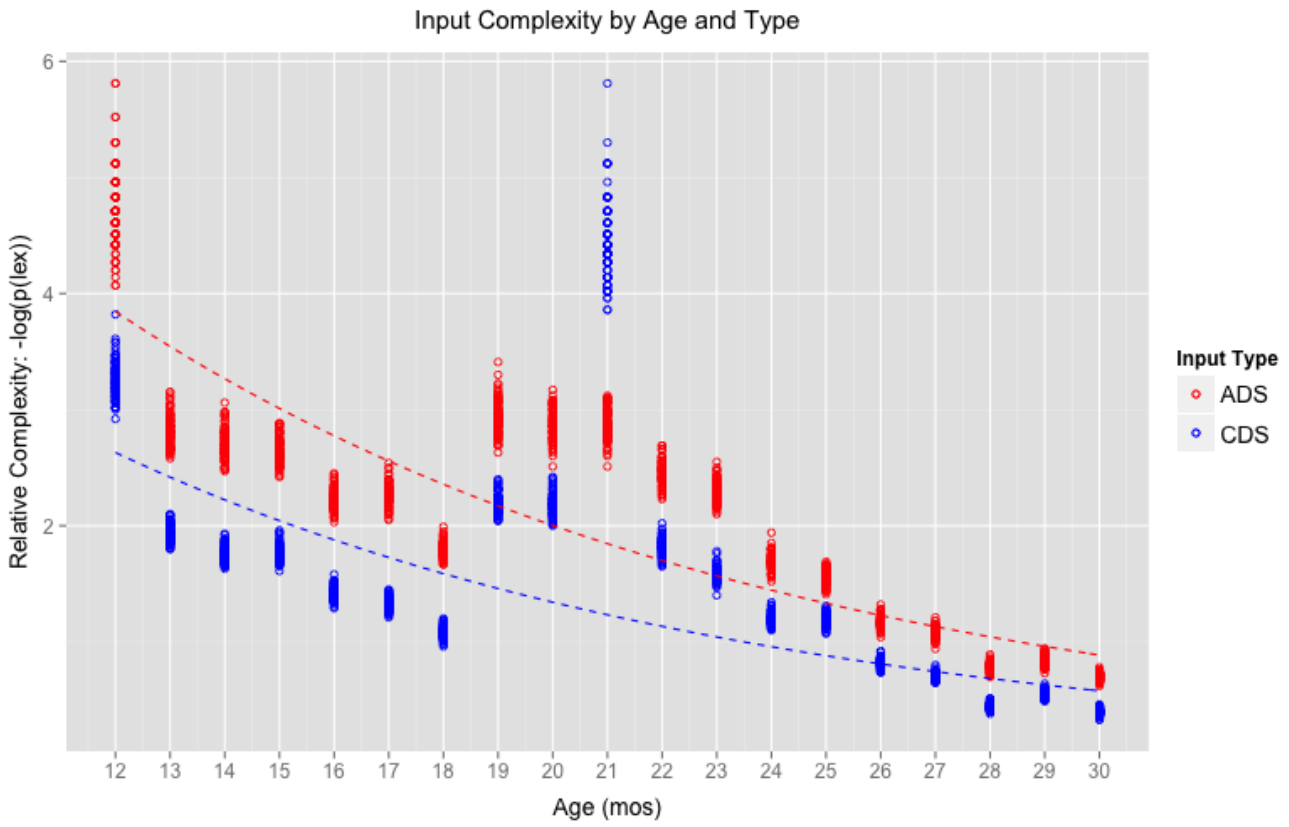


Figure 1: Measure of complexity relative to median child's lexicon at four time points in Child-Directed and Overheard, or Adult-Directed, Speech. Exponential curves for each input type are shown with dotted lines.

Table 1: Coefficients and statistics for exponential regressions ($Y \sim A * \exp(-B * X)$) predicting speech complexity from age for each input type.

Input	A	B	R^2	SE
CDS	1.98	−0.08	0.69	0.32
ADS	2.33	−0.08	0.79	0.24

speech. Looking at Figure 1, we might expect this milestone sometime after the child's third birthday, and certainly not before 30 months.

Our longitudinal analysis of relative complexity sheds light on the puzzle we started with, as well as the trajectory of lexical development. As demonstrated in empirical studies, children are capable of learning from overheard input, but may not recruit this skill typically because the overheard speech they hear is too complex to elicit their attention. While there has been some effort to alter the syntax in which new words are presented in typical overhearing experiments, the scripts for the overheard conversation are designed so that the novel noun the child is supposed to learn is the only unfamiliar word in the dialogue (Akhtar et al., 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe et al., 2012; Martínez-Sussman et al., 2011; Shneidman et al., 2009). Thus, the speech is extremely simplified, and likely at or below the typical complexity of speech directed to children of the participants' ages.

The analysis also provides a new way of looking at child-directed speech. Is attention and an exaggerated acoustic style necessary for the link between child-directed speech and vocabulary development, or is the high proportion of familiar words largely responsible?

During the ages we studied, adult-directed speech was always more complex than child-directed speech, indicating that adults do selectively use words that are accessible to their young interlocutors. However, even the reduced vocabulary apparently used for very young audiences appears highly complex given the lexicon of the typical 12-month-old. This may be where the acoustic profile of child-directed speech critically compensates for unavoidably high lexical complexity in maintaining the child's attention.

By the second year of life, the relative complexity of both types of input fall rapidly as the child's vocabulary increases, and appear to approach each other. That the two are still not equivalent by the age when studies have failed to link overheard input and vocabulary development suggests complexity differences may indeed help account for that finding.

Future Directions

Our complexity measure both gives us leverage on the learning asymmetry between overheard and child-directed speech, and introduces further questions to address. For example, because complexity ratings represent the interaction between adults' choice of vocabulary and children's own lexicons, we can't know how much their rates of decline across time are due to adults' appropriate calibration to the child's knowledge state versus the child's own expanding vocabulary.

In addition, we don't know how much inattention to overheard speech might be influenced by *preference* for child-directed-speech. Do the levels of lexical complexity we found for child-directed speech reflect something universal about children's information processing capacities, or would the 'optimal' level of complexity to attend to and therefore learn from be higher if child-directed speech were unavailable? This question points to the need for a similar analysis in cultures where child-directed speech is not a primary source of children's linguistic input.

Alternative Measures & Sources

There are several methodological alternatives we might consider to complement the current study. To address the limitations of our vocabulary measure, we might conduct the same analysis using another vocabulary measure with more longevity, like the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007). Older children's lexicons might also be approximated using age of acquisition ratings (Kuperman, Stadthagen-Gonzales, & Brysbaert, 2012), scaled with reference to established norms for a given age range. Extending the analysis to later ages is vital to determining if and when child- and adult-directed speech converge in lexical complexity.

Some of the uncertainties remaining in the current analysis might be resolved by limiting our complexity calculation to longitudinal datasets, which would also enable us to more accurately approximate the child's vocabulary development via her production. Others can only be addressed with longitudinal data that capture both the speech directed to a child, and the speech that same child has the opportunity to overhear.

Alternative complexity measures capturing the same idea are also valuable to explore. In our measure, for example, samples with the same proportion of known words, but different levels of diversity of unknown words, are equivalent. This does not reflect the intuition that a sample of speech which contains 600 words you know, but 400 *different* words you don't should be more complex than a sample with 400 of the same, unknown word. Thus, a measure which incorporates the entropy of the unknown words in a sample may even better reflect lexical complexity. Studies interested in the same question may choose to model complexity instead using word-by-word predictability in the input, sampling continuous segments of speech rather than random tokens, and judging complexity based on transitional probability.

Information about optimal levels of complexity might also be informed by children's book standards. Children's books typically have a recommended audience age. It might be informative to calculate our complexity measure for a corpus of children's books either recommended for a given age, or reported by teachers and parents to enthrall children of a given age, and compare it to the age-typical complexity of informal speech input. Previous studies have ascertained that children's books have greater lexical diversity than child-directed speech (Jones & Smith, 2015), but our measure would enable us to assess how calibrated storybooks are for vocabulary learning at different ages.

Experimental Predictions

The current analysis gives us a complexity range for lexical development that can be tested in the lab. If a period of overheard speech were crafted to receive a complexity rating of approximately 0.33 (the mean complexity of child-directed speech at 30 months), could we get a 30-month-old attending to, and consequently learning from, it? What about a younger child with a large vocabulary? The thus-far neglected aspect of the findings linking attention and learnability—that children should also not attend to overly simple input—can now be investigated as well. Will children preferentially attend to slightly more complex, but learnable, overheard speech, compared to completely predictable child-directed speech? Studies with infants terminate the experiment when they look away, so we don't know how children manage their attention when a stimulus continues. We also can't know without empirical support whether our sample size of 1,000 words is a reasonable window in which to assess complexity, and how that window changes as children's attentional resources increase. If children stop tuning in when speech becomes excessively complex, can their attention be regained later on?

Trade-offs between complexity and speech style should be systematically investigated at different ages and vocabularies, along with ways of encouraging learning from indirect speech beyond manipulating lexical complexity. If learning from overhearing is driven at least in part by attention, then we should see learning of a new word by a pig-loving preschooler from a complex overheard dialogue about pigs before learning from a dialogue about architecture of the same complexity. Syntactic complexity could be manipulated as well: unknown words occurring in high proportion in familiar syntactic frames might be easier to learn than words in unfamiliar frames at the same density.

As we move forward in this investigation, it is critical to acknowledge that there are other ways in which child-directed and overheard speech diverge. While children have been shown to be able to infer the meaning of a new word via its discourse context (Sullivan & Barner, 2015), learning new words from overhearing might remain more difficult than from child-directed speech—even at the same complexity level—merely because referents in parent-child interaction are qualitatively different and easier to identify. We can't know how much explanatory power the idea of lexical development based on input complexity might give us, however,

until we expand our analysis to further vocabulary measures, transcripts, and methods.

References

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209.

Akhtar, N., Jipson, J., & Callanan, M. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416–430.

Correa-Chávez, M., & Rogoff, B. (2009, May). Children's attention to interactions directed to others: Guatemalan mayan and European American patterns. *Developmental Psychology*, *45*(3), 630–641.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000–2005). *Santa Barbara corpus of spoken American English, Parts 1–4.* Philadelphia: Linguistic Data Consortium.

Dunn, L. M., & Dunn, D. M. (2007). *PPVT–4: Peabody picture vocabulary test*. Minneapolis, MN: Pearson Assessments.

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (in press). Infant-directed speech is consistent with teaching. *Psychological Review*.

Floor, P., & Akhtar, N. (2006). Can eighteen-month-olds learn words by listening in on conversations? *Infancy*, *9*, 327–339.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (under revision). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language*, *32*(3), 385–397.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5).

Jones, J. L. M. M. N., & Smith, L. B. (2015). The words children hear: picture books and the statistics for language learning. *Psychological Science*, 1–8.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5).

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child Development*, *85*(5), 1795–1804.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

MacWhinney, B. (2000). *The Database* (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Martínez-Sussman, C., Akhtar, N., Diesendruck, G., & Markson, L. (2011). Orienting to third-party conversations. *Journal of Child Language*, *38*(2), 273–296.

Ochs, E. (1982). Talking to children in Western Samoa. *Language in Society*, *11*, 77–104.

Rafferty, A. N., & Griffiths, T. L. (2012). Optimal language learning: The importance of starting representative. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Schieffelin, B. (1990). *The give and take of everyday life: Language socialization of Kaluli children*. Cambridge: Cambridge University Press.

Shneidman, L. A., Arroyo, M. E., Levine, S., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672–686.

Shneidman, L. A., Buresh, J. S., Shimpi, P., Knight-Schwarz, J., & Woodward, A. L. (2009). Social experience, social attention and word learning in an overhearing paradigm. *Language Learning and Development*, *5*, 266–281.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673.

Shneidman, L. A., & Woodward, A. L. (2015). Are child-directed interactions the cradle of social learning? *Psychological Bulletin*.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

Sullivan, J., & Barner, D. (2015). Discourse bootstrapping: Preschoolers use linguistic discourse to learn new words. *Developmental Science*, *19*(1).

Ward, M. (1971). *Them children: A study in language*. New York: Holt, Rinehart, and Winston.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–52.

Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, *17*(3), 323–333.

Degeneracy results in canalisation of language structure: A computational model of word learning

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster LA1 4YF, United Kingdom

Abstract

There is substantial variation in language experience between learners, yet there is surprising similarity in the language structure they eventually acquire. While it is possible that this canalisation of language structure may be due to constraints imposed by modulators, such as an innate language system, it may instead derive from the broader, communicative environment in which language is acquired. In this paper, the latter perspective is tested for its adequacy in explaining the robustness of language learning to environmental variation. A computational model of word learning from cross-situational, multimodal information was constructed and tested. Key to the model's robustness was the presence of multiple, individually unreliable information sources that could support learning when combined. This "degeneracy" in the language system had a detrimental effect on learning when compared to a noise-free environment, but was critically important for acquiring a canalised system that is resistant to environmental noise in communication.

Keywords: canalisation; degeneracy; language acquisition; multiple cues; word learning

Introduction

A key question in the cognitive sciences is how, despite the enormous variation in linguistic experience, each language learner acquires broadly the same language structure, "within a fairly narrow range" (Chomsky 2005). This issue has led to proposals for mechanisms that ensure this "canalisation" of language structure. Traditionally, these mechanisms have been conceived as constraints that apply to structure the language exposure, such as innately specified syntactic or semantic properties. But there is growing realisation that multiple, rich sources of information within the communicative environment may offer substantial, perhaps sufficient constraints to learning.

A similar change in perspective was observed in canalisation in biological evolution. Initial proposals were that canalisation was a consequence of the natural selection of mechanisms that operate to minimise phenotypic variation (Waddington, 1942). However, a more recent explanation is that minimal phenotypic variation is stably achieved as a consequence of interaction between multiple regulators (despite substantial environmental variation) as part of the developmental process of the organism (Siegal & Bergman, 2002). Simulations of the developmental operation of multiple transcriptional regulators found that the greater the interactivity between these sources, the smaller the phenotypic variation resulting from environmental variation.

An analogous perspective can be taken on canalisation of social or cultural systems, such as language, whereby increasing levels of interaction may increase the stability and optimise performance of an information processing system (Bettencourt, 2009). Canalisation of language, long conceived as being a consequence of mechanisms that implement resistance to environmental variation, could instead be the outcome of interacting, multiple sources of information.

Recently, there has been reconsideration of the potential for language learning to be supported by the richness of the language environment. For instance, grammatical category acquisition is not only supported by information from word co-occurrences – the traditional information source for linguistics studies of language acquisition (Redington, Chater, & Finch, 1998) – but also from substantial information in phonotactic and prosodic structure, such as distinct stress patterns on nouns compared to verbs (Monaghan, Christiansen, & Chater, 2007). Furthermore, information about objects and actions within the child's purview may further constrain potential referents for words (Yurovsky, Smith, & Yu, 2013), providing restrictive information about the semantic features associated with particular categories.

There have been several accounts for how such multiple cues may be combined to support learning. The redundancy of different information sources may assist the learner by increasing the saliency of particularly important information present in their environment (Bahrick, Lickliter, & Flom, 2004). Alternatively, the cues may operate summatively (Christiansen, Allen, & Seidenberg, 1998), or they may operate in a hierarchy, such that if one cue is available then it is used in preference to other cues, which are relied upon only if the preferred cues are unavailable (Mattys, White, & Melhorn, 2005).

An alternative possibility, consistent with models of canalisation in biology, is that multiple cues for language learning interact, resulting in a system that is stable in the face of variation in the environment. This property of language is its "degeneracy", defined as "the ability of elements that are structurally different to perform the same function or yield the same output" (Edelman & Garry, 2001). Degeneracy affects not only acquisition – where presence or absence of particular cues will not adversely affect the structure acquired – but also the robustness of the system once the language is acquired, due to reduced dependency on any one information source. Computational models of degeneracy in language and other complex systems have shown that it is important for robustness of

learning (Whitacre, 2010), permitting, for instance, effective processing of speech sounds against background noise (Winter, 2014).

In this paper, a computational model of multiple interacting information sources is presented as a proof of concept that degeneracy can result in canalisation of language structure. The domain of study is word learning, where forms and meanings of words have to be mapped. This task is difficult, because there are numerous possibilities for the target candidate word in multi-word utterances, and multiple possible referents in the environment to which the target word may map (Quine, 1960). However, multiple cues are present both in the spoken language and in the environment that surrounds the learner to assist in this task. This perspective requires extending the notion of degeneracy from examining the redundant, overlapping cues within language structure to examine cues more broadly within the communicative situation.

Within spoken language, information about the grammatical roles for words can be ascertained from distributional information, consequently reducing the number of possible target words that need to be considered. For instance, nouns are frequently preceded by articles (the, a) and these also tend to succeed verbs. Use of such simple distributional information has been shown to assist in determining word referent mappings (Monaghan & Mattock, 2012). Further information for identifying the critical information in an utterance is also available from prosodic information. When teaching a child a new word, the speaker tends to increase the pitch variation, intensity, and duration of the target word within the utterance (Fernald, 1991).

In addition, constraints within the environment help to reduce uncertainty about the potential referents. One of these information cues is derived from cross-situational statistical information, where over multiple situations the learner can increase their association between the target word and target object (McMurray, Horst, & Samuelson, 2012) even when several possible words and referents are present. Such cross-situational learning (Yu & Smith, 2012) can further be supplemented by information that the speaker uses to indicate the field of reference. For instance, speakers tend to use deictic gestures (finger pointing or eye gaze) toward a referent which is being described (Iverson, Capirci, Longobardi, & Caselli, 1999).

However, in isolation, each of these cues is insufficient to perfectly constrain learning: The word succeeding an article is not always a noun – in English adjectives might intervene, and spontaneous language is replete with false starts, and word sequencing errors. Similarly, the loudest word in speech is not always the target word, or a novel word being learned by the listener, and gestural cues are not always reliable. In Iverson et al.'s (1999) study they found that 15% of utterances were accompanied by gestures indicating aspects of the immediate environment to direct children's attention. Yet, such unreliability has profound value for learning. Consider if the child always learned from a speaker who reliably pointed to the intended referent. Then, if ever a situation arose where a referent was not gestured towards, this could impair effective communication, because the cue may be relied upon for effective mapping from word to referent.

There are costs to including multiple cues in the learning situation, because this increases the amount of information needed to be processed in each instance of learning. So, the trade-off between the increased strain on the cognitive systems required by processing of multiple as opposed to single, or no, cues and the potential advantages of interacting information sources for learning must be examined. Specifically, we tested the value of multiple information sources for learning, and we examined the importance of interaction among information sources for linguistic canalisation, i.e., the robustness of learning in the face of environmental variation.

A computational model was constructed to test integration of information received from multiple sources to assist the learning of relations between words and their referents. Two sets of simulations testing the model were conducted. The first assessed the contribution of single cues to word learning. The hypothesis was that adding cues to the input would assist in acquisition of the mapping, with gestural cues assisting in defining the referent, prosodic cues promoting identification of the target word, and distributional cues supporting acquisition of both. However, the reliable presence of cues was hypothesised to result in impaired ability to identify the form-meaning mapping when the cue is no longer present.

The second set of simulations explored the role of multiple cues for learning. The prediction was that multiple cues would further promote learning, but that noisy cues would be most effective for supporting not only effective acquisition but also robustness in the learning, immune from effects of variability in the environment. Thus, a model trained with a degenerate environment should result in a canalised system, being able to effectively map between words and referents even when environmental cues that support this mapping are no longer available.

A Multimodal Model of Word Learning

The starting point for the current model used the hub-and-spoke architecture (Plaut, 2002), where information from different modalities is inputted to a central processing resource, and is thus unconstrained in its integration. These models then determine the optimal way in which information sources can cohere to support learning. The model implementation is closely based on a previous model of multimodal information integration in sentence processing, which was created to simulate behaviour in the visual world paradigm (Smith, Monaghan, & Huettig, 2014). This modeling approach has been effective in demonstrating how and when different information modalities interact in language processing, and how the influence of different modalities on language processing

derive from the nature of the representations themselves, rather than requiring architectural assumptions to be imposed on the system.

The model used here is a subsystem of this larger modeling enterprise, addressing the special case of acquiring word-referent mappings. The model is compatible with previous associative models of word learning (McMurray et al., 2012), as well as being broadly consistent with the principles of statistical models of cross-situational word learning (Yu & Smith, 2012). The model therefore applies these general modeling principles to explore the role of multiple information sources in facilitating, and constraining word learning.

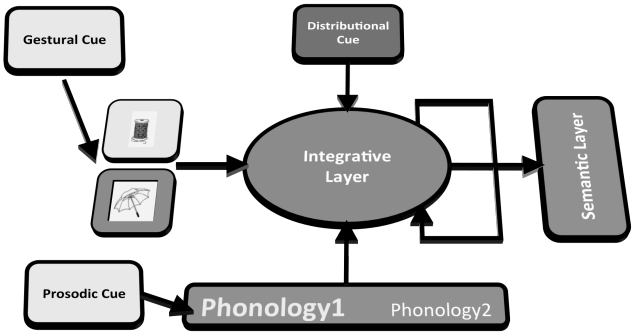


Figure 1: The multimodal integration model of word learning.

Architecture

The model architecture is illustrated in Figure 1. The model is implemented as a recurrent backpropagation neural network. It comprises a central hidden layer of 100 units which received connections from various input modalities, and projected to a semantic layer output.

The phonological input represented two word slots, each of which contained 20 units. The visual input contained two locations each comprising 20 units, where object representations were presented. The semantic layer was composed of 100 units. For some simulations that included a distributional cue, the model also received input from a distributional cue layer, which was composed of 2 units. The integrative layer was also fully self-connected.

Representations

The model was trained to learn 100 words. Representations of each modality of a word was encoded as a pseudopattern so that the properties of the relations between representations could be controlled. The phonological representation of each word was composed of four phonemes, randomly drawn from a set of 10 different phonemes. Each phoneme comprised 5 units, with 2 units active. The visual representation of the word's referent was constructed from 20 units with 8 units active for each representation. The semantic representations were localist, such that one of the 100 units was active for each of the words.

Fifty of the words were randomly assigned to one category, and the remaining fifty were assigned to the other category, such that these categories could be defined by a distributional cue.

Table 1: Proportion of training trials with each cue according to condition.

Condition	Dist Cue	Prosodic Cue	Gestural Cue
No Cue	0	0	0
Single Cues			
Dist Cue	1	0	0
Prosodic Cue	0	1	0
Gestural Cue	0	0	1
Combined Cues			
25% reliability	.25	.25	.25
50% reliability	.50	.50	.50
75% reliability	.75	.75	.75
100% reliability	1	1	1

Training

The model was trained to identify the meaning of the word from phonological and visual representation inputs, for all 100 words. Each trial was a simulation of a cross-situational learning task, where two words and two objects were presented, but only one of the objects was named by one of the words (Monaghan & Mattock, 2012). The model had to learn to solve the task by generating the correct semantic representation for the named object.

For each training trial, a word was randomly selected. Its phonological form was presented at one of the two word slots in the phonological input (position was randomly chosen), and another randomly selected word's phonological form was presented at the other word slot. The object depicting the word's referent was presented at one of the two visual input positions (randomly chosen) and another randomly selected visual representation was presented at the other visual input position.

For the simulations with cues, gesture and prosody were implemented as intrinsic properties of the visual and phonological input representations, respectively, by doubling the activation at the input of the target visual object or the target phonological form. This had the effect of increasing the contribution of the target representation within each representational modality to affect the activation state of the integrative layer, and was a simulation of increased saliency of that representation (i.e., that a gestural cue increases saliency of the target object, and prosodic cue is implemented as an increase in intensity, duration, and pitch of the target spoken word). This is illustrated in Figure 1 as a highlighting of the uppermost object and the first phonological representation as a consequence of gestural and prosodic cues, respectively.

The distributional cue was implemented as an extrinsic cue. If the word was from the first (randomly assigned) category then the first unit in the distributional layer was

active, and if the word was from the second category the second unit was active. This cue could therefore assist the model in determining which was the target object and spoken word, but the cue did not operate within either of these modalities.

The simulations of single cues presented each learning trial with the cue present with 100% reliability (see Table 1). The simulations of multiple cues varied the extent to which the cues were reliably present in each learning situation, from 25%, through to 100% reliability.

Activation cycled in the model for 6 time steps. At time step one, the visual and phonological inputs were presented. For two time steps activation passed from the input to the integrative layer and from the integrative layer to the semantic layer, and from the integrative layer to itself. At time steps 3 to 6 the target semantic representation was presented at the semantic output layer, and activation continued to cycle around the model. The model was trained with continuous recurrent backpropagation through time (Pearlmutter, 1989) with error determined by sum squared error of the difference between the actual and target semantic representations. In one epoch of training, each of the 100 words occurred once as the target. The model was trained up to 100,000 epochs at which point performance for each condition had asymptoted.

Twenty versions of the model with different pseudopattern representations, different randomised starting weights, and different randomised ordering of training patterns were run.

Testing

The model's performance was assessed during training on its ability to produce the target semantic representation for each phonological and visual input. If the activity of the semantic unit corresponding to the target word was more active than any other unit in the semantic layer, then the model was determined to be accurate.

Accuracy during training was assessed, and also the point in training at which the model was able to accurately detect all 100 words for five consecutive epochs.

At the end of training, the robustness of the model's learning was assessed by measuring its accuracy when no cues were present during testing.

Results

Single Cues

The model's accuracy during training when no cues and single cues were present is shown in Figure 2.

An ANOVA with epoch at which the simulation reached the accuracy criterion as the dependent variable, and cue condition (no cue, distributional cue, prosodic cue, gestural cue) as within subjects factor was conducted to test whether the model learned differently according to the presence of cues. The result was significant, $F(3, 57) = 70722$, $p < .001$, $\eta_p^2 = 1.00$. Post hoc tests revealed that the model reached criterion more quickly for the prosodic cue (mean epochs = 35,800, SD = 1,005), and gestural cue (mean = 35,650, SD = 745) conditions than the no cue condition which had not reached criterion by 100,000 epochs (mean proportion correct was .96), both $p < .001$. Though the trajectory of learning was distinct, as shown in Figure 2, the effect of distributional cues was smaller, and not significantly different in time to criterion compared to the no cue condition (mean proportion correct after 100,000 epochs was .99). The prosodic and gestural cues supported learning more than the distributional cue, both $p < .001$, but there was no statistical difference in speed of learning from the prosodic and gestural cues, $p = 1$.

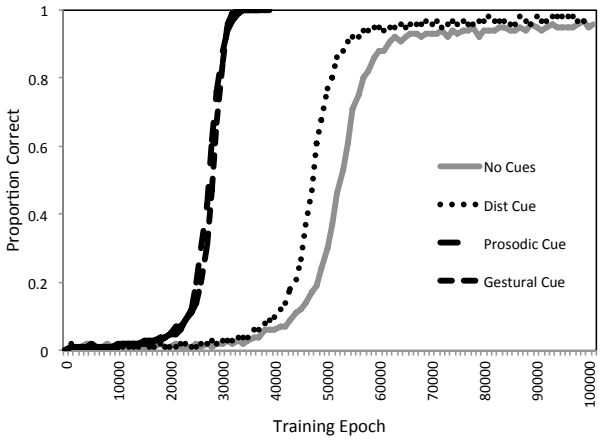


Figure 2: Accuracy during training for the single cues conditions, compared to no cue condition.

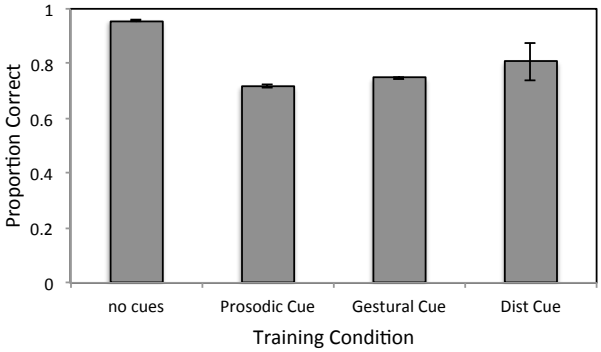


Figure 3: Accuracy after training for the single cues conditions, when no cues are present during testing (Dist = Distributional).

The robustness of the model's learning to omission of cues during testing is shown in Figure 3. An ANOVA on accuracy in the post-learning test with no cues present, and cue condition as within subjects factor was significant, $F(3, 57) = 8.982$, $p < .001$, $\eta_p^2 = .321$. Post hoc tests showed that the distributional cue did not significantly affect robustness of learning compared to the no cue condition, $p = .284$, however, the prosodic and gestural cue both resulted in poorer performance than the no cue condition, both $p < .001$. The gestural cue resulted in more robust learning than

the prosodic cue, p = .001, but these conditions did not differ significantly from the distributional cue condition, both p = 1.

We tested whether the difference between the intrinsic cue conditions (prosodic and gestural cues) was due to their quicker acquisition. We trained every model to the same number of training trials (100,000) then tested robustness of learning. The results were similar. Even with more training, the effect of a single, reliable intrinsic cue was detrimental to the model's ability to map between form and meaning when the cue was not present, $F(3, 48) = 45.62$, $p < .001$, $\eta_p^2 = .740$. Prosodic and gestural cues were now not significantly different than one another, $p = .423$, but were both significantly different than the no cue and the distributional cue conditions, all $p < .001$.

Multiple Cues

The model's accuracy during training for combined cues with different levels of reliability is shown in Figure 4. For epoch taken to reach training criterion, an ANOVA indicated that combined cues with different reliability significantly affected speed of learning, $F(4, 76) = 3855$, $p < .001$, $\eta_p^2 = .99$. Post hoc tests indicated that learning in the no cue and the 25% cue reliability condition were significantly slower than the 50% condition, both $p < .001$, which was in turn slower than the 75% condition, $p < .001$, which was in slower than the 100% perfect reliability multiple cue condition, $p < .001$. Thus, as anticipated, the greater the reliability of information present during learning, the faster the model learned to map between forms and meanings.

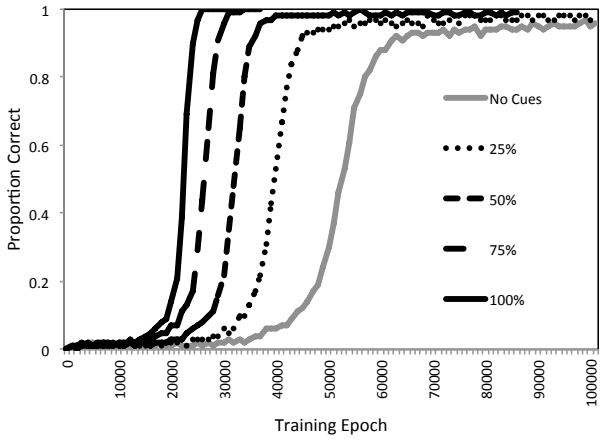


Figure 4: Accuracy during training for the multiple cue conditions, compared to no cue condition.

The robustness of learning was also compared between these conditions. The results are shown in Figure 5. An ANOVA demonstrated that the robustness of performance at testing was affected by the cues present during training, $F(4, 76) = 2.953$, $p = .025$, $\eta_p^2 = .135$. Post hoc tests revealed that the no cue and 50%, 75%, and 100% cue conditions were significantly different, all $p < .001$. The 25% cue condition was not significantly different than any other condition, all $p \geq .718$. As reliability increased from 50% to 75%, the robustness of the model declined, $p < .001$, and similarly declined from 75% to 100% reliability, $p < .001$. Thus, low reliability of cues did not seem to assist in learning quickly or robustly, but once individual cues appeared at least half the time, further increasing the reliability of the cues began to reduce the resistance of the model to the absence of cues after training. 50% reliability appears to be close to the optimal trade-off for speed and robustness of learning.

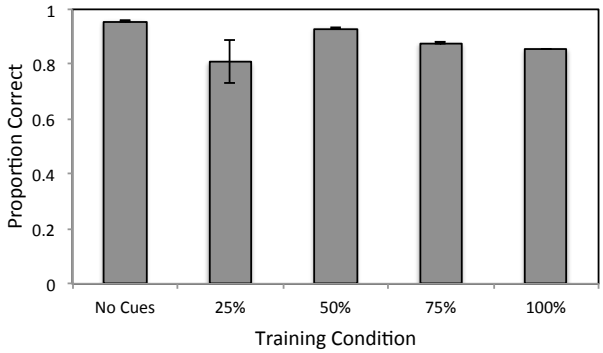


Figure 5: Accuracy after training for the multiple cue conditions, when no cues are present during testing.

Discussion

Language learning occurs in situations where multiple, interacting sources of information are available to support acquisition. Although attending to multiple cues increases the processing load on the individual, this degeneracy in language results in two important advantages for the language learning system.

First, adding a combination of cues to the model's input improves the speed and accuracy of learning to map between representations. Providing some guiding information about the intended object in a scene containing more than one referent, and emphasis of the target word in a multiword utterance, along with additional information about the general category of the target, improves performance. Even when the individual cues occurred only 50% of the time, learning of form-meaning mappings was still significantly enhanced compared to learning in the absence of cues.

This observation that speed and accuracy of language learning is promoted by multiple cues has been explored extensively, and is consistent with several current accounts of multiple cue integration in learning (Bahrick et al., 2004; Christiansen et al., 1998; Mattys et al., 2005; Monaghan et al., 2007). All these theories would predict the growing advantage of learning as cues increase in reliability, as observed in the current simulations.

However, the degeneracy of language also has a second advantage. The learning that is acquired from a degenerate

environment is highly robust (Ay, Flack, & Krakauer, 2007), and the model was able to make use of cues even when they were variably present across communicative situations. However, this multiple cue advantage for robustness was only observed when there was noise in the environment: When the cues occurred with perfect reliability then, even though learning was at its fastest, the acquired system was fragile and prone to error under suboptimal subsequent conditions. Thus, canalisation of language structure in a word learning task can be conceived of as a consequence of the interaction of multiple information sources for learning.

There is therefore a trade-off between speed of initial learning, and the robustness of that learning. The former is supported by perfectly reliable information (see, e.g., Onnis, Edelman, & Waterfall, 2013), and more information resulted in better and better learning. The latter is supported by multiple information sources, but with each individual source being somewhat noisy. The precise point of this trade-off is an issue for further exploration in computational systems, in order to determine the extent to which natural language environments are optimally designed for acquisition.

The simulations presented here suggest that, rather than canalisation being a challenge in the face of environmental variation, it is instead a primary consequence of this variation in a system that is able to integrate multiple information sources.

Acknowledgments

This work was supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. Thanks to Rebecca Frost for comments on this work.

References

Ay, N., Flack, J., & Krakauer, D. (2007). Robustness and complexity co-constructed in multimodal signalling networks *Philosophical Transactions of the Royal Society B: Biological Sciences, 362* (1479), 441-447.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13,* 99-102.

Bettencourt, L. M. A. (2009). The rules of information aggregation and emergence of collective intelligent behavior. *Topics in Cognitive Science, 1*, 598–620.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221-268.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry, 36*, 1-22.

Edelman, G., & Gally, J. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences, 98* (24), 13763-13768.

Fernald, A. (1991). Prosody in speech to children: Prelinguistic and linguistic functions. *Annals of Child Development, 8*, 43-80.

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development, 14,* 57–75.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134,* 477–500.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831-877.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The Phonological Distributional coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259-305.

Monaghan, P. & Mattock, K. (2012). Integrating constraints for learning word- referent mappings. *Cognition, 123,* 133-143.

Onnis, L., Edelman, S., & Waterfall, H. (2011). Local statistical learning under cross-situational uncertainty. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.*

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263-269.

Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19, pp 603-639.

Quine, W.V.O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Siegal, M. L., & Bergman, A. (2002). Waddington's canalisation revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences*, *99*(16), 10528-10532.

Smith, A.C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology, 75,* 28-54.

Waddington, C. H. (1942). Canalisation of development and the inheritance of acquired characters. *Nature*, *150*(3811), 563-565.

Whitacre, J. (2010). Degeneracy: a link between evolvability, robustness and complexity in biological systems *Theoretical Biology and Medical Modelling, 7,* 6.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960-967.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*(1), 21-39.

Yurovsky, D., Smith, L. B. & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*, 959-966.

General Mechanisms Underlying Language and Spatial Cognitive Development

Hilary E. Miller (hemiller@wisc.edu) and Vanessa R. Simmering (simmering@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

Previous research showed that children's spatial language production predicts their spatial skills, but the mechanisms underlying this relation remain a source of debate. This study examined whether 4-year-olds' spatial skills were predicted by their attention to task-relevant information—in tasks that emphasize either memory or language—above and beyond their spatial word production. Children completed three types of tasks: (1) a memory task assessing attention to task-relevant color, size, and location cues; (2) a production task assessing adaptive use of language to describe scenes, varying in color, size, and location; and (3) spatial tasks. After controlling for age, gender, and vocabulary, children's spatial skills were significantly predicted by their memory for task-relevant cues, above and beyond their task-related language production and adaptive use of language. These findings suggest that attending to relevant information is a process supporting spatial skill acquisition and underlies the relation between language and spatial cognition.

Keywords: spatial cognition; short-term memory; language production; cognitive development

Introduction

Spatial skills are fundamental cognitive abilities that support basic behaviors such as perceiving and remembering locations, navigating, and relational reasoning. Additionally, spatial skills predict performance in math and science and entry into STEM fields (Verdine et al., 2014; Wai, Lubinski, & Benbow, 2009). As individual differences in spatial skills arise early in development (e.g., Verdine et al., 2014), it is important to understand the factors that contribute to the development of spatial skills during early childhood in order to devise effective interventions to promote spatial cognition.

Spatial language has been identified as one factor that predicts children's spatial abilities. Multiple studies have identified relations between language and spatial cognition across a variety of spatial tasks (e.g., Dessalegn & Landau, 2008; Pruden, Levine, & Huttenlocher, 2011). Despite the proliferation of research, the causal nature of the link between language and spatial cognition remains unclear. It is possible that language directly causes changes in spatial abilities, spatial abilities cause changes in language, language indirectly relates to spatial cognition through general mechanisms such as attention, or that multiple causal factors interact bi-directionally to shape spatial skills and language. The current study investigated whether a more general mechanism of attention to task-relevant cues might underlie the relation between language and spatial cognition.

Previous research has shown correlations between children's spatial word production and their spatial performance. For example, children who produced particular spatial terms such as "left" and "right" (Hermer-Vazquez, Moffet, & Munkholm, 2001) and "middle" (Simms & Gentner, 2008) tended to perform better on spatial tasks that involve these dimensions. Also, the total number of spatial words that children spontaneously produced during free play predicted their spatial performance (Pruden et al., 2011). These effects are thought to be specific to spatial words, as the relation between spatial word production and children's performance on spatial tasks held even after controlling for spatial word comprehension, IQ, (Hermer-Vazquez et al., 2001) and general receptive vocabulary (Pruden et al., 2011).

Further evidence that language facilitates spatial development comes from paradigms in which children were provided with verbal cues specifying spatial relations among objects or features (e.g., "top", "left") before or during spatial tasks. The results across these studies showed that spatial language cues enhanced children's spatial task performance (e.g., Dessalegn & Landau, 2008; Miller, Patterson, & Simmering, 2016). These findings have been interpreted as evidence that language helps children verbally encode task-relevant spatial information and once children have acquired relevant spatial words they can use language to facilitate their performance (e.g., Hermer-Vazquez et al., 2001; Loewenstein & Gentner, 2005; Pruden et al., 2011).

Other studies showed that such effects were not specific to spatial language, but arose more generally through highlighting task-relevant cues (e.g., Shusterman, Lee, & Spelke, 2011). Specifically, Shusterman et al. (2011) showed that providing children with non-spatial language (e.g., "the red wall can help you find the sticker") that highlighted the utility of particular cues improved their spatial performance. This effect was specific to task-relevant language, as similar but task-irrelevant language did not improve performance (e.g., "look at the pretty red wall").

The range of results across conditions with both spatial and non-spatial words suggests that language can draw children's attention to relevant cues in a spatial task. However, these findings raise questions about whether the relation between language and spatial cognition arises from children's spatial word acquisition. Is it the case that once children can produce relevant language (e.g., spatial words) they use their language knowledge to verbally encode task-relevant information or is it the case that children need to be selective in the types of cues they encode? It is possible that children can produce task-relevant language on their own, but may not use it in the context of spatial tasks. Being provided with language could help children encode relevant information, but their abilities to produce words on their own may not be enough to direct children to encode relevant cues.

Miller, Vlach, and Simmering (in press) tested these two

alternative accounts and showed that the relation between language and spatial cognition does not arise solely from children's abilities to produce spatial words. Rather, it reflects children's abilities to adaptively use language. In this study, 4-year-olds completed a spatial scene description task that assessed their use of task-relevant language. In the task, children described the location of a mouse in a spatial scene (Figure 1). There were three possible types of cues (color, size, or location) that children could use to describe the location of the mouse; across trials the relevance of the color and size cues varied such that on some trials, color and/or size was not a relevant cue for describing the mouse's location (cf. Figure 1A vs 1B or 1C). Results showed that children who used language more adaptively (i.e., provided more relevant than irrelevant cues) performed better on the spatial tasks, even when controlling for their age, gender, vocabulary (both general receptive and spatial productive), and quantity of task-related language produced (including spatial words).

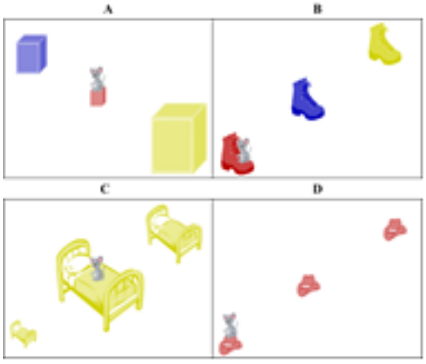


Figure 1: Sample trials of the spatial scene description task with varying numbers of distinctive cues: A) 3 cues = color, size, and/or location, B) 2 cues = color and/or location, C) 2 cues = size and/or location, D) 1 cue = location only.

Miller et al.'s (in press) results suggest that using language in task-relevant ways relates to spatial performance, but it is unclear whether language is the causal factor underlying this link or whether a third factor—not specific to language or spatial skills—contributes to both. Miller et al. hypothesized that children's attention to task-relevant cues supports both language and spatial skills. Some support for this hypothesis comes from research showing that children's spatial task performance can be facilitated by non-spatial language (e.g., Shusterman et al., 2011) and/or non-verbal cues that direct attention to task-relevant features (i.e., by making them more stable or salient; Learmonth, Nadel, & Newcombe, 2002; Learmonth, Newcombe, & Huttenlocher, 2001).

The current study provides a further test of the hypothesis that the link between spatial skills and language arises from a third common factor: the ability to direct attention to task-relevant cues. We tested this by comparing children's spatial task performance to not only their spatial scene descriptions (as in Miller et al., in press), but also a memory task assessing attention to relevant spatial and non-spatial cues (see Figure 2 below). In the memory task, children viewed the scenes from the description task and after a brief delay had to choose which of three pictures matched the memory array based on color, size, or location of the target referent object. We hypothesized that children's memory for task-relevant information in this task would predict their spatial skills above and beyond the factors previously shown to relate to children's spatial cognition: age, gender, PPVT-IV score, task-related production, and adaptation score (Levine, Huttenlocher, Taylor, & Langrock, 1999; Miller et al., in press; Pruden et al., 2011; Voyer, Voyer, & Bryden, 1995).

Method

Participants

Sixty-nine-4-year-olds (M = 4.52, SD = 0.39 years, 31 females) participated in the study. An additional 13 children participated but were excluded due to: incomplete data (4), experimenter error (1), technical problems (2), parental interference (1), non-compliance (1), and not talking during the production task or insufficient vocabulary (3 and 1, respectively, described further below). Participants were recruited from a database compiled by a university research center. Participants received small prizes for participation.

Design and Procedures

Children were tested individually in the spatial scene description task (description task), the spatial scene memory task (memory task), three spatial tasks, and the Peabody Picture Vocabulary Test (PPVT) IV. Tasks were presented in a quasi-random order such that the first, third, and fifth tasks were always spatial tasks and the second and fourth tasks were the description and memory tasks, and the final task was the PPVT-IV. The order of the spatial tasks and the order of the description and memory task were counter-balanced across participants. Children received small prizes between tasks. Caregivers completed a productive spatial vocabulary checklist. Sessions lasted approximately 45 minutes each.

Spatial Scene Description Task Before the description task, children completed a warm-up task to help them understand the task and to feel comfortable talking aloud. In the warm-up task, children viewed 10 PowerPoint slides showing one familiar object on each slide and were asked to describe what they saw to a stuffed animal who was not looking at the screen ("Tell Bucky what you see"). The experimenter discouraged children from pointing during the warm-up and description tasks by instructing them to sit on their hands.

The description task tested children's abilities to disambiguate the target object's location (i.e., mouse) relative to a referent object. The task was conducted in PowerPoint. Each trial showed a picture of a spatial scene in the center of the slide. Each scene included three referent objects (e.g., beds) distributed diagonally across the screen (see Figure 1 above); the diagonal orientation (top-left to bottom-right vs. top-right to bottom-left) varied randomly across trials. The mouse was in a support relation to one referent object (target

referent object). On each trial, the child was asked to "Tell Bucky where the mouse is" to a stuffed animal not facing the screen. The trials varied in the number of relevant cues for describing the mouse's location such that the child could use 3 cues (color, size, and location, Figure 1A), two cues (color and location, size and location; Figure 1B and 1C), or one cue (location, Figure 1D) to describe the mouse's location.

The description task included 24 trials presented in one of four randomized orders. There were 6 trials per cue type, with 2 trials for each possible mouse's location (i.e., front, middle, back referent; see Figure 1). The diagonal alignment of the referent objects allowed children to use a range of different spatial words to describe the referent object's location (e.g., front, middle, back, first, last, left, right, top, center, bottom).

Spatial Scene Memory Task The memory task was modeled after the description task and tested children's memory for the features associated with the mouse's location. The PowerPoint slides from the description task were used in this task (in the same trial order) as memory arrays. Following the memory array presentation, there was a 1 s delay with a blank screen, and then the test array was presented. Test arrays probed memory for one dimension (color, size, or location) of the target referent object (Figure 2). Test arrays included three pictures along the bottom of the screen; these pictures showed only one referent object, and children were instructed to pick the picture that exactly matched the referent (e.g., box, ball, couch) the mouse was on in the memory array. Foil pictures varied along the probe dimension, for example, if the probe was color the referent objects in the foil scenes would be the same size and location as the target referent, but a different color (see Figure 2A). For each cue type (e.g., 2 cue-color and location, Figure 2C) of the memory array, children were presented twice with each probe type. The probe dimensions were randomized across trials and children did not know which probe would be tested in advance.

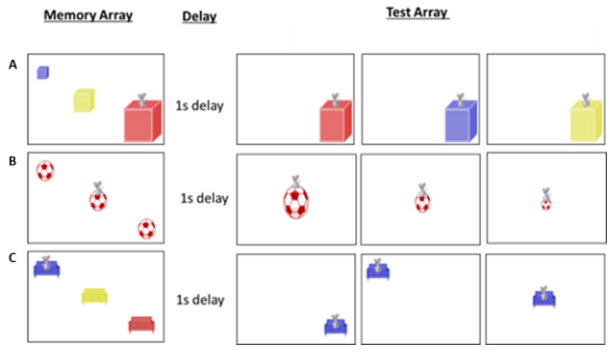


Figure 2: Sample trials of the spatial scene memory task: A) 3-cue memory array, color probe; B) 1-cue memory array, size probe; C) 2-cue memory array, location probe.

Spatial Tasks Children participated in short versions of three spatial tasks (shown in Figure 3): Spatial Analogies Task (Levine et al., 1999; Pruden et al., 2011), Children's Mental Transformation Task (Huttenlocher & Levine, 1990; Pruden et al., 2011), and Feature Binding Task (Dessalegn & Landau, 2008). For the Spatial Analogies and Mental Transformation Task, we adapted the short version from Pruden et al. (2011) and for the Feature Binding Task, we used half the number of trials as in Dessalegn & Landau (2008). In the Mental Transformation Task (10 trials; Figure 3A), children saw two pieces of a shape and selected which of four shapes the two pieces would make if combined. In the Spatial Analogies Task (13 trials; Figure 3B), children viewed a picture depicting two objects in a spatial relation and chose which of four other pictures shared that relation. In the Feature Binding Task (12 trials; Figure 3C), children saw a square with two different colors on opposite sides and chose a matching figure out of three options following a 1 s delay.

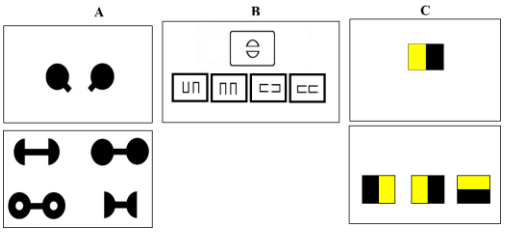


Figure 3: Sample trials of the spatial tasks: A) Mental Transformations; B) Spatial Analogies; C) Feature Binding.

Vocabulary assessments. The PPVT-IV measures receptive vocabulary and involved the children pointing to one of four pictures depicting the target word. The spatial vocabulary checklist measured productive spatial vocabulary and included 80 words taken from both the MCDI: Words and Sentences (Fenson et al., 1994) and from a spatial word coding manual. Caregivers indicated words they have heard their child produce (Cannon, Levine, & Huttenlocher, 2007)

Coding and measurement

Spatial scene description task Each session was transcribed by two research assistants. Transcribers were reliable on 89% of trials and discrepancies were resolved through a third research assistant blind to the first two transcripts. Final transcripts were coded by two different research assistants blind to the study hypotheses (mean 98% reliability across dimensions coded), and disagreements were resolved by the first author. Coders scored the number of times children mentioned color or spatial terms, and whether they used the terms correctly. For spatial terms, we separately categorized both size (e.g., small, medium) and location terms (terms referring to the referent's location in the scene, e.g., *top* chair) from the other types of spatial terms mentioned.

For the regression analysis, we calculated three measures from the coding as in Miller et al. (in press). For task-related production, we tabulated per trial (1) the quantity of non-spatial terms used by averaging the number of color terms produced, and (2) the quantity of spatial terms used by averaging the number of spatial terms (including but not limited to size and location terms). These variables were

created to ensure that any effects found with children's adaptive use of language did not result solely from the number of potentially-relevant words children could produce.

For (3) the adaptation score, we calculated how often each child produced relevant versus irrelevant cues during the description task. On each trial, a child's response was scored for the three cue types (color, size, and location) as a 0 or 1 depending on whether they produced any description of those cues. Multiple descriptions of the same cue type were only counted as one (i.e., "front corner" and "front" would both count as 1 for location cues on that trial). To account for the relevance of the cues, color and size terms were coded as negative if produced on trials when the terms could not differentiate the referents. Specifically, color was coded as negative on 2 cue (size and location) and 1 cue trials (Figure 1C and 1D), and size was coded as negative on 2 cue (color and location) and 1 cue trials (Figure 1B and 1D). On all other trials color and size cues were coded as positive, and location cues were always coded as positive because location was relevant on all trial types. Positive adaptation scores reflected children providing more relevant than irrelevant cues and negative adaptation scores reflected children providing more irrelevant than relevant cues. Scores closer to 0 reflected performance that did not differ by trial type; for example, a child who mentioned only color terms on every trial would receive a score of 0. Scores farther from 0 reflected both the number of cues mentioned and the cue relevance. Thus, the adaptation score was sensitive to the context in which children provided different types of cues.

To ensure that children had sufficient language knowledge to perform the task, we made two types of exclusions. We excluded trials for which the child's caregiver indicated on the checklist that their child did not produce any of the relevant location words (e.g., trials with the target on the middle object if the checklist indicated the child did not produce *middle* or *center*) and the child did not produce these words during the task. We also excluded trials from the description task if the child used an incorrect color, size, or location term (e.g., saying "top", when the object was on the bottom of the display). This resulted in the exclusion of 121 trials from 25 children's data (8% of total trials).

Spatial scene memory task and spatial tasks The experimenter marked responses on a session sheet during the study. Videos were used to code for reliability; 18 participants' sessions per task (26%, with different participants chosen for each task) were checked by a second research assistant, resulting in 99% agreement with disagreements resolved by a third research assistant. We created a spatial composite score by calculating the mean proportion correct in the three spatial tasks after normalizing for different chance levels (see Figure 3 above). We normalized scores by taking each child's score minus chance, then dividing by one minus chance, resulting in scores with 0 as chance and 1 as perfect performance.

For the memory task, we calculated a memory composite score for each child as their overall proportion of correct responses. We also grouped the trial types based on whether the probed dimension varied in the memory array (i.e., Figure 2A shows a trial where color was probed when it varied, and Figure 2B shows a trial where size was probed when it did not vary) to parallel the adaptation score from the description task. As in the description task, the memory arrays varied as to whether color and size cues differentiated the target referent object. However, unlike the description task, all cue dimensions were probed across trials and thus were potentially relevant. As such, instead of taking a difference score of relevant vs. irrelevant cues as in the adaptation scores, we created separate variables based on whether the probed dimension differentiated the target referent object.

We created three variables for this analysis: color/size (C/S) differentiated, C/S undifferentiated, and location differentiated. The C/S differentiated variable was calculated as children's mean proportion correct on color-probe trials on which color varied in the memory array (e.g., Figure 2A) and size-probe trials on which size varied in the memory array. The C/S undifferentiated variable was calculated as children's mean proportion correct on color-probe trials on which color did not vary in the memory array and size-probe trials on which size did not vary in the memory array (e.g., Figure 2B). The location differentiated variable was calculated as children's mean proportion correct on all location-probe trials (e.g., Figure 2C), as location always differentiated the referent objects. Although the location differentiated and C/S differentiated variables were similar (i.e., both include probe dimensions that varied in the memory arrays), we chose to calculate them separately to have equal number of trials (8) included for each variable.

Vocabulary assessments The experimenter marked children's responses while administering the PPVT-IV and terminated testing when the child responded incorrectly on eight or more trials within a 12-trial block (following the standardized instructions). Children's standardized scores were calculated offline using established norms (Dunn & Dunn, 2007). Productive spatial vocabulary was calculated as the proportion of words caregivers checked.

Results

This study investigated the relation between children's spatial skills and their attention to task-relevant information in both memory and language tasks. We hypothesized that individual differences in children's memory for relevant cues would predict their spatial composite score above and beyond demographics, vocabulary, quantity of task-related language production, and adaptive use of language. One participant's data was removed for being an outlier in the regression model (final $N = 68$). As a preliminary analysis, we tested whether there were differences in children's spatial composite scores based on whether they received the description or memory task first, but found no significant difference ($p = .926$) and thus excluded this variable from further analyses.

To evaluate our hypothesis, we conducted a regression analysis with two steps. The first step was conducted as a

replication of Miller et al. (in press) to test whether children's adaptive use of language predicted their spatial task performance. The second step evaluated our hypothesis by testing whether children's memory for task-relevant information predicted their spatial skills above and beyond their adaptive use of language. For the first step, children's spatial composite scores were regressed on demographics (age, gender), vocabulary (PPVT-IV, productive spatial vocabulary), quantity of task related language production (quantity of non-spatial and spatial terms used), and adaptation score. We found that children's adaptation score significantly predicted their spatial composite score (t_{60} = 2.21, p = .031, ΔR^2 = .045), replicating previous findings (Miller et al., in press). For the second step, we added each child's memory composite score to the model. We found that proportion correct on the memory task significantly predicted children's spatial composite score (t_{59} = 2.43, p = .018, ΔR^2 = .050) as shown in Figure 4A, and the variance accounted for by the adaptation score became marginal (p = .090)[1].

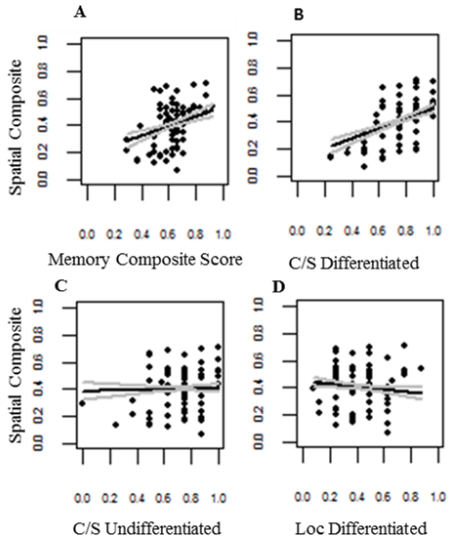


Figure 4: Relation between children's spatial composite score and A) memory composite score, performance on B) C/S differentiated trials, C) C/S undifferentiated trials, and D) location (Loc) differentiated trials. Error bars represent ±1 S.E. for point estimates from the regression models.

To test whether performance across the trial types differentially predicted spatial performance, children's spatial composite scores were regressed on C/S differentiated, C/S undifferentiated, and location differentiated (controlling for demographic, vocabulary, task-related production, and adaptation score variables as before), shown in Figures 4A-4C, respectively. Performance on C/S differentiated trials significantly predicted children's spatial composite scores (t_{57} = 3.77, p < .001, ΔR^2 = .101) but performance on C/S undifferentiated and location differentiated trials did not account for significant proportions of the variance in spatial composite score (p = .777, p = .265, respectively). The results are consistent with the hypothesis that attention to task-relevant cues is a general mechanism supporting spatial cognition and language use.

Discussion

This study tested whether children's attention to task-relevant cues in a memory task was more predictive of their spatial skills than their use of task-relevant cues in a production task. We posited that 4-year-olds' memory for task-relevant cues would predict their spatial composite score above and beyond demographics (age, gender), vocabulary (PPVT-score, productive spatial vocabulary), task-related production (quantity of non-spatial and spatial terms used) and adaptation score. Before adding the memory score, the model showed that adaptive language use predicted children's spatial composite scores after controlling for the other variables, replicating results that adaptive use of language—both spatial and non-spatial—is more predictive of spatial skills than spatial word production (Miller et al., in press).

Adding the memory composite score to the model showed that memory for task-relevant cues predicted children's spatial skills above and beyond all factors included in the model. The amount of variance accounted for by the adaptation score became only marginally significant when the memory composite score was added, indicating overlap in the contributions these measures make toward accounting for variance in children's spatial skills. Overall, these findings suggest that children's attention to cues in a spatial scene is related to their spatial performance, and could potentially account for much of the variance previously thought to be related to language. However, it is important to note that these findings do not exclude the possibility that other factors such as encoding or memory strategies may explain some of the variance in the memory task.

When we grouped memory task performance in a similar way to the description task, we found that only children's memory for color and size cues that differentiated the target referent objects in the memory array, when these dimensions were probed, significantly predicted spatial skills. These results are consistent with our hypothesis that attention to task-relevant cues underlies the relation between spatial skills and language. Children who correctly recognized the target referent object's color or size cues when these cues were differentiated in the memory array likely performed better on the spatial tasks because they are better at attending to cues that distinguish relevant information in the spatial tasks. This ability would also support more relevant language production in the spatial scene description task, leading to the shared variance across these tasks found in our results. Performance on the color and size undifferentiated trials likely did not

[1] The qualitative pattern of results remains the same with the outlier data included. The only quantitative difference is that the adaptation score is marginal in the first step (p = .089) and non-significant in the second step (p = .225) with the outlier data included.

predict spatial performance because in the spatial tasks features that do not differentiate among cues would not be relevant for solving the task. Thus, whether or not children attended to non-differentiated cues should not influence their performance as long as they attend to the relevant/differentiated cues. Although location cues always differentiated the target referent object, performance on location probes likely did not predict children's spatial skills due to poor performance (M = .42, SD = .49; see Figure 4D). Children rarely produced location terms in the spatial scene description task (5% of trials; similar to levels reported by Miller et al., in press), indicating that they may not have attended to location cues in the description or memory tasks. This is interesting because it suggests that attention to locations is not strongly related to spatial skills that rely on transforming or comparing relations among objects.

The current study provides novel insights into the mechanisms underlying the relations between language and spatial cognition. Of the potential causal models laid out in the introduction, our results suggest that a third factor connects spatial language and spatial skills. Further studies will be needed to test potential causal directions among the various factors that support developmental improvements in spatial skills and language use. Our research highlights the importance of considering the multiple cognitive processes that support spatial performance and provides insight into basic cognitive factors that may connect spatial skills to cognitive processes that support achievement in STEM domains

Acknowledgments

Thanks to the families who participated in this research, as well as to the research assistants who aided in data collection. Participant recruitment was partially funded by the Waisman Intellectual and Developmental Disabilities Research Center grant through NICHD (P30HD03352). Subsets of the data were presented at the 9th Biennial Meeting of the Cognitive Development Society and the 56th Annual Meeting of the Psychonomic Society.

References

Cannon, J., Levine, S. C., & Huttenlocher, J. (2007). A system for analyzing children and caregivers' language about space in structured and unstructured contexts. Spatial Intelligence and Learning Center (SILC) Technical Report.

Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. Psychological Science, *19*, 189–195.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, fourth edition: Manual.* Minneapolis, MN: Pearson Assessments.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185. http://doi.org//10.2307/1166093

Hermer-Vazquez, L., Moffet, A., & Munkholm, P. (2001). Language, space, and the development of cognitive flexibility in humans: The case of two spatial memory tasks. Cognition, *79*, 263–299.

Huttenlocher, J., & Levine, S. C. (1990). *Primary test of cognitive skills*. CGTB/McGraw-Hill.

Learmonth, A. E., Nadel, L., & Newcombe, N. S. (2002). Children's use of landmarks: Implications for modularity theory. Psychological Science, *13*, 337–341.

Learmonth, A. E., Newcombe, N. S., & Huttenlocher, J. (2001). Toddlers' use of metric information and landmarks to reorient. Journal of Experimental Child Psychology, *80*, 225–244.

Levine, S. C., Huttenlocher, J., Taylor, A., & Langrock, A. (1999). Early sex differences in spatial skill. Developmental Psychology, *35*, 940-949.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. Cognitive Psychology, *50*, 315–353.

Miller, H. E., Patterson, R., & Simmering, V. R. (2016). The role of language in children's use of spatial relations to remember locations. *Cognition*, *150*, 170–180.

Miller, H. E., Vlach, H. A., & Simmering, V. R. (in press). Producing spatial words is not enough: Understanding the relation between language and spatial cognition. *Child Development*.

Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? Developmental Science, *14*, 1417–1430.

Shusterman, A., Lee, S., & Spelke, E. S. (2011). Cognitive effects of language on human navigation. Cognition, *120*, 186–201.

Simms, N., & Gentner, D. (2008). Spatial language and landmark use: Can 3-, 4-, and 5-year-olds find the middle? In 29th Annual Conference of the Cognitive Science Society (pp. 191–196). Washingtion, D.C.: Lawrence Erlbaum Associates.

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Filipowicz, A. T., & Chang, A. (2014). Deconstructing building blocks: Preschoolers' spatial assembly performance relates to early mathematical skills. Child Development, *85*, 1062–1076.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin, *117*, 250.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, *101*, 817–835.

When the Words Don't Matter: Arbitrary labels improve categorical alignment through the anchoring of categories

Ellise Suffill (E.M.Suffill@sms.ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ UK

Holly Branigan (Holly.Branigan@ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ UK

Martin Pickering (Martin.Pickering@ed.ac.uk)
Department of Psychology, 7 George Square
Edinburgh, EH8 9JZ UK

Abstract

Novel labels provide feedback that may enhance categorical alignment between interlocutors. However, the nature of this feedback may not always be linguistic. Lupyan (2008) has demonstrated the effects of labels on individual categorization, and even non-word labels have seemingly produced greater consistency in sorting strategies (Lupyan & Casasanto, 2014). We extend this to alignment by demonstrating that arbitrary labels can increase sorting consistency to bring people's categories closer together, even without dialogue. Importantly, we argue that increased alignment is not always due to labeling in a linguistic sense. Results suggest that it is not the content of the non-word labels driving the alignment effects, but the very presence of the labels acting as 'anchors' for category formation. This demonstrates a more general cognitive effect of arbitrary labels on categorization.

Keywords: Alignment; Categorization; Labels; Lexical Effects.

Introduction

Shared understanding of the world is crucial for successful communication. Such *alignment* occurs through the formation of shared representations and the development of common ground (Clark & Brennan, 1991; Pickering & Garrod, 2004). Because of this, alignment is often discussed as a product of interaction. However, interaction may not always be necessary: Alignment can be considered in terms of the alignment of information states across individuals, rather than the explicit transfer of this information between them (Pickering & Garrod, 2006). For example, individuals may align without interaction if they simultaneously listen to the same speaker whilst in different rooms.

Alignment has often been studied with respect to phonological, lexical and syntactic information (Branigan, Pickering, & Cleland, 2000; Garrod & Anderson, 1987; Pardo, 2006). However, it may also include the alignment of conceptual information, and, thus, alignment can occur not only based on speakers' externally observable language, but also on their internal representations, such as alignment upon a given situation model (Zwaan & Radvansky, 1998). We posit multidirectional feedback between several levels of linguistic information available to interlocutors. This is in contrast to feed forward accounts, such as that of Levelt, Roelofs and Meyer (1999). Instead, feedback to the conceptual level could potentially occur from any of the other, more linguistic levels. Evidence already exists for the effects of linguistic feedback on alignment across multiple levels of processing: Both word repetition and semantic relatedness can enhance syntactic alignment (Branigan, Pickering, & Cleland, 2000; Cleland & Pickering, 2003). Thus, it is possible that linguistic information such as labels could affect conceptual alignment. We will focus on the effects of word-based information on conceptual alignment, and do so through the use of categories as proxies to concepts.

According to Murphy (2002), categories act as the external application of concepts to groups of items within the world. Storing concepts as categories allows us to deal with novel objects and having exemplars related to each concept aids us in correctly categorizing novel items with other similar and previously encountered items. Thus, categories are essential to the way in which we respond to items. In language, we reinforce these categories through the application of labels that aid us in learning category boundaries (Lupyan, 2006). Thus, it seems plausible that this dependency on labels could affect categorization.

Effects of Existing Labels on Categorization

Lupyan (2008) has provided evidence for the *representational shift* account, which posits that labeling can cause a distortion of items most reliably associated with a category label, in terms of the physical features associated with that specific item. Memory was worse for items that were labeled at a basic level (e.g., chair vs. table), compared to those that were not labeled. Recall decreased more as a result of this labeling. This suggests that applying a category label to an item causes the item to become a mix of its idiosyncratic features and those typically associated with the category. Lupyan posits that this leads to greater

consistency in categories through the selection of more robust category attractors.

Lupyan and Spivey (2010) demonstrated that the effects of labeling are not limited to single items and can work across item groups. They found that hearing a redundant label leads to faster detection of visual cues placed near the 'labeled' items. For example, hearing 'chair' resulted in faster detection of a probe placed near one of multiple chairs in a display, despite participants having already been told to attend to chairs in prior instructions. Thus, labels can facilitate the process of visual perception, by affecting how individuals pay attention to groups of items, even when the labels appear redundant.

Utilizing Novel Labels

Non-word labels can be used in experiments to examine interactions between the characteristics of the labeled stimuli, and the meanings that develop for the novel labels applied to them. Using novel words has the advantage of the labels not being associated to specific entries in the mental lexicon. This gives greater potential for new label meanings to develop on the basis of the stimuli properties. Lupyan and Casasanto (2014) demonstrated that the application of meaningless, non-word labels (e.g., 'fooves' vs. 'crelches') to novel, 'alien' stimuli can facilitate the categorization process. Categorization was based upon physical characteristics, such as the smoothness versus pointedness of the aliens' heads. Results demonstrated that the novel labels worked as successfully as conventional labels (e.g., 'smooth' vs. 'pointy') at increasing sorting consistency. However, the form of the non-word labels strongly reflected sound symbolism for the physical characteristics of the items and, thus, there were expectations regarding how the labels would be applied to the stimuli (e.g., 'foove' referring to 'smooth'; cf. Köhler, 1929). Thus, it is questionable how arbitrary these labels were when placed within such a specific context. If one wanted to investigate the potential effects of more arbitrary labels, then sound symbolism should be reduced as much as possible, and stimuli could equally be made more abstract (i.e. by not having characteristics that can be readily linked to the non-word labels). Then, one could examine whether novel labels intrinsically improve consistency in categorization by developing associated meanings, if indeed labeling effects remained.

Extension to Alignment

In order to examine the nature of novel labeling effects, we developed a task that allowed us to investigate categorical alignment by utilizing non-word labels with more simplistic, abstract stimuli. For half of the experiment, participants sorted stimuli into groups with access to non-word labels, whilst in the other half of the experiment they did not have access to the non-word labels. Half of the participants were exposed to their partner's categorizations throughout, and half were not exposed. Participants were not allowed to explicitly communicate with each other in any part of the categorization task (i.e. no dialogue or gesturing).

We aimed to test whether any labeling effect remained when the non-word labels had as little influence of prior word-meaning associations as possible and when they could not be consistently linked to the stimuli via sound symbolism. Specifically, we asked whether the effect was due to the labels acquiring meaning, or whether it was due to a less linguistic process by which labeling increased consistency in participants' sorting strategies without acquiring meaning.

We also examined whether the effects of labeling were affected by exposure to one another's sorting strategies. The Exposed condition tested for the effect of labels on the alignment of categories through participants seeing their partner's categories at intervals throughout the task. The Non-exposed condition focused on whether alignment would be higher when categorization occurred alongside access to labels, regardless of the fact that participants did not have exposure to each other's categories (i.e., if there would be greater consistency between individuals when they had access to labels, compared to when they did not have labels). If the non-word labels increased sorting consistency by developing an associated meaning, then it is possible that this meaning could be communicated across the individuals within a pair, given the shared nature of words and associated meanings (Laskowski, 2010). This could lead to differences in alignment for the Exposed versus Non-exposed conditions, as information transfer as an effect on alignment could only occur in the Exposed setting.

The items to be categorized were randomly generated, triangular shapes (Laskowski, 2010). The non-word labels were provided in order to help participants divide up the stimuli into more discrete categories. The dependent variable of the study was how aligned participants were in their assignment of items to categories i.e., the extent to which their categorizations overlapped.

Method

Participants

64 British participants (45 female) formed 32 experimental pairs. Ages ranged from 18 to 27 years, with an average of 19.86 years. All participants identified themselves as native, monolingual speakers of English. Written consent was obtained prior to testing. The University of Edinburgh's Ethics Committee approved this study.

Stimuli

Perceptually morphed, triangular shapes (from Laskowski, 2010) were used as experimental items (see Figure 1). Shapes were morphed across the dimensions of size, shape, angles and pointedness. Items were printed on 3x3 inch cards. There were two sets of triangles (Set A and B), each comprising 26 items.

The non-word labels were "TEB" (/tɛb/) and "DUP" (/dʌp/). These were presented as printed labels alongside the

stimuli sets in the With-labels condition, and were absent in the No-labels condition. The labels were constructed to be similar in linguistic qualities, so as to reduce the possible effects of sound symbolism. That is, both /t/ and /d/ are alveolar plosives, /ɛ/ and /ʌ/ are open-mid, unrounded vowels (although they vary slightly in position), and /b/ and /p/ are both bilabial plosives.



Figure 1: Items 1-6 from Set A as examples of stimuli.

Design

There were two independent factors: Exposure (between-participants; Exposed vs. Non-exposed) and Labels (within-participants; With-labels vs. No-labels). In the Exposed conditions participants viewed their partner's categories at intervals throughout the study. In the Non-exposed conditions, participants never saw the other person's categories. For Labels, participants who had access to non-word labels (With-labels) in one block (e.g., trials 1–5) would not have labels (No-labels) in the second block (e.g., trials 6–10), and vice versa. We counterbalanced in which block participants had access to labels and which set (A or B) they sorted first. A barrier was used to obscure the participants from each other. In the Exposed conditions, this barrier remained in place for the sorting phases, but was removed during the intervals so that participants could see each other as well as their item groups. In the Non-exposed condition, the barrier obscured the view of the other participant and their item groups for the entirety of the experiment.

Participants were instructed to categorize the stimuli into two groups using a minimum group size of 9 and a maximum group size of 17, but could use any number between this (e.g. 12 and 12, or 16 and 10, etc.). This is because limiting group sizes to a specific number could have led to more unnatural category formation, by forcing participants to place items under categories that they would not otherwise have chosen to.

Procedure

The experimental set up and procedure was as follows, with each experiment comprising 10 rounds (see Figure 2):

1. Participants were given the instructions; "Sort the triangles into two groups in a way that would make sense to you, as well as to another person".
2. Upon receiving the labels participants were told; "You have these two labels to place upon your groups. Place one label on one group each. You choose how to use them. You can move them across groups between rounds if you wish"
3. Participants faced each other across a table with the barrier in place.
4. Participants sorted one set of A or B into two groups, with a time limit of 2 ½ minutes. All triangles had to be assigned to a group.
5. There was a 30-second interval between sets. In the Exposed conditions, participants saw their partner's groups (plus labels, if they were on the Labels block of the experiment); in the Non-exposed condition participants did not see their partner's sorting strategies.
6. The barrier was replaced (for Exposed conditions) and participants sorted the next 4 sets.
7. After trial 5 (end of block one), participants took a short break from the task in which they filled in demographic information.
8. In the second block, whichever set was not used in block one was sorted over trials 6–10.
9. Participants completed a post-test questionnaire regarding their sorting strategies and whether they applied meaning to the labels.

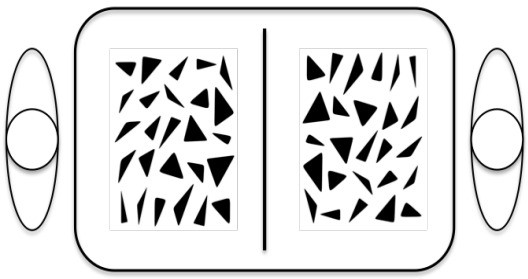


Figure 2: Two participants faced each other across a table with the central barrier in place.

Results

Calculating Alignment Scores

Scoring reflected how similarly the two participants in a pair split items across groups. For each set, a participant would finish with two groups of items. Each item was identifiable to the coder by number. If both Participants A and B placed items 1–8 in 'group 1' and items 18–26 in 'group 2', this would result in an alignment score of 17 for the round. We then reversed the assignment of groups for Participant A and determined the alignment score. We selected the higher of these two scores (as assignment of group number was arbitrary).

Descriptive Statistics

The average alignment scores for each Condition were numerically higher in With-labels blocks relative to No-labels blocks (see Table 1).

Table 1: Mean alignment scores (*SD*) by Block for each Condition of Exposure by Labels.

Condition	Mean Alignment	
	Block1	Block2
Ex+Labels1st	20.15 (4.00)	19.05 (3.82)
Ex+Labels2nd	18.6 (4.30)	19.85 (2.93)
Non-Ex+Labels1st	18.63 (3.74)	17.4 (2.76)
Non-Ex+Labels2nd	18.48 (3.27)	18.6 (3.47)

Linear Mixed Effects (LMER) Analysis

In order to test for the effects of Labels and Exposure on alignment, data were analyzed in R 3.2.1 (R Core Team, 2015) with a linear mixed-modeling approach via the lme4 package, version 1.1-8 (Bates, Maechler, Bolker & Walker, 2013). This approach was chosen as it allowed us to account for random variance due to differences between participant pairs. Under this approach, the threshold for statistical significance was set at $|t| > 2$. A backwards, stepwise elimination approach was used to select factors for the final model, using likelihood ratio tests to compare models. Models included random slopes and intercepts for the variable Labels, with Pair (participant pairs) also entered as a random effect. The initial model included Set (A or B), Labels (With-labels or No-labels), Exposure (Exposed or Non-exposed) and Order (Labels 1st or Labels 2nd). Removing Order as a fixed effect did not significantly reduce model fit ($X^2(1)=0.61$, $p=.44$). Removing Exposure as a fixed effect also did not significantly reduce model fit ($X^2(1)=2.07$, $p=.15$). However, removing Labels as a fixed effect did significantly reduce model fit ($X^2(1)=6.13$, $p=.01$), as did removing Set ($X^2(1)=7.51$, $p=.01$). Including the interaction between Set and Labels did not contribute significantly to model fit ($X^2(1)=2.46$, $p=.12$). The best-fit model therefore included the main effects of Set and Labels as fixed effects (see Table 2). With-labels produced significantly higher alignment scores than No-labels ($\beta=0.93$, SE=0.37, $t=2.52$). Thus, there was a significant effect of Labels, regardless of Exposure. Set A also produced significantly higher alignment scores than Set B ($\beta=1.04$, SE=0.37, $t=2.83$).

Table 2: Beta, standard errors and t-values for fixed effects on alignment score. Model fit by REML.

Fixed Effects	β	S.E.	t
Intercept	-0.98	0.50	-1.96
Set(A)	1.04	0.37	2.83
Labels(With)	0.93	0.37	2.52

Table 3: Variance and residual for random effects. Model fit by REML.

Random Effects		
Pair	Intercept	5.383
	Labels(With)	1.252
Residual		7.67

No. of Observations = 320.

Labels-specific Analysis

A labels-specific analysis was conducted using LMER to assess whether having labels matched across groups led to significantly higher alignment scores. Matched status was dependent on whether the non-word labels matched across the groups used to calculate a pair's alignment scores: that is, whether Participant A placed specific items under TEB and others under DUP in the same manner as Participant B. For example, if Participant A's group 1 was compared to Participant B's group 1 in order to calculate their alignment score, matched status would depend on whether both participants placed the same label upon that group. There were 160 data points overall for the With-labels condition, with 4 similarly sized groups in terms of data points across Match and Exposure conditions (see Table 4).

Table 4: No. of observations and mean alignment (*SD*) for Matched vs. Unmatched labels for Exposed and Non-exposed Conditions

Exposure	Match	*N*	Mean (*SD*)
Exposed	Matched	42	20.76 (2.99)
	Unmatched	36	19.26 (3.87)
Non-exposed	Matched	38	18.81 (3.83)
	Unmatched	44	18.48 (3.51)

Models included random slopes and intercepts for Set, Exposure and Match (Matched or Unmatched), with Pair (participant pairs) entered as a random effect. The initial model included Set, Exposure, Match and Order as fixed effects. Removing Order did not significantly reduce model fit ($X^2(1)=1.02$, $p=.31$). Removing Match also did not significantly reduce model fit ($X^2(1)= 0.48$, $p=.49$). Removing Exposure significantly reduced model fit ($X^2(1)= 4.23$, $p=.04$), as did removing Set ($X^2(1)= 3.91$, $p=.04$). Including the interaction between Set and Exposure did not significantly increase model fit ($X^2(1)=1.00$, $p=.32$). Therefore, the model of best fit included Set and Exposure as a fixed effects (see Table 5). The Exposed condition produced significantly higher scores than the Non-exposed condition ($\beta=2.07$, SE=0.79, $t=2.61$). Set B produced significantly lower scores than Set A ($\beta=-1.74$, SE=0.79, $t=-2.19$).

Table 5: Beta, standard errors and t-values for fixed effects on alignment score. Model fit by REML.

Fixed Effects	β	S.E.	t
Intercept	19.16	0.61	31.28
Set(B)	-1.74	0.79	-2.19
Exposure(Expo)	2.07	0.79	2.61

Table 6: Variance and residual for random effects. Model fit by REML.

Random Effects		
Pair	Intercept	0.74
	Set(B)	4.33
	Exposure(Expo)	9.48
	Match(Match)	2.54
Residual		7.81

No. of Observations = 160.

Qualitative Analysis

Data from post-test questionnaires were analyzed to check for patterns in assigned label meanings. In 32 instances (50% of participants), one or both of the labels were reported to hold no significance to the individual. There was a great deal of variation in the meanings that the remaining participants assigned to the labels, as reflected by the 34 different terms that they used to refer to them. These ranged across references to stimuli size, length, angle size, roundness of corners and even number of items.

Sound symbolism effects associated with the slight differences in vowel position for TEB and DUP appear highly unlikely to explain the effects: There were only 3 instances in which both participants within a pair reported assigning meanings to the label that could be considered semantically similar. For example, one participant associated DUP with 'Fat' and TEB with 'Thin', whereas their partner associated DUP with 'Heavy' and TEB with 'Light'. This high variability in meanings and low occurrence of matches in label associations is consistent with our finding that label content did not significantly contribute to the increased alignment scores across participants in a pair, at least on an explicit level.

Discussion

Pairs of participants were more likely to categorize abstract stimuli in the same way when they had access to non-word labels, than when they did not. This is consistent with Lupyan and Casasanto's (2014) proposals, in that it suggests that arbitrary, non-word labels may gain significance in meaning and, through this, focus the categorization process to create more consistent sorting patterns. In the current study, the labeling effect held regardless of exposure (i.e., whether participants saw each other's sorting strategies, or not), which suggests labels may increase consistency in sorting across individuals, so that people's categorizations are more aligned without interaction. However, examining data only from trials in which participants did have access to labels does suggest an effect of exposure. It appears that whilst arbitrary labels can improve alignment without the need for interaction, having exposure to a partner's sorting strategies can build upon this labeling effect.

Exposure Effects in Alignment

Language transmission studies have demonstrated how novel language systems for open-ended meaning spaces can begin to evolve categorical structure with enough exposure across generations of participants, as sorting patterns are transmitted from one individual to the next, and as a function of how much iteration occurs (Kirby, Cornish & Smith, 2008; Carr, Cornish & Kirby, 2015). This categorical structure appears to develop on the basis of cognitive biases for physical features such as item shape and size.

As such, these processes could occur within the current paradigm, but with information being repeatedly passed between a pair of individuals, rather than being transmitted down from one generation to the next. For internal consistency in the non-exposed condition, iteration would reinforce sorting strategies through repeated exposure to items, in order to increase overall consistency within and across individuals (and this would be facilitated by the presence of non-word labels). As for alignment in the exposed condition, a similar scenario could develop, but with transmission occurring at the interval in which participants transfer information regarding sorting strategies via exposure.

Labeling vs. Anchoring

We found that, whereas alignment scores were higher in the presence of non-word labels, alignment was not affected by specifically which items the labels were applied to. This was corroborated by the qualitative analysis of label meanings reported by participants in the post-test questionnaire. Although it is possible that implicit labeling effects could have occurred, it does not seem that any consistent meaning was tied to TEB or DUP in a way that would enhance alignment.

Instead, the labels may have led to higher categorical alignment through a form of 'anchoring', by providing a point of reference for each category, but without explicitly linking to that category through an associated meaning. In a sense, anchoring can be thought of in terms of Kirsh's (1995) markers, as opposed to perfect reminders. That is, the labels may act as markers that indicate to the individual that there is an important distinction to be made between the items (the two categories into which they are to be sorted), without telling us anything specific about the nature of that distinction. Existing labels (or novel labels that have gained an associated meaning) act as perfect reminders, in that they can communicate something important about the category (and category items) that they are applied to. Thus, the anchoring effect demonstrated here could not be readily referred to as linguistic, but is instead something seemingly more general, perhaps linking to other cognitive functions

commonly considered as external to the linguistic domain. Specific, cognitive functions that language is argued to support include memory, attention and simplification of tasks, as well as the development of representations (Clark, 1998). Given this, the effect could stem from having the label there as a "word", rather than it mattering specifically what that word is.

Conclusion

The effect of labels on categorical alignment is seemingly not always due to a linguistic process of associating concepts to words as linguistic units. In the current study, the labels did not appear to lead to higher categorical consistency and alignment through linguistic labeling effects, but through a more general cognitive process of anchoring, or marking out the distinction between categories. This finding has important implications for studies investigating what may appear to be 'labeling' effects. The non-word labels led to higher alignment even in the absence of exposure between pairs, demonstrating that alignment does not rely exclusively on interaction and sometimes can occur through increased consistency across non-interacting individuals. However, it does appear that exposure to a partner's sorting strategies can build upon the effects of labeling on alignment. Future studies should aim to discover the generalizability of this anchoring effect, by introducing new forms of potential markers to the categorization process, as well as by investigating at what point the markers could potentially gain communicative meaning through interaction.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 613465. A very special thank you to Eleanor Drake and Jia Loy for their support, patience and wit throughout this project.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(4).

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13-B25.

Carr, J. W., Cornish, H., & Kirby, S. (2014, April). The cumulative cultural evolution of category structure in an open-ended meaning space. Paper presented at the 10th International Conference on the Evolution of Language, Vienna, Austria.

Clark, A. (1998). Magic words: how language augments human computation. In P. Carruthers and J. Boucher, eds, *Language and thought: interdisciplinary themes* (pp.162–183). Cambridge, MA: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on Socially Shared Cognition*, *13*, 127-149.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214-230.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, *73*(1), 31-68.

Köhler, W. (1929). *Gestalt psychology*. Oxford, England: Liveright.

Laskowski, C. (2010). *The role of language in conceptual coordination*. Unpublished PhD thesis, University of Edinburgh.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01), 1-38.

Lupyan, G. (2006). Labels facilitate learning of novel categories. In A. Cangelosi, A.D.M. Smith, & K. Smith (Eds.), *The Sixth International Conference on the Evolution of Language* (pp. 190–197). Singapore: World Scientific.

Lupyan, G. (2008). From chair to" chair": a representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General, 137*(2), 348-396.

Lupyan, G., & Casasanto, D. (2014). Meaningless words promote meaningful categorization. *Language and Cognition*, *7*(2), 167-193.

Lupyan, G., & Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. *Attention, Perception, & Psychophysics, 72*(8), 2236-2253.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT press.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382-2393.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169-190.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2-3), 203-228.

R Core Team (2015). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162.

Can the High-Level Semantics of a Scene be Preserved in the Low-Level Visual Features of that Scene? A Study of Disorder and Naturalness

Hiroki P. Kotabe (hkotabe@uchicago.edu)
Omid Kardan (okardan@uchicago.edu)
Marc G. Berman (bermanm@uchicago.edu)

Department of Psychology, 5848 South University Avenue,
Chicago, IL 60637 USA

Abstract

Real-world scenes contain low-level visual features (e.g., edges, colors) and high-level semantic features (e.g., objects and places). Traditional visual perception models assume that integration of low-level visual features and segmentation of the scene must occur *before* high-level semantics are perceived. This view implies that low-level visual features of a scene alone do not carry semantic information related to that scene. Here we present evidence that suggests otherwise. We show that high-level semantics can be preserved in low-level visual features, and that different high-level semantics can be preserved in different types of low-level visual features. Specifically, the 'disorder' of a scene is preserved in edge features better than color features, whereas the converse is true for 'naturalness.' These findings suggest that semantic processing may start earlier than thought before, and integration of low-level visual features and segmentation of the scene may occur after semantic processing has begun, or in parallel.

Keywords: low-level visual features, scene semantics, semantics, scene recognition, visual perception, scene gist, visual processing

Introduction

A scene of an environment contains a lot of information that we perceive as "features," broadly construed. There are lower-level visual features such as edges and colors and higher-level semantic features such as recognizable objects, places, and descriptors (Oliva & Torralba, 2001; Rosch & Mervis, 1975). Here we focus on two specific semantic features of a scene—its level of 'disorder' and its level of 'naturalness'—due to their psychological importance (Berman, Jonides, & Kaplan, 2008; Kotabe, Kardan, & Berman, 2016; E. O. Wilson, 1984; J. Q. Wilson & Kelling, 1982). Traditional non-Gestalt visual perception models suggest that integration of low-level visual features and segmentation of the scene must occur *before* high-level semantic features are perceived (e.g., Biederman, 1987; Treisman & Gelade, 1980; Marr, 1976). This would imply that low-level visual features do not intrinsically carry information about high-level semantic features. Here we question this assumption by asking, can the low-level visual features of a scene preserve any of the high-level semantics of that scene? Furthermore, is it possible that different high-level semantics are preserved in different types of low-level visual features?

The preservation of high-level semantics in low-level visual features would be of import to theories of visual perception, posing a challenge especially to those that assume that semantic processing starts later in visual perception. First, it would suggest that semantic processing may start earlier than thought previously. Second, it would suggest that integration of low-level visual features and segmentation of the scene may occur after semantic processing has begun, or in parallel.

We know of some of work that is relevant to this idea. First, although it may seem improbable that humans can start to process semantics from information carried by low-level visual features, before objects are perceived, we note that the brain is a meaning making machine that can even find meaningful objects in white noise (Gosselin & Schyns, 2003). Furthermore, there is ample evidence that people can rapidly identify the semantic category of a scene—a remarkable feat considering the subtle comparisons one must make among the large number of scenes within a scene category (e.g., imagine the number of scenes one could consider 'natural'), not to mention the large number of scene categories. After only 20 ms of exposure to a scene, people can categorize whether the scene contains an animal or not with about 94% accuracy (Thorpe, Fize, & Marlot, 1996). This is a shorter duration than used in some subliminal priming experiments! After only 27 ms of exposure to a scene, people can recall seeing semantic features, as evidenced by a free-recall experiment (Fei-Fei, Iyer, Koch, & Perona, 2007). After only 33 ms of exposure to a scene, people can not only categorize objects in a scene (e.g., dog) but can even identify within-category kinds (e.g., a German Shepherd) above chance (Grill-Spector & Kanwisher, 2005). Even if scenes are jumbled into six parts and presented for only 50 ms, people can categorize the gist of the scenes better than chance. After 100 ms of exposure to a scene, people can perceive if an object is incompatible within the scene (Biederman, Teitelbaum, & Mezzanotte, 1983). Although we do not examine the time course of visual information processing in this study, we see this research as consistent with the idea that semantic processing starts very early in visual processing.

There is other support for our hypothesis. An electroencephalogram (EEG) experiment showed that low-level category-dependent processing can occur in less than a tenth of a second after presentation of a stimulus (Vanrullen

& Thorpe, 2001). Not only was the presentation rapid, but category-dependent brain processing started soon after exposure, consistent with low-level visual information carrying semantic information. There are studies that suggest that people can identify the semantic category of a scene in the near absence of attention (Fei-Fei, VanRullen, Koch, & Perona, 2005; Li, VanRullen, Koch, & Perona, 2002). Other research suggests that object semantics are processed prior to figure-ground segmentation (Peterson & Gibson, 1994). A patient study suggests that objects can be recognized when early visual processing is intact but recognition of object parts is impaired, consistent with low-level visual features carrying global semantic information (Davidoff & Warrington, 1999). At least one study suggests that it takes the same amount of time to detect an object as it does to categorize it (Grill-Spector & Kanwisher, 2005), which is contrary to semantic processing starting at a higher and more time-delayed level. An fMRI study also supports this idea by showing that scene categories could be decoded from activity in V1 (Walther, Caddigan, Fei-Fei, & Beck, 2009). In fact, the decoding accuracy of V1 (26%) was not that far off from the decoding accuracy of the parahippocampal place area (31%), which is known to be a key region involved in processing scene semantics. All of these studies cast doubt on traditional hierarchical models of visual perception.

Specifically concerning the preservation of semantics in low-level spatial features, Oliva and Torralba (2001) presented the spatial envelope model which proposes that the global spatial layout of a scene, defined by specific low-level visual feature configurations, carries information about the semantic category (e.g., natural vs. built) of that scene (see also Oliva & Torralba, 2006). This computational model suggests that segmentation and the processing of individual objects or regions is not necessary for classifying scenes into semantic categories.

As for the preservation of semantics in low-level color features, although some research suggests that color information is not critical for the rapid categorization of scenes (Delorme, Richard, & Fabre-Thorpe, 2000; Fei-Fei et al., 2005), other research suggests otherwise. Oliva and Schyns (2000) showed that color information helps people categorize scenes into semantic categories when the color information is diagnostic of a semantic category. Follow-up research by Goffaux et al. (2005) provided both behavioral and EEG evidence that diagnostic color information is part of the scene "gist" (Oliva, 2005) that facilitates rapid scene recognition. This is consistent with other research that suggests that prior experience benefits rapid scene understanding (Greene, Botros, Beck, & Fei-Fei, 2015). In fact, Goffaux et al. (2005) showed that atypical scene colors hinder rapid scene recognition.

Specifically concerning the preservation of semantics related with disorder and naturalness in low-level visual features, we showed that the disorder of a scene could be predicted by objective low-level visual features (Kotabe et al., 2016), and we have shown that this is also true for naturalness (Berman et al., 2014). Relatedly, Oliva and Torralba (2001) showed that naturalness could be predicted based on the principal components of power spectra which capture orientation and spatial frequency information. It is unclear, however, whether these results are possible because the disorder and naturalness of a scene systematically varies non-causally with certain low-level visual features (e.g., the low-level visual features relate with objects that convey semantics related with disorder or naturalness, rather than conveying semantics themselves) or because high-level semantics are actually preserved in low-level visual features. Here we test the latter possibility.

Notes on General Method

We sampled broadly from real-world environments by utilizing 260 colored images of environments that ranged from more urban to more natural (according to previously-collected ratings, Berman et al.; Kardan et al., 2015) and from more orderly to more disorderly (according to previously-collected ratings, Kotabe, Kardan, & Berman, 2016) (see Figure 1 for examples; all images can be downloaded here in original resolution: https://goo.gl/IKHXeC). We manipulated these scene images by extracting and scrambling their low-level edge features and their low-level color features. We had people rate these derived stimuli in terms of disorder or naturalness. Together, that gave us both disorder and naturalness ratings for the scrambled-edge stimuli, the scrambled-color stimuli, and the unaltered scene images. Data analysis was conducted on the image-level mean ratings.

Note that we did not use the rapid scene recognition paradigm for this study. Although this paradigm is useful for the study of the time course of visual information processing, it does not directly test whether low-level visual information carries high-level semantic information. It also relies on recognition instead of directly testing perception. The method we used, which involved freely rating semantic dimensions of presented scenes, directly measured the perception of these semantic dimensions. Furthermore, by taking these measurements between-subjects, we eliminated memory issues including the possibility that high-level semantics are preserved in low-level visual features only when one has previously viewed the unaltered scene (thus has memory of the scene and low-level visual features), or when one has previously viewed its low-level visual features in a scrambled stimulus (thus has memory of the low-level visual features).

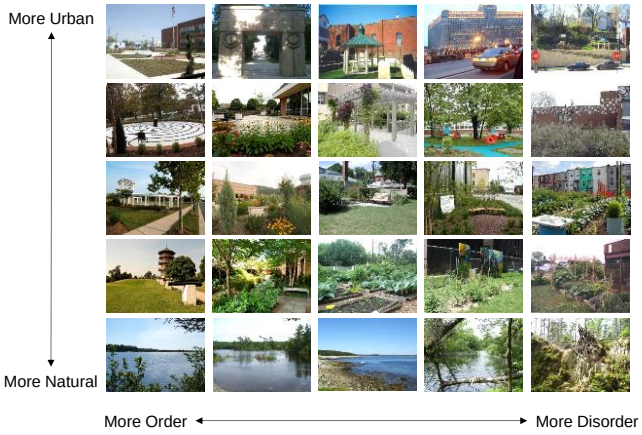


Figure 1: Sample images from our set of 260 scene images varying in naturalness and disorder.

Experiment 1: Is Disorder Preserved in Edges?

We extracted and scrambled the edges and colors of the 260 scene images. We had people rate these derived stimuli in terms of disorder. We then tested the association of these disorder ratings with the disorder ratings of the original scenes to see if disorder was preserved in the low-level visual edge or color features. Based on our previous work, which suggests that edges matter more than colors for the perception of disorder (Kotabe et al., 2016), we predicted that the disorder ratings of the scrambled-edge stimuli would correlate stronger with the disorder ratings of the original scenes than would the disorder ratings of the scrambled-color stimuli.

Method

Participants and design 191 US-based adults (108 men, 82 women, 1 other) were recruited from the online labor market Amazon Mechanical Turk (AMT) and participated in this two-condition (stimuli: scrambled edges vs. scrambled colors) between-subjects experiment.

Extracting and scrambling edges We developed a method to extract and scramble the edges of the scenes such that the average low-level edge properties (e.g., edge continuities, straight and non-straight edge ratios) would be preserved but colors and identifiable segments and objects would be removed. In this method, a matrix, called the mask matrix, was constructed to be the same size as the original images (600*800) with its elements randomly assigned between zero and one. This matrix was then convolved with a median filter sized 30*40 pixels. In this way, patches of 1s and 0s were made randomly and placed at random locations across the mask with random sizes equal to or greater than the 30*40 pixels, with half of every mask having, on average, half a surface of 1s and half a surface of 0s. Next, the edge map of the target image created as in (Berman et al., 2014; Kardan et al., 2015) was randomly rotated either 90 or 270 degrees and overlaid on the 180 degrees rotated edge map, creating a stimulus consisting of less identifiable stimuli with twice as many edges (but same straight and non-straight edge ratios) as the original image. This stimulus was then multiplied (dot product) by the mask so that half of its edges got removed at random. The resulting stimulus had, on average, the same amount of edges with similar edge types, but no identifiable segments or objects from the original image (see Figure 2b for an example).



Figure 2: (a) Example of a highly natural scene, (b) its derived scrambled-edge stimulus, and (c) its derived scrambled-color stimulus.

Extracting and scrambling colors For the scrambled-color stimuli (see Figure 2c), we randomly repositioned windows of 5*5 pixels from the image. The window size was selected so that: 1) segments and objects became non-discernible, and 2) to keep the color texture of the scene visible. For example, using a 1*1 pixel window size resulted in stimuli in which less frequent colors were so scattered that they became invisible to the eye. Using a 10*10 pixel window kept some of the segments or objects identifiable.

Procedure Participants were first given a brief introduction to the image-rating task. They were instructed, "You will be presented with a series of 50 images containing various lines (colors). We simply want you to rate each image in terms of how disorderly or orderly it looks." We intentionally did not define "disorder" or "order" because we were interested in people's natural and spontaneous definitions of disorder, which turned out to be surprisingly uniform. Participants were then randomly presented 50 of the 260 scrambled-edge stimuli (Experiment 1a) or scrambled-color stimuli (Experiment 1b) on a plain white background. The randomization scheme had two layers. First, we randomly selected 10 images from each quintile of urbanness/naturalness. Second, we presented these 50 images in random order. This ensured that each participant would view a wide sample of images from more urban to more natural. For each image, they were instructed to rate the scene in terms of disorder on a seven-point Likert-type scale ranging from "very disorderly" to "very orderly." The task would continue to the next image immediately after a rating was made. By not fixing presentation time, we would not artificially make people view the scenes for shorter or longer than they wanted to, which could have influenced their perceptions.

Results

We correlated the disorder ratings of the scrambled-edge and scrambled-color stimuli with the previously collected

disorder ratings of the original scenes. Disorder ratings of the scrambled-edge stimuli significantly correlated with disorder ratings of the original scenes, $r = .38$, $p < .001$, providing first evidence that disorder was partially preserved in the low-level edge features (see Figure 3a). In contrast, disorder ratings of the scrambled-color stimuli *did not* significantly correlate with disorder ratings of the original scenes, $r = .02$, $p = .731$ (see Figure 3b), suggesting that disorder was not preserved as much in the color features. In support, the difference between these two dependent correlations was statistically significant, $t = 4.43$, $p < .001$, according to Williams' test (1959). One may note that the disorder ratings for the scrambled-color stimuli varied less ($SD = 0.52$) than the disorder ratings for the scrambled-edge stimuli ($SD = 0.90$), which may explain the small correlation of $r = .02$ insofar as reduced variance in X is related to reduced covariance between X and Y, so we conducted Thorndike Case 2 correction for range restriction setting the unrestricted *SD* of disorder ratings for the scrambled-color stimuli equal to the *SD* of disorder ratings for the scrambled-edge stimuli which increased the correlation to $r = .03$, $p = .63$, still consistent with edges preserving disorder semantics better than colors.

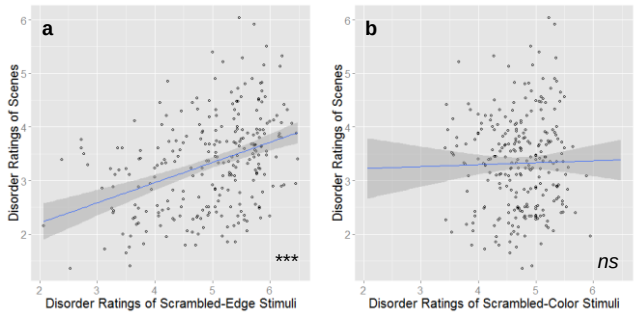


Figure 3: Results of Experiment 1. (a) Disorder ratings of scrambled-edge stimuli significantly correlated with the disorder ratings of scene images. (b) Disorder ratings of scrambled-color stimuli did not significantly correlate with the disorder ratings of scene images. Least-squares lines with 95% confidence bands shown. *** $p < .001$

Because of imperfect linearity, we also tested these associations with two nonparametric tests of association based on rank-order, Spearman's rho (ρ) and Kendall's tau-b (τ). Disorder ratings of the scrambled-edge stimuli were again significantly associated with disorder ratings of the original scenes according to both tests, $\rho = .40$, $p < .001$ and $\tau = .27$, $p < .001$, providing further evidence that disorder was partially preserved in the low-level edge features. In contrast, disorder ratings of the scrambled-color stimuli were again *not* significantly associated with disorder ratings of the original scenes according to both tests, $\rho = .02$, $p = .414$ and $\tau = .01$, $p = .758$, again suggesting that disorder was not preserved as much in the color features. In support, the difference between the two dependent ρs was statistically significant, $t = 4.58$, $p < .001$, according to Williams' test.

These results suggest that high-level semantics related to disorder at the scene-level were preserved in the low-level edge features of the scenes but not as much in the low-level color features of the scenes. This provides evidence that high-level semantics can be preserved in low-level visual features, and more specifically, that some types of low-level visual features carry certain semantic information better than others. But is it possible that different semantic information is preserved better in different low-level visual features? Specifically, is 'naturalness' better preserved in colors than in edges because colors are more diagnostic of naturalness (Oliva & Schyns, 2000)? We tested this possibility in the following experiment. We note that this would be contrary to the spatial envelope model (Oliva & Torralba, 2001) which suggests that naturalness is a perceptual dimension that is well-represented by the *spatial* structure of a scene.

Experiment 2: Is Naturalness Preserved in Colors?

We extracted and scrambled the edges and colors of the 260 scene images. We had people rate these derived stimuli in terms of naturalness. We then tested the association of these naturalness ratings with the naturalness ratings of the original scenes to see if naturalness was preserved in the low-level visual edge or color features. We predicted that the naturalness ratings of the scrambled-color stimuli would correlate stronger with the naturalness ratings of the original scenes than would the naturalness ratings of the scrambled-edge stimuli, under the assumption that colors are more diagnostic of naturalness.

Method

Participants and design 186 US-based adults (118 men, 67 women, 1 other) were recruited from AMT and participated in this two-condition (stimuli: scrambled edges vs. scrambled colors) between-subjects experiment.

Procedure The procedure was the same as in Experiment 1 except that participants rated naturalness on a seven-point Likert type scale ranging from "very urban" to "very natural."

Results

The analysis followed the same procedure as in Experiment 1. Naturalness ratings of the scrambled-color stimuli significantly correlated with naturalness ratings of the original scenes, $r = .24$, $p < .001$, providing first evidence that naturalness was partially preserved in the low-level color features (see Figure 4c). In contrast, naturalness ratings of the scrambled-edge stimuli *did not* significantly correlate with naturalness ratings of the original scenes, $r = -.06$, $p = .358$ (see Figure 4a), suggesting that naturalness was not preserved as much in the edge features. In support, the difference between these two dependent correlations was statistically significant, $t = 3.52$, $p < .001$, according to Williams' test. One may note that the naturalness ratings for the scrambled-edge stimuli varied less ($SD = 0.40$) than the disorder ratings

for the scrambled-color stimuli (*SD* = 0.68), which may explain the small correlation of r = -.06, so we conducted Thorndike Case 2 correction for range restriction setting the unrestricted *SD* of naturalness ratings for the scrambled-edge stimuli equal to the *SD* of naturalness ratings for the scrambled-color stimuli which increased the correlation to r = -.10, p = .11, still consistent with colors preserving naturalness semantics better than edges.

There was an outlying clustering of scenes rated as highly natural (see Figure 4). After removing these with a cutoff of 6.5/7.0 on the naturalness scale (N = 212 remaining), the results provide even stronger support for our hypothesis. Naturalness ratings of the scrambled-color stimuli significantly correlated with naturalness ratings of the original scenes, r = .44, p < .001 (see Figure 4d). In contrast, naturalness ratings of the scrambled-edge stimuli *did not* significantly correlate with naturalness ratings of the original scenes, r = .02, p = .358 (see Figure 4b), suggesting that naturalness was not preserved as much in the edge features. In support, the difference between these two dependent correlations was statistically significant, t = 4.87, p < .001, according to Williams' test.

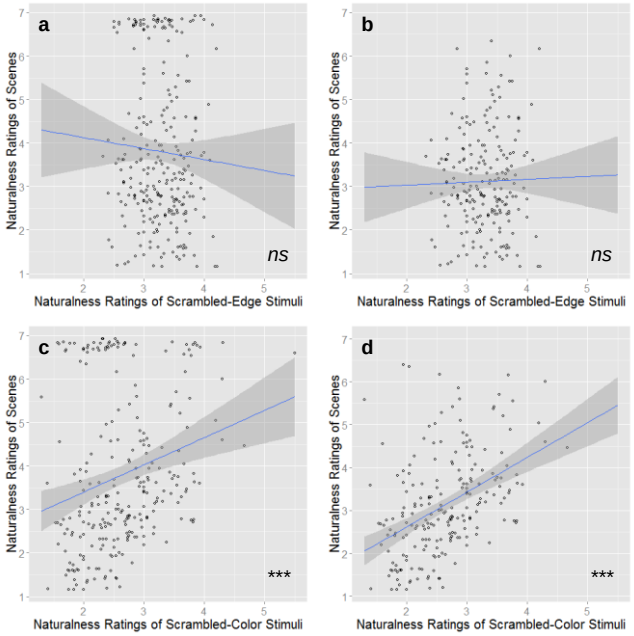


Figure 4: Results of Experiment 2 before and after removing cluster of highly natural scenes. (a-b) Naturalness ratings of scrambled-edge stimuli did not significantly correlate with the naturalness ratings of scene images; (c-d) Naturalness ratings of scrambled-color stimuli significantly correlated with the naturalness ratings of scene images. Least-squares lines with 95% CI shown. *** p < .001

Because of imperfect linearity, we also tested these associations with Spearman's rho and Kendall's tau-b. Naturalness ratings of the scrambled-color stimuli were again significantly associated with naturalness ratings of the original scenes according to both tests, before, ρ = .29, p < .001 and τ = .21, p < .001, and after, ρ = .48, p < .001 and τ = .34, p < .001, removing the cluster of highly natural scenes, providing further evidence that naturalness was partially preserved in the low-level color features. In contrast, naturalness ratings of the scrambled-edge stimuli were again *not* significantly associated with naturalness ratings of the original scenes according to both tests, before, ρ = -.07, p = .288 and τ = -.05, p = .287, and after, ρ = -.01, p = .851 and τ = -.01, p = .879, removing the cluster of highly natural scenes, again suggesting that naturalness was not preserved as much in the edge features. In support, the difference between the two dependent ρs was statistically significant before, t = 4.27, p < .001, and after, t = 5.74, p < .001, removing the highly natural scenes, according to Williams' test.

These results suggest that high-level semantics related to naturalness at the scene-level were preserved in the low-level color features of the scenes but not as much in the low-level edge features of a scene. This further supports our general hypothesis that high-level semantics can be preserved in low-level visual features. It also further supports our more specific hypothesis that some low-level visual features carry certain semantic information better than others.

Conclusion

Together, these experiments provide direct evidence that high-level semantics can be preserved in low-level visual features, and that different high-level semantics can be preserved in different types of low-level visual features. This is evidenced by our two experiments, the first showing that high-level semantics related with disorder were preserved better in low-level edge features than in low-level color features, and the second showing that high-level semantics related with naturalness were preserved better in low-level color features than in low-level edge features. This research adds to the body of literature that is starting to entertain the possibility that object perception and segmentation do *not* need to occur before identifying the semantic category of a scene.

Acknowledgments

This work was supported by a grant from the TKF Foundation and an internal grant from the University of Chicago.

References

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., … Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PLOS ONE*, *9*, e114572.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological Science*, *19*, 1207–1212.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., Teitelbaum, R. C., & Mezzanotte, R. J. (1983). Scene perception: A failure to find a benefit from prior expectancy or familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 411.

Davidoff, J., & Warrington, E. K. (1999). The bare bones of object recognition: Implications from a case of object recognition impairment. *Neuropsychologia*, *37*(3), 279–292.

Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: A study in monkeys and humans. *Vision Research*, *40*(16), 2187–2200.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision*, *7*(1), 10.

Fei-Fei, L., VanRullen, R., Koch, C., & Perona, P. (2005). Why does natural scene categorization require little attention? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, *12*(6), 893–924.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*(6), 878–892.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *14*(5), 505–509.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: Rapid scene understanding benefits from prior experience. *Attention, Perception, & Psychophysics*, *77*(4), 1239–1251.

Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, *16*(2), 152–160.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., … Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom-up processing of the visual features of nature? *Frontiers in Psychology*, *6*, 471.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016). *The order of disorder: Deconstructing visual disorder and its effect on rule-breaking*. Manuscript submitted for publication.

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601.

Marr, D. (1976). Early processing of visual information. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *275*(942), 483–519.

Oliva, A. (2005). Gist of the scene. *Neurobiology of Attention*, *696*(64), 251–258.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176–210.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

Peterson, M. A., & Gibson, B. S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science*, *5*(5), 253–259.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Vanrullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience*, *29*(34), 10573–10581.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B*, *21*, 396–399.

Wilson, E. O. (1984). *Biophilia*. Cambridge, MA: Harvard University Press.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly*, *249*, 29–38.

Is the Self-Concept like other Concepts? The Causal Structure of Identity

Stephanie Y. Chen (stephanie.chen@chicagobooth.edu)
Daniel M. Bartels (bartels@chicagobooth.edu)
Oleg Urminsky (oleg.urminsky@chicagobooth.edu)

University of Chicago Booth School of Business
5807 South Woodlawn Ave.
Chicago, IL 60637 USA

Abstract

We investigate the age-old questions of what makes us who we are and what features of identity, if changed, would make us a different person. Previous approaches to identity have suggested that there is a type of feature that is most defining of identity (e.g., autobiographical memories or moral qualities). We propose a new approach to identity that suggests that, like concepts in general, more causally central features are perceived as more defining of the self-concept. In three experiments, using both measured and manipulated causal centrality, we find that changes to features of identity that are perceived as more causally central are more disruptive to both the identity of the self and others.

Keywords: self-concept; concepts and categories; causal reasoning; personal identity

Introduction

Recent research has found that personal identity influences a wide-range of in-lab and real-world decisions (e.g., Akerlof & Kranton, 2010; Bryan et al., 2011). While this research has explored the relationship between identity and decision making, it largely ignores what underlies these decisions, the representation of the self. How is the self-concept structured and what is most defining of the self-concept?

Understanding these questions is particularly important because beliefs about continuity of identity—whether or not we will be the same person in the future—have been linked to how we value future outcomes. People who believe that their identity will change tend to discount future rewards more steeply and show less willingness to give up immediate rewards to save money than those who believe it will remain stable (Bartels & Urminsky, 2011, 2015). This paper explores how people represent the self and what features of identity people believe they need to retain to remain the same person.

Previous explorations of self-continuity have provided various answers to the question of what is most defining to identity. Philosophers have long suggested that continuity of memories allows for continuity in identity (Locke 1694/1979). In particular, autobiographical memories provide a unique narrative for each individual. Nichols and Bruno (2010) found that disruption to memories disrupted identity judgments not only for one's self but also in judgments of other people. Other explorations of lay intuitions about personal identity have suggested that the things that make us distinctive are those that make us who we are. This includes autobiographical memories, since they are unique for each individual. Other features that also make us distinct and may be essential to our self-concept are personality traits and preferences (Gelman et al., 2007).

More recent research has suggested that lay theories of identity put moral qualities at the center of the self-concept. Strohminger and Nichols (2014) examined this *essential moral self hypothesis* by comparing how changes to moral features and various other types of features (e.g., memories, personality, preferences) impacted identity continuity judgments. They found that changes to the moral features of identity were most disruptive to identity judgments.

To bridge these approaches to identity and provide a new framework for encompassing these discrepant findings, we appeal to the idea, from concept and categories literature, that more causally central features are more defining of a concept (Ahn, et al., 2000; Rehder & Hastie 2001; Sloman, Love, & Ahn, 1998). We propose that representations of the self-concept are like representations of concepts in general and hypothesize that 1) the self-concept incorporates causal relationships between the features of identity and, 2) more causally central features are perceived as more defining of identity. For example, the importance of memories, traits, or preferences for the self-concept depends on how these features are causally related to each other and to other features of identity. Although ideas about causal centrality have been highly influential in the study of concepts, this is the first time that this approach has been incorporated into a theory of personal identity.

Causal Centrality and Continuity of Identity

In the following experiments, we measured beliefs about how defining features of identity are to the self-concept and beliefs about the causal centrality of these features. Beliefs about the how defining features are to identity were measured by asking participants how much a change to a feature would disrupt their identity (i.e., the extent to which a change in a feature would lead them to feel that they were a different person). Changes to features that are more defining to the self-concept should lead to greater perceived disruption to identity. To measure causal centrality, we elicited beliefs about how various features of identity are causally linked either by having participants draw these causal links in a concept map (Experiment 1) or by having them verbally report these links (Experiment 2). Based on participants' reports, we calculated causal centrality of features two ways described below.

Number of Causal Connections Features may be defining to a concept to the extent that they participate in cause-

effect relationships with other features of the concept (Rehder & Hastie, 2001). This measure calculates causal centrality as the number of direct causal relationships a feature participates in as either a cause or an effect.

Causal Depth We also calculate causal centrality according to Sloman et al.'s (1998) dependency model[1]. This model suggests that the causal centrality is determined by causal depth. That is, causes are more central than their effects; the deeper a feature is in the causal chain the more central it is. This model takes into account both the feature's direct and indirect effects, and the strength of these causal links.

Data Analysis In Experiments 1 and 2, for both measures of causal centrality described above, we calculated the Spearman correlation between causal centrality and disruption to identity two ways: averaged across participants and individually for each participant. If more causally central features are more defining to identity, causal centrality and disruption to identity will be positively correlated such that changes to more causally central features would be more disruptive to identity continuity.

Experiment 1

The first experiment examined whether people perceive causally central features of identity as more defining of their own identity. Participants performed two tasks: 1) the concept map task, which measured the causal centrality of 16 features of personal identity and, 2) the identity questionnaire, which measured how defining each feature was to the participant's personal identity.

Method

Design Participants were randomly assigned to one of two between-participants conditions. The conditions served to counterbalance the order the two tasks were performed in.

Participants Participants were 92 University of Chicago students. Twelve participants were excluded due to computer program failures, either to record data or to display randomized features, yielding 80 cases.

Procedure

Concept Map Task Each participant drew a computerized map of the causal links between 16 features of their identity (Table 1). Twelve of the sixteen features were chosen from categories of personal identity identified as important in the prior literature (memories, personality, morality, and preferences/desires; e.g., Strohminger & Nichols, 2014). The remaining four features were intended to be low importance. Two were found, in previous research, to be less important for identity (instances of semantic memories, Strohminger & Nichols, 2014) and two (fillers) were found to be unimportant for identity in a pretest.

Participants used ConceptBuilder software to report beliefs about causal relations (Kim & Park, 2009), first in an unrelated practice task and then to draw the causal map of personal identity using the 16 features. The features were initially presented on the screen in random order. Participants could move the features and draw unidirectional or bidirectional arrows between them to represent cause-effect relationships, as they saw fit (see Figure 1). For each link specified, participants also rated the strength of the causal relationship (1 = weak, 2 = moderate, 3 = strong).

Identity Questionnaire Participants rated each feature on how much a change in the feature would disrupt their identity on a scale of 0 (exactly the same person) to 100 (completely different person).

Table 1: Features used in Experiments 1 and 2.

Features	Category
Cherished memories of time with family	Autobiographical memory
Important childhood memories	Autobiographical memory
Memories of important life milestones	Autobiographical memory
Height	Filler
Level of Hunger	Filler
Level of Wholesomeness	Morality
Level of Honesty	Morality
Level of Loyalty	Morality
Intelligence Level	Personality
Degree of Shyness	Personality
Reliability	Personality
Goals for personal life	Preferences/desires
Favorite Hobbies/Activities	Preferences/desires
Aesthetic Preferences	Preferences/desires
Knowledge of math	Semantic memory
Knowledge of music	Semantic memory

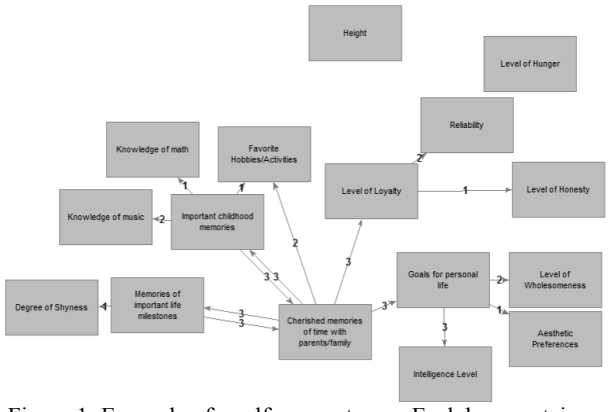


Figure 1: Example of a self-concept map. Each box contains a feature of personal identity. The arrows represent causal relationships between features. The numbers on each arrow indicate the strength of the causal relationship.

Results & Discussion

On average, participants drew 20.0 causal links between the features. This did not significantly differ across conditions ($M_{mapfirst}$ = 21.2, $M_{surveyfirst}$ = 18.7, $t(78) = 1.3$, $p > .05$.

[1] According to this iterative model, C_i, the centrality of feature i, is determined (at each time step) by summing across the centrality of the concept's other features (at time, t), $c_{j,t}$, multiplied by how dependent each feature, j, is on feature i, d_{ij}:

$$c_{i,t+1} = \sum_j d_{ij} c_{j,t}$$

The implementation of the model is a repeated matrix multiplication that comes to a stable ranking within a small number of iterations (Kim & Park, 2009; Sloman et al., 1998).

Number of Causal Connections Consistent with our hypothesis, we found a significant overall Spearman correlation between the average causal connections and rated disruption to identity, $r = .79$, $p < .001$. This positive relationship between causal connections and disruption to identity was observed for 80% of participants. The mean individual-level correlation (within-participant across all features) between feature centrality and importance[2] was significantly positive ($M = .33$), $t(79) = 7.4$, $p < .001$.

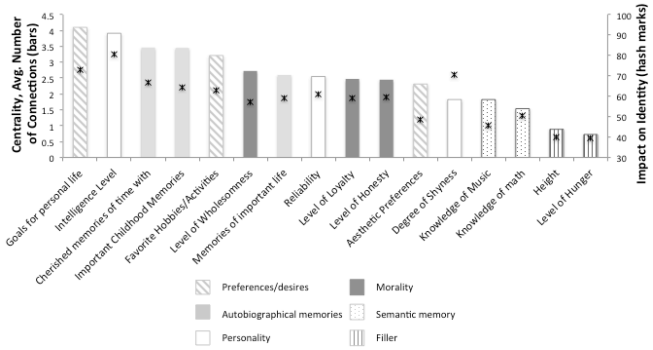


Figure 2: Results of Experiment 1 (number of causal connections). Causal centrality is shown on the left y-axis (bars). The impact a change in each feature had on identity is shown on the right y-axis (x marks).

Causal Depth We also found evidence that this measure of causal centrality was associated with identity judgments. The Spearman rank correlation between causal depth and disruption to identity ratings was significant, $r = .65$, $p = .01$ (see Figure 3). The mean individual-level correlation between feature centrality and disruption to identity was significantly positive ($M = .23$), $t(79) = 5.3$, $p < .001$. This positive relationship between causal depth and disruption to identity was observed for 78% of participants.

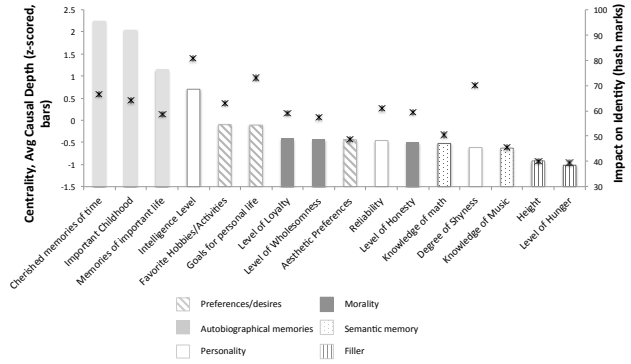


Figure 3: Results of Experiment 1 (causal depth). Causal centrality is shown on the left y-axis (bars). The impact a change in each feature had on identity is shown on the right y-axis (x marks).

Experiment 2

Experiment 1 found that changes in more causally central features of identity were seen as more disruptive to the continuity of one's own identity. Although our perceptions and evaluations of ourselves can be strikingly different from how we perceive others (Pronin, 2008), people use analogies to the self in forming judgments of even dissimilar others (Orhun & Urminsky, 2013). Experiment 2 tested whether our findings generalize to judgments of other people's identities.

Design Participants were assigned to one of three conditions (self, close-other, generic-other).

Participants Two-hundred-fifty participants were recruited via Amazon Mechanical Turk. Eleven participants were excluded before analysis, either because of a scripting error (5), failed attention check (4), or giving all the same answers (2), yielding 239 cases.

Procedure Participants in the self condition completed a causal centrality task and the identity questionnaire from Experiment 1 for features of their own identity. Participants in the close-other condition did the same tasks for a non-romantic close other they specified, while participants in the generic-other condition completed the tasks for a generic other person.

To measure centrality, participants completed a "listing causal relationships" task, using the 16 features of identity from Experiment 1. After a practice task with feedback, participants completed 16 individually randomized trials. In each trial, participants saw a different target feature and indicated which of the other 15 features, if any, were caused by the target feature (see Figure 4). Then, for all the features selected as direct effects of the target feature, participants rated the strength of the relationship.[3]

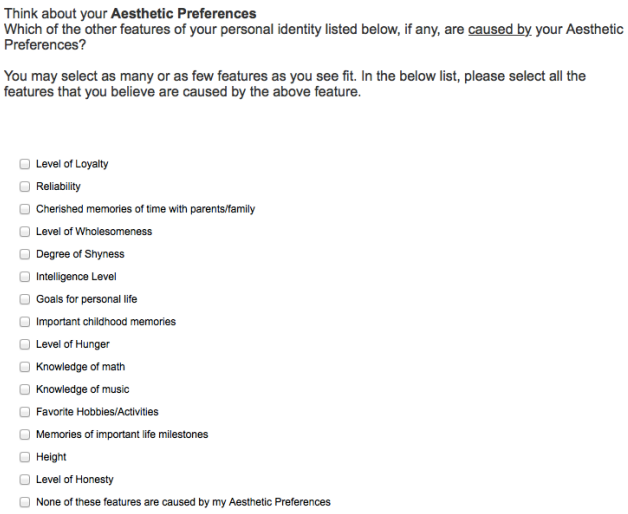


Figure 4: Illustration of listing causal relationships task.

Results & Discussion

On average, participants reported 75.7 causal links between the 16 features of identity. The number of links chosen did not differ by condition ($M_{self} = 71.1$, $M_{close\text{-}other} = 78.1$, $M_{generic\text{-}other} = 77.9$, $F(2, 238) = .69$, $p > .05$), suggesting participants perceived similar causal complexity in others' personal identity as in their own.

[2] All correlations reported in Experiments 1 and 2 are Spearman rho, and Fisher transformations were performed prior to t-tests.

[3] In a separate pre-test, we confirmed that this task yielded similar causal centrality scores to Experiment 1's concept map task.

Table 2: Experiment 2, Summary of Individual-Level Correlation Analysis Results.

	Causal Connections		Causal Depth	
Condition	Mean Spearman Correlation	% Participants with Positive correlations	Mean Spearman Correlation	% Participants with Positive Correlations
Self	.34, t(78)=7.3, $p < .001$	77%	.26, t(78) = 5.2, $p < .001$	72%
Close-other	.38, t(78)=9.1, $p <. 001$	84%	.32, t(78) = 6.6, $p < .001$	80%
Generic-other	.30, t(80)=6.3, $p < .001$	74%	.22, t(80) = 4.8 $p < .001$	72%

Number of Causal Connections Changes in features with more causal connections were rated as more disruptive to identity in all conditions (self: $r = .60$, $p = .02$; close-other: $r = .62$, $p = .01$; generic-other: $r = .44$, $p = .09$). The differences between conditions in disruptiveness of the more causally connected features were not significant (self vs. close-other, $p = .84$, self vs. generic-other, $p = .17$, close-other vs. generic-other, $p = .12$). Likewise, the average individual-level correlations were significantly positive and the majority of participants had a positive correlation between features' causal connections and rated disruptiveness of change in all conditions (see Table 2).

Causal Depth Changes in more causally central features were rated as more disruptive to identity in the self and close-other conditions (self: $r = .49$, $p = .05$; close-other: $r = .65$, $p = .01$) but not in the generic-other condition ($r = .42$, $p = .11$). The correlations in the close-other and generic-other conditions were significantly different ($p = .04$).[4]

In all conditions, the average individual-level correlations were significantly positive and the majority of participants had a positive individual-level correlation between features' causal centrality and rated disruptiveness of change (see Table 2).

Experiment 3

There were two main aims of Experiment 3. The first aim was to manipulate the causal centrality of features in a set vignettes to test whether making a feature more causally central impacts how defining that feature is for identity. In order to do this, we constructed vignettes that described the causal relationships between four salient features of a person. For example, one vignette described four of Jack's features as relating to each other in a common cause structure—Jack's memories of being a lonely child caused his shyness, his preference for solitary activities, and his awkward demeanor (Figure 5, Version A). In order to manipulate whether a feature was causally central or peripheral, we created two versions of each vignette. In another version of the above vignette the position of two target features (shyness and memories) were flipped so that Jack's shyness caused his memories, preferences, and demeanor (Figure 5, Version B). So, the exact same features were counterbalanced to be causally central (cause feature) and causally peripheral (effect feature), to control for any idiosyncratic influences of specific features.

The focal task was to select which of two individuals, one missing the effect feature (e.g., shyness in Version A) and one missing a cause feature (e.g., memories in Version B), was more likely to be the character in the story. As according to both approaches to causal centrality, the cause features is more causally central than the effect feature (it is both involved in more causal connections and deeper in the causal chain), retaining the cause feature should be more important for continuity of identity. So, we predicted that participants would choose the individual who is missing the effect feature (and retains the cause feature) as being more likely to be the character in the vignette.

The second aim of Experiment 3 was to understand more precisely how causal beliefs influence identity judgments. To do this, we created common effect (one effect with three causes) versions of all the vignettes which allows us to distinguish between the two approaches to causal centrality. For example, the common effect version of Jack's vignette presented his childhood memories as an effect of the other three features, including shyness (Figure 5, Version C). As with the common cause vignettes, we created two versions of the common effect vignettes to counterbalance the position of two target features in the causal structure. The other common effect version of Jack's vignette presented his shyness as an effect of the other three features, including his memories (Figure 5, Version D).

Unlike the common cause vignettes, the two accounts of causal centrality make different predictions about which individual should be selected in the common effect vignettes. The causal depth approach predicts that the individual missing the effect is more likely to be judged as the character in the story. This is because features deeper in the causal chain (the cause features) are more important to categorization, and in this case, to identity continuity. In contrast, the number of causal connections approach predicts that participants will tend to pick the person missing the cause feature. This is because the effect feature is linked to all three cause features while each cause is only linked to one other feature, the effect feature. So, based on the number of causal connections measure, childhood memories would be more causally central than shyness in Versions C, whereas shyness would be more causally central than childhood memories in Versions D.

[4] The difference between the self and close-other conditions, and the difference between the self and generic-other conditions were not significant ($p = .14$, $p = .58$, respectively).

Method

Design Participants were randomly assigned to read one of two sets (Set 1 vs. Set 2) of six vignettes which counterbalanced which vignettes described a common effect vs common cause structure.

Participants Sixty participants were recruited via Amazon Mechanical Turk. Four participants were removed before analysis, either for answering the comprehension check question wrong or for failing an attention check.

Materials We constructed six vignettes that described the causal relationships between four features of a person's identity. There were four versions of each vignette that contained the same four features. Two described a common cause structure and two described a common effect structure.

The vignettes were split into two sets (Sets 1 and 2). Each set contained both common cause versions for three vignettes and both common effect versions for the other three vignettes (e.g., for the Jack vignette Set 1 contained Versions A and B, Set 2 contained Versions C and D, see Figure 5). Participants were randomly assigned to read one of the two versions of each vignette in the set (which counterbalanced the position of features in the causal structure, e.g., Version A or B in the common cause version). So, participants read only one version of each of the six vignettes.

Procedure Each participant read the six vignettes in the set that they were assigned to. Diagrams like those in Figure 5 accompanied the vignettes and could be used to answer the questions. The order of presentation of the vignettes was randomized within-participants.

To measure which feature was seen as more defining to identity, after reading each vignette, we asked participants which of the two people was most likely to be the character in the vignette. One person was missing a cause feature; the other was missing an effect feature. Participants read that these people retained the other three features of the character in the vignette. Participants then reported how plausible they felt the vignette was on a scale of 0 (not at all plausible) to 100 (extremely plausible).

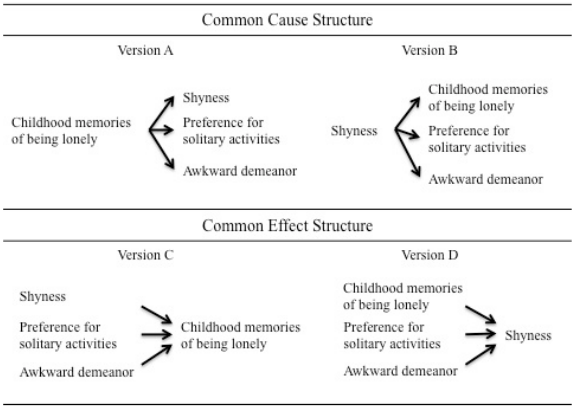


Figure 5: Structure of vignettes used in Experiment 3. There were four versions of each vignette: two versions for each causal structure. Versions A/B and C/D counterbalanced the placement of the target features as cause or effect.

Results & Discussion

The dependent measure was the average of the individual-level percentage of trials in which the participant selected the person missing the effect feature. For the common cause trials, we predicted that participants would pick the person missing the causally peripheral effect feature as the same person rather than the person missing the causally central cause feature. This is what we found. Participants were significantly more likely to select the person missing the effect feature (M = 72%, $t(55) = 5.3$, $p < .01$), replicating the prior findings with a causal manipulation.

For the common effect trials, however, the two approaches to causal centrality yield different predictions. The number of causal connections approach suggests a missing effect should disrupt identity more than a missing cause. So, based on this definition of causal centrality, the results should be the reverse of the common cause trials, with participants picking the person missing the effect *less* than the person missing the cause. In contrast, the causal depth approach implies that a missing cause should disrupt identity more than a missing effect would. This approach predicts that participants will pick the person missing the effect *more* than the person missing the cause, just as they did for common cause trials.

Our results are more consistent with the number of causal connections approach. In the common effect condition, participants were *less* likely to select the person missing the effect feature—the feature that had more connections but was less deep—than the person missing the cause feature (M = 41%), $t(55) = 2.6$, $p = .01$. The average percentage of missing effect selections was significantly different between the common cause and common effect conditions, (Ms = 72% vs. 41%), $t(55) = 5.2$, $p < .01$.

Because different causal structures may also differ in how natural they appear to be (Ahn, 1999), we also examined the plausibility of the two different types of causal structures. The common cause and common effect vignettes were rated as equally plausible ($M_{commoncause}$ = 73.9, $M_{commoneffect}$ = 72.5, $t(55) < 1$, $p > .05$). So, the observed difference in selections between the two conditions cannot be explained by a difference in the believability of the two causal structures. These results experimentally corroborate our correlational findings. Change in more causally central features, as defined by the number of causal connections, is more disruptive to perceived continuity of identity.

Lastly, we looked at the effects of vignette plausibility. Concepts, in general, are influenced by our prior knowledge (Murphy & Medin, 1985). This suggests that prior intuitions about what causal relationships are likely to occur among features of identity may moderate identity judgments. The higher the perceived plausibility, the more likely that participants believed that the feature described as having fewer connections was actually causally peripheral and the more likely that they selected the person missing the causally peripheral feature.

We found the predicted moderation by plausibility. We found a correlation between the plausibility of the story and

the proportion of selections of the person missing the causally peripheral feature with fewer connections (r = .64, p < .001). The average proportion of selections of the person missing the causally peripheral feature was significantly higher among the twelve most plausible vignettes (M = 74.2%) than the twelve least plausible vignettes (M = 55.7%), $t(22) = 2.5$, p = .02.

General Discussion

We found that, as with concepts in general, causal beliefs play a critical role in the representation of the self-concept. People perceived more causally central features as being more necessary for continuity of identity, for both the self (Experiment 1) and others (Experiment 2). Manipulating the causal centrality of a feature changed perceptions of how defining that feature was to identity (Experiment 3).

While both the number of causal connections and causal depth related to how defining a feature was to identity, the results of Experiment 3 suggest that the number of causal connections better describes how causal beliefs influence identity judgments. This suggests that features that either cause many other features or are caused by the combination of many other features (or both) will be most defining of identity. As causes generally occur before their effects, this means that the features that develop early in our lives (e.g., childhood memories) will not necessarily remain the most defining to identity, contrary to what a causal depth approach would suggest. Rather how defining an early-developing feature is to identity will depend on how many other features it causes. Late-developing features will be important if they are caused by the combination of other features and, in turn, cause new features.

Our approach to identity may also have implications for understanding affective responses. In fact, it has been suggested that differences in representations of the self-concept can explain differences in affective responses to stress and failure. People with whose self-aspects are highly overlapping or highly associated tend to have more trouble coping with negative experiences because failure in one aspect is likely to spillover into other aspects of the self (Linville, 1987; McConnell, 2011). Our results suggest that understanding the causal relationships between these aspects may be useful to predicting what types of negative experiences will be most impactful on mood and affect.

Prior research has focused on comparing the individual importance of different types of features. These approaches seem to have missed a critical aspect of representations of identity, beliefs about the causal relationships between features. These relationships influence the extent to which a feature defines identity. These findings are consistent with the general drive to explain the world, including ourselves, using causal relations (Gopnik, 1998; Keil, 2006).

Acknowledgments

We thank Halley Bayer for help with data collection, and Douglas Markant and Sarah Molouki for helpful comments on this work. This research was supported in part by the John Templeton Foundation grant #40357.

References

Ahn, W. (1999). Effect of causal structure on category construction. *Memory & Cognition*, *27*(6), 1008-1023.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41*, 361-416.

Akerlof, G., & Kranton, R. (2010). *Identity Economics: How Our Identities Shape Our Work. Wages, and Well-being.* Princeton, NJ: Princeton University Press.

Bartels, D. M., & Urminsky, O. (2011). On intertemporal selfishness: How the perceived instability of identity underlies impatient consumption. *Journal of Consumer Research*, *38*, 182-198.

Bartels, D. M., and Urminsky, O. (2015). To know and to care: How awareness and valuation of the future jointly shape consumer savings and spending. *Journal of Consumer Research*, 41, 1469-1485.

Bryan, C.J., Walton, G.M., Rogers, T., & Dweck, C.S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, *108*, 12653-12656.

Gelman, S., Heyman, G., & Legare, C. (2007). Developmental changes in the coherence of essentialist beliefs about psychological characteristics. *Child Dev*, 78, 757–774.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*, 101–118.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 227-254.

Kim, N. S., & Park, E. Y. (2009). ConceptBuilder: An open-source software tool for measuring, depicting, and quantifying causal models. *Behavior Research Methods*, *41*, 128-136.

Linville, P. (1985). Self-complexity and affect extremity: Don't put all your eggs in one cognitive basket. *Social Cognition, 3*, 94-120.

Locke, J. (1694/1975). An essay concerning human understanding. Oxford: Clarendon Press.

McConnell, A. R. (2011). The Multiple Self-aspects Framework: Self-concept representation and its implications. *Personality and Social Psychology Review, 15,* 3-27.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289.

Nichols, S., & Bruno, M. (2010). Intuitions about personal identity: An empirical study. *Philosophical Psychology, 23,* 293–312.

Orhun, A. Y., & Urminsky, O. (2013). Conditional Projection: How Own Evaluations Influence Beliefs About Others Whose Choices Are Known. *Journal of Marketing Research*, *50*, 111-124.

Pronin, E. (2008). How we see ourselves and how we see others. *Science*, *320*, 1177-1180.

Rehder, B. & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General, 130,* 323-360.

Sloman, S., Love, B., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22,* 189-228.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159-171.

The Effect of Emotion and Induced Arousal on Numerical Processing

Karina Hamamouche (hamamouc@bc.edu)
Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Michelle Hurst (hurstm@bc.edu)
Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Sara Cordes (cordess@bc.edu)
Department of Psychology, 140 Commonwealth Avenue
Chestnut Hill, MA 02467 USA

Abstract

Prominent theories suggest that time and number are represented by a common magnitude system. However, distinct patterns of temporal and numerical processing occur in the presence of emotional stimuli, calling into question theories of a common magnitude system, while also unveiling questions regarding the mechanisms underlying these temporal and numerical biases. We tested whether numerical processing, like temporal processing, may be impacted by increased arousal levels, yet have a higher threshold level in order to impact estimates. If so, then induced arousal may reverse the typical pattern of numerical underestimation in the presence of emotions. Adults (N = 85) participated in either a stress-induction or a control version of the task. Then, participants completed a numerical bisection task in the presence and absence of emotional content. Increasing arousal had no impact on numerical processing, except in the presence of happy faces, providing further evidence for distinct processing mechanisms.

Keywords: quantity processing; numerical cognition; temporal processing; emotion; stress

Introduction

Temporal and numerical processing is vital for our everyday interactions. How many seconds will it take to cross the street? How many slices of pizza are needed to serve a family dinner? These basic quantitative processes are posited to form the foundations of mathematical thought and have been shown to predict math achievement (Halberda, Mazzocco, & Feigenson, 2008), highlighting the importance of understanding basic quantitative processing.

Prominent theories suggest that the processing of time, number, and space are a part of a common magnitude system (Walsh, 2003; Cantlon, Platt, & Brannon, 2009). Evidence in support of this theory demonstrates analogous performance on timing and counting tasks in both rats and humans (Meck & Church, 1983), and comparable parietal cortex activity during temporal and numerical processing (Walsh, 2003). Further, children suffering from genetic disorders that are known to impact numerical processing (e.g., Turner Syndrome) also experience spatial and temporal deficits, suggesting an overlap among systems involved in quantity processing (Silbert, Wolff, & Lilienthal, 1977).

While many controlled laboratory studies have investigated quantitative processing, this work has largely ignored the fact that temporal and numerical processing in the real world rarely occurs in an emotional vacuum. Some work investigating how emotional stimuli impact numerical and temporal processing has led to distinct theories regarding how these quantities are processed (Droit-Volet & Meck, 2007; Young & Cordes, 2013). The bulk of this work has focused on performance during a bisection task in which participants judge whether a target duration or a target numerosity is more similar to a short/small standard or long/large standard. For example, participants may be presented an array of 7 dots and asked to judge whether it is more similar to the learned standards of 4 dots (small) or 16 dots (large). Bisection task data have been used to assess biases in estimates by measuring the value at which participants are indifferent between the two standards (Point of Subjective Equality or PSE; the point at which 50% of responses are long/large).

Prior studies have revealed that both children and adults exhibit patterns of temporal overestimation (i.e., lower PSEs) in the presence of emotional stimuli (Droit-Volet, 2003; Droit-Volet & Meck, 2007), yet identical emotional content leads to underestimation (i.e., higher PSEs) of numerical values (Baker, Rodzon, & Jordan, 2013; Young & Cordes, 2013). These distinct patterns of temporal and numerical processing in the presence of emotion present a challenge to a common magnitude theory while also posing new questions about the specific mechanisms (i.e., attention, arousal) underlying the processing of different quantities. The current study seeks to explore these mechanisms, specifically arousal, on numerical processing in the context of emotional stimuli.

Temporal overestimation in the presence of emotion has been linked to increased arousal (e.g., Angrilli, Cherubini, Pavese, Manfredini, 1997; Droit-Volet & Meck, 2007; Droit-Volet & Wearden, 2001; Gil & Droit-Volet, 2012;

Ortega & Lopez, 2008). Evidence for this is derived from the fact that temporal dilation under emotion tracks with arousal ratings of emotional stimuli, with the most arousing emotional stimuli resulting in the greatest degree of overestimation (Young & Cordes, 2013). Moreover, increased arousal in neutral tasks has been linked to temporal overestimation. For example, the rapid presentation of a stream of stimuli, thought to elicit increased arousal, has been found to result in the overestimation of the presentation of a simultaneous target during timing tasks (e.g., Ortega & Lopez, 2008). Relatedly, filled intervals (i.e., the consistent presentation of a stimulus throughout the duration to be timed) are estimated to last longer than empty intervals (i.e., intervals demarcated by two distinct stimuli), likely due to increased arousal (see Wearden, Norton, Martin, & Montford-Bebb, 2007). Thus, because increased arousal has been linked to temporal dilation and because temporal dilation under emotional circumstances is greatest in response to the most emotionally arousing stimuli, it is posited that arousal is the source of the temporal overestimation observed under emotional circumstances.

On the other hand, arousal is not thought to be the source of the observed numerical underestimation. Instead, because numerical underestimation is found in the context of any emotional stimulus, Young & Cordes (2013) proposed that heightened attentional focusing drives numerical underestimation in the context of emotion. According to this view, the social salience of emotional content serves to heighten attention. Whether this heightened attention to the social stimulus results in numerical underestimation via simultaneous heightened attention to numerical stimuli (and thus, improved numerical processing), or instead, via distracted attention away from the numerical stimulus (and thus impaired numerical processing), has yet to be determined.

Emotional content results in some degree of arousal during temporal and numerical processing; however, it is unclear why the same emotional stimuli may lead to temporal *over*estimation but numerical *under*estimation. One possibility is that numerical processing has a higher threshold than temporal processing in order for arousal to impact estimates, such that a higher level of arousal is necessary to influence numerical estimates. If so, an overall heightened level of arousal, as created by a stress induction task (e.g., Kirschbaum, Pirke, & Hellhammer, 1993), may elicit a pattern of numerical *over*estimation mirroring that of temporal tasks. Relatedly, arousal levels may be slower to rise in response to arousing stimuli, such that it takes time for emotional stimuli to result in threshold levels of arousal. Given that temporal processing is sequential (occurs over a period of time), whereas numerical processing of arrays (as in previous studies) is simultaneous (requires the simultaneous apprehension of several items within a brief presentation period), arousing stimuli may only appear to impact temporal processing because temporal stimuli require more time to process, giving time for arousal levels to rise. If so, then heightening arousal levels via stress induction prior to the numerical task, thus providing time for arousal levels to rise prior to the start of the numerical task may allow for a true assessment of the impact of arousal on numerical processing.

In the current study, we investigated the effects of induced arousal on a numerical task in the presence and absence of emotional stimuli. Arousal was manipulated prior to the numerical task, therefore eliminating concerns regarding the speed of rising arousal levels. Moreover, to our knowledge, no work has investigated the impact of induced arousal on basic numerical processing, making this investigation worthwhile in its own right. While a plethora of evidence suggests that stress in the form of math anxiety can be detrimental for performance on symbolic math tasks (e.g., Meece, Wigfield, Eccles, 1990), it is unclear what role arousal plays in our most primitive sense of number. Lastly, while studies have investigated the impact of emotion on numerical processing, no work has explored factors contributing to individual differences in these emotional biases. That is, are emotions more likely to impact individuals who already have a less precise representation of number? Or, alternatively, is it that all adults, regardless of numerical precision, are similarly vulnerable to the impact of emotional content?

Methods

Participants

Eighty-five undergraduate students (range: 18-25, M_{age} = 19.17, males = 13) participated in this study for course credit. Seven students completed the study, but were excluded from analyses for not following the instructions (n=3), below chance performance on the standard values of the bisection task (n=2), producing PSEs that were more than 3 standard deviations below the group average (n=1), or computer error (n=1), leaving a final sample of seventy-eight. Participants were randomly assigned to two conditions with a final distribution of Stress Induction (N=40, M_{age}=19.02, males = 7) and Control (N=38, M_{age}= 19.00, males = 5).

Stimuli

Stress task A modified version of the Trier Social Stress Test (TSST) was used (Kirschbaum et al., 1993).

Bisection task A bisection task (similar to the task used by Droit-Volet et al., 2004; Young & Cordes, 2013) was implemented. Numerical stimuli were composed of arrays of dots. The standard small value had a magnitude of 4; the standard large value had a magnitude of 16. Intermediate values corresponded to the magnitudes 5, 6, 8, 10, and 13. The numerical arrays consisted of black dots displayed on a white background. Surface area of each individual item was identical in half of the trials (M = 1147.14). In the other half of trials, cumulative surface area was held constant across trials regardless of set size (M = 1385.86). The bisection task was computerized; stimulus presentation was controlled and responses were recorded by a REALBasic program.

Emotional Faces The face stimuli were identical to those used by Young & Cordes (2013). The face stimuli (happy, neutral, and angry) were selected from the NimStim set (Tottenham et al., 2009). The set of faces were standardized on intensity, attractiveness, arousal, and valence so that each image was a clear representation of a specific emotion.

Procedure

Participants were randomly assigned to the Stress or Control condition. Those in the Stress condition sat next to an experimenter who asked them to count backwards by 13 beginning with the number 10,099 (e.g., 10,099, 10,086, 10,073…). Participants were told their performance would be assessed and if a mistake was made, the experimenter would stop them and make them start over from 10,099. Answers were given orally so that the experimenter could keep track of performance. In the Control condition, the experimenter gave similar instructions, except participants were allowed to use a pencil and paper to write their answers down, were told that their performance would not be assessed, and the experimenter left the room while the participant completed the task. After five minutes, participants were stopped and began the bisection task.

Next, participants were seated in front of the computer and familiarized with arrays of dots that were labeled as the "Small" (4) and "Large" (16) standards. Participants were given 12 practice trials in which, they were presented with dot arrays representing either standard value and were asked to indicate whether it was small or large. Dot arrays were presented for 500 ms. Adults selected their response by pressing either [a] for smaller or ['] for larger on the keyboard. Feedback was provided for practice trials only. Following practice, adults participated in 42 baseline test trials that were identical to practice except dot arrays of representing the five intermediate sizes were intermixed among standard trials (6 trials per 7 set sizes). Participants were asked to indicate whether each array was "more similar to the small or large standard". Following baseline, participants had emotion test trials that were identical to the baseline trials expect that a face appeared for 750 ms prior to each dot array presentation. Faces depicted either happy, angry, or neutral emotions. During the emotion test trials, all set sizes (2 standard values and 5 intermediate values) were presented 18 times each, 6 times per emotion, for a total of 126 trials. Emotion trials were presented in a random order, with trials involving the three emotions intermixed throughout. Adults received feedback on the practice trials, but did not receive feedback on any of the test trials.

Data Analyses

Following past work (Droit-Volet et al., 2004; Young & Cordes, 2013), each participant's PSE was calculated for the baseline trials and each of the three emotions separately as a measure of accuracy. PSEs were computed by determining the equation of the line relating the proportion of trials the participant indicated the array was more similar to the large standard and the size of the dot array, and using that to compute the set size corresponding to a 0.5 proportion of large responses. Moreover, each participant's difference limen (DL; the value half way between the set sizes corresponding to a 75% probability of a large response and a 25% probability of a large response) was also calculated as a measure of precision in responding, with higher DLs corresponding to lower precision in responding.

Results

Baseline Performance (no faces)

First, we analyzed performance on the baseline trials to determine whether increased arousal (due to the stress induction) altered numerical processing in the absence of emotional stimuli. Analyses revealed no differences in the PSE or DL of participants across conditions during the baseline trials (PSE: $t(76) = .610$ $p = .544$; DL: $t(76) = -.283$ $p = .778$) revealing that heightened arousal did not result in biases or altered precision in numerical judgments.

Emotion Test Trials - PSE

We conducted a 2 (Condition: Stress, No Stress) x 3 (Emotion: Neutral, Happy, Angry) repeated measures ANOVA on PSEs obtained from the emotion test trials. Analyses revealed a main effect of emotion, $F(2, 152) = 9.767$, $p < .001$ and a significant condition x emotion interaction on PSE, $F(2, 152) = 3.830$, $p < .024$. There was no main effect of condition, $F(1, 76) = .120$, $p = .730$. In order to investigate the interaction further, we looked at the differences across emotions in each of the two conditions separately.

Control Condition In line with prior findings (Young & Cordes, 2013), there was a main effect of emotion on PSE in the Control condition, $F(2, 74) = 9.039$, $p < .001$. The PSEs corresponding to trials involving happy faces ($M = 8.433$, $SD = 0.94$) and angry faces ($M = 8.332$, $SD = 0.90$) were both significantly higher relative to neutral faces ($M = 8.052$, $SD = 1.10$, p's $< .01$), consistent with a pattern of underestimation under emotional circumstances. Performance did not differ between angry and happy faces ($p > .23$). This finding replicates prior research revealing numerical underestimation in the context of both angry and happy faces relative to neutral faces (Young & Cordes, 2013).

Stress Condition There was also a main effect of emotion on numerical judgments in the Stress condition, $F(2, 78) = 4.542$, $p < .014$, however a different pattern emerged. Angry faces ($M = 8.36$, $SD = 0.84$) were significantly underestimated relative to happy faces ($M = 8.155$, $SD = 0.81$) and neutral faces ($M = 8.09$, $SD = .906$; *p's < .03*). There was no significant difference between happy and neutral faces ($p > .5$).

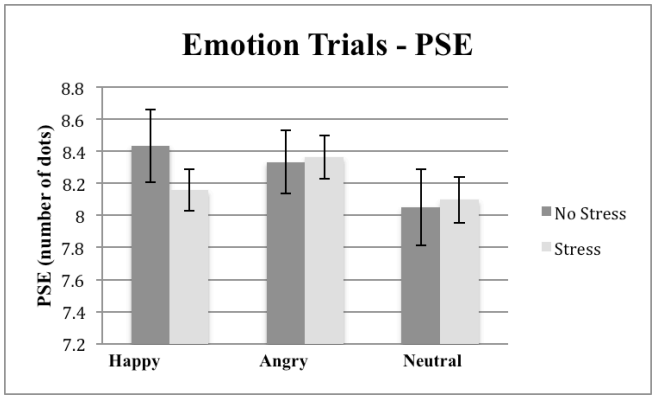


Figure 1. PSE as a function of emotion across conditions.

Emotion Test Trials – DL

We ran a 2 (Condition: Stress, Control) x 3 (Emotion: Neutral, Happy, Angry) repeated measures ANOVA on the DLs obtained on emotion test trials. Analyses revealed a main effect of emotion on DL, $F(2, 152) = 7.765$, $p < .001$, but no main effect of condition $F(1, 76) = .117$, $p = .733$, nor a condition x emotion interaction, $F(2, 152) = 1.722$, $p = .182$. Overall, results revealed increased precision in numerical judgments following presentation of emotional faces ($M_{Happy} = 2.679$, $SD_{Happy} = .265$; $M_{Angry} = 2.647$, $SD_{Angry} = .238$) relative to neutral faces ($M = 2.736$, $SD = .336$, p's $< .04$). There was no significant difference between angry and happy faces ($p = .103$).

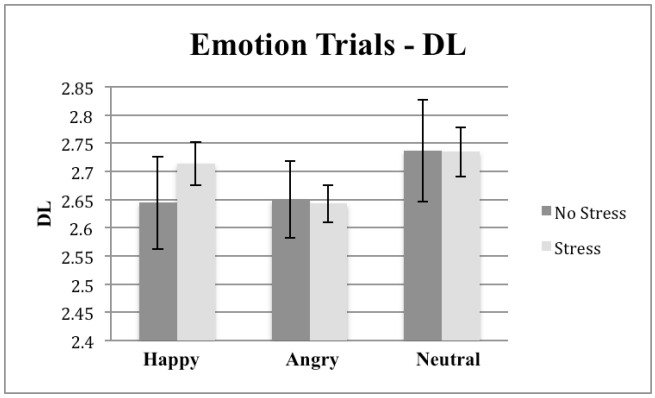


Figure 2. DL as a function of emotion for both conditions.

Individual Differences

Lastly, we explored whether individual differences in precision in the underlying representation of number relates to the magnitude of numerical bias observed in the presence of emotional stimuli. Is it the case that individuals with lower numerical acuity (as assessed via performance on baseline trials) may be more vulnerable to numerical biases in the presence of emotional stimuli? In order to test this question, we computed a measure of numerical bias by subtracting each participant's average PSE across the emotion trials (average of PSE_{Happy} and PSE_{Angry}) and subtracting this from their $PSE_{Neutral}$. Importantly, these analyses only pertain to participants in the Control condition. This measure represents the degree to which participants' underestimated number following emotional stimuli compared to neutral stimuli. Each participant's Baseline DL was used as a measure of numerical acuity in the underlying representation. If those individuals with less precise numerical representations were more prone to numerical biases, we would expect a positive correlation between the participant's Baseline DL and this numerical bias measure. A marginal correlation between $DL_{Baseline}$ and this measure of numerical bias ($r = .298$, $p = .074$) was obtained, suggesting that the magnitude of numerical bias observed under emotional circumstances may be partially predicted by the precision in an individual's underlying representation in number.

Discussion

Prominent theories posit a common magnitude system for representing temporal and numerical magnitudes (Cantlon et al., 2009; Meck & Church, 1983; Walsh, 2003). If this is the case, then temporal and numerical processing should reveal similar biases under identical circumstances. Yet, work reveals that identical emotional stimuli impact temporal and numerical processing in distinct fashions (Young & Cordes, 2013). Given that the pattern of temporal overestimation tracks with the level of arousal of the emotional stimuli (i.e., angry, but not happy, emotional stimuli result in overestimation of durations), it has been posited that increased arousal underlies the observed temporal biases (Droit-Volet & Meck, 2007; Gil & Droit-Volet, 2012). Conversely, given that numerical underestimation is observed in the presence of both arousing (angry) and less arousing (happy) emotional stimuli, it has been hypothesized that changes in attention modulate the numerical underestimation observed (Young & Cordes, 2013). These distinct patterns, however, could potentially be explained by different arousal thresholds for numerical and temporal processing and/or by a delayed arousal response.

In the present study, we examined how induced stress (i.e., increased arousal) impacted subsequent numerical judgments, both in the presence and absence of emotional stimuli. By heightening arousal prior to participation in the numerical task, we were able to assure altered arousal levels in participants at the time of numerical judgments. Moreover, by presenting arousing emotional stimuli (angry faces) following a stress induction task, we maximized arousal in participants prior to engagement in the numerical task to increase the likelihood that any arousal threshold was met. As such, results of this study can speak to how increased arousal impacts numerical processing, providing a true test of the common magnitude system. That is, by maximizing arousal in participants prior to participation in a numerical task, we are able to explore whether numerical biases mimic those found in temporal processing. Yet results did not reveal this to be the case.

Results of our baseline trials revealed that heightened arousal (as induced in our stress task) did not disrupt basic

numerical processing as we had predicted. In fact, results revealed no significant differences between the Stress and Control conditions for baseline responding, suggesting that induced stress did not alter numerical processing in the absence of emotional stimuli.

Importantly, responding during emotional trials in the Control condition mimicked that of previous studies, again revealing a pattern of underestimation following the presentation of either happy or angry emotional stimuli relative to neutral stimuli. Thus, again, in the absence of heightened arousal, participants underestimate number in the presence of emotional stimuli.

However, a different pattern of underestimation following emotional stimuli emerged among those in the Stress condition. Those in the Stress condition only underestimated angry faces compared to neutral faces, with responding on happy trials mirroring that found on neutral trials. These findings are significant for two reasons. First, despite no differences in numerical processing in baseline, findings of a distinct pattern of results across condition make it clear that our stress induction task was effective. Results revealed that induced stress led to a differential pattern of numerical judgments in participants across conditions. Thus, a failure to find baseline differences suggests that induced arousal does not have a strong impact on numerical processing in the absence of emotional stimuli. Second, despite heightened arousal, participants in the Stress condition did not respond with a pattern of overestimation of numerical stimuli, as might have been predicted by a common magnitude system. That is, induced arousal prior to presentation of the numerical task failed to result in numerical biases mimicking those found in temporal tasks, providing strong support against a common magnitude system. Instead, results in the Stress condition revealed a similar (though not identical) pattern of underestimation in the presence of emotional stimuli. However, importantly, in contrast to the Control condition, underestimation was only observed following the presentation of angry faces, not happy ones.

What can explain this novel pattern of results? Findings are likely accounted for by a difference in how participants in the Stress condition may have perceived the emotions presented. It is possible that induced stress may have resulted in an overall negative interpretation of the face stimuli, causing participants to perceive angry faces as angrier, yet happy faces were perceived as more neutral, thus resulting in underestimation of angry trials, yet similar performance on happy and neutral trials. This finding is in line with work demonstrating that highly anxious individuals may be more likely to attend to threatening content, which may be at the expense of attending to the happy stimuli (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, van IJzendoorn, 2007). Other work has shown that stress increases one's sensitivity to threats, but *reduces* specificity (van Marle, Hermans, Qin, & Fernandez, 2009), thus leading to a failure to differentiate happy from neutral faces. Thus, participants in the Stress condition likely perceived the happy faces as being more similar to the neutral faces, thus minimizing any differences observed between the happy and neutral face trials.

Results are also the first to suggest that numerical underestimation biases in the presence of emotional stimuli may reflect heightened numerical acuity. While it has been hypothesized that the numerical biases observed under emotional circumstances may be the results of impaired numerical processing (Rodzon, Baker, & Jordan, 2011; Young & Cordes, 2013), our findings provide evidence for enhanced numerical performance following emotions. Our data reveal significantly smaller DL (i.e., greater precision) after emotional content is presented. This finding is consistent with the literature investigating children's and adult's numerical judgments in the presence of emotional faces (Lewis, Zax, & Cordes, submitted). Prior work has shown that children's numerical judgments become more precise following the presentation of emotional faces (relative to following neutral faces). Prior work has also shown positive impacts of emotions on subsequent tasks. For example, Phelps, Ling, & Carrasco (2006) reported increased contrast sensitivity, and thus enhanced performance on neutral tasks following emotions. Vuilleumier (2005) found similar effects of enhanced performance when emotions preceded and were not concurrent with the task demands. This study joins others suggesting that emotional or threatening content may benefit numerical processing if the emotional content is not inherent in the stimuli to be enumerated (Hamamouche et al., submitted). The adults in our study were also more precise in their numerical judgments following the presentation of emotional faces likely due to heightened attention brought on by the socially salient stimuli, providing further support for the attention model associated with numerical processing.

While many studies have investigated the impact of emotional faces on temporal and numerical processing (e.g., Baker et al., 2013; Droit-Volet, 2003; Droit-Volet & Meck, 2007; Young & Cordes, 2013), it has been unclear if emotions impact an individual's numerical judgments to the same degree, regardless of one's initial numerical representations. We investigated whether those with less precise numerical representations would be more likely to adjust their numerical judgments in the presence of emotion than those with more precise representations. Our data hinted at a possible relationship between participant's baseline precision on the numerical task and the degree to which they underestimated during the emotion trials, indicating that the underestimation effect seen in the presence of emotion may be related to individual differences in numerical representation. However, this correlation was only marginally significant and thus we cannot make strong claims about the potential of this relationship. Future work should explore whether the magnitude of these numerical biases may also be predicted by individual sensitivities to emotional stimuli and/or social stimuli, more generally.

This study is the first of its kind to investigate the interaction of stress and emotion on numerical processing. Although arousal did not impact numerical processing as predicted, this study joins a growing body of literature providing evidence questioning the plausibility of a common magnitude system (e.g., Agrillo, Ranpura, & Butterworth, 2011; Baker, Rodzon, & Jordan, 2013; Young & Cordes, 2013). Future work should investigate the interaction between stress and emotion on temporal processing, in order to further understand basic quantitative processing in real-world situations.

Acknowledgments

Support for the preparation of the manuscript was provided by the Natural Sciences and Engineering Research Council of Canada to MH.

References

Agrillo, C., Ranpura, A., & Butterworth, B. (2011). Time and numerosity estimations are independent: Behavioral evidence for two different systems using a conflict paradigm. *Cognitive Neuroscience, 1*(2), 96-101.

Angrilli, A., Cherubini, P., Pavese, A., & Manfredini, S. (1997). The influence of affective factors on time perception. *Perception & Psychophysics, 59*, 972-982.

Rodzon, K., Baker, J., & Jordan, K. (2011, July). The impact of emotion numerical estimation. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (3552-3557). Austin, TX: Cognitive Science Society.

Baker, J., Rodzon, K., & Jordan, K. (2013). The impact of emotion on numerosity estimation. *Frontiers in Psychology, 4,* 521.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. *Psychological bulletin*, *133*(1), 1.

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in cognitive sciences*, *13*(2), 83-91.

Droit-Volet, S. (2003) Alerting attention and time perception in children. *Journal of Experimental Child Psychology, 85*(4), 372-394.

Droit-Volet, S., Brunot, S., & Niedenthal, P. (2004). Perception of the duration of emotional events. *Cognition & Emotion, 18*, 849-858.

Droit-Volet S., & Meck, W. (2007). How emotions colour our time perception. *Trends in Cognitive Sciences, 1*(12), 504-513.

Droit-Volet, S. & Wearden, J. (2001). Temporal bisection in children. *Journal of Experimental Child Psychology, 55,* 142-159.

Gil, S. & Droit-Volet, S. (2012). Emotion time distortions: The fundamental role of arousal. *Cognition & Emotion, 26*(5), 847-862.

Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*, 665-668.

Hamamouche, K., Niemi, L., Cordes, S., (submitted). Quantifying a Threat: Evidence of a Numeric Processing Bias.

Kirschbaum, C., Pirke, K., Hellhammer, D. (1993). The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28*(1-2), 76-81.

Lewis, E., Zax, A., & Cordes, S. (submitted). The Impact of Emotion on Numerical Estimations: A Developmental Perspective.

Meck, W. & Church, R. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology Animal Behavior Processes, 9*(3), 320-34.

Meece, J., Wigfield, A., Eccles, J. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment and performance in mathematics. *Journal of Educational Psychology, 82*(1), 60-70.

Ortega, L., & Lopez, F. (2008). Effects of visual flicker on subjective time in a temporal bisection task. *Behavioural Processes, 78*(3), 380-386.

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science*, *17*(4), 292-299.

Silbert, A., Wolff, P., Lilienthal, J. (1977). Spatial and Temporal Processing in Patients with Turner's Syndrome. *Behavior Genetics, 7*(1), 11-21.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., ... & Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242-249.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock." *Psychological Monographs: General and Applied*, 77(13), 1-31.

van Marle, H. J., Hermans, E. J., Qin, S., & Fernández, G. (2009). From specificity to sensitivity: how acute stress affects amygdala processing of biologically salient stimuli. *Biological Psychiatry*, *66*(7), 649-655.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences, 9*, 585-594.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space, and quantity. *Trends in Cognitive Sciences, 7*(11), 483-488.

Wearden, J. H., Norton, R., Martin, S. and Montford-Bebb, O. 2007. Internal clock processes and the filled-duration illusion. *Journal of Experimental Psychology, 33*, 716–729.

Young, L. & Cordes, S. (2013). Fewer Things, Lasting Longer: The Effects of Emotion on Quantity Judgments. *Psychological Science, 24*(6), 1057-1059.

Putting the "th" in Tenths: The Role of Labeling Decimals in Revealing Place Value Structure

Abbey M. Loehr (Abbey.Loehr@Vanderbilt.edu)
Department of Psychology and Human Development, 230 Appleton Pl #552
Nashville, TN 37206 USA

Bethany Rittle-Johnson (Bethany.Rittle-Johnson@Vanderbilt.edu)
Department of Psychology and Human Development, 230 Appleton Pl #552
Nashville, TN 37206 USA

Abstract

Language is a powerful cognitive tool. For example, labeling objects or features of problems can support categorization and relational thinking. Less is known about their role in making inferences about the structure of mathematics problems. We test the impact of labeling decimals such as 0.25 using formal place value labels ("two tenths and five hundredths") compared to informal labels ("point two five") or no labels on children's problem-solving performance. Third- and fourth-graders (N = 104) were randomly assigned to one of three conditions (formal labels, informal labels, or no labels) and labeled decimals while playing a magnitude comparison game and number line estimation task. Formal labels facilitated performance on comparison problems that required understanding the role of zero, which highlighted place value structure. However, formal labels hindered performance when explicit understanding of place value magnitudes was required. Findings highlight how the language teachers and students use can impact problem-solving success.

Keywords: mathematics; problem solving; labels; decimals

Introduction

Previous research suggests language may play a critical role in learning and understanding across a variety of domains (Fyfe, McNeil, & Rittle-Johnson, 2015; Miura, Okamoto, Vlahovic-Stetic, Kim, & Han, 1999; Paik & Mix, 2003). Labels in particular have been shown to act as a powerful cognitive tool, recruiting processes that support categorization and relational thinking. For example, providing shared labels encourages children to treat objects similarly and categorize (e.g., Gelman & Markman, 1986; Graham, Kilbreath, & Welder, 2004). Further, children attribute characteristics of ambiguous objects based on their categorical label, rather than relying on perceptual features of the objects (Gelman & Markman, 1986). In addition to supporting categorization, providing shared labels that have a relational meaning enables children to map related sets of objects (Waxman & Gelman, 1986).

Much less is known about the role of shared labels in making inferences about the structure of mathematics problems. Looking for and making use of structure is one of eight mathematical practice standards outlined by the Common Core State Mathematics Standards (2010). For example, mathematically proficient students are able to recognize the relationship between place value location and the value of a digit (e.g., place values decrease from left to right).

Several indirect pieces of evidence suggest that labels may play a role in children's mathematics understanding. First, shared labels facilitated performance on a repeating patterns task (Fyfe et al., 2015). Four- to five-year-olds solved repeating pattern problems and were exposed to either shared, generic labels (e.g., A-B-B-A-B-B) or unshared, specific labels (e.g., blue-red-red-blue-red-red). Children in the formal labels condition solved more pattern problems correctly compared to children in the informal labels condition. The abstract pattern problems required children to make the same kind of pattern as a model pattern by recreating the part that repeats using new materials. The shared labels were generic and arbitrary, which may have helped reveal the structure in the model pattern and generalize it using new materials.

Second, cross-cultural studies suggest differences in number names used in different languages impact mathematics performance. English fraction labels lack information related to the relational magnitudes they represent. In comparison, East Asian languages use verbal names for fractions that explicitly represent part-whole relations. Cross-cultural research compared how Korean, Croatian, and U. S. children performed on a fraction-identification task prior to receiving formal instruction on fractions (Miura et al., 1999). Korean children significantly outperformed Croatian and U.S. children, suggesting differences in fraction labels impacted performance. Additionally, when English-speaking children were provided with fraction names that revealed part-whole relations in a similar way as Korean fraction labels, they outperformed Korean children on a similar fraction-identification task (Paik & Mix, 2003).

Thus, providing children with language that carries meaningful information and can be shared across multiple instances may be one way to support thinking that reveals the mathematical structure of problems. The current study tested how providing different labels for symbolic decimals helps children make inferences about place value structure.

Labels and Decimal Knowledge

How would you say the decimal 0.25? Most adults would name this decimal using informal "point" language (i.e.,

point two five or point twenty five). In contrast, when children learn to name decimals they are taught to use formal place value labels (i.e., twenty five hundredths). Teachers might also use decomposed place value labels by naming each place value separately (i.e., two tenths and five hundredths). The way we describe or label decimals may impact how children make sense of these numbers. The fractional amounts decimals represent are non-intuitive, and as a result, these symbols are often difficult to interpret. In an effort to understand, children often treat decimals like numbers they have lots of experience with – whole numbers (e.g., Stafylidou & Vosniadou, 2004). For example, when children are asked to compare decimal magnitudes they often think 0.25 is greater than 0.9 because of a whole number bias (25 is greater than 9; Resnick et al.,1998).

Labeling decimals using formal, decomposed place value labels might help children understand decimal magnitudes for at least two reasons. First, these labels could help reveal place value structure. Decomposed place value labels assign each digit with an associated value or magnitude, which may encourage children to make place value comparisons by providing a shared place value label. For example, when comparing 0.25 and 0.9, distinct place value labels may encourage children to compare 2 and 9 instead of comparing 25 and 9. Further, providing shared labels promotes relational thinking potentially by revealing the mathematical structure of problems (Fyfe et al., 2015). Second, these labels may help children distinguish decimals from whole numbers by reducing a whole number bias.

There are also compelling reasons to predict that informal point labels will aid or harm thinking. In comparison to formal place value labels, informal "point" labels that reflect familiar language adults use may activate partial understanding of decimal magnitudes children acquire during everyday experiences. Children are exposed to these labels for decimals in everyday environments in which we often label decimal amounts, such as reading thermometers and discussing weight. Mix et al. (2014) found that children as young as 3 years showed surprising understandings of multidigit place values on simple tasks focusing on mappings between spoken number names to written numerals, dots, or block representations. The authors argued that these partial understandings were likely acquired through statistical learning processes that occur in everyday environments rich with multidigit numerals and verbal number names. If children develop these partial understandings in a similar way with decimal magnitudes, labeling decimals using informal, familiar labels could activate this knowledge.

However, using informal labels may harm thinking by activating whole number misconceptions. Labeling digits using only their number names may encourage children to treat decimals like whole numbers. Activating misconceptions has been shown to hinder problem-solving performance (McNeil & Alibali, 2005), in part because children perseverate on using incorrect strategies (Fyfe, Rittle-Johnson & DeCaro, 2012). Thus, informal labels may encourage a whole number bias that interferes with children's problem-solving success.

Current Study

We examined the influence of naming decimals using formal, decomposed place value labels compared to informal, everyday labels or no labels on children's decimal magnitude problem-solving performance. Decimal magnitude knowledge was examined using two main performance measures (i.e., magnitude comparison and number line estimation) and several follow-up transfer tasks. When comparing symbolic decimals, children's success rates and the types of errors they make vary depending on features of the symbolic decimals they are comparing (e.g., Desmet, Grégoire, & Mussolin, 2010; Durkin & Rittle-Johnson, 2015; Resnick et al., 1998). For example, children are influenced by the length or number of digits (using whole number logic, assume decimals with more digits are greater than decimals with fewer digits) and the value of the digits (decimals that include zeros as placeholders are misunderstood).

We hypothesized that formal labels would facilitate performance on magnitude comparison problems that highlight place value structure by including zeros as placeholders and problems that require ignoring a whole number response. Additionally, we predicted that the advantages of informal labels would be counteracted by the activation of whole number misconceptions that interfere with problem solving. Therefore, we predicted children in the informal labels condition would perform similarly to the no labels condition.

Number line estimation was included as a more general measure of symbolic mapping knowledge for decimals. We hypothesized the same pattern of results for this task as the magnitude comparison task but expected effects to be weaker given that symbolic mapping knowledge requires additional knowledge of the specific quantity represented by a given written numeral that may draw more on children's prior knowledge.

Method

Participants

Participants were 121 third- and fourth-grade children. A pretest was given to identify children who did not already demonstrate a high level of decimal magnitude knowledge and thus would perform near ceiling regardless of label condition. Thirteen were excluded from participation because they scored above 75% on the pretest measure. Four additional children were excluded from analysis because they had diagnosed learning disabilities. The final sample included 104 children (*M* age = 9 yrs, 7 mos; 56% female; 26% ethnic minorities).

Design and Procedure

Children completed a brief pretest in their classrooms and participated in a single individual session lasting

approximately 40 minutes. Children were randomly assigned to one of three conditions: formal labels (n = 56), informal labels (n = 55), or control (n = 56). The only difference between conditions was the labels the experimenter and children used to name decimals during the decimal comparison game and number line task. The remaining transfer tasks were administered without using any labels. All tasks were presented on a laptop computer with the exception of the decimal comparison card game. The experimenter read aloud each question and recorded the child's verbal response.

Materials

Pretest An abbreviated version of a validated assessment measured children's decimal magnitude knowledge (Durkin & Rittle-Johnson, 2015). Sample items included comparing decimal magnitudes (e.g., circle the decimal that is greater), identifying decimals worth the same amount (e.g., 0.5 and 0.50), writing a decimal that comes between two decimals (e.g., 0.4 and 0.5), and locating decimals on number lines.

Decimal Comparison Game The decimal labels manipulation occurred while children played a decimal magnitude comparison game (e.g., which decimal is greater?). The game had the same rules as the card game War. Children played the game with the experimenter using a deck of decimal cards, and the player with the greater decimal won each round. Children read aloud the decimal labels printed on the cards before choosing the greater decimal. The printed labels were removed halfway through game play to give children an opportunity to practice generating the decimal labels on their own with feedback from the experimenter. During game play, children compared the magnitudes of 40 pairs of decimals. Pairs were designed to reveal different levels of understanding based on previous research that has identified common errors children make when comparing decimal magnitudes (Desmet, Grégoire, & Mussolin, 2010; Durkin & Rittle-Johnson, 2015; Resnick et al., 1998). Comparisons fell into three different categories. On *benchmark comparisons* (n = 5), children compared a decimal to a familiar 0 or 1 benchmark. The second comparison type included *congruent and incongruent* pairs (n = 17). Congruent comparisons can be solved correctly by comparing decimals as whole numbers (e.g., 0.68 and 0.2; n = 7), whereas incongruent comparisons cannot be solved correctly using whole number rules (e.g., 0.51 and 0.8; n = 10). Finally, *role of zero* comparisons included decimals with a zero in either the tenths or hundredths place. Children often apply rules for the role of zero in whole numbers to decimals. For example, children ignore a leading zero (e.g., 0.04 is the same amount as 0.4) and think a trailing zero increases a decimal's magnitude (e.g., 0.40 is greater than 0.4). Eleven of these pairs had identical non-zero digits (e.g., 0.40 and 0.4 or 0.09 and 0.9). The remaining 7 pairs had different non-zero digits and a zero in the tenths place only (e.g., 0.07 and 0.1 or 0.8 and 0.02), so competing strategies of either comparing the digit values or comparing the length of the decimals could be used.

Decimal Number Line Estimation To measure magnitude knowledge, a 0-1 decimal number line task was created (18 trials; adapted from Siegler, Thompson, & Schneider, 2011). Children were instructed to name each decimal according to their assigned label condition before placing the decimal on the number line. The decimals were taken from previous work and included decimals with one or two digits (Rittle-Johnson, Siegler, & Alibali, 2001). Percent absolute errors (PAE) were calculated reflecting the absolute difference between the student's estimate and the correct location. Lower PAEs indicate more accurate estimates. Each child received an average PAE score across all 18 trials. Because children are often influenced by the number of digits a decimal has, we calculated an average PAE score for hundredths trials (e.g., 0.46; n = 5) and for tenths trials (e.g., 0.2; n = 5). An average PAE score was also calculated for the remaining 8 trials that included decimals with a zero in the tenths or hundredths place (e.g., 0.40 and 0.09) because children often experience confusion about the role of zero.

Decimal Comparison Transfer Decimal comparison transfer items (n = 10; adapted from Durkin & Rittle-Johnson, 2015; Rittle-Johnson, Siegler, & Alibali, 2001) included problems with decimals that included digits in either the thousandths or ones places, which children were not exposed to during the comparison game. Half of the problems were comparisons involving the role of zero (e.g., 3.3 and 3.300).

Place Value Two items assessed children's place value knowledge as used in Rittle-Johnson et al. (2001). These items were administered after the labels manipulation had occurred to determine if using formal labels helped children understand place value concepts. One item presented the number 413.728 and asked how much the 2 was worth from a list of 5 choices: 0.2, 2 tenths, 2 hundredths, 2 tens, or 2 hundreds. The second item asked how many tenths were in 30 hundredths.

Additional Measures Several additional items (n = 13) were included for exploratory purposes. Due to poor reliability and few condition differences, results are not reported for these measures.

Analysis and Results

To examine children's performance on the primary outcome measures (performance on decimal comparison game, decimal number line estimation accuracy, and decimal comparison transfer), a series of ANCOVAs with condition as a between-subject variable were performed. Specifically, condition was dummy coded with formal labels and informal labels entered into the models, and no labels as the reference group. In all models, children's age, grade, and

their score on the pretest were included as covariates. Preliminary analyses revealed no interactions with age, grade, or pretest scores so these interaction terms were not retained in the final models.

Pretest

On the pretest, children answered a minority of problems correctly (M = 32% correct, SD = 14%). Importantly, there were no differences by condition at pretest, $F < 1$.

Decimal Comparison Game Performance

Across conditions children solved about half of the decimal magnitude comparison problems correctly (see Table 1). There were no significant effects between any of the label type conditions for overall performance, $F's < 2.5$. Children performed above chance (33%), $t(103) = 7.05$, $p < .001$, but condition differences were not reliable. Comparison types were designed with common errors and misunderstandings in mind, and some comparisons could be solved correctly using whole number rules. Thus, we compared performance on the three comparison types.

Table 1: Summary of Performance by Condition

Task	Formal *M (SD)*	Informal *M (SD)*	Control *M (SD)*
Comparison Game Accuracy	.53 (.25)	.44 (.18)	.47 (.20)
Benchmark	.74 (.28)	.71 (.27)	.71 (.28)
Congruent	.91 (.26)	.96 (.18)	.94 (.21)
Incongruent	.31 (.43)	.14 (.31)	.16 (.35)
Role of zero			
Same digits	.43 (.38)*	.20 (.34)	.24 (.39)
Different digits	.45 (.38)*	.53 (.28)	.62 (.21)
Number Line Est. PAE	20 (12)*	18 (6)	16 (4)
Tenths	30 (16)	36 (13)	35 (13)
Hundredths	13 (10)	10 (8)	9 (5)
Role of zero	18 (17)*	11 (11)	8 (5)
Comparison Transfer Accuracy	.48 (.16)*	.39 (.13)	.43 (.12)
Congruent & Incong.	.42 (.19)	.39 (.13)	.38 (.12)
Role of zero	.54 (.18)*	.39 (.18)	.47 (.21)

As expected, children's percent correct on *benchmark comparisons* was high and similar across conditions ($F's <$ 1; see Table 1). Children across all three conditions were also successful on *congruent comparison*s that could be solved correctly using a whole number rule (e.g., 0.62 and 0.2; $F's < 1$; see Table 1). However, for *incongruent comparisons* in which a whole number rule produced an incorrect answer (e.g., 0.51 and 0.8), children's percent correct with formal labels was highest and lower with informal and no labels (see Table 1). These condition differences were not reliable, though. Children in the formal labels condition performed at chance (33%), $t(34) = -.26$, p = .80, but children in the informal labels and no labels conditions performed significantly below chance, $t(33) = -3.56$, $p = .001$ and $t(34) = -2.91$, $p = .006$, respectively.

Children's percent correct on *role of zero comparisons* with identical non-zero digit values (e.g., 0.40 and 0.4) was highest with formal labels and lower with informal and no labels (see Table 1). There was a significant effect of formal labels relative to no labels, $F(1, 98) = 4.94$, $p = .03$, $\eta_p^2 = .05$, and no effect of informal labels relative to no labels, $p = .72$. A follow-up analysis revealed a significant effect of formal labels relative to informal labels, $F(1, 98) = 6.55$, $p = .01$, $\eta_p^2 = .06$. Children's accuracy on role of zero comparisons with different non-zero digits (e.g., 0.07 and 0.1) was highest with no labels, lower with informal labels, and lowest with formal labels (see Table 1). There was a significant, negative effect of formal labels relative to no labels, $F(1, 98) = 7.50$, $p = .01$, $\eta_p^2 = .07$. There was no significant effect of informal labels relative to no labels, $p = .21$, or between the two label types, $p = .15$.

To understand the negative effect of formal labels relative to no labels, we examined performance on problems where the correct answer can be achieved using whole number rules. On 4 of these 7 problems, ignoring a zero and choosing the decimal with the greater digit results in the correct answer (e.g. 0.03 and 0.4). Children's percent correct on these 4 problems was highest in the control condition (M = 91%, SD = 38%), lower in the informal labels condition (M = 79%, SD = 37%), and lowest in the formal labels condition (M = 55%, SD = 42%). There was a significant, negative effect of formal labels relative to no labels, $F(1, 98) = 17.64$, $p < .01$, $\eta_p^2 = .15$. There was no significant effect of informal labels relative to no labels, $p = .20$. A follow-up analysis revealed a significant, negative effect of formal labels relative to informal labels, $F(1, 98) = 8.25$, $p = .01$, $\eta_p^2 = .08$. On the remaining 3 problems, the correct answer could not be obtained by ignoring a zero and using whole number rules (e.g., 0.07 and 0.1). Children's percent correct on these problems was highest in the formal labels condition (M = 31%, SD = 41%), lower in the control condition (M = 24%, SD = 38%), and lowest in the informal labels condition (M = 18%, SD = 34%), but condition differences were not reliable, $F's < 1.64$.

Comparison Game Performance Summary Performance on difficult incongruent and role of zero comparisons revealed some positive but mixed effects of providing formal labels. Formal labels led to higher performance on role of zero comparisons that isolate place value, but only when there was no competing digit value information (i.e., only for problems that had identical non-zero digits). For role of zero comparisons that had different non-zero digits, formal labels led to lower performance, potentially by reducing a whole number bias that led to the correct answer.

Number Line Estimation Accuracy

Overall, children's estimations of decimal locations on the number line were inaccurate (see Table 1). There was an unexpected significant, negative effect of formal labels

relative to no labels, $F(1, 98) = 5.22$, $p = .03$, $\eta_p^2 = .05$. There was no significant effect of informal labels relative to no labels and no significant effect between the two label types, *p's* > .21. To understand the negative effect of formal labels relative to no labels, we examined performance on trials in which children may have been influenced by the number of digits in a decimal and could have experienced confusion about the role of zero.

Children's estimates were least accurate for one-digit decimals that only included *tenths*, suggesting these were the most difficult trials (see Table 1). There were no significant effects between any of the three label type conditions, *p's* > .10. Children's estimates for two-digit decimal *hundredths trials* were more accurate than estimates for tenths trials (see Table 1), but condition differences were not reliable for these trials. Children's PAE for two-digit decimal *role of zero trials* were similar to hundredths trials (see Table 1). There was a significant, negative effect of formal labels relative to no labels, $F(1, 98) = 12.56$, $p < .01$, $\eta_p^2 = .11$. There was no significant effect of informal labels relative to no labels, $p = .22$. A follow-up analysis revealed a significant, negative effect of formal labels relative to informal labels, $F(1, 98) = 5.16$, $p = .03$, $\eta_p^2 = .05$.

In general, formal labels impeded children's ability to accurately estimate the location of decimals on a 0-1 number line. This negative effect of formal labels was strongest for role of zero trials and was not present for the most difficult tenths trials.

Decimal Comparison Transfer

Performance across conditions was low on transfer magnitude comparison problems, although performance was significantly above chance (33%), $t(103) = 7.43$, $p < .001$ (see Table 1). There was no significant effect of either label type relative to no labels, *p's* > .14. A follow-up analysis revealed a significant effect of formal labels relative to informal labels, $F(1, 98) = 5.16$, $p = .03$, $\eta_p^2 = .05$.

We also examined performance on congruent and incongruent comparisons and role of zero comparisons. For congruent and incongruent comparisons, there were no effects between any conditions, *p's* < .31. For role of zero comparisons, there was no significant effect of either label type relative to no labels. A follow-up analysis revealed a significant effect of formal labels relative to informal labels, $F(1, 98) = 9.60$, $p < .01$, $\eta_p^2 = .09$.

Place Value Knowledge

Despite exposure to formal place value labels, only a quarter of the children in the formal labels condition were able to use the learned labels to correctly identify the hundredths place value (26%). A similar percentage of children in the informal labels condition (18%) and the no labels condition (6%) were successful on this item, χ^2 (2, $N = 104$) = 5.18, $p = .08$.

Children were much more successful at determining how many tenths were in 30 hundredths. More children in the formal labels condition answered this item correctly (69%) compared to children in the informal labels condition (38%) and no labels condition (43%), χ^2 (2, $N = 104$) = 7.43, $p = .02$. Thus, formal labels seemed to reveal place value structure, as evidenced by understanding the relationship between tenths and hundredths.

Discussion

While shared labels have been shown to support categorization and relational thinking (e.g., Gelman & Markman, 1986; Waxman & Gelman, 1986), less is known about their role in making inferences about the structure of mathematics problems. Several indirect pieces of evidence suggest that labels play a role in mathematics understanding (e.g., Fyfe et al, 2015; Miura et al., 1999; Paik & Mix, 2003).

We found that naming decimals using formal, decomposed place value labels had mixed effects on decimal magnitude problem solving performance. Children who learned to name decimals using formal labels (e.g., "two tenths and five hundredths") compared to informal labels (e.g., "point two five") or no labels were better able to solve decimal magnitude problems that required understanding the role of zero. In particular, they solved slightly more incongruent magnitude comparison problems correctly (e.g., Which decimal is greater, 0.51 or 0.8?), solved more role of zero comparison problems correctly with decimals that had identical non-zero digits (e.g., Which decimal is greater, 0.4 or 0.40? 0.09 or 0.9?), and were better able to determine the relationship between tenths and hundredths. In part, they may have been less likely to treat decimals as whole numbers compared to children in the informal and no labels conditions.

However, there were unexpected negative effects of formal labels compared to informal and no labels. Their performance was lower on role of zero magnitude comparison problems and number line estimation problems that required explicit knowledge of how much tenths and hundredths are worth. This decrement in performance compared to the informal and/or no labels conditions on some tasks may reflect a transitional phase when children's performance becomes worse before it becomes better (e.g., "U-shaped development"; McNeil, 2007; Namy et al., 2004; Siegler, 2005). As children learn that their way of thinking only sometimes leads to correct solutions, they begin to reject that way of thinking; however, rejecting an old way of thinking and generating new, correct ways of thinking are separate processes that develop over time (Siegler, 2005), so children's performance can get worse before it gets better. Noticing place value structure seems to reflect a kind of transitional knowledge important for developing decimal magnitude knowledge. Indeed, understanding the role of zero as a placeholder was found to reflect an intermediate level of decimal magnitude knowledge (Resnick et al., 2016). Further, correctly naming decimals using place value language is predictive of symbolic-mapping knowledge for decimals two years later (Mazzacco & Delvin, 2008).

Compared to formal place value labels, using informal labels could have activated contextual knowledge children acquire through everyday experiences (e.g., Mix et al., 2014). Unfortunately, the tasks used in the current study may not have been suitable for revealing this type of knowledge. In general, children in the informal labels condition performed similarly to those in the control condition. There was some concern that informal labels might activate whole number misconceptions. Children in the informal labels condition performed somewhat worse than children in the no labels condition on transfer role of zero magnitude comparisons. However, in general informal language did not seem to activate misconceptions more so than no labels.

In conclusion, findings from the current study extend previous research on the role of language in mathematics learning, and more specifically the use of labels to reveal the mathematical structure of problems (Fyfe et al., 2015). Identifying mathematically meaningful labels may be a powerful first step in the process of impacting students' problem-solving behavior and understanding.

Acknowledgments

Overall group comparisons on two outcomes were briefly described at the 2016 American Educational Research Association meeting.

References

Desmet, L., Grégoire, J., & Mussolin, C. (2010). Developmental changes in the comparison of decimal fractions. *Learning and Instruction, 20*(6), 521–532.

Durkin, K., & Rittle-Johnson, B. (2015). Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction, 37,* 21-29.

Fyfe, E. R., McNeil, N. M., & Rittle-Johnson, B. (2015). Easy as ABCABC: Abstract language facilitates performance on a concrete patterning task. *Child Development, 86*(3)*,* 927-935.

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104*(4), 1094–1108.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, *75*(2), 409-427.

McNeil, N. M. (2007). U-shaped development in math: Seven year olds outperform nine year olds on mathematical equivalence problems. *Developmental Psychology, 43*, 687-695.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, *76*(4), 883–899.

Mazzocco, M. M. M., & Devlin, K. T. (2008). Parts and "holes": Gaps in rational number sense among children with vs. without mathematical learning disabilities. *Developmental Science*, *11*(5), 681–691.

Miura, I. T., Okamoto, Y., Vlahovic-Stetic, V., Kim, C. C., & Han, J. H. (1999). Language supports for children's understanding of numerical fractions: Cross-national comparisons. *Journal of Experimental Child Psychology, 74*(4), 356-365.

Mix, K. S., Prather, R. W., Smith, L. B., & Stockton, J. D. (2014). Young children's interpretation of multidigit number names: From emerging competence to mastery. *Child Development, 85*(3), 1306-1319.

Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A U-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development, 5*, 37–57.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Mathematics Standards*. Washington, D.C.: Authors.

Paik, J. H., & Mix, K. S. (2003). U.S. and Korean children's comprehension of fraction names: a reexamination of cross-national differences. *Child Development*, *74*(1), 144–154.

Resnick, I., Barbieri, C., Rinne, L., Hurwitz, A., & Jordan, N. C. (2016, April). The role of decimal understanding in fractions understanding and overall mathematics achievement. Poster presented at the 2016 Applying Cognitive Science Research in the Classroom – A Math Center Conference, Arlington, VA.

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education*, *20*(1), 8–27.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346–362.

Siegler, R. S. (2005). Children's learning. *American Psychologist, 60,* 8, 769-778.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–96.

Stafylidou, S., & Vosniadou, S. (2004). The development of students' understanding of the numerical value of fractions. *Learning and Instruction*, *14*(5), 503–518.

Waxman, S.R., & Gelman, R., 1986. Preschoolers' use of superordinate relations in classification and language. *Cognitive Development, 1*, 139–156.

Implicit measurement of motivated causal attribution

Laura Niemi (lauraniemi@fas.harvard.edu)[1], **Joshua Hartshorne (jharts@mit.edu)**[2,3]
Tobias Gerstenberg (tger@mit.edu)[2] & **Liane Young (liane.young@bc.edu)**[3]
[1]Department of Psychology, Harvard University, Cambridge, MA 02138
[2]Massachusetts Institute of Technology, Cambridge, MA 02139
[3]Department of Psychology, Boston College, Chestnut Hill, MA 02467

Abstract

Moral judgment often involves pinning causation for harm to a particular person. Since it reveals "who one sides with", expression of moral judgment can be a costly social act that people may be motivated to conceal. Here, we demonstrate that a simple, well-studied psycholinguistic task (implicit causality) can be leveraged as a novel implicit measure of morally relevant causal attributions. Participants decided whether to continue sentences like *"Amy killed Bob because..."* with either the pronoun *he* or *she*. We found that (1) implicit causality selections predicted explicit causal judgments, (2) selecting the object (victim) for harm/force events (e.g., *kill, rape*) predicted endorsement of moral values previously linked to victim-blame, and (3) higher hostile sexism predicted selecting the female as the cause in male-on-female harm/force. The implicit causality task is a new measure of morally motivated causal attribution that may circumvent social desirability concerns.

Keywords: implicit cognition; causation; psycholinguistics; moral psychology; implicit causality; semantics

Introduction

Blame and condemnation is often placed on the perceived cause of a negative outcome (Malle, Guglielmo, & Munroe, 2014; Cushman, 2008). However, while we often talk loosely about *the* cause of an event (Hilton, 1990), the truth is far more complex. For instance, we might say that Brutus slew Caesar because he despised him. However, Brutus might explain Caesar's death by referring to a different, prior cause: "Caesar was ambitious, so I slew him." In fact, there are many points along the causal chain that links the Big Bang to Caesar's death, any of which could conceivably be referenced in an explanation.

Nonetheless, people tend to focus on certain causes at the expense of others. For example, people typically focus on causes that are more proximal to the event, ignoring those that are more distal (cf. White, 1992; Hilton, McClure, & Sutton, 2010). Both Caesar's birth and his presence in the Forum on the Ides of March were causally necessary for the murder, but there is something unsatisfactory about the explanation that *Brutus slew Caesar because Caesar was born*. Furthermore, when perceiving causal events, humans tend to focus on powers and ignore liabilities (White, 2006, 2007). For example, classic Michottean perception of causation involves asymmetric attention to power over liability: even though the "launched" ball could be construed as *stopping* the other ball, people reliably perceive a "launching" rather than a "stopping" event (Michotte, 1963; Mayrhofer & Waldmann, 2014; White, 2006).

Since causal powers are mapped to "agents" and causal liabilities to "patients" (White, 2006), this suggests that people will construe causality asymmetrically across the agent-patient dyad during social and moral reasoning. In one sense, this seems obvious: powers of the "agent" (e.g., the ability to kill) seem much more likely to be deemed causal than liabilities of the "patient" (i.e., the ability to be killed). Brutus could not have killed an immortal, but Caesar's mortality is unlikely to be considered a candidate explanation (*Brutus killed Caesar because Caesar is mortal*). However, agents are more than their physical powers and liabilities; actions are imbued with social and moral meaning that cannot be boiled down to physical descriptions. While humans may prefer certain kinds of explanations and focus on certain kinds of causes, how the "causal asymmetry" plays out in social-moral domain is far from clear (White, 2006).

People can and do discuss distal causes, patients, and liabilities during evaluation of morally relevant events (Heider & Simmel, 1944). Thus, understanding human blame assignment and condemnation behavior requires rich theories of causation (e.g. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Wolff, 2007). Deepening our understanding of the representation of causation in moral judgment, in turn, is crucial for both our theoretical understanding of moral cognition as well as for public policy and organizational planning.

However, the social importance of moral reasoning makes it difficult to study: subjects may be motivated to give socially desirable answers or conceal their thoughts and attitudes (Alicke, 2000; Banaji & Heiphetz, 2010; Fazio & Olsen, 2003). Prior work shows that attributions often disfavor people representing social groups that participants view in an adversarial manner and favor people representing social groups with whom they are allied (Morgan, Mullen, & Skitka, 2010). Thus, causal attributions can signal people's *specific* personal alliances and hostilities. In public discourse, collaborative settings such as the workplace, as well as adversarial situations, revealing partiality can be socially detrimental.

A key challenge is to find tasks that implicitly measure morally relevant judgments in such a way that circumvents guarding or concealment. There are now multiple well-validated implicit measures that track people's attitudes about and associations between concepts – most notably the Implicit Association Test (IAT) (Nosek, Banaji, & Greenwald, 2002; Banaji & Heiphetz, 2010; Fazio & Olsen, 2003; Nock et al., 2010; Nosek et al., 2002). However, so far, there are no good implicit measures of moral reasoning and blame assignment *per se*. In this study, we show that a well-understood phenomenon from psycholinguistics – implicit causality – presents just such a measure.

Implicit Causality

People have reliable expectations about why some events happened. Most people expect the sentence

1. John frightened Amy because...

to continue with a reference to the sentence subject, John (e.g., *he was carrying a gun*), but

2. John feared Amy because...

to continue with a reference to the sentence object, Amy (e.g., *she had a bad temper*).[1]

Verbs' tendencies to lead people to select the subject or the object are called their "implicit causality biases" on the basis that they reflect automatic intuitions about the (most important) cause of the event in question (Garvey & Caramazza, 1974; Hartshorne & Snedeker, 2012; Rudolph & Forsterling, 1997; VanBerkum, Koornneef, Ottena, & Nieuwland, 2007).There are many subject-biased verbs like *frightened* and object-biased verbs like *feared*, as well as verbs that produce no reliable trend in either direction (Pickering & Majid, 2007; Rudolph & Forsterling, 1997; Hartshorne & Snedeker, 2012; Ferstl, Garnham, & Manouilidou, 2011). The implicit causality task used here involves simply asking participants to decide whether sentences like (1) or (2) continue with the pronoun *he* or *she* (Hartshorne, 2013).

Interest in determining how exactly implicit causality selections map onto causal cognition has produced a large and contentious literature (Bott & Solstad, 2014; Hartshorne & Snedeker, 2012; Hartshorne, 2013; Pickering & Majid, 2007; Brown & Fish, 1983; Rudolph & Forsterling, 1997). This work – which we return to in the Discussion – has largely focused on determining what the differences are between verbs that result in different implicit causality selections (e.g., *frighten* vs. *fear*). Here, we show that *individual variation* in selections provides a window into individuals' (moral) beliefs, values, and reasoning processes.

Explicit and Implicit Judgments

We compare implicit causality bias and explicit judgments about the causal responsibility of the participants in the event, particularly, beliefs indicative of a view that victims of violence "had it coming" and that such events are "victim-precipitated". While a view that a victim "probably had it coming" is likely to be perceived as relatively "safe" to endorse explicitly in an anonymous online survey – which is what we used – these are precisely the sorts of beliefs that people may be motivated to conceal in a non-anonymous context (e.g., the workplace or other setting in which impartiality is the social norm; also, adversarial situations).

Here, we compared implicit causality bias to judgments that the agent (sentential subject) was necessary and sufficient for what happened, and the patient (sentential object) allowed, controlled and deserved the outcome. We expected a preference to select the object ("object-bias", henceforth) across events of harm and force to predict explicit beliefs that agents were less necessary and sufficient, and that patients were more likely to have allowed, controlled and deserved the events. If implicit causality selections predict explicit ratings of agents' and patients' causal capacities – judgments closely tied to moral judgment (Alicke, 1992; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015) – then this provides initial support for the instrument's utility as an implicit measure of morally relevant causal attributions.

Beliefs and Values

We expected that participants would be more likely to demonstrate an object-bias for harm and force verbs when shifting moral responsibility off the person in the subject position and onto the person in the object position would align with their personal beliefs and values. In the current research, we test two case studies of such motivated causal attribution related to (a) moral values, and (b) sexism.

Our predictions related to moral values align with prior work showing that *binding values* – a cluster of highly intercorrelated moral values that censure disloyalty, disobedience to authority, and sexual/spiritual "impurity" (Graham et al., 2011) robustly predict attributions of responsibility and blame to victims (Niemi & Young, in press). If implicit causality biases to objects of events of harm and force are more likely in people higher in binding values, this represents evidence of convergent validity in support for our claim that the implicit causality task taps morally motivated causal attributions.

We also examine the capacity of the implicit causality task to serve as an implicit measure of people's hostility toward a particular social category, using the test case of hostility toward women. To this end, we manipulated the gender of the subject and object in the implicit causality prompts (e.g., "John" *verbed* "Mary"; and vice versa), and measured participants' hostile sexism (Glick & Fiske, 1996). We expected participants who were higher in hostile sexism, which involves antagonistic attitudes toward women, to be more likely to select the object as causal in the implicit causality task only when men were presented as harming women, and not when women were presented as harming men. That is, implicit causality selections should reflect a view of violence against women as more "victim-precipitated" in people high in hostile sexism. Notably, on our account, sexism should only be correlated with object-bias for verbs of harm and force when men are in the subject position and women are in the object position, and not vice versa; whereas binding values should be correlated with object-bias for verbs of harm and force regardless of gender of subject and object.

Method

459 participants were recruited via Amazon Mechanical Turk ($M_{age} = 37.25, SD_{age} = 31.39$; 207 female, 247 male, 5 selected other or missing). 314 additional individuals failed

[1] "Implicit causality" has been used to refer to two different phenomena (Hartshorne, 2013). One involves a tendency for verbs to lead people to refer to the sentence subject or object following the *because* conjunction; this is the phenomenon investigated here. The other involves a tendency for verbs to trigger intuitions about covariation; see Discussion.

attention checks ($n = 189$); didn't complete the study or previously took a related study ($n = 125$), and were excluded from analyses.

Implicit Causality Task

Participants viewed 24 randomized prompts in the format "*[Agent] [verb]*ed *[Patient]* because" and were asked to "Please select which word you think would follow." They were offered the choices "he" or"she" (counterbalanced pronoun order across items).

We varied agent and patient gender between participants, presenting half of the sample with male agents and female patients, and vice versa for the other half. Verbs included 12 conveying harm and force ("harm/force verbs" henceforth); and 12 filler verbs with ranging causal biases (Bott & Solstad, 2014; Hartshorne & Snedeker, 2012). Verbs and the probability of selecting the object ("object-bias") are shown in Table 1.

Measures of Beliefs about Agents and Patients

After completing the implicit causality task, participants were instructed to "Consider a hypothetical event:" and were again presented with the 24 events they had seen during the implicit causality task but this time without the "because" connective (e.g., "George impressed Julie."). Each event was followed by instructions to "Weigh the following possibilities:" and a series of items in the following order:

1. Agent Unnecessary: "Would *[patient]* have been *[verbed]* by someone else?"
2. Agent Sufficient: "Would *[agent] [verb]* someone else?"
3. Patient Control: "Did *[patient]* have control over the occurrence of the event?"
4. Patient Allowing: "Did *[patient]* let the event happen?"
5. Patient Desert: "Could *[patient]* have deserved the event?"

Participants responded using sliding scales anchored at 0 = "Definitely No"; 50 = "Unsure"; 100 = "Definitely Yes". We created variables capturing "Agent Contribution" by averaging the Agent Unnecessary ratings (reverse-coded) and Agent Sufficient ratings (*Cronbach's alpha*=.78); and "Patient Contribution" by averaging the Patient Control, Patient Allowing, and Patient Desert ratings (*Cronbach's alpha*=.79).

Moral Values Questionnaire

Moral values in the five foundations (caring, fairness, ingroup loyalty, authority, and purity) were assessed using the 30-item Moral Foundations Questionnaire (MFQ) (Graham et al., 2011)."Individualizing values" represent the extent of endorsement of caring and fairness values. "Binding values" represent the extent of endorsement of ingroup loyalty, authority and purity values. Participants also provided demographic information including politics, gender and religiosity.

Ambivalent Sexism Inventory

Sexism was assessed with the Ambivalent Sexism Inventory (ASI) (Glick & Fiske, 1996). The ASI measures extent of agreement with 22 statements about women and men in society and allows for the calculation of a hostile sexism score and a benevolent sexism score. Hostile sexism captures antagonistic attitudes toward women and toward women's pursuit of equality, whereas benevolent sexism captures traditional stereotypes about women as requiring protection by men. In order to directly index adversarial attitudes toward women, we used the hostile sexism subscore.

Results

Implicit Causality and Beliefs about Agents and Patients We hypothesized that the implicit causality task provides a window into beliefs that victims of violence "had it coming" – which people may be motivated to conceal in many contexts. Regression analysis with Agent and Patient Contribution entered simultaneously confirmed that object-bias in the implicit causality task was negatively predicted by Agent Contribution ratings ($\beta = -.176, p < .001$) and positively predicted by Patient Contribution ratings ($\beta = .240, p < .001$) for harm/force verbs (Model R^2 =.13).[2] The fact that perceptions of patients as having the capacity to control, allow and deserve events of harm and force traded off with perception of agents as necessary and sufficient in the model ($r = -.483, p < .001$) indicates that the nature of the implicit causality task, which forces a choice between the agent or patient, aligns well with causal intuitions. Patient Contribution ratings were relatively low for events of harm and force[3], which aligns with intuitive jurisprudence: perpetrators of violence and imposition are typically viewed as more causally responsible than victims.

In sum, increased beliefs that patients controlled, allowed and deserved events and decreased beliefs that agents were necessary and sufficient were associated with increased likelihood of selecting people in the object position (patients) rather than people in the subject position (agents) as causal in the implicit causality task for events involving harm and force (see Fig.1a-b). As an illustrative example, ratings of the likelihood that patients *deserved being killed* were higher for participants who continued the sentence "Bob killed Amy because" with the pronoun referring to the patient – *she* ($M = 26.10, SEM = 1.89$) rather than the agent – *he* ($M = 19.62, SEM = 1.79$).

Implicit Causality and Moral Values Prior work has found that greater endorsement of binding values – loyalty, respect for authority and preservation of purity – predicts greater attribution of blame to victims (Niemi & Young, in press). Accordingly, we expected greater endorsement of binding values to correlate with object-bias for harm/force verbs in the implicit causality task. As hypothesized, binding values were correlated with object-bias for harm/force verbs ($r = .294, p < .001$), but not filler verbs ($r = .07, p = .14$). To illustrate, binding values positively predicted the likelihood

[2]Object-bias for filler verbs was positively predicted by Patient Contribution ratings ($\beta = .169, p < .001$), but not Agent Contribution ratings ($p = .54$).

[3]Patient Contribution ratings for harm/force verbs ($M = 31.5, SEM = .77$) were significantly lower than for filler verbs ($M = 55.42, SEM = .62, t(458) = 32.27, p < .001$). See Table 1 for means.

Table 1: Average implicit causality object-bias (OB), Agent Contribution, and Patient Contribution ratings.

Verb	OB	Agent	Patient
Harm/Force			
Clobbered	0.55	62.76	29.26
Coerced	0.30	56.96	41.41
Enslaved	0.36	66.39	22.88
Forced	0.39	61.11	29.65
Influenced	0.31	51.78	58.70
Killed	0.53	67.08	19.66
Manipulated	0.29	57.52	40.69
Raped	0.30	70.15	15.86
Robbed	0.26	66.05	19.92
Stabbed	0.50	66.00	22.47
Strangled	0.54	66.09	22.70
Tempted	0.31	52.75	54.80
Fillers			
Approached	0.44	50.63	50.74
Confused	0.29	51.36	40.53
Congratulated	0.88	46.75	65.67
Delighted	0.30	52.49	62.26
Impressed	0.20	52.90	58.71
Observed	0.62	51.52	41.38
Praised	0.85	49.04	63.94
Punished	0.74	57.34	43.96
Quoted	0.61	51.60	47.38
Skipped	0.60	54.34	37.54
Thanked	0.83	51.45	65.25
Transported	0.73	51.31	67.67

that participants would select the referent to the patient rather than the agent following "*[Agent]* killed *[Patient]* because" (*Odds Ratio (OR)*=1.41, p<.001).

Also as hypothesized, correlations between binding values and object-bias for harm/force verbs were maintained across gender frames, no matter whether male-verbed-female ($r = .399, p < .001$), or female-verbed-male ($r = .201, p = .002$).[4] Individualizing values were uncorrelated with object-bias for harm/force verbs and filler verbs (p's>.46).

To determine the contribution of binding values above and beyond associated beliefs about agents and patients, we entered these variables simultaneously into regression analyses.[5] Analyses revealed that binding values ($\beta = .206, p < .001$), Patient Contribution ($\beta = .195, p < .001$), and Agent Contribution ($\beta = -.151, p = .002$) significantly predicted object-bias for the harm/force verbs (Model R^2=.17).

In sum, increased binding values – in addition to beliefs that patients allowed, controlled and deserved the harm/force events, and that agents were not necessary or sufficient – predicted object-bias for harm/force verbs (see Fig.1c).

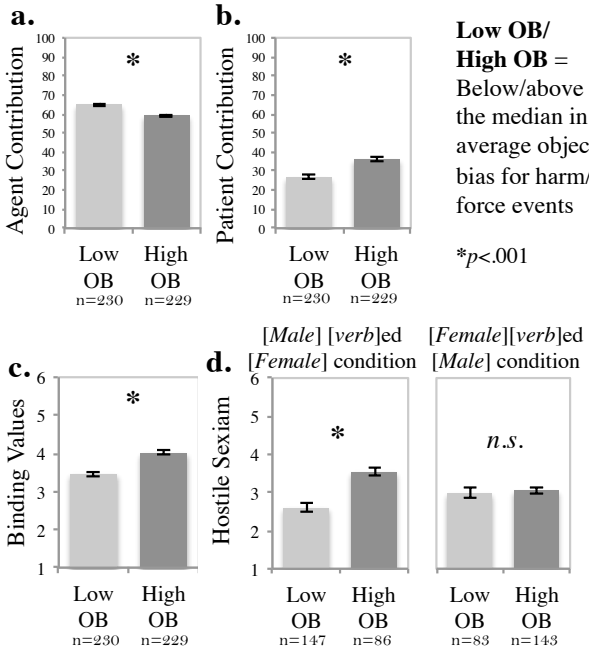


Figure 1: Primary results: Participants who were more likely to select the object as causal in the implicit causality task for harm/force events (e.g., selected "she" not "he" for events such as "Bob killed Amy because ... he or she?") showed (a) reduced ratings of agents' causal contributions, (b) increased ratings of patients' causal contributions, and (c) higher binding values. Participants who were more likely to select the object as causal for harm/force events when agents were male and patients were female showed higher hostile sexism (d, left panel), but not when agents were female and patients male (d, right panel).

Implicit Causality and Sexism Supporting hypotheses, hostile sexism was correlated with object-bias for harm/force verbs when male-verbed-female ($r = .439, p < .001$; Fig.1d, left panel), but *not* when female-verbed-male ($r = .05, p = .45$; Fig.1d, right panel). To illustrate, higher hostile sexism indicated a higher likelihood of selecting "she" following "Bob killed Amy because ... he or she?" ($OR = 1.94, p < .001$), but had no predictive value for selections following "Amy killed Bob because ... he or she?" ($OR = .927, p < .54$). These results indicate that a view of men's violence toward women as "victim-precipitated" in people high in hostility toward women is measurable with the implicit causality task.

Replications

The above study was replicated twice as part of two follow-up studies. We again found significant correlations between binding values and implicit causality object-bias for harm/force events [Replication 1: n=788 ($r = .137, p < .001$); Replication 2: n=249 ($r = .287, p < .001$)] and significant correlations between hostile sexism and implicit causality object-bias for harm/force events when men were sentence subjects and women sentence objects [Replication 1, n=410 ($r = .375, p < .001$); Replication 2, n=121 ($r = .459, p < .001$)], but *not* when women were sentence subjects and men sentence objects ($p's > .84$).

[4]Correlations were significantly different ($Z = 2.32, p = .02$).

[5]Notably, and aligning with prior work showing that binding values predict increased attribution of responsibility and blame to victims and reduced counterfactual focus on perpetrators (Niemi & Young, in press), binding values were positively correlated with Patient Contribution ($r = .276, p < .001$) and negatively correlated with Agent Contribution ($r = -.226, p < .001$) for harm/force verbs.

Discussion

The implicit causality task is a novel measure of motivated causal attribution. Average object-bias in the task (i.e., likelihood of selecting a pronoun referring to the person in the object position to follow *"X verbed Y because"*) positively correlated with beliefs that patients allowed, controlled, and deserved harm/force events. Implicit causality object-bias for harm/force events was also positively correlated with endorsement of binding values, which were previously found to predict victim blaming (Niemi & Young, in press). Finally, implicit causality object-bias for harm/force events when men were sentence subjects and women were sentence objects was positively correlated with hostile sexism. This demonstrates that more specific patterns of hostility toward people in particular social categories (here, women) are traceable with the implicit causality task. We replicated these results twice.

The relationship between implicit causality selections and ratings of agents' and patients' contributions to harm and force indicates that the implicit causality task has predictive validity as an index of people's causal attributions across the agent-patient dyad. This finding has implications for an influential theory of moral cognition – "dyadic morality" – which argues that the *agent-harms-patient* template represents the core structure of immoral events (Gray, Young, & Waytz, 2012; Schein & Gray, 2015; Gray, Schein, & Ward, 2014).

On this account, a universal feature of moral judgment is a tendency for people to justify their moral judgments by identifying harmed victims and asymmetrically loading causality and blame onto harmful perpetrators (Gray et al., 2014; Schein & Gray, 2015). The current results indicate that there is systematic diversity in how people attribute causation across the dyad for events that involve harm and force.

Prior work has argued that implicit causality biases do not reliably predict explicit causal judgments, at the verb-specific level of analysis (Hartshorne, 2013), based on examining the relationship between implicit causality selections and people's ratings of covariation (i.e., ratings of subject as "the kind of person who verbs people" and the object as the "kind of person people verb"; Brown & Fish, 1983). In line with this prior work, we found that participants' ratings of agents' sufficiency and necessity, which involved a counterfactual framing of similar items in our study, did not pattern with implicit causality ratings for *individual verbs*. The capacity of the implicit causality task to tap into people's causal reasoning about agents and patients was revealed when averaging *across* verbs of harm and force. This indicates that the task is particularly well-suited to reveal one's general construal of the nature of dyadic human action involving harm and imposition, namely, how much one views harm as sourced in autonomous agents versus compelled by precipitating patients.

The positive relationship between object-bias in the implicit causality task and binding values aligns with prior findings that binding values predict victim blame via increased attributions of responsibility to victims (Niemi & Young, in press). Why might binding values predict attributions to patients at the implicit and explicit level? Unlike *individualizing values* which promote unconditional care and straightforwardly define agents' harm to patients as *always wrong*, binding values – loyalty, respect for authority, concern for purity – involve moralization of norms which sometimes *require* harm (Graham et al., 2011; Niemi & Young, in press). For example, loyalty to the ingroup may entail harming people belonging to the outgroup; harming or banishing a group member who has brought dishonor on the group may be compatible with moral valuation of purity (e.g., honor killing). Thus, people who highly endorse binding values may deviate from the typical pattern of moral judgment focused on events fitting an *agent-harms-patient* template (Gray et al., 2012; Schein & Gray, 2015). Indeed, aligning with the prior work focused on moral judgment (Niemi & Young, in press), here we found that people higher in binding values were those more likely to show an "inverted" pattern of attribution for events of harm and force whereby patients were selected as causal over agents.

We also found evidence that the implicit causality task tracks more specific motivated attributions. Hostile sexism was correlated with object-bias for events of harm and force when women were in the object position (and men in the subject position), but not when men were in the object position (and women in the subject position). This result indicates that hostility toward a specific social category (here, women) meaningfully predicts implicit attributions of causality across events of harm and force to representatives of that social category in the position of "harmed-patient" – in addition to broader moral commitments (i.e., binding values). Future work will investigate the influence of manipulations of the implied race of the sentence subject and object together with individual differences in racism on implicit causality for events of harm and force. Other work will explore the extent to which viewing agent-patient events as induced by patients serves post hoc justification of moral condemnation or exculpation of agents with whom one feels at risk of being identified (Heider, 1958; Morgan et al., 2010) by experimentally manipulating group membership (e.g., Masten et al., 2010).

Prior work has demonstrated that implicit measures can tap into people's hidden biases (Fazio & Olsen, 2003). The IAT, for example, is a powerful implicit instrument because even if participants infer the experimenter's aim to measure negative attitudes toward a target and attempt to alter their task behavior, they are not likely to be successful. Nevertheless, researchers may be interested in measuring attitudes toward a target more covertly. One advantage of the implicit causality task is that it can be embedded with distractor items, potentially disguising experimenters' aims. And since individual items are answered very quickly, demand characteristics and participant bias can be kept low in this manner without making the task excessively long. Moreover, the ease of administrating the implicit causality task will make it simple to combine with functional magnetic resonance imaging (fMRI), electroencephalography (EEG), and other measures of neural activity in future investigations of the neurobiological mecha-

nisms involved in social judgment and causal attribution. The implicit causality task can also be completed with a pencil and paper, or administered verbally. Thus, it has the capacity to tap attitudes via causal attributions in a broad range of populations, including people unfamiliar with or physically incapable of using technological devices, and people who are not literate, including young children. Cross-linguistic investigations with the implicit causality task are also suitable (cf. Hartshorne, Sudo, & Uruwashi, 2013), which will enable researchers to shed light on cultural differences in motivated causal attribution. Ongoing work is now focused on characterizing the extent of the *implicitness* of the implicit causality task, involving, in part, manipulating task instructions and probing participants' thoughts on the purpose of the task.

Conclusion

We have shown that responses to the implicit causality task predict people's general beliefs about agents' and patients' causal contributions to events involving violence and imposition, as well as their higher-level moral values and social attitudes. The implicit causality task provides a versatile tool for future research into motivated causal attribution.

Acknowledgments The authors gratefully acknowledge the helpful comments of Steven Pinker, Fiery Cushman, Jorie Koster-Hale, Jonathan Phillips, Alek Chakroff, the ECOM Research Group, the Events in Language Reading Group at Harvard, and the Morality Lab at Boston College.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556.

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives in Psychological Science*, *10*(6), 790-812.

Banaji, M. R., & Heiphetz, L. (2010). Handbook of social psychology. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), (p. 348-388). New York: Wiley.

Bott, O., & Solstad, T. (2014). Psycholinguistic approaches to meaning and understanding across languages (studies in theoretical psycholinguistics). In B. Hemforth, B. Mertins, & C. Fabricius-Hansen (Eds.), (p. 213-251). Cham: Springer.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *17*, 237-273.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(353-380).

Fazio, R. H., & Olsen, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology*, *54*, 297-327.

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in english: A corpus of 300 verbs. *Behavior Research Methods*, *43*(1), 124-135.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*(3), 459-464.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Glick, P., & Fiske, S. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491-512.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*, 366–385.

Gray, K., Schein, C., & Ward, A. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600-1615.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101-124.

Hartshorne, J. K. (2013). What is implicit causality? *Language, Cognition and Neuroscience*, *29*(7), 804-824.

Hartshorne, J. K., & Snedeker, J. (2012). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes*, *28*(10), 1474–1508.

Hartshorne, J. K., Sudo, Y., & Uruwashi, M. (2013, feb). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, *60*(3), 179–196. doi: 10.1027/1618-3169/a000187

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65–81.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*(3), 383–400.

Malle, B. F., Guglielmo, S., & Munroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Mayrhofer, R., & Waldmann, M. R. (2014, Sep). Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, *132*(3), 485-490. doi: 10.1016/j.cognition.2014.05.013

Michotte, A. E. (1963). *The perception of causality*. New York: Basic Books.

Morgan, G. S., Mullen, E., & Skitka, L. J. (2010). When values and attributions collide: Liberals and conservatives' values motivate attributions for alleged misdeeds. *Personality and Social Psychology Review*, *36*(9), 1241-1254.

Niemi, L., & Young, L. (in press). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*.

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychological Science*, *21*(4), 511-517.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research and Practice*, *6*(1), 101-115.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*(5), 780-788.

Rudolph, U., & Forsterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin*, *121*(2), 192-218.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147-1163.

VanBerkum, J. J. A., Koornneef, A. W., Ottena, M., & Nieuwland, M. S. (2007). Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, *1146*, 158-171.

White, P. A. (1992). Causal powers, causal questions, and the place of regularity information in causal attribution. *British Journal of Psychology*, *83*(2), 161–188.

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147.

White, P. A. (2007). Impressions of force in visual perception of collision events: A test of the causal asymmetry hypothesis. *Psychonomic Bulletin & Review*, *14*(4), 647-652.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Boredom, Information-Seeking and Exploration

Andra Geana (ageana@princeton.edu)[*+] Robert C. Wilson (bob@email.arizona.edu)[°]
Nathaniel Daw (ndaw@princeton.edu)[*+] Jonathan D. Cohen (jdc@princeton.edu)[*+]

[*]Department of Psychology/[+]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540 USA
[°]Department of Psychology, University of Arizona, Tucson, Arizona 85721 USA

Abstract

Any adaptive organism faces the choice between taking actions with known benefits (exploitation), and sampling new actions to check for other, more valuable opportunities available (exploration). The latter involves information-seeking, a drive so fundamental to learning and long-term reward that it can reasonably be considered, through evolution or development, to have acquired its own value, independent of immediate reward. Similarly, behaviors that fail to yield information may have come to be associated with aversive experiences such as boredom, demotivation, and task disengagement. In accord with these suppositions, we propose that boredom reflects an adaptive signal for managing the exploration-exploitation tradeoff, in the service of optimizing information acquisition and long-term reward. We tested participants in three experiments, manipulating the information content in their immediate task environment, and showed that increased perceptions of boredom arise in environments in which there is little useful information, and that higher boredom correlates with higher exploration. These findings are the first step toward a model formalizing the relationship between exploration, exploitation and boredom.

Keywords: boredom, exploration, information-seeking

Introduction

The complexity and uncertainty of the real world makes choosing between behavioral options challenging. We often have numerous alternatives from which to choose, and we often have incomplete information about many of these. When making a choice, therefore, we must often consider two competing goals: one is earning as much reward as possible, and the other is to gain information about the alternatives, that may improve our choices in the future. This dilemma is known as the exploration-exploitation tradeoff (Cohen, McClure & Yu, 2007) that has been the subject of growing investigation.

Ample evidence has shown that humans and animals engage in information-seeking, even at the cost of current reward (Behrens et al. 2007; Bromberg-Martin & Hikosaka, 2009). Theoretical models show that the information acquired under these scenarios can improve the computation of value estimates, leading to better choices and greater reward over the longer term (Wilson et al. 2014). In this context, it is worth noting that evidence from the literatures on curiosity and creativity suggests that humans and other animals find new information valuable even when it is not possible to immediately use it to acquire better reward (Gottlieb, 2012; Kidd & Hayden 2015), and that this desire for information can lead to the disengagement from activities that are otherwise rewarding. This type of disengagement is a widespread and long-documented phenomenon, and has been considered to be an important factor in boredom (Csikszentmihalyi, 2000; Eastwood et al., 2012). That is, boredom may reflect a bias toward the pursuit of behavior that is not immediately rewarding, but that may provide information useful for increasing long-term reward — in other words, exploration.

Consistent with this proposition, monotonous, repetitive, or insufficiently informative or stimulating tasks increase the perception of boredom (Hill & Perkins 1985; Pattyn et al. 2008), and they are valued and attended to less (Schmidhuber, 1997; Eastwood et al., 2012). Furthermore, observed behavioral correlates of boredom seem to suggest a link to exploration: boredom has been found to prompt the search for new stimulation (Fowler 1967; Meagher & Mason 2012), the increased drive to discover new goals and resources (Eastwood et al. 2012), and the tendency toward innovation and creativity (Bench & Lench 2013).

Complementing these empirical findings, studies in the reinforcement learning and machine learning literatures have suggested that boredom might serve an adaptive function – for instance by signaling an increased opportunity cost of choosing the current option compared to other available options (Kurzban et al. 2013), or by ensuring that too-well-known (i.e. insufficiently informative) options are penalized in value, which can lead to better learning and a higher long-term reward rate (Schmidhuber, 1991; Simsek & Barto, 2006). While these empirical and theoretical lines of work are consistent with the assertion that boredom reflects a signal biasing behavior toward exploration, to date there has been no direct test of this hypothesis. Empirical work has largely been qualitative and observational (Hill & Perkins 1985), while theoretical predictions have not been tested in human participants.

To address this gap in the literature, we conducted three experiments that parametrically manipulated the information content of participants' task environment, examined perceptions of boredom, choices to engage with the task or abandon it in favor of an alternative, and the link between self-reported boredom and overall exploration behavior. Results showed that varying information content elicited correlated changes in boredom, and that people showed increased exploration in response to boring (i.e., less informative) contexts. We conclude by discussing the relationship of our findings to a theoretical model that parallels optimal foraging theory (which focuses on immediate reward) to formalize the value of information (as a proxy for future reward) in explore-exploit decisions.

Experiment 1

This experiment tested for correlations between self-reports of boredom and the amount of useful information that can be gained by continued engagement in the current task. Previous theoretical work has suggested that prediction error (PE) can be used to measure the amount of useful information left to learn in a given task environment (Schmidhuber, 1997). Specifically, it was proposed that asymptotically low prediction errors signal that a good representation of the task has already been learned, and persistently high PEs signal that the task environment is too random and errors cannot be reduced. Both cases were suggested to cause boredom, and the increased drive to disengage and search for more informative alternatives. (Schmidhuber, 1990; Luciw et al., 2013). Accordingly, we used change in prediction error as an index of learning and, correspondingly, participants' estimates of information available in the task environment,

We found that, in line with theoretical work, people's self-reported boredom ratings correlated negatively with change in prediction errors, consistent with the hypothesis that boredom increases as the amount of information that can be gained from the task decreases. When participants played a computer game for which they already had all the information necessary to perform perfectly (so there was nothing left for them to learn), or the outcomes were completely random (so there was no task structure to learn), they reported being more bored than in a task in which they could acquire useful information as they played.

Methods

Participants Twenty-five Princeton University undergraduates (ages 18 to 22) performed the experiment in exchange for course credit.

Task Participants were asked to predict numbers generated by a virtual machine (for a similar design, see Nassar et al.2010). On each trial, they made their predictions by adjusting a vertical slider (the "prediction slider", see fig 1A) between 0 and 100 to indicate the next value that the virtual machine would generate. After they adjusted the slider, they pressed the space key to confirm their prediction, and the machine generated the number for that trial. Games in the task consisted of thirty trials, and changes between games were signaled to the participants; there were twenty-four games in total, with the session lasting approximately one hour.

The critical experimental variable was the difference between the participants' prediction and the actual generated number, that we refer to as the Prediction Error (PE). Participants were rewarded based on these prediction errors: the smaller the error (i.e., the closer their prediction was to the actual number), the more points they received.

The machine generated numbers according to an underlying distribution, which differed between conditions. In the "Gaussian" condition, numbers were generated from a Gaussian distribution with a fixed mean and standard deviation; however, each number was not displayed until after the participant recorded their guess. In the "Certain" condition, numbers were generated from a Gaussian, and displayed on the screen before the participant made their response. In the "Random" condition, numbers were generated uniformly between 0 and 100, but again not displayed until after the participant recorded their guess. Therefore, the underlying distribution of the number-generating machine was such that participants had to either learn the generative process to gradually reduce their prediction error (in the Gaussian condition), they were already told the next number and did not need to learn anything to make perfect predictions (the Certain condition), or the numbers were randomly generated and participants could not reduce their prediction error (in the Random condition). Participants were also asked to self-report their level of boredom on every fourth trial (for a total of ten times throughout each game). They did so by adjusting another slider (the "boredom slider") at the bottom of the screen (figure 1A).

Results

The "Certain" condition elicited the highest boredom ratings in all participants (repeated measures ANOVA, $F(2, 60) = 5.03$, $p = 0.01$). The ratings for the Certain condition were consistently higher than for the Gaussian and Random conditions in both early trials (first six games) and late trials (last six games), as shown in figure 1B, and for the average ratings within a game (1C). The Gaussian condition, by contrast, was consistently rated as the least boring. The Random condition was rated in-between the other two.

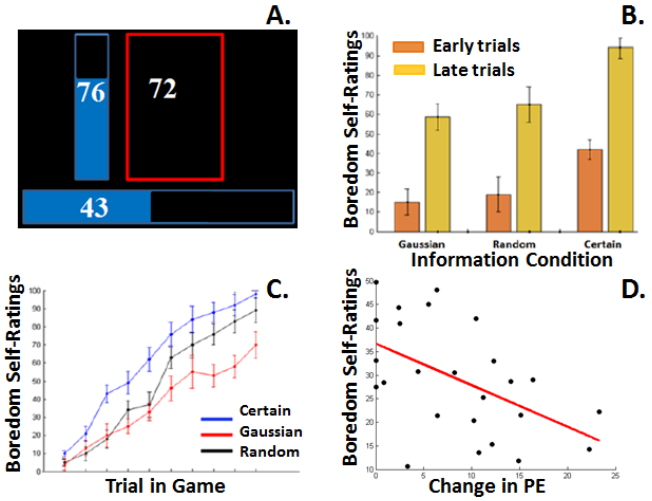


Figure 1: A. **Task d**esign: participants had to predict the next number generated by the virtual machine (the red rectangle). They predicted by adjusting the vertical slider to the left. They rated their boredom using the horizontal slider at the bottom. B, C. Ratings for the Certain condition (no useful information content) were significantly higher both within a game, and for early and late games. D. Average change in prediction error within a game correlated with the average boredom rating for that game.

There was also a significant observed main effect of time on boredom ratings: for all three conditions, later ratings were significantly higher than earlier ratings, both within a game , and across the entire session in early versus late trials (figures 2A and 2B, two-way repeated measures ANOVA, $F(2,18) = 13.39$, $p < 0.01$).

Absolute prediction errors were computed for each game (as the absolute value of the difference in participants' prediction from the number generated on each trial), and the average change in prediction error for each game was computed as the average difference between PEs on consecutive trials. These values were then binned for changes in PE, and the average boredom ratings corresponding to those games were calculated (regardless of which condition those games were in – although, as explained in the methods, the participants were only able to significantly reduce their PE in the Gaussian condition). As shown in figure 1D, there was a significant negative correlation between the change in prediction error and the boredom ratings (R^2=0.2613, $p < 0.01$).

Discussion

This first experiment revealed a correlation between the information content available in a task (operationalized as change in prediction error), and the subjective perception of boredom. The results suggest that the amount of information that can be learned from a task is linked to how boring the task is perceived: the Certain condition – i.e., the one in which there was no useful information to be learned by performing the task, because all the information was already given to participants – elicited the highest boredom ratings (fig. 1B). Conversely, the Gaussian condition, in which it was possible to improve predictions by learning the underlying number-generating distribution, was rated as the least boring. This is consistent with previous theories on 'too much or too little information' causing suboptimal levels of arousal (Schmidhuber, 1997; de Rijk, Schreurs & Bensing 1999), as well as with the notion of "desirable difficulty" – i.e., the notion that there is a certain amount of effortful information-processing that helps learning and is perceived as desirable (Bjork & Bjork, 2011). To our knowledge, this is the first direct empirical demonstration of a correlation between state boredom and a quantitative manipulation of information-content.

Experiment 2

This experiment examined the extent of task disengagement and exploratory behavior in response to information-content. Using the same three conditions that elicited differential boredom levels in participants in Experiment 1, we modified the number-prediction task to allow participants to decide, on their own, whether they wished to persist in the current game or quit and move on to another game. This afforded a more direct examination of the relationship between information content and exploration, and how this traded off against present reward.

Methods

Participants Twenty Princeton University undergraduates (ages 18 to 22) performed the experiment. They were compensated with $12 for their time, plus a performance-dependent bonus of up to $7.

Task Participants played a variant of the number-prediction task used in Experiment 1. However, in this version, games did not have a fixed length. Rather, participants were told that a game could go on for up to one hundred trials, but they could choose to end it earlier and move on to a new game at any time by pressing the "reset" button on the screen (figure 2A). If they pressed the "reset" button, they would see a brief inter-game screen, and then start a new game with a new number-generating process. Participants were told that the task would take approximately fifty minutes, regardless of how many games they went through in that time: the task finished at the end of the current game once the fifty-minute time period was up. There was no boredom slider in this design. In all other respects, the tasks and conditions (Gaussian, Certain and Random) were the same as in Experiment 1. After each game ended (either because the participant pressed the "Reset" button, or after 100 trials), the next game was drawn from one of the three conditions with equal probability.

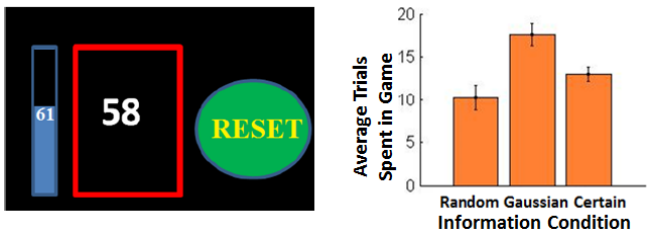


Figure 2: A. Variant of the Experiment 1 task. Participants could click **a** 'Reset' button to end current game and start a new one. B. Participants spent most time in the Gaussian games (where information content was most useful), despite the fact that the Certain games were the most rewarding.

Results

No participants chose to stay in any game for the entire duration of one hundred trials; all pressed the "reset" button to move on to a new game well before the total number of possible trials in the current game had elapsed. There was a significant difference, however, in the average number of trials spent in a game before choosing to switch (repeated measures ANOVA, $F(2,69) = 7.04$, $p < 0.01$; fig 2B), with most participants spending significantly longer on games in the Gaussian condition than in either of the other two conditions ($F(2,69)=9.22$, $p<0.01$). The difference between the Certain and Random conditions was not significant (paired t-test, $t(23) = -1.29$, $p = 0.206$). Furthermore, the probability of switching away after the first trial of a new game was significantly higher for the Certain games than either of the other two conditions; again, the values for the

Certain and Random and Certain conditions were statistically indistinguishable ($t(23) = 0.11$, $p = 0.91$).

Discussion

The results of this experiment suggest that the conditions with low information content, and associated with boredom in Experiment 1, carry a "penalty." The three conditions elicited a U-shaped curve for quitting times that mirrored the curve for boredom in Experiment 1 (figure 2B). This behavior is particularly striking for the Certain condition, in which the potential for reward was highest (participants had access to the correct prediction on every trial). Despite this, none of the participants stayed in a Certain game until it terminated, choosing instead to switch away from these games *more* frequently than from the other two conditions. Thus, participants were willing to take a point loss in order to quit the Certain game early.

This switching behavior resembles exploration – foregoing current reward-maximizing behavior in favor of options associated with a greater likelihood to gain information. There are at least two ways in which such behavior could be viewed as adaptive. First, it might help improve participants' representation of the task environment, and thus make better decisions about which tasks to perform: Experiencing more games could lead to faster discrimination between the Gaussian and Random games, by reducing estimation uncertainty (Payzan-LeNestour & Bossaerts, 2011). This could help participants determine the condition they were in earlier in the game which, in turn, would allow them to quit Random games earlier — a reward-maximizing strategy. Second, even if participants were not consciously trying to improve their representation of the task environment, but rather just aimed to terminate boring games earlier, it is possible that this drive to escape boring situations reflects an endogenous bias toward exploration, acquired over the course of evolution and/or development. Such a bias may reflect the prevailing value of exploration in the real world which is complex and rich in opportunities to gain information. This is consistent with previous work suggesting that humans and animals show an inherent aversion to low-stimulation, information-poor tasks (Fowler, 1965). Experiment 3 was conducted to further explore this possibility and, in particular, the idea that boredom, exploration and the value of information are sensitive to the alternatives available in the environment.

Experiment 3

This experiment examined whether boredom and exploration are dependent not only on the information content of the current task, but also on the (perceived) alternatives in the task environment. This builds on prior suggestions that engagement in a current task engagement is sensitive to global properties of the environment (Fowler 1967; Csikzentymihalyi 1997), and that motivation and boredom are associated with the opportunity cost of current behavior compared to alternative possible behaviors (Eastwood et al. 2012; Kurzban et al. 2013). However, to our knowledge, these ideas have not yet been tested in controlled laboratory experiments..

Here, we tested whether it was possible to change people's perceptions of boredom by manipulating the availability of other more or less attractive options in the task environment, and the extent to which this impacted exploratory behavior.

Methods

Participants Forty participants recruited from the Princeton University undergraduate community received course credit for participating in this experiment.

Task Participants played a task that consisted of two parts (referred to as part A and part B, figure 3A), played in order. They were told at the beginning of the experiment about both parts, and what each part would entail. They were also regularly reminded about part B while playing part A (they received three reminders, every five minutes, for the twenty-minute duration of part A).

Part A involved a two-armed bandit horizon task used in previous work to quantify exploratory behavior (Wilson et al., 2014). Participants had to choose between an ambiguous bandit (from which they had seen only one reward sample), and an unambiguous bandit (from which they had seen three reward samples). The decision horizon of the task was manipulated to be either short (a total of five trials: four forced-choice trials during which participants received three samples from one bandit, and one sample from the other, and one free-choice trial in which they could choose whichever bandit they wished), or long (a total of ten trials: four force-choice, six free-choice trials). This task can be used to quantify exploratory behavior (defined as the frequency with which the ambiguous option is chosen when its estimated value falls below that of the unambiguous option), and has shown that such behavior increases with task horizon. Every seven games, participants received a query screen that asked them to assess task-related factors such as difficulty, the average number of points they earned, or how many games they have played so far. Among these questions, there were regular queries about their level of interest in the game, which they were asked to rate a total of six times over the course of the task. Part A was the same for all participants, and lasted for a total of seventy-two games and approximately twenty minutes.

Part B differed between participants. There were four conditions, each involving one of a set of tasks that had been previously rated by a different sample of participants in a brief interest-rating study. As noted above, participants were instructed from the start that they would be performing Part B after Part A was finished, and they were reminded about it three times during part A. Each participant was randomly assigned to one of the four conditions, as follows: 10 participants watched a "CrashCourse" YouTube video (previously rated as a highly interesting task); 10 participants counted the number of words in a two-page mathematical typography article (a task previously rated as highly boring); 11 participants played a simple color-

matching game (previously rated at medium levels of interestingness), and 9 participants played another round of a bandit task very similar to the one played in part A.

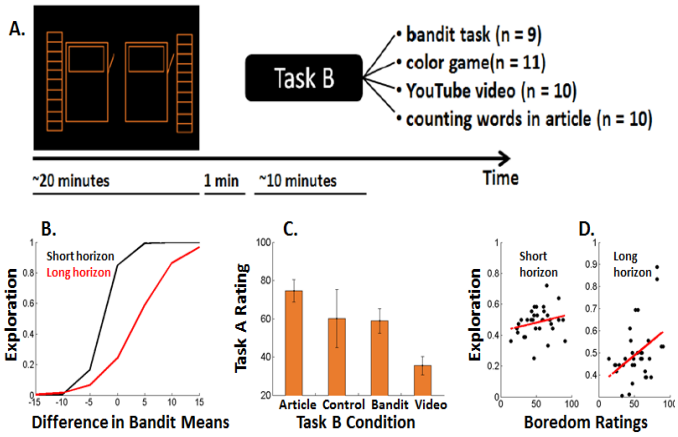


Figure 3. A. Participants played 20 minutes of the two-armed bandit horizon task (Wilson et al., 2014), followed by a break, and 20 minutes of Task B. B. Choice curves for the short horizon (black) and long horizon games (red)show more exploration in long horizon. C. People's ratings of how boring Task A was differed depending on which Task B they had to perform. D. Higher boredom ratings led to more exploratory behavior in the long horizon.

Results

Participants rated the interestingness of the bandit task (played by all participants during part A) differently, depending on the task they were told they would play during part B: the YouTube video (rated as highly interesting), the word-counting task (highly boring), or one of the control tasks (intermediate ratings). A one-way ANOVA showed a significant main effect of condition ($F(2,26) = 9.07$, $p < 0.01$), with participants who expected to watch the video (in part B) rating the bandit task (while performing it during part A) as more boring than those participants who expected to perform the word-counting task (paired t-test, $t(1,15) = 5.21$, $p < 0.01$). The ratings for the two controls fell in an intermediate range (fig 3C).

Exploration in the bandit task was defined as choosing the ambiguous bandit. Participants replicated the results from previous studies (Wilson et al., 2014) regarding the impact of decision horizon on exploration: the decision curve for the long horizon (fig. 3B, red) is flatter and shifted to the right compared to the curve for the short horizon (black), suggesting higher exploration for long horizons, both in the form of decision noise and of information bonus.

Average exploration within each correlated significantly with participants' ratings of the task: the higher the boredom rating, the more likely participants were to explore (figure 3D). This pattern was only observed in the long decision horizon, and not in the short horizon (F (1,33) = 6.01, p = 0.02 for Horizon 6, $F(1,33) = 1.88$, $p = 0.17$ for Horizon 1), and it was observed for estimates of both exploration-related parameters: decision noise and information bonus (which were both significantly higher in the long horizon with high boredom ratings).

Discussion

Experiment 3 showed that it was possible to manipulate participants' perceptions of boredom and bias toward exploration within a given task, based on a manipulation of the task environment (in this case, the next task to be performed). All participants played the same bandit task for the same period of time, but those participants who had been told they would perform an interesting task (watch a YouTube video) after finishing the bandit task, rated the bandit task as significantly more boring than those who had been told that they would have to perform a boring task (word-counting; figure 3C). This is consistent with previous findings regarding the effect of increased available stimulation on relative motivation (Fowler, 1967), as well as the theoretical framework proposed recently by Kurzban et al (2013) that relates boredom to perceived opportunity cost.

A second result was the strong correlation between participants' boredom and their exploratory behavior in the two-armed bandit horizon task: participants who rated the bandit task as more boring also showed significantly higher exploration (fig 3E). Increased exploration in response to a boring situation has previously been suggested in the literature (Cohen, McClure & Yu 2007), but this is the first report of an empiricaclly measured correlation between self-reported boredom and a quantitative measure of exploration.

Interestingly, Task B's identity is, by design, irrelevant to the value of exploration *within* the bandit task (since Task B only occurs after it is over), but nevertheless affected it. This suggests the perceived value of alternatives is a globally estimated quantity that generalizes freely and potentially inappropriately to particular, more constrained choices, much as has been suggested for opportunity costs in other domains (Niv et al., 2007).

General Discussion: An Information-Sampling Account of Boredom and Exploration

The results of the three experiments described above provide convergent support for a relationship between information context, boredom, and exploration. We showed that levels of reported boredom correlate with the amount of useful information that participants can extract from the environment, such that when there is too much or too little available information, they become more bored (Experiment 1). We also showed that perceptions of boredom can be modulated based on the environment in which the task is performed (Experiment 3). This strongly suggests that when people compute the value of staying with the current behavior, they take into account some measure of relative value between the local and global environments. Lastly, we showed that when boredom levels are high (due to low information content), people show a greater propensity to switch away from the current task – both in situations when that may be useful (Experiment 3), and even when it would appear to be suboptimal (Experiment 2).

These findings suggest two factors contribute to boredom and the tendency to switch behavior: a lack of useful information in the current task, and the availability of more (valuable) information outside the current task. This is consistent with both previous theoretical work linking boredom to too much or too little available information (Schmidhuber, 1997; Csikszentmihalyi, 2000), and with more recent accounts that suggest that boredom might signal an increased opportunity cost of performing the current task (Kurzban et al, 2013). Our results are also consistent with the idea that humans need constant access to a certain amount of information in order to maintain a satisfactory level of adaptive behavior (Zakay 2014), and that that information comes in the form of optimal levels of variability in the environment (Kidd et al., 2012).

The work presented here provides an important first step toward formalizing the link between information, boredom and exploration. Our treatment of exploration — as the decision to switch away from a current task with the specific goal of acquiring more information — parallels the formal theoretical treatment of foraging behavior, in the marginal value theorem (MVT; Charnov, 1976), as the decision to switch away from a current task with the specific goal of acquiring greater reward. In related work, we have sought to formalize this definition of exploration in a model that relates it to information acquisition in a way that parallels the definition of foraging for reward in MVT, and to test this model in an empirical study. In aggregate, this line of work promises to offer a normative understanding of human exploratory behavior and, within this framework, to cast boredom as an adaptive response to situations that should favor exploration in the service of learning, and the maximization of long-term reward.

References

Behrens. T.E.J., Woolrich, M.W., Walton, M.E. Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience* 10: 1214 – 1221

Bench, S. W., & Lench, H. C. (2013). On the function of boredom.*Behavioral Sciences*, *3*(3), 459-472.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*

Bromberg-Martin, E. S., & Hikosaka, O.(2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, 119-126.

Cohen, J. D., McClure, S. M. & Yu, A. J. (2007).Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society: Biological Sciences. 362*, 933–942

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety*.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441, 876 – 879

de Rijk, A. E., Schreurs, K. M., & Bensing, J. M. (1999). Complaints of fatigue: related to too much as well as too little external stimulation?. *Journal of Behavioral Medicine*, *22*(6), 549-573.

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged mind defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482-495.

Fowler, H. (1967). Satiation and Curiosity: Constructs for a Drive and Incentive-Motivational Theory of Exploration. *Psychology of learning and motivation*,

Gottlieb, J. (2012). Attention, learning, and the value of information. *Neuron*,*76*(2), 281-295.

Hill, A. B., & Perkins, R. E. (1985). Towards a model of boredom. *British Journal of Psychology*, *76*(2), 235-240.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449-460.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661-679.

Luciw, M., Kompella, V., Kazerounian, S., & Schmidhuber, J. (2013). An intrinsic value system for developing multiple invariant representations with incremental slowness learning.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*(37), 12366-12378.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue?. *Physiology & Behavior*, *93*(1), 369-378.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. PLoS Computational Biology, 7, e1001048.

Schmidhuber, J. (1991, November). Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on* (pp. 1458-1463). IEEE.

Şimşek, O., & Barto, A. (2006). An intrinsic reward mechanism for efficient exploration. *Proceedings of the 23rd international conference*, 833-840.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Zakay, D. (2014). Psychological time as information: The case of boredom.*Frontiers in psychology*, *5*.

Analyzing Distributional Learning of Phonemic Categories in Unsupervised Deep Neural Networks

Okko Räsänen
(okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000,
Aalto 00076, Finland

Tasha Nagamine
(tasha.nagamine@columbia.edu)
Department of Electrical Engineering, Columbia University,
500 W. 120th St., New York,
NY 10027 USA

Nima Mesgarani
(nima@ee.columbia.edu)
Department of Electrical Engineering, Columbia University,
500 W. 120th St., New York,
NY 10027 USA

Abstract

Infants' speech perception adapts to the phonemic categories of their native language, a process assumed to be driven by the distributional properties of speech. This study investigates whether deep neural networks (DNNs), the current state-of-the-art in distributional feature learning, are capable of learning phoneme-like representations of speech in an unsupervised manner. We trained DNNs with unlabeled and labeled speech and analyzed the activations of each layer with respect to the phones in the input segments. The analyses reveal that the emergence of phonemic invariance in DNNs is dependent on the availability of phonemic labeling of the input during the training. No increased phonemic selectivity of the hidden layers was observed in the purely unsupervised networks despite successful learning of low-dimensional representations for speech. This suggests that additional learning constraints or more sophisticated models are needed to account for the emergence of phone-like categories in distributional learning operating on natural speech.

Keywords: statistical learning; distributional learning; language acquisition; phonemic categories; speech perception; categorical perception; connectionism

Introduction

Acquisition of the native language phonemic system is an important step in early language acquisition, enabling a transition from generic auditory perception towards symbolic and generative representation of words and subword units. Although it is known that infants adapt to the distributional characteristics of their native language sound system during their first year of life (Werker & Tees, 1984), it is less obvious whether early perceptual representations of speech actually consist of sequential invariant atomic units such as phones or phonemes before lexical learning, or whether adult-like phonemic system emerges only through extensive experience and learning at multiple levels of language representations (c.f., Werker & Curtin, 2005; see also Räsänen & Rasilo, 2015, for a recent overview).

Since distributional learning can be framed as unsupervised machine learning from speech data, a number of computational studies have investigated how phone categories could be clustered from acoustic speech input only and how selective these automatically discovered sound categories are (e.g., de Boer & Kuhl, 2003; Vallabha, McLelland, Pons, Werker, & Amano, 2007; Kouki, Kikuchi & Mazuka, 2010). These studies have typically limited their analysis to pre-segmented or otherwise carefully selected subsets of speech tokens and/or phone categories. In addition, they have enforced an explicit clustering procedure of potentially infinitely many different acoustic tokens into a finite number of discrete and disjoint phone classes. The general finding has been that these acoustic clusters tend to be selective towards specific phones but are far from a representational system that would be invariant to non-phonological acoustic variability across talkers, speaking styles, and other factors. Due to the challenges in bottom-up clustering speech directly into phonemic categories, a number of computational models (e.g., Feldman et al. 2013; Elsner et al., 2012) and theoretical frameworks (e.g., Werker & Curtin, 2005; Räsänen & Rasilo, 2015) propose that phonemic learning is inherently tied to concurrently emerging lexical knowledge and should not be considered as an isolated process strictly preceding word learning.

Despite the emerging view that phonemic learning cannot be addressed in isolation from lexical learning, it is still important to understand how different aspects of language experience affect the development of speech perception capabilities. One of these aspects is the question of how much of early adaptation to one's native language can still be driven by purely auditory statistics. In the present study, we investigated whether deep neural networks (DNNs), a set of powerful machine learning techniques for feature learning, are capable of extracting phoneme-like representations from continuous speech similarly to their capability of learning mammalian-like visual receptive fields from image data. More specifically, we asked whether the representations resulting from unsupervised distributional learning of speech reflect phonemic contrasts of the language when the network is forced to discover low-dimensional re-presentations of the initially high-dimensional acoustic space, i.e., whether phonemic variation dominates other distributional properties of natural continuous speech.

Deep neural networks and phonemic learning

Deep neural networks, which are artificial neural networks with two or more hidden layers, are the current state-of-the-art in the discovery of non-linear structure (or *features*) from

stochastic data (Hinton, 2014). They have also been shown to provide good approximations for the emergence of increasingly abstract visual features in mammalian visual pathway (e.g., Cichy et al., 2016). In the context of speech, DNNs have become the state-of-the-art acoustic models in standard automatic speech recognition (ASR) systems due to their scalability and representational power in comparison to the previously used shallow models such as Gaussian-mixture models. In addition, purely unsupervised deep autoencoder networks (see Methods) have been shown to be effective for learning low-dimensional representations from high-dimensional acoustic input in the absence of any supporting linguistic information (e.g., Deng et al., 2010).

In the previous work, Nagamine, Seltzer, and Mesgarani (2015) showed that hidden layers of a feedforward neural network become increasingly selective to phone categories and phonetic features when trained on continuous speech. The selectivity observed in the DNN was also found to be similar to the phonemic selectivity observed in the human superior temporal gyrus (Mesgarani et al., 2014). However, Nagamine et al. trained their network in a supervised manner using phonetic labels of the input acoustic vectors as targets for the DNN output layer. This means that the entire network was optimized to perform discrimination of the acoustic input in terms of the given phonetic categories–the standard approach taken in ASR.

In contrast to supervised learning, concurrent phonetic labeling of speech input is not available to infants learning their native language. The previous study therefore leaves open whether similar increasingly abstract phonemic structure can also emerge from purely auditory learning when the neural network attempts to find a low-dimensional but high-fidelity code for the incoming acoustic input. If so, this would provide evidence for how much of the native phonemic invariance properties can be acquired simply by listening to speech in the absence of any further constraints and give insight to the type of "receptive fields" that become responsible for phonemic perception. On the other hand, a failure to learn increasingly invariant phonemic representations from acoustic input would suggest that local short-term dependencies of speech, as captured by the feedforward networks, would be insufficient for the emergence of phonemic categories and that additional constraints from concurrently emerging knowledge at different levels (e.g., Feldman et al., 2013; Räsänen & Rasilo, 2015) or different network topologies are needed in the learning process (see, e.g., Synnaeve et al., 2014).

In order to investigate whether DNNs as hierarchical generative models of speech are capable of acquiring some type of invariance properties with respect to phonetic or phonemic representations of the input speech, we conducted a number of learning simulations using the existing standard unsupervised DNN architectures.

Methods

Speech input to the DNNs was represented using logarithmic Mel-spectral features similarly to the earlier work (Nagamine et al., 2015). The input signal was converted to 10-ms feature frames $\mathbf{x}_t$ using a sliding 25-ms window and computing 24-band log-Mel-spectrum from each window. The features were Z-score normalized across each utterance to ensure proper scaling for neural network input. The final inputs to the networks were formed by concatenating 11 subsequent Mel-spectrum frames $\mathbf{x}_t$ to a single input vector $\mathbf{f}_t = [\mathbf{x}_{t-5}, \mathbf{x}_{t-4}, \ldots, \mathbf{x}_{t+5}]^{\mathrm{T}}$. Unlike Nagamine et al. (2015), we decided to leave out the first and second derivatives of the Mel-spectra since the resulting time-frequency patches already contain local temporal dynamics of the input (as confirmed by the replication of the earlier findings; see the Experiments section).

Three standard DNN architectures were investigated in the present work: **1)** a supervised deep multilayer perceptron (MLP) for classification of speech to phone labels (replication of the previous work by Nagamine et al., 2015), **2)** a stack of unsupervised restricted Boltzmann-machines (RBMs) that learn a generative model over the input data, the entire stack referred to as a deep belief network (DBN), and **3)** an unsupervised deep feed-forward autoencoder network (AEN) that learns to map input speech into a low-dimensional bottleneck-layer and then expand (decode) that representation back to a reconstruction of the original input.

The use of DBNs and AENs to study distributional phonetic learning was motivated by the finding that DBNs are capable of learning increasingly abstract visual features from image data (Hinton & Salakhutdinov, 2006) and achieve superior dimensionality reduction performance in comparison to linear models such as PCA in many tasks. The assumption in the present study is that the phonemic identity of the speech segments might require fewer bits to encode than the details of the acoustic input itself, and therefore a generative network with a decreasing number of nodes in the higher and narrower layers should become more "*phonemic*" in its representation when dimensionality reduction is imposed on the data. This is, of course, only if the variance in the acoustic input is best explained across dimensions correlated with phonemic identities instead of some other low-dimensional description of the input, and that the learning algorithms used to estimate DNN parameters are capable of finding this manifold.

In our experiments, the MLP was trained in the standard way using acoustic feature vectors $\mathbf{f}_t$ as input to the network with H hidden layers, computing the activation of the output layer $\mathbf{h}_{\mathrm{out},t}$ given the input, and then calculating the error of the activation with respect to a target vector $\mathbf{g}_t$ denoting the phonemic identity of the input vector. The weights of the network were then tuned using backpropagation (BP) algorithm in order to minimize the error of the output layer (Rumelhart, Hinton & Williams, 1986).

DBNs were obtained by first training a three-layered stack of RBMs incrementally layer-by-layer, always fully training the parameters of an RBM with one hidden layer at a time (Fig. 1), then freezing those parameters and using the probabilities of the hidden unit activations given the training data as the "visible layer" input to the next hidden layer. As

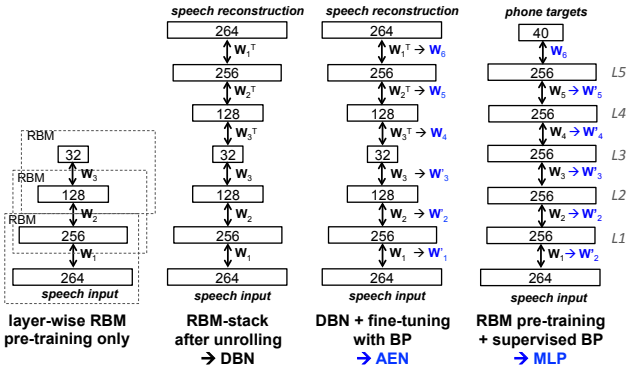


Figure 1: A schematic description of how pre-trained RBMs are "unrolled" to five-layer feedforward DBN and AEN networks used in the present study. Weights after layer-by-layer RBM pre-training are shown with black **W** and weights after error backpropagation (BP) are shown with blue **W**. The baseline supervised MLP topology is shown on the right (adapted from Hinton & Salakhutdinov, 2006).

a result, the stack becomes a hierarchical generative model P(**f** | **h**) over the training data with higher hidden layer activations **h** representing increasingly complex features of the input data (see Hinton & Salakhutdinov, 2006).

In order to obtain five-layered DBNs (as defined in the present study), the stack of RBMs was "unrolled" (Fig. 1) to a feed-forward network by mirroring the structure of the network on top of the low-dimensional *bottleneck*-layer with the output layer corresponding to a reconstruction of the input speech (see Hinton & Salakhutdinov, 2006). In this case, the three first layers correspond to the standard feed-forward activations of the original stacked RBM while the last two layers are identical to the top-down reconstructions of the same model.

Finally, AENs were obtained by fine-tuning the DBN weights using BP in order to minimize the Mel-spectrogram reconstruction error at the output of the network, therefore breaking the weight symmetry of the DBN.

All networks used sigmoid activation functions except for the output layer, which was always linear. In the case of MLP, instead of using the typical softmax output layer with multinomial labels, we experimented with distributed target representation by first assigning each of the unique phones c with a random vector sampled uniformly from $\mathbf{v}_c \sim[0,1] \in R^d$ ($d = 40$). Then the distribution of phones within the input time window (W = 11 frames) was encoded to a target vector $\boldsymbol{g}_t$ as a weighted mean of the random vectors corresponding to the phone labels of the frames within the window:

$$\mathbf{g}_t = \frac{1}{K}\sum_{i=t-(W-1)/2}^{t+(W-1)/2} \mathbf{v}_{t+i} \qquad (1)$$

Since this type of random mapping preserves the approximate mutual distances between representations (Johnson & Lindenstrauss theorem) while the sum of high dimensional random vectors preserves information regarding the individual components (e.g., Kanerva, 2009), the approach enables creation of fully dense target vectors that represent arbitrary distributions of phones, and, in general, provides opportunities to incorporate structured target representations using fixed-dimensional outputs (see Gallant & Okaywe, 2013). DBNs always utilized a Gaussian input layer to accommodate the z-score-normalized input.

For the baseline MLPs, we used the same network configuration as in Nagamine et al. (2015) by using five hidden layers, each consisting of 256 nodes (Fig. 1, right). As for the DBNs and AENs, we initially experimented with a number of different network layouts and layer sizes using a subset of the training data, including bottleneck-architectures with gradually decreasing number of nodes in deeper layers, bottlenecks with different numbers of nodes, and even expanding networks with an increasingly many nodes at higher layers (see Table 1 for a summary). Since there were no major qualitative differences in the findings, one basic bottleneck layout of **d** = [256, 128, 32, 128, 256] nodes per layer for the DBN and AEN was chosen for more detailed analysis.

The dimension of the input layer for all networks and of the output layer for the AENs and DBNs was always 264 (11 frames x 24 frequency bands). In order to ensure that the training of the networks was successful, we always manually verified that the reconstruction or classification error decreased monotonically as a function of the epoch number during BP, and that the generative networks were capable of performing sensible reconstructions from the input Mel-spectrograms.

Data

Two qualitatively different corpora were used in the experiments in order to get a comprehensive picture of the learning process. The TIMIT corpus of American English read speech (Garofolo et al., 1993) was used as the primary dataset since the earlier work was evaluated on the same data and since TIMIT represents natural variation of speech across multiple talkers and dialects. The full training set of 4620 sentences was used to train the DNNs and the test set of 1620 sentences was used in the phonemic invariance analyses. Both sections contain speech from male and female talkers and the data is hand-labeled for phone segments.

In addition, we used enacted child-directed speech from the Caregiver (CG) Y2 UK corpus (Altosaar et al., 2010) to investigate whether results differ for limited-variability speech from a small vocabulary of approx. 80 words, each word repeated multiple times in the training set, and when all speech comes from a single talker ("a caregiver"). For this purpose, 1600 utterances from *Talker-01* of the corpus were used to train the DNNs and a remaining set of 797 utterances were used to probe the phonemic selectivity of different layers. The CG UK Y2 corpus comes with a phone annotation created by forced-alignment from text to speech using an automatic speech recognizer. Due to the simplicity of the material, this reference can also be considered as highly reliable at the level of individual phones.

The data were randomly divided into a set of minibatches for training, each minibatch consisting of 100 samples for each RBM parameter update and 1000 samples for each BP update. In all simulations, BP was always run for 25 epochs similarly to Nagamine et al. (2015) whereas DBN-pretraining consisted of 15 epochs per layer.

Methods for network selectivity analysis

Activations of the networks were analyzed in the context of the underlying phone annotation. Original 61 phone classes of TIMIT annotation were first mapped to the reduced set of 39 phones and with silences and closures excluded (Lee & Hon, 1989). The set of 38 unique phones in the original CG annotation was used in its original form. In order to study phone-specific activations of the networks, only the test data input frames consisting of at least 90.9% of a single phone segment were included in the analysis, corresponding to 19706 samples on TIMIT test set and 4749 samples on the CG *Talker-01* data.

Similarly to Nagamine et al. (2015), the activation of each layer in the context of different phone classes was analyzed using the F-ratio. First, the activation vectors consisting of all nodes within a layer of interest were grouped according to the phone labels associated with the inputs. The cross-phone variance of the node activations was then compared to the intra-phone variance, revealing whether the node activations for different realizations of the same phone are more similar than activations for two any arbitrary phone segments. By measuring the average F-ratio across layers, we can probe whether the activations for different allophones of the same phone class are more consistent in some layers than others. In addition to the F-ratio, we measured the mean mutual information (MI) between node activations and corresponding phone labels to see how many bits of information does each individual node, on average, contain regarding the phone classes of the input vectors.

Finally, k-Nearest Neighbor (KNN) classification of the layer-specific activations into phone categories was performed in order to evaluate how well the full pattern of activation in a layer discriminates between phone classes. More specifically, every activation $\mathbf{h}_{i,t}$ of layer i for input $\mathbf{f}_t$ was used as a single feature vector for classification (e.g., $dim(\mathbf{h}_i) = 256$ in all hidden layers i of the MLP). Four-fold cross-validation performance with 75% of the vectors as training data and 25% of vectors as testing data was then computed. The parameter k was always varied between [1, 10] and the best result across this range was chosen as the classification accuracy for each fold before averaging the results across all folds. In addition to analyzing phone selectivity, we also included analyses of selectivity towards manner of articulation (MOA) and talker gender using the TIMIT data and the same set of measures.

Results

Fig. 2 shows the overall analysis results from the three different networks (supervised MLP, unsupervised DBN and AEN) for the TIMIT data with multiple talkers. Fig. 3 shows the corresponding results for the single-talker IDS speech from the CG corpus. Table 1 shows a summary of KNN-based phonetic discriminability of layer activations for alternative network topologies tested on TIMIT.

The first finding is that the supervised MLP replicates the earlier results of Nagamine et al. (2015) with increasing network layers showing higher selectivity towards phone classes (max. improvement of 11%) and less sensitivity to talker identity (gender), as measured by F-ratio or KNN classification performance. In contrast, no such invariance properties are observed for the unsupervised networks. Although F-ratio of the bottleneck-versions of the AEN and DBN increases during the reconstruction of the input, the activations are not more informative regarding the corresponding phone identity as revealed by decreasing KNN performance. With a fixed or expanding number of nodes in the hidden layers (Table 1), very minor improvements in phone selectivity (max. 2.7%) are observed in comparison to input features but without any abstraction from gender-specific patterns.

In addition, the MI between individual node activations and phone labels is not positively correlated with the discriminability of the overall pattern of activation across all nodes–not even in the supervised case (Figs. 2 and 3). A closer analysis of the distributions of node-specific MI-values in case of the MLP revealed that the MIs become more tightly concentrated towards small values with an increasing layer number. Simultaneously, the number of highly informative individual nodes decreases. This suggests that the representations at higher layers are inherently distributed and the same nodes contribute to encoding of multiple different phone classes. When analyzed individually, each node will naturally show increased selectivity towards an internally coherent subset of all possible speech inputs, but this should not be confused with overall capability of the individual nodes to represent abstracted categorical knowledge.

In order to ensure that the results were not affected by overfitting of the model to the data, we also conducted the same set of analyses for the activations on the training data. No qualitative differences were observed in the results in this case. In addition, we re-ran the experiments using a shorter input window length (5 frames ≈ 50 ms) to ensure that the phonemic structure was not lost due to the inclusion of the neighboring temporal context in the acoustic representations during the training stage. Again, the results were qualitatively similar to those reported with a longer input window.

Finally, since the KNN performance was always higher for a BP-fine-tuned autoencoder in comparison to the pre-training only, we also ran further 100 iterations of the BP-algorithm to see whether the KNN performance of the AEN layers would increase above the original input Mel-spectrum selectivity with the help of extra training. However, the KNN-based selectivity measure flattened out around the level observed in Figs. 2 and 3 and then started to decrease with more training epochs, likely due to overfitting.

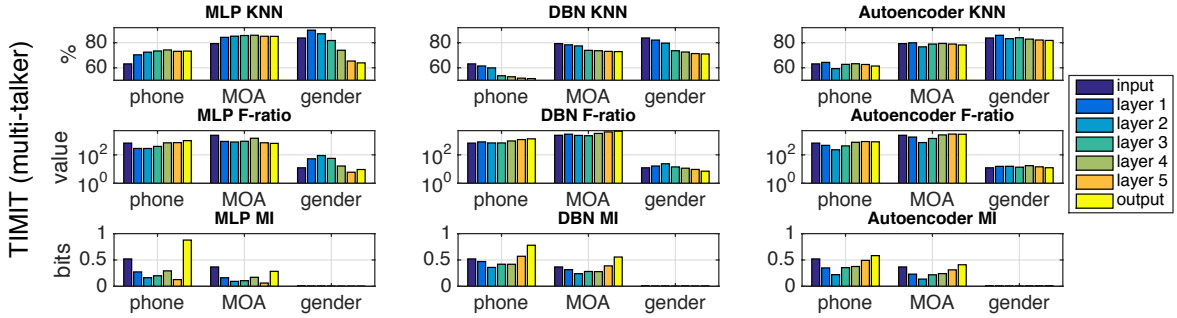


Figure 2: Network invariance with respect to phones, manner of articulation (MOA), and speaker gender for multi-talker training. The supervised MLP network is shown on left, followed by the unsupervised pre-trained DBN (center), and the fine-tuned autoencoder (right). Top: KNN classification performance using layer activations as features. Middle: F-ratio of node activations w.r.t. sample classes of interest. Bottom: mutual information (MI) between classes and node activations. DBN and AEN shown for **d** = [256, 128, 64, 128, 256].

Table 1: KNN performance (% correct) for phone and gender classification in TIMIT data for different hidden layers (L) in other tested unsupervised network topologies.

topology	L1	L2	L3	L4	L5	
DBN [256 256 256 256 256]	**62.0**	59.0	56.7	55.2	54.6	phones
AEN [256 256 256 256 256]	64.0	59.9	64.2	**65.0**	64.6	
DBN [256 512 1024 512 256]	**61.4**	60.2	58.2	57.0	56.0	
AEN [256 512 1024 512 256]	62.4	61.5	64.9	**65.7**	65.5	
DBN [128 64 8 64 128]	**59.4**	56.8	38.5	38.5	38.2	
AEN [128 64 8 64 128]	**60.6**	57.4	55.6	55.5	55.5	
AEN [256 256 256 256 256]	84.7	82.2	84.7	**84.9**	84.9	gender
AEN [256 512 1024 512 256]	84.6	83.7	85.5	**86.5**	86.0	
AEN [128 64 8 64 128]	**79.0**	75.5	71.4	71.2	70.7	

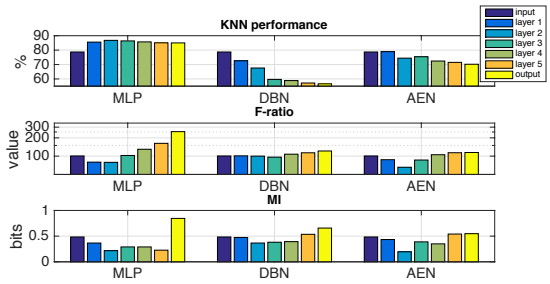


Figure 3: DNN phone invariance measures for the CG single-talker data for different network types and layers. DBN and AEN shown for **d** = [256, 128, 64, 128, 256].

Overall, it seems that the DBN is simply smoothing the input data (lower KNN-performance and more uniform activations in terms of F-ratio at deeper layers) whereas fine-tuning of the AEN leads to low-dimensional but detailed representations that encode both suprasegmental and segmental acoustic details. Unlike the supervised MLP, neither the DBN nor AEN exhibit increased phonemic invariance in comparison to the original input features.

Discussion and conclusions

The present experiments investigated the emergence of phonemic representations in unsupervised deep neural networks using adult-directed speech from multiple talkers similarly to the supervised counterpart performed earlier (Nagamine et al., 2015) and on single-talker data of child-directed speech. The central finding is that the studied deep feedforward networks did not show similar increased selectivity towards phonemic structure that was observed in the networks trained in a supervised manner (Nagamine et al., 2015) or in the auditory neurons of the superior temporal gyrus as analyzed by Mesgarani et al. (2014).

The results are also qualitatively different from the earlier clustering studies that have reported above-chance grouping of acoustic spectra or formants frequencies to disjoint phone-like categories (e.g., de Boer & Kuhl, 2003; Vallabha et al., 2007; Kouki et al., 2010). However, a major difference to the earlier clustering studies is that the DNNs do not force division of the data samples into a finite number of categories similarly to standard clustering algorithms, but learn distributed representations of the statistical structure of the data. In addition, our input data to the learning process was not carefully chosen to represent stable parts of vowel sounds but contained continuously extracted slices of the speech input – a situation that the auditory system has to also face unless further temporal constraints such as syllabic (Räsänen, Frank & Doyle, 2015) or phonetic (e.g., Räsänen, 2014) boundary cues are included in the process. This leaves open whether the phonemic structure in DNNs would become more explicit under more constrained but ecologically plausible learning settings. Another possibility is that the use of recurrent neural network architectures could learn better context-dependent models for speech patterns as they do not assume independence of the neighboring speech frames similarly to the currently studied networks. Although such units would conflict with the idea of a phone or phoneme as a context-independent cluster of spectral properties, a simplification often assumed in early language acquisition research, it is well known that human speech perception also operates on longer time-spans than individual segments.

Results from the supervised paradigm clearly indicate that the selectivity of the internal representations become more phonemic when the target output is also phonemic in nature. In the context of modeling early language acquisition, the targets cannot be discrete phone labels as such. However,

the precise labels could be substituted to other available and correlating information such as larger structural units the input frames belong to (e.g., Elsner et al., 2012; Feldman et al., 2013; Synnaeve et al., 2014) or even the cross-situational referential context in which the speech is observed (Räsänen & Rasilo, 2015). This could lead to similar, albeit slower, learning of representations showing phonemic invariance.

Interestingly, despite the absence of increased phonemic invariance in the unsupervised networks, the findings should still be compatible with the basic idea of distributional adaptation to the native language phonetic system (e.g., Kuhl et al., 2008) since the studied networks learn a generative statistical model over the training input. The input speech reconstructions from the network will depend on the familiarity with the input and are biased towards the statistical patterns of the training data. As long as the perceptual representations for speech input are assumed to correspond to the activations of the hidden layers or the reconstruction itself, the system is less sensitive to phonetic details of "non-native" speech patterns the more it is trained with one language only. This provides an analogy between human distributional learning and overfitting of statistical models to a certain set of training data. However, computational verification and implications of this idea are out of the scope of the present study and should be addressed in the future work.

Acknowledgments

This study was funded by the Academy of Finland project no. 274479 and a grant from National Institute of Health, NIDCD, DC014279 and the Pew Charitable Trusts, Pew Biomedical Scholars Program. Tasha Nagamine was funded by the From Data to Solutions NSF IGERT grant.

References

Altosaar et al. (2010). A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proc. LREC-2010*, Malta, pp. 1062– 1068.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A. (2016). Deep Neural Networks predict Hierarchical Spatio-temporal Cortical Dynamics of Human Visual Object Recognition. arXiv:1601.02970.

de Boer, B., & Kuhl, P., (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4, 129–134.

Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., & Hinton, G. (2010). Binary coding of speech spectrograms using a deep auto-encoder. *Proc. Interspeech-2010*, Makuhari, Japan, 26–30 Sept., pp. 1692–1695.

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. *Proc. ACL-2012*, Jeju, Republic of Korea, pp. 184–193.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for developing lexicon in phonetic category acquisition. *Psychological Review,* 120, 751–778.

Gallant, S., & Okaywe, W. (2013). Representing objects, relations, and sequences. *Neural Computation*, 25, 2038–2078.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., & Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. Philadelphia: LDC.

Hinton, G. E. (2014). Where do features come from? *Cognitive Science*, 38, 1078–1101.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313, 504–507.

Kanerva, P. (2009). Hyperdimensional Computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cogn. Comp.*, 1, 139–159.

Kouki, M., Kikuchi, H., & Mazuka, R. (2010). Unsupervised Learning of Vowels from Continuous Speech Based on Self-Organized Phoneme Acquisition Model. *Proc. Interspeech–2010*, Makuhari, Japan, pp. 2914–2917.

Kuhl, P. K., Conboy, B. T., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions B of the Royal Society*, 363, 979–1000.

Lee, K.-F., & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden-Markov models. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37, 1641–1648.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343, 1006–1010.

Nagamine, T., Seltzer, M. L., & Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. *Proc. Interspeech-2015*, Dresden, Germany, pp. 1912–1916.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533–536.

Räsänen, O. (2014). Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. *Proc. CogSci-2014*, Quebec, Canada, pp. 2817–2822.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, 122, 792–829.

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *Proc. Interspeech-2015*, Dresden, Germany, pp. 3204–3208.

Synnaeve, G., Schatz, T., & Dupoux, E. (2014). Phonetics embedding learning with side information. *Proc. IEEE SLT Workshop,* South Lake Tahoe, NV, pp. 106–111.

Vallabha, G. K., McLelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of National Academy of Sciences*, 104, 13273–13278.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence from perceptual reorganization during the first year of life. *Infant Behavior and Devel*., 7, 49–63.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1, 197–234.

The Structure of Names in Memory:
Deviations from Uniform Entropy Impair Memory for Linguistic Sequences

Melody Dye (meldye@indiana.edu)[1], Brendan Johns (btjohns@buffalo.edu)[2], Michael N. Jones (jonesmn@indiana.edu) [1], & Michael Ramscar (michael.ramscar@uni-tuebingen.de)[3]

[1] Department of Psychological & Brain Sciences, Indiana University, Bloomington

[2] Department of Communicative Disorders and Sciences, University at Buffalo

[3] Department of Linguistics, University of Tübingen

Abstract

Human languages can be seen as socially evolved systems that have been structured to optimize information flow in communication. Communication appears to proceed both more efficiently and more fluently when information is distributed evenly across the linguistic signal. In previous work (Ramscar et al., 2013), we used tools from information theory to examine how naming systems evolved to meet this requirement historically, and how, over the past several hundred years, social legislation and rapid population growth have disrupted naming practices in the West, making names ever harder to process and remember. In support of these observations, we present findings from three experiments investigating name fluency, recognition, and recall. These results provide converging empirical evidence for an optimal solution to name design, and offer a more nuanced understanding of how social engineering has impaired the structure of names in memory.

Keywords: sequence learning; information theory; artificial grammar; associative memory; category fluency

Introduction

Names are central to identity, and the primary means by which we individuate and discriminate ourselves from our peers. However, while they are central to human life, names are unusually difficult to learn and remember. English speakers typically find a person's name more difficult to recall than any other piece of information about them, and names form the majority of naturally occurring 'tip of the tongue' (TOT) experiences (Burke et al., 1991). Moreover, the recall of names is disproportionately impaired in old age, with TOT rates for names increasing more with age than for other words. Indeed, it has been reported that older English speaking adults consider the deterioration in name recollection to be the most disturbing age-related cognitive problem they face, and the aspect of memory that they would most like to improve (Cohen & Faulkner, 1986).

What makes names so challenging? While most nouns are generic, picking out a class of objects, names are *sui generis*, picking out an individual (Cohen, 1990). Given the size of modern populations, the design of an efficient naming system must balance the need to individuate each person against the massive rise in linguistic complexity that unique identifiers would create. Using government databases and historical records, Ramscar et al. (2013) found that historically, Mandarin Chinese and English have converged on a similarly elegant solution to this problem, smoothing identifying information over several name elements each of which is distributed in a highly efficient, Zipfian manner. However, in the recent past, social forces have conspired to disrupt naming practices in the West, altering the shape of these distributions and reducing their efficiency. This presents an excellent opportunity to conduct an *in vivo* experiment on an extant linguistic sub-system, and to investigate how changes to its distributional structure have impacted its representation in memory.

While this paper takes as its focus one highly specialized linguistic subsystem – names – its broader aim is to examine how the descriptive tools offered by information theory may lend insight into how semantic categories are represented in mind.

Principles of Information Theory

In engineering efficient communication systems, the fundamental problem concerns how to code a message generated at point A (the source) into a physical signal that can be transmitted to point B (the destination) as efficiently and reliably as possible, such that the message can still be recovered intact. The answer to this challenge can be found in the set of mathematical principles and conceptual definitions elaborated by Claude Shannon (1948) in his theory of information.

Information theory specifies the theoretical bounds on this project in terms of two measures: Channel capacity defines the maximum rate at which information can be reliably transmitted; Entropy defines the shortest description length at which a random variable can be coded. These principles can be used to assess the design of communication systems of every stripe, including natural language systems.

In the simple case of a ***name grammar***—our focus here—the space of alternatives at the source consists in the population of to-be-named individuals, and the code, in a set of tokens that, in combination, efficiently discriminates among those individuals, allowing for their successful identification. An important question is how such a grammar might be designed such that it minimizes the average length of a name, yet still reliably discriminate the intended individuals. In this, two principles are key: 1) that each element of the grammar approximate a Zipfian distribution, and 2) that the uncertainty over each element remain roughly constant (Ramscar et al., 2013). We assess each precept in turn.

Zipf's Law. Zipf's (1949) Law is a universal scaling law for lexical frequencies and is one of the most striking and empirically robust regularities known to language. For a given sample of text, the frequency distribution of its lexical

items approximates an inverse power law, in which a steep decline in frequency over the highest ranked words eases off as rank grows, producing a long tail with similarly low frequencies. Stated formally: If r is rank, $p(w_r)$ is the probability of a word of rank r, and α is the law's exponent, varying between 0 and 1, then:

$$p(w_r) \propto r^{-\alpha}$$

This principle appears to hold not only for words, but for linguistic units at every scale, such as letters, syllables, and even word meanings.

The existence, and the import, of this peculiar distributional feature has been the subject of some debate. In his 'principle of least effort', Zipf (1949) proposed that communicators seek to balance efficiency on the one hand, and comprehensibility on the other, and that these opposing forces minimize communicative effort over time, giving rise to language's characteristic distribution (an intuition formalized by Ferrer-i-Cancho & Solé, 2003). Other explanations have been given in terms of optimal coding theory (Mandelbrot, 1953), preferential attachment (Simon, 1955), and principles of memory search and information retrieval (Anderson & Milson, 1989). Amidst these diverse accounts, what is shared is a notion that the nature of the distribution is—in some sense—optimal, and accordingly, that deviations from it represent departures from optimality.

Regardless of one's theoretical commitments, it is straightforward to empirically establish the cost of deviating by quantifying the increase in entropy that results from altering the shape of the distribution (c.f. Ramscar et al., 2013). The broader question is how or why this might matter.

Entropy Rate Constancy Over the past decade, linguists have interested themselves in an important corollary of Shannon's work on channel capacity: For a particular linguistic sequence, comprised of a set of distinct elements, information transfer is at its theoretical maximum when the entropy over each element is *constant* (and continuously approximating the channel capacity). A growing number of studies have demonstrated that speakers ably conform to this principle by smoothing information over discourse—avoiding excess peaks or troughs in uncertainty through an array of articulatory strategies and lexical and syntactic choices (c.f. Genzel & Charniak, 2002; Keller, 2004; Aylett & Turk, 2004; Bell et al., 2009; Qian & Jaeger, 2012).

Languages themselves also appear to be structurally adapted for entropy management. Nouns provide a particular challenge in this regard, as they are—in many languages—the most diverse part of speech. Accordingly, lexical uncertainty will tend to be at its highest in any given slot where a noun occurs. The impact this has on processing is well-attested: One of the most common sites of disfluencies is at the determiner preceding a noun, and the less likely the noun, the more likely a disfluency (Clark & Wasow, 1998). Nouns are also the most common sites for incorrect lexical retrieval, and an array of other processing difficulties (Vigliocco et al., 1997). However, various linguistic subsystems—such as grammatical gender—can help manage nominal entropy by using preceding elements to subdivide the space of possible upcoming nouns, thereby reducing uncertainty in context (Dye et al., 2015).

Names may pose yet a greater challenge, as they draw from a wider array of phonological sequences than other nouns (Brennen, 1993) and may be less well discriminated by semantic context. However, they appear to be governed by a similar principle. One way to conceptualize a name grammar, which consists in multiple, sequential elements, is as a ***decision tree***. In a classic tree structure, the root node represents a space of mutually exclusive alternatives that are systematically partitioned at various branching points, giving way to specific outcomes at the terminals (Quinlan, 1986). The outcomes can be understood as the to-be-identified individuals; the branch points, as the requisite yes/no questions asked by a clever guesser attempting to discriminate the named individual. In a well-designed system, the information contributed by each branch should be identical, such that the (conditional) information required from one branch to the next remains constant, while the total (unconditional) information grows linearly as a function of path length.

Ideally then, name elements should serve to iteratively increase the degree to which an individual can be identified, while minimizing the uncertainty over each element in context. Experiment 1 tests this claim by varying entropy rate in an artificial name grammar.

Experiment 1

This study examines whether naming systems are more easily learned and remembered when there is a uniform entropy rate across name elements, such that identifying information is evenly distributed over the full name.

Participants

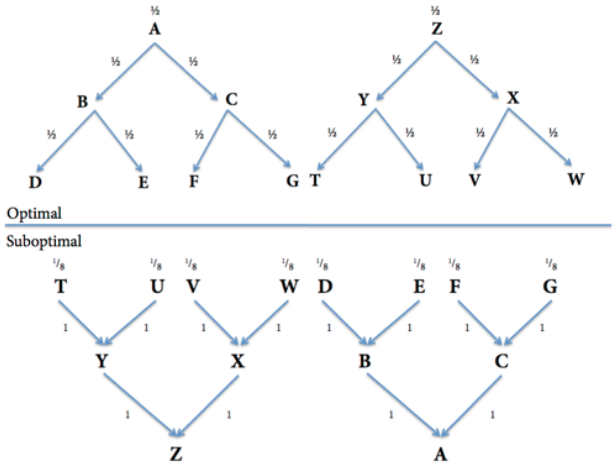
40 students (15 males and 25 females) at Indiana University participated for course credit in an undergraduate psychology course. All were native English speakers, who had completed a prosopagnosia prescreen and simple face recognition task to ensure they could recognize and discriminate faces similar to the experimental stimuli.

Materials

The present study employs a pair of artificial name grammars, comprised entirely of novel names. Each grammar consists in a set of eight full names, drawn from a finite pool of first, middle, and last name elements. The first grammar has a highly efficient information structure, with a small pool of first names followed by successively larger pools of middle and last names. The second grammar, devised to be inefficient, replicates this structure, but with the key difference that the first and last name populations are *reversed*. While names in both grammars convey the same amount of total information, the peak entropy for the suboptimal grammar is three times that of the optimal one. Where the entropy of the first name element is computed independently:

$$H(p) = H(X) = -\sum_{x \epsilon X} p(x) log_2 p(x) = \sum_{x \epsilon X} p(x) \frac{1}{log_2 p(x)}$$

and the entropy of the second and third name elements is conditioned on the preceding string:



$$H(Y|X) = -\sum_{x \epsilon X} p(x) H(Y|X = x) = -\sum_{x \epsilon X} \sum_{y \epsilon Y} p(x, y) log_2 p(y|x)$$

Figure 1: In an optimal name grammar (top), there is a small pool of first names and successively larger pools of second and last name elements. Identifying information is spread evenly over each name element, with elements sharing identical (conditional) probabilities across positions. However, if the grammar is reversed (bottom), then only the first name element is informative; the subsequent elements each have a conditional probability of 1, and thus add no information.

Names Name elements are four-letter, one-syllable nonces drawn from the English Lexicon Project (Balota et al. 2007), with the number of phonological neighbors held constant, and with accuracy and reaction time in pseudolexical decision tasks within one standard deviation of the mean.

Faces Faces are drawn from a body of similarly posed and cropped portraits taken at the Museum of Modern Art in New York City, with faces occupying near-identical spatial positions in each photograph. The set of faces were chosen to be easily discriminable, and were evenly distributed across ethnic groups. To ensure that, for the purposes of the naming study, subjects could easily re-identify the faces in each set, the stimuli were piloted in an old-new recognition task (n=41), with similar looking distractors. Selected faces averaged 91% accuracy.

Procedure

Training At study, participants were asked to learn a series of sixteen face-name pairs, eight of which were female and eight male. The female faces were assigned one set of names, the males another. Which set of name elements was assigned (A or B), and to which face-gender, was counterbalanced across participants. Name assignment within gender was randomized. The choice of grammar (optimal or sub-optimal) was varied across participants, with even assignment according to participant gender.

On each trial, a face appeared on-screen and then a name was read aloud, with the spelling of each name element appearing as it was articulated. The full name was repeated twice in view of the image, followed by a brief mask. The training sequence consisted in three repeated blocks. To control for sequential dependencies at encoding, the order of training was pseudo-randomized, such that faces from each gender were spread evenly across trials, and the same name element was never seen on adjacent trials.

Test At test, subjects completed a cued recall task, in which they were presented with a face and asked to produce its name. Subjects were instructed that partial responses would be counted. Subjects were given 30 seconds in which to complete typing. All pairs were tested exactly once, and response latencies were recorded. Specific face-name pairings were randomized between subjects. Test order was randomized in the same way as training.

Scoring Name accuracy was scored by two raters blind to the hypotheses tested. 1 point was awarded for every correct name element in the correct position. ½ a point was awarded if the element was correct, but in the wrong position, or if the name element was incorrect, but highly similar (phonologically) to the target. Any disagreements between raters were resolved through discussion.

Results

There was substantial variability in subject performance on the recall task in both conditions, with mean performance of 53.6% and a standard deviation of 26.7%. A by items analysis revealed that names trained with the optimal grammar were recalled significantly faster and more accurately than their rearranged versions (Correct Log RT: $t(15)$=-3.36, p=.002; Recall Score: $t(15)$= 3.91, p<.001; Figure 2). A parallel analysis for faces yielded comparable results (Correct Log RT: $t(15)$=-4.42, p<.001; Recall Score: $t(15)$=2.57, p=.01). An independent-samples by subjects analysis conformed with the general direction of results, but did not reach significance (Correct Log RT: $t(34)$=-1.26, p=.108; Recall Score: $t(38)$= 1.04, p=.15).

Discussion

This study revealed that names that complied with entropy rate constancy were more quickly and accurately recalled than those that did not.

The Price of Deviating

Over the past several hundred years, social legislation and rapid population growth have disrupted naming practices in the West. In particular, the rise of nation states has coincided with the adoption of census policies that have made family names hereditary, rendering them fixed, rather than mutable. In English-speaking countries, where the family name has traditionally been positioned last, this has created significant pressure on the first name to provide the individuating information, resulting in a greater diversity of higher entropy first names. Curiously, the collapse of efficient naming has not affected society equally: In the US, women and African Americans are disproportionately burdened with the least efficient names. Evidence from modeling simulations

suggests that this flattening of the distribution should have negative consequences for name memory (Ramscar et al., 2014).

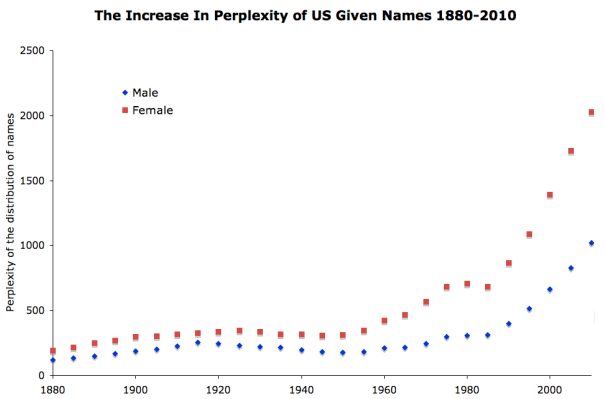


Figure 2: The perplexity of male and female names with a count ≥ 5 in US Social Security applications at 5 year intervals from 1880 to 2010 (Ramscar et al., 2013).

Are male names indeed easier to recollect than female names in American English? One of the more commonly used measures to assess semantic memory is ***category fluency***, in which participants produce as many items from a specific category as possible within a specified time frame. Fluency data are then analyzed by counting the number of items produced, and assessing the factors that influence transitions between category items.

In a fluency task, subjects tend to be superior at producing items that they have more familiarity with, all things being equal. For example, a doctor might more readily reel off a string of diseases, and a chef, a list of cooking implements. Domain-specific knowledge and experience are also reflected in gender differences: Capitani et al. (1999) report that men are advantaged in recalling man-made items, such as tools and vehicles, whereas women show more facility at food stuffs, like fruits and vegetables.

In Experiment 2, subjects of both genders were tested on their ability to produce male and female names. Given sex segregation in peer and social groups across development and early adulthood (Mehta & Strough, 2009), there is good reason to suspect that males and females will have varying experience with names in accordance with their gender. If, on average, male subjects have more experience of male names and female subjects of female names, than a double dissociation in subject performance should be observed, wherein males produce more male names, and females more female names.

An alternative possibility, assessed here, is that there may be an interaction between both the frequency with which a category is sampled and its distributional structure—a structure which may be more or less optimal for retrieval purposes (Anderson & Milson, 1989). If the organization of female names makes them harder to access in memory, this may diminish, or even eliminate, the predicted female advantage for female names.

Experiment 2

This study examines the effect of the distributional structure of male and female names on name fluency. Subjects were asked either to recall 1) the names of as many famous living people as they could, or 2) as many first names as they could.

Participants

146 students (73 males and 73 females) at Indiana University participated for course credit in an undergraduate psychology course. All were native English speakers. Subjects were randomly assigned to one of the task versions. 15 subjects were dropped from analysis for failing to follow instructions (e.g., by mixing genders), leaving n=33 for first names collected separately, 36 for first names collected together, 28 for famous names collected separately, and 34 for famous names collected together.

Design

The fluency tasks were administered in the online survey program Qualtrics. Two versions of each task were administered. In the first version of the task, no gender was specified, and subjects simply recalled as many names as they could within two minutes. In the second, subjects were asked to produce the names for one gender, and then the other, with ninety seconds for each task. Which gender was requested first was counterbalanced according to the gender of the participant.

Results

A 2 (subject gender) x 2 (name gender) repeated measures ANOVA was run over each version of the task design, revealing the same basic pattern of results across tasks: a significant main effect of name gender – more male names were produced overall – and a significant interaction of subject gender and name gender. In each case, post hoc analyses (Tukey HSD) indicated that while female subjects produced similar numbers of male and female names, male subjects produced significantly more male names than female names (p<0.001).

Modeling Fluency

To model participants' search through memory in the first name fluency task, the Luce Choice Axiom was used to quantify the contribution of different linguistic factors to first name retrieval. Given a set of possible responses, the axiom defines the probability of responding to stimulus S_i with response R_j as:

$$P(R_j|S_i) = \frac{\beta_j^{\lambda_0} S(i,j)^{\lambda_1}}{\sum_{k=1}^{n} \beta_k^{\lambda_0} S(i,k)^{\lambda_1}}$$

where β_j is the response bias for item j, and $S(i,j)$ is the similarity between items i and j. In our application of the decision rule, the response set is the full set of first names produced in the task, and $P(R_j|S_i)$ is the probability of producing a particular name given the last name that was

produced. The response bias β_i scales with name frequency, and the item similarity $S(i,j)$ with the semantic similarity between one name and the next. Frequency estimates were calculated from a 20 million sentence Wikipedia corpus, and semantic similarity from the vector cosine of the names' semantic representations, as computed by the distributional semantic model BEAGLE (Jones & Mewhort, 2007) run over that same corpus. The parameters λ_0 and λ_1 control the relative contributions of frequency and semantic similarity in producing the response. For each participant's name sequence, we determined the set of parameters that were most likely to have generated the observed data, using a model comparison technique (see Johns et al., 2013 for detailed methodology).

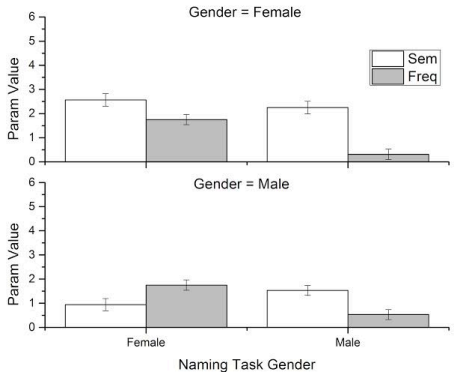


Figure 3: Fluency model of first name production.

Across subject genders, there was a significant effect of the semantic parameter [$F(1,65)=4.65$, $p<0.05$], indicating that overall, female subjects relied more on semantic information in producing first names than their male counterparts. In comparing the two tasks – producing male names vs. producing female names – there were main effects of both semantics [$F(1,65)=8.716$, $p<0.01$] and frequency [$F(1,65)=33.805$, $p<0.001$], suggesting that the influence of these parameters varied by name gender. In particular, male naming relied more heavily on semantics, while female naming relied more on frequency, overall. However, there was also a significant interaction of semantics with subject gender [$F(1,65)=17.417$, $p<0.001$]: female subjects made significantly more use of semantics in generating female names than did male subjects.

Discussion

Across four name fluency tasks, a robust interaction between subject gender and name gender was observed: While female subjects recalled similar numbers of male and female names, male subjects recalled significantly more male names than female names. The absence of a mirror effect across genders is surprising: As a consequence of gender homophily in their social networks, female subjects should have more experience, on average, with female names, which should support fluency. Yet in all but one condition, female subjects actually produced more male names overall.

There were also clear differences in the strategies deployed in generating male and female names from memory. An analysis of the factors contributing to name production indicated that subjects relied more on semantics in generating male names and more on frequency in generating female names. This difference in strategy may be telling. In related work on verbal fluency, Johns et al. (2013) found that subjects who would go on to be diagnosed with mild cognitive impairment showed increased reliance on frequency cues in sampling the semantic space. Similarly, Taler et al. (2013) report that when bilinguals are forced to switch between languages, they reduce their use of semantic cues, favoring frequency instead. This suggests that when the task is made more difficult, or when the semantic space is degraded, speakers switch more towards a baseline sampling procedure that prioritizes frequency as the guiding cue.

Taken together, these findings lend support to the hypothesis that the distributional properties of female names make them harder to recall than male names. However, there is another possibility: Female subjects may simply have more *equal* experience with male and female names, given the overrepresentation of males in the broader cultural canon (e.g., educational materials and popular media). Experiment 3 was designed to assess and control for the effects of prior familiarity.

Experiment 3

If female names are less memorable than male names in American English, they should take longer to recognize in association with an individual, even when familiarity and name length are taken into account. In the first half of the experiment, our subjects were tested on their ability to recognize famous faces, which can be interpreted as a latent name generation task. Subjects viewed images of famous individuals intermixed with similarly posed and photographed non-famous distractors, and were asked to indicate whether they thought they recognized the individual or not. In the second half of the experiment, subjects viewed the same famous individuals, but this time, paired with either their actual name (hit) or a same-gender lure (false alarm). Subjects were asked to determine whether the names and faces were accurately matched or not. Across both tests, accuracy and response latency were recorded.

Participants

48 students (24 males and 24 females) at Indiana University participated for course credit in an undergraduate psychology course. All were native English speakers.

Materials

The famous names collected in Experiment 2, along with search volume data from Google Trends and overall Google hits, was used to create an inventory of a hundred famous individuals that college-age subjects were likely to recognize. **Names** The famous men and women whose names were accurately paired with their pictures (hits) were matched for name length and fame. This was also true for the men and

women whose faces were mismatched with novel names (lures). Novel names were generated using naming data from the last fifty years of the Social Security Index, and were matched on length, and first and last name popularity. Notably, it was not possible to precisely match either set of names for syllable length, although this was balanced to the extent possible. Popular American female names are longer, on average, and contain more syllables than male names.

Faces Photographs of each famous individual were representative forward-facing poses taken from the first page of Google Image search results, and were resized and cropped, such that the face of the individual took up a similar portion of the frame. Photographs of foils were taken from similarly posed and cropped professional portraits taken at the Museum of Modern Art in New York City.

Results

In the face recognition task, a 2 (subject gender) x 2 (face gender) repeated measures ANOVA revealed a significant interaction of subject gender and face gender for both accuracy [$F(1,46) = 19.58$, $p < .001$] and correct log RT [$F(1,46) = 6.59$, $p=.0135$]. Post hoc analyses (Tukey HSD) indicated that female subjects correctly identified more women than men ($p<.05$), and that male subjects not only identified more men than women ($p<.02$), but were also faster in those correct identifications ($p<.001$).

In the name recognition portion, a similar analysis revealed a significant interaction of subject gender and name gender on accuracy [$F(1,46) = 5.937$, $p=.019$], and both a main effect of name gender [$F(1,46) = 18.09$, $p<.001$] and a significant interaction of subject gender and name gender [$F(1,46) = 16.08$, $p<.001$] on correct log RT. Post hoc analyses revealed that female subjects correctly identified the names of more women than men ($p=.045$), and that male subjects not only correctly identified more men than women ($p=.012$), but again, were significantly faster in their identifications of them ($p<.0001$).

A more detailed analysis, in terms of hits and false alarms, revealed a parallel pattern of results, with the additional finding that in the name recognition task, both male *and* female subjects were significantly faster at correctly rejecting male lures.

Discussion

In both recognition tasks, classification accuracy patterned by gender, lending credence to the hypothesis that our male subjects have, on average, more exposure to men and male names, and our female subjects to women and female names. However, while male subjects were consistently faster in their identifications of males across both tasks; the inverse trend did not hold for our female subjects. More tellingly, female subjects actually proved faster at correctly rejecting *male* lures in the name recognition task, suggesting that they found these names easier to process.

General Discussion

Previous empirical work supports the notion that names are uniquely difficult to learn and remember compared to other semantic categories. For example, subjects find it harder to learn to pair faces with names, than to pair faces with occupations, even when the frequencies of names and occupations are held constant, and even when the same lexical items are used (e.g., 'baker' as both a surname and occupation; Cohen & Faulkner, 1986; McWeeny, Young, Hay, & Ellis, 1987). The experiments and analyses presented here offer insight into how the distributional structure of names—and other semantic categories, more broadly—can affect memory for those items across an array of tasks. In particular, selectively increasing entropy over a given element appears to disrupt memory not only for that element, but for the linguistic sequence it belongs to.

Acknowledgments

This research was funded by an NSF graduate fellowship to MD. We thank Richard Shiffrin and lab for comments and discussion.

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis. *Language and Speech*, *47*(1), 31–56.

Balota, D.A. et al. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.

Bell, A. et al. (2008). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Brennen, T. (1993). The difficulty with recalling people's names: the plausible phonology hypothesis. *Memory*, *1*(4), 409–431.

Burke, D. M. et al. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, *30*(5), 542–579.

Capitani, E., Laiacona, M., & Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex*, *35*(2), 273–278.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242.

Cohen, G. (1990). Why is it difficult to put names to faces? *British Journal of Psychology*, *81*(3), 287–297.

Cohen, G. & Burke, D.M. (1993). Memory for proper names: A review. *Memory*, 1, 249-263.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2015). A functional theory of gender paradigms. In F. Kiefer, J.P. Blevins, & H. Bartos (Eds.) *Morphological Paradigms and Functions*.

Ferrer i Cancho, R., & Sole, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788–791.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text (pp. 199–206). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Griffin, Z.M. (2010). Retrieving personal names, referring expressions, and terms of address. *The Psychology of Learning and Motivation*, 53, 345-387.

Johns, B.T. et al. (2013). Using cognitive models to investigate the temporal dynamics of semantic memory impairments in the development of Alzheimer's disease. In the *Proceedings of the 12th ICCM*.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory*.

McWeeny, K. H. et al. (1987). Putting names to faces. *British Journal of Psychology*, *78*(2), 143–149.

Mehta, C.M. & Strough, J. (2009). Sex segregation in friendships and normative contexts across the life span. *Developmental Review*, 29, 201-220.

Qian, T., & Jaeger, T. F. (2012). Cue Effectiveness in Communicatively Efficient Discourse Production. *Cognitive Science*, *36*(7), 1312–1336.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106.

Ramscar et al. (2013) The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems. *Proceedings of the 35th Meeting of the Cognitive Science Society*, Berlin, Germany.

Ramscar, M. et al. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425–440.

Taler, V. et al. (2013). A computational analysis of semantic structure in bilingual fluency. *Journal of Memory and Language*, 69, 607-618.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical Gender Is on the Tip of Italian Tongues. *Psychological Science*, 8(4), 314–317.

Zipf, G.K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley: Cambridge, MA.

What Do We Learn from Rating Metaphors?

Paul H. Thibodeau (paul.thibodeau@oberlin.edu)
Department of Psychology, Oberlin College
120 W. Lorain St; Oberlin, OH 44074 USA

Les Sikos (lsikos1@swarthmore.edu)
Frank H. Durgin (fdurgin1@swarthmore.edu)
Department of Psychology, Swarthmore College
500 College St; Swarthmore, PA 19081 USA

Abstract

What makes some metaphors easier to understand than others? Theoretical accounts of metaphor processing appeal to dimensions like *conventionality* and *aptness* to explain variability in metaphor comprehensibility. In a typical experiment, one group of naive participants rates a set of metaphoric sentences along these dimensions, while another is timed reading the same sentences. Then, the ratings are used to predict response times in order to identify the most relevant linguistic dimension for metaphor comprehension. However, surprisingly high correlations between ratings of theoretically orthogonal constructs and the results of an experiment in which a context manipulation affected ratings of metaphor *conventionality* and *aptness* suggest that these measures should be treated as dependent, rather than explanatory, variables. We discuss the implications of this perspective for theories of language processing.

Keywords: Metaphor, analogy, measurement, conventionality, language

Introduction

What makes some metaphors easier to understand than others? Theoretical accounts of metaphor processing appeal to dimensions like *conventionality* (e.g., Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997) and *aptness* (e.g., Glucksberg & Haught, 2006; Jones & Estes, 2006; Chiappe, Kennedy, & Chiappe, 2003) to explain variability in metaphor comprehensibility. In this context, *conventionality* reflects the familiarity of a metaphor; *aptness* reflects the degree to which a metaphor vehicle captures important features of the topic.

In a typical experiment, one group of naive participants rates a set of metaphoric sentences along these dimensions, while another is timed reading the same sentences. Then, the ratings are used to predict response times in order to identify the most relevant linguistic dimension for metaphor comprehension (e.g., Blank, 1988; Chiappe et al., 2003; Chiappe & Kennedy, 1999; Giora, 1997; Glucksberg, McGlone, & Manfredi, 1997; Jones & Estes, 2005, 2006).

However, surprisingly high correlations between ratings of theoretically orthogonal constructs raise questions about how well naive participants can actually operationalize these linguistic dimensions (Jones & Estes, 2006; Thibodeau & Durgin, 2011). For instance, Table 1 shows correlations between ratings of *familiarity*, *naturalness*, *imageability*, and *figurativeness* from data collected by Cardillo, Schmidt, Kranjec, and Chatterjee (2010), which are all significant at the $p < .001$ level, even though theoretical accounts of metaphor processing have argued that (at least some of) these constructs are orthogonal. A principal components analysis reveals that 70.3% of the variance across the four dimensions is captured by a single variable that loads positively on *familiarity*, *naturalness*, and *imageability*, and negatively on *figurativeness*, $ps < .001$. Similar patterns are found in existing datasets that seek to norm metaphoric stimuli for experimental work (e.g., Campbell & Raney, 2015; Cardillo et al., 2010; Katz, Paivio, Marschark, & Clark, 1988; Roncero & de Almeida, 2014).

Table 1: Correlations between ratings of familiarity, naturalness, imageability, and figurativeness from data collected by (Cardillo et al., 2010).

	Natural	Image	Figurative	PC
Familiarity	.936	.509	-.598	.931
Naturalness		.571	-.579	.942
Imageability			-.361	.706
Figurativeness				-.748

The goal of the present paper is to highlight this issue (i.e., what are we actually measuring when we ask people to rate sentences for *conventionality*, *aptness*, *metaphoricity*, etc?) by showing that ratings of these dimensions change as function of the context in which metaphoric sentences are presented. In the experiment, metaphoric target sentences were situated within one of three contexts: *matched*, *mixed*, or *literal* (see Table 2). In every case, participants read a brief description of a scenario, followed by a sentence that included a target metaphor. In some cases (the *matched* and *mixed* conditions), the initial description (i.e., *Context*) included metaphoric language, which was either consistent with the target metaphor (i.e., in the *matched* condition, the target metaphor extended a previously instantiated conventional metaphor) or inconsistent with the target metaphor (i.e., in the *mixed* condition, the target metaphor came from a different metaphor family than the metaphor used in the initial description of the scenario). In the *literal* condition, the target metaphor followed a non-metaphoric description of the scenario.

Table 2: Example base for generating stimulus items. Context A, B or C was presented with Extension A or B. Target words are indicated by capitalization.

Context
A. In big cities across America, crime has become an epidemic that can't be cured. It is beginning to infect small towns as well.
B. In big cities across America, crime has become a beast that is roaring out of control. It is beginning to prey on small towns as well.
C. In big cities across America, crime has become a problem that can't be solved. It is beginning to affect small towns as well.
Extension
A. There is no ANTIDOTE strong enough to cure it.
B. There is no CAGE strong enough to restrain it.

Experiment

Methods

Participants We recruited 1,200 people to participate in an online questionnaire through Amazon's Mechanical Turk. We restricted our sample to participants living in the United States who had a good performance record (a minimum 90% approval rating on previous tasks). We used Turk's exclusion capabilities (and internal tracking of IP addresses) to prevent people from participating in the task more than once. Upon completing the survey, participants were given a nine digit random number to paste back into the Mechanical Turk interface. Seven participants pasted an incorrect or incomplete completion code into Turk, leaving data from 1193 participants for analysis.

Stimuli The stimuli were modeled after those developed by Thibodeau and Durgin (2008, Exp. 3). One paired-item stimulus base is shown in Table 2. For each stimulus base, two different metaphoric mappings (e.g., A: CRIME IS AN INFECTIOUS DISEASE vs. B: CRIME IS A WILD ANIMAL) were used to generate conceptually parallel initial vignettes that contained two or more context sentences. A third conceptually parallel vignette (C: CRIME IS A PROBLEM) used non-metaphoric language to communicate the same basic situation.

Target sentences were designed to be interpretable following a description that included *matched* or *mixed* metaphors or non-metaphoric, *literal*, language. That is, the study was not designed to to compare across situations that involved arriving at entirely different semantic interpretations of the target sentence. However, we did expect the manipulation to modulate the processing fluency of the target metaphors. Specifically, we expected that the metaphors would be be processed more easily in the *matched* than *mixed* or *literal* contexts (Thibodeau & Durgin, 2008).

A total of 36 paired-item stimulus bases were developed. Three experimental conditions of two items each (a total of 72 items in each condition) were constructed from each stimulus base by pairing contexts and target sentences: *Matched Context* (Context A + Extension A; Context B + Extension B), *Mixed Context* (Context A + Extension B; Context B + Extension A), and *Literal Context* (Context C+ Extension A; Context C + Extension B).

All target metaphors were first instated nominally with a word that was sentence-medial (e.g., "There is no ANTIDOTE/CURE strong enough..."), and words leading up to the critical figurative noun were identical across sentences. These constraints were imposed in order to make the stimuli well-suited for investigation with methodologies such as ERP, since integration processes compete with metaphor comprehension processes when metaphor vehicles are presented at the ends of sentences (Friedman, Simson, Ritter, & Rapin, 1975; Osterhout, 1997). Note that each target word served as its own control because it appeared in all three contexts.

Survery Materials and Procedure Each participant read one item (i.e., context–target pairing) from each of the 36 stimulus bases described above, such that all participants rated 12 *matched*, 12 *mixed*, and 12 *literal* items. The order of items was pseudo-randomized such that no participant saw more than two items from a given condition in a row. Participants were asked to attend to the (uppercase) target word — the word which instantiated the metaphoric extension — when making their ratings.

In an effort to collect both contrastive and holistic ratings, we created two versions of the questionnaire. One asked participants to rate target words along four dimensions: *surprisingness*, *comprehensibility*, *conventionality*, and *metaphoricity*. The other asked participants to rate target words along the single dimension of *aptness* (see definitions of dimensions in Table 3). Following the instructions, participants saw an example question and had an opportunity to re-read the definitions of the dimensions. They then rated two filler items — a *matched* pairing and a *literal* pairing — before going on to rate the 36 experimental items.

Each item was presented, individually, above one or four 7-point Likert scales. Participants were prompted for ratings on each of the dimensions with a question. For example: "Surprisingness: how surprising was the metaphoric WORD as it is currently used?" The scale ranged from "very low" to "very high" for each dimension.

Analysis Principal components analysis (PCA) is a statistical procedure for revealing the internal structure of a dataset containing possibly correlated variables in a way that best ex-

Table 3: Definitions of dimensions as presented to participants in the instructions of the experiment.

Dimension	Definition
Surprisingness	Sometimes metaphors are used very naturally and do not seem surprising; in other cases, it can feel like they come out of nowhere. This is the dimension of 'surprisingness' that we'd like you to rate.
Comprehensibility	By this we simply mean, how easy or difficult it is to understand the statement. Some expressions are easier to understand, like "My mother is a saint," while others might be harder to understand "My mother is a paperclip."
Conventionality	Expressions can also vary in how conventional they are for conveying the idea that they are supposed to communicate. For example, consider the following two descriptions of a person running fast: a conventional or common way, 'he was running like the wind,' and a much less conventional way, 'he was running like a Porsche on a German highway.'
Metaphoricity	Finally, even metaphoric expressions may vary in how much they are figurative expressions rather than literal. For example, if I say that the sun is a ball, someone might interpret that as being literally true if they thought that 'ball' meant any object that was spherically shaped. Another person might think the expression quite metaphorical because balls are toys, and the sun is not a toy. You will be asked to judge the metaphoricity of the expressions below by indicating whether you find them to be very metaphorical or close to literal.
Aptness	Metaphors can vary in the extent to which they capture important features of the topic being described. This is the dimension of 'aptness.' For instance, it would be apt to express that someone is a fast runner by saying they are a rocket, but less apt to express that someone is a fast runner by saying they are an astronaut.

plains the common variance of the data (Dunteman, 1989). A primary goal of the present analysis was to identify and characterize the principal components (PCs) of the ratings, while checking for correlations among the various dimensions.

We test for effects of the context manipulation by fitting separate, by-item, repeated measures ANVOAs for each PC (separate models since the principal components analysis yields orthogonal components).

Results and Discussion

We found that extracting two PCs explained 92.8% of the variance across the five dimensions. The first component explained 74.1% of the variance and loaded highly on four of the dimensions: *comprehensibility*, *conventionality*, *aptness*, and negatively on *surprisingness*; this component loaded less heavily on *metaphoricity* (see Table 4). As expected, these dimensions seem to converge on the notion of processing fluency (ease) with respect to metaphor interpretation. As a result we call this component *Processing Fluency*.

The second component explained 18.6% of the variance across the five dimensions (71.8% of the variance that remained after accounting for the first component), and loaded most strongly on *metaphoricity*. As a result we call this second PC *Figurativeness*. Correlations between the individual rated dimensions and PCs are presented in Table 4. Note that ratings of *aptness* (collected separately) and *comprehensibility* (collected together with other dimensions) correlated exclusively with *Processing Fluency*, whereas *conventionality* and *surprisingness* ratings contributed to *Figurativeness* as well.

Figure 1 shows mean *Processing Fluency* and *Figurativeness* as a function of *Context* (*matched*, *mixed*, or *literal*). An ANOVA by items indicated that *Processing Fluency* was reliably affected by *Context*, $F(2,142) = 133.759$, $p < .001$, $\eta^2 = 0.653$. Paired comparisons further revealed that *Processing Fluency* was higher in the *matched* context than in either the *literal*, $t(71) = 13.519$, $p < .001$, or the *mixed* context, $t(71) = 12.941$, $p < .001$, and that *Processing Fluency* did not differ between the *literal* and *mixed* contexts, $t(71) = 0.995$, $p = .323$.

Table 4: Correlations among rated dimensions and the first two (orthogonal) principal components of the ratings data. All correlations significant at the $p < .001$ level except those noted: $^{*}p < .05$ and $^{\diamond}p > .05$.

.

	Surp	Meta	Conv	Compr	Apt
PC1: Ease	-.919	.436	.933	.969	.928
PC2: Fig	.311	.893	-.187*	.021$^{\diamond}$	.055$^{\diamond}$
Aptness	-.821	.422	.766	.883	
Comprehens	-.848	.428	.897		
Conventional	-.902	.260			
Meta'icity	-.142*				

These findings suggest that a supportive discourse context increased participants' metacognitive sense of processing fluency of the required metaphoric mapping. In principle, such benefits could be argued to emerge from lexical priming alone (e.g., McGlone, 2011). However, Thibodeau and Durgin (2011) found that lexical priming of the metaphor vehicle was

insufficient to produce shifts in ratings of *aptness*, whereas priming the specific metaphor mapping was both necessary and sufficient. For instance, people read a target sentence like "*Education is a lantern*" faster after they had read sentences like "*A mentor is a lantern*" (i.e., when the metaphor vehicle conveyed a similar meaning) but not after they had read sentences like "*A camp light is a lantern*" (i.e., when the metaphor vehicle was used non-metaphorically) or after they had read sentences like "*A flag is a lantern*" (i.e., when the metaphor vehicle conveyed a different meaning). Therefore, a supportive context appears to prime the relationship between the relevant figurative and literal semantic domains, rather than merely the lexical item *per se*. Because the ratings concern metaphor quality (rather than response latency), they show that interpretation of the metaphor (the mapping) is indeed facilitated by the context manipulation.

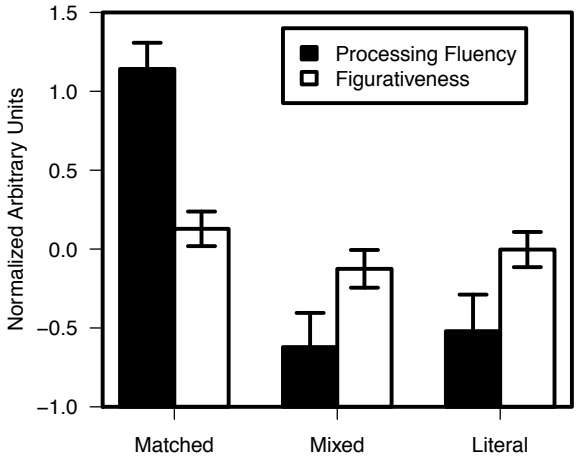


Figure 1: The mean values (across 72 items) of the first and second principal components of the rating data as a function of Context. Standard error bars are shown.

The *Figurativeness* principal component was also reliably affected by *Context*, $F(2,142) = 7.536$, $p < .001$, $\eta^2 = 0.096$. Paired comparisons indicated that *Figurativeness* was judged higher in the *matched* context than in the *mixed* context, $t(71) = 3.838$, $p < .001$. There was some evidence that *Figurativeness* estimates in the *literal* context were lower than those in the *matched* context, $t(71) = 2.099$, $p = .039$, and higher than those in the *mixed* context, $t(71) = 1.818$, $p = .073$.

However, as is evident in Figure 1, the effects of *Context* on *Figurativeness* (effect size for *matched* vs. *mixed* = 0.45) were small compared to the effects of the *Context* on enhancing *Processing Fluency* (effect size for *matched* vs. *mixed* = 1.52). Thus the *Figurativeness* component of participants' ratings may reflect a comparison between the implied meaning of the target word and its literal sense (above and beyond the effect of *Processing Fluency*, *per se*). This comparison may nonetheless be facilitated if the intended metaphoric sense of the target word is made clearer by a *matched* context.

It is important to distinguish between the measure of *Figurativeness* yielded from the principal components analysis and direct ratings of *metaphoricity*. Participants' ratings of *metaphoricity* were affected more strongly by the context manipulation, $F(2,142) = 32.483$, $p < .001$, $\eta^2 = 0.314$, compared to the principal component ($\eta^2 = 0.096$).[1] This is because ratings of *metaphoricity* were correlated with ratings of *aptness*, *surprisingness*, *comprehensibility*, and *conventionality*, and thus loaded onto *Processing Fluency*, $r(71) = .436$, $p < .001$.

In the measure of *Figurativeness* computed from the principal components analysis, on the other hand, variance in ratings of *metaphoricity* that was consistent with the other four dimensions was partialed into the measure of *Processing Fluency*, leaving the *Figurativeness* principal component to account for variance that was mostly unique to an aspect of rated *metaphoricity* (and to a lesser extent rated *surprisingness* and *conventionality*).

General Discussion

In this experiment we showed that asking people to rate metaphors for five distinct qualities produced two orthogonal principal components. The first, which we labeled *Processing Fluency*, seems to represent the ease with which a metaphor can be interpreted (another appropriate label for this dimension might be: *Ease of Interpretation*). The second, which we called *Figurativeness*, seems to represent the extent to which a word's meaning is perceived as clearly figurative rather than literal, once *Processing Fluency* is taken into account. Manipulating the context for a target metaphor had a large impact on *Processing Fluency*, but only a slight impact on perceived *Figurativeness*. That is, extended metaphors were judged by raters to be similarly metaphoric (or not) regardless of context. However, these metaphors were processed more fluently when they were presented in the context of a consistent metaphoric mapping.

One important implication of these findings relates to the explanatory power of off-line ratings of metaphor processing for theories of metaphor comprehension. One goal of language researchers has been to identify a linguistic dimension that explains why some metaphors are easier to process than others. *Aptness* has been highlighted by researchers who argue that metaphors are processed as class inclusions statements (e.g., Glucksberg & Haught, 2006; Jones & Estes, 2006; Chiappe et al., 2003), rather than through a comparison mechanism (e.g., Blank, 1988; Bowdle & Gentner, 2005; Giora, 1997). Although it very well may be the case that such a dimension can be measured and that it can explain, *a priori*, variability in metaphor processing fluency, our results suggest that operationalizing the construct by gathering ratings from

[1] Although pairwise comparisons reveal a similar pattern of results for the direct ratings and the principal component: significant differences between the *matched* and both *mixed*, $t(71) = 7.712$, $p < .001$, and *literal* contexts, $t(71) = 5.961$, $p < .001$; a trending but non-significant difference between the *mixed* and *literal* contexts, $t(71) = 1.711$, $p = .091$.

naive participants is fundamentally flawed.

In contrast to how these ratings are typically used (i.e., as operationalizations of independent, predictor, variables), our work suggests that these dimensions may be more appropriately considered as an indirect measure of processing fluency (Thibodeau & Durgin, 2011). When people are asked to rate sentences for abstract qualities like *conventionality* and *aptness*, they misattribute how easily they processed the sentence for the dimension they are being asked to rate (Alter & Oppenheimer, 2009; Jacoby & Whitehouse, 1989; Jacoby, Allan, Collins, & Larwill, 1988; Kahneman, 2011).

This is problematic because processing fluency is supposedly what we are trying to explain by gathering ratings of the linguistic dimensions in the first place. In other words, using subjective ratings of *conventionality* and *aptness* to predict how quickly people process metaphoric sentences entails showing that an "off-line" measure of processing fluency is related to an "on-line" measure of processing fluency. As a result, it is unclear what we can learn about the mechanisms that support language processing from such experiments.

To address this issue, we suggest designing experiments that actively manipulate the relevant linguistic dimensions or operationalizing the constructs with more objective methods like corpus analysis. For instance, one way to manipulate the familiarity of a metaphor "on-line" is to repeatedly expose people to similar metaphoric expressions (e.g., "A figure skater is a butterfly"; "A ballerina is a butterfly"; Bowdle & Gentner, 2005; Thibodeau & Durgin, 2011; Chettih, Durgin, & Grodner, 2012). One way to measure a construct like *aptness* "off-line" is to use corpus-based metrics like Latent Semantic Analysis (Kintsch, 2000; Kintsch & Bowles, 2002).

We also see value in the use of statistical procedures like PCA for researchers interested in testing questions about how metaphors are processed. On this approach, ratings of dimensions like *conventionality* and *aptness* can be viewed as converging on a notion of *Processing Fluency*. Just as personality instruments and attitudinal surveys often include complementary items that facilitate reliable measurement of personality and attitudinal constructs, asking people to rate various linguistic dimensions of metaphors can elicit more reliable estimates of *Processing Fluency*. The results of the current experiment (as well as analyses of existing data sets like those of Cardillo et al., 2010) suggest that people can reliably differentiate their sense of a sentence's metaphoricity (figurativeness) from a dimension like *Processing Fluency* – particularly when a procedure like PCA is used to make these dimensions orthogonal (i.e., to partial variance in ratings of metaphoricity that seem to be affected by *Processing Fluency* into a measure of *Processing Fluency*). On this approach, one can more confidently use subjective ratings in response time and imaging studies to test mechanistic questions about figurative language processing.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and social psychology review*.

Blank, G. D. (1988). Metaphors in the lexicon. *Metaphor and Symbol*, *3*(3), 21–36.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological review*, *112*(1), 193.

Campbell, S. J., & Raney, G. E. (2015). A 25-year replication of katz et al.'s (1988) metaphor norms. *Behavior research methods*, 1–11.

Cardillo, E. R., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior research methods*, *42*(3), 651–664.

Chettih, S., Durgin, F. H., & Grodner, D. J. (2012). Mixing metaphors in the cerebral hemispheres: What happens when careers collide? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 295.

Chiappe, D. L., & Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, *6*(4), 668–676.

Chiappe, D. L., Kennedy, J. M., & Chiappe, P. (2003). Aptness is more important than comprehensibility in preference for metaphors and similes. *Poetics*, *31*(1), 51–68.

Dunteman, G. H. (1989). *Principal components analysis* (No. 69). Sage.

Friedman, D., Simson, R., Ritter, W., & Rapin, I. (1975). Cortical evoked potentials elicited by real speech words and human sounds. *Electroencephalography and clinical Neurophysiology*, *38*(1), 13–19.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive linguistics*, *8*, 183–206.

Glucksberg, S., & Haught, C. (2006). Can florida become like the next florida? when metaphoric comparisons fail. *Psychological Science*, *17*(11), 935–938.

Glucksberg, S., McGlone, M. S., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of memory and language*, *36*(1), 50–67.

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 240.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*(2), 126.

Jones, L. L., & Estes, Z. (2005). Metaphor comprehension as attributive categorization. *Journal of Memory and Language*, *53*(1), 110–124.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988).

Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol*, *3*(4), 191–214.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*(2), 257–266.

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, *17*(4), 249–262.

McGlone, M. S. (2011). Hyperbole, homunculi, and hindsight bias: An alternative evaluation of conceptual metaphor theory. *Discourse Processes*, *48*(8), 563–574.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and language*, *59*(3), 494–522.

Roncero, C., & de Almeida, R. G. (2014). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior research methods*, 1–13.

Thibodeau, P. H., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language*, *58*(2), 521–540.

Thibodeau, P. H., & Durgin, F. H. (2011). Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, *26*(3), 206–226.

Listener sensitivity to foreign-accented speech with grammatical errors

Yuki Asano & Andrea Weber
(yuki.asano; andrea.weber@uni-tuebingen.de)
English Department, Eberhard Karls University Tübingen, Germany

Abstract

The present accent rating study investigates the interaction between accent strength and grammatical correctness on perceived accentedness. German native (L1) listeners rated German sentences produced by L1 and non-native (L2) speakers. Sentences either contained a grammatical error or were grammatically correct. Results showed that grammatical correctness affected the accent rating of sentences produced by L1 speakers, but not of those by L2 speakers. The inverse influence of grammatical errors on sentences spoken with stronger accents suggests that phonological information plays a more important role for global perception of speech accentedness than grammatical correctness does, revealing a hierarchical importance of factors that form an L2 accent. This finding is in line with recent findings from an online processing ERP study (Hanulíková, van Alphen, van Goch, & Weber, 2012) in which L1 listeners were tolerant towards grammatical errors made by L2 speakers, i.e. showed no P600 effect for grammatically incorrect sentences.

Keywords: perceived accent strength; grammatical error

Introduction

Spoken language is notoriously variable. The auditory signal produced by a speaker can vary due to differences in voice, stress, intonation, and speaking rate, among other factors. Additionally, phonological processes of assimilation or deletion, as well as dialects and foreign accents, can lead to deviations in form from the standard of a target language. Yet, despite these deviations, listeners usually understand spoken language quite easily, suggesting that the human speech processing system is flexible and able to adapt to varying input quickly. Previous studies have shown that listeners are, for example, able to adapt to foreign-accented speech within a few minutes (Bradlow & Bent, 2008; Weber & Pöllmann, 2010). It has also been found that listeners adjust their interpretation of the speech signal so that it is in line with information about the identity of the speaker. As a consequence, sentences like "I am pregnant." are perceived to be anomalous when produced by a male speaker, because listeners are aware of semantic and pragmatic constraints which restrict the use of the adjective pregnant to females (Van Berkum, van der Brink, Tesnik, Kos, & Hagoort, 2008).

In the vein of speaker-specific adjustments in comprehension, Hanulíková et al. (2012) found in a recent ERP study that listeners are surprisingly tolerant towards grammatical errors in foreign-accented speech. When a Turkish L2 speaker of Dutch produced sentences containing a gender-agreement error, L1 Dutch listeners did not show the P600 effect typically observed for this type of syntactic error (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). When the same sentences were produced by an L1 speaker of Dutch, however, there was a clear P600 effect visible in the ERP response of the Dutch listeners. It is possible that Dutch listeners were so habituated to gender-agreement errors in Turkish-accented Dutch that the brain no longer marked them as errors. Interestingly, a comparable and very fast habituation was also visible for accent-free sentences in the experiment. When the study was split into two blocks, the P600 effect observed for L1 speech vanished in the second half of the experiment. This suggests that listeners adjust their model regarding the occurrence of errors even in L1 speech. This finding was in accordance with previous studies (e.g., Hahne & Friederici, 1999). Furthermore, the P600 is sometimes interpreted as an indicator of unexpectedness (e.g., Gouvea, Philipps, Kazanina, & Poeppel, 2010). In this sense, the P600 effect found for L1 speech with grammatical errors can also be interpreted as a surprisal effect.

The question that arises from the study of Hanulíková et al. (2012) is whether grammatical errors have an effect on perceived accent strength, even though they do not disrupt online processing of foreign-accented speech. Previous studies on accentedness mainly employed rating tasks that either focused on effects of speaker background or phonological deviations. One of the decisive factors for speaker background is, for example, age of acquisition (AOA). Flege, Munro, and Mackay (1995) found in a rating experiment with Italian L2 speakers of English that speakers who learned their L2 after the age of 15 hardly ever receive ratings that fall within the L1 English range. Another factor that has been found to affect a speaker's accent in L2 is length of residence (LOR), where LOR means the amount of time spent in a country in which the speaker's L2 is the dominant language. Flege and Fletcher (1992) found that LOR has effects on degree of foreign accent, although this seems to be a slow process and significant effects only showed after roughly 14 years of residence. Adding to that, gender also affects the degree of perceived accentedness. Flege et al. (1995) found that female subjects who did not learn English as their L1, but started learning it as L2 as children received higher pronunciation scores than their male peers. However, in late adolescence this picture reverses, and male subjects begin to receive higher scores when AOA increases. It is unclear, however, whether this is motivated by biological or social factors (Edwards, 2008). A second main strand in previous foreign accent rating studies investigated phonological deviations. For example, Magen (1998) showed that L1 English speakers are sensitive to syllable structure factors, consonant manner, and lexical and phrasal stress in utterances produced by L2 speakers of English with a Spanish background. More generally, it can be said that previous studies have shown a high sensitivity of L1 listeners towards non-nativeness in speech. Non-nativeness is noticed reliably in very short stretches of speech (e.g., the

release burst of stop consonants, see Flege, 1984) and even in a strongly distorted speech signal (e.g., Munro, Derwing, & Burgess, 2010).

L2 speakers typically deviate not only in the phonological form but also in grammar from the target language. How far grammatical correctness of sentences produced by L2 speakers contributes to overall perceived accentedness is still an open question. L1 speech usually manifests itself with an absence of grammatical errors, as is reflected in the strong P600 effect when errors do occur (Hanulíková et al., 2012); in foreign-accented speech, however, grammatical errors, in addition to myriad phonological and phonetic cues can be considered yet another indicator of a speaker's non-nativeness. The current study investigates this matter in a German accent rating study by incorporating grammatical errors into sentences spoken by L1 and L2 speakers of German. The L2 speakers had different L1 backgrounds in order to add variation to the study; the authors judged the L2 speakers to all have noticeable accents in German but with varying accent strength. Following the results of Hanulíková et al. (2012), two outcomes are possible: First, grammatical errors might not influence accent ratings because listeners are familiar with them, and they do not noticeably disrupt online processing. Second, it is also possible that grammatical errors do influence accent ratings, as we know that L1 listeners can reliably detect grammatical errors in foreign-accented speech reliably when asked to do so (Hanulíková et al., 2012) and are generally very sensitive towards nonnativeness in speech (Flege, 1984). Hence, it could be assumed that grammatical errors count as yet another cue towards the perceived foreignness of the speaker.

Accent Rating

Method

Participants Thirty-five monolingual L1 speakers of German (mean age = 24.4; 23 = female, 12 = male) took part in the study.

Materials Fifteen German sentences with varying syntactic structures were constructed as materials. Care was taken to avoid low frequency words that L2 speakers of German might not know. Furthermore, the entire breadth of the German sound inventory was represented in the sentences. Each sentence was prepared both with and without grammatical errors. For the incorrect versions, we used three different kinds of grammatical errors that were conside to be typical of various L2 speakers of German (see also Schwartz & Sprouse, 1994). Three typical error types in L2 German (Heringer, 2001) were chosen to disguise the grammatical error manipulation. The error types were preposition errors, case errors, and verb-agreement errors. Five of the sentences contained a preposition error of the kind given in (1a/b), where (1a) is correct and (1b) incorrect (the relevant preposition is marked in italics). The verb *kommen* 'to come' in combination with *Hause* 'home' requires the preposition *nach* 'to'.

(1a) Du musst schnell *nach* Hause kommen.
(1b) *Du musst schnell *zu* Hause kommen.
(Engl. You need to come home quickly.)

Another five sentences contained a case-error that could either be realized on the determiner or on an adjective preceding a noun, because both determiners and attributive adjectives are inflected for case in German. An example can be found in (2a/b) (a = correct, b = incorrect). The German preposition *hinter* 'behind' selects a dative NP that is marked with the dative definite article; in the case of a masculine noun like *Schuppen* 'shed', the correct form is *dem* (2a).

(2a) Hinter *dem* Schuppen steht ein altes Fahrrad.
(2b) *Hinter *der* Schuppen steht ein altes Fahrrad.
(Engl. Behind the shed, there is an old bicycle.)

Five additional sentences contained an agreement error between the main verb and the subject. An example can be found in (3a/b). Agreement takes place between the subject *du* 'you' (second person singular) and the main verb *spielen* 'to play'. The correct verb form for second person singular is *spielst*. The verb in (3b) is, on the other hand, is ambiguous between first person plural, third person plural, and the infinitive.

(3a) Du *spielst* sehr gut Fußball.
(3b) *Du *spielen* sehr gut Fußball.
(Engl. You play football very well.)

All sentences were recorded with a mobile Zoom H2N recorder in a quiet room in .wav format (24Bit/96kHz). There were two speaker groups. The L1 speaker group consisted of five female students of the University of Würzburg (Germany) whose L1 was German (age range 23-26, mean age = 24.2). The L2 speaker group consisted of five female speakers with a mixed L1 language background (Polish (1x), Kirghiz (1x), Russian (1x), Persian (2x)) of moderate to high proficiency in German with a noticeable accent in their pronunciation. At the time of recording, all L2 speakers studied at the University of Tübingen (in Germany) in various degree programs. All 10 speakers produced each of the 15 sentences in both the correct and incorrect versions (i.e., 30 sentence recordings per speaker). Speakers were asked to read all of the sentences carefully before recording them in order to avoid hesitations and uncertainties. Speakers always recorded the correct and incorrect versions of each sentence in consecutive trials, and the examiner paid attention that both versions were, apart form the grammatical error, comparable in pronunciation and fluency. Multiple tokens of each sentence were recorded, and the best tokens (i.e., recordings with no hesitations or unintentional grammatical errors) were chosen

for presentation in the study. The chosen sentences were normalized for loudness using Samplitude Music Studio 2013.

Procedure The experiment was carried out as an online rating study using the OnExp software (`http://www.lingexp.uni-tuebingen.de /OnExp/`). A link to the experiment was emailed to the participants, all of who were naive to the purpose of the study. Participants first had to indicate their age and language background (including L1 and dialects) in a short questionnaire. Then, they were informed that they would hear short sentences spoken by L1 and L2 speakers of German and that their task would be to rate these sentences for degree of foreign accent. The instructions did not specify that some of the sentences contained grammatical errors. Participants were asked to wear headphones to minimize background noise and distraction during the experiment. They were told that they would have to rate "accentedness" on a 9-point scale where 1 designates "no accent" and 9 "strong foreign accent" (see Southwood & Flege, 1999, for a discussion of rating scales in accent studies). The instructions and the explanation of the rating scale were visible throughout the experiment to prevent subjects from accidentally mixing up the two ends of the scale. The instructions also indicated that they should listen to each sound file only once. Participants were randomly assigned to one of five experimental lists. Each list contained 60 sentences, cross-balanced in a Latin-Square Design for *sentence* (N = 15), *language background of the speaker* (L1 and L2), and *grammatical correctness* (correct and error).

Results

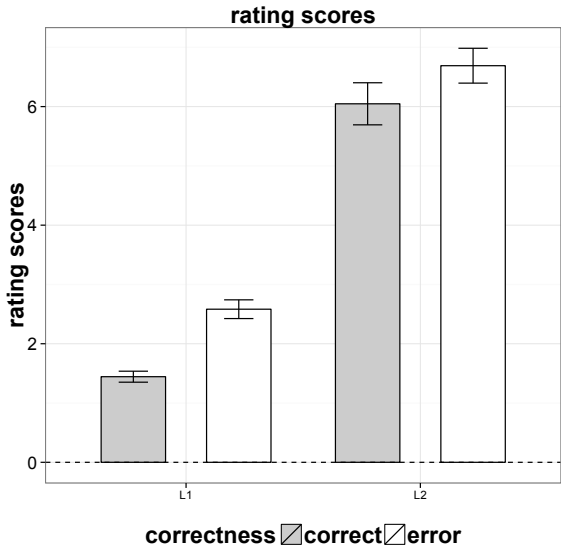
In total, 2100 data points were recorded (35 participants x 60 trials). From these, 5 data points had to be discarded due to the lack of a response. Average rating scores in each *speaker group*, *grammatical correctness* and *type or error* are shown in Table 1.

Table 1: Mean rating scores and standard deviations for each condition.

speaker group	type of error	grammatical correctness	
		correct	error
L1	preposition	1.57 (0.89)	2.81 (1.90)
	case	1.37 (0.81)	2.46 (1.60)
	agreement	1.40 (0.79)	2.43 (1.79)
L2	preposition	5.99 (2.18)	6.69 (2.04)
	case	6.15 (2.11)	6.77 (1.78)
	agreement	6.24 (2.11)	6.63 (2.10)

In the following analyses, we provide statistical results from linear mixed-effects regression (LMER) models providing p-values that are based on null-hypothesis significance testing, as well as from descriptive statistics providing 95% CI error bars as a complementary data analysis, since the use of a statistical significance referring only to p-values has recently been critically discussed due to the low reliability of the obtained p-values (Cohen, 1994; Cumming, 2011, 2013; Loftus, 1993). Cumming (2013) showed *a dance of the p-values* (p. 6) to point out an enormous variation in p-values from less than .001 to 0.75 when replicating an experiment 25 times with two independent groups, each group having an N of 32. CI error bars are informative because they indicate the possible variations of p-values shown in the dance of the p-values, while a single p-value gives virtually no information about the variation among the infinite possible p-values. The LMER-analysis included *accent rating* as dependent measure, *speaker group* (L1 vs. L2), *grammatical correctness* (correct vs. error) and *type of error* (preposition vs. case vs. agreement) as fixed factors and *participant* and *sentence* as random factors including random slopes (Cumming, 2013; Barr, Levy, Scheepers, & Tily, 2013). The analysis showed a significant main effect of *speaker group* (rating scores for L2 speakers were higher than those for L1 speakers, ß= 4.61, SE = 0.10, t = 45.6, $p < 0.001$) and of *grammatical correctness* (rating scores for grammatically incorrect sentences were higher than those for correct sentences, ß= 1.15, SE = 0.10, t = 11.3, $p < 0.001$). Moreover, there was an interaction between *speaker group* and *grammatical correctness* (the difference between the scores of correct and incorrect sentences were greater for L1 speakers than that for L2 speakers, ß= 0.51, SE = 0.14, t = 3.6, $p < 0.001$). No effect of *type of error* was found, see Figure 1.

Figure 1: Raw rating scores and 95% CI bars by L1 speakers (left) and by L2 speakers (right). Within-subject CIs were calculated with `summarySEwithin` function in R-package `Rmisc`.



The box plots confirm the interaction found in the LMER-analysis. Additionally, they illustrate that the rating scores of L2 speakers exhibit a larger range than those of L1 speakers.

Importantly, however, we did not find a ceiling effect in the L2 speaker data. Both in the *correct* and *error* conditions, the upper CIs did not reach the maximum score (=9).

Next, in order to corroborate the interaction found in the LMER-analysis, the rating scores were normalized by subtracting the average scores of grammatically incorrect sentences from those of correct sentences for each speaker, see Figure 2.

Figure 2: Mean normalized rating scores (*correct - error*) and 95% CI bars by L1 speakers (left) and by L2 speakers (right). Within-subject CIs were calculated with `summarySEwithin` function in R-package `Rmisc`.

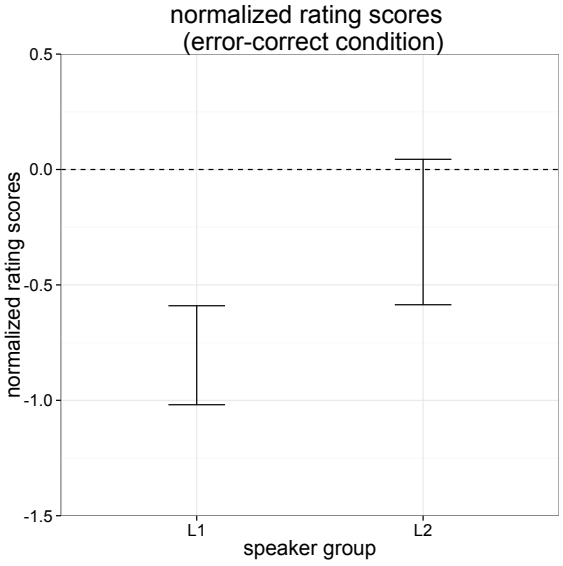


Figure 2 shows that only the CI bars of L1 speakers were placed in the negative area, suggesting that the accent rating scores for grammatically incorrect sentences were higher than those for grammatically correct sentences. The CI bars of L2 speakers crossed 0 and did not overlap with those of L1 speakers, suggesting that the rating scores for grammatically correct and incorrect sentences did not differ from each other and that the normalized scores of L1 and L2 speakers differed significantly from each other. The results from the LMER-analysis and from the CI bars were thus congruent.

Discussion

The results of the present rating study add to our understanding of perceived accent strength. Sentences with grammatical errors were rated as more strongly accented than the same sentences without grammatical errors, but this was true only when the sentences were spoken by L1 speakers. Thus, in the presence of grammatical errors, L1 speech was perceived as more accented. When sentences were spoken by L2 speakers, however, additional grammatical errors did not significantly affect the perceived accent strength. This suggests that phonological and phonetic information is more decisive for perceived accentedness than grammatical correctness is. It is important to note that we did not find a ceiling effect in the L2 speaker data. Therefore, the lack of the *grammatical correctness* effect on L2 speakers' scores cannot be explained by claiming that L2 speakers would have produced grammatically correct sentences with accents so strong that participants would have not been able to give higher scores. Our finding is in line with recent findings from an ERP study (Hanulíková et al., 2012) in which L1 listeners were tolerant towards grammatical errors made by L2 speakers, i.e. showed no P600 effect for grammatically incorrect sentences. We see at least three possible explanations for this interaction between accent strength and grammatical correctness. First, it is possible that in strongly accented sentences, the grammatical error was pronounced less clearly, and therefore the error was less strongly noted, (see, however, Hanulíková et al., 2012). Second, listening to foreign accented speech may require higher demand on attention control (Baddeley & Hitch, 1974; Bialystock, 1992) than listening to L1 speech. This attention mechanism controls the limited cognitive resources in all forms of information processing through shifting efficient attention between foregrounding and backgrounding of task-relevant and -irrelevant information (Isaacs & Trofimovich, 2011; Rosen & Engle, 1998). While paying attention to phonologically deviant forms in L2 speech, listeners' attention capacity for grammatical errors is possibly reduced. In this case, the result suggests that phonological information is perceptually more dominant in global rating of foreign accents than grammatical information is. This knowledge contributes both to empirically defining a hierarchical importance of factors that form foreign accented speech as well as to revealing which of the factors is most dominant in the global perception of accents. However, it is questionable whether an attention effect would emerge in our experimental setting alternating nativeness of the speakers.

Alternatively, experience with grammatical errors in L2 speech led to a reduced surprisal effect and hence weaker adverse influence on the ratings. Based on the mechanism of statistical learning (e.g. Romberg & Saffran, 2010), exposure to foreign accented speech accommodated the listener's sensitivity to processing grammatical errors in L2 speech. This accommodation is advantageous in processing a specific kind of speech effectively (e.g. Eisner & McQueen, 2005; Kraljic & Samuel, 2005; Norris & Cutler, 2003). In order to support this claim, we need to test whether listeners with little experience with the foreign-accented speech or grammatical errors atypical of L2 speakers would fail to show the effect found in the present study. Importantly, our speakers had different L1s, so accent-specific perceptual accommodation cannot explain our result.

Note that in our study we only used three types of grammatical errors that did not necessarily affect the sentence meaning, i.e. sentence meaning could still be inferred. The three types of errors were comparable in their results. We believe that in most cases it was still possible to correctly

interpret the message of the incorrect sentences. However, grammatical errors can also affect the interpretability of a sentence more strongly. If a speaker, for example, uses the wrong gender in an anaphoric pronoun, this would make processing for this sentence more difficult because the correct referent of the pronoun can no longer be assigned as a result of the grammatical error. It is of further interest to investigate the relationship between the impairment of sentence meaning and accent rating performance. Further research in this area could focus on perception differences between other possible grammatical errors. This was not possible in the study presented in this article because of the small number of items that were used to keep the length of the experiment manageable. It could, however, be the case that frequent errors are handled differently by L1 listeners than errors that occur relatively infrequently in L2 speech. Hahne and Friederici (1999) have shown in an EEG-study that listeners seem to adjust their response to errors in L1 as a function of error probability. It would be interesting to see whether this carries over into the perceived accentedness of speech produced with a foreign accent. In this respect, corpora could help determine frequently produced grammatical errors in foreign-accented speech for a certain well-known accent, and these errors could then be compared to errors that are demonstrably infrequent. Apart from overall error frequency, further research could also investigate error frequency within a sentence. Is the reaction of L1 listeners when there is only one grammatical error in a sentence different from a sentence with more than one error? Hence, is there a point at which L1 listeners just put an L2 speaker into a category that tells the parsing system to ignore grammatical errors for the sake of semantic content? Previous research suggests that L1 listeners may develop a focus on meaning in L2 speech (Galloway, 1980), which makes it interesting to see whether there is a threshold for error occurrence after which additional errors are simply ignored by the parser.

Conclusions

The present accent rating study revealed an interaction between accent strength and grammatical correctness which had not been investigated in previous studies. Only in ratings for sentences spoken by L1 speakers we found an effect of grammatical errors. When it comes to L2 speech, which was phonologically accented, grammatical errors did not influence the global perception of accentedness. Thus, in L2 speech, phonological information was more decisive than grammatical information for global perception of accentedness. We discussed this result as a piece of empirical evidence that 1) L1 listeners paid attention to a perceptually more decisive factor than to a less decisive one because the listeners' capacity of attention control is limited in online speech perception and that 2) listeners accommodated their perception to a specific kind of speech in order to effectively process it.

Acknowledgements

We would like to thank Lydia Kiessling and Daniel Baumgartner who helped to design and to prepare the study, including its presentation.

References

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–90). London: Academic Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bialystock, E. (1992). Selective attention in cognitive processing: The bilingual edge. , *83*, 501–513.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaption to non-native speech. *Cognition*, *106*, 707–729.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2013). The new statistics: Why and how [10.1177/0956797613504966]. *Psychological Science*.

Edwards, J. G. H. (2008). Social factors and variation in production in L2 phonology. In (pp. 251–279). John Benjamins.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics*, *67*, 224–238.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of Acoustical Society of America*, *76*, 692–707.

Flege, J. E., & Fletcher, J. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of Acoustical Society of America*, *91*, 307–389.

Flege, J. E., Munro, M. J., & Mackay, I. R. A. (1995). Factors affecting strength of the perceived foreign accent in second language. *Journal of Acoustical Society of America*, *97*, 3125–3134.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *The modern language journal*, *64*, 428–433.

Gouvea, A. C., Philipps, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. In *Language and cognitive processes* (Vol. 25, pp. 149–188).

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*, 194–205.

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, *24*, 878–887.

Heringer, H. J. (2001). *Fehlerlexikon - Deutsch als Fremdsprache: Aus Fehlern lernen: Beispiele und Diagnosen*. Cornelsen Verlag.

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*, 113–140.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 250–256.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*, 381–400.

Munro, M. J., Derwing, T. M., & Burgess, C. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626–637.

Norris, M. J. M., D., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204-238.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley interdisciplinary reviews. Cognitive science*, *1*(6), 906–914.

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language*, *39*, 418–436.

Schwartz, B. D., & Sprouse, R. A. (1994). Word order and nominative case in nonnative language acquisition: a longitudinal study of (L1 Turkish) German interlanguage. In T. Hoekstra & B. D. Schwartz (Eds.), *Language acquisition studies in generative grammar* (pp. 317–68). Amsterdam: John Benjamins.

Southwood, M. H., & Flege, J. E. (1999). Scaling foreign accent: direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, *13*, 335–349.

Van Berkum, J. J., Brown, C., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 443–467.

Van Berkum, J. J., van der Brink, D., Tesnik, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*(580–591).

Weber, A., & Pöllmann, K. (2010). Identifying foreign speakers with an unfamiliar accent or in an unfamiliar language. In *New sounds 2010: The 6th international symposium on the acquisition of second language speech* (pp. 536–541). Poznan, Poland: Adam Mickiewicz University.

A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding

Renqiao Zhang[1*] **Jiajun Wu**[1*] **Chengkai Zhang**[1] **William T. Freeman**[1,2] **Joshua B. Tenenbaum**[1]
[1]Massachusetts Institute of Technology [2]Google Research {andy17, jiajunwu, ckzhang, billf, jbt}@mit.edu

Abstract

Humans demonstrate remarkable abilities to predict physical events in complex scenes. Two classes of models for physical scene understanding have recently been proposed: "Intuitive Physics Engines", or IPEs, which posit that people make predictions by running approximate probabilistic simulations in causal mental models similar in nature to video-game physics engines, and memory-based models, which make judgments based on analogies to stored experiences of previously encountered scenes and physical outcomes. Versions of the latter have recently been instantiated in convolutional neural network (CNN) architectures. Here we report four experiments that, to our knowledge, are the first rigorous comparisons of simulation-based and CNN-based models, where both approaches are concretely instantiated in algorithms that can run on raw image inputs and produce as outputs physical judgments such as whether a stack of blocks will fall. Both approaches can achieve super-human accuracy levels and can quantitatively predict human judgments to a similar degree, but only the simulation-based models generalize to novel situations in ways that people do, and are qualitatively consistent with systematic perceptual illusions and judgment asymmetries that people show.

Keywords: physical scene understanding; neural network; analysis by synthesis; simulation engine; blocks world

Introduction

The outputs of vision include not only the objects in a scene and their spatial relations, but also their physical properties and relations: What is heavy or light? What is balanced or attached, and what isn't? What is likely to fall? What will happen next? When objects move, their motion can be predicted from these physical inferences; motion can also affect our physical judgments when objects move in unexpected ways.

These capacities for physical scene understanding are basic to how we see the world. Precursors to them can be found in infants as young as 3-5 months old, even before children acquire their first words labeling kinds of objects (Carey, 2009; Baillargeon, 2004). Building computational models of these abilities has been a target for recent work in both cognitive science and computational vision (Battaglia, Hamrick, & Tenenbaum, 2013; Gupta, Efros, & Hebert, 2010; Mottaghi, Bagherinezhad, Rastegari, & Farhadi, 2015; Fragkiadaki, Agrawal, Levine, & Malik, 2015; Zheng, Zhao, Yu, Ikeuchi, & Zhu, 2015; Li, Azimi, Leonardis, & Fritz, 2016). In contrast to earlier work on intuitive physics that emphasized explicit reasoning about textbook-style physics problems (McCloskey, 1983), with models focused on people's qualitative judgments (Forbus, 1984; Siegler, 1976), recent studies of physical scene understanding have looked at more rapid, perceptual inferences, which can be parametrically manipulated and modeled quantitatively, and which could serve as the basis for grounded action planning. Several studies have argued that rapid perceptual inferences about the physics of scenes can be explained by positing an "intuitive physics engine" (IPE), a mental system for approximate probabilistic simulation analogous to those used in video-game physics engines (Sanborn, Mansinghka, & Griffiths, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; K. A. Smith & Vul, 2013). These simulation engines approximate object dynamics interacting under Newtonian or other forms of classical mechanics over short time scales, in ways that are perceptually reasonable (if not necessarily physically accurate) and efficient enough to run in real time for complex scenes.

Other authors have suggested that the simulation-based IPE scheme might be prohibitively expensive for brains to implement (Davis & Marcus, 2016). An alternative class of models has been proposed based on stored memories of experienced scenes and physical outcomes, together with pattern recognition algorithms (such as neural networks) for accessing appropriate memory items to predict outcomes in a new scene context (Sanborn et al., 2013; Sanborn, 2014).

Although cognitive scientists have yet to seriously test memory-based alternatives to simulation in physical scene understanding tasks, AI researchers at Facebook recently demonstrated such a possibility in a working system. Lerer, Gross, and Fergus (2016) trained deep convolutional neural networks (CNNs) to make physical predictions directly from visual images, judging for instance whether a stack of blocks will fall, as Battaglia et al. (2013) studied empirically and modeled using approximate probabilistic simulation. The FAIR neural network, named PhysNet, was partly pretrained on ImageNet (Krizhevsky, Sutskever, & Hinton, 2012) and then trained on after a large dataset of synthetic scenes and outcomes. It achieved a high accuracy (89%) on the stability prediction task, generalized to real images reasonably well (67%), and exhibited positive correlations with human responses. This suggests that memory-based systems for visual intuitive physics may be promising at least in AI applications, and perhaps also as cognitive models.

Motivated by the success of CNNs in machine vision object recognition tasks (Krizhevsky et al., 2012), neuroscientists have proposed analogous architectures as accounts of the fast feedforward aspects of human visual object recognition (Yamins et al., 2014; Serre, Oliva, & Poggio, 2007). If CNNs can be successfully applied to physical scene understanding tasks as well, they could offer a compelling alternative to simulation as an account of how people can predict physical outcomes so well, so quickly.

Our goal in this paper is to conduct the first rigorous em-

∗ indicates equal contributions.

pirical comparisons of simulation-based (IPE) and neural-network-based (CNN) models for physical scene understanding. Although CNNs have many appealing features as models of visual cortex, they also have features that are less appealing – and arguably less human-like. They typically require large amounts of training data, which a human might not have access to. Large training sets may be required for any new scenario, even if it is just a simple variation on previously seen cases. For instance, in order to predict whether a pile of four blocks is stable, a CNN may have to see at least thousands of cases that either do or do not fall under gravity. In contrast, an IPE model, just like humans, is able to make many predictions with reasonable accuracies without training, as the simulation engine within encodes abstract physical knowledge that applies to a very wide range of scenes.

Even with a large amount of training data, it is unclear whether the knowledge learned by CNNs may be transferable to some similar cases. Lerer et al. (2016) showed that a network trained on images of two and four blocks could generalize to images of three blocks to some extent, but there is no clear way for a neural network to answer a different but related question to those it is trained for, *e.g.*, in which direction the blocks would fall, unless explicit labels are provided during training. One of the main points in favor of IPE models is their ability to explain how people can easily make many different judgments about very different configurations of blocks, without specific training (Battaglia et al., 2013).

Perhaps most interestingly, people are prone to systematic "physics illusions" that IPE models naturally capture. For instance, stacks of blocks often look to people as if they are sure to fall when they are actually carefully balanced. People do not, however, make the opposite error: They do not systematically mistake unstable stacks for stable ones. Probabilistic simulation-based models are similarly tempted to make this asymmetric pattern of errors (Battaglia et al., 2013): Small amounts of uncertainty in the simulation can make a stable configuration appear unstable, but are unlikely to make an unstable one appear stable. It is unclear whether neural-network-based models can capture these perceptual illusions.

In this paper, we report four experiments comparing the behavior of discriminatively trained neural networks and generative simulation-based models with human judgments on blocks-world physics tasks, addressing the questions above. Exp. 1 evaluates the performance of the IPE model and performance-optimized neural networks in predicting block stability. Exp. 2 explores the role of limiting CNN training data, to see if performance on smaller training sets looks more human-like. Exp. 3 evaluates both model classes for asymmetries in the stability illusions described above. Exp. 4 tests CNNs and IPE models' ability to generalize to situations slightly different from those the CNN was trained on.

The Blocks World

For our experiments, we study a set of seemingly simple but physically rich scenarios: a pile of blocks with one on top of another. Our goal is to study how humans and computational models behave on various tasks given these stimuli, and to reveal possible correlations between them. We now illustrate our stimuli in detail.

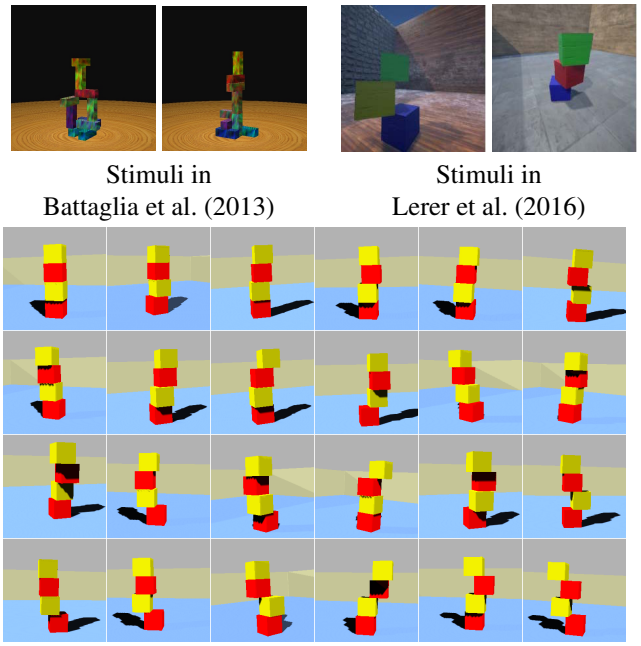


Figure 1: Sample stimuli used by Battaglia et al. (2013), Facebook AI Research (Lerer et al., 2016), and us. Our stimuli are ordered by increasing visual instability (defined in Experiment 3)

For each stimulus, there are four blocks with side length 1 meter piled on the ground, each supporting another on top of it. There is only one block at the same height level. Because laying blocks at uniform random is likely ($p = 75\%$) to result in an unstable system, we draw the horizontal position of a block from a normal distribution with variance 0.29^2 centered at the horizontal position of the block under it, to ensure that there are half stable and half unstable piles in the dataset. Later, we study cases where the number of blocks varies, and for them we update the variance accordingly.

Whether blocks are stable, *i.e.*, groundtruth labels, can be derived from the coordinates of blocks. A block will fall if and only if the center of mass of all blocks above it, including itself, does not fall on top of the block under it.

For rendering, we generate images of resolution 256×256.We place a pile of blocks in a virtual experiment field with a size of 30×30 meters and a height of 4 meters. We have one light source, 16 meters high, to simulate real-life lightening. We also vary the position, focal point, and tilt angle of the camera. We represent its coordinates in cylindrical coordinates (r, θ, z), with origin on the ground right beneath the center of the bottommost block. The camera positions are sampled from $r \sim N(11, 0.3^2)$, $\theta \sim \text{Uniform}(0, \pi/2)$, and $z \sim N(3, 0.01^2)$. We choose these parameters to ensure all blocks are within the view of the camera. The focal point of the camera is set at the center of the pile plus a Gaussian noise with variance 0.2^2. We also tilt the camera; its angle from the

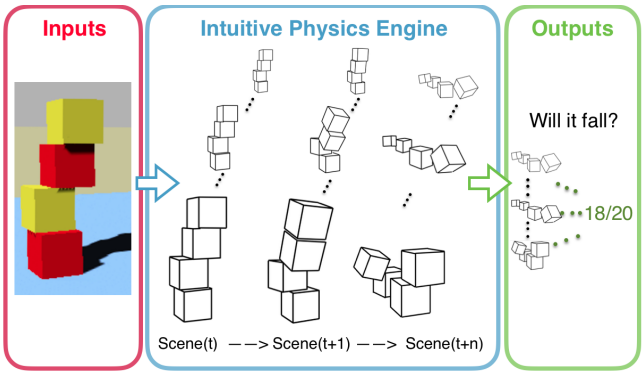


Figure 2: The Intuitive Physics Engine (IPE) model

direction of projection is sampled from $N(0, 2^2)$. We incorporate these variances for evaluating the generalization ability of the models.

Computational Models

We study two classes of computational models. One is the Intuitive Physics Engine (IPE) Model (Battaglia et al., 2013), which aims to simulate humans' reasoning on physical scenes by an approximate probabilistic simulation engine. The other is convolutional neural networks (CNNs), a class of discriminative recognition models that have gained much popularity in AI fields like computer vision in recent years.

The Intuitive Physics Engine Model

The Intuitive Physics Engine (IPE) consists of two components: a Bayesian vision system, which infers the configurations of blocks from given images, and a physical inference system, which calculates the Bayesian posterior probability distribution of physical properties (*i.e.*, stability) by running a number of simulations under perturbation forces and geometric noises. Figure 2 illustrates the IPE model. For more details, please see Battaglia et al. (2013).

For each scene, we render images of the initial state under perspective projection from three fixed viewpoints rotated by 45°. These triplets of images are then fed into the Bayesian vision system, which uses a Metropolis-Hasting (MH) sampling algorithm to infer a Bayesian posterior distribution of the scene's initial state (position, height, and the number of blocks presented). We run the MH sampling for $5,000$ steps, with a 2D Gaussian blurring kernel of width 2 on the observed images, as suggested by Battaglia et al. (2013).

With the inferred initial geometry, we run 20 simulations for each scene using the Open Dynamics Engine (ODE) (R. Smith, 2006). We set the friction coefficient to 0.2, the bounce coefficient to 0.2, and the side-length and density of each block to $1m$ and $500kg/m^3$, respectively. Gravity is set to $9.81m/s^2$ pointing downwards. Before each simulation starts, a horizontal zero mean Gaussian noise σ is added to the positions of blocks. Then the simulation runs at a step size of $10ms$ for 2 seconds. During the first second, a horizontal force with magnitude ϕ is exerted at the center of the bottom face of the bottommost block. The direction of the force is uniformly sampled from $(0, 2\pi)$ and changes at a frequency of 50Hz. We consider a pile unstable if the vertical coordinate of the top block changes by more than 0.2 meters when the simulation ends.

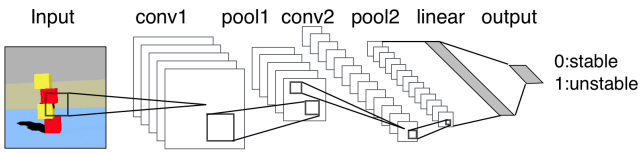


Figure 3: The structure of LeNet

Convolutional Neural Networks

CNNs have gained much popularity in computer vision (Krizhevsky et al., 2012). Here we consider two popular CNN frameworks: the small but powerful LeNet (LeCun, Bottou, Bengio, & Haffner, 1998), and the widely used AlexNet (Krizhevsky et al., 2012).

LeNet, originally proposed for digit recognition, has been widely used as a recognition model in vision because of its effectiveness and simplicity (LeCun et al., 1998). LeNet consists of two convolutional layers, each followed by a pooling layer and an activation layer. There are then two fully connected linear layers at the end. We modify the final layer so that instead of ten outputs for digit classification, the model now has two output units — its confidences on whether the blocks will fall or not. Figure 3 shows the structure of LeNet.

The second is the popular AlexNet (Krizhevsky et al., 2012), which achieves impressive performance on ImageNet classification. AlexNet consists of five convolutional, pooling, and activation layers, and three linear layers at the end. We evaluate both AlexNet pretrained on ImageNet, as well as AlexNet trained from scratch.

We use Torch (Collobert, Kavukcuoglu, & Farabet, 2011) for implementation. We set the learning rate to 0.01 for LeNet and for fine-tuning AlexNet, and to 0.2 for training AlexNet from scratch. We use stochastic gradient descent for training.

Behavioral Experiments

To collect human responses, we first randomly divide all test images into groups, each consisting of 10 images. We then add four *easy* cases (two stable, two unstable), whose stability is visually apparent, into the group. For each group, we collect 80 responses on Amazon Mechanical Turk. We only allow workers with an approval rate $> 90\%$ to submit responses, and we only accept responses from workers that answered all four easy cases correctly.

Experiment 1: Predicting Falling Blocks

In our first experiment, we test the performance of the IPE model and neural networks on images with four blocks, and compare the results with human responses.

Experimental Setup For the IPE model, we consider cases with various levels of geometric Gaussian noises σ and external forces ϕ during physical simulations. We then compare their performance with LeNet, AlexNet, and humans.

		ϕ				
		0	35	40	45	50
σ	0	94.2	87.2	79.5	71.3	63.8
	0.05	91.3	83.4	76.1	69.1	61.8
	0.1	83.2	75.7	70.3	62.6	56.4
	0.15	72.2	66.8	59.4	54.2	51.2
	0.2	58.5	53.8	52.1	51.0	50.9
Corr		≥ 0.45	≥ 0.54	≥ 0.56	≥ 0.58	≥ 0.60

Table 1: Accuracies (%) of the IPE model with different σ and ϕ, and their correlations with human responses. We use $(\sigma, \phi) = (0.1, 40)$ for following experiments.

We use 1,000 test images, each with a pile of four blocks. For neural networks, we build a training set of 200,000 images (disjoint from the test set) with groundtruth labels.

Results and Discussions As shown in Table 1, when no geometric error or external force is added to the IPE model ($\sigma = 0, \phi = 0$), its results almost always match ground-truths (94.2% accuracy). Accuracy decreases as noises increase; however, as previously described in Battaglia et al. (2013), we also observe that correlation between IPE responses and human predictions goes up. For the following experiments, we use an IPE model with $(\phi, \sigma) = (0.1, 40)$ as it matches human performance in terms of both accuracy and correlation.

We compare results for stable and unstable cases separately, and list them in Table 2. We observe that human predictions and the IPE model responses have an asymmetric pattern: they perform well on unstable cases, but for images with a stable pile of blocks, their accuracies are much worse. On the contrary, neural networks do not exhibit a similar pattern; they have roughly the same accuracies for both cases. We will revisit this asymmetry more in Experiment 3.

Experiment 2: Limited Training Data

In our second experiment, we inspect the behaviors of neural networks with different sizes of training sets. As our the IPE model requires only one or a few examples for simulation, its performance does not change with the availability of training data. The same applies to humans.

Experimental Setup Instead of using training sets of 200,000 images, we now only provide the networks with training sets of 100 to 20,000 images. For each scale, we sample five training sets independently, train one network on each set, and compute the average of their performance. The other setup is same as that in Experiment 1.

Results and Discussions As shown in Figure 4, the performance of CNNs decreases as there are fewer training data. Although AlexNet (not pretrained) performs better with 200,000 training images, it also suffers more from the lack of data, while pretrained AlexNet is able to learn better from a small amount of training images. For our task, both models require around 1,000 images for their performance to be comparable to the IPE model and humans. We then evaluate

Method	Stable	Unstable	All
Human	38.0	92.9	65.5
IPE	40.7	99.0	70.3
LeNet (200K)	91.3	89.0	90.1
AlexNet (200K)	91.5	92.3	91.9
AlexNet (Pretrained, 200K)	94.5	94.7	94.6
LeNet (1,000)	68.0	69.3	68.7
AlexNet (1,000)	71.8	70.1	70.9
AlexNet (Pretrained, 1,000)	72.5	74.2	73.4

Table 2: Accuracies (%) of humans, IPE, LeNet, and AlexNet (pretrained and not pretrained), on 200K or 1,000 images. The results on 1,000 images are averaged over five models trained on independently sampled sets.

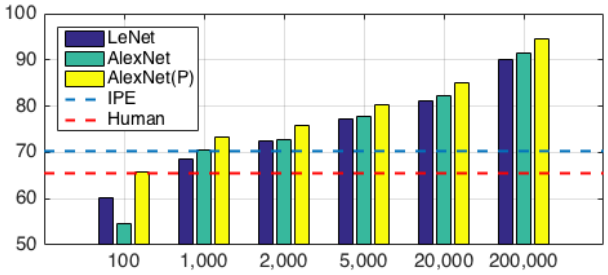


Figure 4: CNN models with different sizes of training sets

the networks trained with 1,000 images. As shown in Table 2, there is still no asymmetric pattern in the responses of the less-trained networks.

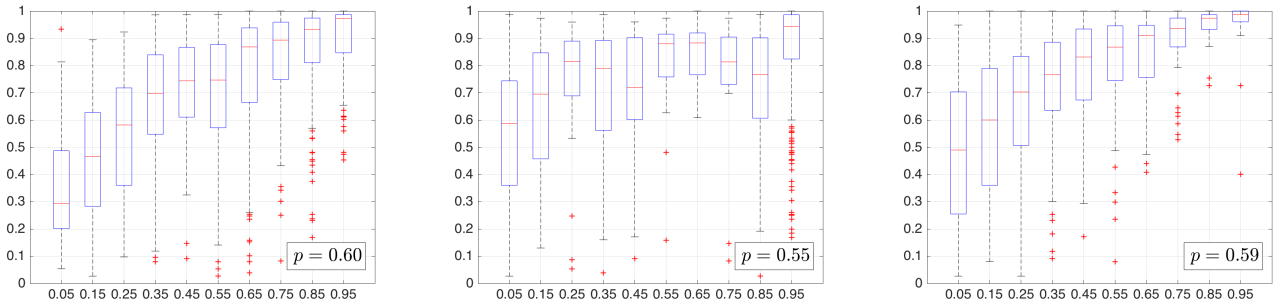
We now look into how each model correlates with human responses in more detail. Figure 5 (a) and (b) demonstrate that the IPE model has a stronger correlation with humans, compared to LeNet trained on the full training set. Another interesting finding is that the less-trained LeNet (c) is more human-like. We will discuss this more in the final section.

Experiment 3: Boundary Cases

We now systematically study the asymmetry we observed in Experiment 1. In particular, we focus on a few groups with *visually unstable* piles, *i.e.*, piles that are carefully balanced and therefore stable, but illusory to humans so that they believe these blocks will fall.

Experimental Setup We define *visual instability*, scaling from 0 to 5, to describe how unstable a pile of blocks looks like. A pile with instability value x means there exists at least one block so that the center of mass of the blocks above it lies $x/10$ meters away from its center on x-y plane. As the side-length of blocks is 1 meter, a pile with a visual instability value 4 looks very unstable to humans, significantly different from one with value 1. Figure 6 shows examples with various visual instability values.

For this experiment, we restrict possible camera positions so that the deviations of blocks can be clearly perceived. We generate four datasets of stable blocks with visual instabilities of 1, 2, 3, and 4 respectively, each with 100 images.



(a) Human (Y-axis) vs IPE (X-axis) (b) Human (Y) vs LeNet 200K (X) (c) Human (Y) vs LeNet 1,000 (X)

Figure 5: From left to right: human responses vs (a) responses of IPE (normalized numbers of moving blocks), (b) LeNet trained on the full training set (200,000 images), and (c) LeNet trained on 1,000 images. Results for AlexNet are similar. We list Pearson's correlation coefficients at the bottom-right corner.

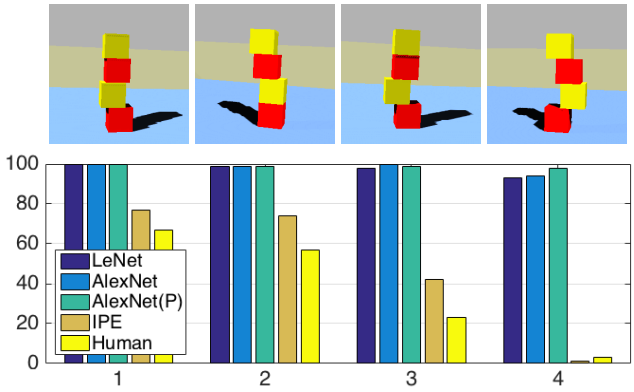


Figure 6: Upper: stable blocks with increasing visual instabilities; Lower: performance of LeNet, AlexNet, pretrained AlexNet, IPE, and humans on the four datasets. Neural networks are trained on 200K images. Behaviors of networks trained on a smaller set (1,000 images) are similar.

Results and Discussions As shown in Figure 6, the performance of neural networks are, in general, better than their performance in Experiment 1, probably because images here are easier as the camera positions are restricted. Also, their performance barely changes for groups with different visual instabilities. Even for the most deceptive group (visual instability 4), a LeNet has an accuracy of 93%. We also test AlexNet (both pretrained and not pretrained) on cases where blocks are unstable but visually stable, and the network, again, gives highly accurate results ($\geq 93\%$).

The performance of IPE and humans, on the other hand, changes drastically across groups. Corresponding to results in Experiment 1, both IPE and humans consistently predict that blocks with visual instability 4 will fall. Their accuracies are higher when visual instability is smaller, but still not close to those of neural networks. This confirms our observation of the asymmetry. More discussions follow in the final section.

Experiment 4: Knowledge Transfer

A possible explanation to humans' one-shot learning ability is based on the concept of transfer learning. In our fourth experiment, we evaluate the behaviors of computational models on tasks involving knowledge transfer.

Experimental Setup For this experiment, we generate 200 test images with three and five blocks, respectively. Examples are shown in Figure 7. We modify the variance of block positions to ensure there are half stable and half unstable cases.

Our Bayesian vision system is extended to include the number of blocks as one parameter in sampling. Because the number of blocks directly determines the total mass, we also vary the magnitude of the perturbation force according to the inferred number of blocks to keep its effect consistent. For neural networks, we simply test the models previously trained on the 200,000 images with four blocks.

Results and Discussions Table 3 shows that while CNNs achieve $\sim 90\%$ accuracies on four-block cases, their performance is much worse on cases where the number of blocks is smaller than that in training examples. Specifically, the predictions of models trained on 200K images are at chance. For cases with more blocks, CNNs, especially pretrained AlexNet, can learn to generalize to some extent. However, their behaviors are different from human responses. In comparison, humans and the IPE model have relatively consistent performance, with slight decreases in accuracies as the number of blocks goes up and the task becomes more difficult.

These experiments demonstrate that the knowledge learned by neural networks cannot be transferred, at least in a straightforward way, to scenarios outside the training set. The IPE model and humans enjoy more flexibility in reasoning in the complex world and solving more general problems.

General Discussion

Following Facebook AI's reported results, we found that convolutional neural networks can be trained to achieve superhuman accuracy levels on stability judgment tasks from raw images (Exps. 1 and 2). CNNs also correlate reasonably well with human intuitions about how likely a stack of blocks is to fall, and once trained, they can respond to new images extremely quickly. However, these features do not automatically make CNNs a good model of people's physical intuitions. They do not capture systematic judgment asymmetries that humans make, which simulation-based IPE models do capture (Exps. 1-3). CNNs also have limited generalization ability across even small scene variations, such as changing

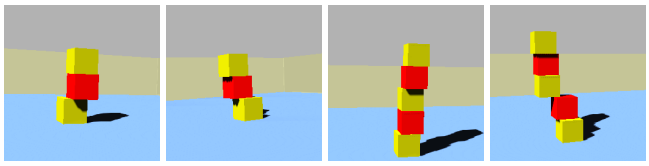


Figure 7: Images with three or five blocks

Model	Training	Test Set			
		3	4	5	Avg
LeNet (200K)	4	50.5	88.5	64.0	67.7
AlexNet (200K)	4	52.5	89.5	65.5	69.2
AlexNet (P, 200K)	4	51.0	95.0	78.5	74.8
LeNet (1,000)	4	57.0	64.0	66.0	62.3
AlexNet (1,000)	4	54.0	62.0	64.5	60.2
AlexNet (P, 1,000)	4	55.0	71.0	72.0	66.0
IPE $(0.1, 10x)$	N/A	72.0	64.0	56.0	64.0
Human	N/A	76.5	68.5	59.0	68.0

Table 3: Results on the task of transfer learning

the number of blocks. In contrast, IPE models naturally generalize and capture the ways that human judgment accuracy decreases with the number of blocks in a stack (Exp. 4).

Taken together, these results point to something fundamental about human cognition that neural networks (or at least CNNs) are not currently capturing: the existence of a mental model of the world's causal processes. Causal mental models can be simulated to predict what will happen in qualitatively novel situations, and they do not require vast and diverse training data to generalize broadly, but they are inherently subject to certain kinds of errors (*e.g.*, propagation of uncertainty due to state and dynamics noise) just in virtue of operating by simulation.

Despite the success of CNNs in accounting for other high-level human perceptual capacities, such as rapid object classification (Yamins et al., 2014), our results suggest that at least some perceptual judgments which people can make in a quick glance are not well explained by current feedforward neural networks. We should not conclude however, that neural networks cannot help to explain how people make intuitive physical judgments. If people do indeed have a "physics engine in the head", somehow this simulator must be implemented in neural circuits. Recurrent neural networks (RNNs) could provide one model for this (Fragkiadaki et al., 2015). It is also possible that CNNs, if trained on more diverse scenes and physical judgments than those studied here and/or pretrained on large-scale image classification tasks (as in Lerer et al., 2016), could capture more of the qualitative inference behavior people show in our tasks. Lastly, CNNs could be useful for visual intuitive physics by quickly estimating the relevant object properties in images needed to represent the world's state in a physics engine, which would then support more sophisticated reasoning and prediction by simulation (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). Going forward we are eager to explore these and other productive lines of exchange between simulation-based generative models and memory-based neural network models.

Acknowledgement We are grateful to Tomer Ullman for helpful discussions. This work is in part supported by NSF Robust Intelligence 1212849 Reconstructive Recognition, the Center for Brain, Minds and Machines (NSF STC award CCF-1231216), and MERL.

References

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In *BigLearn, Neural Information Processing Systems Workshop*.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

Forbus, K. D. (1984). Qualitative process theory. *Artificial intelligence*, *24*(1), 85–168.

Fragkiadaki, K., Agrawal, P., Levine, S., & Malik, J. (2015). Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy newtons: Unifying process and dependency accounts of causal attribution. In *Annual Meetings of the Cognitive Science Society*.

Gupta, A., Efros, A. A., & Hebert, M. (2010). Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*.

Li, W., Azimi, S., Leonardis, A., & Fritz, M. (2016). To fall or not to fall: A visual approach to physical stability prediction. *arXiv preprint arXiv:1604.00066*.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–130.

Mottaghi, R., Bagherinezhad, H., Rastegari, M., & Farhadi, A. (2015). Newtonian image understanding: Unfolding the dynamics of objects in static images. *arXiv preprint arXiv:1511.04048*.

Sanborn, A. N. (2014). Testing bayesian and heuristic predictions of mass judgments of colliding objects. *Front. in Psychology*, *5*.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, R. (2006). *Open Dynamics Engine (ODE)*.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Neural Information Processing Systems*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., & Zhu, S.-C. (2015). Scene understanding by reasoning stability and safety. *International Journal of Computer Vision*, *112*(2), 221–238.

Learning Behavior-Grounded Event Segmentations

Christian Gumbsch (christian.gumbsch@student.uni-tuebingen.de)

Jan Kneissler (jan.kneissler@uni-tuebingen.de)

Martin V. Butz (martin.butz@uni-tuebingen.de)
Chair of Cognitive Modeling, Department of Computer Science and Department of Psychology, Faculty of Science,
Eberhard Karls University of Tübingen, Tübingen, Germany

Abstract

The event segmentation theory (EST) postulates that humans systematically segment the continuous sensorimotor information flow into events and event boundaries. The basis for the observed segmentation tendencies, however, remains largely unknown. We introduce a computational model that grounds EST in the interaction abilities of a system. The model learns events and event boundaries based on actively gathered sensorimotor signals. It segments the signals based on principles of probabilistic predictive coding and surprise. The implemented model essentially simulates, anticipates, and learns event progressions and event transitions online while interacting with the environment by means of dynamic, predictive Bayesian models. Besides the model's event segmentation capabilities, we show that the learned encodings can be used for higher-order planning. Moreover, the encodings systematically conceptualize environmental interactions and they help to identify the factors that are critical for ensuring interaction success.

Keywords: event models; object interaction; predictive encoding; event segmentation; higher order planning; factorization; conceptualization

Introduction

The embodiment turn in cognitive science has emphasized the importance of simulating relevant aspects of the outside environment by means of perceptual symbol systems (Barsalou, 1999). To enable motor-grounded simulations, the inclusion of actions was emphasized (Engel, Maye, Kurthen, & König, 2013). Moreover, the importance of explicit forms of predictions and anticipations has been emphasized, supporting both, cognitive development (Barsalou, Breazeal, & Smith, 2007) and adaptive, goal-directed behavior (M. Botvinick & Weinstein, 2014; Butz, Sigaud, & Gérard, 2003; Sigaud, Butz, Pezzulo, & Herbort, 2013). In fact, recent treatises suggest that predictive coding and anticipations may form the foundations that bring about embodied cognition (Clark, 2013; Friston, 2009; Hohwy, 2013). In this paper, we present an algorithm that models an anticipatory learning system, which develops suitable compositional structures to interact with the environment adaptively and goal-directedly.

The *event segmentation theory* (EST) (Zacks & Tversky, 2001; Zacks, Speer, Swallow, Braver, & Reynolds, 2007) suggests that humans tend to structure the stream of sensory perceptions into events and event transitions. Events were characterized as "a segment of time at a given location that is conceived by an observer to have a beginning and an end" (Zacks & Tversky, 2001, p. 3). In various studies that focused on event structure perception, it was shown that events are characterizable as relatively uniformly unfolding interactions, whereas event boundaries are characterized by sudden, strongly non-linear changes in the unfolding events. While some of these changes seem to be strongly related to movement variables, movement variables alone could not account for all the segmentations that humans indicated (Zacks, Kumar, Abrams, & Mehta, 2009). We propose that event segmentations may be grounded in, and develop from, own sensorimotor experiences.

To investigate this proposition, we introduce a computational cognitive model implementation, which is based on Zacks et al. (2007)'s schematic EST model. The implemented system learns how it is able to manipulate objects solely by actively processing sensorimotor interactions. The system essentially develops a predictive world model, which segments the gathered sensorimotor experiences into events and event transitions from scratch. Events are sets of forward models that are active over an extended period of time while interacting with the environment. Event boundaries mark the beginning and ending of particular events. As a result, the individual events characterize particular object manipulations or simple hand movements, while event boundaries identify types of contact onset and offset events. We show that the developing structures are highly suitable (i) to predict the future sensorimotor progression, including when the next event boundary is probably reached and which event can be expected next, and (ii) to execute higher-order, goal-directed planning. We particularly show that the developing hierarchically organized, event-oriented, behaviorally-grounded structures are highly suitable for executing factorized, hierarchical reinforcement learning (RL) according to the options framework (M. M. Botvinick, Niv, & Barto, 2009; Sigaud, Butz, Kozlova, & Meyer, 2009; Sutton, Precup, & Singh, 1999).

Architecture

According to EST, the processing of sensory inputs is influenced by a set of event models, which predict future sensory input. Information about errors in these predictions is used to adapt and switch between the available event models. We implement this approach by using forward models as event models to generate sensory predictions. Additionally, to form representations of events as a set of forward models, our system builds representation of event boundaries, marking the transition from a forward model to another.

The system consists of four main components. It is schematically shown in Figure 1. The **Predictive System** consists of a set of currently active event models, which pre-

dict the next sensory information $\vec{s}_{t+1}$ based on the current sensory input $\vec{s}_t$ and the executed motor action $\vec{x}_t$. The active event models are updated while interacting with the environment using the error $\vec{e}_{t+1}$ between predicted $\vec{s}'_{t+1}$ and actually encountered next sensory perceptions $\vec{s}_{t+1}$. **Error Detection** mechanisms are used to keep track of the accuracy of the predictions. If a significant prediction error is detected, which may be related to notions of *surprise*, event transitions might be detected, leading to a change in the set of active event models. The **Event Models**-component contains all currently active event models and determines which ones are currently applicable. Moreover, the component learns new event models when necessary. Finally, the **Event Boundary Model**-components contain models that characterize event transitions dependent on the currently active event models, the sensory information, and the surprise signal.

These four components essentially constitute the learning and control architecture. The architecture encodes events in the form of sets of temporal forward models. Event boundaries are characterized by event boundary models, which additionally predict the state of the environment at particular event transitions. While interacting with the environment, the system internally simulates the interaction with the environment and verifies these simulations given sensory observations. Moreover, the system can be used to actively infer actions given desired goal states. Due to the event-based architecture, higher-level inference-based planning is possible.

Predictive system

While simulating the current changes in the environment, the predictive system holds N forward models $M(t) = (M_{1,i_1}(t), ..., M_{N,i_N}(t))$ at a certain point in time t. Each forward model predicts sensory changes given motor commands, that is, $M_{n,i} : \vec{x} \rightarrow \Delta s$. At a certain point in time t, each active forward model $M_{n,i_n}(t)$ receives the motor command $\vec{x}_t$ as an input and predicts the sensory consequences $\Delta s'_{n,t+1}$. The predictions of all N active forward models form the predicted sensory change vector $\Delta\vec{s}'_{t+1}$. The predicted sensory input of the next time step is thus:

$$\vec{s}'_{t+1} \leftarrow \vec{s}_t + \Delta\vec{s}'_{t+1}. \quad (1)$$

After executing $\vec{x}_t$ the real sensory input $\vec{s}_{t+1}$ is used to update the active forward models to improve the respective model predictions. Additionally, each forward model $M_{n,i}$ stores the moving average (over the 100 last steps) of its prediction error $\bar{e}_{M_{n,i}}$ and the variance of that prediction error $\sigma^2_{M_{n,i}}$, estimating the current accuracy of the model predictions.

Choosing sufficiently small step sizes, the velocity kinematics of any dynamic system can be approximated arbitrarily well by a linear model. In the general case, the current velocity kinematics depend on the system state. Seeing that in our simple test scenario this was not the case, though, we assumed non-changing, linear velocity kinematics, which we learned by means of Recursive Least Squares.

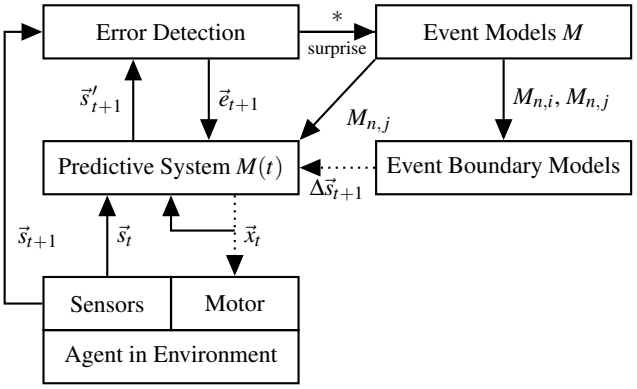


Figure 1: Illustration of the system. Solid arrows symbolize the information flow during forward modeling. In this pathway active forward models predict the next sensory input, are improved by learning and exchanged if necessary. Dashed arrows symbolize the information flow during planning. Here the required sensory change to trigger a desired event is computed and the forward models are used inversely to generate a suitable motor command.

Error detection and switching event models

To this point our architecture uses a set of active forward models to predict the sensory changes during the course of one event. However, if the system is confronted with an event boundary, marking the end of the current event and the beginning of a new one, it should autonomously decide to switch its set of forward models.

According to EST an event boundary is accompanied by a rising error in prediction (Zacks et al., 2007). It has been shown that computational models can use prediction errors to segment a stream of video sequences into separate events (Reynolds, Zacks, & Braver, 2007) or to autonomously identify useful subgoals for higher level motor planning (Butz, Swarup, & Goldberg, 2004). Our architecture uses the error in prediction as a criterion to detect a possible event boundary. The prediction error $e_{n,t}$ for the sensory dimension n at time step t is considered 'surprising' (Butz et al., 2004) if

$$e_{n,t} > \bar{e}_{M_{n,i}} + 2\sigma_{M_{n,i}}, \quad (2)$$

with $M_{n,i}$ being the currently active forward model of dimension n.

If a significant error signal is detected the architecture enters a *searching period*, during which the next active event models are determined. All existing forward models M_n of dimension n and one newly generated forward model are considered. For a fixed number of time steps (10 time steps in our simulations) each model predicts the next sensory input, is updated, and the prediction error is recorded. Afterwards, the mean prediction errors of all models during this searching period are compared and the model $M_{n,j}$ with the smallest mean error is chosen as the new forward model for dimension n in the predictive system. If the winning model $M_{n,j}$

already existed prior to searching, the newly generated model is discarded. If $M_{n,j}$ is new, it is added to the set of possible models M_n for dimension n. All forward model updates on existing models during a searching period are discarded.

Learning event boundary models

The introduced components form a mechanism to detect event boundaries, which can be characterized by an exchange of at least one of the active forward models in the predictive system. For our architecture to be able to act goal-directedly, a representation is necessary that describes at which situation one event boundary occurs. Assuming that an event boundary can be characterized by particular constellations of event-boundary-relevant sensory inputs, we approximate an event boundary by the probability density of sensory constellations that are experienced when a transition occurred. In other words, we are modeling the conditional probability $P(\vec{s} \mid M_{n,i} \to M_{n,j})$, making the assumption that this probability distribution can be reasonably well approximated by a multidimensional normalized Gaussian function $G_{\vec{\mu}_{n,i\to j},\Sigma_{n,i\to j}}(\vec{s}_t)$. This is equivalent to requiring that event boundaries occur close to specific points in sensory space. This assumption holds well in the simple scenario considered. In the general case, other densities may be used such as Gaussian mixture models.

Planning

To be able to trigger desired events, our system can be used in a backwards fashion to plan goal-directed behavior to reach specific event boundaries. To do so, we approximate active inference (Friston et al., 2013) by means of the developing event and event boundary models. As a result, planning consists of two inference stages. First, a target event boundary, or a sequence of event boundaries, is chosen. Next, the necessary motor commands are inferred to reach the next desired event boundary.

Selection of a target event boundary We assume that some of the event boundaries are coupled with positive reward. In our model, event boundaries are characterized by event transitions, such that a particular event transition $M_{n,i} \to M_{n,j}$ is chosen as the goal transition. As a result, the system strives to achieve this transition by attempting to maximize $P(M_{n,i} \to M_{n,j}, \vec{s}_t)$. Higher-level, inference-based planning is used to determine a sequence of event boundaries, which is expected to lead from the current event to the desired event transition.

When $M_{n,i}$ to $M_{n,j}$ is the only desired transition and $M_{n,i}$ is the currently active model, then the system strives to maximize $P(M_{n,i} \to M_{n,j}, \vec{s}_t)$. When multiple event transitions are considered desirable, that is, when transitions from the currently active model $M_{n,i}$ to a set J of potential target models $(M_{n,j})_{j\in J}$ are rewarding, then the transition to the closest mean $\vec{\mu}_{n,i\to j}$ of the associated Gaussian is chosen. When $M_{n,i}$ is currently active, but only transitions $M_{n,k} \to M_{n,j}$ are expected to be rewarding (with $k \neq i$), the system chooses a reachable intermediate transition $M_{n,i} \to M_{n,k}$. Although in our simulation one intermediate transition always suffices, the principle can generally be applied for generating larger sequences of transitions. The approach is also closely related to model-based, hierarchical RL, where extended actions are described as 'options' (Sutton et al., 1999; M. M. Botvinick et al., 2009).

Deriving the motor commands Let us assume that the nth coordinate prediction is based on model i and the transition $M_{n,i} \to M_{n,j}$ is currently desired. The system is supposed to reach a place $\vec{s}'_n$ in sensor space that maximizes the conditional probability of the desired transition:

$$\vec{s}'_n := \arg\max_{\vec{s}} P(M_{n,i} \to M_{n,j} | \vec{s}). \tag{3}$$

This can be reformulated and solved using the Bayes theorem, assuming that also the prior of $\vec{s}_n$ is a Gaussian distribution $P(\vec{s}_n) = G_{\vec{\mu}_{n,i},\Sigma_{n,i}}(\vec{s}_n)$:

$$\vec{s}'_n = (\Sigma_{n,i}\,\vec{\mu}_{n,i\to j} - \Sigma_{n,i\to j}\,\vec{\mu}_{n,i})\,(\Sigma_{n,i} - \Sigma_{n,i\to j})^{-1}. \tag{4}$$

Under the assumption that the prior is approximately uniform (compared to the transition distribution), $\Sigma_{n,i\to j} \ll \Sigma_{n,i}$ this corresponds to the maximum of $P(\vec{s}_t \mid M_{n,i} \to M_{n,j})$. We thus are making small steps in sensory space following the gradient of the transition model:

$$\Delta\vec{s}'_t = \eta\, W\, \nabla P(\vec{s}_t \mid M_{n,i} \to M_{n,j}). \tag{5}$$

If the matrix W is the identity matrix this performs an exact gradient ascend with step width constant η. However, we found better performance when choosing W such that dimensions with large variance are effectively suppressed. The suppression essentially focuses system behavior on the behaviorally relevant input dimensions.

Finally, the desired displacement $\Delta\vec{s}'_t$ in sensory space can be translated directly into a motor command using the inverse of the prediction model:

$$\vec{x}'_t := M^{-1}_{n,i}(\Delta\vec{s}'_t), \tag{6}$$

effectively making the system move towards an area where a desired event boundary is believed to be situated.

Evaluation

Our system was tested in a scenario, where multiple events occur and thus the acquisition of different forward models is necessary to predict the sensory changes. We have therefore chosen a scenario in which a simulated agent interacts with different objects in continuous space. Figure 2 shows the hidden, conceptual structure of the environment, which the model uncovers by the detailed principles.

The agent consists of a hand, able to move freely through a limited workspace, and a stationary mouth area. Three types of differently colored objects (1 type of 'foe' and 2 types of 'food') occur in the simulation. Foe objects have no friction

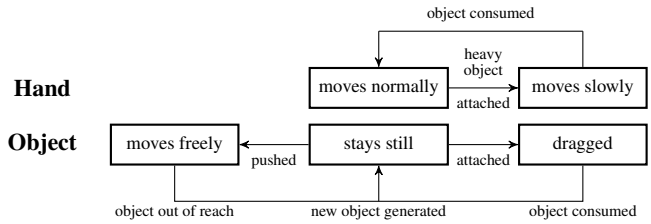


Figure 2: Illustration of the different events (boxes) and event boundaries (arrows), which the introduced system uncovers. The system's 'hand' is able to attach to objects or to push objects, dependent on the object type. A pushed object moves away from the hand until it is out of reach. An attached object moves with the hand until it is consumed. Hand movements are slower when a heavy object is attached.

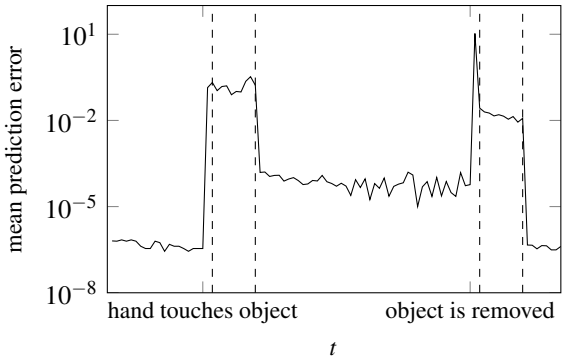


Figure 3: Mean prediction error of all active forward models for one exemplary interaction with a food object. The x-axis displays time with event boundaries highlighted. Dashed lines mark the beginning and end of a searching phase

and slide away without friction when pushed by the hand. They vanish when the distance of the object to the center of the agent's workspace exceeds a threshold. Food objects stick to the hand upon contact and afterwards move along with it. They vanish when they are dragged into the mouth. We use two types of food objects: Light food does not alter the hand movement when attached to it, whereas heavy food slows the hand movement down by a factor of $\frac{1}{2}$. If an object vanishes, a new one is immediately generated at a different position.

In every simulation step t one elementary movement of the hand, described by $\vec{x}_t$, is performed and a sensory input $\vec{s}_t$ is received, which contains all information necessary to predict event boundary occurrences. In particular, $\vec{s}_t$ consists of the position of the hand ($s_{1,t}, s_{2,t} \in [0,100]$), the position of the object ($s_{3,t}, s_{4,t} \in [0,100]$), the position of the object in a hand-centered frame of reference ($s_{5,t}, s_{6,t} \in [-100,100]$), the distance of the object to the center of the workspace ($s_{7,t} \in [0,50]$) and the object's color ($s_{8,t} \in \{0,100,200\}$). $\vec{x}_t$ contains the motor command, which determines the change in hand position (with $\Delta s_{1,t}, \Delta s_{2,t} \in [-0.5,0.5]$). Since the forward models of our architecture must be able to linearly compute the change in sensor information based on $\vec{x}_t$, the vector additionally contains sensory information describing the velocity of the hand during the last object contact (to predict the position of the foe after pushing it) and the velocity of the object in reference to the center of the workspace (to predict changes in the object's distance to the workspace center). A small amount of Gaussian distributed motor noise was added ($\sigma = 0.05$), such that an elementary movement was not completely deterministic.

In our scenario the event boundaries leading to the disappearance of an object and the creation of a new one are rewarded. Therefore the system strives to drag food objects in the agent's mouth and 'kill' foes by pushing them out of the agent's workspace. We chose to reward the event boundaries of the sensory dimension s_3, since the forward models for the object's position need to change at every event boundary and is therefore considered most reliable.

Results

In a first test we evaluated the improvement of the forward models over time by monitoring the prediction error in ten independent simulations. The motor command $\vec{x}_t$ was determined by an informed, hard-coded algorithm, which made the system touch an object with the hand, pushing it away or subsequently dragging it into its 'mouth'. The average prediction error of all active forward models for one exemplary object interaction is plotted in Figure 3. In this example, the hand first moves to the food. After contact, the food sticks to the hand and is moved alongside the hand into the mouth, which results in 'food consumption' and thus food removal. After that, a new object is generated randomly. At the event boundaries (hand touches object, object is removed) the prediction error drastically increases. This ascent is particularly big when the object is removed, since a lot of sensory information changes in this single time step. For the following ten time steps, the system searches for new forward models, such that the prediction error remains large. After that, the best adapted set of forward models is active. The prediction error for all forward models decreases over time. Figure 4 shows the prediction error for some of the forward models over their time of activation. While the prediction errors strongly fluctuate, they all logarithmically converge to 0. All forward models that correctly predict no change in sensory information immediately reach a prediction error of 0 (not shown).

In a second test we evaluated the planning capabilities of our system to use its event and event boundary models to perform goal-directed behavior. The system's goal was to trigger events resulting in the removal of the currently present object. We ran ten simulations, whereas one simulation run consisted of 25 epochs, each consisting of a training and a testing phase. During training, five objects of each type were presented consecutively at random positions. The system was given a time interval of 500 simulation steps to interact with the object. If the system failed to remove the object in the given time period, the hard-coded algorithm used above performed the

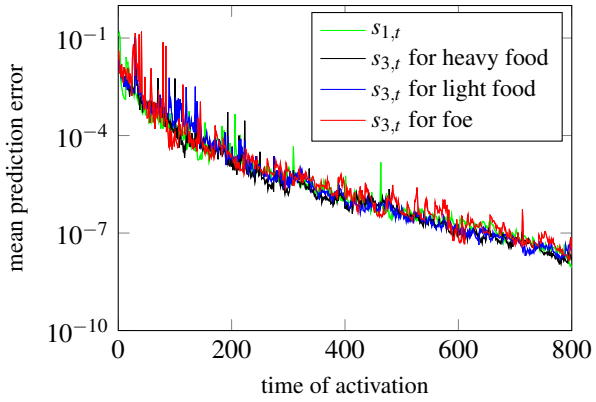
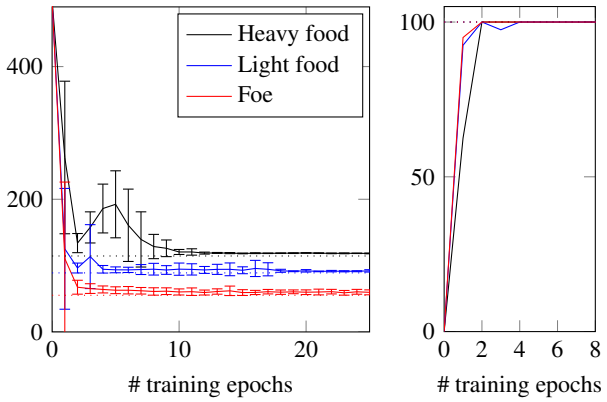


Figure 4: Mean prediction error of one forward model over the time this model is active. Colors indicate the type of the present object and the sensory dimension this forward model predicts. Only the non-trivial cases, in which the object moves, are shown.



(a) Mean number of time steps (b) % successful

Figure 5: Goal-directed behavior during testing epochs; solid lines show system performance; dotted lines show optimal performance; a) mean and standard deviation of time steps required to successfully interact with an object; b) mean percentage of successfully completed object interactions.

required movements. During testing, each object consecutively appeared at four fixed positions. When removing an object during testing the hand was reset to a starting position. Figure 5a shows the mean number of time steps the system needed to remove an object for the different testing epochs. In the first two testing epochs the time required for the interactions drastically decreases. After ten training epochs, the system performs the interactions in nearly minimum time (dashed lines indicate the optimum). Figure 5b shows the percentage of objects removed by the system's hierarchical, goal-directed behavior. Already after the first epoch, the system successfully removes nearly all foe and light food objects. From the fourth epoch onwards, the system removes all objects reliably within the allowed time frame.

To analyze if the system was able to differentiate between relevant and irrelevant sensory dimensions for the prediction of an event boundary, we analyzed the variances of the covariance matrices of each event boundary model after one exemplary run. The mean difference over all Gaussian distributions between biggest and smallest variance in between each Gaussian distribution is 697 – implying that there are drastic differences in relevance for the different sensory dimensions. A more detailed analysis shows that the largest and smallest variance indeed depended on the event boundary. For example, a model describing the event boundary 'hand touches foe' contains the biggest variances for the global position of the object (x-wise 296, y-wise 249) and small variances for the object's color (0.007) and the position of the object in the hand-centered frame of reference (x-wise 19.9, y-wise 21.6). This implies that the object's type and the distance between hand and object are considered relevant for this event boundary, while the exact position of the object is not. In contrast the event boundary 'object is consumed' has the biggest variance for object color (3095) and small variances for the object's position in the hand-centered frame (x-wise 22.9, y-wise 19.3) and for global object position (x-wise 48.7, y-wise 36.5). Here the color of the object is irrelevant because both food objects can be consumed and they differ strongly in color (color difference = 100 in our simulation). Instead, the exact object position is relevant.

Conclusion

Inspired by the event segmentation theory and its schematic model put forward in Zacks et al. (2007), we have developed a computational, motor-grounded event segmentation model. Previous work has shown that statistical analyses of visual changes can be used to categorize segments of video sequences into distinct events (Buchsbaum, Canini, & Griffiths, 2011; Shi, Wang, Cheng, & Smola, 2008; Niebles, Wang, & Fei-Fei, 2008). Additionally and partially in contrast, our model has analyzed spatial, motor-dependent changes by learning predictive forward models and by using the learned forward models to detect event transitions based on a rigorous statistical measure of 'surprise'. Moreover, our system has shown that the learned predictive model cannot only be used to segment sensorimotor time series, but also to plan hierarchically goal-directedly. In the still rather restricted but continuous noisy environmental simulation, our system was able to identify events, which characterized particular object manipulations including 'moving without object contact', 'dragging a light object', 'dragging a heavy object', and 'moving while an object is moving'. Identified event transitions characterized boundary conditions including 'attaching to an object', 'kicking an object', and 'consuming an object'. Event boundary encodings identified those environmental factors that were critical for causing particular event transitions, such that the encodings can be thought of as conceptualizations of environmental interaction options, yielding object concepts, such as 'kickable', 'attachable', or 'draggable'.

Our model is closely related to advances in artificial intelligence and cognitive robotics. Calinon, Guenter, and Billard (2007) have put forward a system that learns a temporal Gaussian Mixture Model from behavioral demonstrations. Imitations of observed environmental interactions were executed using Gaussian mixture regression, focusing control on the relevant interaction aspects. Segmentation and higher level planning, however, were not addressed. Other work has predefined partitions over continuous subspaces during which a particular motor skill could be activated (Konidaris, Kaelbling, & Lozano-Perez, 2014). Partitions were computed by a global clustering algorithm. In contrast, our system learns to partition its environment by means of local measures of surprise based on developing forward models. Moreover, our system factorizes its developing model such that is becomes able to identify those environmental properties that are critical to bring a particular event about.

In sum, the proposed computational model offers an algorithm that can develop suitable event segmentations online from sensorimotor experiences with the environment. The model suggests that EST may be applied to structure own motor behavior and to identify those sensorimotor signals that are critical to accomplish particular environmental manipulations. We are currently working on extending the framework to be able to also solve non-linear control challenges in more complex scenarios in virtual realities.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–600.

Barsalou, L. W., Breazeal, C., & Smith, L. B. (2007). Cognition as coordinated non-cognition. *Cognitive Processing*, *8*, 79-91.

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1655). doi: 10.1098/rstb.2013.0480

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262 - 280.

Buchsbaum, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. In *Annual conference of the cognitive science society* (p. 3162-3167).

Butz, M. V., Sigaud, O., & Gérard, P. (2003). Internal models and anticipations in adaptive learning systems. In M. V. Butz, O. Sigaud, & P. Gérard (Eds.), *Anticipatory behavior in adaptive learning systems: Foundations, theories, and systems* (pp. 86–109). Berlin: Springer.

Butz, M. V., Swarup, S., & Goldberg, D. E. (2004). Effective online detection of task-independent landmarks. In R. S. Sutton & S. Singh (Eds.), *ICML'04 workshop on predictive representations of world knowledge.*

Calinon, S., Guenter, F., & Billard, A. (2007). On learning, representing, and generalizing a task in a humanoid robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *37*, 286–298.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science*, *36*, 181-253.

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? the pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, *17*, 202 - 209.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*, 293 - 301.

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*(598). doi: 10.3389/fnhum.2013.00598

Hohwy, J. (2013). *The predictive mind*. Oxford, UK: Oxford University Press.

Konidaris, G., Kaelbling, L., & Lozano-Perez, T. (2014). Constructing symbolic representations for high-level planning. *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1273-1280.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision*, *79*, 299–318.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*, 613–643.

Shi, Q., Wang, L., Cheng, L., & Smola, A. (2008). Discriminative human action segmentation and recognition using semi-markov model. In *Computer vision and pattern recognition. cvpr 2008. ieee conference on* (pp. 1–8).

Sigaud, O., Butz, M., Pezzulo, G., & Herbort, O. (2013). The anticipatory construction of reality as a central concern for psychology and robotics. *New Ideas in Psychology*, *31*, 217 - 220.

Sigaud, O., Butz, M. V., Kozlova, O., & Meyer, C. (2009). Anticipatory learning classifier systems and factored reinforcement learning. In G. Pezzulo, M. V. Butz, O. Sigaud, & G. Baldassarre (Eds.), *Anticipatory behavior in adaptive learning systems* (p. 321-333). Berlin: Springer.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181-211.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*, 201–216.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*, 273–293.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*, 3–21.

Information-Seeking, Learning and the Marginal Value Theorem: A Normative Approach to Adaptive Exploration

Andra Geana (ageana@princeton.edu)[*+] Robert C. Wilson (bob@email.arizona.edu)[°]
Nathaniel Daw (ndaw@princeton.edu)[*+] Jonathan D. Cohen (jdc@princeton.edu)[*+]

[*]Department of Psychology/Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540 USA
[°]Department of Psychology, University of Arizona, Tucson, AZ 85721

Abstract

Daily life often makes us decide between two goals: maximizing immediate rewards (exploitation) and learning about the environment so as to improve our options for future rewards (exploration). An adaptive organism therefore should place value on information independent of immediate reward, and affective states may signal such value (e.g., curiosity vs. boredom: Hill & Perkins, 1985; Eastwood et al. 2012). This tradeoff has been well studied in "bandit" tasks involving choice among a fixed number of options, but is equally pertinent in situations such as foraging, hunting, or job search, where one encounters a series of new options sequentially. Here, we augment the classic serial foraging scenario to more explicitly reward the development of knowledge. We develop a formal model that quantifies the value of information in this setting and how it should impact decision making, paralleling the treatment of reward by the marginal value theorem (MVT) in the foraging literature. We then present the results of an experiment designed to provide an initial test of this model, and discuss the implications of this information-foraging framework on boredom and task disengagement.

Keywords: exploration, explore-exploit tradeoff, information-seeking, decision making.

Introduction

All organisms face the frequent need to decide between persisting with one behavior (and the known rewards it brings), or switching to another. This tradeoff is well documented in the literature: stay-or-switch behavior has been studied in humans (Behrens et al., 2007) and non-human animals such as primates (Pearson et al. 2009), birds (Krebs, Kacelnik & Taylor 1978), rodents, and even non-vertebrates (Gallistel 1990) including the extent to which this follows optimal sampling strategies (Goldstone & Ashpole, 2004; Daw et al. 2006). When examining such decisions, it is helpful to distinguish between at least two types of circumstances under which an organism might choose to persist in or change its behavior: one involves situations in which rewards are which rewards are known up to stochasticity but either changing (as when foraging from a depleting patch) or varying in quality across options (as in encountering a series of candidate prey), so the decision whether to switch to other alternatives is a way to maximize current reward rate. The other involves situations in which the options' values are imperfectly known (as in bandit tasks) so that switching behavior may not yield immediate benefit, but may provide new information that can support learning and improvements in reward-rate over the longer term.

The first scenario has been extensively studied in the foraging literature. When choosing among behaviors with different reward opportunities (e.g. foraging patches) that are progressively depleting, or in circumstances in which (estimable) changes can happen outside the local environment, it is optimal to switch behavior when it is estimated that the value of the current behavior falls below the mean expected value of the available alternatives (Krebs & Inman, 1992). This policy can be shown to optimize immediate reward rate, as described by the marginal value theorem (MVT, Charnov, 1976), and numerous studies have found that animals' foraging behavior approximates this (Krebs, Kacelnik & Taylor 1978).

Most foraging scenarios of this type occur in environments with well-specified rewards, in which uncertainty usually stems from stable variance or hazard rates (risk), so it is possible to incorporate it into reward expectations through estimates of expected utility and the switching policy given by the MVT is optimal asymptotically (following all possible learning). However, in many circumstances reward opportunities may not only be stochastic, but the properties of this stochasticity may be unknown (Payzan-LeNestour et al., 2013). That is, uncertainty may stem from ambiguity rather than risk. In such circumstances, sampling the environment can provide information that, even if it is associated with immediate sacrifices in predictable reward rate, can be used to learn about the environment and improve reward rate over the longer term. We refer to such information-seeking as "exploration," to distinguish it from foraging choices that we define as the pursuit of alternative behaviors based on decisions involving reward opportunities with known distributional properties (e.g., mean and variance)[1].

The drive toward exploration has been well-documented in both human and animal literatures (e.g., Cohen, McClure & Yu, 2007), and there is rich evidence that under many scenarios, organisms will choose to sample the environment for useful information even at the cost of current or predictable reward (Wilson et al., 2014). This is the case, for instance, when we choose to try out the new special at a restaurant instead of sticking with our favorite dish, or when

[1] We make this qualitative distinction largely for the purpose of clarity, and to guide formal treatment, fully recognizing that real-world circumstances almost certainly fall along a continuum between these extremes and involve a mix of these two type of decisions.

we choose to watch a new show instead of rewatching an episode of an old favorite. In this way, exploration is different from foraging: though both involve the choice between persisting with the current action or switching away to something else, exploration is geared toward acquiring new information, while foraging is geared toward acquiring predictable sources of reward.

However, exploration of this sort has been studied largely in the context of "bandit" tasks – choice among a fixed set of options, whose properties must be learned from sampling – and not in the serial switching scenarios modeled by the MVT. Here, we propose a formal model for exploration that parallels the formulation of the MVT for reward, but applies it to maximizing information alongside reward. In particular, we augment the patch foraging scenario to more explicitly reward information gathering – this models, for instance, development of expertise when encountering a series of options, for instance, learning a trade over successive jobs or improving one's dating skills – and study the behavior of optimal agents. We then present the results of an experiment designed to provide an initial test of this model, and discuss the implications of this information-foraging framework on boredom and task disengagement.

A Normative Model of Exploration

Paralleling circumstances to which the MVT has been applied, we model an environment consisting of local reward patches that offer different reward rates, with the model agent free to either stay within a patch to reap reward (exploit), or switch away to search for other patches, in this case with only partially or unknown characteristics (explore). Each patch is comprised of an environment in which the agent can earn rewards by making accurate predictions of the outcome of a stable stochastic process.

Upon "entering" a patch, the agent does not know the parameters of the stochastic process, but these can be learned by sampling. On each time step spent within a patch, the agent makes a prediction, and receives a reward proportional to the accuracy of the prediction. The longer an agent spends learning about a patch, the better its estimates of the underlying structure can become, and higher the reward it can receive. This distinguishes this task environment from the environment assumed in most studies of foraging: here, the patch becomes more rewarding with the passage of time (and sampling), rather than depleting.

An additional important assumption is that the properties of patches are not independent of one another, but rather reflect properties of a global environment from which they are drawn. Thus, within limits, sampling a local patch can provide information that is relevant to other patches. This is a property of many real-life environments, in which humans sequentially sampling different "patches" learn about the local structure while simultaneously learning about an overarching global structure (Diuk et al 2013). For instance, when going apple-picking, we learn about the quality and availability of fruit in each individual tree (so we could choose to move from a smaller, poorer tree to a better one), but at the same time we are also learning about the overall qualities of the orchard, so next time we go apple-picking me might choose an altogether different orchard.

Under this framework, exploiting a local patch obtained increasing local reward (fig. 1A), but exploring many local patches helps the agent learn the global structure faster, which would in turn allow it to make better choices earlier in the local patches. Depending on goals, therefore, it could be optimal to quit a local patch (even if it was yielding a high reward) and move to another patch, at the potential cost of a lower reward, for the sake of learning about the global environment that could improve returns in the future.

This local/global structure allowed us to model an environment with a distribution of available information paralleling the distribution of available reward in standard foraging environments. In other words, each patch held not only reward (which increased with time spent in patch), but also information (which decreased with time spent in patch). This generated a canonical explore-exploit decision tradeoff: maximize known rewards by staying within a patch (exploitation), or switch patches to acquire information (exploration). We constructed a model of this process, and used it to compare performance with a pure exploitation strategy, a strict foraging model, and human behavior in an empirical version of the task.

Model Assumptions

Each patch represented a stochastic environment in which the agent could earn rewards by making accurate predictions. Each patch had a hidden distribution with mean μ_i and standard deviation σ (which was the same between patches). On each time step spent inside the patch, the agent had to make a prediction relating to this distribution. The agent's reward $r_{t,i}$ was proportional to the accuracy of the prediction (for a similar task design, see Nassar et al.'s (2010) "estimation task"), according to

$$r_{t,i} = \rho - PE_t, \qquad (1)$$

where ρ represented the maximum amount of reward that an agent could earn (if its predictions were fully accurate), and PE_t represented the prediction error, computed as the difference between the agent's prediction *Pr* and the actual number generated in the patch on time step *t*:

$$PE_t = Pr_t - N(\mu_i, \sigma) \qquad (2)$$

Figure 1A shows the increase in reward with time spent in patch. An agent could spend as long as it wanted exploiting a patch, but each patch had a fixed chance of termination λ, meaning that on every time step the patch would end with probability λ, and continue with probability $(1 - \lambda)$.

As explained in the previous section, all patches were connected under a higher-level, global structure. In other words, the underlying patch distribution parameter μ_i came from a global distribution with a (fixed) grand mean *M* and standard deviation *S*. Exploiting a local patch obtained

increasing local reward (fig. 1A), but decreasing information (fig. 1B), while exploring more local patches helps the agent learn the global structure faster. There was also a global reward *R* associated with learning the global mean *M*. Our model tracked several quantities of interest as an agent exploited a patch with the above structure; for simplicity, we approximate the hierarchical estimation problem with nested error-driven updates. First, at each time step it computed an estimate of the local mean for patch *i* at time *t*, $\mu_{i,t}$, as the average of all data points $x_{i,t}$ observed in that patch up to the current time:

$$\overline{\mu_{i,t}} = \frac{\sum_t x_{i,t}}{n} \quad (3)$$

which can also be written in terms of the prediction error *PE* and a learning rate of 1/n, as

$$\overline{\mu_{i,t}} = \overline{\mu_{i,t-1}} + \frac{1}{n} * PE_t \quad (4)$$

(Given the structure of the task, the optimal prediction at any time step was the current estimate of the mean, $\overline{\mu_{i,t}}$, and our model assumed that the agent would always predict that mean). In addition to tracking the mean estimate for the patch, the model also used error-driven updating to estimate the variance of the local patch,

$$\sigma_{i,t}^2 = \sigma_{i,t-1}^2 + \frac{1}{n}\left(\frac{(n-1)*PE^2}{n} - \sigma_{i,t-1}^2\right) \quad (5)$$

which allowed computation of how informative each new data point was, in terms of how much it could reduce variance about the local patch. As the above equation shows, the informativeness of each new data point decreased proportionally to *1/n* (see fig 1B). The model also tracked an estimate of the global mean, in terms of the history of visited patches. Each final mean estimate, μ_i, for the distribution within a patch served as an additional data point for inferring the grand mean M , in the same way that each within-patch data point served to estimate μ_i.

Crucially, the model assumed that upon first entering a new patch, the initial prediction regarding the distribution of that patch (essentially, the prior, before any data points from that patch were observed) was set to the current estimate of the global mean *M*. This provided a way to quantify the value of information in each patch, in terms of expected reward, as the estimated improvement in initial predictions on future patches. That is, the better the estimate of the global mean, the better the agent could do, on average, when entering a new patch. This is because the mean for each patch was drawn from a distribution centered on the global mean, and thus the optimal initial guess (prior) for a given patch was the global mean. Thus, at each time step *t*, the value of acquiring one extra data point in the current patch *i* could be estimated in terms of how much it improved future predictions (i.e. how much closer it moved them to *M*), relative to how much sampling a new patch would improve future predictions. Accordingly, we approximated the information value of staying in a patch with

$$V_{stay} = \frac{1}{(N-1)(n-1)} PE_{i,t} \quad (6)$$

while the value of leaving the patch was

$$V_{switch} = \frac{1}{(N-1)} PE_{i+1,t} \quad (7)$$

where *N* was the current number of patches exploited so far, *n* the current data points observed in the current patch, $PE_{i,t}$ the next estimated prediction error within the current patch, and $PE_{i+1,t}$ the next estimated prediction error assuming that the agent explored a new patch. This relative value between staying (exploiting) and switching (exploring) depended therefore on the current position within the game (*n*), the current position within the patch (*N*), and the two variance estimates for the patch mean (σ^2) and the global mean (S^2), as those variance estimates were used to compute the two prediction errors of interest in the above equation. Given a fixed number of available timesteps, the assumption that the agent could not return to a patch once it switched away, and values for the rewards ρ, *R* and the termination probability λ, we used dynamic programming to compute the value of staying or going at every time step (combining the immediate rewards for estimation within a patch with the approximate future value of learning the global mean, *V*), as a function of how well both the local patch and the global mean were known.

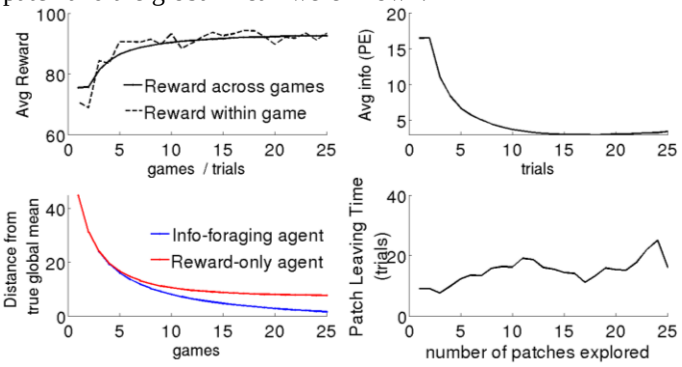


Figure 1: Model results. A. Reward increases with time spent in a game (dotted line), and average reward in a game increases as more patches are explored (solid black line). B. The amount of information obtained from each new sample decreases with trials spent in a game (i.e. patch). C. Information-foraging agent that leaves patches early (blue) learns the global structure parameters earlier than an agent that exploits a patch for all its reward (red). D. Model predicts later patch leaving times as a function of how well the world is known (i.e. how many patches have already been explored).

Results

Figure 1 shows model results (5000 simulations of twenty-five games each, with a maximum of twenty-five trials per game). As per the task structure, reward increases with time in game (fig. 1A, dotted line); furthermore the reward

increases across games as global structure is learned (solid line). The change in prediction error (PE) as well as the reduction in uncertainty from each new data point decrease with time spent in game (fig 1B). Compared to an agent that exhausts all available trials in a patch, our "information-foraging" agent that leaves a patch depending on the relative informativeness of an extra data point within the patch versus a data point in a new patch showed faster learning of the global mean (i.e., learned it both in a shorter number of trials, and approached it faster across games, fig. 1C). Under certain model parameters, it also earned more average reward when compared to a model that only takes into account local reward (i.e. current prediction error). Looking at model predictions for the optimal time to switch away from a patch, as a function of both mean estimated variance of the global mean and, and as a function of time in game (though the two measures are somewhat correlated), the model predicted longer dwell times later in the game, when the uncertainty about the global mean had been significantly reduced (figure 1D).

Experiment 1

The following experiment was designed to test model predictions from our information-foraging model. The main prediction from the model, given the task structure presented to the subjects, was that they would quit high-reward, low-information patches early when the value of gaining information from new patches was higher (i.e., near the beginning of the task, when they had not learned much about the global environment), but spend longer and longer in patches as the usefulness of new information decreased.

Methods

Participants Twenty-five participants were recruited from the Princeton community. They gave informed consent, and were compensated for their time at a rate of $12/hour, plus a bonus of up to $5 for performance.

Task Participants played a game in which they controlled a virtual archer that made his way through enemy territory toward a castle (fig 2A, below). The goal was to defeat an "evil overlord," and the ability to do so could be enhanced by facing waves of "minions" trained by the overlord, and learning about their behavior as an indicator of his. Minions appeared sequentially on the screen, the archer had to fire an arrow to hit the minion, and doing so earned one point. A hit and miss counter was available on the bottom left of the screen, indicating to participants how well they were doing.

Participants were informed that the archer had to confront seven waves of minions before facing the overlord. Each wave consisted of a maximum of thirty-five minions, appearing one by one, from the right of the screen (fig 1A), at a variable location. Participants could adjust the archer's firing position on each trial, to best anticipate where the next minion would appear. At the end of seven waves, the archer confronted the overlord, and had only one shot to either defeat it (i.e., aim the arrow accurately enough to hit it), or be defeated by it (i.e. miss). A reward of 30 points was available for defeating the overlord.

Participants were informed that each wave of minions was trained by a different commander, and that all of the commanders had been trained by the overlord. Commanders shared, but did not perfectly imitate the overlord's preference for location of appearance. Similarly. minions shared, but did not perfectly imitate their commander's location of appearance. These instructions reflected the generative properties of the environment to be learned: the location of appearance of each minion within a wave was drawn from a distribution with a fixed mean and variance, and the means for each wave were drawn from a distribution with the same variance and a mean equal to the preferred location of the overlord.

Before encountering each minion in a wave, participants had two options: They could choose to stay and confront that minion, or choose to "run away" by pressing the large "RUN" button at the top left of the screen. If they chose to run, that wave of minions would end, a screen appeared announcing a new wave (with a new distribution of locations, and they would then have the same two options for each minion in the new wave. Given this task structure, sampling within a wave could lead to progressive improvement in performance (and reward) for a given wave. However, learning the preference of the overlord (associated with a much larger reward) required an appropriate balance of sampling within and across waves, as the informativeness of each data point decreased (see fig. 1B) within a wave, and other waves provided additional information.

Importantly, participants were told that they had only one hundred and fifty arrows to use on the minions – this operationalized the finite number of steps in our model – so they would have to decide how to use those arrows in a way that would give them the best shot of defeating the overlord. The model predicted that they should use fewer arrows (confront fewer minions) in earlier waves, switching waves (i.e., exploring) to maximize information about average locations of the waves (as a predictor of the overlord). At the same time, the model predicted that, as information accrued, and future opportunities to do so diminished (i.e., task horizon shortened), they should use arrows more liberally to earn points within each wave (i.e., exploit).

Results Twenty-five Participants learned the task, as evidenced both by their increasing accuracy in targeting the minions, within a wave (figure 2B) and by the increasingly accurate first location estimate – i.e., change of hitting the first minion – in later waves compared earlier waves (significant linear trend, $F(1,6) = 9.42$, $p = 0.02$, figure 2C).

No participants attempted to defeat all minions in a wave. However, participants' average number of minions attempted within a wave increased in later waves compared to earlier waves (figure 2C). This equates to earlier quitting times earlier in the game. Average likelihood to "run" (i.e. quit the current wave and move on to the next one) increased, for all participants, as a function of the number of

minions they had confronted in the current wave (i.e the number of time steps they had spent there, fig. 2D).

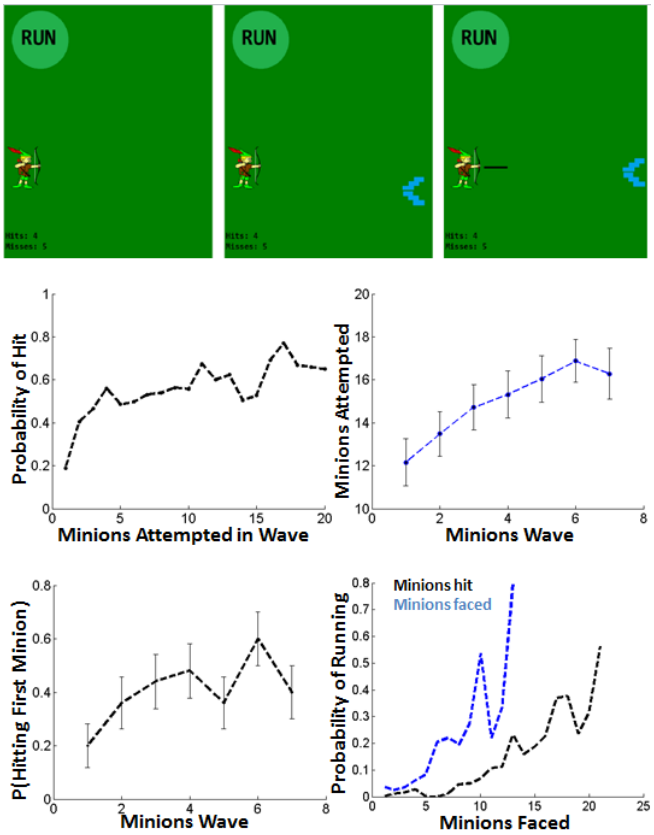


Figure 2: A. The archer task: participants adjust the position of the archer using the mouse, then release the arrow. The minion then comes from the right of the screen. B. Participants' chances of hitting a minion increase with time spent in a minion wave, as they learn to predict the locations of the minions better. C. Leaving times (i.e. the number of minions attempted) increase in later waves compared to earlier waves. D. The chance of hitting the first minion in a wave (i.e. the first sample of that particular wave's location) increases across the game. E. Participants are more likely to run after facing more minions (black line) and after hitting more minions (blue line).

Likelihood to run also increased with the number of minions actually defeated in the current wave (fig. 2E, blue line), which is well-correlated with how well the participants had learned that particular wave; comparatively, the likelihood to run was slower to increase as a function of faced minions (black line). Finally, in line with optimal performance, participants' accuracy in later waves improved on the first trial of a wave (before they got any data from the current minion wave), indicating that they generalized the knowledge about the structure of previous games to make better predictions in the current one (Fig. 2B).

Discussion

Building on ideas from optimal foraging theory and reinforcement learning, we proposed a normative model of adaptive exploration. The model is based on a variant of the Marginal Value Theorem (MVT; Charnov, 1976), in which explore-exploit decisions are driven by estimates of the relative information (rather than reward) associated with each option. The model balances the goals of learning about the properties of the local and global environments, and we showed that in doing so it is capable of optimizing overall reward rate. We tested, and found support for qualitative predictions of the model in an empirical study: human participants exhibited behavior consistent with information-seeking, and stay-leave (explore-exploit) decisions that were sensitive to estimates not only of current reward, but also current information, and that these were integrated into their decision to stay or leave.

The model predictions are consistent with theories of intrinsic motivation stating that the drive to explore arises from an innate need to interact efficiently with the environment (Deci & Ryan, 1985; White, 1959), as well as with the notion of "flow" and the optimal arousal theory of motivation, according to which organisms seek to balance an internal need for optimal levels of information and stimulation (Carrol, Zuckerman & Voegel, 1982; Csizentmihalyi, 2000). Furthermore, results here show that quitting a current high-reward but low-information patch can in fact still lead to higher overall reward than staying in the uninformative patch. This is consistent with model findings that average within-game reward increases across games, as the global structure is learned (Fig 1A). Similar findings have been presented previously in the machine learning literature, in studies showing that artificial agents capable of penalizing a too-well-known option's value can learn a complex grid environment faster and earn higher overall reward (Simsek & Barto, 2006). Ecological models of optimal foraging have also mentioned the "penalty of ignorance", i.e. the benefits that a forager could lose by not improving its information about the world over time (Olsson & Brown, 2006). However, to our knowledge, the model we present here is the first to cast exploration in terms that parallel the role of reward in optimal foraging theory. In this respect, the model provides a bridge between the closely-related literatures on foraging and exploration, and a path toward theoretical integration.

Along similar lines, our model provides a mechanistic, and potentially normative account of the phenomenology associated with boredom. The link between exploratory behavior and boredom has been suggested many times in both human and animal literature (Fowler 1959; Cohen, McClure & Yu, 2007; Meagher & Mason 2012; Kurzban, 2013). Our model formalizes this idea, suggesting that boredom might be considered as signaling the value of exploration; that is, that information provided by the current behavior is below what can be expected from alternatives, and therefore that a switch in behavior is warranted.

Consistent with this suggestion, we have found in a separate set of experiments that boredom is negatively correlated with the rate of information acquisition (learning) in the

current task, is influenced by context, and that participants are willing to sacrifice reward in order to avoid boredom and increase the rate of information acquisition and learning. These observations are consistent with ones from the study reported here. At least initially, participants chose to switch away from a given wave of minions, even as their performance improved, reflecting a valuation of information and learning (and the diminution of "boredom") over immediate reward. This echoes a pervasive pattern of behavior in video games, and explains the need for "levels" to maintain gamers' interest — a phenomenon that is consistent with the current model. Interestingly, however, participants in our experiment chose to stay with a wave of minions *longer* as the task progressed; that is, they appear to have become *less* "bored" as overall time-on-task increased. This seemingly counterintuitive effect was predicted by the theory: as the task horizon shortened, the worth of information diminished relative to immediate reward, thus driving participants to persist (exploit) as the task neared an end.

It is important to note that, from the vantage of the model proposed, the value of exploration is dependent on the structure of the environment (e.g., the amount of time available, as well as the difficulty of the learning problem) and on how well the agent has learned the environment. Our findings apply to the pre-asymptotic phase(s) of learning, when gaining information from exploration can contribute to forming better representations of the environment and ultimately to better strategies for gaining reward. If the world were fully known, the same findings would not hold; however, given the complexity of the real world, it seems likely that organisms spend a considerable fraction of their time in pre-asymptotic phases of learning. That certainly appears to be the case for our collective understanding of how natural agents learn about their environment, and along similar lines, we hope that the work we describe here offers a useful new path for exploring this domain of understanding.

References

Behrens. T.E.J., Woolrich, M.W., Walton, M.E. * Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience* 10: 1214 – 1221

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem.*Theoretical population biology*, *9*(2), 129-136.

Cohen, J. D., McClure, S. M. & Yu, A. J. (2007).Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society: Biological Sciences. 362*, 933–942

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441, 876 – 879

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience*, *33*(13), 5797-5805.

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged mind defining boredom in terms of attention. *Perspectives on Psychological Science*, *7*(5), 482-495.

Fowler, H. (1967). Satiation and Curiosity: Constructs for a Drive and Incentive-Motivational Theory of Exploration. *Psychology of learning and motivation*, *1*, 157-227.

Gallistel, C.R. (1990). *The Organization of Learning*. Cambridge, MA: Bradford Books/MIT Press.

Goldstone, R. L., & Ashpole, B. C. (2004). Human foraging behavior in a virtual environment. *Psychonomic bulletin & review*, *11*(3), 508-514.

Hill, A. B., & Perkins, R. E. (1985). Towards a model of boredom. *British Journal of Psychology*, *76*(2), 235-240.

Kaelbling, L., Littman, M., & Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237–285.

Krebs, J. R., Kacelnik, A. & Taylor, P. (1978) Tests of optimal sampling by foraging great tits. *Nature* 275, 27–31Lai & Robbins 1985

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(06), 661-679.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, *30*(37), 12366-12378.

Olsson, O., & S Brown, J. (2006). The foraging benefits of information and the penalty of ignorance. *Oikos*, *112*(2), 260-273.

Payzan-LeNestour, E., & Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: Estimation uncertainty and "unexpected uncertainty" both modulate exploration. Frontiers in Neuroscience

Pearson, J. M., Hayden, B. Y., Raghavachari, S., & Platt, M. L. (2009). Neurons in posterior cingulate cortex signal exploratory decisions in a dynamic multioption choice task. *Current biology*, *19*(18), 1532-1537.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior (SAB90)*.

Şimşek, O., & Barto, A. (2006). An intrinsic reward mechanism for efficient exploration. *Proceedings of the 23rd international conference*, 833-840.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Learning that numbers are the same, while learning that they are different

Jon A. Willits (jwillits@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St. Bloomington, IN 47408 USA

Michael N. Jones (mnjones@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St. Bloomington, IN 47408 USA

David Landy (dlandy@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St. Bloomington, IN 47408 USA

Abstract
It has been suggested that the way that number words are used may play an important role in the development of number concepts. However, little is currently known about the overall ways in which number words are used in child-directed speech. To address this, we performed an analysis of how number words are used in the CHILDES database. We looked at four statistics: 1) lexical frequency, 2) contextual diversity, 3) word co-occurrence, and 4) distributional similarity, to see if these distributional statistics suggest why some aspects of number acquisition are easy and others are hard, and if these statistics are informative about specific debates in number acquisition. We found that that are many important differences in how small and large number words are used (such as differences in frequency, co-occurrence patterns, and distributional similarity), differences that may play an role in shaping hypotheses about children's acquisition of number concepts. **Keywords**: number representation, language acquisition, concept acquisition, statistical learning, corpus analyses

Introduction

Numbers are an odd category of things. Although often treated as objects themselves, numbers can be observed only as relational features—features of sets of objects. Their structure is as regular one could hope to find in a category, and they are extremely important in modern life. But numbers are very challenging for children and students to understand and use, and many simple-to-phrase questions about the counting numbers remain unanswered even by professional mathematicians. In short, numbers are among the most frequent and regular, and at the same time most abstract and difficult, of the concepts in a young child's environment. Here, we focus on the role of the statistics of the child's linguistic environment to interrogate two key challenges for the learner of numbers: how do we learn that number words form a category, that they are all the same kind of thing? But in contrast, how do children also learn how each number word is different from the others?

One important mechanism by which children can learn about semantic categories is by learning about the distributional patterns of the words that refer to them. Much research has demonstrated that children can learn a lot about a word's semantics, based on the word's pattern of co-occurrence. For example, toddlers can learn whether a verb's meaning is transitive or intransitive by counting the number of nouns that occur in the same utterance as the verb (Gleitman, 1990; Yuan, Fisher, & Snedeker, 2012). Toddlers can also learn about the meaning of a novel noun if it occurs in distributions that are informative about the category to which it belongs (Lany & Saffran, 2011).

If children can infer semantic content from words' lexical distributions – words' frequencies, the words with which they co-occur, and the words to which they are distributionally similar, what can they learn about number concepts based on how number words are used in child-directed speech? What might this say about the development of children's number concepts and number word usage?

The Conceptual Structure of Number

The cognitive structure of numbers is a matter of intense current debate. At this point, a general consensus has emerged that the processing of numbers under about 4 is qualitatively different from that of numbers roughly 4 and over (Feigenson, Dehaene, & Spelke, 2004). Numbers over 4, on this account, are processed by a general magnitude system, which discriminates relatively numerous sets from relatively sparse sets, but does not have distinct ways of capturing precise numbers. This system is generally thought to be 'log-scaled' in the sense that the reliability of discrimination between two sets is a function of the ratio between their numerosities, regardless of the absolute numerosity. At the same time, numbers under 4 are processed by a very different mechanism that individuates particular objects, but may not be all that involved in estimating magnitudes. This mechanism has been tied to subitization (Trick & Pylyshyn, 1994; Revkin et al, 2008), and to object tracking (e.g., Spelke & Kinzler, 2007). On some accounts, linguistic experience is used to coordinate these two systems, bootstrapping a new "positive whole number" system (Carey, 2009), which combines representation of exactness with large magnitudes, comprising concepts like "exactly 7".

The general dual-systems account outlined above has been applied to several behavioral paradigms beyond subitization. First, when asked to give an experimenter a certain number of objects (the *give-N* task), children tend to show discrete changes in capacity, from being able to give one object only successfully, to being able to give one or two, to being able to give up to three, before passing into a final "four or more" stage. A common interpretation is that the first three stages are associated with the small-number processing system, while the final stage reflects invocation of the full count-list and magnitude system into the task.

The Linguistic Structure of Number

Number words are an extremely common part of our

language. For example, in the CHILDES database (MacWhinney, 2000) examined below, the word *one* is slightly more common than the word *mom*, while *two* is slightly more common than *dad*. The words *one* and *two* taken together comprise almost exactly 1% of all the words in CHILDES. And while these two number words are certainly the most common, the remaining number words [0, 3-20, 30, 40, 50, 60, 70, 80, 90], *hundred*, *thousand*, and *million* constitute about an additional 0.5% of the total words in the corpus, a truly large amount of tokens for such a small number of types (31).

The structure of language has been implicated in particular aspects of number learning. For instance, phonological irregularity has been argued to slow the learning of teen words in English (Fuson & Kwon, 1991), and numbers between 20 and 40 in languages where such words are partially irregular or inverted (Brysbaert et al., 1998). Processing of numbers in large ranges (one thousand to one billion) may also be affected by how these words are used (Landy et al., 2013; Landy et al., 2014).

Although number-elaboration accounts rely on language to connect number experiences, the specific ways that numbers are structured in language are less thoroughly explored than number-related behaviors. In particular, the information available in the morphemic structure of the speech stream—though useful in distinguishing categories in language—has to our knowledge not been systematically explored in the case of number words. Although the literature often refers to a "count list" (Carey, 2009), it is unclear how much of a role this list plays in the number-related linguistic experience of young children, though compelling recent work suggests that counting and label-present sets with children have a positive impact on number learning (Gunderson & Levine, 2011). More generally, we do not yet know how the striking behavioral distinctions between numbers under and over 4 are mirrored in the structure of child-directed language, nor how the statistics of language differentiate numbers over and under 10.

Starting from the basis that number words are very common, that they are unevenly distributed and structured, and that they are implicated in number concept learning, we ask what information about numbers exists in the statistical structure of the speech stream, that could potentially help children understand numbers' complex category structure. In particular, we see the central challenge as one of learning that number words form a category, but at the same time learning that and how number words differ from one another. To the degree that number words are used similarly, identifying them as a category should be facilitated. At the same time, the easier it is to learn that items belong to the same category, the harder it should be – at least for a time – to distinguish members of the category from one another (Rogers & McClelland, 2004).

Analysis and Data Set

To address some of these questions and provide an introductory analysis of what children can learn about number words based on their distributional patterns in child-directed speech, we analyzed these distributional patterns in the CHILDES corpus, a collection of transcripts of caregivers and family members interacting with children in a wide range of situations and contexts (MacWhinney, 2000). For our analyses, we built a composite corpus containing all documents involving typically developing children from American English speaking households up to 72 months of age. This resulted in a corpus containing 4,561 documents, 32,580 word types, and 7,016,488 million word tokens. Compared to estimates of the number of words children typically hear (Risley & Hart, 1995), the size of this corpus represents approximately 5-10% of the input a typical child hears by age six, thus providing a very good model of the available input to a child.

We performed four analyses of how number words are used in the CHILDES corpus, to answer the following questions: First, what is the frequency of number words, how do frequencies between different numbers compare to each other and to non-number words, and how do these frequencies change over the course of development? Second, what is the contextual diversity of number usage? Do some numbers occur in broader or narrower ranges of contexts? And how does this change across development? Third, what are the other words with which number words co-occur, and is there a difference in co-occurrence patterns for different kinds of numbers? Finally, how do number words vary in terms of their distributional similarity to other number words? Which numbers can most easily be categorized as a number word based on their distributions? Which numbers are most prototypical? Answers to questions like these can begin to help us understand how the usage of number words contributes to the development of number representations.

Number Frequency and Contextual Diversity

One of the most basic factors that may affect the learnability of number words is how frequently they are used, how frequency varies for different number words, and how frequency varies across development. For an empirical answer to these questions, we took the American English CHILDES corpus and subdivided it into 10 sub-corpora, designed to reflect children's cumulative experience at evenly divided points. Thus, we created corpus M12 composed of all corpora with target children of up to 12 months of age, M18 for all composed of all corpora with target children of up to 18 months of age, and so on, up to M72, composed of all corpora for children up to 72 months of age. We then calculated the frequency of a large set of number words, and converted this frequency count into a "parts per million" count (i.e. divided the frequency count by the corpus size, and then multiplied that proportion by 1,000,000) designed to control for differences in sizes of the specific sub-corpora. The results of this analysis are shown in Figure 1.

Figure 1 demonstrates a number of important properties about the frequencies of number words. As noted above, the frequency of the words *one* is an order of magnitude above the other number words. This reflects the importance of the concept of *one*, but also the fact that one is used in a number of ways that aren't explicitly numerical. However, even these non-numerical uses of one are still informative about numerical content (e.g. "*another one*", "*the one I love*",

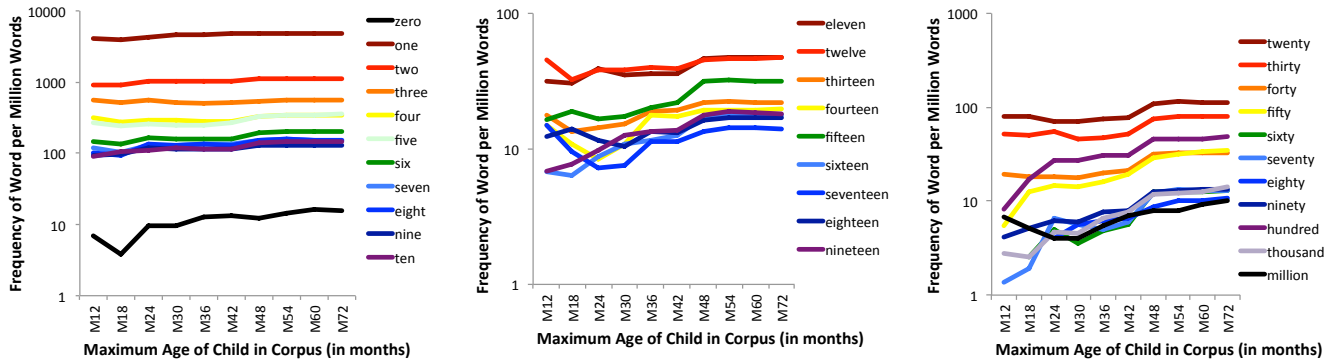


Figure 1. Cumulative frequency per million of various number words across development, plotted on a log10 scale (e.g. M12 is in all corpora for all kids up to 12 months of age, M72 is for all corpora for all kids up to 72 months of age). Note that across the three sub-figures, the scale of the y-axis changes, reflecting the large differences between big and small numbers.

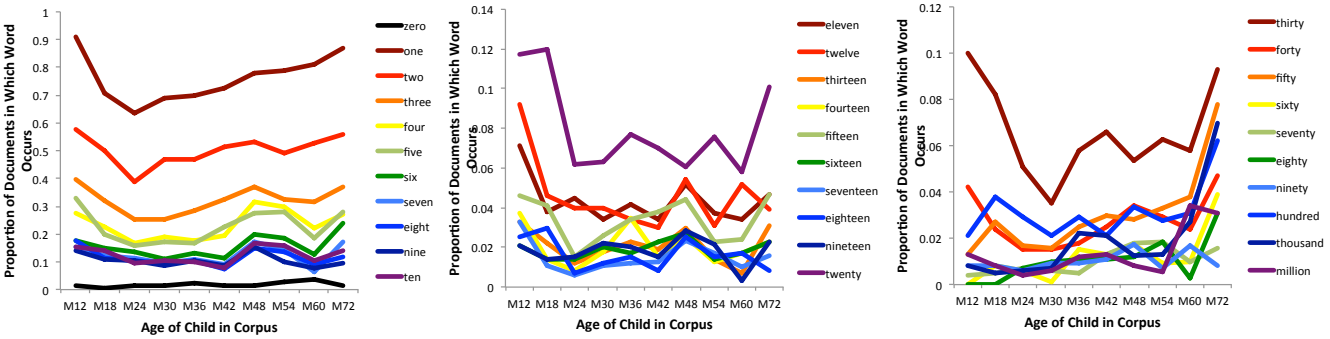


Figure 2. The proportion of documents in which the word occurs, for various number words across development. Note that across the three sub-figures, the scale of the y-axis changes, reflecting the large differences between big and small numbers.

"*one of the toys*", "*it is a good one*"). Beyond one, the frequency of the words are roughly in cardinal order. Of the "small" numbers, 2 is much more frequent than 3, which is much more frequent than 4 and 5, which are much more frequent than everything else. The rank orderings are all very sensible (especially when one considers usages like dates, times, money, and other measurement contexts). Consistent with previous cross-cultural analyses of number word frequencies, small numbers were exponentially more frequent than large numbers (Dahaene & Mehler, 1992).

One notable and surprising finding from these analyses was how little the frequency ratios of the number words changed across the different age samples in the corpus. All numbers become slightly more frequent over time (though for some numbers like *one* this difference is very small). Words for larger numbers do make up a larger proportion of the number words as the children in CHILDES increase in age, although this difference is not as large as it appears given the log-scaled differences between big numbers and small numbers. For example, the numbers one through ten make up 95.0% of all number word usage in the sub-corpora from kids age 12 months and below, whereas they make up 90.7% of the usage for kids age 48 months and above.

We also investigated the contextual diversity (CD) of number word usage, operationally defined as the proportion of different CHILDES documents in which the number word occurred. *High* CD has been found to be a strong predictor (above and beyond word frequency) of lexical access in adults (Adelman & Brown, 2008) and of age of acquisition in children (Hills et al., 2013). However, lower CD has been linked to faster learning of words in a number of behavioral experiments (Akhtar & Tomasello, 1997). Thus, the exact role of contextual diversity on word learning is obviously complex and not yet well understood.

We found that contextual diversity of number words in CHILDES was highly correlated (r = 0.867) with their frequencies (as is always true), with words like *one* and *two* much more diversely used than larger numbers. However, it is also notable that CD of number words showed differences across the corpora from children of different ages, with trajectories that are intriguing with regard to possible developmental trends and the role that usage diversity plays on word learning. For almost all numbers, the contextual diversity as relatively higher in speech to children 12 months and younger, dropped to a lower rate in corpora for children in the middle range of the CHILDES sample, and then rose again in corpora for older children. This is consistent with the idea that when children are first introduced to a word, caregivers use that word in a very high number of contexts, then selectively use that word in a much narrower range of contexts (facilitating the acquisition of the word's specific meaning and core properties), before reverting to a more base-rate level of usage of the word, in terms of diversity of contexts (Tomasello, 2003).

Number Word Co-occurrence

The second set of statistics we investigated were word co-occurrence patterns of the number words, to see if these would be helpful to kids when learning about number word meaning. To do this, we calculated the co-occurrence of number words with other words within a 12-word window. The most frequently co-occurring words with various

number words are shown in Table 1.

When looking at co-occurrence of number words with all other words in the corpus, interesting patterns emerge, which are suggestive of what children can learn about numbers based on these co-occurrence patterns. First, one of the most apparent facts is how similar all the number words were in terms of the words with which they frequently co-occurred. This is a big part of why (as we show in the next section) it would be easy to learn that number words belong to the same category. Next, one can clearly see how grammatical number and conceptual number are interrelated, with the number *one* co-occurring with the plural form "*s*" much less often than other numbers (of which "*s*" would be the most frequent collocation for many). Finally, it is apparent that many of the frequent collocations of numbers are other numbers (and numbers that are nearby in order), evidence of the extent to which counting and other ordering is highly frequent in child directed speech.

Table 1. Highest co-occurring words for various number words within a 12-word window in CHILDES. The morpheme "_S" is demonstrating how often the number word had the plural "s" in its 12-word window, when "s" is treated as a distinct morpheme in the corpus (i.e. broken off of the word to which it was attached).

one	**two**	**three**	**four**
you	_S	one	_S
one	you	_S	three
is	one	two	one
the	two	you	you
that	is	three	two
I	I	is	is
this	the	four	four
_S	that	the	give
eight	**eighteen**	**eighty**	**hundred**
is	twenty	eighty	a
you	is	is	and
eight	_S	you	you
seven	eighteen	and	_S
nine	you	five	is
I	I	I	I
six	seventeen	eight	the
the	the	the	to

The relationship involving the co-occurrence of number words with other number words is shown in for all pairs of numbers in Figure 3. In this Figure you can see many interesting facts about number word co-occurrence. First, the pattern of number words being more likely to co-occur with other number words of similar size is shown by the brightness along the diagonal. Especially interesting is how the breadth of this "counting: window grows as numbers get larger. *Three* is very likely to co-occur with *two*. *Four* is as well, though less so. *Six* and *seven* are barely more likely to co-occur with *two* then we would expect them to based on their baserate frequencies. However, while a simple model based on logarithmic scaling might assume that number words would continue to be more broadly associated as they increased in magnitude, we find instead that particular number ranges seem to form co-occurrence clusters. For numbers in the teens, for example, all the teens are highly likely to co-occur with each other, even 13 and 19. Many other idiosyncratic (but meaningful) highly probable collocations can be seen, including those that encode relations due to time (such as numbers like *six* and *twelve* with *thirty*) and dates (such as *nineteen* and *ninety*).

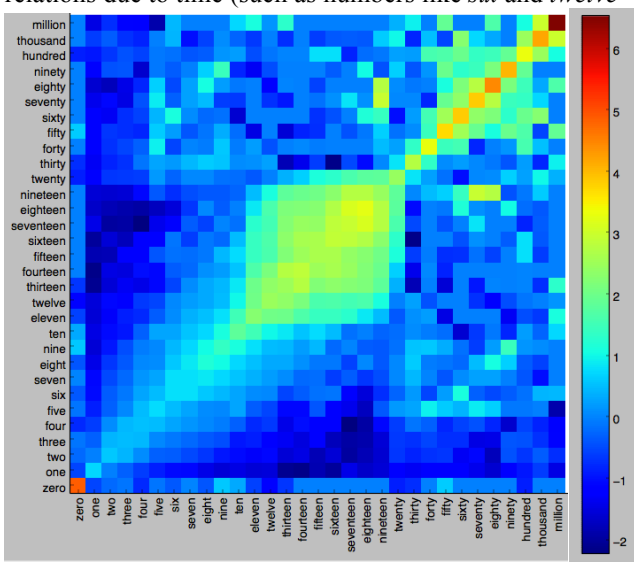


Figure 3. A heat map showing the pointwise mutual information (PMI) scores of number words with other number words at 72 months, demonstrating the degree to which they co-occur occur with each other more or less one would expect due to their base rate frequencies. A value of zero means the two co-occur exactly as often was one would expect, with values > and < than one meaning they co-occur 10 (or 100 for a score of +/- 2, or 1000 for a score of +/-3) times more or less than one would expect them to, given their word frequencies.

As a final analysis of word-occurrence patterns for numbers, and how important number words were to the structure of child-directed speech, we performed a principal component analysis (PCA) of the co-occurrence patterns between the 8,000 most frequent words in the corpus. We then looked at the extent to which number words loaded on the 30 highest components, shown in Figure 4.

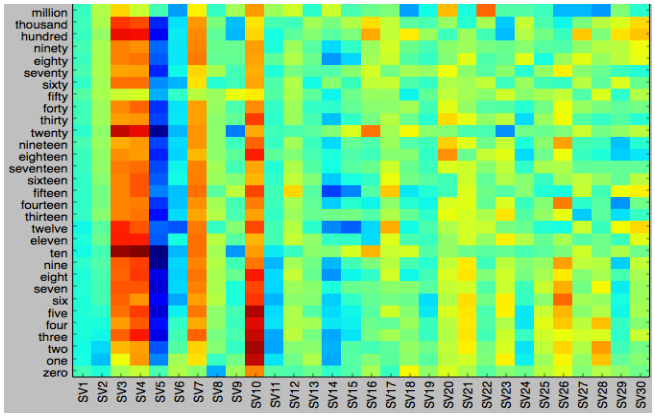


Figure 4. The singular value loadings for various number words across the first 30 singular value dimensions.

Figure 4 makes it clear just how important number words are to the structure of child-directed speech. Excepting principal components 1 and 2 (which tend not to carry structural information in these kinds of analyses), at least some number words load extremely highly on most of the primary principal components). Though beyond the

scope of this short paper, follow-up analyses of these principal components show that many of the components are interpretable as representing contextually different ways in which numbers are used, such as counting, money, time, dates, and measurement. Some of these relationships can be identified in Figure 4 by noting which numbers do, and do not, load on particular principle components.

Number Distributional Similarity

In our final analysis, we attempted to assess the extent to which distributional statistics can be used to learn that numbers belong to the same category, as well as to tell how number words are different from one another. To do this, we derived these similarity scores for each of the 8,000 most frequent words in the corpus, by computing the correlation of each word pair's 30-element principal component vectors (such as those shown in Figure 4). Thus, the extent to which each pair of words was similar depended on the extent to which the two words shared the same co-occurrence patterns in language (Riordan & Jones, 2011).

To assess how useful these similarity scores would be for guessing that number words belong to the same category, we performed a discrimination analysis of using each of the 31 number words from the above analyses, and a set of approximately 600 other words from a collection of 30 other semantic categories (such as mammals and foods). We took each pair of these words (both the number-number pairs as well as the number-"not-number" pairs), and used their similarity score to guess whether that pair should be judged as belonging to the same category or to different categories. We used a constant similarity threshold across all comparisons (r = 0.30), and used the accuracy of these judgments to compute balanced accuracy scores (i.e. we evenly averaged the score on all trials where the correct answer was "yes" with all trials where the correct answer was "no", to account for the fact that there were many more "no" trials). We performed this discrimination analysis on three different sub-corpora to get a sense of how the results changed across speech to children with of different ages. The results of this analysis are shown in Figure 5.

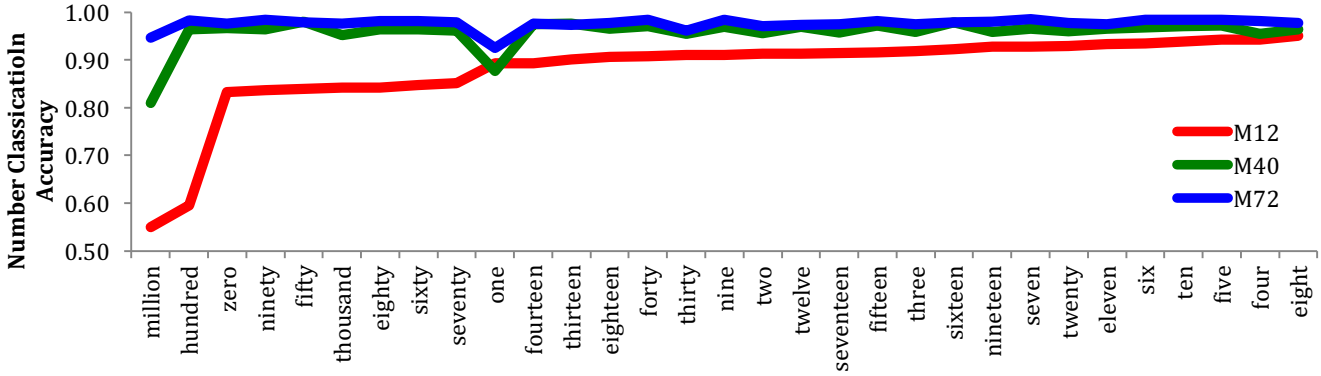


Figure 5. Classification balanced accuracy for various number words using distributional similarity to predict same vs. different category membership, for various number words at various stages of amount of input (up to 12 months, up to 40 months, up to 72 months).

This attempt to classify number words as same and different from other words based on their distributional statistics resulted in a number of interesting findings about the statistical structure of number words. The first is how remarkably successful children could be at classifying number words based on their distributional statistics. This was true even in the "youngest" corpus, with only the child-directed speech to children 12 months and below, with a mean accuracy across all number words of 88.0% (SE = 1.6%). Using all speech to children up to 40 months, accuracy rose to 95.8% (SE = 0.6%). Using all speech to children up to 72 months, accuracy peaked at 97.6% (SE = 0.2%). Number words are used remarkably consistently (in fact, number words can be classified better than words from any other category, Willits & Jones, in review), and learning they belong to the same category based on their usage statistics would be a very easy task. Also interesting, however, were the specific words that were more difficult, especially for the "youngest" corpus. As expected, *one* was difficult, but not the most difficult. Lower frequency (and less consistently used) words like the decade words and *million* were the words that were most difficult to classify.

To assess what children could learn about the internal structure of the number category based on language statistics, we performed a hierarchical cluster analysis of words based on their co-occurrence patterns. This cluster analysis is shown in Figure 6.

As can be seen in Figure 6, the sub-category clusters that emerge are incredibly sensible and informative about what children might be learning about number from number words. The most distinctively used words are *one*, *two*, and *three*, which form their own subcluster. Other small numbers group together, as do the teens, decades, and large numbers. One notable exception is 19, which groups with the decade numbers, driven by the fact that most of the CHILDES data was collected in the 1980s and 1990s.

Discussion

We find that the use of numbers in child-directed speech bears evidence of many of the same distinctions noted in childrens' number-related behavior. In the frequency analyses, the power-law nature of number words' frequency distribution and diversity of usage (with *one*, *two*, *three*, and *four* each being used successively and dramatically more than the next) could be important factor in children's difficulty learning to properly use words for larger numbers.

In the co-occurrence analyses, we discovered many ways in

which the co-occurrence patterns varied, that could lead to differences in learnability. Lower numbered words were relatively less likely to co-occur with one another, suggesting that these words are used in a number of diverse

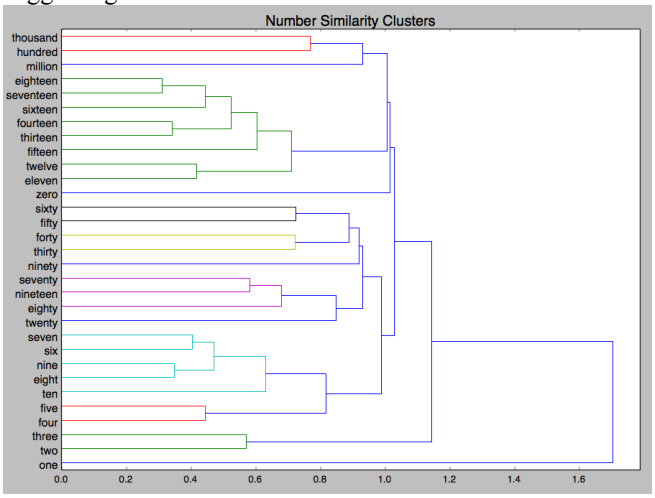


Figure 6. Hierarchical clustering of number words based on their co-occurrence patterns in CHILDES.

ways other than strictly counting and comparatively quantifying, which may lead to some of the difficulty children have at extending their ability to quantify small numbers to larger quantities.

Likewise, the similarity analyses demonstrate a number of facts about number word usage with implications for number concept acquisition. In particular, the two broadest divisions in the hierarchical cluster analysis were between one and other numbers, and between two and three and other numbers. One in particular seems to be used quite differently from other numbers, which is not terribly surprising given its less obviously cardinal roles in language, e.g., as a specification of an agent, or in idioms like "one time". The typical adult number categories (4-10, the teens, the decades, and the larger numbers) form distinct clusters as well, while zero forms its own subcategory.

Previous behavioral evidence has indicated special trouble understanding the computational role of the number words *million* (Landy et al., 2013), *zero* (Evans, 1983), and *hundred* and *thousand* (Rips, 2013; Siegler & Opfer, 2003). Statistical analyses confirm that these are treated differently than other number words: they are more difficult for a statistical algorithm to categorize as numbers based on usage (especially in early months), and form distinct sub-clusters in the hierarchical clustering analysis.

Just as striking is the nature of the clusters. The teen words tend to strongly co-occur, to form a tight cluster, and to contain strong mutual information. One plausible interpretation is that, compared to other numbers, these numbers appear primarily together in count lists, and appear relatively rarely in other contexts.

References

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*, 214–229.

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, *33*, 952-965.

Brysbaert, M., Fias, W., & Noel, M. P. (1998). The Whorfian hypothesis and numerical cognition: is twenty-four processed in the same way as four-and-twenty? *Cognition*, *66*, 51-77.

Carey, S. (2009). *The Origin of Concepts*. Oxford U. Press.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1-29.

Evans, D. W. (1983). *Understanding zero and infinity in the early school years.* Doctoral Dissertation.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307-314.

Fuson, K. C., & Kwon, Y. (1991). Chinese-based regular and European irregular systems of number words: The disadvantages for English-speaking children. Language in mathematical education: *Research and Practice*, 211-226.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, *1*, 3-55.

Gunderson, E. A., & Levine, S. C. (2011). Some types of parent number talk count more than others: Relations between parents' input and children's cardinal number knowledge. *Developmental Science*, *14*, 1021-1032.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language. *Journal of Memory and Language*, *63*, 259-273.

Landy, D., Silbert, N., & Goldin, A. (2013). Estimating large numbers. Cognitive science, 37(5), 775-799.

Landy, D., Charlesworth, A., & Ottmar, E. (2014). Cutting In Line: Discontinuities in the Use of Large Numbers by Adults. In Proceedings of the 36th Annual Conference of the Cognitive Science Society.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning infants' acquisition of lexical categories. *Psychological Science*, *21*, 284-291.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., and Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*, 607-614.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature- based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*, 303-345.

Rips, L. J. (2013). How many is a zillion? Sources of number distortion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1257.

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation evidence for multiple representations of numerical quantity. Psychological science, 14(3), 237-250.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*, 89-96.

Tomasello, M. (2003). Constructing a language: A Usage-based Theory of Language Acquisition. Cambridge, MA: Harvard University Press.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80-102.

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development*, *83*, 1382-1399.

Probabilistic Simulation Predicts Human Performance on Viscous Fluid-Pouring Problem

James Kubricht[⋆1] kubricht@ucla.edu
Chenfanfu Jiang[⋆2] cffjiang@cs.ucla.edu
Yixin Zhu[⋆3] yixin.zhu@ucla.edu
Song-Chun Zhu[2,3] sczhu@stat.ucla.edu
Demetri Terzopoulos[3] dt@cs.ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology [2] Department of Statistics [3] Department of Computer Science
[⋆] Equal Contributors University of California, Los Angeles

Abstract

The physical behavior of moving fluids is highly complex, yet people are able to interact with them in their everyday lives with relative ease. To investigate how humans achieve this remarkable ability, the present study extended the classical water-pouring problem (Schwartz & Black, 1999) to examine how humans take into consideration physical properties of fluids (e.g., viscosity) and perceptual variables (e.g., volume) in a reasoning task. We found that humans do not rely on simple qualitative heuristics to reason about fluid dynamics. Instead, they rely on the perceived viscosity and fluid volume to make quantitative judgments. Computational results from a probabilistic simulation model can account for human sensitivity to hidden attributes, such as viscosity, and their performance on the water-pouring task. In contrast, non-simulation models based on statistical learning fail to fit human performance. The results in the present paper provide converging evidence supporting mental simulation in physical reasoning, in addition to developing a set of experimental conditions that rectify the dissociation between explicit prediction and tacit judgment through the use of mental simulation strategies.

Keywords: Intuitive physics; mental simulation; animation; reasoning

Introduction

Imagine that you are preparing to pour pancake batter onto a griddle. To pour the correct amount, you must decide where to hold the container, at what angle, and for how long. We encounter similar situations frequently in our daily lives when interacting with viscous fluids ranging from water to honey, and with different volumes, contained in receptacles of various shapes and sizes.

In the majority of these situations, we are able to reason about fluid-related physical processes so as to implicitly predict how far a filled container can be tilted before the fluid inside begins to spill over its rim. However, people perform significantly worse when asked to make explicit reasoning judgments in similar situations (McAfee & Proffitt, 1991; Schwartz & Black, 1999). In the well-known Piagetian water level task (WLT; Rebelsky, 1964), participants receive instructions to draw the water level at indicated locations on the inside of tilted containers. Surprisingly, about 40% of adults predict water levels that deviate from the horizontal by 5 degrees or more (e.g., McAfee & Proffitt, 1991). Schwartz and Black (1999) modified the WLT to include two containers, one wider than the other. The investigators asked participants to judge which container would need to be tilted farther before the water inside begins to pour out. Only 34% of the participants correctly reported that the thinner container would need to be tilted farther than the wider one. However, when instructed to complete the task by closing their eyes and imagining the same situation, nearly all (95% of) the participants rotated a thinner, imaginary container (or a real, empty one) farther. These findings suggest that people are able to reason successfully about relative pour angles by mentally simulating the tilting event. An apparent contrast in human performance between an explicit reasoning task and a simulated-doing task has also been found in people's inferences about the trajectories of falling objects (Krist, Fieberg, & Wilkening, 1993; K. Smith, Battaglia, & Vul, 2013). Thus, empirical findings in the literature of physical reasoning suggest that people employ both explicit knowledge about physical rules *and* mental simulation when making inferences (Hegarty, 2004).

The *noisy Newton* framework for physical reasoning hypothesizes that inferences about dynamical systems can be generated by combining noisy perceptual inputs with the principles of classical (i.e., Newtonian) mechanics, given prior beliefs about represented variables (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Sanborn, 2014; Sanborn, Mansinghka, & Griffiths, 2013; K. A. Smith & Vul, 2013). In this framework, the locations, motions and physical attributes of objects are sampled from noisy distributions and propagated forward in time using an *intuitive physics engine*. The resulting predictions are queried and averaged across simulations to determine the probability of the associated human judgment. Bates et al. (2015) extended the framework from physical scene understanding (e.g., Battaglia et al., 2013) to fluid dynamics using an *intuitive fluid engine* (IFE), where future fluid states are approximated by probabilistic simulation via a Smoothed Particle Hydrodynamics (SPH) method (Monaghan, 1992). The particle-based IFE model matched human judgments about final fluid states and provided a better quantitative fit than alternative models that did not employ simulation or account for physical uncertainty.

The present study aims to determine whether a particle-based IFE model coupled with noisy input variables can account for human judgments about the relative pour angle of two containers filled with fluids differing in their volume and

viscosity. The experiment reported here was inspired by previous empirical findings in water-pouring tasks; e.g., participants tilt containers filled with imagined molasses farther than those with an equal volume of water, suggesting that people are able to take physical attributes such as viscosity into account when making fluid-related judgments (Schwartz & Black, 1999). Bates et al. (2015) also found that their participants' judgments were sensitive to latent attributes of the fluid (e.g., stickiness and viscosity).

To quantify the extent that humans employ their perceived viscosity of fluids in subsequent reasoning tasks, we utilized a recent development in graphical fluid simulation (Bridson, 2008; Jiang, Schroeder, Selle, Teran, & Stomakhin, 2015) to simulate the dynamic behavior of fluids in vivid animations. Previous work has shown that realistic animations can facilitate representation of *dynamic* physical situations (Tversky, Morrison, & Betrancourt, 2002). Furthermore, recent research on human visual recognition indicates that latent attributes of fluids (e.g., viscosity) are primarily perceived from visual motion cues (Kawabe, Maruya, Fleming, & Nishida, 2015), so displaying realistic fluid movement is needed to provide the input of key physical properties that enable mental fluid simulations. The present study, which uses a modification of Schwartz and Black's (1999) water-pouring problem coupled with advanced techniques in computer graphics, aims to test the hypothesis that animated demonstrations of flow behavior facilitate inference of latent fluid attributes, which inform mental simulations and enhance performance in subsequent reasoning tasks.

Experiment

Participants A total of 152 participants were recruited from the Department of Psychology subject pool at the University of California, Los Angeles, and were compensated with course credit.

Materials and Procedure Prior to the reasoning task, participants viewed animated demonstrations of the movement of a moderately viscous fluid in two situations. The fluid used in the demonstrations was colored orange and was not observed in the judgment task. In the first demonstration, the fluid pours over two torus-shaped obstructions in a video looped three times and lasting for 11.5 seconds. The flow demonstration videos were presented to provide visual motion cues to inform participants' perceived viscosity. Following the flow demonstration, participants viewed a video of a cylindrical container filled with the same orange fluid tilting at a constant angular rate ($\omega = 22$ deg$\cdot s^{-1}$; see Fig. 1) from the upright orientation of the container and moving towards the horizontal. The video was looped three times for a duration of 14.7 seconds.

Following the demonstration videos, two new fluids were introduced, one with low viscosity (*LV*; similar to water) and one with high viscosity (*HV*; similar to molasses). The *LV* and *HV* fluids were colored either red or green (counterbalanced across subjects). As shown in the top panel of Fig. 1,

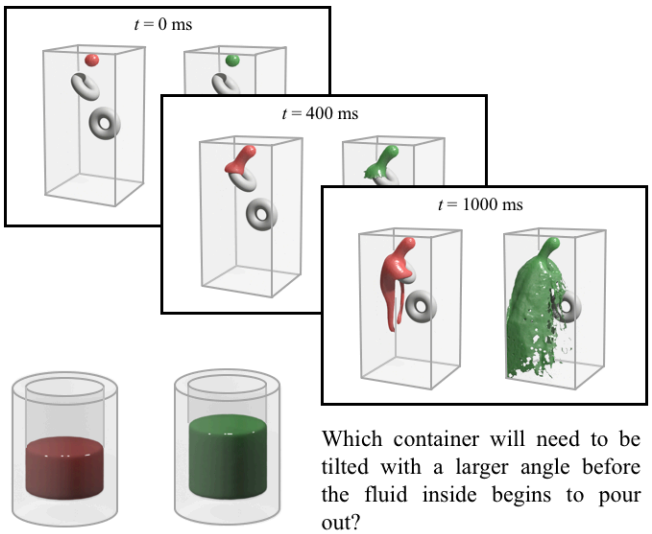


Figure 1: Illustration of flow demonstration video and judgment trial. (Top) Sample frames from the *HV* (red) and *LV* (green) flow video. (Bottom) Tilt judgment trial, where $V_{HV} = 40\%$ and $V_{LV} = 60\%$.

participants viewed a flow video of both the *HV* and *LV* fluids (looped three times) for a duration of 11.5 seconds before each judgment trial[1]. The two flow videos were presented side by side for comparison, and the relative position of each fluid was counterbalanced across subjects. The *LV* and *HV* fluids were selected to readily distinguish the two fluids based on their perceived viscosities, which were inferred from visual motion cues in the flow videos (Kawabe et al., 2015).

In the subsequent reasoning task, participants viewed a static image of two containers side by side filled with the *LV* and *HV* fluids (see bottom panel of Fig. 1). Participants were instructed to assume that each container was tilted simultaneously in the same way as observed earlier for the orange fluid in the tilting demonstration. They were informed that both containers were tilted at the same rate, and were provided with the quantity of fluid in each container. Participants were then asked to report "which container will need to be tilted with a larger angle before the fluid inside begins to pour out" and received no feedback following completion of each trial. The experiment manipulated the volume of the *LV* and *HV* fluids (V_{LV} and V_{HV}, respectively) in each container across the values 20%, 40%, 60%, and 80%, representing the proportion of the container filled. Hence, the experiment consisted of 16 trials presented in a randomized order, including all possible volume pairs between the *LV* and *HV* fluids. The experiment lasted approximately 10 minutes.

Human Results

Fig. 2 depicts human performance for all 16 fluid volume combinations. To assess the relationship between *HV* fluid

[1]Tilting and pouring videos can be viewed online at the following URLs: *http://hydra.math.ucla.edu/~cffjiang/cogsci/*[...]
Tilting videos: [...] = *tilt_movies/*
Flow videos: [...] = *pour_fluid_to_box_varying_viscosity/*.

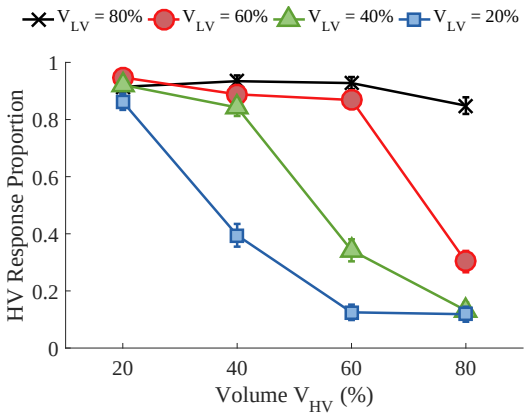


Figure 2: Human *HV* response proportions for all experimental conditions. The volume of the *HV* fluid (V_{HV}) is plotted on the horizontal axis, and separate lines indicate the four possible volumes for the *LV* fluid (V_{LV}).

volume and human judgments, we performed a logistic regression analysis and compared estimated coefficients for each participant across *LV* fluid volume conditions. We found that coefficients for each participant varied significantly between V_{LV} conditions ($F(3,149) = 113.89, p < .001$). In the highest *LV* fluid volume condition ($V_{LV} = 80\%$), V_{HV} had a minor impact on participants' responses ($\bar{\beta}_{80} = .04, \sigma_{\beta_{80}} = .25$) relative to the lowest *LV* fluid volume condition ($V_{HV} = 20\%$; $\bar{\beta}_{20} = .57, \sigma_{\beta_{20}} = .43$). These results demonstrate that participants' responses were increasingly sensitive to *HV* fluid volume for the greater V_{LV} conditions.

Next, we examined whether humans rely on heuristic-based reasoning to make their judgments. One candidate heuristic is that given two containers filled with different volumes of each fluid, the container with lesser fluid volume requires a greater rotation before beginning to pour. While participants adhered to this rule for trials where $V_{HV} < V_{LV}$, their judgments for each of the $V_{LV} < V_{HV}$ trials did not accord to the same heuristic. For example, in trials where V_{HV} = 40%, 60%, and 80% and V_{LV} = 20%, 40%, and 60%, respectively, the lesser-volume heuristic predicts *LV* fluid responses. However, *HV* response proportions for those trials were significantly greater thane zero ($t(151) = 9.92, 8.86, 8.10$, $p < .001$). A second heuristic is to always choose the *HV* fluid as requiring a greater rotation since it moves slower than the *LV* fluid. The above three cases also disagreed with this heuristic since the rule would predict unity *HV* response proportions. In summary, response proportions in the specified trials reveals that participants attended to latent fluid attributes (e.g., viscosity) *and* volume difference when making their tilt angle judgments (see Fig. 3).

Models

Fluid Simulation with Physical Dynamics

The simulation of incompressible flows through numerical evaluation of physical equations has become one of the most significant topics in computer graphics and mechanical engineering. The velocity field of simulated fluids is determined according to the constraints specified in the Navier-Stokes equations:

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v} \cdot \nabla \boldsymbol{v} + \frac{1}{\rho} \nabla p = \boldsymbol{g} + \mu \nabla \cdot \nabla \boldsymbol{v}, \quad (1)$$

$$\nabla \cdot \boldsymbol{v} = 0, \quad (2)$$

where $\boldsymbol{v}$ is the velocity, ρ is the density, p is the pressure, $\boldsymbol{g}$ is the gravitational acceleration (approximately 9.8 m s^{-2}), and μ is the viscosity. Equation (1) is called the momentum equation—it is simply Newton's second law (i.e., $F = ma$) applied to fluid dynamics. Equation (2) is the incompressibility constraint on fluid velocity, where the null divergence of the velocity field corresponds to constant density within volumetric regions. Most liquids need to satisfy this constraint in order to maintain constant volume while moving.

To numerically solve these partial differential equations, we adopt the Fluid Implicit Particle/Affine Particle in Cell (FLIP/APIC) method (Zhu & Bridson, 2005; Jiang et al., 2015; Jiang, 2015), which has become standard in physics-based simulation calculations due to its accuracy, stability and efficiency. Unlike Smoothed Particle Hydrodynamics (SPH), which purely relies on particles to discretize the computational domain, FLIP/APIC uses both particles and a background Eulerian grid. The Navier-Stokes equations are solved on the grid, allowing for accurate derivative calculations, well-defined free surface and solid boundary conditions, and accurate first-order approximation of physical reality. The FLIP/APIC method also circumvents common artifacts of SPH; e.g., underestimated density near free surfaces and weakly compressible artifacts. In fact, the requirement for incompressibility is crucial in the fluid-pouring problem studied in this paper. We choose not to use SPH because it does not guarantee a divergence-free velocity field unless additional computational components are included. FLIP/APIC, however, maintains the benefits of particle-based methods due to its hybrid particle/grid nature. The presence of particles in the current model serves to facilitate visualization and the tracking of material properties. Besides modeling fluid, the state-of-the-art physics-based simulation methods have provided realistic cues for modeling complex tool and tool-uses (Zhu, Zhao, & Zhu, 2015), generic containers (Liang, Zhao, Zhu, & Zhu, 2015) and soft human body (Zhu, Jiang, Zhao, Terzopoulos, & Zhu, 2016).

Realistic visualization of simulated fluids is particularly important for the flow demonstration animations displayed before each trial in our viscous fluid-pouring task. To provide vivid impressions of the motion of the *LV* and *HV* fluids, we adopt OpenVDB (Museth et al., 2013) and utilize the latest particle fluid surfacing methods developed in the field of computer graphics to reconstruct a smooth level set surface from the simulated fluid particles based on their locations.

The favorable efficiency and precision of the FLIP/APIC method allows for effortless generation of ground truth re-

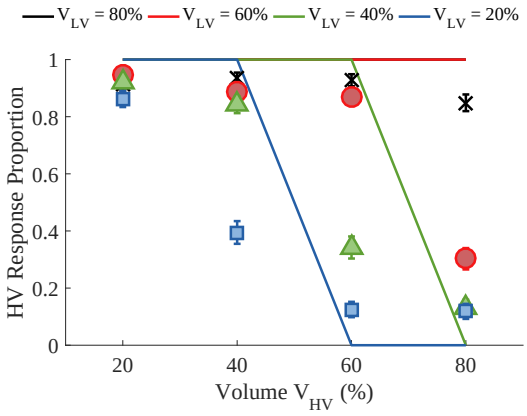


Figure 3: Simulation results from the FLIP/APIC model ($RMSD = 0.6747$). Separate lines indicate model predictions, given ground-truth volume/viscosity values for each fluid. Symbols indicate human response proportions. The deterministic simulation method returns binary predictions and does not provide probabilistic response proportion estimates.

sponses for our task, given that fluid viscosity and volume are known. The particle locations and vessel tilt angles are explicitly recorded at each time step as simulation outputs. Since the FLIP/APIC method does not involve any stochastic processes, the output of each simulation is deterministic. In each simulation, $40,000$ particles (with around 8 particles per grid cell) were used to ensure both stability and convergence.

Intuitive Fluid Engine

Fluid simulation with physical dynamics provides deterministic fluid movements if the ground-truth values of viscosity and volume are known. Hence, the decisions directly derived from the FLIP/APIC fluid simulator are binary judgments (Fig. 3), which implies that the physical simulation with high precision cannot explain humans' probabilistic judgment in the fluid-pouring task. Inspired by the approach of Bates et al. (2015) and the noisy Newton model (e.g., Sanborn et al., 2013), we combine the physical simulator of FLIP/APIC with noisy input variables (i.e., viscosity and volume) to form the Intuitive Fluid Engine (IFE) model, thereby accounting for physical uncertainty and the influence of viscosity and volume on our reasoning task.

The input variables for our IFE are the ground-truth values of volume and viscosity (V_T, μ_T) for the two fluids in each experimental trial. The sampling process of the IFE then samples $N = 10,000$ noisy viscosity and volume pair values $\{(V_i, \mu_i), i = 1, 2, \cdots, N\}$ for each fluid. Each sample (V_i, μ_i) of the 10000 generated noisy input variables is later passed to the FLIP/APIC simulator to produce a binary decision $B_i(V_i, \mu_i) \in \{L, R\}$. The decision is that either the left L or the right R container needs to be tilted with a larger angle before the fluid inside begins to pour out. By aggregating the prediction from all $10,000$ samples and dividing the sum by the number of samples, the IFE outputs the distribution $\boldsymbol{P}(V_T, \mu_T)$ for the given ground-truth values of viscosity and volume:

$$\boldsymbol{P}(V_T, \mu_T) = \begin{cases} P(V_T, \mu_T)_L = \dfrac{\sum_i H(B_i(V_i, \mu_i), L)}{N} \\ \\ P(V_T, \mu_T)_R = \dfrac{\sum_i H(B_i(V_i, \mu_i), R)}{N}, \end{cases} \tag{3}$$

where $H(\Psi, \Theta) = 1$ when $\Psi = \Theta$, and it is 0 otherwise.

To model physical uncertainty, the sampling process of the IFE model is implemented by adding perceptual noises to the ground-truth values of the physical input variables (i.e., viscosity and volume). Noisy volume is generated by adding an offset to its ground-truth value from a Gaussian distribution with mean 0 and variance σ_μ, whereas the noisy viscosity is generated by adding a fixed amount of Gaussian noise on a logarithmic scale (Sanborn et al., 2013): $V_i = f^{-1}(f(V_T) + \varepsilon)$, where V_T is the ground-truth value, $f(V_T) = \log(\omega \cdot V_T + k)$, f^{-1} is the inverse of f, and $\varepsilon \sim \text{Gaussian}(0, \sigma_V)$. The results reported herein chooses $\sigma_\mu = 0.1$, $\sigma_V = 0.1$, $k = 1.5$, $\omega = 1$.

Each simulation required approximately 10 minutes to run on a modern single-core CPU. In order to run a large number of simulations during the sampling process, we discretized the viscosity and volume spaces into finite sets. Specifically, the viscosities are discretized into 8 different cases $(0.1, 1, 10, 100, 200, 500, 1000, 2000)$ and the volumes into 21 different cases (0% to 100% with a step-size 5%). We precomputed the simulation results for each discretized case, and stored the results in a database. During the sampling process, the sampled inputs are numerically rounded to the precomputed discretized cases, where the results can be immediately retried from the database without re-computing.

Non-Simulation Models

To examine whether fluid simulation is necessary to account for how humans reason about fluid behavior, we compare the simulation model with two statistical learning methods—the generalized linear model (GLM) (McCullagh, 1984) and the support vector machine (SVM) (Cortes & Vapnik, 1995). These models are purely data-driven and do not involve any explicit knowledge of physical laws or physical simulation. The selected features for these models include (i) the volumes of fluids in both containers, and (ii) the viscosity ratio between the *LV* and *HV* fluids.

To predict human judgment for the $i^\star$th trial $J_{i^\star}$, both non-simulation models were tested with the $i^\star$th trial, and trained with the remaining 15 trials $\{J_i, i = 1, 2, \cdots, 16, i \neq i^\star\}$. The trained GLM model is directly applied to the test case to predict which container will need to be tilted to a larger angle before the fluid inside begins to pour out. Since the SVM is a discriminative classification method which can only predict discretized labels (i.e., $+1$ indicating selection of the left container and -1 indicating selection of the right container), we introduced perceptual noise (the same method for the IFE) to each test trial to model physical uncertainty. For each test trial, a set with $10,000$ samples was generated. The trained

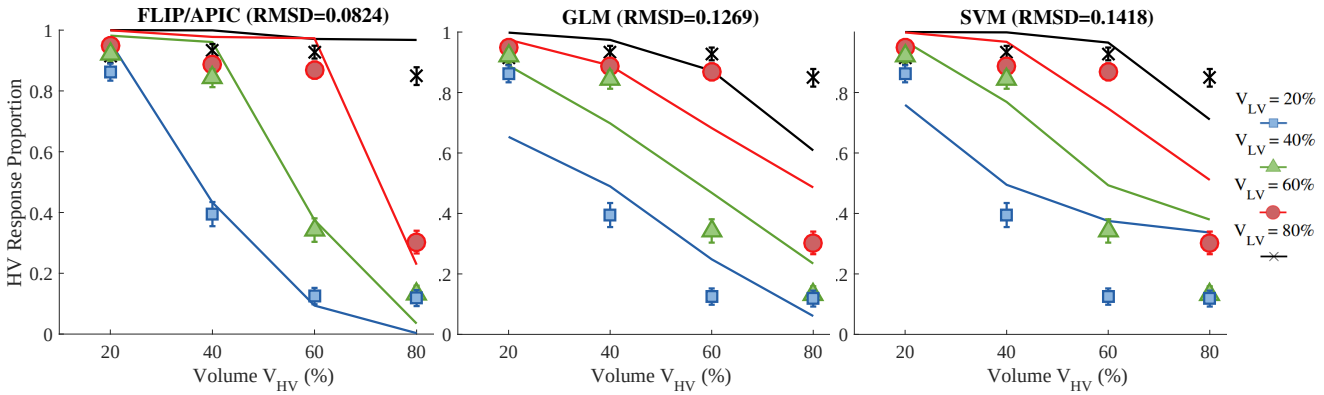


Figure 4: Comparison of results between our three prediction models: (Left) IFE, (Middle) Regression, and (Right) SVM with perceptual noise. Horizontal axes denote *HV* fluid volume; vertical axes denote the predicted proportion of *HV* fluid responses associated with a greater rotation angle. The IFE simulation model outperforms competing data-driven models.

SVM model is then applied to predict the labels ($+1$ or -1) in each sample, which are then aggregated to form the probability distribution for each test trial.

Model Results

We first compared how well different computational models account for human performance for the 16 trials. Fig. 4 depicts results from the IFE, GLM, and SVM models with perceptual noise. Human judgments and model predictions were highly correlated ($r(14) = 0.9954, 0.9488$, and 0.9251, respectively). RMSD (root-mean-squared deviation) between human judgments and the models' predictions are $0.0824, 0.1269$, and 0.1418, respectively. Compared to the purely data-driven models (i.e., the GLM and SVM models), the simulation-based IFE model encodes material properties (e.g., viscosity) and perceptual features (e.g., volume) and provides better approximations to human judgments in the viscous fluid-pouring task. These results again support the role of simulation as a potential mental model that supports human inference in physical reasoning tasks.

Next, we examined whether the IFE model captures human performance on the three trials where the heuristic rule outlined earlier provides incorrect predictions; i.e., those trials where V_{HV} = 40%, 60%, and 80% and V_{LV} = 20%, 40%, and 60%, respectively. Here, the ground-truth model predicts that the *HV* fluid will require a greater angle of rotation before beginning to pour, while the heuristic rule discussed earlier suggests the opposite. *HV* response proportions for these trials were .39, .34, and .30, respectively, and the IFE model returned consistent predictions of .43, .37, and .23. Alternatively, the GLM model predicted response proportions of .49, .47, and .49, and the SVM model predicted response proportions of .49, .49, and .51. Thus, our IFE model captured human performance on the specified trials while competing data-driven methods returned predictions biased toward the ground-truth model and away from the lesser-volume heuristic.

Discussion

Our results from the viscous fluid-pouring task agree with the findings of Bates et al. (2015) in that our probabilistic, simulation-based IFE model outperformed two non-simulation models (SVM and GLM). Our behavioral experiment also indicates that people naturally attend to latent attributes (e.g., viscosity) when reasoning about fluid states following observation of realistic flow demonstration animations. By extending our probabilistic IFE method to a reasoning task, we demonstrated that the noisy Newton framework can account for human performance in a fluid-related judgment task that traditionally precludes mental simulation strategies.

While simulation has been demonstrated as the default strategy in other mechanical reasoning tasks (Hegarty, 2004; Clement, 1994), the participants in Schwartz and Black's (1999) experiments failed to spontaneously represent and simulate physical properties relevant to the water-pouring problem when making their judgments. It is important to note, however, that the present task differs from the traditional water-pouring task in several ways: (i) fluid viscosity and volume (rather than container diameter) varied across trials, (ii) a cup-tilting demonstration was displayed to visualize the rate of simulated rotation, and (iii) motion cues from flow demonstrations informed the perception of latent fluid attributes (e.g., viscosity). Comparison of our study to previous water-pouring studies suggests that the dissociation between explicit physical prediction and implicit judgment reported in the intuitive physics literature could be resolved in some situations by modifying task characteristics and instructions in ways that motivate simulated representation. While our viscous water-pouring problem indicates a set of simulation-inducing task characteristics, further research should aim to determine specific experimental factors that trigger simulation strategies. Specifically, can the conditions employed in the present task extend to classical rigid-body and fluid mechanics problems to resolve the discrepancy between people's

explicit predictions and tacit judgments, and if so, what additional task characteristics serve to facilitate mental simulation?

Classical research in artificial intelligence has traditionally dismissed robust mental simulation as a strategy for physical reasoning due to its inherent complexity, often proposing simplified qualitative models instead (De Kleer & Brown, 1984). While the computational fluid simulations employed in the present study require extensive numerical evaluation to make predictions about future fluid states, humans appear to do so with precision and accuracy in comparatively small amounts of time. Furthermore, their performance in our reasoning task suggests representation of physical quantities that extends beyond qualitative process theory. While human results are generally consistent with physics-based simulation models coupled with noisy input variables, there remain discrepancies between model predictions and human judgments. Hence, future research should aim to address whether humans *simulate* fluid movements using mental models that accord to physical laws or *emulate* fluid dynamics by drawing on their everyday interactions with liquids across diverse physical situations (Grzeszczuk, Terzopoulos, & Hinton, 1998).

Acknowledgments

Support for the present study was provided by an NSF Graduate Research Fellowship, DoD CASIT grant W81XWH-15-1-0147, DARPA SIMPLEX grant N66001-15-C-4035 and ONR MURI grant N00014-16-1-2007.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Proceedings of the 37th annual conference of the cognitive science society.*

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bridson, R. (2008). *Fluid simulation for computer graphics*. CRC Press.

Clement, J. (1994). Use of physical intuition and imagistic simulation in expert problem solving.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

De Kleer, J., & Brown, J. S. (1984). A qualitative physics based on confluences. *Artificial Intelligence*, *24*(1), 7-83.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th annual conference of the cognitive science society.*

Grzeszczuk, R., Terzopoulos, D., & Hinton, G. (1998). Neuroanimator: Fast neural network emulation and control of physics-based models. In *Proceedings of the 25th annual conference on computer graphics and interactive techniques.*

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, *8*(6), 280–285.

Jiang, C. (2015). The material point method for the physics-based simulation of solids and fluids. *PhD dissertation, University of California, Los Angeles*.

Jiang, C., Schroeder, C., Selle, A., Teran, J., & Stomakhin, A. (2015). The affine particle-in-cell method. *ACM Transactions on Graphics (TOG)*, *34*(4), 51.

Kawabe, T., Maruya, K., Fleming, R. W., & Nishida, S. (2015). Seeing liquids from visual motion. *Vision research*, *109*, 125–138.

Krist, H., Fieberg, E. L., & Wilkening, F. (1993). Intuitive physics in action and judgment: The development of knowledge about projectile motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 952.

Liang, W., Zhao, Y., Zhu, Y., & Zhu, S.-C. (2015). Evaluating human cognition of containing relations with physical simulation. In *Proceedings of the 37th annual conference of the cognitive science society.*

McAfee, E. A., & Proffitt, D. R. (1991). Understanding the surface orientation of liquids. *Cognitive Psychology*, *23*(3), 483–514.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, *16*(3), 285–292.

Monaghan, J. J. (1992). Smoothed particle hydrodynamics. *Annual review of astronomy and astrophysics*, *30*, 543–574.

Museth, K., Lait, J., Johanson, J., Budsberg, J., Henderson, R., Alden, M., ... Pearce, A. (2013). Openvdb: an open-source data structure and toolkit for high-resolution volumes. In *Acm siggraph 2013 courses* (p. 19).

Rebelsky, F. (1964). Adult perception of the horizontal. *Perceptual and Motor Skills*, *19*, 371-374.

Sanborn, A. N. (2014). Testing bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in psychology*, *5*.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116.

Smith, K., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the 35th conference of the cognitive science society* (pp. 3426–3431).

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*(1), 185–199.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: can it facilitate? *International journal of human-computer studies*, *57*(4), 247–262.

Zhu, Y., & Bridson, R. (2005). Animating sand as a fluid. *ACM Transactions on Graphics (TOG)*, *24*(3), 965–972.

Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., & Zhu, S.-C. (2016). Inferring forces and learning human utilities from videos. In *Ieee conference on computer vision and pattern recognition (cvpr).*

Zhu, Y., Zhao, Y., & Zhu, S.-C. (2015). Understanding tools: Task-oriented object modeling, learning and recognition. In *Ieee conference on computer vision and pattern recognition (cvpr).*

College Students' Understanding of Linear Functions: Slope is Slippery

Marta K. Mielicki (mmieli2@uic.edu)
Department of Psychology, University of Illinois at Chicago

Jennifer Wiley (jwiley@uic.edu)
Department of Psychology, University of Illinois at Chicago

Abstract

A common obstacle for students in the transition from arithmetic to algebra is developing a conceptual understanding of equations representing functions. Two experiments manipulated isomorphic problems in terms of their solution requirements (computation vs. interpretation) and format to test for understanding of linear functions. Experiment 1 provided problems in a story context, and found that performance on slope comparison problems was low, especially when problems were presented with equations. Experiment 2 tested whether performance on slope comparison problems improves when problem prompts include explicit mathematical terminology rather than just natural language consistent with the problem story. Results suggest that many undergraduate students fail to access the mathematical concept of slope when problem prompts are presented with natural language. Overall, the results suggest that even undergraduate students lack understanding of the slope concept and equations of linear functions, both which are foundational for advanced algebraic thinking.

Keywords: algebraic problem solving, slope, linear functions

Introduction

A common obstacle for students in the transition from arithmetic to algebra is developing a conceptual understanding of equations representing functions. Prior work has demonstrated that students can persist in procedural approaches rather than conceptual approaches (Rittle-Johnson & Alibali, 1999; Rittle-Johnson, Siegler, & Alibali, 2001); process-based approaches rather than object-based approaches (Kieran, 1992); operational or computational approaches rather than relational approaches (Chesney & McNeil, 2014; Kaput, 2000; McNeil & Alibali, 2005) even after algebra instruction. Although much of the research on the transition from arithmetic to algebraic thinking has focused on younger students (middle school and high school), even college students may continue to experience difficulties and lack a mature understanding of linear functions (Hall et al., 1989; Wollman, 1983).

Prior work has suggested that problem presentation plays an important role in solution success. Mevarech and Stern (1997) found that context can affect performance on linear equation problems that are presented graphically. Importantly, the problems that Mevarech and Stern (1997) administered all pertained to the mathematical concept of slope. They found that both younger students (around 12 year olds) and undergraduates performed better when problems were embedded in sparse contexts than when problems were presented in realistic contexts (e.g. graphs depicting how two hoses fill a pool at different rates). Mevarech and Stern (1997) suggested that presenting problems in realistic contexts may overload the problem solver with extraneous information which obscures the underlying mathematical concept and leads to poor performance.

Other work has also found that presentation format can affect undergraduate performance on solving linear equation problems. Mielicki and Wiley (2016) presented undergraduate students with isomorphic problems pertaining to linear functions either in graphical format or with a set of equations. Problems either required computation of a point on a single line (solve for x, solve for y) or required relational reasoning (comparing slopes or comparing points across several linear functions). A main finding was that slope comparison problems were most difficult for students overall, and that these problems were especially difficult when presented in equation format. This finding conforms to the interpretation that Mevarch and Stern (1997) proposed for their findings, and suggests that students may have difficulty accessing relevant mathematical knowledge when slope problems are presented in equation format. Taken together, these findings of significant difference in performance due to problem presentation highlight specific deficits that many undergraduates possess in their understanding of equations representing linear functions.

Experiment 1

The goal of Experiment 1 was to further test the hypothesis that undergraduate students may lack a mature understanding of linear functions expressed as equations, and particularly the mathematical concept of slope. As in Mielicki and Wiley (2016), students were asked to solve isomorphic problems presented either with graphs or with equations. The problems entailed either computation or interpretation across a set of linear functions. Both problem types featured distinct subtypes: computation problems either entailed solving for x or solving y and interpretation problems either entailed comparing the slopes of three linear functions or comparing y values of three linear functions along some range of x values. Examples of each problem type are shown in Figure 1.

Based on a cognitive task analysis, it was expected that the four subtypes of problems would have different cognitive demands in equation and graph format. Overall, computation problems were expected to have higher solution rates than interpretation problems because the former only require consideration of a single linear function whereas the latter require comparing three functions. Between the two subtypes of computation problems, solving for x was expected to be

more demanding than solving for y in equation format, since solving for x requires additional computational steps to isolate the variable (all problems were presented in y = mx + b format). For interpretation problems, the cognitive demands for problems requiring the comparison of slopes should not vary by presentation format if students understand the conceptual meaning of what the quantities in the $y = mx + b$ equation represent. In either format, little calculation is necessary – the solver could easily compare the visual trends of the three lines with a graph, or compare the three values of "m" across equations. In contrast, comparing points might be easier to do in graphical format because in equation format multiple calculations are required to compare y values for three equations. These calculations may introduce additional opportunities for error which are not present when point comparison problems are presented with graphs. It was expected that these differences in computational demand would be reflected in solution accuracy.

Method

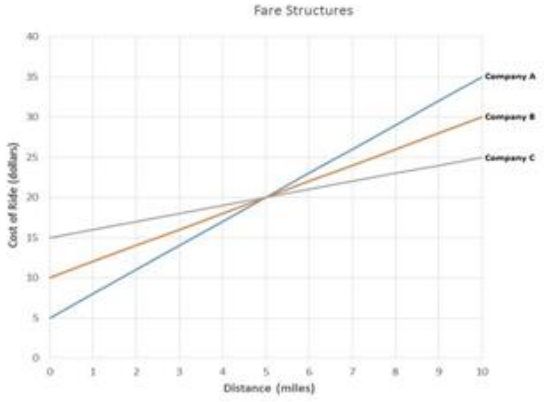
Participants

A sample of 32 (21 female) undergraduate students from the University of Illinois at Chicago participated in exchange for course credit. Participants were mostly first year undergraduate students (ages ranging from 17 to 23), and had taken 1 math course on average since starting college. Many participants (91%) reported a science (Biology, Chemistry, Psychology, Engineering) or health-related (Pre-med, Pre-nursing, Kinesiology) major. No participants reported pursuing a mathematics major.

Materials

Each participant completed 12 interpretation and 12 computation problems. Computation problems either entailed solving for x or solving for y. Interpretation problems either entailed comparing the slopes or comparing the y values of the three linear functions over a range of x values. Of the 24 problems, half were presented with graphs and half were presented with equations (always in $y = mx + b$ format).

In addition to these manipulations, several other presentation features were either varied in the same way for each participant to reduce monotony, or counterbalanced in order to control for order effects. Each problem was presented individually, but pairs of problems were presented with one format (graphs or equations) and the format for each pair was switched for the second half of the problems so that each participant saw three instances of each problem subtype (solve for x, solve for y, slope comparison, point comparison) in each format (graph, equations). Six configurations of linear functions (one for each block of 4 problems) were used in the same order for each participant. Computation problems requiring solving for x were always paired with slope comparison interpretation problems, and computation problems requiring solving for y were always paired with point comparison interpretation problems. Pairs of problems were alternated, and graph versus equation format was alternated between each pair of problems. In addition, each pair of problems was presented with one of two problem scenarios (i.e., real-world contexts such as comparing cab companies).

Solve for x Problem in Graph Format	Solve for y Problem in Symbolic Format
Malik is comparing three cab companies. Each company has a different fare structure for charging customers. Use the graph below to answer the following questions.  If company C charges Malik $25, how many miles did he travel?	Bob is participating in a walkathon, and he has gotten three sponsors to donate money to charity for every kilometer he walks. Each sponsor has a different pledge plan for how much money they will donate. Here are the equations for each sponsor's pledge where y is the amount of money donated in dollars and x is the distance walked in kilometers. Sponsor A: y = 3x + 5 Sponsor B: y = 2x + 10 Sponsor C: y = x + 15 How much will Sponsor C donate if Bob walks 10 kilometers?

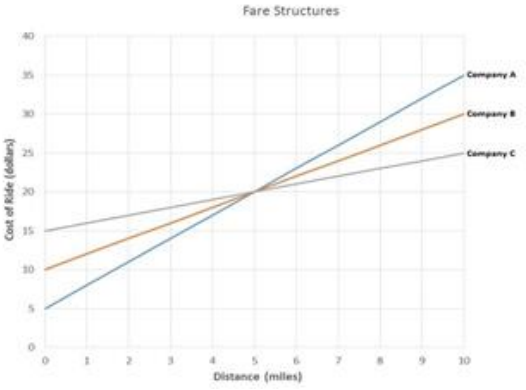
Slope Comparison Problem in Graph Format	Point Comparison Problem in Symbolic Format
Malik is comparing three cab companies. Each company has a different fare structure for charging customers. Use the graph below to answer the following questions.  Which company has the lowest rate per mile?	Bob is participating in a walkathon, and he has gotten three sponsors to donate money to charity for every kilometer he walks. Each sponsor has a different pledge plan for how much money they will donate. Here are the equations for each sponsor's pledge where y is the amount of money donated in dollars and x is the distance walked in kilometers. Sponsor A: y = 3x + 5 Sponsor B: y = 2x + 10 Sponsor C: y = x + 15 Which sponsor will donate the least if Bob walks over 5 kilometers?

Figure 1: Examples of interpretation problems (slope comparison, point comparison) and computation problems (solve for x, solve for y) in either equation or graphical format.

The following were counterbalanced across participants: order of format presentation (graph first versus equation first), order of scenario (scenario A with graph and B with equations versus scenario A with equations and B with graph), order of pair presentation (solve for x/slope comparison pair first versus solve for y/point comparison pair first), and order of problem subtype presentation within pairs of problems (computation first versus interpretation first). This 2x2x2x2 counterbalancing design led to 16 versions, and 2 participants completed each version.

Procedure

All problems were presented one at a time on a computer screen, but participants wrote their responses in an answer booklet that was provided. The answer booklets also included the graphs or equations for each problem so that participants could annotate and interact with the representations as needed. After completing the problems, participants were asked to rephrase a subset of the problems with the prompt, "In your own words, please tell me what you think the problem is asking you to do."

Results

A 2x2 within-subjects ANOVA was conducted with format (graph, equation) and problem type (interpretation, computation) as independent variables and proportion correct as the dependent variable. As shown in Figure 2, the analyses revealed a main effect of representation, $F(1,31) = 27.56.$, $p < .001$, $\eta_p^2 = .47$. Participants solved both types of problems more successfully when problems were presented with graphs than when they were presented with equations. The analysis also revealed a main effect of problem type, $F(1,31) = 68.46$, $p < .001$, $\eta_p^2 = .69$. Participants solved computation problems more successfully than interpretation problems, regardless of presentation format. There was a significant interaction, $F(1,31) = 4.93$, $p < .05$, $\eta_p^2 = .14$. Follow up analyses revealed no differences in accuracy for computation problems in different formats, $t(31) = 1.83$, $p = .08$, and higher accuracy for interpretation problems presented in graphical format relative to equation format, $t(31) = 4.51$, $p < .001$.

Additional 2x2 within-subjects ANOVAs were conducted separately for each problem type because the nested subtypes were not orthogonal. For computation problem subtypes, there was no effect of format, $F(1,31) = 3.36$, $p = .08$, or problem subtype, $F(1,31) = 3.14$, $p = .09$, and no interaction, $F < 1$. For interpretation problem subtypes, there was a main effect of format in favor of graphs, $F(1,31) = 20.31$, $p < .001$, $\eta_p^2 = .40$, and a main effect of problem subtype with higher accuracy on point comparison problems than slope comparison problems, $F(1,31) = 43.36$, $p < .001$, $\eta_p^2 = .58$. The interaction was not significant, $F < 1$.

Participants' rephrasing responses for slope comparison problems were coded based on reported solution methods. Most participants (31% for both formats) rephrased the question without providing any indication of a solution method. Some participants referenced the representation, but did not provide a solution method (6% for graphs, 9% for equations), and others mentioned a method but did not provide enough information to determine whether the method was correct or not (6% for graphs, 3% for equations). The remaining responses were coded as either incorrect solution method, or correct solution method/reference to the underlying concept (slope). For slope comparison problems presented with equations, 38% of participants reported an incorrect solution method relative to 25% of participants for problems presented with graphs. When slope comparison

problems were presented with equations, common incorrect strategies were "plugging in a number" for the x value of all three functions and comparing the results. For slope comparison problems presented with graphs, incorrect strategies were often some variation on "I just looked at the graph to see which line was highest." This approach indicates that students often experienced slope/height confusion (Leinhardt, Zaslavsky, & Stein, 1990). Conversely, 32% of participants reported a correct solution method or actually referenced the concept of slope when problems were presented with graphs relative to 18% when problems were presented with equations.

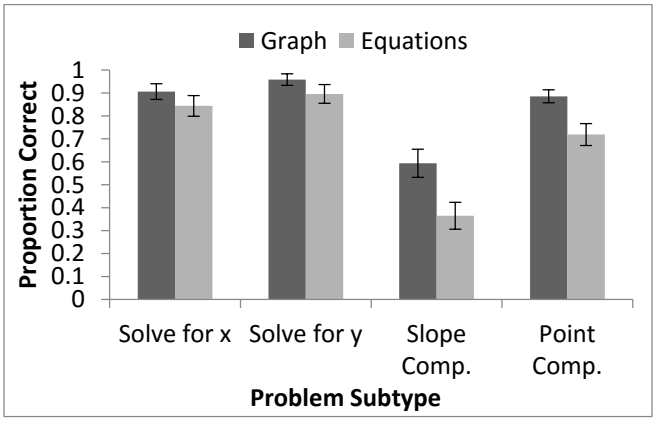


Figure 2: Mean proportion correct for the four problem subtypes presented in graphical and equation format. Error bars represent standard error.

Discussion

One striking result from Experiment 1 was the low overall accuracy on slope comparison problems in both formats, but particularly for slope comparison problems presented with equations. This pattern of results replicates the findings of Mielicki and Wiley (2016).

Participants' rephrasing responses suggest that graphs and equations may make different solution methods more accessible. Importantly, although the solution methods that students used when solving graphical slope problems sometimes led to a correct solution, these solution methods were not always indicative of conceptual understanding. For instance, it was possible for students to confuse slope with height when problems were presented with graphs (Leinhardt, Zaslavsky, & Stein, 1990) and still answer the problem correctly because the line with the highest height (*y* value) also had the highest slope.

However, even though performance was better on the slope comparison problems presented graphically, performance was still lowest overall on this problem type. It is possible that the overall low accuracy on slope comparison problems can be traced to difficulty during the problem comprehension phase. In previous work, it has been demonstrated that students often struggle with translating between stories, equations, and mathematical concepts (Cummins et al, 1988; Mayer, 1982; Nathan, Kintsch, & Young, 1992). In many contexts, students seem to fail to identify relevant mathematical principles. The results of Experiment 1 could be attributed to students not making the connection between the mathematical concept of slope and the demands of the task. If this is the case, then altering the language of the problem prompt to make the mathematical principles more clear should improve performance.

Experiment 2

The main goal of Experiment 2 was to test whether students would be able to solve equations requiring comparisons of slopes more successfully when the comprehension phase of problem solving is supported by presenting problem prompts with explicit mathematical terminology as opposed to natural language (as in Experiment 1).

This study focused specifically on only the two interpretation problem subtypes (point comparison, slope comparison) from Experiment 1. Manipulating the linguistic form of the problem prompt was not expected to have an effect on performance for point comparison problems. In Experiment 1, performance on point comparison problems benefitted from graphical format. This graphical advantage was consistent with the predictions from the cognitive task analysis that point comparison problems presented with equations were more computationally demanding than those presented with graphs. Thus the same pattern of results was expected for point comparison problems presented with mathematical terminology in Experiment 2.

If the graphical advantage for slope comparison problems found in Experiment 1 was due to graphs making different (though not necessarily correct) strategies accessible to participants, then supporting the comprehension phase should eliminate this advantage because participants should rely less on incorrect strategies if they can better access the underlying mathematical concept of slope. Thus, it was predicted that the graphical advantage found in Experiment 1 would be replicated when problems were presented with natural language, but that this advantage would be eliminated when problems were presented with explicit mathematical terminology. In addition, an overall main effect of linguistic form was expected because problem performance should improve overall when the comprehension phase is supported.

Method

Participants

A sample of 32 (20 female) undergraduate students from the University of Illinois at Chicago participated in exchange for course credit. Participants were mostly first year undergraduate students (ages ranging from 17 to 29), and had taken 0 to 3 math courses since starting college. Many participants (84%) reported a science (Biology, Chemistry, Psychology, Neuroscience) or health-related (Pre-dental, Pre-nursing, Kinesiology, Occupational Therapy) major. No participants reported pursuing a mathematics major.

Materials

A subset of slope comparison and point comparison problems from Experiment 1 was used for Experiment 2, and a second version of each problem was created in which the problem prompts were rephrased using explicit mathematical terminology instead of natural language. For example, prompts like "which cab company charges the most per mile?" were changed to "which line has the highest slope?" Each participant completed 8 problems, which were divided evenly between format (graphs, equations), type (point comparison, slope comparison), and linguistic form (natural language, mathematical terminology). These were the main manipulations of interest.

In addition, several presentation features were counterbalanced across participants. Items were blocked by linguistic form, and the order of the blocks was counterbalanced across participants with half completing natural language problems first and half completing mathematical terminology problems first. Within each block, the first and fourth problems were presented with one representation, and the second and third problems were presented with the other representation. Slope comparison and point comparison problems were alternated within each block. The order of problem type presentation and the order of format presentation were counterbalanced across participants. This 2x2x2 design resulted in 8 versions of the task, and 4 participants completed each version.

Procedure

The procedure was the same as the procedure in Experiment 1 except that rephrasing responses were not collected.

Results

A 2x2x2 within-subjects ANOVA was conducted with format (graph, equation); problem type (slope comparison, point comparison); and linguistic form (natural language, mathematical terminology) as independent variables and proportion correct as the dependent variable. As shown in Figure 4, the analysis revealed a main effect of linguistic form with problem prompts presented with mathematical terminology being solved more accurately than problem prompts presented with natural language, $F(1,31) = 8.03$, $p < .01$, $\eta_p^2 = .21$. There was also a main effect of format, with higher accuracy on problems presented with graphs than equations, $F(1,31) = 6.55$, $p < .05$, $\eta_p^2 = .17$. There was no main effect of problem type, $F < 1$. The linguistic form by representation and representation by problem type interactions were not significant, $F < 1$, and the three way interaction also did not reach significance, $F(1,31) = 2.99$, $p = .09$.

However, there was a linguistic form by problem type interaction, $F(1,31) = 7.79$, $p < .01$, $\eta_p^2 = .20$. Follow up paired-sample *t*-tests collapsing across representation revealed no difference in performance on point comparison problems based on linguistic form, $t < 1$; however, presenting problems in mathematical language significantly improved performance on slope comparison problems, $t(31) = 4.45$, $p < .001$.

Because it was predicted *a priori* that the results from Experiment 1 would be replicated for slope comparison problems presented in natural language form, and that mathematical language should eliminate the graphical advantage, paired-samples *t*-tests were conducted to compare performance on slope comparison problems presented with graphs with performance on problems presented with equations for each linguistic form. As predicted, performance was better on slope comparison problems presented with graphs than presented with equations when problems were in natural language form, $t(31) = 2.29$, $p < .05$. There was no difference in performance on slope comparison problems when presented in mathematical terminology, $t < 1$.

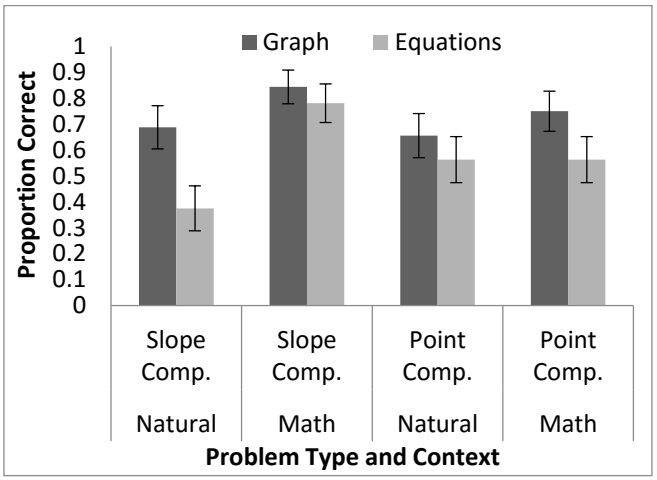


Figure 4: Mean proportion correct for slope comparison and point comparison problems presented in graph and equation format and in natural language or mathematical terminology form. Error bars represent standard error.

Discussion

The main result from Experiment 2 was that accuracy on slope comparison problems, particularly when presented with equations, improved when problems were presented in mathematical language. In addition, the graphical advantage found for slope comparison problems presented with natural language was eliminated when problems were presented with mathematical terminology.

General Discussion

The findings from Experiment 2 demonstrate that students are capable of solving slope comparison problems under certain conditions. Students do possess some understanding of the slope concept, but the limitations of this understanding are brought to light when problems are presented with natural language and particularly in equation format.

The graphical advantage for slope comparison problems found in Experiment 1 could be attributed to graphical format facilitating solution methods that lead to a correct answer without engaging the appropriate mathematical concept.

Student responses to the rephrasing prompts provide some support for this interpretation, but future research will need to address this possibility further by collecting trace data using think-aloud or eye-tracking methodology in order to better understand the ways that students are solving slope comparison problems in both formats.

Taken together, these results suggest that many undergraduates have not made the shift from procedural to conceptual understanding of slope. Students' difficulty recognizing or accessing the slope concept during problem solving highlights specific weaknesses in undergraduates' algebraic understanding. Slope has been widely acknowledged as a fundamental mathematical concept, with important implications for achievement in mathematics. Understanding of slope has been identified as central to success in precalculus and calculus (Carlson, Oehrtman, & Engelke, 2010), which in turn are required for many STEM career paths. Although slope is an important mathematical concept, it is also a notoriously difficult one for students to understand (Stump, 2001). Because linear equations are foundational, understanding and addressing weaknesses in students' conceptions of functions and slope represents an important step towards mending the leaky STEM pipeline.

Acknowledgments

Many thanks to James W. Pellegrino and Mara V. Martinez for their valuable feedback on this project.

References

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*, 113–145.

Chesney, D. L., & McNeil, N. M. (2014). Activation of operational thinking during arithmetic practice hinders learning and transfer. *The Journal of Problem Solving*, *7*, 24-35.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving algebra word problems. *Cognitive Psychology*, *20*, 405–438.

Hall, R., Kibler, D., Wenger, E., & Truxaw, C. (1989). Exploring the Episodic Structure of Algebra Story Problem Solving. *Cognition and Instruction*, *6*, 223–283.

Kaput, J. J. (2000). *Teaching and learning a new algebra with understanding*. Dartmouth, MA: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Kieran, C. (1992). The learning and teaching of school algebra. In D. A. E. Grouws (Ed.) *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics (*pp 390–419). New York: Macmillan Publishing Co., Inc.

Leinhardt, G., Zaslavsky, O., & Stein, M.K. (1990). Functions, graphs, and graphing: Tasks, learning and teaching. *Review of Educational Research*, *60*, 1-64.

Mayer, R. E. (1982). Different problem-solving strategies for algebra word and equation problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 448–462.

Mevarech, Z. R., & Stern, E. (1997). Interaction between knowledge and contexts on understanding abstract mathematical concepts. *Journal of Experimental Child Psychology*, *65*, 68–95.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, *76*, 883-899.

Mielicki, M. K. and Wiley, J. (2016) Alternative representations in algebraic problem solving: When are graphs better than equations? *The Journal of Problem Solving*, *9*, 3-12.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra word problem comprehension and its implications for the design of computer learning environments. *Cognition and Instruction*, *9*,329–389.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other?. *Journal of Educational Psychology*, *91*, 175-189.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*, 346-362.

Stump, S. L. (2001). High school precalculus students' understanding of slope as measure. *School Science and Mathematics*, *101*, 81–89.

Wollman, W. (1983). Determining the sources of error in a translation from sentence to equation. *Journal for Research in Mathematics Education*, *14*, 169-181.

L2 Idiom Processing: Figurative Attunement in Highly Idiomatic Contexts

Sara D. Beck (sara.beck@uni-tuebingen.de)
Collaborative Research Center 833, Wilhelmstrasse 50
72074 Tübingen, GERMANY

Andrea Weber (andrea.weber@uni-tuebingen.de)
English Department, Wilhelmstrasse 50
72074 Tübingen, GERMANY

Abstract

Using cross-modal priming, we investigated the processing of idioms in non-native listeners in varying experimental contexts. As idiomatic processing models have presented evidence for an idiomatic mode of processing that can be activated for non-native speakers in highly figurative contexts (Bobrow & Bell, 1973), this experiment revisits those claims while also examining access to figurative meaning in addition to the literal meaning of individual words within an idiom. This experiment showed increased priming for visual targets related to the figurative meaning of an idiom when the experimental list contained a large proportion of idiomatic sentences compared to when the list contained only a small proportion of idiomatic sentences. Non-native speakers not only showed online access to figurative meaning but were also sensitive to highly idiomatic contexts; though, responses to the targets related to literal meaning of the final word of the idiom were faster in all instances than figuratively-related targets.

Keywords: cross-modal priming; L2 listening; figurative language; idioms; context; attunement

Introduction

While understanding idioms is *a piece of cake* for native speakers of English, non-native (L2) speakers often struggle to recognize and understand them. Not only is figurative language extremely prevalent in everyday English use, but idioms are among the most frequent figurative expressions used by native speakers and an integral part of non-native language competence (see e.g., Cieślicka, 2006; 2013). In defining idioms, researchers generally agree that 1) the meaning of an idiom often differs from the literal meaning of the words comprising the phrase, 2) idioms have fixed structures or structures with limited variation, and 3) idioms are multi-word expressions (Liu, 2008). Although the challenges that these expressions pose for L2 learners are well-documented (see e.g., Cooper, 1999), the underlying processes are still in need of research. While we know that highly proficient L2 listeners can have access to figurative meaning in some instances, it may not be the case for all idioms. The current experiment tested online processing of figurative and literal meaning for L2 listeners in two different contexts: a highly figurative and a less figurative context.

Native Idiom Processing

Idiom processing has been the subject of many L1 studies for quite some time. Idioms are a particularly interesting linguistic phenomenon since many allow for both figurative and literal interpretations. For instance, *a piece of cake* can refer literally to a slice of cake or it can be figurative and mean "easy." This aspect of idioms has been the basis of one of the most studied questions of idiomatic processing: How does the processing of figurative meaning compare to that of literal meaning? A number of models have been proposed to address this issue.

One of the first models of processing developed specifically to address idiomatic processing was proposed by Bobrow and Bell (1973). Like standard pragmatic models, the *Idiom List Hypothesis* assumes that figurative and literal meaning undergo separate processes and, in normal contexts, literal meaning has processing priority over figurative meaning. In the idiomatic mode of processing, idiomatic meanings are retrieved from a list and do not undergo the same composition that literal language does. Following contradictory psycholinguistic evidence, Swinney and Cutler's (1979) *Lexical Representation Hypothesis* proposed that simultaneous processing occurs; however, figurative meaning is accessed first due to lower processing costs. Both theories assume that idioms are stored as one unit and processed as long words; thus, figurative meaning need only be retrieved, while a literal phrase must be retrieved and composed. An alternative proposed by Cacciari and Tabossi (1988), the *Configuration Hypothesis*, assumes that literal word meaning is processed until an idiomatic key in the idiom is reached. At this point, the configuration of words is recognized as an idiom, and the figurative meaning is accessed, giving priority at that point to figurative processing.

Literal-first models have been widely discounted based on the mounting psycholinguistic evidence that figurative meaning is often faster than not only literal meaning in a non-biased sentential context (e.g., Cacciari & Tabossi, 1988; Swinney & Culter, 1979) but also comparative novel phrases (e.g., Tabossi, Fanari & Wolf, 2009). However, it is not clear whether processing occurs simultaneously as two separate processes, one process that differentiates at a recognition point, or even one process influenced by other individual idiomatic properties such as decomposability (see e.g., Gibbs, Nayak & Cutting, 1989), literal saliency (e.g., Cieślicka 2006), or frequency (e.g., Tabossi et al. 2009).

Non-Native Idiom Processing

Much research on non-native processing of idioms focuses on both the comparison of access to figurative and literal meaning in addition to the comparison of idiomatic

processing to novel language. Unlike L1 research, there is more variation in the access to figurative meaning and most L2 results focus on supporting or refuting the existing L1 models of processing. Researchers such as Conklin and Schmitt (2008) and Underwood, Schmitt and Galpin (2004) and Siyanova-Chanturia, Conklin and Schmitt. (2011) investigated access to figurative meaning in comparison to novel phrases using eye-tracking methods. Conklin and Schmitt (2008) found that idioms, examined in their research as a subset of formulaic language, were read more quickly than comparable novel phrases whether used figuratively or literally. Underwood et al. (2004) found a similar advantage based on the number of fixations, but not for total fixation length, suggesting a more complex picture of the processing of idioms and other formulaic language that also accounts for L2 disadvantages. However, neither make any claims about figurative access in comparison to literal access. Siyanova-Chanturia et al. (2011), on the other hand, found no advantage for idioms compared to novel phrases in proficient L2 users and also found that the figurative meanings of idioms required more time to retrieve than the literal interpretation, which seems to support a literal-first model of processing for L2 users.

Some L2-specific idiom processing models have been proposed in addition to the L1 models. The *Idiom Diffusion Model of Second Languages* (Liontas, 2002; 2015) proposes a two-stage comprehension model. The first stage involves prediction, eased by idioms which are the same in a learner's L1 and L2; the second stage is confirmation or replacement and/or reconstruction. Though a comprehension model, it suggests that processing is eased for translatable idioms, supported by an offline study from Irujo (1986) and a timed production task from Liontas (2002). In her *Model of Dual Idiom Representation*, Abel (2003) proposes that the important factor for L2 processing is decomposability, or the relation of the individual constituents to the idiomatic meaning. This model assumes that nondecomposable and frequently encountered idioms are represented by idiom entries in the mental lexicon, as in the *Idiom List Hypothesis*, while nondecomposable idioms are represented by lexical entries of the individual constituent words. Abel's model, however, is based solely on offline ratings. Finally, Cieślicka's (2006) *Literal Salience Model* directly addresses idiom processing and is based on online data. The model suggests that literal meanings remain most salient for L2 users, even for well-known idioms, as they are more likely to be used and encountered by learners. Based on ideas presented by Giora (1997), salient meanings are accessed more quickly than non-salient meanings. The model is based on findings from a cross-modal priming experiment that showed that access to literal meaning occurs prior to figurative meaning for L2 listeners. Using non-biasing sentences followed by literally- and figuratively-related targets, reaction times to literal targets were faster than figurative ones when compared to matched controls.

Research on L2 idiom processing is less developed than and lacks the quantity that L1 research has been afforded. While some evidence supports fast access to figurative meaning, the speed of access in comparison to novel or literal language is still inconclusive. And, like L1 research, it is unclear whether or not figurative language has its own mode of processing.

The Current Experiment

While the L1 *Idiom List Hypothesis* based on research by Bobrow and Bell (1973) has been dismissed, among other reasons (see e.g., Cacciari & Tabossi, 1988; Cacciari, 2014) based on false assumptions about slow access to figurative meaning in the absence of a biasing context, this model is not necessarily refuted by the literature for L2 users and still reflects some intuitions about the way we comprehend idioms. As Swinney and Cutler (1979) also observed, when an individual becomes aware of a highly idiomatic context, L1 listeners often become more attuned to figurative meaning occurring in natural communicative situations and might even fail to see the literal meaning of an idiom. The experiment from Bobrow and Bell (1973) presented idioms with the possibility of both literal and figurative interpretations following a biasing context containing several sentences with either literal-only or figurative-only interpretations. Participants were asked to note which interpretation—literal or figurative—they first perceived. Based on increased literal interpretations first following the literal-only contexts and figurative interpretations first following figurative-only contexts, Bobrow and Bell argued for two separate modes of processing. Following a highly figurative context, the figurative mode of processing becomes active and leads to deviation from a normal literal-first mode of processing. What Bobrow and Bell did not consider, and what Swinney and Cutler's observations suggest, is that the adjustment observed might be due to a contextual adaptation rather than a mode of processing unique to idiom or figurative language processing.

The ability of L2 listeners to adapt to their environment in language comprehension is well-documented, and there is evidence for rapid attunement to varying linguistic situations. Listeners can adapt, for example, to speech rate, surrounding noise and idiosyncrasies of a speaker both in their L1 and their L2 (see e.g., Sebastián-Galles et al., 2000). L2 listeners are also able to quickly attune to changes in their environment such as surrounding noise (McQueen & Huetting, 2012) and foreign accents (e.g., Weber et al. 2014). In addition, sequential effects of item presentation can influence the listening process. Perea and Carreiras (2003) found that listeners are able to shift their response criterion on a trial-by-trial basis to adjust to the lexical status and frequency of a previous trial. While an offline, even conscious, adaptation to the presence of figurative meaning is a common intuition, this phenomenon has not been examined in on online setting, and it warrants further research to determine if a context dependent shift based on the presence of figurative language occurs in L2 listeners—be it via a figurative mode of processing or a contextual figurative attunement.

The current experiment will revisit the idea of an idiomatic mode of processing for non-native listeners activated in a biasing experimental context using current psycholinguistic methods. We address the questions still left open in light of current research on idiom processing: Will a highly literal or idiomatic experimental context also affect online processing of figurative meaning in L2 listeners? If so, do the results support a unique mode of idiomatic processing or attunement to the figurative context?

Method

In the present English cross-modal priming study, we presented idioms with a medium degree of literal and figurative interpretation (neither highly literal nor figurative based on L2 ratings) in one of two experimental contexts to German learners of English (see Beck & Weber, 2014). Participants either encountered target idioms embedded in sentences among more sentences containing idioms and very few literal sentences or among only literal sentences.

Listeners were presented an auditory prime followed by a visual target. In a lexical decision task, German participants had to decide whether or not the visual target was a real word of English or not—reaction times (RTs) to targets are known to be faster when prime and target are semantically related compared to when they are unrelated. Facilitatory priming provides information about the processing of the auditory prime, and faster targets compared to their unrelated controls indicate the activation of meaning. For the current experiment, we were interested in both facilitatory priming for targets related to the figurative meaning of the idioms and targets related to the literal meaning of constituent words compared to matched controls. If the same kind of contextual attunement found by Bobrow and Bell (1973) applies to non-native listeners, then we would expect increased priming effects for figurative targets in an idiomatic experimental context compared to the non-idiomatic experimental context. Additionally, access to literal constituent words can give us more insight into non-native idiom processing.

Participants

Eighty-one native speakers of German were paid a small fee to participate in the experiment. Participants were University of Tübingen students who identified themselves as skilled speakers of English. One participant was excluded as she reported that she was unable to hear the stimuli.

Materials

Twenty-five target idioms were embedded at the end of non-biasing short sentences and presented in one of two varying experimental contexts containing 100 trials. The two experimental contexts differed only in 75 filler items, specifically the amount of sentences with idioms in the fillers (explained below). The target idioms had a VP syntactic structure and were controlled for familiarity, meaningfulness, literality, and translatability (English to German) as rated by L1 and L2 users (Beck & Weber, 2014). Each trial consisted of an auditory sentence prime followed by a visual target. Each sentence prime was paired with four different targets.

The four targets included literal or figurative targets and their respective unrelated control targets. Targets semantically related to the literal meaning were based on the last content word of the sentence and chosen from the Nelson et al. (1998) association norms database. For the idiom *to kick the bucket (*primed in the sentence *His uncle kicked the bucket.),* the literal target PAIL was chosen. The unrelated control word was matched for orthographic complexity and length (BOAT as a control for PAIL). Targets related to the figurative meaning of the idiom were chosen based on relation to the overall meaning of the idiom. For *to kick the bucket*, the target DIE was chosen, as the overall meaning is "to die." Similarly, figurative control targets were also controlled for orthographic complexity and length (ZOO as a control for DIE). The four lists of targets were also controlled for lexical frequency. See Table 1 for reference. The auditory sentence primes remained the same for all four lists. Targets were equally distributed across 8 lists (all four targets in two experimental contexts) and presented 400ms after the offset of the final word in each sentence.

Table 1: Sample of experimental items.

	Literal		**Figurative**	
Stimuli	Target	Control	Target	Control
His uncle *kicked the bucket*.	PAIL	BOAT	DIE	ZOO

The experiment was performed using Presentation® software (Version 17.2, www.neurobs.com). Experimental sentences were recorded by a female speaker of American English (first author) in an experimental lab setting. The eight lists were randomly distributed among participants. Each list began with four practice trials followed by 25 experimental and 75 filler trials (a total of 100 experimental trials) and was presented to an equal number of participants. Conditions were evenly distributed across lists.

High-idiomatic Context This variation included a highly figurative context by increasing the amount of idiomatic trials to a total of 75. In addition to the 25 critical trials, 50 filler sentences also embedded idioms into the end of neutral sentences. Only 25 filler trials did not include idioms.

Fifty of the filler trials contained non-word targets, and the other half contained word targets. The four lists in this variation differed only in experimental target words.

Low-idiomatic Context This variation kept the idiomatic context to the 25 experimental trials by including only literal filler items.

Half of the trials contained non-word targets, and the other half contained word targets. The four lists in this variation differed only in experimental target words.

Procedure

Participants were tested individually in a quiet room. First, participants were given instructions in English on the lexical decision task. Participants were instructed that they would hear sentences directly followed by the appearance of a word or non-word on the computer screen in front of them. Participants were told to listen to sentences and then decide whether the string of letters on the screen was a word or not. Subjects were asked to make their decision by pressing a green button with their dominant hand for 'YES' and a red button with the other hand for 'NO' as quickly and accurately as possible. Participants were also instructed that it was important both to listen and to respond to the visual targets, as they would be asked about what they heard after experiment.

Once participants understood the instructions and answered the instruction questions correctly, they could participate in the priming study. The participants listened over closed headphones, and the visual targets were presented on a laptop screen. The targets appeared on the screen 400ms after the offset of the auditorily presented sentence. The next trial began 1000ms later. Reaction times were measured from the onset of the presentation of the visual stimuli.

The experiment concluded with a short yes/no comprehension test on items and a language background questionnaire (see also Cieślicka, 2006). The entire experiment took about 15 minutes.

Analysis

Ten correct responses with RTs longer than 2000ms (0.5% of the total data) were considered outliers and were removed from these analyses. Additionally, three targets (*CHIME, SHRUB, YARN*) were answered correctly by participants only 50% or less of the total trials and were excluded from the results (29 responses or 1.5% of total data). In total, 2.0% of the data were not included in these analyses.

Table 2: Reaction times (in ms).

	Variation 1 (75% Idioms)		**Variation 2** (25% Idioms)	
	Figurative	**Literal**	**Figurative**	**Literal**
Related	749	709	737	674
Unrelated	813	739	740	711

Analyses of Variance (ANOVAs) were conducted on correct responses of the remaining RTs across participants (F_1) and across items (F_2) to examine the within-subject effects of *figurativeness* (with two levels *figurative* and *literal*) and *relatedness* (with two levels *related* and *unrelated*) and the between-subject effect of *experimental context* (with two levels *high-idiomatic* and *low-idiomatic*) between-participants and between-items. Table 2 reports the mean RTs measured from target onset for each condition, and the corresponding priming effects for RTs are shown in Figure 1.

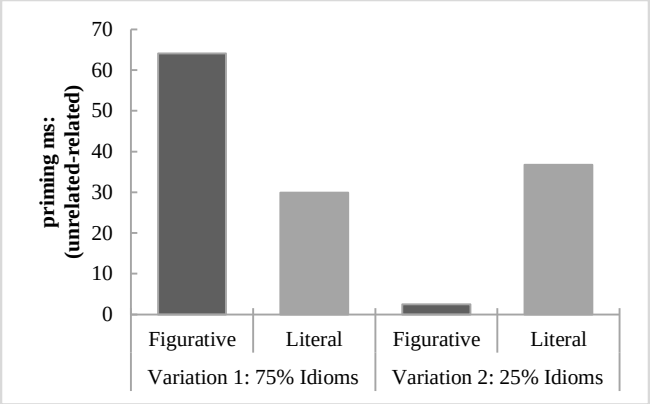


Figure 1: Priming effects (in ms).

Overall ANOVAS on RTs showed main effects of *figurativeness* (F_1[1,78]=20.29, p<.001; F_2[1,42]=11.41, p=.002), *relatedness* (F_1[1,78]=9.89, p<.01; F_2[1,42]=2.80, p>.05), and a weak interaction between *figurativeness*, *relatedness*, and *experimental context* (F_1[1,78]=4.02, p<.05; F_2<1). No other interactions were significant.

To further explore the interaction in the RT analysis, separate analyses for *figurativeness* were conducted. For *literal* targets, there was a main effect of *relatedness* (F_1[1,78]=5.88, p<.05; F_2[1,42]=2.37, p>.05) and no significant interactions. RTs for literally-related targets were faster than RTs for unrelated targets across both experimental contexts. Effects were more consistent across subjects than across items, suggesting variation between individual targets.

For *figurative* targets, there was a main effect of *relatedness* (F_1[1,78]=6.14, p<.05; F_2[1,48]=1.87, p>.05) and an interaction between *relatedness* and *experimental context* (F_1[1,78]=5.24, p<.05; F_2[1,48]=1.22, p>.05). Again, both effects are more consistent across subjects than items. RTs figuratively-related to the prime, as in the literal analysis, were faster than unrelated targets. Thus, we find facilitatory priming for relatedness in both figurative and literal targets.

In order to further explore the interaction in the *figurative* analysis, separate analyses were conducted for each *experimental context*. In the *high-idiomatic* context there was a main effect of *relatedness* (F_1[1,39]=10.58, p<.01; F_2[1,24]=2.75, p>.05). However, in the *low-idiomatic* context, no main effects were present. While facilitatory priming was present for figurative targets in the presence of the high-idiomatic context, this effect disappears in a more literal context for non-native listeners.

Results

Though there was no main effect of experimental context, our results show that the experimental context significantly impacted facilitatory priming for figurative targets. As shown in Figure 1, the facilitatory priming for figurative targets (dark grey bars) varies considerably from one variation to the other, while the literal targets (light grey bars) show no significant changes. In the high-idiomatic context, a facilitatory priming effect for figurative targets of 64ms is

observed compared to a non-significant 3ms in the low-idiomatic context. Figurative targets in this context represented the only condition in which relatedness was not a significant effect. These results provide evidence that even in a non-biasing sentence context, a highly figurative global environment can impact online processing of figurative meaning for idioms. This data is in line with the offline results from Bobrow and Bell for native participants (1973).

The priming effect for literally-related targets showed less variation—an increase of only 7ms—suggesting that while processing for figurative targets was impacted, the processing of individual literal constituents was not significantly impacted. In the case of the low-idiomatic experimental context, we can also argue that where figurative meaning is not facilitated, literal meaning is more dominant. This result also supports the L1 data collected by Bobrow and Bell (1973) as well as the L2 data from. Cieślicka (2006) and Siyanova-Chanturia et al. (2011).

Considering the impact of figurativeness, overall, literally-related targets were faster than figuratively-related targets. While this reflects the same results found by Cieślicka (2006), it does not necessarily imply a processing priority for literal meaning over figurative meaning. While the figuratively-related targets correspond to the overall meaning of the idiom, the literally-related targets correspond only to the literal meaning of the final constituent of the idiomatic phrase. It does, however, indicate that literal processing of constituents is present even when figurative meaning is processed. And, when figurative meaning is not facilitated by context, literal meaning appears to be dominant.

Our results are further supported by those gathered by Sprenger (2003), who also found strong literal constituent activation in production tasks. Our results differ, though, from Rommers, Dijkstra and Bastiaansen (2013). Rommers et al. collected EEG data from a top-down procedure which indicated that related literal targets were not activated in highly predictable idioms when the final word of an idiom was replaced with the related target. Thus, while hearing a word might activate the literal meaning of that word, the functionality of these literal words may be limited or even switched off in some cases.

The consistency of effects—generally stronger across subjects than across items—supports general knowledge of variation across idioms (see e.g., Titone & Connine 1994) and possibly varying association strength of our targets, as these were selected by this author rather than a database. While our idioms were controlled for much of this variation, the presence of remaining differences cannot be excluded.

Discussion

Based on the results of this study, we will briefly consider the fit of our data with L1 and L2 models of idiom processing.

L1 Processing Models

Our results are not compatible with the assumptions of stage models of processing. The *Idiom List Hypothesis* (literal-first) is problematic as our participants activated literal constituent meaning in addition to figurative meaning of the idiom in a high-idiomatic context. Additionally, while priming of figurative meaning increased in the high-idiomatic context, literal constituent meaning still precluded figurative meaning. This hypothesis predicts, however, that that literal meaning should not be activated at all in a high-idiomatic context. While our data does not rule out a second, idiomatic mechanism that might be primed by a global figurative context, we argue that attunement is a more suitable explanation for this phenomenon. The *Lexical Representation Hypothesis*, assuming that figurative meaning is retrieved faster than compositional processing of literal meaning, is likewise not supported by participants' RTs in our experiment. However, the tenant of this model suggesting that composition and idiom retrieval can occur simultaneously cannot and should not be dismissed.

Furthermore, our data cannot make strong claims for or against the *Configuration Hypothesis* as we included only highly familiar idioms, and this model focuses on predictable idioms based on the recognition point of the idiom. Though familiarity generally correlates with predictability (see e.g., Titone, Connine 1994), any further interpretation of compatibility would be far-reaching.

L2 Processing Models

We can make limited claims about the previously discussed L2 processing models, but it seems that our results showing online access to figurative meaning are not compatible with the challenges many of these models present. As we used only non-translatable idioms, we would expect processing difficulties for our participants based on the *Idiom Diffusion Model of Second Languages.* Although we cannot compare our results with translatable idioms to make a stronger claim, the access to figurative meaning makes a case against it.

Our data is generally compatible with the *Literal Salience Model.* Our participants responded faster to literally-related targets than to figuratively-related targets as predicted by this model. However, the results solely from this observation are not compelling enough to interpret that literal meaning has a priority over figurative meaning. Sprenger, Levelt and Kempen (2003) also found literal constituent priming in production tasks and claimed that, rather than a processing priority, activation of constituent word lemmas in addition to a superlemma, a phrasal representation of the idiom on a lexical-syntactic processing level, are activated. Though this production model does not clearly lay out a time course for processing, we assume that activation must spread to a superlemma from individual constituents, and it is possible that this activation occurs more slowly for L2 listeners than activation of an individual word lemma does.

Figurative Attunement

Our findings can generally be interpreted as compatible with current ideas on L2 listening adaptation. Based on the strong differences in processing of figurative language between the experimental variations, we argue that listeners are able to quickly adapt to a figurative context on the processing level.

Conclusion

Though our data does not provide strong evidence for or against idiom processing as its own mode of processing, separate from literal processing, it does provide strong evidence that proficient L2 listeners can have online access to figurative meaning. Additionally, the experiment supports the idea that figurative attunement is possible even in a very short amount of time, furthering evidence that L2 listeners can detect and shift their response-criterion in the presence of a highly figurative context.

Acknowledgements

Thanks to Daniel Baumgartner, James Kemmler, Ida Sanders, and Friederike Sieß for help with data collection.

References

Abel, Beate (2003): English idioms in the first language and second language lexicon: a dual representation approach. In *Second Language Research* 19 (4), pp. 329–358.

Beck, S.; Weber, A. (2014, July): L2 Idiom Processing. Presented at Cross-Linguistics Perspectives on Language Learning and Processing, Tübingen, Germany.

Cacciari, Cristina (2014): Processing multiword idiomatic strings. Many words in one? In *The Mental Lexicon* 9 (2), pp. 267–293.

Cacciari, Cristina; Tabossi, Patrizia (1988): The Comprehension of Idioms. In *Journal of Memory and Language* 27, pp. 668–683.

Cieślicka, Anna (2006): Literal salience in on-line processing of idiomatic expressions by second language learners. In *Second Language Research* 22 (2), pp. 115–144.

Cieślicka, Anna (2013): Second Language Learners' Processing of Idiomatic Expressions: Does Compositionality Matter. In Krystyna Droździał-Szelest, Mirosław Pawlak (Eds.): *Psycholinguistic and sociolinguistic perspectives on second language learning and teaching. Studies in honor of Waldemar Marton*. Berlin, New York: Springer.

Conklin, Kathy; Schmitt, Norbert (2008): Formulaic Sequences: Are They Processed More Quickly than Nonformulaic language by Native and Nonnative Speakers? In *Applied Linguistics* 29 (1), pp. 72–89.

Cooper, Thomas C. (1999): Processing of Idioms by L2 Learners of English. In *TESOL Quarterly* 33 (2), pp. 233–304.

Gibbs, Raymond W.; Nayak, Nandini P.; Cutting, Cooper (1989): How to kick the bucket and not decompose: Analyzability and idiom processing. In *Journal of Memory and Language* 28 (5), pp. 576–593.

Giora, Rachel (1997): Understanding figurative and literal language: The graded salience hypothesis. In *Cognitive Linguistics* 8 (3), pp. 183–206.

Irujo, Suzanne (1986): Don't Put Your Leg in Your Mouth: Transfer in the Acquisition of Idioms in a Second Language. In *TESOL Quarterly* 20 (2), pp. 287–304.

Liontas, John I. (2002): Context and idiom understanding in second languages. In *EUROSLA Yearbook* 2, pp. 155–185.

Liontas, John I. (2015): Straight from the Horse's Mouth: Idiomaticity Revisited. In Roberto R. Heredia, Anna B. Cieślicka (Eds.): Bilingual figurative language processing: Canbridge University Press, pp. 301–340.

Liu, Dilin (2008): Idioms. Description, comprehension, acquisition, and pedagogy. New York: Routledge (ESL & applied linguistics professional series).

McQueen, James M.; Huetting, F. (2012): Changing only the probability that spoken words will be distorted changes how they are recognized. In *Journal of the Acoustical Society of America* 131 (1), pp. 509–517.

Nelson, D. L.; McEvoy, C. L.; Schreiber, T. A. (1998): The University of South Florida word association, rhyme, and word fragment norms. Available online at http://www.usf.edu/FreeAssociation/.

Perea, Manuel; Carreiras, Manuel (2003): Sequential effects in the lexical decision task: The role of item frequency of the previous trial. In *The Quarterly Journal of Experimental Psychology* 56A (3), pp. 385–401.

Rommers, Joost; Dijkstra, Ton; Bastiaansen, Marcel (2013): Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension. In *Journal of Cognitive Neuroscience* 25 (5), pp. 762–776.

Sebastián-Galĺes, N.; Dupoux, E.; Costa, A.; Mehler, J. (2000): Adaptation to time-compressed speech: Phonological determinants. In *Perception & Psychophysics* 62 (4), pp. 834–842.

Siyanova-Chanturia, Anna; Conklin, Kathy; Schmitt, Norbert (2011): Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. In *Second Language Research*, pp. 1–22.

Sprenger, Simone A. (2003): Fixed expressions and the production of idioms. Dissertation. MPI, The Netherlands.

Swinney, David A.; Cutler, Anne (1979): The Access and Processing of Idiomatic Expressions. In *Journal of Verbal Learning and Verbal Behavior* 18, pp. 523–534.

Tabossi, Patrizia; Fanari, Rachele; Wolf, Kinou (2009): Why are idioms recognized fast? In *Memory & Cognition* 37 (44), pp. 529–540.

Titone, Debora A.; Connine, Cynthia M. (1994): Descriptive Norms for 171 Idiomatic Expressions: Familiarity, Composability, Predictability, and Literality. In *Metaphor and Symbolic Activity* 9 (4), pp. 247–270.

Underwood, Geoffrey; Schmitt, Norbert; Galpin, Adam (2004): The Eyes Have It: An Eye-Movement Study into the Processing of Formulaic Sequences. In Norbert Schmitt (Ed.): *Formulaic Sequences*. Amsterdam: John Benjamins.

Weber, Andrea; Di Betta, Anna Maria; McQueen, James M. (2014): Treack or trit: Adaptation to genuine or arbitrary foreign accents by monolingual and bilingual listeners. In *Journal of Phonetics* 46, pp. 34–51.

Synthesized size-sound sound symbolism

Gwilym Lockwood (gwilym.lockwood@mpi.nl)
Neurobiology of Language Department
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525XD Nijmegen
The Netherlands

Peter Hagoort (peter.hagoort@mpi.nl)
Neurobiology of Language Department
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525XD Nijmegen
The Netherlands

Mark Dingemanse (mark.dingemanse@mpi.nl)
Language and Cognition Department
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525XD Nijmegen
The Netherlands

Abstract

Studies of sound symbolism have shown that people can associate sound and meaning in consistent ways when presented with maximally contrastive stimulus pairs of nonwords such as bouba/kiki (rounded/sharp) or mil/mal (small/big). Recent work has shown the effect extends to antonymic words from natural languages and has proposed a role for shared cross-modal correspondences in biasing form-to-meaning associations. An important open question is how the associations work, and particularly what the role is of sound-symbolic matches versus mismatches. We report on a learning task designed to distinguish between three existing theories by using a spectrum of sound-symbolically matching, mismatching, and neutral (neither matching nor mismatching) stimuli. Synthesized stimuli allow us to control for prosody, and the inclusion of a neutral condition allows a direct test of competing accounts. We find evidence for a sound-symbolic match boost, but not for a mismatch difficulty compared to the neutral condition.

Keywords: sound symbolism; iconicity; ideophones; cross-modal correspondences; language

Introduction

Research into iconicity, where aspects of a word's form reflect aspects of its meaning, has considerably nuanced the classical view of words as wholly arbitrary (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Lockwood & Dingemanse, 2015; Perniss, Thompson, & Vigliocco, 2010). Iconicity is found across languages, both spoken (Dingemanse, 2012) and signed (Emmorey, 2014; Perniss & Vigliocco, 2014), and plays a significant role in language acquisition (Imai & Kita, 2014; Perry, Perlman, & Lupyan, 2015; Yoshida, 2012), language evolution (Cuskley & Kirby, 2013; Verhoef, Kirby, & de Boer, 2015; Zlatev, 2014), and language processing (Lockwood & Tuomainen, 2015; Meteyard, Stoppard, Snudden, Cappa, & Vigliocco, 2015; Westbury, 2005); but it is still unclear exactly how.

Studies have shown that people are sensitive to the meanings of sound-symbolic words in a foreign language, associate certain artificial words (henceforth nonwords) with certain properties depending on their vowels and consonants, and learn new words better when there is a sound-symbolic relationship between form and meaning (Aveyard, 2012; Davis, 1961; Dingemanse, Schuerman, Reinisch, Tufvesson, & Mitterer, in press; Kovic, Plunkett, & Westermann, 2010; Lupyan & Casasanto, 2015). The general consensus is that cross-modal correspondences and perceptuo-motor analogies between sounds and meanings provide a way of bridging the two domains in consistent ways (Perniss & Vigliocco 2014). However, there is not yet a satisfactory answer to which cross-modal correspondences are implicated in sound symbolism or how exactly these correspondences help people to make mappings. Many experiments have relied on forced choice decisions where participants judge which nonword goes with which property (Bremner et al., 2013; Davis, 1961; Köhler, 1929; Nielsen & Rendall, 2011, 2013; Sapir, 1929). This sets up a paradigm where participants consistently identify sound-symbolically matching sets of stimuli (e.g. the nonword *bouba* and the round shape, the nonword *kiki* and the spiky shape). The combined weight of these experiments is an affirmation of the existence and prevalence of sound symbolism. However, these studies do not address how the associations affect the participants' choices: does a sound-symbolic match provide a mapping boost helping the participant to choose the matching set of stimuli, or does the sound-symbolic mismatch provide a cue to exclude that set of stimuli, or is it a combination of both? Moreover, it is not always clear whether a mismatch is an actual clash or whether mismatch is simply taken to mean "not matching".

Other experimental designs suggest that it is not as simple as the two-alternative forced choice literature makes out (Monaghan, Mattock, & Walker, 2012; Westbury, 2005).

Rating experiments which vary sound-symbolic representations of size along a graded scale have shown that people judge sound symbolism in a graded fashion rather than simply as being there or not (Thompson & Estes, 2011). A graded model of sound symbolism is more detailed, but leaves the same question open: is it driven equally at both ends of the graded spectrum? Learning experiments have shown that it may be one end of a graded spectrum which drives sound-symbolic associations, such as an association between labial sounds and roundness creating an incidental association between non-labial sounds and spikiness (Jones et al., 2014). While it appears that the spiky—round spectrum does not map directly onto the labial/voiced—non-labial/voiceless spectrum suggested by two-alternative forced-choice studies, it remains to be seen whether this imbalance holds for other domains. Finally, other learning experiments suggest there is a sound-symbolic processing bias, but that it is weak and can be overcome with training (Nielsen & Rendall, 2012).

We ran a similar learning experiment with Japanese ideophones (Lockwood, Dingemanse, & Hagoort, 2016) rather than nonwords. In this study, we taught the ideophones to a group of Dutch participants with no knowledge of Japanese. For half the ideophones, the participants learned the real Dutch translations (e.g. *dik*, or fat, for *bukubuku*, which means fat); for the other half, the participants learned the opposite Dutch translations (e.g. *verdrietig*, or sad, for *ukiuki*, which means happy). In a recognition task, participants remembered the ideophones in the real condition far better than the ideophones in the opposite condition (86.1% recognition accuracy vs. 71.1%). When we repeated the experiment with a set of arbitrary adjectives and another group of participants, there was no sound-symbolic effect across the two conditions (79.1% recognition accuracy in the real adjective condition, 77% in the opposite adjective condition). This is in line with the nonword studies that show a mapping boost for sound-symbolically matching stimuli and a mapping difficulty for sound-symbolically mismatching stimuli, although it is not possible to say whether the effect is driven by one or both of these mapping strategies.

In another sound symbolism study with real words, Nygaard et al. (Nygaard, Cook, & Namy, 2009) found a different result. Participants learned Japanese words with their real translations and their opposite translations equally well, but learned words with random translations less well. They proposed that cross-modal correspondences help sound-to-meaning mappings for both matching and mismatching words, as antonym pairs are conceptually very close. Under this interpretation, sound symbolism in learning tasks is not a graded effect. Rather, the lack of any sound-to-meaning correspondence makes word learning harder than having a mismatching or counterintuitive cross-modal clash to build upon.

While using real words from real languages avoids the ecological validity problem of nonwords, there are other confounds which cannot be completely be ruled out. Firstly, sound-symbolically congruent and incongruent prosody has been shown to affect meaning judgement (Nygaard, Herold, & Namy, 2009). It is possible that our Dutch participants were just picking up on the prosody of the Japanese ideophones rather than the sounds themselves. Secondly, orthography is a constant confound in tasks with both nonwords and real words (Cuskley, Simner, & Kirby, 2015).

This paper builds on Lockwood et al. (2016) by creating nonwords in the shape of Japanese ideophones, synthesizing the sound stimuli, and limiting the meanings to a simple size contrast. This lets us investigate a spectrum of sound-symbolically matching, mismatching, and neutral stimuli. Here, we take neutral to mean that a relation that is neither an obvious match nor an obvious mismatch. The use of a speech synthesizer to generate the sounds eliminates possible prosodic differences which in natural speech may indicate sound-symbolic contrasts (Dingemanse et al. in press). Keeping translations to "big" and "small" lets us work within a well-attested sound-symbolic framework where participants' subjective ratings are constrained and predictable.

Including a neutral condition while ensuring that the mismatch condition is a cross-modal clash (rather than just a lack of cross-modal correspondence) allows us to adjudicate between different theoretical accounts for sound-symbolic effects. If the participants learn matching nonwords better than neutral nonwords, but there is no difference between neutral and mismatching nonwords, this is evidence for a sound-symbolic match boost as in Lockwood et al. (2016) and Jones et al. (2014). If participants learn matching nonwords better than neutral nonwords and neutral nonwords better than mismatching nonwords, this is evidence for a graded sound-symbolic rating effect as in Nielsen and Rendall (2012) transferring to sound-symbolic learning. Finally, if participants learn the neutral nonwords worse than both the matching and mismatching nonwords, this is evidence for cross-modal correspondences boosting learning regardless of whether the associations correspond or clash, as in Nygaard et al. (2009).

Methods

In the main experiment, 30 participants learned 36 nonwords in three learning rounds, and were tested immediately afterwards. We first describe the stimuli design and selection.

Stimuli design

We created nonwords in the CVCV-CVCV pattern found in Japanese ideophones. These nonwords were deliberately created in order to sound big, neutral, or small, based on attested cross-modal correspondences between sound and size. Big-sounding nonwords featured voiced stops and mid/low back vowels. Small-sounding nonwords featured voiceless stops and high front vowels. Neutral-sounding nonwords featured mid-vowels, and had either all voiced, all unvoiced, or a mix of voiced and unvoiced stops. Table 1

shows the distribution of vowels and consonants used in each word type.

Table 1: sound distributions across nonword types

Nonword type	Consonants	Vowels
big-sounding	[b] [d] [g]	[a] [o]
small-sounding	[p] [t] [k]	[i] [ʏ]
neutral-sounding	[p] [b] [t] [d] [k] [g]	[ɛ] [ə]

We wrote a Matlab script to generate all possible combinations of words according to this pattern where the consonant was not repeated (e.g. *bobaboba*), and this resulted in 192 possible nonwords.

We then synthesized the nonwords using the Dutch voice nl2 from the diphone synthesizer MBROLA (Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996). All nonwords were given the same pitch, vowel durations, and prosodic contours.

Stimuli rating pre-test

28 native Dutch speakers listened to each synthesized nonword under the impression that they were size adjectives from a real language. Participants rated how big the word sounded on a Likert scale of 1-7, where 1 represented really small, 4 neutral, and 7 really big. Participants were also told to indicate whether their rating was influenced by a similar-sounding Dutch word in order to detect lexical confounds. We removed 17 nonwords where at least four participants indicated that it reminded them of something.

In the remaining 175 nonwords, participants consistently judged the big-sounding words as big (mean=5.57), the neutral-sounding words as neutral (mean=3.90), and the small-sounding words as small (mean=2.68). This was a highly significant effect according to a one-way ANOVA ($F=694.3$, $p<0.001$), and post-hoc Bonferroni tests showed that this difference was significant between each condition (all ps <0.001). This is shown in Figure 1.

We selected 36 nonwords for the full experiment according to their mean ratings. For the big-sounding nonwords, we chose the 12 highest-rated nonwords; for the small-sounding nonwords, we chose the 12 lowest-rated nonwords; and for the neutral-sounding nonwords, we chose the 12 nonwords which were rated most closely to 4. All 36 nonwords were from the originally designated condition, i.e., all 12 big nonwords were nonwords which we designed to sound big, and so on.

All nonwords meant either *groot* (big) or *klein* (small). This set up three conditions: nonwords that meant big (or small) and sounded big (or small) were sound-symbolically matching, nonwords that meant big (or small) but sounded small (or big) were sound-symbolically mismatching, and nonwords that meant big or small but neither obviously matching nor mismatched were neutral. This is illustrated in Table 2. Correspondences between onsets in the nonwords and translations were controlled across conditions.

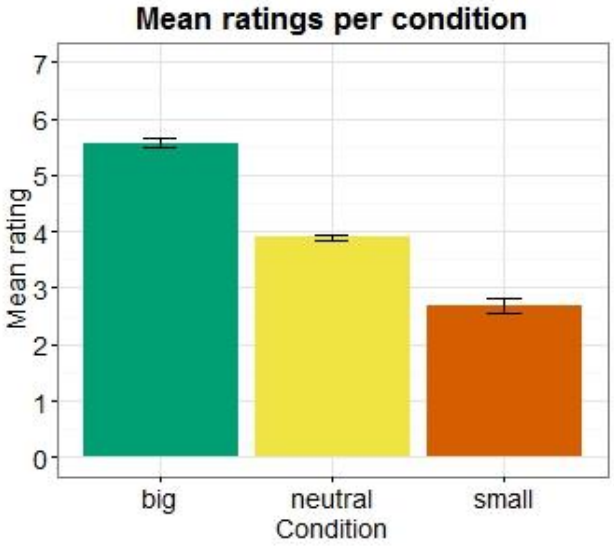


Figure 1: Size ratings per condition

Table 2: examples of learning conditions and nonwords

Condition	Nonword	Translation
Match	badobado	groot
	tʏpitʏpi	klein
Neutral	detədetə	groot
	gɛpɛgɛpɛ	klein
Mismatch	gogagoga	klein
	tipitipi	groot

Main experiment

Participants had three learning rounds in which to learn the nonwords, and then a test round immediately afterwards. They were told that the words came from an African language with a complicated adjective agreement system; in a post-experiment debriefing they were informed that the words were artificial. Item translations were counterbalanced across participants. The procedure is illustrated in Figure 2.

We used Presentation to present the stimuli and record responses. In the learning round, the initial Dutch word was presented for 1000ms with 100ms of jitter, followed by a fixation cross for 1000ms with 100ms of jitter. As the nonword was played over the speakers, a blank screen was presented for 2000ms with 200ms of jitter. This was again followed by a fixation cross. The final screen with the nonword and its Dutch meaning was presented until participants were happy to move onto the next item. Between trials, a blank screen was presented, followed by a fixation cross to announce the beginning of the next trial. Timings in the test round were identical, except that a question mark was presented instead of a blank screen while the nonword played. Participants responded by button press for yes/no answers.

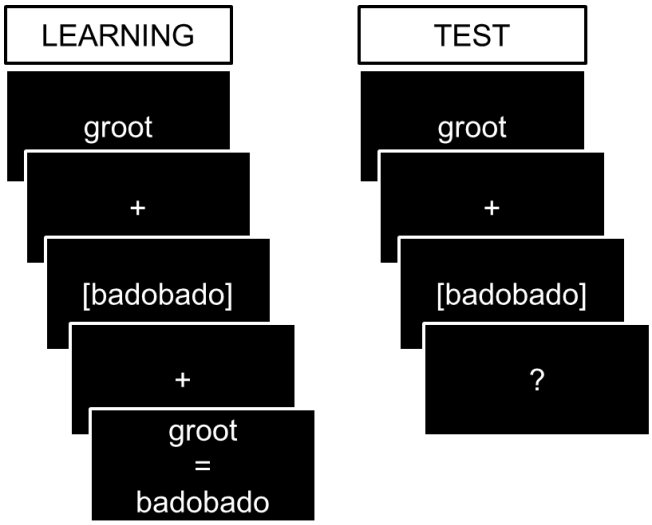


Figure 2: Learning and test round procedures.

We tested 33 native Dutch speaking participants (4m, 29f) aged 18-26 (mean: 21y 4m) with normal or corrected-to-normal vision, recruited from the MPI participant database. Three were discarded due to issues with the Presentation script, leaving us with 30 participants in total. This sample size is identical to Lockwood et al. (2016). However, the reduction in the number of items to learn per condition means that more participants are needed to match the power of that study. Therefore, this experiment is intended as an initial experiment to be replicated with a larger sample size.

Results

Participants identified nonwords at 75.56% accuracy in the match condition, at 66.11% accuracy in the neutral condition, and at 62.50% accuracy in the mismatch condition. This is shown in Figures 3 and 4. Error bars in Figure 4 represent standard error. Mean accuracy was consistent across antonym meanings (match: big = 74.44%, small = 76.67%; neutral: big = 70.56%, small = 61.67%; mismatch: big = 63.89%, small = 61.11%. All ps>0.1).

As the dependent variable was binary—correct or incorrect—we analyzed the responses using a mixed-effects logit model with the glmer function of the lme4 (versions 1.1-8) package in R. The data was modelled by including a per-participant and per-nonword random adjustment to the fixed intercept with a random slope for the fixed effect by participant. The condition was sum contrast coded to compare match to neutral and neutral to mismatch.

Model comparison between a model with condition as a fixed effect and a model with no fixed effect showed that condition was a significant fixed effect (χ^2=8.36, p=0.015). Secondly, the best model included a fixed effect of condition, a random effect by participant with random intercepts and random slopes by condition, and random intercept by nonword. This model showed that participants did better in the match condition than the neutral condition (β=0.48, SE=0.20, p=0.017), but found no evidence for a difference in performance in the neutral and mismatch conditions (β=-0.11, SE=0.21, p=0.60).

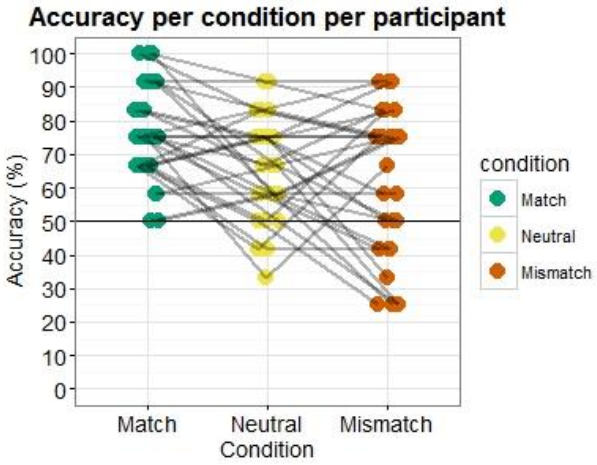


Figure 3: test round results per participant

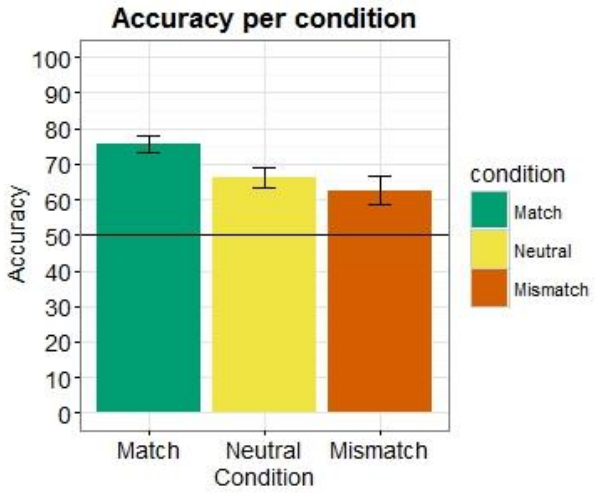


Figure 4: overall test round results

Discussion

Sound symbolism research has shown that cross-modal correspondences help people make mappings between sound and meaning. However, it is unclear whether this is because cross-modal correspondences provide a mapping boost or because a lack of a correspondence causes a mapping difficulty. In this study, we build on previous sound-symbolic word learning research by explicitly controlling the type of sound-symbolic relationship in each condition. Participants learned nonwords which had a variety of sound-symbolic cues to help scaffold word learning. Nonwords in the match condition had cross-modal correspondences between their sounds and meaning; nonwords in the mismatch condition had cross-modal clashes between their sounds and meaning; and nonwords in the neutral condition had neither matching nor mismatching cross-modal information.

Participants learned the nonwords in the match condition better than the nonwords in the neutral condition, but there was no difference in participants' performance in the neutral and mismatch conditions, and nor was there any difference between how well participants learned nonwords meaning big and small. This suggests that sound-symbolic effects in learning, and perhaps other behavioural tasks, are due to cross-modal correspondences providing a mapping boost. It also suggests that cross-modal mismatches do not provide a mapping boost, but nor do they provide an increased mapping difficulty (although mean scores suggest a possible graded effect, which will be examined in a replication with a larger sample size). This provides initial support for Lockwood et al. (2016), Jones et al. (2014), and Imai et al. (2014; Imai, Kita, Nagumo, & Okada, 2008), whose learning experiments have previously suggested that sound-symbolic bootstrapping depends on the boost effect from matching cross-modal correspondences. It also suggests that the graded perception of sound symbolism in rating tasks (such as in Nielsen & Rendall, 2011; Thompson & Estes, 2011, and indeed, the stimuli selection pre-test for this study) does not extend to a graded learning effect. Finally, it provides some evidence against the proposal that any kind of cross-modal associations, corresponding or clashing, are better for facilitating sound-symbolic mappings than no cross-modal associations at all. However, this does not rule out the findings of Nygaard et al. (2009). In their experiments, the learning phase was far longer and continued until participants reached a ceiling effect in their accuracy responses. It is possible that there is an initial sound-symbolic match boost during the first stages of word learning, while any kind of cross-modal association can help scaffold word learning during later stages of learning and consolidation.

A replication of this study with a larger sample size will provide further evidence of whether sound symbolism boosts word learning through cross-modal correspondences rather than other factors. Moreover, it will allow us to explore individual differences in sound symbolism during learning. This study shows that participants learned the matching nonwords better than the neutral nonwords, while there was no evidence for a difference in how well the participants learned neutral and mismatching nonwords. This is most obviously shown in Figure 4. However, the dotplots in Figure 3 suggest that it may not quite be so simple. Participants appear to be split, where approximately half learn the neutral nonwords better than the mismatching nonwords, and approximately half learn the mismatching nonwords better than the neutral nonwords. It is possible that some participants learn words better when there is a cross-modal association between sound and meaning, whether corresponding or clashing, while other participants learn in a way that reflects the graded effect of sound-symbolic perception. A larger sample size in a follow-up replication can explore these individual differences fully.

Finally, this study only addressed size symbolism for consonant voicing and vowel position, and we cannot assume that sound symbolism works this way for all form-meaning mappings.

In summary, we conducted a learning task equally inspired by findings from natural language iconicity and nonword studies. Using synthesized words in a constrained semantic space allowed us to adjudicate between different proposals about how sound symbolism affects learning. The evidence points to a match boost but not a mismatch difficulty, clarifying the role of cross-modal correspondences in sound-symbolic word learning.

Acknowledgments

This work was funded by an International Max Planck Research Scholarship awarded to Gwilym Lockwood and by a Veni grant from NWO (Netherlands Organisation for Scientific Research) awarded to Mark Dingemanse.

References

Aveyard, M. E. (2012). Some consonants sound curvy: Effects of sound symbolism on object recognition. *Memory & Cognition*, *40*(1), 83–92. http://doi.org/10.3758/s13421-011-0139-3

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, *126*(2), 165–172. http://doi.org/10.1016/j.cognition.2012.09.007

Cuskley, C., & Kirby, S. (2013). Synaesthesia, cross-modality, and language evolution. In J. Simner & E. M. Hubbard (Eds.), *Oxford Handbook of Synaesthesia* (pp. 869–907). Oxford: Oxford University Press.

Cuskley, C., Simner, J., & Kirby, S. (2015). Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research*, 1–12. http://doi.org/10.1007/s00426-015-0709-2

Davis, R. (1961). The Fitness of Names to Drawings. a Cross-Cultural Study in Tanganyika. *British Journal of Psychology*, *52*(3), 259–268. http://doi.org/10.1111/j.2044-8295.1961.tb00788.x

Dingemanse, M. (2012). Advances in the Cross-Linguistic Study of Ideophones. *Language and Linguistics Compass*, *6*(10), 654–672. http://doi.org/10.1002/lnc3.361

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*(10), 603–615. http://doi.org/10.1016/j.tics.2015.07.013

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (in press). What sound symbolism can and cannot do. *Language*.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings.* (Vol. 3, pp. 1393–1396 vols.3). http://doi.org/10.1109/ICSLP.1996.607874

Emmorey, K. (2014). Iconicity as structure mapping. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130301. http://doi.org/10.1098/rstb.2013.0301

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298. http://doi.org/10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65. http://doi.org/10.1016/j.cognition.2008.07.015

Jones, M., Vinson, D., Clostre, N., Zhu, A. L., Santiago, J., & Vigliocco, G. (2014). The bouba effect: sound-shape iconicity in iterated and implicit learning. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. (pp. 2459–2464). Québec.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, *114*(1), 19–28. http://doi.org/10.1016/j.cognition.2009.08.016

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Language Sciences*, 1246. http://doi.org/10.3389/fpsyg.2015.01246

Lockwood, G., Dingemanse, M., & Hagoort, P. (2016). Sound-Symbolism Boosts Novel Word Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. http://doi.org/10.1037/xlm0000235

Lockwood, G., & Tuomainen, J. (2015). Ideophones in Japanese modulate the P2 and late positive complex responses. *Language Sciences*, 933. http://doi.org/10.3389/fpsyg.2015.00933

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, *7*(2), 167–193. http://doi.org/10.1017/langcog.2014.21

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*. http://doi.org/10.1016/j.neuropsychologia.2015.01.042

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1152–1164. http://doi.org/10.1037/a0027747

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*(2), 115–124. http://doi.org/10.1037/a0022268

Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition*, *4*(2), 115–125. http://doi.org/10.1515/langcog-2012-0007

Nielsen, A., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *67*(2), 153–163. http://doi.org/10.1037/a0030553

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*(1), 181–186. http://doi.org/10.1016/j.cognition.2009.04.001

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. *Cognitive Science*, *33*(1), 127–146. http://doi.org/10.1111/j.1551-6709.2008.01007.x

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology*, *1*(E227). http://doi.org/10.3389/fpsyg.2010.00227

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130300. http://doi.org/10.1098/rstb.2013.0300

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and Its Relation to Lexical Category and Age of Acquisition. *PLoS ONE*, *10*(9), e0137147. http://doi.org/10.1371/journal.pone.0137147

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225–239. http://doi.org/10.1037/h0070931

Thompson, P., & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2392–2404. http://doi.org/10.1080/17470218.2011.605898

Verhoef, T., Kirby, S., & de Boer, B. (2015). Iconicity and the Emergence of Combinatorial Structure in Language. *Cognitive Science*. http://doi.org/10.1111/cogs.12326

Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language*, *93*(1), 10–19. http://doi.org/10.1016/j.bandl.2004.07.006

Yoshida, H. (2012). A Cross-Linguistic Study of Sound Symbolism in Children's Verb Learning. *Journal of Cognition and Development*, *13*(2), 232–265. http://doi.org/10.1080/15248372.2011.573515

Zlatev, J. (2014). Human Uniqueness, Bodily Mimesis and the Evolution of Language. *Humana.Mente Journal of Philosophical Studies*, *27*, 197–219.

Examining the Specificity of the Seductive Allure Effect

Emily J. Hopkins (hopkinse@sas.upenn.edu)
Deena Skolnick Weisberg (deena.weisberg@psych.upenn.edu)
Department of Psychology, 3720 Walnut St., Solomon Labs
Philadelphia, PA 19104 USA

Jordan C. V. Taylor (jordt@sas.upenn.edu)
Department of Philosophy, Claudia Cohen Hall, Room 433
Philadelphia, PA 19104 USA

Abstract

Previous work has found that people feel significantly more satisfied with explanations of psychological phenomena when those explanations contain neuroscience information — even when this information is entirely irrelevant to the logic of the explanations. This *seductive allure effect* was first demonstrated by Weisberg, Keil, Goodstein, Rawson, & Gray (2008), and has since been replicated several times in independent labs (e.g., Fernandez-Duque, Evans, Christian, & Hodges, 2014; Rhodes, Rodriguez, & Shah, 2014; Weisberg, Taylor, & Hopkins, 2015). However, these studies only examined psychological explanations with added neuroscience information. The current study thus investigated the generality of this effect and found that the seductive allure effect occurs across several scientific disciplines whenever the explanations include reference to smaller components or more fundamental processes. These data suggest that people have a general preference for reductive explanations.

Keywords: seductive allure; explanations; decision-making

Introduction

What is the relationship between the form of an explanation and its content? In ideal circumstances, the quality of an explanation should be determined by its success at generating understanding of the target phenomenon; form should matter less, if at all. However, there are many cases where the form of an explanation erroneously influences people's judgment of its quality, as when people judge longer explanations as better (Kikas, 2003; Langer, Blank, & Chanowitz, 1978). Similarly, people often feel that they have gained a sense of understanding from statements or situations that aren't actually explanatory (see Trout, 2002). For example, both adults (Lombrozo & Carey, 2006) and children (Kelemen, 1999) preferentially endorse teleological explanations that refer to goals or end-states, even when mechanistic explanations would be more appropriate.

One particularly interesting instance of this kind of error is the *seductive allure effect* in psychology: People judge explanations of psychology findings as better when those explanations contain logically irrelevant neuroscience information (Weisberg et al., 2008). That is, people feel that they understand a psychological phenomenon better when it is described using the language of neuroscience, although this language should make no difference. Further, this effect is much stronger for poor-quality, circular explanations. Participants judged bad explanations as significantly better when they contained added neuroscience terminology. Ratings of explanations that were already of high quality — in most cases, these were the explanations that researchers themselves gave for the psychological phenomena — were unaffected by added neuroscience information.

Although this finding has been replicated several times, demonstrating its robustness (Fernandez-Duque et al., 2015; Rhodes, Rodriguez, & Shah, 2014; Weisberg, Taylor, & Hopkins, 2015), it is still unclear why this effect happens. One possibility is that it is specific to psychology and neuroscience; something about neuroscientific language in particular plays a role in improving explanations of psychological phenomena. However, recent work has failed to identify the mechanism by which neuroscience content may have this effect. Although early evidence suggested that neuroscience images influence people's judgments (McCabe & Castel, 2008), these results have failed to replicate (see Farah & Hook, 2013, for review). Additionally, neuroscience jargon (e.g., "fMRI imaging") has no effect over and above references to the brain in plain language (e.g., "brain scans"; Weisberg, Taylor, & Hopkins, 2015, Study 3). Therefore, neither appealing imagery nor scientific jargon is responsible for making neuroscience information seductive. Although it is still possible that some other property unique to the pairing of psychology and neuroscience is responsible for the seductive allure effect, an alternative explanation is that this effect is representative of a more general bias in judging explanations.

The current work investigates one candidate for this general bias: a preference for reductive explanations (see Craver, 2007). Scientific reductionism holds that explanatory elements from one discipline may be *reduced* to elements of a more fundamental or basic discipline if the laws of the "higher" discipline follow as logical consequences of the more fundamental one (Nagel, 1961). People may thus judge explanations of psychological phenomena that contain irrelevant neuroscience information as better because the brain plays this reductive role for psychological states. To test the hypothesis that the seductive allure effect is indicative of a general preference for reduction, we presented subjects with descriptions of

phenomena across a range of sciences in a plausible reductive hierarchy (Figure 1); in this hierarchy, each science is most immediately explainable in terms of the one below it. If people do have a general preference for reduction, the seductive allure effect should be seen any time an explanation contains reference to the next level down on the hierarchy, leading to preferences of chemical explanations for biological phenomena, for example.

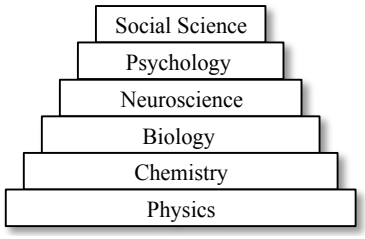


Figure 1: Hierarchy of sciences.

For each phenomenon, we constructed four explanations, according to a Quality (good/bad) x Explanation Level (horizontal/reductive) design. Horizontal explanations refer only to the science from which the phenomenon itself is drawn (e.g., biological explanations for biological phenomenon). Reductive explanations include reference to the next level downwards in the hierarchy (e.g., chemical explanations for biological phenomenon). If participants show a general preference for reduction, they should judge reductive explanations as better than horizontal explanations for all sciences, even though the explanatory content of both is the same. If the seductive allure effect is unique to the pairing of psychology and neuroscience, however, we should observe this preference only for psychology and not for the other sciences.

Method

Participants

Participants were recruited from two different populations: Workers on Amazon's Mechanical Turk (n = 167) and undergraduate students enrolled in psychology classes at the University of Pennsylvania (n = 152). MTurk workers were paid for their participation, and undergraduate students received course credit. Some of these participants (20 MTurk workers and 40 undergraduates) were excluded from the sample for failing attention check questions (described in the Procedure). The final sample used for all analyses thus consisted of 147 MTurk workers and 112 undergraduates. MTurk workers (80 women, 55 men, 12 did not report gender) were 39.8 years of age on average (range: 19-71), and undergraduates (64 women, 44 men, 4 did not report gender) were 19.8 years of age on average (range: 18-23). Most of the MTurk workers (89.8%) had completed at least some college. Among the undergraduates, 36.0% were freshmen, 28.8% were sophomores, 20.7% were juniors, 13.5% were seniors; 1 participant did not report his or her year.

Design

All participants completed an online survey hosted by Qualtrics. The explanations task used a 2 (Explanation level: horizontal, reductive) x 2 (Quality: good, bad) x 6 (Science: physics, biology, chemistry, neuroscience, psychology, social science) design. Explanation level was between-subjects: Participants were randomly assigned to either the horizontal (74 MTurk workers, 54 undergraduates) or reductive (73 MTurk workers, 58 undergraduates) condition. Quality and Science were within-subjects variables: All participants rated two explanations from each science, one good and one bad.

Materials

The Rating Explanations task used 24 phenomena (four per science). The phenomena described concepts, principles, or research findings from each of the six sciences. Each phenomenon had four corresponding explanations: horizontal-good, horizontal-bad, reductive-good, reductive-bad (Table 1). The good versions of the explanations were the ones that researchers or textbooks provided for the phenomena; all explanations were verified by experts in the respective fields. The bad explanations were worded so as to provide no information regarding why the phenomena occurred. They were either circular restatements of the phenomenon, or they provided additional information that was irrelevant with no mechanistic information that could explain the phenomenon. Experts in each field confirmed that the bad explanations were non-explanatory.

Both horizontal-good and horizontal-bad explanations used only terminology and concepts from the same discipline as the phenomenon. That is, biological phenomena were described only in biological terms, chemical phenomena were described only in chemical terms, etc. Explanations in the reductive condition used terminology from the discipline below that of the phenomena in our reductive hierarchy: biological explanations were supplemented with chemistry information, chemistry explanations were supplemented with physics information, etc. For phenomena from the domain of physics, the reductive explanations referred to smaller particles and/or more fundamental forces (e.g., reducing "friction" to "vibration of molecules"). Importantly, the reductive information did not add any additional explanatory information beyond what was already contained in the horizontal explanation; this was also verified by experts.

For each phenomenon, the four versions of the explanation were matched as closely as possible outside of the manipulations for quality and explanation level. The added reductive text was identical for good and bad versions of the explanation. Length of explanation was carefully

Table 1: Sample Phenomenon from Biology

Male anole lizards bob their heads up and down rhythmically as part of a mating ritual to attract females. They typically increase their rate of head-bobbing when they see a female lizard of their species. However, their rate of head-bobbing also increases when they see another male lizard of the same species, even if no female lizards are present. *Why do male lizards bob their heads when other males are nearby?*		
	Good	Bad
Horizontal	This happens because the male lizards are extremely territorial, and head-bobbing is a distinctive behavior typical of this particular species of lizard. During mating season when they are in competition with each other for females, males use various dominance displays to defend their territory. They perceive other males as a threat and engage in increased head-bobbing, which is a sign of aggression.	This happens because the male lizards are seeking mates, and head-bobbing is a distinctive behavior typical of this species of lizard. During mating season when they are trying to attract females, males use a variety of behaviors that are characteristic of anole lizards. They perceive the presence of other males and engage in increased head-bobbing, which is commonly seen during mating season.
Reductive	This happens because the male lizards are extremely territorial. During mating season when they are in competition with each other for females, males use various dominance displays to defend their territory. They perceive other males as a threat and engage in increased head-bobbing, which is a sign of aggression. **Aggressive behavior is known to be associated with elevated levels of testosterone and other aggression-enabling hormones.**	This happens because the male lizards are seeking mates. During mating season when they are trying to attract females, males use a variety of behaviors that are characteristic of lizards. They perceive the presence of other males and engage in increased head-bobbing, which is commonly seen during mating season. **Aggressive behavior is known to be associated with elevated levels of testosterone and other aggression-enabling hormones.**

matched; within a phenomenon, the four versions of the explanation never differed in length by more than 4 words. Additionally, there were no significant differences in average word count among the six sciences.

The 24 phenomena were divided into two pre-determined sets of 12 (two per science), and participants were randomly assigned to receive one of the two sets. Each set was further subdivided into two blocks of six phenomena (one per science); the order in which these two blocks were presented was randomly determined for each participant. Within each block, the six phenomena were presented in a random order. Each participant saw one good and one bad explanation from each science; two combinations of good and bad explanations were pseudorandomly determined ahead of time and participants were randomly assigned to one of the two different permutations. Participants were randomly assigned to either the horizontal or reductive condition, and all 12 explanations that they rated came from their assigned explanation level. This counterbalancing method led to 16 different randomly-assigned presentation orders in a 2 (Item Set: A or B) x 2 (Block Order) x 2 (Good/Bad combination) x 2 (Explanation Level: horizontal, reductive) design.

Procedure

Participants used a sliding scale ranging from -3 to 3 to indicate their ratings of each explanation. They were first given instructions on how to use the slider; this also served as a check that participants were reading instructions. They were told to use the slider to select 0 on the first page in order to proceed with the survey. If they selected anything other than 0, they were directed to another page asking them again to select 0. Participants who did not select the correct response on this second page (3 MTurk workers and 9 undergraduates) were excluded from analyses.

After these general instructions on using the slider, participants were given instructions for the explanations task (modified from Fernandez-Duque et al., 2015):

> You will now be presented with descriptions of various scientific findings. All the findings come from solid, replicable research; they are the kind of material you would encounter in a textbook. You will also read an explanation of each finding. Unlike the findings themselves, the explanations of the findings range in quality. Some explanations are better than others: They are more logically sound. Your job is to judge the quality of such explanations, which could range from very poor (-3) to very good (+3).

On each trial, participants were presented with a description of a scientific phenomenon, which was displayed for 10 seconds before participants could advance to the next screen. On the next screen, an explanation was displayed below the phenomenon, and participants were instructed to rate the quality of the explanation. Participants rated 12 explanations, with an attention check trial administered after the first six (Oppenheimer, Meyvis, & Davidenko, 2009). This trial was similar in format to the others. First, a description of a phenomenon was presented for 10 seconds. When participants advanced to the next screen, instead of seeing an explanation, they saw text

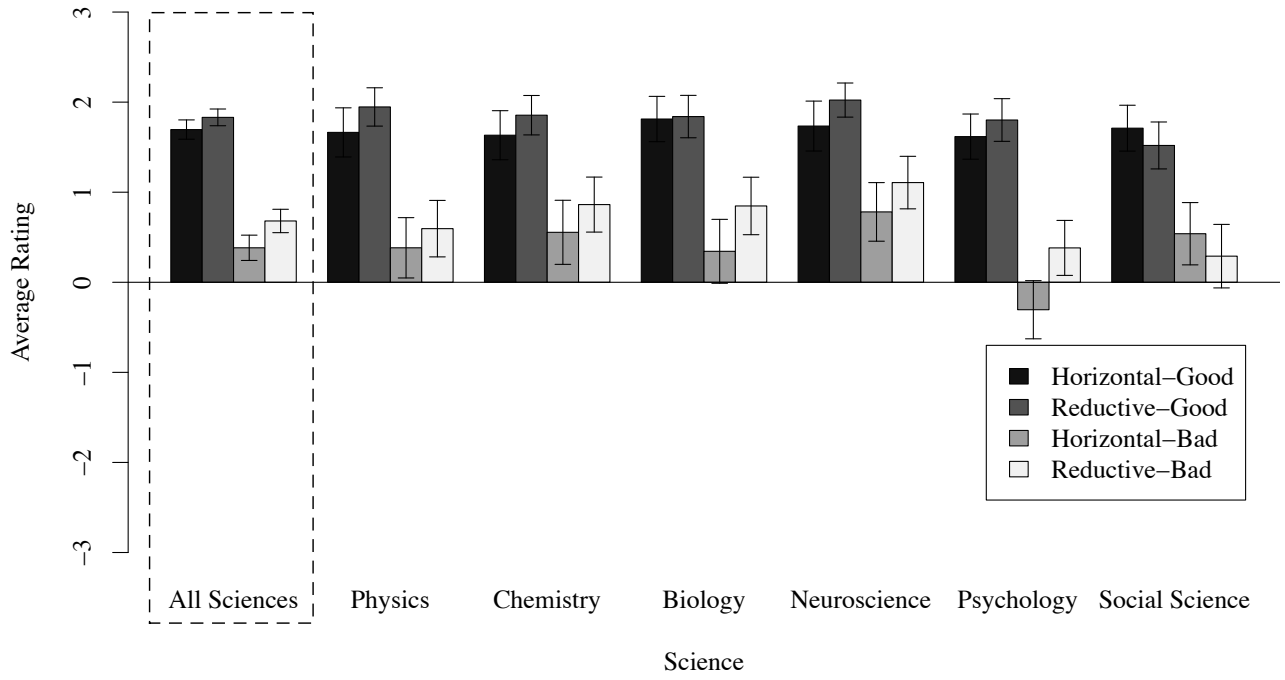


Figure 2: Average ratings of explanations by science, condition, and quality. Error bars are 95% CIs.

instructing them to select 3 on the scale. Participants who did not select 3 (17 MTurk workers and 31 undergraduates) were excluded from analyses.

Perceptions of Science. After the explanations task, participants responded to three questions designed to assess their views of 10 scientific disciplines: physics, chemistry, biology, neuroscience, psychology, sociology, economics, and political science (measure adapted from Fernandez-Duque et al., 2015). Participants rated the perceived scientific rigor of each discipline, the extent of the knowledge gap between a novice and an expert in each discipline, and the societal prestige of each discipline (presented in a random order). For each discipline, the ratings of the three items were made on a 10-point scale, which were summed to create a single score (out of 30).[1]

Results

Data from the explanations task (Figure 2) were analyzed using a mixed-effects linear regression model predicting the rating given on each trial from the sample (MTurk, undergraduates), explanation level (horizontal, reductive), explanation quality (good, bad), and science (physics, chemistry, biology, neuroscience, psychology, and social science). Sample and explanation level were between-participants variables; quality and science were within-participants variables. All possible interactions were tested, but the four-way interaction and most of the three-way interactions did not significantly improve model fit and were dropped. The best-fitting model included random intercepts by participant and item and a random effect of item on the slope for the quality variable. The science variable was backwards-difference coded to test five planned contrasts between pairs of adjacent sciences: chemistry vs. physics, biology vs. chemistry, neuroscience vs. biology, psychology vs. neuroscience, and social science vs. psychology (Figure 1). Significance levels were determined by generating bootstrapped confidence intervals around the regression coefficients.

Main Effects

There were significant effects of Sample (β = -0.25, 95% CI [-0.44, -0.07]), Quality (β = 1.27, 95% CI 1.07, 1.47]), and Explanation Level (β = 0.21, 95% CI [0.02, 0.42]). Undergraduate students (M = 1.01, SD = 1.85) gave significantly lower ratings on average than MTurk workers (M = 1.26, SD = 1.76). Good explanations (M = 1.76, SD = 1.43) were rated significantly higher than bad explanations (M = 0.53, SD = 1.93). Reductive explanations (M = 1.26, SD = 1.71) were rated significantly higher than horizontal explanations (M = 1.04, SD = 1.88). Finally, the contrast between psychology and neuroscience was significant (β = -0.55, 95% CI [-1.06, -0.06]): Neuroscience explanations (M = 1.41, SD = 1.67) were rated higher than psychology explanations (M = 0.88, SD = 1.85). No other contrasts between adjacent pairs of sciences were significant.

[1] Participants also completed measures of reflective thinking, logical reasoning, and general scientific literacy. For brevity, data from these measures will not be discussed here.

Interactions

A significant Sample x Quality interaction (β = 0.54, 95% CI [0.30, 0.76]) indicates that the difference in ratings between MTurk workers and undergraduates was driven primarily by their ratings of the bad explanations. Ratings of good explanations were similar between the two groups (M_M = 1.76 and M_U = 1.77), but the MTurk workers gave higher ratings to bad explanations than undergraduates did (M_M = 0.76 for and M_U = 0.24).

There was also an Explanation Level x Science interaction, wherein the contrast between social science and psychology was significant (β = -0.60, 95% CI [-0.97, -0.27]). This indicates that the effect of reductive information was significantly different between these two sciences. In psychology, as well as in physics, chemistry, biology, and neuroscience, the reductive explanations were rated higher on average than the horizontal explanations. However, the opposite was true for social science: Reductive explanations (M = 0.90, SD = 1.91) were rated *lower* than horizontal explanations (M = 1.13, SD = 1.84).

Finally, there was a significant, three-way Sample x Quality x Explanation Level interaction (β = 0.49, 95% CI [0.01, 0.94]), indicating that the magnitude of the Quality x Explanation Level interaction was larger for MTurk workers than for undergraduates. Separate analyses of the two groups revealed a significant Quality x Explanation Level interaction in the MTurk sample, but not in the undergraduate sample. Among MTurk workers, there was a larger difference between the horizontal and reductive conditions for bad explanations (M_H = 0.52 and M_R = 0.99) than for good explanations (M_H = 1.71 and M_R = 1.80).

Analyses by Science

To further investigate whether the seductive allure effect, which was previously observed with psychology explanations augmented by neuroscience information, occurred for other pairs of sciences, we conducted separate regression analyses for each science testing for main effects of Quality, Explanation Level, and a Quality x Explanation Level interaction. These models also included random intercepts for participant and item to account for repeated measures and differences between the individual stimuli used within each science.

There was a significant, positive effect of Quality in all six sciences, mirroring the strong effect of Quality observed in the earlier regression. The effect of explanation level was statistically significant for psychology (β = 0.43, 95% CI [0.12, 0.71]), and marginally significant (p < .10) for physics, chemistry, biology, and neuroscience. Also consistent with the prior analyses, the Explanation Level effect was in the opposite direction for social science compared to the other five sciences.

As has been observed in some previous studies on this effect (Fernandez-Duque et al., 2015; Weisberg et al., 2008), there was a marginally significant Quality x Explanation Level interaction for biology (β = -0.48, 90% CI [-1.00, -0.01]) and psychology (β = -0.50, 90% CI [-1.00, -0.07]). Post-hoc tests found significant differences between reductive-bad and horizontal-bad explanations in biology, $t(257) = 2.07$, $p < .05$, and psychology, $t(257) = 3.03$, $p < .01$, but there was no significant difference by explanation level for good explanations in either biology or psychology. For stimuli from the other sciences, the effect of explanation level was not moderated by the quality of the explanation.

Perceptions of Science

As described in the Method section, each science was rated on a 10-point scale for three different questions; the three ratings were summed to give a single score out of 30 for each science. The summed scores for sociology, economics, and political science were highly correlated (alpha = .79 for undergraduates and .82 for MTurk workers), and the three were averaged to create a "social science" score. By and large, these ratings mirror our predicted reductive scale of the sciences, with the more fundamental sciences being rated as more rigorous, difficult, and prestigious (Figure 3). The exception to this is neuroscience, which was rated higher than physics, chemistry, and biology. Paired t-tests were conducted on all adjacent pairs of fields; all comparisons were statistically significant (p < .001).

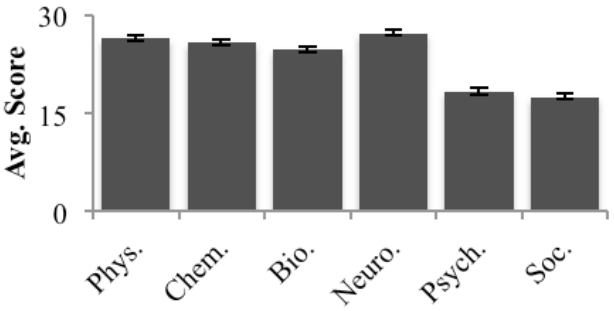


Figure 3: Avg. perception scores. Error bars are 95% CIs.

Discussion

The main goal of the current study was to investigate the generality of the seductive allure effect. Prior research has demonstrated that adding irrelevant neuroscience information to explanations of psychological phenomena makes these explanations seem better to naïve participants. We hypothesized that this effect is due to a general preference for reductive explanations, which should manifest across different scientific domains. Our data support this hypothesis: Participants judged explanations containing irrelevant reductive information as better across a range of sciences. The seductive allure effect is thus not unique to the pairing of psychology and neuroscience.

However, we did find that the preference for reductive information was strongest for the psychology/neuroscience pairing. In addition, the effect for the social science/psychology pairing was in the opposite direction,

with participants preferring the horizontal (non-reductive) explanations in this case. Taken together, these results point to a general disinclination for person-level explanations. This is in line with previous work finding generally poor public opinions about psychology as a science (Keil, Lockhart, & Schlegel, 2010; Lilienfield, 2012).

Regardless, participants were reliably able to discriminate good from bad explanations across all sciences, demonstrating an intact ability to sense explanation quality. However, even bad explanations tended to be rated positively on average. Undergraduate students were more critical of bad explanations than MTurk workers, perhaps because being in an academic environment encourages more skepticism. Interestingly, participants were less critical of bad neuroscience items than bad items from other sciences. Together with the high prestige ratings for neuroscience, this suggests that neuroscience information may exert some unique allure, even if this does not fully explain its appeal in explanations of psychological phenomena.

Future work should investigate this particular effect, as well as why reduction is so appealing as an explanatory form in the sciences. One interesting set of questions concerns the proper level for reduction: Are explanations seen as more appealing when they contain information only from the immediately adjacent science (e.g., chemistry for biology), or would further reduction make explanations seem even better (e.g., physics for biology)? Alternatively, do people prefer explanations that reference an additional field of science, regardless of whether that field is more fundamental? A current study is investigating these questions; participants were asked to select the methods that would be useful for investigating phenomena from various sciences. Preliminary results show that participants most often selected methods from the field of the phenomenon (46% of the time) or the immediately reductive field (37%), suggesting that they believe there is a particular level of reduction for each science that is maximally explanatory.

Finally, future work should examine the potential role of training in ameliorating the seductive allure effect. Previous work (Weisberg et al., 2008, Study 3) found that neuroscience experts were not seduced by irrelevant neuroscience. Ongoing work in our lab expands this investigation to experts in all six of our target sciences to determine the role of expertise: Does training in a particular science protect against the seductive allure effect for that science, or in general? Answering this question can provide further insight into the nature of the effect itself and into people's judgments of scientific explanations in general.

Acknowledgments

This research was funded by the John Templeton Foundation through the *Varieties of Understanding* project. We'd like to thank Don Berry, Martha Farah, Simon Garcia, Cindy Keeler, Phil Nelson, Sam Reznik, and Paul Sneigowski for their assistance with stimulus creation.

References

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Clarendon Press.

Farah, M. J., & Hook, C. J. (2013). The seductive allure of "seductive allure." *Perspectives on Psychological Science*, *8*(1), 88–90.

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience*, *27*(5), 926–944.

Keil, F. C., Lockhart, K. L., & Schlegel, E. (2010). A bump on a bump? Emerging intuitions concerning the relative difficulty of the sciences. *Journal of Experimental Psychology: General*, *139*(1), 1–15.

Kelemen, D. (1999). Why are rocks pointy? Children's preference for teleological explanations of the natural world. *Developmental Psychology*, *35*(6), 1440–1452.

Kikas, E. (2003). University students' conceptions of different physical phenomena. *Journal of Adult Development*, *10*(3), 139–150.

Langer, E. J., Blank, A., & Chanowitz, B. (1978). The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology*, *36*(6), 635–642.

Lilienfeld, S. O. (2012). Public skepticism of psychology: Why many people perceive the study of human behavior as unscientific. *American Psychologist*, *67*(2), 111–129.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

McCabe, D. P., & Castel, A. D. (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition 107*(1), 343-352.

Nagel, E. (1961). *The Structure of Science. Problems in the Logic of Explanation*. New York: Harcourt, Brace & World, Inc.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.

Rhodes, R. E., Rodriguez, F., & Shah, P. (2014). Explaining the alluring influence of neuroscience information on scientific reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1432–1440.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*, 212–233.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–477.

Weisberg, D. S., Taylor, J. C. V., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, *10*(5), 429–441.

More than Words: The Many Ways Extended Discourse Facilitates Word Learning

Sumarga H. Suanda, Linda B. Smith, Chen Yu
{ssuanda, smith4, chenyu}@indiana.edu
Department of Psychological and Brain Sciences, Indiana University
Bloomington, IN, 47405 USA

Abstract

Child-directed speech is often temporally organized such that successive utterances refer to the same topic. This type of *extended discourse* on the same referent has been shown to possess several verbal signatures that could facilitate learning. Here, we reveal multiple non-verbal correlates to extended discourse that could also aid learning. Multimodal analyses of extended discourse episodes reveal that during these episodes, toddlers and parents exhibit greater sustained attention on objects, and greater coordination between their behaviors. The results indicate the interconnections between multiple aspects of the language-learning environment, and suggest that parents' speech may both shape and be shaped by non-verbal processes. Implications for understanding how the learning environment influences development are discussed.

Keywords: language acquisition; word learning; discourse development; child-directed speech; joint attention.

Introduction

Children acquire language in an environment rich with structure and regularities. One of its noticeable structures is its temporal structure, with adjacent utterances frequently referring to the same conversational topic:

Mother: *oh there's a super car?*
Mother: *you like cars don't you?*
Mother: *what are you going to do with it?*
Mother: *are you going to make it go?*
(Messer, 1980)

These extended episodes of verbal discourse on the same referent – what we will call *extended discourse* for short - could facilitate learning in multiple ways. Repeated utterances to one topic give children multiple opportunities to identify the focus of parents' speech, and allow children to deploy comprehension of one utterance in the service of comprehending subsequent utterances (Frank, Tenenbaum, & Fernald, 2013; Messer, 1980; Sullivan & Barner, 2016). Additionally, the repetition of utterance properties (i.e., words, sentence structures) common in extended discourse, has been proposed to aid speech perception (Bard & Anderson, 1983), word segmentation (Onnis, Waterfall, & Edelman, 2008), and syntax learning (Hoff-Ginsberg, 1986).

Previous research thus provides evidence for the idea that the *linguistic* features of extended discourse facilitates learning. Here, we take a different perspective on why extended discourse aids learning. Our perspective is based on the idea - and data - that the language-learning environment is inherently multi-modal and multi-dimensional (Roy, Frank, DeCamp, Miller, & Roy, 2015), and that the verbal discourse children hear is intricately tied to its nonverbal perceptual and social contexts (Adamson & Bakeman, 2006). If this is true, then extended episodes of verbal discourse likely co-occur with, and may even be driven by, extended episodes of sustained attention on the part of the child and extended episodes of joint engagement on the part of the child-parent dyad. Since both of these processes have been linked to healthy language development (e.g., Adamson, Bakeman, & Deckner, 2004; Salley, Panneton, & Colombo, 2013), we suggest then that extended *verbal* discourse may facilitate language learning in part through its underlying *nonverbal* components. In the current study, we test the hypothesis that extended discourse possesses important nonverbal features, including: (1) enhanced child and parent attention to the talked-about object, which means parents' speech is more *referentially transparent* (Cartmill et al., 2013), and (2) a greater degree of joint engagement between children and their parents (Adamson et al., 2004; Tomasello & Farrar, 1986).

The role of referential transparency and joint engagement on word learning is well established. For example, Cartmill and colleagues found that children who heard more referentially transparent speech as toddlers had larger vocabularies as preschoolers (Cartmill et al., 2013). Similarly, in a series of recent studies using mini head-cameras worn by toddlers as they played with novel objects with their parents, Smith, Yu, and colleagues found that the referential transparency of parents' object naming (measured by the visual dominancy of the named object over competitor objects) was predictive of toddlers' object name learning (Pereira, Smith, & Yu, 2014; Yu & Smith, 2012). With respect to joint engagement, Adamson and colleagues (2004) found that the time parents and toddlers spent in joint visual attention was linked to toddler vocabulary development (see also Adamson et al., 2004). Other non-visual forms of joint engagement, such as fluid turn-taking of head and hand action in play, has also been linked to word learning (Pereira, Smith, & Yu, 2008).

In the current study, parents and toddlers were observed as they played with, and as parents talked about, a set of objects (see Figure 1). From these free-flowing object-play interactions, we identified moments of extended verbal discourse, as well as moments of short verbal discourse. We measured, via head-mounted eye tracking, moment-by-moment gaze patterns of parents and their toddlers. We also

measured parents' and toddlers' manual actions. Of interest was whether parents' speech inside extended discourse (compared to short discourse) would be marked by more sustained and greater referential transparency (indexed by gaze and action patterns on the referent object), as well as greater joint engagement (indexed by the coupling of gaze and manual actions between toddlers and their parents).

Methods

Participants

Fifty-two parent-toddler dyads participated (*Mean* toddler age = 17.9mos, *SD* = 4.3mos). Twenty-three toddlers were girls. Data of 17 dyads, were part of a previous report on joint attention (Yu & Smith, 2013).

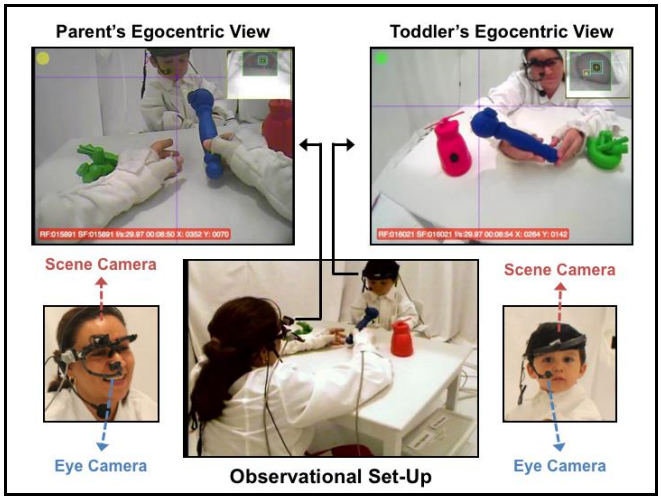


Figure 1. *The observational set up: toddlers and parents played on a table-top in the lab and were equipped with head-mounted eye trackers (bottom row), which produced egocentric views and estimates of gaze direction (top row).*

Apparatus

Figure 1 depicts the set-up. Toddlers sat in a chair across from their parents who sat on floor cushions. Toddlers and parents wore light-weight head mounted eye trackers from Positive Science[1]. As seen in Figure 1, these eye trackers consisted of one outward-facing camera that records the observer's first-person, egocentric views and one inward-facing camera that records the observer's eye movements (for more details, see Franchak, Kretch, Soska, & Adolph, 2011). Both cameras recorded at a temporal resolution of 30 Hz and a spatial resolution of 720x480 pixels. Parents wore a headset equipped with a microphone.

Stimuli included two sets of three novel objects. All objects had a single main color, were similar in size, and were small enough for toddlers to handle. Each object was paired with a novel disyllabic word (e.g., "*habble*", "*tema*").

[1] For 23 dyads, parents were equipped with the Wearcam eye-tracking system. All the current results did not vary as a function of eye-tracking system.

Procedure

After toddlers and parents were fitted with the recording equipment, we placed a set of three objects on the table and instructed parents to play with their children as they normally would, leading to a free-flowing interaction with no constraints on how parents or their children should play, or on what parents should say, with one exception. Prior to play, we told parents that when talking about the objects, to use the names we provided.

The play session consisted of a series of brief trials, each lasting between 1-2 minutes long. On each trial, dyads played with one of two object sets. Object sets were swapped between trials to keep toddlers engaged. Depending on toddlers' compliance, the play session lasted between 2 and 4 trials (*M* = 3.12; *SD* = 1.00); total play duration lasted on average about 5 minutes (*M* = 4 min 57s; *SD* = 89s).

Coding: Parent Speech

Parents' speech during play was fully transcribed. The unit of transcription was the utterance, defined as a string of speech between two periods of silence lasting at least 400ms. Utterances containing reference to one of the objects were marked as *referential utterances*, which included utterances when parents named an object (e.g., "that's a *habble*"), employed a pronoun referring to an object (e.g., "can you push *it*?"), or used an alternate concrete noun referring to an object (e.g., "don't throw *the toy*"). For each referential utterance, we coded the referent by watching the video. On average, there were 95.2 utterances per dyad (*SD* = 36.3), 49.5 of which were referential (*SD* = 23.4).

We then classified each referential utterance as either part of an *extended discourse* or a *short discourse* (Figure 2). To be counted as part of an extended discourse, utterances had to satisfy two criteria. First, utterances had to be part of at least three consecutive utterances referring to the same object. Second, adjacent utterances had to occur within ten seconds of each other. On average, 28.9 utterances per dyad were classified as part of an extended discourse (*SD* = 19.4); 15.7 were classified as part of a short discourse (*SD* = 7.2). An additional 4.7 utterances (*SD* = 4.4) referred to multiple objects and were excluded from analysis.

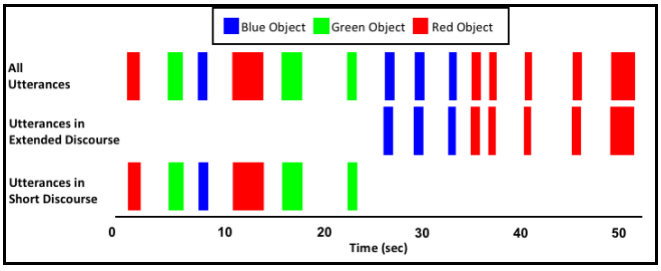


Figure 2. *Representative time series of parents' utterances (top row), including utterances that were part of extended discourse (middle row; see text for criteria) and utterances that were part of short discourse (bottom row).*

Coding: Parent and Child Sensorimotor Behaviors

Gaze Coding. At the end of play, we calibrated both toddlers' and parents' eye trackers by having them fixate to known locations on the table (see Yu & Smith, 2013 for a detailed explanation of calibration). Based on the calibration data, eye-tracking software produced frame-by-frame point-of-gaze estimates, as indicated by the cross-hairs in Figure 1. Eye tracking spatial accuracy is about 3^0 once calibrated (see Franchak et al., 2011).

Using footage from the first-person scene camera (with cross-hair superimposed) and the eye camera (which depicts moment-by-moment eye movements), coders manually annotated the whole session frame-by-frame for parents' and toddlers' target of gaze. For each frame, gaze was coded as one of five possibilities: each of the three toy objects, the partner's face, or elsewhere.

Manual Activity Coding. Coders also watched the session from multiple angles (first-person scene cameras, third-person view cameras) and annotated the session frame-by-frame for moments when toddlers and parents touched each of the three objects. Figure 3 depicts a representative time series of gaze and holding behavior of toddlers and parents over the course of a trial.

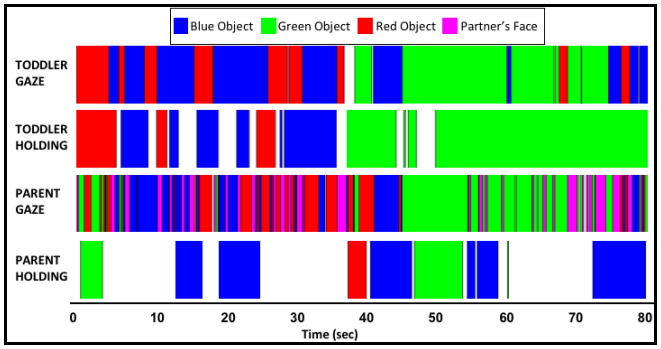


Figure 3. *Representative time series of the sensorimotor behaviors of one toddler (top two rows) and one parent (bottom two rows).*

Results

Referential Transparency in Extended vs. Short Discourse

Is speech in extended discourse more referentially transparent than speech in short discourse? We considered parents' speech to be referentially transparent to the extent that the talked-about (target) objects, and not the distractor objects, were looked at and held by toddlers and parents (see also Frank et al., 2013). Figure 4 depicts the temporal profiles of gaze and holding around utterances that were either part of extended discourse or part of short discourse.

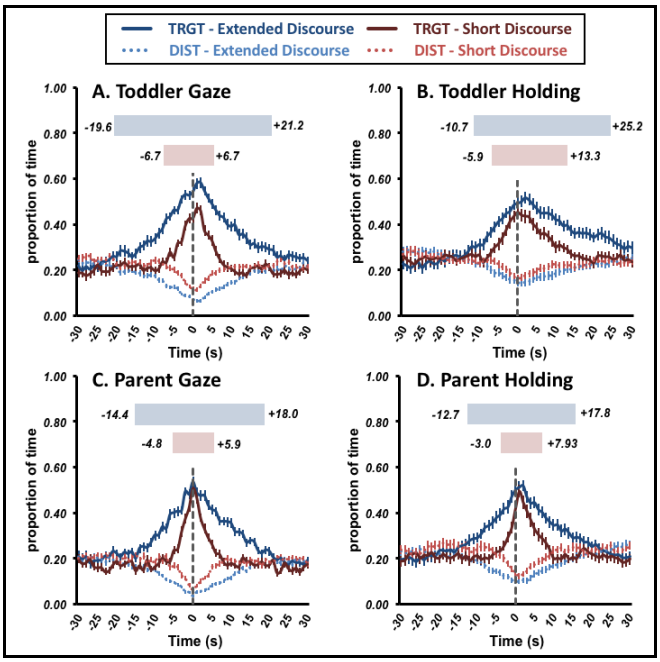


Figure 4. *Temporal profiles of visual and manual attention to target and distractor objects around utterances of different discourse lengths. Vertical dotted lines mark the onset of parents' utterances. Horizontal bands (and adjacent numbers) illustrate the point at which looking (or holding) to target vs. distractors deviates statistically (blue band: extended discourse; red band: short discourse).*

As the figure shows, both utterance types exhibited some referential transparency since both were characterized by more looking and holding to the target than the distractor objects. However, two noticeable patterns distinguished utterances in extended discourse from utterances in short discourse. First, the period of time in which the referent of parents' speech was transparent was longer for utterances in extended discourse than for utterances in short discourse. For utterances in extended discourse, we determined via frame-by-frame *t*-tests that the point in time in which toddlers' gaze at the target object first diverged from the distractor objects was 19.6s prior to utterance onset. Looking time continued to be different until 21.2s after utterance onset. For utterances in short discourse, the period was much shorter: 6.7s both before and after utterance onset. This pattern of more enduring referential transparency in extended discourse was consistent for all toddler and parent measures (see Figure 4).

Second, the figure also shows that even around the moment of utterance, attention to target and distractor objects was more divergent in utterances of extended discourse than in utterances of short discourse. When our analyses honed in on the moments of parents' utterances, which regardless of discourse length are moments when parents are talking about the target object, gaze and holding of the target object was greater in utterances that were part of extended discourse than in utterances that were part of

short discourse (see Table 1). This finding suggests a correlation between the *degree* of toddlers' and parents' nonverbal attention to an object and the duration of verbal discourse on that object. We return to the implications of this correlation in the General Discussion.

Table 1: *Mean gaze and holding of the talked-about object and not talked about object during referential utterances.*

	Extended		Short		Sig.
	TRGT	DIST	TRGT	DIST	
Child Gaze	.56 (.14)	.07 (.05)	.44 (.14)	.12 (.07)	***
Child Holding	.50 (.18)	.15 (.12)	.45 (.19)	.17 (.11)	**
Parent Gaze	.52 (.16)	.04 (.04)	.49 (.16)	.07 (.05)	***
Parent Holding	.52 (.18)	.09 (.08)	.48 (.19)	.13 (.09)	**

Note: **TRGT**: Target object being referred to; **DIST**: Distractor objects not being talked about; **Sig:** significance of paired-samples *t*-tests on the proportion of all object looking or holding that was directed to the target object for utterances in extended vs. short discourse; ***p <= .001, **p <= .01.

Joint Engagement in Extended vs. Short Discourse

To examine the social correlates underlying extended verbal discourse, we asked whether toddlers' and parents' act in a more coupled manner during extended discourse. We considered toddlers and parents to be coupled if their looking and/or holding were synchronously directed towards the same object. As Figure 5 indicates, there are multiple ways dyads could be coupled. Although these forms of coupling are not mutually exclusive (e.g., toddlers might simultaneously be looking at what their parents are holding *and* what their parents are looking at), we chose to analyze them individually because the relevance of each form of coupling has been suggested in the literature (see Yu & Smith, 2013).

For each dyad, we measured frame-by-frame whether or not the dyads were coupled. We then examined the proportion of time that dyads were coupled within each utterance. We considered both the time that dyads were coupled on the talked-about object, or *target coupling*, and the time that dyads were coupled on any object, or *total coupling*[2]. Figure 6 compares the degree of coupling within utterances of extended discourse to coupling within utterances of short discourse. As the figure illustrates, across all forms of dyadic coupling (and across both measurements of coupling), we observed greater coupling in extended discourse than in short discourse. These findings demonstrate that extended discourse co-occurs with social interactions that are richer in joint attention and joint action, illustrating yet another nonverbal reason why participating in extended discourse may benefit children's development.

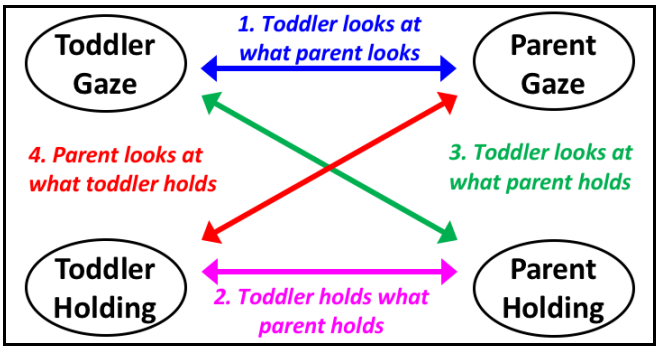


Figure 5. *Four non-mutually exclusive ways toddler and parent behaviors could be coupled.*

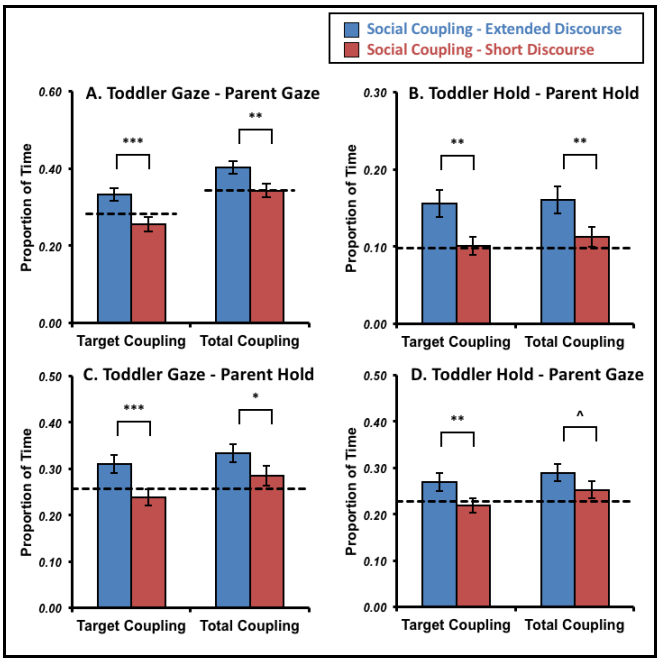


Figure 6. *The degree of social coupling (both to the target object and to all objects) within utterances across discourse types. Horizontal dotted lines reflect baseline coupling across the entire interaction.*

To what extent might the observed greater coupling be due simply to the fact that in extended discourse, toddlers and parents may have been more attentive and active in general (e.g., looking at and manipulating objects more frequently)? If this were the case, greater degrees of coupling in extended discourse than in short discourse would be expected by chance alone. To address this concern and provide a more rigorous test of coupling, we adopted a signal detection analysis that controls for base rate levels of toddlers' and parents' behaviors. Briefly, we classified each

[2] In computing the total coupling between toddler and parent gaze, we considered time spent looking at each other, or *mutual gaze*, as coupled.

frame into *hits* (e.g., toddlers look at an object that parents look at), *false alarms* (toddlers look at an object that parents do not look at), or *misses* (toddlers do not look at an object that parents look at). The *F-score* from this classification scheme served as our primary measure of coupling. The *F-score* is a combination (the harmonic mean) of *Precision* (hits / hits + false alarms), which in our case controls for toddler behavior, and *Recall* (hits / hits + misses), which in our case controls for parent behavior. We measured coupling for each dyad in three contexts: (1) in extended discourse utterances, (2) in short discourse utterances, and (3) across the entire interaction, which reflects baseline degrees of coupling. As Table 2 indicates, toddlers and parents were more coupled during episodes of extended discourse than during episodes of short discourse. Additionally, the degree of coupling in extended discourse was greater than baseline levels of coupling. Thus, this more stringent analysis confirms that episodes of extended verbal discourse are also episodes with rich joint engagement.

General Discussion

In the current study, we show that during the course of parent-toddler object play, multiple aspects of toddlers' language environment converge in real-time, generating segments of interaction that are potential gold mines for word learning. These segments are extended episodes of verbal discourse. Although previous research (e.g., Frank et al., 2013; Messer, 1980) has uncovered linguistic features that make these episodes valuable for learners, we revealed attentional and social correlates to these episodes that would also benefit learning. We found that within extended discourse, toddlers' and parents' attention were focused on the referent object for an extended period of time, making the topic of parents' speech especially clear and transparent. Previous observational (Yu & Smith, 2012), experimental (Hollich et al., 2000; Tomasello & Farrar, 1986), and individual-difference (Cartmill et al., 2013) research provide converging evidence that transparency in parents' speech is important for toddlers' word learning. Thus, extended discourse about an object may facilitate learning through an increase in the referential transparency of parents' speech.

We also observed that within extended discourse, toddlers and their parents displayed great degrees of social coordination. The importance of social coordination, or joint attention, for multiple facets of development is widely accepted and well established (e.g., Moore & Dunham, 1995). Although research in this area typically focuses on the role of joint *visual* attention, recent research suggests the relevance of other forms of joint attention as well. For example, Yu and Smith (2013) found that in the complex contexts of parent-toddler interactions, attending to partner manual activity may be the key way by which they share attention. In a similar vein, Deak and colleagues (2014) argue that following parents' manual actions may be a stepping stone for learning to follow parents' gaze. The fact that in extended discourse we observed heightened rates of many forms of joint attention suggests then that extended discourse may facilitate toddlers' development in part by offering redundant pathways to attention sharing, and by providing fertile grounds for training socio-cognitive development.

The current data do not only show a correlation between the duration of verbal discourse and the *duration* of attention and social coupling. The data also demonstrate a correlation between the duration of verbal discourse and the *intensity* of attention and social coupling. When our analyses honed in on individual utterances, utterances that were part of extended discourse were characterized by *more* toddler attention, *more* parent attention, and *more* social coupling. Considering the correlational nature of these findings, we cannot speak to their precise causal underpinnings. It is possible for example that more focused and sustained toddler attention to an object leads to parents' extended verbal discourse about that object. Alternatively, extended talk may actually play a role in focusing and sustaining toddlers' attention on the talked-about object (see Baldwin & Markman, 1989). And of course it is also possible that the influence is multi-directional; toddlers' attention, toddler-parent coupling and extended verbal discourse may bootstrap each other, producing the temporal

Table 2. *Mean coupling scores (F-Scores) across discourse length and at baseline.*

	Target Coupling			*Total Coupling*		
Form of Social Coupling	Extended	Short	Baseline	Extended	Short	Baseline
Child Gaze – Parent Gaze	.59 (.13)	.52** (.16)	.46*** (.09)	.47 (.12)	.42** (.13)	.42* (.08)
Child Hold – Parent Hold	.30 (.15)	.21*** (.13)	.14*** (.07)	.22 (.15)	.15** (.10)	.14*** (.07)
Child Gaze – Parent Hold	.56 (.13)	.50* (.16)	.37*** (.11)	.47 (.13)	.40** (.15)	.37*** (.11)
Child Hold – Parent Gaze	.50 (.15)	.45* (.14)	.35*** (.10)	.42 (.15)	.36** (.12)	.35* (.10)

Note: Asterisks reflect significance tests between coupling in extended discourse utterances and short discourse utterances, and between extended discourse utterances and baseline; * $p < .05$, ** $p < .01$, *** $p < .001$

dynamics and inter-relations that we observed. Either way, the results highlight a tight link between the verbal and non-verbal aspects of toddlers' language learning experience.

Conclusion

One view of early word learning is that the complexity of toddlers' learning environment makes it an especially challenging task. From a sea of information to which they are exposed - numerous words, several objects, and various social signals - they must figure out how the words they hear relate to the world they see. To make matters worse, toddlers are learning words while they are still figuring out the regularities of their perceptual world and the quirks of their social world. A different view of word learning is that the sea of information provides helpful information and that the multi-tasking makes word learning easier not harder. The current data provide evidence in line with this latter perspective, revealing a learning environment rich with redundancies and correlations between its linguistic, perceptual and social dimensions. Although much more work is needed, it may be that it is through extracting the latent structures from these interconnected dimensions and through solving multiple, mutually-constraining tasks that toddlers come to learn words as effortlessly as they do.

Acknowledgments

We thank many members of the Computational Cognition and Learning Laboratory, especially Maddie Bruce, Danielle Rosenstein and Jessica Steinhiser, for their assistance in this research. This research was supported in part by the National Science Foundation (BCS 092428) and the National Institutes of Health (R01-HD074601, R21-EY017843, K99-HD082358).

References

Adamson, L. B., & Bakeman, R. (2006). Development of displaced speech in early mother-child conversations. *Child Development, 77*, 186-200.

Adamson, L. B., Bakeman, R., & Deckner, D. F. (2004). The development of symbol-infused joint engagement. *Child Development, 75*, 1171-1187.

Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development, 60*, 381-398.

Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language, 10*, 265-292.

Cartmill, E. A., Armstrong III, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences of the United States of America, 110*, 11278-11283.

Deak, G. O., Krasno, A. M., Triesch, J., Lewis, J., & Sepeta, L. (2014). Watch the hands: infants can learn to follow gaze by seeing adults manipulate objects. *Developmental Science, 17*, 270-281.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child Development, 82*, 1738-1750.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning & Development, 9*, 1-24.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology, 22*, 155-163.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H., … Rocroi, C. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development,* 65(3, Serial No. 262).

Messer, D. J. (1980). The episodic structure of maternal speech to young children. *Journal of Child Language, 7*, 29-40.

Moore, C., & Dunham, P. J. (1995). *Joint Attention: Its Origins and Role in Development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition, 109*, 423-430.

Pereira, A. F., Smith, L. B., & Yu, C. (2008). Social coordination in toddler's word learning: interacting systems of perception and action. *Connection Science, 20*, 73-89.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic Bulletin & Review, 21*, 178-185.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences of the United States of America, 112*, 12663-12668.

Salley, B., Panneton, R. K., & Colombo, J. (2013). Separable attentional predictors of language outcome. *Infancy, 18*, 462-489.

Sullivan, J., & Barner, D. (2016). Discourse bootstrapping: preschoolers use linguistic discourse to learn new words. *Developmental Science, 19*, 63-75.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development, 57*, 1454-1463.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition, 125*, 244-262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE, 8*, e79659.

Two Potential Mechanisms Underlying the Link between Approximate Number Representations and Symbolic Math in Preschool Children

Ariel Starr (arielstarr@berkeley.edu)
Helen Wills Neuroscience Institute, 134 Barker Hall
Berkeley, CA 94720 USA

Rachel C. Roberts (rachel.c.roberts@me.com) & Elizabeth M. Brannon (ebrannon@sas.upenn.edu)
Department of Psychology, 3401 Walnut Street
Philadelphia, PA 19103 USA

Abstract

The approximate number system (ANS) is frequently considered to be a foundation for the acquisition of uniquely human symbolic numerical capabilities. However, the mechanism by which the ANS influences symbolic number representations and mathematical thought remains poorly understood. Here, we tested the relation between ANS acuity, cardinal number knowledge, approximate arithmetic, and symbolic math achievement in a one-year longitudinal investigation of preschoolers' early math abilities. Our results suggest that cardinal number knowledge is an intermediary factor in the relation between ANS acuity and symbolic math achievement. Furthermore, approximate arithmetic performance contributes unique variance to math achievement that is not accounted for by ANS acuity. These findings suggest that there are multiple routes by which the ANS influences math achievement. Therefore, interventions targeting both the precision and manipulability of the ANS may prove to be more beneficial for improving mathematical reasoning compared to interventions targeting only one of these factors.

Keywords: approximate number representations; numerical cognition; math cognition

Introduction

In our information-driven society, math ability is essential for success, particularly in the STEM fields that drive the modern economy. However, within the population there exists large variance in math ability, and low math proficiency is associated with poor health and occupational outcomes (e.g., Parsons & Bynner, 2006). Critically, math ability when a child first enters schooling is the strongest predictor of later math and overall academic achievement (e.g., Duncan, Dowsett, & Claessens, 2007). Many cognitive and socioeconomic factors are known to contribute to individual differences in math achievement (e.g., Bull & Scerif, 2001; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Klibanoff, Levine, Huttenlocher, Vasilyeva, & Hedges, 2006). One of these factors is an evolutionarily ancient system for representing approximate quantities. Although educated humans typically think about number using language, we also possess a system for representing number in an approximate, nonsymbolic fashion. This system, termed the approximate number system (ANS), is not dependent on language or formal schooling and is present in a wide variety of nonhuman species (e.g., Dehaene, 1997; Gallistel & Gelman, 1992; Hubbard et al., 2008).

The ANS is frequently hypothesized to be a cognitive foundation for symbolic math abilities. Lending support to this view is the pervasive finding that the acuity of the ANS, typically measured by an individual's ability to compare two arrays of dots, correlates with symbolic math achievement throughout the lifespan (see Chen & Li, 2013 for review, including failures to find this relation). Importantly, ANS acuity prior to the beginning of formal math instruction is predictive of later math achievement (Libertus, Feigenson, & Halberda, 2013; Mazzocco, Feigenson, & Halberda, 2011; Starr, Libertus, & Brannon, 2013; vanMarle, Chu, Li, & Geary, 2014). These studies suggest that the precision of approximate number representations may contribute to children's acquisition of symbolic math principles and influence symbolic math performance throughout the lifespan.

Despite the many studies documenting a link between ANS acuity and symbolic math, the mechanism underlying this relation remains unclear. One possibility is that the precision of the ANS influences children's acquisition of symbolic number representations. Before children can begin learning symbolic arithmetic and other mathematical operations, they must first learn the meaning of number words and Arabic numerals. While there is debate over the nature of the nonverbal representations that first ground the meaning of number words (e.g., Carey, 2004; Gallistel & Gelman, 1992), it is clear that sometime in early childhood, children map number words onto approximate number representations (Siegler & Opfer, 2003). Perhaps children with more precise internal representations of numerical quantities are at an advantage for forming these mappings, which enables them to master the count list and counting principles earlier than their peers with less precise internal quantity representations (Starr et al., 2013; vanMarle et al., 2014). In support of this view, preschoolers' competence with numerical symbols, and particularly their understanding of cardinality, has been shown to mediate the relation between their ANS acuity and math achievement (vanMarle et al., 2014). Thus, the precision of the ANS may influence the rate and fidelity with which children learn the meaning of numerical symbols. Under this scenario, the ANS may influence math achievement indirectly through its connection to symbolic number representations – the ANS

may be influential for grounding the meaning of numerical symbols, but may not directly contribute to symbolic mathematics (see also Lyons & Beilock, 2011).

A second hypothesis is that the manipulability of ANS representations serves as an intuitive basis for symbolic arithmetic (Barth, La Mont, Lipton, & Spelke, 2005). The ANS enables human infants (McCrink & Wynn, 2004), preschoolers (Barth et al., 2005; Gilmore, McCarthy, & Spelke, 2010), and monkeys (Cantlon & Brannon, 2007) to perform approximate arithmetic operations without the use of symbols. Therefore, the ANS supports both quantity representation and quantity manipulation. Children's approximate arithmetic performance at the beginning of kindergarten is predictive of their symbolic math achievement at the end of the academic year (Gilmore et al., 2010). In addition, approximate arithmetic performance has been found to mediate the relation between ANS acuity and symbolic math achievement in grade-school children (Pinheiro-Chagas et al., 2014). Furthermore, training nonsymbolic arithmetic in adults leads to improvements in their symbolic arithmetic performance (Park & Brannon, 2013; 2014). Thus, it may not be the precision of ANS representations but instead the manipulability of ANS representations that influences symbolic math achievement. As a result, children who are more adept at manipulating approximate quantities in arithmetic operations may also be more adept at symbolic arithmetic because of the overlap in cognitive processes required by both forms of arithmetic.

Critically, the above hypotheses need not be mutually exclusive – it is possible that the ANS contributes to symbolic math through multiple routes, including other possibilities not listed above. For example, the ANS may also influence symbolic math abilities through its representation of numerical order (Lyons & Beilock, 2011), or it may serve as an on-line error detection system, enabling erroneous answers to be rejected in favor of more plausible solutions (Lourenco, Bonny, Fernandez, & Rao, 2012). In addition, the relation between the ANS and symbolic math is not necessarily static, and it is possible that the ANS may contribute to symbolic math abilities in different ways throughout development depending on the child's current level of math knowledge.

In the present research, we tested two possible pathways through which the ANS may influence math abilities in a longitudinal investigation of preschoolers' math achievement. Children were tested at 3.5 years of age with a nonsymbolic numerical comparison task, cardinal number knowledge task, symbolic math test, and general IQ test. One year later, when the same children were approximately 4.5 years of age, we tested them with a nonsymbolic numerical comparison task, nonsymbolic approximate arithmetic task, symbolic math test, and general IQ test. Here we report how each of these measures contributes to math achievement at 4.5 years of age, with a specific focus on whether both the precision and manipulability of the ANS influence math ability via number word knowledge and approximate arithmetic.

Methods

Participants

Data from 97 children were included in the 3.5-year analyses (mean age: 3.61 years, range: 3.50-3.89 years, 49 female). Data from 161 children were included in the 4.5-year analyses (mean age: 4.56 years, range: 4.48-4.85 years, 86 female). Data from an additional 56 children at the 3.5-year visit and 10 children at the 4.5-year visit were excluded due to missing data points. Note that although the same children were tested at 3.5 and 4.5 years of age, a greater proportion of children had missing data points at 3.5 years, resulting in a smaller sample size for analyses that involved both the 3.5- and 4.5-year visit compared to analyses that involved only the 4.5-year visit. The most common cause of missing data points was an inability of our model to settle on a value for w, and this affected a greater number of 3.5-year-olds than 4.5-year-olds.

Procedure

Children came into the lab for two visits each lasting less than one hour at 3.5 and 4.5 years of age. At the first 3.5-year visit, children completed a symbolic math assessment and one session of the nonsymbolic number comparison task. During the second visit, 3.5-year-olds completed an IQ assessment, the counting knowledge task, and a second session of the nonsymbolic number comparison task. The order of the tasks within each session was counterbalanced across participants. At 4.5 years, during the first visit children completed the symbolic math assessment, and one session of the nonsymbolic number comparison task. During the second visit, 4.5-year-olds completed the IQ assessment, a second session of the nonsymbolic number comparison task, and the nonsymbolic approximate arithmetic task. The counting knowledge task was only administered at 3.5 years because the majority of 4.5-year-olds perform at ceiling on this task. The approximate arithmetic task was administered only at 4.5 years because 3.5-year-olds were unable to successfully perform the task. At each visit, parents gave written consent to a protocol approved by the local Institutional Review Board and were compensated monetarily and with a small gift for the child.

Nonsymbolic Numerical Comparison Task On each trial, a touchscreen computer displayed two bounded boxes (8 x 9.5 cm) containing arrays of dots. Children were instructed to touch the box that contained more dots and to make this choice without counting. Arrays contained between 4 and 14 dots, and the numerical ratio between the arrays was 1:2, 2:3, 3:4, or 6:7. To control for non-numerical perceptual cues, the parameters of the arrays varied such that the smaller and larger numerical array each had the larger cumulative surface area on 50% of trials. All of the dots within a single array were homogenous in element size and color, and the color of each array varied randomly from trial to trial. Differential audio-visual feedback was provided

after each trial, and children received a small sticker for each correct response to keep them engaged. Children performed practice trials until they made three consecutive correct responses or completed a maximum of ten trials. Children performed 60 test trials in each test session for a total of 120 trials each time point.

To estimate each child's ANS acuity, we used a psychophysical modeling technique (e.g., Halberda & Feigenson, 2008; Piazza et al., 2010) to calculate a Weber fraction (*w*) based on performance in the nonsymbolic numerical comparison task. The resulting value of *w* represents the noise in each participant's internal ANS representations, such that lower values of *w* correspond to less noise (i.e. higher ANS acuity).

Nonsymbolic Approximate Addition Task On each trial, children viewed an animation that consisted of an array of dots moving behind an occluder box, followed by a second array moving behind the same occluder. This animated arithmetic sequence lasted a total of 2000 ms. Children then saw two boxes containing arrays of dots and were instructed to touch the array that contained the same number of dots as had moved behind the occluder box. Answer choices remained on the screen until a decision was made. Correct and incorrect answer choices differed by a 1:2 or 1:4 ratio, and all choice arrays contained 2, 4, or 8 dots. Individual dot size varied across arrays but was homogenous within each array. Differential audiovisual feedback was provided after each trial, and children were rewarded with a small sticker for correct responses. Children performed practice trials until they made three consecutive correct responses or completed a maximum of ten trials. Children then completed a total of 42 test trials.

Counting Knowledge Task This task was modeled after the Give-a-Number task (Wynn, 1992). The experimenter introduced the child to a dinosaur puppet and asked the child to give the dinosaur a certain number of fish by placing them on a plate in front of the puppet. If the child provided the correct number of fish, the trials progressed in the order 1-3-5-6-6. If the child provided an incorrect number of fish on any trial, the child was asked for N-1 fish. The trials proceeded until the child answered correctly at least twice for N and failed at least twice for N+1, or until the child successfully provided six fish twice. The child's score was equal to this final value of N.

Standardized Assessments Children's mathematical ability was assessed with the Test of Early Mathematics Ability (TEMA-3) (Ginsburg & Baroody, 2003), which consists of a series of verbally administered questions that assess age-appropriate counting ability, number-comparison facility, numeral literacy, and basic calculation skills. To assess general intelligence, children completed two verbal and two nonverbal subtests of the Reynolds Intellectual Assessment Scales (RIAS) (Reynolds & Kamphaus, 2003). A composite IQ score was calculated for each child.

Results

The first series of analyses examined the relation between ANS acuity, as indexed by *w*, counting knowledge, and future symbolic math achievement. First we confirmed that ANS acuity at 3.5 years predicted symbolic math achievement one year later at 4.5 years after controlling for IQ (Figure 1). We performed a linear regression analysis with *w* and IQ at 3.5 years entered as possible predictors of math achievement at 4.5 years. Both *w* and IQ emerged as significant predictors of math achievement (Table 1). In a second regression model, we added counting knowledge as an additional predictor. Consistent with the hypothesis that counting knowledge may mediate the relation between *w* and math achievement, when we added counting knowledge to the model its contribution to TEMA was significant, whereas *w* was no longer a significant predictor. In addition, this model captured a larger proportion of variance than the model without counting knowledge ($F = 22.103$, $p < .001$). Further, when counting knowledge and IQ were used to predict TEMA, the residuals did not account for significant variation in *w* ($F(1, 82) = 2.998$, $p = .09$). This suggests that ANS acuity is not directly impacting math performance but instead may be acting via its influence on counting proficiency.

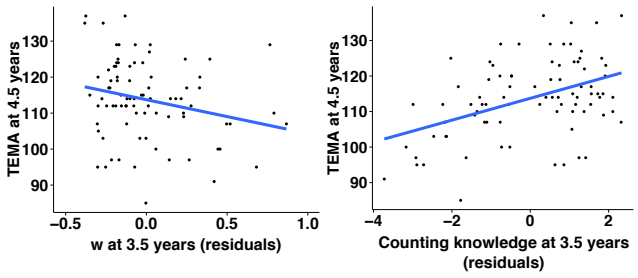


Figure 1. Scatterplots illustrating the relation between *w* at 3.5 years and math achievement at 4.5 years controlling for IQ (left) and between counting knowledge at 3.5 years and math achievement at 4.5 years controlling for both *w* and IQ (right).

Table 1: Regression models predicting math achievement at 4.5 years using measures collected at 3.5 years.

	Model 1		**Model 2**	
R^2	.108		.292	
F-statistics	$F(2, 81) = 6.00$		$F(3, 80) = 12.41$	
p-statistics	$p < .004$		$p < .001$	
Predictor	$\beta_{Adjusted}$	*p*	$\beta_{Adjusted}$	*p*
ANS Acuity	-.245	.024	-.170	.079
IQ	.216	.045	.097	.322
Cardinal knowledge	--	--	.460	< .001

The next series of analyses examined the relation between *w*, approximate arithmetic performance, and symbolic math achievement all at 4.5 years of age (Figure 2). We began by performing a linear regression analysis to confirm that *w* predicts math performance after controlling for IQ (Table 2). In the next regression model, we added approximate arithmetic performance as another possible predictor. In this model, all predictors contributed significant unique variance, and this model explained significantly more variance in math ability than the model without approximate arithmetic ($F = 9.438$, $p = .003$). However, when approximate arithmetic performance and IQ were regressed on symbolic math achievement, these residuals were significantly correlated with *w* ($F(1, 141) = 4.674$, $p = .03$). Therefore, it appears that approximate arithmetic and ANS acuity each contribute unique variance to symbolic math performance at 4.5 years of age.

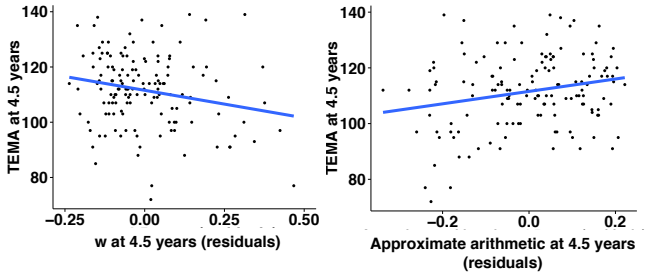


Figure 2. Scatterplots illustrating the relation between *w* and math achievement at 4.5 years controlling for IQ (left) and the relation between approximate arithmetic performance and math achievement at 4.5 years controlling for both *w* and IQ (right).

Table 2: Regression models predicting math achievement using measures collected at 4.5 years of age.

	Model 1		**Model 2**	
R^2	.197		.243	
F-statistics	$F(2, 140) = 18.42$		$F(3, 139) = 16.16$	
p-statistics	$p < .001$		$p < .001$	
Predictor	$\beta_{Adjusted}$	*p*	$\beta_{Adjusted}$	*p*
ANS Acuity	-.219	.005	-.171	.03
IQ	.357	< .001	.321	< .001
Approximate Arithmetic	--	--	.234	.003

In a final combined model, we asked whether cardinal number knowledge and approximate arithmetic performance each contribute variance to children's math achievement[1]. This model revealed that both factors make unique contributions math achievement ($F(4, 104) = 17.74$, $R^2 = .383$, $p < .001$; Cardinal knowledge: $\beta = .419$, $p < .001$; Approximate arithmetic: $\beta = .196$, $p = .019$; ANS acuity: $\beta = -.141$, $p = .082$; IQ: $\beta = .129$, $p = .116$), which suggests that both the precision of children's ANS and their ability to manipulate the ANS each impact symbolic math achievement.

Discussion

The goal of the present research was to investigate the mechanisms by which approximate number representations contribute to preschoolers' emerging symbolic math capabilities. We first replicated previous findings (e.g., Libertus, Feigenson, & Halberda, 2011; Starr et al., 2013; vanMarle et al., 2014) that individual differences in the precision of the ANS are related to symbolic math achievement in preschool-aged children. We found that ANS acuity at age 3.5 was predictive of math achievement one year later after controlling for general IQ. Next we turned to quantitative abilities that may serve as intermediary steps between the ANS and symbolic math. We focused on cardinal number knowledge and nonsymbolic approximate arithmetic, which are abilities that have been previously found to correlate with symbolic math performance in young children. Our results suggest that there may be multiple routes by which the ANS contributes to symbolic math achievement. With regards to the precision of the ANS, it appears that ANS acuity may indirectly influence math achievement via children's acquisition of numerical symbols and the cardinality principle. In addition, nonsymbolic approximate addition performance contributes unique variance above and beyond that contributed by ANS precision. Together, these results suggest that both the acuity and manipulability of the ANS influence children's early math performance.

Our finding that cardinal knowledge is an intermediary step between the ANS and symbolic math is consistent with prior work that has identified cardinal knowledge as a key mediator of the relation between ANS acuity and math achievement in preschool-aged children (Chu, vanMarle, & Geary, 2015; vanMarle et al., 2014; see Schneider, Beeres, Coban, & Merz, 2016 for meta-analysis). In this light, the finding that symbolic numerical comparison is a stronger predictor of symbolic math achievement than nonsymbolic numerical comparison (Schneider et al., 2016) is not necessarily a refutation of the idea that the ANS contributes to symbolic math. Rather, cardinal representations of symbolic numbers may serve as stepping-stones between approximate number representations and symbolic mathematical operations. When children begin to form mappings between numerical symbols and approximate quantities, the ease with which these mappings are formed and the quality of these mappings may be influenced by the precision of children's ANS. Once this mapping is formed and children become proficient with numerical symbols, the

[1] *w* values from the 4.5-year visit were used because the values at 3.5 and 4.5 years of age were significantly correlated ($r = .31$, $p < .01$) and using the 4.5-year values enabled us to include a larger proportion of our sample.

precision of the ANS itself may not directly impact math achievement.

In the present study we focused on children's understanding of cardinality. Therefore, although our results point towards the importance of cardinal number knowledge for early math proficiency, it remains to be seen how other forms of symbolic number knowledge, including ordinality, may relate to math ability. Indeed, symbolic numerical ordering ability has been found to mediate the relation between ANS acuity and symbolic math achievement in adults (Lyons & Beilock, 2011). However, a recent training study with adults found that training adults on a symbolic numeral ordering task did not produce gains in symbolic arithmetic ability (Park & Brannon, 2014). Additional work is therefore needed to investigate how the ANS supports symbolic number representations throughout the lifespan, particularly after children have mastered cardinality, and how this relation contributes to symbolic math achievement.

The present results also provide support for a second route by which the ANS influences math ability. Beyond basic quantity representation, the ANS also supports arithmetic operations in infants, preschool children, and monkeys, all of whom have no understanding of symbolic arithmetic (Barth et al., 2005; Cantlon & Brannon, 2007; McCrink & Wynn, 2004). This capacity may provide an intuitive basis for the acquisition of symbolic arithmetic principles. Consistent with this view, we found that approximate arithmetic ability in 4.5-year-olds was a significant predictor of performance on the TEMA, which is primarily composed of symbolic numerical problems. Further, approximate arithmetic ability predicted unique variance in TEMA scores that was not accounted for by ANS acuity. As evidenced by a recent series of training studies, the link between approximate and symbolic arithmetic extends into adulthood: training adults on a nonsymbolic arithmetic task similar to that employed here led to significant gains in their symbolic arithmetic performance. (Park & Brannon, 2013; 2014). Critically, training on other types of tasks, including nonsymbolic numerical comparison, did not transfer to symbolic arithmetic. Although both nonsymbolic numerical comparison and approximate arithmetic tasks require representing approximate numerical quantities, approximate arithmetic additionally requires the manipulation of those quantities. This manipulation component appears to be a critical factor in the relation between the ANS and symbolic math.

The majority of studies relating the approximate number system to symbolic math have focused on individual differences in the acuity of approximate number representations. However, the manipulability of these representations may be a second mechanism by which the ANS influences symbolic math. Approximate arithmetic may share a cognitive foundation with symbolic arithmetic, making it an attractive target for potential interventions to remediate deficits in math proficiency. A priority for future studies should therefore be to adapt these training paradigms for children and determine if improving children's approximate arithmetic abilities also produces benefits for symbolic math. In addition, the relation between the precision and manipulability of the ANS remains relatively unexplored and is deserving of further study.

The ANS endows young children with a robust sense of quantity prior to beginning formal mathematics training. Although many studies have provided evidence for a correlation between the fidelity of the ANS and symbolic math achievement, there remain key open questions concerning the mechanisms underlying this relation and whether this correlation is indicative of a causal relation. In the present study, we identified two potential pathways through which the ANS influences preschoolers' early math proficiency. The first role for the ANS may occur when children are beginning to learn the meaning of number words, at which point approximate number representations may serve as an anchor for acquiring symbolic number representations. Therefore, the precision of the ANS may influence children's facility with mapping numerical symbols to approximate magnitudes. A second role for the ANS may stem from its ability to support arithmetic operations. The shared demand for manipulating quantities may form a conceptual bridge between nonsymbolic and symbolic arithmetic. The present results therefore suggest a nuanced relation between approximate number representations and symbolic math achievement in which multiple features of the ANS contribute to the emergence of symbolic math ability in young children. In light of these results, interventions designed to target one or both of these pathways may be differentially beneficial for children depending on their level of symbolic number knowledge and mathematical proficiency.

Acknowledgments

We would like to thank Anna Beth Keith, Simone Halliday, and Celia Litovsky for their assistance with data collection, as well as all of the families who participated in this research. This work was funded by NSF REESE and DLS grants and the James McDonnell Scholar Award to EMB, and an NSF GRFP and SRCD SECC Dissertation Research Funding Award to AS.

References

Barth, H. C., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, *102*(39), 14116–14121.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293.

Cantlon, J. F., & Brannon, E. M. (2007). Basic Math in Monkeys and College Students. *PLoS Biology*, *5*(12), e328.

Carey, S. (2004). Bootstrapping & the Origin of Concepts. *Daedalus*, *133*(1), 59–68.

Chen, Q., & Li, J. (2013). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163–172.

Chu, F. W., vanMarle, K., & Geary, D. C. (2015). Early numerical foundations of young children's mathematical development. *Journal of experimental child psychology*, *132*, 205-212.

Dehaene, S. (1997). The number sense: How the mind creates mathematics. Oxford University Press.

Duncan, G. J., Dowsett, C. J., & Claessens, A. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428-1446.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1), 43–74.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and achievement in the first year of formal schooling in mathematics. *Cognition*, *115*(3), 394-406.

Ginsburg, H. P., & Baroody, A. J. (2003). Test of Early Mathematics Ability (3rd ed.). Austin, TX: Pro-Ed.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457-1465.

Hubbard, E. M., Diester, I., Cantlon, J. F., Ansari, D., Van Opstal, F., & Troiani, V. (2008). The Evolution of Numerical Cognition: From Number Neurons to Linguistic Quantifiers. *Journal of Neuroscience*, *28*(46), 11819–11824.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867.

Klibanoff, R. S., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. V. (2006). Preschool children's mathematical knowledge: The effect of teacher 'math talk.'. *Developmental Psychology*, *42*(1), 59–69.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Is approximate number precision a stable predictor of math ability? *Learning and Individual Differences, 25, 126-133.*

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences*, *109*(46), 18737–18742.

Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition*, *121*(2), 256–261.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS ONE*, *6*(9), e23749.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*(11), 776–781.

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science*, *24*, 2013–2019.

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition*, *133*(1), 188–200.

Parsons, S., & Bynner, J. (2006). *Does numeracy matter more?* National Research and Development Centre for Adult Literacy and Numeracy, Institute of Education, University of London.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., et al. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41.

Pinheiro-Chagas, P., Wood, G., Knops, A., Krinzinger, H., Lonnemann, J., Starling-Alves, I., et al. (2014). In How Many Ways is the Approximate Number System Associated with Exact Calculation? *PLoS ONE*, *9*(11), e111155.

Reynolds, C. R., & Kamphaus, R. W. (2003). Reynolds Intellectual Assessment Scales.

Schneider, M., Beeres, K., Coban, L., & Merz, S. (2016). Associations of non‐symbolic and symbolic numerical magnitude processing with mathematical competence: a meta‐analysis. *Developmental Science.*

Siegler, R. S., & Opfer, J. E. (2003). The Development of Numerical Estimation. *Psychological Science*, *14*(3), 237-250.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, *110*(45), 18116–18120.

vanMarle, K., Chu, F. W., Li, Y., & Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Developmental Science, 17(4), 492-505*.

Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220–251.

Construal level affects intuitive moral responses to narrative content

Nicholas A. Lester (nlester@umail.ucsb.edu)
Department of Linguistics, South Hall 3432
Santa Barbara, CA 93106 USA

René Weber (renew@comm.ucsb.edu)
Department of Communication, 4005 Social Sciences & Media Studies Bldg
Santa Barbara, CA 93106 USA

Abstract

The Model of Intuitive Morality and Exemplars (MIME) predicts a mutual dependency between the moral scrutiny of mediated narratives and media exposure. This study proposes moral judgments of media content are not only related to basal moral domain salience and exemplars, but also to the immediate processing state of the individual at the moment of exposure. An experiment manipulating construal level prior to exposure to a mediated narrative was conducted to test this proposal. The results suggest that evaluations of moral violations are modulated by construal level. High-level construal led to harsher, more consistent judgments of domain-violator morality, eliminating the effect of baseline moral intuitions. Low-level construal induced an apparent trade-off in moral evaluation strategy which is sensitive to both narrative outcome and domain salience. When domain violators were punished, intuitive moral salience was negatively correlated with moral evaluations; however, when domain violators were rewarded, the opposite trend emerged. These findings indicate the need for an adjustment to the MIME model to allow for processing states to interact with moral domain salience and moral judgments of media content. They also suggest that the strength and quality of moral intuitions are not robust to broader cognitive processes, but interact with them.

Keywords: Model of intuitive morality and exemplars; moral foundation theory; construal level theory; media enjoyment.

Introduction

Recent work on media perception has emphasized the role of instinctive moral responses in shaping the enjoyment of media narrative. This research explains that the affective dispositions that individuals form towards characters and events in a given dramatic context vary as a function of their basal moral inclinations. For example, characters that violate heavily weighted moral principles tend to be evaluated more negatively than characters that violate less heavily weighted principles (Tamborini, Eden, Bowman, Grizzard, Weber, & Lewis, 2013). These intuitive moral inclinations are thought to interact with experientially derived schemas or *exemplars*, which allow for the fast mapping of moral expectations onto characters (e.g., hero vs. villain), even on the basis of very little evidence. However, the specifics of the timeline for the operation of schema-based and intuition-based processes remains a point of debate (Raney, 2004).

The most comprehensive model to emerge from this strand of research is the Model of Intuitive Morality and Exemplars (MIME; Tamborini, 2013). The MIME integrates short-term and long-term processes to account for individuals' local responses to media content and global patterns of selective media exposure. The short-term processes identified so far have dealt with relatively enduring qualities of the individual (e.g., basal moral intuitions) or properties of the stimulus itself (e.g., similarity to available moral-narrative schemas or presence/absence of moral conflict). However, several studies working from the perspective of Construal-level Theory (CLT) have demonstrated that more transient cognitive states can have powerful effects on moral judgment (Eyal, Liberman, & Trope, 2008; Gong & Medin, 2012; Žeželj & Jokić, 2014). In the present study, we attempt to link these two strands of research.

Intuitive Morality and MIME

The fundamental structure of the MIME is based on the core tenants of Disposition Theory (DT; Zillmann, 2000). Both approaches hold that the enjoyment of dramatic narrative hinges on the satisfaction of moral expectations developed relative to characters and the outcomes they experience. In the classical model, these moral expectations arise from continuous moral scrutiny over the course of the narrative. If we deem a character to be morally righteous, we form a positive affective disposition towards that character. This disposition leads us to empathize with that character, and so to hope that they receive positive outcomes. The opposite situation holds for characters toward which we have developed negative dispositions: we do not empathize with them and hope for negative outcomes. This path from moral scrutiny to empathic response has received empirical support. Weber, Tamborini, Lee, and Stipp (2008) asked participants to rate the characters and plots of a popular soap opera along several dimensions: morality of characters, liking of the characters, and perceived positivity/negativity of character outcomes. Their findings show correlations between perceived character morality, degree of positivity or negativity of outcome relative to perceived character morality, and character liking. But what system informs the constant moral evaluations which underlie the formation of affective dispositions?

Moral Foundations Theory (MFT; Haidt, 2001; Haidt & Joseph, 2007) provides part of the answer. MFT proposes that every moral judgment is in part determined by a spontaneous, emotionally driven response – a gut intuition – that appears immediately and with a simple valence. These intuitions emerge in five distinct domains: care/harm,

fairness/cheating, loyalty/betrayal, authority/ subversion, and sanctity/degradation (Haidt & Joseph, 2007). The care/harm domain relates to empathy and suffering; fairness/cheating pertains to reciprocity rooted in a basic sense of justice; loyalty/betrayal refers to the upholding of social bonds; authority/subversion relates to social hierarchies; and sanctity/degradation captures a general aversion to disgusting objects or actions. These domains are cross-culturally stable; however, the relative weighting of the domains – so-called *domain salience* – tend to vary between cultures, subcultures, and even individuals.

Tamborini (2013) places these five dimensions at the heart of MIME. The baseline salience of each domain in both absolute and relative terms interacts with cultural environment, media exposure, and situational exemplar salience to determine moral appraisals. Expanding on the research of Weber et al. (2008), a few studies have begun to validate the effects of domain salience on the formation of affective dispositions towards characters and outcomes. For instance, Tamborini et al. (2013) presented participants with several brief narratives. In each narrative, the main character commits a transgression relevant to only one of the intuitive moral domains. The narratives were also manipulated for outcome. Under one condition, the main characters experienced positive outcomes; in the other condition, they experienced negative outcomes. Tamborini and colleagues also collected information about how heavily the participants' tended to weight each moral domain when making moral judgments. As predicted, moral transgressions were judged more harshly when participants cared more about the domain.

One possibility not explored in the context of MIME so far is that moral intuitions and exemplar salience might be shaped by a person's immediate *cognitive environment*. Cognitive environment here refers to general properties of an individual's cognitive processing state immediately prior to and concurrent with the media exposure. One reason to believe that cognitive environment could affect even these rapid, automatic moral intuitions comes from the growing literature on CLT (Trope & Liberman, 2010). Studies in this vein have repeatedly shown that judgments from many different cognitive domains depend on how the individual represents or *construes* the target of a judgment before it is passed on to other cognitive systems.

Construal-level Theory

CLT proposes that one and the same stimulus can be processed – or *construed* – at different levels of abstraction (or concreteness). The system that imposes this degree of abstraction, or *construal level*, exerts its effect before other 'downstream' cognitive systems have access to the conceptual material. As such, it acts as a kind of filter, one which determines both what information is to be extracted from the perceptual input and how that information should be integrated into the resulting concept. Construal levels are usually divided into two opposing types, *high-level construals* and *low-level construals*. High-level construals have been shown to result in schematic, homogeneous, and decontextualized representations. Low-level construals, on the other hand, produce specific, heterogeneous, and context-dependent representations. Differences in construal level have been shown to affect downstream cognition in several domains (see Trope and Liberman, 2010, for a review).

CLT and Moral Evaluation

Eyal, Liberman, and Trope (2008; Study 1) tested whether manipulations of construal level would alter participants' evaluations of moral transgressions. Participants were asked to read several vignettes containing morally reprehensible behavior (along with mitigating circumstances that might partially justify that behavior; e.g., a family eating their pet dog after it has been accidentally run over). After reading each vignette, the participants were presented either with an expression summarizing the general moral principle at stake (e.g., *dishonoring the family pet*; high-level construal) or a specific description of the actions involved (e.g., *eating the meat of a dead dog*; low-level construal). Eyal et al. found that participants in the high-level construal condition rated the transgressions as more negative than participants in the low-level construal condition. These results were replicated in several follow-up experiments with different manipulations of construal level.

Eyal and colleagues explain these findings by suggesting that construal level operates independently of and prior to our intuitive emotional responses. High-level construals mask mitigating details while low-level construals draw them into focus. The difference in representation leads to a difference in moral appraisal. Further support for this explanation comes from Agerström and Björklund (2009), who found that high-level construal resulted in harsher appraisals of individuals who acted selfishly under conditions favoring altruistic behavior.

Gong and Medin (2012) were unable to replicate these findings. Using slightly different (though arguably more direct) construal-level priming techniques, they found that high-level construal resulted in *less* harsh appraisals low-level construal. Even in a direct replication of Eyal and colleagues' Study 2, they still found a *vitiating* effect of high-level construal (for similar results see Lammers, 2012).

More recent research has supported this challenge. In replications with higher-power designs ($1-\beta = 0.95$) than the studies reported in Eyal et al. (2008) and Gong and Medin (2012), Žeželj and Jokić (2014) found results mostly in line with those of Gong and Medin. Most importantly, they found harsher evaluations for low-level construal when using direct manipulations of construal level (in agreement with Gong and Medin's Study 1).

But why should low-level construal result in harsher evaluation of moral transgressions? Both Gong and Medin (2012) and Žeželj and Jokić (2014) suggest that focusing on contextual factors may render the transgression all the more vivid (or *imageable*). Thus, instead of attenuating the

severity of the act, low-level construal may actually enhance it, leading to a more powerful intuitive-emotional response.

One aspect of the intuitionist moral approach that has not been addressed in the research on CLT and morality so far is the notion of domain salience. No study has controlled for *which* moral domain was being violated or upheld. Indeed, one reason for the conflicting results may be that the domains represented in the vignettes were weighted differently by the different populations recruited for the different studies. These differences in domain salience might lead to differences in the strength of the construal-level effect, or even to a difference in evaluation strategy. Another factor not accounted for in the CLT morality literature is the possible effect of character outcomes. Many of the vignettes adopted in this literature frame the moral transgression in terms of prior mitigating factors (e.g., the family pet is killed accidentally *prior to* its being consumed). However, to our knowledge, only one of the vignettes used in these studies contained a description of what befalls the character *after* the transgression (this is the cheating vignette, in which a student cheats on a test and is rewarded with good grades; see Gong & Medin, 2012, Appendix A). As character outcomes constitute a kind of detail, and one standing temporally between the violation exposure and the judgment measurement, they may interact with low-level construal to influence the severity of the judgment.

Present Study

The present study addresses the shortcomings in the previous section by blending the methodologies of Gong & Medin (2012) with those of Tamborini et al. (2013). We combine a manipulation of construal level with the same narrative stimuli used in Tamborini et al. (2013). In so doing, we explore whether high- or low-level construal of the narrative content impacts the response of the intuitive moral system.

Hypotheses

Given the conflict outlined above, we restrict ourselves to non-directional hypotheses. Harsher moral appraisals under high-level construal will support Eyal et al. (2008) *inter alia*, who predict that mitigating details become available under low-level construal. We refer to this explanation as the *mitigation principle*. Harsher appraisals under low-level construal will support Gong and Medin (2012) *inter alia*, who predict that the graphic detail of low-level construal intensifies the negative emotional response triggered by the moral violation. We refer to this explanation as the *intensification principle*.

All prior studies on CLT and morality have returned an effect of construal level on moral judgment of events. Therefore, we predict that construal level will result in an aggregate difference in the negativity of moral appraisals, all else being equal. This prediction is formalized as H1 below.

H1: *Moral evaluations for domain violators will differ between a low- and a high construal level at the moment of exposure.*

Based on the findings of Tamborini et al. (2013), we also expect increased domain salience to result in harsher appraisals of moral transgressors (H2), though we expect this effect to interact with construal level in a way consistent with the outcome of H1.

H2: *With increasing domain salience moral evaluations for domain violators will decrease. This relationship will be moderated by construal level at the moment of exposure.*

Weber et al. (2008) found that moral transgressors who were perceived as more severely punished were judged more harshly. Thus, character outcomes may operate directly on both moral judgments and narrative appeal. Based on the logic of H2, this outcome-driven effect should further interact with domain salience and construal level. The higher moral evaluations associated with reward outcomes should diminish more sharply as domain salience increases (contextual factors should have less of an effect for people with strong intuitive reactions). Moral evaluations for punishment outcomes, on the other hand, should already be somewhat low for low-salience domains, and should converge with reward narrative as salience increases. As in H2, we expect construal level to moderate this relationship.

H3: *Moral evaluations of domain violators will be lower/worse for punishment outcomes compared to reward outcomes in narratives. With increasing domain salience, moral evaluations will decrease/worsen more rapidly for reward outcomes than for punishment outcomes. This relationship will be moderated by construal level at the moment of exposure.*

Method

The general paradigm is adapted from Tamborini et al. (2013). First, we assess participants' baseline salience for each of the five intuitive domains using the Moral Foundations Questionnaire (MFQ31; Haidt et al., 2006). Then, we prime the participants for high-level or low-level construal by means of a hypernym-hyponym word categorization task (Fujita et al., 2006, Experiment 3; Gong & Medin, 2012, Study 4). Following the priming task, we present the participants with ten brief 'film synopses' in which the main character violates one of the five moral domains and is either punished or rewarded for that transgression. In the critical task, we asked participants to rate the characters' morality.

Participants

217 undergraduates from a university on the west coast of the United States of America were recruited to participate in this experiment.

Stimuli and Design

Stimuli for the experiment include both priming materials and narrative materials. The priming stimuli consisted of

two sentential frames and a set of 40 common nouns to be categorized. In the low-level condition, participants were presented with a sentential frame designed to elicit a hyponym of the target word: *An example of (a/an)* WORD *is* __. For instance, if provided with *cat*, a participant might successfully respond *tabby* but not *animal*. In the high-level condition, participants saw a frame designed to elicit hypernyms of the targets: *(A/An)* WORD *is an example of* __. In this condition, a correct response to *cat* would be animal, but not *tabby.* Semantic categorization tasks of this kind have elsewhere been shown to effectively modulate construal level (Fujita et al., 2006). The 40 nouns to be categorized were extracted from the MRC Psycholinguistic Database (Wilson, 1988). All words were matched for concreteness and familiarity ratings as provided in the MRC annotation (concreteness: M = 593.55, sd = 29.02; familiarity: M = 583.83, sd = 37.21).

The narrative materials were taken directly from Tamborini et al. (2013). These materials comprise 20 narratives. Each narrative concerns a focal character that violates one of the five moral domains, such that each domain is violated in two different narratives (10 narrative skeletons total). Each of these narrative skeletons consists of three parts: a positive introduction to the character (one or two sentences), a description of the transgression (three or four sentences), and a description of the outcome (one sentence). Therefore, our narratives contain both mitigating and damning details regarding each domain violator. This is critical given that the different explanations invoked by the mitigation principle compared to the intensification principle ultimately turn on the availability of cues capable of mitigating or intensifying the offense. The first two parts for each narrative are identical across all conditions. The third part varies as a function of the outcome manipulation: in the 'reward' condition, the final sentence describes a positive outcome; in the 'punishment' condition, it describes a negative outcome. For a full discussion, see Tamborini et al. (2013).

We employ a 2 x 2 x 5 mixed-effect design. Our between-subjects variable is construal level; our within-subjects variables are outcome and domain. Participants were randomly assigned to one of four possible conditions based on a crossing of narrative block (ordering) and construal-level condition. Each narrative block contained ten narratives (two for each of the five domains). The blocks were counterbalanced across the outcome condition, such that participants in the two block conditions saw the same narratives with opposite (punishment vs. reward) endings. In the low-level and high-level construal conditions, participants completed the hyponym and hypernym word categorization tasks, respectively.

Procedure

All measures and stimuli were presented on a 17-in LCD display with 1366 X 768 screen resolution. The interface and data collection were implemented through OpenSesame 0.27.4 (Mathôt et al., 2012). As in the previous studies using these narratives, participants were told that they would be viewing a number of brief film synopses.

Prior to the main task, each participant completed the Moral Foundations Questionnaire (MFQ31; Haidt, Graham, & Hersh, 2006), a 31-item survey with two parts. In Part One, 15 factors involved in making moral judgments (e.g., "Whether or not someone was harmed") are rated on a scale from 0 = "not at all relevant" to 5 = "extremely important". This section also includes one catch question ("Whether or not someone believed in astrology") designed to help in identifying disengaged participants. In Part Two, 15 statements regarding personal beliefs are rated from 0 = "strongly disagree" to 5 = "strongly agree". Presentation of MFQ items was randomized within each part for each participant, though all participants completed Part One before Part Two.

In the main task, participants read each of the ten narratives described above. Each narrative was presented individually, followed by three questions adopted from Tamborini and colleagues' domain violation questionnaire. The critical question elicited a moral evaluation of the focal character based on their actions (*How moral were* NAME's *actions in this movie plot?*); the others (not evaluated here) targeted perceived ethicality and enjoyment of the narrative. Responses were collected using a 7-point Likert scale running from 1 = 'extremely immoral' to 7 = 'extremely moral.' The order of presentation of both the narratives and the domain violation questions was randomized for each participant. Responses and reaction times were collected. Only the former are considered here.

Data Trimming

Prior to running our data analysis, we omitted several problematic data points on *a priori* grounds. Five participants were excluded from the study due to technical errors. Another eight participants were removed due to unreliable MFQ responses, as indicated by a response of 3 or higher to the 'catch' question. Such a response indicates low engagement with the task. In addition, we removed any responses with response latency less than 2000 ms.

Results

We fit a linear mixed effect models to predict participants' moral evaluations of focal domain violators. Participant and narrative type (the narrative skeleton, two per domain) were included as random effects (intercepts only) to partial out the idiosyncrasies of individual participants and narratives. We also include a fixed effect of domain (harm, fairness, authority, purity, loyalty) as a control to ensure that our results generalize to the total intuitive-moral system. To test H1, we include a main effect of construal level. To test H2, we add a main effect of domain-salience (MFQ scores) and the interaction between construal level and domain salience. Finally, to test H3, we further add a main effect of outcome condition (punishment or reward), along with the three-way interaction between outcome condition, construal level and domain salience. Prior to the regression analysis, the raw

evaluation scores were transformed by adding one to avoid the presence of zeroes.

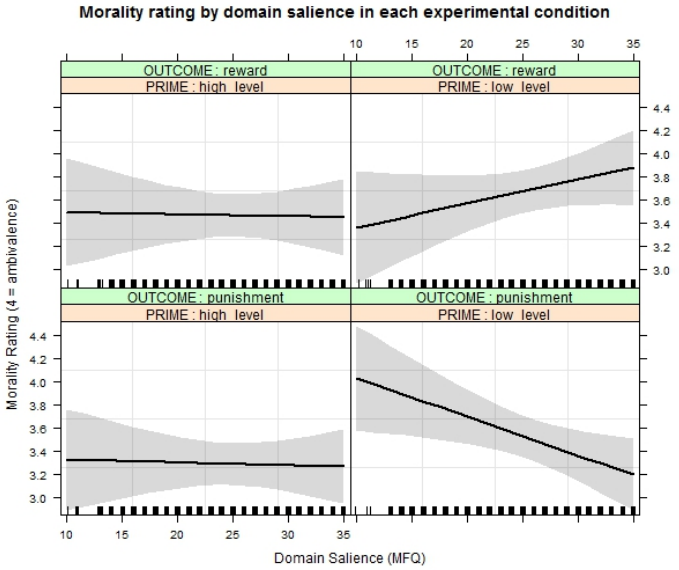


Figure 1: Interaction of construal level, narrative outcomes, and domain salience on morality ratings

H1 and H2

Our model did not uncover a significant main effect of construal level ($F(1, 1259.34) = 0.85$, n.s.). Neither did the two-way interaction between construal level and domain salience reach significance ($F(1, 1379.00) = 0.10$, n.s.). Unexpectedly, we also found no main effect of domain salience ($F(1, 1083.07) = 0.25$, n.s.). Therefore, we do not find support for a simple relationship between construal level and domain salience. It appears that construal level has interfered with general effect of domain salience observed in other studies.

H3

We did uncover a significant three-way interaction of construal level, domain salience and outcome condition ($\beta = \pm 0.05$, $F(1, 1345.58) = 4.55$, $p < 0.03$). These results are summarized in Figure 1.

High-level construals (left two panels) are insensitive to domain salience, as indicated by the essentially zero slope of both regression curves. For low-level construals (right two panels), the picture is more complex. On the one hand, domain violators that were punished for their transgressions were rated more negatively as domain salience increased. On the other hand, domain violators that were rewarded for their transgressions were given increasingly *positive* ratings as domain salience increased.

Discussion and Conclusions

The present experiment provides support for a moderating role of construal level in the formation of moral appraisals during exposure to media content. Whereas previous studies have demonstrated generally negative correlations between moral assessment of domain violators and the intuitive salience for the violated domain, the present study only replicated this pattern under a single condition: low-level construal of narratives with negative outcomes for domain violators. If the character receives a positive outcome under low-level construal, the trend is entirely reversed. Moreover, if the participant is operating under high-level construal with either positive or negative outcome, the effect of domain salience disappears altogether.

Our findings shed light on the current controversy between mitigation-based and intensification-based accounts of the effects of low-level vs. high-level construal. Both accounts emphasize the effect of details extracted through low-level construal, but differ with respect to the nature of this effect. The mitigation principle (e.g., Eyal et al., 2008) states that low-level construal increases moral ratings by focusing details that deflate the intensity of the emotional response. The intensification principle (e.g., Gong & Medin, 2012) states that low-level construal decreases moral ratings by focusing details that exacerbate the negativity of the emotional response. Our results speak in favor of both principles, though each appears to operate only under certain conditions.

First, consider the imperviousness of high-level construal to details within the narratives. These details include the character introductions (which included positive statements about the characters, casting them as upholders of other moral domains) and the outcomes they experienced. No matter how (un)important the domain being violated, these details were not capable of raising morality scores above floor performance. This fact is predicted by both the mitigation and the intensification accounts in that both assume a lack of conceptual articulation for high-level construal. However, the fact that high-level construal consistently produces lower ratings relative to low-level construal is not compatible with the intensification principle, which states that moral evaluations under low-level construal should sink well below – not approach – those of high-level construal. Instead, this relationship meshes with the mitigation principle; lack of access to relevant mitigating details leads to low morality judgments.

But why should high-level construal eliminate the effect of domain salience? To answer this question, we look to Raney's extension to DT. Raney (2004) proposed that not all moral evaluations involve an intuitive moral response to the actual content of the stimulus. Rather, he suggests that we often rely on basic *scripts* or story exemplars to evaluate characters. These scripts arise as we generalize over prior experience with similar narratives. They allow us to rapidly assign moral roles and expectations to characters. Returning to the present study, the abstract representations generated by high-level construal could lead to more focused (i.e., low-competition) activation of archetypal scripts and exemplars through their similarity in form (*conceptual-structural isometry*). Support for this notion comes from the fact that similarity in representational structure has been shown to facilitate (analogical) co-mapping in other

domains (e.g., Falkenheiner, Forbus, & Gentner, 1989). This enhanced connectivity between stimulus and schema could allow for rapid mapping between the domain violator and an abstract 'villain' role, resulting in lowered morality appraisals across the board.

Low-level construal yields a much more complicated pattern of responses. Punishment narratives produce a negative association between domain salience and morality ratings, while reward narratives produce a positive association. Thus, punishment narratives support the intensification principle, while reward narratives support the mitigation principle. Despite the apparent contradiction, both of these trends may be explained by a single mechanism within the MIME: the reciprocal link between exemplars and intuitive moral responses. To understand this point, we must deal with three facts. First, low-level construal is sensitive to domain of violation, indicating engagement with the intuitive moral system (unlike high-level construal). Second, as domain salience increases, so does the impact of the outcome effect, suggesting a 'blind' (i.e., non-monotonic) boost to the signal based on domain salience. Finally, the directions of the effects seem to reflect the quality of the outcome – punished characters are less moral than rewarded characters, even though both committed the same violation. Thus, we propose that the detailed and context-dependent representations of low-level construal activate the intuitive moral system, which introduces a salience-proportional boost to the signal as activation spreads into the exemplar network. This activation spreads relative to the content of the narrative, including the violation and the outcome. The more typical punishment narratives feed activation into a coherent set of compatible schemata dealing with the punishment of moral transgression. In turn, these schemata reinforce negative appraisals of the violator through the broad affective annotation of a common 'villain' role. For reward narratives, however, several incoherent and non-compatible schemata are simultaneously activated: the violation engages violation/punishment schemata like those described above, while the reward outcome connects to virtue/reward schemata. The advantage for the latter schemata may be a result of the relative recency of the outcome as compared to the violation. Such a recency effect could overshadow the competing negative push of the violation to create the overall positive interaction observed here.

The mechanisms sketched above have far-reaching implications for our understanding of construal level, moral appraisal, and response to media content. They raise many important questions related to the dynamics (i.e., flexibility) of the construal level system during extended exposure to morally relevant content and how the local effects reported here translate into long-term tuning of the moral system.

References

Agerström, J., & Björklund, F. (2009). Moral concerns are greater for temporally distant events and are moderated by value strength. *Social Cognition, 27,* 261–282.

Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology*, *44*, 1204-1209.

Falkenheiner, B., Forbus, K. D., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and examples. *Artificial Intelligence*, *41*, 1-63.

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self control. *Journal of Personality and Social Psychology*, *90*(3), 351-367.

Gong, H. & Medin, D. L. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision Making*, *7*(5), 628-638.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814-834.

Haidt, J. & Joseph, C. (2007). The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind*, Vol. 3 (pp. 367–391). New York, NY: Oxford.

Haidt, J., Graham, J., & Hersh, M. A. (2006). *The moral foundations questionnaire*. Retrieved from http://faculty.virginia.edu/haidtlab/mf.html.

Lammers, J. (2012). Abstraction increases hypocrisy. Journal of Experimental Social Psychology, 48, 475-480.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324.

Raney, A. A. (2004). Expanding disposition theory: Reconsidering character liking, moral evaluations, and enjoyment. *Communication Theory*, *14*(4), 348-369.

Tamborini, R. (2013). A model of intuitive morality and exemplars. In R. Tamborini (Ed.), *Media and the moral mind* (pp. 43-74). New York: Routledge.

Tamborini, R., Eden, A., Bowman, N. D., Grizzard, M., Weber, R., & Lewis, J. (2013). Predicting media appeals from instinctive moral values. *Mass Communication and Society*, *16*(3), 325-346.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440-463.

Weber, R., Tamborini, R., Lee, H., & Stipp, H. (2008). Soap opera exposure and enjoyment. A longitudinal test of disposition theory. *Media Psychology*, 4(1), 462-487.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary version 2.00. *Behavior, Research Methods, Instruments, and Computers*, *20*(1), 6-10.

Žeželj, I. L. and Jokić, B. R. (2014). Replication of experiments evaluating impact of psychological distance on moral judgment (Eyal, Liberman, & Trope, 2008; Gong & Medin, 2012). *Social Psychology*, *45*(3),223-231.

Zillmann, D. (2000). Basal morality in drama appreciation. In I. Bondebjerg (Ed.), *Moving images, culture, and the mind* (pp. 53–63). Luton: University of Luton Press.

Beyond Markov: Accounting for Independence Violations in Causal Reasoning

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003 USA

Abstract

Although many theories of causal cognition are based on causal graphical models, a key property of such models—the independence relations stipulated by the Markov condition—is routinely violated by human reasoners. Two accounts of why people violate independence are formalized and subjected to experimental test. Subjects' inferences were more consistent with a *dual prototype model* in which people favor network states in which variables are all present or all absent than a *leaky gate model* in which information is transmitted through network nodes when it should normatively be blocked. The article concludes with a call for theories of causal cognition that rest on foundations that are faithful to the kinds of causal inferences people actually draw.

The last 25 years has seen a dramatic increase in research asking how causal knowledge influences many acts of cognition, including reasoning, categorization, decision making, and learning. Theory in this area has been advanced by use of the formalism known as *causal graphical models* that provides a good first order approximation of human abilities. Nevertheless, there is now considerable evidence that a defining feature of these models—the Markov condition—is routinely violated by reasoners. This article describes these violations and then compares two new models that account for them, albeit in different ways. The diverging predictions of these models are then tested in a new experiment.

Independence Violations in Causal Reasoning

The causal networks in Figs. 1 and 2 provide two prominent examples of how human reasoners violate the independence relations associated with causal graphic models. Starting with the common cause network in Fig. 1 in which X is a common cause of Y_1 and Y_2, whereas Y_1 and Y_2 are unconditionally dependent (e.g., one can reason from the presence of Y_1 to the likely presence of X and then to the likely presence of Y_2) the Markov condition stipulates that they are independent conditioned on X (i.e., $Y_1 \perp Y_2|X$). In this case X "screens off" the Ys from one another (Pearl, 2000; Spirtes, 2000). Because it will be useful to characterize the sign and magnitude of the dependence between two variables, I define a function D (for "delta") that characterizes the difference in the probability of one variable given the presence or absence of another. Moreover, I define that difference in terms of probabilities transformed into log odds. For example, $D(Y_1, Y_2)$ is defined as,

$$D(Y_1, Y_2) = logit\,(p(y_1^1|y_2^1)) - logit\,(p(y_1^1|y_2^0)) \quad (1)$$

where y_i^x denotes $Y_i = x$ (e.g., y_1^1 means Y_1 is present). Because within-network causal relations will have the same properties (e.g., have the same causal strength) in this work, symmetry entails that $D(Y_1, Y_2) = D(Y_2, Y_1)$, which will therefore be abbreviated $D(Y)$.

Analogously, $D(Y_1, Y_2|x^1)$—the difference in (the logit of) the probability of Y_1 given the presence or absence of Y_2, conditioned on x^1—is defined as,

$$D(Y_1, Y_2|x^1) = logit\,(p(y_1^1|y_2^1, x^1)) - logit\,(p(y_1^1|y_2^0, x^1)) \quad (2)$$

and is abbreviated $D(Y|x^1)$.

The normative predictions for $D(Y)$ and $D(Y|x^1)$ are presented in the left chart of Fig. 1 assuming that the common cause graph is parameterized such that $p(x^1) = .50$, that the generative "causal power" associated with both $X \rightarrow Y_1$ and $X \rightarrow Y_2$ is .80, that there are weak alternative causes of Y_1 and Y_2 such that $p(y_1^1|x^0) = p(y_2^1|x^0) = .20$, and that multiple causal influences integrate according to a standard noisy-or function (Cheng, 1997). The figure illustrates the independence relations that characterize a common cause network: Whereas $D(Y) > 0$, $D(Y|x^1) = 0$, that is, the Ys are independent conditioned on x^1.

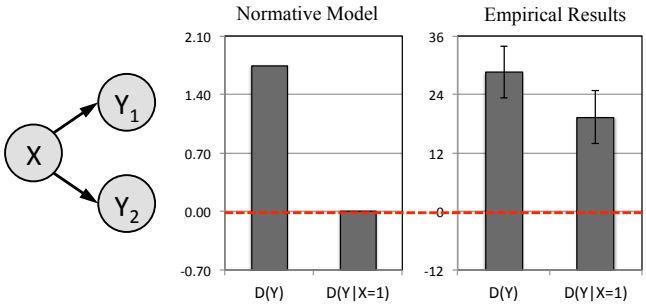


Figure 1. Normative predictions for the common cause model on the left and the corresponding empirical rating from Rehder & Waldmann (2015). Error bars are 95% confidence intervals.

The right chart of Fig. 1 illustrates how that independence relation is typically violated. Rehder & Waldmann (2015) instructed subjects on causal relations that formed a common cause model in the domain of economics, meteorology or sociology. For example, for economics subjects were told that "low interest rates causes small trade deficits" ($X \rightarrow Y_1$) and that "low interest rates causes high retirement savings" ($X \rightarrow Y_2$). The variables senses involved in the causal links were varied (e.g., *high* rather than low interest rates sometimes played the role of X). Subjects were then presented with a series of inferences in which they were asked to rate (on a 0-100 scale) the probability of one variable conditioned on others. We found that the empirical analog of $D(Y|x^1)$ (the difference between $rating(y_i^1|y_j^1, x^1)$ and $rating(y_i^1|y_j^0, x^1)$) was about 18 points (Fig. 1).

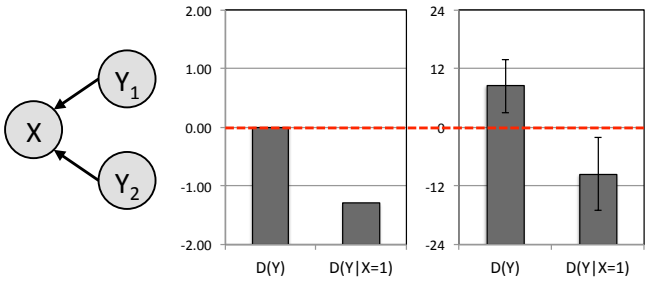


Figure 2. Normative predictions for a common effect model and the corresponding empirical ratings from Rehder & Waldmann (2015). Error bars are 95% confidence intervals.

Fig. 2 presents a common effect network in which X is now a common effect of Y_1 and Y_2. For this network the independence relations are the reverse of those for a common cause network: Whereas Y_1 and Y_2 are unconditionally independent ($Y_1 \perp Y_2$) they become dependent conditioned on X. And, assuming that $Y_1 \rightarrow X$ and $Y_2 \rightarrow X$ are generative and operate independently then Y_1 and Y_2 exhibit an *explaining away* relationship when X is present (the presence of Y_1 makes Y_2 less likely and vice versa). These predictions—$D(Y) = 0$ and $D(Y|x^1) < 1$—are presented in the left chart of Fig. 2 assuming the same parameterization as in Fig. 1 (i.e., $p(y_i^1) = .50$, causal powers of .80, and $p(x^1|y_1^0 y_2^0) = .20$). And the right chart exhibits how those relations are typically violated: Rehder & Waldmann found that $D(Y) > 0$. As an aside, note that, as predicted, $D(Y|x^1)$ was significantly negative, although see Rottman & Hastie (2014) for evidence that explaining away is usually weaker than predicted by the normative model. The pattern of independence violations in Figs. 1 and 2 was also observed by Rehder (2014a) using a forced-choice task.

A number of rationalizations of these independence violations have been offered. For instance, Park & Sloman (2013) showed that those that arise with a common cause network are sometimes partly due to subjects' beliefs that the two causal links could be disabled by a common factor (also see Ali, Chater, & Oaksford, 2011; Lagnado & Sloman, 2004; Fernbach & Rehder, 2013; Mayrhofer & Waldmann, 2015; Rehder, 2014b; Walsh & Sloman, 2008). However, this account fails to explain the violations that occur with the common effect network in Fig. 2 (although it may explain why those violations were numerically about half the size of those in Fig. 1). Relatedly, Rehder & Burnett (2005) explained the large variety of Markov violations they observed by assuming that all variables were related by an underlying common cause (an assumption justified on the basis of the fact that the variables were features of a category; also see Rehder, 2014b). However, this account also fails to explain the results from Rehder & Waldmann (2015), which tested materials that were not category features.

To address these descriptive failures of the normative model, I present two new models of causal reasoning—the *leaky gate model* and the *dual prototype model*. For each I first show that it accounts for the independence violations shown in Figs. 1 and 2. Because the models make the same predictions for those simple networks, I describe two new causal networks for which their predictions diverge. An experimental test of those networks is then reported.

The Leaky Gate Model

The leaky gate model is based on the intuition that information flows along the directed edges of a causal network—in a manner that is reminiscent of "spreading activation" accounts of memory—even in situations where that flow is normatively blocked. It accomplishes this by introducing additional statistical structure associated with every triple of variables A, B, and C in the network that are related such that $A - B - C$ (where the edges represents a directed edge in either direction).

Y_1, X, and Y_2 are so related in the networks in Figs. 1 and 2. The joint distribution specified by the leaky gate model for those networks is derived from the joint specified by the normative model under parameterization θ, referred to as $p_N(X, Y_1, Y_2|\theta)$. The leaky gate model's joint distribution, p_{LG}, augments p_N with an additional *energy function*, $\epsilon_{XY_1Y_2}$ (Koller & Friedman, 2009),

$$log(p_{LG}(X, Y_1, Y_2|\theta, a)) \propto \epsilon_{XY_1Y_2} + log(p_N(X, Y_1, Y_2|\theta)) \quad (3)$$

$$\epsilon_{XY_1Y_2} = \begin{cases} a, & X = Y_1 = Y_2 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where a is a free parameter ≥ 0. $\epsilon_{XY_1Y_2}$ represents an expectation (whose magnitude is represented by a) that X, Y_1, and Y_2 will tend to have the same value, that is, to be all present or all absent. Eq. 3 yields a proper joint distribution after exponentiation and normalization.

The introduction of $\epsilon_{XY_1Y_2}$ is sufficient to reproduce the pattern of independence violations shown in Figs. 1 and 2. When $a = .75$ and the common cause network is instantiated with same parameters as in Fig. 1, the leaky gate model predicts $D(Y|x^1) = .75$ rather than 0. For the common effect network in Fig. 2, it predicts $D(Y) = 1.37$ rather than 0.

The leaky gate model generalizes to more complex networks. Consider the elaborated common cause network in Fig. 3 in which the two effects (Y_1 and Y_2) are themselves the causes of two other variables (Z_1 and Z_2). In this network, there are three triples of connected variables: $Y_1 - X - Y_2$, $X - Y_1 - Z_1$, and $X - Y_2 - Z_2$. Thus, the leaky gate model's joint distribution for this network is,

$$\begin{aligned} log(p_{LG}(X, Y_1, Y_2, Z_1, Z_2|\theta, a)) & \quad (5) \\ \propto \epsilon_{XY_1Y_2} + \epsilon_{XY_1Z_1} + \epsilon_{XY_2Z_2} & \\ + log(p_N(X, Y_1, Y_2, Z_1, Z_2|\theta)) & \end{aligned}$$

$$\epsilon_{XY_1Z_1} = \begin{cases} a, & X = Y_1 = Z_1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

$$\epsilon_{XY_2Z_2} = \begin{cases} a, & X = Y_2 = Z_2 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

and $\epsilon_{XY_1Y_2}$ is given by Eq. 4.

The predictions for the elaborated common cause network are shown in Fig. 3 for both the normative and leaky gate model and a number inference types (it also includes the predictions of the dual prototype model presented in the next section). The top row of charts presents predictions regarding the independence of the Ys—$D(Y)$ and $D(Y|x^1)$—but also those regarding the independence of the Zs—$D(Z)$ and $D(Z|x^1)$. The bottom charts compare $D(Z|x^1)$ with a new kind of inference, namely, $D(Z|x^1 y_1^1 y_2^1)$ in

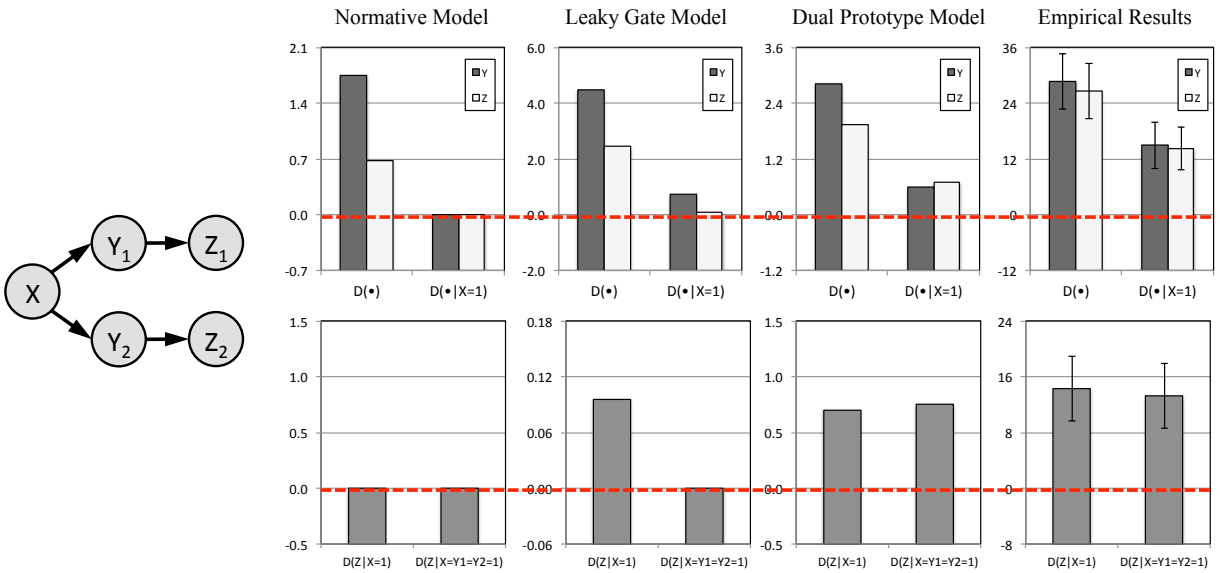


Figure 3. Predictions of the normative model, the leaky gate model, and the dual prototype model for the elaborated common cause network. The "•" in, say, D(•) represents either Y or Z. Empirical results are shown in the last column. Error bars are 95% confidence intervals.

which Z_1 and Z_2 are screened off from one another by *three* variables (X, Y_1, and Y_2) rather than one.

Using the same parameters as in Fig. 1, the normative model predicts that $D(Y) > D(Z)$. Because the causal relations are probabilistic, the Ys provide more information about each other than the Zs. Nevertheless, $D(Y|x^1) = D(Z|x^1) = 0$, because knowledge of X renders the Zs independent along with the Ys. In contrast, the leaky gate model ($a = .75$) predicts that the Ys and Zs are no longer independent. Nevertheless, the magnitude of the dependence between the Ys—$D(Y|x^1)$—is greater than that between the Zs—$D(Z|x^1)$ —which in turn is close to 0. This is so because the probabilistic links between the Ys and Zs further attenuates the flow of information.

Whereas the normative model predicts $D(Z|x^1) = D(Z|x^1y_1^1y_2^1) = 0$, the leaky gate model predicts $D(Z|x^1) > D(Z|x^1y_1^1y_2^1) \approx 0$. When the flow of information is blocked by three variables, virtually no information gets transmitted between the Zs even when the gates are leaky.

Eqs. 5-7 can also be applied to the elaborated common effect model in Fig. 4 in which the causes Y_1 and Y_2 are the effects of Z_1 and Z_2. That figure also shows that whereas the normative model predicts that $D(Y) = D(Z) = 0$ (i.e., the causes are unconditionally independent), the leaky gate model predicts instead that $D(Y) > D(Z) > 0$. It also predicts that explaining away will be weaker (i.e., $D(Y|x^1)$ and $D(Z|x^1)$ will be less negative) than in the normative model.

In summary, the key predictions of the leaky gate model are that $D(Y|x^1) > D(Z|x^1) \geq 0$ for the elaborated common cause model, $D(Y) > D(Z) > 0$ for the elaborated common effect model, and $D(Z|x^1y_1^1y_2^1) \approx 0$ for both.

The Dual Prototype Model

The dual prototype model is based on the intuition that causal inferences reflect a bias to expect that the variables in a network are either all present or all absent. Like the leaky gate model, the joint distribution it defines, p_{DP}, modifies the one specified by the normative model, p_N. For the simple common cause and common effect networks in Figs. 1 and 2, p_{DP} is the same p_{LG} and so reproduces the independence violations in Figs. 1 and 2 in the same manner. Differences between p_{DP} and p_{LG} arise for the networks in Figs. 3 and 4, however. In particular, Eq. 10 is analogous to Eq. 7 but elaborated with an alternative energy function,

$$log(p_{DP}(X, Y_1, Y_2, Z_1, Z_2|\theta, a)) \propto \epsilon_{XY_1Y_2Z_1Z_2} + log(p_N(X, Y_1, Y_2, Z_1, Z_2|\theta)) \quad (10)$$

$$\epsilon_{XY_1Y_2Z_1Z_2} = \begin{cases} a, & X = Y_1 = Y_2 = Z_1 = Z_2 \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

$\epsilon_{XY_1Y_2Z_1Z_2}$ represents the expectation that all five variables will tend to be all present or all absent.

Unlike the leaky gate model, the predictions of the dual prototype model are less sensitive to the distance between two variables in a network. For the extended common cause network (Fig. 3) it predicts a smaller difference between $D(Y)$ and $D(Z)$, that $D(Y|x^1) \approx D(Z|x^1)$, and that $D(Z|x^1y_1^1y_2^1)$ will be greater than 0 (and about equal to $D(Z|x^1)$). For the extended common effect network (Fig. 4) it predicts that $D(Y) = D(Z) > 0$ and $D(Z|x^1y_1^1y_2^1) > 0$. The prediction that $D(Z|x^1y_1^1y_2^1) > 0$ in both networks is especially notable because they imply the presence of independence violations between the Zs even when they are separated by multiple "blockers."

Overview of Experiment

These predictions are tested in an experiment in which subjects learned either the common cause network in Fig. 3 or the common effect network in Fig. 4 and then drew a number of causal inferences, including those necessary to

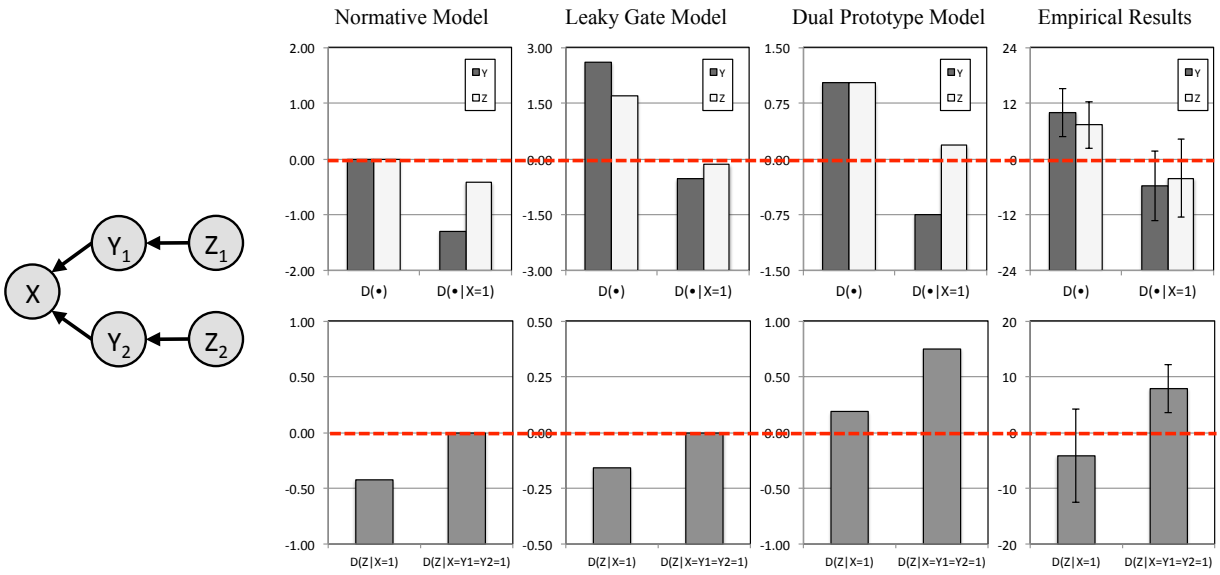


Figure 4. Predictions of the normative model, the leaky gate model, and the dual prototype model for the elaborated common effect network. The "•" in, say, D(•) represents either Y or Z. Empirical results are shown in the last column. Error bars are 95% confidence intervals.

compute the key quantities $D(Y), D(Z), D(Y|x^1), D(Z|x^1)$, and $D(Z|x^1y_1^1y_2^1)$. On the basis of previous research I expected that the independence relations stipulated by the normative model would be violated. The key question is whether those violations match the pattern predicted by the leaky gate model or the dual prototype model.

Method

Materials. Three domains were tested: economics, meteorology, and sociology. The five variables in each domain are shown in Table 1. In the database of materials each variable had the two values shown in Table 1 ("high" or "low" for interest rates) plus a third "normal" value. However, the variables were described as binary to subjects (e.g., interest rates were either high or normal) and the causal relations were described as obtaining between the non-normal values (e.g., low interest rates → small trade deficits). To control for any domain knowledge that subjects might have brought to the experiment, a between-subject factor controlled which variable states were described as causally related and took on the values ++++, ----, -+-+-, and +-+-+, where each +/- picks out the value in Table 1 for variables A, B, C, D, and E, respectively.

As a further safeguard against subjects' potential domain knowledge, a second between-subject factor controlled the assignment of these variables to the causal roles in Figs. 3 and 4. In the common cause condition, the roles of, X, Y_1, Y_2, Z_1, Z_2 were played by A, B, C, D, and E for half the subjects and by A, D, E, B, and C for the other half. In the common effect condition, they were played by D, B, C, A, and E for half the subjects and by D, A, E, B, and C for the other half. This scheme balances the assignment of the variables to the roles of Y and Z so that any differences between inferences involving Ys and those involving Zs cannot be attributed to the particular variables involved.

The description of each causal relation consisted of two sentences, one that stated that one variable caused another and a second that described the mechanism responsible for that relationship. Table 2 presents an example of one of the sets of causal relations that formed a common cause network in the domain of economics. Subjects also viewed a diagram of the causal relations and a third between-subject variable controlled which of four versions of that diagram was presented. The four versions of the common cause diagram are presented in Fig. 5 (names of the actual variables replaced X, Y_1, etc.). These layouts were chosen to ensure that the spatial distance between the two Ys on the screen was the same as between the Zs and that the Ys and Zs appeared an equal number of times in each quadrant of the screen. The four common effect diagrams were the same as those in Fig. 5 with the arrows reversed.

Table 1

Variable	Economics	Meteorology	Sociology
A	Interest rates (low+/high–)	Ozone levels (high+/low–)	Urbanization (high+/low–)
B	Trade deficits (small+/large–)	Air pressure (low+/high–)	Interest in religion (low+/high–)
C	Retirement savings (high+/low–)	Humidity (high+/low–)	Socio-economic mobility (high+/low–)
D	Job mobility (high+/low–)	Wind direction (up+/ down–)	Interest in sports
E	Income taxes (low+/high–)	Air temperature (low+/high–)	Commitment to the rule of law (strong+/weak–)

Procedure. Subjects first studied several screens of information about the domain and then performed the inference test. The initial three screens presented a cover story and a description of the domain's five variables and their

two values. A fourth screen presented the four causal links and a fifth the presented the diagram of those links. When ready, participants took a multiple-choice test of this knowledge. While taking the test, participants could return to the information screens they had studied; however, doing so obligated them to retake the test.

Table 2

Causal relation	Description
$X \rightarrow Y_1$	Low interest rates cause small trade deficits. The low cost of borrowing money leads businesses to invest in the latest manufacturing technologies, and the resulting low-cost products are exported around the world.
$X \rightarrow Y_2$	Low interest rates cause high retirement savings. Low interest rates stimulate economic growth, leading to greater prosperity overall, and allowing more money to be saved for retirement in particular.
$Y_1 \rightarrow Z_1$	Small trade deficits cause high job mobility. The intense demand for exports means that entrepreneurs are forming many new companies, and those new companies must hire workers away from existing companies.
$Y_2 \rightarrow Z_2$	High retirement savings causes low income taxes. When retirement savings are high, states are likely to lower income taxes to encourage spending and so stimulate the economy.

Subjects were then presented with a test that included those inferences needed to compute $D(Y)$, $D(Z)$, $D(Y|x^1)$, $D(Z|x^1)$, and $D(Z|x^1y_1^1y_2^1)$. Each question presented a particular economy (or society or weather system) whose variables were shown in a layout like one of those in Fig. 5. The values of variables whose state was known appeared in its corresponding box, boxes for variables whose values were unknown were blank, and the unknown variable whose value was being requested was filled with "?????." Subjects were asked "What's the probability that [this economy] has X?" where X was the to-be-inferred variable (e.g., high interest rates). Subjects entered their response by moving a tick on a rating scale whose ends were labeled "0%" and "100%". Each inference type was asked in both directions, e.g., subjects were asked both $p(y_1^1|y_2^1)$ and $p(y_2^1|y_1^1)$. The order of the 36 questions was randomized for each subject. Subjects could refer to a printed sheet of the causal diagram during the entire inference test.

Design and participants. The experiment consisted of a 2 (network type: common cause or common effect) × 3 (domain: economics, meteorology, or sociology) × 2 (assignment of variables to roles) between-subject design. 96 New York University undergraduates received course credit for participating and were randomly assigned to these 12 cells subject to the constraint that an equal number appeared in each cell. The four level factors that controlled which variable states were described as causally related and which version of the diagram was presented were both randomly assigned for each subject.

Results

Initial analyses revealed no effects of the between-subject counterbalancing variables and so the results are presented in Figs. 3 and 4 collapsed over those factors. The common cause and common effect conditions are reported separately.

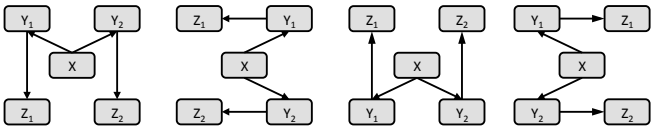


Figure 5

Common cause results. Consistent with previous research, the Ys were treated as dependent even when the state of their common cause X was known: Conditioned on x^1, one Y was rated about 15 points more probable when the other Y was present versus absent. One new finding is that the Zs were also treated as dependent and that the magnitude of those effects, $D(Y|x^1)$ and $D(Z|x^1)$, were about equal. Another is that the magnitude of the dependence between the Zs was the same regardless of whether they were separated by one or three blockers, $D(Z|x^1) \approx D(Z|x^1y_1^1y_2^1)$. These three differences scores were each significantly greater than zero, ps < .0001, and not significantly different from one another, ts (47) < 1. This pattern of results was predicted by the dual prototype model but not the leaky gate model (Fig. 3). One unexpected result is that whereas all three models predicted that $D(Y) > D(Z)$, the corresponding empirical judgments (29 and 26, respectively) did not differ from one another significantly, $t < 1$. Notably, it was the dual prototype model that predicted the smallest difference between $D(Y)$ and $D(Z)$.

Common effect results. Also consistent with previous findings, the causes of a common effect network were treated as unconditionally dependent: One Y was rated about 10 points more probable when the other Y was present versus absent, $p < .001$. Although the degree of dependence between the Zs—$D(Z)$—was a bit smaller (7 points), it was significantly different than 0, $p < .01$, and not significantly different than $D(Y)$, $t(47) = 1.27$, $p = .29$. In addition, $D(Z|x^1y_1^1y_2^1)$ was significantly greater than 0, $p < .001$. These results—$D(Y) \approx D(Z)$ and $D(Z|x^1y_1^1y_2^1) > 0$—were predicted by the dual prototype model but not the leaky gate model (Fig. 4). One unexpected finding is that $D(Y|x^1) \approx D(Z|x^1)$, $t < 1$, whereas the dual prototype model predicted $D(Z|x^1) > D(Y|x^1)$. Notably, although these measures of explaining away were negative, neither differed significantly from zero, ps > .14.

A model fitting exercise in which the three models were fit to the subjects' inferences corroborated the apparent superiority of the dual prototype model. Although space prohibits a full reporting of these results, I found that the common cause inferences were fit best by the dual prototype model ($AIC = 75.4$), followed by the normative (90.0), and leaky gate models (92.0). Common effect inferences were also fit best by the dual prototype model ($AIC = 63.4$), followed by the leaky gate (70.8) and normative models (77.1). The predictions of the fitted dual prototype model deviated from the empirical ratings by an average of 4.1 and 3.1 points (on a 0-100 scale) in the common cause and common effect conditions, respectively, and the correlations with those ratings were .96 and .98.

Discussion

Empirically, this article has both replicated and extended

previous findings. Once again, the two effects of a common cause network were not treated as independent conditioned on their cause. The new finding is that the same is true of the effects of those effects: The Zs were treated as dependent just like the Ys. The two causes of a common effect network (the Ys) were again not treated as unconditionally independent and, in a new result, so too were the causes of those causes (the Zs). Remarkably, in both networks the Zs were treated as dependent even when separated by multiple "blockers" ($D(Z|x^1y_1^1y_2^1) > 0$).

Theoretically, this article has advanced our understanding of such effects via the testing of two formal models. One might have expected that independence violations arise because information flows along a network in cases where it shouldn't. Yet this intuition—formalized as the leaky gate model—predicts that the magnitude of independence violations should decrease as the number of blockers between two variables increases. The fact that the common cause subjects treated the Zs as dependent to the same degree regardless of whether they were separated by three blockers or one provides striking evidence against "spreading activation" sorts of accounts.

Instead, reasoners behaved as if they overestimated the likelihood of two prototypical networks states, one in which all variables were present and another in which they were all absent, a view which explains the absence of an effect of either distance (Ys vs. Zs) or the number of blockers on the magnitude of independence violations. Because this tendency works in the opposite direction of explaining away in a common effect network, the dual prototype model also accounts for the lack of significant explaining away in this and past work (Rehder & Waldmann, 2015; Rehder, 2014a; Rottman & Hastie, 2013).

It is important to note that the dual prototype model does not imply that reasoners ignore the direction of causality in their causal inferences, because that model's joint distribution modifies rather than replaces the joint derived from the normative model. Common cause and common effect networks are theoretically important because they are identical ignoring the direction of causality. Yet in every experimental test of which I am aware (including this one) differences between those networks obtain (compare Figs. 3 and 4). The conclusion to be drawn from this work is not that people aren't sophisticated causal reasoners but rather that their understanding of the statistics implied by causal networks differs from that of network theorists.

One might be tempted to interpret the dual prototype model's success as reflecting subjects' beliefs that the causal links operated deterministically (and that the effect variables had no alternative causes), a situation that would normatively predict that variables were either all present or all absent (at least for the common cause network). Yet when these subjects were asked to infer X given 0, 1, or 2 Ys (i.e., $p(x^1|y_1^0, y_2^0)$, $p(x^1|y_i^0, y_j^1)$, or $p(x^1|y_1^1, y_2^1)$) their ratings (27, 57, and 82 in the common cause condition; 29, 64, and 97 in the common effect condition) reflected a belief that the causal links were probabilistic rather than deterministic (and the fitted causal power parameter was < .86 in both conditions). Nevertheless, an experimental test of the dual prototype model in which causal links are described as probabilistic is called for.

Cognitive theories based on causal graphical models have enjoyed great success accounting for causal-based judgments. Yet, the fact that people fail to honor the independence relations that such graphs express means that the theories rest on shaky foundations. It is time to move them onto on a firmer foundation, one that embodies fundamental empirical facts regarding how people draw causal inferences. Let the formalisms described herein represent modest proposals in that direction.

References

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition, 119*, 403-418.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Fernbach, P. & Rehder, B. (2012). Toward an effort reduction framework for causal inference. *Argument and Computation, 4, 1-25*

Koller, D. & Friedman, L. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lagnado, D.A. & Sloman S.A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*:856–76

Mayrhofer, R. & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science, 39*, 65-95.

Park, J. & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186-216.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Perales, J., Catena, A., & Maldonado, A. (2004). Inferring non-observed correlations from causal scenarios: The role of causal knowledge. *Learning and Motivation, 35*, 115–135.

Rehder, B. (2014a). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*, 54-107.

Rehder, B. (2014b). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of object categories. *Cognitive Psychology, 50*, 264-314.

Rehder, B., & Waldmann, M. (2015). *Failures of Explaining Away and Screening Off in Described versus Experienced Causal Learning Scenarios*. Submitted for publication.

Rottman, B., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109-139.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. New York: Springer-Verlag.

Walsh, C. R., & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. In G. H. Bower, et al. (Eds.), *Memory and Mind: A Festschrift for Gordon Bower* (pp. 345-358).

Curiosity-Driven Development of Tool Use Precursors: a Computational Model

Sébastien Forestier (sebastien.forestier@inria.fr)
Inria Bordeaux Sud-Ouest and Université de Bordeaux, France

Pierre-Yves Oudeyer (pierre-yves.oudeyer@inria.fr)
Inria Bordeaux Sud-Ouest and Ensta-ParisTech, France

Abstract

Studies of child development of tool use precursors show successive but overlapping phases of qualitatively different types of behaviours. We hypothesize that two mechanisms in particular play a role in the structuring of these phases: the intrinsic motivation to explore and the representation used to encode sensorimotor experience. Previous models showed how curiosity-driven learning mechanisms could allow the emergence of developmental trajectories. We build upon those models and present the HACOB (Hierarchical Active Curiosity-driven mOdel Babbling) architecture that actively chooses which sensorimotor model to train in a hierarchy of models representing the environmental structure. We study this architecture using a simulated robotic arm interacting with objects in a 2D environment. We show that overlapping phases of behaviours are autonomously emerging in hierarchical models using active model babbling. To our knowledge, this is the first model of curiosity-driven development of simple tool use and of the self-organization of overlapping phases of behaviours. In particular, our model explains why and how intrinsically motivated exploration of non-optimal methods to solve certain sensorimotor problems can be useful to discover how to solve other sensorimotor problems, in accordance with Siegler's overlapping waves theory, by scaffolding the learning of increasingly complex affordances in the environment.

Keywords: curiosity-driven learning; tool use; goal babbling; overlapping waves; developmental trajectory; HACOB model

Introduction

The understanding of tool use development in young children is one of the key question for the more general understanding of the ontogeny of human cognition. Indeed, a series of abilities are progressively developed from the simplest reaching movements of the arms through more dexterous manipulation of a spoon, towards advanced control of multiple interacting objects. The latter shows an understanding of shapes, forces and other physical properties that can be hierarchically recruited for mental transformations and planning operations which are pillars of human cognition. Child development has first been described as staircase-like successive stages in which all children go through (Piaget, Cook, & Norton, 1952). More recently, different views were developed and describe the structure and variability of observed children's developmental paths. In particular, the development of tool use precursors can be described as three consecutive and overlapping stages of behaviours where sequential learning and goal-directed behaviours play an increasing role (Guerin, Kruger, & Kraft, 2013): body babbling, behaviours with a single object, and behaviours with several interacting objects. A study of free play (Zelazo & Kearsley, 1980) shows that at $9\frac{1}{2}$ months play is mostly composed of tactile examination, waving or mouthing of a single object but simple relational acts of banging two objects together are already present. Later at $13\frac{1}{2}$ months, the study reveals that most children instead prefer to explore the relationships among objects, but still show behaviors of the previous phase. Furthermore, they show that this overlapping phases pattern averaged across children is also present in a longitudinal study of a single child. Another type of behavioral structure is described by Siegler as the variability in a child's set of current methods to solve a problem, which leads to the overlapping waves theory (Siegler, 1996).

In this paper we focus on the study of such progressions between phases of behaviours in a robotic model, and on the use of concurrent methods to solve a problem. We hypothesize that several mechanisms play a role in behavioural progression's structure and in particular 1) the intrinsic motivation to explore through a self-regulation of the growth of complexity of self-selected skills or tasks; 2) the structure of the representation used to encode sensorimotor experience.

Intrinsic motivation, sometimes called "curiosity", have been identified to play a fundamental role in driving spontaneous exploration in infant free play (Kidd & Hayden, 2015). They have been defined as mechanisms that push infants to explore activities for their own sake, driven by the search of novelty, surprise, dissonances or optimal challenge (Gottlieb, Oudeyer, Lopes, & Baranes, 2013). In the last decade, various families of computational models of intrinsic motivation were developed, often based on the formal frameworks of active learning and reinforcement learning (Baldassarre & Mirolli, 2013). One family of models, that has targeted to study the developmental dimensions of intrinsic motivation, has considered a curiosity-driven learning mechanism where the learner actively engages in sensorimotor activities that provide high learning progress, avoiding situations that are too easy or too difficult and progressively focusing on activities of increasing complexity (Gottlieb et al., 2013). Such computational models have shown that developmental trajectories could emerge from the curiosity-driven learning of sensorimotor mappings, in very different settings. In the Playground Experiment (Oudeyer, Kaplan, & Hafner, 2007), a quadruped robot motivated to maximize its learning progress acquired how to use its motor primitives to interact with the items of an infant play mat and a robot peer, following a self-organized learning curriculum. In (Baranes & Oudeyer, 2013), such mechanisms were shown to allow for efficient learning of large repertoires of skills involving high-dimensional continuous actions, as intrinsic motivation guided the system to explore sensorimotor problems of increasing complexity. In a model of active vocal develop-

ment (Moulin-Frier, Nguyen, & Oudeyer, 2014), an agent had to learn how to produce sounds with its vocal tract by self-exploration combined with imitation of adult speech sounds. This model reproduces accurately major phases of infant vocal development until 6 months. In both studies, developmental trajectories are emerging from learning, with both regularities in developmental steps and diversity. Regularities result from the attractor dynamics of the interaction between motivated learning, the body and the environment. The diversity comes from different mechanisms: stochasticity in the algorithms, variability in the environment, and the multiple attractors of the dynamic learning system. Existing models have considered the exploration and learning of sensorimotor correspondences mapping a motor space to a single task/sensory space. However, in the perspective of an open-ended development of reusable skills, and specifically in the development of tool use, multiple interdependent and hierarchically organized task spaces should be available to the agent. For instance, using a tool to act upon an object could make use of previously explored interaction with the tool. Consequently, an intrinsic motivation towards learning progress maximization could particularly be useful in the context of tool use where progress on some high-level task can not happen before progress on lower-level tasks have been made, by focusing training on currently learnable self-generated tasks.

We study aspects of those hypothesis leveraging previous models of curiosity-driven learning and extending them to the active exploration of hierarchical sensorimotor and task spaces. We define a hierarchy of sensorimotor models that structures the sensory space to reflect the interaction of the different items of the environment. The question of the autonomous learning of such a sensorimotor hierarchy is an important one but is not essential to test the questions asked in this paper, so here we provide the hierarchy to the agent as a prior. In this hierarchy of models to explore, different exploration choices are available to the agent: which model to explore, and how to explore that model. The problem of finding an efficient active choice strategy is an instance of strategic learning (Nguyen & Oudeyer, 2012), where different outcomes and strategies are available and the agent has to learn which strategies are useful for which outcomes. This can be viewed as a generalization of active learning methods in machine learning. We define the HACOB (Hierarchical Active Curiosity-driven mOdel Babbling) architecture and compare several possible strategies to study the role of active learning and hierarchical representation in the structuring of developmental trajectories. We compare the different learning conditions in a 2D environment where a simulated arm with three joints plus a gripper can grab one of two available tools to move an out-of-reach object, and we study the structure of behavioural phases during exploration.

To our knowledge, HACOB is the first model of the curiosity-driven development of tool use, and the first to show the autonomous emergence of overlapping phases in the development of simple tool use in a simulated robotic setup. Here we define tool use as the ability to perform different effects on an object with the help of an intermediate object, using some sort of learned inverse mapping. Our model is also the first to account for the intrinsically-motivated parallel exploration of different tools to reach one goal, in line with Siegler's overlapping waves theory. Other models predefine successive phases in object affordances learning (Ugur, Nagai, Sahin, & Oztop, 2015), or do not study the role of intrinsic motivation in tool affordances learning (Stoytchev, 2005), or have only considered the autonomous development of single object manipulation (Gottlieb et al., 2013).

However, here we do not study some important factors in the development of tool use. For instance, young infants need to adapt to the maturation of vision and to a developing body. Also, social guidance through imitation and mimicry is of central importance for the development of tool use but we do not address the question of its modeling in this paper nor of the interplay between social learning and self-exploration.

Methods

Environment

We simulate[1] a 2D robotic arm that can grasp tools that can be used to move an object into different boxes in the environment. In each trial, the agent executes a motor trajectory and gets the associated sensory feedback. Finally the arm, tools and objects are resetted to their initial state. The next sections precisely describe the items of the environment and their interactions. See Fig.1 for an example state of the environment.

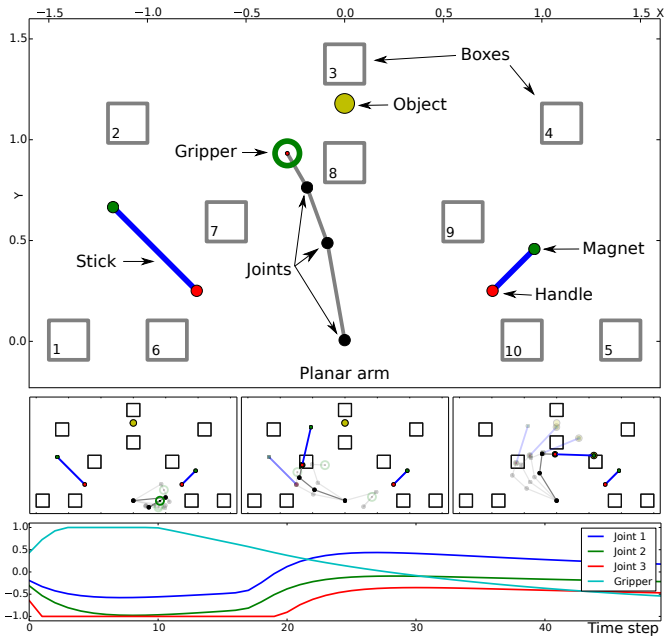


Figure 1: Top: a state of the environment. Middle: position of the arm at time steps 17, 33 and 50, with some intermediate positions, along the 50 steps movement. Bottom: trajectory of each of the four virtual motors, generated by a DMP.

[1] Source code and notebooks available as a Github repository at https://github.com/sebastien-forestier/CogSci2016

Robotic Arm The 2D robotic arm has 3 joints plus a gripper located at the end-effector. Each joint can rotate from $-\pi$ *rad* to π *rad* around its resting position, mapped to a standard interval of $[-1,1]$. The length of the 3 segments of the arm are 0.5, 0.3 and 0.2 so the length of the arm is 1 unit. The resting position of the arm is vertical with joints at 0 *rad* and its base is fixed at position $[0,0]$. The gripper g has 2 possible positions: *open* ($g \geq 0$) and *closed* ($g < 0$) and its resting position is *open* (with $g = 0$). The robotic arm has 4 degrees of freedom represented by a vector in $[-1,1]^4$. A trajectory of the arm will be represented as a sequence of such vectors.

Motor Control We use Dynamical Movement Primitives (Ijspeert, Nakanishi, Hoffmann, Pastor, & Schaal, 2013) to control the arm's movement as this framework permits the production of a diversity of arm's trajectories with few parameters. Each of the 4 arm's degrees-of-freedom (DOF) is controlled by a DMP starting at the rest position of the joint. Each DMP is parameterized by one weight on each of 2 basis functions and one weight specifying the end position of the movement. The weights are bounded in the interval $[-1,1]$ and allow each joint to fairly cover the interval $[-1,1]$ during the movement. Each DMP outputs a series of 50 positions that represents a sampling of the trajectory of one joint during the movement. The arm's movement is thus parameterized with 12 weights, represented by the motor space $M = [-1,1]^{12}$.

Objects and Tools Two sticks can be grasped by the handle side in order to catch an out-of-reach object. A small stick of length 0.3 is located at position $(0.75, 0.25)$ and a long stick of length 0.6 is located at position $(-0.75, 0.25)$ as in Fig. 1. An object (yellow ball), initially at position $(0, 1.2)$, can be caught by the magnetic side of one of the two sticks, moved and possibly placed into one of ten fixed squared boxes. If the gripper is closed near the handle of a stick (closer than 0.2), it is considered grasped and follows the gripper's position and the angle of the arm's last segment until the gripper opens. Similarly, if the magnetic side of a stick reaches the ball (within 0.1), the ball will then follow the magnet. The ten boxes (identified from 1 to 10) are static and have size 0.2. Boxes 1 to 5 can only be reached with the long stick, and the other five boxes can be reached with both sticks.

Sensory Feedback At the end of the movement, the robot gets sensory feedback from the different items of the environment (S, 25D). First, the trajectory of the gripper is represented as the x and y positions and the aperture (1 or -1) of the gripper at 3 time points: steps 17, 33, 50 during the movement of 50 steps (S_{Hand}, 9D). Similarly, the trajectories of the end points of the sticks are 3-point sequences of x and y positions (S_{Stick_1} and S_{Stick_2}, 6D each). It also gets the position of the single object at the end of the movement (S_{Object},2D). The agent receives the identifier (from 1 to 10) of the reached box if one of them has been reached by the ball, 0 otherwise. It also receives the distance between the ball at the end of the movement and the closest box (S_{Boxes}, 2D).

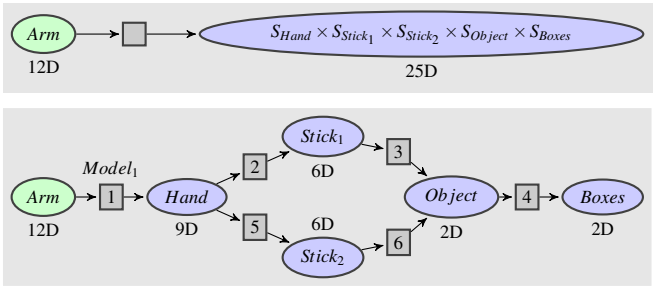


Figure 2: Architectures. Top: Flat. Bottom: Hierarchical.

Learning Architectures

The problem settings for the learning agent is to explore its sensorimotor space and collect data so as to discover how to produce a diversity of effects, and to learn repertoires of skills allowing to reproduce these effects in the form of inverse models. Consequently, the system is not given a priori a single target task to be solved: it rather autonomously selects the sensorimotor problems it will focus on through an intrinsically motivated selection of sensorimotor models.

Flat Architectures We define a flat architecture as directly mapping the motor space M (12D) and the sensory space S (25D). To do so, the agent needs a sensorimotor model that learns the mapping and provides inverse inference of a probable m to reach a given s. The sensorimotor model stores new information of the form (m,s) with $m \in M$ being the experimented motor parameters and $s \in S$ the associated sensory feedback. It computes the inverse inference with the nearest neighbour algorithm: it gets the motor part of the nearest neighbour in S of the given s, and adds exploration noise (Gaussian with $\sigma = 0.01$) to explore new motor parameters.

The agent also needs an interest model that chooses goals in the sensory space. The control condition is a random motor babbling condition (F-RmB) that always randomly chooses new motor parameters m. In the other conditions, the agent performs Goal Babbling, a method by which it self-generates goals in the sensory space and tries to reach them. To generate those goals, different strategies have been studied (Baranes & Oudeyer, 2013). It was shown that estimating the learning progress in different regions of the sensory space and generating the goals where the progress is high leads to a fast learning. However, this cannot be applied in a 25D sensory space as a learning progress signal cannot be estimated in this volume. Thus, in the flat random goal babbling condition (F-RGB), we use a random generation of goals in the sensory space, which was nevertheless proven to be highly efficient in complex sensorimotor spaces (Rolf, Steil, & Gienger, 2010).

Hierarchical Architectures The 25D sensory space can be structured to reflect the interaction of the different items of the environment. Indeed, the arm motor position influences the gripper, which influences one of the tools (but not both at the same time), which influences the position of the object and the filling of the boxes. We thus study here learning architectures that could make use of this sensorimotor structure, and

we call them hierarchical. Those architecture learn 6 models at the same time (see Fig. 2: gray squares are models). Each of those models functions in the same way as the random goal babbling flat architecture (F-RGB). Each model has a motor space (e.g. motor space of model 2 is S_{Hand}), a sensory space (respectively S_{Stick_1}, see arrows in Fig. 2), and can choose goals randomly in this sensory space. At each iteration, the architecture first has to choose the model in which to pick a goal, a procedure that we call Model Babbling. Once a model is chosen, it finds a goal in its sensory space, and infer motor parameters (that can be in the sensory space of a lower-level model) to reach that goal. Then, it passes those parameters as a goal to a lower-level model, which similarly infers motor parameters and passes those ones until the actual *Arm* motor space gets parameters to execute in the environment (with the same exploration noise as in Flat architectures). Model 4 has also to choose which lower-level model to use in order to reach an object end position s_o in S_{Object}, as two models (3 and 6) have S_{Object} as sensory space. Model 4 chooses the tool that enabled reaching s_o as close as possible in the past, e.g. if model 3 has in its history a reached sensory point s closer to s_o than any reached point with model 6, then model 3 is chosen. Finally, when motor parameters m are executed in the environment and feedback s is received, the mappings of all models are updated. However, only the tool-particular models are updated when a tool is currently held.

Random vs Active Model Babbling In a first condition, the agent randomly chooses the model that will find a goal, which is called Random Model Babbling (H-RMB). The problem of Model Babbling is an instance of strategic learning (Nguyen & Oudeyer, 2012), where different outcomes and strategies to learn them are available and the agent learns which strategies are useful for which outcomes. In that paper, they show that an active choice of the outcomes and strategies based on the learning progress on each of them increase learning efficiency compared to random choice. To develop active learning strategies, we first define a measure of learning progress for each of the 6 models. When a model has been chosen to babble, it draws a random goal s_g, and finds motor parameters m to reach it using the lower-level models. The actual outcome s in the sensory space of the model, associated to m might be very different from s_g as this goal might be unreachable, or because lower-level models are not mature enough for that goal. We define the competence associated to a goal s_g as the negative distance between the goal and the reached point, divided by the maximal distance in this space, to scale this measure across different spaces:

$$C(s_g) = -\frac{||s_g - s||}{max_{s'}||s' - s||} \quad (1)$$

and the interest $I(s_g)$ associated to this goal as

$$I(s_g) = |C(s_g) - mean_{kNN}C(s)| \quad (2)$$

where $mean_{kNN}C(s)$ is the mean competence of the ($k = 20$) nearest previous goals (k-Nearest Neighbours algorithm).

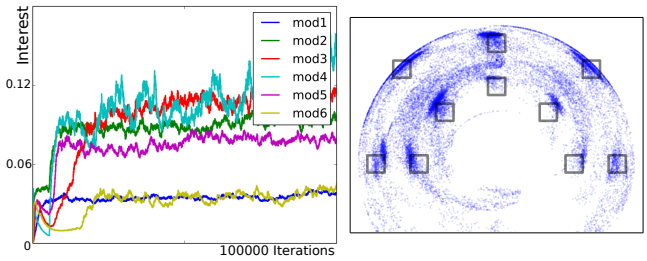


Figure 3: Condition H-P-AMB. Left: Interest of each model. Right: Exploration of the object space: each dot is the position reached with the object at the end of a movement.

The interest of a model is initialized at 0 and updated to follow the interest of the goals (with rate $n = 200$):

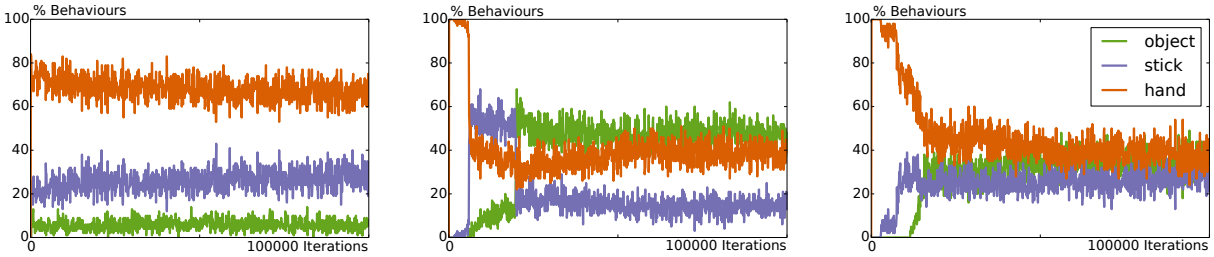
$$I_{model} = \frac{n-1}{n} I_{model} + \frac{1}{n} I(s_g) \quad (3)$$

In condition H-P-AMB, the choice of model is probabilistic and has $\varepsilon = 10\%$ chance to be random, and $(1 - \varepsilon)$ to be proportional to their interest. In condition H-GR-AMB, the choice of model is greedy (model with maximum interest) but also with $\varepsilon = 10\%$ of random choice. Finally, condition H-P-AMB-CTC (Curiosity-driven Tool Choice) is the same as H-P-AMB but the choice of the tool to use (model 3 or 6) is made with probabilities proportional to the interest of the two models, instead of being based on the more competent tool for the given object goal position. We call HACOB this Hierarchical Active Curiosity-driven mOdel Babbling algorithmic architecture with the algorithms H-P-AMB and H-P-AMB-CTC being two variants of the architecture.

Results

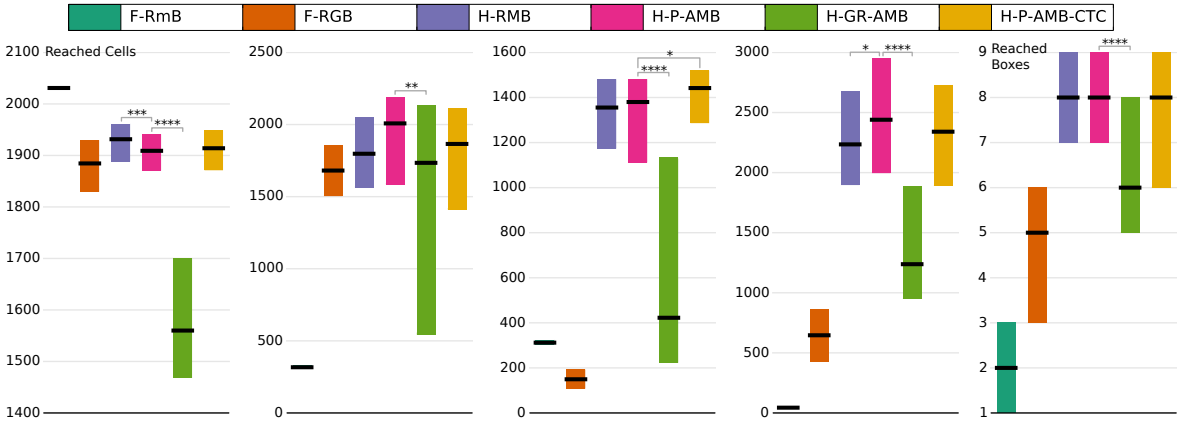
We perform 100 independent simulations of 100000 iterations per condition, starting with 100 iterations of motor babbling. Fig. 3 shows details about one trial of the condition H-P-AMB. We can see the interest of each model during one simulation, and the corresponding explored object space. The interests of models 2 and 5 increase once the arm succeeded to grab the corresponding stick. Following that, the interests of models 3 and 6 increase once the object has been reached.

Structure of the Evolution of Behaviours We provide a measure of three types of behaviours with objects during exploration. In the first category (*hand*) the arm did not grab any stick and thus did not move the out-of-reach object. In the second category (*stick*), the arm did grab one of the two sticks but did not touch the object with it. The third category (*object*) contains the movements where both a stick was grabbed and the object was moved by the stick. Fig. 4 shows a typical evolution of the proportion of the three categories of behaviours. We performed a more detailed analysis (see Table 1) by counting the trials where the evolution of the behaviours were similar to the ones found in infant development of the interaction with object (Guerin et al., 2013). A structure was considered similar to infant behavioral structures if



(a) No developmental structure (b) Developmental stages, abrupt changes (c) Overlapping phases structure

Figure 4: Typical behavioral evolution in the conditions (a) F-RGB, (b) H-GR-AMB, (c) H-P-AMB.



(a) *Hand* (b) $Stick_1$ (c) $Stick_2$ (d) *Object* (e) *Boxes*

Figure 5: Exploration of sensory spaces. Box plots show medians and quartiles of the 100 trials.

it validated each of the following criteria: behaviours of categories *stick* and *object* increase from 0 to more than 10% (potentially after an initial phase with a steady low value), are followed by a curve with small slope and no abrupt changes, and behaviours of category *object* start to raise at least 1000 iterations after *stick* started to raise (see Fig. 4(c) for a valid instance). Also, the median number of abrupt changes across trials are reported in Table 1 (as the sum of steady changes of more than 10% in the three behaviours), with a significant difference between condition H-GR-AMB and others (Mann-Whitney U tests, $p < 10^{-4}$).

Table 1: Behavioural analysis

Condition	Number of Trials validating criteria	Median number of Abrupt changes
F-RmB	0	0
F-RGB	0	1
H-RMB	60	2
H-P-AMB	70	2
H-GR-AMB	7	6
H-P-AMB-CTC	79	1

Exploration Efficiency Also, for each condition we measured the total exploration of the sensory spaces during training. The exploration of the hand, sticks and object spaces is defined as the number of reached cells in a 100×100 discretization of the (X,Y) space of their position at the end of the movement. Boxes' exploration is the number of boxes reached with the object during training. Fig. 5 shows the exploration of the different sensory spaces for each condition. We provide Mann-Whitney U test results of comparisons of total exploration for some pairs of conditions. One star means $p < 0.05$, two: $p < 10^{-2}$, three: $p < 10^{-3}$, four: $p < 10^{-4}$.

Structure of Tool Choice Finally, we compare the structure of tool choice made to reach object goal positions in two conditions for which only this choice differs. Fig. 6 shows the choice of tool to reach a given object goal position in the conditions H-P-AMB and H-P-AMB-CTC. When model 4 is babbling, it infers the best object position s_o to reach a random goal $s_b \in S_{Boxes}$. We plot all the choices that model 4 made during exploration, at position s_o on a 2D space, with color blue if $Stick_1$ was chosen and red if $Stick_2$ was chosen. In condition H-P-AMB, we can see strong boundaries between tool choice regions. By contrast, in condition H-P-AMB-CTC, both tools are chosen in all regions.

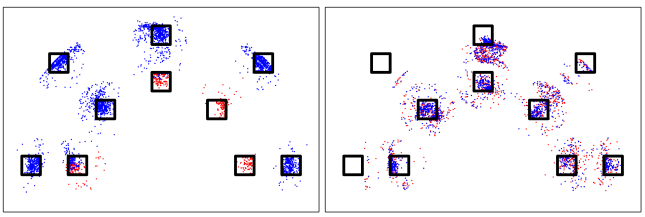


Figure 6: Chosen tool depending on object goal position. Blue: long stick choice. Red: small stick. Left: H-P-AMB, strong boundaries between tool choice regions. Right: H-P-AMB-CTC, parallel exploration of both tools in all regions.

Discussion

Structure of the Evolution of Behaviours Results show different structures of behaviour evolution in the different conditions. Flat architectures cannot efficiently learn in this environment with a high-dimensional sensory space. Therefore, they do not show structure in the behavioural evolution but rather steady proportions of the three behaviours. By contrast, hierarchical condition H-GR-AMB shows successive behavioural steps with abrupt changes, which reflects the greedy choice of model to babble. When one model becomes more interesting than another, it is chosen for a large number of iterations until another model exceeds its interest. Random model babbling shows overlapping phases structures more compatible with infants' studies in the evolution of the three behaviours, but less than active model babbling (60% instead of 70% or 79%). This is because random model babbling does not adapt its choice of models to their interests along development. Indeed, it often explores model 1 even if it is sufficiently explored to make progress on higher-level models, and so explores less the object position space than active model babbling (H-P-AMB). Also, all models are still useful to explore after the number of iterations simulated here so the first behavioural phases (*hand* and *stick*) do not lessen towards the end of the simulations in condition H-P-AMB.

A notable difference between active and random model babbling is on the cognitive or intentional level, as active model babbling monitors the current progress of each model whereas random model babbling does not. Furthermore, in other setups where some tasks are learned much faster than others and where at some point it becomes useless to explore a mastered task, active model babbling should also show an important difference both on a quantitative exploration point of view and on the structure of the evolution of behaviours.

Variability of Strategies to Reach Goals The comparison of conditions H-P-AMB and H-P-AMB-CTC shows that when the agent chooses the tool to reach a given object goal position based on the interest of the corresponding models, both tools are trained to reach all goals instead of training only the best performing tool. Indeed, with the active curiosity-driven choice of tool, the small stick has produced more diverse effects on the object than in the optimal tool's condition (Fig. 5c), even if those effects could also have been generated with the long tool. Curiosity-driven active tool choice shows more child-like results in accordance with Siegler's overlapping waves theory, which describes the use of strategies in infants and explains that non-optimal strategies are also explored as they might turn out to be finally good ones for this problem or for different but related ones.

Finally, we first demonstrated that hierarchical structure is a determining factor for the emergence of structured behavioural phases in our simple tool use setup. Then we showed that the active exploration of this hierarchical structure with curiosity-driven mechanisms combined with goal babbling reinforces this emergence and is essential to efficiently learn in this hierarchy. Although mechanisms such as action observation, sequential learning or causal inference are known to be highly important mechanisms in human development, we thus suggest that curiosity-driven exploration and goal babbling should also be considered as ones of them, but they have comparatively little been studied so far.

References

Baldassarre, G., & Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Springer.

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1).

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11).

Guerin, F., Kruger, N., & Kraft, D. (2013). A survey of the ontogeny of tool use: from sensorimotor experience to planning. *Autonomous Mental Development, IEEE Transactions on*, *5*(1).

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, *25*(2).

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P.-Y. (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in Psychology*, *4*.

Nguyen, S., & Oudeyer, P.-Y. (2012). Active choice of teachers, learning strategies and goals for a socially guided intrinsic motivation learner. *Paladyn*, *3*(3).

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, *11*(2).

Piaget, J., Cook, M., & Norton, W. (1952). *The origins of intelligence in children* (Vol. 8) (No. 5). International Universities Press New York.

Rolf, M., Steil, J., & Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *Autonomous Mental Development, IEEE Transactions on*, *2*(3).

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford University Press.

Stoytchev, A. (2005). Behavior-grounded representation of tool affordances. In *Proceedings of the 2005 ieee international conference on robotics and automation.*

Ugur, E., Nagai, Y., Sahin, E., & Oztop, E. (2015). Staged development of robot skills: Behavior formation, affordance learning and imitation with motionese. *IEEE Transactions on Autonomous Mental Development*, *7*(2).

Zelazo, P., & Kearsley, R. (1980). The emergence of functional play in infants: Evidence for a major cognitive transition. *Journal of Applied Developmental Psychology*, *1*(2).

Spatial Memory and Foraging: How Perfect Spatial Memory Improves Foraging Performance

Bryan Elvis Kerster (bkerster@ucmerced.edu)
Christopher T. Kello (ckello@ucmerced.edu
Cognitive and Information Science, 5200 N. Lake Road
Merced, CA 95343, USA

Abstract

Foraging is a search process common to all mobile organisms. Spatial memory can improve foraging efficiency and efficacy, and evidence indicates that many species—including humans—actively utilize spatial memory to aid in their foraging, yet most current models of foraging do not include spatial memory. In this study, a simple online foraging game was used to attempt to replicate and extend findings from a recent study (Kerster, Rhodes, & Kello, 2016) to further investigate the role of spatial memory in foraging. The game involved searching a simple 2d space by clicking the mouse to try and find as many resources as possible in 300 clicks. Spatial information was displayed that provided complete information about search history in order test how "perfect" spatial memory improves search performance. Over 1000 participants were recruited to participate in the task using Amazon's Mechanical Turk, which allowed this test to be performed across a wide parameter space of different resource distributions. Results replicated many of the findings of earlier studies, and demonstrated that spatial memory can have a dramatic effect on search performance.

Keywords: Foraging; spatial memory; Lévy walks; area restricted search; crowdsourcing

Introduction

Foraging is ubiquitous amongst living organisms, as it is a key task required for survival and procreation. Foraging is the process of searching an environment for resources, such as food or mates. While, foraging generally refers to a physical search process across a landscape, the principles involved are shared across many types of search processes, including memory search, visual search, and problem solving (Cain, Vul, Clark, & Mitroff, 2012; Rhodes, Kello, & Kerster, 2014; Rhodes & Turvey, 2007).

Some researchers have theorized that animal foraging behaviors are memory-less processes known as Lévy walks. Lévy walks are a random walk model where movement lengths are drawn from a Lévy distribution. It is unlikely that animal searches are literally random walks (Pyke, 2015), but the model captures an important aspect of foraging behavior. The distribution of path lengths, which are the lengths of straight movements made before stopping or switching directions, have been observed to follow a power law distribution. This indicates a clustered pattern of movement where a large number of smaller movements are interspersed with occasional larger movements across different scales. Lévy distributions have been observed in the foraging movements of a variety of different animals including albatrosses (Viswanathan et al., 1996), a variety of different fish species (Sims et al., 2008), and have been identified in memory search (Rhodes & Turvey, 2007) and visual search (Rhodes et al., 2014)..

Marginal value theory is an alternative approach to modeling foraging animal movements. This approach treats resources as a series of patches of varying sizes, and abstracts away the direct physical movement aspect of foraging while concentrating on optimizing the time a forager spends in a patch before moving on to another one. Marginal value theorem states that foraging can by optimized by comparing the rate of resource gain per unit time to the average rate. When the rate dips below the average, the forager should seek a new patch (Charnov, 1976).

One of the current leading models of animal foraging expands on marginal value theory by removing the spatial abstractions and implementing a model that attempts to optimize how an organism moves through the environment by relating turning rates to time since the last resource was found (Hills, 2006). Shortly after resources are found, the model turns more frequently keeping it within the resource patch. As resources become sparser, the model will turn less, propelling it on until it finds another resource and presumably another patch. This approach, known as *area-restricted search*, has recently been successful at modeling real world animal foraging data, and has been shown to generate movement patterns very similar to those generated by Lévy flight models (Hills, Kalff, & Wiener, 2013; Kareiva & Odell, 1987).

A number of studies have demonstrated that while there are some fundamental differences between animal foraging and cognitive search tasks, there are also enough similarities that findings related to foraging can be applicable to more abstract cognitive tasks. For example, Rhodes & Turvey (2007) demonstrated that times between word utterances in a category recall task follow a power law distribution as would be predicted by a Lévy walk model. Similar work has also been done from an optimal foraging perspective (Hills, Jones, & Todd, 2012). Visual search also provides a task whose constraints are similar to a foraging task, and unsurprisingly strong similarities in behavior between the tasks has been observed (Wolfe, 2013).

The successes of Lévy walks, marginal value theorem, and area-restricted search notwithstanding, spatial memory does not play a role in guiding search movements in any of these theories, i.e. there is no memory for past search locations. The exclusion of spatial memory stands in

contrast with studies of animal foraging—a number of animals have been observed to utilize spatial memory for such tasks including monkeys, primates, and octopi (Garber, 1989; Mather, 1991). The lack of spatial memory in leading models of foraging helps them to be simple and tractable to analyze, but it appears that organisms with well-developed spatial memories bring those faculties to bear when engaging in a task as inherently spatial as foraging. Additionally, if spatial memory proves important to human foraging it may be informative for theories of visual search as well.

Recent work has shown evidence that humans utilize spatial memory in their search strategies. Kerster, Rhodes, & Kello (2016) demonstrated a significant effect of landmarks in a virtual foraging task. In that task, participants were presented with either a blank screen or a Hubble space image and were asked to click anywhere on the screen to search for hidden resources. Resources were non-renewing and were not correlated to the background images. Analyses of performance showed that participants were able to find significantly more resources when the Hubble images were present. The authors concluded that the images served as landmarks for searchers, aiding their spatial memory. Kerster et al. (2016) formulated a foraging model that utilizes spatial memory as one of its key features. The model produced search patterns similar to those generated by human participants, in terms of the effects of resource density and clustering on performance and search trajectories. One important difference between human participants and the model was that, even though the relative patterns of performance were similar, the model was able to find many more resources on average. The authors conjectured that this difference may be due to the model utilizing perfect spatial memory.

In the present study, we expand on Kerster et al. (2016) by testing human foraging performance under experimental conditions that mimic perfect spatial memory, akin to the foraging model that they formulated. We test whether providing a complete history of a player's past search locations and resources found will improve performance to level of the model, and how search trajectories might change with perfect spatial memory.

Methods

The experimental design used here is based on the foraging game described in Kerster et al. (2016). 1034 participants played a browser-based game written in Adobe Flash. Participants were recruited using Amazon's Mechanical Turk and were paid $0.25 for their time and participation. All data was collected during a single 24 hour period.

Participants were shown a 1280 x 1024 pixel black screen with a score counter in the top left corner which displayed the number of resources found, and a "fuel" display in the top right which displayed how many clicks they had remaining in text and with a depleting meter.

Participants were instructed to find as many resources as they could in 300 clicks, and they were able to click freely anywhere on the screen and were given visual and audio cues when they selected a location with a resource. Resources were hidden from view until found and were non-renewing. See Figure 1 for an example of what the game looks like during play.

Each participant was presented a single trial from one of 9 different resource conditions. Resources conditions were manipulated in a 3x3 design between resource density and clustering. Resource density corresponded with the total amount of resources, and was set to 100, 600, or 1100 resources. Clustering of resources was controlled using a recursive algorithm that produces power law distributions of resource clusters with varying degrees of clustering from very clustered to uniform random. The parameter controlling degree of clustering was set to three distinct levels—highly clustered (0.1), less clustered (0.3), and random i.e. not clustered (0.5).

Prior to beginning of the foraging game, each participant was presented with a short text briefing that provided instructions and framed the experiment as "space exploration." This was followed by a display showing an example distribution of resources. The example was drawn from the same condition as what they would encounter in the game, but the particular distribution shown was different than the one used during the game. Participants were also informed that, if they received a high score, they would have an opportunity to add their initials to the high score list. A high score list was used to increase participant engagement in the game. Participants then had a short practice trial (15 foraging clicks) before the game began. The whole experiment generally took less than five minutes to complete.

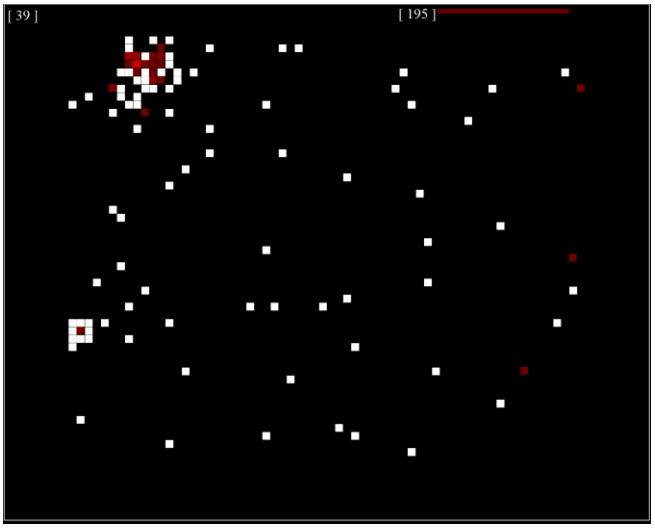


Figure 1: An example of the game's appearance during play. Squares indicate where the player has foraged thus far. White squares indicate nothing was found at that location, and red squares indicate that resources were found, with brighter shades indicating higher values. The number of resources found is shown in the upper left, and the number

of foraging clicks remaining is shown in the upper right (in numerical and bar form).

The foraging game in this experiment differed from the one used in Kerster et al. (2016) in two important ways. First, the search area was divided into an invisible grid of 15 x 15 pixel squares. Each square was searched by clicking over it, and each square could be searched only once. This restriction ensured that that there was no search area overlap among clicks, and therefore no wasted clicks searching in previously searched locations. By contrast, the game used by Kerster et al. (2016) allowed participants to click on locations that were partially or entirely searched already.

```
function distributeResources(rectangle, prob_split)
{
  if (rectangle.stars_remaining < 1) return
  else if (rectangle.size < 1 pixel) {
    place remaining stars at pixel
    return
  }
  else if (rectangle.stars_remaining == 1) {
    place star randomly in rectangle
    return
  }

  // alternate between vertical and horizontal splitting
  (rectangle1,rectangle2) = splitRectangle(rectangle,alternate)

  for each star {
    if (random_prob() < prob_split) star in rectangle1
    else star in rectangle2
  }

  // bias direction is randomized each split
  if (random_prob() < 0.5) prob_split = 1 - prob_split

  distributeResources(rectangle1, prob_split)
  distributeResources(rectangle2, prob_split)
}
```

Figure 2: Pseudocode for the recursive algorithm used to generate clustered resources. The probs_split parameter corresponds to a number between 0 and 0.5 and determines the degree of clustering, where 0 would create a single cluster, and 0.5 creates a uniform random distribution. (Kerster et al, 2016)

The other difference between Kerster et al. (2016) and the present study was the display of participant's previous searches. In the present study, a square was placed after each click on the corresponding location on the grid, and color coding was used to indicate the number of resources found, or no resources found. In particular, the range of brightness/hue levels was normalized by setting the brightest color to the maximum number of resources available in any given square, and setting white to mean that no resources were found. The restriction against overlapping clicks, plus the displayed information about previous locations foraged, effectively created a perfect external spatial memory for participants.

The game was otherwise the same as in Kerster et al. (2016), so we compare the results of the two experiments directly.

Results

Path length measures were computed using the same method described in Kerster, et al. (2016). Euclidean distances were calculated between each successive pair of clicks, and distances were summed into single path lengths for consecutive segments that had less than 10˚ of change between them. The angle threshold was used so that consecutive clicks in a relatively straight line were treated as a single path length. Each play yielded 203 path lengths on average, out of 300 clicks in total. By contrast, the average number of path lengths was 165 in Kerster et al. (2016). The increase in numbers of path lengths is likely the result of changes in foraging strategies between the two experiments (see below), as well as the transition to a fixed grid.

Path length distributions were analyzed to investigate whether they demonstrated heavy tailed properties. A number of studies have shown that the path length distributions of various foraging animals can be quantified in terms of the functional forms of their tails. Exponential, lognormal, and power law functions are commonly used, where the latter two functions feature heavier tails than exponential. Specifically, maximum likelihood methods suggested that foraging animals are sometimes best fit by power laws or truncated power laws (Humphries et al., 2010; Humphries, Weimerskirch, Queiroz, Southall, & Sims, 2012), and other times by lognormal or stretched exponential functions (Breed, Severns, & Edwards, 2015; Edwards, Freeman, Breed, & Jonsen, 2012). Additionally lognormal distributions have been fit to human foraging movements as well (Kerster, Kello, Rhodes, & Bien-Aime, 2013; Kerster et al., 2016). These heavy tailed distributions are indicative of spatial clustering in the search movements. Additionally, it has been argued that power law functions with exponents near two indicate Lévy walks, which are optimal under a certain set of assumptions (Viswanathan et al., 1999), although Lévy walks have been criticized as being overly simplified models of foraging by others (Pyke, 2015).

Aikaike's Information Criterion (Akaike, 1974) was applied to the path length distributions provided by each subject using the same methods described in Kerster et al. (2016) to determine which function best fit each distribution. Like in Kerster et al. (2016) the vast majority of distributions were best fit by the lognormal (92%). The estimated exponents also replicated, and were near the theoretically optimal exponent of 2 for Lévy walks. These findings indicate that people produced search trajectories that were inherently clustered.

Table 1: Comparison of resources found and estimated exponents across conditions between the two experiments. Standard errors are in parentheses.

This experiment replicates many of the key findings of the earlier foraging game, which then raises the question: What effect did changes in the experiment have? The model described in Kerster et al. (2016) also replicated many of these findings, but demonstrated much higher search efficiency as measured by their normalized score, i.e. proportion of available resources found. Kerster et al. proposed that the difference in score between human participants and the model was due to the model having a perfect spatial memory, and non-overlapping foraging locations. The present experiment conferred the same benefits upon human participants, so we expected a large increase in score between this experiment and Kerster et al. (2016).

As predicted, participants demonstrate much higher scores in the current experiment (M=16%, SD=19%) compared to the previous experiment (M=7%, SD=8%) as confirmed by a Welch's t-test ($t(7142)=15.5$, $p < 0.0001$). A qualitative look shows that scores in all conditions except the uniform random distribution of resources were substantially higher (see Table 1). Score is normalized by dividing the total number of resources found by the number of available resources in the condition, so that scores may be directly compared across distributions. The small difference observed for the random condition is because knowledge of prior resource locations provides no information about where to find other resources, thus spatial memory should only be useful in preventing repeated search of the same

	% found		Est exponent	
Density	Current Experiment	Kerster et al., 2016	Current Experiment	Kerster et al. 2016
Sparse	22.8% (0.87)	9.9% (0.81)	1.93 (0.11)	1.84 (0.16)
Medium	28.2% (0.80)	11.3% (0.90)	1.97 (0.12)	1.95 (0.19)
Dense	32.6% (0.75)	13.2% (0.88)	2.06 (0.15)	2.00 (0.16)
Clustering				
Most	55.6% (0.20)	21.7% (0.57)	1.95 (0.10)	1.97 (0.16)
Less	15.9% (0.28)	7.4% (0.23)	2.04 (0.14)	2.00 (0.17)
Random	7.0% (0.14)	5.5% (0.17)	1.98 (0.14)	1.81 (0.18)

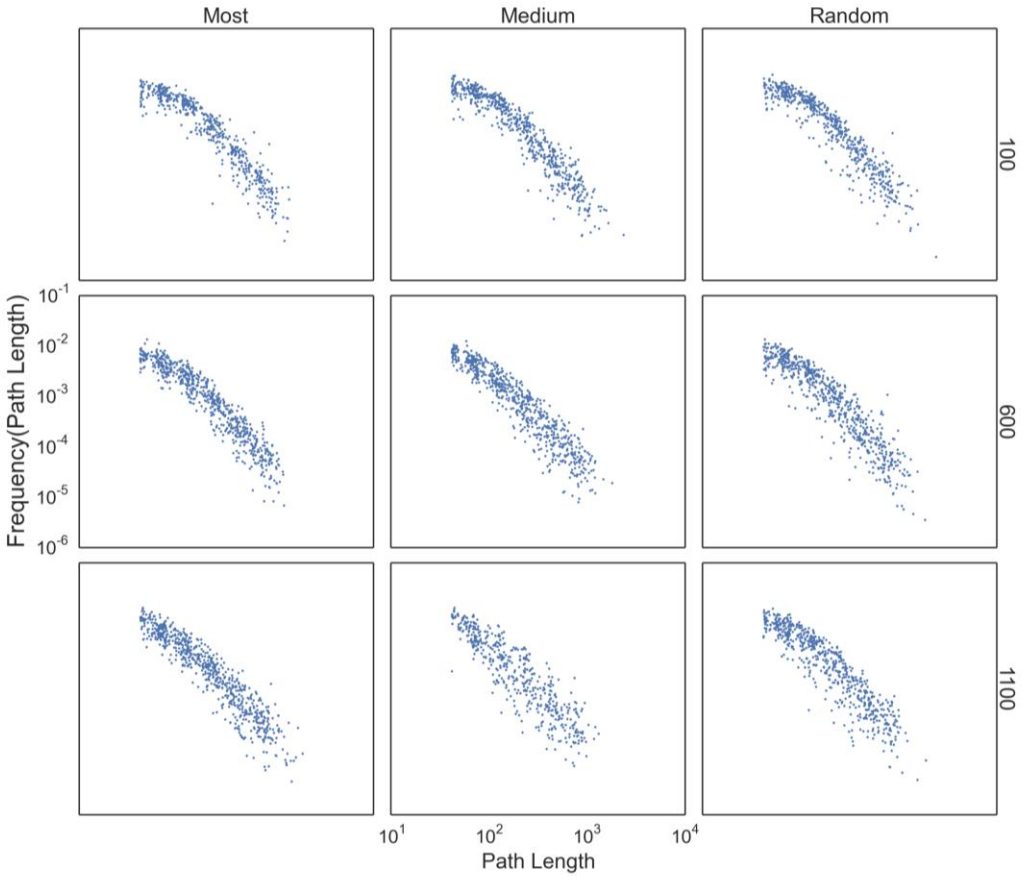


Figure 3: Path length distributions for each trial. A histogram with 10 logarithmically spaced bins was created for each trial. All trials are plotted together for each condition, separated by density and clustering. Axis labels are only displayed once, but each plot has the same axes.

location.

As mentioned earlier, two key changes were made between this experiment and the one reported in Kerster et al, 2016. Without further controls it is difficult to precisely determine how much of the increases in score were due to improvements in spatial memory, and how much were due to increased search efficiency because search locations could no longer overlap. Nevertheless, we developed a simple method to control for search efficiency by scaling normalized score by the number of unique pixels visited. This allows for the relative comparison of scores accounting for inefficiency due to repeatedly searching the same area. This analysis revealed that, for each condition except the random conditions, the majority of observed performance increases (about 70-80%) were due to improvements in spatial memory (see Table 2).

The observed increases in performance raise the question of how search trajectories changed in the present experiment due to perfect externalized spatial memory. To address this question, we used spatial Allan Factor analysis to test whether trajectories were more or less clustered in the present study compared with Kerster et al. (2016). Allan Factor analysis is designed to measure nested clustering in point processes (Allan, 1966), and was used to measure clustering in previous search tasks (Kerster et al., 2016; Rhodes, Kello, & Kerster, 2014).

Table 2: Relative increases in scores accounted for by search efficiency (lack of overlap) and perfect spatial memory

	% increase due to search efficiency	**% increase due to perfect memory**
Density		
Sparse	23.2%	76.8%
Medium	24.5%	75.5%
Dense	26.5%	73.5%
Clustering		
Most	20.3%	79.7%
Less	30.7%	69.3%
Random	77.9%	22.1%

Allan Factor analysis works by tiling the space with squares of side length L, and counting the number of points N (in this case foraged locations) within each square i. The differences between adjacent squares are averaged together, and then normalized by twice the mean.

The Allan factor statistic A(L) provides a measure of spatial variance across a number of given scales L. If foraged locations are randomly distributed then $A(L) \sim 1$ for all L. If locations are clustered across scales then $A(L) > 1$ and increases with L. If the clusters are hierarchically nested across scales then $A(L) \sim L^{\alpha}$ where $\alpha > 0$. This can be expressed as a linear relationship in log-log coordinates.

Allan Factor values A(L) were regressed across scales (L) on each distribution in log-log coordinates to produce a slope value. These values were then compared between the two experiments. Slope values are significantly higher (Welch's two-sided $T(7142)=10.7$, $p < 0.0001$), in the current experiment ($M=1.12$, $SD=0.74$) than in the distributions produced in Kerster et al. (2016) ($M=0.86$, $SD=0.50$) (see figure 4). This indicates that perfect spatial memory led to an increase in the clustering of their searches across scales.

Discussion

The current study is a direct extension of recent work using an online foraging game to explore human foraging behavior while controlling for variables that would be very difficult to control using other methods. The previous study by Kerster et al. (2016) provided evidence that spatial memory is a key feature of human foraging, and the current study confirms and extends those findings. The model used in that study uses perfect memory as a simplification as opposed to trying to implement a particular more realistic type of memory. The manner in which we give perfect to participants is also not meant to mimic any particular theory of spatial memory, but to instead explore how performance is constrained by limitations in memory. Our results demonstrated the magnitude by which foraging performance can be increased when spatial memory is improved. As long as available resources are not randomly located in the environment, search performance can be effectively doubled by providing perfect spatial memory.

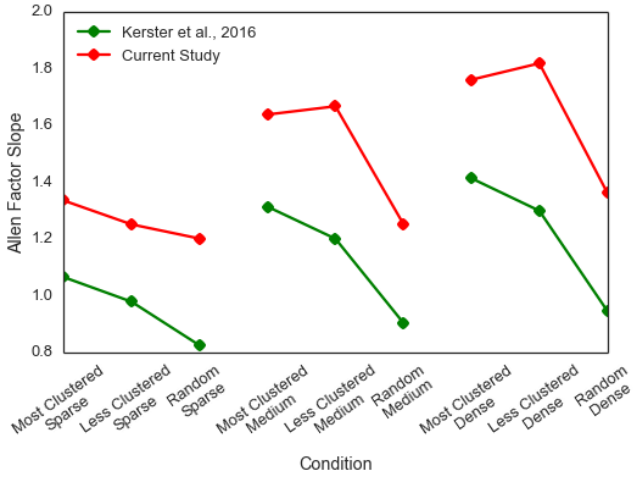


Figure 4: Comparison of Allan Factor slopes across conditions between the two experiments. Both experiments demonstrate similar patterns across condition, but the current study has higher values in all conditions, indicating more nested clustering in foraging movements.

The current study also replicated a number of important findings in the foraging literature. Notably participants generated clustered path length distributions that resemble Lévy walks. This is consistent with observations of foraging animals (Sims et al., 2008), as well as cognitive foraging tasks (Rhodes et al., 2014; Rhodes & Turvey, 2007).

In fact, Allan Factor analyses revealed that participants with perfect spatial memory showed significantly greater clustering across scales in their movements. Improved spatial memory allowed searchers to exploit information about the environment better as they uncovered it through searching. The increased clustering we observed in their movements is likely a result of improved exploitation of the clustering in the environment.

Interestingly, the increases in clustering occurred both as the resource environment became clustered, and as the density increased. The interaction effect observed here is somewhat different from that observed in Kerster, et al. (2016) and may be the result of some factors specific to the task. Further work is necessary to understand exactly how resource density, clustering, and movement constraints interact.

The simple foraging task used here is designed to have constraints in common with a variety of foraging and search tasks, and the findings presented here could be applied to many of these tasks. Some of the more interesting ramifications lie in cognitive tasks, such as visual and memory searches. Both the natural world and our memories are non-randomly distributed, which implies that knowledge of previously searched locations could be a crucial factor in efficient cognitive search.

Acknowledgments

This research was funded in part by a grant from the National Science Foundation, BCS 1031903.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716.

Breed, G. A., Severns, P. M., & Edwards, A. M. (2015). Apparent power-law distributions in animal movements can arise from intraspecific interactions. *Journal of The Royal Society Interface*, *12*(103), 20140927.

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian Optimal Foraging Model of Human Visual Search. *Psychological Science*, (August).

Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical population biology*, *9*(2), 129–136.

Edwards, A. M., Freeman, M. P., Breed, G. A., & Jonsen, I. D. (2012). Incorrect Likelihood Methods Were Used to Infer Scaling Laws of Marine Predator Search Behaviour. *PLoS ONE*, *7*(10).

Garber, P. A. (1989). Role of spatial memory in primate foraging patterns: Saguinus mystax and Saguinus fuscicollis. *American Journal of Primatology*, *19*(4), 203–216.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive science*, *30*(1), 3–41.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal Foraging in Semantic Memory. *Psychological review*.

Hills, T. T., Kalff, C., & Wiener, J. M. (2013). Levy Processes and Area-Restricted Search in Human Foraging. *PLoS ONE*, *8*(4), 1–7.

Humphries, N. E., Queiroz, N., Dyer, J. R. M., Pade, N. G., Musyl, M. K., Schaefer, K. M., … Sims, D. W. (2010). Environmental context explains Lévy and Brownian movement patterns of marine predators. *Nature*, *465*(7301), 1066–9.

Humphries, N. E., Weimerskirch, H., Queiroz, N., Southall, E. J., & Sims, D. W. (2012). Foraging success of biological Lévy flights recorded in situ. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(19), 7169–74.

Kareiva, P., & Odell, G. (1987). Swarms of predators exhibit“ preytaxis” if individual predators use area-restricted search. *American Naturalist*, *130*(2), 233–270.

Kerster, B. E., Kello, C. T., Rhodes, T., & Bien-Aime, R. J. (2013). Adaptive Foraging : Effects of Resource Conditions on Search Paths in a Web-Based Foraging Game. *The Annual Meeting of the Cognitive Science Society*.

Kerster, B. E., Rhodes, T., & Kello, C. T. (2016). Spatial memory in foraging games. *Cognition*, *148*, 85–96.

Mather, J. a. (1991). Navigation by spatial memory and use of visual landmarks in octopuses. *Journal of Comparative Physiology A*, *168*(4), 491–497.

Pyke, G. H. (2015). Understanding movements of organisms: It’s time to abandon the Lévy foraging hypothesis. *Methods in Ecology and Evolution*, *6*(1), 1–16.

Rhodes, T., Kello, C. T., & Kerster, B. (2014). Intrinsic and extrinsic contributions to heavy tails in visual foraging. *Visual Cognition*, *22*(6).

Rhodes, T., & Turvey, M. (2007). Human memory retrieval as Lévy foraging. *Physica A: Statistical Mechanics and its Applications*, *385*(1), 255–260.

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J. a, Pitchford, J. W., … Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour. *Nature*, *451*(7182), 1098–102.

Viswanathan, G., Afanasyev, V., Buldyrev, S., Murphy, E., Prince, P., & Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413–415.

Viswanathan, G., Buldyrev, S. V., Havlin, S., Luz, M. G. E. da, Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911.

Wolfe, J. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *Journal of Vision*, *13*, 1–17.

Temporal Expressions in Speech and Gesture

İdil Bostan (idbostan@ku.edu.tr)
Ahmet Börütecene (aborutecene13@ku.edu.tr)
Oğuzhan Özcan (oozcan@ku.edu.tr)
Koç University Arçelik Research Center for Creative Industries (KUAR), Koç University
Rumelifeneri Yolu Sariyer 34450
Istanbul – TURKEY

Tilbe Göksun (tgoksun@ku.edu.tr)
Department of Psychology, Koç University
Rumelifeneri Yolu Sariyer 34450
Istanbul - TURKEY

Abstract

People use spatial metaphors to talk about temporal concepts. They also gesture frequently during speech. The characteristics of these gestures give information regarding the mental timelines people form to experience time. The present study investigates the expression of temporal concepts on a natural setting with Turkish speakers. We found that Turkish speakers used more metaphoric temporal phrases (e.g., *short period, time flies quickly*) than words referring to time without spatial content (e.g., *today, nowadays*) in a session where they talked about people's fortune. Spontaneous gestures were mainly classified as metaphoric and beat gestures and were mostly produced on the sagittal axis, which contradicts with the previous findings. Yet, we also found that people used vertical axis to represent current and future events. These findings suggest that lateral axis may not always be the most common direction for co-speech temporal gesture use, and the pragmatic constraints of the environment may influence the spatial conceptualization of time.

Keywords: time, spatial metaphors, temporal gestures, Turkish

Introduction

Time is an abstract concept that we cannot physically experience through our senses. The need for a concrete ground to perceive and express time has led to the use of spatial metaphors. *Space* is easier to experience in physical reality, so people usually map their understanding of time onto space (Boroditsky, 2001). People use spatial metaphors to talk about time (e.g., Clark 1973; Evans 2004; Lakoff, 1992). For example, it is common for English speakers to say "move the meeting *forward*" or "this week has been *long.*" Recent research also suggests that people produce spontaneous gestures to depict temporal information in their gestures (e.g., Casasanto & Jasmin, 2012; Cooperrider & Nunez, 2009). Most of the findings stem from experimental paradigms and many studies use native English speakers as participants. In this study, we investigate native Turkish speakers' use of temporal information in speech and gesture in a natural setting.

Most Western cultures locate future ahead and past in the back. Other cultures use different variations of time orientation. In Mandarin, for example, even though there is a use of the front-back analogy, people also produce *up* for future and *down* for past (Boroditsky, 2001). In Aymara language of the Southern America, *front* represents past and *back* represents future (Nunez & Sweetser, 2006). There are many possible explanations to why and how people locate temporal features. One important factor to represent time is the reading and writing direction of speaker's native language (Bergen & Lau, 2012). Most languages perceive earlier events at the left or back of the self and future events at the right or front of the self. Using nonlinguistic tasks, Ouellet and colleagues (2010) found that Hebrew speakers demonstrated the opposite pattern of right to left flow of time in line with their right to left reading-writing direction. Spanish speakers have the opposite representation of time, not only in visual tasks, but also in auditory tasks. Similarly, Fuhrman and Boroditsky (2010) showed that when asked to temporally order pictures, English speakers (left to right) put earlier events to the left side of space, whereas Hebrew speakers (right to left) put the same events to the right side of space (see also Boroditsky, Fuhrman & McCormick, 2011). These findings present the impact of reading-writing direction on other modalities. Thus, in nonlinguistic tasks, the representation of time seems to be influenced by one's native language's characteristics.

People's representation of time based on space can also be apparent in their expressions of time in both speech and gesture (Casasanto & Jasmin, 2012). That is, people talk about time with regard to space in spatial metaphors. In English, these metaphors can be in a deictic form with a reference to a certain time point (e.g., it happened *way back* in the past), or may not include any directionality information at all (e.g., a *long time* has passed). The front-back analogy is the most dominant spatial axis in English language, with a few exceptions of other directions (e.g., the meeting is coming *up*). However, this kind of use is not systematic (Casasanto & Jasmin, 2012).

Do gestural expressions follow verbal temporal information? Speech and gesture form an integrated language system (McNeill, 1992). As a result, gesture is

often used as another source to understand the underlying mechanisms of cognitive functions. Gestures are also spatial in nature and can be suitable to embody the content of temporal speech. Gestures can be classified into four main categories: *iconic* gestures that represent actual objects or actions, *metaphoric* gestures that refer to abstract ideas or concepts, *deictic* gestures that involve pointing to an object or location, and *beat* gestures, which are quick hand movements to supplement information in speech (McNeill, 1992). If speech and gesture are temporally and semantically integrated, then the descriptions of temporal relations should be similar in each modality.

Cooperrider and Nunez (2009) suggest that English speakers' use of gestures is very predictable based on the writing direction and their cultural conceptualizations of time. Previous studies suggest that English speakers gesture both on the left-right and front-back axes depending on the task (Casasanto & Jasmin, 2012). When asked to gesture deliberately, deictic metaphors invoke a sagittal mental timeline, whereas sequential metaphors prime speakers to a lateral timeline. English speakers refer to the sagittal axis (from back to front) rather than lateral axis (from left to right) while speaking metaphorically about time. Yet, they dominantly use lateral axis in their spontaneous co-speech gestures about time (Casasanto & Jasmin, 2012). This dissociation points to an implicit placement of sequential events on the lateral axis.

Turkish is another left to right flowing language based on the reading-writing direction. Turkish speakers also refer to time in spatial metaphors and represent it as a physical entity (Ördem, 2014). Phrases like *O günler arkada kaldı* ("those days are *behind* us") show Turkish speakers' use of sagittal axis while referring to time. Turkish and English speakers' expressions of time display parallels in the conceptualization of time as a spatial motion or a location in space. Additionally, the developmental stages that lead to an adult understanding of time are similar between two language groups when their linguistic comprehension of metaphors regarding time is measured (Özçalışkan, 2004).

Even though there is accumulating evidence on how people talk and gesture about time, detailed investigations of temporal expressions in both speech and gesture in languages other than English are still insufficient and there is room for new findings. In languages, several categories such as duration, tense, sequence time, deictic time, and metaphors are used to define time (Le Guen & Balam, 2012). This paper focuses on two time concepts: duration and spatial metaphors to express time.

To understand the mapping between specific temporal expressions and their gestural construals, in this study we investigate the types and axes of the gestures people perform while referring to the past, present, and future as well as different types of temporal expressions. In a naturalistic study, we provide a comprehensive examination of speech and gesture relations by focusing on both time words without a spatial content (e.g., *today, tomorrow*) and spatial metaphors (e.g., *before, short time*) and providing both quantitative and qualitative analyses. We predict that due to the abstract nature of time concept, regardless of the type of phrase, people would produce more metaphoric gestures than any other types. If individuals' gestural expressions directly map onto verbal expressions, gestures would follow the timeline depicted in speech.

Method

Participants

Eighteen native Turkish speakers participated in the study. They were all undergraduate or graduate students (M_{age} = 22.4, *SD*= 2.1, 15 females). These individuals were the speakers (see more information below). In addition, two individuals participated as listeners.

Materials and Procedure

This study used naturalistic fortune telling sessions as the design. Turkish fortune telling sessions provide a viable setting to assess temporal information. In Turkey, coffee fortune-telling is a widespread practice interwoven in daily life and considered by most people as an occasion for creating social interaction. Moreover, Turkish coffee drinking and fortune telling sessions create an enjoyable, warm and sometimes humorous social environment. In Turkish culture, it is very common for people to tell others' fortune by examining the coffee cup after drinking coffee. Both the fortuneteller and the listener (the person who drank the coffee) are engaged and attentive in this interaction (see Figure 1). The fortunetellers' narratives involve rich verbal and non-verbal information. Furthermore, the contents of fortune telling involve the past, present and future of the fortune receivers' lives, which provide abundant natural temporal data to be examined. That is, fortune telling is a good setting for us to examine temporal speech and gesture interaction.

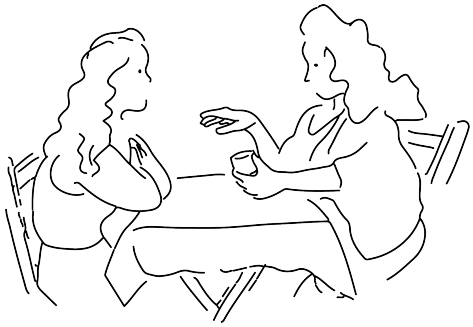


Figure 1: The setting of a coffee cup fortune-telling session

Fortunetellers (speakers in this study) were recruited among graduate and undergraduate students around the university campus who were passionate about fortune telling and performed it occasionally. Listeners of the fortunes were two female students who were willing to hear their fortunes. The listeners were blind to the research questions. In the sessions the fortuneteller, the listener, and the

researcher were present. To maintain an environment of natural conversation, sessions were held in coffee houses around the campus and the city. Researcher did not intervene in the process and merely recorded the session. Fortuneteller and the listener sat face to face, on opposite sides of a table. Their hands were free to move and they were able to hold or let go of the coffee cup whenever they wished to, as would be in a regular coffee drinking setting. Overall, the physical setting of the experiment resembled any two people sitting and chatting in a coffeehouse, with the sole addition of a camera recording the interaction. There was no limit to the duration of the session, but most of them averaged around 10-15 minutes. Before each session the fortunetellers were asked to complete an informed consent form and a questionnaire with demographic questions.

The sessions were video recorded with a smartphone camera. The camera was positioned to frame the hands, arms and faces of both participants to record the body movements. A small tripod was used to capture a stable footage. No instructions were given regarding gesture use.

Coding

Speech. All speech was transcribed by native Turkish speakers. First, phrases that involved temporal sentential meaning were identified such as *bu ara çok şanslı olacaksın* ("you are going to be very lucky these days"). Then, temporal phrases were coded into two levels of meaning: Literal temporal phrases (LP) or metaphoric temporal phrases (MP). LPs were the ones that refer to time and did not involve any spatial content, such as *şimdi* ("now"). MPs had some metaphoric meaning in them, expressed by the spatial words, such as *kısa süre* ("short period"). This was a delicate process since many spatial metaphors of time have been embedded so deeply in Turkish language that they may seem like literal at first glance. For example the word *önce (before)* comes from the root *ön,* which means *front*, a spatial concept (Gentner, Imai & Boroditsky, 2002). Although *önce* seems to be a literal phrase, this little distinction makes it a subtle metaphorical one. Similarly *bir yıl içinde* (in one year) has a spatial component based on the preposition *içinde (in)*. To overcome this problem, we first categorized phrases into three categories as Literal, Metaphorical and Subtle-metaphorical. Yet, after analyzing all of them separately, we found the two metaphorical groups yielded similar results in people's gesturing patterns so we reported them as one MP category. Temporal phrases were also coded according to their temporal orientation of the sentence: phrases regarding past (e.g. *öncesi* "before"*)*, present (e.g. *bugünlerde* "nowadays"*)*, or future (e.g. *gelecekte* "in the future").

Gesture. The gestures that accompanied temporal phrases were identified. A change in the shape of the hand or motion signaled the end of a gesture. Gestures were first categorized according to their types: metaphoric, deictic or beat gestures. Because all iconic gestures referred to an abstract concept, we categorized them as metaphoric (from now on we will only use metaphoric to refer to these types of gestures). Manipulative gestures were also needed to be included because fortune telling takes place by referring to the visual stimuli inside the coffee cup, so it was common for the participants to move or twirl the cup to be able to see better.

Gestures were then categorized according to the axis, in which the gesture took place. There were four different axes: sagittal, lateral, vertical, and diagonal (see Figure 2). The sagittal axis refers to the front-back orientation, lateral axis refers to the left-right orientation, vertical axis refers to the up-down orientation, and diagonal refers to any combination of these axes. After the axis was identified, the direction of the gesture was also coded. This involved several categories within the axis: to *left* and to *right* for lateral, *up* and *down* for vertical, *forwards* and *backwards* for sagittal, and any combination of these six directions for diagonal. If the gesture did not occupy any of these axes, but was only a movement within the same space or a change in the posture of the hand, it was coded as *no-axis*. Mostly beat gestures were coded as no-axis gestures. Finally, for deictic gestures, their reference point was also coded. For example, if they pointed to the ground, the gesture was coded as vertical axis, downward direction.

The interaction between gesture and speech was also coded for the match in meaning between both modalities. From the previous research we know that cultures, who write from left to right tended to put future in the front, up, and to the right, whereas they perceived past in the back, down, and to the left (Bergen & Lau, 2012; Casasanto & Jasmin, 2012; Ouellet & Santiago, 2010). With this expectation, whenever a temporal phrase matched with this formula it was coded a "match" and a "mismatch" when the phrase and gesture carried the opposite meanings. For example, when a forward gesture co-occurred with the word "later", it was a match situation. In contrast, if the gesture points at the back, while saying "later," we coded it as a mismatch.



Figure 2: (top left) lateral axis, (top right) vertical axis, (bottom left) sagittal axis, (bottom right), and diagonal axis as a combination of any other axes

Results

Speech. A total of 227 temporal phrases were identified. Three participants did not use any temporal phrases during their sessions, so they were excluded from the analysis. In total, 64 phrases (28.19%) were LPs and 163 phrases (71.81%) were MPs. The difference between the use of different types of phrases was statistically significant, χ^2 (1, 227) = 43.18, $p < 0.001$. The total of 227 phrases consisted of 72 unique temporal phrases. While some phrases like "*now*" were repeated several times across several participants, some like "*its time will come*" were used only once (see Appendix A for all types of phrases). Of all the phrases, 23% of them referred to the past, 22% referred to the present, and 55% referred to the future.

Gesture. Among the 15 participants, there were 102 gestures that were accompanied by temporal speech. All these remaining participants gestured at least once and at the most 21 times. Of all temporal speech, 45% involved temporal gesture use. Each participant used at least one temporal gesture. Of the 102 gestures, 32 (31.37%) were metaphoric, 8 (7.84%) were deictic, 59 (57.84%) were beat, and 3 (2.94%) were manipulative gestures. Difference among gesture types was significant, χ^2 (3,102) = 73.43, $p <$ 0.0001. Thus, people mostly produced beat and metaphoric gestures during temporal expressions.

People expressed time using either one of the 4 axes (sagittal, vertical, lateral, diagonal) or no axis at all. 60.78% of all gestures were performed without an axis. Among those gestures that were made within an axis, 42.5% were in sagittal, 7.5% were in lateral, 37.5% were in vertical, and 12.5% were in diagonal axes, χ^2 (3,40) = 14.8, $p < 0.001$. The majority of no-axis gestures were beat gestures (87.1%). As seen in Table 1, most metaphoric gestures consisted of sagittal (40.63%), vertical (25%), and no-axis (15.63%) gestures.

Table 1: Gesture Type – Axis Interaction

Axis/ Type	Sagittal	Lateral	Vertical	Diagonal	No Axis
Metaphoric	13	3	8	3	5
Deictic	1	0	5	1	1
Manipulative	1	0	0	0	2
Beat	2	0	2	1	54

The use of gesture types was similar for both temporal speech types, as shown in Table 2. Beat and metaphoric gestures were the most common types with a significant level of difference among gesture types for both LP, χ^2 (3,32) = 23, $p < 0.0001$, and MP χ^2 (3,70) = 55.14, $p <$ 0.0001 (see Table 2).

Table 2: Gesture Types According to Temporal Speech Type

Gesture type	Literal Phrases Number-Percentage	Metaphoric Phrases Number-Percentage
Metaphoric	10 (31.3%)	22 (31.4%)
Deictic	4 (12.5%)	4 (5.7%)
Beat	18 (56.3%)	41 (58.6%)
Manipulative	0 (0%)	3 (4.3%)

Although the use of gesture types did not differ between literal and metaphorical uses of temporal speech, there were differences in the use of gesture axis. Vertical and no axis gestures are the most common in LPs with a significant level of difference, χ^2 (4,32) = 35.5, $p < 0.0001$. While no axis gestures were still the most common type in MPs, the use of vertical gestures decreased and the use of sagittal gestures increased compared to LPs. The difference in the use of gesture axis was also significant for MPs, χ^2 (4,70) = 87.43, $p < 0.0001$.

Detailed analysis of this gesture-speech interaction results provided some interesting findings. Among all 15 vertical gestures, 6 of them accompanied by the word *"now"* and 1 accompanied *"today"* with a downward motion. Although there were more MP gestures in all gesture axes due to higher number of metaphorical phrases, vertical axis was an exception: There were more vertical gestures in LPs compared to MPs.

Table 3: Axes According to Temporal Speech Type

Gesture Axis	Literal Phrases Number-Percentage	Metaphoric Phrases Number-Percentage
Sagittal	2 (6.3%)	15 (21.4%)
Lateral	0 (0%)	3 (4.3%)
Vertical	10 (31.3%)	5 (7.1%)
Diagonal	2 (6.3%)	3 (4.3%)
No Axis	18 (56.3%)	44 (62.9%)

Finally, almost all meaningful gestures that were not beat gestures (except 1 of them) matched with the accompanied speech. For example, when people talked about time on a sagittal axis, they put the future events in front of their bodies and past to their back. Similarly, earlier events were represented on the left and future events were put to the right of the body.

Discussion

The present study investigated temporal speech and spontaneous gestures accompanying temporal phrases during a natural conversational setting in Turkish. We found that Turkish speakers used many spatial metaphors to express time, and the use of these metaphorical references for time was twice as many than the use of literal time words. Almost half of the temporal phrases were accompanied by spontaneous gestures. Beat and metaphoric gestures were the most common gestures among all types.

In contrast to findings in English, (Casasanto & Jasmin, 2012), the sagittal axis was the dominant one used for spontaneous temporal gestures in Turkish along with the use of vertical axis gestures. Previous studies did not report the use of vertical axis in temporal gestures.

In the present study, we made a detailed examination of temporal phrases. In a setting when someone talks about the past happenings and possible future events, time is mainly expressed using space. Phrases that are not necessarily considered as spatial can have spatial meanings in Turkish. In addition, Turkish language is very rich in using spatial metaphors. For example, to talk about a past event that the person has close connection with; one can use several metaphors such as *geçmişe bağlılık* (connection to the past)*, geçmişi anmak* (remembering the past), *geçmişle bağları koparmak* (loosing the connection with the past) or *geçmişte kalmak* (stay in the past). As a result of these, people produce many different types of spatial metaphors in Turkish. We also propose that the level of abstractness in metaphors would differ. In some phrases such as *aradaki 1 ay* (one month in between) may be less abstract than *zaman harcamak* (to spend some time). Yet, in this natural setting we did not observe enough instances to make a distinction for the level of abstractness in spatial metaphors. Future studies need to consider these differentiations in the use of space to describe time and investigate the level of abstractness in metaphors in experimental settings.

People attempt to visually support the conveying temporal message with gestures, which is more difficult to comprehend as an abstract concept. This is true for both literal and metaphoric use of temporal phrases. The use of metaphorical gestures is meaningful in this sense. When people talk about an abstract concept, they may supplement it with co-occurring semantic gestures like metaphoric ones. Yet, people also produced many beat gestures. The use of these meaning-laden beat gestures may help individuals to plan for their conversation and may benefit the speaker thinking about the abstract concept while preparing to talk about it. The temporal synchrony between speech and gesture can provide us information about the different functions of metaphoric and beat gestures, which is a question for future research.

Regarding the gestural axis, our findings contradict the previous research that used story prompts are given in story telling situations (Casasanto & Jasmin, 2012). The sagittal axis was the most commonly produced timeline in spontaneous gestures. In a natural setting when people talk about time without any study prompts, they used very few gestures on the lateral axis. One reason could be due to the nature of the conversation. In this fortune telling sessions as people focus on talking about the future (55% referred to the future events) and the listener sits across the person, it can be inevitable to use sagittal gestures. The teller refers to the person to express ideas about what can happen for future events. Thus, using the sagittal axis in this context can be more pragmatic than using the lateral axis.

In our analyses, we coded 2 additional categories for gesture axis: diagonal and vertical. Overall, people produced very few gestures using a mixture of sagittal and lateral axes (diagonal). Yet, most of the deictic gestures and 25% of the metaphoric gestures (8 out of 34) were produced using a vertical axis. This shows that to represent a present moment or current events such as “now” or “in the near future” people can use a vertical axis to refer to these by pointing the front of their bodies. These gestures are not only simple pointing gestures, thus the exact purpose of this axis also needs further investigation in an experimental setting.

One limitation of the current study is that the majority of temporal gestures come from a few participants, while others perform 3.5 gestures on average per person. Further studies need to work with a larger sample size to overcome these imbalances and have a more homogeneous sample.

Taken together, this study contributes to the literature on speech-gesture interaction in expressing temporal events. As in line with studies conducted in English, we presented evidence for the heavy use of spatial metaphors. However, metaphoric gestures are produced on a sagittal axis during natural conversation settings, suggesting that not all types of settings and/or languages invoke the use of lateral axis for spontaneous temporal gestures.

Acknowledgments

We thank Sabri Börütecene for Figure 1 drawing and Hulusi Kayabaşı for helping us with the gestures in Figure 2. We also thank Sinan Karaçam, Elifnaz Yalçınöz, Bilge Eren, Jiyan Erincik, Şeyma Taşel, and Zeynep Yıldız for their help in transcribing and coding the data.

References

Bergen, B. K., & Lau, T. T. C. (2012). Writing direction affects how people map space onto time. *Frontiers in Psychology*, *3*.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1-22.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently?. *Cognition, 118(1),* 123-129.

Casasanto, D., & Jasmin, K. (2012). The Hands of Time: Temporal Gestures in English Speakers. *Cognitive Linguistics, 23*(4), 643-674.

Casasanto, D., & Bottini, R. (2014). Mirror reading can reverse the flow of time. *Journal of Experimental Psychology: General*, *143*(2), 473.

Cooperrider, K., & Núñez, R. (2009). Across time, across the body Transversal temporal gestures. *Gesture*, *9*(2), 181-206.

Evans, V. (Ed.). (2003). *The structure of time: Language, meaning and temporal cognition*. John Benjamins Publishing.

Fuhrman, O., & Boroditsky, L. (2010). Cross-Cultural Differences in Mental Representations of Time: Evidence From an Implicit Nonlinguistic Task. *Cognitive Science*, *34*(8), 1430-1451.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space → time metaphors. *Language and Cognitive Processes*, *17*(5), 537-565.

Lakoff, G. (1992). The contemporary theory of metaphor. In: Andrew Ortony (ed.), *Metaphor and Thought*, 202–250. Cambridge: Cambridge University Press.

Le Guen, O., & Pool Balam, L. I. (2012). No metaphorical timeline in gesture and cognition among Yucatec Mayas. *Frontiers in Psychology*, *3*(271).

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Núñez, R. E., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive science*, *30*(3), 401-450.

Ouellet, M., Santiago, J., Israeli, Z., & Gabay, S. (2010). Is the future the right time?. *Experimental Psychology, 57,* 308-314.

Ördem, E. (2014). Türkiye Türkçesinde Zaman Metaforları: Bilişsel Dil Bilimsel Bir Yaklaşım. *Journal of Dil Arastirmalari*, (14).

Özçalişkan, Ş. (2004). Time Can't Fly, But a Bird Can: Learning to Think and Talk About Time As Spatial Motion in English and Turkish. *European Journal of English Studies*, *8*(3), 309-336.

Appendix A.

Time expressed in literal and metaphorical phrases.

Literal Phrases

…'dan beri *"since"*
1.5 ay *"1.5 months"*
2 vakte kadar *"until 2 times"*
Aylardan beri *"for months"*
Aynı zamanda *"at the same time"*
Bir anda *"suddenly"*
Bir gün *"one day"*
Bir süre daha *"a while longer"*
Bugün *"today"*
Eskisi *"old"*
Geçmiş zaman *"past time"*
Geçmişte *"in the past"*
Gelecek *"future"*
Ne zaman … o zaman *"when ... then"*
Olacak zaman *"times to come"*
Şimdi(lik) *"now (for now)"*
Şu an/saat/tarih *"this moment/hour/date"*
Zamanında *"in the past"*

Metaphorical Phrases

1 yıl içinde *"in 1 year"*
3-4 ay sonra *"3-4 months later"*
Aradaki 1 ay *"1 month in between"*
Arkası(ndan) *"after/behind (after that)"*
Az önce *"a moment ago"*
Bir ara *"some time"*
Bir süre sonra *"a while later"*
Biraz zaman önce *"a while ago"*
Bu ara/dönem/sıralar *"these days/nowadays"*
Daha önce *"before"*
Geçmiş bir yerde kalmış *"past is stuck"*
Geçmişe bağlılık *"connection to the past"*
Geçmişi anmak "*remembering the past"*
Geçmişle bağları koparmak *"losing connection with past"*
Geçmişte kalmak *"stay in the past"*
Geçtiğimiz 2 yıl içinde *"in the last 2 years"*
Geride kalmak *"left behind"*
İleride(si) *"ahead"*
Kısa süre/vade *"short period/duration"*
Kısa zaman sonra *"short while later"*
Önceki süreç *"previous term"*
Öncesi *"before"*
Önümüzdeki 3 ay/dönem *"next 3 months/term"*
Önündeki süreç *"period ahead"*
Orta vade *"medium duration"*
Sonra *"Later"*
Sonraki günler *"Days after"*
Sonrası(nda) *"After (that)"*
Şu sıralar *"recently"*
Uzun süre/vade *"long period/duration"*
Vakit vermek *"to give time"*
Vakti olmak *"to have time"*
Yakın zaman *"close time"*
Zaman almak *"to take time"*
Zaman geçirmek *"to spend time"*
Zaman girmiş "time came between"
Zaman sıkıştırıyor *"time rushes"*
Zaman tanımak *"to give time"*
Zamana bırakmak *"to let time go by"*
Zamanı gelmek *"time has come"*
Zamanı var *"there is time"*

Children's Awareness of Authority to Change Rules in Various Social Contexts

Xin Zhao (xz437@cornell.edu)
Department of Human Development, Martha Van Rensselaer Hall, Ithaca, NY 14853 USA

Tamar Kushnir (tk397@cornell.edu)
Department of Human Development, Martha Van Rensselaer Hall, Ithaca, NY 14853 USA

Abstract

To investigate children's awareness of authority to change rules, we showed children (ages 4-7) videos of one child playing a game alone or three children playing a game together. In the group video, the game rule was initiated either: by one of the children, by three children collaboratively or by an adult. They then were asked whether the characters in the videos could change the rules. Children believed that the character could change the rule when playing alone. Their responses to the group video depended on how the rule was initiated. They attributed authority to change rules only to the child who initiated the rule, unless the rule was created collaboratively. We also asked children whether they could change norms (school/moral/artifact norms) in daily life; and found moral/artifact distinction in children's endorsement of norm changing. These results suggest that children recognize flexibility in changing rules even in preschool years.

Keywords: cognitive development, social cognition, normative reasoning, authority, moral development

Introduction

As adults, we recognize that rule following is important, but can also reason about cases in which rules can (and may need to) be changed. Interest in the origins of children's understanding of rules and norms dates back to Piaget, who interviewed children about rules in marble games (Piaget, 1965). Based mainly on naturalistic observations and interviews, Piaget argued that children's understanding of rules starts out rigid and then becomes flexible. Young children start from viewing rules as eternal, fixed and unalterable, while later (at around 10 years old) understand that rules are alterable based on mutual agreement.

This idea that young children view rules as fixed has some support from recent research showing how important rule following and rule enforcing is to very young children. For example, from the second year of life, children begin to be aware of and respect various rules in their families (e.g. Dunn & Munn, 1985, 1987). Observational studies of family interactions have shown that toddlers talk about permissibility of actions and use social rules to explain and justify their behaviors in their conflicts with parents and siblings (Dunn & Munn, 1985, 1987). A series of experimental studies have shown that toddlers enforce arbitrary rules even on strangers (Rakoczy, Warneken, & Tomasello, 2008). Children spontaneously protest and criticize rule violators, but do not criticize same act that does not constitute a violation of a rule (Rakoczy, 2008; Wyman, Rakoczy, & Tomasello, 2009).

Also, younger children are more likely than older children to emphasize the role of rules and social norms in their causal-explanatory reasoning about human behaviors. For example, Kalish and his colleagues (Kalish & Lawson, 2008; Kalish & Shiverick, 2004) have found that four-year-olds are more likely to make predictions about people's future behaviors based on social norms over personal preferences. It is not until age eight that children shift away from norm-based explanations to explanations based on personal preferences.

Paralleling the way preschoolers understand the norms governing human behavior, they exhibit a "functional fixedness" in their understanding of the norms governing artifact usage. In some cases, when shown a function of an artifact, they refused to use the object for alternative feasible purposes or other objects for the same purpose (e.g. Casler & Kelemen, 2005; Kelemen, 2004). In other cases, though they allow multiple uses, they believe that only one is the norm (e.g. Vredenburg, Kushnir & Casasola, 2014).

From a simple reading of these examples, it seems that children are quite rigid in their beliefs about rules. But even toddlers, and certainly preschoolers, show some flexibility as well. For example, they show context-sensitivity: distinguishing between moral rules, which pertain to violations of common good, justice and others' wellbeing (e.g. harming and not sharing), and conventional rules, which are arbitrarily decided by social groups. These distinctions influence how young children judge rule violations in terms of seriousness, contingency and generalizability; they rate conventional transgressions as less serious, more contingent on the presence or absence of rules and less generalizable across different contexts than moral transgressions (e.g. Nucci & Nucci, 1982; Smetana, 1981; Turiel, 1998).

There are other, more indirect indications that young children may not be so rigid. For one thing, the success of the "protest" paradigm rests on fact that children learn new rules quickly and easily (e.g. Rakoczy, Brosche, Warneken, & Tomasello, 2009; Rakoczy et al., 2008). Also, at five years old, children spontaneously create their own rules to coordinate themselves in games, and even negotiate rules collaboratively with others in cases of conflict (Göckeritz, Schmidt, & Tomasello, 2014; Köymen et al., 2014). Finally, children often make up rules in make-believe games when they play alone or with friends (Taylor, 1999).

We offer a proposal to resolve these conflicting findings: our hypothesis is that children's understanding of norms and rules is flexible, and that fixedness arises due to

considerations of *authority* that govern the formation and/or enforcement of rules. Children's considerations of authority have been studied in cases of object ownership (e.g. Neary & Friedman, 2014) and creator intent (e.g. Diesendruck, Markson, & Bloom, 2003), and in such cases young children consider authority when making decisions about who can possess an object, determine its purpose, or distribute resources. Thus, we propose that, as a guiding principle, children consider the issue of authority over rules (for example: were the rules made by an individual, by collaborative agreement among members of a group, or by some other means?) in deciding whether the rules are fixed or flexible.

In order to test this idea in the present study, we ask young children an important question about flexibility – that is, whether rules can change and who can change them. Importantly, we vary the way in which the rules are introduced and examine whether children endorse rule changes, under what conditions they do, and whether their endorsements change with age. We choose to study children ages 4 through 7, as this is a time in which children's normative reasoning undergoes important developments (e.g Kalish & Shiverik, 2004)

Our main comparisons of interest come from the manipulation of individual or collective authority. First, we manipulated whether the game was solitary or social. Based on prior work, children may be more rigid when rules apply to a social context than to a solitary individual (where, for example, the rule is indistinguishable from a personal preference).

Second, we manipulated whether the rule was initiated by an adult authority figure or by the children themselves. Once again, based on prior work, children should be more rigid when the rule comes from an adult authority figure.

Third, we manipulated whether the rule was initiated by one of the children (i.e. a "leader" and his/her "followers") or by all of the children through a collaborative agreement. Since there was no *adult* authority, it is an open question whether children would view the authority of the "leader" child as enough to warrant permission to change the rules. Also, though past work has shown that children appreciate and honor joint commitments (e.g. Warneken, Gräfenhain, & Tomasello, 2012), it is still an open question whether collaborative agreement could lead them to endorse *any child* changing or of *no child* doing so.

Our task began with children viewing videos of other children playing games. The solitary and group contexts were presented within subjects. The different group contexts were randomly assigned between subjects. The rule of the game was initiated either: 1) by one of the children, 2) by collaborative agreement of all the three children or 3) by an adult. Children were then asked whether each character in the videos could change the rules.

We also surveyed children regarding whether they believed children could change norms in daily life (school norm, moral norm, and artifact norm). We included this additional measure for two reasons. One, we were interested in whether children would ever endorse *children* changing rules that are made *by adults*. Recent studies (Chernyak & Kushnir, 2013; Chernyak, Kushnir, Sullivan, & Wang, 2013) have found that preschool children do not tend to endorse the freedom to violate social and moral norms, but that this changes with age. Also, consistent with social domain theory (e.g. Turiel, 1998; Smetana, 1981), children are more likely to say that they can choose to violate social and artifact conventions (such as what to wear, where to eat, etc) than moral norms (harm, fairness, helping). We expect to see that if children endorse norm changing in our task at all, it would be in accordance with this moral/conventional distinction.

Importantly, too, we were interested in whether there was any consistency in individual children's beliefs about game rule changing and norm changing. Thus, we explored the question of whether children's attitudes about the rules of the games in our artificial lab task would be in accordance with their beliefs about the types of moral and conventional norms they encountered in their own daily lives.

Method

Participants

Sixty-nine children aged 4-7 years old (M = 5.57, SD=0.99) participated. Participants were recruited from preschools, afterschool programs and museums in a small university town. Two additional children participated but were replaced because of the experimenter's mistakes.



Figure 1. Screenshots of the Videos Used in the Video Task (Above: alone video; bottom: group video).

Materials

The materials included 8 videos showing children playing games and a questionnaire with 3 vignettes. The videos were filmed with children of around 9-10 years old who were able to act consistently across videos. Half of the videos were boys, the other half were girls. Two videos were of a boy/girl playing a stacking game alone (*Alone Video*), six videos of three boys/girls playing a sorting game together (*Child Rule, Collaborative Rule, Mom Rule*). The genders of the characters were counterbalanced across participants; if participants saw the girls together, they saw the boy alone, and vice versa. The alone videos were about 20 seconds each, and the group videos were about 34 seconds each.

Procedure

All children were interviewed in a quiet corner or a separate room at the local schools or the museum. The interview consisted of two parts: a video task and a questionnaire. All the participants completed the video task first, following by the questionnaire.

Video task Each child was shown two videos (order counterbalanced). Children were randomly assigned to view one of the three group videos (*Child Rule, Collaborative Rule, Mom Rule*) where three children were playing a sorting game (sorting balls with different colored stickers into baskets) together. Each child also saw the *Alone Video* where one child was playing a stacking game (stacking blocks of a certain color into a tower) alone.

Child Rule: Before showing the video, the experimenter pointed to each of the characters in the video and introduced their names, and said, "They are friends and they are going to play a game together." Then she pointed to the child in the middle and said "John/Sophie has a rule for this game. Let's see what rule John/Sophie has for this game." Then the experimenter played the video. Figure 1 demonstrated the screenshots of the videos. In the video, there were a bunch of balls with either yellow or blue stickers on the table. The three characters sat in a triangle facing each other. The child sitting in the middle said, "I have a rule for this game. Blue stickers, here; yellow stickers here. Let's play together." Then the three children played together following the rule and finished sorting the balls.

Collaborative Rule: The procedure was similar to the Child rule condition except the experimenter introduced the video by saying, "They are going to make up a rule *together*." In the video, the center child proposed the rule, then looked at the other two children for agreement. One at the time each of the other two children made eye contact and nodded to the center child, indicating agreement, before they began the game. We used this type of silent "assent," rather than having all three children converse and decide collectively (e.g. Köymen et al., 2014) in order to keep the videos as similar as possible.

Mom Rule: The participant was told, "John's mom has a rule for this game." In the video, a woman walked in the screen and made eye contact with the children and said, "I have a rule for this game" then stated the same rule as in the other two conditions, "blue stickers here; yellow stickers here." Immediately after the mom left, the children began playing, in exactly the same way as the other two group videos.

Dependent measures: After showing each video, the experimenter asked the participants if they remembered the rule mentioned in the video to make sure that they were paying attention and understood the video. If the participant didn't answer it correctly, the video was shown again. Children correctly answered the rule of the game 93.5% of the times when shown the video for the first time. Remaining participants correctly answered the rule after being shown twice. After that, the participants were asked two critical questions (order counterbalanced). One about the center child "Now John/Sophie wants to change the rule of this game. He wants the rule to be that yellow stickers here and blue stickers here (switching the places putting the color of the stickers), Can John/Sophie just change the rule?" The other question is about one of the other two children (non-center child): "Now Andy/Julia wants to change the rule of this game. Can he/she just change the rule?" In the mom rule condition, in addition to the two questions about the center child and one of the non-center children, the experimenter also asked whether the mom could change the rule if she wanted. The participants were also asked "why" after the yes/no responses.

Alone Video: In the Alone Video, the experimenter first said, "This is John/Sophie. John/Sophie is playing alone. He/She has a rule for this game. Let's see what rule he/she has for this game." In the video, there were a bunch of either red or green blocks on the table. The child said, "I have a rule for this game. Red ones make a tower, green ones stay flat." Then the child finished playing with the blocks following the rule. After being shown the video, the participants were asked the same rule-changing question as in the group video (but the rule change was to stack the other colored blocks).

Questionnaire After the video task, all participants were also read a questionnaire. The questionnaire consisted of 3 items in the following general format:

"There is a rule at Lily's school. The rule is that kids sit in red chairs during lunch. Lily always follows this rule and sits in the red chair during lunch. But now, Lily wants to change the rule. She wants the rule to be that kids sit in green chairs during lunch. Can Lily just change the rule? Why?"

We included three items: school norm, moral norm, artifact norm. In *School norm,* the character in the story wants to change an arbitrary rule at school (sitting in red chairs during lunch). In *Moral norm*, the character in the story wants to change a moral norm (sharing with siblings). In *Artifact norm*, the character in the story wants to change a known artifact convention (using umbrella when it's raining). The participants were also asked "Why" to explain their responses. The orders of the three items were fully counterbalanced across participants.

Coding

For each question, participants were given a score of "0" if they answered "no" and "1" if they answered "yes". The first author and a research assistant coded participants' responses. Reliability between coders was 99.08%.

Results

We first checked if there were any order effects or gender effects and found no effects of order or gender in any of the questions (Fisher's exact test, p's $> .05$). So we combined the data of different orders and genders together in the following analysis.

Video task

Figure 2 demonstrated the proportion of "yes" responses to the rule changing questions in each condition.

Child Rule In the child rule condition, the significant majority (77%) of the children answered that the child who made up the rule (center child) can change the rule, Binomial $p = .017$. About half of those who answered "yes" (47%) gave the explanations "he made up the rule" or "it's his game." In contrast, only a few children (36%) answered that the child who did not make up the rule (non-center child) could change the rule, Binomial $p = .286$. Out of those who answered "no," half gave the explanations "he didn't make up the rule" or "he's not the boss." A McNemar's test showed that significantly more children attributed an ability of changing the rule to the child who made up the rule than the children who didn't make up the rule, $p = .012$. There was no correlation with age in this condition (p's > .05).

Collaborative Rule In the collaborative rule condition**,** the majority (75%) of the children answered that both the center child and the non-center child could change the rule of the game, Binomial p's < .023. McNemar's test showed no significant difference between the number of participants who said the center child could change the rule and the number of participants who said the non-center child could change the rule ($p = 1.0$). We then added up each participant's "yes" responses to the center child and non-center child together, and found that this score was negatively correlated with age in months ($r = -.406$ $p = .049$). This suggested that young children were more likely to say that both children could change the rule, while older children were more likely to say that *neither* could do so.

Mom Rule In the mom rule condition, 39% of children answered that the center child could change the rule, and 35% answered that the non-center child could change the rule, Binomial p's > .05. A significant majority (87%) of the children answered that the mom could change the rule, Binomial $p < .001$. The most frequent explanations they gave referred to the fact that the mom made up the rule (25%) and that the mom was the authority figure (e.g. "Because she's the boss", 15%). McNemar's tests showed that children were more likely to think that mom could change the rule than either the center child ($p = .001$) or the non-center child ($p < .001$), but there was no significant difference in proportion of "yes" responses to questions about the center child and the non-center child ($p = 1.0$), Also, no age effect was found in this condition (p's > .05).

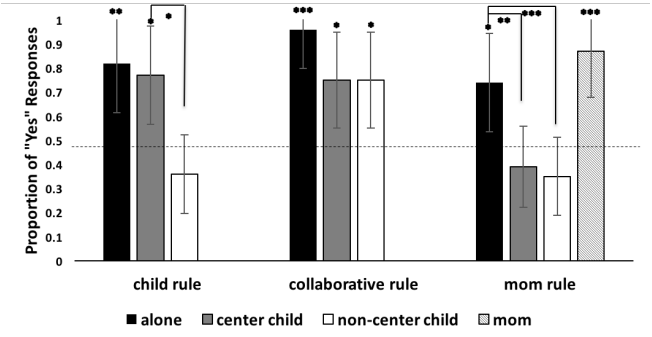


Figure 2. Proportion of "Yes" Responses to the Rule Changing Questions in Each Condition (bars represent 95% confidence intervals for each mean; asterisks indicate a significant difference from chance: (*) p<.05, (**) p<.01, (***) p<.001, using Binomial tests).

We also compared children's responses to the rule-changing questions in the three different conditions. Compared to the Mom Rule condition, participants in the Child Rule and Collaborative Rule conditions were significantly more likely to answer that the center child could change the rule (Fisher's exact test, p's < .05*)*. Participants in the Collaborative Rule condition were significantly more likely to respond that the non-center child could change the rule compared to the participants in the Child Rule condition (Fisher's exact test, $p = .016$) and the Mom Rule condition (Fisher's exact test, $p = .008$).

Alone Video We also looked at children's responses to the alone video, where one child was playing alone. An overwhelming majority of the participants answered that the child in the video could just change the rule, Binomial p's < .05. The most frequent explanations they gave referred to the fact that the child made up the rule (e.g. "Because it's her game", 36.2%) or that the child was by herself (e.g. "Because she's playing alone", 13.8%). McNemar's tests showed that children were more likely to say that the center child could change the rule when he/she was playing alone than playing in group when the rule was made by mom ($p = .008$). The difference was marginally significant in the Collaborative Rule condition ($p = .063$) and not significant in the Child Rule condition ($p = 1.0$). There was no correlation with age in *Alone Video* (p's > .05).

Questionnaire

The questionnaire consisted of three questions respectively referring to school norm (sitting in the red or green chairs during lunch), moral norm (sharing or not sharing with brothers) and artifact norm (holding an umbrella or holding a bucket in the rain). Figure 3 showed the proportion of "yes" responses the participants gave in each of the stories. Binomial tests showed that the significant majority of the children (65%) answered that the character in the story was not able to change the rule at school ($p = .015$). A significant majority of the children (74%) answered that the character in the story was not able to change the moral rule, ($p < .001$). However, about half of the children (49%) said that the character in the story could change artifact rule ($p = 1.0$). McNemar's tests also revealed that significantly more children answered that the child in the stories could change the artifact rule than the school rule ($p = .013$) or the moral rule ($p < .001$). No correlation with age was found in any of the three stories (p's > .05).

We then examined the relationship between children's responses to the rule-changing questions in the games and their responses to rule-changing questions about everyday life. We added up each participant's "yes" responses to the three norm questions and found that this score was significantly positively correlated to their responses to the rule-changing question of the center child in the Child rule condition and the Mom Rule condition, even when age was

controlled (Child Rule: r = .473, p = .03; Mom Rule: r = .821 p < .001). The correlation was marginally positive in the Collaborative Rule condition (r = .377, p = .076). This score was also positively correlated to children's responses to the non-center child in all the three group games, when age was controlled (p's < .05). Thus, children who thought that the everyday norms could be changed also tended to endorse rule changing in the video task.

In a second analysis, we gave children a "1" for every response consistent with the distinction between artifact norm from moral or school norm (i.e. a "1" for "no" to moral and school and a "1" for "yes" to artifact). We found that this score was marginally negatively correlated with their responses to the non-center child changing rules in the Mom Rule game (r = -.49, p = .021), while not significantly correlated with the responses to the center child or any child in the Collaborative Rule or the Child Rule game (p's > .05). This suggests that those children who make stronger distinction between the moral/school and artifact norms endorsed less authority for the children to change rules created by adults.

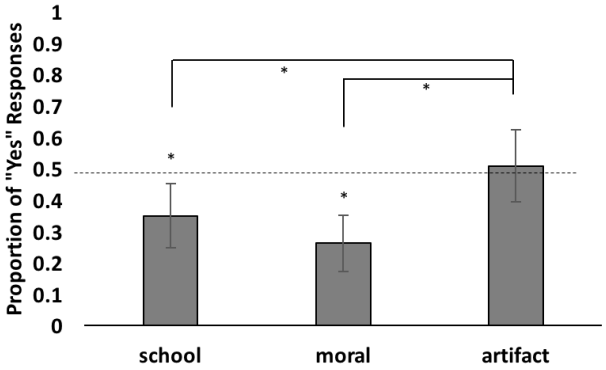


Figure 3. Proportion of "Yes" Responses to the Rule Changing Questions for the School Norm, Moral Norm and Artifact Norm Items (asterisks indicate a significant difference from chance: (*) p<.05, (**) p<.01, (***) p<.001, using Binomial tests).

General Discussion

Our results reveal that young children's beliefs about who can and cannot change rules are sensitive to considerations of the authority over the formation of the rules. Children recognized that game rules could be changed when a child had made up the rule for his/her own solitary play. But when the rules were made for playing with a group, children took into account how the rules were initiated even though all the other information (e.g. the content of the rule, the participants of the game) were kept the same across contexts. When the rule was initiated by a single player, children believed only that player had the authority to change the rule, and not his/her playmates. When the rule was made through a collaborative agreement among all three players, children said that any could change the rule. When the rule was initiated by an adult (mom of one of the children), children said that only the adult had the authority to change the rule of the game, and not any of the players.

Our results complement prior work focused on children's respect and enforcement of rules by showing that children recognize that there is flexibility in changing rules, provided the agent of change has some authority to do so. The recognition of *who* has that authority aligns well with existing knowledge in several different areas of research, including children's understanding of authority, of ownership and creator intent, and of the norms of collaboration. For example, children's responses that rules which comes from mom cannot be changed by children is consistent with the evidence shown in previous research that young children recognize parents as important authority figures (Laupa, 1995). Children's belief that a rule initiated by a single child for group play can only be changed by that child is consistent with previous work on children's understanding of ownership of objects (e.g. Neary & Friedman, 2014) and creator intent (e.g. Diesendruck, Markson, & Bloom, 2003). Just like with objects, young children may develop awareness of ownership and/or creation of rules: they understand that the person who "owns" the rule (in the sense that they created it) has the authority to change the rule, while other people do not have the authority.

Children's response that any group members has the same authority to change a collaborative rule is consistent with prior work on children's understanding of joint commitments to collaborate (e.g. Warneken et al., 2012). It is noteworthy, however, that there were age differences, where young children were more likely to say that *both* children could change the rule and older children were more likely to say that *neither* could do so. This may have been a difference in the way children interpreted this one instance of agreement. Younger children may have taken the set up to imply agreement in general: that is, that if one member of the group changed the rule the other members would agree. Older children, on the other hand, may require more evidence to assume that children who agree once will always agree with their playmates. This would be consistent with the fact that, with age, children accumulate experiences in which friends do not necessarily agree. It would be interesting to explore this hypothesis directly. For example, we could examine how children's changing beliefs about friendship relate to their responses to our task. We could also explicitly manipulate whether or not other players always agree on the rule across games, only sometimes agree, or never agree.

Finally, it was interesting to find that, in some cases, children said that they could change the norms in their everyday lives. Consistent with prior work on freedom on choice (Chernyak et al, 2013) and with the predictions of social domain theory (e.g. Turiel, 1998; Smetana, 1981), children were more likely to say that they could change artifact norms than moral norms.

The tendency to say "yes" to game rule changing and also "yes" to norm changing, however, suggests that perhaps

children didn't distinguish between moral norms, conventional norms, and game rules initiated by children. This is less of a worry, however, when combined with the result that children whose responses were most consistent with this moral/conventional distinction were most likely to endorse the mom's authority in the game scenario, but no more likely to endorse the child's. This final result indicates that children do distinguish between children's rules for their own game and rules set by an authority figure, but leaves open questions about the nature and extent of this distinction.

In summary, our results demonstrate that belief in flexibility in changing rules is present even in the preschool period. Young children understand the importance of following and enforcing rules and norms, but at the same time can also reason about how rules can be changed and how advances take place in human activities. Critically, this study shows that children consider the issue of *authority* over rules in deciding whether rules can change and who can change them. Taken together, this study suggests that the important feature of human thinking - that we make rational and flexible reasoning about human activities - is present in young children' beliefs about rules and norms.

Acknowledgments

We gratefully thank the members of the Early Childhood Cognition Lab, and especially Cindy Wu, Sammi Wong and Mona Chen for assistance with recruitment and coding. We thank the staff of the Sciencenter and preschools, and the children and parents who participated in this research.

References

Casler, K., & Kelemen, D. (2005). Young children's rapid learning about artifacts. *Developmental Science*, *8*, 472–480.

Chernyak, N., & Kushnir, T. (2013). The Self as a Moral Agent: Preschoolers Behave Morally but Believe in the Freedom to Do Otherwise. *Journal of Cognition and Development*, *15*, 453–464.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A Comparison of American and Nepalese Children's Concepts of Freedom of Choice and Social Constraint. *Cognitive Science*, *37*, 1343–1355.

Diesendruck, G., Markson, L., & Bloom, P. (2003). Children's Reliance on Creator's Intent in Extending Names for Artifacts. *Psychological Science*, *14*, 164–168.

Dunn, J., & Munn, P. (1985). Becoming a Family Member: Family Conflict and the Development of Social Understanding in the Second Year. *Child Development*, *56*, 480–492.

Dunn, J., & Munn, P. (1987). Development of justification in disputes with mother and sibling. *Developmental Psychology*, *23*, 791–798.

Göckeritz, S., Schmidt, M. F., & Tomasello, M. (2014). Young children's creation and transmission of social norms. *Cognitive Development*, *30*, 81-95.

Kalish, C. W., & Lawson, C. A. (2008). Development of Social Category Representations: Early Appreciation of Roles and Deontic Relations. *Child Development*, *79*, 577–593.

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, *19*, 401–416.

Kelemen, D. (2004). Are Children "Intuitive Theists"? Reasoning About Purpose and Design in Nature. *Psychological Science*, *15*, 295–301.

Köymen, B., Lieven, E., Engemann, D. A., Rakoczy, H., Warneken, F., & Tomasello, M. (2014). Children's Norm Enforcement in Their Interactions With Peers. *Child Development*, *85*, 1108–1122.

Laupa, M. (1995). Children's reasoning about authority in home and school contexts. *Social Development*, *4*, 1–16.

Neary, K. R., & Friedman, O. (2014). Young Children Give Priority to Ownership When Judging Who Should Use an Object. *Child Development*, *85*, 326–337.

Nucci, L. P., & Nucci, M. S. (1982). Children's Social Interactions in the Context of Moral and Conventional Transgressions. *Child Development*, *53*, 403–412.

Piaget, J. (1965). The moral judgment of the child (1932). *New York: The Free*.

Rakoczy, H. (2008). Taking fiction seriously: Young children understand the normative structure of joint pretence games. *Developmental Psychology*, *44*, 1195–1201.

Rakoczy, H., Brosche, N., Warneken, F., & Tomasello, M. (2009). Young children's understanding of the context-relativity of normative rules in conventional games. *British Journal of Developmental Psychology*, *27*, 445–456.

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*, 875–881.

Smetana, J. G. (1981). Preschool Children's Conceptions of Moral and Social Rules. *Child Development*, *52*, 1333–1336.

Taylor, M. (1999). *Imaginary companions and the children who create them.* Oxford University Press.

Turiel, E. (1998). The development of morality. *Handbook of child psychology*.

Vredenburgh, C., Kushnir, T., & Casasola, M. (2015). Pedagogical cues encourage toddlers' transmission of recently demonstrated functions to unfamiliar adults. *Developmental science*, *18*, 645-654.

Warneken, F., Gräfenhain, M., & Tomasello, M. (2012). Collaborative partner or social tool? New evidence for young children's understanding of joint intentions in collaborative activities. *Developmental Science*,*15*,54–61.

Wyman, E., Rakoczy, H., & Tomasello, M. (2009). Normativity and context in young children's pretend play. *Cognitive Development*, *24*, 146–155.

An Ecological Model of Memory and Inferences

Daniela Link (daniela.link@unil.ch) and Julian N. Marewski (julian.marewski@unil.ch)
Department of Organizational Behavior, Université de Lausanne, 1015 Lausanne, Switzerland

Lael J. Schooler (lschoole@syr.edu)
Department of Psychology, Syracuse University, Syracuse, NY 13244, USA

Abstract

In this paper, we develop a memory model that predicts retrieval characteristics of real-world facts. First, we show how ACT-R's memory model can be used to predict people's knowledge about real-world objects. The model assumes the probability of retrieving a chunk of information about an object and the time to retrieve this information depend on the pattern of prior environmental exposure to the object. Second, we use frequencies of information appearing on the Internet as a proxy for what information people would encounter in their natural environment, outside the laboratory. In two Experiments, we use this model to account for subjects' associative knowledge about real-world objects as well as the associated retrieval latencies. Third, in a computer simulation, we explore how such model predictions can be used to understand the workings and performance of decision strategies that operate on the contents of declarative memory.

Keywords: ACT-R; declarative memory; decision making; fast-and-frugal heuristics; Internet; strategy selection

The Importance of Memory for Inferences

Many of our every-day decisions are based on declarative knowledge retrieved from long-term memory. For example, a consumer who decides between different car brands will retrieve knowledge, such as information about the price segment, brand image, or fuel efficiency, to decide which brands to consider more closely. In judgment and decision research, there is a rich literature on how people infer unknown criteria, such as the quality of a car, based on object attributes used as *cues* (e.g., Gigerenzer, Hertwig, & Pachur, 2011). The kind of *cue-knowledge* a person retrieves when making a decision will likely depend on the information related to the decision objects she has encountered before, say, on the Internet. The person will then use this cue-knowledge as input of decision strategies when making inferences about unknown *criteria*, such as the quality of a car. A detailed cognitive model of how environmental patterns are reflected in memory, tied to models of decision making, could hence help to understand how human decision making depends on and is adapted to the environment.

Modeling Declarative Knowledge in ACT-R

In the cognitive architecture *ACT-R*, knowledge about the world is represented in declarative memory (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004). The basic unit of knowledge in declarative memory is the *chunk*. A chunk combines a set of elements into a long-term memory unit, where different concepts are configured together in the chunk's *slots*. New declarative knowledge is added to memory by encoding representations of objects that are attended in the environment. For example, the knowledge that the city of Berlin has an airport can be represented in a chunk with the following structure:

BERLIN-AIRPORT

ISA	CITY_FACT
CITY	BERLIN
ATTRIBUTE	AIRPORT
RELATION	HAS

The chunk is of type CITY_FACT. Its slots contain the city BERLIN, the attribute AIRPORT, and the relation HAS. If the same constellation of concepts is encountered and attended again, rather than creating a duplicate chunk, the memory entry of an existing chunk will be strengthened, and as a result, will become more readily accessible in memory.

In addition to symbolic information (the chunk-type and slot values), each chunk encodes subsymbolic information about the likelihood that the chunk will be needed to reach one of the system's processing goals – the chunk's *activation* (Anderson & Milson, 1989). A chunk's activation, in turn, is probabilistically related to its retrieval and the time required for retrieval. Table 1 summarizes the relevant equations for ACT-R's declarative memory system (see Anderson et al., 2004 for details on ACT-R's theory of declarative memory).

Table 1: Equations relevant for memory retrieval

Equation number	Equation
1) Activation	$A_i = B_i + \sum_{j=1}^{m} W_j S_{ji}$
2) Base-level Activation	$B_i = \ln n / (1-d) - d \ln L$
3) Associative Strength	$S_{ji} = \ln \frac{P(i\|j)}{P(i)}$
4) Attention Weighting	$W_j = {}^{W}/_{m}$
5) Retrieval Probability	$p_i = \frac{1}{1+e^{-(A_i-\tau)/s}}$
6) Retrieval Time	$T_i = F e^{-A_i}$

Note. Equation 2 is an approximation of base-level activation.

In the following, we explore to what extent ACT-R's memory model can be used to predict people's knowledge about real-world objects when using the Internet as a mirror

of the environment. Implementing the above formulas in Matlab, we aspire to develop a convenient method for modeling memory contingent on frequencies of objects and their attributes in people's natural environment, outside the laboratory. As we will illustrate, alike efforts may be helpful when using ACT-R's memory system to understand, for instance, the workings and performance of decision strategies.

Predicting Retrieval from Internet Frequencies

We used frequencies of mentions of objects on the Internet as proxies to what information people would encounter in their natural environment, outside the laboratory (henceforth *web frequency*). Specifically, we searched for the names of cities (e.g., Berlin), the names of city attributes (e.g., airport) and the combination of cities and attributes (e.g., Berlin airport) on www.en.wikipedia.org in English, using the Wikipedia API tool to find the total number of hits for pages containing our search term. Search was performed on September 14th 2014.

To calibrate and test our model, in two Experiments, we collected behavioral data on people's knowledge of European cities by asking them for pairs of cities and attributes whether or not they had heard of the city having the attribute before (e.g., "Does Berlin have an airport?").

Memory Activation of Knowledge

We assume that when a person believes she has encountered a city together with an attribute before, this implies a successful retrieval of a chunk *i* encoding the relation between the city and the attribute. Given this assumption, each time a person tries to retrieve a combination of a city and an attribute, according to Equation 1, two sources of activation for chunk *i* are (a) the *base-level activation* of chunk *i* whose slots contain the city name and the attribute name and (b) *spreading activation* from the city and attribute names included in the retrieval request.

A chunk's base-level activation is a function of the number of encounters, *n*, with the object or relation encoded in the chunk (Equation 2). We approximate the frequency with which a city and an attribute occur together in one context by the number of times both concepts co-occur on one page in the knowledge base Wikipedia ($N_{\text{city\&cue}}$). As a simplification, we assume the time of creation to be equal for all chunks and hence replace the lifetime of the chunk, *L*, by a constant. The decay parameter *d* is usually set to .5. The base-level activation of the chunk *i* can hence be written as:

$$B_{\text{i,web}} = c_0 + \ln N_{\text{city\&cue}}, \tag{7}$$

where the constant c_0 absorbs the value for the term ln 2 - .5 ln *L*.

For the spreading activation, we assume that the chunks encoding the city and the attribute forming part of the retrieval request spread activation to chunk *i*. Following Equation 3, the associative strength between a city and an attribute depends on the number of times the city and the attribute co-occur together relative to each of the concepts' base-rates of occurring (Schooler & Anderson, 1997). *P* (*city*|*cue*) is the probability that the city occurs, given the presence of the attribute. When dividing this conditional probability by *P* (*city*), we get a measure for how much more likely the city is to occur when the attribute is present than when it is not. The associative strength between the attribute and the city is the logarithm of this odds ratio:

$$S_{cue,i} = \ln \frac{P(city|cue)}{P(city)} = \ln \frac{\frac{P(city \cap cue)}{P(cue)}}{P(city)} = \ln \frac{P(city \cap cue)}{P(city)*P(cue)}. \tag{8}$$

We estimate the probability of a city or attribute occurring, (*P* (*city*), *P* (*cue*)) by dividing the frequency of its occurrence in the knowledge base (N_{city}, N_{cue}) by the total size of the knowledge base N_{all}. We approximate the size of the knowledge base by the number of hits for pages in Wikipedia on which the word *and* appears (N_{and}), so we can write:

$$S_{cue,i} = \ln \frac{\frac{N_{\text{city\&cue}}}{N_{\text{and}}}}{\frac{N_{\text{city}}}{N_{\text{and}}} * \frac{N_{\text{cue}}}{N_{\text{and}}}} = \ln \frac{N_{\text{city\&cue}} N_{\text{and}}}{N_{\text{city}} N_{\text{cue}}}. \tag{9}$$

It can be shown mathematically that $S_{city,i} = S_{cue,i}$. Assuming that the attention weights W_j from the *m* sources of activation sum to 1 (cf., Anderson, Reder, & Lebiere, 1996) and activation spreading from the city and the attribute with equal proportions, the total activation for chunk *i*, as estimated from the web, can be written as:

$$A_{i,web} = c_0 + \ln N_{\text{city\&cue}} + \ln \frac{N_{\text{city\&cue}} N_{\text{and}}}{N_{\text{city}} N_{\text{cue}}}. \tag{10}$$

We assume the memory activation A_i resulting from encounters with information in a person's environment to be a linear function of the activation $A_{i,web}$ as estimated from web frequencies:

$$A_i = c + b\, A_{i,web}. \tag{11}$$

The parameters *c* and *b* serve as scaling parameters describing the unknown relation between how often a person encounters an object in her environment and the web frequency of the corresponding search term.

Given these assumptions, the formula approximating memory activation for chunk *i* by web frequencies of corresponding search terms *N* can be written as:

$$A_i = c + b \left(\ln N_{\text{city\&cue}} + \ln \frac{N_{\text{city\&cue}} N_{\text{and}}}{N_{\text{city}} N_{\text{cue}}} \right). \tag{12}$$

Retrieval Probability & Retrieval Latency of Knowledge

We use the chunk's activation estimated from web frequencies to predict our participants' *retrieval probabilities* of city-attribute associations according to Equation 5:

$$p_i = \frac{1}{1+e^{\frac{-(c+b\left(\ln N_{\text{city\&cue}} + \ln \frac{N_{\text{city\&cue}} N_{\text{and}}}{N_{\text{city}} N_{\text{cue}}}\right) - \tau)}{s}}}, \tag{13}$$

as well as corresponding *retrieval times* according to Equation 6:

$$T_i = Fe^{-\left(c+b\left(\ln N_{\text{city\&cue}} + \ln\frac{N_{\text{city\&cue}}N_{\text{and}}}{N_{city}N_{cue}}\right)+\varepsilon\right)}, \quad (14)$$

where τ is the *retrieval threshold* and ε is the *retrieval noise* generated from a logistic distribution with a mean of zero and a variance of $\sigma^2=\frac{\pi^2}{3}s^2$. In our model, we assume noise not only in the activation level but also in the retrieval threshold (cf., Marewski & Schooler, 2011), where the total retrieval noise parameter, *s*, is an aggregate of the *criterion noise parameter*, s_τ, and *activation noise parameter*, s_A, so that

$$s = \sqrt{s_\tau^2 + s_A^2}. \quad (15)$$

The response times are assumed to be the sum of *perceptual-motor times*, *I*, such as visual encoding and key pressing, and memory retrieval time:

$$RT_i = I + T_i. \quad (16)$$

Empirical Data

Participants

One hundred twenty-eight (Exp. 1) and 73 subjects (Exp. 2) participated in an experiment conducted at the University of Lausanne, Switzerland. Participants received a flat fee of CHF 5 ($ 5.14) plus a bonus of up to CHF 33 ($ 33.90) depending on the consistency of their responses in the main task and a short control task at the end of the experiment.

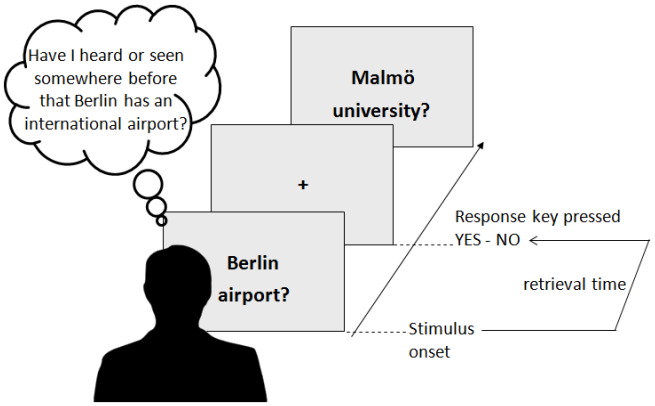


Figure 1: Illustration of the city-attribute task. Attributes were *international airport*, *university*, *high-speed train line*, *major league soccer team*, *company*, *underground*, *cathedral*, and *harbor* (Exp. 1).

Stimuli and Procedure

We assessed retrieval rates and response time distributions for people's knowledge about 95 European cities on 8 attributes in Experiment 1 and 7 in Experiment 2. The difference between Experiment 1 and Experiment 2 was that we dropped the attribute *harbor* from the list of attributes for which knowledge was tested. Specifically, participants saw city-attribute pairs, one at a time (Figure 1). Participants were asked to respond by key press either with "Yes" (they could remember having heard or seen somewhere before that the city possessed the attribute) or with "No" (they could not recall any such instances). For each trial, we recorded both (i) subjects' responses and (ii) the time that elapsed between the presentation of a city-attribute pair and a keystroke. Additionally, for each city, we asked subjects whether they recognized the city name and whether they knew anything else about the city. In total, subjects responded to 950 (Exp. 1) or 855 (Exp. 2) trials.

Model Fits and Predictions

We fitted the parameters of the memory model for chunks encoding knowledge about cities to the data from Experiment 1. Leaving these parameters fixed, we used our model to predict (i.e., for a different set of participants) memory performance in Experiment 2.

Model Calibration on Experiment 1

Post-hoc, the cities "Nice" and "Derby" were excluded because web frequencies also included results for the adjective "nice" and the sport "derby". Also all Swiss cities were excluded from the list because knowledge about these cities reflected personal experience rather than knowledge acquired through the media. 88 cities were included in the final sample. To calibrate the model, we first fit Equation 13 to the observed retrieval rates from Experiment 1. We set the total retrieval noise *s* to the value (.83) used by Marewski and Schooler (2011) and anchored the activation scale by setting the expected value of the retrieval criterion distribution, τ, to zero, so that an object with an activation of 0 would have a 50% chance of being retrieved (cf. Marewski & Schooler, 2011). With a simple regression conducted on the log-odds form of Equation 13, we estimated the constant *c* (-6.11) and the scaling parameter *b* (.69). The Pearson correlation between empirical retrieval rates and simulated retrieval probabilities is r = .72.

With these parameters fixed, in a second calibration step, we fit Equation 16 to the response time distributions for successful retrievals ("Yes" responses) in Experiment 1. Of course, response latency is not a perfect proxy for retrieval time. The total response time includes other components such as the time it takes to read a word and the time to press a key. To model these non-retrieval times, we assume response times are the sum of retrieval times plus perceptual-motor times (Equation 16). We model perceptual-motor times by drawing from a uniform distribution with boundaries of $t \pm t/2$, where *t* is set to the mean time as simulated by the ACT-R production rules necessary for performing the task excluding the memory retrieval (1.01 s).

Subsequently, we fit the latency factor *F* (.80) and the criterion noise parameter s_τ (.67) to the response time distributions of the 704 items of the city-attribute task in Experiment 1. This, as implied by Equation 15, fixes the activation noise parameter to s_A = .49. We did so by minimizing the sum of maximum vertical distances between the empirical and predicted cumulative response time distributions (cf., Voss, Rothermund, & Voss, 2004),

weighted by the empirical retrieval rates for each of the items. The simulation calculates, for each item, the expected value of retrieval time from the proportion of an item's activation distribution that falls above a retrieval criterion sampled from the retrieval criterion distribution. To simulate response time distributions, we took a total of 10,000 samples from the retrieval criterion distribution per item. In sum, four parameters were estimated to predict retrieval probability and retrieval latency. In addition, four parameters were fixed (i.e., not fitted to the data). Table 2 gives an overview of the parameters and their values.

Table 2: Parameters of the memory model

Parameter	Value
Parameters estimated from retrieval probabilities	
Constant, c	-6.11
Scaling parameter, b	.69
Parameters estimated from response time distributions	
Latency factor, F	.80
Criterion noise parameter, s_τ	.67
Fixed parameters	
Total retrieval noise parameter, s	.83
Expected value of retrieval criterion distribution, τ	0
Activation noise parameter, s_A	.49
Expected value of the perceptual-motor time distribution	1.01 s

We calculated medians of the empirical and simulated response time distributions excluding city-attribute pairs for which less than three participants responded with "Yes". We then smoothed the empirical and simulated medians with a running window of size five. The weighted (by the number of "Yes" responses) correlation between empirical and simulated smoothed median response times is r = .62.

Model Predictions for Experiment 2

Leaving these parameter values unchanged, we predict memory performance in Experiment 2. Figures 2 and 3 show the predicted and observed retrieval rates and response time distributions, respectively.

Figure 2 plots retrieval as a function of activation. The points represent the empirical retrieval rates (proportion of "Yes" responses), the S-shaped curve shows the predicted retrieval probabilities based on Equation 13. The Pearson correlation between empirical retrieval rates and predicted retrieval probabilities is r = .72.

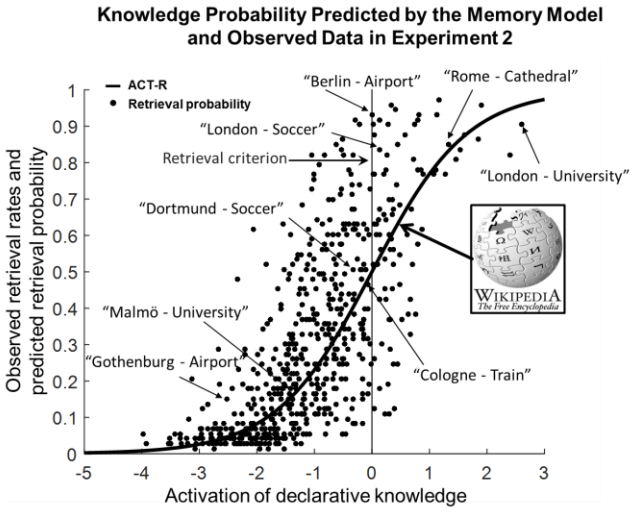


Figure 2: Observed retrieval rates and predicted retrieval probabilities for knowledge about 88 cities (Exp. 2) computed over 73 participants. Retrieval rates are plotted as a function of the expected value of the knowledge activations for 616 city-attribute pairs. The vertical line shows the expected value of the retrieval criterion.

Figure 3 plots response times for positive responses ("Yes") given to the city-attribute task of Experiment 2 as a function of the corresponding chunk's expected value of activation. The points represent the empirical quartiles of response time distributions, the solid lines show the quartiles of predicted response time distributions based on ACT-R's retrieval mechanism (Equation 14). As can be seen, while generally increasing with decreasing activation, median response times are not a simple monotonic function of a chunk's expected value of activation. Chunks will be retrieved when their momentary activation exceeds the retrieval threshold. As we assume noise in a chunk's activation as well as in the retrieval criterion, chunks with a low expected value of activation sometimes exceed the retrieval criterion, at a momentary activation that is likely higher than the expected value of their activation. For that reason, predicted response times flatten out towards the lower end of the activation scale. As Figure 3 shows, our memory model is able to capture the increase in median and spread of response time distributions with decreasing activation of memory chunks. Response time distributions based on a low number of "Yes" responses are noisier and less well predicted by our memory model than those calculated from a high number of responses. Excluding city-attribute pairs for which less than three participants responded "Yes", the weighted (by the number of "Yes" responses) correlation between empirical and predicted smoothed median response times is r = .34.

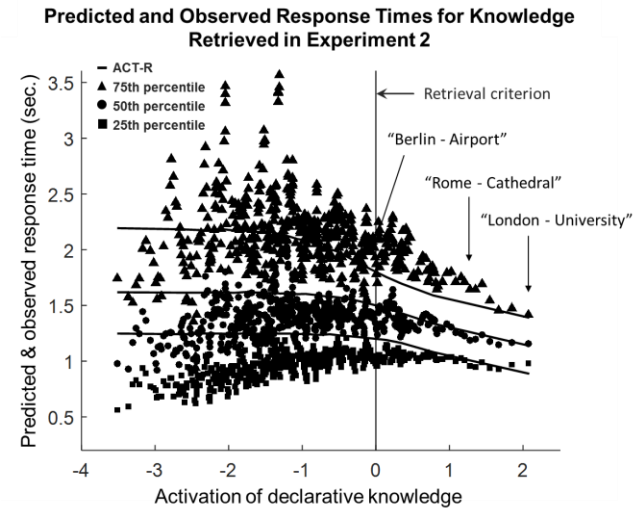


Figure 3: Dots show the 25th, 50th, and 75th percentiles of measured response times for the 572 city-attribute pairs of Experiment 2 where at least three participants responded with "Yes" in the city-attribute task. The solid lines show the quartiles of response times predicted by the model. The x-axis plots the expected value of the chunk's activation, as derived from web frequencies. Response times and activations are smoothed with a moving average. The vertical line shows the expected value of the retrieval criterion.

Application to Decision Making

We extend previous efforts to populate the contents of ACT-R's declarative memory with records that reflect the associated objects' statistical patterns of occurrence in the real world (Marewski & Schooler, 2011; Salvucci, 2014). We believe modeling efforts of this kind lend themselves to many possible applications. In what follows, we illustrate just one.

Much research in the cognitive and decision sciences has explored how people infer objects' (e.g., cities') values on unknown criteria (e.g., population size, wealth) from the objects' attributes, used as cues. Within the *fast-and-frugal heuristics* research program (e.g., Gigerenzer et al., 2011) several strategies describing how people make such inferential decisions have been suggested. The *take-the-best heuristic* (Gigerenzer & Goldstein, 1996), for example, retrieves knowledge about objects' attributes in order of strength of relation to the criterion. This strength of relation is measured as *cue validity*, or as the probability that a city A has a higher value on the criterion (e.g., population) to be inferred than city B, given that A has a positive value on a cue (e.g., has a university) and B a negative or unknown value on that cue (e.g., has no university). Starting with the most valid cue, take-the-best prescribes comparing objects successively on cues in order of decreasing validity, until one cue is identified that allows for making a decision.

While there is evidence that people actually use strategies like take-the-best (e.g., Bröder & Gaissmaier, 2007; Walsh & Gluck, 2016), and are able to adapt their strategy choice to the statistical structure of the environment (e.g., Rieskamp & Otto, 2006), what is known as the *strategy selection problem* remains a serious modeling challenge in the cognitive decision science and beyond (see Marewski & Link, 2014 for an overview).

In addressing that modeling challenge, one strand of research explores how environments are reflected in the memory system, that is, how statistical properties of the environment translate into retrieval probabilities and retrieval latencies of decision-relevant information. In interaction with the memory system, so the rationale goes, the environment carves out for each strategy a *cognitive niche* (Marewski & Schooler, 2011). In so doing, that interplay likely restricts the consideration set of strategies that can be applied to make a decision. Second, among the set of *applicable* strategies, currencies like the strategies' speed of execution, required effort, and accuracy influence selection.

The memory model introduced in this paper simulates which knowledge a person will likely retrieve when confronted with a decision problem. In doing so, the model generates knowledge which can serve as input for different decision strategies. Given the rules prescribed by a particular strategy, one can make predictions on how a strategy will operate, based on the input provided by the memory model. In this way, the model aids exploring whether a strategy will be applicable, how much effort executing that strategy will require (e.g., the number of cues that must be retrieved before a decision can be made), and how accurate the resulting decisions might be.

To illustrate this, Figure 4 explores the niche of the take-the-best heuristic: Panel A depicts the probability of applicability of this heuristic, B the mean cue validity of the discriminating cue, and C the mean accuracy across paired comparisons of 88 cities included in Experiments 1 and 2. Cue validities were calculated from the actual attributes of the cities for a comparison of city size. The probability of attribute-knowledge retrieval was simulated based on the memory model calibrated to the retrieval rates observed in Experiment 1. The cities have been grouped into 22 equally sized bins according to their rank of environmental frequency (approximated by web frequencies).

As can be seen in Panel A, the probability that take-the-best can make a decision increases with the environmental frequencies of both cities. This relationship is paralleled by the effort required to execute take-the-best (Panel B): Fewer cues need to be checked (i.e., the discriminating cue is of high validity) as the environmental frequencies of the cities increase. In areas where both cities are of low environmental frequency, the applicability of take-the-best is at its lowest, and in the cases where that heuristic is applicable, it needs to examine several cues before a decision can be made. As one might expect, the heuristic's accuracy (Panel C) generally rises with the validity of the discriminating cue. However, accuracy is low when both cities have about the same environmental frequency.

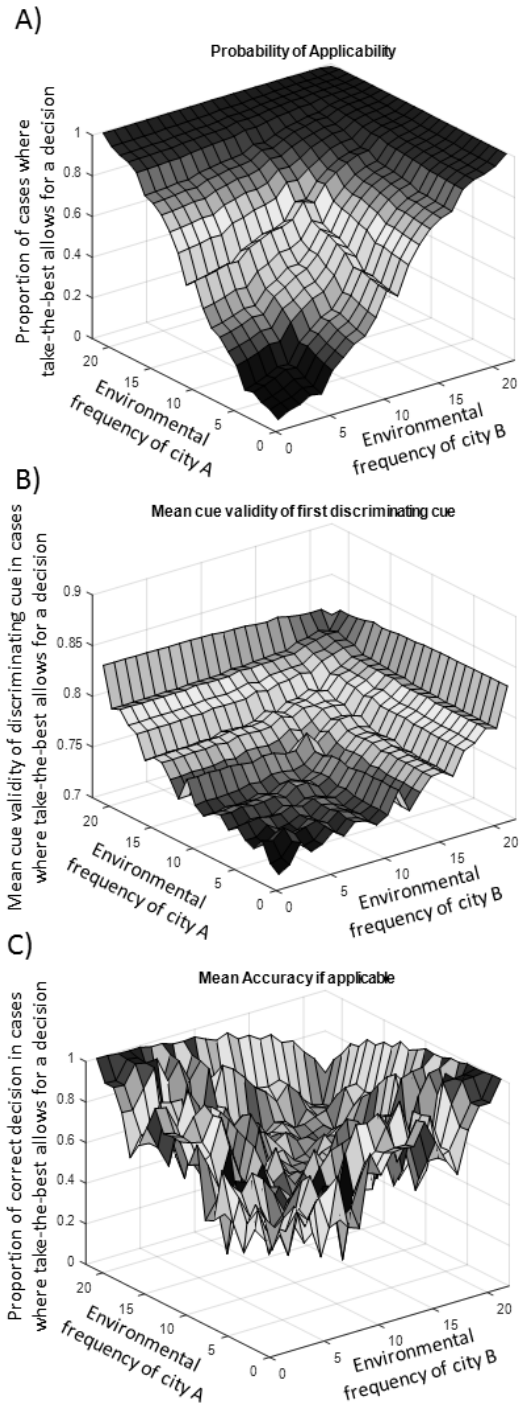


Figure 4: Simulation of inferences about city size made by the take-the-best heuristic. The 88 cities are grouped into 22 equally-sized bins according to their rank of environmental frequency. Bin numbers are shown on the horizontal axes. The vertical axis shows the mean applicability (A), cue validity of the first discriminating cue (B) and accuracy of inferences (C) across 10,000 simulated subjects for an exhaustive pairing of cities within each of the bins. Note that these simulations are exploratory.

Outlook and Conclusion

We are working on implementing simulations of memory-based inferences to, eventually, predict when people will use which decision strategy in a given environment. We hope that such modeling efforts will, one day, invite insights into how the environment, in interaction with the memory system, aids adaptive strategy selection.

Acknowledgments

This work was supported by a grant from the Swiss National Science Foundation [100014_144413].

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703–719.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review,111,* 1036–1060.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology, 30*, 221–256.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multi-attribute decisions. *Psychonomic Bulletin & Review,14*, 895–900.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York: Oxford University Press.

Marewski, J. N., & Link, D. (2014). Strategy selection: An introduction to the modeling challenge. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*, 39–59.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*, 393–437.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207–236.

Salvucci, D. D. (2014). Endowing a cognitive architecture with world knowledge. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1353-1358). Austin, TX: Cognitive Science Society.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32*, 219–250.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206-1220.

Walsh, M. M., & Gluck, K. A. (2016). Verbalization of decision strategies in multiple-cue probabilistic inference. *Journal of Behavioral Decision Making, 29*, 78-91.

Conversational expectations account for apparent limits on theory of mind use

Robert X. D. Hawkins, Noah D. Goodman
{rxdh,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Theory of mind is a powerful cognitive ability: by the age of six, people are capable of accurately reasoning about others' beliefs and desires. An influential series of language understanding experiments by Keysar and colleagues, however, showed that adults systematically failed to take a speaker's beliefs into account, revealing limitations on theory of mind. In this paper we argue that these apparent failures are in fact successes. Through a minimal pair of replications comparing scripted vs. unscripted speakers, we show that critical utterances used by Keysar and colleagues are uncooperative: they are less informative than what a speaker would actually produce in that situation. When we allow participants to naturally interact, we find that listener expectations are justified and errors are reduced. This ironically shows that apparent failures of theory of mind are in fact attributable to sophisticated expectations about speaker behavior—that is, to theory of mind.

Keywords: Theory of mind; social cognition; pragmatics

Introduction

Humans can accurately and intelligently reason about the mental states of other humans. Among other things, this ability – called *theory of mind* (Premack & Woodruff, 1978) – allows us to infer the underlying beliefs and intentions that motivate others' actions, and to use these inferences to predict future actions (Baker, Saxe, & Tenenbaum, 2009). Children acquire this ability by at least age six (Wellman, Cross, & Watson, 2001) and it serves as an important landmark in the developmental trajectory of intuitive theory use (Gopnik & Wellman, 2012). While theory of mind use often appears to be automatic and effortless, Keysar and colleagues (Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; Lin, Keysar, & Epley, 2010) have argued that it is actually the opposite, even for adults: we are "mindblind" by default and only overcome our egocentric biases through an effortful process of perspective-taking. In other words, while adults are *capable* of applying theory of mind reasoning, we do not always apply it reliably.

In this paper we argue that the apparent failures used to support this view are in fact successes for sophisticated social reasoning. In particular, we argue that critical utterances used by Keysar and colleagues are *uncooperative*: they are less informative than what a speaker would actually produce in that situation; listeners who are sensitive to the pragmatics of the situation expect these more informative utterances and produce "errors" when their expectations are flouted.

The argument offered by Keysar and colleagues is based on an elegant experimental paradigm, where participants played a simple communication game with a confederate. The two players were placed on opposite sides of a 4×4 grid containing a set of everyday objects (see Fig. 1). The confederate played the role of 'director,' giving instructions about how to move objects around a grid, and the participant played the role of 'matcher,' attempting to follow these instructions. For example, the objects in one trial included a cassette tape. The director gave an instruction like 'move the tape up one square,' referring to the cassette. Critically, some objects were occluded such that only the matcher could see them, creating an asymmetry in the players' knowledge. To perform accurately on critical trials, the matcher would need to apply theory of mind to reason about which objects were shared and which were private. For example, imagine a roll of tape were placed in an occluded slot: if a participant failed to account for the director's (partial) knowledge, she might interpret 'tape' to mean the occluded roll of tape (which the director couldn't possibly know about). Indeed, Keysar et al. (2003) found that participants attempted to move the hidden item in 30% of cases: 71% of participants attempted to move this hidden item at least once (out of four critical cases) in the experimental condition, compared to 0% in a control condition where there was no ambiguity over the referent. Additionally, eye-tracking data showed that participants considered the hidden item more often and for longer in the experimental condition than the control condition.

While these results are compelling, the paradigm has been criticized from few different angles. Heller, Grodner, and Tanenhaus (2008) have pointed out that in many cases, the hidden object was a better fit for the referring expression than the one in common ground (e.g. the hidden roll of tape vs. the cassette tape for "the tape"), making the hidden object *a priori* more likely to be the referent; it would generally be fairer to compare two objects that fit the referring expression equally well. We validate this argument by empirically measuring relative fit of the expressions to the target and distractor items. Moreover, Hanna, Tanenhaus, and Trueswell (2003) argued that the viewpoint asymmetry paradigm is somewhat unnatural: common ground is typically built incrementally over the course of an interaction rather than presented all at once, and it is rare for a shared display to differ in perceptual accessibility.

In this paper, we offer an additional factor that helps account for Keysar's results. Theory of mind as applied within language understanding depends on an accurate model of what a speaker would say in different situations. Given an utterance, a listener can then reason backward to the most plausible situation (Grice, 1975; Clark, 1996; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). This suggests that we consider whether the utterances produced by the confederate in Keysar's critical conditions were actually what a speaker in that context would be expected to say. If not, then perhaps

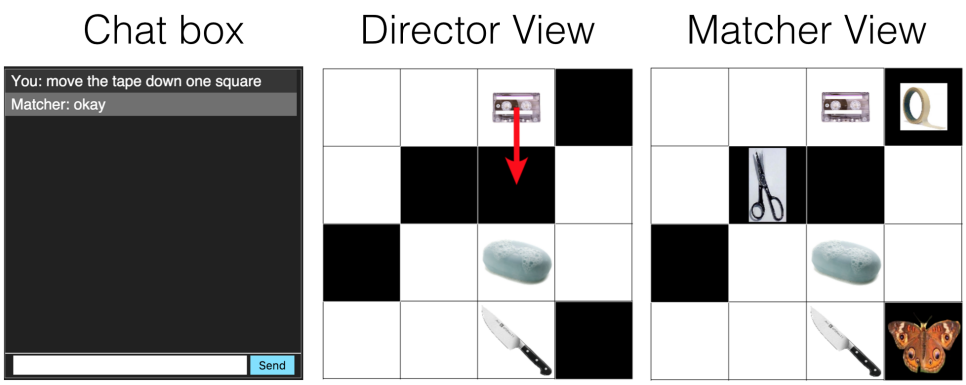


Figure 1: Interface used in the reported experiments. Objects behind the black squares were hidden from the director.

listeners are making choices that are in fact consistent with a correct pragmatic interpretation of the confederate's (uncooperative) utterance. More precisely, when both players know that objects are occluded in the display, the speaker may tend to add additional precision to references in order to avoid confusion. If the listener *expects* the speaker to do this, they will pragmatically pick the *a priori* more likely referent of the referring expression, which in critical trials will be the occluded object. In other words, it is precisely *because* the listener takes the speaker's mental state into consideration that they are tricked by an uncooperative confederate into choosing the wrong item.

We began by replicating Keysar et al. (2003) in a multiplayer web experiment. We recruited participants to be both director and matcher (instead of using a confederate), but instructions for critical items, as well as a random subset of filler items, remained scripted as in the original study. We replicated the original finding, but noted a tendency of directors to be overinformative in unscripted filler trials. We then ran the same experiment without using any scripted instructions, observing unconstrained director utterances. We found much greater precision in unconstrained director utterances, which match targets much better than distractors, and better performance of the matchers. This minimal pair of experiments demonstrates that listener mistakes are at least partially due to the pragmatics of the task, ironically showing that apparent failures of theory of mind are in fact attributable to sophisticated expectations about speaker behavior—that is, to theory of mind.

Expt. 1: Scripted Replication

Participants

We recruited 34 participants (17 pairs) from Amazon Mechanical Turk. All participants were from the U.S. Three pairs were excluded for making 2 or more errors on non-critical items.

Materials & Procedures

Participants interacted in a real-time, multi-player environment on the web (Hawkins, 2015). Pairs of participants—assigned randomly to 'director' and 'matcher' roles—interacted with one another through a web interface, shown in Fig. 1. On the left side of the screen, participants could freely type messages to one another; on the right side the screen, players could view a set of objects placed in a 4 x 4 grid. Five of the grid cells were occluded from only the director's perspective, and the remaining 11 were visible to both matcher and director. Six or seven objects were displayed in the grid at a given time. One of these objects was a 'target', such as a cassette tape, placed in an unoccluded cell such that both participants could see it. Another object, such as a roll of tape, was placed in an occluded slot such that it was only visible to the matcher. The rest of the objects were unrelated 'fillers' placed in random locations. We used the same set of targets and occluded alternatives as Keysar et al. (2003), but we were unable to obtain the filler objects from the original experiment and created our own. Before entering the game environment, every participant independently passed a short quiz about the task's instructions, ensuring that they understood the interface. Among other items on the quiz, we verified that both participants understood that items behind black cells were only visible to the matcher.

The experiment was composed of eight items, with each item using a different set of objects. Each item included one 'critical pair' of objects, one of them the target and the other hidden, such as the cassette tape and the roll of tape. For each item, we gave the director a series of four instructions to move objects around, which were displayed as a series of arrows pointing from some object to an unoccupied cell. To collect clean mouse-tracking data, we began every instruction by asking the matcher to click a small circle in the center of the grid. After this small circle was clicked, the director was allowed to communicate the next instruction and we started recording from the matcher's mouse. One of the instructions was a 'critical instruction,' which referred to the target object. For half the instructions, directors were free to communicate however they wished. For the other half, including all the critical instructions, their messages to the matcher were prescripted using the precise wording from Keysar et al. (2003). For example, when giving instructions on how to move the cassette tape, the director would be forced to use the ambiguous utterance "Move the tape down one square." (That is, in

	% attempted at least once			% attempted at least twice			% of total cases		
	Orig.	Expt. 1	Expt. 2	Orig.	Expt. 1	Expt. 2	Orig.	Expt. 1	Expt. 2
Experimental	71	93	61	46	57	32	30	43	24
Baseline	0	7	0	0	0	0	0	2	0

Table 1: Side-by-side comparison of error rates in Keysar et al. (2003) and our two replications. The first two sections show the percentage of *participants* attempting to move the occluded distractor at least once, or twice, of the four possible cases. The third section shows the percentage of *all experimental trials* that the participant actually tried to move the occluded object.

these conditions, the scripted message would automatically appear in the director's chat window, and they would have to click 'send' for the experiment to continue.)

We collected baseline performance for each condition by replacing the hidden alternative (e.g. a roll of tape) with an object that did not fit the critical instruction (e.g. a battery); we used the same unambiguous replacements as Keysar et al. (2003). Each participant received half the items in the experimental condition and half in the baseline condition. The assignment of items to conditions was randomized across participants, and the order of conditions was randomized under the constraint that the same condition would not be used on more than two consecutive items. All object sets, object placements, and corresponding instruction sets were the same for all participants.

This paradigm differs from those used by Keysar et al. (2003) in three primary ways. First, participants were not seated across from each other at a table: they each saw a view of the 4 x 4 grid on their screen and communicated via a text box. Second, we did not use a trained confederate. We randomly assigned one of the players to the role of the instruction giver, and maintained the original wording by scripting a subset of their instructions. Finally, the hidden object was not placed in a bag, in which respect our design more closely resembles Keysar et al. (2000).

Results and discussion

In Table 1 we show the error rates on critical items, and compare to the data from Keysar et al. (2003). We find that 93% of participants (all but one) attempted to move the hidden distractor at least once in the Experimental condition, out of four possible items, compared to only 7% (only one) in the baseline condition. This is similar to the effect observed by the authors in the original study, which found 71% and 0%. Our errors were larger across the board, perhaps due to the interface or the population, but the gap between the two conditions is roughly the same size. In Table 2, we break down the pattern of errors by item. We note that several items have much higher error rates than others – for example, 75% of participants in the experimental condition of item 6 made an error (the "whiteboard eraser" vs. the "pencil eraser") while only 17% of participants in item 8 made an error (the "computer mouse" vs. the "toy mouse"). Informally, it seems as though the more difficult items are the ones where the utterance fits the distractor better than the target. We empirically substantiate this observation in our results for Expt. 2 below.

This item-wise variability suggests that the dependent variable highlighted in the original study (i.e. "percentage of participants who moved the critical item at least once") is somewhat problematic: it could look like 100% of participants made errors even if they all made those errors on one particularly difficult item. Indeed, if we exclude the three 'hard' items where over 60% of participants in the experimental condition made errors, this dependent variable drops from 93% to only 43% of participants.

As a proxy for the eye-tracking analyses reported by Keysar et al. (2003), we conducted a mouse-tracking analysis. We define the decision window as the span of time between the point when the matcher received their instruction message and when they started moving an object. If it took them multiple attempts to move the correct object, we restricted our analysis to the first attempt. Within the decision window, we computed the total amount of time spent hovering over the cell containing the target and divided by the total length of the decision window to get a measure of the relative time spent considering the target. We had to exclude an additional 3 participants for this analysis, because the timestamps for director and matcher did not align and we could not establish the decision window properly. A paired-samples t-test found that people tended to spend less time hovering over the target cell on critical experimental trials than on baseline trials, $t(10) = -2.65, p = 0.02$ (see Fig. 2), indicating that the presence of a hidden distractor interfered with participants ability to directly choose the target.[1] .

By running this replication as a multi-player web experiment we have available an additional source of data beyond the original experiments: half of the instructions were unscripted, providing observations of natural production of referential descriptions for filler items. Informally, we noted a tendency toward additional, possibly unnecessary, precision in descriptions. Instead of "move the stuffed animal down", participants said "move the stuffed panda bear down." Or, instead of saying "move the plane to the right" when there is only one plane, participants said "move the red airplane to the right." Perhaps directors were taking the time to make more precise descriptions because they believed it was contextually relevant: both parties know that there are hidden objects in the environment increasing the chance of miscommunication from imprecise descriptions. If the matcher expected the di-

[1] This analysis includes participants who actually made errors, since the data is too sparse to exclude them.

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
	instruction	"glasses"	"bottom block"	"tape"	"large measuring cup"	"brush"	"eraser"	"small candle"	"mouse"
	target	sunglasses	block (3rd row)	cassette	medium cup	round hairbrush	board eraser	medium candle	computer mouse
	hidden distractor	glasses case	block (4th row)	scotch-tape	large cup	flat hairbrush	pencil eraser	small candle	toy mouse
Expt. 1	# incorrect	0	5	1	3	2	6	6	1
	# correct	6	3	1	8	2	2	4	6
Expt. 2	# incorrect	0	1	1	3	9	7	1	5
	# correct	12	14	11	13	7	8	12	8

Table 2: Item-wise error rates for critical trials

rector to be precise, then they would be justified in picking the first or best object that meets the description (rather than worrying excessively about occluded cells). That is, the scripted instructions used by the director for critical trials may have been *uncooperative* for this situation, and thus led matchers astray. We tested this prediction in Expt. 2, where we removed the scripted instructions and allowed speakers to refer to items however they wished. By the reasoning above we expected to see more precise descriptions by unscripted directors and fewer errors by matchers in the critical trials.

Expt. 2: Unscripted Replication

Participants

We recruited 64 participants (32 pairs) from Amazon Mechanical Turk, roughly doubling the sample size from Expt. 1. All participants were from the U.S. Three participants were excluded for making 2 or more errors on non-critical items, and one additional participant was excluded because they were not a native English speaker.

Materials & Procedures

Everything was the same as Expt. 1, except we did not use scripted messages for critical instructions.

Results

Error rates are reported in Table 1, alongside the results from Keysar et al. (2003) and our scripted replication in Expt. 1.

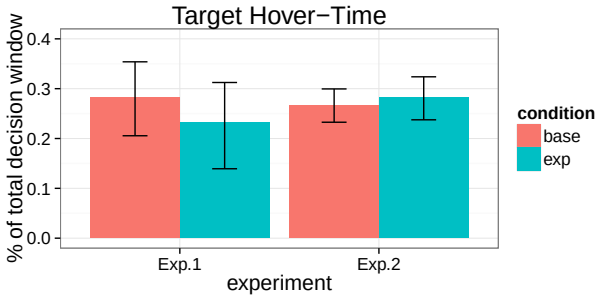


Figure 2: Mean percentage of time spent hovering over target cell in the two conditions for Expt. 1 (left) and Expt. 2 (right). Error bars are boostrapped 95% confidence intervals.

Participants never moved the hidden object in the baseline condition, and total error rates for experimental trials are significantly lower than the rates found in Expt. 1, $\chi^2(1) = 5.35, p = 0.02$.

We find that patterns of errors in Expt. 2 diverge significantly from a uniform distribution across items, $\chi^2(7) = 24.8, p < 0.001$. Looking more closely at these patterns, we see that the only items where errors are consistently made are those where the more precise utterances used remain ambiguous. For example, in item 5, both the target and distractor are hair brushes: one is round and one is flat. Many participants produced the sub-class label "hair brush," which was more precise than the scripted basic-level "brush" from Expt. 1, but the two objects were still confusable at the sub-class level. If we remove the two most difficult items (the "hair brushes" and the "erasers"), the total percentage of errors on experimental trials drops from 24% to 10% and the percentage of participants making at least one error drops from 61% to 32%.

When we conducted a mouse-tracking analysis identical to the one reported for Expt. 1, we found no significant decrease in target hover time between experimental and baseline trials $t(27) = 0.89, p = 0.38$. To directly test for differences in hover time patterns across the two experiments, we used a mixed-effects model with a random intercept for game ID and an interaction between condition and experiment on target hover time. We found a marginally significant interaction, $b = 0.09, t = 1.89, p = 0.066$ (see Figure 2), providing some evidence that the presence of a hidden distractor no longer interfered target selection in Expt 2.

Next, we test whether these improvements in performance are in fact due to more informative speaker behavior. We recruited twenty judges on Amazon Mechanical Turk, who provided ratings for how well the 71 unique labels used by speakers across both experiments (including scripted labels) fit the target and hidden distractor objects. Their responses were given on a slider with endpoints labeled "not at all" and "perfectly." Inter-rater reliability was relatively high, with intra-class correlation coefficient of 0.6 (95% CI = $[0.54, 0.66]$). In a mixed model including random intercepts for raters and items, we found a significant crossover interaction of ex-

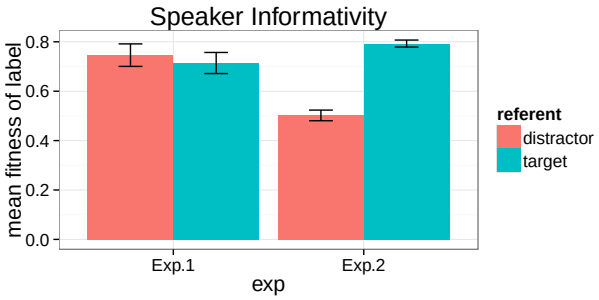


Figure 3: Mean fitness ratings provided by judges on Amazon Mechanical Turk. Error bars are bootstrapped 95% confidence intervals.

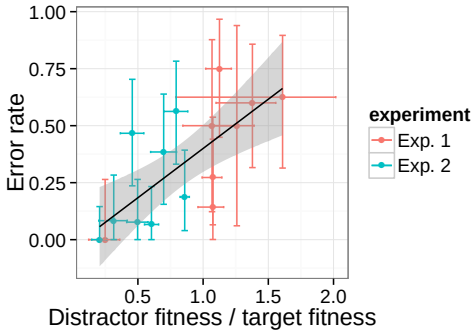


Figure 4: Item-wise error rates compared with label fitness ratios, across both experiments. Error bars are bootstrapped 95% confidence intervals for fitness and 95% highest posterior density intervals for error rates.

periment and referent on mean fitness rating (see Fig. 3), $b = 0.32, t = 9, p < 0.001$. In Expt. 1, the scripted label fit the hidden distractor just as well or better than the target, but in Expt. 2, the unconstrained labels fit the target much better and the hidden distractor much worse. In other words, the scripted labels used in Keysar's studies were less informative than speakers normally produce in this scenario.

Does differential informativity of scripted utterances account for the variability across items that we noted in Expt. 1? In Figure 4, we compare the item-wise error % and ratio of target fit to distractor fit. We find that across both experiments, participants have significantly higher error % on items where the speaker's label fits the distractor better than the target, $b = 0.43, t(14) = 3.97, p = 0.001$, capturing a significant portion of variance, $R^2 = .5, F(1,14) = 0.001$. This suggests that item-wise variability across the two experiments is primarily driven by relative informativity of speaker utterances.

After separately establishing that matchers in Expt. 2 make fewer mistakes, that directors in Expt. 2 produce more informative utterances, and that informativity captures item-wise variability in error rates in both experiments, we tested the link between these effects in our aggregated data: does label informativity generally predict errors on critical trials? We used a mixed effects logistic regression model to estimate the effect of target and distractor fit on the probability of making a critical error in both experiments, including a random intercept for game ID. We found that participants are less likely to make errors when the target fit is higher, $b = -0.8, z = -4.1, p < 0.001$ and more likely to make errors when the distractor fit is higher, $b = 1.7, z = 3.7, p < 0.001$. Furthermore, a model including target fit and distractor fit in addition to item-level fixed effects is significantly better than a model including item alone, $\chi^2(2) = 36.2, p < 0.001$, implying that speaker informativity captures residual variance beyond the item-wise effects reported above.

General Discussion

Pragmatic language understanding requires sophisticated social reasoning. To interpret an utterance, a listener must consider how a speaker is likely to behave in context. The rational use of theory of mind for a listener thus depends on her expectations about the speaker: If, in a particular context, she expects a speaker to provide sufficiently informative utterances, she may be justified in neglecting his epistemic state when resolving reference.

In our replication (Expt. 1), we found evidence that listeners neglect the speaker's epistemic state, as Keysar and colleagues claim. We also found that when the speaker's utterance fit the hidden distractor better than the target, matchers were more likely to make errors. Indeed, we found that the extremely heterogeneous pattern of errors across items was well explained by the relative fit of the utterance to target and distractor. This suggests that reference disambiguation was driven primarily by *a priori* label fitness rather than consideration of occluded vs. mutual visible items. However, this does not necessarily imply limitations on theory of mind—if speakers could be expected to naturally provide expressions which apply better to the target, then this behavior would be appropriate. In Expt. 2, we found that speakers did naturally produce more precise, informative utterances than required and these unconstrained utterances fit the target significantly better than they fit the hidden distractor. For example, no speaker in Expt. 2 produced "the bottom block," which was used as a critical instruction in Expt. 1. Instead, they said "the bluish block with a B" or "block with the blue writing," which relied on less confusable perceptual features and thereby decreased error rates. Thus, the errors observed in Expt. 1 can be explained as a result of an uncooperative confederate (speaker)—the experiment set up certain pragmatic expectations through the task context, then deliberately flouted them in critical trials.

The Gricean maxim of quality dictates that cooperative speakers should make their utterance as informative as is required for current purposes, but no more so. If the referring expressions used in Expt. 1 uniquely pick out the target from the speaker's perspective, then why would unscripted speakers in Expt. 2 be more informative than this? One possibility is that awareness of the complex mismatch in epistemic state with the listener leads the speaker to provide extra information in an attempt to avoid miscommunication. Another possibility is that more general dynamics of language are in play. Studies of over-informativity in referring expressions

have uncovered a general tendency of speakers to provide redundant information. This tendency can depend on a number of situational factors, such as the tendency to mention perceptually salient features to speed up the identification process (Koolen, Gatt, Goudbeek, & Krahmer, 2011). The exact origin is a subject that must be followed up by future research (e.g. Gann & Barr, 2014), but it is clear that the tendency of speakers to produce highly informative referring expressions is useful to, and relied on by, listeners.

This finding is consistent with a recent proposal by Heller, Parisien, and Stevenson (2016) that referring expressions are interpreted by probabilistically integrating multiple sources of information: when conversational expectations lead participants to expect over-informative utterances, they (rationally) place relatively less weight on features of the environment determining common ground, such as shared perceptual access. Further work in a wider variety of tasks is necessary to pin down the various factors determining the relative weighting of these different sources of information, but we have argued that pragmatics play a crucial role.

It is worth noting several significant differences between our study and Keysar et al. (2003). The primary difference between our study and the original, of course, is that it was run on the web with participants connected through a virtual environment, instead of face-to-face in a room. We believe we addressed the major concern about exploring theory of mind in web experiments – that participants do not truly believe they are interacting with another human – by allowing instantaneous, responsive, real-time interaction. On the other hand, it is known that textual communication, as in our chat box, can differ from face-to-face verbal communication. Additionally, aspects of the interface such as the graphical representation of occluded cells may be less intuitive, or require more training, on the web than in the lab.

A related difference is our decision not to use a confederate. While confederates are useful for reducing variation across instances of the experiment and delivering carefully targeted manipulations, their use may have unexpected consequences. Beyond the difficulties of conducting large-scale experiments with confederates, it is difficult to exactly replicate all the subtleties of the confederate's behavior that might influence results. The pair of experiments we report is a reminder that manipulations administered by a trained confederate can interact in unexpected ways with a participant's social and communicative expectations. Regardless of the experimental context, it is illuminating to see how real participants naturally interact.

Acknowledgements

We're grateful to Boaz Keysar for providing select materials for our replication. Expt. 1 was originally conducted under the supervision of Michael Frank, with early input from Desmond Ong. This work was supported by ONR grants N00014-13-1-0788 and N00014-13-1-0287, and a James S. McDonnell Foundation Scholar Award to NDG. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Clark, H. H. (1996). *Using language*. Cambridge university press Cambridge.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gann, T. M., & Barr, D. J. (2014). Speaking from experience: Audience design as expert performance. *Language, Cognition and Neuroscience*, *29*(6), 744–760.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085–1108.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (pp. 43–58). New York: Academic Press.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25 - 41.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551–556.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(04), 515–526.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 655–684.

Simpler structure for more informative words: a longitudinal study

Uriel Cohen Priva (uriel_cohen_priva@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, RI 02912, USA

Emily Gleason (emily_gleason@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, RI 02912, USA

Abstract

As new concepts and discoveries accumulate over time, the amount of information available to speakers increases as well. One would expect that an utterance today would be more informative than an utterance 100 years ago (basing information on surprisal; Shannon, 1948), given the increase in technology and scientific discoveries. This prediction, however, is at odds with recent theories regarding information in human language use, which suggest that speakers maintain a somewhat constant information rate over time. Using the Google Ngram corpus (Michel et al., 2011), we show for multiple languages that changes in lexical information (a unigram model) are actually negatively correlated with changes in structural information (a trigram model), supporting recent proposals on information theoretic constraints.

keywords: information rate, information theory, Google

Introduction

Most of Campbell's condensed soup cans in Andy Warhol's famous 1962 work show between four and seven words on the front of the can. Currently, a typical Campbell's can has more than ten. The "cream of mushroom with roasted garlic" soup is "great for cooking", and the can additionally specifies its net weight, as shown in Figure 1. Campbell now offers more soup varieties, more than 80, compared to only 32 in 1962.



Figure 1: Andy Warhol's Campbell's tomato soup (left), and modern Campbell's cream of mushroom with roasted garlic (right)

On every front, the development of human society is accompanied by information growth. There are more books to read today, objects to use, and apps to download. In information theory (Shannon, 1948), the information encoded in some event is its negative log probability – unpredictable events are more informative. Imagine if every exchange of words in English were recorded. In terms of information theory, some exchanges would be more informative than others. Predictable utterances provide less information than unpredictable ones. Repetitive "how are you"s do not contribute much information, but utterances such as Neil Armstrong's "small step for man" do. It is reasonable to expect that a random sample of exchanges collected today will contain more information than an equal-sized sample collected a hundred years ago: the modern sample may contain *unfollow*, *Higg's boson* and *politically correct* – lexical items, scientific discoveries and social concepts that were first used or discovered within the past 100 years.[1]

If a million word sample collected today contains more information than it did a hundred years ago, the expectation is that information rate (entropy), will be higher as well. However, the prediction that information rate has been rising is incompatible with recent findings in psycholinguistics. It has been proposed that speakers manipulate their speech so that they will not exceed or fall below acceptable information rates, such as by omitting, reducing, or hypo-articulating low-information linguistic material and expanding or hyper-articulating high-information linguistic material (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). Expansion and reduction have been demonstrated for individual segments (Cohen Priva, 2015; R. van Son & van Santen, 2005; R. J. J. H. van Son & Pols, 2003), syllables (Aylett & Turk, 2004), morphemes (Kuperman, Pluymaekers, Ernestus, & Baayen, 2007; Kurumada & Jaeger, 2015; Pluymaekers, Ernestus, & Baayen, 2005), and words (Arnon & Cohen Priva, 2014; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Jurafsky, Bell, Gregory, & Raymond, 2001; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 2011; Seyfarth, 2014). Such effects have even been demonstrated at the edge of clauses (Jaeger, 2010; Levy & Jaeger, 2007; Norcliffe & Jaeger, 2014), suggesting that information theoretic considerations are also driven by syntactic information. Other studies suggest that higher-level syntactic considerations affect the duration of individual words within that construction (Gahl & Garnsey, 2004; Kuperman & Bresnan, 2012). This trend is even suggested in the Campbell's soup example presented in the first paragraph– there are more possible choices of soup, and additional words are re-

[1] This is not to say that some words do not fall out of fashion– see Petersen, Tenenbaum, Havlin, and Stanley (2012).

quired to describe most choices, spreading out the amount of information per symbol.

What factors contribute to the limitations on information rate? Current research considers at least two. First, speakers may be unable to speak faster or provide more information due to speaker-internal limitations, such as time constraints for motor planning or a cap in cognitive ability (Bell et al., 2009; the within-speaker model in Jaeger, 2010, pp. 50–51). Alternatively, the limit may focus on the communication channel – even if speakers are able to produce high information rates, their listeners may be unable to follow what is being said at such rates (Jaeger, 2010; Pate & Goldwater, 2015). Both explanations predict that information rate will not exceed certain thresholds, even if more information does become available.

How can information rate be held constant given an increase in available information? Several factors affect the amount of information provided by speakers. Consider the phrase *it is raining*. The information provided by an utterance as a whole is the negative log probability of observing the utterance, combining the probability of the content and the probability of the structure. One aspect of this is *world knowledge*. If it is -10 degrees outside, then *it is raining* becomes a highly unlikely utterance, whereas *it is snowing* becomes far more likely. Studying this type of information is beyond the scope of this paper, but it is possible to measure *lexical* and *structural* information. Lexical information is derived from the frequency or probability of individual words. The word *precipitating* is less frequent than the word *raining* (despite denoting a larger set of events). The phrase *it is precipitating* is therefore lexically more informative than *it is raining*. Structural knowledge would tell us that *it is raining* is more common than *raining it is*, and so the untopicalized form is more probable and less informative than the second, topicalized form. Therefore, one of the ways language use can change to accommodate the rising amounts of information is by reducing structural complexity. Increase in available information tends to increase the information provided in any utterance, but using more probable (less informative) structures would balance this increase. Does language compensate for the rising availability of information by reducing structural complexity?

We test this hypothesis using a longitudinal study of language by contrasting the entropy of a three-word language model with the entropy of a single-word language model (unigram model, Jurafsky & Martin, 2000). A three-word (trigram) language model determines the probability of the appearance of a word using the expected negative log probability of observing a word given the two preceding words (1). This method is similar to the one used in Genzel and Charniak (2002). A unigram model determines the probability of a word using no context (2). Both models take yearly entropy as the weighted average surprisal of all words in the corpus for that year. If more information is available, the diversity of the lexicon will be higher, but if the language focuses on a more restricted subset of available information, the entropy of the single-word model would drop. In contrast, the three-word model factors in both available information and the structural complexity of the language. If these two values were independent, they should be positively correlated– if context were not available, then the best estimate for the trigram model (1) would be the unigram model (2), and therefore a rise in unigram entropy would predict a rise in trigram entropy. However, we instead expect that one would come at the expense of the other, and that increasing the amount of available information should lead to a reduction in trigram entropy to keep information rate within acceptable ranges.

(1) Trigram entropy

$$E\left[-\log \Pr(\text{word}|\text{two previous words})\right]$$

(2) Unigram entropy

$$E\left[-\log \Pr(\text{word})\right]$$

Another possible hypothesis is that information rate constraints would have no effect on textual data. After all, readers (and writers) can theoretically slow down and speed up as they will, in order to digest (or produce) denser or more informative words and structures (although, see Genzel & Charniak, 2002, who found evidence for *entropy rate constancy* in text). This expectation has the same prediction as the null hypothesis: increase in unigram information rate would lead to a rise in structural information rate. A negative correlation between unigram and trigram entropy would suggest that writers still tend not to exceed some level of information rate.

Methods and materials

The Google Ngram corpus

Historical spoken data was not systematically collected, but written data is available. The Google Ngram corpus (Lin et al., 2012; Michel et al., 2011) provides yearly frequency counts for sequences of words, and has previously been used to study related phenomena, such as the lifecycle of words (Petersen et al., 2012). The corpus contains several subsets that limit the type of word sequences to words that were published in a specific language, or a specific country. For example, the American English subset includes only word sequence counts of English books that were published in the United States. A typical datum in the Google Ngram corpus for the American English subset might contain a three-word sequence, such as "take aerial photographs", followed by two numbers, e.g. "1992 23". This would mean that the sequence *take aerial photographs* appeared 23 times in all the books scanned by Google that were published in 1992 in English in the United States.

We focus on data from the 20th century, for which data is available for the greatest number of languages. We exclude data from 2000 and onwards, as suggested by the authors of corpus (supplementary material of Michel et al., 2011). We

excluded languages that had too little data in the 20th century (Simplified Chinese, Hebrew),[2] and languages for which no single country is dominant (Spanish).[3] English data was split by the corpus into American English and British English, and English was therefore included. The exact same methodology, as detailed below, was replicated for each of the remaining languages: American English, British English, French, German, Italian, and Russian. For both trigrams and unigrams, we excluded words that mixed letters and numbers, as too many of those seemed like data from tables, rather than language use. [4]

Calculating trigram entropy

Trigram surprisal was estimated as the maximum likelihood estimate (MLE) of observing the third word in a three word sequence given all the possible words that could follow the previous two words. For example, to calculate the probability of the word *photographs* appearing in the context *take aerial*, the frequency of *take aerial photographs* is divided with the frequency of *take aerial* followed by any word. The negative log of the probability provides the number of bits the word *photographs* provided in that context. The average number of bits per word is the entropy of the corpus given the model.

MLE estimates were used rather than models incorporating smoothing or backoff (Jurafsky & Martin, 2000, ch. 4), as such methods explicitly integrate information from lower-order n-grams to the probability calculations of trigrams. Thus, they already factor out cases in which a word's frequency is biased by the context in which it appears (e.g. *Francisco* is frequent, but almost always preceded by *San*). The proposed account predicts that new words are likely to be structurally accommodated by facilitating (restrictive) contexts. Switching to smoothed models could mask this effect.

Calculating unigram entropy

For unigram entropy, the first words of each trigram in the trigram model were counted. The first word was chosen since trigrams in the 2012 version of the Google Ngram data do not span sentence boundaries (this is the version used here; Lin et al., 2012), and we did not want to bias the sample towards sentence-final words, which are likely to be less informative if our hypothesis is correct. The surprisal of observing a particular word in any context was taken to be the negative log of the number of times the word was observed, divided by the number of times each word was observed (MLE of word probability).

[2]Hebrew and Chinese had comparable results to the other languages when using only data for the last 30 years of the 20th century.

[3]We did run the study for Spanish, with comparable results to the other languages.

[4]Median total number of trigrams per year after exclusions (in millions): American English: 623.11; British English: 221.1; French: 202.8; German: 151.84; Italian: 54.97; Russian: 82.83. Median number of unique trigrams per year (in millions): American English: 80.22; British English: 35.75; French: 30.11; German: 26.97; Italian: 14.2; Russian: 19.29.

Statistical method

For each language we measured the relationship between trigram entropy and unigram entropy in a linear regression, with trigram entropy as the predicted variable and unigram entropy as the main predictor. The log number of unique trigrams per year, the log of the total number of trigrams in the corpus, and the log number of unique unigrams were used as controls, as well as the log number of volumes and log number of pages that were included in the original Google Ngram corpus. The total number of unigrams was identical to the total number of trigrams, and was not used (as unigrams were taken from the trigram dataset, see previous section). The greater the unique number of trigrams relative to the total number of trigrams, the higher the entropy is expected to be, everything else being equal (as the entropy of a uniform distribution over $n+1$ outcomes is higher than a uniform distribution over n outcomes). There is no concrete prediction for the total number of trigrams as a predictor, but it is expected to capture some of the variance that is associated with having more books (or topics of discussion). The number of unique unigrams expresses an alternative (though less accurate) estimate for the richness of the lexicon than unigram entropy. All the predictors and predicted values are *time series*, and are not considered to be independent from their previous values (e.g. the correlation between trigram entropy in year n and year n-1 is 0.99 for American English). Therefore, the regression used the differences between each pair of consecutive years for all variables. All counts-based controls were logged, as the logged values correlated better with trigram entropy (e.g. pearson r 0.88 for the correlation between log unique number of trigrams and trigram entropy, but only 0.81 for its unlogged counterpart in American English), and therefore consistute more appropriate controls.

Results and discussion

For all languages, changes to unigram entropy were strongly negatively correlated with changes to trigram entropy (American English: β=-0.44, SE=0.05, t=-8.753, $p<10^{-12}$; British English: β=-0.51, SE=0.034, t=-15.05, $p<10^{-15}$; French: β=-0.55, SE=0.055, t=-9.925, $p<10^{-15}$; German: β=-0.46, SE=0.07, t=-6.532, $p<10^{-8}$; Italian: β=-0.76, SE=0.058, t=-13.148, $p<10^{-15}$; Russian: β=-0.55, SE=0.042, t=-12.945, $p<10^{-15}$), suggesting that structural (or transitional) complexity is reduced when lexical complexity rises, even when the size of the corpus is controlled for. Figure 2 plots for American English the partial correlation between changes in unigram entropy and the residual changes in trigram entropy after other predictors were controlled for. Figure 3 shows the equivalent relationship for German, the language with the least significant relationship between unigram entropy and trigram entropy.

For all languages used, the number of unique trigrams was positively correlated with trigram entropy (American English: β=0.6, SE=0.049, t=12.26, $p<10^{-15}$; British English: β=0.71, SE=0.035, t=20.624, $p<10^{-15}$; French: β=0.76,

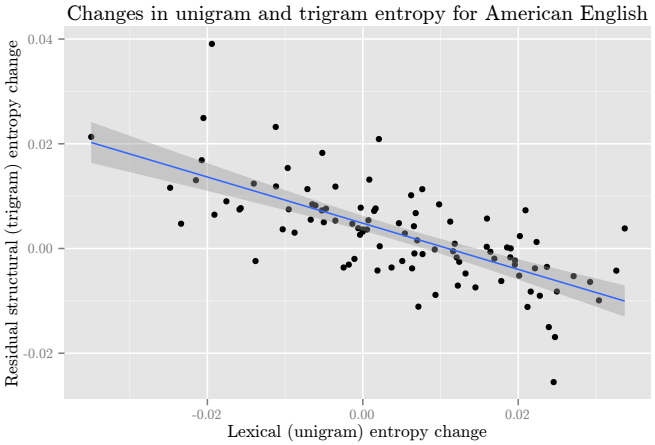


Figure 2: A plot showing the relationship in American English between changes in unigram entropy on the x-axis, and residual changes in trigram entropy on the y-axis, after other predictors are controlled for.

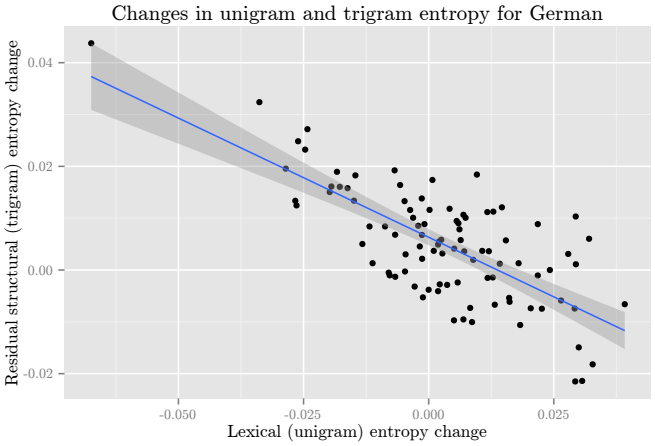


Figure 3: A plot showing the relationship in German between changes in unigram entropy on the x-axis, and residual changes in trigram entropy on the y-axis, after other predictors are controlled for.

SE=0.065, t=11.777, $p<10^{-15}$; German: β=0.86, SE=0.11, t=7.852, $p<10^{-11}$; Italian: β=0.84, SE=0.078, t=10.794, $p<10^{-15}$; Russian: β=1.161, SE=0.093, t=12.537, $p<10^{-15}$), as expected. The total number of trigrams was significantly negatively correlated with trigram entropy in every language except British English (American English: β=-0.24, SE=0.058, t=-4.142, $p<10^{-4}$; French: β=-0.23, SE=0.051, t=-4.531, $p<10^{-4}$; German: β=-0.22, SE=0.054, t=-3.983, $p<0.001$; Italian: β=-0.47, SE=0.05, t=-9.286, $p<10^{-14}$; Russian: β=-0.36, SE=0.057, t=-6.326, $p<10^{-8}$). In all languages except Italian and Russian, the number of unique unigrams had no effect on trigram entropy (positive correlation for Italian: β=0.28, SE=0.12, t=2.376, $p<0.05$; negative correlation for Russian: β=-0.31, SE=0.11, t=-2.687, $p<0.01$). The number of books was positively correlated with trigram entropy for Russian only (β=0.066, SE=0.03, t=2.19, $p<0.05$), and the number of pages was positively correlated with trigram entropy for only Italian (β=0.23, SE=0.065, t=3.454, $p<0.001$). No other languages were affected by book or page count.

In all languages, unigram and trigram entropy change over time. There are clear drops in trigram entropy, e.g. all Western world countries have a drop in trigram entropies in the 1970s, despite an increase in the size of the corpus. This in itself, prior to controlled analysis, is an interesting finding. It would suggest that the acceptable range for information rate may sometimes drop. Figure 4 plots the change in residual unigram entropy over time, controlling for the total number of unigrams and the number of unique unigrams in the corpus. Figure 5 plots the change in residual trigram entropy over time, controlling for the total number of trigrams and the number of unique trigrams in the corpus.

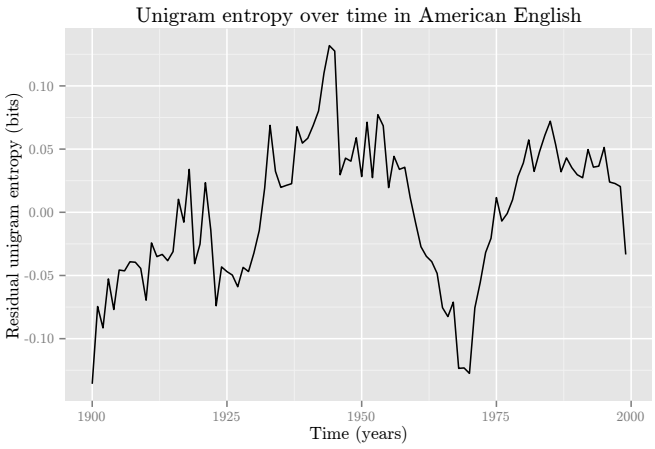


Figure 4: Residual values of unigram entropy in American English during the 20th century. The x-axis is years. The y-axis is the residual unigram (lexical) entropy after controlling for parameters signifying the size of the corpus: log number of unique unigrams, log number of unigrams.

Because of the individual language differences, we additionally combined all languages in a mixed-effects regres-

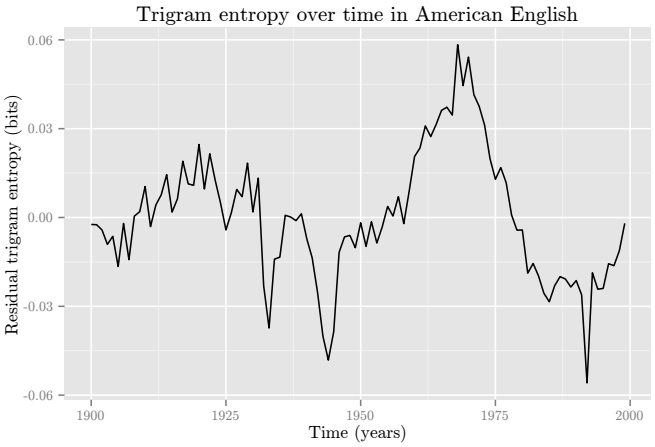


Figure 5: Residual values of trigram entropy in American English during the 20th century. The x-axis is years. The y-axis is the residual trigram (structural) entropy after controlling for parameters signifying the size of the corpus: log number of unique trigrams, log number of trigrams.

sion, post-hoc, using random effects for language, with trigram entropy as a random slope. The controls for this regression were identical for those used in the individual linear regressions, and the results were nearly identical. Critically, changes to unigram entropy were still strongly negatively correlated with changes to trigram entropy (β=-0.55, SE=0.0161, t=-34.359, $p<10^{-15}$). The number of unique trigrams and number of volumes were positively correlated with trigram entropy (β=0.96, SE=0.021, t=45.410, $p<10^{-15}$ and β=0.048, SE=0.0107, t=4.496, $p<10^{-5}$, respectively). The total number of trigrams and the number of unique unigrams were negatively correlated with trigram entropy (β=-0.29, SE=0.0191, t=-14.916, $p<10^{-15}$ and β=-0.088, SE=0.0238, t=-3.705, $p<0.001$, respectively). Number of pages was not significant.

Summary

Changes to unigram entropy and trigram entropy were negatively correlated, the opposite of what the null hypothesis expects: When amounts of lexical information rise, structural information drops. This constitutes strong evidence for accounts that expect language to restrict the amount of information provided at a given time (Aylett & Turk, 2004; Jaeger, 2010; Levy & Jaeger, 2007). It is quite surprising that the transitional or structural properties of language should change in response to the increasing amount of information, as predicted by information theoretic accounts, and yet this is the case for all the languages studied here.

These findings open the door to studies of other trade-offs in long term information rate. We use transitional probabilities here as an estimate of structural complexity (more complex transitions would indicate more complex structure), but it would also be interesting to use parsed corpora to look at the frequency of different grammatical constructs. Does a rise in information also predict a decrease in complex grammatical structures, such as complex clauses? Our hypothesis would suggest so. The Google Ngram corpus does contain basic syntactic information, in the form of rough part-of-speech tags (e.g. noun vs. verb, but not preterite vs. participle; Lin et al., 2012), and those too can perhaps be used to infer complexity in future studies. In a separate study, Cohen Priva (under revision) shows that speech rate (another form of information rate) is negatively correlated with both lexical and syntactic information rates.

Both unigram and trigram information rates rose during certain years and fell in others. There were large scale dips in both unigram and trigram entropy in all languages at different times, suggesting the existence of additional factors that play a part in determining the acceptable information rate for a language at any particular time. The timing of the drops in information rate is potentially quite telling– in most of the languages analyzed here, there are large dips in trigram entropy around 1915-1920 and 1935-1945, perhaps corresponding to the two world wars. If information rate corresponds to large-scale societal attitudes, then this would give us a new tool to study societal change. Fluctuations in information rate may be used to track society-level mood, or even predict future events. For instance, does information rate rise before democratic revolutions or following them? Being allowed to express one's opinion will mean that more opinions will be expressed, but perhaps the expression of new opinions is what brings about democratic revolutions in the first place.

Acknowledgments

This work greatly benefited from discussions with Fiery Cushman, Dan Jurafsky, Elinor Amit, Scott AnderBois, Laura Kertz, and David Badre.

References

Arnon, I., & Cohen Priva, U. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon*, *9*(3), 377–400. doi: 10.1075/ml.9.3.01arn

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Cohen Priva, U. (under revision). *Not so fast: Fast speech correlates with lower lexical and structural information.* Brown University manuscript.

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language*, *80*(4), 748–775.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the Association for Computational Linguistics* (pp. 199–206).

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jurafsky, D., Bell, A., Gregory, M. L., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: Benjamins.

Jurafsky, D., & Martin, J. (2000). *Speech and language processing: an introduction to Natural Language Processing, computational linguistics, and speech recognition*. New York: Prentice Hall.

Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, *66*(4), 588–611. doi: 10.1016/j.jml.2012.04.003

Kuperman, V., Pluymaekers, M., Ernestus, M., & Baayen, H. (2007). Morphological predictability and acoustic duration of interfixes in Dutch compounds. *The Journal of the Acoustical Society of America*, *121*(4), 2261–2271. doi: 10.1121/1.2537393

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178. doi: 10.1016/j.jml.2015.03.003

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Scholkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems (NIPS)* (Vol. 19, pp. 849–856). Cambridge, MA: MIT Press.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the acl 2012 system demonstrations* (pp. 169–174).

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. doi: 10.1016/j.cognition.2012.09.010

Michel, J.-B., Shen, Y. K., Presser Aiden, A., Veres, A., Gray, M. K., Pickett, J. P., ... Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Norcliffe, E., & Jaeger, T. F. (2014). Predicting head-marking variability in yucatec maya relative clause production. *Language and Cognition*, *FirstView*, 1–39. doi: 10.1017/langcog.2014.39

Pate, J. K., & Goldwater, S. (2015). Talkers account for listener and channel characteristics to communicate efficiently. *Journal of Memory and Language*, *78*, 1–17. doi: http://dx.doi.org/10.1016/j.jml.2014.10.003

Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*, 146–159.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155. doi: 10.1016/j.cognition.2014.06.013

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

van Son, R., & van Santen, J. (2005). Duration and spectral balance of intervocalic consonants: a case for efficient communication. *Speech Communication*, *47*, 100–123.

van Son, R. J. J. H., & Pols, L. C. W. (2003). How efficient is speech? *Proceedings of the Institute of Phonetic Sciences*, *25*, 171–184.

The Influence of Language-specific Auditory Cues on the Learnability of Center-embedded Recursion

Jun Lai (J.Lai@tilburguniversity.edu)
Tilburg centre for Cognition and Communication and Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, the Netherlands

Chiara de Jong (c.dejong_3@tilburguniversity.edu)
Tilburg centre for Cognition and Communication, Tilburg University, the Netherlands

Dingguo Gao (edsgao@mail.sysu.edu.cn)
Department of Psychology, Sun Yat-sen University, China

Ren Huang (huangren@mail2.sysu.edu.cn)
Department of Psychology, Sun Yat-sen University, China

Emiel Krahmer (E.J.Krahmer@tilburguniversity.edu)
Tilburg Center for Cognition and Communication, Tilburg University, the Netherlands

Jan Sprenger (J.Sprenger@tilburguniversity.edu)
Tilburg Center for Logic and Philosophy of Science, Tilburg University, the Netherlands

Abstract

The learnability of center-embedded recursive structures has attracted much attention (Corballis, 2007; Friederici, 2004; Rey, Perruchet, & Fagot, 2012). However, most of the previous studies adopted the artificial grammar learning paradigm (Reber, 1967) and did not apply natural language stimuli. Rather, they applied synthetic meaningless training materials, which hardly represent the richness and complexity of natural language. Accordingly, in the current study, we attempt to tighten the link between artificial language learning and natural language acquisition in the auditory modality, by enriching our learning environment with phonological cues that occur in natural, spoken information; in particular, Chinese tones. In a grammaticality judgment task, we examined the syntactical processing by participants from different language backgrounds. Through the cross-language comparison between Chinese and Dutch native speakers, we aim to test the influence of language-specific phonological cues on processing complex linguistic structures. The results showed that tones had a more beneficial learning effect for Chinese than for Dutch participants. In other words, when participants learned a new language, they were likely to bring their own language routines implicitly from the familiar native language into processing the unfamiliar one.

Keywords: Phonological cues; Language-specific; Artificial language; Syntactical processing; Auditory modality

Introduction

The ability to use and understand hierarchical structures has been proposed to be a unique characteristic of human primates (Hauser, Chomsky, & Fitch, 2002; Fitch & Hauser, 2004). One of these complex structures is center-embedded recursion, which has attracted attention from various research domains (Corballis, 2007; Friederici, 2004; Rey, Perruchet, & Fagot, 2012). However, when center-embedded structures occur in natural language, such as the English sentence "*The dog that the cat chased ran away"*, it appears to be difficult to understand and to process for human language users.

Although a large number of studies have investigated the processing of center-embedded recursion using artificial grammar learning paradigm, whether this type of structures can be learned still remains controversial (Conway & Christiansen, 2005; Conway, Ellefson & Christiansen, 2003; Li, Jiang; Guo; Yang & Dienes, 2013; Vicari & Adenzato, 2005). Moreover, most previous studies focused on the learning of center-embedded recursion in the visual modality (de Vries, Monaghan, Knecht, & Zwitserlood, 2008; Rey, et al., 2012), whereas learning the structures in the auditory modality, which deserves more research attention, was virtually neglected (Conway et al., 2003; Mueller, Bahlmann, & Friederici, 2010). This is quite surprising since the auditory modality is arguably more important than the visual modality for first language acquisition during infancy. Hence, experiments that aim at a proper understanding of how complex linguistic structures are acquired have to include it.

A recent study (Lai, Krahmer, & Sprenger, 2015) has demonstrated that learning such structures is indeed possible in the auditory modality within a 30-miniute exposure. Listeners were able to learn an artificial language, consisting of sequences with center-embedding (for example, non-word sequences following the structure of A_1B_1, $A_1A_2B_2B_1$, $A_1A_2A_3B_3B_2B_1$). Moreover, the results suggested that facilitative cues from previous experiments in the visual modality (Lai & Poletiek, 2011, 2013; Lai, Krahmer, & Sprenger, 2014), which were observed in the learning of visual center-embedded recursion, were also helpful in the auditory modality (Lai, et al., 2015). These cues pertained to 1) the order of the stimuli in the learning environment

(incrementally increasing the complexity of the training input); 2) the total amount of stimuli presented (few specimen rather than many); 3) the frequencies with which the training stimuli were repeated (differential rather than uniform). All three factors turned out to improve the performance of the participants in judgments of grammaticality significantly, and performance was best when all three cues were combined.

More precisely, in Lai et al. (2014, 2015), there were three groups: a) a Starting-small (SS) group, which was trained with staged input. Participants saw 144 learning exemplars, ranging from the easiest to the most difficult; b) the Starting-less (SL) group, which saw less unique exemplars (36), but all exemplar were repeated for an equal number of time; c) the Starting-high (SH) group, which also received less unique exemplars (36), but exemplars were repeated for an unequal number of times. Results showed that all groups performed significantly above chance level. In addition, it was found that humans were able to extract center-embedded recursion in an auditory learning environment, when the training input was arranged incrementally with increasing complexity. The starting small facilitation effect thus also holds true for the auditory modality. Moreover, the SH group performed significantly better than both the SS and SL group. The small diversity in exemplars of the SH group did not hinder learning CE recursion, but helped participants to focus on regularities. Additionally, the repetition of that limited amount of exemplars enables learners to become acquainted with the grammatical structures and to solidify their memory of the recursive structures.

The adoption of artificial grammar learning paradigm has been criticized for having low ecological validity and can hardly mimic the complete natural language acquisition procedure (Forkstam, Jansson, Ingvar, & Petersson, 2009). The simplified artificial language does not have the richness and complexity of the real natural language in various dimensions, such as the large amount of vocabulary, semantics, and phonology, etc. (Arciuli & Torkildsen, 2012). Therefore, it remains under speculation to which degree results from artificial language learning can be generalized to the actual language acquisition process. Accordingly, the purpose of the current study is to create a stronger link between artificial language learning and natural language learning. It is crucial that the statistical learning study that we conduct can not only simulate the richness of natural language environment, but also maintain its particular advantage, namely, the strict control over prior language knowledge and the learning material (Arciuli & Torkildsen, 2012). Therefore, we retain the artificial grammar paradigm, but we supplement it by adding a property of natural languages, namely, phonological cues. More precisely, using the auditory modality, Chinese tones were added to the artificial learning input. Firstly, we examine whether the phonological property from natural language would influence the processing of auditory center-embedded recursion. Secondly, we investigate whether listeners from different language background would process the complex structures differently. Thirdly, we studied whether the facilitative cues from our previous study (Lai et al., 2015) also worked in the presence of phonological cues. Previous studies showed that learners relied on native speech routines in their own language while learning to segment a new language (Vroomen, Tuomainen, & de Gelder, 1998, Saffran, Werker, & Werner, 2006). For example, Vroomen et al. (1998) found that word stress and vowel harmony had a differential impact on Finnish, Dutch, and French listeners. Similarly, in word segmentation studies, it has been shown that adult learners profited from a "metrical segmentation strategy", which treated language-specific factors as signals in recognizing word boundaries (Saffran, et al., 2006; Tyler, 2006). For example, English listeners tended to consider strong syllables as the onsets of upcoming words, since English is a stress-timed language and routinely most of English words start with strong syllables (Cutler, & Norris, 1988).

These previous studies indicated that phonological cues from natural language indeed affected participants differently according to their natural language. The above phenomena were observed in word segmentation tasks. We examine whether they also exist in syntactical processing tasks. We focus on tones in Chinese. We hypothesize that tones, which are present in Chinese but absent in most European languages, such as Dutch, would have a different influence on Chinese and Dutch listeners, when they are processing center-embedded structures in the auditory modality. In a grammaticality judgement task, we exposed Dutch and Chinese participants to two sets of artificial input, one with and the other without Chinese tones. Chinese are expected to make use the existence of tones, while Dutch are not. We also test whether the optimal learning strategies in our previous experiment (Lai et al., 2015) succeed when the input resembles natural language more closely.

Methods

Participants

Fifty Dutch speakers (25 female, mean age 21.68 year, SD 2.12) from Tilburg University and fifty Chinese speakers (38 female, mean age 20.38 year, SD 3.17) from Sun Yat-sen University participated in this study for course credit. None of the Dutch participants learned Chinese before the experiment, and vice versa. No participants had prior knowledge about the experiment. All the participants had normal hearing abilities.

Materials and design

Non-word sequences were generated by a center-embedded recursive rule, possessing the pattern A_1B_1, $A_1A_2B_2B_1$ or $A_1A_2A_3B_3B_2B_1$ (c.f. Lai, et al., 2015). There were two sets of stimuli: one was the non-tonal set, which was recorded by a native Dutch speaker; whereas the other one is the tonal set, which was recorded by a native Chinese speaker. Both speakers were instructed to read sequences in a natural way

as if pronouncing natural speech in their own language. The speed of reading and intonation were held constant. For example, each sequence was pronounced syllable by syllable (approximately 400 ms per syllable). We also checked the validation of our experimental material by conducting an intelligibility test prior the main experiment. The test showed that all Dutch utterances were intelligible for Dutch native speakers, and Chinese utterances were also clear to Chinese native speakers. Listeners in the intelligibility test would not participate in the main experiment.

For the stimuli set, the sequences consisted of two, four, or six consonant-vowel syllables. The syllables were constructed using six consonants (/b/, /p/, /d/, /t/, /n/, and /m/) and two vowels (/i/, /e/, /o/, and /a/). We created a set of A-syllables (i.e. bi, be, di, de, ni, ne), and a set of B-syllables (i.e. po, pa, to, ta, mo, ma). The A-B mapping was not random. Not every A can go with every B-syllable. Instead, the pairs of consonants (b-d, d-t, n-m) determined the relatedness of AB pairs. None of the sequences existed in Dutch or Chinese lexicons.

For the Chinese stimuli set, the same set of (lexical) syllables was used, while adding Chinese tones to the syllables. A tonal syllable is characterized by two factors, namely pitch height (fundamental frequency) and contour (with a level, low-rising, high-falling, or dipping shapes) (Wang & Saffran, 2014). Mandarin Chinese employs tones to differentiate lexical meanings. These special variations in Chinese tones appear as "truly foreign acoustic cues to the ears of non-tonal speakers" (Wang & Saffran, 2014). In the current design, two distinctive tones, i.e. Tone 2 (low-rising) and Tone 4 (high-falling), were selected to differentiate Category A syllables from Category B syllables. Syllables with tones were strictly selected to avoid producing words with semantic meanings.

Sequences were read out with a natural intonation. The pattern was held consistent across all stimuli. Both Dutch and Chinese utterances were presented at the same volume and in the same manner to participants in all conditions. We measured the pitch contour of the recorded sequences by examining their acoustical parameter, namely, the fundamental frequency (F0), as depicted in Figure 1.

As displayed in Figure 1, for four-syllable sequences, Dutch- and Chinese utterances possessed distinctive features. On the one hand, the pitch of Dutch utterances was comparatively flat. The pitch of the third syllable was slightly lower than the others, marking the start of the second half of the sequences. (For the 6-syllable ones, the lower pitch started at the fourth syllable). The pattern was held consistent. On the other hand, for each Chinese utterance, the first half had a low-rising trend, and the second half had a high-falling pattern.

Procedure

Participants were randomly assigned into one of the four groups: Dutch listener --Tonal stimuli, Dutch listener --Non-tonal stimuli, Chinese listener --Tonal stimuli, and Chinese listener --Non-tonal stimuli. The experiment consisted of two phases, learning and testing. Participants were informed that they were participating in a simple language learning task. They were not provided with any information about the underlying center-embedded recursive grammar. In the learning phase, participants were instructed to listen to the artificial sequences attentively. Each trial starts with a beep (400 ms), followed by a learning sequence in a syllable-by-syllable method and then the inter-stimulus interval (1000 ms).

Regarding to the complexity of the exemplars, there was an equal number of all three levels, i.e. zero-, one-, and two-level of embedding (LoE) (for example, *bepa*, *beditopa*, *bedinimatopa*). In the learning phase, in total there were 144 sequences, consisting of 36 unique exemplars with a

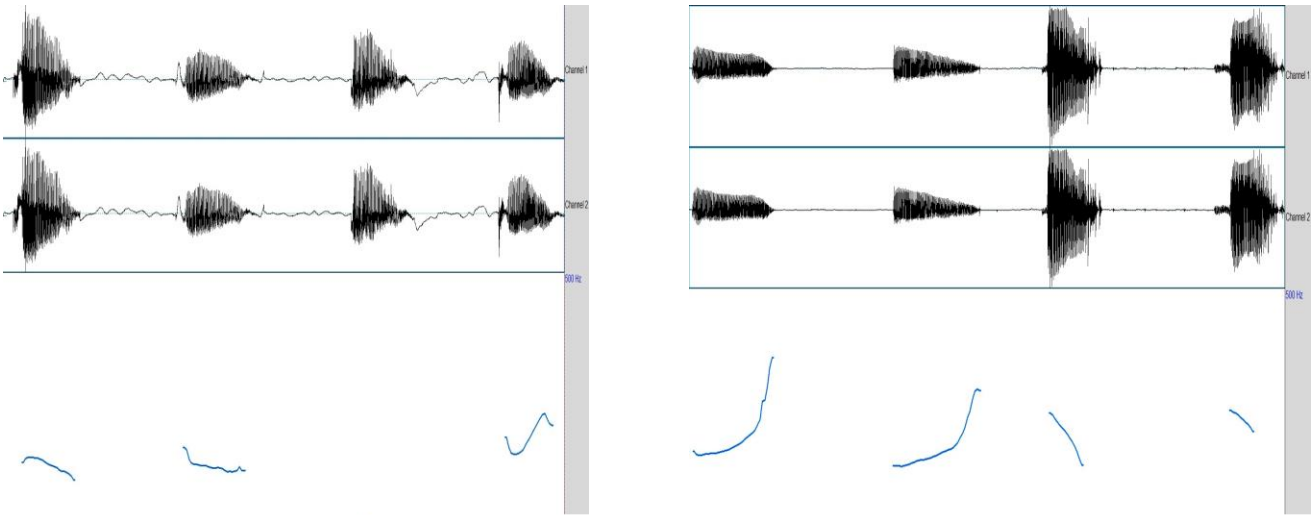


Figure 1. Comparison of fundamental frequency (F0) contours between Dutch and Chinese sequences. For example, the F0 of a 4-syllable sequence (dibipata), on the left is the Dutch utterance and on the right is the Chinese one with tones.

repetition. Therefore, there were 12 unique sequences for each LoE. Exemplars were repeated according to their occurrence probability, which was determined by the center-embedded recursive grammar. High probable exemplars were repeated more frequently than low probable ones. Specifically, 0-LoE sequences were repeated more times than 1-LoE ones, which in turn occurred more times than 2-LoE ones. In the testing phase, there were 72 unique sequences, half grammatical and half not. Ungrammatical items were produced by mismatching the related syllables with unrelated ones. The test sequences, which also had three levels of complexity, occurred in a random manner, instead of being arranged from the simplest to the most difficult. Participants were required to make judgments whether the test sequences were governed by the same rule as the one in the learning phase. If they agreed that the test sequences were generated by the same rule that generated the learning sequences, then they pressed "YES"; if not, then "NO".

Prior to the formal start of test phase, a practice session with four trials familiarized the participants with the forthcoming procedure. Practical trials would not occur again in formal tests. After the experiment, participants were debriefed.

Results

A one-sample t-test showed that all groups performed significantly above chance level accuracy: Dutch participants listening to Non-tonal set, M= .60, SD= .08, t (24) = 6.25, $p < .001$; Dutch participants listening to Tonal set, M= .58, SD= .10, t (24) = 4.00, $p < .001$; Chinese participants listening to Non-tonal set, M = .61, SD= .10, t (24) = 5.50, $p < .001$; Chinese participants listening to Tonal set, M= .69, SD= .14, t (24) = 6.79, $p < .001$.

An ANOVA showed that there was a main effect of native language, F (1, 96) = 8.22, p = .005, η_p^2 = .08. Generally, Chinese speakers (M = .65, SE = .02) scored significantly higher than Dutch speakers (M = .59, SE= .02). There was no main effect of tonality of input, F (1, 96) = 1.20, p = .276, η_p^2 = .12, but crucially the interaction between native language and tonality was significant, F (1, 96) = 5.35, p = .023, η_p^2 = .053 (as shown in Figure 2).

When Chinese speakers listened to stimuli with tones, they scored higher than they did when listening to stimuli without tones, t (48) = 2.14, p = .037, r^2 = .09. By contrast, Dutch speakers performed similarly, irrespective of whether they were exposed to stimuli with or without tones, t (48) = 1.01, p = .320, r^2 = .02. Therefore, the presence of tones advanced the learning performance of Chinese participants, but not of Dutch participants.

Furthermore, as depicted in Figure 3, we observed a main effect of grammaticality, F (1, 192) = 26.60, p = .000, η_p^2 = .122, and a main effect of group, F (3, 192) = 3.44, p = .018, η_p^2 = .051 (Table 1), but no significant interaction, F (3, 192) = .250, p = .861, η_p^2 = .04. Generally, scores on grammatical items (M = .67, SE= .01) were significantly higher than on ungrammatical items (M = .57, SE= .01), $p < .001$.

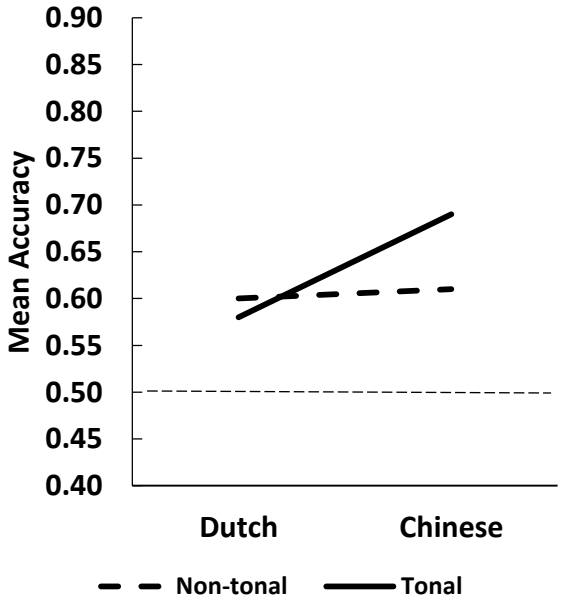


Figure 2. Mean accuracy of Dutch and Chinese participants on tonal and non-tonal test sequences. The dotted line represents chance level (M= .50).

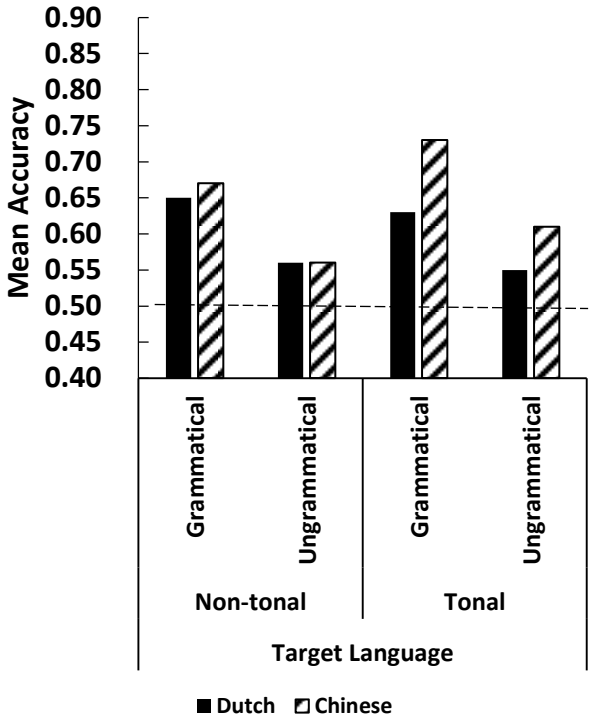


Figure 3. Mean accuracy of Dutch and Chinese participants on the grammaticality of both tonal and non-tonal test sequences. The dotted line represents chance level (M= .50).

Discussion

In a complex linguistic structure learning experiment, we added Chinese tones onto the non-word artificial input and investigated whether listeners' learning performance would be influenced by their native language background. Firstly, we replicated the significant learning effect obtained by the starting-high strategy in previous research (Lai et al., 2014; 2015). All groups achieved better than chance accuracy in discriminating grammatical sequences from non-grammatical ones. The successful learning overall is potentially explained by the efficient interaction of three factors: a) the incremental ordering (the most fundamental associations being presented first, then dependencies with one embedding, and in the end the most complex ones with two embedding) helped learners deconstruct the complexity, by arranging the learning input in a more efficient and easier way; b) the small set of exemplars helped participants focus on the underlying regularities. As a result, participants avoided being easily confused by the diversity and variations of the large amount of unique exemplars; c) the repetition of exemplars helped them consolidate memories. Therefore, the current results contribute to the debate about the conditions under which complex statistical patterns can be learned best (Conway et al., 2003; Perruchet & Rey, 2005). For instance, unlike the current study with a relatively small set of stimuli, Gomez (2002) and Gomez and Maye (2005) suggested that higher variability in exemplars can actually help participants learn regularities. They found that a token "*X*" can facilitate learning of *"aXb"* structures. Being sensitive to the degrees of stimulus complexity, statistical learning deserves further research on the composition of learning input.

Crucially, our results revealed that the nature of participants' native language, instead of the target language, had an important influence on syntactical processing of center-embedded recursive structures. Chinese participants performed better, when listening to stimuli with tones than without. Chinese participants might possess comparatively higher sensitivity towards the tonal variations. This indicates that when the target artificial language contained language-specific cues, participants made use of these phonological cues that they were familiar with from their native language. In contrast, these lexical tones had no noticeable impact on Dutch participants when listening and processing recursive structures, presumably because tones are not present in Dutch. Regarding this aspect of second language acquisition, our results confirm that learners might benefit in learning the new language, when the target language shares properties from their native language.

Recent cross-linguistic research has also probed into the interplay between prior linguistic experience and subsequent (second) language learning (Onnis & Thiessen, 2013). For instance, in a tonal word vs. non-word discrimination task, Mandarin bilinguals and monolinguals largely outperformed English monolinguals (Wang & Saffran, 2014). On the one hand, this might be due to the prior knowledge matching with the target language or not; on the other hand, it might be that the underlying cognitive mechanism for tonal language users leads to better performances in more domain-general cognitive tasks, compared to non-tonal language users, as Bidelman, Hutka, and Moreno (2013) suggested. Our results, which showed that Chinese scored general higher than Dutch participants in this task, were consistent with this potential account. Since our stimuli were recorded by different speakers (Chinese for the tonal set and Dutch for the non-tonal set), it is also conceivable that other factors, besides the presence of tones, might influence learning performance. Further research on language-specific statistical biases is needed.

Furthermore, the previous literature mainly focused on the contribution of phonological cues in speech processing, e.g. phrase discrimination from sound stream (Bion, Höhle, & Schmitz, 2007). Our results manifested that phonological cues can facilitate syntactical processing in a higher level, namely, phrase structure grammar (Chomsky, 1957), which produces hierarchical center-embedding.

In addition, being consistent with previous research (Lai, et al., 2014; 2015), we found that generally participants were more accurate in recognizing grammatical sequences than ungrammatical ones. However, our current data showed that with the help of tones, Chinese participants improved in detecting ungrammatical sequences.

Conclusion

The present study displayed the crucial influence of language-specific cues on learning center-embedded structures via the auditory modality. The phonological tone cue had a different impact on speakers with a tonal and non-tonal background. As predicted, when the target language shares the phonological characteristics of participants' native language, learning was enhanced. This finding sheds further light on second language acquisition. Further more studies are needed to incorporate more language-specific cues in order to conduct the comparison between artificial language learning and second language acquisition.

References

Arciuli, J., & Torkildsen, J. V. (2012). Advancing our understanding of the link between statistical learning and language acquisition: the need for longitudinal data. *Frontiers in Psychology, 3*:324.

Bach, E., Brown, C., & Marslen-Wilson, W. (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, *1*(4), 249-262.

Bidelman, G.M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: evidence for bidirectionality between domains of language and music. *PLos ONE* 8:e60676.

Bion, R.A.H., Höhle, B., & Schmitz, M. (2007). The role of prosody on the perception of word-order differences by 14-month-old German infants. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International*

Congress of Phonetic Sciences (pp. 1537–1540). Saarbrucken, Germany: International Congress of Phonetic Sciences.

Chomsky, N. (1957). *Syntactic Structures*. The Hague/Paris: Mouton.

Corballis, M.C. (2007). Recursion, language, and starlings. *Cognitive Science, 31*, 697-704.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology-Learning Memory and Cognition, 31*(1), 24-39.

Conway, C. M., Ellefson, M. R., & Christiansen, M. H. (2003). When less is less and when less is more: Starting small with staged input. In *Proceedings of the 25th annual conference of the cognitive science society* (pp. 270-275). Mahwah, NJ: Lawrence Erlbaum.

Cutler, A., & Norris, D. (1988). The role of the strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, 14, 113-121.

De Vries, M. H., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition, 107*(2), 763-774.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science, 303*(5656), 377-380.

Forkstam, C, Jansson, A, Ingvar, M., & Petersson, K.M. (2009). Modality transfer of acquired structural regularities. *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1686-1691). Austin, TX: Cognitive Science Society.

Friederici, A.D. (2004). Processing local transitions versus long-distance syntactic hierarchies. *Trends in Cognitive Sciences, 8*, 245-247.

Gomez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Gomez, R.L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*, 183-206.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*(5598), 1569-1579.

Lai, J., Krahmer, E. J., & Sprenger, J. M. (2014). Studying Frequency Effects in Learning Center-embedded Recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 797-802). Austin, TX: Cognitive Science Society.

Lai, J., Krahmer, E. J., & Sprenger, J. M. (2015). The learnability of Auditory Center-embedded Recursion. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp.1237-1242). Austin, TX: Cognitive Science Society.

Lai, J., & Poletiek, F. H. (2011). The impact of adjacent-dependencies and staged-input on the learnability of center-embedded hierarchical structures. *Cognition, 118*(2), 265-273.

Lai, J., & Poletiek, F. H. (2013). How "small" is "starting small" for learning hierarchical centre-embedded structures? *Journal of Cognitive Psychology, 25*(4), 423-435.

Li, F. F., Jiang, S., Guo, X. Y., Yang, Z. L., & Dienes, Z. (2013). The nature of the memory buffer in implicit learning: Learning Chinese tonal symmetries. *Consciousness and Cognition, 22*(3), 920-930.

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of Embedded Syntactic Structures Depends on Prosodic Cues. *Cognitive Science, 34(2)*, 338-349.

Onnis, L., & Thiessen, E.D. (2013). Language experience changes subsequent learning. *Cognition, 126(2)*, 268-284.

Perruchet, P., & Rey, A.,(2005). Does the mastery of center-embedded linguistic structures distinguish humans form non-human primates? *Psychonomic Bulletin and Review, 12(2)*, 307-313.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 77*, 317-327.

Rey, A., Perruchet, P., & Fagot, J. (2012). Centre-embedded structures are a by-product of associative learning and working memory constraints: Evidence from baboons (Papio papio). *Cognition, 123*(1), 180-184.

Saffran, J.R., Werker, J.F., & Werner, L.A. (2006). The infant's auditory world: Hearing, speech, and the beginning of language. In R. Siegler & D. Kuhn (Eds.,), *Handbook of child development* (6th ed., pp 58-108). New York: Wiley.

Tyler, M.D. (2006). French listeners can use stress to segment words in an artificial language. *Proceedings of the 11th Australasian International Conference on Speech Sci. & Tech.*, edited by Warren P., & Watson, C.I. (Australasian Speech Sci. & Techno. Assoc. Inc., Auckland, New Zealand), 222-227.

Vicari, G., & Adenzato, M. (2014). Is recursion language-specific? Evidence of recursive mechanisms in the structure of intentional action. *Consciousness and Cognition, 26*, 169-188.

Vroomen, J., Tuomainen, J., & de Gelder, B. (1998). The rules of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language, 38*(2), 133-149.

Wang, T.L., & Saffran, J.R. (2014). Statistical learning of a tonal language: the influence of bilingualism and previous linguistic experience. *Frontiers in Psychology, 5*.

Zimmerer, V. C., Cowell, P. E., & Varley, R. A. (2011). Individual behavior in learning of an artificial grammar. *Memory & cognition, 39*(3), 491-501.

U-INVITE: Estimating Individual Semantic Networks from Fluency Data

Jeffrey C. Zemla (jzemla@brown.edu)
Yoed N. Kenett (yoed_kenett@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Providence, RI 02912 USA

Kwang-Sung Jun (kjun@discovery.wisc.edu)
Wisconsin Institute for Discovery
University of Wisconsin-Madison, Madison, WI 53706 USA

Joseph L. Austerweil (joseph_austerweil@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Providence, RI 02912 USA

Abstract

Semantic networks have been used extensively in psychology to describe how humans organize facts and knowledge in memory. Numerous methods have been proposed to construct semantic networks using data from memory retrieval tasks, such as the semantic fluency task (listing items in a category). However these methods typically generate group-level networks, and sometimes require a very large amount of participant data. We present a novel computational method for estimating an individual's semantic network using semantic fluency data that requires very little data. We establish its efficacy by examining the semantic relatedness of associations estimated by the model.

Keywords: semantic networks; memory retrieval; fluency; random walk; probabilistic modeling

Introduction

Semantic memory is the system of memory that stores concepts and facts. Although the way in which semantic memory is organized into categories and subcategories remains an open question (Jones, Willits, & Dennis, 2015), one common approach is to represent it as a network comprised of nodes (a word or concept) and edges between nodes that signify that the two concepts are associated. However, how do we estimate a given individual's semantic network?

A growing body of work has related statistics of semantic networks (e.g., centrality) to cognitive phenomena, such as language development, memory retrieval and creative thinking (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Hills, Todd, Lazer, Redish, & Couzin, 2015). Most of this work has analyzed aggregated group-based networks, which cannot be used to understand individual differences. Currently, only one study has examined individual differences in semantic networks (Morais, Olsson, & Schooler, 2013). Here, we present a novel probabilistic method to estimate an individual's semantic memory structure efficiently using data from a semantic fluency task.

The semantic fluency task (listing of items in a category) has a long history in cognitive psychology (Henley, 1969). Typical subjects show a distinct behavioral pattern in this task, reporting items in clusters (sub-categories) and switching to new clusters when subsequent items are hard to retrieve (Troyer, Moscovitch, & Winocur, 1997). For instance, when typical subjects list animals, they may list several farm animals before switching to zoo animals. This clustering and switching behavior has been used to make inferences about the cognitive processes and representations underlying search through semantic memory.

Abbott, Austerweil and Griffiths (2012, 2015) proposed a model of semantic memory retrieval that accounts for this clustering and switching behavior (though see Hills, Jones, & Todd, 2012 for an alternative model that also accounts for this behavior). Given a semantic network, data are generated by taking a censored random walk on that network: Starting from a category's node, their model moves over random edges, emitting the labels of any nodes (e.g., "turkey") in its path if they are in the target category and have not been visited previously. The result is a fluency list that contains no duplicate items and is arranged by the order in which the items were first encountered. Jun et al. (2015) proposed a computational method based on this process to infer a semantic network from fluency data. Their method, initial-visit emitting random walk (INVITE) is based on the principle that multiple fluency lists from the same network can be used to infer that semantic network.

In this paper, we build on the INVITE model to develop a novel method for estimating an individual's semantic network.. We begin by presenting a few possible methods to estimate semantic networks, including INVITE and our approach. Next we evaluate these approaches and show that our approach can efficiently recover a network from simulated fluency data. Finally, we present an experiment where we collected fluency data from participants and examined the semantic similarity of edges in networks generated by the different approaches.

Estimating Semantic Networks

To estimate a network from fluency lists, we assume items are retrieved according to a censored random walk on a network (Abbott et al., 2015) and invert this process using Bayes' rule. Formally, let **G** be the participant's semantic network, $\vec{X}^m$ be the *m*th list produced from the participant (a censored random walk from the network), $\vec{S}^m$ be the inter-item response times (IRTs) of the *m*th list (so S_k^m is the time between response *k* and *k*-1 in list *m*), $\vec{Z}^m$ be the *m*th uncensored random walk on **G** (all category members visited by the random walk regardless of whether they were previously said) and $c(\vec{Z}^m)$ be a censoring function applied to the uncensored random walk such that it returns the censored walk, (i.e., $c(\vec{Z}^m) = \vec{X}^m$). We assume that each IRT S_k^m is gamma-distributed (e.g., Luce, 1986) with parameters $\tau_k^m - \tau_{k-1}^m$ and β. The former parameter reflects the number of censored items between two unique items in the *m*th uncensored list, where τ_k^m is the index of the *k*th unique item reported in the *m*th uncensored list. β is a parameter that controls the amount of variability in response times. Intuitively, the first parameter increases the expected IRT (so as the number of censored items between two uncensored items increases, the expected IRT increases) and β controls the variance (see Figure 1).

We examine this model as well as a restricted model that does not include response times, and a naïve random walk model that assumes no censoring occurs.

The Naïve Random Walk Model

The naïve random walk procedure (RW) ignores the censoring procedure and places edges between all successive items in every fluency list as if there were no censored items. For example, if we have a single list "*dog*, *cat*, *mouse*", our network would consist of three nodes and two edges, *dog-cat* and *cat-mouse*. When few lists are available, the RW procedure is a close fit to the most likely network. However, when many lists are available, a network estimated using this procedure quickly becomes over-connected, resulting in a network that contains many false edges. The RW procedure is inconsistent, meaning that it is not guaranteed to converge and, in fact, it will typically become less accurate as the number of lists increases.

The INVITE Model

Jun et al. (2015) proposed a method to invert the generative process used by Abbott et al. (2015): Given a participant's semantic fluency data were produced by a random walk on a network, what is the most probable network? Given *M* fluency lists, each denoted as $\vec{X}^m = (X_1^m \ldots X_{Nm}^m)$, we seek a network **G** that maximizes the likelihood of the data:

$$\prod_{m=1}^{M} \prod_{k=1}^{N_m - 1} \mathbb{P}(X_{k+1}^m | X_{1:k}^m) \qquad (1)$$

where N_m denotes the length of the *m*th censored list and X_k^m denotes the *k*th item from the *m*th list. Hereafter, we remove this superscript for readability when it is clear from context. The key to INVITE is to generate each item in a fluency list from a different random walk – one that treats visited nodes as transient and unvisited nodes as absorbing. To form each random walk, we re-arrange the states in the transition matrix of **G** so that the rows and columns are in list order, e.g., G'_{12} denotes a transition from X_1 to X_2:

$$\mathbf{G}' = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$

where **Q** denotes transitions between previously emitted items (transient states), **R** denotes transitions from previously emitted items to novel items (absorbing states), and **0** and **I** (the identity matrix) ensure the random walk is absorbing. **G'** is updated after each step in a list and reconfigured after each list.

Thus, we calculate $\mathbb{P}(X_{k+1}|X_{1:k})$ as the probability of starting at X_k and being absorbed by X_{k+1} given transient states $X_{1:k}$ and absorbing states $X_{k+1:Nm}$. This is computed using the fundamental matrix (Doyle & Snell, 1984) of **G'**, $\mathbf{N}=(\mathbf{I}-\mathbf{Q})^{-1}$, where N_{ij} denotes the expected number of times a walk starting from state *i* visits state *j* before being absorbed. Thus,

$$\mathbb{P}(X_{k+1}|X_{1:k}) = \begin{cases} \sum_{i=1}^{k} N_{ki} R_{i1} & \text{if } \mathbf{N} \text{ exists} \\ 0 & \text{otherwise} \end{cases}$$

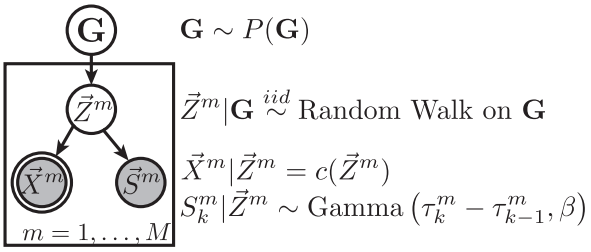


Figure 1: Left: Graphical model representing our computational method. The box is a plate, which means that the variables are copied and conditionally independent given the network **G** for each list *m* from one to the total number of lists *M*. The shaded nodes are observed. The double circle denotes a deterministic function. Right: The generative process for the model. iid means independent and identically distributed.

The U-INVITE Model

Our model, U-INVITE, improves performance of the INVITE model given a small number of lists. It does so by extending INVITE in two ways: (1) by using the time between responses (IRTs), and (2) assuming that the network is undirected and unweighted (the probability of transiting to any connected node is equal).

Inter-item Response Times As shown in Jun et al. (2015), INVITE works particularly well when the number of lists is large. When the number of lists is small, there may not be enough information in the order of the items to accurately

estimate the network. One way we resolve this is by incorporating IRTs into the emission probability. Rather than maximizing Equation 1, we maximize the following equation:

$$\prod_{m=1}^{M}\prod_{k=1}^{N_m-1}\sum_{r=1}^{\infty}\mathbb{P}(X_{k+1}^{m}\text{ in r steps}|X_{1:k}^{m})^{\theta}\mathbb{P}(S_{k+1}^{m}|r,\beta)^{(1-\theta)}$$

That is, we weight the probability of being absorbed by X_{k+1} in r steps by the probability of observing an IRT of S_{k+1} given an r-step walk. We calculate the probability of observing an IRT given r steps using a gamma distribution parameterized by β. We add an additional free parameter θ to adjust the influence of IRTs on the model, i.e., when θ is 1 the model ignores IRT information. Although this equation contains an infinite summation, we have found that limiting this to $r = 20$ works as an efficient approximation in practice, as most chains are absorbed by the next state in fewer than 20 steps. Rather than computing the probability of being absorbed by X_{k+1} in the limit, we compute:

$$\mathbb{P}(X_{k+1}\text{ in r steps}|X_{1:k})=\sum_{i=1}^{k}\left(\mathbf{Q}^{r-1}\right)_{ki}R_{i1}$$

We assume a uniform prior for **G**, and that $\mathbb{P}(X_1^m|\mathbf{G})$ is uniformly distributed for all M lists.

Unweighted Networks and the Search Procedure In addition to timing information, we include additional constraints to estimate networks efficiently: We assume that the random walk is unweighted and undirected. Although these assumptions may seem psychologically unrealistic, Abbott et al. (2015) found that both weighted and unweighted semantic networks captured human performance in semantic fluency tasks well. Whether human semantic networks are unweighted or weighted is an unsettled question and orthogonal to the purpose of our paper (a method for estimating weighted networks with IRT information could be created by deriving a MLE estimator without constraints on the transition matrix, as in Jun et al., 2015). Its strength enables us to estimate networks efficiently from censored lists. The original INVITE allows weighted edges, adding additional degrees of freedom that need to be inferred. Further, the transition matrix inferred by INVITE is fully-connected; to convert it into a network that is not fully-connected would require an additional thresholding process (where estimated edge weights lower than an additional threshold parameter are removed from the final network and then appropriately normalized). For these reasons, it is difficult to compare networks constructed by INVITE and U-INVITE, and we do not provide a direct comparison of the algorithms in this paper.

To find the network that maximizes the likelihood of the data, we use a stochastic search procedure with smart initialization. Using an initial network constructed with the RW procedure, we toggle one or more edges and compute the new network's probability, accepting the change when the new network is more probable given the data. We favor toggling edges that connect two items present in multiple lists, as these "hub nodes" have a larger effect on the network's posterior probability. Specifically, we set a fixed probability (P_{hub}=.8) that we toggle an edge that connects two hub nodes, or otherwise toggle an edge at random. We also favor toggling one edge at a time, as the probability of producing a network that has zero probability (cannot produce the data) increases rapidly as the number of simultaneous edge changes is increased. At each update, we toggle 1+D edges simultaneously where D is sampled from a Geometric distribution with P_{geom}=.2. These free parameters affect only the time to convergence, and ensure that the search procedure will converge in the limit. We run this procedure repeatedly until we have tried 1500 updates without finding a network with a higher likelihood. We found this stopping criterion to be robust for the toy networks estimated in this article.

Simulations

Varying the Number of Lists

We used simulated data to estimate the accuracy of four different models as a function of the number of fluency lists used to fit the network. We compared RW, U-INVITE, and U-INVITE with IRTs. We report results using two possible values of θ in the IRT method: 0.5 (the IRT5 model) and 0.9 (the IRT9 model).

We generated 10 toy small-world networks, consisting of 15 nodes each, using the Watts-Strogatz procedure (Watts & Strogatz, 1998). Previous literature has suggested that human semantic networks are small-world like (e.g., Borge-Holthoefer & Arenas, 2010), being highly clustered yet having a low shortest path length between any two nodes. We chose parameters for the Watts-Strogatz procedure to generate networks that were roughly comparable in node degree and clustering coefficient to what has been reported previously for human semantic networks (Steyvers & Tenenbaum, 2005). Our toy networks had an average node degree of 4 and a mean clustering coefficient of 0.29.

We varied the number of fluency lists used to estimate the network from 2 to 35. Lists were generated by starting at a random node in the network and taking a random walk until all of the nodes were traversed, then extracting only the first visit of each node from the list. Each list was truncated to roughly 70% of its length, or 11 items, with the restriction that each node in the network is traversed at least once in the set of lists. This truncation process mimics human-generated data reported later (i.e., each list contains approximately 70% of the total items listed by a participant). Simulated IRTs were generated from a gamma distribution, using the number of steps between two items in a walk and β=1.1 as parameters.

We calculated the cost of each reconstructed network using Hamming distance, or the number of edges that would need to be added or removed to convert it to the original network. The results, shown in Figure 2, demonstrate that U-INVITE does converge to the original network, though incorporating IRTs can lead to convergence with fewer lists. While the IRT5 model performs reasonably well, we found that it was outperformed by the IRT9 model, which assigns a higher weight to the item order than to the IRTs.

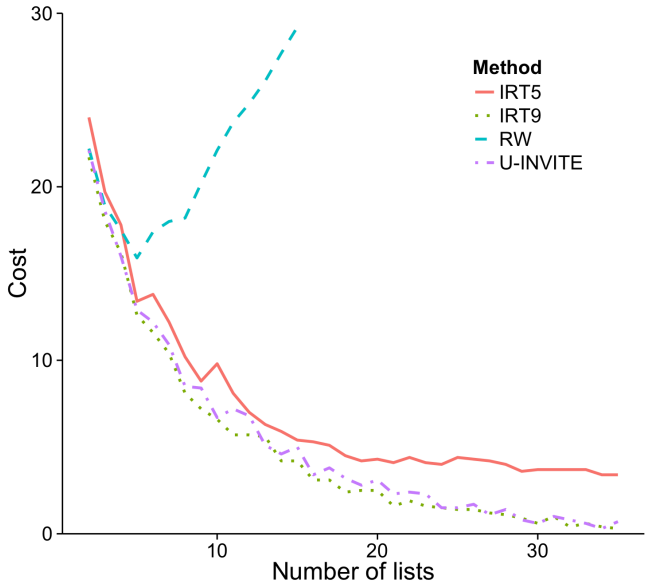


Figure 2: As the number of fluency lists increases, our methods tend toward zero error. Not shown: the RW method increases linearly to about 50 by 35 lists.

Model Comparison Given Three Lists

We conducted an additional simulation using only three fluency lists to examine whether IRTs improve network estimation when only a small number of lists are used. Using the same procedure as above, we generated 300 toy networks with 15 nodes each, and reconstructed each network using each of the four methods. A one-way repeated measures ANOVA was conducted to examine the effect of the different methods on the cost of estimating the original network (Table 1). This analysis revealed a significant main effect of method, $F(3, 897) = 26.13$, $p < 0.001$, $\eta^2 = .08$. Post-hoc analyses revealed that this main effect was due to the IRT9 model outperforming the other three models (all p's < 0.001). No significant differences were found between the average cost of the RW, U-INVITE, and IRT5 models.

We classified the edges in the reconstructed networks to indicate whether an edge was present in both the original and reconstructed network (Hit), in the original but not the reconstructed network (Miss), or in the reconstructed but not the original network (False Alarm). A one-way repeated measures ANOVA was conducted for Hits/Misses, $F(3, 897) = 262.06$, $p < 0.001$, $\eta^2 = .47$, as well as False Alarms, $F(3, 897) = 415.79$, $p < 0.001$, $\eta^2 = .51$. We found that compared to U-INVITE, the IRT9 model contains significantly more hits (and fewer misses). However, we found no difference in the number of false alarms. This indicates that incorporating IRTs improves upon U-INVITE by accurately detecting more genuine edges, while keeping the number of false alarms constant. In contrast, the RW method generates substantially more false alarms compared to U-INVITE and IRT9. Finally, the IRT5 model resulted in the highest amount of hits and fewest misses, but also the most false alarms. In future work, we will explore why this is the case and how to weight IRT and order information optimally. Next, we compare the different methods on real behavioral results from a semantic fluency task where participants generate multiple fluency lists.

Table 1: Results of estimating networks from three lists. 300 networks were generated and each method was used to estimate these toy networks. Values denote average scores (standard deviation in parentheses).

Measure	RW	U-INVITE	IRT5	IRT9
Cost	19.2 (3.3)	18.9 (3.6)	19.1 (3.7)	18.2 (3.5)
Hits	17.7 (1.8)	16.3 (2.0)	18.3 (1.9)	16.9 (2.0)
Misses	12.3 (1.8)	14 (2.0)	11.7 (1.9)	13.0 (2.0)
False Alarms	6.9 (1.9)	5.0 (2.2)	7.5 (2.4)	5.0 (2.1)

Experiment: A repeated semantic fluency task

Methods

Participants We recruited twenty participants from Amazon Mechanical Turk (11 male, mean age 31.75) who were located in the United States.

Procedure Participants were given a category label (e.g., *animals*) and asked to generate as many items from that category as possible in three minutes. Each participant completed nine lists in total, three for each of three categories (animals, fruits, and vegetables). The order of the lists was pseudo-randomized so that each triad of lists contained one of each category, and participants never completed the same category twice in succession.

Each response was hidden from view after it was entered to reduce cueing effects from previously entered items. Participants were instructed to list each item no more than once within a list, but that they could repeat themselves on subsequent lists.

Results

We present the results solely from the *animal* category, which generated the most responses. Twenty participants generated 280 unique animals in total (average 54.5 per participant and 33.7 per list). Participants were largely successful at avoiding repetitions, repeating fewer than one animal per list on average. All repetitions were removed from the data set prior to analysis.

To validate our method, we examined the similarity between connected nodes using the BEAGLE lexical semantic database (Jones & Mewhort, 2007). The database estimates the semantic similarity between two words from their statistical co-occurrence in a large corpus of text. For example, *dog-cat* has a high BEAGLE similarity whereas *dog-toad* has a low BEAGLE similarity.

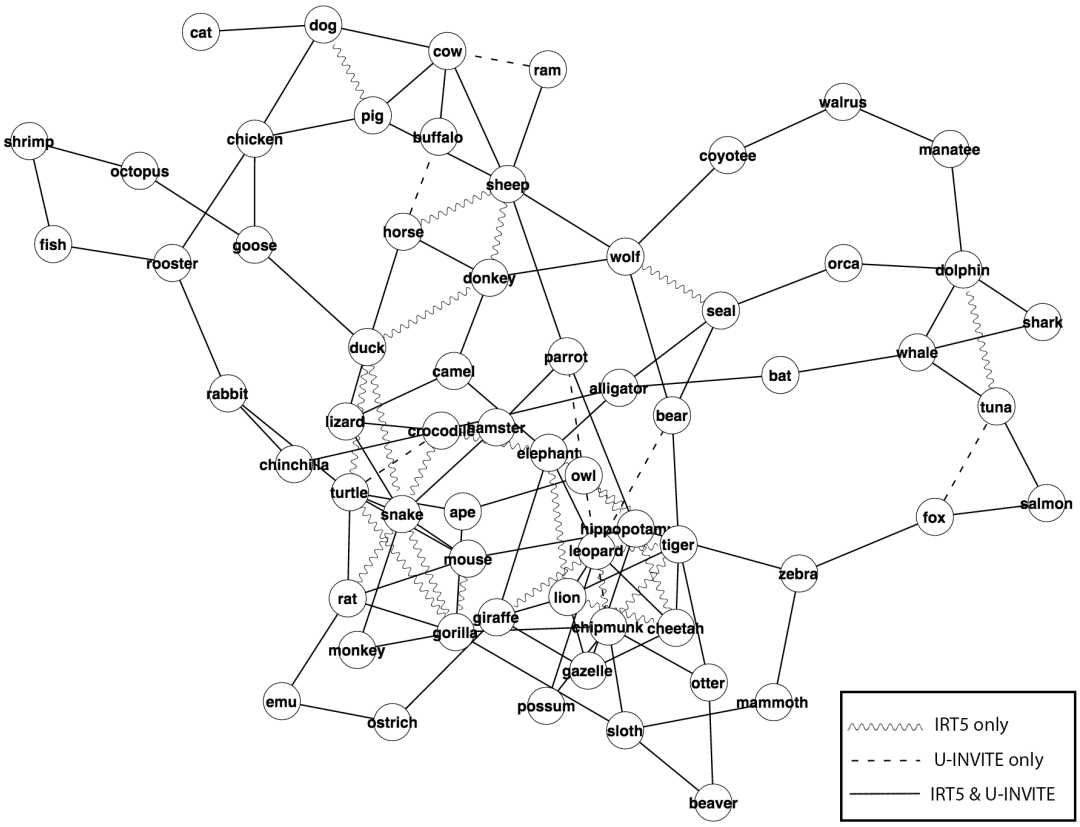


Figure 3: Network reconstruction with U-INVITE and IRT5 models of the semantic network of a single participant. Edge style denotes model success at estimating edges: Solid line: edges estimated by both methods; Dashed line: edges estimated only by the U-INVITE model; Sinusoidal line: edges estimated only by the IRT5 model.

We computed the BEAGLE value for all edges in each network, except those that could not be computed because one of the nodes was not in the BEAGLE database (accounting for 6.4% of edges). For each participant, we also calculated the BEAGLE value that would be expected between a random pair of nodes in the network (i.e., the average BEAGLE value across all possible edges in the network). We then subtract this score from the BEAGLE value of each edge in the network to generate a BEAGLE difference score (BDS), where positive scores indicate the two connected animals are more similar than would be expected by chance.

All five methods generated sensible networks[1]: the average BDS of all edges for each participant and each method (80 networks) was higher than would be expected by chance. U-INVITE generated networks that were similar to the RW method. Across all twenty participants, U-INVITE added zero edges compared to the RW method, and removed only 39 edges (roughly 2.4% of all edges). This is probably because U-INVITE needs more data to make accurate estimates. Jun et. al (2015) found that for toy networks, INVITE was typically a poor estimator of the true network when the number of lists was small.

Compared to U-INVITE, the IRT9 model added 23 edges (1.4%) and removed 22 edges (1.4%). We calculated the average BDS (per participant) of edges present in IRT9 but not in U-INVITE, and found that these edges had a BDS significantly greater than expected by chance, indicating that the added edges connect nodes that are semantically similar, $M_{BDS} = .068$, $t(12) = 2.21$, $p = .047$. However, we also found that edges present in U-INVITE but not in IRT9 were more similar than expected by chance, $M_{BDS} = .063$, $t(10) = 3.23$, $p = .009$. There was no statistical difference between the average BDS of edges added compared to edges removed ($p = .89$).

The IRT5 model made substantial changes to the network compared to U-INVITE (see Figure 3 for an example). Across all participants, the IRT5 method added 486 edges (30.1%) that were not present in the U-INVITE network and removed 150 edges (9.3%) of the edges that were present in the U-INVITE network. The edges added by IRT5 have an average BDS significantly greater than expected by chance, $M_{BDS} = .015$, $t(19) = 2.91$, $p = .017$. However, as with IRT9, we also found the reverse to be true: Edges that were removed from the U-INVITE model have a higher BDS score than was expected by chance, $M_{BDS} = .025$, $t(18) =$

[1] Networks for each participant and each method are available online at http://research.clps.brown.edu/austerweil/UINVITE16/

4.34, $p < .001$. Comparison between the edges removed and edges added showed no statistical difference ($p = .22$).

Conclusions

Advancements in network science and probabilistic modeling enable scientists to investigate how the structure of semantic memory contributes to language development, creativity and intelligence, and memory retrieval (De Deyne, Kenett, Anaki, Faust, & Navarro, in press). However, current research has been limited to group analyses, which cannot account for individual differences. Our method estimates an individual's semantic memory structure based on multiple semantic fluency responses.

Our approach extends that of Jun et al. (2015) by constraining the estimated networks to be unweighted and undirected, and incorporating response time information. We found that our method accurately recreated small-world networks, which have consistently been found to resemble human networks (Borge-Holthoefer & Arenas, 2010). Further modifications to our model may improve its accuracy. For instance, we may use a more realistic response time function other than a gamma distribution, such as the ex-Gaussian distribution (Heathcote, Popiel, & Mewhort, 1991). We also plan to examine more realistic process models (e.g., an imperfect censoring function would allow us to model perseverations in semantic retrieval). Finally, we plan to examine how to weight IRT information optimally and perform additional validations of our method.

Developing methods to estimate an individual's network representation from fluency data has great potential across cognitive science. They will allow us to examine individual differences in semantic networks and relate them to neurocognitive variables that affect memory search and executive functions in typical and clinical populations (Faust & Kenett, 2014). For instance, Alzheimer's and semantic dementia patients show marked disruption in performance on a semantic fluency task (Rohrer, Salmon, Wixted, & Paulsen, 1999). We hope that our method can be used to improve our understanding of impaired and unimpaired cognitive search.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. *Advances in Neural Information Processing Systems, 25*, 3050-3058.

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review, 122*(3), 558-569. doi:10.1037/a0038693

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in Cognitive Science. *Trends in Cognitive Sciences, 17*(7), 348-360. doi:10.1016/j.tics.2013.04.010

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy, 12*(5), 1264-1302.

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., & Navarro, D. J. (in press). Large-scale network representations of semantics in the mental lexicon. In M. N. Jones (Ed.), *Big data in cognitive science: From methods to insights*: Psychology Press.

Doyle, P. G., & Snell, J. L. (1984). *Random walks and electric networks*.

Faust, M., & Kenett, Y. N. (2014). Rigidity, chaos and integration: Hemispheric interaction and individual differences in metaphor comprehension. *Frontiers in Human Neuroscience, 8*(511), 1-10.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: an example using the Stroop task. *Psychological Bulletin*, *109*(2), 340-347.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior, 8*(2), 176-184.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431-440. doi:10.1037/a0027373

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences, 19*(1), 46-54.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1-37. doi:10.1037/0033-295X.114.1.1

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. Busemeyer & J. Townsend (Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232-254).

Jun, K.-S., Zhu, X., Rogers, T. T., Yang, Z., & Yuan, M. (2015). Human memory search as initial-visit emitting random walk. *Advances in Neural Information Processing Systems (NIPS)*.

Luce, R. D. (1986). *Response times* (No. 8). Oxford University Press.

Morais, A. S., Olsson, H., & Schooler, L. J. (2013). Mapping the structure of semantic memory. *Cognitive Science, 37*(1), 125-145. doi:10.1111/cogs.12013

Rohrer, D., Salmon, D. P., Wixted, J. T., & Paulsen, J. S. (1999). The disparate effects of Alzheimer's disease and Huntington's disease on semantic memory. *Neuropsychology, 13*(3), 381-388.

Steyvers, M., & Tenenbaum, J. B. (2005). The large scale structure of semantic networks: Statistical analysis and a model of semantic growth. *Cognitive Science, 29*(1), 41-78.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology, 11*(1), 138-146.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(4), 440-442.

Contextual Events and Their Role in a Two-Choice Joint Simon Task

Steve Croker (s.croker@ilstu.edu), J. Scott Jordan (jsjorda@ilstu.edu), Daniel S. Schloesser (dschloe@ilstu.edu), Vincent Cialdella (vtciald@ilstu.edu), Alex Dayer (adayer@ilstu.edu)
Department of Psychology, Illinois State University
Campus Box 4620
Normal, IL 61790-4620 USA

Abstract

We examined the effects of individual versus joint action on a Simon task using motion tracking to explore the implicit cognitive dynamics underlying responses. In both individual and joint conditions, participants were slower to respond, and were differentially attracted to the distracter response location, when the spatial component of the stimulus was incompatible with the response location. When two people completed similar two choice tasks together, the results were not statistically different from the individual condition, even though the magnitude of the stimulus-response compatibility effect was slightly larger. Neither was there an increased effect when the partner had no stimulus-response conflict to resolve. We found no evidence for an action conflict when the responses of the two partners were different. These data imply that the literature regarding the Joint Simon task is still in the process of determining the relevant events that interact with and support joint action.

Keywords: Joint action; Simon effect; motion tracking

In the two-choice Simon task, the spatial dimension of the stimulus leads to an automatic activation of the response corresponding to that spatial information (e.g., a stimulus located to, or pointing to, the left, primes a left response). If the correct response is located to the same side as the stimulus, responses will be faster than when the spatial features of stimulus and response do not overlap (Kornblum, Hasbrouq, & Osman, 1990). Sebanz, Knoblich, and Prinz (2003, 2005) had participants engage in a Simon task either as individuals or as dyads. The task required that participants respond to pictures of hands wearing a colored ring pointing to the left or right. Response buttons were located to the left or right of the screen. When participants did the task together, one individual was responsible for responding to one component of the stimulus (e.g., respond right for a green ring) while the other was responsible for responding to another component (e.g., respond left for a red ring). Essentially, each participant engaged in a go/no-go task. That is, they responded when the stimulus matched the condition of their rule (i.e., a go trial) and did not respond when the stimulus did not match the condition of their rule (i.e., a no-go trial). In each trial, the direction of response indicated by the ring color was either *compatible* with the spatial dimension of the stimulus (e.g., left response and hand pointing left) or *incompatible* (e.g., left response and hand pointing right). For dyads, response times (RTs) were significantly longer on incompatible trials than compatible trials. This Simon effect did not occur when participants engaged in the same go/no-go task alone. In other words, the spatial compatibility of the pointing hand only affected RTs when participants completed the task with another person. Sebanz et al. argue that the joint Simon effect (JSE) emerged in the dyad condition because participants automatically represented the action of their partner. Consequently, when their partner's co-represented action conflicted with their own, they had to resolve a response conflict.

An alternative account of the JSE is presented by Dolk, Hommel, Prinz, and Liepelt (2013; Dolk et al., 2014). According to their referential coding hypothesis, the JSE does not emerge via co-representation of another's action, but rather the co-actor is treated as a spatial reference frame for coding one's actions. In the individual go/no-go task, there is only one response, which is not spatially coded. When a co-actor is introduced, a response is coded as "left" or "right" relative to the co-actor. Dolk et al. also suggest that, on this view, the source of the reference frame need not be an intentional agent. To test this claim, they replaced the co-actor with objects such as a waving cat, a clock, and a metronome, and found that the JSE still appeared.

Croker, Jordan, Schloesser, and Cialdella (2015) had participants engage in a Simon task alone or with a confederate. However, whereas most researchers have each participant effectively engage in a go/no-go task, Croker et al. asked participants to respond to one of two choices on every trial. In their task, participants always followed the same response rule (i.e., right for a green ring and left is the ring was not green). In the *individual* condition, participants completed this task alone. In the *joint-color* condition, a confederate worked alongside the participant and responded left for red rings and right if the ring was not red. In the *joint-direction* condition, a confederate worked alongside the participant and responded to the direction of the pointing hand. This *joint-direction* condition created two kinds of trials: trials in which the two stimulus dimensions primed the same response for the participant and confederate (i.e., non-conflict trials) and trials in which the stimulus primed opposite responses for the participant and confederate (i.e., conflict trials). In addition to requiring participants to respond on every trial, Croker et al. diverged from the Sebanz et al. (2005) paradigm by having participants indicate their responses using a computer mouse, while the confederate made responses on a button box. To make a response, participants moved a mouse cursor from the bottom center of the screen to a response location in the left or right upper corner of the display. In addition to providing RTs, the use of mouse trajectory responses—instead of button presses—

allowed the authors to obtain an implicit measure of the cognitive dynamics underlying response choice: maximum deviations toward the distracter response (MDs). In order for movement data to be informative, it is necessary to make both response options available to participants. In go/no-go tasks, in which only one response location is utilized (even if the alternative is displayed), participants move the mouse in a straight line from the start location to the response location, even on spatially incompatible trials. A two-choice task means that participants always have to consider both response locations, and we can observe the extent to which a movement is attracted to the alternate response location prior to its arrival at the correct response. Croker et al. replicated the basic Simon effect in both the RT and MD data across all three conditions; both RTs and deviations toward the distracter responses were greater on incompatible trials than compatible trials. On *compatible* trials, mouse trajectories were attracted toward a single strong attractor for that problem. On *incompatible* trials there was greater competition between the two response options, as different dimensions of the stimuli (i.e., color and direction) primed different responses, and there was greater deviation towards the distracter response. Croker et al. also found that RTs were increased, and MDs were decreased, in the joint conditions. This finding was interpreted as evidence for the role of inhibition; participants took longer to respond, but took more direct routes to the response, suggesting that the increased RTs were due to the presence of a co-actor, and not due to increased response-path lengths.

The purpose of the current study is to extend the work of Croker et al. (2015) in two ways. First, to examine the temporal dynamics of *both* partners' response selection in a two-choice Joint Simon task. In addition to the research described above, many other studies have shown a joint Simon effect (JSE) for dyads sharing a task in a go/no-go fashion (e.g., Malone, Castillo, Kloos, Holden, & Richardson, 2014; Welsh, 2009), but in the present study, we look at two participants who are both engaged in two-choice tasks. Croker et al. (2015) were only able to record movement trajectories for one person in joint-action conditions. In the present study, we use motion tracking to record response trajectories of two participants working alongside one another. Second, Croker et al. failed to replicate some of Sebanz et al.'s (2005) findings, concluding that such effects may have emerged specifically out of the unique constraints of the Sebanz et al. task. By collecting data from two participants both engaged in arm reach responses (as opposed to mouse and button responses) we can determine the extent to which the mouse trajectory data and arm reach trajectory data are consistent, or whether Croker et al.'s findings were are also a function of the specific task context.

If the inhibition account of the Joint Simon data proposed by Croker et al. (2015) generalizes beyond the task constraints of that experiment then, in addition to anticipating a replication of the basic Simon effect (i.e., larger RTs and MDs for incompatible trials than compatible trials), we predict that the presence of a co-actor will cause greater automatic inhibition, and thus we expect both partners to exhibit longer RTs in the joint conditions than the individual condition, but we do not expect to see increased MDs. Alternatively, given Malone et al.'s (2014) assertion that Joint Simon effects emerge out of a "…dynamical (time-evolving) interpersonal coupling that operates to perturb the behavior of socially situated actors…" (p. 1), the social availability of each actor's movement dynamics in the present task may lead to the two participants' dynamics becoming coupled. This would be consistent with Malone et al.'s finding that the trial-to-trial RTs collected in a Joint Go/No task were more correlated over the long-term than RT streams collected from pseudo pairs (i.e., randomly paired data from the Individual condition). Such time-evolving, interpersonal coupling in the present task might lead to the participants' real-time dynamics synergistically influencing each other and, as a possible result, generating response synchrony that serves to negate any Joint Simon effects. Such an outcome would imply that the effects deemed due to inhibition in the Croker et al. study do not generalize to a situation in which both actors respond on every trial via socially-observable movements.

Method

Participants and materials

Forty participants (32 female; mean age = 20.2 years) from Illinois State University volunteered as participants for extra course credit. Stimuli consisted of a set of four images, and one image appeared center screen on each trial. Every image was comprised of two stimulus dimensions: hand direction and ring color. In each trial, a hand was presented pointing either left or right wearing a ring that was either green or red. This yields four possible stimulus combinations: left-pointing hand/green ring, left-pointing hand/red ring, right-pointing hand/green ring, and right-pointing hand/red ring. We used a modified version of the Sebanz et al. (2005) paradigm to accommodate motion tracking technology. Stimuli were presented on a screen using Inquisit software (http://www.millisecond.com/), and pairs of participants responded using a wooden hammer to tap either a left or right wooden target (23" apart and 26" above the start location). A Polhemus Patriot motion sensor was attached to each hammer in order to capture the motor dynamics of each response (see Figure 1). Participants were told to respond as quickly as possible and to return to the starting location after each trial. On trials in which one or both of the participants responded incorrectly, a red X was presented on the corresponding side of the screen.

Procedure

After completing four practice trials, participants completed three blocks (100 trials each) of the Sebanz et al. task. In each block, all four images (left-pointing hand/green ring, left-pointing hand/red ring, right-pointing hand/green ring, and right-pointing hand/red ring) were presented 25 times each in a randomized order.

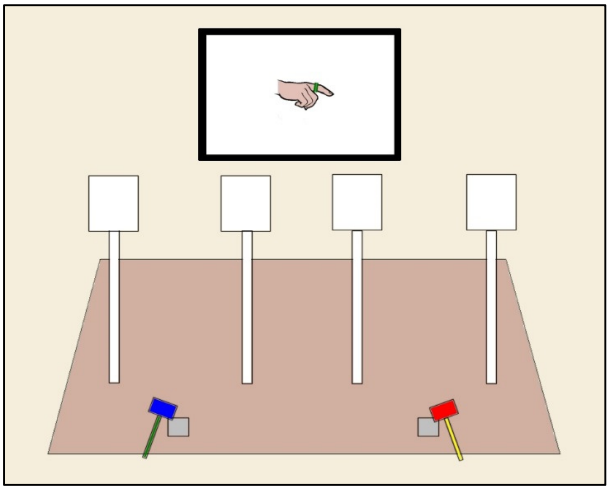


Figure 1: Experimental setup. Each participant responded to a left or right response location using a wooden hammer with a motion sensor attached to it, starting from a point marked on the table (grey square).

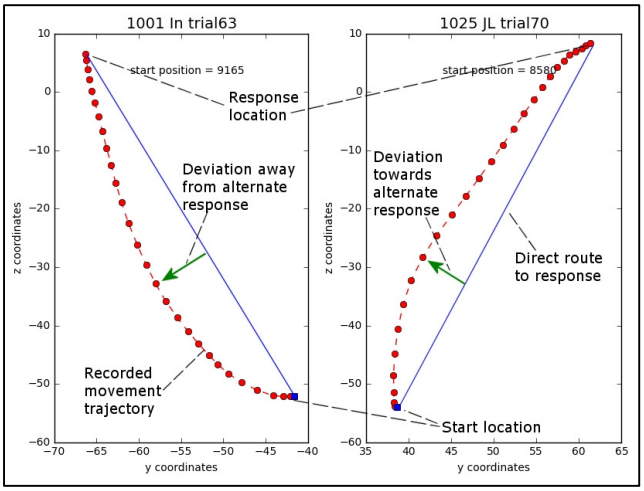


Figure 2: Sample trajectories: (a) a left response with a negative deviation; (b) a right response with a positive deviation.

In the *individual* block, participants completed the task on their own. They were asked to tap the right target whenever the ring was green, and the left target if the ring was not green. In the *joint-color* and *joint-direction* blocks, participants completed the task with a partner. They were instructed that their partner would respond on every trial to the same stimulus at the same time, but that the partner was following a different rule. In the *joint-color* block, participants responded to ring color. Each participant's rule was different, but their actions were the same. Partner A's task was to respond to red rings by tapping the left target and to non-red rings by tapping the right target. Partner B's task was to respond to green rings by tapping the right target and non-green rings by tapping the left target. There were two *joint-direction* conditions: *joint-left* and *joint-right*. In both *joint-direction* conditions, one partner responded to the spatial component of the stimulus (direction) while the other responded to the non-spatial component (color). In the *joint-left* condition, the partner responding to direction (i.e., Partner A) had the task of responding to hands pointing left by tapping the left target and to hands not pointing left by tapping the right target. Partner B's task was to respond to green rings by tapping the right target and to non-green rings by tapping the left target. In the *joint-right* condition, partner A's task was to respond to hands pointing right by tapping the left target and to hands not-pointing right by tapping the right target. Partner B's task was the same as Partner B's task in the *joint-left* condition: to respond to green rings by tapping the right target and to non-green rings by tapping the left target. Partner A completed the individual condition before the joint conditions and partner B completed the individual condition after the joint conditions. Joint conditions were counter-balanced between participants.

Design and Analyses

The dependent variables were *response time* and *maximum deviation* from an idealized trajectory toward the correct response. Streaming *x*, *y*, and *z* coordinates from each pair of participants were recorded at a rate of 60Hz using a Polhemus Patriot magnetic motion tracking system and Liberty 8 Capture software (Richardson, 2007). As the movement of interest is in the *y* (left-right) and *z* (up-down) planes, we did not include *x*-coordinate data (forward-backward) in our analyses. We recorded a continuous data stream for each condition (i.e., 100 trials), and defined trial onset as the point at which the tangential velocity of the movement in the *y* and *z* planes exceeded 5% of the maximum, and trial offset as the point at which tangential velocity dropped below 5% of the maximum, after passing one third of the maximum value. This procedure ensured that we captured just the outward portion of the motion, from the start location to the response location. We computed maximum deviation by comparing the *y*,*z* coordinates of the movement trajectory for each trial with a direct line between the start and end points of each trajectory. The greatest distance between the actual and idealized trajectories is the maximum deviation. A positive MD value indicates deviation towards the distracter response, and a negative deviation indicates deviation away from the distracter (see Figure 2). Only correct responses were included for analysis, but error rates were small (0.6%). Some trials had to be excluded due to noisy data resulting from electromagnetic interference to the magnetic field used by the motion tracking system; in a few cases this meant eliminating all trials for a given participant. After data were excluded, we had 38 participants in the individual condition, and 37 in each of the joint-color and joint-direction conditions.

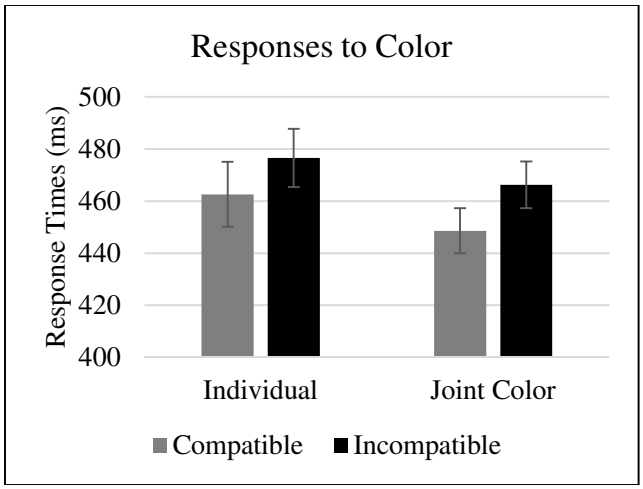


Figure 3: Response times (with standard error bars)

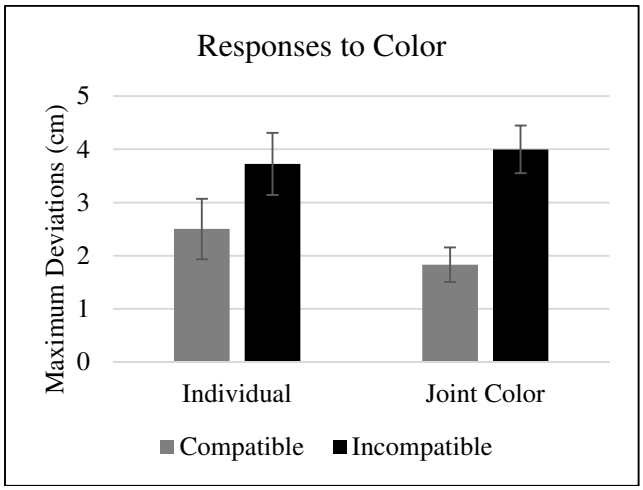


Figure 4: Maximum deviations (with standard error bars)

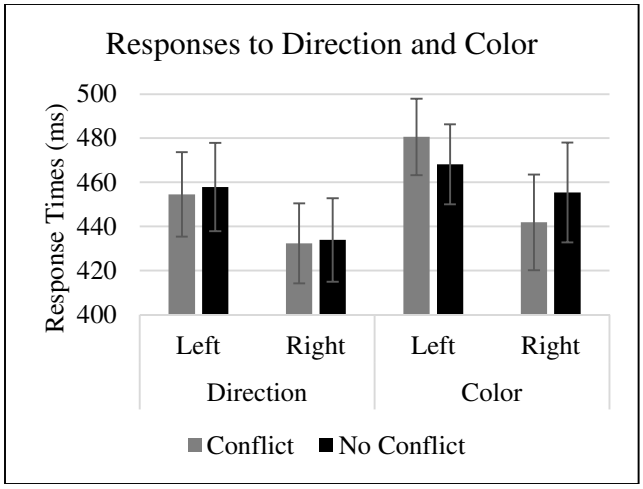


Figure 5: Response times for conflict vs. no-conflict responses for the joint-left and joint-right conditions (with standard error bars).

Results

Individual and Joint-Color Conditions

The individual condition enables us to explore whether we observe the basic Simon effect, and whether motion tracking gives us any additional insight into the effect. The joint-color task, in which partner A responds left to red, and partner B responds right to green, allows us to determine whether the addition of a co-actor with a different task rule, but the same actions, increases the size of the effect. We found an effect of spatial compatibility both for response times (RTs), $F(1,35) = 30.303$, $p < .001$, $\eta_p^2 = .464$, and maximum deviations (MDs), $F(1,35) = 22.825$, $p < .001$, $\eta_p^2 = .395$, for both the individual and joint-color conditions (see Figures 3 and 4). The increase in MDs on spatially incompatible trials shows that participants were differentially attracted to the unselected, incorrect response on these trials, suggesting an early response to the task-irrelevant spatial component of the stimulus, which was later inhibited.

The difference in RTs between compatible and incompatible trials was larger in the joint-color condition than the individual condition (18ms vs. 14ms), as was the difference in MDs. Although the differences were numerically greater in the joint-color condition, we found no statistical difference between the individual and joint-color conditions and, importantly, no interactions between condition and spatial compatibility for either dependent variable.

Joint-Direction Condition

The joint-direction condition allows us to explore the effects of a co-actor who is following a different task rule and sometimes producing different actions. On some trials, both participants respond to the same location (*no-conflict*); on others, they respond to different locations (*conflict*). Participants who completed the *joint-left* task in the joint-direction condition were instructed to respond left to left-pointing hands, which means that all their responses were spatially compatible with the stimulus. Participants who completed the *joint-right* task in the joint-direction condition were instructed to respond left to right-pointing hands, meaning that all trials were incompatible. We conducted mixed ANOVAs on the effects of task (direction vs. color—between subjects), direction-compatibility for the partner responding to the direction of the hand (left vs. right—between subjects), and action conflict between the two participants (conflict vs. no-conflict responses—within subjects) on RTs and MDs.

For RTs, there was an interaction between task, direction-compatibility, and action conflict, $F(1,33) = 9.249$, $p = .005$, $\eta_p^2 = .219$, whereby conflict and no-conflict trials differed for the color-task participants only, and the direction of the effect difference varied according to the direction-compatibility of their partner (see Figure 5). The two stimuli to which the participant given the right-direction task responded to the *same* location as the color-task participant are right/green and left/red. The two stimuli to which the left-direction response

was *different* to the color-task response are also right/green and left/red, both of which are spatial conflict trials for the color-task participant. Therefore, the interaction between action conflict and direction-compatibility can be explained as a simple stimulus-response compatibility effect. The participants given the direction task did not respond differentially on conflict or no-conflict trials, suggesting that they did not need to distinguish between own- and other-responses. Note that we do not expect to see stimulus-response compatibility effects for these participants because all responses in the left-direction condition are spatially compatible with the stimuli, and all responses in the right-direction condition are spatially incompatible with the stimuli.

There was an interaction between direction condition and response type for MDs, $F(1,33) = 16.314$, $p < .001$, $\eta_p^2 = .331$, whereby MDs were largest for conflict responses in the left-direction condition and no-conflict responses in the right-direction condition. As with the RT data, these larger MDs correspond to spatially-incompatible trials for the participant given the color-task, yet there was no interaction with participant task (direction vs. color), probably because this pattern also held for participants in the left-direction task. There was also an interaction between direction condition and task, $F(1,33) = 5.75$, $p = .022$, $\eta_p^2 = .148$, such that MDs were larger in the left-direction condition than the right-direction condition for participants given the color task, but there were no differences between these two conditions for participants given the direction task (see Figure 6).

In order to test the prediction that the presence of a co-actor would lead to larger RTs, but not MDs, for the color-task participant, we conducted 2 (task: individual vs. joint-direction) x 2 (spatial compatibility) repeated-measures ANOVAs. We found no main effect of task on RTs, $F(1,17) < 0.01$, $p = .989$, $\eta_p^2 = .000$. As with the joint-color condition, the difference between compatible and incompatible trials was numerically larger in the joint-direction condition than the individual condition (13ms vs. 8ms), but the condition x compatibility interaction was non-significant, $F(1,17) = 1.36$, $p = .260$, $\eta_p^2 = .074$. There was no effect of task on MDs, $F(1,17) = 0.40$, $p = .533$, $\eta_p^2 = .023$.

Finally, given that the role of partner A was just to respond to direction, we conducted mixed ANOVAs on the effects of stimulus direction, stimulus color, and direction-compatibility of the task for partner A only. There was an interaction between stimulus direction and condition for both RTs, $F(1,17) = 6.018$, $p = .025$, $\eta_p^2 = .261$, and MDs, $F(1,17) = 39.657$, $p < .001$, $\eta_p^2 = .700$. RTs and MDs were greater for left stimuli for participants responding to right and vice versa. For the stimuli in which the direction of the hand was congruent with the rule, the MDs are actually negative – participants were repelled from the distracter response (see Figure 7).

Discussion

As expected, the spatial compatibility of a stimulus had an effect on RTs and MDs in the individual condition. This is

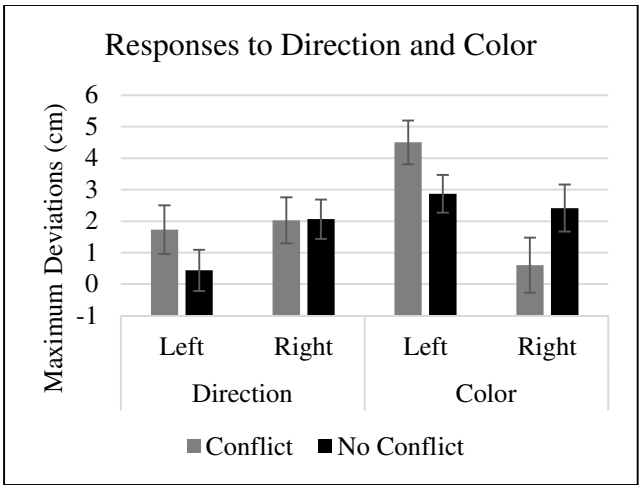


Figure 6: Maximum deviations for conflict vs. no-conflict responses for the joint-left and joint-right conditions (with standard error bars).

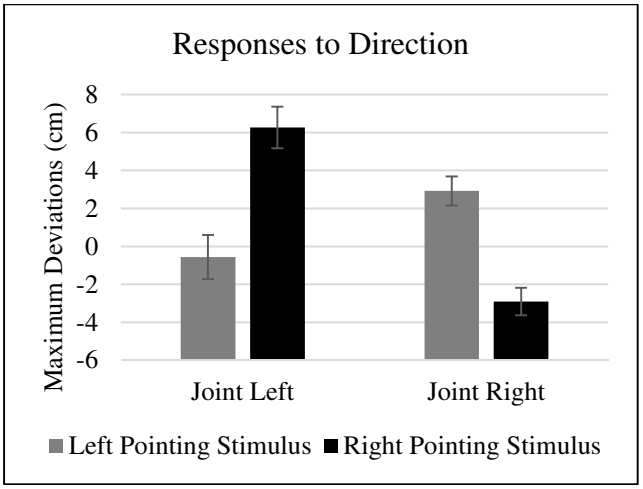


Figure 7: Maximum deviations for left vs. right stimuli for partner A (with standard error bars).

the basic Simon effect whereby participants take longer to respond to stimuli with conflicting spatial and non-spatial cues. The MD data give us direct evidence that participants are differentially attracted to the distracter response on spatially incompatible trials; reach trajectories showed early attraction to the response indicated by the direction of the hand, prior to moving toward the correct response location indicated by the color of the stimulus. This effect persisted when participants worked alongside a partner, regardless of whether their partner was responding to the color or spatial dimension of the stimulus. Our prediction that RTs would begreater in the joint conditions was not supported. This is surprising, as we previously found increased RTs for participants responding to color using a mouse tracking paradigm (Croker et al., 2015). However, aside from the difference between moving a mouse cursor through two dimensions and a motion sensor through three dimensions,

the major difference between the two studies is the response modality of the co-actor. Croker et al. had a confederate respond with a button press whereas, in the present study, co-actors responded by generating a reach behavior. These behaviors were emitted by both participants simultaneously, in full view of the other. Perhaps the difference in findings between Croker et al. (2015) and the present study is due to an entrainment that may have emerged between the movements of the two participants as they generated responses in the format demanded by the present task.

Contrary to the predictions of both the action co-representation and referential coding accounts, we found no effect of conflict vs. no-conflict trials in the joint-direction condition. If it is the case that participants represent or encode both their action and that of their partner, we would expect faster and more direct responses on trials where both participants respond to the same location than trials where one responds to the left and the other to the right. Participants did not take longer to respond when their action was the opposite of their partner. However, we also found no support for the prediction, based on our inhibitory account, that RTs would be greater in the joint condition than the individual condition.

The MD data present a slightly more complicated picture. We found evidence for a basic Simon effect for the color-task participants, in that MDs were larger for spatially incompatible trials, yet there was no interaction with task (direction vs. color). As shown in Figure 6, the lack of an interaction is probably due to the difference in MDs between conflict and no-conflict trials for participants in the left-direction condition. It could be the case that participants given the left-direction task also encoded the spatial incompatibility for their partners in the conflict trials, and reflected their partner's shifting trajectory as their partners resolved the spatial stimulus-response incompatibility. If this were the case, we would then also expect to see a similar effect for the no-conflict trials for participants given the right-direction task; an effect we did not observe. However, all trials are spatially incompatible in the right-direction task, and the response patterns of these participants may reflect this fact. Rather than a response *cost* for the left-direction participants on conflict trials, there may be a *facilitation* on the no-conflict trials, in which neither participant has to resolve a spatial conflict, either on the basis of partner response or stimulus-response compatibility. This interpretation is consistent with the fact that the MDs for participants given the direction task are smallest on left-direction/no-conflict trials. We also found that participants given the direction task exhibit small, even negative, deviations when the stimulus direction matches the target direction given in the task instructions. Participants given the "respond left to left hands" rule showed a spatial bias towards the left for all stimuli, and participants given the "respond left to right hands" rule showed a bias to the right. This finding suggests that, even though the instructions can be interpreted as stating how to respond to both left- and right-pointing hands, the movement data reflect the fact that only one spatial direction was explicitly mentioned.

In sum, using a novel motion-tracking methodology, we found that movement trajectories offer confirmatory evidence that the response time cost for spatially incompatible trials in the Simon task is due to an initial attraction to the direction indicated by the stimulus. In addition, the present data seem to constitute yet another study that reveals the contextually sensitive nature of Joint Simon effects. That is, while the data of Sebanz et al. (2003, 2005) are consistent with a co-representation account, the data of Dolk et al. (2013, 2014) with a referential coding account, the data of Croker et al. (2015) with an inhibition account, and the data of the present study with an entrainment account, it seems all of these accounts fail to survive generalization to different task constraints. While this lack of replication makes it difficult to assert the "correctness" of any one account, it simultaneously clarifies the immense complexity of interacting factors that contextually and contingently interact and support the phenomenon of joint action.

References

Croker, S., Jordan, J. S., Schloesser, D., & Cialdella, V. (2015). The effects of racial similarity and dissimilarity on the joint Simon task. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio, P. P. (Eds.) Proceedings of the 37th Annual Meeting of the Cognitive Science Society (pp. 489-494). Austin, TX: Cognitive Science Society.

Dolk, T., Hommel, B., Prinz, W., & Liepelt, R. (2013). The (not so) Social Simon effect: A referential coding account. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1248-1260.

Dolk, T., Hommel, B., Colzato, L.S., Schütz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: A review and theoretical integration. *Frontiers in Psychology, 5*:974.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility – A model and taxonomy. *Psychological Review, 97*, 253-270.

Malone, M., Castillo, R. D., Kloos, H., Holden, J. G., & Richardson, M. J. (2014). Dynamic structure of joint-action stimulus-response activity. *PloS One*, *9*, e89032.

Richardson, M. J. (2007). Liberty 8 Capture [Computer Software]. http://www.xkiwilabs.com/downloads_polhemus.html

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition*, *88*, 11-21.

Sebanz, N., Knoblich, G., & Prinz, W. (2005). How two share a task: Corepresenting stimulus-response mappings. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 1234-1246.

Welsh, T. N. (2009). When 1+ 1= 1: The unification of independent actors revealed through joint Simon effects in crossed and uncrossed effector conditions. *Human Movement Science*, *28*, 726-737.

A Framework for Evaluating Speech Representations

Caitlin Richter[1], **Naomi H. Feldman**[2*], **Harini Salgado**[3], **Aren Jansen**[4]
[1]Department of Linguistics, University of Pennsylvania, Philadelphia, PA 19104
[2]Department of Linguistics and UMIACS, University of Maryland, College Park, MD 20742
[3]Pomona College, Claremont, CA, 91711
[4]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD 21211
*Address for correspondence: nhf@umd.edu

Abstract

Listeners track distributions of speech sounds along perceptual dimensions. We introduce a method for evaluating hypotheses about what those dimensions are, using a cognitive model whose prior distribution is estimated directly from speech recordings. We use this method to evaluate two speaker normalization algorithms against human data. Simulations show that representations that are normalized across speakers predict human discrimination data better than unnormalized representations, consistent with previous research. Results further reveal differences across normalization methods in how well each predicts human data. This work provides a framework for evaluating hypothesized representations of speech and lays the groundwork for testing models of speech perception on natural speech recordings from ecologically valid settings.

Keywords: speech perception, speaker normalization, Bayesian modeling, approximate inference

Listeners track statistical distributions of sounds in their language. Adults are sensitive to these distributions when perceiving speech (Clayards, Tanenhaus, Aslin, & Jacobs, 2008), and infants' discrimination is influenced by these distributions (Maye, Werker, & Gerken, 2002). Statistical properties of the input can differ depending on the dimensions that listeners extract from the speech signal. For example, acoustic characteristics of vowels are highly variable when represented by their formant frequencies, but the variability is greatly reduced when they are represented by the z-score of their formant frequencies relative to other vowels by the same speaker (Figure 1; see also Cole, Linebaugh, Munson, & McMurray, 2010). Because the distributional characteristics of speech depend so heavily on the dimensions used, understanding the dimensions that listeners extract from the speech signal is a critical part of understanding phonetic learning and perception.

In this paper we introduce a novel approach to evaluating hypotheses about the dimensions that support listeners' perception. We adopt a cognitive model of speech perception from Feldman, Griffiths, and Morgan (2009), which predicts that listeners' perception is biased toward peaks in the acoustic distribution of sounds in their input. This model provides a formal link between the distribution of sounds in the input and listeners' discrimination abilities. We measure the input from a speech corpus and use the model to predict listeners' discrimination behavior. Different ways of representing the speech signal imply different distributions of sounds in the corpus, yielding different predictions about listeners' discrimination. We are interested in learning which representations of speech best predict listeners' actual discrimination.

We model AX discrimination tasks, in which listeners decide whether two sounds are acoustically identical. Previous

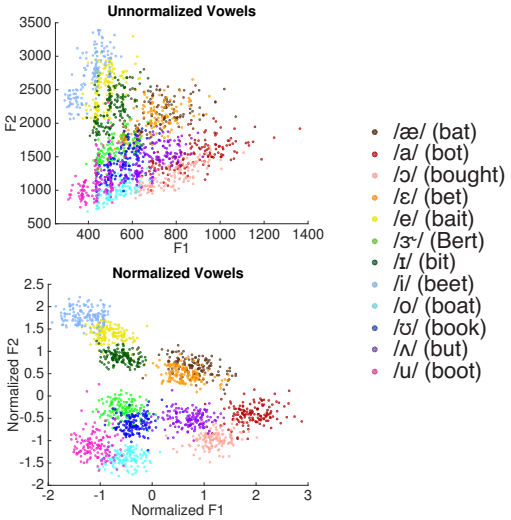


Figure 1: Acoustic characteristics of vowels produced in hVd contexts by men, women, and children from Hillenbrand et al. (1995), plotted as raw formant frequencies (top) and z-scored formant frequencies (bottom). If listeners' perception is biased toward peaks in these distributions, these feature spaces make different predictions about listeners' performance in perceptual discrimination tasks.

approaches to evaluating speech dimensions have instead focused on categorization tasks, in which listeners decide which category a sound belongs to (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011). Discrimination provides several advantages over categorization: it is a more fine-grained measure than categorization, and can be reliably measured in both adults and infants, even when listeners do not have well-formed categories for a given set of sounds. In addition, whereas building a categorization model requires speech recordings to have been labeled with phoneme identities, building a discrimination model does not, as we explain below.

As an initial case study, we examine speech dimensions defined by two speaker normalization techniques: z-scoring, which has been proposed in the cognitive science literature (e.g., Lobanov, 1971), and vocal tract length normalization (VTLN), which is widely used in automatic speech recognition (ASR) systems (e.g., Wegmann, McAllaster, Orloff, & Peskin, 1996). We find that both normalization methods yield a robust improvement over unnormalized representations in predicting listeners' discrimination, consistent with previous research. We also find that VTLN predicts human data better than z-

scoring, despite being less effective at eliminating speaker variability. These results illustrate our method for evaluating these dimensions against discrimination data and provide clues to the dimensions that guide listeners' perception. Adapting a cognitive model to operate over speech recordings also lays the groundwork for testing models of speech perception in more ecologically valid settings, by enabling cognitive scientists to make use of the same rich corpus data that is often used by researchers working in automatic speech recognition.

Speaker normalization

We begin by characterizing the speech representations tested in this paper. Speech contains commingled effects of linguistic, paralinguistic, and purely physical sources of variation. The goal of speaker normalization is to find representations that diminish some of the variability in the speech signal, like that from the speaker's body, while retaining task-appropriate information such as the variability that represents different phonemes. While some models have questioned whether listeners normalize across speakers (Johnson, 1997), most evidence suggests some degree of normalization. Normalized representations have been found to improve phonetic categorization (Cole et al., 2010; Nearey, 1978), increase a categorization model's match with human behavior (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011), and improve the performance of speech recognizers (Wegmann et al., 1996; Povey & Saon, 2006). We test the effects of two specific methods for speaker normalization, z-scoring and vocal tract length normalization, in a cognitive model of vowel discrimination.

Within-speaker z-scoring (Lobanov, 1971) has been suggested for descriptive sociolinguistic research (Adank, Smits, & Hout, 2004), as it highlights learned linguistic content while removing speaker-body confounds from vowel formants. A related manipulation, linear regression, has also been shown to improve cognitive models of fricative perception (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011).

Vocal tract length normalization (VTLN) is a technique developed for automatic speech recognition. VTLN compensates for speaker differences in vocal tract length by stretching or compressing the frequency axis of the productions of each speaker (Wegmann et al., 1996). The aim of VTLN is to adjust the corpus so that it is as if all the speakers had identical vocal tract lengths. VTLN is widely successful in ASR systems, where it substantially decreases the word error rate (e.g., Giuliani, Gerosa, & Brugnara, 2006). In performing this normalization, we use a procedure from Wegmann et al. (1996) adapted for an unsupervised setting, which selects frequency adjustments for speakers on the basis of their /i/ productions by maximising the similarity of all /i/ tokens across speakers.

We apply both normalization methods to vowels that are represented by mel frequency cepstral coefficients (MFCCs; Davis & Mermelstein, 1980). MFCCs are widely employed as an input representation in ASR systems (although we do not implement an ASR system here). MFCCs are a 12-dimensional vector that describe a timepoint of speech by capturing information about the spectral envelope, reflecting vocal tract shape. Thus, they capture information similar to formant frequencies, but have the advantage that they can be computed automatically from the speech signal, without being subject to the error inherent in automatic formant tracking.[1]

Table 1: Effect of normalization on K-L divergence.

	MFCCS (unnormalized)	Z-score normalized	VTLN normalized
Gender KLDiv	7.84	4.58	6.14
Dialect KLDiv	4.41	2.09	4.19

Effects of normalization Our method for testing hypothesized representations of speech against human perception relies on the idea that different representations of speech yield different distributions of sounds in the input. To examine the distribution of sounds in the input, we computed MFCCs, z-scored MFCCs, and MFCCs with VTLN from vowel recordings in the Vowels subcorpus of the Nationwide Speech Project (NSP; Clopper & Pisoni, 2006). This corpus contains ten different vowels pronounced in the context hVd (hid, had, etc.) by 5 female and 5 male speakers from each of 6 dialect regions of the United States. Each of these 60 speakers repeats each of the 10 hVd words 5 times, for a total of 3000 hVd tokens balanced across vowel, gender, and dialect.[2]

We characterize the effects of each normalization method on the distribution of vowels in the NSP corpus by computing symmetrized Kullback-Leibler divergence, a measure of difference between two probability distributions (Wang, Kulkarni, & Verdú, 2006). Lower K-L divergence indicates greater similarity between two distributions. We estimated K-L divergence across gender and across dialect (averaged over 15 pairwise comparisons of 6 dialect regions). Male and female speakers differ in their vocal tract lengths, and thus normalization algorithms would be expected to increase similarity between their vowel productions. Dialects also differ in their pronunciations of different vowels; although this variation is not related to vocal tract length, it may nevertheless be impacted by normalization algorithms that seek to neutralize speaker differences.

K-L divergence between genders and between dialect pairs is highest in MFCCs with no speaker normalization, reflecting the effects these factors have on the original acoustic signal (Table 1). Both VTLN and z-scoring reduce K-L divergence between genders, as predicted, so that male and female speakers saying the same vowel appear more similar after either of these normalizations than they are in unnormalized MFCCs. Z-scoring using all 10 NSP vowels also reduces K-L divergence between dialect pairs. In contrast, VTLN matching

[1] Z-scoring has previously been applied to formant frequencies, but we show in the next section that it is also effective at normalizing across speakers when applied to MFCCs. Vowel-intrinsic normalization methods such as formant ratios were not tested here because they are not straightforward to apply to MFCCs.

[2] While this corpus does not correspond exactly to listeners' experience with language, conducting initial simulations with a corpus of vowels in neutral contexts allows us to investigate algorithms for speaker normalization while sidestepping issues of how listeners generalize across phonological contexts.

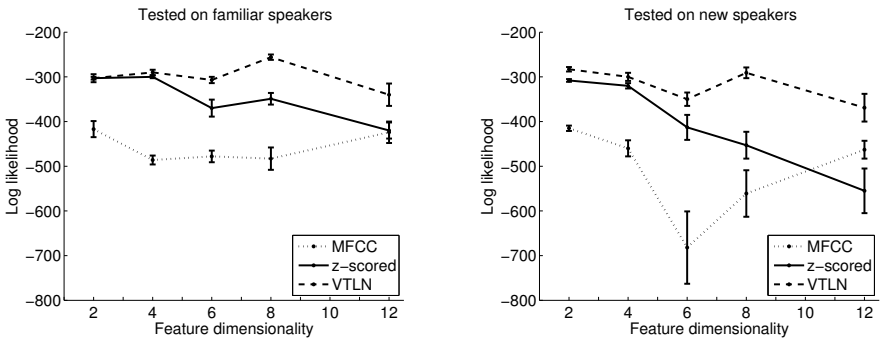


Figure 2: Model fit to human data, when generalizing to familiar speakers (left) and new speakers (right). Results for 1-dimensional features are not shown on this scale as they are extremely poor, with log likelihoods of -1300 to -1600.

across speakers on the basis of /i/, which differs little across the dialects (Clopper & Pisoni, 2006), has minimal effects on dialect K-L divergence: cross-dialect differences remain nearly as distinct after VTLN. Overall, when z-scoring by speaker K-L divergence is lowest (though nonzero; gender and dialect information remains in the representation); VTLN removes somewhat less of the gender and dialect variation.

In summary, MFCCs, z-scored MFCCs, and MFCCs with VTLN each correspond to different distributions of vowels in the input, with z-scoring being the most effective at increasing the overlap between the distributions of vowels spoken by different speakers. The next section describes a cognitive model that uses these input distributions to quantitatively predict listeners' vowel discrimination in the laboratory.

Cognitive model

The model of discrimination we adopt has been shown to accurately predict listeners' discrimination of both vowels and consonants (Feldman et al., 2009; Kronrod, Coppess, & Feldman, 2012), but has not yet been implemented directly on speech recordings. The model formalizes speech perception as an inference problem. Listeners perceive sounds by inferring the acoustic detail of a speaker's target production through a noisy speech signal. Because listeners need to correct for uncertainty in the speech signal, their perception is biased toward acoustic values that have high probability under their prior distribution. This creates a dependency between the listeners' prior distribution and their perception of sounds.

Formally, speakers and listeners share a prior distribution over possible acoustic values that can be produced in the language, $p(T)$. Prototypical sounds in the language have highest probability under this distribution, but the distribution is nonzero over a wide range of acoustic values, corresponding to all the ways in which speech sounds might be realized. When producing a sound, speakers are assumed to sample a target production T from this distribution. The target production can carry meaningful information aside from category identity, such as dialect information or coarticulatory information about upcoming sounds, making its acoustic value something that listeners wish to recover. The stimulus S heard by listeners is similar to the target production T, but is assumed to be corrupted by a noise process defined by a Gaussian likelihood function, $p(S|T) = \mathcal{N}(T, \Sigma_S)$. Both T and S are d-dimensional vectors, where d is the dimensionality of the feature space. In our simulations, the feature space is defined by either MFCCs, z-scored MFCCs, or MFCCs with VTLN.

Listeners hear S and reconstruct T by drawing a sample from the posterior distribution, $p(T|S) \propto p(S|T)p(T)$. We refer to a listener's sample from the posterior distribution as a percept. The percept is a continuous acoustic value, rather than a category label; this is consistent with a large body of evidence showing that listeners recover fine-grained acoustic detail from the speech signal (e.g., Pisoni & Tash, 1974).

The model can be used to predict listeners' discrimination behavior in the laboratory. In AX discrimination tasks, listeners hear two stimuli and decide whether they are acoustically identical. The model assumes that for each stimulus, listeners sample a percept from their posterior distribution, $p(T|S)$. They then compute the distance between their percepts of the two stimuli and compare it to a threshold ε. If the percepts are separated by a distance less than ε, listeners respond *same*; otherwise they respond *different*. Given these assumptions, the proportion of *same* responses for two stimuli, S_1 and S_2, is predicted to follow a binomial distribution whose parameter is the probability that the percepts for the two sounds are within a distance ε of each other, $p(|T_1 - T_2| < \varepsilon | S_1, S_2)$. The noise covariance Σ_S and the response threshold ε are free parameters which are optimized to best predict discrimination data.

Whereas previous work with this model has estimated listeners' prior distribution from perceptual categorization data, here we estimate listeners' prior distribution directly from production data in the NSP corpus; we make the simplifying assumption that the prior distribution directly mirrors the distribution of sounds in the input. Previous work has also assumed that listeners' prior distribution is a mixture of Gaussians, with one Gaussian distribution corresponding to each phonetic category. We avoid making this assumption by using an exemplar-based implementation of the model.

Shi, Griffiths, Feldman, and Sanborn (2010) showed that exemplar models provide a way of approximating Bayesian inference. Specifically, exemplar models implement a form of approximate inference known as importance sampling. To

use importance sampling for our simulations, we need a set of exemplars $\{T^{(i)}\}$ that are sampled from listeners' prior distribution $p(T)$. We assume the vowels in the NSP constitute this set of exemplars. We then weight each exemplar by its likelihood with respect to a stimulus S, $p(S|T^{(i)})$, and select an exemplar according to its weight. An exemplar from the corpus sampled in this way behaves as though it were drawn from the posterior distribution $p(T|S)$. This method does not require us to know a parametric form for the prior distribution $p(T)$, because the prior distribution is represented only through exemplars. In addition, it does not require the exemplars from the corpus to have category labels, as the weights $p(S|T^{(i)})$ are defined by the model's Gaussian likelihood function, corresponding to the speech signal noise.

We estimate the model's probability of responding *same* on each trial by using importance sampling to obtain 100 pairs of percepts corresponding to the pair of stimuli presented in that trial. The proportion of these pairs of percepts that are within distance ε of each other provides an estimate[3] of the probability of responding *same* on that trial. We use these probabilities to predict listeners' actual *same-different* responses in an experiment. We implement the model several times with different speech representations: MFCCs, z-scored MFCCs, or MFCCs with VTLN. Comparing model likelihoods across the three speech representations allows us to ask which representations best predict listeners' discrimination.

Simulations

Simulations implemented the perceptual model with normalized and unnormalized representations, comparing model predictions to human discrimination data. To the extent that different representations of speech yield different distributions of sounds in a corpus, they should make different predictions about the biases that listeners will exhibit in a speech perception experiment. Representations that yield more accurate predictions can be assumed to contain information more similar to the dimensions that listeners use in speech perception.

Human perceptual data We use vowel discrimination data from an AX discrimination task conducted by Feldman et al. (2009) in quiet listening conditions. Twenty participants heard a continuum of 13 isolated vowels that were synthesized to simulate a male speaker. First and second formants of these stimuli ranged linearly in mels from /i/ (as in 'beet') to /e/ (as in 'bait'). Participants heard all ordered pairs of stimuli and judged whether each pair was acoustically identical. MFCCs, z-scored MFCCs, and MFCCs with VTLN computed from these thirteen stimuli serve as input to the model, as the stimulus S. Model predictions are then compared with listeners' *same-different* responses to each pair of stimuli.

Speech representations The exemplars that serve as the model's prior distribution are vowels from the Vowels subcorpus of the NSP. Vowels were represented either as MFCCs, z-scored MFCCs, or MFCCs with VTLN, computed at their midpoint. Although speech recognition systems typically use 12 MFCC dimensions, we additionally include simulations that omit subsets of the higher dimensions, as the lower dimensions are better able to capture information from formants traditionally used to describe vowel quality.

The NSP is an ideal case for the z-scoring normalization, because each speaker says the same tokens the same number of times. However, MFCCs for the stimulus 'speaker' were only available for the /i/-/e/ vowel continuum. Because the stimuli were originally synthesized according to average formant values for male speakers, we handled this missing data by normalizing the stimuli according to average z-scoring factors of the 30 male NSP speakers. Due to the reliance of our VTLN procedure on only /i/ tokens, this normalization was straightforward to apply to the stimuli.

Fitting parameters The NSP corpus was divided into two balanced, equally sized sets of exemplars. Two methods were used for dividing the corpus. In one case, the two halves contained different exemplars from the same speakers, while in the other case the two halves contained exemplars from different speakers (balanced for gender and dialect region). Each division of the corpus was created once, and used for simulations with all speech feature types.

In each simulation, one set of exemplars was used to fit parameters: the response threshold ε and the noise variance Σ_S (constrained to be diagonal) were selected using Markov chain Monte Carlo to maximize model likelihoods given perceptual data.[4] The other set of exemplars was used to compute model likelihoods at test. The roles of the two sets of exemplars were then reversed, resulting in 2-fold cross-validation. Each set of exemplars served as a test set for 10 simulations. Points and error bars in Figures 2 and 4 represent means and standard error calculated across all 20 simulations; the relatively small error bars indicate that results were consistent across replications.[5]

Results Model performance is assessed by computing the log likelihood of the model, given the human data. Higher log likelihoods indicate a closer match to perceptual data.

Previous work with categorization models suggests that normalized representations are more consistent with listeners' perception than unnormalized representations (Apfelbaum & McMurray, 2015; McMurray & Jongman, 2011) . We replicate this result with our method: in almost all cases, unnormalized MFCCs have the lowest likelihoods among the three representations tested (Figure 2). We also find that MFCCs normalized by VTLN outperform z-scored MFCCs, although z-scoring within speakers neutralizes more inter-speaker variation as measured by K-L divergence (Table 1). Thus, the better performance of the VTLN models is not merely an artifact of acoustic similarity among vowels in the corpus; this also may imply that human representations underlying vowel discrimi-

[3] We use add-one smoothing to compute this estimate.

[4] Z-scoring and VTLN were applied prior to parameter fitting; therefore neither adds free parameters to the model, and no optimization to fit human data is involved in their application.

[5] Numerical values fit for model parameters were also consistent across the 20 replications for each speech feature type.

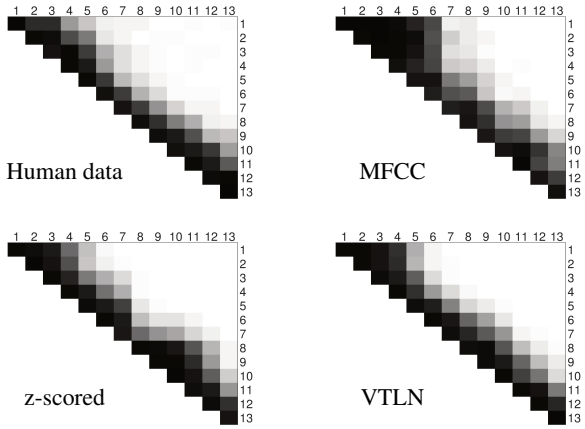


Figure 3: Confusion matrices for how often each pair of stimuli is judged to be the same (black) vs. different (white) by humans and by 4-dimensional models tested on familiar speakers. Axis labels denote stimulus numbers from an AX trial.

nation do not completely normalize across speakers. Examples of *same-different* responses from models using each type of feature are shown in Figure 3, together with human data.

Models using speech representations of one dimension are always extremely poor; the first cepstral coefficient on its own does not provide enough information for the perceptual task. In some simulations, such as for z-scored features, test likelihood decreases with higher numbers of dimensions beyond a certain point, likely due to overfitting of parameters to particular exemplars.[6] It appears that the lower MFCC dimensions, particularly the second dimension, contain information relevant to listeners' discrimination of an /i/-/e/ continuum.

Although the NSP's speaker and phoneme labels are not used by our cognitive model, we take advantage of this information in analyses for insight on the types of exemplars sampled as percepts by the model. Across all models, the 100 percepts drawn from the posterior distribution for each stimulus contained on average 30 different exemplars, indicating that a number of exemplars (from different speakers) are treated as linguistically similar to each other. As a measure of model quality and interpretability, we examine the identity of the exemplars sampled by the model during each perceptual judgment (Figure 4). The percentage of samples that belong to the classes of vowels along the /i/-/e/ continuum (NSP 'heed', 'hayed', and 'hid' tokens; henceforth referred to as high front vowels) gives information on model quality, because all the stimuli are perceived by US English speakers as falling along this continuum. The proportion of times a model samples female exemplars to recover the linguistic target for this experiment's male-speaker stimuli gives an indication of the model's ability to generalize linguistic content across genders.

Models using unnormalized MFCCs tend to make the least use of female exemplars, indicating that this representation does not recognize very much similarity between male and female speakers saying the same vowel. Models with z-scored features are closest to sampling 50% female exemplars; this confirms that the z-scored representation is highly effective at neutralizing difference between speakers of different genders (Table 1), although it is not the representation that gives the best match to human perceptual performance (Figure 2).

While simulations with 2 through 6 dimensions consistently treated the experimental stimuli as being most similar to high front vowels in the corpus, simulations with higher orders of cepstral coefficients did not (Figure 4), reinforcing the importance of the lower MFCC dimensions in capturing listeners' perception of these stimuli. We suspect that this behavior emerged due to the artificial synthesis of the experimental stimuli, which resulted in these high front vowel stimuli being most similar to low back vowels from the corpus in two of the higher MFCC dimensions. This can cause the model to perceive stimuli as low back vowels in cases where it generalizes along those dimensions. This underscores the difficulty of bringing together ecologically valid speech corpora with the type of controlled stimuli typically used in experimental settings, and illuminates areas in which future research may provide insight by addressing these challenges.

Finally, we can compare the log likelihoods from Figure 2 to a benchmark showing ideal model performance. Feldman et al. (2009) estimated listeners' prior distributions for the /i/ and /e/ categories from perceptual categorization data, rather than from speech recordings. Using their estimate, the model yields average log likelihood of -233, somewhat higher than those obtained here. A corpus-based model should approach this value as the distribution of sounds in the corpus approaches the prior distribution that listeners use in perceptual tasks.

Discussion

In this paper a novel method was introduced for evaluating hypotheses regarding the dimensions that guide listeners' speech perception. A cognitive model of AX discrimination was implemented directly on speech recordings and used to evaluate two speaker normalization methods. Both normalization methods improved the model's fit to perceptual data, consistent with previous research. Between the two normalization methods, VTLN outperformed z-scoring, despite being less effective at collapsing gender and dialect differences.

The advantage of VTLN over z-scoring in modeling human perceptual data suggests that VTLN allows the model to generalize across speakers in a way that is more similar to human perceivers. For example, listeners may track statistical distributions of speech in ways that allow them to collapse across gender while retaining differences across dialects. Nevertheless, a prior distribution estimated from perceptual categorization data still outperforms the prior distributions measured from a corpus; none of the three representations tested here match human perception exactly. Instead, these simulations provide initial clues to the dimensions that guide listeners' perception

[6] The corresponding training likelihoods for z-scored dimensions remained stable or even increased at higher numbers of dimensions. Similarly, the low likelihood observed at six dimensions for MFCCs was due to several runs that achieved high likelihood on the training exemplars and low likelihood on the test exemplars.

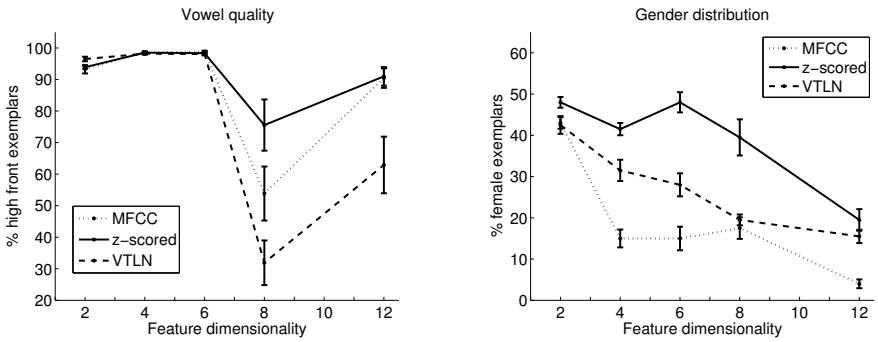


Figure 4: Secondary evaluations, showing how often the stimuli were correctly perceived as high front vowels (left) and how often the model based perceptions on female exemplars (right). Data are averaged across the familiar and new speaker test cases, which were very similar on these measures.

while also allowing us to validate a novel method for assessing speech representations against human data.

Our method provides a general tool for investigating proposals regarding the dimensions that guide listeners' perception. The model's prior distribution can in principle be estimated from any speech corpus, and does not require phoneme labels. This ability to make use of unlabeled corpora provides an advantage over evaluation methods that rely on categorization. In addition, the importance sampling approximation used here can be implemented on any speech representation for which a likelihood function $p(S|T)$ can be computed between stimuli and exemplars and for which a distance metric between exemplars in the corpus can be compared to a threshold ε, and thus can be used even for representations that lack a fixed or finite set of dimensions. This flexibility makes it a promising tool for exploring cross-linguistic differences in listeners' sensitivity to perceptual dimensions, as well as for evaluating theories of dimension learning against children's discrimination data.

To our knowledge, this is also the first time the model from Feldman et al. (2009) has been implemented on speech recordings. Modifying models to operate over corpora of natural speech allows them to make use of ecologically valid datasets, and can thus facilitate a richer understanding of the way in which listeners' perception is shaped by their environment.

Acknowledgments We thank Josh Falk for help piloting the model, Phani Nidadavolu for help computing speech features, and Hynek Hermansky, Bill Idsardi, Feipeng Li, Vijay Peddinti, Amy Weinberg, and the UMD probabilistic modeling reading group for helpful comments. This work was supported by NSF grant BCS-1320410.

References

Adank, P., Smits, R., & Hout, R. v. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, *116*, 3099-3107.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin and Review*, *22*, 916-943.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804-809.

Clopper, C. G., & Pisoni, D. B. (2006). The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, *48*, 633-644.

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167-184.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Proceedings of the IEEE*, 357-366.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752-782.

Giuliani, D., Gerosa, M., & Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech & Language*, *20*(1), 107–123.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099-3111.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (p. 145-165). New York: Academic Press.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A unified model of categorical effects in consonant and vowel perception. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, *49*, 606-608.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219-246.

Nearey, T. M. (1978). *Phonetic feature systems for vowels* (Vol. 77). Indiana University Linguistics Club.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, *15*, 285-290.

Povey, D., & Saon, G. (2006). Feature and model space speaker adaptation with full covariance Gaussians. *Proceedings of Interspeech*.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*, 443-464.

Wang, Q., Kulkarni, S. R., & Verdú, S. (2006). A nearest-neighbor approach to estimating divergence between continuous random vectors. *Convergence*, *1000*(1), 11.

Wegmann, S., McAllaster, D., Orloff, J., & Peskin, B. (1996). Speaker normalization on conversational telephone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 339-341.

Facilitating Spatial Task Learning in Interactive Multimedia Environments While Accounting for Individual Differences and Task Difficulty

Dar-Wei Chen (darwei.chen@gatech.edu)
School of Psychology, Georgia Institute of Technology,
654 Cherry Street, Atlanta, GA 30332-0170, USA

Richard Catrambone (richard.catrambone@psych.gatech.edu)
School of Psychology, Georgia Institute of Technology,
654 Cherry Street, Atlanta, GA 30332-0170, USA

Abstract

Two experiments examined the effects of interactive tutorial features (compared to "passive" features) on learning spatial tasks, an area seldom explored in interactivity research. Experiment 1 results indicated that for simple spatial tasks, interactive tutorials hindered learning for participants of higher spatial ability but improved learning for lower-ability participants. This interaction can be explained by "compensation," the notion that people of higher ability can compensate for poor external support (passive tutorials) while people of lower ability need the better support. It is likely that the increased cognitive load of interactivity (Kalyuga, 2007) hindered high-spatial participants on a relatively easy task. In Experiment 2, task difficulty was increased, and the results revealed that the interactive tutorial produced better learning than the passive tutorial, regardless of spatial abilities. With the relatively difficult task, the benefits of interactivity became clearer because most people actually needed the interactive features despite the associated cognitive load.

Keywords: user interactivity; learning technology; spatial learning; multimedia; individual differences

Background

Education has changed markedly since "paper and pencil" was the dominant form of learning. An internet connection is now the only requirement for learners to reach previously-inaccessible worlds of information. Therefore, not surprisingly, technology-driven learning is on the rise. For example, a survey by the Sloan Consortium showed that almost three-quarters of universities report increasing demand for online courses (Parry, 2010).

Improving the quality of technology-driven learning is in the interest of educators and the public at large. One of the more intriguing features of educational technology is multimedia and any discussion about improving educational technology would be incomplete without studying how features of multimedia (materials that incorporate elements such as text, images, animation, video, interactivity, etc.) can be harnessed for positive learning outcomes.

One of the current discussions in technology-driven learning concerns the impact of user interactivity on learning outcomes. Interactivity can be defined as "reciprocal activity between a learner and a multimedia learning system, in which the [re]action of the learner is dependent upon the [re]-action of the system and vice versa" (Domagk, Schwartz, & Plass, 2010, p. 1025). Three aspects of interactivity proposed by Moreno and Mayer (2007) are: pacing (controlling speed of information presentation), manipulating (controlling aspects of information presentation), and navigating (selecting information sources); these aspects were investigated in this study.

Some research suggests that students learn best when material is interactive; that is, when the user has a relatively high degree of control over the material and the actions of the user and material are closely related to the actions of the other. For example, Schwan and Riempp (2004) found that people using interactive learning tools could accurately complete knot-tying tasks with half of the practice time as those using non-interactive tools; they posited that interactivity allowed participants to tailor their information gathering by helping them more easily distribute study time to difficult tasks. A study by Khalifa and Lam (2002) showed that the understanding exhibited by users in "interactive conditions" was indicative of knowledge about how concepts link together; the users in "passive conditions" understood the material at merely a "list-like" level. In research about learner control, students have been found to better optimize their experiences and generalize findings when in control (Gureckis & Markant, 2012).

In some ways, interactive materials could confer educational benefits similar to those of "minimal guidance" interventions, which generally hypothesize that people learn more effectively when required to discover information on their own, as opposed to being directly instructed. Interactive interventions and minimal-guidance interventions allow learners to take control of the learning process, leading them to assess their knowledge gaps and methods to overcome them, both of which are productive activities (Chi, 2000). For example, with some interactive materials, a learner might have to choose a method of gathering information that will yield the information he or she needs; when directly instructed, learners tend not to have to assess their own learning as much.

However, not all research demonstrates consistent benefits for usage of interactive materials. Sometimes, interactivity becomes a mental burden on learners by introducing "non-essential extraneous processing load" (Kalyuga, 2007). For example, Moreno and Mayer (2005) found that interactivity helps people achieve "meaningful learning" only when sufficient guidance exists within the

interface. Therefore, interactivity seems not to be an inherently positive feature in learning materials; instead, the amount and aspects of interactivity are important keys in creating useful multimedia instruction.

The present study provides a nudge in clarifying some aspects of multimedia interactivity with respect to their impact on learning of spatial tasks. The chosen task for this study was partially solving Rubik's Cube, a spatial task that requires pattern recognition and inference-making. This task is of interest because it is unlike tasks that are driven by declarative knowledge; much of the existing interactive learning research already focuses on declarative knowledge. One of the goals in this study was to determine whether some aspects of multimedia interactivity could be used to foster understanding of tasks that are driven by spatial pattern recognition and inference-making, as opposed to primarily declarative knowledge (e.g. memorizing the capitals of all fifty states) or procedures (e.g. operating a vacuum cleaner). Another goal of this study is to more realistically represent passive conditions by allowing users the option of exercising basic control over videos as they would have in most online learning interfaces. Many studies in the past have created passive conditions in which users did not have the option of controlling any aspect of videos, situations that do not often arise with the modern internet.

Experiment 1

Participants used either an interactive or a "passive" video-based tutorial to learn how to create "the cross" on Rubik's Cube. Creating the cross involves two types of pieces: center pieces (in the middle of each cube face) and edge pieces (pieces with colors on two sides, as opposed to corner pieces that have colors on three sides). Participants were to A) place edge pieces around the yellow center such that a yellow "plus sign" was formed, and B) align the secondary (non-yellow) colors of those edge pieces with the colors of the adjacent centers. Figure 1 visually explains the cross.

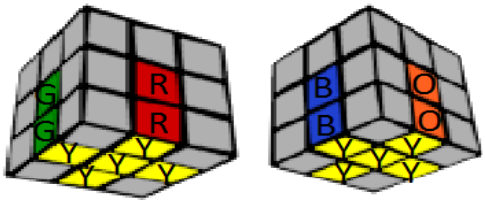


Figure 1: "The cross," shown from two angles
(Y = yellow, G = green, R = red, B = blue, O = orange)

Method

Participants Participants were 31 college students (18-22 years in age; 15 interactive, 16 passive) who had no self-reported prior experience in systematically learning the Rubik's Cube (e.g. looking up instructions online, learning from a friend). The students received course credit for their participation and were randomly assigned to conditions.

Materials Two types of tutorials (shown in Figure 2):

- The interactive tutorial presented a series of Java applets featuring an on-screen cube whose faces would turn depending on user input. Each major step (and the moves within each step) were presented in succession. Learners could manipulate the learning pace through the use of the "step-by-step" buttons, navigate cleanly between move sequence animations through the use of the indexed steps, and manually rotate the virtual cube.
- The "passive" tutorial included a series of videos that are comparable in user control to videos commonly found on websites. That is, the videos are played straight through by default unless rewound, fast-forwarded, or paused. All information available in the interactive tutorial was available through the videos, although the videos afforded less control in its presentation of information.

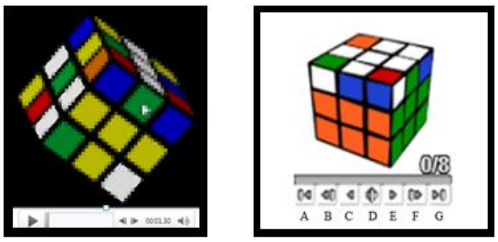


Figure 2: Passive tutorial (L) and interactive tutorial (R)

Design and Procedure The effects of tutorial type on participants' learning were examined with a between-subject design. Before starting the tutorial, each participant completed a demographics form and a spatial ability test (on-screen paper is folded, has a hole punched through it, and participants must predict what the paper looks like when unfolded again; Ekstrom et al., 1976).

The study started with participants being allotted eight minutes to read a Rubik's Cube introduction while having a cube available to use. After the reading period, participants were given 20 minutes to access their assigned tutorial.

After the tutorial phase, participants moved onto the assessment, in which they were given scrambled cubes (i.e., none of the edge pieces were already placed into the cross) and instructed to construct the cross within a four-minute period (cubes were physical, not virtual); inference-making was tested as the scrambled cube configuration differed from the cubes originally given to participants. Participants' times were recorded if they were able to create the cross.

Assessment Scoring Scheme Performances on the assessment were scored using a two-tiered scheme. This scheme was developed by one of the researchers and reviewed for face and external validity by 2007 Florida Open Rubik's Cube champion Andrew Chow (A. Chow, personal communication, November 27, 2012). Participants could score a maximum of six points in Tier 1. It encompasses the first step of creating the cross without regard for matching center colors. Points are awarded for each individual yellow edge piece placed around the yellow

center; the third and fourth edge pieces are weighted more heavily because placing those sometimes involves moving other edge pieces out of place, thus requiring more awareness. Tier 2 has a maximum of four points, and it involves matching the secondary colors of the edge pieces to the colors of their adjacent centers. The possible scores for a cube in Experiment 1 are shown in Table 1.

Table 1: Scoring possibilities for Experiment 1 (Tier 1 points must be complete before Tier 2 points can be earned).

Tier	Cross pieces in place	Number of matching centers	Score
Tier 1	1	0	1
	2	0	2
	3	0	4
	4	0	6
Tier 2	4	1	6
	4	2	8
	4	4	10

Results and Discussion

Main Effect of Tutorial Interactivity No significant differences in performance were found between the interactive condition participants (M = 8.33, SD = 2.29) and passive condition participants (M = 8.81, SD = 2.23), F (1, 29) = 0.349, MSE = 5.10, $p > 0.05$, d = 0.21 (non-parametric Mann-Whitney U test similarly non-significant, p = 0.572). The difference in solve time between those in the interactive condition (M = 93.29s, SD = 55.71) and those in the passive (M = 84.5s, SD = 41.04) was also found to be non-significant, F (1, 11) = 0.157, MSE = 2,458.08, $p > 0.05$.

A few explanations are possible for the relative similarity in performances between the two conditions. One revolves around a limitation of the study: The given assessment task might have been completed too easily by the participants, leading to ceiling effects in the achievement scores – more than two-thirds of the participants (21 out of 31) achieved a perfect score. With a majority of scores bunched at the high end of the scale, leaving little room for performance differences to be expressed in the data, any actual effects of the tutorial types would have been difficult to find. Experiment 2 addressed this issue by requiring participants to complete a more difficult cube-related task within the same time constraints.

In terms of the study manipulation, the possibility exists that for this relatively easy task, the differences between the tutorials were not large enough to elicit significantly different user actions (future studies should implement manipulation checks regarding how learners actually used the respective tutorials); that is, the way the participants used the interactive tutorial might not have been significantly different from how participants used the passive tutorial. For example, interactive participants had the option of allowing whole sequences of moves to play consecutively without stopping, using the interactive applets almost like videos. For a relatively easy task, such an approach might have been sufficient.

However, even if participants did use the interactive tutorial differently from the passive one, those differences might not have revealed themselves in the scores. For example, one of the main distinguishing features of the interactive tutorial was the presence of "step-by-step buttons" that allowed participants to scroll through individual moves with self-selected pacing and relative ease. However, as access to internet videos becomes increasingly commonplace, especially among college students, perhaps the passive condition's tasks of rewinding or fast-forwarding to search through a video is no longer much of a mental burden relative to searching through the use of the buttons. In fact, the interactive condition could even have introduced its own larger mental burdens with the emphasis it placed on user control of the tutorial (Kalyuga, 2007), negating any potential benefits of the interactivity.

Individual Differences in Spatial Ability As might be expected, a participant's spatial ability had a significant positive correlation with his or her achievement score (r = 0.415, p = 0.05); spatial ability also had a significant negative correlation with solve time (r = -0.495, p = 0.016). In short, these results indicate that participants of high spatial ability generally performed better on the assessment task and finished the task more quickly than those of low spatial ability. In this study, "high-spatial" participants were those with spatial abilities above the median and "low-spatial" participants were those lower than the median. Furthermore, using a linear regression, spatial ability was found to uniquely account for a significant amount of achievement score variance that condition could not account for itself (r-$change$ = 0.154, p = 0.033). To explain findings such as this one, Mayer and Sims (1994) posit that people with high spatial ability are able to achieve – while using fewer cognitive resources – the same understanding of the multimedia instructions as people with low spatial ability; therefore, they can transfer more of their resources to the actual task at hand (the cube, in this case).

Interaction Between Condition and Spatial Ability. As stated previously, the difference in mean achievement scores between the two tutorial types was not statistically significant. However, an ANOVA did demonstrate a significant interaction between tutorial type and spatial ability, F (1, 26) = 4.27, MSE = 4.12, p = 0.049. More specifically, the interaction suggested that low-spatial participants benefited more from interactivity than high-spatial participants did; that is, high-spatial participants scored about 46% higher than low-spatial participants when using the passive tutorial, while the differences between the participants were negligible when they used the interactive tutorial (larger sample sizes are needed for post-hoc tests to be conducted, however). Figure 3 illustrates this finding.

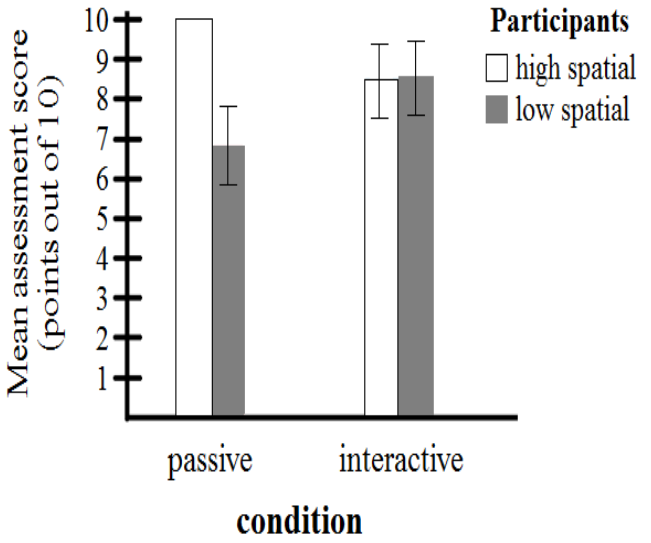


Figure 3: Interaction between condition and spatial ability

An explanation for this result is "compensation," the notion that people of high spatial ability can compensate for ostensibly weaker external support (as experienced in the passive condition) while people of low spatial ability benefit from stronger external support (interactive condition). In this particular study, high-spatial participants were likely able to better mentally visualize and compensate for the information that the passive condition did not present as well, while low-spatial participants were significantly aided by interactivity because they had lower capacities to compensate and fill the information gaps themselves. In other words, interactivity was unnecessary for high-spatial participants, but made a positive difference for those on the lower end of the spectrum. This conclusion aligns with a finding from Hoffler and Leutner (2010) that high-spatial people generally learned well from either multiphase diagrams (low support) or animations (high support), whereas low-spatial people needed the animations to perform relatively better than they did with the diagrams.

Experiment 2 examined whether the effects of interactivity and spatial ability would change on a more difficult task (which also served to counter aforementioned ceiling effects). It was hypothesized that, in a task that could not be easily completed without good support, the benefits of interactivity would become more evident because the benefits would outweigh the associated cognitive load.

Experiment 2

Procedures for Experiment 2 were identical to those of Experiment 1 with the exception of the task assigned to the participants. The participants used either an interactive tutorial or a passive tutorial to learn how to create "the first layer," which is one step further than the cross. Creating the first layer involves creating the cross and then placing the appropriate corner pieces around it such that the yellow side is complete and the colors of the corner pieces match the adjacent edge pieces. Figure 4 illustrates the first layer.



Figure 4: The "first layer" shown from three angles.

Method

Participants Participants were 47 college students (18-22 years in age; 23 interactive, 24 passive) who had no self-reported prior experience in systematically learning the Rubik's Cube. The students received course credit for their participation and were randomly assigned to conditions.

Materials. All of the content related to the cross was identical to that of Experiment 1. The tutorials were extended to include the additional concepts that participants had to learn in order to create the first layer.

Design and Procedure. Participants experienced the same procedures as the participants in Experiment 1. That is, even though Experiment 2 required participants to learn more material, they were still allowed just 20 minutes during the tutorial phase and 4 minutes for the assessment. The time limit was held constant between experiments to increase the likelihood that the task for Experiment 2 was indeed more difficult than the task for Experiment 1.

Assessment Scoring Scheme. The 10-point scheme for the cross was kept in place, but 10 more points were added to account for the additional points possible for adding corner pieces; therefore, 20 points was the maximum possible score for creating the first layer. Table 2 outlines the 10 additional points for the corner pieces.

Table 2: Scoring possibilities for corner pieces in Experiment 2 (to be used if cross is completed).

Cross pieces in place	Number of matching centers	Corner pieces in place	Score
4	4	0	10
4	4	1	13
4	4	2	16
4	4	3	18
4	4	4	20

This scheme was reviewed by 2007 Florida Open Rubik's Cube champion Andrew Chow (A. Chow, personal communication, May 11, 2015).

Results and Discussion

Main Effect of Tutorial Interactivity Participants using the interactive tutorial (M = 11.04, SD = 6.06) significantly

outperformed those using the passive tutorial (M = 7.46, SD = 5.27), t (45) = 2.17, p = 0.04, d = 0.63 (see Figure 5). Solve times were not compared due to the low number of participants who were able to actually complete the task.

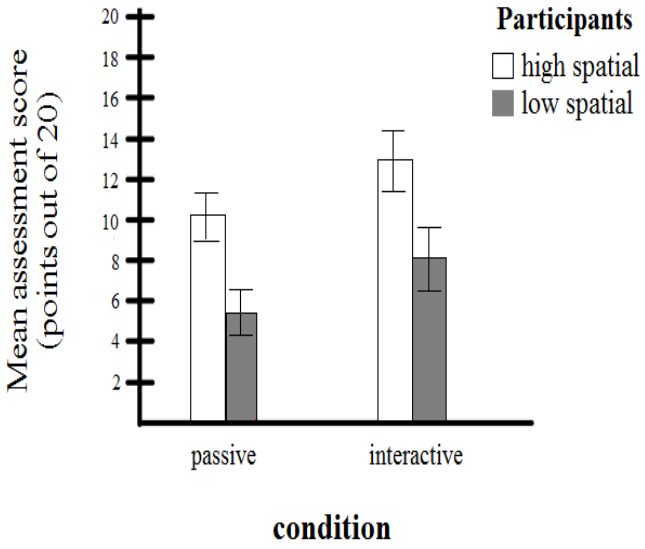


Figure 5: Assessment scores by condition and spatial ability

Participants using the interactive tutorial were originally hypothesized to score more highly on the assessment, and the results in Experiment 2 support that hypothesis. The interactive condition participants likely performed better because their tutorials encouraged user control of the learning process (Zhang et al., 2006). For example, as the interactive interface stopped after each individual move of a sequence, participants might have been relatively likely to self-question and reflect on their understanding of the move. If they did not understand, the "go back one move" button easily allowed participants to see the move again. The videos in the passive condition did not stop at the conclusion of individual moves unless the participant stopped it, and precise rewinding to view a move again was not as easy as in the interactive condition.

Another feature that encouraged reflection was the cube rotation mechanism, which allowed interactive condition participants to manually rotate the on-screen Rubik's Cube at any time to view the positions of any pieces that were of interest. This feature helped participants to become more aware of their knowledge gaps because they had to deliberately find the information to fill those gaps when they were stuck. In the passive condition, the participants potentially received all of the same information because their on-screen cubes were rotated automatically, but the participants were perhaps less likely to know which knowledge gap was being filled by the presented information given their reduced control over the tutorial.

Both of these interactive features must be more deeply examined in the future for a better understanding of how learners used them for reflection. Of course, the interactive features existed in the interactive condition tutorials for Experiment 1 as well, but they produced different performance effects because of the lower task difficulty in Experiment 1. Interactivity proved to be useful for both high- and low-spatial participants in Experiment 2 because the task was difficult enough to require external support; interactivity was not a superfluous feature for the high-spatial participants like in Experiment 1. Interactivity was not necessarily useful to high-spatial participants in Experiment 1 because the mental burdens of interactivity (Kalyuga, 2007) outweighed the benefits on a task they likely could have done well on without much support; low-spatial participants needed the support (compensation).

The reflection presumably encouraged in the interactive condition might have aided processes of metacognition. In the passive condition, participants might have known that a move was to be done, but be relatively unsure about the reasoning because they were less likely to be confronted with their gaps in knowledge at the right time. Participants in the interactive condition were more likely to learn *why* a move was done because the interface encouraged them to think about it through A) the step-by-step emphasis in the interface, and B) the user control of turning the cube for needed information. This deeper-level knowledge was likely the reason that interactive participants performed better on the assessment; the assessment required transfer and fundamental knowledge because participants were given a newly-scrambled cube for the assessment that differed in initial state from the one they used in the tutorial, diminishing the effectiveness of rote memorization.

Metacognitive processes demand mental resources, and sometimes can hinder performance (Kalyuga, 2007). However, the benefits of metacognition apparently outweighed the drawbacks in Experiment 2, corroborating the findings of other studies about the effectiveness of metacognition (e.g., van den Boom, Paas, van Merrienboer, & van Gog, 2004; Kramarski & Michalsky, 2010).

Individual Differences in Spatial Ability Spatial ability was correlated significantly with assessment score (r = 0.45, p < 0.01), in line with results from Experiment 1. However, unlike Experiment 1, no statistically-significant interaction was found between tutorial type and spatial ability, F (1, 43) = 0.01, MSE = 28.00, p = 0.92. Therefore, it can be concluded that the effects of tutorial type were relatively similar across participants of varying spatial abilities.

Video Games and Spatial Ability. A significant correlation was found between time spent playing video games and assessment score (r = 0.44, p < 0.01). The data from the present study do not indicate the nature of the causality, although the correlation between video game hours and spatial ability was also significant (r = 0.37, p < 0.05).

General Discussion

Perhaps the most intriguing finding from these two studies is the difference in data patterns between Experiment 1 and Experiment 2. With the relatively easy task in Experiment 1,

the results were consistent with the theory of compensation: high-spatial participants appeared to perform worse with the interactive features while low-spatial participants performed better with them (interaction effect), F (1, 26) = 4.27, MSE = 4.12, p = 0.049. When the task increased in difficulty for Experiment 2, the effectiveness of the interactive features was no longer dependent on a participant's spatial ability. Instead, there was a main effect of tutorial type, with both high-spatial participants and low-spatial participants receiving an equal boost from the interactive features.

Interactivity is a broad concept not limited to the aspects discussed here: pace, information manipulation, and navigation. Future studies implementing different aspects of interactivity could yield somewhat different results, and more granular data regarding tutorial usage (e.g., how aspects of interactivity affect user actions, cognitive load data) could help researchers identify and describe more exactly the low-level mechanisms driving the effects found here. Another concern is that Rubik's Cube is not necessarily representative of the many types of tasks, or even spatial tasks, that exist. Future research should identify the aspects in interactivity that improve learning, the people who benefit most from using those aspects of interactivity, and situations in which interactivity is most appropriate.

In 2011, almost one-third of US college students had taken at least one online course (Online Learning Consortium, 2012), and that percentage is growing. Countries like Great Britain (invested $100 million in 2011 to boost online learning) and Australia (20% growth between 2007 and 2012) are also seeing online education as a not just a reasonable alternative to "traditional" schools, but a necessity in modern learning (International College of Economics and Finance, 2012). Learning technologies offer many conveniences over standard materials such as portability and information access. They also provide opportunities for user interaction that traditional textbooks cannot match. However, as demonstrated in this paper, interactivity is not "one size fits all." Accounting for individual differences and task-specific details can help educators harness the powers of interactivity to improve learning technologies and outcomes for all users.

References

Chi, M. T. H. (2000). Self-explaining: the dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in instructional psychology, (pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates.

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior, 26* (5), 1024-1033.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Service.

Gureckis, T.M. & Markant, D.B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspectives on Psychological Science, 7* (5), 464-481.

Hoffler, T. N. & Leutner, D. (2010). The role of spatial ability in learning from instructional animations – Evidence for an ability-as-compensator hypothesis. *Computers in Human Behavior, 27* (1), 209-216.

International College of Economics and Finance (2012). 8 countries leading the way in online education. *ICEF Monitor*. Retrieved from: http://monitor.icef.com/2012/06/8-countries-leading-the-way-in-online-education/

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educ. Psychol. Rev., 19*, 387-399.

Khalifa, M. & Lam, R. (2002). Web-Based Learning: Effects on Learning Process and Outcome. *IEEE Transactions on Education, 45* (4), 350-356.

Kline, K. & Catrambone, R. (2011). Learning from Multiphase Diagrams: Effects of Spatial Ability and Visuospatial Working Memory Capacity. In *Proceedings of the HFES 55th Annual Meeting*, 570-574.

Kramarski, B. & Michalsky, T. (2010). Preparing preservice teachers for self-regulated learning in the context of technological pedagogical content knowledge. *Learning and Instruction, 20* (5), 434-447.

Mayer, R. E. & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86* (3), 389-401.

Moreno, R. & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journ. of Educational Psychology, 97* (1), 117-128.

Moreno, R. & Mayer, R.E. (2007). Interactive multimodal learning environments. *Educ. Psychol. Rev., 19*, 309-326.

Online Learning Consortium. (2012). Changing Course: Ten Years of Tracking Online Education in the United States. *Online Learning Consortium*. Retrieved from: http://onlinelearningconsortium.org/survey_report/changing-course-ten-years-tracking-online-education-united-states/

Parry, M. (2010, January 26). Colleges See 17 Percent Increase in Online Enrollment. *The Chronicle of Higher Education*. Retrieved from: http://chronicle.com/blogs/wiredcampus/colleges-see-17-percent-increase-in-online-enrollment/20820

Schwan, S. & Riempp, R. (2004). The cognitive benefits of interactive videos: Learning to tie nautical knots. *Learning and Instruction, 14*, 293-305.

Van den Boom, G., Paas, F., van Merrienboer, J. J. G., & van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: effects on students' self-regulated learning competence. *Computers in Human Behavior, 20* (4), 551-567.

Zhang, D., Zhou, L., Briggs, R.O., Nunamaker, J.F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information and Management, 43*, 15-27.

A Neural Dynamic Model Parses Object-Oriented Actions

Mathis Richter (mathis.richter@ini.rub.de)
Jonas Lins (jonas.lins@ini.rub.de)
Gregor Schöner (gregor.schoener@ini.rub.de)

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44870 Bochum, Germany

Abstract

Parsing actions entails that relations between objects are discovered. A pervasively neural account of this process requires that fundamental problems are solved: the neural pointer problem, the binding problem, and the problem of generating discrete processing steps from time-continuous neural processes. We present a prototypical solution to these problems in a neural dynamic model that comprises dynamic neural fields holding representations close to sensorimotor surfaces as well as dynamic nodes holding discrete, language-like representations. Making the connection between these two types of representations enables the model to parse actions as well as ground movement phrases—all based on real visual input. We demonstrate how the dynamic neural processes autonomously generate the processing steps required to parse or ground object-oriented action.

Keywords: relations; neural process model; action parsing; dynamic field theory; grounded cognition; cognitive schemas

Introduction

If you were to describe the arrangement of furniture in your office you would probably make use of the spatial relations between different items. You may recognize without effort that "the bookshelf is to the left of the desk" although this relationship is not directly specified by perception and requires active construal. In fact, relational processing may be foundational to higher cognition (Halford, Wilson, & Phillips, 2010).

Evaluating even simple relations requires several coordinated steps (Logan & Sadler, 1996): (1) binding each object to a role (here, the desk is the *reference object*, the bookshelf is the *target object*); (2) centering the reference frame on the reference object; (3) applying a relational operator (here, "to the left of") to the target object in that frame. We have previously realized these computational steps in a neural process model (Lipinski, Schneegans, Sandamirskaya, Spencer, & Schöner, 2012; Richter, Lins, Schneegans, Sandamirskaya, & Schöner, 2014) that mapped spatial relational phrases supplied in a language-like, discrete form, to a visual scene representation, effectively *grounding* amodal input in perception. Conversely, the model was able to complete partial phrases by referring to the visual scene.

Here, we extend this model to the parsing of object-oriented actions and the grounding of action phrases, such as "the red ball is moving toward (or away from) the yellow ball". This is challenging for a neural process model as the perceptual representations are inherently transient. We add two components that extract motion directions through a neural dynamic model of motion detection (Berger, Faubel, Norman, Hock, & Schöner, 2012) and transform the visual scene to an intrinsic reference frame (van Hengel, Sandamirskaya, Schneegans, & Schöner, 2012) based on the motion direction of a reference object. The new model integrates this form of action parsing with the grounding of spatial relationships of the earlier model.

We use dynamic field theory (DFT; Schöner, Spencer, & the DFT Research Group, 2015) as a theoretical framework. DFT describes neural population activity by activation fields that are defined over metric feature dimensions and evolve continuously in time through a neural dynamics. While the fields capture representations in a modal form close to the sensorimotor surfaces, neural nodes sharing the same dynamics enable modeling discrete, amodal representations. Mutual coupling between fields and nodes allows for interaction between these two kinds of representations. By using only tools from the DFT repertoire we arrive at a seamless process account that is pervasively neural. This requires that we solve the following fundamental problems that arise from restrictions of neural mechanism and are not commonly addressed in accounts of higher cognition.

First, information represented by neural activity cannot be freely moved within and between neural populations, because neural connectivity is fixed. In visual cortex, for instance, visual objects are represented in neural maps. Applying a neural operator to a location or an object in such a map is possible only if it is connected to that location. Connecting operators to every location in a map would require unrealistic neural resources. The alternative is to connect the operator to only one default region, a virtual fovea, and shift the representations of objects to that region. This is analogous to the concept of an attentional neural pointer of Ballard, Hayhoe, Pook, and Rao (1997) and is achieved in our framework by steerable neural mappings (Schneegans & Schöner, 2012).

Second, for similar reasons of limiting the required neural resources, the nervous system represents high-dimensional visual information in multiple low-dimensional neural feature maps, in particular in the early tiers of the cortical hierarchy. To refer to any particular object, corresponding representational pieces must be bound together. In a neural implementation of the classical idea of binding through space (Treisman & Gelade, 1980), we endow every feature map with a spatial dimension shared across maps and process objects sequentially in time (Schneegans, Spencer, & Schöner, 2015).

Third, the discrete processing steps this implies and that are critical to all of higher cognition are natural in information processing accounts but hard to achieve in neural process models, in which neural activation evolves continuously in time under the influence of input and recurrent connectivity. In our model, discrete events emerge from continuous neural

dynamics through dynamic instabilities, at which the match between neural representations of *intentional states* and their *conditions of satisfaction* are detected (Sandamirskaya & Schöner, 2010).

Finally, the problem of preserving role-filler binding (Doumas & Hummel, 2012) at the interface between the modal and the amodal representations is also solved by sequential processing.

Methods

Dynamic field theory describes processes that characterize neural activity at the population level. Models in DFT are based on activation patterns defined as dynamic fields, $u(x,t)$, over continuous feature dimensions, x, (e.g., color or space). These activation patterns evolve in time, t, under the influence of lateral interactions and external input based on the following integro-differential equation

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int g(u(x',t))\, w(x-x')\, dx'.$$

Here, the activation's rate of change, $\dot{u}(x,t)$, depends on $u(x,t)$ itself, on a time constant, τ, a negative resting level, h, and external input, $s(x,t)$, from sensors or other fields. Lateral interaction is determined by convolving the output of the field, $g(u(x,t))$, a sigmoid function with threshold at zero, with an interaction kernel, $w(\Delta x)$. The kernel combines local excitation and surround inhibition along the field's feature dimension.

When presented with localized input above the output threshold, lateral interaction leads to an instability, in which a subthreshold solution becomes unstable and the field moves to a new attractor, a self-stabilized activation peak. From such instabilities, neural events emerge at discrete times from the time-continuous dynamics of the fields. These events are critical for organizing sequential processes in DFT models.

Depending on the tuning of their interaction kernel, dynamic fields may either support multiple peaks or may be selective and only create a single peak that suppresses all others. Fields may also be tuned to hold self-sustained peaks that remain even after input is removed. Fields can be defined over single or multiple dimensions. Dynamic nodes share the fields' dynamic characteristics but do not span a feature dimension. Instead, they represent the 'on' or 'off' state of discrete elements within an architecture.

DFT architectures consist of multiple fields and nodes that are interconnected, where the output of one field is input to another field. Fields of different dimensionalities may be connected along the shared feature dimensions.

Architecture

The DFT architecture shown in Fig. 1 can deal with two types of tasks. First, it can ground a language-like phrase such as "the red object moving toward the yellow object", that is, it can find the objects in the scene that correspond to the phrase. Second, it can generate a phrase such as the one above from observing a video. Solving these tasks within a single neural architecture requires integrating various components, which we describe in more detail now.

Perception

The architecture receives video input from a camera or video file. This input is fed into two three-dimensional perception fields (top right of Fig. 1) that hold a representation of the scene. Both fields share the spatial dimensions of the camera image but the *perception color field* represents the color of objects in the scene and the *perception movement field*, new over the previous model (Richter et al., 2014), represents their movement direction. To create the input to the perception fields, each video frame goes through several preprocessing steps. For the color field, the preprocessing is first based on generic image processing algorithms. After these, activation is generated that scales with the color saturation of objects in the scene. For the movement field, the preprocessing consists of a neural dynamic implementation of the counter-change model of motion perception (Berger et al., 2012). Both perception fields always have stable peaks of activation when there are colored or moving objects in the scene. They project activation into the spatial attention fields along the two spatial dimensions and act as a saliency mechanism. They also project directly into the reference and target field and enable these fields to track moving objects even if spatial attention is currently focused elsewhere.

Attention

The core of the attentional system consists of two three-dimensional *attention fields*. They are defined over the same dimensions as the two perception fields but their activation remains below threshold unless additional input arrives from a feature attention field or a spatial attention field.

A pair of one-dimensional fields spans each feature dimension (color and movement direction): the *intention field* represents feature values for guided search and impacts on the three-dimensional attention fields; the *condition of satisfaction (CoS) field* matches input from the attention fields against what is represented in the intention field.

Two *spatial attention fields* are defined over the two spatial dimensions of the camera image. One field allows for multiple simultaneous peaks and projects into the reference and target fields. The other only allows for a single peak; it can be boosted to induce a selection decision on multiple candidate objects. A peak generated in this spatial attention field suppresses activation at all other locations in the other spatial attention field. It further projects into the three-dimensional attention fields, enabling peaks to form there that represent the feature values at the selected location (which then impact on the CoS fields). This implements a neural mechanism of feature binding across space (Schneegans et al., 2015).

Coordinate transformation

The two-dimensional *reference field* and *target field* each represent the spatial position of their respective objects. The tar-

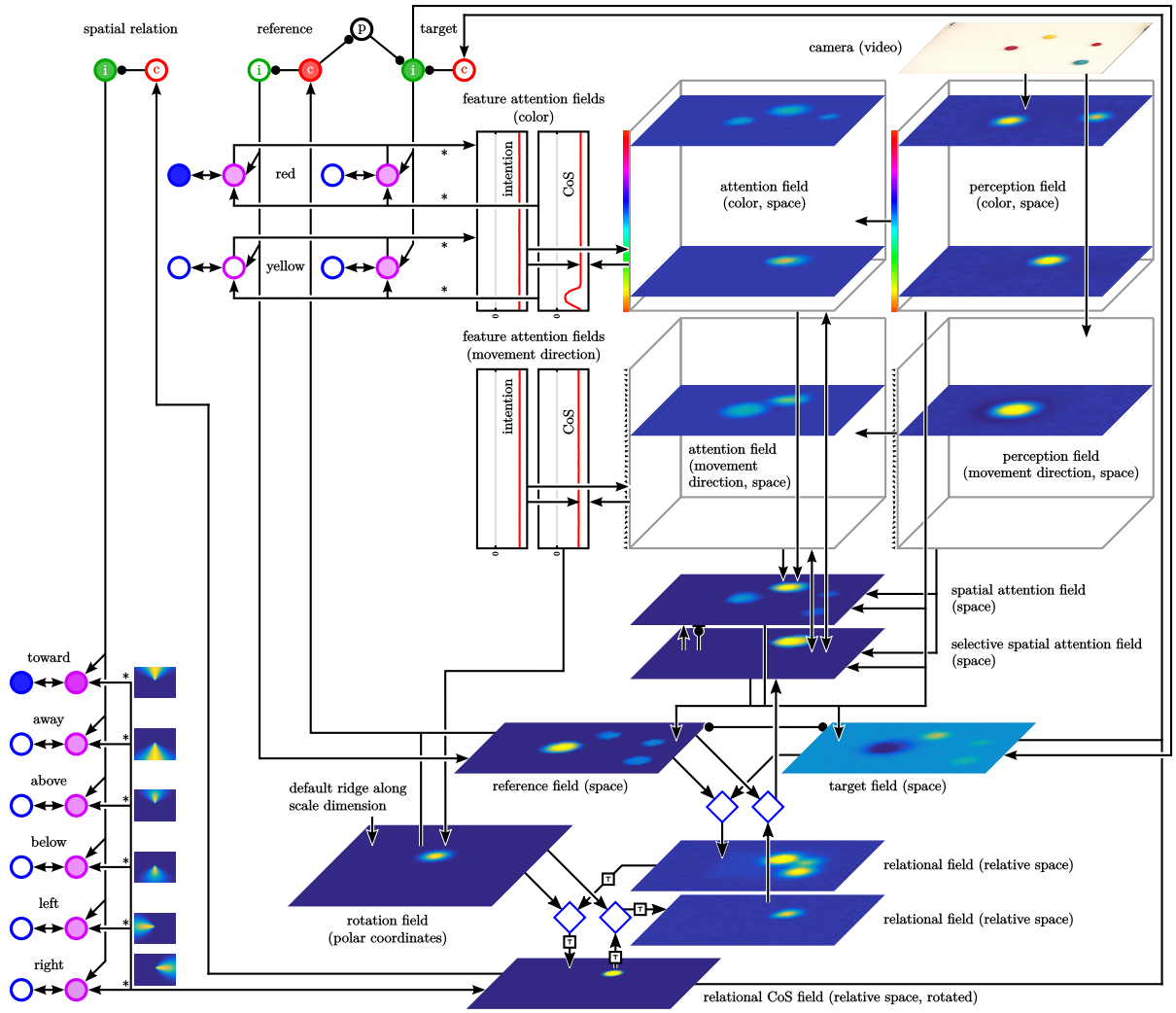


Fig. 1: Architecture with activation snapshots while it is generating a phrase about a video. Fields are shown as color-coded activation patterns; for three-dimensional fields, two-dimensional slices are shown. Node activation is denoted in opacity-coded circles. Spatial templates are illustrated as color-coded weight patterns (bottom left). Excitatory synaptic connections are denoted by lines with arrow heads, inhibitory connections by lines ending in circles. Transformations to and from polar coordinates are marked with a 'T'. Coordinate transformations are denoted as diamonds.

get field projects into the *relational field* via a steerable neural mapping (upper left blue diamond in Fig. 1) that shifts the representation of the target objects so that it is centered on the reference object. This transformation to a new reference frame is implemented as a convolution for performance reasons.

The shifted representation of the target objects is then rotated around the reference object. This transforms the target representation into an intrinsic reference frame defined by the reference object's movement direction. This rotatory transformation, new over the previous model, is realized by a steerable neural mapping that shifts activation patterns along the azimuth of the polar coordinate representation of the relational field (lower left blue diamond in Fig. 1). The extent of the shift is determined by the movement direction of the reference object, which is held by the *rotation field*.

The rotated target representation is projected into the *relational CoS field*. A second input to this field from spatial concept nodes encodes the associated spatial templates through weight patterns (illustrated in the lower left of Fig. 1). Overlap of the two inputs leads to a peak that represents the selected target. The steerable neural maps thus make it possible to apply the relational operator encoded in the fixed weight patterns to objects at any visual location in any orientation, implementing neural pointers.

The relational CoS field projects into the selective spatial attention field via reverse transformations for rotation and shift (upper and lower right diamonds in Fig. 1). Selective spatial attention projects into the three-dimensional attentional fields, forming peaks there that in turn project to the

feature fields, which may activate production nodes.

Concepts

Concepts like 'red' or 'toward' are represented by discrete nodes (denoted by circles in Fig. 1) that project with patterned synaptic weights into their respective feature fields. The nodes come in pairs: *memory nodes* (blue circles) act as an interface to a user who may activate them as input or observe them as output; *production nodes* (pink circles) gate the impact of their respective memory nodes onto the architecture. Note that there are copies of such pairs of nodes for each role that a concept may appear in (e.g., two pairs for 'red', as reference and as target), enabling role-filler binding. The synaptic weight patterns between nodes and fields could be learned by Hebbian learning rules but are hand-tuned here.

Process organization

The processes within the architecture are organized by instabilities of neural nodes that switch components 'on' or 'off'. These discrete events thus emerge from the time-continuous neural dynamics. Process organization is based on a structural principle borrowed from behavioral organization (Richter, Sandamirskaya, & Schöner, 2012). The core structure is the *elementary behavior*, which consists of two dynamic substrates. The *intention node* (green circle in Fig. 1) determines whether a process is active and has impact on connected structures. The *condition of satisfaction node* (CoS, red circle) is activated once a process has terminated and inhibits the intention node, turning the process off. Here, we employ elementary behaviors that control the grounding of the reference object (reference behavior), the target object (target behavior), and the spatial relation term (spatial relation behavior) (top left in Fig. 1). Role-filler binding is preserved during grounding by processing reference and target objects sequentially, organized by the *precondition* node (black circle) that inhibits the intention node of the target behavior until the reference behavior has terminated.

Results

In the following, we describe the dynamic processes that unfold within the architecture as it executes tasks. The results come from numerical solutions of the architecture's differential equations.[1] To simplify visual object recognition, we use a scene with uniformly colored objects on a white background.

Parsing an action

Fig. 2 illustrates the processes within the architecture as it generates a phrase about a video in which a red ball rolls toward a yellow ball (see top right of Fig. 1).

At $t = 0$ we give a boost into the architecture, which impacts the intention nodes of all behaviors. After this boost, the architecture runs autonomously in continuous time, without any further intervention from user or program. First, the reference object is grounded; the target behavior is inhibited by the precondition constraint until the reference behavior is finished. Without information about which objects to describe, the architecture decides based on their saliency. At t_1, the selective spatial attention field shows a saliency advantage for the moving red object in the lower left corner.

At t_2, the spatial attention field has made a selection decision and formed a peak. This creates a self-sustained peak in the reference field, selecting the moving object as reference. It also activates the production node 'reference: red' (top of Fig. 2) by projecting activation into the color CoS field via the attention color-space field (both not shown in Fig. 2; see Fig. 1). At the same time, the rotation angle field (not shown in Fig. 2) forms a representation of the object's movement direction, which it receives from the attentional movement-space field. It will later be used as a parameter to rotate the target objects. At this point, the architecture has grounded the reference object. That is, it has formed a connection between the continuous representations in the fields and the discrete representations in the nodes.

At t_3, the behavior to ground the reference object has been inhibited by its CoS node and the behavior to ground the target object has become active. However, even though the reference behavior is inactive, the peak in the reference field is still tracking the position of the moving object, because it receives input from the perception fields. Contrary to the reference behavior, the selective spatial attention field is not boosted during the target behavior, allowing multiple target candidates to be projected to downstream fields. The target field has formed three peaks at the positions of the remaining objects. The field's output is transformed and projected into the relational field, where the target positions are now represented relative to that of the reference object. This representation is rotated around the reference object and projected into the relational CoS field.

At t_4, the relational CoS field has formed a peak at the target position that overlaps most with the spatial template for the relation 'toward'. This activates the corresponding production node 'spatial: toward'.

At t_5, the activation from the relational CoS field is transformed and projected back into the selective spatial attention field, from there into the attentional color-space field, and from there into the target field as well as the color CoS field. The peak in the color CoS field activates the production node 'target: yellow'.

At this point, the architecture has produced the relational phrase 'red toward yellow' and has created a grounding of this phrase in sensorimotor representations.

Grounding a phrase

The architecture can also ground a phrase that it is given by user input. Due to space constraints, we cannot describe the process at the same level of detail. The process is very similar to that of grounding spatial relations reported earlier (Richter et al., 2014). The user supplies the phrase by activating mem-

[1] The architecture is implemented and simulated using the C++ framework *cedar* (Lomp, Zibner, Richter, Rano, & Schöner, 2013).

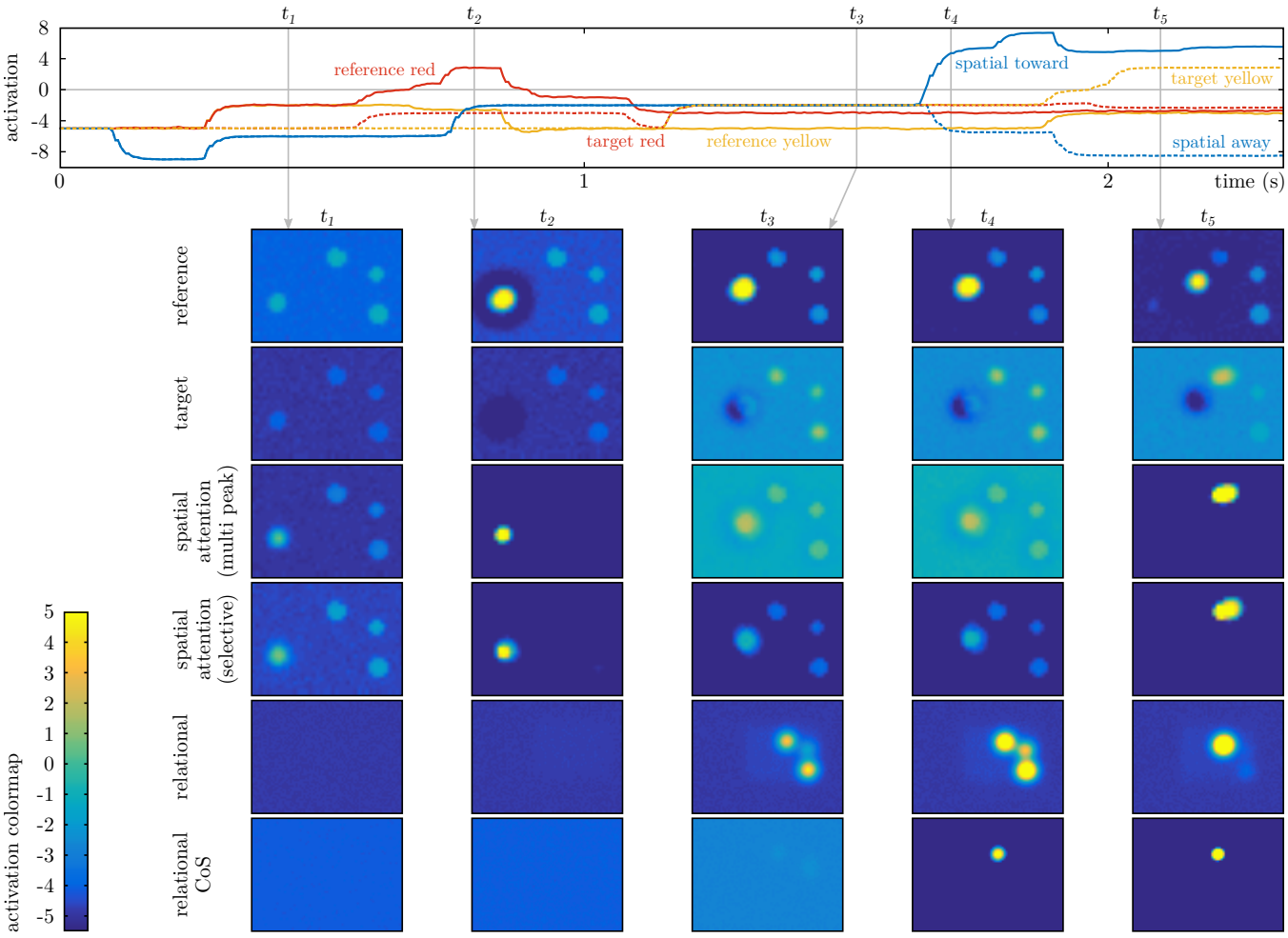


Fig. 2: Activation time courses of relevant production nodes (top) and activation snapshots of relevant fields at five points in time (bottom). Fields are color-coded using the color map on the bottom left.

ory nodes through manual boosts. Visual search for objects is then guided, as opposed to bottom-up saliency-driven. For instance, to ground the reference object, its red color is represented in the color intention field, bringing up peaks of red objects in the attentional color-space field—analogously with yellow objects for the target. Similarly, the template for spatial relations preshapes the relational CoS field and only allows peaks that overlap with the template. The grounding is established once a representation in the fields has been established for each element of the supplied phrase.

Discussion

We have extended a neural process model of spatial relations to include parsing of object-oriented actions and grounding of movement phrases. In the model, space-time continuous activation patterns are both coupled to sensory input and linked to neural representations of cognitive schemas like *move toward* or *move away from*. This provides a neural processing account of the interaction between sensorimotor activation, conceptual processing, and language, that theories of perceptual symbols (Barsalou, 2008) and embodied construction-grammar (Bergen & Chang, 2013) postulate. The integrative nature of the model leads us to confront fundamental issues such as the neural pointer problem, the binding problem, and how discrete processing steps emerge from time-continuous neural dynamics. Our solutions derive from the conceptual commitments of the theoretical framework of dynamic field theory.

We build on existing modeling approaches to the grounding of language that are neurally inspired but do not typically adhere to neural principles as consistently. For instance, the Neural Theory of Language (Feldman, 2006) is a hybrid framework that combines neural network concepts with ideas that are not compatible with neural process thinking. Similarly, Madden, Hoen, and Dominey's (2010) model for embodied language complements neural networks with algorithms that are not neurally based. Some models invoke neural concepts to account for psychophysical data. For instance, Regier and Carlson (2001) use the notion of an attentional vector sum to capture spatial terms. Such models are

not typically embedded into architectures that autonomously generate the complete sequence of processing steps required to ground and generate language.

The ambition of a neural process account for higher cognition is shared with the group of Eliasmith (2013). Their Neural Engineering Framework (NEF) enables spiking neural networks to realize vector symbolic architectures (Gayler, 2004). In this substantially different approach, concepts and objects are represented by high-dimensional vectors through an encoding and decoding stage and transient neural patterns computed by superposition and projection. DFT, in contrast, is based on self-stabilized activation patterns defined over a few feature dimensions. Whether DFT and NEF can span the same range of cognitive phenomena—which approach is more consistent with neural reality is open for now.

The current model is open to extension in various directions, such as incorporating learning, scaling the number of concepts, and building more complex sequences of processing steps. Higher-order schemata (e.g., source-path-goal; Lakoff & Johnson, 1999) may be realized similar to what we demonstrated here. Exploiting the working memory implicit in our representations may enable us to link to relational mental models (Knauff, 2013).

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behav Brain Sci*, *20*(4), 723–42; discussion 743–67.

Barsalou, L. W. (2008). Grounded cognition. *Annu Rev Psychol*, *59*, 617–45.

Bergen, B., & Chang, N. (2013). Embodied construction grammar. In T. Hoffmann & G. Trousdale (Eds.), *Oxford Handbook of Construction Grammar.* Oxford University Press.

Berger, M., Faubel, C., Norman, J., Hock, H. S., & Schöner, G. (2012). The counter-change model of motion perception: an account based on dynamic field theory. In *ICANN* (pp. 579–586).

Doumas, L. A. A., & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford Handbook of Thinking and Reasoning.* Oxford University Press.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press.

Feldman, J. A. (2006). *From molecule to metaphor: A neural theory of language*. Cambridge, MA: MIT Press.

Gayler, R. W. (2004). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *CoRR*, *abs/cs/0412059*.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. *Trends Cogn Sci*, *14*(11), 497–505.

Knauff, M. (2013). *Space to reason: A spatial theory of human thought*. Cambridge, MA: MIT Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. *J Exp Psychol Learn*, *38*(6), 1490–1511.

Logan, G. D., & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (pp. 493–529). Cambridge, USA: MIT Press.

Lomp, O., Zibner, S. K. U., Richter, M., Rano, I., & Schöner, G. (2013). A software framework for cognition, embodiment, dynamics, and autonomy in robotics: cedar. In V. Mladenov (Ed.), *ICANN* (pp. 475–482). Heidelberg: Springer.

Madden, C., Hoen, M., & Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain Lang*, *112*(3), 180–8.

Regier, T., & Carlson, L. A. (2001). Grounding spatial language in perception: an empirical and computational investigation. *J Exp Psychol Gen*, *130*(2), 273–98.

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello & et al. (Eds.), *36th CogSci* (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *IEEE/RSJ IROS* (pp. 2457–2464).

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: how instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–79.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biol Cybern*, *106*(2), 89–109.

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating what and where: Visual working memory for objects in a scene. In G. Schöner, J. P. Spencer, & the DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory.* Oxford University Press.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychol*, *12*(1), 97–136.

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: intrinsic frames, the term "between", and autonomy. In *IEEE RO-MAN* (pp. 150–157).

Vector Space Semantic Models Predict Subjective Probability Judgments for Real-World Events

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA.

Abstract

We examine how people judge the probabilities of real-world events, such as natural disasters in different countries. We find that the associations between the words and phrases that constitute these events, as assessed by vector space semantic models, strongly correlate with the probabilities assigned to these events by participants. Thus, for example, the semantic proximity of "earthquake" and "Japan" accurately predicts judgments regarding the probability of an earthquake in Japan. Our results suggest that the mechanisms and representations at play in language are also active in high-level domains, such as judgment and decision making, and that existing insights regarding these representations can be used to make precise, quantitative, a priori predictions regarding the probability estimates of individuals.

Keywords: Judgement and decision making; Subjective probability; Semantic representation; Semantic space models

Introduction

Subjective probability judgment plays an important role in everyday cognition and behavior. Our interactions with the world around us are guided by the probability estimates we place on its largely uncertain events. These estimates, in turn, stem from our knowledge of the world, and the cognitive mechanisms that we possess for learning, representing, and applying this knowledge.

The study of subjective probability judgment has yielded a number of valuable insights regarding how individuals assign probabilities to uncertain events (Kahneman & Tversky, 1973; Lichtenstein, Fischhoff & Phillips, 1977; Tversky & Koehler, 1994; Wallsten & Budescu, 1983). One of the most important of these insights involves the use of simple heuristics, such as those relying on associations between the various objects or concepts involved in the judgment task (Kahneman & Frederick, 2002; Sloman, 1996). The use of association-based heuristics is relatively effortless, though it can lead to biases in specific settings. This is one reason why, for example, individuals commit the conjunction fallacy in the Linda problem (Tversky & Kahneman, 1983), which asks them to judge whether Linda, a female activist concerned with issues of social justice, is more likely to be a bank teller or a feminist bank teller. Here the description of Linda is strongly associated with feminism, making participants believe that the probability of Linda being a feminist bank teller is higher than her being a bank teller, despite the fact that all feminist bank tellers are in fact also bank tellers.

The events considered in most research on associative judgment are abstracted or artificial. These types of tasks are valuable, as they allow for rigorous tests of scientific hypotheses. However, in order for association-based heuristics to be considered good models of subjective probability judgment, they should be able to predict the specific probabilities individuals assign to the occurrence of real-world events, that is, the types of events that individuals encounter and evaluate on a day-to-day basis. Thus we should not only be able to state that individuals place a higher probability on Linda being a feminist bank teller compared to a bank teller, but also predict the explicit probabilities individuals attach to, for example, various outcomes in current affairs or popular culture. These types of events are often of the form "X happens to Y" (e.g. an earthquake occurs in Japan), and associative heuristics predict that the association between X and Y (e.g. "earthquake" and "Japan") is used by individuals to judge the probabilities of these types of events.

Predicting real-world judgments is not trivial: Although associative heuristics are easy to apply, the information that these heuristics utilize is fairly complex. Thus, even though decision makers may use the strength of association between "earthquake" and "Japan" to predict the probability of there being an earthquake in Japan, it is not immediately clear what determines these associations, and in turn what the decision maker's actual probability judgment about an earthquake in Japan will be.

Associative processing is also of interest in the study of language, and there have been many recent advances in understanding the determinants of association, or more generally, semantic relatedness, as it applies to people's comprehension and use of words (Bullinaria & Levy, 2007; Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landaur & Dumais, 1997; Lund & Burgess, 1996). The key insight underlying these advances is that the representation of words depends on the statistical structure of the environment in which these words occur (see also Firth, 1957 and Harris, 1954). Studying the distribution of words in the types of settings people encounter on a day-to-day basis can uncover the representations that people have of everyday words, and in turn the semantic relationships and associations between these words, and the objects and concepts they represent.

Models that build semantic representations using the distribution of words often characterize each word in their vocabulary as a vector in a highly multidimensional space. The proximity between the vectors of two words corresponds to the relatedness or association of the words, so that synonyms and other closely related words are frequently located near each other. Vector space semantic

models are typically trained on very large natural language text corpora, and subsequently have large vocabularies, which can be used to make predictions regarding judgments of word similarity, the strength of word priming, and related psycholinguistic phenomena, for nearly all the words commonly used in a given language. The predictions of these models have been shown to be highly accurate, suggesting that the representations recovered by these models provide a good characterization of the representations underlying semantic processing in language use. For this reason, these models are also popular in machine learning and artificial intelligence, particularly in applications related to computer processing of natural language (see Turney & Pantel, 2010).

Hare, Jones, Thomson, Kelly, and McRae (2009) have shown that the word associations captured by vector space approaches are able to account for priming effects regarding event representation. Relatedly, Paperno, Marelli, Tentori, and Baroni (2014) have found that word association correlates very strongly with explicit probability judgments of word co-occurrence. These results suggest that the representations and associations that guide word use and comprehension may also be the ones involved in making probability judgments for real-world events, and that vector space semantic models could in turn be used to predict these probability judgments. Thus, for example, we could obtain an estimate of the actual probability individuals assign to there being an earthquake in Japan by examining the (linguistic) association between "earthquake" and "Japan" generated by vector space semantic models.

In this paper we test this idea by studying subjective probability judgments about different countries and different famous people. Our tests utilize vector representations released by Google Research, which are based on the recurrent neural network methods proposed by Mikolov and coauthors (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Across eight studies we ask participants to assign probabilities to various natural events happening in these countries (e.g. earthquake in Japan) and to these people (e.g. Jon Stewart becoming president), and we predict judged probabilities using the distance between the vectors for the various words and phrases that make up the events.

Methods

Participants

Our tests involve eight distinct studies with 200 participants each in Studies 1-4 and 100 participants each in Studies 5-8, leading to a total of 1,200 participants (overall mean age = 34.81, 51% male). These participants were recruited from Amazon Mechanical Turk, and performed the studies online, for which they were each compensated $0.50. Our studies included an attention check question and we excluded the 31 participants who failed this attention check across the studies. The number of participants in the above studies was determined prior to running the studies, and the specific numbers were chosen as they were round numbers that allowed us to obtain a sufficient number of estimates for each event.

Materials and Procedure

The first four of our studies asked participants to judge the probability that various man-made and natural disasters would happen in the different countries of the world. For each of the countries offered to the participants, they were asked to assess the probability that the country would experience a terrorist attack in the next week (Study 1), be in a state of war at the start of 2016 (Study 2), experience an earthquake over the next year (Study 3), or experience an epidemic over the next year (Study 4). Each participant in each study was given a list of 30 countries chosen at random from the 193 countries that were members of the United Nations when the studies were run.

The remaining four studies asked participants to make judgments about various famous people in the United States. For each person offered to the participant, he or she was asked to assess the probability that the person would be the U.S. president in 2020 (Study 5), win a Nobel Prize in 2020 (Study 6), win a Grammy Award in 2020 (Study 7), or win an Academy Award in 2020 (Study 8). The list of famous people used in this study was obtained from a separate pool of MTurk participants (mean age = 36.81, 56% male) who were each asked to write the names of ten highly recognizable people in the United States that were still alive. The 50 most frequent names generated by these participants were used in Studies 5-8. Again, each participant in Studies 5-8 was asked to make judgments about 30 people chosen randomly from our list of 50 people.

Probability judgments for each of the countries in Studies 1-4 and each of the people in Studies 5-8 were made on a slider scale between 0-100%. The 30 events offered to each participant were presented one after the other, on separate screens, in a random order. With the 30 judgments for each of the 200 participants, we obtained approximately 30 probability estimates for each of the events considered in Studies 1-4. Likewise, with the 30 judgments for each of the 100 participants, we obtained approximately 60 probability estimates for each of the events considered in Studies 5-8.

Overview of Analysis

In this paper we utilize a set of pretrained vector representations recently released by Google Research (code.google.com/p/word2vec). These vectors have been trained on a 100 billion word subset of the Google News corpus, by continuous bag-of-words (CBOW) and skip-gram techniques of Mikolov et al. (2013a, 2013b). This approach relies on a recurrent neural network that, for the CBOW technique, attempts to predict words using other words in their immediate context, and for the skip-gram technique, attempts to the do the inverse of this. These representations have a vocabulary of 3 million words and phrases, including countries and names with two or more component words, such as "United States" and "Jon Stewart". The recent successes of Mikolov et al.'s methods,

the large amount of training data and resulting vocabulary used in the word representations, and the fact that these representations have been obtained from a news corpus, make them particularly valuable for the tests we are conducting.

Each of the 3 million vectors we use is described on 300 dimensions, and the linguistic association between any two words or phrases can be computed using the distance between their corresponding vectors in this 300 dimensional space. In this paper we use the distance between "terrorism", "war", "earthquake", and "epidemic" and the words corresponding to the 193 countries to predict the probabilities that participants assign to the disasters happening in the countries in Studies 1-4. Likewise we use the distance between "president", "Nobel Prize", "Grammy Award", and "Academy Award" and the names of the 50 famous people to predict the probabilities that participants assign to the people winning these awards in Studies 5-8. Thus for example, we can calculate the association of "earthquake" and "Japan" or of "President" and "Jon Stewart" using the distance between each of these pairs of vectors, and in turn use this distance to predict the probability people assign to there being an earthquake in Japan, or to Jon Stewart becoming president. The metric of distance we consider is cosine similarity, so that the distance between any two vectors a and b is given by $dist(a,b) = a \cdot b/(||a|| \cdot ||b||)$. This metric varies between -1 and +1 (with -1 capturing orthogonal vectors and +1 capturing vectors with identical directions). Additional details about Mikolov et al.'s Word2Vec training techniques can be found in Mikolov et al (2013a, 2013b). Note that in analyzing Studies 5 and 6, we remove famous people who have previously won Nobel Prizes or have previously served two terms as the president of the United States (these awards or positions cannot be won again in the future). We also exclude participant judgments regarding St. Vincent and Grenadines, as this country is not represented in the set of word vectors released by Google Research.

Results

Overview of Data

Recall again that there are about 30 participant judgments of event probability for each of the 193 events in Studies 1-4, and about 60 participant judgment of event probability for each of the 50 events in Studies 5-8. In this paper our main dependent variable will be the probability estimate for an event obtained by averaging all the probability estimates made by participants for that event. We find that these average probability estimates vary substantially with the event that participants are required to judge, with, for example, the average probability assigned to there being an earthquake in Japan over the next year being 55.03% (N = 33, SD = 25.24) and the average probability assigned to there being an Earthquake in Norway being only 11.88% (N = 18, SD = 16.78).

Average event probabilities are highly dispersed for judgments regarding Grammy and Academy awards in Studies 7 and 8. In these studies, the average probabilities assigned to different people winning these awards appear to be roughly uniformly distributed between 0% and 70%. In contrast we observe the lowest dispersion in average estimates of epidemic and earthquake probabilities in Studies 3 and 4, in which average event probabilities are clustered between 20% and 40%.

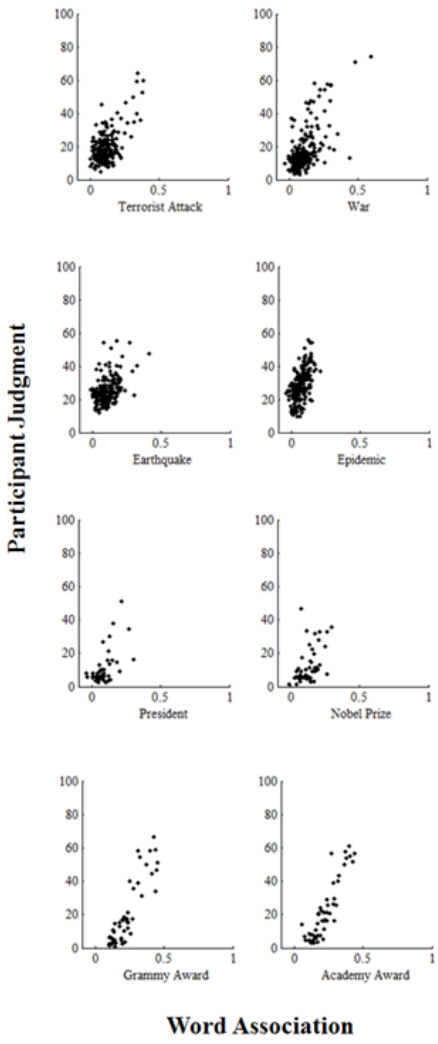
The main independent variable in this paper will be the linguistic association between the word and phrases in an event. Again, this measure is the cosine similarity between the disasters and countries (in Studies 1-4) or the awards and people (in Studies 5-8). Although cosine similarity can vary between -1 and +1, associations between the words in 941 out of the 968 events in our studies were positive. The distributions of these associations were, as with probability estimates, most dispersed for events involving Grammy and Academy awards in Studies 7 and 8, and least dispersed for events involving epidemics and earthquakes in Studies 3 and 4. The distribution of average event probabilities and word associations for the different events in our studies can be observed in Figures 1A-1H.

The Predictive Power of Word Associations

Can associations predict event probabilities? Recall that for each of the 968 events across our studies we have both the average probability assigned to the event by our participants, and the association (or, more formally, cosine similarity) between the words in the event, specified by our set of vector representations. A first step in our analysis is examining the correlation between word associations and the average estimates of participants. A standard test using Pearson's correlation reveals positive, significant correlations between these two variables in each of our studies ($p < 0.001$). Overall our approach does best in Studies 7 and 8, which involve judgment of popular culture, particularly the probability of various people winning Grammy Awards and Academy Awards. Here word associations and average participant probability estimates have correlations of 0.89 and 0.90 respectively. Our approach is also highly successful at predicting judgments of terrorism and war in different countries, in Studies 1 and 2, with correlations of 0.62 and 0.64 respectively. Predictions regarding judgments of presidential victories and Nobel Prizes for different people in Studies 5 and 6 achieve correlations of 0.59 and 0.48 respectively. The model does worst in predicting judgments of natural disasters in different countries, with correlations for earthquakes in Study 3 and epidemics in Study 4, being 0.44 and 0.57. Scatter plots displaying the relationship between word associations and participant judgments can be seen in Figures 1A-1H , and the correlations outlined here are summarized in Table 1.

It is interesting to note the above analysis does not involve any model fitting, and the word associations we use are already built in to the vector representations released by Google Research. Ultimately, the results discussed here (such as correlations of 0.89 and 0.90 between word

associations and participant probability judgments for popular culture events) emerge from what is essentially a zero parameter model.



Figures 1A-H: Scatter plot of the word associations (in terms of cosine similarity) generated by the model and average participant probability estimates (in terms of percentage) for the events in Studies 1-8 respectively.

Some model fitting could, however, help us better understand the properties of the approach we are proposing. For this purpose we first consider a simple linear model, which transforms the cosine similarity measure of word association, which ranges from -1 to 1, into a probability judgment scale, which ranges from 0 to 100. Note that such a linear fit would not change correlations or their significance levels, but would allow for a better understanding of the degree of variance in the data explained by our approach. After fitting linear models for the eight studies, using a basic linear regression, we unsurprisingly find highly significant relationships between word associations and participant judgments ($p < 0.001$ for each of the studies). Overall the R^2 statistics for these fits vary between 0.19 (for earthquake judgments in Study 3) and 0.80 (for Academy Award judgments in Study 8). More details about these fits are provided in Table 1.

The differences in the correlations and fits across the eight studies could be attributed to varying uses of association-based heuristics in different domains. Perhaps participants are just more likely to apply associative processing when making judgments of pop culture, as in Studies 7 and 8, compared to judgments regarding natural disasters in different foreign countries or winners of Nobel Prizes, as in Studies 3, 4, and 6. Alternately, it is possible that we have the highest correlations in Studies 7 and 8 because our participants have more knowledge about popular culture than they do about natural disasters or Nobel Prizes. Due to their increased knowledge they are more likely to make fine grained probability assessments in Studies 7 and 8, allowing for a cleaner dataset on which to predict probability judgments. Indeed, as discussed above, average probability estimates for the events in Studies 7 and 8 have a much higher spread than average estimates in Studies 3 and 4.

The above analysis has only attempted to predict the average probability estimate placed on the events by our participants. This type of aggregation is desirable for many reasons (see e.g. Wallsten, Budescu, Erev & Diederich, 1997 for a review). However it ignores the variance across participants in their judgments of the probabilities. Do word associations provide a good account of probability judgment when we allow for participant-level heterogeneity? We can test this on participant-level data using a linear regression model with participant-level random intercepts. As many individual participant estimates, unlike aggregate estimates, lie at the boundaries of the probability scale (i.e. at 0% and 100%) the regression we use permits a censored dependent variable using the Tobit method (Tobin, 1958). With these controls we find that cosine similarity has a strong positive significant relationship with probability estimates in each study ($p < 0.001$ for each study), showing that associations can predict not just aggregate probability judgments but also probability judgments on an individual level.

Study	Event	Target	ρ_{linear}	ρ_{log}	$\rho_{logistic}$	R^2_{linear}	R^2_{log}	$R^2_{logistic}$
1	Terrorist attack next week	Country	0.62	0.40	0.65	0.38	0.16	0.42
2	Being at war at start of 2016	Country	0.63	0.47	0.63	0.40	0.22	0.40
3	Earthquake next year	Country	0.44	0.37	0.44	0.19	0.14	0.20
4	Epidemic next year	Country	0.57	0.48	0.57	0.32	0.24	0.32
5	US president in 2020	Person	0.59	0.48	0.59	0.35	0.23	0.34
6	Nobel Prize in 2020	Person	0.47	0.41	0.49	0.23	0.17	0.24
7	Grammy Award in 2020	Person	0.89	0.85	0.88	0.79	0.72	0.78
8	Academy Award in 2020	Person	0.90	0.81	0.91	0.81	0.65	0.84

Table 1: Summary of correlations (ρ) and R^2 values from linear, logarithmic, and logistic model fits for the events in Studies 1-8. Note that all model fits involve two free parameters, and that the correlations correspond to those displayed in Figures 1A-H. All of these correlations are highly significant ($p < 0.001$).

Testing Non-linear Relationships

Psychophysical judgments are often characterized by non-linear relationships, as with the Weber-Fechner law, and it is possible that the associations decision makers perceive between the words in the event at hand are transformed non-linearly before being mapped onto probability judgments. We can test this by comparing the fits of the above linear model with a group of similarly parameterized non-linear models. The first model we consider takes a natural-log transformation of the associations for each event. These transformed values are then fit by minimizing mean-squared error. As with the untransformed linear regression, this is a two parameter model. Thus if we write the average probability estimates as y and the word associations for an event as x, our logarithmic model would attempt to find parameters β_0 and β_1 to fit $y = \beta_0 + \beta_1 \ln(x)$. Note that the log transform cannot be used on negative numbers. A very small minority the cosine similarity values for the events are in fact negative. These have been ignored in our analysis (none of the results change if we use more complex log-based functions for transforming these negative numbers).

The second model we consider is a logistic curve. Such sigmoidal (s-shaped) curves are frequently used to obtain choice probability estimates in discrete choice experiments, as their outputs are bounded by 0 and 1, and the logistic curve is perhaps the most commonly used of all of these (E.g. with Luce's choice rule). We fit the two-parameter logistic curve by minimizing mean-squared error, with the cosine similarity values as our independent variable and the average participant probability estimates as our dependent variable. Here if we write the average probability estimates as y and the word associations for an event as x, our logistic model would attempt to find parameters β_0 and β_1 to fit $y = 1/(1+\exp\{-\beta_0 - \beta_1 x\})$.

As both our logarithmic and our logistic models involve two parameters, their predictive accuracy can be directly compared with those of the linear model described above. Ultimately we find that the logistic and the linear models perform about equivalently, providing nearly identical correlations and R^2 values. The logarithmic model, in contrast provides a much more inferior fit than the linear and logistic models, with correlations and R^2 values as low as 0.41 and 0.17 for the Nobel Prize judgments in Study 6. Despite this fact, the predictions of all models have statistically significant relationships with the judgments of participants ($p < 0.001$ for all models in all studies).

A brief examination of the parameter values generated by our logistic fits explains why they perform as well as the linear models. These parameters are typically such that the inputs to the logistic function fall within its middle, linear range, implying that the logistic function behaves roughly like a linear model. Ultimately, it seems that our measure of word association maps linearly onto the probability judgments of our participants. A summary of the fits for the logarithmic and logistic models is provided in Table 1.

Discussion and Conclusion

In this paper we find that vector semantic space approaches can predict the probability judgments that people make about various events in the world. More specifically, the associations between words and phrases, as assessed by a set of vector representations released by Google Research and trained using the recurrent neural network methods proposed by Mikolov et al. (2013a, 2013b), correlates very heavily with the probabilities that people assign to natural events described using those words and phrases. Furthermore, model fitting indicates that this relationship is linear, rather than logarithmic or sigmoidal.

There are some limitations to the approach we have proposed. For example, vector space semantic models cannot by themselves modify their output to control for the length of time the events are supposed to occur. Thus, the approach described above would give the same probability estimate for an earthquake in Japan in the next one year, as it would for a similar earthquake in the next five years. There is some evidence that human probability judgment doesn't sufficiently account for magnitudes, such as durations of the events (see Fredrickson & Kahneman, 1993), but there is no doubt that our predictions could be improved by allowing for a secondary system that adjusts the estimates obtained through semantic relatedness based on event duration, as well as other non-sematic features of the event at hand.

Performance could also be improved by fitting the vector space models to the data. Recall that the above analysis only performs a transform of cosine similarity to predict probability judgment. It alters neither the number of dimensions used in the model, nor the size of the context window to train the models, nor the weights placed on these dimensions to judge semantic distance (both of which are specified a priori). A more sophisticated approach that trains the vector space models on the probability estimates of participants, would no doubt provide better predictions regarding their subjective probability estimates.

There are also boundary conditions. For example, it is unlikely that the approach outlined in this paper would be able to successfully describe probability judgments involving symbolic reasoning or more complex delilberative processing. However, despite these limitations, our results have some important implications. Firstly they provide new techniques for predicting real-world judgments. These predictions are quantitative, in that they attempt to capture the exact numerical probability assigned to an event. These predictions are also domain-general, in the sense that they can be applied to a number of different types of real-world events. Ultimately, the vector space models that we use have very large vocabularies, and are able to provide a precise measure of association between any two words or phrases in their vocabularies, and subsequently precise probability judgments for simple events composed of these words and phrases. Existing judgment models do not have this important property.

In addition to numerous practical applications to areas such as risk perception and communication (Slovic, 2000), these quantitative predictions can be used to more rigorously study the processes already known to characterize probability judgment (Kahneman & Tversky, 1974; Lichtenstein et al. 1977; Tversky & Koehler, 1994; Wallsten & Budescu, 1983), and also potentially uncover novel effects and regularities. They can also be used to predict everyday decisions involving these events, such as, for example, the purchasing of insurance. These types of real-world decisions are of considerable scholarly interest.

Our results also highlight the power of association-based heuristics in making probability judgment (Kahneman & Frederick, 2002; Sloman, 1996). Though the descriptive power of these heuristics is accepted, this paper is the first to show that a semantic instantiation of these heuristics can be used to predict the actual probabilities that decision makers assign to natural everyday events. Likewise, these results illustrate the power of vector semantic space models. These approaches not only predict judgments of word meaning and use in linguistic domains, but also judgments involving complex events in the real-world. It is important to note that many of the results could not be obtained by simpler approaches that use, for example, only the co-occurrence of words to judge associations. "Jay Z" and "Joe Biden" may never directly co-occur with "Academy Award", but participants nonetheless ascribe a higher probability to Jay Z, a musician, winning an Academy Award, relative to the Joe Biden, a politician.

Finally, the link between the representations and associations used to assess word meaning and the representations and associations used to make probability judgments, observed in this paper, suggests the existence of a single system of learning, storing, and retrieving knowledge for both language use and for high-level judgment. Subjective probability judgment does not rely on knowledge that is fundamentally different from the type of knowledge used to understand and generate language. This in turn implies a close connection between two important and influential areas in psychology. Future work should attempt to build more general models of semantic cognition; models which are not only able to explain how people acquire the knowledge of word meanings, but also how people use these word meanings to form their beliefs about the objects and events they encounter in the world.

References

Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: A computational study." Behavior Research Methods 39.3 (2007): 510-526.

Firth, J. R. (1957). A synopsis of linguistic theory 1930 – 1955. In Studies in linguistic analysis (pp. 1–32). Oxford, England: Blackwell Publishers.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology,65(1), 45-68.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211-235.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. Cognition, 111(2), 151-167.

Harris, Z. (1954). Distributional structure. Word, 10, 146 – 162.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114(1), 1.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited. In Heuristics and Biases: The Psychology of Intuitive Judgment.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2).

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art (pp. 275-324). Springer Netherlands.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, 28(2), 203-208.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems.

Paperno, D., Marelli, M., Tentori, K., & Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. Cognitive Psychology, 74, 66-83.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1), 141-188.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychological Review, 90(4), 293.

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. Psychological Review, 101(4), 547.

Wallsten, T. S., & Budescu, D. V. (1983). State of the art—Encoding subjective probabilities: A psychological and psychometric review. Management Science, 151-173.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 10(3), 243-268.

Noisy Parameters in Risky Choice

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania, Philadelphia, PA.

Graham Loomes (graham.loomes@wbs.ac.uk)
Warwick Business School, University of Warwick, Coventry, UK.

Abstract

We examine the effect of variability in model parameters on the predictions of expected utility theory and cumulative prospect theory, two of the most influential choice models in decision making research. We find that zero-mean and symmetrically distributed noise in the underlying parameters of these models can systematically distort choice probabilities, leading to false conclusions. Likewise, differences in choice proportions across decision makers might be due to differences in the amount of noise affecting underlying parameters rather than to differences in actual parameter values. Our results suggest that care and caution are needed when trying to infer the underlying preferences of decision makers, or the effects of psychological, biological, economic, and demographic variables on these preferences.

Keywords: Decision making; Random utility; Random preference; Risky choice; Prospect theory

Introduction

Research on risky choice has relied heavily on the use of deterministic models. Perhaps the two most widely used models today are expected utility theory (EUT) (von Neumann & Morgenstern, 1947) and cumulative prospect theory (CPT) (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). When their functional forms are specified and parameterized, deterministic models can make precise quantitative predictions. However, such models fail to capture an important aspect of choice behavior: namely, that choice is stochastic, and decision makers may respond differently when given exactly the same choice problem on more than one occasion within a short space of time (see Rieskamp et al., 2006 for a discussion).

Modelling stochastic risky choice requires a representation of each individual's preferences as probability distributions. From an early stage in the development of the literature, there were two ways in which this type of modelling was accomplished. One approach involved adding some 'error' specification to each individual's deterministic 'core' preferences (e.g. Luce, 1959). We shall refer to this approach as Fechnerian – a broad term which covers a number of ways in which some form of 'add-on' term might be specified. Another approach allowed the parameters of an individual's preference function to vary from one moment to another, thereby opening up the possibility that if the same choice were presented at different times in the course of an experiment, it might be resolved differently on each occasion (Becker et al. 1963). We shall refer to this way of modelling noise as the random preference or random parameter (RP) approach.

To the extent that preferences can be seen as being constructed, influenced either by contextual and environmental cues, imperfect recall, or momentary fluctuation of attention, it might seem that the RP approach is conceptually more appropriate, as it permits variability in preference through variability in the parameters of the individual's preference functions. However, people may also vary in the way they record their decisions, due to varying degrees of complexity of the stimuli, varying degrees of motivation or engagement, 'interference' from previous decisions within the session or from some other unobservable features of the stimuli or task. Since these factors are largely outside the core structure of people's preferences, they may be better modelled by some Fechnerian noise term that operates in addition to, but independently of, the intrinsic variability in preference (see Birnbaum, 2011, for a discussion).

So it seems plausible that the observed variability in individuals' repeated choices may arise from a number of sources which are not mutually exclusive. Despite this, the great majority of studies inferring preference functions from choice data use specifications which operate as if some form of Fechnerian noise is the only stochastic component of the decision process: very few consider the possibility of variability in terms of interactions between an RP core and Fechnerian factors. Yet if observed choices actually entail multiple sources of variability, there may be serious consequences for theoretical inference if we try to force data into a conceptually inadequate specification.

Risky Decision Models

We can write the gambles in the choices studied in this paper as $X = (x_1, p_1; x_2, p_2)$, so that X can be seen as offering rewards x_1 and x_2 with probabilities p_1 and p_2. Both EUT and CPT describe decision makers' preferences between pairs of such gambles using utility functions. Using a simple power value function formulation for the utility of any payoff x, the utility of a particular gamble under EUT is given as: $U(X|\alpha) = p_1 \cdot x_1^{\alpha} + p_2 \cdot x_2^{\alpha}$. The utility for X according to the power value function formulation of CPT, with Prelec's one-parameter probability weighting function, is similarly: $U(X|\alpha,\gamma) = \pi(p_1) \cdot x_1^{\alpha} + (1 - \pi(p_1)) \cdot x_2^{\alpha}$ where $x_1 \geq x_2 \geq 0$ and $\pi(p_1) = e^{-(-\ln p_1)^{\gamma}}$.

α captures the shape of the subjective value function for payoffs, with $\alpha < 1$ describing concave value functions that

correspond to risk averse preferences under EUT, and $\alpha > 1$ describing convex value functions that correspond to risk seeking. γ captures the shape of the probability weighting function, with $\gamma < 1$ generating an overweighting (underweighting) of small (large) probabilities, and $\gamma > 1$ generating the opposite. Typically, decision makers behave as if they have both $\alpha < 1$ and $\gamma < 1$ when the above gambles are explicitly presented to them.

In binary choice, these models predict that X is chosen over Y whenever $U(X|\alpha,\gamma) > U(Y|\alpha,\gamma)$. As these models are deterministic, they need to be modified in order to capture probabilistic choice data. The most common approach to doing this has been to assume that the utilities for the two gambles are each subject to Fechnerian noise ε, with $E[\varepsilon] = 0$, so that X is chosen over Y whenever $[U(X|\alpha,\gamma) + \varepsilon_X] - [U(Y|\alpha,\gamma) + \varepsilon_Y] > 0$. Supposing further that ε_X and ε_Y are independent of each other, we can define $\varepsilon = \varepsilon_X - \varepsilon_Y$. Then the probability of choosing X is the probability that $U(X|\alpha,\gamma) - U(Y|\alpha,\gamma) + \varepsilon > 0$. When ε is distributed according to the Gumbel distribution, this leads to the logit choice rule specified by Luce (1959): $\Pr[X \text{ chosen}] = f(\theta \cdot U(X|\alpha,\gamma) - \theta \cdot U(Y|\alpha,\gamma))$, where f is the logistic function. This error specification, when applied by itself, predicts that the modal choice will always be the option with the higher utility according to the deterministic core.

Now let us consider separately the effect of noise in the decision maker's parameters, as represented in the RP approach. With a simple additive formulation for parameter noise, the above equations can be rewritten with $\alpha = \alpha^* + \eta_\alpha$ and $\gamma = \gamma^* + \eta_\gamma$, where η_α and η_γ are noise terms with $E[\eta_\alpha] = E[\eta_\gamma] = 0$ (Becker et al., 1963; Loomes & Sugden, 1995). η_α and η_γ are liable to vary from trial to trial, and thus α and γ, and subsequently $U(X|\alpha,\gamma)$ and $U(Y|\alpha,\gamma)$, also tend to vary from trial to trial. However, since the expected values of α and γ are $E[\alpha] = \alpha^*$ and $E[\gamma] = \gamma^*$, α^* and γ^* characterize the central tendency of a decision maker's underlying preferences. If variability came only from sources represented by noisy parameters, the probability of X being chosen would be determined by the proportion of cases where $U(X|\alpha,\gamma) - U(Y|\alpha,\gamma) > 0$ over the range of feasible values of η_α and η_γ, weighted by their probabilities.

However, we now suppose that RP variability is combined with other sources of variability captured by a Fechnerian specification of the kind proposed by Luce (1959), we have, for each α and γ pair: $\Pr[X \text{ chosen}] = \Pr[U(X|\alpha,\gamma) - U(Y|\alpha,\gamma) + \varepsilon > 0]$. Now on those occasions where the α and γ drawn from the distributions of parameters are such that $U(X|\alpha,\gamma) - U(Y|\alpha,\gamma)$ is small, there is a relatively high chance (though still less than 0.5, of course) that Fechnerian variability will result in choosing the option with the lower $U(.)$, whereas there is a smaller chance of that happening when RP variability produces a larger $U(X|\alpha,\gamma) - U(Y|\alpha,\gamma)$ difference.

A combination of Fechnerian and RP specifications offer a more adequate account of choice data involving dominated gambles than each of these approaches alone. Models with only standard Fechnerian noise predict much higher frequencies of violations of transparent dominance than are generally observed, while RP-only models predict that dominance is never violated at all, contrary to the evidence (Loomes, 2005; see also Busemeyer & Townsend, 1993). Scholars have also found that allowing for both types of noise is necessary to provide a good quantitative account of behavior (Blavatskyy & Pogrebna, 2010; Loomes, 2005).

Effects of Noisy Parameters

Fechnerian noise and random parameters are necessary to characterize probabilistic choice. Yet despite this, much empirical decision research neglects the effect of RP when deriving predictions from deterministic models. Such neglect is no doubt due to convenience. Luce's choice rule has an analytical representation, which greatly facilitates model fits and related analyses.

This neglect may also reflect the intuition that unsystematic variability in parameter values, independent of Fechnerian noise, has no systematic effect on choice, so that modal choices can be used to make qualitative inferences about how underlying preferences rank the available options, regardless of the randomness in underlying parameters. This paper tests this intuition and finds that it is incorrect. Unlike the Fechnerian noise term ε, the parameter noise terms η_α and η_γ affect utility non-linearly. Even if they have a zero mean and are symmetric, changing their variance can alter both absolute choice probabilities and the ordering of relative choice probabilities. Thus it is possible that we observe X being chosen more frequently than Y, with $\Pr[X \text{ chosen}] = \Pr[U(X|\alpha,\gamma) - U(Y|\alpha,\gamma) + \varepsilon > 0] > 0.5$, but the central tendency of the decision maker's underlying preferences more frequently favours Y over X, with $U(Y|\alpha^*,\gamma^*) > U(X|\alpha^*,\gamma^*)$ and $\Pr[U(Y|\alpha,\gamma) > U(X|\alpha,\gamma)] > 0.5$.

Risk Attitudes

As an illustration of this, let us now consider the choice between a risky gamble X offering a 50% chance of obtaining \$10 and a 50% chance of obtaining \$0, and its safe expected value equivalent Y offering \$5 with certainty. We assume that a decision maker's central tendency is described by the power form of EUT and that his choices display both Fechnerian and RP noise. Keeping things simple, we suppose that the Fechnerian noise is modelled via a Luce choice function (with $\theta = 1$) while the RP component involves $\alpha^* = 0.9$ with η_α being distributed uniformly in the interval [-0.5, 0.5]. Supposing $\alpha^* < 1$ implies that the decision maker's underlying preferences more often than not entail risk aversion. However, when each of the possible realizations of $[U(X|\alpha,\gamma), U(Y|\alpha,\gamma)]$ are embedded in the Luce formulation, the above assumptions give $\Pr[X \text{ chosen}] = 0.53 > 0.5$. Thus, despite underlying preferences predominantly supporting Y, the decision maker chooses X more frequently than Y.

This happens due to the nonlinearity of utility difference in α. In the Luce choice rule assumed above, the probability of choosing X is an increasing function of $f(\alpha) = U(X|\alpha,1) - U(Y|\alpha,1) = 0.5 \cdot 10^\alpha - 5^\alpha$. f is convex in α for the range of α we

are considering. This implies that the expected value of f is greater than f applied to the expected value of α, i.e. $E[U(X|\alpha,1) - U(Y|\alpha,1)] > U(X|\alpha^*,1) - U(Y|\ \alpha^*,1)$. In this case we obtain $E[U(X|\alpha,1) - U(Y|\alpha,1)] > 0$, resulting in a higher choice probability of X. This is despite $E[\alpha] = \alpha^* < 1$, and that subsequently $U(X|\alpha^*,1) - U(Y|\alpha^*,1) < 0$.

The point is expanded upon in Figure 1, where we plot the probability of choosing X = ($10, 0.5; $0, 0.5) over Y = ($5, 1) according to power function EUT with only Fechnerian noise (using a Luce rule with θ = 1) and compare that with a specification where RP (with η_α distributed uniformly in the interval [-0.5,0.5]) interacts with Fechnerian noise. The Fechnerian-only model entails Pr[X chosen] less than, equal to or greater than 0.5 according to whether α^* is less than, equal to or greater than 1, as shown by the solid line in Figure 1. However, this is not the case when α is noisy. As shown by the broken line in Figure 1, there are a range of values of α^* between 0.87 and 1 where Pr[X chosen] > 0.5. Over this range, the decision maker's expected modal choice suggests risk seeking whereas the central tendency of his preferences, as represented by α^*, represents risk aversion or risk neutrality.

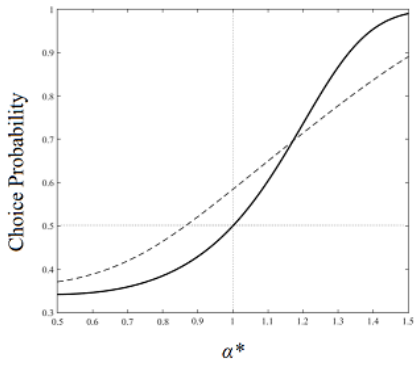


Figure 1. The probability of choosing a risky gamble X over its expected value Y for varying values of α^*, plotted with only Fechnerian noise (solid line) and with both Fechnerian and RP noise (dashed line). Here we can observe a higher choice probability of X over Y for some values of $\alpha^* < 1$ in the presence of RP noise.

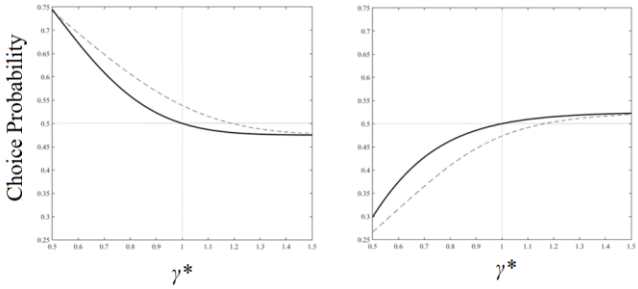
The important point of this illustration is to show that when there is both Fechnerian and RP noise, we cannot make reliable inferences about the decision maker's risk attitude using only modal choice frequencies, even if we assume independence between the different sources of noise. Moreover, as we show in the next two sections, the co-existence of Fechnerian and RP noise in conjunction with EUT core preferences can generate patterns of choice that have been interpreted as providing support for non-EU models such as CPT. In related work (not reported here) we also illustrate the pernicious effect of RP noise on parameter recovery and quantitative model fitting.

Probability Weighting: The Four-Fold Pattern

We now turn to cases where probabilities may be transformed nonlinearly as with CPT using the single parameter Prelec formulation given earlier. When $\gamma < 1$, this function overweights low probabilities and underweights high probabilities. Typically the crossover point is around 0.37, with $\pi(p) > p$ for $p < 0.37$ and $\pi(p) < p$ for $p > 0.37$. Such an 'inverse-S' function is crucial in capturing the famous four-fold pattern of risky choice (Tversky & Kahneman, 1992).

In the positive domain considered here, this pattern corresponds to modal choices favouring a risky gamble that offers a high payoff with a small probability relative to a sure option with the same expected value (which looks like risk-seeking), while at the same time generating modal choices favouring a sure option over a risky gamble with the same expected value that offers a large probability of a slightly higher payoff and a small probability of a considerably lower payoff (which looks like risk aversion). Thus in the choice between a risky gamble X^I offering a 1% chance of obtaining $10 and a 99% chance of obtaining $0, and its safe expected value equivalent Y^I offering $0.10 with certainty, decision makers typically choose X^I. In contrast, in the choice between a risky gamble X^{II} offering a 99% chance of obtaining $10 and a 1% chance of obtaining $0, and its safe expected value equivalent Y^{II} offering $9.90 with certainty, decision makers typically choose Y^{II}.

This pattern cannot be generated by deterministic EUT or EUT with only standard Fechnerian noise. But let us consider a setting with both Fechnerian and RP noise. For simplicity, we fix α = 1 so the utility function is linear, and we allow RP noise only in the γ parameter, with η_γ being distributed uniformly in the interval [-0.5, 0.5]. For Fechnerian noise, we use the Luce function with θ = 1, as in the previous section.



Figures 2a and 2b. The probability of choosing a low-probability risky gamble X^I over its expected value Y^I (left) and the probability of choosing a high-probability risky gamble X^{II} over its expected value Y^{II} (right), for varying values of γ^*. These figures are plotted with only Fechnerian noise (solid line) and with both Fechnerian and RP noise (dashed line).

Figure 2a shows the probability of choosing X^I over Y^I and Figure 2b shows the probability of choosing X^{II} over Y^{II}. As expected, a model with Luce noise but without noisy parameters and with $\gamma^* = 1$ entails for both pairs a 0.5 chance of choosing each option. For all $\gamma^* < 1$, the risky option is the modal choice in Figure 2a while the sure amount is the modal choice in Figure 2b. However, when γ itself exhibits symmetric noise and when this source of variability interacts with Fechnerian noise, the effect – as shown by the broken line – is to shift the transformation

path up in Figure 2a and down in Figure 2b: that is, the interaction increases the choice probability of X^I over Y^I and for Y^{II} over X^{II} for all γ^* considered. One implication of this is that at the point where $\alpha = 1$ and $\gamma^* = 1$ – that is, in the case where the deterministic model entails a risk neutral expected utility maximizer – the modal choices exhibit the 'mixed attitude to risk' typical of the data and widely viewed as supportive of CPT with $\gamma < 1$. Indeed, there is a range of γ^* between 1 and 1.15 for which the decision maker's expected modal choice generates a preference for X^I over Y^I and for Y^{II} over X^{II}, a behavioral pattern associated with the overweighting of small probabilities, whereas the central tendency of his preferences represents an underweighting of small probabilities.

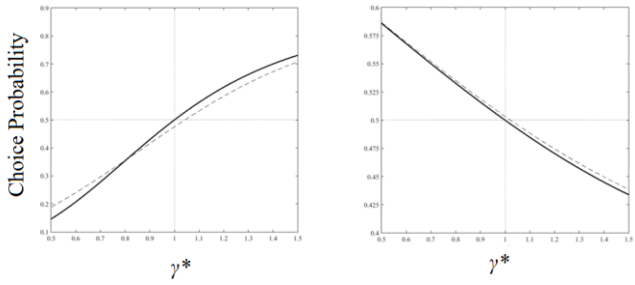
Probability Weighting: Common Ratio Effect

The probability weighting biases assumed by Prospect Theory are also necessary for it to account for the common-ratio effect (Kahneman & Tversky, 1979). The classic common-ratio case involves choices between two pairs of lotteries. One pair offers a gamble $X^{III} = (x, p; 0, 1\text{-}p)$ versus $Y^{III} = (y, 1)$ where p is typically around 0.8 and where y is equal to, or a little below, the expected value of X^{III}. In the example we consider in this section, our scaled-up pair is a choice between a risky gamble X^{III} offering an 80% chance of obtaining \$10 and a 20% chance of obtaining \$0, and its safe expected value equivalent Y^{III} offering \$8 with certainty. Here decision makers typically choose Y^{III}.

The second pair involves scaling down the probabilities of the positive payoffs in the first pair by some factor and correspondingly increasing the probabilities of 0 in both options to give a choice between $X^{IV} = (x, \lambda p; 0, 1\text{-}\lambda p)$ and $Y^{IV} = (y, \lambda; 0, 1\text{-}\lambda)$. Scaling down X^{III} and Y^{III} by a typical factor – letting $\lambda = 0.25$ – gives X^{IV} offering a 20% chance of obtaining \$10 and an 80% chance of obtaining \$0 versus its relatively safe expected value equivalent Y^{IV}, a 25% chance of obtaining \$8 and a 75% chance of obtaining \$0. In such scaled-down pairs, decision makers typically choose X^{IV} much more frequently: indeed, it is quite common for a sample of experimental participants to make Y^{III} the modal choice in the first pair but make X^{IV} the modal choice in the second pair. This is inconsistent with EUT, which assumes that individuals' preferences are linear in probabilities. Thus, scaling down the probabilities in this way reduces the EU of each option but leaves the ordering unchanged: a risk averse individual will prefer the Y option in both pairs, while a risk seeking individual will consistently choose the X options. In a deterministic world of EU maximizers, whatever proportion of the sample chooses X^{III} in the scaled-up pair should also choose X^{IV} in the scaled-down pair.

As with the four-fold pattern described above, the observed change in modal choices can be accommodated by CPT with $\gamma < 1$, leading to the overweighting of low probabilities and the underweighting of high probabilities. This is illustrated by the solid lines in Figures 3a and 3b in which we fix $\alpha = 1$, assume a Luce noise term with $\theta = 1$, and let γ range between 0.5 and 1.5. Here a model with Fechnerian noise but without RP noise and with $\gamma^* < 1$, entails that Y^{III} is the modal choice in Figure 3a while X^{IV} is the modal choice in Figure 3b.

But now suppose we allow both Fechnerian noise and RP noise to co-exist and interact. As above, to allow for noise in the γ parameter, we suppose η_γ is distributed uniformly in the interval [-0.5, 0.5]. Under these assumptions, we have a shift in the choice probabilities so that the interaction increases the choice probability of Y^{III} over X^{III} and of X^{IV} over Y^{IV} for γ^* in the neighbourhood of 1. Again, one implication of this is that at the point where $\alpha = 1$ and $\gamma^* = 1$ – that is, in the case where the deterministic model entails a risk neutral expected utility maximizer – the modal choices exhibit the reversal in choice probability observed in behavioral experiments, with a preference for Y^{III} in the scaled-up pair, but a preference for X^{IV} in the scaled-down pair. Thus we see that even though the central tendency values of the parameters entail not just EUT but expected value maximization, modal choices display a mixture of risk aversion and risk seeking in a manner resembling CPT with $\gamma < 1$. Indeed, as above, there is a range of γ^* between 1 and 1.02 for which the central tendency of the decision maker's underlying preferences, as represented by γ^*, represents an underweighting of small probabilities, whereas the behavioral pattern generated by these γ^* is commonly associated with the overweighting of small probabilities.



Figures 3a and 3b. The probability of choosing a scaled-up risky gamble X^{III} over its expected value Y^{III} (left) and the probability of choosing a scaled-down risky gamble X^{IV} over its expected value Y^{VI} (right), for varying values of γ^*. These probabilities are plotted with only Fechnerian noise (solid line) and with both Fechnerian and RP noise (dashed line).

Discussion

Differences between Decision Makers

We have seen that it is unsafe to infer an individual's underlying preferences from modal choice patterns, if we expect both RP noise and Fechnerian noise to play a role in the choice process. What is true for individuals may also be true for studies which draw conclusions about differences in preferences between different groups of people based either on differences in choice proportions or else on the basis of 'representative agent' assumptions. Between them, the disciplines of psychology, neuroscience, marketing, and economics have produced a large number of studies examining the relationship between risk preference and a wide variety of demographic, social, biological, cognitive,

emotional and neural variables. Much of this work makes the implicit or explicit assumption that differences in choice probabilities between different groups reflect differences in underlying utility/value functions and/or probability weighting preferences.

For example, based on choice proportions, men are considered to be more risk seeking than women (Charness & Gneezy, 2012); Chinese are considered more risk seeking than Americans (Hsee & Weber, 1999); the nucleus accumbens is seen as influencing risk seeking choices whereas the anterior insula is seen as influencing riskless choices (Kuhnen & Knutson, 2005); high incentives are associated with more risk aversion than low incentives (Holt & Laury, 2002); and decision makers under high time pressure are seen as being more risk averse than decision makers under low time pressure (Zur & Breznitz, 1981). Likewise stress is seen as affecting the amount of probability weighting in gains and losses (Porcelli & Delgado, 2009), the degree of striatal activity is assumed to influence the overweighting of small probabilities (Hsu et al., 2009), framing the decision as involving precaution is assumed to lead to the overweighting of small and medium-sized probabilities (Kusev et al., 2009), and decision feedback is considered to lead to objective probability weighting (Jessup et al., 2008). Finally, based on one of the most striking findings in contemporary decision making research, it is often assumed that decision makers tend to weigh probabilities differently when gamble payoffs and probabilities are *described* compared to when these payoffs and probabilities are *experienced* (Hertwig, 2015).

However, as we have shown, differences in choice proportions may be due not to differences in central tendency parameter values but rather to differences in the amount of variability in those underlying parameters. To illustrate, let us return to Figure 1. The horizontal axis in that Figure represents varying values of α^* under EUT and the vertical axis represents the choice probability for the gamble *X* corresponding with those different values of α^*. The two lines reflect varying levels of parameter noise in the model. Suppose that a male decision maker chooses *X* with frequency 0.52 while a female decision maker chooses *X* with frequency 0.48. If the degree of parameter noise were the same for both individuals, such a difference could reasonably be attributed to differences in underlying α^*, with the male having an increased propensity for risk seeking. But if the male's underlying preferences involve more parameter noise (the dotted line) than the female's (the solid line), then the opposite would be the case: the male's α^* would be approximately 0.9 as compared with the female's α^* of about 0.95. The same holds for inferences regarding probability weighting, as in Figure 2a and 2b.

It might be argued that there is no reason to suppose that males' parameter values are noisier than females' values. But until comparisons of choice frequencies were made, there was no strong *a priori* reason to suppose that gender systematically affected risk preferences. Why should it be preferences that are affected by gender (or age or time pressure, etc.) rather than the variability in parameter noise? Of course, this is not to imply that the conclusions drawn in the above papers are necessarily wrong. Nonetheless, many of those conclusions rely critically on the claim that it is preferences that are driving observed behavioral differences. To our knowledge none of this work explicitly considers the possibility that the changes in choice proportions observed across decision makers may be attributed to noise rather than to underlying preference. Indeed, in some of the settings considered above, a change to the amount of noise displayed by decision makers might be a more compelling explanation for observed behavioral differences, as compared with a change in the underlying parameter values.

Other Domains

Our analysis has focused upon binary risky choice, a domain in which within-person variability of choice has been widely observed. However, the potential confounding effects of interactions between sources of noise are not limited to risky choice: such effects are liable to distort inferences regarding underlying preferences in all non-linear utility-based models.

Consider, for example, the exponential discounting model (Frederick et al., 2002), which is commonly used to model choices between rewards occurring at differing periods of time. One of the most frequent criticisms of this model is that it cannot account for an increased preference for a proximate reward over a delayed reward as the lengths of the delay diminish by some common amount. For example a decision maker may prefer \$10 in five weeks to \$5 in four weeks, but also prefer \$5 immediately to \$10 in one week. This is typically attributed to hyperbolic discounting and present-bias. However, the findings in this paper suggest it may be possible to explain these reversals using only Fechnerian and RP noise without assumptions of present bias. If this is the case then it means that observed differences in these types of decisions that have been ascribed to demographic, biological, neural, cognitive, emotion, social, and task-based factors, may not necessarily reflect the impact of those factors on discount rates but might to some extent reflect differences in the effects of noise. Similar results may also hold for multi-attribute choice, altruistic choice, and strategic decision making.

Beyond Deterministic Models

Models such as expected utility theory and cumulative prospect theory are deterministic, and their predictions depend heavily on experimenter assumptions regarding their stochastic specifications. Many researchers have already argued that it is as important to choose the right type of noise as it is to choose the right core model, in order to accurately fit choice data (see e.g. Loomes, 2005). It has even been shown that the relationship between these deterministic models and their stochastic implementations is so great that it is possible to alter the relative fits of these

models by altering assumptions regarding their underlying sources of noise (Blavatskyy & Pogrebna, 2010).

In this paper we take this point further: the predictions of deterministic utility-based models are so dependent on their stochastic specifications that psychologically desirable assumptions about these specifications (such as noise in preferences) can alter the modal choice predictions associated with these models. These effects pose a strong challenge for utility models of preferential choice. How useful are models like EUT and CPT if their key predictions can be reversed by introducing some noise into the deliberation process?

This problem is ultimately endemic to deterministic models of choice, and cannot be remedied by the application of more rigorous methodological tools. In our opinion it suggests that theoretical research on decision making should attempt to move beyond these types of models when attempting to describe choice. There have already been a number of advances in modelling the cognitive basis of the stochastic choice process (see Rieskamp et al., 2006 and Oppenheimer & Kelso, 2015). Cognitive models of stochastic choice make explicit assumptions about how noise enters into deliberation, and how it interacts with preference, choice, and even decision time and confidence. In allowing stochasticity to play a central role in choice, these models are naturally able to capture a large range of behavioral effects that currently lie outside the descriptive scope of deterministic models. Indeed some of these models even try to explain key decision making anomalies using only unsystematic noise, rather than specific restrictions on value functions or probability weighting (Bhatia, 2014). Most importantly, however, the predictions of cognitive stochastic choice models are clearly defined. Additional assumptions about the sources of variability in choice are not necessary, and thus do not have the potential to reverse the key predictions of these models. Future research should consider using these types of psychologically-grounded stochastic choice models to understand the behavior of decision makers.

References

Becker, G. M., DeGroot, M. H., & Marschak, J. (1963). Stochastic models of choice behavior. Behavioral Science, 8(1), 41-55.

Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. Psychonomic Bulletin & Review, 21(5), 1095-1111.

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: comment on Regenwetter, Dana, and Davis-Stober (2011). Psychological Review, 118(4), 675-683.

Blavatskyy, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: why both are important for analyzing decisions. Journal of Applied Econometrics, 25(6), 963-986.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, 100(3), 432.

Charness, G., & Gneezy, U. (2012). Strong evidence for gender differences in risk taking. Journal of Economic Behavior & Organization, 83(1), 50-58.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. Journal of Economic Literature, 351-401.

Hertwig, R. (2015). Decisions from experience. The Wiley Blackwell handbook of judgment and decision making, 239-267.

Holt, C. A., & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. American Economic Review, 902-904.

Hsee, C. K., & Weber, E. U. (1999). Cross-national differences in risk preference and lay predictions. Journal of Behavioural Decision Making, 12.

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. Psychological Science, 19(10), 1015-1022.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263-291.

Kuhnen, C. M., & Knutson, B. (2005). The neural basis of financial risk taking. Neuron, 47(5), 763-770.

Kusev, P., van Schaik, P., Ayton, P., Dent, J., & Chater, N. (2009). Exaggerated risk: prospect theory and probability weighting in risky choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(6), 1487.

Loomes, G. (2005). Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. Experimental Economics, 8(4), 301-323.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. European Economic Review, 39(3), 641-648.

Luce, R.D. (1959). Individual choice behavior. John Wiley and Sons, New York.

Oppenheimer, D. M., & Kelso, E. (2015). Information Processing as a Paradigm for Decision Making. Annual Review of Psychology, 66, 277-294.

Porcelli, A. J., & Delgado, M. R. (2009). Acute stress modulates risk taking in financial decision making. Psychological Science, 20(3), 278-283.

Rieskamp, J., Busemeyer, J. and Mellers, B., (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. Journal of Economic Literature, 44, 631–61.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323.

Zur, H. B., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. Acta Psychologica, 47(2), 89-104.

Do Do Do, The The The:
Interactivity and Articulatory Suppression in Mental Arithmetic

Frédéric Vallée-Tourangeau
Department of Psychology, Kingston University
Kingston-upon-Thames UNITED KINGDOM KT1 2EE
f.vallee-tourangeau@kingston.ac.uk

Miroslav Sirota
Department of Psychology, University of Essex
Colchester UNITED KINGDOM CO4 3SQ
msirota@essex.ac.uk

Gaëlle Vallée-Tourangeau
Department of Management, Kingston University
Kingston-upon-Thames UNITED KINGDOM KT2 7LB
g.vallee-tourangeau@kingston.ac.uk

Abstract

Doing long sums in the absence of complementary actions or artefacts is a multi-step procedure that quickly taxes working memory; congesting the phonological loop further handicaps performance. In the experiment reported here, participants completed long sums either with hands down—the low interactivity condition—or by moving numbered tokens—the high interactivity condition—while they repeated 'the' continuously, loading the phonological loop, or not. As expected, articulatory suppression substantially affected performance, but more so in the low interactivity condition. Independent measures of basic arithmetic skill and mathematics anxiety moderated the impact of articulatory suppression on performance in the low but not in the high interactivity condition. These findings suggest that working memory resources are augmented with interactivity, underscoring the importance of characterizing the properties of the system as it is configured by the dynamic agent-environment coupling.

Keywords: Interactivity, Mental Arithmetic, Articulatory Suppression, Working Memory, Systemic Cognition

Introduction

Different components of working memory are engaged in doing long sums without external aids or complementary actions (Raghubar, Barnes, & Hecht, 2010). The exact involvement of these components depends on the complexity of the arithmetic task, the presentation format and modality of presentation, as well as the agent's level of mathematical competence (DeStefano & LeFevre, 2004). Take the task of adding a long series of single-digit numbers presented visually all at once in a random pattern. The requisite arithmetic skills to compute the correct total are certainly mastered by numerate young adults. However, calculating the correct answer in a multi-step procedure requires temporary storage and executive skills: interim totals are calculated and rehearsed, numbers tagged as having been added, others tagged as not, attention allocated to certain areas of the visual presentation or switched to others to identify what number or the easiest number to add next, arithmetic knowledge retrieved from long term memory to facilitate the identification of congenial sub-totals. Some of these processes rely on the sub-vocal rehearsal of cumulative interim sums. It is no surprise that loading the phonological loop in dual-task paradigms interferes with mental arithmetic that requires counting (Fürst & Hitch, 2000; Logie, Gilhooly, & Wynn, 1994).

Complementary Actions and Interactivity

The role of working memory in mental arithmetic is traditionally established with an experimental procedure that limits or prevents participants from modifying the problem presentation in working out an answer. These research efforts reflect a commitment to a representational and internalist model of cognitive processing. In order to create an unadulterated window onto the processes implicated in mental arithmetic and to permit the clinical precision of their segmentation, simple problems devoid of content are presented in a manner that cannot be modified by the agent. However, once released from the confines of the cognitive psychologist's laboratory, mental arithmetic is often situated (Lave, 1988) and naturally supported by a range of complementary actions, such as pointing, to guide attention and bind elements in a functional sequence (Kirsh, 1995, Carlson, Avraamides, Cary, & Strasberg, 2007). Gesturing can also lighten the cognitive load: individuals who were not allowed to

gesture while explaining how they solved a mathematical problem also exhibited poorer recall on an interfering memory task compared to individuals who were allowed to gesture (Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001).

Interacting with an external representation of the problem can modify its physical presentation and the agent's "mental" computations are reflected in the dynamic changes in the problem's appearance. Imagine, again, adding a long series of single digit numbers, however, this time, the numbers do not configure a static visual presentation, but rather adorn the face of wooden tokens creating a malleable physical configuration which participants can modify as they work on the problem. The calculation unfolds along a spatio-temporal itinerary wrought by the agent's actions. These actions modify the problem presentation and in doing so the problem is restructured, added numbers can be physically demarcated, no longer exerting attentional pull, congenial interim totals (e.g., 8+7) are identified and physically segregated, shifting the affordances of what to do next, guiding the agent to identify complementary sub-totals (e.g., 9+6), that inter-lock to create easy-to-remember provisional sums (e.g., 30), improving efficiency and reducing error. Thus, the dynamic reconfiguration of the problem guides, in part, the allocation of attentional resources and strategy selection (Vallée-Tourangeau, 2013).

The Present Experiment

The present experiment employed a dual-task procedure to explore the impact of articulatory suppression in a mental arithmetic task. The task involved adding 11 single-digit numbers presented either as a static configuration (a low interactivity condition) or as a set of number tokens that could be manipulated in calculating the answer (a high interactivity condition). Participants completed the task either with articulatory suppression—by repeating aloud 'the' continuously— or without. Thus the experiment employed a 2(Interactivity: Low, High) × 2(Articulatory Suppression: Without, With) design with both factors as repeated measures. Past research findings led us to expect poorer performance with articulatory suppression, but better performance with interactivity. If interactivity augments an agent's working memory capacity, the impact of articulatory suppression on performance should be mitigated in the high interactivity condition, such as to result in an interaction between the two factors: The performance advantage conferred by a high degree of interactivity should be greater with articulatory suppression than without.

We also profiled participants in terms of their (i) basic arithmetic skills, (ii) level of mathematics anxiety, and (iii) executive function with an attention-switching task. We used these concomitant variables to determine whether they were moderators of the impact of suppression on mental arithmetic, and whether the moderation was the same in the low and high interactivity condition. We expected all three variables to moderate the impact of suppression primarily in the low interactivity condition; if a higher degree of interactivity augments working memory resources, then participants' performance would be more resilient and the moderating properties of these factors might be attenuated.

Method

Participants

Fifty-two Kingston University psychology undergraduate and postgraduate students (45 females) participated in the experiment in exchange for course credits (M_{age} = 21.8, *SD* = 4.0).

Materials and Measures

Arithmetic Task. Participants were invited to add series of 11 single digits. For each sum the digits were arrayed in a random cloud pattern, and were presented either on a sheet of A4 or as identically-arranged wooden tokens. Participants were instructed to calculate the sum as quickly as they could and announce their answer to an experimenter. They did so either with their hands flat on the table top in front of them and were not allowed to use their fingers to count or point (low interactivity) or by moving the tokens about as they saw fit in producing an answer (high interactivity). Performance on this arithmetic task was measured in terms of accuracy—percentage correct and absolute calculation error—solution latencies and efficiency. Participants' efficiency at calculating the sums was measured as the ratio of their accuracy—percentage correct—over the resources invested in arriving at the answer. The later was operationalized as the proportion of time taken to announce an answer out of the maximum of time to do so as indexed by the average latency of the slowest quartile. Thus, if a participant's accuracy was 80%, taking an average of 60 seconds to announce an answer, and that the average latency for the slowest quartile was 80 seconds, that participant's efficiency ratio would be 80%/(60/80) or 80%/75%, hence 1.067. A ratio of 1 or greater indicates efficient performance, whereas a ratio below 1 indicates inefficient performance.

Mathematics Anxiety Scale. Participants completed a 25-item Mathematics Anxiety Scale-UK (MAS-UK; Hunt, Clark-Carter & Sheffield, 2011). The questionnaire invited participants to imagine how anxious they would feel in certain situations (1 = "not at all" and 5 = "very much"), such as "Working out how much your shopping bill comes to" or "Taking a maths exam".

Basic Arithmetic Skill. Basic arithmetic skill was measured by having participants complete as many of 45

simple expressions (such as 11-9 = ?) in a 60-second period.

Executive Function: Shifting. The plus-minus task reported in Miyake, Friedman, Emerson, Witzki and Howerter (2000) was employed to measure attention switching skills. With three different series of 30 double-digit numbers, participants were instructed to add 3 to each in the first series, subtract 3 to each in the second series, and alternate between adding and subtracting 3 with the third series. The switching cost, measured in seconds, was the difference in completion time for the third series minus the average completion time for the first two. Larger differences indicated poorer switching skills.

Procedure

Ten different sums of 11 single-digits were created: none of the sums were the same and totals ranged from 57 to 80. From these, five were randomly selected and allocated to the low interactivity condition, and the other five to the high interactivity condition for each participant. Participants completed these five sums twice within each level of interactivity: Once with articulatory suppression, once without. Thus the design employed was a 2 (Interactivity: Low, High) × 2 (Articulatory Suppression: Without, With) repeated measures. The order of the four conditions for each participant was constructed as follows: One of the four conditions was randomly selected to be the first condition experienced by the participant. Once that first condition was identified, the order of the other three was determined by the following constraint: conditions with the same level of interactivity could not be presented in succession (e.g., the two high interactivity conditions experienced consecutively). For example, if the condition with low interactivity and with articulatory suppression was the first condition experienced by a participant then the remaining conditions could be presented in the following order: (ii) high interactivity without articulatory suppression; (iii) low interactivity without articulatory suppression; (iv) high interactivity with articulatory suppression. As a result, participants never calculated the same set of five sums in succession. The first presentation of a condition with articulatory suppression was always preceded by a training task during which participants were asked to write successive subtractions of 3 starting from 100 for one minute while continuously repeating 'the'. The instructions read: "You will be presented with 5 addition problems involving single digit numbers. For each of the problems, you must add the digits as quickly and as accurately as you can". The low interactivity instructions then read "During this task you will be required to keep your hands flat on the table and must not move them for the duration of this task" while the high interactivity instructions read: "During this task you will be able to manipulate the tokens as you see fit for this task". Participants experienced each interactivity condition twice, once with articulatory suppression, once without. The articulatory suppression instructions read: "You will also be required to repeat the word "the" throughout the duration of the task as you did in the practice task. You can start repeating the sound and calculating the sums when prompted". Participants were instructed to announce their answer to the experimenter once completed. In the articulatory suppression conditions if more than two seconds elapsed without participants engaging in the secondary task, they were prompted to comply with the task. Finally, the presentation of each condition was separated by the completion of either the Basic Arithmetic Skill test, the Mathematics Anxiety Scale, or the Attention Switching task; the order of these three tasks was counterbalanced across participants.

Results

Mental Arithmetic Performance

Participants' performance was measured in terms of the percentage of sums correctly solved (out of five), the average absolute calculation error, the average latency to solution, and the efficiency ratio in each of the four conditions.

Percentage Correct. The mean percentage of correct additions in the four experimental conditions are reported in the top portion of Table 1. As expected, participants were better at providing correct answers in the absence of articulatory suppression; however performance was always better in the high interactivity condition. In addition, the decline in performance with articulatory suppression appeared steeper in the low interactivity condition. A 2×2 repeated measures analysis of variance (ANOVA) confirms these impressions: The main effect of suppression was significant, $F(1, 51) = 60.1$, $p < .001$, $\eta_p^2 = .54$, as was the main effect of interactivity, $F(1, 51) = 13.6$, $p = .001$, $\eta_p^2 = .21$; the interaction was also significant, $F(1, 51) = 4.06$, $p = .049$, $\eta_p^2 = .07$.

Absolute Calculation Error. The mean absolute calculation error in the four conditions are plotted in Figure 1. These data illustrate a substantial effect of suppression, with larger deviations from the correct answers recorded with suppression than without. Errors were generally smaller in the high interactivity condition, and more important, articulatory suppression appeared not to have as dramatic an impact on calculation accuracy in the high interactivity condition. In a 2×2 repeated measures ANOVA the main effects of suppression, $F(1, 51) = 36.2$, $p < .001$, $\eta_p^2 = .42$ and interactivity, $F(1, 51) = 9.69$, $p = .003$, $\eta_p^2 = .16$, were significant, as was the interaction, $F(1, 51) = 6.02$, $p = .018$, $\eta_p^2 = .11$.

Solution Latency. The mean solution latencies are reported in the middle portion of Table 1. In the absence of articulatory suppression, solution latencies were similar in the low and high interactivity conditions. And while

participants were generally slower with articulatory suppression, they were slowest in the high interactivity condition. The 2×2 repeated measures ANOVA revealed a significant main effect of suppression, $F(1, 51) = 16.4$, $p < .001$, $\eta_p^2 = .25$, a significant main effect of interactivity, $F(1, 51) = 11.5$, $p < .001$, $\eta_p^2 = .19$, as well as a significant interaction, $F(1, 51) = 18.6$, $p < .001$, $\eta_p^2 = .27$.

Table 1: Mean and standard deviation for the percentage correct, latency (measured in seconds), and efficiency ratio for the five sums in the low and high interactivity condition without and with articulatory suppression.

	Articulatory Suppression			
	Without		With	
	Percentage Correct			
Interactivity	*M*	*SD*	*M*	*SD*
Low	56.0%	27.7%	26.9%	31.7%
High	63.5%	28.2%	44.6%	29.5%
	Latencies			
	M	*SD*	*M*	*SD*
Low	38.1	16.4	41.5	21.2
High	37.4	14.8	51.2	27.2
	Efficiency Ratio			
	M	*SD*	*M*	*SD*
Low	1.20	0.99	0.68	0.95
High	1.16	0.70	0.99	0.84

Efficiency Ratio. The mean efficiency ratios are reported in the bottom portion of Table 1: Participants were much more efficient in the absence of articulatory suppression, and efficiency declined sharply with suppression. However, in the high interactivity condition, participants' efficiency ratio remained good even with articulatory suppression. A 2×2 repeated measures ANOVA supported these impressions: The main effect of suppression was significant, $F(1, 51) = 32.8$, $p < .001$, $\eta_p^2 = .39$, but the main effect of interactivity was not, $F(1, 51) = 2.61$, $p = .113$, $\eta_p^2 = .05$; however, the interaction between suppression and level of interactivity was significant, $F(1, 51) = 7.47$, $p = .009$, $\eta_p^2 = .13$.

Moderators of the Impact of Suppression on Calculation Error

Participants were profiled in terms of their basic arithmetic skills, level of math anxiety and executive function using an attention switching task. To test our moderation hypotheses we conducted a moderation analysis for within-subject design using ordinary least square regression with difference scores, as proposed and formalized by Judd, Kenny and McClelland (2001). This was a preferred solution for the current experiment with a sample size that is not optimal for a multilevel modelling. Judd et al. suggested that the moderation in within-subject designs occurs when a concomitant variable (e.g., basic arithmetic skill, or math anxiety level) predicts differences in performance between two conditions. Thus we determined how these variables moderated the difference in performance with and without articulatory suppression (within each level of interactivity, and then by collapsing level of interactivity). Table 2 reports the correlations between each of the concomitant variables and the difference in absolute calculation errors between the condition with articulatory suppression and the condition without, when interactivity level is low, high, and when collapsing over the two levels of interactivity ($df = 50$ for all correlation coefficients).

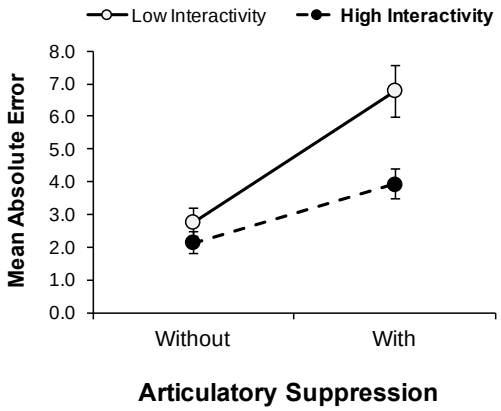


Figure 1: Mean absolute deviation (with standard errors) in the low and high interactivity condition as a function of the absence and presence of articulatory suppression.

Basic Arithmetic Skill. Overall, the increase in absolute calculation error when collapsing across interactivity conditions was moderated by basic arithmetic skills, $r = -.37$, $p = .007$, that is the higher was participants' arithmetic skill, the smaller the increase in calculation error with articulatory suppression. This relationship was also observed in the low interactivity condition, $r = -.28$, $p = .042$ but less so in the high interactivity condition, $r = -.27$, $p = .056$.

Math Anxiety. When collapsing the data over both interactivity conditions, levels of mathematic anxiety moderated the impact of articulatory suppression, $r = .34$, $p = .014$; that is, the higher the level of math anxiety, the higher the increase in calculation error with articulatory suppression. However, this overall pattern obscures a more interesting pattern across levels of interactivity. Thus in the low interactivity condition math anxiety was a significant moderator of the increase in error with suppression, $r = .41$, $p = .003$, but not in the high interactivity condition, $r = .01$, $p = .951$.

Attention Switching. As the correlation coefficients reported in the bottom row of Table 2 indicate, scores on the attention switching test did not moderate the increase in calculation error with articulatory suppression.

Table 2: Correlation between increase in absolute calculation error as a function of articulatory suppression and Basic Arithmetic Skill (BAS), Math Anxiety Score (MAS), and Attention Switching Score (SWITCH) in the low and high interactivity condition, and overall (df = 50).

	Interactivity Condition		
	Low	High	Overall
BAS	-.28 *	-.27	-.37 **
MAS	.41 **	.01	.34 *
SWICH	-.01	.04	.01

Note. * $p < .05$, ** $p < .01$

Discussion

This experiment explored how interactivity could mitigate the impact on mental arithmetic performance of a reduction in working memory resources through articulatory suppression. We predicted that performance overall would be influenced by both articulatory suppression and interactivity, and more important, that these two factors would interact such that the impact of articulatory suppression would be more pronounced in the condition with low interactivity. This is what we observed: While mental arithmetic performance was always poorer with articulatory suppression, the deterioration of accuracy was always significantly greater when participants completed the sums with their hands palm down on the table top. The repeated-measures design ensured that differences across conditions were not in themselves a reflection of differences in arithmetic skills or working memory capacity across participants.

The data reported here on the effect of articulatory suppression on counting performance corroborate previous findings (see Raghubar et al., 2010). However, interactivity attenuated the impact of a secondary task that taxed the phonological loop which reduced participants' ability to rehearse interim totals or plan counting strategies sub-vocally. The possibility of restructuring the physical problem presentation over the course of the calculation ensured that the participants could reconfigure the environment in a manner that compensated for the reduction in internal working memory capacity. This is not to say that working memory was augmented such as to soak up completely the resources depletion caused by articulatory suppression since performance was affected by the secondary task, but nonetheless it was sufficiently robust to ensure efficient calculations, as reflected by the efficiency ratio measure that did not dip substantially below 1 as it did in the low interactivity condition with articulatory suppression.

Accuracy in the high interactivity conditions dropped by 20% with articulatory suppression, and latency increased by nearly 14 seconds on average (a 37% increase in latency). In contrast, the latency across the low interactivity conditions increased by 3.4 seconds on average with articulatory suppression (a 9% increase in latency). At first, the latency data might suggest participants did not fully engage with the secondary task in the low interactivity condition, yet accuracy was down by 30% and absolute calculation error were four times as large with articulatory suppression in the low interactivity condition (see Fig. 1). Rather, what these relatively short latencies indicate was that the task was very hard in the low interactivity condition with articulatory suppression: participants abandoned more quickly than in the high interactivity condition and were more likely to guess the answer. It is interesting to note that participants' level of mathematics anxiety was a significant moderator of the impact of suppression on calculation error, but only in the low interactivity condition. This suggests that math anxious participants might have guessed more in low interactivity condition, reducing problem latency but also increasing error. This pattern has been previously reported in the math anxiety literature (e.g., Ashcraft & Klause, 2007). Thus, in this simple arithmetic task, the reduction of internal working memory capacity through articulatory suppression had its most deleterious effect on participants with higher levels of math anxiety.

Participants' basic arithmetic skills moderated the impact of articulatory suppression. This pattern was marginally more pronounced in the low than in the high interactivity condition. Along with the moderating influence of math anxiety on the impact of suppression across levels of interactivity, these data suggest that a higher degree of interactivity produced more resilient performance irrespective of differences in skills and anxiety. Finally, the attention switching scores did not moderate the impact of suppression on performance either in the low or high interactivity condition. This finding is a little puzzling. To the extent that this task gauges participants' ability to switch their attention, we expected these scores to correlate positively with changes in performance as a function of suppression; they did not. A more precise measure of attention switching, perhaps using an automated task, or a composite score from different attention switching tasks, might offer a more informative window on how switching skills might moderate the influence of articulatory suppression on mental arithmetic.

Overall, these moderation patterns are interesting. Psychometric efforts to unveil the cognitive capacities and dispositions subserving performance in a domain must be interpreted relative to a context of reasoning. In the experiment reported here, the context varied in terms of the degree of interactivity it afforded and the cognitive resources that could be deployed on the primary task. Math anxiety may be an important moderator of mental arithmetic performance (Ashcraft, 2002), but allowing participants to interact with a malleable problem presentation attenuates the impact of math anxiety on

performance. Similarly, a high interactivity context elevates performance such that participants with poorer arithmetic skills perform in a manner similar to participants with stronger skills.

The data presented here validate a systemic perspective on cognition, one that seeks to describe cognitive products and processes of a system configured by the dynamic coupling of an agent and his or her physical environment (Vallée-Tourangeau & Vallée-Tourangeau, 2014; Vallée-Tourangeau, Abadie, & Vallée-Tourangeau., 2015). In the present mental arithmetic task, interacting with the physical problem presentation transformed an agent's ability to solve these problems. The resulting performance invites a characterization of the cognitive capacities of the system rather than of the agent. Profiling an agent's cognitive resources—such as working memory capacity—with tasks that eliminate interactivity with a physical problem presentation will paint an inaccurate and perhaps distorted picture of how these resources are deployed, augmented and transformed once the agent is released from the psychologist's laboratory and is embedded in the physical world.

Acknowledgements

We thank Emily Cox-Crowley, Lisa Guthrie, Alexandra Licudi, and Francesca Tromans for their help with the preparation of the experimental material and the recruiting and running of the participants.

References

Ashcraft, M. H. (2002). Math anxiety: Personal, educational and cognitive consequences. *Current Directions in Psychological Research, 11*, 181-185.

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review, 14*, 243-248.

Carlson, R. A., Avraamides, M. N., Cary, M., & Strasberg, S. (2007). What to the hands externalize in simple arithmetic? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 747-756.

DeStefano, D., & LeFevre, J. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16*, 353-386.

Fürst, A. J., & Hitch, G. J. (2000). Separate roles for executive and phonological components of working memory in mental arithmetic. *Memory & Cognition, 28*, 774-782.

Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011) The development and part validation of a UK scale for mathematics anxiety. *Journal of Psychoeducational Assessment, 29*, 455-466.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science, 12*, 516-522.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. Psychological Methods, 6, 115-134.

Kirsh, D. (1995). Complementary strategies: Why we use our hands when we think. In J. M. Moore & J. L. Lehman (Eds.), *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 212-217). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. New York: Cambridge University Press.

Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition, 22*, 395-410.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*, 110-122.

Vallée-Tourangeau, F. (2013). Interactivity, efficiency, and individual differences in mental arithmetic. *Experimental Psychology, 60*, 302-311.

Vallée-Tourangeau, F., Sirota, M., & Villejoubert, G. (2013). Reducing the impact of math anxiety on mental arithmetic: The importance of distributed cognition. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society* (pp. 3615-3620). Austin, TX: Cognitive Science Society.

Vallée-Tourangeau, F., & Vallée-Tourangeau, G. (2014). Diagrams, jars and matchsticks: A systemicist's toolkit. *Pragmatics & Cognition, 22*, 187-205.

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General, 144*, 581-603.

Between versus Within-Language Differences in Linguistic Categorization

Anne White (anne.white@ppw.kuleuven.be)
Gert Storms (gert.storms@ppw.kuleuven.be)
Faculty of Psychology and Educational Sciences, University of Leuven
Tiensestraat 102, B-3000 Leuven, Belgium

Barbara C. Malt (barbara.malt@lehigh.edu)
Department of Psychology, Lehigh University
17 Memorial Drive East, Bethlehem, PA 18015-3068, USA

Steven Verheyen (steven.verheyen@ens.fr)
Institut Jean-Nicod, PSL Research University, École Normale Supérieure
Pavillon Jardin, 29 rue d'Ulm, 75005 Paris, France

Abstract

Cross-linguistic research has shown that boundaries for lexical categories differ from language to language. The aim of this study is to explore these differences *between* languages in relation to the categorization differences *within* a language. Monolingual Dutch- (N=400) and French-speaking (N=300) Belgian adults provided lexical category judgments for three lexical categories that are roughly equivalent in Dutch and French. Each category was represented by good, borderline, and bad examples. A mixture modeling approach enabled us to identify latent groups of categorizers within a language and to evaluate cross-linguistic variation in relation to within-language variation. We found complex patterns of lexical variation within as well as between language groups. Even within a seemingly homogeneous group of speakers sharing the same mother tongue, latent groups of categorizers display a variability that resembles patterns of lexical variation found at a cross-linguistic level of comparison.

Keywords: artifact categories; cross-linguistic differences; semantic variation; vagueness

Introduction

People of different languages and cultures share a perception of the similarity among entities within at least some domains (e.g., common household containers: Ameel, Storms, Malt, & Sloman, 2005; Malt, Sloman, Gennari, Shi, & Wang, 1999; color: Roberson, Davies, Corbett, & Vandervyver, 2005: human locomotion: Malt, Ameel, Imai, Gennari, Saji, & Majid, 2014; and spatial relations: Munnich, Landau, & Dosher, 2001). Despite the shared non-linguistic appreciation of these domains, its relation with linguistic categorization is complex: Linguistic categories do not map directly onto similarity clusters (Ameel et al., 2005; Malt et al., 1999).

In different languages the world is carved up differently. This cross-linguistic variation has been shown for domains as varied as color, causality, mental states, number, body parts, containers, motion, direction, and spatial relations (Malt & Majid, 2013; Malt et al., 2015). Malt and colleagues, for instance, described how different languages label a set of household containers and found that not all languages observe the same distinctions, despite perceiving the similarity of the objects in the same way (Ameel et al., 2005; Malt, Sloman, & Gennari, 2003; Malt et al., 1999). For example, the Dutch word for bottle (*fles*) encompasses objects that in French are either called *bouteille* or *flacon*. Not only are there differences in the number of distinctions made in different languages, there is crosscutting in the way exemplars of a category are grouped together as well (Malt et al., 2003). The roughly equivalent French *bouteille* and Dutch *fles* demonstrate a difference in how they map onto a shared similarity space, which reflects a cross-linguistic difference in meaning representation. Additionally, the categories *fles* and *bouteille* each include a different number of objects, indicating differences in category extension as well.

Although these cross-linguistic differences have received growing attention in recent years, within-language variation exists as well (McCloskey & Glucksberg, 1978; Verheyen, Hampton, & Storms, 2010). Inter-individual differences in linguistic categorization have been described in the relation to vagueness (Black, 1937; Verheyen & Storms, 2013). A distinction is made between vagueness in criteria and vagueness in degree (Devos, 2003). The former is involved when individuals use different criteria to determine if an object belongs to a category. When individuals agree on the criteria for category membership but use a different cut-off for separating members from non-members, the latter type of vagueness is in play. In seemingly homogeneous groups of speakers of the same language, groups that display one or both types of differences have been identified (Verheyen & Storms, 2013).

The aim of this study is to quantitatively explore the extent and nature of differences in categorization between two language groups with respect to the variation existing between latent groups of categorizers within a language. More specifically this study evaluates the degree of within-language variability in relation to the degree of cross-linguistic variability for roughly equivalent categories. To this end we collected category judgment data from Belgian Dutch and French speaking participants, who share a similar environment and who perceive similarity in the tested domain in much the same way (Ameel et al., 2005).

A mixture modeling approach was used in order to identify latent groups of categorizers in a seemingly homogeneous group, that is, adult speakers of the same mother tongue. The mixture model partitions a participant sample into subgroups of individuals who display similar categorization behavior. By doing so, it identifies subgroups that use different criteria in making their category decisions (Verheyen & Storms, 2013; Verheyen, Voorspoels & Storms, 2015). We compared the categorization patterns of Dutch-speaking and French-speaking Belgian participants for the roughly equivalent categories *doos* and *boîte* (similar to English *box*), *fles* and *bouteille* (similar to English *bottle*) and *pot* and *pot* (similar to English *jar*). We also identified latent groups within each language, and the differences between the latent within-language groups were compared to those between languages.

Method

Participants

We collected data from approximately 400 monolingual Dutch-speaking and 300 monolingual French-speaking Belgian adult participants. Data from participants who were determined to be bilingual were discarded, as well as data of participants under the age of 17, resulting in the sample sizes displayed in Table 2.

Materials

The stimulus material consisted of an existing set of pictures of household containers described by Ameel et al. (2005) expanded with new stimuli, totaling to 192 stimuli. The new pictures were made according to the guidelines used by Ameel et al. (2005). The objects were photographed in color against a neutral background with a constant camera distance to preserve relative size. A ruler was included in front of each object to provide additional size information.

Four lexical categories were investigated, selected based on unpublished naming data for the full set of 192 common household objects. The four most frequently generated category names across the complete set by 32 monolingual Dutch-speaking adult participants were *fles*, *pot*, *bus*, and *doos*. Because it was not feasible to conduct a categorization experiment with multiple categories for the complete set, a selection of 40 stimuli per category was made. In order to make an adequate selection of the stimuli, a pilot category judgment task with approximately 30 Dutch speaking participants was conducted. In the pilot study participants had to decide if a given name was suited for a presented object by responding 'yes' or 'no'.

The stimuli for this study were selected based on the results of the pilot study according to the following criteria. The selected set of stimuli spanned the full range of proportion of yes-responses, varying from approximately 0.10 to 0.90. This selection contained a mixture of clear members, borderline members, and clear non-members, spanning the range of shapes and sizes for the category. The earlier collected naming data of Dutch and French participants were also taken into account. Some objects that showed incongruities in the use of category labels between Dutch and French were included. For instance, a cooking *pot* was called *pot* by almost all Dutch participants, while in French this is rarely called *pot*.

For the French version of the task, four categories were selected based on naming data of adult monolingual French participants, analogous to the category selection for the Dutch version. The four most generated category names were *bouteille*, *pot*, *flacon*, and *boîte*. Because these categories do not map directly onto the categories for the Dutch task (Malt et al., 1999), the French task was composed as follows. The stimulus set for *bouteille* and *flacon* both consisted of the objects presented in the Dutch category *fles*. For the French category *boîte* the items belonging to the Dutch category *doos*, and for the French category *pot* the items belonging to the Dutch category *pot* were presented. The French speaking participants judged the same set of objects as the Dutch speaking participants, with the exception of two objects that were only presented in the Dutch category *bus*. Further descriptions will be limited to the roughly equivalent categories *doos-boîte*, *fles-bouteille* and *pot-pot*. The categories *flacon* and *bus* will be disregarded in the discussion of the results, since they do not have a roughly equivalent category in Dutch and French, respectively.

Procedure

The linguistic categorization task was conducted via Qualtrics and the link to the task was distributed via social networks (e.g. Facebook), both for the French-speaking and Dutch-speaking participants. After informed consent, demographical information was collected (age, gender, education level, and mother tongue). The participants were instructed to decide for each item in the series whether or not it belonged to a particular category. The four categories were presented to participants in random order, as were the pictures within each category. Above every picture the question 'Is this an X?' was displayed, which participants could respond to by choosing 'yes' or 'no'. The instructions explicitly stated that pictured objects could belong to one or more categories and that there were no right or wrong answers. The full survey took between 10 to 15 minutes to complete.

Model analyses

Item response theory (IRT) modeling can be used as a formalization of the Threshold Theory of Semantic Categorization (Hampton, 1995; Verheyen, Hampton, & Storms, 2010). The Threshold Theory (Hampton, 1995, 1998, 2007) accounts for vagueness in degree by allowing individuals to use a different cut-off or threshold along the criterion for category membership. In that case participants diverge only with respect to the category's extension.

The use of a mixture IRT model allows for the identification of subgroups within a seemingly homogeneous group, that use different criteria in their

category judgment. By allowing for subgroups, the assumption that all participants employ the same criterion is relaxed, and thus vagueness in criteria is accounted for. Within each of the identified groups, individuals can still differ in terms of their categorization cut-off.

Individual categorization decisions Y_{ci}, where c refers to a categorizer and i to an item, serve as input for the mixture model. When an item i is endorsed as a category member by categorizer c, Y_{ci} takes value 1; when it is not endorsed as a member of the target category it takes value 0. Each of these categorization judgments is regarded as an outcome of a Bernoulli trial with equation (1) modeling the probability of a positive response.

$$\Pr(Y_{ci} = 1) = \frac{e^{\alpha_g(\beta_{gi} - \theta_c)}}{1 + e^{\alpha_g(\beta_{gi} - \theta_c)}} \quad (1)$$

For each latent group g of categorizers a separate criterion is extracted. The values for the parameters β_{gi}, θ_c, and α_g are estimated by application of the model to an item by participant categorization matrix. The position of each item i along the criterion of group g is indicated by the estimate β_{gi} and represents the extent to which that item meets the group's criterion. θ_c represents categorizer c's threshold or cut-off: the extent to which items need to meet the criterion to be endorsed as category members. The relative position of θ_c and β_{gi} defines the probability of endorsement: the more β_{gi} exceeds θ_c, the higher the probability that the item will be endorsed as a category member. Conversely, the more to the left of θ_c β_{gi} is positioned, the lower the probability that the item will be endorsed. A separate α_g for each group determines the shape of the response function (Verheyen & Storms, 2013; Verheyen, Voorspoels, & Storms, 2015).

For each of the six (3 categories x 2 languages) data sets, the model in Equation (1) was estimated with 1, 2, 3, 4, and 5 different groups. The parameters in Equation (1) were estimated in a Bayesian manner, using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) running 3 chains of 10,000 samples each with a burn-in of 4,000 samples. The chains were checked for convergence and label switching. The reported results are based on the posterior means for the models that yielded the smallest Bayesian Information Criterion (BIC). These models yield the best approximation of the categorization data when both model fit and complexity are taken into account. See Verheyen, Voorspoels, and Storms (2015) for example code, rationale for prior specification, and model selection simulations.

Results

Cross-linguistic differences

We start by presenting the results for all participants of a language group. Figure 1 displays per item the categorization proportion of French participants for the French category *boîte* and of Dutch participants for the Dutch category *doos*. This graph confirms the idea that *boîte* and *doos* are roughly equivalent categories since the categorization proportions for French and Dutch seem to follow a roughly similar rising trend. However, several notable differences in categorization proportions can be observed as well.

The categorization proportions for the French *boîte* (M=0.45) are on average higher than the categorization proportions for the Dutch *doos* (M=0.40) ($t(39) = 3.287$, $p < 0.01$). The observation that *boîte* has a larger category extension than *doos* suggests a degree difference: the category *boîte* is somewhat larger than the category *doos*.

This is not the only difference between the two categories. There are also indications that the two language groups use different categorization criteria since the shape of the proportion curves is not the same. If the x-axis of the graph would be organized according to the categorization proportions of the French participants, one would obtain a different order of the items along the axis. This is demonstrated by an imperfect correlation of 0.78 between *doos* and *boîte*. Some of the French proportions are even lower than the corresponding Dutch ones, despite the established degree difference. So although the French category *boîte* includes more objects in general, some objects are not considered to be as good a category member as they are in the smaller Dutch category *doos*.

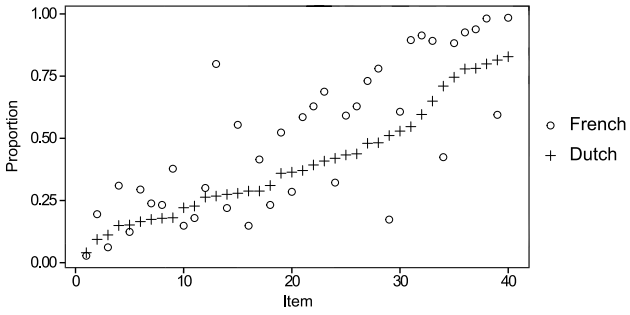


Figure 1: Categorization proportions per item for the French category *boîte* and the Dutch category *doos*. The order of the items is determined according to the Dutch categorization proportions.

As to the other categories[1], *bouteille* (M=0.24) displays significantly lower categorization proportions than *fles* (M=0.38) ($t(39) = -6.213$, $p < 0.001$), and *pot* (F) (M=0.33) does not display a significant difference with *pot* (D) (M=0.35) ($t(39) = -1.354$, $p = 0.1837$). This indicates a difference in degree for the categories *fles-bouteille*, but not for *pot-pot*. The correlations between *fles-bouteille* (0.83) and *pot-pot* (0.74) suggest that both category pairs show a difference in criteria.

[1] Due to space limitations only the graphs for the categories *doos* and *boîte* are displayed. Similar graphs were made for the categories *fles–bouteille* and *pot–pot*. These figures were similar to those presented and lead to the same conclusions. A discussion of the results using *doos–boîte* was preferred since this category yields the same number of latent groups in Dutch and French and thus allows for a straightforward comparison.

Within-language differences

To study within-language differences we identified latent groups of categorizers using the mixture IRT-approach. Table 1 shows the BIC values for every category, for partitionings[2] in one to five groups. The solution with the lowest BIC (indicating the appropriate number of subgroups to consider) is in bold typeface.

Table 1: BIC values for five partitions of the categorization data with the number above each column representing the number of groups.

category	1	2	3	4	5
boîte	11256	10870	**10695**	10917	11159
doos	16838	16142	**16000**	16213	16221
pot (F)	10121	**10097**	10290	10530	10772
pot (D)	14170	**13788**	14040	14296	14549
bouteille	9015	**8645**	8859	9088	9323
fles	15650	14844	**14802**	15054	15310

Table 2 presents the total number of participants per category and also the total number of participants per latent group. Since the latent groups were determined by the mixture model, the participants are not necessarily evenly divided over the identified groups corresponding to the partitionings with the lowest BIC value.

Table 2: Overview of the number of respondents per group.

category	**all**	Group 1	Group 2	Group 3
boîte	**322**	60	71	191
doos	**448**	77	167	204
pot (F)	**310**	27	283	/
pot (D)	**424**	192	232	/
bouteille	**308**	110	198	/
fles	**436**	47	180	209

Figure 2 and Figure 3 display the categorization proportions in the three latent subgroups for the French category *boîte* and the Dutch category *doos*. The amount of variability within each language is striking. Even within a language, there appears to be only limited consensus with respect to categorization decisions. While one may expect some disagreement in the middle part of the curve for the borderline objects, there is also considerable disagreement at the ends of the curve, which should hold the clear members and clear non-members of the category for which one would expect to find strong agreement among speakers of the same mother tongue.

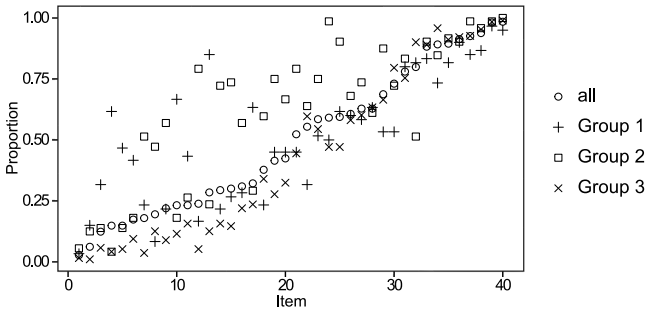


Figure 2: Categorization proportions for the French participants for the category *boîte* per item and per latent group with *all* referring to the average categorization proportion of the entire sample of French participants.

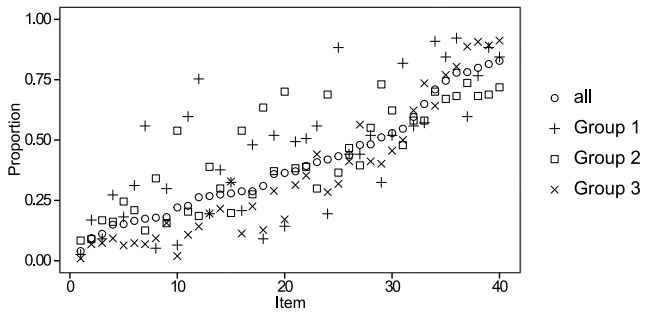


Figure 3: Categorization proportions for the Dutch participants for the category *doos* per item and per latent group with *all* referring to the average categorization proportion of the entire sample of Dutch participants.

This finding is not the result of the mixture analysis yielding smaller groups of categorizers who show more variability in their categorization data. The reliabilities[3] shown in Table 3 refute the possibility that the variability displayed in the graphs is due to unreliability. All reliabilities exceed 0.95 demonstrating that the participants in a subgroup performed in much the same way.

Table 3: Reliabilities per latent group of categorizers and for the complete group of participants per category.

category	**all**	Group 1	Group 2	Group 3
boîte	**0.995**	0.961	0.981	0.995
doos	**0.993**	0.975	0.981	0.992
pot (F)	**0.993**	0.956	0.999	/
pot (D)	**0.995**	0.991	0.993	/
bouteille	**0.992**	0.983	0.990	/
fles	**0.995**	0.968	0.991	0.992

[2] For the categories *fles* and *pot* (D) a group consisting of, respectively, only one and ten participants was identified. In those cases the analyses were repeated without these participants, resulting in the numbers shown in Table 2.

[3] Reliability was evaluated by applying the split-half method, followed by the Spearman–Brown correction. The displayed reliability is the average reliability across 10,000 random splits.

If the mixture analyses succeeded in identifying homogeneous groups of categorizers, the reliability within each group of latent categorizers should be higher than the reliability for the language group as a whole (if the groups are equated for number of participants). To evaluate whether this is true, a sampling procedure was used. Table 4 displays the average split-halves reliability across 10,000 random splits of the data of 25 randomly drawn participants from either the complete language group (all) or one of the latent groups within a language.

With the exception of two groups (*doos* group 2 and *boîte* group 1) the data indeed show the pattern we expected, showing that the variability we observe in the graphs is not random variation but reflects meaningful between-group-differences within a language. This is the case for the latent groups within both the Dutch and French language groups.

Table 4: Average split-half reliability across 10 000 random samples of 25 participants.

category	**all**	Group 1	Group 2	Group 3
boîte	**0.940**	0.911	0.946	0.963
doos	**0.892**	0.926	0.883	0.938
pot (F)	**0.920**	0.952	0.924	/
pot (D)	**0.924**	0.934	0.938	/
bouteille	**0.908**	0.930	0.922	/
fles	**0.917**	0.940	0.936	0.939

The sampling procedure was repeated for the two language groups together, that is, drawing 10,000 random samples of 25 participants out of the complete set for both languages together. The split-half reliability is 0.904 for all category pairs (*doos-boîte*, *fles-bouteille*, and *pot-pot*). One would expect that adding another language group to the dataset adds considerable variability to the data. Therefore, reliability should show a notable decrease in comparison with the reliability calculated within a language group (Table 4, first column). However, the reliability calculated over language groups is only slightly lower compared to the reliability calculated within one language group. This finding suggests that the within-language variability is comparable to the cross-linguistic variation.

Describing the identified variability between latent groups of categorizers in terms of differences in criteria and degree can be done both within and between languages. The strength of the correlations between categorization proportions can be interpreted as the extent to which different criteria are used. Differences in means between categorization proportions reflect a difference in degree.

The correlations in Table 5 vary from 0.20 to 0.86. Within-language correlations are displayed in the light gray area, and cross-language correlations are displayed in the dark gray area. The mixture analyses indeed succeeded in separating maximally different groups within one language, since the highest correlation between the categorization proportions of two groups of the same language is 0.77. It is quite striking that the correlations found for latent groups within a language do not exceed the correlation between the language groups (0.78). It also becomes clear that there are groups of categorizers who show a higher correlation with latent groups of the other language compared to correlations within their language group. For example, the categorization proportions for the Dutch *doos* of Group 1 resemble the proportions for the French *boîte* of Group 2 more than they resemble the other groups of their own language. Table 5 is a clear demonstration of the complexity of within-language and cross-linguistic variation in categorization patterns.

Table 5: Correlations of the categorization proportions for Dutch (D) and French (F) participants per latent group for the category pair *doos-boîte*.

	D2	D3	F1	F2	F3
D1	0.369	0.748	0.417	0.856	0.596
D2		0.706	0.825	0.200	0.473
D3			0.717	0.707	0.818
F1				0.436	0.772
F2					0.757

Taking into account differences in degree makes the comparison even more complex, since the strength of the correlation is not related to whether or not there is a difference in means. For example, comparing the average categorization proportions of D1 with those of F1 results in $t(39) = -4.7616$, $p < 0.0001$, whereas the comparison D2 – F3 results in $t(39) = 0.3549$, $p = 0.7246$. Both show a clear difference in criteria, but only the former shows a significant difference in degree. Of the groups that show a better correspondence in the used criteria, comparing the average categorization proportions of D3 and F3 shows no degree difference ($t(39) = -0.1741$, $p = 0.8627$), whereas the comparison D1-F2 does ($t(39) = -5.843$, $p < 0.0001$). Similar observations can be made for the comparison of latent groups of the same language.

Drawing a straightforward conclusion regarding cross-linguistic differences becomes more complex if one takes into account that there are latent groups that show higher correspondence with latent groups of another language group than they do with latent groups of the same language. Averaging over latent groups will in this case distort the comparison on a cross-linguistic level. This applies for both differences in criteria and degree.

Conclusion and discussion

The purpose of this study was to evaluate cross-linguistic lexical categorization differences relative to the categorization differences that exist between latent groups within each language. When comparing the variability displayed in Figure 1 versus Figures 2 and 3, the degree of variability within a language is higher than expected. One would expect the variability within a language to be less pronounced in comparison to cross-linguistic differences. Non-linguistic appreciation of properties of domains seems

to be universal (at least for some domains, including the one studied here), but the relation of this non-linguistic understanding to linguistic categorization is complex. That is, linguistic categories do not map directly onto similarity clusters (Malt et al., 1999). These complex patterns of lexical variation for categories of everyday objects emerge not only between languages but within a language as well. Especially the latter differences seem to be more complex than earlier assumed. Vagueness in degree and criteria seem to cause complex patterns of lexical variation between latent groups of categorizers that resemble the patterns of lexical variation at a cross-linguistic level.

The amount of variability observed within one language poses a challenge for cross-linguistic research. It is common practice in cross-linguistic research not to take into account within-language differences and to average across all individuals within a language, provided the sample comes from a restricted geographic region, implying a shared dialect. This may lead to conclusions that do not hold for the latent groups a language might harbor. For example, based on Figure 1 one might believe that the difference between the categories *doos* and *boîte* mainly consists of a difference in degree. The correlation of 0.78 between the language groups is imperfect but points out that there is a substantial agreement between both language groups as well. However, taking into account the within-language differences, it becomes clear that this conclusion could vary a great deal depending on the combination of latent groups, since these correlations vary from 0.20 till 0.86.

Future research will pinpoint possible causes for the observed variation. A possible path involves relating personal characteristics (age, gender, education level) and item characteristics to the parameter estimates of the different latent groups.

How are we able to manage the considerable inter-individual differences during the communication process and prevent a breakdown in communication? Possible answers to this question may lie in the way polysemy or words with new meanings such as eponyms are dealt with during communication using processes of sense creation and selection (Clark & Gerrig, 1983; Foraker & Murphy, 2012). Even for common nouns, referring to familiar objects in their most literal sense, these processes are relevant in the context of inter-individual differences.

Acknowledgments

Anne White is a research assistant at the Research Foundation-Flanders (FWO-Vlaanderen). Steven Verheyen was funded by ANR project TriLogMean (ANR-14-CE30-0010).

References

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language, 53*, 60–80.

Black, M. (1937). Vagueness. An exercise in logical analysis. *Philosophy of Science, 4*, 427-455.

Clark, H. H., & Gerrig, R. J. (1983). Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior*, 22(5), 591-608.

Devos, F. (2003). Semantic vagueness and lexical polyvalence. *Studia Linguistica, 57*, 121-141.

Foraker, S., & Murphy, G. L. (2012). Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4), 407-425.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language, 34*, 686–708.

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*, 137–165.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, 31, 355-384.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhaltery, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Malt, B., Ameel, E., Imai, M., Gennari, S. P., Saji, N., & Majid, A. (2014). Human locomotion in languages: Constraints on moving and meaning. *Journal of Memory and Language, 74*, 107–123.

Malt, B., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language, 49*(1), 20–42.

Malt, B. C., Gennari, S., Imai, M., Ameel, E., Saji, N., & Majid, A. (2015). Where are the concepts? What words can and can't reveal. In E. Margolis & S. Laurence (Eds.), *The conceptual Mind: New directions in the study of concepts*. Cambridge, MA: MIT Press.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40*, 230-262.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition, 6*, 462-472.

Munnich, E., Landau, B., & Dosher, B. A. (2001). Spatial language and spatial representation: a cross-linguistic comparison. *Cognition, 81*, 171-208.

Roberson, D., Davies, I. R. L., Corbett, G. G., & Vandervyver, M. (2005). Free-Sorting of colors across cultures: Are there universal grounds for grouping? *Journal of Cognition and Culture, 5*, 349 – 386.

Verheyen, S., Hampton, J. A., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the Rasch model. *Acta Psychologica, 135*, 216-225.

Verheyen, S., & Storms, G. (2013). A Mixture approach to vagueness and ambiguity. *PLoS ONE, 8*.

Verheyen, S., Voorspoels, W., & Storms, G. (2015). Inferring choice criteria with mixture IRT models: A demonstration using ad hoc and goal-derived categories. *Judgment and Decision Making, 10*, 97-114.

Event participants and linguistic arguments

Roxana-Maria Barbu (RoxanaMariaBarbu@cmail.carleton.ca)
Institute of Cognitive Science, 1125 Colonel By Drive
Ottawa, ON K1S 5B6 Canada

Ida Toivonen (Ida.Toivonen@carleton.ca)
Institute of Cognitive Science & School of Linguistics and Language Studies, 1125 Colonel By Drive
Ottawa, ON K1S 5B6 Canada

Abstract

Although there is a clear and intuitive mapping between linguistic arguments of verbs and event participants, the mapping is not perfect. We review the linguistic evidence that indicates that the mapping is imperfect. We also present the results of a new experimental study that provides further support for a dissociation between event participants and linguistic arguments. The study consists of two tasks. The first task elicited intuitions on conceptual event participants, and the second task elicited intuitions on linguistic arguments in instrument verbs and transaction verbs. The results suggest that while instrument phrases and currency/price phrases are considered necessary event participants, they are not linguistic arguments.

Keywords: event participants; linguistic arguments; syntax; semantics; psycholinguistics

Introduction

The relation between events and event participants is expressed in language as the relation between verbs and their arguments (Rissman, Rawlins, & Landau, 2015). This paper discusses the mapping between event participants and linguistic arguments. We argue that certain phrases (e.g., instrument and currency/price phrases) are difficult to classify as arguments or adjuncts because they are conceptualized as event participants but are not syntactically required.

When a verb is used in a sentence, that verb is typically accompanied by words or phrases that express who or what was engaged in or affected by the activity denoted by the verb. Consider, for example, the verb *praise* as it is used in (1):

(1) Lisa praised the child last evening.

In (1), *Lisa* and *the child* are directly involved in the *praising* activity, and those noun phrases are referred to as *arguments*. The noun phrase *last evening* is not directly involved in the activity denoted by the verb, and it is referred to as an *adjunct*.

The number and type of a verb's arguments partly depends on the nature of the event that the verb denotes. For example, a *running* or *strolling* event involves fewer participants than a *throwing* or a *putting* event.[1] Similarly, a basic motion verb such as *run* or *stroll* is likely to have fewer arguments than a caused motion verb like *throw* or *put*. The arguments of a motion verb refer to someone moving (*Karim* in (2)) and perhaps also a goal or a path (*to the store*). The arguments of a caused motion verb refer to whoever (or whatever) caused the motion (*Linda* in (3)), the entity moved (*the potatoes*) and perhaps the goal or path (*in the bucket*).

(2) Karim ran to the store.

(3) Linda threw the potatoes in the bucket.

Despite the clear and intuitive connection between arguments and event participants, the two notions are independent: there is not always a perfect mapping between linguistic arguments and event participants. We will provide a short review of work from the theoretical linguistics literature that illustrates this point. We then present the results from a new psycholinguistic study with two tasks. One task elicited intuitions about how many and what type of people or things are conceptually needed for different events to take place. The other task gauged how many and what type of people or things are syntactically needed in order for a sentence to be considered linguistically complete. The two tasks yielded different results, and the study thus provides experimental support for a separation between the concept of event participants and the concept of linguistic arguments, assuming that linguistic arguments are syntactically needed for the sentence to be complete (we revisit this assumption later). Before we turn to our experiment, we discuss the difference between arguments and adjuncts, and we also present linguistic evidence that verbal arguments are different from event participants.

Linguistic arguments

Tallerman (2005, 98) defines linguistic arguments as phrases that are selected by the head (often a verb) and that have an especially close relationship with the head. This definition is similar to other characterizations in the literature, but there is no universally agreed upon definition of argumenthood. However, scholars of language generally agree that verbs come with restrictions concerning their argument structure. This assumption is shared among theoretical linguists, descriptive linguists, psycho- and neurolinguists, computational linguists, and philosophers of language. By argument structure, we mean a set of elements linguistically required by the verb. There is debate on exactly how to determine what the argument structures of different verbs look like, and what information needs to be included in the structures.

Argument structure has been given distinct formal treatments within different strains of formalist/generative frameworks. Examples include SUBCAT lists in Head-Driven

[1] By participant, we do not necessarily mean a person. Here, we use the word in a general sense. For our present purposes, a participant can be a person or a thing, or perhaps a place or a time.

Phrase Structure Grammar, argument lists in Lexical-Functional Grammar, and theta grids in Principles and Parameters Theory. Yet other theoretical conceptualizations of argument structures are presented in Valency Grammar, Dependency Grammar, and Role and Reference Grammar. We will not make an attempt to review the vast literature where these topics are treated, but Levin and Rappaport Hovav (2005) is a good starting point. We turn instead to the distinction between arguments and adjuncts.

Consider again example (1), *Lisa praised the child last evening*, which includes two arguments (*Lisa*, *the child*) and one adjunct (*last evening*). Unlike arguments, adjuncts are not selected by the head; they are modifiers and added to clauses more freely. Example (1) is relatively straightforward, but it is sometimes difficult to determine whether a phrase is an argument or an adjunct.[2] For example, there is evidence that *to the park* is an argument in *Sandy ran to the park*, but *in the park* is not an argument in *Sandy ran in the park* (Van Luven, 2014; Kearns, 2011, 39), although this is not immediately obvious. Another tricky case is the passive *by*-phrase (e.g., *by Tara* in *the book was read by Tara*).

There is a battery of syntactic tests sensitive to whether a given phrase is an argument or an adjunct. The tests are useful, but do not always give consistent results (Christie, 2015; Needham & Toivonen, 2011). The two most commonly used diagnostics are conceptual necessity and syntactic obligatoriness. Phrases that are conceptually necessary in order for the event to take place are arguments. The syntactic counterpart to the conceptual necessity requirement states that arguments are obligatorily present in the linguistic string. We will argue below that there is an important distinction between event participants and linguistic arguments, and the two tests, conceptual necessity and syntactic obligatoriness, actually target different notions. However, the two notions often align, and when they do, they successfully differentiate between arguments and adjuncts. For example, the conceptual necessity and syntactic obligatoriness tests work well for (1): in order for a *praising* event to take place, it seems necessary that a praiser and something praised be present. It does not seem necessary to know about the point in time when the praising took place. Of course, most events take place at some time, but *praise* is not a verb that directly concerns time (by contrast, the verb *last* requires a temporal complement: *the meeting lasted one hour*). The time phrase is also not syntactically obligatory (overly expressed) in conjunction with *praise*: *last evening* can be omitted from (1).

What is conceptually necessary is not always expressed. For example, in order to *yell*, there has to be something that you yell, but it is nevertheless fully grammatical to say *Joe was yelling* without mentioning *what* he was yelling. The two tests thus do not always align. Jackendoff (2002, 138–149) argues that the conceptual necessity test is semantic and the obligatoriness test is syntactic. We take a slightly different position: conceptual necessity concerns how we understand different types of events and is not a linguistic notion.

Syntactic obligatoriness is, however, a linguistic notion, but it is not foolproof. Adjuncts are not obligatory (though see Goldberg & Ackerman, 2001 for some interesting potential examples of obligatory adjuncts). Arguments may be optional, but if a complement is obligatory, then it is an argument. The experiment we report in this paper taps into the different intuitions behind these two tests. We focus specifically on phrases that have been identified as difficult to classify with respect to argumenthood. We are not directly concerned with optional direct objects of verbs like *eat* and *wash*.

There are several other syntactic diagnostics that distinguish arguments from adjuncts (at least in English). These tests involve manipulating the structure of the sentence in ways sensitive to whether phrases are arguments or adjuncts. For example, *pseudo-clefting* is possible with adjunct but not argument targets (Hedberg & DeArmond, 2009):

(4) What Lisa did last evening was praise the child.

(5) *What Lisa did the child was praise last evening.

According to this diagnostic, if a phrase can occur after *do* in a pseudo-cleft, then that phrase is an adjunct. If the occurrence of a given phrase after *do* renders the phrase ungrammatical, then that phrase is an argument. Similarly, adjuncts but not arguments can appear after *do so* in VP anaphora:

(6) Lisa praised the child last evening and Bob did so today.

(7) *Lisa praised the child and Bob did so the teenager.

The exact explanations for how these tests work vary depending on theoretical framework. However, the VP-anaphora test can be understood quite straightforwardly with minimal appeal to technical machinery: *do so* is an anaphor that refers back to a verb and its arguments (or a minimal VP). In (7), *do so* therefore refers back to *praise the child* and the sentence is ungrammatical for the same reason as **Bob praised the child the teenager*: arguments cannot be repeated.

The difference between linguistic arguments and event participants

The examples in (2–3) above illustrate that the type of event a verb denotes can be a predictor of how many and what type of argument(s) a verb takes. Basically, an event seems to involve (perhaps entail) certain participants, and the number and type of participants depend on the nature of the event. However, the mapping between participants and linguistic arguments is not deterministic, and different verbs denoting similar events do not necessarily take the same argument structure. We will illustrate this point with a few oft-cited examples.

The first example has to do with verbs used to describe the ingestion of food. Examples include *eat, devour, dine, munch, gobble*, and *nibble*. These verbs describe different types of

[2]The blurry line between arguments and adjuncts has been discussed by Forker (2014); Hedberg and DeArmond (2009); Tutunjian and Boland (2008); Dowty (2003); Koenig, Mauner, and Bienvenue (2003); Croft (2001); Schütze and Gibson (1999); Whaley (1993); Grimshaw (1990); and many others.

eating events, and seem to involve an eater and something that is eaten. All the verbs take a subject (the eater), but they differ in how the complement (what is eaten) is linguistically expressed. The verb *eat* takes an optional object (8) (as indicated by the parentheses), the verb *devour* takes an obligatory object (9–10), and *dine* cannot take a direct object (11–12).

(8) Mike ate (an apple).

(9) Sally devoured the pizza.

(10) *Sally devoured.

(11) Jasleen and Sam were dining.

(12) *Jasleen and Sam were dining a great meal.

The verbs *munch, nibble* and *gobble* can take a complement, but not as a plain noun phrase object directly after the verb. The preposition *on* is required for *munch* and *nibble*:

(13) Fiona munched/nibbled on some potato chips.

(14) *Fiona munched/nibbled some potato chips.

Similarly, the particle *up* is required for *gobble*. Together, the examples in (8–14) illustrate that classifying a verb as a verb of ingestion might tell you that the event described by the verb involves two participants (an eater and something eaten). However, it does not tell you the exact number or types of arguments the verb takes.

Another example concerns pairs of verbs that describe the same event from different perspectives: *flee* and *chase*, *win* and *beat*, *borrow* and *lend*, *receive* and *give*, etc. These pairs describe the same general events: a chasing event, a winning event, a lending event, or a giving event (see, e.g., Gleitman, 1990). However, the verbs that describe the same event differ in their argument structure. Compare *flee* and *chase*:

(15) The hare fled (from the fox).

(16) The fox chased the hare.

The verb *flee* requires the individual being chased (or expelled) to be expressed, but the individual or thing the subject is fleeing from is only optionally expressed as a prepositional phrase. The verb *chase* requires both the chaser and the fleer to be expressed, and both are expressed as noun phrases. In this case, the individual who is fleeing is a direct object and not a subject. We conclude, as we did with the ingestion verbs, that knowing what type of event the verb refers to is not enough to predict whether or how those participants are realized as linguistic arguments.

This is not to say the realization of arguments is completely arbitrary. Many interesting generalizations have been proposed connecting specific semantic characteristics to the syntactic realization of arguments. For example, verbal aspect can be signaled in the realization of arguments (see, e. g., the papers in Demonte & McNally, 2012; Levin & Rappaport Hovav, 2005 and Tenny, 1994). However, it is clear from that literature that only some distinctions are *linguistically* relevant: knowing what participants are involved in an event is not enough to pinpoint the precise characteristics of linguistic arguments. Examples of relevant work include Pylkkänen (2008); Grimshaw (2005); Levin and Rappaport Hovav (2005); Wechsler (1995); Tenny (1994); Marantz (1993); Bresnan and Moshi (1990); Bresnan and Zaenen (1990); Grimshaw (1990); and Jackendoff (1990, 1983).

Experiment

The aim of the study was to explore speakers' intuitions about event participants and linguistic arguments. On the one hand, we wished to explore how many and what participants were considered *conceptually* necessary in order for events of different types to take place. On the other hand, we wanted to gauge what participants were considered *linguistically* necessary in conjunction with the verbs corresponding to the different events. Our experiment is similar to those of Rissman et al. (2015) and Koenig et al. (2003), as they also explore the status of instruments by eliciting speakers' intuitions.

Our studies differed in methodology: Rissman et al. (2015) explored speakers' semantic judgements; however, they instructed speakers with respect to what arguments are. Koenig et al. (2003) instructed speakers to add something that makes sense and is grammatical to the end of a phrase. Koenig et al. (2003), Rissman et al. (2015), and our study all yielded similar results.

Methods

Seventy-nine students with no background in Linguistics were recruited from Carleton University, using the school's psychology recruitment system. As compensation for participating in the experiment, students received 0.75% class credit towards a psychology or neuroscience course. All students were English speakers between the ages of 18 and 24.

The experiment consisted of two tasks. In task 1, speakers were instructed to think about an event such as *writing* and to state the participants required for the event to occur. Specifically, speakers were instructed to imagine a *verb*-ing event taking place in a large box and were asked to answer the following question: What needs to be in the box (people, objects, places etc.) for the *verb*-ing event to take place? Before beginning the experiment, speakers were provided with two sample events, *eating* and *dancing*. To accentuate the open-ended nature of the task, speakers were provided with a series of answers ranging in the number of participants such as *a person*; *a person* and *food*; *a person*, *food*, and *place* for an eating event.

In task 2, speakers were presented with the beginning of a sentence and asked to complete the sentence however it made most sense to them. Once again, they were provided with sample answers in order to illustrate the range of possibilities. Given the phrase *the woman ate*, possible ways to complete it include but are not limited to: not adding anything; *an ice cream; an ice cream with a spoon; an ice cream from Vito's Gelateria; an ice cream sitting on her balcony*.

This study focuses on two particular classes of verbs which we call *instrument verbs* and *transaction verbs*. Instrument verbs denote events that typically require some kind of tool or instrument. The verbs included in this task were *cut*, *draw*, *paint*, *scrape*, *scratch*, *scrub*, *sketch*, and *write*. It is not clear whether instruments are arguments or adjuncts of verbs such as *cut* or *draw* (see Rissman et al., 2015; Rissman, 2013; Koenig et al., 2003; and others). For example, in a sentence like (17), it is not clear whether *with her new knife* is an argument or an adjunct:

(17) Rita cut the bread with her new knife.

We wanted to test whether the unclear status of instruments might be due to a misalignment between conceptual participants and linguistic participants. For example, perhaps instruments are conceptually necessary in order for the event to take place, but they are not necessary as syntactic arguments in the clause.

Transaction verbs describe activities where the possession of an object changes from person A to person B, and person A receives something in exchange. The transaction verbs included in the study were *sell*, *buy*, and *purchase*, and these verbs are exemplified in (18–19):

(18) Bo sold the bags (to Jay) (for $870).

(19) Jay bought/purchased the bags (from Bo) (for $870).

The goal and price phrases are usually considered optional for the verb *sell*. Similarly, the source and price are optionally mentioned in clauses headed by the verbs *buy* and *purchase* (Stamenov, 1997).

It is important to note that some verbs allow an instrument or a price as a subject or an object, and we did not include such verbs here. Examples include *the key opened the door* where *the key* is an instrument, and *this book costs $20* where *$20* is a price.

Speakers' responses were coded based on the expected results. For task 1, we predicted someone performing the action, something the action was performed on, an instrument, and new categories were added as needed (e.g., time, place, manner). Given a *scrubbing* event, answers coded as having one conceptual participant include *sponge*; *a brush*; *cleaning materials*. Answers coded as having two conceptual participants include *something used to scrub, and something to be scrubbed*; *dirty dish and soap*; *a person, a scrub*. Answers coded as having three conceptual participants include *scrubber, a mess, an item to scrub*; *surface, person, and sponge*; *person, something to be scrubbed, an object to scrub with like a sponge*.

Since the speakers were given the beginning of the sentence in task 2, one linguistic participant (the subject) was already provided. An answer was coded as having only one linguistic participant if the speaker did not add any linguistic material to the phrase provided (e.g., *Mike scrubbed*). Answers interpreted as having two linguistic participants typically include the object, such as *the floor*; *toilets*; *the window*. Answers interpreted as having three linguistic participants include *mud off the floor*; *the vomit off the classroom floor*. For more detail about the method, please see Barbu (2015).

Results

Instrument verbs We expected responses to tasks involving instrument verbs such as *scrub* to include some sort of tool or instrument such as *sponge* in addition to someone doing the scrubbing and something being scrubbed. We were interested in whether speakers would be more likely to mention instruments in one of the two tasks. Recall that task 1 was designed to gauge what participants are conceptually necessary for the event to take place, whereas task 2 specifically targeted what arguments should naturally be expressed in the linguistic string.

In task 1 speakers' responses averaged out to 1.97 conceptual participants per event. In task 2, speakers mentioned 2.24 linguistic participants on average, so the average is slightly higher than the average for conceptual participants provided in task 1. While the number of conceptual and linguistic participants did not vary greatly between the two tasks, the specific participants listed did. Interestingly, there was a striking difference between the two tasks. While instruments were mentioned for each event in task 1, this was not the case in task 2, where an instrument was mentioned for only one of the events, *drawing*, and only by 5% of the speakers. Table 1 illustrates the number of mentions of an instrument in task 1 (conceptual participants) and task 2 (linguistic participants). The responses are given per verb. For example, when presented with a *writing* event, 90% of the speakers mentioned an instrument (e.g., *pencil, pen*) in task 1, while no speakers mentioned one in task 2.

Table 1: Mentions (%) of instruments as conceptual participants (Task 1) and linguistic participants (Task 2)

Event	Conceptual	Linguistic
scratch	39%	0%
scrape	47%	0%
paint	80%	0%
sketch	85%	0%
write	90%	0%
scrub	95%	0%
cut	95%	0%
draw	100%	5%

In sum, while speakers often mentioned instruments as necessary in order for a particular activity (e.g., *scrubbing*) to take place, they did not provide instrumental phrases (e.g., *with a sponge*) as linguistic material to complete sentences headed by the verb *scrub*.

Transaction verbs For transaction events, expected responses include two people (someone selling, buying, or purchasing something to or from someone else), the transferred

item, and some sort of monetary compensation. Answers coded as having one conceptual participant include *money*; *products*; *credit cards*. Answers coded as having two conceptual participants include *something to buy, someone to buy it*; *shopaholics, items for sale*; *someone purchasing something*. Answers coded as having three conceptual participants include *a person, something to sell, someone to sell it to*; *items to sell, money, people*; *people, items, place*. Answers coded as having four conceptual participants include *objects to sell, seller, buyer, place*. In task 1, speakers mentioned on average 2.8 conceptual participants per event.

Recall that one linguistic participant (e.g., the car dealer) and a verb were already provided for task 2. Answers coded as having two linguistic participants include *a car*; *a broken car*; *12 cars*. Answers coded as having three linguistic participants include *20 cars throughout the day*; *10 Ferraris today*; *a car for more than he thought it was worth*. In task 2, speakers provided an average of 2.3 linguistic participants per event.

The number of conceptual and linguistic participants did not vary greatly between the two tasks, but the types of participants listed did. For example, there was a significant difference between the two tasks with respect to currency. While in task 1 at least some speakers mentioned some form of currency, almost none of the speakers did in task 2. For example, when asked to list the participants involved in a *buying* event, 47% of the speakers listed some form of currency in task 1, while none of the speakers listed currency in task 2. Table 2 summarizes the mentions of currency for tasks 1 and 2.

Table 2: Mentions (%) of currency as conceptual participant (Task 1) and linguistic participant (Task 2)

Event	Conceptual	Linguistic
purchase	83%	0%
buy	47%	0%
sell	21%	5%

Table 2 illustrates that currency was mentioned much more in task 1 than in task 2, and on our interpretation, this means that currency is a necessary event participant at a conceptual level, but is not a linguistic argument.

Discussion

A simple analysis based solely on the number of participants/arguments mentions suggests that the results of the two tasks of this study are quite similar. However, a more careful analysis reveals important differences between the two tasks, both for instrument verbs and transaction verbs. For instrument verbs, instruments were commonly mentioned in task 1, while being nearly absent in task 2. These results suggest that an instrument is perceived as being a necessary component of events such as *writing* and *scrubbing*, but the instrument phrase is nevertheless not a syntactic argument of the verbs used to describe those events. Similarly, transaction events are perceived as involving some form of money, yet phrases that refer to money are not syntactic arguments.

Despite the fact that at least some speakers mentioned an instrument for each of the verbs in this class, there is a difference between verbs. This is expected, as conceptual participants are strongly related to the meaning of the verb, and explanations for these differences might be found in the lexical semantics of the specific verbs used to elicit judgments. For example, all subjects listed some kind of instrument for a *drawing* event, but only 39% of subjects listed an instrument for a *scratching* event. This difference might be due to body parts not being conceptualized as instruments (see, e.g., Rissman et al., 2015; Carlson & Tanenhaus, 1989). When asked to think about a *scratching* event, it is quite likely that speakers thought of fingernails or claws. However, it is more difficult to think of *drawing* events where a body part is used as an instrument, although it is of course possible to draw with your finger on a steamed up window.

The syntax of instrument and currency phrases

Our results suggest that speakers view instruments for verbs like *scrub* and payments for verbs like *buy* as conceptual event participants, even though they are not linguistically necessary complements of the verbs. Recall that obligatory phrases are arguments, but arguments are not necessarily obligatory (e.g., the optional object of *eat*). Are instrument and currency phrases arguments, even though they are not obligatorily expressed? In order to answer this question, we would need a full linguistic investigation. Here, we will simply present a tentative answer based on the syntactic argumenthood tests introduced above, pseudo-clefting and VP-anaphora. Introspective judgements of the grammaticality of pseudo-cleft and VP-anaphora examples suggest that instrument and currency phrases are not syntactic arguments.

Example (20) shows a pseudo-clefted version of *Suzie scrubbed the floor with a brush* and is grammatical. The VP-anaphora example in (21) is also grammatical. Arguments cannot appear in these syntactic frames, so these examples suggest that instruments are adjuncts, assuming that *with a brush* in this example is representative of instruments.

(20) What Suzie did with the brush was scrub the floor.

(21) Suzie scrubbed the floor with a brush and Tom did so with a sponge.

Examples (22–23) target the currency phrase *for $500*. Example (22) is a pseudo-clefted version and (23) is a VP-anaphora version of *Tom bought a bike for $500*. The grammaticality of these examples indicates that currency/price phrases are adjuncts, not arguments.

(22) What Tom did for $500 was buy a bike.

(23) Tom bought a bike for $500 and Sue did so for $1000.

The syntactic diagnostics based on pseudo-clefting and VP-anaphora point to the same conclusion as the results of task

2 in our experiment: instrument and currency phrases are not linguistic arguments, even though they are conceptually necessary event participants.

Conclusion

This paper addressed the following question: Is there a distinction between event participants and linguistic arguments? We revisited arguments from the linguistics literature that have previously pointed to an imperfect mapping between the two notions. We also presented the results of a new study consisting of one task targeting conceptual event participants and another task targeting syntactic arguments. The results can be interpreted as capturing differences between conceptual and linguistic participants in instrument verbs and transaction verbs. Despite the open-ended nature of the tasks, there was agreement across speakers that while instruments and currency are conceptual participants, there is no need for them to be overtly stated. On our interpretation, the experimental data presented here suggests that while instruments and currency are conceptualized as event participants and potential semantic arguments, syntactically they behave as adjuncts. This conclusion is supported by the results of two traditional syntactic argumenthood tests.

References

Barbu, R.-M. (2015). *Verbs and participants: Non-linguists' intuitions*. Unpublished master's thesis, Carleton University.

Bresnan, J., & Moshi, L. (1990). Object asymmetries in comparative Bantu syntax. *Linguistic Inquiry*, *21*, 147–85.

Bresnan, J., & Zaenen, A. (1990). Deep unaccusativity in LFG. In K. Dziwirek, P. Farrell, & E. Mejias-Bikandi (Eds.), *Grammatical relations: A cross-theoretical perspective* (pp. 45–57). Stanford, CA: CSLI.

Carlson, G., & Tanenhaus, M. (1989). Thematic roles and language comprehension. In G. Carlson & M. Tanenhaus (Eds.), *Linguistic structure in language processing: Studies in theoretical psycholinguistics* (pp. 263–89). Dordrecht and Boston: Kluwer Academic Publishers.

Christie, E. (2015). *The English resultative*. Doctoral dissertation, Carleton University, Ottawa.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Demonte, V., & McNally, L. (2012). *Telicity, change, and state: A cross-categorial view of event structure*. Oxford: Oxford University Press.

Dowty, D. (2003). The dual analysis of adjuncts and complements in categorial grammar. In E. Lang, C. Maienborn, & C. Fabricius-Hansen (Eds.), *Modifying adjuncts.* Berlin: Mouton de Gruyter.

Forker, D. (2014). A canonical approach to the argument/adjunct distinction. *Linguistic Discovery*, *12*(2), 27–40.

Gleitman, L. (1990). The structural sources of verb meaning. *Cognition*, *1*(1), 3–55.

Goldberg, A., & Ackerman, F. (2001). The pragmatics of obligatory adjuncts. *Language*, *77*(4), 798–814.

Grimshaw, J. (1990). *Argument structure*. Cambridge, MA: MIT Press.

Grimshaw, J. (2005). *Words and structure*. Stanford: CSLI.

Hedberg, N., & DeArmond, R. C. (2009). On complements and adjuncts. *Snippets*, *19*, 11-12.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Kearns, K. (2011). *Semantics, 2nd edition*. New York: Palgrave Macmillan.

Koenig, J.-P., Mauner, G., & Bienvenue, B. (2003). Arguments for adjuncts. *Cognition*, *89*, 67–103.

Levin, B., & Rappaport Hovav, M. (2005). *Argument realization*. Cambridge: Cambridge University Press.

Marantz, A. (1993). Implications of asymmetries in double object constructions. In S. A. Mchombo (Ed.), *Theoretical aspects of Bantu grammar* (pp. 113–50). Stanford, CA: CSLI.

Needham, S., & Toivonen, I. (2011). Derived arguments. In M. Butt & T. H. King (Eds.), *Proceedings of the LFG11 conference* (pp. 401–421). Stanford, CA: CSLI.

Pylkkänen, L. (2008). *Introducing arguments*. Cambridge, MA: MIT Press.

Rissman, L. (2013). *Event participant representations and the instrumental role: A cross-linguistic study*. Doctoral dissertation, Johns Hopkins University, Baltimore.

Rissman, L., Rawlins, K., & Landau, B. (2015). Using instruments to understand argument structure for gradient representation. *Cognition*, *142*, 266–290.

Schütze, C., & Gibson, E. (1999). Argumenthood and english prepositional phrase attachment. *Journal of Memory and Language*, *40*, 409–432.

Stamenov, M. I. (1997). *Language structure, discourse and the access to consciousness* (Vol. 12). Amsterdam: John Benjamins.

Tallerman, M. (2005). *Understanding syntax, 2nd edition*. London: Hodder Arnold.

Tenny, C. L. (1994). *Aspectual roles and the syntax-semantics interface*. Dordrecht: Kluwer.

Tutunjian, D., & Boland, J. E. (2008). Do we need a distinction between arguments and adjuncts? Evidence from psycholinguistic studies of comprehension. *Language and Linguistic Compass*, *2*, 631–646.

Van Luven, K. (2014). *The argument status of directional PPs.* (B.A. thesis, Carleton University)

Wechsler, S. (1995). *The semantic basis of argument structure*. Stanford, CA: CSLI.

Whaley, L. (1993). *The status of obliques in linguistic theory*. Doctoral dissertation, State University of New York, Buffalo.

Decision-Making and Biases in Causal-Explanatory Reasoning

Samuel G. B. Johnson[1], Marianna Zhang[2], & Frank C. Keil[1]
(samuel.johnson@yale.edu, mariannaz@uchicago.edu, frank.keil@yale.edu)
[1]Department of Psychology, Yale University, 2 Hillhouse Ave., New Haven, CT 06520 USA
[2]Department of Psychology, University of Chicago, Chicago, IL 60637 USA

Abstract

Decisions often rely on judgments about the probabilities of various explanations. Recent research has uncovered a host of biases that afflict explanatory inference: Would these biases also translate into decision-making? We find that although people show biased inferences when making explanatory judgments in decision-relevant contexts (Exp. 1A), these biases are attenuated or eliminated when the choice context is highlighted by introducing an economic framing (price information; Exp. 1B–1D). However, biased inferences can be "locked in" to subsequent decisions when the judgment and decision are separated in time (Exp. 2). Together, these results suggest that decisions can be *more* rational than the corresponding judgments—leading to choices that are rational in the output of the decision process, yet irrational in their incoherence with judgments.

Keywords: Decision-making; causal reasoning; inductive reasoning; explanation; behavioral economics.

Introduction

Our decisions often depend on prior inferences. For instance, a patient decides on a treatment that matches the disease likeliest to ail her, based on diagnostic tests; an investment banker chooses an asset allocation expected to maximize profits, based on past returns; a consumer chooses the toothpaste likeliest to keep his teeth white, based on persuasive advertising.

As others have argued, such inference-based decisions are often *causal* (e.g., Sloman, 2005). The patient's treatment will *cause* her to recover; the investment banker's choice will *cause* profits to be maximized; the toothpaste will *cause* the consumer's teeth to be white. Much is known about how people make causal predictions and evaluate causal explanations (e.g., Rottman & Hastie, 2014; Sloman, 2005; Waldmann & Holyoak, 1992). In particular, recent work has triangulated a set of heuristics used in making diagnostic inferences, including causal explanations (Johnson, Rajeev-Kumar, & Keil, 2014, 2015a; Khemlani, Sussman, & Oppeheimer, 2011; Lombrozo, 2007). How do these mechanisms translate into choice behavior?

One possibility is that we use these mechanisms to arrive at judgments, and then translate those judgments into decisions. (This is roughly the view of classical decision theory; e.g., Jeffrey, 1965.) Although this pathway from judgment to decision is itself normative in preserving coherence, it can lead to errors in decisions to the extent that the judgments are themselves biased. Since judgments arrived at through diagnostic reasoning are subject to systematic biases (e.g., Khemlani et al., 2011; Lombrozo, 2007), one would expect decisions that depend on those judgments to also be biased.

Alternatively, decision-making may recruit additional mechanisms beyond judgment. In dual process terms, we could think of the intuitive judgments as relying on System 1, and the biases result because they are not corrected by System 2 (Kahneman, 2003). If decision-making recruits additional System 2 resources that are not available in judgment, then the decisions may be less biased, or even unbiased.

Some previous results are consistent with this more nuanced picture of judgment and decision. For example, in addition to judgments leading to our decisions, our decisions also seem to affect our judgments (Johnson, Rajeev-Kumar, & Keil, 2015b). When people choose a course of action that is more consistent with one diagnostic judgment rather than another, people tend to think that the corresponding judgment is more likely to be true—even if the reason for choosing the corresponding action is independent of the judgment (i.e., the stakes are higher for being wrong given the other choice). Likewise, people are more likely to search for information that confirms a decision (Fischer & Greitemeyer, 2010) and judge disconfirmatory evidence more harshly (Chaxel, Russo, & Kerimi, 2013). All of these findings point to a bidirectional relationship between judgment and decision.

Although it is well-known in behavioral economics circles that monetary incentives improve performance in decision-making contexts (e.g., Levitt & List, 2007), it is less clear whether the 'pseudoincentive' of a decision-making *task* (with the same monetary compensation as a judgment task) would be sufficient to induce System 2 monitoring.

We examine these issues in two sets of studies. First, we test whether a bias against explanations making unverified predictions propagates from judgment to decision (Exp. 1A and 1B), and test boundary conditions of these effects (Exp. 1C and 1D). Second, we look at individual differences in a context where *both* judgments and decisions are elicited from the same participants and are separated in time (Exp. 2).

Experiments 1A and 1B

When interpreting evidence to distinguish between hypotheses, people are unwilling to settle for ignorance (Khemlani et al., 2011; Sussman et al., 2014).

For example, suppose that you are hunting. There are two types of deer in the forest, one with white spots on its tail (species *W*) and another without spots (species *N*), which roam the forest in equal numbers. Species *W* has a

wide explanatory scope, because it can explain more potential features (i.e. white spots on tail) than species *N*, which has a *narrow* scope.

Suppose that, due to the policies of your local government, the deer have overlapping hunting seasons, but species *W* must be shot with a bow-and-arrow, whereas species *N* must be shot with a gun. Now, suppose you see a deer in the distance, but its tail is occluded by a tree. Do you shoot with a gun or with a bow-and-arrow?

In reasoning through problems like this, people attempt to *infer* whether this particular deer *would* have spots, if the tree were not in the way. Unfortunately, since the forest has equal numbers of *W* and *N* deer, this strategy is not helpful—it has exactly an equal chance of having spots (if it is *W*) or not (if it is *N*). Nonetheless, people do not settle for ignorance, and use the base rate of the diagnostic feature—in this case, the proportion of deer *in general* that have white spots on their tails—to guess whether this *particular* deer will have white spots (Johnson, Rajeev-Kumar, & Keil, 2015a). Since most deer do not have white spots on their tails, people erroneously infer that this deer is also unlikely to have white spots, and will conclude it is more likely to belong to species *N*. Because most effects and features are relatively uncommon in general, people generally are averse to explanations with a wide scope of unverified predictions (see also Johnson, Johnston, Koven, & Keil, 2015, for evidence of this bias in 4-year-old children).

We make inferences in large part so that we can make choices in the world. In this case, the inference ought to influence whether you use a gun or a bow-and-arrow to shoot the deer. More generally, we often make decisions in economic contexts which depend on explanatory inference where evidence is unavailable. Would a bias against wide scope explanations, making unverified predictions, also arise in decision-relevant contexts?

To test this, participants in Exp. 1 read about situations where two different explanations (one wide and one narrow) had different choice implications. For example, suppose you've been having problems with your robotic lawnmower—it has been running into trees and making strange noise. There are two possible problems that could lead to this behavior—it could be a faulty hesolite axle (which makes no other predictions) or a faulty transduction spindle (which also makes the prediction that the spindle should remain cool during use). However, because safety precautions make it impossible to lift the lawnmower's lid, you cannot check whether or not the spindle is cool.

Thus, based on previous research, we would predict that participants should favor the narrow explanation that did *not* make the unverified prediction, even if the two explanations actually have equal posterior probabilities, given the information in the problem. Exp. 1 tested this prediction in two different ways. Some participants were asked to make an explicit causal/explanatory judgment, identifying which part was the most likely cause of the problem (Exp. 1A). Previous research suggests that participants should favor the narrow explanation here. Other participants were asked to choose which replacement part they would buy (Exp. 1B). If their causal inferences translate directly into decisions, they should also favor the narrow explanation here. Conversely, if the decision-making context leads them to recruit System 2 resources that correct for bias, then they should be more likely to provide normative responses.

Method

We recruited 383 participants from Amazon Mechanical Turk (N = 186 for Exp. 1A, N = 197 for Exp. 1B); 48 were excluded from analysis due to poor performance on check questions (see below).

Participants in both experiments completed 5 items (concerning robotic lawnmowers, pest control, junk mail, television repair, and household detergents) in a random order. For example, in Exp. 1A, the lawnmower item read:

> Imagine your autonomous robotic lawnmower hasn't been working. It's definitely a problem with either the transduction spindle or the hesolite axle. These two problems occur equally often.
>
> A faulty hesolite axle causes disorientation and makes noise.
>
> A faulty transduction spindle causes disorientation, makes noise, and stays cool during use.
>
> Your lawnmower has been running into trees and making strange noise, but you can't tell whether the transduction spindle stays cool during use because the lawnmower's lid cannot be opened during use as a safety precaution.

That is, the *narrow* explanation (*faulty hesolite axle*) makes two confirmed predictions (*disorientation* and *noise*). The other *wide* explanation (*faulty transduction spindle*) makes the same two confirmed predictions, plus one latent or unverified prediction (*stays cool*). The order of the wide and narrow explanations was counterbalanced for each participant.

In Exp. 1A, participants answered a *cause* question, reporting which causal explanation they favored (e.g., "Which part do you think caused the problem?"), on a scale from 0 ("Definitely transduction spindle") to 10 ("Definitely hesolite axle").

The items in Exp. 1B were the same, except they were also given some additional information about the decision-making context, focusing on what interventions they could make to solve the problem. For the lawnmower example, participants read:

> To fix it, you must replace one of the parts and check if the lawnmower is fixed. You can buy a new transduction spindle for **$40** or a new hesolite axle for **$40**.

They then answered a *choice* question, reporting which choice they would make (e.g., "Which part would you buy?") on a scale from 0 ("Definitely buy transduction

spindle") to 10 ("Definitely buy hesolite axle"). For both Exp. 1A and 1B, the left/right orientation of the scales was adjusted to match the order in which the explanations were listed, and the default setting on all scales was the midpoint.

At the end of each study, participants were asked to check off of a list of items that had appeared throughout the study, as an attention check. Participants incorrectly answering more than 30% of these questions were excluded from analysis.

Results

All measures were scaled so that negative scores correspond to *narrow* scope inferences or decisions, and so that positive scores correspond to *wide* scope inferences or decisions.

When participants were asked to evaluate explanation in Exp. 1A, they had a significant bias toward the negative latent scope explanation [M = -0.25, SD = 0.84; $t(167)$ = -3.87, $p < .001$, d = -0.30]. This bias is consistent with previous work on causal explanation (e.g., Khemlani et al., 2011), where people tended to favor explanations that did not posit unverified predictions.

However, when participants were asked to choose between potential interventions based on explanations, in Exp. 1B, they no longer had any bias [M = 0.00, SD = 1.41; $t(166)$ = 0.02, p = .98, d = 0.00]. This led to a significant difference between Exp. 1A and 1B [$t(333)$ = 2.00, p = .047, d = 0.22]. See Table 1 for means and confidence intervals across Experiments 1A–D.

Exp.	DV	Prices	Mean	CI
1A	Cause	Yes	-0.25	(-0.38, -0.12)
1B	Choice	Yes	0.00	(-0.21, 0.22)
1C	Cause	No	-0.03	(-0.16, 0.11)
1D	Cause	Cheap	-0.14	(-0.31, 0.03)

Table 1: Results of Experiments 1A–D

Discussion

These results suggest a nuanced role for explanatory inference in decision-making. Exp. 1A demonstrated that a signature bias of explanatory reasoning—found previously in causal diagnosis (Khemlani et al., 2011), categorization (Sussman et al., 2014), stereotyping (Johnson, Kim, & Keil, 2016), and causal strength judgment (Johnson, Johnston, Toig, & Keil, 2014)—also appears in the kinds of causal reasoning problems that feed directly into decision-making.

However, somewhat surprisingly, this bias did not translate into biased decisions in Exp. 1B. Taken together, participants in these experiments indicated that N was a more likely cause than W, yet they were equally likely to intervene on N and W. These decisions at once violate and affirm the tenets of rationality: They *violate* rationality in the sense that individuals' decisions were inconsistent with their beliefs; yet, they *affirm* rationality in the sense that their decisions were unbiased. This unbiased decision, while inconsistent with their beliefs, is rational taken in isolation.

Something about making an inference-based decision, rather than a mere inference, appears to be pushing people toward more rational behavior. In dual process terms (Kahneman, 2003), one possibility is that explanatory heuristics produce System 1 responses which can be overridden by System 2 monitoring. Perhaps the stakes of decision-making invoke more monitoring of intuitive judgments, leading to more normative responses. Exps. 1C and 1D explore implications of this account.

Experiment 1C

Exp. 1C aimed to pinpoint which difference between Exp. 1A and 1B drove the difference in outcomes. These studies differed in two ways: (1) They used different dependent measures and tasks (a causal diagnosis versus a choice); and (2) They invoked different judgment contexts (a reasoning context versus a choice context) in that Exp. 1B provided information about interventions to fix the problem, such as the prices of the options. Which of these factors led to the biased inferences in Exp. 1A but unbiased decisions in Exp. 1B?

On the one hand, it may be the *task* itself (causal diagnosis versus choice) that is crucial. On the assumption that decision-making invokes more System 2 monitoring than mere inference, it seems plausible that the nature of the question itself is driving the results: Forcing participants to appreciate the stakes of the problem by using a decision process may lead them to more normative responses.

Alternatively, the mere *context* of making an economic decision could suffice to raise the stakes. The contextual information supplied in Exp. 1B indicated that the judgment *implied* a course of action, and perhaps that implication is sufficient even in the absence of an overt decision.

In Exp. 1C, we distinguished between these factors by using the same dependent measure as Exp. 1A (a causal diagnosis) but including the contextual information from Exp. 1B, to establish the decision-making context. If the task itself led to more rational judgment, then we would expect biased judgment in Exp. 1C (as in Exp. 1A); but if the choice context is sufficient to invoke rational judgment, then we would expect unbiased judgment (as in Exp. 1B).

Method

We recruited 206 participants from Amazon Mechanical Turk; 21 were excluded from analysis due to poor performance on check questions.

The procedure was identical to Exp. 1B, including the same paragraph of contextual information ("To fix it…"; see Exp. 1B methods). However, the dependent measure was the same causal question used in Exp. 1A ("Which part do you think caused the problem?").

Results and Discussion

Participants' judgments were normative, even though these judgments were causal inferences rather than choices. That is, participants' judgments were unbiased [M = -0.03, SD = 0.96; $t(184)$ = -0.37, p = .71, d = -0.03]. Correspondingly, the causal judgments in Exp. 1C (with the choice context) differed significantly from the causal judgments in Exp. 1A (without the choice context) [$t(351)$ = 2.34, p = .020, d = 0.25] but not from the choices in Exp. 1B [$t(350)$ = -0.22, p = .82, d = -0.02]. See Table 1.

These results suggest that a judgment that *implies* a decision is sufficient to induce System 2 monitoring, just as much as a decision itself. Exp. 1D further probes the boundary conditions of this normative choice behavior.

Experiment 1D

What is it about a choice context that induces System 2 monitoring? It could be that having to make a decision, regardless of the stakes, is sufficient to induce monitoring. Alternatively, it could be that the importance of the choice could be the key factor, in which case the economic stakes of the choice should be critical.

Exp. 1D sought to tease apart these mechanisms by introducing "dirt cheap" prices. If *any* choice is sufficient to induce monitoring, regardless of the stakes, then we should expect unbiased inferences. Conversely, if it is the stakes themselves that are critical, then we should expect the bias to return when they are minimized.

Method

We recruited 198 participants from Amazon Mechanical Turk; 18 were excluded from analysis due to poor performance on check questions.

The procedure was identical to Exp. 1C, except the prices were lowered to "dirt cheap" levels. For example:

> To fix it, you must replace one of the parts and check if the lawnmower is fixed. From the local junkyard, you can buy a replacement transduction spindle for **$0.75** or a replacement hesolite axle for **$0.75**.

The methods were otherwise identical to Exp. 1B.

Results and Discussion

The results were mixed. On the one hand, participants' causal judgments were somewhat non-normative, leading to a marginally significant bias [M = -0.14, SD = 1.15; $t(179)$ = -1.66, p = .100, d = -0.12]. However, despite the significant difference between Exp. 1A and 1C, the current bias did not significantly differ from either experiment [$t(346)$ = 1.00, p = .32, d = 0.11 and $t(363)$ = -1.05, p = .29, d = -0.11, respectively]. See Table 1 for means and comparisons across experiments.

These results are not conclusive, but they are suggestive. The marginally significant bias seems to suggest that extremely low stakes allow for some degree of System 1 bias that is uncorrected by System 2 monitoring. However, the results falling midway between Exp. 1A and 1C (albeit not significantly differing from either) suggests that both mechanisms may be at play in bias reduction: The stakes appear to play a role, but the mere act of implying a choice also appears to play a role.

Experiment 2

We have been describing the theoretical picture supported by these results in dual process terms—that people make intuitive judgments which are then corrected by more explicit reasoning when making decisions. A more radical view of these results is that causal-explanatory reasoning is simply not a force in decision-making, or that decision-making relies on separate reasoning processes, as opposed to the heuristics known to be used in explanatory reasoning (e.g., Johnson, Rajeev-Kumar, & Keil, 2014, 2015a; Lombrozo, 2007). Could this view be right?

Exp. 2 capitalized on the fact that reasoners do not make uniform judgments in the face of explanations varying in scope—indeed, Exp. 1A revealed considerable variability in judgments (SD = 0.84) despite the mean favoring the narrow explanation. That is, participants varied greatly in the magnitude and even direction of their bias (see Johnson, Rajeev-Kumar, & Keil, 2014, 2015a for discussion of the mechanisms underlying this bias, which can lead to biases in either direction, depending systematically on individuals' prior beliefs).

In Exp. 2, participants were asked to make a judgment (as in Exp. 1A) followed by a choice (as in Exp. 1B). If the unbiased choices in Exp. 1B occurred because people are relying on a different computational system for choice that circumvents diagnostic judgment heuristics, then individuals who make biased diagnostic inferences in judgment would be unlikely to make the same biased inferences in choice, or should at least be far less biased. Conversely, if the reasoning mechanisms are the same, then once locked into a judgment, a participant would likely make a choice that matches that judgment.

Method

We recruited 299 participants from Amazon Mechanical Turk; 1 was excluded from analysis due to poor performance on check questions.

The procedure combined the dependent measures of Exp. 1A and 1B, in a within-subjects design. Participants were randomly assigned to one of the five vignettes, and completed both the *cause* question (from Exp. 1A) and the *choice* question (from Exp. 1B), in that order.[1] The procedure was otherwise identical to the other experiments.

Results and Discussion

Among the 113 participants who favored the narrow

[1] The reverse order was not used because this order does not test our hypothesis—if participants made the choice first, then the congruence between choice and inference would be explained by our earlier results showing that economic contexts lead to debiasing.

explanation in responding to the cause question [M = -2.36, SD = 1.51], these participants also tended to *choose* the option corresponding to that diagnosis [M = -2.09, SD = 2.11]. Likewise, among the 101 participants who favored the wide explanation in responding to the cause question [M = 2.13, SD = 1.42], these participants also tended to choose the option corresponding to that diagnosis [M = 1.91, SD = 1.91]. In fact, these choices were just as strong as their initial diagnoses [$t(112)$ = 1.56, p = .12 and $t(100)$ = 1.33, p = .19, respectively], indicating little evidence for less biased decisions than judgments, even though regression toward the mean would push judgments toward less bias.

We draw two conclusions from these results. First, even though decision-making leads to error-correction when made in the absence of an explicit judgment, errors can be "locked in" by first making an explicit judgment. That is, participants were no less biased in making decisions than they were in making judgments in this task, where their decisions followed explicit judgments. Second, despite the unbiased choices in previous studies, these results suggest a strong relationship between diagnostic causal reasoning and subsequent decisions that depend on those causal judgments: Analyses of individual participants revealed that those whose causal judgments were biased in one direction tended to likewise make decisions that were biased (just as much) in the same direction. Although choices were unbiased at the aggregate level in previous studies, likely due to adjustments caused by System 2 error-correction, choices are nonetheless strongly associated with their antecedent causal judgments.

This study is subject to the limitation that choices were made immediately after judgments on a very similar scale, which may lead to anchoring and other scale-use issues. Hence, future work should correct this problem by using less alignable scales, or by using an intermittent task to reduce carry-over effects. Nonetheless, the finding that there was no significant regression toward the mean between tasks suggests that anchoring-and-adjustment cannot be a complete explanation for the current results: There could indeed have been anchoring, but there was little or no adjustment.

General Discussion

Decisions are often predicated upon causal judgments. Yet, the heuristic mechanisms underlying causal judgments often lead to biased inferences. Would these biases translate into decision-making?

At least in the case of the bias against explanations making unverified predictions (Khemlani et al., 2011), the answer appears to be 'no'. Although participants made biased judgments in choice-relevant inference problems (Exp. 1A), these biases were eliminated when making choices based on those inferences (Exp. 1B). These unbiased responses also carried over to causal judgments that were accompanied by information contextualizing the choice as an economic decision (such as prices; Exp. 1C). The bias appeared to return when the stakes of the choice were greatly lowered (Exp. 1D), although the bias was of a smaller magnitude than it had been when the choice context was omitted altogether.

These results together suggest that choice contexts can attenuate or eliminate diagnostic reasoning biases. This effect is most likely attributable in part to increases in System 2 monitoring when the choice context is made salient (regardless of stakes), and in part due to accentuated monitoring caused by higher stakes.

Nonetheless, these results do not undermine the claim that choices depend on diagnostic reasoning processes. Indeed, Exp. 2 asked participants to make both a judgment and a decision, and found that participants who made biased judgments were also likely to make biased decisions, in the same direction. This finding indicates that participants' decisions are based on antecedent judgments. In addition, in contrast to Exp. 1, the choices were *just as biased* as the judgments, suggesting that the act of making an inference can "lock in" the relevant decision, when the judgment and decision are separated in time.

These results contribute to debates concerning human rationality. On the one hand, our results affirm mainstream views in behavioral economics, which have a generally low opinion of human decision processes. This is true in two senses in the current work: First, inferences were biased in a decision-relevant context (Exp. 1A), and these biased judgments could be "locked in" to biased decisions when the judgment and decision were separated in time (Exp. 2). Second, when the decision was not preceded by an explicit judgment, the decision was inconsistent with its antecedent judgment, suggesting incoherence in the decision-making process, in violation of traditional normative models (e.g., Jeffrey, 1965).

Nonetheless, these results are hopeful in a different sense, and more friendly to classical views of human decision faculties. Economists are fond of critiquing lab experiments (including many behavioral economics studies) because they often fail to reflect the incentives present in the marketplace which can create pressure for more optimal behaviors (Levitt & List, 2007), especially at the aggregate level of institutions such as the stock market. Thus, the argument goes, suboptimal behavior in lab contexts can give way to more optimal behavior in economic contexts. Our results go a step further: Not only can market mechanisms potentially drive more rational behavior, but the psychological mechanisms underlying choice behavior appear to induce error-monitoring processes that can lead people to behave more rationally, even in the absence of economic incentives.

This work can be expanded upon in several ways. First, it should extend to other reasoning biases. For instance, people are biased in some contexts to favor overly simple explanations (Lombrozo, 2007), and in other contexts to favor overly complex explanations (Johnson, Jin, & Keil,

2014). Further, the size and direction of these biases can be influenced by normatively irrelevant factors. We have collected some preliminary data, suggesting that these biases are attenuated in choice contexts, just as scope biases were shown to be attenuated in the current results. Likewise, people tend to 'digitize' their beliefs, holding propositions to be either certainly true or certainly false, rather than coming in degrees (Johnson, Merchant, & Keil, 2015; Murphy & Ross, 1994). Perhaps people would be less likely to digitize in a choice context, leading to more normative (Bayesian) behavior.

Second, more work is needed to uncover the mechanisms underlying the bias attenuation. We have argued that System 2 error-monitoring is the most plausible explanation, but further work should pinpoint the mechanism and pinpoint the boundary conditions. Explicit responses (such as think-aloud protocols) would be one useful source of evidence, as would other manipulations designed to change the stakes and context (building on Exp. 1C and 1D). Such studies could help to pinpoint the conditions under which biases are and are not attenuated, with potential implications for real-world choice behavior and for debates on the limits of human rationality—as well as the limits on those limits.

References

Chaxel, A., Russo, J.E., & Kerimi, N. (2013). Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making*, *8*, 561–576.

Fischer, P., & Greitemeyer, T. (2010). A new look at selective-exposure effects: An integrative model. *Current Directions in Psychological Science*, *19*, 384–389.

Jeffrey, R.C. (1965). *The logic of decision*. New York, NY: McGraw-Hill.

Johnson, S.G.B., Jin, A., & Keil, F.C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Johnston, A.M., Toig, A.E., & Keil, F.C. (2014). Explanatory scope informs causal strength inferences. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Kim, H.S., & Keil, F.C. (2016). Explanatory biases in social categorization. *Proceedings of the 38th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2015). Predictions from uncertain beliefs. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2014). Inferred evidence in latent scope explanations. *Proceedings of the 36th Conference of the Cognitive Science Society*.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015a). *Sense-making under ignorance*. Manuscript under review.

Johnson, S.G.B., Rajeev-Kumar, G., & Keil, F.C. (2015b). Belief utility as an explanatory virtue. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Johnston, A.M., Johnson, S.G.B., Koven, M.L., & Keil, F.C. (2015). Probability versus heuristic accounts of explanation in children: Evidence from a latent scope bias. *Proceedings of the 37th Conference of the Cognitive Science Society*.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*, 1449–1475.

Khemlani, S.S., Sussman, A.B., & Oppenheimer, D.M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–35.

Levitt, S.D., & List, J.A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*, 153–174.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–57.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–93.

Rottman, B.M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*, *140*, 109–139.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford, UK: Oxford University Press.

Sussman, A.B., Khemlani, S.S., & Oppenheimer, D.M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Waldmann, M.R., & Holyoak, K.J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222–236.

Does Chess Instruction Enhance Mathematical Ability in Children? A Three-Group Design to Control for Placebo Effects

Giovanni Sala (giovanni.sala@liv.ac.uk)
Department of Psychological Sciences, University of Liverpool, Bedford Street South, Liverpool, UK

Fernand Gobet (fernand.gobet@liv.ac.uk)
Department of Psychological Sciences, University of Liverpool, Bedford Street South, Liverpool, UK

Roberto Trinchero (roberto.trinchero@unito.it)
Department of Philosophy and Education, University of Turin, Via Gaudenzio Ferrari 9, Turin, ITA

Salvatore Ventura (salvatore.ventura@unimi.it)
University of Milan, ITA

Abstract

Pupils' poor achievement in mathematics has recently been a concern in many Western countries. In order to address this issue, it has been proposed to teach chess in schools. However, in spite of optimistic claims, no convincing evidence of the academic benefits of chess instruction has ever been provided, because no study has ever controlled for possible placebo effects. In this experimental study, a three-group design (i.e., experimental, placebo, and control groups) was implemented to control for possible placebo effects. Measures of mathematical ability and metacognitive skills were taken before and after the treatment. We were interested in metacognitive skills because they have been claimed to be boosted by chess instruction, in turn positively influencing the enhancement of mathematical ability. The results show that the experimental group (participants attending a chess course) achieved better scores in mathematics than the placebo group (participants attending a Go course) but not than the control group (participants attending regular school lessons). With regard to metacognition, no differences were found between the three groups. These results suggest that some chess-related skills generalize to the mathematical domain, because the chess lessons compensated for the hours of school lessons lost, whereas the Go lessons did not. However, this transfer does not seem to be mediated by metacognitive skills, and thus appears to be too limited to offer educational advantages.

Keywords: chess; mathematics; transfer; education.

Introduction

Recently, pupils' poor achievement in mathematics has been the subject of debate both in the United States (Hanushek, Peterson & Woessmann, 2012; Richland, Stigler, & Holyoak, 2012) and in Europe (Grek, 2009). Policy makers and researchers have investigated several alternative methods and activities with the aim of improving the effectiveness of mathematical teaching. Teaching chess in schools is one of these activities. Chess has recently become part of the school curriculum in several countries, and several large studies and educational projects involving chess are currently ongoing in Germany, Italy, Spain, Turkey, the United Kingdom, and the United States. Moreover, the European Parliament has expressed its positive opinion on using chess courses in schools as an educational tool (Binev, Attard-Montalto, Deva, Mauro, & Takkula, 2011).

Chess as Educational Tool: The available Evidence

Several studies have tried to demonstrate the potential benefits of chess training on various cognitive abilities such as attention (Scholz et al., 2008), development of spatial concepts (Sigirtmac, 2012), general intelligence (Hong & Bart, 2007), and metacognition (Kazemi, Yektayar, & Abad, 2012). Other studies focused on academic variables, such as reading and mathematics (Christiaen & Verhofstadt-Denève, 1981).

Recently, several studies have investigated the positive influence that chess could exert on children's mathematical abilities (Barrett & Fish, 2011; Kazemi et al., 2012; Sala, Gorini, & Pravettoni, 2015; Scholz et al., 2008; Trinchero, 2012; Trinchero & Sala, in press). Barrett and Fish (2011) examined the effect of chess instruction on adolescents receiving special education services, using TAKS (Texas Assessment of Knowledge and Skills) scores in mathematics. The chess-treatment group showed an overall better performance than the control group. In Kazemi et al. (2012), the same pattern occurred with a sample of typically developing male students from primary and secondary schools. Scholz et al. (2008) examined the possible benefits of chess instruction on a sample of primary school children with learning disabilities (IQ between 70 and 85), with disappointing results. Finally, Sala's and Trinchero's studies (Sala, Gorini, & Pravettoni, 2015; Trinchero, 2012; Trinchero & Sala, in press) focused on the effect of chess instruction on primary school children's mathematical problem-solving ability. In these studies the chess-treatment groups systematically outperformed the control groups.

The Lack of a Placebo Group

Following the above overall positive results, the view of the chess community has been that chess practice increases academic performance because chess is an intellectually demanding and stimulating game (Bart, 2014; Garner, 2012; Root, 2006). Such optimistic view has been challenged, *ante litteram*, by Gobet and Campitelli's (2006) review of the lit-

erature regarding the effectiveness of chess instruction in enhancing children's academic and cognitive abilities. Gobet and Campitelli highlighted that – in spite of some the promising results reported in the reviewed experiments – the benefits of chess practice were yet to be clearly established, because of the poor experimental design used in most studies. In particular, most studies did not control for placebo effects – that is, effects that are not due to the treatment per se but that are due to other, uncontrolled aspects of the experimental design. Potential mechanisms behind placebo effects include instructors' motivation, the state of excitement and attention induced by a novel activity, and teachers' expectations.

More recently, a meta-analysis (Sala & Gobet, 2016) has lent further support to Gobet and Campitelli's conclusions. In that meta-analysis, 24 studies passed the selection criteria, with 40 effect sizes. The overall effect size of chess instruction was moderate, with $g = 0.34$. It was also found that studies with at least 25 hours of chess instruction were more likely to have a positive effect on mathematics, reading, and cognitive abilities ($g = 0.43$) than studies with less than 25 hours treatment ($g = 0.30$). Only one study (Fried & Ginsburg, n.d.), which was interested in perceptual and visuo-spatial skills, included a placebo control group. This study showed no difference between the chess group and the placebo group, which consisted of attending counselling sessions. Unfortunately, Fried and Ginsburg's (n.d.) study did not investigate the effect of chess instruction on children's mathematical ability. Therefore, this study cannot corroborate or disprove any claims about the effectiveness of chess in enhancing mathematical abilities.

The lack of a placebo control group – i.e., a group attending other activities non-related to chess – has thus been identified as the most serious flaw of the previous studies in the field. These studies often compared the practice of the game of chess to regular school activities. However, chess practice may exert a moderate positive influence on children's academic just because it is a novel activity in their curriculum that is fun, and this may induce a state of increased attention and motivation.

From an educational (and practical) perspective, this might be irrelevant. Policy makers and educators might be interested in the effectiveness of chess instruction, regardless of any chess-specific or non-chess-specific element causing the improvement in pupils' academic skills. Assuming this "whatever works" perspective, the only element of interest would be the size of the effect – i.e., *how much* children attending chess courses improve their academic skills, such as mathematical literacy.

Nonetheless, evaluating whether the benefits of chess practice on children's academic skills are due to elements specific or non-specific to the game of chess is an important theoretical question. In fact, assuming that skills acquired in chess lead to benefits in domains such as mathematics clearly implies the presence of far transfer. Far transfer occurs when a set of skills acquired in one domain (e.g., chess) generalizes to other domains (e.g., mathematics) that are only loosely related to the source domain.

The Problem of Transfer in Chess

Since transfer is a function of the extent to which two domains share common features (Thorndike & Woodworth, 1901), far transfer rarely occurs, as shown by substantial research in education and psychology (Donovan, Bransford, & Pellegrino, 1999; Gobet, 2015a,b). Moreover, in line with Thorndike and Woodworth's (1901) hypothesis, several studies have shown that chess players' skills tend to be context-bound, suggesting that it is difficult to achieve far transfer from chess to mathematics. In her classic study, Chi (1978) demonstrated that chess players' (both adults and children) memory for chess positions did not extend to the recall of digits. Chess players outperformed non-chess players in remembering chess positions, but no difference occurred with lists of digits. The same result was obtained in a study carried out by Schneider, Gruber, Gold, and Opwis (1993). Waters, Gobet, and Leyden (2002) found that chess players' perceptual skills did not generalize to visual memory of shapes. More recently, Bühren and Frank (2010) found that chess grandmasters did not outperform chess amateurs in the economic game known as beauty contest. Finally, Unterrainer, Kaller, Leonhart, and Rahm (2011) found that chess players' planning abilities did not transfer to the Tower of London, a test assessing executive function and planning skills.

The Present Study

The research on the transferability of chess players' skills to other domains offers a complementary perspective to the research on the benefits of chess instruction on children's academic (e.g., mathematics) achievement. Whilst the latter has provided encouraging results, the former has offered results that seem to preclude any generalizability of chess-specific skills.

A possible explanation may be that chess instruction shares with the domain of mathematics some general features, such as quantitative relationships (e.g., the value of the chess pieces) and problem-solving situations (e.g., tactics), which can in turn generalize to mathematics. However, this transfer can occur only when these skills are at the beginning of their development, and hence still generalizable. Put simply, children studying the four operations may take advantage of an activity (i.e., chess) dealing with the calculation, for example, of the pieces' values. By contrast, it is hard to believe that knowing advanced chess techniques, such as Lucena's method in Rook endgames, might be beneficial for college students' skills in calculus or combinatorics. In the former case, there may be an overlap between the domain of chess and mathematics, in the latter there is none. This may explain why the studies dealing with the effect of chess instructions on children's mathematical ability have provided positive results, whilst studies regarding the generalizability of chess masters' skills have showed no effect.

An alternative explanation has been suggested by Kazemi et al. (2012), according to which chess instruction improves children's metacognition – i.e., the ability to monitor one's own cognitive processing (Brown, 1987; Flavell, 1979) – which, in turn, enhance their mathematical ability. Kazemi et

al. (2012) found that youngsters who attended a chess course outperformed the control group not only in mathematical ability, but also in metacognitive skills. Once again, unfortunately, this study did not control for possible placebo effects.

In any case, since metacognitive skills have been claimed to be one of the most important predictors of mathematical ability (Desoete & Roeyers, 2003; Veenman, Van Hout-Wolters, & Afflerbach, 2006), and since playing chess is an activity for which the self-monitoring of one's thinking processes is essential (De Groot, 1965), Kazemi et al.'s (2012) hypothesis is plausible. Obviously, the null hypothesis – i.e., chess instruction has no effect on children's mathematical and/or cognitive abilities, and the observed benefits are marginal and only due to placebo effects – may be valid too.

Since there is a need, in this field of research, to clarify whether the observed positive influence of chess practice is due to placebo effects or to chess training itself, we ran a three-group design study. Beyond the usual experimental and control groups – attending a chess course and regular school activities, respectively – an active control group was added to control for potential placebo effects. Moreover, the participants of this study were given – along with a test of mathematical ability – a questionnaire assessing metacognitive skills, in order to evaluate whether metacognition is the link connecting chess instruction to the improvements in mathematical ability, as proposed by several authors in the field.

This work thus investigates (a) whether chess instruction enhances children's mathematical ability, (b) whether this effect is mediated by metacognition, and (c) whether the effect of chess instruction on the above two variables is genuine or due to placebo effects.

Method

Participants

Fifty-two fourth graders in three classes of a primary school in Italy took part in this experiment. The mean age of the participants was 9.32 years (*SD* = 0.32). Parental consent was asked and obtained for all the participants.

Material

A six-item test was designed to test participants' mathematical ability (range score 0 – 6). The items used were all from the Italian version of the IEA-TIMSS international survey among fourth graders (Mullis & Martin, 2013), a test with good psychometric properties. These items were chosen because they measure not only procedural mathematical knowledge (e.g., 3 + 7 = ?), but also problem-solving ability. In fact, all the items required solving a mathematical problem starting from a given set of data. An example of such items is provided in Figure 1.

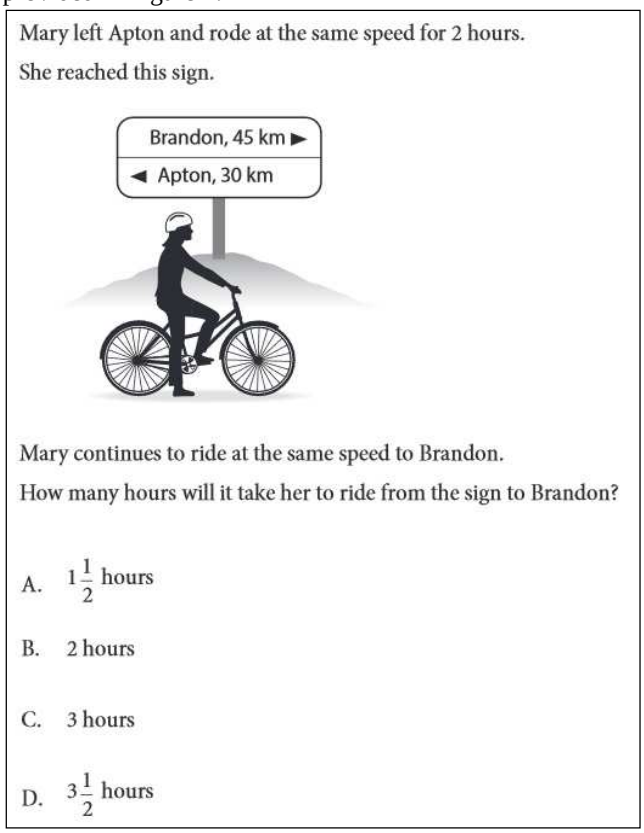


Figure 1. An example of the items used in the test of mathematical ability.

To assess participants' metacognitive skills, we used the Italian version of Panaoura and Philippou's (2003) questionnaire (15-item version; range score 15 – 75). Participants were given 45 minutes for completing the battery of tests.

Design

The three classes were randomly[1] assigned to three groups:

- One class attended 15 hours of chess lessons during school hours, along with regular school activities (experimental group).
- One class attended regular school activities only (control group).
- One class attended 15 hours of Go[2] lessons during school hours, along with regular school activities (placebo group).

Importantly, the two interventions – i.e., chess and Go courses – replaced part of the hours originally dedicated to

[1] Randomizing participants instead of classes is unpractical in schools, especially when interventions replace part of the curricular hours. The participants were assigned to the three classes according to accidental sampling – i.e., the teachers did not adopt any particular didactic criterion (e.g., the students struggling with math in a particular class).

[2] Go (or Baduk) is an Asian strategic board game for two players. The objective of the game is to surround a larger portion of the board (called *Goban*) than the opponent. To pursue this aim, the two players place *stones* (the game pieces) on the board.

mathematics and sciences. This way, we could confront the effectiveness of chess (and Go) instruction with the traditional way of teaching mathematics or mathematics-related subjects – such as sciences.

The chess and Go lessons followed a pre-arranged teaching protocol, which consisted of the basic rules of the games, tactical exercises, and playing complete games. All these activities focused mainly on problem-solving situations, such as finding the correct move and evaluating the advantages/weaknesses in a particular position. It must be noticed that the two courses (chess or Go) did not deal with or introduce any mathematics-related topics, unless these were part of the respective games (e.g., in chess, a Bishop is worth three Pawns). That is, the two courses had only chess- and Go-related contents. In order to rule out possible extraneous effects related to instructor personality (e.g., Pygmalion effect), the chess and Go interventions were carried out by the same instructor, who is an experienced teacher and both a chess and Go trainer. The tests of mathematical ability and metacognition were administered twice, once before the beginning of the course and once after the end. The experimental design is summarized in Figure 2.

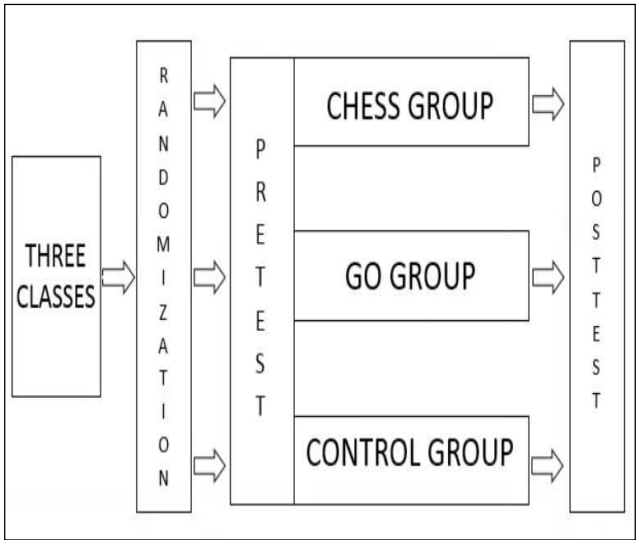


Figure 2. The experimental design.

Finally, the chess/Go trainer, the teachers, the parents, and the children involved reported no unpleasant issues. On the contrary, the people who took part in this project expressed a positive feedback.

Results

Mathematical Ability

No significant differences between the three groups were found in the pretest scores ($F(2, 51) = 1.030$, $p = .365$). A univariate analysis of covariance (ANCOVA) was used to evaluate the role of group (independent variable) and mathematics pretest scores (covariate) in affecting mathematics post-intervention scores (dependent variable).

The results showed a significant effect of the covariate ($F(1, 48) = 21.834$, $p < .001$), and a significant effect of group ($F(2, 48) = 3.371$, $p = .043$). The pairwise comparisons showed that the control group outperformed the Go group ($p = .017$), the chess group marginally outperformed the Go group ($p = .088$), whereas no significant difference was found between the control and the chess group ($p = .487$). The results are summarized in Table 1.

Table 1. Mathematical ability scores in the three groups.

Group	Pretest	Posttest	Adjusted mean
Chess	2.13 (1.26)	2.50 (1.41)	2.30
Go	1.81 (1.08)	1.62 (1.20)	1.63
Control	1.53 (1.13)	2.40 (1.55)	2.60

Note. Standard deviations are shown in brackets.

Metacognitive Skills

No significant differences between the three groups were found in the pretest scores ($F(2, 51) = 0.487$, $p = .617$). A univariate analysis of covariance (ANCOVA) was used to evaluate the role of group (independent variable) and metacognition pre test scores (covariate) in affecting metacognition post-intervention scores (dependent variable).

The results showed a significant effect of the covariate ($F(1, 48) = 47.809$, $p < .001$), and no significant effect of group ($F(2, 48) = 0.367$, $p = .694$). The pairwise comparisons showed no differences between the three groups. The scores are summarized in Table 2.

Table 2. Metacognitive-skill scores in the three groups.

Group	Pretest	Posttest	Adjusted mean
Chess	55.2 (11.0)	57.0 (10.5)	56.3
Go	52.7 (9.2)	54.8 (8.6)	55.8
Control	55.3 (6.5)	58.3 (6.0)	57.6

Note. Standard deviations are shown in brackets.

Discussion

According to the results presented in this paper, chess seems to be more effective in enhancing children's mathematical skills than Go, but not than regular school activities. This outcome – which is consistent with the aforementioned reviews (Gobet & Campitelli, 2006; Sala & Gobet, 2016) – might be discouraging for researchers and teachers who have upheld the implementation of chess instruction in school curricula. However, the fact that the placebo group (Go instruction) underperformed in this experiment – whilst the chess group equalled the performance of the control groups – suggests that some chess-related skills generalized to the domain of mathematics (Scholz et al., 2008), and therefore that the benefits of chess instruction, albeit limited, are not only the mere by-product of placebo effects.

With regard to metacognitive skills, children do not seem to benefit from any advantage from chess instruction. In fact, the participants of the three groups performed equally, both

in the pretest and in the posttest, suggesting that metacognition does not represent the cognitive link between chess instruction and mathematical ability.

Strengths and Limitations of the Study

In spite of the long history of research on the benefits of chess instruction on mathematical ability, which started more than thirty years ago with Christiaen and Verhofstadt-Denève (1981), the current study is the first to use a placebo group. In addition to its design, it has several strengths. First, the same instructor taught the experimental and placebo groups. Second, the interventions were implemented during school hours, and replaced part of the lessons dedicated to the teaching of mathematics or mathematics-related subjects (i.e., sciences), in order to compare directly chess and Go instruction to ordinary school teaching. Finally, the design included both a cognitive (i.e., metacognition) and an academic (i.e., mathematics) variable, in order to search for a possible causal link between chess instruction and enhancement of children's mathematical ability.

The study also suffers from a few weaknesses. The sample size was relatively small, which affected statistical power. Randomization was done at the class level rather than the individual level. (However, as remarked in footnote 1, randomization at the individual level is nearly impossible in the context of real schools.) Finally, only one measure of mathematical ability and metacognitive skills, respectively, was used.

Recommendations for Future Research

The present study supports the hypothesis according to which chess skill transfers to mathematical ability. Importantly, because the effect was not observed in the Go condition, this generalization of chess skill does not depend on placebo effects. However, this far transfer seems – when it occurs – to be limited in size, which is in line with substantial previous research in the field (e.g., Donovan et al., 1999). Furthermore, the results of this study do not corroborate Kazemi's et al. (2012) idea that chess instruction fosters children's mathematical ability by enhancing their metacognitive skills.

Given the near-absence of studies controlling for placebo effects in this line of research, it is essential to replicate and extend this work. First, since the duration of treatment seems to be positively related to the effect of chess instruction on cognitive and academic outcomes (Sala & Gobet, 2016), future studies should directly manipulate this variable, in order to understand the optimal duration of chess courses. Second, given that the positive effect of chess instruction does not appear superior to the regular curricular activities, it would be interesting to compare the effect of chess interventions held during school hours with the effect of chess interventions held during extra-curricular hours. Third, there has been little research that has explicitly tried to teach mathematics using chess. Possible examples include illustrating the Cartesian graph with the chess board and introducing the concept of block distance – as opposed to Euclidean distance – with the movement of the King. As it is known that awareness makes transfer more likely (Gick & Holyoak, 1980; Salomon & Perkins, 1989), it is plausible that making explicit the links between chess and mathematics could facilitate transfer. Finally, other activities could be used with the placebo groups, such as other board games (e.g., checkers) and music.

The chess-in-school field of research has been nearly exclusively interested in establishing the presence of benefits of chess instruction for curricular topics (mostly mathematics and reading) and general cognitive abilities (e.g. intelligence and creativity). However, very little research has been carried out on the mechanisms (presumably) leading to such benefits. A crucial aim for this field of research, then, is to develop a detailed causal model explaining the cognitive processes that mediate learning and transfer. With such a model, more precise hypotheses could be tested than it is currently the case.

Acknowledgments

The authors gratefully thank the principal, the teachers, the parents, and the children involved in this study.

References

Barrett, D. C., & Fish, W. W. (2011). Our move. Using chess to improve math achievement for students who receive special education services. *International Journal of Special Education, 26*, 181–193.

Bart, W. M. (2014). On the effect of chess training on scholastic achievement. *Frontiers in Psychology*, 5:762.

Binev, S., Attard-Montalto, J., Deva, N., Mauro, M., & Takkula, H. (2011). Declaration of the European Parliament, 0050/2011.

Brown, A. (1987). Metacognition, executive control, self-regulation and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding* (pp. 65–116). Hillsdale, NJ: Erlbaum.

Bühren, C., & Frank, B. (2010). Chess players' performance beyond 64 squares: A case study on the limitations of cognitive abilities transfer. *Joint Discussion Paper Series in Economics*, 19-2010.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73–96). Hillsdale, NJ: Erlbaum.

Christiaen, J., & Verhofstadt-Denève, L. (1981). Schaken en cognitieve ontwikkeling [Chess and cognitive development]. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden, 36*, 561–582.

De Groot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

Desoete, A., & Roeyers, H. (2003). Can off-line metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*, 188–200.

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academies Press.

Flavell, J. (1979). Metacognition and cognitive monitoring. *American Psychologist, 34*, 906–911.

Fried, S., & Ginsburg, N., (undated). *The effect of learning to play chess on cognitive, perceptual and emotional development in children*.

Garner, R. (2012). Chess makes a dramatic comeback in primary schools. Retrieved from: http://www.independent.co.uk/news/education/education-news/chess-makes-a-dramatic-comeback-in-primary-schools-8301313.html

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355.

Gobet, F. (2015a). Cognitive aspects of learning in formal and non-formal contexts: Lessons from expertise research. *British Journal of Educational Psychology, Monograph Series II: Number 11, Learning beyond the Classroom*, 23–37.

Gobet, F. (2015b). *Understanding expertise: A multi-disciplinary approach*. London: Palgrave.

Gobet, F., & Campitelli, G. (2006). Educational benefits of chess instruction. A critical review. In T. Redman (Ed.), *Chess and education. Selected essays from the Koltanowski Conference* (pp. 124–143). Dallas, TX: University of Texas at Dallas.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*, 23–37.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2012). *Achievement growth: International and US state trends in student performance*. Harvard's Program on Education Policy and Governance.

Hong, S., & Bart, W. M. (2007). Cognitive effects of chess instruction on students at risk for academic failure. *International Journal of Special Education, 22*, 89–96.

Kazemi, F., Yektayar, M., & Abad, A. M. B. (2012). Investigation the impact of chess play on developing meta-cognitive ability and math problem-solving power of students at different levels of education. *Procedia - Social and Behavioral Sciences, 32*, 372–379.

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Panaoura, A., & Philippou, G. (2003). *The construct validity of an inventory for the measurement of young pupils' metacognitive abilities in mathematics*. Paper presented at the International Group for the Psychology of Mathematics Education Conference, Honolulu, HI.

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In T. N. Postlethwaite & T. Husen (Eds.), *International Encyclopedia of Education* (pp. 6452–6457). Oxford: Elsevier.

Richland, L. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist, 47*, 189–203.

Root, A. W. (2006). *Children and chess: A guide for educators*. Westport, Ct: Teacher Ideas Press.

Sala, G., & Gobet, F. (2016). Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. *Educational Research Review, 18*, 46–57.

Sala, G., Gorini, A., & Pravettoni, G. (2015). Mathematical problem solving abilities and chess: An experimental study on young pupils. *SAGE Open, July-September*, 1–9.

Schneider, W., Gruber, H., Gold, A., & Opwis, K. (1993). Chess expertise and memory for chess positions in children and adults. *Journal of Experimenal Child Psychology, 56*, 328–349.

Scholz, M., Niesch, H., Steffen, O., Ernst, B., Loeffler, M., Witruk, E., & Schwarz, H. (2008). Impact of chess training on mathematics performance and concentration ability of children with learning disabilities. *International Journal of Special Education, 23*, 138–148.

Sigirtmac, A. D. (2012). Does chess training affect conceptual development of six-year-old children in Turkey? *Early Child Development and Care, 182*, 797–806.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 9, 374–382.

Trinchero, R. (2012). *Gli scacchi, un gioco per crescere. Sei anni di sperimentazione nella scuola primaria* [Chess, a game to grow up with. Six year of research in primary school]. Milan: Franco Angeli.

Trinchero, R., & Sala, G. (2016). Can chess training improve Pisa scores in Mathematics? An experiment in Italian primary schools. *Eurasia Journal of Mathematics, Science & Technology Education, 12*, 655–668.

Unterrainer, J. M., Kaller, C. P., Leonhart, R., & Rahm, B. (2011). Revising superior planning performance in chess players: The impact of time restriction and motivation aspects. *American Journal of Psychology, 124*, 213–225.

Veenman, M. V. J., Van Hout-Wolters, B., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14.

Waters, A. J., Gobet, F., & Leyden, G. (2002). Visuo-spatial abilities in chess players. *British Journal of Psychology, 30*, 303–311.

Are Symptom Clusters Explanatory?
A Study in Mental Disorders and Non-Causal Explanation

Daniel A. Wilkenfeld (daniel.wilkenfeld@berkeley.edu)*
Jennifer Asselin (asselin.7@osu.edu)**
Tania Lombrozo (lombrozo@berkeley.edu)*
*Department of Psychology, UC Berkeley
**Department of Philosophy, The Ohio State University

Abstract

Three experiments investigate whether and why people accept explanations for symptoms that appeal to mental disorders, such as: "She experiences delusions because she has schizophrenia." Such explanations are potentially puzzling, as mental disorder diagnoses are made on the basis of symptoms, and the DSM implicitly rejects a commitment to some common, underlying cause. Do laypeople nonetheless conceptualize mental disorder classifications in causal terms? Or is this an instance of non-causal explanation? Experiment 1 shows that such explanations are indeed found explanatory. Experiment 2 presents participants with novel disorders that are stipulated to involve or not involve an underlying cause across symptoms and people. Disorder classifications are found more explanatory when a causal basis is stipulated, or when participants infer that one is present (even after it's denied in the text). Finally, Experiment 3 finds that merely having a principled, but non-causal, basis for defining symptom clusters is insufficient to reach the explanatory potential of categories with a stipulated common cause. We discuss the implications for accounts of explanation and for psychiatry.

Keywords: explanation, understanding, mental disorders

Anecdotal evidence suggests that people consider diagnostic categories to be explanatory. For instance, one might explain a patient's high blood sugar levels by appeal to diabetes. This is also the case for mental disorders. A blurb about the film *A Beautiful Mind* claims that "the film displays the idea that Nash is a genius *because he has schizophrenia*" (Covell, 2013, emphasis added). In the wake of a mass shooting, it is common for people to cite the shooter's mental illness in explaining the atrocity (Craghill & Clement, 2015).

Consider the most basic form of such explanatory claims: those that appeal to a diagnostic category to explain the presence of a symptom. "He suffers from hallucinations because he has schizophrenia." "She is persistently sad because she has depression." Such claims may be common, but they're also potentially puzzling. The *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, known as the DSM-5, catalogues recognized mental disorders along with criteria for diagnosis. Importantly, these criteria involve the presence, duration, and severity of various *symptoms*, and are explicit that it is these symptoms which define the disorder: "A mental disorder is a syndrome *characterized* by clinically significant disturbance of an individual's cognition, emotion regulation, or behavior…" (American Psychiatric Association, 2013, p. 20, emphasis added). Explaining a symptom by appeal to a diagnostic category thus borders on a tautology – it's (almost) like explaining that someone has little money *because he is poor* (that's just what it *means* to be poor).

What, then, is the value of these explanations? One possibility is that laypeople's beliefs about mental disorders depart from the stipulations of the DSM – as even experts' do when it comes to causal relationships between symptoms (Kim & Ahn, 2002a, 2002b). Rather than thinking of mental disorders in terms of symptom clusters, laypeople may treat diagnostic labels more like medical disease labels, which often pick out some underlying condition that's *causally responsible* for observable symptoms. Past work finds that while clinicians generally reject the idea that mental disorders share causal "essences" (Ahn, Flanagan, Marsh, & Sanislow, 2006), laypeople do not (Cooper & Marsh, 2015). On this view, then, mental disorder classifications could be explanatory because they (perhaps indirectly) offer *causal* explanations. But another possibility is that laypeople accept at least some *non-causal* explanations (e.g., Prasada & Dillingham, 2006), and that DSM diagnostic categories support them.

Despite these connections to prior work, no research – to our knowledge – has investigated whether and why mental disorder classifications are found explanatory. Given their explicitly non-causal basis, DSM categories provide fruitful terrain in which to explore questions about the relationship between explanation and causation. Is a disease classification only explanatory when it picks out some common, underlying causal structure? Or can mere symptom clusters – clusters that are not tied by causal etiology – explain the presence of the symptoms associated with the corresponding diagnosis? We explore these questions in three experiments.

Experiment 1

In Experiment 1 we verify a presupposition of our project: that laypeople do, in fact, find mental disorder categories explanatory. To do so, we assess people's willingness to accept claims of the form: "Alex experiences hallucinations because she has schizophrenia." In addition to looking at absolute levels of agreement with such statements, we include medical diseases and non-explanations for comparison.

Methods

Participants Fifty-three adults (23 male, 30 female, mean age = 34) were recruited through the Amazon Mechanical Turk marketplace (MTurk) and participated in exchange for

monetary compensation. In all experiments, participation was restricted to users with an IP address within the United States and an approval rating of at least 95% based on at least 50 previous tasks. An additional seven participants were excluded prior to analysis for failing to consent, failing to complete the experiment, or giving an incorrect response to one of the reading comprehension or attention check questions (detailed below). For all experiments, those who had completed another experiment in this line of research were ineligible.

Materials and Procedures Participants were presented with 12 vignettes in random order. In each vignette, participants were told that a character displayed a symptom, went to a doctor, and was correctly diagnosed with a disease that had five listed symptoms. They were also told the character was suffering from a second, unrelated problem. For example, participants read:

> Rachel…has been experiencing a decreased range of motion (in her fingers)....She was correctly diagnosed with arthritis. Symptoms of arthritis include decreased range of motion, pain, swelling, stiffness, and inflammation.
>
> As it happens, Rachel also has headaches, which are not a symptom of arthritis.

After each vignette, participants were asked two reading comprehension questions, which asked both whether the main symptom (e.g., "decreased range of motion") and the unrelated symptom (e.g., "headaches") were symptoms of the disorder mentioned—anyone who answered either question incorrectly for any vignette was excluded from further analysis. (Participants were also excluded if they failed an attention check based on Oppenheimer, Meyvis, and Davidenko, (2009), at the end of the task.)

Participants were then asked to evaluate an explanation claim on a 7-point Likert scale from "Strongly Disagree" (1) to "Strongly Agree" (7):

> [Name] has [symptom] *because* s/he has [disease].
> (e.g., Rachel has a decreased range of motion *because* she has arthritis.)

In eight of the cases – four based on medical conditions ("Medical") and four based on mental disorders ("Mental") – the symptom asked about was in fact the first one listed for the disorder (e.g., whether someone had a skin rash because of measles or hallucinations because of schizophrenia). In the remaining four cases ("Control") – two medical and two mental – the symptom was from the unrelated condition (e.g., "Rachel has headaches because she has arthritis"). The control items were included to ensure that participants did not provide indiscriminately high ratings.

Results & Discussion

Responses to the "because" statements were averaged for each participant into three sets: those for the medical, mental, and control vignettes. These average ratings were then analyzed with a repeated-measures ANOVA with vignette type as a within-subjects factor, revealing a significant effect, $F(2, 104) = 571.644$, $p < .001$, $\eta_p^2 = .917$. Bonferroni pairwise comparisons revealed no significant difference ($p = .071$) between the Medical ($M = 6.35$, $SD = .611$) and Mental ($M = 6.14$, $SD = .779$) disorders, but that both differed significantly (ps $< .001$) from control cases ($M = 1.87$, $SD = .759$). Moreover, the ratings for Mental disorders were significantly above the scale midpoint ($p < .001$), suggesting that participants indeed found references to mental disorder categories explanatory.

Experiment 2

Having established that mental disorder classifications are found explanatory, we consider (in Experiment 2) whether this depends upon the causal structure of the category. Specifically, we varied whether diagnosis was based on the presence of a common cause or on a cluster of symptoms. For comparison, we also included a condition in which the disease was diagnosed in an arbitrary fashion. To limit effects of prior knowledge about mental disorders, we introduced fictional disorders on an alien planet.

Methods

Participants Participants were 141 adults (63 male, 77 female, 1 other, mean age 32) who were recruited through Amazon Mechanical Turk and participated in exchange for monetary compensation. Sixty-seven participants were excluded prior to analysis based on the same criteria as Experiment 1.

Materials Participants were presented with a vignette in which an alien was diagnosed with a mental disorder. Participants were divided into three groups ("Condition"). One group was told the diagnosis was made based on the presence of a particular cause ("*stipulated cause,"* N = 34), a second group was presented with diagnoses based on symptom clusters ("*symptom,"* N = 53), and a third group was presented with diagnoses based on which disorder name was pulled out of a hat ("*random draw*," N = 54).[1] Participants received symptoms corresponding to one of three disorders, which were modeled on depression, schizophrenia, or borderline personality disorder. For example, participants in the symptom/depression group read:

> John is an alien on the planet Zorg. Lately he has been experiencing a number of troubling symptoms, including persistent sadness.

[1] An equal number of participants was included in each condition prior to exclusions; however, participants in the *cause* condition did significantly worse ($p < .001$) on a true/false question about whether the disorder was diagnosed on the basis of common symptoms (this was supposed to be false for them). There were no other significant differences on reading comprehension or attention checks (ps $> .240$).

Recently, John went to the doctor to find out what was wrong. The doctor consulted her medical textbook, which is used by all doctors on Zorg as the standard for defining and diagnosing illness. ***It says that a doctor should diagnose a given mental illness when and only when the patient exhibits some number (but not necessarily all) of the symptoms on a list of symptoms corresponding to that illness.***

For example, it says that Gordon's Disease should be diagnosed when a patient has some number (but not necessarily all) of the symptoms of persistent sadness, trouble falling asleep, difficulty maintaining a stable weight, light-headedness, and difficulty concentrating. The book is very clear that ***the disease always has some of these symptoms.*** However, it doesn't always have the same cause in different people, or even the same cause for all symptoms within a given person. For example, it could be caused by a virus in some people, but by a genetic predisposition in others. Or even for the same person, some symptoms could be caused by a virus, and others by a genetic predisposition***. All that matters for having the disease is having the right set of symptoms***.

John does in fact have a number of these symptoms, so the doctor diagnoses him as having Gordon's Disease.[2]

All participants then rated their agreement with three statements about the disorder's explanatory status (in random order, not labeled for participants) from 1 ("Strongly Disagree") to 7 ("Strongly Agree"):

Because: [Name] is [symptom] *because* he has [disorder].

Understand: I *understand* why [name] is [symptom].

Token Cause: [Disorder] is the *cause* of [name's] [symptom].

Participants were also asked to generalize properties across individuals with the same diagnosis, but in the interest of space, we do not report these results here.

After these agreement and generalization ratings, participants were asked to answer the following question about the basis for the disorder, rated on a 7-point scale:

Inferred Common Cause: How likely do you think it is that there is a common cause behind all cases of [disorder]?

Participants were excluded from analyses if they could not correctly answer any of three true/false questions regarding the basis on which the disorder was diagnosed. For example, they had to answer whether "Gordon's Disease is diagnosed on the basis of symptoms sharing a particular cause." (The correct answers varied by condition.) At the end of the task, participants were asked what real disorder they thought was closest to the one in the vignette. Only 38 of 141 participants correctly identified the disorder with which they were presented, and whether people identified the disorder had no impact on other responses. Finally, participants supplied standard demographic information, reported any problems with the survey, and had the same attention check used in Experiment 1.

Results & Discussion

Responses were analyzed in a mixed ANOVA with condition (3: *random draw*, *symptom*, *stipulated cause*) and disorder (3: *depression*, *schizophrenia*, *borderline personality*) as between-subjects factors, and statement (3: *because*, *understand*, *token cause*) as a within-subjects factor. This analysis revealed a significant main effect of condition, $F(2, 132) = 25.80$, $p < .001$, $\eta_p^2 = .281$. Tukey post-hoc tests indicated significantly higher ratings for *stipulated cause* than *symptom* ($p < .001$), which were in turn significantly higher than *random draw* ($p < .005$) (See Figure 1). There were no other significant effects. Because statement did not interact with condition, we averaged the three responses to create a single "explanation score".

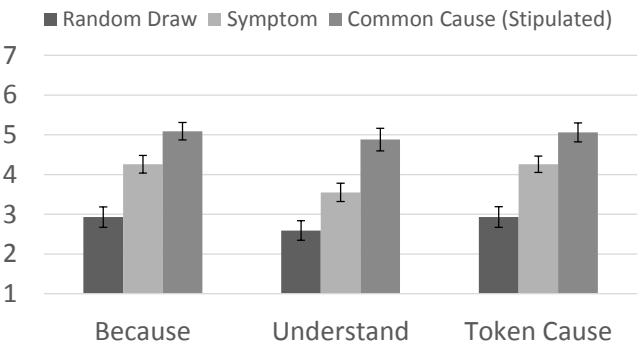


Figure 1: Experiment 2 Mean DVs by Condition. Error Bars ± 1 SEM

These results suggest that whether a diagnostic category is considered explanatory depends, at least in part, on the basis for category membership. The results also suggest that sharing a common cause is not necessary for conferring some explanatory potential: participants in the *symptom* condition gave significantly higher ratings than those in the *random draw* condition, although their ratings did not differ from the scale mid-point ($p = .890$.) However, it could be that these participants assumed the presence of a common cause, even though the vignette stipulated in those cases that none was present.

[2] We used the locution that a person "has" a disorder. Reynaert and Gelman (2007) found systematic differences in beliefs about disorder permanence depending on whether a noun-phrase, adjective-phrase, or possessive-phrase ("has") was used, which could also affect explanation judgments. This could not, however, account for differences we find *across* our conditions. Reynaert and Gelman (2007) cite the *Publication Manual* of the American Psychological Association as suggesting it's best to frame disorders in terms of possessive phrases (e.g., "has Gordon's Disease").

To investigate whether this occurred, we first examined responses to *inferred common cause* as a function of condition using a one-way ANOVA with *condition* as a between-subjects factor and *inferred common cause* as a dependent variable (see Figure 2). This analysis revealed a significant main effect, $F(2, 138) = 34.26$, $p < .001$, $\eta_p^2 = .332$; Tukey post-hoc tests revealed that all differences were highly significant (ps < .001).

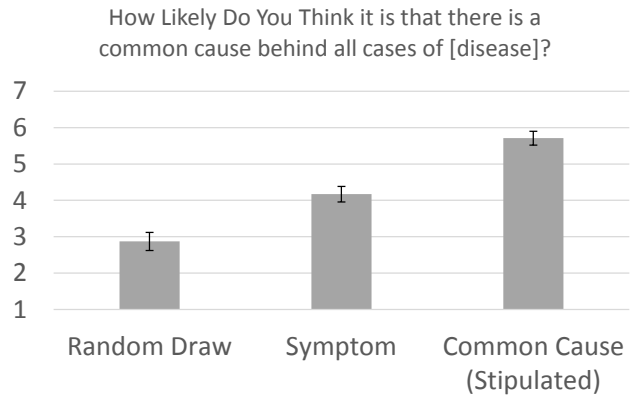


Figure 2: Experiment 2 Mean Inferred Common Cause ratings by Condition, from low (1) to high (7) likelihood. Error Bars ± 1 SEM

Although inferences to a common cause varied significantly across conditions, mean ratings reveal that many participants in the *symptom* condition (and even some in the *random draw* condition) believed there was a reasonable probability that one existed. Could it be, then, that diagnostic categories were only found explanatory to the extent that people believed the category picked out a common cause, regardless of the condition to which they were experimentally assigned?

To test this possibility, we ran a hierarchical regression. An initial model used the variable *inferred common cause* to predict explanation scores (see Figure 3). This model was highly significant ($p < .001$), with $R^2 = .35$. A second model that also included condition, coded as two dichotomous variables, accounted for more variance, with $R^2 = .40$ (this increase in R^2 was significant as assessed by a significant change in F-scores, $p < .001$). In this second model, *inferred common cause* had an unstandardized coefficient of .38 ($p < .001$), being in the stipulated cause condition (yes=1, no=0) had an unstandardized coefficient of .41 ($p = .184$), and being in the random condition (yes=1, no=0) had an unstandardized coefficient of -.72 ($p = .008$).

In sum, people's assessments of explanations were affected by *both* their inferences regarding the existence of a common cause and by the experimental manipulation of diagnostic procedure. This suggests that some factor (or factors) other than the presence of a common cause – and that varied across conditions – played a role in modulating judgments. In Experiment 3, we examined whether a relevant difference between the *common cause (stipulated)* condition and the *symptom* condition was the fact that in the former case, the symptom cluster had a non-arbitrary basis (the common cause) and presumably practical implications for treatment.

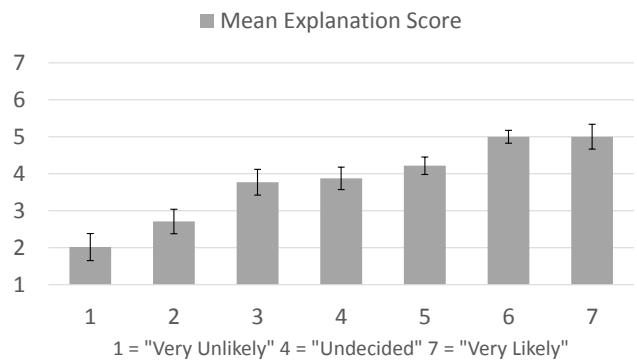


Figure 3: Experiment 2 Mean Explanation Score as a Function of Response to *Inferred Common Cause*.

Experiment 3

In this experiment, participants considered mental disorder classifications based on symptom clusters (without an underlying common cause), but in one case, the symptom cluster was motivated by non-arbitrary practical considerations, and in the second case, it was described as the result of historical accident. If having some principled basis for defining a symptom cluster – even when it's not causal – is sufficient to support good explanations, we would expect higher explanation ratings in the former case than in the latter.

Method

Participants Ninety-two adults (43 male, 49 female, 1 other, mean age 38) were recruited through MTurk and participated in exchange for monetary compensation. Fifty-nine participants were excluded prior to analysis based on the same criteria used in Experiments 1 and 2.

Materials Participants were presented with a single vignette in which an alien was diagnosed with a mental disorder, where the disorder was characterized by always sharing some cluster of the same symptoms. Participants were divided into two experimental groups ("Condition"). In the *reason* condition (N = 47), participants were told that the symptoms were grouped together into one disorder on the bases of treatment and prognosis. In the *no reason* condition (N = 45), participants were told that the symptoms were grouped together by historical accident. Both conditions stressed that there was no shared causal mechanism behind all instances of the disorder. As in Experiment 2, the vignettes were based on one of three mental disorders from the DSM: depression, schizophrenia, or borderline personality disorder.

In the *reason* condition involving depression, for example, participants read:

> Even though the symptoms of Gordon's Disease do not share a common cause, they were grouped together for principled reasons. When these symptoms occur in conjunction, they interfere with multiple facets of life in a way that interferes with close personal relationships and

with daily routines, making it difficult for people to obtain the social support and physical well being that could facilitate treatment. As a result, people who exhibit clusters of these symptoms are at heightened risk for suicide, and also need to pursue alternative treatment plans. It's because patients with these symptoms experience similar risks and are best suited to particular treatments that they're grouped under a common disorder.

For participants in the *no reason* condition, the relevant paragraph was replaced by the following:

> The symptoms of Gordon's Disease were grouped together under one disorder by historical accident, rather than for any principled reason. Had 18th century doctors documented the associated symptoms in a different order, the disease might have been defined by a very different cluster of symptoms.

All participants were then asked to rate their agreement with the measures used in Experiment 2, and also completed comprehension and attention checks. As in Experiment 2, only a minority of participants (33 out of 92) were able to identify the real disorder, and whether or not the disorder was correctly identified had no impact on responses.

Results & Discussion

Responses were analyzed in a mixed ANOVA with condition (2: *reason*, *no reason*) and disorder (3: *depression*, *schizophrenia*, *borderline personality*) as between-subjects factors, and statement (3: *because*, *understand*, *token cause*) as a within-subjects factor (See Figure 4). This analysis did not find a significant effect of condition (p = .830): overall, mean responses to the three statements were the same across the reason (M = 3.809, SD = 1.391) and no reason (M = 3.882, SD = 1.409) conditions. However, there was a significant interaction between condition and statement, (p = .004): only "understand" ratings decreased numerically from the *reason* to the *no reason* condition, though this decrease was not itself significant (p = .195), nor were the numerical increases found for "because" (p = .530) and "token cause" (p = .170). So while the significant interaction is suggestive and merits further scrutiny, we collapsed the three ratings into a single explanation score (as in Experiment 2) for subsequent analyses.

As in Experiment 2, there was a reliable association between explanation scores and the probability assigned to *inferred common cause*, r = .281, p = .007. Interestingly, responses to *inferred common cause* did not themselves vary across conditions (p = .888).

In sum, when evaluating whether a diagnostic category offers an explanation for its symptoms, participants were insensitive to the question of whether the symptoms were grouped on the basis of a (non-causal) principled reason or a historical accident. However, once again, there was a reliable association between explanation ratings and inferences concerning the existence of a common cause.

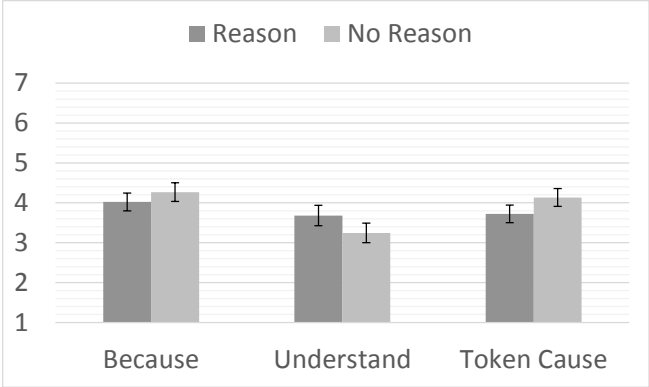


Figure 4: Experiment 3 Mean DVs by Condition. Error Bars ± 1 SEM

General Discussion

Across three experiments, we find that people are willing to explain symptoms by appeal to mental disorder classifications, but that their willingness to do so depends in large part on causal assumptions that are not endorsed by the DSM, and indeed denied in some of our vignettes. Specifically, Experiment 1 used real mental disorder classifications, and found that these classifications were not only found explanatory, but as explanatory as medical diagnoses. Experiment 2 used fictional disorder names in an alien context, which allowed us to stipulate the causal basis for each disorder classification. We found that classifications were most explanatory when they corresponded to a common cause. However, participants also found classifications based on symptom clusters significantly more explanatory than those based on a random draw. This pattern of results was largely, but not exclusively, driven by participants' assumptions about the presence of a common cause: even when the characterization of the disorder denied a common cause, participants in the symptom condition often inferred that one existed. Finally, Experiment 3 found that a non-arbitrary but non-causal basis for grouping symptoms was insufficient to improve the explanatory potential of a symptom cluster.

One robust finding from Experiments 2 and 3 is that people consider diagnostic categories more explanatory when they correspond to a common cause across cases. This is broadly consistent with causal accounts of explanation in philosophy (e.g., Woodward, 2003), according to which explanations identify one or more causes of what's being explained. However, causal accounts of explanation are both more and less demanding. On the one hand, a causal explanation need not identify a *single* common cause. It's typically sufficient to pick out a cause (or causes) that operated in the case being explained. It's not entirely clear, though, what this means when the explanation invokes a *diagnostic category* with some causal basis rather than the causes themselves. It could be that people take the diagnostic category to pick out a circumscribed *set* of causes, and thus find it explanatory by virtue of its implicit causal content, even when this content falls short of identifying a single common cause.

In another sense, causal accounts of explanation may be insufficiently demanding: inferring a common cause in our *symptom* conditions was typically enough to boost explanation ratings just above the scale mid-point, but it did not subsume the effects of experimental condition. It could be that a diagnostic category is more explanatory when it is itself *characterized* in causal terms, which is a condition that goes beyond the mere existence of a common cause.

Another possibility is that (some of) our participants were engaged in a genuinely non-causal form of explanation. Indeed, while the dominant approach to explanation within philosophy is causal, there are a variety of alternatives (Woodward, 2014), including those that focus on unification (e.g., Friedman, 1974), pragmatic import (e.g., Wilkenfeld, 2014), or argumentative structure (e.g., Hempel, 1965). Similarly, some psychological approaches to explanation suggest more formal accounts (e.g., Prasada & Dillingham, 2006).

While it's certainly possible that participants endorsed mental disorder classifications as explanatory for some non-causal reason, the findings from Experiment 3 speak against the most obvious possibilities. In particular, introducing a non-arbitrary basis for the symptom cluster, as we did in the *reason* condition, should have made the diagnostic categories better candidates for explanation on most accounts: the manipulation made the symptoms more inferentially useful, introduced pragmatic import, and arguably suggested some basis for unification. There's clearly more work to be done, but our initial findings present a puzzle for non-causal approaches.

How would clinicians respond in our task? Given prior work suggesting that expertise is associated with weaker beliefs in causal essences underlying mental disorders (Ahn et al., 2006; Cooper & Marsh, 2015), we speculate that clinicians would be less likely to endorse the explanations offered here. Similarly, we speculate that if laypeople knew what mental disorders were really thought to be by clinicians (Cooper & Marsh, 2015), they would not consider them (as) explanatory. However, our findings also suggest that dislodging laypeople's causal-essentialist beliefs about mental disorder classifications is no easy task. In the *symptom* condition of Experiment 2, and in both conditions of Experiment 3, many participants believed a common cause was likely, even though it had been explicitly denied.

While psychiatrists are largely interested in a classification scheme that best serves diagnosis and treatment, rather than one that reflects lay intuitions, the present results might have important implications. There's evidence that treatments are less effective if the patient does not believe the treatment will work (e.g., Seligman 1991), and people's intuitive beliefs about mental disorders have implications for their views about the efficacy of different kinds of treatment. It follows that lay beliefs could inform psychiatric practice.

The impact of our results on philosophy is more pronounced. People's persistence in seeing causes where there are (by stipulation) none to be had—and the fact that that tendency seems to account for much of their explanatory judgments—underscores the central role of causation in explanation.

References

Ahn, W. K., Flanagan, E. H., Marsh, J. K., & Sanislow, C. A. (2006). Beliefs about essences and the reality of mental disorders. Psychological Science, 17(9), 759-766.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders: DSM-5. Washington, D.C: American Psychiatric Association.

Cooper, J. A., & Marsh, J. K. (2015). The influence of expertise on essence beliefs for mental and medical disorder categories. *Cognition, 144*, 67-75.

Covell, K. (2013, May 23). A beautiful mind [Presentation]. Retrived from https://prezi.com/vfuqvvu3pzxq/a-beautiful-mind/

Craighill, P., & Clement, S. (2015, October 26). What Americans blame most for mass shootings. *Washington Post*. Retrieved from http://tinyurl.com/oj79jrw

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71*(1), 5-19.

Harvey, A.G., Soehner, A., Lombrozo, T., Bélanger, L., Rifkin, J. & Morin, C.M. (2013). 'Folk theories' about the causes and treatment of insomnia. Cognitive Research and Therapy, 37, 1048-1057. doi:10.1007/s10608-013-9543-2

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science by Carl G. Hempel*. New York: Free Press.

Kim, N. S., & Ahn, W. (2002a). The influence of naive causal theories on lay concepts of mental illness. American Journal of Psychology, 115(1), 33-66.

Kim, N. S., & Ahn, W. (2002b). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. Journal of Experimental Psychology: General, 131(4), 451.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73-112.

Reynaert, C. C., & Gelman, S. A. (2007). The influence of language form and conventional wording on judgments of illness. Journal of Psycholinguistic Research, 36(4), 273-295.

Seligman, M. E. (1991). *Learned optimism*. New York: Knopf.

Strevens, M. (2004). The causal and unification approaches to explanation unified—causally. Noûs, 38(1), 154-176.

Wilkenfeld, D. A. (2014) Functional explaining: A new approach to the philosophy of explanation. *Synthese*, *191*(14), 3367-3391.

Woodward, J. (2003). *Making things happen: A theory of causal explanation* Oxford University Press.

Woodward, J. (2014). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Not all overlaps are equal: Social affiliation and rare overlaps of preferences

Natalia Vélez (nvelez@stanford.edu), Sophie Bridgers (sbridge@stanford.edu), & Hyowon Gweon (hyo@stanford.edu)
Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305 USA

Abstract

Shared preferences are a critical component of social attraction. Knowing that someone likes the same things as you do is indicative of broader underlying similarities that support successful social partnerships. However, not all overlaps in preferences are equally informative. Here we propose that the rarity of overlaps in preferences may be a particularly salient cue for social affiliation. We find evidence that people are sensitive to the rarity of overlaps in preferences and affiliate themselves (Experiment 1) or predict others' affiliations (Experiment 2) with potential social partners who share a relatively rare preference. Because preferences provide information about both what people know and what they like, we also tested the effect of overlaps in knowledge (without taste) and overlaps in taste (without knowledge) to understand why we are drawn to people who share our preferences.

Keywords: social categories; preferences; probabilistic reasoning, social affiliation

Introduction

We are drawn to people who are similar to us. We find them attractive and trustworthy, and we find ourselves wanting to befriend them and learn from them (Boer et al., 2011). Of the many traits we might share with others, preferences exert a particularly powerful pull. Initial conversations with potential friends are filled with exchanges of one's likes and dislikes (Rentfrow & Gosling, 2006). We delight in discovering overlaps in these preferences, bonding over being fans of the same musicians or even liking the same local coffee shop.

However, not all overlaps in preferences are equally meaningful; some are better indicators of a successful friendship. Overlaps in favorite movies may be more meaningful than overlaps in favorite days of the week, and meeting someone who loves your favorite movie may be more exciting than finding someone who somewhat enjoyed your favorite movie. While there may be many factors that modulate the personal significance of shared preferences, here we focus on one factor that may have a general and profound influence on how we perceive and interpret such commonalities: the *prevalence*, or *rarity*, of overlaps in preferences.

Imagine meeting someone at a party who is raving about his iPhone. Even though you like your iPhone, too, this coincidence might not strike you as particularly meaningful; after all, many people also like iPhones. However, if you met someone who is enthusing about his Raspberry Pi (a tiny programmable computer), which you also like, somehow you might ascribe more meaning to this shared preference. Since Raspberry Pi users are much less common than iPhone users, meeting someone who shares your preferences for Raspberry Pi would also be much rarer than meeting someone who also appreciates iPhones. Here we propose that sensitivity to the distribution of preferences among a population informs our reasoning about people who share these preferences.

There are reasons to believe that people might be sensitive to the rarity of preferences. Developmental research suggests that humans are adept statistical reasoners even from an early age, readily differentiating between statistically common and uncommon events. Even preverbal infants use distributions of objects in the population to infer the likelihood of a sample from the population (Xu & Garcia, 2008; Xu & Denison, 2009), and attribute preferences or intentionality to agents whose behaviors violate random sampling (Kushnir et al., 2010; Gweon et al., 2010). Beyond sensitivity to the relative frequencies of physical events, teens and adults readily estimate the frequencies of different behaviors and preferences in the population. For example, sharing less mainstream tastes is found to be a good predictor of stable friendships in adolescents (Selfhout et al., 2009).

Why do shared preferences occupy such a privileged place in our social interactions? First, preferences might reveal much deeper information about values, personality, and past experiences than shallow similarities such as physical appearance (e.g. Rentfrow & Gosling, 2006). Second, by identifying those who like what we like, we could make better decisions about whom to approach for more information in the future (Fawcett & Markson, 2010). In fact, an overlap in preferences doesn't simply reflect shared tastes (i.e., liking); it also implies shared knowledge. After all, in order for you to like tinkering with Raspberry Pi, you had to know about it in the first place. It is possible the broader commonalities we infer from shared preferences may either be driven more by the overlap in knowledge, or by the overlap in tastes.

Shared knowledge between two individuals may be a useful indicator of similar personal histories or shared social experiences that allowed both individuals to have acquired that knowledge. By contrast, shared tastes alone may only reflect a relatively narrow set of commonalities such as similar personality traits or sensory experiences. Imagine meeting someone who tells you a lot about Raspberry Pi without expressing a clear preference, and someone who has never heard of it but after trying yours, really likes it; chances are you would assume more common ground with the former than the latter. Thus people might consider shared knowledge as more predictive of successful social affiliation than shared tastes. Consistent with this idea, a recent study by Soley and Spelke (2016) finds that children indeed place more weight on shared musical knowledge than shared musical taste when choosing whom to befriend.

The current study is inspired by the intuitive significance of rare preference overlaps in our everyday lives, and is motivated by both classic and recent research on our ability to estimate, represent, and use information about approximate statistical properties of events and behaviors. In a series of

experiments, we ask whether the distribution of preferences in a population influences people's reasoning about the significance of preference overlaps. Furthermore, given prior work on the role of shared beliefs (Soley & Spelke, 2016), we attempt to separate the effect of shared preferences into "shared tastes" and "shared knowledge." We predicted that (1) rare overlaps in preferences are stronger cues for social affiliation than common overlaps, (2) shared knowledge has a larger impact on people's social affiliation judgments than do shared tastes, and (3) both the effect of shared knowledge and tastes would be systematically modulated by prevalence.

Experiment 1

In Experiment 1, we used a first-person paradigm to ask whether people use the prevalence of their own preferences as a cue for social affiliation. Participants first provided their preferences and generated prevalence estimates for each; they then chose someone to talk to between two agents who each shared one of their preferences. We predicted that, even though both agents have overlapping preferences with the participant, the relative prevalence of these preferences would influence the participant's choice: namely, people might prefer agents who share rarer preferences. In Experiment 1a, we looked at the effect of shared preferences, broadly construed; in Experiment 1b, we used identical task structures to tease apart the effect of shared knowledge and shared tastes.

Methods: Experiment 1a

Participants 300 adults participated in an online study on Amazon Mechanical Turk. All participants in this and subsequent experiments had U.S. IP addresses and provided informed consent in accordance with the requirements of Stanford IRB. 50 participants provided contradictory responses (see Procedure, below) and were excluded from the analyses.

Procedure Participants were first asked to select their favorite activity amongst three options: listening to music, watching movies, or reading books. These domains were selected based on pilot data and past research indicating that overlaps in these domains are stronger cues to social affiliation than overlaps in other domains such as food (Rentfrow & Gosling, 2003). Within their favorite domain, participants were asked to list their top five favorite items (i.e., songs, movies, or books) in no particular order. We then asked how "widely liked" each item is. Participants provided this popularity estimate in two steps: first, they rank-ordered each item from most to least widely liked by dragging and dropping each item into a list; second, they provided a numeric estimate of how many of 100 randomly selected people would like each item. People who provided inconsistent responses between these two questions (i.e., estimating that more people would like the lowest-ranked item than the highest-ranked item) were excluded from analysis.

In the final trial, participants were given a binary choice between two chat avatars of people they could talk to. Each avatar had a prompt above it saying "I like X!", where X was

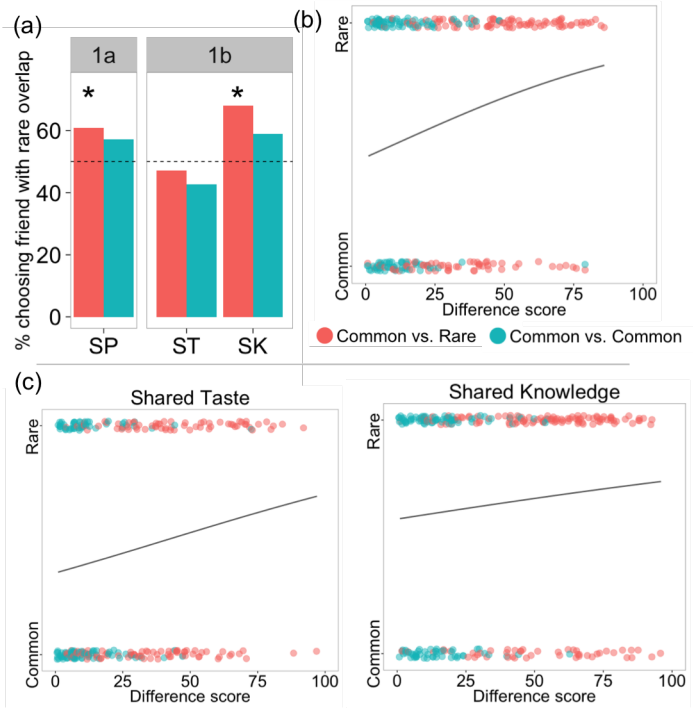


Figure 1: Experiment 1 results. (a) Percentage of participants who chose the agent with rare preference overlap. Asterisks indicate deviations from chance ($p < 0.05$). (b) Relationship between difference scores and friend choice: each point indicates one participant's difference score and friend choice; line indicates the logistic regression fit. (c) Logistic regression fits for Shared Taste (left) and Shared Knowledge (right).

one of the items the participant had listed. Participants were split into two between-subjects conditions. In the Common vs. Common (C vs.C) condition, participants were given a choice between an agent who liked the most widely-liked item and an agent who liked the *second* most widely-liked item. In the Common vs. Rare (C vs. R) condition, participants chose between an agent who liked the most widely-liked item and an agent who liked the *least* widely-liked item.

Methods: Experiment 1b

Participants We recruited 602 adults from Amazon mTurk. 78 participants were excluded for providing inconsistent responses (See Exp.1a Procedure).

Procedure The procedure was identical to that of Exp.1a except for a few prompts. Participants in the "Shared Knowledge" (SK, $n = 269$) group were asked how *widely known* each item is (i.e., how many of 100 people would know about each item). Participants in the "Shared Taste" (ST, $n = 255$) group were asked how *likable* each item is (i.e., how many of 100 people would like each item if given a chance to watch, listen to, or read it for the first time). Finally, participants were given a choice between two agents who shared the participant's knowledge (SK) or tastes (ST). As in Exp. 1a, both

SK and ST groups were split into C vs. C and C vs. R conditions.

Results and Discussion

Experiment 1a We hypothesized that participants would be more likely to choose the agent who shared their rarer preference, and this tendency would be modulated by the relative popularity of the two items. The results were consistent with our hypotheses. In the C vs. C condition, where the difference in popularity between the two agents' preferences was negligible, participants' responses did not differ from chance (57.14%; binomial test: $p = 0.142$). However, participants in the C vs. R condition were more likely to choose the agent who shared their rarer preferences (Figure 1(a), 60.84%; binomial test: $p = 0.012$).

We further examined whether participants' social affiliations were predicted by their fine-grained estimates of how widely shared their preferences are among the general population. For each condition, we took the difference of the numerical estimates of popularity (henceforth "Difference Scores") either between the first and second most widely liked items (C vs. C condition) or between the first and last items (C vs. R condition) to ask if this predicted people's choices. Even after controlling for condition, there was a significant relationship between the difference scores and participants' friend choice: the higher the gap in popularity between the two items, the more likely participants were to choose the agent with the rarer preference (logistic regression: $z = 3.446, p < 0.001$; Figure 1(b)).

Experiment 1b Liking something typically means that one also has prior knowledge of it. Under this assumption, Experiment 1b aimed to tease apart two factors that underlie the effect of shared preferences on social affiliation: shared knowledge (SK; without explicit mentioning of liking) and shared taste (ST; without prior knowledge). The results revealed interesting differences between the two factors. Participants in the SK group consistently chose the agent who knew about the less widely-known item in the C vs. R condition (68.03%; binomial test: $p < 0.001$); this effect was weakly present even in the C vs. C condition (59.02%; $p = 0.06$) where the difference between popularity was small. Consistent with this preference for the rare-overlap agent across both conditions, the Difference Scores did not predict people's choices ($z = 1.54, p = 0.12$) (Figure 1(c)). By contrast, the participants in the ST group did not differ from chance in either condition (C vs. C: 42.61%, $p = 0.14$; C vs. R: 47.14%, $p = 0.55$); however, the difference scores still predicted their friend choice (logistic regression: $z = 25.2, p = 0.01$).

Collectively, these results are consistent with our hypothesis that rare preference overlaps exert a stronger influence than common preference overlaps on people's judgments for social affiliation. In Exp. 1a, we found that the *degree* of rarity influenced people's choices: participants consistently chose the agent with the rarer preferences when the difference in rarity of the two agents' preferences was large (C vs. R), but not when this difference was small (C vs. C), and the magnitude of the difference predicted people's choices. Results from Exp. 1b echoed and extended these results, revealing differences between the two subcomponents of shared preferences: shared knowledge and shared tastes. Overall, shared knowledge had a much larger influence on people's choices compared to shared tastes.

One unique strength of this task is that it harnessed participants' own preferences and real-world knowledge: participants expressed their own preferences and generated their own estimates about their prevalence among the general population. Using this naturalistic approach, we were able to confirm the intuitive significance of rare preference overlaps.

However, these strengths were accompanied by some limitations. First, we may have neglected asymmetries between domains (music, books, movies). For example, people consider music preferences as particularly emblematic of their personalities, followed closely by movies, but far ahead of other domains such as food (Rentfrow & Gosling, 2003); thus people's prior knowledge in each domain and their beliefs about their diagnosticity might have influenced their judgments about the significance of an overlap. Second, we had little control over what population people used to generate the prevalence estimates. While it is indeed fascinating to wonder what cognitive mechanisms underlie our ability to estimate, for example, that 95 of 100 people like the *Harry Potter* books but only 8 of 100 like *The Silmarillion*, this method presumably introduced some variability in the exact nature of the population people drew from to generate these estimates. Furthermore, the self-generated nature of these estimates limited our ability to systematically manipulate the rarity of preferences. In Experiment 2, we use a complementary third-person task to directly address these issues.

Experiment 2

In Experiment 2, we sought a complementary way of investigating the role of rare preference overlaps in a tightly controlled context. We introduced participants to an alien planet whose inhabitants expressed their preferences for two novel games. Participants were then asked to pair these aliens with potential friends. Instead of generating their own prevalence estimates, participants were given explicit information about the popularity of each game, which was varied systematically across conditions. Thus this task allowed us to eliminate the role of people's existing knowledge and preferences, and have a tighter control over the prevalence information.

Methods

Experiment 2a: Methods

Participants 692 adults were recruited from Amazon Mechanical Turk.

Procedure Participants were introduced to a faraway planet called "Gazoom," inhabited by friendly aliens (Gazorps). Every Gazorp on the planet was asked whether they like to play two fun games, "wumbus" and "jibboo", and the results of

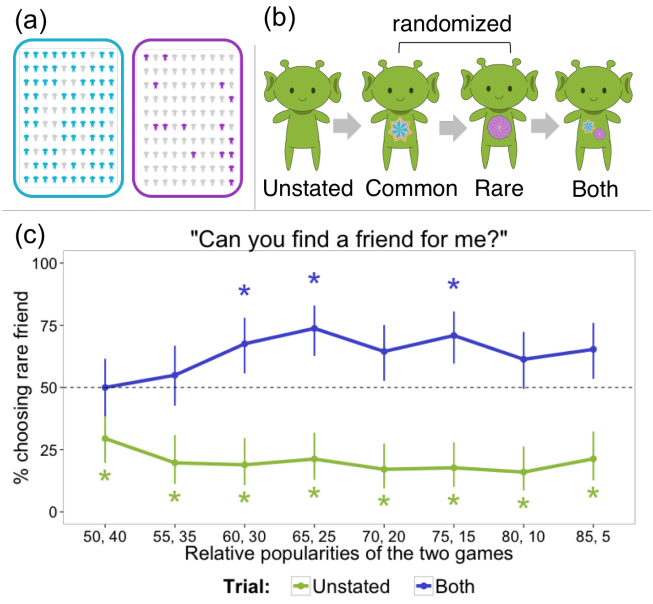


Figure 2: Experiment 2a results. (a) Example population information (75-15). (b) Trial order, showing the preferences of the target Gazorp on each trial. (c) Percentage of participants choosing the friend who likes the rare game. Error bars denote 95% CI. Asterisks denote significant deviations from chance, after Bonferroni correction.

the poll were presented on two separate pages. On each page, participants saw the logo for either wumbus or jibboo above the gray silhouettes of 100 Gazorps, a census representing the entire population of Gazoom. The Gazorps who liked each game were then filled in with the color of the game to indicate the game's popularity. Participants were shown a reminder of the census results throughout the study (Figure 2(a)).

In each of the four test trials, participants saw a target Gazorp who asked the participant to "find a friend" for her among two other Gazorps. Participants were asked to select a friend for: (1) a Gazorp with no stated preference, (2) a Gazorp who liked the majority-preferred game, (3) a Gazorp who liked the minority-preferred game, and (4) a Gazorp who liked *both* games (Figure 2(b)). In all trials, participants were given the choice between a potential friend who liked wumbus or a potential friend who liked jibboo. The preferences of all Gazorps were clearly indicated using verbal prompts and by overlaying the game logo over the image of the Gazorp.

Participants were randomly assigned to one of nine conditions, which differed in the number of aliens (out of 100) who liked each game: 45-45, 50-40, 55-35, 60-30, 65-25, 70-20, 75-15, 80-10, or 85-5. The order of the first and last trials was fixed, and the two middle trials were shuffled. All other aspects of the design were randomized, including: the logo for the games, the game preferred by the majority, the order in which the two polls were presented, the names of the target Gazorps, and the position of potential friends in the test trials.

Experiment 2b: Methods

Participants 1803 adults were recruited from Amazon Mechanical Turk. 228 participants were excluded for failing a manipulation check (see Procedure below).

Procedure The procedure was identical to that of Experiment 2a except for small changes in the prompts. In the Shared Knowledge (SK) group (N=823), Gazorps were asked which games they *knew* how to play, and the census results displayed how many Gazorps knew how to play each particular game. In the Shared Taste (ST) group (N=752), Gazorps were asked to try two new games that they had never played before, and the census results showed how many Gazorps liked each game after trying it for the first time. Accordingly, in the four test trials, participants were given the choice between potential friends who knew different games (SK) , or liked different games after trying them (ST).

Results and Discussion

Experiment 2a We first examined participants' responses in the middle two trials. Because the target Gazorp had a single, clearly stated preference, we expected participants to select the friend who shared the target Gazorp's preference. Indeed, 603 of 692 participants matched target Gazorps with friends who liked the same games in both trials; 52 matched in one trial, and 38 matched in neither trial ($\chi^2(2) = 901, p < 0.001$). This confirmed that the participants understood the task, and that they robustly chose the Gazorp who shared the target's preference regardless of the rarity.

Our main interest was in people's responses in the other two trials, where the target Gazorps' preferred games were uninformative. We first looked at the control (45-45) condition, where the two games were equally popular ($N = 84$). Given no difference in popularity, we predicted people to choose randomly between the two friends. This allowed us to looked for three potential sources of response bias: (1) the friend's position (left or right); (2) the names of the games (wumbus or jibboo); and (3) the color of each game (blue or purple). Across all three criteria, participants' responses did not differ from chance (all $p > 0.10$), suggesting that neither of the two potential friends was more appealing *a priori*.

In the remaining 8 conditions ($N = 608$), we asked whether participants' choice of friends for the target Gazorp was influenced by prevalence information. In the first trial where the target Gazorp's preference was *unstated*, we predicted that participants would choose the Gazorp with the more common preference; in the last trial where the target liked both games, we predicted that participants would choose the Gazorp with the rarer preference. Finally, we predicted that people's responses would be systematically influenced by the relative differences in the popularity of the games.

As predicted, participants' responses varied based on the trial type (Unstated vs. Both)(Figure 2 (c and b)), suggesting that they used the population information. Collapsing across all conditions, in the first trial (Unstated), 79.77% of par-

ticipants chose the friend with the most common preference (485/608; binomial test: $p < 0.001$), and difference in the popularity of two games did not predict participants' choices (logistic regression: $z = 1.43, p = 0.151$). By contrast, in the last trial (Both), 63.65% of participants chose the friend with the rare preference (387/608; binomial test: $p < 0.001$), and the difference in the popularity significantly predicted these choices ($z = 1.97, p = 0.049$).

Experiment 2b Our results in Exp 2b aligned well with those in Exp 2a. In the middle two trials where the target had a preference for one of two games, participants consistently chose the friend who knew about (Shared Knowledge) or liked (Shared Taste) the same game. In the SK group, 618 matched in both trials, 82 in one, and 123 in none ($\chi^2(2) = 648, p < 0.001$); in the ST group, 588 matched in both trials, 114 in one, and 64 in none ($\chi^2(2) = 683, p < 0.001$).

When the target Gazorp's preferences were unstated (first trial), most participants chose the friend who knew or liked the common game (SK: 67.83%; ST: 77.08%; binomial test: all $p < 0.001$). Conversely, when the target Gazorp liked or knew about both games (last trial), most participants chose the friend who knew or liked the rare game (SK: 67.07%; ST: 57.45%; all $p < 0.001$).

We then used multinomial logistic regression to examine whether the relative popularity of the two games affected friend choice in the Unstated and Both trials. In the Unstated trial, participants in both SK and ST groups were more likely to choose the friend who knew or liked the more common game as the difference in the prevalence between the two games increased (SK: $z = 3.42, p < 0.001$; ST: $z = 3.94, p < 0.001$). In the Both trial, differences in popularity had the opposite effect: participants were more likely to choose the friend who liked or knew about the *less* common game as the difference in the prevalence between the two games increased (ST: $z = -3.25, p = 0.001$; SK: $z = -4.05, p < 0.001$). These results show that when the target Gazorp knew or liked both games, participants chose the friend who knew or liked the rarer game. A complementary way of looking at this is to examine each level of popularity separately and to test *at which levels* participants' friend choices differed from chance (Figure 3). We accounted for multiple comparisons using Bonferroni correction ($p < 0.006$). This analysis revealed a striking asymmetry between SK and ST groups. Participants in the SK group were highly influenced by the rarity information, choosing the friend who knew the rare game in most conditions, except when the difference between the two games was minimal (50-40: 55.29%, $p = 0.39$; all other $p < 0.003$). However, participants in the ST group chose the friend who liked the rare game when the difference in popularity between the two games was at its most pronounced (85-5: 66.67%, $p = 0.001$; all other $p > 0.008$).

General Discussion

In a series of experiments, we found converging evidence that overlaps in rare preferences exert a stronger influence

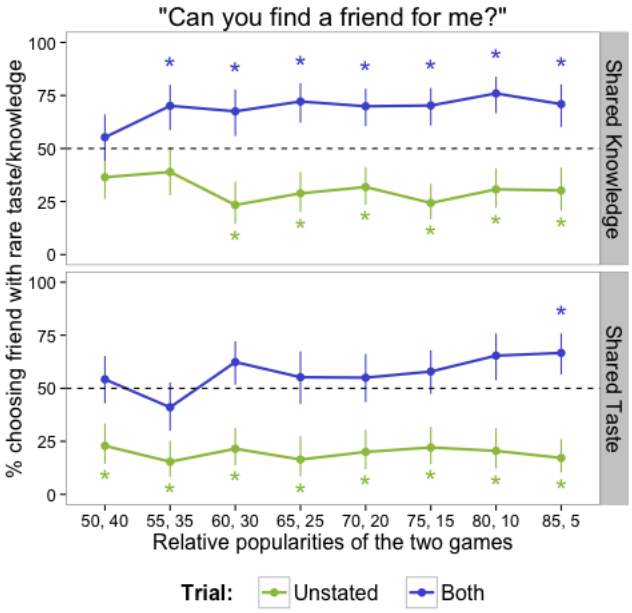


Figure 3: Experiment 2b results. Percentage of participants in the SK (top) and ST (bottom) groups who chose the friend who liked/knew about the rare game.

on social affiliation than overlaps in relatively common preferences. In Experiment 1, we found that participants preferred to affiliate with agents who shared a rare preference with them, based on their own estimates of how widely their preferences were shared among the general population. In Experiment 2, we systematically manipulated the prevalence of preferences in a novel population and asked participants to make third-party judgments. Using this complementary task, we again found that participants preferentially matched agents with friends who shared the rarer preference in games. Indeed, this was true only if the target agent had a preference for both games; in the absence of information about his preferences, participants recommended the friend with the more common preference. Finally, across both experiments, we found that rare overlaps in *knowledge* were particularly attractive, compared to overlaps in taste without prior knowledge.

Why do we find rare overlaps in preferences significant and indicative of social affiliation? Encountering someone who shares a preference with you may support a statistical inference about the causal mechanisms leading to this coincidence (see Griffiths & Tenenbaum, 2007); perhaps overlaps in rare preferences signal stronger, deeper similarities in values (Boer et al., 2011), personality traits (Rentfrow & Gosling, 2006), and cultural knowledge (Soley & Spelke, 2016). Even though overlaps in common preferences might also reflect similarities in particular domains, they may be less diagnostic of a shared social history, as such preferences

span a larger, potentially more heterogeneous, set of people. For instance, a wide range of social experiences might lead people to know about (and like) iPhones, while a much more selective and distinctive set of experiences and values might lead people to know about (and like) programming in Raspberry Pi. Gershman and Gweon (under review) propose that we might use information about others' preferences to infer an underlying network of social connectivity between ourselves and others. Under this framework, finding someone who shares a rare preference might indicate a tighter social connection than finding someone with a common overlap.

Two caveats should accompany our findings. First, prior studies showed that U.S. adults tend to prefer unique (rare) items in a uniform group of objects compared to East Asians (Kim & Markus, 1999), suggesting that the degree to which item selection reflects individuals' preferences might depend on the shared cultural values. Thus we acknowledge that the effect of rarity might be culturally variable. Second, while we tried to separate knowledge from liking in our Shared Knowledge tasks, we cannot entirely rule out the effect of liking. The agents expressed knowledge without any references to liking the target items; however participants might have assumed that someone who says "I know about X!" is communicating at least a slight preference for X. The effect of knowledge might be weaker if agents expressed an explicit indifference, and even weaker if they expressed a dispreference. Future work should explore these possibilities.

Our results raise additional interesting questions for future work, including how people draw estimates about the prevalence of preferences, and what factors might bias these estimates. Social psychology research has shown that adults' estimates of the frequencies of different attitudes in a population are self-serving, and are typically biased to justify one's own choices (e.g., Monin & Norton, 2003). If rarity is an influential cue for social affiliation, we might also exaggerate or underestimate the rarity of our preferences to make certain overlaps appear more or less meaningful. Such biases are easily imagined in situations where people try to establish a bond with another person. Consider someone on a first date on the lookout for similarities, exclaiming: "You like pizza! What a coincidence, I do too!" By (intentionally or unintentionally) misconstruing an obviously common preference as rare, one might insinuate (of convince oneself of) the possibility that there must be many other overlaps yet to be discovered. We look forward to further testing these intuitions.

In sum, our study shows how our real-world knowledge (e.g., prevalence of traits, preferences) might underlie our intuitive sense of what interpersonal similarities are meaningful, and provide compelling evidence for the idea that shared rare preferences are strong cues for social compatibility. By combining intuitions from social psychology and cognitive science, these results provide new insights into the hidden social structure underlying shared preferences.

Acknowledgments

We are grateful to Alex Shaw, Zoe Liberman, and Josh Tenenbaum for fun, helpful discussions about this work. This research was supported by NSF Graduate Research Fellowships to S.B. and N.V., and the Varieties of Understanding grant from the John Templeton Foundation to H.G.

References

Boer, D., Fischer, R., Strack, M., Bond, M. H., Lo, E., & Lam, J. (2011). How shared preferences in music create bonds between people: Values as the missing link. *Personality and Social Psychology Bulletin*, 1-13.

Fawcett, C. A., & Markson, L. (2010). Children reason about shared preferences. *Developmental Psychology*, *46*, 299-309.

Gershman, S., & Gweon, H. (under review). Learning the structure of social influence.

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*, 180-226.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*, 9066-9071.

Kim, H., & Markus, H. R. (1999). Deviance or uniqueness, harmony or conformity? a cultural analysis. *Journal of Personality and Social Psychology*, *77*, 785-800.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*, 1134-1140.

Monin, B., & Norton, M. I. (2003). Perception of a fluid consensus: Uniqueness bias, false consensus, false polarization, and pluralistic ignorance in a water conservation crisis. *Personality and Social Psychology*, *29*, 559-567.

Rentfrow, P. J., & Gosling, S. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*, 1236.

Rentfrow, P. J., & Gosling, S. D. (2006). Message in a ballad: The role of music preferences in interpersonal perception. *Psychological Science*, *17*, 236-242.

Selfhout, M. H. W., Branje, S. J. T., ter Bogt, T. F. M., & Meeus, W. H. J. (2009). The role of music preferences in early adolescents' friendship formation and stability. *Journal of Adolescence*, *32*, 95-107.

Soley, G., & Spelke, E. S. (2016). Shared cultural knowledge: Effects of music on young children's social preferences. *Cognition*, *148*, 106-116.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*, 97-104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 5012-5015.

What Causal Illusions Might Tell us about the Identification of Causes

Robert Thorstad (rthorst@emory.edu)
Phillip Wolff (pwolff@emory.edu)
Emory University Department of Psychology

Abstract

According to existing accounts of causation, people rely on a single criterion to identify the cause of an event. The phenomenon of causal illusions raises problems for such views. Causal illusions arise when a particular factor is perceived to be causal despite knowledge indicating otherwise. According to what we will call the Dual-Process Hypothesis of Causal Identification, identifying a cause involves two cognitive processes: 1) an automatic, intuitive process that identifies possible causes on the basis of perceptual cues (spatial and temporal) and 2) a slow, reflective process that identifies possible causes on the basis of causal inference, in particular, a consideration of possible mechanism. Consistent with this hypothesis, we found that in response to a causal illusion shown in a naturalistic setting, people's initial judgments of causation were higher than their ultimate judgments of causation (Experiment 1). Using an online measure of the time-course of people's causal judgments, we found that people initially view animations of causal illusions as causal before concluding that they are non-causal (Experiment 2). Finally, we obtained similar results using a deadline procedure, while also finding that the lower the cognitive reflectiveness (as measured by the CRT), the stronger people's impressions of causation were (Experiment 3). Implications for different classes of theories of causation are discussed.

Keywords: causal reasoning; thinking & reasoning

Introduction

In 1977 New York City experienced a major blackout. Remarkably, some individuals felt, at least momentarily, that the blackout was caused by their own actions. For instance, Sparrow (1999) reports a child who hit a ceiling light fixture with a paddle ball at the exact moment the lights went off, and an opera singer who touched a door just as the power went out. One person exclaimed after plugging in a toaster, "I blew out the whole neighborhood!" (Sparrow, 1999). In situations such as this, people may experience strong feelings of causation while at the same time knowing that such feelings are unwarranted. We will term this experience of causation absent a plausible mechanism a *causal illusion*.

The fact that people can have conflicting judgments about the existence of a causal relationship is consistent with the idea that judgments of causation may be based on two kinds of processes: 1) a fast and intuitive process that identifies potential causes on the basis of perceptual cues, temporal cues in particular, and 2) a slow and reflective process that identifies potential causes on the basis of causal mechanisms, and in particular, how the entities in an event might be spatially arranged in order to allow for the transmission of energy or force (Wolff & Shepard, 2013). According to what we will call the *Dual-Process Hypothesis of Causal Identification*, these two processes occur regularly in people's analyses of everyday events. The two processes may not be easily recognized as distinct processes because in most cases they lead to the same conclusion. In the case of causal illusions, the conclusions of the two processes diverge, and hence their presence is revealed.

The distinction between intuitive and reflective processes is not new. Various forms of this distinction can be found in the perception, reasoning, and social cognition literatures. According to dual-processing theories, System 1 represents the statistical structure of the environment in a sub-symbolic format and involves processes that are implicit, unconscious, and heuristic, while System 2 represents the environment in a symbolic format and involves processes that are explicit, sequential, and rule-based (Sloman, 2015; Evans, 2008; Kahneman, 2003; Stanovich & West, 2000). In the causation literature, the distinction between intuitive and reflective processes is implied in the work of Schlottmann and Shanks (1992), who proposed a distinction between perceived and judged causation. This distinction is also indirectly suggested by the existence of a major partitioning of the causation literature into two areas: research on the perception of causality, which concerns processes that operate independently of background knowledge (e.g., Hubbard, 2013; Michotte, 1963; Rips, 2011; Scholl & Tremoulet, 2000; White, 2006) and research on causal reasoning and learning that involves the use of prior knowledge in the creation of new causal relations (e.g., Ahn et al, 1995; Lien & Cheng, 2000; Goldvarg & Johnson-Laird, 2001; Mayrhofer and Waldmann, 2015; Sloman & Lagnado, 2015; Wolff & Barbey, 2015). Implicit in the causation literature is the view that different kinds of causal reasoning might occur in the intuitive and reflective systems.

According to Sloman (2015) the intuitive system does not represent statistical associations directly, but rather generates such associations though the representation of causal structure. According to Sloman, the intuitive system is capable of making relatively sophisticated distinctions, such as whether an outcome was merely observed or the result of an intervention. In Sloman's (2015) proposal, the reflective system is slower than the intuitive system and capable of exerting some control over the intuitive system, though not completely. When in conflict, people will usually chose the conclusion generated from the reflective system over that generated by the intuitive system. In Sloman's (2015) view, the intuitive system is impressive in how much causal reasoning it is able to perform correctly.



Figure 1: An example causal illusion. A man appears to open elevator doors using only his hands.

We do not necessarily disagree with Sloman's (2015) proposal on how causal reasoning might be instantiated in the intuitive and reflective systems. Like Sloman (2015), we see the intuitive system as able to perform certain types of causal reasoning. Where we may disagree is in the relative level of sophistication of the reasoning in the intuitive system. Causal illusions, we propose, occur in situations in which the intuitive system is getting things wrong, and we speculate that such illusions are relatively common.

To date, however, no studies have directly examined the unique contributions made by the intuitive and reflective systems to the identification of a cause. According to our Dual-Process Hypothesis, the intuitive and reflective systems are expected to come to different conclusions about a cause in situations in which perceptual cues conflict with inferences from background knowledge. In particular, conclusions generated by the intuitive system are expected to occur earlier in the time-course of processing than conclusions generated by the reflective system. This basic prediction was tested in the following three experiments.

Experiment 1: "Jedi Powers"

In Experiment 1, we aimed to establish the phenomenon of causal illusions, and to begin to investigate the time-course of causal identification. In this study participants experienced an unexpected causal illusion on their way to our lab: a man appeared to open an elevator door by merely gesturing with his hands (Figure 1). Importantly, the man made no physical contact with the doors or with any of the buttons in the elevator. Unbeknownst to the participants, the doors were opened by a confederate outside of the elevator pushing the elevator button. Our main prediction was that the intuitive system would lead to feelings of causation that would ultimately be reduced by the reflective system. Participants' impressions of causation were measured in the subsequent interview, which included a short questionnaire.

Methods

Participants. 23 undergraduate participants were recruited for a study on perception and tested in groups of 1-4. One participant was excluded because they personally knew the confederate.

Causal Illusion. Participants were informed that the study would take place in our lab on another floor and followed a research assistant to the building elevator. There, a confederate pretended to re-open the elevator doors using only his hands (see Figure 1). The confederate re-opened the doors a total of three times, and maintained a neutral expression. In reality, the doors were controlled by an unseen confederate; no participants reported discovering the unseen confederate.

Causality Ratings. After the causal illusion, participants completed two sets of written ratings. (A) *Causality Description*: participants answered three questions about their experience: "what do you think you saw?," "please describe how this impression unfolded over time," and "did you think the man caused the elevator doors to open?" (B) *Time-Course Ratings*: participants rated the following statements on a 1-5 Likert scale: "to what extent did you [feel for a moment / ultimately conclude] that the man caused the elevator doors to open?"

Causality Coding. Two raters unfamiliar with the experiment coded participants' written descriptions for two features. (A) *Link Rating*: did the written descriptions mention a link between the man and the doors opening? (B) *Causality Rating*: did the written descriptions attribute causality to the man in causing the doors to open?

Results

Causality Descriptions. If causal illusions are in part causal, participants should spontaneously attribute causality to the Jedi in their causal descriptions. Participants' written descriptions frequently attributed causality to the confederate. Example descriptions were that "A man was controlling the doors of the elevator with his hands," "The man in the elevator kept causing the door to stay open on the wrong floor, like magic," and "A man […] was able to open the doors simply by moving his hands." Coders rated a mean of 87% of written descriptions as causal (inter-rater reliability=89%, kappa=0.33), suggesting that participants spontaneously perceived causal illusions as causal.

Time-Course Ratings. According to the dual-process hypothesis, causal illusions should create a strong *initial* impression of causality, but weaker *subsequent* impression of causality, due to the conflict between intuitive and reflective systems. Participants' time-course ratings supported this account: participants rated a stronger initial than ultimate impression of causality, $t(21)=3.72$, $p<0.01$ (Figure 2). This difference in causality ratings was consistently observed in individual participants. A majority of participants (13/22) rated a stronger initial than ultimate impression of causality, while just 2 participants rated a stronger ultimate than initial impression.

Because the number of participants varied from trial to trial, we conducted an additional analysis to ensure that participants' ratings were not driven by the impressions of other participants. There was no effect of the number of participants on either participants' initial impression of causality, $F(3,18)=1.10$, $p=0.37$, or on their ultimate impression of causality, $F(3,18) < 1$, n.s..

Discussion

There are two main results of Experiment 1. First, the results show that causal illusions give rise to impressions of causality, even in the absence of a possible mechanism. The majority of participants' written descriptions attributed causality to the confederate, and causal illusions received high initial ratings of causality. Second, the results provide initial support for the predictions of the dual-process hypothesis. Causal illusions created a stronger initial than ultimate impression of causality, consistent with a conflict between a fast intuitive system using perceptual cues and a slower reflective system using background knowledge. This effect was consistent across participants and rarely occurred in the other direction.

A limitation of Experiment 1 is that we were only able to measure participants' impressions of causation well after the occurrence of the event, rather than as they actually unfolded in real-time. The results are vulnerable, then, to the possibility that participants really did not have a strong initial impression of causation, but merely attributed this impression after the fact. In Experiment 2, this limitation was addressed through the use of a real-time measure of participants' impressions of causation.

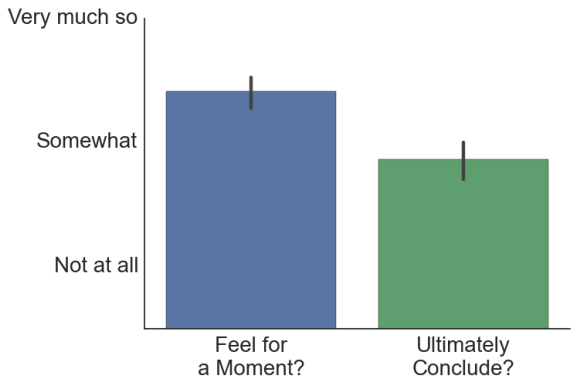


Figure 2: Participants rated a stronger initial than subsequent impression of causality when experiencing causal illusions. Error bars +/- 1 SEM.

Experiment 2: Time Course

In Experiment 2, participants viewed three kinds of events: causal, non-causal, and causal illusions (see Figure 3A). The temporal unfolding of the three kinds of events was exactly the same, except for the position of the causer in the scene. For example, in one set of animations, participants saw a record begin to turn. In the causal version, the hand made physical contact with the record. In the non-causal version of the event, the hand did not move. In the causal illusion version of the event, the hand moved, but did not make physical contact with the record. The experiment included six different sets of animations. As participants watched an animation, they indicated the degree to which they felt the event was causal or non-causal by how far they turned a dial to the right (causal) or left (non-causal) (see Figure 3D). Participants could change how far they turned the response dial at any point during the presentation of the animation. The dual-process hypothesis predicts that for the causal event, participants should move and keep the dial in the cause direction. For the non-causal event, participants should move and keep the dial in the non-causal direction. Critically, for the causal illusions, the dual-process hypothesis predicts that participants should *initially* move the response dial to cause, but *later* move the response dial to non-cause.

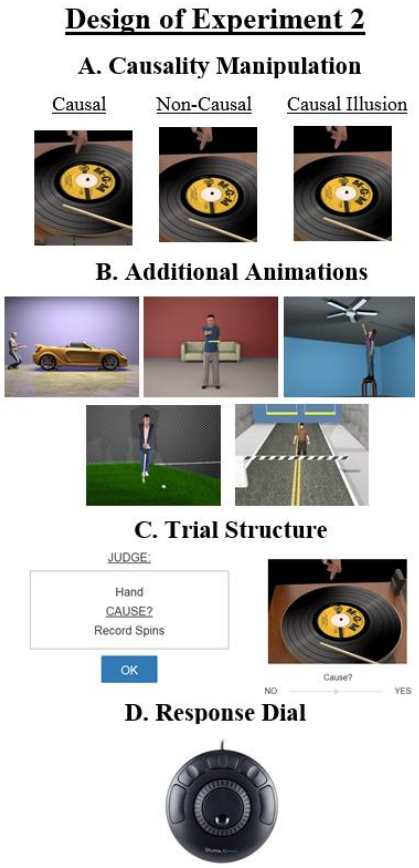


Figure 3**:** Experimental design for Experiment 2.

Methods

Participants. 60 adult participants were tested individually in the lab. 3 participants were excluded for failing to respond to at least 2 animations.

Animations. Animations were constructed using 3D Studio MAX animation software and rendered in Mental Ray or VRAY to increase realism. Six experimental animations were created (Figure 3A,B); each animation had a causal, non-causal, and causal illusion variant. All animations were

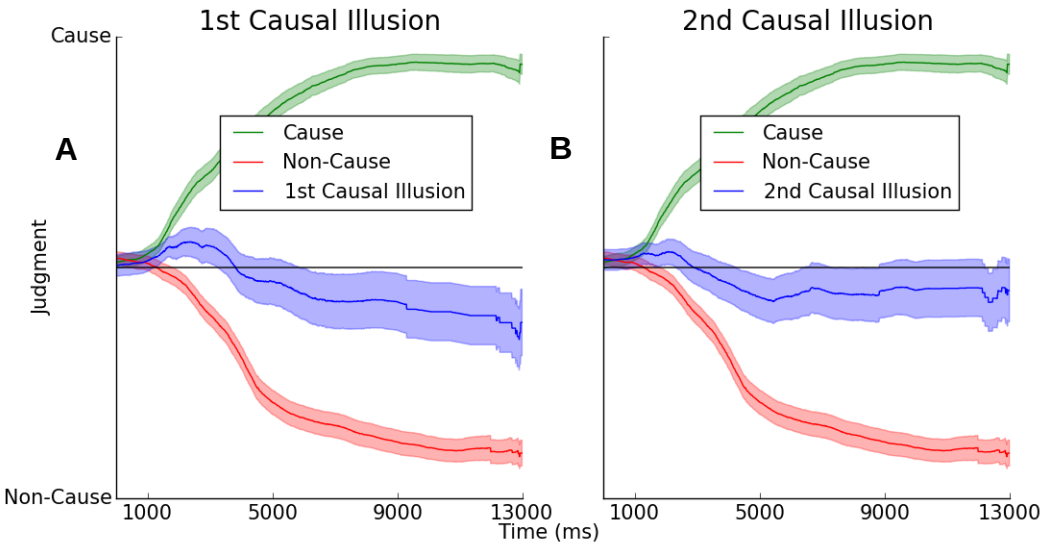


Figure 4: The average decision function for causal, non-causal, and causal illusion animations (shading +/- 1 SEM). Causal illusions are plotted separately for the 1st and 2nd causal illusions in Figure 4A and 4B, respectively.

matched at the single-frame level for duration (7s) and for when the resulting effect occurred (4s). Animations were displayed at a rate of 30 frames/s. An additional 5s of a still first frame was added to the beginning of the animations to allow participants to orient to the scene. 6s of a still last frame was also added to the end of the animation to allow the causal impression additional time to unfold. Six practice animations were created with clear cases of causation and non-causation to allow participants to become comfortable with the procedures in the experiment.

Trial Structure. The experiment was conducted on lab computers using a javascript application. Participants first completed 6 practice trials with clear causal or non-causal animations. Before the animation played, one of the objects in the scenes was named. Participants were instructed to evaluate the causality with respect to this object (see Figure 3C). Participants then viewed the animation, and were instructed to use a response dial to judge causality in real time. Participants responded "yes" or "no" by moving the dial right and left, respectively. Dial position was recorded as a continuous variable from +15 (yes) to -15 (no) every 10 ms. Participants were instructed that the position of the dial should always reflect their current opinion of whether a particular event was caused by the named object.

Data Analysis. Data analysis was conducted using custom Python scripts. For each participant, we averaged the responses to causal, non-causal, and causal illusion animations separately to create a single average decision for each animation type. We considered only data recorded at or after the time the effect occurred, resulting in 3 separate 13s-long decision functions, 1 for each animation type, with position recorded every 10 ms.

Results

The average decision function for each type of animation is shown in Figure 4. Significance tests were conducted by binning decisions at the group level using 100 ms bins, and testing whether the mean of that bin (which could range from +15 to -15) differed from 0 (i.e., no decision), using single-sample t-tests, α=0.05, 2-tailed.

As expected, causal animations were quickly judged as causal, with decisions differing significantly from 0 beginning 800 ms after the effect occurred. Non-causal animations were quickly judged as non-causal, with decisions differing significantly from 0 beginning 1,200 ms after the effect occurred. Critically, the predictions of the dual-process hypothesis were also supported for the causal illusions, provided the analyses are restricted to only the first causal illusion participants saw. In particular, for the 1st causal illusion, as seen in Figure 4A, causal illusions were *initially* judged as causal 1,160 ms after the effect occurred, and then later judged noncausal at 12,700 ms after the effect occurred. The results from all of the causal illusions show a similar pattern, but are not significantly different from 0. The results suggest that participants viewed the animations in a more reflective manner once they saw one of the causal illusions. In particular, participants may have been able to use remembered knowledge about the 1st causal illusion to suppress the causal impression of the 2nd causal illusion. This result is consistent with the dual-process hypothesis, since prior inferences may have had a chance to interact with perceptually-based judgments.

Discussion

The main predictions of the dual-process hypothesis were supported. For causal illusions, but not for clear causes or non-causes, participants made two opposite decisions over the course of the single trial. Participants initially judged the 1st causal illusion to be causal, but subsequently judged it to be non-causal. In addition to providing evidence for the

existence two kinds of processes, the results from this experiment provide some indication about when in time the two kinds of causal processes take place. In particular, it appears that the results of the intuitive process become available at around about 1000 ms, whereas the results from the reflective process might not be felt until 1300 ms. With these temporal benchmarks in place, we can test the dual-process hypothesis using a standard methods for examining the temporal time-line of processing, as demonstrated in the next experiment.



Figure 5: The deadline procedure in Experiment 3. Participants made a yes/no decision about causation under either a short or long response deadline.

Experiment 3: Thinking Styles

Experiment 3 had two aims. First, we sought to provide converging evidence for the existence of multiple processes in the interpretation of causal events using a deadline methodology, as well as by examining whether people's impressions of causation might depend on their cognitive reflectivity.

As in the previous experiment, participants viewed causal, non-causal, and causal illusion animations. In the current experiments, however, participants were prompted to make a yes/no decision about causation after either a short or long response deadline (Figure 5). The dual-process hypothesis predicts that participants will be more likely to judge causal illusions as causal under a short than long response deadline.

In addition to completing a deadline procedure, participants also completed a measure of reflective thinking style, specifically the Cognitive Reflection Test (CRT7; Toplak et al, 2014). If heuristic processes are responsible for the impression that causal illusions are causal, than participants higher in cognitive reflectiveness may be less susceptible to causal illusions.

Methods

Participants. 51 adult participants were tested individually in the lab.

Stimuli. Stimuli were the animations used in Experiment 2, as well as 4 new animations for a total of 10 different animations, each with a causal, non-causal, and causal illusion variant. The new animations were added to ensure results are not specific to items from Experiment 2.

Deadline procedure. The deadline procedure is outlined in Figure 5. Participants were first prompted which events to judge. Participants were then presented with an animation with the same time-course as in Experiment 2, except that the animation terminated 1s after the effect occurred. Participants then used the left and right arrow keys arrow keys to judge (yes/no) whether the first event caused the second event. Half of participants (*short-deadline* group) had 1s to respond; half of participants (*long-deadline* group) had 10s to respond. Response deadline was manipulated between participants because of possible order effects discovered in Experiment 2. All participants saw each of 30 animations twice in random order for a total of 60 trials.

Individual Differences Measures. After the deadline task, participants completed other tasks in our lab, and then completed the revised 7-item version of the Cognitive Reflection Test (CRT-7; Toplak et al, 2014).

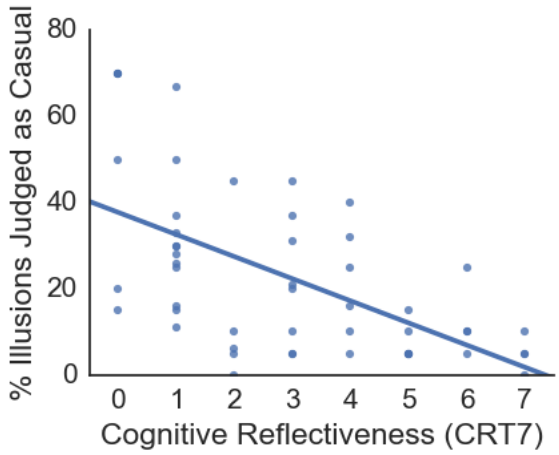


Figure 6: Participants high in cognitive reflectiveness were less likely to endorse causal illusions as causal.

Results

Just as predicted by the dual-process hypothesis, ratings of causation were higher for causal illusions when the deadline was short (M = 25.0%) than long (M = 18.2%), $t(9) = 3.76$, $p < 0.01$. In contrast, ratings of causation did not differ between the two deadlines for the animations showing clear cases of causation, $t(9) = 1.33$, $p = 0.215$, or non-causation, $t(9) = 0.27$, $p = 0.793$.

An analysis of participants' individual differences provided further support for the dual-process hypothesis. In these analyses, we correlated participants' CRT7 scores with the percent of times participants judged causal illusions as causal. As shown in Figure 6, cognitive reflectiveness correlated negatively with ratings of causation, $r(50) = -0.579$, $p < 0.001$. The effect was specific to causal illusions: cognitive reflectiveness was unrelated to judgments of causation in response to the causal, $r(50) = 0.03$, $p = 0.036$, and non-causal stimuli, $r(50) = -0.247$, $p = 0.081$. This pattern of results suggests that cognitive reflectiveness is a strong predictor of resistance to causal illusions.

General Discussion

In a series of three experiments, the predictions of the dual-process hypothesis were supported. In Experiment 1, we documented the phenomenon of causal illusions outside the lab. We showed that participants had a strong initial

impression of causation, but a weaker ultimate impression of causation, supporting the claim that two cognitive processes are involved in causal reasoning. In Experiment 2, we replicated this dual pattern of response using a novel online measure of decision-making. Additionally, we showed an order effect, where participants were most susceptible to causal illusions for the very first illusion viewed. Finally, in Experiment 3, we demonstrated the same dual pattern of response using a deadline procedure. Additionally, we showed individual differences in susceptibility to causal illusions, where individuals high in reflective thinking were less susceptible to causal illusions.

An alternative interpretation of our results is that the intuitive process really processes dependency, not perceptual cues (Cheng, 1993). On this account, participants may judge a contingency between sub-events in a causal illusion: that the effect is more likely to occur after rather than before the causer moves. However, this interpretation does not easily account for our finding that the impression of causation declines over time for causal illusions (Experiments 1 and 2). With repeated presentations of a causal illusion, contingency should remain constant or increase; however, participants' judgments of causation decreased with repeated presentations.

The results have implications for theories of causal identification. As noted by Sloman and Lagnado (2015) and Copley and Wolff (2014), theories of how people identify causes fall into two categories: 1) Dependency theories which characterize causal relations with respect to statistical, logical, or counterfactual dependency relations, and 2) Process theories which characterize causal relations with respect to the notions transmission and force. The results reported in this paper might be viewed as problematic for all of these theories, because these theories define causation with respect to only one process. On the other hand, the results from this paper might be viewed as an opportunity. It may be that the different processes implies in this paper point to the need for multiple theories of causation. As suggested by Sloman (2015) dependency theories might be well suited for explain intuitive processing, whereas as suggested by Wolff and Shepard (2013), process theories might be well suited for reflective processes used in the identification of a mechanism that allows for the transmission of energy or force. An examination of the time-course of the processing of dependency relations and of forces may provide some resolution to the feasibility of a pluralistic theory of causation.

Acknowledgments

This research was supported by NSF grant #134088 to the second author. The authors also with to thank Yiwei Gao for programming assistance and Jason Shepard for helpful comments on the project.

References

Ahn, W., Kalish, C.., Medin, D., & Gelman, S. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352.

Cheng, P. (1993) Separating causal laws from causal facts: pressing the limits of statistical relevance. In D. Medin (Ed.), *The psychology of learning and motivation* (pp. 215-264). San Diego, CA: Academic Press.

Copley, B. & Wolff, P. (2014). Theories of causation can and should inform linguistic theory. In. B. Copley, F. Martin, & N. Duffield (Eds.), *Causation in grammatical structures* (pp. 11-56). Oxford: Oxford UP.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278.

Goldvarg, E. & Johnson-Laird, P. (2001). Naïve causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565-610.

Hubbard, T. (2013). Phenomenal causality I: varieties and variables. *Axiomathes*, *23*, 1-42.

Kahneman, D. (2003). A perspective on judgment and choice. *American Psychologist, 58*, 697 – 720.

Lien, Y., & Cheng, P. (2000). Distinguishing genuine from spurious causes: a coherence hypothesis. *Cognitive Psychology, 40*, 87 – 137.

Mayrhofer, R. & Waldmann, M. (2015). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science*. Advance online publication.

Michotte, A. (1963). New York: Basic Books.

Rips, L. J. (2011). Causation from perception. *Perspectives on Psychological Science*, *6*, 77–97.

Schlottmann, A. & Shanks, D. (1992). Evidence for a distinction between judged and perceived causality. *Quarterly JEP*, *44A*, 321-342.

Scholl, B. & Tremoulet, T. (2000). Perceptual Causality and Animacy. *TICS*, *4*(8), 299-309.

Sloman, S. (2015). Two systems of reasoning, an update. In J. Sherman, B. Gawronski, & Y Trope (Eds.), *Dual process theories of the social mind*. Guilford Press.

Sloman, S. & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*, 223-247.

Sparrow, J. (1999). The blackout history project. Retrieved from http://sloan.stanford.edu/SloanConference/papers/.

Stanovich, K. & West, R. (2000). Individual differences in reasoning: implications for the rationality debate. *Behavioral Brain Sciences*, *23*, 645–726.

Toplak, M., West, R., & Stanovich, K. (2014). Assessing miserly information processing: an expansion of the CRT. *Thinking & Reasoning*, *20*(2), 147-168.

White, P. (2006). The role of activity in visual impressions of causality. *Acta Psychologica, 123*, 166–185.

Wolff, P., & Barbey, A. (2015). Causal reasoning with forces. *Frontiers in Human Neuroscience, 9*.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. *Psychology of Learning and Motivation, 58*, 167 – 202.

Quantifying Joint Activities using Cross-Recurrence Block Representation

Tian Xu (txu@indiana.edu)
Computer Science and Cognitive Science Department, 1101 East 10th Street Indiana University
Bloomington, IN, 47405
Chen Yu (chenyu@indiana.edu)
Psychological & Brain Sciences, Cognitive Sciences, and informatics Department, 1101 East 10th Street Indiana University
Bloomington, IN, 47405

Abstract

Humans, as social beings, are capable of employing various behavioral cues, such as gaze, speech, manual action, and body posture, in everyday communication. However, to extract fine-grained interaction patterns in social contexts has been presented with methodological challenges. Cross-Recurrence Plot Quantification Analysis (CRQA) is an analysis method invented in theoretical physics and recently applied to cognitive science to study interpersonal coordination. In this paper, we extend this approach to analyzing joint activities in child-parent interaction. We define a new representation as Cross Recurrence Block based on CRQA. With this representation, we are able to capture interpersonal dynamics from more than two behavioral streams in one Cross Recurrence Plot and derive a suite of measures to quantify detailed characteristics of coordination. Using a dataset collected from a child-parent interaction study, we show that these quantitative measures of joint activities reveal developmental changes in coordinative behavioral patterns between children and parents.

Keywords: cross-recurrence quantification analysis; child parent interaction; multimodal behavioral data; interpersonal interaction; hand coordination; statistical methods

Introduction

Humans are social animals and coordination is the foundation of everyday social interactions (Kendon, 1970). We exhibit a remarkable ability to coordinate our behaviors with our social partners to achieve common goals at different cognitive levels and time scales, from motor physiological coordination, such as dancing (Kimmel, 2012), to goal-oriented dialogue and collective task completion (Fusaroli and Tylén, 2015). During social interaction, the dynamics between interactors evolves and escalates over time in a complex way. To understand how information is exchanged and communication is unfolded through behavioral cues, we need to examine the characteristics of coordination in great details.

Recently, technological advances in sensing and computing devices allow us to collect high-density and large-volume behavioral data to study interpersonal coordination (Dale et al., 2011). However, due to the complexities and stochastic nature of human behaviors, new ways in data collection have also presented data-mining challenges to cognitive scientists (Fusaroli et al., 2014) – with a large amount of data collected, how we can effectively discover novel and reliable patterns to advance our understanding of human coordinated behaviors. For example, in a dyadic interaction, two social partners most often communicate via several expressive channels, such as language, facial expression, body movements, and manual actions. Those intrapersonal and interpersonal behaviors happen in fractions of a second, and together form a dynamically interactive loop to continuously influence each other at every moment (Louwerse et al., 2012).

The primary goal of the present study is to introduce a new method to quantify engagement and joint activities in social interaction. The method is built upon Cross Recurrence Plot Quantification Analysis (CRQA), which has been introduced in cognitive science to analyze synchronization and alignment in interpersonal coordination (Riley et al., 2011). The proposed method extends classic CRQA approaches by using a new representation of joint events: Cross Recurrence Block. Multiple behavioral measures can be derived from this new representation, revealing fine-grained quantitative patterns that cannot be detected with standard statistical measures of individual and joint behaviors.

Cross Recurrence Plot Analysis

Recurrence plot Quantification Analysis (RQA) and Cross-Recurrence plot based Quantification Analysis (CRQA) have been proposed as nonlinear methods that provide informative data visualization and rigorous quantification of dynamical systems and their trajectories with which many properties would otherwise be lost due to averaging with traditional correlation analysis (Zbilut et al., 1998; Marwan et al., 2007). In particular, CRQA is able to reveal and describe the shared dynamic trajectory between two different data series by constructing the states that both systems visit over time. While first introduced as a form of generalized cross-correlation between two continuous data streams, CRQA has been extended to nominal time series with categorical values using a radius of zero in computing state matches between two systems (Coco and Dale, 2014). These methods have been widely used in many areas, such as the analysis of pattern and rhythm in sound and music (Cooper and Foote, 2002), visual search scan pattern analysis (Anderson et al., 2012) and discourse conceptual structure analysis in doctor-patient conversations (Angus et al., 2012). Despite the analytical power of CRQA as a non-linear method for extracting inter-system dynamics, the process of constructing Cross Recurrence Plot (CRP) can be relatively simple and straightforward. In the next section, we will introduce the plot

construction process and CRQA with the empirical dataset from a parent-child toy play interaction study.

Generating Cross-Recurrence Plots from Data

Figure 1 shows both the experimental setup and multimodal data streams collected in our study. We invited 24 dyads of children and their parents to the lab (12 children at 12 months old age, 12 children at 24 months old age). During the experiment, parents and children played with two sets of toys in free-flowing interaction for about 6 minutes (Yu, Smith, 2013). The toys were painted into single colors (blue, green or red) to facilitate automatic image processing. We recorded the entire interaction from multiple cameras, including both the child's and parent's first-person viewpoints, a third-person view, and a bird-eye view. The present study focuses on manual activities generated by parents and children in toy play. Human coders went through the videos from multiple viewpoints and manually annotated frame-by-frame about which object was held by the child and the parent with each of their hands. As shown in Figure 1(b), holding actions of both the child and parent were represented as four categorical temporal data streams with different color indicating different Region-Of-Interest (ROI) in each moment. The

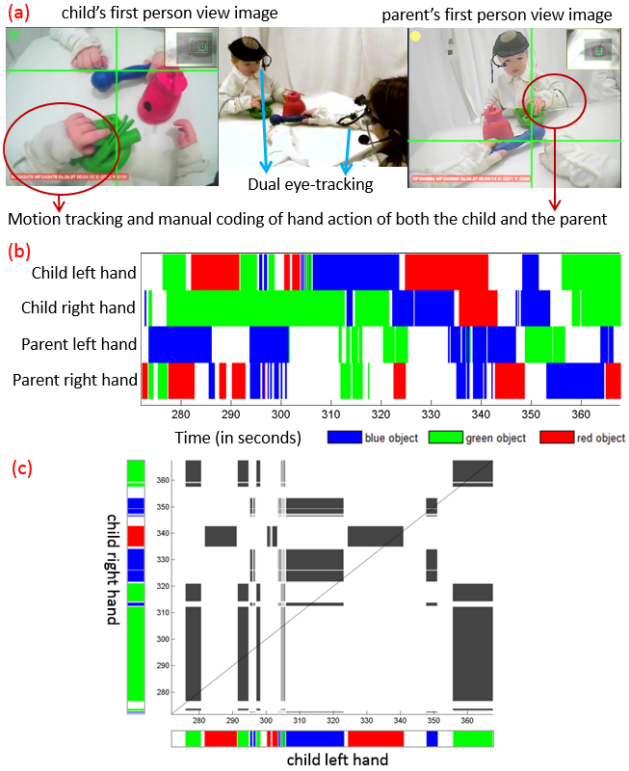


Figure 1: (a) Our multisensory experimental set-up wherein the child and parent engaged in free object-play interaction. Both child's and parent's gaze, holding action and speech events were coded into nominal temporal data streams; (b) Coded manual holding action behavioral streams during one experimental trial; (c) A Cross Recurrence Plots constructed using a child's two hand action streams with the data shown in (b).

sampling rate of all four time series was 30Hz. Each data point in a time series has been assigned with one of the four possible ROI values: 1-3 (blue, green, red) indicating the target object held by one of the interactors, and 0 (white) meaning empty hand.

Figure 1(c) shows a Cross Recurrence Plot (CRP) constructed with the child's left hand holding data stream on the x-axis and the child's right hand holding data stream on y-axis. A CRP is basically a M×N matrix, in which N is the number of data points of the temporal stream on x-axis and M is the number of data points on y-axis. For any cell in the matrix at row i column j, that point $\boldsymbol{R}_{i,j}$ in CRP is assigned to 1 if the child's right hand at time t_i and the child's left hand at time t_j were both holding the same object; $\boldsymbol{R}_{i,j}$ will be 0 otherwise (Marwan et al., 2007).

With standard CRQA, we can calculate various measurements such as recurrence rate, determinism, etc. (Zbilut et al., 1998). For example, recurrence rate calculates the density of recurrence plots that reflects the degree of coordination and coupling between the behavioral streams on x-axis and y-axis. In Richardson and Dale (2005), the authors calculated the recurrence rate within a 30-second temporal window to reflect the coordinated gaze behaviors between the speaker and listener; In Yu and Smith (2013), the authors calculated correlation scores with different time delays and recurrence rates within multiple temporal windows to reveal how eye-hand coordination between child and parent influences the child's visual attention towards objects.

Cross Recurrence Block Representation

In standard CRP, $\boldsymbol{R}_{i,j}$ can be either 1 (the behavior of the interactor on x-axis at time t_i and his partner on y-axis at time t_j fall in the same category), or 0 (no match in their behaviors). In such cases, a plot is usually visualized using two colors, black and white (Marwan et al., 2007). The first contribution of our new method here is to extend the value range of $\boldsymbol{R}_{i,j}$ to include the different behavioral categories; and as a result, the CRP becomes colorful – each color depicts not only a match between two data points from two time series at a certain point in time but also the exact behavioral category that is matched. Figure 2 shows the colored version of CRP constructed with child's holding data shown in Figure 1(c). In this plot, we used held objects as behavioral categories.

Traditionally, the data structure of Cross Recurrence Plot is represented as a matrix composed of individual points and continuous lines. Therefore, quantitative measurements derived are based on those two representations. As illustrated in Figure 2, a set of temporally adjacent recurrence points form a block (labeled as $Block_A$ with green color). When we look back to the original data streams recording the held objects in the child's two hands (shown at the bottom of Figure 2), one can easily see that $Block_A$ is the product of a sequence of two individual holding events: the child was holding the green object from 276s to 280.5s in his left hand, and his right hand also grabbed the same object at 276.8s and continued to hold it much longer until 312.2s. One can

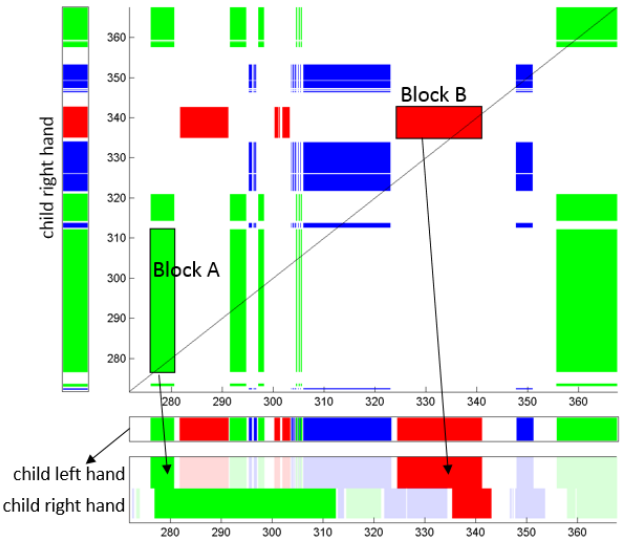


Figure 2: Cross Recurrence Plot with added color information indicating the target object of each joint activity sequence.

intuitively see that $Block_A$ captures a sequential pattern of two manual actions. Similarly, $Block_B$ represents another pair of sequential holding events: the child was playing with the red object using their left hand first from 324.3s to 340.8s, then their right hand let go of the blue object to jointly hold the same red object from 335.1s to 343.6s.

As illustrated by the two examples above, we propose a new representation based on CRP, viewing and grouping temporally adjacent points as blocks with width and height. Such blocks represent joint activities that can be directly mapped back to temporal events in the behavioral data streams used to construct a cross recurrence plot. Formally, we define such blocks as **Cross Recurrence Block (CRB)** that are made of a set of points in a CRP sharing the same categorical value and being temporally adjacent to each other. As shown in Figure 2, a CRB can vary in size and shape; if the height of one CRB is just one data point, this CRB is equivalent to a horizontal line; if, in an extreme case, both the height and width of one CRB is equal to one data point, this CRB is in fact just one single point. With this general representation, the same criterion and measurements can be applied to all CRBs, in contrast to point and line representations which always require different sets of parameters.

In implementation, every CRB can be simply represented as a vector $<x_{min}, y_{min}, x_{max}, y_{max}, cv>$ describing the location of this CRB on the cross recurrence plot and its categorical value. Thus, any recurrence matrix/plot can be economically stored as a list of CRBs: $\{CRB_1, CRB_2 \ldots CRB_x\}$. This method dramatically reduces both storage and computation costs, compared with the traditional way of saving the entire M×N matrix.

To summarize, the new representation CRB has three advantages: 1) it contains the information about categorical values in two data streams, not just whether it is a match or not; 2) it can be directly linked to joint action sequences in behavioral data; and 3) it is an efficient data structure to represent and store the entire plot for any further computation.

CRP with more than two data streams

An additional advantage of CRB is to capture complex joint activities. Due to the data structure of a single matrix used in standard CRQA, each plot can only be constructed to reflect the interaction dynamics between two data streams (Marwan et al., 2007; Fusaroli et al., 2014). With CRB representation, blocks can overlap with each other. Therefore, multiple data streams can be used in either x-axis or y-axis. Figure 3(a) shows a CRP constructed with two holding action streams from child on the x-axis and two holding action streams from parent on the y-axis. For any point at (row_i, $column_j$) in the current CRP, we compare the i_{th} data points in all the data streams on y-axis at time t_i with all the j_{th} data points in temporal streams on x-axis at t_j. The categorical value of that point can be a list if there are more than one match.

$$\boldsymbol{R}_{i,j} = \{cv_{i_1}, cv_{i_2}, \ldots, cv_{i_k}\} \cap \{cv_{j_1}, cv_{j_2}, \ldots, cv_{j_k}\} \qquad (*)$$

cv_{i1} is the categorical value of i_{th} data point in the first data stream on y-axis; cv_{j1} is the categorical value of j_{th} data point in the first data stream on x-axis.

For example, point *p'* in Figure 3(a) has two categorical value matches (blue and green) because the child and parent were engaged in playing both the green and blue objects intermittently. More specifically, such joint play activities were usually formed because both partners played with one object at the beginning, then jointly switched to another object, and then jointly switched back to the one they played previously. In this case, the engagement between the first period and the third period on the original object form a large cross recurrent block which is overlapped with the cross recurrent block of the second in-between object. In implementation, after obtaining the two matching categorical values of *p'*, we group them into the existing CRBs. In this example, point *p'* belongs to two blocks: a part of green CRB_1, and also a part of blue CRB_2. In this way, point *p'* contributed to the overlapping portion between CRB_1 and CRB_2. The overlapping situation between two CRBs with multiple categorical value assignments of point *p'* is illustrated in Figure 3(b) with each layer consisting of one CRP with one different categorical value.

As shown in Figure 3(b), the recurrence matrix could also be stored in 3 dimensions – one for each categorical value. However, in practice, using M×N×C as the recurrence data structure (*C* is the total number of categorical values in behavioral data streams) would require a large storage space. In addition, this 3D matrix can be sparse, and therefore both storage and computing costs can be high when operated on a 3D sparse matrix representation. Instead, using CRBs as the basic building component to represent a cross recurrence plot has two advantages: 1) **completeness**: it allows multiple categorical value assignments of any point in a recurrence matrix because it can handle overlaps between different

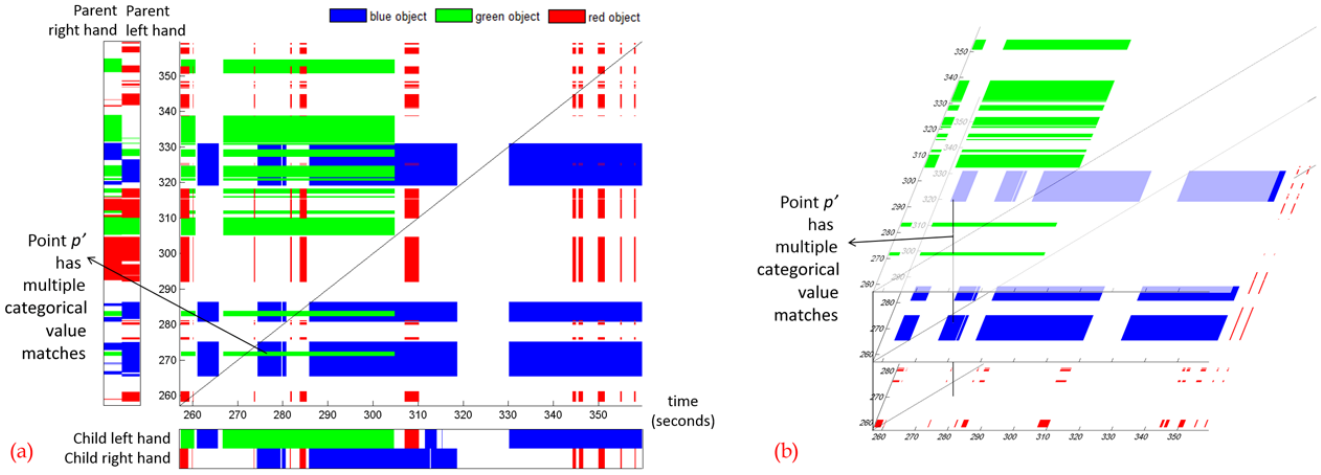


Figure 3: (a) A Cross Recurrence Plot constructed with two hand action streams from the child on the x-axis and two hand action streams from the parent on the y-axis. **C**ross **R**ecurrence **B**lock (CRB) representation allows a single plot to capture the dynamics from more than two behavioral modules. (b) Point *p'* has two matching categorical values and constitute one overlap point between two CRBs.

blocks, and 2) **efficiency**: it avoids unnecessary storage and computing costs.

With our new representation, Cross Recurrence Block, a variety of measurements can be extracted to reflect detailed interaction patterns. Next, we will introduce a suite of quantitative engagement measurements derived from CRBs.

Coordination beyond Synchronization

Interpersonal coordination is conducted in many complex forms that goes beyond synchronized behaviors wherein two partners were engaged in the same object/task at the exactly same time. In daily conversations and joint activities, we lead and follow each other's behavioral cues, take turns, diverge and converge from one target onto another. In this section, we will show that the CRB representation can be used to reveal fine-grained patterns of real-time coordination.

Width and Height

For each block, the width equates the horizontal length of each CRB and reflects how long the behavioral module(s) of $agent_x$ on the x-axis has participated in this joint action episode. Similarly, the height measure states how long the behavioral module(s) of $agent_y$ on the y-axis contributed to the same joint activity.

For example, in Figure 4, $Block_B$ is formed by the child's holding the blue object from 285.9s to 318.6s and then the parent's holding the same object from 280.5s to 286.3s. Thus, $Block_B$'s width is 32.7s and its height is 5.8s. Similarly, the sequence of holding events that formed $Block_C$ is that the child was holding the red object first from 307.1s to 310.2s, then the parent picked up the same red object and held it from 292.1s to 304.6s. As a result, $Block_C$'s width is 3.1s and its height is 12.5s.

Shape

For every joint action, if the child was holding the object longer than the parent, then the formed CRB appears wider, as horizontally shaped, such as $Block_B$ in Figure 4; instead, if the parent was holding the object longer, the corresponding CRB will be more vertically shaped, such as $Block_A$ and $Block_C$. The frequencies of horizontally and vertically shaped CRBs show the overall influence of each agent in joint activities throughout the whole interaction.

Time lag

The relative position of each CRB respect to the diagonal line reflects the leading and following relationship between the two participants. For each CRB, the lower left corner indicates the starting point of this joint activity sequence and the upper right corner reflects the ending point where both agents exit the joint action episode one after another (see $Block_C$ in Figure 4). Thus, the start time lag is calculated by subtracting $time_x$ from $time_y$ of the lower left corner of the

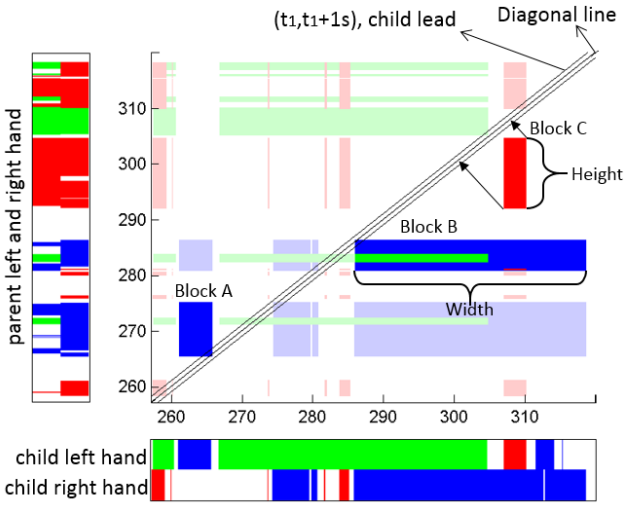


Figure 4: Illustration of Cross Recurrence Block based quantitative measurements

block: $start_time_lag = time_y - time_x$. If the time lag is positive, then it means that for this action sequence, $agent_x$ acts earlier than $agent_y$, who is the initiator of this activity episode. For the end time lag, similarly we calculate it by subtracting $time_x$ from $time_y$ of the CRB's upper right corner. With visualization, one can see the leading and following relationship simply by judging the positions of the start and ending points relative to the diagonal line in the plot.

With highlighted synchronized diagonal line and time shifted diagonal lines in Figure 4, we can intuitively identify the leading and following agent of each joint activity episode. For example, in $Block_A$, most part of which is above the diagonal line, was formed by the sequence of events that the child held the blue object first, then passed the object onto the parent's hand. In contrast, in $Block_B$, the parent was playing with the blue object first at 280.8s, then the child joined the object play at 285.8s, and continued playing with the same object until 318s, long after the parent switched to play with other objects at 286.3s. The start time lag of $Block_B$ is -5.0s, meaning the parent's action preceded the child's by 5 seconds; and the end time lag is -31.7s, meaning the parent switched to other objects first.

In addition, the numerical value of this measure and CRB's relative position to the diagonal line directly display the temporal relevancy of the event sequence that constitutes each individual CRB. The CRBs in upper left and lower right corners of the plot are formed by events far away in time. In this paper, we mainly focused on CRBs that overlapped with the diagonal line and calculated their CRB based quantitative measurements, which will be shown in the next section.

Results

With the holding action dataset collected from our child-parent toy play study in two different age groups (12 and 24 month), our main goal here is to demonstrate that standard measures of individual and joint behaviors didn't reflect different coordinative patterns between the two age groups, while the measures based on CRB representation revealed fine-grained patterns showing different interaction dynamics in child-parent play.

To compare with CRB-based measures, we first computed standard behavioral statistics, such as the mean duration and frequency of the child's and parent's holding events, mean duration and proportion of time of joint holding events wherein the child and the parent were holding the same object with either or both of their hands. The mean results and standard errors are listed in Table 1. None of those measures is statistically significant.

Next, we applied **C**ross **R**ecurrence **B**lock based **Q**uantification **A**nalysis (CRBQA) to the dataset, constructed CRP with four holding action data streams for each individual trial based on CRB data structure, and calculated quantitative measures including width, height, start and ending time differences, frequency of vertical and horizontally shaped CRBs overlapped with the diagonal line.

Table 1. List of behavioral measures of the child's and parent's holding events between two age groups

Behavioral measure	12 month	24 month	stats
Duration of child holding (seconds)	4.67±0.51	3.42±0.43	t(23)=1.87 p=0.07
Duration of parent holding (seconds)	1.91±0.14	2.40±0.21	t(23)=-1.90 p=0.07
Frequency of child holding (per minute)	15.49±1.98	15.26±1.30	t(23)=0.09 p=0.93
Frequency of parent holding (per minute)	22.42±1.56	22.39±1.90	t(23)=0.01 p=0.99
Duration of joint holding (seconds)	0.82±0.07	10.71±0.05	t(23)=1.23 p=0.23
Joint holding Proportion of time	11.27±1.65	11.24±1.22	t(23)=0.01 p=0.99

The CRB based results are shown in Table 2. Width indicates the duration of child's participation in forming CRBs, while height states for the parent's holding duration. Horizontal CRBs are blocks wherein the child's holding lasted longer than the parent, and vice versa for the vertical CRBs. With start time lag, positive values indicate that the joint holding sequences were initiated by the child and how many seconds afterwards the parent followed up by holding the same object; for ending time difference, positive values show that the child stopped holding the object first and negative values show that the parent switched to a new target object first when the child was still holding the same object.

Table 2. List of measures extracted with Cross Recurrence Block based Quantification Analysis of the child's and parent's holding events between two age groups

CRBQA	12 month	24 month	stats
Width (seconds)	11.89±2.34	7.96±1.45	t(23)=1.43 p=0.17
Height (seconds)	3.25±0.30	7.33±1.43	t(23)=-2.80 p<0.01
Frequency of horizontal CRBs (per minute)	4.51±0.56	3.82±0.33	t(23)=1.05 p=0.30
Frequency of vertical CRBs (per minute)	2.22±0.54	4.57±0.91	t(23)=-2.23 p<0.05
Start time lag (in seconds)	4.55±1.75	-0.01±1.32	t(23)=2.08 p<0.05
End time lag (in seconds)	-4.09±0.93	-0.63±1.26	t(23)=-2.21 p<0.05

Among 6 measures based on CRB representation, four of them show significant differences between the two groups. 12-month-old infants seem to hold objects much longer in joint activities compared with their parents (11.89 vs. 3.25 seconds). At the 24 month, they seemed to more or less become equal partners (7.96 vs. 7.33 seconds). For the measure of start time lag, positive values indicate that the joint holding sequences were initiated by the child. In the 12-month-old group, after the child started playing with a target object with their hands, on average, 4.55 seconds later, the parent followed up by holding the same object. For end time lag, the negative values for 12 month old group meant that the parent switched to a new target object first 4.09 seconds on average before the child terminated his manual actions on the current object. For the 24-month-old group, there was not much difference on temporal leading and following relationships between the dyad. This suggested that in the 12-month-old group, the parent was mainly following the child's lead, and in most joint plays, the child started holding a target object first and continued to play with it even after the parent switched to hold another object. These examples here show promise that the CRB-based approach can reveal fine-grained patterns of joint activities that we may not be able to obtain otherwise.

Conclusion

In this paper, we introduced a novel method to analyze coordinated behaviors between two social partners. This approach is built upon Cross Recurrence Quantification Analysis. The three key contributions here are: 1) to encode state matches between two agents in different categorical values and colors in a Cross Recurrence Plot (CRP); 2) to use a block representation to capture interaction with more than two data streams in one plot; 3) to derive a suite of measures based on CRB representation. We used holding action dataset from a child-parent interaction toy play study to demonstrate that this method can capture characteristics and dynamics beyond two temporal streams, revealing developmental changes of coordinated behaviors between 12 month and 24 month. One future direction in this particular application is to show coordinated patterns extracted based on CRB representation play a critical role in child development. For example, if better coordination leads to better development, then we should be able to use some derived measures to predict later development outcome. Moreover, this method can be applied to basically all interaction contexts with categorical behavioral data on various time scales. Any form of interpersonal or intrapersonal engagement shared between different behavioral models can be easily visualized and quantified with our method, such as gaze, speech data, pointing gestures, etc. As we move toward data-intensive science to understand human cognition, we will rely more heavily on developing and utilizing data analysis methods, such as the one presented here, to look for findings that may be undiscovered but highly robust. We envision that this method to be a useful and intuitive analytical tool for interaction study community.

Acknowledgments

We thank Steven Elmlinger, Melissa Elston, Charlotte Wozniak and Seth Foster for collection of the data. This work was funded by the National Institutes of Health Grants R01 HD074601 and R21 EY017843.

References

Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., & Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior research methods*, *45*(3), 842-856.

Angus, D., Smith, A., & Wiles, J. (2012). Conceptual recurrence plots: Revealing patterns in human discourse. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(6), 988-997.

Coco, M. I., & Dale, R. (2014). Cross-Recurrence Quantification Analysis of Behavioral Streams: An R Package. *Frontiers in Psychology*, *5*.

Cooper, M. L., & Foote, J. (2002). Automatic Music Summarization via Similarity Analysis. In *ISMIR*.

Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*, *21*(04), 1153-1161.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In *Translational Recurrences*, 137-155. Springer International Publishing.

Fusaroli, R., & Tylén, K. (2015). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. *Cognitive science*.

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta psychologica*, *32*, 101-125.

Kimmel, M. (2009). Intersubjectivity at close quarters: how dancers of Tango Argentino use imagery for interaction and improvisation. *Cognitive Semiotics*, *4*(1), 76-124.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*(8), 1404-1426.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, *438*(5), 237-329.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science*, 29(6), 1045-1060.

Riley, M. A., Richardson, M. J., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in psychology*, *2*, 38.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244-262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11).

Zbilut, J. P., Giuliani, A., & Webber, C. L. (1998). Detecting deterministic signals in exceptionally noisy environments using cross-recurrence quantification. *Physics Letters A*, *246*(1), 122-128.

`

Temporal Horizons and Decision-Making: A Big Data Approach

Robert Thorstad (rthorst@emory.edu)
Phillip Wolff (pwolff@emory.edu)
Emory University Department of Psychology

Abstract

Human behavior is plagued by shortsightedness. When faced with two options, smaller rewards are often chosen over larger rewards, even when such choices are potentially costly. In three experiments, we use big data techniques to examine how such choices might be driven by people's temporal horizons. In Experiment 1, we determine the average distance into the future people talk about in their tweets in order to determine the temporal horizon of each U.S. state. States with further future horizons had lower rates of risk taking behavior (smoking, binge drinking) and higher rates of investment (e.g., education, infrastructure). In Experiment 2, we used an individual's tweets to establish their temporal horizon and found that those with longer temporal horizons were more willing to wait for larger rewards. In Experiment 3, we were once again able to predict the choice behaviors of individuals from their tweets, this time showing that those with longer future horizons were less likely to take risks. The findings help establish a powerful relationship between people's thoughts about the future and their decisions.

Keywords: prospection; future thinking; big data.

Introduction

People often act impulsively. They eat unhealthy foods, gamble, overspend, and engage in behaviors that are likely to have a negative impact on their wellbeing. Such shortsighted behaviors demonstrate the phenomenon of *delay discounting*, a type of reasoning in which long-term benefits or risks are minimized in favor of smaller short-term rewards. Delay discounting is seemingly non-normative and maladaptive. Immediate and distant futures are equal parts of one's life as a whole, so it would appear to be in people's best interest to wait for larger rewards. And yet, people continue to make poor decisions, despite the long-term risks and in spite of the potential benefits of waiting for larger rewards. In this paper, we use big data techniques to address the question of why delay discounting might occur.

We hypothesize that delay discounting may be due, in part, to the way individuals think about the future. Specifically, we hypothesize that individuals with a long *temporal horizon* – who tend to think far into the future – will be less likely to discount future rewards.

There is mixed evidence in the literature for the role of future thinking in delay discounting. There is some evidence that individuals who tend to think about the future more in general (e.g. have a future time perspective) are less likely to discount future rewards (Steinberg et al, 2009; Daugherty & Brase, 2010), although effect sizes tend to be small. Additionally, there are reports that encouraging individuals to think about the future more vividly (e.g. episodic future thinking) reduces delay discounting (Peters & Buchel, 2010; but see also Kwan, Craver, Green, Myerson, & Rosenbaum, 2013).

Findings for temporal horizon, however, are more mixed. Heerey, Matveeva, & Gold (2011) report that schizophrenic patients with a longer time horizon are less likely to discount future rewards. However, the authors find no relationship between time horizon and discounting for healthy adult controls. Additionally, Fellows & Farah (2005) report that patients with lesions to ventromedial prefrontal cortex have a shortened time horizon but show normal delay discounting. These results provide mixed evidence for the role of time horizon in delay discounting.

One reason for these mixed results may be the problem of how to measure temporal horizon. Temporal horizon is usually measured by asking participants to imagine future events and then estimate their distance in the future. However, such explicit tasks may not accurately reflect participants' time horizon. Additionally, time horizon may be subject to state effects that are not captured in a single laboratory measurement. The current research addresses these limitation through an analysis of people's naturally occurring language on twitter about future events. By measuring future horizon in this way, we can more accurately assess its potential impact on people's choices about immediate and future rewards. Additionally, because twitter data is longitudinal in nature, we can address potential state effects by measuring a participant's time horizon over many different time points.

Experiment 1: Discounting in Populations

The hypothesis that delay discounting depends on future horizon should extend beyond the choices of individuals to the collective choices of an entire population. Because various statistics of impulsive / long-range thinking are already available for populations, we began our investigation by analyzing the temporal horizon and collective choices of entire states. To conduct this kind of analysis we first collected a large number of tweets from each state. The tweets were then automatically analyzed for whether they referred to the future or past, as well as their distance into the future or past. Using these analyses, we could determine each state's *future orientation*, that is, the degree to which the tended to talk about the future or the past, as well as each state's *temporal horizon*, that is, distance into the future and past. Once each state's temporal horizon was determined, we could examine the extent to which each state's future horizon correlated with various kinds of collective choices. In particular, we examined the association between future horizon and various kinds of

`

Future Orientation

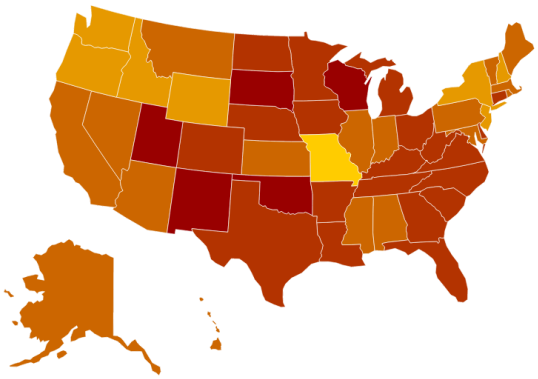


Figure 1: Mean future orientation of each U.S. state. Colors range from yellow (least future) to dark red (most future). Note that the 4 states excluded from analysis are dummy-coded with the mean value.

Temporal Horizon

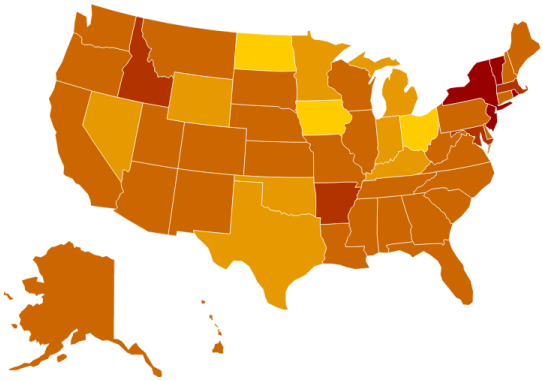


Figure 2: Mean temporal horizon of each U.S. state. Colors range from yellow (shortest horizon) to dark red (longest horizon). Note that the 4 states excluded from analysis are dummy-coded with the mean value.

impulsive behaviors, such as binge drinking rates, cigarette smoking rates, and drunk driving rates. We also examined the association between future horizons and healthy choice behaviors, such as seatbelt usage, education spending, and highway spending. Because of the theoretical novelty of temporal horizon, we also compared our measure of future horizon with several demographic variables including political orientation and economic activity.

Methods

Tweet Collection. We collected 8,550,131 tweets from April to July 2015 using the Twitter Developer API. Tweets were restricted to English language only. Tweets were collected for equal durations from each U.S. state.

Future Orientation Classification. We defined the future orientation of a state as the number of tweets about the future, divided by the number of tweets that could be classified as either past or future (future / (future+past)). We classified tweets as about the past, future, or neither using a custom-built classifier. For each tweet, we performed part-of-speech parsing using the Stanford Parser (Chen & Manning, 2014). The result is a sentence represented as a combination of part-of-speech-tags (e.g. "NP" for noun phrase) and words. We then developed a set of 112 lexical and syntactic rules to classify the tweet as past, future, or neither. For example, the tweet "Maybe ill stay home w mom tomorrow" was classified as future because it contains a singular noun "home" dominating the lexical item "tomorrow" (the relevant rule is NN|RB > tomorrow). In total, about 31% of tweets were classified as future and about 23% of tweets were classified as past. The future classifier was validated against human ratings of 1,000 sentences, showing 76.61% agreement with human raters (precision=0.75, recall=0.90). This agreement with human raters compares favorably to other attempts in the literature (e.g. Nakajima et al, 2014), and approaches human accuracy (human-to-human percent agreement = 86.44%). This classifier is available for use on our website: http://mindandlanguagelab.com/futureAnalysis.

Temporal Horizon Classification. For future tweets only, we classified their temporal horizon using a keyword approach. We created a list of 58 twitter-appropriate keywords marking explicit periods of time: for example, "tomorrow", "tmrw," etc. For each keyword, we estimated the average number of minutes in the future it occurs. For example, "tomorrow" occurs on average 1440 minutes (24 hours) in the future. For each future tweet, we used regular expression matching to identify temporal keywords in the tweet. For example, "maybe ill stay home with mom tomorrow" matched the temporal noun "tomorrow." 258,281 tweets had identifiable time horizons under this criterion. For each tweet that matched one or more temporal keywords, we averaged the resulting number of minutes to estimate its time horizon. Because time horizon had a large range (180-259,200 minutes) we took the natural logarithm of the result. We then averaged temporal horizon at the state level, creating a single score representing the average temporal horizon of each U.S. state. To ensure accurate estimation, we excluded states with fewer than 100 tweets with identifiable time horizons (3 States excluded: AK, ME, MT). We also excluded 1 state (HI) because its time horizon was more than 5 SD above the mean; this may be due to a large number of tweets from non-residents on vacation.

Demographic Measures. To explore the characteristics of states with different time horizons and future orientation, we collected state-level economic, political, religiosity & well-being indices from Gallup inc. We separately correlated time horizon and future orientation at the state level with each measure.

Risky Decision-Making. We hypothesized that time horizon might relate to risk, such that states with longer time horizon would take fewer risks. To evaluate this hypothesis, we collected 6 state-level indices of risky behavior from publicly available government data. We collected 3 indices of risky behavior (binge drinking,

cigarette smoking, and drunk driving) and 3 indices of behavior we considered the opposite of risky (seatbelt usage, education spending, and highway spending). Education and highway spending data were collected from the US Census; the other indices were collected from the Center for Disease Control. We then created a *risk composite* by separately normalizing the risk indices, correcting sign such that positive numbers represent risk, and averaging the resulting z-scores for each state.

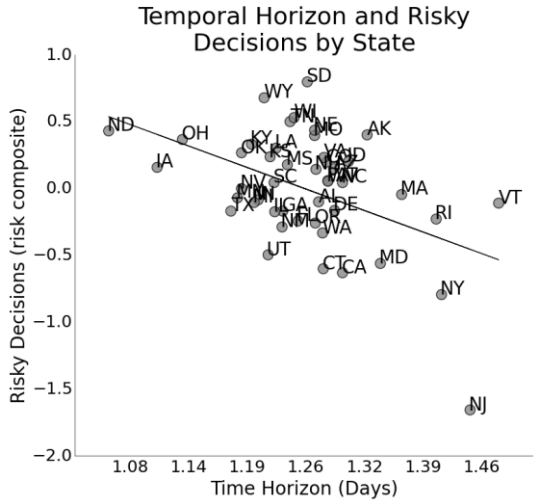


Figure 3: Temporal horizon and risky decisions by state. Labels represent state abbreviations. Note that the horizontal axis is plotted in logarithmic scale.

Results

Temporal Horizon by State. We first present the average temporal horizon for each U.S. state. The mean temporal horizon was 1.27 days, see Figure 1. The U.S. census region with the longest time horizon was the Northeast; the Midwest had the shortest time horizon.

Future Orientation by State. We next present the average future orientation for each state. The mean future orientation was 31.55% (*SD* = 1.02%). Figure 2 displays the mean future orientation for each U.S. state. To evaluate the relationship between time horizon and future orientation, we calculated the Pearson correlation between states' temporal horizon and future orientation. There was a negative relationship between future orientation and temporal horizon, $r(46) = -0.411$, $p<0.01$. One potential explanation for this finding is that thoughts about the close future may be more frequent than thoughts about the far future.

Demographic Measures. To begin to investigate the characteristics of states with different time horizon and future orientation, we conducted an exploratory correlation analysis with 29 economic, political, religiosity, and well-being measures from Gallup, inc. States with a long time horizon were more politically liberal than states with a short time horizon, $r(45) = 0.533$, $p < 0.001$. By contrast, states with greater future orientation, were less politically liberal than less future-oriented states, $r(45) = -0.398$, $p < 0.01$.

Table 1: Results of a linear regression predicting risky decision-making at the state level. Future time horizon was still significantly negatively related to risk when including demographic and computational controls.

	β	p
Future Horizon	-0.368	0.023 *
Population	0.575	0.482
Gender	0.175	0.175
Median Age	0.015	0.916
GDP	-0.930	0.260
Past Horizon	-0.057	0.688

Risky Decision-Making. The main hypothesis evaluated was that states with long time horizon will take fewer risks than states with short time horizon. To evaluate this hypothesis, we created a composite of 6 risky decision-making indices at the state level, and separately correlated the composite with time horizon and future orientation. States with a long time horizon took significantly fewer risks than states with a short time horizon, $r(45) = -0.467$, $p < 0.01$ (Figure 3). Because New Jersey had a risk composite score more than 3 SD less than the mean, we verified that the relationship still holds excluding New Jersey, $r(45) = -0.352$, $p < 0.05$.

To verify that this relationship was not due to a simple tendency to think more about the future, we correlated states' future orientation with risk composite score. There was no relationship between future orientation and risk, $r(46) = 0.271$, $p = 0.069$.

In order to rule out alternative explanations for the relationship between time horizon and risk, we conducted a linear regression analysis controlling for several additional variables (Table 1). To control for demographic differences between states, we controlled for gender (percent male), population, median age, and GDP at the state level. To validate our computational method, we included an additional control by calculating a state's past horizon. To do this, we performed every computational step used to calculate a state's future horizon, except using tweets classified as past, not future. To evaluate whether future time horizon is negatively related to risk, we separately fit linear regression models with all control variables, with and without future time horizon. Adding future time horizon significantly improved model fit: $R^2 = 0.401$ (with time horizon) versus 0.315 (without time horizon), F-test: $F(1,39) = 5.564$, $p < 0.05$.

Discussion

In Experiment 1, we evaluated the hypothesis that having a long temporal horizon increases sensitivity to potential future costs. To do this we built a classifier to identify how near or far U.S. states tend to think in the future, and correlated the results with an index of risky decision-making.

The main result of Experiment 1 was that states with longer time horizons took fewer risks than did states with shorter time horizons. This result was specific to temporal

horizon, rather than future orientation in general. This result remained significant when including several demographic and computational controls.

One limitation of Experiment 1 was that analysis was conducted at the level of states, not individuals. As such, it was impossible to obtain an experimental measure of risky decision-making. The aim of Experiment 2 was to address this limitation by linking individual participants' temporal horizon to an experimental measure of future decision-making. Additionally, we aimed to extend our results beyond future costs to potential future benefits. If having a long time horizon increases sensitivity to potential future costs, does it also increase sensitivity to potential future benefits?

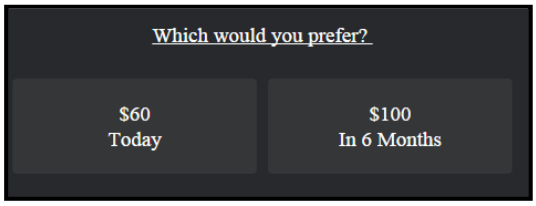


Figure 4: A sample delay discounting trial. In this trial, the participant is asked to choose between $60 today and $100 after a delay of 6 months.

Experiment 2: Discounting in Individuals

The aim of Experiment 2 was to test the hypothesis that individuals with a long temporal horizon are less likely to discount future rewards. Participants completed a delay discounting task where they made a series of 60 choices between smaller present and larger future rewards (Figure 4). Participants also provided their twitter handle, allowing identification of their individual temporal horizon. If temporal horizon affects delay discounting, participants with a long temporal horizon should be more likely to choose to wait for future rewards.

An additional aim of Experiment 2 was to investigate the stability of temporal horizon over time. To do this, we separately classified participants' temporal horizon using only recent tweets from the past 2 and 4 weeks, as well as using tweets from all dates.

Methods

Participants. 198 participants were recruited via Amazon Mechanical Turk and received $1 for participation. 29 participants were excluded for failing to provide a valid twitter handle (N=7), providing a "protected" twitter account that could not be processed (N=14), or failing to complete the delay discounting task (N=8).

Delay Discounting. Delay discounting questions were composed by fulling crossing 6 delay lengths (1 week, 6 months, 1 year, 5 years, 10 years, 20 years) with 10 immediate reward amounts ($1, $5, $10, $20, $40, $60, $80, $90, $95, and $99). The delayed reward was always $100. For example, on one trial, participants chose between $60 today and $100 in 5 years. Participants completed 60 delay discounting trials in random order. We excluded data from 13 participants who chose the future reward less than 5 times.

Temporal Horizon Classification. Participants were asked to provide their valid twitter handle with at least 50 tweets. Temporal horizon was classified using the same methods as in Experiment 1, with 2 additions. First, we added an additional preprocessing step to remove URLs, hashtags (#), and twitter user mentions (@) using regular expression matching. Second, to investigate the stability of temporal horizon over time, we separately classified participants' temporal horizon using only tweets from within 7 days, 14 days, 30 days, 60 days, 180 days, and 360 days from the experiment, as well as data from all time periods. To ensure accurate estimation of time horizon, we excluded participants with fewer than 10 tweets (at any date) for which time horizon could be calculated.

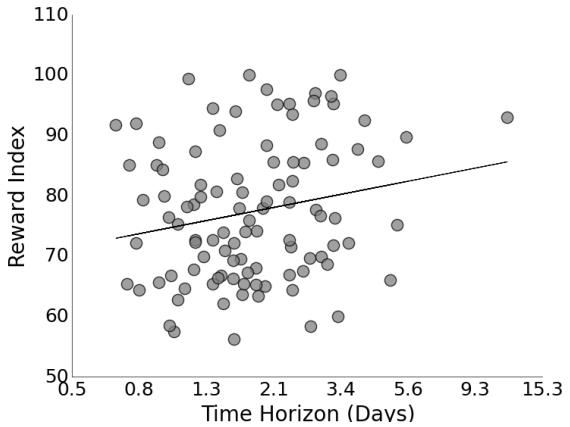


Figure 5: Participants with long time horizon were more willing to wait for future rewards. The horizontal axis is plotted in logarithmic scale.

Results

First, we evaluated whether participants' temporal horizon affected delay discounting. For each participant, we calculated a *reward index*, e.g. the total money earned on the delay discounting task, divided by the total possible earnings. A large reward index indicates a willingness to wait for future rewards. Participants with a long time horizon were more likely to wait for future rewards, $r(100) = 0.204$, $p < 0.05$ (Figure 5).

Second, we asked whether participants' decisions were better predicted by their recent time horizon, or by their time horizon based on all tweets. We did not find that the temporal horizon associated with more recent tweets was more predictive of delay discounting than the temporal horizon associated with all tweets (all ps > 0.2). This could imply that temporal horizon is a trait-like characteristic that is relatively stable in an individual. However, it is also the case that the number of recent tweets was relatively low, which could reduce the ability to find an effect for recent tweets.

`

Discussion

The main finding of Experiment 2 was that participants' temporal horizon affected their delay discounting. Using all tweets to classify temporal horizon, participants with a long temporal horizon were more likely to wait for future rewards. This result supports the hypothesis that one reason participants discount future rewards is their cognitions about the future.

Experiment 2 demonstrated that temporal horizon affects decision-making in situations where a benefit occurs only in the long future. In Experiment 3, we wondered whether the same might be true for future costs.

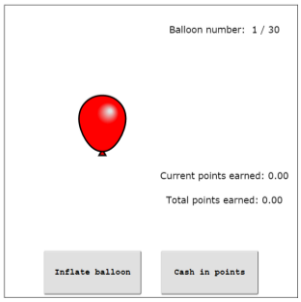


Figure 6: The BART task. Participants earned points for inflating balloons, and at any point could bank points earned and proceed to the next trial.

Experiment 3: Risk Taking in Individuals

The aim of Experiment 3 was to explore the relationship between temporal horizon and decision-making where the future outcome is a cost, not a benefit. In risky decision-making, participants must trade off between a potential gain in the present, and a loss in the future. In Experiment 2, we found that a long temporal horizon increased participants' sensitivity to future gains. However, in Experiment 1, we also found at the level of states that a long temporal horizon increased states' sensitivity to future costs, e.g. risks. The aim of Experiment 3 was to test at the individual level whether a long temporal horizon might increase participants' sensitivity to future costs.

In Experiment 3, participants completed the Balloon Analogue Risk Task (BART; Lejuez et al, 2003). In this task, participants are presented with a series of 30 balloons (Figure 6). Participants earn points every time they inflate the balloon, but they take a risk by doing so: the balloon may pop, resulting in no points for the trial. Alternatively, at any time participants may bank points currently earned and proceed to the next trial.

To evaluate whether future horizon affects evaluation of future costs, we also asked participants to provide their twitter handle. For each individual, we identified their temporal horizon and future orientation on the basis of their tweets. If temporal horizon affects evaluation of future costs, participants with a long temporal horizon should take fewer risks in the BART task.

Methods

Participants. 83 participants were recruited via Amazon Mechanical Turk and received $1 for participation. There were 3 exclusion criteria. Participants were excluded for failing to provide a valid twitter handle (*N*=9), providing a "protected" twitter account that could not be processed (*N*=8), or failing to complete the BART task (*N*=1).

Temporal Horizon and Future Orientation Classification. Participants were asked to provide their valid twitter handle with at least 50 tweets. Temporal horizon and future orientation classification were performed for each individual twitter account using the methods from Experiment 1, with 2 changes. First, we removed hashtags (#) and user mentions (@) using the same methods as Experiment 2. Second, to improve classification accuracy at the individual level, we simplified the temporal horizons classifier by restricting temporal keywords to those referencing tonight (e.g. "tonight," "2nite," etc.) and references to "tomorrow." The total number of tweets collected was 117,281. 14 participants were excluded because their twitter account did not provide any tweets with identifiable time horizon.

BART Task. Participants completed a 30-trial version of the BART task, using a javascript implementation by Timo Gnambs (2013). Balloon explosion points were drawn from a uniform distribution from 1-128 clicks. There were 3 balloon colors (10 balloons each) with differing points gained per click: 0.5, 1, or 5 points. We calculated two measures of participants' risk taking. (1) *Adjusted pumps*: mean number of inflations per balloon, excluding trials where the balloon exploded. (2) *Number of explosions*: number of balloons inflated until explosion.

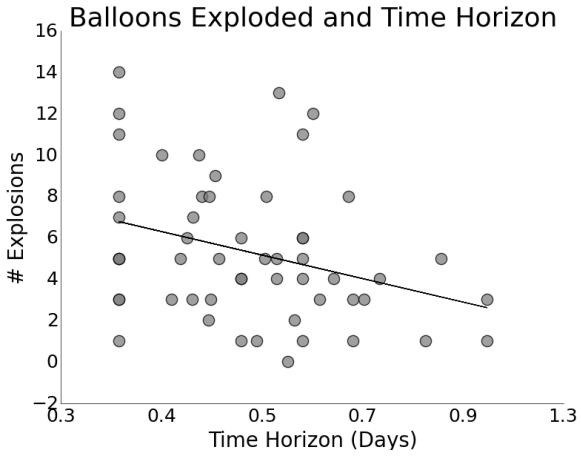


Figure 7: Participants with a long time horizon exploded fewer balloons in the BART task. The horizontal axis is plotted in logarithmic scale.

Results

The main hypothesis evaluated was that participants with a long time horizon will take fewer risks. Supporting this prediction, participants with a longer time horizon exploded

`

fewer balloons, $r(54) = -0.321$, $p < 0.05$ (Figure 7) and trended towards inflating balloons fewer times, adjusted pumps: $r(54) = -0.230$, $p = 0.094$. This supports the claim that participants with a long time horizon are more sensitive not only to future benefits, but also to future costs.

We also verified that our results were driven by temporal horizon, rather than a mere tendency to think about the future. There was no relationship between future orientation and BART performance, r(54)=0.027, p=0.846 (explosions); r(54)=0.067, p=0.632 (adjusted pumps). This lack of result supports the conclusion that temporal horizon, and not a mere tendency to think about the future, increases sensitivity to future costs.

Discussion

The main result of Experiment 3 was that having a long temporal horizon increases sensitivity to potential future costs, not just benefits. Individual participants' temporal horizon, but not their future orientation in general, predicted risks taken in the BART risk task.

General Discussion

The results provide support for the hypothesis that delay discounting is determined, at least in part, by a person's future horizon. In Experiment 1, we showed that U.S. states with longer time horizons, but not greater future orientation, were more likely to engage in risky and impulsive decisions. In Experiment 2, we found that people with longer future horizons, as revealed by their tweets, were more likely to wait for future rewards. In Experiment 3, we found that people with longer horizons were more likely to avoid risks.

Although not directly examined in this paper, the experiments provide some insight into the length of people's future horizon. For example, in Experiment 1 we observed that the average temporal horizon was approximately 1.27 days. In contrast, in Experiment 2 we found a slightly longer temporal horizon: 2.01 days. Given that the temporal horizon in Experiment 2 was almost twice as long as in Experiment 1, the difference between these experiments may appear surprising. However, the difference between these experiments can probably be explained by differences in sampling. In Experiment 1, we randomly selected tweets from the twitter stream, and as a consequence no doubt sampled frequent twitter users more often than infrequent twitter users. It could be that frequent twitter users tend to have a shorter temporal horizon than infrequent users. The way we sampled in Experiment 2 was less susceptible to this potential bias because twitter users were selected based on their participation in a separate experiment. Arguably, then, the mean temporal horizon from Experiment 2 may be more representative of a typical temporal horizon.

One question not resolved by the current findings is whether temporal horizon is a relatively stable characteristic of individuals, i.e. a trait, or whether temporal may vary as a function of context and experience, making it more like a state. To the extent that temporal horizon is a state, it may be possible to have an impact on people's decision-making processes by changing temporal horizon. On the other hand, it could be that the distance people look into the future is largely independent of context, and as a consequence these biases may be a relatively stable characteristic of individuals. Regardless, by looking at people's temporal horizon we can see how decisions may depend on people's temporal biases.

Acknowledgments

This research was made possible through the support of a grant from the John Templeton Foundation, "Prospective Psychology Stage 2: A Research Competition" to Martin Seligman. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. The authors wish to thank Bridget Copley for assistance developing the future orientation classifier.

References

Chen, D. & Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of EMNLP 2014*.

Daugherty, J. & Brase, G. (2010). Taking time to be heathy: predicting health behaviors with delay discounting and time perspective. *Personality and Individual Differences*, *48*(2), 202-207.

Fellows, L. & Farah, M. (2005). Dissociable elements of human foresight: a role for the ventromedial frontal lobes in framing the future, but not in discounting future rewards. *Neuropsychologia*, 43(8), 1214-1221.

Gnambs, T. (2013). Balloon Analogue Risk Task [Computer software]. Retrieved from timo.gnambs.at.

Heerey, E., Matveeva, T., & Gold, F. (2012). Imagining the future: degraded representations of future rewards and events in schizophrenia. *Journal of Abnormal Psychology*, *120*(2), 483-489.

Kwan, D., Craver, C., Green, L., Myerson, J., & Rosenbaum, R. (2013). Dissociations in future thinking following hippocampal damage: evidence from discounting and time perspective in episodic amnesia. *JEP: General*, *142*(4), 1335-1369.

Lejuez, C. et al. (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and clinical psychopharmacology*, 11(1), 26.

Nakajima et al. (2014). Investigation of future reference expressions in trend information. *Proceedings of the AAII Spring Symposium Series*, 31-38.

Peters, J. & Buchel, C. (2010). Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, 66(1), 138-148.

Steinberg, L., Graham, S., O'Brein, L., Woolard, J., Cauffman, E., & Banich, M. (2009). Age differences in future orientation and delay discounting. *Child Development*, *80*(1), 28-4.

Using Subgoal Learning and Self-Explanation to Improve Programming Education

Lauren E. Margulieux (l.marg@gatech.edu)
Georgia Institute of Technology, School of Psychology
Atlanta, GA 30332-0170 USA

Richard Catrambone (richard.catrambone@psychology.gatech.edu)
Georgia Institute of Technology, School of Psychology
Atlanta, GA 30332-0170 USA

Abstract

The present study explored passive, active, and constructive methods of learning problem solving procedures. Using subgoal learning, which has promoted retention and transfer in procedural domains, the study compared the efficacy of different methods for learning a programming procedure. The results suggest that constructive methods produced better problem solving performance than passive or active methods. The amount of instructional support that learners received in the three different constructive interventions also affected performance. Learners performed best when they either received hints about the subgoals of the procedure or received feedback on the subgoal labels that they constructed, but not when they received both. These findings suggest that in some cases constructing subgoal labels is better than passively or actively engaging with subgoal labels. There is an optimal level of instructional support for students engaging in constructive learning and that providing too much support can be equally as detrimental as providing too little support.

Keywords: subgoal learning; self-explanation; worked examples; computing education.

Introduction

Students in higher education need to be able to learn independently, at least in part. As the number of students pursuing bachelor's and advanced degrees increases, so does the ratio of students to instructors and the number of online courses. These factors make direct interaction between students and instructions increasingly limited and self-guided learning increasingly valuable. To help students be more independent learners, support from researchers and instructional designers is needed. The present research examined a new strategy to support independent learning: the integration of subgoal learning and self-explanation.

Subgoal Learning

Subgoal learning refers to a strategy used predominantly in STEM fields that helps students to deconstruct problem solving procedures into subgoals to better recognize the structural components of the problem solving process (Atkinson, Catrambone, & Merrill, 2003; Catrambone, 1998). Subgoals are functional pieces of procedures used to solve problems that contain one or more individual steps, such as solving for a variable in a calculus problem.

Research suggests that when instructions help students learn the subgoals of a procedure, students are better able to transfer knowledge to solve novel problems. Catrambone and Holyoak (1990) found that when instructional materials highlighted the subgoals of a procedure, learners were more likely to correctly apply it to problems that used the same procedure but had different contextual features (e.g., problems about birthdays versus those about football) or had modified or new steps. Subsequent studies (Catrambone, 1994, 1996, 1998; Margulieux & Catrambone, 2014; Margulieux, Guzdial, & Catrambone, 2012) have consistently found that subgoal-oriented instructions improved problem solving performance across a variety of STEM domains.

Subgoal-oriented instructions are typically implemented as worked examples. Worked examples give learners concrete examples of the procedure being used to solve a problem. Because problems necessarily include a context, such as birthdays or football, worked examples include context-specific information. Eiriksdottir and Catrambone (2011) argued that, when studying examples, learners tend to focus on superficial features rather than the structural features because superficial features are easier to grasp and novices do not have the necessary domain knowledge to recognize the structural features of examples (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). A focus on superficial features leads to ineffective organization and storage of information that, in turn, leads to ineffective recall and transfer (Bransford, Brown, & Cocking, 2000).

To promote deeper processing of worked examples and, thus, improve retention and transfer, worked examples have been manipulated to promote subgoal learning. Subgoal labeling is a technique used to promote subgoal learning that has been used to help learners recognize the structural structure of the procedure being exemplified in worked examples (e.g., Catrambone, 1994, 1996, 1998). Subgoal labels are function-based instructional explanations that describe the purpose of a subgoal to the learner.

Catrambone (1998) found that learners who received labels that were abstract (e.g., Ω) and had greater prior knowledge performed better than those who received labels that were context-specific (e.g., isolate x) on problem solving tasks that were given after a week-long delay or that required using the procedure differently than demonstrated in the examples. Catrambone (1998) argued that learners with sufficient prior knowledge were able to correctly explain to themselves the purpose of the subgoal. He argued that prompting self-explanation of the subgoal by providing a label that did not explain the subgoal's function was more effective than providing an informative label.

The findings from Catrambone's (1998) research align with a growing body of evidence that learning is more effective when students actively or constructively engage with content rather than passively receive content. This body of evidence is summarized by Chi (2009) and used to support her Interactive-Constructive-Active-Passive (ICAP) framework. In this framework, Chi (2009) characterized four types of learning based on students' engagement with content: interactive, constructive, active, and passive (see Figure 1 for definitions and examples).

Passive	Active	Constructive	Interactive
Receiving information without physical activity (e.g., listen to a lecture)	Receiving information with physical activity (e.g., take notes on a lecture)	Individually producing information beyond that which is provided (e.g., connect concepts to prior knowledge)	Collaboratively producing information beyond that which is provided (e.g., discuss concepts)

Figure 1. Definitions of passive, active, constructive, and interactive learning based on the ICAP framework (Chi, 2009).

Using this framework to compare the learning outcomes from various learning activities, Chi (2009) found that interactive and constructive learning were the most effective, active learning was the second most effective, and passive learning was the least effective. Most research about subgoal learning, besides Catrambone (1998), has provided meaningful subgoal labels that explain the function of subgoals to learners. Providing labels to learners promotes passive learning, which is the least effective method of learning. The present study explored whether more engaging methods of learning subgoals, such as self-explanation of the subgoals of a procedure, would improve novel problem solving.

Present Study

The present study prompted participants to learn the subgoals of a procedure through a worked example that either encouraged passive, active, or constructive learning. The problem solving domain for the present study was programming. Because participants were novices, the present study used a drag-and-drop programming language to teach programming concepts. Drag-and-drop programming languages are more easily understood by novice learners because they can select and drag pieces of code from a menu, which does not require learning the syntax and semantics of a programming language (Hundhausen, Farley, & Brown, 2009). The programming language used in the present study was Android App Inventor, which is used to create applications (apps) for Android devices. Participants used App Inventor to create an app that has buttons that play sounds when pressed (see Figure 2 for excerpt).

The subgoals of the procedure were identified using the Task Analysis by Problem Solving (TAPS) procedure (Catrambone et al., 2016) that has been used in prior research (e.g., Margulieux & Catrambone, 2014). In the passive learning condition, participants were given subgoal labels created by the experimenters, as is conventional in prior subgoal research (e.g., Catrambone, 1998). These subgoal labels will also be created through the TAPS procedure (Catrambone et al., 2015).

In the active learning condition, participants were given the worked example grouped by subgoals and asked to select a subgoal label from a list of labels that matched the purpose of the group. The list contained only labels that were viable options, meaning the list did not include distractor items. This active method of self-explaining was equivalent to the active self-explanation methods used by Aleven and Koedinger (2002) and Conati and VanLehn (2000). The method matches Chi's (2009) definition of active learning as a method that requires activity from the learner but not construction of new information.

In the constructive learning conditions, participants were asked to create their own subgoal labels to explain the subgoals of the procedure. To train participants to construct their own subgoal labels, they were given subgoal label training. Only the constructive groups received this training. The passive and active groups received a comparable task: analogy training. Training for analogies (e.g., water : thirst :: food : hunger) was considered equivalent because both analogies and subgoal labeling requires people to consider the underlying relationship between words and come up with a new word that describes that relationship.

The three constructive learning conditions prompted participants to construct their own subgoal labels. They differed on the amount of guidance that participants received while constructing labels. In the guided constructive conditions, participants were given the worked example with the solution steps grouped by subgoal, and the example indicated which subgoals achieved the same functions. For instance, all of the subgoals denoted as "Label 1" achieve the same function though the contexts are different (see Figure 2). In the guided constructive with hints condition, participants were given hints about the similarities among different instances of the same subgoal. In the guided constructive without hints condition, participant did not receive these hints. In the unguided constructive condition, participants received a worked example that did not indicate which steps belonged to which subgoals. Participants in this condition had to identify the subgoals for themselves and create labels for them.

The amount of guidance that participants received during instruction differed based on whether they received feedback. Some studies in the self-explanation literature have found that feedback supports self-explanation because it reinforces correct explanations and reduces floundering

(Aleven & Koedinger, 2002; Conati & VanLehn, 2000). Conversely, other studies have found that feedback creates overreliance on instructional information provided via feedback and, overall, hinders self-explanation (Schworm & Renkl, 2006). Based on this conflicting evidence, feedback was considered an important feature to vary in the present study. Instructions for participants who received feedback had another copy of the worked example that included subgoal labels created by the experimenters. For the passive condition, this copy was exactly the same as the initial worked example. For the active and constructive conditions, the copied example with experimenter-created subgoal labels provided feedback to the participants about whether they selected the correct labels or created similar labels. Participants who received feedback were asked to compare their labels to those created by the experimenter to prompt them to reflect on the similarities or differences between the two. Instructions for participants who did not receive feedback included only the worked example with the passive, active, or constructive interventions. These participants were asked to re-read the example to make time on task more similar to that of participants who received feedback. The exception was that participants in the passive and no feedback condition were not asked to re-read the example to make their experience different from those in the passive with feedback condition. Due to this difference, the time on task was different, providing some insight into whether time on task affects performance for this task.

Given Labels (Passive)	Placeholder for Label (Active and Constructive)
Problem: Create an app that plays a drum sound when the image of a drum is touched. Handle Event Click on "My Blocks" to see the blocks for components created Click on "clap“ Drag out a when clap.Touched block Set Output Click on “clapSound” Drag out call clapSound.Play Connect it after when clap.Touched	Problem: Create an app that plays a drum sound when the image of a drum is touched. Label 1: _____ Click on "My Blocks" to see the blocks for components created Click on "clap“ Drag out a when clap.Touched block Label 2: _____ Click on “clapSound” Drag out call clapSound.Play Connect it after when clap.Touched

Figure 2. Worked example formatted with given labels or placeholders for labels.

Because the worked example was long, participants received only one worked example. Giving one worked example provided a unique opportunity to ensure that participants in the feedback condition did not overly rely on feedback. Participants were not told that they would receive feedback until they completed the task, meaning that they did not know to expect feedback.

The guidance provided by feedback was expected to interact with subgoal learning method. Withholding feedback can lead to incorrect explanations and floundering, but giving feedback can hinder self-explanation and lead to overreliance on feedback (Renkl, 2002; Schworm & Renkl, 2006). Learners making self-explanations can flounder because self-explanation, especially constructive explanations, requires some insight, meaning that learners have to recognize connections between pieces of information that are not necessarily apparent from the instructions (Wylie & Chi, 2014). Durso, Rea, and Dayton (1994) found that insight resulted from mental restructuring of knowledge that made connections between previously disjointed pieces of information. Durso et al. (1994) also found that if participants were given the solution to the problem at hand, mental restructuring did not occur. Durso et al.'s (1994) findings can explain why receiving information that could have been constructed through self-explanation does not allow for mental restructuring. Therefore, extra guidance from feedback on self-explanations was not always expected to lead to better learning outcomes, especially when learners received high levels of guidance during self-explanation.

Method

Participants

Each of the 10 conditions had 20 participants (N = 200). Participants were students at a mid-sized, southeastern, technical institute. Participants were required to have no experience with App Inventor and could not have taken more than one computer science or programming course. They also completed a multiple-choice pre-test to ensure that they did not have prior knowledge of the procedure. The majority of participants (91%) scored a zero on the pre-test, and no participants scored higher than one point. Participants completed a demographic questionnaire. Demographic variables and pre-test scores were analyzed to determine that groups were equivalent, but this analysis will not be reported due to space limitations.

Design

The experiment was five-by-two factorial, between-subjects design: subgoal learning method (passive, active, guided constructive with hints, guided constructive without hints, or unguided constructive) was crossed with feedback (no feedback or feedback). Dependent measures that are included in this paper were performance on the problem solving tasks and the subgoal labels that participants construct. Other measures are not discussed in this paper due to space restrictions.

Procedure

Sessions took between 80 and 110 minutes, depending on how quickly participants complete each of the tasks. First,

participants completed the demographic questionnaire and pre-test. Then participants started the instructional period, which took 40 to 55 minutes. All manipulations occurred within the instructional period.

The instructional period started with an overview video of the App Inventor interface that was the same across all participants. The video did not include information about the procedure being taught, but it was intended to help participants familiarize themselves with the problem solving space in which they would be working. After the introductory video, participants received either subgoal label or analogy training. Next, participants received the worked example. The worked example listed the steps taken to create a Music Maker app that plays musical sounds when images of musical instruments are pressed or the device is shaken. The format of the worked example depended on participants' assigned method of subgoal learning. The passive method gave participants subgoal labels (see passive condition in Figure 2), and the other methods gave participants spaces to fill in subgoal labels (see constructive condition in Figure 2), except for the unguided constructive condition, which had only the listed steps. For the active method, participants had a word bank with labels that they could select. In the guided constructive with hints condition, additional text highlighted similarities between all subgoals called "Label 1." This guidance was given for each subgoal.

When participants finished the first pass through the worked example, they were either prompted to re-read the example for the no feedback condition, or they were given the worked example with the experimenter-created subgoal labels for the feedback condition. Participants in the feedback condition were told that the subgoal labels in the second copy of the worked example were created by subgoal label experts. Then they were asked to compare the labels that they made or selected to those given in the second example. To ensure that participants paid attention to the worked example and could complete tasks in the App Inventor interface, they were asked to complete practice problems before finishing the instructional period.

Following the instructional period, participants completed problem solving tasks that measured learning. During this assessment period, participant did not have access to the instructional materials. They were told of this restriction at the beginning of the session. The problem solving tasks asked participants to modify or add components to their Music Maker app. Of the five tasks, two required contextual transfer from the worked example, meaning that the superficial features of the app components were different (e.g., exchange a drum sound for a cymbal sound) but the procedural steps used to create them were the same. The remaining three tasks required procedural transfer from the worked example, meaning that the individual steps used to create the app components were different but the procedure used to create them was structurally the same. For instance, the worked example showed steps to make a sound play when an image is clicked, and a problem solving task asked participants to make a label display text when an image is clicked. Participants had up to 25 minutes to complete the problem solving tasks.

Results and Discussion

For the problem solving tasks, participants received a score for number of correct steps taken towards problem solutions. For each correct action, such as adding code to play a drum sound, participants earned one point. Because the tasks involve multiple steps, scoring based on steps rather than whole answers provided more sensitivity. The maximum possible score was 25. Performance on the problem solving tasks depended on the interaction of subgoal learning method and feedback, $F(4, 190) = 3.39$, MSE = 23.6, $p = .01$, partial $\eta^2 = .067$. Due to the disordinal nature of this interaction, the main effects will not be reported to avoid confusion in interpreting the results (Maxwell & Delaney, 2004). Simple main effects comparisons were used to determine the effect of feedback on each method of learning subgoals. This analysis found that feedback affected the guided constructive groups, but it affected them in different ways (see Table 1). Participants in the guided constructive with hints conditions performed statistically better when they did not receive feedback than when they did, whereas participants in the guided constructive without hints conditions performed statistically better when they received feedback than when they did not.

Table 1: Simple main effects analysis of subgoal learning methods on problem solving performance. * indicates statistical significance at the .05 level.

Learning Method	Mean for No Feedback	Mean for Feedback	Mean Difference	Std. Error
Passive	15.5	17.4	-1.90	1.54
Active	18.0	16.1	1.95	1.54
Guided Constructive with Hints	21.0	17.5	3.50*	1.54
Guided Constructive without Hints	18.0	21.5	-3.54*	1.54
Unguided Constructive	18.0	18.3	-0.30	1.54

A simple main effects comparison was used for the feedback variable to explore the relative efficacy of different methods of learning subgoals,. Method of learning subgoals affected performance for groups that received feedback, $F(4, 190) = 3.54$, MSE = 23.6, $p = .008$, partial $\eta^2 = .069$, and groups that did not receive feedback, $F(4, 190) = 3.27$, MSE = 23.6, $p = .013$, partial $\eta^2 = .064$. Based on pairwise comparisons within the two types of feedback groups, those in the guided constructive with hints and without feedback condition performed statistically better than those in the passive condition without feedback, Mean Difference = 5.55, $p = .004$. Furthermore, participants in the guided constructive without hints condition and with feedback performed statistically better than those in the active condition with feedback, Mean Difference = 5.45, $p = .005$. These results suggest that, within both the feedback and no feedback groups, the best performing conditions scored statistically significantly better than those in the worst performing conditions. The other conditions that scored in the middle were not statistically better or worse than the best or worst performing conditions.

Overall, this pattern of results matched the expected pattern of results, suggesting that there is an optimal level of support for learning subgoals. In particular, the disordinal effect of feedback on the guided constructive groups suggests that learners perform best with just enough support and providing too much support hinders learning. Based on these results, it was concluded that providing hints for learners constructing subgoal labels and providing feedback on constructed labels are both techniques that can help learners to perform better on later problem solving, but providing both types of support could hurt performance.

Participant-Created Labels

To determine the quality of participant-created labels, they were qualitatively analyzed. Each label was analyzed as one unit (i.e., each word within a label was not analyzed individually), and each participant was categorized based on all of their labels collectively. In nearly all cases, all of the labels that a participant created fell into one of the following categories. The coding scheme included categories for whether labels were context-specific, context-independent, or incorrect. Context-specific labels included information about the specific instance of the subgoal and, therefore, could be applied only to that one instance. For example, the participant-created label "name and add picture to image sprite" could be applied only to the steps that named and added a picture to an Image Sprite. For a participant's labels to be classified as context-specific, at least 80% of their labels had to include information about the context. Context-independent labels, on the other hand, did not contain any information about the specific instantiation of that subgoal. For example, the participant-created label "add properties to app" is context-independent because it can be applied to any property, such as the name and picture of an Image Sprite, that is being added to the app. To be classified as context-independent, at least 80% of labels had to not include information about the context. Context-specific labels were considered to be of a lower quality than context-independent labels because they cannot be applied to novel problems. Context-independent labels indicate a more conceptual understanding of the procedure that is more easily applied to solving new problems.

Incorrect subgoal labels were those that were execution-based instead of function-based, such as "click on menu," or those that did not describe the correct function. To be classified as incorrect, more than one label had to meet either of these criteria. For the unguided constructive conditions, many of the subgoals that participants identified included many more steps than the subgoals created by experimenters. For example, some subgoals that participants grouped were more than 20 steps long, whereas the longest experimenter-grouped subgoal was seven steps. In all cases, the participant-created labels for these higher level subgoals were context-specific. For example, one participant identified a subgoal that was 24 steps long and labeled it "make the correct sounds play according to whatever input is received." To distinguish these labels from the other context-specific labels, these labels were classified as higher-level context-specific labels. The higher-level context-specific labels were considered lower quality subgoal labels than the context-independent or –specific labels. One of the benefits of learning the subgoals of a procedure is that subgoals break up long procedures into functional pieces that are easier to adapt to novel problems. The higher-level subgoals were not identifying these functional pieces but instead describing the procedure that was being executed.

Most participants in the guided constructive with hints conditions created context-independent labels (69%). Some of these participants created context-specific labels (22%) or incorrect labels (8%). Many participants in the guided constructive without hints conditions created context-independent labels (49%). A proportion of these participants created context-specific labels (27%) or incorrect labels (24%). The majority of participants in the unguided constructive conditions created higher-level context-specific labels (79%). A small number of these participants created context-independent labels (9%), context-specific labels (9%), and incorrect labels (3%).

Most of the participants in the guided constructive with hints conditions created subgoal labels that were similar to the experimenter-created labels, meaning that they created labels that aligned with those created through an intensive task analysis with a subject-matter expert. Therefore, it is not surprising that these participants did not benefit from feedback (i.e., experimenter-created labels) and performed well on the novel problem solving tasks. In fact, the condition that did not receive feedback performed better that the condition that did. Because participants created high quality labels, comparing their labels to the experimenter-created labels in the feedback might not have been as

beneficial as reviewing the labels that they constructed, as participants in the no feedback condition did. Comparing labels might have caused participants to unjustifiably question or doubt their understanding of the procedure, whereas reviewing their own labels would reinforce the mental representations that participants developed.

In summary, the results suggest that constructive methods of learning subgoals (i.e., self-explaining subgoals) are the most effective, but they require some instructional support. Either receiving feedback on constructed labels or receiving hints while constructing labels, but not both, led to the best problem solving performance. Participants who received hints while constructing labels were more likely to construct high quality labels than participants who did not receive hints. These participants performed better when they did not receive feedback than when they did, suggesting that the feedback provided too much instructional support to promote constructive learning. In contrast, participants who did not receive hints performed better when they received feedback than when they did not, suggesting that the feedback was necessary for the best performance when participants did not receive hints.

Acknowledgments

Our thanks to the reviewers who helped to improve this paper. We also thank Mark Guzdial and Briana Morrison for their help in guiding this research and Wesley Hughes and Ben Freed for their help in collecting and coding the data.

References

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*, 147-179.

Atkinson, R. K., Catrambone, R., & Merrill, M. M. (2003). Aiding transfer in statistics: Examining the use of conceptually oriented equations and elaborations during subgoal learning. *Journal of Educational Psychology, 95*(4), 762-773.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (2000). How people learn: Brain, mind, experience, and school: Expanded edition. Retrieved from http://www.nap. edu/catalog.php?record_id=9853

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory and Cognition, 22*, 605-615. doi:10.3758/BF03198399

Catrambone, R. (1996). Generalizing solution procedures learned from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1020-1031. doi:10.1037/0278-7393.22.4.1020

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127*, 355-376.

Catrambone, R., Gane, B. D., Adams, A. E., Bujak, K. R., Kline, K. A., & Eiriksdottir, E. (2016). Task analysis by problem solving (TAPS): A method for uncovering expert knowledge. Under review.

Catrambone, R. & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147-1156.

Chi. M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73-105.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. International *Journal of Artificial Intelligence in Education, 11*, 389-415.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science, 5*(2), 94-98.

Eiriksdottir, E., & Catrambone, R. (2011). Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors, 53*(6), 749-770. doi:10.1177/0018720811419154

Hundhausen, C. D., Farley, S. F., & Brown, J. L. (2009). Can direct manipulation lower the barriers to computer programming and promote transfer of training?: An experimental study. *ACM Transactions in CHI, 16*(3).

Margulieux, L. E. & Catrambone, R., (2014). Improving programming instruction with subgoal labeled instructional text. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.) Proceedings of the 35th Annual Conference of CogSci (pp. 952-957). Austin, TX: Cognitive Science Society.

Margulieux, L. E., Guzdial, M., & Catrambone, R. (2012). Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. Proceedings of the Ninth Annual International Conference on ICER (pp. 71-78). New York, NY: ACM. doi:10.1145/2361276.2361291

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). New York, NY: Psychology Press.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*, 529-556.

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers & Education, 46*, 426-445.

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. Mayer (Ed.) The Cambridge Handbook of Multimedia Learning, 2nd Edition (pp.413-432). Cambridge University Press.

The Effects of Gender Stereotypes for Structure Mapping in Mathematics

Kreshnik Nasi Begolli (keko@temple.edu)
Temple University, Department of Psychological, Organizational, and Leadership Studies, 1301 W. Cecil B. Moore Avenue Philadelphia, PA 19122 USA

Brooke Herd (bherd@uci.edu), Hannah Sayonno (hsayonno@uci.edu), Susanne Jaeggi (smjaeggi@uci.edu)
University of California, Irvine, School of Education, 3200 Education, Irvine, CA 92697 USA

Lindsey Engle Richland (lrichland@uchicago.edu)
University of Chicago, Department of Comparative Human Development, 5730 S. Woodlawn Ave., Chicago, IL 60637

Abstract

Fear of a negative stereotype about one's performance can lead to temporary underperformance on tests; e.g. women may underperform on a math test when prompted to think about gender. The current study extends this literature to examine whether stereotype threat not only leads to underperformance on tests, but also may impact reasoning and learning more broadly. We focus in particular on the effects of stereotype threat on analogical learning, a complex reasoning process that imposes a high working memory load. In this study, we examined the effects of gender stereotypes when females were asked to learn by comparing the mathematical concepts of combinations and permutations. Overall, participants given a threat before learning gained less from the instruction, as reflected by assessments administered immediately after the lesson and after a 1-week delay. This could lead to systematic differences in the quality of abstract representational knowledge for individuals from negatively stereotyped groups.

Keywords: Gender stereotype threat; analogy; comparison; mathematics education; video stimulus; working memory

Stereotype threat refers to the phenomenon in which members of a group fear that others will hold negative assumptions about their skills or capacities, by feeling additional pressure of conforming or disproving such beliefs (Spencer, Steele, & Quinn, 1999). Within education, one existing stereotype is that women naturally have lower abilities in mathematics than men, particularly on higher level mathematics concepts (Maloney & Beilock, 2012; McJunkin, 2009; Spencer, Steele, & Quinn, 1999).

The current literature suggests that working memory (WM) overload may be a primary mechanism explaining detrimental effects of gender stereotype threat on performance (Beilock, Rydell, & McConnell, 2007; Schmader & Johns, 2003). Participants are compromising their WM ability by anxious ideation about the activated stereotype, thus occupying their WM resources through thoughts such as "I had better perform well to disprove their stereotype about me as a woman." This ideation, while motivational, also engages verbal working memory, thus reducing available resources for task engagement.

Most studies of stereotype threat have focused on impact on test performance, examining the potential for a threat imposed immediately before a test to temporarily reduce performance and produce an unrepresentative measure of participants' performance (see Steele & Aaronson, 1995). Very little work has explored whether these threats might also impact learning, however, which would be a less temporary impact and might lead to more systematic differences in knowledge gains by negatively stereotyped groups. Thus, more research is needed to examine the effects of stereotype threat on learning.

To close this gap in the literature, this study examined the effects of gender stereotype threat on learning mathematics during high working memory load instructional analogies. Instructional analogies are common within mathematics education internationally (Richland, Zur, & Holyoak, 2007), and comparing problems or solutions is a common pedagogical recommendation (e.g., Common Core State Standards Initiative, 2012; National Research Council, 2001). However, for students to gain from these comparisons, they must have adequate available WM resources to represent two or more systems of relations, discover their alignments and map between the systems to draw inferences based on alignments (or misalignments) leading to encoding of a mathematical schema (Begolli, Richland, & Jaeggi, 2015; Gentner, 1983; Gick & Holyoak, 1983; Morrison, Doumas, & Richland, 2011; Waltz, Lau, Grewal, & Holyoak, 2000).

Structure mapping is particularly taxing when learners are not provided with adequate instructional supports to draw attention to key relational correspondences or when the content is not highly familiar (Begolli & Richland, 2016; Richland & McDonough, 2010; Rittle-Johnson & Star, 2007). Under higher WM demands, novices are more likely attend to and retain object features of compared representations rather than abstracting and retaining the common relational structure. In a previous experiment, Richland and McDonough (2010) tested the role of high versus low instructional cues within a video lesson in which a teacher made an analogy between the mathematical concepts of combinations and permutations. Students in the low-cued lesson (hypothesized to require higher WM demands) performed worse on cross-mapping problems on a posttest, suggesting their knowledge was less schematized.

Working memory is theorized to lie at the core of the observed effects of gender stereotype threat and learning from comparisons. Building on this work, the current study examined whether the pressure imposed by a threat of

gender stereotypes could reduce WM capacity that students normally rely on to learn a mathematical schema through structure mapping.

To test this hypothesis, this study exposed students to either a threat or neutral video lesson on permutations and combinations concepts. In order to ensure that the instructional analogy imposed high WM demands, we draw our stimuli from the low-cued video lesson utilized by Richland and McDonough (2010).

This work has theoretical implications for informing our understanding of the effects of stereotype threat on learning through structure mapping, possibly due to underlying shared mechanisms. Thus, laying the groundwork for further research to provide evidence linking research on structure mapping and stereotype threat. Further, this work has direct implications for informing teachers about the adverse effects of cultural pressures on learning processes.

Experiment: Impact of Gender Stereotypes for Instructional Analogy

Method

The research study was a randomized experimental study, with half of the participants being administered a gender stereotype threat and the other half receiving no threat. All participants received the same instruction subsequently. Learning was measured on an immediate posttest and one week later. Participants were not cognizant of the study's purpose of examining stereotype threat, as the study was advertised as a GRE math lesson on Combinations and Permutations.

Participants. Participants were 45 women who participated in the study either for course credit or monetary compensation. Only women who had not taken a college level statistics course were invited to participate in the study, to reduce the possibility of prior knowledge of permutations and combinations. The final analyses included 26 females in the threat condition and 19 in the non-threat condition (five participants in the non-threat condition did not complete all phases of the study).

Procedure. Participants were directed to a Google Forms Survey, which collected participants' demographic information and screened them based on their previous knowledge of statistics. If invited to participate, on Day 1, all participants were given 15-minutes to complete a pretest. They next answered a questionnaire, viewed an instructional lesson, and completed a posttest. Participants were randomly assigned to either the threat or non-threat conditions. The videotaped lesson and questionnaire were varied according to condition. Five to seven days later, participants were given 15-minutes to complete a delayed posttest. At the end, participants were debriefed about the true nature of the experiment.

Questionnaire. The questionnaire for the threat (experimental) group, primed female identity in order to reinforce the gender stereotype, taken from (Shih, Pittinsky, & Ambady, 1999). The control group received a similar questionnaire after their pretest, with the same format and number of questions. However, the questions only pertained to neutral topics, such as whether the participant watched cable television (Shih et al., 1999).

Instructional Videos. Each student watched a videotaped lesson on two mathematical concepts commonly included on the GRE: permutations and combinations. The instructional portion of the video lesson was identical to the low-cued video lesson used by Richland & McDonough (2010). From this, two versions of the video lesson were developed: 1) the threat condition containing introductory text providing a stereotype threat, and 2) the non-threat condition with a neutral introductory text. For the threat students, the video lesson began with a screen stating "Please watch the following 15-minute video very carefully," which was followed by "The lesson and the following tests have been shown to display gender differences in the past, exhibiting lower scores for women than men." The wording used to induce threat was modeled according to a study done by McJunkin (2009), which examined females' performance after being exposed to stereotype threat. The words remained on the screen for 30 seconds before the video lesson on combinations and permutations, which continued for approximately 13 minutes.

The video lesson for the control group began with "Please watch the following 15-minute video very carefully," and followed with "This lesson and the following tests have displayed no gender differences in the past; scores between women and men have been relatively equal"

During the instructional portion, an instructor discussed two mathematics problems and their solutions. The first exemplified the concept of combinations and the second the concept of permutations. Permutations involve finding the number of arrangements of a set of objects where the position (or order) is important. Combinations involve calculating the number of arrangements possible in which position (or order) is not important. Importantly, the instructor aligned and mapped between similarities in the permutation and combination concepts, while highlighting that the combination problems require an additional procedural step.

Table 1. Proportion correct at each time point by condition and question type. Standard deviations in parenthesis.

Condition	Pretest		Immediate						Delayed					
			Facilitory		Misleading		Unrelated		Facilitory		Misleading		Unrelated	
Threat	.10	(.16)	.69	(.35)	.33	(.24)	.40	(.32)	.46	(.34)	.40	(.28)	.50	(.28)
No Threat	.14	(.18)	.84	(.29)	.45	(.28)	.55	(.33)	.58	(.34)	.55	(.28)	.61	(.27)

Mathematics Assessment. The mathematics assessments had four types of problems: facilitory similarity, misleading similarity, unrelated, and distractors, discussed next.

Facilitory similarity. These questions refer to problems in which the story context and mathematical structure correspond between the video lesson and the posttest. This meant that the story context of the permutation problem in the instructional lesson (placements in a race), corresponded with the permutation problem on a posttest (placements in a race), and the combination problem story context during instruction was similar to the combination story context on the posttest.

Misleading Similarity. These questions refer to problems in which the story context of problems in the instruction were cross-mapped with problem contexts on the posttest. For instance if students learned the concept of permutations in the context of runners in a race, those students may attempt the same solution strategy on the posttest when faced with a problem that actually involved combinations but was set in the context of runners in a race, or vice versa. This meant that story context during instruction, such as teaching permutations in the context of a race, was followed by a combination problem on the posttest in the context of a race. If participants had coded the abstract relations within the problems they should perform similarly on the facilitory and misleading posttest problems. Those who had retained the surface features should be more likely to preform worse on misleading problems.

Unrelated Context. These problems were dissimilar in surface features to problems that were reviewed in the video, which tested participants' ability to apply combinations and permutations solutions in a completely different context.

Distractor problems. These problems involved factorials and advanced combinatorics problems, and were not analyzed.

Pretest. The pretest had 6 questions, 3 testing combinations and 3 testing permutations concepts. Because students did not watch the video lesson, these questions cannot be discussed in terms of facilitory and misleading similarities and were used as a baseline.

Immediate and Delayed Posttest. Both posttests had 10 questions, which contained the same story contexts, but were presented in an altered order with different numbers. Both tests contained 2 facilitory similarity problems, 2 misleading similarity problems, and 2 unrelated problems. Four problems functioned as distractors and assessed advanced combinatoric problems and factorials.

Answers from participants were coded based on whether the correct solution strategy was used to solve the questions. Correct strategy setup was scored as correct. Calculation errors were not scored, because our interest was to examine whether students knew which solution strategy was appropriate given the context. The assessments and coding was drawn directly from Richland & McDonough (2010).

Results

Raw means of the mathematics measures at each test point are summarized in Table 1. The random assignment of students was successful, with no significant differences in prior knowledge between the two groups (threat vs. no-threat) at pretest $F(1, 43) = .881$, $p = .353$.

We conducted a 2 (time of test: immediate to posttest) x 3 (question type: facilitory, misleading, and unrelated) x 2 (condition: Threat vs. No Threat) repeated measures analysis of covariance with time of test and surface similarity as within subject factors and condition as a between subject factor. Pretest score was utilized as a covariate and was a significant predictor when used in the model $F(1, 42) = 10.388$, $p = .002$, $\eta^2 = .198$.

As expected and shown in Figure 1, students performed worse under threat. There was a main effect of condition $F(1, 42) = 4.312$, $p = .044$, $\eta^2 = .093$ such that students who did not experience the gender threat used the correct strategy more often on all the problems ($M = .57$, SD = .31) than those who were under threat ($M = .44$, SD = .24). Also, problems with the same storyline and mathematical structure on both the lesson and the posttest were easier to solve than problems that were cross-mapped. Thus, there was a main effect of question type $F(2, 84) = 11.794$, $p = .006$, $\eta^2 = .223$, where a larger proportion of facilitory similarity problems were setup correctly than misleading similarity problems or unrelated problems (see Table 1).

Follow up pairwise comparisons confirm that facilitory similarity problems were set up correctly more often than both misleading similarity ($p < .001$) and unrelated problems ($p = .001$). There was also a trend suggesting that misleading similarity problems were more likely to lead to errors when compared to unrelated problems ($p = .051$).
There was no main effect of time $F(1, 42) = 2.041$, $p = .161$, $\eta^2 = .046$, such that students' overall scores were not significantly different between immediate test ($M = .56$, SD

=.21) and delayed test (M =.44, SD =.25). However, there was an interaction between time of test and question type $F(2, 84) = 7.968$, $p = .001$, $\eta^2 = .280$. Students were better at using the correct solution strategies for facilitory problems immediately after the lesson than 1-week later. The inverse was true for misleading and unrelated problems. The decrease for facilitory similarity is likely because the problem story contexts during instruction became less salient over time and thus less helpful at posttest (see Table 1).

There was no interaction between question type and condition F (2, 84) = .005, p = .995, η^2 = .003 and no interaction for time and condition F (2, 42) = .015, p = .682, η^2 = .004. No other interactions were expected to be significant and our results confirmed this expectation.

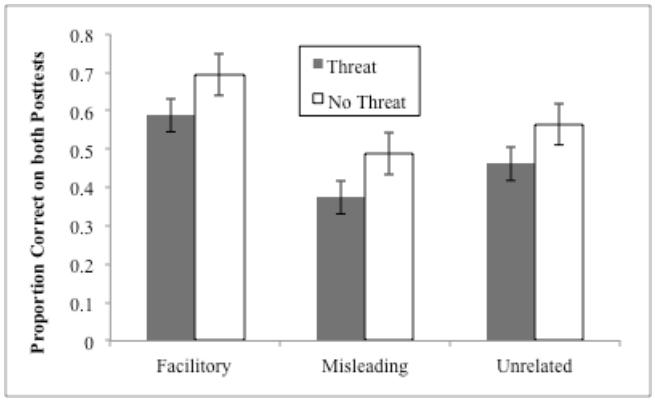


Figure 1. Adjusted means for accuracy by condition for each question type across immediate and delayed posttests using pretest score as covariate. The bars represent standard errors.

Discussion

The study reveals that learning from analogy may be susceptible to pressures from social context, and that effects can be lasting rather than temporary. Many studies have examined the role of stereotype threat on performance, but this study is one of the first to show that threat of negative stereotypes can impact *learning*. Unlike previous accounts documenting the detrimental effects of stereotype threats on women's mathematics test performance (McJunkin, 2009; Steele & Aronson, 1995), inducing the threat before a learning event and examining outcomes after a 1-week delay provides new insight into the impact stereotypes can have for schema formation.

Our data across two time points support that students who were exposed to stereotype threat were significantly worse at learning combinations and permutations concepts in comparison with students who had not been exposed to stereotype threat, and that these differences lasted over a week delay.

The mathematics outcomes align with current views suggesting that concern about negative stereotypes may particularly impair performance on advanced mathematics concepts, more than lower working memory problems that merely require straightforward calculations (see Malone & Beilock, 2012).

A potential concern regarding our paradigm is that participants' immediate posttest performance may have been impaired not due to learning differences, but due to the same immediate performance pressure effects observed in traditional studies of stereotype threat on test performance (Beilock, Rydell, and McConnel, 2007). At the same time, any state differences based on the threat should have dissipated by a 1-week delay. Thus, our results supported our hypothesis, in that stereotype threat impeded the learning of mathematics through structure mapping, rather than artificially decreased test performance. An alternative explanation for decreased performance could be that the effects of threat did not impede structure-mapping processes per se, but broader learning and/or encoding, considering the effects of threat were similar on facilitory and misleading problems.

In line with previous findings in the analogy literature, students in our sample were better able to solve problems where the context between the lesson and the posttest was aligned. This shows that students were successful at encoding the solution strategies involving permutations and combinations concepts. However, students performed worse on problems where the context was cross-mapped between the video lesson and the posttest or the problem context was unrelated to problems in the video lesson.

The question type differences are also consistent with Richland and McDonough's (2010) findings. Misleading similarity problems may have placed a greater demand on WM, particularly inhibitory processes, since students would need to inhibit their inclination to solve based on surface similarities, in order to apply the correct solution strategy. Yet, Richland and McDonough (2010) postulate that higher WM demands, induced by the lack of instructional supports, led students to misapply solution strategies only on misleading similarity problems. It is possible that if students in our study learned from a lesson with high instructional supports, the effects of threat would dissipate for the facilitory questions. Thus, manipulating WM demands through the video lesson in addition to the problem types (facilitory & misleading) and assessing WM or inhibitory processes would strengthen our findings. In general, our results also point towards WM as an important component of analogical thought or broader learning processes, and highlight that these may be mechanisms through which stereotype threat reduces long-term learning and schema formation.

Implications for Future Research

The paradigm used in this study, administering a stereotype threat before an opportunity for analogical learning, provides a platform for ongoing studies that aim to

shed light on the impact of pedagogical supports and WM resources on structure-mapping and relational reasoning.

For example, Beilock & Carr (2005) found that stereotype threat affects mathematics performance for women with higher WM capacities. This suggests that women who are most likely to succeed underperform when under threat. While this has tremendous implications for test taking, it may have even greater implications if students face threat throughout their learning experience. Thus, students with greater potential could learn significantly less, which could accumulate over time, resulting in significant achievement gaps. Our data coupled with WM measures and a larger sample size may provide insight into the mechanisms and behavioral effects of stereotype threat on learning.

On the other hand, there is mounting evidence that instructional supports offload WM demands, which may interact with the effects of gender stereotypes. The lack of instructional supports may be particularly beneficial for lower performing students and those with lower WM abilities. Considering that gender stereotypes most significantly affect females with high WM capacities, we may find that the effects of threat differ based on instructional supports and WM capacity. In the current study, we only used the lowly-cued version of Richland & McDonough's (2010) video lesson when inducing stereotype threat. A continuation of this study will compare the effects of stereotype threat when learning from a highly-cued video lesson. The potential implications of this work are that competent teachers, who implicitly induce threat, may help lower performing students at the expense of higher performers.

Overall, these future studies seek to untangle the effects of stereotype threat and its possible cognitive mechanisms, under varying instructional contexts, and clarify their role for learning mathematics through structure-mapping. While we discuss mathematics more broadly, combinations and permutations represent only one mathematical domain, and further studies are needed to understand whether these results generalize to other mathematical concepts. Despite this, it is important to note that the stimulus used (the math video lesson) does approximate a true lesson experience. The instruction was not text-based, and came from a live instructor, thus suggesting greater ecological validity for generalizing to teacher actions in a classroom setting.

This research has important implications for teachers in the classroom. Teachers must be aware of the culture they are creating within their classroom and should support and facilitate learning through teaching methods that utilize high instructional supports and reduce any hindrances of societal pressures that might reduce learners' abilities.

In sum, these research findings not only reveal the potentially profound effects of stigma in a learning context but also suggest various research directions for further studies. This research creates a platform for further investigating the role of working memory as an underlying mechanism of the pressures of social stigma. Furthermore, our results lead to new research questions regarding stereotype threat and teaching methods. Questions are left to be asked regarding the impact of high or low instructional teaching methods in aiding or preventing stereotype threat from impeding learners, as well as whether individual characteristics, such as innate and perceived ability and motivation, can alter the effects stereotype threat can have on a learner.

References

Begolli, K. N., & Richland, L. E. (2016). Teaching Mathematics by Comparison: Analog Visibility as a Double-Edged Sword. *Journal of Educational Psychology*, *108*(2), 194–213.

Begolli, K. N., Richland, L. E., & Jaeggi, S. M. (2015). The Role of Executive Functions for Structure-Mapping in Mathematics. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warfaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Vol. 1, pp. 1–6). Austin, TX: Cognitive Science Society.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail working memory and "choking under pressure" in math. *Psychological Science*, *16*(2), 101–105.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: mechanisms, alleviation, and spillover. *Journal of Experimental Psychology. General*, *136*(2), 256–276. http://doi.org/10.1037/0096-3445.136.2.256

Common Core State Standards Initiative. Common core state standards for mathematics (2012). Common Core State Standards Initiative.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive Science*, *7*(2), 155–170. http://doi.org/10.1016/S0364-0213(83)80009-3

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. http://doi.org/10.1016/0010-0285(83)90002-6

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, *16*(8), 404–406.

McJunkin, L. M. (2009). Effects of stereotype threat on undergraduate women's math performance: Participant pool vs. classroom situations. *Emporia State Research Studies*, *45*(2), 27–31.

Morrison, R. G., Doumas, L. A. A. L. a a, & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, *14*(3), 516–529. http://doi.org/10.1111/j.1467-7687.2010.00999.x

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics*. The National Academies Press. Retrieved from http://www.nap.edu/catalog/9822/adding-it-up-

helping-children-learn-mathematics
Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*(1), 28–43. http://doi.org/10.1016/j.cedpsych.2009.09.001
Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128.
Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561–574. http://doi.org/10.1037/0022-0663.99.3.561
Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440.
Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*(1), 80–83.
Shih, M., Pittinsky, T. L., & Trahan, A. (2006). Domain-specific effects of stereotypes on performance. *Self and Identity*, *5*(1), 1–14.
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.
Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*. http://doi.org/10.1037/0022-3514.69.5.797
Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*(7), 1205–1212.

Improving with Practice: A Neural Model of Mathematical Development

Sean Aubin (saubin@uwaterloo.ca)
Aaron R. Voelker (arvoelke@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada N2L 3G1

Abstract

The ability to improve in speed and accuracy as a result of repeating some task is an important hallmark of intelligent biological systems. Although gradual behavioural improvements from practice have been modelled in spiking neural networks, few such models have attempted to explain cognitive development of a task as complex as addition. In this work, we model the progression from a counting-based strategy for addition to a recall-based strategy. The model consists of two networks working in parallel: a slower basal ganglia loop, and a faster cortical network. The slow network methodically computes the count from one digit given another, corresponding to the addition of two digits, while the fast network gradually "memorizes" the output from the slow network. The faster network eventually learns how to add the same digits that initially drove the behaviour of the slower network. Performance of this model is demonstrated by simulating a fully spiking neural network that includes basal ganglia, thalamus and various cortical areas. Consequently, the model incorporates various neuroanatomical data, in terms of brain areas used for calculation and makes psychologically testable predictions related to frequency of rehearsal. Furthermore, the model replicates developmental progression through addition strategies in terms of reaction times and accuracy, and naturally explains observed symptoms of dyscalculia.

Keywords: neural engineering framework; semantic pointer architecture; nengo; cognitive modelling; mathematical ability; dyscalculia; skill consolidation

Introduction

Adaptability and scalability are two central challenges in cognitive modelling. Building a model that performs some behaviour is a daunting enough task, even without considering how it might improve over time. This is made significantly more challenging when building a large-scale biologically plausible model that uses spiking neurons to represent and communicate information over time.

One example of such a biologically plausible model is the Semantic Pointer Architecture Unified Network (Spaun), which is currently the largest behaving model of the human brain (Eliasmith et al., 2012). This model flexibly performs eight different cognitive tasks by receiving images of handwritten digits through its simulated retina and outputting responses by drawing using its simulated arm. This model makes some progress in addressing the challenge of scalability (Eliasmith, 2013), but lacks the ability to learn from previously experienced cognitive tasks to permanently improve its performance (Spaun only changes its long term connection weights during a simple reinforcement learning task). In this work, we show that one task from Spaun's repertoire, *addition-by-counting*, can be extended to benefit from this cognitive ability.

For addition-by-counting, the model is presented with two digits and asked to draw the digit that corresponds to the sum of these two digits. The "counting" strategy methodically computes the result by adding one to a digit, for the number of times indicated by the other digit. In Spaun, this is accomplished by a number of structures that function together. Visual cortical areas compress the representation of a given image into a semantic representation, referred to as a *semantic pointer*. Prefrontal cortical areas transform these pointers, while maintaining partial results in working memory, and a basal ganglia and thalamus control loop selects actions to coordinate and drive the behaviour of the system. The cortical areas of Spaun should, but are currently unable, to recognize and learn from previous instances of this problem. Including such learning should translate to performance improvements in both speed and accuracy.

DeWolf & Eliasmith (2013) have presented a neural model in which a simple motor skill is consolidated into cortex via repeated practice. However, the skill being learned in that case cannot be extended to solve the addition task because it is unclear how a simple motor action could be generalized to counting from any digit to any other. Consequently, here we consider how consolidation can happen for more sophisticated representations, and argue that such representations may help to explain more complex cognitive phenomena. This is demonstrated by proposing a spiking neural model that exhibits gradual performance improvement on the addition-by-counting task.

Mathematical Development

The development of numeracy in children is not a simple progression of skills. For instance, children learn to count before they understand how sets relate to numbers. As well, it is thought that after learning to count, they progressively learn the relationship between set sizes and numbers until the age of five, after which they have a complete understanding of the number scale (Sarnecka & Carey, 2008). This same sort of complex learning development can be observed in children learning addition. Typically, children progress through various strategies before finally memorizing the results of addition, as shown in Table 1 (Siegler, 1987).

The *Counting* strategy involves choosing the larger number and incrementing it a number of times equal to the smaller number. The *Recall* strategy is where the two numbers form an association to a previously memorized answer, which is then recalled from long-term memory. Other strategies have been identified, although we focus here on recall and count-

Table 1: Percentage of addition strategy use by grade level (Summarized from Siegler (1987)).

Grade level	Counting	Recall	Guess or no response	Other
Kindergarten	30 %	16 %	30 %	24 %
Grade 1	38 %	44 %	8 %	10 %
Grade 2	40 %	45 %	5 %	11 %

Table 2: Median solution times (seconds) per addition strategy use by grade level (Summarized from Siegler (1987)).

Grade level	Counting	Recall
Kindergarten	6.0 s	3.9 s
Grade 1	6.9 s	2.1 s
Grade 2	3.9 s	1.8 s

Table 3: Percentage of errors per addition strategy use by grade level (Summarized from Siegler (1987)).

Grade level	Counting	Recall
Kindergarten	19 %	29 %
Grade 1	4 %	17 %
Grade 2	3 %	7 %

ing, since the greatest developmental change is seen with these two strategies. Additionally, for the sake of simplicity, we focus on the progression from the counting strategy to recall, using sums less than ten. As demonstrated below, the model is initially entirely reliant on the counting strategy, but gradually learns the recall strategy, causing an increase in accuracy and a decrease in reaction times, consistent with the data shown in Table 2 and Table 3.

To this point we have characterized the typical developmental path for children. However, there are individuals who suffer from dyscalculia. Dyscalculia is a learning disability characterized by various problems with numeracy, one of which can be understood as difficulty making the transition from counting to retrieval. Our model demonstrates how, without a parallel learning mechanism, symptoms of dyscalculia can arise. In particular, our model explains increased activation of the prefrontal cortex compared to individuals with normal numeracy (Kucian & von Aster, 2015) and a lack of progression to recall based strategies.

Neural Representation of Digit Semantics

To build our model, we make use of the Neural Engineering Framework (NEF; Eliasmith & Anderson (2003)) and the Semantic Pointer Architecture (SPA; Eliasmith (2013)) that were both used to build Spaun.

The NEF may be understood as a "neural compiler" for mathematical functions using vector spaces. The presented model relies on two key principles of the NEF as a method for constructing networks of spiking neurons and connection weights from a mathematical description of the system. The first principle describes how a vector can be mapped onto a distributed representation over neurons. The second principle characterizes how connections between neurons can transform these vectors.

A vector $\mathbf{x}(t)$ is represented by encoding it into the spiking activity of a population of neurons. Each neuron i has an encoding vector $\mathbf{e}_i$ (which can be understood as a preferred direction in the vector space), a gain α_i, and a background current J_i^{bias}. These parameters determine how the input vector is translated into the input current $J_i(t)$ to a neural nonlinearity $G_i[\cdot]$. For our work, this neural nonlinearity is the leaky integrate-and-fire (LIF) model, which converts the input current into a neural spike train $a_i(t)$.

$$a_i(t) = G_i[J_i(t)], \;\; J_i(t) = \alpha_i \mathbf{e}_i \cdot \mathbf{x}(t) + J_i^{bias} \tag{1}$$

To decode an approximation of the vector back from these spike trains, they are first convolved with a low-pass filter $h(t)$ (a decaying exponential modelled after the postsynaptic current) and then multiplied by a decoding vector $\mathbf{d}_i$.

$$\hat{\mathbf{x}}(t) = \sum_i \mathbf{d}_i (a_i * h)(t) \tag{2}$$

The decoders $\mathbf{d}_i$ are found using regularized least squares optimization to minimize the error over the range of inputs $\mathbf{x}$:

$$\int (\mathbf{x} - \hat{\mathbf{x}})^2 \, d\mathbf{x}. \tag{3}$$

To describe how two neural ensembles are connected, we define a weight matrix as the outer product of the encoders and decoders $\omega_{ij} = \mathbf{e}_i \bigotimes \mathbf{d}_j$. The second principle then shows that the input current to neuron i may be rewritten as a weighted summation of its postsynaptic potentials, allowing for the connection of multiple populations of neurons:

$$J_i(t) = \sum_j \alpha_i \omega_{ij} (a_j * h)(t) + J_i^{bias}. \tag{4}$$

To apply arbitrary transformations to the vector using these connections, we can minimize the decoding error for the desired function:

$$\int (f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2 \, d\mathbf{x}. \tag{5}$$

Together, these two principles allow us to construct spiking neural models of arbitrary functions of vector spaces.

To apply the NEF to cognitive models, the SPA suggests specific architectural components and organization as well as a general kind of representation called a semantic pointer. Semantic pointers are compressed neural representations that can be efficiently manipulated, transformed, and dereferenced to retrieve deep semantic information. It has been

suggested that this approach addresses the symbol grounding problem by defining the semantics of a neural representation as resulting from the compression of sensory information as well as conceptual relations (Eliasmith, 2013). Consequently, the SPA lends itself especially well to representing concepts which are grounded in multiple modalities but unified in a single representation. In this model, the semantic pointers are used to represent digits, which have been shown to be grounded in auditory and visual modalities (Nieder, 2012), but also include conceptual relationships (e.g. TWO comes after ONE but before THREE). Conceptual relationships are captured in the SPA by adopting a compression operator from a specific type of Vector Symbolic Architecture (Gayler, 2004) called a Holographic Reduced Representation (Plate, 1995).

Modelling the Counting Strategy

Although Spaun serves as the starting point for our model, it is not practical to simulate all 2.5 million of its neurons. Instead, a counting circuit based on Spaun's design was implemented. As shown in Figure 1, after a question input is received, the procedure consists of three main steps:

1. The digit contained in the working memory neural population is routed to a "transformation system".
2. The digit is transformed by a heteroassociative "incrementing memory" to produce the semantic pointer corresponding to the incremented digit.
3. The incremented number is returned to working memory.

This process is continued until the "Counts finished" are equal to the "Total counts to take". At which point the value in "Count result" is routed as the final answer to the output.

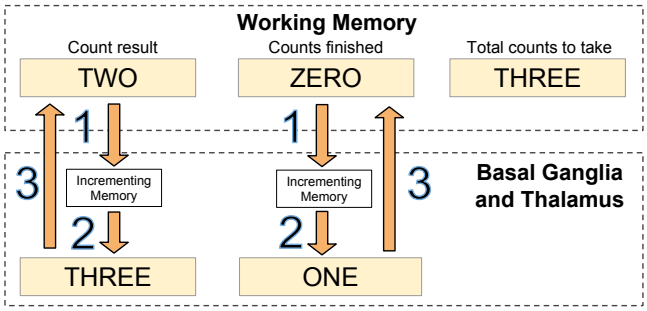


Figure 1: High-level overview of the addition-by-counting procedure. See text for details.

These steps are controlled by an action selection system implemented in the basal ganglia and thalamus (Stewart et al., 2010) in a manner similar to Spaun.

To increment a digit, a predefined heteroassociative memory is used to associate each semantic pointer with its incremented pointer. This is implemented using a single population of neurons, with encoders tuned to each vector of the numerical vocabulary and decoders chosen to apply a transform to the output to compute the incremented semantic pointer. Finally, the output layer uses mutual inhibition to form a winner-take-all (WTA) mechanism, which ensures the transformation system returns only a single pointer. This design is different from Spaun's, which relied on the repeated convolution of a single vector to indicate a count. While both approaches are viable, an associative memory is more effective for simpler, low-dimensional models.

When using a heteroassociative memory, the semantic pointer corresponding to each digit is taken from a random ten-dimensional orthonormal basis. This ensures that no prior information is given to bias the relationship between the ten digits. Initially, the model only knows how to increment a digit, as is the situation for learning addition during childhood using solely the counting strategy.

The learned heteroassociative memory is identical to the aforementioned predefined memory, except the decoders can be learned with experience, so that no action selection mechanism is required to cycle through mappings until the desired result is reached.

Memorization via Reinforcement Learning

The counting circuit that employs the predefined memory is slow, since information is routed back and forth at the rate of subvocal rehearsal. We thus refer to the counting portion of the model as the "Slow-Net". In the model, other cortical areas consolidate the function of the Slow-Net by memorizing its eventual responses. The portion of the network responsible for the storage and subsequent retrieval of originally counted answers will be referred to as the "Fast-Net".

The purpose of the Fast-Net is to learn to associate the two digits provided as input to the Slow-Net with the Slow-Net's eventual response. This heteroassociative memory is highly nonlinear with respect to the vectors that represent each digit. Knight et al. (submitted 2016) has shown that by combining the supervised Prescribed Error Sensitivity (PES; MacNeil & Eliasmith (2011)) learning rule with the unsupervised Vector Oja (Voja; Voelker et al. (2014)) learning rule, we can scalably and efficiently learn such complex functions in the NEF. The Fast-Net leverages the unsupervised learning rule to efficiently represent the incoming addends and uses the supervised learning rule to learn the correct sum from the Slow-Net output.

Specifically, the PES learning rule minimizes the difference between the output of a neural population and its desired value, by adjusting its decoders ($\mathbf{d}_i$ from (2)) in response to an error signal. However, for discontinuous high-dimensional functions such as the desired heteroassociative memory, this supervised learning rule is insufficient. This is because a neuron will often fire in response to multiple inputs, in which case its decoder will be adjusted to completely overwrite the past association. To avoid this, a neuron should only fire for a single input, which is achievable by selecting the encoders ($\mathbf{e}_i$ from (1)) to be equal to the semantic pointers for each of the digits ahead of time, as is done in the predefined heteroasso-

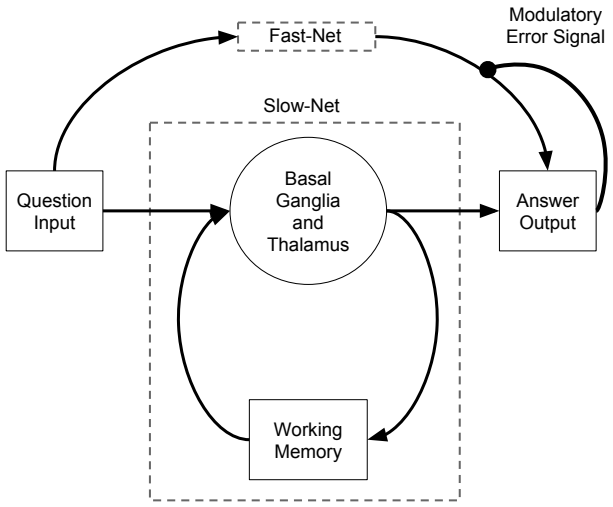


Figure 2: A high-level view of the model, featuring the parallel Slow-Net and Fast-Net.

ciative memory. However, this would assume that the area of cortex designated to learn the addition task is already aware of the set of possible inputs to the Slow-Net.

To keep our approach general, the Voja rule learns to form a sparse encoding of the possible inputs as they are presented to the Fast-Net. This is achieved by adjusting the encoders of any active neurons to become selective to only the current input. This allows PES to learn the correct output without overwriting past associations, while being able to scale to recall over 100,000 associations (Knight et al., submitted 2016).

In our model, to enable simultaneous learning and operation, the Fast-Net is placed in parallel to the Slow-Net, as shown in Figure 2. Both networks receive the same input, but the answer from the Slow-Net is projected to modulate the output of the Fast-Net. This *modulatory error signal* is triggered whenever the Slow-Net responds with an answer, which then provides the feedback for learning via PES. In summary, the population of neurons in the Fast-Net memory continually learns to represent the input digits by adjusting its encoders, while its decoders adjust to associate its input with the output of the Slow-Net.

The learning rate of the heteroassociative memory can be adjusted to model developmentally plausible learning. At high learning rates the model learns mappings after being shown only a single example and at lower learning rates it gains confidence gradually, covering the spectrum of human variability and demonstrating the versatility of the heteroassociative memory. Given the focus of this paper on increasing the accuracy and speed of reaction, while ignoring realistic amounts of rehearsals required to learn, an artificially high learning rate was chosen for the simulation.

Results

The model was built and simulated using Nengo (Bekolay et al., 2014). The results of the Slow-Net, which implement the counting strategy, are shown in Figure 3.

As expected, the magnitude of error between the answer from Fast-Net and the correct answer decreases with rehearsal, as shown in Figure 4. Each epoch of training consists of 20 randomized example additions with no repetition. Give the fast learning rate of the model, after each epoch there is a significant drop in error (during each subsequent epoch, the model is seeing problems it has encountered before). Note that after the magnitude of error drops below 0.5, the Fast-Net response will drive the model's response instead of Slow-Net, since it has the correct answer. Which network drives the overall response is determined by the basal ganglia. Notably, any decrease of error magnitude past 0.5 reflects an increase in the certainty of the answer.

Additionally, once the Fast-Net becomes confident enough in its responses to over-ride the Slow-Net, the speed of responses becomes faster and more uniform, as shown in Figure 5. The confidence of the response is determined by its similarity to any known number representation (correct or incorrect). If the response is incorrect, environmental feedback will drive the Fast-Net to change its answer. Given the high learning rates, such errors are rapidly corrected. The fact that reaction times regularize with transition to a recall-based strategy matches experiments investigating addition strategies (Siegler, 1987).

Neuroanatomical Mappings and Dyscalculia

As discussed in Spaun's mapping of counting (Eliasmith et al., 2012), parietal areas are more active for stable, learned transformations while prefrontal areas are more active for transient, working memory representations. Given that the Fast-Net responses eventually replace those from the Slow-Net, we would expect that with practice brain activity will move from the prefrontal cortex to frontal-parietal areas.

Although this is quite difficult to measure during the natural development of a child, this hypothesis is supported by the observation that those with dyscalculia show greater activation of the prefrontal cortex compared to individuals with normal numeracy (Kucian & von Aster, 2015). Although this model make no claims about why dyscalculia occurs, given that it is a complicated disability usually accompanied by various comorbidities, it does provide a possible explanation as to why such compensation occurs. Specifically, those with dyscalculia are unable to consolidate the functional role of the prefrontal cortex during the counting task within the frontal-parietal region and must instead rely on their working memory. Given an excessively noisy input, inaccurate feedback, or inappropriate modulation of the error signal, the Fast-Net could fail to learn the mapping between addends and sum. Consequently, there would be limited progression from counting to recall and a continued dependence on working memory.

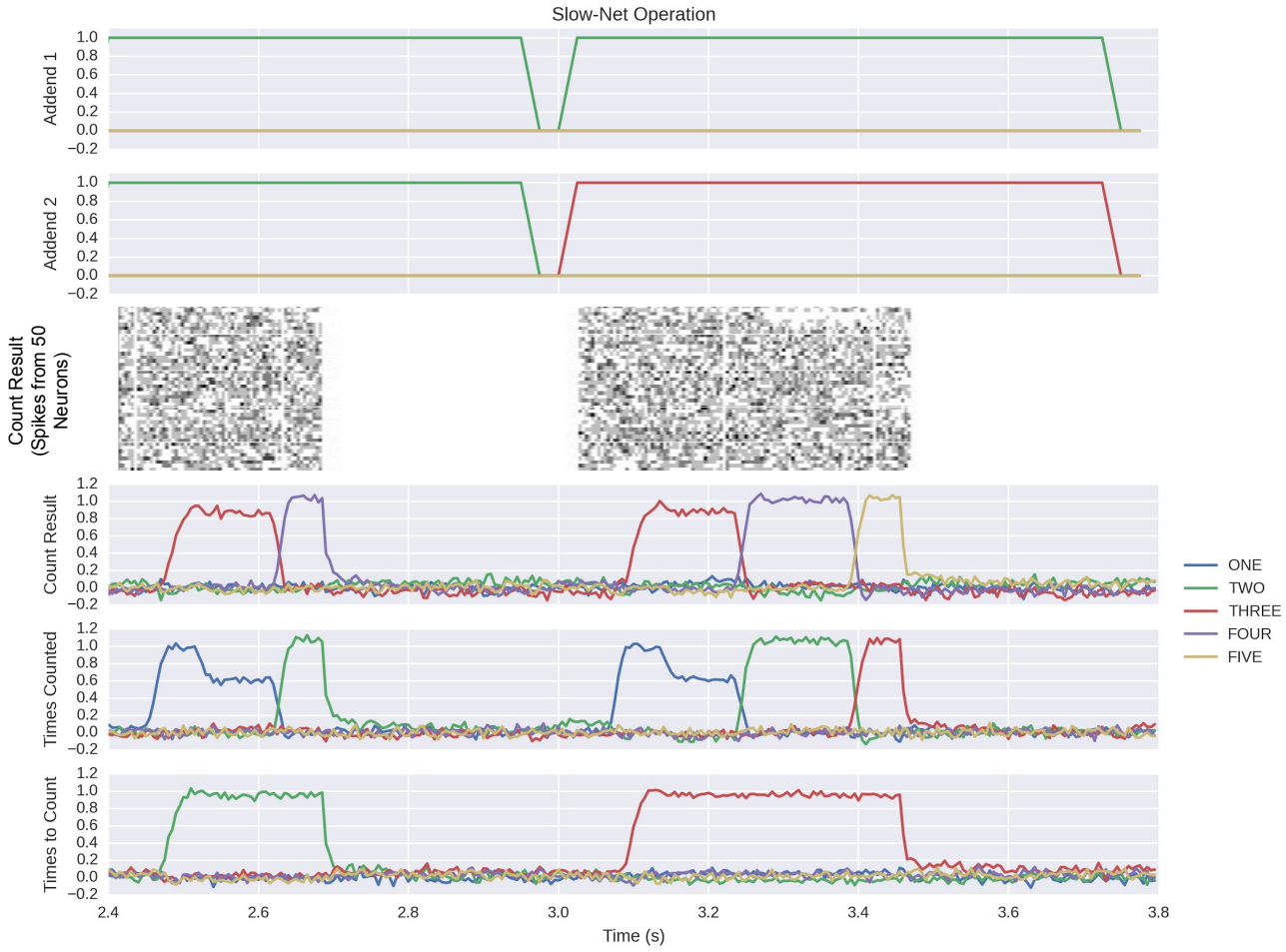


Figure 3: The Slow-Net (counting network) answering the questions $2+2$ and $1+3$. The plots show similarity of neural activity to semantic pointers over time (colours indicate which pointer). As shown in "Count Result" the model is progressing through each intermediate digit before reaching the answer. Lines are the similarity between the spiking pattern in the area and the ideal spiking pattern for each number in the numerical vocabulary.

Discussion and Conclusions

We presented a biologically plausible model that progresses from a methodical counting procedure to recalling the response for an addition task. As specified in Table 3 and shown in Figure 4, the accuracy of the Fast-Net memory progresses from noisy to accurate. Once a sufficient accuracy threshold is reached, the memory takes over the process of addition from the Slow-Net, increasing the reaction times.

These simulations suggest the following prediction: that performance should increase in proportion to the frequency of the presented addends. That is, improvement in learning should be proportional to the amount of rehearsal, as well as the familiarity of the given addends. For example, continual presentation of sums with the first addend of "3" should allow for faster learning of other sums involving the addend "3", since the encoders of the heteroassociative memory will have already adapted to represent this digit.

The model as presented is clearly limited in several respects. For instance, numerical representation in the brain consists of more structure than the randomly selected orthonormal vectors used here. This assumption is reasonable for small magnitudes, but is untenable for numerical representation in general. For instance, there is evidence that neurons tuned to numerical size comparisons are proportional to a log scale and can be highly sensitive to task saliency (Nieder & Dehaene, 2009). Capturing such properties will require different numerical representations.

Finally, this model only describes a handful of the brain areas associated with numerical calculation. One of the most surprising areas involved in addition is the cerebellum. It has been suggested that cerebellar activity might be a developmental artifact persisting from when addition is first learned

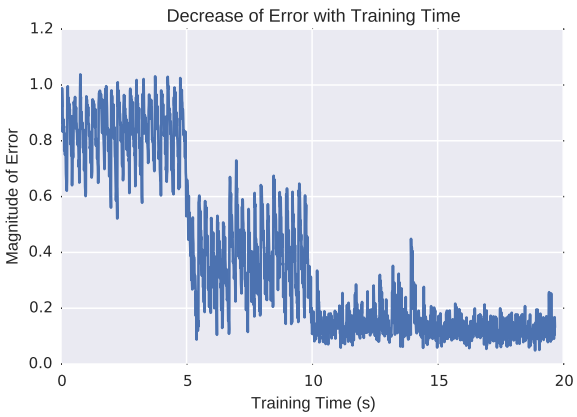


Figure 4: Error magnitude in Slow-Net decreasing with training received from the feedback of each trial. Note the drop in error after each training epoch.

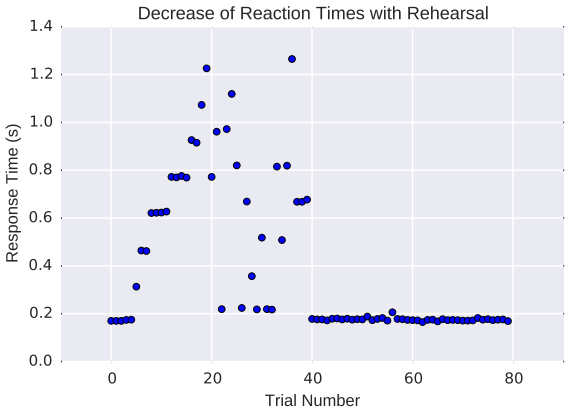


Figure 5: Reaction times decreasing with rehearsal as the Fast-Net takes over for the Slow-Net for increasingly more additions.

as a physical combination of objects (Arsalidou & Taylor, 2011). To model cerebellar involvement, counting and grouping objects would need to be rehearsed via explicit motor plans. This could be captured with a more in-depth model that includes a motor control hierarchy and a visual system similar to those found in Spaun.

Acknowledgments

This work was supported by CFI and OIT infrastructure funding as well as the Canada Research Chairs program, NSERC Discovery grant 261453, AFOSR grant FA8655-13-1-3084 and NSERC graduate funding.

References

Arsalidou, M., & Taylor, M. J. (2011). Is 2+ 2= 4? meta-analyses of brain areas needed for numbers and calculations. *Neuroimage*, *54*(3), 2382–2393.

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., ... Eliasmith, C. (2014). Nengo: A python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*(48). doi: 10.3389/fninf.2013.00048

DeWolf, T., & Eliasmith, C. (2013). A neural model of the development of expertise. In *12th international conference on cognitive modelling.* Ottawa, Ontario.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*, 1202-1205. doi: 10.1126/science.1225266

Gayler, R. W. (2004). Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. *arXiv preprint cs/0412059*.

Knight, J., Voelker, A. R., Mundy, A., Eliasmith, C., & Furber, S. (submitted 2016). Efficient SpiNNaker simulation of a heteroassociative memory using the Neural Engineering Framework. In *The 2016 international joint conference on neural networks (IJCNN).* IEEE, submission N-16403.

Kucian, K., & von Aster, M. (2015). Developmental dyscalculia. *European journal of pediatrics*, *174*(1), 1–13.

MacNeil, D., & Eliasmith, C. (2011). Fine-tuning and the stability of recurrent neural networks. *PloS one*, *6*(9), e22885.

Nieder, A. (2012). Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proceedings of the National Academy of Sciences*, *109*(29), 11860–11865.

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual review of neuroscience*, *32*, 185–208.

Plate, T. A. (1995). Holographic reduced representations. *Neural networks, IEEE transactions on*, *6*(3), 623–641.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, *116*(3), 250.

Stewart, T. C., Choo, X., & Eliasmith, C. (2010). Symbolic reasoning in spiking neurons: A model of the cortex/basal ganglia/thalamus loop. In *32nd annual conference of the cognitive science society.*

Voelker, A. R., Crawford, E., & Eliasmith, C. (2014). Learning large-scale heteroassociative memories in spiking neurons. In *13th international conference on unconventional computation and natural computation.* London, Ontario.

Examining Referential Uncertainty in Naturalistic Contexts from the Child's View: Evidence from an Eye-Tracking Study with Infants

Yayun Zhang and Chen Yu
yayzhang@indiana.edu, chenyu@indiana.edu
Department of Psychology & Brain Science, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

Young infants are prolific word learners even though they are facing the challenge of referential uncertainty (Quine, 1960). Many laboratory studies have shown that human infants are skilled at inferring the correct referent of an object from ambiguous contexts (Swingley, 2009). However, little is known regarding how children visually attend to and select the target object among many other objects in view when parents name it during free play interactions. In the current study, we explored the looking pattern of 12-month-old infants using naturalistic first person images with varying degrees of referential ambiguity. Our data suggest that infants' attention is selective and they tend to only select a small subset of objects to attend to at each learning instance despite the complexity of the data existed in the real world. This work allows us to better understand how perceptual properties of objects in infants' view influence their visual attention, which is also related to how they select candidate objects to build word-object mappings.

Keywords: statistical learning; word-referent mapping; learning mechanisms

Introduction

Infants encounter words in complex environment and one challenge in early word learning is that of referential uncertainty: how infants manage to find the right word-referent pairs in the noise (Quine, 1960). Many studies have shown that human infants are able to infer the correct referent of an object from ambiguous contexts (Swingley, 2009; Waxman & Booth, 2001). Using the cross-situational word-learning task, Smith and Yu (2008) have found that infants can learn word-referent pairs by computing distributional statistics across the co-occurrences of words and referents at multiple naming moments, suggesting that infants attend to and systematically store the co-occurrence information during training. Additional evidence demonstrates that infants keep track of not only the strongest available associations but also low-frequency information, which further supports the notion that infants are sensitive to the co-occurrence statistics between words and referents and they keep track of a system of associations (Vouloumanous & Werker, 2009).

While laboratory tests have led to significant advancement in our understanding of the underlying word-learning mechanism, cognitive scientists have also started to investigate word learning using more naturalistic data. One interesting ongoing discussion in the literature is centered on the question of how noisy our daily environment is. Medina, Snedeker, Trueswell, and Gleitman (2011) argue that learners encounter words in complex environments where infinite referents might be treated as the label's correct referent, therefore co-occurrences in the real world are too noisy to be effectively learned by human learners. In their study using the "Human Simulation Paradigm" (HSP). They showed adults video clips of parent interacting with an infant. The original sound of the video was muted and a beep was inserted at the onset of the label when parent named an object. Adult learners were asked to watch these videos and guess the intended referent by the parent at the moment of the beep. They found that participants were not able to aggregate information and learn the correct word-referent mapping across trials. The researchers concluded that because there are potentially too many candidate referents which could be mapped on to a label, it is impossible for learners to continuously store and update the word-object co-occurrences across word learning moments and make appropriate decisions based on aggregated statistics.

However, other investigators reached different conclusions by using variants of the HSP method. Yurovsky, Smith and Yu (2013) used training videos from both the observers' view (captured by a tripod-mounted camera) and the child's view (captured by a head-mounted camera) to study how uncertain participants were when asked to make explicit hypothesis regarding the intended referent by the parent. They found that about 50% of the naming episodes by mothers to toddlers were not ambiguous to the adults, who could accurately guess the target referent. They then investigated whether participants were able to learn artificial language labels by integrating statistics across the most ambiguous naming events, which were instances that most adults could not guess the correct target referent. Significant learning was found only from the child's perspective but not from the observer's perspective, suggesting that the kind of input children experience may facilitate statistical aggregation. In a related follow-up study done by Zhang, Yurovsky and Yu (2015), participants were presented with a mixture of ambiguous and unambiguous first person videos and were asked to make guesses about the correct referent on a trial-to-trial basis. Their results suggest that word-learning is a continuous process that learners make progress gradually by integrating previous knowledge. Being able to remember and carry over partial knowledge, despite the uncertainty of the information at a moment, could facilitate learning and partial knowledge can be especially helpful when the learning situations are ambiguous.

Several recent studies have investigated word learning from learners' own perspective by placing lightweight head-cameras and eye-trackers on children while they interact with their parents. For example, researchers have found that referential uncertainty in 1½ year olds infants' own visual field is significantly reduced at the sensory level. The clutter and distraction in child's visual field are effectively reduced when objects are close to their eyes and head as close objects are visually large and can block the view of potential distracters (Yu & Smith, 2012). In addition, when parents played with and talked about novel objects with their toddlers, the visual properties (e.g. object's image size or centeredness relative to other objects) of the target object during naming predicted children's later novel object-name learning (Pereira, Smith & Yu, 2014). These perceptual cues available in children's view may play an important role in children's internal statistical computations. These findings are quite informative considering the previous assumption that cluttered everyday environment can cause a high degree of referential uncertainty. There is a need to take the learners' view into account and to study the visual input directly perceived by the learners, because ultimately the statistical information that makes contact with children's learning system matters the most. Even though low referential ambiguity facilitates word learning and parents do create relatively clean and unambiguous naming moments (Frank, Tenebaum & Fernald, 2013), naturalistic learning situations vary in their quality with some being more ambiguous than others, and some may facilitate learning and some may not.

In addition, reserachers have also started to investigate whether visual attention plays a role in infants' word learning process (e.g. Smith & Yu, 2013). Although past research on visual attention indicates that statistical word-learning is constrained by infants' developing attention system (Yu & Smith, 2011), little is known regarding how children visually attend to and select the target object among many other objects in view when parents name it during everyday interactions. Given that infants' attention is selective as they voluntarily direct their attention to certain aspects of the environment moment by moment, will they select a subset of information to attend to at each naming instance and aggregate their knowledge over time? Is it the case that when the adults utter a new word during interaction, children pay attention to a lot of objects, happenings, and properties that can possibly be a match, therefore word learning by aggregating information is so hard and impossible? In the current study, we explored the looking pattern of 12-month-old infants using naturalistic images with varying degrees of referential ambiguity in order to examine whether perceptual properties of objects in children's own view during naming moments would influence how young infants select candidate objects to build word-object mappings.

Experiment 1

The paradigm of presenting dynamic natural first-person scenes obtained from an infant's first person perspective to another age-matched infant while gathering on-line eye-tracking data has been done successfully in other studies (e.g. Aslin, 2009). Different from the original paradigm that used dynamic videos, we used still images in the current study. Our goal of Experiment 1 was to measure 12-month-old's looking behaviors during free viewing of natural word-learning scenes. Specifically, we wanted to compare and examine how different visual properties of target objects at naming moments influence the way infants allocate their attention. No spoken label was provided in this condition.

Participants. Twenty-five 12-month-old infants (10 female, ages ranged from 11.7 to 13.2 months, M_{age} = 12.28, SD_{age} = .43) participated the study. Parental consent was obtained for all participants in compliance with the IRB of Indiana University. All children received a gift for their participation.

Materials. Forty-four images were selected from a set of naming moment vignettes collected by Yurovsky et al. (2013) for their original study. This set of vignettes included play sessions from eight parent-child dyads. All vignettes were captured from children's first person view using head-mounted cameras during toy play with their parents and each vignette was 5 seconds long with the target name's onset occurred at the third second. As shown in Figure 1, we selected one frame from each 5-second naming window and systematically varied both the size (big vs. small) and location (centered vs. off-centered) of the target toy to create 4 experimental conditions (Figure 1). The four frames of the same object were selected from different naming moments.



Figure 1: Sample images from target object "ball" for four experimental conditions.

As shown in table 1: 1) there were many objects in view (ranging from 10 to 15) to which infants could direct their attention, which suggests that overall there was a high degree of ambiguity and uncertainty in all 4 experimental conditions; 2) there were also distinct differences in visual complexity and uncertainty among the four conditions. The target objects seem to be more visually dominant and salient in the big/centered condition and therefore more likely to attract infants' attention while the targets in small/off-centered

Table 1: Visual property details averaged across all 11 images in each condition.

	Mean number of objects in view	Mean proportional size of target object	Mean distance of target from center (max=1)
Big/Centered	10	22.30%	0.008
Big/Off-centered	11	16.30%	0.124
Small/Centered	14	4.70%	0.013
Small/Off-centered	15	5.20%	0.305

condition were embedded in a set of objects in view, therefore less noticeable. These learning scenes with varying degrees of uncertainty allow us to examine how infants direct their attention in those different contexts.

Apparatus. The learners' eye gaze was measured by a Tobii 1750 eye tracker. The principle of this corneal reflection tracking technique is that an infrared light source is directed at the eye and the reflection of the light on the corneal relative to the center of the pupil is measured and used to estimate where the gaze is fixated. The eye-tracking system recorded gaze data at 50 Hz (accuracy = 0.5°, and spatial resolution = 0.25°) as a learner watched an integrated 17 inch monitor with a resolution of 1280 × 1024 pixels. E-prime software was used to present the stimuli and to automate the recording of eye location with the eye tracker software.

Procedure. Infants were seated on their caregivers' laps approximately 60cm from the monitor in a quiet room. Parents were instructed to keep their child seated, facing forward and refrain from talking to them or direct their attention. We also told parents to either look down or close their eyes throughout the entire procedure so as to not to influence their infant's behavior.

The point of gaze was calibrated with a toy animation that appeared randomly at five locations (four corners and center) across the screen, one at a time. After successful calibration, the first trial began with the centered presentation of an animation to orient infants' attention to the screen. As soon as infants looked at the center, pre-selected first person view images would be presented full-screen. In total, 44 images (11 toys, each has 4 conditions) were displayed for 7 seconds each. The temporal order of images was pseudorandomized so that images of the same object and images of the same condition do not appear consecutively. The first attention grabbing slide was interspersed every 4 trials to maintain child's attention. While infants were attending to the images, they also heard soft music played in the background. The entire testing session was about 6 minutes long.

Results and Discussion. Because perfect tracking in a continuous mode is not possible due to technical limitation of the eye-tracker, involuntary head movement or loss of attention, we only included trials with more than 50% of gaze data points. Included trials have an average of 80% of gaze data points. For data analysis, we fit linear mixed effect model to the data by using size or location as the fixed factor, and subject and item as random factors. In Experiment 1, we focused on analyzing gaze data to 1) quantify the degree of uncertainty; and 2) to measure how much time infants attend to the target objects.

Quantifying Degree of Uncertainty. To investigate whether the number of objects attended by infants differed when target size or location changes, we first measured the total number of objects attended and found that even given that there were more than 10 objects in view, and also given plenty of viewing time (7 seconds per image) to attend to many objects, infants only selectively attended 3-6 objects per trial ($M_{big/centered}$=3.74; $M_{big/off\text{-}centered}$=5.24; $M_{small/centered}$= 5.11; $M_{small/off\text{-}centered}$=5.94). As shown in Figure 2A, we did not find a significant main effect for size or location (size: β=1.13, p=.05; location: β=.75, p=.21).

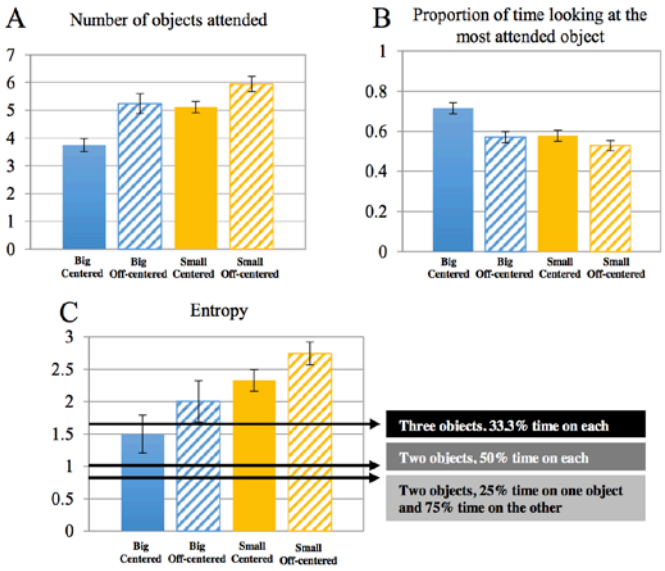


Figure 2: A. Mean number of objects attended; B. Mean proportion of time infants look at the most attended object; C. Mean entropy (averaged across trials in each condition).

Knowing that the subset of objects infants attended to was quite small, we further examined how infants allocated their attention among the subset of objects they chose. Do infants attend to those objects equally frequently or do they only primarily attend to one or two objects? To answer this question, we measured infants' proportion of time looking at the most attended object and found that across all four conditions, as shown in Figure 2B, infants spent more than 50% of time looking at one selected object ($M_{big/centered}$=.71; $M_{big/off\text{-}centered}$=.57; $M_{small/centered}$= .58; $M_{small/off\text{-}centered}$=.53). By fitting lmer models, we found that if the target size was big, infants spent more time looking at their most attended object (β = -.13, p<.01). However, location does not have an impact (β = -.07, p=.13). This finding suggests that even infants focused on only a few objects per trial, they predominantly only look at one object at least half of the time.

Next, to further capture the uncertainty that infants faced within a trial, we calculated entropy based on their looking times. In information theory, Entropy can be used to describe the uncertainty given a distribution. In the present case, given

n objects in view, we calculated $I = \sum_{i=1}^{n} P_i log \frac{1}{P_i}$ where p_i is the proportion of time looking at object i. This Entropy measure captures the dynamics of attention as it takes into account not only the number of objects attended but also the looking duration on each object. For example, if one looked at two objects equally, entropy equals to 1. If one looked at two objects, but one look is much longer (75% of the time) than the other (25% of the time), then entropy value would get lower and equals to 0.81 as it was a less uncertain situation compare with looking at two objects equally. If one looked at three object equally, then entropy gets higher and equals to 1.56 (Figure 2C). Thus, both more looks and a more even distribution of looks will cause the increase of entropy with high uncertainty while fewer and uneven looks will cause the decrease of entropy with low uncertainty. As shown in Figure 2C, entropy measures in all four conditions were relatively low, suggesting low uncertainty based on infants' looking behavior ($M_{big/centered}$=1.50; $M_{big/off\text{-}centered}$=2.00; $M_{small/centered}$= 2.33; $M_{small/off\text{-}centered}$=2.75). We then assessed whether size or location of target objects influenced how uncertain learners were based on entropy value. We found that infants tended to be more uncertain (higher entropy) when target size was small (β=.42, p<.05), but location was not a significant factor (β=.26, p =.15).

These results are quite informative as they support the idea that the visual dynamics of children' visual field might not be as noisy as people previously believed if we consider statistical learning from an embodied view (Yu & Smith, 2012). The present results also show that children's selective attention may simplify the learning problem even more because they only look at a subset of objects in their visual field and spend most time attending to only one of the selected objects.

Target Look. Because all frames are taken from natural naming moments, there is a correct named target for each scene even though participants were not aware of which object was being named. We next explored whether learners attended to the correct target or not without labels and we used two different ways to quantify this measure: 1) proportion of time infants look at the correct target in a given trial; 2) if we treat the object that was attended the most by infants as the one selected by them as the target, then how likely the object they select is the correct target.

By examining whether size or location of target objects influences how long infants look at the target object ($M_{big/centered}$=.34; $M_{big/off\text{-}centered}$=.15; $M_{small/centered}$=.24; $M_{small/off\text{-}centered}$=.11, Figure 3A), we observed that infants looked at the target object significantly longer when it is big in view (β= -.90, p<.001) and when it is centered in view (β=-1.01, p<.01). These results suggest that if the target's size is big or if it is centered relative to other objects in the visual field, infants are more likely to pay attention to that object and treat it as a potential referent if naming occurs.

Because infants were not aware that there was a potential target object in each scene, we were interested to see whether their most attended object during free viewing was likely to be the target, if so, whether the visual properties of the target toy influence their accuracy. Our results indicate that only location (β=-.32, p<.001) but not size (β=-.19, p=0.05) was a significant predictor ($M_{big/centered}$=.66; $M_{big/off\text{-}centered}$=.29; $M_{small/centered}$=.56; $M_{small/off\text{-}centered}$=.22, Figure 3B). This analysis provides evidence that when the target object is off-centered, infants are less likely to attend to it and treat it as a potential target even it is still big in view, which suggests that infants may have a center bias when free viewing natural scenes.

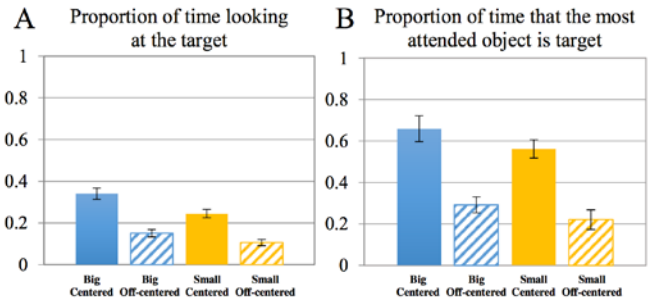


Figure 3: A. Mean proportion of time looking at the target object; B. Mean proportion of time that the most attended object is target (averaged across trials in each condition).

By analyzing free-viewing gaze data without naming in Experiment 1, we found that visual properties, such as size and location of objects in infants' own view can influence the way learners visually attend to those objects and the number of objects they are able to attend to is quite limited due to selective attention.

Experiment 2

Because we are interested in examining referential uncertainty during naming moments, in the second study, we investigated whether hearing a label during free-viewing would have an impact on how infants allocate their attention and whether their looking pattern changes after the label (e.g. look at more or fewer objects, stay longer or shorter on previously attended objects).

Participants. Twenty-three infants (11 females) between 11.4 and 12.6 months of age (M_{age} = 12.2, SD_{age} = .31) were recruited from the same population as in Experiment 1, none of these children participated in the previous experiment.

Materials. The same 44 images used in Experiment 1 were used in Experiment 2. A female native English speaker recorded the 44 labeling sentences that were infant directed. Toys' English labels were used. As shown in Figure 4, all labeling utterances were about 1 second long, with the onset of the utterance occurred at exactly the fourth second of each 7-second trial, so there were 3 seconds of silence both before and after the labeling sentences. To keep infants attentive, the same object was labeled using different sentence structures in different conditions, such as "Look at the __!" "There is a __!" See the__!" "It's a__!" and same sentence structure does

not occur consecutively. Same background music with lower volume was used.

Procedure. The procedure was the same as Experiment 1.

Results and Discussion. Mean percentage of gaze points contained across all usable trials is 83%. As shown in Figure 4, we are mainly interested in two types of comparisons: 1) compare looking behaviors happened in the last 3 seconds of the silence condition and the last 3 seconds of the label condition. This comparison controlled for the amount of visual experience infants received, the only difference between the two conditions was whether or not a label was presented; 2) calculate looking pattern changes between the first and the last 3 seconds of viewing for each condition, then compare the changes between silence and label conditions. In Experiment 2, we emphasized on analyzing gaze data to examine whether label would influence: 1) the number of objects infants select to attend; 2) the proportion of time infants attend to the correct target. For all subsequent analyses, we fit lmer models to the data and used label as the fixed factor and subject and item as random factors.

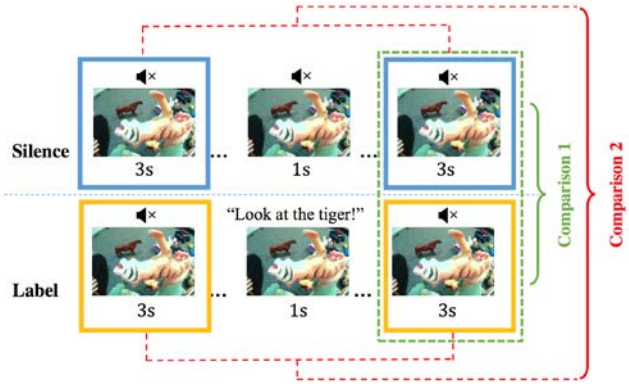


Figure 4: Types of comparisons implemented in Exp 2.

To compare looking behaviors observed in the last 3 seconds of the silence condition and the last 3 seconds of the labeled condition, we did not find a significant main effect of label on the number of objects attended (β=-.13, p=.46, Figure 5A), suggesting that label doses not influence how many objects infants choose to pay attention to. By comparing the number of objects attended before and after the label (first vs. last 3 seconds of the label condition), we found that the average number of new objects (the ones they did not attend to before the label) they chose to attend after the label was 1.33. The average number of new objects attended in the silence condition (first vs last 3 seconds of the silence) is 1.32, which is not significantly different from the label condition (β=.03, p=.77, Figure 5B). Our data suggest that infants do not change their looking patterns by selecting fewer or more objects to attend to because of the label. This is probably because if infants do not already know the referent's name, even with a label, there is still no clear indicator of which object might be the correct target. Although the information selected within a learning moment can be quite narrow that only a few objects are first attended and then stored in the memory, infant might still try to maintain more flexible visual attention and sample relatively broad co-occurrence data when the additional cues provided (e.g. label) could not help them narrow down the information selected further at the moment.

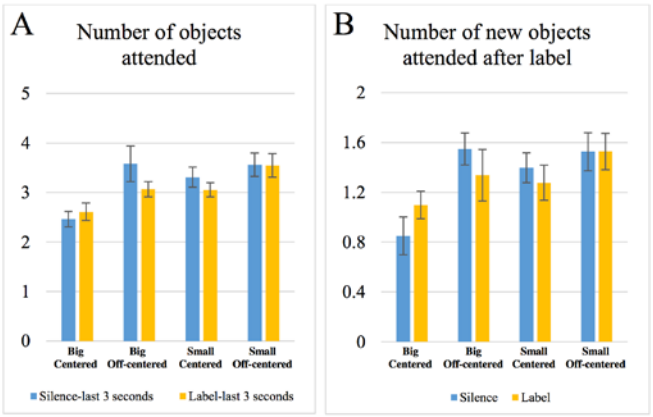


Figure 5: A. Mean number of object attended; B. Mean number of new objects attended after label in 4 conditions.

Because the labeling utterances are referring to the correct target in view, we further examined which object infants chose to pay attention to the longest and whether that object was the correct target. As shown in Figure 6, label does not influence the proportion of time infants look at the most attended object (β=.02, p=.21) nor the proportion of time that their most attended object is target (β=-.03, p=.32), suggesting that infants did not change their looking patterns dramatically after hearing the label.

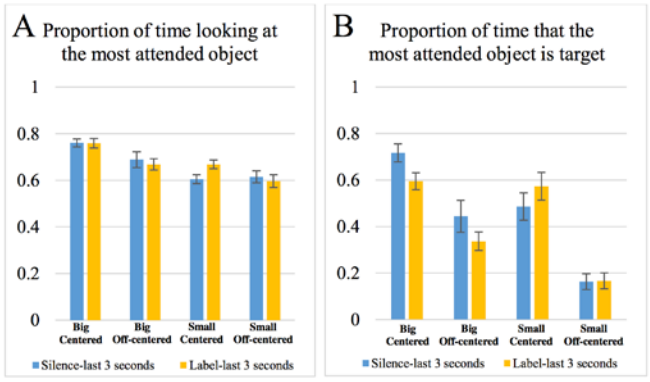


Figure 6: A. Mean proportion of time looking at the target object; B. Mean proportion of time that the most attended object is target in 4 experimental conditions.

The results in Experiment 2 are consistent with previous studies on parents' object labels in free play (e.g. Tomasello & Todd, 1983). They found that parents who use more follow-in labeling, which is the case the child is already attending to the object before labeling, have children with larger vocabulary, suggesting that just following and labeling what they have already attended to (instead of redirecting child's attention) would be quite effective because children would not switch their attention after hearing a label.

General Discussion

Our results show that perceptual properties of objects in infants' own view during naming moments dramatically influence how they select candidate objects to be considered to build word-object mappings. Experiment 1 results demonstrate that data available to statistical learners are not the data in the real world, but only a small subset of that data that is made into the learners' perceptual system at each learning moment. Such information is filtered through not only the dynamics of first person views, but also the learner's own developing attention system because it is not possible for infants to attend to everything in their own view. Thus, to address the question of whether natural learning moments are too complex for statistical learners to keep track of lots of information over time, we provide evidence to show that what the learners attend to at naming moments is not a large number of objects, but rather they attend to a small sample of available information in the world. This filtering process significantly simplifies the amount of information available for learners to carry over from one moment to the next, and to further process and integrate statistical evidence in their cognitive systems.

The quantitative results derived from gaze data can advance our understanding of the referential uncertainty problem encountered in real-life situations. At the same time, they are also in line with the previous results found using the cross-situational learning paradigm (Smith & Yu, 2008). The way infants learn word labels from real life learning moments might be similar to the way they learn words in cross-situational learning (CSL) tasks as in both cases they allocate their attention to only a few objects in view at a moment. Given that infants are able to learn the correct object-label mappings by aggregating information across trials in CSL tasks, it would be interesting to see whether they are also able to learn correct object names by collecting and accumulating information selected from first person scenes that resemble real-world learning situations. In addition, many adult studies using various paradigms (e.g. Yu & Smith, 2007; Zhang, Yurovsky & Yu, 2015) have shown that word-referent learning is a continuous statistical learning process and individual's ability to remember and carry over knowledge from past learning instances facilitates subsequent learning. One possible future direction along this line would be to design a word-learning experiment using first person view naming instances. By measuring learners' eye movement during training and comparing that with their learning outcome may allow us to understand real-time learning mechanisms, such as how statistical learners aggregate information moment by moment and whether the information learners select to attend to during training would link to what they learn at the end.

Despite the fact the environment young learners encounter is very complex and noisy, they are able to use selective attention to filter and clean up the inputs before processing them in their cognitive system. It is important to examine the underlying learning mechanisms by measuring and analyzing statistical information that is selected by, further stored and retained in the sensory, attentional, and memory processes as it is through the interactions of all these cognitive components in the learning system that young learners acquire the knowledge of solving word-learning problems and build their vocabularies.

Acknowledgments

This research was supported by NIH R01 HD074601. Special thanks to Lillian Hogan for data collection.

References

Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry and vision science: official publication of the American Academy of Optometry*, *86*, 561.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*, 1-24.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 9014-9019.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, *21*, 178-185.

Quine, W. V. (1960). Word and Object. MIT Press. *Cambridge, MA*.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558-1568.

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development*, *9*, 25-49.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*, 3617-3632.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, 197-211.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414-420.

Yu, C., & Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. *Developmental Science*, *14*, 165-180.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*, 244-262.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959-966.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental psychology*, *45*, 1611.

Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive psychology*, *43*, 217-242.

Zhang, Y, Yurovsky, D., & Yu, C. (2015). Statistical Word Learning is a Continuous Process: Evidence from the Human Simulation Paradigm. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

The Naïve Utility Calculus unifies spatial and statistical routes to preference

Julian Jara-Ettinger*, Felix Sun*, Laura Schulz, & Joshua B. Tenenbaum
Department of Brain & Cognitive Sciences, MIT. Cambridge, MA. 02139
* These authors made an equal contribution.

Abstract

Humans can seamlessly infer what other people like, based on what they do. Broadly, two types of accounts have been proposed to explain different aspects of this ability. A first account focuses on inferences from spatial information: agents choose and move towards things they like. A second account focuses on inferences from statistical information: uncommon choices reveal preferences more clearly compared to common choices. Here we argue that these two kinds of inferences can be explained by the assumption that agents maximize utilities. We test this idea in a task where adult participants infer an agent's preferences using a combination of spatial and statistical information. We show that our model predicts human answers with higher accuracy than a set of plausible alternative models.

Keywords: Computational modeling; Naïve Utility Calculus; Theory of mind; Social cognition.

Introduction

As humans, we understand that other people have minds, and we can infer what they know and what they want by watching their behavior. Imagine, for instance, that a man walks towards a cookie jar, opens it, peeks in, and then closes it again. Although we cannot see the inside of the man's mind or of the cookie jar, we nevertheless suspect that the man likes cookies, that he planned to eat a cookie, that he believed there were cookies in the cookie jar, and that he was wrong: the cookie jar was empty.

Our ability to infer other people's preferences, in the service of interpreting their actions and predicting their future behavior, is at the heart of this ability. A large body of work suggests that preference inferences rely on spatial information. When we watch an agent navigate, a first focus is on the path's *end state*: Agents navigate to complete goals that fulfill their desires (Woodward, 1998). A second focus is on the path's *directedness*: We expect agents to navigate efficiently, and we use this expectation to attribute goals (Gergely & Csibra, 2003). Thus, if an agent does not take the shortest path towards a goal this implies there is a constraint in the way (Csibra, Biró, Koós, & Gergely, 2003), a subgoal that the agent completed within the path (Baker, Saxe, & Tenenbaum, 2009), or that the actions themselves are the goal (Schachner & Carey, 2013).

When we infer preferences, however, we not only rely on what agents choose; we also take into account what they don't choose. Suppose that an agent can pick a fruit from a bag filled with a hundred apples and one orange. If the agent takes an apple, she doesn't necessarily like them better than oranges. But if she takes the only orange, then she probably likes them better than apples. Intuitively, the second situation reveals a stronger preference, even though the agent could have chosen either fruit in both cases. In other words, the strength of the preference inference depends on the statistical information of the possible choices.

The ability to infer preferences using spatial and statistical information are both at work from early in life. Infants as young as three months old expect agents to navigate efficiently to some extent (Skerry, Carey, & Spelke, 2013) and show a robust expectation by their first birthday (Gergely & Csibra, 2003). Similarly, the ability to draw inferences from statistical information has its roots in infancy and it plays a role in how we learn what other people like (Kushnir, Xu, & Wellman, 2010; Wellman, Kushnir, Xu, & Brink, 2016), how we learn about the world (Gweon, Tenenbaum, & Schulz, 2010), and even how we learn the meaning of new words (Xu & Tenenbaum, 2007).

Together, these two lines of evidence suggest a dual system for inferring preferences: one that relies on spatial information, and one that relies on statistical information. But real-world situations do not break down so cleanly. Agents usually combine both spatial and statistical distributions of potential rewards in their environment, and so should our judgments about their preferences from observing their actions.

Here we propose that, rather than being supported two systems of knowledge, preference inferences from spatial and statistical information are derived from a single intuitive theory of agents: the naïve utility calculus (Jara-Ettinger, Gweon, Tenenbaum, & Schulz; 2015; Jara-Ettinger, Tenenbaum, & Schulz, 2015). Critically, our goal here is not to compare the naïve utility calculus with formal theories of decision-making, but with other theories of intuitive decision-making. Here we show how the naïve utility calculus (NUC) supports inferences from spatial and statistical information. We test our proposal by implementing and comparing a spatial inference model, a statistical inference model, and a NUC model against adult performance on a preference-inference task. We end by discussing the implications of our findings on understanding the development of commonsense psychology.

The Naïve Utility Calculus

A growing set of studies suggests that humans reason about agents in terms of utility maximization (Jara-Ettinger et al., 20215; Jern et al, 2011; Johnson & Rips, 2015; Lucas et al, 2014). Specifically, humans have an intuitive theory of how utilities are comprised of costs and rewards, and how, together, they guide what others do. According to this *Naïve Utility Calculus*, agents act by estimating the costs and rewards associated with each possible plan, and by selecting the plan with the highest utility (the difference between rewards and costs). That is, when people watch an agent,

they assume that her behavior yielded high utilities, and they use this assumption to infer the agent's competence (her costs) and her motivation (her rewards). To illustrate inferences through the NUC, consider an agent who chooses an apple over an orange. This implies that the utility for the apple, *U(a)*, is higher or equal than the utility for the orange, *U(o)*. By decomposing each utility into its costs (*C(a)* and *C(o)*, respectively) and rewards (*R(a)* and *R(o)*, respectively), the agent's choice implies that $R(a)\text{-}C(a) \geq R(o)\text{-}C(o)$. If both fruits were equally easy to get, then $C(a)=C(o)$ and, therefore, $R(a) \geq R(o)$. That is, when two options are matched for costs, agents choose what they like best. Suppose instead that the apple, the agent's choice, was more costly to get than the orange. Because $R(a)\text{-}C(a) \geq R(o)\text{-}C(o)$ and $C(a)>C(o)$, then $R(a)>R(o)$. That is, agents unambiguously reveal their preferences when they choose the more costly option. Last, if the apple was easier to get, then $R(a)\text{-}C(a) \geq R(o)\text{-}C(o)$, and $C(a)<C(o)$. Under these circumstances, *U(a)* (*R(a)-C(a)*) may be higher than *U(o)* (*R(o)-C(o)*) because the apple's reward (*R(a)*) was high, or because the orange's cost (*C(o)*) was high. Thus, when agents choose low cost options their preferences are not revealed.

Although its developmental origins are unclear, the NUC is at work from early childhood, supporting fundamental inferences by age five and with some aspects already at work by age two (Jara-Ettinger, et al. 2015).

Inferences from spatial information

The NUC explains why humans are sensitive to spatial information. Suppose an agent takes a sequence of actions to complete a goal. If the agent maximized utilities, then two things must be true. First, the reward must outweigh the costs. Otherwise, the plan's utility would be negative and the agent could obtain a final higher utility by not acting at all. Second, the agent must be minimizing costs: the smaller the costs the agent incurs, the higher the utility she obtains. In spatial contexts, cost minimization reduces to efficient navigation. Thus, expecting agents to maximize utilities implies that a path's *directedness* and *end state* can help reveal preferences. If, however, humans do so through a naïve utility calculus, then, as the example above reveals (see apple-orange example), humans should also be sensitive to a third feature of spatial navigation: its *cost*.

Inferences from statistical information

Inferences from statistical information ultimately rely on the assumption that rare choices reveal stronger preferences. Although intuitive, the causes underlying this assumption are unclear. The NUC, however, naturally produces this expectation. Suppose that an agent can take any object from a box. If she doesn't have a preference, then taking whichever object is easiest to get maximizes her utilities (if all objects have the same reward, the option with the lowest cost yields the highest utility). In contrast, if the agent prefers one type of object to the others, then she will have to incur a higher cost in terms of time, effort, attention, and distance to locate the object of the desired category and to retrieve it. The less common an object's category is, the higher the cost the agent must incur to locate it and obtain it. Thus, retrieving rare objects suggests that the agent incurred a higher cost, and, if agents maximize utilities, this cost is only warranted if the reward associated with the rare object is higher than the reward associated with more common objects.

Computational modeling

If humans infer preferences using their NUC, then a formal implementation should quantitatively predict adult preference judgments. Alternatively, if humans infer preferences through simpler ways, then simpler models should predict human inferences with equal or better accuracy. To test if participants integrate spatial and statistical information through the assumption of utility maximization we ran a preference-inference task with adult participants and we compared their performance to five computational models: our full Naïve utility calculus model and two NUC lesioned models, as well as two alternative models. These models are based on the proposals of how infants infer preferences from spatial information (spatial model) and from statistical information (statistical model). Next, to test if participants infer these preferences in a Bayesian way, we compared participants' self-reported confidence judgments with estimates from each model.

Alternative models

Spatial model The spatial model formalizes the proposal that goals directly reveal preferences. As, such, it uses a limited source of spatial information: the end state. This model assumes that the distribution of choices an agent makes matches her underlying preferences. For instance, if an agent collects two red objects and one blue object, then the reward for collecting a red object is $Rred = 2R/3$ and the reward for collecting a blue object is $Rblue = R/3$, where R is a constant set to 1 (changing the value of R does not change our results as model comparison was done by z-scoring model predictions. See Results).

Statistical model The statistical model is based on proposals for how people infer preferences by relying on statistical information (Gweon, et al, 2010; Xu et al, 2007). These models were formulated in simpler domains than the one we test in our experiment so we extended them to fit our experimental design. Our model assumes that, before taking any actions, the agent goes through a decision making process to find objects that will give her high rewards. Specifically, the statistical model assumes that the agent considers one object at a time (at random) and decides whether to collect it or not based on its reward. That is, when the agent considers taking an object from category *k*, she selects it with a probability proportional to its reward. These assumptions imply that more common object categories are more likely to be considered and that more rewarding kinds of objects are more likely to be collected, once the agent considers them. As such, selecting an object from a rare category suggests that the agent prefers it to more common objects that the agent likely considered collecting first. In this model, the observer assumes that the agent considers each object with uniform probability (if

there are *n* objects, the agent considers each object with probability 1/n).

Given the theory of how an agent chooses what to collect, we use Bayesian inference to recover the agent's preferences given her choices. Specifically, because in our experiment we use two types of objects (see Stimuli), we use Bayes' rule to estimate the relative magnitude of one reward type over the other (with 0 indicating that the first category contains all the rewards, 0.5 indicating that both categories are equally rewarding, and 1 indicating that the second category contains all the rewards), using a uniform prior.

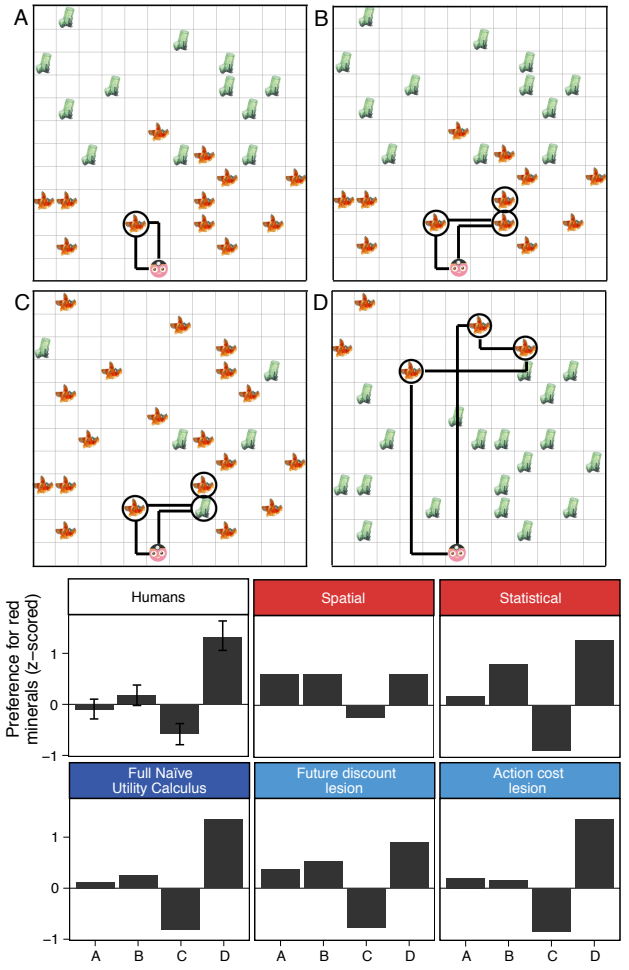


Figure 1. Stimuli examples along with participant judgments and model predictions. Because judgments are z-scored, positive and negative values are relative to the average preference inference and do not correspond to preferring red or green minerals, respectively. The negative-valued prediction in this plot indicates a weaker preference for red minerals compared to the average inference.

Naïve Utility Calculus models

The last three models are implementations of the naïve utility calculus (NUC), but they integrate costs in different ways, enabling us to understand how humans may reason about costs, rewards, and utility maximization. All models are formulated as generative models that predict agent choices given their preferences, and the inference from choices to preferences is done through Bayes' rule with a uniform prior over the possible distribution of rewards over the object categories.

Full Naïve Utility Calculus The full NUC model assumes that agents maximize utilities. Costs are function of the number of actions the agent takes (set to be a constant cost per action = 0.01; our conclusions are robust to parameter changes) and rewards are exponentially discounted over time. Intuitively, the future discount corresponds to the assumption that the longer an agent takes to reach a reward, the less likely the reward will still be there. Thus, this model relies on spatial information in three ways: first, it expects agents to navigate efficiently because smaller sequences of actions incur fewer costs (minimizing costs), and because collecting objects faster results in higher rewards (maximizing exponentially discounted rewards); second it assumes that the agent's goals have sources of rewards; and last, and in contrast to the alternative models, it assumes that longer distances reveal stronger preferences. More formally, the cost of actions and rewards in objects are integrated into a utility function ($U=R-C$) and the utility-maximizing actions are derived through a Markov Decision Process. Further details about the computational implementations of the naïve utility calculus can be found in Jara-Ettinger et al (2015).

Future-discount lesion The future-discount lesion is identical to the NUC model but rewards aren't discounted over time. Thus, this model integrates statistical information in a full manner, and spatial information in a simplified manner. The model expects agents to navigate efficiently only because lower costs lead to higher utilities, but not because longer distances increases the chance of losing the target reward.

Action cost lesion Conversely, the action cost lesion model is identical to the NUC but it ignores action costs. Nevertheless, the model assumes that the agent's rewards are discounted over time. This model therefore integrates spatial information through the expectation that agents act efficiently because the longer it takes them to reach a reward, the less likely it will still be there when they arrive.

Experiment

To test our models, we designed a simple task where participants watched a miner collect minerals in mines with variable distributions of minerals.

Stimuli

Figure 1 shows examples of the stimuli. Each stimulus consisted of an animated display of an agent (the miner) entering a mine (a 12x12 grid world) and collecting green and/or red minerals. Each map contained 24 minerals in the same locations (which were chosen at random and kept constant across stimuli), but the proportion and the distribution of these minerals varied. The proportion varied according to three levels: more green than red (20 green and 4 red), more red than green (4 green and 20 red), or an equal number of each (12 of each). The distributions of these minerals varied according to three levels: red minerals closer, green minerals closer, or all minerals intermixed.

This generated a total of nine different maps. By varying the proportion of the objects, we can test how statistical information influences preference inferences; by varying the location of the objects we can test how spatial information influences preference inferences.

The miner's paths were obtained by computing the shortest path an agent would need to take to collect all minerals of one kind, or to collect the closest minerals (which could be a combination of red and green minerals). These paths were generated in accordance to three conditions. In the first condition, the miner collected one mineral and exited the mine. In the second condition, the miner collected three minerals in a single trip and then exited the mine. And in the last condition the miner collected three minerals, but had to return to the mine's exit after collecting each object. Thus, the first and second conditions test how the amount of data an observer receives influences observers' inferences, and the second and third conditions together test how the costs of collecting the minerals influence observers' inferences. The combination of the two agent types (strong preference or no preference) with the nine maps produced a total of 18 test paths per condition.

Participants

90 U.S. residents (as determined by their IP address) were recruited and tested through Amazon's Mechanical Turk platform (Mean age = 33 years. Range = 20 - 59 years).

Procedure

Participants were randomly assigned to the one mineral condition, to the three minerals in one trip condition, or to the three minerals in three trips condition (N = 30 participants per condition). Thus, each participant only completed one-third of the trials. Participants first completed a brief tutorial that explained the task. Next, participants completed a questionnaire with three questions to ensure they understood the task. Participants who responded all questions correctly were given access to the experiment, and participants who made at least one error were redirected to the beginning of the tutorial.

In the test stage, participants saw an animated display of the miner collecting the minerals and had to respond four questions. The first two questions were multiple choice control questions asking about the proportion and distribution of the minerals. Participants who answered these questions incorrectly were asked to re-examine the stimulus. The third question asked participants to rate the miner's preference using a slider that ranged from "Red is much more valuable" (coded as a 0) to "Green is much more valuable" (coded as a 1). The last question asked participants to rate their confidence in the preference judgment using a slider that ranged from "Not at all" (coded as a 0) to "Extremely confident" (coded as a 1).

Results

Figure 2 shows the results from the experiment. As expected, the formalizations of spatial and statistical accounts matched the qualitative pattern of participant judgments: they predicted strong and weak preferences accurately. However, as Figure 2 shows, the NUC model captured human judgments with higher precision. To evaluate model performance more precisely we computed each model's correlation with average human judgments (z-scored within each participant and averaged; see Table 1.

Model	Correlation (95% CI)
Spatial	.84 (0.79,0.92)
Statistical	.81 (0.74,0.90)
Naïve Utility Calculus	**.97 (0.96,0.98)**
Future-discount lesion	.93 (0.90,0.96)
Action cost lesion	.96 (0.92,0.97)

Table 1. Model correlations with participant responses along with 95% bootstrapped confidence intervals.

Comparison with alternative models

Overall, the NUC model had the highest correlation (r=0.97) between its predictions and participant responses. To evaluate this correlation we bootstrapped the correlation difference between the NUC and the alternative models. The NUC reliably outperformed the spatial model (correlation difference=0.12; 95% CI=(0.03,0.18)) and the statistical model (correlation difference=0.16; 95% CI=(0.05,0.22)).

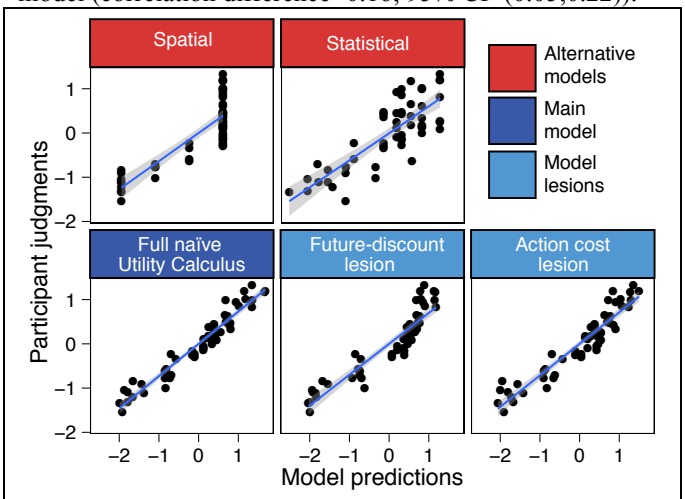


Figure 2. Experiment results. In each plot, each black dot represents a stimulus. The x-axis shows the model's prediction (z-scored) and the y-axis shows average participant judgments (z-scored within each participant and averaged).

Figure 1 shows four example trials that reveal how the NUC outperforms the alternative models. The spatial model fails to capture differences between trials A, B, and C, as it is not sensitive to the amount of evidence. The statistical model roughly captures human responses, but it attributes a stronger preference to the miner in trial B, as it neglects the spatial distribution. In contrast, the NUC models show sensitivity to the amount of data, the spatial information, and the statistical information.

Comparison with model lesions

Both model lesions had a lower correlation with participant judgments compared to the full NUC model (see Table 1). Removing the future-discount parameter led to a significant decrease in the model's correlation with human

judgments (correlation difference = .042; 95% CI = (0.004,0.072)). This suggests that participants are sensitive to an exponential discounting of the mineral rewards over the length of the miner's trajectory. Similarly, removing the cost of travelling decreased the model's correlation with human judgments (difference = .022; 95% CI = (-0.007,0.047)). However, 13% of the mass of the 95% confidence interval was on the negative region. This suggests that integrating a linear cost over the future-discount may better fit human judgments, but the results are inconclusive.

Confidence judgments

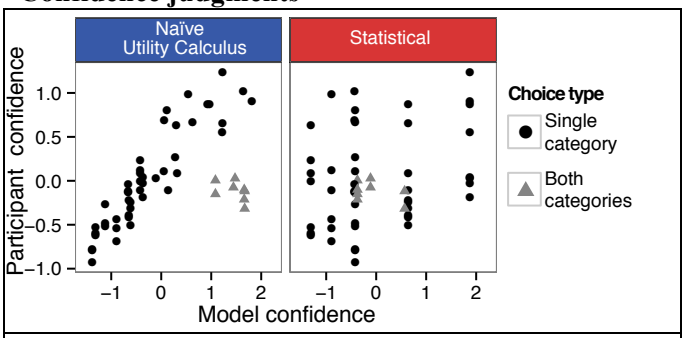


Figure 3. Confidence judgments. The models' confidence ratings were obtained by computing the standard deviation of the posterior distribution of each stimulus, multiplying them by -1 (so as to match the qualitative order in participant judgments), and then z-scoring the values. Participant confidence judgments were z-scored within participant and averaged.

Our evidence so far suggests that humans infer preferences through the assumption of utility-maximization. Nevertheless, this inference is not necessarily Bayesian. Participants may, for instance, approximate the responses from a Bayesian models through simpler heuristics. To explore this possibility, we asked participants to report confidence judgments on each trial (see Methods section) and we compared them with a rough measure of each model's uncertainty: the posterior distribution's standard deviation. If participants are inferring preferences in a probabilistic manner, then the NUC's standard deviation should correlate with participant confidence judgments. However, if participants infer preferences through some heuristics that approximate Bayesian inference, then their confidence should not necessarily be related to the one in our model. Moreover, the statistical model, being Bayesian, also produces confidence judgments (the spatial model generates a single inferred estimate with full confidence), enabling us to further test its validity.

Figure 3 shows each model's negative standard deviation along with participants' confidence judgments. Although the alternative models all captured preference inferences in a coarse way (see Figure 2), their measures of confidence did not resemble participant's confidence judgments (see Figure 3). In contrast, the NUC model and its lesions predicted with far higher accuracy participants' confidence judgments. Table 2 shows the correlations and confidence intervals. Although the NUC's correlations were reliably greater than 0, Figure 3 reveals that it failed to capture the variation in a small set of stimuli (the results were qualitatively identical for the NUC model lesions). Post-hoc inspection of these outliers revealed that they were all cases where the miner had selected a combination of red and green minerals (because of the way we generated the stimuli, the miner only took a combination of red and green minerals whenever these were the closest and the agent had no preference; see Stimuli section). Consistent with this, we found that when we decomposed the stimuli into trials where the agent collected only one type of mineral (Single category), the NUC model and its lesions showed high correlations and performed roughly as well. In contrast, in the stimuli where the agent collected various kinds of minerals (Both categories), none of the models predicted human confidence judgments (see Table 2). Nevertheless, it is important to note that this subset of stimuli consists of seven data points, making it difficult to draw conclusions from the correlations.

Model	Correlation (95% CI)	Single category correlation	Both categories correlation
Statistical	0.28 (0.04,0.56)	0.29 (0.05,0.57)	-0.43 (-1,0.01)
Naïve Utility Calculus	0.65 (0.49,0.83)	0.91 (0.88,0.95)	-0.45 (-1,-0.04)
Future-discount lesion	0.33 (0.12,0.51)	0.84 (0.79,0.89)	-0.45 (-1,-0.04)
Action cost lesion	0.68 (0.53,0.86)	0.91 (0.88,0.96)	-0.32 (-0.98,0.18)

Table 2. Correlation between the standard deviation of the model's posterior distribution and participant confidence judgments, along with 95% bootstrapped confidence intervals. The first column shows the overall correlations, and the last two columns show the correlations after splitting the stimuli into the group where the miner only collected one type of mineral (single category) and when the miner collected a combination of red and green minerals (both categories). The spatial model is not presented as it only produces a point estimate rather than a probability distribution.

Discussion

Here we reviewed evidence that, from early in life, humans can infer preferences using statistical and spatial information, and we proposed that these two types of inferences are driven by the naïve utility calculus (NUC) – our intuitive theory of how agents select their goals by estimating and maximizing utilities. We tested our proposal by implementing a formal model of the naïve utility calculus and comparing it to other accounts that rely on spatial and statistical information separately. Our results show that adults were both sensitive to the spatial and statistical information of an agent's behavior, and that this variation was best captured by the NUC model.

Critically, all accounts fit participant judgments qualitatively. Thus, implementing formal computational models was critical for generating precise predictions and assessing whether they explained variation in human judgments in a fine-grained manner. Our results show that

the NUC model significantly outperformed the alternative models at a detailed level.

In order to better understand the NUC's performance we implemented two model lesions. In one model lesion we removed the future discount parameter (future-discount lesion) and in the second model lesion we removed the cost for traveling (action cost lesion). Critically, both model lesions were still sensitive to the statistical information, and they both expected the agent to navigate efficiently. The NUC correlated with human responses better than both of the model lesions, but this difference was only reliable when comparing the NUC model with the cost sensitive lesion and not when comparing it with the action cost lesion. Our results suggest that a non-linear reward discount is critical for how humans reason about efficiency. However, once a model integrates a future-discount parameter, adding a cost of traveling only produces a modest improvement.

Although the alternative models roughly predicted human responses, a comparison of the models' posterior standard deviation (a measure of the model's uncertainty) against participant confidence judgments revealed strong discrepancies. In contrast, the NUC and its lesions predicted our participant's confidence judgments for a large set of stimuli (see Figure 3 and Table 2, columns 1 and 2). Nevertheless, all models failed to capture human confidence judgments in the trials where the miner collected a combination of red and green minerals closest to the mine's entrance (see last column of Table 2). In these situations, the NUC models were confident that the miner liked both minerals roughly as much, and that she was therefore collecting the closest ones. Participants made similar judgments, but they were less confident. One possible explanation for this discrepancy is that our model assumes that the cost for traveling is fixed and observable, whereas participants may not. Instead, participants may be uncertain about how exhausting it is to travel the mine, and this may lead to a confound in the miner's behavior: she might be taking the closest minerals because she likes all minerals just as much, or because she finds traveling deep into the mine to be very costly. A richer version of the NUC that integrates uncertainty over the costs and rewards is needed to evaluate this possibility.

Altogether, our results show that the NUC explains why and how humans rely on spatial and statistical information when inferring preferences. Empirical results show that the ability to infer preferences from spatial information and from statistical information arises in early childhood (Gweon et al., 2010; Gergely & Csibra, 2003). However, these sources of information have been studied separately, and different accounts have been proposed to explain how we draw these inferences. Our finding that inferences from statistical and spatial inferences are unified in adults raises two hypotheses about the development of this reasoning. A first possibility is that the NUC is already at work in infancy. If so, infants may use it to solve tasks involving spatial information (e.g., Gergely & Csibra, 2003), and tasks involving statistical information (e.g., Kushnir, et al, 2010). A second possibility, however, is that the NUC emerges later in life. Under this account, infants must rely on simpler expectations about agents to reason about spatial and statistical information (perhaps driven by two separate systems of understanding; a spatial one and a statistical one). If this is true, then the proposed explanations for how infants use spatial and statistical information (formalized in the spatial and the statistical models), may be correct and serve as the bedrock for a richer unifying intuitive theory: the naïve utility calculus.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216, and by the Simons Center for the Social Brain.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Csibra, G., Bıró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111-133.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *TiCS*, (7).

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *PNAS*, *107*(20).

Jara-Ettinger, J. Schulz, L. E. & Tenenbaum, J. B. (2015). The naive utility calculus: Joint inferences about the costs and rewards of actions. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14-23.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not So Innocent Toddlers' Inferences About Costs and Culpability. *Psychological science*, *26*(5), 633-640.

Jern, A., Lucas, C. G., & Kemp, C. (2011). Evaluating the inverse decision-making approach to preference learning. In *NIPS*.

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive psychology*,*77*, 42-76.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134-1140.

Lucas, Christopher G., Thomas L. Griffiths, Fei Xu, Christine Fawcett, Alison Gopnik, Tamar Kushnir, Lori Markson, and Jane Hu. "The child as econometrician: A rational model of preference understanding in children."*PloS one*.

Schachner, A., & Carey, S. (2013). Reasoning about 'irrational' actions: When intentional movements cannot be explained, the movements themselves are seen as the goal. *Cognition*.

Skerry, A. E., Carey, S. E., & Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *PNAS*, *110*(46).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. (2016). Infants use statistical sampling to understand the psychological world. *Infancy*.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Task-set Selection in Probabilistic Environments: a Model of Task-set Inference

Ian W. Eisenberg (ieisenbe@stanford.edu)
Department of Psychology, Stanford University

Russell A. Poldrack (russpold@stanford.edu)
Department of Psychology, Stanford University

Abstract

To act effectively in a complicated, uncertain world, people often rely on task-sets (TSs) that define action policies over a range of stimuli. Effectively selecting amongst TSs requires assessing their individual utility given the current world state. However, the world state is, in general, latent, stochastic, and time-varying, making TS selection a difficult inference for the agent. An open question is how observable environmental factors influence an actor's assessment of the world state and thus the selection of TSs. In this work, we designed a novel task in which probabilistic cues predict one of two TSs on a trial-by-trial basis. With this task, we investigate how people integrate multiple sources of probabilistic information in the service of TS selection. We show that when action feedback is unavailable, TS selection can be modeled as "biased Bayesian inference", such that individuals participants differentially weight immediate cues over TS priors when inferring the latent world state. Additionally, using the model's trial-by-trial posteriors over TSs, we calculate a measure of decision confidence and show that it inversely relates to reaction times. This work supports the hierarchical organization of decision-making by demonstrating that probabilistic evidence can be integrated in the service of higher-order decisions over TSs, subsequently simplifying lower-order action selection.

Keywords: task-sets; structure learning; Bayesian cognition; model-based; decision making

Introduction

Humans face the daily challenge of making decisions in an uncertain and open-ended world. In such a world, caching independent stimulus-response mappings is impractically slow and fails to capitalize on the structure inherent in natural tasks. Many actions learned at one time may generalize to new environments: our experiences dealing with petulant adults may help us placate ill-tempered children or vice versa, expertise using our own computer seamlessly translates to computer expertise in general, and so on. In general, the organized structure of the world allows agents to fruitfully group learned actions into higher-order action policies which can be retrieved to avoid redundant learning. Such action policies are often referred to as task-sets (TSs), and their use potentially eases the learning problem by providing flexible representations that can be leveraged across environments.

Much of the computational work on TSs relates to model-based decision making (Daw et al., 2012; Solway & Botvinick, 2012). In this framework, a person learns a goal-sensitive action policy based on an internal model of the world, which consists of a set of states, transition probabilities, and actions. Through exploration and feedback, the agent gradually develops an action policy which determines their behavior. While this general concept has been informative in outlining how TSs may be learned given a model of the world, model-based decision making has largely ignored how the agent's internal models are developed and selected when multiple models may apply in a particular environment (so-called "structure learning").

Some research has directly addressed the problem of structure learning, proposing that agents simultaneously infer the latent causal structure of the world while identifying the appropriate TS given that inferred structure (Gershman & Niv, 2010; Redish et al, 2007). From this perspective, structure learning is intimately tied with stimulus-response learning, leading to the compelling prediction that people will reuse TSs whenever they infer that the latent structure of the world conforms to the structure in which the TS was first learned.

Inferring the latent world state is closely related to categorization (Gershman et al. 2010), the cognitive process by which people use an organizational framework to assign discrete instances (objects, events, emotions, etc.) to groups that are functionally or perceptually equivalent on some level of abstraction (Anderson, 1991; Shafto et al. 2011). The central idea is that latent categories stochastically generate observable features conforming to some characteristics distribution. If people represent, on some level, a generative model of the environment that constitutes a hypothesis space over possible categories, then they can infer the underlying category given uncertain evidence (Fei-Fei et al., 2007; Tenenbaum et al., 2006). Moreover, they can categorize novel observations by appealing to these generative models. For decision-making, useful categorical boundaries are defined by states which call for different action policies. To capture TS selection in such a scenario, we use a task where the agent knows that multiple task-relevant states exist, such that establishing the latent world state is equivalent to establishing the best TS. If the agent can uncover the structure underlying state transitions, they can greatly simplify the task and improve their performance.

Empirical and computational support for probabilistic inference over TSs comes from work by Collins and colleagues, who have shown that people reuse TSs in an approximately optimal way based on contextual support (Collins & Koechlin, 2012; Collins & Frank, 2013). Collins & Koechlin have put forward a model where a small number of TSs are held in a working memory-like cache

where they are evaluated to assess their individual "reliabilities." TSs are selected when they prove reliable, and otherwise discarded, replaced by new TS propositions constructed combinatorially from TSs held in long term memory. In a similar vein, Frank & Badre (2012) propose a Bayesian "mixture of experts" model of TS selection (see Doya et al., 2012 for a similar idea). In this framework, multiple competing TS hypotheses govern people's behavior as they search for higher-order rules in a hierarchically structured decision making task (Badre et al. 2010).

These experiments used deterministic cues to indicate the appropriate TS, a simplification that potentially obscures the factors underlying TS inference in general. In this project, we aimed to resolve these problems and clarify the process underlying TS inference. We introduce a novel task-switching paradigm that required participants to reason over probabilistic environmental cues to select the appropriate TS on a trial-by-trial basis. With this paradigm, we anticipated that participants would use multiple sources of information when selecting TSs such that their decisions related both to contextual cues and TS transition probabilities. As these different sources contribute to behavior in subtle ways, we develop an explicit quantitative model to assess the information participants access to infer TSs. We hypothesize that while there will be substantial individual differences in how people integrate information, TS inference can be characterized by Bayesian inference with minimal free parameters reflecting individual information-processing biases.

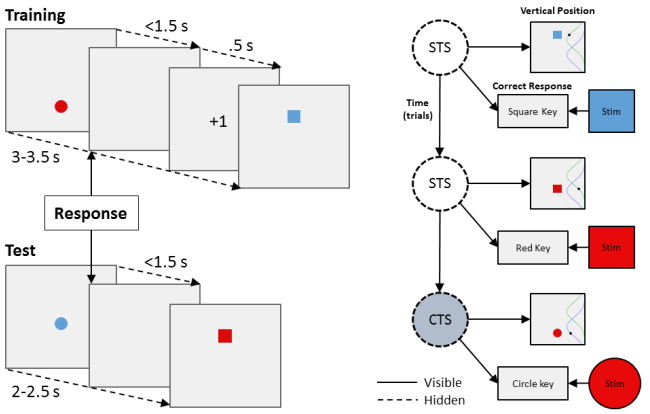


Figure 1: (*Left)* On each trial, a shape appears at one of 12 vertical positions. The participant responds with one of four keys corresponding to the two features of the stimuli (red, blue, circle, square). During training they get deterministic feedback after they respond. (*Right)* Schematic of the latent trial structure. Stimulus vertical position is drawn one of two distributions (shown in green and blue) corresponding to the current TS (STS: green, CTS: blue). The current TS has a 10% chance of changing from trial-to-trial, otherwise it remains the same. The stimuli are randomly drawn on each trial and are observable while the current TS is latent. Thus to correctly respond to the stimulus the participant must infer the current TS based on the stimulus position and task history.

Method

Task Description

49 participants completed the Probabilistic Context Task, a task-switching experiment (Figure 1) composed of two phases: training (832 trials: 45 min) and testing (800 trials: 30 min) On each trial, participants were required to select one of four keys in response to two-dimensional stimuli varying in color (red or blue) and shape (circle or square). Each key was mapped to one of these feature (e.g. key 1 for blue stimuli, key 2 for circles), which were randomized across participants. Participants had 1.5 s to respond. The correct response was determined by a latent TS that established the relevant feature, which changed from trial-to-trial. There were two TSs: the shape TS (STS) and the color TS (CTS). Correct responses conformed to both the stimulus and the TS (pressing the red key for a red circle while the CTS was operating). Correct responses earned a point which was presented for .5 s during the training phase followed by a variable intertrial interval. During the testing phase participants received no feedback. Overall, each training trial lasted 3-3.5 s, and each test trial lasted 2-2.5 s.

While there were no deterministic cues indicating the current TS, the task was designed to allow inference of the trial-by-trial TSs using probabilistic information. TSs switched probabilistically on each trial such that $P(TS_{t-1} = TS_t)$, the probability of the TS remaining the same from trial-to trial, was 90%, referred to for the remainder of the paper as the recursion probability. Additionally, on each trial the stimulus's vertical position on the screen was drawn from a truncated Gaussian distribution (limits 1 and -1, corresponding to the top and bottom of the screen, respectively) parameterized by the current TS. Stimulus position Gaussians had the same standard deviation (σ=.37), but had different means (μ = .3 or -.3) depending on the TS. These Gaussians were discretized into 12 bins spanning the screen. The TS that was primarily associated with the top of the screen was counterbalanced across participants. For simplicity, for the remainder of the paper we will assume the STS was primarily associated with the top of the screen.

Before training participants were explicitly told about both color and shape TSs and were given an opportunity to practice using a separate set of practice stimuli and key mappings. When training started, participants knew that one key would correspond to each stimulus feature (red, blue, circle square), and only one TS operated on each trial, but did not know what the response mappings were or how the TSs switched. They were told that their goal was to learn on which trials they should respond based on color or shape. They were also told that they should use the feedback to learn during training, but not to rely on it, as it would be removed during the test. Participants were encouraged to respond as quickly and accurately as possibly. Participants also knew that their performance during training and test determined their bonus payment, which could range from

$0-$5. A post-task questionnaire probed participant's explicit understanding of the task, including their estimates of TS transition probabilities.

In summary, during the training phase, participants received deterministic feedback which they could use to learn the mapping between stimuli features and response keys as well as the determinants underlying TS switches. Because feedback was omitted during the subsequent test phase, participants had to respond based on their understanding of the TSs' relationship with the probabilistic cues in the environment (stimulus position and the previous operating TS), or based on irrelevant factors fabricated by the participant (e.g. deterministic switches every 5 trials, switch after a red square). Above chance performance during the test phase would indicate that the participant had internalized some true aspects of the task structure.

Behavioral Analysis

Participants responses were assigned to either the CTS or STS based on whether one of the two colors keys or shapes keys was pressed, respectively. Because we were interested in TS selection, it was necessary that participants knew which keys corresponded to which features by the beginning of the test phase. To ensure this, we coded each response as either conforming or not conforming to one of the two dimensions of the stimulus and excluded participants whose average stimulus conformance fell below 75% during the test phase. This exclusion criterion ensured that all analyzed participants knew the two appropriate responses for each stimulus (e.g. either the red or circle keys for a red circle)

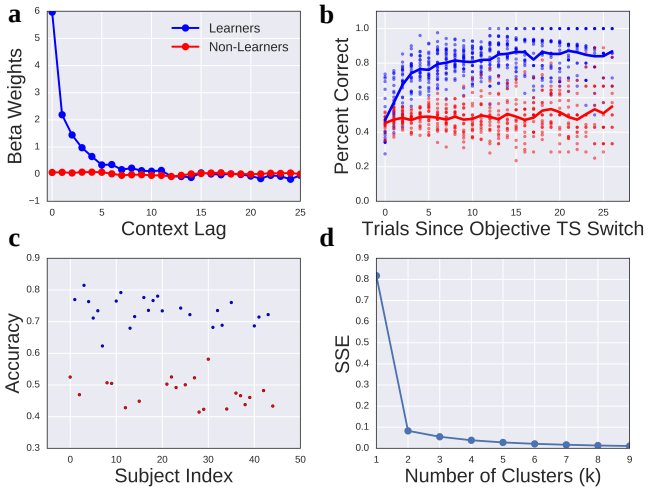


Figure 2: Summary of learners (blue) and non-learners (red). *(a)* Output of regression predicting participant choice by current context and context history. *(b)* Participant accuracy as a function of trials since objective (latent) TS switch. Each point is an individual participant's accuracy at that delay. *(c)* Clustering of participants using k-means on participants accuracy with two centroids initialized at .49 and .51. *(d)* Sum of squared errors for different k values with random initialization averaged over 1000 iterations.

and only had to determine the operating TS to successfully respond. Of our 49 participants, 4 were excluded based on this criterion. All remaining participants conformed to the stimulus >90% of the time. For the remaining participants, we collapsed their responses from the four choices to the two TSs resulting in a binary choice vector across trials.

Our analysis was principally concerned with how people weighed the probabilistic information potentially relevant to TS selection. Therefore it was necessary that behavior related in some way to task variables. We used k-means clustering to divide the participants based on overall accuracy, resulting in a clear separation between two groups of participants: 24 "learners" and 21 "non-learners" (Figure 2c). While we presume that there is structure in all participant's behavior, this paper is solely exploring the correspondence between task structure and behavior, rather than a complete evaluation of participant behavior. Modeling work was therefore restricted to the learners.

Prior to modeling we fit mixed-effects logistic regressions to both groups to assess the impact of context, context history and prior choice on TS selection.

Computational Modeling Optimal TS inference during test can be formalized as Bayesian inference over probabilistic cues and task history. The optimal prior over TSs on $trial_t$ is proportional to the posterior over TSs after $trial_{t-1}$. Specifically, the prior is the product of the transition probability matrix between TSs and the posterior vector over TSs. In other words, if the person was absolutely confident in the TS on $trial_{t-1}$, then the prior conforms to the transition probabilities associated with that TS; if the person was completely unsure on $trial_{t-1}$, the prior over TSs is uniform. This prior information is then combined with the stimulus position's likelihood under each TS's positional distribution to arrive at a posterior over TSs on each trial.

Our main hypothesis is that most participants will integrate both transition probabilities and positional distributions to select TSs, but individuals may be biased in their weighting such that their choices unequally favor the probabilistic cue. This is a soft form of base-rate neglect, where the prior is down-weighted in favor of the likelihood. Our model (the *bias-2* observer, below*)* instantiates this idea by fitting two variables, r_1 and r_2, which together define a participant's transition probability matrix.

$$P(TS)_t = \frac{\begin{pmatrix} r_1 & (1-r_2) \\ (1-r_1) & r_2 \end{pmatrix} \cdot \begin{pmatrix} P(TS_1)_{t-1} \\ P(TS_2)_{t-1} \end{pmatrix} * \begin{pmatrix} P(context_t|TS_1) \\ P(context_t|TS_2) \end{pmatrix}}{N}$$

We can define a number of cases corresponding to different inference strategies: if r_1 and r_2 equal .9 (the true recursive probability) the participant is Bayes optimal; if r_1 and r_2 are less than .9 the participant overweights the probabilistic cue (with $r_1 = r_2 = .5$ being the special "base-rate neglect" case); if r_1 and r_2 are greater than .9 the participant overestimates the transition probabilities. In

reality, the transition matrix should be symmetric, but we estimate the transitions associated with each TS separately to allow for participant bias such that they prefer to choose one TS over the other without any evidence.

Simpler models are possible by fixing some of the parameters. We contrast the *bias-2* observer to three related models: one where the transition matrix is forced to be symmetric (*bias*-1), an optimal observer where the transition matrix is fixed based on the training run (*optimal)*, and a base-rate neglect model with a fixed transition matrix as defined above (*base-rate neglect).*

The likelihood on each trial was calculated based on TS positional distributed, which were defined by the mean and standard deviation for each TS observed during the training phase. While this assumption is optimistic in regards to the task statistics participants encoded during training, as long the participant's estimation errors are not systematically biased away from the true statistics, the models should reflect participant performance in the aggregate.

Each model resulted in a vector of trial-by-trial TS posteriors for the testing phase of each participant. While an optimal decision maker would select the most likely TS, we assume some noise in translating posterior estimates into action. Thus we fit an ε parameter to each model which reflects the probability of randomly choosing a TS. This lead to a final vector of trial-by-trial TS choice likelihoods, which were used to fit each model.

Individual participant parameter estimates were fit using python's lmfit module's L-BFGS method, with the cost equal to the -log likelihood of that participant's TS selections. Model selection was accomplished using Bayesian information criterion (BIC: Schwartz, 1978), and by fitting the models on either the first or last half of the data and testing on the left out half. When discrete TS choices were needed, they were defined by the maximum likelihood on each trial across the posterior.

We were also interested in whether more difficult decisions were related to reaction times (RT), as predicted by a number of studies relating choice confidence and RT (e.g. Henmon, 1911; Roitman & Shadlen, 2002). We defined a trial-by-trial estimate of model choice confidence based on the average distance from .5 across the TS posteriors, ranging from 0 (indifference) to 1 (certainty). Because there were only two possible TSs, this is equivalent to calculating the distance from the choice boundary between the two TSs. We assessed this relationship with a mixed-effects linear regression, using the lme4 package.

Results

Context and aspects of context history significantly predicted TS choice for learners ($p < .001$), but not for nonlearners (Figure 2a). In addition, prior choice was significantly predictive in both groups. When included in the same model, prior choice abolished the effect of context history on participant choice in the learner group.

Model comparison across the population showed that the *bias-2* observer was a significantly better fit than any of the comparison models (Table 1). Moreover, individual analysis showed that both the *bias-2* and *bias-1* observers fit better than the base-rate neglect or optimal models for participants. Converting the *bias-2* posteriors into TS choices, we found that model choices matched individual participant's choices well (μ=87.6%, σ=4.5). Competing models were also relatively successful at capturing participant choices (bias-1: μ=86.1%, σ=5.0; optimal: μ=82.7%, σ=5.3%). Each model's likelihood for individual participants is shown in Figure 2.

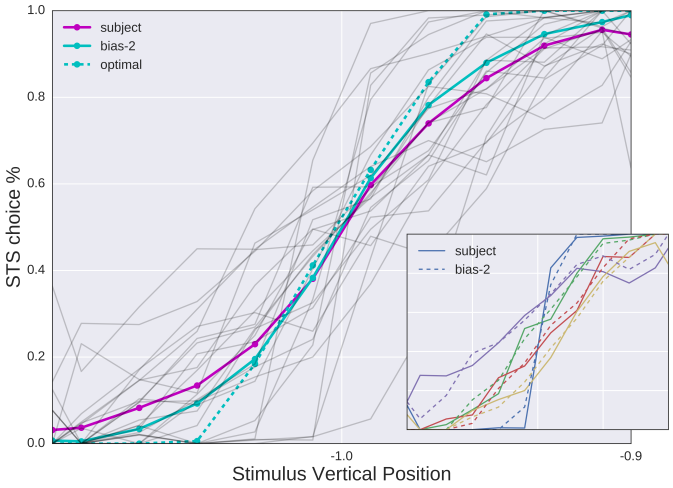


Figure 3: Percentage of trials for each of the 12 vertical position bins where responses reflected STS selection. The stimulus was never shown exactly at the midline. The purple line shows the average percentage chosen across all participants. The teal lines show the bias-2 and optimal model performance. Though not shown the bias-1 lies between these two curves. Individual participants curves are shown in light gray. (*Inset)* 5 example participant fits.

Parameter estimates showed no systematic preference for one TS over another as measured by the parameter estimates of r_1 (μ=88.9%, σ=13.8) and r_2 (μ=86.8%, σ=16.9). Overall, the population average transition matrix is quite similar to the true recursion probability of 90%, though there is a slight population-wide bias to overvalue the stimulus's vertical position. However, while this population-wide estimate is close to the true statistics (and therefore close to optimality), there is incredible variability across participants indicating that the population summary may mask consistent inferential biases that are distributed around an optimal strategy. While we interpret these differences as relation to biases in the TS inference process, it is possible that they instead stem from individual differences in encoding the environmental structure, which would affect TS inference. To address this we looked at the participant-reported estimates of the task statistics during a post-task questionnaire. These estimates were less accurate (STS: μ=68.2%, σ=21.5; CTS: μ=70.6%, σ=19.8) than the parameter estimates and were not significantly related to the bias-2 observer parameter estimates (CTS: r = -.13(24), p = .53; STS: r = .25(24), p = .23).

Table 1: Model *BIC across participants, n = 18,906*

Model	BIC
Bias-2	11905
Bias-1	12841
Optimal	14470
Base-Rate Neglect	18167

To visualize the model fits, we calculated the proportion of times participants selected the STS at each vertical position (Figure 3). All models predict that the STS should be chosen more frequently for higher contextual values. Also shown are individual traces (in gray) highlighting the large heterogeneity in individual performance as well as example individual bias-2 fits (inset), which demonstrate the flexibility of the model to capture these large individual differences.

Finally, decision confidence as estimated by *bias-2* was inversely related to RT (β = -.35(.01), t = -25.50). The regression predicts that when choice confidence equaled 1, participants responded 254 ms faster than when it equaled 0. Random effects analysis showed that this trend was true for all but one subject. Five representative participants are shown in Figure 4.



Figure 4: Two sample participant reaction times plotted against the *bias-2* model confidence. 0 indicates that both CTS and STS had a posterior probability of .5, while 1 indicates that either CTS or STS had a posterior probability of 1. Individual regressions are also shown.

Discussion

Using a novel task-switching task, we investigated how people integrate probabilistic evidence in the service of task-set (TS) selection. We found that of the people classified as “learners”, most based their decisions on both the probabilistic cues and transition probabilities, consistent with their internalizing the latent structure of the environment. On the population level, participants appeared to correctly identify the true statistics of the environment, giving the impression that they behaved in accordance with optimal Bayesian inference. However, individual participants differed greatly in their weighting of the two sources of information, such that some overvalued the probabilistic cue when making their choice.

The importance of this distinction is particularly clear when predicting RT from model estimates of trial-by-trial choice confidence. As choice confidence is a continuous metric, it is particularly sensitive to specific trial sequences, as well as parameter estimates. During test, each trial’s TS is estimated based on the encoded transition probabilities, the posteriors over TSs on the previous trial, and the probabilistic cue. Thus it is imperative to have an individual-specific estimate of the encoded transition probabilities to analyze trial-by-trial performance. In this task, the estimate of model confidence inversely related to RT. This parameter was defined by the absolute distance from the choice boundary between the two TSs, suggesting that this distance may relate to the speed of evidence accumulation in a way analogous to perceptual decision tasks (Roitman & Shadlen, 2012). Evidence accumulation in higher-level decision making has been suggested before by Shadlen & Kiani (2013), where they forward the idea that accumulators may serve as a general algorithmic description of many cognitive computations. The relationship between RT and choice confidence would support this description.

A related idea is that RT relates to decision conflict. Difficult decisions are, by definition, closer to the choice boundary indicating that the evidence does not clearly favor a particular action. On a neural level this conflict may stem from the concurrent representation of multiple TSs which must compete in a winner-take-all fashion before gating lower-level actions (Collins & Frank, 2013). If this competition is probabilistically resolved in proportion to each TSs representational strength, this idea is just a restatement of evidence accumulation for mutually exclusive alternatives.

The best fit model had two free parameters, which together represent a bias towards the STS or CTS (reflected in an asymmetrical transition matrix) and the encoded recursion probability. Differences in the recursion probabilities may either reflect individual differences in encoding of the task statistics or biased weighting during decision-making. For instance, if participants encoded the true transition probabilities, but only attended to the stimulus position when making a choice, the model would estimate an “encoded” recursion probability of .5. While it is impossible to completely disentangle these two alternatives, the lack of correspondence between the parameter estimates and the participant estimates on the post-task questionnaire suggests a decision bias, rather than an encoding bias. However, due to the possibility that encoded task statistics are not directly available to semantic retrieval during the questionnaire, we cannot rule out either possibility.

Regardless of whether variability is linked to encoding or the decision process, an obvious question emerges: what underlies these individual differences? Participants undoubtedly arrived at the experiment with different prior expectation for the kinds of rules that may be operating within a psychology experiment. While we attempted to normalize their expectations by orienting them to the TSs of interest (shape or color), the prior expectations for higher order rules may have prevented some people from appropriately integrating certain information. This may

partially explain why some people were unable to learn any rule at all - their prior beliefs constrained the search space, preventing the encoding of the relevant variables.

Similarly, early identification of a particular pattern may have stifled later learning - a type of confirmation bias that may have attentional roots. Participants who identified the relationship between TS and vertical position may have been less motivated to search for more complicated relationships. While we expect that the relationship between transition probabilities and TS selection relates more to unconscious statistical reasoning than explicit rules, it may be that explicit adherence to a particular rule overwhelmed other potential factors. In addition, lower level processes like perseverance may compete with these cognitive strategies, as suggested by the significant relationship between prior choice and TS choice in the non-learner group. Further work exploring their effects may refine our understanding of TS selection and allow us to account for the behavior of the substantial portion of non-learners.

In this work we compared model choices to participants with the simple maximum likelihood linking function. Our success in fitting participants without relying on a softmax rule indicates that this decision behavior may deviate from the probability matching widely reported in the decision-making literature (Erev & Barron, 2005). From the perspective of hierarchical reinforcement learning, there is no particular reason to believe that a single decision rule underlies decision-making at various levels of abstraction. It is possible that TSs are selected by a qualitatively different process than lower-level action selection, as proposed by Collins & Koechlin (2012). One alternative hypothesis is that higher-order action constructs like TSs are simply less noisy than lower-order decisions. Conflict resolution would consistently favor the stronger (more supported) TS, leading to the appearance of maximization behavior without appealing to fundamentally different computations. Future work could pursue this hypothesis by selectively degrading the observable evidence that contributes to TS selection.

Conclusion

We have shown that people can successfully leverage probabilistic information to infer a decision-relevant world state. While the group results seem to indicate that people act in accordance with Bayesian optimality, individual analysis reveals large heterogeneity in the inference strategy. The bias-2 model was able to capture much of this variation, suggesting that TS inference can be viewed as a biased integration over multiple information sources.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98(3), 409–429.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. Neuron, 66(2), 315–326.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: creating, clustering, and generalizing TS structure. Psychological Review, 120(1), 190–229.

Collins, A., & Koechlin, E. (2012). Reasoning , Learning , and Creativity : Frontal Lobe Function and Human Decision-Making, 10(3).

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Raymond, J. (2012). Model-based influences on humans' choices and striatal prediction errors, 69(6), 1204–1215.

Doya, K., Samejima, K., Katagiri, K., & Kawato, M. (2002). Multiple Model-Based Reinforcement Learning. Neural Computation, 14, 1347–1369.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. Psychological Review, 112(4), 912–931.

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. Computer Vision and Image Understanding, 106(1), 59–70.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cerebral Cortex (New York, N.Y. : 1991), 22(3), 509–26.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. Psychological Review, 117(1), 197–209.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology, 20(2), 251–6.

Henmon, V. a. C. (1911). The relation of the time of a judgment to its accuracy. Psychological Review, 18(3), 186–201.

Redish, a D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychological Review, 114(3), 784–805.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 22(21), 9475–9489.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. Neuron, 80(3), 791–806.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. Cognition, 120(1), 1–25.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychological Review, 119(1), 120–54.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7), 309–318.

N400 amplitudes reflect change in a probabilistic representation of meaning: Evidence from a connectionist model

Milena Rabovsky (rabovsky@stanford.edu)
Steven S. Hansen (sshansen@stanford.edu)
James L. McClelland (mcclelland@stanford.edu)
Department of Psychology, Stanford University

Abstract

The N400 component of the event-related brain potential is widely used in research on language and semantic memory, but the cognitive functions underlying N400 amplitudes are still unclear and actively debated. Recent simulations with a neural network model of word meaning suggest that N400 amplitudes might reflect implicit semantic prediction error. Here, we extend these simulations to sentence comprehension, using a neural network model of sentence processing to simulate a number of N400 effects obtained in empirical research. In the model, sequentially incoming words update a representation capturing probabilities of elements of sentence meaning, not only reflecting the constituents presented so far, but also the model's best guess at all features of the sentence meaning based on the statistical regularities in the model's environment internalized in its connection weights. Simulating influences of semantic congruity, cloze probability, a word's position in the sentence, reversal anomalies, semantic and associative priming, categorically related incongruities, lexical frequency, repetition, and interactions between repetition and semantic congruity, we found that the update of the predictive representation of sentence meaning consistently patterned with N400 amplitudes. These results are in line with the idea that N400 amplitudes reflect *semantic surprise*, defined as the change in the probability distribution over semantic features in an integrated representation of meaning occasioned by the arrival of each successive constituent of a sentence.

Keywords: neural network model; sentence comprehension; language; event-related potentials; N400; semantic surprise

Introduction

Language and meaning processing have been investigated with event-related brain potentials (ERPs), providing continuous time-resolved measures of electrical brain activity, and with models that characterize the processing of language and meaning, either in terms of the principles that govern these processes or the processes that implement them. Integration of these approaches could constrain selection among alternative models of the computations and the processes that implement them, while also providing for a possible integrated explanation of the diverse set of empirical phenomena that have been discovered through ERP research. In this spirit, the current work builds on other work described below, aiming to contribute to a better understanding of the computational principles and functional processes underlying the N400 ERP component, an electrophysiological indicator of meaning processing (see Kutas & Federmeier, 2011, for review).

The N400 is a negative deflection at centro-parietal electrode sites peaking around 400 ms after the presentation of a meaningful stimulus. N400 amplitudes have been shown to be modulated by a wide variety of variables. For example, the N400 is modulated by contextual fit, with larger amplitudes to incongruent as compared to congruent sentence continuations, such as when the sentence "He spread the warm bread with…" is completed by the word "socks" instead of "butter". The N400 also shows larger amplitudes to congruent continuations with lower as compared to higher cloze probability; decreasing amplitudes over the course of a sentence; smaller amplitudes for targets after semantically or associatively related as compared to unrelated primes; smaller amplitudes for repeated words as compared to a first presentation; and smaller amplitudes for words of high as compared to low lexical frequency.

Despite the large body of data on N400 amplitude modulations and widespread agreement that the N400 is related to semantic processing, the computational principles and processing mechanisms underlying N400 amplitude generation are as yet unclear. Various verbally-formulated theories are currently under active debate proposing, e.g., that N400 amplitudes reflect lexical and/or semantic access, semantic integration/ unification, semantic binding, or semantic inhibition (reviewed by Kutas & Federmeier, 2011). The development of an explicit computational account that addresses this diverse range of phenomena would thus appear to be a worthwhile goal.

There are at least two ways in which one could seek to better understand the N400 component by means of computational modeling. First, one might try to implement a neurobiologically realistic model of the brain processes underlying the N400 component, an approach that makes it possible to also model the morphology of the ERP waveform (Laszlo & Plaut, 2012; Laszlo & Armstrong, 2014). Another approach is to directly relate variations in N400 amplitudes to measures obtained from functional-level models of cognitive processes. While this approach may entail losing some possibly interesting information with respect to neural realization, it allows the modeling process to focus on the goal of better understanding the cognitive functions underlying N400 amplitudes. Many neural network models are of this functional type, in that the model is viewed as conforming to a computational principle characterized at the functional level. The principle is often articulated in terms of the goal to maximize consistency

with information in the environment which can then be formalized in terms of the more specific goal of minimizing *prediction error* (Hinton, 1987; Rumelhart et al., 1995). Focusing on this functional level, Rabovsky & McRae (2014) used an attractor network model of word meaning to investigate which measure in the model covaries with N400 amplitudes over a series of typical N400 word processing paradigms. They consistently observed a close correspondence between N400 amplitudes and network error at the semantic layer of representation, and took this correspondence to suggest that N400 amplitudes reflect implicit semantic prediction error. Here we extend this approach to the processing of words in sentences.

What kind of model would be most appropriate to simulate N400 amplitudes in sentences? Simple recurrent network models (SRNs; Elman, 1990) are typically trained to predict the next word in sentences based on the preceding context so that network error in these models reflects an implicit prediction error which could correlate with N400 amplitudes. Indeed, such a correlation was recently reported by Frank et al. (2015) who used four information measures derived from three probabilistic language models as predictors for six ERP deflections (including the N400). However, the prediction error in SRNs trained to predict the next input based on the preceding context is not specific to semantics but rather reflects *word surprisal* (the negative log of the probability of a word in a specific context) and is affected by both syntactic and semantic expectation violations (Levy, 2008). As the N400 is a functionally specific indicator of meaning processing (Kutas & Federmeier, 2011) while syntactic violations typically modulate different ERP components, we therefore decided against using such an SRN, and instead simulated N400 amplitudes using a model that is specifically trained to understand and predict sentence meaning, the Sentence Gestalt (SG) model (McClelland et al., 1989).

The SG model minimizes the mismatch between its estimates of the probability of semantic features of events given the words presented so far in a sentence and the observed probabilities of these features in the meanings of sentences, such that, once it has learned, its estimates after each new word encountered as a sentence unfolds should come close to matching the true probabilities. Thus the model can be characterized as an implicit probabilistic model of sentence comprehension: The model's outputs can be seen as representing conditional probability distributions over possible semantic features of the events described by the sentence up to and including the latest word. Furthermore, the magnitude of the update of the hidden unit state produced by the presentation of the latest word can be characterized as reflecting the change in this probability distribution produced by the word. We use a measure of this update we call *semantic surprise,* based on a measure that has been called Bayesian surprise (Itti & Baldi, 2005). Formally, the semantic surprise (*SemS*) produced by the *nth* word in a sentence is defined as the difference between the probability distribution over semantic feature representations consistent with the sentence through the *nth* word and the distribution consistent with the sentence through the preceding word, as measured by the Kullbach-Leibler divergence:

$$SemS_n = \sum_r p(r|n) \log_2(p(r|n)/p(r|n-1))$$

Here r indexes alternative possible patterns of semantic features of the event being described by the sentence and $p(r|n)$ and $p(r|n\text{-}1)$ denotes the probability of that pattern given the sentence up through word n and n-1 respectively.

The change in activation at the hidden (SG) layer of the model reflects this semantic surprise, and, as shown in a series of simulations of empirical N400 effects described below, models the N400, suggesting that the N400 is itself a measure of semantic surprise.

The Sentence Gestalt model

The SG model does not assume that sentences are represented in a specific propositional format. Instead, it is based on the idea that the task of sentence processing consists in processing sequences of incoming words to build representations enabling correct responses to various probes, and the model is allowed to find the best way to build these representations in order to meet the imposed demands through adjustments of connections between simple processing units organized in layers. A detailed description of the model is provided elsewhere (McClelland et al., 1989); we briefly sketch it here. For the current simulations, the model was re-implemented in the PDPTool software, V3 (http://web.stanford.edu/group/pdplab/pdphandbookV3/).

Architecture. The model can be conceptualized as consisting of two parts (see Fig. 1). The first part sequentially processes each incoming word (presented at the input layer) to update activation in the SG layer which represents the model's best guess interpretation of the meaning of the sentence as a whole, using the previous activation of the SG layer together with the activation induced by the new incoming word to produce the updated SG layer activation. The second part of the model is used primarily for performance assessment and training, decoding the content from the SG layer by probing it concerning the event described by the sentence.

Environment and training. It is important to note that the statistical regularities underlying the model's best guess interpretation of the meaning of the sentence at a given point in its presentation are determined by the training set so that the effects on semantic surprise depend on the training set as well. There are two different approaches to training which are complementary in that they each have strengths and weaknesses. First, models can be trained on large-scale training corpora approximating real life language environments of human participants. While this approach allows for simulation of empirical experiments with the exact same stimuli on a single-trial basis, the factors

responsible for the effects produced by the model may remain somewhat opaque. A second possible approach is to train models on a synthetic training set which implements variation among certain dimensions considered to be relevant for the target empirical phenomenon (the N400, in our case) and/or the theory advanced to explain the empirical phenomenon (semantic surprise, in our case). While this approach is limited in its capacity to fully explain specific empirical data points, it is more transparent concerning the general factors and principles responsible for the effects produced by the model. Because the main aim of the current study is to advance a theory concerning the functional basis of N400 amplitudes by highlighting the common core shared by the different dimensions that have been shown to modulate them, this transparency concerning the responsible factors is of primary importance to our goals. Thus, we trained our model on a small synthetic training set, aiming to create statistical regularities in the training set that allowed us to run simulation experiments containing manipulations corresponding to manipulations in empirical N400 experiments. We observe, based on these simulation experiments, that variables or dimensions that influence N400 amplitudes in the world influence semantic surprise in our model in the same way, suggesting that N400 amplitudes reflect semantic surprise.

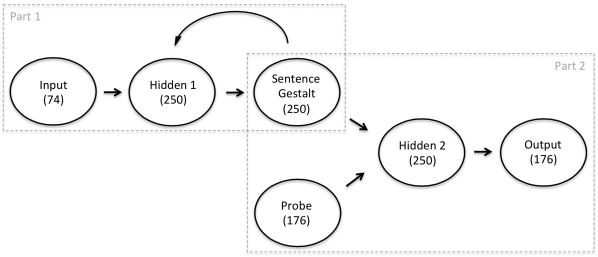


Fig. 1: The Sentence Gestalt (SG) model.

The model environment consists of sentences (presented word by word at the input layer) such as 'At breakfast, the man eats eggs' each paired with a corresponding event (a set of role-filler pairs, e.g., agent – the man), probabilistically generated online during training according to pre-specified constraints. After each presented word (represented by a word-specific unit at the input layer), the model is probed concerning the event described by the sentence. Responding to a probe consists in completing a role-filler pair when probed with either a thematic role (i.e., agent, action, patient, location, or situation; each represented by an individual unit at the probe and output layer) or a filler of a thematic role (e.g., the man, to eat, the eggs, etc.). For the filler concepts, we used feature-based semantic representations that were handcrafted so that members of the same semantic category shared some semantic features. For example, somewhat similar to the representations used by Rogers and McClelland (2008), all living things shared a semantic feature ('can grow'), all animals shared an additional semantic feature ('can move'), all birds shared one more semantic feature ('can fly') and then the canary had two individuating features ('can sing' and an item-unique individuating feature) so that the robin and the canary shared three of their five semantic features while the salmon and the canary shared two features, the rose and the canary shared only one feature, and the jeans and the canary did not share any features. While the labels for the features are irrelevant for model behavior, the aim in constructing the representations was to create graded similarities between concepts roughly corresponding to real world similarities. A comparison between a similarity matrix of the concepts based on the hand-crafted semantic representations and representations based on semantic word vectors derived from co-occurrences in large text corpora (Pennington, Socher, & Manning, 2014) suggested a reasonable correspondence (r = .73). Such feature-based semantic representations were also employed in the original version of the model; this allows us to capture the influence of semantic similarity over and above the influence of co-occurrence in language as implemented via the presented sentences (enabling simulation of categorically related semantic incongruities; Sim. 1).

After each presented word, the model is probed for each thematic role and each filler of each role-filler pair involved in the described event, and the model's activation at the output layer is compared with the correct output. Error is then back-propagated through the entire network and connections are adjusted to minimize the difference between model-generated and correct output (we used cross-entropy error, a learning rate of 0.00005 and momentum of 0.9). Because the model is probed concerning the described event after every single presented word, it anticipates the meaning of each sentence as early as possible, so that the activation at the SG layer (and accordingly at the output layer in response to the presented probes) becomes tuned to the regularities in the corpus. For example, the model learns that a sentence beginning "The woman writes…" more often describes the woman writing an email than an sms, and encodes this regularity in the connection weights, resulting in probabilistic pre-activations of units in the SG layer before email or sms appear in the sentence. Indeed the model's connection weights capture the base-rate probabilities of the semantic features of each of the roles in the sentence, so that when probed with a role prior to the presentation of the first word of a sentence the pattern over the filler units corresponds approximately to the overall probability across the entire environment that the feature will be present in the filler of the probed-for role.

Since the minimum of the cross-entropy error is reached when the network's estimates of feature probabilities match the actual probabilities of those features, the change in the network's estimates occasioned by each successive word should match the change in these actual probabilities (Rumelhart et al., 1995). Treating the semantic feature probabilities as conditionally independent given the words seen so far, this change in estimates of feature probabilities

can be shown to correspond to the $SemS_n$ measure defined in the introduction.

Thus far we have described how changes in the activation of semantic features in the model's output layer should correspond to the semantic surprise. However, we do not assume that these semantic feature activations are actually computed during sentence processing. Instead, we propose that the pattern of activation over the SG layer (together with the connection weights in Part 2 of the model) implicitly represent this probability distribution in such a way that the update at the SG layer mirrors the update in the actual probability distribution over features. We use the following cross-entropy measure to characterize this update:

$$\sum_i a_i(n) \log\left(\frac{a_i(n)}{a_i(n-1)}\right) + \left(1 - a_i(n)\right)\log\left(\frac{1 - a_i(n)}{1 - a_i(n-1)}\right)$$

Here i ranges over all of the SG layer units, $a_i(n)$ represents the activation of unit i based on the current word and $a_i(n-1)$ represents the activation of unit i based on the previous word. Similar results are obtained using the sum over SG units of the absolute value of the difference between $a_i(n)$ and $a_i(n-1)$.

Simulations

Sim. 1: Categorically related semantic incongruities. Federmeier and Kutas (1999) presented sentence pairs such as "They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of…" and observed gradually increasing N400 amplitudes from congruent sentence continuations ("palms") to unexpected continuations which were members of the same semantic category as the expected continuation and thus shared semantic features ("pines") to incongruent continuations.

To simulate these data, we trained the model such that one member of each semantic category (i.e., trees, drinks, etc.) was never presented in the same sentence context (i.e., as a patient of the same action) as the other category members so that it was completely unexpected in that context. For the simulation experiment, we presented the model with 10 such unexpected sentence continuations which were categorically related to the congruent continuations, as well as 10 congruent continuations, presented with a probability of .8 during training when the specific combination of agent and action had been presented, and 10 incongruent sentence continuations which were never presented as patients of the specific action during training and did not share any semantic features with the congruent continuations.

As shown in Fig. 2A, semantic surprise induced by the critical words gradually increased from congruent continuations to unexpected continuations categorically related to the congruent continuations, $t_{(9)} = 5.23$, $p < .0001$, and from those to unexpected continuations unrelated to the congruent continuations, $t_{(9)} = 6.30$, $p < .0001$, in line with N400 amplitudes.

The model also successfully captures other typical N400 effects in sentences such as cloze probability effects, with larger N400 for low cloze as compared to high cloze probability sentence continuations (**Sim. 2**; Kutas & Hillyard, 1984), sentence position effects with decreased N400 amplitudes over the course of a sentence (**Sim. 3**; Van Petten & Kutas, 1990), and influences of so-called semantic illusions or reversal anomalies, i.e. only a very slight increase of N400 amplitudes in sentences such as 'For breakfast, the eggs eat…' as compared to 'For breakfast, the boys eat…' while the increase in 'For breakfast, the boys plant…' is much larger (**Sim. 4**; Kuperberg et al., 2003; also simulated by Brouwer, 2014, PhD thesis, see below). Even though the SG model is designed as a model of sentence processing, word pairs and isolated words should be processed by the same system so that we also used the model to simulate N400 effects outside of sentence context. We describe the simulation of semantic priming in detail.

Sim. 5: Semantic priming. Bentin et al. (1985) observed smaller N400 amplitudes to target words presented after semantically related primes (i.e., primes from the same semantic category as the targets) as compared to unrelated primes. To simulate these data, we presented the model with 10 word pairs where the referenced concepts were members of the same semantic category and thus shared semantic features at the output layer (e.g., monopoly – chess) and 10 word pairs where the primes and targets from the related pairs were re-assigned such that there was no semantic similarity between prime and target. As shown in Fig. 2B, semantic surprise was smaller for targets after semantically related as compared to unrelated primes, $t_{(9)} = 5.14$, $p < .0001$, in line with N400 amplitudes (Bentin et al., 1985). The SG model additionally captures several other N400 effects in word processing such as associative priming, with smaller amplitudes to targets after associatively related (e.g., play – chess) as compared to unrelated primes (**Sim. 6**; Kutas & Hillyard, 1989), repetition priming, i.e. smaller N400 amplitudes to immediately repeated words as compared to target words presented after unrelated primes (**Sim. 7**; Nagy & Rugg, 1989), and smaller amplitudes to words of high as compared to low lexical frequency (**Sim. 8**; Van Petten & Kutas, 1990), captured through the encoding of base rate probabilities of features in the model's connection weights.

Finally, probability distributions can change and anticipatory preparedness to likely upcoming features depends on constant adaptation of represented probabilities based on new experiences. In neural network models, this adaptation is driven by the difference between expected and observed outcomes. Thus, if N400 amplitudes reflect this difference then larger N400 amplitudes should entail enhanced adaptation. Simulation 9 focuses on this relation.

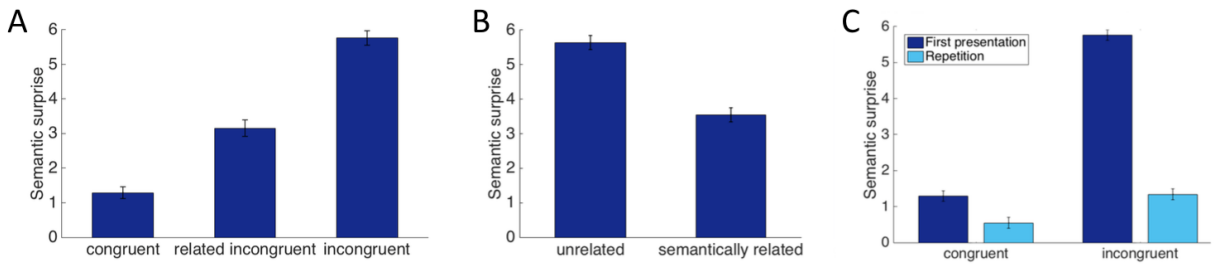


Fig. 2: Influences of (A) categorically related semantic incongruities, (B) semantic priming, and (C) repetition X congruity.

Sim. 9: Semantic incongruity and repetition. The influence of semantic incongruity on N400 amplitudes is reduced by repetition due to a stronger repetition-induced reduction of amplitudes for incongruent continuations (Besson et al., 1992). Repetition effects have been simulated as consequences of connection adjustments induced by the prior presentations of the item within the experiment. We thus presented the 10 congruent and 10 incongruent sentences from Sim. 1 twice while learning was operative (learning rate = 0.0005) so that the first presentation served not only as an experimental condition but also produced connection adjustments.

As shown in Fig. 2C, the difference in semantic surprise between the critical congruent vs. incongruent words was smaller during repetition. A rmANOVA confirmed a significant interaction between repetition and congruity, $F(1,9) = 87.18$, $p < .0001$, reflecting a stronger influence of repetition for incongruent, $F(1,9) = 82.35$, $p < .0001$, $\eta_p^2 = .90$, as compared to congruent, $F(1,9) = 16.47$, $p = .003$, $\eta_p^2 = .65$, sentence continuations, in line with N400 data.

These results nicely illustrate the intrinsic relationship between semantic surprise and adaptation. However, there is a subtle issue with this simulation, namely that the output layer which drives learning represents the events described by the sentences such that the simulation assumes these events to be observed while processing the sentences. This is not true for the empirical experiment. Thus, in a prospective version of this simulation the error signal that drives learning should be derived from the difference between SG activation before and after the critical words.

Discussion

The goal of the present study was to investigate the functional basis of the N400 ERP component by relating N400 amplitudes to a computational model of sentence comprehension, the Sentence Gestalt (SG) model. Across a series of simulations of N400 effects, we consistently observed a correspondence between N400 amplitudes and semantic surprise as represented by the magnitude of the update of hidden unit activations that implicitly represent probability distributions over semantic features.

N400 amplitudes have been previously linked to changes in lexical activation (Brouwer, 2014, PhD thesis; see also Rabovsky & McRae, 2014, for discussion, and Crocker et al., 2010, for a model of the P600 component). Brouwer (2014) focused on reversal anomalies and took the very small increase of N400 amplitudes in sentences such as 'For breakfast, the eggs eat…' (Kuperberg et al., 2003) to indicate that the N400 component does not reflect semantic integration but the retrieval of lexical information. He further suggests that semantic integration is linked to the P600 component (which is increased in reversal anomalies). Our account differs from Brouwer's in that it specifies a single integrated representation of meaning which is updated whenever a word is presented, with the extent of this update reflected in N400 amplitudes. When the word is presented in a sentence, the update is seen as an update of an integrated representation of the meaning of the sentence. The integration process is relatively heuristic and may not accord with syntactic constraints in constructions such as reversal anomalies. Indeed, analysis of our model's output layer suggests that it experiences a 'semantic illusion' in that it continues to assign the eggs to the patient instead of the agent role even after the word 'eat', in line with the suggestion that language processing can be shallow (Ferreira et al., 2002). Our model does not address the P600. It is possible that the P600 effect in reversal anomalies reflects a re-assignment of the eggs to an agent role. Alternatively, the comprehender may continue to see the eggs as being eaten, with the P600 reflecting detection of a syntax error. As a third possibility, the P600 may reflect conflict monitoring triggered by competing interpretations, one arising from a heuristic process and the other arising from a controlled process (van Herten et al., 2006). Further research seems required to better understand the P600.

As noted above, the SG representation together with the model's weights latently predict the semantic features of each role filler in the sentence based on prior constituents, and the update of the SG due to the next constituent adjusts these latent predictions. Latent prediction in that sense means that the SG model (and presumably the brain) becomes tuned through experience to be prepared to respond to likely upcoming semantic features with little additional effort. This kind of latent prediction seems to range from the pre-activation of specific semantic features in sentence context (e.g., for the categorically related incongruities, where less semantic update is necessary when an unexpected sentence continuation shares semantic features with an expected continuation; Sim. 1) to the latent structure in connection strengths and default activation that leads to less semantic update when processing a high frequent as compared to a low frequent word in isolation

(Sim. 8). This characterization should make clear that prediction in that sense does not refer to explicit intentional prediction of specific items but rather to a general configuration of the system optimized for upcoming semantic information. This entails that semantic activation changes induced by new incoming input primarily reflect the discrepancy between probabilistically anticipated and encountered features, i.e. semantic surprise, in accordance with predictive coding (Friston, 2005). In line with this view, N400 amplitudes in many paradigms appear to have one thing in common, namely that they seem to be a function of the fit between the semantic features that are implicitly expected based on previously experienced regularities and those activated by the current stimulus.

In sum, the present study aimed to contribute to a better understanding of the functional basis of the N400 ERP component by relating N400 amplitudes to an implemented model of sentence comprehension. Across a series of simulations of N400 effects, we consistently observed a correspondence between N400 amplitudes and the update of conditional probability distributions over semantic features. Besides demonstrating that the SG model naturally captures electrophysiological indicators of internal cognitive dynamics during language comprehension, these results are in line with the idea that N400 amplitudes reflect semantic surprise as the extent of change induced by an incoming stimulus in a probabilistic representation of meaning.

Acknowledgments

Funded by the European Union's Horizon 2020 programme, Marie Sklodowska-Curie grant 658999 to Milena Rabovsky.

References

Bentin, S., McCarthy, G., & Wood, C.C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroen Clin Neuro, 60,* 343-355.

Besson, M., Kutas, M., & Van Petten, C. (1992). An event-related potential (ERP) analysis of semantic congruity and repetition effects in sentences. *JoCN, 4,* 132-149.

Brouwer, H. (2014). *The electrophysiology of language comprehension: A neurocomputational model.* PhD thesis.

Crocker, M.W., Knoeferle, P., & Mayberry, M.R. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain Lang, 112,* 189-201.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14,* 170-211.

Federmeier, K. & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *J Mem Lang, 41,* 469-495.

Ferreira, F., Bailey, K.G.B., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11,* 11-15.

Frank. S.L., Otten, L.J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain Lang, 140,* 1-11.

Friston, K. (2005). A theory of cortical responses. *Philos T R Soc B, 360,* 815-836.

Hinton, G. E. (1989). Connectionist learning procedures. *Artif Intell*, *40*(1), 185-234.

Itti, L. & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *NIPS* (Vol. 19), 547-554, Cambridge, MA: MIT Press.

Kuperberg, G.R., Sitnikova, T., Caplan, D., & Holcomb, P.J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cogn Brain Res, 17,* 117-129.

Kutas, M. & Federmeier, K. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol, 62,* 621-647.

Kutas, M. & Hillyard, S.A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203-205.

Kutas, M. & Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 161-163.

Kutas, M. & Hillyard, S.A. (1989). An electrophysiological probe of incidental semantic association. *JoCN, 1,* 38-49.

Laszlo, S. & Armstrong, B. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of ERP reading data. *Brain Lang, 132,* 22-27.

Laszlo, S. & Plaut, D. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain Lang, 120,* 271-281.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

McClelland, J.L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Lang Cognitive Proc, 4,* 287-335.

Nagy, M.E. & Rugg, M.D. (1989). Modulation of event-related potentials by word repetition: The effects of inter-item lag. *Psychophysiology, 26,* 431-436.

Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global vectors for Word Representation. *EMNLP*.

Rabovsky, M. & McRae, K. (2014). Simulating N400 amplitudes as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132,* 68-89.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. *Backpropagation: Theory, Architectures and Applications*, 1-34.

Van Herten, M., Chwilla, D.J., & Kolk, H.H.J. (2006). When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *JoCN, 18,* 1181- 1197.

Rogers, T.T. & McClelland, J.L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behav Brain Sci, 31,* 689-749.

Van Petten, C. & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Mem Cognition, 18,* 380-393.

Asking and evaluating natural language questions

Anselm Rothe[1], **Brenden M. Lake**[2], **and Todd M. Gureckis**[1]
[1]Department of Psychology, [2]Center for Data Science, New York University

Abstract

The ability to ask questions during learning is a key aspect of human cognition. While recent research has suggested common principles underlying human and machine "active learning," the existing literature has focused on relatively simple types of queries. In this paper, we study how humans construct rich and sophisticated natural language queries to search for information in a large yet computationally tractable hypothesis space. In Experiment 1, participants were allowed to ask any question they liked in natural language. In Experiment 2, participants were asked to evaluate questions that they did not generate themselves. While people rarely asked the most informative questions in Experiment 1, they strongly preferred more informative questions in Experiment 2, as predicted by an ideal Bayesian analysis. Our results show that rigorous information-based accounts of human question asking are more widely applicable than previously studied, explaining preferences across a diverse set of natural language questions.

Keywords: Bayesian modeling; active learning; information search; question asking

Cognitive science and machine learning have both explored the ability of learners to ask questions in order to gain information. In an active learning setting, a learning machine is able to query an oracle in order to obtain information that is expected to improve performance. This is often contrasted with passive learning where examples are presented without the advantage of active control. Machine learning research has shown that active learning can speed acquisition for a variety of learning tasks (Cohn, Atals, & Ladner, 1994; MacKay, 1992; Settles, 2009). Interestingly, humans seem to benefit from active learning in similar ways (Castro et al., 2008; Markant & Gureckis, 2015).

Although these studies have revealed common computational principles between human and machine active learning, they have largely sidestepped a hallmark human ability: the capacity for asking rich, sophisticated, and even clever questions using language. Active learning algorithms and most psychological studies emphasize relatively simple types of stereotyped, non-linguistic queries (essentially "What is the category label of document X?", Angluin, 1988; Markant & Gureckis, 2015). In contrast, people can ask far richer questions which more directly target the critical parameters in a learning task. For example, when learning about categories of animals, a child not only can point at examples and request a category label (e.g., "What is that object over there?") but can also ask about characteristic features (e.g., "Do all dogs have tails?"), typical examples (e.g., "What does a lemur look like?"), related categories (e.g., "How do alligators and crocodiles differ?"), and other types of questions which constrain the space of possible concepts (Graesser, Langston, & Bagget, 1993; Mills, Legare, Grant, & Landrum, 2011).

Despite this observation, little is known about how humans generate and evaluate natural language questions, particularly from a computational perspective. For example, how do people search an infinite space of possible questions? How do people evaluate different types of questions within a common currency? In the present paper, we begin to try to answer these questions by comparing the preference people have for asking certain natural language questions to different ways of valuing questions according to an ideal Bayesian analysis.

Studying question asking in the Battleship game

We examine question asking in a simple active learning task called the Battleship game due to its superficial similarity to a single-player version of the popular children's game (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014). The goal of the game is to determine the location and size of 3 non-overlapping ships on a 6×6 grid (Figure 1). The ships are horizontal or vertical and can be between 2 and 4 tiles long. During the standard game, a participant sequentially clicks on tiles to reveal either the color of the underlying ship part or an empty water part (*sampling phase*, Figure 1). An efficient active learner seeks out tiles that are expected to reduce uncertainty about the ship locations and avoids tiles that would provide redundant information (e.g., when the hidden color can be inferred from the already revealed tiles). At a certain point, the sampling is stopped and participants are asked to fill in the remaining tiles with the appropriate color, based on their best guess (*painting phase*, Figure 1). The score they receive is a combination of the number of observations made in the sampling phase and the number of correctly painted tiles.

The task is well suited for the present study because the underlying hypothesis space of possible ship configurations is relatively large (1.6 million possible game boards) but is easy to explain to participants prior to the start of the task. In addition, the game is interesting and fun for participants while being amenable to an ideal observer analysis (see below). The major innovation of the present paper is that we set up situations where participants can ask any question they want in natural language (e.g., "Are the ships touching?" or "What is the total area of the ships?"). This allowed us to study rich, natural language question asking in the context of a well understood active learning task.

Models of question evaluation

At times we notice somebody ask a question that strikes us as especially clever. But why do some questions seem better than others? Here we describe a set of models which provide an objective "yardstick" for evaluating the quality of participant's questions with respect to the goals of the task. In a given trial, a player must learn a hidden configuration corresponding to a single hypothesis h in the space of possi-

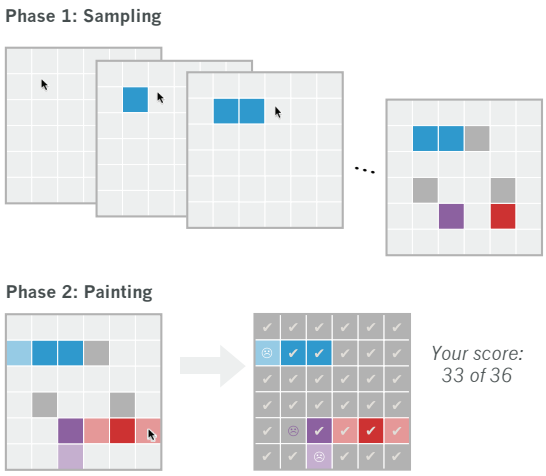


Figure 1: With the goal to locate all three ships (blue, red, purple) on the grid, a participant uncovers tile by tile. At a certain point, this *sampling phase* is stopped and the participant guesses the color of the remaining tiles. For each correctly painted tile one point is awarded.

ble configurations H. We model her prior belief distribution over the hypothesis space, $p(h)$, as uniform over ship sizes. The prior is specified by first sampling the size of each ship from a uniform distribution and second sampling uniformly a configuration from the space of possible configurations given those sizes. The player can make a query x (turning over a tile or asking a natural language question) and receives the response d (the answer). The player can then update her posterior probability distribution over the hypothesis space by applying Bayes' rule,

$$p(h|d;x) = \frac{p(d|h;x)p(h)}{\sum_{h' \in H} p(d|h';x)p(h')}. \quad (1)$$

The semi-colon notation indicates that x is a parameter rather than a random variable. The posterior $p(h|d;x)$ becomes the next step's prior $p(h|D;X)$, with X representing all past queries and D representing all past responses,

$$p(h|d,D;x,X) = \frac{p(d|h;x)p(h|D;X)}{\sum_{h' \in H} p(d|h';x)p(h'|D;X)}. \quad (2)$$

The likelihood function $p(d|h;x)$ is $\frac{1}{n}$ if d is a valid response to the question x (and zero otherwise). The normalizing constant, n, depends on the type of question asked. For example when asking for the coordinates of any one of the tiles that contain a blue ship n is defined by the number of blue ship tiles in the true configuration. However, for most queries that we collected $n = 1$. The posterior predictive value of a new query x resulting in the answer d is defined as $p(d|D;x,X) = \sum_{h \in H} p(d|h;x)p(h|D;X)$.

Expected Information Gain (EIG). According to EIG, the value of a query x is the expected reduction in uncertainty about the true hypothesis, averaged across all possible answers A_x of the query:

$$EIG(x) = \sum_{d \in A_x} p(d|D;x,X)\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big]$$

where $I[\cdot]$ is the Shannon entropy. EIG is closely related to machine learning approaches to active learning (Settles, 2009) and has a long history of study as a model of human information gathering (Oaksford & Chater, 1994).

Expected Savings (ES). According to ES, a query x is valued according to the expected reduction of errors in the painting task (Figure 1) averaged across all possible answers A_x of the query

$$ES(x) = \sum_{d \in A_x} p(d|D;x,X)\Big[EC[p(h|D;X)] - EC[p(h|d,D;x,X)]\Big].$$

Here $EC[p(h|v)]$ are the Expected Costs when coloring tiles in the painting task according to belief distribution $p(h|v)$,

$$EC[p(h|v)] = \sum_i \sum_l p(l|v;i) \times [C_{hit}\ p(l|v;i) + C_{miss}(1 - p(l|v;i))],$$

where the belief that tile i has color l is given by $p(l|v;i) = \sum_{h \in H} p(l|h;i)p(h|v)$. The choice to actually paint the tile in that color is here given by $p(l|v;i)$ again because we assume a probability matching choice function. $C_{hit} = 0$ and $C_{miss} = 1$ indicate the costs associated with painting a tile correctly or incorrectly, respectively.

Experiment 1 – Question Generation

There is an infinite number of questions that can be asked in any situation. However, most of them would have little or no information value while others would be highly informative. Our first experiment explored how people generate free-form, natural language questions in a modified version of the Battleship game. Our ultimate goal is to relate open-ended natural language questions to models of information utility.

Participants. Forty participants recruited on Amazon Mechanical Turk, with restriction to the United States pool, were paid a base of \$2 with a performance based bonus of up to \$3.60. Participants were awarded a bonus of \$0.20 for each generated question that was in line with the task rules, encouraging a minimum level of question quality without providing monetary incentives for especially rich and creative questions.[1]

Method. Before eliciting the natural language questions, a number of safeguards were implemented to help the participants understand the task. These included detailed tutorial-like instructions that explained the task and comprehension quizzes to verify understanding. In addition, key task information remained visible throughout the whole experiment.

[1] We decided against paying people based on question quality. For example, participants would have to reason about what we, the experimenters, expect to be good questions.

In a warm-up phase participants played five rounds of the standard Battleship game (i.e., turning over tiles to find the ships) to ensure understanding of the basic game play. Then, in the main phase, participants were given the opportunity to ask free-form questions in 18 trials. We defined 18 different "contexts" which refer to partially revealed game boards (see Figure 2A).

At the beginning of a trial, we introduced participants to a partly revealed game board by letting them click on a predetermined sequence of tiles (which are the past queries X and answers D in Equation 2). We chose this format of tile-uncovering moves, resembling the warm-up phase, to give the impression that a human was playing a game that was paused in an unfinished state. Subsequently, as a comprehension check, participants were asked to indicate the possible colors of each covered tile (e.g., whether the tile could be hiding a piece of the red ship). The task would only continue after all tiles were indicated correctly (or a maximum of six guesses were made).

Next, participants were given the following prompt: "If you had a special opportunity to ask any question about the grid, ships, or tiles - what would you ask?" (represented as x in Equation 2). A text box recorded participants' responses. The only two restrictions were that combinations of questions were not allowed (i.e., putting two questions together with "and" or "or") and questions had to be answerable with a single piece of information (e.g., a word, a number, true/false, or a single coordinate). Thus, participants could not ask for the entire latent configuration at once, although their creativity was otherwise uninhibited. Due to practical limitations participants asked only one question per trial, no feedback was provided and there was no painting phase. We emphasized to participants that they should ask questions as though they were playing the game they already had experience with in the earlier part of the experiment.

Question asking contexts. To produce a variety of different types of partial knowledge states or "contexts" from which people could ask questions, we varied the number of uncovered tiles (6 or 12), the number of partly revealed ships (0 to 3), and the number of fully revealed ships (0 to 2). These factors were varied independently while excluding impossible combinations leading to a total of 18 contexts/trials.

Results

We recorded 720 questions (18 trials $\times$ 40 participants). Questions that did not conform with the rules or that were ambiguous were discarded (13%) along with (3%) which were dropped due to implementation difficulties. The remaining 605 questions (84%) were categorized by type (see Table 1).

Question content. As a first stage of our analysis, we manually coded commonalities in the meaning of questions independent of the specific wording used. For example, the questions "How many squares long is the blue ship?" and "How many tiles is the blue ship?" have the same meaning for our purposes and were formalized as shipsize(blue), where shipsize is a function with parameter value blue. Since the function shipsize also works with red and purple as parameter values, it represents a cluster of analogous questions. Within these functional clusters we then considered the frequency by which such questions were generated across the 18 contexts to get a sense of participant's question asking approach (first column in Table 1).

Table 1: The natural language questions obtained in Exp. 1 were formalized as functions that could be understood by our model. The table shows a comprehensive list. Column N reports the number of questions people generated of that type. Questions are organized into broad classes (headers) that reference different aspects of the game.

N	**Location/standard queries**
24	What color is at [row][column]?
24	Is there a ship at [row][column]?
31	Is there a [color_incl_water] tile at [row][column]?
	Region queries
4	Is there any ship in row [row]?
9	Is there any part of the [color] ship in row [row]?
5	How many tiles in row [row] are occupied by ships?
1	Are there any ships in the bottom half of the grid?
10	Is there any ship in column [column]?
10	Is there any part of the [color] ship in column [column]?
3	Are all parts of the [color] ship in column [column]?
2	How many tiles in column [column] are occupied by ships?
1	Is any part of the [color] ship in the left half of the grid?
	Ship size queries
185	How many tiles is the [color] ship?
71	Is the [color] ship [size] tiles long?
8	Is the [color] ship [size] or more tiles long?
5	How many ships are [size] tiles long?
8	Are any ships [size] tiles long?
2	Are all ships [size] tiles long?
2	Are all ships the same size?
2	Do the [color1] ship and the [color2] ship have the same size?
3	Is the [color1] ship longer than the [color2] ship?
3	How many tiles are occupied by ships?
	Ship orientation queries
94	Is the [color] ship horizontal?
7	How many ships are horizontal?
3	Are there more horizontal ships than vertical ships?
1	Are all ships horizontal?
4	Are all ships vertical?
7	Are the [color1] ship and the [color2] ship parallel?
	Adjacency queries
12	Do the [color1] ship and the [color2] ship touch?
6	Are any of the ships touching?
9	Does the [color] ship touch any other ship?
2	Does the [color] ship touch both other ships?
	Demonstration queries
14	What is the location of one [color] tile?
28	At what location is the top left part of the [color] ship?
5	At what location is the bottom right part of the [color] ship?

At a broader level, there are natural groups of question types (Table 1). While this partitioning is far from the only possible scheme, it helps to reveal qualitative differences between questions. An important distinction contrasts *location/standard queries* with *rich queries*. Location queries ask for the color of a single tile and are the only question type afforded by the "standard" Battleship task (Gureckis & Markant, 2009; Markant & Gureckis, 2012, 2014). Rich queries incorporate all other queries in Table 1 and reference

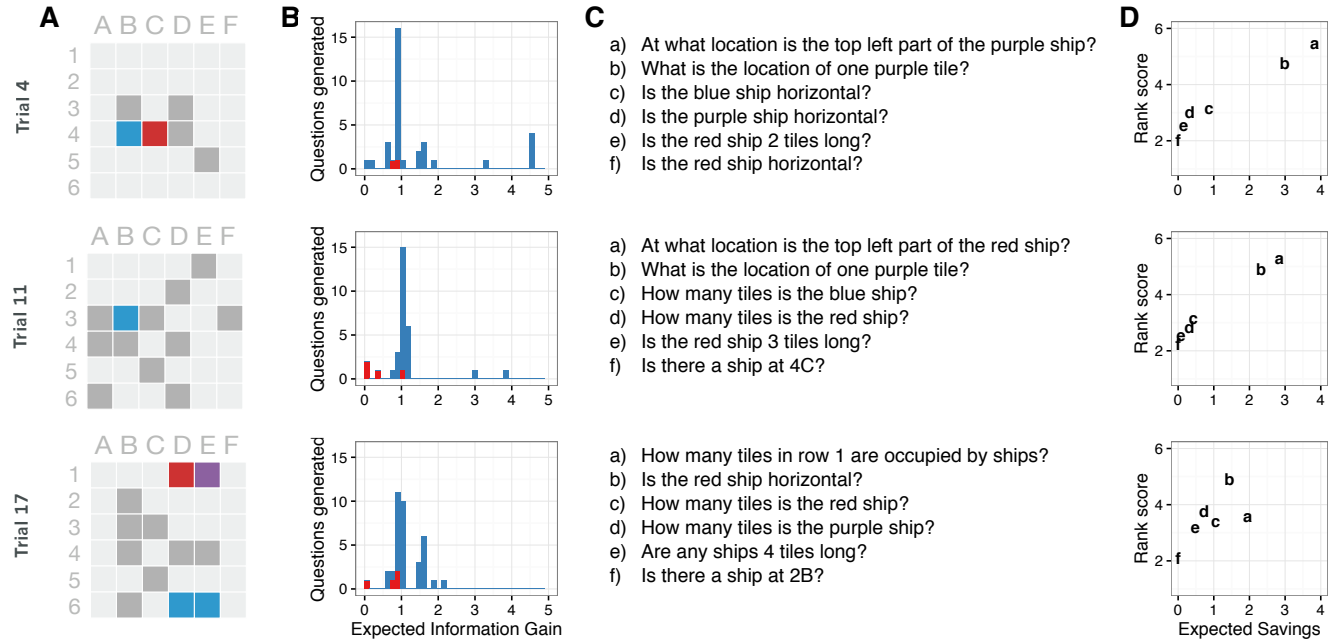


Figure 2: Three selected trials exemplifying (A) the partly revealed configuration, (B) the qualities of the questions (as measured by the Bayesian Expected Information Gain model) that participants generated in Experiment 1 (red = simple queries, blue = rich queries), (C) the six questions that were sampled from those obtained in Experiment 1 and presented in Experiment 2, (D) the values that participants (y-axis) and the Bayesian Expected Savings (ES) model (x-axis) assigned to these questions (higher score means better question).

more abstract properties of the game. Of the rich queries, *demonstration queries* ask for an example given a reference label. In the case of battleship, demonstration queries ask for an example tile of a ship, whereas most other rich queries ask about a part of feature of gameboard configuration. Demonstration queries can be especially helpful in active learning settings where the set of positive examples is relatively small (Cakmak & Thomaz, 2012; Hendrickson, Navarro, & Perfors, in press), as is the case in Battleship. Examples of high value demonstration queries are shown as the first two questions of Trial 4 and Trial 11 in Figure 2C.

Question frequencies. Among all 605 questions, only 13% were of the location/standard query type. In other words, being freed from the constraints of typical active learning studies, our participants creatively invented a vast array of questions. Only 47 questions (8%) were demonstration queries despite the fact that these can be especially useful (see below). In sum there were 139 unique questions that were repeated with different frequencies. The most popular questions was "How many tiles is the red ship?" ($n = 66$), followed by the same question asking about the purple ($n = 64$) or blue ship ($n = 55$). When grouping the questions, almost half of all generated questions ($n = 289$) addressed the size of one or several ships (Table 1). Another large group of questions targeted the orientation of the ships ($n = 116$).

Question quality. A more interesting analysis concerns the overall *quality* of these questions (as objectively assessed by the models described above).[2] One intriguing hypothesis is that there should be a positive relationship between the frequency by which a question is generated in a given context and the objective quality of the question. We used the EIG model to evaluate all 605 questions in their respective contexts. However, counter to our hypothesis, the objectively best questions were generated rarely (see Figure 2B for three example trials). The worst questions were also not generated often, while the most generated questions were in the intermediate range. Indeed, for all 18 contexts, each frequency distribution has a high peak and this peak is always below the maximum end of the distribution.

Interestingly, location/standard queries were generally inferior to richer queries. The mean EIG for location/standard queries was 0.77 compared to 1.26 for the rich queries, demonstrating the effectiveness of the more sophisticated queries (red vs. blue in Figure 2B).

Context specificity. A good question in a certain context is not necessarily a good question in a different context. To estimate the context sensitivity of the generated questions, we permuted the configurations each question was associated with across all 605 questions and evaluated the EIG for each new configuration-question pair. The average EIG across questions in the original data set was larger than in all 100 permutation sets, $p < 0.01$. Thus, people produced a range of questions that were both rich and context sensitive.

[2] Since both models make similar predictions for this section, we only report results for the EIG model here and save the model comparison for the analysis in Experiment 2.

Experiment 2 – Question Evaluation

In Experiment 2 we look more closely at how people *evaluate* natural language questions by having participants select what they viewed as the best question from a set taken from Experiment 1. In addition, we provide participants with the answer to that question.

Participants. 41 participants on Amazon Mechanical Turk were paid \$6 with a potential performance-based bonus of up to \$3.60. The higher payment compared to Exp. 1 was due to a longer experiment duration.

Method. The materials and procedure were nearly identical to Exp. 1, except that participants, rather than generating free-from rich questions from scratch, chose from a list of natural language questions asked by participants in Exp. 1. They received the answer to that question and could utilize this information in a subsequent painting phase.

Participants viewed the same 18 board configurations (contexts) along with a selection of six natural language questions which were sub-sampled from the full list of human-generated questions from the corresponding context in Exp. 1. They were asked to rank the questions for quality by positioning them from best to worst in a sortable list. After ranking they were provided with the answer to the top-ranked question and then had to do the painting task.

The reduction was necessary, as the intention of this experiment was to study question evaluation without the burden of having to consider a large number of possible questions. For sub-sampling, we used a simple algorithmic procedure designed to include the most frequently generated questions, the highest quality questions (according to EIG), and some questions that were neither frequent nor high quality.[3]

To ensure people read each question they ranked, they were asked to classify each question by the form of its possible answers (either a color, a coordinate on the grid, a number, or yes/no, which span all possible answers to the questions in Table 1). Only after a correct classification they were able to continue, in case of a wrong classification they had to wait for 5 seconds.

In contrast to Experiment 1, the bonus was tied to the performance in the painting phase. For each correctly painted tile, we awarded a potential bonus of \$0.10. The bonus was only paid for a single trial, selected by lottery at the end of the experiment. This allowed us to award a higher bonus per tile and also kept people motivated up to the very last trial.

Results

In each trial, participants ranked six rich questions by quality. Subsequently, they received the answer to the top-ranked question. In our analysis, a higher rank score represents a better question (i.e., 6 for the highest position and 1 for the lowest) and we will treat the top-ranked question as the "chosen" question.

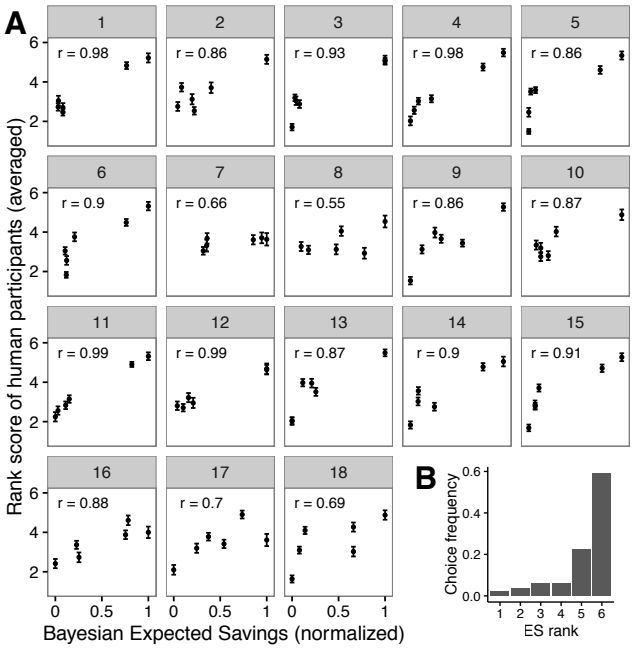


Figure 3: (A) Human rank scores (Experiment 2) and Bayesian Expected Savings (ES) model scores for the six questions per trial. Error bars show 1 se. (B) Human choice frequencies of the six questions ranked by ES, collapsed across all trials. Rank ties were resolved such that the choice counts were split between the ranks.

Figure 3A shows the high correlations between rank scores and question qualities, as measured by the Expected Savings (ES) model (average Pearson correlation $r = 0.84$). This is remarkable given that the model scores come from a Bayesian ideal-observer analysis without any free parameters.

People predominantly chose the best question (Figure 3B). The best question was the most selected question in all trials except 16 and 17, where the second-best question was favored slightly more often. Even more, the ES model scores are well reflected in the choice distributions within each trial (not shown), such that lower ES questions were selected less often (average Pearson correlation $r = 0.87$).

The correlations with the EIG scores are somewhat lower ($r = 0.70$ and 0.75 for ranking and choice, respectively). The different preferences of the models become clear with the example of the question "How many tiles are occupied by ships?" In many contexts this question has a high EIG value because it allows the learner to rule out many hypothesized configurations but a low ES value because such abstract information does often not help much with the painting task. For a more careful comparison of which model provides the best fit to the human rankings, the model utilities were transformed into choice probabilities via a softmax function with one free temperature parameter. For each model, the parameter was fit per participant to the choice data from Experiment 2. We found that ES had a higher log-likelihood for 30 out of 41 participants (73%). In addition, we looked at the

[3]The free-from questions for each context in Experiment 1 were placed in a 2D space with EIG and generation frequency as dimensions. We then sampled 1000 six question subsets and took the sample with the largest average pairwise distance between questions in the subset.

log-likelihood differences between the two models and set an arbitrary threshold to 1.6. We found an above-threshold difference in favor of ES for 26 participants (63%) but for zero participants in favor of EIG. Previous work has suggested that EIG provides a better account of human active learning than ES (Markant & Gureckis, 2012), but this work only considered location/standard queries as opposed to the rich questions we considered here.

Discussion and Conclusions

While humans and machines seem to benefit from active learning in similar ways, people ask far richer and more sophisticated types of questions. Previous experimental and computational work has used Bayesian analysis to model how people play 20 Questions with either pre-determined sets of questions (Cohen & Lake, 2016) or a small number of hypotheses (Ruggeri & Feufel, 2015; Ruggeri, Lombrozo, Griffiths, & Xu, 2015). However, people require no such restrictions. They can construct open-ended queries to resolve uncertainty in massive and novel hypothesis spaces. In this paper, we studied a probabilistic reasoning task which aimed to capture as much complexity as possible while remaining amenable to ideal Bayesian analysis. Most natural language questions can be precisely interpreted as constraints on the hypothesis space, allowing various measures of question quality to be computed exactly.

We draw a number of conclusions from two experiments. When freed from the typical constraints of active learning studies, people generated natural language questions from a rich space of possibilities spanning multiple qualitative types. In every studied context, there were a number of rich queries that were more informative than the best standard queries. Furthermore, questions were highly context sensitive and tuned to the particular partially observed game states that participants saw (as opposed to heuristic selections ignoring the current context). We found that the highest information questions were rarely generated spontaneously (Exp. 1), yet this was not because people do not recognize the quality of the questions (Exp. 2). Importantly, we were able to capture people's evaluations of natural, rich questions by a Bayesian model with zero parameters. In our setting, people's preferences are better described by the cost-sensitive measure Expected Savings rather than the cost-insensitive measure Expected Information Gain.

We also hope these findings will help inspire new, more human-like active learning algorithms. Some queries resemble the features that the generative model of a game board configuration has 'built-in' (e.g., the size or the orientation of a ship), relating to active learning algorithms that ask feature relevance queries for the purpose of classification (Settles, 2011). Unlike these algorithms, however, our participants referred to features that are inductive in nature (e.g., about ships touching each other, about one ship being larger than another, or about ships having parallel orientation). These features are interesting because they reference configural or emergent features which are not explicit in the Bayesian model. By querying these emergent properties, people must have either synthesized new features or transferred structure from related tasks. We hope this type of work will inform more human-like question asking machines.

Acknowledgments. This research was supported by NSF grant BCS-1255538, the John Templeton Foundation Varieties of Understanding project, a John S. McDonnell Foundation Scholar Award to TMG, and the Moore-Sloan Data Science Environment at NYU.

References

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319-342.

Cakmak, M., & Thomaz, A. L. (2012). Designing robot learners that ask good questions. In *Proceedings of the seventh annual acm/ieee international conference on human-robot interaction* (pp. 17–24). doi: 10.1145/2157689.2157693

Castro, R., Kalish, C. W., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2008). Human Active Learning. *Advances in Neural Information Processing Systems 21*.

Cohen, A., & Lake, B. M. (2016). Searching large hypothesis spaces by asking questions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Cohn, D., Atals, L., & Ladner, R. (1994). Improving generalization with Active Learning. *Machine learning*, *15*(2), 201–221.

Graesser, A. C., Langston, M. C., & Bagget, W. B. (1993). Exploring information about concepts by asking questions. *The Psychology of Learning and Motivation*, *29*, 411–436.

Gureckis, T. M., & Markant, D. B. (2009). Active Learning Strategies in a Spatial Concept Learning Game. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (in press). Sensitivity to hypothesis size during information search. *Decision*.

MacKay, D. J. C. (1992). Information-Based Objective Functions for Active Data Selection. *Neural Computation*, *4*(4), 590–604. doi: 10.1162/neco.1992.4.4.590

Markant, D. B., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Markant, D. B., & Gureckis, T. M. (2014). A preference for the unpredictable over the informative during self-directed learning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Markant, D. B., & Gureckis, T. M. (2015). Is It Better to Select or to Receive? Learning via Active and Passive Hypothesis Testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122.

Mills, C. M., Legare, C. H., Grant, M. G., & Landrum, A. R. (2011). Determining who to question, what to ask, and how much information to ask for: The development of inquiry in young children. *Journal of Experimental Child Psychology*, *110*(4), 539–560.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Ruggeri, A., & Feufel, M. A. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00918

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Settles, B. (2009). *Active Learning Literature Survey* (Tech. Rep.). University of Wisconsin-Madison.

Settles, B. (2011). Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1467–1478).

Sequential images are not universal, or Caveats for using visual narratives in experimental tasks

Neil Cohn (neilcohn@visuallanguagelab.com)
Tilburg center for Cognition and Communication (TiCC), P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract

Sequential images have frequently been used as experimental stimuli in the cognitive and psychological sciences to explore topics like theory of mind, temporal cognition, discourse, social intelligence, and event sequencing, among others. The assumption has been that sequential images provide a fairly universal and transparent stimuli that require little to no learning to decode, and thus are ideal for non-verbal tasks in developmental, clinical, and non-literate populations. However, decades of cross-cultural and developmental research have actually suggested something different: that sequential image comprehension is contingent on exposure and practice with a graphic system. I here review this literature and advocate for more sensitivity to the "fluency" needed to understand sequential images.

Keywords: visual narratives; sequential images; experimental methods; cross-cultural cognition

Introduction

Researchers in the cognitive and psychological sciences have often relied on wordless sequential images like comics in experimental tasks. Most often, these tasks focus on using sequential images for elicitation, while aiming to investigate some other aspect of cognition. We find sequential images used in popular procedures for theory of mind (Baron-Cohen, Leslie, & Frith, 1986; Sivaratnam, Cornish, Gray, Howlin, & Rinehart, 2012), temporal cognition (Núñez & Cooperrider, 2013), discourse (Gernsbacher, 1985), social intelligence (Campbell & McCord, 1996), and event sequencing (Tinaz, Schendan, Schon, & Stern, 2006), not to mention the longstanding use of the Picture Arrangement Task as a measure in the WAIS-IQ test battery and other clinical assessments (Kaufman & Lichtenberger, 2006; Ramos & Die, 1986). The assumption has been that sequential images provide a universally understood medium of communication that is transparent to participants. They are believed to be understandable by everyone—including young children—with little learning or decoding necessary since they are assumed to rely on perceptual processing alone (e.g., McCloud, 1993).

This assumed transparency is likely underscored by the iconicity of images and a logic that may go something like this: Sequential images typically depict events visually, and everyone understands events transparently via perceptual processing; thus sequential images should be understandable to everyone. This thinking has lead to their frequent use in experiments for children and clinical populations, and in anthropological work, especially with non-literate populations.

However, empirical research about the comprehension of sequential images challenges these assumptions. Though some of this research is several decades old, they are supplemented by recent growing work on visual narratives, and the overall lessons remain relevant and important for consideration. Here, I will review this literature on sequential image comprehension, focusing on cross-cultural research, developmental research, and recent neurocognitive research. Altogether, such work will show that sequential images require exposure and practice with a graphic system in order to establish a "fluency" in comprehending this "visual language" (Cohn, 2013). Because of this, researchers should question the assumptions they hold about sequential images, and should be more careful in their use of visual narratives as stimuli for experimental tasks.

Understanding sequential images

Only recently has research on sequential image comprehension begun to explore the structure of visual narratives in the context of the cognitive sciences (Cohn, 2013). At their most complex, drawn sequential images—as found in comics—use a visual narrative structure that packages meaningful information into hierarchic constituents (Cohn, Jackendoff, Holcomb, & Kuperberg, 2014). Such structures extend beyond just the meaningful understandings of events between images (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012), and seem to engage similar neural mechanisms as language processing (Cohn et al., 2014; Cohn et al., 2012; Magliano, Larson, Higgs, & Loschky, 2015). However, the foundations for sequential image comprehension are more basic.

Consider the sequence in Figure 1, which appeared as a stimulus in a study by Byram and Garforth (1980), discussed below. If a person recognizes this as a visual narrative sequence, it shows a boy watering plants under the hot sun. He begins to sweat, as indicated by water droplets jumping off his head in the second image. In the final panel, he is therefore shown with shirt off, watering himself instead of the plants. The ability construe the meaning of these successive images requires knowledge of causation both between the character and the sun, and between the contents of each image. Narratively, this sequence shows a set-up in the first image, an initiation of the event in the second image, and a climax in the final panel (Cohn, 2013). Such narrative states are cued by a variety of factors, including the change in position of the character between frames, and the conventionalized sweat drops leaping off the boy's head (Cohn, 2013; Forceville, 2011).

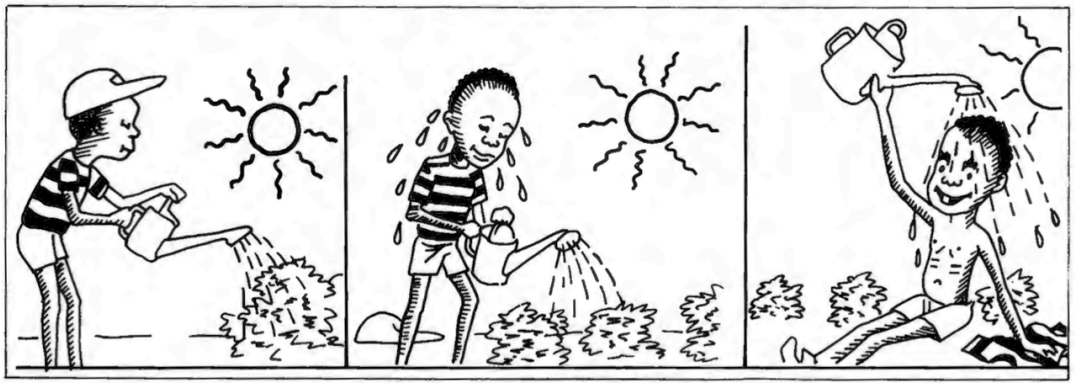


Figure 1: A simple sequence of images used in cross-cultural research in Botswana, from Byram and Garforth (1980).

Despite these narrative and causal relations, at the most simple level, understanding of sequential images requires only two basic constraints (Bornens, 1990; Cohn, In prep). First, a comprehender needs to understand that the characters and elements found in one image are the same referential entities as those in prior and subsequent images. This is the ***continuity constraint***. In Figure 1, this constraint guides our recognition that the character in the first frame is the same as the character in the second and third frames. Without adhering to the continuity constraint, we might interpret each frame as depicting a different scene and thus a different character. In this case, Figure 1 would not show one boy three times, but would show three separate boys.

Second, a comprehender must also recognize that these continuous referential elements take on different temporal, causal, spatial, and/or narrative states in comparison to those in the prior and subsequent images. In Figure 1, this ***activity constraint*** allows us to recognize not only that it is the same character across images, but that this character is in different *related* states in each frame (here, temporal/narrative states). Without this constraint, the images may be recognized as all showing the same referential elements (characters, background, etc.) but would not be distinguished as conveying a *sequence* with linkages between them. In other words, each image would be a scene unto itself. In Figure 1, each image would be its own scene of the same boy, if adhering to the continuity constraint, or of different boys, if flouting the continuity constraint.

The combination of the continuity and activity constraints provide the foundation for all other understanding of sequential images *as a sequence*. These very basic constraints are necessary for comprehenders to further construe more complex aspects of sequential images. For example, higher levels of sequential image comprehension involve the mapping of visual cues in images to narrative roles, the organization of those images into hierarchic constituents, and modification of those sequences using complex framing devices (Cohn, 2013, In prep). As will become clear in the next sections, most limitations in sequential image understanding do not reach these more complex levels. Rather, they demonstrate problems with these simple constraints on continuity and activity. Both high-level and low-level constraints operate on individual image sequences (like strips) and internally to the sequences of longer works (like comic books).

Cross-cultural fluency

The universality of sequential images is first called into question by considering cross-cultural research. In work from the 1960s-1980s, researchers examined wordless sequential images as a means to communicate non-verbally in cross-cultural settings, often with rural, non-literate individuals. Such work often had a practical intent for purposes of educational and/or humanitarian efforts on behalf of these populations, and similar efforts persist in contemporary efforts to use comics as a "universal" communicative tool (e.g., www.comicsunitingnations.org). However, this earlier work found sequential images to be fairly inadequate materials, because various populations did not construe the meaning of these stimuli as expected.

In many cases, respondents had difficulty interpreting the sequential images as a sequence. For example, a study in Nepal found that many respondents did not understand that images repeated the same characters across a 3-image long sequence (Fussell & Haaland, 1978). Similar strain in interpretations were shown by respondents in Papua New Guinea (Bishop, 1977), where some individuals had difficulty construing sequences as conveying temporal orders (Cook, 1980), though this effect was improved with practice and familiarity with Western pictures and comics. An older study in Kenya likewise found that people had trouble recognizing a sequence of images as being in a sequence (Holmes, 1963). Additional work in Africa found that rural Bantu populations (Zulu and Tsonga) in South Africa interpreted sequential images as conveying temporal states of the same character (continuity and activity constraints) less than European counterparts, though this ability increased with age and acculturation (Duncan, Gourlay, & Hudson, 1973). Later work in South Africa found similar interpretations between South Africans and their British counterparts, with effects modulated by exposure to graphics and literacy (Liddell, 1996, 1997). Comparable accounts in Africa were found with the Basotho people (Jenkins, 1978) and in Botswana (Byram & Garforth, 1980).

In these studies, individuals did not seem to be construing sequential images *as a sequence*, and sometimes "misinterpreted" the contents of the individual images themselves (Duncan et al., 1973; Fussell & Haaland, 1978). Most often, these individuals interpreted each image as its own scene, and did not follow the continuity constraint binding each image to the next. It is noteworthy that these individuals typically lived in more rural communities, with little or no exposure to sequential images in the form of comics or picture books. Those who did have access to such materials—along with literacy—had a greater likelihood of making the sequential interpretations. Also, while many of these studies are fairly old and may not reflect the current status of such populations' understandings, the main point remains: Not everyone can comprehend sequences of images.

Some comparable findings have occurred in more recent work on temporal cognition. This research uses a "card sorting task" which asks participants to spatially arrange visual events, with the idea that the layout (vertical, horizontal, circular, etc.) can inform about possible spatial metaphors underlying temporal cognition (Núñez & Cooperrider, 2013). However, results using these tasks in native communities have been mixed, and, again, often depend on participants' familiarity with comics and/or written language (Le Guen & Pool Balam, 2012; Levinson & Majid, 2013). Such limitations have lead to critiques that "in small-scale communities…[these tasks may not be] well-suited…because they presuppose familiarity with materials and practices that, in fact, require considerable cultural scaffolding" (Núñez & Cooperrider, 2013, p. 225).

This necessity for exposure to a graphic system has been echoed in cross-cultural research on *production* of sequential images. For example, Japanese children, who are immersed in a culture of comics, draw far more advanced visual narratives than children from other countries like the United States or Egypt (Wilson & Wilson, 1987). A study in Egypt has made this role the most salient (Wilson, 2016). Suburban Egyptian children, who had exposure to illustrated books and comics, drew narratives with a similar proficiency as children in the United States. This starkly contrasted with the children from rural villages, who had little access to such visual culture, where *only 4-8%* of children drew coherent sequential visual narratives. In these cases, the images in the sequences were isolated objects or events, or had only loose semantic relations. This resembles the inability to establish continuity across sequential images, here in production rather than comprehension. Such findings reinforce that sequential image understanding cannot rely on the architecture of the visual perceptual system alone (Duncan et al., 1973; Fussell & Haaland, 1978).

Development of fluency

Because exposure and practice appear necessary for the understanding of sequential images, it implies the need for learning of this system of communication. We should then ask: what is the developmental trajectory for such comprehension? This question becomes particularly important given the widespread use of visual narratives in tasks for children (e.g., Baron-Cohen et al., 1986; Sivaratnam et al., 2012), particularly those used to ostensibly assess the age of onset of *other* cognitive functions, such as Theory of Mind or IQ.

In fact, developmental studies on sequential image comprehension date back almost 100 years. Research by Piaget and colleagues (Krafft & Piaget, 1925; Margairaz & Piaget, 1925), argued that not until the ages of 7 or 8 did children recognize that characters occurring across images were the same character at different states, rather than numerous different characters (i.e., the continuity constraint).

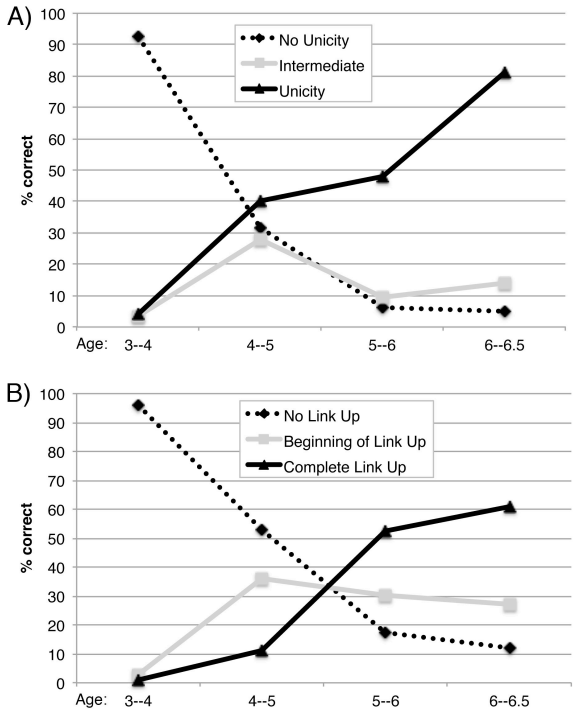


Figure 2: Graphs of data adapted from Bornens (1990). "Unicity" (A) refers to the ability to recognize that characters in one image were the same as those in a prior image (the "continuity constraint"). "Link up" (B) refers to the recognition that characters in one image were in a successive state than those in a prior image (the "activity constraint"). In both cases, these abilities are poor under age 5, but proficient by age 6.

More recent research has suggested the understanding of the continuity constraint in children of much younger ages. Bornens (1990) assessed children's verbalized understandings of several sequential image scenarios for both their ability to grasp "unicity" (i.e., continuity constraint) and "link up" (i.e., activity constraint) of characters across images. Figure 2 adapts the results from this study into graphs (provided numerically in the paper). As depicted in Figure 2A, children at or below the age of 4 had significant difficulty in maintaining continuity of characters between images, but started understanding that

characters across images repeat in different states by age 4 and 5, before progressing to a full understanding at age 6. A similar finding appeared for children's ability to "link up" the actions of the events between images (the continuity constraint), as depicted in Figure 2B. It is also noteworthy that these findings were modulated by children's socio-economic status, with those from less "culturally privileged" environments not reaching proficiency until between ages 5 and 7, perhaps suggesting a difference in exposure to visual material, as in the findings of visual narrative "fluency" discussed above.

The ages of onset for sequential image comprehension in this study are consistent with other findings. For example, arranging images into a logically and/or temporally ordered sequence is difficult for 2 and 3 year olds (Weist, Atanassova, Wysocka, & Pawlak, 1999; Weist, Lyytinen, Wysocka, & Atanassova, 1997), but can be done proficiently by 4 or 5 year olds (Fivush & Mandler, 1985; Friedman, 1990; Weist et al., 1999; Weist et al., 1997). Children by around the age of 5 can also reconstruct logically ordered sequences better than random image sequences (Brown, 1975). Additional work has shown that the ability to recognize and infer missing images from a visual narrative begins around age 5 (Brown & French, 1976; Schmidt & Paris, 1978; Schmidt, Paris, & Stober, 1979), though this ability is modulated by the degree to which sequences maintain a continuity of common characters across images (Kunen, Chabaud, & Dean, 1987). Both picture arrangement and the ability to infer omitted information appear to continue developing into later years, though age alone does not determine proficiency: it is modulated by experience with comics (Nakazawa, 2016).

Such findings overall suggest that the continuity and activity constraints come online between the ages of 4 and 5. However, given the findings in cross-cultural research, such ages likely reflect age of onset when children have cultural exposure to sequential images.

Experience in fluent populations

The studies discussed above imply that the understanding of sequential images requires exposure to visual narratives, and in such context, follows a consistent developmental trajectory. What about the understanding of sequential images by adults in cultures with rich visual narratives? Sequential images should pose no problems to experimental participants so long as they are healthy, college-aged adults from visually rich cultures, right?

First, there are no guarantees that participants have acquired the requisite "fluency" to gain proficiency in comprehending sequential images, even if they belong to a broader culture that has rich visual narratives. Without assessment, we cannot simply assume that everyone in a given culture will behave uniformly. While no studies have yet to explicitly study the range of such fluencies, anecdotes abound from individuals who claim that they "can't understand comics."

Second, more empirically grounded data has shown that even when the basic ability to understand sequential images seems intact (i.e., the continuity and activity constraints), and visual narratives take more complex characteristics, various aspects of comprehension are modulated by frequency of comic reading. For example, Nakazawa (2016) has designed the Chiba University Comics Comprehension Test (CCCT), which involves a battery of tests including picture arrangement and fill-in-the-blank tasks. The CCCT has been used to assess the comprehension of sequential images in both children and adults. In general, Nakazawa (2016) has found that the ability to understand comics is modulated by age and expertise, with college students (who most frequently read comics) having greater proficiency than both younger children/teenagers and older adults. However, Japanese college students were shown to be more proficient in their visual narrative comprehension than American college students (Nakazawa & Shwalb, 2012). This difference was attributed to the relative pervasiveness of manga readership throughout Japanese society, in contrast to the more niche subculture of comics readership in the United States.

In addition, recent research has included a measure of "comic reading experience" in both behavioral and neurocognitive testing of visual narrative processing. The Visual Language Fluency Index (VLFI) questionnaire assesses the frequency with which individuals read and draw a variety of types of visual narratives (comic books, comic strips, graphic novels, Japanese manga), and their self-rated "expertise" in that understanding. This information is combined in the following formula to compute a "VLFI score", weighted more to comprehension (comic reading) than production (comic drawing):

$$\left(\begin{matrix}Mean\ Comic\ Reading\ Freq.\\ x\ Comic\ reading\ expertise\end{matrix}\right) + \left(\frac{Comic\ Drawing\ Freq.\ x\ Drawing\ Ability}{2}\right)$$

This metric has been shown to correlate with a range of manipulations to sequential images—particularly more complex aspects of narrative structure—in measurements including response times, viewing times, and the amplitudes of neural responses in event-related brain potentials (Cohn & Kutas, 2015; Cohn & Maher, 2015; Cohn et al., 2012). Resources related to the VLFI can be found at: http://www.visuallanguagelab.com/resources.html.

Such findings suggest that, even with proficiency in the basic understandings of sequential images with healthy, college-aged adults, experience modulates the processing of visual narratives. Thus, if sequential images are used as an elicitation for probing *other* cognitive abilities, there is no guarantee that participants will be uniform in their comprehension of those stimuli without some form of assessment.

Conclusion

Altogether, the studies in this review suggest that sequential image understanding is a culturally learned ability contingent on exposure and practice with visual narratives,

such as comics and illustrated picture books. Such findings distinctly contrast with the general assumptions that sequential images are transparently understood, universal across cultures, and require no explicit learning (e.g., McCloud, 1993). Given that many experimental procedures have used sequential images as stimuli, researchers should be sensitive to these issues when designing studies and tasks. Such findings also call into question the degree to which results found in prior experiments are reflective of the actual cognition that is being tested (theory of mind, time-space metaphors, IQ, etc.) or if they reflect the ability of participants to understand sequential images at a basic level. For example, do picture arrangement tasks really inform about space-time metaphors (Núñez & Cooperrider, 2013) or IQ (Kaufman & Lichtenberger, 2006; Ramos & Die, 1986), or are these confounded by visual narrative fluency?

In many cases, drawn sequential images may not be necessary, as technological advancement makes the use of videos far easier for tasks requiring elicitation. While more complicated aspects of visual narrative patterning do appear difficult for naïve film viewers to understand, lower level issues like the continuity or activity constraints appear to be spared (Ildirar & Schwan, 2015; Schwan & Ildirar, 2010). This may be due to filmic use of basic percepts (instead of drawings) and their pervasive temporality and movement, which requires less decoding than drawn conventions.

Nevertheless, this does not mean that experimentation should avoid sequential images completely. If sequential images are used as stimuli, researchers are encouraged to assess the "fluency" of their participants in this "visual language" (i.e., using the VLFI or CCCT), just as they would be expected to assess participants' proficiency in other expressive systems, like language.

Given that research on sequential images has only recently progressed with seriousness in the cognitive sciences, future work should examine additional behavioral "fluency" assessments beyond the VLFI or CCCT. It is also unclear whether fluency varies based on the specific tasks involved in sequential images (e.g., picture arrangement vs. elicitation). In addition, studies will need to investigate just how general cognitive processes (memory, perception, attention) interact with visual narrative specific aspects of sequential image comprehension and fluency. Together, such efforts will further inform our understandings of sequential image comprehension, and the degree to which they may be used in other experimental tasks.

References

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. *British Journal of Developmental Psychology, 4*(2), 113-125.

Bishop, A. (1977). Is a Picture Worth a Thousand Words? *Mathematics Teaching, 81*, 32-35.

Bornens, M.-T. (1990). Problems brought about by "reading" a sequence of pictures. *Journal of Experimental Child Psychology, 49*(2), 189-226. doi:http://dx.doi.org/10.1016/0022-0965(90)90055-D

Brown, A. L. (1975). Recognition, Reconstruction, and Recall of Narrative Sequences by Preoperational Children. *Child Development, 46*(1), 156-166. doi:10.2307/1128844

Brown, A. L., & French, L. A. (1976). Construction and Regeneration of Logical Sequences Using Causes or Consequences as the Point of Departure. *Child Development, 47*(4), 930-940. doi:10.2307/1128428

Byram, M. L., & Garforth, C. (1980). Research and testing non-formal education materials: a multi-media extension project in Botswana. *Educational Broadcasting International, 13*(4), 190-194.

Campbell, J. M., & McCord, D. M. (1996). The WAIS-R Comprehension and Picture Arrangement subtests as measures of social intelligence: Testing traditional interpretations. *Journal of Psychoeducational Assessment, 14*, 240-249.

Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. London, UK: Bloomsbury.

Cohn, N. (In prep). What can you draw without narrative? A hierarchy of grammatical complexity for sequential images. *Cognitive Science*.

Cohn, N., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2014). The grammar of visual narrative: Neural evidence for constituent structure in sequential image comprehension. *Neuropsychologia, 64*, 63-70. doi:10.1016/j.neuropsychologia.2014.09.018

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia, 77*, 267-278. doi:10.1016/j.neuropsychologia.2015.08.026

Cohn, N., & Maher, S. (2015). The notion of the motion: The neurocognition of motion lines in visual narratives. *Brain Research, 1601*, 73-84. doi:10.1016/j.brainres.2015.01.018

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology, 65*(1), 1-38. doi:10.1016/j.cogpsych.2012.01.003

Cook, B. L. (1980). Picture communication in the Papua New Guinea. *Educational Broadcasting International, 13*(2), 78-83.

Duncan, H. F., Gourlay, N., & Hudson, W. (1973). *A Study of Pictorial Perception Among Bantu and White School Children*. Johannesburg: Witwaterstrand University Press.

Fivush, R., & Mandler, J. M. (1985). Developmental Changes in the Understanding of Temporal Sequence. *Child Development, 56*(6), 1437-1446. doi:10.2307/1130463

Forceville, C. (2011). Pictorial runes in *Tintin and the Picaros*. *Journal of Pragmatics, 43*, 875-890.

Friedman, W. J. (1990). Children's Representations of the Pattern of Daily Activities. *Child Development, 61*(5), 1399-1412. doi:10.1111/j.1467-8624.1990.tb02870.x

Fussell, D., & Haaland, A. (1978). Communicating with Pictures in Nepal: Results of Practical Study Used in Visual Education. *Educational Broadcasting International, 11*(1), 25-31.

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology, 17*, 324-363.

Holmes, A. C. (1963). *A study of understanding of visual symbols in Kenya*. London: Oversea Visual Aids Centre.

Ildirar, S., & Schwan, S. (2015). First-time viewers' comprehension of films: Bridging shot transitions. *British Journal of Psychology, 106*(1), 133-151. doi:10.1111/bjop.12069

Jenkins, J. (1978). Using Pictures in Non-Formal Education. *Educational Broadcasting International, 11*(1), 32-38.

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing Adolescent and Adult Intelligence* (3rd ed.). Hoboken: Wiley.

Krafft, H., & Piaget, J. (1925). La notion d'ordre des événements et le test des images en desordre chez I'enfant de 6 a 10 ans [The concept of order of events and muddled picture test with 6-10 year olds]. *Archives de Psychologie, 19*, 306-349.

Kunen, S., Chabaud, S. A., & Dean, A. L. (1987). Figural factors and the development of pictorial inferences. *Journal of Experimental Child Psychology, 44*(2), 157-169. doi:http://dx.doi.org/10.1016/0022-0965(87)90028-2

Le Guen, O., & Pool Balam, L. I. (2012). No metaphorical timeline in gesture and cognition among Yucatec Mayas. *Frontiers in Psychology, 3*. doi:10.3389/fpsyg.2012.00271

Levinson, S. C., & Majid, A. (2013). The island of time: Yélî Dnye, the language of Rossel Island. *Frontiers in Psychology, 4*. doi:10.3389/fpsyg.2013.00061

Liddell, C. (1996). Every picture tells a story: South African and British children interpreting pictures. *British Journal of Developmental Psychology, 14*(3), 355-363. doi:10.1111/j.2044-835X.1996.tb00711.x

Liddell, C. (1997). Every Picture Tells a Story—Or does it?: Young South African Children Interpreting Pictures. *Journal of Cross-Cultural Psychology, 28*(3), 266-283.

Magliano, J. P., Larson, A. M., Higgs, K., & Loschky, L. C. (2015). The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. *Memory & Cognition*, 1-13. doi:10.3758/s13421-015-0558-7

Margairaz, E., & Piaget, J. (1925). La structure des réits et I'interpretation des images de Dawid chez I'enfant [The structure of children's stories and interpretation of pictures of Dawid]. . *Archives de Psychologie, 19*, 211-239.

McCloud, S. (1993). *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.

Nakazawa, J. (2016). Manga literacy and manga comprehension in Japanese children. In N. Cohn (Ed.), *The Visual Narrative Reader* (pp. 157-184). London: Bloomsbury.

Nakazawa, J., & Shwalb, D. W. (2012). *Japan and the US comparison of university students' Manga reading literacy.* Paper presented at the Proceedings of Annual Conference of 54th Japanese Association of Educational Psychology.

Núñez, R., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences, 17*(5), 220-229. doi:http://dx.doi.org/10.1016/j.tics.2013.03.008

Ramos, M. C., & Die, A. H. (1986). The Wais-R Picture Arrangement Subtest: What do Scores Indicate? *The Journal of General Psychology, 113*(3), 251-261. doi:10.1080/00221309.1986.9711036

Schmidt, C. R., & Paris, S. G. (1978). Operativity and Reversibility in Children's Understanding of Pictorial Sequences. *Child Development, 49*(4), 1219-1222. doi:10.2307/1128764

Schmidt, C. R., Paris, S. G., & Stober, S. (1979). Inferential distance and children's memory for pictorial sequences. *Developmental Psychology, 15*(4), 395-405.

Schwan, S., & Ildirar, S. (2010). Watching Film for the First Time: How Adult Viewers Interpret Perceptual Discontinuities in Film. *Psychological Science, 21*(7), 970-976. doi:10.1177/0956797610372632

Sivaratnam, C. S., Cornish, K., Gray, K. M., Howlin, P., & Rinehart, N. J. (2012). Brief Report: Assessment of the Social-Emotional Profile in Children with Autism Spectrum Disorders using a Novel Comic Strip Task. *Journal of Autism and Developmental Disorders, 42*(11), 2505-2512. doi:10.1007/s10803-012-1498-8

Tinaz, S., Schendan, H. E., Schon, K., & Stern, C. E. (2006). Evidence for the importance of basal ganglia output nuclei in semantic event sequencing: An fMRI study. *Brain Research, 1067*(1), 239-249. doi:http://dx.doi.org/10.1016/j.brainres.2005.10.057

Weist, R. M., Atanassova, M., Wysocka, H., & Pawlak, A. (1999). Spatial and temporal systems in child language and thought: a cross-linguistic study. *First Language, 19*(57), 267-308. doi:10.1177/014272379901905701

Weist, R. M., Lyytinen, P., Wysocka, J., & Atanassova, M. (1997). The interaction of language and thought in children's language acquisition: a crosslinguistic study. *Journal of Child Language, 24*(01), 81-121. doi:doi:null

Wilson, B. (2016). What happened and what happened next: Kids' visual narratives across cultures. In N. Cohn (Ed.), *The Visual Narrative Reader* (pp. 185-227). London: Bloomsbury.

Wilson, B., & Wilson, M. (1987). Pictorial Composition and Narrative Structure: Themes and Creation of Meaning in the Drawings of Egyptian and Japanese Children. *Visual Arts Research, 13*(2), 10-21.

Deciding to Remember: Memory Maintenance as a Markov Decision Process

Jordan W. Suchow (suchow@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, USA

Abstract

Working memory is a limited-capacity form of human memory that actively holds information in mind. Which memories ought to be maintained? We approach this question by showing an equivalence between active maintenance in working memory and a Markov decision process in which, at each moment, a cognitive control mechanism selects a memory as the target of maintenance. The challenge of remembering is then finding a maintenance policy well-suited to the task at hand. We compute the optimal policy under various conditions and define plausible cognitive mechanisms that can approximate these optimal policies. Framing the problem of maintenance in this way makes it possible to capture in a single model many of the essential behavioral phenomena of memory maintenance, including directed forgetting and self-directed remembering. Finally, we consider the case of imperfect metamemory — where the current state of memory is only partially observable — and show that the fidelity of metamemory determines the effectiveness of maintenance.

Keywords: memory maintenance, Markov Decision Process, cognitive control, working memory

Introduction

Working memory is a storage system that actively holds information in mind and allows for its manipulation, providing a workspace for thought (Baddeley, 1992). Its capacity is strikingly limited, perhaps to only a few sights or sounds. Using working memory is effortful: pupils dilate, skin conductance rises, and secondary tasks become harder to perform well (Kahneman, 1973). Much of the research on working memory has focused on characterizing its limits and determining what gives rise to them. For example, working memory capacity is known to be lower in young children and the elderly (Dobbs & Rule, 1989), correlates strongly with a person's fluid intelligence (Conway et al., 2003), is affected by sleep schedule (Steenari et al., 2003), and can be impaired in people with mental disorders such as schizophrenia (Goldman-Rakic, 1994). From this work, we have learned a considerable amount about how much can be remembered and who is best at remembering it.

Information held in working memory is malleable (Jonides et al., 2008). It can, for example, be remembered and forgotten intentionally through the processes of directed forgetting and directed remembering, which prioritize some experiences over others for later access (Muther, 1965; Bjork et al., 1968). These directed maintenance mechanisms are closely related to cognitive control and to the top-down processes that determine our conscious thoughts from moment to moment (Macrae et al., 1997). At times, these control processes can backfire, causing unwanted thoughts and memories to linger despite our best intentions (Wegner, 2009).

Given the flexibility available to the working memory system, a question naturally arises: What is the optimal way to maintain memories? What is the space of possible maintenance strategies, and how successful is each of them in retaining information over short durations?

We approach this question by likening working memory maintenance to a sequential decision process in which, at each moment, a cognitive control mechanism decides which memories to prioritize. We focus on a particular kind of sequential decision process known as the *Markov decision process* (MDP) (Puterman, 1994), which provides an abstract mathematical framework for describing decision-making in a setting that is partly under control of the decision-maker (here, the maintenance process) and partly under control of the environment (here, the degradation process). Besides being well suited to describing the problem of memory maintenance, the MDP has the added benefit of being one of the most well-understood models in the mathematics and psychology of reinforcement learning. Thus, having established the connection, existing concepts and tools from reinforcement learning can be brought to bear on the dynamics of memory maintenance.

The plan of the paper is as follows. First we describe the essential behavioral phenomena of memory maintenance and control. Then we formulate the problem of memory maintenance as an MDP. The next section describes the form of solutions to the maintenance problem — a maintenance policy — and proceeds by computing the optimal policy under various reward functions. Next, we show how the optimal policy, and cognitively plausible approximations thereof, can reproduce the behavioral phenomena described earlier. Before concluding, we extend our framework to the case of imperfect metamemory, describing memory maintenance in a partially observable mind — i.e., in situations where the maintenance system has incomplete or uncertain information about the current status of actively-held memories.

Memory maintenance and control

The essential behavioral phenomena of active memory maintenance and control involve monitoring, prioritizing, and controlling memories.

Directed remembering

Memories can be forgotten intentionally. In experiments on this process of "directed forgetting", participants study some information and are then directed to remember or forget specific elements of what was studied (Muther, 1965;

Bjork et al., 1968). Memory tends to be better for the to-be-remembered information than for the to-be-forgotten information.

For example, in Woodward & Bjork (1971), participants studied a list of words and were later asked to recall as many of them as possible. This is the popular *free recall* paradigm used extensively in studies of long-term memory. Following each word's presentation, a cue appeared instructing the participants to remember or to forget the word. Later, participants were asked to recall all the words from the studied list, regardless of how those words initially had been marked. The recall task was challenging. Critically, its difficulty depended on how the word had been marked: words marked as to-be-remembered were recalled 23.3% of the time, whereas those marked as to-be-forgotten were recalled only 4.7% of the time. This is the hallmark of directed forgetting, which has been demonstrated in both long- and short-term memory (Woodward & Bjork, 1971; Bjork et al., 1968).

Williams et al. (2012) demonstrated directed remembering in visual working memory. In the experiment, participants held in mind the colors of one or two colorful squares. On trials when two objects were presented, a cue would sometimes appear 1 s into the retention interval, informing the participant of which object would be tested a few seconds later. This hint afforded participants the opportunity to alter their maintenance behavior accordingly. The probability of remembering the tested object was highest when there was only one object, lower when there were two but one of them was cued, and even lower when there were two but none was cued. In a second experiment, Williams et al. showed that the benefit afforded to the cued object comes at a cost: the non-cued object is almost entirely forgotten. This is a strong demonstration of flexible redirection of memory maintenance and its effects.

Monitoring

Monitoring comes in the form of *metamemory*, an awareness of one's memories and the systems that store them. Metamemory is often studied in the context of long-term memory, where it is invoked to explain phenomena such as tip-of-the-tongue states and the feeling of knowing (Wellman, 1977). Healthy individuals have a rich set of metamemory skills that guide learning, decision making, and action (Metcalfe & Shimamura, 1994). Neurological diseases, such as Alzheimer's and Korsakoff's syndrome, adversely affect metamemory judgments, causing a mismatch between what is remembered and what is believed to be remembered (Pannu & Kaszniak, 2005).

Self-directed remembering

Maintenance can also be directed by internally generated signals — "self-directed remembering" Suchow (2014). In one experiment, instead of redirecting maintenance to an externally cued objects, participants were given a cue to redirect maintenance to the best- or worst-remembered object. The fidelity of memory was better after maintenance had been redirecting than in a baseline where it was not (Suchow, 2014).

The Markov Decision Process

A Markov decision process is defined by a state space, a set of possible actions, a transition model, and a reward function. Each is defined in turn below:

State space We suppose that there is a memory-supporting commodity, akin to attention, that can be divided into *quanta*, each of which is assigned to a particular memory. The quanta might, for example, represent cycles of a time-based refreshing process (Vergauwe et al., 2009) or neural populations in prefrontal cortex that represent "token" encodings of visual events (Bowman & Wyble, 2007). The more of the commodity assigned to a memory, the stronger and more robust it is. The state of working memory is then an allotment of the quanta to each memory, which may receive the entire commodity, only a portion of it, or perhaps none at all. The state space thus forms a $(K-1)$ regular discrete simplex, where K is the number of memories held in working memory and where the discretization is determined by the number of quanta N.

Actions At each time step, the maintenance process selects a quantum as the recipient of maintenance. Thus the set of possible actions A is of size N, one action per quantum, and does not depend on the state.

Transition model The transition model specifies the probability of moving from one state of memory to another and is thus a formal model of memory degradation. We will make use of the transition model proposed in Suchow (2014) — i.e., a Moran process, a model of evolution in finite populations that originated in population genetics (Moran, 1958) and which has been used to describe dynamic processes in diverse settings. Under the Moran process, at each time step a quantum degrades because another quantum interferes with it or replaces it. The degraded quantum is chosen randomly, uniformly across all the quanta. The interfering (or replacing) quantum is determined by the action chosen by the maintenance process. We can write the state as an allotment of quanta to memories, $s = [n_1, n_2, ..., n_K]$, summing to N, the number of quanta. At each time step, one of the n's is incremented and one is decremented. The incremented n is determined by the chosen action — if the chosen action maintains a quantum belonging to that memory, it is deterministically incremented. The decremented n is chosen with probability proportional to n because the quanta are all equally likely to degrade. This defines a transition model $P(s' \mid s, a)$, which gives the probability of landing in state s' given that the agent took action a while in state s.

Reward function By definition, the agent's goal is to maximize the total reward that is received. The reward function is a mapping from states to an amount of reward that is received for landing in that state. In the case of most working memory tasks, which are episodic (in the sense that information arrives all at once and is then discarded at the end of the trial), and which have a retention interval that is known

to the participant, the reward function is time-varying, taking on a value of zero everywhere until the moment of the test, at which point it becomes positive for some states and (possibly) zero for others. For simplicity, we assume that the retention interval is chosen in such a way (e.g., from an exponential distribution) that the reward function is stationary. The specifics of the reward function inevitably depend on the demands of the task and are usually implicit in the experiment's design and feedback mechanism. For example, tasks using the "continuous partial report paradigm" require participants to hold information in mind for a fixed duration, e.g., 2000 ms, with reward provided in proportion to the similarity between the participant's response and the true value. Other tasks provide all-or-none feedback.

We will consider three reward functions relevant to the goals of a memory maintenance system. The first applies to tasks with an all-or-none design in which the memorizer receives full credit for having remembered enough about the cued memory to access it (i.e., having at least k quanta assigned to it at the time of the test, where k is the strength of the weakest accessible memory) and otherwise receives no reward. This reward function is appropriate when scoring performance using a high-threshold model, considering only the probability of remembering while ignoring accuracy. In the second, the memorizer is rewarded for having at least one sufficiently strong memory (i.e., one with greater than some threshold number of quanta), but where remembering something about everything is unnecessary. In the third, there is an imbalance across memories in the reward given for remembering them: some are more valuable than others.

Maintenance policies

The Markov decision process is a general framework for describing the problem of sequential decision making, but it does not specify the particular strategy used by the agent to make a decision. That strategy is defined by a policy, a function that specifies an action (or probability distribution over actions) for each possible state. Much of modern research on MDPs focuses on finding the optimal policy, one that maximizes the (possibly time-discounted) reward.

The simplest maintenance policies do not depend on the current state of memory. Rather, they produce the same behavior in every state. Borrowing terminology from game theory, we call these maintenance policies *unconditional*. An example of an unconditional maintenance policy is `all-i`, which always selects the *i*th quantum as the target of maintenance. A second unconditional strategy is `random`, which selects a target at random, uniformly over all quanta — this maintenance policy is equivalent to a neutral Moran process.

Conditional policies, in contrast, depend on the state. In the context of memory maintenance, consider for example the strategy `all-j`, which selects a quantum uniformly from among those assigned to memory j if one exists, otherwise choosing randomly among all the quanta.

The optimal policy is conditional. Using dynamic programming, we computed the optimal policy for a time-discounted variant of the above MDP under each of the reward functions described above, setting $N = 10$, $K = 3$, and the discount factor to 0.99. The optimal policy is different under each reward function, reflecting the differing demands of the task. When the reward function encourages having at least one highly-stable memory, the optimal policy tends to maintain memories that are already stable, preferring to select a quantum assigned to a memory with an above-median allocation of quanta 64% of the time. In contrast, when the reward function encourages good performance on the task, which requires storing more than just one memory, the optimal policy tends to maintain memories that are least stable, preferring to select a quantum assigned to a memory with an above-median allocation of quanta only 29% of the time. When the reward function encourages prioritization of a particular memory, the optimal policy deterministically maintains that memory so long as it has not fully degraded, in which case it chooses randomly among the others — this is the `all-j` maintenance policy described above. At a minimum, then, any cognitive implementation of memory maintenance must be able to selectively maintain memories according to their strength and according to their identity.

The optimal policy can be approximated by a simple strategy that rests on plausible cognitive mechanisms, inspired by a psychological principle known as Luce's choice axiom (Luce, 1959; Herrnstein, 1961). According to the axiom, when faced with a choice among alternatives, a decision-maker will exhibit 'matching behavior', selecting options with probability proportional to their value. Matching behavior was originally studied in the context of learning theory, where value is defined as the expected reward (Estes, 1957; Sutton & Barto, 1998). Thus if two levers offer rewards in a ratio of 2:1, an individual that displays matching behavior will press the more rewarding lever twice as often. Here, value is akin to memory strength and is defined by the number of quanta dedicated to a memory.

In practice, it is common to consider a generalization of matching behavior in which a real-valued parameter L determines the decision-maker's sensitivity to the signal. In this so-called "softmax" generalization of matching behavior, the probability of selecting option a from the set of alternatives A is given by

$$P(a) = \frac{v(a)^L}{\sum_{b \in A} v(b)^L}, \tag{1}$$

where $v(x)$ is the strength of the signal generated by x and where L determines the decision maker's sensitivity to the signal (Sutton & Barto, 1998).

Five values of L are particularly significant. When $L = 0$, the process is unconditional (i.e., insensitive to the signal). This corresponds to a neutral process. When $L = 1$, the process gives preference to objects in proportion to how strongly they are currently represented. When $L \to \infty$, the winner takes all. In contrast, when $L = -1$, the process gives preference to objects in proportion to how *weakly* they are currently repre-

sented, and in the limit $L \to -\infty$, the loser takes all.

Equation 1 defines the `Luce` family of maintenance policies that will be examined empirically in the following sections.

Reproducing the behavioral phenomena

The benefit of directed remembering

We simulated performance of a memorizer who uses the `Luce` family of maintenance policies in the directed-remembering task from Exp. 1 of Williams et al. (2012). Unlike the `random` policy, which is unable to direct maintenance to the cued object, policies in the `Luce` family can (Fig. 1).

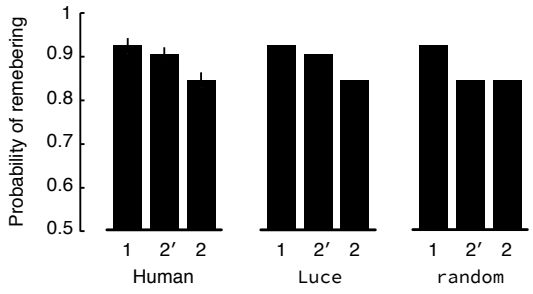


Figure 1: Reproducing Experiment 1 from Williams et al. (2012), with the `Luce` policy. In condition 1, participants remember 1 object. In condition 2, they remember 2. In condition 2′, they intially remember 2 and then direct maintenance to the one that is cued.

The cost of directed remembering

Next, we simulated performance of a memorizer who uses the `Luce` family of maintenance policies in the directed-remembering task from Exp. 2 of Williams et al. (2012). The `random` policy remains unable to direct maintenance to the cued object — thus, in comparison to human performance, the cued object is remembered too poorly, and the non-cued object is remembered too well. In constrast, policies in the `Luce` family demonstrate both effects (Fig. 2), directed maintenance to the cued object and drawing it away from the non-cued object.

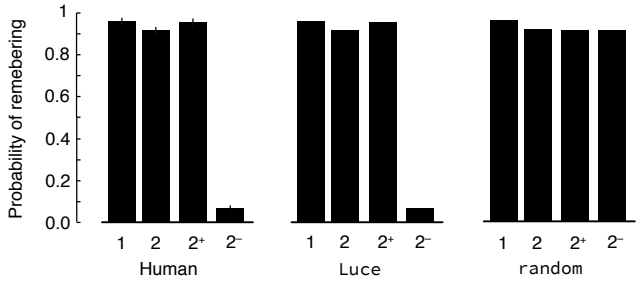


Figure 2: Reproducing Williams et al. (2012), Exp. 2, with the `Luce` policy. Objects that do not receive the benefit of preferential maintenance are rapidly lost.

Self-directed remembering

Lastly, we simulated performance of a memorizer who uses the `Luce` family of maintenance policies in the self-directed remembering task of Suchow et al. (2014). The `random` policy cannot direct maintenance. In constrast, policies in the `Luce` family can redirect maintenance to to best- or worst-remembered object through an appropriate choice of L (Fig. 3).

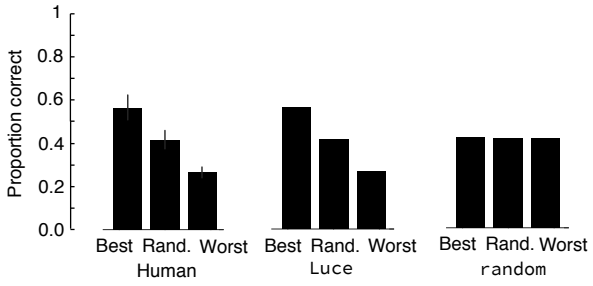


Figure 3: Reproducing self-directed remembering from Suchow et al. (2014), with the `Luce` policy. Maintenance can be redirected based on a metamemory signal.

Predicting new phenomena

Graded directed remembering

Given the apparent flexibility of directed remembering, it may be possible to give graded preference to some objects over others. There is strong evidence that such graded preferences are possible during encoding. The `Luce` family of policies can be extended to give graded preference to certain memories over others. To do this, we first define a priority function f that assigns a score to each memory. For example, memories A, B, and C may receive scores of 4, 3, and 1, meaning that A has $4\times$ the priority of C and B has $3\times$ the priority of C. Quanta are selected with probability proportional to the priority score of the memory to which it is assigned. For a system with N quanta, of which n_{A} are assigned to memory A, n_{B} to B, and n_{C} to C, the probability of selecting a quantum q that is of type j is given by

$$P(q) = \frac{f(j)}{\sum_{j \in \{\mathrm{A,B,C}\}} f(j) n_j}. \tag{2}$$

This is equivalent to adding selective pressures to the neutral process and allows for prioritization and graded directed-remembering.

Partially observable minds

The framework of a Markov decision process makes a strong commitment to the accessibility of the memory state to the memory maintenance system: it assumes perfect, real-time, no-cost metamemory. However, metamemory is imperfect (Flavell & Wellman, 1977).

By generalizing the MDP to a partially-observable world, we can accommodate situations of imperfect or costly

metamemory. A partially observable world is one in which the agent does not know exactly what state it is in, making it impossible to directly carry out conditional policies that depend on the state. Often the agent has available some instrument (a "sensor") for measuring or sensing the state. In the case of memory maintenance, the sensor is metamemory. The agent uses the sensor to update its beliefs about the state. Thus the *partially observable Markov decision process* (POMDP) extends the MDP through the introduction of a sensor model, which describes the information about the state that is provided by each observation, and a belief state, which is a probability distribution over the state space that embodies the agent's beliefs about the current state (Monahan, 1982). The Dirichlet distribution is a convenient representation of uncertainty about the state of memory resource allocation because it is the conjugate prior for multinomial data.

In a partially observable mind, inefficiencies of metamemory limit the efficacy of flexible maintenance behaviors. This is because in a world where the future depends on the past, one who does not even know the present cannot suitably plan for what is to come. We demonstrate this dependence by defining a simple metamemory agent and then simulating its behavior with different levels of efficiency. Metamemory observations made by the agent come in the form of object labels sampled with probability proportional to their strength (that is, the number of quanta assigned to them). This defines the sensor model. The agent is initially unaware of the allocation of the commodity, represented by a belief state initially set to a Dirichlet distribution with concentration parameters 1, 1, and 1, which is equivalent to a uniform distribution over all possible allocations. At each time step, the agent makes m observations. We assume that the metamemory system has no memory of its own and thus considers only the observations made at the current time step (see below for a brief discussion of optimal filtering, in which the metamemory system also considers past observations). To avoid the problems caused by sampling zero quanta of a certain type, we use additive smoothing by adding one to all the counts. These counts are used by the `Luce` policy, with exponent 1. The efficiency of metamemory can be varied by altering the number of observations made at each time step. This formulation makes it possible to vary efficiency between two extremes. At one extreme, $m = 0$ and the agent gains no information about the state. At the other extreme, in the limit $m \to \infty$, the agent has perfect information about the state. Intermediate efficiencies lead to intermediate performance (Fig. 4).

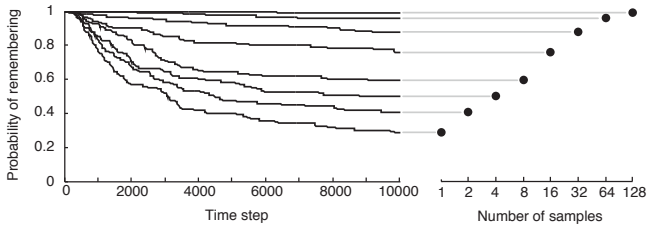


Figure 4: Inefficiencies of metamemory limit the efficiency of memory maintenance. On the left are forgetting functions for a simulated agent whose memory is only partially observable. At each time step the agent draws m quanta (with replacement) and observes their assignment. Selection happens according to the procedure in the main text. On the right, performance increases with the number of samples taken. Simulations were run with settings $N = 128$, $K = 12$, and $L = -1$.

Discussion

In this paper, we approached the problem of memory maintenance by demonstrating an equivalence to a Markov decision process in which, at each moment, a cognitive control mechanism selects a memory as the target of maintenance. The challenge of remembering is then finding a maintenance policy well-suited to the task at hand. We computed the optimal policy under various conditions and defined plausible cognitive mechanisms, embodied by the `Luce` policy, that can approximate these optimal policies. Framing the problem of maintenance in this way makes it possible to capture in a single model many of the essential behavioral phenomena of memory maintenance, including directed remembering, priority-based directed remembering, and self-directed remembering. Finally, we considered the case of imperfect metamemory — where the current state of memory is only partially observable — and show that the fidelity of metamemory determines the effectiveness of maintenance.

Perhaps the biggest payoff that comes from framing the problem of memory maintenance in this way is the set of new questions that it makes possible to ask.

For example, one might ask where maintenance policies come from. Specifically, how are they learned? Methods such as temporal difference learning have emerged as candidate learning mechanisms used in the brain to learn policies that guide behavior, and it has become popular to relate this particular class of learning algorithms to known reward circuitry in the brain (O'Doherty et al., 2003). Particularly relevant is Todd et al. (2008), which discusses learning to use working memory by temporal difference methods. Specifically, temporal difference learning can be used to shape representations in the prefrontal cortex so that they are useful for working memory (Todd et al., 2008). Also relevant is O'Reilly & Frank (2006), which develops an "actor/critic" model of the neural substrates of working memory and cognitive control, showing that an active gating mechanism that controls the contents of working memory can be learned through learning mechanisms from reinforcement learning (O'Reilly & Frank, 2006).

Finally, it may be useful to consider other resource allocation tasks that are similar in structure to that of memory maintenance — e.g., scheduling and queuing. Much of the original work on these problems came from the field of operations research, which originated from military planners in WWII and which today considers the optimal solutions to decision

making and resource allocation tasks in a variety of settings, often in the context of organizational behavior (Taha, 2007) or electronic systems (Åström & Wittenmark, 2011). Having made the link to these related problems, it may be fruitful to consider known solutions as candidate psychological mechanisms. For example, queuing theory is a set of tools for considering resource allocation tasks that feature the continuous arrival of entities that require the resource (e.g., callers to a company's customer support center) (Kleinrock, 1975). Most of the popular working memory tasks are episodic, with information arriving all at once and then being discarded at the end of the trial. Our visual experience is not always so episodic; rather, it is sometimes necessary to update the contents of working memory with new information or redirecting maintenance when the goals change (Matthey et al., 2012). Looking towards queuing theory, for example, may provide insight into this problem of maintenance in the face of continuously-arriving information.

Acknowledgments

This work was partially supported by the National Science Foundation (grant SPRF-IBSS-1408652 to T.L.G and J.W.S).

References

Åström, K. J., & Wittenmark, B. (2011). *Computer-controlled systems: Theory and design*. Courier Dover Publications.

Baddeley, A. (1992). Working memory. *Science*, *255*, 556-559.

Bjork, R. A., Laberge, D., & Legrand, R. (1968). The modification of short-term memory through instructions to forget. *Psychonomic Science*, *10*, 55-56.

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, *114*(1), 38–70.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552.

Dobbs, A. R., & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500-503.

Estes, W. K. (1957). Of models and men. *American Psychologist*, *12*(10), 609-617.

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (p. 3-33). Erlbaum.

Goldman-Rakic, P. S. (1994). Working memory dysfunction in schizophrenia. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *6*, 348-357.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*(3), 267272.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193–224.

Kahneman, D. (1973). *Attention and effort*. Prentice Hall.

Kleinrock, L. (1975). *Queueing systems. Volume 1: Theory*. Wiley.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Ford, R. L. (1997). On regulation of recollection: The intentional forgetting of stereotypical memories. *Journal of Personality and Social Psychology*, *72*(4), 709-719.

Matthey, L., Bays, P., & Dayan, P. (2012). Probabilistic palimpsest memory: Multiplicity, binding and coverage in visual short-term memory. In COSYNE (p. III-48).

Metcalfe, J. E., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing.* The MIT Press.

Monahan, G. E. (1982). State of the art—a survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, *28*(1), 1–16.

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60-71.

Muther, W. S. (1965). Erasure or partitioning in short-term memory. *Psychonomic Science*, *3*, 429-430.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, *18*(2), 283–328.

Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review*, *15*(3), 105–130.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Steenari, M.-R., Vuontela, V., Paavonen, E. J., Carlson, S., Fjällberg, M., & Aronen, E. T. (2003). Working memory and sleep in 6-to 13-year-old schoolchildren. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(1), 85–92.

Suchow, J. W. (2014). *Measuring, monitoring, and maintaining memories in a partially observable mind*. Doctoral dissertation, Harvard University.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction)*. The MIT Press. Hardcover.

Taha, H. A. (2007). *Operations research: An introduction*. Pearson/Prentice Hall.

Todd, M. T., Niv, Y., & Cohen, J. D. (2008). Learning to use working memory in partially observable environments through dopaminergic reinforcement. In *Advances in neural information processing systems* (pp. 1689–1696).

Vergauwe, E., Barrouillet, P., & Camos, V. (2009). Visual and spatial working memory are not that dissociated after all: A time-based resource-sharing account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1012-1028.

Wegner, D. M. (2009). How to think, say, or do precisely the worst thing for any occasion. *Science*, *325*, 48–50.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development*, *48*, 13–21.

Williams, M., Hong, S. W., Kang, M. S., Carlisle, N. B., & Woodman, G. F. (2012). The benefit of forgetting. *Psychonomic Bulletin & Review*, *20*(2), 348-355.

Woodward, A. E., & Bjork, R. A. (1971). Forgetting and remembering in free recall: Intentional and unintentional. *Journal of Experimental Psychology*, *89*(1), 109-116.

When to Block versus Interleave Practice? Evidence Against Teaching Fraction Addition before Fraction Multiplication

Rony Patel (rbpatel@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Ran Liu (ranliu@andrew.cmu.edu)
Kenneth Koedinger (koedinger@cmu.edu)
Departments of Psychology and Human-Computer Interaction, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

In practice, mathematics education is blocked (i.e., teaching one topic at a time; CCSS, 2010), but research generally promotes interleaving (i.e., teaching multiple topics together; Rohrer & Taylor, 2007). For example, fraction arithmetic is blocked with students being taught fraction addition before fraction multiplication. Since students often confuse fraction operations to produce arithmetic errors, interleaved fraction arithmetic instruction might be more productive than blocked instruction to teach students to discriminate between the operations. Additionally, a cognitive task analysis suggests that fraction multiplication may be a prerequisite to fraction addition and thus reversing the blocking order may enhance learning. Two experiments with fraction addition and fraction multiplication were run. Experiments 1 and 2 show that interleaved instruction is generally better than the current blocked instruction. Experiment 2 provides evidence that blocking that reverses the standard order -- providing practice on fraction multiplication before fraction addition -- produces better learning.

Keywords: blocking; interleaving; fractions

Introduction

A number of researchers recommend interleaving of practice, whereby multiple topics are taught and practiced together (Kornell & Bjork, 2008; Rohrer & Taylor, 2007). Others have pointed to circumstances where blocking, in which instruction/practice on each type of problem is grouped together and is much more typical in school instruction, is better for learning (Carvalho & Goldstone, 2014). The Knowledge, Learning, and Instruction Framework (Koedinger, Corbett, & Perfetti, 2012) suggests, more generally, that effectiveness of an instructional recommendation depends on a fit with the specifics of the nature of knowledge to be acquired and what learning processes are needed to acquire such knowledge. For example, when the target knowledge involves a need to differentiate similar situations to respond appropriately, interleaving can help the induction of appropriate cues (e.g., as might be represented in the if-parts of production rules) for such differentiation. Fraction addition tasks (e.g., 1/3 + 1/2 = ?) and fraction multiplication tasks (e.g., 1/3 * 1/2 = ?) are a good example in that the tasks look quite similar, only differing the operator (+ vs. *), but the appropriate responses are quite different (converting to common denominators vs. multiplying the numerators and denominators). Thus, this domain provides a fitting context in which to explore theoretical conditions that can explain and predict when interleaving will be better than blocking. We present two experimental studies contrasting interleaved versus blocked practice of these two topics within fraction arithmetic.

We also illustrate how a cognitive task analysis of this domain suggests a reasonable alternative to the typical and recommended approach of teaching fraction addition before fraction multiplication. The analysis not only suggests that fraction multiplication is simpler than fraction addition, but that it may be a prerequisite to the extent fraction multiplication is used to implement the fraction conversion steps required in the fraction addition procedure. The second of the two studies explores this hypothesis.

Practitioners Block, Researchers Interleave

In practice, mathematics is taught in a blocked fashion. The Common Core State Standards (CCSS, 2010) have been adopted by 42 of the 50 states, suggesting the CCSS are representative of the standards for when mathematics topics are generally taught in the United States. Based on these recommendations, fractions should be taught across 3 years, with different topics in different years. Note that these standards do not force blocking or blocking in this order, but the textbooks aligned with the Common Core follow the blocked fraction addition to fraction multiplication sequence. In contrast, research in interleaving and blocking mathematics knowledge shows advantages of interleaving over blocking (Rohrer & Taylor, 2007). Interleaving may aid learning of strategy selection because it provides negative feedback on over-generalized induction of when to perform an operation that blocking does not (cf., Li, Cohen, & Koedinger, 2012).

Sequencing Fraction Arithmetic Instruction

In 2008, the National Mathematics Advisory Panel concluded, "The most important foundational skill not presently developed, appears to be proficiency with fractions" (p. 18). Fraction arithmetic errors are commonly the product of incorrect strategy choices, not computational errors or fact recall errors (Siegler & Pyke, 2013). A

common incorrect strategy choice students make on fraction addition problems is to independently add the numerators and denominators of the addends, for example, getting 2/5, when adding 1/2 and 1/3 (Ni & Zhou, 2005). To make this issue more complicated, this independent whole number strategy is the correct method for solving fraction multiplication problems (e.g., 1/3 times 2/5 does equal (1*2)/(3*5)). Learning fraction multiplication might encourage students to make this error once again in fraction addition problems even if they did learn the correct fraction addition strategy. In addition, fraction addition knowledge may also interfere with fraction multiplication knowledge (Siegler et al., 2011). For example, students sometimes find a common denominator before solving a fraction multiplication problem and then only multiply the numerators resulting in an incorrect answer (e.g., 1/2 * 1/3 = 3/6 * 2/6 = 6/6 = 1).

Given the potential sources of confusion and the presence of interference errors, interleaving fraction arithmetic operations might be beneficial relative to blocking operations, especially for learning good strategy choices. Blocking can have the benefit, however, of strengthening prerequisite subskills before moving on to more complex skills (VanLehn, 1987). Contrary to typical practice, fraction multiplication may be a prerequisite to fraction addition since a variant of fraction multiplication is needed to get equivalent fractions in fraction addition problems with different denominators Further, fraction multiplication provides whole number multiplication practice that may support finding common denominators in fraction addition.

Current Studies

Experiment 1 compares blocked fraction addition problems before fraction multiplication problems to interleaving both problem types to see if interleaving alleviates interference errors between the two types of problems and if blocked instruction causes them. Experiment 2 compares blocked fraction addition to fraction multiplication, blocked fraction multiplication to fraction addition, and interleaving both problem types to see if progressive skill building (i.e., blocked fraction multiplication to fraction addition) outperforms the other blocked condition and reduces the interference amongst the fraction operations.

Experiment 1

Methods

Participants. Participants were 70 sixth grade students in a school in the Greater Pittsburgh Area. 35 students were randomly assigned to each of the two conditions in the study. These students had fraction instruction, but were far from mastery.

Procedure and Design. All participants were run during their regularly scheduled math class with the rest of their classmates. The study took approximately 80 minutes (i.e., two classroom periods). Students worked independently on a computer, using an online intelligent tutoring system. All students had a pretest, midtest, and posttest with two instruction periods between the assessment periods. The inclusion of the midtest allowed us to track changes over time.

Conditions. *Blocked Condition (Fraction Addition before Fraction Multiplication):* Students see 24 fraction addition problems with correctness feedback at every step in the first instruction period followed by 24 fraction multiplication problems with correctness feedback in the second instruction period.

Interleaved Condition: Students are presented the same 24 fraction addition problems and 24 fraction multiplication problems with correctness feedback from the blocked condition in random order. This design was chosen as opposed to perfectly interleaving to avoid students from learning that they simply needed to use the alternating strategies with every other problem. Twelve addition and 12 multiplication problems are randomly presented in each instruction block.

Instruction. All students were presented with a tutor interface for practicing fraction addition problems with same denominators, fraction addition problems with different denominators, and fraction multiplication problems with different denominators (created with CTAT; Aleven, McClarren,, Sewall, & Koedinger, 2006). For all three problem types, students were provided a problem and provided a space to input a numerator and denominator answer. Students were also provided a checkbox they could use to indicate whether the fractions in the problem need to be converted before the problem could be solved (for fraction addition problems with different denominators). See Figure 1 below for example problems. If students were incorrect completing any particular step, the system would mark that step red. Students were required to input the correct answer for any given step before moving on to the next step. If he or she was stuck, the system provided up to three hints per step. The first hint was always abstract (e.g., for fraction multiplication problems, multiply the denominators to get the denominator answer). The second hint provided the concrete instantiation (e.g., in this case, that means multiplying 2 and 3). The final hint provided the answer (e.g., please input 6 into highlighted cell). Once students provided the correct answer, the step would turn green to acknowledge a correct answer.

Assessment Tasks. *Fraction Addition:* Participants were presented 4 fraction addition problems (2 same denominator and 2 different denominators). Students were scored on accuracy and incorrect strategy choice. Accuracy was based on students having the correct magnitude (i.e., if the answer was 1/2, all equivalent fractions, even if they are not reduced, count as a valid answer). In this case, an incorrect strategy choice reflects a strategy choice that would be

successful for a fraction multiplication problem. This includes the independent whole number strategy (adding the numerators and denominators independently) or multiplying either the two numerators or the two denominators while adding the others.

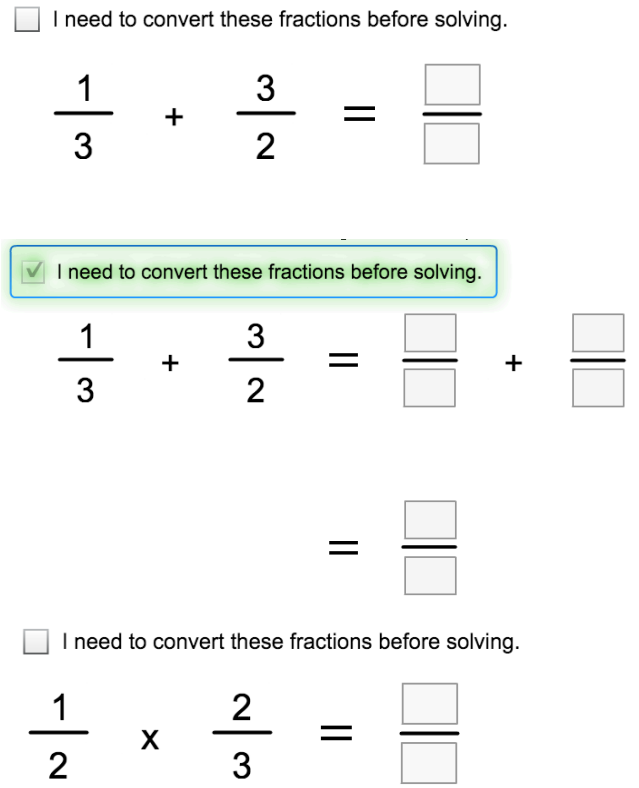


Figure 1. Example instructional interfaces of (a) a fraction addition problem with different denominators before converting, (b) a fraction addition problem with different denominators after the student requests to convert denominators, and (c) a fraction multiplication problem. Students were prompted to solve the following fraction arithmetic problem for every problem provided.

Fraction Multiplication: Participants were presented 4 fraction multiplication problems (2 same denominator and 2 different denominators). Students were scored on accuracy and incorrect strategy choice. Accuracy was based on students having the correct magnitude. In this case, an incorrect strategy choice reflects a strategy choice that would be successful for a fraction multiplication problem. This includes finding a common denominator and only multiplying the numerators (assuming a common denominator was not present already), multiplying the numerators (or denominators) and adding the denominators (or numerators).

Fraction Division: Participants were presented 4 fraction division problems (2 same denominator and 2 different denominators) at posttest only. Students were scored on accuracy.

Results

Blocked Addition-to-Multiplication VS. Interleaved. In a regression, both pretest proportion correct and condition (interleaved = 1) were used to predict posttest proportion correct (R^2 = 32.54%, $F(2,67) = 16.16$, $p < .01$). Pretest was a significant predictor (i.e., students with higher scores pretest scores did better on the posttest; $B = 0.50$ (0.10), $t = 5.06$, $p < 0.01$). Condition was a marginally significant predictor (i.e., interleaved marginally outperformed blocking at posttest; $B = 0.08$ (0.05), $t = 1.80$, $p = 0.08$). The interleaved condition had a posttest accuracy of 79% compared to the blocked addition-to-multiplication posttest accuracy of 68%.

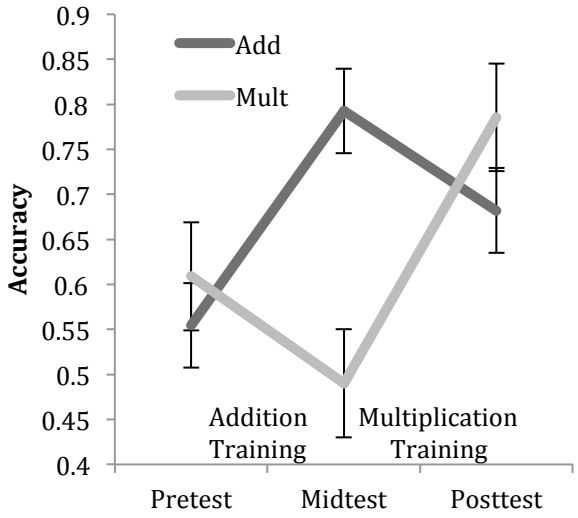


Figure 2. Students in the blocked addition-to-multiplication condition in Experiment 1 performed better on fraction addition after fraction addition training (between the pretest and the midtest), while doing worse on fraction multiplication. After fraction multiplication training (between midtest and posttest), students performed better on fraction multiplication and worse on fraction addition. Error bars reflect +/- 1 SE.

Transfer to Fraction Division. An independent t-test was conducted to compare posttest division proportion correct between conditions. The interleaved group significantly outperformed the blocked addition-to-multiplication group (70% vs. 57%; $t = -2.27$, $df = 66$, $p = 0.03$).

Evidence for Interference. Figure 2 shows how blocked fraction addition practice (before the midtest) changes performance on fraction addition and fraction multiplication and how fraction multiplication practice (between mid and posttest) produces further changes. Notice how there is positive transfer within the two task types, but negative transfer or interference between them. Paired t-tests demonstrate that fraction addition practice between the pretest and midtest produces improvement on fraction addition (56% vs. 82%, $t = -4.61$, $df = 37$, $p < 0.01$), but

decline on fraction multiplication (62% vs. 48%, $t = 2.58$, $df = 37$, $p = 0.01$). Similarly, practice on fraction multiplication between the midtest and the posttest produces improvement on fraction multiplication (48% vs. 82%, $t = -3.34$, $df = 34$, $p < 0.01$), but decline on fraction addition (82% vs. 70%, $t = 2.63$, $df = 34$, $p = 0.01$).

Insights from Process Data. We analyzed the learning process data to further investigate why the interleaved practice may have yielded better learning. To do so, we fit a logistic regression model predicting first-attempt correctness on each step of each problem. The predictors included an intercept parameter for each student, an intercept parameter for each problem step (e.g., entering the correct numerator on a multiplication problem) for each condition (blocked vs. interleaved), and a slope parameter for each type of problem step for each condition. Examining the beta estimates revealed striking condition-driven differences in one particular type of skill reflecting the very initial step in which students must check a box to indicate whether to convert the fractions. In particular, the slope estimates for this skill were lower in the interleaved than in the blocked condition, suggesting a slower progression in mastering this skill during learning (Figure 3). Yet, the slower progression on this skill resulted in better pre-post gains.

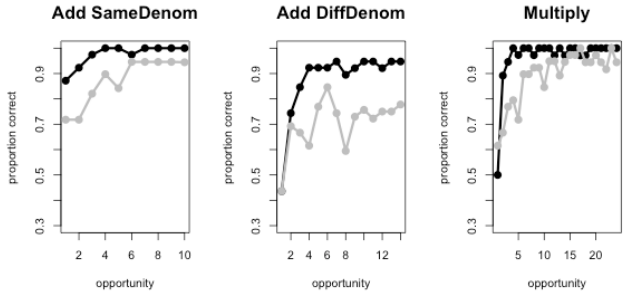


Figure 3. Process data learning curves for the skill of deciding whether to convert fractions. Severe interference when switching problem types is apparent in the blocked condition (black) but is more incremental in the interleaved condition (grey).

We interpreted this as evidence of local trial-to-trial interference between problem types, in the step where students must decide the appropriate procedure to carry out given the problem (i.e., they had to pay attention to the operator to see if it was a '+' or 'x'). Despite evidence of more local interference and slower progression in this procedure selection step, students get much more explicit practice handling the effects of such interference. We hypothesize that this "procedure selection" practice underlies the superior pre-post gains observed in the interleaved condition.

Figure 3 also illustrates the striking interference that occurs for this procedure selection skill when there are shifts between problem types in the blocked addition-to-multiplication condition. This interference is apparent at the sharp learning curve dips that occurs in the fraction addition block at the point where students switch from adding same-denominator fractions (which do not require checking the box to convert denominators) to adding different-denominator fractions (which do require checking the box), and again at the point where students switch to multiplying fractions (which do not require checking the box).

Discussion and Motivation for Experiment 2

We saw support for the hypothesis that interleaved fraction addition and multiplication practice produces better learning than does conventional blocked ordering, particularly on the measure of transfer to fraction division. We found clear evidence of interference that we have previously modeled as over-generalized induction of the if-part of production rules in blocked practice (Li et al., 2012). Interleaved practice provides frequent negative feedback that can be used to refine the if-part to prevent over-generalization errors.

In Experiment 2, we replicated the two conditions from Experiment 1 and included fraction division at all assessment points. Further, based on cognitive task analysis supported by the tutor process data, we also added a third condition that provides blocked instruction in a non-conventional ordering, going from fraction multiplication to fraction addition. In the tutor process data, we observed interference on specific steps within fraction arithmetic procedures such as students incorrectly trying to apply the independent whole number strategy to a fraction addition problem (Liu, Patel, & Koedinger, 2016). These errors were more likely if there had been practice on fraction multiplication problems prior to the fraction addition problem. We also observed student difficulties with finding a common denominator, but also with converting to equivalent fractions. Since this step involves fraction multiplication, at least in principle, we wondered whether there may be positive transfer from fraction multiplication to fraction addition.

Experiment 2

Methods

Participants were 118 sixth grade students in a school in the Greater Pittsburgh Area. 59 students were in the interleaved condition, while 59 students were in the blocked condition. In the blocked condition, 29 students were given fraction addition training before fraction multiplication training, and the other 30 students were given fraction multiplication training before fraction addition training.

The main difference from Experiment 1 is that this experiment included an additional blocked condition, moving from fraction multiplication to fraction addition. In this condition, students saw 24 fraction multiplication problems followed by 24 fraction addition problems with correctness feedback on every step of all problems. This experiment also employed a delayed posttest 3 days after the study concluded, and included fraction division problems at

all time points. All remaining methods were the same as in Experiment 1.

Results

Interleaved vs. both Blocked Conditions. A regression was run with pretest proportion correct and condition (1 for interleaved and 0 for both blocked conditions) predicting proportion correct on the delayed posttest (R^2 = 55.44%, $F(2,105) = 65.32$, $p < 0.01$). Both pretest ($B = 0.73$ (0.06), $t = 11.40$, $p < 0.01$) and condition ($B = 0.06$ (0.03), $t = 2.28$, $p = 0.02$) were significant predictors. The interleaved condition had a posttest accuracy of 79% compared to the blocked posttest accuracy of 74%.

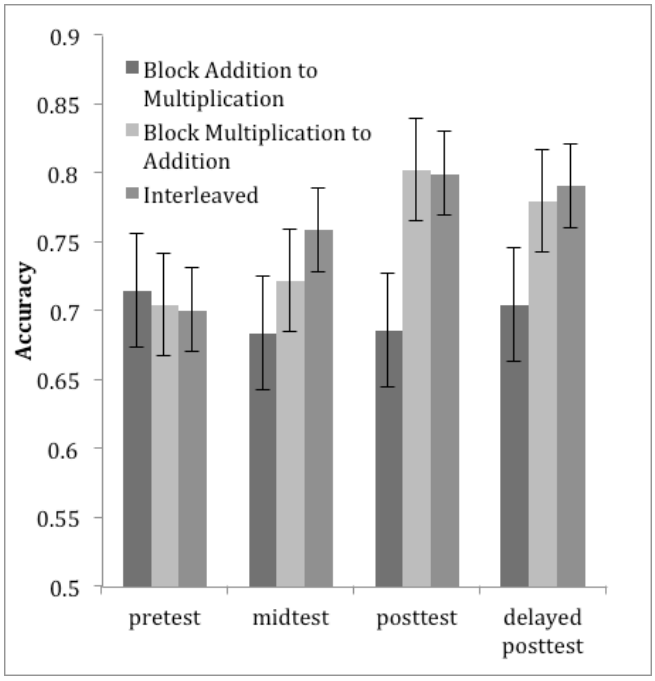


Figure 4. The three sequencing groups in Experiment 2 started with similar accuracies at pretest, but the interleaved and blocked multiplication-to-addition conditions outperformed the standard blocked addition-to-multiplication on both the posttest and delayed posttest.

Comparing Blocking Orders Individually. Separating the blocked conditions by order type reveals a more interesting story (see Figure 4). In a regression, pretest and condition as a factor (all conditions were compared to the blocked addition-to-multiplication condition) were used to predict the delayed posttest (R^2 = 57.13%, $F(3,104) = 47.34$, $p < 0.01$). Pretest was a significant predictor ($B = 0.73$ (0.06), $t = 11.59$, $p < 0.01$). On average, the students in the blocked multiplication-to-addition condition outperformed the students in the blocked addition-to-multiplication condition by 8% (78% vs. 70%; $B = 0.07$ (0.04), $t = 3.57$, $p = 0.04$). On average, the students in the interleaved condition outperformed the students in the blocked addition-to-multiplication condition by 9% (79% vs. 70%; $B = 0.10$ (0.03), $t = 3.12$, $p < 0.01$). Note: The interleaved condition did not outperform the blocked multiplication-to-addition condition (79% vs. 78%; $B = 0.02$ (0.03), $t = 0.88$, $p = 0.38$).

Blocked Multiplication-to-Addition Does Not Cause Interference Errors. Four paired t-tests were conducted to test changes in fraction multiplication and fraction addition from pretest to midtest and midtest to posttest in the blocked multiplication-to-addition condition (the blocked addition-to-multiplication followed similar interference patterns as experiment 1). Specifically, these t-tests were designed to test if learning fraction multiplication causes worse performance on fraction addition (and vice versa) due to overgeneralization errors. After being taught fraction multiplication between the pretest and midtest, students got better at fraction multiplication (79% vs. 91%, $t = -1.93$, $df = 29$, $p = 0.06$) and stayed the same at fraction addition (55% vs. 58%, $t = -0.43$, $df = 29$, $p = 0.66$). After being taught fraction addition between the midtest and the posttest, students got better at fraction addition (58% vs. 78%, $t = -3.33$, $df = 27$, $p < 0.01$) and stayed the same at fraction multiplication (91% vs. 90%, $t = -0.07$, $df = 27$, $p = 0.95$).

Blocked Addition-to-Multiplication Practice Interferes with Fraction Division. The previous study showed an advantage of interleaving to blocking in terms of fraction division accuracy at posttest. This advantage appears to be due to interference in the blocked addition-to-multiplication condition. Changes in division accuracy were only observed in the blocked addition-to-multiplication condition between the pre- and mid-test (72% vs. 50%, $t = 2.68$, $df = 26$, $p = 0.01$). This decline in performance appears to result from over-generalization of fraction addition procedures.

General Discussion

Perhaps the most interesting result suggests that the U.S. curriculum teaches fraction addition and fraction multiplication in the least effective way out of the three alternatives we explored. Across both experiments, blocked fraction addition-to-multiplication practice produced significantly less learning than both interleaved practice and blocked fraction multiplication-to-addition practice.

Consistent with theory (Li et al, 2012; Rohrer & Taylor, 2007), we observed evidence of interference between practice trials of different kinds of fraction operations. That interference was particularly pronounced for fraction multiplication and division after an initial block of fraction addition practice and for fraction addition after a second block of fraction multiplication practice (see interference sections under the results). The generally slower learning rate in the interleaved condition in Experiment 1 is consistent with interference from over-generalization during interleaved practice, but the better posttest performance suggests these over-generalizations were at least somewhat remediated by error feedback during interleaved practice.

We did not find clear evidence of interference in the blocked multiplication-to-addition condition. Perhaps the

lower cognitive load of starting with a simpler task, namely fraction multiplication, left students with capacity for more deliberate and accurate reasoning about strategy selection. Thus, they may have been better able to avoid over-generalizing the fraction multiplication strategy (independent whole number operations) to fraction addition. Further, when moving to fraction addition, the prior fraction multiplication practice may have aided performance of key steps, particularly finding a common denominator (which requires whole number multiplication) and converting the numerators to maintain equivalent fractions (which arguably requires faction multiplication, e.g., 1/2*2/2 = 2/4). With better prior preparation, students may have had more cognitive headroom to more deliberately avoid over-generalizing similar to above, but in this case from fraction addition to fraction multiplication or division. While this argument is, in broad strokes, consistent with VanLehn (1987), a precise application of the SimStudent model in Li et al (2012) would still predict interference due over-generalization. Thus, more work is warranted. A replication of this reverse blocking effect is desirable.

Applications to Education

In the current U.S. standards, fraction arithmetic training follows the same order as whole number arithmetic training (addition, subtraction, multiplication, and finally, division). While there may be some appeal to ordering the operations the same way, our results imply that interleaving practice among different operations yields superior learning outcomes. Interleaved practice forces students to explicitly practice recognizing *when* to carry out which procedure, an aspect that is important for overcoming interference. An alternative implication is that re-structuring blocked instruction to progress from simpler subskills to more complex skills (i.e., practicing fraction and whole number multiplication before adding fractions with different denominators), could be more optimal than what is currently implemented in classrooms.

Generalizing to Other Domains

Our studies targeted the fraction arithmetic domain and past experience and theory give us pause in suggesting cross-domain generalizations (Koedinger et al, 2012). We specifically chose fraction addition and multiplication because they present highly similar tasks features that require different strategic responses. Other task type pairs that have this characteristic are good candidates for applying interleaving. For example, tasks types of finding the area versus perimeter of figures are highly similar (with the words "area" or "perimeter" sometimes being the only difference) and thus interleaving practice of these task types should be effective at reducing over-generalization errors (i.e., computing the area on a perimeter problem). On the other hand, we also found evidence consistent with some forms of blocking being as effective as interleaving. The benefits of blocking may be particularly enhanced when subskills accumulate into more complex tasks. Thus, we recommend careful cognitive task analysis to support the decision of when to block or interleave.

Acknowledgments

Thanks to Sharon Carver, Vincent Aleven, and Mimi McLaughlin for their advice. Support came from grants R305B090023 and R305B110003 from the Department of Education's Institute of Education Sciences.

References

Carvalho, P.F., & Goldstone, R.L. (2014). Putting category learning in order: category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & cognition*, *42*(3), 481-495.

Common Core State Standards Initiative. (2010). Common Core State Standards for mathematics. Washington, D.C.: National Governors Association Center for Best Practices and the Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/CCSSI_Math Standards.pdf

Li, N., Cohen, W., Koedinger, K.R. (2012). Problem order implications for learning transfer. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, 7315 (pp. 185-194). Berlin: Springer.

Liu, Patel, & Koedinger (2016). Modeling Common Misconceptions in Learning Process Data. Submitted for publication.

Koedinger, K.R., Corbett, A.T. & Perfetti, C. (2012). The knowledge-learning-instruction(KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36(5), 757-798*.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories is spacing the "enemy of induction"? *Psychological science*, *19*(6), 585-592.

National Mathematics Advisory Panel (NMAP). (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. Washington, DC: Department of Education.

Ni, Y., & Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist*, *40*(1), 27-52.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481-498.

Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental psychology*, *49*(10), 1994.

Siegler, R.S., Thompson, C.A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*, 273–296.

VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence*, *31*(1), 1-40.

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2006, June). The cognitive tutor authoring tools (CTAT): preliminary evaluation of efficiency gains. In *Intelligent Tutoring Systems* (pp. 61-70). Springer Berlin Heidelberg.

Helping people make better decisions using optimal gamification

Falk Lieder (falk.lieder@berkeley.edu)
Helen Wills Neuroscience Institute, University of California at Berkeley, CA, USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California at Berkeley, CA, USA

Abstract

Game elements like points and levels are a popular tool to nudge and engage students and customers. Yet, no theory can tell us which incentive structures work and how to design them. Here we connect the practice of gamification to the theory of reward shaping in reinforcement learning. We leverage this connection to develop a method for designing effective incentive structures and delineating when gamification will succeed from when it will fail. We evaluate our method in two behavioral experiments. The results of the first experiment demonstrate that incentive structures designed by our method help people make better, less short-sighted decisions and avoid the pitfalls of less principled approaches. The results of the second experiment illustrate that such incentive structures can be effectively implemented using game elements like points and badges. These results suggest that our method provides a principled way to leverage gamification to help people make better decisions.

Keywords: Gamification; Decision-Making; Bounded Rationality; Reinforcement Learning; Decision-Support

Introduction

Most decisions have both immediate and delayed consequences. For instance, investing in a pension plan entails less fun today than buying a wide-screen TV but a higher standard of living 25 years later. Unfortunately, the immediate outcomes often dominate people's considerations because future outcomes are discounted disproportionately (Berns, Laibson, & Loewenstein, 2007). One of the reasons for this *myopia* is that optimal long-term planning is often intractable because the number of possible scenarios grows exponentially as you look ahead further. Consequently, people have to rely on fallible heuristics to limit the length and number of scenarios they consider (Huys et al., 2015). While these heuristics can make us short-sighted in some situations (Huys et al., 2012), they work very well in others (Gigerenzer, 2008). Thus, perhaps, it is not our heuristics that are broken but the decision environments in which they fail (Gigerenzer, 2008; McGonigal, 2011).

The myopic nature of human decision-making (Berns et al., 2007) suggests that decision environments can be repaired by aligning each action's immediate rewards with the value of its long-term consequences. This could be achieved through *gamification* (Deterding, Dixon, Khaled, & Nacke, 2011). Gamification is the use of game elements such as points, levels, and badges in non-game contexts like education, the work place, health, and business. This approach has become extremely popular in the past five years. It is now widely used to engage people and nudge their decisions (Hamari, Koivisto, & Sarsa, 2014), and it has also inspired tools that help people achieve their goals and improve themselves (McGonigal, 2015; Kamb, 2016; Henry, 2014).

While gamification can have positive effects on motivation, engagement, behavior, and learning outcomes (Hamari et al., 2014), it can also have unintended negative consequences (Callan, Bauer, & Landers, 2015; Devers & Gurung, 2015). Unfortunately, it is currently impossible to predict whether gamification will succeed or fail (Hamari et al., 2014; Devers & Gurung, 2015), and there is no principled way to determine exactly how many points should be awarded for a given action. Here we address these problems by connecting the practice of gamification to the theory of pseudo-rewards in reinforcement learning. We leverage this connection to offer a mathematical framework for gamification and a computational method for designing optimal incentive structures. Our method offloads the computations of long-term planning from people by building the optimal decision into the incentive structure so that foresight is no longer necessary. This helps people make better decisions that are less short-sighted.

The plan for this paper is as follows: We start by introducing the theory of pseudo-rewards from reinforcement learning. We then apply this theory to derive a method for designing optimal incentive structures. Finally, we test the effectiveness of our method in two behavioral experiments. We close with implications for decision support and gamification.

Formalizing Gamification

Each sequential decision problem can be modeled as a *Markov Decision Process* (MDP)

$$M = (\mathcal{S}, \mathcal{A}, T, \gamma, r, P_0), \tag{1}$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $T(s,a,s')$ is the probability that the agent will transition from state s to state s' if it takes action a, $0 \leq \gamma \leq 1$ is the discount factor, $r(s,a,s')$ is the reward generated by this transition, and P_0 is the probability distribution of the initial state S_0 (Sutton & Barto, 1998). A *policy* $\pi : \mathcal{S} \mapsto \mathcal{A}$ specifies which action to take in each of the states. The expected sum of discounted rewards that a policy π will generate in the MDP M starting from a state s is known as its *value function*

$$V_M^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right]. \tag{2}$$

The optimal policy $\pi^\star_M$ maximizes the expected sum of discounted rewards, that is

$$\pi^\star_M = \arg\max_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(S_t, \pi(S_t), S_{t+1})\right], \tag{3}$$

and its value function satisfies the Bellman equation

$$V^\star_M(s_t) = \max_a \mathbb{E}\left[r(s_t, a, S_{t+1}) + \gamma \cdot V^\star_M(S_{t+1})\right]. \tag{4}$$

We can therefore rewrite the optimal policy as

$$\pi^\star_M(s) = \arg\max_a \mathbb{E}\left[r(s_t, a, S_{t+1}) + \gamma \cdot V^\star_M(S_{t+1})\right], \tag{5}$$

which reveals that it is myopic with respect to the sum of the immediate reward and the discounted value of the next state.

Here, we leverage the MDP framework to model game elements like points and badges as *pseudo-rewards* $f(s,a,s')$ that are added to the reward function $r(s,a,s')$ of a decision environment M to create a modified environment $M' = (\mathcal{S}, \mathcal{A}, T, \gamma, r', P_0)$ with a more benign reward function $r'(s,a,s') = r(s,a,s') + f(s,a,s')$. From this perspective, the problem with misaligned incentives is that they change the optimal policy $\pi^\star_M$ of the original decision problem M into a different policy $\pi^\star_{M'}$ that is optimal for the gamified environment M' but not for the original environment M. To avoid this problem we have to ensure that each optimal policy of M' is also an optimal policy of M.

Fortunately, research in reinforcement learning has identified the necessary and sufficient conditions pseudo-rewards have to satisfy to achieve this: according to the *shaping theorem* (Ng, Harada, & Russell, 1999) adding pseudo-rewards retains the optimal policies of any original MDP if and only if the pseudo reward function f is *potential-based*, that is if there exists a *potential function* $\Phi : \mathcal{S} \mapsto \mathbb{R}$ such that

$$f(s,a,s') = \gamma \cdot \Phi(s') - \Phi(s), \tag{6}$$

for all states s, actions a, and successor states s'.

Pseudo-rewards can be shifted and scaled without changing the optimal policy, because linear transformations of potential-based pseudo-rewards are also potential-based:

$$a \cdot f(s,a,s') + b = \gamma \cdot \Phi'(s') - \Phi'(s), \tag{7}$$

$$\text{for } \Phi'(s) = a \cdot \Phi(s) - \frac{b}{1-\gamma}. \tag{8}$$

Shifting all pseudo-rewards $f(s,a,s')$ by the rewards $r(s,a,s')$ also retains the optimal policy, because it is equivalent to multiplying all rewards by 2 and positive linear transformations of the rewards preserve the optimal policy (Ng et al., 1999).

If gamification is to help people achieve their goals, then the pseudo-rewards added in the form of points or badges must *not* divert people from the best course of action but make its path easier to follow. Otherwise gamification would lead people astray instead of guiding them to their goals. Hence, the practical significance of the shaping theorem is that it gives the architects of incentive structures a method to rule out incentivizing counter-productive behaviors:

1. Model the decision environment as a MDP.
2. Define a potential function Φ that specifies the value of each state of the MDP.
3. Assign points according to Equation 6.

This method may be useful to avoid some of the dark sides of gamification (Callan et al., 2015; Devers & Gurung, 2015). To make this proposal more concrete and actionable, the next section presents a method for designing good potential functions.

Designing Optimal Incentive Structures

While the shaping theorem constrains pseudo-rewards to be potential-based there are infinitely many potential functions that one could choose. Given that people's cognitive limitations prevent them from fully incorporating distant rewards (Huys et al., 2012; Berns et al., 2007), the modified reward structure $r'(s,a,s')$ should be such that the best action yields the highest immediate reward, that is

$$\pi^\star_M(s) = \arg\max_a r'(s,a,s'). \tag{9}$$

Here, we show that this can be achieved with our method by setting the potential function Φ to the optimal value function $V^\star_M$ of the decision environment M, that is

$$\Phi^\star(s) = V^\star_M(s) = \max_\pi V^\pi_M(s). \tag{10}$$

First, note that the resulting pseudo-rewards are

$$f(s,a,s') = \gamma \cdot V^\star_M(s') - V^\star_M(s), \tag{11}$$

which leads to the modified reward function

$$r'(s,a,s') = r(s,a,s') + \gamma \cdot V^\star_M(s') - V^\star_M(s). \tag{12}$$

Hence, if the agent was myopic its policy would be

$$\begin{aligned} \pi(s) &= \arg\max_a \mathbb{E}\left[r(s,a,s') + \gamma \cdot V^\star_M(s') - V^\star_M(s)\right] \\ &= \arg\max_a \mathbb{E}\left[r(s,a,s') + \gamma \cdot V^\star_M(s')\right]. \end{aligned} \tag{13}$$

According to Equation 5, this is the optimal policy $\pi^\star_M$ for the original decision environment M. Thus, people would act optimally even if they were completely myopic. And they should perform equally well, if they do optimal long-term planning to fully exploit the gamified environment M' or learn its optimal policy $\pi^\star_{M'}$ through trial-and-error, because the shaping theorem (Eq. 6) guarantees that the gamified environment M' has the same optimal policy, that is $\pi^\star_{M'} = \pi^\star_M$.

This suggests that potential-based pseudo-rewards derived from $V^\star_M$ should allow even the most short-sighted agent who considers only the immediate reward to perform optimally. In this sense, the pseudo-rewards defined in Equation 11 can be considered optimal. In addition, optimal pseudo-rewards accelerate learning when the agent's initial estimate of the value function is close to 0 (Ng et al., 1999).

Computing the optimal pseudo-rewards requires perfect knowledge of the decision environment and the decision-maker's preferences. This information may be unavailable in practice. Yet, even when the optimal value function $V_M^\star$ cannot be computed, it is often possible to approximate it. If so, the approximate value function $\hat{V}_M$ can be used to approximate the optimal pseudo-rewards (Eq. 11) by

$$\hat{f}(s,a,s') = \gamma \cdot \hat{V}_M(s') - \hat{V}_M(s). \tag{14}$$

For instance, you can estimate the value of a state s from its approximate distance to the goal $s^\star$:

$$\hat{V}_M(s) = \hat{V}_M(s^\star) \cdot \left(1 - \frac{\text{distance}(s,s^\star)}{\max_s \text{distance}(s,s^\star)}\right), \tag{15}$$

where $\hat{V}_M(s^\star)$ is the estimated value of achieving the goal. Based on previous simulations (Ng et al., 1999), we predict that approximate pseudo-rewards (Eq. 14) can have beneficial effects similar to those of optimal pseudo-rewards but weaker. We tested these predictions in two behavioral experiments.

Experiment 1: Modifying Rewards

Methods

We recruited 250 adult participants on Amazon Mechanical Turk. Participants were paid \$0.50 for playing the game shown in Figure 1. In this game, the player receives points for routing an airplane along profitable routes between six cities. In each of the 24 trials the initial location of the airplane was chosen uniformly at random, and the task was to earn as many points as possible. Participants were incentivized by a performance dependent bonus of up to \$2. This game is based on the planning task developed by Huys et al. (2012). Our version of this task is isomorphic to a MDP with six states, two actions, deterministic transitions, and a discount factor of $\gamma = 1 - 1/6$. The locations correspond to the states of the MDP, the two actions correspond to flying to the first or the second destination available from the current location, the routes correspond to state-transitions, and the points participants received for flying those routes are the rewards. The current state was indicated by the position of the aircraft and was updated according to the flight chosen by the participant. The number of points collected in the current trial was shown in the upper right corner of the screen. After each choice there was a 1 in 6 chance that the game would end and the experiment would advance to the next trial, and a 5 in 6 chance that the participant could choose another flight. The participants were instructed to score as high as possible, and their financial bonus was proportional to the rank of their score relative to the scores of all participants in the same condition. The optimal policy in this MDP is to move counter-clockwise around the circle in all states except *Williamsville* and *Brownsville* (see Figure 1). Importantly, at *Williamsville* the optimal policy incurs a large immediate loss, and no other policy achieves a positive reward rate.

Participants were randomly assigned to one of four conditions: In the control group, there were no pseudo-rewards

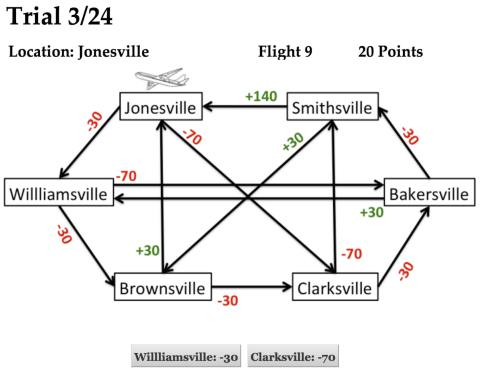


Figure 1: Interface of the control condition of Experiment 1. The map shows the unmodified rewards r.

Pseudo-Rewards	Smiths-		Jones-		Williams-		Browns-		Clarks-		Bakers-	
None	140	30	−30	−70	−30	−70	−30	30	−30	−70	−30	−70
Optimal	2	−76	2	−5	−12	2	−4	2	2	0	2	−42
Approximate	8	−102	−22	−4	−22	−4	36	38	−34	−16	24	−32
Non-Potential-Based	119	9	−51	−41	−51	−41	−1	9	−51	41	−1	9

Table 1: Rewards in Experiment 1. The first entry of each cell is the (modified) reward of the counter-clockwise move and the second one is the (modified) reward of the other move.

(Figure 1). In this condition finding the optimal path required planning 4 steps ahead. In the three experimental conditions the rewards were modified by adding pseudo-rewards. To keep the average reward constant, the pseudo-rewards were mean-centered by subtracting their average; since mean-centering is a linear transformation this retained the guarantees of the shaping theorem (see Eq. 7). Next, the mean-centered pseudo-rewards were added to the rewards of the control condition (see Figure 1) yielding the modified rewards shown in Table 1, and the flight map was updated accordingly. In all other respects the interfaces of the experimental conditions were exactly the same as the interface of the control condition (see Figure 1). All participants were thus unaware of the existence of pseudo-rewards. In the first experimental condition the pseudo-rewards were derived from the optimal value function according to the shaping theorem (Eq. 11). In this condition, looking only 1 step ahead was sufficient to find the optimal path. The second experimental condition used the approximate potential-based pseudo-rewards defined in Equation 14 with the distance-based heuristic value function defined in Equation 15 where $s^\star$ was *Smithsville*, $\Phi(s^\star)$ was its highest immediate reward (i.e., $+140$), and $\text{distance}(a,b)$ was the minimum number of actions needed to get from state a to state b. The resulting pseudo-rewards simplified planning but not as much as the optimal pseudo-rewards. Finding the optimal path required planning 2-3 steps ahead and the immediate losses were smaller. In the third experimental condition, the pseudo-rewards violated the shaping theorem: the pseudo-reward was $+50$ for each transition that reduced the distance to the most valuable state (i.e. *Smithsville*) but there was no penalty for moving away from it.

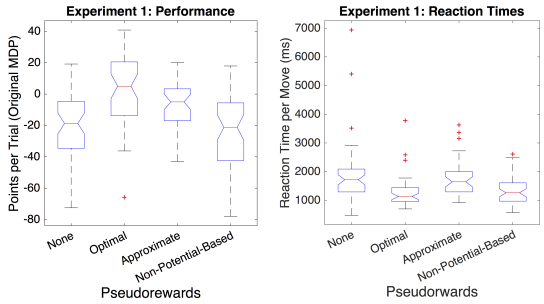


Figure 2: Performance and reaction times in Experiment 1.

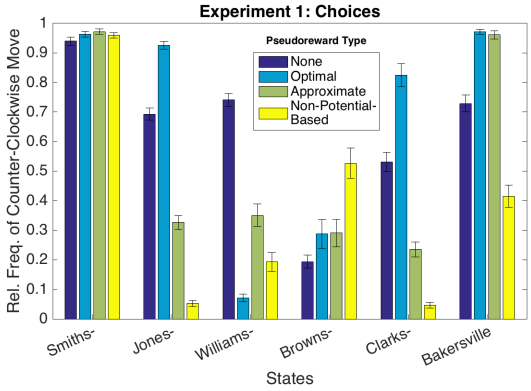


Figure 3: Choice frequencies in each state of Experiment 1 by condition. Error bars enclose 95% confidence intervals.

Results and Discussion

The average completion time of the experiment was 13:37 min. The median response time was 1.3 sec. per choice. We excluded 3 participants who invested less than one third of the median response time of their condition and 11 participants who scored lower than 95% of all participants in their condition. The boxplots in Figure 2 summarize the median performance and reaction times of participants in the four conditions. The median performer of the control group *lost* 18.75 points per trial. By contrast, the majority of the group with optimal pseudo-rewards achieved a net *gain* in the unmodified MDP (median performance: $+5.00$ points/trial). The median performance in the group with approximate potential-based pseudo-rewards was -5.00 points per trial, and in the group with non-potential-based pseudo-rewards the median performance was -21.25 points per trial. A Kruskal-Wallis ANOVA revealed that the type of pseudo-rewards added to the reward function significantly affected people's performance in the original MDP ($H(3) = 40.35, p < 10^{-8}$) as well as their reaction times ($H(3) = 29.96, p < 10^{-5}$). Given that the pseudo-reward type had a significant effect, we performed pairwise Wilcoxon rank sum tests to compare the medians of the four conditions. The non-potential-based pseudo-rewards failed to improve people's performance ($Z = 0.72, p = 0.47$). By contrast, the approximate potential-based pseudo-rewards succeeded to improve their performance ($Z = 2.86, p = 0.0042$) and led to better performance than the heuristic pseudo-rewards that violated the shaping theorem ($Z = 3.61, p = 0.0003$). People performed even better when gamification was based on the optimal pseudo-rewards: adding optimal pseudo-rewards led to better decisions than adding the approximate potential-based pseudo-rewards ($Z = 2.68, p = 0.0074$), presenting the true reward structure ($Z = 4.76, p < 10^{-5}$), or adding the non-potential-based pseudo-rewards ($Z = 5.34, p < 10^{-7}$).

In addition, some pseudo-rewards accelerated the decision process (Figure 2): optimal pseudo-rewards decreased the median response time from 1.72 to 1.14 sec/decision ($Z = -4.19, p < 0.0001$), and non-potential-based pseudo-rewards decreased it to 1.12 sec/decision ($Z = -3.38, p = 0.0007$). But approximate potential-based pseudo-rewards had no significant effect on response time (1.65 sec/decision; $Z = -0.28, p = 0.78$).

Next, we inspected how the pseudo-rewards affected the choices our participants made in each of the six states (see Figure 3). The optimal pseudo-rewards significantly changed the choice frequencies in each of the six states and successfully nudged participants to follow the optimal cycle *Smithsville* $\rightarrow$ *Jonesville* $\rightarrow$ *Williamsville* $\rightarrow$ *Bakersville* $\rightarrow$ *Smithsville* (see Figure 1). Their strongest effect was to eliminate the problem that most people would avoid the large loss associated with the correct move from *Williamsville* to *Bakersville* ($\chi^2(2) = 1393.8, p < 10^{-15}$). The optimal pseudo-rewards also increased the frequency of all other correct choices along the optimal cycle, that is the decisions to fly from *Bakersville* to *Smithsville* ($\chi^2(2) = 326.5, p < 10^{-15}$), from *Smithsville* to *Jonesville* ($\chi^2(2) = 7.9, p = 0.0191$), and from *Jonesville* to *Williamsville* ($\chi^2(2) = 299.8, p < 10^{-15}$). In addition, the optimal pseudo-rewards increased the frequency of the correct move from *Clarksville* to *Bakersville* ($\chi^2(2) = 92.0, p < 10^{-15}$). The only negative effect of the optimal pseudo-rewards was to slightly increase the frequency of the suboptimal move from *Brownsville* to *Clarksville* ($\chi^2(2) = 13.2, p = 0.0013$). By contrast, the non-potential-based pseudo-rewards misled our participants to follow the unprofitable cycle *Jonesville* $\rightarrow$ *Clarksville* $\rightarrow$ *Smithsville* $\rightarrow$ *Jonesville* by raising the frequency of the reckless moves from *Jonesville* to *Clarksville* ($\chi^2(2) = 1578.6, p < 10^{-15}$) and from *Clarksville* to *Smithsville* ($\chi^2(2) = 813.7, p < 10^{-15}$). The effect of the approximate pseudo-rewards was beneficial in *Smithsville*, *Williamsville*, and *Bakersville*, but negative in *Jonesville*, *Brownsville*, and *Clarksville* (see Figure 3). This explains why only potential-based pseudo-rewards had a positive net-effect on performance (Figure 2).

Finally, we investigated learning effects by comparing the average choice frequencies in the first five trials versus the last five trials. While people's decisions improved with learning in the conditions with potential-based pseudo-rewards, learning had a negative effect when the pseudo-rewards violated the shaping theorem: In the condition with non-potential-based pseudo-rewards learning reduced the frequency of the correct choice in *Jonesville* ($\chi^2(2) = 9.22, p = 0.01$). By con-

trast, in the condition with optimal pseudo-rewards learning improved people's choices in *Smithsville* ($\chi^2(2) = 13.02, p = 0.0015$), and in the condition with approximate potential-based pseudo-rewards learning improved people's choices in *Jonesville* ($\chi^2(2) = 11.44, p = 0.0033$). In the control condition, learning made people more likely to take the correct action in *Williamsville* ($\chi^2(2) = 24.16, p < 0.0001$) and *Bakersville* ($\chi^2(2) = 22.74, p < 0.0001$) but less likely to take the correct action in *Clarksville* ($\chi^2(2) = 8.80, p = 0.0123$). In summary, while learning with potential-based pseudo-rewards guided people closer towards the optimal policy, non-potential-based pseudo-rewards lured them away from it. This is consistent with the shaping theorem's assertion that pseudo-rewards have to be potential-based to always retain the optimal policy.

In summary, we found that pseudo-rewards can help people make better decisions–but only when they are designed well. The results support the proposed method for designing incentive structures: Assigning pseudo-rewards according to the shaping theorem avoided the negative effects of non-potential-based pseudo-rewards. Furthermore, using the optimal value function as the shaping potential lead to the greatest improvement in decision-quality.

Experiment 2: Explicit Pseudo-Rewards

To assess the potential of gamification for decision-support in real life, Experiment 2 conveyed pseudo-rewards by points without monetary value.

Methods

We recruited 400 participants on Amazon Mechanical Turk. Participants were paid \$2.50 for about 20 min of work plus a performance dependent bonus of up to \$2 that averaged at \$1. We inspected the data from the 355 participants who had not participated in earlier versions of the experiment and excluded 19 participants who responded in less than one third of the median response time of all participants or performed worse than 95% of the participants in their condition.

The task from Experiment 1 was modified by adding stars in two new experimental conditions (see Figure 4). To make it easy for participants to add rewards plus pseudo-rewards, the rewards were scaled down by a factor of 10 and the optimal pseudo-rewards were recomputed according to Equation 11. The pseudo-rewards awarded for each action were then shifted by the expected return of the optimal policy (\$0.90) so that they predict how much money the player will earn in the long-run if they choose that action and act optimally afterwards. The experiment had four conditions: In the control condition no pseudo-rewards were presented. Three experimental conditions presented the optimal pseudo-rewards in three different formats: condition 2 embedded the pseudo-rewards into the reward structure as in Experiment 1; condition 3 presented them separately in the form of stars; and in condition 4 the number of stars communicated the sum of pseudo-reward plus immediate reward. All numbers were rounded to one significant digit. In the conditions with stars

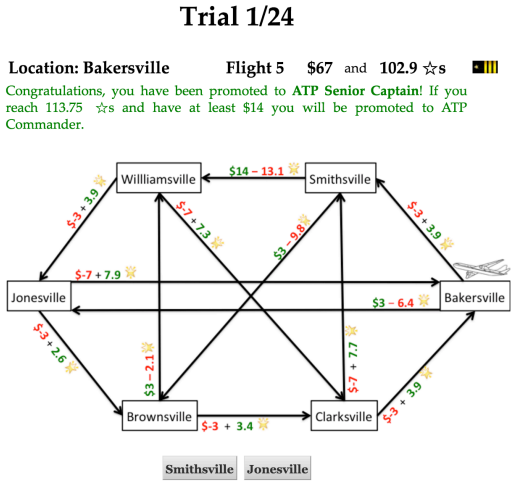


Figure 4: Screenshot of Experiment 2.

the instructions stated that the stars were designed to help the pilots make better decisions and explained their meaning. In addition, the instructions included the tip that the better flight is the one with the higher sum of stars plus dollars (condition 3) or that it is the one awarded more stars (condition 4). The stars had no monetary value, but they determined whether and when the character played by the participant would be promoted. The character could rise from *Trainee* to *ATP senior captain* via 15 intermediate levels. The number of points required to reach the next level increased according to the difficulty curve proposed by Bostan and Öğüt (2009). Participants were told how many stars and dollars were required to reach the next level (see Figure 4). The current score and the shoulder batch corresponding to the current level were shown at the top of the screen, and a feedback message appeared whenever the character was promoted. The player started the game with +\$50 so that their balance would remain positive as they learned to play the game. In all conditions a quiz ensured the participants understood the task and its incentives before they could start the game. This quiz comprised three questions on how the participant's financial bonus would be determined and three questions testing they understood the mechanics of the task.

Results and Discussion

Overall, presenting pseudo-rewards in one of the three formats significantly improved people's performance ($Z = 3.43$, $p = 0.0006$). Most importantly, adding points (i.e., stars) without monetary value can be just as effective as directly modifying the reward structure of the environment: Integrated pseudo-rewards significantly increased people's performance from -0.73 dollars/trial to $+0.17$ dollars/trial ($Z = 3.69, p = 0.0002$). The resulting level of performance was not significantly different from the performance in the condition with embedded pseudo-rewards ($+0.42$ dollars/trial, $Z = 0.52, p = 0.62$). By contrast, presenting pseudo-rewards separately failed to significantly increase people's performance

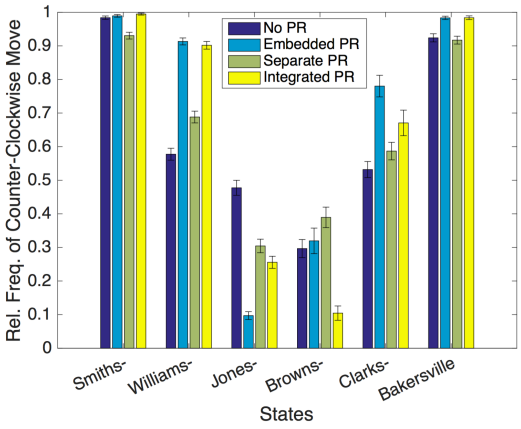


Figure 5: Choice Frequencies in Experiment 2: Effects of stars and badges on performance

(median performance: -0.5 dollars/trial; $Z = 0.22, p = 0.83$).

Inspecting the choice frequencies (Figure 5) confirmed that the three presentation formats had significantly different effects: Embedded pseudo-rewards and integrated pseudo-rewards were more beneficial than separately presented pseudo-rewards in all 6 states (all $p \leq 0.0218$). Embedded pseudo-rewards were more beneficial than integrated pseudo-rewards in Jonesville and Clarksville (both $p \leq 0.0001$), but integrated pseudo-rewards were more beneficial than embedded pseudo-rewards in Brownsville ($p < 10^{-9}$). Furthermore, participants were significantly faster when pseudo-rewards were embedded in the decision environment than when they were presented separately ($Z = -4.06, p < 0.0001$) or in the integrated format ($Z = -2.78, p = 0.0053$).

In conclusion, adding points that convey the sum of optimal pseudo-rewards plus immediate reward can be as effective as changing the reward structure itself.

Conclusion

We have proposed a general method for improving incentive structures based on the theory of MDPs and the shaping theorem. Its basic idea is to offload the computation necessary for long-term planning into the reward structure of the environment such that people will act optimally even when they consider only immediate rewards. The results of Experiment 1 provide a proof of principle that our approach can help people make better sequential decisions. Our findings suggest that the shaping theorem can be used to delineate when gamification will succeed from when it will fail and to design incentive structures that avoid the perils of less-principled approaches to gamification. Experiment 2 illustrated that the incentive structures designed with our method can be implemented with game elements like points and badges. In both experiments the pseudo-rewards helped people overcome their short-sighted tendency to avoid an aversive action with desirable long-term consequences in favor of immediate reward–a cognitive limitation that can manifest in procrastination and impulsivity. Therefore, our method might be useful for improving inter-temporal choice. Our findings are consistent with the view that the limitations of human decision-making can be overcome by reshaping incentive structures that make us prone to fail into ones that our heuristics were designed for. Our method achieves this by solving people's planning problems for them and restructuring their incentives accordingly. The program providing the pseudo-rewards can be seen as a cognitive prosthesis because it compensates for people's cognitive limitations without restricting their freedom. In conclusion, optimal gamification may provide a principled way to help people achieve their goals and procrastinate less.

Acknowledgments. This work was supported by grant number ONR MURI N00014-13-1-0341. We thank Rika Antonova, Ellie Kon, Paul Krueger, Mike Pacer, Daniel Reichman, Stuart Russell, Jordan Suchow, and the CoCoSci lab for feedback and discussions.

References

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice–toward an integrative framework. *Trends in Cognitive Sciences*, *11*(11), 482–488.

Bostan, B., & Öğüt, S. (2009). Game challenges and difficulty levels: lessons learned from RPGs. In G. K. Yeo & Y. Cai (Eds.), *Learn to Game, Game to Learn: Proceedings of the 40th Conference of the International Simulation and Gaming Association.*

Callan, R. C., Bauer, K. N., & Landers, R. N. (2015). How to avoid the dark side of gamification: Ten business scenarios and their unintended consequences. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 553–568). Cham: Springer.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15).

Devers, C. J., & Gurung, R. A. (2015). Critical perspective on gamification in education. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 417–430). Cham: Springer.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford: Oxford University Press.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?–A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034).

Henry, A. (2014). *The best tools to (productively) gamify every aspect of your life.* Retrieved from `http://lifehacker.com/the-best-tools-to-productively-gamify-every-aspect-of-1531404316`

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*, *8*(3), e1002410.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098–3103.

Kamb, S. (2016). *Level up your life: How to unlock adventure and happiness by becoming the hero of your own story.* Emmaus: Rodale Books.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world.* New York: Penguin.

McGonigal, J. (2015). *SuperBetter: A revolutionary approach to getting stronger, happier, braver and more resilient–powered by the science of games.* London, UK: Penguin Press.

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th Annual International Conference on Machine Learning* (pp. 278–287). San Francisco: Morgan Kaufmann.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA, USA: MIT press.

Learning biases may prevent lexicalization of pragmatic inferences: a case study combining iterated (Bayesian) learning and functional selection

Thomas Brochhagen (t.s.brochhagen@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
Science Park, Kruislaan 107, 1098 XG Amsterdam, The Netherlands

Michael Franke (mchfranke@gmail.com)
Department of Linguistics, University of Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany

Robert van Rooij (r.a.m.vanRooij@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
Science Park, Kruislaan 107, 1098 XG Amsterdam, The Netherlands

Abstract

Natural languages exhibit properties that are difficult to explain from a purely functional perspective. One of these properties is the systematic lack of upper-bounds in the literal meaning of scalar expressions. This investigation addresses the development and selection of such semantics from a space of possible alternatives. To do so we put forward a model that integrates Bayesian learning into the replicator-mutator dynamics commonly used in evolutionary game theory. We argue this synthesis to provide a suitable and general model to analyze the dynamics involved in the use and transmission of language. Our results shed light on the semantics-pragmatics divide and show how a learning bias in tandem with functional pressure may prevent the lexicalization of pragmatic inferences.

Keywords: semantics; pragmatics; iterated learning; evolutionary game theory; scalar expressions

Introduction

Why are natural languages structured the way they are and not differently? In particular, what factors are involved in the selection of semantic structure from a space of possible alternatives? A number of recent studies have began to investigate such issues, pertaining to the development and emergence of linguistic features (see Steels 2015 and Tamariz & Kirby 2016 for recent overviews). While different methodologies have been put forward to this end, the overarching argument is one of competing pressures: natural languages need to enable successful communication, as well as be suited for acquisition. That is, they need to be well-adapted to the communicative needs of their users but they also need to be learnable in order to survive their faithful transmission across generations.

The present investigation focuses on the interplay of such selective forces by means of a case study on the lexicalization of semantic upper-bounds. Using evolutionary game theory, framed in terms of language use and learning, we investigate the prevalence of a lack of upper-bounds in the literal meaning of scalar expressions. The innovation of the model lies in its combination of functional pressure on successful communication, effects of learning biases on (iterated) Bayesian language learning (Griffiths & Kalish, 2007), and probabilistic models of language use in populations with distinct lexica (Frank & Goodman, 2012; Franke & Jäger, 2014; Bergen et al., to appear). Our results show that a learning bias for simple semantic representations coupled with communal language exposure can lead to a prevalence of a lack of semantic upper-bounds provided bounds can be inferred via pragmatic reasoning. This, in turn, offers an explanation to why certain pragmatic inferences fail to lexicalize.

Conveying and lexicalizing upper-bounds

A considerably large class of natural language expressions do not lexicalize an upper-bound. That is, they are truth-conditionally compatible with more informative or "stronger" alternatives. Here, informativity is understood as an order over entailment. Examples in English include, among many others, numerals such as *five* and *six*, scalar adjectives like *cold* and *big*, as well as quantifiers like *some* and *many*. The commonality of these expressions lies in their literal meaning being compatible with the truth of stronger (relevant) alternatives. For example, if it is true that 'Bill read five books' it may well be true that 'Bill read six/seven/... books'. Crucially, the latter entails the former; if it is true that 'some students came to class' then this would also hold if it were true that 'all students came to class'. Prima facie this pattern may seem surprising. If it were true that 'some (but not all) students came', then it would putatively serve interlocutors better if *some* semantically ruled out the stronger *all* case.

In practice, however, what is communicated often goes beyond the literal meaning of expressions. In particular, mutual reasoning between interlocutors can lead to the enrichment of semantic content (Grice, 1975). Such *pragmatic reasoning* is driven by interlocutors' mutual expectations of rational language use. In this case the use of a less informative expression when a more informative one could have been used can license a defeasible inference that the stronger alternative does not hold (cf. Horn 1972; Gazdar 1979). In this way, a hearer who assumes the speaker to be knowledgeable and cooperative, i.e., able and willing to supply all relevant information at her disposition, can infer *some* to be pragmatically strengthened to convey an upper-bounded interpretation.

This kind of reasoning was already hinted at above: a speaker would have used the more informative alternative if

it were true. Since she did not, her addressee can infer that it does not hold. Conversely, a speaker who reasons about her addressee can rely on her drawing this inference. Notwithstanding, while pragmatic reasoning provides an answer to how interlocutors may derive and convey upper-bounds, the question why they are not part of the literal meaning still stands. More generally, while a divide between semantics and pragmatics is commonly agreed upon by linguists, under consideration of purely functional pressures this raises the challenge of providing justifications for the former's structure.

We see two main explanations for a lack of semantic upper-bounds in scalar expressions. The first is that pragmatic reasoning offers a general mechanism to strengthen the meaning of a wide range of expressions when the conditions outlined above hold. Consequently, cases where cooperativity or knowledge are not likely to be given are non-committal to whether stronger alternatives hold. If for all the speaker knows 'some students came' but she does not know whether 'all came', then the compatibility of *some* with (possibly) *all* succinctly conveys the speaker's uncertainty about the latter. Given that scalar expressions occur in contexts in which their upper-bounded reading is absent, one could argue for a functional advantage of a lack of semantic upper-bounds: If expressing such a state of affairs is relevant and contextual cues provide enough information for a hearer to discern when a bound is conveyed pragmatically, then doing so is preferred over enforcing the bound overtly through a longer (more complex) expression, e.g. by stating 'some but not all' explicitly. That is, all else being equal, speakers prefer to communicate as economically as possible, and pragmatic reasoning enables them to do so. Additionally, this can be contrasted with the hypothetical alternative of lexicalizing two expressions – one with and one lacking an upper-bound. Four conditions may pressure language to English-like semantics over this alternative: (i) contextual cues are very reliable, morphosyntactic disambiguation is either (ii) not frequently necessary or (iii) not very costly, or (iv) having larger lexica is more costly than morphosyntactic disambiguation. In a nutshell (i) and (ii) place a heavy burden on the ability to retrieve contextual cues to a degree that is unlikely to undercut the benefit of safe communication with more expressions. As for (iii) and (iv), these seem mostly like technical solutions without a proper empirical basis.

A second explanation targets the contrast of the underlying semantic representation of upper-bounds and a lack thereof, positing a learning advantage for the latter. That is, by virtue of their simpler representations, expressions lacking upper-bounds are more readily learnable than their bounded counterparts. If a bound can be supplied pragmatically, this difference may explain the prevalence of such semantics. In the following we explore this hypothesis. While we do not want to argue that functional pressures may not play a role, we see a clear benefit in exploring whether learnability would not give us additional leverage. Furthermore, it should also be stressed that these explanations may well be complementary in that a full-fledged account can reasonably be expected to involve an interplay between them.

Communication and selective dynamics

We employ evolutionary game theory to investigate the selection dynamics of language fragments that capture the contrast between upper-bounded meanings and their unbounded counterparts. Evolutionary game theory offers general and precise means to model the dynamics of linguistic pressures (Nowak & Krakauer, 1999; Huttegger & Zollman, 2013). More concretely, our model uses the replicator-mutator dynamics (see Hofbauer & Sigmund 2003 for an overview). In contrast to previous models we integrate (iterated) Bayesian learning in the dynamics (Griffiths & Kalish, 2007), as well as probabilistic language users of varied degrees of pragmatic sophistication (Frank & Goodman, 2012; Franke & Jäger, 2014) together with multiple lexica (Bergen et al., to appear). In this way the model synthesizes core insights of previous proposals of language use and learning.

Linguistic interactions are represented by signaling games (Lewis, 1969). An interaction involves two players; a speaker and a hearer. The speaker has some private information about the state of the world and tries to convey this information to the hearer by sending a message. The hearer receives the message and interprets it. Communication is successful if the message's interpretation matches the speaker's information, i.e., when the message is interpreted correctly. We restrict our attention to games with two states, $S = \{s_1, s_2\}$, and two messages, $M = \{m_1, m_2\}$. This allows for a minimal contrast between weaker (non-)upper-bounded expressions and stronger alternatives in states were a bound is to be conveyed. Players base their choices on lexica that specify the semantics of expressions. Formally, a lexicon L is a Boolean $(|S|, |M|)$-matrix such as L_a and L_b below.

$$L_a = \begin{array}{c} \\ s_1 \\ s_2 \end{array} \begin{array}{c} \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array} \qquad L_b = \begin{array}{c} \\ s_1 \\ s_2 \end{array} \begin{array}{c} \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

According to L_a, message m_1 is true in s_2 and false in s_1. The converse holds for m_2. In L_b, however, m_1 is true in both s_1 and s_2. In this way L_a stands for a language that encodes an upper-bound for m_1, e.g. *some but not all*, whereas L_b stands for one where m_1 lacks such a bound, e.g. *some*.

To contrast literal language use to its pragmatic counterpart we consider two kinds of players. *Literal players* communicate according to the semantics of their lexica. *Pragmatic players* reason about their interlocutors and base their choices on this reasoning. In other words, pragmatic speakers take into account how listeners would interpret a message and pragmatic listeners take the speaker into account when interpreting. This kind of signaling behavior shares the core features common to models of rational language use such as Rational Speech Act models (Frank & Goodman, 2012) and Quantal- and Best-Response models (Franke, 2009; Franke &

Jäger, 2014). Following this line of research, player behavior can be captured by a hierarchy of reasoning types. Literal players constitute the bottom of the hierarchy, level 0, and are solely guided by the semantics of their language. Players of level $n+1$ behave rationally according to level n behavior of their interlocutors. Presently, it suffices to consider players of no order higher than 1 as the cases considered here offer little room for further pragmatic refinement. Literal hearer and speaker behavior are summarized in (1) and (2), and their pragmatic counterparts in (3) and (4).

$$R_0(s|m;L) \propto P^*(s)L_{sm} \quad (1)$$

$$S_0(m|s;L) \propto \exp(\lambda\, L_{sm}) \quad (2)$$

$$R_1(s|m;L) \propto P^*(s)S_0(m|s;L) \quad (3)$$

$$S_1(m|s;L) \propto \exp(\lambda\, R_0(s|m;L)^{\alpha}) \quad (4)$$

Speakers strive to maximize their communicative success but may occasionally make mistakes in computing their expected utility. This is regulated by the soft-max parameter λ, $\lambda > 0$ (Luce, 1959; Sutton & Barto, 1998). Intuitively, the larger λ the more faithful a speaker's choices are to the maximization of expected success. A second parameter α, $\alpha \in [0,1]$, controls the tension between semantics and pragmatics. Lower α-values lead to more literal signaling whereas larger values lead to stronger pragmatic behavior. $P^* \in \Delta(S)$ is a prior over states. For simplicity in the following we assume this prior to be common and uniform. Lastly, a *player type* is a combination of signaling behavior, i.e., either literal or pragmatic, and a lexicon.

With these player types at hand population-level dynamics can be analyzed. The dynamics involve two key components: communicative fitness and lexicon learnability. The fitness of a type i is given by its expected utility. That is, the fitness of type i, f_i, in a population x is the sum of its weighted expected utility, $f_i = \sum_j x_j U(x_i,x_j)$, where x_j is the proportion of players of type j in x and $U(x_i,x_j)$ is the symmetrized expected utility of x_i and x_j.[1] A type's fitness indicates how well it communicates within a population. The average fitness of the population is Φ, $\Phi = \sum_i x_i f_i$.

The second component is given by a learning matrix Q, which adds (iterated) Bayesian learning to the dynamics (Griffiths & Kalish, 2007). It operationalizes the assumption that new generations of learners combine prior learning biases with observations over language use. This determines the probability of adopting a type. As a consequence, lexica that better explain observations are more likely to be passed onto the next generation faithfully. Here we assume that the adoption of a lexicon is not solely dependent on how well a parent's lexicon explains the data but also on its proportional representation in the population. The intuition is that it may be possible that a lexicon that is – in principle – well-suited for acquisition may nevertheless not be adopted when only few types of the preceding population used it. More concretely, Q_{ji} specifies the probability that a child of type j will be of type i. This is weighted by the probability of being exposed to type i, $N_{ji} = x'_i Q_{ji}$ where x'_i is proportion of i in the previous generation. Q itself depends on observations of linguistic utterances and a learning prior. The set of possible observations is given by all combinations between state-message pairs, $O = \{\langle\langle s_1,m_i\rangle, \langle s_2,m_j\rangle\rangle | m_i,m_j \in M\}$. A member of O encodes that a teacher produced m_i in state s_1 and m_j in s_2. It encodes one witnessed message for each state. A datum d is a sequence of length k of members of O. Learners witness such data sequences. Accordingly, $Q_{ji} \propto \sum_d P(d|t_j)P(t_i|d)$ with $P(t_i|d) \propto P(t_i)P(d|t_i)$. The prior $P(t)$ gives the probability of a type without witnessing data, i.e., it encodes the learning bias of players prior to linguistic exposure. The likelihood $P(d|t)$ gives the probability of data being produced by a type. The posterior thusly yields a combination of prior expectations and data that captures a type's fit to the data. This, in turn, regulates the transmission of its lexicon. Putting these components together, the discrete selection dynamics are captured by $\hat{x}_i = \sum_j N_{ji} \frac{x_j f_j}{\Phi}$ (cf. Hofbauer & Sigmund 2003).

As mentioned above, the prior captures the learning bias of players. For simplicity we assume it to be solely dependent on lexica and not on a type's signaling behavior. In particular it biases learners for simpler semantic representations over more complex ones. In the following the prior encodes the intuition that the semantic representation of an upper-bound is more complex than a lack thereof. Since semantics are only implicitly represented through a lexicon's Boolean matrix, the bias is regulated by a parameter c that operates over cells of a lexicon.[2] As illustrated above by lexica L_a and L_b, we let s_1 stand for a "some but not all"-state and s_2 for an "all"-state. Accordingly, the prior biases players against lexica in which a message m holds true only of the former and not the latter, as in L_a. The literal meaning of English *some* corresponds to a message true of s_1 and s_2 in such a fragment. All other semantics are a priori equally probable. Thus, $P(t_i) \propto n - c \cdot r$ where n is the total number of states and r that of upper-bounded messages only true of s_1 in t_i's lexicon.

Analysis

We consider a total of 12 player types, obtained by pairing six lexica with literal and pragmatic signaling behavior. The lexica are specified in Table 1. Lexica L_1 to L_3 are not optimal for communication because they assign the same meaning to

[1] $U(x_i,x_j) = [U_S(x_i,x_j) + U_R(x_i,x_j)]/2$. $U_S(x_i,x_j)$ and $U_R(x_i,x_j)$, the expected utility of speaker type x_i and hearer type x_j and vice versa, are respectively $\sum_s P^*(s) \sum_m S_n(m|s;L) \sum_{s'} R_o(s'|m;L)\delta(s,s')$ and $U_S(x_j,x_i)$ for n and o being the reasoning level of i and j, and $\delta(s,s') = 1$ iff $s = s'$ and 0 otherwise.

[2] In principle this contrast could be made more precise with an adequate representational language, e.g., through measures over representational complexity. There is a growing effort to develop such empirically testable representational languages. For instance, the so-called *language of thought* has been put to test in various rational probabilistic models that show encouraging results (see e.g. Katz et al. 2008; Piantadosi et al. under review, 2012 and references therein). We think that our assumption is well-warranted as a working hypothesis and decide against such an enrichment given that the introduction of a larger framework would also require further assumptions and justifications.

$$L_1 = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad L_3 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$L_4 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad L_5 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad L_6 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

Table 1: Space of possible lexicon fragments considered.

all their messages. They were included to showcase the selection process for a larger hypothesis space. L_4 and L_5 are the crucial lexica that represent upper-bounded semantics and a lack thereof, respectively. Lastly, L_6 is similar to L_5 in that two messages are true of the same state but differs from it in assigning upper-bounded semantics to m_2.

In the following we focus our analysis on the contrast between literal and pragmatic types using lexica L_4 and L_5. Note in particular that the linguistic behavior of pragmatic L_4 and L_5 can be close to indistinguishable as both are able to convey upper-bounds. The former semantically, the latter pragmatically. Depending on the parameters, however, there might be slight differences between the probability with which speakers would (erroneously) use a semantically false description and that they would (erroneously) use a pragmatically suboptimal description. This contrasts with literal L_5, which lacks the means to convey an upper-bound by m_2. Differences between pragmatic L_4 and L_5 are expected to mainly depend on the learning bias. Things are less clear for literal L_5 contrasted with literal/pragmatic L_4. The former has a learning advantage but is expected to fare worse in terms of communicative fitness in virtue of ambiguous m_2.

The learning matrix Q was computed with data sequences of length 20. Pilot simulations showed that changes in sequence length influence the population in a predictable way: smaller values lead to more heterogeneous populations whereas larger ones lead to more pronounced differences. This is expected insofar as the likelihood that a sequence of length 1 was produced by any type is relatively uniform (modulo prior) whereas the likelihood of types with lexica L_1 - L_3 to produce, for instance, a sequence of 10 observations consistently with the same state-message combination is less likely than for pragmatic players using L_4 - L_6 or literal L_4. Thus, while noteworthy, sequence length has no direct bearing on the main contrast of interest. Similar considerations hold for α and λ – set to 1 and 50 in the following. Overall, lower rationality in λ or more pragmatic violations in α lead to a higher selection of lexica with semantic upper-bounds. The fitness of pragmatic behavior increases with higher λ-/α-values. In other words, these parameters level the functional contrast between L_4 and L_5. Due to space restrictions we leave a more detailed analysis of their interaction to future research and focus on the learning bias instead.

All simulations were initialized with an arbitrary distribution over types, constituting the population's first generation, and were run for 20 generations. This corresponds to a developmental plateau after which no noteworthy change was registered.

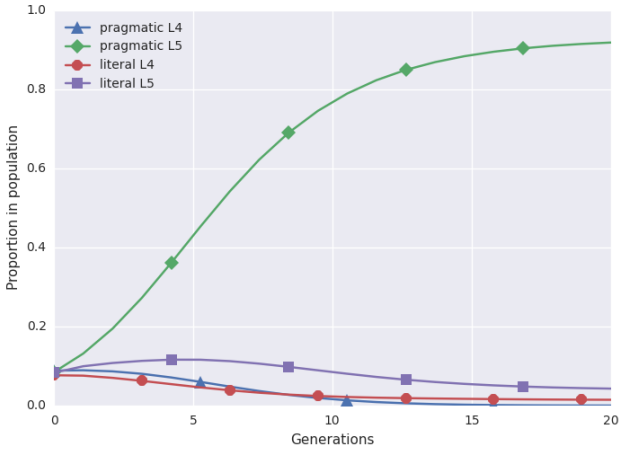


Figure 1: Mean proportions of target types across 20 generations of 1000 populations ($\alpha = 1, \lambda = 50, k = 20, c = 0.85$).

	0	.01	0.03	0.05	0.07	0.09	0.85
lit. L_1	.001	.002	.002	.004	.004	.002	.004
lit. L_2	.001	.002	.001	.001	.001	0	0
lit. L_3	.001	.003	.001	.002	.002	.001	.004
lit. L_4	.193	.193	.165	.177	.187	.142	.022
lit. L_5	.022	.022	.039	.04	.033	.049	.042
lit. L_6	.023	.023	.028	.027	.028	.025	.002
prag. L_1	.001	.001	.001	.002	.004	.003	.005
prag. L_2	.001	.001	.001	.001	.001	0	0
prag. L_3	.001	.002	.002	.002	.004	.002	.006
prag. L_4	.257	.234	.23	.209	.205	.208	0
prag. L_5	.249	.26	.29	.328	.329	.379	.914
prag. L_6	.25	.237	.24	.207	.202	.188	0

Table 2: Mean proportion of types in 1000 populations after 20 generations across different bias values c ($\alpha = 1, \lambda = 50, k = 20$).

Results. We report the mean results obtained from 1000 independent simulations for each set of parameters. Figure 1 illustrates the development of the target types over 20 generations with $c = 0.85$, i.e., a strong bias against semantic complexity. Figure 2 shows the effect of the learning bias after 20 generations for $c \in [0, 1]$. More detailed results for all types across a sample of c-values are presented in Table 2.

In sum, the results show that in the present setup a weak bias is sufficient to lead to a selection of L_5 over L_4. This effect increases with the bias' strength provided L_5 users are pragmatic. That is, such a learning bias may prevent the lexicalization of pragmatic inferences and explain the prevalence of L_5-like semantics. However, note that literal L_5 is underrepresented across all values of c. This highlights that, while the bias plays a major role for the contrast between L_4 and L_5, in itself it does not enable types that fail to convey an

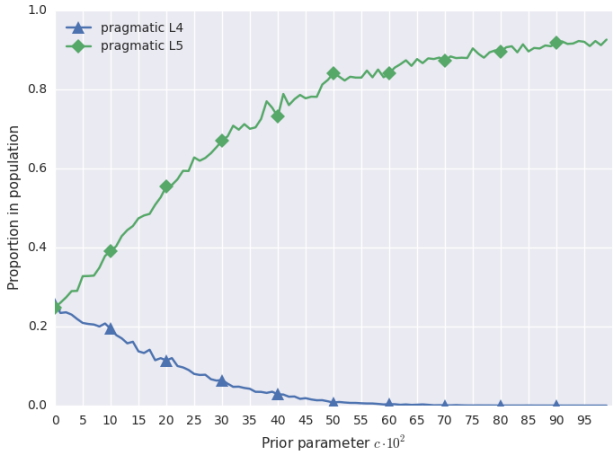


Figure 2: Mean proportion of pragmatic L_4 and L_5 types after 20 generations with $c \in [0,1]$ in 1000 populations ($\alpha = 1, \lambda = 50, k = 20$).

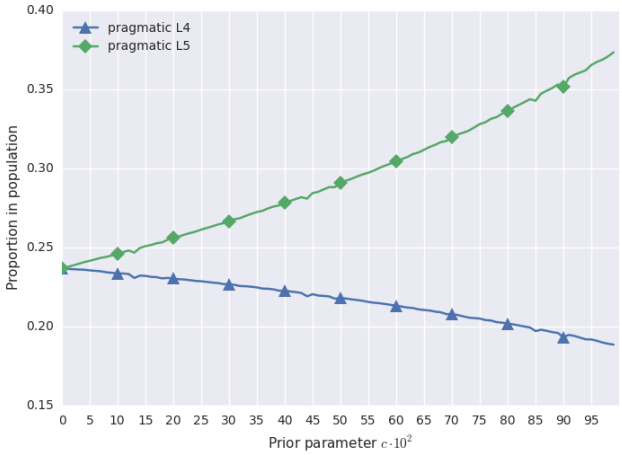


Figure 3: Mean proportion of pragmatic L_4 and L_5 types after 20 generations of parental teaching with $c \in [0,1]$ in 1000 populations ($\alpha = 1, \lambda = 50, k = 20$).

upper-bound to establish themselves in the population. This suggests pragmatic reasoning to be paramount to the selection of L_5-like semantics. That is, neither the learning bias nor functional pressure alone but their combination lead to the systematic lack of semantic upper-bounds in scalar expressions.

Discussion

Broadly speaking these results confirm our expectations that a lack of semantic upper-bounds coupled with pragmatic reasoning can overcome selective pressures and stabilize in a population provided there is a bias for simpler representations. This outcome is particularly encouraging in light of the potential additional advantages of a lack of semantic upper-bounds discussed above that were not considered here.

The model gives a justification for lexicalization patterns found in natural language, as well as for the failure to lexicalize certain pragmatic inferences. That is, while the puzzle raised by semantics is hard to explain by purely functional means, we suggest part of the answer to lie in learnability. Simpler semantic representations are more likely to be learned, and pragmatic reasoning can counteract functional disadvantages otherwise incurred. This result is of particular relevance for the longstanding assumption of a divide and interaction between semantics and pragmatics and offers an account of why (certain) pragmatic inferences are not part of the literal meaning of expressions. It furthermore leaves open the possibility of such inferences to fossilize when they do not compete against a lexical simplicity bias.

A virtue of this model is that it allows for analysis specific modifications and extensions. Straightforward extensions include larger hypothesis spaces, as well as larger or different lexicon fragments. Another worthwhile modification relates to possible disadvantages of pragmatic reasoning. We tacitly assumed pragmatic reasoning to come at no cost. However, there is experimental evidence for the assumption that the pragmatic derivation of upper-bounds costs effort and takes additional processing time (cf. Neys & Schaeken 2007; Huang & Snedeker 2009). This raises the question at which point such usage-based cost undercuts the learnability advantage of simpler semantic representations.[3]

Lastly, a crucial component of this model is the learning matrix Q together with its weighted variant N. The latter describes "communal teaching" where learning a lexicon is influenced by its proportional representation in the preceding generation. This diverges from the mean field dynamics of classical (dynamic) evolutionary game theory and iterated learning. Replacing the weighted matrix N by Q, i.e. not weighting the learning matrix, allows for an inspection of the model's predicted outcome for a more classical case of "parental teaching". Figure 3 shows the effect of the learning bias for this case (cf. Figure 2).

Note first that the main result of the prior driving selection of L_5 over L_4 holds here as well. However, the resulting populations are more heterogeneous than in the communal setting. The intuition behind communal learning was that learning is influenced by the proportion in which languages are used in a population. However, a stage of exposure to communal language use is currently not represented in the model but only implicit in the weights applied to the Bayesian learning matrix. This step may be regarded as conceptually problematic given that it assumes "implicit exposure" to the true distribution of types in the population. Notwithstanding, we see it as a first step in the right direction to capture what constitutes

[3] In the present setup this modification has a straightforward effect: A penalty for pragmatic signaling lowers the fitness of pragmatic types, to the advantage of literal types. However, the penalty needs to be substantial to counteract the functional advantage pragmatic L_5 has over all but L_4 together with its learning advantage.

an important factor in learning pressures. Further, it retains but strengthens the results obtained from a more standard learning setup. It is nevertheless important to highlight the weaker results obtained from parental learning as they suggest a missing piece to our overall account. Ideally one would expect the dynamics to favor a well-adapted lexicon over time even when only learning from parents, offering a stronger justification for the advantage of a lexicon without a population's involvement in the learning process. It is possible that a more complex setup where the speaker wants to convey not only bounded but also upper-unbounded states may shed light on this issue, as discussed in §2. Alternatively, a similarity bias over lexica may enable for a gradual adoption of particular languages. A more involved analysis and comparison between static learning matrices and weighted variants, as well as an explicit representation of exposure to linguistic data from the population raise many interesting technical and conceptual challenges we hope to see addressed in the future.

Conclusion

The development of natural languages is driven by complex intertwined pressures. Drawing from past insights we put forward a model that allows for a general and malleable integration of core aspects involved in this process. Chiefly, the model combines functional pressure, iterated Bayesian learning, and probabilistic speaker and hearer models. In particular, this analysis addresses longstanding issues concerning the semantics-pragmatics divide. It shows that when pressured for learnability and expressivity, the former force drives for simpler semantic representations inasmuch as pragmatics can compensate for them in language use. Consequently, semantic patterns can be explained in virtue of the linguistic behavior of their users and their complexity. Furthermore, the ease of acquisition of simpler semantics offers an answer to why natural languages do not lexicalize certain pragmatic inferences.

Acknowledgments

TB acknowledges funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 /*ESSENCE*/. MF gratefully acknowledges financial support by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63). RvR acknowledges support by the 'Language in Interaction' Gravitation project, funded by NWO.

References

Bergen, L., Levy, R., & Goodman, N. D. (to appear). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M. (2009). *Signal to act: Game theoretic pragmatics*. PhD thesis, University of Amsterdam.

Franke, M., & Jäger, G. (2014). Pragmatic back-and-forth reasoning. *Semantics, Pragmatics and the Case of Scalar Implicatures*, 170–200.

Gazdar, G. (1979). *Pragmatics, implicature, presuposition and logical form*. New York: Academic Press.

Grice, P. (1975). Logic and conversation. In *Studies in the ways of words* (pp. 22–40). Cambridge, MA: Harvard University Press.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, *40*(4), 479–520.

Horn, L. R. (1972). *On the semantic properties of logical operators in english*. Bloomington, IN: Indiana University Linguistics Club.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Huttegger, S. M., & Zollman, K. J. S. (2013). Methodology in biological game theory. *The British Journal for the Philosophy of Science*, *64*(3), 637–658.

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge: Harvard University Press.

Luce, D. R. (1959). *Individual choice behavior: a theoretical analysis*. Wiley.

Neys, W. D., & Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology*, *54*(2), 128–133.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028–8033.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (under review). *Modeling the acquisition of quantifier semantics: a case stuy in function word learnability.*

Steels, L. (2015). *The talking heads experiment: Origins of words and meanings*. Language Science Press.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge, MA, USA: MIT Press.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37–43.

A Computational Model of Perceptual Deficits in Medial Temporal Lobe Amnesia

Patrick Sadil (psadil@psych.umass.edu)
Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way
Amherst, MA, 01003 USA

Rosemary A. Cowell. (rcowell@psych.umass.edu)
Department of Psychological and Brain Sciences, University of Massachusetts, 135 Hicks Way
Amherst, MA, 01003 USA

Abstract

Damage to the Medial Temporal Lobe (MTL) impairs declarative memory and perception. The Representational-Hierarchical (RH) Account explains such impairments by assuming that MTL stores conjunctive representations of items and events, and that individuals with MTL damage must rely upon representations of simple visual features in posterior visual cortex. A recent study revealed a surprising anti-perceptual learning effect in MTL-damaged individuals: with exposure to a set of visual stimuli, discrimination performance worsened rather than improved. We expand the RH account to explain this paradox by assuming that visual discrimination is performed using a familiarity heuristic. Exposure to a set of highly similar stimuli entails repeated presentation of simple visual features, eventually rendering all feature representations equally (maximally) familiar and hence inutile for solving the task. Since the unique conjunctions represented in MTL do not occur repeatedly, healthy individuals are shielded from this perceptual interference. We simulate this mechanism with a neural network previously used to simulate recognition memory, thereby providing a model that accounts for both mnemonic *and* perceptual deficits caused by MTL damage using a unified architecture and mechanism.

Keywords: Neural Network; memory; visual perception; Medial Temporal Lobe; hierarchical object representations

Introduction

Revisiting the Classical, Modular Account of Memory and Perception Long-term, declarative memory has long been known to depend on medial temporal lobe (MTL) structures (Scoville & Milner, 1957; Squire & Zola-Morgan, 1991). However, evidence now suggests an additional role for MTL in other cognitive functions such as visual perception of objects and scenes (Barense et al., 2012; Bartko, et al., 2007; Lee et al., 2005).

A theory termed the Representational Hierarchical (RH) account explains both mnemonic and perceptual deficits caused by MTL damage (Bussey & Saksida, 2002; Cowell, Bussey, & Saksida, 2006, 2010). The RH account assumes that the ventral visual stream contains a hierarchical organization of representations that continues into MTL. Early stages of the pathway (e.g., V1, V2, V4) are assumed to represent simple visual features (e.g., color, orientation), and these simple features are brought together into conjunctions of increasing complexity in anterior brain regions. The hierarchy culminates in MTL where the conjunctions correspond to whole objects, scenes, or complex episodic events. The RH account claims that conjunctive representations in MTL are important whenever a cognitive task – perceptual or mnemonic – uses stimuli with overlapping features, such that individual feature representations in posterior regions provide ambiguous information (Bussey & Saksida, 2002).

Paradoxical Finding of an Exposure-induced Deficit in Visual Discrimination Performance Barense et al. (2012) documented a new and puzzling way in which MTL lesions impair visual discrimination. MTL amnesics and healthy controls were asked to judge whether pairs of simultaneously presented abstract stimuli were the same or different (Fig. 1). In the High Ambiguity condition, each pair of to-be-discriminated stimuli shared 2 out of 3 explicitly defined features, whereas in the Low Ambiguity condition, the items in a pair did not share any of the defined features. Participants were to declare 'mismatch' if any of the 3 features differed in a pair. Amnesic participants were unimpaired at discriminating Low Ambiguity objects, but discrimination of High Ambiguity objects deteriorated dramatically in the second half of trials in this condition (Fig. 2).

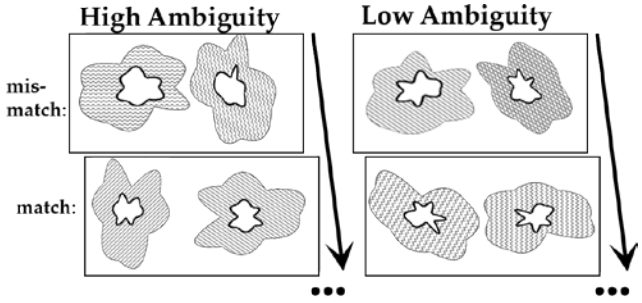


Figure 1: Experiment 3 of Barense et al. (2012). Stimuli were defined by 3 features: inner shape, outer shape, and fill pattern. High Ambiguity, mis-matching pairs shared 2 of these features, but Low Ambiguity pairs share 0.

Barense et al. explained their data in terms of the RH account: Individuals with MTL damage lack conjunctive representations of objects usually stored in perirhinal cortex (PRC), a structure in MTL. Objects are instead represented as individual simple features in posterior visual cortex. In the task of Barense et al., each stimulus is a unique conjunction of features, but the features comprising the stimuli repeat often over trials. After viewing many items, feature-level interference renders feature representations in posterior visual

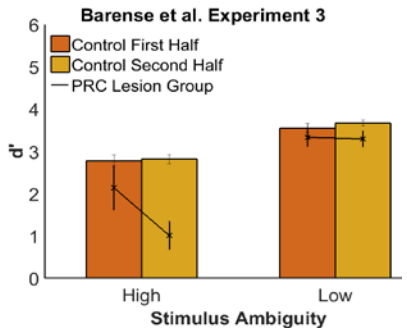


Figure 2: Data from Barense et al. (2012). Subjects with perirhinal (PRC) lesions were impaired at discriminating High Ambiguity stimuli in the second half of trials. Significance was assessed via Crawford's t-test for each Lesion participant separately (Control n=8; Lesion n=2).

cortex inadequate for solving difficult (High Ambiguity) discriminations. Control subjects resolve this interference with a unique conjunction in PRC for each stimulus, but if PRC is damaged, performance is impaired.

However, the *decrease* in MTL amnesics' performance with increasing exposure to the stimuli contrasts with well-established perceptual learning effects. Perceptual learning is usually explained by assuming that experience increases the separability of stimuli, either because stimulus representations become less overlapping (Saksida, 1999; Schoups, Vogels, Qian, & Orban, 2001; Yang & Maunsell, 2004), or because the weights via which representations influence decision-making are optimized (Kumano & Uka, 2013; Liu, Dosher, & Lu, 2015). But if discrimination relies on the separability of stimulus representations, and exposure differentiates representations, then damage to the brain areas containing those representations might be expected to *abolish any improvements* conferred by exposure, but it is not clear why brain damage should cause exposure to *hurt* performance. Put another way, even if individuals with MTL damage possess only feature representations, it is not apparent why exposure should cause feature representations to become *more* overlapping, rather than *less*.

How can this contradiction be resolved? In a previous neural network instantiation of the RH account (Cowell et al., 2006) we accounted for impairments in *recognition memory* induced by MTL damage by considering the *familiarity signal* evoked by stimulus representations in the brain. Exposure to many items sharing visual features entails frequent repetition of the features. Eventually, the representations of all features in posterior visual cortex appear familiar, causing individuals with MTL damage (who possess only feature representations) to perceive all items as equally familiar, impairing recognition memory. This surprising mechanism has been empirically supported (e.g, McTighe, et al., 2010). Here, we invoke the same mechanism to explain the *visual discrimination* impairments reported by Barense et al. To apply this account to visual discrimination, we make an assumption about the strategy used to solve the task: we propose that participants did not discriminate stimuli based on their representational overlap, but instead used a *familiarity heuristic* to decide whether two items were identical (Goldstein & Gigerenzer, 2002).

A Familiarity Heuristic Resolves the Paradox of Exposure-induced Discrimination Impairments We assume that, in difficult discrimination tasks like that of Barense et al. (2012), participants search for a mismatch between two stimuli. To do so, they visually scan back and forth between the stimuli; if switching from one item to the other elicits an impression of novelty relative to the item just examined, this is taken as evidence for a mismatch. That is, the new item appears "unfamiliar" to the extent that it differs from the item inspected immediately before switching. If switching elicits a decrease in familiarity that exceeds some threshold, the two stimuli are judged to mismatch. Hence, the model's signal for familiarity – the 'tunedness' of the stimulus representations (Cowell et al., 2006) – can be used to perform discrimination.

Just as in the memory simulations of Cowell et al. (2006), representations in the model can 'saturate' (i.e., reach maximum) in terms of familiarity. In a discrimination task in which all stimuli are similar, the stimulus features appear repeatedly, resulting in all feature representations becoming highly tuned, i.e., familiar. After sufficient interference, all feature representations possess the same, maximum, level of familiarity and can no longer be used to discriminate two objects via a familiarity heuristic. In contrast, because individual objects do not repeat across trials, familiarity for whole objects does not saturate and so remains useful for discrimination: conjunctive representations in PRC shield a person from perceptual interference. When these conjunctive representations are compromised by MTL damage, visual discrimination is impaired. A key tenet of the RH account is that memory and perception share common neural mechanisms. Our model embodies this by using a mnemonic signal (familiarity) to solve a visual perceptual task.

Description of the Model

Model Architecture We use the model of Cowell et al. (2006) with only minor modifications. The network contains a PRC layer and a layer corresponding to posterior ventral visual stream (Fig. 3). Visual objects are instantiated as 8-dimensional vectors, given as input to the model. We assume that posterior regions represent simple conjunctions of two individual visual dimensions, so the Posterior layer is divided into 4 grids: nodes in each posterior grid receive 2 input dimensions and combine them into a simple conjunction, termed a 'feature'. In contrast, because PRC is assumed to represent whole objects, all 8 input dimensions converge into a single conjunction in the PRC layer. Thus, nodes in the PRC layer contain a unique, conjunctive representation of a four-featured object, whereas the Posterior layer represents those four, 2-dimensional features separately. We simulate PRC damage by removing the PRC layer.

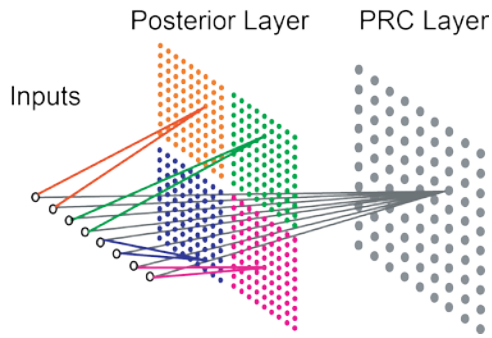


Figure 3: Model architecture. The PRC layer is a single Kohonen grid, representing an object as a unique conjunction. The Posterior layer contains 4 Kohonen grids, each representing a 2-dimensional visual 'feature'.

All model layers are constructed from Kohonen grids, which mimic properties of cortex, including information processing mechanisms such as Hebbian learning and lateral inhibition (Kohonen, 1984). A Kohonen grid (or self-organizing map) is trained by successively presenting stimulus inputs and incrementally adapting the weights of the grid units on each presentation. As the grid learns, its stimulus representations are sharpened or 'tuned': A novel stimulus elicits a broadly distributed pattern of activity, whereas a stimulus that has been repeatedly presented elicits a highly selective activation pattern with a peak over one set of units, relative to the activation levels elsewhere across the grid. In our model, the selectivity of the activation indexes familiarity (Cowell et al., 2006; Norman & O'Reilly, 2003).

Each dimension of an input stimulus can take one of four values in a given stimulus: 0.05, 035, 0.65, or 0.95. This scheme yields $4^8 = 65{,}536$ unique object stimuli, which are represented holistically on the PRC layer, but only $4^2 = 16$ unique, 2-dimensional features on each Posterior grid. This scheme reflects a key assumption of the model: that there is a vast number of possible visual objects in the world, but they are composed from a small number of visual elements.

Initialization and Pretraining All grids contained $200x200$ nodes whose weights were initialized with random values between 0 and 1. Networks were pre-trained for 500 cycles according to the standard Kohonen learning rule,

$$w_i(t+1) = w_i(t) + f(r,t) * (stim - w_i(t)) \quad (1)$$

and,

$$f(r,t) = \eta(t) * v(r,t), \quad (2)$$

where w_i refers to the weights of node i, t is the current cycle, *stim* is stimulus input, $\eta(t)$ is the learning-rate, r is the city-block distance of node i from the most strongly active (winning) node, and $v(r,t)$ is a neighborhood function that scales the learning rate (Eq. 3).

In the pre-training phase, both η and v decrease with each cycle. The neighborhood is defined by a Gaussian function:

$$v(r,t) = \exp\left(-\left(\frac{r}{G(t)}\right)^2\right) \quad (3)$$

where, G(t) = 0.5 + 10 t^{-B}, and B is a constant determining the rate of shrinkage of the neighborhood function. The learning rate decreases as $\eta(t) = t^{-A}$, where the constant A determines the rate of decrease.

In each pre-training cycle, the network sees a different, unique stimulus. For simulations of the Barense et al. (2012) experiments, η and G are constants fixed at the values of the final training cycle ($\eta = \eta(500)$; $G = G(500)$).

Measuring Familiarity Activation (a) of nodes was determined by the sigmoidal function:

$$a_i = 1 \Big/ \left(1 + \exp\left(-k * \ln\left(\frac{1}{dist}\right)\right)\right) \quad (4)$$

where k is a constant that determines the steepness of the sigmoidal function, $dist$ is the mean squared error between a node's weights (w_i) and the stimulus input vector.

The familiarity of a stimulus is given by the selectivity of its activation pattern, calculated separately in each grid. Selectivity is the activation of the peak (the summed activation of the winning node and its nearest 4 neighbors) divided by the summed total activation of the grid. Thus, via normalization, higher familiarity corresponds to more 'tuned' representations. Stimulus familiarity is measured separately in each grid in the network, yielding a single, object-level familiarity score from the PRC layer and four separate feature-level familiarity scores in the posterior layer; when comparing two stimuli, the posterior feature-level familiarity scores are compared separately for each pair of features. For Control networks, this includes 4 Posterior grids and 1 PRC grid; for Lesioned networks, only 4 Posterior grids. In all simulations, $k = 0.08$, $B = 0.3$, and $A = 0.6$.

Simulating Visual Discrimination Behavior

Fixations In Barense et al. (2012), participants decided whether two stimuli were the same or different. Eye-tracking data from control subjects revealed a higher ratio of within-stimulus to between-stimulus fixations in High Ambiguity trials (~1.2) than in Low Ambiguity trials (~0.6). Barense et al. conjectured that this reflected a greater tendency to 'bind' features together in the High Ambiguity condition. Accordingly, we hypothesized that differential fixation ratios might make an important contribution to task performance (e.g., sampling stimuli with a higher within: between ratio may enable more reliable PRC representations) and aligned model parameters with these empirical data. In addition, on trials declared as a 'match', participants made about 25 and 20 fixations for High and Low Ambiguity, respectively.

In the model, one stimulus fixation entails 20 encoding cycles (Eq. 1). Stimuli are 'sampled' via fixations, using a probabilistic rule for switching back and forth between two items in a pair, with switch probability derived from the empirical within:between ratio for each experimental condition (1.2 or 0.6). On any trial, the maximum number of fixations (i.e., the point at which the search for a mismatch terminates and a 'match' is declared, see below) is 25 and 20 in the High and Low Ambiguity conditions, respectively (since the stopping point is somewhat arbitrary and not specified by our theory, we used empirical parameters to determine this).

Discrimination Decisions On each trial, two stimuli are presented. One is arbitrarily selected first (Item A) and the model samples it via successive fixations until switching to the other stimulus, Item B. Upon switching, the model assesses evidence for a 'mismatch' by computing a novelty score and comparing it to a threshold. The novelty score is a measure of 'familiarity change' obtained by taking the familiarity of Item A and subtracting the familiarity of the Item B. When the two items are identical, the novelty score is zero; when they are different, Item B should have slightly lower familiarity than Item A, yielding a positive novelty score. If the novelty score in any individual grid (4 Posterior grids or, when available, PRC) exceeds the threshold, the items are declared to 'mismatch'. At each switch between the two stimuli, if the network finds no evidence for a mismatch, the network continues fixating. In the next comparison, Item B serves as the previously inspected stimulus and Item A as the newly fixated stimulus. Comparisons proceed until either a mismatch is declared or the maximum number of fixations is reached, whereupon a 'match' is declared.

Criterion shift We assume that participants adjust their decision rule as their stimulus representations adapt. That is, if participants begin to perceive *all* items as looking more similar, they require less evidence to declare that two items are mismatching. To effect this, the threshold value of novelty required to declare a pair of items as mismatching (i.e., the *decision criterion*) is allowed to shift by setting it equal to the average of the novelty ('familiarity change') score on the previous six trials. In addition, for each decision, noise drawn from the uniform distribution ($\pm 1e^{-6}$) is added to the threshold. As familiarity change scores decrease, this noise swamps the discrimination signal present in that score. The starting threshold value (i.e., threshold on Trial 1) is set to 2x the maximum noise: $2e^{-6}$.

Simulation 1

Target Empirical Data In *Experiment 3* of Barense et al. (2012), participants indicated whether two simultaneously presented visual stimuli were a match or a mismatch. Stimuli were trial-unique items constructed of 3 features (Fig. 1). In Low Ambiguity trials, the pairs shared none of these 3 features. In High Ambiguity trials, the items shared 2 out of 3 features. Individuals with PRC damage performed similarly to controls at discriminating Low Ambiguity pairs, but for High Ambiguity pairs their performance was intact at first but fell sharply in the second half of the task (Fig. 2).

Stimuli and Task All stimuli were trial unique. Low Ambiguity stimulus pairs shared no 2-dimensional features, whereas High Ambiguity pairs shared 3 out of 4 features (we used 4 total features for consistency with Cowell et al., 2006). Owing to assumptions governing stimulus construction (see *Model Architecture*, above) there are many possible unique stimuli, but the features comprising them appear repeatedly. In addition, we further constrained the feature set to reflect the high degree of feature-overlap in Barense et al. by constructing stimuli (unique 4-featured objects) using only 6 out the 16 possible 2-dimensional features for each Posterior grid (where a 'feature' is a conjunction of 2 input dimensions). This yielded 6^4 total possible objects, with high feature-overlap among them. As in Barense et al., each condition contained 36 'match' and 36 'mismatch' trials.

Results Networks with no PRC layer ('PRC Lesion Group'), like humans with PRC damage, were impaired relative to controls at High but not Low Ambiguity. Critically, the impairment was worse in the second half of trials (Fig. 4). We do not report statistics on simulated data because the results are highly reliable such that variance scales arbitrarily with the number of networks run. Instead, we focus on qualitative patterns, which match those of the patient data: the interaction between Lesion Group, Stimulus Ambiguity and First/Second Half.

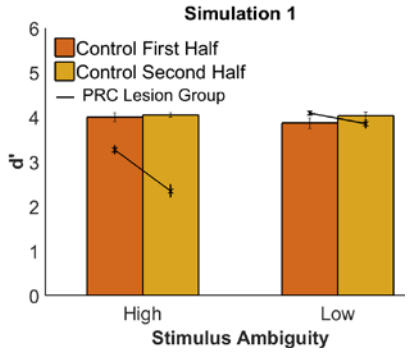


Figure 4: Figure 4: Simulated data for Experiment 3 of Barense et al. (2012). Compare to the empirical data in Figure 2. n = 50 networks per group.

Simulation 1 Discussion The model's simulation of a discrimination deficit for High Ambiguity stimuli in the second half of trials, following PRC damage, hinges on three assumptions: (1) participants solve the task using a familiarity heuristic; (2) task stimuli are unique but contain many low-level features that repeat over trials such that all features eventually appear familiar; (3) stimuli are represented in PRC as whole conjunctions but in posterior regions as individual features: when PRC is missing, discrimination performance is impaired once all features are maximally tuned.

Performance in lesioned networks relies upon posterior feature representations. At the start of the task, individual features are not highly tuned. On each new trial, the features of the first stimulus inspected increase in familiarity. When the network switches to inspect the second stimulus, if that stimulus is not identical to the first, its features are lower in familiarity and a mismatch is correctly declared. However, after many trials, all features have been repeatedly encoded by the network. Now, at the start of a trial, there can be no increase in familiarity when the network inspects the first stimulus. When it switches to the second stimulus, even if that item differs from the first, its features have equal familiarity – the maximum value. At this point, a familiarity heuristic can no longer detect mismatching stimulus pairs. The feature-level

interference has more effect at High than Low Ambiguity because High Ambiguity pairs share 3 out of 4 features whereas Low Ambiguity pairs share no features. Networks search for *any* pair of features across the two stimuli that differ. Since there are 4 mismatching features in Low pairs, the network is more likely to discover at least 1 feature that has not yet reached maximum familiarity.

Performance in control networks is maintained throughout the task because of conjunctive representations in the PRC layer. Individual stimuli are trial-unique, i.e., whole conjunctions are never repeated, so representations in PRC never reach maximum familiarity. At the start of each new trial, the PRC representation for the first stimulus inspected increases in familiarity during the inspection. When the model switches to inspect the second stimulus, if the second stimulus differs from the first, the second will exhibit lower familiarity and the pair will be declared to mismatch.

Simulation 2

Target Empirical Data In *Experiment 4* of Barense et al. (2012), all subjects completed 3 blocks: Low Interference, then High Interference, then a *second* Low Interference block. High Interference was the same as High Ambiguity in *Experiment 3*: pairs of abstract stimuli shared 2 out of 3 features. In Low Interference blocks, two-thirds of trials contained a pair of color photographs of real-world objects. Photo images were trial unique and shared few low-level features with the abstract stimuli of critical comparison trials (Fig. 5). *Experiment 4* replicated the result of *Experiment 3*: MTL-damaged patients showed impaired discrimination at High Interference, but not at Low Interference, even in block 3 of the test session (Fig. 6, left). This suggested that the impairment seen in *Experiment 3* was not due to fatigue.

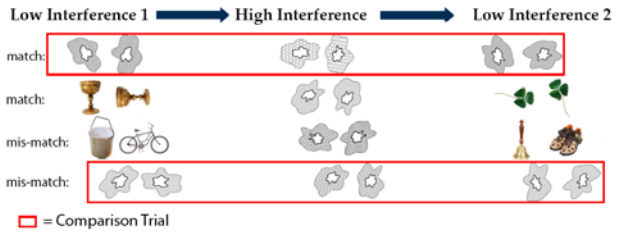


Figure 6: Experiment 4 of Barense et al. (2012). Low Interference blocks used photo stimuli in 2/3 of trials, which shared few features with stimuli on critical comparison trials.

Stimuli and Task As in Simulation 1, abstract stimuli were modeled by using only 6 of the 16 possible stimulus features on each Posterior grid to construct four-featured stimulus wholes, yielding high feature-overlap among stimuli. To reflect the assumption of Barense et al. that abstract stimuli shared no 2-dimensional features with photo stimuli, we constructed photo stimulus inputs from the remaining 10 abstract features (i.e., a non-overlapping set of features).

Networks performed 3 discrimination blocks. A block contained 88 trials, in which every third trial was a critical comparison pair of abstract stimuli (15 matching, 15 mismatching). In High Interference blocks the remaining 58 trials contained extra pairs of High Ambiguity objects. In Low Interference blocks the remaining 58 trials contained pairs of 'photo' stimuli. As in Barense et al., performance was judged on critical trials, which occurred at every third position in the sequence and contained a pair of stimuli sharing many (3 out of 4) features. The difference between High and Low Interference was that, for Low, the trials interposed between critical trials contained items sharing no features with critical-trial stimuli whereas, for High, interposed trials contained items similar to critical-trial stimuli.

Results Networks with no PRC layer were able to discriminate at the level of Control networks in Low Interference blocks (the first and third blocks), but their performance was impaired in High Interference (second) block (Fig. 6), mirroring the results of Barense et al. (2012).

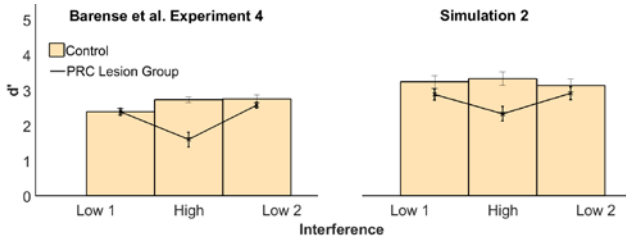


Figure 5: Barense et al., Experiment 4: empirical data (left) and model simulations (right). n = 50 networks per group.

Simulation 2 Discussion The same mechanism that produced the effects of PRC lesions in Simulation 1 also drives PRC lesion deficits in Simulation 2. At High Interference, because all trials contain the same class of stimuli, stimulus features appear repeatedly over trials, and posterior feature representations reach maximum familiarity. Once this occurs, there is no increase in familiarity when a network inspects a new stimulus at the start of a trial. Consequently, a network without a PRC layer cannot detect 'novelty' (a drop in familiarity) upon switching to the other stimulus.

At Low Interference, two-thirds of trials contain photos composed of different features than the critical-trial stimuli. The critical-trial stimulus features repeat too infrequently for their representations to reach maximum familiarity, hence lesioned networks remain unimpaired.

Discussion

Barense et al. (2012) reported a striking perceptual deficit in patients with MTL damage: the *accumulation* of perceptual experience impairs visual discrimination. This result is paradoxical because perceptual discrimination typically *improves* with exposure to the stimuli. Barense et al. argued that MTL-lesioned patients suffer from accumulated feature-level interference, which − in the absence of conjunctive MTL representations − cannot be overcome by feature representations in posterior visual cortex. Although we concur with this explanation, we suggest that it is incomplete.

Standard theories of perceptual learning claim that experience improves discrimination performance by reducing the

overlap between stimulus representations (i.e., training increases representational separation). In such theories, the assumed mechanism for visual discrimination is that discriminability is proportional to the overlap between representations (Saksida, 1999; Schoups et al., 2001). But such proportionality could not account for the data of Barense et al.: if exposure separates representations, even feature-based discrimination should improve with exposure because even feature representations should get *less* overlapping with training. To explain why the performance of MTL patients in Barense et al. got *worse* after exposure to the stimuli, a theory based on representational overlap would require the counter-intuitive assumption that while cortical representations of stimuli underlying perceptual learning often become *less* overlapping with exposure, posterior feature representations in this task became *more* overlapping.

In our account of the data, we eschew representational overlap as the mechanism for visual discrimination. Instead, we take the account offered by Barense et al. (i.e., amnesics suffer from the loss of conjunctive MTL representations) and combine it with a less commonly invoked discrimination mechanism: a *familiarity heuristic*. Under this account – as in prior instances of the RH account applied to *memory* (Cowell et al., 2006; McTighe et al., 2010) – representations of features, but not conjunctions, reach maximum familiarity as interference accrues. Thus, after MTL damage, perceptual experience impairs *visual discrimination performance.*

Before concluding, we clarify two important points. First, although we simulate only two layers in the ventral pathway, this is a subset of the full hierarchy of representations, which includes simpler layers prior to our model's posterior layer and more complex layers such as hippocampus after PRC. Other tasks may require other layers. For example, a discrimination task involving whole objects that repeat would require hippocampal representations capable of combining objects with (e.g.) context, to shield participants from object-level interference that would detrimentally affect PRC representations. Second, we do not suggest that a familiarity heuristic must be used in all discriminations, including easy tasks in which stimuli differ on the basis of simple, salient features such as color: such tasks could be solved by a more standard discrimination mechanism of assessing representational overlap. Our model suggests only that a familiarity heuristic is used for difficult discriminations between similar stimuli such as used in Barense et al.

In sum, we simulated the visual discrimination performance of MTL amnesics with a neural network model that differs only in trivial details from the recognition memory model of Cowell et al. (2006). To our knowledge, this is the first computational model to simulate both mnemonic *and* perceptual deficits caused by MTL damage using a unified architecture and mechanism.

Acknowledgements

We thank Morgan Barense for sharing her data and stimuli, and for helpful feedback. This work was supported by NSF CAREER Award 1554871 to R.A.C.

References

Barense, M. D., Groen, I. I. A, Lee, A. C. H., Yeung, L. K., Brady, S. M., Gregori, M., Kapur, N., Bussey, T. J., Saksida, L. M., Henson, R. N. A. (2012). Intact memory for irrelevant information impairs perception in amnesia. *Neuron*, 75(1), 157–167.

Bartko, S. J., Winters, B. D., Cowell, R. A., Saksida, L. M., & Bussey, T. J. (2007). Perirhinal cortex resolves feature ambiguity in configural object recognition and perceptual oddity tasks. *Learning & Memory*, 14(12), 821–32.

Bussey, T. J., & Saksida, L. M. (2002). The organization of visual object representations: a connectionist model of effects of lesions in perirhinal cortex. *The European Journal of Neuroscience*, 15(2), 355–364.

Cowell, R. A, Bussey, T. J., & Saksida, L. M. (2006). Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *The Journal of Neuroscience*, 26(47), 12186–12197.

Cowell, R. A, Bussey, T. J., & Saksida, L. M. (2010). Functional dissociations within the ventral object processing pathway: cognitive modules or a hierarchical continuum? *Journal of Cognitive Neuroscience*, 22(11), 2460–2479.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1), 75–90.

Kohonen, T. (1984). Self-Organizing Maps (Third Edit). Berlin-Verlag: SPRINGER.

Kumano, H., & Uka, T. (2013). Neuronal mechanisms of visual perceptual learning. *Behavioural Brain Research*, 249, 75–80.

Lee, A. C. H., Bussey, T. J., Murray, E. A., Saksida, L. M., Epstein, R. A., Kapur, N., Hodges, J. R, Graham, K. S. (2005). Perceptual deficits in amnesia: challenging the medial temporal lobe "mnemonic" view. *Neuropsychologia*, 43(1), 1–11.

Liu, J., Dosher, B. A., & Lu, Z. (2015). Augmented Hebbian reweighting accounts for accuracy and induced bias in perceptual learning with reverse feedback. *Journal of Vision*. 15(2015), 1–21.

McTighe, S. M., Cowell, R. A, Winters, B. D., Bussey, T. J., & Saksida, L. M. (2010).Paradoxical False Memory for Objects After Brain Damage. *Science*, 330(6009), 1408–1410.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review*, 110(4), 611–646.

Saksida, L. M. (1999). Effects of similarity and experience on discrimination learning: a nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 25(3), 308–323.

Schoups, A, Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6846), 549–553.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, 20(11), 11–21.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. Science, 253(5026), 1380–6.

Yang, T., & Maunsell, J. H. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *Journal of Neuroscience*, 24(7), 1617–1626.

Linguistic input is tuned to children's developmental level

Daniel Yurovsky
yurovsky@stanford.edu
Department of Psychology
Stanford University

Gabriel Doyle
gdoyle@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Children rapidly learn a tremendous amount about language despite limitations imposed on them by their developing cognitive abilities. One possible explanation for this rapid learning is that caregivers tune the language they produce to these limitations, titrating the complexity of their speech to developmentally-appropriate levels. We test this hypothesis in a large-scale corpus analysis, measuring the contingency between parents' and children's speech over the first 5 years. Our results support the linguistic tuning hypothesis, showing a high degree of mostly parent-led coordination early in development that decreases as children become more proficient language learners and users.

Keywords: Language acquisition, cognitive development, computational modeling

Introduction

Children acquire a tremendous amount of language in their first few years of life. By the time they are able to run down the street, typically developing children have over a thousand words in their productive vocabularies (Mayor & Plunkett, 2011). They can combine these words to produce new, meaningful multi-word utterances (Lieven, Salomo, & Tomasello, 2009). They can even learn new words from just the syntactic constructions in which they occur (Yuan & Fisher, 2009). What explains this rapid developmental progression?

Young children have access to a surprising range of learning mechanisms. By 8-months, infants can use distributional properties of language to segment discrete words from continuous speech (Saffran, Aslin, & Newport, 1996), and by 12 months can use these same kinds of cues to learn ordering regularities in artificial grammars (Gomez & Gerken, 1999) and mappings between words and objects (Smith & Yu, 2008). But while these and other competencies are available early, the amount of information that children can actually learn in lab experiments often strikingly limited. For instance, children's learning of new words from distributional properties of language is highly constrained by their developing attentional and memory systems (Vlach & Johnson, 2013).

How can children learn so quickly when their learning abilities are so constrained? One possibility is that the speech that children hear in the world differs systematically from speech used to test their learning in the laboratory. Indeed, the language that parents produce to their children—across a variety of levels and structures—appears to contain many redundant cues that facilitate learning (Gogate, Bahrick, & Watson, 2000; Thiessen, Hill, & Saffran, 2005; Yurovsky, Yu, & Smith, 2012). However, in some cases child-directed speech appears systematically different in ways that do not support learning. For instance, child-directed speech typically contains simpler and less variable syntactic structures. This simplicity helps children learn simple structures, but also makes it harder for them to learn more complex ones (Montag, Jones, & Smith, 2015; Montag & MacDonald, 2015). The input that is best for children's learning thus depends on what they already know, and what they are trying to learn next.

The explanation for rapid language acquisition may thus be neither in the learner nor in the input, but in the coordination between learner and input: Parents may tune the complexity of their language to the developing abilities and needs of their children. This *linguistic tuning* hypothesis is intuitively appealing, and was supported by some early evidence (Snow, 1972). However, two pieces of contradictory evidence diminished initial enthusiasm for it. First, parents do not appear to use simpler words when speaking to younger children (Hayes & Ahrens, 1988). Second, parents rarely show sensitivity to their children's syntactic errors by correcting them, and children are resistant to the few corrections they get (Brown & Hanlon, 1970; Newport, Gleitman, & Gleitman, 1977).

More recent work has suggested other—perhaps more subtle—ways in which parents might tune the language they produce in response to children's developing knowledge and learning mechanisms. Although they do not correct syntactic errors, parents are much more likely to repeat and reformulate children's ungrammatical utterances, providing a potential corrective signal (Hirsh-Pasek, Treiman, & Schneiderman, 1984; Chouinard & Clark, 2003). And although parents' vocabulary choices may be due at least in part to the content they need to convey, there may still be ways for adults to tailor the structure of their utterances to children's perceived competence. In the current paper, we pursue this question, asking whether parents choose their non-content words in a manner contingent on their children's developmental level.

To test whether parents tune their speech to that of their children, we conduct a large-scale corpus study using *linguistic alignment*, a measure of how much speakers change the way they talk to accommodate their conversational partners. Critically, alignment is a local measure: High alignment results not from choosing words that are simpler overall, but from choosing words that are more contingent on one's conversational partner, and thus easier to process in context (cf. Hayes & Ahrens, 1988). We predict that caregivers should alter their levels of alignment across development, aligning more to younger children who need more linguistic support.

Linguistic alignment

When we use language to communicate, we use the words we say to convey the message we intend. Some of these words will be critical for getting the message across. Consider the

Naima:	Eating **that**. Eating **some of that**.
Mom:	**Some of this**? **You** know what **that** is? **It** is sweet potato.
Naima:	**I** am **a** bear **that** eats.
Mom:	**You're a** bear **that** eats what? What do **you** eat little bear?
Naima:	Fresh pear

Table 1: An exchange between Naima (20-mo.) and her mom in the Providence Corpus. Bolded words are LIWC category members included in the model (Pennebaker et al., 2007).

conversation in Table 1, taken from the Providence corpus (Demuth, Culbertson, & Alter, 2006). In her first response, Naima's mom has little choice but to say "sweet potato" if she wants to inform Naima that they are eating sweet potato. However, she could perfectly well have left out the words "some," "that," and "this," or exchanged them for others and still conveyed the identity of the food on Naima's plate.

Speakers are influenced in their production choices by a variety of factors, ranging from the phonological to the sociolinguistic. We focus here one particular reason for these choices: Contingency on a conversational partner. When we talk, we tend to re-use use each-other's expressions, aligning to each other. This kind of alignment appears to be a pervasive property of human social interaction and linguistic communication (Giles, Coupland, & Coupland, 1991; Garrod & Pickering, 2004). Further, alignment appears to be useful, facilitating fluent processing of speech, and increasing the probability of successful communication and accomplishment of joint goals (Ireland et al., 2011; Fusaroli et al., 2012). Critically, alignment is directional: Even in the same conversation, some speakers will align more than others. For instance, alignment varies across a social hierarchy, with less powerful speakers aligning more to powerful speakers (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013). Thus, linguistic alignment can measure a speaker's effort to coordinate with a conversational partner. We leverage this property to measure the extent to which parents are altering the way they speak to coordinate with their developing children.

We explicitly focus on the words that are least critical for conveying the content of the message. If Mom produces "sweet potato" after Naima does, we know only that they are discussing the same object. However, if Mom produces a function word like "of" after Naima does, she is more likely to be using a similar structure that facilitates Naima's language processing. To capture this idea, we chose as our target words a set of 676 words falling into 14 categories identified as strictly non-topical style dimensions (Linguistic Inquiry and Word Count; Pennebaker et al., 2007). We perform our analyses at the level of categories, to capture both exact repetitions of a conversational partner's words and also reformulations and expansions (Chouinard & Clark, 2003). For example, in Mom's response to Naima, both "this" and "that" would count as instances of the impersonal pronoun category (Table 2). These 14 LIWC categories have been used by us and others in previous work examining alignment in a variety of contexts from social media to supreme court proceedings (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012; Doyle, Yurovsky, & Frank, 2016).

Model

The linguistic tuning hypothesis predicts that parents should tune the complexity of their language to the developing cognitive and linguistic capacities of their children. It thus predicts that parents should show high alignment to their young children, but should gradually reduce their levels of alignment as children develop. To test this hypothesis, we extended the Hierarchical Alignment Model introduced by Doyle et al. (2016). Our model estimates the extent to which a speaker's use of function words is influenced by their conversational partner's use of words in the same category. In Mom's first response to Naima, for instance, she uses the quantifier "some" and the indefinite pronoun "this" (Table 1). Would Mom have been less likely to use these words if Naima had not produced a quantifier and an indefinite pronoun in her previous utterance? And would she be less likely if Naima were older?

The model attempts to predict on an utterance-by-utterance level whether a speaker will produce a word belonging to each of the 14 function word categories. The probability of producing a word is controlled by two independent factors: (1) the speaker's baseline probability of producing that word, and (2) the speaker's tendency to align to their conversational partner, producing words from categories that their partner just produced. Thus, the primary computation in the model is essentially the same as standard logistic regression, which predicts a binary response (production) from a linear combination of predictors (word frequency and alignment probability). The remaining machinery of the model allows frequency and alignment estimates to be pooled hierarchically across categories, speakers, and ages in a principled way.

Model Details

The full graphical representation of our model is shown in Figure 1. The model operates over a representation of utterances as binary vectors indicating the presence or absence of each of the 14 LIWC categories. The probability of producing each category in each utterance is then predicted via two parameters: (1) The speaker's baseline probability for using that word category (η^{base}), and (2) The speaker's change from this baseline due to interacting with the listener (η^{align}). For replies to utterances not containing a category, the category's parameter is produced by taking the inverse logit of its baseline log odds $\left(logit^{-1}\left(\eta^{base}\right)\right)$. If the utterance follows an utterance that contains the category, we say that its probability of production is the inverse logit of the sum of the baseline and alignment log odds $\left(logit^{-1}\left(\eta^{base}+\eta^{align}\right)\right)$.

Because the LIWC categories vary widely in the production frequencies, we draw the log odds of each from an independent uninformative prior $(Uniform(-5,5))$, which covers more than the range of observed probabilities without putting too much mass on extremely large or small val-

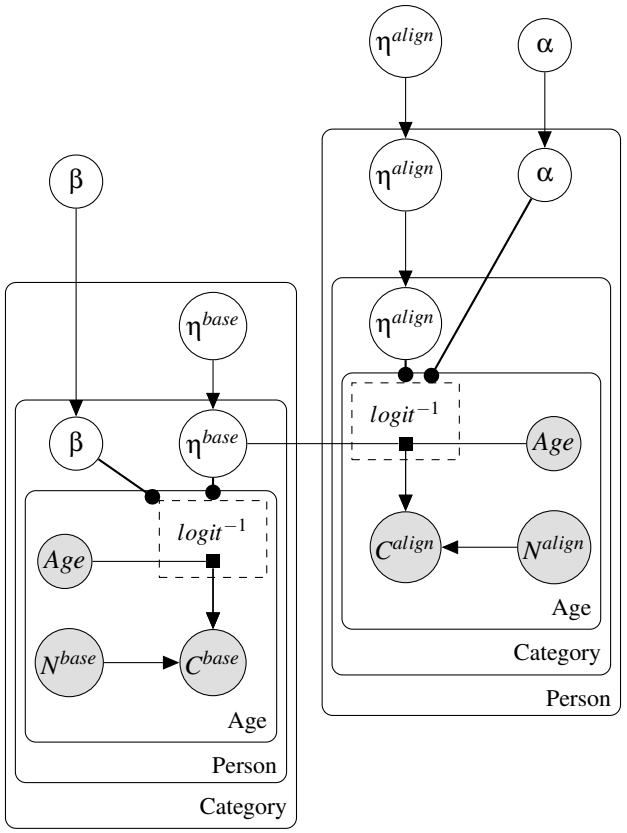


Figure 1: The Hierarchical Alignment Model (HAM) we use to analyze linguistic alignment in CHILDES. Speaker's word choices are modeled as having two influences: (1) Their baseline probability of using each word category (η_{base}), and (2) their increase from baseline due to their conversational partner's use of each category in the previous utterance (η_{align}).

ues. We put a conservative prior on alignment, drawing $\eta^{align} \sim Normal(0, .5)$, regularizing it strongly towards zero.

To pool data across participants for robust estimation, we estimate all parameters hierarchically. We say that there is a population-level of alignment, which generates speaker-levels of alignment, which generate category-levels of alignment. This decision allows us to make principled inferences both about how much parents align to their children in general, and about how much specific parents align to their specific children. For baseline probabilities, we instead nest people within categories as the empirical baseline production probabilities vary widely across categories.

Finally, we let both the probability of using any of these categories (β), and the probability of aligning (α) vary over development. Inspection of posterior parameter estimates in a model with independent age estimates showed a linear relationship for both in this age range, so for simplicity we model β and α as linear scalars of η^{base} and η^{align} respectively.

Using this model, we can test two distinct hypotheses about the way that parents might coordinate to their children's developmental level. First, if β is non-zero, then we can infer

Category	Examples	Adult	Child
Article	*a, an, the*	.31	.13
Certainty	*always, never*	.05	.01
Conjunction	*but, and, though*	.22	.06
Discrepancy	*should, would*	.11	.04
Exclusive	*without, exclude*	.12	..04
Inclusive	*with, include*	.21	.07
Indefinite pronoun	*with, include*	.45	.19
Negation	*not, never*	.21	.11
Preposition	*to, in, by, from*	.38	.14
Quantifier	*few, many*	.13	.04
Tentative	*maybe, perhaps*	.11	.03
1st person singular	*I, me, mine*	.05	.04
1st person plural	*we, us, ours*	.09	.02
2nd person pronoun	*you, yourself*	.33	.06

Table 2: LIWC categories for estimating linguistic alignment, with examples and production probabilities for both adults and children in CHILDES.

that speakers change their baseline likelihood of producing the function words in the 14 LIWC categories over the course of children's development. If parents' β is positive, they are using simpler words earlier in development. Second, if α is non-zero, then we can infer that speakers change their level of alignment over development. If parents' α is negative, they are aligning more to their younger children.

Analysis

Data

To maximize the power and generalizability of our analysis, we selected all English-language transcripts available in CHILDES containing conversations between a parent and a target child who was 12–60-mos. of age (MacWhinney, 2000). This yielded 3,851 total transcripts across 417 unique children. The number of transcripts per child varied widely, ranging from 1 ($n = 164$) to 440 ($n = 1$), with a median of 26.

For each transcript, we first combined all successive utterances from the same speaker into one utterance. We then transformed each utterance into a binary vector with a value for each of the 14 LIWC categories (Pennebaker et al., 2007). We then formatted the utterances as a series of message-reply pairs, in which each utterance was treated as a reply to the previous utterance, and a message for the next utterance.

Method

For each pair of speakers *A* and *B* in a transcript, we computed four counts for each of the 14 LIWC categories: The number messages from *A* to *B* containing the category (N^{align}), the number of messages from *A* to *B* not containing the category (N^{base}), the number of replies containing the category to messages containing the category (C^{align}), and the number of replies containing the category to messages not containing the category (C^{base}). To generate robust parameter estimates, we aggregated counts across all transcripts for the same parent and child into six-month age bins, yielding 8 bins (youngest:

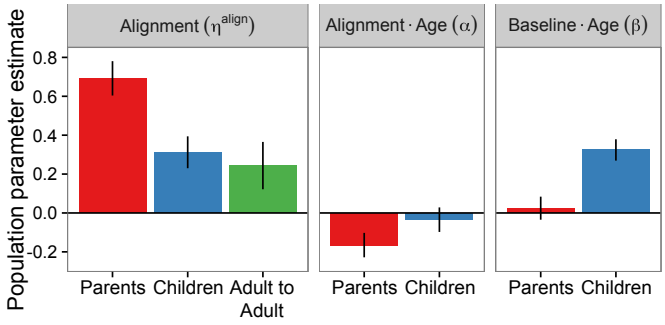


Figure 2: Posterior parameter estimates for alignment (η^{align}), developmental change in alignment (α), developmental change in baseline function word production (β) for both parents and children, as well estimated alignment between parents for a baseline. Bars indicate means, error-bars indicate 95% highest posterior density intervals.

12–18 mo., oldest 54–60 mo.). When estimating α and β—the age-related scalars—we numbered these bins from 1 to 8 and then subtracted the mean bin number from each (4.5). This centers the intercept (η^{align}) at the middle value, yielding the smallest average predictive error for other age bins.

We then fit the Hierarchical Alignment Model to the data separately for children and adults. Posterior distributions for all parameters were estimated using a Hamiltonian Monte Carlo sampler (Carpenter et al., in press) with three independent chains, and 500 samples in each chain. The first 100 samples of each chain were discarded to ensure sufficient burnin based on inspection of trace plots that typically showed convergence after 50–75 samples. In addition, to provide a baseline for comparison, we also estimated alignment for the parent-parent interactions in the corpus. All data and analysis code are available in a public github repostory at `http://github.com/langcog/alignment`.

Results and Discussion

Overall, alignment parameter estimates were above zero for both parents and children, suggesting that both groups aligned (Figure 2). Parents aligned reliably more to their children than children aligned to parents, and parents also aligned reliably more to their children than to each other. Thus, in the aggregate, we can conclude that parents coordinate to their children more than they coordinate to other adults.

The linguistic tuning hypothesis makes two predictions. First, parents might change the words they produce, using simpler words with younger children. If so, we would expect an increase in their likelihood of producing optional function words (positive β). In line with previous analyses, we find no evidence of this (Newport et al., 1977; Hayes & Ahrens, 1988). (We do, however, find a large and reliable increase in *children's* use of these words as they grow older, as would be expected based on their developing linguistic competence.)

Second, parents could use the same words, but be less contingent on children's production—repeating less, clarifying less, rephrasing less. This second prediction is precisely what we observe in both the aggregate parameter estimates and the developmental parameter estimates (Figure 4). Parental alignment decreases reliably over development, and children's alignment shows a similar trend. Thus, parent-child conversations become gradually less coordinated over development, settling to the adult-adult baseline as children need less scaffolding to be successful communicators.

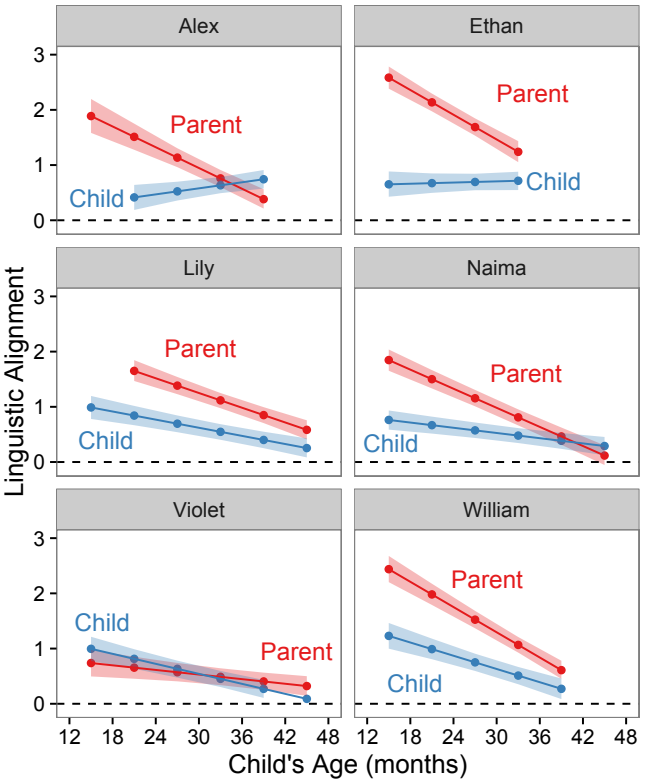


Figure 3: Changes in alignment for 6 children and their parents measured longitudinally in the Providence corpus (Demuth et al., 2006). Points indicate the mean of the posterior distribution, shaded regions indicate 68% highest probability density intervals.

In addition to estimating population-level parameters for alignment and its change over development, our hierarchical model also estimates parameters for each of the individual adults and children in CHILDES. Examining these parameters shows both consistency and variability across dyads (Figure 3). Across children, parental alignment is consistently highest early in development, but both children and parents vary in their level of alignment and in the rate at which it changes across development. These reliable individual differences in alignment present a possible vehicle for studying individual differences in children's language acquisition.

LIWC Category Validation Control. Our measure of alignment is sensitive to repetitions, reformulations, expansions, and other conversational turns that result in contingency of function word categories. But, the alignment we observed in parent-child interactions could have been driven

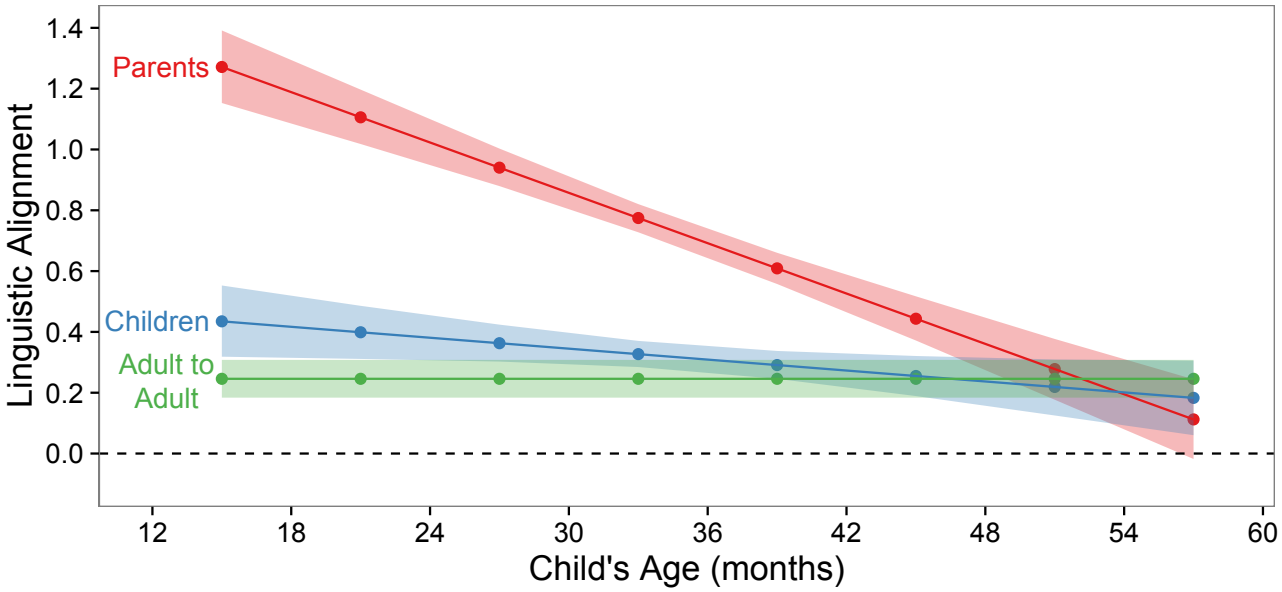


Figure 4: Model-estimated changes in linguistics alignment by 6-month-window. Over development, both parents and children decrease in their linguistic alignment to each other until they align at adult-adult levels. Points indicate the mean of the posterior distribution, shaded regions indicate 68% highest probability density intervals, equivalent to one standard deviation.

entirely by repetition. To control for this, we constructed a set of shuffled LIWC categories by randomly permuting the original word-category assignments. If these shuffled categories yield the same levels of alignment as the original categories, we would conclude that linguistic alignment between parents and children is happening at the word level rather than the category level. Compared to 40 different random shuffles, parents' alignment on the original categories (η^{align}) was greater than 37, and developmental change in alignment (α) was greater than all 40 ($p < .001$). Similarly, children's alignment on the LIWC categories was greater than all 40 shuffled categories ($p < .001$). Thus, the developmental changes in alignment we observed are not just due to more exact repetitions in early childhood, but also to higher-level linguistic coordination.

General Discussion

Although even the youngest infants are equipped with the capacity to learn from the language they hear, their learning is highly constrained by their developing attentional and memory systems (Vlach & Johnson, 2013). How do children acquire language so rapidly despite these cognitive constraints? One hypothesis is that the the language they hear is tuned to their learning capacities, providing the right kind of information at the right time (Snow, 1972; Vygotsky, 1978).

Previous attempts to test this hypothesis have produced a mixture of evidence, with some in strong support of the linguistic tuning hypothesis (e.g. Hirsh-Pasek et al., 1984; Chouinard & Clark, 2003), others in strong opposition (e.g. Brown & Hanlon, 1970; Newport et al., 1977), and yet others finding mixed support and interesting individual differences (e.g. Sokolov, 1993; Dale & Spivey, 2006). Our work leverages the power of hierarchical Bayesian models to integrate together data from two orders of magnitude more children than previous work, providing a more robust test of tuning.

We find, at the population level, that parents indeed provide linguistic input that is calibrated to children's development. The structural elements of their utterances are strongly contingent on children's previous productions early in development, but gradually decline to adult-adult levels of coordination. In line with previous work, we do not find evidence of change at a global level: Parents do not use different words when talking to children of different ages (Hayes & Ahrens, 1988). Instead, calibration is local—sensitive to the on-going conversation and attuned to children's processing in context.

Instead of across-the-board simplification, linguistic tuning may be particularly beneficial; language at just the right level of complexity may produce desirable difficulties for language learning. Critically, linguistic tuning does not require that parents consciously have as their goal the optimization of children's learning. They need not be teachers; they need only be communicators. If parents want to communicate with their children, and their children need significant linguistic support, they will have no choice but to align.

Acknowledgments

We are grateful to Aaron Chuey and Jake Prasad for help with an earlier version of the model. This work was supported by NIH NRSA F32HD075577 to DY and NSF BCS#1528526 and #1456077 to MCF.

References

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the developmenf of language* (Vol. 8, pp. 11–53). New York: Wiley.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (in press). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence . *Journal of Child Language*, *30*, 637–669.

Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*, 391–430.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on world wide web* (pp. 699–708).

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, *49*, 137–173.

Doyle, G., Yurovsky, D., & Frank, M. C. (2016). A robust framework for estimating linguistic alignment in Twitter conversations. In *Proceedings of the 25th international conference on world wide web* (pp. 637–648). International World Wide Web Conferences Steering Committee.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological Science*, *23*, 931–939.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*, 8–11.

Giles, H., Coupland, J., & Coupland, N. (1991). *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71*, 878–894.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of Child Language*, *15*, 395–410.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown & Hanlon revisited: mothers' sensitivity to ungrammatical forms. *Journal of Child Language*, *11*, 81–88.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*, *22*, 39–44.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*, 125–143.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*, 481–507.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. 3rd edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from cdi analysis. *Developmental Science*, *14*, 769–785.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear picture books and the statistics for language learning. *Psychological Science*, *26*, 1489–1496.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*, 447–468.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. A. Ferguson (Ed.), *Talking to children language input and interaction* (pp. 109–149). Cambridge, UK: Cambridge University Press.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC*. Austin, TX: LIWC.net.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*, 549–565.

Sokolov, J. L. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, *6*, 1008–1023.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*, 53–71.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*, 375–382.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process.* Cambridge, MA: Harvard University Press.

Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*, 619–626.

Yurovsky, D., Yu, C., & Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. *Frontiers in Psychology*, *3*.

The Plausible Impossible: Causal Constraints on Magical Reasoning

Andrew Shtulman (shtulman@oxy.edu)
Caitlin Morgan (cmorgan@oxy.edu)
Department of Psychology, Occidental College
1600 Campus Road, Los Angeles, CA 90041

Abstract

A common intuition, often captured in fiction, is that some impossible events (e.g., levitating a stone) are "more impossible" than others (e.g., levitating a feather). We investigated the source of this intuition, hypothesizing that graded notions of impossibility arise from explanatory considerations logically precluded by the violation at hand but still taken into account. Studies 1-2 involved college undergraduates (n = 192), and Study 3 involved preschool-aged children (n = 32). In Study 1, participants saw pairs of magical events (spells) that violated one of 18 causal principles—six physical, six biological, and six psychological—and were asked to indicate which spell would be more difficult to learn. Both spells violated the same causal principle but differed in their relation to a subsidiary principle. Participants' judgments of spell difficulty honored the subsidiary principle, even when participants were given the option of judging the two spells equally difficult. Study 2 replicated the effects of Study 1 with Likert-type ratings, and Study 3 replicated those effects in children. Taken together, these findings suggest that events that defy causal explanation are interpreted in terms of explanatory considerations that hold in the absence of such violations.

Keywords: causal inference, explanation, imagination

Introduction

The animator Walt Disney began his career making cartoons of a very different style than those he made at the end of his career. Disney's early cartoons were surreal: clothes jumped from clothes lines and ran around the yard; sausages jumped from grills and danced in a kick line; pianos turned insolent and bit their players. Before making feature-length films, Disney decided that his cartoons had to be "plausibly impossible": they could violate some of the audience's real-world expectations but not too many. In the movie Snow White, for instance, it was deemed plausibly impossible for forest animals to communicate with Snow White but implausibly impossible for them to double in size or to ooze through keyholes (Lane, 2006).

The idea that something could be plausibly impossible is paradoxical. An impossible event, from a psychological point of view, is one that violates an expectation about the world deemed true by necessity—an expectation that holds in all conceivable permutations of reality (Nichols, 2006; Shtulman & Carey, 2007; Shtulman, 2009). Violations of this nature are presumably black-and-white; an event either violates a "law of nature" and cannot occur in the real world or it violates no such law and can occur in the real world, whatever the likelihood. Considerations of plausibility—or subjective probability of occurrence—make sense when applied to events that can occur but makes much less sense when applied to events that cannot.

Still, fictional narratives frequently make distinctions between impossible events that are plausible and those that are not. In the fictional world of Star Wars, for instance, the Jedi master Yoda teaches the Jedi apprentice Luke Skywalker to levitate stones before he teaches him how to levitate an entire starship. In the fictional world of Harry Potter, the potions instructor Severus Snape teaches his students how to brew forgetfulness before he teaches them how to brew endurance. And in the fictional world of Cinderella, Cinderella's fairy godmother turns a pumpkin into a stagecoach and a horse into a coachman rather than turning a horse into a stagecoach and a pumpkin into a coachman.

Starships are heavier than rocks; endurance is less attainable than forgetfulness; and pumpkins resemble stagecoaches more than they resemble coachmen. But why should these considerations influence our perception of patently impossible events? The magic of levitation severs the connection between weight and liftability; the magic of potions severs the connection between hard work and attainability; and the magic of transfiguration severs the connection between physical resemblance and mutability. Nevertheless, the genre of magical realism is full of impossible events that honor severed connections like these. Honoring such connections is arguably what constitutes the "realism" in magical realism.

What psychological principles might account for graded notions of impossibility, as exemplified by the popular fictional narratives above? Psychologist interested in the mental representation of impossible events have typically analyzed such events in terms of ontological violations. An ontology is a basic category of existence (e.g., "object," "animal," "number"), and an ontological violation is a violation of one of those categories' core properties (e.g., that objects are solid or that animals can die). Ontological violations have been shown to affect an event's memorability (Boyer, 1999). Concepts that violate one or two ontological commitments (e.g., a talking tree) are more memorable than those that violate several (e.g., a talking tree that floats in the air and never dies) or those that violate none (e.g., a flowering tree), which presumably affects how often, and how reliably, these concepts are passed from one person to another and from one generation to another.

Such memory effects have been documented both for artificial materials, constructed in the lab (Barrett & Nyhof, 2001), and for more authentic materials, like Aesop's fables (Upal, 2011) and Grimm's fairytales (Norenzayan, Atran, Faulkner, Schaller, 2006). These effects tell us that an event's memorability is determined by its conformity to our ontological commitments, but they tell us less about an event's plausibility. Indeed, memorability and plausibility

Table 1: The six pairs of spells in each domain, grouped by the (irrelevant) causal constraint they exemplify.

Domain	Causal constraint	Spell
Physics	Object size	Making a (bush, tree) invisible
	Object weight	Making a (basketball, bowling ball) float in the air
	Object shape	Turning a broom into a (shovel, bucket)
	Object complexity	Shrinking a (chair, computer) to half its size
	Object density	Walking through a wall made of (wood, stone)
	Object value	Turning a lump of coal into a lump of (silver, gold)
Biology	Evolutionary similarity	Turning a person into a (monkey, pig)
	Developmental similarity	Turning an adult back into a (teenager, child)
	Ailment severity	Curing a person's (hiccups, arthritis)
	Organ size	Mending a broken (finger, arm)
	Organ complexity	Growing an extra (toe, eye)
	Organ plasticity	Making a person's (hair, teeth) grow longer
Psychology	Knowledge entrenchment	Making a person forget his own (phone number, name)
	Knowledge complexity	Teaching a monkey to do (arithmetic, calculus)
	Skill difficulty	Teaching a cow how to (skip, tap dance)
	Affect intensity	Making someone (smile, laugh)
	Trait stability	Increasing a person's (memory, intelligence)
	Language comprehension	Teaching a person to (read, speak) a foreign language

may be driven by independent factors. Impossible events that violate a minimal number of ontological commitments are not always plausible (in the sense of being believable), and impossible events that are plausible do not always violate a minimal number of ontological commitments (Gervais & Henrich, 2010). For example, fictional characters like vampires and zombies violate a "cognitively optimal" number of ontological commitments but are not commonly believed to exist, whereas religious entities like the Abrahamic God—an omnipotent, omniscient, omnipresent, omnibenevolent, invisible, and immortal being—violate several ontological commitments but are commonly believed to exist (Shtulman, 2008; Shtulman & Lindeman, 2016).

Here, we explore another property of conceptual representations that may account for graded notions of impossibility: the interconnectedness of causal knowledge. Recent research on the structure of causal knowledge suggests that much of this knowledge is organized in abstract, coherent networks of directed causal relations, both for children (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004) and for adults (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). These networks are derived from patterns of covariation between putative causes and putative effects and can be used to generate explanations, predictions, interventions, and counterfactuals.

Critically, with respect to the phenomena of interest (plausible impossibility), these networks involve multiple relations among multiple concepts. A network that represents our knowledge of physical objects, for instance, might have links between concepts as varied as mass, weight, volume, density, buoyancy, solidity, cohesion, contact, and support. Given a representation of this nature, we suspect that severing one connection in the network, as done when contemplating magic, would leave other connections intact. Severing the connection between contact and support, for instance, would leave the connection between weight and support intact, thus yielding the intuition that levitating a starship would be more difficult than levitating a stone.

In the studies that follow, we test this idea across multiple tasks and multiple domains, on the prediction that what makes a magical event plausibly impossible is its conformity to a larger network of causal expectations of which the violated expectation is just one constituent. The cover story for all studies was that participants were reviewing the curriculum for Hogwarts School of Witchcraft and Wizardry, the fictional school in J. K. Rowling's "Harry Potter" series. Participants were asked to evaluate the difficulty of various spells on the Hogwarts curriculum. In all studies, we found that participants relied on explanatory considerations relevant to the spells' outcomes but precluded by the spells' causal structure nonetheless.

Study 1

Method

The participants in Studies 1 through 3 were undergraduates at Occidental College. They were recruited from introductory psychology courses and compensated for their participation with extra credit in those courses. They completed the study in the form of an online questionnaire.

Sixty-four undergraduates participated in Study 1. They were shown 18 pairs of spells, ostensibly culled from the Hogwarts curriculum, and asked to determine which spell in each pair would be more difficult to learn. The pairs were constructed such that both spells violated the same deep-seated causal principle—and were thus impossible for the same reason—but differed from in surface-level properties relevant to a subsidiary principle. For instance, the pair "making a basketball float in the air" and "making a bowling ball float in the air" both violate the principle that

unsupported objects fall, but they differ in relation to the subsidiary principle that weight affects an object's liftability. Of interest was whether participants' responses would honor causal principles not explicitly suspended in the spell descriptions but dependent on the suspended principles nonetheless.

Six pairs of spells involved physical principles, six involved biological principles, and six involved psychological principles. All 18 pairs can be found in Table 1. Within each pair, the spells were designed to differ as minimally as possible. Basketballs and bowling balls, for instance, differ substantially in weight but do not differ substantially in size or shape (both of which might also influence an object's liftability). Also, it should be noted that the outcome of each spell was not always impossible, but it would be impossible to produce that outcome instantaneously or extemporaneously. For example, "making someone laugh" and "making someone smile" are not impossible in principle but are impossible to do instantaneously.

Participants received the 18 pairs of spells in one of two random orders. Half of the participants were asked to indicate which spell would be more difficult to learn (Study 1a), and half were asked to do the same but were given the option of indicating that both spells were "equally difficult" to learn (Study 1b).

Results

In Study 1A, participants' judgments of spell difficulty aligned with the spells' implicit causal ordering 85% of the time for physical spells, 86% of the time for biological spells, and 82% of the time for psychological spells. All three percentages were significantly greater than expected by chance (50%): physics: $t(31) = 10.88$, $p < .001$; biology: $t(31) = 14.04$, $p < .001$; psychology: $t(31) = 10.86$, $p < .001$. In addition, most participants (88%) demonstrated the anticipated effect for a significant number of spells (13 or more, binomial probability < .05), and all spells elicited the anticipated effect for a significant number of participants (21 or more, binomial probability < .05).

In Study 1B, participants' judgments of spell difficulty aligned with the spells' implicit causal ordering 57% of the time for physical spells, 72% of the time for biological spells, and 66% of the time for psychological spells. These percentages are lower than those in Study 1A, but participants were given three response options rather than two. Accordingly, all three percentages were significantly greater than expected by chance (33%): physics: $t(31) = 5.49$, $p < .001$; biology: $t(31) = 11.67$, $p < .001$; psychology: $t(31) = 9.38$, $p < .001$. Most participants (72%) demonstrated the anticipated effect for a significant number of spells (10 or more, binomial probability < .05), and most spells (78%) elicited the anticipated effect for a significant number of participants (15 or more, binomial probability < .05).

Discussion

Participants demonstrated a statistically reliable sensitivity to the spells' implicit causal ordering, judging spells that represented a greater departure from reality as "more difficult" in all three domains. Might these findings be an artifact of the task? The task was designed, after all, to highlight a single difference between each pair of spells, and participants may have heeded that difference only because it was highlighted for them.

One point against this interpretation is that participants continued to heed that difference even when given the option of indicating that both spells would be equally difficult to learn (Study 1B). Nevertheless, we attempted to provide stronger evidence of causality-based reasoning in Study 2, by asking participants to evaluate the 36 spells from Study 1 as individuals rather than as members of a pair.

Study 2

The goal of Study 2 was to replicate the findings of Study 1 using Likert-like ratings of spell difficulty rather than pairwise comparisons. We collected those ratings in one of three ways: by asking participants to rate the two versions of each spell back-to-back (Study 2A), by asking participants to rate the two versions of each spell intermixed with the other spells (Study 2B), and by asking participants to rate either the more-extreme version of each spell or the less-extreme version but not both (Study 2C).

Method

The participants in Study 2 were 128 undergraduates; 32 participated in Study 2A, 32 participated in Study 2B, and 64 participated in Study 2C. Each rated the difficulty of the spells presented in Table 1 on a seven-point scale, from "slightly difficult" to "extremely difficult." Studies 2A and 2B employed within-participants design, whereas Study 2C employed a between-participants design.

Results

In Study 2A, participants' difficulty ratings for the more-extreme spells were significantly higher than their difficulty ratings for the less-extreme spells in all three domains (physics: $M = 4.1$ vs. 3.6, $t(31) = 4.95$, $p < .001$; biology: $M = 4.7$ vs. 4.0, $t(31) = 7.56$, $p < .001$; psychology: $M = 4.2$ vs. 3.6, $t(31) = 7.21$, $p < .001$).

These differences remained significant even when the two versions of each spell were randomly intermixed with other spells (Study 2B), making their comparison less explicit (physics: $M = 4.3$ vs. 3.8, $t(31) = 5.75$, $p < .001$; biology: $M = 5.3$ vs. 4.4, $t(31) = 11.29$, $p < .001$; psychology: $M = 4.5$ vs. 3.8, $t(31) = 7.65$, $p < .001$). And they remained significant even when participants rated the less-extreme version of each spell or the more-extreme version but not both (Study 2C), as assessed with independent-samples comparisons (physics: $M = 4.1$ vs. 3.4, $t(62) = 3.02$, $p < .01$; biology: $M = 5.1$ vs. 3.8, $t(62) = 6.64$, $p < .001$; psychology: $M = 4.4$ vs. 3.2, $t(62) = 5.30$, $p < .001$).

These effects were observed at the item level as well. In Study 2A, participants rated the more-extreme spell in each pair as significantly more difficult than the less-extreme spell for 94% of pairs (17 of 18). This was true for 78% of pairs

(14 of 18) in Study 2B and for 83% of pairs (15 of 18) in Study 2C. All three percentages were significantly greater than expected by chance (binomial probability < .05).

Discussion

Participants' intuitions about spell difficulty honored causal principles precluded by the spell's causal structure even when those intuitions were assessed spell-by-spell, rather than in comparison to its causal match. These effects have proven robust across tasks and across domains, but they have only been documented, to this point, among adults. It is possible that adults have come to perceive magic as causally-constrained through exposure to magical narratives in books, shows, or films. Study 3 explores this possibility by testing participants who have acquired significantly less exposure to such narratives: preschool-aged children.

Study 3

The question of interest in Study 3 was whether four- and five-year-old children view magical events through a causal lens, similar to adults. Children of this age are adept at distinguishing possible events from impossible events, labeling only impossible events as "magic" (see Woolley, 1997, for a review), but it is unclear how they reason about the content of such events. Do they, like adults, believe that some impossible events are more impossible than others?

Method

Thirty-two children between the ages of four to five (*M* age = 4 years, 8 months) participated in Study 3. Approximately half were male, and half were female. They were recruited from a preschool in Southern California and tested on site. Participants were shown the same 18 pairs of spells used in Studies 1 and 2. Each spell was visually depicted on an index card. Participants were presented the cards in pairs and asked to sort them into each of two containers. One container was labeled with a picture of Harry Potter and the other was labeled with a picture of Professor Dumbledore. Participants were instructed on how to use the containers as follows:

"Have you ever heard of Harry Potter? Harry Potter is a boy in a story who has magical powers and goes to a school called Hogwarts School of Witchcraft and Wizardry, where he learns how to cast different spells. Here are two buckets with pictures of wizards on them. This bucket shows Harry who is a young wizard and not very good at magic yet, and this one shows Professor Dumbledore, one of Harry's teachers at the school. He is an older wizard who is much better at magic. Over here I have a pile of cards with different spells on them. Some of the spells are easy, which means that even young wizards can do them, but some of the spells are hard, which means that only older wizards can do them. Can you help me figure out which spells are easy and which spells are hard? You will put the easy spells in this bucket that shows Harry, and you will put the hard spells in this bucket that shows Dumbledore."

The experimenter verified that participants understood the instructions by asking them to point to the bucket for the easy spells and then point to the bucket for the hard spells. The experimenter then presented participants with each pair of spells in one of two random orders. Feedback was not provided during the sorting process, though participants who struggled with the task were assured that there are no correct answers and encouraged to try their best.

Results

Participants sorted the spells in accordance with their implicit causal ordering significantly more often than expected by chance ($M = 10.7$, $t(31) = 4.14$, $p < .001$). This effect was obtained for each domain (physics: $M = 3.7$, $t(31) = 2.96$, $p < .01$; biology, $M = 3.5$, $t(31) = 2.22$, $p < .05$; psychology, $M = 3.5$, $t(31) = 1.96$, $p < .05$) and for at least three of the six spells within those domains.

The effect was smaller for children than it was for adults (Cohen's $d = 0.74$ vs. 2.70), but it was present and increasing with the age. Indeed, the older a child was (in months), the more likely he/she sorted the less-extreme spell into the Harry-Potter container and the more-extreme spell into the Dumbledore container ($r(30) = 0.44$, $p < .05$).

Discussion

By age five, children appear to hold causality-based intuitions regarding the plausibility of impossible events, assessing how difficult it would be to bring about a magical event on the basis of causal principles that should not logically pertain to that event. These intuitions are in place not only before children have received formal instruction on causal principles but also before they have received much exposure to the genre of magical realism. Certainly, children of this age have had some exposure—they were, after all, familiar with the story of Harry Potter—but it is unlikely that such exposure could account for our findings, as our spells were novel and our task was novel as well.

That said, children's sensitivity to the causal structure of our stimuli increased over the ages sampled, from 3 years and 10 months to 5 years and 7 months. This is a period over which children are becoming increasingly sensitive to another aspect of magical phenomena: the difference between a magic trick and "real" magic (Chandler & Lalonde, 1994; Rosengren & Hickling, 1994). That is, five-year-olds tend to recognize that visual illusions, like the illusion of one object passing through another or the illusion of one object turning into another, are brought about by sleight of hand or trick apparatuses, whereas four-year-olds tend to claim that such illusions are genuine instances of magic. This developmental change, from identifying illusions with magic to identifying illusions with trickery, is presumably driven by an increased awareness of the causal constraints on real-world events, and our task may have tapped into the same change, albeit less directly.

General Discussion

In three studies, we found that participants used real-world causal expectations to interpret events that presumably fall outside those expectations: magical events. We obtained

these findings regardless of whether our task involved forced-choice comparisons, opt-out comparisons (in which ties were allowed), or Likert-type ratings; regardless of whether our stimuli involved physical principles, biological principles, or psychological principles; and regardless of whether our participants were preschoolers or college-educated adults.

As a whole, these finding imply that causal knowledge is not easily suspended. Events that explicitly violate everyday causation still elicit causal inferences, possibly because severing one link in a causal network still leaves the rest of the network intact. Events like "walking through a wall made of wood" and "walking through a wall made of stone" have no explanation—they are deemed impossible by the laws of nature—but we still rely on explanatory considerations like hardness and density to interpret those events, perceiving the latter as more difficult to achieve than the former.

Our findings accord with a growing body of literature demonstrating that products of the imagination—fiction, fantasy, pretense, superstition, ritual—are structured by causal constraints on reality. For instance, it's been shown that we prefer stories in which humans transform into animals (a close ontological match) to stories in which humans transform into plants (Kelly & Keil, 1985), and we prefer stories in which animals transform into humans (an ontological promotion) to stories in which humans transform into animals (Griffiths, 2015). We prefer stories that violate contingent truths (e.g., that Washington DC is the capital of the US) to those that violate mathematical truths (e.g., that two plus two equals four; Weisberg & Goodstein, 2009). We prefer extraterrestrial creatures that honor the biological properties of terrestrial creatures—e.g., bilateral symmetry, cephalization—to those that do not (Ward, 1994). And we prefer divine agents (gods) that honor the properties of human psychology—e.g., that knowledge depends on perception, that perception depends on attention, that attention depends on interest—to those that violate such expectations (Lane, Evans, Brink, & Wellman, 2016; Purzycki, 2013).

Our findings also accord with the emerging consensus that causal constraints on imagination operate even in young children (Buchsbaum, Bridgers, Weisberg, & Gopnik, 2012). For instance, two-year-old children recognize that, when pretending to pour tea from an empty teapot into an empty cup, tea has been transferred from the teapot to the cup and tea will spill out of the cup if the cup is overturned (Harris, Kavanaugh, & Meredith, 1994). Four-year-old children can distinguish the properties of one pretend world (the world of Batman) from another (the world of SpongeBob) and keep those properties separate when drawing inferences about what is likely to be true in those worlds (Skolnick & Bloom, 2006). And four-year-old children distinguish fictional stories that resemble reality (stories about finding ladybugs and climbing trees) to those that do not (stories about finding fairies and talking to trees), privileging the former as a more secure source of information than the latter (Richert & Smith, 2011; Walker, Gopnik, & Ganea, 2015).

Our findings extend this literature by highlighting a property of causal knowledge whose effects on imagination have yet to be explored: its interconnectedness. Causal knowledge is organized in abstract, coherent networks, and our findings suggest that this form of organization constrains our interpretation of events that violate such knowledge. Of course, our findings provide only indirect evidence of the influence of causal structure on the interpretation of causal violations. Future research could explore that influence more directly by measuring or manipulating the causal knowledge relevant to a particular causal violation. For instance, a child who has yet to discern the relation between density and buoyancy should have no expectations regarding the role of density in magical events pertaining to buoyancy (e.g., a spell for making aluminum float in water vs. a spell for making lead float in water), whereas a child who has discerned the relation between density and buoyancy should hold such expectations.

Our findings also highlight a quirk in how we reason about physical possibility. Reasoning about physical possibility is a form of modal cognition. Reasoning about moral permissibility is also a form of modal cognition. The two forms of reasoning share a number of similarities (Perkins, 1983; Sinnott-Armstrong, Raffman, & Asher, 1995). Both are concerned with what is normatively true about the world rather than what is descriptively true; both entail the application of preexisting commitments (moral rules, physical laws) to hypothetical situations; and both can be expressed with the same modal verbs ("can," "could," "might," "must," "should"). Indeed, studies that have explored the two forms of reasoning in conjunction have found parallels between them—developmental parallels in the types of events judged impossible or impermissible (Browne & Woolley, 2004; Chernyak, Kushnir, Sullivan, & Wang, 2013) and cognitive parallels in the considerations underlying those judgments (Shtulman & Tong, 2013).

That said, there is a key difference between the two forms of reasoning: impermissibility comes in degrees but impossibility does not—at least not on the surface. We regularly rank some impermissible actions (e.g., murder) as "more wrong" than others (e.g., stealing), and when describing impermissible events, we use hedges like "a little wrong," "sorta wrong," or "mostly wrong." We tend not to use hedges when describing impossible events, though; the phrases "a little impossible," "sorta impossible," or "mostly impossible" generate a tenth as many Google hits as generated by their moral counterparts. Thus, what is odd about the phenomenon at hand is not just that we apply causal expectations to a-causal events but also that we treat the distinction between possibility and impossibility as binary rather than graded.

In conclusion, children and adults alike have consistent expectations about the plausibility of magical events. Suspending disbelief in one causal violation does not lead to widespread suspension of disbelief; other causal expectations are maintained, even if those expectations are no longer appropriate. When Walt Disney specified that his feature-length films had to be plausibly impossible, he pinpointed an intuition that is consistently honored in fiction but not well

understood from a psychological point of view, even today. Future research on graded notions of impossibility promises to shed light both on the structure of causal cognition and the structure of imagination.

References

Barrett, J. L., & Nyhof, M. A. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of Cognition and Culture, 1*, 69-100

Boyer, P. (1999). Cognitive tracks of cultural inheritance: How evolved intuitive ontology governs cultural transmission. *American Anthropologist, 100*, 876-889.

Browne, C. A., & Woolley, J. D. (2004). Preschoolers' magical explanations for violations of physical, social, and mental laws. *Journal of Cognition and Development, 5*, 239-260.

Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B*, 367, 2202-2212.

Chandler, M. J. & Lalonde, C. E. (1994). Surprising, magical, and miraculous turns of events: Children's reactions to violations of early theories of mind and matter. *British Journal of Developmental Psychology, 12*, 83-95.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A Comparison of American and Nepalese Children's Concepts of Freedom of Choice and Social Constraint. *Cognitive Science, 37*, 1343-1355.

Gervais, W. M., & Henrich, J. (2010). The Zeus problem: Why representational content biases cannot explain faith in gods. *Journal of Cognition and Culture, 10*, 383-389.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-32.

Griffiths, T. L. (2015). Revealing ontological commitments by magic. *Cognition, 136*, 43-48.

Harris, P. L., Kavanaugh, R., & Meredith, M. (1994). Young children's comprehension of pretend episodes: Integration of successive actions. *Child Development, 65*, 16-30.

Kelly, M. H., & Keil, F. C. (1985). The more things change...: Metamorphoses and conceptual structure. *Cognitive Science, 9*, 403-416.

Lane, A. (2006, Dec. 11). Wonderful world: What Walt Disney made. *The New Yorker*.

Lane, J. D., Evans, E. M., Brink, K. A., & Wellman, H. M. (2016). Developing concepts of ordinary and extraordinary communication. *Developmental Psychology, 52*, 19-30.

Nichols, S. (2006). Imaginative blocks and impossibility: An essay in modal psychology. In S. Nichols (Ed.), *The architecture of the imagination* (237-255). Oxford, UK: Oxford University Press.

Norenzayan, A., Atran, S., Faulkner, J., & Schaller, M. (2006). Memory and mystery: The cultural selection of minimally counterintuitive narratives. *Cognitive Science, 30*, 531-553.

Perkins, M. (1983). *Modal expressions in English*. London, UK: Pinter.

Pieraut-LeBonniec, G. (1980). *The development of modal reasoning: Genesis of necessity and possibility notions*. London, UK: Academic Press.

Purzycki, B. G. (2013). The minds of gods: A comparative study of supernatural agency. *Cognition, 129*, 163-179.

Richert, R. A., & Smith, E. I. (2011). Preschoolers' quarantining of fantasy stories. *Child Development, 82*, 1106-1119.

Rosengren, K. S. & Hickling, A. K. (1994). Seeing is believing: Children's explanations of commonplace, magical, and extraordinary transformations. *Child Development, 65*, 1605-1626.

Shtulman, A. (2008). Variation in the anthropomorphization of supernatural beings and its implications for cognitive theories of religion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1123-1138.

Shtulman, A. (2009). The development of possibility judgment within and across domains. *Cognitive Development, 24*, 293-309.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development, 78*, 1015-1032.

Shtulman, A., & Lindeman, M. (2016, in press). Attributes of God: Conceptual foundations of a foundational belief. *Cognitive Science*.

Shtulman, A., & Tong, L. (2013). Cognitive parallels between moral judgment and modal judgment. *Psychonomic Bulletin & Review, 20*, 1327-1335.

Sinnott-Armstrong, W., Raffman, D., & Asher, N. (1995). *Modality, morality, and belief*. Cambridge, UK: Cambridge University Press.

Skolnick, D., & Bloom, P. (2006). What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction. *Cognition, 101*, B9-B18.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27*, 453-489.

Upal, M. A. (2011). Memory, mystery and coherence: Does the presence of 2-3 counterintuitive concepts predict cultural success of a narrative? *Journal of Cognition and Culture, 11*, 23-48.

Walker, C. M., Gopnik, A., & Ganea, P. A. (2015). Learning to learn from stories: Children's developing sensitivity to the causal structure of fictional worlds. *Child Development, 86*, 310-318.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology, 27*, 1-40.

Weisberg, D. S., & Goodstein, J. (2009). What belongs in a fictional world? *Journal of Cognition and Culture, 9*, 69-78.

Woolley, J. D. (1997). Thinking about fantasy: Are children fundamentally different thinkers and believers from adults? *Child Development, 68*, 991-1011.

Statistical learning bias predicts second-language reading efficiency

Luca Onnis (lucao@ntu.edu.sg)
Division of Linguistics and Multilingual Studies, Nanyang Technological University
Singapore, 637332

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University
6525 HT Nijmegen, The Netherlands

Hongoak Yun (oaktreepark@gmail.com)
Department of Media and Communication, Konkuk University
Seoul, Korea, 143-701

Matthew Lou-Magnuson (matt0020@ntu.edu.sg)
Division of Linguistics and Multilingual Studies, Nanyang Technological University
Singapore, 637332

Abstract

Statistical learning (SL) is increasingly invoked as a set of general-purpose mechanisms upon which language learning is built during infancy and childhood. Here we investigated the extent to which SL is related to adult language processing. In particular, we asked whether SL proclivities towards relations that are more informative of English are related to efficiency in reading English sentences by native speakers of Korean. We found that individuals with a stronger statistical learning sensitivity showed a larger effect of conditional word probability on word reading times, indicating that they more efficiently incorporated statistical regularities of the language during reading. In contrast, L2 English proficiency was related to overall reading speed but not to the use of statistical regularities.

Keywords: statistical learning; sequential learning; reading; sentence processing; bilingualism.

Introduction

Human languages are learnt and processed in real time. Speech is the ultimate fleeting experience, as it dissipates as soon as it is produced. And while printed text is more stationary, proficient readers process words sequentially at a very fast pace, with relatively few gazes spent looking back to reread previous words. The inherent fleeting nature of language and the great efficiency that humans exhibit in learning and using languages suggest that the brain recruits mechanisms employed for processing sequential information. These mechanisms may involve the ability to unconsciously track and extract patterns of regularities across sensory modalities, and to abstract over these patterns (for reviews see Gomez & Gerken, 2000; Perruchet & Pacton, 2006).

Because of the probabilistic sequential nature of language processing, recent research has attempted to establish links between mechanisms for language learning and processing and of so-called statistical learning (SL), by relating individual variance in SL tasks with individual variance in tasks of language learning. The rationale of such approaches is to show that some measure of statistical learning ability, as assessed in tasks requiring implicitly learning relations among probabilistic sequences, is correlated with performance on one or more tasks involving language. Languages exhibit statistical properties at different levels of analysis, which make them potentially learnable from experience. In the early stages of language development infants and toddlers take in a considerable amount of this statistical structure. Infants exhibit individual differences in statistical learning skills that may modulate language development trajectories (e.g., Arciuli, & von Koss Torkildsen, 2012; Benasich et al., 2006; Kidd, 2012).

Here we take a similar approach in seeking evidence for a relation between statistical learning and second-language reading in adults. Arguably statistical language learning does not stop in the early years of childhood. Studies with older children have also linked poor implicit statistical skills with language and/or reading difficulties (Evans, Saffran, & Robe-Torres, 2009; Yim & Windsor, 2010) and adult native speakers are even sensitive to the particular statistical distribution of sentence structures within an experimental session, and adapt their processing preference accordingly (Fine, Jaeger, Farmer, & Quin, 2013). Thus, sensitivity to the statistical structure of language is likely to support not only children learning a language, but also adults using it daily. Indeed, direct predictive relations between statistical learning scores and online sentence processing and other linguistic tasks exist now both for children and adults (Yim & Windsor, 2010). In addition, neurophysiological data suggest that similar neural mechanisms appear to serve both syntactic processing of language and statistical learning of sequential patterns (Abla, Katahira, & Okanoya, 2008; Christiansen, Conway, & Onnis, 2012). Tracking implicit sequential regularities in linguistic and nonlinguistic stimuli seems to be independent of factors other than language performance, such as age, nonverbal IQ, and memory (Yim & Rudoy, 2010; Kaufman et al., 2010).

Here we are interested in capturing individual differences in statistical learning, language proficiency, and language comprehension, and we aim to correlate the three. Adult non-native speakers of English who show a stronger "English-

like" statistical learning bias in an artificial grammar task which is specifically designed as a litmus test to gauge the strength of preexisting experience with statistical regularities of English, are expected to be more sensitive to the statistics of English in an online reading task. Our study aims to achieve the following goals: 1) further support the view that statistical learning skills underlie not only language learning in childhood, but also language processing in adults; and 2) contrast the processing effect of individual differences in statistical learning vs. second-language (L2) proficiency on language reading.

Language-specific statistical learning

To measure individual statistical learning in second language speakers we used a task devised by Onnis & Thiessen (2013). The rationale for the task is to capture potential proclivities towards language-specific statistical relations. Natural languages differ in the statistical regularities available to learners. One such difference relates to the predominant directionality of conditional relations among words. For example, while "the" does not strongly predict "dog" (because many words can follow "the"), "dog" more strongly retrodicts "the." Learners are sensitive to informative relations in both directions (Jones & Pashler, 2007): both infants and adults are able to segment fluent speech into words on the basis of either forward-going relations among syllables, or backward-going relations (Pelucchi, Hay, & Saffran, 2009; Perruchet & Desaulty, 2008). The predominant directionality of relations among elements of the input differs between natural languages. One example of this is described in linguistic terms as the "headedness" of a language. The head of a phrase is the word that defines the syntactic function of the phrase (e.g., the verb in a verb phrase). Some languages (such as English) are classified linguistically as head-initial, meaning that the head of the phrase tends to occur before complement items (e.g., "going" in "going home"), while other languages are head-final and show the opposite word-order tendency. English for example arranges prepositional phrases like "to school" such that the head "to" precedes the noun "school", while other languages favor postpositional organization (as in Korean 학교-에 *hakkjo-e* 'school to'). An intuitive prediction derived from the linear organization of the input is that English word clusters are more syntactically cohesive in a backward-going direction. For example, in a phrase like "to school," "to" does not strongly predict any word – because many nouns can follow "to" – but "school" more strongly retrodicts "to" because there is a relatively smaller set of words that can precede "school." As these examples demonstrate, learners of different languages may experience different degrees of forward-going and backward-going cohesiveness. To assess this possibility, Onnis and Thiessen (2013) performed a corpus analysis of English (a predominantly head-initial and prepositional language) and Korean (a predominantly head-final and postpositional language). The results indicated that in English, high backward transitional probabilities and low forward transitional probabilities were a better indicator of phrase cohesiveness than high forward transitional and low backward probabilities; in Korean, the opposite pattern held true. Thus, language-specific information latent in the linear order of words partially predicts phrase structure in language.

Differences in statistical patterns between languages may, in turn, alter statistical learning itself. Sensitivity to backward-going regularities may be more adaptive for learners in an English environment than for learners in a Korean environment. Consistent with this hypothesis, Onnis and Thiessen (2013) found differences between English and Korean speakers when they were exposed to an auditory artificial grammar with conflicting forward and backward transitional probabilities, as in the training sample in a):

a) Training sample: .. *fushezirafunizitifugezibu* ..

Crucially, the artificial grammar was such that whenever forward transitional probability (fwd-TP) was low between any two adjacent syllables, backward transitional probability (back-TP) was high (e.g., fwd-TP(*zi*|*she*) = .33) and back-TP(*she*|*zi*) = 1), and vice versa (e.g., fwd-TP(*ra*|*zi*) = 1 and back-TP(*zi*|*ra*) = .33).

Two parses of sample a) into bisyllabic units are equally possible. One parse segments the signal such that the two syllables of a word have a high forward probability and a low backward probability (the HiLo pattern), while in the other parse the word-internal forward probabilities are low and the backward probabilities are high (the LoHi pattern).

b) Possible Parse I (HiLo): ..*fushe zira funi ziti fuge zibu* ..
c) Possible Parse II (LoHi): .. *shezi rafu nizi tifu gezi* ..

At test, two word groupings corresponding to the HiLo and LoHi patterns were pitted against each other in a two-alternative forced-choice task. A participant's *statistical learning bias* was defined as the proportion of LoHi choices over the sum of test trials presented to them. While both language groups had experienced the same grammar, native English speakers predominantly grouped syllables on the basis of high backward probabilities (as in example c above), while native Korean speakers preferred a grouping on the basis of high forward probabilities (as in example b). By contrast, with either visual or tonal non-linguistic stimuli, English and Korean speakers performed equivalently. The fact that the difference in performance between English and Korean speakers is limited to linguistic input, and consistent with the predominant directionality of their native language, suggests that the difference is due to linguistic experience.

Thus, the findings of Onnis & Thiessen (2013) suggest that SL can adapt to the statistical structure of linguistic input in ways that lead learners to have different expectations about novel subsequent input. To further support this claim, we wanted to find out whether individuals' degree of SL bias is correlated with statistical sensitivity to natural language in an online language comprehension task.

Statistical learning and language processing

Statistical patterns play an important role in language learning. Language *processing*, too, has been shown to depend on statistical information. Because of the time-dependent and sequential nature of both speech and reading, language comprehension is an inherently sequential process that probabilistically anticipates upcoming material. Hence, the cognitive processing effort for a piece of language (e.g., a word) should depend on its occurrence probability. At its most fundamental, a word's occurrence probability is simply its frequency in the language. Indeed, this frequency predicts the time required to recognize the word. When the word forms part of a sentence or text, reading time on the word is logarithmically related to its occurrence probability given the preceding context (Smith & Levy, 2013).

Estimating occurrence probabilities of words in context requires a probabilistic language model that implements knowledge about the language's statistics. Possibly the simplest language model is the bigram model, which assumes that a word's probability depends only on its overall frequency and on the immediately preceding word. Hence, word probabilities under such a model equal forward transitional probabilities, which have been found to predict reading times (McDonald & Shillcock, 2003). More sophisticated models can capture language statistics more accurately, resulting in more accurate reading time predictions (e.g., Frank & Bod, 2011). For the current study, however, we limit ourselves to a bigram model because the SL bias in our artificial grammar learning study is defined in terms of transitional probabilities.

Method

Participants. Fifty-seven adult native speakers of Korean (45 women; age M = 22.6, SD = 2.7) were recruited at four universities in Seoul (Konkuk University, Ewha Womans University, Sogang University, and Seoul National University). To qualify for the study their TOEFL score of English as a Second Language score should be over 600 (old version), and they should have spent at least three years in English-speaking country or environment. They participated in a statistical learning task, a sentence reading task, and an English proficiency self-assessment task (in this order). They were tested individually in a quiet room at their own university and were paid 10,000 Korean Won for their participation.

Statistical Learning Task

Materials. For the artificial grammar, the same materials as Onnis & Thiessen's (2013) Experiment 1 were used. The grammar lexicon was composed of eight monosyllabic words (fu, zi, shae, ni, ge, ra, ti, bu). To generate the training materials these words were arranged in a seamless sequence according to the rules of a stochastic Markovian grammar chain. The process started by choosing one of the eight possible words at random, and then generating the next word according to two probabilistic sequence constraints: whenever the forward probability between any two adjacent words was low (fwd-TP = .33), the backward probability was high (back-TP = 1), and vice versa. A sample of this template sequence is " …fu shae zi ra fu ni zi bu fu ge zi ra fu ni zi bu fu ge zi ti fu shae zi …". Frequencies of individual words, bigrams (two-word sequences), and their associated transition probabilities are summarized in Table 1.

Table 1: Summary of statistical relations among adjacent words in the grammar used in the the statistical learning task. Fwd-TP and Back-TP indicate forward and backward transitional probabilities among words in each bigram. Freq1 and Freq2 indicate frequency of occurrence of first and second word in each bigram, while Freq Bigram indicates how often each bigram occurred during training.

Bigram	Type	Fwd-TP	Back-TP	Freq 1	Freq 2	Freq Bigram
fu shae	LoHi	0.36	1.00	133	48	48
fu ni	LoHi	0.32	1.00	133	42	42
fu ge	LoHi	0.32	1.02	133	42	43
zi ra	LoHi	0.35	1.00	132	46	46
zi ti	LoHi	0.31	1.00	132	41	41
zi bu	LoHi	0.34	1.00	132	45	45
shae zi	HiLo	1.00	0.36	48	132	48
ni zi	HiLo	1.00	0.32	42	132	42
ge zi	HiLo	1.00	0.32	42	132	42
ra fu	HiLo	1.00	0.35	46	133	46
ti fu	HiLo	1.00	0.31	41	133	41
bu fu	HiLo	1.00	0.34	45	133	45

The actual sequence was realized using the speech synthesizer MBROLA, and concatenating the eight words to form a pauseless 3.5 minute speech stream of 711 words, with 80 ms for consonants and 260 ms for vowels. Because we were interested in the perception of grouping boundaries as driven by statistical biases alone, MBROLA did not use any prosodic or temporal cues to grouping boundaries. In addition, the sequence faded in and out for 5 s, giving the impression of an unbounded stream. The Italian diphone set in MBROLA was chosen to make the words dissimilar enough to Korean, but still clearly perceivable, and to engage participants in an "alien language" learning task. Finally, we ensured that all syllable sequences were phonotactically legal in Korean.

To verify whether participants preferred a specific pattern of transitional probabilities after exposure to the training phase, at test two types of bigrams were pitted one against the other in a forced-choice task, corresponding to a pattern of high fwd-TP and low back-TP (dubbed "HiLo" bigrams) versus the opposite "LoHi" bigrams. For example, the LoHi bigram 'fu shae' was presented against the HiLo bigram 'shae zi'. Six test pair trials were presented in random order, while the order within a pair was counterbalanced by repeating each test pair twice, for a total of 12 test trials. Note that HiLo and LoHi bigrams were composed of the same pseudowords and had been presented with an equal frequency at training. Hence, the only statistics useful to systematically choose one

type over the other would have to be a preference for the patterns of transition probabilities giving rise to the bigrams.

Procedure. Participants first listened to the training stream for 3.5 min, after which they were presented with the forced-choice task between pairs of LoHi and HiLo bigrams. For each pair they were asked to choose which sequence formed a grouping in the novel language they had just heard. We coded 1 for responses consistent with the English bias (preference for LoHi bigrams), and 0 for responses consistent with head-final languages such as Korean (namely, HiLo bigrams). We then defined a participant's *statistical learning bias* as the proportion of LoHi choices over the 12 test trials presented. The *strength* of the learning bias, computed as $(\text{learning bias} - 0.5)^2$, quantifies bias extremeness towards either LoHi or HiLo preference.

Sentence Reading Task

Materials. Sentences came from the 361-sentence UCL corpus (Frank, Monsalve, Thompson, & Vigliocco, 2013) explicitly created to evaluate language models on word-reading times. These sentences were drawn from original English narratives. Each participant was randomly assigned to one of ten groups, each containing 36 unique test sentences in English from the University College London UCL corpus, and five practice sentences. Test sentences were presented in random order. The words were displayed one at a time, progressing across the screen in their natural position with successive presses of the spacebar. Approximately half of the sentences were followed by a yes-no question regarding the content of what was just read in order to maintain the attention of the participants.

Proficiency assessment task

Participants self-assessed their proficiency in English listening, speaking, reading, and writing, as well as their accent, on a 7-point scale. All Pearson correlation coefficients between each pair of ratings was significantly positive (all $p < .0005$), and thus we took the average rating for each participant as a single measure of second language proficiency.

Results

A more efficient reader should adapt her reading times to the words' log-transformed occurrence probability, such that more probable words are read more quickly (Levy, 2008; Smith & Levy, 2008). Hence, we take the extent to which higher log-transformed forward transitional probability (as opposed to base word frequency) predicts shorter reading time as indicative of a reader's efficiency.

Naturally, we expect participants with higher English proficiency to read faster. In addition, they may also read more efficiently, in the sense that they display a more negative effect of forward probability on reading time. Alternative, non-exclusive possibilities are that participants read English faster and/or more efficiently if they have a more English-like (i.e., larger) learning bias or a stronger (i.e., more extreme) learning bias. We further expect any effects of English proficiency and learning bias to be independent from each other, although proficiency and learning bias may themselves be correlated.

Learning bias and proficiency. The mean statistical learning bias (towards the English-like LoHi patterns), learning bias strength (towards either pattern), and L2 proficiency score were, respectively, 0.575 (SD = 0.197), 0.044 (SD = 0.051), and 5.144 (SD = 0.721). Proficiency was not significantly correlated with learning bias ($r = .08$; $p > .5$) nor with strength of learning bias ($r = -.10$; $p > .4$). We note that the mean SL bias preference is closer to the English-expected pattern than what Onnis & Thiessen (2012; OT dataset) found. A comparison of the data distributions in the two dataset indeed suggests a major difference, notably a bimodal distribution, with the two modes located at 0.4 and 0.6, i.e. on each side of the chance level value of 0.5 in the current data. Conversely, the OT dataset had a single mode at 0.3. The bimodal distribution in our data suggests that the absolute strength of learning bias may be a better measure reflecting sensitivity to SL.

Data preprocessing. Data on a complete sentence was excluded if any RT on a word was extreme (below 80 ms or above 3000 ms). Furthermore, we did not include data on sentence-initial and sentence-final words, words followed by a comma, and clitics. This left a total number of 23,640 data points for analysis.

Reading time analysis. To investigate how readers' sensitivity to language statistics is related to their learning bias and English proficiency, the collected data were analyzed by linear mixed-effects regression. Subject-specific predictors (fixed effects) were: statistical learning bias (SLBIAS), strength of the bias (SLBIAS^2), and English proficiency (PROFICIENCY). Item-specific predictors were: word position in the sentence (WORDPOS), number of letters of word (LENGTH), log-transformed word frequency (WORDFREQ), and the log-transformed forward transitional word probability (FORWPROB). Word frequency and forward transitional probability were computed from word and bigram counts in the written-text part of British National Corpus. Properties of the previous word (PREVLENGTH, PREVWORDFREQ, PREVFORWPROB) were also included to take into account potential spillover in the reading times. As trial-specific predictor, RT on the previous word (PREVRT) was included to factor out the auto-correlation between consecutive key presses (Baayen & Milin, 2010). In addition, the model included by-subject and by-item (i.e., word token) random intercepts and by-subject random slopes of all predictor variables except for the subject-specific ones. RTs were log-transformed and all independent variables were standardized.

The first model that was fitted included the two-way interactions between each of the three subject-specific predictors and each of the six item-specific predictors (PREV)LENGTH, (PREV)WORDFREQ, and (PREV)FORWPROB. Next, non-significant ($|t| < 2$) interactions were removed one at a time, starting with the least significant interaction. Table 3 shows the resulting model's fixed-effects coefficients with

corresponding *t*- and *p*-values (*p*-values are obtained by treating *t*-values as *z*-scores, which is justified by the very large amount of data).

Table 3: Regression model fitted to log-transformed RTs

Variable	Coeff.	*t*-value	*p*-value
(intercept)	6.192	339.9	< .0001
PREVRT	0.091	14.7	< .0001
WORDPOS	−0.015	−3.5	< .0005
LENGTH	0.041	9.8	< .0001
PREVLENGTH	−0.006	−2.2	< .03
WORDFREQ	−0.006	−1.3	.2
PREVWORDFREQ	−0.013	−2.5	< .02
FORWPROB	−0.025	−6.7	< .0001
PREVFORWPROB	−0.013	−3.1	< .002
PROFICIENCY	−0.087	−4.0	< .0001
SLBIAS	0.009	0.4	.7
$SLBIAS^2$	0.001	0.0	.9
PROFICIENCY × LENGTH	−0.021	−5.0	< .0001
$SLBIAS^2$ × FORWPROB	−0.007	−2.8	< .005

As expected, LENGTH is positively related to RT: Longer words take longer to read. In addition, there are reliable negative effects of PREVWORDFREQ and (PREV)FORWPROB on RT: words are read faster if they are more frequent or their occurrence is more likely given the previous word.

The strong negative effect of PROFICIENCY means that participants who self-assessed their level of English as higher read more quickly, which validates the proficiency measure. L2 proficiency also modulates the effect of LENGTH in that less proficient participants (relative to more proficient ones) display increased difficulty with longer words.

Crucially, statistical learning bias is not significantly related to RT nor does it play a role in any interaction effect. In contrast, the strength of learning bias modulates the effect of FORWPROB: The stronger the learning bias (i.e., higher $SLBIAS^2$), the more sensitive reading becomes to forward probability (the effect of FORWPROB is more negative)

Discussion

In this study we investigated the association of individual differences in second-language processing with individual differences in a probabilistic sequence learning task and in second-language proficiency. The rationale is that if SL subserves language, and learning languages implies the discovery of language-specific distributional relations, then SL biases that match the statistical structure of a specific language increase efficiency when processing – here: reading – that language.

Our study extends on recent literature relating SL and literacy development. Arciuli and Simpson (2012) found a correlation between a non-linguistic SL task and measures of reading abilities derived from standardized reading tests in both elementary school children and adult native speakers. In addition, Spencer, Kashak, Jones and Lonigan (2014) established correlations between SL measures and early skills related to literacy development, notably oral language abilities, vocabulary knowledge, and phonological processing. With respect to second language learning, Frost, Siegelman, Narkiss, and Afek (2013) provided evidence that SL predicts word decoding abilities in a second language.

Our results further suggest that the ability to track statistical relations in sequenced patterns may not only be useful in *learning* a language early in life – the focus of previous research – but is also significantly correlated with the ability to *process* natural language as adults. In addition, the above studies established relations between SL and either broad measures of literacy outcomes, such as scores of standardized reading tests, or measures of single-word orthographic knowledge. The present study allowed a finer-grained examination of the role of statistical learning in more naturalistic reading conditions. We examined how biases in statistical learning reflecting the optimization of language-specific knowledge are related to second language proficiency and efficiency in real-time reading. We found that participants whose learning bias more closely matched the English-like head-first pattern were better able to use word predictability (operationalized as forward transitional probability) in real-time sentence processing. Concurrently, those participants showed a weaker effect of the words' base frequency on RTs. This is consistent with our interpretation that having a more English-like SL bias makes one closer to an "ideal" expectation-based English-language processor, who is sensitive to the words' conditional probabilities rather than base frequencies (which are already incorporated in the conditional probability measure).

Although more proficient participants generally read faster, and particularly so on longer words, there was no interaction between L2 proficiency and word frequency or forward transitional probability, suggesting that –perhaps surprisingly– increased proficiency is not reflected in more accurate knowledge or use of English language statistics. Conversely, participants with more positive learning bias did not read more quickly. The absence of this main effect is hard to explain considering our claim that more positive learning bias correlates with more efficient reading, as one would expect more efficient readers (i.e., those who make more optimal use of language statistics) to be faster readers, too. Finally, our results show that L2 proficiency and statistical learning bias are independent factors: They did not correlate across participants nor did they significantly interact.

Conclusion

Finding correlations between artificial grammar learning proclivities and language processing abilities contributes to validating the statistical learning approach to language. Future work could help establish whether people who are more sensitive to statistical sequential information make better language learners, and ultimately language users. In addition, understanding exactly what type of statistical information is required to optimize language tasks such as reading would greatly expand our knowledge of the mechanisms required for language processing. This line of research is not only useful to inform theories of language in

the brain, but has potential practical applications. For example, it may be possible to assist inefficient second-language readers by helping them process statistical information more optimally. Furthermore, our statistical learning task could be used to predict delays in language development, in cases where direct assessment of language is difficult (toddlers, or multilinguals for which assessment of language delays is confounded with proficiency in a given language).

Acknowledgments

This study was funded by: Singapore Ministry of Education's Tier 1 grant #M4011320 to LO; European Union Seventh Framework Programme grant 334028 to SF; Gravitation Grant 024.001.006 from the Netherlands Organization for Scientific Research to the Language in Interaction Consortium. We thank Shimon Edelman for commenting on a previous draft of this manuscript.

References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*, 952–964.

Arciuli, J., & Simpson, I. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*, 286–304.

Arciuli, J., & von Koss Torkildsen, J. (2012). Advancing our understanding of the link between statistical learning and language acquisition: the need for longitudinal data. *Frontiers in Psychology*, *3*. 324.

Benasich, A. A., Choudhury, N., Friedman, J. T., Realpe-Bonilla, T., Chojnowska, C., & Gou, Z. (2006). The infant as a prelinguistic model for language learning impairments: predicting from event-related potentials to behavior. *Neuropsychologia*, *44*, 396–411.

Baayen, R. H. & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Christiansen, M. H., Conway, C. M., & Onnis, L. (2012). Similar neural correlates for language and sequential learning: evidence from event-related brain potentials. *Language and Cognitive Processes*, *27*(2), 231-256.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*, 321–335.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE, 8*, e77661.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science, 22*, 829–834.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*, 1182–1190.

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*, 1243–1252.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5), 178–186.

Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? Comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*, 295–300.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*, 321–340.

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*, 171–184.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126-1177.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research, 43*, 1735–1751.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*, 268–284.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*, 1299–1305.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10, 233–238.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Smith N. J., Levy R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.

Yim, D., & Windsor, J. (2010). The roles of nonlinguistic statistical learning and memory in language skill. *Korean Journal of Communication Disorders*, *15*, 381–396.

The Permeability of Fictional Worlds

Meghan M. Salomon (msalomon@u.northwestern.edu)
Lance J. Rips (rips@northwestern.edu)
Department of Psychology, Northwestern University,
2029 Sheridan Road, Evanston, IL 60208-2710, USA

Abstract

Real people sometimes appear in fiction, for example, Napoleon in War and Peace. Readers may also believe that a person who never actually appears in a novel could potentially appear there. In two experiments, we find evidence that readers think that a real person could appear in specific novels and physically interact with a character. This effect is magnified when the person and character share spatial and temporal elements of their setting.

Keywords: fictional worlds; world knowledge; novels

Fictional Worlds

Just as people develop and maintain representations of real-world situations (e.g., Zwaan & Radvansky, 1998), previous research has established that children and adults keep track of fictional worlds: They possess mental representations of the universe in which a fictional work plays out. Investigators have argued that these representations portray a fictional world as discrete, self-contained, and distinct from the real world and from other fictional worlds. For example, preschoolers know that fictional characters like Batman can't touch or see them and can't touch or see characters like Elmo, indicating that these fictional worlds are distinct (Skolnick & Bloom, 2006; Skolnick-Weisberg & Bloom, 2009). Children can also extrapolate information from fictional situations and generalize them as "fantasy rules," using those rules to predict future fictional events (Van de Vondervoort & Friedman, 2014). In this way, children can learn from fantasy and use that information in fantastical settings. However, children tend not to transfer fictional ideas to real world situations (Richert, Shawber, Hoffman, & Taylor, 2009).

Fictional worlds aren't impermeable, however. Readers approach fictional stories with assumptions based on their knowledge of reality (Gerrig, 1993; Gerrig & Allbritton, 1990; Pollard-Gott, 1993; Gerrig & Rapp, 2004). Readers know to import general truths of reality, such as gravity and physiology, into a fictional text (Lewis, 1978; Skolnick-Weisberg & Goodstein, 2009). They assume that Jay Gatsby can sit upright in his chair and that Scarlett O'Hara breathes. Authors also rely on readers' knowledge of reality in importing more specific details into their fiction (Fillmore, 1981). For example, real people sometimes turn up in novels. Napoleon appears in War and Peace and Houdini in Ragtime (see Foulds, 2015). People may also believe that although a real individual doesn't actually appear in a fictional context, she potentially could. Queen Victoria never explicitly graces a Sherlock Holmes story, but readers may recognize this as possible.

What principles govern readers' judgments about whether real and fictional people can mingle? We report two experiments that examine the idea of "one-way permeability": People should be more apt to think that a real person could appear in a fictional world than the reverse. Perhaps paradoxically, Sherlock Holmes could potentially meet Queen Victoria, but Queen Victoria could hardly meet Sherlock Holmes. What is true in a fictional world may depend on the similarity in time and space between the fictional setting and the real one (Lewis, 1978), but in an asymmetric way. Experiment 1 looked at the effect of spatial distance on permeability, and Experiment 2 at temporal distance.[1]

Methodology

Participants in both studies read brief descriptions of novels, such as Jane Eyre and The Great Gatsby, that specified their author, protagonist, and publication date. The description also mentioned real people (e.g., Calvin Coolidge), along with a brief identifier for each. Participants then answered questions of the type: "Consider the world of Jay Gatsby [Calvin Coolidge]. Would Calvin Coolidge [Jay Gatsby] also exist in the same world?" The instructions stated that "being in the same world" means the individuals "could, at least in theory, meet and physically interact with one another." Participants were also cautioned, "… we do not want you to think of a person that merely has the same characteristics as that character…[n]or … a picture, movie, or other representation of the character. Rather you should think of the character him- or herself."

Experiment 1: Spatial Proximity

We expect that people will agree that real individuals *could* appear in a novel, even though they never actually appear there. But we also expect important restrictions to govern these cases. Just as two real people from the same

[1]Although the experiments we report here focus on simple overlap in space and time, we don't mean to imply that these are the only variables that affect readers' judgments of permeability. Many factors could have such an effect, including ideological, cultural, and technological ones. For example, our intuition is that a Russian czar of the 18th Century is less likely to appear in an 18th Century American novel than a Russian premier of the 20th Century in a 20th Century American novel. We leave these possibility for further research.

country are more likely to interact than two real people from different countries, a real person and a fictional character from the same country are more likely to interact than a similar pair from different countries. For example, although Calvin Coolidge could turn up in the world of Jay Gatsby, it is less likely that Joseph Stalin would appear there. Of course, crossings in the opposite direction should be even more limited, in line with one-way permeability: Coolidge could appear in Gatsby's world, but Gatsby could hardly appear in Coolidge's.

Method

Participants We recruited 30 participants (18 female), aged 18-63, (M=39.17, SD=12.34) through Mechanical Turk.

Procedure Participants in this study answered a series of questions about the novels that appear in Table 1. At the beginning of each block of trials, they read information about one of these novels (author, publication date, and protagonist). The preliminary information also described four real people—political leaders—two from the same country in which the novel was set, and two from a different country. Within each pair, one leader was relatively well known; the other less well known (see Table 1 for a complete list). Participants were given brief descriptions of the leaders and were told that all of them were alive at the time the novel was set.

Following the preliminary briefing, participants answered eight questions about the relations between the fictional characters and the real people. Participants decided whether, for example, Calvin Coolidge [Jay Gatsby] would exist in the world of Jay Gatsby [Calvin Coolidge], and whether Joseph Stalin [Jay Gatsby] would exist in the world of Jay Gatsby [Joseph Stalin]. For example, one question read, "Consider the world of Jay Gatsby. Would Calvin Coolidge also exist in the same world?"

There were seven blocks of trials, corresponding to the seven novels of Table 1. The blocks appeared in a random order, as did the eight questions within each block. After completing the questions, participants indicated which of the Table 1 books they had previously read. The experiment took approximately 20 minutes to complete.

Design The questions varied direction (could a fictional person appear in the real world vs. a real person in a fictional world), location (were the fictional character and the real person from the same country vs. a different country), and fame of the real-world individual (more famous vs. less famous). All factors varied within-subject.

To confirm our judgments of which individuals were more famous, we conducted a norming study. Thirty-one Mechanical Turk participants rated the fame of each real-world political leader from Table 1, on a scale from 1 ("least famous") to 7 ("most famous"). Those we designated more famous received a mean rating of 3.99, and those we designated less famous a mean rating of 1.75, $F(1,30) = 307.72$, $p < .001$. The names of the more and less famous political leaders appear in Table 1, with the more famous leader preceding the less famous one within each cell of columns 2 and 3.

Results and Discussion

Participants were significantly more likely to say that a real person exists in a fictional world (M=67% "yes" responses, SE=3%) than that a fictional person exists in the real world (M=38%, SE=3%), $F(1, 29) = 13.91$, $\eta_p^2 = .32$, $p < .001$. For example, participants were more likely to say that JFK exists in the world of Atticus Finch than that Atticus Finch exists in JFK's world. This supports the idea that fictional worlds are more permeable than the real world and, further, that readers can equip the fictional world of a novel with real people, even though the novel never explicitly mentions them. This main effect is shown in Figures 1a and 1b as the difference between the left- and the right-hand sides of the graphs.

Participants were also more likely to say that people/characters from the same country exist in the same world (M=59% "yes" responses, SE=3%) than that people/characters from different countries exist in the same world (M=46%, SE=3%), $F(1,29) = 23.02$, $\eta_p^2 = .44$, $p < .001$. For example, participants were more likely to say that JFK exists in the world of Atticus Finch than that Khrushchev exists in that world. Thus, readers take geographical distance into account in determining whether fictional and real-world individuals can intermingle. This suggests that participants interpreted our question (e.g., Would Calvin Coolidge exist in the world of Jay Gatsby?) in a probabilistic way (depending on spatial distance and other relevant factors) rather than in an absolute way (e.g., that it is not *impossible* that Coolidge could exist in Gatsby's world). Just as people think that the chance of two real-world people meeting is greater the smaller their spatial separation, people think that the chance of a fictional and a real-world person being in the same world is greater the smaller the distance between their spatial locations.

But although participants believed that spatial distance was relevant to whether real and fictional people could coexist, they did not believe that the real person's fame affected this issue, $F(1, 29) = 0.18$, $\eta_p^2 = .006$, $p = .678$. Participants' judgments did not depend on whether a real person was very well known or less well known. Even a lesser-known real person of the appropriate place and era is a candidate for existing in a fictional world. No interactions with fame were significant in these data.

Finally, we compared responses from participants who had read a book to those of participants who hadn't. We wanted to ensure that participants who had read, for example, Jane Eyre answered questions about this book in the same way as other participants. The results of this comparison for the main conditions appear in Table 2 and indicate very similar findings for readers and nonreaders. Reader or nonreader status did not significantly interact with either the effect of country or the effect of direction of transfer (all F's < 1). This similarity suggests that the descriptions of the novels in the instructions (or

participants' general knowledge about the novels) provided nonreaders with enough information to make sensible judgments.

Although participants judged real people more likely to appear in fiction than fictional characters in the real world, we did not find an absolute ban on fiction-to-real-world transfers. Participants judged that a fictional character could appear in the world of a real political leader on nearly 40% of trials. Some of these responses might be due to simple misreading or other low-level errors. For example, participants may have failed to distinguish the direction of the question (real-to-fiction vs. fiction-to-real). However, some participants may actually believe that the characters could surface in the real world, and we discuss this possibility in the General Discussion.

These results support our hypotheses in two ways. First, readers do import more into fiction than just (causally) necessary truths of reality. Their willingness to do so, however, depends on geographic relevance. Readers are likely to import real people into fictional worlds who are relevant to the novel's setting, even when these people are never mentioned in the text. This indicates that readers are attuned to the relevance of space, not only within the world (Zwaan & Radvansky, 1999), but also across worlds.

Figure 1a: Real people and characters in different countries

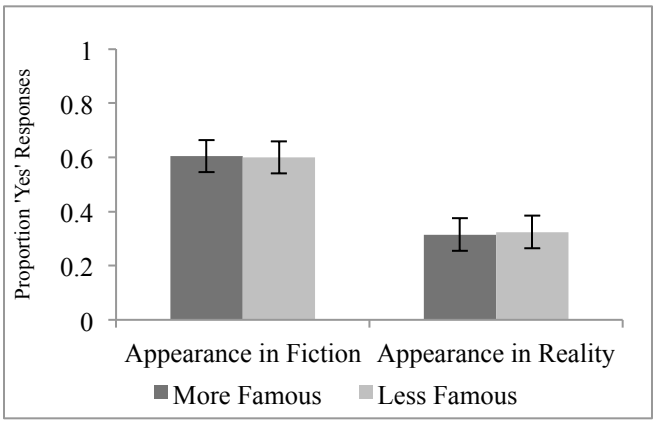


Figure 1b: Real people and characters in same country

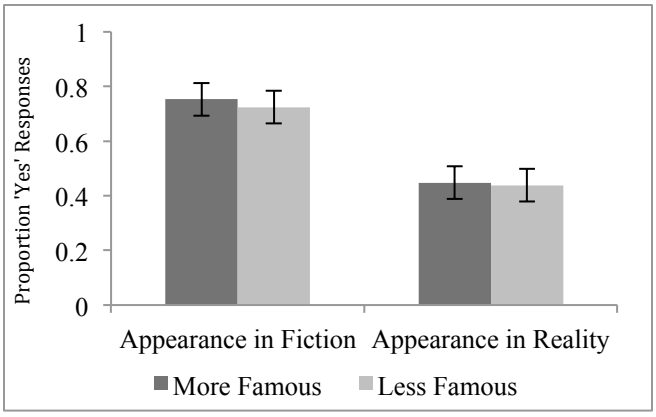


Table 1: Novels used in Experiment 1

Novel	Real Person (Same Location)	Real Person (Different Location)
Jane Eyre (Jane Eyre)	Queen Victoria; Earl Russell	James Polk; John Calhoun
Moby Dick (Captain Ahab)	Millard Fillmore; Stephen Douglas	Napoleon; General Jacques Leroy de Saint Arnaud
Sherlock Holmes (Sherlock Holmes)	William Gladstone; Charles Parnell	Grover Cleveland; Levi Morton
The Great Gatsby (Jay Gatsby)	Calvin Coolidge; Charles Evans Hughes	Joseph Stalin; Nikolai Bukharin
A Christmas Carol (Ebenezer Scrooge)	Prince Albert; Lord Melbourne	Louis Philippe; Francois Guizot
Breakfast at Tiffany's (Holly Golightly)	Dwight Eisenhower; Estes Kefauver	Queen Elizabeth II; Harold Macmillan
To Kill a Mockingbird (Atticus Finch)	John Kennedy; Edmund Brown	Nikita Khrushchev; Georgy Malenkov

Table 2: Percentage of Readers' and Nonreaders' Agreement that Real and Fictional People could be in the Same World, Experiment 1

Direction of transfer	Spatial separation	Non-readers	Readers
Fiction-to-real	Different countries	33.2%	30.1%
	Same country	45.1	43.2
Real-to-fiction	Different countries	61.5	58.5
	Same country	75.4	71.6

Experiment 2: Temporal Proximity

Experiment 1 showed that spatial restrictions govern how easily a real person can appear in a novel. Temporal factors should produce similar effects, perhaps in a more dramatic way. Although Robert E. Lee might appear in a novel set during the Civil War, such as Gone with the Wind, we would not expect Franklin Roosevelt to appear there. Time travel occurs in some science fiction and fantasy, but in the more realistic novels we used here, we don't expect to find real people from a different era.

Method

Participants We recruited 32 participants (14 female), aged 22-62 (*M*=37.16, *SD*=10.29) through Mechanical Turk. One subject was eliminated for answering "no" to all questions.

Materials. In Experiment 2, we paired novels published in the same year, one set in the past (e.g., Gone with the Wind) and the other set at about the time of publication

(e.g., In Dubious Battle). For each pair, we chose one political leader who was alive at the time of the earlier setting (e.g., Robert E. Lee) and one who was alive at the time of the later setting (e.g., Franklin Roosevelt). We then asked about the relationship between these real and fictional characters (in both orders), as in Experiment 1. For example, participants decided whether Robert E. Lee would exist in the world of Scarlett O'Hara, whether FDR would exist in the world of Scarlett O'Hara, whether Robert E. Lee would exist in the world of Jim Nolan, and whether FDR would exist in the world of Jim Nolan. We also asked the reverse of each of these questions (e.g., whether Scarlett O'Hara would exist in the world of Robert E. Lee). The question wording was the same as in Experiment 1 (e.g., "Consider the world of Scarlett O'Hara. Would Robert E. Lee also exist in the same world?"). Table 3 lists the full set of novels, characters, and real people. At the end of the experiment, participants indicated which of the novels mentioned in the experiment they had read.

Procedure Apart from these differences in materials, the experiment proceeded in the same way as did Experiment 1.

Design The questions varied direction (could a fictional person appear in the real world vs. a real person in a fictional world), era of the fictional character (historical vs. modern, relative to the date of publication), and era of the real person (historical vs. modern, relative to the date of publication). All factors were within-subject.

Results and Discussion

Participants were much more likely to say that two people exist in the same world if they were from the same era (M=82% "yes" responses, SE=2%) than if one person existed before the other (M=26%, SE=2%). The results are shown in Figures 2a and 2b: "Yeses" were more common when the time of the real people and characters matched than when their times mismatched, $F(1,28) = 88.53$, $p < .001$. For example, participants thought that Robert E. Lee was more likely than FDR to exist in the same world as Scarlett O'Hara.

As in Experiment 1, participants judged it easier for real people to appear in fiction than for fictional characters to appear in the real world. In the present experiment, however, this effect appeared only for situations in which the person and the character were from the same era (M=88%, SE=2% for real-to-fiction direction vs. M=76%, SE=2% for fiction-to-real direction). When the eras of the character and the real person mismatched, participants were unlikely to say that either could exist in the world of the other (M=26%, SE=2% for both the real-to-fiction and the fiction-to-real directions). This produced a three way interaction of direction, era of the fictional character, and era of the real person that was on the cusp of significance, $F(1,28) = 4.02$, $p = .05$. These results echo those of Experiment 1 in suggesting that participants were not just considering whether real people could appear in fiction in an all-or-none way. Instead, their decisions depended on probability or relevance, based on temporal overlap. Although both Robert E. Lee and FDR *could* appear in the world of Gone with the Wind in a logical or metaphysical sense of "could," participants were also taking into account the relative likelihood of such an appearance, based on temporal distance.

As in Experiment 1, we compared responses from participants who had read a given book to those from participants who hadn't. Although there were too few readers for some of the books to permit a meaningful statistical analysis, the trends for readers and nonreaders are quite similar, as Table 4 shows.

In Experiment 1, we found that participants believed real people were more likely to be present in fictional worlds when the real people were in the same locale as the characters. The data from Experiment 2 reinforce the importance of setting by showing that readers believe that real people are more likely to exist in a fictional world when the time of the novel aligns with that of the real person. When the times mismatch, there's little possibility of world crossing in either direction.

General Discussion

Both experiments confirmed one-way permeability: Participants in Experiment 1 were more likely to agree that a real person could exist in the world of a fictional character than that a fictional character could exist in the world of a real person. In Experiment 2, this difference appeared when the time of the real and fictional people matched, but not when it mismatched.

Permeability was also greater in the second experiment when both the characters and real people were from the past than when both were from the time the novel was written. Participants agreed on 86% of trials that two individuals from the past could inhabit the same world, but agreed on 79% of trials that two individuals from the time of publication could inhabit the same world, $F(1,28) = 5.74$, $p = .024$, by a planned comparison. This is a possible effect of genre: Historical novels may be more permeable than contemporary ones. The author of a historical novel typically has in mind a definite real-world place and time, and the importance of this setting (and accompanying realism) may invite readers to believe that aspects of that setting carry over to the world of the novel. For example, the importance of the Civil War milieu to Gone with the Wind may be a pragmatic indication that the reader should incorporate real-world facts from that milieu into the novel (see Byrne, 1993). By contrast, a particular time and place may be less crucial to contemporary novels, providing readers with fewer guidelines about which aspects of reality will continue to hold in the fictional world.

In thinking about the match between fictional worlds and the real world, however, we should be careful to observe that similar settings may be effective because they determine which causal influences from the real world could penetrate the fictional world. For example, readers believe it is unlikely that Jay Gatsby would encounter Abraham

Lincoln due to the temporal discrepancy. But the temporal mismatch may be important because of causal laws that govern people's lifetimes, not because of the absolute difference in era. It could well be that the Emancipation Proclamation influences the world of Jay Gatsby, despite the difference in time between the issuing of the Proclamation and the Jazz Age.

We noted that participants in Experiment 1 sometimes judged that fictional characters could appear in the real world. Figures 2a and 2b show that the same was true in Experiment 2, especially when the eras of the characters and real people matched. In line with one-way permeability, these judgments were less common than those in which real people were deemed part of a fictional world, but what prompted these decisions? One possibility is a confusion between different categories of hypothetical people. Consider a couple who decides to have a family and even decides on a name for their first-born daughter, say, "Olivia" (cf. Perry, 2001). If the couple happens to break up before they can start their family, then we might say that Olivia would have existed in the real world, but for the unfortunate break up. This case differs, however, from that of hypothetical people who are parts of a fictional work, like the Olivia in Twelfth Night or the characters in our experiments. The fictional status of these individuals prohibits them from being part of the real world (a prohibition that applies to Shakespeare's Olivia but not to the couple's Olivia), and enforces one-way permeability. It seems possible, however, that some participants answered the questions about whether a fictional character would have existed in the real world in a way that is appropriate to a merely hypothetical person. These participants may decide on the basis of the character's realistic qualities that he or she would have existed under the right circumstances, ignoring the character's fictional status. In this vein, one participant commented, "A well written character, if they could come out of the pages, would be able to interact with a real person in their era."

In sum, fictional worlds are not completely isolated. People believe that real-world figures (and presumably other real-world objects and events) could potentially appear in fictional worlds, even if they don't appear overtly. While readers sometimes export information from a fictional text (e.g., Gerrig & Prentice, 1991), these results indicate they also import people from reality.

Acknowledgments

We thank three anonymous reviewers for helpful comments on this paper. Work on the paper is due in part to the US Department of Education, Institute of Education Sciences, Multidisciplinary Program in Education Sciences, Grant Award # R305B140042.

Figure 2a: Historical Novels

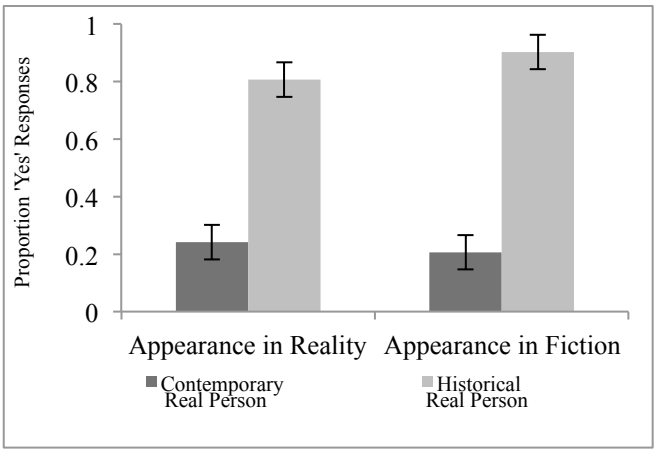


Figure 2b: Contemporary Novels

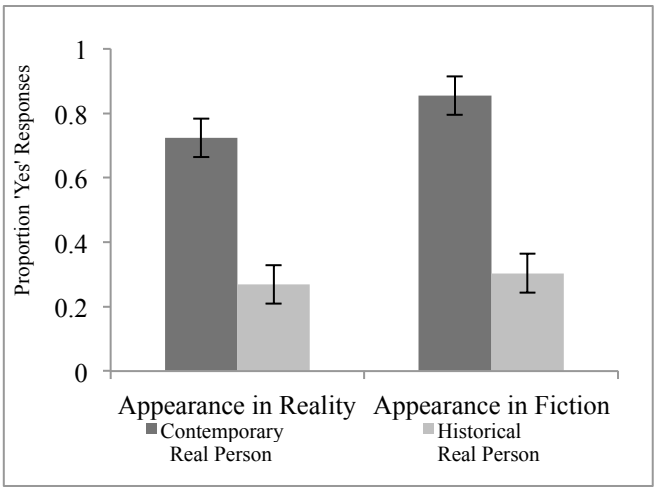


Table 3: Novels used in Experiment 2

Historical Novel	Contemporary Novel	Historic Figure	Contemporary Figure
Gone with the Wind (Scarlett O'Hara)	In Dubious Battle (Jim Nolan)	Robert E. Lee	Franklin Delano Roosevelt
Cold Mountain (W. P. Inman)	The Partner (Patrick Lanigan)	Abraham Lincoln	Bill Clinton
The French Lieutenant's Woman (Sarah Woodruff)	Travels with my Aunt (Henry Pulling)	Queen Victoria	Queen Elizabeth II
War and Peace (Pyotr Kirillovich Bezukhov)	The Idiot (Lev Nikolayevich Myshkin)	Alexander I	Alexander II
Blood Meridian (The Kid)	White Noise (Jack Gladney)	Andrew Jackson	Ronald Reagan

Table 4: Percentage of Readers' and Nonreaders' Agreement that Real and Fictional People could be in the Same World, Experiment 2

Direction of transfer	Temporal relation	Non-readers	Readers
Fiction-to-real	Modern character Modern real person	71.5%	80.0%
	Modern character Historical real person	26.9	26.7
	Historical character Modern real person	24.4	23.1
	Historical character Historical real person	79.0	88.5
Real-to-fiction	Modern character Modern real person	86.2	80.0
	Modern character Historical real person	29.2	40.0
	Historical character Modern real person	20.2	23.1
	Historical character Historical real person	88.2	100.0

References

Byrne, A. (1993). Truth in fiction: the story continued. *Australasian Journal of Philosophy, 71,* 24 – 35.

Fillmore, C. J. (1981). Ideal readers and real readers. In D. Tannen (Ed.), *Analyzing discourse: Text and talk* (pp. 248-270). Washington D. C.: Georgetown University Press.

Foulds, A. (2015). Writing real people. In Z. Leader (Ed.), *On life-writing* (pp. 99-111). Oxford, UK: Oxford University Press.

Gerrig, R. J. (1993). *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.

Gerrig, R. J., & Allbritton, D. W. (1990). The construction of literary character: A view from cognitive psychology. *Style*, *24*, 380-391.

Gerrig, R. J., & Prentice, D. A. (1991). The representation of fictional information. *Psychological Science, 2*, 336-340. doi: 10.1111/j.1467-9280.1991.tb00162.x

Gerrig, R. J., & Rapp, D. N. (2004). Psychological processes underlying literary impact. *Poetics Today*, *25*, 265-281.

Lewis, D. (1978). Truth in fiction. *American Philosophical Quarterly, 15*, 37-46.

Perry, J. (2001). *Reference and reflexivity*. Stanford, CA: CSLI Publications.

Pollard-Gott, L. (1993). Attribution theory and the novel. *Poetics*, *21*, 499-524.

Richert, R. A., Shawber, A. B., Hoffman, R. E., & Taylor, M. (2009). Learning from fantasy and real characters in preschool and kindergarten. *Journal of Cognition and Development, 10*, 41-66.

Skolnick, D., & Bloom, P. (2006). What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction. *Cognition, 101*, B9-B18. doi: 10.1016/j.cognition.2005.10.001

Skolnick-Weisberg, D., & Bloom, P. (2009). Young children separate multiple pretend worlds. *Developmental Science, 12*, 699-705. doi: 10.1111/j.1467-7687.2009.00819.x

Weisberg, D. S., & Goodstein, J. (2009). What belongs in a fictional world? *Journal of Cognition and Culture*, *9*, 69-78.

Van de Vondervoort, J. W., & Friedman, O. (2014). Preschoolers can infer general rules governing fantastical events in fiction. *Developmental Psychology*, *50*, 1594.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185. doi: 10.1037/0033-2909.123.2.162

The development of heuristics in children: Base-rate neglect and representativeness

Samantha Gualtieri (sgualtieri@uwaterloo.ca) & Stephanie Denison (stephanie.denison@uwaterloo.ca)
University of Waterloo, Department of Psychology

Abstract

This paper examines the development of the representativeness heuristic in early childhood. Using a novel paradigm, we investigated 3- to 6-year-old children's ability to use base-rate and individuating information in their predictive inferences. In Experiment 1, we presented children with base-rate and individuating information separately to test their ability to use each independently. In Experiment 2, we presented children with base-rate and individuating information together. Two critical trial types were used, one in which the base-rate information and individuating information pointed to the same response and one in which the base-rate and individuating information pointed to conflicting responses. Results suggest that children progress to adult-like heuristic-based responding at 6 years of age.

Keywords: decision making, base-rate neglect, heuristics

Introduction

From their first year of life, human infants are able to use numerical information in their inductive inferences. For instance, infants expect a random sample drawn from a larger distribution to resemble the larger distribution (Teglas et al., 2007; Xu & Garcia, 2008). However, adults tend to under use numerical base-rate information and over rely on other types of information in their predictive inferences (Kahneman & Tversky, 1973). These two literatures pose a conundrum for cognitive developmentalists: if humans can use base-rates to make inferences as early as 12 months, why are adults so inclined to neglect them and when does the tendency to do so emerge?

In their seminal work, Kahneman and Tversky (1973) outlined notable cases of adults neglecting base-rates. In the classic lawyer-engineer problem, participants read a personality description that was randomly selected from a sample of 70 lawyers and 30 engineers. This description was of a person who was conservative, enjoyed puzzles, and did not care for social issues. When asked to report whether the person was a lawyer or an engineer, participants reported it was more likely that the person was an engineer. That is, people neglected base-rates (i.e., the number of lawyers and engineers) and over used individuating information (i.e., the personality description) in their estimates. Kahneman and Tversky (1973) referred to this as the representativeness heuristic: people over rely on individuating information that fits their representation of a group's characteristics and ignore important factors, such as base-rates, resulting in biased judgments in some cases.

The vast literature examining normative (often also referred to as analytic) versus non-normative (often also referred to as heuristic) responding often includes two assumptions (see Kokis et al., 2002 for a review). The first assumption, which has received mixed empirical support, is that heuristic responding should decrease in adult samples as intelligence increases (adults of higher intelligence should override a heuristic response more often, in favour of the normative response). A related, though largely untested assumption is that heuristic responding should decrease with age, as the ability to override the heuristic response should improve over development. However, it is difficult to draw conclusions from the few studies that have investigated the development of base-rate neglect and the representativeness heuristic in childhood, as their designs lack features that have proven vital to understanding the adult work. One pivotal aspect of Kahneman and Tversky's (1973) problems is the familiarity of the group information to participants. In the previous lawyer-engineer example, adults believe the person is an engineer, as the description they are given closely resembles their representation or stereotype of an engineer. In order to use this heuristic, one has to be familiar with how individuating information can be used to classify a specific case. It is important to consider young children's experience with group information when examining their use of the representativeness heuristic (Stanovich, West, & Toplak, 2011). For example, Jacobs and Potenza (1991) and Davidson (1995) found age-related *decreases* in base-rate use, with the youngest children (6- and 7-year-olds) providing more normative responses than older children and adults in tasks that were adapted for children. These results were interpreted as revealing a counter-intuitive idea: younger children actually make better predictive judgments than older children and adults, as they were more likely to provide the response that was closer to the base-rates. However, participants in these studies were presented with group information that may have been unfamiliar to the youngest children in the sample (e.g., cheerleader and band member stereotypes). Since it is reasonable to assume that the youngest children did not have the relevant group information to begin with, children's use of base-rates may have arose from being unaware of the category information, rather than overriding a heuristic response with a normative one (Stanovich et al., 2011).

One previous study has highlighted this important problem by attempting to manipulate the familiarity of stereotype information across ages. De Neys and Vanderputte (2011) investigated 5- and 8-year-old children's responses to base-rate problems that used both familiar and unfamiliar group information. Their results supported the claim that younger children might be providing normative responses only when they are unfamiliar with the presented stereotypes (and thus only have base-rates to go on). However, attempting to find ages at which these stereotypes are emerging would help determine if the representativeness heuristic is used when both types of information are available to the child.

Using a novel paradigm, we examined young children's use of base-rate and individuating information in their

inferences. In Experiment 1, we established whether 3- to 5-year-old children could use base-rate and individuating information when presented alone. No previous studies have established whether children could use base-rate and individuating information separately in their tasks, which is vital to understanding how children consider this information when presented together. In Experiment 2, we presented 3- to 6-year-old children with both base-rate and individuating information. Analogous to the adult literature, we investigated children's inferences when presented with individuating information that conflicted with base-rate information. To examine the integration of information in young children further, we also included problems in which the two types of information did not conflict.

Experiment 1: Methods

Participants

Children in both experiments were tested individually in lab, at their school or daycare, or at a local museum. Forty-eight children participated in the base-rate condition, including 16 three-year-olds, 16 four-year-olds, and 16 five-year-olds (*Mean age* = 4 years, 4 months, *range* = 3 years, 0 months, 9 days to 5 years, 11 months, 20 days, *female* = 20). An additional child was tested but excluded for failing to correctly identify the majority group on both problems (see Procedure for details). Forty-eight children participated in the individuating condition, including 16 three-year-olds, 16 four-year-olds, and 16 five-year-olds (*Mean age* = 4 years, 6 months, *range* = 3 years, 0 months, 27 days to 5 years, 11 months, 29 days, *female* = 24). An additional child was tested but excluded for non-compliance.

Procedure, Design and Predictions

Participants in both conditions completed two problems, presented consecutively. The experimental session was presented to the children on a laptop, using a PowerPoint presentation, which was narrated live by an experimenter. On the first slide, children were told that they were going to hear about some robots on another planet, and that they would have to answer some questions about the robots after.

In the base-rate condition, each participant completed a color and a shape problem, in counterbalanced order (see Figure 1 for a diagram of the base-rate condition procedure). Two problem-types (color and shape) were used to maintain children's attention. In the color problem, following the initial introduction, participants next saw a slide with two robots standing side by side at a library, one robot wearing red and another robot wearing purple. On the following two slides, participants saw each robot by itself, and the researcher pointed out what color the robot was wearing. The next slide included both of the robots standing side by side again, and the researcher asked the child to point to the one wearing red, and the one wearing purple. On the next slide, participants saw a group of ten robots at the library, eight wearing red, and two wearing purple. Children were asked to indicate which type of robot there was more of. After the child provided their answer, the experimenter indicated that there were lots of robots wearing red, and very few wearing purple. Following this, children saw a slide with a single robot wearing a white coat, making its type unclear. Children were asked to recall which type of robot there was more of in the group. Participants were then asked to indicate which type of robot they thought the one wearing the white coat was. This procedure was also followed for the shape problem; however, participants saw one type of robot wearing hearts and another type of robot wearing stars. To ensure participants did not confuse the robots with those from the previous problem, the shape problem took place at a grocery store.

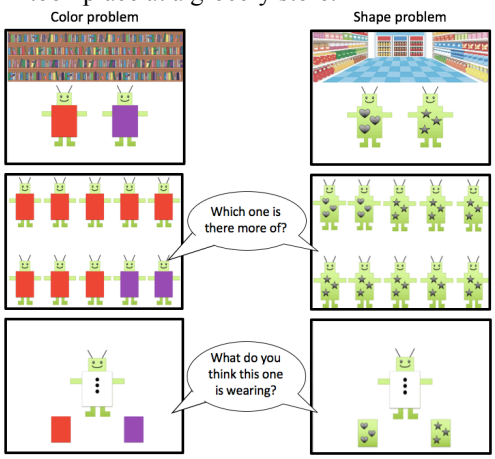


Figure 1: Base-rate condition procedure.

In the individuating condition, each participant completed a trait and a gender problem in counterbalanced order (see Figure 2 for a diagram of the individuating condition procedure). Young children have successfully considered gender stereotypes and trait-based categories in other work (Heyman & Gelman, 2000; Martin & Ruble, 2004). As adults are familiar with occupational stereotypes prior to completing the lawyer-engineer problem, we chose both of these trial types based on young children's familiarity with these categories. In the trait problem, following the initial introduction, participants saw a slide with two robots standing side by side at a park, one robot wearing blue and another robot wearing green. They were told that most of the time the robots wearing blue were nice, and the ones wearing green were naughty. On the next slide, the robot wearing blue was holding flowers, and participants were told it would bring flowers to another robot on its birthday. On the following slide, the robot wearing green was hiding presents, and participants were told that this robot would hide another robot's birthday presents. Participants then saw the blue and green robot standing side by side on the following slide. Participants were asked to indicate which type of robot was nice most of the time, and which type of robot was naughty most of the time. Following this, children saw a slide with a single robot wearing a white coat, making its type unclear. Children were asked to recall which traits were associated with each color. The experimenter then provided some additional information about the robot wearing the white coat, which

matched one of the traits. Half of the children heard a description of nice behavior, as the experimenter said that it liked to clean up the park and help its friend play on the swing. The other half of the children heard a description of naughty behavior, as the experimenter said that it liked to make a mess of the park and scare other robots. Importantly, there was no mention of either the word "naughty" or "nice"; children were only told about the mystery robot's behaviour. Participants were asked to indicate which type of robot they thought the one wearing the white coat was. This procedure was also followed for the gender problem, however participants saw one type of robot wearing yellow and another type of robot wearing orange. They were told that most of the time the robots wearing yellow liked to play with toys that girls like, and those wearing orange liked to play with toys that boys like. When introduced to the yellow robot individually, it had mostly toys typically associated with girls (i.e., a unicorn and a doll). When introduced to the orange robot individually, it had mostly toys typically associated with boys (i.e., a dinosaur and a helicopter). The individuating information was as follows: Half of the children heard a description more typical of girls, as the experimenter said it liked to play dress-up and house. The other half of the children heard a description more typical of boys, as the experimenter said that it liked to play with trucks and train sets. To ensure participants did not confuse the robots with those from the previous problem, the gender problem took place at a school.

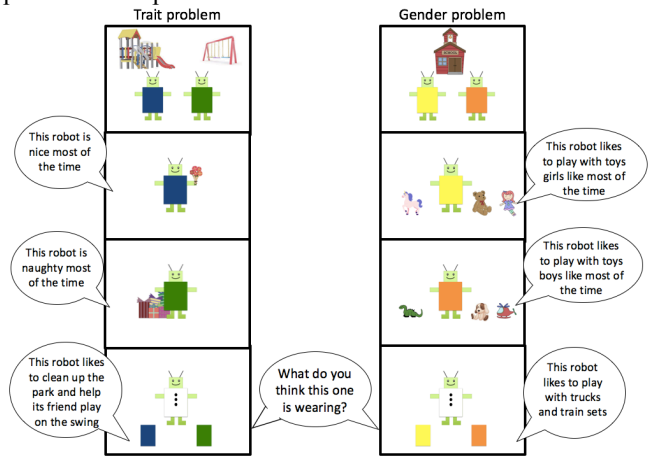


Figure 2: Individuating condition procedure.

Design In both conditions, the position of the two types of robots was counterbalanced in each problem (e.g., half of the children saw the orange robot on the left and half saw it on the right when they were introduced). The experimenter always began by telling the child about the robot on the left. In the base-rate condition, the robots that comprised the majority group were counterbalanced across both problems. The position of the majority group was also counterbalanced in both problems, as the majority was either presented as the first eight or last eight robots in the group of ten. In the individuating condition, the individuating information given about the robot wearing the white coat was counterbalanced as representative of each type.

Predictions If children can use base-rate information to make a predictive inference, then they should choose the type of robot that corresponds to the majority group in that condition. If children can use individuating information (specifically, the individuating information about these trait-based categories and gender stereotypes) to make a predictive inference, then they should choose the color corresponding to the individuating information given in that condition. Thus, children received a score of 1 for a correct response in each condition on each problem.

Experiment 1: Results

Base-rate condition. A repeated-measures Analysis of Variance (ANOVA) with score on each problem (color, shape) as a within-subjects factor and age (3-, 4-, 5-year-olds) as a between-subjects factor revealed a significant interaction (F (2, 45) = 7.82, p = .001, η^2_p = .26). No main effects of age (p = .11), or problem type (p = .43) were found. Post hoc analyses revealed that the interaction was driven by differences in performance on the shape problems, as 3-year-olds significantly differed from 4-year-olds ($Mean_{Difference}$ = -.44, p = .01) and 5-year-olds ($Mean_{Difference}$ = -.44, p = .01).

Overall performance was significantly different from chance, as children tended to indicate that the robot belonged to the majority group (M = 1.5, SD = .77, t (47) = 4.49, p < .001). As we were interested in development across age, we also examined children's performance at each age separately. T-tests revealed that the performance of 4-year-old (M = 1.56, SD = .73, t (15) = 3.09, p = .007) and 5-year-old (M = 1.75, SD = .68, t (15) = 4.39, p = .001) children were significantly different from chance. Three-year-olds' overall performance did not differ from chance (M = 1.19, SD = .83, p = .34). To investigate this further, we compared 3-year-olds' performance on color versus shape problems with a paired-samples t-test to determine whether they found one problem more difficult than the other. There was a significant difference in performance (t (12) = 2.31, p = .040), as 3-year-olds performed above chance on color problems (M = .75, SD = .45, t (16) = 2.24, p = .041), but not on shape problems (M = .44, SD = .51, p = .63).

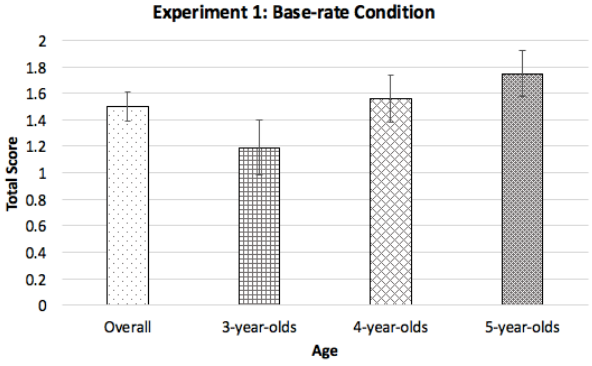


Figure 3: Base-rate condition results by age.

Individuating condition. A repeated-measures ANOVA with score on each problem (gender, trait) as a within-subjects factor and age (3-, 4-, 5-year-olds) as a between-subjects factor revealed a main effect of age on performance (F (2, 45) = 7.12, p = .002, η^2_p = .24), and no main effect of problem type (p = 1), or interaction (p = .44). Post-hoc tests revealed the main effect of age was driven by differences in the performance of 3- and 5-year-old children ($Mean_{Difference}$ = .88, p = .001).

Overall performance was significantly different from chance, as children indicated that the robot belonged to the group the individuating information corresponded with (M = 1.42, SD = .74, t (47) = 3.91, p < .001). Additional t-tests revealed 5-year-old children performed significantly above chance (M = 1.88, t (15) = 10.25, p < .001). However, 4-year-olds were only marginally different from chance (M = 1.38, p = .08), and 3-year-olds did not exceed chance (M = 1, p = 1).

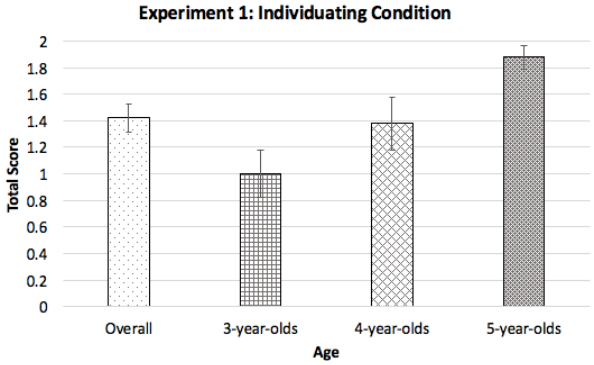


Figure 4: Individuating condition results by age.

Experiment 1: Discussion

Using a novel paradigm, we established that children are able to use base-rate and individuating information separately to make a predictive inference, but that these abilities are still developing between 3 and 5 years. When base-rate information was made salient, 4- and 5-year-old children used base-rates to predict group membership at above chance levels. Although 3-year-old children's overall performance was not above chance, they were able to use base-rate information when group membership was presented using color. When individuating information was made salient, 5-year-old children used the individuating information to predict group membership at above chance levels. Four-year-olds approached significance, using the information correctly approximately 70% of the time. However, 3-year-old children performed at chance levels, suggesting that they are unable to use individuating information to make a predictive inference in our task.

Having examined the ability to use each type of information separately in young children, the next step is to extend the task to an analogue of the classic lawyer-engineer problem in Experiment 2. We continued to test 3-, 4- and 5-year-olds in Experiment 2, as these age groups displayed a range of abilities for using the individuating information presented to them. Each age group successfully used base-rate information when presented using color, so all problems in Experiment 2 use color. This set-up, in which all ages can rely on base-rates, thus holding that ability more or less constant but varying the ability to use individuating information, could be particularly informative. It allows investigation of questions such as: At what ages are children able to integrate information from both of these sources? Will children show bias based on representativeness as soon as they can reliably use individuating information (at 5 years of age)? Or does this bias develop over time (will there be a delay in its appearance)? To examine this progression, the age range is extended to 6 years.

Experiment 2: Methods

Participants

Sixty-four children participated, including 16 three-year-olds, 16 four-year-olds, 16 five-year-olds, and 16 six-year-olds (*Mean age* = 5 years, 1 month, *range* = 3 years, 6 months, 24 days to 6 years, 11 months, 27 days, *female* = 31). An additional seven children were tested, but were excluded due to interference from another child (n = 3) or failing to correctly identify the majority group in both problems (n = 4).

Procedure, Design and Predictions

Participants completed two problems, presented consecutively. As in Experiment 1, children completed a trait and a gender problem, using the same color, group, and individuating information as before. Each child saw one color pair and the associated group information on the first trial, and another color pair and associated group information on the second trial.

Using the trait problem as an example, children were first introduced to the two types of robots, and were told about the traits that corresponded to each color, just as in Experiment 1 (e.g., nice blue and naughty green robots). Once again, they heard some information about the robots' behavior at a birthday party. Following this, participants saw the robots side by side and were asked to indicate which robot was nice and which one was naughty most of the time. Next, participants saw a group of ten robots, with eight wearing blue, and two wearing green. Children were asked to indicate which type of robot there was more of at the park. After the child provided their answer, the experimenter indicated that there were lots of robots wearing blue, and just a few wearing green. Following this, children saw a slide with a single robot wearing a white coat, making its type unclear. To ensure both pieces of information were equally salient to the child, the experimenter reminded the child in a counterbalanced order that most of the time the blue robots were nice and the green robots were naughty, and that there were more robots wearing blue at the park. In each problem, base-rate and

group information were explicitly mentioned twice to the child prior to hearing the individuating information about the robot. The experimenter then provided some additional information about the robot wearing the white coat, which corresponded to one of the group traits. As in Experiment 1, for the "nice" trait, the experimenter said that it liked to clean up the park and help its friend play on the swing. For the "naughty" trait, the experimenter said that it liked to make a mess of the park and scare other robots. Participants were asked to indicate which type of robot they thought the one wearing the white coat was.

Each participant completed a conflict and a no-conflict problem. In the conflict problem, the base-rate conflicted with the individuating information given about the unknown robot. For instance, if there were eight nice robots, the unknown robot was described as liking to make a mess of the park and scare other robots. In the no-conflict problem, the base-rate and individuating information both cued the same group. For instance, if there were eight nice robots, the unknown robot was described as liking to clean up the park and help its friend play on the swing.

Design The order in which the child completed the conflict and no-conflict problems was counterbalanced across participants. Further, the group information associated with the robots in the conflict and no-conflict problems was counterbalanced (i.e., half of the children completed a conflict problem about the trait story, and the other half completed a no-conflict problem about the trait story). The order of reminding about base-rate and group information was also counterbalanced. All factors that were counterbalanced in Experiment 1 were also counterbalanced here (i.e., color and position of majority, the position of the two types of robots, the type of robot the individuating information represented).

Predictions In both problems, the normatively correct answer is to produce a response that aligns with the base-rates (and in the no-conflict problems, this also corresponds with a response from individuating information). In no-conflict problems, children should provide the response consistent with both the base-rate and individuating information, as these will point to the same response. In conflict problems, if children are biased toward using individuating information when it conflicts with the base-rate, they should choose the color corresponding to the individuating information, rather than the base-rate information, similarly to adults in the classic tasks. However, if children do not yet have this reasoning bias, or if they do not fully grasp the stereotype/category information, then they might provide responses consistent with base-rates, rather than individuating information. As each child completed one conflict and one no-conflict problem, they received a score of 0 or 1 for each problem type. In both problems, children received a score of 1 if they chose the majority group.

Experiment 2: Results

A repeated-measures ANOVA including score on each problem type (conflict, no-conflict) as a within-subjects factor and age (3-, 4-, 5-, 6-year-olds) and information recapped first (base-rate, individuating) as between-subjects factors revealed a significant interaction of age and problem type (F (3, 56) = 5.66, p = .002, η^2_p = .23). There was a main effect of problem type (F (1, 56) = 18.67, p < .001, η^2_p = .25). There were no other main effects or interactions.

The interaction was driven by an effect of age on the no-conflict problems ($Mean_{No\text{-}conflict}$ = .70, SD = .46; F (3, 60) = 5.91, p = 0.001). Post-hoc analyses revealed 6-year-olds' scores on no-conflict problems significantly differed from 3-year-olds ($Mean_{Difference}$ = .44, p = .02) and 4-year-olds ($Mean_{Difference}$ = .56, p = .002). The difference between 4- and 5-year-old children on no-conflict problems also approached significance ($Mean_{Difference}$ = -.38, p = .06; see Figure 5). There was no effect of age on performance in the conflict problems ($Mean_{Conflict}$ = .38, SD = .49; p = .38).

Additional t-tests revealed age differences on conflict and no-conflict problems. On conflict problems, 3-year-olds (M = .44, SD = .51, p = .63), 4-year-olds (M = .44, SD = .51, p = .63), and 5-year-olds (M = .44, SD = .51, p = .63) did not exceed chance. Six-year-old children were significantly different from chance (M = .19, SD = .4, p = .007). On no-conflict problems, 3-year-olds (M = .56, SD = .51, p = .63) and 4-year-olds (M = .44, SD = .51, p = .63) did not exceed chance. Five-year-olds were significantly different from chance (M = .81, SD = .40, t (15) = 3.10, p = .007), and six-year-olds were at ceiling (M = 1).

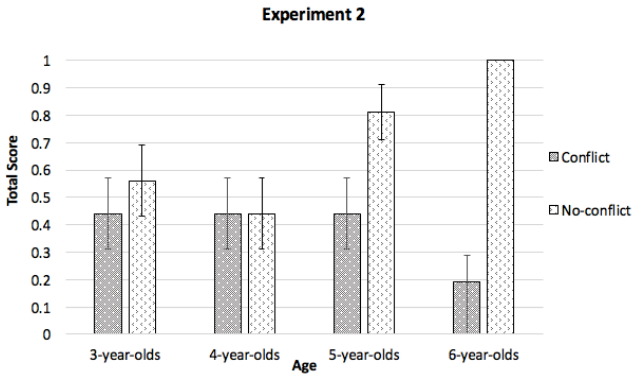


Figure 5: Experiment 2 results by age.

Experiment 2: Discussion

The results of Experiment 2 revealed interesting age differences in children's use of base-rates. Three- and four-year-olds performed at chance levels on both no-conflict and conflict problems, even though the normative response in the no-conflict problems corresponded with both base-rate and individuating information. Perhaps the 4-year-olds, who were starting to use individuating information on its own, but not entirely reliably, face issues integrating both pieces of information at first, due to a cognitive decoupling problem (see Stanovich, et al., 2011). The responses of 5-

and 6-year-old children in our task most strongly resembled the typical patterns of adults[1]. In no-conflict problems, both 5- and 6-year-old children provided the normative response at above chance levels, suggesting that by 5, children can fully handle the task. Interestingly, 5-year-old children did not show a pronounced bias for relying on individuating information, as they were at chance on conflict problems. Six-year-olds showed the predicted pattern from adults on the conflict problems, suggesting that this bias is fairly well ingrained, even at this young age.

General Discussion

In two experiments, we were able to verify methods for examining base-rate neglect and the representativeness heuristic in young children. We established the ages at which children can use base-rate and individuating information, when presented separately and made highly salient, in Experiment 1. Based on the extent to which children were able to use base-rate and individuating information in Experiment 1, we found interesting age differences in performance when both pieces of information were available in Experiment 2. Although three-year-olds could use base-rate information independently, they did not show a preference for base-rate information on conflict or no-conflict problems. Four-year-olds, who were able to use both base-rate and individuating information independently, also struggled when both pieces of information were presented together. Five-year-old children performed above chance on no-conflict problems, though they did not consistently rely on base-rate or individuating information on conflict problems. By the age of 6, children's responses revealed a pattern reminiscent of adult performance. These age differences provide useful insight on the representativeness heuristic, suggesting that its emergence may not be as straightforward as previously thought. Further research with 4- and 5-year-old children would provide insight on the nuances in the development of this bias.

Much additional research can be conducted to further examine the development of base-rate neglect. For example, ongoing work in our lab is investigating a second important feature of the adult work on representativeness that has been, to our knowledge, untested in children. That is, of course in many cases, it is reasonable (and highly computationally efficient) to rely on the type of individuating information that is provided in these vignettes, as the information appears to be accurate and reliable. However, adults will neglect base-rates in favor of individuating information even when that individuating information is presented as unreliable or uninformative for classification. This suggests that adults have an overwhelming expectation that social information should always be heavily relied on. Have 6-year-old children developed this expectation? Or will they be more sensitive to factors like reliability as this bias is developing? Current experiments in our lab are examining this and related questions that should provide further insight into the development of reasoning biases.

[1] Adults (N = 36) completed an experiment identical to Expt. 2 (i.e., same gender and trait conflict and no-conflict problems). Space does not permit a full report of the experiment, but adults produced the expected results (giving the normative response for no-conflict and the non-normative response for conflict problems).

Acknowledgments

We thank the families for participating and Emily Chan and Elizabeth Attisano, for their assistance with data collection.

References

Davidson, D. (1995). The representativeness heuristic and the conjunction fallacy effect in children's decision making. *Merrill Palmer Quart, 41,* 328–346.

Heyman, G.D., & Gelman, S.A. (2000). Preschool children's use of trait labels to make inductive inferences. *J Exp Child Psychol, 77*, 1 – 19.

Jacobs, J.E., & Potenza, M. (1991). The use of judgment heuristics to make social and object decisions: A developmental perspective. *Child Dev, 62*, 166–178.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Kokis, J., Macpherson, R., Toplak, M., West, R.F., & Stanovich, K.E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *J Exp Child Psychol, 83,* 26–52.

Martin, C.L., & Ruble, D. (2004). Children's search for gender cues. *Curr Dir Psychol Sci, 13*, 67 – 70.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Dev Rev, 31*, 103-118.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *P Natl Acad Sci USA, 104,* 19156–19159.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *P Natl Acad Sci USA, 105,* 5012–5015.

Combining Multiple Perspectives in Language Production: A Probabilistic Model

Mindaugas Mozuraitis
Dept. of Computational Linguistics & Phonetics, Saarland University
mindauga@coli.uni-saarland.de

Suzanne Stevenson
Dept. of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Daphna Heller
Dept. of Linguistics
University of Toronto
daphna.heller@utoronto.ca

Abstract

While speakers tailor referring expressions to the knowledge of their addressees, they do so imperfectly. Our goal here is to provide an explanation for this type of pattern by extending a probabilistic model introduced to explain perspective-taking behavior in comprehension. Using novel production data from a type of knowledge mismatch not previously investigated in production, we show that production patterns can also be explained as arising from the probabilistic combination of the speaker's and the addressee's perspectives. These results show the applicability of the multiple-perspectives approach to language production, and to different types of knowledge mismatch between conversational partners.

Keywords: language production; computational modeling; reference; audience design; common ground; perspective-taking; pragmatics; probabilistic models.

Introduction

The process whereby speakers tailor their utterances to the knowledge state of their addressee is known as "audience design". Much research on audience design has focused on reference – i.e., the labeling of objects. Reference is an ideal test bed for audience design, because of the clear action associated with referring: the speaker needs to produce a linguistic form that enables her addressee to identify the intended object. For example, a speaker should only use the name *Aloysius* if she can assume that her addressee will be able to map this name onto the intended person. More generally, in order to be understood, a speaker should rely on shared information when choosing a referring expression (Clark & Marshall, 1981).

Indeed, psycholinguistic research has shown that speakers rely on shared information when choosing a referring expression (e.g., Nadig & Sedivy, 2002). However, their referring expressions are not always based on shared information alone. For example, when producing language under time pressure, speakers do not distinguish between shared information and their own privileged knowledge (Horton & Keysar, 1996). But even when speakers do distinguish the two, their utterances are sometimes formed using privileged information. For example, when there is one shared triangle and a larger triangle known only to the speaker, speakers sometimes say *the small triangle*, even though for the addressee it would be sufficient to use the unmodified expression *the triangle* (Wardlow Lane & Ferreira, 2008; Yoon et al., 2012). Furthermore, when their addressee does not know a name, speakers sometimes include it nonetheless, along with the descriptive information that would allow the addressee to identify the referent (Isaacs & Clark, 1987; Heller et al., 2012). In sum, speakers do not fully adapt to addressees' knowledge, and production often reflects some egocentric tendencies.

The goal of this paper is to propose an explanation for these "mixed" patterns. Specifically, we extend to cover language production a computational model previously introduced to explain reference comprehension through the probabilistic combination of speakers' and addressees' perspectives (Heller et al., 2016). We first present a new production experiment, and then turn to modeling its results. While much research on reference production has considered knowledge mismatch due to differences in visual perspectives (e.g., Horton & Keysar, 1996; Nadig & Sedivy, 2002; Wardlow Lane & Ferreira, 2008; Yoon et al., 2012), our experiment involves a case where knowledge differs with respect to the *function* of objects. The predictions of our model, which combines the contribution of the speaker's perspective and the addressee's perspective, are a good fit to the human data.

Production Experiment

Our experiment investigates the production of referring expressions in cases where the speaker and the addressee have differing knowledge about the *function* of an object. To create such knowledge mismatch, we use Visually-Misleading Objects [VMOs] whose appearance does not match their function, such as a crayon that is shaped to look like a Lego block (similar objects were used in Mozuraitis et al., 2015, a comprehension study). One novel aspect of our design is that it examines the effects of knowledge mismatch about the VMOs' function *indirectly*: As shown in Figure 1(a), these objects were not described by participants (i.e., they were not the target). Instead, they were paired with a target (a typical object) that either matched their appearance (e.g., an actual Lego block) or that shared their function (e.g., a regular crayon).

What happens when both interlocutors have been demonstrated the function of the VMO (**the shared condition**)? If the VMO is categorized based on its appearance (e.g., as a Lego), then there are two Legos in the ***appearance*** condition (Fig. 1(a-1)), and a referring expression should include a modifier to distinguish the appearance-target from the contrast (e.g., *the Lego on your left*). If the VMO is categorized based on its function (e.g., as a crayon), then there are two crayons in the ***function*** condition (Fig. 1(a-3)), thus a modified referring expression should be used to distinguish the function-target from the contrast (e.g., *the longish crayon*). Importantly, these are not mutually exclusive: it is possible that both appearance and

function play into the categorization of objects, and speakers will use a modified expression in both cases.

(a) Target x Contrast conditions (b) Example critical display

	Target	Contrast
1)	Appearance	VMO Lego-crayon
2)	Appearance	Control
3)	Function	VMO Lego-crayon
4)	Function	Control

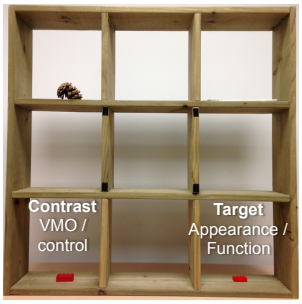


Figure 1: Sample materials

How will this pattern change (if at all) when only the speaker has been shown the function of the VMOs (the privileged condition)? As demonstrated in Mozuraitis et al., (2015), when people are *not* demonstrated the function of these objects, they categorize them based on appearance (e.g., as a Lego). Thus, if speakers (participants) tailor the referring expression to the perspective of the addressee (who can be assumed to think that the VMO is a Lego), they should use a modified expression when referring to the appearance-target (e.g., *the Lego on your left*), but not with the function-target (which they will simply call *the crayon*).

Method

Participants We report data from 80 native English speakers from the University of Toronto community, paid $10 each. An additional 7 participants were excluded because they failed to follow instructions (n = 3) or they reported a suspicion that their partner was not naïve (n = 4).

Materials and design All displays contained four objects (e.g., in Figure 1b the distractor objects are a pinecone and a wand). Three factors were manipulated in a 2 x 2 x 2 design: Knowledge state (shared vs. privileged; between-participants), type of target (appearance vs. function; within-participants), type of contrast (visually-misleading vs. control; within-participants).

Knowledge state. In the shared conditions (n=40), at the beginning of each trial, the function of the VMO (and other objects) was demonstrated non-verbally: e.g., the experimenter drew on a piece of paper with the crayon that looks like a Lego, while both the speaker (participant) and the addressee (confederate) were facing the objects. In the privileged conditions (n=40), the function of objects was demonstrated only to the speaker, while the addressee (confederate) faced away from the display (and wore headphones). In the latter condition, the goal was to give clear and consistent cues to the speaker that the addressee is unaware of the true (unexpected) function of VMOs.

Type of target. In the appearance conditions, the target shared the appearance of the VMO (e.g., a regular Lego with the Lego-crayon; see Fig. 1(a-1)). In the function conditions, the target shared the function of the VMO (e.g., a regular crayon with the Lego-crayon; see Fig. 1(a-3)).

Type of contrast. In the visually-misleading (critical) conditions, the contrast was one of 8 VMOs (appearance-function): Lego-crayon, paintbrush-eraser, baseball-yoyo, cigarette-pencil, book-box, lighter-pencil sharpener, pen-screwdriver, lightbulb-candle. In the control conditions, the VMOs were replaced by objects from an unrelated category that were similar to the VMOs in size and color (Fig. 1(a-2) and 1(a-4)). The control conditions were intended to provide baseline modification rates for the target objects when the display does not contain any contrasting objects.

A list design cycled the pairing of target and contrast objects such that participants saw a given array only once. Across participants, each array occurred in all of the experimental conditions. Across the experiment, target and contrast objects appeared in each of the four display positions equally often. To counteract any contingencies created by the critical items, we added 24 filler displays. To ensure that VMOs are not always relevant for the instruction, two fillers had a VMO paired with an object matching its appearance or function, but neither was the target. To have speakers refer to strange objects, eight fillers had a difficult-to-name object (not VMOs), with six of those as the target. To divert attention from appearance and function contrasts, ten fillers had a pair of objects contrasting in materials, with four as the target. The final four fillers had four unrelated objects. Trials were presented in random order with no two consecutive critical trials.

Procedure Participants (always the speaker) were led to believe the confederate was a naïve participant, and that they were assigned to their respective roles arbitrarily. Following Kuhlen & Brennan (2013), we used confederates blind to the purpose of the experiment. Two confederates participated in the shared condition. A third confederate participated in the privileged condition, to ensure she had not previously seen the function of the VMOs.

Partners were seated on opposite sides of a table with a 3x3 cubbyhole display (Fig. 1(b)); the middle cubbyhole was covered to avoid eye contact that might reveal referential intent. Each trial began by the experimenter demonstrating the function of objects non-verbally, and placing them in the display. Then, the speaker saw a picture on a computer monitor (not visible to the addressee), indicating the object to be moved and its target location. Speakers were given no further instructions, except to refrain from pointing or other gestures. Confederates did not act until speakers completed instructions (they were instructed to this effect during their training).

Results

The dependent variable was whether speakers' referring expressions for the target included a modifier (e.g., *the Lego on your left* or *the smaller Lego*, coded as 1) or not (e.g., *the Lego*, coded as 0). The data were analyzed using a mixed-effects logistic regression model with participants and items as crossed, independent, random effects. We used models

with the maximal random-effect structure that led to convergence. The model that converged included random intercept for participants and random intercept for items.

The 2x2x2 model (summarized in Table 1) revealed a main effect of contrast type: as expected, speakers were overall more likely to use modifiers in referring to the target when it occurred with a VMO than with a control object (.82 vs. .16). There were 3 significant interactions: Target type x Contrast type, Target type x Knowledge state, and the 3-way interaction. Here we focus on unpacking the 3-way interaction, both for theoretical reasons, and because the 2-way interactions are subsumed by it.

Table 1: The 2x2x2 model. Significant effects are bolded.

Effect	*β*	*SE*	*z*	*p*
Knowledge	0.29	0.32	0.90	0.367
Target	-0.32	0.26	-1.24	0.215
Contrast	**4.06**	**0.33**	**12.46**	**<0.001**
Knowledge x Target	**2.20**	**0.53**	**4.18**	**<0.001**
Knowledge x Contrast	-0.93	0.53	-1.74	0.081
Target x Contrast	**1.40**	**0.52**	**2.69**	**0.007**
Knowledge x Target x Contrast	**3.50**	**1.05**	**3.33**	**<0.001**

Separate follow-up analyses were conducted for the two types of contrasts – see Figure 2. When the contrast was a **control** object, there was a main effect of target type (β=-1.07, *SE*=0.37, *z*=2.88, *p*=0.004), indicating that participants were overall more likely to use a modified expression to refer to the function-target rather than the appearance-target (.22 vs. .11); recall these are simply different objects and thus the pattern is not meaningful beyond providing a baseline. The main effect of knowledge state was marginal (*p*=0.074): there was a trend for speakers modifying more in the privileged conditions. However, the Target type X Knowledge state interaction was not significant (*p*=0.581), indicating that, in the absence of a contrasting object, the knowledge state manipulation did not change the modification pattern differentially for the function and appearance targets.

When the contrast was a **VMO**, the Target type X Knowledge state *was* significant (β=4.08, *SE*=0.83, z=4.90, *p*<0.001; main effects were not, *p*s>0.250). Pairwise comparisons revealed that **when the target matched the VMO in function**, participants were *less* likely to use modified expressions to describe it in the privileged condition (.65 vs. .92; β= -1.95, *SE*=0.49, *z*=-3.95, *p*<0.001). This pattern indicates that speakers adapted to the addressee's perspective: they were less likely to use a modifier to differentiate the target (e.g., a typical crayon) from the VMO (e.g., the Lego-crayon) when they had no reason to assume that the addressee knew these shared their function. When **the target matched the VMO in appearance**, the pattern was reversed: participants were *more* likely to use modifiers in the privileged condition (.94 vs. .75; β=2.33, *SE*=0.92, *z*=2.53, *p*=0.012). This pattern further demonstrates that speakers adapted to the addressee's perspective: they were more likely to use a modifier to differentiate the target (e.g., a regular Lego) from the VMO (e.g., the Lego-crayon), when they knew the addressee did not know the real function of the VMO, and thus would expect the VMO to have a function consistent with its appearance.

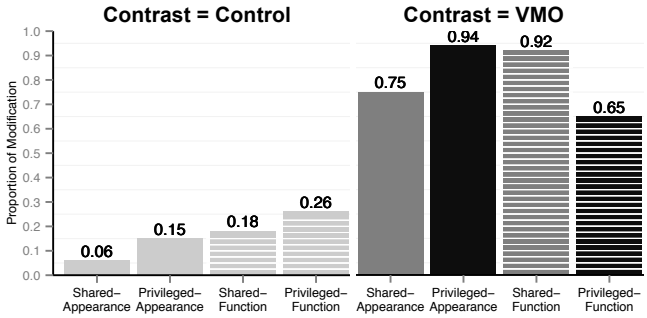


Figure 2: Modification rates produced.

Taken together, the crossover interaction indicates that speakers were sensitive to the knowledge state of their addressee in tailoring referring expressions. But note that if speakers in the privileged condition completely adapted to the addressee's perspective, they should not modify at all in the VMO-privileged-function condition (rightmost column in Figure 2): this is because the function-target (a crayon) would not need to be distinguished from the VMO (the Lego-crayon that the addressee did not know was a crayon). However, speakers used more modifiers in this case than in the corresponding control condition (.65 vs. .26; β=1.84, *SE*=0.41, *z*=4.52, *p*<0.001). This difference reveals that speakers did not completely adapt to the addressee's perspective. Instead, this pattern reflects the consideration of the addressee's perspective along with their own.

The Probabilistic Model

The probabilistic model of Heller et al. (2016) was developed to account for apparently-inconsistent results in the literature on perspective-taking in comprehension. The approach models reference resolution (i.e., the comprehension of a referring expression) as the probability P(*obj*|*RE*) that a certain object *obj* is the referent intended by the speaker, given a referring expression *RE*. Following standard practice, the probability on the left is rewritten using Bayes rule, as a product of two probabilities; importantly, the component probabilities are then conditioned on the *domain of reference* (*d*):

$$\mathrm{P}(obj|RE) =^{\mathrm{def}} \alpha \mathrm{P}(RE|obj,d=e)\mathrm{P}(obj|d=e) + (1-\alpha)\mathrm{P}(RE|obj,d=c)\mathrm{P}(obj|d=c) \quad (1)$$

An important aspect of this model is the conditioning on the domain *d*. This conditioning captures the fact that reference depends not just on what the intended referent is, but also on what other objects need to be distinguished from the referent. The likelihood P(*RE*|*obj*,*d*) captures the preference for using the referring expression *RE* to describe the various objects *obj* in domain *d*; the prior P(*obj*|*d*) captures the probability of *obj* being referred to in the context of domain *d*.

A second important aspect of the model is that reference resolution is guided by *both* the common ground perspective *c* (those entities which both the speaker and the addressee

can see; which was identical to the speaker's perspective), and the egocentric perspective *e* (all the objects the addressee can see). This contrasts with other Bayesian models of reference, which have not considered the combined influence of the speaker's and addressee's perspectives (e.g., Frank & Goodman, 2012; Goodman and Stuhlmüller, 2013; Kehler & Rohde, 2013). These two domains are weighed in formula (1) by α and (1-α), respectively, to ensure the combination of the component probabilities forms a probability distribution.

In order to model **language production**, we extend Heller et al.'s (2016) model in two ways. First, since we are modeling the speaker's choice of referring expression rather than the addressee's search for a referent, the probability of interest is P(*RE*|*obj*) rather than P(*obj*|*RE*) – that is, we directly model the preference for various referring expressions *RE* assuming that the object *obj* to be referred to is a given. In this case, there is no need to apply Bayes rule. Second, we observe that the use of the domain-weighting constant α in the Heller et al. formulation can be avoided: we can rewrite P(*RE*|*obj*) by marginalizing over all possible values of the domain variable *d* (where *D* is the set of possible referential domains for production):

$$P(RE|obj) = \sum_{d \in D} P(RE|obj,d)P(d) \quad (2)$$

Marginalizing over *d* provides a well-motivated way within probability theory to condition the probability of the *RE* on the domain *d* when its value is not a given. By capturing the degree of influence of each domain *d* in the probability P(*d*), the formula in (2) more naturally encodes the kind of domain weighting that Heller et al. achieved with the ad hoc constant α in formula (1). This approach also provides a more general formulation that can be readily extended to the case of more than two domains. It is important to note that formula (1) of Heller et al. can be recast within this same marginalization approach as:

$$P(obj|RE) \propto \sum_{d \in D} P(RE|obj,d)P(obj|d)P(d) \quad (3)$$

where $D=\{e, c\}$, and P(*d*=*e*) replaces α and P(*d*=*c*) replaces (1-α). Thus, together the two probability formulas we propose, (2) and (3), provide a unified way of modeling both production and comprehension of referring expressions under the influence of multiple perspectives.

Instantiating the Probability Model

The Variables in the Probability Formula

The Object *obj* The given object *obj* in our probability formula of Eqn. (2) will denote the *referent*: the object the speaker will label with the goal of the addressee choosing it.

The Referring Expression *RE* Because the dependent variable we used in the experimental results was modification rate, we consider *RE* for the purposes of modeling to represent modified referring expressions. That is, instead of determining the preference for a *specific* referring expression, we use the probability formula to calculate the probability of using modification, P(*RE*=*MOD*|*obj*=*target*), abbreviated as P(*MOD*|*target*).

The Domains *D* We consider the set *D* of multiple domains that influence the speaker's formulation of referring expressions to consist of two domains: the domain *s* is the speaker's (egocentric) perspective, and the domain *a* is the addressee's perspective. Thus, $D=\{s, a\}$. In general, *d*=*s* corresponds to the objects and their properties known by the speaker, and *d*=*a* corresponds to the objects and properties known by the addressee. Importantly, when the addressee did not know the true function of a VMO (privileged condition), we assume the domain *a* reflects the speaker's assumption about the addressee's false belief. (Note that Heller et al. (2016) used different labels for the perspectives of the interlocutors in modeling comprehension: domain *e* for the egocentric domain of the addressee and domain *c* for the common ground, which was identical to the speaker's perspective.)

The Resulting Probability Formula With the above variable values, we can instantiate formula (2) as:

$$P(MOD|target) = \sum_{d \in \{s, a\}} P(MOD|target, d)P(d) \quad (4)$$

Showing the sum over the two possible domains yields the following formula to model the experimental data:

$$P(MOD|target) = P(MOD|target, d=s)P(d=s) + P(MOD|target, d=a)P(d=a) \quad (5)$$

Estimating the Probabilities

P(*MOD*|*target*, *d*) Because our goal is to use the multiple domains approach to explain data patterns obtained under knowledge mismatch (i.e., those in the critical Privileged conditions) we derive probabilities from the Control and Shared conditions, and see whether their application in the Privileged conditions obtains the observed data pattern.

Specifically, we take each probability P(*MOD*|*target*,*d*) in one of the critical experimental conditions to reflect two influences. First, a "baseline" level of modification holds for different objects regardless of the presence of a VMO; this baseline is taken directly from the modification rates in the four Control conditions (we include these because of the variation). Second, when the target needs to be distinguished from a contrasting VMO, as in the Shared conditions, there is an additional level of modification. We thus take the Shared experimental conditions to reflect a "typical" modification rate for when the target shares only appearance or only function with the VMO (Shared-Appearance and Shared-Function, respectively).

Next, we use these values to predict the behavior observed in the two VMO-Privileged conditions. We set the value of the probability P(*MOD*|*target*,*d*) in each of the Privileged conditions to be the sum of:

(i) the baseline level of modification obtained in the corresponding Control condition, plus:

(ii) the amount of modification due to the target sharing either the Appearance or the Function of the contrast object, from the corresponding Shared condition.

These component estimates are shown in Fig. 3a. The resulting values for the Privileged conditions are shown below in Modeling the Production Data, where we consider the speaker's and the addressee's perspectives (other panels of Fig. 3, described below).

P(*d*) As in Heller et al. (2016), we consider that the language user's weighting of each domain is not directly observable. We determine the range of values of P(*d*) that yields a fit to the empirical data, and see if this range fits our hypothesis that speakers use both domains in formulating referring expressions. Because we assume that *d* can only take on the values *s* and *a*, those values exhaust the probability space, and so P(*d*=*s*)+P(*d*=*a*)=1, or P(*d*=*a*)=1–P(*d*=*b*). Given that there is only one parameter to consider here (the other value is the additive inverse), the model needs to account for *two* patterns of modification (i.e., in Privileged-Appearance and in Privileged-Function) with *one* parameter setting for weighing perspectives. Thus, our experimental design provides a critical test of the model.

Modeling the Production Data

To model the data, we must consider what the speaker's and addressee's perspectives are regarding the objects and their relevant properties, with attention to the relation of the VMO to the target. This relation dictates the component probabilities that must be added to the baseline preference for modification. Specifically, does the target share Appearance with the VMO, share Function with the VMO, or share neither? The answer depends on the perspective.

Speaker's Perspective (P(*MOD*|*target*,*d*=*s*): Fig. 3b) Because the speaker always knows the true function of the VMOs, the speaker's perspective is identical in the Shared and Privileged conditions. We calculate P(*MOD*|*target*, *d*=*s*) in the Appearance conditions by adding the shared-appearance probability (.69) to each corresponding baseline, and in the Function conditions by adding the shared-function probability (.74) to each corresponding baseline.

Addressee's Perspective (P(*MOD*|*target*,*d*=*a*): Fig. 3c) In contrast to speakers, the addressee's perspective changes across the knowledge state manipulation. In Privileged-Appearance, the similarity in appearance between the target and the VMO (e.g., a Lego and a Lego-crayon) should lead the addressee to assume that the objects *also share function*. Thus, we add the **shared-function** probability (.74) to the baseline to mimic the addressee's assumed perspective (despite the fact that in actuality the objects share only their appearance). In Privileged-Function (e.g., a crayon target and Lego-crayon contrast), the addressee's false belief should lead to an assumption that the VMO shares neither appearance nor function with the target, and thus a **value of 0** is added to the baseline.

Combining the Two Perspectives (P(*MOD*|*target*): Fig. 3d) Our proposal is that the speaker probabilistically combines both their own perspective with the addressee's perspective in considering the preference P(*MOD*|*target*) for a referring expression. Assuming that the probabilities in Fig. 3b and 3c are weighted equally in Eqn. 5 (P(*d*=*s*)=P(*d*=*a*)=0.5), we obtain the values for P(*MOD*|*target*) across the four experimental conditions illustrated in Fig. 3d. The levels of probability for the Privileged conditions (which are the values we aim to predict) have a very good fit to the behavioral data (Fig. 2; recall that the values for the Shared conditions are set according to the human data.) With 0.25 < P(*d*=*a*) < 0.65, the model yields values in the ranges consistent with the 95% confidence intervals for the Privileged conditions; the best fit is obtained with P(*d*=*a*)=0.47, which is very close to the equal combination in Fig. 3d. Importantly, because we are predicting two values (Privileged-Appearance and Privileged-Function) with one parameter (P(*d*=*a*)), this is not trivial (if it were one data point, there would always be *some* value for P(*d*=*a*) that would achieve the fit).

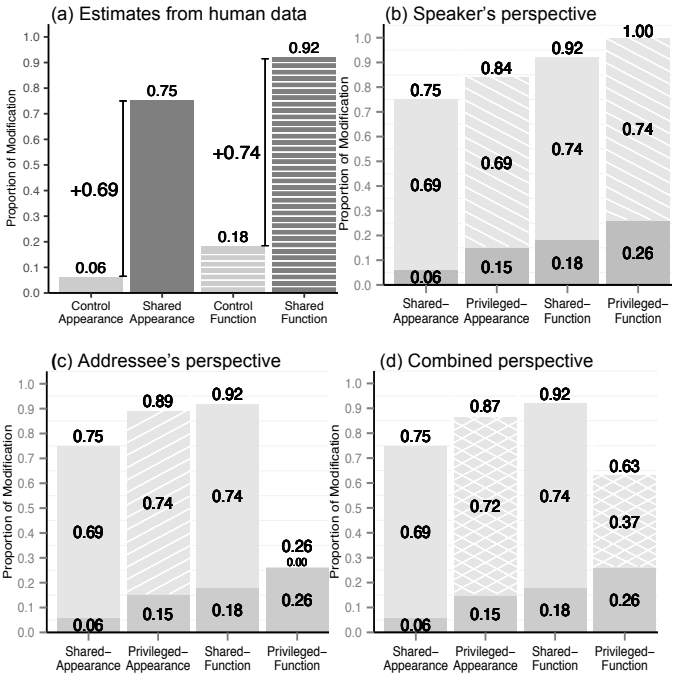


Figure 3: Modeling

Discussion

We present the first model of reference production that probabilistically combines the influence of multiple perspectives. Our model explains the modification patterns observed in our production experiment as a result of speakers considering both their own perspective and the addressee's perspective in choosing the form of a referring expression. The modeling results support the hypothesis that speakers use *both* perspectives, because weighing the two perspectives about equally gives a good match to the human data. A model where speakers are fully egocentric (equivalent to P(*d*=*a*) being close to 0) or fully adapt to the addressee's perspective (equivalent to P(*d*=*a*) being close to 1) cannot predict the pattern observed. Furthermore, this result is attractive because we are using the same approach that successfully modeled perspective-taking in comprehension (Heller et al., 2016).

Why would interlocutors use both their own perspective along with their partner's? One reason is that it may be cognitively taxing to completely suppress one's own perspective. It has already been shown in the literature that perspective-taking abilities are tied to inhibition control (Brown-Schmidt, 2009; Nilsen & Graham, 2009), and to the salience of one's own perspective in the situational context (Wardlow Lane & Ferreira, 2008).

There are also *linguistic* reasons, however, for why interlocutors might use both perspectives, namely performing felicitous moves in conversation. For example, producing a meaningful assertion requires choosing information from one's own perspective and assessing that one's partner does not know the information; formulating a question, in contrast, involves identifying a knowledge gap in one's perspective and assessing that the partner *does* have this information.

Finally, even when tailoring referring expressions, it may not be ideal for speakers to focus on the addressee's perspective alone: what we have been calling "the addressee's perspective" is just a *hypothesis* on the part of the speaker about the addressee's knowledge. While this hypothesis may have been developed based on strong cues (e.g., the fact that the addressee did not see the experimenter demonstrate the function of VMOs), other cues in the situation may suggest to the speaker that the addressee actually shares their knowledge (e.g., they may note small perceptual cues on the VMOs that are suggestive of their unexpected function). Given the uncertainty in assessing the addressee's knowledge state, the speaker may do well to consider their own perspective as relevant to expressing their intent.

Note that this consideration and integration of multiple communicative contexts is different from that explored in Goodman and Stuhlmüller (2013). There, probabilistic weighing is used to model *uncertainty about the other partner's state of knowledge*, with the goal of maximizing adaptation to the partner. This fundamentally differs from our approach of weighing and integrating the differing perspectives *of the two partners*. In future work, insights from the Goodman and Stuhlmüller model of uncertainty with respect to a single perspective can be integrated within our approach that combines multiple perspectives.

Thus, although on the surface it may seem ideal to choose a referring expression that is fully tailored to the addressee's perspective (cf. Clark & Marshall, 1981), our view is that balancing the different demands of conversation requires actively maintaining – and integrating – a representation of both partners' perspectives. In other words, perspective-taking behavior is achieved neither by focusing on shared information nor by trying to fully adapt to one's partner, but rather by probabilistically integrating the perspectives of all interlocutors. Future work will aim to disentangle the considerations that lead to the weighing of perspectives, as well as the effect of feedback on weighing and re-weighing.

Acknowledgments

We are grateful to Rafiya Asad for her help with data collection and coding, and we acknowledge support from NSERC and SSHRC of Canada.

References

Brown-Schmidt, S. (2009). The role of executive function in perspective-taking during on-line language comprehension. *Psycho. Bulletin & Review 16*, 893-900.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & I. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63).

Frank, M. C. & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science 336*, 998.

Goodman, D. N., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science, 5*, 173–184.

Heller, D., Gorman, K. S. & Tanenhaus, M. K. (2012). "To name or to describe: shared knowledge affects referential form". *Topics in Cognitive Science*, *4*, 290-305.

Heller, D., Parisien, C. & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition, 149,* 104–120.

Horton, W. S. & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91-117.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*, 26–37.

Kehler, A. & Rohde, H. (2013). A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation, *Theoretical Linguistics*, *39*, 1-37.

Kuhlen, A. K. & Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, 20, 54-72.

Mozuraitis, M., Chambers, G. C., & Daneman, M. (2015). Privileged vs. shared knowledge about object identity in real-time referential processing. *Cognition*, *142*, 148-165.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13,* 329–336.

Nilsen, E., & Graham, S. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cognitive Psychology 58*, 220-249.

Wardlow Lane, L. & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *6*, 1466-1481.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*, 699–707.

Statistical Learning Ability Can Overcome the Negative Impact of Low Socioeconomic Status on Language Development

Leyla Eghbalzad (leghbalzad1@student.gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302

Joanne A. Deocampo (jdeocampo@gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302

Christopher M. Conway (cconway@gsu.edu)
Department of Psychology, Georgia State University, P.O. Box 5010, Atlanta, GA 30302

Abstract

Statistical learning (SL) is believed to be a mechanism that enables successful language acquisition. Language acquisition in turn is heavily influenced by environmental factors such as socioeconomic status (SES). However, it is unknown to what extent SL abilities interact with SES in affecting language outcomes. To examine this potential interaction, we measured event-related potentials (ERPs) in 38 children aged 7-12 while performing a visual SL task consisting of a sequence of stimuli that contained covert statistical probabilities that predicted a target stimulus. Hierarchical regression results indicated that SL ability moderated the relationship between SES (average of both caregiver's education level) and language scores (grammar, and marginally with receptive vocabulary). For children with high SL ability, SES had a weaker effect on language compared to children with low SL ability, suggesting that having good SL abilities could help ameliorate the disadvantages associated with being raised in a family with lower SES.

Keywords: statistical learning; language development; socioeconomic status, event-related potentials (ERP); cognitive development

Introduction

All typically developing children learn how to comprehend and produce language, suggesting the existence of common biological and/or environmental mechanisms for language development. On the one hand, language acquisition may depend on intrinsic factors such as mental and genetic components (e.g., Chomsky, 1965; Crain & Lillo-Martin, 1999). On the other hand, language development may rely less on internal factors and more on external interactions with social and environmental contexts (Bronfenbrenner, 1979, 1988). The universality and variability in language development suggest that a combination of these perspectives might provide the most appropriate approach for studying language development (Hoff, 2006). In particular, it may be beneficial to study language development in children by focusing on the interaction between intrinsic (e.g., cognitive skills) and extrinsic factors (e.g., social/linguistic environment).

Social environmental factors appear to be essential for the development of language (Kuhl, 2010). For instance, according to the "social gating" hypothesis (Kuhl, 2007), social interactions influence learning in children by increasing their attention span and, therefore, the amount of knowledge retained from the environment. Social environmental factors such as socioeconomic status (SES) also impact language learning (Feldman et al., 2003; NICHHD, 2000; Hoff et al., 2012). SES consists of many components and each component could potentially influence various aspects of child development differently. Primary Caregivers' education level and income are among the most important indicators of SES (Roberts et al., 1999). Children who live in low SES families are reported to have less exposure to linguistic stimulation, which could detrimentally impact their language development (Rowe and Goldin-Meadow, 2009; Sheridan et al., 2012). In addition, there is increasing evidence suggesting that environmental factors can impact brain regions that are associated with executive functions and language. For instance, the prefrontal cortex in children seems to be strongly impacted by their SES (Sheridan et al., 2012). Consequently, children with low SES do not perform as well as children with higher SES on tasks that represent cognitive control, memory, and language (Farah et al., 2006).

Certain biological factors and cognitive mechanisms also play an important role in language development in children. Specifically, statistical learning (SL) abilities appear to be essential for detecting and encoding structured patterns of information in the environment, including language (Conway et al., 2010; Saffran, 2003). Indeed, research suggests that SL is a crucial component of language processing in newborns (Saffran, Aslin & Newport, 1996; Shafto, Conway, Field & Houston, 2012), children (Kidd & Arciuli, 2015; Lum et al., 2012), and adults (Christiansen, Conway, & Onnis, 2012; Misyak, Christiansen, & Tomblin 2010). Shafto, Conway, Field, and Houston (2012) demonstrated an empirical link between visual SL in infants and their subsequent vocabulary development. SL appears to be used to learn the underlying patterns inherent in linguistic signals, which facilitates the prediction of upcoming units (Conway et al. (2010). In an electroencephalography study, Christiansen et al. (2012) reported that the same neural mechanisms appear to be utilized for processing syntactic rules of language and SL. Overall, these studies suggest that SL may be a prerequisite

for language learning and that variations in SL can affect language development.

In sum, language development appears to be highly dependent on both the environment that the child is raised in and his or her cognitive skills. What is not known is the extent to which these two factors might interact to impact language development. For instance, it is possible that having better intrinsic learning abilities could help offset the deleterious effects of being raised in an impoverished social environment.

The specific aim of this study is to examine the relationship between the neural mechanisms of SL, social environmental factors, and language in typically developing children. Specifically, we explored the possible impact of SL as a moderator of the relationship between SES and language development in children. We measured SL by using the event-related potential (ERP) technique while children were performing a computerized visual SL task. We also measured children's vocabulary and grammatical judgment using standardized language assessments.

Method

Participants

We recruited 42 typically developing children aged 7-12 from the Atlanta metropolitan area with English as their native and only language (age mean = 9 years; 25 male). Four participants were excluded from this study, one due to computer software difficulties during ERP data acquisition, and three due to having too many noisy trials in the ERP task (see EEG section below). The final analyses were done using data from 38 participants (Age mean = 9 years; 23 male, 15 female; 13 black or African American; 18 white; 7 more than one race). Participants and their parent/caregiver came to the Psychology Department at Georgia State University for their laboratory visit. During their visit, both parents and children were informed about the goal and details of the study and provided written informed consent and assent to participate. Participants were offered a toy, worth $10 for participating. Additionally, the parents received monetary compensation of $50 for the session.

Socioeconomic Status (SES)

Parents of the participants completed a questionnaire regarding their socioeconomic status (SES) and demographics. This questionnaire consisted of questions about their individual and household income, education, and demographics of the PC and secondary caregiver (SC). In the analysis, we used average of both PC and SC education level as a measure of SES. Each caregiver's education level was measured using the following scale: 0= Less than High School, *N*=5; 1= High School, *N*=14; 2= Some college, *N*=8; 3= Associate's degree, *N*=3; 4= Bachelor's degree, *N*= 15 ; 5= Master's degree, *N*=16; 6= PhD, *N*=5 ; 7= Professional degree, *N*=10. Household income was not used in the analyses due to missing data from more than half of the participants.

Statistical Learning Task

The visual SL task was based on a task recently developed by Jost et al. (2015), which in turn is similar to the classic visual oddball paradigm, but with statistical regularities embedded in the stimuli. We made the Jost et al. (2015) task more child-friendly by making it into a game with a background story ("the Magician task"). This task was presented as a game on a laptop computer. In this task, children were told a story about an inconsistent magician who tries to make food for his children using his magic hat. Children viewed a stream of flashing stimuli consisting of hats of different colors presented with a black background one at a time. Occasionally, a target hat with food was presented within the stream. Children were instructed to "catch" the presented food by pressing a button. Participants were not told that hats of different colors each differentially predicted the probability of occurrence of the target hat. Each target followed a predictor in the sequence with three conditions: high (90% probability of target following), low (20% probability of target following), and no predictor (target presented with no preceding predictor). Each experimental condition (high, low, and no) contained 60 trials, for 180 trials total. Each stimulus was presented on the screen for 500 milliseconds and was followed by a black screen for 500 milliseconds. Six blocks were separated by 30-second breaks during which children watched a short cartoon related to the magician story. Figure 1 shows a schematic presentation of the magician task. It took the participants about 20 minutes to complete the task after net application. If children learned the probabilistic patterns between each type of predictor and the target, it was expected that there would be significant differences in their response times (RTs) to the targets and/or the amplitude differences of ERPs of the predictors based on whether a trial was a high-probability (HP), low-probability (LP), or no-predictor (NP) type. Either of these differences would be evidence of SL[1].

Electroencephalography (EEG) Recording

We collected EEG data measuring changes in electrical potential on the scalp during the statistical learning task using a 32-channel high-density EGI (Electrical Geodesics, Inc.) sensor net and followed standard net application techniques for the EGI system. EEG data were collected in a sound-attenuated room. We used the NetStation 4.3.1 acquisition software (Electrical Geodesics, Inc.) to transform and record the data to digital form. Before starting the SL task, participants were instructed to sit still and avoid excessive blinking. Data were acquired with a 0.1 to 30 Hz bandpass filter and digitized at 250 Hz.

[1] The no-predictor condition in the "Magician" task is the same as the standard stimuli, which means participants saw it more frequently than the high- and low- probability conditions; therefore, ERP responses to the no-predictor condition may be influenced by this difference in frequency of occurrence.

Impedances were kept below 50 kΩ. ERP recordings were time-locked to the onset of each predictor stimulus and continued for 1500ms after onset for a total segment length of 1700ms. In the no-predictor condition, the ERPs were time-locked to the standard preceding the target stimulus. After data acquisition, data from channels with poor or noisy signals was replaced with data induced from surrounding sensors using the computational MATLAB software (version R2012b 8.0.0783; MathWorks). Additionally, trials containing muscle activity such as eye movement and blinks were removed using an artifact detection process.

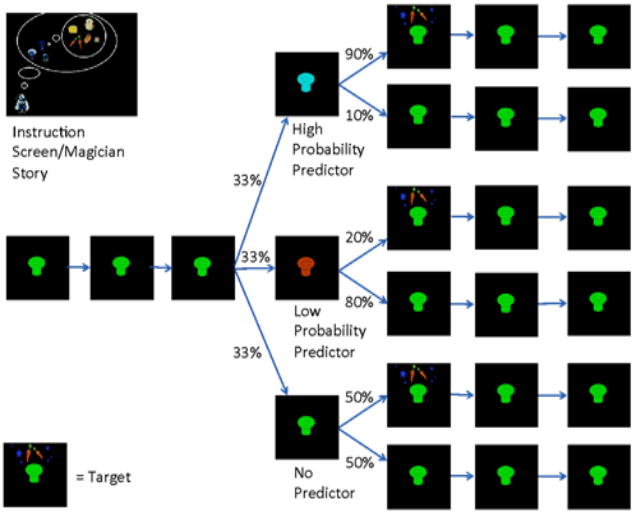


Figure 1: A schematic representation of the statistical learning "Magician" Task.

Language Assessments

The Grammaticality Judgment subtest of the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 1999) was administered as an assessment of syntactic language development. In this test, a sentence with or without grammatical errors was read to the child, and the child was asked whether it sounded correct and if not to fix it by changing only one word. This assessment was administered in a separate room with a trained experimenter after removal of the EEG sensor net. The standardized scores of this test were computed based on participants' age.

Children's receptive language was measured using the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007). During this test, an experimenter showed the participants 4 pictures and asked them to point to the picture that best represented the presented word. This assessment was administered by a trained experimenter in a quiet room before the EEG net application.

Results

The descriptive statistics of the language tests are reported in Table 1. We used the standard scores for these measures which take age into account. The average score for both the grammaticality judgement subtest of the CASL and PPVT is 100 with a standard deviation of 15 points. The participants' average scores were slightly higher than average on the grammaticality judgement and the PPVT tasks (yet still within normal limits); but there was a wide range of scores for both language tests.

Table 1: Means and standard deviations for language tasks

Assessment	*Mean*	*SD*	*Range*
PPVT Standard Score	*111.58*	*19.20*	*73-146*
Grammaticality Judgment Standard Score	*105.53*	*13.42*	*69-132*

N=38

Statistical Learning Measure

Based on Jost et al. (2015), who observed a P300-like ERP component in the posterior region of the scalp in conjunction with SL in the 400-700 milliseconds window following the predictor onset, we focused our analyses on a pre-defined region of 6 electrodes in the posterior region for the same time window (see Figure 2). To assess the behavioral and neural correlates of learning during the SL task, we ran 2 one-way ANOVAs to determine whether the 3 probability conditions (high, low, and no) in ERP amplitudes and reaction times (RT) were significantly different from one another.

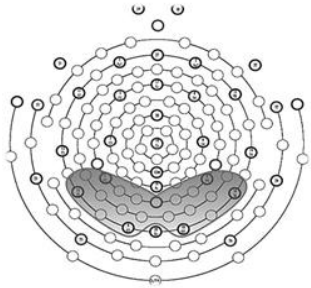


Figure 2: The map of 32-sensor EEG net with posterior region of interest highlighted.

ERP Amplitudes Figure 3 displays the grand average ERP waveforms in the posterior region. Visual inspection suggests that there may be a late positivity roughly 400-700 msec for the high and low predictor conditions. This was confirmed with a one-way ANOVA comparing ERP amplitudes for the 3 probability conditions in the 400-700 ms time-window after predictor onset, which revealed a significant effect of probability condition, $F(2, 74) = 16.60$, $p < .000$. Paired-sample t tests with Sidak adjustment revealed that the ERP wave amplitude was significantly higher for the high-probability condition ($M = 2.42$, $SD = 2.52$) compared to the low-probability condition ($M = 1.59$, $SD = 2.39$), $t(37) = 2.41$, $p < .05$, and no-predictor condition ($M = 0.28$, $SD = 1.94$), $t(37) = 5.19$, $p < .001$. The ERP wave amplitude was also significantly higher for

the low-probability condition compared to the no-predictor condition, t (37) = 3.60, p < .001. These results provide neurophysiological evidence that children demonstrated sensitivity to the different probability conditions, measured by the EEG data, which is consistent with the findings of Jost et al. (2015). These results suggest that as a group, children's learning of the predictor-target statistical patterns was reflected by a larger amplitude for the high predictor stimuli, and to a lesser extent, for the low predictor stimuli.

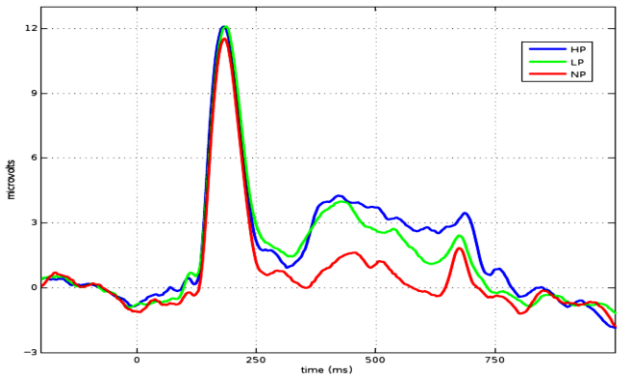


Figure 3: ERP waveform in the posterior region showing 3 different probability conditions. High- probability line is on the top (blue), low-probability line is in the middle (green), and no-predictor line (baseline) is at the bottom (red).

Reaction Times Similarly, the behavioral analyses provide evidence of statistical learning. The results of the second one-way ANOVA comparing RT in each predictor condition showed that participants responded significantly differently to the 3 predictor conditions, F(2, 70) = 31.04, p < .001, sphericity assumed. Paired-sample *t* tests revealed that the RT was significantly lower to the target following the high-probability stimuli (M = 388.97, SD = 78.84) compared to when the target followed the low-probability stimuli (M = 465.08, SD =65.89,), *t* (35) = - 4.96, p < .001, and the no-predictor stimuli (M = 493.20, SD = 67.59), *t* (35) = - 6.21, p < .001. The RTs were also significantly lower for the target when it followed the low-probability condition compared to the no-predictor condition, *t* (35) = - 4.18, p < .001.

Furthermore, difference scores were created for both the ERPs and the RT data between the high and no predictor conditions to explore the magnitude of the learning effects. Because the target was preceded by a standard in the no-predictor condition, we used it as the measure of baseline for both ERPs and RTs as it essentially constituted a measure of ERP or RT that would occur without any predictor. Thus, we defined SL by the difference between baseline and high-predictor conditions. This created one variable for ERP amplitude difference, high probability – no-predictor (H-N), and one variable for RT difference, no-predictor – high probability (N-H). Note that for the RTs, the difference scores were calculated to be positive since the high predictor condition elicited lower RTs than the no predictor condition.

Correlations

The relationship between ERP amplitude difference scores, RT difference scores, and language assessments were examined using Pearson's correlation analyses. The H-N ERP variable (M = 2.11, SD = 2.59) was significantly positively correlated with N-H RT (M = 104.23, SD = 100.71), r = .54, p = .001. Consistent with past research, SES (M = 3.67, SD = 2.06) was positively correlated with scores on PPVT (M = 111.58, SD = 19.20), r = .63, p < .001, and Grammaticality Judgement test (M = 105.53, SD = 13.42), r = .59, p < .001. However, SES was not correlated with any of the SL measures (*r*'s < 0.3, *p*'s > .07). Surprisingly, we did not find significant correlations between SL and either language measure (*r*'s < 0.3, *p*'s > .07). Partial correlation analyses with age as the controlled variable, did not result in any significant changes in these correlations.

Moderation Analyses

Hierarchical multiple regression analyses were conducted to examine if SES as the independent variable predicts language outcome in children and whether SL ability modifies this relationship. The predicting variables were standardized (converted to z scores) prior to data analyses. The results reveal that SL ability (H-N ERP variable) moderated the relationship between the caregivers' education level and grammar, R^2_{adj} = .50, $F(3, 34)$ = 13.13, p < .001, and marginally moderated the relationship between the caregivers' education level and receptive vocabulary in children, R^2_{adj} = .45, F (3, 34) = 11.18, p = .058 (see Figure 4). In addition, the SL reaction time for N-H condition was a significant moderator of the relationship between SES and grammar scores, R^2_{adj} = .367, $F(2, 33)$ = 7.75, p < .001, but not with receptive vocabulary, R^2_{adj} = .373, F (2, 33) = 8.33, p = ns. We controlled for potential covariates in each model by including 3 widely-used cognitive measures: Stroop (inhibitory control and attention), Block Design (visuo-spatial memory), and Digit Span (short-term memory) in each regression analysis separately. None of these measures were significant predictors of variance in grammar scores in addition to SL. However, Block Design and Digit span tasks were significant predictors of variance in receptive vocabulary scores, β = 7.42, p < .05 and β = 7.47, p < .05, respectively. In addition, we controlled for age as a potential covariate in all analyses which did not result in any significant explanation of variance in either language measure. The overall effect that can be seen in Figure 4 is that SES has a larger effect on language scores for the children with lower SL ability; SES has a much weaker influence on children with higher SL. This figure also shows that there is a good distribution of caregivers' average education level in our sample.

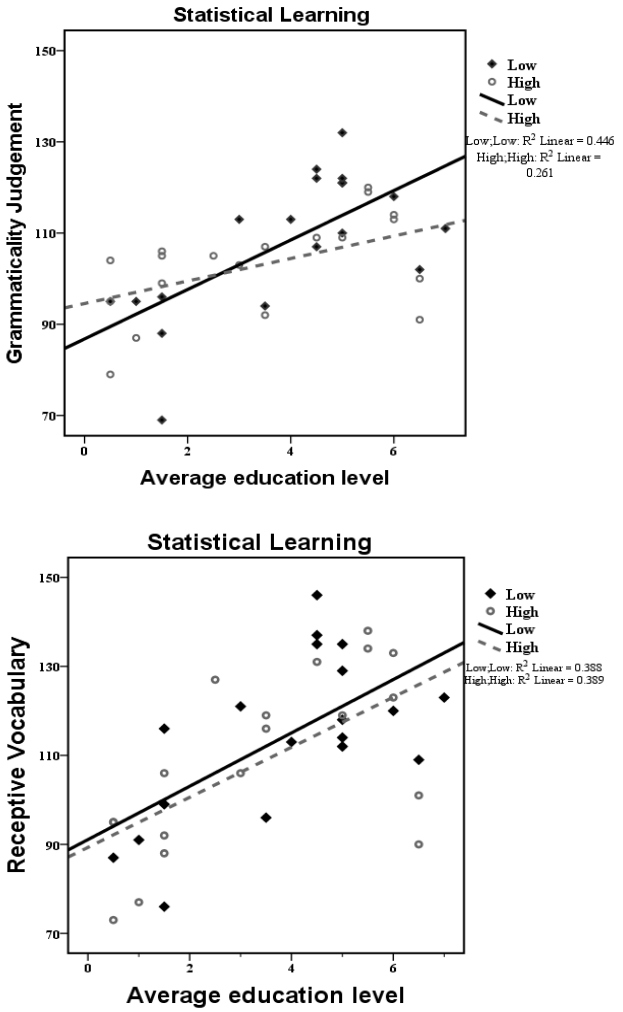


Figure 4: Scatter plots of the interactions between language scores and SES for low and high SL. For the purpose of illustration, SL variable was separated into low SL and high SL by a median split of the data.

Discussion

In this study, we investigated whether SL moderates the known relationship between SES and language outcome. In the SL task, children demonstrated different sensitivity to the different probability conditions, indicating learning of the statistical probabilities. The reaction time results were consistent with the ERP results in demonstrating evidence of SL. Consistent with previous findings, there was a positive relationship between children's SES level and their language ability (Feldman et al., 2003; Hoff & Tian, 2005; NICHHD, 2000). These results replicate previous studies demonstrating a relationship between primary caregiver's education level and language development in children (Stanton-Chapman et al., 2002). Children with more highly educated mothers demonstrated better language skills in both receptive vocabulary and grammar measures compared to those children whose mothers are not highly educated.

More importantly, the results of the moderation analysis revealed that children with high SL appeared to have more robust language ability that was less affected by their SES. In other words, the negative effect of low SES on language appeared to be dampened by high SL ability. On the other hand, for children with lower SL ability, their language scores were much more sensitive to the effects of SES. Thus, children who were raised in less advantaged families showed more typical language development if they had good SL skills whereas if they had low SL their language scores were lower. These results are the first to suggest that intrinsic cognitive abilities, specifically SL, may play a moderating role in the relationship between SES and language skills in children. The negative effect of low SES on language is more apparent when a child's SL ability is low compared to when SL ability is high.

Results also showed that SL had a stronger moderating effect for grammar compared to receptive vocabulary scores. This distinction could possibly be explained by the declarative/procedural model of language, which posits that procedural learning and grammar share a common neurological substrate (Ullman, 2004). Thus, it makes sense that SL ability has a greater moderating effect on grammatical ability compared to vocabulary.

It is important to mention that sample size of 38 participants may be relatively small for the type of analyses used in this study, and so future research with a larger sample size is needed to confirm the observed interaction between SES, SL, and language. In addition, although up to this point we have considered SL an intrinsic or biological factor and SES an environmental one, it is also possible that children's SL ability may have been shaped by the environment they are raised in while differences in SES level could be due to biological or genetic factors.

In sum, this research provides an important examination of the relationship between learning abilities, the socio-linguistic environment, and language development in children. The results suggest that having good SL abilities can help ameliorate the language disadvantages associated with being raised in a lower SES home environment, offering intriguing new ways to think about the relations between learning, language development, and the social/linguistic environment in which a child is raised. One possible implication of these findings is the possibility of designing intervention programs for children of families with low SES. Recent research has demonstrated that it may be possible to improve SL abilities through targeted computerized training (e.g., Smith, Conway, Bauernschmidt, & Pisoni, 2015). Thus, by promoting SL abilities in children raised in low SES families, it may be possible to facilitate children's development by minimizing the impact of being raised in a less than optimal social and linguistic home environment.

Acknowledgements

We would like to thank all of the participants in this study and their parents as well as our funding source, The National Institutes of Health (Grant R01DC012037).

References

Bronfenbrenner, U., & Morris, P. (1998). *The Ecology of Developmental Processes*. In R. M. Lerner (Ed.), Theoretical models of human development (Vol. 1, 5th ed., pp. 993–1028). New York: Wiley

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language*. Circle Pines, MN: American Guidance Service.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Christiansen, M.H., Conway, C.M., & Onnis, L. (2012). Similar neural correlates of language and sequential learning: Evidence from event-related potentials. *Language and Cognitive Processes, 27*(2), 231-256.

Conway, C.M., Bauernschmidt, A., Huang, S.S., & Pisoni, D.B. (2010). Implicit statistical learning in language processing: word predictability is the key. *Cognition, 114*, 356–371.

Conway, C.M., & Pisoni, D.B. (2008). Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences, 1145*, 113–131.

Crain, S., & Lillo-Martin, D. (1999). *Linguistic Theory and Language Acquisition.* Cambridge, MA: Blackwell.

Dunn, L.M. & Dunn, D.M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Circle Pines, MN: American Guidance Service.

Farah, M. J., Shera, D. M., Savage, J. H., Betancourt, L., Giannetta, J. M., Brodsky, N. L., & ... Hurt, H. (2006). Childhood poverty: specific associations with neurocognitive development. *Brain Research, 1110*(1), 166-174.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Colborn, D. K., Janosky, J., Kurs-Lasky, M., et al. (2003). Parent-reported language skills in relation to otitis media during the first 3 years of life. *Journal of Speech, Language, and Hearing Research, 46*(2), 273–287 http://dx.doi.org/10.1044/1092-4388(2003/022)

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26*(1), 55-88.

Hoff, E., Laursen, B., Bridges, K. (2012) Measurement and model building in studying the influence of socioeconomic status on child development. In: *The Cambridge handbook of environment in human development*, pp 590–606. Cambridge: Cambridge UP.

Hoff, E., & Tian, C. (2005). Socioeconomic status and cultural influences on language. *Journal of Communication Disorders, 38*, 271–278.

Jost, E., Conway, C. M., Purdy, J. D., Walk, A. M., & Hendricks, M. A. (2015). Exploring the neurodevelopment of visual statistical learning using event-related brain potentials. *Brain Research, 1597*, 95-107. doi:10.1016/j.brainres.2014.10.017

Kidd, E., Arciuli, J. (2015) "Individual Differences in Statistical Learning Predict Children's Comprehension of Syntax."*Child Development, 78* (1), 184-193.

Kuhl, P. K. (2007). Is speech learning 'gated'by the social brain?.*Developmental science*, *10*(1), 110-120.

Kuhl, P.K. (2010). Brain mechanisms in early language acquisition. *Neuron, 67*, 713-727.

Lum, J.A.G., Conti-Ramsden, G., Page, D. & Ullman, T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*, 1138 – 1154.

Misyak, J. B., Christiansen, M., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, 1. DOI: 10.3389/fpsyg.2010.00031.

NICHD Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development, 71*, 960–980.

Roberts, E., Bornstein, M. H., Slater, A. M., & Barrett, J. (1999). Early cognitive development and parental education. *Infant and Child Development, 8*, 49–62.

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science, 323*, 951–953.

Saffran, J.R., 2003. Statistical learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110-114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, (5294).

Shafto, C.L., Conway, C.M., Field, S.L., & Houston, D.M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy, 17*, 247-271.

Sheridan, M. A., Sarsour, K., Jutte, D., D'Esposito, M., & Boyce, W. T. (2012). The impact of social disparity on prefrontal function in childhood. *Plos ONE, 7*(4), 1-13. doi:10.1371/journal.pone.0035744

Smith, G.N.L., Conway, C.M., Bauernschmidt, A., & Pisoni, D.B. (2015). Can we improve structured sequence processing? Exploring the direct and indirect effects of computerized training using a mediational model. *PLoS ONE*, *10*(5): e0127148.

Stanton-Chapman, T. L., Chapman, D. A., Bainbridge, N. L., & Scott, K. G. (2002). Identification of early risk factors for language impairment. *Research in Developmental Disabilities, 23*, 390–405.

Teinonen T, Fellman V, Naatanen R, Alku P, Huotilainen M. 2009. Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neurosci. 10*(21).

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative /procedural model. *Cognition, 92*, 231–270.

The Influence of Group Interaction on Creativity in Engineering Design

Trina C. Kershaw[1], Rebecca L. Peterson[1], & Sankha Bhowmick[2]
tkershaw@umassd.edu, rpeterson@umassd.edu, sbhowmick@umassd.edu
Departments of [1]Psychology and [2]Mechanical Engineering, University of Massachusetts Dartmouth
285 Old Westport Road, North Dartmouth, MA 02747 USA

Abstract

Group work is frequently part of idea generation, despite evidence that group interaction may reduce productivity during brainstorming sessions. Idea quantity is one aspect of creativity, but the originality of ideas generated is also important. In this paper, we examine how different aspects of group interaction, such as who makes the most contributions to an idea and the number of group members contribute to an idea, impact the originality of concepts generated by engineering students. We found that the most original concepts were produced when the concept originator was the top contributor to the design, and when the majority of group members contributed to the concept, particularly among senior students. These results are discussed in relation to previous work and suggestions are made for future research that assesses the interaction between design fixation and group processes.

Keywords: creativity; group processes; engineering

Brainstorming (Osborn, 1957) has long been a popular method of idea generation, despite pervasive evidence that individuals produce more ideas than groups (Diehl & Stroebe, 1987, 1991; Mullen, Johnson, & Salas, 1991; Paulus & Brown, 2003). Research about group dynamics during group idea generation sessions suggests that fewer ideas are developed due to production blocking (Diehl & Stroebe, 1987). While group members may be able to generate many ideas, a delay between generating ideas and expressing them, such as when waiting for someone else to finish speaking, leads to fewer ideas being produced by groups compared to individuals (Diehl & Stroebe, 1991).

While there is much research on the negative aspects of brainstorming, there can also be benefits to the technique. For example, Nijstad and Stroebe (2006) found that exposure to group members' ideas led to generation of related ideas, thus suggesting that the other ideas were serving as retrieval cues for related semantic information in long-term memory. Brainstorming is also a popular technique because of higher levels of satisfaction of working in a group rather than working alone (Paulus, Dzindolet, Poletes, & Camacho, 1993). Nijstad, Diehl, and Stroebe (2003) suggest that the experience of fellow group members generating similar ideas reduces an individual's performance anxiety.

Because traditional brainstorming has a mix of positive and negative aspects, several alterations have been proposed to improve the process. One method is brainwriting, in which group members write out their ideas rather than speaking them (VanGundy, 1983). Brainwriting can reduce production blocking because group members do not have to wait to express their ideas. Another suggestion is to have individuals generate ideas alone, and then share them with the group (Diehl & Strobe, 1987). Baruah and Paulus (2008) found that participants who generated ideas alone, then in a group, had a higher quantity of ideas than participants who followed the opposite sequence.

Suggestions such as writing out ideas and working alone before interacting in a group are part of an idea generation technique that is frequently used in engineering research, the 6-3-5 method (Rohrbach, 1969). The 6-3-5 method involves a group of 6 individuals who each write out 3 ideas. The ideas are then circulated through the group so that each group member can comment on and modify the ideas of the other 5 group members. We use a modified version of the 6-3-5 procedure, designed by Otto and Wood (2001), in which group sizes can range from 4-6 participants, ideas are expressed as annotated sketches instead of as words, and the ideas only make three circulations through the group. The modified version of the 6-3-5 method not only involves the suggestions of written instead of oral production of ideas and individual work prior to group interaction, but the groups are *together nominal* groups, in which group members sit together but do not speak to each other. In the modified 6-3-5 method, both the initial individual ideation and subsequent idea circulation periods are performed in silence, rather than with spoken interaction between group members. Mullen et al. (1991) found that together nominal groups had less loss of productivity than traditional brainstorming groups. Further, idea generation techniques that use a combination of annotated sketches plus rotational sharing of others' ideas led to a greater quantity of ideas than idea generation techniques that used words only, sketches only, or gallery sharing of others' ideas, in which generated ideas are all viewed at once (Linsey et al., 2011).

Thus far, our discussion has focused on the quantity of ideas generated, which is only one aspect of creativity. The quality of ideas is also important. In engineering, there is a strong need for the creation of innovative products that can compete in the global marketplace (Duderstadt, 2008). In recent work, we have focused on the development of creativity in undergraduate engineering students (Kershaw et al., 2014, 2015). We have measured the originality of concepts that students produce as solutions to engineering problems that involve designing next-generation products such as alarm clocks or litter collection systems. Thus far, our work has shown that senior students produce more original concepts than freshmen, in line with other similar studies (e.g., Atman, Chimka, Bursic, & Nachtmann, 1999) and well-established theories of skill acquisition (e.g., Fitts & Posner, 1967), which suggest that higher levels of engineering skill

and knowledge are associated with greater levels of creativity (cf. Ericsson, 1999).

Our work thus far, however, has not examined the role of group factors in creativity. While we have used the 6-3-5 idea generation method, which involves together nominal groups, we have only examined differences between groups, such as curriculum level (freshman vs. senior), and have not examined processes that are occurring within the groups that may influence the originality of generated ideas. In the following study, we examine two aspects of group interaction: which group members have the most creative inputs for a concept, and the number of group members that contribute to the overall originality of a concept.

In the 6-3-5 method, multiple group members view each concept and have an opportunity to modify it, but it is possible that some group members may contribute more to the final design than others. Blair and Hölttä-Otto (2012) found that final designs, once they had been circulated through the 6-3-5 group, received higher originality scores than the initial versions of the same concepts. This suggests that the contributions of the concept originator may not be as influential on the originality of the design as the contributions of other group members. Likewise, the cohesiveness of the group may affect overall creativity. Groups that value high standards for performance (Paulus & Brown, 2003) or have experience working together (Milliken, Bartel, & Kurtzberg, 2003) tend to produce more creative ideas.

Based on previous literature, we predicted that groups in which multiple individuals contributed to the originality of a concept would produce the most creative solutions to engineering problems. Based on our previous research, we expected that group factors would interact with students' curriculum level: freshman or senior. Participants who were seniors were enrolled in a year-long capstone design course that required extensive group work. As they were more accustomed to working in groups, and were familiar with the high technical standards expected in engineering, we expected that interactive senior groups would produce the most original concepts.

Method

Participants

Participants were 180 undergraduate students enrolled in freshman and senior undergraduate engineering courses at the University of Massachusetts Dartmouth (UMD). The freshmen (n = 100) were enrolled in an introduction to engineering course that included students in all engineering majors, including mechanical, electrical and computer, and bioengineering. The seniors (n = 80) were enrolled in a senior mechanical engineering design course, which is the capstone course of their curriculum. All students completed the concept generation procedure as part of an in-class exercise about creativity. No demographic data were collected about the participants.

Materials and Procedure

Participants were separated into groups of 4-6 students and provided with a consent letter. They also received blank paper and a colored marker for drawing their concepts. Each group received a sample product for the redesign exercise. For this experiment, the sample product was a simple litter collection device (see Figure 1). After receiving the materials, the experimenters explained the concept generation exercise to the students.



Figure 1: Simple Litter Collection Device

All participants followed the modified 6-3-5 concept generation method (Otto & Wood, 2001). Each participant had 10 minutes to individually sketch 3 concepts of a next generation litter collection system, annotating his/her designs as necessary. After the individual ideation period, participants circulated their concepts clockwise to the next student in the group. Students were instructed that they had 5 minutes to modify the concepts they received, make suggestions or other comments, or sketch a new concept on the same page. Concept sketches continued to be circulated until they returned to the original owner, or at least were passed to 3 different participants within the group. This procedure is the same as was reported in Kershaw et al. (2015).

Analysis

Originality Coding: Concept Level. The originality metric (Table 1) was derived from the CEDA instrument developed by Charyton, Jagacinski, and Merrill (2008). Several modifications were made: first, we only used the originality scale, and not the other items. Second, we modified the 11-point originality scale to a standard 5-point Likert scale, based on higher inter-rater agreement findings by Genco, Johnson, Hölttä-Otto, and Seepersad (2011). Third, we developed a decision tree to assess the originality of the concept as a whole (c.f. Kershaw et al., 2015; Figure 2).

Table 1: Originality Metric

Originality Metric	
0	Common
2.5	Somewhat interesting
5	Interesting
7.5	Very Interesting
10	Exceptional

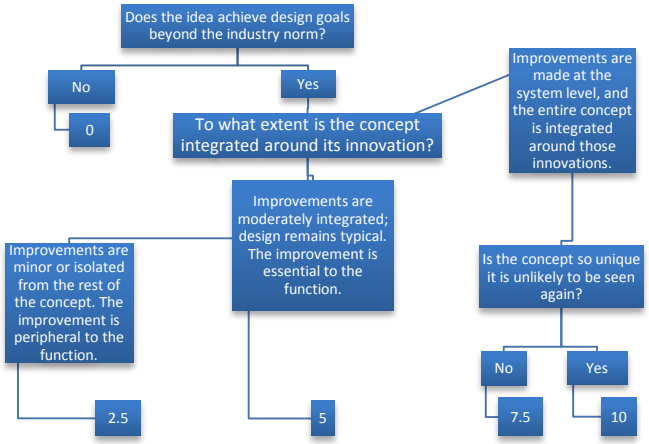


Figure 2: Originality Coding Decision Tree

Of the 465 concepts produced by the participants, a subset of 90 concepts was coded by the first and second authors and a research assistant. The agreement between raters was evaluated using Cohen's (1968) weighted kappa. After the training round of 90 concepts, inter-rater agreement between the three raters exceeded 0.7. The raters discussed any remaining differences and identified a mutually agreeable final rating for the subset of concepts. After the training round, the remaining concepts were coded by the second author or a research assistant.

Originality Coding: Individual Contributions to Concepts. In the following analysis, we assessed the contribution of each individual to each concept produced by the group using the decision tree shown in Figure 2. This analysis was not the improvement that each group member made to the overall concept beyond the originator, but instead, the unique contribution of each individual. We were able to track each individual's contribution because each group member was assigned a different color marker. As was done for the concept-level coding, a subset of the individual contributions were coded by the authors to determine inter-rater agreement. After review of the contributions of 35 individuals to 90 concepts, inter-rater agreement (kappa) between the three authors was .85. The second author then coded the remaining individual contributions to the concepts.

Top Contributor Analysis. Once individual contributions were coded, we examined the weight of each group member's contribution to the originality of the concept. Did the originator of a concept create its most creative aspect? Or was the originality of a concept more due to the contributions of other group members? Or were there equal contributions among group members?

The top contributor scores were derived from examination of the individual-level coding. The designation of the top contributor was scored as follows: if the concept's originator had the highest originality score, a 0 was scored. If a different group member had the highest score, a 1 was scored. If multiple group members had the same top score (e.g., two group members had individual-level originality scores of 2.5), then a 2 was scored. The authors scored a subset of the concepts together but the majority of the scoring was completed by the third author.

Group Contribution Coding. Each of the concepts was coded for the degree to which group members contributed to it. For each concept, the total number of group members, number of group members who had an originality score for a concept, and number of group members who contributed to the concept were tracked. It is important to note the difference between having a score for a concept and actually contributing to it. For example, a group member could have written "good idea" or drawn a smiley face on a concept. This person would have a score, but would not have contributed to the overall originality. Level of group contribution was coded dichotomously: a concept either showed evidence of group members working together (1) or was the product of one group member (0), with the other members simply commenting on the design. A group contribution score was calculated by dividing the number of contributors to the concept by the total number of group members. The second author completed the group contribution coding under the direction of the first author.

Coding Demonstration. A demonstration of the concept-level originality coding, individual-contribution originality coding, top contributor analysis, and group contribution coding is applied to the concepts shown in Figures 3 and 4. The concept shown in Figure 3 received an overall originality score of 5. It is an improvement to the function of a standard litter picker. This design received an originality score of 5 because, according to the decision tree (Figure 2), it shows moderate improvements over a standard litter picker, and these improvements are somewhat integrated. For example, the modified ends are expandable and include sensors that detect particular types of materials. At the same time, much of the design remains typical. It is clear from Figure 3 that the originator of the design, green, has copied his/her drawing from the design of the provided sample litter picker.

At the individual level, green is the originator of the concept, and modifies the basic litter picker by suggesting that different ends are needed to collect more litter. Purple suggests that these different ends should be expandable. Green and purple's ideas were both minor, isolated improvements to the basic litter picker and thus both were given an individual originality score of 2.5. Teal suggests that sensors could be incorporated into these larger, expandable ends to detect particular types of material, such as gold. Teal's contribution suggests a more system-level than isolated change, yet most of the design remains standard. Thus, teal was given an individual originality score of 5, and is also the top contributor to this design. Because red only commented without making any modifications, he/she received an originality score of 0.

In terms of degree of group contribution, this concept showed a high level of interaction between most of the group members, and thus received a group contribution value of 1. Three of the four members of the group contributed to the concept. Thus, it received a group contribution score of .75 (3 contributors/4 total group members).

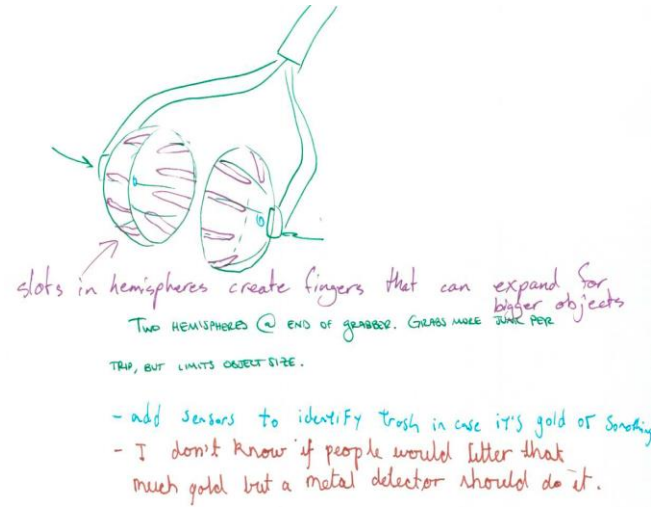


Figure 3: Sample Concept

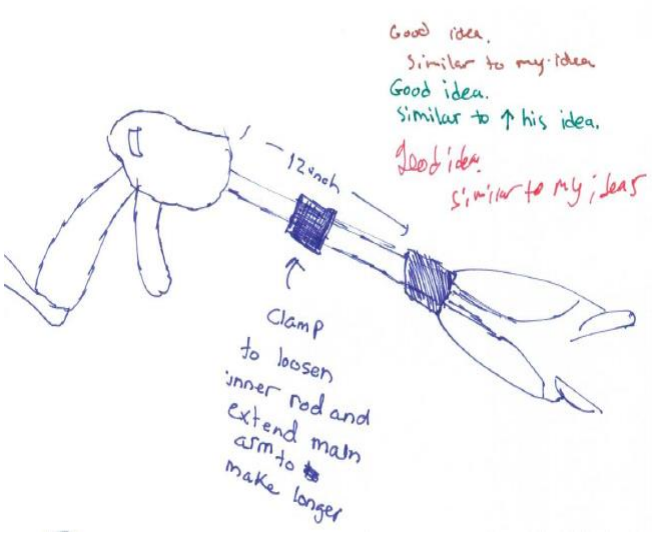


Figure 4: Sample Concept

As a second example, the concept shown in Figure 4 received an overall originality score of 2.5. This design is a standard litter picker that is very similar to the sample that was provided to students, but has a minor addition of telescoping. According to the decision tree (Figure 2), minor or isolated improvements such as telescoping features should receive an overall originality score of 2.5. Because this concept design has only one minor addition to a standard litter picker it received an originality score of 2.5.

At the individual level, this design is an example of one group member, in this case the originator, being responsible for the overall originality while the other members are supportive of the design choices. Blue is the originator and contributed the telescoping feature to the standard picker design and received the individual score of 2.5 for a basic and minor addition to the litter picker. Blue is the top (and only) contributor to the concept. Brown, green, and red each received individual scores of 0 because they did not contribute to the design by drawing in improvements; they only commented in agreement with the design blue made. Because only one group member contributed to the originality of the concept, it received a group contribution value of 0, and a group contribution score of .25 (1 contributor/4 group members).

Results

Participants produced 465 litter collection system concepts. Thirty-eight concepts were removed from final analysis, 11 generated by freshmen and 17 by seniors, because they were not engineering solutions (ex. train a vulture to eat trash), leaving 427 concepts for analysis.

Results regarding the effects of curriculum level on concept-level and individual-level originality scores were discussed in Kershaw et al. (2015). In summary, there was not a significant difference in overall originality between freshmen and seniors. As noted earlier in the paper, however, the goal of the current set of analyses was to examine how interactions between group members affected originality, which was not examined in previous work.

Top Contributors to Originality

For the following analyses, only concepts with originality scores of 2.5 or above were used. If a concept received a score of 0, meaning that it did not achieve design goals beyond industry standards, then there could not be a top contributor. We removed 137 concepts that received originality scores of 0, thus leaving 290 concepts in the analysis. Descriptive statistics for the top contributor analyses are in Table 2.

Table 2: Summary of Top Contributor Originality Scores

	M	*SD*	n Concepts
Freshmen			
• Concept Originator	3.78	1.78	111
• Other Group Member	2.86	1.20	21
• Multiple Group Members	3.08	1.50	13
Seniors			
• Concept Originator	3.71	1.75	101
• Other Group Member	3.10	1.09	21
• Multiple Group Members	3.15	1.55	23

First, we examined the distribution of top contributors within each curriculum level. A chi square test of independence indicated that there was no difference in the distribution of top contributor types (original, other group member, or multiple group members) between freshmen and seniors, χ^2 (2, N = 290) = 3.25, p = .20. A second chi square test showed that having the design originator as the top contributor to originality was more common than having another group member or multiple group members, χ^2 (2, N = 290) = 206.59, p = .0001.

Second, we examined the relationship between curriculum level (freshmen vs. seniors), contributor type (originator, other group member, or multiple group members), and concept-level originality. A two-way between-subjects analysis of variance (ANOVA) was conducted. There was no main effect of curriculum level and no interaction between curriculum level and contributor type, Fs < 1. There was a significant main effect of contributor type, F (2, 284) = 5.12, p = .01, η_p^2 = .04. Post-hoc Tukey tests indicated that groups in which the concept originator was the top contributor had higher originality scores than groups in which a different group member was the top contributor, p = .02, 95% CI = .11-1.44. There were no other significant differences between the contributor types.

Level of Group Contribution

To test the effects of the level of group contribution on originality, all 427 concepts were included. A one-way ANOVA conducted to compare freshmen and seniors indicated that freshmen had higher group contribution scores (M = .87, SD = .24, 95% CI = .81-.88) than seniors (M = .78, SD = .25, 95% CI = .74-.81), F (1, 425) = 8.89, p = .003, η_p^2 = .02, indicating a higher degree of group interaction on the concepts produced by freshmen. A correlation analysis with all students showed a non-significant relationship between group contribution scores and originality, r (427) = .07, p = .16. Because other analyses had shown cohort differences, we also examined this relationship within the freshman and senior groups. Within the freshmen, the level of group contribution was not correlated with originality, r (220) = .01, p = .89. In contrast, within the seniors, the level of group contribution was significantly correlated with originality, r (207) = .14, p = .05.

Discussion

The results indicate that it was most common for the originator of a concept to be the top contributor to a concept than a different member of the group or several group members. This pattern was shown in both freshmen and seniors. We also found that higher originality scores were achieved in groups in which the concept originator was the top contributor. The dominance of the originator as the top contributor, and higher originality scores in concepts with originators as top contributors, indicates that individuals may generate the most original ideas, in contrast to groups. This finding is in contrast with our predictions and with previous data obtained at UMD using similar samples and design problems, in which group members' contributions were found to raise the originality scores of concepts above what was generated by the concept originators (Blair & Hölttä-Otto, 2012). This finding, however, is in line with previous findings that show fewer ideas produced by groups rather than individuals (e.g. Mullen et al., 1991).

One of the chief differences between the current research and that reviewed by Mullen et al. (1991) is that our focus was on originality, not quantity. Further, there are some key differences between our work and that reported by Blair and Hölttä-Otto (2012). First, Blair and Hölttä-Otto only analyzed a set of 15 litter collection system concepts, compared to the 427 concepts that were analyzed for this paper. Second, Blair and Hölttä-Otto applied a feature-level analysis of originality to the concepts instead of using the decision tree we used. Thus, their determination of originality was based on how a given design satisfied different features, such as trash removal or storage, while our determination was based on how well the ideas in a particular design were integrated. One can make the argument that group design may enhance feature-level improvement but not an overall improvement in creativity. Clearly the metrics of the coding were different and one needs to carefully compare the two to ascertain whether there are actually fundamental differences in our results versus those of Blair and Hölttä-Otto (2012).

Another explanation of the top contributor results is related to design fixation. While results indicate that the originator was the predominant contributor in design creativity, it is important to point out that there may have been multiple members in the group who contributed to originality points. However, they were unable to contribute at a higher level of creativity compared to the originator. It may have been difficult for students to improve on an existing design and earn creativity points as opposed to earning points by designing from scratch. So, the lack of contribution could be because students were led into a particular line of thinking based on an existing design. This would be related to design fixation, a common problem in engineering design that can occur because of exposure to group members' ideas. For example, Kohn and Smith (2011) found that more exposure to typical design solutions produced by group members led to lower novelty, yet had no effect of quantity of ideas produced. Further analysis is necessary to assess the role of design fixation in group interactions within our current data set.

In contrast to the findings regarding the top contributor in each group, the level of group contribution was related to originality, but only among the seniors. Senior groups who had higher levels of interaction produced more original concepts. This finding is in line with previous findings that groups with high performance standards and experience working together will produce more creative ideas (Milliken et al., 2003; Paulus & Brown, 2003). Our previous research had shown that seniors had higher originality scores compared to freshmen (Kershaw et al., 2014). Although we did not replicate that finding here, the group contribution results suggest that something different is happening within the senior groups than the freshman groups. Our seniors completed the litter collection system exercise within the project groups within their capstone course. Group dynamics that were already established could have played a role, in that some project groups could be more egalitarian in design decisions while others may have had a clear leader who dominated group decisions. Alternatively, individual ability differences could affect group outcomes, with higher levels of intelligence, need for cognition, or flexibility affecting the

originality of designs that groups produced. Further research is needed to explore these factors.

Acknowledgements

The authors thank Molly McCarthy for assistance with the originality coding, and Adam Young for assistance with development of the decision tree. We also thank UMD engineering instructors Don Foster, Marc Richman, Mehdi Raessi, and Steve Warner for access to their classes.

The authors acknowledge support from NSF Grant DUE-1140424. Any opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

References

Atman, C.J., Chimka, J.R., Bursic, K.M., & Nachtmann, H.L. (1999). A comparison of freshman and senior engineering design processes. *Design Studies, 20*, 131-152.

Baruah, J., & Paulus, P.B. (2008). Effects of training on idea generation in groups. *Small Group Research, 39*, 523-541.

Blair, B.M., & Hölttä-Otto, K. (2012). Comparing the contribution of the group to the initial idea in progressive idea generation. *Proceedings of the ASME 2012 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. Paper DETC2012-70309.

Charyton, C., Jagacinski, R.J., & Merrill, J. (2008). CEDA: A research instrument for creative engineering design assessment. *Psychology of Aesthetics, Creativity, and the Arts, 2*, 147-154.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213-220.

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology, 53*, 497-509.

Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of Personality and Social Psychology, 61*, 392-403.

Duderstadt, J.J. (2008). *Engineering for a changing world: A roadmap to the future of engineering practice, research, and education*. Ann Arbor, MI: The University of Michigan.

Ericsson, K.A. (1999). Creative expertise as superior reproducible performance: Innovative and flexible aspects of expert performance. *Psychological Inquiry, 10*, 329-333.

Fitts, P.M., & Posner, M.I. (1967). *Human performance*. Belmont, CA: Brooks Cole.

Genco, N., Johnson, D., Hölttä-Otto, K., & Seepersad, C.C. (2011). A study of the effectiveness of the empathic experience design creativity technique. Paper Number: DETC2011-021711. *ASME IDETC Design Theory and Methodology Conference*, Washington, DC.

Kershaw, T.C., Peterson, R.L., McCarthy, M.A., Young, A.P., Seepersad, C.C., Williams, P.T., Hölttä-Otto, K., & Bhowmick, S. (2015). A cross-sectional and longitudinal examination of the development of innovation capability in undergraduate engineering students. *Proceedings of the ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. Paper DETC2015-47650.

Kershaw, T.C., Seepersad, C.C., Hölttä-Otto, K., Williams, P.T., Young, A.P., Bhowmick, S., & McCarthy, M.A. (2014). The effects of the undergraduate curriculum and individual differences on student innovation capabilities. *Proceedings of the ASME 2014 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*. Paper DETC2014-35540.

Kohn, N.W., & Smith, S.M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology, 25*, 359-371.

Linsey, J.S., Clauss, E.F., Kurtoglu, K., Murphy, J.T., Wood, K.L., & Markman, A.B. (2011). An experimental study of group idea generation techniques: Understanding the roles of idea representation and viewing methods. *Journal of Mechanical Design, 133*, 1-15.

Milliken, F.J., Bartel, C.A., & Kurtzberg, T.R. (2003). Diversity and creativity in work groups: A dynamic perspective on the affective and cognitive processes that link diversity and performance. In P.B. Paulus & B.A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 32-62). New York: Oxford University Press.

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology, 12*, 3-23.

Nijstad, B.A., Diehl, M., & Stroebe, W. (2003). Cognitive stimulation and interference in idea-generating groups. In P.B. Paulus & B.A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 137-159). New York: Oxford University Press.

Osborn, A.F. (1957). *Applied imagination*. New York: Scribner.

Otto, K. N., & Wood, K. L. (2001). *Product design: Techniques in reverse engineering and new product development.* Upper Saddle River, NJ: Prentice Hall.

Paulus, P.R., & Brown, V.R. (2003). Enhancing ideational creativity in groups: Lessons from research on brainstorming. In P.B. Paulus & B.A. Nijstad (Eds.), *Group creativity: Innovation through collaboration* (pp. 110-136). New York: Oxford University Press.

Paulus, P.B., Dzindolet, M.T., Poletes, G., & Camacho, L.M. (1993). Perceptions of performance in group brainstorming: The illusion of group creativity. *Personality and Social Psychology Bulletin, 19*, 79-89.

VanGundy, A.B. (1983). Brainwriting for new product ideas: An alternative to brainstorming. *Journal of Consumer Marketing, 1*, 67-74.

There is more to gesture than meets the eye: Visual attention to gesture's referents cannot account for its facilitative effects during math instruction

Miriam A. Novack[1] (mnovack1@uchicago.edu)
Elizabeth M. Wakefield[1] (ewakefield@uchicago.edu)
Eliza L. Congdon[1] (econgdon@uchicago.edu)
Steven Franconeri[2] (franconeri@northwestern.edu)
Susan Goldin-Meadow[1] (sgm@uchicago.edu)

[1]Department of Psychology, University of Chicago,
5848 S University Ave, Chicago, IL 60637 USA

[2]Department of Psychology, Northwestern University
2029 Sheridan Rd, Evanston, IL 53706 USA

Abstract

Teaching a new concept with gestures – hand movements that accompany speech – facilitates learning above-and-beyond instruction through speech alone (e.g., Singer & Goldin-Meadow, 2005). However, the mechanisms underlying this phenomenon are still being explored. Here, we use eye tracking to explore one mechanism – gesture's ability to direct visual attention. We examine how children allocate their visual attention during a mathematical equivalence lesson that either contains gesture or does not. We show that gesture instruction improves posttest performance, and additionally that gesture *does* change how children visually attend to instruction: children look more to the problem being explained, and less to the instructor. However looking patterns alone cannot explain gesture's effect, as posttest performance is not predicted by any of our looking-time measures. These findings suggest that gesture does guide visual attention, but that attention alone cannot account for its facilitative learning effects.

Keywords: Gesture; eye tracking; learning; visual attention

Introduction

Teachers use more than words to explain new ideas; they often accompany their speech with gestures – hand movements that express information through both form and movement patterns. Teachers gesture spontaneously in instructional settings (Alibali et al., 2014) and controlled experimental studies have found that children are more likely to learn novel ideas from instruction that includes speech and gesture, than speech alone (e.g., Ping & Goldin-Meadow, 2008; Singer & Goldin-Meadow, 2005; Valenzeno, Alibali, & Klatzy, 2003).

Gesture might improve learning by conveying multiple ideas simultaneously (Singer & Goldin-Meadow, 2005), engaging the motor system (Macedonia, Muller, & Friederici, 2011), and linking abstract ideas to concrete objects in the environment (Valenzeno et al., 2003). One understudied potential benefit of gesture is that it engages and directs visual attention. Here we used instruction on the concept of mathematical equivalence as a case study to test how gesturing towards a novel mathematical equation affects not only children's learning outcomes, but also their allocation of visual attention across the equation.

There are reasons to think that gesture's ability to direct visual attention to relevant objects may underlie its positive effects. Because gesture is a spatial, dynamic social cue, it can focus a listener's visual attention on a specific part of the visual environment. Even young infants will shift their visual attention in response to gesture (Rohlfing, Longo, & Bertenthal, 2012). This could, in turn, increase the likelihood that children would focus on crucial aspects of a problem being taught, and would thus learn more from instruction. Learners likely need to attend to the critical information in an instructional context in order to learn from it. For example, toddlers are more likely to learn pairings between objects and labels if their attention is focused on the object while it is being labeled (Yu & Smith, 2012). If gesture during instruction highlights important features of the problem and causes learners to visually fixate on these features while relevant information is being provided in speech, that increased looking should lead to better learning.

Previous work using eye tracking to understand how people process gesture has focused on visual processing of naturally produced gesture during face-to-face communication, such as when watching a person tell a story. Most of this work has been descriptive, documenting where interlocutors focus their visual attention during communication rather than documenting how patterns of visual attention affect comprehension. Overall, the findings suggest that looking directly toward a speaker's hands is actually quite rare (e.g., Gullberg & Holmqvist, 2006; Gullberg & Kita, 2009). Instead, listeners prefer to look mostly at a speaker's face and spend little time overtly attending to gesture. On the rare occasions when interlocutors do look directly at a gesture, it is typically because the speaker himself is looking towards his own

hands, or is holding a gesture in space for an extended period of time (Gullberg & Kita, 2009).

While this descriptive work on visual attention to gesture during spontaneous discourse is informative, we cannot assume the findings will be consistent in instructional settings. First, unlike the discourse studies described above, classroom teachers often gesture towards or near objects (Alibali & Nathan, 2012). In fact, most of the behavioral work that investigates the utility of teachers' gestures has been in situations where gestures are performed in reference to objects (e.g., Ping & Goldin-Meadow, 2008; Singer & Goldin-Meadow, 2005; Valenzeno, et al., 2003). For example, children learn more when a teacher gestures toward a math problem that is written on a chalkboard (Singer & Goldin-Meadow, 2005). This means that in most formal instructional settings, learners have three demands on their visual attention capacities – the instructor who is speaking, the gestures she produces, and the objects she is gesturing toward. Thus, the way in which gesture affects allocation of visual attention in these situations may differ drastically from other kinds of conversational settings.

Second, and more importantly, the way gesture captures or directs visual attention during instruction may have different cognitive implications than how gesture functions in discourse. Specifically, learning involves more than just comprehension of the content of a message; it requires that learners integrate the presented information with their existing knowledge to arrive at a novel conceptual state. This is a non-trivial difference between comprehension and learning, and it may mean that gesture necessarily serves a different function in an instructional context than it does during casual conversation. If learners are sensitive to this, then we might expect that the way gesture affects visual attention during instruction will meaningfully map onto learning outcomes.

In the current study, we ask how gesture directs visual attention for 8-10 year-old children who are learning how to solve missing addend equivalence problems (e.g., *2+5+8 = __+8*). We use eye tracking to compare children's visual attention to instructional videos with either speech alone, or speech with accompanying gesture. Previous work using a similar paradigm has found that giving children relatively brief instruction, using example problems, and allowing children to solve additional problems themselves results in an increased understanding of mathematical equivalence. Importantly, incorporating gesture into instruction boosts this understanding (e.g., Singer & Goldin-Meadow, 2005) relative to instruction with speech alone. In the present study, we use a *grouping* gesture during instruction. This gesture involves producing a V-point to the first two numbers in a missing addend equivalence problem followed by a point to the blank space. This V-point gesture represents the idea that one can solve the equation by adding, or grouping, the first two addends and putting that total in the blank. This V-point gesture is one produced spontaneously by children who already understand how to solve these sorts of problems (e.g., Perry, Church, & Goldin-Meadow, 1988) and has also been shown to lead to learning when taught to children (Goldin-Meadow, Cook & Mitchell, 2009). Furthermore, this particular gesture is of interest because it contains both *deictic* properties (pointing to specific numbers) and *iconic* properties (representing the idea of grouping through its form). Therefore, the benefits of learning from this type of gesture could arise from looking to the gesture itself, from looking to the numbers that the gesture is referencing, or from some combination therein.

Methods

Participants

Data from 50 participants were analyzed for the present study. Children between the age of 8 and 10 (*mean age* = 8.8 years) were recruited through a database maintained by the University of Chicago Psychology Department and tested in the laboratory. The sample includes 26 children in the Speech+Gesture Condition (14 females) and 24 children in the Speech Alone Condition (14 females). All children in the current sample scored a 0/6 on a pretest, indicating that they did not know how to correctly solve mathematical equivalence problems at the start of the study. Prior to the study, parents provided consent and children gave assent. Children received a small prize, and $10 compensation for their participation.

Materials

Pretest/Posttest. The pretest and posttest each contained 6 missing addend equivalence problems, presented in one of two formats. In Form A, the last addend on the left side of the equals sign was repeated on the right side (e.g., a+b+c= __+c') and in Form B, the first addend on the left side of the equals side was repeated on the right side (e.g., p+q+r = p'+__). Both pretest and posttest consisted of 3 of each problem type.

Eye Tracker. Eye tracking data were collected via corneal reflection using a Tobii 1750 eye tracker with a 17 inch monitor. Tobii software was used to perform a 5-point calibration procedure using standard animation blue dots. This was followed by the collection and integration of gaze data with the presented videos using Tobii Studio (Tobii Technology, Sweden).

Instructional videos. Two sets of 6 instructional videos were created to teach children how to solve Form A missing addend math problems (e.g., 5+6+3=__+3) – one set for children in the Speech Alone condition and one set for children in the Speech+Gesture condition. All videos showed a woman standing next to a Form A missing addend math problem, written in black marker on a white board. At the beginning of each video, the woman said, "Pay attention to how I solve this problem", and then proceeded to write the correct answer in the blank (e.g., writing 11 in the previous example). She then described how to solve the

problem, explaining the idea of *equivalence*: "*I want to make one side equal to the other side*. 5 plus 6 plus 3 equals 14, and 11 plus 3 is 14, so *one side is equal to the other side.*" During this spoken instruction, the woman kept her gaze on the problem. In the Speech+Gesture videos, the woman accompanied her speech with a gesture strategy. When she said "*I want to make* ***one side****…"*, she simultaneously produced a V-point with her index and middle figure to the first two addends, then, as she said "…the ***other side***" she moved her hand across the problem, bringing her fingers together to point to the answer with her index finger. She produced no gestures in the Speech Alone videos. To ensure that the speech was identical across the two training conditions, the actress recorded a single audio track for each problem, prior to filming. Each of the twelve videos was approximately 25 seconds long.

Procedure

Children first completed a written pretest containing 6 missing addend math problems. All children in the current sample scored 0/6[1]. The experimenter then wrote children's (incorrect) answers on a white board and they were asked to explain their solutions.

Next, children sat in front of the eye tracking monitor, approximately 18 inches from the screen, and were told they would watch instructional videos that would help them understand the type of math problems they had just solved. Their position was calibrated and adjusted if necessary, then they began watching the first of the 6 instructional videos (either Speech Alone, or Speech+Gesture, depending on the assigned training condition). At the conclusion of each video, children were asked to solve a new missing addend problem on a small, hand-held whiteboard, and were given feedback on whether or not their answer was correct (e.g., "that's right, 10 is the correct answer" or "no, actually 10 is the correct answer"). All problems shown in the instructional videos were Form A, and all problems that children had the opportunity to solve were Form A.

After watching all 6 instructional videos and having 6 chances to solve their own problems during training, children completed a new, 6-question paper-and pencil posttest. The posttest, like the pretest, included 3 Form A problems and 3 Form B problems. As children saw only Form A problems during training, we refer to these as "Trained" problems and Form B as "Transfer" problems.

Results

Behavioral Results

Training. Figure 1 shows the proportion of participants in each condition who answered problems correctly during training. A mixed-effects logistic regression predicting the log-odds of success on a given training problem with problem number (1-6) and condition (Speech Alone, Speech+Gesture) as fixed factors and subject as a random factor revealed a positive effect of training problem (β=0.91, *SE*=0.15, *z*=6.21, *p*<.001), indicating that children became more likely to correctly answer problems as training progressed. There was, however, no effect of condition during training (β=0.03, *SE*=0.72, *z*=0.04, *p*=.96, indicating that learning rates during training did not differ by condition. By the final training problem, over 90% of participants in both groups were answering the problems correctly, which suggests that both types of instruction were equally comprehensible.

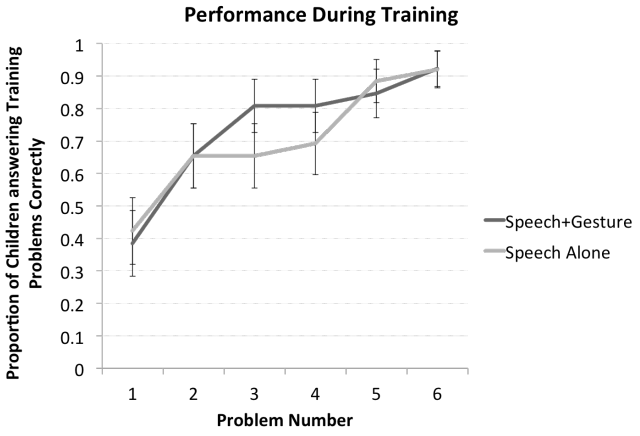


Figure 1. Performance during training on practice problems. Learning increased across the 6 problems, but was not different across the two training conditions.

Posttest. Although the groups did not differ in performance at the end of training, their scores on an immediate posttest reflected an advantage of having learned through Speech+Gesture instruction (see Figure 2). Participants in the gesture condition answered significantly more problems correct at the posttest (*M*=4.11, *SD*=2.04) than participants in the speech condition (*M*=2.64, *SD*=2.08). A mixed-effects logistic regression with problem type (Form A: trained, Form B: transfer) and Condition (Speech+Gesture, Speech Alone) as fixed factors and subject as a random factor showed a significant effect of condition (β = -2.60, *SE* =0.99, *z*=2.59, *p*<.01) indicating that posttest performance in the Speech+Gesture Condition was better than performance in the Speech Alone Condition. There was also a significant effect of problem type (β=2.27, *SE*=0.43, z=5.31, *p*<.001), demonstrating that performance on Form A (trained problems) was better than performance on Form B: (transfer problems). There was no significant interaction between Condition and Problem Type (β=0.29, *SE*=0.79, *z*=-0.37, *p*=0.71).

[1] Children who answered pretest problems correctly (n=59) were still run in the study but are excluded from the current analyses.

[2] There was a gesture space in the Speech Alone video, despite

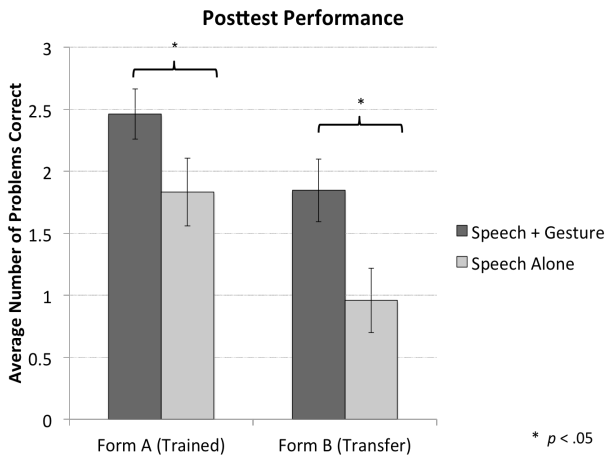


Figure 2. Posttest performance by condition and problem type. Error bars represent +/−1 standard error of the mean.

Eye-Tracking Results

We used a multistep process to analyze the eye tracking data: (1) Areas of interest (AOIs) were generated for the instructor, problem and gesture space[2] (See Figure 3) using Tobii Studio. Fixations outside of these AOIs were collapsed into "Other". (2) Data were extracted and processed, such that the AOI a participant fixated in could be determined at 50 msec intervals across the entire length of each problem. (3) Time segments of interest, during which a particular event was happening in the videos (e.g., the instructor stating the equalizer strategy, "*I want to make one side equal to the other side*") were identified, and total gaze duration during a given time segment in each AOI were computed. (4) We calculated the proportion of time a participant spent in each AOI within each segment collapsed across all six problems. For each participant, eye tracking data were excluded if visual inspection showed that the calibration was off. On average, children in the Gesture Condition contributed data from 4.96 (*SD* = 1.34) trials, and children in the Speech Condition contributed data from 4.90 trials (*SD* = 1.34).

Allocation of visual attention across conditions.

To determine whether patterns of visual attention differed when children were instructed through Speech+Gesture vs. Speech Alone, we considered the proportion of time children spent in each AOI for two time segments of interest. The **strategy** segment encompassed time when the instructor stated the equalizer strategy: *I want to make one side, equal to the other side.* During this segment, spoken instruction was identical across conditions, but children in the Speech+Gesture condition also saw co-speech instructional gestures. As the strategy was explained twice per problem, data from these epochs were combined into one segment of interest. The **explanation** segment encompassed time when the instructor elaborated on the strategy, highlighting the particular addends in the problems (e.g., "*5 plus 6 plus 3 is 14, and 11 plus 3 is 14*"). This segment was visually identical across the experimental groups, allowing us to ask whether the presence of gesture during the preceding **strategy** segment caused children in the Speech+Gesture condition to focus their visual attention in the subsequent **explanation** segment differently than those in Speech Alone instruction.

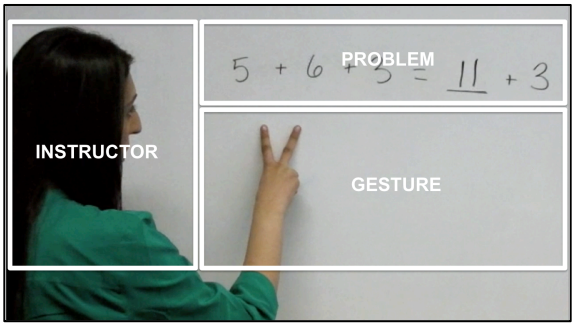


Figure 3. Still shot taking during a gesture segment, with AOIs overlaid.

Strategy segment. Figure 4 shows the proportion of time children spent looking in each of the AOIs during the strategy segment in each condition. On average, children in the Speech+Gesture condition spent a greater proportion of time looking to the problem itself compared to children in the Speech Alone condition (60% versus 48%) (β=0.11, *SE*=0.05 *t*=2.39, $p<0.05$). In contrast, children in the Speech Alone condition allocated more visual attention to the instructor, compared to children in the Gesture condition (47% vs. 18%) (β =-0.29, *SE*=0.04 *t*=-6.19, $p<0.01$). Finally, children in the Speech+Gesture condition spent 19% of the time looking to the Gesture space. Unsurprisingly, children in the Speech Alone condition spent significantly less time (3%) in this AOI (β=.16, *SE*=0.02 *t*=5.63, $p<0.01$) as there was nothing there to draw their attention. Together, these results suggest that gesture does affect visual attention in an instructional context, leading participants to look more to the objects being referenced, and less to the instructor herself.

Explanation segment. Figure 4 also shows the proportion of time spent in each AOI during the explanation segment, with children across both conditions splitting their time evenly between the instructor and the problem. Analyses indicated that there were no differences in looking times to the AOIs by Condition during the explanation segments.

[2] There was a gesture space in the Speech Alone video, despite the fact that there was never any gesture produced in those videos.

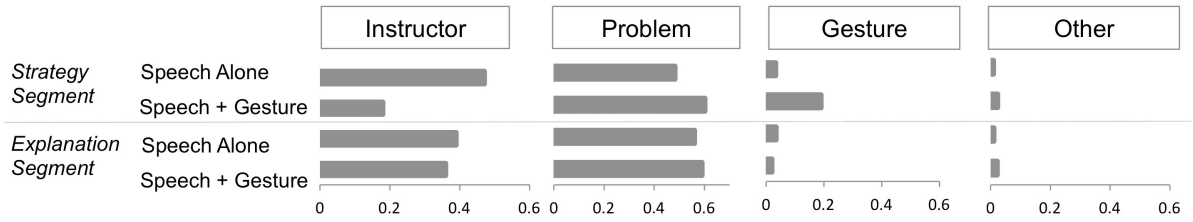


Figure 4. Average proportion of gaze duration across all 6 problems during strategy and explanation segments.

Relation between visual attention and learning.
Given the condition differences between the allocation of visual attention during the strategy segment of instruction, we were interested in whether the focus of attention elicited by the presence of gesture predicted learning outcomes. To explore this we conducted a regression to determine whether looking towards the problem itself (which children did more in the Speech+Gesture condition) predicted posttest performance. Proportion of time looking to the problem did not predict performance on the posttest (β=2.53, *SE*=1.89, *t*=1.34 *p*=0.18). In other words, the presence of gesture did lead children to look more to objects referenced by gesture but that increase in looking was not responsible for the increase in learning outcomes. Focusing just on the gesture condition, we see that children spend relatively little time looking directly at the gesture (only about 19%), and the amount of looking to the gesture itself, while it is being produced, has no relation to learning outcomes within the gesture condition (β=1.96, *SE*=4.47, *t*=0.44 *p*=0.66).

Discussion

Although decades of work have found that gesture supports learning when added to instructional contexts, this was the first study to ask how gesture during instruction guides visual attention and facilitates learning through an attentional mechanism. Our behavioral results replicate previous work (e.g., Singer & Goldin-Meadow, 2005). We show that children who learn from watching speech+gesture instruction have more robust learning than children who learn from speech alone, as demonstrated by higher performance on a posttest. Importantly, and surprisingly, we also add a novel finding to the behavioral literature. Whereas most researchers consider posttest performance alone as a measure of learning, we asked how children's performance changed *during* instruction. We show that learning rates *during* instruction did not differ across the two groups, but only emerged after a change in context (i.e., moving from sitting in front of the eye tracker to a desk), and when intermittent reminders of the strategy were not present. This suggests that our learning paradigm may only produce fragile, temporary learning outcomes, but that the addition of gesture to the instruction can help solidify that knowledge. This short-term retention effect corroborates previous work showing that the effects of gesture are particularly good at promoting long-lasting learning (e.g., Cook, Mitchell, & Goldin-Meadow, 2008).

Our eye tracking results demonstrated that at a global level, gesture directs visual attention towards spoken referents in a formal, instructional context, and that children are more likely to focus on referents of gesture than gesture itself. This is interesting, given that the Speech+Gesture videos contained more *items* (i.e., moving hands) for children to look at than the Speech Alone videos, and yet, children in this condition focused the majority of their attention on the problem. Relatedly, it is notable that there was relatively little overt focus on the gesture form, even though previous work suggests that the form of the gesture itself is important for learning in this task (Goldin-Meadow et al., 2009). Finally, in terms of general looking patterns, we found that although gesture affects visual attention when it is being produced, it does not affect visual attention of subsequent speech-only instruction, as seen from our analysis of the explanation segment of instruction.

Our looking time findings suggest similarities between natural communicative gesture, and purposeful, instructional gesture. Like work on communicative gesture, we find that looking directly at gesture is relatively uncommon. However, our results may suggest a difference between natural and instructional gesture contexts: even though fixation on gesture is relatively rare, gesture in instructional contexts may draw more attention than gesture in natural communication. When gesture was present in the current study, all children in the sample looked directly at it, at least for some amount of time. In a study of gesture in discourse, only 9% of gestures were ever fixated (Gullberg & Holmqvuist, 2006). This difference may be attributable to the way gestures were used in our instruction that differ from their use in discourse. In our videos, gestures were front-and-center – they were in the middle of the screen, while the instructor was faced away from the child, providing a cue to their importance. In contrast, in previous studies of communicative gesture in discourse, participants see face-to-face communication, where the face may take center stage. Further work examining more types of instructional gesture (and perhaps less salient instructional gestures) may reveal what is driving this difference.

In our final analysis, we asked whether attention to the problem during the strategy segment of instruction led to better posttest performance, with the rationale that finding this link would suggest that at least part of the facilitative effects of gesture in previous studies is driven by its ability to guide attention. Although we did not find evidence that gesture enhances learning by highlighting important features

of a problem, and increasing fixation to those features, it may be possible for gesture to highlight important *relational* aspects of a problem, which will be examined in future work. For example, adults solving these same kinds of math problems are less likely to make errors if they traversed the equal sign, a gaze pattern that may be highlighting the relational structure of the equation (Chesney et al., 2013). Thus it is possible that gesture could lead to useful eye-movement patterns not captured by the current analysis, which could in turn support learning outcomes.

It also remains possible that the effect of gesture on visual attention is not the main mechanism through which gesture facilitates learning. For example, Ping & Goldin-Meadow (2008) found that 5-6 year olds were just as likely to improve their understanding of Piagetian conservation after a lesson that included gesture, irrespective of whether or not the objects to which the gestures referred (i.e., glasses that contained water) were present. In another study, Goldin-Meadow, Cook, & Mitchell (2009) taught children how to solve missing addend equivalence problems by producing a grouping gesture either to the correct, or incorrect addends to be grouped. Remarkably, children learned even if they had produced the V-point to the wrong addends, suggesting that directing visual attention to the wrong place does not disrupt gesture's positive effects on learning. Still, it seems likely that visual attention is part of the story. In fact, in the example given above, Goldin-Meadow et al. (2009) found that although children could learn from an 'incorrect' gesture, they benefitted more from the same gesture, used to highlight grouping of the correct addends, and, presumably, draw visual attention to these addends.

In the present study, we have established that instructional gesture *does* drive children to look at a novel equation differently, and children show increased learning after this type of instruction; we have just also shown that this shift in global looking pattern does not provide a simple causal explanation for this cognitive effect. Future work will consider how more nuanced aspects of visual attention, such as whether it helps children synchronize their looking with spoken instruction, as well as ways in which the ability to guide visual attention may combine with other features of gesture to support learning.

Acknowledgments

Funding for this study was provided by NICHD (R01-HD47450, to Goldin-Meadow), NSF BCS 1056730, and the Spatial Intelligence and Learning Center (SBE 0541957, Goldin-Meadow is a co-PI) through the National Science Foundation. We also thank Kristin Plath, William Loftus, and Aileen Campanaro for their help with data collection, and Amanda Woodward for the use of the eye tracker.

References

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in Mathematics Teaching and Learning: Evidence From Learners' and Teachers' Gestures. *Journal of the Learning Sciences*, *21*, 247–286. doi:10.1080/10508406.2011.611446

Alibali, M. W., Nathan, M. J., Wolfgram M. S., Church, R. B., Jacobs, S.A., Johnson Martinez, C. & Knuth, E. J. (2014) How Teachers Link Ideas in Mathematics Instruction Using Speech and Gesture: A Corpus Analysis, *Cognition and Instruction, 32*, 65-100. doi: 10.1080/07370008.2013.858161

Chesney, D. L., McNeil, N. M., Brockmole, J. R., & Kelley, K. (2013). An eye for relations: eye-tracking indicates long-term negative effects of operational thinking on understanding of math equivalence. *Memory & Cognition, 41*, 1079-1095.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*, 1047-1058. doi: 10.1016/j.cognition.2007.04.010

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009).Gesturing gives children new ideas about math. *Psychological Science, 20*, 267–272. doi:10.1111/j.1467-9280.2009.02297.x

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of nonverbal behavior*, *33*, 251-277. doi: 10.1007/s10919-009-0073-2

Gullberg, M., & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition, 14*, 53–82. doi: http://dx.doi.org/10.1075/pc.14.1.05gul

Macedonia, M., Muller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32*, 982-998. doi: 10.1002/hbm.21084

Rohlfing, K. J., Longo, M. R., & Bertenthal, B.I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. *Developmental Science, 15,* 426-435. doi:10.1111/j.1467-7687.2012.01139.x

Singer, M. a, & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. *Psychological Science*, *16*, 85–89. doi:10.1111/j.0956-7976.2005.00786.x

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988).Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359–400. doi:10.1016/0885-2014(88)90021-4

Ping, R. M., & Goldin-Meadow, S. (2008). Hands in the air: using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology*, *44*, 1277–87. doi: 10.1037/0012-1649.44.5.1277

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*, 187–204. doi: 10.1016/S0361-476X(02)00007-3

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*, 244-262. doi:10.1016/j.cognition.2012.06.016

Measuring and modeling distraction by self-referential processing in a complex working memory span task

Jeroen Daamen (daamenjeroen@gmail.com)
Marieke van Vugt (m.k.van.vugt@rug.nl)
Niels Taatgen (n.a.taatgen@rug.nl)
Institute of Artificial Intelligence & Cognitive Engineering, Nijenborgh 9
9747 AG, Groningen, The Netherlands

Abstract

Two experiments using novel complex working memory span tasks were performed, both requiring the participants to remember a span of letters whilst being distracted by the processing of words. Word processing could either be self-referential (SRP) or not. In the first experiment recall performance was compared between SRP and non-SRP conditions using the same words. In the second experiment, we compared SRP and non-SRP in two tasks equalized in semantic processing but using different words. In both experiments recall performance was significantly lower after SRP compared to non-SRP, indicating that SRP has a disruptive effect on the recall task. A cognitive model implemented in PRIMs, using goal competition during SRP, interfering with rehearsal of letters, could account for the observed experimental results. If SRP interferes with subsequent tasks in this manner it should also interfere with tasks other than recall, such as SRP occurring in daily life.

Keywords: self-referential processing, distraction, cognitive modeling, complex working memory

Introduction

Distractions from ongoing tasks form a problem in our day-to-day lives, it reduces our productiveness and can have negative consequences on task performance. Distractions can come from our perceptual inputs (external) or from our mind (internal), the latter can be viewed as distractions by self-generated thought, that is "*mental contents that are not derived directly from immediate perceptual input*" (Smallwood & Schooler, 2015). These thoughts can occur as part of a task, for example when one needs to construct an internal representation or mentally weigh the different factors of a decision, or they can be task independent when they stray from the task at hand, for example when we let our minds wander. In demanding tasks such as driving a car or piloting a plane the consequences of internal distractions can be severe (Casner & Schooler, 2014; Yanko & Spalek, 2013 respectively).

Having objective measures of distractions caused by self-generated thought and having a better understanding of the mental processes underlying them will allow us to better understand and prevent undesired effects of mental distraction during important tasks. For this reason we set out to measure and model the distractions caused by self-generated thought during self-referential processing (SRP), the processing of information in relation to the self (see introduction Northoff et al., 2006). We therefore designed a novel experimental paradigm to measure the distracting effects SRP on one's ability to recall presented letters, using a complex working memory (CWM) span task as a basis. We hypothesized that distraction by SRP leads to worse CWM span task performance than distraction by a non-SRP task as SRP may lead to task-unrelated self-generated thoughts instead of task-related maintenance rehearsal.

Methods

Two CWM span task experiments were performed varying only slightly in set-up. The first experiment will be described in detail, for the second experiment only the differences with the first will be noted.

Experiment 1

Participants

Subject recruitment for both experiments was done via a Facebook post on the "*Paid research participants Groningen*" group offering 10 euros for those who decided to participate (experiment duration approximately 1 hour). 27 participants were included in the first experiment (19 female, age 22.3 ± 2.7). Only native Dutch speakers were included in the experiment. Informed consent was obtained from all participants.

Task

In this experiment participants were required to remember presented letters while processing presented words (Figure 1). It was created using PsychoPy (Peirce, 2007).

The screen background was dark grey and all text was presented in white (Gill Sans MT, font height ~1cm). The experiment consisted of 12 blocks, with one block containing each combination of span and condition once. The spans used were 3, 4 and 5, as is common in CWM span tasks (Conway et al., 2005). For the storage task, participants needed to remember letters that were presented one at a time on the screen for 1s, and between each presentation there was 4s of self-paced processing of word stimuli (SRP or neutral – see below). Before each letter presentation the screen was blank for 1s to allow for rehearsal. We included these delays on purpose to maximize the potential for distraction by SRP.

Each trial started with showing the participant the current condition. For the SRP condition this was "Does this word describe you? (Yes/No)", for the neutral condition this was "Does the word contain the letter 'a'? (Yes/No)". The letter 'a' was chosen for the neutral condition because it was

present in roughly half of the used word stimuli (48.0%). Both sentences were in Dutch. Participants were able to respond by pressing the left (labeled 'NO') or right 'ctrl' (labeled 'YES') buttons.

In the next phase a random letter stimuli was presented in the center of the screen for 1.0s. Before the presentation the screen was blank for 1.0s. Directly after the letter presentation followed 4.0s of self-paced processing tasks (SRP or non-SRP). As soon as a participant responded to a word the next word would be presented. If there were less than 700ms remaining, the screen would stay blank for the remaining time to prevent participants being flashed by a stimulus at the end of the phase. These phases were repeated a number of times equal to the current span.

The recall phase was indicated by a number of underscores equal to the current span. The underscores were replaced by the user's input as they started typing. Error correction was possible by using the backspace key. When they entered the last letter the feedback was presented. Participants were instructed to guess if they couldn't remember a letter. The participants were shown how well they did on the storage task in the form of "[x] out of [span] letters correct". They also received their average response time in the processing task as well as their percentage of correctly judged processing items for the neutral condition. Due to the subjectivity of the SRP condition there was no score shown. A pilot study showed that participants were consistent with their previous responses in the SRP condition, indicating that feedback on this was not critical.

Scoring

The storage task was scored using partial-credit unit scoring (Conway et al., 2005). That is, the score for each trial was calculated as number of items in correct serial position divided by the span of that trial. The processing task was scored using the percentage of correctly processed items. If the last processing item did not receive a response before going to the next phase, this item wasn't taken into account for the final score.

Stimuli

To-be-remembered stimuli were chosen from the set of all consonants (i.e., B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, and Z). Within one trial no letters were repeated. No vowels were used to prevent easy grouping of letter stimuli by remembering them as words. The used word stimuli were derived from the 50 item International Personality Item Pool questionnaire (IPIP) used for measuring the Big-Five factor markers as reported by Goldberg (1992). These words were translated into Dutch.

Experiment 2

Experiment 2 differed from Experiment 1 in one key aspect, namely the non-SRP condition.

Participants

30 native Dutch participants (18 female, age 22.4 ± 4.0) were included for the second experiment. Participants from the first experiment were excluded. Written consent was obtained from all participants.

Non-SRP condition

A potential objection to the non-SRP processing task in Experiment 1 is that it does not involve semantic processing of the presented word. For that reason, we repeated the experiment with a non-SRP condition in which participants answered the question "Does this object fit in a shoebox?". Another advantage of this task is that there is no confusion possible between conditions as different word stimuli are used for the SRP and non-SRP conditions. For the "shoebox task", we used translated nouns from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). 50 words were selected to which the answer was an unambiguous yes, and another 50 to which the answer was an unambiguous no.

Data analysis

The data of the experiments were analyzed using R (R Core Team, 2015). Participants with 5% of response times < 200ms were excluded from the analysis as well as participants with a mean neutral condition processing score over all trials < 85% as this would indicate that the subject was not performing the tasks. The response inconsistency for the SRP condition was also evaluated; if a participant was equally likely to respond yes or no over multiple repetitions of the same word, this would indicate that he is not performing the required task. Yes responses were scored 1 and no responses -1, their inconsistency for a given stimuli was then calculated as the variance over their responses. This results in a value of 0 when all responses are the same and 1 when there are as many yes as no responses. Participants with mean inconsistency over all word stimuli > 0.5 were excluded.

This resulted in two exclusions for experiment 1, one scored at chance in the neutral processing phase and one had

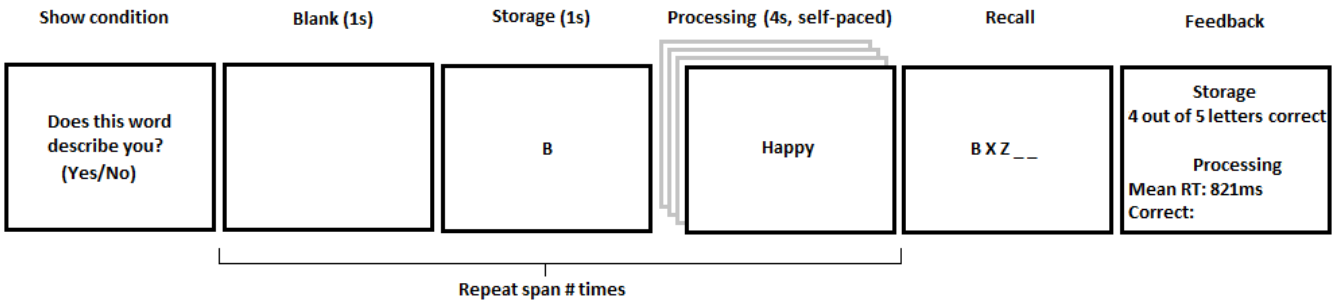


Figure 1: Overview of a single trial (SRP condition, span 5).

unrealistic low response times in the SRP condition. For experiment 2 two participants were excluded from the data analysis because they scored below 85% in the processing phase in the neutral condition.

Results

To analyze the effect of distraction by SRP we looked at the average recall score each participant attained per condition for each span. A difference in score between the SRP and natural condition indicates that one of type of distraction has a larger effect on recall than the other.

Figure 2 shows the mean score per condition per span, including 95% confidence interval bars. For all spans the average score for the SRP condition was lower than the neutral condition. To analyze this effect we performed a logistic mixed effects analysis in R (LME, Bates, Maechler, Bolker, & Walker, 2014) of the relation between condition and partial-credit score. The span and condition were entered as fixed effects and the intercepts for each participant as random effect. The reported p-values result from an ANOVA between the complete model and the model without the effect of condition. Using this approach we found that the score for the SRP condition was significantly lower than for the neutral condition ($\chi2(1) = 6.45$, $p = 0.011$). However, after including average response time for each trial to the previous model we no longer found a significant effect of condition on score ($\chi2(1) = 0.77$, $p = 0.38$). This means that the difference between the conditions can also be explained by a difference in difficulty of the processing task.

To examine whether the observed difference between the SRP and non-SRP was an artifact of the non-semantic nature of the non-SRP task, we conducted experiment 2, in which we replaced syntactic with semantic judgments. We again found that the score for the SRP condition was significantly lower than for the neutral condition ($\chi2$ (1) = 27.5, $p < 0.001$; figure 2). After including the average response time per trial in the model there was still a significant effect of condition on score ($\chi2$ (1) = 6.68, $p = 0.0097$). Therefore, in experiment 2 the difference in accuracy cannot be explained by the difficulty of the processing task.

Model

The results of experiment 2 have been modeled using the primitive elements model of skill (PRIMs; Taatgen, 2013), which has previously been successful in modeling visual distraction (Taatgen, Katidioti, Borst, & van Vugt, 2015). The key component of our distraction model is the idea that the current goals of the model activate mental operators to achieve that goal. However, the task-related operators have to compete with operators that are not activated by the goal, but by other influences. These influences can be external (e.g., a distracting visual stimulus), or internal (e.g., a task-relevant memory trace that has an unexpected association with a distracting train of thoughts). Experiment 2 was chosen as a target for modeling because it had the clearest results and didn't allow the participant to be confused between the SRP and neutral conditions (since each condition used a clearly different set of words).

Design and key elements

The model of this task requires several distinct components. It needs to store sequential information, which are the presented letters and their order. In addition to that it needs the ability to rehearse and the ability to report this information. Finally it needs to be able to process and respond to presented words, the exact mechanism of which will need to differ for the SRP and neutral condition. We will discuss these components in detail one by one.

The way we store sequential information is still unclear, there is evidence that we can chunk multiple items, treating them as one, and store positional information about each item within that chunk (Dehaene, Meyniel, Wacongne,

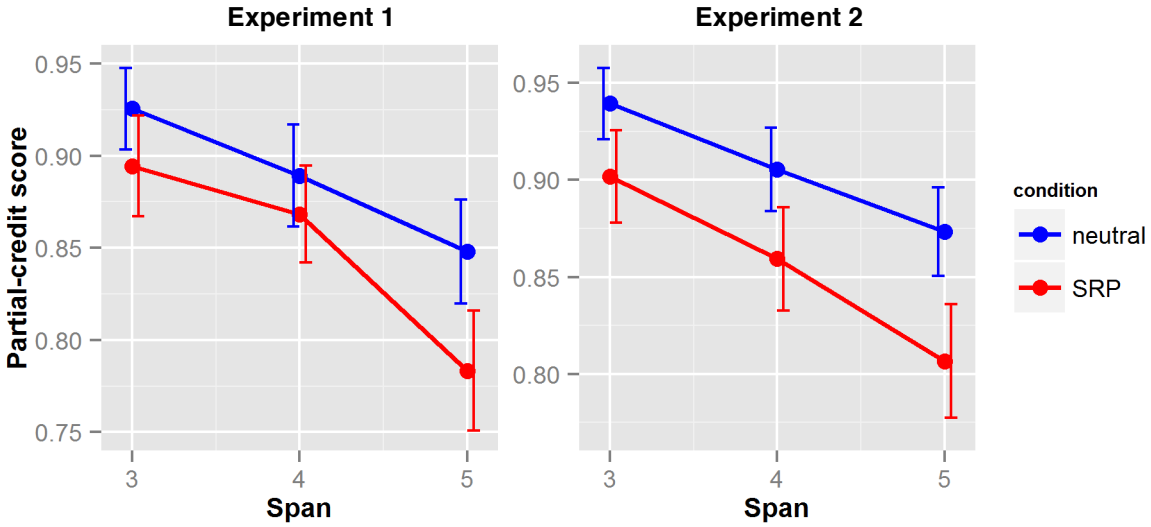


Figure 2: Recall accuracy as a function of span, comparing SRP and non-SRP (neutral) conditions. Data are shown for experiment 1 (left) and 2 (right), showing 95% confidence intervals.

Wang, & Pallier, 2015; Ladd & Woodworth, 1911). We modeled this by having position-specific operators. That is, there are separate operators for storing and retrieving each individual serial position of a letter. When storing, a chunk is created that contains positional information about an item, the item itself and a reference to the current goal chunk. The positional information and the reference to the current goal chunk is later used to try to retrieve the chunk by the position-specific operators. This reference is needed because otherwise chunks from previous trials will be recalled. If it is not successfully retrieved the model moves on to the next operator, which tries to retrieve the chunk containing the next position.

In the design of the experiment we included a one-second break between processing and the presentation of the next letter stimulus, which the participants could use to rehearse. We hypothesized that distraction caused by the processing in the SRP condition interfered with this rehearsal process. During this period, distraction competes with rehearsal. Because the processing phase ended if there was less than 700ms remaining for a to be presented stimulus, this time was added to the inter-trial interval, making this phase on average 1.3s in duration.

Distraction was modeled as follows. During the processing phase of the task, memory retrievals are required to answer either the shoebox question or the self-referential question. The assumption of the model is that self-referential words are associated with operators that elaborate these words. Elaboration involves additional memory retrievals, and representing the results in working memory, which in its turn can prime additional elaboration.

The probability that elaboration happens is much larger in the SRP condition, which is modeled by strength of association between the SRP words and an elaboration operator.

Once the blank period starts, there is a chance that a distracting fact remains in WM. This distracting fact spreads activation to operators to further think about that fact, which have to compete with the task-relevant operators to rehearse the letters. A successful distraction will therefore impede mental rehearsal, and lower the working-memory score. Once rehearsal wins the competition, the distracting fact is removed from WM and rehearsal continues normally. It is worth mentioning that this implementation is thus not strictly based on a competition between goals but between operators. This was done mainly because distraction periods are so short that it hard to justify it as being an active goal. The term 'goal competition' was used because this is more common in the literature.

The only difference between the models of the SRP and non-SRP conditions is that the words in the shoebox condition do not spread activation to the distraction operators, while the SRP words do.

In PRIMs the speed at which operators execute is increased over multiple runs. This means that at first the operators execute relatively slow which is unrealistic for our participants who already have substantial real life experience remembering, retrieving and rehearsing items in memory. Therefore, the model is trained for 15 trials before the experimental run starts. The complete set of models is available from the author on request.

Model results

The model was run 500 times and the results were compared to the results of experiment 2. Figure 3 shows that the partial score is matched quite nicely by the model; showing a clear difference between the two conditions. Figure 4 shows that the model has some trouble matching the recall score per serial position. The primacy effect matches reasonably well but the recency effect is too strong in the model.

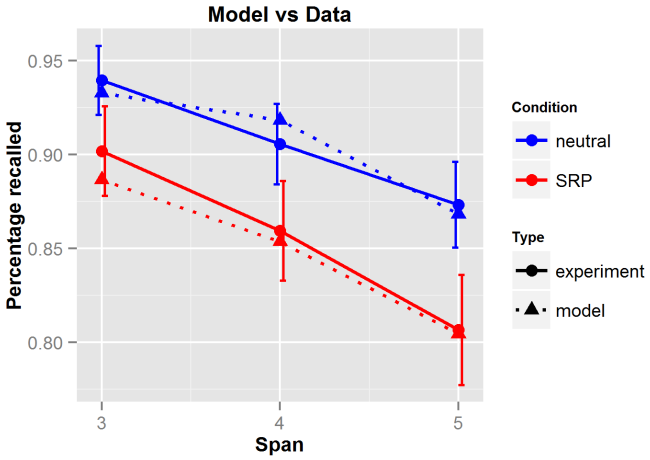


Figure 3: Comparison of partial-credit score between the experimental data (solid lines) and the model (dashed lines). 95% confidence interval bars are plotted for the experiment data only.

Effect of parameters

The model seems to do quite well in accounting for the observed difference in recall score between the two conditions. Various model parameters affect the model fits. Increasing the noise in memory tends to make the effect of span on score smaller, decreasing the slope between score and span. Decreasing the multiplication factor that scales memory retrieval time affects the difference between the two conditions since with a lower latency, the rehearsal process is faster allowing for more repetitions to be squeezed into the same amount of time. This benefits the neutral condition more than the SRP condition. Lowering the retrieval threshold increases the partial score for each span since with a lower threshold chunks are more easily recalled. This effect seems to be stronger for higher spans, most likely due to a ceiling effect at span 3.

Discussion

This study attempted to measure and model distraction caused by SRP on a memory task. To this extent two CWM span task experiments were conducted, comparing

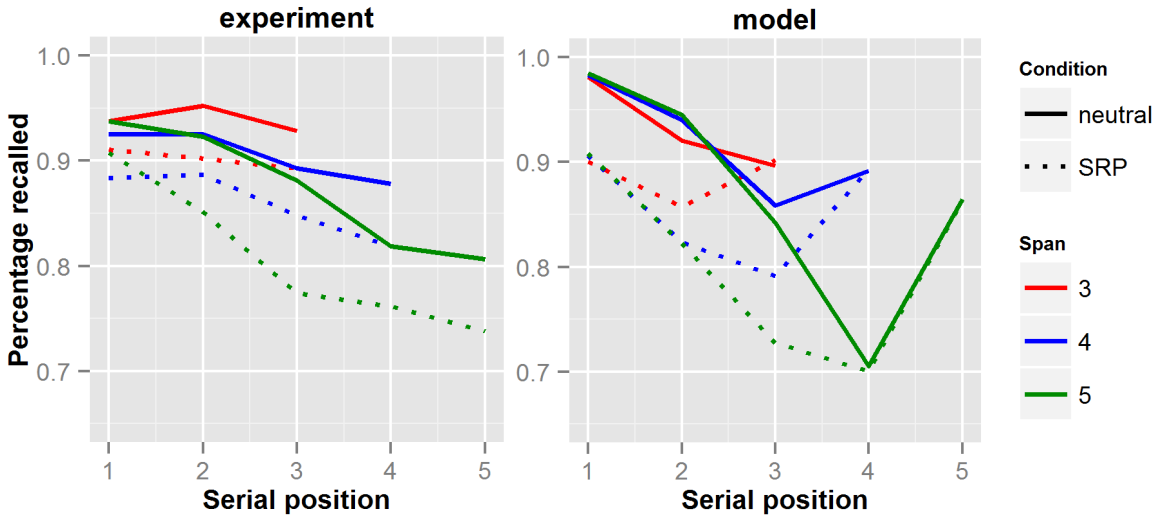


Figure 4: Percentage of successful recall by serial position for the experimental data (left column) and the model data (right column), separately for the SRP (dotted lines) and neutral (solid lines) conditions.

distraction by SRP (SRP condition) to distraction by other forms of processing (neutral condition) measured by the recall score on the CWM span task. The neutral processing consisted of letter recognition (experiment 1) and semantic processing (experiment 2). We found that in both experiments, SRP is associated with worse performance in the recall task than the other processing task. In experiment 2 we found that after including the average response time still a significant amount of variance in the recall score data could be explained by the condition factor. This was not so for experiment 1, most likely due to confusion between conditions by participants, see below. The experiment was modeled in PRIMs using a competitive goal approach in which distraction caused by SRP prevented letter rehearsal in the SRP condition. This model does quite well in accounting for the observed difference in recall score between the two conditions.

In the first experiment some participants reported that they sometimes forgot the condition associated with the current trial. This could happen due to the fact that the word stimuli for both conditions were the same and the condition was only indicated at the start of the trial. It is easy to detect when participants confuse the neutral with the SRP condition, as this means they would have scored at chance level. To see how often this might have happened we compared the number of neutral trials with <65% correct for both experiments. For experiment 1 this happened in 15 of the 900 trials (1.67%), and for experiment 2 this was 1 of 1044 trials (0.0958%). This indicates that indeed some confusion has occurred in experiment 1 but only in a small fraction of the trials. One can assume that these numbers are similar for the SRP condition (i.e. participants treating the SRP condition as the neutral condition), but this is hard to measure due to subjectivity of the responses in the SRP condition. Note that this doesn't negatively affect the main finding, if anything it only underestimates the true effects.

Further data analysis showed that the responses of the participants in the SRP condition could explain some of the level of distraction. The word pairs were labeled 1 and -1 for the positive and negative counterpart respectively. Responses were labeled 1 and -1 for yes and no respectively. Multiplying the responses by the word labels, averaging them per word, and then averaging this value for all words provides a basic indication about how positive a person thinks about himself, ranging from -1 (totally negative) to 1 (totally positive). This value was 0.6639 ± 0.224 for experiment 1 and 0.7102 ± 0.121 for experiment 2 (mean ± SD). Adding this 'positivity value' to the LME provided a significantly better fit for experiment 2 ($\chi 2$ (1) = 3.9773, p = 0.04612). Further exploring the nature of this effect—which was not found in experiment 1—is an interesting avenue for future research.

Our model does quite well in accounting for the effect of condition found in experiment 2, but it does have some limitations. Firstly, the model only accounts for retrieval errors, other types of mistakes such as transposition errors, item confusion or protrusion are not taken into account. Secondly, the model doesn't account for what is happening in the mind when it is distracted, in the current implementation the model just 'pauses' for a moment. This has the desired effect of preventing rehearsal but is not a plausible explanation of the mental processes happening during this time.

If the found decrease in recall performance is indeed caused by remaining emotions and thoughts after SRP, it gives rise to two interesting propositions. Firstly, this means that this effect could possibly be reduced by reducing SRP, for example, mindfulness training (Goldin, Ramel, & Gross, 2009). Secondly this means that performance in other types of secondary tasks (for example a processing task instead of recall) will also be affected by SRP.

In conclusion, we found that SRP has a negative effect on recall performance in a CWM span task. These findings can

be accounted for using a cognitive model made in PRIMs, in which SRP causes a distracting fact to enter WM, which subsequently interferes with rehearsal by activating competing-but task unrelated-operators. If our mechanism is correct, this implies that the disruptive effect of SRP should also extend to other types of secondary tasks and that the effect might be reduced by actively increasing participant's SRP through for example mindfulness. These results show that CWM span tasks can be used as an objective measure of distractions caused by self-referential thought, and that PRIMs can increase our understanding of the mental processes underlying them. Together these will allow us to better understand and prevent undesired effects of mental distraction during important tasks.

Acknowledgments

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen and Marie Curie grant ACCDECMEM to Marieke van Vugt.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R Package Version 1.1-7*, 0. doi:citeulike-article-id:1080437

Casner, S. M., & Schooler, J. W. (2014). Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(3), 433–442. doi:10.1177/0018720813501550

Conway, A. R. a, Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. doi:10.3758/BF03196772

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees. *Neuron*, *88*(1), 2–19. doi:10.1016/j.neuron.2015.09.019

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375–399. doi:10.3758/BF03203275

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*. doi:10.1037/1040-3590.4.1.26

Goldin, P., Ramel, W., & Gross, J. (2009). Mindfulness Meditation Training and Self-Referential Processing in Social Anxiety Disorder: Behavioral and Neural Effects. *Journal of Cognitive Psychotherapy*, *23*(3), 242–257. doi:10.1891/0889-8391.23.3.242

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage*, *31*(1), 440–457. doi:10.1016/j.neuroimage.2005.12.002

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. doi:10.1016/j.jneumeth.2006.11.017

R Core Team. (2015). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.* Retrieved from http://www.r-project.org/

Smallwood, J., & Schooler, J. W. (2015). The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology*, *66*, 487–518. doi:10.1146/annurev-psych-010814-015331

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–71. doi:10.1037/a0033138

Taatgen, N. A., Katidioti, I., Borst, J., & Vugt, M. Van. (2015). A Model of Distraction using new Architectural Mechanisms to Manage Multiple Goals. *ICCM Conference*, 264–269.

Woodworth, R. S., & Ladd, G. T. (1911). Elements of physiological psychology. Scribners: New York. Back to the Top.

Yanko, M. R., & Spalek, T. M. (2013). Route familiarity breeds inattention: A driving simulator study. *Accident Analysis & Prevention*, *57*, 80–86. doi:10.1016/j.aap.2013.04.003

Shakers and Maracas: Action-based Categorisation Choices in Triads Are Influenced by Task Instructions

Nicholas J. Shipp (n.j.shipp@herts.ac.uk)
Department of Psychology, University of Hertfordshire
Hatfield, AL10 9AB, UNITED KINGDOM

Frédéric Vallée-Tourangeau (f.vallee-tourangeau@kingston.ac.uk)
Department of Psychology, Kingston University,
Kingston upon Thames, KT1 2EE, UNITED KINGDOM

Susan H. Anthony (s.h.1.anthony@herts.ac.uk)
Department of Psychology, University of Hertfordshire
Hatfield, AL10 9AB, UNITED KINGDOM

Abstract

The forced-choice triad task has become increasingly popular in use over recent years. While it is seen as being a categorisation task (Lin & Murphy, 2001) variation in task instructions often leads to different results. Shipp, Vallée-Tourangeau, and Anthony (2014) used the triad task to show that when participants are asked to choose an option that 'goes best with the target', they are more likely to select the choice that shares an action relation when it also shares taxonomic information. However using the instruction to select the item that "goes best" is vague and might encourage a strategy other than a categorical decision. The present experiment used the same triads as in Shipp et al. to test whether participants would match items based on shared actions or shared taxonomic relations when given specific categorisation instructions. The task instructions were manipulated so that participants either selected the item that "goes best", "goes best to form a category" or is "most similar" to the target. The results found instances where the instructions of "goes best to form a category" led to a higher probability that participants would select the action choices over the instructions of "goes best". However when participants were encouraged to use similarity overall action choices were lower. Therefore the triad task does encourage a natural categorisation strategy and differences in task instructions across research are a result of the stimuli used.

Keywords: Action; Triads; Context; Instructions.

Introduction

Over years of research a variety of tasks have been used to measure the way in which objects are categorised including feature/exemplar listing tasks and sorting tasks. The forced-choice triad task has become a more popular tool in categorisation research over recent years and has been successfully used to show the influence of thematic relations in conceptual knowledge (Lin & Murphy, 2001; Mirman & Graziano, 2012; Kalénine & Bonthoux, 2008; Simmons & Estes, 2008).

Triad tasks are formed from a target word presented with one of two choice options sharing a particular relation to the target. The choice that participants select is seen to reflect how information is mentally organized. Lin and Murphy (2001) researched thematic choices in adults' conceptual knowledge where previously it was thought that only young children make use of thematic relations (Inhelder & Piaget, 1964; Kalénine & Bonthoux, 2008; Olver & Hornsby, 1967; Smiley & Brown, 1979). Lin and Murphy manipulated the choices where one shared a taxonomic relation to the target and one shared a thematic relation. For example the target of *bee* presented with *flies* (taxonomic relation) and *honey* (thematic relation). The results showed that when participants were asked to select the item that "goes best" with the target to form a category, the majority of choices were for the thematically related item (62%). In a second experiment Lin and Murphy used the same triad task adjusted the task instructions. When participants were instructed with "which two best form a category" the percentage of participants selecting the thematically related item decreased (49%).

Using the same triad task and "goes best with" instructions, Shipp, Vallée-Tourangeau, and Anthony (2014) identified conditions under which participants would draw upon action knowledge. Shipp et al. matched items together because they share the same physical action to operate them. For example *rifle* and *water pistol* share an action where both are operated with a 'trigger' action. Shipp et al. designed three sets of triads to measure how often participants selected the item sharing an action to the target. In the first set of triads, referred to as the Same Category Object (SCO) triads, the objects came from the same taxonomic category. However one of the options also shared an action with the target. These were designed to test for the additive effects that action might have in the same way that previous research had shown that shared themes have an additive effect when presented with taxonomic relations (Wisniewski & Bassok, 1999). An example of an SCO triad shows the target of *pencil* presented with *elastic band* (taxonomic) and *paintbrush* (taxonomic + action). In the second set of triads only one of the choice options shared a taxonomic choice where the other only shared an action. These Different

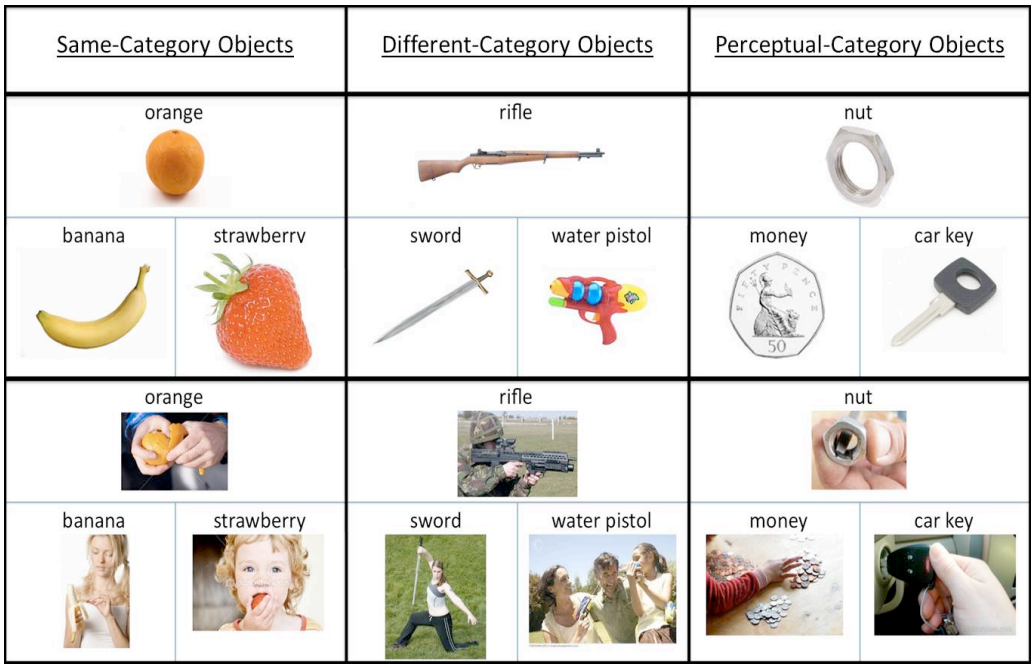


Figure 1. Examples of stimuli used in the experiment. From left to right: Same Category Object triad, Different Category Object triad, and Perceptual Category Object triad in the context-lean condition (top panels) and in the context-rich condition (bottom panels).

Category Object (DCO) triads were designed to test competition effects of action against taxonomic relations. For example the target of *rifle* was presented with *sword* (taxonomic only) and *water pistol* (action only). The triads were further manipulated by context whereby triad pictures were either shown to participants as three isolated objects against a white background (context-lean) or shown used by an agent for its functional purpose (context-rich). The results showed that participants were significantly more likely to select the action item on the SCO triads in either context (context-lean = 61%, context-rich = 70%). However the overall percentages of action choices on the DCO triads were low with a significant increase from the context-lean (32%) to the context-rich (53%) condition. This would therefore suggest that action is less likely to present as a basis for category membership on its own but has an additive effect when presented with taxonomic information. In addition a third set of triads were designed to assess the confounding variable that objects that share an action will invariably share perceptual properties. For example the *rifle* and *water pistol* look similar as they are designed around the trigger/handle component. Therefore it is possible that participants selected the action item because they looked similar rather than sharing an action. The Perceptual Category Object (PCO) triads were designed such that none of the three items shared a taxonomic relation, but one choice option shared perceptual properties to the target and the other choice shared an action to the target. If participants were selecting objects in the triads because of perceptual properties rather than shared action, then they would be less likely to select the action items on the PCO triads. The results deflected this possibility showing that in the context-rich condition participants were more likely to select the action choice (69%) showing the strong role that action plays in the triad task, particularly when presented within a functional context. In a related study, Tsagkaridis, Watson, Jax and Buxbaum (2014) used a triad task to show that participants are more likely to select thematically related items when they are used together for a purpose, such as *wine bottle* and *corkscrew*.

Task instructions are particularly important in the experiments reviewed thus far, but these are inconsistently formulated. Research has shown that variations in the task instructions lead to different choices selected on the triad task as shown by Lin and Murphy (2001) when participants were asked to select the item that "goes best" or "which two form a category". Simmons and Estes (2008) also showed different levels of thematic preference across different experiments when participants were instructed to select the item "most similar to" (46%), "most different to" (39%) or "most like to" (57%). Mirman and Graziano (2012) specifically used the instructions of "goes best" so as not to cause a taxonomic bias within the task. While the triad task is seen as a categorisation task (Estes, Golonka & Jones, 2011; Golonka & Estes, 2008; Lin & Murphy, 2001; Murphy, 2001; Simmons & Estes, 2008) the task appears to favour a thematic strategy unless participants are given more explicit category instructions. However such instructions are somewhat unclear in what they ask participants to do. It could be that participants rather than categorising the objects are selecting the item most similar to the target. While categorisation and similarity are seen as related processes where models of categorisation rely on similarity (Medin & Schaffer, 1978; Rosch & Mervis;

1975; Rosch, Simpson & Miller, 1976) there are dissociations between them (Goldstone, 1994; Iachini, Borghi & Sense, 2008; Rips, 1989; Smith & Sloman, 1994). A potential criticism of Shipp et al. is that the instruction of "goes best" could favour an action strategy, and participants are not engaging in a categorisation task. Therefore it is an empirical question as to whether the same pattern would be found in performance on the same triads when the instructions favoured a categorisation strategy. The aim of the present experiment was to test the effects of varying instructions on the action choices made in the triad task previously used by Shipp et al. (2014). The same triads from Shipp et al. were used and presented either in the context-lean or context-rich conditions previously used. However three sets of instructions were used where participants were asked either to select the item that "goes best", "goes best to form a category", or "is most similar to the target". If it is the case that participants on the triad task use a categorisation strategy then there should be little difference in action-based choices between "goes best to form a category" and "goes best". If however participants are completing the triad task using a similarity strategy then there should be little difference between "goes best" and "is most similar to". Several predictions can be made based on the findings of Shipp et al. as follows; it was predicted that (i) selection of the action-related item would be highest in the SCO triads and lowest in the DCO, replicating the 'additive' effect found previously, and (ii) the action choice percentages would be higher in the context-rich than context-lean conditions. Given that previous research views the triad task as a categorisation task (Estes et al., 2011; Lin & Murphy, 2001) it was predicted that no differences would be found between the "goes best" and "goes best to form a category" instructions in terms of the influence on action choices.

Method

Participants

Ninety undergraduate Psychology students (65 females, M_{age} = 21.19, SD = 6.12) took part in the experiment in return for course credit.

Materials

The same 30 triads (10 x SCO, 10 x DCO, 10 x PCO) triads used in Shipp et al. (2014) were used here presented within subjects to participants (see Fig. 1). The SCO triads consisted of a target with two choice options sharing a taxonomic relation to the target, but one also sharing an action with the target. The DCO triads consisted of a target with a choice option sharing a taxonomic relation only, and one sharing an action relation only. The PCO triads consisted of a target with one choice option sharing an action with the target, and one sharing perceptual properties. The triads were once again presented in a between subjects manner either in the context-lean or context-rich conditions. No differences were present in the program used except for

Table 1. *Full list of the Same Category Object (SCO) triads used in the experiment*

Target Item	Choice Items	
	Taxonomic Choice	Taxonomic and Action Choice
Pencil*	Elastic band	Paintbrush
Glass	Jug	Cup
Spatula	Grater	Saucepan
Pin**	Screw	Plug
Orange*	Banana	Strawberry
DVD player*	Television	CD player
Bed*	Wardrobe	Sofa
Leaflet	Poster	Newspaper
Spade*	Shears	Trowel
Ketchup*	Vinegar	Salt

Table 2. *Full list of the Different Category Object (DCO) triads used in the experiment*

Target Item	Choice Items	
	Taxonomic Choice	Action Choice
Fax machine*	Telephone	Photocopier
Screwdriver	Hammer	Key
Drink bottle**	Mug	Jam jar
Rifle**	Sword	Water pistol
Computer**	Printer	Piano
Calculator*	Set square	Mobile phone
Book	Ipod	Wallet
Paperclip*	Ruler	Clothes peg
Deodorant*	Hair gel	Insect repellent
Knife*	Ladle	Saw

Table 3. *Full list of the Perceptual Category Object (PCO) triads used in the experiment*

Target Item	Choice Items	
	Perceptual Choice	Action Choice
Axe	Cane	Tennis racket
Baseball bat*	Wrapping paper	Mace
USB pen*	Chewing gum	Phone charger
Clarinet	Wooden spoon	Balloon
Nut	Money	Car key
Present**	Storage box	Shoe
Cocktail shaker*	Vase	Maracas
Gun	Hairdryer	Cleaning spray
Peppermill**	Spray paint	Hair wax
Handbag*	Cheese grater	Cookie jar

Note. *Indicates those triads where choice selection significantly differed from chance with greater selection for the action choice. **Indicates those triads where choice selection significantly differed from chance with greater selection for the taxonomic (SCO/DCO) or perceptual choice (PCO).

the instructions given at the beginning.

Procedure

The procedure was the same as that used in Shipp et al. (2014). The triads were presented using Superlab on a 15" Macintosh laptop. The program began with a single practice triad followed by the 30 test items. A fixation cue was presented at the top of the screen on each trial for 1000ms. The fixation cue was replaced by the target word and picture (dependent on the condition the participant were assigned to). After 1500ms the two choice options appeared beneath the target alongside the appropriate images. Participants were either instructed to select the choice item that "goes best with the target", "goes best with the target to form a category" or "is most similar to the target". In the same manner as Lin and Murphy (2001) participants in the category instruction condition were presented with a definition in order to emphasise the categorical nature of the task. The instructions stated, "A category is defined as a set of things that share some commonalities - be it functions, purposes, physical and perceptual characteristics, or behavioural predispositions". Participants were instructed to press the 'a' key to choose the item on the left-hand side of the screen and the 'l' key for the item on the right-hand side of the screen. The action item was counterbalanced so that in half of the trials it appeared on the left side or right side of the screen. After they had made their choice the triad disappeared and the fixation cue appeared again for the next triad. Thus the design of the current experiment was a 3 (instructions) x 2 (context) x 3 (triads) mixed design with instructions and context as between subjects factors and triads as a repeated measures factor.

Results

The mean proportion of action responses was calculated for the SCO, DCO and PCO triads across context and instructions. As was found in Shipp et al., (2014) participants showed a tendency to select the action choice more with the SCO (66%) than with the PCO (57%) and DCO (54%) triads, and more so in the context-rich (67%) condition than context-lean (51%). In addition action choices were higher following the Goes Best to Form a Category (GBFC, 63%) instructions than in the Goes Best (GB, 58%) and the Most Similar (MS, 56%) instructions. Due to the dichotomous nature of the dependent variable the proportions of the choices made on the triads was initially analysed using single sample t-tests with a theoretical mean of 0.5. The results showed that in only nine (of thirty) triads did the choices not significantly differ from chance (see Tables 1, 2 and 3). In order to make the data fully comparable with the previous work of Shipp et al. (2014) these triads were left in for the remainder of the analysis.

A 3x2x3 mixed analysis of variance was conducted on the mean percentage of action choices across the triads. The analysis revealed a significant main effect of Context with a higher number of action choices in the context-rich condition, $F(1, 84) = 44.70$, $p < .001$, $\eta^2 = .35$. The main effect of Triads was also found to be significantly different, $F(2, 168) = 17.91$, $p < .001$, $\eta^2 = .18$. Post hoc analysis using the Bonferroni adjustment found that the action responses on the SCO triads were significantly higher than both the DCO triads ($p < .001$) and the PCO triads ($p < .001$). No difference was found between the DCO triads and the PCO triads ($p = .86$). The main effect of Instructions was not significant, $F(2, 84) = 2.70$, $p = .073$, $\eta^2 = .06$, nor was the two-way interaction between Context and Instructions, $F < 1$, The two-way interaction effect between Context and Triads was significant, $F(2, 168) = 11.47$, $p < .001$, $\eta^2 = .12$. In all three triads the mean percentage of choices was higher in the context-rich condition than in the context-lean condition, but the effect was stronger in the SCO ($p < .001$) and PCO triads ($p < .001$) than in the DCO triads ($p = .035$). The two-way interaction between Instructions and Triads was also significant; $F(4, 168) = 2.51$, $p = .044$, $\eta^2 = .06$. Post hoc analysis revealed that the only differences were found on the PCO triads where GBFC led to a higher mean proportion of action choices compared to GB ($p = .033$) and MS ($p < .001$). The difference between GB and MS was marginally significant ($p = .051$).

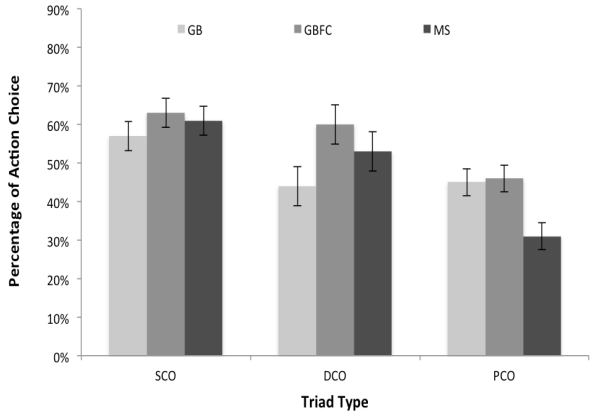


Figure 2. Mean percentage of action choices in the context-lean condition with the Same Category Object (SCO), Different Category Object (DCO), and Perceptual Category Object (PCO) triads between the Goes Best (GB), Goes Best to Form a Category (GBFC) and Most Similar (MS) instructions. Error bars are standard errors of the mean.

However of main interest here is the three-way interaction between Context, Instructions and Triads which was significant, $F(4, 168) = 5.01$, $p = .001$, $\eta^2 = .11$. Figures 2 and 3 show the mean proportion of action choices in each triad across the different instructions, spilt across the two contexts. Looking at the proportion of action choices in the context-lean condition (see Fig. 2) no differences were found between the instructions on the SCO triads. Post hoc analysis found that on the DCO triads the category instructions (GBFC) led to higher action choices than GB ($p = .028$). In addition on the PCO triads the similarity instructions (MS) led to

lower action choices than GB (p = .006) and GBFC (p = .003).

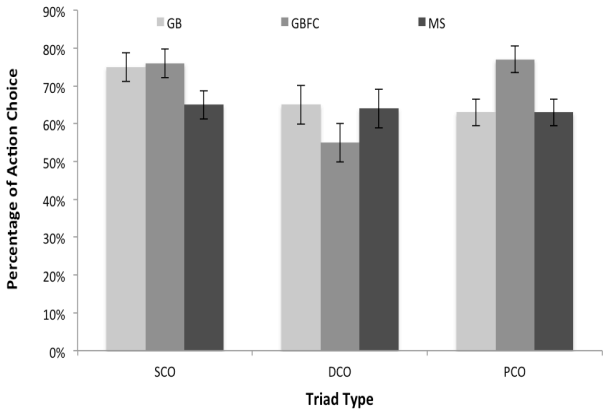


Figure 3. Mean percentage of action choices in the context-rich condition on the Same Category Object (SCO), Different Category Object (DCO), and Perceptual Category Object (PCO) triads between the Goes Best (GB), Goes Best to Form a Category (GBFC) and Most Similar (MS) instructions. Error bars are standard errors of the mean.

Examining the context-rich condition (see Fig. 3) no instruction related differences were found on the DCO triads. On the SCO triads the MS instructions led to significantly lower action choices than GB (p = .049) and GBFC (p = .0.36), but no difference was found between GB and GBFC (p = .90). In addition on the PCO triads the category instructions (GBFC) led to higher action choices than GB (p = .006) and MS (p = .006). The data here show that under no conditions did the GB instructions lead to statistically higher action choices than the GBFC instructions, and the similarity instructions led to proportionally lower action choices overall.

The overall mean proportion of action choices on the triads were recalculated without those triads which did not significantly differ from chance (see Tables 1, 2 and 3). The analysis was repeated and while the means across all of the triads increased, the exact same pattern was found overall with a significant three way interaction between Triads, Context and Instructions, $F(4, 168) = 3.11$, $p = .017$, $\eta^2 = .07$.

Discussion

The results reported here replicate Shipp et al. (2014) where participants were more likely to select the action choice when it also shared taxonomic information. This can be seen in the triads where the highest mean proportion of action choices was seen in the SCO triads compared to the DCO. In addition action choices were higher presented in a functional context. This effect was stronger for the SCO and PCO triads compared to the DCO triads.

However of main interest was the effect of task instructions. A potential criticism of the results reported in Shipp et al. (2014) was that whilst claims were made by those authors regarding the role of action in a categorization task, the instructions "goes best with" might have not have encouraged participants to see the task as one of categorization and, on the contrary, encouraged the use of a non-categorical strategy and more reliance on action. The significant three-way interaction reported here deflects this criticism. If it were the case that participants were not using a categorisation strategy then action choice preferences with GBFC instructions would be significantly lower than with the GB instructions. Overall the GB instructions did not lead to significantly higher action choice frequencies than the GBFC instructions. In contrast there are examples within the conditions where GBFC instructions actually led to higher action choice frequencies than GB instructions. Therefore rather than the more ambiguous GB instructions inflating action choice frequencies, it appears that these instructions are, if anything, reducing the probability of picking the action item. In addition, if it were the case on the triads that GB promoted a similarity strategy rather than a categorisation then the MS instructions would result in choice preferences more similar to those obtained with the GB instructions. However the results show that action frequencies were lower when participants were invited to select the most similar item. This further suggests that the triad task is a categorisation task and whilst it may draw on the participants understanding of similarity, performance does not rely solely on this.

An interesting finding here relates to the use of the GBFC instructions on the DCO triads. With these triads and instructions selection of the action choice was fairly high in both the context-lean (60%) and the context-rich (59%) conditions. Theoretically speaking it would be predicted that action choice frequencies should be fairly low with the DCO triads when participants are asked to group by category as the taxonomic item was designed to share category membership with the target, and the action choice was not. For example *rifle* and *sword* are both weapons and therefore when asked "goes best to form a category" participants should be more likely to select *sword* over *water pistol* as they are both members of the same category of 'weapons' (see Rosch, 1975). There are two possible explanations for the unexpected pattern. The first is due to the fact that taxonomic information contains not only information about category membership and functional information, but also perceptual information. If participants are directed to form a category between the target and choice item they should be influenced by perceptual information. Objects which are operated in a similar manner method such as *rifle* and *water pistol* will often share perceptual characteristics as they are designed to work within the ergonomic confines of the human body (i.e., designed around a handle and 'trigger' action). Therefore the action item, to some extent, shared perceptual characteristics with the target object. Further research using the same triad task where participants gave written protocols supports this explanation (Shipp, Vallée-Tourangeau, & Anthony, in prep). On the DCO triads participants often gave either an action *or* perceptual reason for matching the action-related item

to the target, particularly in the context-lean condition. Participants gave fewer perceptual reasons in the context-rich condition compared to the context-lean, possibly due to the reduced visual aspects of the objects themselves when they are held by an agent. For example, the handle of the rifle can no longer be seen in the context-rich condition because the hand of the agent blocks it. As such participants report selecting the action choice in the context-lean condition because of the shared perceptual aspects, but in the context-rich report selecting it because of the shared action between them.

The second possible explanation is that participants are creating goal-derived categories. For example some participants might be grouping *water pistol* with *rifle* because of the general goal of "things used for shooting" in preference to the functional category of weapon. This type of goal is highlighted by the context shown and therefore this, in combination with fewer obvious shared perceptual features, might increase the salience of such goal-derived categories. This would also explain why a high percentage of action choices were seen in the PCO triads where none of the items shared a standard category membership with the target. An example of this would be with *cocktail shaker* as the target and *maracas* as a choice option where participants might derive a sense of category membership based on the goal of "things that make a noise when shaken". The most likely option however is that both of these explanations work together when participants make their choices.

In conclusion, the results show previous concerns that action choices were inflated by the "goes best" instructions have been alleviated following the comparison of such instructions with choice preferences elicited with "goes best to form a category" instructions. However what is not clear here and needs further investigation is the type of category participants create on the fly when engaging with a triad tasks, whether these are categories that cohere in terms of their semantic or goal-derived features. There is yet more to be understood about performance on the forced-choice triad task.

References

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic Thinking: The Apprehension and Consequences of Thematic Relations. *Psychology of Learning and Motivation-Advances in Research and Theory*, *54*, 249-294.

Goldstone, R. L. (1994). The role of similarity in categorization – Providing a groundwork. *Cognition, 52,* 125-157

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1454.

Iachini, T., Borghi, A. M., & Senese, V. P. (2008). Categorization and sensorimotor interaction with objects. *Brain and Cognition*, *67*, 31-43.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child.* London: Routledge & Kegan Paul.

Kalénine, S., & Bontoux, F. (2008). Object manipulability affects children's and adult's conceptual processing. *Psychonomic Bulletin & Review, 15,* 667-672.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology-General, 130*, 3-28.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General, 141,* 601-609.

Olver, R. R., & Hornsby, J. R. (1967). On equivalence. In J. S. Bruner, R. R. Olver, & P. M. Greenfield (Eds), *Studies in cognitive growth* (pp. 68-85). New York: Wiley.

Rips, L. J. (1989). The psychology of knights and knaves. *Cognition, 31*, 85-116.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192-233.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human perception and performance*, *2*, 491-502.

Shipp, N. J., Vallée-Tourangeau, F., & Anthony, S. H. (2014). The context-dependent nature of action knowledge. In P. Bello, M. Guarani, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2925-2930). Austin, TX: Cognitive Science Society.

Shipp, N. J., Vallée-Tourangeau, F., & Anthony, S. H. (in prep). Strategies for making forced-choice categories: An item analysis of Shipp, Vallee-Tourangeau, and Anthony (2014).

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition, 108,* 781-795.

Smiley, S. S., & Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations – nonmonotonic relations trend from preschool to old age. *Journal of Experimental Child Psychology, 28*, 249-257.

Smith, E. E., & Sloman, S. A. (1994). Similarity-based versus rule-based categorization. *Memory & Cognition, 22*, 377-386.

Tsagkaridis, K., Watson, C. E., Jax, S. A., & Buxbaum, L. J. (2014). The role of action representations in thematic object relations. *Frontiers in Human Neuroscience*, *8*, 1-12.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology*, *39*, 208- 238.

Training Prospective Abilities through Conversation about the Extended Self

Nadia Chernyak (chernyak@bu.edu)[1] , Kathryn A. Leech (kleech@umd.edu)[2,3], Meredith L. Rowe (meredith_rowe@gse.harvard.edu)[3]

[1]Department of Psychological and Brain Sciences, Boston University, Boston, MA 02215
[2]Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD 20742
[3]Harvard Graduate School of Education, Cambridge, MA 02138

Abstract

Prospection is an important cognitive achievement, and is related to uniquely human abilities such as planning, delay of gratification, and goal attainment. While prospection develops rapidly during early childhood, little is known about the mechanisms that support its development. Here we explored whether encouraging children to talk about their extended selves (self outside the present context) boosts their prospective abilities. Preschoolers (N = 81) participated in a 5-minute interaction with an adult in which they were asked to talk about events in the near future, distant future, near past, or present. Compared with children discussing their present and distant future, children asked to discuss events in their near future or near past displayed better planning and prospective memory. Additionally, those two conditions were most effective in eliciting self-projection (use of personal pronouns). Results suggest that experience communicating about the close-in-time, extended self contributes to children's future-oriented thinking.

Keywords: prospection, future thinking, preschoolers, conversation, social context, extended self

Introduction

The ability to think about, plan for, and envision our future selves is an important cognitive achievement of early childhood. Prospection has been proposed to be a uniquely human ability and is critical for a variety of positive outcomes including goal-attainment and self-regulation (Atance & O'Neill 2001). Recent work on prospection has consistently found that prospective abilities develop rapidly during the preschool age (see Atance 2008) and continue developing through middle childhood (Lagattuta & Sayfan, 2011). While the structure and developmental timeline of young children's prospective abilities have received recent attention, relatively less is known about the mechanisms supporting their development. In our work, we explore the extent to which talking about one's extended-self improves children's prospective abilities.

Early precursors of prospection appear frequently in young children's verbal utterances and actions. Children begin to use future-oriented terms (e.g., "might", "may", "could") around the age of 2 (e.g., Bowerman, 1986; Atance & O'Neill, 2005) and are able to talk about the events of "tomorrow" by age 3 (Hayne, Gross, McNamee, Fitzgibbon, & Tammee, 2011). Yet it is not until later that children begin to act in service of their future selves: by the late preschool period, children show marked improvements in action-based measures of prospection such as delay of gratification (e.g., the ability to inhibit a desirable response in favor of a future reward; Mischel, Shoda, & Rodriguez, 1989), decreased temporal discounting (i.e., valuing future rewards over present rewards; Steinberg et al., 2009), planning (e.g., Atance & Meltzoff, 2005), and prospective memory (remembering to carry out intended plans at future time points; Gaujardo & Best, 2000). Recent work has investigated the extent to which such prospective abilities are associated (Atance & Jackson, 2009; Nigro, Brandimonte, Cicogna, & Cosenza, 2014; Neroni, Gamboz, & Brandimonte, 2014), underpinned by other cognitive competencies (e.g., language development, memory, or theory of mind; Hanson, Atance, & Paluck, 2014), or are linked to cognizing about the past (Coughlin, Lyons, & Ghetti, 2014; Cuevas, Rajan, Morasch, & Bell, 2015; Schacter, Addis, & Buckner, 2007).

One powerful predictor of children's social, cognitive, and linguistic abilities is their day-to-day social communicative context. For example, the quantity and quality of vocabulary input that parents provide to children predicts children's own vocabulary growth (see Hoff, 2006; Rowe, 2012); encouraging children and parents to talk about mental states predicts children's understanding of the mind (Lu, Su, & Wang, 2008; Reese, Sparks, & Leyva, 2010; Taumoepeau & Reese, 2013), and making even small changes in children's linguistic input has powerful effects on children's conceptual understanding (Rhodes, Leslie, & Tworek, 2012).

Such training studies are powerful in two respects: first, they are able to provide a basis for creating more formalized interventions targeting children's conceptual development. Second, they can help uncover the causal mechanisms underlying conceptual development. For example, in an important study, Rhodes and colleagues (2012) found that exposing children to generic talk in a short storybook task led to an increase in children's essentialist thinking, suggesting that generics and essentialism are causally related.

Inspired by the previous narrative work, we were interested in whether practice with projecting oneself into the future scaffolds children's prospective abilities. We designed a short training study in which we asked children to discuss and generate self-relevant future events. Prior theoretical work suggests that practice with simulating and anticipating future events helps motivate us to better prepare for those events (e.g., Taylor, Pham, Rivkin, & Armor, 1998). We reasoned that young children, who are still developing the ability to discuss their futures, and may therefore be unlikely to do so spontaneously, would be particularly likely to benefit from such intervention. On this account, simulating oneself in the future may help better

motivate young children to act in service of their future selves.

Our training study also allowed us to test several possibilities of how and why future-oriented talk might scaffold children's prospective abilities. One possibility is that simulating oneself in any context outside of the present helps children reason about themselves outside the here and now and make decisions on behalf of their extended selves (extended-self talk hypothesis). In support of this possibility, decontextualized talk (talk outside the here and now) in many forms (e.g., explanations, abstractions, future and past events) is shown to be a very powerful predictor of children's language and cognitive development (e.g., Rowe, 2012; Demir, Rowe, Heller, Levine & Goldin-Meadow, 2015). Yet another possibility, however, is that extended-self talk has to be restricted in content in order to scaffold prospective abilities (future-oriented talk hypothesis). Projecting oneself in the future specifically (rather than the past) might help anticipate future states, prepare for upcoming future events, or simply bring to mind one's future self. The concept of one's "future self" is taken out of an abstract, hypothetical state and brought to mind concretely through conversation and episodic mental simulation. Work with adults has shown that even brief reminders of one's future self specifically improves delay-of-gratification by helping adults feel closer to their future selves (see Hesrschfield, 2011). On this account, we would not expect any and all forms of extended-self talk to be similarly motivating, since talk about the past does not provide the benefit of anticipating upcoming future events. Finally, hybrid accounts are also possible: Because cognizing about the future and past are thought to rely on the same cognitive competencies (e.g., Schachter et al. 2007) discussing the extended self (in the future or past) might improve prospective abilities, but only in as much as the extended self is perceived as being relevant and concretely tied to one's present self (self-relevant extended-self talk hypothesis). In support of this possibility, Herschfield (2011) reports a link between adults' ability to delay gratification and how closely they believed future selves to be related to present selves. In the context of our work, this hypothesis predicts that discussing extended self events that are nearer in time to one's present self are more likely to feel self-relevant, would be particularly motivating for young children, and thus serve as salient reminders to act in service of one's future selves.

To distinguish among these different hypotheses, we designed a training study in which 3-5-year-old children were exposed to one of four different types of conversation about themselves. In one group (near future talk group), children were asked to generate events in their near future (within the next 24 hours) by being prompted to talk about "things that will happen later today." In a control group (present talk group), children were asked to talk about things in their present, contextualized context (e.g., "What do you see around you right now?"). In addition, we were interested in whether any future talk scaffolds children's abilities, or if future talk has to be temporally contiguous and closely related with children's present selves. We thus included a distant future talk group in which children were asked to discuss events that would occur after the next 24 hours (e.g., "things that will happen several weeks from now"). Finally, because cognizing about the future has been hypothesized to relate to the same cognitive processes as cognizing about the past we included a near past talk condition in which children generated events within the last 24 hours (e.g., "things that happened earlier today"). Immediately following training, children were tested on a broad range of prospective tasks.

Our primary question was whether the training groups would differ from one another on our prospective measures. In particular, we tested for three possible hypotheses using linear contrasts (weights explained below):

1. Extended-Self Hypothesis: Any conversations about the extended self, or self in the non-present (past, near future, or distant future), should boost prospective abilities. In this case, the near past, near future, and distant future conditions (weighted -1 each) should outperform children in the present condition (weighted +3) on prospective abilities.
2. Future-Oriented Hypothesis: Any conversation specifically about the future should boost prospective abilities. In this case, children in the near future and distant future conditions (weighted -1 each) should outperform children in the present or near past conditions (weighted +1 each).
3. Self-Relevant, Extended-Self Hypothesis: Conversation about the extended self that are close in time to the present self should boost prospective abilities. In this case, children in the near future and near past conditions (weighted -1 each) should outperform children in the present and distant future conditions (weighted +1 each).

Method

Participants

Participants were 81 three-to-five-year-olds (45 female; 36 male; Mean age = 4.40 years; Range = 3.0 – 5.71 years) recruited from six separate preschool centers. Demographics on individuals were not obtained, but two centers self-identified as serving low/lower-middle class communities, two served primarily upper-middle class communities, and two served mixed (both types of) communities.

Procedure

All children were tested in a separate room or quiet corner at their local preschool. One experimenter conducted the introduction and training phase, and a second experimenter, who remained blind to the training condition the children had just participated in, conducted the assessment phase. Sessions were videotaped for later coding and transcription. Sessions were coded by NC; a condition-blind research assistant then coded 25% of the data (reliability = 95%).

Timeline Introduction All children began by being introduced to the concept of linear time (e.g., Busby Grant & Suddendorf, 2009). Children were shown a rectangle divided into three colored squares signifying three distinct time periods ("before now", "now", and "after now"). The experimenter then placed the word "now" on the middle square and said, "this is everything that's happening right now" and then proceeded to list three examples of events in the present context (e.g., like us playing this game right here or your class playing outside or my friend (referring to the second experimenter) over there working"). Examples were modified slightly to fit the present context. The experimenter then asked the child to identify which square should signify "before now", and which square should signify "after now". Corrective feedback was provided until each child correctly identified "before now" and "after now", and the experimenter affixed the words "before" and "after" to their respective squares, such that the timeline showed "before now" "now" and "after now" in succession.

Training Period At this point, children were randomly assigned into one of four conditions (described below), in which they participated in a brief conversation with the experimenter about a specified time period.

1. Near Future Condition (n = 21): In the near future talk condition, children were told that they would be talking about events that are going to happen "after now". The experimenter then listed three examples of near future events (next 24 hours) in increasing temporal order: ("After now, are things like right after this game when you will go back to class, later today when you will go home from school, or even a really long time from now when will you go to bed tonight"). To encourage future self-projection, the experimenter then asked the child to draw a picture of him/herself in the last exampled future time period: "Can you draw a picture of yourself going to bed tonight?" After the child completed the drawing, she placed it on the square labeled "after now", and reaffirmed that it belongs on that square ("We're going to put this right here because this is going to happen after now!").

The child was then cued to generate events in his/her near future. The experimenter asked the child to list some events that would happen in three distinct time periods all taking place within the next 24 hours: (a) "right after" this game when the child goes back to his/her class (e.g., "What are some things you'll do right after this game, like when you go back to class?"), (b) "later today" when the child goes home from school, and (c) "a long time from now" when the child goes to bed tonight. The experimenter asked the question pertaining to each temporal cue and then encouraged the child to continually generate events (e.g., "and what are some other things you'll do later today?"). The experimenter proceeded to the next question/time period once the child had either: (i) repeatedly stated s/he could not generate further events; or (ii) generated five events.

2. Near Past Condition (n = 20): The near past talk condition proceeded exactly as the near future condition, except that the experimenter referred to events that happened in the preceding (rather than following) 24 hours. The experimenter pointed to the square labeled "before now" and stated she and the child would be discussing things that happened "before now". She then listed three examples of near past events (past 24 hours), which were matched to the near future events ("Before now, are things like right before this game when you were back in your class, earlier today when you first woke up, or even a really long time ago when you went to bed last night"). As in the near future condition, the experimenter then asked the child to draw a picture of him/herself going to bed last night and placed the drawing on the square titled "before now". The child was then asked to generate events during three time periods that took place within the past 24 hours: (a) "right before" this game when the child was in class, (b) "earlier today" when the child first woke up, and (c) "a long time ago" when the child went to bed last night.

3. Distant Future Condition (n = 20): The distant future talk condition proceeded in the same form as the near future talk condition, with the following modifications: First, the experimenter listed examples taking place after the proceeding 24 hours ("After now, are things like tomorrow when you will wake up in the morning, a few weeks from now when you will [celebrate Thanksgiving], or even a really long time from now when you are all grown up"). For the second example ("a few weeks from now"), we used a well-known upcoming holiday (e.g., Thanksgiving, Valentine's Day, Fourth of July), which varied depending on the day of the year that the child was tested. The experimenter then asked the child to draw a picture of him/herself when s/he is "all grown up" and placed the picture on the square labeled "after now". Finally, the experimenter asked the child to generate events in three distinct time periods: (a) "tomorrow" when the child first wakes up, (b) "a few weeks from now" when the child celebrates [an upcoming holiday], and (c) "a really long time from now" when the child is all grown up.

4. Present Condition (n = 20): The present condition was matched to the other three, except that children were told they would be talking about the square labeled "now". The experimenter then listed three examples of things in the child's present context, including something that the child could see around him/her ("things like what you see around you – like this game"), hear around him/her ("things like what you hear around you – like your class playing outside"), and feel around him/her ("things like what you feel around you – like this hard floor"). Examples were modified slightly to fit the context (e.g., the experimenter always used a prominent sound, such as children playing outside or teachers talking that could be easily heard by both herself and the child). The child was then asked to draw a picture of him/herself as s/he is "right now", and the picture was placed on the "now" square. The experimenter then asked the child to talk about the present context and

generate things that s/he (a) sees around him/her right now, (b) hears things around him/her, and (c) feels around him/her. As with the all other conditions, the experimenter gave the first prompt ("What are some things you see around you right now?"), and encouraged the child to generate examples. The experimenter proceeded to the next prompt once the child generated five examples or repeatedly stated s/he could not generate any further examples.

Assessment Following the training period, a new experimenter (blind to the child's training condition) assessed the child on a series of prospective tasks:

1. Prospective Memory Task: Following a procedure adapted from Guajardo and Best (2000), children were shown a wooden box and told there was a gift inside ("I have a gift for you in this box when we are all done with this game."). The experimenter then mentioned that the child had to remind her to open the box at the end of the game. To increase motivation, the experimenter then told the child that she often has trouble remembering things, and provided a cue that the child could use ("when I say 'we're all done', you have to remind me to open the box and give you your gift"). At the end of the game, the experimenter made sure to explicitly state the promised cue "We're all done!". If the child did not remind the experimenter within 10 seconds following the cue, she provided a second reminder "Did you have to remind me of anything?" and waited 10 more seconds. If the child still did not remember, she opened the box and retrieved the gift for the child.

Each child received a Prospective Memory Score between 0 and 2. Children were given a full score of 2 if they successfully remembered to tell the experimenter to open the box during the proper cue (after the experimenter said "we're all done!"), a score of 1 if they remembered after the second reminder ("Did you have to remind me of anything?"), and a score of 0 if they did not remember.

2. Mental Time Travel Task: We used two items adapted from Atance & Meltzoff's (2001) mental time travel task, in which children viewed a scene (e.g., a forest) and were told to imagine themselves planning to walk through it ("Let's pretend that you are going to walk across this road through the forest. Let's get ready to go!"). They were then shown three items – an item needed for a possible future state (e.g., water for drinking; correct response), an item that was semantically associated with the scene (e.g., a plant), and a distractor item (e.g., a present). The items were labeled and children were asked to provide an item selection ("Which of these things do you need to bring with you?") and a justification for their item selection ("And why do you need to bring the [chosen item]?").

Scoring. Children received a Mental Time Travel Correct Item Selected Score between 0-2 corresponding to the number of times children had selected the correct item across the two trials. In addition, children's explanations were coded according to whether they appropriately referred to a functional future use of that item (e.g., "I might get thirsty so I need to drink"; "the jacket because it's so cold outside"). Children received a Planning Explanation Score of 0-2.

3. Additional Tasks: We used three additional tasks testing temporal discounting and children's concept of linear time. These tasks did not show condition differences and are not discussed or further analyzed here.

Results

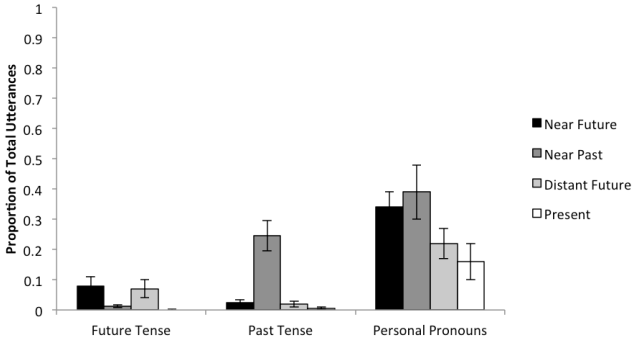
Talk Produced During Training

We looked at the types of events that children generated during the training session. Preliminary results revealed no effects of gender, age, or school center. We therefore collapsed across these variables in the following analyses.

The amount of utterances or the number of events that children generated did not vary across conditions (both p's $> .15$), confirming that all four of our training sessions were equated for the overall amount of talk children produced.

We then looked at the proportion of utterances employing the future and past tense (Figure 1). Note that for all events, children could use either the proper tense to which the time period referred (e.g., "I will sleep") or could answer without using the proper tense (e.g., "sleeping"). An ANOVA on the proportion of children's utterances containing the future tense revealed significant condition effects, $F(3,75) = 3.63$, $p = 0.02$, $\eta_{partial}^2 = 0.13$. In particular, children in the two future conditions (near future and distant future) produced a greater proportion of future-tense utterances than the near past and present: linear contrast $t(75) = 3.26$, $p = 0.002$. Similarly, an ANOVA on the proportion of past tense utterances revealed a significant effect of condition, $F(3,75) = 15.97$, $p < .0001$, $\eta_{partial}^2 = 0.39$. In particular, the near past condition differed significantly from the other three: linear contrast $t(75) = 6.90$, $p < .0001$. Therefore, while overall use of future tense was low (comprising less than 10% of the utterances children produced), our training successfully induced children to use it. Children produced future tense utterances in the two future conditions (near future, distant future), and past tense utterances in the one past condition (near past).

We also looked at the amount of self-projection involved when generating events. As a proxy for self-projection, we looked at children's use of personal pronouns (e.g., "I", "me") during the training session. Note that children could generate events either without the use of personal pronouns ("sleep") or with ("I'm gonna be sleeping"). In particular, we were interested in whether the two conditions in which children were asked to discuss their self-relevant (close-in-time) extended self might produce greater self-projection in comparison to their distally-temporal selves. An ANOVA on the proportion of utterances containing personal pronouns revealed significant differences across conditions, $F(3,75) = 2.79$, $p = 0.05$, $\eta_{partial}^2 = 0.10$. Planned linear contrasts showed that children in the near past and near future conditions used a greater proportion of personal pronouns in their utterances than those in the distant future or present condition, $t(75) = 2.75$, $p = 0.008$. Therefore,



talking about the close-in-time, extended self caused a greater amount of self-projection.

Figure 1: Proportion of (Bars Represent Standard Error) Future Tense, Past Tense, and Personal Pronouns Out of Total Utterances Across Conditions

Effects of Training

We then looked at the effect of our training on children's prospective abilities. Preliminary analyses revealed no effects of gender, so data was collapsed along this variable. We did, however, find, significant effects across ages and between preschool centers. We therefore control for age (as a covariate) and preschool center in all of our analyses. Figure 2 shows the analyses.

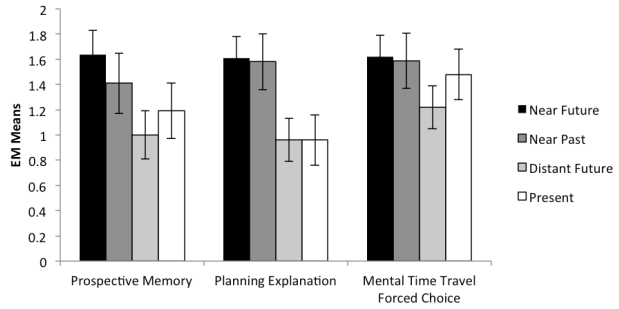


Figure 2: Estimated Marginal Means (Bars Represent Standard Errors) of Prospective Tasks Across Conditions

For Prospective Memory, there was a significant effect of Condition, $F(3,58) = 3.40$, $p = 0.02$, $\eta_{partial}^2 = 0.15$., School Center, $F(5,58) = 3.33$, $p = 0.01$, $\eta_{partial}^2 = 0.22$., and Age, $F(1, 58) = 4.01$, $p = 0.05$, $\eta_{partial}^2 = 0.07$. Planned linear contrasts supported the self-relevant extended-self hypothesis, $F(1,58) = 6.38$, $p = 0.01$.

Similarly, the Planning Explanation Score showed a significant effect of Condition, $F(3,59) = 2.91$, $p = 0.04$, $\eta_{partial}^2 = 0.13$, School Center, $F(5,59) = 3.99$, $p = 0.003$, $\eta_{partial}^2 = 0.25$, and Age, $F(1,59) = 4.61$, $p = 0.04$, $\eta_{partial}^2 = 0.07$. Planned linear contrasts once again supported the self-relevant, extended-self hypothesis, $F(1,59) = 16.62$, $p = 0.001$. There were no significant effects for the Mental Time Travel Forced Choice Scores (all p's > 0.05).

Discussion

Recent work in developmental psychology has taken an interest in the mechanisms that drive the development of young children's prospective abilities. Here we find that a short conversation about one's "extended-self" improved children's prospective memories and planning abilities. Our work suggests that experience communicating and thinking about children's extended selves may prime them to make decisions on behalf of their extended selves.

Across several measures, we also find support for our self-relevant, extended-self hypothesis: reminders of one's extended self are only useful when the extended self was temporally contiguous to the present self. Conversation about the temporally-contiguous extended self (near future and near past) improved planning abilities relative to conversation about the distally-related extended self (distant future). Moreover, children engaged in higher self-projection when discussing their close-in-time extended selves. Our results mimic prior work showing that adults' abilities to engage in saving for their future selves were predicted by how closely-related they believed their future selves were to their present selves (Herschfield, 2011). Moreover, brief visual reminders of one's extended self (i.e., age progressed portraits of one's future self) also helped improve saving behavior (Herschfield et al., 2011). Our work suggests that a similar mechanism may also account for children's prospective abilities – reminders of one's extended self may help activate concepts about the future self, or may make the extended-self appear closely-related to one's present self.

We note that our approach offers an important method for studying individual differences in how frequently children (and adults) conceive of their extended selves. We found that even a few directives from an adult could prime children to think about their future selves, but that the extent to which children engaged in true future-oriented thinking (i.e., used future tense) and engaged in self-projection varied across children and conditions. In particular, talking about close-in-time events caused greater use of personal pronouns. Such events may have been more readily recognizable to children as closely associated to their present selves, whereas temporally distant events (e.g., "adulthood") may have felt fundamentally distinct from and incompatible with children's present selves (Carey, 1985).

We provided several cues to help children understand, discuss, and visualize their extended selves: a linear time line, an adult's label of future time points (e.g., "later today"), the child's own self-generated discussion, and a pictoral label (that the child drew). The variety and number of these cues may have been particularly helpful in drawing children's attention to their future (or past) selves. Future work may also help to disambiguate the relative impact of each of these cues on children's prospective abilities.

Our work suggests that there is a strong role of social context in activating children's abilities to engage in future-oriented thinking and planning. Even brief conversations with adults help scaffold, shape, and active concepts about one's extended self. Training young children to to cognize, remember, and discuss their extended selves may ultimately

help them make future-oriented decisions that benefit those extended selves.

Acknowledgments

We would like to thank Julia Wang and So Yeon Shin for assistance with data collection and coding, participating preschools and directors for kindly lending space, and the Harvard Graduate School of Education for start-up funds to the third author to make the study possible.

References

Atance, C. M. (2008). Future thinking in young children. *Current Directions in Psychological Science*, *17*, 295-298.

Atance, C. M., & Jackson, L. K. (2009). The development and coherence of future-oriented behaviors during the preschool years. *Journal of Experimental Child Psychology*, *102*, 379-391.

Atance, C. M., & Meltzoff, A. N. (2005). My future self: Young children's ability to anticipate and explain future states. *Cognitive Development*, *20*, 341-361.

Atance, C. M., & O'Neill, D. K. (2001). Episodic future thinking. *Trends in Cognitive Sciences*, *5*, 533-539.

Atance, C. M., & O'Neill, D. K. (2005). Preschoolers' talk about future situations. *First Language, 25*, 5-18.

Bowerman, M. (1986). First steps in acquiring conditionals. In E. C. Traugott, A. ter Meulen, J. S. Reilly & C. A. Ferguson (Eds.), *On Conditionals* (pp. 285-307). Cambridge, UK: Cambridge University Press.

Busby Grant, J., & Suddendorf, T. (2009). Preschoolers begin to differentiate the times of events from throughout the lifespan. *European Journal of Developmental Psychology*, *6*, 746-762.

Carey, S. (1985). *Conceptual Change in Adulthood.* Boston, MA: MIT Press.

Coughlin, C., Lyons, K. E., & Ghetti, S. (2014). Remembering the past to envision the future in middle childhood: Developmental linkages between prospection and episodic memory. *Cognitive Development*, *30*, 96-110.

Cuevas, K., Rajan, V., Morasch, K. C., & Bell, M. A. (2015). Episodic memory and future thinking during early childhood: Linking the past and future. *Developmental Psychobiology*, *57,* 552-565.

Demir, Ö. E., Rowe, M. L., Heller, G., Goldin-Meadow, S., & Levine, S. C. (2015). Vocabulary, syntax, and narrative development in typically developing children and children with early unilateral brain injury: Early parental talk about the "there-and-then" matters. *Developmental Psychology*, *51*, 161-175.

Guajardo, N. R., & Best, D. L. (2000). Do preschoolers remember what to do? Incentive and external cues in prospective memory. *Cognitive Development*, *15*, 75-97.

Hanson, L. K., Atance, C. M., & Paluck, S. W. (2014). Is thinking about the future related to theory of mind and executive function? Not in preschoolers. *Journal of Experimental Child Psychology*, *128*, 120-137.

Hayne, H., Gross, J., McNamee, S., Fitzgibbon, O., & Tustin, K. (2011). Episodic memory and episodic foresight in 3-and 5-year-old children. *Cognitive Development*, *26*, 343-355.

Hershfield, H. E. (2011). Future self-continuity: how conceptions of the future self transform intertemporal choice. *Annals of the New York Academy of Sciences*, *1235*, 30-43.

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L. L., & Bailenson, J. N. (2011). Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research*, *48*, S23-S37.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55-88.

Lagattuta, K. H., & Sayfan, L. (2011). Developmental changes in children's understanding of future likelihood and uncertainty. *Cognitive Development*, *26*, 315-330.

Lu, H., Su, Y., & Wang, Q. (2008). Talking about others facilitates theory of mind in Chinese preschoolers. *Developmental Psychology*, *44*, 1726-1736.

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science*, *244*, 933-938.

Neroni, M. A., Gamboz, N., & Brandimonte, M. A. (2014). Does episodic future thinking improve prospective remembering? *Consciousness and Cognition*, *23*, 53-62.

Nigro, G., Brandimonte, M. A., Cicogna, P., & Cosenza, M. (2014). Episodic future thinking as a predictor of children's prospective memory. *Journal of Experimental Child Psychology*, *127*, 82-94.

Reese, E., Sparks, A., & Leyva, D. (2010). A review of parent interventions for preschool children's language and emergent literacy. *Journal of Early Childhood Literacy*, *10*, 97-117.

Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*, 13526-13531.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*, 1762-1774.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. *Nature Reviews Neuroscience*, *8*, 657-661.

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., & Banich, M. (2009). Age differences in future orientation and delay discounting. *Child Development*, *80*, 28-44.

Taylor, S. E., Pham, L. B., Rivkin, I. D., & Armor, D. A. (1998). Harnessing the imagination: mental simulation, self-regulation, and coping. *American Psychologist*, *53*, 429-439.

Taumoepeau, M., & Reese, E. (2013). Maternal reminiscing, elaborative talk, and children's theory of mind: An intervention study. *First Language*, *33*, 388-410.

How people differ in syllogistic reasoning

Sangeet Khemlani[1] and P.N. Johnson-Laird[2,3]
sangeet.khemlani@nrl.navy.mil, phil@princeton.edu
[1]Naval Research Laboratory, Washington, DC 20375 USA
[2]Princeton University, Princeton NJ 08540 USA
[3]New York University, New York, NY 10003, USA

Abstract

Psychologists have studied syllogistic inferences for more than a century, but no extant theory gives an adequate account of them. Reasoners appear to reason using different strategies. A complete account of syllogisms must therefore explain these strategies and the resulting differences from one individual to another in the patterns of conclusions that they draw. We propose a dual-process theory that solves these two problems. It is based on the manipulation of mental models, i.e., iconic simulations of possibilities. We also propose a new way in which to analyze individual differences, which depends on implementing a stochastic computer program. The program, *mReasoner*, generates an initial conclusion by building and scanning a mental model. It can vary four separate factors in the process: the *size* of a model, its *contents*, the propensity to consider *alternative models*, and the propensity to *revise its heuristic conclusions*. The former two parameters control intuitive processes and the latter two control deliberative processes. The theory accounts for individual differences in an early study on syllogisms (Johnson-Laird & Steedman, 1978). The computational model provides an algorithmic account of the different processes on which three subsets of participants relied (Simulation 1). It also simulates the performance of each individual participant in the study (Simulation 2). The theory and its implementation constitute the first robust account of individual differences in syllogistic reasoning.

Keywords: syllogisms, mental models, mReasoner, individual differences, deduction, counterexamples.

Introduction

It may be surprising to practitioners of the cognitive sciences that syllogistic reasoning is difficult to explain. After all, the first empirical study on syllogisms was carried out over a hundred years ago (Störring, 1908). Twelve separate theories now exist to explain reasoning by syllogism. And syllogisms are very simple inferences, such as:

(1) All of the architects are bankers.
Some of the bankers are not chefs.
What, if anything, follows?

The two premises each contain a single quantified term, such as, "all of the architects", and they can be in one of four separate *moods*, shown below (with their Scholastic abbreviations in parentheses):

All a are b. (*Aab*)	No a are b. (*Eab*)
Some a are b. (*Iab*)	Some a are not b. (*Oab*)

There are 64 possible pairs of syllogistic premises, depending on the moods of premises and the arrangement of the terms *a*, *b*, and *c* (i.e., the *figure* of a syllogism):

Figure 1	*Figure 2*	*Figure 3*	*Figure 4*
a – b	b – a	a – b	b – a
b – c	c – a	c – a	b – c

There are nine possible responses to (1), i.e., conclusions in four moods and two orders of end terms, *a* and *c*, and the response that no valid conclusion follows (hereafter, NVC). But, typically, reasoners do not consider all nine responses in their spontaneous conclusions; they generate just one or two. As a meta-analysis of six studies shows (Khemlani & Johnson-Laird, 2012), the most common response to (1) is that *Some of the architects are not chefs* – an error, since it is possible that all the artists are chefs, and so no definite conclusion follows validly.

In a typical study, responses to a syllogism vary from one individual to another. For instance, reasoners draw the erroneous conclusion to (1) about half of the time, but they also make a different error and conclude that *Some of the architects are chefs*. Only about a fifth of participants (university students) respond correctly that there is no valid conclusion, where a *valid* conclusion describes any conclusion that is true in all cases in which the premises are true (see Jeffrey, 1981). In logic, valid inferences can be drawn from any set of premises, including (1), from which it follows validly: *Possibly, all of the architects are chefs*. But, most experiments ask participant to draw definite conclusions, and it is rare for reasoners to infer spontaneous conclusions about possibilities (cf. Evans, Handley, Harper, & Johnson-Laird, 1999).

The psychologist's task is to explain the robust patterns of inference across this small, restricted set of 64 problems. The difficulty is that reasoners approach the problems with different abilities and appear to develop different strategies. Perhaps as a result, none of the twelve theories surveyed in the meta-analysis provides an adequate account of overall performance. The variability in reasoners' responses was enough to convince some theorists that the only way to understand how people reason syllogistically is to examine their individual differences (Stenning & Cox, 2006). To address the deficit, we developed a new theory – one that explains reasoners' most common responses as well as their individual idiosyncrasies for all 64 syllogisms.

In what follows, we review recent investigations into individual differences in syllogistic reasoning, and then describe a computational theory of syllogisms. Next, we report two analyses of the results from an early experiment, based on the computational theory. One analysis accounts for the variation in performance among three subsets of participants; and the other simulates the performance of the individual participants.

Individual differences in syllogistic reasoning

Several proposals describe individual differences in syllogistic reasoning. For example, some reasoners appear to be more proficient at the task than others (Galotti, Baron, & Sabini, 1986; see also Bucciarelli & Johnson-Laird, 1999). Bara et al. (1995) measured separately several factors, such as the ability to understand quantified assertions, but found that only one correlated with syllogistic performance, namely, working memory capacity. It accounts for a small amount of the variance in accuracy. Likewise, as Galotti et al. argued, good reasoners appear to consider more alternatives than do poor reasoners.

Ford (1995) argued that participants' descriptions and diagrams, and their justifications of their inferences, suggest that they adopt two different sorts of mental representation: diagrams and verbal representations (see also Bacon, Handley, & Newstead, 2003). She based her argument on participants' spontaneous use of a verbal strategy that substitutes the terms of the premises to yield a conclusion, and argued that some reasoners use the substitution strategy more than others. But, as Johnson-Laird and Bara (1984) suggest, a substitution-based strategy may be compatible with one representation.

One prominent attempt to model individual differences is due to Stenning and Cox (2006). They gave participants an ancillary "immediate inference" task (Newstead & Griggs, 1983) to ascertain whether they interpret assertions in systematically different ways. They proposed that some reasoners were more hesitant to draw valid conclusions, whereas others were more rash in drawing invalid conclusions. These patterns correlated with their subsequent syllogistic reasoning. The results were compelling enough for the authors to claim that analysis of aggregate performance is fundamentally "unjustified and misleading" (p. 1477). What the authors did not do, however, was to present an account of the syllogistic reasoning of the individual participants in their study.

Bucciarelli and Johnson-Laird (1999) also argued that aggregate analyses of reasoning often fail to capture meaningful differences in inferential behavior. In their studies, they discovered that participants seldom have a fixed interpretation for each syllogistic premise, and that they differ in the strategies they adopt. For example, they differ in which premise they interpret first and how they go about searching for alternative representations. But, these authors also did not present an analysis of the individual participants in their study.

In sum, no existing theory provided a compelling account of the computations that underlie the differences among individuals. Part of the difficulty may be a methodological problem: how does one analyze and assess the reliability of an explanation of the variation in performance of a complex skill such as syllogisms? With the exception of Polk and Newell (1995), psychologists had not hitherto made much progress towards an appropriate methodology. The first step is to frame a theory that is able to explain individual differences. In the next section, we describe such a theory.

A computational theory of syllogisms

mReasoner (Khemlani & Johnson-Laird, 2013) is a unified computational implementation of mental model theory, which posits that reasoning depends on the construction and manipulation of *mental models*, i.e., iconic simulations of possibilities (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird, 2006; Johnson-Laird, Khemlani, & Goodwin, 2015). The theory and its implementation are based on three fundamental principles:

- Mental models represent *possibilities*: a given assertion refers to a set of discrete possibilities that are observed or imagined (Johnson-Laird, 2006).
- The principle of *iconicity:* A mental models is *iconic* as far as possible in that its structure is isomorphic to the structure of what it represents (see Peirce, 1931-1958, Vol. 4). But, models can also include abstract symbols, e.g., the symbol for negation (Khemlani, Orenes, & Johnson-Laird, 2012).
- The principle of *dual processes:* reasoning, including syllogisms, is based on two interacting sets of processes: intuitions yield an initial conclusion by building and scanning a single model; and deliberations search for counterexamples to intuitive conclusions and, where possible, formulate alternative ones (Khemlani & Johnson-Laird, 2013; Khemlani et al., 2015).

The computational model makes syllogistic inferences by first constructing a small set of tokens that denote the entities referred to the premises. For example, mReasoner can build the following initial model for (1):

```
architect   banker
architect   banker
architect   banker   ¬chef
                      chef
```

where '¬' denotes the mental symbol for negation. The system then scans the model for an intuitive conclusion. It scans in the direction in which the model was built. In the model above, for instance, the system builds tokens for *architects* first, *bankers* second, and *chefs* third. Hence, the program draws an initial conclusion in the figure *a-c*, i.e., it concludes that *some of the architects are not chefs*. This conclusion matches the preponderance of conclusions that reasoners spontaneously generate. For other sorts of syllogisms, the system draws initial intuitive conclusions in the *c-a* figure, again depending on how the model was constructed.

Because the intuitive conclusion depends on just a single model, the system generates it quickly. But, as the example illustrates, the conclusion may be invalid. To correct the error, the program can call on a deliberative component to search for counterexamples to conclusions (Johnson-Laird, 2006). It operates by modifying the initial model using a finite set of search strategies (Bucciarelli & Johnson-Laird, 1999; Khemlani & Johnson-Laird, 2013). When the deliberative system is engaged, it can find a counterexample to the conclusion that *some of the architects are not chefs*:

```
architect   banker    chef
architect   banker    chef
architect   banker    chef
            banker   ¬chef
```

This model falsifies the intuitive conclusion, and so the program responds that no valid conclusion holds.

mReasoner's machinery for inference operates stochastically (see Khemlani et al., 2015). It builds models and searches for counterexamples based on four separate parameters:

1. The **λ parameter** controls the *size* of a mental model, i.e., the maximum number of entities it represents. It does so by basing the size on a sample drawn from a Poisson distribution of parameter λ. Hence, λ can be set to an approximation of a positive real number.
2. The **ε parameter** governs the model's contents, which are drawn from the most common set of possibilities corresponding to a particular assertion (the *canonical* set), or else the *complete* set of possibilities consistent with the assertion. For example, in the case of *All a are b*, reasoners tend to consider only one canonical possibility: *a's* that are *b's*. But the complete set of possibilities allows for *b's* that are not *a's*. The ε parameter sets the probability of drawing from the complete set. It ranges from [0, 1].
3. The **σ parameter** describes mReasoner's propensity to engage its deliberative component, i.e., its counterexample search mechanisms. It ranges from [0, 1].
4. The **ω parameter** is a nested parameter; it describes what happens when mReasoner finds a counterexample to its intuitive conclusion. When ω = 0, the system reports that no valid conclusion follows. When ω = 1, it *weakens* its initial conclusion and searches for counterexamples of the weakened conclusion, if possible. For example, it can transforms *All a are c* to a weaker claim, namely *Some a are c*. When ω is between 0 and 1, it is the probability of weakening the conclusion.

At present, mReasoner provides the closest fit to the results from the data presented in the meta-analysis on syllogistic reasoning (Khemlani & Johnson-Laird, 2012) compared to alternative theories. Figure 1 depicts the data from the meta-analysis across the 64 syllogisms, as well as mReasoner's closest-fitting simulation of the data, and shows a strong correlation between the two (r = .82). But, is the system capable of accounting for individual differences in syllogistic reasoning? In the following sections, we extend its analysis to different subsets of individuals (Simulation 1) and then to the idiosyncrasies of individuals' inferences (Simulation 2).

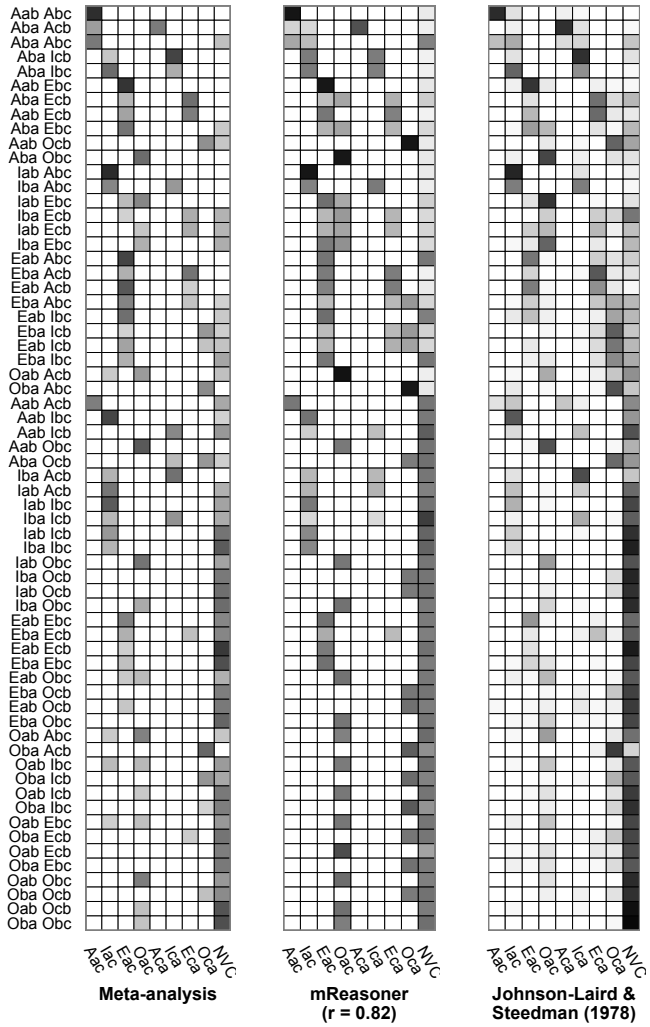


Figure 1. The percentages of responses to 64 syllogisms: a) in the meta-analysis in Khemlani and Johnson-Laird (2012); b) in mReasoner's best fit for the meta-analysis data (r = .82); and c) in the results of an experiment (Johnson-Laird & Steedman, 1978). *Aac* = All of the A are C, *Iac* = Some of the A are C, *Eac* = None of the A is a C, *Oac* = Some of the A are not C, and NVC = no valid conclusion. Each of the 64 pairs of premises occurs in a row, and each of the possible responses occurs in a column. The upper 27 rows denote syllogisms with a valid conclusion and the lower 37 denote NVC syllogisms. The grey scale in each cell indicates the proportion of corresponding conclusions (black = 100% and white = 16% or below). Hence, nearly 100% of participants in Johnson-Laird and Steedman (1978) responded that no valid conclusion (NVC) follows from the bottom-most syllogism, Oba Obc.

Simulation 1

We sought to analyze the data from an early experiment on syllogistic reasoning (Johnson-Laird & Steedman, 1978; see Figure 1) in order to model the participants' most frequent strategies. The participants were students at Columbia University (tested under the aegis of Janellen Huttenlocher), and they performed better than in any other study of all 64 problems. As Figure 1 shows, they made NVC responses more often than in the meta-analysis as a whole (45% vs. 30%, Wilcoxon test, z = 6.19, p < .001). Hence, the sample was biased towards higher performing, more deliberative reasoners.

We carried out an exploratory cluster analysis (Hartigan & Wong, 1979) to discover similarities in participants' patterns of reasoning. The data from the study were pooled over the 64 syllogisms and subjected to the *partitioning around medoids* (PAM) clustering algorithm, which is used to estimate the optimal number of distinct clusters in a given dataset (Kaufman & Rousseeuw, 1990). The analysis estimated that the optimal number of clusters in the data was 3. We used this estimate to constrain a hierarchical cluster analysis on the full range of participants' responses separated by the 64 syllogisms (see Hartigan,

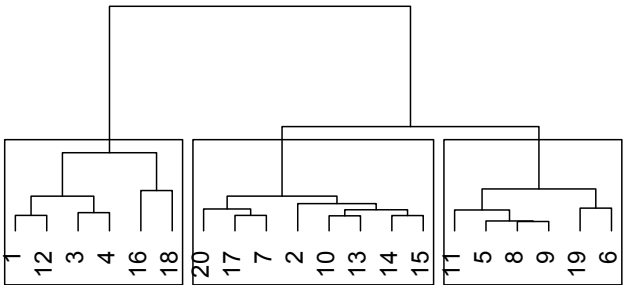


Figure 2. Dendrogram of a hierarchical cluster analysis performed on the 20 participants' propensity to yield 9 syllogistic reasoning responses pooling across 64 syllogisms from the data in Johnson-Laird and Steedman (1978). Each leaf in the tree reports a participant's unique identifying number. The analysis yielded three clusters of performance.

1975). Figure 2 shows how the cluster analysis grouped the 20 participants.

The three clusters suggest that there were systematic differences between subsets of individuals. To characterize them, we used mReasoner to simulate the three subsets by choosing appropriate parameter settings. If systematic differences exist, the simulations that best fit the data for each subset should differ, and the parameter settings of the best-fitting simulations should characterize the procedures on which the subsets relied.

Method and procedure

To simulate the three subsets of participants' performance derived from the cluster analysis, mReasoner generated simulated datasets for every possible combination of quantized settings of its four parameters. For each unique parameter setting, the system generated a dataset in which it carried out 64 syllogisms 100 times. The parameter settings were quantized to span their ranges as follows:

λ (size): *2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0*
ε (canonicality): *.0, .2, .4, .6, .8, 1.0*
σ (counterexample search): *.0, .2, .4, .6, .8, 1.0*
ω (weakening conclusions): *.0, .2, .4, .6, .8, 1.0*

Hence, the system generated 7 x 6 x 6 x 6 = 1512 separate simulated datasets. A grid search located the best fitting parameter setting for each of the three subsets of strategies.

Results and discussion

Figure 3 shows the data aggregated across the three subsets yielded by the cluster analysis, along with mReasoner's best-fitting simulations of those data, and Table 1 reports the parameter settings of the simulations. The parameter settings predict characteristics of the participants that comprise each subset.

As Figure 3 shows, reasoners' responses in Subset 1 appeared to vary more than in any other subset. The

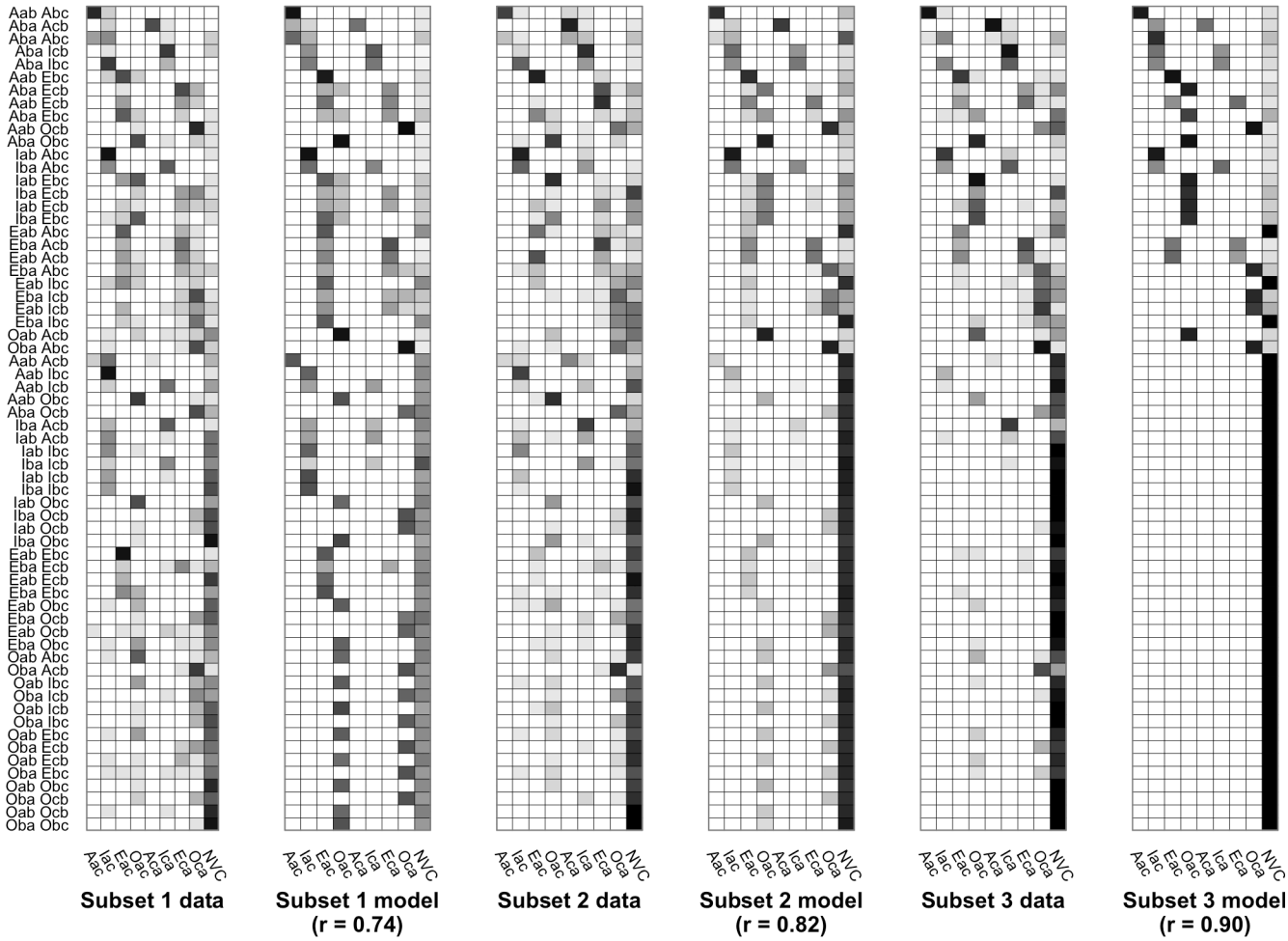


Figure 3. The percentages of responses to the syllogisms for three subsets of participants in Johnson-Laird and Steedman (1978), the corresponding mReasoner simulations, and their correlations with the data (see the text for an interpretation).

	Parameter settings and fit					
Subset	λ	ε	σ	ω	*r*	*Performance*
1	2.0	0.0	0.4	0.6	.74	Intuitive
2	3.0	0.6	0.8	0.6	.82	Intermediate
3	2.0	0.0	1.0	0.8	.90	Deliberative

Table 1. The parameter settings of mReasoner's best-fitting simulations for each of the three subsets of participants, the Pearson correlation of the simulated data with the actual data from the corresponding subset, and an assessment of each subset's performance based on an interpretation of the parameter settings.

parameter settings for Subset 1 explained this variability: reasoners appeared to build relatively small models, since the optimal parameter setting of λ, the size of the model, was 2.0: a model representing only 2 individuals. Since ε = 0, it meant that participants stuck to canonical possibilities, and were unable to explore the problem space fully. The relatively low value of the σ parameter (0.4) reinforces this interpretation; a value higher than 0.5 was likely to lead to a search for counterexamples that could have corrected errors. We accordingly characterize the participants in Subset 1 as *intuitive* reasoners. The term is not meant to impugn their intelligence, but rather reflects difficulty to consider alternative possibilities.

Subset 2 has an optimal setting for λ of 3.0, so the participants built larger initial models. They have a relatively high value of ε (0.6), and so they often considered non-canonical possibilities. Likewise, their propensity to search for counterexamples was high (σ = 0.8), which explains why they accurately inferred that the invalid problems had no valid conclusion more often. Hence, these individuals appear to be of *intermediate* ability.

Subset 3 has two parameter settings that were almost mirror images of those for Subset 1. They reflect a tendency to consider small, canonical models at the outset (λ = 2.0, ε = 0.0), but they always searched for counterexamples (σ = 1.0), and when they found one, they were likely to consider alternative weaker conclusions (ω = 0.8). We refer to the participants in this subset as *deliberative* reasoners.

The divergent sets of strategies yield systematic, behavioral predictions between the subsets of individuals, particularly between Subsets 2 and 3. For example, because the participants in Subset 2 consider more possibilities at the outset, they should provide correct responses faster than the members of Subset 2.

These simulations provide a computational explanation for why some reasoners are better than others. But, within these subsets, individuals are also likely to differ, and so our second simulation examined individual results.

Simulation 2

Simulation 2 used the same procedure and modeling technique as Simulation 1, but instead of comparing mReasoner's simulated datasets to reasoning performance at the subset level, we compared the predictions of the datasets to performance against the each of the 20 participants' data.

Method

For each of 1512 unique settings of the four parameters, the system generated a simulated dataset in which it carried out 64 syllogisms 100 times. An automated analysis discovered the parameter settings of the best-fitting simulations for each of the 20 participants.

Results and discussion

Table 2 provides the parameter settings and fit statistics of the best fitting simulations. The mean correlation between the best-fitting simulated datasets and participants' performance was .70, and ranged from .54 to .87. The mean of the worst-fitting simulated datasets and participants data was .25 (Wilcoxon test, $z = 3.9$, $p < .0001$) and the significant difference between the two sets of correlations suggests that the model's settings have drastic and qualitative effects on the patterns of inference it is capable of simulating. The correlations were lower than those for the analyses of subsets, and they reflect the reduction in power between subset and individual results. Reasoners' responses are much less systematic than when they are aggregated; nevertheless, all but four of the best-fitting simulations achieved a correlation of .60 or higher. Hence, while the variation in participants' responses was high, the computational model was capable of accounting for a significant proportion (about 50%) of their variance.

The optimal parameter values obtained from the analysis provide insight into participants' behaviors. For instance, if there exists significant concordance between the individual parameter values, then it would be because the participants are behaving similarly, i.e., no significant subsets or

	Parameter settings and fit statistics					
Participant	λ	ε	σ	ω	r_{BEST}	r_{WORST}
1	4.0	0.2	0.6	0.6	.69	.38
2	3.0	0.6	0.8	0.6	.69	.28
3	2.5	0.4	0.6	0.8	.59	.37
4	4.0	0.6	0.6	1.0	.57	.33
5	2.0	0.0	1.0	0.8	.86	.10
6	2.0	0.0	1.0	0.6	.81	.13
7	3.0	0.6	0.8	0.6	.63	.25
8	2.5	1.0	0.8	0.4	.82	.20
9	2.5	0.6	0.8	0.8	.80	.20
10	2.5	0.6	0.8	0.8	.72	.22
11	2.0	0.0	0.8	1.0	.80	.16
12	2.5	0.0	0.6	0.8	.65	.34
13	4.5	0.4	0.8	0.4	.74	.30
14	3.5	0.8	0.8	0.2	.75	.26
15	2.5	0.0	0.8	0.4	.65	.23
16	2.5	0.4	0.4	0.4	.58	.36
17	4.5	0.8	0.6	0.6	.65	.30
18	2.0	0.0	0.4	0.0	.54	.35
19	2.5	0.0	1.0	0.4	.87	.07
20	2.0	0.0	0.8	0.6	.62	.15

Table 2. The parameter settings of mReasoner's best-fitting simulations for each of the participants in the analysis, and the Pearson correlations of the best- and worst-fitting simulation with the data from the corresponding participant. The highlighted rows denote settings in which the correlation between mReasoner's best-fitting simulation and the data was < .60.

individual differences exist. In fact, the data show no significant concordance (Kendall's $W = .17$, $p = .89$) and is consistent with the cluster analysis reported earlier.

In sum, mReasoner successfully simulates the patterns of inference produced by individual reasoners.

General discussion

The mReasoner program is a computational implementation of the mental model theory of reasoning. It includes accounts of various sorts of deduction including syllogistic reasoning. It models in a stochastic way both intuitive and deliberative reasoning. In the past, investigators have lacked a systematic way in which to capture individual differences (though cf. Polk & Newell, 1995). In our view, the appropriate methodology consists in five main steps:

1. Obtain detailed data on the performance of a set of participants in carrying out the relevant task, such as their accuracy in drawing conclusion from the 64 different sorts of premise.
2. Develop a theory of their performance, and implement a computer program with stochastic parameters for the key factors that should determine performance, such as whether or not a search is made for alternative models.
3. Carry out a cluster analysis on the participants' data in order to determine whether there are systematic differences among different subsets of participants in their responses to the task.
4. Simulation 1: carry out an automated search through quantized settings of the parameters in order to determine whether it is possible for the settings to account for the differences amongst the subsets that step 3 revealed.
5. Simulation 2: carry out a similar search in order to account for the data from the individual participants.

The mReasoner system provided a close match to aggregated data from six syllogistic reasoning studies. We therefore carried out the five-step procedure described above in an analysis of the experimental results from one study of 20 participants. The cluster analysis yielded three main subsets of participants, and Simulation 1 showed that the theory could characterize them: intuitive reasoners who maintain small mental models and tend not to search for alternative models; intermediate reasoners who build larger, more varied models but do not search for counterexamples; and deliberative reasoners, who actively engage in a search for counterexamples, and weaken their conclusions in the light of them. The analysis of participants' individual reasoning patterns likewise showed a close match between the simulation and their data.

How might the theory be improved? It fails to capture certain systematic patterns in participants' responses; the theory makes the wrong prediction for syllogisms such as, e.g., Oba Acb. Hence, a more accurate theory needs to explain these and some other aberrant results. Of course, it remains a tentative achievement to fit data from just one domain of reasoning. The present analyses show promise that the computational model we describe is flexible enough to account for individual differences in syllogisms; a true test of its power is one in which it can model differences across a broad swath of reasoning domains.

Acknowledgements

This research was supported by a grant from the Office of Naval Research awarded to SK and by NSF Grant No. SES 0844851 to PJL. We are grateful to Ruth Byrne, Monica Bucciarelli, Geoff Goodwin, Tony Harrison, Laura Hiatt, Robert Mackiewicz, Marco Ragni, and Greg Trafton. Special thanks goes to Thad Polk for having kept the raw data from Johnson-Laird and Steedman (1978) and providing us with them.

References

Bacon, A., Handley, S., & Newstead, S. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking & Reasoning, 9,* 133-168.

Bara, B., Bucciarelli, M., & Johnson-Laird, P.N. (1995). The development of syllogistic reasoning. *American Journal of Psychology*, *108*, 157-193.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23,* 247–303.

Evans, J.St. B. T., Handley, S.J., Harper, C.N.J., & Johnson-Laird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1495-1513.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54,* 1–71.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General, 115,* 16–25.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Johnson-Laird, P. N. (2006). *How we reason.* Oxford, England: Oxford University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984b). Syllogistic inference. *Cognition, 16,* 1–61.

Johnson-Laird, P. N., Khemlani, S., & Goodwin, G.P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19.

Johnson-Laird, P. N., & Steedman, M. J. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64–99.

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding groups in data.* Wiley & Sons: New York, NY.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138, 427-457.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, 4, 1-20.

Khemlani, S., Lotstein, M., Trafton, J.G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *Quarterly Journal of Experimental Psychology*, 68, 2073–2096.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology, 24,* 541–559.

Newstead, S. E., & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and V erbal Behavior , 22,* 535–546.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.

Stenning, K., & Cox, R. (2006). Reconnecting interpretation to reasoning through individual differences. *Quarterly Journal of Experimental Psychology, 59,* 1454–1483.

Störring, G. (1908). Experimentelle Untersuchungen u¨ber einfache Schlussprozesse [Experimental investigations of simple inference processes]. *Archiv fu¨r die gesamte Psychologie, 11,* 1–27.

Grounded Distributional Semantics for Abstract Words

Katsumi Takano (katsumi@utm.inf.uec.ac.jp)
Akira Utsumi (utsumi@uec.ac.jp)

Department of Informatics, The University of Electro-Communications
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

Since Harnad (1990) pointed out the symbol grounding problem, cognitive science research has demonstrated that grounding in perceptual or sensorimotor experience is crucial to language. Recent embodied cognition theories have argued that language is more important for grounding abstract than concrete words; abstract words are grounded via language. Distributional semantics has recently addressed the embodied nature of language and proposed multimodal semantic models. However, these models are not cognitively plausible because they do not address the recent embodiment view of abstract concepts. Therefore, we propose a novel multimodal distributional semantics in which abstract words are represented indirectly through grounded representations of their semantically related concrete words. A simulation experiment demonstrated that the proposed model achieved better performance in computing the word similarity than other multimodal or text-based distributional models. This finding suggests that the indirect embodiment view is plausible and contributes to the improvement of multimodal distributional semantics.

Introduction

Distributional semantics is a computational approach to constructing semantic representations of words by observing distributional statistics of word occurrence in large collections of text. The lexical meaning of a word is represented by a high-dimensional vector in a semantic space, and the degree of semantic relatedness between any two words can be easily computed from their vectors, for example as the cosine of two vectors. Because of its simplicity and versatility, distributional semantics has recently received enormous attention, and a variety of methods have been proposed and refined (Turney & Pantel, 2010; Jones, Willits, & Dennis, 2015). Recent advances of neural network language models or word embedding models (Mikolov, Chen, Corrado, & Dean, 2013) are also regarded as notable approaches of distributional semantics.

However, distributional semantics has been criticized as psychologically implausible because it is ungrounded or disembodied (Glenberg & Robertson, 2000). Embodied cognition theory argues that language is grounded in our sensory perception and experience, but distributional semantic models learn from only linguistic information and do not use sensorimotor information. Although it is controversial whether distributional semantics cannot essentially simulate human semantic memory, it is no doubt that it cannot represent the meaning of some kinds of words, in particular concrete words, just as they are represented in human semantic memory.

For this reason, recent studies of distributional semantics have addressed the use of sensorimotor or perceptual information to learn psychologically plausible lexical representation. For example, Andrews, Vigliocco, and Vinson (2009) used perceptual features collected as featural norm to ground a language-based topic model on perceptual experience, and demonstrated that the integrated model outperformed the language-based topic model. Although the featural norm used in their study is represented by language, recent attempts have been directed toward multimodal distributional semantics (for a review, see Baroni, 2016), in which textual information is integrated with multimodal information computed directly from nonlinguistic inputs such as visual (Bruni, Tran, & Baroni, 2014; Kiela, Hill, Korhonen, & Clark, 2014), auditory (Kiela & Clark, 2015), or olfactory (Kiela, Bulat, & Clark, 2015) ones.

Their primary research purpose is to improve the accuracy of distributional semantic models by using multimodal information and it has often been demonstrated that multimodal information can improve the performance of language-based distributional semantics. However, this is not always the case. Kiela et al. (2014) showed that, although the inclusion of visual input improves the representations of concrete words, it degrades the representations of abstract words. This finding is highly important, because it seems to suggest that abstract words may not be embodied in the same way as concrete words are embodied. It has also been pointed out by recent development of the embodied view of language (e.g., Louwerse, 2011; Borghi & Cangelosi, 2014; Thill, Padó, & Ziemke, 2014). These views argue that abstract words are embodied differently from concrete words, and acquisition of abstract words depends more heavily on linguistic experience. To provide a psychologically more plausible model, distributional semantics must deal with the problem of abstract words, and at the same time, such a theory can provide an effective computational framework for testing the validity of the embodiment theory of language.

In this paper, therefore, we propose a novel distributional semantic model by taking into account the recent embodied view of abstract concepts. We focus on visual images as a source of perceptual information, as used in many other studies on multimodal distributional semantics, and propose a mechanism of embodiment of abstract concepts via language, in which an embodied vector representation of an abstract word is obtained from the embodied representations of concrete words that are semantically related to that abstract word. We then test the psychological plausibility of the proposed method using human similarity ratings of English word pairs.

Embodiment of Abstract Words

Abstract concepts pose a serious challenge to the embodied theory of language. As Dove (2015) pointed out, it is difficult to see how representations grounded in perceptual experience can capture the content of abstract concepts or words such as

truth and *justice*. Even if abstract concepts somewhat depend on perceptual experience, the way in which such abstract concepts are grounded seems to differ from concrete concepts. This intuition has been supported by a number of empirical findings obtained in cognitive neuroscience. Neuroimaging studies have demonstrated that processing of abstract concepts elicits greater activation of the left perisylvian language network (such as the left inferior frontal gyrus and the left superior temporal cortex) as compared to processing of concrete concepts (e.g., Binder, Desai, Graves, & Conant, 2009; Wang, Conder, Blitzer, & Shinkareva, 2010). Vigliocco et al. (2014) also found greater engagement of the rostral anterior cingulate cortex, an area associated with emotional processing, for abstract than concepts words.

Some embodied theories claim a general mechanism of embodiment common to both concrete and abstract concepts. For example, Barsalou's (1999) theory of perceptual symbol systems argues that concepts (and meanings as well) are not amodal, formal symbols, but rather inherently modal, perceptual symbols, which refer to perceptual states (or neural representations) in sensorimotor systems. These perceptual symbols become integrated into a simulator for a concept that enables us to simulate sensorimotor experience of perceiving and/or acting on exemplars of that concept. According to Barsalou (1999), abstract concepts are represented in the same perceptual symbol systems, although the simulator of abstract concepts is constructed from not only sensorimotor but also introspective states or events. Barsalou (1999) explains this by using examples such as *truth* and *negation*, but the explanation is relatively vague and limited to some logical or affective concepts.

Recently, a number of studies have attempted to resolve the problem of abstract concepts while maintaining the embodied view by emphasizing the role of language in processing of abstract concepts. For example, Word-as-Tools (WAT) theory (Borghi & Cangelosi, 2014) claims that words function as tools to extend our cognition and such the linguistic function is more crucial for abstract than for concrete words representation. Dove (2014) argues for a similar view that language provides an important means of extending our cognitive or mental capabilities by enabling access to an embodied representational system that exists independently of language. The basic idea underlying these views is that abstract words are grounded in sensorimotor or perceptual experiences, but the embodiment is indirect, rather than direct in the case of concrete words. This "indirect embodiment" view is proposed more clearly by Thill et al. (2014). They propose a "division of labor" approach between a perceptual layer that associates basic, concrete concepts with perceptual features and a relational layer that grounds more complex and abstract concepts in relation to basic concepts. Gleitman et al.'s (2005) explanation of how "hard words" are acquired is closely related to the indirect embodiment view. In the early stage of lexical acquisition, the meaning of concrete words is acquired directly from perceptual information via word-to-world pairing. In the later stage, the meaning of hard words, which is not easily accessible through perception, is acquired by a structure-to-world mapping procedure that combines linguistic observations with cooccurring perceptual experience. From the distributional semantics side, Louwerse (2011) also put forward the indirect embodiment view, namely the symbol interdependency hypothesis, to refute a criticism against distributional semantics made by the embodied view. According to the symbol interdependency hypothesis, language comprehension is symbolic through interdependencies of amodal linguistic symbols, while it is indirectly embodied through the references linguistic symbols make to perceptual representations. Hence, "language has evolved to become a communicative short-cut for language users and encodes relations in the world, including embodied relations (*ibid.*, p.279)."

However, despite receiving much attention from cognitive science over the recent years, the indirect embodiment view of abstract words has not been addressed by computational studies on distributional semantics, as pointed out in the introduction. Although multimodal distributional semantics incorporates the embodied view of language into an algorithm for constructing word vectors, it cannot deal with the problem of abstract words. Kiela et al.'s (2014) approach clearly shows this difficulty; to improve the accuracy of multimodal word vectors, they excluded visual information when constructing vectors of abstract words, while including visual information for concrete words. For distributional semantics to be a psychologically more plausible model, it is highly necessary to integrate the indirect embodiment view into multimodal distributional semantics. This is what we aim to accomplish in this paper.

Grounded Distributional Semantics

Our grounded distributional semantic model, whose algorithm is shown in Figure 1, is based on the indirect embodiment view that grounding of abstract words is mediated by language. Therefore, different grounding methods are used for concrete and abstract words. For a concrete word, its vector representation is constructed by combining a text-based vector (computed from a text corpus) with a visual vector computed from images for that word (i.e., Steps 3(a) and 4 of Figure 1). This method does not differ from Kiela et al.'s (2014) and other standard methods for multimodal distributional semantics. On the other hand, the vector representation of an abstract word is constructed by combining a text-based vector with visual vectors of concrete words semantically related to that abstract word (i.e., Steps 3(b) and 4).

An important part of this algorithm is the classification between concrete and abstract words at Step 2. In this paper, we use the image dispersion method proposed by Kiela et al. (2014). This method is motivated by the intuition that the diversity of images retrieved for a particular word depends on its concreteness. It is generally expected that images for concrete concepts such as "*apple*" are more similar to one another than those for abstract concepts such as "*happiness*." The degree of image dispersion for a concept is defined as the average pairwise cosine similarity between all the vector representations in the set of images for that concept. All words are sorted in descending order of image dispersion, and top

1. Construct a linguistic semantic space DSM_L from a text corpus.
2. Divide the vocabulary V into concrete words V_C and abstract words V_A (i.e., $V = V_C \cup V_A$ and $V_C \cap V_A = \phi$).
3. Construct a visual semantic space DSM_V using visual images.
   (a) For a concrete word $w_i^C \in V_C$, compute a visual vector $\boldsymbol{y}(w_i^C)$ directly from the images labeled with w_i^C.
   (b) For an abstract word $w_i^A \in V_A$, compute a visual vector $\boldsymbol{y}(w_i^A)$ as $\sum_{w_j^C \in SN(w_i^A)} \boldsymbol{y}(w_j^C)$ where $SN(w_i^A) \subset V_C$ is a set of n semantic neighbors of (i.e., n concrete words semantically related to) w_i^A. Semantic neighbors of w_i^A are computed in the linguistic semantic space DSM_L.
4. Construct a grounded semantic space DSM_G by combining DSM_L and DSM_V as follows. For each word $w_i \in V$, normalize $\boldsymbol{x}(w_i) \in DSM_L$ and $\boldsymbol{y}(w_i) \in DSM_V$, and concatenate them.

Figure 1: Overall algorithm of the proposed distributional semantic model.

$100p_{abs}\,(0 < p_{abs} < 1)$ percent are judged to be abstract and constitute V_A. Kiela et al. (2014) confirmed that the degree of image dispersion is a valid measure for concreteness by showing a high correlation between concreteness ratings and dispersion values. Furthermore, as compared to concreteness rating data, image dispersion is more appropriate for judging the importance of embodiment in the acquisition of word meanings, because the consistency among images for a word is likely to imply that this word refers to a discernible object or action in the world that we can easily recognize through perceptual experience.

Determining semantic neighbors (or in other words, "mediator words") of an abstract word is also a key step for successful modeling. We apply a simple method that chooses n concrete words most similar to a target abstract word. However, this simple method may not work well in some cases. Some highly abstract words (e.g., *truth*, *wisdom*) may not have semantically similar concrete words, but this method forces such abstract words to be represented by visual vectors of n concrete words, which may be relatively more similar than other concrete words but absolutely dissimilar to the abstract words. Hence, we consider another method, in which $N(> n)$ semantic neighbors of an abstract word are selected from the whole vocabulary V, and then n most concrete words are selected from N semantic neighbors according to the degree of image dispersion. If n concrete words cannot be selected (i.e., N semantic neighbors contain less than n concrete words), no visual vector is attached to that abstract word in the same way as Kiela et al.'s (2014) method.

Evaluation Experiment

Test Data

As test data for evaluating distributional semantic models, we used two word similarity norms, namely WordSim-353 (Finkelstein et al., 2001) and SimLex-999 (Hill, Reichart, & Korhonen, 2015), which are well-known gold-standards widely used in the computational linguistics community. These data contain English word pairs (353 pairs in WordSim-353 and 999 pairs in SimLex-999) with human similarity ratings. The similarity ratings in WordSim-353 are regarded as measures of semantic relatedness (or in other words, associative similarity), while the ratings in SimLex-999 reflect semantic similarity (or taxonomic similarity). For example, a pair of *clothes – closet* is rated 8.00 in WordSim-353, but 1.96 in SimLex-999. These two words are related because the word "closet" refers to a small room or cabinet for hanging "clothes," but their meanings are not similar.

Multimodal Semantic Space

We generated a text-based semantic space (i.e., DSM_L in Figure 1) from the British National Corpus with 100M word tokens as follows. First, we constructed a word-cooccurrence matrix with 43,528 different words (and thus the dimensions of the matrix were 43,528 $\times$ 43,528). The cooccurrence frequency was counted within a sentence. Second, the word-cooccurrence matrix was weighted by positive pointwise mutual information (PPMI), which generally achieves better performance (Bullinaria & Levy, 2007). Finally, the weighted matrix was smoothed by singular value decomposition (SVD) and the dimension of row word vectors was reduced to 300.

To construct a visual semantic space (i.e., DSM_V in Figure 1), we collected 30 images for each word using *Flickr* image retrieval, from which a visual vector for that word was constructed. As a method for visual vector computation, we used the well-known Bag of Visual Words (BoVW) approach (Sivic & Zisserman, 2003). BoVW computes a vector representation for each image by clustering into visual words a set of local descriptors obtained from the image dataset, and then by computing a histogram of visual word counts. In this paper, we extracted SURF descriptors (Bay, Tuytelaars, & Van Gool, 2006) using a dense keypoint sampling method. Each local descriptor is represented by a 3×128-dimensional vector, because a SURF feature is normally represented in a 128-dimensional vector and computed for each of the three components of HSV. All the SURF feature vectors were clustered into 300 visual words, and each image is represented by a 300-dimensional histogram vector of visual words. Finally, a visual vector for each word is computed by summing the visual vectors of all the images retrieved by that word, and then by weighting the summed vector by PPMI.

Method

According to the proposed method shown in Figure 1, we constructed several grounded semantic spaces by changing the values of the parameters p_{abs} (from 0.10 to 0.95 in steps of 0.05), n (from 1 to 5), and N (from 10 to 50 in steps of 10, 100, and all). To evaluate the relative performance of the proposed model, we also generated the standard multimodal representations and Kiela et al.'s (2014) filtering-based representations. In the standard method, every word is represented by concatenating its linguistic vector with a visual vector computed directly from images of that word. In Kiela

Table 1: Correlation coefficients between the cosine similarity computed by the distributional semantic models and the pairwise similarity ratings of two datasets. For the proposed model, the highest correlation is shown boldfaced for each of the two datasets. For the other models (including the proposed model with N =all), results obtained with the same parameters as the proposed model are selectively shown.

Model	WordSim-353	SimLex-999
Text-based DSM_L	0.525	0.248
Visual DSM_V, Standard	0.227	0.118
Visual DSM_V, Proposed ($p_{abs} = 0.50$, $n = 2$, N =all)	0.257	0.112
Visual DSM_V, Proposed ($p_{abs} = 0.95$, $n = 2$, N =all)	0.185	0.085
Grounded DSM_G, Standard	0.476	0.235
Grounded DSM_G, Kiela et al. (2014) ($p_{abs} = 0.50$)	0.510	0.262
Grounded DSM_G, Kiela et al. (2014) ($p_{abs} = 0.95$)	0.532	0.245
Grounded DSM_G, Proposed ($p_{abs} = 0.50$, $n = 2$, N =all)	0.478	0.220
Grounded DSM_G, Proposed ($p_{abs} = 0.95$, $n = 2$, N =all)	0.390	0.179
Grounded DSM_G, Proposed ($p_{abs} = 0.95$, $n = 2$, $N = 30$)	**0.551**	0.260
Grounded DSM_G, Proposed ($p_{abs} = 0.80$, $n = 5$, $N = 30$)	0.535	**0.278**

Note. p_{abs} = the proportion (i.e., threshold) of abstract words; n = the number of semantic neighbors (i.e., mediator words) from which visual vectors of abstract words are computed; N = the size of population from which n mediator words are chosen.

et al.'s (2014) filtering-based method, concrete words, which are determined by image dispersion, are represented in the same way as the standard method (i.e., concatenation of their linguistic and visual vectors), while abstract words are represented only by their linguistic vectors.

The performance of each semantic space was measured by Pearson's correlation coefficient between the computed cosine values and the semantic similarity ratings in the test data. If the proposed semantic space can improve the performance as compared to the standard multimodal and Kiela et al.'s (2014) semantic spaces as well as to the text-based semantic space, it would demonstrate that the indirect embodiment view of abstract words is supported computationally.

Result

Table 1 shows the correlation coefficients between the cosine similarity computed by the distributional semantic models and the similarity ratings of the test data. Overall, the proposed grounded semantic space achieved better performance on both datasets than the text-based model, standard multimodal model, and Kiela et al.'s (2014) model. This result indicates that the proposed model is made psychologically more plausible by language-mediated visual representations of abstract words, thus providing computational evidence for the indirect embodiment view of abstract words.

The result that the standard multimodal model did not outperform the text-based model is consistent with Kiela et al.'s (2014) result; this indicates that inclusion of perceptual input does not always improve the quality of lexical meaning representations. To overcome this difficulty, Kiela et al. (2014) did not use perceptual information to represent image-dispersed words, but we use language-mediated perceptual information for those words, in accordance with the recent idea of embodiment of abstract concepts. Hence, the observed better performance of our model against Kiela et al.'s (2014) model demonstrates the plausibility of our refined representations and the indirect embodiment view.

On the other hand, the proposed method yielded worse performance when all abstract words are represented using the visual vectors of their semantic neighbors (i.e., in the case of N = all in Table 1). This result implies that indirect embodiment via language is not always effective; Kiela et al.'s (2014) method of not including visual information works better in some cases. In other words, some abstract words (e.g., *truth*, *wisdom*) need not to be grounded in perceptual experience and their meanings may be able to be captured primarily by linguistic information.

The proposed grounded model achieved the best performance at very high p_{abs} ($p_{abs} = 0.95$ for WordSim-353 and $p_{abs} = 0.80$ for SimLex-999). This result indicates that only a relatively small number of words need to be grounded directly in perceptual information; most words are likely to be grounded via language or need not to be grounded. It may also implies that word concreteness does not necessarily determine the necessity of grounding in perceptual experience. Table 2 lists some example words that are computed as concrete and abstract by the proposed model. Some highly concrete words (e.g., baby, computer, seafood) whose concreteness ratings (1 to 5) (Brysbaert, Warriner, & Kuperman, 2014) are higher than 4.5 and many moderately concrete words (e.g., currency, interview, summer) whose concreteness ratings are higher than 3.0 are classified as abstract. Nevertheless, indirect grounding is effective in representing these words. The proposed model improved the accuracy of similarity computation for word pairs in WordSim-353 such as *governor–interview*, *currency–market*, *summer–drought*,

Table 2: Examples of concrete and abstract words in WordSim-353 classified by the proposed DSM_G ($p_{abs} = 0.95$, $n = 2$, $N = 30$). For an abstract word, its mediator words (i.e., concrete words used for computing its visual vectors) are also shown in parentheses. Note that no mediator words are shown for some abstract words because N semantic neighbors of those words do not include n concrete words.

Concrete words in V_C
animal, bread, car, carnivore, collection, food, forest, impartiality, jaguar, mammal, marathon, market, moon, office, shower, television, tiger, water, zoo
Abstract words in V_A
activity, alcohol, astronomer (*lunar, saturn*), baby, bird, book (*collection, print*), center, chance, children (*playgroup, whist*), cognition (*representational, intentionality*), computer, currency (*ecu, lira*), discovery, drought (*soil, rainforest*), experience, family (*home, bear*), flood (*stream, storm*), fruit (*bean, salad*), girl, governor (*governorship, appoint*), index, interview (*quote, television*), king, law, liquid (*water, bubble)*, love, marriage, media (*magazine, television)*, museum (*collection, painting*), music, psychology, reason, seafood (*salad, pasta*), sprint (*marathon, athlete*), summer (*garden, lake*), treatment, word

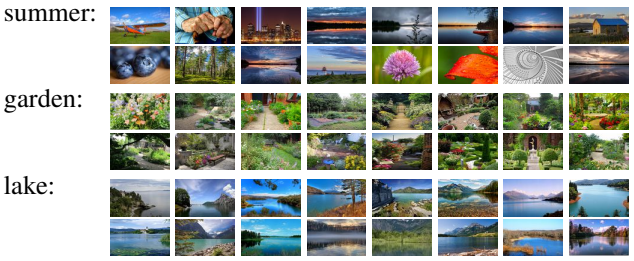


Figure 2: Examples of images for an indirectly grounded word *summer* and its mediator words *garden* and *lake* used in our multimodal distributional semantic model.

and *seafood–food*, as compared to the text-based model and the standard multimodal model. For example, as shown in Figure 2, a variety of images associated with the word *summer* seem not to represent one discernible concept, and thus they are somewhat harmful for an appropriate visual representation of *summer*. On the other hand, images for the mediator words *garden* and *lake* are highly consistent, and these visual information can represent some aspects of summer.

Hence, it may be inappropriate that we refer to the partitioning of vocabulary (V_A and V_C at Step 2 of Figure 1) in terms of abstractness. Rather, what this partitioning really reflects is the necessity of grounding in perceptual experience. Note that this finding is consistent with the existing findings on lexical acquisition (e.g., Gleitman et al., 2005). "Easy" words (mostly nouns), which express concrete basic-level concepts, are learned at the early stage of word learning, and then many other words are learned through the knowledge of easy words, leading to a rapid vocabulary growth.

Discussion

In this paper, inspired by the recent view of the embodiment of abstract words, we have proposed a novel approach to multimodal distributional semantics. In the proposed method, the grounded visual representation of an abstract word is learned through semantically related mediator words, rather than directly from images for that word as used in the standard multimodal distributional semantics. The plausibility of the proposed method is tested and justified by comparing the cosine similarity computed by distributional semantics with human semantic similarity ratings. This finding can be interpreted as supporting the indirect embodiment view of abstract words.

However, our model is not so sophisticated and rather preliminary. It did not achieve a statistically significant improvement over the existing multimodal semantic models. We can point out several issues to improve the quality of the model. One important issue to address for future research is a method for choosing mediator words whose visual vectors form the grounded representation of an abstract word (at Step 3(b) of Figure 1). In this paper, we simply apply semantic similarity computed by the text-based model as a measure of choosing mediator words, but a number of empirical studies have suggested that some other factors affect the embodiment of abstract words. For example, emotion (or affect) is suggested to play an important role in processing abstract words (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011; Vigliocco et al., 2014). Affective experience is crucial in the grounding of abstract concepts, and thus it is likely that affectively similar words are effective mediators of embodiment for abstract words. Age-of-acquisition may also be related to the embodiment of abstract words, because basic words learned at the early stage of lexical acquisition are represented primarily perceptually, while other words learned at the later stage are acquired through the knowledge of basic words (Gleitman et al., 2005; Thill et al., 2014).

Another important research topic is the use of motor information for distributional semantics. Multimodal semantic spaces learned from static images are less than grounded in motor experience. To overcome this difficulty, we can consider distributional semantics using feature vectors extracted from videos or continuous image sequences.

Further refinement can be achieved in the proposed model if we consider different ways of indirect grounding according to the degree of abstractness for words. For example, the number of mediator words can be determined depending on the abstractness of words. We can also combine an original visual vector directly computed from the images for an abstract word with those for mediator words if the original vector is somewhat useful for representing the meaning of the abstract word.

Through these improvements, we can obtain a psychologically more plausible distributional model, and multimodal distributional semantics would hold promise as the computational basis for a psychologically testable theory of embodied

cognition and symbol grounding.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers JP15H02713, JP15K12045, and SCAT Research Grant.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–498.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Linguistic Issues in Language Technologies*, *10*(1), 3–13.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer vision — ECCV 2006* (pp. 404–417). Springer.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796.

Borghi, A. M., & Cangelosi, A. (2014). Action and language integration: From humans to cognitive robots. *Topics in Cognitive Science*, *6*, 344–358.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Jounral of Artificial Intelligence Research*, *49*, 1–47.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–991.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Dove, G. (2014). Thinking in words: Language as an embodied medium of thought. *Topics in Cognitive Science*, *6*, 371–389.

Dove, G. (2015). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. *Psychonomic Bulletin & Review*, Online First Articles.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placeing search in context: The concept revisited. In *Proceedings of the 10th international conference on world wide web* (pp. 406–414).

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.

Glenberg, A., & Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *43*, 379–401.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). New York, NY: Oxford University Press.

Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (pp. 231–236).

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing (emnlp2015)* (pp. 2461–2470).

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 835–841).

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of workshop at the international conference on learning representation (iclr).*

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ieee international conference on computer vision (ICCV 2003), volume 2* (pp. 1470–1477).

Thill, S., Padó, S., & Ziemke, T. (2014). On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, *6*, 545–558.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Vigliocco, G., Kousta, S.-T., Rosa, P. A. D., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*, 1767–1777.

Wang, J., Conder, J. A., Blitzer, D. N., & Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Human Brain Mapping*, *31*, 1459–1468.

Active Overhearing: Development in Preschoolers' Skill at 'Listening in' to Naturalistic Overheard Speech

Ruthe Foushee (foushee@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94704

Abstract

Overhearing can be seen as active learning, and overheard speech provides an increasingly viable source of linguistic input across development. This study extends previous results showing learning from overhearing simplified, pedagogic speech to a more ecologically valid context. Children learn multiple words and facts corresponding to novel toys either through an overheard phone call or through direct instruction. Remarkably, 4.5–6-year-olds learned four new words equally well in both conditions. Their performance on a set of six facts was even better, especially when taught directly. Analysis of the videos revealed that older children with high test accuracy both looked toward the experimenter often, and tracked objects as she discussed them. 3–4.5-year-olds only learned facts from overhearing, and exhibited greater varability in attention. These results suggest learning from overhearing is driven by attention to the indirect input, and may be a skill that undergoes substantial development during the preschool years.

Keywords: active learning; lexical development; overhearing

Introduction

While studies of children's word-learning primarily examine dyadic interactions where adult and child are engaged in joint attention, this represents but one of the myriad contexts in which children across the world learn their first words (Correa-Chávez & Rogoff, 2009; Ochs, 1982; Schieffelin, 1990; Ward, 1971). Not only does the degree to which children's early word-learning occurs in dyadic labeling contexts vary substantially across cultures (and even households), but the link between joint attention and vocabulary development arguably wanes in the second year of life (Akhtar & Gernsbacher, 2007). An existing body of work has demonstrated children's ability to learn individual items from simplified speech. The current study extends this work by providing a more ecologically valid test in an effort to understand the potential utility of learning from overhearing in the wild. We explore the idea that the disassociation between joint attention and vocabulary development even in cultures where a child typically receives substantial child-directed speech might mark the beginning of her playing a more active role in her own lexical development. Specifically, we examine preschooler's acquisition of new words by overhearing non-child-directed speech through the lens of active, or self-directed, learning.

Learning new words and information through overhearing can be seen as a case of active learning for several reasons. For one, children's information-gathering in this context is self-directed (Gureckis & Markant, 2012), and their learning seems to be linked to the degree to which they attend to the indirect speech (Martínez-Sussman, Akhtar, Diesendruck, & Markson, 2011). If a feature of self-directed learning, as suggested by Bruner (1961), is also that an individual improves at the very skill of it through practice, then cross-cultural variation in child-directed speech customs, which uniformly result in linguistic proficiency on the part of the child, might represent evidence of children having 'learned to learn.' More specific evidence for this idea comes from Shneidman et al. (2009), who found children who spent more time alone with multiple adults performed better in an overhearing task. Finally, recognizing an 'information gap' (Loewenstein, 1994) like the unknown names of novel objects, as well as how to fill it (i.e., by overhearing) requires the metacognitive awareness that may well be considered a characteristic of an active learner.

Experimental studies of overhearing have demonstrated that children can learn new words from indirect speech as early as 18 months (Floor & Akhtar, 2006; Gampe, Liebal, & Tomasello, 2012). In these studies, learning of a word for a novel object by children exposed to the object-word mapping via overhearing is compared to that of children who were taught the mapping via direct address (Akhtar, Jipson, & Callanan, 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe et al., 2012; Martínez-Sussman et al., 2011; ?, ?). These studies have found learning from a third-party adult conversation even in the presence of a distracting toy (Akhtar, 2005). While these studies have successfully demonstrated that toddlers can learn new words without joint attention, they leave room for improvement in terms of the naturalistic nature of the overhearing opportunity. In many cases, the child witnesses a new word being taught pedagogically to another adult, and the word often appears in child-directed speech and/or simplistic linguistic frames: either explicit labeling contexts (e.g., "This is a *blicket*."), or embedded in a directive (e.g., "Can you put the *toma* in here?"). The current study follows Gampe et al. (2012), which avoids such linguistic frames because they are rarely used between adults and therefore likely fail to simulate the daily opportunities children have to learn new vocabulary from overhearing. Both explicit

labeling and directives might make it unnecessarily simple for the child to attend to the novel word. We similarly used adult-directed, rather than child-directed speech. While children undoubtedly have ample opportunity to overhear adults talking to other children, we saw overhearing adult-directed speech as a more stringent test of children's capacity.

The current study also improves upon previous investigations of overhearing by increasing the number of learning targets, with the assumption that the inter-adult speech children have access to will contain not one, but many words which are unfamiliar. As in Martínez-Sussman et al. (2011), both words and facts are included to compare learning of new vocabulary to learning of information more generally. We selected a phone call, used in related studies of social cues to word learning (Baldwin et al., 1996; Bannard & Tomasello, 2012), as our overhearing context. Discussions with parents during piloting indicated that this was a familiar situation, and in fact, several parents' cell phones rang during the course of the study. By using a phone conversation, the overheard input could be highly controlled and non-pedagogical while still being representative of naturalistic, adult-directed speech. It is worth noting that while undoubtedly present in many homes, this is a source of linguistic input that has previously been discounted, along with speech to pets, in examinations of effective input for vocabulary development, as it does not involve a present human interlocutor (Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013).

In line with the increased complexity of the learning situation, we tested older children than the youngest involved in previous studies. Participants in Experiment 1 were 4.5–6 years old, and those in Experiment 2 were 3–4.5. Both studies seek to address (1) whether children can learn new words and facts from naturalistic overheard speech, and (2) how self-directed and didactic language-learning compare during this period of development.

Experiment 1

Participants

Participants were 48 children between 4.5 and 6 years of age (26 female; M = 5.22 years, SD = 0.51 years). An additional three children participated but were excluded for failing familiar trials (1), refusing to complete the study (1), or having already witnessed a sibling participate (1). Participants were tested in a quiet space at UC Berkeley, a preschool, or a children's museum. Parents and siblings were often present, and sessions were typically video recorded by at least one camera, contingent on parental consent.

Stimuli & Procedure

Stimuli were six toys, four novel, and two familiar (a plush dog and a toy cup of milk), shown in Figure . Participants were randomly assigned to one of two conditions: Overhearing or Didactic, and were trained on one of two mappings between novel labels and objects. Target words and facts corresponding to one of the mappings appear in Table 1.



Figure 1: Stimuli used in Experiment 1.

Table 1: Words and facts used in Experiment 1.

Word	Fact
fep	from Disneyland
pimwit	my sister loves
toma	uncle gave me
zav	found in the garden
dog	bring to school
cup	had for two years

Overhearing Participants in the Overhearing condition sat across a low table from a confederate. The experimenter entered the room separately, placing a box with the six toys on the table, and introducing them with "These are my toys!" The confederate then removed each toy from the box individually, commenting that they were unfamiliar to her ("I've never seen these toys before, these are [Experimenter]'s toys!"). The experimenter sat on a chair against the wall to the child's left, "working" on her laptop. After all the toys were out, the confederate encouraged the child to continue playing, and went to fill out paperwork against an adjacent wall. One minute later, the experimenter's phone rang, and she delivered a script casually describing each of the toys in the context of a naturalistic conversation with a friend. Following a brief greeting and exchange of pleasantries, she discussed each of the novel objects in segments approximately 15 seconds long, during which she referred to physical properties of the objects (e.g., color, shape), and mentioned each word five times and each fact once. She avoided making eye contact with the child. At the end of the phone call, she briefly mentioned the words and their associated facts again, before telling the friend she had to hang up to play a game and joining the child at the table for the test phase.

Didactic Children in the Didactic condition heard the same script, but directly addressed to them. The experimenter spoke enthusiastically, made eye contact, and directed their attention to each object as it was labeled. Properties that appeared in the script (e.g., "has stickers that you can take on and off") were pedagogically demonstrated for the child. As in the Overhearing condition, the learning phase was

followed immediately by the test phase.

Test Phase The test phase was identical in both conditions, and consisted of three blocks of six trials each. These 18 test trials were presented in one of two pseudorandom orders, one the reverse of the other. The first two blocks tested word-learning, while the third tested learning of facts. On each trial, the child was asked to place the toy associated with a particular word or fact into a container, which varied by block (i.e., "Can you put the *zav / one I found in the garden* in the bowl/box/hat?"). Responses were recorded and coded for correctness. Children who failed trials requesting familiar objects by name were excluded. At the end of the test phase, the experimenter asked the child to identify her favorite toy, and noted her response.

Videocoding Videos of sessions in the Overhearing condition were coded in Datavyu to analyze children's behavior during the experimenter's phone call. The experimenter's speech was coded into segments distinguishing the social portions of the call from the learning-related content of the script, which was further divided into periods comprising discussion of each individual toy. Periods in which the child turned her head to the experimenter, along with the specific object or objects with which the child was playing were coded without audio or reference to the speech coding. Given that Martínez-Sussman et al. (2011) found a positive relationship between the child's gaze to the experimenter and learning, we predicted that children who looked more toward the experimenter as she spoke would show greater accuracy at test. Our study involved many objects, however, so visual attention to the experimenter might actually jeopardize identifying the appropriate object to which to map an overheard novel word or fact. We therefore coded their touch behavior and later related it to the segments of the experimenter's speech as a proxy from which to infer children's accurate tracking of the objects as the experimenter discussed them. We also predicted that children who demonstrated better tracking of the objects would perform better at test.

Results & Discussion

Test Performance Children learned both novel words and facts above chance (0.25 and 0.17, respectively) in both conditions (Overhearing, words: 43% accuracy, $t(191) = 5.22$; facts: 67% accuracy, $t(143) = 12.68$. Didactic, words: 41% accuracy, $t(191) = 4.53$; facts: 88% accuracy, $t(143) = 25.61$; all $p's < 0.001$). While children in the Didactic condition performed significantly better than those in the Overhearing condition on facts ($t(256.32) = -4.33$; $p < 0.001$), there was no difference in performance between the two conditions in word-learning ($p = 0.6$). Finally, performance on word-learning was significantly worse than that on fact-learning in both conditions (Overhearing: $t(315.86) = 4.30$; $p < 0.001$; Didactic: $t(330.37) = 10.28$; $p < 0.001$). Performance means by learning target and condition appear in Figure 2.

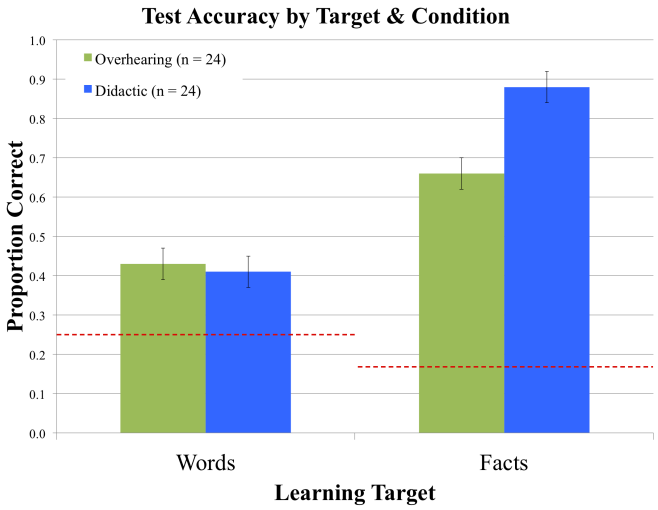


Figure 2: Experiment 1 mean accuracy at test by learning target and condition. Chance for each target type is indicated with a dashed line, and error bars indicate SEM.

We further fit a mixed effects logit model predicting test accuracy (incorrect = 0, correct = 1) from an interaction between condition (Didactic or Overhearing) and learning target (word or fact), with random intercepts for subject to the data, excluding the trials testing comprehension of the familiar object labels used as controls. Fact-learning was better than word-learning in both conditions ($B = 2.43$, $SE = 0.01$, $z = 8.01$, $p < 0.001$), and condition only made an impact on accuracy on facts, such that being in the Overhearing condition decreased one's odds of accuracy on fact performance ($B = -1.42$, $SE = 0.38$, $z = -3.73$, $p < 0.001$), but not on word performance ($B = 0.11$, $SE = 0.26$, $z = 0.44$, $p = 0.66$). This model resulted in a significantly better fit than the null model with no predictors and random intercept ($\chi^2(3) = 36.62$, $p < 0.001$, AIC for model with interaction: 816.7, AIC for null model: 1100.5), as well as a model which included both learning target and condition, but not their interaction ($\chi^2(1) = 14.67$, $p < 0.001$, AIC for model without interaction: 829.4). Models including these predictors and preferred item, requested item, age, gender, order (1 or 2), learning order (taught in the first or second half of learning phase) mapping (1 or 2) also did not result in significantly better fits, suggesting these factors were not significant influences on children's performance.

To test the potential impact of the inclusion of facts relating to familiar versus novel objects, we fit an additional model with condition and familiarity-of-object to just the fact-learning data. A model which also included an interaction between condition and familiarity resulted in a significantly better fit ($\chi^2(1) = 4.25$, $p < 0.05$, AIC without interaction: 295.65, AIC with interaction: 293.40).

The familiarity of the object affected children's accuracy on fact-learning only in the overhearing condition, where a fact's correspondence to a novel, rather than familiar, object decreased their odds of demonstrating learning of the mapping at test ($B = -1.40$, $SE = 0.68$, $z = -2.06$, $p < 0.05$).

Video Measures To obtain a single measure of each child's attention to the content of the phone call, we first calculated the proportion of each segment of the call during which the experimenter was discussing one of the novel objects, and the child was touching that same object. From this, we subtracted the mean proportion the child was playing with that novel object during the five segments of the call in which the experimenter was *not* discussing that object. If the child exclusively attended to that object when it was the one the experimenter was describing, the score would remain unchanged. If, on the other hand, the child merely liked that object, and played with it regardless of what the experimenter spoke about, or played with non-matching objects generally more often, the score would be negative or zero. A score of 1 would indicate the child only ever touched that object while the experimenter spoke about it, and any positive score below 1 would imply she did this more often than not. Seventeen out of the 19 participants for whom we received video consent received positive scores on this measure, and one of the 19 did not touch the objects at all (range: $-0.04 - 0.63$, $M = 0.31$, $SD = 0.18$).

In addition, there was substantial variation in the proportion of the call's duration children spent orienting their gaze toward the experimenter ($0 - 0.47$, $M = 0.14$, $SD = 0.14$). Here, the mixed effects logit model with the best fit to the data for novel word and fact accuracy in video recorded children in the Overhearing condition included an interaction between the child's gaze and touch measures, such that their odds of accuracy increased the higher the two scores were ($B = 15.728$, $SE = 7.0$, $z = 2.25$, $p < 0.5$). While the gaze measure alone had no significant impact ($B = -1.49$, $SE = 1.71$, $z = -0.87$, $p = 0.38$), the touch measure otherwise slightly decreased subjects' odds of accuracy on test as it increased ($B = -4.09$, $SE = 1.58$, $z = -2.58$, $p < 0.01$). The interaction between the two validate the measure in that their relation makes sense given the demands of the task. While it might be detrimental to look only at the experimenter, or interact exclusively with the toys as they are discussed, a learner who distributes her attention between the experimenter discussing the objects, and the objects themselves, is most likely to learn the mappings.

Experiment 2

Having demonstrated that 4.5–5-year-olds can learn multiple new words from a single overhearing exposure as well as they can through didactic instruction, we pursued the development of this skill in younger children. Will we see more of a trade-off in learning between conditions in younger ages? Several children in the Overhearing condition of Experiment 1 made triumphant declarations at test of having "spied" to discover the answers, suggesting they were unaware the phone call was for their benefit, but recognized it was relevant to their situation. Will younger children have the same realization and overhear to learn more about the objects?

Participants

54 children ages 3–4.5 years participated (25 female; M = 3.89 years, SD = 0.40 years). An additional six children participated, but were excluded due to failing the familiar trials (4), not finishing the task (1), or experimenter error (1).

Stimuli & Procedure

The method for Experiment 2 is identical to the previous experiment, except that the number of novel words was reduced by one to make it more appropriate for a younger age range. The children therefore learn—either through overhearing or didactic instruction—three novel words and five novel facts, and receive 15 test trials rather than 18.

Results & Discussion

Test Performance Like their older counterparts, both Overhearing and Didactic participants performed above chance (0.20) on fact-learning (Overhearing: 49% accuracy, $t(108) = 5.95$. Didactic: 71% accuracy, $t(114) = 12.11$; both $p's < 0.001$). Where the younger children differ, however, is in their word-learning through overhearing. While children in the Didactic condition performed above chance (0.33) on word-learning (51% accuracy, $t(137) = 4.24$; $p < 0.001$), those in the Overhearing condition did not (36% accuracy, $p = 0.59$). Performance on both word- and fact-learning was better in the Didactic condition (words: $t(261.36) = -2.60$, $p < 0.01$; facts: $t(216.91) = -3.54$, $p < 0.001$). Mean accuracy on words and facts for the two conditions is plotted in Figure 3.

As in the previous experiment, models with Condition (Didactic or Overhearing), learning target (words or facts) and random intercepts for subject were fit to the data for facts and novel words. Both were significant predictors, and indicated that being in the Overhearing condition decreased one's odds of accuracy at test ($B = -1.02$, $SE = 0.29$, $z = -3.5$, $p < 0.001$), as did the target being a novel word, rather than fact ($B = -0.89$, $SE = 0.19$, $z = -4.78$, $p < 0.001$). The model which included an interaction between Condition and learning target did not result in a significantly better fit ($\chi^2(1) = 0.97, p = 0.33$, AIC without interaction: 755.93, AIC with interaction: 756.97), suggesting that the impact of condition did not differ substantially by learning target, as it had in the previous experiment. Adding mapping, order, learning order, age, gender, preferred item, or requested item as predictors did not result in models with significantly better fits, suggesting they were not particularly influential in children's performance. The selective effect of familiarity in the overhearing condition was replicated for this age group. A model with random intercepts for subjects, familiarity, condition, and an interaction between the two as predictors of test performance was fit to the fact-learning data. The

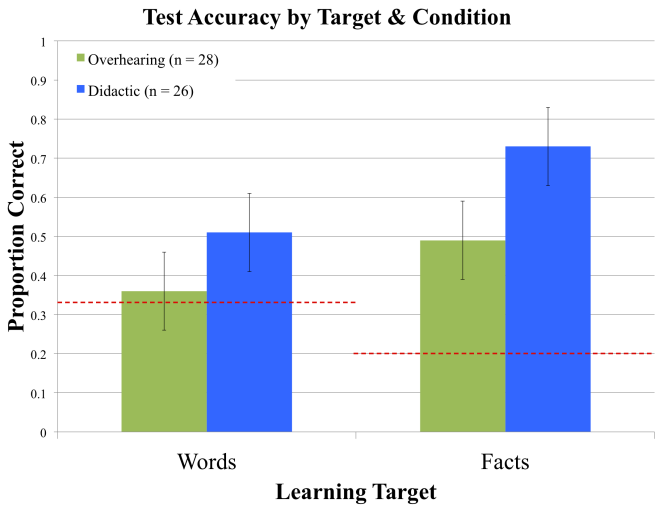


Figure 3: Experiment 2 mean accuracy at test by learning target and condition. Dotted lines show chance, and error bars indicate SEM.

interaction between familiarity and condition was a significant predictor of test accuracy ($B = -1.75$, $SE = 0.58$, $z = -3.04$, $p < 0.01$), though neither was independently (familiarity: $B = 0.63$, $SE = 0.40$, $z = 1.55$, $p = 0.12$; condition: $B = -0.17$, $SE = 0.48$, $z = -0.36$, $p = 0.72$). The model with the interaction (AIC: 337.88) resulted in a better fit than the same model without it ($\chi^2(1) = 9.53$, $p < 0.01$, AIC: 345.41), and a model with only condition ($\chi^2(2) = 10.56$, $p < 0.01$, AIC: 344.44) or familiarity ($\chi^2(2) = 21.63$, $p < 0.001$, AIC: 355.50) as predictors.

Video Measures The 23 children in the Overhearing condition whose parents consented to videorecording show similar range in both behavioral proxies for attention. Children looked toward the experimenter for up to half the duration of the phone call ($0.01 - 0.49$, $M = 0.15$, $SD = 0.18$), and received varied scores on the touch measure ($-0.28 - 0.43$, $M = 0.11$, $SD = 0.20$). Unlike the older age group, five children received negative scores, one never touched any of the objects, and 16 children scored positive values. Already, the greater number of children with negative values on this measure suggests that the younger age group may not have been tracking the overheard mappings as consistently as their older counterparts.

Consistent with this, entering children's gaze proportions and touch measures into a mixed effects logit model with correct response as the binary outcome variable and random intercepts for subject revealed that neither measure, nor an interaction between them, was significant. Thus, not only were the younger children seemingly tracking the objects less, the attentional coordination they *were* doing was less effective for learning, as it was not related to their later performance at test.

General Discussion

The previous two studies have shown that preschoolers can learn a substantial amount of linguistic information via naturalistic overheard speech. Their ability to do so, however, is developing during this period, and their differential performance by age and learning target leave open several questions. First, what accounts for the persistent advantage of fact-learning over word-learning? Other findings of this same asymmetry have attributed it to the difficulty of learning a new phonological form. While that might explain why children performed better on fact accuracy overall, it does not explain why the older children in the Didactic condition performed far better on facts, but equivalently on words. If what we are seeing is a ceiling effect on learning four new words, then why are both groups at that ceiling, if learning from overhearing is sufficiently difficult to make performance on facts significantly worse? Are words and facts qualitatively different in other ways, perhaps in children's level of motivation to learn them?

Given that both words and facts were presented together, we cannot draw strong conclusions about differences in their ease of learning via different methods. Indeed, the familiarity of the objects was found to improve fact-learning in the Overhearing condition, showing that the two are not entirely dissociable. While there was enough information in the overheard script to map the words and facts independently to the objects (confirmed by the difference in performance between the two), learning of one might influence the other. Learning which objects the facts using familiar nouns referred to, for example, could be done without looking to the physical objects to make the correct association. This might explain why object familiarity conferred a fact-learning advantage particularly in the Overhearing condition. To understand the ways in which learning both types of information about the set of objects may have affected performance, we are currently replicating this study with exclusively novel labels as learning targets. If the difficulty with learning novel words in the Didactic condition stemmed from acquiring novel phonological forms, we might expect children's performance to be equivalent to their performance here. If, however, the memory demands of learning both words and facts corresponding to multiple objects imposed severe memory demands, we might expect children to perform better in the words-only study. Alternatively, learning the new words associated with facts which were themselves easier to remember may have been helpful, predicting children should perform worse in the follow-up.

In the Overhearing condition of this experiment, looks toward the experimenter were equivalent in the two age groups. This may have reflected not only attention to the experimenter's speech, which does not in fact require visual access, but instead attempts to establish joint focus with the speaker, or to determine her referential intent. In this context, in contrast to previous studies, the experimenter's reference to the objects at hand was carried in the content of her speech alone, and not in her gaze to the relevant objects. In a sim-

ilar context, where objects were labeled without joint attention, but with visible focus on the objects by a single speaker, infants were able to learn new word mappings (Baldwin et al., 1996; Bannard & Tomasello, 2012). When that cue to referential intent was absent by placing the speaker behind a screen, however, infants did not demonstrate learning of the mappings in an explicit pointing task, but did so in a looking task (Bannard & Tomasello, 2012). It may be, therefore, that younger children in our study had greater difficulty establishing specifically word-to-object mappings without the experimenter looking toward the objects, but would have been able to show some association in a subtler test.

We conclude from the video measures that successful acquisition of object-label mappings from overheard speech is due to appropriately coordinating one's attention between the referents and the speaker. This is not an empty claim—much of language-learning, and even the sort of learning that can occur from overhearing (e.g., phonology: Au et al., 2002), happens implicitly. Children in this study needed to recognize that the phone call that was taking place was relevant to their situation and attend to it in order to learn about the toys they were playing with, even in the absence of social cues to reference from the experimenter. The older group reliably did. The age-related difference we see in learning from overhearing in this study may reflect general development in the mastery of self-directed learning and attention management.

Working in an active learning framework provides several additional follow-ups. For example, one proposed advantage of self-directed learning has been retention. If we use the same procedure, then, and call participants back after several days, we might predict less of a memory loss in children from the Overhearing condition. Another possibility stems from the relation between test accuracy and the videocoding measures. If learning in this context is driven by explicitly distributed attention, are there ways to increase metacognitive awareness of the learning opportunity available in overhearing? Children might attend more if their parents were the speakers, for example. Given how much ambient speech they are exposed to, it might be prudent to preferentially attend to a familiar speaker. Alternatively, the 'information gap,' but not the means to close it, might be made either more apparent or more urgent, perhaps through an explicit goal introduced at the beginning of the study, or a topic about which the child is already interested.

As expected, joint attention and direct address seem to be decreasingly important for word-learning, while we do not see evidence for the same trend with more general fact-learning. We are continuing to undertake new studies, including those outlined here, to better understand the contextual factors that enable children to be active learners from environmental input, linguistic and otherwise.

References

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209.

Akhtar, N., & Gernsbacher, M. A. (2007). Joint attention and vocabulary development: A critical look. *Language and Linguistic Compass*, *1*(3), 195–207.

Akhtar, N., Jipson, J., & Callanan, M. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416–430.

Au, T., Knightly, L. M., Jun, S.-A., & Oh, J. S. (2002). Overhearing a language during childhood. *Psychological Science*, *13*(3), 238–243.

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, *67*(6), 3135–3153.

Bannard, C., & Tomasello, M. (2012). Can we dissociate contingency learning from social learning in word acquisition by 24-month-olds? *PloS One*, *7*(11).

Bruner, J. (1961). The act of discovery. *Harvard Educational Review*, *31*(1), 21–32.

Correa-Chávez, M., & Rogoff, B. (2009, May). Children's attention to interactions directed to others: Guatemalan mayan and European American patterns. *Developmental Psychology*, *45*(3), 630–641.

Floor, P., & Akhtar, N. (2006). Can eighteen-month-olds learn words by listening in on conversations? *Infancy*, *9*, 327–339.

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language*, *32*(3), 385–397.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.

Martínez-Sussman, C., Akhtar, N., Diesendruck, G., & Markson, L. (2011). Orienting to third-party conversations. *Journal of Child Language*, *38*(2), 273–296.

Ochs, E. (1982). Talking to children in Western Samoa. *Language in Society*, *11*, 77–104.

Schieffelin, B. (1990). *The give and take of everyday life: Language socialization of Kaluli children*. Cambridge: Cambridge University Press.

Shneidman, L. A., Arroyo, M. E., Levine, S., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672–686.

Ward, M. (1971). *Them children: A study in language*. New York: Holt, Rinehart, and Winston.

Towards a Cognitively Realistic Representation of Word Associations

Ivana Kajić (ivana.kajic@plymouth.ac.uk)[1,2]
Jan Gosmann (jgosmann@uwaterloo.ca)[1]
Terrence C. Stewart (tcstewar@uwaterloo.ca)[1]
Thomas Wennekers (thomas.wennekers@plymouth.ac.uk)[2]
Chris Eliasmith (celiasmith@uwaterloo.ca)[1]
[1] Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada N2L 3G1
[2] School of Computing, Plymouth University
Plymouth, Drake Circus, United Kingdom PL4 8AA

Abstract

The ability to associate words is an important cognitive skill. In this study we investigate different methods for representing word associations in the brain, using the Remote Associates Test (RAT) as a task. We explore representations derived from free association norms and statistical n-gram data. Although n-gram representations yield better performance on the test, a closer match with the human performance is obtained with representations derived from free associations. We propose that word association strengths derived from free associations play an important role in the process of RAT solving. Furthermore, we show that this model can be implemented in spiking neurons, and estimate the number of biologically realistic neurons that would suffice for an accurate representation.

Keywords: semantic spaces; vector representations; spiking neurons; insight; remote associates test

Introduction

Creating word associations is an important skill for the development of many cognitive abilities. Language acquisition is highly dependent on the ability to associate words (Elman et al., 1997), and the associative organisation of children's language is known to facilitate learning of new words and syntax (Brown & Berko, 1960; Hills, 2013). Furthermore, word associations have been shown to play a role in analogical problem solving important for inference and concept attainment (Powell & Vega, 1971).

Despite their importance in a variety of cognitive tasks, it is not clear how associations are represented in the brain. This question is of interest to researchers in cognitive and computer science. From a cognitive science perspective, understanding word representations has been relevant for the modelling and explaining of psycholinguistic phenomena (Landauer & Dumais, 1997; Jones & Mewhort, 2007; Steyvers, Shiffrin, & Nelson, 2004). In computer science, machine learning has been concerned with optimal word representations for automated natural language processing and text comprehension.

In this study we investigate biologically plausible representations of word associations. To this end, we analyse two sources of information about word associations and different forms of encoding of the associations. We compare these methods on predicting human performance on the Remote Associates Test (RAT). In particular, we note that some methods of representation may be better for solving this task, but those methods may not do as good a job at predicting human performance. Finally, we take the representation method that is closest to human performance and implement it using spiking neurons. Not only does this demonstrate that this algorithm could be implemented biologically, but it also allows us to determine how many neurons would be needed to represent these associations, given realistic biological constraints.

Remote Associates Test

The RAT was conceived to study the ability of an individual to form new associations among seemingly unrelated words (Mednick, 1962). The test consists of a list of word problems and each problem contains three cue words (e.g., *call*, *pay*, *line*). The task is to find a word (*phone*) associated with all three cue words within a time limit of up to several seconds. The words can form a word phrase (phone line), or a compound word (payphone). Individuals scoring higher on the test are assumed to more easily create uncommon and less stereotypical associations between pairs of words. The RAT has been used in psychology and cognitive neuroscience to study creative thinking and insight. Because the RAT problems differ in difficulty, they give us information about which associations are common and therefore easier to come up with. This allows us to infer which ways of representing associations are likely used in the brain as they should reproduce the same patterns of easy and hard problems.

For comparison to human data, this work uses data set from the experimental condition where subjects were given only a few seconds to solve a problem. This is meant to address situations where people are solving RAT based on insight, rather than explicitly searching for solution. Longer time periods would allow analytical solving, rather than relying on unconscious information such as word association, and lead to higher scores on the RAT (Bowden & Jung-Beeman, 2003; Kounious & Beeman, 2014).

Related Work

A number of studies have investigated the influence of word representations in in a wide range of semantic memory tasks (Landauer & Dumais, 1997; Steyvers et al., 2004; Jones & Mewhort, 2007). Common to all approaches is the representation of words as vectors, whose relationships can be quantified using linear algebra methods. Latent Semantic

Analysis (LSA) evaluates word occurrences in large corpora of text and derives vector representations for each word. The words similar in meaning will have similar vector representations. LSA has been successfully applied to explain a variety of psycholinguistic phenomena (Landauer & Dumais, 1997; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Instead of using large corpora of text to derive the semantic spaces, another approach is to use free association norms (FAN; Nelson, McEvoy, & Schreiber, 2004). The association strengths between word pairs have been derived experimentally, by asking each participant to provide the first word which comes to their mind given a word cue. By applying dimensionality reduction techniques on this data, the word association space (WAS) was created and used to predict the performance on semantic memory tasks such as recognition and recall (Steyvers et al., 2004). Corpus-based approaches have been shown to solve the RAT with solution rates higher than humans (Toivonen, Gross, Toivanen, & Valitutti, 2013; Klein & Badia, 2015). However, very few studies have investigated models which match the performance on the RAT with the human performance (but see, Kajić & Wennekers, 2015; Bourgin, Abbott, Griffiths, Smith, & Vul, 2014; Gupta, Jang, Mednick, & Huber, 2012). In this study we analyse what kind of biologically plausible representations yield a performance comparable to human performance on this test.

Representation of associations

We use two datasets to construct the word representations. The first dataset are the free association norms (Nelson et al., 2004), containing association strengths for over 5000 word cue-target pairs. The strengths between the words are organized in an asymmetric association matrix FAN_{asym}, with rows representing cues and columns representing targets used in the free association experiment. The asymmetry is a result of non-reciprocal association strengths. For example, given the cue *left*, 94% subjects respond with *right*, however, given the cue *right*, 41% subjects respond with *left* and 39% subjects respond with *wrong*. Formally, this is a difference in the forward (cue to target) and the backward (target to cue) association strength. In addition to the asymmetric matrix, the symmetric matrix FAN_{sym} is created by adding the asymmetric matrix and its transpose.

The second form of association information is derived from the Google Books Ngram Viewer dataset (version 2 from July 2012; Michel et al., 2011). An n-gram is a sequence of n words, and this dataset provides occurrence frequencies of n-grams across over 5 million books published up to 2008. The set of words used in this study is restricted to the same words that have been used in the FAN data. Furthermore, we only used n-gram frequencies from 2008. For every combination of two words w_1 and w_2 the corresponding entry in the matrix NG_{asym} was set to the sum of occurrences of the 2-gram (w_1, w_2) and the 1-gram $w_1 w_2$ in the corpus. Each row of the matrix was then normalized to sum to one. The symmetric matrix $NG_{\text{sym}}$is computed in the same way as the FAN_{sym}matrix. In the rest of the analysis we will merely use NG_{sym} which includes the backward strength between word co-occurrences. This is necessary to solve the problems where only the second part of the compound word is given as one of the three cues (e.g., *board* for *blackboard*). Even though the NG matrices give co-occurrence counts, we will use the terms *association matrix* and *association strength* as they are used in the same manner as the FAN association matrix.

There are two commonly used approaches to represent word associations. First, we can directly use the association matrix. That is, we represent a word as a localist vector (all zeros except for a single one for the word itself), and then, to perform the association we multiply the word by the association matrix. The non-zero entries in the resulting vector represent the word associates. Alternatively, we can embed the associates in a vector space. That is, instead of representing the full association matrix, we compress that matrix into a lower-dimensional space. In certain cases this approach can adjust the similarity space between the words to uncover latent structure among the associations. For example, this is the basis for Latent Semantic Analysis (Deerwester et al., 1990), where similar words are made more similar and less similar words less similar. In particular, we use singular value decomposition (SVD) to take the 5018-dimensional localist word representation and compress it into an D-dimensional distributed representation (where D is varied between 128 and 4096).

Preliminary evaluation

To determine which representational approach gives the best performance on the task, we use the problem set from Bowden and Jung-Beeman (2003). Out of 144 RAT problems, we used the 117 problems for which the cues and the target exist in the set of free association norms. We took the sum of the vector representations of each cue word, and multiplied it by the association matrix. The resulting vector was compared to the vectors for all of the possible response words. In the ideal case, the correct solution word would be the most similar to this output value. However, we also determined if the solution word was in the top 2, 3, 5, and 10 most similar words, as reported in Table 1.

The results indicate that the solution appears as the top-ranked word more often for Google n-grams (NG_{sym}) than for association norms (11 solutions in the first position versus

Table 1: Target positions for 117 RAT problems

	Within top				
Association matrix	1	2	3	5	10
FAN_{asym}	5	12	16	31	49
FAN_{sym}	4	5	6	14	36
NG_{sym}	11	15	16	22	35

5 and 4 solutions for symmetric and asymmetric matrices). However, if we allow for the solution word to be in the top 5 or 10, then the FAN_{asym} association matrix performs best.

Figure 1 also includes the results from applying SVD to the association matrices. Contrary to expectation, SVD does not improve the performance on the RAT test except in a few specific circumstances. In particular, if we need the solution word to be in the top 3 words and we are using the Google n-grams (NG_{sym}), then using 512-dimensional SVD provides a slight improvement over the full asymmetric FAN matrix.

In majority of cases, the statistical n-gram data performs better than the free association norms. However, this only shows increased performance on the task, not whether the n-gram approach performs similarly to people on this task. To address this question, we compare the two approaches with the human performance.

Matching human performance

Instead of analysing which method gets the most correct solutions on the RAT, we now explore which method yields the results most similar to human results. That is, we are interested to find which method is better in solving problems that humans find easy, and worse in solving problems humans find hard. To do so, we predict a probability of producing the correct solution within a 2 s time limit. We use the same set of problems as in the previous section, and match to the percentage of people solving each problem (Bowden & Jung-Beeman, 2003).

Let $s(w,v)$ be the associative strength from word w to v. Given the three cues c_k with $k = 1,2,3$ each word w_i in the vocabulary is activated according to

$$a(w_i) = \sum_{k=1}^{3} \alpha_k \cdot s(c_k, w_i) \tag{1}$$

where α_k are free parameters intended to model the effect that subjects might differently prioritize the problem cues. We set $a(c_k) = 0$ for the cues to prevent them from appearing high in the results. Moreover, we fix $\alpha_1 = 1.0$ as a scaling of all α_k with a constant will produce the same predictions.

Given that w_s is the solution word, we calculate the predicted probability for producing the correct answer as

$$P = \beta \cdot \frac{a(w_s)}{\sum_i a(w_i)} \tag{2}$$

with β being another free parameter. Note, that we are not calculating the probability of each individual word being given as answer, but the probability of producing the correct vs. the wrong response. Because of that, β is not fixed to one, but should be chosen such that $P \leq 1$.

We did curve fits to the data from Bowden and Jung-Beeman (2003) by minimizing the root mean square error between the proportion of participants solving the problem within the time limit and our predicted solving probability. For the curve fits we used the association strengths from the original FAN_{asym}, FAN_{sym}, and NG_{sym} matrices. In addition, we used the 768-dimensional NG^{768}_{sym} matrix which gave improved performance in Figure 1.

The resulting parameter values are given in Table 2. Representations derived from free norms yield a better fit on this data set ($r^2 = 0.58$) compared to the n-gram data ($r^2 = 0.30$). There was no difference between the asymmetric and symmetric FAN matrices. Interestingly, the second cue gets consistently a higher weight. We speculate that this is caused by this cue appearing in the center of the screen with the other cues above and below. For n-gram fits the parameters α_2 and β are large, but because of the low r^2-value, these values cannot be seen as meaningful. For visual inspection, we have plotted the model fits using free association norms and Google n-grams in Figure 2. Further error analysis revealed that the Google n-grams underestimate the solution probabilities of easy items (more than 32% solved by humans) while at the same time predicting a non-zero probability for items unsolved by humans.

All solutions in the data set used are based on compound words which explains that n-gram data can solve more and harder RAT problems. This also means that the insight process in RAT solving could be based on such co-occurrence information. But the results provide evidence that this is not the case and that the insight process is based on associations closer to the associations produced in an unconstrained free association task. These kinds of associations are likely to be based on additional semantic information not available to purely statistical approaches.

Biological plausibility

While the previous sections argue that people use word association data of the form seen in the free association norms to perform the RAT task, there is the separate question of whether such an operation can be implemented in the brain. Can neurons in the brain precisely implement the mathematical matrix operations described above? How many neurons would be needed to implement it accurately enough?

To determine this, we implemented the above algorithm using two groups of spiking Leaky Integrate-and-Fire (LIF) neurons, with synaptic connections from the first group to the second group. The first group represents the input (the sum of the vectors from the three cue words), and the second group represents the output (the result after multiplying by the association matrix).

To allow a group of neurons to represent a vector (which, in

Table 2: Model fits and best fitting parameters

Association matrix	α_1	α_2	α_3	β	r^2
FAN_{asym}	1.0	2.06	1.20	1.13	0.58
FAN_{sym}	1.0	2.50	1.63	2.86	0.58
NG_{sym}	1.0	13.45	1.25	3.55	0.30
NG^{768}_{sym}	1.0	11.88	1.00	8.23	0.22

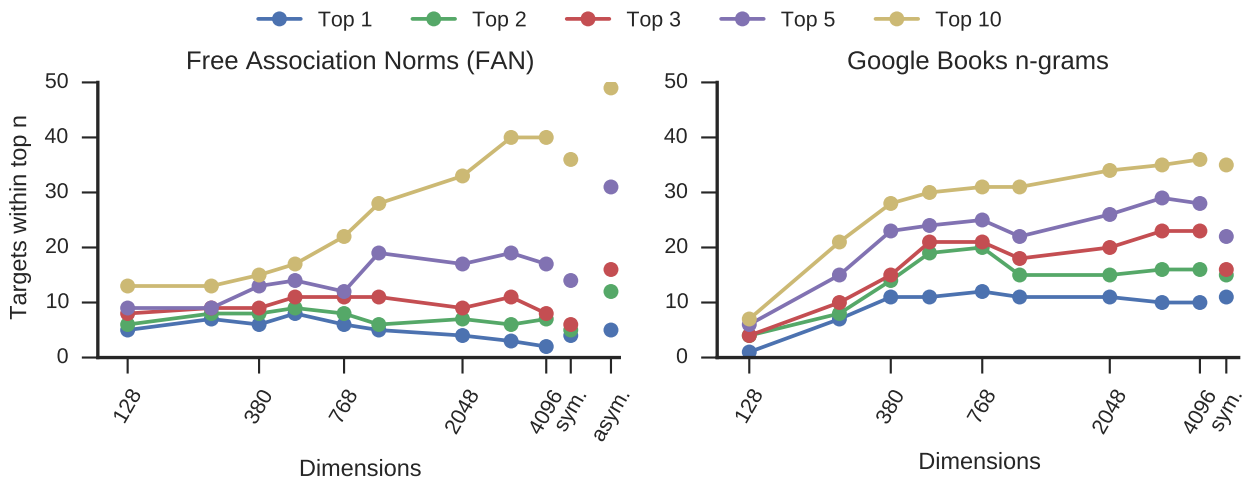


Figure 1: The number of correct solutions within the top n most similar words over 117 problems. The left plot shows the results with representations based on free association norms. The right plot is based on Google n-gram data. The isolated points at the end of the x-axis in both graphs represent the original symmetric and (for FAN) asymmetric matrices. All other data points are computed by applying the SVD to the matrices.

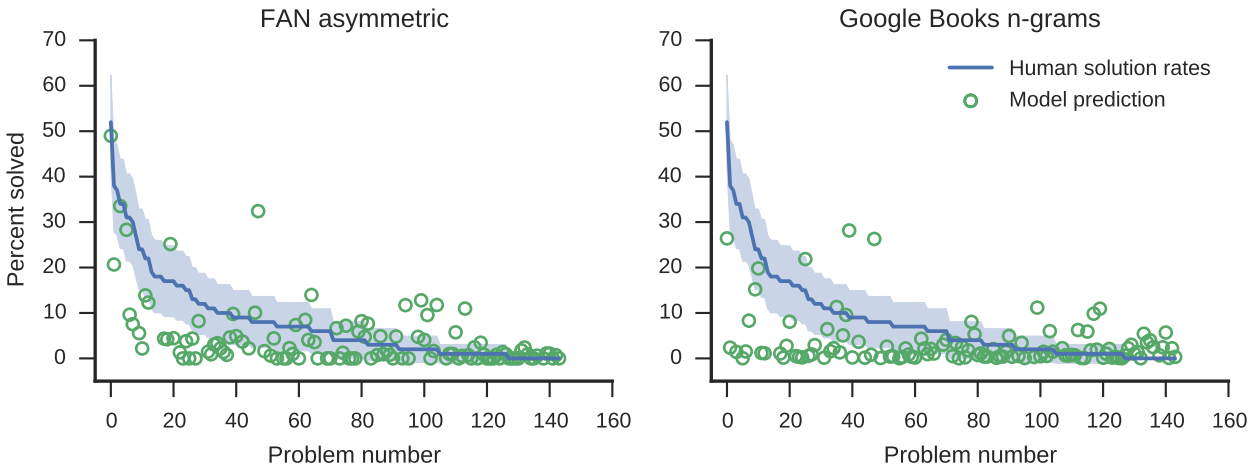


Figure 2: Model fits to human data (blue line with shaded 95 % confidence interval). Human data are percentages of participants solving a RAT problem in 2 s and they have been sorted in descending order, so that problems solved by more participants have lower problem indices. Every green circle represents the probability of producing the correct solution to a RAT problem predicted by the model.

turn, represents a word), each neuron in the group has a randomly chosen preferred vector. This is the vector for which this neuron will fire maximally. For other vectors, the neuron will fire less frequently. This is a generalization of the standard idea of neurons having a preferred stimulus or representation, as seen in motor cortex, visual cortex, and throughout the brain. Importantly, since each neuron's preferred vector is randomly chosen, the neurons will provide a distributed representation, even if the vector representation is localist. For example, if the represented vector is [0, 1, 0, 0], then one neuron might have a preferred vector of [0.1, 0.3, 0.8, 0.5] and so it would fire slowly (the similarity between the two vectors is 0.3, as measure by the dot product), while another neuron might have a prefered vector of [0.2, 0.9, 0.1, 0.4], so it would fire quite frequently (dot product of 0.9). We have previously shown that such a representation is extremely robust to neural damage and consistent with observed patterns of neural activity (Stewart, Bekolay, & Eliasmith, 2011).

Given this approach to representation, we need to connect the first group of neurons to the second group of neurons in such a way that if we cause the first group of neurons to fire as they should when representing a particular cue word vector x, then this should cause the second group of neurons to fire with the pattern for the vector that is the result of multiplying x by the association matrix. We do this by setting the synaptic weights between the first and second groups. Many techniques could be used to perform this task (including standard backpropagation learning rules), but here we simply treat it as a least-squares optimization problem and directly solve for the best set of connection weights for this task. This overall approach is known as the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003).

To test the model we used three RAT problems of easy, intermediate, and hard difficulty, as shown in Figure 3B. We estimate the output similarities from the spiking output with the methods of the NEF and compare it to the analytical result. The accuracy of the neural representation increases as the number of neurons increases. The root mean square error with the analytical result, relative to the word most similar to the cues, is in the range from 4.5 % to 3.0 % depending on the number of neurons ranging from 100 360 up to ten times as many neurons. Thus, we can approximate the model equations with biologically realistic spiking neurons with minor deviation.

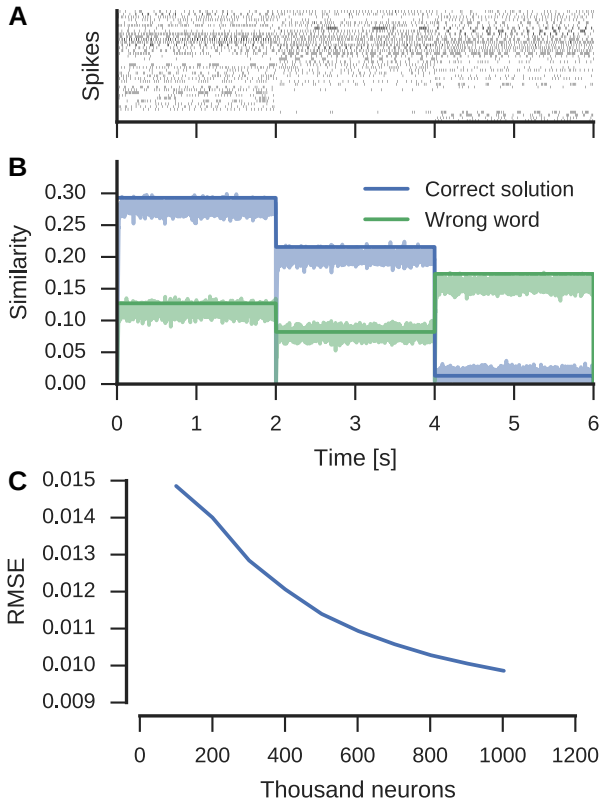


Figure 3: Example run of the neural network model for three RAT problems of easy, intermediate and hard difficulty. (**A**) Spike patterns of a subset of the neurons in the *Solution* ensemble. (**B**) Similarity of the representation in the *Solution* ensemble with the correct solution (blue) and most similar wrong word (green). The solid lines give the analytical result, whereas the semi-transparent lines give the network output. (**C**) Root mean square error between neural network output and analytical calculation as we change the number of neurons.

Discussion

In this study we have done a computational analysis of two different sources of word associations and described how well they predict human performance on the RAT. We have shown that statistical language data like n-grams allow the highest solution rates on this task, consistent with the previous work (Klein & Badia, 2015; Toivonen et al., 2013). However, further analysis revealed that the better prediction of the human performance is obtained with the free association norms.

First, we discovered structural differences between the n-gram data set and association norms by applying dimension reduction with SVD. Previous studies have shown that the dimensionality reduction on association norms can be used to accurately predict human responses on certain episodic memory tasks such as recognition memory and cued word recall (Steyvers et al., 2004). Our analysis provides evidence that dimensionality reduction does not improve performance on the RAT. Moreover, it impairs the performance when the target is among the ten most similar words to the cue words. This indicates that for some RAT problems the important associations are contained in links which are not present in a low-dimensional representation. This is reminiscent of the finding that direct association strengths are the best predictor of intrusion rates in free recall (Steyvers et al., 2004). Whether there is a connection between the associative mechanisms in the RAT and free recall with intrusion remains to be explored in the future work. The dimensionality reduction on n-grams revealed a considerable amount of redundant infor-

mation: the original 5018 dimensional word vectors can be reduced to at least 768 dimensions without large differences in the results. Moreover, the SVD can even lead to improvement when looking at targets which appear within the three most similar words to the cues. Second, the modelling analysis showed that n-gram data, yielding best scores on the RAT, is a worse predictor of human performance.

The FAN data model was a better fit to human solution probabilities in the RAT. As expected, the model was not able to solve any problems for which there was no association between the cues and the target in the association norms indicating that free norms might not be the only source of information. However, for the other problems we have demonstrated that free associations play an important role in the insight process.

Finally, we demonstrated the biological plausibility of this approach in a spiking neural model of the insight solution process of the RAT. The model shows the expected behaviour and is more likely to produce the correct solution for easy RAT problems. In the future, we plan to match the model more rigorously to human data and extend it with recurrent processing to explore a variety of mechanisms and representations involved in the memory search.

Notes

The model and data analysis source code are available at `https://github.com/ctn-archive/kajic-cogsci2016`.

Acknowledgments

This work has been supported by the Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN (CogNovo, grant number 604764), the Canada Research Chairs program, the NSERC Discovery grant 261453, Air Force Office of Scientific Research grant FA8655–13–1–3084, CFI and OIT. This work made use of SHARCNET and Compute Canada computer resources.

References

Bourgin, D. D., Abbott, J. T., Griffiths, T. L., Smith, K. A., & Vul, E. (2014). Empirical Evidence for Markov Chain Monte Carlo in Memory Search. In *Annual Conference of the Cognitive Science Society.*

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *35*(4), 634–639.

Brown, R., & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, *31*(1), 1–14.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1997). *Rethinking innateness: A connectionist perspective on development (neural networks and connectionist modeling)*. MIT Press.

Gupta, N., Jang, Y., Mednick, S. C., & Huber, D. E. (2012). The road not taken: Creative solutions require avoidance of high-frequency responses. *Psychological Science*, *23*(3), 288 – 294.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child- versus adult-directed language. *Journal of Child Language*, *40*, 586–604.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Kajić, I., & Wennekers, T. (2015). Neural network model of semantic processing in the remote associates test. In *Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches, 29th Annual Conference on Neural Information Processing Systems (NIPS 2015).*

Klein, A., & Badia, T. (2015). The usual and the unusual: Solving remote associates test tasks using simple statistical natural language processing based on language use. *The Journal of Creative Behavior*, *49*(1), 13–37.

Kounious, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, *65*, 71–93.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220-232.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... others (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Powell, A., & Vega, M. (1971). Word association and verbal analogy problems. *Psychonomic Science*, *22*(2), 103–104.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2011). Neural representations of compositional structures: Representing and manipulating vector spaces with spiking neurons. *Connection Science*, *22*, 145-153.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237–249.

Toivonen, H., Gross, O., Toivanen, J. M., & Valitutti, A. (2013). On Creative Uses of Word Associations. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis* (Vol. 190, p. 17-24). Springer Berlin Heidelberg.

Don't Blink!
Evaluating Training Paradigms for Overcoming the Attentional Blink

Trudy Buwalda (t.a.buwalda@rug.nl)
Jelmer Borst (j.p.borst@rug.nl)
Marieke van Vugt (m.k.van.vugt@rug.nl)
Niels Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen, the Netherlands

Abstract

A lot of people show a decline in performance when they have to report a second target stimulus in a stream of distractor stimuli. Curiously, this decline only happens when the second target appears approximately 200-500ms after the first target. Recently, Choi, Chang, Shibata, Sasaki, and Watanabe (2012) have shown that a short, one-hour training can eliminate this "attentional blink". Up to now, it is still unclear why this training works. In this paper, we have evaluated a range of different training paradigms to test several hypotheses about the mechanism behind the reduction of the attentional blink. Our results show that none of these training paradigms have a large training effect when administered in isolation. The training by Choi et al. (2012) outperforms them all. The most likely explanation for this effect are temporal expectations relative to the first target.

Keywords: Attentional Blink, Training, Strategy Choice, Temporal Expectations

Introduction

The attentional blink is a deficit in reporting a second target stimulus in a rapid stream of visual stimuli, first reported by Raymond, Shapiro, and Arnell (1992). When a stream of visual stimuli is shown at a rate of approximately 10 items per second (see Figure 1), it is easy to spot one divergent, or target, stimulus. However, when a second target stimulus is presented after the first target, people often do not report this second target. The second target can be presented at different distances, or lags, from the first target (T1). The distance between two targets influences the difficulty with which the second target (T2) is reported. When the lag is between two and five stimuli, accuracy of reporting the second target is considerably reduced.

For a long time, the attentional blink has been thought of as a robust effect that stems from a structural limitation, which cannot be trained away by practice. Therefore it was remarkable when Choi et al. (2012) reported a training paradigm that significantly reduced the attentional blink. This training paradigm is very similar to the original task. In this training, the second target in a stream was made more salient by giving it a red color. In addition, during training the second target was always presented as the second item from the first target (lag 2), but during the pre- and posttest other lags were present as well. After training on this paradigm for only an hour, the attentional blink was almost eliminated and this effect lasted for several months.

Choi et al. (2012) concluded that the results after the colored-target training are due to an increase in temporal resolution: the ability to distinguish between the presented stimuli, which could be the result of top-down attentional processes and bottom-up perceptual processes. From these two hypotheses, Choi et al. (2012) favored a more top-down explanation. We will look at both aspects more closely, using two different experimental conditions.

In this paper, we investigate the mechanisms behind the succes of colored-target training by Choi et al. (2012). To this end, we conducted an experiment with 125 participants, that a) replicates Choi's colored-target training, and b) tests four possible explanations of the training effect. The four explanations that we test are speed of processing, a change in strategy, the implicit use of feedback, or temporal expectations relative to the start of the stream. Before we introduce the experiment, we will first discuss those explanations in more detail.

Processing Speed

A possible explanation for an increase in temporal resolution is a bottom-up increase in perceptual processing speed. To test this, we have trained people to perceive target stimuli faster. We have used a task that was developed by Choi et al. (2012), called the *letter-mask task*. In this task, a letter is presented, followed by a mask. The letter has to be reported as quickly as possible. Presentation time of the letter is constantly updated to represent the accuracy of the participant, i.e. reduced when accuracy increased and vice versa. Choi et al. (2012) found that people improved on this task after the colored-target training.

Because the tasks adapts itself to the level of the participant by modifying the presentation time of the letter, it would be very suitable as a training task as well. We hypothesized that after training on the letter-mask task, discrimination of stimuli in the attentional blink task would be increased, resulting in a similar effect as the colored-target training.

Additionally, if speed of processing is indeed the relevant training aspect in the letter-mask task, we would expect to find no effect when the presentation time of the letter is kept constant. Therefore we also use a letter-mask training with a constant presentation time of 100ms.

Strategy

The second interpretation of temporal resolution is as a top-down process. By using a different cognitive strategy, top-down control can reduce the attentional blink.

Evidence that strategy may play a role in the attentional blink comes from earlier studies that have shown that the at-

tentional blink can be reduced by forcing people to approach the task in a different manner. For example, several studies have shown that a second task, increases performance on the attentional blink task (Olivers & Nieuwenhuis, 2005; Taatgen, Juvina, Schipper, Borst, & Martens, 2009). This could indicate that a disruption of the general strategy people use for the attentional blink actually improves performance.

A study by Ferlazzo, Lucido, Di Nocera, Fagioli, and Sdoia (2007) showed that the attentional blink is significantly reduced when people are instructed to report the two targets as one entity: instead of reporting two letters, they were asked to report a syllable. This shows that the attentional blink is not only caused by low-level attentional and visual processes, but that top-down strategies play a role as well.

To test whether the colored-target task induces a similar strategy change, we have used a training paradigm similar to the tasks used in Ferlazzo et al. (2007). Instead of having to report each target separately, participants had to combine both targets into one entity. In this training task, two digits have to be identified in a stream, the combination of digits had to be reported as a word. For example, the digits 5 and 9 had to be reported as the word fifty-nine. We therefore refer to this task as the *number-word task*.

Feedback

In addition to the explanations for the temporal resolution hypothesis, as posed by Choi et al. (2012), we hypothesized that implicit feedback in the colored-target task may play a role in the reduction of the blink, perhaps by affecting the strategy particpants used. An important aspect of the colored-target task (Choi et al., 2012) is that the second target always appears at the same lag, lag 2. Combined with the salient color, this makes a second target very hard to miss. This setup ensures that a participant knows that something was missed if only one target was perceived. To test the role of feedback in the reduction of the blink, we used a training task in which explicit feedback was given to the participant.

Temporal Expectations

Finally, we also looked at the temporal expectation hypothesis. This hypothesis was put forward by Tang, Badcock, and Visser (2014). They found that varying the number of distractors before T1 in the pre- and posttest, or the number of distractors between T1 and T2 (the lags) in the training reduced the effect of the colored target training. They suggest that the colored-target training paradigm merely shows that strong expectations about timing can ameliorate the attentional blink indirectly, but cannot eliminate the structural limitations. In addition, Willems, Damsma, Wierda, Taatgen, and Martens (2015) found that in their study the colored target was not necessary for reducing the attentional blink, but training on a single lag was. In both interpretations temporal expectations are a stimulus driven process that arises from the fixed temporal locations of the targets during training.

Temporal expectations may exist from the first stimulus of the processing stream, or the expectation may be relative to the first target. We will focus on the first interpretation of temporal expectations: expectations from the start of the stimulus stream. To test this, the first target in the stream of the pre- and posttest can have three different positions in the stream: item 3, item 5, or item 7. If the temporal expectations created by the colored-target training are indeed relative to the beginning of the stream, we expect to find a larger effect of training for position five, the position that was used in the training.

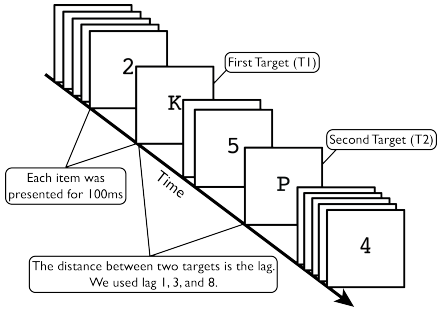


Figure 1: The attentional blink task used in the pretest, the posttest and the control condition.

The Current Experiment

To test these hypotheses, we designed an experiment with different training conditions, but an identical pre- and posttest. The pre- and posttest consist of the standard attentional blink task (see Figure 1). However, to test for temporal expectations from the beginning of the stream, we manipulated the position of T1, so it could appear as either the third, the fifth or the seventh item in the stream. The training conditions we used were a control condition, identical to the pre- and posttest, the colored-target condition, to replicate the effect found by Choi et al. (2012), and the four training manipulations mentioned above: letter-mask, letter-mask 100ms, feedback, and number-word.

Methods

All participants performed three tasks in the experiment:the pretest, a training, and the posttest. The pre- and posttest consist of the RSVP task commonly used in attentional blink experiments. The training task could be one of six tasks, all of which are described in the Tasks section.

Participants

In total 125 people participated in the experiment, 25 participants per training condition. Overall, the mean age was 21.84 (SD: 2.67), 85 of the participants were female. All participants had normal or corrected-to-normal visual acuity. Written informed consent was obtained prior to the experiment. No participants were excluded.

Apparatus and Stimuli

All tasks were presented using the software package E-Prime (Schneider, Eschman, & Zuccolotto, 2002), version 2.0.10.

The stimuli were black on a white background in a bold 12-point Courier New font and consisted of consonants, excluding the "Q", "V", and "Y", and digits, excluding "0" and "1". Viewing distance was approximately 60 cm.

Tasks

The participants did two different tasks in three parts: the attentional blink task in the pre- and posttest and the training task in the training. The training task was different in every condition. In the control condition, participants performed the same task in the training as in the pre- and posttest. The other conditions consisted of the letter-mask task, a variation of the letter-mask task, the attentional blink task with feedback, the colored-target task, and the number-word task.

The Attentional Blink Task The attentional blink task was used in pre- and posttest, and as a training for the control group. This task required the identification of two letter targets amongst a stream of rapidly presented digit distractors. Each item was presented for 100ms. The first target (T1) was presented as either the third, the fifth, or the seventh item in the stream, with an equal number of trials in each group. The second target appeared either immediately following T1 (lag 1), as the third item following T1 (lag 3), or as the eighth item following T1 (lag 8). As mentioned in the introduction, an attentional blink is likely to occur at lag 3 (300ms after T1 onset) and unlikely to occur at lag 1 (100ms after T1) or lag 8 (800ms after T1). Targets were chosen at random from the set of letter stimuli, with the constraint that no letters were repeated within a trial and distractors were randomly chosen from the set of digits such that two successive digits were never identical. No feedback was given.

The Attentional Blink Task with Feedback The attentional blink task with feedback is the same task as the attentional blink task, but feedback is given at the end of each trial. The feedback consists of a number of points per correctly reported target. The participant could obtain 5 points per target, so that a maximum of 10 points per trial could be reached.

The Letter-Mask task The LM task required the identification of only one letter, the target, which was initially presented for 70ms, and followed by a mask: the pound sign #. The target was chosen at random from the set of letter stimuli. To ensure that the task was equally difficult for all participants, the duration of the target was updated on every fourth trial, depending on the accuracy of the participant. If the accuracy was lower than 75%, 3ms were added to the target duration. When the accuracy was above 80%, 3ms were subtracted from the target duration.

The Letter-Mask 100ms task The letter-mask 100ms task is similar to the letter-mask task. The only difference is the presentation time of the stimuli. In the letter-mask 100ms task, every letter was presented for 100ms, the same presentation duration as the stimuli in the attentional blink task. The mask was always presented for 100ms.

The Colored-Target Task The Colored-Target task is based on the training task used by Choi et al. (2012). This is similar to the attentional blink task, but the second target is given a salient red color, and the distribution of the lags is different. In the colored-target task only lag 3 is used.

The Number-Word Task In the number-word task participants are presented with two digits, each for 100ms and followed by a mask (#). The task was to identify the two-digit number that those two digits form. The answer had to be selected from four options and was always the two digit number in words. For example, if the two digits are 5 and 9, the correct answer would be fifty-nine. The other three answer options would always be: first digit and a randomly digit, a randomly selected digit and the second digit, and the two randomly chosen digits used in the previous answers. Answers were presented in a random order. The response had to be given by selecting answer a, b, c, or d, and pressing the corresponding key on the keyboard. Feedback was given in the form of text and points.

Procedure and Design

The experiment was composed of three parts, a pretest in which people performed the attentional blink task, a training in which people practiced on either the letter-mask task, the letter-mask 100ms task, the feedback task, the colored-target task, the number-word task or the control task (the attentional blink task). It ended with a posttest that was identical to the pretest. Before the pretest, a short practice block was given. See Figure 1 for the procedure of the standard attentional blink task. The pretest started with a practice block that was followed by four experimental blocks of 90 trials each. The practice block consisted of 9 trials, one trial for each combination of T1 position (3, 5, or 7) and lag (1, 3, or 8). All experimental blocks contained 90 trials, 10 trials for each combination of jitter and lag. Within blocks the lags were chosen at random.

Prior to each pretest trial, a fixation cross was shown in the middle of the screen, participants were asked to start the trial by pressing space bar. After pressing space bar, the fixation cross stayed on the screen for a duration of 1000ms, followed by a blank screen. After 100ms a stream consisting of 22 stimuli appeared in the center of the screen. Each stimulus was presented for 100ms without inter stimulus interval. After each stream participants were asked to report two letters, by pressing the corresponding keys on the keyboard. They were instructed to press space bar whenever they did not see a target. Although they were encouraged to enter the targets in the order of presentation, responses were counted correct in either order. Participants did not get feedback on their performance.

The training part of the experiment also started with a practice block, followed by four experimental blocks. The practice block consisted of 9 trials, while each of the experimental blocks was 126 trials.

As in the pretest, each trial started with a fixation cross in

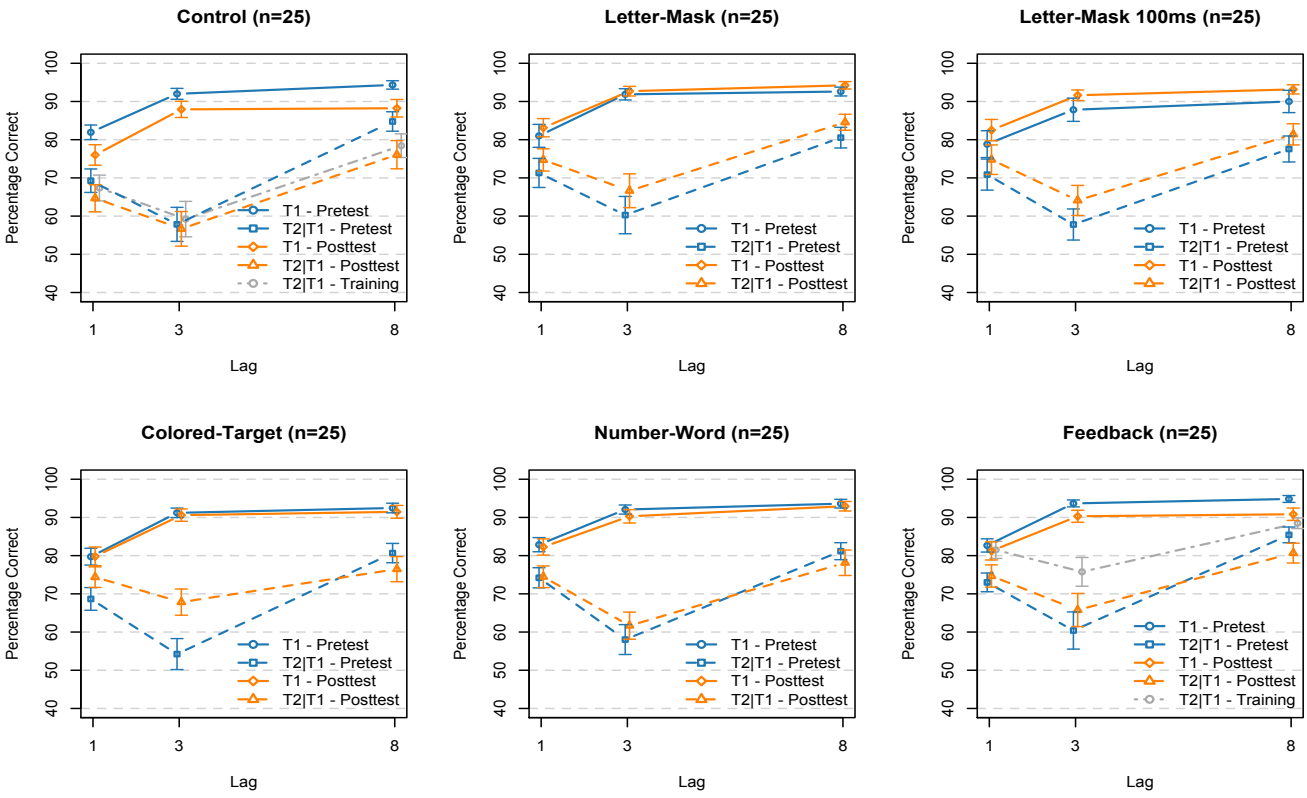


Figure 2: Blink sizes in pre- and posttest for all training conditions and T1 positions. Blue lines show the results for the pretest, while orange lines depict the results for the posttest. Dashed lines are $T2|T1$ results. For the control condition and the feedback condition, mean T2 accuracy during training is shown as a dotted gray line. Error bars depict standard errors.

the middle of the screen. After the fixation cross, either the stimuli stream or the target letter appeared on the screen. Afterwards, participants had to enter the key(s) corresponding to the target. In the letter-mask task, one response had to be given, and participants were instructed to do this as fast as possible. In the number-word task, the response had to be given in the form of a multiple-choice question. In the letter-mask task, the letter-mask 100ms task, the number-word task and the feedback task feedback was given in the form of points: 5 points were awarded for every correct response. The posttest was identical to the pretest, except the practice block. The experiment took approximately two hours to complete.

Results

Results are shown in Figure 2. When applicable, T2 accuracy during training is shown as a gray line. In all training conditions, an attentional blink is present in the pretest (blue lines): performance on T2 is lower on lag 3 than on lag 1 and 8. There seems to be a small reduction in blink size for most training conditions in the posttest (orange lines), and a larger reduction for the colored-target condition.

To establish whether each training regime differs from the control condition, linear mixed effect models (R package lme4) with the dependent factor blink size were fitted on the data, and the independent factors Part (pretest and posttest) and Training (control and the training of interest). The letter-mask and letter-mask 100ms condition are combined in one model, as they test the same question: the processing speed hypothesis. T2 accuracy was only calculated when T1 was correct, this is denoted as $T2|T1$. Blink size is calculated by subtracting the mean $T2|T1$ accuracy on lag 3 from the mean $T2|T1$ accuracy on lag 1 and 8. We used likelihood ratio tests to test for each factor, comparing the full model to the model with that factor removed.

As shown in Table 1, there is a significant effect of Part in all models: the blink is larger in the pretest than in the posttest. However, the colored-target training is the only training condition that shows a decrease in blink size compared to the control condition.

To test whether the temporal expectation hypothesis uses implicit time estimation from the start of the stimulus stream, we have tested the effect of T1 position on the difference in blink size between pre- and posttest for the colored-target and control condition, see Figure 3. The model includes the random factor Subject, and fixed factors Part (pre- and posttest), T1 position (3, 5, and 7), and Training (control and colored-target).

There is a significant effect of Part ($\chi^2(6) = 64.38, p < .01$)

Table 1: Results of the linear mixed-effect models for the decrease in blink size by training, compared to the control condition.

Training	χ^2	Difference in Df	p-value
Colored-target			
Part	60.20	2	$< .01$
Training	16.71	2	$< .01$
Part x Training	16.37	1	$< .01$
Feedback			
Part	23.82	2	$< .01$
Training	2.66	2	$> .1$
Part x Training	2.57	1	$> .1$
Letter-Mask			
& Letter-Mask 100ms			
Part	12.49	3	$< .01$
Training	.61	4	$> .1$
Part x Training	0.10	2	$> .1$
Number-Word			
Part	16.05	2	$< .01$
Training	1.08	2	$> .1$
Part x Training	1.05	1	$> .1$

and of Training ($\chi^2(6) = 18.82, p < .01$). The main effect of T1 position is significant ($\chi^2(8) = 16.29, p = .04$). There is only one interaction effect, between Training and Part ($\chi^2(3) = 17.54, p < .01$). This means that we replicate the effect found in Choi et al. (2012), the colored-training greatly reduces the attentional blink. However, this effect does not seem to be mediated by the position of T1.

Post-hoc testing revealed that the later T1 was presented, the more participants thought to see non-existing targets before T1, so called 'ghost targets'. We tested for these ghost targets by looking at pretest responses where T1 was reported as T2 and T2 was incorrect These ghost-targets were reported more for position 5 than for position 3 ($\beta = -.4, SE = .07, z = -5.66$), and more for position 7 than for position 5 ($\beta = .3, SE = .06, z = 5.30$).

Discussion

First of all, we replicated the results by Choi et al. (2012). After training on the colored-target task, the blink is significantly reduced as compared to the control condition. We have tested several explanation for this phenomenon. Each explanation will be discussed below.

Processing Speed

We tested whether the reduction of the attentional blink could be explained by a speed up in target processing by training participants on the letter-mask task with an adaptive presentation time and the letter-mask task with a constant 100ms presentation time. Training on these tasks did not reduce the blink size, as compared to the control condition. This indicates that the effect of processing speed is either very small, or nonexistent and cannot explain the effect of the colored-target training.

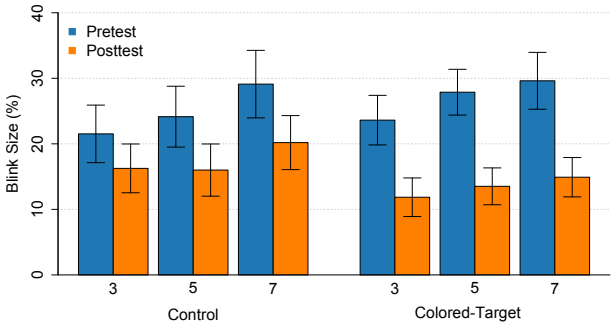


Figure 3: The influence of T1 position on blink size.

Strategy

Our next training condition, the number-word task, tested the influence of a strategy change. Training on the number-word task also did not reduce the attentional blink. Of course, strategy change is very broad, and it is impossible to test all strategies. We can therefore not exclude an influence of strategy in the colored-target training. However, it is clear that this training did not improve the attentional blink in the same way as the colored-target training. There are at least two possible explanations for this. Either the strategy used in the number-word task is too similar to the strategy used in the standard attentional-blink task, or the superficial properties of the task are too different from the attentional blink task. The more dissimilar a training task is, the more training is needed to generalize strategies from that task (Taatgen, 2013). In that case, a one hour training might be too short to find an effect of strategy in this task.

Feedback

Because of the implicit feedback given in Choi et al. (2012), we hypothesized that this feedback might play a role in reducing the attentional blink. We found that feedback during training did not improve performance on the posttest. However, feedback did improve performance during the training task (see the gray lines in Figure 2). This result was surprising, because it is generally assumed that feedback does not influence performance on the attentional blink task (e.g. Martens, Wolters, & van Raamsdonk, 2002). However, the effect of feedback is not a lasting effect. When the feedback is not given anymore, the blink size is still reduced, but this did not differ from the control condition.

Temporal Expectations

We have also tested whether temporal expectations from the start of the stimulus stream influence the decrease in blink size after training. We found that our manipulation of T1 position only marginally impacted performance, but the effect was robust. It was present in all training conditions. A

later T1 position was associated with a larger blink size in the pretest, as well as in the posttest. However, the effect of T1 position does not mediate the effect of training.

The temporal expectation hypothesis was first posited by Tang et al. (2014). They did find an influence of T1 position on the effect of training. When T1 position was varied in pre- and posttest, but not in training, the attentional blink became smaller but was not eliminated.

One difference between their study and our study is the different T1 positions that were used. Whereas we used positions 3, 5, and 7, Tang et al. (2014) used positions 2 to 5. However, they did not look at the influence of T1 position in more detail. We can therefore not directly compare the results found in the current study to the results of Tang et al. (2014).

Although we did not find an effect of temporal expectation on the size of the blink in this experiment, we did find an effect of T1 position on the blink size in general. One possible explanation for this effect could be 'ghost-targets', distractor stimuli that are interpreted as target stimuli. Many of our participants noted that they often saw more than two targets in the stream. Since they could only report two targets in every trial, they may have reported the ghost target instead of the real target. In trials where T1 is presented at a later position, there is a larger possibility of these ghost targets appearing before the real targets, thus increasing the blink.

That we did not find an effect of T1 position does not mean that temporal expectations do not play a role in the colored target training. In the current experiment we have only tested one possible expectation: the time between T2 and the start of the stream. Temporal expectations can also play a role between T1 and T2. Tang et al. (2014) have looked at this by manipulating the training session. When the colored-target training incorporated multiple lags, instead of only one short lag, the effect of training was diminished. Similarly, Willems et al. (2015) found that training the key manipulation in the colored-target training was the single training lag, as opposed to the color of the second target. Taken together, these results indicate that temporal expectations probably play a role in the colored-target training, but that these expectations are relative to the first target and not relative to the start of the stimulus stream. Furthermore, we have shown here that blink size can depend on the position of the first target of the stimulus stream. The later this target is presented, the bigger the blink is. A possible explanation for this effect is that a participant sees 'ghost-targets', while no target is present.

Conclusion

All training manipulations of strategy, feedback, letter-mask, and number-word, give similar results: there is a small increase in T2 accuracy after training, but this increase does not differ from the control condition, while the colored-target training does result in a significantly reduced attentional blink. These results either indicate that none of our training manipulation seem to be a factor in the training-induced reduction of the attentional blink, or the influence of strategy is too small to be distinguished in our sample. Either explanation means that we should look into other explanations for the colored-target training effect.

The most likely candidate for the explanation of the training effect of the colored-target training is the temporal expectation hypothesis (Tang et al., 2014; Willems et al., 2015). However, as opposed to what is expected based on research in time estimation (e.g. Taatgen & van Rijn, 2011), these expectations are not relative to the first stimulus on the screen, but relative to the first task-relevant stimulus on the screen.

Acknowledgements

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen.

References

Choi, H., Chang, L.-H., Shibata, K., Sasaki, Y., & Watanabe, T. (2012). Resetting capacity limitations revealed by long-lasting elimination of attentional blink through training. *Proceedings of the National Academy of Sciences*, *109*(30), 12242–12247.

Ferlazzo, F., Lucido, S., Di Nocera, F., Fagioli, S., & Sdoia, S. (2007). Switching between goals mediates the attentional blink effect. *Experimental Psychology*, *54*(2), 89–98.

Martens, S., Wolters, G., & van Raamsdonk, M. (2002). Blinks of the mind: memory effects of attentional processes. *Journal of experimental psychology. Human perception and performance*, *28*(6), 1275–1287.

Olivers, C. N. L., & Nieuwenhuis, S. (2005). The beneficial effect of concurrent task-irrelevant mental activity on temporal attention. *Psychological Science*, *16*(4), 265–269.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849–860.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime: User's guide*. Psychology Software Incorporated.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review*, *120*(3), 439–471.

Taatgen, N. A., Juvina, I., Schipper, M., Borst, J. P., & Martens, S. (2009). Too much control can hurt: A threaded cognition model of the attentional blink. *Cognitive Psychology*, *59*(1), 1–29.

Tang, M. F., Badcock, D. R., & Visser, T. a. W. (2014). Training and the attentional blink: limits overcome or expectations raised? *Psychonomic bulletin & review*, *21*(2), 406–11.

Willems, C., Damsma, A., Wierda, S. M., Taatgen, N., & Martens, S. (2015). Training-induced Changes in the Dynamics of Attention as Reflected in Pupil Dilation. *Journal of Cognitive Neuroscience*, *27*(6), 1161–1171.

Allocation of attention during auditory word learning

Keith S. Apfelbaum (apfelbaum.3@osu.edu)
Department of Psychology, 1835 Neil Ave.
Columbus, OH 43210 USA

Vladimir Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, 1835 Neil Ave.
Columbus, OH 43210 USA

Abstract

The deployment of selective attention has been studied in depth as a mechanism of visual categorization for decades. However, little work has investigated how attentional mechanisms operate for non-visual domains, and many models of categorization tacitly presume domain-general attention use. In three experiments, we investigated whether learners deploy attention to novel auditory features when learning novel words in a similar fashion to the prevailing visual categorization findings. These studies yielded evidence of non-isomorphism, as selective attention in the auditory domain shows high context specificity, in contrast to the wide generalization of attention in the visual domain.

Keywords: selective attention; auditory attention; categorization; category learning; word learning

Introduction

Selective attention plays a crucial role in efficient categorization. Selective attention allows an organism to focus on features that reliably differentiate items from different categories, and ignore irrelevant features. By using only relevant features, organisms can efficiently classify items (i.e., decide what category the items belong to) without interference from irrelevant information. Classic findings in category learning show that adult learners readily and flexibly apply selective attention to learn category structures (Shepard, Hovland, & Jenkins, 1961). Attending selectively to features that are more predictive of category membership drastically improves learning compared to relying on all features equally. Such effects are thought to occur regardless of the type of feature; as long as features are not integral, learners readily show selective attention to the features (Best, Yim, & Sloutsky, 2013; Kruschke, 2003).

Selective attention's benefit for category learning is intuitive: in any category learning task where some features are relevant and others are irrelevant, a learner is best served by ignoring the irrelevant ones. However, despite this general benefit of selective attention, attention could be deployed in any number of ways: it could operate by changing attention to entire dimensions, or by shifting attention to regions within a dimension; it could be a finite resource, such that increased attention to one dimension necessitates decreased attention to another, or dimensions could be independent; it could readily generalize across contexts, or be tightly linked to contexts; and so on. Critically, these differences could emerge across sensory domains – for example, visual category learning may differ from auditory lexical categories. The goal of research presented here is to test the isomorphism of mechanisms of categorization across modalities by examining how selective attention operates in auditory lexical categorization.

The bulk of research on selective attention has relied on visual stimuli. Although these studies are extremely informative, they leave a critical question of how categorization proceeds in non-visual domains. For example, little work has investigated selective attention in auditory learning of phonological or lexical categories. Understanding selective attention in this domain is critical to elucidating the nature of categorization. Isomorphism of selective attention would rapidly broaden our understanding of the processes of auditory categorization. Studying attention in the visual domain is simpler in many respects, such as the ability to explicitly measure attention through eye-tracking (Rehder & Hoffman, 2005). Similarly use of auditory tasks, such as measuring auditory processing in young children during sleep, could help answer visual categorization questions. However, a lack of isomorphism would suggest a need for in depth investigations of auditory attention, akin to those in the visual domain, to understand points of demarcation between domains, and to describe the mechanisms of auditory category learning.

Coarse grains of analysis suggest parallels between attention in the two domains. Language learners perform a type of dimensional attention to emphasize relevant dimensions and down-weight irrelevant ones (Apfelbaum & McMurray, 2015; Francis & Nusbaum, 2002; Toscano & McMurray, 2010), akin to dimensional weighting in visual categorization tasks (e.g., Nosofsky, 1986). Additionally, language development may pass from early, holistic representations of multiple stimulus features to more selective, phonetic-based representations (Werker & Curtin, 2005), parallel to the shift from distributed to selective attention in visual category learning (Sloutsky, 2010).

However, several domain differences may result in differential use of attention. Visual attention can rely on physical reorientation of the eyes to filter out irrelevant visual features; the auditory system lacks such physical means of selection. Additionally, auditory lexical categories rely on highly overlearned phonetic features; the same small set of features is used to differentiate all words throughout the lifespan of a monolingual speaker; visual categories include a wider variety of features, and sometimes even necessitate creation of novel features on the fly (Schyns,

Goldstone, & Thibaut, 1998). Finally, the temporal nature of auditory stimuli may encourage flexible attention deployment; static visual stimuli allow resampling information after initial filtering, whereas ignored auditory information may be unavailable if it later proves relevant.

The nature of selective attention use for auditory lexical categories thus remains an open empirical question. Although some work has demonstrated that selective attention occurs within this domain (Francis & Nusbaum, 2002; Holt & Lotto, 2010; Idemaru & Holt, 2011), no studies have examined if this attention is accomplished in the same manner as in visual categorization. The present experiments offer initial insight into the validity of an amodal theory of selective attention that subserves both visual and auditory-lexical category learning.

Experiment 1

Selective attention during category learning is signaled by increased attention to one dimension accompanied by decreased attention to other dimensions. Measuring the deployment of selective attention relies on tracking either how learners attend to features during processing directly, such as with eye-tracking (Rehder & Hoffman, 2005), or inferring attention from performance measures, such as learning curves or discrimination thresholds (Kruschke, 2003). A challenge for gauging isomorphism across domains is using a task that allows comparable dependent variables; although eye-tracking allows direct measure of visual selective attention, no comparable direct measure of attention in the auditory domain is possible.

Additionally, the features themselves are not directly comparable. Auditory features operate differently (e.g. their temporal nature) and may be processed fundamentally differently. Attempting to equate features across domains is speculative at best. Instead, we emphasize qualitative comparisons: patterns of selective attention are consistent across different visual features, so differences that arise in our selection of auditory features would signal a diversion from the constant visual effects.

We thus used an inferential approach to gauging auditory selective attention, allowing ready comparison with patterns seen previously in the visual domain. Visual studies typically show rapid optimization to relevant category features, and concomitant decreased attention to irrelevant dimensions (Best, Yim, & Sloutsky, 2013). This pattern exhibits highlighting or down-weighting entire dimensions; when attention is allocated to a dimension in one context or one task, this pattern generalizes across contexts and values of the dimensions (Best et al., 2013; Kruschke, 2003).

We adapted the learned inattention paradigm to study selective attention allocation in the auditory lexical domain. We measured how discrimination thresholds for novel phonological contrasts changed based on their relevance for an auditory category learning task. Specifically, we sought evidence that learners improve discrimination of contrasts that are relevant to category learning, and simultaneously decline in discrimination of category irrelevant contrasts.

Methods

Participants Data from 56 participants were included in analysis. The participants received partial course credit for a psychology course at Ohio State University. All reported being native speakers of English, and all had normal hearing and normal or corrected-to-normal vision. An additional 14 participants failed to complete the study due to computer error, exceeding time limitations to complete the task, or failure to learn the categories within 140 trials.

Stimuli Two 11-step continua were generated based on a single recording of the non-word *dubo*. The first continuum derived from adjusting the duration of the first vowel in *dubo*; the second consisted of different pitch patterns on the second vowel. These features were selected because they are not primary cues to phonetic distinctions in English; learners are more likely to show changes in attention to these features than primary phonetic features with which they have a lifetime of experience.

The continua were generated by excising the vowels from the recording, and then manipulating them to create equal-spaced steps before re-splicing the vowels back into the original recordings. The vowel-duration continuum on the first vowel consisted of durations ranging from 66 ms to 152 ms in approximately 8.5 ms steps. For the pitch continuum on the second vowel, all steps had a vowel-initial pitch of 184 ms. At the lowest step, this pitch slowly fell to 130 Hz; at the highest step it rose to 233 Hz. Pilot testing showed that these step sizes resulted in approximately equal discrimination. We fully crossed the two continua, creating 121 unique stimuli (11 vowel-one durations X 11 vowel-two pitches). The stimuli were visually and auditorily inspected to ensure that they sounded natural and were free of artifacts from the acoustic manipulations.

Procedure The procedure used a pre-test, training, post-test design. Pre- and post-test were identical; these segments measured participants' discrimination thresholds separately for each contrast to determine whether discrimination was affected by categorization training. Specifically, we measured whether discrimination improved for a contrast that was relevant for categorization and simultaneously decreased for a contrast that was irrelevant. For pre-and post-test trials, participants were required to identify whether two stimuli were acoustically identical or different. On *different* trials, stimuli were chosen using an adaptive staircase procedure to estimate discrimination thresholds. Same and different trials were randomly interleaved. Trials for each staircase continued until the participant (1) reversed accuracy on different trials four times; (2) successfully discriminated the smallest possible distinction (i.e. a one-step difference); or (3) failed to discriminate the largest possible distinction.

Discrimination thresholds of the two vowel contrasts were tested independently; when testing discrimination of the duration contrast, the pitch contrast was held constant, so discrimination could only emerge on the basis of a single contrast. For each contrast, discrimination was measured

from the lowest step and from the highest step; the discrimination threshold for that contrast was the average estimated threshold from each of these steps. There were thus four separate estimated thresholds (two each for the duration and the pitch contrast). The staircases for the four threshold estimations were randomly interleaved.

Training consisted of teaching participants two categories defined by one of the novel features. One of the two contrasts was selected as relevant to category identity, and the other was irrelevant. Assignment of which contrast was relevant was counterbalanced across participants: 28 participants had vowel-one duration relevant and the other 28 had vowel-two pitch relevant. Participants were informed that they were going to learn two words from a language that uses different sounds than English. Each trial, they saw two visually distinct cartoon alien creatures and heard a stimulus; they clicked on the alien they believed the word identified, and then received feedback.

The relevant dimension determined which alien was correct; one alien was associated with the lowest step on this dimension, and the other with the highest step. No other steps of the relevant contrast were presented. The step for the irrelevant dimension was randomly selected on each trial, so this dimension was unpredictive of category membership. Participants continued receiving trials until they reached 80% accuracy across a 40-trial window. If they failed to reach this criterion in 140 trials, they were removed from the experiment and their data were excluded.

After categorization training, participants completed the post-test. This test was conducted exactly as in the pre-test, allowing comparison of pre-training thresholds with those after one dimension was made relevant for categorization.

Results and discussion

We first analyzed performance during categorization trials. Participants quite quickly reached the accuracy criterion, with a mean trials-to-criterion of 47.2 (*SD*=19.0). Thirty-nine of the 56 participants reached criterion by the 40th trial. Participants for whom duration was relevant reached criterion slightly more quickly than those for whom pitch was relevant (*M*=42.2 and 52.1, respectively; *t*(54)=-1.99, *p*=.052), suggesting that the pitch contrast may have been slightly more difficult than the duration contrast.

The critical analysis of discrimination thresholds showed a classic pattern of learned inattention (Figure 1); discrimination of the relevant contrast became more sensitive after training (as indicated by decreased discrimination thresholds), and the irrelevant contrast showed poorer discrimination after training. We analyzed these data using mixed effects models (ANOVA yielded similar results). The DV was step-size of the discrimination threshold for the contrast as estimated from the staircase procedure. We contrast-coded feature type (irrelevant: -.5; relevant: +.5) and test-type (pre-test: -.5; post-test: +.5) for use as fixed factors. Participant was included as a random intercept, as was the slope of relevance for participants. Including which feature was relevant as a random intercept did not improve model fit by χ^2 test (*p*=.31), so this was not included in the analyses; including this random effect produced extremely similar results.

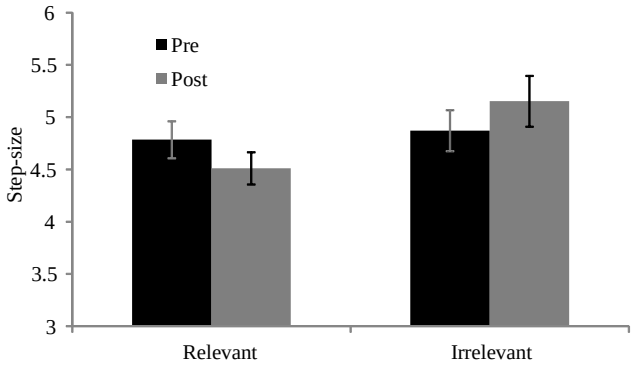


Figure 1. Mean step size of reversals as a function of relevance and test-type. Error bars represent the standard error of the mean.

Results confirmed the pattern seen in Figure 1. Specifically, there was no main effect of feature type (*B*=-.363, *SE*=.299, *t*(55)=-1.22, *p*=.23) nor of test type (*B*=.0011, *SE*=.082, *t*(168)=.014, *p*=.99). However, there was a significant interaction of feature type and test type (*B*=-.56, *SE*=.16, *t*(168)=-3.43, *p*<.001). Simple effects testing showed a significant decrease in discrimination threshold for the relevant feature (*B*=-.28, *SE*=.11, *t*(839)=-2.59, *p*=.0098), and a significant increase in discrimination threshold for the irrelevant feature (*B*=.28, *SE*=.12, *t*(839)=2.30, *p*=.022).

The results present evidence for selective attention in learning auditory lexical categories defined by novel features. After training, participants' ability to recognize small acoustic changes in a feature that had been relevant for categorization improved, whereas their ability to recognize changes in a second feature that was irrelevant declined. This pattern of results mirrors patterns from visual category learning (Best et al., 2013; Dopson, Esber, & Pearce, 2010; Hoffman & Rehder, 2010; Kruschke & Blair, 2000; Mackintosh & Little, 1969).

Although these results are suggestive of isomorphism of selective attention across domains, evidence of generalization is necessary before concluding that isomorphism exists. Findings in the visual domain demonstrate attention to entire dimensions; when a dimension is attended, processing benefits persist in later tasks, even for new values of the dimension and new contexts. Experiment 1 used identical stimuli during training and test – we tested the same values of the same features in the same contexts. To further investigate selective attention in the auditory lexical domain, we considered two types of generalization. Experiment 2 examined generalization of selective attention to a novel voice (a change in irrelevant context). Experiment 3 examined generalization to a novel word (a change in relevant context).

Experiment 2

Experiment 2 followed the design of Experiment 1, but incorporated a change in irrelevant context. Specifically, training was done in one voice, whereas testing was conducted in a different voice. The same words and features were used in the two segments, so learning could readily generalize if learners deploy attention to abstract dimensions. However, if selective attention operates in a context-specific manner, generalization may be unattested.

Methods

Participants Data from 81 participants were included in the analysis of Experiment 2. Thirty-seven received male-voice stimuli during pre- and post-test, and female-voice stimuli during training; the remaining 44 received the reverse. Six additional participants were excluded from analysis for exceeding time limitations to complete the task, or failure to learn the categories within 140 trials.

Stimuli The female-voice stimuli were identical to those in Experiment 1. For the male-voice stimuli, we acoustically manipulated the female-voice stimuli to approximate a male voice. We used the Change Gender option in Praat, which scales the formant values in a stimulus. We found that a ratio of 1/1.1 altered the percept of talker gender without impacting clarity of the stimuli. We thus created a male-voice version of every stimulus from Experiment 1.

Although this method is less natural than recording two different talkers, it allows close control of the manipulated features. The Change Gender function does not impact duration or overall pitch, so the stimuli in the two voices maintained identical values of the features despite the percept of different voices. The distribution of cues was thus identical between the two stimulus sets; learners can thus generalize even on the basis of specific feature values, but must do so in a novel context.

Procedure The procedure was identical to Experiment 1, except that the voice used in pre- and post-test was the opposite gender as that used in training.

Results and discussion

Analysis of the training portion of the study again showed very rapid learning (M=45.8; 64 of 81 participants reached criterion by trial 40). No difference in speed of learning was seen between the voices used in training.

The discrimination thresholds showed sensitivity to category relevance, as the threshold decreased for the relevant feature (Figure 2). However, no concomitant increase in threshold was seen for the irrelevant feature; this feature also showed a (very small) decrease. We analyzed these data using a mixed effects model with similar structure to the one used in Experiment 1. In addition to the random effects used for that model, we added training voice as a random intercept, and a random slope of relevance for training voice. Including training voice instead as a fixed factor yielded extremely similar results, with no main effects or interactions of training voice.

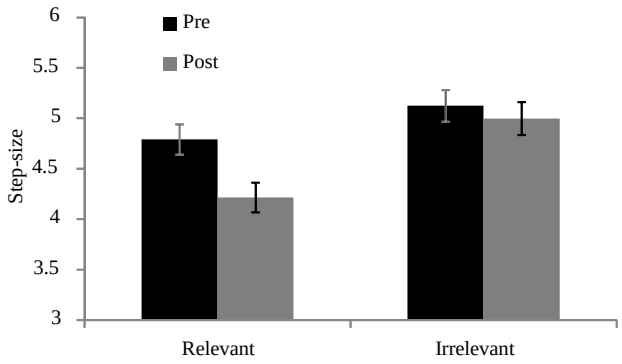


Figure 2. Mean step size of reversals as a function of relevance and test-type. Error bars represent the standard error of the mean.

This analysis revealed a main effect for test type (B=-.35, SE=.069, t(2460.3)=-5.05, p<.0001), with lower thresholds at post-test than at pre-test. There was no effect of relevance (B=-.50, SE=.28, t(6.0)=-1.82, p=.12). However, there was a significant two-way interaction of test type and relevance (B=-.45, SE=.14, t(2460.3)=-3.24, p=.0012). Simple effects analyses used to investigate this interaction showed a significant effect of test type for the relevant contrast (B=-.57, SE=.094, t(1230.0)=-6.10, p<.0001); thresholds were lower at post-test than they were at pre-test. However, the model for the irrelevant feature showed no effect of test type (B=-.13, SE=.10, t(1230.3)=-1.23, p=.22); performance for the irrelevant feature was similar at pre-test and post-test.

These results differ from classic findings of learned inattention. Although the relevant feature showed evidence of improved processing, this was not accompanied by the expected drop in performance for the unattended feature. This finding is quite surprising; nearly all models of attention argue for a limited quantity of attention, such that attention to one dimension necessitates decreased attention to others. Our results violated this coupling, suggesting that attention in the auditory lexical domain may not engage the same coupling across dimensions.

Experiment 2 thus provided only partial evidence of isomorphism across domains. Learners generalized highlighting of a relevant dimension across talker voices, but did not simultaneously generalize learned inattention to the irrelevant dimension. This pattern suggests that attention in the auditory lexical domain may operate in a more context-specific manner than in the visual domain.

Experiment 3

Experiment 3 examined generalization more thoroughly, by investigating generalization across a relevant context. Specifically, training and testing used the same features, but embedded in different words. During training, vowel duration and pitch contour were manipulated in one word (e.g. *dubo*), whereas testing employed these same features in a second word (e.g. *bugo*). This more directly measures

whether learners are deploying attention to abstract dimensions; for example, if they are learning about vowel duration in general, patterns of attention should generalize across lexical contexts. However, if attention is deployed in a context-sensitive manner, then attention to one word learned in training may not generalize to other words at test.

Methods

Participants Data from 60 participants were included in the final analysis. Thirty participants were included in each training-word condition (*dubo* train/*bugo* test or vice versa). Data from five additional participants were not included in the final sample for exceeding time limitations to complete the task, or failure to learn the categories within 140 trials.

Stimuli The stimuli included the original stimuli from Experiment 1 using *dubo*, as well as new stimuli constructed using the same features and methods of generating these features, based on the word *bugo*. The new base word was recorded by the same talker in the same session. This second word includes the same vowels in similar phonological contexts (surrounded by voiced obstruents). As such, although generalization must occur across words, the contexts are quite similar, and the values of the features were quite similar. If generalization is unattested across these similar words and feature values, it is unlikely to arise across the much broader diversity of words in the language.

Procedure The procedure was identical to Experiment 1, except that different words were used in pre-/post-test and training.

Results and discussion

Analysis of the training data again showed that participants rapidly reached the accuracy criterion (*M*=48.4 trials; 42 of 60 participants reached criterion by the 40th trial).

Analysis of the critical pre- and post-test discrimination thresholds showed marked improvement in discrimination for *both* relevant and irrelevant contrasts (Figure 3). Statistical analyses were conducted using mixed effects models with similar structures to those in the prior experiments, but adding a random intercept of word used during training and a random slope of training word for which feature was relevant. Including word as a fixed factor instead yielded similar results, with no significant effects or interactions of word.

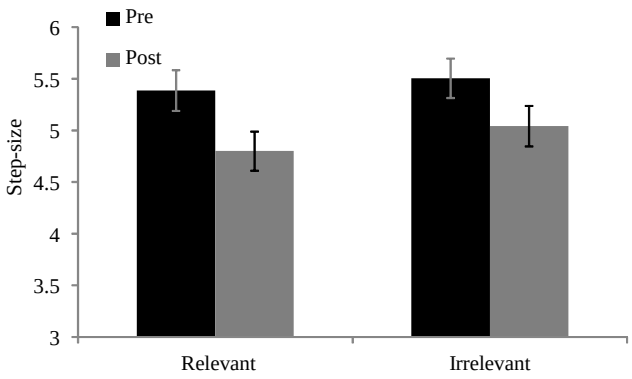


Figure 3. Mean step size of reversals as a function of relevance and test-type. Error bars represent the standard error of the mean.

The model revealed a main effect of test type (*B*=-.53, *SE*=.085, *t*(1798.0)=-6.24, *p*<.0001), as discrimination thresholds declined from pre- to post-test. However, there was no effect of relevance (*B*=-.17, *SE*=.47, *t*(1.2)=-.37, *p*=.76), nor was the interaction significant (*B*=-.14, *SE*=.17, *t*(1798.0)=-.85, *p*=.40). The relevance of the features during training had no effect on discrimination performance, suggesting that category learning did not result in generalization of selective attention. Instead, performance improved similarly for relevant and irrelevant features.

These results demonstrate no generalization of selective across lexical contexts. Instead, sensitivity improves for both relevant and irrelevant features (perhaps because of learning about the task). Whereas visual category learning tasks show generalization of selective attention across contexts (e.g., Best et al., 2013; Kruschke, 2003), the auditory lexical domain seems to display much stronger context specificity.

General Discussion

We examined the degree to which auditory lexical category learning exhibits isomorphic use of selective attention to visual category learning. Although we found evidence that selective attention aids learning of auditory lexical categories, its deployment proved more context sensitive than for visual categories. Learners in Experiment 1 highlighted relevant features and decreased attention to irrelevant features; however, this pattern only weakly generalized to a novel voice, and showed no apparent generalization to a novel word. Visual category learning shows ready generalization; in fact, visual category learning studies often use training in one context, and then measure continued selective attention in a new context (Best et al., 2013; Kruschke, 2003). This suggests a clear demarcation between the two domains.

Phonological information seems to generalize eventually; when learning novel words, learners bring prior learning about feature relevance to bear. If generalization were never attested, learners would have to learn for each word that features like voice are irrelevant. Yet our results show that at least initially, learners treat novel features with high context specificity. This specificity is similar to developmental evidence that infants may treat words as functionally independent, without generalizing features (Apfelbaum & McMurray, 2011; Werker & Curtin, 2005). However, these similar patterns of specificity among adults are surprising. In the visual domain, even for novel features learners exhibit patterns of learned inattention (Best et al., 2013). Generalized selective attention across contexts seems the default in vision. For auditory lexical categories, it appears to be non-default. This finding may lend credence

to theories of voice-specific priming, in which words are obligatorily processed with indexical information as part of the representation (Papesh, Goldinger, & Hout, 2016). Similar results suggest that talker information and phoneme information are asymmetrically related because of talker normalization processes (Mullennix & Pisoni, 1990); talker information may be difficult to ignore to focus on abstract features because these features are often altered by talker characteristics. However, these talker-specificity effects are less straightforward as an explanation of the lack of generalization across lexical contexts.

This context specificity may arise because of the closed-set nature of features used to differentiate lexical items. Languages use a relatively small set of phonemes to differentiate an enormous set of words. Listeners receive copious exposure to this small set, allowing learning of the relevant features to become quite entrenched. Processing features differently (as in the present study) may prove quite difficult. Indeed, inflexibility of processing speech features may be why learning the phonology of a second language is markedly difficult for adults (Jusczyk, 1993).

Yet even if such an explanation holds for why learners show poor generalization across contexts in these experiments, it nonetheless represents a departure from processing in the visual domain. Auditory lexical category learning may rely on a core set of inflexible features, whereas visual category learning is flexible in the face of novel features. Selective attention in the auditory lexical domain would thus operate within a constrained context for novel features, and only generalize for known features, instead of generalizing readily. As such, selective attention across these domains shows a lack of isomorphism; the way attention operates depends on the domain.

References

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–38. doi:10.1111/j.1551-6709.2011.01181.x

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, *22*(4), 916–943. doi:10.3758/s13423-014-0783-2

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: developmental differences between adults and infants. *Journal of Experimental Child Psychology*, *116*(2), 105–19. doi:10.1016/j.jecp.2013.05.002

Deng, W., & Sloutsky, V. M. (2015). The Development of Categorization: Effects of Classification and Inference Training on Category Representation. *Developmental Psychology*, *51*(3), 392–405.

Dopson, J. C., Esber, G. R., & Pearce, J. M. (2010). Differences in the associability of relevant and irrelevant stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(2), 258–267. doi:10.1037/a0016588

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319–340. doi:10.1037/a0019042

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227. doi:10.3758/APP

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–56. doi:10.1037/a0025641

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, *21*, 3–28.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175. doi:10.1111/1467-8721.01254

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645. doi:10.3758/BF03213001

Mackintosh, N. J., & Little, L. (1969). Intradimensional and extradimensional shift learning by pigeons. *Psychonomic Science*, *14*(1), 5–6.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–61.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41. doi:10.1016/j.cogpsych.2004.11.001

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *The Behavioral and Brain Sciences*, *21*(1), 1–54.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. doi:10.1037/h0093825

Sloutsky, V. M. (2010). From Perceptual Categories to Concepts: What Develops? *Cognitive Science*, *34*(7), 1244–1286. doi:10.1111/j.1551-6709.2010.01129.x

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. doi:10.1111/j.1551-6709.2009.01077.x

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234.

Examining Cardiac and Behavioral Responses in a Modality Dominance Task

Christopher W. Robinson (robinson.777@osu.edu)
Krysten R. Chadwick (chadwick.63@osu.edu)
Jessica L. Parker (parker.1026@osu.edu)
Department of Psychology, The Ohio State University at Newark
1179 University Dr, Newark, OH 43056, USA
Scott Sinnett (ssinnett@hawaii.edu)
Department of Psychology, University of Hawaii at Manoa
2530 Dole Street, Honolulu, HI 96822, USA

Abstract

The current study examined cardiac and behavioral responses to changing auditory and visual information while using modified oddball tasks. When instructed to press the same button for auditory and visual oddballs, auditory dominance was found with cross-modal presentation slowing down visual response times and decreasing visual accuracy. When instructed to make separate responses to auditory and visual oddballs, visual dominance was found with cross-modal presentation slowing down response times and decreasing auditory accuracy. However, examination of cardiac responses that were time-locked to stimulus onset show cross-modal facilitation effects, with discrimination of oddballs and standards occurring earlier in the course of processing in the cross-modal condition than in the unimodal conditions. These findings shed light on potential mechanisms underlying modality dominance effects and have implications on tasks that require simultaneous processing of auditory and visual information.

Keywords: Cross-modal processing; Sensory Dominance; Attention.

Introduction

While most of our experiences are multisensory in nature, historically, most research has focused on processing within a single sensory modality. Over the last 40 years there has been a growing body of research examining how sensory systems process and integrate incoming information (see for example Driver & Spence, 2004; Posner, Nissen, & Klein, 1976; Spence, 2009; Wickens, 1984), with some multisensory environments facilitating learning (e.g., Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Massaro, 1998) and others attenuating learning (Sloutsky & Napolitano, 2003; see also Robinson & Sloutsky, 2010a for a review). For example, intersensory redundancy, when the same information can be conveyed in different sensory systems, can often facilitate learning and speed up responses (Bahrick & Lickliter, 2000; Giard & Peronnet, 1999). However, in many situations, information presented to one sensory modality is irrelevant or may even conflict with information presented to a different sensory system. In these latter situations, modality dominance effects can be observed, with one modality attenuating encoding and/or responding to information in the other modality (see Spence, Parise, & Chen, 2012 for a review).

One commonly used paradigm to study modality dominance is the Colavita visual dominance task (Colavita, 1974; Colavita, Tomko, & Weisberg, 1976; Colavita & Weisberg, 1979; Egeth & Sager, 1977). In this task, participants are presented with auditory or visual information and instructed to quickly respond by pressing one button when they hear an auditory stimulus and a different button when they see a visual stimulus. On a small percentage of trials, the auditory and visual stimuli are presented at the same time. Participants often miss these cross-modal trials by only pressing the visual button, as opposed to pressing both buttons, or a third button associated with a cross-modal stimulus (see Sinnett, Spence, & Soto-Faraco, 2007). Research using variations of this task consistently points to visual dominance, with most of the sensory and attentional manipulations weakening but not reversing the effect (but see Ngo, Cadieux, Sinnett, Soto-Faraco & Spence, 2011). While visual dominance effects are robust and well-studied, underlying mechanisms are poorly understood (see Spence et al., 2012 for a review).

The current study expands on this literature in several important ways. Robinson, Chandra, and Sinnett (2016) recently demonstrated that it is possible to reverse modality dominance in an oddball paradigm by having participants make the same response to auditory and visual oddballs. More specifically, participants were repeatedly presented with the same sound, picture, or sound-picture pairing, and were required to inhibit responses to this stimulus (standard). They were also instructed to press a button on a keyboard as quickly as possible if the picture, sound, or both the picture and sound changed (visual, auditory, or cross-modal oddballs, respectively). While pairing the pictures and sounds together slowed down response times to visual oddballs, it often had no negative effect on auditory processing (i.e., response times to auditory oddballs did not differ when presented with or without the visual standard). The first goal of the current study was to test the generalizability of this finding by using a slightly different procedure with more salient and familiar visual stimuli. It is possible that auditory dominance was found because the visual stimuli used in Robinson et al. were monochromatic, unfamiliar images.

The second goal was to examine if auditory and visual dominance effects can be modulated by top-down attentional control, or if instead proceed with attention having no effect. While attentional manipulations often fail to reverse auditory and visual dominance (Napolitano &

Sloutsky, 2004; Ngo, Sinnett, Soto-Faraco, & Spence, 2010; Sinnett et al., 2007), it is possible that individual differences in attentional control modulate or reverse the effect. For example, one proposed mechanism underlying visual dominance is that participants strategically bias their attention in favor of visual input to compensate for the low alerting properties of visual input (Posner et al., 1976). Thus, it is possible that participants who are better at selectively attending to visual input are more likely to show visual dominance effects and/or less likely to be distracted by conflicting auditory information. At the same time, it has also been argued that auditory dominance may stem from auditory stimuli *automatically* grabbing attention and attenuating or delaying visual processing (Robinson & Sloutsky, 2010a). If auditory dominance effects stem from auditory stimuli automatically engaging attention, then individuals with high or low attentional control may show the same pattern of results. To examine effects of attentional control on modality dominance, we collected individual differences in resting Heart Rate Variability (HRV) prior to the experiment. The underlying idea is that the prefrontal cortex plays a significant role in executive functions such as selective attention and emotional regulation. HRV may serve as a proxy for individual variability in executive functions because parasympathetic activity adds short term variability to the heart beat via the vagus nerve and participants with higher HRV often perform better on a variety of executive function tasks (Hansen, Johnson, & Thayer, 2003; Thayer & Lane, 2000). They may also show a different pattern of results on modality dominance tasks.

The final goal of the current research was to examine real-time psychophysiological responses to changing auditory and visual information to possibly gain insight into the time course of cross-modal interference effects. It is well documented that infants' heart rates slow down when actively processing visual information (see Richards & Casey, 1992 for a review), and using a modified oddball task, heart rate also appears to slow down to novel, less frequent sounds than to more frequent sounds (Robinson & Sloutsky, 2010b). By time-locking heart rate with the onset of standard and oddball items, the current study examined the feasibility of using changes in adults' cardiac responses as a measure of auditory, visual, and cross-modal processing. It was hypothesized that cardiac responses would differ for standards and oddballs; thus, potentially providing an additional measure of discrimination. We also examined if comparing cardiac responses to auditory and visual oddballs when presented cross-modally with the respective unimodal baselines would provide converging evidence of modality dominance effects.

Experiment 1

Experiment 1 employed a cross-modal oddball task and participants made the same response to auditory and visual oddballs. It was hypothesized that cross-modal presentation would have a greater cost on visual processing.

Method

Participants Thirty-eight adults (23 Females, M = 19.1 years) participated in Experiment 1. Participants were undergraduate students at The Ohio State University Newark who received course credit in exchange for participation.

Apparatus A Dell Latitude E6430 laptop computer with DirectRT software was used for stimulus presentation and to record response times and accuracies. Visual stimuli were presented on a Dell P2212hB monitor and auditory stimuli were presented via Kensington 33137 headphones at approximately 65 dB. A Dell Latitude E6430 laptop computer with Mindware software was used to record electrocardiograms. Two Ag-AgCl electrodes were placed on the participants' right collarbone and left lower rib, and a reference electrode was placed on the participants' right lower rib. Electrocardiograms were collected using a BioNex acquisition unit with a BioNex Impedance Cardiograph and GSC amplifier. DirectRT on the stimulus presentation laptop sent event markers to Bionex; thus, time-locking electrocardiograms with stimulus presentation.

Materials The stimulus pool consisted of five visual and five auditory stimuli. Visual stimuli (see Figure 1) were approximately 400 x 400 pixels and pulsated centrally on a computer monitor for 750 ms, with a random 600-900 ms Inter-Stimulus Interval (ISI). The auditory stimuli consisted of bear, frog, elephant, cat, and dog sounds, which were taken from Marcell et al. (2000) and were shortened to 750 ms using Audacity software. As in basic oddball paradigms, one stimulus was frequently presented (approximately 90%, standard) and other stimuli were less frequent (approximately 10%, oddballs). The standard was a dog bark, an image of a dog, or the dog and bark were paired together. The auditory and visual oddballs were an elephant, frog, cat, and bear.



Figure 1. Visual stimuli used in Experiments 1 and 2

Procedure The study consisted of four phases. In the first phase, participants sat still for five minutes while the computer recorded resting HRV. Participants then completed three different oddball tasks on the computer, while their heart rate was monitored. The current study deviated from traditional oddball paradigms in that a trial was defined as a series of stimuli with either a standard or oddball at the end of the series (e.g., 5 standards → oddball), as opposed to each stimulus being a trial. This manipulation gave the heart at least 6 s to respond to an oddball before encountering another oddball (assuming two short oddball sequences were presented back to back).

In the auditory oddball condition, there were 16 standard trials and 16 oddball trials. On auditory oddball trials, a dog bark was presented either four or five times, followed by

one of the other animal sounds (oddball). On auditory standard trials, participants heard four or five dog barks, followed by another dog bark (standard). Half of the trials consisted of four standards followed by a standard or oddball, and the remaining trials consisted of five standards followed by a standard or oddball. DirectRT sent an event marker to Bionex at the onset of the last standard or oddball in each trial. The unimodal visual condition was similar to the auditory condition, with the exception that standard and oddballs were pictures, not sounds. Thus, for each condition, we measured how quickly participants pressed a button when they encountered an oddball and how quickly the heart differentiated standards and oddballs.

In the cross-modal condition, the trials consisted of four or five standard image-sound pairs (dog-dog bark) followed by another image-sound pair that was either a standard or an oddball. Each participant had a total of 96 trials (48 standard and 48 oddball). Sixteen of the oddball trials had only a visual change (visual oddball), while an additional 16 trials only had an auditory change (auditory oddball). Lastly, there were also 16 double oddball trials, where both auditory and visual stimuli changed. As in the unimodal conditions, each stimulus was presented for 750 milliseconds, with a 600-900 ms ISI.

Results and Discussion

Behavioral Analyses Overall, participants correctly reported when the auditory component changed and when both modalities changed (hit rate > .99 across both unimodal and cross-modal conditions). However, cross-modal presentation interfered with visual oddball detection, with visual hit rate in the unimodal visual condition (M = .99) exceeding the cross-modal condition (M = .95), t (37) = 2.05, p = .048, suggesting that cross-modal presentation attenuated responding to visual but not auditory oddballs.

Additional analyses focused on response times in the cross-modal condition when only the auditory or visual component changed, and these response times were compared to the respective unimodal baselines. A 2 (Modality: Auditory vs. Visual) x 2 (Presentation Mode: Unimodal vs. Cross-modal) repeated measures ANOVA revealed an effect of Presentation Mode, F (1,37) = 33.63, p < .001, and a Modality x Presentation Mode interaction, F (1,37) = 10.91, p = .002. While auditory discrimination times in the unimodal condition (M = 459 ms, SE = 9.92) were faster than in the cross-modal condition (M = 474 ms, SE = 9.72), t (37) = 2.10, p = .042, the interaction suggests that the cost of cross-modal presentation was more pronounced in the visual condition, with visual discrimination in the unimodal condition (M = 443 ms, SE = 6.95) being faster than the cross-modal condition (M = 488 ms, SE = 8.84), t (37) = 7.17, p < .001. Thus, accuracy and RT data show that cross-modal presentation attenuated visual processing more than auditory processing, a finding consistent with auditory dominance. However, there was no slow down when both modalities changed (M = 424 ms, SE = 11.34). In fact, response times on these trials were faster than all trial types, ts > 2.26, ps < .05, suggesting that the slowdown occurs because of the conflicting information (e.g., auditory standard elicits no response; whereas, visual oddball elicits button press), as opposed to cross-modal presentation increasing task demands more generally.

HRV Analyses To examine the relationship between HRV and modality dominance, we calculated a measure of resting HRV for each participant during the five minute baseline phase. Mindware software was used to isolate the baseline phase and to detect and remove artifacts. Root Mean Square of the Successive Differences (RMSSD) was calculated for each participant, with higher values indicating more variability in resting heart rate. A median split was used to classify each participant as having low or high HRV. Response times broken up by HRV are reported in Figure 2. A 2 (HRV: Low vs. High) x 2 (Modality: Auditory vs. Visual) x 2 (Presentation Mode: Unimodal vs. Cross-modal) mixed-factors ANOVA revealed no significant effects or interactions with HRV, F's < 1.60, ps > .214, suggesting that both groups showed the same overall pattern.

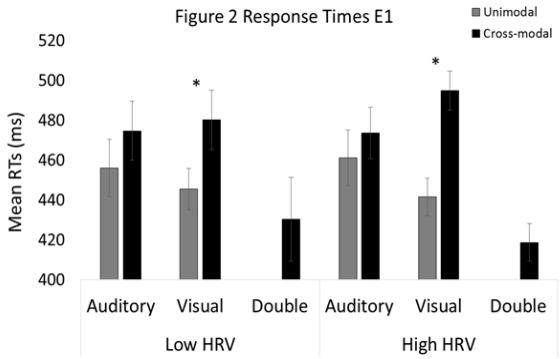


Figure 2. Mean response times across trial types, conditions, and HRV in Experiment 1. Error Bars denote Standard Errors and "*" denotes cross-modal RTs differ from unimodal RTs, ps < .001.

HR Analyses Weighted Inter-Beat Intervals (IBIs) were exported every second, and difference waveforms were calculated by subtracting pre-stimulus IBI (Weighted IBI from -1 s to stimulus onset) from each 1 s IBI bin post stimulus. Note that IBIs reflect the time between heartbeats; thus, increases in IBI reflect slowed heart rate, and difference IBIs greater than 0 reflect heart rate slowed compared to pre-stimulus levels. Paired t tests comparing standard and oddball IBIs were conducted each second to determine how quickly the heart differentiated oddballs from standards.

As can be seen in Figures 3A and 3B, cardiac responses to oddballs and standards differed at 4 s after stimulus onset in the auditory condition and 5 s after stimulus onset in the visual condition. Note that these effects were primarily driven by heart rate acceleration to oddballs; whereas, infants show slower heart rate to less frequent oddballs (Robinson & Sloutsky, 2010b). One explanation for this difference may stem from using passive looking time tasks with infants and speeded response time tasks with adults.

However, it is also worth noting that the discrimination of auditory and visual oddballs occurred earlier in the cross-modal condition (2 s after stimulus onset) compared to the unimodal conditions. Thus, behavioral data point to cross-modal interference with cross-modal presentation slowing down visual response times, but changes in time-locked cardiac responses show facilitation, with discrimination occurring earlier in the course of processing when information is presented to both sensory modalities.

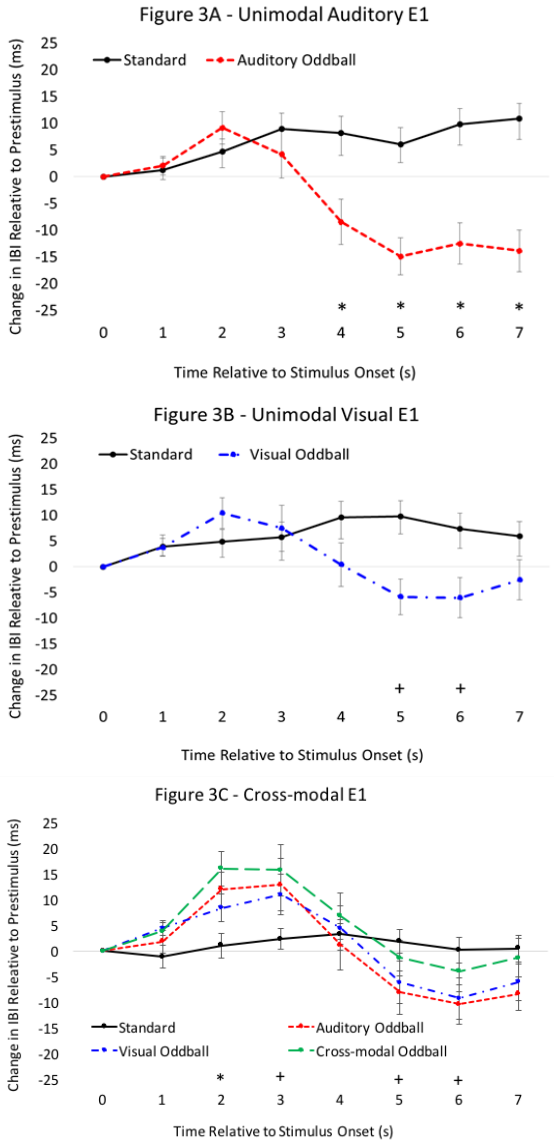


Figure 3. Cardiac responses across time. Error Bars denote *SE*s, and "+" and "*" denote auditory and visual oddballs differed from standard, *p*s < .05 and .007, respectively. Bonferonni corrections require a *p* value of .007 to reach significance.

Experiment 2

The primary goal of Experiment 2 was to further examine modality dominance effects, while using a task that is more similar in structure to the traditional visual dominance tasks (Colavita, 1974). It was hypothesized that requiring separate responses to auditory, visual, and cross-modal oddballs would result in visual dominance, with participants making more visual based errors when both modalities change (c.f., Robinson et al., 2016; Sinnett et al., 2007).

Method

Participants, Materials, and Procedure Twenty-seven new participants (15 Females, M = 23.97 years) from The Ohio State University Newark participated in Experiment 2. The stimuli and procedure was identical to Experiment 1 except that participants were instructed to press 1 on the number pad if the auditory component changed, 2 if the visual component changed, and 3 if both modalities changed (button assignment was counterbalanced across participants). Participants in the unimodal condition were only instructed to press one of the buttons.

Results and Discussion

Behavioral Analyses Accuracies in the current experiment were in the opposite direction compared to Experiment 1. While hit rates exceeded .99 when detecting visual oddballs, cross-modal presentation attenuated auditory hit rates, with auditory oddball detection in the unimodal auditory condition (M = .99) exceeding auditory hit rates in the cross-modal condition (M = .78), t (26) = 6.57, p < .001.

To examine Colavita visual dominance effects, we examined errors made on double oddballs. The overall error rate to double oddballs was 15%. Of the 66 errors made, there were 11 misses where participants failed to make any response. On the remaining trials, participants pressed only the visual button 41 times and only the auditory button 14 times, resulting in a visual modality bias, χ^2 (1, N = 27) = 12.30, p < .001.

Additional analyses focused on response times. A 2 (Modality: Auditory vs. Visual) x 2 (Presentation Mode: Unimodal vs. Cross-modal) repeated measures ANOVA only revealed an effect of Presentation Mode, F (1,26) = 449.17, p < .001, which suggests that cross-modal presentation equally affected response times in both modalities.

To make direct comparisons across Experiments, we submitted accuracies and RTs to two 2 (Experiment: 1 vs. 2) x 2 (Modality: Auditory vs. Visual) x 2 (Presentation Mode: Unimodal vs. Cross-modal) mixed-factors ANOVAs. We focus only on the effects and interactions with Experiment. All main effects and interactions were significant for accuracy, Fs (1,63) > 11.52, ps < .001, but only a main effect of Experiment and an Experiment x Modality interaction were found for RT, Fs (1,63) > 178.96, ps < .001.

HRV Analyses As in Experiment 1, we collected resting heart rate, calculated RMSSD for each participant, and used a median split to classify each individual as low or high HRV. A 2 (HRV: Low vs. High) x 2 (Modality: Auditory vs. Visual) x 2 (Presentation Mode: Unimodal vs. Cross-

modal) mixed-factors ANOVA revealed no significant effects or interactions with HRV, F's < 2.31, p's > .141, suggesting that both groups showed the same overall pattern (Figure 4).

We also examined if HRV could predict the type of errors made on double oddballs. The low HRV group made 18 visual-based errors and 4 auditory-based errors. The high HRV group made 23 visual-based errors and 10 auditory-based errors. A *Fisher's* exact test revealed no differences between the proportion of visual-based errors, $p = .36$.

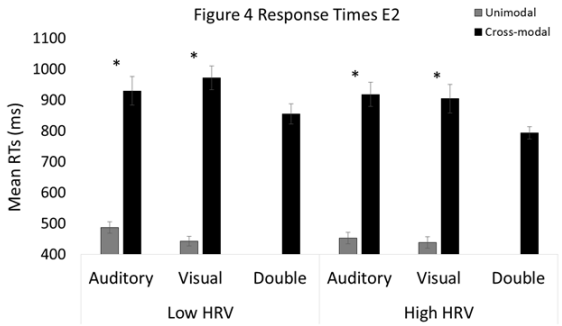


Figure 4. Mean response times across trial types, conditions, and HRV in Experiment 2. Error Bars denote Standard Errors and "*" denotes cross-modal RTs differ from unimodal RTs, ps < .001.

HR Analyses Time-locked cardiac responses to standards and oddballs are reported in Figure 5 and significant paired t tests are reported on the x axis. While discrimination was not as robust as in Experiment 1, the same pattern emerged with discrimination of auditory and visual oddballs being more robust and occurring earlier in the course of processing in the cross-modal condition than in the unimodal condition.

General Discussion

The Colavita visual dominance effect (Colavita, 1974) has been robustly replicated in the literature for the past several decades (see for example, Ngo et al., 2010; Sinnett et al., 2007; Spence et al., 2012 for a review). Indeed, while Ngo et al. (2011) did manage to reverse the effect (only under extreme conditions), it was not until recently (Robinson et al., 2016) that visual dominance has been consistently reversed. The first goal of this experiment was to extend these findings by using a slightly different procedure with more salient visual stimuli. In doing so, auditory dominance was again demonstrated when looking at both response latency and accuracy (Experiment 1). That is, responses to visual oddballs were slowed down when presented concomitantly with auditory standards, when compared to visual oddballs presented in silence. Additionally, more errors were made to visual oddballs when paired with an auditory standard than when presented in silence. This demonstration of auditory dominance dovetails with other research using a similar oddball/change detection paradigms (Robinson et al., 2016; Sloutsky & Napolitano, 2003).

In Experiment 2, auditory dominance reverted to visual dominance when participants were required to use multiple response keys. Opposite to Experiment 1, participants made more errors to auditory oddballs when paired with the visual

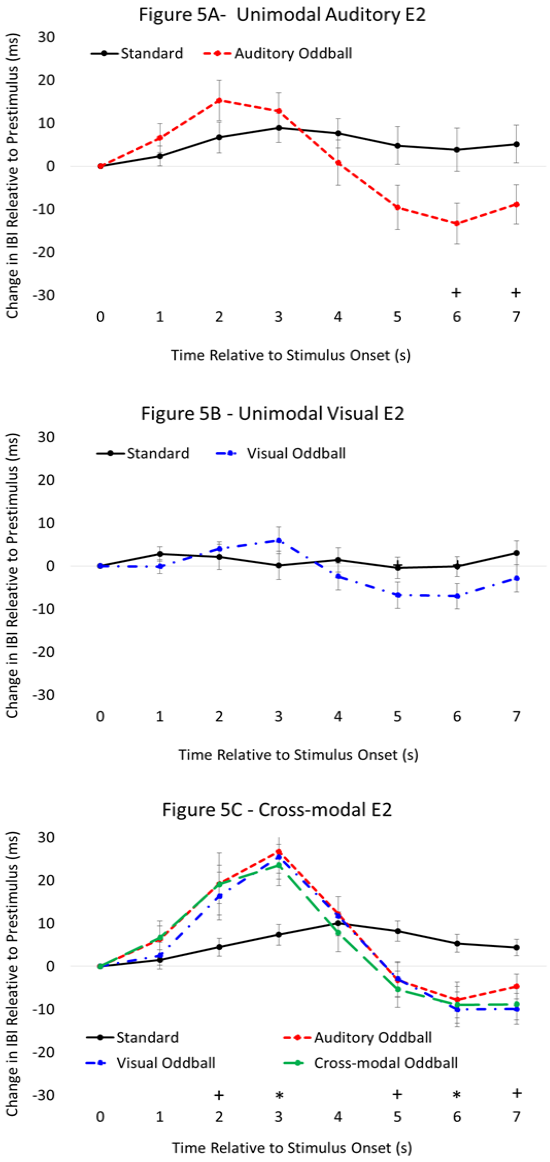


Figure 5. Cardiac responses across time. Error Bars denote SEs, and "+" and "*" denote auditory and visual oddballs differed from standard, ps < .05 and .007, respectively.

standard when compared with auditory oddballs presented without images. Visual dominance was further reflected in the percentage of visually-based errors (responding with the visual response button only) to double oddballs.

The second goal of the current study was to explore whether auditory or visual dominance effects can be modulated by top-down attentional control. To address this, we used HRV as a proxy for top-down attentional control, as previous research (Hansen, Johnson, & Thayer, 2003; Thayer & Lane, 2000) has demonstrated that high HRV is correlated with increased performance on tests of executive functioning. Interestingly, when performing a median split

on our participants (see Figures 2 and 4), HRV did not seem to modulate dominance type. These findings are consistent with previous research showing that attentional manipulations do not reverse modality dominance (Napolitano & Sloutsky, 2004; Ngo et al., 2010; Sinnett et al., 2007), and suggest that factors other than endogenous attention may modulate dominance effects.

The third goal of this project was to examine real-time psychophysiological responses to changing auditory and visual information to possibly gain insight into the time course of cross-modal interference effects. While both studies report slower behavioral responses in the cross-modal conditions, cardiac responses to auditory and visual oddballs were actually faster than the cross-modal condition. Interestingly, these early cardiac responses in the cross-modal condition were decelerations, not accelerations. While future research is needed, it is possible that both auditory and visual dominance reflect competition while participants are making a decision and/or initiating a response, and that early cardiac decelerations reflect more robust or possibly even faster encoding in the cross-modal conditions.

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15(9), 839–843.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16, 409-412.

Colavita, F.B., Tomko, R., & Weisberg, D. (1976). Visual prepotency and eye orientation. *Bulletin of the Psychonomic Society*, 8, 25-26.

Colavita, F.B., & Weisberg, D. (1979). A further investigation of visual dominance. *Attention, Perception & Psychophysics*, 25, 345–347.

Driver, J., & Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In C. Spence & J. Driver (Eds.), Crossmodal space and crossmodal attention. Oxford, UK: Oxford University Press.

Egeth, H.E., & Sager, L.C. (1977). On the locus of visual dominance. *Attention, Perception & Psychophysics*, 22, 77-86.

Giard, M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473-490.

Hansen, G., Johnsen, M.W., J.J., & Thayer, J.F. (2003). Cardiac vagal tone is correlated with selective attention to neutral distractors under load. *Psychophysiology*, *50*, 398–406.

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation Naming of Environmental Sounds. *Journal of Clinical and Experimental Neuropsychology*, *22*:6, 830 - 864.

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Napolitano, A., &Sloutsky, V.M. (2004). Is a Picture Worth a Thousand Words? The Flexible Nature of Modality Dominance in Young Children. *Child Development*, *75*(6), 1850-1870.

Ngo, M. K., Cadieux, M. L., Sinnett, S., Soto-Faraco, S., & Spence, C. (2011). Reversing the Colavita visual dominance effect. *Experimental Brain Research*, 214(4), 607-618.

Ngo, M. K., Sinnett, S., Soto-Faraco, S., & Spence, C. (2010). Repetition blindness and the Colavita effect. *Neuroscience Letters*, 480, 186–190.

Posner, M.I., Nissen, M.J. & Klein, R.M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, 83, 157-171.

Richards, J. E., & Casey, B. J. (1992). Development of sustained visual attention in the human infant. In B. A. Campbell, H. Hayne, & R. Richarson (Eds.), *Attention and information processing in infants and adults*, pp. 30-60. Hillsdale, NJ: Erlbaum.

Robinson, C.W., Chandra, M., & Sinnett, S. (2016). Existence of competing modality dominances. *Attention, Perception, & Psychophysics*, *78*, 1104-1114.

Robinson, C. W., & Sloutsky, V. M. (2010a). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135-141.

Robinson, C. W., & Sloutsky, V. M. (2010b). Attention and cross-modal processing: Evidence from heart rate analyses. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp 2639-2643). Austin, TX: Cognitive Science Society.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, 69, 673–686.

Sloutsky, V.M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*(3), 822-833.

Spence, C. (2009). Explaining the Colavita visual dominance effect. Progress in Brain Research, 176, 245–258.

Spence, C., Parise, C., & Chen, Y. C. (2012). The Colavita visual dominance effect. In M.M. Murray, & M.T. Wallace (Eds.), The Neural Bases of Multisensory Processes (pp. 529-556). Boca Raton, FL: CRC Press.

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorder*, *61*, 201–216.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), Varieties of attention (pp. 63–101). Orlando, FL: Academic Press.

From Words to Behaviour via Semantic Networks

Armand S. Rotaru (ucjt401@ucl.ac.uk) **Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)**
Faculty of Brain Sciences, University College London,
WC1H 0DS, London, United Kingdom

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University,
6525 HT Nijmegen, the Netherlands

Abstract

The contents and structure of semantic networks have been the focus of much recent research, with major advances in the development of distributional models. In parallel, connectionist modeling has extended our knowledge of the processes engaged in semantic activation. However, these two lines of investigation have rarely brought together. Here, starting from a standard textual model of semantics, we allow activation to spread throughout its associated semantic network, as dictated by the patterns of semantic similarity between words. We find that the activation profile of the network, measured at various time points, can successfully account for response times in the lexical decision task, as well as for subjective concreteness and imageability ratings.

Keywords: computational modelling; semantic networks; text corpora; lexical decision; concreteness; imageability

Introduction

In the last 15 years, a great deal of effort was invested in collecting extensive behavioural norms, for lexical semantic tasks such as free association (Nelson, McEvoy, & Schreiber, 2004), similarity judgement (Bruni, Tran, & Baroni, 2014; Silberer & Lapata, 2014), feature generation (McRae, Cree, Seidenberg, & McNorgan, 2005; Recchia & Jones, 2012; Vinson & Vigliocco, 2008). In addition, large norms have been obtained for tasks that rely primarily on orthographic and phonological processing, but also include a semantic component, such as lexical decision (Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012).

This wealth of data has allowed researchers to start exploring the ties that link language to perception and action, in a more methodical and in-depth manner than was previously possible. At a general level, especially within the fields of computational linguistics and natural language processing, representational similarity analysis has been employed in order to study verbal and visual semantic representations across domains of knowledge (Kriegeskorte, Mur, & Bandettini, 2008; for a recent review, see Kriegeskorte & Kievit, 2013). This approach is inspired by several embodied theories of cognition in which the semantic system is considered to rely on integrated modal (especially visual) and amodal representations (Barsalou, Santos, Simmons, & Wilson, 2008; Louwerse, 2007; Vigliocco, Meteyard, Andrews, & Kousta, 2009). The research following said approach has shown that unimodal (i.e., verbal or visual), but especially multimodal (i.e., verbal-visual) distributional models (for a detailed review, see Bruni, Tran, & Baroni, 2014) can provide a good account of human task performance in a number of semantic tasks. Such studies demonstrated that integrating information from two modalities provides a better account of behavioural data than that offered by the individual modalities, across a wide range of models and integration methods, even for abstract concepts, such as *peace* and *freedom* (Bruni, Tran, & Baroni, 2014; Hill & Korhonen, 2014; Hill, Reichart & Korhonen, 2014). The results are consistent with those of previous studies (Andrews, Vigliocco, & Vinson, 2009; Louwerse, 2011; Maki & Buchanan, 2008; Riordan & Jones, 2011; Sadeghi, McClelland, & Hoffman, 2015; Steyvers, 2010), indicating that language and perception can be seen as highly redundant, yet complementary, sources of semantic information.

Differences in the reliance upon one or the other modalities, as well as in degree and strength of association to other concepts, have been argued to underscore difference across domains of knowledge. For example, representational richness has been argued to underlie the distinction between concrete and abstract concepts, whereby concrete concepts are richer than abstract ones when it comes to perceptual and motor elements, but poorer with respect to introspective and linguistic elements (see Gee, Nelson, & Krawczyk, 1999; Hill, Korhonen, & Bentz, 2014; Pecher, Boot, & Van Dantzig, 2011; Vigliocco et al., 2009; Wiemer-Hastings & Xu, 2005). A large number of studies have used comprehensive behavioural norms and subjective ratings to evaluate the role of semantic richness, using different measures of richness such as number of features as well as contextual and semantic diversity, to name a few (for reviews, see Jones, Johns, & Recchia, 2012; Mirman & Magnuson, 2008; Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012). Not surprisingly, the results paint a rather complex picture, where semantic richness effects are both task-general and task-specific, have both an early and a late impact on task behaviour (Hargreaves & Pexman, 2014), and either facilitate or hinder task performance (Mirman & Magnuson, 2008).

Here, we attempt to bring a fresh perspective in the study of how concepts (both concrete and abstract) are represented and, crucially, processed, by developing a computational model that accounts for previous findings by incorporating

both structural and dynamical elements. In particular, we explore the extent to which we can predict response times and accuracies in visual word recognition (i.e., lexical decision), as well as both concreteness and imageability ratings, starting from distributional models of semantics (Mandera, Keuleers, & Brysbaert, 2015; Westbury et al., 2013) supplemented by simple assumptions concerning the dynamic spreading of activation during processing.

Method

Model

We derive semantic richness measures of words from a probabilistic model of semantic processing, in the following manner: (a) pre-process the written part of the British National Corpus (Leech, Garside, & Bryant, 1994), by converting all the words to lowercase, eliminating punctuation marks and removing words whose absolute frequencies are less than 5; (b) construct 300-dimensional vector representations for the words in the BNC, by employing the Skipgram[1] model (Mikolov, Chen, Corrado, & Dean, 2013); (c) compute a representational similarity matrix *DM* from said vectors, using vector cosine as a measure of similarity between the vectors (i.e., words); (d) set to zero all the negative values in *DM*, as a means of reducing the amount of noise present (i.e., vector cosines which carry very little semantic information); (e) normalize the rows of the matrix, such that each row sums to one, and that any value $DM(I,J)$ can be interpreted as the strength of the directional connection from word W_I to word W_J ; (f) consider the discrete Markov chain associated with *DM*, which we denote as *MARKOV*(*DM*), and compute the state of *MARKOV*(*DM*) at steps $K = 1$ through $K = 7$, namely $S_K(DM)$; (g) for each word *W* and each *K* between 1 and 7, count the number of close neighbours of *W* ($numNeigh_K$). A word *V* is considered a close neighbour of *W* if $P(V|S_K(DM)) > thresh_K$, where $thresh_K$ are lower thresholds.

In the end, we are left with seven free parameters (i.e., $thresh_{1\text{-}7}$) and seven semantic richness measures (i.e., $numNeigh_{1\text{-}7}$), as well as with a few fixed parameters for the underlying Skipgram model.[2] Although our richness measures are all derived in a very similar manner, they have rather different interpretations, at least from a graph-theoretical perspective (Koschützki, Lehmann, Peeters, Richter, Tenfelde-Podehl, & Zlotowski, 2005). The meaning of each measure is briefly described in Table 1.

Table 1. Semantic richness measures computed by our model, and their tentative interpretation. For clarity, only the distinguishing aspects of each measure are presented.

Semantic richness measure	Graph theoretical interpretation
$numNeigh_1$	# of close neighbours
$numNeigh_2$	# of connections between close neighbours # of distant neighbours # of connections between close and distant neighbours
$numNeigh_3$	# of connections between distant neighbours # of connections between distant and close neighbours
$numNeigh_{4\text{-}7}$	# of close direct and indirect neighbours # of very distant neighbours

Data Analysis

We focus on four dependent measures: concreteness ratings (Brysbaert, Warriner, & Kuperman, 2014), imageability ratings (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), and both accuracies and response times from a lexical decision task, for a subset of 2,328 words from Keuleers, Lacey, Rastle, and Brysbaert (2012).

We include the following baseline variables: (log) contextual diversity, (log) frequency (van Heuven, Mandera, Keuleers, & Brysbaert, 2014), familiarity (Gilhooly & Logie, 1980; Stadthagen-Gonzalez & Davis, 2006), age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), (squared) hedonic valence (Warriner, Kuperman, & Brysbaert, 2013), number of letters, Coltheart's N (i.e., the number of words that can be produced by substituting one letter of a given word for any other, such that the result is a valid word; Coltheart, Davelaar, Jonasson, & Besner, 1977), orthographic Levenshtein distance (OLD20; the average orthographic editing distance between a word and its twenty closest neighbours in the lexicon; Yarkoni, Balota, & Yap, 2008), and phonological Levenshtein distance (PLD20; the average phonological distance between a word and its twenty closest neighbours in the lexicon; Suárez, Tan, Yap, & Goh, 2011). In addition we include semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013) as a baseline measure. This latter has been argued to capture basic semantic differences across concepts as represented in distributional semantic networks. Our variables of interest are the seven measures of semantic richness (i.e., $numNeigh_{1\text{-}7}$).

We run two multiple linear regressions, one for the baseline variables, and one for the complete set of predictors (i.e., the baseline variables and our semantic richness measures). Since our richness measures are very strongly correlated with one another, we partial out any variance shared with other predictors, such that $numNeighR_K$ = Res

[1] We prefer the Skipgram model for two main reasons: firstly, it is nearly state-of-the-art when it comes to accounting for behavioral data (Baroni, Dinu, & Kruszewski, 2014; Mandera, Keuleers, & Brysbaert, 2015); secondly, several freely available, computationally efficient and well documented implementations of the model exist (https://code.google.com/p/word2vec/).

[2] We use the Skipgram implementation provided by the *gensim* tool (Řehůřek & Sojka, 2010), with the following parameter values: *alpha* = 0.025 (initial learning rate), *window* = 5 (radius of sliding window), *sample* = 0 (amount of downsampling), *negative* = 0 (amount of negative sampling), and *iter* = 1 (number of iterations over the entire corpus).

($numNeigh_K$ ~ *Baseline* + $numNeighR_1$ + … + $numNeighR_{K-1}$), for all values of *K* between 1 and 7. Therefore, our predictors consist of *Baseline* and $numNeighR_{1\text{-}7}$, whereas our dependent variables are *Log RT*, *Accuracy*, *Concreteness* and *Imageability*.

We employ one half of the words for model tuning, and the other half for model evaluation. We derive the optimal values for our predictors by using a variant of the simplex method (Lagarias, Reeds, Wright, & Wright, 1998), with (negative) total amount of variance explained serving as the objective function. In order to avoid local minima, we run 100 iterations of the optimisation process, and keep only the best result.

Results

The results are displayed in Tables 2 and 3. Our semantic richness measures can account for a significant amount of variance in concreteness and imageability ratings, as well as in response times in the lexical decision task. However, they do not explain variance in lexical decision accuracy over and above the baseline measures (Table 2). Table 3 shows the regression weights for all predictors and dependent variables.

Table 2. Percentage of variance accounted for by two models: a baseline model, and a combined one, consisting of all the predictors in the baseline model plus the semantic richness measures $numNeighR_1$ through *numNeighR7* (all values are significant at .001 level, except for accuracy in the "combined – baseline" comparison)

Dependent variable	Baseline	Combined	Combined – baseline
Log RT	47.93	49.80	1.87
Accuracy	27.09	27.81	0.72
Concreteness	35.40	58.59	23.19
Imageability	31.90	53.24	21.34

Discussion and Conclusions

We develop a model that takes into account both the structural properties of semantics networks, as well as their dynamic aspects, by considering the flow of semantic activation generated by the automatic processing of individual words. An important result of looking at both structure and dynamics is that it allows us to assess the effects of both direct and indirect, mediated semantic relations between words, rather than limiting our analysis to strong, direct semantic links. Our results suggest that the explanatory power of text-based semantic representations is currently being underestimated, as a consequence of not taking into consideration the additional information provided by spreading activation mechanisms. By ignoring these simple processes, the extra information they generate would have to be integrated into the representations by

Table 3. Regression weights and their associated significance values, namely <0.1 (†), <0.05 (*), <0.01 (**), and <0.001 (***). Log RT = (log) response time; ACC = accuracy; CONC = concreteness; IMAG = imageability.

Outcome / Predictor	Log RT	ACC	CONC	IMAG
(Intercept)	6.611 ***	.676 ***	7.143 ***	7.405 ***
Semantic diversity	.009	-.022 **	-1.219 ***	-1.401 ***
Log contextual diversity	-.025 ***	.026 ***	-.385 ***	-.539 ***
Log frequency	-8.07e-4	-.008 *	.190 ***	.288 ***
Familiarity	-.035 ***	.024 ***	.169 ***	.373 ***
Age of acquisition	.003	-.003	-.313 ***	-.477 ***
Squared hedonic valence	-.004 ***	.002 *	-.090 ***	.025 †
Number of letters	.007 **	.005 *	.040	.031
Coltheart's N	.001 †	-1.48e-4	.012 †	.025 **
OLD20	.002	-6.54e-5	-.148	.058
PLD20	.012 *	-8.50e-4	-.263 ***	-.342 ***
$NumNeighR_1$	-.006 **	.004 *	.181 ***	.401 ***
$NumNeighR_2$	.004 *	-.003 †	-173 ***	-.231 ***
$NumNeighR_3$	-.001	-3.41e-4	-.132 ***	-.271 ***
$NumNeighR_4$	-.001	-1.93e-4	-.250 ***	-.116 ***
$NumNeighR_5$	3.31e-4	-2.91e-4	-.263 ***	-.218 ***
$NumNeighR_6$	.002	-7.74e-4	-.057 **	.014
$NumNeighR_7$	-2.12e+6 **	-.002	.150 ***	.167 ***

design, which would lead to the conflation of representations and processes.

Based on the results presented in Table 2, it seems that our model is considerably more suitable for predicting concreteness and imageability ratings, than reaction time and accuracy in the word recognition task. We believe that this phenomenon might be due to differences between the requirements of the lexical decision task on the one hand, and those of the concreteness/imageability rating task, on the other. Since our model assumes that the string of letters received as input is already a word, it is not surprising that it fares rather poorly in predicting lexical decision response time and accuracy. In contrast, the rating task involves making a considerably more elaborate discrimination, one between concrete/imageable and abstract/non-imageable words, all of which are present in our model (Buchanan, Westbury, & Burgess, 2001).

Beyond the promising initial results, we believe that our model has a number of advantages, which recommend it as a potentially useful tool in the study of semantic processing. In our opinion, the main quality of our model is that it ties together a number of competing modelling approaches, and combines many of their strengths, while avoiding most of their limitations.

Firstly, our model has a pronounced connectionist and/or dynamical systems flavour to it (Anderson, 1983; for a review, see McClelland et al., 2010), whereby the dynamics of the model can be interpreted in terms of "spreading activation". In this case, activation flows from an initial concept to its neighbours, then to the neighbours of its neighbours, and so on, until the system reaches a global "attractor" state (i.e., an eigenstate). However, unlike other existing models (Chen & Mirman, 2012; Hoffman & Woollams, 2015; Rogers & McClelland, 2004), it has a large number of nodes and feedforward/feedback/recurrent connections, making it slightly more realistic and comprehensive. As a result, it might provide better insight into the distinct contribution of structural and task related aspects of semantic behaviour. One potentially promising approach in this regard comes from network science and the theory of stochastic processes, two methodologies which have attracted an increasing amount of attention in cognitive science (De Deyne & Storms, 2008; Ferrer i Cancho & Solé, 2001; Gruenenfelder, Recchia, Rubin, & Jones, in press; Steyvers & Tenenbaum, 2005; Utsumi, 2015; for a general review of network-based analyses of cognition, see Baronchelli, Ferrer i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). Another possibility might be to use a respond-to-signal paradigm (Ratcliff, 2006; Hargreaves & Pexman, 2014), which would provide additional quantitative insights on the accumulation of task-specific and task-independent information during task performance (e.g., in the word naming or the lexical decision tasks).

Secondly, our model can be seen as a probabilistic one (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), such that at each step, the model makes use of its underlying Markov chain, namely *MARKOV*(*DM*), in order to perform multi-step inferences. In contrast to other probabilistic models, such as Topics (Griffiths, Steyvers, & Tenenbaum, 2007), our model is non-hierarchical and does not undergo any form of dimensionality reduction, which means that the inferences are easier to interpret and that less semantic information is lost. Said inferences allow us to assess the strength of both direct and indirect semantic relations between words (Steyvers, Shiffrin, & Nelson, 2004; Howard, Shankar, & Jagadisan, 2011), for instance by testing whether certain words and/or associations between words are crucial for successfully carrying out a semantic task. Moreover, we can also examine the manner in which semantic cues restrict and guide the inference process, as is the case in tasks such as semantic fluency (Hills, Jones, & Todd, 2012), continued free association (De Deyne & Storms, 2008), and extralist cued recall (Nelson, Kitto, Galea, McEvoy, & Bruza, 2013).

Finally, our model is relatively simple, from a structural point of view, and is completely transparent in terms of its parameters. Taken together, these features make our model easy to run, and facilitate comparisons across different subsets of participants, stimuli and tasks. Also, as a results of its simplicity, our current model is very much open to extensions, for instance in order to increase its neuropsychological plausibility.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261-295.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463-498.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Baronchelli, A., Ferrer i Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348-360.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In *Symbols and Embodiment: Debates on meaning and cognition*.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research*, *49*(1), 1-47.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known

English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531-544.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417-430.

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213-231.

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1482), 2261-2265.

Gee, N. R., Nelson, D. L., & Krawczyk, D. (1999). Is the concreteness effect a result of underlying network interconnectivity? *Journal of Memory and Language*, *40*(4), 479-497.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395-427.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (in press). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive Science*.

Hargreaves, I. S., & Pexman, P. M. (2014). Get rich quick: The signal to respond procedure reveals the time course of semantic richness effects during visual word recognition. *Cognition*, *131*(2), 216-242.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of EMNLP*.

Hill, F., Korhonen, A., & Bentz, C. (2014). A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, *38*(1), 162-177.

Hill, F., Reichart, R., & Korhonen, A. (2014). Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, *2*, 285-296.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431-440.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718-730.

Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 385-402.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science*, *3*(1), 48-73.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. Canadian *Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 115-124.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287-304.

Koschützki, D., Lehmann, K. A., Peeters, L., Richter, S., Tenfelde-Podehl, D., & Zlotowski, O. (2005). Centrality indices. In *Network analysis*. Springer: Berlin Heidelberg.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401-412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder--Mead simplex method in low dimensions. *SIAM Journal on Optimization*, *9*(1), 112-147.

Leech, G., Garside, R., & Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th Conference on Computational Linguistics* (vol. 1).

Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In *Handbook of Latent Semantic Analysis*.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273-302.

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598-603.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, *68*(8), 1623-1642.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and

dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348-356.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR*.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 65-79.

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition*, *41*(6), 797-819.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Pecher, D., Boot, I., & van Dantzig, S. (2011). Abstract concepts: sensory-motor grounding, metaphors, and beyond. In *The Psychology of Learning and Motivation* (vol. 54).

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195-237.

Recchia, G., & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6*.

Řehůřek, R., & Sojka, P. (2010) Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303-345.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52-61.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234-243.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41-78.

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*(3), 605-611.

van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Utsumi, A. (2015). A Complex Network Approach to Distributional Semantic Models. *PLoS ONE*, *10*(8), e0136277.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219-247.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183-190.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*.

Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, *29*(5), 719-736.

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. (2012). An abundance of riches: cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*(72).

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Evolution of polysemous word senses from metaphorical mappings

Yang Xu (yang_xu_ch@berkeley.edu)
Department of Linguistics, Cognitive Science Program, UC Berkeley

Barbara C. Malt (barbara.malt@lehigh.edu)
Department of Psychology, Lehigh University

Mahesh Srinivasan (srinivasan@berkeley.edu)
Department of Psychology, UC Berkeley

Abstract

What forces have shaped the evolution of the lexicon? Languages evolve under the pressure of having to communicate an unbounded set of ideas using a finite set of linguistic structures. This suggests *why* the transmission of ideas should be compressed such that one word will develop multiple senses. Previous theory also suggests *how* a word might develop new senses: Abstract concepts may be construed in terms of more concrete concepts. Here, we bring these two perspectives together to examine metaphorical extensions of English word meanings over the past millennium, analyzing how senses from a source domain are extended to new ones in a target domain. Using empirical and computational methods, we found that metaphorical mappings are highly systematic and can be explained in terms of a compact set of variables. Our work shows how metaphor can provide a cognitive device for compressing emerging ideas into an existing lexicon.

Keywords: Word meaning; semantic change; polysemy; metaphorical mapping; systematicity

Words are fundamental components of language, but their meanings are not stable. For example, the English word *grasp* originally conveyed a physical action, as in "grasp a fruit," but was later extended to express an abstract sense of understanding, as in "grasping of an idea." The "physical action" sense first appeared around 1300, and the "understanding" sense emerged around AD 1600 (Christian, Roberts, Samuels, Wotherspoon, & Alexande, 2015a). The synchronic product of such historical sense extensions, known as polysemy - that a single word form can express multiple distinct but related senses - is widespread in natural language (Lakoff & Johnson, 1980; Brugman, 1988; Sweetser, 1991; Geeraerts, 1997). We investigate the extent to which historical polysemous patterns can be predicted by focusing on metaphorical mapping.

Metaphorical mapping is a key structuring force in sense extension and and semantic change (e.g., Lakoff & Johnson, 1980; Sweetser, 1991). It operates by mapping an existing sense of a word from its own *source* domain to another *target* domain based on structural similarities between the two domains. For instance, the "physical action" sense of *grasp* can be thought of as metaphorically extended to "understanding," namely holding onto an idea. The historical process through which polysemy develops raises the following question: Is the evolutionary path that metaphorical senses follow unpredictable, or is there a *systematicity* in how new senses develop from existing ones, driven by overarching evolutionary forces? We address this question by bringing together two influential theories that were not previously in contact.

On the one hand, a prominent theory of the nature of human cognition - Conceptual Metaphor Theory (Lakoff & Johnson, 1980; Reddy, 1979) - holds that thought is grounded in metaphor. By this account, metaphorical mappings in thought should occur in directions that are cognitively natural. For example, such a theory postulates that abstract thought is facilitated by linkage to concrete concepts (e.g., "understanding" construed in terms of concepts that are more concrete or directly related to human experience - a "physical action" of holding onto something). This perspective implies systematicity in the development of the lexicon, because it suggests that mappings will tend to be from concrete to abstract and not *vice versa*. However, it suffers in two important respects as an account of polysemy. First, variables proposed to be cognitively priviledged in metaphorical mapping (discussed below) have not been assessed against the historical record of lexical change, and therefore their explanatory powers and interrelations in accounting for empirical data with respect to the lexicon remain unknown. Second, the conceptual theory itself does not motivate why metaphor should serve as a key mechanism for the evolving lexicon, because its goal is to explain thought, not word meanings *per se*.

On the other hand, a growing line of research on principles of language evolution provides clues for addressing the *why* question. By this account, linguistic structures evolve under the dual pressures of communicative needs and cognitive constraints (e.g., learnability), such that languages should trade off between competing pressures of expressivity and compressibility (Kirby, Tamariz, Cornish, & Smith, 2015). Specifically, language evolution must allow for meaningful and informative communication of ideas while providing sufficient compression of ideas into existing linguistic structures (e.g., via compositional means) to ensure that languages do not grow without bound and are therefore learnable. Although such theories have been applied to explain synchronic features of language such as word length (Piantadosi, Tily, & Gibson, 2011) and the structure of semantic domains (e.g., Regier, Kemp, & Kay, 2015), they have not been used to help understand the role of metaphorical mapping (cf. Geeraerts, 1997; Blank & Koch, 1999) in the diachronic development of lexicons.

We offer such a perspective on metaphorical mapping by bringing together these perspectives. We propose that the

unique nature of metaphor in lexical evolution is precisely explained by functional needs for expressive communication and cognitive constraints for learnability. In one respect, metaphorical mapping facilitates meaningful communication of emerging ideas, such that when a speaker uses a word in a novel sense metaphorically, the listener should find it cognitively effortless to decode the intended meaning (cf., Traugott, 2003). In another respect, metaphorical mapping serves as a compression device that effectively folds emerging meanings into existing words without requiring construction of word forms *de novo*. As such, it serves as a strategic device that trades off between communicative and cognitive constraints. Our proposal predicts that metaphorical sense mappings should occur in systematic ways by conforming to both communicative and cognitive constraints.

We critically assess this idea by analyzing a large set of metaphorical mappings between source and target domains and spanning 1300 years, dating from Anglo-Saxon English to the present. We seek to predict the historical ordering of sense extension by testing whether some domains are more likely to be the starting point of an historical trajectory (the source domain) while others are more likely to be the endpoint (target domain). To our knowledge, this is the first large-scale study evaluating the predictability and directionality of polysemous metaphorical mappings against the record of historical change in a lexicon.

Candidate variables

We identify six candidate variables for metaphorical mapping based on communicative and cognitive considerations. Three of the variables we test are suggested by Conceptual Metaphor Theory (Lakoff & Johnson, 1980; Reddy, 1979). Applied to the development of word senses, meanings that are easier to understand and more richly experienced - because they are more concrete and tied to bodily experience - might develop earlier in historical time and might be especially accessible, well-structured, and conceptualized. As such, they may lend themselves to extension via metaphorical processes because such extensions would provide a high degree of efficiency and expressivity. These considerations provide the following predictions:

Concrete → Abstract. Word senses that refer to things perceived through the sensory systems should serve as a source of metaphorical mappings, relative to ones labeling less perceptible referents (Lakoff & Johnson, 1980).

Embodied → Disembodied. Word senses that refer to things that are more directly and viscerally experienced through our bodies should serve as a source of metaphorical mappings, relative to senses that label referents less directly experienced through our bodies (Lakoff & Johnson, 1980).

External → Internal. Word senses that refer to entities in the external world should serve as a source of metaphorical mappings, relative to senses that label internal, mental entities (e.g., emotions or feelings). This would be predicted if internal entities are understood in terms of external things (Sweetser, 1991), and/or if external entities are more easily lexicalized than internal entities (because external things are easier to indicate ostensively).

An additional three variables can be motivated based on the proposal that metaphorical extensions tend to be communicatively efficient and expressive, which includes describing target domains in vivid and emotionally-valenced ways.

Animate → Inanimate. Insofar as animate entities hold a special status within our mental lives (Silverstein, 1976; Traugott, 2003), words referring to them may be more expressive than those referring to less animate entities, and thus may be recruited by speakers to vividly convey salient features of an intended referent to addressees. Thus, words that refer to animate entities could serve as a source of metaphorical mappings, relative to words that label inanimate entities.

Less valenced → More valenced. If metaphorically-derived senses arise in part because of their expressive power, we might expect derived senses to be more emotionally valenced than originating senses. This predicts that when word senses from source domains are extended, the resulting senses in the target domain will be more emotionally valenced (Ullmann, 1957).

More Intersubjective → Less intersubjective. Intersubjectivity refers to the degree to which people experience something the same way and agree about the nature of that experience (Traugott, 2003). For example, most people would agree about whether or not a chair is wooden (making it more intersubjective) but not necessarily about whether it is beautiful (making it less intersubjective). Word meanings that are more intersubjective can be easier to understand or establish labels for;[1] They would thus tend to serve as a source of metaphorical mappings, relative to senses with less intersubjective meanings.

In sum, at least six potential variables are worthy of evaluation. Some of these variables are likely to be correlated with one another—an important aspect that we address in our analysis. For example, many of the same word meanings are likely to be external, concrete, intersubjective, and embodied. Despite this fact, these variables are not identical and can in principle be teased apart from one another. For example, some highly concrete word meanings, like *table* and *arm* differ in how embodied they are: e.g., arms are more viscerally and directly experienced than tables.

Materials and methods

To explore metaphorical sense mappings in history, we drew on the corpus of metaphorical mappings of English provided by the Mapping Metaphor project (2015). We describe this database and the empirical methods for obtaining behavioral ratings of the six variables.

Historical database We obtained data from the Metaphor Map of English database (Christian, Roberts, Samuels, Wotherspoon, & Alexande, 2015b), which identifies metaphorical links among different semantic domains (as defined within the database) over more than a millennium. This

[1] Personal communication via Eve Sweetser.

database provides metaphorical sense changes identified and coded from 793,742 word forms and 225,131 semantic domains (e.g., some domains include "textiles," "digestive organs," "anger," "pride," etc.) in the Historical Thesaurus of English project (Christian et al., 2015a), based on the *Oxford English Dictionary* and A Thesaurus of Old English (Roberts, Christian, & Grundy, 2015). Each semantic domain defines a category of meaning that word senses can evolve from (source domain) or towards (target domain). We used all domains having full date and word information 1) source domain, 2) target domain, 3) earliest of period of extension between these domains, and 4) some sample words that participated in this extension. In sum, this data set contains records of historical metaphorical sense mappings among 400 semantic domains. The data set contains mappings spanning an 1100 year period, from the Old English period around AD 800 through to the present era, around AD 1950. Within the database, the metaphorical mappings were summarized separately for the Old English period (i.e., before AD 1100), and in 50-year steps for the subsequent 800 years, providing 18 unique historical time points for our analysis. In total, the database lists over 5,000 pairs of domains, indicating the historical direction of sense extension among each pair of domains. Three types of directions were recorded: 1) A→B, i.e. A is source and B is target; 2) A←B, i.e. A is target and B is source; 3) A↔B, i.e. bidirectional. Our analysis in the following section aims to predict these relationships among domains based on the ratings described.

Ratings We gathered ratings of the 400 semantic domains along the six target variables discussed above through an online survey. For each variable, participants rated each of the 400 domains on a 1-7 scale. Participants first read a definition of each domain (e.g., "Plant" - A living thing that grows in the ground, usually has leaves or flowers, and needs sun and water to survive). These definitions were assembled by consulting the *Merriam-Webster Dictionary*. After reading the definition of the domain, participants were asked to rate the domain on one of six variables. For example, for concreteness, participants selected a number between 1 and 7, where 1 represented "highly abstract," 7, "highly concrete," and 4, "intermediate." Similarly, for valence, 1 represented a "highly negative" emotional response , 7 "highly positive," and 4, "intermediate" or neutral. Because a large number (400 domains×6 variables) of ratings needed to be completed, each participant rated a block of 40 domains randomly sampled from the 400 for a single variable. Data were collected from 1448 participants using the Qualtrics survey software,[2] disseminated via Amazon's Mechanical Turk.[3] Data from participants whose native language was not English, and from those who did not respond correctly to three "catch" trials assessing attention to the task, were discarded. We obtained on average 18 (SD=2) ratings for each of the domain-variable questions with a standard deviation of 1.6 (SD=0.32) in inter-subject agreement.

[2]http://www.qualtrics.com/

[3]https://www.mturk.com/mturk/

Computational analyses and results

To investigate systematicity in metaphorical mapping, we performed two analyses to account for 1) diachronic ordering and 2) synchronic asymmetry in metaphorical sense mappings recorded in the historical database. We describe each of these analyses.

Diachronic ordering of metaphorical mappings. We used a set of parameter-free models to predict directions of metaphorical mappings among semantic domains. Specifically, for each of the variables described above, we created a model that specified the predicted direction of mapping between a pair of semantic domains by calculating the difference in average empirical ratings of these two domains along the variable dimension. For example, for the animacy variable, we predicted the direction of mapping to be Animate→Inanimate, such that the domain that was rated as more animate on average should serve as the source, and the domain that was rated less animate on average should serve as the target. Our models also predicted bidirectional mappings (accounting for 6.76% of the available mappings) if the average ratings of two domains were equal. Since the valence predictor is polarized (i.e. a concept can be either positive or negative), we calculated the absolute value of its ratings, and for this model predicted the direction of mapping betwen domains to follow Less valenced→More valenced. Thus, a domain rated 7 (highly positive) would be treated as equally-valenced as a domain rated 1 (highly negative). Table 1 specifies the predictions of each model along with example mappings that the models successfully predicted in the database.

Each model made a prediction about the direction of extension (the source-target relation) at different historical epochs for a given pair of domains (e.g. A and B), out of three possibilities: $A \rightarrow B$, $A \leftarrow B$, or $A \leftrightarrow B$. To establish a baseline, we considered a random model that predicted the direction of mappings arbitrarily. The results appear in Figure 1a. Overall, all models predicted directional change above chance. In particular, externality (accuracy = 68.1%) and concreteness (67.6%) were roughly equivalent to one another, with each exceeding chance-level accuracies (33.3%) by twofold. These were followed by intersubjectivity, valence, and embodiment variables with accuracies all above 50%, while animacy (48.8%) explained the data the least. To verify that the specific predictive directions we proposed for these models were indeed more dominant than their opposite, e.g. Abstract→Concrete, we ran these models by reversing their predictive arrows. In each of these cases, we found that the predictive accuracy was substantially worse and closer to chance: animacy (44.2%), concreteness (25.4%), embodiment (40.5%), externality (25%), intersubjectivity (35.9%), valence (37%). These results support the view that historical metaphorical mappings are systematic rather than arbitrary, such that words are more likely to be extended from some domains toward other domains, compared to the re-

Table 1: Model predictions and examples of model-predicted metaphorical sense mappings in English.

Model prediction	**Existing sense** (rating on dimension: 1-7) → **New sense** (rating)	**Attested word**	**Beginning of new sense**
Animacy			
Animate→Inanimate	Birds(6.7) → Night(1.7)	owl	1400-1450
	Baby and young person(6.7) → Courage(2)	lad	1550-1600
Concreteness			
Concrete→Abstract	Reflection (of light)(6) → Virtue(1.6)	clear	1350-1400
	Plant(6.8) → Intellect(1.8)	vegetary	1550-1600
Embodiment			
Embodied→Disembodied	The human body(6.8) → Kinship and relationship(3.2)	fleshly/flæsclic	Old English
	Bodily tissue(6.1) → Individual colours(2.1)	incarnate	1500-1550
Externality			
External→Internal	Light(6.4) → Thought(1.1)	reflect	1550-1600
	Fireworks(6.7) → Esteem(1.4)	sky-rocket	1850-1900
Intersubjectivity			
Agreeable→Disagreeable	Place and position(5.2) → Supernatural(3.1)	presence	1650-1700
	Mathematics(5.9) → Wisdom(3.5)	calculative	1750-1800
Valence			
Neutral→Valenced	Relative position(4.1) → Excitement(6.4)	up	1300-1350
	Granular texture(3.6) → Moral evil(1.3)	dusty	1600-1650

verse. The fact that the variables we considered helped predict the direction of metaphorical mappings also suggests that the communicative and cognitive factors we described constrain metaphorical extension.

Next, we aimed to explore whether the variables we assessed act in complementary ways, and account for unique variance in explaining directions of metaphorical extension. Specifically, we performed a correlation analysis and a follow-up residual predictive analysis. Figure 1b shows the inter-correlations between all pairs of variables. Externality and concreteness were most strongly correlated (Pearson $r = 0.84$). We thus expected these variables to make similar predictions about mapping directions and for one to make little independent contribution to explaining the directionality of metaphorical mappings over the other. Intersubjectivity was most strongly correlated with both externality ($r = 0.44$) and concreteness ($r = 0.47$) among the remaining pairwise correlations. Animacy, embodiment and valence were generally less correlated with other variables.

To take into account these correlations, we applied a residual prediction procedure to analyze the total variance explained from all of these variables as follows. First, we found the variable that best predicted the directions of all available metaphorical mappings. We then iteratively searched for the variable that best predicted the remaining mappings until there existed no further variables. Figure 1c summarizes the total variance explained via this procedure. Strikingly, the variables together accounted for over 90% of all available mappings. Externality stood out as the single most accurate predictor, explaining 68% of mappings. Embodiment (14%) and valence (5%) explained the majority of the remaining data not predicted by externality, while animacy accounted for 2.3%. Finally, concreteness and intersubjectivity both explained no more than 1% of data. This is not to suggest that these latter variables play no role in explaining the data, but only that their power in explaining the residual variance was likely subsumed by the externality variable, with which they were highly inter-correlated.

Together, our findings suggest that externality, embodiment, and valence define a relatively orthogonal space that can help explain the directionality of the large majority of metaphorical mappings in the historical dataset. These findings provide the first large-scale evidence for the idea that metaphorical senses have been developed in systematic and constrained ways.

Synchronic asymmetry in metaphorical mappings. Our analysis so far has focused on metaphorical mappings at different historical points, but the outcome of these diachronic changes should also be reflected in more global, time-independent asymmetries as to which domains tend to be the sources of metaphorical extension, and which the targets. To explore this further, we examined the extent to which the variables we proposed would explain the asymmetry of semantic domains. Namely, can we predict which domains tend to be sources and which tend to be targets considering all available metaphorical mappings across time points?

To address this question, we first needed to specify a quantifiable indicator of source-target asymmetry for each semantic domain. To do so, we created an asymmetry index (*AI*), formally defined as the difference between two probabilities:

$$AI = p(source) - p(target) \quad (1)$$

Here $p(source)$ is the probability of a domain serving as a source of metaphorical mappings across all time points in history (i.e. a synchronic summary of "source" strength), and

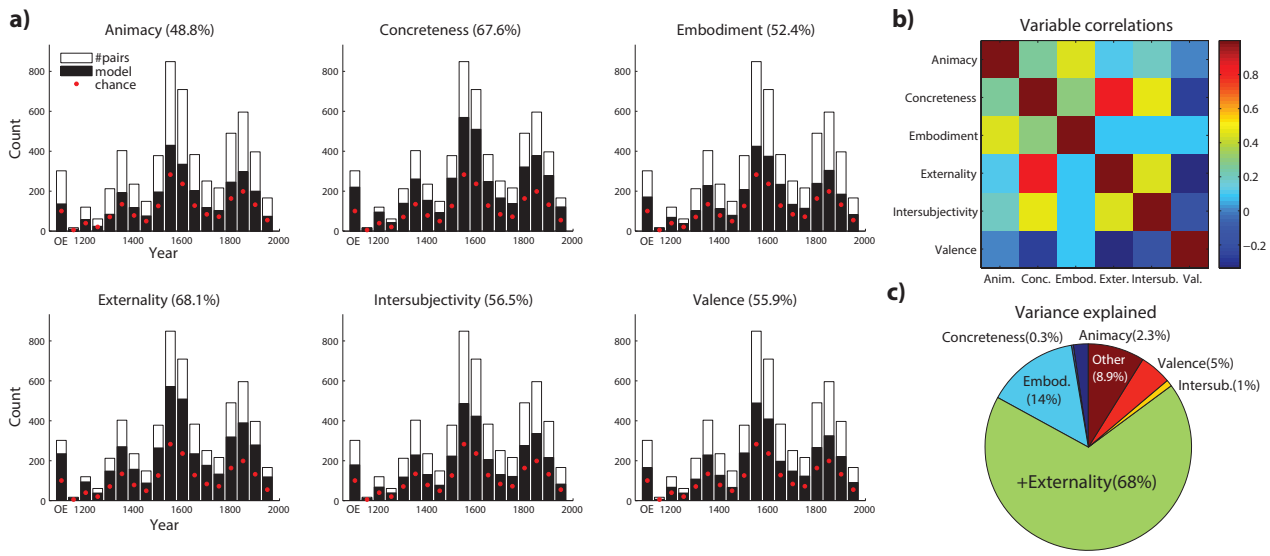


Figure 1: Summary of results on predicting historical order of metaphorical sense mappings. a) Prediction from individual variables. "OE" stands for Old English. b) Correlations among the variables. c) Variance explained by the variables.

$p(target)$ is the probability of a domain being a target. A high positive value of AI indicates a strong source domain, whereas a low negative value of AI indicates a weak source (or strong target) domain. We computed these probabilities by summing all cases of metaphorical mapping where a domain served as a source or target respectively in history and normalized them by the total number of mappings. Table 2 lists the strongest source and target domains in the dataset based on this measure.

Having obtained the asymmetry indices, we then correlated the mean empirical ratings of each variable with the degree of source-target asymmetry in these domains. Figure 2a summarizes the results. All variable ratings except for those of animacy correlated significantly with the asymmetry indices ($p < 0.01$ for the five predictive variables). Specifically, externality (Pearson $r = 0.46$) and concreteness ($r = 0.42$) showed the highest strengths of correlation, similar to their leading roles in predicting direction of metaphorical mappings. Intersubjectivity ($r = 0.22$) and embodiment ($r = 0.13$) were moderately correlated with the AI. These results suggest that domains that are external, concrete, embodied, and share a conceptualization across people, tend to serve as the source of metaphorical mappings across history. Finally, valence negatively correlated ($r = -0.19$) with the asymmetry index, suggesting that more valenced domains tend to be the target, also consistent with our previous findings.

Again taking into account the intercorrelated nature of the variables, we performed a multiple linear regression to examine the relative contributions of the six variables. We expected the most important (and orthogonal) predictor ratings to be weighted the highest when regressed against the asymmetry index. Figure 2b shows the relative contributions of these variables as reflected by their weights. The regression fit has a Pearson $r = 0.42$ ($p < 0.001$). The candidate with the largest absolute weight was externality ($p < 0.001$), followed by embodiment ($p < 0.09$) and valence ($p < 0.26$) although neither contributed significantly beyond externality. This result is largely consistent with the previous results where we found that the same set of variables accounted for most of the variance in in explaining metaphorical mapping directions during different historical epochs.

Table 2: Listing of the strongest source and target domains based on the source-target asymmetry index.

Strongest sources	**Strongest targets**
Textiles	Excitement
Supernatural	Pride
Digestive organs	Anger
Hardness	Hatred & hostility
Softness	Bad
Ruminants	Behaviour & conduct
Cultivated plants	Money
Wetness	Literature
Darkness	Fear
Solidity & density	Vigorous action & degrees of violence

Taken together, our diachronic and synchronic analyses suggest that metaphorical sense extensions are highly systematic and support efficient compression of emerging senses along a compact set of variable dimensions.

Conclusion

The present study focused on one of the principal mechanisms of polysemy generation: metaphorical sense mapping. We found that mappings recorded in a large historical database of English can be explained by a compact set of variables. Our investigation lends convincing support that

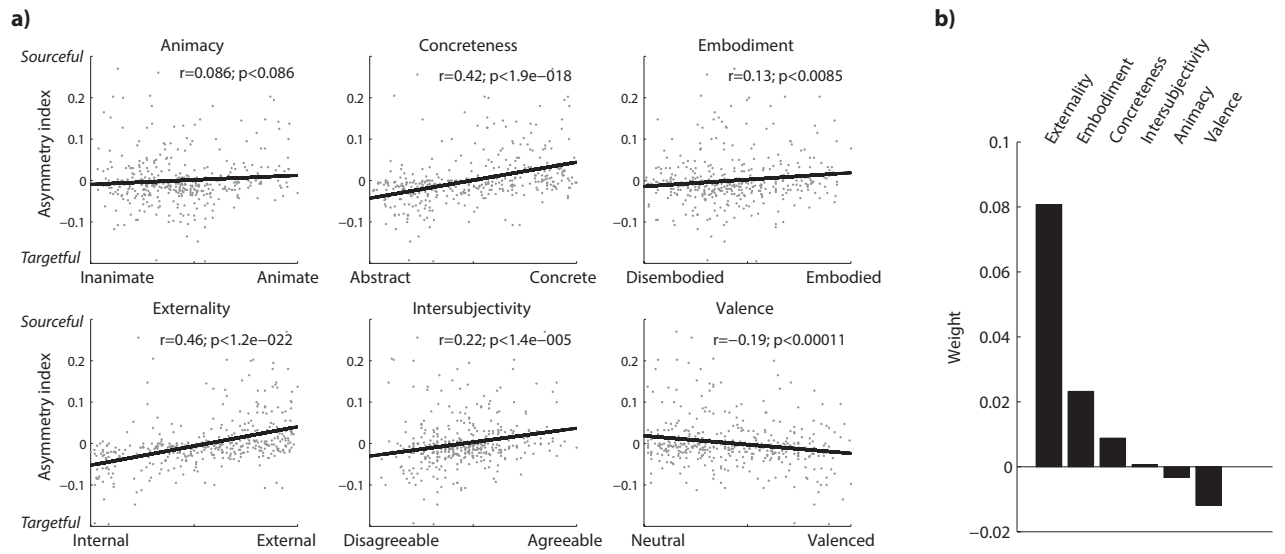


Figure 2: Summary of results on predicting source-target asymmetry of domains. a) Prediction from individual variables. b). Relative weighting of the variables in multiple linear regression.

the evolution of polysemy proceeds in systematic ways. It also offers a novel perspective on the nature of metaphor that bridges theories of conceptual metaphor and language evolution. That is, metaphor provides a cognitive device for compressing emerging concepts into existing words, hence facilitating a finite, learnable lexicon.

Our work also opens many questions for future research. For example, we assumed sense change at the domain level, whereas a full account should explain semantic change at the word level. Future work can also explore cognitive machineries other than metaphor that support polysemous sense extensions, the socio-cultural factors that explain the rise or fall in communicative needs for different semantic domains, and the degree to which these findings generalize beyond English. The current work has provided an empirical approach for in-depth explorations into the nature of meaning.

Acknowledgments

We thank Eve Sweetser and George Lakoff for helpful input and Jessie Salas and Elizabeth Kaplan for assistance with experiment development and data collection. This project was funded by NSF award SBE-1041707 to the Spatial Intelligence and Learning Center (SILC) and a University of California Faculty Startup Award to M. Srinivasan.

References

Blank, A., & Koch, P. (1999). *Historical semantics and cognition*. Walter de Gruyter.

Brugman, C. M. (1988). *The story of over: Polysemy, semantics, and the structure of the lexicon*. Taylor & Francis.

Christian, K., Roberts, J., Samuels, M., Wotherspoon, I., & Alexande, M. (2015a). *The Historical Thesaurus of English, version 4.2*. Glasgow: University of Glasgow. (http://historicalthesaurus.arts.gla.ac.uk/)

Christian, K., Roberts, J., Samuels, M., Wotherspoon, I., & Alexande, M. (2015b). Metaphor Map of English. In *Mapping Metaphor with the Historical Thesaurus.* Glasgow: University of Glasgow. (http://mappingmetaphor.arts.gla.ac.uk)

Geeraerts, D. (1997). *Diachronic prototype semantics*. Oxford University Press.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. The University of Chicago Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. *Metaphor and Thought*, *2*, 164–201.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The Handbook of Language Emergence*, *87*, 237.

Roberts, J., Christian, K., & Grundy, L. (2015). *A Thesaurus of Old English*. Glasgow: University of Glasgow. (http://oldenglishthesaurus.arts.gla.ac.uk/)

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. Dixon (Ed.), *Grammatical categories in Australian languages.*

Sweetser, E. (1991). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure* (Vol. 54). Cambridge University Press.

Traugott, E. C. (2003). From subjectification to intersubjectification. *Motives for Language Change*, *124*, 139.

Ullmann, S. (1957). *The principle of semantics*. Glasgow: Jackson, Son & Co.

Using Violations of Fitts' Law to Communicate during Joint Action

Cordula Vesper (VesperC@ceu.edu)
Laura Schmitz (Schmitz_Laura@phd.ceu.edu)
Günther Knoblich (KnoblichG@ceu.edu)
Department of Cognitive Science, Central European University
Október 6 utca 7, Budapest 1051, Hungary

Abstract

When people perform joint actions together, task knowledge is sometimes distributed asymmetrically such that one person has information that another person lacks. In such situations, interpersonal action coordination can be achieved if the knowledgeable person modulates basic parameters of her goal-directed actions in a way that provides relevant information to the less knowledgeable partner. We investigated whether systematic violations of predicted movement duration provide a sufficient basis for such communication. Results of a joint movement task show that knowledgeable partners spontaneously and systematically violated the predictions of Fitts' law in order to communicate if their partners could not see their movements. Unknowing partners had a benefit from these violations and more so if the violations provided a good signal-to-noise ratio. Together, our findings suggest that generating and perceiving systematic deviations from the predicted duration of a goal-directed action can enable non-conventionalized forms of communication during joint action.

Keywords: Joint action; signaling; coordination strategy; cooperation; communication; social cognition.

Introduction

When two or more people perform joint actions together, communication is often key to successful coordination. An obvious case is having a conversation (Clark, 1996), for instance, discussing the steps necessary to prepare dinner. But communication can also occur non-verbally, such as when someone waves to inform another of her presence or when nodding to indicate approval. These gestures are, like spoken language, purely communicative because they do not serve to achieve a specific action outcome – their exclusive purpose is to inform another person. However, there are many cases where the same action serves an instrumental purpose and informs another person at the same time: If a passenger occupying the window seat on a train starts stands up in a demonstrative way, then the instrumental purpose of her action is to leave her seat. At the same time she informs the person occupying the aisle about her intention to leave. Thus, there is a class of actions that concurrently serve instrumental as well as communicative goals (Pezzulo, Donnarumma, & Dindo, 2013).

What are the specific circumstances that trigger this class of actions? Previous studies have focused on joint actions where communication is needed because one person lacks information required to achieve a joint goal and therefore requires a knowledgeable partner to provide this information. It has been shown that knowledgeable partners exaggerate grip aperture (Sacheli, Tidoni, Pavone, Aglioti, & Candidi, 2013) or modulate kinematic properties such as movement amplitude or direction (Pezzulo et al., 2013; Vesper & Richardson, 2014) to convey knowledge their partners lack, even in interaction with young infants ('motionese'; Brand, Baldwin, & Ashburn, 2002).

There is also evidence that unknowledgeable partners perceive such modulations. For instance, it has been demonstrated that observers are sensitive to subtle kinematic differences in performance (Becchio, Sartori, Bulgheroni, & Castiello, 2008; Sartori, Becchio, & Castiello, 2011). However, in order to understand that a movement modulation is communicative, an observer will also need to understand the intended meaning of the modulation. Whereas verbal language and most gestures are conventional and thus based on associations between an arbitrary (linguistic) code and its meaning (Scott-Phillips, 2015), this may not be required to understand communicative modulations of instrumental actions. Instead, observers may understand kinematic signals by making use of their own motor system to predict the unfolding action of a communicator (Pezzulo et al., 2013; Wilson & Knoblich, 2005; Wolpert, Doya, & Kawato, 2003). Systematic deviations from these predictions may be taken as conveying particular meaning.

The aim of the present study was to ask whether movement duration provides a basis for establishing communication between a Leader and a Follower in this way. We hypothesized that Leaders should modulate movement speed to indicate target locations unknown to a Follower and that Followers should be able to use this information to choose where to move. Such communication should not depend on the Follower having visual access to the Leader's actions if the task context is shared. The reason is that motor simulation can be used to predict movement duration even in the absence of visual input (Umiltà et al., 2001; Vesper, van der Wel, Sebanz, & Knoblich, 2013).

Present study

We investigated whether and how people would communicatively modulate movement duration in a joint setting. Previous research demonstrated that observers can estimate with some precision the duration of (partially) hidden actions. For instance, observers can accurately predict when someone will re-appear behind an occluding object (Graf, Reitzner, Corves, Casile, Giese, & Prinz, 2007; Sparenberg, Topolinski, Springer, & Prinz, 2011). Here, we tested the hypothesis that joint action partners would use this temporal

prediction ability in the service of communication. Specifically, we hypothesized that actors would convey distance information by systematically modulating movement duration.

To this end, we instructed pairs of participants to perform goal-directed hand movements from a starting location to one of three target locations (Figure 1), where matching targets was the goal of the joint action. One member of the dyad was the 'Leader' and knew the target location; the other member was the 'Follower' and did not know the target location. The task was sequential. First the Leader moved to the target, then the Follower attempted to move to the same target. Importantly, an externally triggered tone marked the start of a new trial and a second tone was triggered when Leaders arrived at the target location. There were three different joint conditions (Table 1): In 'Vision', Leaders and Followers could see each other. In 'Pitch', co-actors could not see each other but a tone of different pitch for each of the three locations sounded when the Leader arrived at a particular target location. In 'None', no such immediate source of information was available as Followers could not see the Leaders and Leaders' target hits always produced tones with the same pitch.

For the None condition, we predicted that Leaders would communicatively use the interval between start and arrival tone to create a source of information that would help Followers perform their task. For the two baseline conditions Vision and Pitch, we did not expect any communicative modulation of movement speed from Leaders since both conditions contained explicit target information (either in visual or auditory form) to be picked up by Followers, making additional communication irrelevant (Wilson & Sperber, 2004). Thus, we expected Followers to make use of the respective source of information by simply observing Leaders' movements or by discriminating the different target pitches.

The distance of target locations from the starting point and the target size was chosen so that the duration of Leaders' movements was expected to be equal to all three target locations according to Fitts' law (Fitts, 1954). Building on this law, that defines a tradeoff between distance and target size, the targets were proportionally larger for longer distances between start and target location. Previous research has shown that Fitts' law holds in performance and motor imagery (Decety & Jeannerod, 1995) and also in action observation (Grosjean, Shiffrar, & Knoblich, 2007).

We expected Leaders to strategically modulate their movements to create distinguishable time intervals for the three different target distances in order to provide target information to Followers. Leaders' communication could be quantified as systematic violations of Fitts' law. We expected that such violations would occur in the None condition but not in the other two conditions. An alternative possibility for communication in the None condition is that Leaders delay initiation of their movement and keep movement duration unchanged. To tease apart these two possibilities, we analyzed the movement onset (interval between the external tone and the start of the movement) separately from the movement time (interval between the initiation of the movement and arriving at the target). If participants modulated movement time rather than movement onset this would imply that they chose to convey communicative information via the same channel used for the instrumental action, although other options are available that clearly separate communicative and instrumental information.

Table 1: Design and main hypotheses.

	Vision	Pitch	None
Visual information available?	**Yes**	No	No
Pitch information available?	No	**Yes**	No
Modulation of Leader's action duration expected?	*No*	*No*	***Yes***

Method

Participants

Eleven women and thirteen men participated in randomly-matched pairs (three women only pairs, four men only pairs). Participants were between 21 and 33 years old (M = 26.4 years, SD = 3.0 years), right-handed and had normal or corrected-to-normal vision. The members of three pairs knew each other before the experiment. They gave prior written informed consent, received monetary compensation and were debriefed about the study purpose at the end of the experiment. The experiment was performed in accordance with the Declaration of Helsinki. In each pair, one participant was randomly assigned to the experimental role of 'Leader' and the other to the role of 'Follower'.

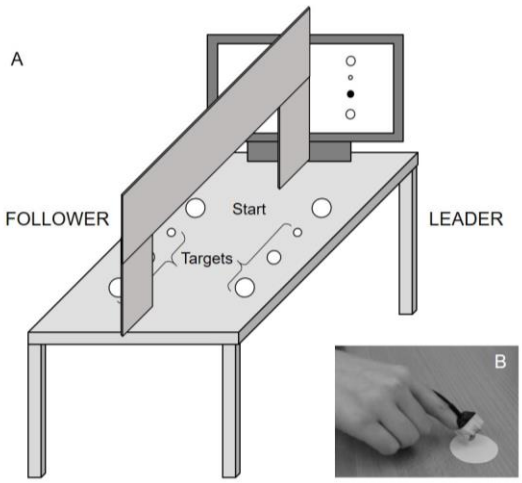


Figure 1: A: Schematic drawing of the experimental setup. Only Leaders received target information on the computer screen. B: Placement of the motion capture sensor.

Apparatus

An interactive real-time motion-capture setup was created for the purpose of the present experiment (Figure 1A). It

consisted of a table with a row of four circles on each long side. The circle diameters were 4.8 cm for the start locations and 1.6 cm, 3.2 cm and 4.8 cm for the three targets. All circles were centrally aligned with a center-to-center distance of 20 cm. The index of difficulty (ID) according to Fitts' law (Fitts, 1954; ID = $\log_2(2A/W)$ with A: movement amplitude, W: target width) was 4.64 for all three targets.

The two participants were seated comfortably at the table. A cardboard partition in the middle of the table had an opening in the middle (85 cm long, 35 cm high) that could be covered with a black opaque cloth to prevent visual contact between participants. The partition also separated the stimulus display on a 24" Asus computer screen (resolution 1920 x 1080 pixels, refresh rate 60 Hz) such that Leaders and Followers could be presented with different information. The interactive setup was controlled online with a Polhemus G4 electro-magnetic motion capture system (www.polhemus.com) that recorded participants' movement data with a constant sampling rate of 120 Hz. Movement sensors were taped centrally onto the nail of each participant's right index finger (Figure 1B). The experimental procedure and data recording was controlled by Matlab 2014a.

Procedure

Participants received written instructions, which were verbally repeated before each block of the experiment. Leaders were explicitly made aware that they would need to help their partner: "Your partner will not know which target position is the correct one. For your partner, the only information about the target can come from your action. Your goal is to help your partner such that she/he will be able to reach the correct target position as fast as possible." Correspondingly, the Followers' instructions emphasized that their partner would have relevant task information: "Your partner will know which target position is the correct one, but you will not know this in advance – for you, the only information about the target can come from your partner's action. Your partner's goal is to help you such that you will be able to reach the correct target position as fast as possible." Participants were also informed that they were not allowed to speak to each other.

The first block, an individual training ('Individual'), was completed only by the participant with the Leader role, while the Follower waited in a separated part of the room. Afterwards, both participants performed three blocks of trials together with short breaks in between (Table 1). In the 'Vision' block, co-actors could see each other's hand movements. In the 'Pitch' block, visual access was prevented but Leaders' target arrival triggered differently pitched tones. In the 'None' block, neither visual access nor pitch information was available. The order of the three joint blocks was counterbalanced across participant pairs.

Each block began with a short calibration procedure to acquire the spatial coordinates of participants' finger positions at start and targets to guide the online control of the experiment by the motion capture system. Then, after three training trials to allow participants to get acquainted with the block's specific procedure, 72 experimental trials were performed (24 trials per target, in random order). The experiment took about one hour in total.

All trials followed the same procedure: Participants first moved with their index finger to the starting location as prompted on the computer screen. Once the Leader (individual training) or both Leader and Follower (joint conditions) were in the start location, the Leader's side of the computer screen displayed the target location and a short tone was played (80 ms, 659 Hz). The Leader now moved to the target at her own speed. Upon target arrival, which was detected by online evaluation of the real-time motion tracking data, a second short tone was played. Its frequency depended on the respective condition: In Individual, Vision and None, the same tone was played for all targets (659 Hz). In Pitch, the frequency varied for the three targets (1109 Hz for the first target, 1319 Hz for the second target, 1661 Hz for the third target). The Follower's task was to then perform a speeded hand movement from her own starting location to the same target as the Leader. Subsequently, the screen would turn green or red (for 300 ms), indicating whether Leader and Follower had moved to matching or non-matching targets, respectively. After an inter-trial interval of 700 ms, the next trial began.

In all blocks, Leaders were instructed to not touch the target locations directly and instead end their movements at a point slightly above the table. This was done to prevent any noise when Leaders hit a target which could potentially give directional auditory cues to Followers. Followers were instructed to touch the targets directly.

Data preparation and analysis

From Leaders' movement time series, three time intervals were extracted. 'Time-to-target' (TT) was defined as the interval between the computer-generated start tone and Leaders' movement offset, i.e. the moment when they reached a target position (offset criterion based on the measured calibration points: horizontally inside a radius of 0.8 cm / 1.6 cm / 2.4 cm and vertically below 1 cm). 'Movement onset' (MO) was defined as the interval between the computer-generated start sound and the Leaders' movement onset (onset criterion based on the measured calibration points: horizontally outside of a 2.4 cm radius or vertically above 1 cm), while 'movement time' (MT) was defined as the interval between Leaders' movement onset and offset. Thus, TT equaled the sum of MO and MT. All trials in which Leaders moved to the wrong target or in which TT exceeded two standard deviations around the mean were excluded per Leader and condition from further analysis (4.1 % of all data).

From the remaining trials, we calculated signal-to-noise ratios (SNR) as measures for Leaders' signal clarity. Specifically, the SNR of TT combines the difference between the mean TTs for the three different targets and the variability of these TTs. Thereby, it captures in one measure how distinct Leaders' timing (= signal) is in relation to its

variability (= noise). SNR was calculated for each participant and each condition as the averaged difference between mean TTs for adjacent targets, divided by the overall standard deviation of all TTs (across targets), as described by the equation

$$SNR_{TT} = \frac{((M(TT_{tg2}) - M(TT_{tg1})) + (M(TT_{tg3}) - M(TT_{tg2}))) / 2}{SD(TT_{all\ targets})}$$

where *M* designates the mean, *SD* the standard deviation, and tg1 to tg3 refer to the three target locations. Higher SNR values indicate a clearer signal. In order to test which part of the movement was modulated, we calculated a SNR not only for TT (SNR_{TT}) but correspondingly also for MO (SNR_{MO}) and MT (SNR_{MT}).

For the analysis of movement velocity, we first filtered all trajectories using a 4th-order Butterworth digital filter with cut-off at 10 Hz and then calculated Leaders' mean velocity along the horizontal axis on which the targets were aligned. Finally, to assess joint task performance, trials in which Followers moved to the same target as the Leaders were classified as a match and when they moved to a different target as a mismatch. Based on this, a percentage of target matches per total number of trials was calculated. All data preparation was done with Matlab 2015a and significance testing with IBM SPSS 22.

Results

Modulation of action duration

To investigate whether Leaders adapted their action performance to inform Followers about the target location, we first compared the signal-to-noise ratio for the overall time-to-target (SNR_{TT}; corresponding to the complete interval between the two tones) in the three joint conditions. As predicted, the SNR_{TT} was significantly higher in None (2.33) compared to Vision (.86), $t(11) = 4.89$, $p < .001$, Cohen's $d = 2.95$, and Pitch (1.04), $t(11) = 4.71$, $p < .001$, Cohen's $d = 2.84$. This indicates that Leaders provided a clearer (i.e. more distinct and more consistent) signal when Followers did not have other means to determine to which target they should move.

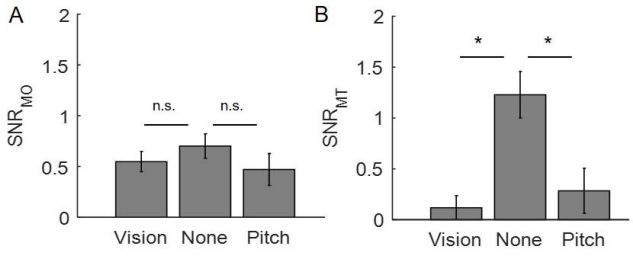


Figure 2: Results of the signal-to-noise ratio analysis for A: movement onset (SNR_{MO}) and B: movement time (SNR_{MT}). Error bars show the standard error.

In a next step, we performed the same analyses separately for Leaders' signal-to-noise ratios of the movement onset (SNR_{MO}) and the movement time (SNR_{MT}) to determine whether they rather chose to wait longer before moving or to slow down their movements. For SNR_{MO} (Figure 2A), None (.7) was not significantly different from Vision (.55), $t(11) = 1.36$, $p > .2$, Cohen's $d = .82$, or from Pitch (.47), $t(11) = 1.25$, $p > .2$, Cohen's $d = .75$. In contrast, for SNR_{MT} (Figure 2B), None (1.23) was significantly larger than Vision (.12), $t(11) = 5.89$, $p < .001$, Cohen's $d = 3.55$, and also than Pitch (.28), $t(11) = 4.55$, $p < .001$, Cohen's $d = 2.75$. Thus, Leaders adapted the execution part of their movements to provide a communicative signal instead of waiting longer before initiating the movement.

Violation of Fitts' law

Given that Leaders chose to provide a communicative signal by changing their movement time, we tested the follow-up hypothesis that they would do so by moving with constant mean velocity irrespective of the target location, which would effectively create a violation of Fitts' law. To this end, we compared Leaders' movement times and mean velocities in None to baseline performance acquired from Leaders' individual training, for which, based on previous research, we expected Fitts' law to hold. Accordingly, we conducted within-subjects ANOVAs to test whether there were significant interactions of the factors condition (Individual, None) and target (first, second, third).

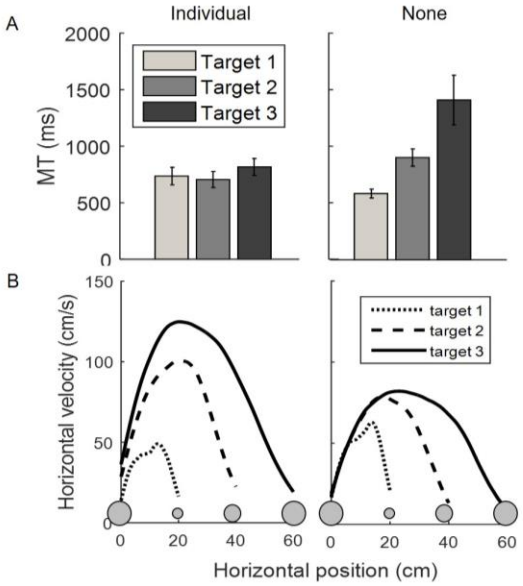


Figure 3: A: Movement times and B: time-normalized grand-average velocity profiles in Individual and None conditions, shown separately for each target.

In line with our predictions, there was a significant interaction of condition and target for movement time (Figure 3A), $F(2,22) = 11.76$, $p < .01$, $\eta_p^2 = .52$, indicating that MT was more strongly influenced by the target position in None than in Individual. There was also a main effect of target, $F(2,22) = 14.92$, $p < .001$, $\eta_p^2 = .57$, and a close-to-significant effect of condition, $F(1,11) = 4.66$, $p = .054$, $\eta_p^2 = .3$.

The analysis of mean velocity (Figure 3B) showed a corresponding pattern of results, suggesting that Leaders indeed created duration differences between targets. In particular, there was a significant interaction effect of condition and target, $F(2,22) = 16.46$, $p < .001$, $\eta_p^2 = .6$, as well as main effects for target, $F(2,22) = 33.4$, $p < .001$, $\eta_p^2 = .75$, and condition, $F(1,11) = 5.03$, $p < .05$, $\eta_p^2 = .31$. Further separate one-way ANOVAs for each condition confirmed that velocity was significantly different for the three targets in the individual baseline, $F(2,22) = 52.06$, $p < .001$, $\eta_p^2 = .83$ (all pair-wise comparisons $p < .001$), but not for None, $F(2,22) = 3.07$, $p > .07$, $\eta_p^2 = .22$ (all pair-wise comparisons $p > .2$).

Joint task performance

Finally, we analyzed the effects that Leaders' signaling performance had on the joint task accuracy, i.e. on how well Followers understood the communicative signal and moved to the correct target location. An analysis of the percentage of target matches showed that dyads' performance suffered from the lack of immediately available perceptual information: Dyads had significantly fewer target matches in None (63.1 %) than in Vision (94.1 %), $t(11) = -6.87$, $p < .001$, Cohen's $d = -4.14$, or Pitch (80.4 %), $t(11) = -3.87$, $p < .01$, Cohen's $d = -2.33$.

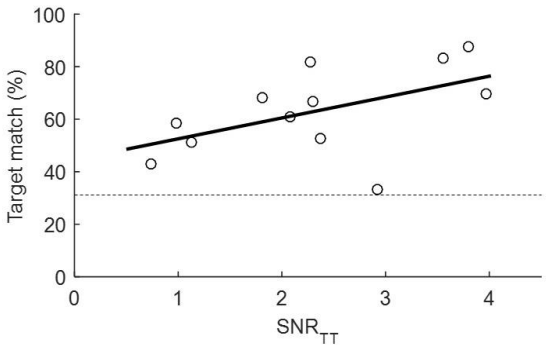


Figure 4: Leaders' signaling behavior as indicated by the signal-to-noise ratio of time-to-target (SNR_{TT}) plotted against dyads' joint performance, i.e. how well Followers managed to move to the corresponding target. The dotted line shows chance performance at 33%.

In the next step, we tested whether those dyads whose Leader provided a better signal in the None condition also succeeded better in moving to the same target locations. To achieve this we correlated pairs' percentage of target match with Leader's signal-to-noise ratio of the overall time-to-target, expecting a positive correlation of the two. Although the correlation (Figure 4) did not reach significance, $r = .521$, $p = .08$, the rather high correlation coefficient suggests a relationship between how good a Leader signals and how well a Follower understands and uses the given information.

Discussion

To increase our understanding of how communication based on instrumental actions is created and used, the present study investigated whether movement duration provides a basis for establishing communication between a Leader and a Follower in a joint action. In our task, Leaders performed movements to target locations unknown to Followers who then attempted to quickly move to the same target location. Crucially, by adjusting target sizes and distances according to Fitts' law (Fitts, 1954), we effectively created a situation in which information could not be transmitted directly via the visual or auditory modality and where established communication systems could not be used. We hypothesized that Leaders would create a new communication system by modulating movement duration to inform Followers and that Followers would be able to use this information to choose which location to move to.

In line with our predictions, we found that Leaders modulated the duration of their movements to indicate different target locations but only if the Follower had no direct visual or pitch information about their actions. This finding demonstrates the specificity of Leaders' communication to contexts in which receiving information was relevant for the joint action partner (Wilson & Sperber, 2004). The present study extends previous work on signaling in interaction contexts that highlighted the role of spatial movement parameters (Pezzulo et al., 2013; Sacheli et al. 2013; Vesper & Richardson, 2014) by demonstrating that movement duration provides a further potential communication channel. Possibly, both producing and understanding communicative signals with action duration is based on motor simulation processes (Wolpert et al., 2003) in which differences between internally predicted and actual action duration would be taken as communicative deviations from natural performance.

Our task was designed such that Leaders had two options of modulating movement duration – either by waiting longer before initiating their movement or by slowing down the movement. The results show that Leaders chose the latter option. This is important for two reasons: First, this result supports the idea that signaling is based on generating predictions and exaggerating aspects of motor performance to allow another person to distinguish between action alternatives. Rather than strategically delaying action onset, Leaders' ongoing actions were systematically sped up or slowed down, although the Follower could not perceive the movements themselves. Second, informing Followers required Leaders to accept a violation of Fitts' law for their own movement execution in order to create distinguishable action durations. An analysis of the velocity profile of Leaders' movements confirmed this by showing that Leaders kept their mean velocity to all targets constant, although the targets greatly differed in size. In fact, keeping velocity constant across targets may have been the most straightforward way to create distinct movement durations for different target locations.

Successful communication presupposes that recipients benefit from the information that was communicated. This was clearly the case for the Followers in our experiment who benefitted from the Leaders' communication when

trying to match the target location. Although dyads' task accuracy was lower compared to the conditions where explicit visual or pitch information about the target location was available, all pairs except one managed to perform better than chance. Note that the chance performance in one pair is entirely due to a Follower who, throughout the entire interaction, failed to understand signals with a very good signal-to-noise ratio (see Figure 4). Although follow-up studies would be required to characterize good Leaders and Followers, mentalizing and perspective-taking abilities are likely to play an important role as both informing and understanding modulations of movement as a communicative act likely requires taking the partner's task knowledge and access to perceptual information into account (see also Volman, Noordzij, & Toni, 2012).

Were Leaders aware of their signaling strategy? Although we do not have quantitative evidence, it is likely that actors in the present study were indeed aware that they modulated movement duration to communicate target location to Followers. Most Leaders reported during debriefing that they distinguished between different targets by using different movement times. Similarly, Follows were aware of what feature of the Leaders' action they used to decide to which target they would move. However, there is a possibility that awareness is not always necessary to be successful with the type of communication observed here: Because communication is embedded into the execution of a joint action, people may not always realize that they facilitate performance for a partner as long as the partner detects deviations from standard performance and effectively uses the information to achieve the jointly planned outcome.

Taken together, the present study provides evidence that joint action coordination can benefit from communicative modulations that violate predictions about instrumental actions. Generating and perceiving systematic deviations from the predicted duration of a goal directed action was sufficient to enable an effective non-conventionalized form of communication during joint action.

Acknowledgments

This research was supported by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n° [609819], SOMICS, and by ERC grant agreement n°616072, JAXPERTISE.

References

Becchio, C., Sartori, L., Bulgheroni, M. & Castiello, U. (2008). The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. *Consciousness and Cognition, 17*, 557-564.

Brand, R. J., Baldwin, D. A. & Ashburn, L. A. (2002). Evidence for `motionese': modifications in mothers' infant-directed action. *Developmental Science, 5*, 72-83.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Decety, J., & Jeannerod, M. (1995). Mentally simulated movements in virtual reality: Does Fitts' law hold in motor imagery? *Behavioral Brain Research, 72*, 127–134.

Fitts, P. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*, 381-391.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M. & Prinz, W. (2007). Predicting point-light actions in real-time. *NeuroImage, 36*, T22–T32.

Grosjean, M., Shiffrar, M., & Knoblich, G. (2007). Fitt's law holds in action perception. *Psychological Science, 18*, 95-99.

Pezzulo, G., Donnarumma, F. & Dindo, H. (2013). Human Sensorimotor Communication: A Theory of Signaling in Online Social Interactions. PLoS ONE, 8, e79876.

Sacheli, L., Tidoni, E., Pavone, E., Aglioti, S. & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research, 226*, 473-486.

Sartori, L., Becchio, C., & Castiello, U. (2011). Cues to intention: the role of movement information. *Cognition, 119*, 242–252.

Scott-Phillips, T. C. (2015). *Speaking Our Minds.* Palgrave Macmillan.

Sparenberg, P., Springer, A. & Prinz, W. (2011). Predicting others' actions: evidence for a constant time delay in action simulation. *Psychological Research, 76*, 41-49.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C. & Rizzolatti, G. (2001). "I know what you are doing": a neurophysiological study. *Neuron, 32*, 91-101.

Volman, I., Noordzij, M. L. & Toni, I. (2012). Sources of variability in human communicative skills. *Frontiers in Human Neuroscience, 6*, 310.

Vesper, C. & Richardson, M. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research, 232*, 2945-2956.

Vesper, C., van der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2013). Are you ready to jump? Predictive mechanisms in interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 48-61.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In Horn, L.R. & Ward, G. (eds.) *The Handbook of Pragmatics.* Oxford: Blackwell, 607-632.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460–473.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and interaction. *Philosophical Transactions of the Royal Society of London B, 358*, 593–602.

Perceived Momentum Influences Responsibility Judgments

Jeffrey R. Parker (jeffparker@gsu.edu)
Marketing Department, Georgia State University
PO Box 3991, Atlanta, GA 30302 USA

Iman Paul (iman.paul@scheller.gatech.edu)
Marketing Department, Georgia Institute of Technology
800 West Peachtree NW, Office 4000, Atlanta, GA 30308 USA

Nicholas Reinholtz (nicholas.s.reinholtz@colorado.edu)
Marketing Department, University of Colorado Boulder
419 UCB, Boulder, CO 80309 USA

Abstract

This work examines the extent to which people hold independent sequential events (e.g., players making correct/incorrect guesses) responsible for overall outcomes (e.g., the team winning/losing the game). Two types of events are found to garner the majority of responsibility for overall outcomes: (1) final events and (2) events that are perceived to disrupt momentum (e.g., an incorrect guess after a sequence of correct guesses). While previous research has shown that final events tend to be perceived as more responsible for overall outcomes, the current experiments are the first to document the role of perceived momentum on responsibility judgments. Specifically, we demonstrate that the effect is mediated by perceived momentum changes after the time of the event and moderated when exogenous factors (e.g., a delay between events) disrupt perceived momentum. We discuss how these findings relate to *pivotality*, the *counterfactual simulation model*, and the role of *unexpectedness* in responsibility judgments.

Keywords: perceived momentum; responsibility judgments; perceived causality

Introduction

Many *outcomes* are the result of a collection of sequential *events*: Sports championships are determined by the results of games played during the regular and playoff seasons, the success or failure of parlay bets depends on the results of multiple independent events (games, races, or the like), final grades are a function of the student's performance on multiple assignments and exams, and the winners of games such as chess are determined by the set of moves each player has made during the match.

We examine how people assign responsibility for a given outcome to the events (which occurred sequentially) that resulted in that outcome. Of particular interest is how people assign responsibility to *aligned events*. Aligned events are those events whose valence corresponds with that of the outcome. For instance, imagine a group of people playing a guessing game in which each player must guess which of two options is the "correct" option. The players make their guesses in secret and sequentially (one after the other). If a certain percentage of the group guesses correctly (incorrectly) then the group will win (lose) the game. Consequently, if the game outcome is a win (loss), the aligned events would be the correct (incorrect) guesses. Taking the loss outcome as an example, our research is interested in the amount of responsibility for the group's loss that is assigned to each incorrect guess (or, equally, to each player who gave an incorrect guess).

Responsibility judgments are at least partially derived from causal inferences (Chockler & Halpern, 2004): It is unlikely that one will be held responsible for an outcome if one's actions did not make some difference to that outcome (Gerstenberg, Halpern, & Tenenbaum, 2015). Further, the closer a given individual's actions are to being *pivotal* to the outcome, the more responsible the individual will be held for the outcome (Bartling, Fischbacher, & Schudy, 2015; Chockler & Halpern, 2004; Lagnado, Gerstenberg, & Zultan 2013). However, both causal and responsibility judgments are highly nuanced, drawing on many inputs, as exemplified by the *counterfactual simulation model* (CSM; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015).

Using the current context of events and outcomes, the CSM distinguishes between four different counterfactual contrasts. *Whether-causation* refers to the perceived probability that the outcome would have been different had a specific event been altered or removed. In contrast, *how-causation* pertains to the perceived probability that the outcome would have occurred in exactly the same manner—not whether it would have occurred or not—had a specific event been altered or removed. *Sufficiency* is the perceived probability that the same outcome would have been reached if all candidate causes other than the target event were removed. Finally, *robustness* is the perceived probability that a causal relationship between a specific event and the outcome would remain unchanged were the circumstances (i.e., the nature of the alternative candidate causes) altered in some small way.

The current research examines contexts like the guessing game described above in which the outcome would change if any one of the aligned events had a different result.

Consequently, the aligned events are all equally pivotal to the outcome and are all whether-causes. Moreover, all of the aligned events are necessarily how-causes as each influences the manner by which the outcome is reached. Lastly, none of the aligned events is a sufficient or robust cause of the outcome as changing any other aligned event would change the outcome.

Importantly, the equivalence of the aligned events in terms of pivotality and the CSM's counterfactual contrasts does not depend on the timing of the events in the sequence leading to the outcome. More concretely, a given event's pivotality, whether-causation, how-causation, sufficiency, and robustness will be the same whether that event was the first or last event in the sequence. Yet, as is argued and demonstrated below, responsibility judgments are impacted by the sequence in which events occur.

The Sequence of Events

Taking the loss outcome in the guessing game described above as an example once more, we contrast two specific sequences of events for a game in which the team is allowed no more than three incorrect guesses. The aligned events (guesses) for this example are bolded in each sequence ("I" = incorrect, "C" = correct).

1. **All-Late**: C_1-C_2-C_3-C_4-C_5-C_6-**I_1-I_2-I_3-I_4**
2. **One-Late**: **I_1-I_2-I_3**-C_1-C_2-C_3-C_4-C_5-C_6-**I_4**

There are three things to note about the above sequences. First, four events (incorrect guesses) align with the overall outcome of the game (loss) in both sequences: Had there only been three incorrect guesses, the team would have won in either case. Second, the players made their guesses in secrecy and, therefore, could not know the other players' guesses or how the group was performing until the conclusion of the game (i.e., the guesses were explicitly independent). Third, since the players' guesses were collected sequentially, the outcome (loss) was not determined until the tenth player made his guess.

Given this context, prior research would suggest that the final aligned event (I_4) will tend to be judged more responsible for the outcome of the game (Miller & Gunasegaram, 1990) in both sequences. Yet, no research to date would predict that responsibility judgments for the aligned events will differ between the all-late and one-late sequences. In fact, since changing the result of any aligned event would change the game's outcome, models of responsibility judgments relying on pure pivotality (Chockler & Halpern, 2004), counterfactual pivotality (Zultan, Gerstenberg, & Lagnado, 2012), or the CSM would predict no differences in responsibility judgments across the two sequences of events. However, we contend that differences in the timing of the aligned events between the two sequences can meaningfully impact causal responsibility judgments via perceptions of momentum.

Perceived Momentum

Although rooted in physics, the concept of *momentum* also holds a place in social cognition and refers to the tendency of, or expectation for, a person or group to repeat recent success or failure. In other words, repeated success (or failure) is expected to beget a higher likelihood of subsequent success (failure).

When sequential events are not independent (e.g., the same sports team playing repeated games) perceptions of momentum may be grounded in rational reasoning. Indeed, in such contexts, it is reasonable to predict that earlier success will increase the likelihood of subsequent success via the boosted confidence or increased learning and ability of the individual or group. Empirical support for actual momentum in such contexts is still a matter of debate (Gilovich, Vallone, and Tversky, 1985; Green and Zweibel, 2013; Miller and Sanjurjo, 2015), but perceptions of momentum persist nonetheless.

We contend that people overgeneralize their beliefs and perceptions of momentum from contexts where it could reasonably exist (i.e., when events are not independent) to contexts where it cannot (i.e., when events are objectively and explicitly independent). We return to our guessing game example in which players make guesses sequentially and in secret. A series of repeated correct guesses in this game may be the result of many factors (e.g., obvious answers or talented players), but it cannot be the result of the players or the team gaining momentum via confidence or increasing ability since each player's guess is objectively independent of those of the other players. Yet, this series of repeated correct guesses can easily be mapped onto contexts in which people normally (and, perhaps, normatively) perceive momentum. Hence, observing repeated successes (or failures) may result in perceptions that the team of players, despite making objectively independent guesses, does have momentum.

Perceiving momentum, even when events are objectively independent, would in turn lead to the inference that succeeding events are more likely to stay the course (e.g., success following repeated successes) than to deviate from it (e.g., failure following repeated successes). Thus, deviations from the direction of perceived momentum should be considered more unexpected. Since aligned events that are more unexpected tend to be held more responsible for outcomes (Gerstenberg, Halpern, & Tenenbaum, 2015), aligned events that deviate from perceived momentum should be considered more responsible for the outcome.

This reasoning suggests that the degree of responsibility assigned to aligned events may differ between the *all-late* and *one-late* sequences described above. Previous research suggests that the last aligned event (I_4) is disproportionately likely to be assigned blame in both conditions. However, if people's responsibility judgments are influenced by the perception of momentum, and consequent relative unexpectedness of the aligned events, the tendency to assign blame to the last event should be

stronger in the one-late sequence than in the all-late sequence. In the one-late sequence, the result of the final event terminates a series of six favorable results that preceded it. Conversely, perceived momentum should lead to a higher likelihood of assigning responsibility to the first aligned event (I_1) in the all-late sequence (in which its result deviates from the results of the previous six events) than in the one-late sequence (in which it is not preceded by any events). The interim aligned events (I_2 and I_3) should be relatively uninfluenced by the sequence as they are always preceded by at least one other aligned event.

Summary

We examined whether people perceive momentum in a series of events where the results were sequential and objectively independent. We predicted an event aligned with the overall outcome (e.g., an incorrect guess in a game whose outcome was a loss) would be assigned greater responsibility if that aligned event directly followed a series of unaligned events (e.g., an incorrect guess after a series of repeated correct guesses).

The experiments that follow support this prediction. Experiments 1 and 2 show that aligned events were assigned significantly greater responsibility when directly preceded by a series of unaligned events. Experiment 3 finds that judgments of perceived momentum mediate this effect. Experiment 4 finds the effect can be moderated by disrupting perceived momentum. Finally, Experiment 5 finds that the expectation of success for a given event is higher, all else being equal, when that event is preceded by successes versus failures.

Experiments

Experiments 1 and 2 tested the proposition that

(1) the tendency to hold the *final aligned event* responsible for the overall outcome would be higher in the one-late (vs. all-late) sequence (both described above) and

(2) the opposite would hold for the *first aligned event*.

Experiment 1: The Influence of Event Sequence on Responsibility Judgments (Losses)

Method. One hundred eighty-eight paid Amazon Mechanical Turk (AMT) participants were told of a hypothetical team of 10 players playing a guessing game in which each player guessed which of two colors a computer had randomly chosen before the game began. The players were indicated to have made their guesses sequentially and in secret, such that no player knew what previous players had guessed or how the team was doing at the time they made their guess. If the team had three or fewer incorrect guesses it would win. Otherwise, it would lose. All participants were told that four players had made incorrect guesses and that the team had lost the game.

Participants were randomly assigned to either the one-late or all-late sequence described above. To reiterate, (1) the first three guesses and the final guess were incorrect (i.e., were aligned with the outcome) in the one-late sequence, while (2) the final four guesses were incorrect (aligned with the outcome) in the all-late sequence.

Participants were shown the complete set of results (correct/incorrect guesses) in a table ostensibly organized in the order in which players had made their guesses. They were then asked which of the four players who had guessed incorrectly was the most to blame (i.e., was most responsible) for the team's loss. Participants could indicate that either all four players were equally to blame or that no player was to blame if they desired. (In all experiments, the percent of participants indicating equal/no blame did not significantly differ between the one-late and all-late sequences. Thus, although these equal/no-blame responses are retained and included in the data analyzed in each experiment, they are not discussed further.)

Results. As expected, the percentage of participants indicating that the player making the final incorrect guess (i.e., the final aligned event) was the most to blame for the team's outcome was significantly lower in the all-late (44%) versus the one-late (66%) sequence ($\chi^2(1) = 7.97$, $p < .005$). Conversely, a larger proportion of participants felt the player making the first incorrect guess (i.e., the first aligned event) was the most to blame in the all-late (25%) versus the one-late (1%) sequence ($\chi^2(1) = 26.10$, $p < .001$). These results are consistent with our contention that people (1) perceive momentum in contexts where sequential events are objectively independent and (2) use this perceived momentum in forming responsibility judgments for the overall outcome.

Experiment 2: The Influence of Event Sequence on Responsibility Judgments (Wins)

Method. One hundred ninety-nine paid AMT participants participated in Experiment 2, which was identical to Experiment 1 except the following: (1) the team needed at least four correct answers to win, (2) the team's outcome was a win, and (3) the correct and incorrect responses were switched in the two sequences (e.g., the all-late sequence featured six incorrect guesses before concluding with four correct guesses). Thus, the aligned events in Experiment 2 were the four players that had guessed correctly. After viewing the results table, participants were asked which of the four players that had guessed correctly deserved the most credit (i.e., was most responsible) for the team's win.

Results. Replicating Experiment 1 in the domain of wins, the percentage of participants indicating that the player making the final correct guess (i.e., the final aligned event) deserved the most credit for the team's outcome was significantly lower in the all-late (68%) versus the one-late (84%) sequence ($\chi^2(1) = 6.82$, $p < .02$). Conversely, a larger proportion of participants felt the player making the first

correct guess (i.e., the first aligned event) was the most to credit in the all-late (16%) versus the one-late (0%) sequence ($\chi^2(1) = 17.23$ $p < .001$). Again, it seems that participants perceived momentum in the described game and used this in forming their responsibility judgments. Experiment 2 also demonstrated that this result is not limited to contexts with negative outcomes.

Experiment 3: Testing the Role of Perceived Momentum via Mediation

While the results of the first two experiments are consistent with our contention that people perceive momentum in contexts where sequential events are objectively independent and use this perceived momentum in forming responsibility judgments, they do not provide any direct evidence of this process. Accordingly, Experiment 3 was designed to explicitly measure perceptions of momentum in the sequences examined in Experiment 1.

Method. Two hundred ninety-four paid AMT participants participated in Experiment 3, which used the same sequences (one-late vs. all-late), team outcome (loss), and dependent variable (most to blame) as Experiment 1. However, prior to indicating which aligned event (incorrect guess) was the most to blame for the outcome, participants first indicated how much momentum they felt the team had *after* each player's guess (-5 = strongly negative momentum, 0 = no momentum, +5 = strongly positive momentum). This measure of perceived momentum was collected to test whether changes in perceived momentum mediated the influence of sequence (one-late vs. all-late) on responsibility judgments.

Results. As in Experiments 1 and 2, a smaller proportion of participants felt the player making the final incorrect guess (i.e., the final aligned event) was the most blame in the all-late (37%) versus the one-late (60%) sequence ($\chi^2(1) = 14.93$, $p < .001$). Conversely, a larger proportion of participants felt the player making the first incorrect guess (i.e., the first aligned event) was the most to blame in the all-late (38%) versus the one-late (12%) sequence ($\chi^2(1) = 27.22$, $p < .001$).

We further explored the role of perceived momentum in forming responsibility judgments. For both sequences, perceived momentum peaked after the sixth consecutive success (see Figure 1). Examining the individual responses of the 83% of participants who indicated that a specific event (i.e., player) was most to blame, 60% assigned blame to the event that corresponded with the biggest drop in perceived momentum. Of the 17% of participants indicating equal/no blame, 39% also reported no changes in perceived momentum over the course of the events, further suggesting a link between perceived momentum and assigned blame. Notably, only 7% of the total number of participants indicated no changes in perceived momentum across the sequences of events.

We conducted a series of regressions to examine the mediating role of perceived momentum on the likelihood of assigning blame to the fourth aligned event (which was the tenth event in both sequences). First, a logistic regression with sequence (one-late vs. all-late; effect coded 1 and -1 respectively) as the sole predictor showed a significant effect on assigned blame ($\beta_{sequence} = .46$, $z = 3.83$, $p < .001$; this mirrors the results from the previously described chi-square test). Second, a regression with sequence (effect coded as before) as the sole predictor showed a significant effect on the *change* in perceived momentum associated with the final aligned event ($\beta_{sequence} = -1.84$, $t = -13.57$, $p < .001$). Finally, when both sequence (effect coded as before) and change-in-momentum were used as predictors in a logistic regression, change-in-momentum was a significant predictor of blame ($\beta_{change\text{-}in\text{-}momentum} = -.18$, $z = -3.41$, $p < .001$) while sequence no longer held significant predictive power ($\beta_{sequence} = .14$, $z = .91$, $p = .36$). This suggests perceived momentum may play a mediating role in the formation of responsibility judgments.

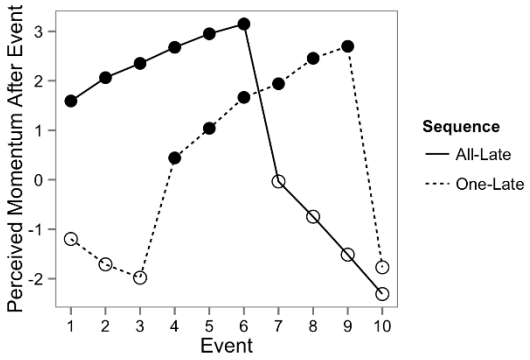


Figure 1. Perceived momentum in Experiment 3. Non-solid circles indicate incorrect guesses (aligned events).

Experiment 4: Testing the Role of Perceived Momentum via Moderation

If responsibility judgments are partially derived from perceptions of momentum, then disrupting the momentum perceived to exist in a series of events should mitigate its impact on responsibility judgments. Experiment 4 tested this prediction.

Method. One hundred ninety-five paid AMT participants participated in Experiment 4, which used the same sequences (one-late vs. all-late), team outcome (loss), and dependent variable (most to blame) as Experiment 1. However, Experiment 4 added a second factor that was crossed with the sequence factor: whether or not participants were told the game was unexpectedly delayed immediately prior to the 10th player's guess (i.e., the final aligned outcome in both sequences). It was expected that an unexpected delay in the game would disrupt the perceived momentum occurring the in the game at that

point. Accordingly, the sequence of events should be less informative for responsibility judgments for the final aligned event when it is known that this event was preceded by an unexpected delay in the game.

To summarize, Experiment 4 used a 2 (sequence: one-late vs. all-late) x 2 (game delay: yes vs. no) between-subjects design. We predicted a significant effect of sequence on blame assigned to the final player when there was no game delay (as in previous studies), but not when there was a game delay.

Results. As expected, there was a significant interaction between the sequence (dummy-coded; 1 = one-late, 0 = all-late) and game-delay factors (dummy-coded; 1 = no delay, 0 = delay; $\beta = 1.85$, $p < .002$). When there was no game delay, a smaller proportion of participants felt the player making the final incorrect guess (i.e., the final aligned event) was the most blame in the all-late (47%) versus the one-late (76%) sequence ($\chi^2(1) = 9.16$, $p < .002$). However, when there was a game delay, the influence of sequence on responsibility judgments for the final player was muted and even slightly reversed (all-late = 60% vs. one-late = 48%; $\chi^2(1) = 1.44$, $p < .08$). We contend that the unexpected game delay disrupted perceived momentum and, necessarily, mitigated the influence of event sequence on responsibility judgments.

Experiment 5: Evidence of Perceived Momentum via Expectations of Success versus Failure

We have argued that perceiving momentum in a sequence of events leads to the inference that success is more likely after a series of repeated successes and vice versa. Consequently a failure after multiple successes is more unexpected and assigned greater responsibility for the outcome (Gerstenberg, Halpern, & Tenenbaum, 2015). While the results of the preceding experiments have supported this contention indirectly, Experiment 5 explicitly tested whether people are biased towards expecting a success (failure) to follow a series of successes (failures).

Method. One hundred sixty-nine paid AMT participants participated in Experiment 5, which used the same sequences (one-late vs. all-late) and team outcome (loss) as Experiment 1. However, although participants were told of the team's loss, they were not given all of the events' results. Instead, participants (randomly assigned to either the one-late or all-late sequence) were shown the results of eight events, while the results of the other two events were hidden. See Table 1 ("I" = incorrect, "C" = correct; bolded events = aligned events).

After seeing the table of event results, participants were told that the team had lost by exactly one incorrect guess. Thus, one of the hidden results was necessarily an incorrect guess and the other a correct guess. The primary dependent variable in this experiment asked participants which of the two hidden results they believed was the incorrect guess (participants had to choose one or the other).

Results. We analyzed the percentage of participants choosing event 10 as the incorrect response across sequences (one-late vs. all-late). Since perceived momentum would lead one to believe a correct guess is more likely after a series or correct answers, we expected and found that a smaller proportion of participants chose event 10 as being the incorrect response in the one-late (40%) versus the all-late (57%) sequence ($\chi^2(1) = 4.93$, $p < .03$). Thus, it would appear that people perceive momentum in contexts where sequential events are objectively independent.

Table 1: Sequence Stimuli for Experiment 5

	Event Results	
Event (Guess)	One-Late	All-Late
1	**I_1**	C_1
2	**I_2**	C_2
3	**I_3**	C_3
4	?	?
5	C_1	C_4
6	C_2	C_5
7	C_3	**I_1**
8	C_4	**I_2**
9	C_5	**I_3**
10	?	?

General Discussion

Responsibility and causality judgments have received a lot of attention within the cognitive science field (e.g., Halpern & Pearl, 2005; Miller & Gunasegaram, 1990). The results presented here add to that literature. Two main findings expand the current understanding of causal responsibility judgments. First, people often perceive momentum where it cannot exist. We show this in a context where sequential events are explicitly and objectively independent. Second, since perceived momentum implies that it should be more likely for an event to stay the course (i.e., have the same result as a preceding series of repeated results), aligned events whose results deviate from the previous course are more unexpected and, therefore, tend to be attributed more responsibility for the overall outcome.

The role of perceived momentum in responsibility judgments was empirically supported in three ways: (1) via mediation analysis (Experiment 3), (2) via moderation of perceived momentum (Experiment 4), and (3) via measured expectations of success/failure as a function of the results of the series of events preceding the target event (Experiment 5).

These results add an important level of nuance to responsibility judgments not captured by the counterfactual

simulation model (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015) or models relying on perceived pivotality (Chockler and Halpern 2004). The aligned events in the sequences examined in these studies were equally pivotal to the outcome and were all whether- and how-causes that were equally (in)sufficient and (not) robust causes. Yet, the degree of responsibility each was assigned varied as a function of where it appeared in the overall sequence of events.

This work also extends that of Gerstenberg, Halpern, & Tenenbaum (2015) by demonstrating that the unexpectedness of a given event can be influenced by an implicit, inference-based cue: perceived momentum.

In future work, we will more explicitly examine counterfactual reasoning in the contexts examined here. Specifically, while most factors associated with counterfactual reasoning (e.g., pivotality) were held constant in the above studies, it is possible that an unexpected result is one that is associated with more plausible counterfactuals, which should increase the perceived responsibility of that event (consistent with our findings). Additionally, sequences other than those used in the current set of studies will be examined to further understand the role of perceived momentum in responsibility judgments.

Acknowledgments

The authors would like to thank the participants in the CB Group at Georgia Tech University for their invaluable feedback on this project.

References

Bartling, B., U. Fischbacher, &. S. Schudy (2015). Pivotality and responsibility attribution in sequential voting. *Journal of Public Economics*, *128*, 133-139.

Chockler, H., & J. Y. Halpern (2004). Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, *22*, 93-115.

Gerstenberg, T., J. Y. Halpern, & J. B. Tenenbaum (2015). Responsibility judgments in voting scenarios. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 788-793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., N. D. Goodman, D. A. Lagnado, & J. B. Tenenbaum (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 782-787). Austin, TX: Cognitive Science Society.

Gilovich, T., R. Vallone, & A. Tversky (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17* (3), 295-314.

Green. B. S., J. Zwiebel (2013), "The Hot Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Baseball. *working paper*.

Halpern, J. Y., & J. Pearl (2005). Causes and Explanations: A Structural-Model Approach. Part 1: Causes. *British Journal for the Philosophy of Science, 56* (4), 843-887.

Lagnado, D. A., T. Gerstenberg, & R. Zultan (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37,* 1036-1073.

Miller, D. T., & S. Gunasegaram (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology, 59*(6), 1111-1118.

Miller, J. B., & A. Sanjurjo (2015). Surprised by the Gambler's and Hot Hand Fallacies? A Truth in the Law of Small Numbers. *working paper*.

Zultan, R., T. Gerstenberg, & D. A. Lagnado (2012). Finding fault; Causality and counterfactuals in group attributions. *Cognition, 125* (3), 429-440.

A computational investigation of the Sapir-Whorf hypothesis: The case of spatial relations

Christine Tseng[1] **(ctseng@caltech.edu), Alexandra Carstensen**[2] **(abc@berkeley.edu), Terry Regier**[3] **(terry.regier@berkeley.edu), Yang Xu**[3] **(yang_xu_ch@berkeley.edu)**

[1] Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125 USA
[2] Department of Psychology, University of California, Berkeley, CA 94720 USA
[3] Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Abstract

Investigations of the Sapir-Whorf hypothesis often ask whether there is a difference in the non-linguistic behavior of speakers of two languages, generally without modeling the underlying process. Such an approach leaves underexplored the relative contributions of language and universal aspects of cognition, and how those contributions differ across languages. We explore the naming and non-linguistic pile-sorting of spatial scenes across speakers of five languages via a computational model grounded in an influential proposal: that language will affect cognition when non-linguistic information is uncertain. We report two findings. First, native language plays a small but significant role in predicting spatial similarity judgments across languages, consistent with earlier findings. Second, the size of the native-language role varies systematically, such that finer-grained semantic systems appear to shape similarity judgments more than coarser-grained systems do. These findings capture the tradeoff between language-specific and universal forces in cognition, and how that tradeoff varies across languages.

Keywords: Linguistic relativity; Sapir-Whorf hypothesis; semantic universals; name strategy; categorization; spatial relations; computational models.

Introduction

Languages partition human experience into semantic categories in different ways. For example, the Mandarin Chinese spatial term *shang4* denotes a set of spatial relations that is roughly equivalent to those described by English *on* and *above* combined. Do such differences affect how speakers of different languages apprehend and think about the world?

The Sapir-Whorf hypothesis (Sapir, 1929; Whorf, 1956) is commonly framed in terms of this question. When the question is posed this way, in simple yes-or-no terms, it invites an equally simple answer: that language either does or does not influence cognition. However, empirical studies have provided conflicting answers to this question across a variety of semantic domains (Roberson, Davies, & Davidoff, 2000; Gilbert, Regier, Kay, & Ivry, 2006; Brown, Lindsey, & Guckes, 2011; Malt, Sloman, Gennari, Shi, & Wang, 1999; Munnich, Landau, & Dosher, 2001; Kranjec, Lupyan, & Chatterjee, 2014; Majid, Bowerman, Kita, Haun, & Levinson, 2004). This inconsistent pattern of results suggests that a slightly more complex formulation of the hypothesis may be warranted. Concretely, rather than asking whether language does or does not shape cognition, it may be useful to ask under which circumstances it does, to what extent, and whether that extent itself varies across languages in a systematic way, because of general principles. We pursue these questions here.

Our empirical focus is the semantic domain of topological spatial relations. Earlier investigations have revealed wide yet constrained cross-language variation in spatial semantic categories (Levinson & Meira, 2003), and have also revealed that non-linguistic cognition about such spatial relations reflects both universal forces and some influence of native language (Khetarpal, Majid, Malt, Sloman, & Regier, 2010; Carstensen et al., under revision). However, while documenting the interesting interplay of universal and linguistic forces in apparently non-linguistic spatial cognition, such earlier studies did not explore that interplay using a computational model, and did not illuminate under which circumstances language shapes cognition.

In addressing these open questions, our theoretical starting point is an influential proposal from the literature. In a classic study of language and color cognition, Kay and Kempton (1984) proposed what they called the *name strategy*: that language will affect cognition when non-linguistic aspects of cognition are ambiguous, uncertain, or otherwise ineffective. Their empirical findings were consistent with this idea. We instantiate the name strategy in a computational model, apply the model to data from the spatial domain, and explore the above open questions in terms of the theoretical framework this model provides.

In what follows, we first describe the data we consider, which are drawn from five languages: Dutch, English, Chichewa, Mandarin, and Maihiki. We then present our computational model, and show how it instantiates the name strategy. We then analyze the data through the lens of the model. To preview our results, we find: (1) that across all five languages, native language plays a small but significant role in predicting spatial similarity judgments, consistent with earlier findings, and (2) that the size of the native-language role varies systematically across languages, such that finer-grained semantic systems appear to shape similarity judgments more than coarser-grained systems do. We argue that these findings contribute to the debate over the Sapir-Whorf hypothesis by grounding the tradeoff between universal and language-specific forces in a computational model based on an independently proposed principle: the name strategy.

Data

To investigate the relation of spatial language and cognition, we compare linguistic and nonlinguistic categorization of spatial scenes across five languages: Dutch, English, Chichewa, Mandarin, and Maihiki. Maihiki is an underdocumented language of Peruvian Amazonia, presently being investigated by Lev Michael and colleagues at Berkeley. The spatial scenes were those of the Topological Relations Picture Series (TRPS) (Bowerman & Pederson, 1992). This stimulus set contains 71 spatial relational scenes, each of which depicts a spatial relationship between a figure object and a ground object. Figure 1 shows a small subset of the TRPS scenes, and the semantic categories in which these scenes fall for some of the languages we consider. It can be seen that there is considerable variation in spatial categories across these languages.

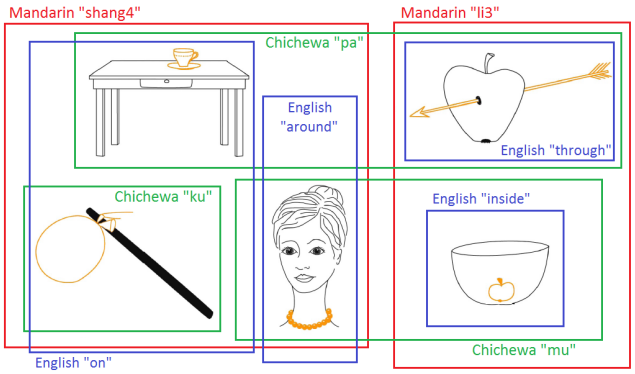


Figure 1: Cross-linguistic naming variation in the spatial domain. Each scene depicts a spatial relation between a *figure* object (in orange) and a *ground* object (in black). The scenes are grouped differently by different languages.

The data on which we rely were collected previously by Khetarpal et al. (2009, 2010, 2013), Carstensen (2011), and Carstensen et al. (under revision). Below we briefly describe the data collection procedure used in those prior studies.

Participants

A total of 47 native English speakers (24 from Khetarpal et al., 2010; 23 from Carstensen et al., under revision), 24 native Dutch speakers (Khetarpal et al., 2009), 38 native Chichewa speakers (Carstensen, 2011), 7 native Maihiki speakers (Khetarpal et al., 2013), and 17 native speakers of Mandarin Chinese (Carstensen et al., under revision) pile-sorted and then named spatial scenes. All English, Dutch, and Chichewa speakers were tested in their native languages and home countries (the United States, the Netherlands, and Malawi, respectively). Mandarin Chinese speakers were recruited on the UC Berkeley campus and tested in their native language. Speakers of Maihiki were tested in their home country of Peru, but as this is an endangered language with very few speakers, they were tested in Spanish, in which all Maihiki participants were also fluent. All participants first took part in a nonlinguistic pile-sorting task and subsequently completed a naming task; English speakers' pile-sorting data is from Khetarpal et al. (2010) but their naming data is from Carstensen et al. (under revision) for which naming instructions are more closely aligned with those of the other studies.

Pile-sorting task

Participants sorted the 71 scenes in the TRPS into piles based on the similarity of the spatial relations depicted in the scenes. Each scene showed an orange figure object positioned relative to a black ground object and participants were instructed to group the scenes into piles based on the similarity of these spatial relations, such that the relation was similar for all cards in a given pile. Participants were informed that they could make as few or as many piles as they chose, rearrange their piles as they felt necessary, and could take as much time as they wanted.

Naming task

After completing the sorting task, the same participants were asked to name the spatial relation depicted on each card. For languages other than Maihiki, labels picking out the figure and ground objects were supplied in the participant's native language and the participant filled in a blank to complete a sentence specifying the figure object's location relative to the ground object. For example, for the scene depicting a cup on a table, English-speaking participants were presented with the partial sentence "The cup is (blank) the table", and were asked to fill in the blank. Maihiki speakers were asked to produce full sentences, supplying names for the figure and ground objects and describing the spatial relation between them; verbal clarification of the scenes was given in Spanish when necessary. In keeping with earlier work, the labels produced in the naming task were sanitized to collapse over responses that differed in components without spatial meaning (e.g. variations in verb tense).

Treatment of the data

We aggregated these data into separate language-specific and universal components, for use in our computational analyses below. For each language l, we constructed a 71×71 *co-naming* matrix L^l, such that entry L^l_{ij} of that matrix contained the proportion of speakers of language l who supplied the same spatial term for scenes i and j. For instance, if 12 out of 16 speakers of language l supplied the same spatial term in l for scenes i and j, then $L^l_{ij} = 12/16 = 0.75$. Thus the L^l matrix summarizes which scenes tended to receive the same name vs. different names in language l. We analogously constructed, for each language l, a 71×71 *co-sorting* matrix S^l, such that entry S^l_{ij} of that matrix contains the proportion of speakers of language l who sorted scenes i and j into the same pile. Finally, we approximated a universal similarity space, U, by taking the average over the S^l matrices across languages l. Earlier studies (Khetarpal et al., 2010; Carstensen et al., under revision) found that spatial sorting patterns were

broadly similar across languages,[1] although they did reflect the sorter's native language to some extent. Thus, U is an attempt to retain what is common and discard what is different about pile-sorting behavior across languages.

Computational formulation

The core principle behind our analyses is the name strategy of Kay and Kempton (1984). Figure 2 illustrates this principle in the context of a simple pile-sorting scenario. Suppose there are three stimuli (in our case three spatial scenes) here labeled 1, 2, and 3, and suppose that the task is to sort stimulus 3 into one of two existing piles, which presently contain stimuli 1 and 2 respectively. Assume further that the three stimuli are equally distant from each other in a universal similarity space, so that it is entirely ambiguous on that basis which pile stimulus 3 should be sorted into. Finally, assume that the speaker's native language partitions these stimuli into two categories $N1$ and $N2$, where stimuli 1 and 3 are co-named under $N1$, and stimulus 2 is named under $N2$. The name strategy holds that in such cases, where universal non-linguistic structure yields ambiguity or is otherwise ineffective, linguistic category structure may provide additional information to resolve the issue. In this case, stimuli 3 and 1 are co-named, which should encourage stimulus 3 to be sorted into pile 1, tipping the balance in that direction. The bars in the bottom panel of Figure 2 contrast the choice probabilities over the two piles for a hypothetical model that relies only on a universal similarity space, with those for a model that relies only on the category structure of the language.

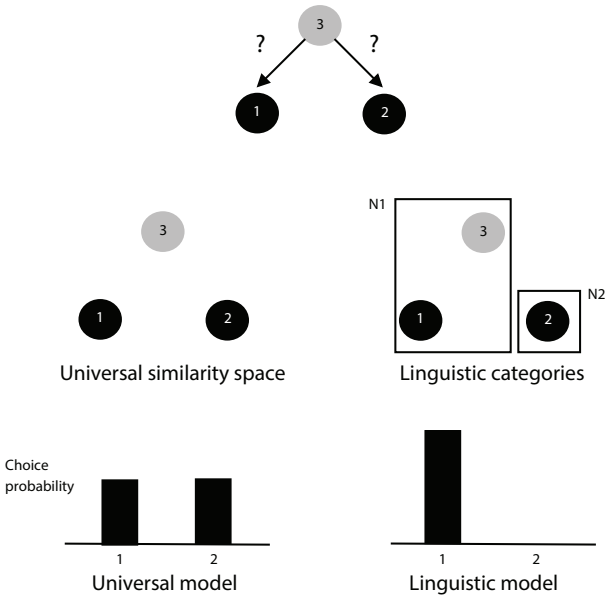


Figure 2: Illustration of the name strategy.

To capture the idea of the name strategy, we formulate models that can be used to predict an individual's pile-sorting behavior. These predictions are based on universal similarity structure U, and on language-specific naming information L^l, both as specified above, where l is the native language of the individual in question. We instantiate the name strategy through a *residual predictive analysis*: we first note how much of an individual's sorting behavior can be predicted by U, and then determine how much of that individual's as-yet-unexplained (residual) sorting behavior can be predicted by L^l, beyond what U can predict. This two-step procedure requires that we specify a *universal model* based on U that we employ in the first step, and a *language-specific model* based on L^l for the residual analysis, i.e. the items left unexplained by U. This procedure provides a way to quantify the relative contributions of universal and language-specific forces, and to assess the degree to which those relative contributions may vary across languages.

For each individual, we know the number of piles that individual produced, and which scenes were sorted into each pile. We sought to recapitulate the sorting process of that individual. Concretely, for each query scene, we sought to predict which pile that scene was sorted into, using a leave-one-out procedure. That is, for each scene i, we held out pile membership for i and sought to predict pile membership for i, based on pile membership for all other scenes $j \neq i$.[2] We cast this prediction in probabilistic terms, and predicted that this individual would place spatial scene i in that pile c (for non-linguistic category) that yielded the highest posterior probability $p(c|i)$:

$$p(c|i) \propto p(i|c)p(c) \propto f(i|c) \tag{1}$$

Here, we assume the individual has no preference for any pile to begin with, and we therefore place a uniform prior on c. In determining the likelihood $f(i|c)$, we considered three possible strategies on which individuals might rely in sorting scenes into piles, shown in Equation 2 below. These are: (a) a fixed clustering strategy: sort by average similarity between a query scene i and all existing scenes j in a given pile c; (b) a fixed nearest-neighbor or chaining strategy: sort based on maximum similarity between a query scene i and any of the existing scenes j in a given pile c; and (c) a hybrid strategy that varies on a trial-by-trial basis: choose between the clustering and chaining strategies according to which yields higher likelihood. We formalize these strategies by specifying the likelihood function as follows:

$$f(i|c) = \begin{cases} \frac{1}{|c|}\sum_{j\in c} sim(i,j), & clustering \\ \max_{j\in c} sim(i,j), & chaining \\ \max(f(i|c)_{clustering}, f(i|c)_{chaining}), & hybrid \end{cases} \tag{2}$$

[1] We verified the universal tendencies in pile sorting by correlating the S matrices (upper triangular parts due to symmetry) between each pair of languages. We confirmed that pile sorting is largely similar across speakers of different languages reflected in mean Pearson's $r = 0.98$ ($SD = 0.01$) among the pairwise correlations.

[2] Ideally, we would like to predict scene-pile assignments in the sequence in which they occurred during the experiment. However, only the end-state of the pile-sort was recorded, not the sequence that led to it, so we used the leave-one-out procedure which is unaffected by the sorting sequence.

For each scene assignment, we chose the strategy that yielded the highest predictive accuracy. To distinguish between the universal and language-specific models, we took $sim(i,j) = U_{ij}$ for the universal model, and $sim(i,j) = L^l_{ij}$ for the linguistic model, where l is the individual's native language and U and L^l are as defined above.

To model a given individual's behavior, we first conducted a predictive analysis using the universal model, and noted which of that individual's scene assignments were predicted correctly and which incorrectly. We then conducted a residual predictive analysis using the language-specific model, on those scenes that were incorrectly predicted by the universal model, and again noted which scenes were predicted correctly and which incorrectly.

Analyses and results

We conducted such analyses for each speaker of each language. We then examined the results of those analyses with a view to answering the open questions posed at the beginning of this paper. We did so in three sets of followup analyses, which we present below.

Relative contributions of universal and language-specific forces

One open question is the magnitude of the relative contribution of universal and language-specific forces to allegedly non-linguistic tasks such as pile-sorting, when assessed using the method outlined above.

Figure 3 summarizes the results of the residual predictive analyses just described. For each language, the black bar shows the accuracy of the universal model in predicting scene-pile assignments, averaged across speakers of that language. The white bar stacked on top of it shows the accuracy of the language-specific model in predicting residual scene-pile assignments, i.e. those that the universal model failed to predict, again averaged across speakers of that language.[3] To establish a baseline in this language-specific residual predictive task, we considered chance predictions from a model that chooses scene-pile assignments randomly from among the available piles. In this case, the chance-level accuracy for each individual is $\frac{1}{k}$ where k is the number of piles that individual generated during pile sorting, and the dashed horizontal line for each language shows the chance level of residual predictive accuracy for that language, averaged across speakers of the language.

Overall, the universal model accounts for a substantial proportion of the pile-sort data for each language, suggesting strong universal tendencies in people's similarity judgments about spatial relations. At the same time, for each language, the language-specific model predicts residual scene-pile assignments at rates above chance. To assess whether this effect of language is significant at the level of individual speakers, we examined the proportion of individuals

[3]Similar results were also obtained when, for each speaker, we left pile-sort data from that speaker's native language out of the universal similarity matrix U on which we base predictions.

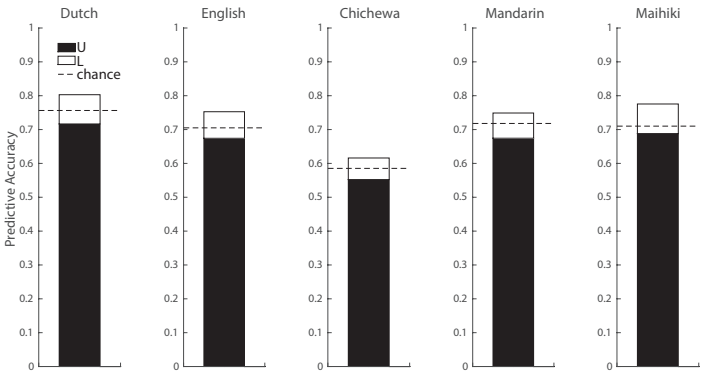


Figure 3: Summary of results of residual predictive analyses.

for whom the language-specific model exhibits above-chance residual predictive accuracy. We found that this proportion is high across speakers of all the languages we considered (Chichewa: 30/38; Dutch: 21/24; English: 20/23; Maihiki: 6/7; Mandarin: 15/17; Bonferroni-corrected $p < 0.05$ for each language except for Maihiki (uncorrected $p = 0.05$) under binomial tests assuming 0.5 probability of success per speaker). Taken together, these findings provide evidence in support of the Sapir-Whorf hypothesis in this domain, and contextualize it relative to what can be accounted for by universal forces alone. It should be emphasized that this is a rather conservative test for an effect of language, in that any scene-pile assignment that would be correctly predicted by both the universal model and the language-specific model will be credited here to the universal model, as it was run first, in keeping with our instantiation of the name strategy. Thus, it may be safest to think of these residual predictive accuracies as providing a lower bound on the size of the language-specific contribution.

Native language compared with other languages

Support for the Sapir-Whorf hypothesis would be strengthened if we had evidence that an individual's native language outperformed other languages in residual predictive accuracy. For example, is residual Maihiki pile-sorting better predicted by Maihki naming than it is by Chichewa naming, or Mandarin, or some other language? This pattern would be expected if the sorter's native language is in fact being called upon during the putatively non-linguistic sorting process. We turn next to consider this question.

To test this, we re-ran the residual predictive analyses described above, but for each individual, instead of using that individual's native-language naming information (in the form of L^l for native language l), we used naming information from each other language (i.e. L^k for each language $k \neq l$). For example, to predict residual pile-sorting data for Dutch speakers, we used four different linguistic models based on naming from each of the four alternative languages other than Dutch, while keeping all other procedures unchanged. For this and remaining analyses, we focused on individuals that exhibited

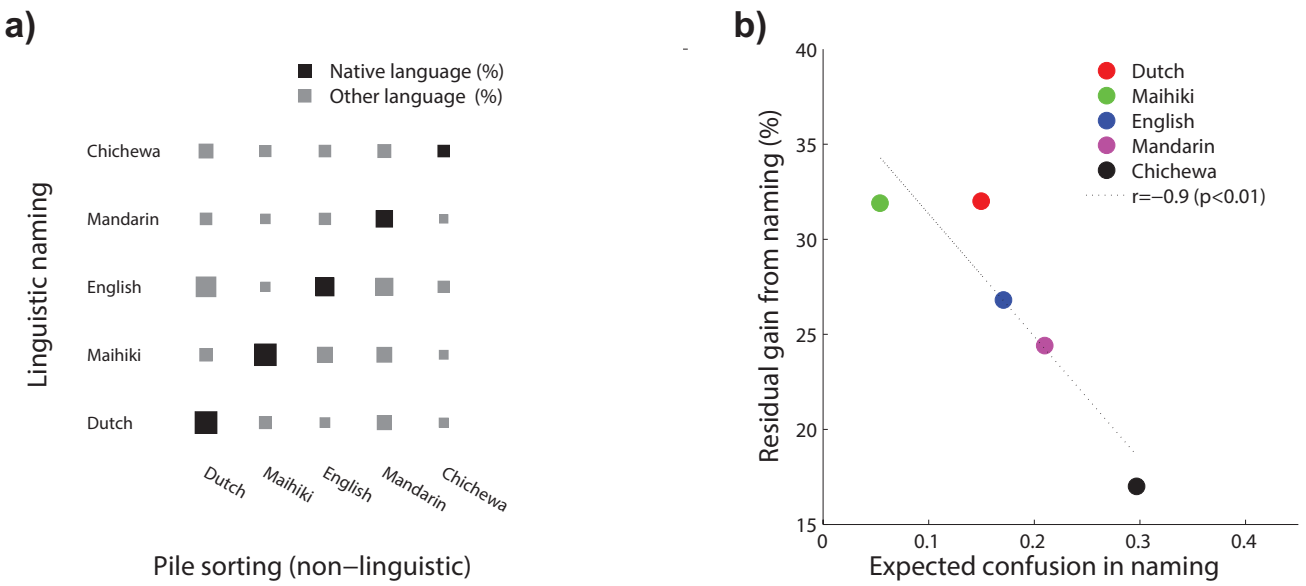


Figure 4: Summary of results of cross-language comparisons. (a) Within-language and cross-language predictive accuracies on residual scene-pair assignments. Diagonal elements (black squares) reflect predictive accuracies from native languages. Off-diagonal elements (gray squares) reflect predictive accuracies from other languages. Square size is proportional to percentage gain in predictive accuracy in the residual analysis. (b) Relationship between naming confusion (semantic coarse-grainedness; horizontal axis) and percentage gain in residual prediction from each language (vertical axis).

above-chance accuracies in residual prediction.

Figure 4(a) presents the results of this cross-language predictive analysis. The plot summarizes how well different linguistic naming matrices L (by row) predict residual scene-pile assignments in pile-sorting by speakers of different languages (by column). Square size is proportional to percentage gain in predictive accuracy in the residual analysis. Diagonal elements (in black) represent performance given native-language information, and the four off-diagonal elements (in gray) within each column represent performance given non-native-language information, on the same pile-sort data. In almost all cases, native-language models outperformed the non-native-language models in predicting residual pile-sorting data by speakers of that native language (19 out of 20 pairwise comparisons; the exception is that English naming predicts sorting data from Mandarin speakers slightly better than Mandarin naming itself). This finding supports the suggestion that native language was recruited in pile-sorting. We note also that the percentage gain in the residual prediction differs across languages. In Figure 4(a), the diagonal elements are sorted by accuracy, and it can be seen that Dutch predicts its speakers' residual pile-sort data the best, and Chichewa predicts its speakers' residual pile-sort data the worst. In our final analysis, we asked whether this variation in cross-language predictive accuracy is systematically linked to the nature of the semantic systems involved.

Semantic grain and linguistic relativity

Our final analysis builds on earlier explorations of *semantic grain* in sorting and naming (Khetarpal et al., 2010; Carstensen et al., under revision). A natural possibility is that fine-grained semantic systems may have a greater effect on non-linguistic spatial similarity judgments (reflected in higher residual predictive accuracy) than coarse-grained systems do. The rationale is that a fine-grained semantic system offers more opportunities to resolve ambiguous cases. This proposal follows from the name strategy, where Kay and Kempton (1984) found that the degree of linguistic effect on how speakers resolve ambiguity in color judgements (e.g. distinguishing colors near the blue-green boundary) depends on the fine-grainedness of color naming systems. Similarly, in the case of pile-sorting of spatial scenes, we expect scene pairs that cannot be resolved linguistically by a coarse-grained system because they fall in the same category in that system, can be resolved linguistically by a fine-grained system because they fall in different categories in that system.

To test this prediction, we examined the relationship between native-language residual predictive accuracy (from the previous analyses) and the semantic grain of each language. We assessed semantic grain for each language l by measuring the expected amount of *naming confusion*, or co-naming, in l's co-naming matrix L^l, averaging together the co-naming proportions across all unique pairs of scenes i, j:

$$nc(l) = \text{average}_{(i,j)} L^l_{i,j} \tag{3}$$

Here, $nc(l)$ measures the extent to which different stimuli tend to receive the same name and thus be linguistically indistinguishable in language l. Coarse semantic grain corresponds to a high value for $nc(l)$ (because different scenes i, j will often receive the same names in a coarse-grained system,

and thus have high co-naming values $L^l_{i,j}$), and fine semantic grain corresponds to a low value for $nc(l)$.

Figure 4(b) shows that there is a strong negative correlation (Pearson's $r = -0.9$; $p < 0.01$ from a permutation test with 10,000 samples) between expected naming confusion (coarse-grainedness) for a language, and residual gain in predictive accuracy (size of the Sapir-Whorf effect) for that language.[4] Dutch and Maihiki–the two most fine-grained systems in our data–yield the highest predictive accuracies in the residual analysis. In contrast, Chichewa–which has a comparatively coarse-grained spatial semantic system–yields the lowest residual predictive accuracy. We have considered only five languages here, and future work can usefully examine the robustness of this finding with a larger set of languages. Nevertheless, these findings provide initial support for the prediction that semantically finer-grained languages will tend to have a greater effect on non-linguistic judgments than coarser-grained languages will.

Conclusion

We have presented two main findings: (1) Language appears to play a small but significant role in shaping spatial similarity judgments, in line with earlier studies, and (2) the extent of this linguistic effect varies as a function of the semantic grain of one's native language. We have arrived at these findings by pursuing a proposal from the literature, Kay and Kempton's (1984) name strategy, and by instantiating that proposal in a computational analysis. Similar ideas appear elsewhere in the literature (e.g. Vong et al., 2015), and we hope that our work will encourage further formal analyses of the link between cross-language semantic structures and cognition.

Acknowledgments

We thank Lev Michael and Grace Neveu for access to the Maihiki data, and Asifa Majid for access to the Dutch data. This project was funded by NSF award SBE-1041707, the Spatial Intelligence and Learning Center (SILC), and NSF Graduate Research Fellowship grant DGE-1106400 (to AC).

References

Bowerman, M., & Pederson, E. (1992). Topological relations picture series. In *Space stimuli kit 1.2* (p. 51). Nijmegen: Max Planck Institute for Psycholinguistics.

Brown, A., Lindsey, D., & Guckes, K. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, *11*.

Carstensen, A. (2011). *Universals and variation in spatial language and cognition: Evidence from Chichewa.* Undergraduate thesis, University of California, Berkeley.

Carstensen, A., Khetarpal, N., Majid, A., Malt, B., Sloman, S., & Regier, T. (under revision). Cross-language universals and variation in cognition: The cases of space and artifacts.

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, *103*, 489-494.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*, 65-79.

Khetarpal, N., Majid, A., Malt, B., Sloman, S., & Regier, T. (2010). Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Kranjec, A., Lupyan, G., & Chatterjee, A. (2014). Categorical biases in perceiving spatial relations. *PLoS ONE*, *9*, e98604.

Levinson, S. C., & Meira, S. (2003). Natural concepts in the spatial topological domain–adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*, 485-516.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108-114.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Munnich, E., Landau, B., & Dosher, B. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, *81*, 171-207.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369-398.

Sapir, E. (1929). The status of linguistics as a science. *Language*, 207–214.

Vong, W. K., Navarro, D. J., & Perfors, A. (2015). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, 1–9.

Whorf, B. (1956). The relation of habitual thought and behavior to language. In J. Carroll (Ed.), *Language, thought, and reality* (p. 134-159). Cambridge, Massachusetts: The MIT Press.

[4] We replicated this finding using a discrete measure for semantic grain based on height following Khetarpal et al., (2010), where we found a strong negative correlation between mean coarseness in naming for a language and residual gain (Pearson's $r = -0.92$).

Effects of Auditory Input on a Spatial Serial Response Time Task

Christopher W. Robinson (robinson.777@osu.edu)
Department of Psychology, The Ohio State University at Newark
1179 University Dr, Newark, OH 43056, USA

Jessica L. Parker (parker.1026@osu.edu)
Department of Psychology, The Ohio State University at Newark
1179 University Dr, Newark, OH 43056, USA

Abstract

The current study examined how relevant and irrelevant auditory stimuli affect the speed of responding to structured visual sequences. Participants were presented with a dot that appeared in different locations on a touch screen monitor and they were instructed to quickly touch the dot. Response times sped up over time, suggesting that participants learned the visual sequences. Response times in Experiment 1 were slower when the dot was paired with random sounds, suggesting that irrelevant sounds slowed down visual processing/responding. Dots in Experiment 2 were paired with correlated sounds (both auditory and visual information provided location information). While the redundant intersensory information did not speed up response times, it did partially attenuate auditory interference. These findings have implications on tasks that require processing of simultaneously presented auditory and visual information and provide evidence of auditory interference and possibly dominance on a task that typically favors the visual modality.

Keywords: Cross-modal processing; Sensory Dominance; Attention.

Introduction

Many important tasks rely on detecting statistical regularities in the environment. For example, speech segmentation, word learning, and category learning require a person to abstract transitional probabilities of speech sounds, detect that some words co-occur with specific objects, and learn that some features are necessary or probabilistically relevant for a given category, respectively. Moreover, much of this learning appears to happen automatically with young infants quickly learning artificial categories (Younger & Cohen, 1983) and learning the transitional probabilities of speech sounds after two minutes of exposure to an artificial language (Saffran, Aslin, & Newport, 1996). Much of this learning can also happen without attention, with children learning probabilities of auditory sequences even when the primary task was visual in nature and they were not instructed to pay attention to the auditory sequences (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997, but see Toro, Sinnett, & Soto-Faraco, 2005).

In addition to perceiving and abstracting statistical regularities, there are also occasions when the structure involves motor memory. For example, playing musical instruments, typing, swinging a golf club, driving a manual transmission, etc., require coordinated movements, which eventually become automated and consume few attentional resources. Many studies have studied perceptual-motor learning in adults. For example, research using a Serial Response Time Task (SRTT) often consists of presenting visual information to spatially distinct locations, and participants are instructed to quickly respond to this information (e.g., Dennis, Howard, & Howard, 2006; Nissen & Bullemer, 1987; Song, Howard, & Howard, 2008). Unbeknownst to participants, the visual sequences are often structured and follow a statistical pattern, but see Vadillo, Konstantinidis, and Shanks (2016) for a recent review of implicit vs. explicit learning. While many studies have used variants of a SRTT, very little is known regarding how information from other sensory modalities affects learning of these structured visual sequences. Therefore, the primary goal of the current study is to examine how relevant and irrelevant auditory information affect learning and responding to visually structured sequences.

Over the last 40 years, there is a considerable amount of research showing that when simultaneously presented with auditory and visual information, the visual modality dominates the auditory modality (Colavita, 1974; Colavita, Tomko, & Weisberg, 1976; Colavita & Weisberg, 1979; Egeth & Sager, 1977). For example, in a classical Colavita task, participants are instructed to quickly respond to auditory and visual information by quickly pressing one button when they hear a sound and by pressing a different button when they see an image/flash (Colavita, 1974). On some of the trials, auditory and visual information are presented at the same time. Participants often miss these cross-modal trials by only pressing the visual button; therefore, it was concluded that the visual modality dominated the auditory modality. The Colavita visual dominance effect and variations of this task consistently point to visual dominance, with stimulus and attentional manipulations often weakening but not reversing the effect (see Sinnett, Spence, & Soto-Faraco, 2007; Spence, Parise, & Chen, 2012 for reviews). While numerous sensory, attentional, and motor mechanisms have been put forward to account for visual dominance, underlying mechanisms are poorly understood.

Recent findings provide some support for auditory dominance; however, these studies often test infants and children rather than adults (Lewkowicz, 1988a; 1988b; Robinson & Sloutsky, 2004; Sloutsky & Napoliltano, 2003; Sloutsky & Robinson, 2008). For example, infants and children are often better at discriminating pictures when presented in silence than when the same pictures are paired

with sounds or words (Lewkowicz, 1988a; 1988b; Robinson & Sloutsky, 2004; 2010a; Sloutsky & Robinson, 2008). At the same time, the pictures appear to have no negative effect on auditory processing; thus, multisensory presentation attenuated visual but not auditory processing. To account for this finding, Robinson and Sloutsky (2010) have posited that sensory modalities may share the same pool of attentional resources and compete for attention. Furthermore, due to the transient and dynamic nature of auditory input, it may be adaptive to first allocate attention to auditory input before it disappears. Increased attention automatically deployed to the auditory modality may come with a cost - attenuated or delayed visual processing.

While this competition for attention explanation (Robinson & Sloutsky, 2010) may account for some of the developmental findings, there are only a few studies pointing to auditory interference in adults, and these studies do not use a traditional Colavita paradigm that require participants to quickly respond to multisensory information. For example, in sound induced flash illusion, participants are presented with a series of beeps and flashes (Shams, Kamitani, & Shimojo, 2000; 2002). There are often no response time constraints and participants are not asked to attend to (or report on) the auditory information. In these tasks, the auditory information influences the number of flashes reported. For example, if participants see two flashes but hear three beeps, they might report seeing three flashes.

Cross-modal presentation can also affect visual statistical learning. In a cross-modal statistical learning task, participants were presented with streams of auditory, visual, or cross-modal sequences (auditory and visual sequences were presented at the same time), and participants were either tested on the auditory or visual sequences (Robinson & Sloutsky, 2013). Increasing the task demands by randomizing one of the streams attenuated visual but not auditory statistical learning. In other words, participants learned the visual sequences when presented in silence or when paired with correlated sounds, but randomizing the auditory stream attenuated visual statistical learning. Randomizing the visual stream had no negative effect on auditory statistical learning.

One possible explanation that may account for the auditory interference/dominance effects in the sound induced flash illusion and cross-modal statistical learning tasks (Robinson & Sloutsky, 2013; Shams, Kamitani, & Shimojo, 2000; 2002) is that both tasks rely almost exclusively on temporal processing. According to the Modality Appropriateness Hypothesis (Welch & Warren, 1980), the modality that is most suitable for a given task will dominate. Given that the visual system is better at processing location information (Alias & Burr, 2004) and the auditory modality is better at processing temporal information (Burr, Banks, & Morrone, 2009), it is not surprising to see the auditory modality dominate in temporal tasks such as statistical learning (Conway & Christiansen, 2005).

In the current study, we employed a spatial SRTT to determine if auditory stimuli also affect the speed of responding to visual input on a task better suited for the visual modality. Participants were presented with a sequence of dots that appeared in different locations on a touch screen monitor. Participants either heard a random sequence of sounds (Experiment 1), a correlated sequence of sounds (Experiment 2), or the visual sequences were presented in silence (Experiments 1 and 2). According to the Modality Appropriateness Hypothesis (Welch & Warren, 1980), the task should be well suited for the visual modality; thus, the auditory input should have no negative effect on visual response times. However, if auditory stimuli automatically engage attention and pull attention away from the visual modality (c.f., Robinson & Sloutsky, 2010), then it is possible that auditory stimuli will slow down visual responses and/or slow down learning rate on a visual-spatial task.

Experiment 1

Method

Participants Twenty-nine undergraduate students from The Ohio State University-Newark (12 Females, $M = 18.26$ years) participated in Experiment 1, from which they gained research assignment credit for the Introduction to Psychology course. One participant was tested but not included in the analyses due to a reported hearing loss.

Apparatus The experiment was created using OpenSesame software and ran on a Dell Optiplex 9010 computer with an Intel Core i7 processor. The visual stimuli were shown on a 22" Planar PXL2230 1920 x 1080 touch screen monitor. Participants used the touch screen to respond to the visual stimuli. The auditory stimuli were presented via Kensington KMW33137 headphones at approximately 65-68 dB.

Materials and Design The visual stimulus was the default fixation stimulus generated in OpenSesame. The fixation stimulus was a filled white circle (dot) with an 8 pixel radius and a 2 pixel hole, and it was presented on a black background. The dot appeared at 12 different locations on the monitor, and each participant saw two sequences, as represented by the xy coordinates in Table 1. Sequence order (sequence 1 vs. sequence 2) was randomized for each participant with approximately half of the participants seeing sequence one first, and the other half seeing sequence two first.

Condition (silent vs. sound) was also manipulated within subjects. In the sound condition, each visual stimulus was paired with a tone and tones were presented at the following 12 different frequencies: 200 Hz, 400 Hz, 600 Hz, 800 Hz, 1200 Hz, 1200 Hz, 1400 Hz, 1600 Hz, 1800 Hz, 2000 Hz, 2200 Hz, 2400 Hz, and 2600 Hz. For approximately half of the participants, the tones were paired with sequence one, and for the remaining participants, tones were paired with sequence two.

Procedure The experiment consisted of a silent condition and a sound condition. In the silent condition, the visual sequences were presented in silence, and in the sound condition, visual sequences were paired with the tones. The

fixation dot appeared in a repeating pattern of 12 locations (see sequences in Table 1), and the same sequence repeated 20 times in each condition, giving a total of 240 trials per condition, and 480 trials total in the experiment. Participants were instructed to touch the dot on the touch screen monitor as quickly as possible. The dot stayed on the screen until the participant touched that location. In the sound condition, participants were told that they would hear a sound, but that they were to respond only to the visual stimuli as in the silent condition. The tones were randomly paired with visual sequences and the pairings switched on every trial. For example, on trial 1, a 200 Hz tone may have been presented when the dot appeared in location 1. On the next trial, location 1 may have been associated with a 1600 Hz tone, etc. The order was counterbalanced among the participants so that approximately half of the participants received the sound condition first, and the other half received the silent condition first. In addition, the visual stimulus pattern was also counterbalanced among participants so that half of the participants experienced the pattern in reverse order.

XY Coordinates of the Visual Stimulus		
Stimulus	Sequence 1	Sequence 2
1	(283, 116)	(1350, 524)
2	(456, 564)	(936, 703)
3	(708, 826)	(1436, 764)
4	(1436, 116)	(484, 118)
5	(1720, 516)	(34, 328)
6	(1233, 732)	(284, 764)
7	(284, 764)	(1233, 732)
8	(34, 328)	(1720, 516)
9	(484, 118)	(1436, 116)
10	(1436, 764)	(708, 826)
11	(936, 703)	(456, 564)
12	(1350, 524)	(283, 116)

Table 1. Above is the 12 location pattern of the visual stimuli in Experiment 1.

Results and Discussion

Reaction times were calculated for each stimulus and mean response times were averaged for each trial (12 stimuli per trial). See Figure 1 for mean response times and standard errors across the 20 trials. As can be seen in the figure, responses sped up over time, suggesting that some learning occurred, and response times were generally faster in the silent condition.

Log transformed response times were submitted to a 2 (silent vs. sound) x 4 (block 1, block 2, block 3, block 4) repeated measures ANOVA. A block was defined as five trials (e.g., block 1 = trials 1-5, block 2 = trials 6-10, etc.). The results showed a significant effect of condition, F (1, 28) = 4.65, p = 0.04, which shows that response times in the sound condition (M = 812.40 ms, SE = 20.26) were slower than the silent condition (M = 765.00 ms, SE = 16.93). There was also a significant effect of time, F (3, 84) = 32.71, p <.001, showing that reaction time sped up across the blocks. The means and (SEs) for blocks 1 – 4 were 832.50 ms (15.15), 802.60 ms (17.69), 764.50 ms (16.84), and 754.70 ms (15.87), respectively. Paired samples t-tests, using log transformed response times, were conducted between the blocks, showing block 1 was significantly slower than block 2, t (28) = 4.26, p < 0.001, and block 2 was significantly slower than block 3, t (28) = 3.53, p = 0.001. The difference between block 3 and block 4 did not reach significance. There was also a significant Condition x Block interaction, F (3, 84) = 5.65, p = 0.001. As can be seen in Figure 1, response times in the sound condition were significantly slower than the silent condition in block 1 (trials 1-5), t (28) = -2.91, p = .007, and marginally slower than silent in block 2 (trials 6-10), t (28) = - 1.98, p = .057. The sound and silent conditions did not differ in blocks 3 or 4.

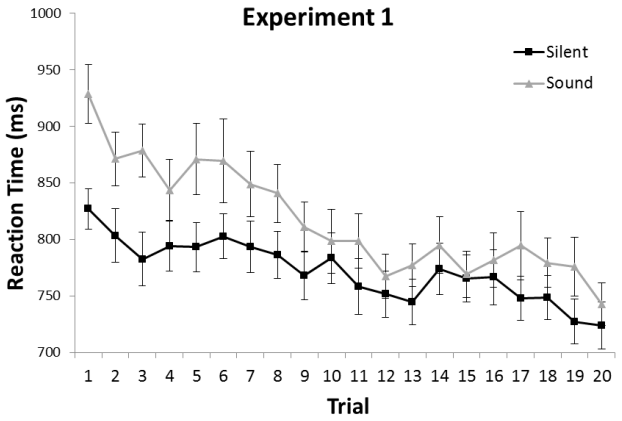


Figure 1. Mean response times across trials and condition. Error Bars denote Standard Errors.

In summary, the visual modality is typically well suited for processing of spatial information (Welch & Warren, 1980) and participants were clearly learning the visual sequences, as indicated by a speed up in response times. That said, the current experiment provides support for auditory dominance, with irrelevant tones slowing down responses to structured visual sequences. Even though participants were told to ignore the sounds, they couldn't, at least in the early stages of learning. By the end of the experiment, there was no significant difference in response times between the sound and silent condition. One possibility is methodological in nature and due to ceiling effects. However, it is also possible that this weakened interference stemmed from the visual task becoming more automated and less prone to cross-modal interference. It will be important to address this issue in future research.

Experiment 2

The goal of Experiment 2 was to determine if any auditory stimulus would slow down visual responses or if this effect was restricted to irrelevant tones. Participants in Experiment 2 were presented with the same two visual sequences used in Experiment 1 and one of the sequences was presented in

silence and one was paired with tones. In contrast to Experiment 1, the tones in this experiment were correlated with the location of the visual dots (e.g., every time a dot appeared in location 1, participants heard tone 1, etc.). If the presence of any auditory stimulus grabs attention and slows down visual processing, then the correlated/redundant auditory information may also slow down response times to the structured sequences. However, it is also possible that the interference is restricted to irrelevant or conflicting tones. If this is the case, then correlated sounds may not interfere with visual responses, with comparable response times in the silent and sound conditions. It is also possible that intersensory redundancy may also speed up processing (Colonius & Diederich, 2006; Giard & Peronnet, 1999), with response times in the correlated sound condition being faster than the unimodal visual baseline.

Method

Participants, Materials, and Procedure Thirty-one undergraduate student at The Ohio State University-Newark (13 Females, M = 19.17 years) participated in the study. In return, these students got credit for their research assignment in the Introduction to Psychology course. Data from one participant was not included due to reported hearing loss. Experiment 2 used the same visual and auditory stimuli used in Experiment 1. The procedure for Experiment 2 was exactly the same as the procedure for Experiment 1, but each visual stimulus location was paired with a specific auditory stimulus. Just as in Experiment 1, participants were instructed to touch the dot on the screen as quickly as possible when they see it appear.

Results and Discussion

Reaction times across the 20 trials are reported in Figure 2. As can be seen in the figure, response times sped up across training and auditory interference effects decreased compared to Experiment 1.

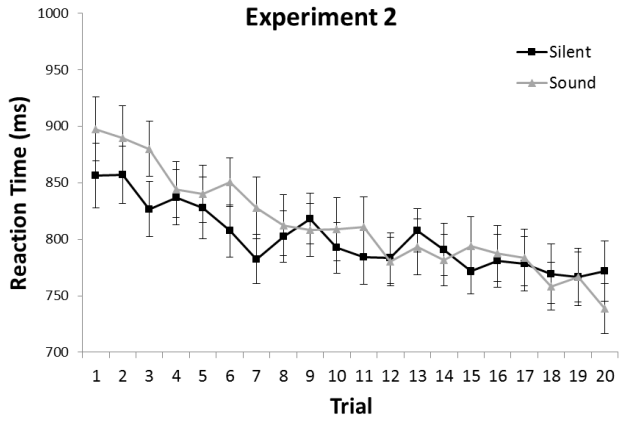


Figure 2. Mean response times across trials and condition. Error Bars denote Standard Errors.

Log transformed response times were averaged across five trials, and means were submitted to a 2 (silent vs. sound) x 4 (block 1, block 2, block 3, block 4) repeated measures ANOVA. The analysis only revealed an effect of time, F (3, 90) = 55.74, p <.001, showing that reaction time decreased across the blocks. The means and (*SE*s) for blocks 1 – 4 were 855.00 ms (19.98), 810.90 ms (19.67), 789.90 ms (18.69), and 770.60 ms (20.50), respectively. Paired samples t-tests, using log transformed data, were conducted between the blocks, showing block 1 to be significantly slower than block 2, t (30) = 7.50, p < 0.001. Block 2 was significantly slower than block 3, t (30) = 4.04, p < 0.001, and block 3 was significantly slower than block 4, t (30) = 3.02, p = 0.005. This displays a reliable pattern of learning across the blocks of the experiment. Although participants were initially faster in the silent condition at responding to visual input, the nonsignificant effect of condition and condition x block interaction suggest that cross-modal interference attenuated when using correlated sounds.

General Discussion

Many tasks require a person to divide attention across sensory modalities. The research on adults' processing of simultaneously presented auditory and visual information consistently points to visual dominance, with the visual modality dominating processing or responding (Colavita, 1974; Colavita, Tomko, & Weisberg, 1976; Colavita & Weisberg, 1979; Egeth & Sager, 1977). However, there are several recent studies highlighting situations where the presence of an auditory stimulus interferes, delays, or alters visual processing (Robinson & Sloutsky, 2013; Shams, Kamitani, & Shimojo, 2000; 2002). However, these studies employ tasks that rely almost exclusively on the processing of temporal information, which appears to be better suited for the auditory modality (Conway & Christiansen, 2005). The current study used a spatial SRTT, which should be better suited for the visual modality (Welch & Warren, 1980). In both of the reported experiments responding to a dot appearing in a predictable sequence sped up over time, suggesting that participants were learning the visual sequences. At the same time, response times were slower in Experiment 1 when the dot was paired with irrelevant tones compared to a silent condition, especially early in the experiment, which suggests that the auditory information was interfering with visual responses. Experiment 2 expands on this finding by showing that auditory interference is attenuated when using correlated sounds – sounds that also provided information regarding the location of the visual stimulus.

The finding that random sounds slowed down visual processing and/or responding more than correlated sounds is important and provides important insights into the nature of the cross-modal interference. For example, one explanation that can account for the findings in Experiment 1 is that the auditory and visual modalities share the same pool of resources and sensory modalities are competing for these resources (Robinson & Sloutsky, 2010). Moreover, because

auditory stimuli are dynamic and transient in nature it may be adaptive to allocate attention to this class of stimuli before they disappear. Thus, under high cognitive load conditions where resources are depleted or when examining speeded responses, auditory stimuli might automatically grab attention and attenuate or delay visual processing (Robinson & Sloutsky, 2010). The finding the auditory interference attenuated in Experiment 2 when sounds were correlated with visual information suggests that cross-modal interference effects may be occurring later in the course of processing, with only irrelevant and/or conflicting auditory information slowing down responding to visual input.

Several issues need to be examined in future research. First, irrelevant slowed down visual processing, and there was some evidence that correlated sounds weakened the effect. One possibility is that the effect is specific to auditory interference. However, it is also possible that irrelevant auditory stimuli increase task complexity or simply add conflicting information, which in turn slows down responses. For example, imagine a similar SRTT where dots vary in color. In the irrelevant condition, dot color varies randomly and does not predict spatial location. In the correlated/relevant condition, dot color is correlated with location. Finding that random colors also slow down learning would suggest that interference stems from increased task demands or any conflicting information, and the slowdown is not directly tied to auditory dominance per se. It will also be important to examine if interference effects are asymmetrical in nature - a signature pattern of modality dominance effects. The current study only examined the effect of sounds on visual sequence learning. Thus, to claim auditory dominance, future research will need to show that the presence of the visual sequence has no negative effect on auditory sequence learning.

Another issue that will need be addressed involves resolving the discrepancy between the current findings and the findings from the pip and pop effect (Van der Burg, Olivers, Bronkhorst, & Theeuwes, 2008). In the pip and pop task, participants are presented with numerous lines on a monitor and they have to quickly find a line that is perfectly vertical or perfectly horizontal, and then report out the line's orientation. The task is challenging when only presented visually; however, changing the color of the lines and synchronizing line color with an audible click significantly speeds up target detection. Interestingly, visual correlated cues do not speed up detection. This suggests that the clicking sound may result in automatic intersensory integration and highlight relevant visual information in complex visual scenes.

The intersensory integration account would predict that correlated sounds and possibly even random sounds should both facilitate target detection and speed up response times in the current study because of the close temporal presentation of the dot and sound. However, this was not the case. While this will have to be examined in future research, we believe the discrepant findings stem from the complexity of the visual scene. In the current speeded response task, the visual stimulus should automatically pop out because there is only one stimulus presented at a time. As visual scenes become more complex, such as in the pip and pop task, redundant intersensory information may make certain parts of the visual scene become more salient. This account is somewhat consistent with the Intersensory Redundancy Hypothesis (Bahrick & Lickliter, 2000), which states that intersensory redundancy automatically directs attention to the redundant information. For example, when infants are required to learn amodal information such as the tempo of a hammer tapping on a table, presenting this information visually and auditorily facilitates learning compared to when the same tempo is only presented visually or auditorily (Bahrick & Lickliter, 2000).

In summary, while most of the research in adult populations points to visual dominance (see Sinnett, Spence, & Soto-Faraco, 2007; Spence, Parise, & Chen, 2012 for reviews), the current study provides some support for auditory interference on a task that is typically better suited for the visual modality. These findings have implications for tasks that require quick processing and responding to multisensory information and shed light on potential mechanisms underlying modality dominance effects.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257-262.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190.

Burr, D., Banks, M.S., & Morrone, M.C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49-57.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16, 409-412.

Colavita, F.B., Tomko, R., & Weisberg, D. (1976). Visual pre-potency and eye orientation. *Bulletin of the Psychonomic Society*, 8, 25-26.

Colavita, F.B., & Weisberg, D. (1979). A further investigation of visual dominance. *Attention, Perception & Psychophysics*, 25, 345–347.

Colonius, H., & Diederich, A. (2006). The race model inequality: Interpreting a geometric measure of the amount of violation. *Psychological Review*, *113*, 148–154.

Conway, C.M., & Christiansen, M.H., (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24-39.

Dennis, N.A., Howard, J.H., & Howard, D.V. (2006). Implicit sequence learning without motor sequencing in young and old adults. *Experimental Brain Research*, *175*, 153–164.

Egeth, H.E., & Sager, L.C. (1977). On the locus of visual dominance. *Attention, Perception & Psychophysics*, 22, 77-86.

Giard, M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in

humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473-490.

Lewkowicz, D. J. (1988a). Sensory dominance in infants: 1. Six-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 155-171.

Lewkowicz, D. J. (1988b). Sensory dominance in infants: 2. Ten-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 172-182

Nissen, M.J., Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Robinson, C. W., & Sloutsky, V. M. (2013). When audition dominates vision: Evidence from cross-modal statistical learning. *Experimental Psychology*, *60*, 113-121.

Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135-141.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*, 1387-1401.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: *Listening (and learning) out of the corner of your ear. Psychological Science*, 8, 101–105.

Shams, L., Kamitani, S., Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature*, *408*, 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147-152.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, 69, 673–686.

Sloutsky, V.M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*(3), 822-833.

Song, S., Howard, J.H., & Howard, D.V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research*, *189*, 145–158.

Spence, C., Parise, C., & Chen, Y. C. (2012). The Colavita visual dominance effect. In M.M. Murray, & M.T. Wallace (Eds.), The Neural Bases of Multisensory Processes (pp. 529-556). Boca Raton, FL: CRC Press.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*, B25-B34.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychological Bulletin and Review*, *23*, 87-102.

Van der Burg, E., Olivers, C.N., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 1053-1065.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638-667.

Younger, B. A., & Cohen, L. B. (1983). Infant perception of correlations among attributes. *Infant Development*, *54*, 858-860.

Memory for exemplars in category learning

C. E. R. Edmunds (ceredmunds@gmail.com)
School of Psychology, Plymouth University
Plymouth, PL4 8AA, UK

Andy J. Wills (andy.wills@plymouth.ac.uk)
School of Psychology, Plymouth University
Plymouth, PL4 8AA, UK

Fraser N. Milton (F.N.Milton@exeter.ac.uk)
School of Psychology, University of Exeter
Exeter, EX4 4QG, UK

Abstract

Some argue that category learning is mediated by two competing learning systems: one explicit, one implicit (Ashby et al., 1998). These systems are hypothesised to be responsible for learning rule-based and information-integration category structures respectively. However, little experimental work has directly investigated whether people are conscious of category knowledge supposedly learned by the implicit system. Here we report one experiment that directly compared explicit recognition memory for exemplars between these two category structures. Contrary to the predictions of the dual-systems approach, we found preliminary evidence of superior exemplar memory after information-integration category learning compared to rule-based learning. This result is consistent with the hypothesis that participants learn information-integration category structures by using complex rules.

Keywords: category learning; memory; dual-systems; recognition;

One approach to categorization assumes that generalisation from past experiences to novel ones is mediated by two competing systems: one explicit and one implicit. The COVIS (COmpetition between Verbal and Implicit Systems) model is a popular instantiation of this approach (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby, Paul, & Maddox, 2011). In this model, the explicit, verbal system is described as learning by testing hypotheses using working memory. This system is assumed to optimally learn rule-based category category structures that can be easily verbalised such as the unidimensional structure shown in Figure 1A. In contrast, the implicit system is described as using procedural learning to associate areas of stimulus space with a motor response. This system is assumed to optimally learn category structures that cannot be easily verbalised such as the information-integration category structure shown in Figure 1B. Critically, the implicit system is assumed to "produce category knowledge that is opaque to declarative consciousness" (p.1, Smith et al., 2015).

Whether COVIS, and more broadly a dual-systems approach, adequately explains the processes of category learning is still a matter for debate. Proponents of COVIS argue that the case is closed, that the evidence for dual-systems approaches is overwhelming and that the field should move on to more interesting questions, such as those concerning the exact nature of the systems and how they interact with each other (Ashby & Maddox, 2011). In support of this view there is a large quantity of evidence that has been used to support COVIS (for a review see Ashby & Maddox, 2011). However, much of this evidence has been questioned (Edmunds, Milton, & Wills, 2015; Newell, Dunn, & Kalish, 2011; Stanton & Nosofsky, 2007; Zaki & Kleinschmidt, 2014). Also, despite the volume of studies, there is very little experimental work that directly investigates the key theoretical assumption that the learning of the implicit system is not available to consciousness. Instead, the focus has been on demonstrating that information-integration category structures are learned procedurally or demonstrating that learning of rule-based and information-integration categories are dissociable using experimental, neuropsychological or neuroscientific methods (Ashby & Maddox, 2005, 2011; Price, Filoteo, & Maddox, 2009). Therefore, the claim that the case is closed may be premature.

In the current study, we directly examine whether participants have conscious access to information about the information-integration categories they have learned. COVIS predicts that they do not, but some recent behavioral and neuroimaging evidence suggests otherwise. Behaviorally, Edmunds et al. (2015) found that the vast majority of participants were able to report a clear explicit strategy after training, regardless of whether they had been learning an information-integration or rule-based category structure. This was despite having met the criteria usually used in the COVIS literature to check that participants in the information-integration category structure condition are using the implicit system. This check uses a model-based strategy analysis inspired by General Recognition Theory (GRT; Ashby & Gott, 1988), a multidimensional version of signal detection theory. The failure of the model-based strategy analysis here may be because its output depends strongly on the set of strategies the modeller chooses to use, with the estimated proportion of "implicit" responders reducing substantially if more complex rule-based models are included (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015).

Turning to neuroimaging evidence, in a recent study from our lab we found greater activation of the medial temporal lobe in information-integration category learning, rela-

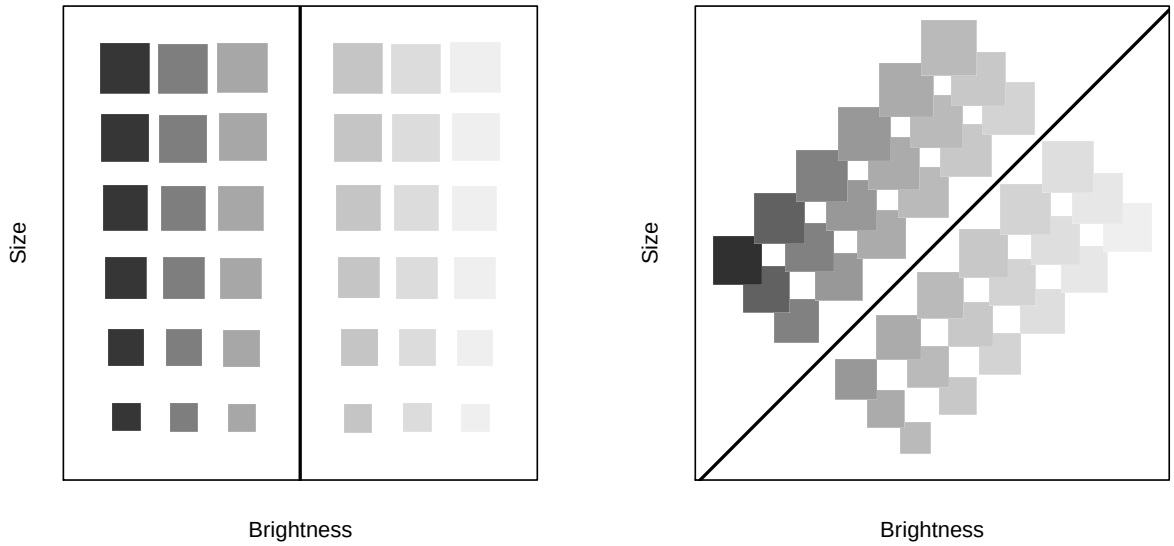


Figure 1: a) The unidimensional category structure with a rule based on size. b) The information-integration category structure with a positive category boundary.

tive to a well-matched rule-based category learning condition (Carpenter, Wills, Benattayallah, & Milton, in press). The medial temporal lobe has long been considered critical for explicit memory. Therefore, Carpenter et al.'s results suggest that information-integration category learning involves explicit memory processes to a greater extent than rule-based category learning.

One hypothesis that explains these two findings is that both rule-based and information-integration category structures are learned through the application of simple rules but that, for information-integration ones, those rules are supplemented by explicit memory of specific examples. For example, participants may store the examples that are exceptions to the simple rule. This hypothesis is also consistent with other evidence that found that participants who use a rule-plus-exception strategy have greater recognition memory for the exceptions to a simple category rule (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004). As participants learning information-integration categories would have to remember more exceptions than those learning a rule-based structure, we would also expect greater recognition performance for information-integration learners.

However, this evidence is not conclusive. Neither Edmunds et al. (2015) nor Carpenter et al. (in press) directly measure explicit access to category knowledge. Further, some of the results of Carpenter et al. are at variance with a previous neuroimaging study of rule-based and information-integration category learning (e.g. Nomura et al., 2007). Carpenter et al. argue that the differences between their study and that of Nomura et al. are due to methodological problems with the Nomura et al. study. This argument is supported by the fact that Carpenter et al.'s results are broadly in line with the only other closely-related neuroimaging study (Milton & Pothos, 2011). Nevertheless, more direct evidence is needed.

In the following experiment, we directly examined recognition memory for exemplars in rule-based and information-integration category learning. Recognition memory is commonly assumed to be a test of explicit memory processes (Gabrieli & Fleischman, 1995). If rule-based structures are learned explicitly and information-integration category structures are learned implicitly, as predicted by COVIS, then one would predict, if anything, better recognition memory performance for participants in the rule-based condition than for participants information-integration condition.

In contrast, our hypothesis is that that participants learn information-integration category structures explicitly using simple rules bolstered by memory for exceptions to those rules. This strategy would allow participants in the information-integration condition to score as highly as if they used the optimal diagonal decision bound, however it would also increase demands on recognition memory as participants would have to remember the exceptions. An alternative similar hypothesis is that participants in the information-integration condition may be using complicated rule strategies such as a conjunction rule. If this is the case, participants would still have to pay more attention to stimulus features information-integration than in the unidimensional condition as they would have to compare each stimulus to multiple boundaries. Whereas, for the unidimensional condition they only have to focus on one stimulus dimension. Either

way, we would expect to see better recognition memory in the information-integration condition than in the rule-based condition.

Method

Participants

Forty-two undergraduate psychology students were recruited from the Plymouth University participation pool and compensated with partial course credit.

Stimuli and category structures

The stimuli were 36 grey squares that varied in brightness and size displayed on a white background. The stimuli seen by each participant depended on which category structure they learned.

Half the participants were randomly assigned to learn a unidimensional rule-based category structure and the other half to learn an information-integration category structure as illustrated in Figure 1. The orientation of the category boundaries in abstract stimulus space were counterbalanced within conditions resulting in two unidimensional category structures–with a rule based solely on either the brightness (11 participants) or size of the square stimuli (10 participants)–and two information-integration category structures–where the optimum boundary had either a positive (10 participants) or negative gradient (11 participants). In addition, the stimuli were log-scaled so that all adjacent stimuli were approximately equally perceptually discriminable.

The abstract representation of the information-integration positive category structure is identical to that used by Spiering and Ashby (2008) with 6 stimuli added to bring the total number of stimuli up to 36. These stimuli were added to facilitate the random selection of a third of stimuli as "unseen" items for the recognition task. The remaining category structures are rotations ($\pi/4$, $\pi/2$, $3\pi/4$ rad) of this original structure around the origin and then translated so that 'center of gravity' of the points remained the same.

Procedure

The experiment was split into three phases: category training, recognition test and finally, category test.

Category training In this phase, participants were trained on two thirds of the available stimuli. The training stimuli were selected randomly for each participant subject to several constraints: 1) that those stimuli selected were symmetrical around the category boundary and 2) that no adjacent stimuli of similar difficulty were removed. In total there were 360 training trials, split into 3 blocks of 120 trials. In each block, the 24 stimuli were shown 5 times in a random order. On each trial, the participants looked at the stimulus until they made a response using either the "Z" key for Category A or the "/" key for Category B. Participants were unable to respond until at least 500ms had passed. Then, either "Correct" in green or "Incorrect!" in red was displayed for 500ms. A blank white screen was displayed between each trial for 500ms. Throughout the experiment, the labels "Category A" and "Category B" were displayed on the bottom left and right of the screen respectively. If participants took longer than 5000ms to respond, no corrective feedback was given, instead the message "PLEASE RESPOND FASTER" was displayed for 500ms.

Recognition test In this phase participants judged whether each stimulus was "old" and appeared in the training phase, by pressing the "O" key, or was "new" and had not been shown in the training phase, by pressing the "N" key. The words "New" and "Old" were presented on the bottom left and right of the screen respectively. After this, participants judged the confidence they had in their old-new judgement on a Likert scale that varied from 1 (=guessed) to 5 (=certain) by pressing the corresponding number key. Each of the 36 stimuli were presented three times in a randomised order. No feedback was given.

Category test In this phase, participants were asked to judge the category membership of all 36 stimuli, not just those they had seen in the category training phase. The procedure was identical to that of the training phase, apart from there was no feedback. Each of the 36 stimuli were presented three times in a random order.

Verbal report questionnaire At the end of the experiment, participants were asked to complete a questionnaire that asked them to describe in detail the strategy that they used. This was to determine whether the participants could explicitly report the strategy they used and whether any participants used a rule-plus-exception strategy. The questionnaire asked them to "Imagine that another participant was asked to complete the experiment as you did. What instructions would you give them so that they could exactly copy your pattern of responding?"

The verbal reports were coded by CERE and AJW.

Analysis

All data analyses were conducted in R (R Core Team, 2014). For every condition and phase of the experiment, all trials were removed for which the reaction times were outliers in that condition (i.e. outside 1.5 times the interquartile range above the upper quartile and below the lower quartile).

Results

One participant was removed from the unidimensional condition because their accuracy score was consistently below chance (i.e. 50%), resulting in 20 and 21 participants in the unidimensional and information-integration category structure conditions respectively.

Performance

Category learning We found a statistically significant difference in categorization accuracy at test, $F(1,39) = 13.51$, $p < .001$. Proportion correct was higher for the unidimensional category structure, $M_{UD} = 0.87$, $SD = 0.07$, than for

the information-integration category structure, $M_{II} = 0.78$, $SD = 0.11$.

Recognition To determine overall memory performance, we calculated d' values for each participant.

We found that there was a statistically significant difference in d' between the two category structure conditions, $t(39) = 2.04$, $p = .048$. Specifically, d' was higher for the information-integration category structure, $d' = 0.01$, $SD = 0.02$, than for the rule-based category structure, $d' = 0.00$, $SD = 0.01$. Further, d' was significantly greater than chance in the information-integration category structure condition, $t(20) = 2.98$, $p = .007$, but not in the unidimensional category structure condition, $t(19) = 0.67$, $p = .511$.

Strategy analyses

The performance analyses above indicate a slight memory advantage for stimuli in the information-integration category structure condition compared to those in the unidimensional category structure condition. In this section, we investigate possible sources of this advantage.

Model-based analysis One possibility consistent with COVIS is that the category structure manipulation failed to result in a corresponding shift in category learning system. In other words, participants in the information-integration condition could have been using the sub-optimum, explicit system. If this were the case, then it would not be surprising that participants had explicit memory for category information. This is always a concern for experiments in the COVIS literature. The usual solution is to conduct a model-based analysis to determine which strategies participants are using to learn the structure. If the majority of participants in the information-integration condition are identified by the analysis as using the optimum diagonal strategy, then proponents of COVIS would conclude that those participants are using the implicit system.

In the model-based analysis typically used in the COVIS literature (Ashby & Gott, 1988), four types of model are fitted to the data from each participant. These model types are qualitatively different types of optimum decision boundaries that split the stimulus space into two, with "Category A" responses on one side and "Category B" responses on the other.

The *unidimensional* models assume that the stimuli are categorised on the basis of a single stimulus dimension. In this case, there are two possible unidimensional models: the stimuli can be split either on the basis of brightness or size. A unidimensional rule based on brightness would be "Place the light squares in Category A and the dark ones in Category B". This would be represented in stimulus space as a vertical or horizontal line. This model has two parameters: the value at which the boundary crosses the axis and perceptual noise.

The *conjunction* models assume that participants make a decision for each stimulus dimension and then combines them to determine category membership. A conjunction rule in this case might be "Place the light, small squares in Category A. Everything else is in Category B." This model can have up to three parameters: a decision criterion on each dimension and a noise parameter. Four versions of the conjunction model were included corresponding to each corner of the stimulus space.

The *general linear classifier* (GLC) models assume that the decision boundary between the categories can be described by a diagonal line between them. This is the optimum strategy for the information-integration condition. This model can have up to three parameters: the gradient, intercept and a noise parameter.

Two types of *random* models were also included: unbiased and biased. In the unbiased model, it is assumed that for every stimulus the participant is equally likely to pick either category. There are no parameters in this model. In the biased random model, it is assumed that for every stimulus the participant in likely to ascribe it to Category A with a certain probability (i.e. 30%). This model has one parameter, the proportion of Category A responses, and is a more general version of the unbiased random model, for which the parameter is equal to 50%.

The data from each participant was fitted to each of these models. The degree of fit was measured by the Bayesian Information Criterion (Schwarz, 1978). The results from this analysis, which was performed using the grt package in the R environment Matsuki (2014), are reported in Table 1.

Table 1: Strategies identified in each condition according to the model-based analysis.

	Strategies			
Condition	UD	GLC	CJ	RND
UD	14	2	3	-
II	6	10	4	1

Category structures: UD=Unidimensional, II=Information-integration. Models: UDX=Unidimensional based on brightness, UDY=Unidimensional based on size, GLC=General linear classifier, CJ=Conjunction RND=Random (both types).

Here we can see that our data meets the criterion commonly used in the COVIS literature: the majority of participants in the information-integration condition have been found to be using the optimum GLC (or diagonal) strategy. Researchers in the COVIS framework would normally conclude from this that people in the information-integration condition were using the optimum implicit system.

Verbal reports The model-based analysis indicates that our data meet the minimum requirements for a COVIS study: there was an obvious effect of strategy type depending on which category structure participants learned. However, as mentioned in the introduction, the results of this analysis have been found to depend on the models included in the analysis

Table 2: Strategies identified in the verbal report questionnaire

	Strategies							
	UD	seqUD	CJ	CJ2	RuleX	Implicit	Other	None
UD	15	1	1	0	0	1	0	2
II	4	4	3	5	1	0	1	4

Category structures: UD=Unidimensional, II=Information-integration. Strategies: UD=Unidimensional, seqUD=Sequential unidimensional, CJ=Conjunction, CJ2=Double conjunction, RuleX=Rule-plus-exception.

(Donkin et al., 2015). Additionally, our previous research found that participants identified as diagonal classifiers could also report using complex rule-based strategies (Edmunds et al., 2015). In this section we examine the strategies that participants report using to see if participants in the information-integration condition report using complex rule-based strategies.

Participants reported several different types of strategy. A report was classified as a *unidimensional* strategy if it described a rule based only of the stimulus dimensions such as "the dark squares are Category A and the light ones are Category B."

A report was classified as a *sequential unidimensional* strategy if the participants first categorised the stimuli at the extreme ends of one stimulus dimension and then defined a second unidimensional rule, on the other stimulus dimension, for the stimuli in the middle of the dimension. For example, a participant might say: "The very small stimuli were in Category A, and the very large in Category B. For the middle sized stimuli, the light ones were in Category A and the dark in Category B."

A report was classified as a *conjunction* strategy if the participant described an AND rule on the basis of two stimulus dimensions such as "Stimuli that are both small and dark are in Category A; else Category B."

A report was defined as a *double conjunction* strategy if the participant described two opposing corners of the stimulus space, but failed to define the other areas of the space. For example: "Large and dark patterns go into B. Small and light colours into A." As can be seen in this example, the participant fails to describe what category small dark stimuli would be in.

A report was classified as a *rule-plus-exception* strategy if the participant reported a simple rule with some exceptions. For example, "Light stimuli were usually Category A and dark stimuli Category B. However, one light medium sized stimulus was in Category B."

A report was classified as an *implicit* strategy if the participant recommended "not thinking too much" or to "rely on instinct" or similar phrases.

The inter-rater reliability for whether or not a participant reported a strategy was perfect. Similarly, both coders agreed perfectly on the strategies participants used.

As we can see in Table 2, no participants reported using an implicit strategy in the information-integration condition. This replicates our finding that the model-based analysis does not correspond well to the strategies participants report (Edmunds et al., 2015). Furthermore, only one participant in the information-integration condition used a rule-plus-exception strategy. This indicates that the advantage in memory may not be due to the use of a particular strategy, but because of the need in complex strategies to attend closely to the stimuli in order to categorise them by comparing to multiple decision bounds.

Discussion

Summary

A key dual-system model of category learning, COVIS, predicts that categorization is mediated by two competing learning systems: one explicit and one implicit. These two systems are hypothesised to optimally learn two different types of category structure. The explicit verbal system optimally learns rule-based category structures, whereas the implicit system optimally learns information-integration category structures. A key feature of this model is that category knowledge learned using the implicit system is unavailable to consciousness (Smith et al., 2015). In contrast, behavioral and neuroscientific work from out lab indicates that participants learning information-integration categories are aware of category knowledge and may be using explicit memory to facilitate categorization (Carpenter et al., in press; Edmunds et al., 2015). This experiment aimed to directly test these possibilities by comparing participants' performance on an old-new recognition task after learning either a unidimensional rule-based category structure, or an information-integration one.

Contrary to the predictions of COVIS, we found superior memory for exemplars after learning an information-integration category structure compared to a rule-based one. This indicates that participants may learn information-integration category structure using complex rule-based strategies rather than implicitly. This would result in superior exemplar memory for items when learning a information-integration category compared to a unidimensional structure as participants would have to attend more closely to the stimuli's features in order to compare them to multiple decision boundaries. Previous conclusions in the COVIS literature

concerning the presence of implicit-like category learning may be due to an over-reliance on the assumption that a limited model-based analysis can provide evidence for implicit responding (Donkin et al., 2015).

One apparent limitation of the current study is that response accuracy in the category test phase is lower in the information-integration condition than in the rule-based condition. In an ideal comparison between rule-based and information-integration learning, the conditions would be matched for error rate. However, it seems likely that improving overall performance on the information-integration could only improve memory for the exemplars as this would involve using a more refined rule-based strategy. Thus, it seems unlikely that such a change would qualitatively alter our conclusions.

Another potential limitation is that recognition performance is poor in both conditions of the current experiment. This may be due to the stimuli being perceptually very similar to one another. Further work might increase discriminability by adding additional features that were not predictive to category membership. We are currently investigating this possibility.

In conclusion, this experiment finds preliminary evidence that participants learning the information-integration category structure do so explicitly. This conclusion is in contrast to the prediction of the COVIS model, which assumes that the information-integration structure is learned implicitly.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178.

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 147–161.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 1–13). New York: Cambridge University Press.

Carpenter, K. L., Wills, A. J., Benattayallah, A., & Milton, F. N. (in press). A comparison of the neural correlates that underlie rule-based and information-integration category learning. *Human Brain Mapping*.

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 933–948.

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures . *The Quarterly Journal of Experimental Psychology*, *68*(2), 1203–1222.

Gabrieli, J. D. E., & Fleischman, D. (1995). Double dissociation between memory systems underlying explicit and implicit memory in the human brain. *Psychological Science*, *6*(2), 76–82.

Matsuki, K. (2014). grt: General recognition theory [Computer software manual].

Milton, F., & Pothos, E. M. (2011). Category structure and the two learning systems of COVIS. *European Journal of Neuroscience*, *34*(8), 1326–1336.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). *Systems of Category Learning. Fact or Fantasy?* (Vol. 54). Elsevier Inc.

Nomura, E. M., Maddox, W. T., Filoteo, J. V., Ing, a. D., Gitelman, D. R., Parrish, T. B., ... Reber, P. J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex*, *17*(1), 37–43.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 548–568.

Price, A., Filoteo, J. V., & Maddox, W. T. (2009). Rule-based category learning in patients with Parkinson's disease. *Neuropsychologia*, *47*(5), 1213–1226.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534–553.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Smith, J. D., Zakrzewski, A. C., Herberger, E. R., Boomer, J., Roeder, J. L., Ashby, F. G., & Church, B. A. (2015). The time course of explicit and implicit categorization. *Attention, Perception, & Psychophysics*, *77*(7), 2476–2490.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information integration, but not rule-based category learning. *Psychological Science*, *19*(11), 1169–1177.

Stanton, R. D., & Nosofsky, R. M. (2007). Feedback interference and dissociations of classification: evidence against the multiple-learning-systems hypothesis. *Memory & Cognition*, *35*(7), 1747–1758.

Zaki, S. R., & Kleinschmidt, D. F. (2014). Procedural memory effects in categorization: evidence for multiple systems or task complexity? *Memory & Cognition*, *42*(3), 508–24.

Chinese and English speakers' neural representations of word meaning offer a different picture of cross-language semantics than corpus and behavioral measures

Benjamin D. Zinszer (bzinszer@ur.rochester.edu)
Andrew J. Anderson (andrewanderson@ur.rochester.edu)
Rajeev D. S. Raizada (rajeev.raizada@gmail.com)
Rochester Center for Brain Imaging
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627 USA

Abstract

Speakers of Chinese and English share decodable neural semantic representations, which can be elicited by words in each language. We explore various, common models of semantic representation and their correspondences to each other and to these neural representations. Despite very strong cross-language similarity in the neural data, we find that two versions of a corpus-based semantic model do not show the same strong correlation between languages. Behavior-based models better approximate cross-language similarity, but these models also fail to explain the similarities observed in the neural data. Although none of the examined models explain cross-language neural similarity, we explore how they might provide additional information over and above cross-language neural similarity. We find that native speakers' ratings of noun-noun similarity and one of the corpus models do further correlate with neural data after accounting for cross-language similarities.

Keywords: cross-language semantics, multivoxel pattern analysis, semantic models, concept representation

Introduction

Multi-voxel pattern analyses and neural decoding methods have recently enabled cognitive neuroscientists to predict patterns of brain activity for word stimuli by generalizing from a training set of words and their associated functional neuroimaging data to new words (e.g., Mitchell et al., 2008) or new participants (e.g., Raizada & Connolly, 2012). One application of this inference has been using functional brain data to link words across languages, known as neural translation. Studies of bilingual speakers have shown decodable associations between a speaker's mental representations of the same word in each of their two languages (Buchweitz et al., 2012; Correia et al., 2014). In recent work, we translated a small lexicon of seven words between native speakers for Chinese and English by comparing each group's functional brain activity in representational similarity space (Zinszer et al., 2015).

Findings in neural translation are generally consistent with hypotheses that suggest semantic representations should be strongly correlated across speakers based on shared extrinsic (e.g., sensory) information (Binder & Desai, 2011; Hauk et al., 2006). For example, the appearance, sound, and general functions of a dog would be roughly the same for speakers of any language. Zinszer et al.'s (2015) neurally-based translation provides evidence that word representations do encode this perceptual information. Thus, on one hand, neural representations of translation equivalent words should be the same to the extent that experience in the world is shared across speakers of different languages. On the other hand, research in translation ambiguity shows that even translation equivalents evoke unique semantic information across languages (Degani & Tokowicz, 2010; Malt & Majid, 2013). Highly accurate models of word meaning should reflect some language-specific information and diverge across languages.

In corpus-based models, word co-occurrence statistics differ between translation-equivalent words when those words are used in systematically different ways across languages. This problem is widely known in cross-language information retrieval (CLIR) and has inspired a number of approaches for improving translation accuracy (Zhou et al., 2012). However, many translation models rely on parallel documents and contextual clues to compare translations for a particular token. This information is not available to human participants or models for the isolated single-word stimuli used in most neural decoding studies.

Corpus-based semantic models reflect something important about word meaning as it is represented in the brain, as evidenced by their successful application to neural decoding (as in Mitchell et al., 2008), but the cross-language semantic differences that have vexed machine translation do not seem to prevent neural translation of small lexicons (Zinszer et al., 2015). In this sense, corpus-based models of word meaning may over-estimate these differences relative to speakers' actual mental representations.

In this study, we contrast three classes of model for word meaning: (1) Neural data for seven translation-equivalent concrete nouns in Chinese and English obtained by Zinszer et al. (2015), (2) two simple corpus-based models for the seven words in each language, and (3) behavioral models of word representation elicited from native speakers of each language. We compare these models using representational similarity analysis, evaluating cross-language similarity within the models and comparing the corpus- and behavior-based models to the neural data.

We predicted that corpus- and behavior-based models of word representation would reflect the translation relationships between words through high cross-language

correlations, although we expected this correlation to be somewhat attenuated in the corpus models given the nuances of translation ambiguity captured in the source material. We expect these models to be strongly correlated to the neural data. However, if the corpus- or behavior-based models also capture language-specific aspects of meaning, we predict that the models will correlate with variance in neural data *not* explained by the cross-language similarity between the neural data of the two groups.

Method

Target Nouns

The target nouns were seven translation-equivalent words in English and Chinese. These words were selected based on four requirements: (1) concrete nouns, (2) monosyllabic in both languages, (3) represented by a single Chinese character, and (4) unlikely for English translations to be known by the Chinese participants (see Table 1 for list). To verify criterion (4), Chinese participants in the fMRI study completed a brief quiz in which they wrote the English translations for twenty Chinese words. Mean translation accuracy was 1.56 of the 7 target nouns. In the following analyses, we describe a variety of different multivariate representations for these target nouns based on data from neuroimaging, corpus, and behavioral measures.

Table 1. Target nouns (first and bolded) and semantic associates in English and Chinese (with pinyin)

English	Chinese
axe, knife, sword	**斧 (fǔ)**，锯，剑
broom, mop, vacuum	**帚 (zhǒu)**，拖把，簸箕
gown, dress, robe	**袍 (páo)**，裙，礼服
hoof, paw, foot	**蹄 (tí)**，爪，脚
jaw, mouth, bone	**颚 (è)**，口，骨
mule, horse, donkey	**骡 (luó)**，马，驴
raft, boat, barge	**筏 (fá)**，船，艇

Neural Representations

Functional MRI data were acquired from Zinszer et al.'s (2015) neural translation experiment. In that experiment, eleven native speakers of English (4 M / 7 F) and Mandarin Chinese (3 M / 8 F) at Dartmouth College completed a simple semantic relatedness task while undergoing fMRI. Participants viewed 49 words (7 target nouns, 42 filler) in pseudorandom order, repeated over seven runs. To ensure semantic processing of the stimulus words, they were periodically asked to rate the semantic relatedness of a word to the preceding word. Individual participants' neural responses were estimated for the seven target nouns based on a GLM model of functional activity. Each participant's functional responses were abstracted into similarity space (see "Representational Similarity Analysis") and these similarity structures were averaged across participants within the same language group. Similarity structures were estimated for whole brain data as well as for each of 96 ROIs in the Harvard-Oxford neuroanatomical atlas. Further details of these procedures are described in the original study (Zinszer et al., 2015).

Corpus-Based Representations

Following the method of Mitchell et al. (2008), semantic representations were constructed for each target noun based on their co-occurrence rates with 25 verbs in a large corpus of webpages. The 25 Chinese verbs were obtained from five Chinese-English bilinguals who translated the English verbs independently and were then aggregated based on majority consensus (see Table 2). In a few cases, two English verbs most frequently translated to the same Chinese verb (e.g., hear and listen as 听). In these cases, a lower frequency or compound Chinese verb with similar meaning was substituted to maintain 25 unique dimensions.

We used the Leeds University query tool for two similar Internet corpora in Chinese (90 million words) and English (160 million words; http://corpus.leeds.ac.uk/internet.html; Sharoff, 2006) to obtain concordance counts in a three word window for each noun-verb pair. Count data vectors were normed to unit length, yielding a 25-dimensional vector for each noun.

Due to the low frequency of some target nouns, we also generated broadened representations from the same corpus data. For each target noun, two associated nouns were selected to help bring out defining elements of the target noun relative to noise-level co-occurrences (e.g., *mop* and *vacuum* capture information about shape and function of *broom*). The three 25-d vectors for the target noun and its two associates were averaged to produce the broadened representation. Table 1 lists associates of each target noun.

Table 2. Verbs in English and Chinese

English	Chinese	English	Chinese
see	看	enter	进入
say	说	move	移动
taste	尝	listen	倾听
wear	穿	approach	接近
open	开	fill	填
run	跑	clean	清理
near	靠近	lift	举
eat	吃	rub	擦
hear	听	smell	闻
drive	驾驶	fear	怕
ride	骑	push	推
touch	碰	manipulate	操纵
break	打破		

Behavior-Based Representations

Eleven native English speakers (4 M / 7 F) and eleven native Chinese speakers (6 M / 5 F) at the University of Rochester (Rochester, NY, USA) completed two semantic

relatedness judgment tasks in their respective native languages.

In the first task (hereafter the NN ratings), participants judged the semantic relatedness for every pairwise combination of the seven target nouns (21 total comparisons). Order of words in each pair and order of pairs were randomized for each participant. Semantic relatedness judgments were made on a scale of 1 (unrelated) to 5 (highly related) and averaged over participants.

After an unrelated intervening task (visual categorization of animal-like figures), participants made binary semantic relatedness judgments for every target noun with each of the 25 verbs (175 total binary ratings, hereafter the NV ratings). Responses were averaged across participants to yield decimal values for each NV pair.

Representational Similarity Analysis

Representational similarity analysis compares the way a set of referents (such as the seven target nouns) is organized in different representational spaces. For example, multivariate brain activation patterns and corpus-based models described in the preceding sections can be compared to each other when each is abstracted into similarity space. This similarity space is composed of Pearson pairwise correlations between the multivariate representations for each of the seven target nouns, resulting in a 7x7 similarity matrix in which each noun is described by the correlation of its multivariate representation to those of the other six nouns. The correlation values are further transformed using Fisher's *r*-to-*z* for normalizing correlation coefficients. We can then compare two 7x7 similarity spaces by Pearson correlation of their 21 unique values (the lower-left triangle of the matrix).

Results

We computed similarity structures for the each measure of neural and semantic relatedness. The NN ratings were used directly by averaging the English and Chinese participants' responses to the semantic relatedness task for each noun .

Cross-Language Correlations

For all measures, 21 unique values from the 7x7 similarity matrices (as in Figure 1) were correlated across languages, such as NN English vs. NN Chinese. Table 3 reports these cross-language correlations. The whole brain fMRI data and both sets of behavioral ratings (NN and NV) reflected strong cross-language correlations (r<0.60, p<=0.003). The Leeds corpus measure showed no such cross-language correlation, although the Broadened Leeds representations were moderately correlated (r=0.44, p<0.05).

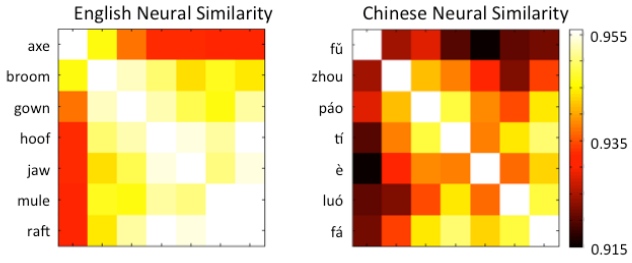


Figure 1. Neural similarity in each language group. Comparable similarity matrices were computed for the target nouns in each corpus and behavioral model, allowing cross-language and cross-model comparisons.

Table 3. Cross-language correlations for each measure of word representation

Measure	*r*	*p*
MRI - Whole brain	0.89	< 0.001
NN rating	0.85	< 0.001
NV rating	0.61	0.003
Broadened Leeds	0.44	0.045
Leeds	-0.08	0.726

Model-to-Brain Correlations

Next we compared each semantic model to fMRI data from the corresponding language to measure the degree to which patterns of functional activity correspond to the words' semantic representations in the models. These correlations are reported in Table 4 for the whole brain data, and summary statistics are provided for correlations across 96 anatomical ROIs as defined by the Harvard-Oxford brain atlas (http://www.fmrib.ox.ac.uk/fsl/).

Table 4. Correlation of each semantic model to whole brain neural data and summary statistics for correlations at ROIs.

	Whole brain		Harv.-Oxf. ROIs		
Model	***r***	***p***	**mean *r***	***s.d.***	**max |*r*|**
English					
NN	0.15	0.52	0.16	0.13	0.47
NV	0.11	0.64	0.05	0.13	0.49
Broad	-0.24	0.30	-0.14	0.17	0.60
Leeds	-0.24	0.30	-0.13	0.15	0.41
Chinese					
NN	-0.11	0.62	-0.09	0.13	-0.36
NV	-0.08	0.72	-0.10	0.14	-0.43
Broad	-0.47	0.03	-0.45	0.11	-0.69
Leeds	-0.17	0.46	-0.15	0.10	-0.38

All semantic models and the two sets of whole brain data were plotted using multi-dimensional scaling for cosine similarity (Figure 2, next page). As suggested by the correlations to brain data reported in Table 4, Chinese and English speakers' patterns of functional response were closer to one another than they were to any of the semantic models. Behavior-based models (NN and NV) clustered together in the MDS projection, and the NN models were both the most similar across languages and the closest approximations of the brain data. The Leeds corpus-based representations were distantly separated from one another and from the brain data, but the Broadened model in

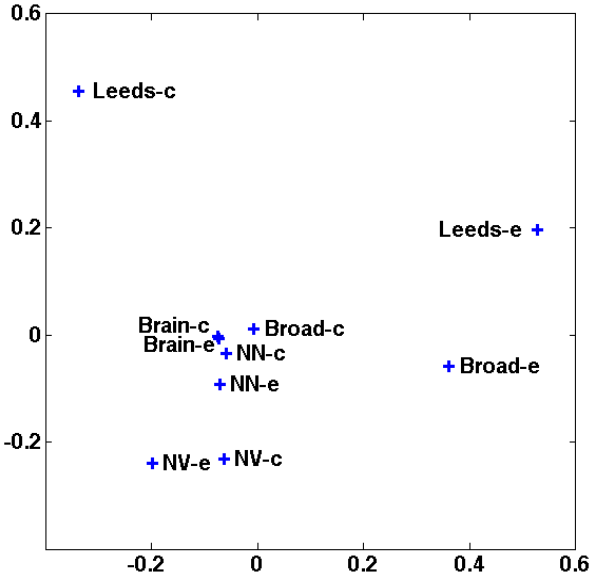


Figure 2. MDS plot of similarity space for semantic models and whole brain representations in Chinese (c) and English (e). Brain-e and Brain-c representations are overlapping.

Chinese more strongly (though negatively) correlated with the brain data.

Residuals Analysis

The four semantic models offered a poor account of the neural responses in both languages, suggesting that the cross-language similarity supporting neurally-based translation was not explained by any of the proposed corpus- or behavior-based models. In our next analysis, we turned our attention to the language-specific components of the neural representations by isolating the residuals of the neural representations in each language that remained after accounting for cross-language neural correlations.

In this residuals analysis, we again correlated the semantic models (behavior- and corpus-based) to neural data, but we evaluated the models against only the unexplained variance in the Chinese and English neural data that remained after cross-language correlation (that is, the residuals of the correlation). Here we ask to what degree our semantic models explain the remaining variance between the Chinese and English neural representations after accounting for the language-invariant components.

For each ROI, we saved the residuals from the correlation between the Chinese and English neural representations and used them in lieu of the original neural representations. We then repeated the correlations reported in the preceding section. We found that across the 96 ROIs, correlations were generally weak and distributed around zero (see Table 5).

Because this analysis is principally concerned with the language-specific contribution of the semantic models after controlling for cross-language similarity, we searched for ROIs where the *both* correlation between the English model and English residuals *and* the correlation between the Chinese model and Chinese residuals were significant. Thus, each ROI has a p-value indicating the probability of both languages' models producing significant correlations under the null hypothesis. Using this constraint, a critical p-value of $\sqrt{\alpha}$ in *each* language produces a joint probability of Type I error equal to α for an ROI if the null hypothesis is true. Bonferroni correction for multiple comparisons across the 96 ROIs yields a threshold of $\alpha = 0.00052$ for each ROI, and thus we highlighted ROIs where *both* the Chinese and English correlations were $p < \sqrt{0.00052}$ or 0.023.

Table 5. Summary statistics for correlations of each semantic model to neural data in Harvard-Oxford ROIs.

	Harv.-Oxf. ROIs		
Model	**mean *r***	***s.d.***	**max |*r*|**
English			
NN	0.14	0.18	0.56
NV	-0.02	0.20	0.42
Broad	-0.03	0.23	0.63
Leeds	0.07	0.19	0.40
Chinese			
NN	-0.11	0.19	-0.66
NV	-0.08	0.21	-0.55
Broad	-0.30	0.18	-0.69
Leeds	-0.10	0.15	-0.39

Table 6. Cortical regions in which a semantic model significantly correlates with neural representations in the same language after controlling for cross-language neural similarity.

			Model-to-brain	
Model	**Harvard-Oxford ROI**		**Eng. *r***	**Chi. *r***
NN	3	R Middle Frontal Gyrus	0.52	-0.53
	16	R Postcentral Gyrus	0.56	-0.66
Broad	40	R Frontal Operculum	0.64	-0.53
	46	L Supracalcarine Gyrus	0.51	-0.61

The NN behavioral model and Broadened Leeds corpus model yielded significant results in both languages. Table 6 lists the regions where English and Chinese models explained additional variance in their respective languages beyond the cross-language neural similarity. We also investigated whether the semantic models in each language exclusively correlated to the neural data for that language, or whether they also correlated with the opposite language. This test of double-dissociation identifies whether the semantic model is using language-specific information to explain variance in the target language (e.g., English model only correlates with English neural data) or language-invariant information (e.g., English model correlates with both English and Chinese neural data). In all four regions of interest identified in Table 6 we found that at least one of the models (English or Chinese) significantly correlated with the neural data in both languages (p<0.05).

Discussion

In this paper, we compared four semantic models (two corpus-based and two behavior-based) in Chinese and English with neural similarity structures that have previously been used to translate between speakers of English and Chinese. The strong cross-language correlations in the neural data reported by Zinszer et al. (2015) indicate that important language-invariant information underlies the neural responses to word stimuli. Cross-language correlations revealed in the behavioral models also supported this conclusion, but these models did not correlate with the neural responses, pointing to another source of information. Given the previous success of using corpus-based models of word meaning to decode neural data (e.g., Mitchell et al., 2008), we reasoned that language-specific aspects of word meaning might be captured by our models and offer insight into the what aspects of the neural representations were language-specific and what aspects were attributable to universally available conceptual information (i.e., world-knowledge).

Representations Shared Across Languages

The strong cross-language correlations in the neural data indicate that language-invariant semantic information underlies a significant portion of the neural responses to word stimuli. Although neither the corpus nor behavioral models showed the same degree of cross-language convergence, the NN ratings were very close despite the potential differences in word meanings across languages (Degani & Tokowicz, 2010). In this sense, directly querying native speakers' intuitions about the similarity in word meanings (the NN model) produced the most accurate reflection of cross-language similarities in the neural representation of word meaning. Analogous ratings of object similarity have proved useful in identifying individual differences in the neural representations of objects (Charest et al., 2014), so this close link between direct behavioral query of a representation (the NN ratings) and neural similarity is not surprising.

Surprisingly, correlations between the whole-brain data and all four semantic models proved almost entirely insignificant, with the exception of the Broadened Leeds model's correlation with the Chinese neural data. Searching the 96 cortical regions of the Harvard-Oxford atlas did not yield strong correlations between the neural representations and most of the semantic models. Only the Broadened Leeds corpus showed an overall trend towards correlating with the Chinese data, but none of these correlations survived correction for multiple comparisons.

Taken as a whole, these results are suggestive of at least two important underlying sources of information. The regularities in similarity structures between languages suggest that both the neural data and behavioral model can only be explained by language-invariant representations, and the failure of these two measures to correlate with one another suggests that they may capture different aspects of semantic knowledge. The corpus-based models did not show as much cross-language regularity, nor did they correlate with the neural data, leaving their meaning ambiguous in this context.

Language-Specific Representations

The initial corpus-based semantic models were extremely divergent from one another across languages, suggesting that they may encode highly language-specific aspects of word meaning. These models perhaps even exaggerate the differences between speakers' semantic representations in Chinese and English, in light of the high correlations across the behavioral models. However, this divergence is not as surprising when one considers the broad structural differences between written Chinese and English.

Although neither the corpus nor behavioral models showed explanatory power for the neural representations, we show in the residuals analysis that some divergent information is contained in this neural signal. The noun-noun similarity ratings and Broadened Leeds corpus models both added explanatory power over and above the direct comparison of neural representations in each language, confirming that these models encode real, neurally-implemented information that is not shared across the speakers' neural responses in each language. However, this explanatory power was (paradoxically) not language-specific. Every region in which the models significantly correlated with their own language was also explained by at least one model from the other language. This observation challenges the assumption that the NN or Broadened Leeds semantic models represent purely language-specific information. An alternate explanation may be that language-invariant information encoded in these models is differentially represented in the brains of speakers of one language or the other.

The ROIs showing the strongest such relationships are also intriguing. Bilateral postcentral gyri produced strong correlations between languages in Zinszer et al.'s (2015) cross-language comparison of neural data, a fact that was taken as evidence for language-independent somatosensory involvement. The correlation between representations in the L Supracalcarine gyrus and the Broadened Leeds corpus are somewhat puzzling, given this region's importance in visual processing. The corpus model is not particularly visually oriented, and one would not expect visual correlates of meaning to be more apparent in a model of word co-occurrence than in a behavioral model or cross-language neural similarity.

Future Work

Several perplexing and tantalizing questions arise from these comparisons of neural, corpus, and behavioral models of word meaning in Chinese and English. One major limitation of the present study is the amount of statistical power available from an analysis of only seven words in each language. While many of our investigations yielded null results, this finding could arise from a genuinely null

relationship or from the impoverished representation of each language in this highly constrained set of stimuli.

Further, the meaning of significant negative correlation between the broadened Chinese corpus model and Chinese neural data is not immediately clear. This correlation indicates regularity in both structures (less similarity in the corpus reliably corresponds to greater similarity in the brain), but we propose that these measures may be mutually informative in future studies of neural decoding. Our own informal investigations suggest that the co-occurrence relationships of interest may be more remote than detectable in the three word window, thus skewing the present model based on syntactic constraints.

Finally, while we know that functional responses of Chinese and English speakers contain information that can discriminate between words in both languages, this shared information is still not adequately explained by any individual model described in this paper. What then do each of the neural, behavioral, and corpus models represent? Recent neural decoding research has focused on decomposing the neural signal into interpretable components of meaning (e.g., sensory modalities or cognitively plausible features). An integrative approach has indeed proved important for comparing experiential and corpus-based representations (Andrews et al., 2009). Thus evaluating a more elaborated set of stimulus words with a combination of these more detailed models may be instrumental in unlocking these various sources of semantic information encoded in the brain.

Conclusions

We know that cross-language neural similarity is meaningful because it permits decoding across languages, but whatever the source of these cross-speaker and cross-language regularities, it is not directly derived from speakers' intuitions about semantic relatedness nor from corpus statistics, nor are these sources of information irrelevant to neural representation of semantics since they do provide some explanatory power beyond the cross-language correlations. Alternate representational models and more representative stimulus words may yet provide better descriptions of cross-language semantics and their respective implementations in the brains of a language's speakers. Future research will test such models' ability to decode word-elicited concepts across languages and clarify this highly complex emerging picture.

Acknowledgements

We are grateful to Donias Doko and Carol Jew for data collection, Xixi Wang for language consulting, and Peiyao Chen, Anqi Li, Yinghui Qiu, and Tianyang Zhang for translation.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463-493.

Binder, J. R., & Desai, R. H. R. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–36. doi:10.1016/j.tics.2011.10.001

Buchweitz, A., Shinkareva, S. V, Mason, R., Mitchell, T. M., & Just, M. A. (2012). Identifying bilingual semantic neural representations across languages. *Brain and Language*, *120*(3), 282-9. doi:10.1016/j.bandl.2011.09.003

Charest, I., Kievit, R. a., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. doi:10.1073/pnas.1402594111

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *The Journal of Neuroscience*, *34*(1), 332–8. doi:10.1523/JNEUROSCI.1302-13.2014

Degani, T., & Tokowicz, N. (2010). Semantic ambiguity within and across languages: an integrative review. *Quarterly Journal of Experimental Psychology, 63*(7), 1266–303. doi:10.1080/17470210903377372.

Hauk, O., Johnsrude, I., & Pulvermu, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*. doi:10.1002/wcs.1251

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A, & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, *320*(5880), 1191–5. doi:10.1126/science.1152876

Raizada, R., & Connolly, A. (2012). What makes different people's representations alike: Neural similarity space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience*, *24*(4), 868–877.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, (eds), *WaCky! Working papers on the Web as Corpus.* Gedit, Bologna, http://wackybook.sslmit.unibo.it/

Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, *45*(1), 1–44. doi:10.1145/2379776.2379777

Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., & Raizada, R. (2015). You say potato, I say tudou: How speakers of different languages share the same concept. *Proceedings of the 37th annual conference of the Cognitive Science Society.*

Modeling developmental and linguistic relativity effects in color term acquisition

Barend Beekhuizen & Suzanne Stevenson
Department of Computer Science
University of Toronto
{barend,suzanne}@cs.toronto.edu

Abstract

We model two patterns related to the acquisition of color terms in Russian and English: children produce overextension errors for some colors but not others, and language-specific distinctions affect color discrimination in a non-linguistic task. Both effects, as well as a reasonable convergence with adult linguistic behavior, are shown by a Self-Organizing Map trained on naturalistic input. We investigate the effect of different ways of representing colors, i.e., as perceptual features or in terms of the cognitive biases on categorization extracted from crosslinguistic color naming data. We also consider the influence of color term frequency. Our results suggest effects of all three of term frequency, cognitive biases, and perceptual features.

Keywords: color terms, language acquisition, linguistic relativity, Typological Prevalence Hypothesis

Introduction

Languages vary in how they carve up lexical semantic domains; e.g., for situations where English uses the preposition *on*, Dutch uses *aan* and *op*. The crosslinguistic variation in lexical semantic divisions raises two interesting questions regarding the language user. First, are these various divisions all equally easy to learn, and if not, what drives the difference in ease of acquisition? Second, does acquiring a language-specific system affect other parts of cognition, a position known as 'linguistic relativity' (Gumperz & Levinson, 1996)? The two questions are related, as the acquisition of language-specific semantic divisions can be expected to go hand in hand with any extra-linguistic effects of acquiring such systems.

In this paper, we use a computational model to study both the acquisition of color terms and behavior on a non-verbal color discrimination task. Our goal is to propose a unified account of these two phenomena by simulating them within a single computational word-learning model. We explore what factors drive the two phenomena, considering both the features with which we represent color in the model and the varying frequencies of the color terms.

As there is an understanding of how color is represented on a perceptual level (e.g., Fairchild, 1998), we can use perceptual features as one representation of color. We also explore features motivated by the Typological Prevalence Hypothesis, which holds that crosslinguistically more common divisions are more cognitively accessible and thus easier to learn (Gentner & Bowerman, 2009). This exploration is motivated by positive results of this approach in the domain of spatial adpositions (Beekhuizen, Fazly, & Stevenson, 2014). Specifically, we use a representation based on the crosslinguistic biases in the divisions of the color space, derived from elicitation data (Kay, Berlin, Maffi, Merrifield, & Cook, 2009).

We find that both types of features yield a good fit to the developmental pattern of color term acquisition as well as to the behavioral pattern, with term frequency having an impact on some results. In some cases, we find that using the features together produces a better fit to human data, indicating that both perceptual properties and crosslinguistic biases may play a complementary role in learning a system of color terms.

Empirical findings and related work

Color term acquisition The development of children's use of color terms is slow, and across languages displays many errors where children use one color term where adults would use another ('overextension errors') (e.g. Harkness, 1973; Roberson, Davidoff, Davies, & Shapiro, 2004). In line with the Typological Prevalence Hypothesis, it has also been argued that the crosslinguistic patterns in color systems are reflected in the pattern of acquisition – i.e., children learn crosslinguistically rarer divisions later than more common ones (Dougherty, 1978). Here, we look at two languages for which detailed developmental color naming data is available, English (Bateman, 1915) and Russian (Davies, Corbett, McGurk, & MacDermid, 1998).

Color discrimination Winawer et al. (2007) (henceforth: W07) ask whether having two primary color terms (*sinij* 'dark blue' and *goluboj* 'light blue' in Russian) versus one (*blue* in English) affects non-linguistic color discrimination. They presented adult monolingual speakers of Russian and English with triplets of a stimulus color chip, an identical target chip and a different distracter. Participants were asked to decide which of the target and distracter was identical to the stimulus, and response latency was measured. On each trial, the distracter was either 'within' the same (participant-determined) category of dark or light blue as the stimulus, or 'across' the category boundary. Each distracter was also either 'near' or 'far' from the stimulus chip on the color scale.

Three conditions were explored: with a verbal dual task, a spatial dual task, or no interfering task. In the latter two conditions, Russian speakers picked the target faster when it was located 'across' the category boundary from the distracter, but only in the 'near' (harder discrimination) cases. (Participants in both languages are slower at picking the target chip when it is 'near' the distracter compared to when it is 'far.') The across-category advantage in Russian was not found under the verbal interference task, and English speakers showed no effect of category in any condition, indicating that the Russian category-advantage is a linguistic influence on lower-level processing.

Related modeling work To our knowledge, the only other attempt at modeling linguistic relativity is Colunga and Gasser (1998), who train a neural network on artificial languages and semantic domains to study both the effects of ease of acquisition and cognitive consequences of acquiring semantic divisions. Our model has a similar architecture and displays similar effects, but is trained on naturalistic data.

An earlier attempt at modelling color term acquisition is Belpaeme and Bleys (2005), who present a multi-agent model that represents color in an $L^*a^*b^*$ space (see below), although they do not focus on the developmental trajectories of learners or on behavioral linguistic relativity effects. Our approach can be considered as complementary, focusing on the cognition and behavior of an individual learner rather than on biases in the emergence of community-wide systems.

Beekhuizen and Stevenson (2015) used the Generalized Context Model (GCM; Nosofsky, 1987) to simulate the developmental English color naming data of Bateman (1915). While this approach showed interesting preliminary results, GCM is limited in its ability to acquire language-specific attention weighting. We require a model able to incrementally acquire and represent varying attentional weights over subintervals of the values of a dimension, possibly independently of values on other dimensions. The Russian 'blue's are a case where such representational potential is needed: attention to a part of the luminance scale is heightened, but only for blue hues. Self-Organizing Maps (SOMs; Kohonen, Schroeder, & Huang, 2001), explored for language acquisition by, e.g., Li and Zhao (2013), constitute a class of models that can capture such effects, while also having the potential to show developmental effects due to their incremental nature.

Our Computational Model

Self-Organizing Map

A Self-Organizing Map M is a neural network consisting of an $m \times n$ grid of neuron cells $[c_{11}, c_{12}, \ldots c_{mn}]$, where every cell consists of a vector of feature values. At every iteration i of training, an input stimulus s, with values for the same set of features, is compared to all cells $c \in M$, and is subsequently mapped to the cell to which it is most similar, called the Best Matching Unit (BMU) cell for s, or c_s. The values of c_s as well as its neighboring cells are then updated with the values of s. This way, M will come to display a topology that reflects the similarity among the input items.

Formally, $c_s = \arg\min_{c \in M} d_{\text{feat}}(c, s)$ where $d_{\text{feat}}(c, s)$ is the Euclidean distance between the feature values of c and s. All cells are updated in proportion to their map distance from c_s:

$$c_{jk}^{i+1} = c_{jk}^{i} + h_{jk}^{i} \times (s - c_{jk}^{i}) \quad (1)$$

$$h_{jk}^{i} = \alpha \cdot \exp\Big(-\frac{d_{\text{map}}(c_{jk}, c_s)}{2 \times {\sigma_i}^2}\Big) \quad (2)$$

That is, h_{jk}^{i} yields the excitation of the neuron cell c_{jk} given a center of activation at the coordinates of c_s, taking into account their distance in the map grid given by d_{map}. Here $\alpha = [0, 1]$ is a learning rate parameter, and σ_i the neighborhood radius of c_s, given by the exponential function $\sigma_i = \sigma_0 \times \exp(-\frac{i}{\lambda_\sigma})$; σ_0 and λ_σ are constants defining the intercept and slope of the function yielding the neighborhood radius. To observe developmental effects, slow learning is needed, and therefore we set $\alpha = .05$, $\sigma_0 = 1$, $\lambda_\sigma = 2000$, and train 8×8 maps.

Feature Representations

We formulate acquisition of color vocabulary as a categorization task that associates a color term (category label) with a color stimulus (a set of color property features). An input item consists of a representation of the properties of a Munsell color chip (a property-feature vector) paired with a color term (a term-feature vector). Each cell of the SOM represents a learned association between a set of property-feature values and a distribution over the terms in the term-feature vector.

The term-feature vector has length $|T|$, where T is the set of primary color terms in a language. To represent term t_i in an input item, the ith feature is set to a value a in $[0, 1]$, and all others set to 0; e.g., in a system with 4 terms, input $t_2 = [0, a, 0, 0]$. The parameter a (in our experiments set to .2) reflects the relative importance of term features in training. The term-feature vector of each cell of the SOM will come to hold a distribution over terms, which we normalize to arrive at a probability of a term for a cell, $P(t|c)$ (see below).

The property-feature vector represents the set of stimuli of Munsell color chips, S, in one of two forms. First, we test the idea that the cross-linguistic tendencies in the semantic distinctions are telling of the extra- or pre-linguistic cognitive biases of language learners (cf. the Typological Prevalence Hypothesis). As in Beekhuizen and Stevenson (2015), we operationalize this idea with Principal Component Analysis (PCA) over the World Color Survey data (Kay et al., 2009), which contains color terms for 330 Munsell color chips in 110 languages. The closeness of a pair of chips in the resulting space reflects the frequency with which they are labeled with the same term, and the space thereby represents the crosslinguistic tendencies to group chips under a particular term. (More details can be found in Beekhuizen & Stevenson, 2015.) If the extension of a color term – i.e., the set of chips labeled with that term – is spread widely over the PCA space, it is assumed to be harder to learn than if a set of the same size were spread less widely over the PCA space. We refer to property features based on the PCA components as the conceptual, or `conc`, features.

We also can represent the various color chips at a purely perceptual level. We use the coordinates of the chips in $L^*a^*b^*$ space, which is thought to encode the perceptual dissimilarity between colors (Fairchild, 1998). This feature set will be referred to as the perceptual, or `perc`, features.

Both the property-feature spaces were normalized such that the mean for each feature is .5 and the values are in $[0, 1]$. SOMs are initialized with values of 0 for term features and .5± a very small random value for property features.

Sampling for training data

Input items are sampled as a pair of a color term $t \in T$ and a stimulus color chip $s \in S$ from the distribution $P(t,s) = P(s|t)P(t)$. We obtain the conditional probability distributions for $P(s|t)$ from adult elicitation data in English (Berlin & Kay, 1969) and in Russian (Davies & Corbett, 1994).[1] (For the latter data in Yxy coordinates, we convert those coordinates into $L^*a^*b^*$, and identify the Munsell chip with the closest $L^*a^*b^*$ value.)

As one estimation of $P(t)$, we used the relative term frequency over all color terms. For English, these were taken from the child-directed speech portion of the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) of CHILDES (MacWhinney, 2000). For Russian, lacking a corpus of child-directed speech of suitable size, we use the relative term frequencies reported in Vamling (1986).[2] We also assess sampling according to a uniform distribution for $P(t)$. These two conditions are called `corpus` and `uniform`.

Experimental Methods

For all experiments, we run 30 simulations for each of the six combinations of `features={perc,conc,perc+conc}` and `sampling={corpus,uniform}`. At every test moment (every 100 input items), we present the model with an unlabelled color chip s (i.e., a property-feature vector with no term features) and extract the most probable term that the model associates with those property features. We obtain the Best Matching Unit for s as $c_s = \arg\min_{c \in M} d_{\text{feat}}(c,s)$, where only the property features of c are compared to those of s. The model response for s, term t_s, is extracted from the probability distribution over the terms T for c_s:

$$t_s = \arg\max_{t \in T} P(t|c_s) \tag{3}$$

$$P(t|c_s) = \frac{\text{value}(t,c_s)}{\sum_{t' \in T} \text{value}(t',c_s)} \tag{4}$$

where $\text{value}(t,c_s)$ is the value for feature t in cell c_s.

Evaluating linguistic convergence

To evaluate whether the model obtains an adult level of understanding of the color terms, we test it with color stimuli corresponding to the complete set of color chips S_{adult} for which we have adult responses ($|S_{\text{adult}}| = 49, |T| = 12$ for Russian; $|S_{\text{adult}}| = 211, |T| = 11$ for English). Model convergence with adult linguistic behavior is then given by:

$$\textbf{score}_{\textbf{C}} = \frac{|S_{\text{correct}}|}{|S_{\text{adult}}|} \tag{5}$$

at each test moment, where S_{correct} is the set of test stimuli for which $t_s = t_{\text{correct}}$, and t_{correct} is the modal adult response for the given chip. To avoid accidental local optima, we average $\textbf{score}_{\textbf{C}}$ over the most recent 20 test moments. We let the model run until it ceases to improve $\textbf{score}_{\textbf{C}}$ for 10K inputs.

Evaluating linguistic development

In the child color naming data, several types of patterns are observed: For some color stimuli, children produce hardly any or no errors, whereas for others, overextensions are observed, sometimes even more frequently than the correct term. Our goal is to assess the fit between the model's distribution over terms, $P(t|c_s)$ (Eqn. 4), for each stimulus s at various points in learning, and the relative dominance of terms exhibited by children at various points in development.

To that end, we compare the ranking of terms based on $P(t|c_s)$ to a ranking derived from child elicitation data (ranked by the number of children producing an error for that color in Bateman, 1915 and Davies et al., 1998). For every color stimulus presented to children from n age groups, we find the n consecutive, equal-sized bins of test moments for which the predicted ranking for that stimulus matches optimally the observed ranking of each age group for that color.[3] Each bin contains at least 5 test moments, to avoid finding unrealistically narrow 'age groups' in the model data. The model ranking of terms is given by $P(t|c_s)$ averaged over all test moments in that bin, across 30 simulations. The low values in this pooled probability distribution ($P(t|c_s) < .05$) are rounded down to 0 to avoid diluting the ranking metric with insignificant predictions; similarly, we consider only errors occurring a minimum of 3 times in the child data. The two – model and observed – rankings are then compared using Kendall's τ_b, which we use as our evaluation measure.

Evaluating color discrimination

We take the final state of the SOM to correspond to adult organization of the color terms. Reflecting the hypothesis that linguistic knowledge affects the extra-linguistic task of color discrimination, we take the closeness between the BMUs of two stimuli in our learned SOM to correspond to the degree of difficulty people show in discriminating them. We convert the 20 stimuli of W07 into our representation of color properties, yielding the vector $S_{\text{disc}} = [s_1, \ldots, s_{20}]$. Following W07, we consider two stimuli s_i and s_j to be 'near' if $j = i+2$, and 'far' if $j = i+4$. To find the perceived distance between the target and distracter, $s_{\text{t}}, s_{\text{d}} \in S_{\text{disc}}$, we take their SOM distance $d_{\text{map}}(c_{s_t}, c_{s_d})$ (as defined above). The greater the distance, the easier to discriminate the target from the distracter.

We find the category boundary in the model by having it predict the most likely Russian term per stimulus in S_{disc}, and placing the boundary between the last light blue (*goluboj*)

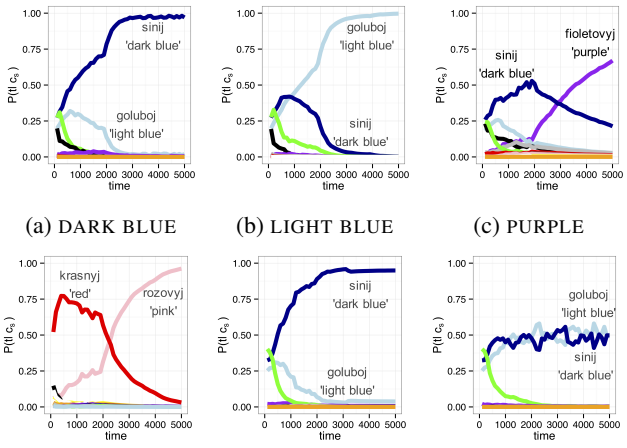
[1] This formulation of $P(s|t)$ is informative about the mapping of terms to colors: a chip s_1 labeled half the time as *blue* is less likely to be sampled for the term *blue* than a chip s_2 labeled 100% of the time as *blue*. However, $P(s|t)$ says nothing about how frequently the colors are discussed with that label: if s_1 is more frequent in the world, usages of *blue* may refer to it more than to s_2. At this point we know of no way to estimate a sampling of colors people refer to.

[2] The different sources of frequency data may differentially affect outcomes in the two languages, an issue for future research.

[3] Since the actual frequencies of colors and their co-occurrence with terms is unknown (see footnote 1), and such patterns will certainly affect learning, we think it overly strict to require the model to align all test stimuli in parallel. Future work will explore more directly the impact of color frequencies on our model's results.

	Russian		English	
	corpus	uniform	corpus	uniform
perc	.84 (.04)	.89 (.03)	.91 (.03)	.93 (.02)
conc	.86 (.03)	.87 (.02)	.92 (.02)	.93 (.02)
perc+conc	.89 (.03)	.92 (.03)	.95 (.02)	.97 (.01)

Table 1: Results for convergence: mean and standard deviation of $\textbf{score}_\textbf{C}$ (Eqn. 5), over 30 simulations.



(a) DARK BLUE (b) LIGHT BLUE (c) PURPLE

(d) PINK (e) DARK BLUE (f) LIGHT BLUE

Figure 1: $P(t|s)$ over time for example stimuli in Russian for `(perc+conc,corpus)` (a-d) and `(conc,corpus)` (e-f).

response and the first dark blue (*sinij*) response.[4] English does not have these distinct terms, but the observed category boundaries for Russian and English hardly differ according to W07. We thus use as the English boundary the mean location of the Russian category boundary under the given combination of `features` × `sampling`. A target–distracter pair, s_t–s_d, is considered 'within'-category if s_t and s_d are on the same side of the boundary, and 'across'-category otherwise. Analogously to W07, the map distances for the 8 'near' and 8 'far' pairs closest to the category boundary were calculated from the model for all simulations.

We evaluate whether the model's behavior corresponds to human behavior in W07 by seeing if the same significant effects are found: we compare the $d_{\text{map}}(c_{s_t}, c_{s_d})$ values (using t-tests) between near and far cases, and between within- and across-category cases, and see whether these two interact.

Results: convergence and development

The model reaches its closest fit to adult behavior after some 20K (Russian) or 30K (English) input items. Table 1 shows that the model captures adult behavior well; a naive baseline always guessing the most frequent term would reach a $\textbf{score}_\textbf{C}$ of .20 (English) or .22 (Russian). We find that `uniform` sampling achieves slightly closer to adult naming behavior. With

[4]We discard 4% of Russian simulations which did not have a sequence of only *goluboj* followed by only *sinij*.

	Russian		English	
	corpus	uniform	corpus	uniform
perc	.91	.86	.96	.95
conc	.91	.89	.91	.90
perc+conc	.90	.89	.98	.96
error-free learner	.81	.81	.95	.95

Table 2: Results for development (mean τ_b over stimuli and age groups).

`sampling=corpus`, more frequent terms take up more of the SOM, leaving less space for less frequent terms to capture their full extension (and hence they are often mislabeled by more frequent neighboring terms). One area for further exploration is whether learners have to be relatively immune to frequency when processing color terms, as otherwise less frequent color terms may not be strongly represented.

Furthermore, we find that there is little difference in $\textbf{score}_\textbf{C}$ in either language between the `perc` and `conc` features alone. However, the model performs somewhat better with both used together. This suggest that the cross-linguistic conceptual space and the perceptual features are complementary, contributing somewhat different information to learning.

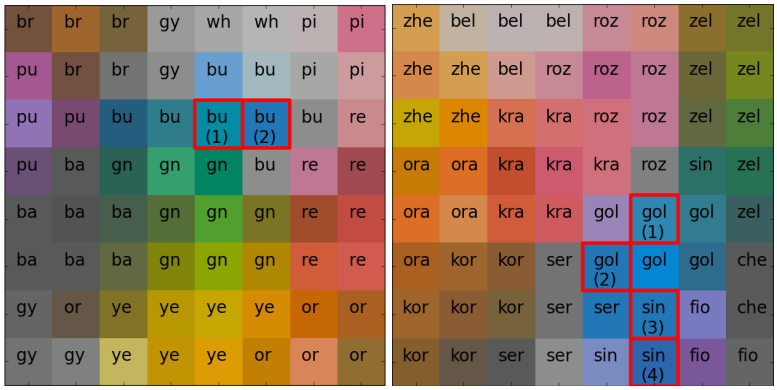
Turning to the development results next (Table 2), we find that the model has a good fit to observed patterns of developmental behavior. An error-free learner – one always predicting the correct term with a probability of 1 – has average τ_b values of .81 (Russian) and .95 (English), and so a large part of the global score comes from the model correctly simulating adult behavior rather than the overextension patterns. In most cases, however, the model surpasses these scores, indicating that it does capture some overextension patterns, especially in Russian, which has many more such errors.

For English, the two `perc` settings give a better fit than the `conc` settings. Considering the match with children's development on particular colors helps understand why. For English, Bateman (1915) presented children (age 6–12) with 8 color chips. The model only displays the correct overextensions of *blue* to PURPLE with `(perc+conc, corpus)` and `(perc, corpus)`, and fails to simulate the correct pattern for ORANGE in both `conc` settings. The other color terms were learned with the correct developmental pattern under all settings: For BLACK, WHITE, RED, and BLUE, no or hardly any overextensions were found either in children or in the model, and the few observed overextensions for YELLOW and GREEN were predicted in any parameter setting.

For Russian, we observe a difference between `corpus` and `uniform` sampling. Davies et al. (1998) presented 3- to 5-year-olds with 12 color chips. 3-year-olds label LIGHT BLUE and PURPLE more frequently *sinij* 'dark blue' than the correct terms, but do not label DARK BLUE as *goluboj* 'light blue' or *fioletovyj* 'purple' as often. Furthermore, 3-year-olds more frequently use *krasnyj* 'red' than *rozovyj* 'pink' for PINK. The model predicts these effects completely under

	W07	model
Russian	near > far	✓
	within > across	✓
	near-within > near-across	✓
	far-within ≈ far-across	✗
English	near > far	✓
	within ≈ across	✓
	near-within ≈ near-across	✓
	far-within ≈ far-across	✓

(a) Model match to W07; ✓/✗ indicate a match/non-match, respectively, across all 6 feature/sampling settings for Russian, and for 5 of 6 settings in English.



(b) Converged maps for English (left) and Russian (right). S_{disc} stimuli are mapped to the highlighted cells. Cell labels are taken from the first few letters of the most likely term for the cell. bu=*blue*; gol=*goluboj*, sin=*sinij*.

Figure 2: Results for modeling the discrimination experiment.

`(perc+conc, corpus)` (Fig. 1a-d), and generally predicts the observed ranking better when using `corpus`, suggesting that term frequencies explain some of the error pattern.

Feature sets display a subtle effect for Russian as well. The asymmetry for DARK BLUE and LIGHT BLUE disappears when we train on `perc`, as DARK BLUE and LIGHT BLUE are (too) easily discriminated in the perceptual space. This is different for the conceptual features: as many languages group DARK BLUE and LIGHT BLUE under one term, the inferred cognitive bias is to group them together. Figures 1e-f show that for `conc`, the asymmetry is present, but *goluboj* never gets fully learned. A combination of both feature sets thus seems necessary to understand this effect: perceptual dissimilarity is needed to discriminate them, but cognitive biases bias the learner against forming two categories. The asymmetry may then emerge because of the slightly higher term frequency of *sinij* (.08) over *goluboj* (.06).

Results: discrimination

Figure 2a summarizes the findings of W07 in their color discrimination task (1st column), along with an indication of whether the model results match those findings (2nd column). For example, the entry for Russian of "near-within > near-across" means that people found the near-within cases harder to discriminate than the far-within cases (a statistically-significant difference); "far-within ≈ far-across" means the difference between those two cases for people was not statistically significant. For the former, our model also found a statistically significant effect in the same direction, and for the latter, the lack of an effect in the same direction.[5]

The fact that the model matches "near > far" for both languages supports our assumption that map distance in the learned SOM is a good proxy for discriminability of stimuli. Importantly, the model matches the main finding of W07 that distracters in a different category from the target are more easily discriminated than distracters in the same category, for Russian but not for English (the "within > across" and "within ≈ across" rows in Figure 2a).

To illustrate why this happens, Figure 2b presents a typical converged map for English and for Russian. For English, chips s_1:s_4 are mapped to the cell marked with (2), and chips s_5:s_{20} to the cell marked with (1). Because the category boundary is placed between s_{11} and s_{12} of S_{disc}, all pairs of targets and distracters are mapped to the same cell (1), whether across-category *or* within, and such pairs are indiscriminable for the learner. For Russian, the different shades of blue cannot be compressed on the SOM as much as in English, because there are two terms that need to be discriminated: English *blue* is the most likely term in 5 cells, whereas Russian *sinij* and *goluboj* combined are the most likely terms in 10 cells. Thus in Russian, we see that the 20 S_{disc} stimuli are mapped to a larger part of the SOM (cells (1)–(4) in that map) than the English stimuli, and distances across the categories – from cells (1)-(2) to (3)-(4) – are further than within categories (within (1)-(2) or within (3)-(4)).

Finally, the model generally fails to predict the empirical interaction whereby Russian displays a significant within-across difference for near but not far cases. Under all settings, the model predicts both differences to be significant. We do find a trend in the right direction: for all settings, the within-across difference is greater in the model for the near cases than for the far cases.

Discussion

In this paper, we looked at the developmental pathway of color term acquisition and the effects of acquiring the color term system of a particular language on a non-verbal discrimination task. A Self-Organizing Map (SOM) trained on naturalistic input models three effects: (1) some patterns of overextension errors in linguistic development and (2) subsequent convergence in Russian and English, as well as (3) a higher ability to discriminate light blue from dark blue stimuli in Russian, but not English. Our model thus provides a

[5] Recall that a closer SOM distance, $d_{\text{map}}(c_{s_t}, c_{s_d})$, means the target and distracter stimuli s_t, s_d are "harder to discriminate".

mechanistic conception of learning that gives a unified explanation of both linguistic development and linguistic relativity. The idea that between-language variation is represented by the varying amount of information compression on the SOM (due to the different patterns of words with stimuli across languages) gives us an explanatory principle that could be applied to domains beyond color.

We asked whether possible cognitive biases inferred from crosslinguistic categorization tendencies (cf. Gentner & Bowerman, 2009, reflected in our 'conceptual features', play a role, or whether perceptual features of color best explain the effects. Both feature sets contribute to the explanation of linguistic development: in some cases (naming PURPLE in English), the error pattern is predicted only when the perceptual features are present. For others, leaving out the conceptual features hurts the fit with the observed data (naming DARK BLUE in Russian), suggesting that these biases do play a role.

We also investigated frequency effects: The model fails to predict common overextension patterns in both languages when not taking term frequency into account. Nonetheless, sampling on the basis of corpus frequencies makes the model converge less well to adult behavior for infrequent terms, suggesting that, over development, learners may need to be decreasingly sensitive to term frequency.

One issue we did not explore is different initializations of the SOMs. As children experience color prior to acquiring terms for them, it is possible that the map is already 'pre-organized' by exposure to color stimuli without associated color terms. We plan on studying further whether such pre-linguistic exposure affects the developmental patterns.

Finally, we looked at the converged states of the SOMs in predicting color discrimination behavior across languages, finding a weak preference for models trained on perceptual features. Since we are able to track the development of the SOM, we can also investigate the effect of language-specific lexical semantic systems on extra-linguistic behavior over developmental time (see, e.g., McDonough, Choi, & Mandler, 2003, for such developmental effects in another domain). In the future, we plan to explore suitable semantic domains for evaluating how well our model simulates linguistic relativity effects over the course of acquisition.

Acknowledgements

We acknowledge the support of NSERC of Canada and thank the four CogSci reviewers for helpful suggestions.

References

Bateman, W. G. (1915). The naming of colors by children: The Binet test. *The Pedagogical Seminary*, *22*(4), 469–86.

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning Meaning without Primitives: Typology Predicts Developmental Patterns. In *Proceedings CogSci.*

Beekhuizen, B., & Stevenson, S. (2015). Perceptual, conceptual, and frequency effects on error patterns in english color term acquisition. In *Proceedings CogACLL.*

Belpaeme, T., & Bleys, J. (2005). Explaining universal colour categories through a constrained acquisition process. *Adaptive Behavior*, *13*(4), 293–310.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their universality and evolution*. Berkeley, CA: UC Press.

Colunga, E., & Gasser, M. (1998). Linguistic relativity and word acquisition: a computational approach. In *Proceedings CogSci.*

Davies, I., & Corbett, G. (1994). The basic color terms of Russian. *Linguistics*, *32*, 65–89.

Davies, I., Corbett, G., McGurk, H., & MacDermid, C. (1998). A developmental study of the acquisition of Russian colour terms. *J. Child Lang.*, *25*, 395–417.

Dougherty, J. (1978). On the significance of a sequence in the acquisition of basic color terms. In B. Blount & M. Sanches (Eds.), *Sociocultural Dimensions of Language Change* (pp. 133–48).

Fairchild, M. (1998). *Color Appearance Models*. Reading, MA: Addison-Wesley.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others. The Typological Prevalence Hypothesis. In J. Guo et al. (Ed.), *Crosslinguistic Approaches to the Psychology of Language. Research in the Tradition of Dan Isaac Slobin* (pp. 465–80). New York, NY: Psychology Press.

Gumperz, J., & Levinson, S. (Eds.). (1996). *Rethinking Linguistic Relativity*. London, UK: CUP.

Harkness, S. (1973). Universal aspects of learning color codes: A study in two cultures. *Ethos*, 175–200.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *World Color Survey*. Stanford, CA: CSLI.

Kohonen, T., Schroeder, M. R., & Huang, T. S. (Eds.). (2001). *Self-Organizing Maps* (3rd ed.). Springer.

Li, P., & Zhao, X. (2013). Self-organizing map models of language acquisition. *Frontiers in Psychology*, *4*(828).

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Erlbaum.

McDonough, L., Choi, S., & Mandler, J. (2003). Understanding spatial relations: Flexible infants, lexical adults. *Cognitive Psychology*, *46*(3), 229-59.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *J. Exp. Psychol.*, *13*(1), 87–108.

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2004). The development of color categories in two languages: a longitudinal study. *J. Exp. Psychol. Gen.*, *133*(4), 554–71.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *J. Child Lang.*, 127–152.

Vamling, K. (1986). A note on Russian blues. *Slavica Lundensia*, *10*, 225–233.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *PNAS*, *104*(19), 7780–5.

Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions

Caroline Graf, Judith Degen, Robert X.D. Hawkins, Noah D. Goodman
cgraf@uos.de, {jdegen,rxdh,ngoodman}@stanford.edu
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Nominal reference is very flexible—the same object may be called *a dalmatian*, *a dog*, or *an animal* when all are literally true. What accounts for the choices that speakers make in how they refer to objects? The addition of modifiers (e.g. *big dog*) has been extensively explored in the literature, but fewer studies have explored the choice of noun, including its level of abstraction. We collected freely produced referring expressions in a multi-player reference game experiment, where we manipulated the object's context. We find that utterance choice is affected by the contextual informativeness of a description, its length and frequency, and the typicality of the object for that description. Finally, we show how these factors naturally enter into a formal model of production within the Rational Speech-Acts framework, and that the resulting model predicts our quantitative production data. **Keywords:** referential expressions, levels of reference, basic level, experimental pragmatics, computational pragmatics

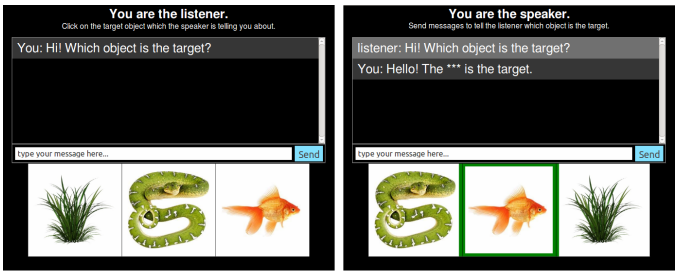


Figure 1: Screenshots from speakers' and listeners' points of view, showing role names and short task descriptions, the chatbox used for communication and a display of three pictures of objects. The referent was identified to the speaker by a green box.

Referring to objects is a core function of human language, and a wealth of research has explored how speakers choose referring expressions (Herrmann & Deutsch, 1976; Pechmann, 1989; van Deemter, Gatt, van Gompel, & Krahmer, 2012). However, most of this literature has focused on the addition of modifiers (as in the choice between "the dog", "the brown dog", and "the big brown dog", e.g., Sedivy, 2003; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Here we investigate how speakers choose a simple nominal referring expression—what governs the choice of calling a particular object "the dalmatian", "the dog", or "the animal" when all are literally true? That is, what governs the choice of the taxonomic level at which an object is referred to? Noun choice can be seen as the most basic decision in forming a referring expression. Like modification, these choices differ in their specificity; unlike modification, the number of words used does not differ—in English, *some* noun must be chosen. In this paper we provide experimental evidence from a coordination game regarding the flexible choice of nominal referring expressions and explain this data with a probabilistic model of pragmatic production.

Previous evidence about the generation of referring expressions suggests that choice of reference level will depend on the interplay of several factors. Grice's Maxim of Quantity (Grice, 1975) implies a pressure for speakers to be sufficiently *informative*. For instance, a speaker who is trying to distinguish a dalmatian from a German Shepherd would be expected to avoid the insufficiently specific term "dog" (Brennan & Clark, 1996). On the other hand, recent work in experimental pragmatics has shown that the choice of referring expression depends on the *cost* of utterance alternatives (Rohde, Seyfarth, Clark, Jäger, & Kaufmann, 2012; Degen, Franke, & Jäger, 2013); sometimes, speakers are willing to produce a cheap ambiguous utterance rather than a costly (e.g. long or difficult-to-retrieve) unambiguous one. Finally, classic work on concepts suggests that *typicality* of a referent within its category affects the choice of reference (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). In particular, speakers will generally choose to refer at the *basic level* (e.g. "dog"), but may become more specific for objects that are atypical for the basic level term.

To evaluate the impact of these factors on nominal reference we constructed a two-player online game (Fig. 1). Participants saw a shared context of objects, one of which was indicated as the referent only to the speaker. The speaker was asked to communicate this object to the listener, who then chose among the objects. Critically, the speaker and listener communicated by free use of a chat window, allowing us to gather relatively natural referring expressions. We manipulated the category of distractor objects and used items that varied in utterance complexity and object typicality. This allowed us to evaluate whether each factor influences the referring expressions generated by participants. We expect that speakers will (1) tend to avoid longer or less frequent terms, and (2) will pragmatically prefer more specific referring expressions when the target and distractor(s) belong to the same higher-level taxonomic category or when distractors are more typical members of that category level.

A promising modeling approach for capturing the quantitative details of human language use is the Rational Speech-Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). The RSA framework has been applied to many language interpretation tasks (e.g. Goodman & Stuhlmüller, 2013; Kao, Wu, Bergen, & Goodman, 2014), but relatively rarely to production data (but see Franke,

2014; Orita, Vornov, Feldman, & Daumé III, 2015). We describe an RSA model of nominal reference that includes informativeness, cost, and typicality effects. A speaker in RSA is treated as an approximately optimal decision maker who chooses which utterance to use to communicate to a listener. The speaker has a utility which includes terms for the cost of producing an utterance (in terms of length or frequency) and the informativeness of the utterance for a listener. The listener is treated as a literal Bayesian interpreter who updates her beliefs given the truth of the utterance. These truth values are usually treated as deterministic (an object either is a "dog" or it is not); here we relax this formulation in order to incorporate typicality effects. That is, we elicit typicality ratings in a separate experiment, and model the listener as updating her beliefs by weighting the possible referents according to how typical each is for the description used. We evaluate the quantitative model predictions against our production data. The model also allows us to evaluate the need for each extra component—typicality, length, frequency—and determine whether the empirical bias toward reference at the basic level (Rosch et al., 1976) can be accounted for without building it in as a separate factor.

Experiment: nominal reference game

Methods

Participants and materials We recruited 56 self-reported native speakers of English over Mechanical Turk. Participants completed the experiment in pairs of two, yielding 28 speaker-listener pairs.

Stimuli were selected from nine distinct domains, each corresponding to distinct basic level categories such as "dog." For each domain, we selected four subcategories to form our target set (e.g. "dalmatian", "pug", "German Shepherd" and "husky"). Each domain also contained an additional item which belonged to the same basic level category as the target (e.g. "greyhound") and items which belonged to the same supercategory but not the same basic level (e.g. "elephant" or "squirrel"). The latter items were used as distractors.

Each trial consisted of a display of three images, one of which was designated as the target object. Every pair of participants saw every target exactly once, for a total of 36 trials per pair. These target items were randomly assigned distractor items which were selected from four different context conditions, corresponding to different communicative pressures (see Fig. 2). We refer to these conditions with pairs of numerals specifying which levels of the taxonomy are present in the distractors: (a) **item12**: one distractor of the same basic level and one distractor of the same superlevel (e.g. target: "dalmatian", distractor 1: "greyhound", distractor 2: "squirrel"), (b) **item22**: two distractors of the same superlevel, (c) **item23**: one distractor of the same superlevel and one unrelated item and (d) **item33**: two unrelated items.

Furthermore, the experiment contained 36 filler items, in which participants were asked to produce referential expressions for objects which differed only in size and color. Images

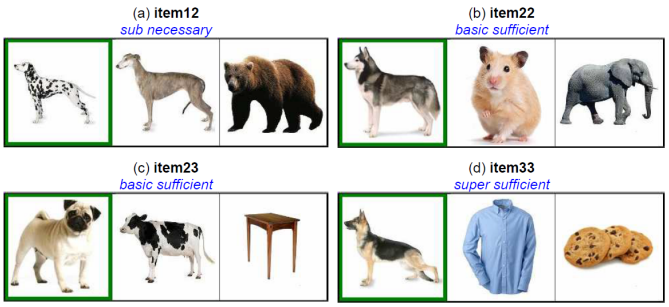


Figure 2: The four context conditions, exemplified by the *dog* domain. The target is outlined in green; the types of distractors differ with condition (see text).

from filler trials were not reused on target trials. Trial order was randomized.

Procedure Pairs of participants were connected through a real-time multi-player interface (Hawkins, 2015), with one member of each pair assigned the speaker role and the other to the listener role. Participants kept their allotted roles for the entire experiment. The setup for both the speaker and the listener is shown in Fig. 1. Each saw the same set of three images, but positions were randomized to rule out trivial position-based references like "the middle one." The target object was identified by a green square surrounding it for the speaker (but not listener). Players used a chatbox to send text messages to each other. The task was for the speaker to get the listener to select the target object.

Annotation To determine the level of reference for each trial, we followed the following procedure. First, trials on which the listener selected the wrong referent were excluded, leading to the elimination of 1.2% of trials. Then, speakers' and listeners' messages were parsed automatically; the referential expression used by the speaker was extracted for each trial and checked for whether it contained the current target's correct sub, basic or super level term using a simple grep search. In this way, 66.2% of trials were labelled as mentioning a pre-coded level of reference. In the next step, remaining utterances were checked manually to determine whether they contained a correct level of reference term which was not detected by the parsing algorithm due to typos or grammatical modification of the expression. In this way, meaning-equivalent alternatives such as "doggie" for "dog", or contractions such as "gummi","gummies" and "bears" for "gummy bears" were counted as containing a level of reference term. This caught another 13.8% of trials. A total of 20.0% of correct trials were excluded because the utterance consisted only of an *attribute* of the superclass ("the living thing" for "animal"), of the basic level ("can fly" for "bird"), of the subcategory ("barks" for "dog") or of the particular instance ("the thing facing left") rather than a category noun. These kinds of attributes were also sometimes

mentioned in addition to the noun in the trials which were included in the analysis—4.0% of sub level terms, 12.6% of basic level terms, and 46.2% of super level terms contained an additional modifier. On 0.5% of trials two different levels of reference were mentioned; in this case the more specific level of reference was counted as being mentioned in this trial.

Typicality norms To examine the influence of typicality on speaker behavior, we obtained typicality estimates in a separate norming study. 240 participants were recruited through Mechanical Turk. On each trial, we presented participants with an image from the main experiment and asked them "How typical is this for X?", where X was a category label at the sub-, basic-, or super- level. They then adjusted a slider bar ranging from *not at all typical* to *very typical*.

Due to the large number of possible combinations of objects, we only collected norms for certain combinations of objects and descriptions: for each target (e.g., dalmatian), we collected typicality at all three levels ("dalmatian," "dog," and "animal"). For each distractor of the same superclass as the target (*distsamesuper*, e.g., a kitten), we collected typicality at all three levels of the *target*. For each distractor of a different superclass (*distdiffsuper*, e.g., a basketball) we only collected typicality at the super- level of the target ("animal") and assumed lowest typicality at the other levels. This resulted in the following distribution of 745 norms: *target-sub* (36), *target-basic* (36), *target-super* (36), *distdiffsuper-super* (168), *distsamesuper-sub* (331), *distsamesuper-basic* (93), and *distsamesuper-super* (45).

Each participant provided typicality ratings for 7 *target*, 10 *distdiffsuper*, and 28 *distsamesuper* cases (randomly sampled from the total set of items). Each case received between 6 and 27 ratings. Raw slider values ranged from 0 (not typical) to 1 (very typical); average slider values were used as the typicality values throughout our results.

Results

Proportions of sub, basic, and super level utterance choices in the different context conditions are shown in the top row of Fig. 3. The sub level term was preferred where it was necessary for unambiguous referent identification, i.e., when a distractor of the same basic level category as the target was present in the scene (item12, e.g. target: dalmatian, distractor: greyhound). Where it was not necessary (i.e., when there was no other object of the same basic level category present, as in conditions item22, item23 and item33), there was a clear preference for the basic level term. The super level term was strongly dispreferred overall, though it was used on some trials, especially where informativeness constraints on utterance choice were weakest (item33).

To test for the independent effects of informativeness, length, frequency, and typicality on sub-level mention, we conducted a mixed effects logistic regression. Frequency was coded as the difference between the sub and the basic level's log frequency, as extracted from the Google Books Ngram

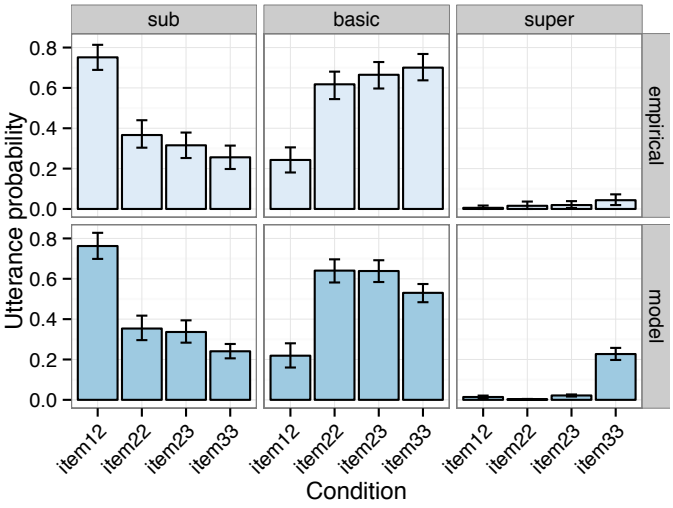


Figure 3: Empirical utterance probabilities (top row) and model posterior predictive MAP estimates (bottom row) by condition, collapsed across targets and domains. Error bars indicate bootstrapped 95% confidence intervals.

English corpus ranging from 1960 to 2008. Length was coded as the ratio of the sub to the basic level's length.[1] That is, a higher frequency difference indicates a *lower* cost for the sub level term compared to the basic level, while a higher length ratio reflects a *higher* cost for the sub level term compared to the basic level.[2] Typicality was coded as the ratio of the target's sub to basic level label typicality. That is, the higher the ratio, the more typical the object was for the sub level label compared to the basic level. For instance, the panda was relatively atypical for its basic level "bear" (mean rating 0.75) compared to the sub level term "panda bear" (mean rating 0.98), which resulted in a relatively *high* typicality ratio.

Condition was coded as a three-level factor: *sub necessary*, *basic sufficient*, and *super sufficient*, where item22 and item23 were collapsed into *basic sufficient*. Condition was Helmert-coded: two contrasts over the three condition levels were included in the model, comparing each level against the mean of the remaining levels (in order: *sub necessary*, *basic sufficient*, *super sufficient*). This allowed us to determine whether the probability of type mention for neighboring conditions were significantly different from each other, as suggested by Fig. 3.[3] The model included random by-speaker and by-domain intercepts.

A summary of results is shown in Table 1. The log odds

[1] We used the mean empirical lengths in characters of the utterances participants produced. For example, the minivan, when referred to at the subcategory level, was sometimes called "minivan" and sometimes "van" leading to a mean empirical length of 5.64. This is the value that was used, rather than 7, the length of "minivan".

[2] We replicate the well-documented negative correlation between length and log frequency ($r = -.53$ in our dataset).

[3] Adding terms that code the ratio of the sub vs super level frequency and length did not lead to an improvement of model fit.

Table 1: Mixed effects model summary.

	Coef β	SE(β)	p
Intercept	−0.30	0.35	>0.4
Condition sub.vs.rest	2.46	0.24	<**.0001**
Condition basic.vs.super	0.52	0.23	<**.05**
Length	−0.52	0.14	<**.001**
Frequency	−0.02	0.08	>0.78
Typicality	4.17	0.84	<**.0001**
Length:Frequency	−0.30	0.11	<**.01**

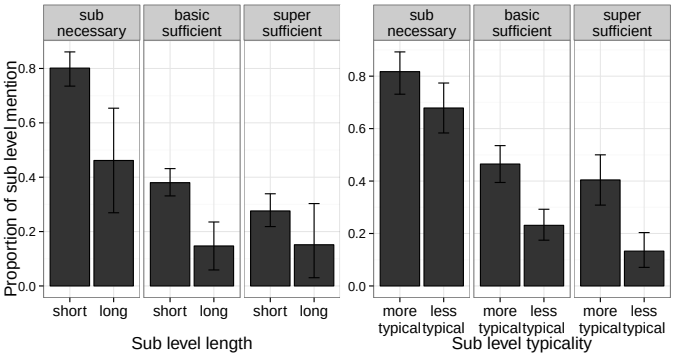


Figure 4: Probability of using sub, basic and super level terms. Left: when the sub length is relatively short (.67,2] or long [2,4.67) compared to the basic level term length. Right: when the target object was relatively more [1.06,1.91) or less (.88,1.06] typical for the sub compared to the basic level term.

of mentioning the sub level term was greater in the *sub necessary* condition than in either of the other two conditions, and greater in the *basic sufficient* condition than in the *super sufficient* condition, suggesting that the contextual informativeness of the sub level mention has a gradient effect on utterance choice.[4] There was also a main effect of typicality, such that the sub level term was preferred for objects that were more typical for the sub level compared to the basic level description (Fig. 4). In addition, there was a main effect of length, such that as the length of the sub level term increased compared to the basic level term ("chihuahua"/"dog" vs. "pug"/"dog"), the sub level term was dispreferred ("chihuahua" is dispreferred compared to "pug", Fig. 4). Finally, while there was no main effect of frequency, we observed a significant length by frequency interaction, such that there was a frequency effect for the relatively shorter but not the relatively longer sub level cases: for shorter sub level terms, relatively high-frequency sub level terms were more likely to be used than relatively low-frequency sub level terms.

Unsurprisingly, there was also significant by-participant and by-domain variation in the log odds of sub level term mention. For instance, mentioning the subclass over the basic level term was preferred more in some domains (e.g. in the "candy" domain) than in others. Likewise, some domains had a greater preference for basic level terms (e.g. the "shirt" domain). Using the superclass term also ranged from hardly being observable (e.g. the "flower" domain) to being used more frequently (e.g. in the "bird" domain). Nevertheless, mentioning the sub level term was always the most frequent choice where a distractor of the same basic level was displayed. Furthermore, it was the case in all domains that the sub level term was mentioned most frequently and the basic level least frequently in just this condition, compared to the other three conditions.

Modeling level of reference

We formulated a probabilistic model of reference level selection that integrates contextual informativeness, utterance cost, and typicality. As in earlier Rational Speech-Acts (RSA) models (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013), the speaker seeks to be informative with respect to an internal model of a literal listener. This listener updates her beliefs to rule out possible worlds that are inconsistent with the meaning of the speaker's utterance. Rather than assuming that words have deterministic truth conditions, as has usually been done in the past, we account for typicality by allowing each label a graded meaning. For instance, the word "dog" describes a dalmatian better than a grizzly bear, but it also describes a grizzly bear better than a tennis ball. The speaker also seeks to be parsimonious: the speaker utility includes both informativeness and word cost; cost includes both length and frequency.

Formally, we start by specifying a literal listener L_0 who hears a word l at a particular level of reference in the context of some set of objects O and forms a distribution over the referenced object, $o \in O$:

$$P_{L_0}(o|l) \propto [\![l]\!](o).$$

Here $[\![l]\!](o)$ is the lexical meaning of the word l when applied to object o. We take this to be a real number indicating the degree of acceptability of object o for category l. We relate this to our empirically elicited typicality norms via an exponential relationship: $[\![l]\!](o) = \exp(\text{typicality}(o,l))$.[5] This relationship is motivated by considering the effect of a small difference in typicality on choice probability: in our elicitation experiment a small difference in rating should mean the same thing at the top and bottom of the scale (it is visually equivalent on the slider that participants used). In order for a small difference in typicality rating to have a constant effect on relative choice probability (which is a ratio), the relationship must be exponential.

Next, we specify a speaker S_1 who intends to refer to a particular object $o \in O$ and chooses among possible nouns $l \in$

[4] Importantly, model comparison between the reported model and one that subsumes basic and super under the same factor level revealed that the three-level condition variable is justified ($\chi^2(1) = 5.7$, $p < .05$), suggesting that participants don't simply revert to the basic level unless contextually forced not to.

[5] Cases where typicality was not elicited were assumed to have typicality 0.

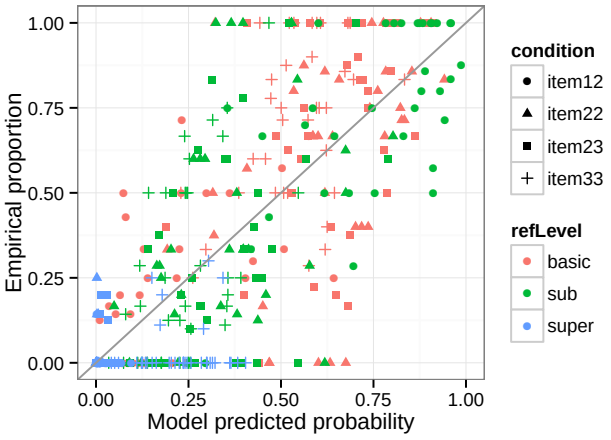


Figure 5: Mean empirical production data for each level of reference against the MAP of the model posterior predictive at the by-target level.

$\mathcal{L}(o)$. We take $\mathcal{L}(o)$ to be the three labels for o at sub, basic, and super level. The speaker chooses among these nouns in a way that is influenced by informativeness of the noun for the literal listener ($\ln P_{L_0}(o|l)$), the frequency ($\hat{c}_f$) and the length ($\hat{c}_l$), each weighted by a free parameter:

$$P_{S_1}(l|o) \propto \exp(\lambda \ln P_{L_0}(o|l) + \beta_f \hat{c}_f + \beta_l \hat{c}_l)$$

Length cost $\hat{c}_l$ was defined as the empirical mean number of characters used to refer at that level and frequency cost $\hat{c}_f$ was the log frequency in the Google Books corpus from 1960 to the present.

We performed Bayesian data analysis to generate model predictions, conditioning on the observed production data (coded into sub, basic, and super labels as described above) and integrating over the three free parameters. We assumed uniform priors for each parameter: $\lambda \sim Unif(0,20)$, $\beta_f \sim Unif(0,5)$, $\beta_l \sim Unif(0,5)$. We implemented both the cognitive and data-analysis models in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, electronic). Inference for the cognitive model was exact, while we used Markov Chain Monte Carlo (MCMC) to infer posteriors for the three free parameters.

Point-wise maximum a posteriori (MAP) estimates of the model's posterior predictives at the target level (collapsing across distractors for each target, within each condition) are compared to empirical data in Fig. 5. On the by-target level the model achieves a correlation of $r = .79$. Looking at results on the by-domain level (collapsing across targets) and on the by-condition level (further collapsing across domains, as in Fig. 3) yields correlations of .88 and .96, respectively. The model does a good job of capturing the quantitative patterns in the data, especially considering the sparsity of our data at the by-target level. One clear flaw is that the model predicts greater use of the super level label than people exhibit. Further systematic deviation appears likely for specific

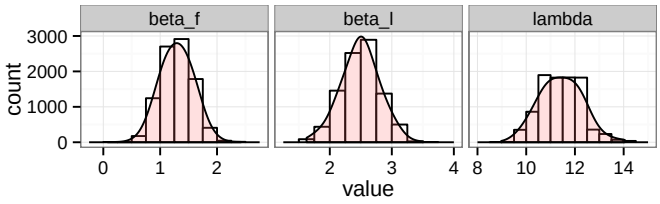


Figure 6: Posterior distribution over model parameters. Maximum a posteriori (MAP) $\lambda = 10.8$, 95% highest density interval (HDI) $= [9.7, 12.8]$; MAP $\beta_l = 2.5$, HDI $= [1.9, 3.1]$; MAP $\beta_f = 1.3$, HDI $= [0.8, 1.8]$.

items. On examination, candy items like "gummy bears" or "jelly beans" were particularly problematic, being referred to primarily by their sub level term in all contexts.

Parameter posteriors are presented in Fig. 6. Informativeness is weighted relatively strongly, while length is weighted somewhat more strongly than frequency. Note that the 95% highest density intervals (HDIs) for all three weight parameters exclude zero, indicating that some contribution of each is useful in explaining the data. In order to ascertain whether typicality was indeed contributing to the explanatory power of the model, we ran an additional Bayesian data analysis with an added typicality weight parameter $\beta_t \in [0,1]$. This parameter interpolated between empirical typicality values (when β_t=1) and deterministic (i.e. 0 or 1) *a priori* values based on the true taxonomy (when β_t=0). We found a MAP estimate for β_t of .94, HDI $= [0.88, 1]$, strongly indicating that it is useful to incorporate empirical typicality values. Finally, we ran a model including a parameter weighting the *product* of frequency and cost, corresponding to the interaction term in our regression analysis. Its posterior distribution was strongly peaked at 0, indicating that any contribution of the interaction is already captured by other aspects of the model.

Discussion and conclusion

The choice speakers make of how to refer to an object is influenced by a rich variety of factors. In this paper, we specifically investigated the choice of level of reference in nominal referring expressions. In an interactive reference game task in which speakers freely produced referring expressions, utterance choice was affected by utterance cost (in terms of length and frequency), contextual informativeness (as manipulated via distractor objects), and object typicality. The interplay of these factors is naturally modeled within the RSA framework, where speakers are treated as choosing utterances by soft-maximizing utterance utility, which includes terms for informativeness and cost. In previous formulations of RSA models, informativeness was determined by a deterministic semantics; here we "softened" the semantics by allowing nouns to apply to objects to the extent that those objects were rated as typical for the nouns. The resulting model provided a good fit to speakers' empirical utterance choices, both qualitatively and quantitatively.

The model predicts a well-documented preference for

speakers to refer to objects at the basic level when not constrained by contextual considerations (Rosch et al., 1976). In our model, this preference emerges naturally from cost considerations: basic-level labels tend to be shorter and more frequent than sub and super level terms. However, speakers did not always use the basic level term, even when unconstrained by context. In certain cases where object typicality was relatively high for the sub level term compared to the basic level term, that term was preferred (as was the case for "panda bear"), suggesting an interesting interplay between typicality and level of description. While our results show that a model can capture several basic-level phenomena through frequency, length, and typicality features, it leaves open the origin and causal role of these linguistic regularities. Future research will be needed to determine how linguistic regularities are related to conceptual regularities and why.

An interesting analogy can be drawn from choosing a noun to choosing a set of adjectives; that is, between selection of a level of reference in simple nominal referring expressions and selection of a set of features to include in modified referring expressions. For the latter, a much discussed phenomenon is that of *overinformative* modifier use (Gatt, Krahmer, van Deemter, & van Gompel, 2014)—for example, saying "big blue" when all objects in the context are blue. The preference for the basic level in the *super sufficient* condition and the still substantial use of sub level terms in the *basic sufficient* condition can also be considered overinformative. However, we showed that a Rational Speech-Acts model using non-deterministic semantics, derived from typicality estimates, predicts that speakers *should* use these more specific descriptions. The extent to which similar considerations may apply to modified referring expressions should be explored. Future research should also examine the interaction of these choices: circumstances under which speakers choose a modifier and how nominal and modifier choice interact.

Acknowledgments

This work was supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and an SNF Early Postdoc. Mobility Award to JD. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

References

Brennan, S. E., & Clark, H. H. (1996, nov). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology. Learning, memory, and cognition*, *22*(6), 1482 – 1493.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. In *Proceedings of the 35th annual conference of the cognitive science society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society.*

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, *29*(8), 899–911.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–84.

Goodman, N. D., & Stuhlmüller, A. (electronic). *The design and implementation of probabilistic programming languages.* Retrieved 2015/1/16, from `http://dippl.org`

Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics*, *3*, 41–58.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Huber.

Kao, J., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(33), 12002–12007.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Orita, N., Vornov, E., Feldman, N., & Daumé III, H. (2015). Why discourse affects speakers' choice of referring expressions. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1639–1649.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. In *Proceedings of the 16th workshop on the semantics and pragmatics of dialogue* (pp. 107 – 116).

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Sedivy, J. C. (2003, jan). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of psycholinguistic research*, *32*(1), 3–23.

van Deemter, K., Gatt, A., van Gompel, R. P. G., & Krahmer, E. (2012, apr). Toward a computational psycholinguistics of reference production. *Topics in cognitive science*, *4*(2), 166–83.

Do classifier categories affect or reflect object concepts?

Laura J. Speed[1] **Jidong Chen**[2] **Falk Huettig**[3] **Asifa Majid**[1,3,4]
(l.speed@let.ru.nl) (jchen@csufresno.edu) (Falk.Huettig@mpi.nl) (asifa.majid@let.ru.nl)
[1]Centre for Language Studies, Radboud University, Nijmegen, NL
[2]California State University, Fresno, USA
[3]Max Planck Institute for Psycholinguistics, Nijmegen, NL
[4]Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL

Abstract

We conceptualize objects based on sensorimotor information gleaned from real-world experience. To what extent is conceptual information structured according to higher-level linguistic features? We investigate whether classifiers, a grammatical category, shape the conceptual representations of objects. In three experiments native Mandarin speakers (a classifier language) and native Dutch speakers (a language without classifiers) judged the similarity of a target object with four objects (presented as words or pictures). One object shared a classifier with the target, the other objects did not. Overall, the target object was judged as more similar to the object with the shared classifier than distractor objects in both Dutch and Mandarin speakers, with no difference between the two languages. Thus, even speakers of a non-classifier language are sensitive to object similarities underlying classifier systems, and using a classifier system does not exaggerate these similarities. This suggests that classifier systems reflect, rather than affect, conceptual structure.

Keywords: classifiers; object concepts; Mandarin; Dutch; linguistic relativity; language and thought

Introduction

When asked to describe the similarity between a knife and a sword, one might describe their visual properties: they are both "*sharp*", "*metallic*", "*shiny*", etc. Likewise for the similarities between a saw and a pair of scissors, there are numerous similarities. The present research investigates whether the addition of a shared grammatical category between nouns would serve to increase conceptual similarity of objects too. Do conceptual representations reflect universal sensory and motor regularities? Or can they also be structured according to higher-level linguistic information?

For decades psychologists, linguists, and anthropologists have debated whether or not the language we speak can affect the way we think about the world. Evidence for some effect of language on thought has been provided for many domains, such as color (e.g., Davidoff, Davies, & Roberson, 1999; Gilbert, Regier, Kay, & Ivry, 2006; Winawer, Witthoft, Frank, Wu, Wade, & Boroditsky, 2007), spatial cognition (e.g., Levinson, 2003; Majid, Bowerman, Kita, Haun, & Levinson, 2004), and time (e.g., Boroditsky, 2001; Boroditsky, Fuhrman, McCormick, 2010), to name but a few.

One linguistic domain in which effects of language on thought have been extensively investigated is grammatical gender. Grammatical gender divides nouns into classes according to the behavior of associated words (e.g., articles, adjectives; cf. Corbett, 2006). In some languages, nouns are grammatically classified according to sex, i.e., masculine or feminine. For example *apple* is masculine in German, *der Apfel*, but feminine in French, *la pomme*. Grammatical gender tends to be semantically arbitrary for objects without a natural gender; however, it has been shown to affect how speakers of such languages think about objects. For example, Spanish and German speakers are more likely to ascribe male qualities to grammatically masculine objects and female qualities to grammatically feminine objects (Boroditsky, Schmidt, & Phillips, 2003).

We may ask, however, whether such linguistic features are, in fact, changing the conceptual structure of objects, or whether they are merely being used as strategic devices during language processing. One way to test this is to assess the effect of such grammatical systems using tasks that recruit language to a greater or lesser extent. Phillips and Boroditsky (2003) provided some evidence that grammatical gender affects object concepts for German and Spanish speakers. They found effects of grammatical gender on judgments of similarity between people and objects even when the task was completed in English (a language with no gender system), when performing a non-linguistic task (using pictures), and during a verbal interference task (suggesting real conceptual change). On the other hand, Bender, Beller, and Klauer (2011) suggest grammatical gender is only available as a syntactic property, and does not change conceptual representations. Lexical decisions to nouns were faster when they had been preceded by words matching in grammatical gender but not for words matching in semantic gender (natural gender). That is, congruence in the syntactic features (grammatical gender) facilitated responses, but congruence between grammatical gender and semantic gender (which would reflect conceptual similarity) did not affect response time. Recently, Bender, Beller, and Klauer (2016) have suggested grammatical gender effects may in fact be due to "personification" (with nouns being associated to personified allegories such as *Lady Liberty*), rather than grammatical information, per se.

Here we look at a different grammatical category associated with nouns, i.e., numeral classifiers. In languages with numeral classifiers, their use is obligatory when a specific quantity is indicated, with certain quantifiers, and with demonstratives. Numeral classifiers are said to refer to

a specific feature (e.g., material, shape, size) of the entity associated with the corresponding noun (cf., Allan, 1977). For example, the Mandarin classifier *tiao2* refers to long, rope-like objects such as legs, snakes, and rivers. Although classifier categories have underlying conceptual meanings, there can be large variability with respect to category size and coherence, with some classifiers covering a broad range of objects and taxonomic categories (e.g., Saalbach & Imai, 2012). So, although *tiao2* is used with "long" things, it can also be used for less prototypically long things, such as dogs, underwear and a piece of news.

It appears that speaking a language with a classifier system can affect how speakers of a language conceptualize objects. Lucy and Gaskins (2003), for example, compared speakers of Yucatec Maya (an indigenous language of Mexico that contains numeral classifiers) with English speakers. Speakers of the two languages judged the similarity of objects differently: Yucatec Mayans had a preference for matching objects according to material, whereas English speakers matched the same objects according to shape. Zhang and Schmitt (1998) and Saalbach and Imai (2007) found Mandarin speakers rated pairs of nouns that shared a classifier as more similar than nouns that did not share a classifier, but speakers of non-classifier languages (English and German) rated both pairs similarly.

But, do classifier systems really *affect* the conceptual organization of object concepts, or do they *reflect* conceptual organization instead? Classifiers are usually not arbitrarily related to features of objects. They pick out common features of the entities they classify (cf., Allen, 1977; Lakoff, 1987). There is evidence to suggest there is some conceptual salience to the features classifiers denote. For example, Clark (1976) noted the parallel between the conceptual features frequently found in classifier languages and the conceptual features salient to children in early word learning. Classifiers often denote features such as "*long*", "*round*", or "*animate*"; the exact features on which children base their over-generalization errors. No language has been found with a classifier system that distinguishes referents on the basis of color, and similarly children do not use color as a basis of their over-generalizations.

Thus, instead of classifiers influencing the organization of object concepts, classifiers could reflect the organization of object concepts, thus providing evidence for the effect of thought on language. Classifier categories may be structured around natural similarities in the world. If so, speakers of non-classifier languages should perceive relations between objects sharing classifiers similarly to those speaking classifier languages. Consistent with this, Saalbach and Imai (2005) found speakers of German, as well as Mandarin, judged objects sharing a classifier as more similar than pairs of objects that did not share classifiers.

If classifiers reflect universally recognized similarities between objects, how does this reconcile with previous findings of linguistic relativity (e.g., Zhang & Schmitt, 1998; Saalbach & Imai, 2007)? The answer may lie in the experimental details. The majority of previous experiments investigating effects of classifier systems present people with words from the participants' language; e.g., Mandarin speakers are presented with names of objects in Mandarin, and English speakers with names in English. So, in one sense, it is not surprising speakers of classifier languages are sensitive to the grammatical information associated with nouns. If we think of current psycholinguistic models of language use, then when a noun is activated grammatical information forming part of the lemma level representation would also be activated. Linguistic relativity effects could be explained by the activation of the relevant grammatical feature associated with the word form, rather than differences in the conceptual representation of objects, per se (cf., Kousta, Vinson, & Vigliocco, 2008). To test whether classifiers, in fact, affect the structure of object concepts or whether they serve only to facilitate processing of words with shared grammatical features, parallel experiments with linguistic and non-linguistic stimuli need to be conducted.

We investigated one classifier language – Mandarin – and one language without classifiers – Dutch. Participants were presented with a target object and had to rate its similarity to four other objects. Using a rating task means more fine-grained differences can be detected, compared to, for example a forced-choice task (e.g., Lucy & Gaskins, 2003; Saalbach & Imai, 2007, Experiment 1), where participants have to choose one object over another. One of the four objects used a noun that possessed the same classifier as the target object, the other three objects did not. If classifiers affect the way objects are thought of, we could expect speakers of Mandarin would judge the target object and classifier object as more similar than the other objects, but Dutch speakers would not, because having a shared classifier will increase perceived similarity. If, on the other hand, classifiers reflect real-world similarities amongst objects, then we would expect both Mandarin and Dutch speakers to judge the target object and classifier object as more similar than the distractor objects. A third possibility is that both Mandarin and Dutch speakers judge the classifier object as more similar to the target than the distractors, because of real-world similarities that the classifier system is built upon, but that this effect is greater in Mandarin than Dutch speakers (cf., Saalbach & Imai, 2007, Experiment 2 and 3). Further, if classifier systems affect object representations, rather than lexical representations, then we should see differences between Mandarin and Dutch speakers on judgments of pictures, as well as words. In the following three experiments we assess the effect of classifiers on similarity judgments with pictures and words.

Experiment 1. Picture-picture similarity

Method

Participants

25 native Mandarin speakers and 24 native Dutch speakers participated in the experiment (average age 23, range 18-

33). All speakers were familiar with other languages. Mandarin speakers came from different dialect backgrounds (e.g., Shandong, Henan, Shanxi, Jilin, Hubei, Jiangsu), but were all educated in Putonghua (standard Mandarin) and mainly use this language daily, both at work and at home. Dutch speakers were all multilingual with English, German, Spanish and French. None of the Dutch participants were familiar with Mandarin, or any other language with classifiers.

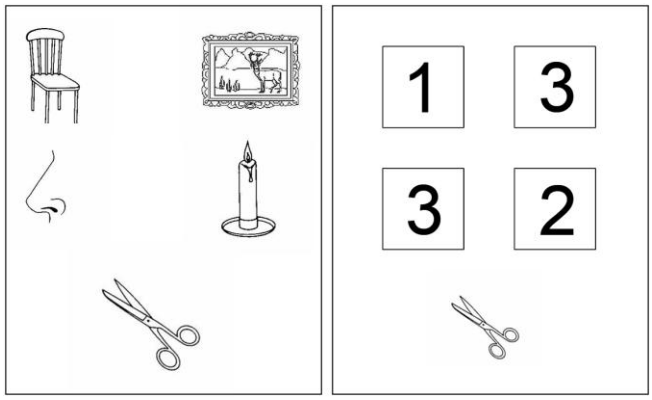


Figure 1. Example page of the stimulus booklet (left) with target (scissors), classifier match (chair) and comparison objects, with response sheet (right).

Stimuli

Line drawings of everyday objects, familiar to both groups of participants, were used. Each trial consisted of a target object and four comparison objects. The target object shared a classifier with one of the comparison objects. A norming study was first conducted in which a separate set of twelve native Mandarin speakers named 240 line drawings of concrete and imageable objects. Classifier choice for a noun was not strict (i.e., nouns could take different classifiers), but there was a clear dominant classifier for each object. Based on the norming results, we chose noun pairs that shared the same dominant classifier. The remaining three comparison objects were "distractors", and did not share a classifier with the target. On each trial, pictures of the four comparison objects were presented on a piece of paper, laid out in a two-by-two grid. The target object was centered under the comparison objects (see Figure 1). Participants were given a separate response sheet, consisting of four empty boxes in the same configuration as the comparison objects, with a picture of the target object at the bottom of the sheet. The experiment consisted of 48 trials. The classifier match appeared equally often in each of the positions within the gird, and the position of an each item was counter-balanced across participants.

Procedure

Participants were instructed, in their native language, to judge how similar each target object was to each of the other objects. They were asked to indicate their judgments in the corresponding boxes in their answer sheet, with 0 indicating "no similarity" and 10 "identical in similarity".

Results

Previous studies have shown Chinese participants give overall higher similarity ratings than Westerners (Saalbach & Imai, 2007). In order to control for any differences in how the similarity scale was used across individuals, we transformed each participant's rating scores into standardized *z*-scores. The transformed data were then analyzed with a language (Mandarin vs. Dutch) by object (classifier match vs. distractor 1 vs. distractor 2 vs. distractor 3) mixed ANOVA, treating participants and items as random effects.

There was a main effect of object type on similarity ratings $F_1(3, 141) = 40.05$, $p < .0001$, $\eta_{2p} = .46$; $F_2(3, 144) = 5.72$, $p < .0001$, $\eta_{2p} = .11$. Simple planned comparisons revealed the classifier match received higher similarity ratings than any of the other distractors: classifier vs. distractor-1 $F_1(1, 47) = 67.64$, $p < .001$, $\eta_{2p} = .6$; $F_2(1, 48) = 17.02$, $p < .001$, $\eta_{2p} = .26$, classifier vs. distractor-2 $F_1(1, 47) = 57.4$, $p < .001$, $\eta_{2p} = .557$; $F_2(1, 48) = 12.37$, $p < .001$, $\eta_{2p} = .21$, classifier vs. distractor-3 $F_1(1, 47) = 105.18$, $p < .001$, $\eta_{2p} = .69$; $F_2(1, 48) = 11.05$, $p = .002$, $\eta_{2p} = .19$. There were no significant differences between the distractor objects.

There was no significant main effect of language $F_1 < 1$; $F_2 < 1$. Crucially, there was no significant language by object type interaction $F_1(3, 141) = 1.4$, $p = .25$, $\eta_{2p} = .03$; $F_2 < 1$ (see Figure 2).

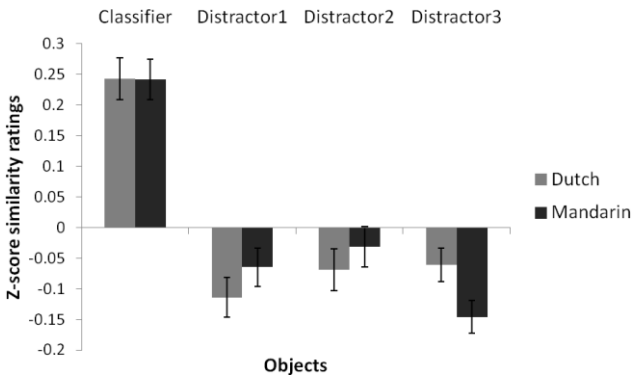


Figure 2. Mandarin and Dutch speakers' picture-picture similarity judgments (*z*-score) Experiment 1 (bars = 1*SE*).

Experiment 2. Word-picture similarity

We found a robust effect of classifier pair similarity, but no indication of a difference between Mandarin and Dutch speakers in Experiment 1. This suggests object representations are not affected by classifiers, but does not rule out lexical-level language-specific effects. In Experiment 2 we presented the target object as a word, in order to encourage activation of the classifier categorization.

Participants

A different set of 25 Mandarin speakers and 24 Dutch speakers (average age 23 years, range 18-28) participated in

the experiment. The demographic characteristics and recruitment procedure was the same as in Experiment 1.

Stimuli
The stimuli were as in Experiment 1 but now the target was presented as a word (without the classifier) in the native language and the comparison objects were presented as pictures.

Procedure
Participants were asked to judge similarity of the target object to the comparison objects on a scale from 1 (no similarity) to 10 (identical in similarity).

Results
Data was analyzed as in Experiment 1. There was a main effect of object type on similarity ratings $F_1(3, 141) = 52.92$, $p < .001$, $\eta_{2p} = .53$; $F_2(3, 144) = 8.66$, $p < .001$, $\eta_{2p} = .15$. Simple planned comparisons revealed the classifier match received higher similarity ratings than any of the other distractors: classifier vs. distractor-1 $F_1(1, 47) = 70.79$, $p < .001$, $\eta_{2p} = .6$; $F_2(1, 48) = 20.57$, $p < .001$, $\eta_{2p} = .30$, classifier vs. distractor-2 $F_1(1, 47) = 73.05$, $p < .001$, $\eta_{2p} = .61$; $F_2(1, 48) = 20.61$, $p < .001$, $\eta_{2p} = .30$, classifier vs. distractor-3 $F_1(1, 47) = 73.71$, $p < .001$, $\eta_{2p} = .61$; $F_2(1, 48) = 11.54$, $p < .001$, $\eta_{2p} = .19$. There were no significant differences between the distractor objects.

There was again no main effect of language $F_1 < 1$; $F_2 <$ 1and no significant language by object type interaction $F_1(3, 141) = 1.54$, $p = .21$ $\eta_{2p} = .03$; $F_2 < 1$ (see Figure 3).

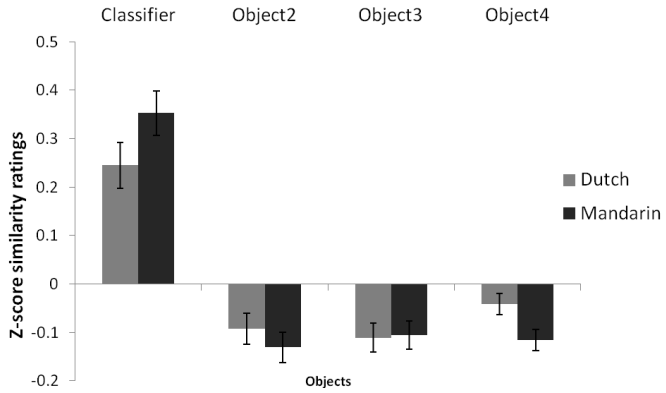


Figure 3. Mandarin and Dutch speakers' word-picture similarity judgments (z-score) Experiment 3 (bars = 1*SE*)

Experiment 3. Word-word similarity

Again, in Experiment 2 we find a robust classifier effect, but no difference between Dutch and Mandarin speakers. As a final test of a classifier effect we maximized the linguistic context by presenting both the target and the comparison objects as words.

Participants
A different set of 25 Mandarin speakers and 24 Dutch speakers participated in the experiment (average age 23 years, range 18-28). The demographic characteristics and recruitment procedure was as in Experiment 1 and 2.

Stimuli
The stimuli were as in Experiment 1 and 2, but participants were presented with words instead of pictures. Mandarin participants were presented with nouns in Mandarin characters and Dutch participants with the Dutch nouns.

Procedure
Participants were asked to judge similarity of the target object to the comparison objects on a scale from 1 (no similarity) to 10 (identical in similarity).

Results
Data was analyzed as in Experiment 1 and 2. There was a main effect of object type on similarity ratings $F_1(3, 141) = 52.92$, $p < .001$, $\eta_{2p} = .53$; $F_2(3, 144) = 9.6$, $p < .001$, $\eta_{2p} = .17$. Simple planned comparisons revealed the classifier match received higher similarity ratings than any of the other distractors: classifier vs. distractor-1 $F_1(1, 47) = 145.22$, $p < .001$, $\eta_{2p} = .6$; $F_2(1, 48) = 27.25$, $p < .001$, $\eta_{2p} = .36$, classifier vs. distractor-2 $F_1(1, 47) = 201.19$, $p < .001$, $\eta_{2p} = .81$; $F_2(1, 48) = 16.95$, $p < .001$, $\eta_{2p} = .26$, classifier vs. distractor-3 $F_1(1,47) = 80.56$, $p < .001$, $\eta_{2p} = .63$; $F_2(1, 48) = 7.99$, $p = .007$, $\eta_{2p} = .19$. There were no significant differences between the distractor objects.

There was again no main effect of language $F_1(3, 141) = 1.54$, p = .21, $\eta_{2p} = .03$; $F_2 < 1$ and crucially no significant language by object type interaction $F_1(3, 141) = 1.54$, $p = .21$ $\eta_{2p} = .03$; $F_2 < 1$ (see Figure 3).

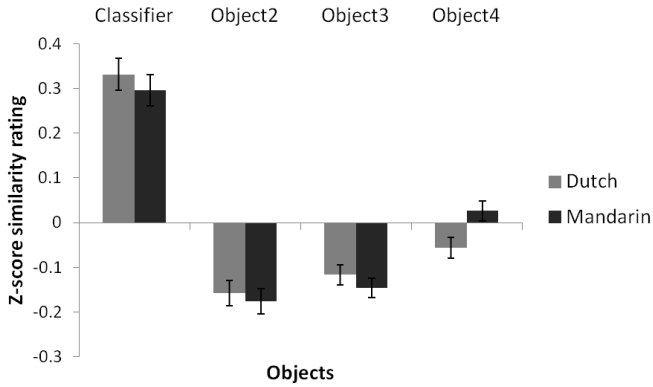


Figure 4. Mandarin and Dutch speakers' word-word similarity judgments (z-score) Experiment 2 (bars = 1*SE*).

Discussion

Across three experiments we found objects were judged as more similar to a target object when they shared a classifier in Mandarin compared to when they did not. However, this effect was observed both in a language that uses classifiers – i.e., Mandarin – as well as a language that does not have a classifier system – i.e., Dutch. Moreover, the magnitude of this effect did not differ between Mandarin and Dutch

speakers. From these results we conclude classifiers do not impact overall object similarity.

Why do some studies find differences in similarity ratings between languages with and without classifiers, but we do not? Our study focused on overall similarity of objects, and used stimuli (e.g., visual pictures) which emphasized shape. We did not use real objects, unlike Lucy and Gaskins (2003), who found what distinguished Yucatec from English speakers was attention to shape vs. material. It is possible, therefore, that we missed a critical ontological distinction. Nevertheless, other studies have focused on pictorial or linguistic stimuli and have reported linguistic relativity effects (Zhang & Scmitt, 1998; Saalbach & Imai, 2007). So this cannot be the whole story.

It is possible a shared classifier is more salient when there are fewer objects to judge. For example, when comparing only two objects at a time, with no distractors, a greater number of features of a word (such as grammatical category) can be attended to. However, when there are four objects to compare simultaneously, it is likely that only the most salient similarities are attended to. It is also possible that we failed to find an effect of language on classifier similarity judgments because we did not present the words with their classifiers, but as bare nouns. Huettig, Chen, Bowerman, and Majid (2010) found people look towards objects sharing a classifier with a spoken word more than visual distractors, but only when the classifier was explicitly used, and Gao and Malt (2009) found heightened classifier effects when a classifier was present in a sentence compared to when it was not. Similarly, Vigliocco, Vinson, Indefrey, Levelt, and Hellwig (2004) found grammatical gender effects in speech substitution errors only when nouns were produced with determiners marked for gender, but not for single noun phrases or noun phrases with indefinite determiners not marked for gender. Thus, grammatical information may need to be explicit and salient to affect categorization.

It is also possible that using similarity judgments as a measurement is not sensitive enough to reveal differences between Dutch and Mandarin speakers. That is, since classifiers are mostly built on shared real-world features, the similarity between the target object and the classifier object in terms of these features may be quite obvious for explicit similarity judgments. That is, in some sense, the judgments could be at ceiling level. Perhaps a more low-level, automatic task, closer to real-world language use, could reveal the linguistic advantage of a classifier system. However, Saalbach and Imai (2011) found the opposite results: Chinese speakers showed an enhanced classifier similarity effect compared to German speakers in a similarity rating task, but not a speeded word-picture matching task.

Similarly, Gao and Malt (2009) propose there are three classifier categories: "well-defined" categories in which there is a clear feature that all objects sharing the classifier possess; "prototype" categories in which there is a typical feature, but also a gradient of typicality in category membership; and "arbitrary" categories for which there are no typical features defining membership. One might predict that objects belonging to the "well-defined" classifier category could easily be grouped in terms of similarity by speakers of a non-classifier language, but only speakers of that specific classifier language could group the "arbitrary" category correctly. Thus, it is possible that the classifiers used in the present study fit more into the "well-defined" category. It could be predicted that an advantage for speakers of a classifier language could be found if the classifier pairs used shared "arbitrary" classifiers. However, Goa and Malt (2009) only found an advantage for the "well-defined" category amongst Mandarin speakers when testing recall of nouns with shared classifiers.

So, does language influence thought? Our results suggest that a grammatical feature, classifiers, does not have an influence on the way that objects are categorized, as measured by global similarity judgments. However, it is likely that other forms of linguistic information do affect the way objects are thought about. For example, effects of grammatical gender on object categorization appear to be well-attested. As noted above, classifiers are said to highlight only one, or a few, features of an object (e.g., "shape"; cf. Allen, 1977), and not information relevant to the entire concept. This contrasts with grammatical gender, which could affect the way all features of an object are conceptualized. For example, Spanish speakers described the word *key*, a noun with female grammatical gender, with female characteristics "*golden, intricate, little, lovely, shiny and tiny*" (Boroditsky, Schmidt & Phillips, 2003). Thus, gender attributes have knock-on effects onto other features: e.g., size, hedonics, texture, visual appearance (perhaps through connotative meaning; cf. Osgood, Suci, & Tannenbaum, 1957).

Previous research on the cognitive consequences of speaking a classifier language have been mixed, with the suggestion that effects are modest. In three experiments we failed to find language-specific heightened sensitivity of object similarity for speakers of a classifier language compared to speakers of a non-classifier language, using pictures and words. We, therefore, conclude that classifier systems do not affect overall conceptual representations of objects. Previous evidence could be the result of strategic or explicit use of classifier information that does not reflect typical categorization processes. Alternatively, classifiers may affect conceptualization of specific object features, which cannot be observed with overall similarity judgments (e.g., Lucy & Gaskins, 2003).

Both speakers of a classifier language and a non-classifier language judged objects sharing a classifier as more similar than objects that did not share a classifier. Thus, classifier systems do not impact overall object similarity, but instead likely reflect the way the world is organized.

Acknowledgements

This research was funded by the Max Planck Gesellschaft. LS and AM are supported by The Netherlands Organization for Scientific Research: NWO VICI grant.

References

Allan, K. (1977). Classifiers. *Language*, *53*(2), 285-311

Bender, A., Beller, S., & Klauer, K. C. (2011). Grammatical gender in German: A case for linguistic relativity? *The Quarterly Journal of Experimental Psychology*, *64*(9), 1821–1835.

Bender, A., Beller, S., & Klauer, K. C. (2016). Lady Liberty and Godfather Death as candidates for linguistic relativity? Scrutinizing the gender congruency effect on personified allegories with explicit and implicit measures. *The Quarterly Journal of Experimental Psychology*, *69*(1), 48–64.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English Speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2010). Do English and Mandarin speakers think about time differently? *Cognition*, *118*(2), 123–129.

Boroditsky, L., Schmidt, L., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.) *Language in mind: Advances in the study of language and thought* (pp. 61–80). Cambridge, MA: MIT Press.

Chao, Y-R. (1968). *A grammar of spoken Chinese*. Berkeley, CA: University of California Press.

Corbett, G. G. (2006). Gender, grammatical. *The Encyclopedia of language and linguistics. 2nd Edition. Oxford: Elsevier*, (pp.749–756).

Clark, E. V. (1976). Universal categories: On the semantics of classifiers and children's early word Meanings. In A., Juilland (Ed.) *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday*. *Vol* 3, (pp. 449-62) Saratog, CA: Anma Libri.

Davidoff, J., Davies, I. & Roberson, D. (1999) Colour categories of a stone-age tribe. *Nature, 398*, 203-204.

Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes*, *24*(7-8), 1124–1179.

Gilbert, A. L., Regier, T., Kay. P. & Ivry R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences, 103*(2), 489–494.

Huettig, F., Chen, J., Bowerman, M., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture*, *10*(1), 39–58.

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2008). Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 843–858.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Levinson, Stephen C. (2003). *Space in language and cognition*. Cambridge University Press: Cambridge.

Lucy J. A, & Gaskins S. (2003) Interaction of language type and referent type in the development of nonverbal classification preferences. In D. Gentner & S. Goldin-Meadows (Eds). *Language in mind: Advances in the study of language and thought.* (pp. 465-492). Cambridge, MA: MIT Press.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8*(3), 108-114.

Osgood, C.E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. In *Proceedings of the 25th annual meeting of the Cognitive Science Society.* Lawrence Erlbaum Associates Mahwah, NJ.

Saalbach, H., & Imai, M. (2005). Do classifier categories structure our concepts? In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates Mahwah, NJ.*y*

Saalbach, H., & Imai, M. (2007). The scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General*, *136*(3), 485-501.

Vigliocco, G., Vinson, D. P., Indefrey, P., Levelt, W. J. M., & Hellwig, F. (2004). Role of grammatical gender and semantics in German word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 483–497.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785.

Zhang, S., & Schmitt, B. (1998). Language-dependent classification: The mental representation of classifiers in cognition, memory, and ad evaluations. *Journal of Experimental Psychology: Applied*, *4*(4), 375–385.

Exploring the Cost Function in Color Perception and Memory: An Information-Theoretic Model of Categorical Effects in Color Matching

Chris R. Sims (chris.sims@drexel.edu)
Applied Cognitive & Brain Sciences, Department of Psychology, Drexel University

Zheng Ma (zma4@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Sarah R. Allred (srallred@scarletmail.rutgers.edu)
Department of Psychology, Rutgers University–Camden

Rachel A. Lerch (rachel.a.lerch@drexel.edu)
Applied Cognitive & Brain Sciences, Department of Psychology, Drexel University

Jonathan I. Flombaum (flombaum@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University

Abstract

Recent evidence indicates that color categories can exert a strong influence over color matching in both perception and memory. We explore this phenomenon by analyzing the cost function for perceptual error. Our analysis is developed within the mathematical framework of rate–distortion theory. According to our approach, the goal of perception is to minimize the expected cost of error while subject to a strong constraint on the capacity of perceptual processing. We propose that the cost function in color perception is defined by the sum of two components: a metric cost associated with the magnitude of error in color space, and a cost associated with perceptual errors that cross color category boundaries. A computational model embodying this assumption is shown to produce an excellent fit to empirical data. The results generally suggest that what appear as 'errors' in working memory performance may reflect reasonable and systematic behaviors in the context of costs.

Keywords: color perception; visual working memory; information theory; rate–distortion theory

Visual working memory is central to daily life. Even extremely simple tasks, such as visually comparing the size or color of two objects, requires the storage and manipulation of perceptual information in working memory. Given the central role of visual working memory in natural tasks, it is quite surprising that this system is also quite limited. Previous studies have demonstrated that the capacity of visual working memory for simple unitary features such as color or orientation is on the order of 2–4 bits (Sims, Jacobs, & Knill, 2012; Sims, 2015). With such a strong constraint on information processing, it seems especially paramount that the brain use its available working memory capacity in an efficient manner. But what defines an 'efficient' perceptual system?

A natural and intuitive answer is that visual working memory is used efficiently when it minimizes task-relevant costs and errors. According to this perspective, the key to understanding perceptual processing is identifying the particular *cost function* that it seeks to minimize. Abstractly, if a sensory signal x is misperceived or misremembered as a different signal y, then there exists some subjective cost (or disutility) associated with this error, and this can be quantified by some function $\mathcal{L}(x,y)$. An efficient perceptual system is one that minimizes the expected cost according to a particular cost function, while subject to a constraint on the capacity of the perceptual channel. Recent work (Sims et al., 2012; Sims, 2015) has demonstrated that this problem statement corresponds quite naturally to a branch of information theory known as rate–distortion theory (Berger, 1971). Rate–distortion theory concerns the optimal solution to the problem of minimizing the costs of communication error, subject to constraints on available capacity.

The goal of the present paper is to apply rate–distortion theory in order to identify the cost function that drives color matching in perception and memory. For example, if a particular shade of red is misperceived or misremembered as a slightly different shade of red, how costly is that error to the brain? Although this may seem like a trivial question, color perception–even in simple laboratory contexts–exhibits many subtle properties that are not completely understood (Allred & Flombaum, 2014). Of particular relevance is the finding that categories can strongly influence perception and memory (Huttenlocher, Hedges, & Vevea, 2000; Bae, Olkkonen, Allred, & Flombaum, 2015). The question considered in this paper is how color categories influence both color perception and color memory, as formalized within the mathematical framework of rate–distortion theory.

Bae and colleagues (2015) reported a series of four experiments examining how categories influence color perception and memory, and also developed a computational model to account for their results. According to their model (referred to as CATMET), colors are encoded in two separate channels in perceptual processing: one channel encodes a category-based representation, while the other encodes a metric-based representation of color on a continuous scale. Color perception results from the heuristic integration of these two channels to form an estimate of the afferent sensory signal.

The current paper seeks to explain these same experimental data, but using an alternative modeling approach based on rate–distortion theory. The goal of this effort is not to supplant the CATMET model. Rather, rate–distortion theory can offer an explanation at Marr's computational level of anal-

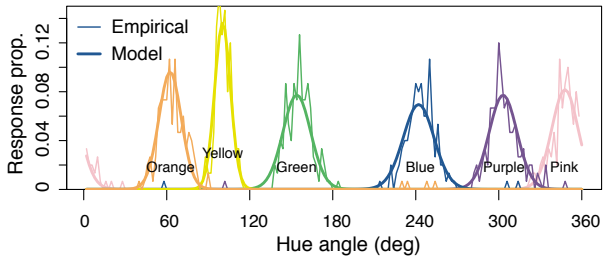


Figure 1: Empirical data and model fit for the color identification experiment conducted by Bae et al. (2015).

ysis (Marr, 1982). Whereas the CATMET model provides a process-level explanation for *how* categories influence perception, the goal is to propose an explanation as to *why* categories influence perception. The model developed in this paper represents a preliminary step in that direction.

To us, it is intuitive and plausible that there is a cost to committing categorical errors in perception. Consider the task of selecting fruit based on ripeness. Both ripe and unripe fruit may encompass a wide range of hues, but for the purpose of finding the best one to eat, perceptual errors that do not cross this category boundary have little practical implication. Much more nutrition is provided by yellow bananas and red berries than green bananas and berries; hence it seems plausible that perception should be sensitive to category boundaries.

In a nutshell, we propose that color perception is the result of the rational minimization of a particular cost function, subject to a constraint on capacity. In the model we will describe, the cost function driving perception is the sum of a metric error term (the distance between a stimulus and its perceived value in color space) as well as an additional cost when perceptual errors cross color category boundaries.

Before describing the mathematical details of the model, we first introduce the experimental results to be explained.

Experimental results

Bae et al. (2015) conducted four experiments examining the relationship between color categories, and color matching in perception and memory. In the current paper we restrict our attention to three of these datasets: color identification, undelayed estimation, and delayed estimation.

In the *color identification* experiment, subjects were presented with a color wheel containing 180 equiluminant colors varying only in hue, along with 6 color category labels (pink, orange, yellow, green, blue, purple). Subjects were asked to simply click on a point along the color wheel to indicate the best example for each of the six color categories. The results from this experiment are shown in Figure 1. Subjects were highly consistent in their identification of color categories.

In the *undelayed estimation* experiment, subjects were presented with a color patch as well as a color wheel from which the color was sampled. The task for the subject was to choose a point along the color wheel to indicate the best color match for the given probe. This procedure was repeated using 180 different color targets, collecting a large number of trials per subject. For complete experimental methods the reader is directed to the source publication (Bae et al., 2015). Note that since the color patch and color wheel remained visible throughout the duration of each trial, it would seemingly be an easy task for subjects to click on the exact matching color on each trial. The *delayed estimation* experiment was methodologically similar, except that in this experiment the color patch disappeared during the response portion of the trial and subjects had to rely on a memory representation of the probe color.

The important results from the undelayed and delayed estimation experiments are illustrated in Figure 2. Figure 2a shows the overall histogram of responses grouped into 180 equal-width bins. Although probe stimuli were sampled uniformly along the color wheel, responses are clearly nonuniformly distributed. The magnitude of this effect is greater in the delayed experiment. Figure 2b shows the mean bias observed for each of the 180 probe stimuli. Positive values indicate responses that were clockwise, on average, relative to the probe stimulus. The bottom panel shows the circular standard deviation of the response distribution. The figure shows that response variability also varied systematically across hues.

Figure 3a gives an overhead view of the both datasets in their entirety, showing the conditional response distribution for each of the 180 stimuli used in the estimation experiments. An unbiased perceptual system would exhibit a straight diagonal line; in contrast, human performance shows a consistent pattern of distortion (bias) and variability.

The empirical results summarized in Figures 2 and 3 demonstrate unquestionably that color perception and color memory both show strong stimulus-specific properties (see also Allred & Flombaum, 2014). Bae and colleagues (2015) previously developed a model that could produce these effects; our goal is to model the effects specifically as the consequence of a cost function, framed by rate-distortion theory.

An information-theoretic model of color perception and color memory

Rate–distortion theory concerns the optimal solution to minimizing costs according to a particular cost function, subject to a constraint on channel capacity. We assume that color perception can be modeled as a communication channel where some input signal x is perceived or remembered as a possibly different signal y. In the experiments under consideration, stimuli are color hues drawn from a circular color wheel; hence x and y can be considered as the angle of a given stimulus, and the response angle around this color wheel. Our model assumes a cost function consisting of two terms: a metric cost related to the angular difference between x and y, and a categorical cost that is based on the probability that x and y would be assigned different color category labels. Hence,

$$\mathcal{L}(x,y) = f(y-x) + P(C_x \neq C_y). \quad (1)$$

The first term represents the metric cost of error. Potential

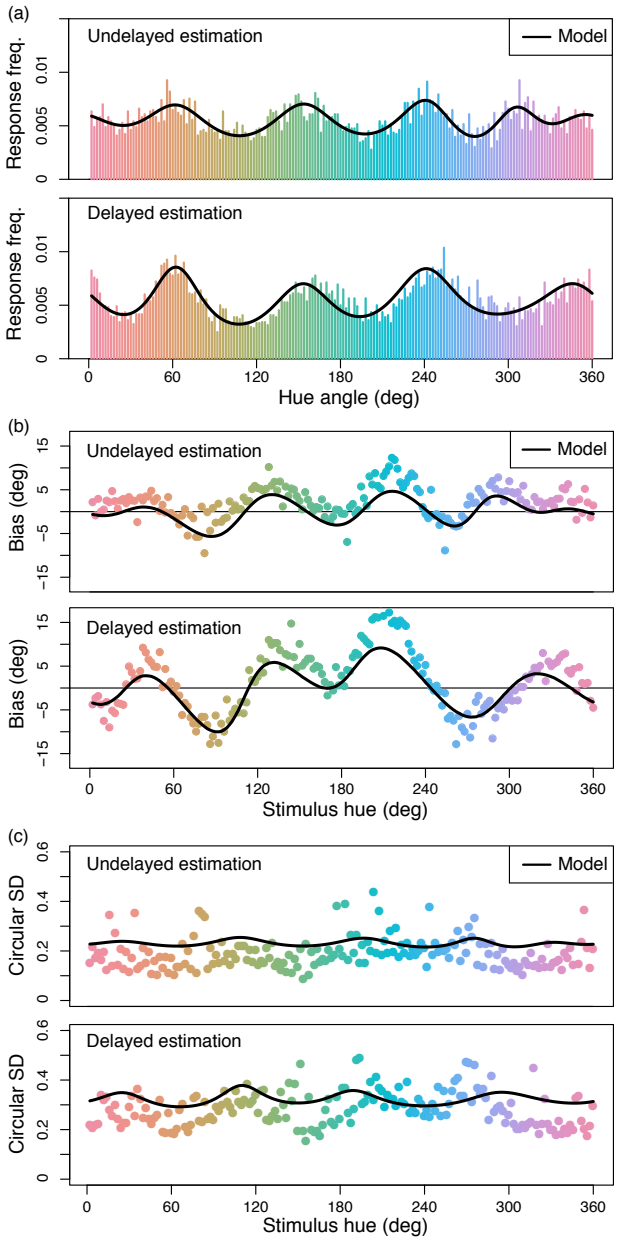


Figure 2: Results from the estimation experiments conducted by (Bae et al., 2015). (a) Response frequency, demonstrating systematic biases in color perception and memory. (b) The magnitude of bias for each stimulus hue. (c) The circular standard deviation in the response distribution for each stimulus hue. In all panels, empirical data is given by colored markers/lines while the black curve shows the information-theoretic model fit to the data.

candidates for this function include the squared error, $(y-x)^2$ or absolute error $|\,y-x\,|$. However, since x and y are points in a circular space we first assume a metric cost function based on the cosine of the difference between x and y:

$$f(y-x) = \frac{1}{2}\left[1-\cos(y-x)\right]. \tag{2}$$

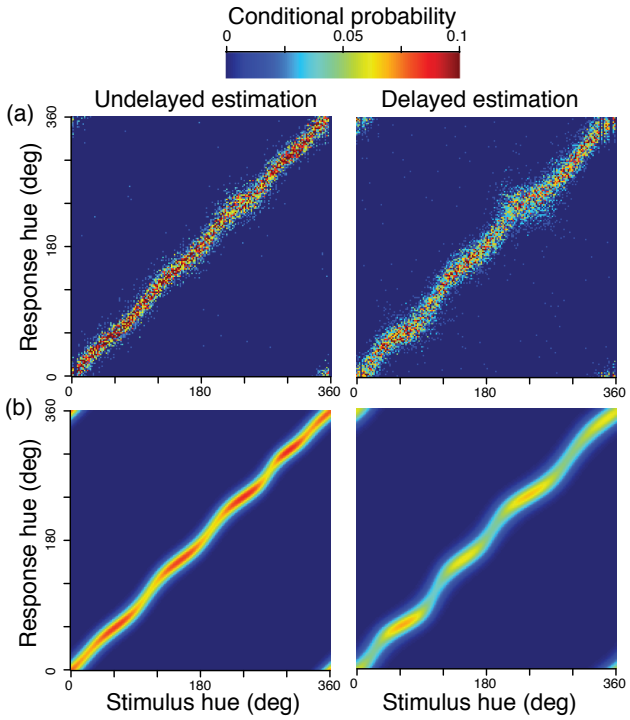


Figure 3: (a) Conditional probability distributions showing the response distribution for each of the 180 stimuli used in the undelayed (left panel) and delayed (right panel) estimation experiments. (b) Model fit to the experimental data.

This function is zero when $x = y$, and reaches a maximum of 1 when x and y differ by the maximum of 180 degrees. The second term in the cost function given by Eq 1 indicates the probability that x and y would be assigned different category labels, indicated by C_x and C_y. Specifying this requires a model of how color hues are mapped on to color categories. Our model assumes that different hues in color space are better or poorer examples of each color category. The strength or 'goodness' of a given hue for a particular color category k is modeled as a scaled Von Mises distribution:

$$\phi_k(x) = \frac{\alpha_k}{e^{\tau_k}} \exp\left(\tau_k \cos(x-\mu_k)\right). \tag{3}$$

Hence, each color category is described by three parameters: its central location (μ), the category precision (τ, the inverse of width), and the maximum strength of the color category (α). If there are K categories, then a given hue angle corresponds to a vector of category strengths, $\langle \phi_1(x), \phi_2(x), \ldots, \phi_K(x) \rangle$. The probability that the hue x is assigned to category k is modeled using the softmax equation:

$$P(C_x = k) = \frac{\exp(\phi_k(x))}{\sum_{j=1}^{K} \exp(\phi_j(x))}. \tag{4}$$

This process allows for noise in color category assignment: when all $\alpha_k = 0$, category assignment is performed at chance. With this generative model of color category assignment, the probability that two hues x and y are assigned different labels

is equal to one minus the probability that they are given the same label:

$$P(C_x \neq C_y) = 1 - \sum_{k=1}^{K} P(C_x = k)P(C_y = k). \quad (5)$$

Equations 1–5 fully describe the cost function $\mathcal{L}(x,y)$ that we assume the perceptual channel seeks to minimize. Specifically, the goal is to minimize the expected cost, subject to a constraint on available channel capacity,

$$\begin{aligned} &\text{Minimize } \mathrm{E}[\mathcal{L}(x,y)] \text{ w.r.t. } p(y \mid x) \\ &\qquad = \sum_x \sum_y \mathcal{L}(x,y)p(y \mid x)p(x) \\ &\text{Subject to } \mathrm{I}(x,y) \leq \mathcal{C}, \end{aligned} \quad (6)$$

where the minimization is performed over the space of conditional probability distribution $p(y \mid x)$ (this conditional probability distribution specifies the probability that the channel produces an output y for a given input x). The second line of this equation specifies that the mutual information $I(x,y)$ must be less than a specified channel capacity $\mathcal{C}$. Equation 6 represents a standard poblem statement in rate–distortion theory (Berger, 1971). For readers unfamiliar with information theory, in the present case it only matters that the solution to this constrained optimization problem represents an information channel that minimizes expected cost according to a given cost function, subject to a specified constraint on channel capacity. Our model obtains a solution to this equation using an efficient numerical algorithm due to Blahut (1972).

With a cost function specified, it is possible to examine the predictions of the model by comparing the distribution $p(y \mid x)$ (the distribution of channel outputs for a given stimulus input) against empirical data. However, the model just described requires the specification of three parameters for each color category (μ_k, τ_k, α_k), along with a constraint on available memory capacity ($\mathcal{C}$).

Our approach is to estimate the mean (μ_k) and precision (τ_k) of each color category using data from the color identification experiment (Figure 1). To do so requires one additional modeling assumption. If a color category is defined by its central location μ and precision τ, it is necessary to describe how a single color is selected from this category as the best example of the category. Our current model assumes that color identification is also based on the softmax equation. Now however, rather than selecting between categories, the goal is to select a hue that best represents a particular category k. Mathematically, this is stated as

$$P(\theta \mid k) = \frac{\exp(\beta \cdot \phi_k(\theta))}{\sum_\Theta \exp(\beta \cdot \phi_k(\Theta))}. \quad (7)$$

This introduces one additional parameter, β, which controls the noise in color identification (as $\beta \to \infty$, the model deterministically selects the peak of the color category, μ_k).

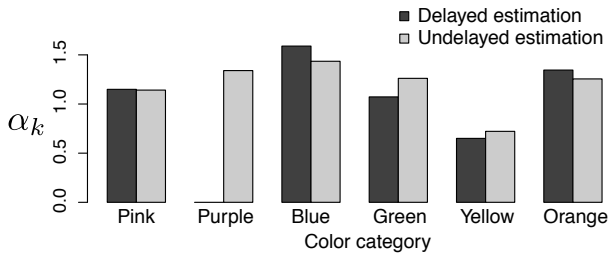


Figure 4: Color category strengths, α_k, fit to the undelayed and delayed estimation experiments.

In the model a single parameter β is used for all color categories. In summary, the parameters μ_k, τ_k and β were estimated from the color identification experiment. The best fitting parameters were determined by maximum likelihood estimation using numerical optimization. As shown in Figure 1, this model produces a close fit to the data.

With these parameters fixed, the category strengths (α_k) and capacity $\mathcal{C}$ were fit to the undelayed and delayed estimation experiments via maximum likelihood estimation. Model predictions are based directly on the optimal channel distribution $p(y \mid x)$ obtained from Equation 6. Corresponding model fits are shown in Figures 2 and 3. In terms of parameter estimates, the primary difference between the undelayed and delayed conditions is the estimated memory capacity. For the undelayed condition, channel capacity was estimated as 3.02 bits; for the delayed condition estimated capacity was 2.60 bits. These estimates are well in line with previous analyses of visual working memory capacity (Sims et al., 2012; Sims, 2015). What is notable is that in the undelayed experiment, capacity is still strongly limited even while stimuli remain continuously visible throughout each trial. This underscores the fact that perception has limited channel capacity in an information sense. In other words, there is always uncertainty in the interpretation of sensory signals.

The color category strengths, α_k are illustrated in Figure 4. Recall that these values determine the maximum strength of each color category, which in turn influences the probability of category assignment via Equation 4. The parameter estimates from the two experiments are highly similar, with one notable exception: in the delayed estimation condition of the experiment, the purple color category exhibits no influence over perception. This can also be seen in the histograms in Figure 2a. The reason for this difference between conditions is not clear; perhaps the influence of categories on perception shows strong individual differences, or perhaps the difference is due to the memory retention interval imposed by the delayed estimation experiment. Exploring this question remains a topic for future research.

Figure 5 shows the estimated cost functions for the undelayed and delayed experiments. The function $\mathcal{L}(x,y)$ is visualized as a two-dimensional heat map, where colors correspond to the cost of a particular perceptual error. Inter-

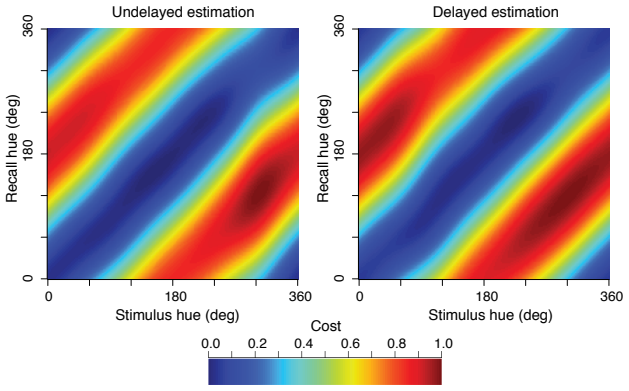


Figure 5: Cost functions estimated from the undelayed and delayed estimation experiments. The plots are re-scaled so that costs lie in the range (0, 1).

Comparison	Model	ΔBIC Delayed	ΔBIC Undelayed
(a)	Category + Metric	**0**	**0**
	Metric only	1159.71	294.92
	Category only	10682.77	7697.35
(b)	Cosine cost	151.94	246.93
	Parameterized cost	**0**	**0**

Table 1: Model comparisons based on Bayesian Information Criterion (BIC) scores, for the delayed and undelayed estimation experiments. Models with the lowest relative score (ΔBIC) are favored, indicated with bold type. Comparisons: (a) Examining whether category and metric costs of error are both necessary to account for performance. (b) Comparing the choice of alternative metric cost functions.

estingly, the cost function is highly similar between the two experiments; the change in performance is almost entirely due to the decrease in channel capacity in the delayed estimation experiment. According to the model, as capacity decreases, the relative influence of category errors on perception increases (as illustrated in Figure 2).

Model extensions and additional analyses

The model we have described so far assumes that the perceptual cost function is the sum of two terms, a metric and a categorical cost. Are both of these components necessary to explain performance in the information-theoretic model?

To answer this question, we fit two additional information-theoretic models: one that uses only a metric cost, and one that uses only a category cost. Relative performance of the three models was assessed using the Bayesian Information Criterion (BIC score). The results indicate that the model combining both categorical and metric costs is strongly supported over the two alternatives (BIC scores are reported in Table 1a; $\Delta\text{BIC} > 10$ is typically interpreted as strong support). This result is particularly important, as nearly all existing models of visual working memory ignore or overlook the influence of categories on perception (for further discussion, see Allred & Flombaum, 2014; Bae et al., 2015).

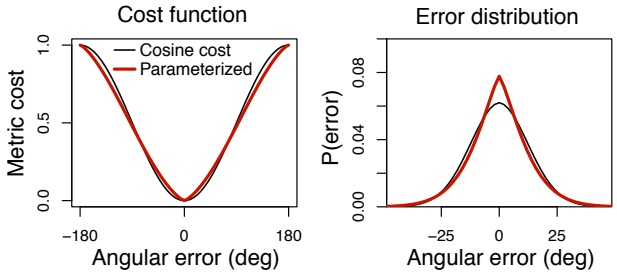


Figure 6: Left: Comparison of two different metric cost functions: a cosine cost function and a parameterized function fit to the data from the delayed estimation experiment. Right: Predicted error distributions for each cost function. In both cases a channel capacity of 2.75 bits is assumed.

One additional assumption of the model that we examine is the form of the metric cost function. The model described so far has assumed a cosine cost function (Equation 2). We developed an alternate model that includes category error costs, as well as a parameterized metric cost function of the form

$$f(y-x) = \frac{\delta z^{\gamma}}{\delta z^{\gamma} + (1-z)^{\gamma}}, \quad z = \frac{|\, y-x \,|}{\pi}, \tag{8}$$

where $|\, y-x \,|$ represents the absolute angular difference. This cost function introduces two additional parameters into the model, δ and γ, controlling the shape of the function. The category cost was left unchanged from the original model. Relative BIC scores comparing the cosine and parameterized metric cost functions are reported in Table 1b. The results indicate that the parameterized cost function offers a superior account for both the delayed and undelayed experiments.

The differing form and predictions of the two cost functions are illustrated in Figure 6. Although the two cost functions appear similar, the parameterized function predicts an error distribution with a sharper peak and heavier tails than a cosine cost function. This result is also in line with previous findings in the visual working memory literature (Sims, 2015; Van den Berg, Shin, Chou, George, & Ma, 2012).

Finally, we note that in the current paper model parameters were fit to the aggregated data from all participants. An important direction for future research is to examine individual differences in the influence of color categories as they relate to differences in memory capacity.

Conclusions

This paper argues that a useful approach for understanding perception is to understand the cost function that it seeks to minimize. We proposed a specific cost function for the perception and memory of color, based on a combination of a metric cost, and a cost associated with memory errors that cross category boundaries. This hypothesis was explored using a branch of information theory known as rate–distortion theory, which concerns optimal communication or information transmisison subject to strong limits on channel capacity.

We demonstrated that the application of this framework was able to provide a close quantitative fit to experimental data previously collected (Bae et al., 2015).

There is substantial evidence to suggest that categories can influence perception (Goldstone & Hendrickson, 2010). For example, Huttenlocher et al. (2000) found that observers' memory for the size of simple shapes is influenced by learned categories. Feldman, Griffiths, & Morgan (2009) conducted a rational analysis of perception of speech sounds in noise, and demonstrated that an optimal solution results in the reduced discriminability near prototypical vowel sounds. At the same time, there is conflicting evidence about the role of categories in color perception specifically. Witzel and Gegenfurtner (2013) argue that category effects are not inherent to perception, rather that these effects can be explained by attention to categorical distinctions which stem from linguistic category boundaries. We believe the model developed in the current paper may offer a productive theoretical tool for elucidating the influence of categories on perception.

The current model was partly inspired by an existing model of how categories influence color working memory (CATMET; Bae et al., 2015). Whereas CATMET assumes that color matching results from the integration of evidence from two independent information sources (category and metric representations), the current model assumes a single perceptual channel that optimizes a cost function with two terms. At present there is no consensus regarding the nature of high-level color representation in the brain to inform the selection between these two approaches (for related work see Stoughton & Conway, 2008; Wade, Augath, Logothetis, & Wandell, 2008; Bird, Berens, Horner, & Franklin, 2014). In the absence of clear physiological evidence for a dual channel representation, a one channel model is more parsimonious. A more meaningful distinction between the CATMET model and the current approach, however, lies in their intended roles as process-level and computational explanations for behavior, respectively (Marr, 1982). Unlike CATMET, the current approach explains biases in perception from the rational perspective of minimizing the costs of error.

A closely related approach is a Bayesian model of color perception developed by Persaud & Hemmer (2014). Their model assumes that color perception is the result of Bayesian inference, combining noisy sensory evidence with prior knowledge of color categories. Both Bayesian inference and rate–distortion theory offer normative theoretical frameworks. Bayesian models of perception typically make assumptions regarding the nature of noise that limits the fidelity of perceptual processing. Information theory, by contrast, assumes a limit on channel capacity; the nature of noise in the channel is optimized with respect to a given cost function. Hence, rate–distortion theory offers a more direct approach to studying perception as an adaptive system. We believe that the computational approach we are developing represents a promising framework for understanding color perception as a boundedly optimal system.

Acknowledgments

This work was supported by NSF BCS 0954749 (SA).

References

Allred, S. R., & Flombaum, J. I. (2014). Relating color working memory and color perception. *Trends in cognitive sciences*, *18*(11), 562–565.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall.

Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, *111*(12), 4590–4595.

Blahut, R. E. (1972). Computation of channel capacity and rate–distortion functions. *IEEE Transactions on Information Theory*, *IT-18*(4), 460–473.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of experimental psychology: General*, *129*(2), 220.

Marr, D. (1982). *Vision*. New York, NY, USA.

Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1162–1167).

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of vision*, *15*(3), 1–27.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological review*, *119*(4), 807.

Stoughton, C. M., & Conway, B. R. (2008). Neural basis for unique hues. *Current Biology*, *18*(16), R698–R699.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785.

Wade, A., Augath, M., Logothetis, N., & Wandell, B. (2008). fMRI measurements of color in macaque and human. *Journal of Vision*, *8*(10), 6.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of vision*, *13*(7), 1.

Abstraction in time: Finding hierarchical linguistic structure in a model of relational processing

Leonidas A.A. Doumas (alex.doumas@ed.ac.uk)
Andrea E. Martin (andrea.martin@ed.ac.uk)
Department of Psychology
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ, United Kingdom

Abstract

Abstract mental representation is fundamental for human cognition. Forming such representations in time, especially from dynamic and noisy perceptual input, is a challenge for any processing modality, but perhaps none so acutely as for language processing. We show that LISA (Hummel & Holyaok, 1997) and DORA (Doumas, Hummel, & Sandhofer, 2008), models built to process and to learn structured (i.e., symbolic) representations of conceptual properties and relations from unstructured inputs, show oscillatory activation during processing that is highly similar to the cortical activity elicited by the linguistic stimuli from Ding et al. (2016). We argue, as Ding et al. (2016), that this activation reflects formation of hierarchical linguistic representation, and furthermore, that the kind of computational mechanisms in LISA/DORA (e.g., temporal binding by systematic asynchrony of firing) may underlie formation of abstract linguistic representations in the human brain. It may be this repurposing that allowed for the generation or emergence of hierarchical linguistic structure, and therefore, human language, from extant cognitive and neural systems. We conclude that models of thinking and reasoning and models of language processing must be integrated—not only for increased plausiblity, but in order to advance both fields towards a larger integrative model of human cognition.

Keywords: computational models, sentence processing, analogy, relational reasoning, concepts, binding, temporal asynchrony, oscillations, computational neuroscience

Introduction

Abstract hierarchical representations are the hallmark of human language (Chomsky, 1957). Forming such representations is certainly necessary during language processing (see Martin, 2016 for a possible process model). But what are the computational origins of such an ability? One possibility is that the brain repurposed a mechanism or process already at its disposal when abstraction become an efficient solution to a problem posed by the environment, such as communicating information across time and space, or predicating novel arguments that you have never encountered before.

Ding et al. (2016) recently showed evidence of cortical tracking of abstract, hierarchical linguistic structures in oscillatory patterns in data from an electrocorticography (ECoG) and a magnetoencephalography (MEG) experiment. This tracking crucially could not be attributed to processing of acoustic information, transitional probability, or word predictability (Ding et al., 2016). Strikingly, LISA (Hummel & Holyoak, 1997, 2003), a symbolic-connectionist model of analogy making, and DORA (Doumas, Hummel, & Sandhofer, 2008), a symbolic connectionist model of relational reasoning, predict such a representational pattern. We tested the hypothesis that the representational patterns produced by LISA and DORA during their processing will give rise to the hierarchical structures matching the linguistic structures observed by Ding et al. (2016) without any formal or structural changes to the model. Such an approach would be particularly compelling because it shines a light both on how the brain might parse language (i.e., the class of possible parsing mechanisms underlying cortical tracking of linguistic representations as seen in Ding et al., 2016), and about how linguistic structures might have come to be the way there are. In order to test these predictions, we simulated oscillatory unit data in LISA/DORA using the same sentence stimuli as Ding et al. (2016). We tested whether LISA/DORA parsed the sentences correctly, and we observed the pattern of unit firing LISA/DORA exhibited while processing the sentences.

The LISA/DORA models

LISA (Hummel & Holyoak, 1997, 2003) is a model of analogy and relational reasoning. DORA (Doumas et al., 2008) is an extension of LISA that learns structured (i.e., symbolic) representations of relations from unstructured (holistic flat feature vector) inputs. That is, DORA provides an account of how the structured relational representations LISA uses to perform relational reasoning are learned from examples. Both LISA and DORA are symbolic-connectionist models, or models based on traditional connectionist computing

principles that effectively implement symbolic representations as they solve the binding problem and can, therefore, represent compositional structures (see Doumas & Hummel, 2005, 2012).

DORA accounts for over 25 phenomena from the literature on relational learning, as well as its development (e.g., Doumas & Hummel, 2010; Doumas et al., 2008; Lim et al., 2014; Morrison et al., 2012; Sandhofer & Doumas, 2008). In addition, as DORA learns relational representations, it comes to take LISA as a special case, and can simulate the additional 30-plus phenomena in relational thinking simulated by LISA. The description of LISA/DORA that follows is a brief overview due to space constraints. For full details of the models and their operations see Doumas et al. (2008) and Hummel and Holyoak (1997, 2003).

LISAese Representations In LISA (and DORA *after* it has gone through learning) relational structures are represented by a hierarchy of distributed and localist codes (see Figure 1). At the bottom, "semantic" units represent the features of objects and roles in a distributed fashion. At the next level, these distributed representations are connected to localist units (POs) representing individual predicates (or roles) and objects. Localist role-binding units (RBs) link object and predicate units into role-filler binding pairs. At the top of the hierarchy, localist P units link RBs into whole relational propositions.

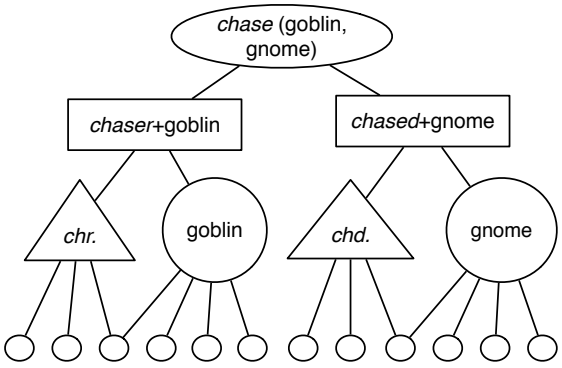


Figure 1. Representation of a LISA/DORA representation of the proposition *chase* (goblin, gnome). We use different shapes to represent units in different layers (ovals for P units, rectangles for RB units, triangles and large circles for PO units, and small circles for semantic units) for the purposes of clarity. In the model these units are simply nodes in different layers of the network.

Propositions are divided into two mutually exclusive sets: a driver and one or more recipients. In LISA/DORA, the sequence of firing events is controlled by the driver. We take the driver to be the focus of attention in LISA/DORA (i.e., what LISA/DORA is attending to at a specific moment). The driver contains one (or at most three) proposition(s). Activation flows from the driver units to their semantic units. Units in the driver and recipient are connected to the same pool of semantic units. Thus, units in the recipient become active in response to the pattern of activation imposed on the semantic units by the active driver proposition. The flow of activation from driver to recipient through shared semantic units is important for many of LISA and DORA's processes including comparison, analogical mapping, relation learning, schema induction, and memory retrieval. We will not discuss these processes further as they are not important for the purposes of the current paper, but full details may be found in Hummel & Holyoak, 1997, 2003) and Doumas et al. (2008).

Representing binding information What is most important about LISA/DORA for the purposes of the present paper is the manner in which the models solve the binding problem. As noted above, LISA and DORA are symbolic-connectionist models. That is, they are based on traditional connectionist computing principles (i.e., layers of interconnected nodes passing activation via weighted connections that are modified via Hebbian learning), but unlike traditional connectionist systems, they can process symbolic structure.

Processing symbolic structure requires that representational elements in a system can be composed into meaningful structures in a manner that does not violate the independence of those elements (see e.g., Markman, 1999; Russell & Norvig, 2003). For example, representing a relational proposition like *chase* (goblin, gnome) requires representing that *chasing*, a goblin, and a gnome are all present, and that goblin is bound the *chaser* role and gnome to the *chased* role. Importantly, the binding of *chaser* to goblin must not change the fundamental meaning of either what it means to be a goblin or what it means to be a *chaser*—i.e., the binding system must not violate role-filler independence.

In LISA and DORA roles and fillers are represented independently in the PO and semantic units. In order to behave symbolically, however, when a proposition in the driver becomes active, role-filler bindings must be represented dynamically on these units (i.e., POs and semantic units; see Hummel & Holyoak, 1997). Both LISA and DORA use time to carry this dynamic binding information.

Binding information is represented in LISA with bound role-filler pairs firing in synchrony. To illustrate, when a proposition like *chase* (goblin, gnome) becomes active in the driver (Figure 2a), the units representing *chaser* and goblin become active and fire together (representing the binding between *chaser* and goblin; Figure 2a[i]). Subsequently, the units representing

chased and gnome become active and fire together (representing the binding between *chased* and gnome; Figure 2a[ii]). Bound role-filler pairs fire together, and out of synchrony with other bound role-filler pairs. These distinct firing bursts allow LISA to code bindings between roles and their fillers, and process these structures symbolically, forming the basis of LISA's capacity to solve analogical mappings, and perform relational inference (see Hummel & Holyoak, 2003)

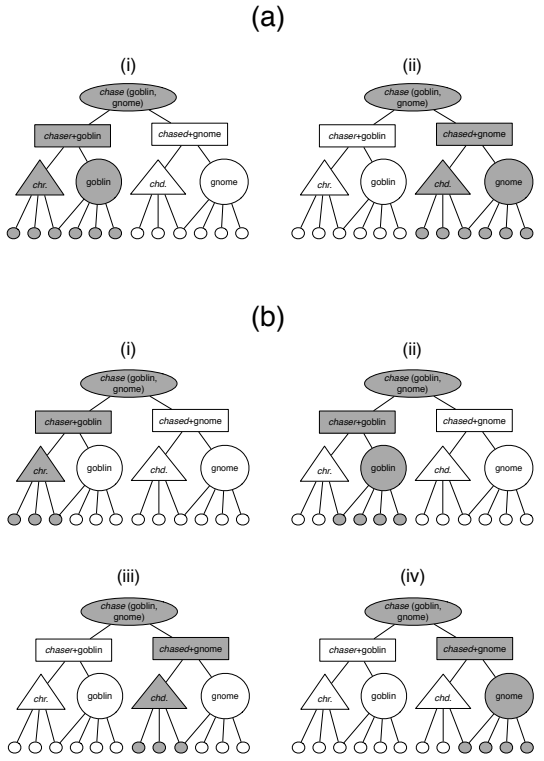


Figure 2. Dynamic binding in LISA and DORA. (a) Binding in LISA. (i) To bind the role *chaser* to goblin, units coding for *chaser* and goblin (as well as those coding conjunctively for *chaser*+goblin) fire. (ii) To bind *chased* to gnome, units coding for *chased* and gnome (as well as those coding conjunctively for *chased*+gnome) fire. (b) Binding in DORA. (i-ii) To bind the *chaser* role to goblin, units coding for *chaser* (as well as those coding conjunctively for *chaser*+goblin) fire, followed by the units coding for goblin (as well as those coding conjunctively for *chaser*+goblin). (iii-iv) To bind the *chased* role to gnome, units coding for *chased* (as well as those coding conjunctively for *chased*+gnome) fire, followed by the units coding for gnome (as well as those coding conjunctively for *chased*+gnome).

In DORA, binding information can be carried either by synchrony (as in LISA) or by systematic asynchrony of firing, with bound role-filler pairs firing in direct sequence.[1] During asynchronous binding, when a proposition like *chase* (goblin, gnome) becomes active in the driver (Figure 2b), the units representing *chaser* fire (along with units conjunctively coding for *chaser*+goblin and for the *chase* (goblin, gnome) proposition; Figure 2b[i]), followed directly by the units representing goblin (along with units conjunctively coding for *chaser*+goblin and for the *chase* (goblin, gnome) proposition; Figure 2b[ii]), representing the binding of *chaser* to goblin. Then, the units representing *chased* fire (along with units conjunctively coding for *chased*+gnome and for the *chase* (goblin, gnome) proposition; Figure2b[iii]), followed directly by the units representing gnome (along with units conjunctively coding for *chased*+gnome and for the *chase* (goblin, gnome) proposition; Figure 2b[iv]), representing the binding of *chased* to gnome. In short, bound role-filler pairs fire in direct sequence, and out of synchrony with any other bound role-filler pairs. These patterns of sequential oscillation dynamically code role-filler bindings in DORA, and underlie DORA's capacity to use the representations that it learns to support relational reasoning (e.g., analogical mapping, schema induction, and relational induction; see Doumas et al., 2008) and to learn structured relational representations from unstructured object representations.

Crucially, sequential firing of related constituent elements is a necessary property of binding via synchrony and systematic asynchrony. When LISA/DORA perform any structured processing, a pattern will invariably emerge wherein bound elements within a larger compositional proposition will fire in direct sequence and at a different time-scale than units coding for conjunctions of independently bound elements and full propositional compounds. In the following section we show that the pattern produced by LISA/DORA as it processes compositional structures matches very closely the temporal pattern of spike activity observed in Ding et al.'s (2016) when people process compositional propositions.

Simulation

Ding et al. (2016) presented auditory strings of synthesized speech in Mandarin Chinese in an MEG experiment, and strings of synthesized speech in American English in an ECoG experiment. They manipulated the structural relationship between the units in the auditory string, i.e., the syllables. In one condition, there was no meaningful relationship

[1] Asynchrony-based binding allows roles and fillers to be coded by the same pool of semantic units, which allows DORA to learn representations of relations from representations of objects (Doumas et al., 2008).

between the strings of syllables, in the second condition, phrases were formed from adjacent syllables, and in the third condition, sentences emerged from the string of syllables. Using this design, they observed peaks in the MEG-based oscillatory response on the timescale of syllabic rate (4Hz), phrasal rate (2Hz), and sentential rate (1Hz). Importantly, for trials with Mandarin sentences only speakers of Mandarin (compared to English speakers in a control group) showed tracking of phrasal and sentential representations in the form of peaks at 2Hz and 1Hz, respectively, although both English and Mandarin speakers showed tracking of speech/acoustic-syllabic stimuli regardless of language comprehension. The ECoG data from English speakers showed a similar pattern to the Mandarin MEG data, but without direct one-to-one acoustic-syllabic-to-phrase correspondence. Importantly, Ding et al. also observed cortical activity coding for sentence structure when English speakers tracked sentences of varying syllabic durations. For example, when English speakers tracked sentences with a noun phrase followed by a verb phrase wherein the initial noun phrase was three or four syllables (e.g., "the gold lamp", or "mahogany desk"), cortical activity tracked the entire phrasal structure, with a burst firing for the duration of the phrasal unit. Ding et al. controlled for effects of predictability in a string by showing that tracking of phrasal and sentence forms is not confounded by transitional probability.

Ding et al.'s results suggest definite structural form emerging during sentence processing. Specifically, beyond processing information at the level of syllables (or the basic features of a sentence tracked even by non-speakers of a language), speakers of a language process information that appears to track phrase structure and sentence structure. Moreover, when processing simple 2 argument verb structures, the structural pattern that emerges is two significant cortical response peaks (seemingly capturing phrasal information) firing within (at twice the rate) of a single cortical response peak (seemingly capturing sentence information).

We simulated the Ding et al. (2016) studies using the same English sentences used in their experiments 5 and 6 (with native English speakers). All of these sentences took the form modifier-noun-verb-noun, forming sentences like, "new plans give hope", and "dry fur rubs skin". LISA/DORA can represent sentences of this type in two ways. Most simply, such sentences could be represented with the modified noun represented as a single object containing both the semantics of the object and the modifier (see Figure 3a). Alternately, LISA/DORA can represent hierarchical propositions by representing propositional structures as arguments of other propositional structures. For example, to represent "dry fur rubs skin" the modified noun phrase "dry fur" can be represented explicitly by the propositional structure *dry*(fur), which can then serve as the argument of the agent role of the *rubs* relation (see Figure 3b; details of higher-order structure representation in LISA and DORA can be found in Hummel & Holyoak, 1997 and Doumas et al., 2008). We have previously hypothesised that LISA/DORA can alternate between these types of representation depending on the properties of the current task (e.g., Doumas et al., 2008; Rabagliati et al., submitted). Specifically, when modifier information must be considered explicitly, the later type of representation (as in Figure 3b) might be employed. Alternately, when the modifier information can be considered implicitly, the former type of representation (as in Figure 3a) can be employed. For the purposes of the current simulations both types of representations would work, however we used the hierarchical representations (i.e., Figure 3b) to code the sentences following our assumption that participants coded the modifier-noun structure explicitly.

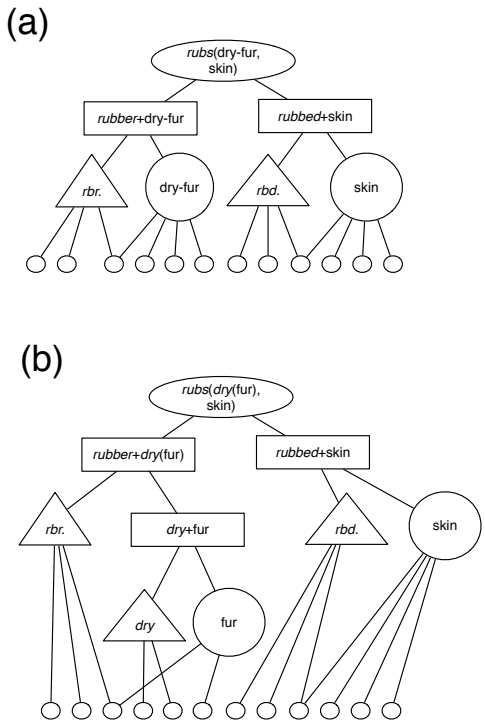


Figure 3. Representations of the sentence "dry fur rubs skin" in LISA/DORA. (a) A representation where the dry-skin modified noun is represented as a single unit connected to the semantics of both dry and skin. (b) A higher-order representation of the sentence where the modified noun is represented as a predicate structure, *dry*(skin) taken as an argument of the agent role of the *rubs* relation.

It is important to note that the DORA model can learn all of the representations used in the current simulation from experience. As demonstrated previously (e.g., Doumas & Hummel, 2010; Doumas et al., 2008; Hamer & Doumas, 2013; Lim et al., 2014; Sandhofer & Doumas, 2008), DORA can learn explicit structured (i.e., symbolic) representations of verb structures like *give*, *rubs*, or *chases*, and of single-place modifiers like *dry*, *new*, or *golden* from experience with objects in the world involved in those relations or with those feature. For the present study we hand-coded these representations, as the process of learning was not the focus of the current simulations.

To simulate Ding et al.'s experimental procedure we allowed LISA/DORA to process Ding et al.'s English sentences one at a time. Representations of the sentence structures entered the driver (i.e., were attended to). LISA/DORA processed the sentences as it normally would (i.e., the units fired to represent and encode binding information; see above). We tracked firing rate of all the nodes in the driver as LISA/DORA processed the sentences. Because of the controlled length and structure of the sentences, DORA, like the participants in the Ding et al. experiments, took the same amount of time to process each sentence. The results of the simulation and the comparison to the patterns observed by Ding et al. are presented in Figure 4. Interestingly, the pattern of firing of the nodes in the various layers of LISA/DORA very closely mirror the patterns observed by Ding et al. Specifically, just like the human participants, DORA showed an activation burst in it's P units that lasted throughout the processing of the sentence (i.e., firing at the 1Hz range), activation bursts at twice the rate of the whole sentence burst (i.e., the RB unit firing in the 2Hz range), and activation bursts at 4 times the rate of the whole sentence burst (i.e., the PO units firing in the 4 Hz range).

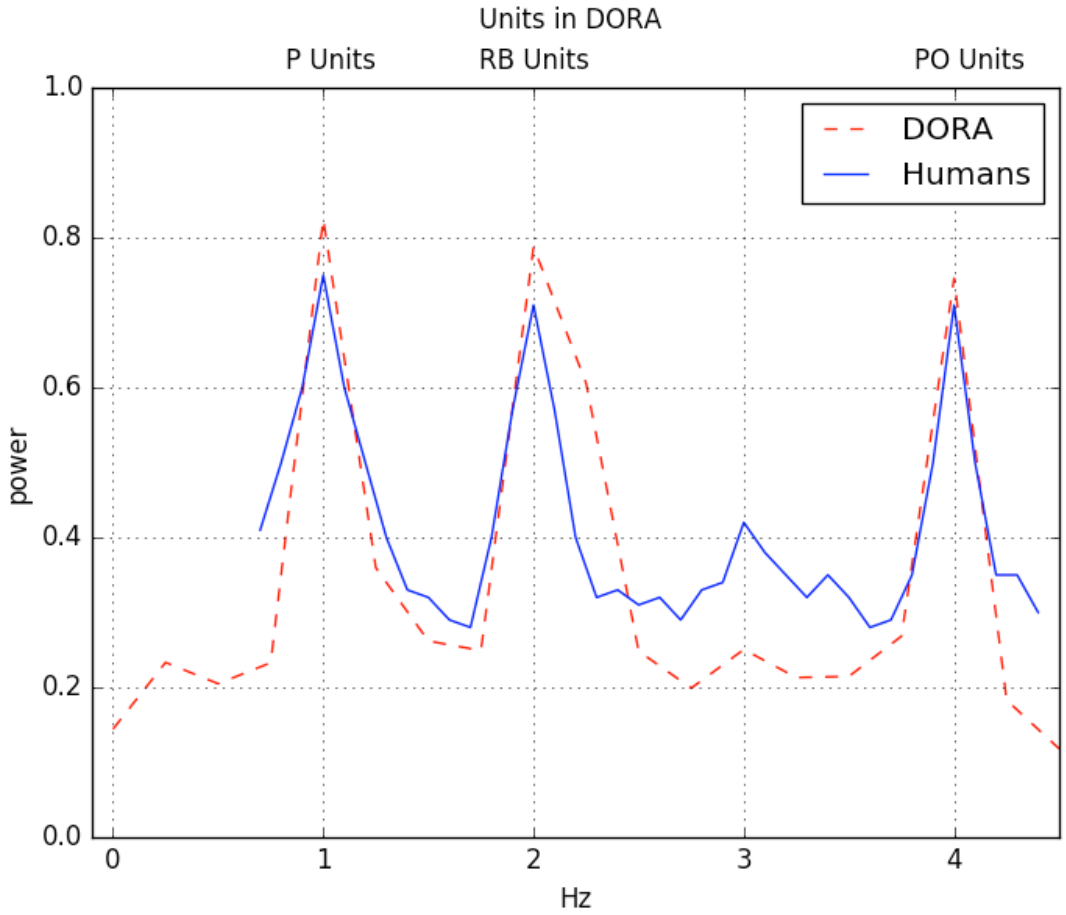


Figure 4. The solid line represents cortical power of participants listening to 4 word sentences played for 1 second in Ding et al. (2016). High cortical firing is evident at the 1Hz (the duration of the sentence), 2Hz, and 4 Hz range. The dashed line depicts firing in LISA/DORA while processing the same sentences used in Ding et al. There is evidence of units firing for the duration of the sentence, at intervals of half the length of the sentence, and at intervals lasting 1 quarter of the length of the sentence.

Conclusion

We have shown that abstract, hierarchical linguistic representation can be acquired, represented, and processed by LISA/DORA, models that were built for completely different purposes (analogy making in the case of LISA and relation learning in the case of DORA). Specifically, we have shown that the oscillatory activation patterns in LISA/DORA that arise as a natural consequence of the models performing dynamic binding appear to very closely fit data from human cortical tracking of hierarchical linguistic units (Ding et al., 2016).

It is interesting that models built for completely different purposes so successfully perform another task without modification. It is notable that extant models of sentence processing would likely not generalise so seamlessly to tasks such as analogical reasoning. Both probabilistic grammar (e.g., Levy, 2008) and connectionist approaches (see Joanisse & McClelland, 2015 for a review) must either be given a set of explicit grammatical phrase structure rules, or must learn the statistical specifics of a particular syntactic structure or parsing problem, rendering them unable to generalize to problems that routinely violate statistical and featural regularity like analogical reasoning (see, e.g., Holyoak, 2012).

We take our results to provide computational support the general claim—see, e.g., Penn, Holyoak, and Povinelli (2008)—that the ability to form and represent relational roles may underlie a number of our uniquely human cognitive capacities such as language. It is, perhaps, telling that the very same mechanisms that are necessary for processing relational structure and performing relational cognition seem to so closely simulate language processing.

Given our results, we suggest that models of relational reasoning and language processing might fruitfully be integrated. Such an integrative approach offers the possibility of producing powerful, neurophysiologically and cognitively plausible models that can perform well on multiple problems. We aim to further articulate the model by testing DORA on natural speech input, varied syntactic structures, in rich discourse contexts, on multilingual input, and with different assumptions about existing knowledge representations.

References

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1), 158-164.

Doumas, L. A. A. & Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. *Cognitive Science*, 34, 698-712.

Doumas, L.A.A., Hummel, J.E., & Sandhofer, C.M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*,*115,* 1-43.

Hamer, A. J. & Doumas, L. A. A. (2013). Discovering quantification and number in a role-filler model. *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220-264.

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3), 235-247.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.

Lim, A., Doumas, L. A. A., & Sinnett, S. (2014). Supramodal Representations in Melodic Perception. *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*.

Markman, A. E. (1999). *Knowledge representation*. Taylor & Francis: New York, NY.

Martin, A. E. (2016). Language Processing as Cue Integration: Grounding the Psychology of Language in Perception and Neurophysiology. In press at *Frontiers in Psychology: Language Sciences.*

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Brain and Behavioral Sciences*, 31, 109-178.

Russell, S. & Norvig, P. (2003). *Artificial Intelligence: A modern approach*. Pearson.

Sandhofer, C. M. & Doumas, L. A. A. (2008). Order and presentation effects in learning categories. *Journal of Cognition and Development*, 9, 194-221.

Modeling sampling duration in decisions from experience

Nisheeth Srivastava, Johannes Müller-Trede
UC San Diego, 9500 Gilman Drive
La Jolla, CA 92093 USA

Paul Schrater
University of Minnesota
Minneapolis, MN 55455 USA

Edward Vul
UC San Diego, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Cognitive models of choice almost universally implicate sequential evidence accumulation as a fundamental element of the mechanism by which preferences are formed. When to stop evidence accumulation is an important question that such models do not currently try to answer. We present the first cognitive model that accurately predicts stopping decisions in individual economic decisions-from-experience trials, using an online learning model. Analysis of stopping decisions across three different datasets reveals three useful predictors of sampling duration - relative evidence strength, how long it takes participants to see all rewards, and a novel indicator of convergence of an underlying learning process, which we call predictive *volatility*. We quantify the relative strengths of these factors in predicting observers' stopping points, finding that predictive volatility consistently dominates relative evidence strength in stopping decisions.

Keywords: response time; decision-making; evidence accumulation; sequential sampling; decisions from experience

Introduction

In orienting decision research from analyzing static economic choices towards more dynamic decisions akin to those people face in everyday life, the decisions-from-experience (DFE) paradigm presents an important step forward (Hertwig, Barron, Weber, & Erev, 2004). The DFE paradigm may be thought of as a modification of the classic certainty equivalence method commonly used procedure to elicit utility functions for money (see von Winterfeldt & Edwards, 1993). Participants in typical certainty equivalence experiments choose between a risky option that pays H with a probability p and L with probability $1-p$ and a safe option that always pays M, where $H > M > L$. They learn about the two options' payouts and the associated probabilities from explicit descriptions.

Decisions from experience modify this protocol: Participants are not shown payouts or probabilities, and must learn about them experientially. Several interesting observations emerge from research on DFE. Subjects sample more variable options and options with higher stakes for longer, for example (Lejarraga, Hertwig, & Gonzalez, 2012). They also appear to underestimate the probability of rare events (Hertwig et al., 2004), though much less so with increasing experience (Zhang & Maloney, 2012).

In the particular variety of DFE we consider throughout this paper, typically known as the sampling paradigm, participants are permitted to sample each of two options without consequence as long as they like, before finally committing to a binding choice. This protocol is particularly interesting since it closely mirrors the flow of information in many everyday settings–learn from the environment *ad libitum*, then make a choice. Importantly, such choices are actually composed of two decisions: a latent decision to terminate learning, and an overt decision to choose the risky or the safe option, based on the learned information.

Efforts to model the overt decision about which option to choose have been relatively successful (Erev et al., 2010). Little attention has been paid, however, to modeling the earlier latent decision about how long to sample information (or when to stop learning). Research on DFE has used simple statistical approaches as place-holders, assuming an underlying probability distribution over sampling lengths and fitting this distribution to the empirical distribution of sampling lengths observed in the data (Gonzalez & Dutt, 2011). Recently, Markant et al. (2015) proposed a model that jointly predicts choices and sampling length distribution. However, since their model uses *known* lottery stakes, it cannot speak to the effects of the specific characteristics of the actual, realized samples that comprise individual learning experiences in the DFE paradigm.

In this paper, we investigated variables that, on theoretical grounds, are expected to predict sampling lengths in DFE, without assuming *a priori* knowledge of lottery stakes and probabilities. Our analyses of three different datasets revealed the influence of two important variables on sampling duration. One is the difference in the options' expected values during sampling. The second is a measure of outcome uncertainty we call predictive *volatility*, which tracks abrupt changes in the magnitude of prediction error an observer experiences while learning about a DFE decision. We further developed a computational model of sequential sampling for DFE that uses these predictors to make accurate sampling length predictions for individual DFE trials.

Predictors of sampling duration

While largely neglected in the context of DFE, sampling durations have played a key role in research on perceptual decision tasks. Several of the leading theories in this domain utilize drift-diffusion models (Forstmann, Ratcliff, & Wagenmakers, 2016). In such models, the principal variable of interest tends to be the relative evidence strength in favor of either outcome (Bitzer, Park, Blankenburg, & Kiebel, 2014), measured in log posterior odds. The larger the log odds favoring a particular option, the more decisive–and quicker–the evidence accumulation in its favor. In DFE, relative evidence strength corresponds to the difference between the imputed value of the two options. Theories of perceptual decisions applied to DFE thus suggest that smaller differences in the options' imputed values should be associated with longer sampling durations.

Formally, we track the expected value difference (EVD) between the two gambles, measured at every sample for each

DFE trial,

$$EVD = pH + (1-p)L - M, \quad (1)$$

where $p = \frac{|H|}{|H|+|L|}$, and $|\cdot|$ is the number of times an outcome has occurred in the sequence up to the time at which the measurement is taken. In the following, we refer to this quantity as the EVD predictor.

Information theory, we suggest, points to a second, complementary factor that might influence sampling durations. From an information-theoretic perspective, the DFE observers' goal is to efficiently learn the reward rate of the gamble(s) they are sampling. The decision to terminate information gathering may then be seen as an agent's rational response to a learning procedure that has saturated. In practice, observers, while trying to learn useful models of their environment, have access to the prediction errors in such models. Unexpected increases in prediction error magnitude (as illustrated in Figure 1) signal the presence of unlearned environmental dynamics, stimulating rational observers to sample more.

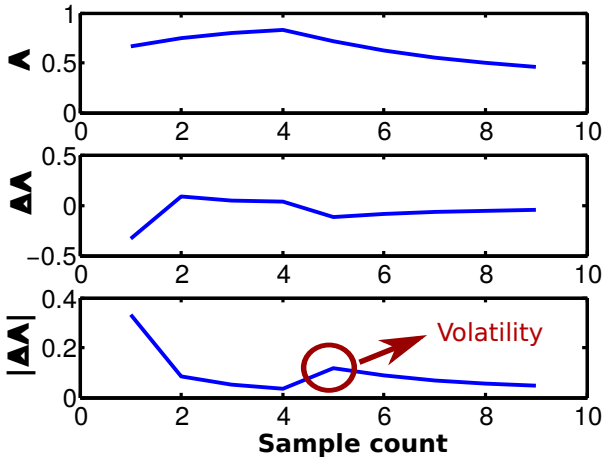


Figure 1: An intuitive view of predictive volatility. When the prediction error in a sequential learning process increases abruptly, it is reasonable to infer that the process has yet to converge. This indicator of the need to keep learning is what we call predictive volatility.

To understand the learning process, we can track the evolution of the learned parameter in a simple parametric observer model over individual sampling sequences. If a learning procedure is efficient, the prediction error is expected to show an asymptotic gradual decrease, reflecting that estimation is increasingly precise as more data are sampled. Critically, we assume that human observers are intuitive statisticians in this particular sense – they are implicitly aware that when learning is efficient, prediction error gradually declines. But in individual learning sequences, this decline is not always monotonic. We call deviations from the prediction error's expected trajectory episodes of *predictive volatility*. Observers who are sensitive to the expected trajectory of prediction error may rationally treat such episodes of volatility as evidence that learning has not yet completed, and respond by sampling for longer.

Formally, we model learning in DFE as a statistically efficient observer sequentially updating estimates of the mean parameter Λ of a Poisson distribution tracking the frequency with which the high outcome of the risky option occurs in the sampling sequence. With every new sample, the parameter estimate shifts by some quantity $\Delta\Lambda$. In absolute values, this quantity is mathematically–and, we argue, psychologically–expected to decay over time, so that $|\Delta\Lambda_t| < |\Delta\Lambda_{t-1}|$. Deviations from this trend constitute predictive *volatility*. Such deviations suggest to an observer that some aspects of the task may remain unlearned, and therefore justify continued sampling. In the case of human observers, the deviations must be sufficiently large to be noticeable, which suggests that volatility is perhaps best thought of as a binary variable which is either present or absent. This led us to formally operationalize volatility as

$$v(t) = 1 \text{ iff } |\Delta\Lambda_t| > \kappa \times |\Delta\Lambda_{t-1}|, \text{ and 0 otherwise.} \quad (2)$$

The constant $\kappa > 1$ determines the smallest noticeable deviation. All our analyses use $\kappa = 2$; substantially larger values would degrade the information present in this signal (since such large fluctuations are statistically infeasible in DFE reward rate estimation, where the set of possible outcomes is very limited); substantially smaller values would add noise to the signal, in the form of volatility false positives. Across an entire sampling sequence, the cumulative effect of such episodes is measured by the trial volatility load

$$V = \sum_t v(t), \quad (3)$$

and is referred to, in our following analysis, as the *volatility* predictor.

Finally, observers who know (i.e., have learned) that the standard DFE paradigm pits a safe option against a risky option (see above) may want to see all three reward outcomes at least once before terminating sampling. Depending on how skewed the risky option's odds are, this can take a relatively long time. The number of samples it takes to see all three reward outcomes at least once thus also contains valuable information about the length of the sampling sequence. It enters our predictive model in the form of a *counting* predictor, counting the number of samples it took a participant to see all three options at least once.

Results

We present two sets of results. We first demonstrate, using a proportional hazards regression analysis, that both the magnitude of difference in expected value and the amount of volatility seen in the sampling sequence influence sampling durations as predicted by our theory. Model selection reveals that volatility plays a more influential role in this process.

These results, however, are calculated using *post hoc* predictors that an actual observer would not have access to in

real-time. Our second set of results uses sequential counterparts to these predictors to develop a sequential model that simulates the trajectory of the stopping probability of any DFE trial, sensitive to the influence of both differences in expected value and episodes of volatility experienced in real-time.

Data

We procured data from two sources: the decisions-by-sampling condition from the Technion Prediction Tournament, which involved two sets (an estimation and a competition set) of 40 participants each solving 30 such problems, and a sample of 37 participants solving 19 different DFE problems we collected in the DFE condition of a different experiment (Experiment 2 in Lejarraga & Müller-Trede, 2016). We refer to the Technion estimation dataset as **TE**, the Technion competition dataset as **TC**, and our own sample as **LM**.

Experimental protocols were largely identical across the datasets.[1] Participants could sample both options in each lottery pair as often as they liked, and subsequently committed to one final draw that would correspond to their actual payout. All participants were compensated via a random incentive scheme, and earned real money corresponding to their payout in one randomly selected choice problem. We note that participants in LM revisited each choice problem in a group setting between individual trials (for details, see Lejarraga & Müller-Trede, 2016), whereas participants in TE and TC did not.

Volatility matters more

What can the expected value difference between options tell us about the decision to stop sampling? Recall that small EVDs may trigger further sampling, whereas large EVDs supply a reason to choose one option over the other (and thus to terminate sampling). The average EVD predictor, defined as $\frac{1}{T}\sum_t^T |EVD_t|$ for sequences of duration T, should then be negatively correlated with sampling lengths. If observers use the weight of economic evidence to decide when to terminate information search, greater average magnitudes of the expected value difference should correspond to earlier sampling termination and vice versa.

We tested this hypothesis by running a Cox proportional hazards regression, assessing the direction and magnitude of effect the average expected value difference measured during a sampling sequence has on the hazard rate across all trials per subject. The top panels in Figure 2(A) show that, as expected, EVD consistently increases hazard rates across participants in all three datasets.

We ran a similar proportional hazard regression using volatility load as a predictor of sampling sequence lengths. As the bottom row in Figure 2(A) demonstrates, volatility load consistently retards hazard rates in participants across all three datasets tested. Panel B in Figure 2 compares the regression coefficients obtained when we use both predictors–EVD and volatility–normalized and combining all participants in each dataset. In all three cases, the normalized predictors are uncorrelated ($r < 0.05, p > 0.25$), so the regression coefficients (β-weights) are informative about the relative importance of the two predictors (Nathans, Oswald, & Nimon, 2012). Volatility consistently dominates EVD as a predictor across all three datasets.

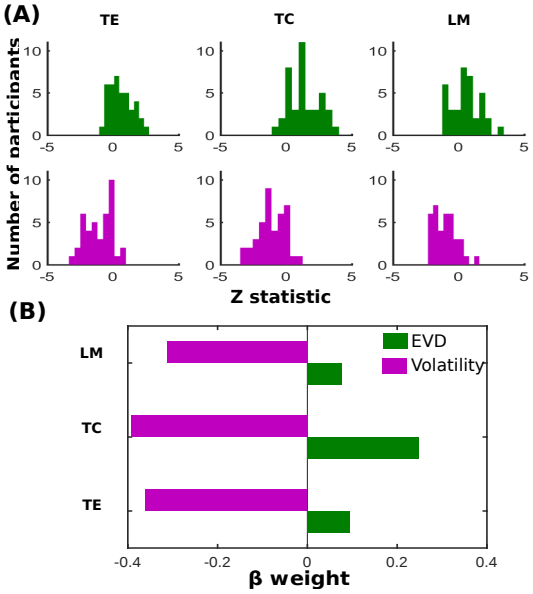


Figure 2: **(A)** Histograms of subject-wise Z-statistics obtained by Cox regression of predictors against sampling sequence lengths. Negative Z-statistics indicate that the predictor reduces the hazard rate, yielding longer sampling durations than baseline expectations. **(B)** Coefficients from Cox multiple regressions using normalized EVD and volatility predictors for all three datasets.

Table 1: Model selection using BIC. Lower values are better within individual datasets. ΔBIC from best model reported in brackets.

	TE	**TC**	**LM**
EVD	14312 (+165)	14643 (+177)	7090 (+63)
Volatility	14157(+10)	14529 (+63)	7031(+4)
Both	14147	14466	7027

A similar conclusion can be drawn by computing the Bayesian Information Criterion (BIC) for regressions using the two predictors indvidually, and then together. Not only is the full model preferable by BIC, corresponding ΔBIC values show that the volatility-alone model is considerably closer to the full model than the EVD-alone model, suggesting that it is a more powerful predictor. Together, these analyses demonstrate (i) that EVD and volatility are independent sources

[1] Participants in the LM experiment "chose" by alloting fractions of an allocation budget to either option; TE and TC used binary choices. Importantly for our purposes, the sampling procedure was the same in all three studies.

of information for predicting sampling duration and (ii) that volatility is a more informative predictor for sampling durations than EVD.

DFE sampling durations are *post hoc* predictable

How well can our account predict actual sampling durations in the DFE paradigm? A simple additive model combining these two predictors,

$$\text{duration} = \text{volatility} + \beta\ \text{EVD}$$

yields correlations $r = \{0.56, 0.53, 0.45\}$ with human sampling lengths for the TE, TC and LM datasets respectively, suggesting that these predictors can explain around 20-30% of the variance in sampling durations for human observers in DFE.[2] For lack of competitors–to the best of our knowledge, ours is the first model for predicting DFE sampling durations at the individual trial level–we cannot assess this performance comparatively.

To further improve predictive ability, we can add the counting predictor to the model. Its incorporation yields an augmented linear model

$$\text{duration} = \text{count} + \alpha\ \text{volatility} + \beta\ \text{EVD}$$

which improves the correlations with human data to $r = \{0.56, 0.69, 0.71\}$ for the TE, TC, and LM datasets respectively.[3]

Note that adding the counting predictor did not improve the data-model correlation for the TE dataset. This is because participants in this dataset strongly violated the expectation that participants would want see all three outcomes at least once. Of the 1200 total trials in this dataset (40 participants $\times$ 30 problems), as many as 742 trials (62%) were terminated without having seen all three outcomes, including 192 (16%) that were terminated after drawing just two samples altogether. For comparison, participants in the TC and LM datasets terminated 41% and 14% of all trials before seeing all three possible outcomes, respectively. We suspect that the higher sampling effort in the LM dataset may reflect additional intrinsic motivation participants in that study derived from the repeated social interactions between choice problems.

In the other two datasets, adding the counting predictor boosts the data-model correlation to ≈ 0.7, so the augmented model explains around 50% of the variance in those data. The large improvement in predictive ability somewhat overstates the predictor's true explanatory value, however. To see why, note that the sample count at which all three options have been seen once, by definition, cannot exceed the overall sampling sequence length. The counting predictor is thus upper-bounded by the independent variable it is used to predict. The resulting absence of counting predictor values greater than the actual sampling length inflates its correlation with the independent variable. Hence, while the counting predictor *prima facie* adds substantial predictive value to our model, it does so for reasons that need not be theoretically insightful.

[2] Best fit $\beta = 0, -0.1, -0.4$.

[3] For best fit values of $\alpha = \{3.8, 5.9, 2.6\}$ and $\beta = \{-0.2, -0.4, -0.9\}$ respectively.

Real-time stopping point prediction

While we show that sampling lengths in DFE are substantially predictable *post hoc* using objectively observable predictors, not all these predictors are available to decision-makers at the time of making their decisions. Neither the average EVD magnitude nor the cumulative volatility load across the complete sequence is available to an observer who is currently sampling. In the following analyses, we used elements of our predictors that are available to observers in real-time, and test how well they predict eventual stopping decisions.

To do so, we use computational models that perform the same sampling task as the observer, stepping through each trial sample by sample, predicting sequence lengths indirectly by estimating stopping probabilities λ_t at each sample. To make this analysis feasible, we make the simplifying assumption that there are no individual differences across participants within datasets. Doing so yields multiple data points for each DFE problem, instead of just one per problem-by-participant pair. This allows us to construct a stopping point distribution for each problem.

We then use these stopping point distributions to fit a basic piece-wise constant hazard model that assumes the stopping probability increases linearly with the sampling count t, i.e.,

$$\lambda_{t+1} - \lambda_t = \delta, \tag{4}$$

and fit $\{\lambda_0, \delta\}$ for each unique problem using a grid search, maximizing the statistical indistinguishability (measured using a two-sided T test; typical values $p > 0.95$) between the empirical stopping point distributions and the model's predicted stopping point distribution, averaged across multiple simulations (N=1000). We kept these $\{\lambda_0, \delta\}$ values fixed in the subsequent models we describe below, fitting only the additional parameters.

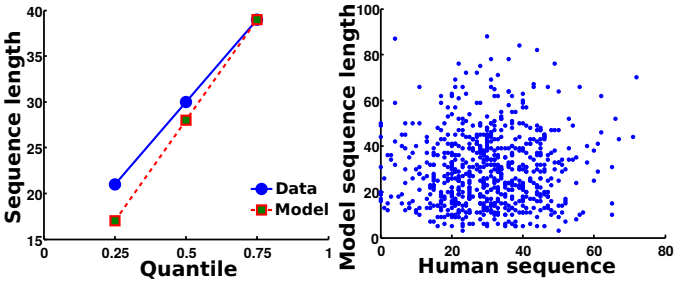


Figure 3: While simple statistical models of sampling lengths can reproduce population-level statistics *(left)*, they *(right)* fail to predict sampling sequence lengths for individual trials. Results shown for LM dataset.

This simple model theoretically and empirically resembles previous sampling length models proposed in the literature,

which assess model fit by testing whether it produces the same statistical distribution of sampling lengths as the underlying data (Gonzalez & Dutt, 2011; Markant, Pleskac, Diederich, Pachur, & Hertwig, 2015). Figure 3 illustrates that our baseline model closely approximates the empirical distribution of sampling lengths, much like existing models (compare left panel with Figure 1 in Markant et al., 2015). It is a poor predictor of sampling lengths at the individual trial level ($r = 0.03$, right panel), however, which illustrates a basic limitation of this modeling approach.

Table 2: Best fit correlations of sampling durations predicted by sequential models with human data in all three datasets.

Models	**TE**	**TC**	**LM**
Baseline	-0.04	0.02	0.03
Baseline + Vol	0.20	0.11	0.19
Baseline + EVD	0.15	0.11	0.06
Baseline + EVD + Vol	0.26	0.18	0.21
Baseline + EVD + Vol + Counting	**0.44**	**0.38**	**0.41**

Next, we added a volatility predictor to the model. In particular, we assumed that the baseline stopping probability would increase by Δ every time the observer encounters volatility in the sampling sequence,

$$\lambda_{t+1} - \lambda_t = \delta + v(t)\Delta. \tag{5}$$

Here, volatility refers to single episodes of volatility within a sampling sequence as defined in Equation 2, not the cumulative quantity ("load") measured across the entire sequence. If volatility retards the termination probability as predicted, negative values of Δ will yield greater correlations of the model's sample sequences with human data. As Figure 4 illustrates, observer models that reduce stopping probability when encountering volatility ($\Delta < 0$) indeed provide the best fit to the data in all three datasets. Our hypothesis about the influence of volatility thus finds clear support in the data.

We ran a similar analysis to measure the sample-by-sample impact of the EVD predictor. We assumed that incoming signals of greater relative evidence strength would affect the stopping probability following a logistic relationship, with larger values having a disproportionally larger effect. Thus,

$$\log \frac{\lambda_t}{1-\lambda_t} = \log \frac{\lambda'_t}{1-\lambda'_t} + k \log |d_t + 1|, \tag{6}$$

where λ'_t is obtained from Equation 4, k is fitted to the data to maximize the model-data correlation, and d_t is the EVD calculated using the sequence up to the tth sample. Table 2 provides a summary of the results.

To combine the influence of volatility and EVD, we revisited Equation 6 with λ' calculated via Equation 5, and with $\{\Delta, k\}$ as free parameters. The best overall model fit yielded weakly positive correlations across the three datasets.[4]

[4]Best fit parameter values for all three datasets: $\Delta = \{-0.20, -0.15, -0.125\}, k = \{0.02, 0.02, 0.02\}$.

Finally, when we incorporated the counting predictor into our model in the form of a real-time decision threshold–if all options seen at sample t, terminate with probability λ_t, otherwise, terminate with probability λ_0–the correlations improved substantially, to $\{0.44, 0.38, 0.41\}$, for the same parameter values as in the previous model. Unlike in the aggregate analysis, the effect of the counting predictor in sample-by-sample data does not necessarily suffer from the problem of artifactual inflation of correlation. These numbers thus present a fair picture of the predictability of sampling durations using only real-time information. Empirical estimates of the test-retest reliability of participants' sampling lengths would be required to assess how well our model's performance matches the best possible performance.

Discussion

This paper develops theory and algorithms to predict sampling durations in economic decisions from experience in which observers freely sample options before committing to a binding choice. We argue and then empirically demonstrate that a combination of evidence strength, predictive volatility, and simply tracking how long it takes participants to observe the entire reward structure of the particular DFE problem they are solving goes a considerable way in explaining individual decisions to stop sampling. Previous attempts to modeling information search in DFE succesfully reproduced distributions of sampling duration but were effectively random in trial-level predictions. Our account matches these previous attempts in distribution-level performance and surpasses them in making reasonably accurate trial level predictions. Finally, since it yields direct stopping point predictions, our model could easily be combined with choice models that respect the epistemic limitations of sampling-based DFE such as primed sampling or natural means (Erev et al., 2010) to make joint predictions of choice and sampling duration.

Of the three predictors we examine, one–the number of samples required to observe all possible outcomes–is specific to DFE. Its predictive power is substantial, which is interesting because it suggests that participants in DFE experiments both discern an aspect of the task structure not explicitly described to them (i.e., that each choice is between a safe option and a risky option with exactly two possible outcomes), and adaptively react to it. The other two predictors–evidence strength and volatility–are substantially more general and may thus be used to predict sampling duration in other experimental modalities. For example, Juni et al. have demonstrated that observers sample for longer when encountering noisier stimuli in a visuomotor estimation task (Juni, Gureckis, & Maloney, 2011). In this modality, greater stimulus noise corresponds directly to lower evidence strength, and as in our case, lower evidence strength is associated with longer sampling.

Our discovery of the significant influence of predictive *volatility* on sampling duration warrants further investigation. Predictive volatility is a highly accessible information signal

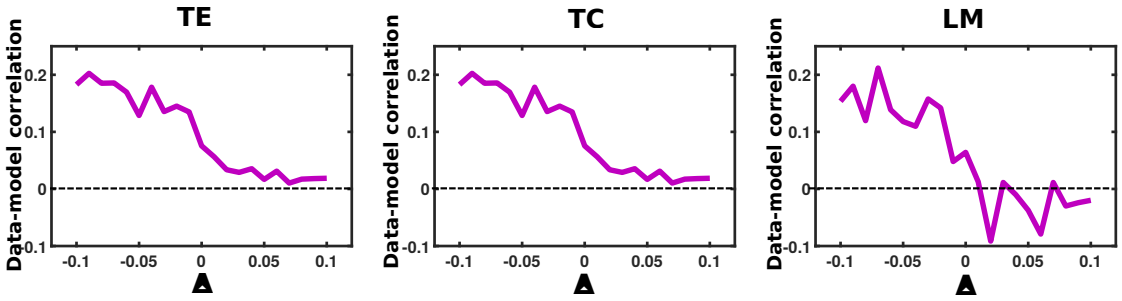


Figure 4: Model-data correlations for observer models fitted using different increments to stopping probability when encountering volatility.

that observers could draw on in a variety of real-world decisions from both experience and memory. To date, researchers have modeled response durations as arising from either evidence accumulation rising to a fixed threshold (Forstmann et al., 2016), or from a time-sensitive threshold collapsing to meet accumulating evidence (Thura, Beauregard-Racine, Fradet, & Cisek, 2012). Our results suggest that thresholds need not stay fixed or fall over time. Instead, they could rise and fall adaptively within trials in response to sequence-dependent predictive volatility. Volatility's representational generality, alongside our demonstration of its consistent and considerable impact on DFE stopping point decisions, invites further exploration in other experimental designs.

The role of predictive volatility in determining when to terminate sampling could also streamline the functional interpretation of cortico-striatal dopaminergic activity in decision-making (Schultz, Dayan, & Montague, 1997). Dopamine has been experimentally associated with encoding both reward and prediction error. The latter association appears to be more robust, however, in that it is congruent with a larger literature on the role of prediction error in multiple motor, cognitive and perceptual functions (Friston et al., 2012). Our account provides a rationale for why dopaminergic activity could be temporally correlated with the choice process without actually encoding reward: It may instead play a critical role in the latent decision to make a choice, and to terminate information search.

References

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a bayesian model. *Frontiers in human neuroscience*, *8*.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., ... Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, *67*, 641–666.

Friston, K., Shiner, T., FitzGerald, T., Galea, J., Adams, R., Sporns, O., et al. (2012). Dopamine, affordance and active inference. *PLoS Comput Biol*, *8*(1), e1002327.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Dont stop til you get enough: adaptive information sampling in a visuomotor estimation task. In *33rd Annual conference of the cognitive science society* (pp. 2854–2859).

Lejarraga, T., Hertwig, R., & Gonzalez, C. (2012). How choice ecology influences search in decisions from experience. *Cognition*, *124*(3), 334–342.

Lejarraga, T., & Müller-Trede, J. (2016). When experience meets description: How dyads integrate experiential and descriptive information in risky decisions. *Management Science (in press)*.

Markant, D., Pleskac, T. J., Diederich, A., Pachur, T., & Hertwig, R. (2015). Modeling choice and search in decisions from experience: A sequential sampling approach. In *37th annual conference of the cognitive science society.*

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, *17*(9), 1–19.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930.

Von Winterfeldt, D., & Edwards, W. (1993). *Decision analysis and behavioral research*. Cambridge, MA (USA) Cambridge Univ. Press.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*.

Experience as a Free Parameter in the Cognitive Modeling of Language

Brendan T. Johns,[1] Michael N. Jones, [2] & Douglas J. K. Mewhort[3]
btjohns@buffalo.edu, jonesmn@indiana.edu, mewhortd@queensu.ca
[1]Department of Communicative Disorders and Sciences, University at Buffalo, Buffalo, NY
[2]Department of Psychological and Brain Sciences, Indiana University, IN
[3]Department of Psychology, Queen's University, Kingston, ON

Abstract

To account for natural variability in cognitive processing, it is standard practice to optimize the parameters of a model to account for behavioral data. However, variability reflecting the information to which one has been exposed is usually ignored. Nevertheless, most language theories assign a large role to an individual's experience with language. We present a new way to fit language-based behavioral data that combines simple learning and processing mechanisms using optimization of language materials. We demonstrate that benchmark fits on multiple linguistic tasks can be achieved using this method and will argue that one must account not only for the internal parameters of a model but also the external experience that people receive when theorizing about human behavior.

Keywords: Cognitive modeling; Model optimization; Language processing; Corpus-based models.

Introduction

Models of cognition often have to deal with troublesome sources of variance that other fields (e.g., physical systems) do not. For example, no two individuals process a stimulus in the same way, and the same individual rarely processes the same stimulus identically at multiple times. In addition to individual differences and temporal stability, there is true random and measurement variance. While many of the sources of variance can be represented by free parameters, much of what may be systematic variance ends up being encapsulated by an overall noise parameter, often thought to reflect the inherent stochastic nature of the response process (Shiffrin, Lee, Kim, & Wagenmakers, 2008).

Almost every cognitive model contains free parameters, coefficients that are initially unknown, but are estimated from the observable data. The exact values for free parameters do not change the model's architecture—the theory that the model formalizes should be independent of its parameter values—but the settings do change a model's behavior. Hence, researchers use estimation methods to find the set of parameters that maximize a model's fit to data, and those parameter estimates are often allowed to vary across different data sets to which the model is applied.

A tacit assumption in cognitive models is that behavioral differences across individuals or tasks can be explained by differences in process parameters. But an alternative source of variance, often ignored, comes from differences in the subject's individual learning history or variance in memory representations selected for a task, independent of changes in the process parameters.

Theories of cognition commonly assume that aspects of the external world are stored internally. The storage assumption applies to memory (Anderson & Schooler, 1991), perception (Barsalou, 1999), and linguistic organization (Landauer & Dumais, 1997). In effect, the assumption acknowledges that human beings are embedded in a structured physical environmental that informs learning and that constrains behavior. As Simon (1969; p. 53) makes clear, "The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves".

Although we acknowledge natural variation in processing mechanisms, we explore in this paper the other source of variation—the environmental information to which people have been exposed—on lexical tasks. Subjects differ in what they know, and the differences should cause a corresponding change in behavior. Of course, the effect of variable knowledge depends on the specific task. For example, accumulated linguistic knowledge likely affects a lexical decision experiment more than a non-verbal perceptual identification task, so including linguistic knowledge when modeling lexical decision makes a good deal of sense.

One way to build linguistic information into a model is to use a representation of word meaning constructed from a standard corpus, such as the TASA corpus (introduced by Landauer & Dumais, 1997). The TASA corpus includes paragraphs from textbooks, from grades 1 to 12 and has been used as the gold standard in tests of co-occurrence models (e.g., Landauer & Dumais, 1997; Jones & Mewhort, 2007); it has frequently been integrated into processing models in cognate areas (e.g., Johns, Jones, & Mewhort, 2012). Although the TASA corpus is likely representative of the linguistic experiences that subjects have experienced, it is not intended to map exactly onto the experiences of specific individuals. Indeed different groups of subjects may have had exposure to wildly different linguistic sources, depending on culture, geography, educational system, and so forth. Hence, for any group of subjects, there is a natural variation in their knowledge, variation that should impact the behavior of those subjects on specific laboratory tasks.

Paradoxically, most theorists recognize that knowledge is central to performance in standard laboratory tasks but rely on a single corpus to model cognition. By relying on a single corpus, they ignore an important source of variability, namely the different knowledge that individuals bring to the task. If one could estimate a group of subjects' average linguistic experience, it should be possible to account for behavioral data at a more refined level.

The present article describes a new method for taking into account both the internal parameters of a model as well as the environmental information that defines a subject's unique

experience. Unfortunately, one cannot track a subject's linguistic history. As a proxy of that history, however, we collected a very large diverse set of language materials and combined it with a simple mechanism to uncover the most informative texts for a specific set of data.

To demonstrate the generality of our method, it was applied to both lexical organization and lexical semantic data. Modelling each of these tasks requires information to be accumulated into the mental lexicon, after which it will be acted upon to accomplish a specific task. The aim is to show that, by combining experiential fitting with realistic cognitive models across multiple areas, benchmark accounts of language-based behaviors can be attained.

Corpora and Data Fitting Methodology

To estimate the type of linguistic information used in a certain behavioral task, we use a wide variety of large language sources. These sources are then split into smaller sections, and it is iteratively determined which sections maximized the fit of a model to a set of data. At the end of the iterative process, the algorithm will have determined the most informative set of texts needed to explain performance on the task. Here, we describe the training materials and specifics of the method used to fit the models.

Training Materials

The texts come from five different sources: (a) Wikipedia (Shaoul & Westbury, 2010), (b) Amazon product descriptions (attained from McAuley & Leskovec, 2013), (c) 1,000 fiction books, (d) 1,050 non-fiction books, and (e) 1,500 young-adult books. All of the books were attained from e-books, and the vast majority were written in the last 50 years by popular authors. The set of sources—from an online encyclopedia to books targeted at young adults, to marketing materials for a large range of products—was designed to represent a broad set of possible experiences that an individual might have with written language. It is impossible, of course, to span the entire range of possible linguistic information, but the present materials represent a substantial range of texts, one that should give experiential fitting a fair test. To equate each source's contribution, each was trimmed to six million sentences, for a total of 30 million sentences across all texts (approximately 400 million words).

The data-fitting method will determine which set of texts is the most informative for fitting a particular experiment, just as statistical methods are used to estimate the optimal free parameters of a model. The corpora were split into small sections of 50,000 sentences yielding 120 sections for each corpus, for a total of 600 different sections across the corpora. Each section is large enough to allow for a measure of how much linguistic information the section contains, but is small enough that the different sections can still be combined to determine an optimal set of language.

Data Fitting Methodology

The goal of the data-fitting algorithm is to determine the combination of the sources that gives the best fit for a specific model to a set of data. To do so, we used a hill-climbing algorithm iteratively to select the sections that maximize the model's likelihood of generating various behavioral datasets.

A hill-climbing algorithm is an iterative local search algorithm, where a model is fit by incrementally improving its fit to a set of data. Once an increase in fit is no longer possible, the algorithm terminates. For experiential fitting, the first iteration selects the section that provides the best fit. Subsequent iterations add additional sets on top of the previously selected sections, to construct an overall training corpus. Once a section has been selected, it can no longer be used (sampling without replacement). Hence, the training materials increase their resolution continuously, in correspondence with the structure of the set of data attempting to be modeled. Fitting ends when the addition of a further section into the overall corpus does not increase the fit of the model to the data. To avoid getting stuck on local maxima, 10 unique starting points were made, in a rank order of the best fitting sections. The best fit will be displayed in the below simulations.

Discussion

To explore the power of experiential fitting, a large amount of text was assembled across a number of different sources. To determine the optimal set of linguistic data to explain a set of data, the texts were split into smaller pieces, and a hill-climbing algorithm was used iteratively to find the selection of text that maximally increased the fit of a model to a set of data. One could think of the process as a kind of parameter fitting (see Shiffrin, et al., 2008), but instead of optimizing the internal parameters to explain a set of behavioral data, the procedure optimized the structure of the external world (i.e. linguistic information). Optimizing the linguistic information allows us to determine the power gained by accounting for the variance in linguistic experience to an explanation of human behavior. That is, if linguistic behavior is related to the structure of linguistic experience, determining the optimal set of language materials with which to train a model should provide a substantial increase in the fit of the model.

Lexical Semantics

Models of semantic memory, particularly Latent Semantic Analysis (Landauer & Dumais, 1997), have strongly influenced studies of the effect of linguistic experience. LSA showed that a simple averaging mechanism, when combined with sufficient amounts of language information (derived from a large text corpus), can construct a representation of the meaning of words that is closely matches how people use language.

The model used here is derived from BEAGLE (Jones & Mewhort, 2007), a random vector accumulation model. The BEAGLE model is based on using sentential information in the learning process. In this model, words are represented by two vector types: a static environmental vector, that represents the perceptual (visual/auditory) aspects of a word, and dynamic context/order vectors, which mark both co-

occurrence and simple syntactic usages of a word. Each time a word is seen in a corpus of text, the dynamic vectors are updated. For context information, updating is done by summing the environmental vectors of the other words that occurred in the sentence with it (with high frequency function words removed). Accordingly, the context representation accumulates pure co-occurrence information. Order vectors, by contrast, accumulate rudimentary syntactic information, by recording the position of words that surround the usage of a word. The original BEAGLE model used circular convolution to form an n-gram representations of a sentence. Here, we will use a simplified form of the model (see Recchia, Sahlgren, Kanerva, & Jones, 2015), because the simplified form is less computationally expensive. We refer the reader to Recchia, et al. (2015) for a complete description of the simplified form of BEAGLE. In the following simulations, order, context, and the complete (the sum of the order and context vectors) representations were used.

The model was tested using two different data types: (a) synonym tests, and (b) item-level semantic priming. Following Landauer and Dumais (1997), synonym tests have become a standard test for models of semantic representation. In the synonym test, subjects are required to pick the word from a set of four that is most similar in meaning to a target word. A real-world example is the Test of English as a Foreign Language (TOEFL). Landauer & Dumais used 80 questions from the TOEFL and reported that LSA achieved an accuracy of 55% on this test.

Semantic priming will also be analyzed, a type of data that semantic space models have had success in accounting for (Jones, Kintsch, & Mewhort, 2006). In behavioral experiments, subjects are asked to perform simple tasks, such as lexical decision, but the target word is preceded by a prime. The prime can be semantically related or not, and the benefit provided by the prime is measured in terms of the speedup seen when the target is preceded by a semantically related item versus a semantically unrelated word.

Hutchison, Balota, Cortese, & Watson (2008) have shown that models of semantic representation may succeed at the mean level across items but fail at the individual word level. They examined priming in lexical decision for 300 different items and found that semantic variables offered minimal fits to the data, with forward association strength having the best correlation to overall levels of priming at $r = 0.164$, $p < 0.01$, while LSA had a non-significant correlation of $r = 0.053$. Clearly semantic priming data are difficulty to account for at an item level; hence, the data provide an excellent test for the power of experiential fitting.

Data Fitting Methodology

As noted earlier, all corpora were split into sections of 50,000 sentences and vector sets were generated for all the three types of information created by BEAGLE (context, order, and complete). The hill-climbing algorithm selected semantic vectors to maximize the model's performance on the TOEFL test, and rank correlation in semantic priming. That is, the necessary semantic information was refined iteratively to maximize the fit to the data. Iterative refinement halted when adding new material failed to increase the quality of the fit. To minimize the chance of falling into a local minimum during the fit, 10 random starts were used. To form a comparison, 50 resamples of the full corpus (of 30 million sentences), and the average performance increase across each 50,000 section of this corpus was recorded. This will provide a measure of how successful the model is independent of experiential optimization.

Results

The top panel of Figure 1 shows accuracy on the TOEFL test as a function of the number of sentences included in the fit; it shows results for the three kinds of information (item, order, and complete) along with a control condition in which the sections of text were assembled randomly. For the random corpora, the results were concatenated at 10 million sentences in order to aid in visualization. The complete model achieved the best performance at 97% accurate at only 1.1 million sentences. The context representation maximized at 92% accurate at 3 million sentences, while the order representation maximized at 82% accurate at 2.1 million sentences. For the random corpora, the average maximum performance was 57%, consistent with the past results (Jones & Mewhort, 2007). The complete representation is essentially performing at the same level as a native English speaker, an impressive level of performance for a rather simple model.

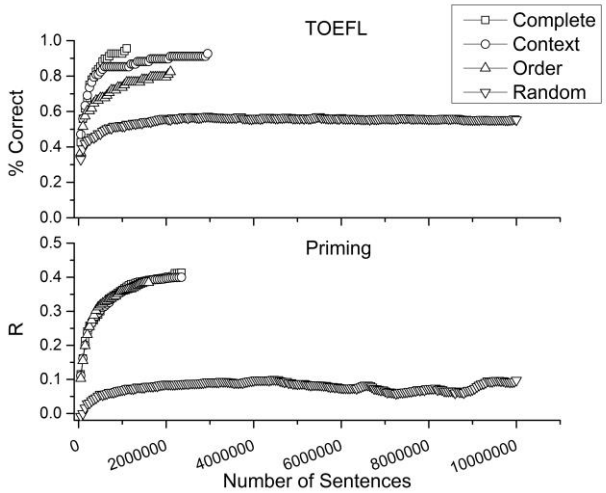


Figure 1. Results of BEAGLE and experiential optimization.

We also examined Hutchison, et al.'s (2008) item-level semantic priming results. Recall that it has been challenging for semantic-space models to account for item-level results. Figure 3 shows the fitted correlation as a function of the number of sentences for item, order, complete (combined) and random controls. As is shown in the bottom panel of Figure 1, all three representation types (item, order, and combined) provided a good fit to the item-level data in semantic priming. There was not a great deal of difference among the non-random representations, the complete model did offer the best fit at $r = 0.412$, $p < 0.001$. Note that the complete model provided a better fit than all the semantic

variables tested by Hutchison, et al. (2008). Indeed, it approached the fit of their 18-variable regression (r = 0.5).

Discussion

Semantic space models have been fundamental in exploring the influence of the linguistic environment on human behavior. This section explored the power that comes from combining a simplified form of a popular semantic space model (Jones & Mewhort, 2007) with experiential fitting, with the result being benchmark fits for every dataset analyzed. What this suggests is that the representations that people form in semantic memory are heavily influenced by the content of experience, and by constructing corpora that reflects this experience, better representations can be constructed.

However, one question about this method that needs to be determined is what source of variance is exploited in these simulations. It needs to be shown that the method is sensitive to group characteristics, where groups of subjects who have likely had different linguistic experiences, are found to have different corpora statistics by the experiential fitting method. That is, the method does not just exploit noise in the different datasets, but is actually approximates the type of experiences a group of subjects may have had.

Lexical Organization

A prominent area in the study of word recognition has focused on examining the influence of environmental variables on the retrieval of words from the mental lexicon. Classically, word frequency has been the most important lexical variable used to examine lexical retrieval, based on findings that higher frequency words are processed more efficiently. This has led to word frequency to be considered a central information type to models of lexical retrieval.

The exact nature of frequency effects has recently been questioned on several grounds. In one line of research, Adelman, Brown, and Quesada (2006) demonstrated that a measure that builds a word's strength in memory by counting the number of contexts that a word occurs in (operationalized as the number of document occurrences across a corpus) provides a superior fit to retrieval times than word frequency; this finding has been replicated across different corpora and datasets (Adelman, et al., 2006; Brysbaert & New, 2009). This measure is commonly referred to as a word's contextual diversity (CD).

However, Adelman et al.'s (2006) document count measure ignores the semantic diversity of the contexts that a word occurs in. To examine this possibility more closely, Jones, Johns, and Recchia (2012) conducted an artificial language learning experiment that manipulated word frequency and contextual diversity, such that certain words occurred with different sets of words (high semantic diversity), while others repeatedly occurred with the same set (low semantic diversity). Although there was no effect of diversity for low-frequency words, high frequency words were retrieved more quickly when they had been learned across multiple diverse contexts, indicating that processing savings occurred only with a change in context. On the basis of these results, and a corpus analysis, Jones, et al. (2012) proposed a new model that builds a more accurate measure of a word's strength in memory, entitled the semantic distinctiveness memory (SDM) model.

The SDM builds a word's strength in memory by weighting each new context by how much unique information that context provides about the meaning of the word. Across various corpora, this model was able to account for a larger amount of variance to a mega dataset of lexical decision and naming times over word frequency and a document count. Additionally, Johns, et al. (2012) demonstrated that the advantage for a semantic diversity count extends to spoken word recognition performance. Johns, Dye, and Jones (2016) have extended the results of the artificial language experiment of Jones, et al. (2012) with natural language materials and found similar results.

The simulations in this section will compare WF, CD, and SDM magnitudes, in combination with experiential fitting. The main source of data is 40,000 lexical decision times attained from the English Lexical Project (ELP; Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007). This is a standard dataset that has been used to differentiate different lexical information sources (Brysbaert & New, 2009; Jones, et al., 2012). Additionally, a set of 2,900 lexical decisions times for young and old adults attained from Balota, Cortese, and Pilotti (1999) was used to test the sensitivity of the experiential fitting method to different subject groups.

Data Fitting Methodology

Because the SDM uses paragraphs, the fitting method split each corpora into 3,000 paragraphs/documents (roughly equivalent to 50,000 sentences). For the Wikipedia corpus, this was a single document in the encyclopedia. For the Amazon product descriptions, one product description was considered a separate document. For the books, due to how they are formatted, there was no simple method to split them into paragraphs. Instead a moving window, with a size of 15 sentences, was used to assemble paragraph-like units.

Typically, the SDM is trained on a whole corpus, as the model is dynamic: previously experienced information is used to determine what should be stored for any new context. However, the model is quite computationally complex, so magnitudes were derived separately for each section. Overall magnitudes were then the sum of the different selected sections, which was also done for the WF and CD variables. These variables were transformed with a natural logarithm before assessing the correlation to the data.

Results

The results of the experiential fitting method on the z-transformed ELP lexical decision time data are displayed in the top panel of Figure 2. Only the results of the SDM are displayed in this figure, because all three measures produce similar results (explored further below). This result is contrasted with the fit that CD values from the SUBTLEX corpus (Brysbaert & New, 2009) provides for this data set, as

it provides the best fits to this data currently available. The figure demonstrates that the use of experiential fitting allows for a large increase in fit for retrieval latencies, even when compared against a very well-constructed corpus, as it outperformed SUBTLEX by a large margin. Additionally, the randomized corpora also achieved a correlation that equaled the SUBTLEX corpus (Brysbaert & New, 2009), demonstrating that the source materials that the experiential fitting method was using was of very high quality.

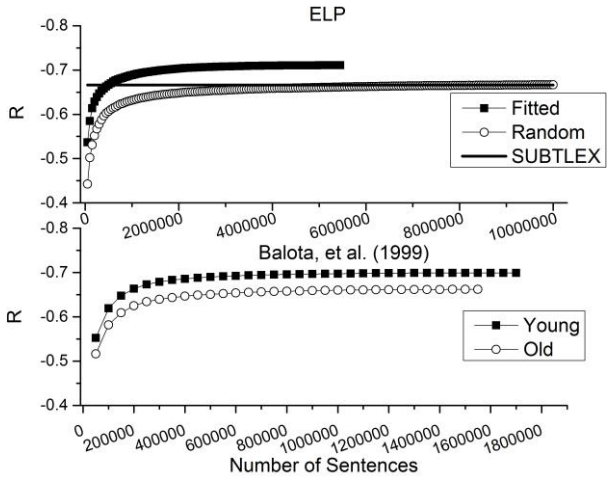


Figure 2. Results of SDM with experiential optimization.

As has been pretty previously shown, magnitudes from the SD model had the highest correlation, with an r =-0.708, p<0.001, compared with an r =-0.702, p<0.001 for contextual diversity, and r =-0.701, p<0.001 for word frequency. As a comparison, the correlation for CD values from SUBTLEX is an r=-0.666, p<0.001. The SDM model providing the superior fit is consistent with past results (Jones, et al., 2012; Johns, et al., 2012), but the interesting aspect of this simulation is the power that experiential fitting provided to all three variables.

As noted previously, there is still a question of what the source of variance that the method is capitalizing on, as it is possible that it is not capitalizing on group or individual characteristics, but instead random noise within the different datasets. As a test of this, 2,900 lexical decision times were attained from Balota, et al. (1999) for younger and older adults. The bottom panel of Figure 2 displays the fits to Balota et al.'s data for the SD model, and demonstrates that a high level of fit was attained for both subject groups, but with a higher fit to younger than older subjects, a standard finding. However, a more interesting analysis is to examine the composition of the resulting corpora for the two subject groups. To do this, the proportion of the different sources that was selected was recorded across 20 runs of the hill-climbing algorithm. These runs were done by removing the previously selected first section for the current run, so that each run begins differently. The results of this analysis are contained in Figure 3.

There was no difference in proportions selected for the non-fiction, Wikipedia, and Amazon sections, but there was a highly significant difference for the young adult sections [F(1,39)=203.51, p<0.001] and the fiction sections [F(1,39)=219.45, p<0.001]. These differences emerge because the young subject group had a higher proportion of young adult sections, while older adults were better described by the fiction sections. Given the composition of the different corpora, this suggests that the retrieval time data of these different groups are sensitive to the statistics of different linguistic sources that the subjects have experienced: young adults are better described by simpler examples of language as encoded in young adult books, but older adults are better accounted for by more linguistically diverse fiction and literature books. At least anecdotally, this is consistent with the type of linguistic experiences these subjects likely had.

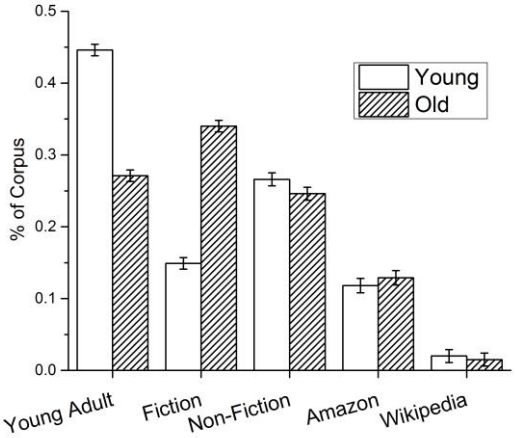


Figure 3. Proportion of different sections selected for young and old subjects.

Discussion

This section demonstrates that the use of experiential fitting can be expanded easily to examine lexical retrieval. There is a rich history of using environment variables (i.e. word frequency) to examine word retrieval patterns, with recent research pointing to the importance of contextual and semantic variables in the construction of a word's strength in the mental lexicon (Adelman, et al., 2006; Jones, et al., 2012). It was found that the SDM model, previously shown to provide a superior fit to large scale lexical decision data than word frequency or a document count, when combined with experiential fitting, provides a better accounting than previously published norms. Additionally, in an examination of young and older adult lexical decision data (Balota, et al., 1999), it was found that the method was sensitive to group characteristics, suggesting that the method is fitting to the experiences that a group of subjects may have had with language.

General Discussion

The current article describes a new method for optimizing cognitive models through experiential fitting, where the

information that a model "knows" is manipulated to provide the best fit to a set of data. The manipulation was done by assembling very large sets of texts spanning multiples areas, including an online encyclopedia, product descriptions from Amazon, and sets of fiction, non-fiction, and young adult books. These corpora were split into small sections, and a hill-climbing algorithm was used to determine the best combination of these materials for a specific model and set of data. It was demonstrated that this method, combined with experience-based cognitive models, provided benchmark fits to multiple types of lexical information.

The underlying philosophy of our method is similar to standard parameter fitting methods (Shiffrin, et al., 2008), which assume that there is natural variability in the parameters that define the cognitive processes that underlie behavior. Similarly, experiential fitting is designed around the idea that there is natural variability in the knowledge bases that different subjects groups have (and also in individual subjects) that leads to variability in behavior.

One of the exciting aspects of this technique is that it provides a mechanism by which to discriminate the varying contributions of internal cognitive mechanisms and external information, an old goal in the cognitive sciences (Anderson & Schooler, 1991; Simon, 1969). If one accepts that language is dictated by a complex interaction of biological and cultural evolution (Christiansen & Chater, 2008), then it is necessary to determine how much of the complexity in human behavior is derived from evolved mechanisms in the brain and how much is provided by the heavily structured environment in which humans are embedded. The simulations reported here provide substantial evidence that the information used to train a model is very important to a model's behavior, just as human behavior is sensitive to the knowledge that a person has gained. The simulation reported in Figure 3, where the experiential fitting method found different corpus constructions to explain younger and older adult's lexical decision data is a promising first step that group-level experiences can be estimated with this method.

More generally, this work points to the usefulness of building cognitive models around a learning mechanism that is capable of extracting information from large text-bases, an issue that has been explored in greater detail elsewhere (e.g. Johns, Jones, & Mewhort, 2012). By basing a model's performance in the learning of large-scale environmental information, it provides a stronger case for the plausibility of a model, as it is capable of scaling to levels of data input that a typical person may receive.

As Simon (1969) described, in order to provide a complete account of behavior, it is necessary to understand both the internal mechanisms and the environmental information that people use to behave. This is especially important in the study of language, as the vast majority of psycholinguistic theories have focused on the internal mechanisms that are responsible for linguistic behavior, while the influence of environmental information has been downplayed.

Downplaying environment information was necessary in early work because we lacked both large amount of texts and computational resources, but neither of these factors are limitations anymore. It is readily possible to examine the impact of linguistic information on human behavior, and by optimizing the linguistic information to which a model is exposed, it provides a powerful test of a model's ability to account for behavioral data.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision time. *Psychological Science*, 17, 814-823.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396–408.

Balota, D.A., Cortese, M.J., & Pilotti, M. (1999). Item-level analyses of lexical decision performance: Results from a mega-study. In *Abstracts of the 40th Annual Meeting of the Psychonomics Society* (p. 44). Los Angeles, CA: Psychonomic Society.

Balota, D. A., Yap, M J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39, 445-459.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.

Brysbaert, M., & New, B. (2009). Moving beyond Kucèra and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977-990.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489-558.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, 61, 1036-1066.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, 65, 486-518.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *Journal of the Acoustical Society of America*, 132, EL74-EL80.

Johns, B. T., Dye, M., Jones, M. N. (2016). The influence of contextual variability on word learning. *Psychonomic Bulletin and Review*.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55, 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66, 115-124.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

McAuley, J., and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (RecSys), 165–172.

Recchia, G. L., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. *Computational Intelligence & Neuroscience*. doi.org/10.1155/2015/986574

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, 42, 393–413.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.

Simon, H. A. (1969). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

An Information-Processing Account of Representation Change: International Mathematical Olympiad Problems are Hard not only for Humans

Takuya Matsuzaki (matuzaki@nuee.nagoya-u.ac.jp)
Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, JAPAN

Munehiro Kobayashi (munehiro-k@math.tsukuba.ac.jp)
University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8571, JAPAN

Noriko H. Arai (arai@nii.ac.jp)
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8634, JAPAN

Abstract

In this paper, we present a new information-processing model of math problem solving in which representation change theory can be implemented. Specifically, we divided the problem representation process into two. One is to straightforwardly translate problem texts into formulas in a conservative extension of Zermelo-Fraenkel's set theory, and the other is to interpret the translated formulas in local mathematical theories. A ZF formula has several interpretations, and representation change is thus implementable as a choice of an appropriate interpretation. Adopting the theory of real closed fields as an example of local theory and its quantifier elimination algorithms as an approximate process of searching for solutions, we develop a prototype system. We use more than 400 problems from three sources as benchmarks: exercise books, university entrance examination, and the International Mathematical Olympiad problems. Our experimental results suggest that our model can serve as a basis of a quantitative study on representation change in the sense that the performance of our prototype system reflects difficulties of the problems quite precisely.

Keywords: problem solving; information-processing model; insight; representation change

Introduction

Some math problems are much more difficult than others to solve even though they do not require higher levels of mathematical knowledge or techniques. Nine dot problem and mutilated draughtboard problem are examples of such problems. Where does the difficulty come from?

In classical information-processing models, the difficulty of a given problem is explained by its computational complexity: the cost of search (Kaplan & Simon, 1990; MacGregor, Ormerod, & Chronicle, 2001). In contrast, Gestalts explain the phenomena by the term *insights* (Isaak & Just, 1995; Ohlsson, 1992). A problem is called an *insight problem* when solving it requires a key feature of the problem to be recognized or restructured (representation change).

One of the major criticisms of classical information-processing account is that it has no mechanism to implement representation change since problem solving is understood as a search within a well-defined problem space (Öllinger, Jones, & Knoblich, 2014). If one tries to enlarge the framework (theory) of the problem to implement representation change inside it, then search space explosion is almost always inevitable. For example, it is a well-known fact that almost all the mathematical activities can be formalized in *Zermelo-Fraenkel's set theory* (ZF), thus representation change can be formalized as proof search in ZF. However, we cannot expect the search to be terminated in a realistic time since the search space of ZF is too vast. On the other hand, the representation change account also has some downsides. It does not provide any process model, and the analysis remains qualitative but not quantitative (MacGregor et al., 2001).

In this paper, we present a new information-processing model that enables us to include the representation change account. On the basis of the flow chart of insight problem solving (Öllinger et al., 2014), we first specify the perceptual process as translation of a given problem into a formula in ZF. We extend the language of ZF so that the translation is kept as straightforward as possible. In other words, we assume that the perceptual process requires no insight but rather corresponds to natural language and image processing. This is worth mentioning since the inputs of the existing information-processing account are usually not obtainable without insight regardless of the theories in which the problems are represented (Newell & Simon, 1972; Chou, 1988; Kerber & Pollet, 2006). The obtained ZF formula is considered to be the *primary* problem representation. There are usually many possible interpretations of the primary problem representation in different mathematical local theories. For example, the mutilated draughtboard problem can be embedded to not only propositional logic but also Peano Arithmetic and Presburger Arithmetic. The possible interpretations of the primary representation are called *secondary* representation.

We take the theory of real closed field (RCF) as an example of local theories and implement an interpretation process from the primary to secondary representation. We adopt a quantifier elimination (QE) algorithm as an approximate process of searching for solutions (Iwane, Yanami, & Anai, 2014) and develop a prototype system to solve geometry and introductory calculus problems.

We manually formalize more than 400 math problems from three different sources in our extended ZF language as a benchmark. The problems are translated so that they can be obtainable automatically from the problem text using state-of-the-art natural language processing theories and techniques (Kamp & Reyle, 1993; Steedman, 2001; Zettlemoyer & Collins, 2005). One source of the problems is exercise books, another is university entrance examinations, and

Suppose that x and y are real numbers. Find the range of a satisfying $x^2+ax+1>0$ for all x.

$$\forall x \in \mathbb{R}(x^2+ax+1>0) \quad \Leftrightarrow \quad a^2-4<0$$
$$\Leftrightarrow \quad -2<a<2$$

Suppose that x and y are real numbers. Find the range of a such that there exists x satisfying $x^2+ax+1<0$.

$$\exists x \in \mathbb{R}(x^2+ax+1<0) \quad \Leftrightarrow \quad a^2-4>0$$
$$\Leftrightarrow \quad a<-2 \vee a>2$$

Figure 1: Problem solving and quantifier-elimination

the other is the International Mathematical Olympiad (IMO). Though all the problems require mathematical knowledge and techniques no higher than high-school level, they have different levels of difficulty. We naturally assume that more insight problems can be found in the IMO than in the other two because IMO problems are known to be solvable by only a few mathematically talented students. The highlight of our paper is the experimental results on the benchmark. This is the first paper to report the automated problem solving results on not only a few problems or a set of artificial problems but a large number of real high-school-level problems.

Preliminaries

Let us first redefine what we mean by "mathematical problem solving." A math problem is usually expressed as a combination of sentences, formulas, and figures. In principle, it can be expressed as a logical formula in a theory. A theory consists of a set of symbols called a *language* and a set of *axioms*. A language consists of *constants*, *variables*, *relations*, *functions*, and *logical symbols*. Constants and variables are *terms*, and also $f(t_1,\ldots,t_n)$ is a *term* if f is an n-ary function symbol and all t_is are terms. $R(t_1,\ldots,t_n)$ is an *atomic formula* if R is an n-ary relation symbol and all t_is are terms. For example, $2x+1=y$ and $x>y+z$ are atomic formulas in arithmetic. In the first-order mathematical logic, *formulas* are defined recursively from atomic formulas and logical symbols. In classical logic, we have seven connectives: $\wedge$ (and), $\vee$ (or), $\neg$ (not), $\rightarrow$ (implies), $\leftrightarrow$ (if and only if), $\forall$ (for all), and $\exists$ (there exists). The last two connectives, $\forall$ and $\exists$, are called *quantifiers*. When a variable is quantified, it is called a *bound variable*. For example, the variable x is bound in the formula $\exists x(f(a)=x)$ though a remains *free* (not bound). A formula containing no free variable is called a *sentence*. A formula containing no quantifier is called *quantifier-free*. The set of seven connectives are known to be complete in a sense that any mathematical assertion can be expressed as a first-order formula provided that an appropriate language and a set of axioms are given.

A mathematical problem is *solved* when we find a formal procedure to show the problem is equivalent to a quantifier free formula of the simplest form. Fig. 1 gives examples.

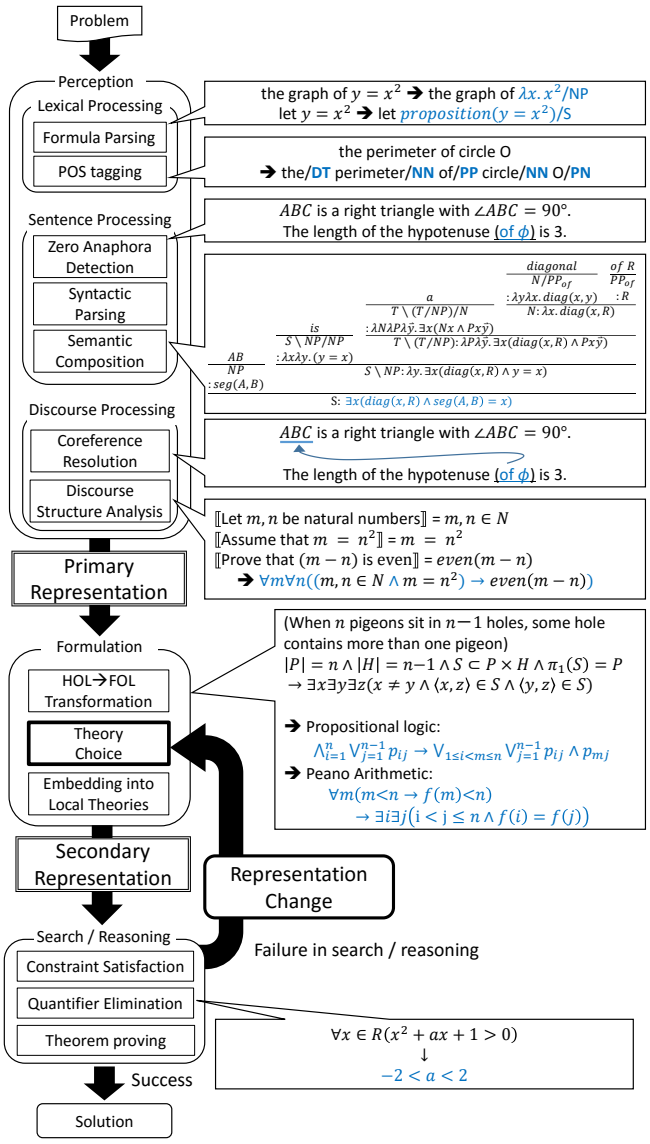


Figure 2: End-to-end problem solving model

Specifically, we say that a problem is *proved* when we show that a given problem is equivalent to True.

A theory is called *decidable* when there is an algorithm to determine whether any sentence is true or not. Gödel's incompleteness theorem shows that any theory containing Peano Arithmetic is undecidable.

Propositional logic, RCF, and Presburger Arithmetic are rare exceptions that are known to be decidable. However, computational complexity of the decision procedures is quite high. The theoretical lower bound of the decision procedure for propositional logic is superpolynomial to the size of input formulas assuming that P≠NP, and those for RCF and Presburger Arithmetic are doubly exponential (Tarski, 1951; Fischer & Rabin, 1974). These lower bounds reflect the phenomena of search space explosion.

An End-to-end Math Problem Solving Model

Fig. 2 presents an overview of our problem solving model. It consists of three modules. The perception module translates a problem into a primary representation expressed in ZF by language processing. The formulation module transforms the primary representation to another formula in ZF that is interpretable in a local theory such as RCF. Finally, the search/reasoning module works on the secondary representation. Once a failure is detected in the reasoning, the process backtracks to the formulation module and seeks another problem representation that makes the reasoning easier. The rest of this section provides more details on the three modules.

Perception Module

The perception module is organized along a hierarchy in natural language: words, sentences, and discourses (i.e., sequences of sentences). The lexical processing unit identifies the parts-of-speech and other syntactic properties of the words and math formulas in a problem. Since math formulas have their own grammar, they are analyzed by a specialized parser. Fig. 2 provides an example in which the same formula $y = x^2$ has different syntactic roles, noun phrase (NP) and embedded sentence (S), in accordance with its context.

The sentence processing unit translates each sentence in the problem into a formal representation. We assume a grammar-driven translation model here, which composes the semantic representation of a sentence along its syntactic structure (Carpenter, 1997; Heim & Kratzer, 1998). Specifically, we developed a Japanese grammar in the formalism of Combinatory Categorial Grammar (CCG) (Steedman, 2001). Fig. 2 depicts the process of semantic composition with CCG for the sentence "*AB* is a diagonal of *R*."

We need to detect omissions (zero pronouns) in the text before the semantic composition. Our current implementation detects them using a list of words and their syntactic arguments (i.e., case frames). Fig. 2 provides an example where an omission ("of ϕ," where ϕ, a zero pronoun, stands for *something*) is detected as the argument of 'hypotenuse.'

The discourse processing unit combines the sentence-level semantic representations into a single formula. We adopt the discourse representation theory (Kamp & Reyle, 1993) as the basic mechanism of the inter-sentential composition. Fig. 2 depicts an example where the semantic representations of three sentences are combined into one with the two connectives $\wedge$ and $\rightarrow$, and two universal quantifications ($\forall m \forall n$). The discourse processing unit also determines the antecedents of the anaphoric expressions including zero pronouns.

Formulation Module

The formulation module receives a primary problem representation and transforms it into a secondary representation that is amenable to reasoning. The process consists of two steps. One is the transformation of the higher-order formulas produced by the perception module to first-order formulas in ZF. The other is the transformation of the ZF formulas to those interpretable in a local theory, such as RCF and propositional logic. The former is usually a routine procedure for a person with the necessary math knowledge. The latter requires a target theory to be chosen beforehand. In the experiments, we chose RCF as the target local theory and confirmed that many pre-university math problems can be mechanically reformulated in RCF. This suggests that, once an appropriate local theory is chosen, the reformulation can be modeled as a heuristic search that seeks a formula in the local theory that is equivalent to the primary representation.

What is missing in our prototype implementation is a mechanism to choose an appropriate local theory. Our hypothesis is that it is the key ability in the representation change, which truly requires 'insight.' Our information-processing model thus serves as a test bed for *computational* models of insight problem solving by plugging-in a theory choice model to it. The rest of this section summarizes the implementation of the two reformulation steps and elucidates the contribution of the problem solving model.

Higher-order to first-order transformation The primary representation often includes higher-order elements (λ-abstractions), which denotes functions (e.g., $\lambda x.x^2$) and conditions (e.g., $\lambda y.(|y| < 1)$). They are necessary to translate the natural language expressions such as "The function that maps $x \in \mathbf{R}$ to x^2" and "The absolute value of y is less than 1. The same condition also applies to x." We eliminate such higher-order elements to obtain a first-order formula. In the current implementation, this is done by iteratively applying a handful of transformation rules such as β-reduction and variable elimination by substitution ($\exists x(x = \alpha \wedge \phi(x)) \Leftrightarrow \phi(\alpha)$).

Reformulation in RCF In the prototype implementation, a primary representation is rewritten into the language of RCF. The first-order language of RCF consists of polynomial equations and inequalities, logical connectives, and quantifiers. We developed a set of axioms that define various math concepts in the (higher-order) language of RCF, such as:

$$\forall x \forall f(\text{minimize}(x, f) \leftrightarrow \forall x'(f(x) \leq f(x'))).$$

The primary representation is iteratively rewritten with these axioms until an equivalent formula is found in the first-order language of RCF. There is no theoretical guarantee that such a formula will be eventually found even when it exists. We empirically examined how often it succeeds in the experiments.

Where in the process does *insight* come? The vocabulary of a problem usually tells us in which theory it should be solved. However, this is not always the case. For instance, the wording in the mutilated draughtboard problem does not suggest it should be formulated in arithmetic but not in propositional logic. Human solvers thus usually start by searching for the solution in propositional logic, putting dominoes on the board in trial-and-error manner. It is inevitable to change

Table 1: Subject areas of the benchmark problems

	Ex	Univ	IMO
Algebra	0	10	21
Linear Algebra	14	62	0
Geometry	81	65	94
Pre-calculus	0	75	0
Calculus	6	33	0
total	101	245	115

Table 2: Overall benchmark results

Problem Source	Solved (%)	Solved Time (sec) min/med/avg/max	Failed FM (%)	Failed TO (%)	Failed WR (%)
Ex	75.2	1 / 4 / 20 / 1069	7.9	13.9	3.0
Univ	65.3	1 / 7 / 38 / 1061	7.3	22.9	4.5
IMO	26.1	2 / 10 / 56 / 513	10.4	60.0	3.5

the representation of the problem to solve it in a realistic time. When and where does representation change happen in cognitive process? The main contribution of our processing model lies in pinning it down to a specific step in the problem formulation process, namely the theory choice.

Our information-processing model helps discriminate between different kinds of 'insight' problems. Nine-dot problem and mutilated draughtboard problem have been considered typical insight problems of the same kind. However, the reasons why people have difficulties are different in nature. Failure in solving nine-dot problem is at least partially due to the ambiguity of the term "line (segment)". Disambiguation of terms is a part of the perception process, but not of the formulation or representation change in our model. In contrast, they fail to solve mutilated draughtboard problem because they cannot choose an appropriate theory to solve it only from the superficial properties of the problem.

Solution Search/Reasoning

In the current implementation, we adopt a QE algorithm for RCF (Iwane et al., 2014) as an example for solution search. Note that we do not argue the QE algorithm per se is the model of human answer-deduction process. We utilize it to approximately measure the difficulty of mathematical reasoning on a given problem representation. The computational cost of the QE algorithm is quite sensitive to the problem representation; its time complexity is doubly exponential to the number of the variables in the representation. We regard a long running time of the algorithm as a sign of the *impasse* in the reasoning, which has been considered as a trigger of representation change (e.g., (Öllinger et al., 2014)). In the experiments, we examined to what extent this failure detection mechanism correctly reflects the difficulty of the problems.

Experimental Procedure

Aim of the Experiment

We developed a prototype implementation of the model described in the previous section. The theory choice process and representation change mechanism is not yet implemented. The aim of the experiment is to test if we can use the model as a basis for developing a computational model of theory choice and representation change. We thus need to verify: A) the model can solve many *non-insight* problems, which do not require representation change and B) the response of the model correlates with the difficulty of the problems. B) means the model is usable to *quantify* the difficulty of a problem in a certain representation, which is a requirement for a quantitative study on representation change.

Material

We collected more than 400 problems taken from three sources: exercise books (**Ex**), Japanese university entrance exams (**Univ**), and International Mathematical Olympiads (**IMO**). The **Ex** problems were sampled from a popular exercise book series. The problems in the books are marked with one to five stars in accordance with their difficulty: one to three stars signify textbook exercise level and four and five stars signify university entrance exam level. We sampled approximately the same number of problems from those marked with one, two, and three stars. The **Univ** problems were taken from the past entrance exams of seven top Japanese national universities. The **IMO** problems were taken from the past IMOs held from 1959 through 2014.

We examined the problems and exhaustively selected those that can be formulated (by humans) in the theory of RCF. The distinction between RCF and non-RCF problems was made solely on the basis of the essential mathematical content of the problems. The selected problems thus contain problems in several subject areas as shown in Table 1.

The problems were manually formalized in a higher-order language. Operators, who all majored in computer science and/or mathematics, were trained to translate the problems as faithfully as possible to the original natural language statements following the design of the perception module.

Experimental Results

The prototype system was run on the benchmark problems with a time limit of 3600s per problem. Table 2 shows the number of successfully solved problems; minimum, median, average, and maximum (wallclock) time spent on solved problems; number of failures in the reformulation of the primary ZF representation in RCF (FM); number of failures due to timeout (TO); and wrong answers (WR). Wrong answers were due to bugs in the current implementation.

Overall, the performances on the **Ex**, **Univ**, and **IMO** problems seem to well reflect the inherent differences in their difficulty levels. We conducted χ^2-test on the difference in the rates of success on them. The difference between **IMO** and other sources were statistically significant ($p < 0.01$) though that between **Ex** and **Univ** was not ($p = 0.09$).

We further examined how well the system performance correlates with the fine-grained difficulty level assessed by human experts. Table 3 lists the performance figures for

Table 3: Results for **Ex** problems by number of stars

#★	Succeeded		Failed	
	Success %	Time (sec) min/med/avg/max	FM (%)	TO (%)
1	82.4 (28/34)	1 / 4 / 5 / 39	11.8	5.9
2	73.5 (25/34)	2 / 4 / 6 / 39	5.9	11.8
3	69.7 (23/33)	2 / 4 / 51 / 1069	6.1	24.2
4	63.2 (24/38)	2 / 6 / 36 / 589	10.5	23.7
5	54.3 (19/35)	3 / 10 / 198 / 3245	2.9	42.9

Table 4: Syntactic profiles of the formalized problems

	Ex	**Univ**	**IMO**
# of $\forall$	2.2	2.0	5.8
# of $\exists$	5.3	9.3	3.1
# of λ	1.3	2.1	0.1
# of relations	12.5	19.8	13.8
# of functions	19.9	36.3	21.9
# of bound variables	8.8	13.4	9.1
# of free variables	3.0	3.1	1.8

the **Ex** problems (one to three stars) and additional problems sampled from those marked with four and five stars in the same exercise books. The overall correlation between the difficulty level and the system performance is clear although the difference in the success rates was statistically significant only between the problems with one star and five stars ($p < 0.05$, χ^2-test).

Analysis of the Experimental Results

Can we estimate the difficulty of a problem just by seeing it? If we can, the difficulty of the problems shall be attributed more to its inherent search cost (e.g., the time complexity determined by the number of variables) rather than the necessity of representation change. Table 4 presents several syntactic features of the benchmark problems. The figures are averaged over the problems taken from each source. It reveals that the syntactic features of the **IMO** problems are not very different from the exercise problems in **Ex** except for the distribution of variable binders ($\forall$, $\exists$, λ).

In addition to the basic features listed in Table 4, we may be able to estimate the difficulty of a problem by the vocabulary (i.e., distribution of function/relation symbols). To see this, we trained a binary classifier that predicts whether or not a problem can be solved by the prototype system in one hour. We used the features in Table 4 and the number of each symbol in a problem as the input and trained the classifier on the results of the benchmark test. Table 5 lists the precision and the recall of the classification obtained by 5-fold cross-validation. The definitions of the precision and recall are: precision = $TP/(TP+FP)$ and recall = $TP/(TP+FN)$, where TP (resp. FP) is the number of problems correctly (resp. wrongly) predicted 'solvable' and FN is the number of problems wrongly predicted 'unsolvable.' The overall prediction accuracy in Table 5 is way above the majority baseline of 57% but the accuracy is not very high especially on **Univ** and **IMO** problems.

Table 5: Accuracy of the solvability prediction

Source	Precision	Recall
Ex	88% (67/ 76)	93% (67/ 72)
Univ	73% (116/160)	78% (116/149)
IMO	57% (17/ 30)	47% (17/ 36)
All	75% (200/266)	78% (200/257)

The analysis presented above revealed that certain types of the difficulty are not captured by the superficial properties of the problems including the problem size and the vocabulary. This is a partial indication of the necessity of representation change or other kinds of insight for solving the problems. A future work is to examine such problems and clarify why they are difficult and what kinds of theory choice appear in human solutions of such problems.

Discussions

A first-order theory consists of a language and axioms. A formal theory is expressed in propositional logic, the first-order predicate logic, or higher-order predicate logic (typed lambda calculus). In our model, we set the primary representation expressed in the first-order ZF. Thus, there are three kinds of theory changes: axiom change, language (and axiom) change, and change from propositional to predicate logic.

Axiom Change

There are infinitely-many possible representations for propositional logic. Among them we can find analytic tableaux (cut-free LK), resolution, and Frege system (LK). The former two are the major systems used as the basis for automated theorem proving. Cut rule (axiom) allows one to introduce "lemmas" to prove theorems.

The pigeonhole principle is known to require exponential size proofs both in analytic tableau (Cook & Reckhow, 1979) and resolution (Haken, 1985), but it has polynomial-size proofs in Frege system (Buss, 1987). This is because we can introduce concepts of "addition", "subtraction" and "counting", and manipulate them to do some "restricted arithmetic" in Frege system. However, search-cost for appropriate cut-formulas is extravagant, and there is almost no hope that someone comes up with appropriate cut-formulas.

Another way to shorten proofs is to introduce a "symmetry rule" (Arai, 1996) as a new axiom. Propositional variable $p_{i,j}$ stands for "the ith-pigeon sitting in the jth-hole", and "$\bigvee_{j=1}^{n} p_{1,j}$" for "the first pigeon sitting in some hole" when expressing the pigeonhole principle in propositional logic (Fig. 2). If we have to check all the possible pigeons' positions, proofs blow up exponentially. Proofs will be shortened if we, without loss of generality, assume that the first pigeon sits in the first hole. In other words, "insight" realizing that a given problem has the property of symmetricity helps us to escape from an exhaustive search. There are some heuristics known to detect symmetricity, and it is implemented on computer (Arai & Masukawa, 2000).

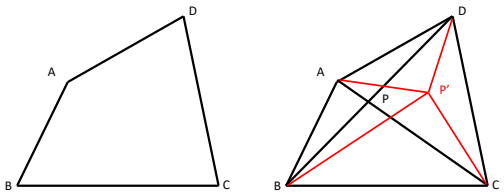


Figure 3: Solution to the quadrangle problem

Language Change

Elementary (Euclidean) geometry is known to be embeddable into the Cartesian coordinate system, and finally to RCF. However, languages and sets of axioms are different. As a result, the difficulties of problems do not remain the same.

Consider the following problem: "Let *ABCD* be a quadrangle. Find the point P that minimizes the sum of *AP*, *BP*, *CP*, and *DP*." Fig. 3 illustrates that the intersection of the diagonals minimizes the sum because of triangle inequality.

Insight may be required to line up the intersection of the diagonals as a candidate for *P*. However, the idea is easier to conceive when it is represented in Euclidean Geometry than in RCF since the intersection of the diagonals has salience in Euclidean Geometry.

Propositional or Predicate?

Mutilated draughtboard problem ($2n \times 2n$ version) is a good example of problems which is solvable when one changes the setting radically. The problem requires exponential-size proofs in resolution and analytic tableaux. It is not known whether or not it has short proofs in tableaux with symmetry rule. However, it has a short proof in arithmetic.

Conclusion

An end-to-end model of math problem solving has been presented. In the model, representation change is explained as the result of a choice of a local theory and the reformulation of a primary problem representation in it. Experimental results on more than 400 problems show that our prototype implementation reflects the difficulties of the problems quite precisely. Specifically, IMO problems require the system "theory change" more often than others when interpreting the timeout as "impasse". It indicates the model correctly captures the difficulty of the problems and hence it can serve as a basis of a quantitative study on representation change. Future work includes further analysis of the difficulty of math problems in light of our information-processing account and development of computational models of theory choice.

References

Arai, N. H. (1996). Tractability of cut-free gentzen type propositional calculus with permutation inference. *Theoretical Computer Science*, *170*(1), 129–144.

Arai, N. H., & Masukawa, R. (2000). How to find symmetries hidden in combinatorial problems. In *Proceedings of the eighth symposium on the integration of symbolic computation and mechanized reasoning.*

Buss, S. R. (1987). Polynomial size proofs of the propositional pigeonhole principle. *The Journal of Symbolic Logic*, *52*(04), 916–927.

Carpenter, B. (1997). *Type-logical semantics*. MIT Press.

Chou, S.-C. (1988). *Mechanical geometry theorem proving* (Vol. 41). Springer Science & Business Media.

Cook, S. A., & Reckhow, R. A. (1979). The relative efficiency of propositional proof systems. *The Journal of Symbolic Logic*, *44*(01), 36–50.

Fischer, M. J., & Rabin, M. O. (1974). Super-exponential complexity of presburger arithmetic. In *Proc. of the siam-ams symposia in applied mathematics* (Vol. 7, pp. 27–41).

Haken, A. (1985). The intractability of resolution. *Theoretical Computer Science*, *39*, 297–308.

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Wiley.

Isaak, M. I., & Just, M. A. (1995). Constraints on thinking in insight and invention. In R. J. Sternberg & J. E. Davidson (Eds.), *The nature of insight* (pp. 281–325). MIT Press.

Iwane, H., Yanami, H., & Anai, H. (2014). Synrac: A toolbox for solving real algebraic constraints. In *Mathematical software–icms 2014* (pp. 518–522). Springer.

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*. Kluwer Academic.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive psychology*, *22*(3), 374–419.

Kerber, M., & Pollet, M. (2006). A tough nut for mathematical knowledge management. In *Mathematical knowledge management* (pp. 81–95). Springer.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 176.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104) (No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In *Advances in the psychology of thinking* (pp. 1–44). Harvester Wheatsheaf.

Öllinger, M., Jones, G., & Knoblich, G. (2014). The dynamics of search, impasse, and representational change provide a coherent explanation of difficulty in the nine-dot problem. *Psychological research*, *78*(2), 266–275.

Steedman, M. (2001). *The syntactic process*. MIT Press.

Tarski, A. (1951). *A decision method for elementary algebra and geometry*. University of California Press.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proc. of the 21st conference in uncertainty in artificial intelligence* (pp. 658–666).

Inductive Ethics: A Bottom-Up Taxonomy of the Moral Domain

Justin F. Landy (justinlandy@chicagobooth.edu)
University of Chicago
5807 S Woodlawn Avenue, Chicago, IL 60637 USA

Daniel M. Bartels (bartels@uchicago.edu)
University of Chicago
5807 S Woodlawn Avenue, Chicago, IL 60637 USA

Abstract

Moral Foundations Theory (MFT) posits that people moralize at least six distinct kinds of virtues. These virtues are divided into "individualizing" and "binding" virtues. Despite widespread enthusiasm for MFT, it is unknown how plausible it is as a model of people's conceptualizations of the moral domain. In this research, we take a bottom-up approach to characterizing people's conceptualization of the moral domain, and derive a taxonomy of morality that does not resemble MFT. We find that this model more accurately reflects people's theories of morality than does MFT.

Keywords: morality; inductive reasoning; concepts; categorization; taxonomies

Introduction

How do people conceptualize the structure of the moral domain? Despite the recent explosion of research in the cognitive science of morality, a satisfactory empirical answer to this foundational question has not yet emerged.

An early attempt to understand conceptualizations of morality focused on the distinction between acts that violate moral principles, and acts that violate social conventions (Turiel, 1983). "Domain Theory" (DT) posits that moral violations concern "justice, rights, and welfare" (Turiel, 1983, p. 3), and that other prohibited actions, though condemnable, only constitute violations of convention, and could be permissible under alternative normative systems.

However, more recent evidence suggests that at least some people treat some acts that cause no objective harm and violate no rights as being truly morally wrong; people's moral domains are more complex than DT would suggest (Haidt & Hersh, 2001; Haidt, Koller, & Dias, 1993; Landy, 2016; Royzman, Landy, & Goodwin, 2014). The most prominent model of this complexity is Moral Foundations Theory (MFT; Graham, Haidt, & Nosek, 2009; Haidt & Joseph, 2004). In brief, MFT posits that there are at least six "moral foundations" – distinct, sometimes competing, virtues that we are innately prepared to moralize. These six virtues are divided into "individualizing" foundations that are concerned with the rights and welfare of individuals – harm prevention ("care"), fairness, and liberty – and "binding" foundations that are concerned with preserving the moral community – loyalty, respect for and obedience to authority, and bodily and spiritual purity ("sanctity"). Figure 1 presents the taxonomic structure of MFT.

MFT was developed by joining insights from cultural psychology and anthropology with evolutionary reasoning, and it has not been tested as a model of people's moral concepts. Thus, despite the theory's popularity, it remains unknown whether it is a plausible model of people's cognitive representations of the moral domain. In this research, we test the plausibility of MFT as such a model, using methods from the study of inductive reasoning. Specifically, we use people's inductive judgments about likely behaviors to derive a bottom-up model of people's taxonomy of the moral domain, and compare this with the theoretical taxonomy proposed in MFT.

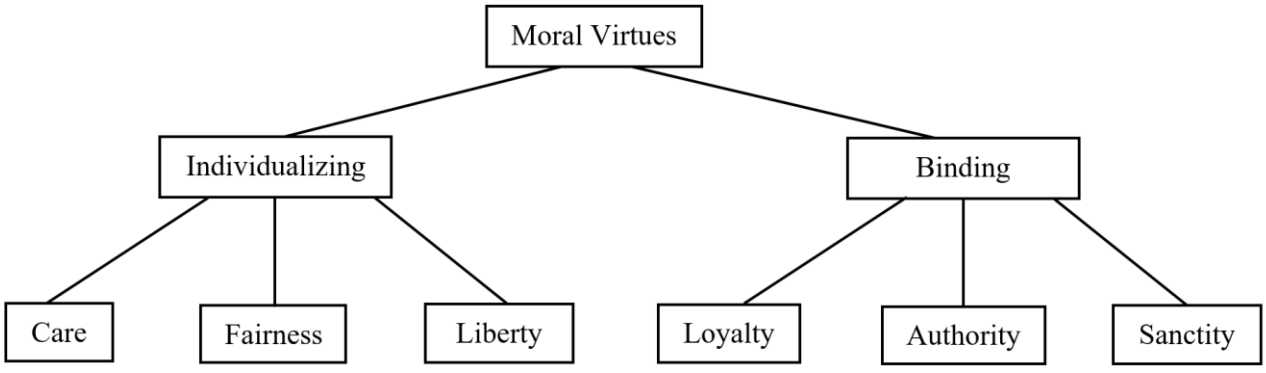


Figure 1: Moral Foundations Theory.

Category-Based Induction and Moral Virtues

The study of category-based induction (Osherson, Smith, Wilkie, López, & Shafir, 1990) is based on the premise that concepts are represented taxonomically, and we accept this as a working assumption here.[1] On this assumption, the strength of inductive inferences that a person makes from one object to another is indicative of how closely related the objects are in that person's taxonomy. An example, adapted from Osherson et al. (1990), illustrates this: consider the premise "Robins use serotonin as a neurotransmitter." Among the (infinite) possible conclusions, "sparrows use serotonin as a neurotransmitter" is typically considered more likely to be true than "geese use serotonin as a neurotransmitter." This is because robins and sparrows are closer to one another in people's taxonomies of birds than are robins and geese. More formally, there is greater premise-conclusion similarity in the former case than in the latter. Robins and sparrows might, for instance, both belong to the superordinate category "songbirds", whereas geese would belong to a separate superordinate category.

We applied this same logic to taxonomies of moral virtues, with the specific aim of testing the validity of MFT as a model of these taxonomies. Consider the premise "Joe commits a fairness violation." If MFT's taxonomy is a reasonable model of people's conceptualizations of the moral domain, then the conclusion "Joe would also commit a care violation" should be considered more likely to be true than "Joe would also commit a loyalty violation." Et cetera.

In Study 1, we had participants rate the likelihood that a person would engage in a wide variety of actions that exemplify the six moral foundations (conclusions), given information about a previous behavior (premise). From these likelihood ratings, we extracted a bottom-up taxonomy of the moral domain. In Study 2, we used a similar task, but also included ratings of the baseline likelihood of each behavior (i.e., the likelihood of a conclusion in the absence of any premise). The increases in inductive strength gained from the inclusion of premises more closely resembled the predictions of our taxonomy than MFT. Finally, in Study 3, we presented participants with two premises, rather than one. Participants' likelihood judgments more closely resembled the predictions of our taxonomy than MFT.

Stimulus Development

Stimuli in all three studies consisted of the long form of the Moral Violations Database – Severity Equated (MVD-SE), a subset of the Moral Violations Database (MVD), a set of nearly 250 behavioral descriptions normed on several criteria, including moral wrongness and representativeness of each moral foundation. Most of the stimuli included in the MVD are original, or are modified forms of the Moral Foundations Vignettes (Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015). The development of the MVD and MVD-SE is detailed elsewhere (Landy & Bartels, 2016), so we only briefly summarize it here.

The MVD-SE contains seven behaviors violating each foundation, drawn from the larger MVD. These behavioral descriptions passed a two-step validation process: one sample rated how well each behavior exemplified each moral foundation, then a second sample assigned each behavior to the foundation that it best exemplified, in a forced-choice task. Behaviors that were rated above the scale midpoint for a foundation by the first sample, and assigned to that foundation by a majority of the second, were considered validated. From these validated stimuli, seven were chosen to represent each foundation (e.g., a person drives past a man on an empty road who is clearly injured (care), hires their nephew instead of a more qualified job applicant (fairness), forces their daughter to enroll as a pre-med student in college (liberty), sends out an email calling their boss an "idiot" (authority), makes critical comments about their home country (loyalty), or looks at pornography in which an adult model has been digitally altered to look like she is 13 years old (sanctity)). These stimuli uniquely exemplify the moral foundations, and provide broad conceptual coverage of each one (e.g., the liberty stimuli include both overbearing parents and overreaching politicians). Moreover, the mean moral wrongness ratings for the foundations are extremely closely equated (5.20-5.26, on a 1-9 scale). The MVD-SE also includes seven non-moral actions, which extensive pretesting has found to be morally inert (e.g., "a person goes parasailing"), and seven counter-normative behaviors that do not exemplify any moral foundation (e.g., "while in a rush, a person bumps into someone on the street, but does not say 'excuse me'"). These counter-normative actions largely consist of violations of polite etiquette.

Study 1

Method

Participants (N = 367) were recruited online through Amazon Mechanical Turk in exchange for monetary compensation. The study was completed online.

Each participant made 64 likelihood judgments, one for each possible premise/conclusion combination of the eight conceptual categories in the MVD-SE (e.g., authority/authority, authority/non-moral, etc.). Premises and conclusions were randomly sampled from the MVD-SE for each question, with the restriction that the premise and conclusion could not be the same action. Questions took the following form: "A person hires their nephew for a job, instead of a more qualified applicant. Given this information, how likely is it that, if they were driving along an empty road and saw a man who was clearly injured, this person would drive past the man and not stop to help him?" (this is one of 7 premises x 7 conclusions = 49 possible

[1] There are, of course, numerous other models of how concepts are represented (see Medin, Rips, & Smith, 2005, for a review). We follow the majority of the research on category-based induction, which we see our methodological approach as deriving from, in focusing on taxonomic representations.

Table 1: Mean conceptual relatedness scores.

	Care	Fairness	Liberty	Authority	Loyalty	Sanctity	Non-Moral	Counter-normative
Care	44%	25%	22%	34%	29%	14%	14%	30%
Fairness		49%	23%	34%	31%	11%	18%	30%
Liberty			45%	18%	17%	10%	13%	18%
Authority				52%	32%	15%	16%	33%
Loyalty					45%	12%	14%	22%
Sanctity						36%	8%	14%
Non-Moral							36%	16%
Counter-Normative								41%

fairness/care questions). Likelihood ratings were made using a sliding scale (0% = "There is no chance this person would do this"; 100% = "This person would definitely do this").

Results

For present purposes, we computed a measure of conceptual relatedness between categories of actions by multiplying likelihood estimates to and from pairs of categories. For instance, if a participant rated the likelihood of committing a fairness violation, knowing that a person had committed a care violation, as 70%, and the likelihood of committing a care violation, knowing that a person had committed a fairness violation as 50%, that participant's fairness/care relatedness score would be 70% x 50% = 35%.

We submitted the mean relatedness scores (presented in Table 1) to a hierarchical cluster analysis using between groups linkage.[2] In agreement with the pattern of means in Table 1, violations of care, authority, fairness, and loyalty, and counter-normative actions were close to one another in Euclidean space and clustered together early in the analysis. In contrast, violations of liberty and sanctity, and non-moral actions were quite distant from all other categories. Figure 2 presents a dendrogram illustrating this analysis.

We confirmed this result by subtracting relatedness scores from 100%, and submitting the resulting dissimilarity scores to multi-dimensional scaling.[3] We restricted our analysis to a two-dimensional solution for ease of presentation, and treated the dissimilarity scores as ordinal variables.[4] As shown in Figure 3, and consistent with the above analyses, violations of care, authority, fairness, and loyalty, and counter-normative actions are quite close to one another in the resultant two-dimensional space, with liberty violations, and especially sanctity violations and non-moral actions, more distant. Model stress was .10, which is generally considered acceptable (see, e.g., Kruskal, 1964a, 1964b; Rosenberg, Nelson, & Vivekananthan, 1968).

These analyses converge on the conclusion that care, fairness, authority, and loyalty violations, and counter-normative actions, are quite closely related to one another in people's taxonomies of morality, while liberty and sanctity violations and non-moral actions are less closely related. It seems reasonable, therefore, to model the virtues of care, fairness, authority, and loyalty, along with politeness, as belonging to a single superordinate category. We conceptualize this category as "obedience to rules". We think that this captures what sets these virtues apart from liberty and sanctity – liberty has to do with not *creating* rules that are burdensome or oppressive for others, and

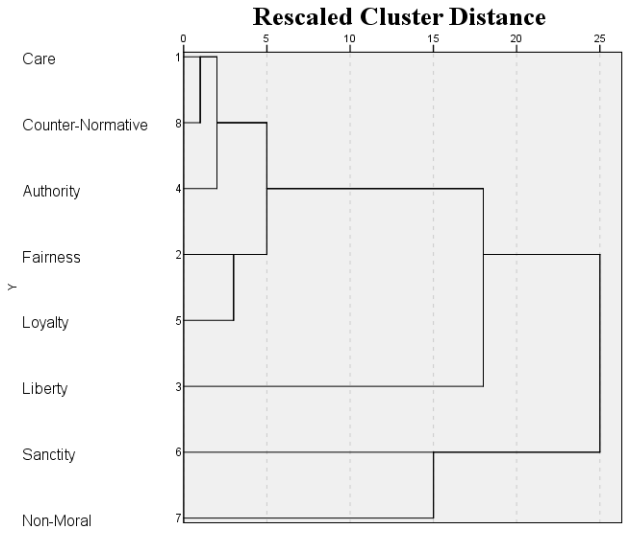


Figure 2: Dendrogram illustrating hierarchical cluster analysis of relatedness scores. X-axis represents squared Euclidean distances between agglomerated clusters.

[2] The results are essentially identical when Ward's method is used instead. We present the results of the analysis using between groups linkage because the resulting dendrogram makes the relationships between categories easier to visualize.

[3] Identical results are obtained if the dissimilarity scores are calculated by subtracting relatedness scores from the maximum observed relatedness (52%) than from the maximum possible relatedness (100%). We therefore focus on the conceptually simpler analysis.

[4] The pattern of results is the same – indeed, it is somewhat clearer – if the dissimilarity scores are treated as interval or ratio variables, however, the model stress is unacceptably high under these assumptions (.20 and .35, respectively).

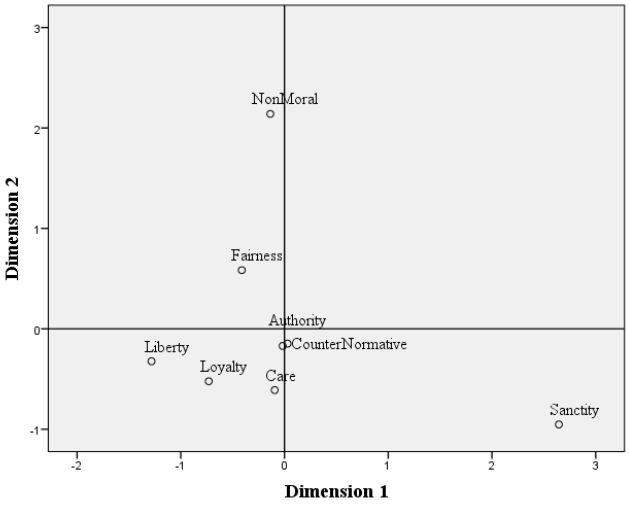


Figure 3: Two-dimensional solution derived from multi-dimensional scaling of relatedness scores.

sanctity violations tend to be so unusual that explicit rules forbidding them (e.g., "Thou shalt not write erotic poetry about thy cat") are probably rarely articulated. Our bottom-up taxonomy of the moral domain is presented in Figure 4. Note that this taxonomy does not resemble MFT, and that the individualizing-binding distinction did not emerge in our analyses of people's judgments.

Study 2

Method

Participants (N = 359) were recruited in the same manner as in Study 1. Participants from Study 1 could not take part in this study.

Each participant made 42 total likelihood judgments, one for each possible premise/conclusion combination of the six moral foundations, and six baseline likelihood judgments with no premise. Politeness and non-moral characteristics are not included in MFT's taxonomy; therefore we did not include the counter-normative and non-moral actions from Study 1 in Studies 2 and 3, as they are not useful for testing the relative predictive validities of the two taxonomies. Premises were randomly selected for each question. Rather than randomly select the conclusion for each question, however, each participant was randomly assigned one of seven conclusions from each foundation, which appeared in all likelihood judgments for that foundation. That is, each participant saw the same conclusion from every foundation seven times, so that their premised judgments were directly comparable to their baseline likelihood judgments. Likelihood ratings were made on the same sliding scale as in Study 1.

Results

We created a measure of inductive strength gained from knowledge of a premise by subtracting baseline likelihood judgments from premised judgments (e.g., if a participant rated the baseline likelihood of a person hiring their nephew instead of a more qualified applicant at 30%, and the likelihood of this, given that the person had driven past an injured man on an empty road without stopping to help, at 65%, the gain in inductive strength would be 35%).

Both taxonomies classify 18 premise-conclusion pairs as belonging to the same superordinate category (e.g., authority and care are both part of obedience to rules in our taxonomy, and loyalty and sanctity are both binding virtues in MFT), and 18 as belonging to different superordinate categories (e.g., liberty and sanctity in both taxonomies). Thus, we calculated the average inductive strength gained from premises that belong to the same superordinate category as the conclusion, versus premises that do not, in each taxonomy. We expected a Premise x Taxonomy interaction, such that more inductive strength would be gained from within-category premises, versus between-category premises, according to our taxonomy, versus MFT.

Means and standard deviations are presented in Table 2. Substantially more inductive strength was gained from within-category premises than between-category premises in both taxonomies (within-subjects ANOVA: $F(1, 358) = 469.46$, $p < .001$, $\eta^2_p = .57$), and there was no mean difference in inductive strength gained across taxonomies ($F(1, 358) = .036$, $p = .850$). However, as expected, the

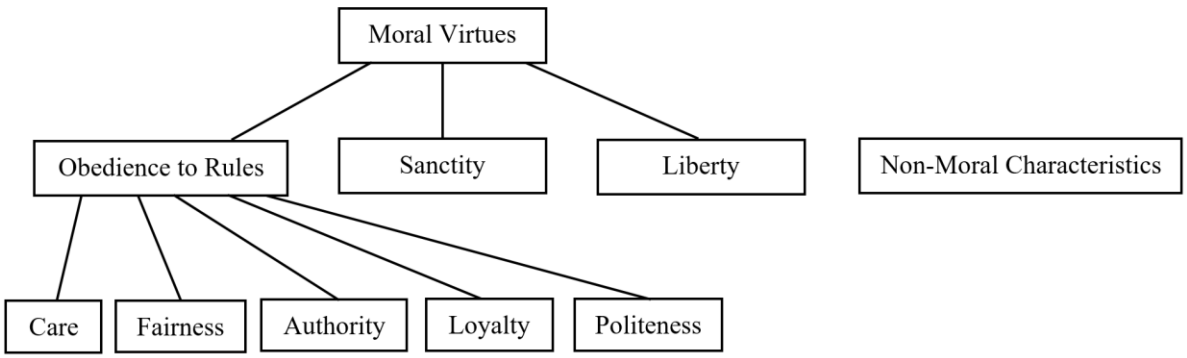


Figure 4: A bottom-up taxonomy of the moral domain.

Table 2: Mean gain in inductive strength from within-category premises and between-category premises.

	MFT	Bottom-Up
Within-Category	11.63 (12.07)	13.30 (13.65)
Between-Category	5.27 (11.29)	3.64 (11.51)

difference in strength gained from within-category and between-category premises was significantly larger for our taxonomy than for MFT, again suggesting that our taxonomy is a better model of conceptualizations of morality than MFT (interaction $F(1, 358) = 21.92$, $p < .001$, $\eta^2_p = .058$).

We next examined cases in which the two taxonomies make differing predictions regarding what premise should be more informative about a given conclusion. For example, in our taxonomy, fairness belongs to the same superordinate category as authority, while sanctity does not; therefore, our taxonomy predicts that the gain in inductive strength for authority conclusions will be greater when participants are given fairness premises than when they are given sanctity premises. In MFT, the reverse is true; MFT therefore makes exactly the opposite prediction. There are eight such combinations, presented in Table 3. Paired- sample t-tests of the gains in inductive strength generally agreed with the predictions of our taxonomy – five of eight significantly supported it, and none significantly supported MFT.

Finally, we constructed a matrix comparing the categorizations in our taxonomy with those in MFT. Premise-conclusion pairs which are in the same superordinate category in our taxonomy but not in MFT (e.g., care/authority) were coded as 1, pairs which are in the same superordinate category in MFT but not in our taxonomy (e.g., care/liberty) were coded as -1, and pairs that both taxonomies categorize in the same way were coded as 0. We computed a correlation between this matrix of categorizations and expressed gains in inductive strength for each participant. A positive correlation indicates that a participant's judgments conform more to the predictions of our derived taxonomy than to those of MFT, and a negative correlation indicates the opposite.

Table 3: Within-Subjects *t*-tests of Study 2 Predictions. *Note:* *$p < .05$; ***$p < .001$.

	Within-Category Premises		
Conclusion	**MFT**	**Bottom-Up**	***t*(358)**
Care	Liberty	Authority	3.99***
Care	Liberty	Loyalty	2.17*
Fairness	Liberty	Authority	4.18***
Fairness	Liberty	Loyalty	3.62***
Authority	Sanctity	Care	4.90***
Authority	Sanctity	Fairness	.40, *ns*
Loyalty	Sanctity	Care	-.008, *ns*
Loyalty	Sanctity	Fairness	-.88, *ns*

Two-hundred-twenty participants out of 358 (61%)[5] expressed a positive correlation, which a binomial test suggests is unlikely to be due to chance, $p < .001$. Moreover, the median correlation, $r = .042$, is significantly larger than 0, by a one-sample Wilcoxon signed-rank test, $p < .001$. Finally, a one-sample t-test performed on the Fisher-transformed correlations indicates that the mean, $z = .047$, is significantly greater than zero, $t(357) = 5.15$, $p < .001$, $d = .27$. Thus, participants' judgments conformed more to the predictions of the taxonomy derived in Study 1 than to the predictions of MFT.

Thus, regardless of how they are analyzed, the gains in inductive strength from learning about a prior behavior consistently resembled the predictions of our taxonomy more than those of MFT. This provides confirmatory evidence that our derived framework better describes people's conceptualizations of morality than does MFT.

Study 3

Method

Participants ($N = 363$) were recruited in the same manner as in the previous studies. Participants from Studies 1 and 2 could not take part in this study.

There are six combinations of two premises and a conclusion in which the two taxonomies make differing predictions about whether the conclusion belongs to the same superordinate category as the premises (see Table 4).

Participants made 24 likelihood judgments of the same form as in Studies 1 and 2, but with two premises instead of one. The premise-conclusion combinations were randomly selected, with the restriction that each participant received four instances of each of the six combinations for which the two taxonomies make differing predictions.

Results

As in Study 2, we created a matrix of categorizations derived from the two taxonomies. Premise-conclusion combinations that belong to the same superordinate category in our taxonomy, but not MFT, were coded as 1, whereas combinations that belong to the same superordinate category in MFT, but not our taxonomy, were coded as -1.

Table 4: Study 3 Predictions.

Premises	**Conclusion**	**Taxonomy Predicting Same Category**
Care/Fairness	Liberty	MFT
Care/Fairness	Authority	Bottom-Up
Care/Fairness	Loyalty	Bottom-Up
Authority/Loyalty	Sanctity	MFT
Authority/Loyalty	Care	Bottom-Up
Authority/Loyalty	Fairness	Bottom-Up

[5] One participant responded "50%" to every question, expressing no variance in her judgments. Her data were excluded from these analyses.

We then computed within-subjects correlations between these codes and participants' likelihood judgments, as in Study 2. A positive correlation indicates that a participant's judgments conform more to our predictions than to MFT, while a negative correlation indicates the opposite. Three hundred-forty-eight participants (96%) expressed a positive correlation, while only 15 (4%) expressed a negative or zero correlation, a result which is highly unlikely to be due to chance, binomial test $p < .001$. Moreover, the median correlation, $r = .45$, is significantly greater than zero, by a one-sample Wilcoxon signed-rank test, $p < .001$. Finally, the mean Fisher-transformed correlation, $z = .48$, is significantly greater than zero, $t(362) = 35.13$, $p < .001$, $d = 1.84$. As in Study 2, participants' judgments conformed more to the predictions of our derived taxonomy than to those of MFT.

General Discussion

Three studies converged on the conclusion that Moral Foundations Theory (MFT) does not describe people's taxonomies of moral virtues well. In Study 1, we derived a taxonomy of virtues, which does not resemble MFT. Study 2 examined the increases in inference strength that come from knowledge of prior behavior, and found that the strength of these gains aligned more closely with the predictions of our taxonomy than those of MFT. Lastly, Study 3 found that ratings of the likelihood of behaviors, given information about two prior behaviors, conformed more to the predictions of our taxonomy than MFT. It is important to note, however, that our results speak *only* to the plausibility of MFT as a model of people's theories of morality, and do not bear on its validity as an evolutionary model of variation in moral virtues.

We do not claim that our derived taxonomy represents the most comprehensive model possible of people's theories of morality. Indeed, because we used stimuli that were already known to be uniquely good exemplars of the moral foundations, we may have left out elements of the moral domain that MFT overlooks. We think that a particularly strong candidate for such an overlooked virtue is honesty, which is sometimes considered part of the fairness foundation, but seems intuitively to be valued even in the absence of fairness concerns (see Landy & Uhlmann, 2016, for a discussion of honesty in folk virtue ethics). Developing a fully bottom-up model of the moral domain that does not inherit the assumptions of any theory is a difficult task, but such a model would be very informative and could help to develop new theoretical advances. Therefore, we see the development of a more comprehensive mapping of concepts of morality as an important direction for future research.

In conclusion, MFT does not seem to model people's conceptualizations of the moral domain especially well. In particular, the distinction between individualizing and binding virtues does not seem to reflect a psychologically real division that people make. By providing a more accurate picture of how people parse their moral worlds, this research helps to clarify a fundamental question in the cognitive science of morality.

Acknowledgments

We thank Halley Bayer for assistance in conducting this research, and Stephanie Chen and the members and guests of the Morality Research Lab for valuable feedback.

References

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods, 47,* 1178-1198.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029-1046.

Haidt, J. & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology, 31,* 191-221.

Haidt, J. & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133,* 55-66.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or, is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613-628.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: I. *Psychometrika, 29,* 1-27.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: II. *Psychometrika, 29,* 115-129.

Landy, J. F. (2016). *Form and content in the categorization of moral violations.* Chicago, IL: University of Chicago.

Landy, J. F. & Bartels, D. M. (2016). *The Moral Violations Database.* Chicago, IL: University of Chicago.

Landy, J. F. & Uhlmann, E. L. (2016). *Morality is personal*. Forthcoming in K. Gray & J. Graham (Eds.), *The atlas of moral psychology.* New York: Guilford.

Medin, D. L. & Rips, L. J. (2005). Concepts and categories: Memory, meaning, and metaphysics. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 37-72). Cambridge, UK: Cambridge University Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97,* 185-200.

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9,* 283-294.

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making, 9,* 176-190.

Turiel, E. (1983). *The development of social knowledge: Morality and convention.* Cambridge, UK: Cambridge University Press.

The Interaction of Memory and Attention in Novel Word Generalization: A Computational Investigation

Erin Grant, **Aida Nematzadeh**, and **Suzanne Stevenson**
Department of Computer Science
University of Toronto
{eringrant, aida, suzanne}@cs.toronto.edu

Abstract

People exhibit a tendency to generalize a novel noun to the basic-level of a hierarchical taxonomy – a cognitively salient category such as "dog" – with the degree of generalization depending on the number and type of exemplars. Recently, a change in the presentation timing of exemplars has also been shown to have an effect, surprisingly reversing the prior observed pattern of basic-level generalization. We explore the precise mechanisms that could lead to such behavior by extending a computational model of word learning and word generalization to integrate cognitive processes of memory and attention. Our results show that the interaction of forgetting and attention to novelty, as well as sensitivity to both type and token frequencies of exemplars, enables the model to replicate the empirical results from different presentation timings. Our results reinforce the need to incorporate general cognitive processes within word learning models to better understand the range of observed behaviors in vocabulary acquisition.

Keywords: novel word generalization; word learning; computational modeling

Introduction

A number of computational models have successfully mimicked child behaviors in learning the meaning of words from ambiguous input (e.g., Siskind, 1996; Yu & Ballard, 2007; Frank et al., 2007; Fazly, Alishahi, & Stevenson, 2010). However, one challenge in word-meaning acquisition that has received less attention is that of *novel word generalization*: i.e., correctly identifying the level of a hierarchical taxonomy that a word refers to. After hearing it only a few times, how does the child determine, for example, that the word *dog* refers to Dalmatians, all dogs of different breeds, or any kind of animal? This issue poses difficulties to the learner because the accumulated evidence can be compatible with more than one of these choices. In this example, all dogs are also animals, and thus the meaning "animal" might also be consistent with all the usages of the word *dog*.

Xu and Tenenbaum (2007) (henceforth XT07) studied novel word generalization in both children and adults by observing decisions about category membership for novel objects in various experimental settings. One of their important findings concerned how people responded having seen 1 vs. 3 labeled exemplars of a certain kind of entity within a taxonomy. For example, having seen a single Dalmatian labeled as a *fep*, people assumed that the novel word *fep* could refer to the general category of dogs. However, if people saw several Dalmatians called *fep*, they apparently recognized that it would be a *suspicious coincidence* if *fep* meant "dog", but only one breed of dog was observed. In such cases, people had a lesser tendency to generalize to the higher level category than after seeing a single exemplar.

Spencer, Perone, Smith, and Samuelson (2011) (henceforth SPSS11) investigated the effect of presentation timing in the same task. XT07 had presented multiple exemplars of a novel word simultaneously. SPSS11 found that instead presenting exemplars in sequence reverses the suspicious coincidence effect. That is, after sequentially viewing three exemplars consistent with a more specific level of the taxonomy (e.g., three dogs of a single breed), people have a **greater** tendency to generalize to the higher category than after seeing one exemplar. SPSS11 explained this reversal as an interaction of word learning with the more general cognitive processes of attention and memory, which differ in their operation across the presentation types: People attend to and remember finer-grained similarities among objects when viewed simultaneously (e.g., that they are all Dalmatians), while the sequential presentation leads people to focus on the general commonalities of the objects (e.g., that they are all dogs).

Our goal in this paper is to provide a computational model that accounts for both the XT07 and SPSS11 findings in a well-motivated manner, by incorporating memory and attentional constraints into an incremental model of word learning and word generalization. It is desirable to integrate together all these pieces – novel word generalization, incremental word learning, and memory and attention – because: (i) word generalization is part and parcel of learning the meaning of words, since it allows the abstraction of meaning from a sequence of specific experiences, and (ii) many word-learning behaviors are influenced by the general cognitive processes of memory and attention (*e.g.*, Vlach et al., 2008; Samuelson & Smith, 2000). Importantly, by explicitly specifying such mechanisms within a computational model, we contribute to the precise understanding of the interactions between them that are required to account for empirical data.

Suspicious Coincidence: Data and Models

XT07 and SPSS11 explored how people generalized a novel word like *fep* to various levels of a taxonomy of objects (including animals, vehicles, and vegetables). Basic-level categories (e.g., dogs or trucks) are those whose members share a significant number of salient attributes; subordinate categories (e.g., Dalmatians or bulldozers) occur lower in the hierarchy, and their members share many fine-grained attributes; superordinate categories (e.g., animals and vehicles) are higher than the basic-level, and their members have fewer attributes in common (Rosch, 1973).

For the sake of space, we focus only on two of the train-

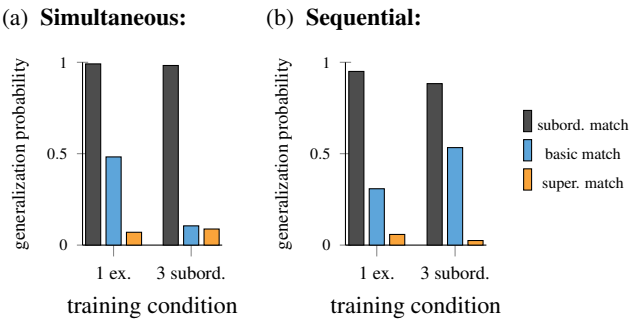
ing conditions in XT07 and SPSS11– the "1-example" and "3-subordinate" conditions – in which the suspicious coincidence effect and its reversal are seen.[1] The 1-example condition has one training trial in which participants observe a single object (*e.g.*, a Dalmatian) that is labeled with a novel word (such as *fep*). In the 3-subordinate condition, participants observe three instances from the same subordinate category (*e.g.*, three different Dalmatians) labeled with the novel word. In XT07's experiment, all three instances were presented *simultaneously*. SPSS11 included a condition in which the three instances are shown and labeled *sequentially*. (Simultaneous and sequential are the same for one example.)

After training, participants select all and only objects that they think are *feps* from a set of test items. Each test object is assessed as exactly one of the following types of match:

- a **subordinate match** has the same subordinate category as a training object (*e.g.*, a Dalmatian).
- a **basic-level match** has the same basic-level category as a training object (*e.g.*, a dog, but not a Dalmatian).
- a **superordinate match** has the same superordinate category as training objects (*e.g.*, an animal other than a dog).

Since SPSS11 replicated the pattern found by XT07 in their simultaneous presentation condition, we report only the results of SPSS11, as shown in Fig. 1.

Figure 1: **SPSS11 Behavioural Data**



SPSS11 data for (1a) simultaneous and (1b) sequential presentations. Each bar is the percent of chosen test objects of each type of match: subord(inate), basic(-level), or super(ordinate). Differences in 1-ex. across the two experiments were not statistically significant.

In the 1-example condition people generalized the novel word to refer to both subordinate matches (*e.g.*, Dalmatians) and (to a lesser extent) basic-level matches (*e.g.*, other kinds of dogs), but not to the superordinate matches (*e.g.*, other animals). This is in line with the idea that people tend to generalize a novel word to a basic-level category such as "dog" because of the perceptual salience of this level of categorization (*e.g.*, Markman, 1991).

In the 3-subordinate condition, when objects are presented simultaneously (Figure 1a), the generalization to the basic level is attenuated compared to the 1-example condition. XT07 explained this behavior as the suspicious coincidence effect. However, when objects are presented sequentially (Figure 1b), there was a surprising reversal of this effect.

While SPSS11 outline possible memory and attentional processes to explain their results, we know of no computational model that can account for both sets of data. XT07's Bayesian model formed hypotheses over a detailed hierarchical taxonomy to account for their own data, but it cannot model the difference between presentation timings, as SPSS11 note. The computational word learner of Nematzadeh, Grant, and Stevenson (2015) (henceforth NGS15) can model the XT07 results without the need for elaborated knowledge of the hierarchy or a built-in basic-level bias. Instead, the results of the model arise from a general type-token frequency interaction of the sort that commonly arises in explanations of linguistic phenomena (e.g., Bybee, 1985; Croft & Cruse, 2004). However, the timing of presentations also has no effect on the NGS15 model, and so the reversal of the suspicious coincidence effect is not achieved. In the next section, we explain how the NGS15 model can be naturally extended to integrate memory and attention, and therefore sensitivity to presentation timing.

Our Computational Model

We start with the NGS15 model because it uses an incremental word learning framework that mimics a range of behaviors in vocabulary acquisition (*e.g.*, Fazly, Alishahi, & Stevenson, 2010; Fazly, Ahmadi-Fakhr, et al., 2010). This framework has recently been extended to incorporate the effects of memory and attention on word learning (Nematzadeh, Fazly, & Stevenson, 2012), presenting a natural opportunity for integrating these processes within word generalization. We describe the NGS15 model, then the novel extensions that enable our model to replicate the SPSS11 data.

Learned Meanings in the NGS15 Model

The NGS15 model is a cross-situational learner that tracks weighted co-occurrences of words and semantic features across its input as in Fazly, Alishahi, and Stevenson (2010). The input to the model is intended to reflect the naturalistic input a child is exposed to, which consists of linguistic input (the words a child hears) paired with nonlinguistic data (the things a child perceives). An input pair is the set of words U_t and the set of semantic features S_t observed at time t:

U_t: { *look*, *a*, *fep* }
S_t: { PERCEPTION, LOOK, ..., DALMATIAN, DOG, ANIMAL }

The output of the model at each time t is a set of meaning probabilities, $P_t(f_i|w_j)$, for each feature f_i and each word w_j observed up through time t. The set of all conditional probabilities $P_t(f_i|w_j)$ for w_j represents the meaning of w_j.

The representation of meaning in NGS15 reflects the structure of taxonomic knowledge. Meaning features are arranged into *feature groups*, each corresponding to a level of the taxonomic hierarchy, as shown in Figure 2.For each word

[1] Our model replicates the results of XT07 and SPSS11 for all training conditions, but we only report the results for these two here.

w_j, a meaning probability distribution, $P_t(.|w_j)$, is calculated for each feature group; that is, $P_t(.|w_j)$ is normalized over the features in a group, rather than over all meaning features. The result is that features at the same level of the hierarchy, such as DALMATIAN and POODLE, or DOG and CAT, compete for probability mass; this ensures that such features, which are mutually incompatible given their taxonomic relationship, cannot simultaneously have high probability. Features at different levels of the hierarchy are in different feature groups and thus do not compete for probability mass; this ensures that meaning probabilities such as $P_t(\text{DALMATIAN}|fep)$, $P_t(\text{DOG}|fep)$, and $P_t(\text{ANIMAL}|fep)$ can all be highly activated if *fep* is intended to refer to a Dalmatian (which is also both a dog and an animal). In this approach, the meaning of a word is the set of n distributions, $P_t(.|w_j)$, one per feature group in a taxonomy with n levels.

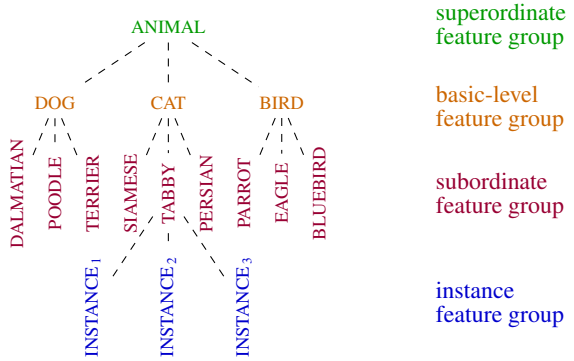


Figure 2: A portion of the taxonomy used in this paper.

The NGS15 Learning Algorithm

The input to the model is processed in an incremental two-step bootstrapping framework: Words and features that co-occur are **aligned** (associated) in proportion to the current meaning probabilities, which are then **updated** with the new evidence regarding strength of association, as follows.

The Alignment Step. For an input pair at time t, the model calculates a probabilistic strength of aligning (associating) each word $w_j \in U_t$ with each feature $f_i \in S_t$:

$$a_t(f_i, w_j) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in U_t} P_{t-1}(f_i|w')} \tag{1}$$

These alignment strengths are incrementally accumulated as:

$$\begin{aligned} \text{assoc}_t(f_i, w_j) &= \text{assoc}_{t-1}(f_i, w_j) + a_t(f_i, w_j) \\ &= \sum_{t' \in \mathcal{T}} a_{t'}(f_i, w_j) \end{aligned} \tag{2}$$

where $\mathcal{T}$ is all times at which f_i and w_j have co-occurred.

In our model, if more than one instance of feature f_i occurs with word w_j at time t, multiple instances of $a_t(f_i, w_j)$ are recorded. For example, in the simultaneous presentation of three exemplars with the word *fep*, the alignment strength $a_t(f, fep)$ will be added three times to the association score for each feature f in the input.

Update of Meaning Probabilities The model next uses the association scores to update the meaning probabilities. Each meaning probability $P_t(f_i|w_j)$ represents the magnitude of the f_i–w_j association *relative to* the association strength between w_j and other features within the same feature group $\mathcal{G}$ as f_i:

$$P_t(f_i|w_j) = \frac{\text{assoc}_t(f_i, w_j) + \gamma_{\mathcal{G}}^t}{\sum_{f_m \in \mathcal{G}} \text{assoc}_t(f_m, w_j) + k_{\mathcal{G}} \gamma_{\mathcal{G}}^t} \tag{3}$$

Here $k_{\mathcal{G}}$ and $\gamma_{\mathcal{G}}^t$ are smoothing terms: $k_{\mathcal{G}}$ reflects the expected number of features in $\mathcal{G}$ and $\gamma_{\mathcal{G}}^t$ represents the *a priori* tendency to observe a feature in $\mathcal{G}$. While $k_{\mathcal{G}}$ is a fixed parameter, $\gamma_{\mathcal{G}}^t$ is a function of the number of observed types within the feature group $\mathcal{G}$, and thus changes over time (see NGS15).

The $\gamma_{\mathcal{G}}^t$ parameters are key to the generalization behavior of the NGS15 model because they influence how much probability mass is allocated to a feature previously unseen with a word (cf. Eqn. 3 when the assoc score is 0). A higher value for $\gamma_{\mathcal{G}}^t$ leads to more probability mass allocated to previously unseen features in group $\mathcal{G}$, allowing for more generalization to new features in that group. Because $\gamma_{\mathcal{G}}^t$ increases with the number of types, it captures the oft-observed tendency in language that people more readily generalize categories for which a greater variety of types of items has been observed. The model matches the child data from XT07 by equating $\gamma_{\mathcal{G}}^0$ across feature groups. But to match the adults, who show a stronger basic-level bias, the model required that the $\gamma_{\mathcal{G}}^0$ parameters be initialized to successively higher values for feature groups successively lower in the hierarchy, entailing that, e.g., it is easier to generalize a novel word to a new breed of dog not seen in training (basic-level generalization), than to a new kind of animal not seen in training (superordinate generalization).

Our Extensions to Integrate Memory and Attention

To render the model sensitive to presentation timing, we adopt the general approach of Nematzadeh et al. (2012), which integrates memory and attention seamlessly into the cross-situational word-learning mechanism. The approach was shown to account for spacing effects in word learning, which are closely related to the presentation timing factors considered by SPSS11. However, the methods must be extended to adequately meet the needs of word generalization in the NGS15 model; we describe those extensions here.

Modeling the Effects of Forgetting. To model the effect of memory, we use the association score formulation of Nematzadeh et al. (2012), which implements "forgetting" by applying a decay factor to each alignment probability (cf. Eqn. 2 above):

$$\text{assoc}_t(f_i, w_j) = \sum_{t' \in \mathcal{T}} \frac{a_{t'}(f_i, w_j)}{(t - t' + 1)^{d_{a_{t'}}}} \tag{4}$$

Each alignment in the sum is scaled by the temporal distance between the current time t and the time t' that the alignment

was made, exponentiated to a decay function $d_{a_{t'}}$ that is inversely proportional to the strength of alignment.

However, we must extend this decay formulation to accommodate our hierarchical knowledge of feature groups.[2] In particular, we find that using the same decay rate across all feature groups is not sufficient. As noted above, appropriate word generalization in the NGS15 model requires that lower levels in the taxonomy be more "open" to generalizing to new features than higher levels in the taxonomy. It is important to note that the decay of alignments also influences "openness" to generalization because it shifts probably mass away from observed word–feature pairs onto unseen events. Thus, to appropriately reflect the nature of the hierarchy – that openness increases with greater depth in the taxonomy – we must parameterize decay by feature group. Just as feature groups lower in the taxonomy must have successively higher γ values to indicate more "openness" to generalization, lower feature groups also require higher decay rate parameters.

We thus use the following formulation of decay:

$$d_{a_t} = \frac{d_{\mathcal{G}}}{a_t(f_i, w_j)} \tag{5}$$

where $d_{\mathcal{G}}$ controls the rate of decay for features in feature group $\mathcal{G}$, and is set successively higher for lower-level feature groups in the taxonomy.

Modeling Attention to Novelty. Building on research showing that people attend more to novel stimuli in learning (e.g., Snyder et al., 2008; MacPherson & Moore, 2010; Horst et al., 2011), we use the general idea of Nematzadeh et al. (2012) in allocating more strength to alignments that are more novel (cf. Eqn. 1):

$$a_t(f_i, w_j) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in U} P_{t-1}(f_i|w')} \cdot \text{novelty}_t(f_i, w_j) \tag{6}$$

In this model, $\text{novelty}_t(f_i, w_j)$ was inversely proportional to how recently w_j had been observed, and thus focused solely on novelty of words; the novelty of the feature f_i was not considered. We must broaden this approach because the experiments here are focused on a single novel word.

Here instead we consider the novelty of the observed *word–feature pairing*, and again draw on considerations of type–token frequencies, as in other aspects of the NGS15 model. Specifically, we scale the alignment strength by the ratio of the token frequency of f_i–w_j observations at time t to the total frequency of all such observations, by formulating $\text{novelty}_t(f_i, w_j)$ as:

$$\text{novelty}_t(f_i, w_j) = \frac{\text{token}_t(f_i, w_j)}{\sum_{t' \in \mathcal{T}} \text{token}_{t'}(f_i, w_j)} \tag{7}$$

where $\text{token}_t(f_i, w_j)$ is the number of tokens of feature f_i that occurred at time t with word w_j.

[2] Nematzadeh et al. (2012) used a single meaning probability distribution over all features – i.e., there are no feature groups.

This formulation achieves attention to novelty as follows. Generally, earlier observations of feature f_i with word w_j will have a stronger alignment than later observations, where the increased number of observations will increase the denominator of $\text{novelty}_t(f_i, w_j)$, and lead to attenuation of the alignment strength. Note that when the co-occurrence of f_i with word w_j is truly novel – *i.e.*, the first time they are observed together – the strength of alignment is undiminished, since the numerator and denominator of the novelty factor are equal in the initial observation of f_i with w_j.

Summary of Novel Extensions to the NGS15 Model In summary, we have extended both the model of NGS15, and the memory and attention mechanisms of Nematzadeh et al. (2012), by: (i) incorporating a forgetting mechanism that is sensitive to the taxonomic level of a feature group, which reflects the needs of taxonomic structure and the process of novel word generalization; and (ii) formulating a mechanism for attention to novelty of word–feature pairings, rather than just to recency of words, consistent with the key role of word–feature association statistics in the model.

These mechanisms have a direct impact on the processing of stimuli in simultaneous vs. sequential presentations in a novel word generalization task. The forgetting mechanism ensures that more general features, such as the kind of animal observed (*e.g.*, dog or cat), are remembered better than more detailed features, such as particular breeds of dogs. The attention-to-novelty mechanism has the consequence that successive observations of word–feature pairings in a sequential presentation scenario are "discounted" with respect to earlier presentations. We demonstrate in our experiments below that, together, these mechanisms interact to enable the model to account for both the suspicious coincidence effect in a simultaneous presentation as found by XT07, and its reversal in a sequential presentation as found by SPSS11.

Methodology

We follow the methods of NGS15, adapted where needed for our extended model on the SPSS11 data.[3]

Training the Model. We use a taxonomy with three levels, corresponding to the subordinate, basic, and superordinate categories of animals. This yields four feature groups, one per category level plus an "instance" group to distinguish multiple objects of the same subordinate category. See Figure 2. In each U_t–S_t input pair, U_t consists of the novel word, and S_t is a set of four features (one per feature group) representing a unique instance of the same subordinate category across all training trials; for example:[4]

U_t: { *fep* }
S_t: { INSTANCE$_1$, DALMATIAN, DOG, ANIMAL }

[3] Our code and data are available at https://github.com/eringrant/novel_word_generalization.

[4] Each FEATURENAME stands for all features of an object at that level of the hierarchy. Such features could be replaced with an appropriate *set* of features without changing the model results.

In the 1-example condition, training consists of just one such U_t–S_t pair. In the 3-subordinate condition, training has three such U_t–S_t pairs, differing only in the unique instance feature (*i.e.*, INSTANCE$_1$, INSTANCE$_2$, INSTANCE$_3$) in each S_t. In the simultaneous condition, the three U_t–S_t pairs are all presented at the same time t. In the sequential condition, the three U_t–S_t pairs are presented one at a time, at t, $t+1$, and $t+2$.

Testing the Model. After training, the level of generalization of the novel word is assessed against test objects, each of which is a subordinate match, a basic-level match, or a superordinate match; for example:

subord. match:	{ INSTANCE$_4$, DALMATIAN, DOG, ANIMAL }
basic match:	{ INSTANCE$_5$, POODLE, DOG, ANIMAL }
super. match:	{ INSTANCE$_6$, TOUCAN, BIRD, ANIMAL }

We adapt the P_{gen} formula of NGS15 to test whether the model generalizes the learned meaning of the novel word w to the various levels of match at test time T (after training):

$$P_{gen}(m|w) = \frac{\text{avg}_{Y \in m}\, P_T(Y|w)}{\text{avg}_{Y' \in \{\text{sub.}\}}\, P_T(Y'|w)}$$

Here $P_T(Y|w)$ is the probability of a test object Y given w, and m is the set of test objects at a certain level of match. The measure in the numerator of P_{gen} is the average such probability across test matches at that level, $\text{avg}_{Y \in m}\, P_T(Y|w)$. This is not directly comparable to the empirical data, which are the percentages of test objects selected from each type of match. To obtain a comparable measure, we scale each probability (for each level of match) by the probability of the subordinate matches in that condition, $\text{avg}_{Y' \in \{\text{sub.}\}}\, P_T(Y'|w)$ (the denominator of P_{gen}). Thus $P_{gen}(m|w)$ is the relative average preference for test items at level m. This renders the subordinate match probability as 1.0 (reflecting that people generally pick close to 100% of the subordinate test items), and shows the other type of matches relative to that amount.

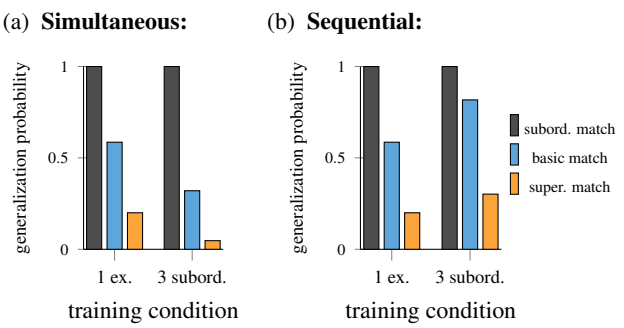
Model Parameters. Since the SPSS11 participants are adults, we use the *adult parameter settings* of NGS15 for the four $\gamma_{\mathcal{G}}^0$ parameters and the four $k_{\mathcal{G}}$ (one each per feature group), which are tuned to achieve a match to adult data of XT07. For our decay parameters, we use:

$$d_{\text{inst}} = 0.8 \quad d_{\text{subord}} = 0.5 \quad d_{\text{basic}} = 0.05 \quad d_{\text{super}} = 0.01$$

Model Results and Discussion

Figure 3 shows the results of our model in the simultaneous and sequential conditions; cf. Figure 1 for the human behavioral data in SPSS11. Following simultaneous presentation of training input (Figure 3a), our model shows the suspicious coincidence effect: Generalization to the basic level is inhibited in the 3-subordinate condition as compared to the 1-example condition. In contrast, sequential presentation reverses the suspicious coincidence effect (Figure 3b): the model exhibits greater basic-level generalization in the 3-subordinate condition. Thus, these results replicate the qualitative pattern evident in the behavioural data of SPSS11.

Figure 3: **Our Model Data**



Our model data for (3a) simultaneous and (3b) sequential presentations. Each bar is the probability of a type of test match: *i.e.*, subord(inate), basic(-level), or super(ordinate), scaled by the subordinate match probability (see text).

The interaction (between presentation type and amount of training) seen in the human data arises as a result of a corresponding interaction in the model. Consider each 3-subordinate condition (simultaneous and sequential) compared to the 1-example condition. **In the simultaneous 3-subordinate case**, the attentional mechanism yields higher alignment strengths between the word and features because their three co-occurrences are all novel at the single presentation time; in addition there is little forgetting because the items are all seen at time t and test is at time $t+1$. This yields stronger subordinate alignments compared to the 1-example case, and therefore somewhat less basic-level generalization.

By contrast, **in the sequential 3-subordinate case**, the word–feature co-occurrences are less salient because they decrease in novelty over the three presentation times. In addition, greater forgetting occurs because there is more time between the (first two) presentation times and test time ($t+4$). In this case, because subordinate features decay faster than basic features, the interaction yields weaker subordinate features compared to the 1-example case, and more basic-level generalization is achieved.

The interaction of memory and attention effects are required to obtain this pattern of results in the model. If the model includes only the decay mechanism, differentiated by taxonomic level, this enables it to focus more on abstract than specific features, and consequently raises the basic generalization closer to the level of the subordinate generalization in *all* conditions. On the other hand, using the attention mechanism alone enables the model to distinguish the sequential and simultaneous conditions, but it cannot on its own raise the basic generalization high enough. Only when the two are used together does the model produce the reversal of the suspicious coincidence effect in the sequential presentation.

The necessity of both memory and attention is suggestive of how word learning occurs in people. In particular, the attention mechanism in the model focuses more probability onto word–feature co-occurrences in their earlier presentations, simulating the general tendency for people to attend more to less-familiar things. In addition, the mechanism that

increases the decay rate for lower-level features in the taxonomy simulates the tendency in people to remember abstract features of objects over very specific features. SPSS11 contended that people are able to attend to specific features in the simultaneous condition due to the close spatial and temporal proximity of the items, and correspondingly attend only to the abstract commonalities of items in sequential presentations. Our model explains this effect as the result of general memory and attention mechanisms that have been shown to play a role in word learning more widely (cf. Nematzadeh et al. (2012); Nematzadeh, Fazly, and Stevenson (2013)). Interestingly, attention in our model is a function of the token frequency of word–feature co-occurrences (as opposed to a fixed parameter) and is therefore a response to the statistics of the data, as are other components of our word generalization formulation. All this further supports that attention, memory and statistical learning interact to produce the suspicious coincidence effect and its reversal across presentations.

Conclusions and Future Work

Novel word generalization – understanding how a word maps to the appropriate level of a taxonomic hierarchy – is an important aspect of novel word learning, but one that has not received much attention in the word-learning community. We propose a unified model of word learning that accounts for the various observed patterns of novel word generalization – in particular, the suspicious coincidence effect (Xu & Tenenbaum, 2007) and its reversal under differing presentation conditions (Spencer et al., 2011). We extend the model of Nematzadeh et al. (2015) with a novel integration of the general cognitive mechanisms of memory and attention, and show that our model's success is a result of the interaction of forgetting and attention to novelty of word–feature co-occurrences. Our approach builds on the earlier NGS15 model in highlighting the importance of type and token frequency patterns in the input to capturing interesting generalization effects, but here these patterns are manifest in our formulation of memory and attention mechanisms.

In incorporating these cognitive processes into our model, we drew on the approach of Nematzadeh et al. (2012), whose model had been shown to account for various spacing effects in word learning (see also Nematzadeh et al., 2013). Much further work is needed to explore whether our model can explain other such effects. For example, Vong, Perfors, and Navarro (2014) showed that people's categorization of novel object instances depends on the distribution of training examples both that are labelled with a word as well as those that are unlabelled. Currently, our model only takes into account word–feature co-occurrences, and is therefore insensitive to features that occur without a word label. We will need to consider how to integrate learning from unlabelled data in order to better model how statistical word learning interacts with object categorization, as it does in people.

References

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia: Benjamins.

Croft, W., & Cruse, A. (2004). *Cognitive linguistics*. Cambridge University Press.

Fazly, A., Ahmadi-Fakhr, F., Alishahi, A., & Stevenson, S. (2010). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. In *CogSci Proceedings* (Vol. 10, pp. 2362–2367).

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *NIPS Proceedings* (Vol. 20, pp. 457–464).

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition*, *118*(2), 234 - 244.

MacPherson, A. C., & Moore, C. (2010). Understanding interest in the second year of life. *Infancy*, *15*(3), 324–335.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. MIT Press.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). A computational model of memory, attention, and word learning. In *CMCL Proceedings* (pp. 80–89).

Nematzadeh, A., Fazly, A., & Stevenson, S. (2013). Desirable difficulty in learning: A computational investigation. In *CogSci Proceedings* (pp. 1073–1078).

Nematzadeh, A., Grant, E., & Stevenson, S. (2015). A computational cognitive model of novel word generalization. In *EMNLP Proceedings* (pp. 1795–1804).

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (p. 111-144).

Samuelson, L. K., & Smith, L. B. (2000). Grounding development in cognitive processes. *Child Develop.*, 98–106.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Snyder, K. A., Blank, M. P., & Marsolek, C. J. (2008). What form of memory underlies novelty preferences? *Psychological Bulletin and Review*, *15*(2), 315–321.

Spencer, J. P., Perone, S., Smith, L. B., & Samuelson, L. K. (2011). Learning words in space and time: Probing the mechanisms behind the suspicious-coincidence effect. *Psychological science*, *22*(8), 1049–1057.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The relevance of labels in semi-supervised learning depends on category structure. In *CogSci Proceedings* (p. 1718-1723).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psych. Rev.*, *114*(2), 245–272.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.

Environmental Orientation Affects Emotional Expression Identification

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
Derek Shapiro (dshapiro723@gmail.com)
Kevin B. Collister (kevin.b.collister@gmail.com)
SUNY Purchase College, Department of Psychology
735 Anderson Hill Road, Purchase, NY 10577, USA

Paul H. Thibodeau (pthibode@oberlin.edu)
Oberlin College, Department of Psychology
101 N. Professor St., Oberlin, OH 44074, USA

Abstract

Spatial metaphors for affective valence are common in English, where up in space=happy/positive and down in space=sad/negative. Past research suggests that these metaphors have some measure of psychological reality: people are faster to respond to valenced words and faces when they are presented in metaphor-congruent regions of space. Here we explore whether the *orientation* of a stimulus – rather than its *position* – is sufficient to elicit such spatial-valence congruency effects, and, if so, which spatial reference frame(s) people use to represent this orientation. In Experiment 1, participants viewed images of happy and sad profile faces in different orientations and had to identify the emotion depicted in each face. In Experiment 2, participants completed this task while lying down on their sides, thereby disassociating environmental and egocentric reference frames. Experiment 1 revealed a metaphor-congruent interaction between emotion and orientation, while Experiment 2 revealed that this spatial-valence congruency effect was only reliable in the environmental frame of reference.

Keywords: spatial metaphor; valence, emotional expression identification; spatial reference frames

Introduction

We often talk about abstract domains like time, emotion, consciousness, health, and social status using spatial metaphors (Lakoff & Johnson, 1980). For example, we organize the concept of emotional valence around a vertical spatial dimension, where a higher spatial position connotes happiness and positivity while a lower spatial position connotes sadness and negativity. Thus we say things like "She *sank* into a *deep depression* and is feeling quite *low*, very *down* in the dumps; we need to give her a *lift, raise* her spirits, and *boost* her self-esteem until she's *flying high*." This particular mapping between space and valence may have its origins in everyday embodied experiences: sad feelings are associated with a drooping posture and drowsiness, while happy feelings are associated with a more alert and erect bearing (though it's rare to see people literally jumping with joy).

Interestingly, research has found that this association between space and valence goes beyond language: people seem to automatically activate metaphor-congruent spatial representations in the course of processing valenced stimuli (and vice versa. e.g., Brookshire, Ivry, & Casasanto, 2010; Casasanto & Dijkstra, 2010; Lynott & Coventry, 2013; Meier & Robinson, 2004).

In one study, participants were faster and more likely to retrieve positive memories while making concurrent motor movements upwards, and faster and more likely to retrieve negative memories while making concurrent motor movements downwards (Casasanto & Dijkstra, 2010). In another experiment, Meier and Robinson (2004) found that participants were faster to identify positive words like *hero* when they appeared at the top of the screen than when they appeared at the bottom of the screen, while the reverse was true for negative words like *liar*. A follow-up study revealed that simply attending to higher or lower regions of space facilitated the subsequent processing of valenced words presented centrally in a metaphor-congruent fashion. More recently, Lynott and Coventry (2013) extended these findings by using non-linguistic stimuli: in their study, participants were faster to respond to happy faces that appeared at the top of the screen compared to sad faces that appeared at the top of the screen and happy faces that appeared at the bottom of the screen. Taken together, these findings provide compelling evidence for the cognitive reality of the spatial-valence metaphorical mapping.

The present work builds on these findings by testing whether people are also sensitive to the *orientation* of valenced stimuli. That is, one way to test for spatial-valence congruency effects is to manipulate the *position* of stimuli in a display (or the direction of movement towards a particular position) – the method used by other researchers tackling this phenomenon. Another way of testing for spatial-valence congruency effects is to manipulate the *orientation* of a stimulus. In addition to presenting a face at the top of a computer screen (position), "up" can be cued by presenting an upward gazing face (orientation). Furthermore, spatial relations like *upright* and *orientation* must be defined with respect to a particular frame of reference; objects that are upright with respect to a computer screen (environmental frame of reference) would appear upside-down to a person standing on their head (egocentric frame of reference). In other words, spatial relationships are multifaceted; a fuller consideration of this nuance can help us understand how people use space to represent valence.

Across two experiments, we investigated whether the *orientation* of a stimulus – rather than its *position* in space –

would be sufficient to elicit such spatial-valence congruency effects, and, if so, which spatial reference frame people use to represent this orientation. In Experiment 1, participants viewed images of happy and sad profile faces in different orientations and had to identify the emotion depicted in each face. We expected a metaphor-congruency effect: that people would be faster (and more accurate) to respond to upward gazing happy faces and downward gazing sad faces, and slower (and less accurate) to respond to upward gazing sad faces and downward gazing happy faces. This finding would provide additional evidence that representations of valence are grounded in conceptions of space, and it would demonstrate that spatial-valence congruency effects are not limited to spatial located, but extend also to spatial orientation.

In Experiment 2, participants completed this task while lying on their right sides, thereby disassociating environmental and egocentric reference frames. In a majority of everyday experiences, environmental and egocentric reference frames are highly correlated – most of the time we see faces that are upright in the world (environmental reference frame) while we sit or stand in an upright position (egocentric reference frame). This study design allowed us to investigate whether the representation of valence is more strongly tied to the reference frame of the world (environmental) or the individual (egocentric).

Given the importance of egocentric reference frames in face perception (e.g., Rossion, 2008; Troje, 2003), one possibility is that the spatial-valence mapping will be defined with respect to the orientation of the participant. On the other hand, our experience with faces in the world, which are normally upright (even if we are tilted or on our side), may tie the spatial-valence mapping to an environmental frame of reference. Indeed, some metaphors in English seem to reference the environmental frame specifically, as when we say, "things are looking *up*." Identifying the reference frame(s) in which the spatial-valence mapping is defined can help us understand how these representations are learned and when they influence our behavior (Davidenko & Flusberg, 2012).

Experiment 1

Methods

Participants We recruited 81 participants (59 female) from the Introduction to Psychology Participant Pool at SUNY Purchase College. The average age was 19 (*SD*=1.2), and participants received course credit for their participation.

Materials & Procedure The experiment was created using PsychoPy software (Pierce, 2007) and was administered on a 21.5" iMac desktop computer. Face stimuli were drawn from the Karolinska Directed Emotional Faces Database (KDEF; Lundqvist, Flykt, & Öhman, 1998). We selected 40 profile faces from the database for use in the study: 10 male faces and 10 female faces each expressing both happiness and sadness. The images were cropped in Adobe Photoshop using the same ovular template to highlight the face and keep each picture the same size.

On every trial, a black fixation cross appeared at the center of the screen, which had a light gray background. After 500 milliseconds, one of the face images appeared at the center of the display in one of five possible orientations (0° = upright; -90º, -45º = looking downwards; 45º, 90º = looking upwards; see Figure 1). All 40 face images appeared in each of the 5 orientations, for a total of 200 trials. Half of the faces were presented facing to the left, and half were presented facing to the right (counterbalanced); the order of trials was fully randomized.

Participants were instructed to respond as quickly and as accurately as possible as soon as the face image appeared during a trial, pressing one button on the keyboard if the face was "happy" and another button if the face was "sad." Participants used the "f" and "j" keys to respond (counterbalanced across participants).

Participants were also randomly assigned to one of two stimulus duration conditions. In the *Unmasked* condition (N=40), the face image remained on the screen until participants pressed a response key. In the *Masked* condition (N=41), the face image remained on the screen for 100 milliseconds and was then replaced by a scrambled version of one of the images, created in Photoshop. This manipulation was included to test whether the effects of metaphorical spatial congruence on emotional expression identification emerge early in visual processing (i.e., within the first 100 milliseconds).

Before the main experimental task, participants completed 8 practice trials consisting of upright, front view, cartoon faces (two sad faces and two happy faces, each presented twice) to acclimate them to the task.

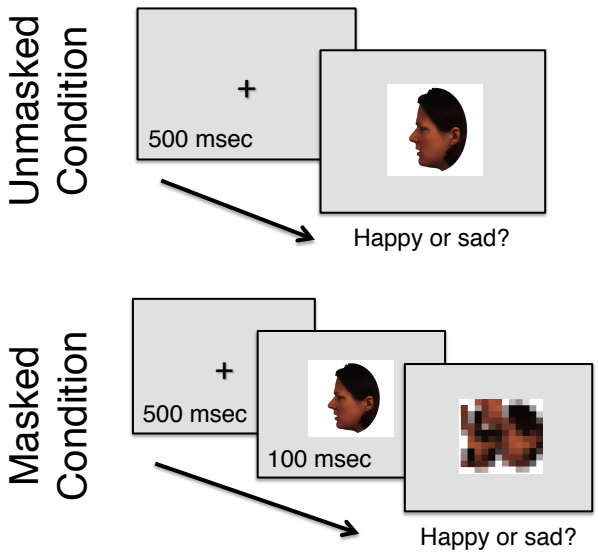


Figure 1. Schematic diagram of trial structure for both stimulus duration conditions in Experiment 1.

Results

Response times faster than 200 milliseconds and slower than five standard deviations above the overall mean RT were removed from analysis (<1% of all trials). Accuracy was very good overall (M=93.7%, SD=4.3).

Our initial analysis included only trials where participants correctly identified the emotional facial expression. Using reaction time as our dependent variable, we ran a 2 (emotion: happy vs. sad) X 5 (orientation: -90º, -45, 0º, 45º, 90º) repeated measures ANOVA with stimulus duration condition included as a between-subjects factor. Of particular relevance to our theoretical question was a predicted metaphor-congruent interaction between the emotion and orientation of the face, which was statistically significant, $F(4, 316)=5.74$, $p<0.001$, $\eta^2=.068$[1] (see Figure 2). Planned contrasts revealed that participants were faster to recognize happy faces oriented at 45º, $t[80]= 2.28$, $p=.025$, and 90º, $t[80]=3.07$, $p=.003$. There were no differences in recognition time by emotional expression at other orientations (-90º, -45º, 0º), ts < 1, ps > .3.

In addition to the predicted interaction, the model revealed a main effect of orientation, $F(4, 316)=36.88$, $p<0.001$, $\eta^2 = .317$, consistent with prior work on the effects of orientation on face perception (e.g., Davidenko & Flusberg, 2012): participants were fastest to respond to upright faces (0º) and were progressively slower to respond as the faces were rotated away from upright. No other main effects or interactions were statistically significant, ps > .1.

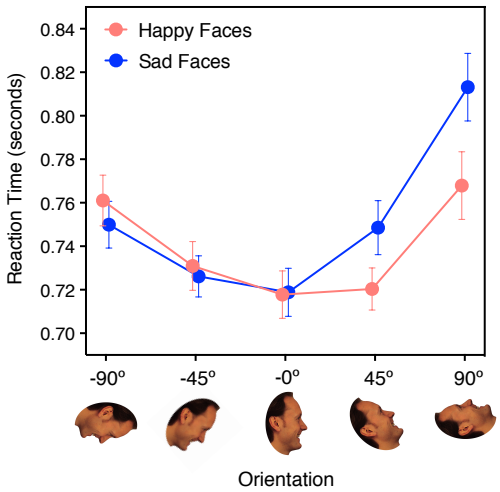


Figure 2. Mean reaction times for happy and sad faces for each stimulus orientation in Experiment 1, collapsed across stimulus duration condition. Error bars represent 95% CIs

An analysis of error trials revealed a similar pattern. Using error frequency as the dependent variable, we conducted another 2 (emotion) X 5 (orientation) repeated-measures ANOVA with condition as a between-subjects factor. Consistent with the analysis of RTs, we found that people made more errors on trials that presented downward gazing happy faces than trials that presented downward gazing sad faces, and vice versa for upward gazing faces, $F(4, 316)=4.84$, $p<0.001$, $\eta^2=.056$. Planned contrasts revealed that participants made more errors in recognizing happy faces oriented at -90º, $t[80]= 5.23$, $p<.001$, -45º, $t[80]=3.28$, $p=.002$, 0º, $t[80]=4.35$, $p<.001$, and 45º, $t[80]=2.94$, $p=.004$; there were no differences in error rates by emotional expression 90º, $t[80]=0.05$, $p=.958$.

The model also revealed differences between the *Masked* and *Unmasked* conditions. Not surprisingly, participants in the *Unmasked* condition (mean accuracy=96.4%, SD=2.39) made fewer errors than those in the *Masked* condition (mean accuracy=91%, SD=4.13), $F(1, 79)=47.76$, $p<0.001$, $\eta^2=.379$. Since the performance of participants in the *Unmasked* condition was close to ceiling, the interaction between emotion and orientation was only present for participants in the *Masked* condition (i.e., the 2-way interaction between emotion and orientation was qualified by a 3-way interaction between emotion, orientation, and condition, $F(4, 316)=3.89$, $p<0.005$, $\eta^2=.045$[2]).

In addition, this model revealed that people made more errors for happy face trials (M=92.4% accuracy, SD=5.64) than for sad face trials (M=95.2% accuracy, SD=4.59)[3], $F(1, 79)=20.1$, $p<0.001$, $\eta^2=.201$, and more errors as the faces were rotated away from upright, $F(4, 316)=13.6$, $p<0.001$, $\eta^2=.136$. As shown in Figure 3, the effect of orientation was only apparent for those in the *Masked* condition, $F(4, 316)=10.66$, $p<0.001$, $\eta^2=.104$. See Figure 3.

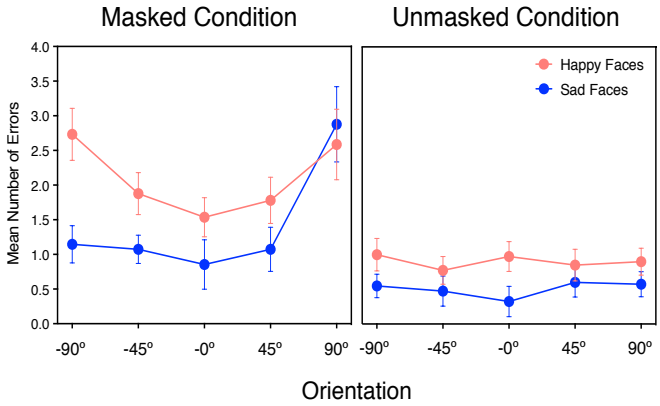


Figure 3. Mean number of errors for happy and sad faces for each stimulus orientation in each condition in Experiment 1. Error bars represent 95% CIs

[1] Though Mauchly's test indicated that the assumption of sphericity was violated in this and several of the following analyses, the F and p-values remain nearly identical under both Greenhouse-Geisser and Huynh-Feldt corrections in all cases.

[2] Results of planned contrasts do not change when focusing exclusively on data from the *Masked* condition: ts > 2.5, ps < .016 for orientations < 90º; t = .62, p = .539 at 90º.

[3] This may reflect a slight response bias in our sample to perceive sadness in others (or negativity more generally), or it may signal that our stimuli were not equally discriminable based on emotion.

Discussion

In Experiment 1, we asked whether the *orientation* of a face (i.e. where the face is *looking,* as opposed to its *position* in space) would be sufficient to elicit spatial-valence congruency effects on performance in an emotional expression identification task. The answer was a clear *yes*: participants were faster and more accurate to identify emotional expressions when the faces were oriented towards metaphor-congruent regions of space. This was true whether the stimuli were masked after 100 milliseconds or remained visible until response, suggesting that metaphorical spatial representations of valence are activated quickly and automatically when people view emotional stimuli (and vice versa; cf., Brookshire, Ivry, & Casasanto, 2010).

Interestingly, these effects seemed to be driven largely by a *decrease* in performance for sad faces facing upwards: while RTs and error rates (in the *Masked* condition) for happy faces increased symmetrically as the images were rotated upwards and downwards away from upright, RTs and error rates for sad faces dramatically increased on upward rotations (i.e., metaphor-incongruent orientations). This is somewhat surprising, as the only other published work on spatial-valence congruency effects that used happy and sad face stimuli found a response time *advantage* for happy faces positioned in metaphor-congruent regions of space (i.e., the top of the display), rather than a metaphor-incongruent decrease in performance for sad faces (Lynott & Coventry, 2013). These researchers interpreted these findings as evidence for a "polarity" account of spatial-valence congruency effects, which is an issue we return to in the general discussion (cf., Dolscheid & Casasanto, 2015).

The results of Experiment 1 cannot address one key question: which way is *up*? Spatial relations like *up*, *down*, and *orientation* must be defined with respect to a particular frame of reference. When participants are seated at a computer in a typical lab study like Experiment 1, several spatial reference frames are conflated: faces that are oriented upwards with respect to the computer screen, the room itself, and the directional pull of gravity (environmental frames) are also orientated upwards with respect to the participant (egocentric reference frames). This makes it impossible to determine which reference frame(s) participants are using to represent the orientation of the faces (and thus which reference frame is driving the observed spatial-valence congruency effects).

Fortunately, there is a simple method for disassociating environmental and egocentric reference frames: tilt your head 90º to one side. Now faces that appeared to be gazing upwards in the environment will appear to be upright or upside-down in your egocentric frame of reference (depending on which way you tilt your head). Interestingly, prior research has shown that people process faces independently in both the environmental and egocentric reference frames: Davidenko & Flusberg (2012) found that people were better at classifying and remembering images of faces that were egocentrically upright (as compared to egocentrically inverted) as well as environmentally upright (as compared to environmentally inverted), though effects in the environmental reference frame were reliably smaller. In Experiment 2, participants completed the same task as in Experiment 1 while lying down on one side.

Experiment 2

Methods

Participants We recruited 85 participants (59 female) from the Introduction to Psychology Participant Pool at SUNY Purchase. The average age was 19.2 (*SD*=2.63) and participants received course credit for their participation.

Materials & Procedure The experiment was similar in design to Experiment 1, with a few key differences:

Instead of sitting on a stool at a computer workstation, participants began the experiment by sitting upright on a futon positioned at the back of the lab room. The computer running the experimental software was positioned on a low table in front of the futon. Participants first completed the same 8 practice trials featuring front-view cartoon faces that participants completed in Experiment 1. The only difference was that they used only their left hand to make the speeded response, using the "1" and "2" keys on the keyboard (counterbalanced across participants). After the practice trials, participants were instructed to lay down on their right side with their head resting horizontally on a flat pillow facing the computer screen.

For Experiment 2 we used eight out of the ten male and eight out of the ten female profile faces that we had used in Experiment 1, each one again appearing with both a happy and sad expression[4]. On any given trial, one of the 32 individual profile images (8 males, 8 females, 2 expressions each) appeared in one of 8 possible orientations (see Table 1 and Figure 4). Participants saw each of the 32 faces in each of the 8 possible orientations, for a total of 256 trials, the order of which was randomized across participants.

Table 1. Stimulus orientations in Experiment 2

	Egocentric Orientation	Environmental Orientation
1	upright	90º
2	upside-down	90º
3	upright	-90º
4	upside-down	-90º
5	90º	upright
6	90º	upside-down
7	-90º	upright
8	-90º	upside-down

[4] This was to keep the experiment short enough to complete in a reasonable time frame, since each face appeared 8 times in Experiment 2 compared to 5 times in Experiment 1. The 4 faces we eliminated for Experiment 2 were chosen based on pilot subject ratings (1-10 scale) of how happy and sad the expressions looked. We selected the two male and two female faces that scored lowest on these ratings.

Note that when participants lay on their right side to view these images, faces that were oriented upwards or downwards in one frame of reference (i.e., rotated 90º or -90º in that frame) were always either perfectly upright or perfectly upside-down in the other frame of reference[5]. This decoupling of the environmental and egocentric reference frames allowed us to investigate independent spatial-valence congruency effects in both frames of reference.

Results & Discussion

The data from four participants were removed from analysis because the computer crashed mid-session (N=1), the participant was under 18 and could not give legal consent to participate (N=2[6]), or the participant's error rate was extremely high, representing a clear outlier (32% errors; N=1). Accuracy for the remaining 81 participants was quite good (*M*=96%, *SD*=3.14). Response times less than 200 milliseconds and greater than five standard deviations above the overall mean RT across all participants and trials were removed from analysis (<1% of all trials).

Past research suggests that there are independent effects of spatial orientation on face perception in the environmental and egocentric reference frames (Davidenko & Flusberg, 2012). Therefore, we analyzed the trial data separately for each frame, including only those trials where participants correctly identified the emotional facial expression.

Environmental Frame We first conducted a 2 (emotion: happy vs. sad) X 2 (orientation: -90º in the environment vs. 90º in the environment) repeated measures ANOVA with mean RT as the dependent variable. There was no main effect of emotion, as participants had similar reaction times to happy and sad faces, $F(1, 80)=0.75$, $p=0.39$. There was a marginal effect of orientation, as participants were slightly slower to respond to faces looking upwards in the environment (90º) compared to faces looking downwards in the environment (-90º), $F(1, 80)=3.29$, $p=0.074$. Crucially, there was a significant metaphor-congruent interaction between emotion and orientation, $F(1,80)=4.48$, $p=0.038$, $\eta^2=.053$: participants were marginally slower to respond to sad faces gazing upwards compared to happy faces gazing upwards, $t[80]=1.82$, $p=.073$, and to sad faces facing downwards, $t[80]=2.94$, $p=.004$, in the environment; there was no difference in recognition time for downward facing faces by emotional expression, $t[80]=.14$, $p=.887$; participants responded similarly fast to happy faces facing upwards and downwards in this environment, $t[80]=.36$, $p=.719$ (see Figure 4).

Egocentric Frame We repeated this analysis for trials where the faces were oriented upwards or downwards in the egocentric frame of reference. There were no main effects of emotion or orientation, nor was there an interaction between the two (all F's < 0.2, all p's > 0.7; see Figure 4).

In both the environmental, $F(1, 80)=27.93$, $p<.001$, $\eta^2=.259$, and egocentric, $F(1, 80)=6.53$, $p=.013$, $\eta^2=.075$, frames of reference, analyses of error rates revealed a main effect of emotional expression: in both frames of reference people were more accurate recognizing happy faces.

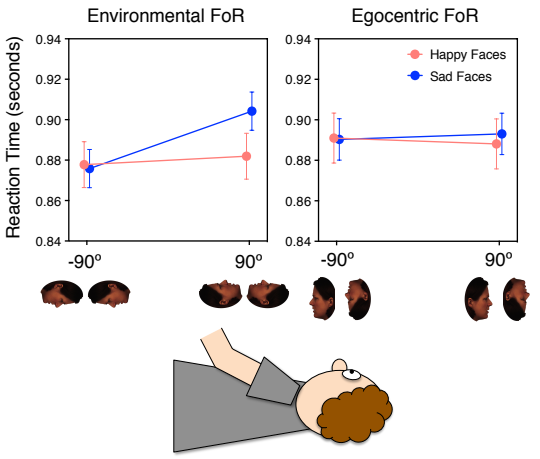


Figure 4. Mean RTs for happy and sad faces for each stimulus orientation and each frame of reference (environmental on the left) in Experiment 2. Error bars represent 95% CIs.

General Discussion

Spatial metaphors for affective valence are common in English, where *up* in space connotes happy or positive feelings ("things are looking *up*!") and *down* in space connotes sad or negative feelings ("I'm *down* in the dumps"). Past research suggests that this association is not merely a matter of language; rather, it offers a window into how people (metaphorically) represent the concept of emotional valence. For example, people are faster to respond to positive and negative words and faces when they are presented in metaphor-congruent regions of space (Lynott & Coventry, 2013; Meier & Robinson, 2004). In the present study, we explored whether the *orientation* of a stimulus – rather than its *position* in space – is sufficient to elicit such spatial-valence congruency effects, and, if so, which spatial reference frame people use to represent this orientation.

In Experiment 1, participants viewed images of happy and sad profile faces in different orientations and had to identify the emotion depicted in each face. Results revealed a significant spatial-valence congruency effect on performance: participants were faster and more accurate to respond when faces were oriented towards metaphor-

[5] As it turns out, when people tilt their head to one side, their eyes rotate several degrees in the opposite direction, a phenomenon known as Ocular Counter-Roll (OCR). At 90º rotation, this effect is very small (roughly 4º), and it does not appear to explain the effects of environmental orientation on face processing (see Davidenko & Flusberg, 2012). Because we observe an interaction between emotion and orientation in our data, OCR cannot account for our findings, since it should equally affect all faces.

[6] A valuable legal and ethical lesson for undergraduate research assistants!

congruent regions of space. In Experiment 2, participants completed the same task while lying down on their side, thereby disassociating environmental and egocentric reference frames. Results indicated that this spatial-valence congruency effect was only reliable in the environmental frame of reference, suggesting that (metaphorical) representations of the spatial dimension of emotional valence are constructed with respect to the environment.

Of course, multiple environmental reference frames were conflated in the present study design (e.g., faces that were environmentally upright were upright with respect to the computer display, the lab room, and the directional pull of gravity), so future work is required to tease apart which one(s) people are using to structure affective valence. Nonetheless, the present findings are notable in part because other research has found that effects of spatial orientation on face perception and memory are typically larger in the egocentric frame (e.g., Davidenko & Flusberg, 2012; Troje, 2003).

Also of note, in both experiments the spatial-valence congruency effects appeared to be driven by a decrease in performance for sad faces looking upwards in the environment. As mentioned in the discussion of Experiment 1, this result is somewhat surprising, as other researchers have observed similar congruency effects driven by a relative *increase* in performance for happy faces and positively valenced words that appear in higher regions of space (Lakens, 2012; Lynott & Coventry, 2013). One possibility is that differences in stimuli may account for these disparate findings: It may be that people simply respond to profile faces differently than they do to front view faces and common English words, perhaps due to the fact that judging emotions based on profile views is not a common activity.

No matter the explanation, these data may pose a challenge to some theories that have been put forth to explain spatial-valence congruency effects. In particular, the "polarity-based" perspective suggests that stimulus dimensions (including space and valence) are always anchored at a default endpoint (+pole) that is typically more frequent and unmarked linguistically (Lakens, 2012; Lynott & Coventry, 2013). In the case of valence, for example, "happy" is the default +pole (you can negate the unmarked term *happy* – *un*happy – but not the term *sad*; *un*sad is not an English word). The polarity account attributes spatial-valence congruency effects to a generic processing advantage for +polar items, and since *up* is another example of a +polar endpoint, this means that people should be fastest to respond to happy stimuli in higher regions of space (but see Dolscheid & Casasanto, 2015, for evidence against a universal polarity correspondence account of metaphor-congruency effects). In the present study, however, we do not observe this sort of processing advantage, but rather a processing *cost* for sad faces oriented towards the metaphor-incongruent upper-regions of the environment.

That being said, there is one way to interpret the present findings that would actually support the polarity account: if people are generally much worse at perceiving upward than downward gazing faces, then in fact it *would* be the case that we are seeing a processing advantage for happy faces looking upwards. However, based on other research on orientation effects in face processing and other (unpublished) findings from our lab, we do not think this is the most parsimonious explanation of the current findings. Still, more work may be required to fully rule out this possibility, and to fully explain why different performance asymmetries emerge in studies of spatial-valence congruency.

References

Brookshire, G., Ivry, R., & Casasanto, D. (2010). Modulation of motor-meaning congruity effects for valenced words. In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1940-1945). Austin, TX: Cognitive Science Society.

Casasanto, D., Dijkstra, K. (2010). Motor action and emotional memory. Cognition, 115, 179 – 185.

Davidenko, N. & Flusberg, S. J. (2012). Environmental inversion effects in face perception. *Cognition, 123*(3), 442-447.

Dolscheid, S., & Casasanto, D. (2015). Spatial congruity effects reveal metaphorical thinking, not polarity correspondence. *Frontiers in Psychology*, *6*(1836), 1-11. doi: 10.3389/fpsyg.2015.01836

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 726–736

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago and London: University of Chicago Press.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces - KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Lynott, D., & Coventry, K. (2013). On the ups and downs of emotion: testing between conceptual-metaphor and polarity accounts of emotional valence–spatial location interactions. *Psychonomic Bulletin & Review, 21*; 218-226.

Meier, B.P., Robinson, M.D. (2004). Why the sunny side is up, associations between affect and vertical position. Psychological Science, 15, 243 – 247.

Peirce, J.W. (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2):8-13.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologia, 128*(2), 274–289.

Troje, N. F. (2003). Reference frames for orientation anisotropies in face recognition and biological-motion perception. *Perception, 32*(2)*,* 201–210.

Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so

Kathryn D. Schuler (kds71@georgetown.edu)
Department of Neurology, 4000 Reservoir Road NW
Washington, DC 20057 USA

Charles Yang (charles.yang@ling.upenn.edu)
Department of Linguistics, Computer Science & Psychology, 3401 Walnut Street
Philadelphia, PA 19104 USA

Elissa L. Newport (eln10@georgetown.edu)
Department of Neurology, 4000 Reservoir Road NW
Washington, DC 20057 USA

Abstract

During language acquisition, children must learn when to generalize a pattern – applying it broadly and to new words ('add *–ed*' in English) – and when to restrict generalization, storing the pattern only with specific lexical items. One effort to quantify the conditions for generalization, the Tolerance Principle, has been shown to accurately predict children's generalizations in dozens of corpus-based studies. This principle hypothesizes that a general rule will be formed when it is computationally more efficient than storing lexical forms individually. It is formalized as: a rule *R* will generalize if the number of exceptions does not exceed the number of words in the category *N* divided by the natural log of *N* (*N*/ln*N*). Here we test the principle in an artificial language of 9 nonsense nouns. As predicted, children exposed to 5 regular forms and 4 exceptions generalized, applying the regular form to 100% of novel test words. Children exposed to 3 regular forms and 6 exceptions did not extend the rule, even though the token frequency of the regular form was still high in this condition. The Tolerance Principle thus appears to capture a basic principle of generalization in rule formation.

Keywords: artificial language; language acquisition; productivity; morphology; computational modeling.

Introduction

When children learn a language, they do not just memorize words or sentences; they acquire the patterns by which words and sentences are formed. We call these *rules* – for example, "add *–ed*" for the past tense of a verb or "add *–s*" to make a noun plural. However, in some cases there are not broad patterns for how words change their form; sometimes there are many idiosyncratic ways in which individual words form the plural or the past tense, such as *mouse* / *mice* or *go* / *went* – and these must indeed be memorized.

Several types of evidence show that children acquire rules when they are available. For example, young children make overgeneralization errors, extending rules to words that are actually lexical exceptions (e.g. *daddy goed to the store; I saw two mouses yesterday)* (Marcus et al. 1992, Pinker 1992, 1995, Yang 2002, Maslen et al. 2004). Furthermore, when children are asked to produce novel lexical forms in controlled experiments, they can spontaneously apply these rules to nonsense words they have never heard before (Berko 1958). In the famous 'wug' test, children were exposed to novel items from various linguistic categories and were asked to provide their inflected forms. (For example, *This is a wug. Now there are two of them. There are two* ___.) Children demonstrated that they had acquired productive rules by applying regular inflections (e.g. *wugs*) in these novel cases (Berko 1958).

However, not all linguistic rules are productive. Some inflected forms are idiosyncratic to a single lexical item, as noted above. In addition, some rules apply to only a restricted subset of lexical items, like *sing/sang/sung* and *ring/rang/rung*, but are not productive in that they do not apply broadly to new words. Children presented with novel forms like *gling* (similar to irregulars like *ring* and *sing*) do not produce *glang* as the past-test form (Berko, 1958). Though there are a handful of examples of irregular rules being generalized during language acquisition (e.g. *wipe-wope*, *think-thunk*) (Pinker 1999), this type of over-irregularization is unattested in the CHILDES corpus (Yang 2015). In other data, over-regularization errors involving *-ed* are relatively common (about 8%; Maslen et al. 2004), whereas analogical errors following irregular patterns are exceedingly rare (no more than 0.2%; Xu & Pinker 1995). Importantly, this absence of over-irregularization is not exclusive to English acquisition; it has been documented in children's naturalistic production data in many other languages (see e.g. Clahsen and Penke 1992 on German, Allen 1996 on Inuktitut, Clahsen et al. 2004 on Spanish, Caprin and Guasti 2009 on Italian, Demuth 2003 on Bantu languages, Deen 2005 on Swahili, among others).

The tendency to generalize some rules but to restrict others motivates the question: what governs when children will form productive rules during language acquisition? Researchers have been investigating rule learning in language acquisition for many years, but most work has focused on the difference between learning rules and learning the exceptions (not on the difference between learning regular and irregular rules). Though it is generally agreed that learners must memorize the idiosyncratic exceptions (e.g., *go-went*), how regular and irregular rules are handled is still debated. Some have proposed that the irregular rules must be memorized in the same way that the exceptions are (Pinker and Ullman 2002). Others have proposed that everything must be memorized, including both the regular and irregular rules (McClelland and

Patterson 2002). While this argument is not resolved, the available behavioral evidence, most clearly from children's production, points to a strong distinction between rules that are productive and those that are not. There must therefore be a mechanism during learning that governs when productive rules will be formed.

Recently Yang has proposed the Tolerance Principle (2005, 2016) – a model of productivity based on the acquisition literature that takes this categorical distinction into account. It quantifies the precise number of exceptions that a productive rule can tolerate before it becomes computationally less efficient than storing all of the lexical items individually. The Tolerance Principle accurately predicts children's generalizations in corpus data from dozens of rules/patterns in a number of languages, demonstrating that it is a viable model of productivity in language acquisition. However, in order to have adequate data for predictions, corpus analyses combine data from multiple children at different ages, not all of whom may show the same behaviors; and one can only test those patterns that happen to occur in real languages. Here we ask whether the Tolerance Principle can predict when children will generalize a productive rule in an artificial language learning experiment, where we can manipulate the precise number of lexical items that obey a rule or are exceptions. Our results indicate that children do indeed form a productive rule when the Tolerance Principle predicts that they will, applying the regular form to 100% of novel test words. When the Tolerance Principle predicts that no productive rule should be formed, children do not extend the rule, even though the token frequency of the most regular form was still high in this condition. In contrast, adult participants learning the same artificial language appear to extend the rule at the same level as the rule's token frequency in the language, approximating a well-studied phenomenon known as *probability matching* (Hudson Kam & Newport, 2005, 2009). We present this as evidence that the Tolerance Principle appears to capture a basic principle of generalization in rule formation in children, and suggest that adults adopt a different strategy during rule learning.

The Tolerance Principle

The Tolerance Principle (Yang, 2005, 2016) is a learning model that quantifies the precise conditions for generalization during language acquisition. It hypothesizes that a general rule will be formed when doing so is computationally more efficient than storing lexical forms individually. The model computes this computational efficiency by calculating the time complexity of applying a rule compared with accessing individual lexical forms. To illustrate, imagine that a learner is faced with a potential rule – for example, the English 'add *–ed* to make a verb in the past tense.' The English learner has encountered many items that obey this rule (regular forms) as well as many items that do not (irregular forms or exceptions). To be maximally efficient in formulating the past tense of verbs, the learner can do one of two things:

(1) *Store all lexical forms individually*: store every item individually in a list ranked by frequency, searching the list every time there is an occasion to express the past tense of a verb.
(2) *Form a productive rule*: store only the exceptions in a frequency-ranked list. To express the past tense, the learner searches the list of exceptions first. If the target verb is not among these exceptions, the learner applies the rule 'add *–ed*.'

The Tolerance Principle computes the time complexity required for each of these operations and assumes that the learner will adopt the optimal (i.e., faster) strategy. Productive rules, then, are formed only when it is more computationally efficient for the learner.

Formally, if *R* is a rule that may apply to *N* lexical items and there are *e* exceptions to this rule, the time required to access the rule can be expressed as $T(N, e)$. If *R* is productive, as in (2) above, then the rule is not applied until the learner has first evaluated and rejected every exception (*e*) on the list. In other words, applying a productive rule consumes *e* units of time. The time required for exceptions, on the other hand, is determined by the lexical item's frequency (i.e., its rank in the list of exceptions). To compute the time complexity $T(N, e)$, Yang (2016:48) calculates "the weighted average of time units over the probabilities of these two sets of items." If R is unproductive, as in (1) above, then all *N* items are treated as exceptions and are listed in order of frequency. The time complexity under these circumstances can be expressed $T(N, N)$, as the number of exceptions *e* is equivalent to the number of items in the list *N*. It is conjectured that the learner compares the time complexity required to form a productive rule, $T(N, e)$, with the time complexity required when all *N* items are stored individually as lexical exceptions, $T(N, N)$. By solving this equation for *e,* the Tolerance Principle computes the precise number of exceptions that a productive rule can tolerate before its formation becomes computationally inefficient.[1] This solution is as follows:

(3) **Tolerance Principle**: Let *R* be a rule that is applicable to *N* items, of which *e* are exceptions. *R* is productive if and only if $e \leq \theta_N = N/\ln(N)$.

In other words, it is only more efficient to form a productive rule when the number of exceptions is less than the number of items divided by the natural log of the number of items. To illustrate, imagine a category of 9 items. Given a rule *R* that may apply to these 9 items, the Tolerance Principle predicts that 4.096 (or $\theta_9 = 9/\ln 9$) exceptions will be tolerated before forming a productive rule becomes less efficient than storing individual items. This means that learners will form a productive rule if there are 4 or fewer exceptions to the rule *R*, but not if there are 5 or more. Importantly, this implies that the distinction between forming a productive rule and storing individual lexical items is a categorical one.

[1] See Yang (2016) for the full derivation of the Tolerance Principle and the supporting psycholinguistic evidence from lexical and morphological processing.

There is a theoretical tipping point at which forming a productive rule becomes less computationally efficient than the alternative strategy. The Tolerance Principle allows us to compute this tipping point.

How well does this model hold up to empirical investigation? Yang has tested the model on corpus data in a number of rule acquisition scenarios and found that the Tolerance Principle predicts productive rule formation surprisingly accurately. For example, consider the English productive rule 'add *–ed* to make a verb past tense.' Yang analyzed the 5 million words of child-directed English from CHILDES (MacWhinney 2000) and found 1022 unique past-tense verbs. By the Tolerance Principle, the English 'add *–ed*' rule should tolerate 147 verbs that are exceptions in a class of 1022 lexical items. Yang's analysis found only 127, well below the tipping point (θ_{1022}=147) for computational efficiency. The irregular rules/patterns in English, however, do not fare so well. Even the irregular class that has the highest homogeneity, the *ing-ang* class (such as *sing-sang* and *ring-rang*), has too many exceptions. The CHILDES English input corpus has 8 verbs that end in *ing*, but only three change the past tense to *ang* (*ring*, *sing*, *spring*) and five do not (*bring*, *fling*, *sting*, *swing*, *wing*), exceeding the threshold of θ_8=3. Thus the *ing-ang* pattern is predicted to be unproductive, in accord with children's productions described in the Introduction.

Though Yang has provided substantial evidence from corpus analyses to validate his account, further empirical investigation is necessary to demonstrate that children indeed follow the Tolerance Principle in acquiring productive rules during language acquisition. Here we will apply two well-known acquisition paradigms – artificial language learning and the "wug" test – to submit the Tolerance Principle to further experimental scrutiny. We use an artificial language paradigm to precisely control the input to child learners, providing them with highly controlled frequencies and numbers of words that follow a rule and words that are exceptions. This allows us to create conditions where the Tolerance Principle predicts productive rule formation (as in the *–ed* example above), and where it predicts that a rule will be unproductive (as in the *ing-ang* example above). We ask whether the Tolerance Principle correctly predicts when a pattern in an artificial language is widespread enough for a child to form a productive rule, using a "wug" test to assess whether children have formed a productive rule (one that applies to novel lexical items) or have restricted generalization.

Behavioral Data

Participants

Fifteen children (mean = 7.48 years, range = 5.08 – 8.92 years) and twenty adult controls participated in this experiment. An additional 3 children began the experiment but did not complete it, and an additional 4 children participated but were excluded from analysis for failure to understand the task (quantified as a failure to produce the correct noun on at least 50% of the test trials). Children were recruited from Washington DC metro area schools and were run either at their school or in our lab. Adult participants were recruited and run online using Amazon Mechanical Turk[2]. All participants were native English speakers with normal hearing and normal to corrected-to-normal vision. Child participants received stickers and a set of small toys for their participation. Adult participants received compensation at a rate of $10/hour.

Stimuli

Description of the language We designed two artificial languages: one in which the Tolerance Principle predicts that learners should form a productive rule and one where learners should not form a productive rule. To do so, we first created a rule *R* for a category of 9 nonsense nouns. The rule was: "To make a noun plural, add *ka*." Next we used the Tolerance Principle to calculate the number of regular forms vs. exceptions a productive rule can tolerate in a category of 9 nouns. Using the predicted value of 4.096 exceptions, we created two conditions: one where a productive rule should be formed (5 regulars, 4 exceptions), and one where a productive rule should not be formed (3 regulars, 6 exceptions).

To create our exposure corpus, we assigned each noun a plural marker that either followed the rule (add *ka*) or was an exception (add *po, tay, lee bae, muy,* or *woo*), depending on the condition. Then we used these nouns and markers to create an exposure corpus of 72 sentences (24 singular and 48 plural). All sentences began with the same nonsense verb *gentif*, meaning "there is/are". Singular sentences were unmarked ("*gentif* + NOUN") and paired with one image of the corresponding object. Plural sentences were formed "*gentif* + NOUN + MARKER" and paired with 2, 4, or 6 images of the corresponding object. There were thus 18 possible sentences in the language: 1 singular and 1 plural sentence for each noun.

We generated the exposure corpus by allowing noun frequency to vary along a Zipfian distribution, with nouns taking the regular form (*ka*) as the most frequent in both conditions. Thus the second most frequent noun was presented half as often as the most frequent noun, the third most frequent noun was half as often as the second, and so on. This is important because the distribution of word frequency in natural language is approximately Zipfian, and the computation underlying the Tolerance Principle assumes that word frequency follows this pattern. Making the regular form the most frequent ensured that its token frequency was high in both conditions.

Procedure

Exposure Each participant was presented with the 72 exposure sentences in random order. On each trial, participants saw a picture of 1, 2, 4, or 6 instances of a noun

[2] Adults run on Mechanical Turk give the same results as adults run in the lab on the same paradigm as children.

and hear the corresponding singular (for 1) or plural (for 2, 4, or 6) sentence. They were asked to repeat the sentence aloud (or type it into a response box on Mechanical Turk) before moving on to the next trial. Every 18 trials they were given a short break. Children were offered a sticker during breaks to encourage them to continue in the experiment.

Production Test After exposure, we used a "wug" test to assess whether children had formed a productive rule (Berko, 1958). During this test, participants were given singular sentence-image pairs containing novel nonsense nouns they had not heard during exposure and were asked to provide the plural form. Each participant completed 12 production test trials, 2 for each of 6 novel nouns. To prevent participants from using a plural form based only on the precise number of instances shown in that trial, the test items contained 3 or 5 instances of the novel noun (whereas there were 2, 4, or 6 in the exposure set). Finally, all participants were given a rating test to ensure they had learned the nouns and markers they were exposed to.

Results & Analysis

For each production test trial, participants were asked to produce the plural form of a novel noun they heard only in a singular form. These novel productions allowed us to assess whether participants formed a productive rule. Recall that the Tolerance Principle predicts that there will be a categorical distinction between productive and unproductive (lexically specific) rules. In our artificial language, a productive rule should be formed if more than 4.096 nouns obey the rule (as in our 5 regular/4 exception condition), but not if fewer than 4.096 nouns do (as in our 3 regular/6 exception condition). When a productive rule is formed, it should be applied to 100% of novel nouns, as is the case for English past tense 'add *–ed*.'

To determine whether participants formed a productive rule, we performed a one-tailed t-test against the hypothesized value of 100%. Participants who have formed a productive rule should, according to the Tolerance Principle, mark these novel plural sentences with *ka* 100% of the time. On the other hand, participants who have not formed a productive rule should use the *ka* inflection significantly less than 100% (and perhaps no more frequently than other inflection forms are used).

Focusing first on the child production data, Figure 1 shows the percentage that each inflection type was produced during the production test for participants in the 5 regular/4 exception condition (5R/4E) and the 3 regular/6 exception condition (3R/6E). These data show that children in the 5R/4E condition mark novel nouns with the *ka* inflection on 91.7% of plural trials. This value is not statistically different from 100% (t=1.00, p=0.18). In contrast, in the 3R/6E condition, children mark novel nouns with the *ka* inflection on only 16.9% of plural trials. This value is substantially and significantly different from 100% (t=6.81, p<0.0001). Children thus appear to have formed a productive rule when the Tolerance Principle predicts that they will (in the 5 regular / 4 exception condition), but not when it predicts that they will not (in the 3 regular / 6 exception condition).

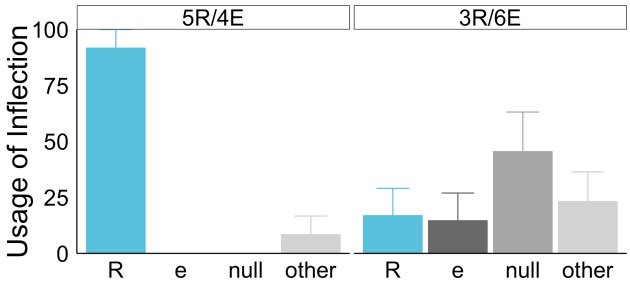


Figure 1: Percentage of each type of inflection added to novel nouns by children when their exposure contained 5 regulars/4 exceptions compared with 3 regulars/6 exceptions. *R* = the regular form; *e* = any exception form; null = no plural marker (unmarked); other = any marker not present in the exposure (e.g. English [+s]).

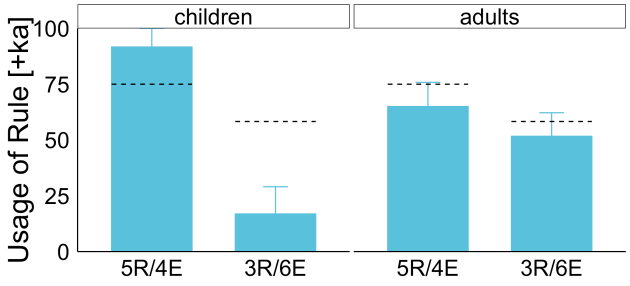


Figure 2: Percentage of regular inflection *ka* applied to novel nouns by children and adults when their exposure contained 5 regulars/4 exceptions compared with 3 regulars/6 exceptions. Dashed line indicates the token frequency of the *ka* inflection in the input.

Table 1: Number of children who applied the rule to 100% of plural test nouns or to fewer than 100% of plural test nouns in each condition.

	5R/4E	3R/6E
# children using rule 100%	6	1
# children using rule <100%	1	7

This strong result is further underlined by looking at the data from individual children. Table 1 shows the number of children who used the *ka* inflection on 100% of test trials in the 5 regular/4 exception condition compared with the 3 regular/6 exception condition. Six out of 7 children produced the *ka* inflection on 100% of production trials in the 5 regular/4 exception condition, while only one out of 8 children did so in the 3 regular/6 exception condition.

Turning next to the adult production data, we find a somewhat different pattern of results. Adults in the 5R/4E condition mark novel nouns with the *ka* inflection on 65.0% of plural trials; unlike children, this value is significantly different from our 100% productivity criterion (t=3.23,

$p<0.01$). Like children, adults in the 3R/6E mark novel nouns with *ka* significantly less than 100% of plural trials ($t=4.59$, $p<0.001$) - only 51.7%. That is, for adults, this contrast is much weaker, and when we compare the use of the *ka* inflection between adults to the children, as in Figure 2, we see striking differences between the two. The Tolerance Principle effect is much more pronounced in children, who exhibit a much more categorical response in their use of the *ka* inflection. Indeed, children, but not adults, show a significant difference between the use of *ka* in the 5R/4E condition and that in the 3R/6E condition (children: $t=4.91$, $p<0.001$, adults: $t=0.89$, $p=0.39$).

One possible explanation is that adults are not obeying the Tolerance Principle and are instead producing *ka* with the same frequency they heard this inflection in their input. This behavior is known as *probability matching* (Hudson Kam & Newport, 2005, 2009). Recall that the nouns in our artificial language follow a Zipfian distribution, with rule-following nouns being the most frequent in the distribution. Thus the token frequency of the *ka* inflection is fairly high in both conditions: 75% of the plural exposure sentences in the 5R/4E condition, and 58.3% of plural exposure sentences in the 3R/6E condition. To determine whether this could explain the difference between the two groups, we analyzed both child and adult use of *ka* against the token frequency of *ka* in the exposure for the two conditions. We found that only adults match the token frequency in both the 5R/4E ($t=0.92$, $p=0.19$) and 3R/6E conditions ($t=0.63$, $p=0.27$). In contrast, the child data is not consistent with a probability matching interpretation. Children in the 5R/4E condition produce the *ka* inflection significantly more than the input frequency ($t=2.00$, $p<0.05$) and in the 3R/6E condition produce the *ka* inflection significantly less than the input frequency ($t=3.40$, $p<0.01$).

Conclusions & Discussion

Here we have asked whether the Tolerance Principle accurately predicts when children will generalize a productive rule in an artificial language learning experiment and when they will restrict generalization. The Tolerance Principle is based on the hypothesis that productivity emerges when it is the most efficient strategy for learners to access lexical forms. The model allows us to calculate the number of lexical exceptions there can be to a given rule before it becomes more efficient for the learner to simply memorize each lexical form individually. For our 9-noun artificial category, the Tolerance Principle predicts that a productive rule will be formed when there are fewer than 4.096 lexical items that are exceptions. We found that, just as predicted, learners formed a productive rule when there were 4 lexical items that were exceptions to the rule, but not when there were 6. These results suggest that the Tolerance Principle has accurately captured something significant about the conditions for generalization during learning.

Importantly, the criterion we used to assess whether learners formed a productive rule was categorical. Our analysis asked whether learners extended the rule to 100% of the test trials – the most rigorous possible test of productivity. We found that, while both children and adults were more likely to extend a productive rule when there were 4 exceptions than when there were 6, only children displayed a categorical distinction between forming a productive rule and not forming one. As predicted by the Tolerance Principle, almost every child exposed to 5 regulars/4 exceptions extended the rule to 100% of test trials, while almost no children exposed to 3 regulars/6 exceptions did (see Table 1). These results for children are in accord with those cited above for regulars like *–ed* versus irregulars like *sing/sang*. While children in the 3R/6E condition did occasionally use *ka* on novel items, their results are only what one would expect under conditions of uncertainty. In the absence of a productive rule, children appear to be unsure how to mark novel forms, most often using no plural marker (null) and otherwise selecting at random from among the various markers they heard during exposure (the *ka* inflection as well as the exceptions).

The striking difference we observed between children and adults, shown in Figure 2, led us to ask whether an alternative model would better predict adult behavior in this task. The Tolerance Principle is based on the number of lexical items (types) that observe or violate a pattern, which is consistent with most approaches to productivity from a wide range of perspectives (e.g., Plunkett & Marchman 1991). By contrast, adults appear to follow more closely the token frequencies. Recall that the token frequency of the *ka* marker was high in both the 5R/4E and the 3R/6E conditions, due to the Zipfian distribution of noun frequency in our exposure corpus – 75% and 58.3%, respectively. We found that adults produced the *ka* marker with the same frequency they heard this marker in their input. This *probability matching* behavior has been observed in adult learners in many other experiments in our lab (e.g. Hudson Kam & Newport, 2005, 2009). There are several ways to interpret the differences between children and adults in the context of computational efficiency, particularly given the greater cognitive resources of adult learners (Newport, 1990). One possibility is that probability matching may be the more efficient computational strategy for adult learners. This would imply that the Tolerance Principle is exclusive to children, capturing a basic principle of generalization in rule formation for very young learners. A related possibility is that only children learn or produce forms categorically. On this interpretation, the difference between 5R/4E and 3R/6E influences both child and adults learners, but only children show this difference in such an extreme contrast in their output. A final possibility is that perhaps adults only engage probability-matching behavior when they are learning from a very small number of items; probability matching may be easy and efficient for adults when there are only a few items to keep track of. However, when there are many items to track, it may no longer be efficient to track and closely match input probabilities. This latter interpretation is supported by the results in studies on a different topic, the effect of inconsistent input. In these

studies children and adults look very different in experiments when the learning task involves only a small number of items, with adults probability matching and children maximizing the majority form in their productions. However, when the number and complexity of the items and their variations become very large, adults begin to behave more like children (Hudson-Kam & Newport 2009). In ongoing work we are exploring this by giving adult participants our Tolerance Principle tasks with a very large set of nouns. We hypothesize that, at this higher level of complexity, adults may begin to exhibit the type of categorical behavior predicted by the Tolerance Principle.

As mentioned above, children and adults behave differently both in the present experiments and in the literature on inconsistent input. Though our results suggest that adults adopt a similar strategy to perform these two different tasks, children appear to be handling them differently. In the literature on inconsistent input, children produce the form that appears most often in their exposure nearly 100% of the time (Hudson-Kam & Newport 2005, 2009). While the results of our 5R/4E condition are similar to these findings, the results from our 3R/6E condition are quite different. Although the *ka* marker is the most frequent marker in the 3R/6E exposure (58% of tokens), children produce this form with much lower probability than they were exposed to. This suggests that children are not simply forming productive rules based on what appears most frequently in their input, as they do when they are faced with patterns that are inconsistently marked or probabilistic. In the present paradigm, in contrast with our studies of inconsistent input, each lexical item is marked in a consistent way across trials. When children are exposed to lexically consistent patterns, they form productive rules based on the number of lexical items that observe these patterns, as predicted by the Tolerance Principle. To examine these issues further, we are conducting an experiment in which token frequency is matched across conditions and another in which the rule is assigned to very low frequency nouns.

As noted above, the computations behind the Tolerance Principle are based on the assumption that productive rules emerge when they are the most efficient strategy for learners. Successful models of many cognitive and neural processes are based on the same notion. This implies that the Tolerance Principle may be applicable to rule acquisition and generalization in domains other than language. Future work is required to determine whether this exciting prospect is empirically supported.

Acknowledgments

This research was supported by NIH grant HD-37082 to E. Newport and by Georgetown University. We thank Sarah Furlong, John Davis, and Breana Downey for their important assistance in recruiting and running the participants, and the members of the Language and Development Lab for helpful comments on the research.

References

Allen, S. (1996). *Aspects of argument structure acquisition in Inuktitut.* Johns Benjamins.

Berko, J. (1958). The child's learning of English morphology, *Word,* 14, 150-177.

Caprin, C. and Guasti, M.T. (2009). The acquisition of morpohosyntax in Italian: A cross-sectional study. *Applied psycholinguistics*, 20, 23-52.

Clahsen, H. and Penke, M. (1992). The acquisition of agreement morphology and its syntactic consequences.: New evidence on german child language from the simone corpus. In Meisel, J., editor, *The acquisition of verb placement*, pp 181-234. Kluwer, Dordrecht.

Clahsen, H., Aveledo, F. & Roca, I. (2002). The development of regular and irregular verbs inflection in Spanish child language. *Journal of Child Language,* 29, 591-622.

Deen, U. (2005). *The acquisition of Swahili*. John Benjamins Publishing.

Demuth, K. (2003). The acquisition of Bantu languages. *The Bantu languages*, pp 209–222. Curzon Press. Surrey, UK.

Hudson Kam, C.L., & Newport, E.L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1, 151-195.

Hudson Kam, C.L., & Newport, E.L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30-66.

Marcus, G., Pinker, S., Ullman, M.T., Hollander, M., Rosen, J., & Xu, F. (1992). *Overregularization in language acquisition.* Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

Maslen, R., Theakston, A.L., Lieven, E.V., and Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research*, 47(6): 1319-1333.

MacWhinney, B. (2000). *The CHILDES Project.* Lawrence Erlbaum.

McClelland, J. & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences,* 6, 465-472.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science*, 14, 11-28.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books, New York.

Pinker, S., and Ullman, M.T. (2002) The past and future of past tense. *Trends in Cognitive Science*, 6, 456-463.

Plunkett, K., and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception. *Cognition,* 38 (1): 43-102.

Yang, C. (2005). On productivity. *Linguistic variation yearbook*, 5(1), 265-302.

Yang, C. (2015). For and against frequencies. *Journal of Child Language.* 42, 287-293.

Yang, C. (2016). *The price of productivity*. MIT press.

Variability in category learning: The Effect of Context Change and Item Variation on Knowledge Generalization

Dustin D. Finch (ddfinch@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th St.
Bloomington, IN, 47401 USA

Paulo F. Carvalho (pcarvalh@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th St.
Bloomington, IN, 47401 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th St.
Bloomington, IN, 47401 USA

Abstract

We explore how context change and item variation during natural category learning influence memory and generalization to new examples. Participants studied either images of the same bird or varied birds from each of several categories. These images could be presented in a constant background color or different background colors. During test, birds were presented in only one of the studied background colors. Performance at test depended on the context overlap between study and test, with better performance when there was minimal context change during study. Also, contrary to previous findings, we found that learners generalized better when items were repeated during study and remembered old items better when items were varied during study. When there is a moderate degree of context change, there is no benefit of repetition or variation for either novel or old items. These results indicate that context change and item variation have complementary effects on learning.

Keywords: concept learning, memory, context, variability, repetition.

Introduction

In our every-day life, we have all recognized someone without being able to "place" the person or remember their name. For example, even though we may have been visiting their office for years, we might fail to remember the name of our dentist when we see her at the grocery store. A related feeling is reported by students in preparation for exams. They are able to effectively remember the content the night before the exam in their dorm room, but that knowledge seems to have vanished when they are tested in an exam context different from their dorm room study context.

These are real-life examples of a property of our memory system – context dependency. When trying to remember something, reinstating the conditions of the encoding situation (i.e., the context) helps retrieve the memory. For example, Godden and Baddeley (1975; see also, Godden & Baddeley, 1980; Smith, Glenberg, & Bjork, 1978) had people learn a list of words in one of two contexts: either underwater or on land. When asked to recall this list, people recalled more words when the testing context was the same as the study context (e.g., study done underwater and recall done underwater as well), compared to when the testing context was different (e.g., study done underwater and recall done on land). This effect has been repeatedly found across different tasks, paradigms and labs (see Smith, 2013 for a recent review and Smith & Vela, 2001 for a meta-analyses), including educational settings (Abernethy, 1940; but see Saufley, Otaka, & Bavaresco, 1985).

Why is memory context dependent? One leading theory is that as memories are created, background information (e.g., how you feel, the smell, sounds, or the room you are in) is encoded into that memory. This background information can then be used as cues to retrieve that memory (Hintzman, 1986; Bjork and Bjork, 1992; Smith 2013). Reinstating the study conditions at test makes these cues available, improving recall.

However, context dependency can have negative consequences, because a specific memory might become too tied to a specific context. For example, a student might study the mitochondrial structure in their room while listening to Pearl Jam. When later testing themselves on the mitochondrial structure while listening to Pearl Jam and still in their room, they are able to successfully retrieve the information studied. However, when asked to write about the process in a quiet classroom surrounded by 30 other students, retrieving this memory might be difficult. One possible solution to this issue is to decontextualize the specific memory.

Memory is considered to be decontextualized when removing the contextual cues associated with it does not affect retrieval. As the example above shows, decontextualization is particularly important when the test situation is different from the study situation. One way to do this is by varying study context and then testing people's memory in a new context (Smith et al., 1978; Smith, 1979; Smith & Handy, 2014; for evidence in classroom settings see Smith & Rothkopf, 1984). Smith et al. (1978) tested this by having a group of people learn a list of words in two separate

rooms and a different group who studied the words in one room only (both groups of learners were asked to leave the room and return during learning). All participants were tested in a novel room. People recalled more words when they studied in two different rooms compared to studying in one room only. Studying in two different contexts led to the creation of decontextualized memories that were more easily retrieved in a new context.

Thus, previous research has demonstrated that our memory includes information about the learning situation, which can have a positive or negative effect on later retrieval. A possible explanation for this effect is derived from the MINERVA 2 model (Hintzman, 1986). This model suggests that every memory is encoded into its own memory trace which includes both target and context information. When accessing previous memories, common properties between the traces will be more salient at retrieval. In this way, when information is studied in one context only, all encoded information (both contextual information as well as target content) are highly salient at retrieval. Conversely, if the target information is studied across several contexts, it will be more salient because it is the only common information across all contexts. This will contribute to good recall, even when context cues are not available at retrieval.

A similar idea has been proposed in the context of generalization of category learning to new items (Hintzman, 1984). A goal of most category learning is to extend a category learning experience to novel items (Goldstone, Kersten, Carvalho, 2013). One way to improve category generalization is by increasing the variation across the studied items. When people study a category by seeing several different examples of the category, they are better at generalizing this knowledge to new items of the category compared to when they see only a few items repeated (controlling for total exposure time; e.g., Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Posner & Keele, 1968; Wahlheim, Finn, & Jacoby, 2012). For example, learning the category dog by seeing examples of *German Shepherds, Bulldogs, Chihuahuas,* and *Poodles* would result in better generalization to new members of the category dog (e.g., *Pugs*) compared to studying only examples of *Chihuahuas*. Seeing several different items allows people to notice and extract the overlapping properties among all examples. This allows for better subsequent generalization (Hintzman, 1984).

In sum, both context change and item variation can be seen as different, perhaps complementary, ways of increasing study variability with positive results. Here we propose to extend previous research on context dependent memory and item variation to a situation where generalization of the knowledge is required. We analyze two main questions: (1) can context change improve not only memory retrieval but also knowledge generalization to novel items, and (2) can context change and item variation work together to result in improved generalization or is this high level of combined variability detrimental for memory and generalization?

We analyze these two questions by having learners learn bird species by studying examples of each species and manipulating the number of different items studied as well as the study context of the items. According to previous research with memory tasks, and the predictions of MINERVA 2, we expect to see better generalization in situations for which there is a high congruency between study and test contexts. Moreover, previous research suggests that both item variation and context variation will promote better generalization to new items at test, by increasing study variability and the detection of the category-relevant properties shared by members of a category (Hintzman, 1984).

An Experiment

In this experiment participants learned twelve different species of birds by studying examples of each species. We manipulated the context of study by changing the color of the background on which the bird was presented. In addition to manipulating the context of study we also manipulated the amount of item variation presented within each category.

Methods

Participants. A total of 238 undergraduate students at Indiana University volunteered to participate in this experiment in return for partial course credit. Participants were randomly assigned to one of the following conditions: No Context Change (N = 73), Medium Context Change (N = 82), and High Context Change (N = 83).



Figure 1: Examples of the birds shown during the study. The first row shows the same bird with the same background color, illustrating what participants would see in the Repetition and No Context Change Condition. The second row shows a different bird with the same background color, illustrating the Variation and No Context Change Condition. The third row shows a different bird in two different background colors, illustrating the Variation and Medium Context Change Condition. The last row shows different birds with different background colors, illustrating the Variation and High Context Change condition.

Stimuli & Materials. The stimuli were 72 images of birds from twelve different species (6 images from each bird species; see Figure 1 for examples), available as part of a database of bird images (Wahlheim, Dunlosky, & Jacoby, 2011). We selected the species Black-headed Grosbeak (*P. melanocephalus*), Blue Jay (*C. cristata*), American Goldfinch (*S. tristis*), American Tree Sparrow (*S. arborea*), Bank Swallow (*R. riparia*), American Robin (*T. migratorius*), Blue Bunting (*C. parellina*), Clark's Nutcracker (*N. columbiana*), Brambling (*F. montifringilla*), Bachman's Sparrow (*P. aestivalis*), Brown-chested Martin (*P. tapera*), and Bluethroat (*L. svecica*).

During the task, participants classified the images using the scientific classification of the species. We created four different versions of each image by changing the background color to one of the following: red, green, blue, pink (see Figure 1). Thirty different images were chosen to be presented during the study phase, and the remaining 42 images were presented only during test. Which images were chosen to be used during study and test was counterbalanced across participants. The stimuli were presented and responses from the participates were collected using a web-browser and the experiment software Collector (a free and open-source platform to run experiments using a web-browser available from https://github.com/gikeymarcia/Collector).

Procedure. An overview of the procedure used in this Experiment is presented in Figure 2. There were two phases to the experiment: Study and Test. Participants completed 48 study trials and 72 test trials.

During the study phase a picture of a bird and its species name would appear on the screen for eight seconds. Afterwards, the name was replaced by a Judgment of Learning prompt. Participants were asked to provide a judgment of learning on a 0 (0% confidence of correct classification)-100 (100% confidence of correct classification) scale and were told to try to use the whole scale while providing their judgments.

Immediately following the Study phase participants played Tetris for one minute until starting the test phase. During the Test phase, on each trial a bird was presented in the center of the screen, along with the name of the twelve species participants had studied. Participants were asked to click on the species name to which they believed the bird belonged. No feedback was given during the test.

For study and test, one third of the participants saw the images presented to them always in the same background color (No Context Change). The color of the background was counterbalanced across participants. Another third of the participants saw the images in two background colors (Medium Context Change). Half of the items in each category were in one color (e.g., red) and the other half in another color (e.g., green). Which two colors were selected was counterbalanced across participants. The last third of the participants always had a new background color for the items studied in each category (High Context Change). Participants never saw a repeated background color within each category, but colors were repeated across categories. Test items were presented always in the same background color – a color chosen from one of the studied colors and counterbalanced across participants.

During study we manipulated the number of different items participants saw for each category. For six out of the twelve categories, one item from each category was repeated four times (Repetition Condition). The other six categories had item variation in which each category had four different studied items (Variation Condition). Which item was repeated was counterbalanced across participants.

During Test we manipulated whether the item was old (studied) or was novel. For categories studied in the Repetition Condition participants saw one old item and five novel items. For categories studied in the Variation Condition, participants saw four old items and two novel items. This was done in order to keep constant the total number of test items across conditions. For both the study and test phases, trials were presented in random order.

		Study Phase				Old Items Test				Novel Items Test				
Context Change Condition	Variation Condition	Pres. 1	Pres. 2	Pres. 3	Pres. 4	Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4	Item 5
No Context Change	Repetition	Ex.1	Ex.1	Ex.1	Ex.1	Ex.1				Ex.2	Ex.3	Ex.4	Ex.5	Ex.6
	Variation	Ex.1	Ex.2	Ex.3	Ex.4	Ex.1	Ex.2	Ex.3	Ex.4				Ex.5	Ex.6
Medium Context Change	Repetition	Ex.1	Ex.1	Ex.1	Ex.1	Ex.1				Ex.2	Ex.3	Ex.4	Ex.5	Ex.6
	Variation	Ex.1	Ex.2	Ex.3	Ex.4	Ex.1	Ex.2	Ex.3	Ex.4				Ex.5	Ex.6
High Context Change	Repetition	Ex.1	Ex.1	Ex.1	Ex.1	Ex.1				Ex.2	Ex.3	Ex.4	Ex.5	Ex.6
	Variation	Ex.1	Ex.2	Ex.3	Ex.4	Ex.1	Ex.2	Ex.3	Ex.4				Ex.5	Ex.6

Figure 2: Schematic representation of the experimental design for study of the categories in each of the study conditions. Each square represents one example of a bird in the species. The color of the square represents the background color in which the bird was presented.

Results

Judgments of Learning Participants were more confident on their ability to correctly classify repeated items (M = 53.48, SD = 22.46) at test compared to varied items (M = 47.92, SD = 21.48), F (1,235) = 38.61, p < .0001. Moreover, participants' Judgments of Learning for repeated and varied categories were influenced by the degree of context change presented, F (2,235) = 5.52, p = .005. When participants studied in the No Context Change and Medium Context Change conditions, they were more confident in their ability to correctly classify repeated items (M = 55.91, SD = 21.06, and M = 55.95, SD = 22.71, respectively; both ps < .0001) compared to varied items (M = 47.59, SD = 20.80, and M = 48.80, SD = 21.85). However, when there was High Context Change, participants rated their ability to later classify both types of items equally (M = 48.90, SD = 22.94, and M = 47.35, SD = 21.93 for repeated and varied items, respectively; p = .349).

Test Performance We started by looking at the effect that changing backgrounds had on overall test performance (see Figure 3). The results show that the amount of Context Change had an overall effect on test performance, F (2,235) = 3.61, p = .029. Planned comparison tests indicate that studying items with four different backgrounds (M = 65.27, SD = 30.31) worsens performance compared to two backgrounds (M = 69.41, SD = 30.62; p < .0001), or one background (M = 69.76, SD = 30.41; p = .001). Participants' performance was equivalent when comparing one background to two backgrounds (p = .531).

Moreover, participants are better at classifying old items (M = 0.51, SD = 0.50), compared to novel items (M = 0.43, SD = 0.50), F (1,235) = 97.09, p < .0001 (see Figure 3). Importantly, item variation and repetition had different effects for old and novel items, F (1,235) = 42.36, p < .0001. When classifying novel items, participants are better for categories studied with item repetition (M = 0.45, SD = 0.50) compared to categories studied with item variation (M = 0.41, SD = 0.49; p = .026). Conversely, the opposite pattern is seen for old items (M = 0.47, SD = 0.50, and M = 0.54, SD = 0.50, for repetition and varition respectively; p < .0001).

Lastly, the effect of item variation depends on the background change condition and whether the stimulus is old or novel, F (2,235) = 14.14, p < .0001. To investigate this three-way interaction, we calculated the difference between performance for categories studied with variation and categories studied with repetition (Variation – Repetition). We used this score to investigate differences when classifying old and novel items and among different background change conditions (see Figure 4).

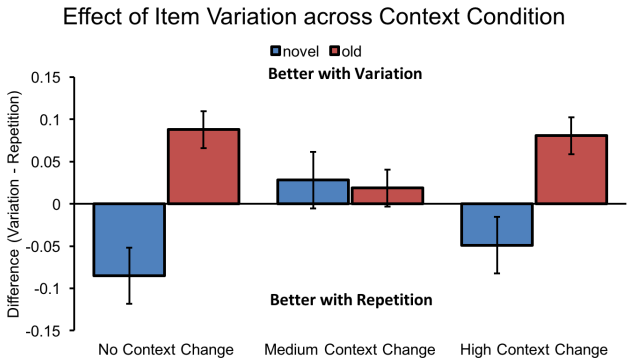


Figure 4: Difference in test performance for novel and old items between categories studied with variation and categories studied with repetition (Variation – Repetition). Positive values represent better categorization with item variation while negative values represent better categorization with item repetition. Error bars represent standard errors of the mean.

The results indicate that when studying items in the No Context Change or High Context Change conditions, pariticipants performed better at classifying novel items for categories studied with repetition, while for old items they performed better for categories studied with variation (both ps < .0001). However, when studying the species in the Medium Context Change condition, participants' performance was equivalent for old and novel items (p = .720), and equivalent for categories studied with repetition and variation for both old (t-test comparing with 0: t (81) = 0.72, p = .469), and novel items (t-test comparing with 0: t (81) = 1.08, p = .282).

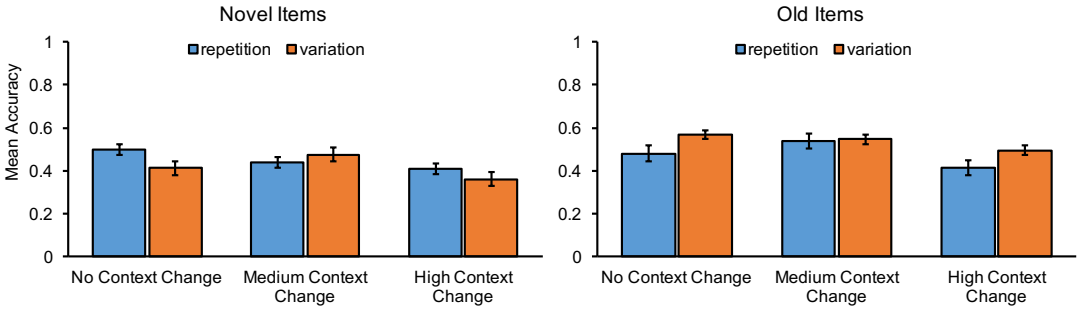


Figure 3: Performance during the test phase by amount of context change during study and item novely at test. Change performance in task is 0.08. Error bars represent standard errors of the mean.

Discussion

In this study we investigated how context change and item variation during study of natural categories (bird species) influence later memory and generalization to new examples.

We found evidence for context dependency in a generalization task. Learners performed better when there was No Context Change and Medium Context Change. As the overlap between study and test conditions (background colors) increased so did test categorization performance. For No Context Change and Medium Context Change conditions there was a high degree of overlap between study and test contexts. For High Context Change there is only a small degree of overlap between study and test. To the best of our knowledge, this is the first demonstration of this context reinstatement effect in a categorization task, and is in agreement with previous research showing a positive relationship between increased overlap of study and test conditions and memory performance (e.g., Smith et al., 2014). Moreover, it extends previous evidence to performance in classification of novel items (generalization) that go beyond remembering specific instances.

As mentioned in the introduction, previous research has shown that adding item variation improves both recognition memory for old items placed in new contexts and categorization of novel items. (e.g., Wahlheim et al., 2012). However, contrary to previous evidence, here we found that learners generalized better when items were repeated during study and remembered the categories of old items better when items were varied during study. There are several reasons for these results. It is possible that the context change manipulation and the fact that participants are studying categories with and without variation simultaneously might have impacted the results. It is possible that the change in contexts lead learners to try to group items into categories by background color, leading to overall lower attention to the similarities between items of the same category but with different background colors. However, the fact that we saw similar results for participants who studied the categories in one context only indicates that this might not be the only factor influencing these results.

Another possibility is related to the high degree of similarity within items of each category and between the studied items and the novel items presented at test. It has been proposed before that the benefits of item variation during study are related to the degree of dissimilarity within the category and between the studied and test items (Hahn, Bailey, & Elvin, 2005; Stewart & Chater, 2002). In our study, the items within each category were highly similar (see Figure 1). Moreover, the novel items used at test were very similar to the studied items as well. This high level of similarity might have reduced the possible benefits of increased item variation at study, because it is relatively easy to identify similarities across items of the same category during study and to extend this knowledge to very similar items at test. It is possible that if more dissimilar items had been used at test and/or during study the results would have been different. Post-hoc analyses looking at the effect of item similarity on test performance suggest that this might be the case. We used an indirect measure of category similarity, how confusable a category is with another category, based on participants' Reponses to compare similar and dissimilar categories. When we compared test performance between categories frequently confused with another category (e.g., *C. parallina* and *C. cristata*) and categories that were not confused with each other (e.g., *L. svecica and N. columbiana*), we see that the pattern of results described is reversed only when the categories are easier to discriminate (not as frequently confused). For these categories, participants remember better the categories of old items repeated during study and generalized the categories to novel items better when items were varied during study.

It is, however, interesting to note that item variation improved memory for old items in this experiment relative to repeating the old items. To our knowledge, this is the first demonstration of such an effect. Item variation might have worked to promote more efficient episodic memory for each of the items presented, making them more distinct and therefore more memorable (Nairne, 2006; Rawson & Van Overschelde, 2008; Schmidt, 1985; von Restorff, 1933). Congruent with this hypothesis, the benefit of item variation is numerically larger when there is No Context Change during study, compared to when there is High Context Change, perhaps because context change in itself would also work to make items more memorable by increasing attention to the item. However, learners seem to be unaware of the potential benefits of variation, generally showing higher degrees of confidence in their ability to later categorize items studied repeatedly than varied items, similar to what has been shown in previous research (e.g., Wahlheim et al., 2012).

Finally, the relative benefit of item *repetition* compared to item *variation* when classifying old and novel items at test is modulated by the amount of context change present during study. We found that when there is a medium degree of context change (i.e., half the trials were presented in one context and the other half in another context), there is no benefit of repetition or variation for either novel or old items. One possible reason for these results is that the partial context overlap led participants to ignore the context changes occurring as they presented a source of confusability, focusing instead on the properties of the items presented. In these conditions, both repetition and variation would work to promote better episodic memory allowing for an efficient generalization at test as well.

Overall, these results are preliminary evidence that context change and item variation influence learning with consequences for knowledge generalization. Moreover, it provides initial evidence of an interaction between the effects of context change and item variation during study on learners' ability to subsequently generalize their knowledge. An effect that students seem unware of. This has potential implications for educational settings, where context change (e.g., lecture classroom, laboratory section, library) and the examples given during study (e.g., different examples of the

same concept or repeated examples across contexts) are frequent occurrences that can be combined for the best learning.

Acknowledgments

This research was supported in part by National Science Foundation REESE grant 0910218 and Department of Education IES grant R305A1100060 to RLG, and Graduate Fellowship SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology (FCT) to PFC. Stimuli images courtesy of the Aging, Memory, & Cognitive Control Lab (http://psych.wustl.edu/amcclab/AMCC). We thank Alifya Saify for her help with stimuli creation.

References

Abernethy, E. M. (1940). The effect of changed environmental conditions upon the results of college examinations. *The Journal of Psychology, 10*(2), 293-301.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes, 2*, 35-67.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101*(1), 116.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology, 66*(3), 325-331.

Godden, D., & Baddeley, A. (1980). When does context influence recognition memory? *British Journal of psychology, 71*(1), 99-104.

Goldstone, R. L., Kersten, A., Carvalho, P. F. (2013). Concepts and Categorization. In Healy, A. & Proctor, R. *Handbook of Psychology Vol. 4: Experimental Psychology* (pp. 607-630). New York City, NY: Wiley.

Hahn, U., Bailey, T. M., & Elvin, L. B. (2005). Effects of category diversity on learning, memory, and generalization. *Memory & Cognition, 33*(2), 289-302.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96-101.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. *Distinctiveness and memory*, 27-46.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77(3p1)*, 353.

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language, 58*(3), 646-668.

Saufley, W. H., Otaka, S. R., & Bavaresco, J. L. (1985). Context effects: Classroom tests and context independence. *Memory & Cognition, 13*(6), 522-528.

Schmidt, S. R. (1985). Encoding and retrieval processes in the memory for conceptually distinctive events. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11,* 565-578.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 460.

Smith, S. M. (2013). Effects of environmental context on human memory. In T. Perfect & D. Lindsay (Eds.). *The SAGE Handbook of Applied Memory* (pp. 162-182). London, UK: Sage Publications.

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(6), 1582.

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1*(3), 341-358.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review, 8*(2), 203-220.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342-353.

Smith, S. M., Handy, J. D., Angello, G., & Manzano, I. (2014). Effects of similarity on environmental context cueing. *Memory, 22*(5), 493-508.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(5), 893.

Von Restorff, H. (1933). Über die wirkung von bereichsbildungen im spurenfeld [Analysis of processes in the memory trace: 1. On the effect of group formations on the memory trace]. *Psychologische Forschung, 18*(1), 299-342.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*(5), 750-763.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect. *Memory & Cognition, 40*(5), 703-716.

Performance Pressure and Comparison in Relational Category Learning

John D. Patterson (jpatter4@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13905 USA

Abstract

An important objective in higher-order cognition research is to understand how relational categories are acquired and applied. Much of the research on relational category learning has investigated the role of within-category comparison opportunities in category acquisition and transfer – guided by predictions from structure mapping theory that alignment leads to highlighting and abstraction of shared relational structure (Gentner, 1983). Recent research has yielded a within-category comparison advantage under the supervised observational learning mode (relative to twice as many single-item trials), but not under the supervised classification mode (Patterson & Kurtz, 2015). In the present study we investigate the role that pressure to succeed at the training task – a critical difference between the two learning modes – plays in the apparent ineffectiveness of learning by comparison within the classification mode. In a 2x2 between-subjects design we crossed two levels of performance pressure (elevated and standard) with two presentation formats (single-item and within-category pairs). The main findings are: (1) a significant interaction showing a negative impact of increased performance pressure for single-item learners, but not for comparison learners; and (2) a theoretically predicted, but empirically elusive effect of comparison over single-item in the classification mode. We conclude that: (1) performance pressure exerts a deleterious effect on relational category learning (in accord with findings in the attribute category literature) that opportunities to compare may compensate for; and (2) pressure to perform does not appear to underlie lackluster comparison + classification performance (relative to observational learning). Further, we offer new evidence on the role that within-category comparison plays in relational category learning.

Keywords: relational categories; structural alignment; comparison; classification learning; transfer; performance pressure

Introduction

Categorization and comparison are two mechanisms that play a central role in human learning, comprehension, and knowledge use. The study of categorization has largely focused on attribute-based categories, or categories whose members belong based on a shared set of intrinsic features. Attribute categories have been studied under a diverse set of circumstances ranging from inference of missing features (Markman & Ross, 2003) to category construction (Ahn & Medin, 1992). By far however, the most prevalent paradigm of study has been the traditional artificial classification learning (TACL) paradigm. In its most common form, the paradigm operates as follows: a single stimulus is presented, the participant is asked to classify the item into one of two category options, a response is selected, and corrective feedback is given. The study of attribute categories has led to a considerable body of knowledge and has provided a viable testing ground to evaluate formal models of categorization (e.g., ALCOVE, Kruschke, 1992; DIVA, Kurtz, 2007; SUSTAIN, Love, Medin, & Gureckis, 2004).

However, attribute-based understandings alone cannot adequately characterize the richness of human category knowledge. Beyond our attribute-based understanding, we are sensitive to, and knowledgeable about, the ways in which objects and attributes in the world relate to one another. Accordingly, the categorization literature has placed increasing emphasis on the study of relational categories (Gentner & Kurtz, 2005; Markman & Stilwell, 2001). Members belong to relational categories, not based on their attributes, but instead based on shared relational structure. For example, consider the relational noun *barrier*. A thunderstorm, for instance, might be a *barrier* to a baseball game. Just the same, a sick child could be a *barrier* to you going to work. Considering the attributes of these two instances of *barrier* reveals that they are greatly disparate; they would be poor candidates for shared membership based on their attributes. However, their overlapping relational structure – that they both occupy an obstructing role between two other things – grants them membership in a shared relational category.

A key question in the study of relational categories is how we come to learn them. A strong candidate mechanism is comparison. A substantial body of evidence from the relational reasoning literature has found that comparison promotes learning and transfer of shared relational structure (Alfieri, Nokes-Malach, & Schunn, 2013, Gick & Holyoak, 1983; Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 1999; see Loewenstein, 2010 for a review). The benefits of comparison can be explained through the process of structural alignment (Markman & Gentner, 1993). Comparison encourages the alignment of instances' relational predicates, serving to preferentially highlight shared relational structure that is not readily salient when considering either instance alone. Importantly, comparison fosters abstraction of shared relational structure, thereby promoting later transfer. Thus, it is predicted that comparison of instances belonging to the same relational category should produce greater highlighting and learning of shared relations than sequential presentations. Indeed, the study of relational categories has shown same-category comparison to produce these effects in children (Gentner & Namy, 1999; Son, Smith, & Goldstone, 2011) and in adults (Kurtz, Boukrina, & Gentner, 2013; Patterson & Kurtz, 2015).

Though the support for comparison's central role in relational learning is extensive, research in relational category learning has exposed possible boundary conditions of its potency. Borrowing from the tradition of the attribute category literature, the study of relational categories has frequently used the TACL paradigm. However, when a straightforward pairing of pure within-category comparison with TACL has been used, it has failed to produce the predicted comparison advantage. An attempt by Kurtz and Gentner (1998) found that a straightforward merging of same-category comparison with classification required twice as many stimulus exposures to produce an effect of comparison over sequential presentations. Moreover, we conducted a series of follow-up attempts that failed to find an effect of comparison when exposure was equated – despite efforts to enhance the invitation to compare by eliciting similarity ratings or correspondences between elements.

Recent research, however, has shown that learning mode can influence the impact of comparison. Patterson and Kurtz (2015) found robust comparison advantages relative to equivalent single-item presentation on within- and across-domain tests of knowledge when learning under a supervised observational mode (passive study of labeled examples). However, no effect was found in the supervised classification mode – suggesting that some facet(s) of the classification mode attenuates the power of comparison. A key difference between the two modes is that task performance – tied to the guess-and-correct cycle of TACL – is emphasized in the classification mode. When combining comparison and classification, this may elicit competition for resources or a strategic element that favors one component or the other. By receiving accuracy feedback on each trial, it is possible that participants' theory of task shifts relative importance to the guess-and-correct cycle over comparison. We can refer to this as a *weighting effect.*

Alternatively, the apparent incompatibility between classification and comparison may be explained by a *distraction effect* (Beilock & Carr, 2005; Markman, Maddox, & Worthy, 2006). According to the *distraction hypothesis*, peoples' available cognitive resources decline as pressure to perform increases – a 'choking effect' that is accredited to the occupation of executive attention with task-irrelevant thoughts and performance-related worries. Given the greater performance demands of the classification mode, it is possible that the pressure to perform results in the occupation of cognitive resources needed for comparison. Indeed, working memory resources have been marked as important to the structural alignment process (Waltz, Lau, Grewal, & Holyoak, 2000). Waltz et al. (2000) found that when participants were under load that occupied either phonological or executive working memory, noticing of relational correspondences was significantly attenuated and noticing of simpler-to-detect attribute correspondences increased.

In the present study, we sought to determine whether the ineffectuality of comparison in the classification mode can be traced to *weighting* or *distraction* effects by varying the degree of performance pressure in the classification task. The global factor, termed 'performance pressure', was manipulated by testing an elevated pressure condition operationalized by: (1) an instruction asserting that achieving the highest possible performance was the goal, (2) trial feedback in red/green colors (incorrect/correct) to underscore evaluation, and (3) presentation of a running accuracy percentage at the end of each trial. The standard pressure group did not receive (1) or (3) and received feedback in black text. The manipulation was intended to further accentuate the guess-and-correct cycle and increase perceived pressure to perform. It would also be helpful to make the classification task *less* performance-centric, but reducing pressure would seem to require changing the essential nature of the task or employing a relatively weak manipulation. If comparison in the classification mode is ineffective due to either a *weighting* or *distraction* effect elevated performance pressure should lead to poorer learning outcomes than standard pressure. To be clear, both possible effects would be predicted to result in the same performance outcome – differentiating between the two is not of immediate concern.

In addition to exploring the comparison + classification incompatibility, the present study addresses the influence of performance pressure on relational category learning more generally by looking at its effect on single-item learning. Though addressed in the attribute category literature, this issue has not been studied in the relational category realm. Markman, Maddox, and Worthy (2006) showed that participants under very high pressure to perform (i.e., contingent monetary incentives) were hindered, by a posited *distraction effect*, in their ability to learn sequentially presented rule-based attribute categories, but were facilitated in learning information integration categories. Relational categories are generally considered to be verbalizable and rule-like in nature (Gentner & Kurtz, 2005). To the extent that the rule-like nature of attribute and relational categories is similar, a *distraction* effect marked by reduced learning outcomes should be expected for single-item relational category learning under elevated pressure.

In a 2x2 design, the effect of performance pressure (standard, elevated) on relational category learning was assessed across two presentation formats (single-item, within-category comparison). Additionally, we measured participants' regulatory focus – their degree of approach (promotion) and avoid (prevention) motivation using Higgins, Friedman, Harlow, Idson, Ayduk, and Taylor's (2001) regulatory focus questionnaire (RFQ). The match between a person's regulatory focus and their perceived nature of the reward structure has been shown to be important to cognitive tasks, including category learning (see Maddox, Baldwin, & Markman, 2006). To control for how individual differences in participants' regulatory focus might interact with the pressure manipulation, we incorporated their RFQ data into our statistical models.

Method

Participants

93 undergraduates from Binghamton University participated toward partial fulfillment of a course requirement.

Materials

The RFQ contained 11 questions, six measuring participants' level of promotion focus and five measuring their level of prevention focus. A five-point Likert scale was used to collect participants' RFQ responses. The training and within-domain test phase stimuli consisted of 36 unique, Stonehenge-like rock arrangements (see Figure 1). Rocks varied in their size, shape, color, and spatial location. The stimuli comprised three relational categories: *monotonicity* – defined by a monotonic decrease in height of the arrangement from left to right, *support* – characterized by the presence of one rock being supported by two other rocks, and *mirrored stack* – consisting of two same color rocks of similar size and shape arranged in a stack. Each arrangement belonged to only one of the three categories. A 24 item subset of 36 stimuli was used for training and 12 were reserved for use at test; subsets were balanced by category. The subsets matched those used in Kurtz, Boukrina, and Gentner (2013) and the subset was held constant across participants. For comparison conditions, training stimuli were presented in same-category pairs. Pairs were randomly generated for each participant on each pass through the training set.

To evaluate participants' ability to transfer category knowledge to a novel domain, 15 mobile-like stimuli were used (see Figure 1). Each mobile conformed to one of the three categories from training, five per category. The mobiles differed considerably from the training/test stimuli in their surface characteristics (color/shape of objects) as well as in the spatial orientation of the category-defining core which was reflected over the X-axis.

Design and Procedure

In a between-subjects design, participants were randomly assigned to one of four training conditions using the supervised classification task. Two conditions presented pairs of items for same-category comparison: elevated performance pressure comparison (n = 23) and standard performance pressure comparison (n = 23). The other two conditions trained with twice as many single-item trials: elevated performance pressure single-item (n = 24) and standard performance pressure single-item (n = 23). Prior to the training instructions, all participants completed the RFQ (Higgins et al., 2001). Following the RFQ, all participants received an archeological cover story and instructions indicating that their goal was to figure out what makes a rock arrangement belong to a particular type and that they would be tested on their knowledge later. Participants in comparison conditions received an additional instruction indicating that looking at the two arrangements together can be helpful for learning. A further instruction was given to the elevated performance pressure groups, indicating (1) that a performance tracker would show them the percentage of trials they had gotten correct after each trial, and (2) that their job was to achieve the highest percentage they could.

Training – Comparison Conditions Training consisted of two passes through 12 paired stimulus trials, totaling 48 stimulus exposures. Each trial showed two stimuli, side-by-side, that remained visible until the trial was complete. Participants were asked to make a joint category decision for the presented items and selected their response with the mouse. The joint decision represents a deviation from Patterson and Kurtz (2015), however preliminary work showed performance under joint and independent responses to be nearly identical. Time was unconstrained for the category decision. Following the response, participants were given feedback for two seconds indicating: (1) whether they were correct and (2) the correct category of the items. In the elevated performance pressure condition, participants received feedback in either green (correct) or red (incorrect) and at the end of the trial participants saw a screen for two seconds showing the percentage of trials correct to that point before advancing to the next trial.

Training – Single-item Conditions Training consisted of two passes through 24 randomized, single-item trials, totaling 48 stimulus exposures. Outside of the number of trials and language adjusted to accommodate single, instead of paired, items in the decision query, the single-item conditions were conducted identically to the comparison conditions.

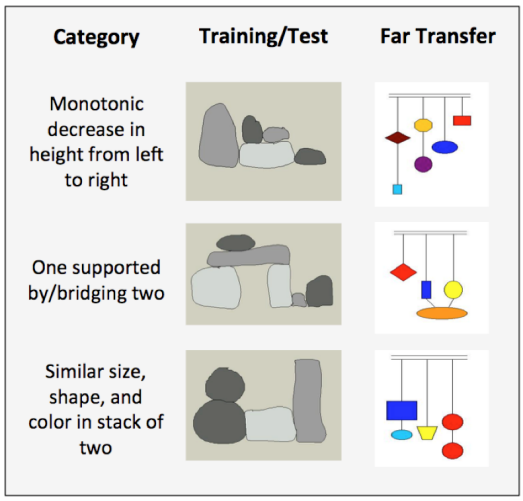


Figure 1: Sample stimuli for each category in each phase.

Assessment After the training, all conditions performed an identical assessment sequence. Participants first received a within-domain test consisting of the 24 'old' rock arrangements from training and 12 previously unseen

arrangements. Old and new items were interspersed randomly. The transfer assessment, consisting of 15 randomly ordered mobile stimuli, was presented after the within-domain test. Both the within-domain and transfer assessments were conducted using a category endorsement task: on each trial an item was presented along with a query asking whether the item belonged to a given category and participants responded with either "yes" or "no." Given our primary interest in how well acquired category knowledge could be extended to new cases, old test items were presented once each while the new within-domain and far transfer items were each presented twice – once with an accurate category label and once with an incorrect label.

Results

The trial-wise accuracy data were modeled using binomial generalized linear mixed effect regressions via the lme4 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2015) packages for the R environment (R Core Team, 2015). Models examining main effects included trial number and the effect of interest as fixed effects. Models examining interactions included trial number, presentation condition, performance pressure condition, and the presentation condition by performance pressure interaction as fixed effects. All models controlled for participants' regulatory focus in the random effects structure by including random intercepts for the 16 levels of prevention focus and 15 levels of promotion focus that resulted from the sample's RFQ data. In the interest of brevity, results of the RFQ will not be discussed. Adjusted means and standard errors for accuracy data can be seen in Table 1.

Training

Modeling of the performance data revealed three effects in the training phase. First, trial number was found to be significantly predictive of accuracy (β = 0.04, SE = 0.003, Wald Z = 11.86, p < .0001), indicating that participants' accuracy increased as they progressed through training. Second, a main effect of presentation condition was observed (β = 0.53, SE = 0.14, Wald Z = 3.84, p = .0001), showing that comparison learners were more accurate than their single-item counterparts. Third, the analysis revealed a marginal presentation condition by performance pressure interaction (β = 0.34, SE = 0.18, Wald Z = 1.90, p = .057). The interaction was driven by an effect of performance pressure for the single-item condition (standard pressure > elevated pressure; β = -0.22, SE = 0.11, Wald Z = -2.06, p = .04), but not for the comparison condition (β = -0.11, SE = 0.14, Wald Z = -0.77, p = .44). In a follow-up test to the interaction, both comparison conditions outperformed their single-item counterparts (elevated pressure: β = 0.87, SE = 0.14, Wald Z = 6.20, p < .0001; standard pressure: β = 0.53, SE = 0.14, Wald Z = 3.84, p < .001).

Old Within-domain Items

Modeling of the old-item accuracy data revealed only a significant effect of trial number (β = -0.01, SE = 0.01, Wald Z = -2.76, p = .006), indicating that participants' accuracy was slightly lower for items encountered later in the queue compared to those encountered earlier. This effect may be associated with old items being interspersed with new ones. There was also a marginal main effect of performance pressure (β = -0.22, SE = 0.13, Wald Z = -1.80, p = .07), marking higher accuracy for the standard pressure group compared to the elevated pressure group. In addition, there was a marginal interaction mirroring that in the training phase (β = 0.41, SE = 0.24, Wald Z = 1.69, p = .09) – showing a significant difference between levels of performance pressure for the single-item condition (standard pressure > elevated pressure; β = -0.43, SE = 0.17, Wald Z = -2.44, p = .01), but not for the comparison condition (β = 0.02, SE = 0.17, Wald Z = 0.10, p = .92).

Novel Within-domain Items

Analyses of the new, within-domain items yielded one reliable effect: a presentation condition by performance focus condition interaction (β = 0.73, SE = 0.22, Wald Z = 3.40, p = .001). Again, the interaction was characterized by a reliable difference between levels of performance pressure for the single-item presentation group (standard pressure > elevated pressure; β = -0.52, SE = 0.15, Wald Z = -3.41, p = .001), but not for the comparison group (β = 0.21, SE = 0.15, Wald Z = 1.34, p = .18). Examining the interaction further, a comparison advantage was found for the elevated pressure group (β = 0.56, SE = 0.18, Wald Z = 3.11, p = .002), but not the standard pressure group (β = -0.17, SE = 0.18, Wald Z = -0.98, p = .33). Since the comparison conditions did not differ, it suggests that the interaction and comparison advantage were driven by poor performance among elevated pressure single-item learners, rather than exceptional performance among elevated pressure comparison learners.

Table 1: Adjusted condition means and standard errors across all performance phases.

	Single-item		Comparison	
	Standard Pressure	Elevated Pressure	Standard Pressure	Elevated Pressure
Training	.59(.04)	.54(.04)	.71(.04)	.73(.04)
Old test	.86(.02)	.80(.03)	.81(.03)	.81(.03)
New test	.78(.03)	.68(.03)	.75(.03)	.79(.03)
Transfer	.64(.04)	.57(.05)	.79(.03)	.82(.03)

Transfer

The transfer results are visualized in Figure 2. Three reliable effects emerged in the analysis of the transfer data. First, an effect of trial number (β = 0.02, SE = 0.005, Wald Z = 4.73, p < .0001) demonstrated that participants' accuracy increased slightly over the course of the transfer assessment. The effect suggests that domain experience helped

participants to recognize the deep similarity between the two domains and apply their category knowledge.

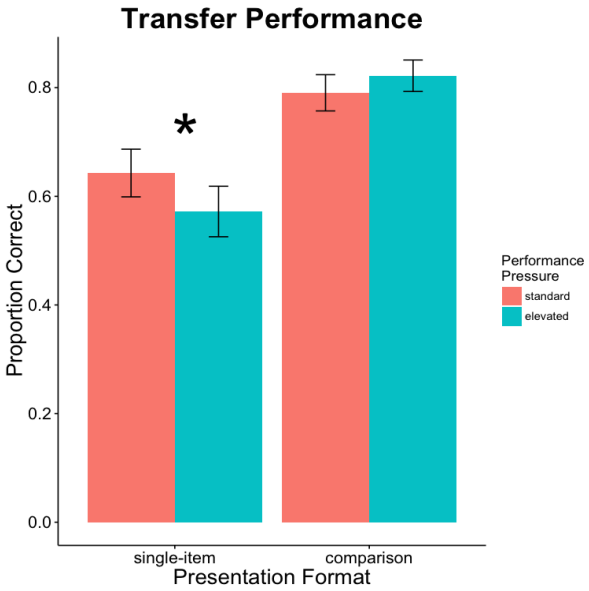


Figure 2: Transfer accuracy by condition. Error bars show +/- 1 SE.

Second, a highly reliable main effect of presentation condition was observed (β = 0.99, *SE* = 0.16, Wald *Z* = 6.08, $p < .00001$), indicating that participants in the comparison group were better at extending their knowledge to a novel domain than those in the single-item group. Third, a significant presentation condition by performance pressure interaction was observed (β = 0.50, *SE* = 0.19, Wald *Z* = 2.63, $p = .01$). The interaction was characterized by pressure-related decrements for the single-item group (standard pressure > elevated pressure; β = -0.29, *SE* = 0.14, Wald *Z* = -2.19, $p = .03$), but not the comparison group (β = 0.20, *SE* = 0.14, Wald *Z* = 1.46, $p = .14$). Critically, follow-up tests examining the effect of comparison revealed that both standard (β = 0.74, *SE* = 0.19, Wald *Z* = 3.94, $p = .0001$) and elevated pressure (β = 1.24, *SE* = 0.19, Wald *Z* = 6.56, $p < .0001$) comparison led to highly reliable advantages over their single-item counterparts.

Discussion

The goal of this study was to evaluate the effect that performance pressure exerts on relational category learning. The results convincingly show that performance pressure has a deleterious effect on relational category learning – leading to learning decrements that were realized across training, within-domain test (both old and new items), and at transfer. However, the pressure-related deficits were restricted to learners who trained via sequential presentations. The effect of performance pressure on single-item learning shows a consistency between the domains of attribute and relational category learning. The similarity between pressure effects for relational and attribute category learning (Markman, Maddox, & Worthy, 2006) suggests that similar systems and/or resources are employed for learning the both category types – perhaps mutually depending on an explicit, verbal system like that posited by the COVIS model (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Future research should further explore the conditions under which the behavior of attribute and relational categories converges and diverges.

A specific interest embedded in our broader goal was to investigate if *weighting* or *distraction* effects might underlie the ineffectuality of comparison in the classification mode as seen by Patterson and Kurtz (2015). If cognitive resources that are necessary for comparison are retrained on the guess-and-correct cycle of classification *or* if they become occupied with worry under pressure to perform, then our performance pressure manipulation ought to have led to poorer comparisons and poorer learning. However, accentuating performance and increasing pressure was found to have no effect on the learning outcomes of comparison learners; this finding casts tentative doubt on the *weighting* and *distraction* accounts of lackluster comparison performance in the classification mode (relative to observational).

One possible rationale for why pressure did not affect comparison learning is that the manipulation was not sufficiently strong to elicit an effect. Other studies (e.g., Markman, Maddox, & Worthy, 2006) have employed stronger manipulations, such as contingent monetary incentives, to achieve high levels of pressure. It seems likely that, at greater levels of pressure, a *distraction* effect would occur, and learning outcomes under comparison would suffer. However, it is clear that our manipulation was sufficiently strong to elicit consistent performance pressure deficits for the single-item group. This suggests that comparison served a compensatory role to the otherwise negative effects of pressure. By this account, elevated pressure learners may have faced a *weighting* or *distraction* effect, but the resources allocated to the alignment process were sufficient to avoid performance decrements. To further elaborate on this effect, future studies should evaluate how differing levels of performance pressure affect alignment-based relational category learning.

Of considerable note, this study also showed a reliable effect of same-category comparison in the supervised classification mode (under standard pressure) at training and transfer over its single-item counterpart. To our knowledge, this is unprecedented. Finding a comparison effect at training and transfer was puzzling, since comparison advantages were not found in Patterson and Kurtz (2015). One notable difference is that the two studies utilized different analysis techniques. Follow-up analyses revealed that the effects from the present study remained after dropping the random effects structure; this prompted us to reanalyze Patterson and Kurtz (2015) using a logistic regression framework as in the present study. Though comparison in the observational mode was the clear winner, showing advantages over single-item at every phase, the

results showed a reliable comparison effect in the classification mode over single-item at training and a marginal effect (p = .09) at transfer. This finding provides further evidence that comparison is an effective means to promote relational category learning. In addition, it echoes the need for researchers to utilize analysis techniques that adequately reflect the richness of the data. For category learning research, analyzing the data trial-wise using logistic regression increases the number of observations available for model fitting, thereby increasing sensitivity over analysis methods that aggregate across trials.

References

Ahn, W., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81-121.

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: a meta-analytic review. *Educational Psychologist*, *48*(2), 87-113.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.

Bates, D., Maechler, M., Bolker, B., Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail working memory and "choking under pressure" in math. *Psychological Science*, *16*(2), 101-105.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*. Washington, DC: American Psychological Association.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487-513.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1-38.

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*, 3-23.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*, 560–576.

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303–1310.

Kurtz, K. J., & Gentner, D. (1998). Category learning and comparison in the evolution of similarity structure. *In Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests for random and fixed effects for linear mixed effect models.

Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. *Psychology of Learning and Motivation*, *53*, 149-182.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*(4), 586-597.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309-332.

Maddox, W. T., Baldwin, G. C., & Markman, A. B. (2006). A test of the regulatory fit hypothesis in perceptual classification learning. *Memory & Cognition*, *34*(7), 1377-1397.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431-467.

Markman, A. B., Maddox, W. T., & Worthy, D. A. (2006). Choking and excelling under pressure. *Psychological Science, 17*(11), 944-948.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*(4), 592-613.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*, 329– 358.

Patterson, J.D., & Kurtz, K.J. (2015). Learning mode and comparison in relational category learning. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2011). Connecting instances to promote children's relational reasoning. *Journal of Experimental Child Psychology*, *108*(2), 260-277.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*(7), 1205-1212.

Intermediate Judgments Inhibit Belief Updating: Zeno's Paradox in Decision Making

James M. Yearsley (james.m.yearsley@vanderbilt.edu)
Department of Psychology, Vanderbilt University, Nashville TN, USA

Emmanuel M. Pothos (emmanuel.pothos.1@city.ac.uk)
Department of Psychology, City University London, London, EC1V 0HB UK

Abstract

Rational agents should update their beliefs in the light of new evidence. Equally, changes in belief should depend only on the quality of the evidence, and not on factors such as the order in which the evidence is acquired, or whether intermediate judgements are requested during evidence acquisition. In contrast we show that requests for intermediate judgments can inhibit belief updating for real decision makers, which represents a new type of decision making fallacy. This behaviour is paradoxical from the point of view of classical Bayesian models, but we show that it is consistent with an a priori, parameter free prediction of a cognitive model based on quantum theory.

Keywords: cognition; decision making; quantum probability.

Introduction

That decision makers should update their beliefs in the light of new evidence is one of the cornerstones of what it means to be rational (Anand, 1993). Equally, for rational agents, the degree of belief change should depend only on the quality of the new evidence acquired and not on other factors such as the order of evidence acquisition or requests for intermediate judgements (Oaksford & Chater, 2009). However it is known that real decision makers do not always conform to the strictest standards of rationality (Tversky and Khaneman, 1974). One notable example of this are order effects, where presenting the same evidence in different orders can have a large influence on the change in belief generated (Trueblood & Busemeyer, 2011).

In this contribution we examine another effect, which is that of requests for intermediate judgments during the process of evidence acquisition. Since requesting such a judgment does not provide a decision maker with any extra relevant information about the problem, making intermediate judgments should have no effect on the process of belief updating for a rational decision maker. Thus at least the simplest classical Bayesian cognitive models should predict no effect of intermediate judgments on the final belief state.

In contrast, in cognitive models where judgment is considered a constructive process, intermediate judgments may well have an effect on belief updating (Schwarz, 2007; White, Pothos & Busemeyer, 2014). One such class of models, which have received much attention recently, are those based on the mathematics of quantum theory (Busemeyer & Bruza, 2011). These models not only predict an effect of intermediate judgments on belief updating, but allow us to make a priori, parameter free predictions of the precise relation between the change in belief and the number of intermediate judgments which can be tested by a suitable experiment. The startling conclusion is that in quantum models, requests for intermediate judgments can strongly inhibit belief updating. In physical systems this singular phenomena is called the quantum Zeno effect, because of the (loose) analogy with Zeno's second paradox; for this reason we term this effect Zeno's paradox in decision making.

In one experiment, we compare the predictions of the quantum model with a simple matched classical Bayesian model and find that the quantum model outperforms the Bayesian one by a large margin. Indeed, the predictions of the Bayesian model are shown to be qualitatively inconsistent with the data. We briefly discuss whether allowing probabilities in a classical model to depend on the memory of a previous judgment can reproduce the quantum behavior.

Experimental Investigation

We performed two identical experiments as a replication exercise. We report only the first experiment here, the conclusions of the second were the same and are reported in full in Yearsley & Pothos (2016).

We recruited 450 experimentally naïve participants, from Amazon Turk. Participants were 49% male and 50% female (1% did not respond to the gender question). Most participants' first language was English (98%) and the average age was 34.8. The experiment lasted approximately 10 minutes; participants were paid $0.50 for their time.

The experiment was implemented in Qualtrics. Our paradigm extends one of Tetlock (1983), which was designed to test for primacy effects in decision making. After some initial screens regarding ethics information and consent, all participants saw the same initial story, regarding Smith, a hypothetical suspect in a murder:

"Mr. Smith has been charged with murder. The victim is Mr. Dixon. Smith and Dixon had shared an apartment for nine months up until the time of Dixon's death. Dixon was found dead in his bed, and there was a bottle of liquor and a half filled glass on his bedside table. The autopsy revealed that Dixon died from an overdose of sleeping pills. The autopsy also revealed that Dixon had taken the pills sometime between midnight and 2am. The prosecution claims that Smith slipped the pills into the glass Dixon was drinking from, while the defense claim that Dixon deliberately took an overdose."

Participants were then given a short set of questions regarding some details of what they had just read, in order to check that they were engaging with the task. These questions

were intended to reinforce memory of the story details and to check for participants who were not concentrating on the experiment. The small number of participants who failed to correctly answer these questions were excluded from subsequent analysis. Participants were subsequently asked whether they thought Smith was likely to be guilty or innocent, based on the information provided in the vignette, and to provide a brief justification for their response, as a further check that they were adequately concentrating on the task and to reinforce memory for the response. The first response is critical, since all model predictions are based on knowledge of the initial (mental) state. Most participants (95%) initially assumed innocence, and so we excluded participants who initially assumed guilt. Participants then saw a screen reminding them of their response.

Participants were split into six groups. The first group was presented with 12 pieces of evidence suggesting that Smith was guilty (participants were told they would only see evidence presented by the prosecution and not by the defense). Each piece of evidence was designed (and pilot tested) to be individually weak, but cumulatively the effect was quite strong. For example one piece of evidence read "Smith had a previous conviction for assault." After reading all 12 pieces of evidence, participants were again asked whether they thought Smith was guilty or innocent, and again asked to justify their choice. Participants in the other five groups were shown the same evidence in the same way, and asked to make the same final judgment, but were also asked to make intermediate judgments (and justify their responses). These intermediate judgments were worded in the same way as the initial and final ones, and were requested at intervals of either 1, 2, 3, 4 or 6 pieces of evidence. A small number of participants gave justifications for their judgments that suggested they were not properly engaging with the task, and were therefore excluded from the analysis. Together with those who failed to correctly answer the questions about vignette, and those excluded from the present analysis because they initially judged 'Guilty,' this left 425 good participants.

The order of presentation of the evidence was partly randomized. The pieces of evidence were split into four blocks of three pieces of evidence each. The order of the blocks was fixed, but the order of the pieces of evidence within each block was randomized. The reason we randomized evidence order in this way, rather than say simply randomizing the order of presentation of all pieces of evidence, is that there are a total of 12!, or about 480 million, possible orderings of the evidence, so it is impossible to capture a representative sample of the orderings by simple randomization.

After the main part of the experiment, participants were shown a list of the pieces of evidence they had encountered, and were asked to rate the strength of each piece of evidence on a (1-9) scale. (The full list of pieces of evidence is presented is Yearsley & Pothos, (2016.))

Zeno's paradox in quantum decision making

We want to explain why the Zeno effect arises in quantum models of decision making using a simple example. In a realistic decision making scenario, such as the experiment we discussed above, the modeling is necessarily more complex and this can make the origin of the effect harder to see.

Consider a 2D quantum system, with a state space spanned by two orthogonal states I and G, corresponding to the beliefs that Smith is either Innocent or Guilty. Presentation of evidence is represented by a rotation of the state such that an initial state I evolves towards G, with pieces of evidence.

We are interested in the probability that a measurement of the state will reveal I, at each of $N \geq 1$ judgments at $T/N, 2T/N ... T$ is (where T is the total number of pieces of evidence) In analogy with the physics case, this may be called the *survival* probability after N judgments. For a typical time independent Hamiltonian we have;

$$Prob('Survival,'N) = Prob\Big(I \text{ at } \frac{T}{N}, I \text{ at } \frac{2T}{N} ... I \text{ at } T | I \text{ at } 0\Big) = \cos^{2N}\Big(\frac{\gamma}{N}\Big) \quad (1)$$

Here γ is a constant encoding the effect of the evidence in the absence of intermediate judgments. As the number of judgments, N, increases, there is a decreasing probability that the system will change from I to G. As $N \to \infty$, the probability that the system will change state vanishes, even after large number of pieces of evidence. This is the famous QZ effect (Misra & Suarshan, 1977), often described informally as proof that 'a watched pot never boils'.)

The derivation leading to Eq.(1) involves a number of assumptions that will not hold in realistic decision making settings. However we can still predict a weakened QZ effect, as a slowing down (in a specific way) of the evolution of the measured opinion state, even under more realistic conditions. We will outline the argument below, full details are given in Yearsley & Pothos (2016).

Two assumptions need to be relaxed. First, realistic measurements are not perfectly reliable. For each measurement, there is a small probability that a participant will incorrectly provide a response not matching his/her cognitive state. This is problematic when several identical measurements are made, since error rates may compound. Imperfect measurements require the use of positive-operator valued measures (POVMs), instead of projection operators. Instead of freezing as $N \to \infty$, some evolution may still occur, but it will depend only on details of the imperfect measurements (Anastopoulos & Savvidou, 2006). The effect of imperfect judgments is encoded by a simple POVM operator with one free parameter, ε. The parameter $0 \leq \varepsilon \leq 1$ reflects how error-less measurements are. For example, if a participant considers Smith innocent, then the probability of responding innocent is only $1-\varepsilon$, leaving a probability to respond guilty of ε.

Second, evolution of cognitive variables is better modeled by a time dependent unitary evolution to capture the fact that the weight given by participants to a piece of evidence de-

pends on its position in the sequence of evidence, implying a primacy or recency effect. We must also take account of the fact that not all pieces of evidence will be regarded as equally strong by participants.

A form for the time dependent unitary evolution general enough for our purposes is

$$U(t_m,t_n) = \exp(-i\sigma_x B(t_m,t_n)), \tag{2}$$

where σ_x is one of the Pauli matrices and

$$B(t_m,t_n) = \alpha \sum_{i=m+1}^{n} a_i e^{-\beta(i-m-1)^2} \tag{3}$$

with α and β real numbers.

The function $B(t_m,t_n)$ specifies the angle a participant's cognitive state is rotated through when presented with pieces of evidence t_m through t_n. Here the a_i represent the strengths of the individual pieces of evidence, as measured in isolation. Thus the first piece of evidence in a sequence is given a weight a_1 the second is given weight $a_2 e^{-\beta}$, and so on. Depending on the value of β this form for $B(t_m,t_n)$ can encode a primacy ($\beta > 0$) or a recency ($\beta < 0$) effect. Using the above, we can show that:

$$\begin{aligned} Prob(I \text{ at } t | I \text{ at } 0) =& (1-\varepsilon)^2 \cos^2(B(0,t)) \\ &+ \varepsilon(1-\varepsilon)\sin^2(B(0,t)) \end{aligned} \tag{4}$$

Eq(4) allows us to determine ε and B(0,t), from empirical classical data on the probability of judging Smith's innocence, assuming innocence initially, and varying the number of pieces of evidence presented (without intermediate judgments). We can also use Eq(4), together with some assumptions about the way judgments change the cognitive state classically, to construct a Bayesian model of the same decision making process. We will do this below, but note that in the case of no intermediate judgments the QT and Bayesian models will coincide. This means that we can use data obtained in the absence of any intermediate judgments to fix all the parameters in both the QT and Bayesian models. Our central predictions, of the specific way in which intermediate judgments affect opinion change, will therefore be parameter free.

The Quantum Model

We now state the prediction of a QZ effect in this decision making setting, the full derivation is presented in Yearsley & Pothos (2016). The result is that a participant deciding Smith's innocence will be less likely to change his/her initial opinion as the number of intermediate judgments increases. Specifically, the expression for the survival probability is:

$$\begin{aligned} Prob^{Q}('survival', N) &= Prob\Big(I \text{ at } \frac{T}{N}, I \text{ at } \frac{2T}{N}, ... | I \text{ at } T\Big) \\ &= (1-\varepsilon)^{N+1} \sum_{i=0}^{N-1} \cos^2\left(B\left(\frac{iT}{N}, \frac{(i+1)T}{N}\right)\right) \\ &+ \varepsilon(1-\varepsilon)^N \sin^2\left(B\left(\frac{(N-1)T}{N}, T\right)\right) \\ &\times \sum_{i=0}^{N-2} \cos^2\left(B\left(\frac{iT}{N}, \frac{(i+1)T}{N}\right)\right) + O(\varepsilon^2) \end{aligned} \tag{5}$$

The first term in this expression corresponds to the probability that the cognitive state is always consistent with innocent, and all the judgments reflect this. The second term corresponds to possibility that the state does change between the second to last and final judgments, but the participant nevertheless responds 'innocent' due to the imperfect measurements. Further terms would correspond to more judgments not matching the cognitive state, or the state changing back from innocent to guilty, these terms are negligible compared to those included in Eq.(5). If $\varepsilon = 0$, $\beta = 0$ and the a_i's are equal then Eq(5) reduces to Eq(1).

Constructing a matched classical model

The QT model assumes evidence changes the opinion state (determined by Eq(4)), judgments may be imperfect, and judgments are constructive. The third property is the characteristically quantum one, so with the first two elements, we constructed an alternative, Bayesian model for survival probability. It is helpful to denote by I_B the event where a participant believes Smith is innocent, and by I_R the event where a participant responds Smith is innocent, similarly for guilty.

The expression we are interested in is the Bayesian analogue of Eq.(5); the survival probability after T pieces of evidence have been presented, given that N judgments have been made. This is

$$\begin{aligned} &Prob^{C}('survival', N) = \\ &\quad Prob\Big(I_R \text{ at } T, I_R \text{ at } \frac{(N-1)T}{N}, I_R \text{ at } \frac{T}{N} | I_R \text{ at } 0\Big) \end{aligned} \tag{6}$$

We want to construct this so that it matches the quantum expression in the case of no intermediate judgments (N=1). We will sketch how to do this here, full details are given in Yearsley & Pothos (2016).

As already noted, because Eq(4) does not involve any intermediate judgments it may be interpreted classically. We can therefore read off,

$$\begin{aligned} Prob(I_B \text{ at } t | I_B \text{ at } 0) &= \cos^2(B(t,0)) \\ Prob(G_B \text{ at } t | I_B \text{ at } 0) &= \sin^2(B(t,0)) \\ Prob(I_R \text{ at } t | I_B \text{ at } t) &= (1-\varepsilon) \\ Prob(G_R \text{ at } t | I_B \text{ at } t) &= \varepsilon \\ Prob(G_R \text{ at } t | G_B \text{ at } t) &= (1-\varepsilon) \\ Prob(I_R \text{ at } t | G_B \text{ at } t) &= \varepsilon \end{aligned} \tag{7}$$

The probabilities involving transitions from Guilty cognitive states to Innocent ones are assumed to be 0. Eq.(4) is therefore also our Bayesian survival probability for the case of no intermediate judgments.

To compute the survival probability when there are intermediate judgments made we need to know the appropriate function for the evolution of the state. The natural classical analogue of Eq.(3), $B^C(t_m,t_n)$, is given by

$$B^C(t_m,t_n) = \sum_{i=m+1}^{n} a_i e^{-\beta(i-1)^2} \tag{8}$$

This differs from $B(t_m,t_n)$ only in the fact that the function multiplying the evidence strength depends only on how many pieces of evidence have been presented before it, and not on whether any intermediate judgments have been made. Note that $B^C(0,t_m) = B(0,t_m)$ since the quantum and classical models agree in the absence of intermediate judgments. In particular this means fitting either function to the data in the absence of intermediate judgments produces the same set of parameters, α,β.

In fact we could use the function $B(t_m,t_n)$ in the Bayesian analysis if we desire, despite the fact it is poorly motivated. It turns out that the Bayesian models perform better when using $B^C(t_m,t_n)$, so we will work exclusively with this.

We can use the information above to derive a prediction for the Bayesian survival probability. Doing so involves two assumptions, first that ε is small, and secondly that the probabilities involving transitions from Guilty cognitive states to Innocent ones are negligible. We can then show (Yearsley & Pothos, 2016;

$$\begin{aligned} Prob^C('survival', N) &= (1-\varepsilon)^{N+1}\cos^2(B^C(0,T)) \\ &+ \varepsilon(1-\varepsilon)^N \sin^2\left(B^C\left(\frac{(N-1)T}{N},T\right)\right) \\ &\times \cos^2\left(B^C\left(0,\frac{(N-1)T}{N}\right)\right) + O(\varepsilon^2) \end{aligned} \tag{9}$$

We are now ready to test the Bayesian and QT predictions in a realistic decision making scenario.

Results and model fits

Empirical assessment involved two steps. First, without intermediate judgments (ie at the first judgment made after having seen some evidence) the data is classical and simply informs us how opinion changes with evidence. Using Eq(4), we can determine ε and $B(t_1,t_2)$ i.e., the parameter specifying the POVMs for Smith's innocence, guilt and the function specifying the way evidence alters the opinion state (the same parameter values are used in the Bayesian and QT models). Second, we examined whether intermediate judgments produce the QZ effect (slowing down of opinion change, as predicted by the QT model, Eq(5)) or not (in which case the Bayesian model should fit better). The predictions about intermediate judgments from the models were assessed after parameter fixing, the first step; they are *a priori and parameter free*.

In order to determine $B(t_1,t_2)$, we first need to know the a_i's for each piece of evidence. These are the parameters indicating the relative strength of each piece of evidence and they were fixed directly, using the participant ratings for each piece of evidence at the end of the task.

The best fit parameters were obtained by minimizing the sum of the squared deviations between the predictions of Eq(4) and the data. Considering the $t = 3$ data point an outlier, the best fit for Eq(4) is obtained with $\alpha = 0.091, \beta = 0.010$ and $\varepsilon = 0.030$, giving an R^2 of .996 and a BIC of -27.8. (BICs computed following Jarosz & Wiley (2014).) The results of the fitting are shown in Fig(1).

Since the data can be thought of as arising from a binomial distribution the most informative way to display a confidence interval is to imagine using the data to update an initially uniform prior for the value of the survival probabilities, and plot the 95% Highest Density Interval (HDI) of the resultant posterior. Error bars in all figures therefore refer to these HDIs.

For small t, $Prob(I \text{ at } t | I \text{ at } 0)$ is non-linear and (extrapolated) not equal to 1 at $t = 0$. This result justifies our assumption of imperfect measurements. In Fig(1), for large t, $Prob(I \text{ at } t | I \text{ at } 0)$ is close to linear with increasing t. Linearity implies that belief change is proportional to the number of pieces of evidence, which seems an obvious expectation for a rational participant (while the belief state is far from guilty). Note that the best fit value of β is positive, confirming our expectation of diminishing returns (equivalently, there is a primacy effect, regarding evidence strength.)

Now that the model parameters have been fixed for both the QT and Bayesian models, we can use Eq(3) and Eq(4) to compute survival probabilities, for different numbers of intermediate judgments.

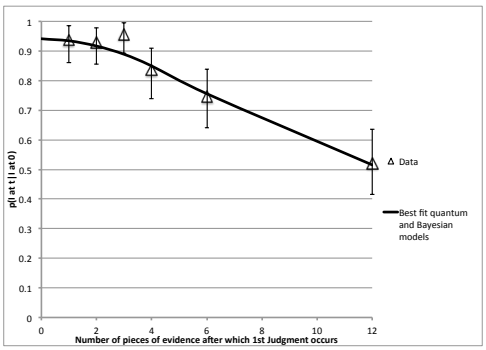


Figure 1: Setting the parameters (opinion change without intermediate judgments): $Prob(I \text{ at } t | I \text{ at } 0)$, for the first judgment a participant made, after having seen different numbers of pieces of evidence. Note the obvious outlier at three pieces of evidence. Data points are the probability computed as an average over participant choices (Number of Participants =64, 71, 70, 73, 71, 76 for each data point) and error bars show the 95% HDI of the posterior.

Empirical results for $Prob('survival', N)$ show a QZ effect, as survival probability increases with small N (Fig(2)). The

classical intuition is reduction of survival probability with more intermediate judgments, because of a probability of error at each judgment. The data clearly favor the QT model: the Bayes factor is 3.4×10^5 (Bayes Factor computed following Jarosz & Wiley (2014).) The dip in the survival probability for large N is an effect of the imperfect judgments.

There is an alternative test of the QT vs Bayesian models. We can employ Eq(5) and Eq(9) to compute survival probabilities for the condition where there is a judgment after every piece of evidence (number of pieces of evidence presented T, and number of judgments N, vary, but T/N fixed to 1). Again, the data clearly favor the QT model (Fig(3)). The Bayes Factor in this case is 8.2×10^9.

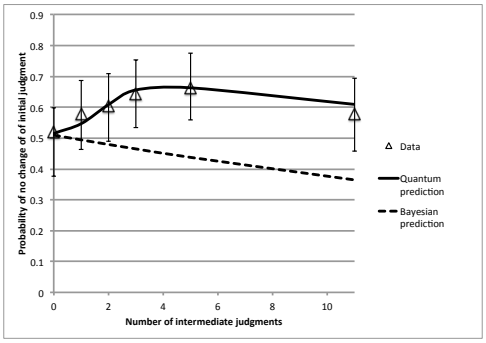


Figure 2: Evaluating the models: Survival probability for N intermediate judgments, for the QT, Bayesian models, against empirical results ; Data points are the probability computed as an average over participant choices (Number of Participants =76, 71, 73, 70, 71, 64) and error bars show the 95% HDI of the posterior.

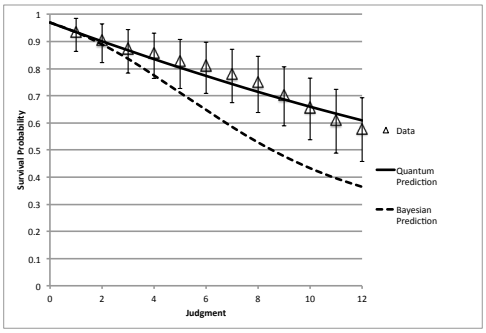


Figure 3: Evaluating the models: Survival probability after each judgment, for the condition with 12 judgments. Data points are the probability computed as an average over participant choices (Number of participants =64 for all data points) and error bars show the 95% HDI of the posterior.

In summary, the data show a clear QZ effect, with survival probability generally increasing with the number of intermediate judgements. Furthermore the behavior is in excellent quantitative agreement with the predictions of the QT model.

Classical Models with Memory Effects

We can use the behavior of the QT model to learn about the structure of any classical model which aims to reproduce these results. Such a model could be constructed by introducing extra variables not directly measured, but that have an effect on the judgment probabilities. This would be post hoc from the point of view of our experimental set up, but it might be possible to motivate the existence and behavior of these extra variables from other considerations. One natural possibility is the memory of any previous judgments taken. This might be motivated by arguing decision makers wish to appear consistent in their judgments, so there is a bias towards aligning a judgment with any previous one[1]. Note that although such a model may be constructed in a classical way, it is still manifestly non-Bayesian, since belief updating depends on factors other than the evidence seen.

We can use the predictions of the QT model to explore how such a model would behave. It is useful to look at the idealized case, Eq(1) and focus on the case of a single intermediate judgment. With T pieces of evidence and no intermediate judgment the survival probability is,

$$Prob(I \text{ at } T | I \text{ at } 0) = \cos^2(T). \tag{10}$$

and this must hold for the classical and quantum models. With a single intermediate judgment the QT probability is,

$$\begin{aligned} Prob^Q\Big(I \text{ at } T, I \text{ at } \frac{T}{2} | I \text{ at } 0\Big) &= \cos^4\Big(\frac{T}{2}\Big) \\ &= Prob^Q\Big(I \text{ at } T | I \text{ at } \frac{T}{2}\Big) Prob^Q\Big(I \text{ at } \frac{T}{2} | I \text{ at } 0\Big) \\ &= \Big[Prob^Q\Big(I \text{ at } \frac{T}{2} | I \text{ at } 0\Big)\Big]^2. \end{aligned} \tag{11}$$

In order that any classical model agree with the predictions of the QT one, Eq.(11) would have to hold for this model. Suppose we have a classical model containing an extra variable that plays the role of the 'memory' of the judgment made at $T/2$. Then this memory variable would have to 'screen off' the prior evidence so that transition probabilities between $T/2$ and T are identical to those between 0 and $T/2$ before any evidence has been seen. In other words, to agree with the QT model predictions, a classical model incorporating memory cannot simply have the memory function an additional piece of evidence. Instead the memory must override any evidence seen prior to the judgment.

There is a second property less often discussed but also necessary for the QZ effect. Proof that Eq.(1) tends to 1 as the number of judgments increases is based on the following,

$$\begin{aligned} \lim_{N\to\infty} Prob\Big(I \text{ at } \frac{T}{N}, I \text{ at } \frac{2T}{N} \ldots I \text{ at } T\Big) &= \lim_{N\to\infty} \cos^{2N}\Big(\frac{\gamma}{N}\Big) \\ &= \lim_{N\to\infty} (1 - \frac{1}{2}(\gamma/N)^2 + O(N^{-4}))^{2N} \\ &= \lim_{N\to\infty} \big[1 - \gamma^2/N + O(N^{-3})\big] = 1 \end{aligned} \tag{12}$$

[1]This possibility was first suggested to us by Gordon Logan.

Crucially this depends on the transition probabilities being non-linear for small amounts of evidence. In other words, the QZ effect holds in part because in the QT model beliefs are 'sticky.' That is, is takes more evidence to change a belief state that is close to a definite state than it does to change a belief that is highly uncertain. We allowed our Bayesian model to mimic this property in the model comparison, but this 'stickiness' of beliefs is in itself a fundamentally non-Bayesian property that must be incorporated into any model aiming to reproduce the QZ effect.

To summarise this section, there is no doubt that one could construct a classical model of belief updating that would mimic the predictions of the QT model (and so fit at least this data and the second set reported in Yearsley & Pothos 2016) by considering the effects of remembering previous judgments. However in addition to any concerns about such a model being post hoc, it would also have to have the two features outlined above; that memories of intermediate judgments screen off evidence, and that beliefs are sticky. It would be interesting to try and construct a classical model wth these properties, although one wonders to what extent this would simply amount to redescrbiing the QT model.

Concluding remarks

Understanding the way beliefs change as a result of accumulating evidence is essential for modeling decision making. Our results suggest that opinion change depends not just on the evidence, but can also be strongly effected by making intermediate judgments, a phenomena we call Zeno's paradox in decision making. This behavior is at odds with Bayesian models of cognition, but agrees quantitatively with the predictions of a model based on QT. Because the QT model was fixed with classical data, this striking prediction follows from a structural feature of quantum theory, the collapse postulate, and not from parameter fixing. Our results show models of decision making need to incorporate influences arising from the process of making judgments.

They also have practical implications. The paradigm we employed, albeit simplified, does have analogies with realistic assessment of evidence; if e.g. juries are expected to reach unbiased conclusions, then the effect of requests for intermediate opinions should be factored in. Likewise, the advent of interactive news web sites (e.g., bbc.co.uk) means that readers can express opinions on news items while reading them, directly and through social media. We raise the possibility that this sort of overlap between acquiring information and expressing an opinion may prevent change in opinion, even in the presence of compelling evidence.

More generally, behaviors paradoxical from Bayesian perspectives have often been interpreted as boundaries in the applicability of probabilistic modeling. Strictly speaking this is not true, since one can always augment Bayesian models with extra variables or interactions, however such models may lack predictive power, or simply be too post hoc. The QT cognition program provides an alternative: perhaps some of these paradoxical findings reveal situations where cognition is better understood using QT, without the need to introduce extra unobserved degrees of freedom. Evidence for the collapse postulate in decision making constitutes a general test of the applicability of QT principles in cognition and adds to the growing body of such demonstrations (Pothos & Busemeyer, 2013). An important future direction will be to understand whether the adaptive arguments used to motivate Bayesian models of cognition (Oaksford & Chater, 2009) need to be adapted to account for phenomena such as the Zeno effect.

Acknowledgments

EMP and JMY were supported by Leverhulme Trust grant RPG-2013-00. Further, JMY was supported by an NSF grant SES-1556415 and EMP was supported by Air Force Office of Scientific Research (AFOSR), Air Force Material Command, USAF, grants FA 8655-13-1-3044. The U.S Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright notation thereon.

References

Anand, P. (1993). Foundations of Rational Choice under Risk. Oxford: OUP.

Anastopoulos, C. and Savvidou, N. (2006). Time-of-arrival probabilities and quantum measurements. J. Math. Phys. 47, 122106.

Busemeyer, J. R. & Bruza, P. (2011). Quantum models of cognition and decision. CUP: Cambridge, UK.

Jarosz, AF and Wiley, J (2014). What are the odds? A practical guide to computing and reporting Bayes factors. Journal of Problem Solving, 7, 2-9.

Misra, B. & Sudarshan, E.C.G. (1977). The Zeno's paradox in quantum theory. Journal of Mathematical Physics, 18, 756-763.

Oaksford, M. & Chater, N. (2009). Prcis of Bayesian rationality: the probabilistic approach to human reasoning. Behavioral and Brain Sciences, 32, 69-120.

Pothos, E.M. & Busemeyer, J.R. (2013). Can quantum probability provide a new direction for cognitive modeling? Behavioral & Brain Sciences, 36, 255-327.

Schwarz, N. (2007). Attitude construction: Evaluation in context. Social Cognition, 25, 638-656.

Tetlock, P.E. (1983). Accountability and the Perseverance of First Impressions. Soc. Psych. Quarterly, 46, 285-292.

Trueblood, J. S. & Busemeyer, J. R. (2011). A comparison of the belief-adjustment model and the quantum inference model as explanations of order effects in human inference. Cognitive Science, 35, 1518-1552.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

White, L. C., Pothos, E. M., & Busemeyer, J. R. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. Cognition 133, 48.

Yearsley JM & Pothos EM (2016). Zeno's paradox in decision making. Proc. R. Soc. B, **283**, 20160291.

Working memory encoding of events and their participants: a neural network model with applications in sensorimotor processing and sentence generation

Martin Takac (takac@cs.otago.ac.nz)[1,2], **Alistair Knott (alik@cs.otago.ac.nz)**[1]
Dept of Computer Science, University of Otago, New Zealand[1], Centre for Cognitive Science, Comenius University, Bratislava[2]

Abstract

In this paper we present a model of how events and their participants are represented in working memory (WM). The model's central assumption is that events are experienced through sequentially structured sensorimotor (SM) routines—as are the individuals that participate in them. In the light of this assumption, we propose that events and individuals are stored in WM as *prepared SM routines*. This proposal allows a new mechanism for binding representations of individuals to semantic roles such as AGENT and PATIENT. It also enables a novel account of how expectations about forthcoming events can influence SM processing in real time as events are perceived. Finally, it supports an account of the interface between WM representations and language.

Keywords: event perception; working memory; embodied cognition; neural networks; syntactic heads

Introduction: semantic working memory

In this paper we present a model of how the brain encodes events and their participants in working memory (WM). The WM medium that stores events was dubbed the 'episodic buffer' by Baddeley (2000). We adopt the slightly broader term **semantic WM**, because our model represents individuals and their properties as well as events. (The term also emphasises the non-phonological character of the WM medium we are modelling.) For Baddeley, the episodic buffer is an interface medium, linking to three distinct neural systems: the **sensorimotor** (**SM**) system, through which events are directly experienced, the **episodic memory** system, in which they are stored in long-term memory (LTM), and the **language** system, through which they are communicated. Here we consider the semantic WM system as it interfaces with the SM system and language; in a companion paper (Takac and Knott, this volume), we consider its interface with LTM.

Experimental work on the WM system has focussed on relatively simple representations: representations of spatial location, visual properties, and prepared motor actions. However, semantic WM is also assumed to be the medium where such representations are combined or **bound** into complex semantic structures. A binding mechanism is crucial for representing events: in particular, representations of individuals must be bound to semantic roles such as AGENT and PATIENT. Note that representations of individuals have their own internal structure, which must be created through some form of binding. An individual has properties (shape, type etc), but also a spatial location. And it can be a singular entity, or a group. It is important that the mechanism binding individuals to semantic roles can operate on compositionally structured representations of individuals, as well as on atoms. Experiments have not revealed much about this binding mechanism, but it is a key topic for neural network research (e.g. van der Velde and de Kamps, 2006; Stewart and Eliasmith, 2012).

In the current paper, we will introduce a new network model of semantic WM, which makes some new proposals about the binding operations that create event representations. We show that the event representations in this model are well suited to support accounts of the role of semantic WM in online SM experience and in sentence processing.

Background: a model of event perception

WM representations of experienced events have to be *created during experience*. Events take time to occur, so the SM processes through which they are experienced must be similarly extended in time. The founding assumption in our model is that event-perception processes have a well-defined temporal structure—and that the mechanism representing events in WM capitalises on this well-defined structure. In this section we outline what this structure is; for details, see Knott (2012).

We argue that perceiving an episode involves a relatively discrete *sequence* of SM operations. This assumption rests on some well-accepted findings about perceptual processes. Firstly, there is good evidence that focal attention must be allocated to an individual in order to process it in any detail (see Walles *et al.*, 2014 for a summary). If an event involves several participant individuals, therefore, the observer must attend to them one by one, rather than in parallel. Secondly, when an event is perceived, participants playing certain semantic roles are recognised first. For transitive events, we argue the AGENT participant must be attended to before the PATIENT (Knott, 2012).[1] If the observer is executing the action, this is because the decision to act must precede selection of a target; if the agent is watching an action, it is because s/he must monitor the agent to identify the intended target (Webb *et al.*, 2010). Thirdly, a representation of the motor action cannot be evoked until the target object has been attended to. In action execution, the agent must activate a representation of the target object before its motor affordances can be computed (Johansson *et al.*, 2001); in action perception, the observer must compute the trajectory of the agent's hand onto the target (e.g. Oztop *et al.*, 2004). If these assumptions, which are individually quite well accepted, are brought together, an interesting model of event perception emerges, in which apprehending a transitive event involves a sequence of three SM operations: attention to the agent, then attention to the target, then activation of a motor programme. The idea that events have a characteristic temporal structure is certainly present in other models of event perception, in particular that of Reynolds *et al.* (2007). For Reynolds *et al.*, these sequen-

[1] Our terms 'agent' and 'patient' refer to Dowty's (1991) more general concepts 'proto-agent' and 'proto-patient'.

tial regularities relate primarily to the structure of an agent's *movements*: they are the kind of regularities that the 'biological motion' system becomes attuned to. In our model, such regularities are encoded within the action representation system, as discrete actions. But there is more to an event than an action. In our model, experiencing an event also involves a higher-level sequence, of relatively discrete *SM operations*. One of these is the activation of an action representation. But this operation must be *preceded* by an action of attention to the agent, and then an action of attention to the patient (if there is one). In our model, the notions of agent and patient are in fact *defined* by the serial order of attentional operrations in this SM sequence: the (proto-)agent is the first individual attended to; the (proto-)patient is the second.

Alongside this model of event perception, we also assume that the perception of each participant in an event involves its own canonically-structured sequence of SM operations. It is well established that in order to classify an object, an observer must first direct focal attention to the region of space it occupies. But observers can also attend to a region of space containing a homogeneous group of objects. Walles *et al.* (2014) argue that in between focal attention and object classification there is an intervening attentional operation that selects a spatial *scale* at which the classifier will be deployed, determining whether the classifier identifies the local or global form (Navon, 1977) of the attended stimulus. This operation determines whether a single individual is classified or a homogeneous group of individuals. In summary, perception of an individual involves a SM routine comprising three operations: selection of a salient region of space, then selection of a classification scale (determining whether a singular or plural stimulus will be classified), and finally activation of an object category. Event perception, in turn, is a higher-level sequential SM routine, some of whose elements have their own sequential structure.

WM representations as prepared sequences

We propose that representations in semantic WM exploit the sequential structure of perceptual processes. Specifically, we propose that WM representations of both individuals and episodes take the form of *prepared sequences*. This proposal is attractive for several reasons. For one thing, it offers a clear account of how semantic WM representations can influence SM processing: a prepared SM sequence is an 'executable' structure, that can initiate sequentially structured SM activity (including actions). For another thing, it suggests an account of a puzzling recent finding: stimuli held in WM appear to be transiently *reactivated* in SM areas during the delay period (see e.g. Meyers *et al.*, 2008). If WM representations are prepared SM routines, that can be executed in simulation, then active simulation processes could occur during the delay period, resulting in these transient patterns of SM activity. Finally, the proposal places semantic WM representations within a class of neural representation that is relatively well understood. We know a lot about how prepared sequences of attentional or motor operations are represented, particularly in macaques. The relevant representations are predominantly in prefrontal cortex (PFC), which is also a key site for semantic WM. A particularly interesting result is from Averbeck *et al.* (2002). They showed that the PFC assembly that stores a prepared sequence of SM operations contains sub-assemblies representing each individual operation—and moreover, that these sub-assemblies are active *in parallel* in the structure representing a planned sequence, even though they represent operations that are active one at a time. Our model will make use of this finding.

A model of semantic role-binding using sequentially-structured WM representations

Modelling semantic WM representations as prepared sequences suggests a novel account of how semantic roles are bound to participants in representations of events. Our account makes use of three ideas, which we introduce below.

The key idea is that the binding mechanism is implemented as part of the active process of rehearsing SM routines, rather than within a static representational structure. The classic binding problem arises because the SM media representing an individual's properties (location, shape etc) naturally represent just *one* individual: if the properties of several individuals are represented, it is hard to specify which properties belong to which individual. If a WM event representation supports the simulation of a sequential SM routine in which representations of agent and patient are active in these media *at different times*, many of these problems go away.

Of course, the event representation must still make reference to both participants, so it can activate these temporally separated representations. The second idea in our binding scheme is that event representations represent participants using *pointers* into the medium representing individuals—and that there are separate pointers for agent and patient. The pointers are active simultaneously in a WM event representation, but they are only *followed sequentially*, when an event is rehearsed. (Event representations thus conform to the properties of prepared sequences identified by Averbeck *et al.*) In neural networks terms, agent and patient are coded 'by place' in our WM event representations, in separate groups of units. Place coding of this kind is not normally seen as a viable way of implementing role-binding: a simple place-coding scheme suffers from the fact that there is nothing in common between representations of John-as-agent and John-as-patient. But if the place-coded representations of agent and patient just hold pointers into the medium representing individuals, which are activated at different times, this problem does not arise.

The third idea in our binding scheme is that the place-coded pointers in WM event representations do not point directly to SM media representing individuals, but rather to a *WM* medium holding representations of individuals. Recall that representations of individuals also have internal structure: we proposed above that the WM representation of an individual is also stored as a prepared, replayable SM routine.

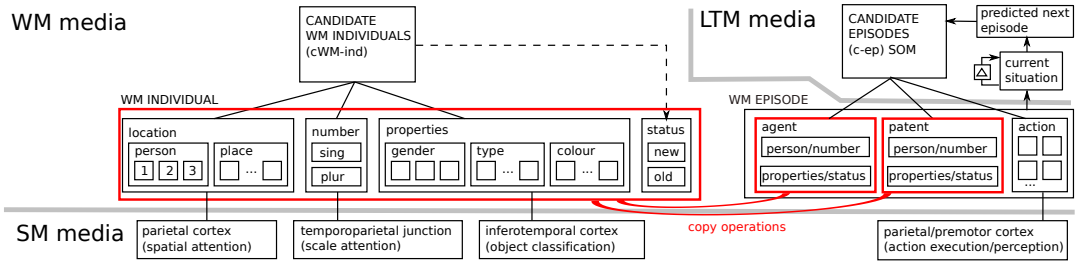


Figure 1: Architecture of the model of semantic WM

In our model, WM representations of recently-perceived individuals are held in a separate WM medium: the agent and patient representations in a WM event point to, and sequentially re-activate, representations within this WM medium. During rehearsal of a WM episode, these sequentially reactivated representations create opportunities for secondary rehearsal operations, simulating the steps involved in perceiving the participant individuals. This scheme introduces a measure of hierarchy in the model of role-binding, enabling the representations filling semantic roles to have a degree of internal structure—an important requirement, as noted earlier.

A model of WM episodes and WM individuals

Our model is illustrated in Figure 1. SM media are below the grey line; WM media are above it. The WM system representing individuals is on the left, and that representing events (or 'episodes', as we call them here) is on the right. The copy operations implementing pointers are highlighted in red.

The WM medium on the left holds a representation of a single selected individual, a **WM individual**, stored as a prepared sequence of a **location**, a **number** (i.e. classification scale) and a set of perceptual **properties**. These three representations are activated in parallel in the WM medium, but when the prepared sequence is executed or rehearsed, they activate associated first-order representations in the attentional and classification systems one at a time, as discussed above.

The media representing a WM individual provide input to another layer, the **candidate WM individuals** (**cWM-ind**) layer, which stores *combinations* of location, number and type over a short time period, and thus represents the set of recently-attended individuals. A partially specified WM individual can function as a query to the cWM-ind layer: if we specify a location, we may be able to retrieve an associated type and number (and vice versa). If we can, then the individual retrieved is classed as 'old'; if we cannot, it is classed as 'new'. These attributes are recorded in a **status** field of the WM individual, which is not part of the prepared sequence. Queries formed from partially-specified WM individuals can be used to generate expectations about the location or properties of individuals in the current scene, as we discuss below.

The **WM episodes** system is structurally similar to the WM individuals system. It holds a representation of a single selected episode, a 'WM episode', stored as a planned sequence of operations activating an **agent**, a **patient** and an **action**. As noted above, the agent and patient media hold content-addressed pointers to representations in the WM individual medium. All the media within a WM episode are active in parallel, but when a WM episode is executed or rehearsed, the representations they point to become active sequentially: the 'agent' and 'patient' media activate two successive representations in the WM individual medium, and then the 'action' medium activates a representation in the (pre)motor system.

The prepared operations in a WM episode also provide input to a layer holding episode representations learned over a longer timespan, the **candidate episodes** (**c-ep**) layer. This layer is a self-organising map (**SOM**): when exposed to training episodes, it learns to represent episodes as localist units, organised so that similar episodes are close together in the map. Each unit can encode a particular combination of representations in the agent, patient and action media, and thus can represent a complete episode by itself. Note this localist scheme is enabled by our model of binding: the 'agent' and 'patient' fields of a WM episode index their fillers *by place*, so carry information about both roles and fillers. Clearly, we cannot represent every possible episode using localist units. But that is not the purpose of the c-ep SOM: its role is rather to represent the episodes that occur frequently, so these can provide a top-down bias on SM processing during experience. Since the c-ep SOM uses localist representations, it can also represent *multiple* expected episodes simultaneously: a useful property, as we will show.

A final component of the network is a layer representing the **current situation**. In our model, this is the hidden layer of a recurrent network that learns to predict the *next episode*, given the episode that has just occurred, plus a copy of its hidden layer at the previous time point. The current situation network learns to predict a *distribution* of possible next episodes in the c-ep SOM (exploiting its ability to represent multiple episodes). (It is somewhat analogous to Reynolds *et al.*'s (2007) recurrent network for event representation, but Reyonlds *et al.*'s network predicts the next *component* of an episode, rather than the next discrete episode.)

One useful feature of our model is that that the c-ep SOM can learn *generalisations* over episodes. One kind of generalisation is hard coded in the model: the copies of WM individuals created in the agent and patient fields ignore location information, so representations of episodes in the SOM abstract over the location of participants. In our model, expectations

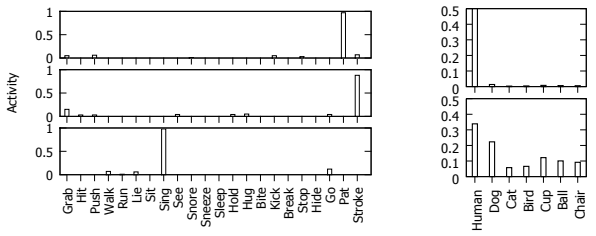
about the locations of objects are dealt with in the WM individuals system, as we will illustrate below. This step considerably reduces the combinatorial possibilities that need to be represented in the SOM. But the SOM also learns generalisations of its own. The ability to generalise is a standard feature of learning in SOMs, since episodes that are sufficiently similar will activate the same localist unit. In particular, since the representations of agents and patients providing input to the SOM are distributed, the SOM can learn to abstract away from the properties of token individuals and represent the participants of episodes as types, as we will show.

Here are some technical details about the network's architecture. The WM individual layer consists of localist sets of feature units for person (1/2/3), number (Sg/Pl), gender (Male/Female/Neuter) and status (new/old). Each set of units can either encode a single property unambiguously, or a probability distribution over properties. The type area also contains feature sets coding animacy, and object type (person/dog/cat/bird/cup/ball/chair). Object location, situated on a 100×100 grid, is coded by a population of 6×6 neurons with Gaussian receptive fields evenly covering the grid. Colour is coded by a population of 11 neurons with Gaussian receptive fields in 3D RGB space maximally responding to standard 11 basic colours (see Figure 3a). Such population coding is neurally plausible and there is a straightforward mathematical way of computing the likelihoods of different stimuli given the activities of neurons in the population (Jazayeri and Movshon, 2006). SM representations (below the grey line in Figure 1) are isomorphic to the WM areas they interface with. Likewise, agent and patient layers of a WM episode are isomorphic to the relevant parts of a WM individual. The 'action' area consists of 22 localist units for actions (see the *x*-axis legend in Figure 2a) and 11 units for their distributed featural codes. The cWM-ind layer is a variable-sized convergence zone of units fully connected with the WM individual layer. When a novel candidate individual is encountered, a new unit in the cWM-ind layer is recruited and the current values of WM individual units are copied into its connection weights (one-shot learning). The c-ep layer is a SOM with 400 units. Each unit also maintains a scalar weight reflecting the frequency of 'hits' for this unit, i.e. the number of times it was the most active unit. These frequency weights serve as priors for computing the Bayesian probability that the current input corresponds to an episode represented by a particular unit (for details see Takac and Knott, 2016). The network that represents the current situation is a recurrent SOM (a 'Merge SOM', Strickert and Hammer, 2005). This provides input to a layer of linear perceptrons which are trained to predict the next episode. Details of all these networks can be found in Takac and Knott (2016).

Training and testing of the network

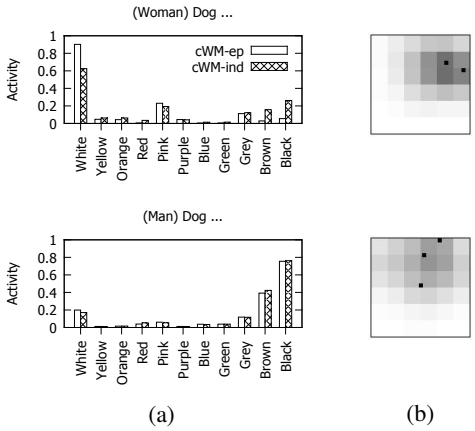
Training We trained the network by simulating SM experience of a sequence of episodes. Each episode is represented in the SM system as a complex sequence of SM operations. We begin by describing the properties of the individuals that featured in episodes. We created a fixed population of token individuals: each with a type, a number, and location and colour properties that are stochastically chosen based on its type. Locations are quasi-randomly generated as positions on a 100×100 grid (which in the system's 6×6 location medium are represented using coarse coding). Colours are stochastically generated from Gaussian distributions centred on 11 basic colours. We then generated a stream of episodes involving these token individuals. Each episode is presented to the WM system as a sequence of SM input items. Episodes are of three types: **transitive** (agent→patient→trans-action), **intransitive** (agent→intrans-action) and **causative** (agent→patient→cause-signal→unaccusative-action). In each case the agent and patient signals has a sequential structure of their own, namely location→number→type/properties. Each of these latter sequences is sent to the WM individuals medium, activating the different components of a WM individual representation one by one. When complete, the WM individual was first passed as a query to the cWM-ind layer, to find out whether the individual it represents has already been encountered. For each candidate unit currently active in the cWM-ind layer, we compute the likelihood that it corresponds to the current stimulus in the WM individual (Jazayeri and Movshon, 2006). This reduces to the average pairwise Kullback-Leibler divergence between the respective areas of the WM individual and the candidate unit weights (see Takac and Knott, 2016 for details). If a likely-enough candidate is returned, it is updated if necessary and the WM individual's status is set to 'old'; otherwise a new entry in the layer is created and the WM individual's status is set to 'new' (candidate units that have not been updated for N episodes are removed). The WM individual is then copied (along with its status) to the appropriate layer in the WM episode medium: either the 'agent' layer or the 'patient' layer. When a complete episode has been presented to the system, the layers in the WM episode are passed as input to the c-ep SOM. This SOM learns in the standard way. Note that while learning in the cWM-ind layer happens in a 'one-shot' manner, it only happens gradually in the c-ep SOM, current situation and next episode media.

Testing the sequence-based binding scheme To demonstrate the new binding scheme, we must show how the WM representations created during experience of an episode allow it to be *replayed*. To test this, after each episode is presented, the WM episode medium is used as input to a replay process, in which the layers in this medium activate the representations they point to one by one. When the activated representations are in the WM individual layer, they are used as a query to retrieve a location representation (recall that the location of individuals is not copied to the WM episode). If the binding scheme is effective, this process should regenerate the same sequence of first-order SM signals that was presented to the network during experience of the episode. In our tests, the sequence was perfectly reconstructed for 99.6% of episodes; this shows that our proposed binding mechanism is effective.



(a) (b)

Figure 2: (a) Action types predicted in the c-ep layer for 3 episode fragments. From top to bottom: man→dog→?, man→cat→?, bird→?. (b) Patient types predicted for Sequence *A* (top) and Sequence *B* (bottom).



(a) (b)

Figure 3: (a) Expectations about the colour of the patient generated by the c-ep and cWM-ind layers for episodes woman→dog (top) and man→dog (bottom). (b) Expectations about location of the patient generated by the cWM-ind layer for these episodes. Darker areas mean stronger expectations. Black dots represent actual locations of currently present white dogs (top) and black dogs (bottom).

Testing the network's prediction/generalisation abilities The network can make several kinds of prediction; we will focus on three progressively more complex predictions. Firstly, the c-ep SOM can make predictions about the episode currently being experienced, as experience is under way. Predictions about actions are easiest to demonstrate, since it represents actions directly. To evaluate these predictions, we introduced some regularities in the episodes presented to the system. Birds always sang (bird→sing); also when people interacted with dogs and cats, they always patted dogs and stroked cats (person→dog→pat, person→cat→stroke). We presented the c-ep SOM with episodes involving these participants, leaving the action field blank, and generated a distribution of expected episodes in the SOM; from this we reconstructed a distribution of expected actions, by linear combination of the weight vectors of SOM units, weighted by unit activity. Figure 2a shows these distributions are correctly weighted towards the actions encountered during training.

The c-ep SOM can also make predictions about the agents and patients of episodes. These are more complex, because its predictions are relayed to the WM individuals system, which refines them based on its own knowledge. To illustrate this process, we introduced a further regularity into the training episodes: in all episodes involving people interacting with dogs, the dogs were black if the agent was a man, and white if the agent was a woman; additionally, people always appeared in the top-left quadrant of the spatial array, and animals in the top-right quadrant. We then generated an underspecified representation in the WM episode: in the agent part, we activated a representation of a person (either man or woman), and in the patient part we activated the type 'dog', unspecified for colour. We used this representation to generate a distribution in the c-ep SOM, from which we reconstructed a predicted distribution of patient features. This whole distribution was copied to the WM individual medium, where it activated a distribution of units in the cWM-ind network. This distribution was used top-down to reconstruct a distribution over expected locations for the patient individual, and to refine the distribution over expected types and properties. Figure 3a shows activity in the colour-coding features of the resulting WM individual expectation. The system correctly predicts a colour centred on black in RGB space for man→dog episodes, and on white for woman→dog episodes. Importantly, the cWM-ind layer can also generate expectations about the location of the dog—see Figure 3b. There is a general bias towards the quadrant containing animals, since dogs always appear in this quadrant. But there are also specific biases towards the location of the black or white dogs that the system has recently encountered, that are *based* on its expectations about the colour of the patient dog.

Finally, the network can make predictions about the *next* episode, using its representation of the current situation. To show this, we presented it with a sequence of training episodes, with constraints on transitions between episodes: when a person hit a dog and then patted the same dog ('sequence *A*'), the dog always bit the person; when a person patted a dog after some other episode ('sequence *B*'), any random episode could follow. We tested the network by presenting sequences *A* and *B*, propagating activity through the Current situation and Next episode prediction layers to obtain a prior distribution over predicted next episodes in the c-ep layer. From this we reconstructed an expected distribution of agents, patients and actions in the WM episode. The network correctly predicted 'dog-bite-man' after Sequence *A*, and made a more neutral prediction after Sequence *B* (see Figure 2b). The learned update rule encodes something like 'If a man hits a dog then pats it, *the dog bites the man*.'

Roles for the network in language processing

As discussed at the outset, semantic WM representations must do service in language processing as well as in SM processing. Our WM model was designed with this in mind.

We envisage several linguistic roles for the network. The cWM-ind medium can function as the medium storing salient discourse referents, and the current situation medium can be understood as holding a representation of the current discourse context; for details on these ideas, see Takac and Knott (2016). We also envisage that the WM model plays a role in sentence processing. Specifically, we propose that generating a sentence reporting an episode stored in WM involves *rehearsing* this episode, in a special cognitive mode in which SM/WM representations can trigger output phonology. We built a neural network model of sentence generation implementing this idea (Takac *et al.*, 2012). Within this model, we argue that the syntactic concept of a **head** can actually be derived from the architecture of the semantic WM system. In syntactic theory, information conveyed by a syntactic head spreads through its local syntactic domain (e.g. a clause or noun phrase (NP)). This spreading process is seen most clearly in agreement phenomena: for instance, subject-verb agreement within a clause, or determiner-noun agreement within a NP). In a NP, agreement rules relate to the head features **person**, **number**, **type**, semantic **gender** and **definiteness**: exactly the information that is maintained tonically in a WM individual during its rehearsal.[2] Heads in a clause can convey all this information about the subject and object, and additionally information about the type of the action: exactly the type of information maintained tonically in a WM episode during its rehearsal. We argue that syntactic heads have an extended syntactic domain because they are read from WM representations that are tonically active during rehearsal, and so can influence phonology at multiple points. Again see Takac and Knott (2016) for further discussion.

Discussion

In this paper we propose that events and their participants are represented in WM as prepared SM sequences. This has several benefits. It permits a direct account of how WM event representations are created during experience, and of how they in turn influence event perception by generating top-down expectations. It enables a new model of role-binding, that allows hierarchical representations of event participants, and localist representations of candidate events, through a novel use of indexing and place-coding. This in turn allows the model to represent large *distributions* of expected events: a very useful ability, which the models of van der Velde and de Kamps (2006) and Stewart and Eliasmith (2012) do not have. Finally, the network supports several aspects of a model of language processing; most interestingly, it creates a framework within which aspects of syntactic structure can seen as deriving from the structure of the semantic WM system.

The model as illustrated here leaves many open questions. One question concerns space requirements. Our place-coding scheme for event participants requires creating several separate *copies* of the WM individual medium, which is expensive as regards storage space. However, storage is within acceptable limits when scaled up to a memory of realistic size, as discussed in Takac and Knott (2016). Another question concerns recursive representations. While the model described here implements a notion of *hierarchical* representations, these representations are not recursive. Again, this issue is addressed in Takac and Knott (2016), where we argue that the current scheme extends well to an account of complement clauses, subordinate clauses, and relative clauses.

[2] 'Person' is conveyed as a special kind of location information: the location can be a point in external space (in which case the point is also specified) but it can also be a direct reference to the speaker or interlocutor, whose location is presumed to be independently given. Type and semantic gender are expressed within the complex of perceptual properties. Definiteness is expressed by status (new or old).

Acknowledgements

This work was supported by the New Zealand Marsden Foundation through Grant 13-UOO-048. Martin Takac was also partially supported by Slovak KEGA grant 017UK-4/2016.

References

Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *PNAS*, **99**(20), 13172–13177.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *TICS*, **4**(11), 417–423.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, **67**(3), 547–619.

Jazayeri, M. and Movshon, A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, **9**(5), 690–696.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, J. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience*, **21**(17), 6917–6932.

Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA.

Meyers, E., Freedman, D., and Kreiman, G. *et al*. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J.Neurophys*, **100**, 1407–19.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, **9**, 353–383.

Oztop, E., Bradley, N., and Arbib, M. (2004). Infant grasp learning: a computational model. *Exp. Brain Research*, **158**, 480–503.

Reynolds, J., Zacks, J., and Braver, T. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, **31**, 613–643.

Stewart, T. and Eliasmith, C. (2012). Compositionality and biologically plausible models. In *The Oxford Handbook of Compositionality*. Oxford University Press, New York.

Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. *Neurocomputing*, **64**, 39–71.

Takac, M. and Knott, A. (2016). A simulationist model of episode representations in working memory. Technical Report OUCS-2016-01, Dept of Computer Science, University of Otago.

Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition*, **125**, 288–308.

van der Velde, F. and de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, **29**, 37–108.

Walles, H., Robins, A., and Knott, A. (2014). A perceptually grounded model of the singular-plural distinction. *Language and Cognition*, **6**, 1–43.

Webb, A., Knott, A., and MacAskill, M. (2010). Eye movements during transitive action observation have sequential structure. *Acta Psychologica*, **133**, 51–56.

What do you expect from an unfamiliar talker?

Dave F. Kleinschmidt[1], and **T. Florian Jaeger**[1,2,3]
{dkleinschmidt, fjeager} @ mail.bcs.rochester.edu
[1]Department of Brain and Cognitive Sciences, [2]Department of Computer Science, and [3]Department of Linguistics,
University of Rochester, Rochester, NY, 14627 USA

Abstract

Speech perception is made much harder by variability between talkers. As a result, listeners need to adapt to each different talker's particular acoustic cue distributions. Thinking of this adaptation as a form of statistical inference, we explore the role that listeners' prior expectations play in adapting to an unfamiliar talker. Specifically, we test the hypothesis that listeners will have a harder time adapting to talkers whose cue distributions fall outside the range of normal variation across talkers. We also show that it is possible to *infer* listeners' shared prior expectations based on patterns of adaptation to different cue distributions. This provides a potentially powerful tool for directly probing listeners' prior expectations about talkers that does not rely on speech produced by many different talkers, which is costly to collect and annotate, and only indirectly related to listeners' subjective expectations.

Keywords: Cognitive Science, Linguistics, Psychology, Language understanding, Learning, Speech recognition, Bayesian modeling, Experimental research with adult humans

Introduction

A longstanding problem in speech perception is how listeners manage to cope with substantial differences in how individual talkers produce speech. Recent evidence suggests that one strategy listeners employ is to *rapidly adapt* to unfamiliar talkers (Bertelson, Vroomen, & de Gelder, 2003; Clarke & Garrett, 2004; Kraljic & Samuel, 2007, among others). Such adaptation can be understood as a form of statistical inference. This insight is captured by a recent proposal, the *ideal adapter* framework (Kleinschmidt & Jaeger, 2015). Each talker's particular accent (way of talking) can be formalized as the distribution of acoustic cues that they produce for each phonetic category (or other underlying linguistic unit). Listeners are taken to adapt to an unfamiliar talker via *distributional learning*, inferring the underlying talker-specific cue distributions from the talker's productions.

Critically, this statistical inference process draws on implicit beliefs about *how* talkers tend to differ from each other. As a consequence, adaptation to an unfamiliar talker should depend on a listener's prior experience with other talkers, rather than only on the speech produced by the unfamiliar talker themselves. Specifically, a listener's experience with other talkers provides the starting point for the distributional learning required for adaptation, or, in Bayesian terms, a *prior belief* about the probability of different possible accents (cue distributions). More informative prior beliefs can substantially reduce the amount of direct evidence needed to converge on accurate beliefs about the current talker's cue distributions.

Code and data to generate this paper is available from github.com/kleinschmidt/cogsci_2016

The goals of the present work are two-fold. First, we test a critical prediction of the ideal adapter framework. To the extent that a listener's prior beliefs are informative, they must take some probability *away* from unlikely accents. Confronted by a talker whose accent falls well outside the range of what they expect based on their previous experience, the ideal adapter framework predicts that a listener will require more evidence to adapt, leading to slowed or incomplete adaptation. There is some evidence that this is the case. For instance, Idemaru and Holt (2011) found that listeners have difficulty adapting to a talker who produces anti-correlated distributions of two cues that are typically positively or uncorrelated. Sumner (2011) found that listeners had trouble adapting to a talker who produced a distribution of cues for the /b/ and /p/ sounds that had substantially lower means than a typical talker.

However, no studies have systematically probed whether and how a listener's prior expectations constrain phonetic adaptation, or even what kind of prior beliefs listeners have. To that end, we expose listeners to a range of different accents, which differ (only) in the cue distributions for /b/ and /p/. By parametrically manipulating these distributions, we create a range of accents that are more or less similar to what a typical talker of English produces. We then assess the degree to which listeners adapt their beliefs about the novel talker's cue distributions, depending on the *a priori* typicality of these distributions.

To anticipate the results, we find that typicality of the novel talker's cue distribution predicts the degree to which listeners adapt to the talker. This suggests that listeners not only have beliefs about the cue distributions for a *particular* single talker (as suggested by previous work; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015; Kronrod, Coppess, & Feldman, 2012), but also have implicit beliefs about the ways in which talkers tend to *differ* from each other, and hence what to expect from an unfamiliar talker. This leads to the second question we address here: what is the content of listeners' prior beliefs about inter-talker variability? To this end, we use a Bayesian belief-updating model to work backwards from listeners' adaptation behavior across talkers and *infer* listeners' shared prior beliefs. This approach provides a more direct assessment of listeners' subjective prior beliefs than production data from many different talkers, which moreover requires costly and time-consuming annotation. If successful, this would provide a powerful method for investigating listeners' prior beliefs that could be applied to other cues, categories, and even social variables (like gender, native language

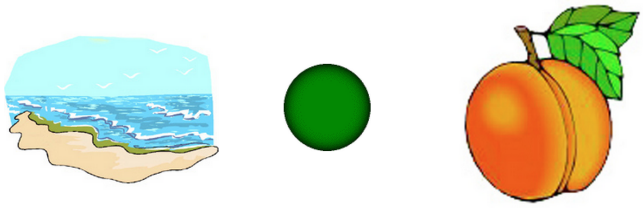


Figure 2: Example trial display (beach/peach). Listeners first click on the green button to play the word, then click on one picture to indicate what they heard.

background, etc.).

Experiment

We tested the role that listeners prior expectations play in adapting to an unfamiliar talker by exposing them to one of five synthetic "accents" (Figure 1). These accents differed only in the distribution of voice onset time (VOT), the primary cue to word-initial stop consonant voicing in English (e.g., "beach" vs. "peach"). Adaptation was assessed based on listeners' classification function, or how they labeled each VOT as /b/ or /p/.

Methods

Subjects We recruited 169 subjects via Amazon's Mechanical Turk, who were paid $2.00 for participation, which took about 20 minutes. We excluded subjects who participated more than once ($n = 4$) or whose accuracy at 0 ms and 70 ms VOT—as extrapolated via a logistic GLM—was less than 80% correct ($n = 28$; $n = 1$ for both reasons). Excluded subjects were roughly equally distributed across conditions (maximum of 5 in 0ms /b/ VOT condition, and minimum of 1 in 20ms /b/ VOT condition). After these exclusions, data from 138 subjects remained for analysis for analysis.

Procedure Our distributional learning procedure is described in Kleinschmidt, Raizada, and Jaeger (2015), and is based on Clayards et al. (2008). On each trial, two response option images appeared, which corresponded to one of three /b/-/p/ minimal pairs (beach/peach, bees/peas, or beak/peak). Subjects then clicked on a central button which played the corresponding minimal pair word, and then clicked on the picture to indicate whether they heard the /b/ or /p/ member of the minimal pair (Figure 2). Subjects performed 222 of these trials.

Each trial's word was synthesized with a voice onset time (VOT) that was randomly drawn from a bimodal distribution, with low and high VOT clusters implicitly corresponding to /b/ and /p/, respectively. This distribution defined the *accent* that the subject heard, and each subject was pseudorandomly assigned to one of five accent conditions (Figure 1). Each of these accents implies a different /b/-/p/ category boundary, and listeners' adaptation was evaluated by comparing their /b/-/p/ classification function (fitted by a logistic GLM) to the classification function derived from the typical talker's cue distributions (corresponding to no adaptation) and the exposure distribution (corresponding to maximal—but not necessarily optimal—adaptation).

/b/ mean VOT	Mean shift	95% CI
-10	29%	13–44%
0	—	—
10	80%	57–104%
20	53%	45–62%
30	49%	40–57%

Table 1: Percentage of boundary shift from typical talker to each exposure talker (see Figure 3), averaged over subjects with 95% bootstrapped confidence intervals. 0% shift corresponds to no adaptation at all, while 100% corresponds to perfect adaptation, ignoring any prior beliefs. Typical and exposure talker boundaries were too close together to reliably determine boundary shift percentage in the 0ms condition.

Results and Discussion

Figure 3 shows the classification functions for each individual listener. In each accent, these classification functions tend to fall in between the boundaries predicted by the typical talker distributions and the boundaries implied by the exposure distributions. We can quantify this by the percentage of the predicted shift in category boundary from the classification function for the typical talker to the boundary implied by the input distribution (Table 1). A 0% shift corresponds to no adaptation at all, while a shift of 100% corresponds to complete adaptation to the exposure distributions, with no (remaining) influence of any prior beliefs.

In all conditions, the average shift percentage was between 0% and 100% (except the 0ms shift condition, which is so close to the typical talker that estimating the percentage is numerically unstable). More interestingly, the more extreme conditions show less complete adaptation than the less extreme conditions. Together, these results suggest that listeners' adaptation was constrained by their prior expectations (given the finite amount of evidence they received about the unfamiliar talker). This provides qualitative evidence that listeners combine their prior expectations with observed cue distributions in order to rapidly adapt to unfamiliar talkers, as predicted by the ideal adapter framework (Kleinschmidt & Jaeger, 2015).

Inferring prior beliefs about talker variability

Our second goal in this paper is to test whether it is possible to infer listeners' prior beliefs about talker variability, based on their patterns of adaptation to different accents. To that end, we use a variant of a Bayesian belief-updating model that has previously provided a good account of how listeners incrementally update their beliefs in order to rapidly adapt to an unfamiliar talker (Kleinschmidt & Jaeger, 2015). This

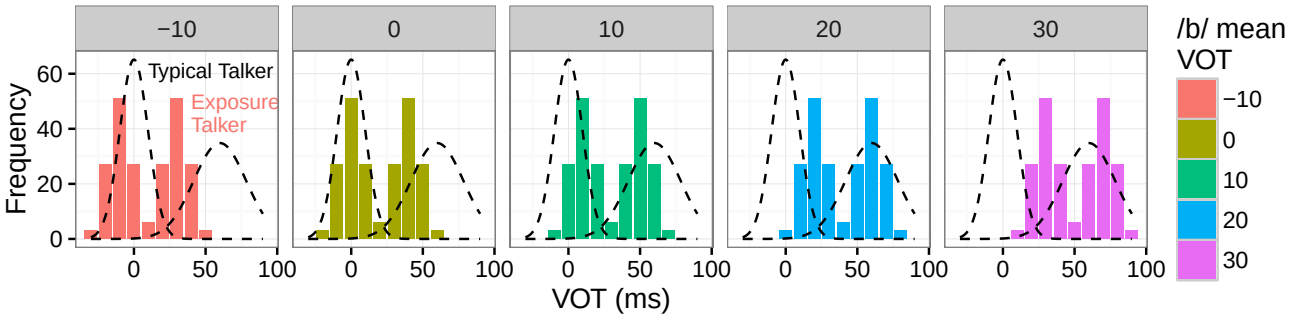


Figure 1: Each subject heard one of these five synthetic accents, which differ only in the distribution of VOTs of the word-initial stops. Black dashed lines show VOT distributions from a hypothetical typical talker (as estimated by Kronrod et al., 2012). Note that the 0 and 10ms shifted accents are reasonably close to this typical talker, while the −10, 20, and 30ms shifted accents deviate substantially.

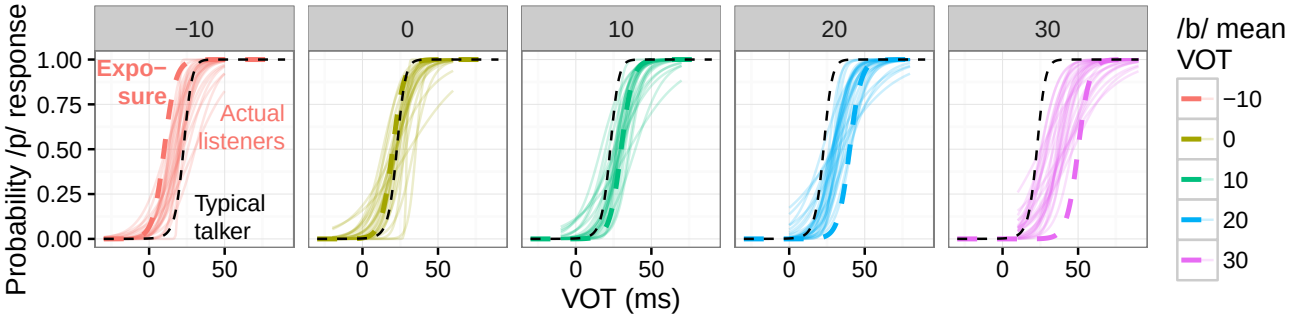


Figure 3: Listeners' responses, smoothed with logistic functions (thin lines), compared with the classification functions expected based on a typical talker (no learning; dashed black lines) and complete adaptation to the exposure distributions (thick dashed colored lines). Listeners' actual category boundaries lie between the typical talker and exposure talker boundaries (see Table 1).

previous modeling work has treated the content of listeners prior beliefs—the category means and variances they think are most likely—as known and fixed, setting them based on pre-adaptation classification data, and then fitting the confidence in those prior beliefs as a free parameter. Here, we wish to fit both the content of (prior expected mean and variance of each category) and the confidence in prior beliefs, based on adaptation data presented above.

Methods

We denote the listener's beliefs about the *current* talker's generative model as the parameters of a two-component mixture of gaussians

$$\theta = \{\mu_b, \sigma^2_b, \mu_p, \sigma^2_p\} \quad (1)$$

where μ is a category's mean VOT and σ^2 is its variance. As in Kleinschmidt and Jaeger (2015), we use an independent, conjugate Normal-χ^{-2} prior for each category, with parameters (Gelman, Carlin, Stern, & Rubin, 2003)

$$\phi = \{\mu_{0,b}, \sigma^2_b, \mu_{0,p}, \sigma^2_p, \kappa_0, \nu_0\} \quad (2)$$

$$\theta|\phi \sim \prod_{c\in\{b,p\}} \text{Normal}(\mu_c|\mu_{0,c}, \frac{\sigma^2_c}{\kappa_0})\chi^{-2}(\sigma^2_c|\sigma^2_{0,c}, \nu_0) \quad (3)$$

where μ_0 and σ^2_0 are a category's prior expected mean VOT and variance, respectively, and κ_0 and ν_0 are the listener's confidence in these prior expectations, measured as pseudo-counts. Note that, as in previous modeling work in this framework, these prior confidence parameters are shared between the two categories. Preliminary simulations showed that it wasn't possible to uniquely identify the model using separate prior confidence parameters for the two categories.

To estimate the listeners' prior beliefs, we infer values for these parameters given the observed adaptation behavior (category responses y and input VOTs x) using Bayesian inference, marginalizing over θ:

$$p(\phi|x,y) \propto p(y|\phi,x)p(\phi) \quad (4)$$

$$\propto \int d\theta p(y|\theta,x)p(\theta|x,\phi)p(\phi) \quad (5)$$

We two simplifying assumptions. First, we assume that the order of the trials does not matter. Second, we assume that listeners pick up on the cluster structure of the input they receive, accurately detecting the mean and variance of each cluster, but for categorizing they use their *posterior* beliefs

Parameter	Expected	95% HPD Int.	Units
κ_0	243	163–495	observations
$\mu_{0,b}$	-11	-28–3	ms VOT
$\mu_{0,p}$	56	44–72	ms VOT
ν_0	772	493–1180	observations
$\sigma_{0,b}$	19	16–23	ms VOT
$\sigma_{0,p}$	16	14–20	ms VOT

Table 2: Expected values and 95% highest posterior density intervals for the prior parameters, given the adaptation data.

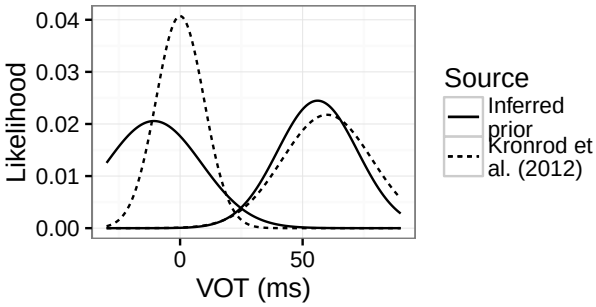


Figure 5: Expected cue distributions based on the prior beliefs inferred here from behavioral adaptation data (solid lines). Plotted with VOT distributions measured by Kronrod et al. (2012) based on a combination of classification and discrimination behavior (dashed lines).

which combine their prior beliefs with these observed statistics. Backing off these assumptions is possible, but computationally more demanding, and we leave it for future work to determine how they might affect our results. Finally, we also add a lapse rate parameter (maximum a posteriori value of 5%), that allows for some proportion of responses to be attributed to random guessing (e.g., because of attentional blinks, see Clayards et al., 2008 for a discussion).

The posterior distributions of each of these parameters (the shared prior beliefs plus lapsing rate) were estimated using MCMC with the Stan software package (Stan Development Team, 2015). Weakly informative hyperpriors were used that were centered at 0 with standard deviations of 100 for the prior expected means and variances (making them roughly constant over reasonable values) and 888 (four times the total number of trials that listeners heard) for the prior confidence pseudocounts (which is essentially uniform on the whole range from completely ignoring prior beliefs to not adapting at all). The prior for the lapsing rate was uniform on $[0,1]$. We ran four chains for 1000 samples each, discarding the first 500 as burn-in for a total of 2000 samples overall. This sampler converged well and achieved good mixing (maximum $\hat{R} = 1.01$; Gelman & Rubin, 1992).

Results

The first way we evaluate this model is to ask how well it fits listeners' behavior. Figure 4 shows listeners' average classification functions, compared with the posterior predictive classification functions from the belief-updating model. The first thing to notice is that the model fits the data well (log-likelihood ratio vs. an intercept-only binomial null model of 12385, and Spearman's $\rho = 0.9$, $p < 10^{-10}$ in both cases) capturing the different classification functions that result from exposure to each input distribution. This in and of itself is an interesting result: it shows that there does exist some set of prior beliefs such that the range of adaptation behavior we observed can be explained by a model where the listeners assigned to the different accent conditions all start from a common set of prior beliefs.

The second way to evaluate this model is based on the prior beliefs it infers listeners to have. Table 2 shows the posterior expectation and 95% highest posterior density intervals for each of the prior belief parameters given the adaptation data above. The behavioral data was consistent with high confidence in prior beliefs with the prior confidence about category variances ($E(\nu_0) = 772.39$) higher than confidence in the category means ($E(\kappa_0) = 243.26$). Both of these (measured in pseudo-observations) are larger than the number of trials that listeners heard in the experiment (222), which means that as far as the belief-updating model is concerned, listeners updated beliefs reflected their prior beliefs as much as (in the case of the means) or more than (for the variances) the distributions they actually observed. This is consistent with the qualitative finding that listeners' category boundaries are intermediate between the boundaries corresponding to a typical talker and the experimental exposure talker.

Figure 5 shows the cue distributions corresponding to the posterior expected values of the prior expected mean and variance parameters given the behavioral data, compared with the distributions corresponding to one typical talker (as determined by a combination of classification and discrimination data, Kronrod et al., 2012, see also Lisker & Abramson, 1964). The inferred prior beliefs are in reasonably good agreement with this typical talker, with one exception: the mean for /b/ is slightly lower ($E(\mu_{0,b}) = -10.69$ ms), and the standard deviation of /b/ is slightly higher ($E(\sigma_{0,b}) = 19.4$ ms).

Discussion

These modeling results show that the pattern of adaptation behavior observed above is consistent with a belief-updating model of phonetic adaptation that combines prior expectations with input statistics in order to infer the current talker's cue distributions. Specifically, it shows that there exists a single set of prior beliefs that captures the range of adaptation to different input distributions that stretches from nearly complete adaptation to partial adaptation at best.

These prior beliefs are reasonably consistent with other attempts to determine what listeners think the underlying cue distributions are (Kronrod et al., 2012), as well as the distributions produced by actual talkers (Lisker & Abramson, 1964). In fact, the prior expected VOT distribution for /p/

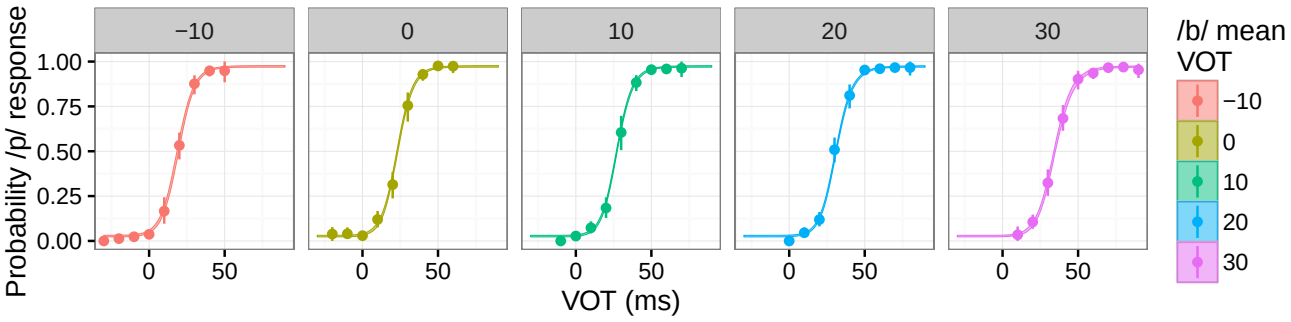


Figure 4: The classification functions (shaded ribbons, 95% posterior predictive intervals) predicted by the belief updating model fit listeners' responses well (dots with lines showing bootstrapped 95% confidence intervals).

that our model inferred is almost identical to that observed by both Kronrod et al. (2012) and Lisker and Abramson (1964). The distribution for /b/ deviates from prior work, however. One possible reason for this is that a substantial minority of English speakers produce pre-voiced /b/ (Lisker & Abramson, 1964), which is characterized by a lower (negative) VOT and a higher variance (often higher even than /p/). That is, across talkers, the /b/ VOT distribution parameters (mean and variance) have a *bimodal* distribution. We assumed a single, unimodal prior distribution, and the prior beliefs we inferred to be most likely are consistent with a compromise between the two types of /b/ distributions that talkers actually produce.

This possibility suggests two directions for future work. First, large-scale corpus studies of VOT distributions are needed to determine to what extent the distribution of /b/ mean VOTs is really bimodal across talkers. However, the only study of this type we are aware of considers only non-negative VOTs (Chodroff, Godfrey, Khudanpur, & Wilson, 2015) and thus excludes talkers who pre-voice their /b/. Second, more modeling work is needed to test whether a multimodal prior is justified given the adaptation data, and if so, whether it would change the inference about listeners' prior expectations for /b/.

The other major result of this modeling is that listeners have high confidence in their prior expectations about the VOT distributions of /b/ and /p/, acting as if they had already observed around 200–800 samples from each category (for the category means and variances, respectively) from the unfamiliar talker they encountered in our experiment.

At first blush, this conflicts with previous work on another phonetic contrast, /b/-/d/, which found confidence values that were one or two orders of magnitude smaller than those inferred here (Kleinschmidt & Jaeger, 2015). Interestingly, the /b/-/d/ contrast is cued by spectral cues (formant frequency transitions) which generally vary substantially across talkers (e.g., Peterson, 1952). The acoustic cues to the /b/-/p/ contrast used in the current study do not show as much variability across talkers (e.g., Allen, Miller, & DeSteno, 2003; Chodroff et al., 2015). When there is little variability across talkers, past experience with other talkers' VOT distributions is highly informative about the distributions that an unfamiliar talker will produce, requiring less adaptation. Likewise, when there is more variability across talkers, listeners need to rely more on the current talker's cue distributions and less on their prior experience. Thus, the apparent discrepancy between the confidence that listeners place in their prior beliefs in the current study and in Kleinschmidt and Jaeger (2015) is actually in line with an *ideal adapter* which combines prior beliefs with current experience weighted according to confidence. This idea finds further empirical support in Kraljic and Samuel (2007), who found that after the same amount of exposure, listeners recalibrate a /d/-/t/ contrast (analogous to the /b/-/p/ contrast used here) much less than an /s/-/ʃ/ contrast (where the latter exhibits larger variability across talkers; e.g., Newman, Clouse, & Burnham, 2001).

Conclusion

A central prediction of the ideal adapter framework (Kleinschmidt & Jaeger, 2015) is that listeners adapt to unfamiliar talkers by combining their prior beliefs with observed evidence about that particular talker's cue distributions. In this paper, we have shown first that for a range of different accents (cue distributions), listeners' behavior in a distributional learning experiment reflects a compromise between what would be expected for the cue distributions produced by a typical talker and the exposure talker. Second, the range of adaptation behavior observed across the various accents that listeners heard can be captured by a belief-updating model with a single set of prior expectations that are updated based on experience with the exposure talker.

These results emphasize the importance of listeners' prior expectations for robust speech perception in the face of talker variability. Even if all the listener knows about the talker is that they are speaking English, they can still benefit from prior experience with other speakers of English to provide an informative head start for adaptation. The modeling framework we use has the additional advantage of allowing us to *infer* what cue distributions listeners believe an unfamiliar talker will produce. This provides a potentially powerful—and heretofore missing—tool for probing listeners' prior ex-

pectations, based only on comprehension data. These beliefs reflect what listeners have learned about the variability they can expect across talkers, and probing how this internal model is related to the *actual* variability across talkers (measured via speech production data) is an important next step in advancing our understanding of robust speech perception.

More generally, prior knowledge is increasingly understood to play in important role in a number of perceptual and memory domains (e.g., Brady & Tenenbaum, 2013; Froyen, Feldman, & Singh, 2015; Orhan & Jacobs, 2011). Distributional learning provides an approach to probing prior expectations about the *statistics* of the sensory world, which, as in speech perception, are critical to effectively coping with non-stationarity in sensory statistics.

Acknowledgements

This work was partially funded by an NSF Graduate Research Fellowship to DFK and NICHD R01 HD075797 as well as an Alfred P. Sloan Fellowship to TFJ. The views expressed here are those of the authors and not necessarily those of the funding agencies.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544. doi: 10.1121/1.1528172

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85–109. doi: 10.1037/a0030779

Chodroff, E., Godfrey, J., Khudanpur, S., & Wilson, C. (2015). Structured Variability in Acoustic Realization : A Corpus Study of Voice Onset Time in American English Stops. In The Scottish Consortium for ICPhS (Ed.), *Proceedings of the 18th international congress of phonetic sciences.* Glasgow, UK: the University of Glasgow.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647. doi: 10.1121/1.1815131

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi: 10.1016/j.cognition.2008.04.004

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–2213.

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian Hierarchical Grouping : Perceptual Grouping as Mixture Estimation. *Psychological Review*, *122*(4), 575–597.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second ed.). Taylor & Francis.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–56. doi: 10.1037/a0025641

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. doi: 10.1037/a0038695

Kleinschmidt, D. F., Raizada, R., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. In R. Dale et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. doi: 10.1016/j.jml.2006.07.010

Kronrod, Y., Coppess, E., & Feldman, N. H. (2012). A Unified Model of Categorical Effects in Consonant and Vowel Perception. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 629–634). Austin, TX: Cognitive Science Society.

Lisker, L., & Abramson, A. (1964). A cross-language sudy of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Newman, R. S., Clouse, S. a., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196. doi: 10.1121/1.1348009

Orhan, A. E., & Jacobs, R. A. (2011). A Nonparametric Bayesian Model of Visual Short-Term Memory. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2451–2456). Austin, TX: Cognitive Science Society.

Peterson, G. E. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175. doi: 10.1121/1.1906875

Stan Development Team. (2015). *Stan: A C++ Library for Probability and Sampling, Version 2.9.0.*

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–6. doi: 10.1016/j.cognition.2010.10.018

Working Memory Affects Attention to Loss Value and Loss Frequency in Decision-Making under Uncertainty

Bo Pang (pangbo@tamu.edu), Kaileigh A. Byrne (kbyrne21@tamu.edu), Darrell A. Worthy (worthyda@tamu.edu)
Department of Psychology, TAMU 4235, College Station, TX 77843 USA

Abstract

Decision-making under uncertainty is pervasive. This work sought to understand the role of working memory (WM) in loss sensitivity by utilizing two widely used tasks, the Iowa Gambling Task (IGT) and the Soochow Gambling Task (SGT), and manipulating WM with a dual-task paradigm. We hypothesized that WM load would reduce attention to both loss value and frequency in the decision-making tasks. To better delineate the psychological processes underpinning choice behavior, we developed an Expectancy-Frequency-Perseveration (EFP) model which parsimoniously captures three critical factors driving choices: expected value, frequency of gains and losses, and perseveration. Behavioral and computational modeling results indicate that WM load compromised performance in the IGT due to reduced attention to loss value but enhanced performance in the SGT because of diminished attention to loss frequency. Our findings suggest that WM heightens attention to losses, but that greater attention is given to loss frequency than loss value.

Keywords: decision-making under uncertainty, working memory, loss, frequency, Iowa Gambling Task

Introduction

Many decisions in everyday life involve some degree of uncertainty, either small decisions, such as whether to pick a new restaurant for dinner or eat at one you frequently visit, or major ones, such as whether to invest in a risky hedge fund for a foreign oil company or invest in safer alternatives such as bonds. The Iowa Gambling Task (IGT; Bechara, Damasio, Damasio, & Anderson, 1994) and the Soochow Gambling Task (SGT; Chiu et al., 2008) mimic real-life decision-making situations involving uncertainty. In the present work we utilize these two tasks to examine how working memory load affects sensitivity to losses versus gains, a critical factor in many decision-making situations.

The IGT is widely used to examine choice behavior in various clinical populations (e.g., brain damage, substance abuse), developmental samples, and in healthy adults (Buelow & Suhr, 2009). The IGT manipulates the uncertainty of premises and outcomes, as well as gains and losses provided by each deck. In this task, players choose between four decks of cards, which yield both gains and losses. Unbeknownst to players, Decks A and B are disadvantageous because they have a negative net expected value, while Decks C and D are advantageous because they have a positive net expected value. The task is initially challenging because the disadvantageous decks consistently yield larger gains (100 versus 50 points), yet they also give larger losses.

Although it is assumed that players make choices by comparing expected values for options in the IGT, some authors present critiques that gain-loss frequency plays an important role in choice behavior in the IGT (Chiu et al., 2008; Steingroever, Wetzels, & Horstmann, 2013). In the IGT, Decks A and C give frequent losses (on 50% of trials), while Decks B and D give less frequent losses (on 10% of trials). A tendency to avoid decks with frequent losses will not have an effect on the net amount of points gained since the high-frequency loss decks are split across the advantageous and disadvantageous decks. The SGT was recently developed to further distinguish the influence of expected value and gain-loss frequency (Chiu et al., 2008). In this task the two advantageous decks (C and D) also give the most frequent losses – on 80% of trials compared to only 20% of trials for the disadvantageous decks (A and B). Thus, a tendency to avoid decks that give frequent losses will lead to poor performance, but a tendency to focus on the net expected value will lead to good performance. Comparing performance in both tasks allows for better inference about the mechanisms that contribute to decision-making performance. Attention to the net expected values provided by each deck should lead to good performance in both tasks, while attention to the frequency of losses should have little effect on performance in the IGT, but a negative effect on performance in the SGT.

Working memory (WM), a central component of cognitive executive function, is crucial to a variety of higher-order cognitive tasks. It might contribute to the decision-making processes in the IGT and the SGT. Recently, a number of studies have examined the role of WM on choice behavior in the IGT or its variants, but the findings have been somewhat mixed (e.g., Hinson, Jameson, & Whitney, 2002; Turnbull, Evans, Bunce, Carzolio, & O'Connor, 2005). Thus, the first goal of the current research was to examine whether completing the IGT requires WM by utilizing the original IGT and a dual-task paradigm. We also evaluated the contribution of WM on choice behavior in the SGT, which is the first attempt to do so. As such, results from these two tasks would provide convergent evidence for the role of WM in decision-making.

Perhaps of greater importance, we sought to pinpoint the specific mechanisms through which choice behavior in the IGT and SGT are reliant on WM. Sensitivity to potential losses is crucial to make reasonable choices in the IGT and other decision scenarios. With impaired attention to losses, decision makers might base their choices mainly on the gains that they expect to receive. In the context of the IGT, such a strategy leads to preference for the disadvantageous decks because these options consistently yield larger gains than the advantageous decks (e.g., Cella, Dymond, Cooper, & Turnbull, 2012). WM might be important to maintain attention to losses, especially in such a complicated task. For example, in a study demonstrating that WM load interferes

with performance in a three-deck variant of the IGT (Hinson et al., 2002), participants under low WM load produced different anticipatory emotional reactions to different decks as the task progressed. These prospective emotional reactions are theorized as somatic markers and assumed to facilitate decision-making under uncertainty (Bechara & Damasio, 2005; Damasio, 1994). Participants under high WM load, however, did not exhibit distinguishable emotional reactions, which might be due to inadequate sensitivity to losses when WM resources were taxed by a WM-demanding concurrent task. We speculated that intact WM resources may enable decision makers to adequately attend to gains and losses and maintain them across trials, making it possible that participants represent appropriate expected values for each option. However, insufficient WM resources may impair players' attention to losses and bias them towards gains, causing inaccurate expected values and poor performance in both tasks.

Besides expected value, attention to the frequency of losses versus gains might also be an important component of decision-making that is mediated by WM. Tracking the frequency and choosing the option with infrequent losses and frequent gains is a heuristic-based strategy, which appears simple but might be WM-sensitive. Indeed, previous studies have shown that participants with intact WM resources prefer heuristic-based strategies, such as win-stay, lose-shift, by remembering the outcomes from past trials more than participants under WM load (Worthy, Otto, & Maddox, 2012). In contrast, Worthy et al. (2012) found that participants under WM load preferred a strategy that implicitly track expected values of options. One possibility is that the frequency of losses is more salient than the net value of each option. WM load may attenuate participants' ability to attend to both the value and the frequency of losses, but because the frequency of losses is more salient than the value of losses, participants with intact WM resources will attend more to the frequency of losses than to the value of losses. This would lead to a pattern of behavior where participants performing the task under no-load conditions would perform better than participants under WM-load on the IGT due to superior attention to the loss value of each deck, but worse on the SGT due to enhanced attention to the frequency of losses.

Computational Modeling

In addition to our behavioral approach of utilizing tasks where attention to loss frequency differentially affects performance, we also utilize computational modeling to isolate and identify these specific psychological mechanisms underpinning choice behavior. A range of computational models have been applied to IGT and SGT data. Prospect valence learning (PVL) models, including PVL-Delta and PVL-Decay, and value-plus-perseveration (VPP) models have been most popular in recent work (Ahn et al., 2008; Worthy, Pang, & Byrne, 2013). As with most reinforcement learning (RL) models (Sutton & Barto, 1998), the basic assumptions behind the PVL models are that outcomes of past decisions are integrated to determine expected values for each option and that decision makers tend to choose options with larger expected value (Ahn et al., 2008). The VPP model further accounts for both the tendency to choose the option with the highest expected value and to perseverate or stay with the same option that was selected on the past trial (Worthy et al., 2013). However, as discussed above, attention to gain-loss frequency is another central source of choice behavior, which may also be affected by WM load. As such, in this work, we developed a new model which accounts for attention to expected value, to the frequency of gains versus losses, and perseveration, which we believe are three critical mechanisms in gambling tasks. We call this the Expectancy-Frequency-Perseveration (EFP) model.

The PVL models and the VPP model have been extensively discussed in many recent articles (e.g., Ahn et al., 2008; Worthy et al., 2013). Readers are referred to these articles for the models' details. In what follows, we first introduce the EFP model and evaluate it against other models by applying all the models to a large data set of IGT data. We then use the EFP model to simulate choice behavior in the IGT and SGT. This model explicitly accounts for the attention people give to the frequency of net gains versus losses so that we could make predictions as to how attention to gain-loss frequency would affect decision-making performance in the tasks.

EFP Model In contrast to the PVL models with a single expected value term and the VPP model with expected value and perseveration terms, the EFP model includes three terms to account for three critical components of choice behavior: expected value, gain-loss frequency, and perseveration. Increasing the number of terms may ostensibly improve the fit of the model or lead to overfitting simply because the model has too many parameters. Considering this, we sought to design a model to capture these three important psychological components while keeping it as parsimonious as possible.

This model first assumes that after making a choice and receiving feedback ($gain(t)$ and $loss(t)$), the utility $u(t)$ for the choice made on trial *t* is given by:

$$u(t) = gain(t) - \rho \cdot |loss(t)| \quad (1)$$

Here ρ represents a loss aversion parameter ($0 \leq \rho \leq 5$) that governs the sensitivity of losses compared to gains. A value of ρ greater than 1 indicates that an individual is more sensitive to losses than gains, and a value less than 1 indicates greater sensitivity to gains than to losses. Note that the EFP model assumes that the subjective utility is linearly proportional to the actual payoff amount, in contrast to the PVL models that use a nonlinear function (see Ahn et al., 2008). One major reason for the nonlinear function is to implicitly account for the gain-loss frequency. The EFP model, however, explicitly captures the gain-loss frequency (see below) and thus a shape parameter is not necessary. Additionally, using a linear function improves the parsimony.

The EFP model then assumes that the utility $u(t)$ is used to update expectancies $E_j(t)$ for the chosen option, *i,* on trial *t*. It utilizes the Delta rule (Sutton & Barto, 1998) which

assumes that expectancies are recency-weighted averages of the rewards received for each option:

$$E_i(t) = E_i(t-1) + \phi \cdot [u(t) - E_i(t-1)] \quad (2)$$

Here ϕ represents the recency parameter ($0 \leq \phi \leq 1$) that describes the weight given to recent outcomes in updating expectancies. As ϕ approaches 1, greater weight is given to the most recent outcomes in updating expectancies, indicating more active updating of expectancies.

The perseveration term in the VPP model was designed to model the tendency to perseverate following gains and to switch following losses. Thus, it also implicitly captures the frequency of gains and losses. In the EFP model we decompose the tendency to select the option with infrequent losses and frequent gains and the tendency to perseverate. The frequency term for chosen option *i*, on trial *t*, differed based on whether the net outcome, *x(t)*, was positive or negative:

$$F_i(t) = \begin{cases} (1-\phi)\cdot F_i(t-1)+1 & \text{if } x(t)\geq 0 \\ (1-\phi)\cdot F_i(t-1)-1 & \text{if } x(t)<0 \end{cases} \quad (3)$$

The frequency value increases by 1 following a net gain or decreases by 1 following a net loss. Instead of using a separate parameter to capture the weight to previous information as in the VPP model, the EFP model utilizes $1 - \phi$. Here ϕ has the same meaning as in Equation 2, accounting for weight given to recent information. Thus, utilizing the same recency parameter for both the value updating function and the gain-loss frequency function increases the parsimony of the EFP model.

The perseveration term for chosen option *i*, on trial *t*, is determined by:

$$P_i(t) = \gamma \quad (4)$$

The tendency to perseverate or switch is denoted by γ which varies between -100 and 100. This perseveration term simply gives a bonus or a reduction to the value of the option that was selected on the last trial, and thus indicates a general tendency to stay or switch to a different option on each trial.

The overall value of each option was determined by taking a weighted average of the expected value and the frequency plus the perseveration strength of each *j* option:

$$V_j(t) = \omega \cdot E_j(t) + (1-\omega) \cdot F_j(t) + P_i(t) \quad (5)$$

where ω ($0 \leq \omega \leq 1$) quantifies the weight given to the expected value for each option versus the weight given to the frequency of losses versus gains provided by each option.

Finally, these overall values $V_j(t)$ were entered into a Softmax rule function to determine the probability of selecting each option, *j*, on each trial, *t*:

$$Pr(G_j(t)) = \frac{e^{\theta(t)\cdot V_j(t)]}}{\sum_{j=1}^{4} e^{[\theta(t)\cdot V_j(t)]}} \quad (6)$$

$$\theta(t) = 3^c - 1 \quad (7)$$

Here *c* ($0 \leq c \leq 5$) represents the response consistency or exploitation parameter. Lower values indicate more random responding over the course of the task.

Model Evaluation We then compared the fit of the EFP, VPP, PVL-Delta, and PVL-Decay models by employing a large IGT dataset (*N*=504) of healthy participants. We fit each participant's data by maximizing the log-likelihood for each model's prediction on each trial. We used Bayesian Information Criterion (BIC; Schwarz, 1978) to assess the relative fit of the model. BIC penalizes models with more free parameters. For each model, *i*, BIC_i is defined as:

$$BIC_i = -2\log L_i + V_i \log(n)$$

where L_i is the maximum likelihood for model *i*, V_i is the number of free parameters, and *n* is the number of trials. Smaller BIC values indicate a better fit to the data. The EFP model exhibited the smallest median BIC value (see Table 1), indicating that it provides a better fit to the data than other models. Also, note that the VPP model fits were close to those of the EFP model, followed by the PVL-Decay model. The PVL-Delta model had a poorer fit compared to these models.

Table 1: Median BIC values for each model

Model	EFP	VPP	PVL-Delta	PVL-Decay
	242.58	242.85	256.76	246.05

Simulations and Predictions To predict the effects of WM load on participants' performance, the proportion of trials when the good decks are selected minus the proportion of trials that the bad decks are selected, we next utilized the EFP model, to simulate choice behavior in the IGT and SGT. Specifically, we focused on how WM load would affect performance by biasing attention to either the value or frequency of losses versus gains. We conducted 2,000 simulations for the EFP model for each of the two tasks by systematically varying either the loss aversion parameter or the weight parameter for value versus gain-loss frequency, while fixing other parameters at reasonable values. In the loss aversion simulations, the loss aversion parameter varied from 0 to 5 in increments of .5, other parameters used were .5 for recency, 1 for exploitation, .5 for weight (indicating equal attention to the expected value and gain-loss frequency), and 0 for perseveration. Figure 1 (left panel) displays the results for the loss aversion parameter. As attention to losses increased, performance improved. This relationship held for both the IGT and SGT, although the slope for the SGT was slightly steeper. In the weight parameter simulation, the weight parameter varied from 0 to 1 with an increment of .1, other parameters were .5 for recency, 1 for exploitation, 1 for loss aversion (indicating equal weighting of the value of gains and losses), and 0 for perseveration. Figure 1 (right panel) depicts the simulation results. Clearly, increased weight to expected value enhanced performance in the SGT. A similar tendency was present for the IGT but was considerably weaker. It appears that weight to expected value versus gain-loss frequency does not strongly impact performance in the IGT. This notion is consistent with the payoff structure of the IGT. Both advantageous and disadvantageous IGT decks include one option with high-frequency losses and another option with low-frequency losses. As such, reliance on the frequency of gains versus losses does not have a strong impact on performance in the IGT.

As discussed earlier, we predict that WM load would compromise attention to losses, or reduce the loss aversion parameter value in the EFP model, and drive participants

away from the frequency heuristic, or increase the weight parameter value in the EFP model. According to the simulations, in the IGT decreased loss aversion is associated with poorer performance, while greater weight to expected value does not strongly impact performance. Thus, we predicted that WM load would impair performance in the IGT. In the SGT, reduced loss aversion is also associated poorer performance, but increased weight to expected value versus gain-loss frequency leads to better performance. As such, WM load should either improve or compromise performance in the SGT, depending on whether participants performing the task without WM load attend more to the value or to the frequency of losses.

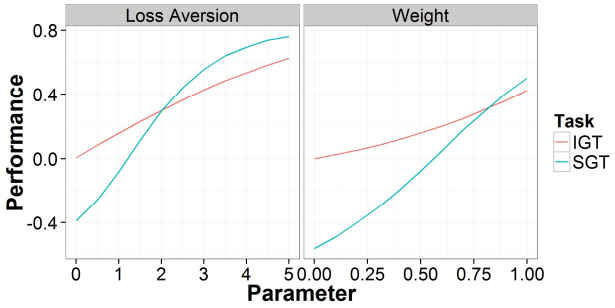


Figure 1. Performance in the IGT and SGT from the EFP model's simulations with varying loss aversion parameter (left panel) and weight parameter (right panel) values.

Experiment In this work, we aimed to examine whether WM contributes to decision-making under uncertainty by utilizing the IGT (Experiment 1A) and SGT (Experiment 1B), and to investigate what specific decision-making mechanisms WM load affects. To manipulate WM, we used a numerical Stroop task which has been used in previous experiments (e.g., Worthy, Otto, & Maddox, 2012). Specifically, in the single task (ST) condition participants only performed the decision-making task, the IGT (Experiment 1A) or the SGT (Experiment 1B), while in the dual task (DT) condition participants concurrently performed the decision-making task and the numerical Stroop task. Two potential mechanisms were tested that might drive the effects of WM load on the decision performance: attention to losses and gain-loss frequency. We combined behavioral analysis and computational modeling to evaluate these possible mechanisms. The computational models described above were fit to the data, and best-fitting parameter estimates, indicative of specific psychological components involved in decision-making processes, from the best-fitting model were compared between the ST and DT conditions. This procedure enabled us to infer the mechanisms whereby WM load affected decision-making behavior.

Experiments 1A 1B

Method

Participants 169 participants (101 females) recruited from an introductory psychology course at Texas A&M University participated in the experiment for course credit. Participants were randomly assigned to either the ST or DT condition in Experiments 1A and 1B.

Materials and Procedures Participants performed the experiment on PCs using Psychtoolbox for Matlab (version 2.5). In Experiment 1A, participants in the ST condition performed the computerized IGT (Bechara et al., 1997). On each of 100 trials four decks of cards appeared on the screen and participants were prompted to select one deck. Upon each selection the computer screen displayed the gain and loss, if applicable, and net value beneath the card decks. The task was self-paced, and participants were unaware of how many card draws they would receive. The schedule of gains and losses was identical to those used in the original IGT (Bechara et al., 1994).

In the DT conditions, in addition to the IGT participants performed a numerical Stroop task concurrently. The memory task required participants to remember which of two numbers was physically larger and which was larger in numerical value while performing the IGT. At the beginning of each trial, two numbers for the concurrent memory task were presented on each side of the screen, one number on each side, for 300 ms. Participants were then allowed to make a selection from among four decks of cards, followed by feedback as mentioned above. A new screen then appeared that queried participants with either *VALUE* or *SIZE*, and they selected either *Left* or *Right* to indicate which side had the number largest in either numerical value or physical size. Upon making a selection, they were told whether they were correct or not, and then the next trial began. Participants were told that they should focus on achieving good performance on the numerical Stroop task and "use what you have left over" for the decision-making task. To allow them to become familiar with the procedure, participants were given 10 practice trials. The practice trials were the same as the formal ones except that each selection on the IGT resulted in zero points regardless of which deck they selected.

The materials and procedures in Experiment 1B were identical to those in Experiment 1A except that participants performed the SGT (see Chiu et al., 2008) instead of the IGT.

Results and Discussion

Experiment 1A We examined performance by 20-trial blocks in the IGT (see the right panel of Figure 2). A mixed ANOVA with WM load (ST versus DT) as a between-subjects factor and Block (five 20-trial blocks) as a within-subject factor revealed a significant main effect of WM load, $F(1, 84) = 12.35$, $p = .001$, $\eta_p^2 = .13$, but not of block, $F < 1$. Moreover, there was a significant interaction, $F(4, 336) = 6.74$, $p < .001$, $\eta_p^2 = .07$. To examine this interaction, we looked at the simple effect of block within each WM load condition using trend analysis. For ST participants, performance improved linearly as the task progressed, $F(1, 42) = 12.27$, $p = .001$, $\eta_p^2 = .23$. In contrast, participants in the DT condition showed a linear downwards trend in performance across blocks, $F(1, 42) = 5.25$, $p = .03$, $\eta_p^2 = .11$. That is, participants learned to perform better over time,

whereas WM load impaired normal progress and led to worse performance across blocks.

Experiment 1B The same ANOVA for SGT performance revealed a significant main effect of WM load, $F(1, 81) = 6.50$, $p = .01$, $\eta_p^2 = .07$, and for block, $F(4, 324) = 14.76$, $p = .00$, $\eta_p^2 = .15$. The WM load X Block interaction was also significant, $F(4, 324) = 2.95$, $p = .02$, $\eta_p^2 = .04$. To examine this interaction, we looked at the simple effect of WM load within each block using *t*-tests. As can be seen in Figure 2 (right panel), although participants in both the ST and DT conditions appeared to learn to perform better across the task, DT participants performed better compared to ST participants in the first three blocks (*p*s < .01), but ST participants reached a similar performance level as DT participants in the last two blocks (*p*s > .72). This result suggests that WM load affected decision-making early in the SGT such that WM load improved performance.

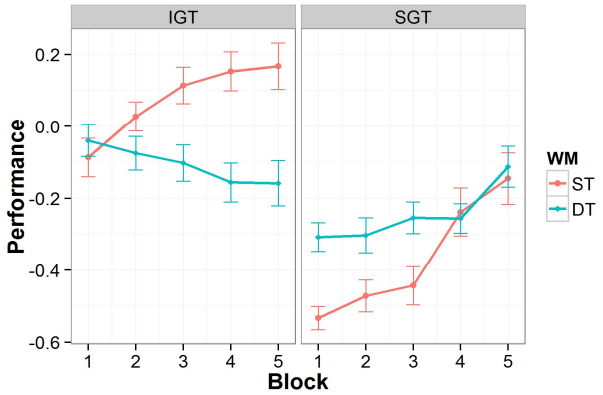


Figure 2. The performance in the IGT (right panel) and in the SGT (left panel) by 20-trial block for each WM condition (ST versus DT).

Computational Modeling

Table 2: Median BIC values for each model as a function of the task and WM condition

	Model	IGT	SGT
ST	EFP	**261.71**	257.96
	VPP	262.68	258.76
	Delta	277.96	264.07
	Decay	271.16	**252.85**
DT	EFP	**273.90**	**269.86**
	VPP	286.20	277.73
	Delta	284.18	274.67
	Decay	285.47	275.60

Model Selection

Four models, the EFP model, the two PVL models, and the VPP model, were evaluated with the method as above. Median BIC values are listed in Table 2. In the both WM conditions for the IGT and the WM load condition for the SGT, the EFP model had the smallest median BIC values, suggesting that this model fit the data best in these conditions. In the ST condition for the SGT, the EFP model exhibited larger median BIC value than the PVL-Decay model. Considering the model fitting results here and those from the aforementioned large IGT dataset, it seems clear that the EFP, PVL-Decay, and VPP models provide a better fit than the PVL-Delta model.

Modeling Results

We next compared the parameter estimates of the EFP model between the ST and DT conditions for each task to examine the effects of WM load on specific psychological processes related to decision-making. Table 3 lists the average best fitting parameter values of the EFP model for each task under each WM condition. Nonparametric Mann-Whitney *U* tests were used because the best-fitting model parameters were not normally distributed. In the IGT, ST participants' data were best fit by higher loss aversion parameter values than DT participants' data, $U = 569$, $p = .002$. This suggests that ST participants were more attentive to losses than DT participants, thus providing direct evidence to support the impaired loss sensitivity hypothesis. Furthermore, participants under WM load showed marginally significantly higher weight to RL expected value, $U = 708$, $p = .057$, providing some evidence to support our prediction that WM load would cause less reliance on a frequency-based strategy.

In the SGT, ST participants exhibited lower values for the weight parameter than did DT participants, $U = 645$, $p = .04$. This suggests that participants with compromised cognitive resources were less likely to utilize the frequency heuristic in the SGT, which is consistent our predictions the IGT results. Although we observed a trend that ST participants showed a higher loss aversion parameter estimates than did DT participants, this difference did not reach significance, $U = 697$, $p = .13$. Moreover, data from ST participants were best fit by higher recency parameter values than data from DT participants, $U = 643$, $p < .05$. This results suggests that ST participants were more attentive to recent outcomes, which might allow participants with intact WM resources to more actively update expectancies compared to participants under WM load.

Table 3: Average parameter estimates as a function of the task and WM condition

Parameters	IGT		SGT	
	ST	DT	ST	DT
ϕ	0.40 (0.39)	0.32 (0.39)	0.40 (0.39)	0.19 (0.27)
ρ	2.11 (2.28)	0.61 (1.43)	1.56 (1.96)	0.97 (1.69)
γ	−0.41 (28.56)	−10.77 (26.92)	20.74 (37.56)	−0.96 (25.69)
ω	0.26 (0.39)	0.37 (0.43)	0.31 (0.41)	0.42 (0.46)
c	0.70 (0.96)	0.57 (0.94)	0.47 (0.48)	0.53 (0.71)

Standard deviations are listed in parentheses.

General Discussion

Our results provide clear evidence that WM contributes to choice behavior in decision-making under uncertainty. Results from Experiment 1A indicated that intact WM is necessary to do well on the IGT. These results were at odds with Turnbull et al. (2005), possibly because random number generation and articulatory suppression tasks used in Turnbull et al. did not tax enough WM resources to interfere with decision making processes, and/or their experiment lacked statistical power due to relatively small sample sizes (n = 25 for each group). On the other hand, our findings confirm implications from previous behavioral studies using variants of the IGT (instead of the original IGT; Hinson et al., 2002) and from brain lesion studies with DLPFC damage patients (Fellows & Farah, 2005). As such, this work contributes to the IGT literature by lending direct support to the idea that choice behavior in the IGT is dependent on WM. Moreover, results from Experiment 1B showed that WM load influenced performance in the SGT. This is the first work demonstrating that WM contributes to choice behavior in this frequently used decision-making task. Therefore, this work provides convergent evidence to support the notion that WM plays a role in decision-making under uncertainty.

Further, our results provide considerable insight into the mechanisms through which WM contributes to decision-making under uncertainty. First, we found that participants with compromised WM resources performed worse in the IGT and had data better fit by smaller loss aversion parameter values, suggesting that WM enables decision-makers to adequately attend to losses in decision-making under uncertainty. In contrast, we found that participants with impaired WM resources performed better overall in the SGT and exhibited greater weight to expected value versus gain/loss frequency, indicating that another role of WM in decision-making under uncertainty is to attend to the frequency of gains versus losses, rather than just the net expected value of each alternative. Much evidence suggests that a prediction error, the difference between the outcome received and the expected value for a given option, is tracked by the ventral striatum, a subcortical region implicated in implicit, procedural learning (e.g., Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). In many RL models, including models used in this article, these prediction errors are used to update the expected value for the option that was chosen on each trial. Given the ability of subcortical regions to track expected value, people may be able to implicitly learn which option provides the largest expected value across the task. However, with intact WM resources decision makers prefer heuristics such as the frequency heuristic which can be efficient in many situations, but counterproductive in situations like the SGT where gain-loss frequency is directly opposed with expected value.

Collectively, this work demonstrates that WM strongly contributes to choices involving uncertainty. Intact WM resources mainly enable decision makers to maintain adequate attention to loss value and loss frequency, with loss frequency receiving greater attention than loss value.

References

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32*, 1376–1402.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*, 7–15.

Bechara, A., Damasio, H., Tranel, D., & Damasio, a R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293–1295.

Buelow, M. T., & Suhr, J. A. (2009). Construct validity of the Iowa Gambling Task. *Neuropsychology Review, 19*, 102–114.

Cella, M., Dymond, S., Cooper, A., & Turnbull, H. (2012). Cognitive decision modelling of emotion-based learning impairment in schizophrenia: The role of awareness. *Psychiatry Research, 196*, 15–19.

Chiu, Y.-C., Lin, C.-H., Huang, J.-T., Lin, S., Lee, P.-L., & Hsieh, J.-C. (2008). Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa Gambling Task? *Behavioral and Brain Functions, 4*, 13.

Fellows, L., & Farah, M. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex, 15*, 58–63.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience, 28*, 5623–5630.

Hinson, M., Jameson, T., & Whitney, P. (2002). Somatic markers, working memory, and decision making. *Cognitive, Affective & Behavioral Neuroscience, 2*, 341–353.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Steingroever, H., Wetzels, R., & Horstmann, A. (2013). Performance of Healthy Participants on the Iowa Gambling Task. *Psychological Assessment, 25*, 180–193.

Sutton, A. G., & Barto, R. S. (1998). *Reinforcement Learning: An Introduction*. Cambridge: MIT press.

Turnbull, H., Evans, Y., Bunce, A., Carzolio, B., & O'connor, J. (2005). Emotion-based learning and central executive resources: an investigation of intuition and the Iowa Gambling Task. *Brain and Cognition, 57*, 244–247.

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(6), 1640–1658.

Worthy, D. A., Pang, B., & Byrne, K. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology, 4*, 1–9.

Adapting Deep Network Features to Capture Psychological Representations

Joshua C. Peterson (jpeterson@berkeley.edu)
Joshua T. Abbott (joshua.abbott@berkeley.edu)
Thomas L. Griffiths (thomas_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Deep neural networks have become increasingly successful at solving classic perception problems such as object recognition, semantic segmentation, and scene understanding, often reaching or surpassing human-level accuracy. This success is due in part to the ability of DNNs to learn useful representations of high-dimensional inputs, a problem that humans must also solve. We examine the relationship between the representations learned by these networks and human psychological representations recovered from similarity judgments. We find that deep features learned in service of object classification account for a significant amount of the variance in human similarity judgments for a set of animal images. However, these features do not capture some qualitative distinctions that are a key part of human representations. To remedy this, we develop a method for adapting deep features to align with human similarity judgments, resulting in image representations that can potentially be used to extend the scope of psychological experiments.

Keywords: deep learning; neural networks; psychological representations; similarity

Introduction

The resurgence of neural networks in the form of *deep learning* has continued to dominate object recognition benchmarks in the field of computer vision, often attaining near or above human-level accuracy for a variety of perceptual tasks, most notably through recent advances in classifying thousands of objects within natural images (Krizhevsky, Sutskever, & Hinton, 2012; He, Zhang, Ren, & Sun, 2015). Part of the success of these models is due to their ability to learn effective feature representations of high-dimensional inputs (e.g., complex color images); a challenge that human perception must also confront (Austerweil & Griffiths, 2013). As a result, cognitive scientists have started to explore how the representations learned by these networks can be used in models of human behavior for perceptual tasks such as predicting the memorability of objects in images (Dubey, Peterson, Khosla, Yang, & Ghanem, 2015) and predicting judgments of category typicality (Lake, Zaremba, Fergus, & Gureckis, 2015).

While deep learning models continue to mimic a growing list of human-like abilities, a number of core questions remain unanswered about the relevance of these models to actual human cognition and perception. For instance, features of the input learned using these networks excel in predicting certain human judgments, but how are these feature representations related to human psychological representations? At first glance, it would seem that the ability of these representations to predict typicality judgments and stimulus memorability would constitute robust evidence of their relevance to people, however recent work has shown that neural networks that classify images can be systematically deceived by imperceptible image transformations (Szegedy et al., 2013), casting doubt on their similarity to humans.

Understanding the relationship between the representations found by deep learning and those of humans is an important question in cognitive science, and could potentially benefit artificial intelligence. However, independent of this question, simply having a good approximation to how people represent images would allow cognitive scientists to test psychological theories using complex, realistic stimuli. Indeed, tasks such as creating stimulus sets that uniformly span psychological space are far from trivial.

In this paper, we address this question directly by examining how well features extracted from state-of-the-art deep neural networks predict human similarity judgments. An initial evaluation shows that these features account for a significant amount of variance in human judgments, but fail to capture qualitative distinctions that are key to human representations. We then develop a method for adapting deep network features to better predict human similarity judgments, and show that this approach can reproduce those qualitative distinctions. These results suggest that while raw features produced by deep learning may not be suitable for use in modeling cognition, they can be modified to bring them into close alignment with human representations.

Deep Representations

In general, deep neural networks (DNN) are neural networks that have depth in terms of their number of hidden layers between input and output (Bengio, 2009). In the past few years, training such networks to understand aspects of large, complex data sets has led to a number of advances in vision and language applications (LeCun, Bengio, & Hinton, 2015).

In computer vision, the majority of this progress has been driven by a particular DNN called a convolutional neural network (CNN) (LeCun et al., 1989). CNNs get their name from the use of convolutional layers, which learn a set of image filters that produce feature maps of spatially-organized inputs like images. This allows for a drastic decrease in the number of parameters the network must learn, which would otherwise explode exponentially in a fully connected network with high-dimensional inputs. The typical CNN architecture includes a series of hidden convolutional layers, followed by a smaller number of fully connected layers, and finally a layer that generates the final output or classification. While CNNs were initially developed over two decades ago, they came to mainstream popularity in 2012 when a 7-layer architecture named AlexNet (Krizhevsky, Sutskever, & Hin-

ton, 2012) won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), reducing the previous winner's error rate by an uncommonly large margin. Since then, a deeper CNN has won the contest every year, currently dominated by Microsoft's 150-layer network which obtained a best-of-top-5 error rate of 4.94%, surpassing the accuracy of non-expert humans at 5.1% (He, Zhang, Ren, & Sun, 2015).

Interestingly, CNNs produce much more than just their outputs (e.g., a category label for an image); they can also return feature representations at each layer of the network. The "deep representations" learned by these networks have proven useful in predicting human behavior. Dubey, Peterson, Khosla, Yang, & Ghanem (2015) used representations extracted from the last fully-connected layer of a CNN to predict the intrinsic memorability of objects. That is, the objects that humans are jointly likely to remember or forget in a large complex natural scene database. The correlation between estimates of memorability and the original memorability scores for each object matched human consistency (i.e. the correlation between memorability scores of random splits of the full sample of subjects). Similarly, Lake, Zaremba, Fergus, & Gureckis (2015) were able to reliably predict human typicality ratings of eight object categories using the same network and features, and called for cognitive scientists to pay attention to deep learning since categorization is a foundational problem in the field.

Deep representations are also beginning to interest the neuroscience community. For example, CNN activations have been used to predict monkey IT cortex activity (Yamins et al., 2014), as well as both low- and high-level activity in human visual areas (Agrawal, Stansbury, Malik, & Gallant, 2014). Delving deeper, Khaligh-Razavi & Kriegeskorte (2014) found that a CNN best explained IT cortex representations out of a set of 37 well-known models from both the computer vision and neuroscience fields, although no model completely explained all of the variance, unsupervised models being the worst of all of them.

Although CNN representations currently do the best job of predicting neural activity as measured by Blood Oxygenation Level Dependent (BOLD) response, this does not guarantee that we can explain psychological representations as a result. In fact, Mur et al. (2013) was partly successful in predicting human similarity judgments (a classic index of psychological representations) from IT cortex representations. However, the key categorical distinctions in the human representations were not well predicted: human IT cortex representations were more similar to monkey IT cortex representations than they were to human psychological representations. In the remainder of the paper, we use a similar approach to evaluate how well deep network features align with human psychological representations, and to explore how the correspondence between the two can be increased.

Evaluating Representations

Our first step is to evaluate the potential correspondence between deep network features and psychological representations. Unlike neural representations, psychological representations cannot be measured directly. However, both spatial and hierarchical psychological representations for N objects can be recovered given an $N \times N$ matrix of similarity judgments using methods such as multidimensional scaling and hierarchical clustering (Shepard, 1980). We thus reduce the problem to one of capturing human similarity judgments, subjecting both human judgments and model predictions to these different methods of extracting representations. We approach this problem by taking the inner-product of the deep feature representations of each pair of images (a measure of similarity between two vectors). We then compute the correlation between these pairwise vector similarities and human similarity judgments for the same stimulus pairs, which gives us a measure of the correspondence we want to evaluate.

Stimuli. Our stimulus set consisted of 120 color photographs of animals (sample images are shown in Figure 1). Images were cropped to 300×300 pixels, resulting in close-ups of either the animal's face or body. The set was constructed to include both inter- and intraspecies variation.

Behavioral Experiment. We collected pairwise similarity ratings for our animal stimulus set through Amazon Mechanical Turk. Participants were instructed to rate the similarity of four pairs of animal images on a scale from 0 (not similar at all) to 10 (very similar). We paid workers \$0.02 per set of four comparisons. Before each task, eight examples were shown to help prevent bias in early judgments. Amazon workers could repeat the task with new animal pairs as many times as they wanted. There were $7,140$ possible image comparisons, each of which was rated by 10 unique participants, for a total of $71,400$ ratings from 209 different participants. The result was a 120×120 similarity matrix after averaging over judgments.

Feature Extraction. We extracted features for each image in our data set using three different popular *off-the-shelf* CNNs of varying complexity that were pretrained in Caffe (Jia et al., 2014). Specifically, we used CaffeNet (based on original AlexNet), VGG16 (Simonyan & Zisserman, 2014), and GoogLeNet (Szegedy et al., 2014), the layer depths of which were 7, 16, and 22 respectively. GoogLeNet and VGG16 achieve roughly half the error rates of AlexNet. Each network had already been trained to classify 1000 object categories from previous ILSVRC competitions. A feedforward pass of each flattened image vector into each network yields feature responses at each layer. For our analysis, we extracted the last layer of each network before the classification layer. For CaffeNet and VGG16, this is a 4096-dimensional fully-connected layer, while the last



Figure 1: Samples from the set of 120 animal photographs.

Table 1: Correlations between human and deep similarities.

	CaffeNet	Google	VGG	HOG+SIFT
R^2	.32	.35	.43	.008

layer in GoogleNet is a 1000-dimensional average pooling layer. Lastly, we also extracted Histograms of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) representations for comparison since such features represent the generic representations of choice for tasks in computer vision prior to the popularity of deep learning.

Results

Table 1 gives performance (R^2) for each model. Raw representations from all three networks show medium to high correlations with the human data. In general, deeper networks with better ImageNet classification accuracy like GoogLeNet and VGG16 did better than CaffeNet, which is considerbly more shallow. The HOG+SIFT baseline did surprisingly poorly, explaining very little variance as compared to the deep representations, suggesting that while these features are useful for many computer vision tasks, they differ in large part from the representations humans employ when judging animal similarity.

Although the VGG representation explained a fair amount of variance, further analyses revealed that the most crucial structural aspects of the human representations were not preserved. The first and second panels of Figure 2 show multidimensional scaling (MDS) solutions for the original human data and the predictions from the unaltered deep representations. While the structure of the MDS solutions for the predicted judgments looks reasonable (e.g., zebras are next to other zebras), major categorical divisions are not preserved. Hierarchical clusterings of the actual and predicted human judgments (the first and second panels of Figure 3) show a similar pattern of results: human judgments exhibit several major categorical divisions, whereas much of this structure is lost in the predicted data.

Adapting Representations

After quantifying the discrepancy between deep and human representations, we can attempt to bring them into closer alignment. First, consider that the final hidden layer feature representation in a neural network can be thought of as the input to a final linear classification layer, such that the problem solved by the final weight matrix is a linear transformation (which is then often scaled by a softmax function to covert to class probabilities). This can be thought of as a rescaling of the final stimulus representation to solve the categorization problem. This suggests that we should not think about the features extracted by the network as a static representation, but as the ingredients for a transformation that solves a problem. Thinking in these terms, we show that we can easily solve for a linear transformation that better captures human similarity judgments.

Similarity Model. Any similarity matrix $\mathbf{S}$ can be decomposed into the matrix product of a feature-by-object matrix $\mathbf{F}$, its transpose, and a diagonal weight matrix $\mathbf{W}$,

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}^T \tag{1}$$

This formulation is similar to that employed by additive clustering models (Shepard & Arabie, 1979), wherein $\mathbf{F}$ represents a binary feature identity matrix (and is similar to Tversky's (1977) model of similarity). When used with continuous features, this approach is akin to factor analysis. Given an existing feature-by-object matrix $\mathbf{F}$, the diagonal of $\mathbf{W}$ can be solved for using linear regression where the predictors for each similarity s_{ij} are the product of the values of each feature for the objects i and j. When $\mathbf{W}$ is the identity matrix, this reduces to the model evaluated in the previous section.

$$s_{ij} = \sum_{i=1}^{N_f} w_k f_{ik} f_{jk}. \tag{2}$$

The result is a convex optimization problem that can be solved straightforwardly, allowing us to find a transformation of the deep features with a closer correspondence to human similarity judgments.

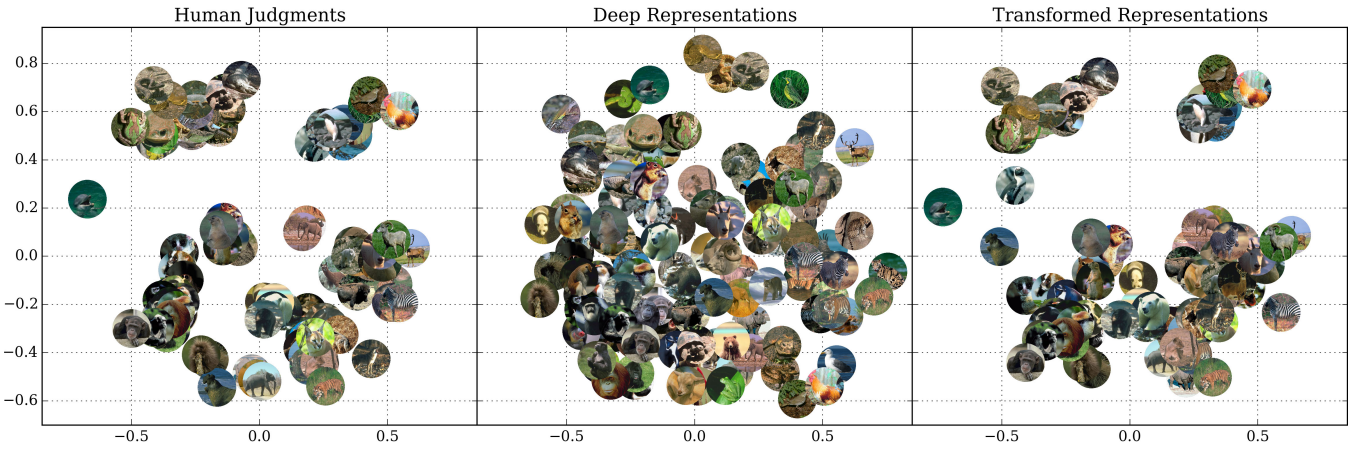


Figure 2: Multidimensional scaling solutions for similarity matrices obtained from human judgements (left), non-transformed deep representations (center), and transformed deep representations (right).

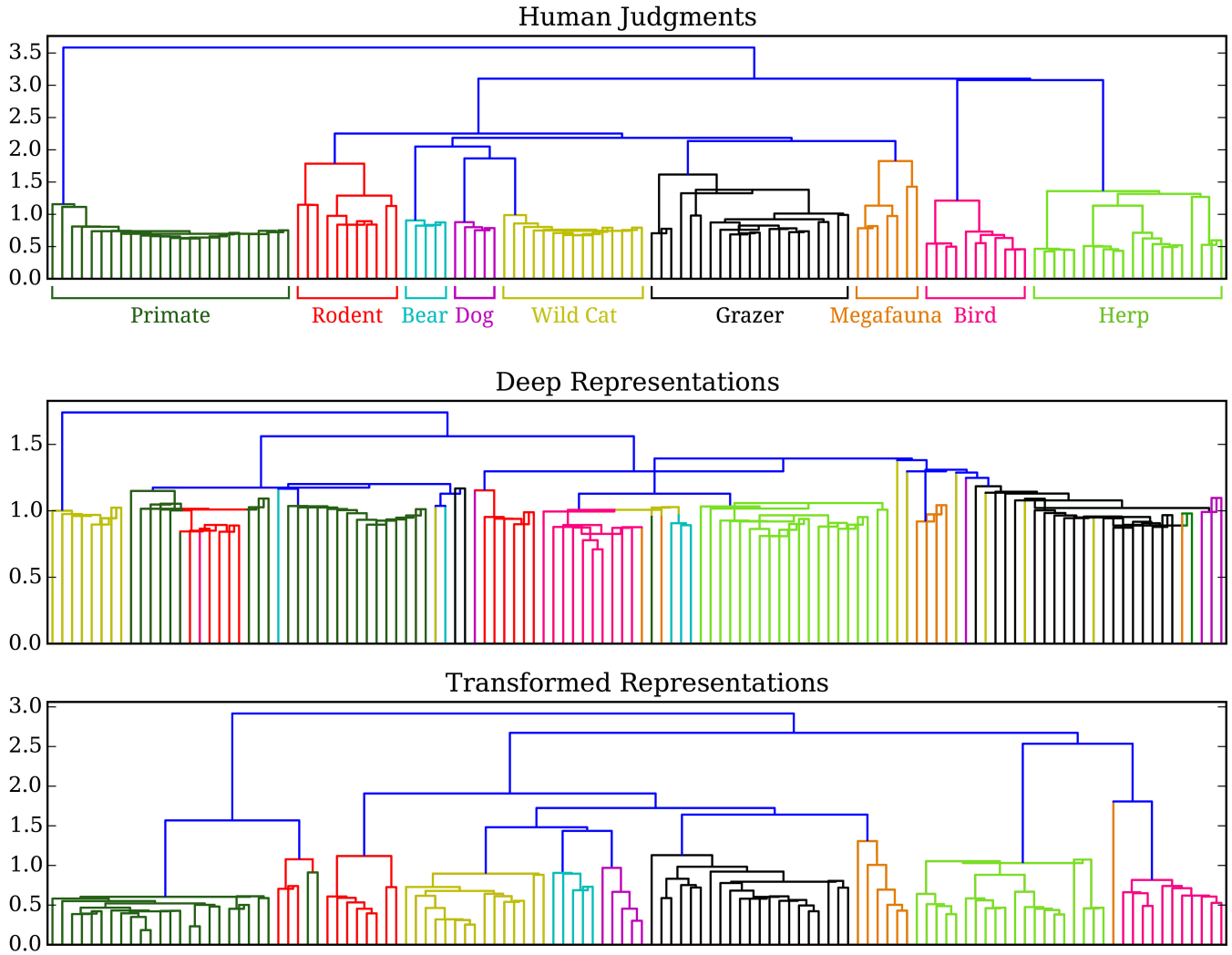


Figure 3: Hierarchical clustering of human judgements (top), deep representations (middle), and transformed representations (bottom). Human judgments resulted in nine interpretable clusters, grouped by color and semantic category label in the top panel. The leaves of the deep and transformed representation clusterings are color-coded relative to the human judgments.

Table 2: Model performance using adjusted CNN features.

	CaffeNet	Google	VGG	SIFT
R^2	.69	.72	.84	.09

Analysis. With such a large number of predictors, regularization is critical to avoid overfitting. We used ridge regression ($L2$ regularization) and performed grid search on cross-validated generalization performance to find the best regularization parameter. We predicted only the upper triangle of the similarity matrix since the matrix is symmetric. Each model was evaluated via its generalization performance in 6-fold cross-validation. We did this for the feature vectors extracted at each layer of the network.

As an additional control against overfitting, we compared model performance with several baselines. In Baseline 1, we shuffled the rows of the feature matrix such that the feature representation of one image was replaced with that of a different randomly chosen image. In Baseline 2, the columns of the feature matrix were randomly permuted for each row separately. Lastly, Baseline 3 simply combined the shuffling schemes from the first two baselines. In all three cases, the randomized feature matrices were subjected to the same set of analyses as the true features, allowing us to check for spurious correlations.

Results

Table 2 shows performance for each network using our adjustment of the representations. The R^2 values reported are the average values across all six folds of the crossvalidation. All five models performed considerably well, each showing improvement over the original non-weighted models. Most notably, VGG16 performed best, accounting for 84% of the variance. Training using the estimated regularization parameter on the entire dataset yielded an R^2 of 91%. In contrast, all three baseline models explained essentially no variance ($R^2 < 0.01$), suggesting that our results were not spurious correlations resulting from our large sets of predictors.

Crucially, the MDS solution for the improved predictions is almost identical to the original human spatial representation. The same improvements were found in hierarchical clusterings of actual and predicted similarity matrices (1st and 3rd panels of Figure 3), this time largely in the form of top-level parent nodes.

Feature Analysis. While higher layers in CNNs tend to produce the most generic high-level features for domain transfer across image applications, the choice of feature depth is ultimately dependent on the task (Sainath, Kingsbury, Mohamed, & Ramabhadran, 2013). This implies that layer responses at different depths may explain different types of human similarity judgments (e.g. tasks that involve comparing visual features versus conceptual information). We examined

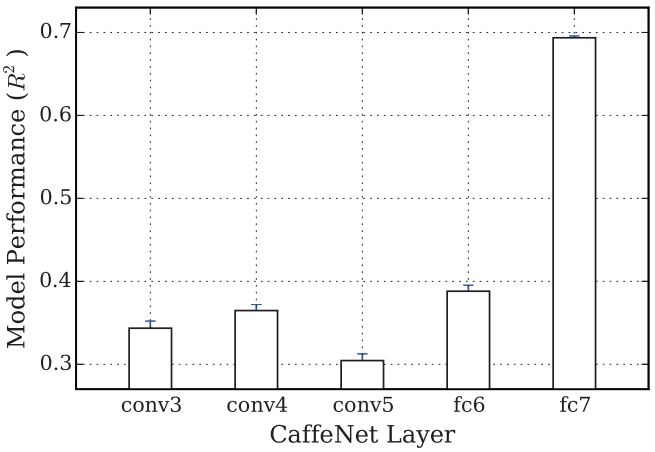


Figure 4: Model performance as a function of feature layer depth in CaffeNet.

our model's performance in predicting similarity judgments as a function of feature depth using CaffeNet, given its more straightforward architecture and manageably-sized layers. Specifically, we compared performance across the last three convolutional layers and the last two fully-connected layers. The results are shown in Figure 4. Performance does appear to correspond strongly to layer depth, although fully connected layers perform much better than convolutional layers, suggesting that human similarity judgments may not be explained well from simpler image features.

Reweighted Classification. We investigated the effect of our fine-tuned representations on a separate animal classifier, using a new animal data set consisting of 1,740 images from 19 animal classes (*bear, cougar, cow, coyote, deer, elephant, giraffe, goat, gorilla, horse, kangaroo, leopard, lion, panda, penguin, sheep, skunk, tiger, zebra*) (Afkham, Targhi, Eklundh, & Pronobis, 2008). We used multinomial logistic regression with 6-fold cross-validation to classify animals using fine-tuned representations as predictors. We fine-tuned these representations by pairwise multiplying the original VGG16 representations with the square-root of the weights obtained through prediction of the human similarity data. However, because some of the weights of the original solution are negative, we used Elastic Net regression to solve for weights constrained to be positive. We ran the same model using the original unaltered VGG16 representation to serve as baseline performance. The original model performed very well (average $R^2 = .94$), whereas the fine-tuned model performed consistently worse ($R^2 = .89$).

Discussion

This analysis constitutes the first formal comparison of deep representations to human psychological representations. Initial results using currently high-performing convolutional neural networks show that the two representations were moderately correlated, but diverge in terms of crucial structural

characteristics, a problem exhibited by similar experiments using neural representations as opposed to deep features (Mur et al., 2013).

Our method of overcoming this problem, by a parsimonious adjustment of the feature representation inspired by a classic model of similarity, appears to have been largely successful. Indeed, the human representations were almost completely reconstructed by our adjusted CNN features. Using features extracted from deep convolutional neural networks provides an opportunity to estimate psychological representations from real, raw sensory inputs (e.g. pixels). However, one potential limitation of this work is the generalizability of the transformation acquired to broader stimulus contexts. Testing this question will require replication and transfer across several domains. To the extent that this can be established, we envision our method as a standard tool for studying cognitive science using natural stimulus sets, on par with modern artificial intelligence.

Beyond this, we see potential for such an interface between cognitive science and artificial intelligence to be exploited for the benefit of each. While our attempt to improve a common categorization objective in computer vision (i.e. one-versus-all classification) using human-tuned representations was not successful, it does raise interesting distinctions between the computational problems solved by humans and CNNs. After all, the full breadth of human categorization behavior exhibits complex patterns such as overlapping class assignments, which are not likely to be well-represented when the learning objective is defined through images and objects characterized by a single label. Further, one might ask if poor categorization performance of the one-versus-all kind is the price paid for a more flexible system of categorization with respect to a set of complex objects that can be partitioned using several "good" configurations, depending on the context and task at hand. Given this possibility, one should be careful not to equate CNN classification performance with human categorization abilities in general.

Acknowledgments. This work was supported by grant number FA9550-13-1-0170 from the Air Force Office of Scientific Research. We thank Alex Huth for help with image selection.

References

Afkham, H. M., Targhi, A. A. T., Eklundh, J.-O., & Pronobis, J.-O. E. A. (2008). Joint visual vocabulary for animal classification. In *19th International Conference on Pattern Recognition (ICPR)* (pp. 1–4).

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *arXiv preprint arXiv:1407.5104*.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., & Ghanem, B. (2015). What makes an object memorable? In *International Conference on Computer Vision (ICCV)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Cognitive Science Society*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., & Ramabhadran, B. (2013). Learning filter banks within a deep neural network framework. In *Ieee workshop on Automatic Speech Recognition and Understanding* (pp. 297–302).

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Neurophysiological Effects of Negotiation Framing

Peter Khooshabeh (khooshabeh@ict.usc.edu) [1,2,3]
Rebecca Lin (rebecca.lin@psych.ucsb.edu) [2]
Celso de Melo (demelo@usc.edu) [3]
Jonathan Gratch (gratch@ict.usc.edu) [3]
Brett Ouimette & Jim Blascovich (blascovi@psych.ucsb.edu) [2]

[1] US Army Research Laboratory, Human Research and Engineering Directorate, Playa Vista, CA 90094 USA
[2]Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106 USA
[3]Institute for Creative Technologies, University of Southern California, Playa Vista, CA 90094

Abstract

In this study, we manipulated gain/loss framing context during a simulated negotiation between a human user and a virtual agent. Task instructions placed users either in a loss or gain framed context, such that those in the loss frame had to minimize expenses whereas those in the gain frame had to maximize profits. The virtual agent displayed facial emotions so that we could also test how interpersonal emotions interact with framing. Results suggest that individuals are more motivated to minimize their losses than maximizing their gains. The loss frame caused individuals to demand more during the negotiation, hence to minimize expenses. Neurophysiological results suggest that cardiovascular patterns of challenge (i.e., positive motivations) were present in the loss frame condition, most strongly when the virtual human smiled. We discuss these results in regards to Prospect Theory. This work also has implications for designing and rigorously evaluating human-like virtual agents.

Keywords: Prospect Theory; Negotiation; Context and Emotion; Gain/Loss Framing; Human-agent interaction

Introduction

Facing a situation with almost certain loss looming, individuals are generally more likely to engage cognitive processes to take risks. Conversely, facing almost certain gain, individuals are more likely to be risk averse. A concrete example of this is to consider a situation in which two individuals are presented with $100. One of them, Bob, is given the option of either (A) definitely keeping $40 or (B) taking a gamble with a 75% probability of winning the entire $100 (and 25% probability of losing everything). The other individual, John, is given the option of either (A) definitely losing $60 or (B) taking a gamble with a 75% probability of winning the entire $100. On average, research in experimental economics, game theory, and psychology has shown that individuals in Bob's decision frame would more likely be risk averse and choose (A), whereas individuals in John's situation would more likely be risk seeking and choose (B).

The goal of this work is to understand the behavioral and neurophysiological impacts of framing in an interpersonal decision making task, such as negotiation. Framing phenomena have been well studied, and have led to Prospect Theory (Kahneman & Tversky, 1979), which suggests that that people are more motivated to take greater risks in order to avoid losses, because "*Losses loom larger than gains* (Kahneman & Tversky, 1979, p. 279)." Related research has focused on the effects of framing on negotiation behavior. For example, how a negotiation is framed can affect how an individual approaches, views and responds to the negotiation partner. For example, Bazerman, Magliozzi and Neale (1985) manipulated framing in a free-market negotiation context, where participants could engage in multiple negotiation transactions, by instructing half of the participants to maximize their profits (gain frame) and the other half to minimize expenses (loss frame). Their results indicated that individuals in the gain frame completed more negotiations compared to those in the loss frame, suggesting that the framing context affected negotiation outcomes.

Another relevant cue in negotiation is the emotional state of the negotiator. Carnevale (2008) investigated the relationship of both negotiation frame outcome and felt affect. He manipulated affect using a mood induction paradigm that involved giving participants a small, clear plastic bag full of chocolates. Results from Carnevale's (2008) study suggested that participants demanded more in the loss frame than in the gain frame when they were not given candy. Conversely, in the positive affect condition, the opposite occurred such that individuals in the gain frame demanded more in the negotiation.

Carnevale's work suggested that positive affect led to a reference point shift to the right of the origin in the classic Prospect Theory value function, and this induced a downsizing of loss differences and an upsizing of gain differences. This explains why individuals in a gain frame demanded more relative to those in the loss frame. Inducing positive affect caused individuals to behave in the opposite direction in the loss frame; they demanded less in the negotiation task than those in the control condition, who did not experience the mood induction paradigm.

In addition to felt affect playing a role in negotiation, emotional displays by a negotiation partner can conceivably affect the goals and related emotions of a negotiator. It is

expected that two parties will display emotions when negotiating over a set of issues, whether their goal is to obtain the best outcome for themselves or arrive at an integrative solution that maximizes the joint gain. Effective negotiation involves not only expressing appropriate emotional responses, but also understanding the effects of others' emotions on the self (de Melo, Carnevale, Read, & Gratch, 2014; Khooshabeh et al., 2013). Emotional appraisals can be inferred from behavior during a task or by asking participants using subjective questionnaire survey instruments. It is also plausible to index the effect of social cues, such as emotions, on negotiators at an implicit level by using advanced neurophysiological measures.

Inferring Psychological States from More than Just Behavioral Performance Measures

Many researchers have demonstrated the role of emotions as social information in motivated performance tasks such as negotiations (Choi, de Melo, Khooshabeh, Woo, & Gratch, 2015; van Kleef, De Dreu, & Manstead, 2004). In addition to behavioral performance during a negotiation task, other measures can index the social effects of emotion, such as psychologically driven neurophysiological states. The bio-psychosocial (BPS) model (Blascovich & Mendes, 2010) is a theoretical account of how a complex, multivariate pattern of cardiovascular responses indicates states of task engagement as well as motivational polarity; i.e., "challenge" and "threat." A challenge state indexes when an individual appraises her resources as exceeding demands in a situation, thereby having positive motivation and higher coping potential, whereas threat indicates when an individual appraises demands in a situation as exceeding resources.

Briefly, the BPS model is based on the neuroendocrine underpinnings (i.e., Dienstbier, 1989) of cardiovascular responses involving the sympathetic-adrenal-medullary (SAM) and hypothalamic pituitary-adrenal-cortical (HPA) axes. Both challenge and threat states involve the activation of the SAM axis, while only the threat state involves both axes. Accordingly, activation of common SAM axis neural and adrenal medullary endocrine processes affect cardiovascular responses underlying both challenge and threat, including increased heart rate (HR) and increased ventricular contractility (VC; i.e., decreased pre-ejection period or "PEP"), both of which index task engagement. Cardiac output (CO) and total peripheral resistance (TPR) patterns differ depending on motivational state. A challenge state results in decreased TPR and an increase in CO, whereas a threat state leads to little or no change or a decrease in CO and little or no change or an increase in TPR (Blascovich & Mendes, 2010).

Psychologically, challenge motivation occurs when an individual's consciously and/or unconsciously evaluated resources outweigh consciously and/or unconsciously evaluated task demands. Threat occurs when resources are evaluated as not meeting task demands. For example, situations that involve uncertainty, which can result from unexpected social cues, have been found to implicitly increase demand evaluations relative to resources (Mendes, Blascovich, Hunter, Lickel, & Jost, 2007). The BPS model of challenge and threat gives insights into appraisals of resource and demand ratios through the lens of a complex set of cardiovascular responses.

The BPS model also gauges the effect of outcome framing when there is no other social counterpart in a decision making context. In a study by Seery, Weisbuch and Blascovich (2009), participants were given a cognitive test. Half the participants were told that they could make as much as five dollars and would receive a monetary reward of $0.50 for each item they answered correctly. The other half of the participants was told that they would begin the task with five dollars and would incur a penalty of $0.50 for each item they did not answer correctly. Expected values in both conditions were the same; however one outcome was framed as a gain and the other as a loss. Their results indicated that participants in the gain frame experienced challenge (decreased TPR), while those in the loss frame experienced threat (increased TPR). This suggests that the BPS model has been validated to detect psychological states that correspond to decision frames. However, it remains to be seen whether psychophysiological states in this unitary cognitive task generalize to dyadic decision making tasks such as negotiation.

Hypotheses

Loss framing will lead to greater demand Based on Prospect Theory, we predicted that individuals in the loss frame condition would be motivated to minimize their expenses compared to those in the gain frame because losses loom larger. Accordingly, those in the loss frame are predicted to make fewer concessions (i.e., greater demands).

Affective Facial Expressions If positive facial expressions from a negotiation opponent have the same effect on observers as a positive mood-induction, then individuals in the happy/gain condition should make more demands in the negotiation based on Carnevale (2008). As Carnevale (2008) noted, "An interesting follow-up experiment may be to examine reference dependence in expected negotiation outcomes driven by perceived negative and positive affect of the adversary. (p. 60)". The connection between mood-induction using chocolate versus a positive facial expression is potentially tenuous, but this warrants an empirical study of whether positive facial expressions simply induce positive moods. For this study, affect was manipulated via the virtual agent who displayed either angry, happy, or neutral facial expressions (control), which were previously validated (Khooshabeh et al., 2013). We included an angry expressive

virtual agent as an exploratory investigation on the effects of negatively valenced interpersonal affect.

Framing and Motivational State According to Seery et al. (2009), framing outcomes in terms of a potential for loss led to lower CO and greater TPR (consistent with relative threat, which is more negatively valenced than challenge). But it is possible that the mechanism by which gain/loss framing worked in the context of their singular, cognitive motivated performance task would be different in situations that involve dyadic negotiation as a function of affect information transmitted by one partner or the other. In particular, given that losses loom larger, individuals should be more positively motivated (i.e., challenged) to overcome expenses (as in the loss frame) than maximize profits (as in the gain frame). Moreover, the frame effect can be interpreted in terms of risk in the negotiation. Negotiators are generally risk tolerant in a loss frame (Carnevale, 2008), thereby making greater demands (i.e., fewer concessions) and risking non-agreement; the opposite takes place in a gain frame, such that negotiators are risk averse, hence making more concessions to reach agreement. Based on this *risk framing hypothesis*, we predict that loss framing should cause relative challenge compared to gain framing. Similarly, gain framing should cause greater relative threat compared to loss framing.

Methods

Participants

Participants were 162 undergraduate students (90 males, 72 females) at the local university, ranging in age from 18 to 24 (M = 19.2, SD = 1.22) and were granted course credit for their participation.

Design

Participants were instructed to maximize points or minimize loss. Specifically, their role was to sell home appliances by negotiating multiple issues, which included delivery time, discount terms, and financing terms. The participants used the mouse to make offers to the interactive virtual agent (buyer) which displayed either an angry, neutral or happy facial expression after participants made an offer on rounds 1, 3 and 5. The virtual buyer conceded two levels on one of the issues per round, which is deemed to be moderate in prior research (van Kleef et al., 2004); the task lasted six rounds where the participants would make counter-offers to the virtual agent's initial and ensuing offer. Participants could reach an agreement with the virtual agent by either accepting the virtual agent's offer or lowering their offer to match a previous offer from the virtual agent.

In order to analyze the full effect of the virtual human's facial expressions across the whole task, an exclusion criterion was applied to remove participants who agreed to the virtual agent's offer before the sixth round. The reason for doing that was because those participants would not have experienced the entirety of the virtual agent's emotional facial displays. The study used a 2 (gain frame/loss frame) x 3 (angry/happy/neutral) factorial between-subjects design. Participants were randomly assigned to conditions via computer program based on the "round robin" technique.

Frame Participants were randomly assigned to complete the task with instructions that effectively put them in a loss or gain frame context. They were informed that 8000 points represented the highest profit they could obtain. In the gain frame, they were informed that the worst negotiation outcome was zero points on each of three issues and the best outcome was the maximum 8000 points. They were instructed to maximize profits. In contrast, loss frame participants were told to minimize their expenses. They were told that -8000 represented the highest expenses, and hence, the worst negotiation outcome. An expense of zero points on each issue represented the best negotiation outcome in the loss frame condition (see Table 1).

An example will help illustrate how the same negotiation outcomes would result in a mathematically equivalent number of points in each frame. For example, a participant might arrive at a negotiation outcome of level 3 for the issue of delivery time, 5 for discount terms, and 4 for finance terms. In the loss frame, this corresponds to expenses of -4900 (-1200 – 1200 – 2500 = -4900). The -4900 expense would cut into a total possible profit of 8000, so the resulting profit would be 3100 points (8000 – 4900 = 3100). The same level 3-5-4 for the three issues would result in 3100 points in the gain frame (400 + 1200 + 1500 = 3100).

Participants and the virtual agent could vary each of their offers on each of the three issues from levels of 1 to 9, with 9-9-9 representing the optimal profit (see Table 1); the negotiation behavior of the virtual agent was fixed such that the initial offer was 1-2-1 and he conceded two levels for one of the issues on each round (see Khooshabeh et al, 2013). The negotiation task interface reflected in real time the number of points the participant earned or lost after each offer that the virtual agent made and with each counter-offer the participant provided.

Table 1. Expenses incurred at each level in the loss frame; (Parentheses: Profit at each level in the gain frame).

Level	Delivery Time	Discount Terms	Financing Terms
1	-1600 (0)	-2400 (0)	-4000 (0)
2	-1400 (200)	-2100 (300)	-3500 (500)
3	-1200 (400)	-1800 (600)	-3000 (1000)
4	-1000 (600)	-1500 (900)	-2500 (1500)
5	-800 (800)	-1200 (1200)	-2000 (2000)
6	-600 (1000)	-900 (1500)	-1500 (2500)
7	-400 (1200)	-600 (1800)	-1000 (3000)
8	-200 (1400)	-300 (2100)	-500 (3500)
9	0 (1600)	0 (2400)	0 (4000)

Physiological responses Participants' cardiovascular responses were measured during the task using a BIOPAC MP150 system and related amplifiers (Goleta, CA). Electrocardiographic (ECG) signals were recorded using an electrocardiograph amplifier (Model ECG100C). Blood pressure was recorded via the continuous non-invasive arterial pressure (CNAP) monitor (Model NIPB100F). Impedance cardiography was collected using a tetra-polar aluminum/mylar tape electrode system (Model NICO100C) secured around the participant's neck and torso and an impedance cardiograph.

Procedure

Participants were brought to the lab. Prior to beginning the experiment they completed an informed consent form. A female research assistant then placed impedance tape and the other electrodes on the participant's neck and torso followed by a blood pressure cuff on the brachial artery of the participant's upper arm (for calibration purposes only) and an additional cuff on the radial artery of the participant's index finger. After recording a five-minute resting baseline of physiological data, the participant was instructed to begin the negotiation task on a computer provided by the lab.

The participants first read instructions that described the scenario and task at hand and then were asked four follow-up questions to ascertain that they understood the instructions. As an added incentive to motivate successful performance during the task, participants were told that they had a chance of winning the distinction of "top negotiator" by both (1) reaching an agreement with the virtual agent negotiation partner and (2) earning the highest profit. They first engaged in a practice round to familiarize themselves with the negotiation task interface before beginning the negotiation task. Unbeknownst to participants beforehand, the task ran for 6 rounds and the virtual agent's offers were preset. Upon completion of the negotiation task, participants were asked to complete survey questionnaires and then were debriefed and thanked for their participation.

Results

Negotiation Performance

First, all the demand scores were normalized so that they were on the same scale ranging from 0-8000 possible points available in the negotiation task. The scores in the gain frame were already in this format; we normalized the scores of participants in the loss frame by adding a value of 8000 to them so that they would be in the same range as those in the gain frame. A 3 (Emotion: Angry, Happy, Neutral) x 2 (Frame: Gain vs. Loss) ANOVA was conducted on the demand scores.

We computed average demand over the six rounds of the negotiation. There was no effect of emotion, F (2, 156) = 1.48, p = .23, η_p^2 = .019. There was a significant effect of frame, F (1, 156) = 10.9, p = .001, η_p^2 = .07. Participants in the loss condition demanded more on average (M = 5872, SE = 249) compared to those in the gain condition (M = 4695, SE = 256). This supports the *risk framing hypothesis* that predicted loss framing would motivate participants to demand more in the negotiation. There was no significant interaction of the emotion and frame variables, F (2, 156) = 1.52, p = .22, η_p^2 = .019.

Another measure of negotiation performance was the amount by which the initial offer differed from the final offer in the negotiation. This demand difference (concession) score was computed by subtracting the number of points demanded in the last round from those demanded in the first round. The ANOVA on the demand difference score indicated that there was no significant effect of emotion (F < 1) and no effect of framing (F < 1), but there was a marginally significant interaction of emotion and framing, F (2, 156) = 2.85, p = .061. An exclusion criterion was applied to remove participants who agreed to the virtual agent's offer before the sixth round. The reason for doing that was because those participants would not have experienced the entirety of the virtual agent's emotional facial displays. The interaction emotion and framing was significant based on this exclusion criteria, F (2, 147) = 3.35, p = .038, η_p^2 = .044. Demand difference in the gain frame, anger emotion condition (M = 1300, SE = 290) was greater than the demand difference in the gain frame, neutral emotion condition (M = 245, SE = 309), F (2, 147) = 3.16, p = .045, η_p^2 = .041 (Figure 1). This suggests that participants conceded more to the angry virtual agent in the gain condition, which replicates previous results (van Kleef et al., 2004). However, in the loss condition, the virtual agent's emotional facial display did not affect the demand difference (Figure 1). Therefore, the gain/loss context mediates the effect of anger emotional facial expressions.

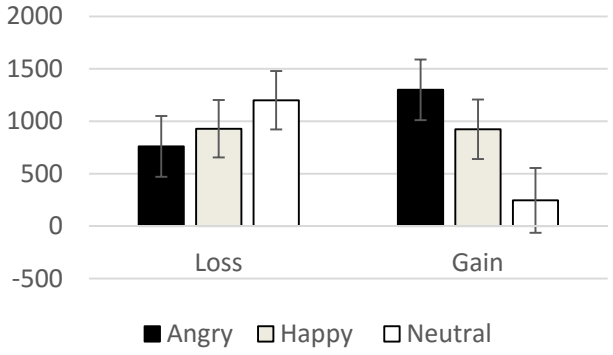


Figure 1. Concession computed by demand difference (Round 1 offer - Round 6 offer). Greater values indicate greater concession. Standard error bars.

Psychophysiological States

The behavioral results from the negotiation task replicated previous work by Carnevale (2008). He found that average

demand over the repeated rounds of the negotiation was greater in the loss frame condition than in the gain frame condition. This suggests that participants were more motivated to minimize their expenses in the loss condition compared to maximizing their profits in the gain condition. Based on these results and the *risk framing hypothesis*, we predicted that the negotiators' neurophysiological state would indicate challenge motivation in the loss frame.

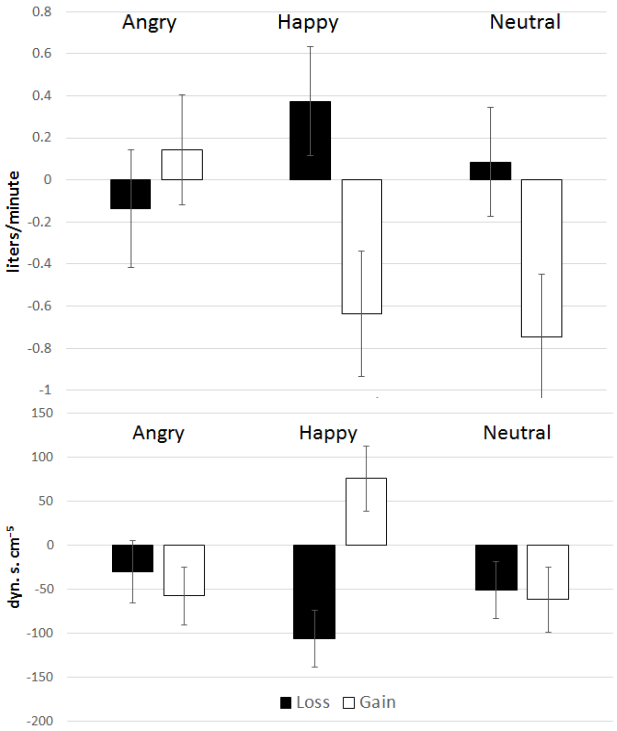


Figure 2. Cardiovascular responses. Top: Cardiac Output (CO); Bottom: Total Peripheral Resistance (TPR)

Physiological reactance scores were calculated by subtracting the CO and TPR measures of last minute of baseline from those in the first minute of the task.[1] A multivariate ANOVA was conducted on these measures with the independent variables of Frame (gain vs. loss) and Emotion (angry, happy, and neutral). The multivariate tests indicated a significant effect of Frame, F (2, 106) = 3.6, p = .03, η_p^2 = .063, and a significant Frame by Emotion interaction, F (4, 214) = 4.0, p = .004, η_p^2 = .07. The effect of Emotion was not significant, F < 1.

Univariate tests indicated that the effect of Frame was significant for the CO reactivity measure, F (1, 107) = 5.32, p = .023, η_p^2 = .047. In the loss condition, CO reactivity was positive (M = .12, SD = .88), which indicated that individuals in the loss frame were challenged. In the gain condition, CO reactivity was negative (M = -.36, SD = 1.5), which indicated that individuals in the gain frame were threatened.

Univariate tests of the Frame by Emotion interaction indicated that the effect was significant for the CO reactivity measure, F (2, 107) = 4.64, p = .042, η_p^2 = .06, and the TPR reactivity measure, F (2, 107) = 5.6, p = .005, η_p^2 = .095. Bonferroni corrected simple-effects analyses were conducted for each measure. The results suggested that neurophysiological reactance to gain and loss frames differed only in the happy emotion condition, F (2, 106) = 8.75, p < .001, η_p^2 = .142. Increased CO reactivity (M = .374, SD = .996) and decreased TPR (M = -106, SD = 83.4) in the loss/happy condition indicated a challenge motivational state. In contrast, CO reactivity decreased (M = -.634, SD = .889) and TPR increased (M = 75.4, SD = 205) in the gain/happy condition, which indicated a threat motivational state (Figure 2).

Discussion

In summary, the results here indicate both conscious and non-conscious reactions to contextual cues such as framing. The behavioral performance during the negotiation task suggests that participants in the loss frame were more demanding relative to those in the gain frame. Analysis of neurophysiological states also supports this behavioral finding that those in the loss frame were more motivated to profit from the negotiation. Cardiovascular measures indicate that individuals in the loss frame exhibited a pattern of physiological responses indicative of a challenge motivational state.

Seery et al. (2009) did not report whether framing affected task performance on the cognitive test whereas our study showed a similar pattern across both the cardiovascular responses and the behavioral performance in the negotiation task. In particular, the results in this study suggest that individuals in a loss frame are not only in a challenge motivational state, but also demand more in the negotiation. These results provide a novel contribution to Prospect Theory, namely that peripheral neurophysiological measures can index when individuals are more motivated to take greater risks in order to avoid losses, because "*Losses loom larger than gains* (Kahneman & Tversky, 1979, p. 279)." Cardiovascular physiology is a nonconscious autonomic process, but negotiation behavior is under conscious control. The fact that two drastically different types of data produced congruent findings serves to strengthen our results.

A positive emotional facial expression displayed by the embodied virtual agent did not seem to have the same effect as a positive mood induction (Carnevale, 2008). There are at least two possible factors that could account for this. First, interpersonal emotional expressions can have complex effects on observers beyond mere emotional contagion as

[1] Due to equipment failure, physiological data for 47 participants was lost.

identical facial expressions can lead to different interpretations based on context (Barrett, Mesquita, & Gendron, 2011; Choi et al., 2015; Khooshabeh et al., 2013; Szczurek, Monin, & Gross, 2012). Therefore, it is possible that the appraisal of the embodied agent's smile in the loss frame condition could be different from the appraisal in the gain frame condition due to changes in context. The second factor that might account for why a positive facial expression did not have the same effect as Carnevale's mood induction paradigm could be due to slight differences in the experimental protocol. Embodied virtual agents in this study displayed emotional facial expressions at a few distinct times throughout the negotiation. Conversely, the positive mood induction in Carnevale's study was only done before and after the instructions, but not during the actual negotiation. This difference in the timing of the affect might account for the different outcomes in framing compared to Carnevale (2008).

Results can help generate design guidelines for virtual agent developers to create negotiation scenarios that convey a gain frame if the goal is to make negotiators demand less from a negotiation agent. For the design of training systems, the goal might be to motivate negotiators more positively when they are learning effective negotiation. To accomplish that, negotiation tutors could engender a neurophysiological state of positive motivation (i.e., challenge).

References

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in Emotion Perception. *Current Directions in Psychological Science*, *20*, 286–290. http://doi.org/10.1177/0963721411422522

Bazerman, M. H., Magliozzi, T., & Neale, M. A. (1985). Integrative bargaining in a competitive market. *Organizational Behavior and Human Decision Processes*, *35*(3), 294–313. http://doi.org/10.1016/0749-5978(85)90026-3

Blascovich, J., & Mendes, W. B. (2010). Social psychophysiology and embodiment. In *Handbook of Social Psychology* (5th Edition, pp. 194–227). New York: Wiley.

Carnevale, P. J. (2008). Positive affect and decision frame in negotiation. *Group Decision and Negotiation*, *17*(1), 51–63. http://doi.org/10.1007/s10726-007-9090-x

Choi, A., de Melo, C. M., Khooshabeh, P., Woo, W., & Gratch, J. (2015). Physiological evidence for a dual process model of the social effects of emotion in computers. *International Journal of Human-Computer Studies*, *74*, 41–53. http://doi.org/10.1016/j.ijhcs.2014.10.006

de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, *106*(1), 73–88. http://doi.org/10.1037/a0034251

Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for physical and mental health. *Psychological Review*, *96*(1), 84–100.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Khooshabeh, P., de Melo, C. M., Volkman, C. B., Gratch, J., Blascovich, J. J., & Carnevale, P. (2013). Negotiation strategies with incongruent facial expressions of emotion cause cardiovascular threat. In *Cognitive Science Society* (pp. 770–775). Berlin, Germany.

Mendes, W. B., Blascovich, J., Hunter, S. B., Lickel, B., & Jost, J. T. (2007). Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners. *Journal of Personality and Social Psychology*, *92*(4), 698–716. http://doi.org/10.1037/0022-3514.92.4.698

Seery, M. D., Weisbuch, M., & Blascovich, J. (2009). Something to gain, something to lose: The cardiovascular consequences of outcome framing. *International Journal of Psychophysiology*, *73*(3), 308–312. http://doi.org/10.1016/j.ijpsycho.2009.05.006

Szczurek, L., Monin, B., & Gross, J. J. (2012). The Stranger Effect: The Rejection of Affective Deviants. *Psychological Science*, *23*(10), 1105–1111. http://doi.org/10.1177/0956797612445314

van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, *86*(1), 57–76.

Generalization of within-category feature correlations

Nolan Conaway & Kenneth J. Kurtz
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

Abstract

Theoretical and empirical work in the field of classification learning is centered on a 'reference point' view, where learners are thought to represent categories in terms of stored points in psychological space (e.g., prototypes, exemplars, clusters). Reference point representations fully specify how regions of psychological space are associated with class labels, but they do not contain information about how features relate to one another (within- class or otherwise). We present a novel experiment suggesting human learners acquire knowledge of within-class feature correlations and use this knowledge during generalization. Our methods conform strictly to the traditional artificial classification learning paradigm, and our results cannot be explained by any prominent reference point model (i.e., GCM, ALCOVE). An alternative to the reference point framework (DIVA) provides a strong account of the observed performance. We additionally describe preliminary work on a novel discriminative clustering model that also explains our results.

Keywords: categorization; generalization; formal modeling

Introduction

Research on human classification learning is fundamentally interested in questions of representation: How do people represent categories? How does a category's internal structure influence its subjective difficulty? How do people generalize their knowledge about categories? Current research addressing these questions is centered around a 'reference-point' framework, whereby people are thought to acquire category knowledge associating stored perceptual referents (e.g., prototypes, exemplars) with individual categories. The success of reference point models (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1984; J. D. Smith & Minda, 2000) is unparalleled within the field, and as a result these models are widely considered to be definitive accounts of how categories are learned and represented (for reviews, see Murphy, 2002; Pothos & Wills, 2011; see also Kurtz, 2015).

Although reference point models differ from one another in a variety of ways, these models are comparable in that they assume that categories are represented by one or more points in a psychological space. On the extremes, prototype models represent categories with a central tendency (i.e., the average across known members), whereas exemplar models use specific observations. Many successful reference point models employ a selective attentional mechanism, enabling them to weight the importance of each stimulus dimension (Kruschke, 1992; Medin & Schaffer, 1978).

Importantly, however, reference point representations do not incorporate all aspects of class structure. While points of reference can be used to encode information about how regions of space are associated with known categories, they do not contain information about how features relate to one another (either globally or within a class). By consequence, reference point models are only sensitive to correlations between features insofar as those correlations are reflected in the distances between stored reference points.

Although there is evidence that people make use of feature correlations in natural concepts (Malt & Smith, 1984), research on correlation learning in a traditional artificial classification learning (TACL) setting has been mixed. Whereas Medin, Altom, Edelson, and Freko (1982) reported evidence that feature correlations influence classification of artificial categories, Murphy and Wisniewski (1989) later expanded upon that study and found little evidence of correlation learning, unless features are expected to be correlated. Finally, Anderson and Fincham (1996) reported that participants used correlations to infer values of missing features; though note that traditional reference point models are unable to simulate feature inference (Lee & Navarro, 2002).

In this paper, we report a classification learning experiment demonstrating that people represent correlations between features, beyond what can be explained in terms of stored reference points. In addition to providing evidence that learners do acquire knowledge about correlations between features, the classification performance we report demonstrates a systematic failure in the reference point framework. We bolster our empirical results with simulations using the Generalized Context Model (GCM; Nosofsky, 1984), an exemplar model that embodies the central tenets of the reference point view.

We also report simulations using the *DIV*ergent *A*utoencoder model (DIVA; Kurtz, 2007, 2015), a autoassociative network model that stands as a similarity-based alternative to the reference point framework. The DIVA model is fully instantiated as a connectionist network: as in traditional multilayer perceptron (MLP) architectures, DIVA is initialized with a input units encoding feature values, as well as a collection of hidden units enabling the learning of an internal representation (Rumelhart, Hinton, & Williams, 1986). DIVA's primary point of departure from these models lies in its learning objective: instead of learning representations to predict class responses, the DIVA model learns auto-associatively to predict feature values along divergent, category-specific output channels. Thus, DIVA's category representations are acquired for the purposes of making *feature* predictions rather than *class* predictions. Classification decisions are made using a secondary response rule: the probability of any given classification depends on the relative amount of feature prediction error (reconstruction error) across all categories, with better reconstructions leading to increased probability.

DIVA's design principles offer a unique account of human

category learning: rather than assuming people learn to predict class responses through association to stored points of reference, DIVA proposes that people learn representations of the observed regularities within each class. Accordingly, with regard to learning feature correlations, DIVA's predictions sharply contrast from those made by reference point models. Specifically, because DIVA is trained on feature prediction (rather than class prediction), it strongly relies on within-class feature correlations to aid learning. Thus, whereas reference point models do not encode any information about feature correlations, DIVA relies on those correlations in the service of minimizing feature prediction error.

Finally, we conclude our report with preliminary work on a novel discriminative clustering model that explains our results without acquiring knowledge of how features are correlated. Development of this model is ongoing, though our simulations results indicate that it may succeed as an account of human classification more generally.

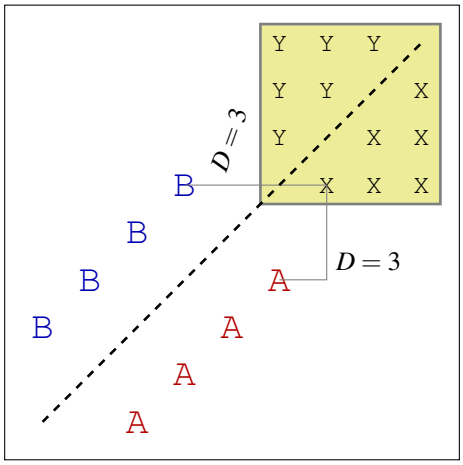


Figure 1: The Diagonal classification. X and Y indicate critical generalization items. Annotations illustrate equal city-block distance between the critical items and known category members.

The Current Study

In a *Diagonal* classification, the two categories ('A' and 'B') are organized along a diagonal boundary (see Figure 1). This classification is notable for its likeness to Information-Integration (Ashby & Maddox, 1990) and Condensation categories (Gottwald & Garner, 1972; Kruschke, 1993). The features are perfectly correlated within each category, but the training exemplars are isolated in one region of the stimulus space, allowing for generalization items that follow the diagonal boundary (labeled X & Y in Figure 1).

Our DIVA simulations revealed that the model typically extends the diagonal boundary to these items – exemplars on the Category A side of the extended boundary are more likely to be classified as members of Category A than exemplars on the other side, and vice-versa (i.e., $A \rightarrow X$, $B \rightarrow Y$). DIVA's performance is concisely explained by its design principles: because DIVA is trained on feature prediction (rather than class prediction), the model learns that the features can be used to predict one another. After training, DIVA reconstructs novel items following its knowledge of each category's feature correlations, affording an extension of the diagonal boundary.

We found that the GCM could not mimic DIVA's performance. Instead of extending the diagonal boundary outward, the GCM classifies each critical item with equal probability. A close examination of the Diagonal classification explains the GCM's performance: each of the critical items is equally distant (under a city-block metric) to known members of both categories. While a Euclidean metric would enable the GCM to extend each category to its critical items, the use of a city-block metric reflects the separable stimulus dimensions used in the behavioral experiment reported below (see Garner, 1974). City-block distance was also supported by the results of an independent pairwise similarity-rating study – stimuli X & Y were not rated as more similar to items on their own side of the diagonal boundary.

With evenly distributed selective attention, the GCM classifies each item with a probability of 0.5. Unequal allocation of attention results in uniform changes to the classification gradient, but not generalization of each category along the diagonal. For example, greater allocation to the vertical dimension results in increased Category B probability for the entire collection of critical items. Finally, other types of reference points (i.e., prototypes, clusters) also lead to the same performance. The critical items area are equally close to the category prototypes, as well as a variety of cluster configurations. The GCM's results therefore characterize a set of predictions made by reference point models more generally.

Examining these predictions more methodically, we conducted a 'grid-search' to evaluate DIVA and GCM performance under a range of parameter settings. At each point in the search, the models were tested on classification of novel items X & Y. To quantify each parameterization's degree of diagonal extension, we calculated the difference score in the Category A classification probability for the critical items on either side of the boundary ($X - Y$). Positive difference scores indicate systematic generalization of each class, and neutral (≈ 0) scores indicate uniform generalization. Plotting these scores as a density curve across points in the search (Figure 2) reveals strongly systematic behavior: whereas DIVA nearly always generalized each class outward, the GCM produced identical responses to X & Y under every parameter setting.

In what follows, we report a behavioral study testing the predictions made by DIVA and the GCM on generalization of the Diagonal classification. If human learners represent categories solely in terms of reference points, then generalization should be uniform: participants should be no more likely to produce a Category A response to items on the A-side of the diagonal than on the opposite side (i.e., $X \approx Y$). However, if participants acquire knowledge about how the features relate to one another within each category (as predicted by DIVA),

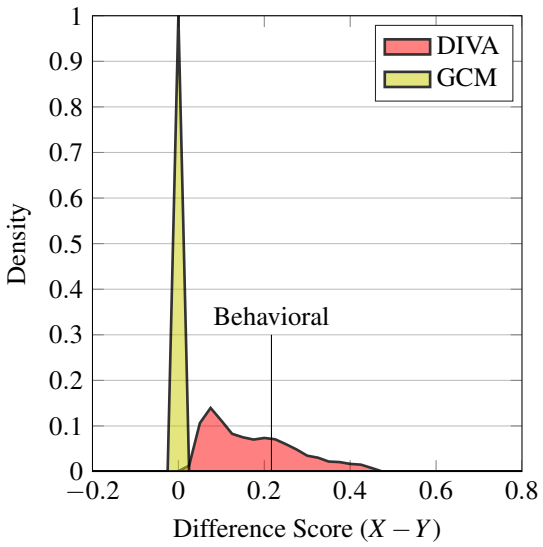


Figure 2: 'Grid search' simulations with DIVA and the GCM. Densities in this plot reflect predictions about the classification of critical items, across many different parameter settings. Positive scores indicate that each category was generalized to the items on its side of the boundary, $A \rightarrow X$, $B \rightarrow Y$.

then we should observe systematic generalization of the diagonal boundary outward (i.e., $A \rightarrow X$, $B \rightarrow Y$).

Participants and Materials. 30 undergraduates from Binghamton University participated in fulfillment of a course requirement. Stimuli were squares varying in shading and size (see Figure 3). The separability of these features justifies a city-block metric (Garner, 1974). An independent scaling study verified that the dimensions are nearly equal in perceptual salience. The assignment between perceptual and conceptual dimensions was counterbalanced across participants.

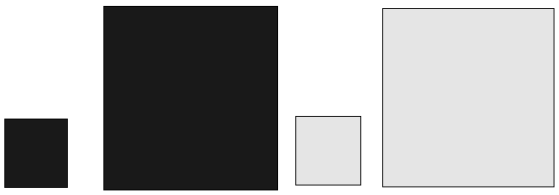


Figure 3: Sample stimuli.

Procedure. Participants completed 96 training trials (12 blocks consisting of the 8 training examples). On each trial, a stimulus was presented on a computer screen and learners were prompted to make a classification decision by clicking one of two buttons (labeled 'Alpha' and 'Beta'). After selecting a class label, learners were given feedback on their response. Following the training phase, participants completed 81 generalization trials consisting of items sampled at 9 positions on each dimension. All of the training examples were included (intermixed). Feedback was not provided during the generalization phase. Participants were informed that there would be test trials prior to beginning the experiment.

Results. By the end of the training phase, most participants had successfully mastered the categories. On average, participants were 89.2% ($SE = 2.5$) accurate during the last training block. Only two of 30 participants failed to reach greater than 6/8 correct during the final training block. Aggregate training data is depicted in the left panel of Figure 4.

As a test of whether learners extended the diagonal boundary, we compared the average number of 'A' classification responses made to critical items X & Y (Figure 4). Learners were more likely to produce an 'A' response to X than to Y, $t(29) = 5.02$, $p < 0.001$, $d = 0.56$. We then compared the difference score we obtained behaviorally ($X - Y$) to our earlier results with DIVA and the GCM: as shown in Figure 2, the difference score we observed cannot be produced by the GCM, but is fully explained by DIVA. Aggregate generalization data is depicted in Figure 5.

Summary

We provided classification training to human learners on a Diagonal classification (Figure 1). Learners systematically extended each category to novel items that are equidistant to known exemplars from both categories. The generalization we observed can be considered evidence that learners acquire knowledge about feature correlations, and they apply that knowledge during generalization. Our results are also inconsistent with the notion that category knowledge solely consists of exemplar, prototype, or cluster reference points.

Our results are concisely explained by the DIVA model. Because DIVA's is principally autoassociative, the model relies on within-class internal regularities (such as feature correlations) to support reconstruction learning. In our above simulations, we found that the model frequently generalized the diagonal boundary outward, just was we observed in the behavioral study. However, it is customary to assess models using a post-hoc parameter fitting process. In the next section, we describe a more formal examination of the performance by DIVA and the GCM.

Simulations

The overall goal of the following simulations is to formally evaluate the performance of DIVA and the GCM in terms of quantitative fit to our observed generalization behavior. Before proceeding, it is worth noting that DIVA and the GCM are not fully comparable: unlike the GCM, DIVA's category representations are constrained via back-propagation learning (Rumelhart et al., 1986), and its performance is stochastic. However, model performance on generalization testing can still be compared to assess whether our results can be explained under a reference point scheme.

We used parameter optimization techniques to find each model's best fit to the observed generalization data using of a mean-squared error (MSE) metric. We used a hill climb-

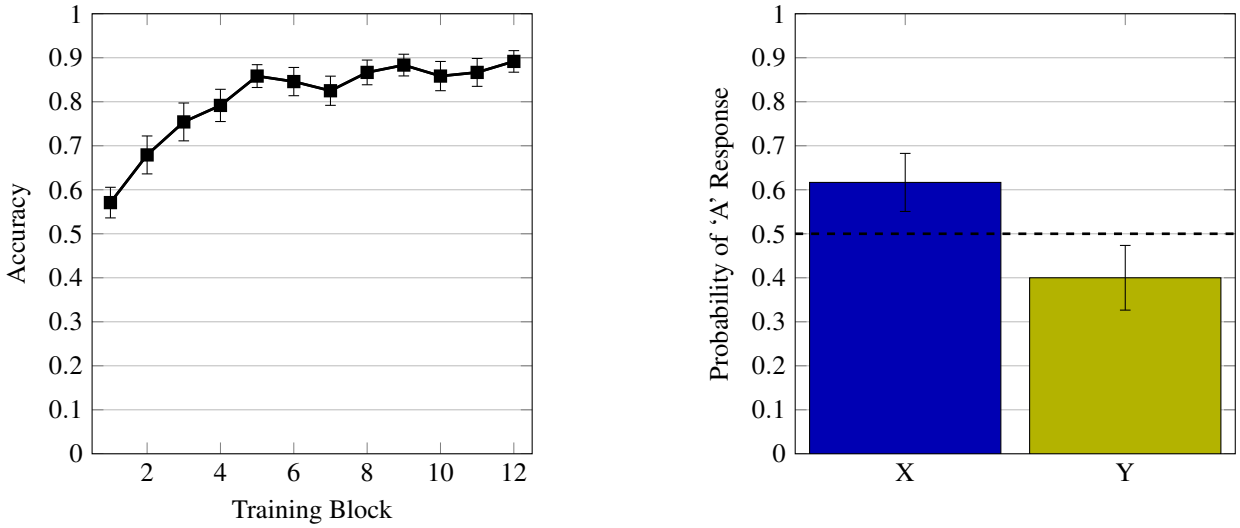


Figure 4: *Left*: Behavioral training accuracy. *Right*: Behavioral responses to critical items *X* & *Y*. Error bars reflect ±1 SE.

ing procedure to fit the GCM over four parameters: exemplar specificity (c), response determinism (γ; Nosofsky & Zaki, 2002), and attention strengths for features 1 and 2 (W_1 and W_2). DIVA's behavior is stochastic, however, which precludes the use of hill-climbing. As a result, we search for DIVA's best fit using a 'grid-search' technique to generate predictions along a range of settings for its four parameters: number of hidden units, learning rate, initial weight range, and a focusing parameter, β (Kurtz, 2015). At each point, DIVA was initialized 2000 times with random small-valued weights and a random presentation sequence. The model used logistic hidden units and linear outputs.

Overall, DIVA was able to provide a stronger fit to the full behavioral gradient than the GCM. DIVA's best fit ($MSE = 0.006$) was achieved with 3 hidden units, learning rate = 0.55, weight range ± 0.5, and $\beta = 3$. The GCM's best fit ($MSE = 0.0075$) was achieved with $c = 5.304$, $\gamma = 1.055$, $W_1 = 0.507$, $W_2 = 0.493$. Model performance was further differentiated the responses to the critical items (rather than the entire gradient): DIVA achieved $MSE = 0.0037$, and the GCM achieved $MSE = 0.02$. Beyond quantitative fitting, however, it is important to acknowledge that DIVA's predictions match the qualitative patterns of observed performance – the model extends each category to novel items along the diagonal boundary. Conversely, the GCM's generalization is completely neutral for these items. Each model's best performance is depicted in Figure 5.

A discriminative clustering account.

Although traditional reference point models are unable to explain our results, we have recently implemented a novel discriminative clustering account that successfully captures the observed generalization performance. The advance made by this account lies in the realization that the reference points associated with each category need not be localized as the exemplars, the category's central tendency, or the central tendency of select clusters of exemplars. Instead, the location of reference points may discriminatively reflect the proximity of opposite-category members. In doing so, this model may provide a account of the effects of contrast categories on conceptual representation (Davis & Love, 2010; Levering & Kurtz, 2006; Palmeri & Nosofsky, 2001).

Overall, the design of the discriminative clustering model is similar to SUSTAIN (Love et al., 2004): the model's reference point representation consists of a collection of clusters, and classification is based on each category's association the clusters. The model begins training with no internal representation, and recruits clusters in when it makes a poor classification decision. The primary departure from SUSTAIN concerns the localization of the clusters: on each trial, clusters belonging to the correct category are moved toward the presented exemplar, and clusters belonging to incorrect categories are move away from the presented exemplar. By allowing the clusters to move discriminatively from members of the opposite category, they become more similar to critical items on the category's side of the diagonal boundary, producing the observed pattern of generalization. Sample performance from this model is depicted in Figure 6.

Discussion

The reference point framework has dominated research and theory in category learning for the past 30 years. Reference point representations are useful in that they specify how regions of space are associated with class labels, but they do not encode information about global or within-class regularities, such as feature correlations.

We reported an experiment suggesting that human learners acquire knowledge of within-class feature correlations. Specifically, after training on a Diagonal classification, participants systematically generalized in accord with each cat-

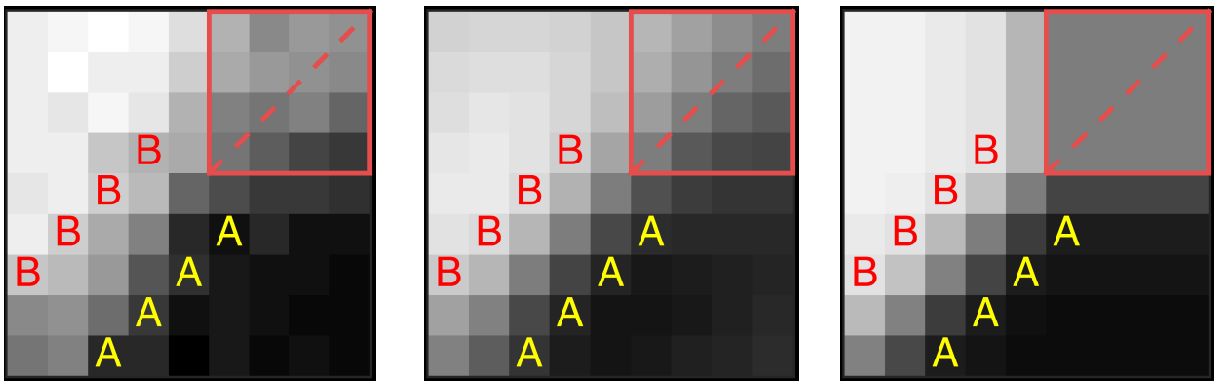


Figure 5: *Left*: Behavioral generalization. *Center & Right*: Best-fit predictions from DIVA and the GCM.

egory's internal structure to exemplars that are equally similar to members of both categories. The observed performance is inconsistent with prominent reference point models (e.g., Kruschke, 1992; Love et al., 2004; Nosofsky, 1984; J. D. Smith & Minda, 2000), and indicates that learners may acquire knowledge about categories that cannot be represented under a reference point scheme. The *Div*ergent *A*utoencoder model (DIVA; Kurtz, 2007, 2015) succinctly explains the observed performance: because DIVA is chiefly an autoassociator, the model depends on within-class regularities (such as feature correlations) to aid in feature prediction. Accordingly, DIVA's generalization shows strong sensitivity to the within-category feature correlations.

Although existing reference point models fail to match our results, we introduced a novel discriminative clustering model capable of producing the observed generalization. Development of this model is ongoing, though its results here show promise. Unlike existing cluster-based approaches (i.e., Love et al., 2004), the cluster locations in our model are optimized both for similarity to same-category exemplars, and for dissimilarity to opposite-category exemplars. Thus, the model's clusters are gradually moved outward in the stimulus space, taking on the value of a category ideal. From this location, clusters are more similar to novel items on their side of the diagonal boundary, affording generalization of each class. Future work will attempt to dissociate the predictions made by DIVA and the discriminative clustering model.

This report adds to an accumulating body of evidence against the idea that category learners solely acquire knowledge in the form of reference points. Research in function learning (DeLosh, Busemeyer, & McDaniel, 1997) has, for example, demonstrated key flaws in the account of extrapolation put forward by reference point similarity – studies on on rule-based generalization have revealed similar flaws in a category learning context (e.g., Erickson & Kruschke, 2002). Traditional reference point approaches are also unable to explain the effects of category variability on generalization (Cohen, Nosofsky, & Zaki, 2001; E. E. Smith & Sloman, 1994). Finally, our recent work (Conaway & Kurtz, 2015) has uncovered unique generalization behavior that cannot be explained via similarity to reference points. Taken as a whole, these findings raise a substantive challenge to theories of human category learning based on similarity to reference points.

Acknowledgments

We thank Sarah Laszlo and Gregory Murphy for helpful comments on this work. We additionally thank Sarah Laszlo for access to her server. We also thank the members of the Learning and Representation in Cognition (LaRC) lab at Binghamton University, as well as the BU Modeling meeting.

References

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of experimental psychology: Learning, Memory, and Cognition*, *22*(2), 259.

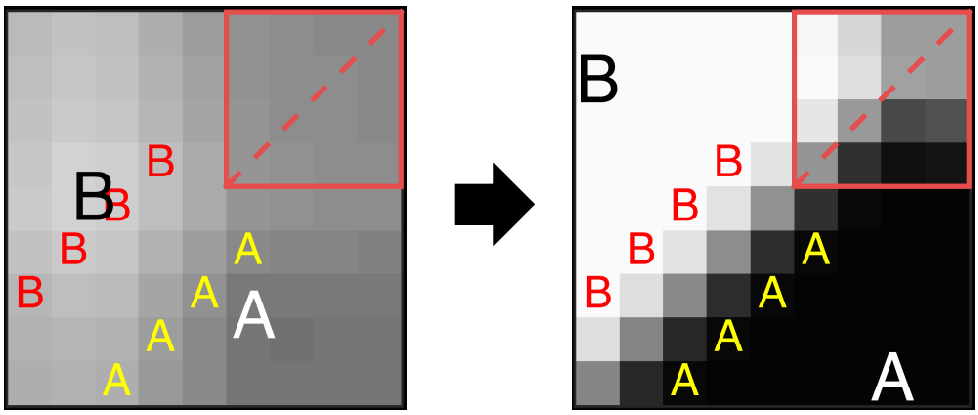
Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 598–612.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175.

Conaway, N. B., & Kurtz, K. J. (2015). A dissociation between categorization and similarity to exemplars. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 435–440). Austin, TX: Cognitive Science Society.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: the sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*, 968–986.



Trial 8 Post-training

Figure 6: Sample performance from the discriminative clustering model. Learned clusters are labeled using large, black and white 'A' and 'B' characters. The model is capable of learning multiple clusters, but a single-cluster solution is typical for this classification. *Left*: generalization at the beginning of Block #2. *Right*: generalization at the end of the training phase.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*(1), 160–168.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gottwald, R. L., & Garner, W. R. (1972). Effects of focusing strategy on speeded classification with grouping, filtering and condensation tasks. *Perception and Psychophysics*, *11*, 179-182.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin and Review*, *14*, 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin and Review*, *9*, 43–58.

Levering, K., & Kurtz, K. J. (2006). The influence of learning to distinguish categories on graded structure. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1681–1686). Mahwah, NJ: Erlbaum.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Malt, B. C., & Smith, E. E. (1984). Correlated properties in natural categories. *Journal of verbal learning and verbal behavior*, *23*(2), 250–269.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 37.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, *85*, 207–238.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT press.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. *Advances in cognitive science*, *2*, 23–45.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*(1), 104–114.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, memory, and cognition*, *28*(5), 924.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology: Section A*, *54*(1), 197–235.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 3.

Deconstructing *tomorrow*: How children learn the semantics of time

Katharine A. Tillman[1], Tyler Marghetis[2], David Barner[1], & Mahesh Srinivasan[3]

[1]U. California, San Diego
Department of Psychology
{katillman, barner}@ucsd.edu

[2]Indiana University, Bloomington
Psychological & Brain Sciences
tmarghet@indiana.edu

[3]U. California, Berkeley
Department of Psychology
srinivasan@berkeley.edu

Abstract

Deictic time words (e.g., "tomorrow," "yesterday") refer to time periods relative to the present moment. While children produce these words by age 2-3, they use them incorrectly for several more years. Here, as a case study in abstract word learning, we explored what children know about these words during this delay. Specifically, we probed children's knowledge of three aspects of meaning: deictic (past/future) status, sequential ordering (e.g., "tomorrow" is after "yesterday"), and remoteness from *now*. We asked 3- to 8-year-olds to place these words on a timeline extending from the past (left) to the future (right). Even 4-year-olds could meaningfully represent the words' deictic status and order, and by 6, the majority displayed adult-like performance. Adult-like knowledge of remoteness, however, emerged independently, after age 7. Thus, even while children use these terms incorrectly, they are gradually constructing a structured semantic domain, including information about the deictic, sequential, and metric relations among terms.

Keywords: time; word learning; development; abstract concepts; timeline

1. Introduction

When learning a new word, children face an inductive problem: How broadly should this new word be extended? This problem is especially challenging in cases where a word's semantic boundaries cannot be inferred from perception of the world. For instance, *deictic time words* like "yesterday" and "tomorrow", which label periods of time relative to the present, lack perceptible referents and are used to refer to a changing set of experiences. Tomorrow's events, for instance, will soon become yesterday's. To make inferences about word meanings in such cases, children could rely on the structure of language itself—e.g., using linguistic cues to learn how these words are related to one another within a common semantic class—and thus construct partial meanings on the way to adult-like semantics (e.g., Carey, 2009; Gleitman et al., 2005). However, while previous studies indicate a long gap between when children initially produce deictic time words and eventual adult-like usage (Ames, 1946; Busby Grant & Suddendorf, 2011; Harner, 1975; 1981), we know relatively little about the inductive process through which they are acquired. Here, as a case study of abstract word learning, we explore children's gradual construction of deictic time word meanings between ages 3 and 8.

Although many children use words like "yesterday" and "tomorrow" as early as age 2 or 3, they make frequent errors in both their use and comprehension for several subsequent years (Ames, 1946; Busby Grant, & Suddendorf, 2011; Harner, 1975). According to parental report, while two thirds of 3-year-olds produce the word "yesterday", fewer than 20% use the word correctly; by age 5, more than 80% of children produce "yesterday", but still, fewer than 60% use it correctly (Busby Grant & Suddendorf, 2011). When asked to name an event that occurred "yesterday" or one that will occur "tomorrow," only about a quarter of 3-year-olds can provide reasonable answers, and less than 70% of 5-year-olds can do so (Busby & Suddendorf, 2005). However, while such studies suggest that children struggle with deictic time words for years, they tell us little about what children *do* and *do not* know about these words, or about how they are learned.

While there has not been systematic study of children's partial knowledge of deictic time words during the delay between initial production and adult-like usage, there are hints that children may independently acquire information about different facets of their meaning. These facets include a word's **deictic status** (e.g., "yesterday" is in the past; "tomorrow" is in the future), its sequential **order** relative to other time words (e.g., "yesterday" is a time after "last week"), and its **remoteness** from the present (e.g., "yesterday" is one day from today). In one study, 3-year-old English speakers appeared to have partial knowledge of the deictic status of "yesterday" and "tomorrow", understanding that these terms refer to a non-present time, without knowing whether they refer specifically to the past and future, respectively (Harner, 1975). Children's spontaneous speech errors also suggest partial meanings. For example, children overextend deictic time words within either the past or future (e.g., Harner, 1981; Nelson, 1996), suggesting that children may acquire deictic status prior to remoteness.

Understanding the nature of children's errors – and the partial word meanings they implicate – could provide critical insight into the inductive hypotheses children make about the meanings of deictic time words and the cues children use to guide them. Here we will consider three potential word learning strategies, which are not mutually exclusive. One possibility is that children make inferences about time word meaning based on grammatical cues in the sentence, i.e., using *syntactic bootstrapping* (Gleitman et al., 2005). In this case, we would expect native-English-speaking children to have early knowledge of **deictic status**, which is expressed by the English tense system. A second possibility is that children use linguistic cues to group terms into a common semantic class (e.g., terms referring to the past) and use lexical contrast to learn their relations within it

("last week" is before "yesterday"; see Carey, 2009; Shatz et al, 2010; Tillman & Barner, 2015) This account predicts early knowledge of **order**. A third possibility is that children use associations between time words and experienced events, i.e., *event mappings,* to estimate the distance from the present indicated by each word, predicting early knowledge of **remoteness**. In order to test these accounts, the present study aimed to separately characterize children's knowledge of these different facets of these terms' meanings. To do so, we asked children to map these words onto a nonverbal representation: a spatial timeline.

Several studies suggest that by age 4 or 5, children can use spatial scales to differentiate the times of events (Busby Grant & Suddendorf, 2009; Friedman, 2002; Friedman & Kemp, 1989; Hudson & Mayhew, 2011). However, no studies have probed children's knowledge using a conventional horizontal timeline, and most have explicitly *avoided* deictic time words. Further, previous timeline studies were limited in their ability to tease apart children's understanding of the different semantic facets of these terms. For instance, they used timelines depicting either the past or the future, but not both, making it impossible to gauge children's knowledge of deictic status, and they were categorical, e.g., "a short time ago" vs. "a long time ago," making it difficult to probe children's knowledge of sequential order or remoteness precisely, because children could place multiple terms within the same category (e.g. Busby Grant & Suddendorf, 2009; Friedman, 2002; Friedman & Kemp, 1998).

Here, we assessed children's comprehension of different facets of the meaning of deictic time words using a new timeline task. We tested adults, and children over a wide age range (age 3 to 8), in order to probe for possible partial knowledge of terms during the long gap between initial production and adult-like usage. Participants used colored pencils to mark where deictic time words (e.g., "yesterday") and events (e.g., the participant's last birthday) should go on horizontal timelines that extended continuously from the past ("when you were a baby") on the left, to the future ("when you'll be a grown-up") on the right. Importantly, the present moment ("right now") was also marked, dividing the line into the past and future.

Critically, this paradigm allowed us to independently assess children's knowledge of a deictic time word's deictic status (i.e., past vs. future), sequential order, and remoteness from the present. Knowledge of the deictic status of a word was indexed by its correct placement to the left or right of the midpoint, regardless of its placement relative to other words. Knowledge of sequential order was indexed by the ordering of words along the line — for example, whether "last week" was placed before (i.e., to the left of) "yesterday," ignoring either term's relation to the present or the distances between them. And knowledge of remoteness from the present was indexed by the spacing of terms along the line — for example, by looking at how far "last week" is placed from "now", compared to the placement of "yesterday" relative to "now."

Our paradigm thus probed children's developing understanding of deictic time words, including the sequence by which they acquire different aspects of mature, adult-like meanings. For example, if a child acquires deictic status prior to the other facets, as predicted by a syntactic bootstrapping account, we would expect them to correctly assign all past terms to the past but perhaps to *fail* to respect their ordering or distance from the present, e.g., by placing "yesterday" much further in the past than "last year".

2. Methods

2.1 Participants

Participants included 109 children between 3;0 and 8;11 years of age, recruited from the San Diego, CA and Berkeley, CA areas. Additionally, we tested 37 adult controls from the UCSD Psychology Department subject pool. Children were tested in lab or at local daycares, schools, and museums, and adults were tested in lab. Informed consent was obtained from adults and parents of participating children. Adults were given course credit; children were given a small prize.

2.2 Materials and procedure

Participants responded by drawing on three 13.5cm left-to-right timelines, printed down the center of a 8.5" × 11" sheet of paper. On each timeline, a vertical tick at the midpoint indicated the present; an icon of a baby indicated the past; an icon of an adult indicated the future (Fig. 1).

To begin, the experimenter [E] gave the paper and some colored pencils to the child. While gesturing to the appropriate sections of the top line, E stated: "Look, this is a timeline. It shows *when* different things happen. The line starts in the past and it goes to the future. So, it goes from when you were a baby all the way to when you're going to be a grown up. And here in the middle is right now. Each time has its own place on the line. You're going to show me *when* different things happen by showing me *where* they go on the line. Look, when you were a baby goes here [E draws a vertical line on the left end point to demonstrate the procedure] and when you are going to be a grown up goes here [E draws a vertical line at right endpoint]. And right now goes here [E draws line at midpoint]. I'm going to give you a pencil, and your job will be to draw an up-and-down line to show me where each thing goes. Ready?" At this point, E introduced the first item, "When [did you] [eat breakfast today]? Think about when you [ate breakfast today]? Draw a line for when you [ate breakfast today]." The child then marked the line.



Figure 1: Timeline used by participants to indicate the relative locations of deictic time words and events.

After all four events were placed on the top timeline, the procedure was repeated for the remaining two timelines, with trials in the form, "Now you're going to show me where [last week] goes. Where does [last week] go? Can you draw a line for [last week]?" Each child placed 8 deictic time words and 4 events on three timelines (Table 1). Participants always received the Events line first but the order of the other two lines was counterbalanced between subjects. For each line, half of the subjects received the items in the order shown in Table 1, and the other half received the reverse order.

Table 1: Target items used in timeline tasks

Timeline	Item 1	Item 2	Item 3	Item 4
Events	Breakfast	Next birthday	Dinner	Last birthday
Time words 1	Last week	Tomorrow	Tonight	This morning
Time words 2	Next week	Next year	Yesterday	Last year

2.3 Coding. For each timeline, the distance in centimeters from the left endpoint of the line to each color-coded mark was measured, as well as its distance, positive (right) or negative (left), from the midpoint of the line

3. Results

To characterize children's knowledge of deictic time words, we undertook the following analyses: First, we assessed comprehension of the deictic status, order, and remoteness of the deictic time words (and, by way of comparison, the life events). Second, we determined the ages of acquisition of these facets of meaning, pinpointing the age at which the majority of children displayed adult-like comprehension. Finally, we calculated the contingencies between adult-like knowledge of these facets of meaning: i.e., the degree to which adult-like knowledge of deictic status, order, and temporal remoteness predicted one another.

3.1 Facets of meaning

3.1.1 Deictic status. For each timeline and subject, we calculated the mean accuracy for all items' placement relative to "now" (e.g., "tomorrow" should be in the future), and then calculated mean deictic status accuracy for each subject and each type of timeline (i.e., Deictic vs. Event). We then analyzed Deictic Status accuracy with a mixed ANOVA, with Time Type (Deictic *vs.* Event) as a within-subjects factor and Age (3 through 8 years old, and adults) as a between-subjects factor.

There was no effect of Time Type, suggesting that children were equally able to represent the past/future status of time words and events. The only effect to approach significance was the main effect of Age, $F(6, 138) = 26.3$, $p << .001$ (Fig. 2). While 3-year-olds performed at chance overall, $t_{14} = .14$, $p > .8$, 4-year-olds were better than chance, $t_{16} = 4.7$, $p << .01$. Mean accuracy improved monotonically from 3- to 7-years-old, $M_3 = .50 < M_4 = .67 < M_5 = .77 < M_6 = .84 < M_7 = .94$ at which point mean performance was no longer distinguishable from adults ($M_{adults} = .94$), $t_{56} = -0.1$, $p > .9$.

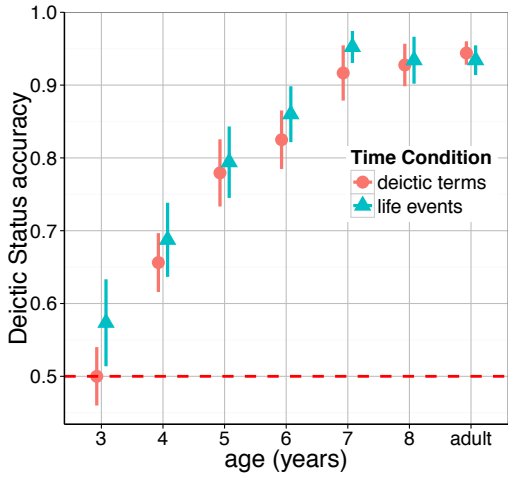


Figure 2: *Deictic Status Knowledge.* Accuracy improved with age, for both time words (e.g., "tomorrow", circles) and events (e.g., "last birthday," triangles). Error lines = SEM; dashed line = chance performance.

3.1.2 Order. We assessed knowledge of relative sequential order—independent of deictic status—in two ways. The first method examined the rank ordering of all four items on a timeline. The second, designed to minimize concerns about working memory demands, evaluated knowledge of sequential order on a trial-by-trial basis, by comparing the placement of each item to the placement of the immediately preceding trial (e.g., if "last week" is tested just after "tomorrow," it should be placed to the left of "tomorrow"). Both analyses yielded the same pattern of results. Here we report only the latter analysis.

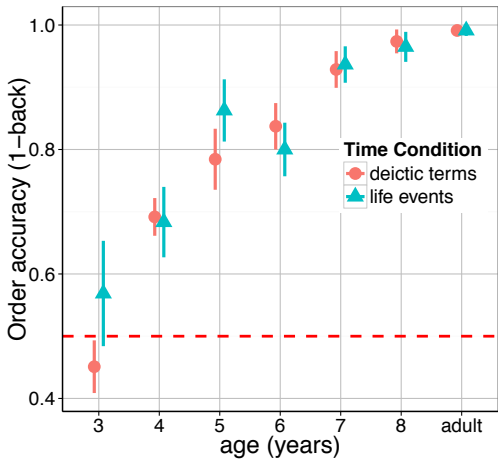


Figure 3: *Order Knowledge.* Performance improved with age, for both deictic time words and events. Error lines = SEM; dashed line = chance performance.

For each child and Time Type, we calculated mean accuracy on this 1-back measure of order knowledge. Once again, there was only a main effect of Age, $F(6, 138) = 33.1$, $p << .01$ (Fig. 3). Just as with Deictic Status, four-year-olds were significantly better than chance, $t_{16} = 4.1$, $p << .01$, while three-year-olds were not, $t_{14} = -0.6$, $p > .5$. Compared to adults, seven-year-olds were significantly worse, $t_{56} = -2.8$, $p < .01$, but eight-year-olds were indistinguishable, $t_{52} = -1.7$, $p > .1$.

3.1.3 Remoteness. We next evaluated knowledge of the temporal remoteness of each item from "now." In essence, we wanted to know to how well children used spatial distance to represent temporal remoteness. Since we were interested in the use of relative and not absolute space, we first standardized the total amount of the line used by each child by dividing each distance from zero by the maximum distance at which any item was placed by that child. Distances thus ranged from 0 to 1. This approach acknowledges that the absolute location of, e.g., "last year", relative to the line's endpoint (infancy) is likely to differ with age, letting us focus on whether participants differentiated how much closer e.g., "yesterday" is to the present, relative to wherever they assigned "last year."

To characterize the maturity of children's representations of remoteness, for each child, we used multiple regression to see whether each item's distance was predicted by (1) the item's correct order, relative to the midpoint, and (2) the mean distance, from the midpoint, where adults placed that item. As a measure of each child's knowledge of temporal remoteness, we used the strength of the relationship between the child's placements and adult-like placements, after factoring out the children's knowledge of order (i.e., semi-partial correlation squared). Mean remoteness knowledge for each age group is shown in Fig 4. Children's knowledge of remoteness improved gradually across our entire age-range, with even 8-year-olds performing significantly differently from adults.

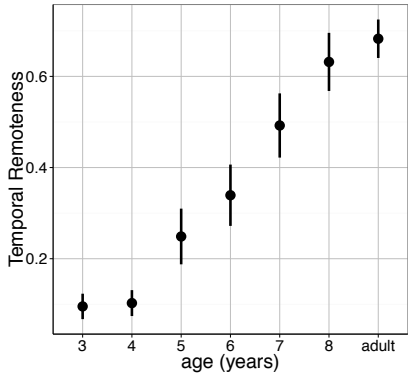


Figure 4: *Remoteness Knowledge*. Children improved gradually but did not reach adult-like performance. Error lines = SEM.

3.2 Order of acquisition

Having characterized children's developing understanding of deictic status, order, and remoteness, we next investigated the order in which these facets of the meaning of deictic time words are learned. We used a threshold approach, comparing the ages at which the majority (50%) of children had made the transition from partial to adult-like understanding of each facet of meaning.

First, based on the continuous measures of their knowledge of deictic status, order, and remoteness, each participant was characterized, via k-means clustering, as a "knower" vs. "non-knower" of that domain (Fig. 5, black dots). "Knowers" included all of the adult participants and the children who clustered with them. Then, for each facet of meaning, we modeled the transition to adult-like knowledge using a Weibull function, where age predicted whether or not a child exhibited adult-like knowledge. This allowed us to estimate the age at which the majority of children acquire adult-like knowledge of each facet, independently of the other two.

The age at which the majority of children transitioned to adult-like knowledge of deictic status and order was the same, around age 6 (Deictic Status: 6.0, 95% bootstrapped confidence interval: [5.45, 6.39]; Order: 5.9 years, [5.42, 6.37]). By contrast, most children did not transition to adult-like knowledge of temporal remoteness until more than a year later, at 7.30 years ([6.81, 7.66]). Since the confidence interval for remoteness does not overlap with the other confidence intervals, this delay is statistically significant.

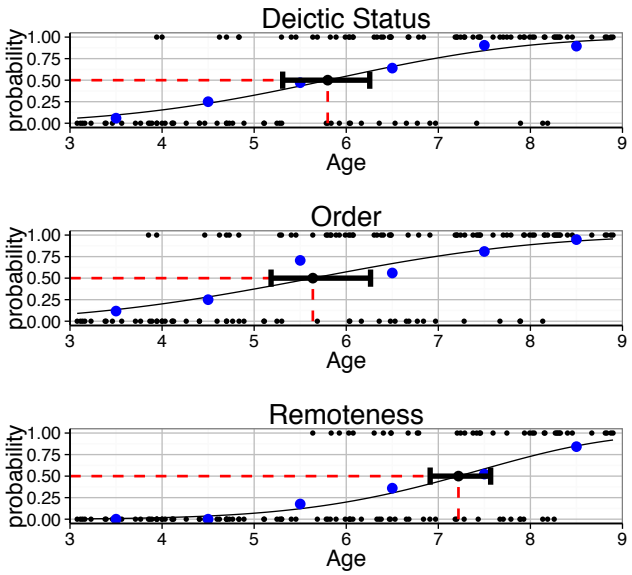


Figure 5: *Age of acquisition for three facets of time word meaning*. Each black dot is an individual child, categorized as either a "knower" (top) or "non-knower" (bottom) of that dimension. Blue dots indicate the mean probability of being a knower in that age group. Red dotted lines indicate age-thresholds for $p = 0.5$; black horizontal error bars indicate bootstrapped confidence intervals on those age-thresholds.

3.3 Learning contingencies

The previous analysis revealed that on their sixth birthday, only a quarter of children exhibit an adult-like grasp of the remoteness of deictic time words, while half of

children know their deictic status, and half know their relative order. But this analysis does not tell us whether the very same children who understand one facet of deictic terms' meanings also understand other facets.

To address this, we calculated the conditional probability of being each type of "knower," given one's "knower" status on each other facet of meaning (Fig. 6). Deictic status and order were highly linked: children who knew one of them almost always knew the other as well. However, while it was extremely uncommon for a child to be a remoteness-knower without also being an order knower (97%) and deictic-status knower (88%), the reverse was not true. The average deictic-status-knower had only a 53% chance of being a remoteness-knower, while the average order-knower had only a 57% chance of being a remoteness-knower. This is exemplified by the top timeline in Fig. 6b, produced by a 5-year-old who knows both deictic status and order, but is far from adult-like in her representation of remoteness. Compare this to the bottom timeline, from an adult, where items are placed based not only on their deictic status and order, but also their relative remoteness (i.e., yesterday is close to now; next week is farther away; and next year even farther). Together with the cross-sample age-of-acquisition data, these results reveal a clear developmental trajectory in which both deictic status and order emerge early and in synchrony, while knowledge of temporal remoteness is developed independently and often much later.

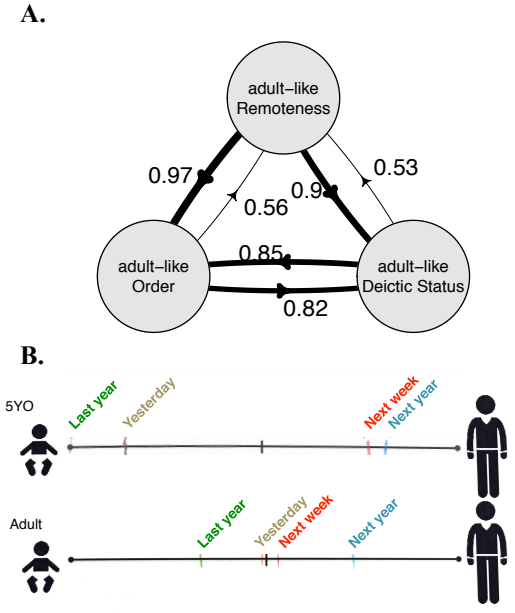


Figure 6: *A. Contingencies among facets of meaning*. Arrows from A to B denote direction of influence. Line width indicates conditional probability of knowing facet B, given knowledge of facet A (scaled from .5 to 1). *B. Example timeline data* from a 5-year-old and an adult. The child exhibits knowledge of deictic status and order, but not remoteness.

4. Discussion

Here, as a case study in abstract word learning, we explored the prolonged trajectory over which deictic time words, like "yesterday," are learned. Prior studies have shown that children produce their first deictic time words at age 2 or 3, but make errors in their use until age 5 or older (e.g., Busby Grant & Suddendorf, 2011). The current study further investigated this delay between production and comprehension, by independently assessing children's knowledge of three facets of the meanings of these words: deictic status, order, and remoteness. Despite their speech errors, we found that even 4-year-olds have partial understanding of the words' deictic status (e.g., that "yesterday" was in the past) and order (e.g., that "yesterday" was before "this morning"), and that most children master these elements of meaning by their 6th birthday. However, our results suggest that fully adult-like mastery of deictic time words, including knowledge of temporal remoteness (and the ability to represent it spatially), may be achieved even later than prior work has revealed. While the majority of children showed evidence of mastery of all three facets of word meaning between their 7th and 8th birthdays, some were not adult-like on our task even shortly before age 9.

What can the present findings tell us about how deictic time words are acquired? Unlike most terms in the early lexicon, their meanings cannot be constrained simply by observing the perceptual structure of the world and performing simple "word-to-world" mappings. In particular, constraints like the *whole-object bias* and the principle of *mutual exclusivity,* which can explain much about how toddlers rapidly acquire hundreds of new nouns, are counterproductive in this case (see Markman, 1994 for discussion of word learning biases). Not only are the referents of deictic time words non-objects, but to the extent that these words *do* denote perceptual experiences (i.e., lived or anticipated events), those associations are transient. There are many-to-many relations between deictic time words and events—any number of events could occur *tomorrow*, and each one of them will cease to be *tomorrow,* and become *yesterday*.

It has been proposed that, to acquire "hard words" such as these, children may rely on structure of language itself (e.g., Gleitman et al., 2005). We posited two potential mechanisms by which linguistic cues could support time word learning -- *syntactic bootstrapping* and *lexical ordering* -- and the present findings provide support for both. In particular, the early acquisition of deictic status supports a syntactic bootstrapping hypothesis, because deictic status is indicated by the morphological tense system of English. Long before they use deictic time words accurately, children comprehend the distinction between past and future tenses (e.g., Weist et al., 1991). For instance, they understand that adding *-ed* to a regular verb indicates that the action has already occurred, and could use this cue to infer the deictic status of a term like "yesterday" if it appears in the same sentence. However, tense alone is insufficient to acquire either order or remoteness -- the tense cues in, e.g., "she played yesterday and worked last week", can't tell us which event happened first, or how long ago either event was.

The fact that order is learned simultaneously with deictic status indicates that other inferential processes are also at work. Under the *lexical ordering* account, other linguistic cues might help children place time words in a common semantic category (e.g. all past words) and then learn their relative order within it. For instance, children might infer that time words form a common class from their appearance in similar sentence frames, and direct contrast could support learning their order (e.g., "That party wasn't *yesterday*, it was all the way *last week*"). Prior work on the acquisition of duration words, such as 'week' and 'year,' also supports this account: at 4, children grasp that duration terms comprise a category (Shatz et al., 2010), and by 5 they begin to work out their ordering (i.e., *year* > *week* > *hour* > *minute*; Tillman & Barner, 2015).

Understanding the temporal remoteness of deictic time words trails knowledge of the other facets of time-word meaning by over a year. Unlike deictic status and order, which could plausibly be inferred from language alone, it is difficult to imagine how children might learn the remoteness of time words from speech input without relying on considerable lived experience. Under the *event mapping* account, after developing a sense of how long ago their last birthday was, a child might be able to associate a deictic term like "last year" with the remoteness of an event, like a party, said to occur then (i.e., by abstracting the time of the event, relative to its time of utterance, as the relevant dimension). However, the finding that remoteness is acquired so late suggests that such mappings are an inefficient means of learning the set of deictic time words, perhaps due to the pitfalls of standard word-to-world mappings described above.

Interestingly, the transition to mature knowledge of remoteness occurs around the same time that clock reading becomes a major focus in standard elementary school curricula, in Grade 2 (Common Core State Standards Initiative, 2012). This raises the possibility that adult-like representations of remoteness depend upon exact definitions for time words. In the case of duration words, children have little understanding of the absolute duration of each word (e.g., *hour=minute*×60; *year=day*×365) until around age 7, when they are taught definitions. If adult-like representations of remoteness also rely on this understanding, we would expect children to fail until the same age, as we observe. In both cases, children infer early on that time words are inter-defined, but often cannot use them accurately to reference events until much later.

Here, we leveraged the spatial timeline as a tool to characterize children's knowledge of time words more precisely than prior methods. However, this method necessarily conflates children's semantic understanding of the target items with their ability to express that knowledge spatially. Confusion about timelines could potentially mask semantic knowledge of time words. Though evidence suggests children as young as 3 can comprehend continuous spatial number-lines (e.g. Bartelleti et al., 2010), future research will be needed to directly test the possibility that this timeline method underestimates children's knowledge of deictic time words.

Strikingly, we have mapped out a developmental trajectory for time-word learning that can take children *five years* or more. Arguably, this process represents a profound conceptual breakthrough. In order to use deictic time words like adults do, a child must have the idea that there is dimension of time that is separate from the events that occupy it. While this might challenge the child's early assumptions about what words can label, mastering this insight may provide a transformational framework for interpreting the past and planning for the future.

References

Ames, L. B. (1946). The development of the sense of time in the young child. Journal of Genetic Psychology, 68, 97–125.

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental psychology*, *46*(2), 545.

Busby, J., & Suddendorf, T. (2005). Recalling yesterday and predicting tomorrow. *Cognitive Development*, *20*(3), 362-372.

Busby Grant, J., & Suddendorf, T. (2009). Preschoolers begin to differentiate the times of events from throughout the lifespan. *European Journal of Developmental Psychology*, *6*(6), 746-762.

Busby Grant, J., & Suddendorf, T. (2011). Production of temporal terms by 3-, 4-, and 5-year-old children. *Early Childhood Research Quarterly*, *26*(1), 87-95.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Common Core State Standards Initiative *(2012). Common core state standards for mathematics*.

Friedman, W. J. (2002). Children's knowledge of the future distances of daily activities and annual events. *Journal of Cognition and Development*, *3*(3), 333-356.

Friedman, W. J., & Kemp, S. (1998). The effects of elapsed time and retrieval on young children's judgments of the temporal distances of past events.*Cognitive Development*, *13*(3), 335-367.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. Language Learning and Development, 1, 23–64.

Harner, L. (1975). Yesterday and tomorrow: Development of early understanding of the terms. *Developmental Psychology*, 11(6), 864–865

Harner, L. (1981). Children's understanding of time. Topics in language disorders, 2(1),51-66.

Shatz, M., Tare, M., Nguyen, S. P., & Young, T. (2010). Acquiring non-object terms: The case for time words. *Journal of Cognition and Development*,*11*(1), 16-36.

Hudson, J. A., & Mayhew, E. M. (2011). Children's temporal judgments for autobiographical past and future events. *Cognitive Development*, *26*(4), 331-342.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua*, *92*, 199-227.

Nelson, K. (1998). *Language in cognitive development: The emergence of the mediated mind*. Cambridge University Press.

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. *Cognitive psychology*, *78*, 57-77.

Weist, R. M., Wysocka, H., & Lyytinen, P. (1991). A cross-linguistic perspective on the development of temporal systems. *Journal of child language*, *18*(01), 67-92.

Spatial Interference and Individual Differences in Looking at Nothing for Verbal Memory

Alper Kumcu (axk393@bham.ac.uk)
School of Psychology, University of Birmingham
Edgbaston, Birmingham, B15 2TT UK

Robin L. Thompson (r.thompson@bham.ac.uk)
School of Psychology, University of Birmingham
Edgbaston, Birmingham, B15 2TT UK

Abstract

People tend to look at uninformative, blank locations in space when retrieving information. This gaze behaviour, known as looking at nothing, is assumed to be driven by the use of spatial indices associated with external information. We investigated whether people form spatial indices and look at nothing when retrieving words from memory. Participants were simultaneously presented four words. During retrieval participants looked at the relevant, blank location, where the probe word had appeared previously, longer than the other blank locations. Additionally, word presentation was sometimes followed by a visual cue either co-located or not with the probe word. Valid cues functioned as visual reinforcement while invalid cues caused interference. Finally, participants with better visuospatial memory looked less at the relevant, blank location, suggesting a dynamic relationship between so-called "external" and "internal" memory. Overall findings suggest an automatic, instantaneous spatial indexing mechanism for words and a dynamic looking at nothing behaviour.

Keywords: looking at nothing; spatial indexing; mental representation; visuospatial memory; verbal memory; spatial interference; individual differences

Introduction

The human mind can anchor spatially-located information to external spatial locations. This mechanism has been expressed within a visual processing model, where the *location* of an object is separated from the *visual feature*s of it (Marr, 1982). This view, expanded into an exhaustive *spatial indexing model* (Pylyshyn, 2001), assumes that the visual system is able to individuate spatial relations before discerning a visual pattern and immediately index the locations of such patterns.

Additionally, within the model, spatial indices remain attached to a particular object independent of its movements and visual properties. Spatiotemporal continuity occurs even when the visual information disappears, as often manifested in mental imagery studies (i.e., Brandt & Stark, 1997). Accordingly, spatial indices tied to external visual and verbal information trigger eye movements when a mental representation is reactivated. Thus, when retrieving information from memory, people tend to exploit location-based indices and look at the seemingly uninformative, empty locations where the information originally occurred even if location is irrelevant to the task. This behaviour is known as *looking at nothing* (Ferreira, Apel, & Henderson, 2008).

In their pioneering study, Richardson & Spivey (2000) documented the use of spatial information and looking at nothing in verbal memory. Four faces randomly appeared on different quadrants of a two by two grid along with four corresponding spoken facts (e.g., "Shakespeare's first plays were historical dramas; his last was the Tempest"). On the next screen, a statement (e.g., "Shakespeare's first play was the Tempest") probed participants' memory for verbal information. During recall, there were significantly more looks at the blank quadrant where the face associated with the probed semantic information had been when compared to other quadrants. Thus, people did not just look at any *nothing* when answering the questions. Rather, they looked at an invisible spatial index, which was previously allocated to the information (Spivey & Geng, 2000).

Looking at nothing may be best thought of as an interface between internal and external worlds. Ferreira et al. (2008) proposed an *integrated memory architecture*, where external cues and internal representations work hand in hand to retrieve information as efficiently as possible (see also Richardson, Altmann, Spivey, & Hoover, 2009). More precisely, the integrated memory account combines visual/auditory and spatial information in the external world with visual, linguistic, spatial and conceptual counterparts in the mental world. When part of an integrated representation (linguistic information) is reactivated, the other parts (spatial information) are retrieved as well.

In the current study, we address the looking at nothing triangle, which is composed of (A) actual looking behaviour, (B) spatial indices and (C) mental representations. Unlike any other looking at nothing studies, we used single, visual words as retrieval material instead of visual objects (i.e., Johansson & Johansson, 2014; Martarelli & Mast, 2013; Spivey & Geng, 2000; Vankov, 2009; Wantz, Martarelli, & Mast, 2015) or auditorily presented statements which are explicitly associated with visual objects (i.e., Hoover & Richardson, 2008; Richardson & Kirkham, 2004; Richardson & Spivey, 2000; Scholz, Mehlhorn, & Krems, 2011, 2014). Our main motivation was to reveal automatic and instantaneous spatial indexing and clear-cut looking at nothing behaviour guided specifically by verbal memory in a visuospatial context. Additionally, we focus on two separate but interrelated questions with regard to different vertices of this triangle within the scope of dynamicity of looking at nothing.

Spatial Interference and Spatial Indexing

First, we probe (1) how (in)congruency of interfering spatial cues affects spatial indexing and looking at nothing. A plethora of studies show that the human mind is susceptible to spatial manipulation in tasks even when the location is irrelevant to successful performance. People react faster and perform better when there is spatial compatibility between cue and probe (the Simon effect; Simon & Rudell, 1967). Although there is evidence for a Simon-like effect in spatial indexing (Vankov, 2009), relatively little is known about the role of spatial interference and congruency in looking at nothing.

Understanding spatial interference and congruency is important for defining a spatial encoding mechanism. Looking at nothing has been observed even when the locations associated with the to-be-retrieved information moved and thus updated the spatial indices. Both adults and 6-month-olds were able to track moving events and looked at the updated locations, indicating a flexible and dynamic spatial indexing structure (Richardson & Kirkham, 2004). In light of such evidence, we might argue that if looking at nothing is sensitive to systematic cue manipulation (i.e., Wühr & Ansorge, 2007), that is, to (in)congruencies with the original indices, then the link between a spatial indexing mechanism and looking at nothing might be stronger than previously thought. More precisely, such a directional link might suggest that not only the existence but also the magnitude of looking at nothing is determined by the strength and stability of spatial encoding.

Interfering cues are also important in understanding the interplay between (B) spatial indices and (C) mental representations. If looking at nothing depends on internal operations working with external spatial cues, then spatial codes which are updated with interference should lead to less looking at nothing. This is predicted based on the assumption that, in such as a case, space becomes unreliable: i.e., there is competition between the mental representations corresponding to the spatial indices for words and for cues, respectively. In contrast, a valid spatial code should strengthen the association between mental representations corresponding to the word and to its location, which, in turn, should be reflected in looking behaviour. While stability of spatial indexing across time has been shown (Martarelli & Mast, 2013; Wantz et al., 2015), spatial stability and the role of interference in looking at nothing still remain largely unknown.

Integrated Memory and Individual Differences

Additionally, we investigate the relationship between (A) looking at nothing and (C) mental representations by asking (2) whether there are individual differences in looking at nothing behaviour based on "internal" visuospatial memory. We hypothesize that the cognitive system uses both internal and external cues to access memory traces. Therefore, external cues may be used to relieve internal operations, and people with relatively worse visuospatial memory should rely more on looking at nothing behaviour (and vice versa). Determining the existence or absence of a correlation between individual visuospatial memory differences and memory-driven eye movements is essential to understanding the intrinsic nature of looking at nothing and its relation with mental representations. If a correlation is found, it will provide further evidence for the integrated memory account (Ferreira, et al., 2008). There is already growing evidence that looking at nothing changes according to internal demands. For example, people tend to exhibit less looking at nothing as they are asked to study and recall the same sentences over and over again, suggesting less reliance on external cues as the task becomes easier through repetition (Scholz et al., 2011). However, not much is known about how differences in internal memory map onto differences in looking at nothing behaviour within the scope of integrated memory operations.

Our experimental paradigm diverges from the previous looking at nothing studies in the following ways: (1) Single words were used instead of visual objects as information to be retrieved. Looking at nothing is fundamentally a visuospatial phenomenon. However, what makes the words recognizable visually is both controllable and not usually related to vision (Harm & Seidenberg, 2004). Thus, we aimed to disentangle the retrieval items from the visual environment to be able to observe more refined behaviour. We also aimed to systematically manipulate and control the stimuli and rule out any item-related effect on the memory load and thus, looking at nothing. Accordingly, words were controlled for a number of memory-related variables. (2) Participants were exposed to retrospective memory interference which was irrelevant to the main task. We expected to push out old information (i.e., encoded words) from the episodic buffer (Baddeley, 2000) and encourage participants to depend on spatial indices for the retrieval of words. (3) Words to be retrieved were presented simultaneously and were not explicitly associated with any kind of visual object. Rather, participants processed the words together and were expected to form immediate indices based on the word location. This is how verbal information is processed in real-world cognitive tasks such as reading (see Fischer, 1999), thus making the task more naturalistic. We aimed to unearth more ecologically valid, systematic and robust looking at nothing behaviour.

Method

Participants

The experiment was carried out with forty-eight students at the University of Birmingham (six males; M_{age} = 19.92, *SD* = 1.96, range: 18 – 27, four left-handed). All participants were monolingual native speakers of British English as determined with the Language History Questionnaire (version 2.0; Li, Zhang, Tsai, & Puls, 2013). Participants reported normal or corrected-to-normal vision, no speech or hearing difficulties and no history of any neurological disorder. They received either £6 (n = 12) or course credit (n = 36) for participation. All participants were fully informed about the details of the experimental procedure and gave written consent. Post-experiment debriefing revealed that all participants were naïve to the purpose of the experiment.

Materials

There were 192 trials involving 864 unique nouns in total. Trials were evenly divided into two groups (n = 96) as experimental (positive probe) trials and fillers. Probe words in the experimental trials were among the four study words in the encoding phase, whereas a different, not seen, word was probed in fillers. Words in the experimental trials (n = 384) were drawn from the extensions of Paivio, Yuille and Madigan norms for 925 nouns (Clark & Paivio, 2004). The word pool was filtered to exclude words shorter than 3 letters and longer than 6 letters. Imageability, frequency (logarithmic values of occurrences per million in Kučera & Francis, 1967; and the CELEX database, Baayen, Piepenbrock, & Gulikers, 1995), age of acquisition, concreteness, availability (Keenan & Benjafield, 1994), length in letters and number of syllables were identified as major predictors of verbal memory (Rubin & Friendly, 1986) and used to control the experimental stimuli.

The subset was then grouped into quadruples and trial sets were identified. Words within quadruples were matched on age of acquisition, availability, concreteness, imageability, length in letters, log frequency and number of syllables (all SDs < 2.00 and all SEs < 1.00). Words were further controlled so that no word started with the same letter, rhymed or related semantically with any other in the quadruple. Monosyllabic, disyllabic and trisyllabic words were evenly distributed [e.g., (3, 3, 3, 3), (1, 2, 1, 2) or (3, 2, 3, 2) etc.]. The word in each trial set with the median imageability value was selected as the probe among four words leaving the others as distractors (see Rubin & Friendly). Welch's t-tests revealed no significant difference between the probe and distractor words in any of the variables (all ps > .05). Thus, any word among the four words in each trial set was as likely to be remembered as any other word. Words in filler trials were drawn from the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). They were also controlled to develop a consistent stimuli set. Words were grouped into quintuples and matched on log frequency in CELEX database (all SDs < 0.60 and all SEs < 0.30). Welch's t-tests revealed that there was no significant difference between the probe and the study words in frequency, length in letters or number of syllables (all ps > .05).

We formed 192 unique mathematical equations (e.g., (5X2) – (1+8) = 1) to present as memory interference between encoding and retrieval phases (see Conway & Engle, 1996 for a similar design). Half of the equations were correct. Incorrect equations were further divided into two equal groups: The results were either plus or minus one of the correct result.

Apparatus

Stimuli were presented on a TFT LCD 22-inch widescreen monitor operating at 60 Hz with a resolution of 1680 x 1050 pixels (501.7 mm x 337.4 mm). The monitor was placed 640 mm in front of the participant. A chin and forehead rest was used to reduce head movements. Participants' eye movements were monitored using SR EyeLink® 1000 (sampling rate: 1000 Hz, spatial resolution < 0.5°, http://sr-research.com/eyelink1000.html). Viewing was binocular but only the left eye was monitored. Auditory material was produced by a native female speaker of British English in a sound attenuated room and recorded using Audacity (version 2.1.10, http://web.audacityteam.org). Participants responded (yes/no they had seen the word) by pressing one of two keys on a standard keyboard. Eye movement data were analysed using the SR EyeLink® Data Viewer (version 2.4.0.198, http://www.sr-research.com/accessories_EL1000_dv.html). No drift or blink correction procedure was applied. Data were analysed and visualised with R (version 3.2.3) (R Core Team, 2015).

Procedure

A pre-experiment questionnaire involving Language History Questionnaire and Edinburgh Handedness Inventory (Oldfield, 1971) was administered.

Eye tracking started with a standard nine-point calibration and validation, which confirmed high data quality (average calibration error < 1° and maximum calibration error < 1.50°). The experiment was composed of five consecutive phases. **Fixation**: A fixation cross appeared at the centre of the screen for 500 ms. **Encoding**: Participants were presented four words on a 2 x 2 grid for 1600 ms. Words (Times New Roman, font size = 40) were centrally placed in rectangular boxes (285 x 85 in pixels, 7.6° x 2.4° of visual angle). **Cueing**: A flashing black dot appeared in cue trials for 1000 ms either in the same (valid cue) or in the diagonal quadrant (invalid cue) as the original location of the probe word in the encoding phase. There was also a third condition where no cue was presented between encoding and interference. The cue manipulation was a between-subjects variable. An equal number of random participants (n = 16) saw the probe word with valid cue, invalid cue or without any cue. **Interference**: Participants were presented a mathematical equation and asked to identify whether the equation was correct or not within 10,000 ms. **Retrieval**: The probe word was auditorily presented as participants looked at the blank grid with empty boxes. Participants were asked to make an unspeeded yes/no judgement to determine whether they had seen the probe word among the four words shown in the encoding phase within 10,000 ms (or they timed-out).

The order of trials and equations were fully randomised. The location of all words in all conditions was counterbalanced with Latin Square design to control gaze biases so that each word appeared an equal number of times in each location of the grid. The experiment was divided into four equal blocks and there was a short pause between blocks. A typical session lasted approximately 60 minutes. Overall accuracy in interference equations was 81.19% and 86.07% in the verbal recognition test, suggesting that participants were attending to the task.

Following the experiment, a computerized version of the Corsi block-tapping task (Milner, 1969) operated on PEBL (Psychology Experiment Building Language, version 0.13, test battery version 0.7, http://pebl.org) (Mueller & Piper, 2014) was used to measure visuospatial short-term memory.

Results

Dwell time percentage (i.e., percentage of total time - in milliseconds - spent on a specific interest area) was used as the main gaze measure and dependent variable because it is immune to differences in task duration. Accordingly, four rectangular interest areas corresponding to the quadrants were identified. All interest areas were of the same size (502 x 368 in pixels, 13.4° x 10.6° of visual angle). They framed the rectangular boxes and were not contiguous. A circular interest area with a diameter of 40 pixels (1.1° of visual angle) was also defined at the centre of the grid. Dwell time percentages accrued on the interest areas during the retrieval phase (from the presentation of the probe word until the participant's response) were calculated. Fixations were a minimum duration of 40 ms. Fixations outside the interest areas (6.54%) were omitted.

Looking at Nothing with Spatial Interference

We began by investigating whether there was a difference between spontaneous looking times across quadrants and cue conditions during the retrieval phase. Dwell time percentages allocated to three irrelevant quadrants were averaged into one irrelevant quadrant and analysed against the relevant quadrant across the three different cue conditions. A mixed analysis of variance showed a main effect of quadrant $F1(1, 141) = 14.40$, $p < .001$; $\eta_p^2 = .09$; $F2(1, 573) = 15.85$, $p < .001$, $\eta_p^2 = .03$ and an interaction effect of cue condition $F1(2, 141) = 3.60$, $p = .03$, $\eta_p^2 = .05$; $F2(2, 573) = 2.89$, $p = .06$, $\eta_p^2 = .01$. Paired t-tests revealed that participants looked significantly longer at the relevant quadrant than the average of three irrelevant quadrants when retrieving the probe word in all conditions together and in the valid cue condition (see Table 1).

Table 1: Differences between dwell time percentages of relevant and irrelevant quadrants under different cue conditions ($df = 47$).

Condition	Relevant	Irrelevant	t	p	d
Valid	0.22	0.17	3.66	.00	0.6
Invalid	0.19	0.19	0.64	.53	0.1
No Cue	0.19	0.17	1.93	.06	0.3

Participants did not look at the relevant quadrant significantly longer in the invalid cue condition. Therefore, dwell time percentages in the invalid cue condition were further analysed to understand the impact of the invalid cue. A repeated measures analysis of variance showed that there was not a significant difference between dwell time percentages across cue quadrant (i.e., where the cue was presented in the cueing phase) ($M = 0.20$, $SD = 0.08$), relevant quadrant ($M = 0.19$, $SD = 0.08$) and irrelevant quadrant ($M = 0.18$, $SD = 0.06$) $F1(2, 94) = 1.33$, $p = .27$; $F2(2, 382) = 1.95$, $p = .15$.

Dwell time percentage spent on the relevant, blank quadrant decreased significantly across blocks $F(3, 141) = 4.33$, $p = .006$, $\eta_p^2 = .08$.

Individual Differences in Looking at Nothing

We investigated whether looking at nothing behaviour changes according to the visuospatial memory differences of participants. Overall, there was a significant, positive correlation between dwell time percentage spent on the central interest area and visuospatial memory measured with Corsi block-tapping test r_s (46) = .34, p = .02 such that participants with better visuospatial memory tended to look more at the centre of the screen and did not look at "nothing" (i.e., relevant, blank quadrant) by definition. An additional variable, *looking at nothing strength*, was formulated by simply subtracting dwell time percentage spent on the central interest area from dwell time percentage spent on the relevant quadrant. As expected, there was also a significant, negative correlation between looking at nothing strength and visuospatial memory r_s (46) = - .29, p = .04 (see Figure 1).

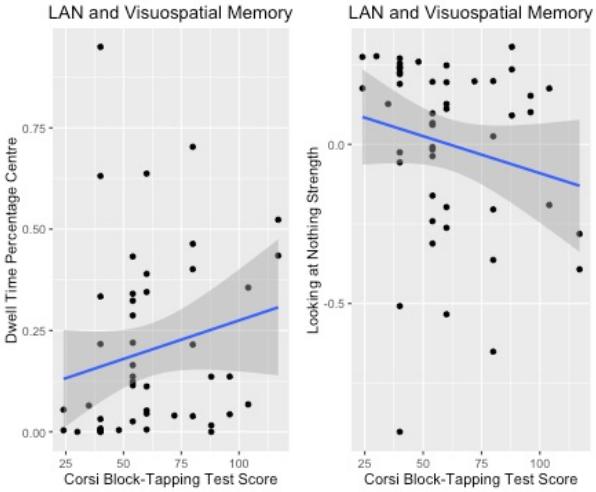


Figure 1: Scatterplots showing the correlations between looking at nothing and visuospatial memory.

To analyse the effect of the interfering cue on the correlation between looking at nothing and visuospatial memory, participants were divided into two equal groups: (1) good (memory score, $M = 79.08$ $SD = 19.54$) and (2) poor (memory score, $M = 43.04$ $SD = 9.43$) visuospatial memory. Welch's t-tests showed that participants with better visuospatial memory looked significantly less at the relevant, blank quadrant when retrieving the probe word during the invalid cue condition compared to participants with poor visuospatial memory (Good $M = 0.17$, $SD = 0.07$, Poor $M = 0.22$, $SD = 0.07$) $t(45.95) = 2.49$, $p = .02$, $d = 0.7$ but not in valid or no cue condition ($ps > .05$). The effect of invalid cue was confirmed with Spearman correlations. As found in all conditions together, there was a significant, positive correlation between dwell time percentage spent on the central interest area and visuospatial memory r_s (46) = .33, p = .02 and also a significant, negative correlation between looking at nothing strength and visuospatial memory r_s (46) = - .35, p = .01 in invalid cue condition.

Discussion

In the present study, we investigated the spatial indexing and looking at nothing processes with (in)congruent spatial

cues for simultaneously presented single words in a recognition memory test, and whether visuospatial memory differences between participants correlate with looking at nothing behaviour. Participants instantly formed spatial indices corresponding to simultaneously presented words even though the locational information was unnecessary for the successful completion of the task. They also looked at relevant, blank locations significantly longer than the other, irrelevant blank locations when they were asked to retrieve the probe word.

We replicated a looking at nothing effect with marginal significance in a "pure" looking at nothing condition where no cue was presented. Given that the presented words were not explicitly associated with visual cues in our experiment, results might be interpreted as further evidence for the automaticity and availability of spatial indices (Vankov, 2009), which, guide eye movements to empty locations. Along with that, the novelty of the present research lies in the use of words instead of visual objects. The fact that participants used the visuospatial channel to access verbal memory traces suggests that looking at nothing is a distinctive, memory-oriented behaviour and might be even more robust than previously documented. One might argue that the looking at nothing effect in this study can be accounted for by an attentional mechanism initiated by the interfering cues. However, participants also looked at the relevant, blank quadrant without any cue in the pure looking at nothing condition with a marginally significant difference and a small effect size. Further, they did not look at the previous location of the cue longer than the other quadrants in the invalid cue condition. Therefore, we conclude that spatial indices formed for single words can reliably orient eye movements to blank locations in a recognition memory task.

Results from the cue manipulation confirmed our hypothesis in general and resulted in a Simon-like effect (Wühr & Ansorge, 2007). As expected, participants formed spatial indices for cues. When the spatial index corresponding to the probe word and the index corresponding to the cue matched (as in the valid cue condition), the looking at nothing effect was amplified. However, when these indices were in competition (as in the invalid cue condition), the initial index was updated and eye movements to the relevant, blank location were disrupted. Taken together, cue manipulation results demonstrate that the link between spatial indexing and looking at nothing is indeed dynamic and systematic. Similarly, we observed a decrease in looking at nothing towards the end of the experiment, which was in favour of the previous findings (Scholz et al., 2011) with the exception that participants in our study studied different items throughout the experiment. One explanation could be that as the participants gradually became accustomed to the task, mental load decreased and so did the necessity to draw information from space.

The relation between mental load and reliance on space was also reflected in individual differences results. To our knowledge, this is the first evidence showing individual differences when looking at nothing. Participants with better visuospatial memory, thus richer internal sources, relied less on the spatial indices and consequently looking at nothing. The tendency of looking at the centre was revealed among these participants probably because they were faster in general and did not have the necessity to refer to any blank quadrants including the relevant one. Although the correlations were either weak or moderate, this alone might be regarded as clear evidence for the integrated memory model, where external (space) and internal (mental representations) elements of memory work together (see Jackendoff, 1996). More importantly, differences were mostly pronounced in the invalid cue condition. Thus, in the event of spatial interference and confusion, participants with better visuospatial memory seemed to disengage from the space by ignoring any deictic code either attached to words or cues, and turned to internal sources. This is particularly important considering that, looking at the centre was not a general trend among all participants in invalid cue condition. The correlation might also suggest that looking at nothing is not a by-product but a functional (Johansson & Johansson, 2014; Scholz et al., 2014) and even a strategic behaviour which systematically changes not only with the processing demands of the task but also from individual to individual. In a nutshell, our results suggest that there is a balanced trade-off between internal and external sources driven by task conditions in order to make the most of environmental opportunities and cognitive capacity.

Looking at nothing is a unique case in that it demonstrates how the cognitive system can maximize efficiency by spreading the cognitive problem across three domains with the act of looking, the environment with the spatial indices and mental representations in the brain. Further, looking at nothing in this study can be regarded as an example of very efficient multimodal coordination given that participants studied verbal information and made use of the visuospatial canvas when auditorily probed. In this regard, the study is expected to have broader implications towards a new and unorthodox understanding of cognition.

References

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database. Philadelphia, PA.: Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. doi:10.1016/S1364-6613(00)01538-2

Brandt, S. A., & Stark, L. W. (1997). Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene. *Journal of Cognitive Neuroscience*, *9*(1), 27–38. doi:10.1162/jocn.1997.9.1.27

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc*, *36*(3), 371–383. doi:10.3758/BF03195584

Conway, A. R., & Engle, R. W. (1996). Individual differences in working memory capacity: more evidence for a general capacity theory. *Memory*, *4*(6), 577–590. doi:10.1080/741940997

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, *12*(11), 405–10.

doi:10.1016/j.tics.2008.07.007
Fischer, M. H. (1999). Memory for word locations in reading. *Memory*, *7*(1), 79–116. doi:10.1080/741943718
Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375–399. doi:10.3758/BF03203275
Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review*, *111*(3), 662–720. doi:10.1037/0033-295X.111.3.662
Hoover, M. A., & Richardson, D. C. (2008). When facts go down the rabbit hole: contrasting features and objecthood as indexes to memory. *Cognition*, *108*(2), 533–542. doi:10.1016/j.cognition.2008.02.011
Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In P. Bloom, M. F. Garrett, L. Nadel, & M. A. Peterson (Eds.), *Language and space* (pp. 1–30). Cambridge, MA: The MIT Press.
Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, *30*(6), 1053–1079. doi:10.1207/s15516709cog0000_86
Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science*. doi:10.1177/0956797613498260
Keenan, T. R., & Benjafield, J. G. (1994). An additional measure of availability derived from the Oxford English Dictionary. *Psychonomic Bulletin & Review*, *1*(2), 255–257. doi:10.3758/BF03200777
Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present Day American English*. Providence: Brown University Press.
Li, P., Zhang, F., Tsai, E., & Puls, B. (2013). Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition*, *17*(03), 673–680. doi:10.1017/S1366728913000606
Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.
Martarelli, C. S., & Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological Research*, *77*(3), 303–309. doi:10.1007/s00426-012-0439-7
Milner, B. (1969). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, *27*, 272 – 277.
Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods*, *222*, 250–9. doi:10.1016/j.jneumeth.2013.10.024
Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*, 97–113. doi:10.1016/0028-3932(71)90067-4
Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1-2), 127–158. doi:10.1016/S0010-0277(00)00156-6
R Core Team (2015). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Richardson, D. C., Altmann, G. T. M., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: a response to Ferreira et al.: taking a new look at looking at nothing. *Trends in Cognitive Sciences*, *13*(6), 235–6. doi:10.1016/j.tics.2009.02.006
Richardson, D. C., & Kirkham, N. Z. (2004). Multimodal events and moving locations: eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General*, *133*(1), 46–62. doi:10.1037/0096-3445.133.1.46
Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition*, *76*(3), 269–295. doi:10.1016/S0010-0277(00)00084-6
Rubin, D. C., & Friendly, M. (1986). Predicting which words get recalled: measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, *14*(I), 79–94. doi:10.3758/BF03209231
Scholz, A., Mehlhorn, K., & Krems, J. F. (2011). Looking at nothing diminishes with practice. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1070–1075). Austin, TX: Cognitive Science Society.
Scholz, A., Mehlhorn, K., & Krems, J. F. (2014). Listen up, eye movements play a role in verbal memory retrieval. *Psychological Research*, *12*(20), 1–10. doi:10.1007/s00426-014-0639-4
Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *The Journal of Applied Psychology*, *51*(3), 300–304. doi:10.1037/h0020586
Spivey, M. J., & Geng, J. J. (2000). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological Research*, *65*(4), 235–241. doi:10.1007/s004260100059
Vankov, I. (2009). Mind the Gap: The Cost of Looking at Nothing, or the Performance Implications of Memory-induced Attention Shifts. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 1318–1323). Austin, TX: Cognitive Science Society.
Wantz, A. L., Martarelli, C. S., & Mast, F. W. (2015). When looking back to nothing goes back to nothing. *Cognitive Processing*, *17*(1), 105–114. doi:10.1007/s10339-015-0741-6
Wühr, P., & Ansorge, U. (2007). A Simon effect in memory retrieval: evidence for the response-discrimination account. *Psychonomic Bulletin & Review*, *14*(5), 984–988. doi:10.3758/BF03194132

Bayesian Pronoun Interpretation in Mandarin Chinese

Meilin Zhan (mezhan@ucsd.edu)[1], **Roger P. Levy (rplevy@mit.edu)**[1,2], **Andrew Kehler (akehler@ucsd.edu)**[1]

[1]Department of Linguistics
University of California, San Diego
San Diego, CA 92093 USA

[2]Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Kehler and Rohde (2013) proposed a Bayesian theory of pronoun interpretation where the influence of world knowledge emerges as effects on the prior and the influence of information structure as effects on the likelihood: $P(\text{referent}|\text{pronoun}) \propto P(\text{pronoun}|\text{referent})P(\text{referent})$. Here we present two experiments on Mandarin Chinese that allow us to test the generality of the theory for a language with different syntactic-semantic associations than English. Manipulations involving two different classes of implicit-causality verbs and passive vs. active voice confirmed key predictions of the Bayesian theory: effects of these manipulations on the prior and likelihood in production were consistently reflected in pronoun interpretation preferences. Quantitative analysis shows that the Bayesian model is the best fit for Mandarin compared to two competing analyses. These results lend both qualitative and quantitative support to a cross linguistically general Bayesian theory of pronoun interpretation.

Keywords: Bayesian modeling; pronoun interpretation; Mandarin Chinese

Introduction

Successful language understanding requires comprehenders to resolve uncertainty in language. One source of potential uncertainty emerges from pronouns (e.g., *he*, *she*, *they*, *it*) since pronouns carry little information and usually do not fully specify the intended referent semantically (e.g., *Jane invited Anne to her house and she had a great time*). Nevertheless, humans generally interpret pronouns rapidly and accurately. Previous work has proposed a variety of factors that influence how comprehenders resolve pronouns. One set of approaches has focused on grammatical factors such as subjecthood preferences (Crawley et al., 1990), first-mention preferences (Gernsbacher & Hargreaves, 1988), and parallel grammatical role preferences (Smyth, 1994). Others have focused on information structural features such as topicality. Centering Theory (Grosz et al., 1995), for instance, uses information-structural relationships within and between utterances and the grammatical roles of potential referents to guide pronoun interpretation. Another approach has argued for the role of world knowledge in referent assignment. The coherence-driven approach (Hobbs, 1979), for example, models pronoun interpretation as a side-effect of the inference processes that underpin general discourse processing.

Recent studies on pronoun interpretation (Kehler et al., 2008; Kehler & Rohde, 2013) integrate aspects of the information structural approach and coherence-driven approach by way of a probabilistic Bayesian framework. Kehler and Rohde (2013) proposed that comprehenders reverse-engineer speakers' intended referents in terms of Bayes' rule, as shown in (M1). The posterior $P(\text{referent}|\text{pronoun})$ represents the interpretation bias: upon hearing a pronoun, the probability of the pronoun referring to a particular referent. On the other hand, the likelihood $P(\text{pronoun}|\text{referent})$ represents the production bias: the probability of the speaker using a pronoun to refer to an intended referent. The prior $P(\text{referent})$ denotes the NEXT-MENTION BIAS: the probability of a specific referent getting mentioned next in the context, independent of the form of referring expression used.

$$P(\text{referent}|\text{pronoun}) = \frac{P(\text{pronoun}|\text{referent})P(\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{pronoun}|\text{referent})P(\text{referent})} \quad \text{(M1)}$$

Equation (M1) says that the relationship between pronoun interpretation and pronoun production follows Bayesian principles, without further specifying which factors affect each term in the numerator. Here we refer to this as the WEAK form of the Bayesian hypothesis. Kehler & Rohde (2013) further suggest that the factors conditioning the prior and the likelihood are different: the influence from world knowledge emerges as effects on the prior, and the influence from information structure as effects on the likelihood. We refer to this specification as the STRONG form of the Bayesian hypothesis. This Bayesian model successfully accounts for a range of English data. However, little work has investigated whether the model generalizes across different languages.

The Present Study

This paper has two objectives. The first is to test the generality of the Bayesian pronoun interpretation theory crosslinguistically, specifically in the case of Mandarin Chinese. Experiment 1 serves as a base-line test case, where we use a passage completion paradigm that allows us to tease apart the influences of world knowledge from those of grammatical and information structural factors. We provide both qualitative and quantitative model evaluations for the experimental data to test the Bayesian account.

The second objective is to further test the theory by way of a syntactic manipulation, specifically the voice used in a context sentence (passive or active). Here we follow up on two results found by Rohde & Kehler (2014) for English; to facilitate our descriptions, for the passive voice (e.g. *Jane was amazed by Sue*) we use the "first noun phrase" NP_1 to refer to the clause's *logical object* (the *syntactic subject* Jane), and the "second noun phrase" NP_2 to refer to the *logical subject* Sue. First, on the assumption that being the syntactic subject of a passive clause is a stronger indicator of topichood than being the syntactic subject of an active clause in English (because the speaker chose a syntactically marked con-

struction to get the logical object into the syntactic subject position), their theory predicts that the rate of pronominalization of the syntactic subject in the passive will be higher than that in the corresponding active case, a prediction that was confirmed by their study. Second, they also found an unexpected effect whereby the voice manipulation influenced next-mention biases. Specifically, Rohde et al. found that passivization increased the rate of next-mention for NP_2, the logical subject. This result was unexpected on the STRONG form of the Bayesian theory. However, Rohde et al. also found that pronoun interpretation reflected this effect of passivization on next-mention preference, as predicted by the WEAK form of the theory.

Here we examine whether similar effects are found for Mandarin Chinese. The syntactic-semantic properties of Chinese passives make these constructions a good test case. Passive voice in Mandarin Chinese is generally realized via the *bei* construction, with linear arrangement NP_1 bei NP_2 verb (e.g., Li & Thompson, 1981). In this construction, NP_1 is the logical object followed by the passive signaling word *bei*, which introduces the logical subject NP_2. The *bei* construction conveys the notion of AFFECTEDNESS (LaPolla, 1988); it describes an event in which the logical object is affected physically or psychologically in some way (Li & Thompson, 1981). We expect that the conveyance of AFFECTEDNESS may increase the probability of mentioning the logical object in the next sentence. If this expectation is borne out in the data, it affords the opportunity for an additional test of the WEAK Bayesian theory, which predicts that any effect of passivization on next-mention preferences should have a corresponding effect on pronoun interpretation preferences. We test these predictions in Experiment 2.

Experiment 1

Experiment 1 provides a first test case for investigating the generality of the Bayesian pronoun interpretation theory in Mandarin Chinese. We used a passage completion paradigm with a 2-by-2 design that allowed us to tease apart the influence of world knowledge-based inference that emerges from verb semantics, and the influence of information structural and grammatical factors by conditioning on the grammatical role of the referent in the analysis.

Methods

Participants We recruited 50 self-reported native Mandarin speakers over Witmart (a China-based online crowdsourcing platform). Each of them was paid $4 for their participation in the experiment.

Materials and procedures Participants completed two-sentence passages, writing a second sentence after a transitive-verb context sentence with two like-gender animate arguments. We used a Prompt Type by Verb Type design. Prompt Type has two levels: The Free condition (1) included no material in the second sentence prompt, allowing us to estimate next-mention preferences $P(\text{referent})$ by analyzing the first-mentioned referent in each condition, and the likelihood $P(\text{pronoun}|\text{referent})$ that a pronoun is produced to refer to that referent by analyzing the referential forms that participants used. The Pronoun condition (2) included an overt pronoun in the second sentence, allowing us to measure empirical pronoun interpretation preferences $P(\text{referent}|\text{pronoun})$.

The verbs in the first sentence were taken from one of two implicit causality (IC) classes (Garvey & Caramazza, 1974; Brown & Fish, 1983). The use of IC verbs allowed us to manipulate the prior, with IC-1 and IC-2 favoring the syntactic subject (NP_1) and the non-subject (NP_2) next-mentions respectively. A norming study (N=45) was conducted prior to Experiment 1 to ensure the verbs we selected have a clear bias towards re-mentioning either the subject or the non-subject. We selected sixteen subject-biased IC-1 verbs and twenty object-biased IC-2 verbs for the main experiments based on the results of the norming study.

(1) Meihui dadong-le Jieyi. ... [IC-1, Free]
Meihui impressed Jieyi. ...

(2) Meihui dadong-le Jieyi. Ta ... [IC-1, Pronoun]
Meihui impressed Jieyi. She ...

(3) Meihui jiegu-le Jieyi. ... [IC-2, Free]
Meihui fired Jieyi. ...

(4) Meihui jiegu-le Jieyi. Ta ... [IC-2, Pronoun]
Meihui fired Jieyi. She ...

Because IC verbs exhibit the most polarized effects on next-mention biases when the follow-on sentence provides an explanation of (i.e., a cause or reason for) the event described by the prompt sentence (Rohde, 2008), we instructed participants to limit their continuations to such explanations. Each participant completed 36 target items and 36 filler items with pseudo-randomization. Prompt Type varied within participants and within items; Verb Type varied only within participants. Each item was presented via the web interface on a separate page with a text box where the participants were instructed to write their continuations.

Coding The responses were coded by two trained native speakers, the first author and a UC San Diego graduate student who was blind to the hypothesis of the study. Each coder went through the responses independently to code the first-mentioned referent (or the assignment of the pronoun in the case of the Pronoun prompt condition) in each continuation as one of the five categories: *NP_1, NP_2, both, unclear,* and *other*. If the two coders did not agree on a reference, or there was not enough information available to identify the intended referent, the response was coded as *unclear*. For continuations in the Free prompt conditions, choice of referring expressions were coded as *name, overt pronoun, null pronoun,* and *other*.

Results and discussion

The analysis only included continuations for which the first-mentioned entity was NP_1 or NP_2; hence continuations were excluded if the referent was coded as *unclear* (3.8%), *both*

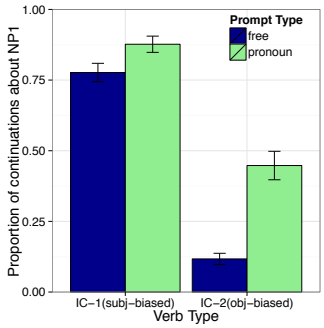


Figure 1: Proportion of continuations about the syntactic subject NP_1, by Verb Type and Prompt Type, in Experiment 1

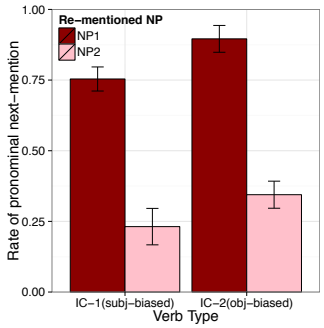


Figure 2: Rate of pronominalization by Verb Type and Re-mentioned NP in Experiment 1

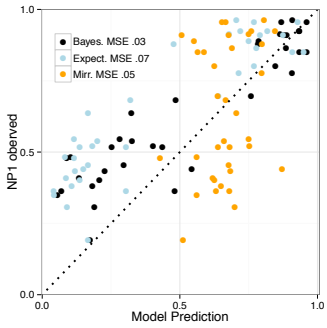


Figure 3: Quantitative model comparisons for Experiment 1

(0.8%), or *other* (2.7%). Continuations were also excluded if the choice of referring expression was other than a name, an overt pronoun, or a null pronoun (1.0%). After applying these restrictions, 1651 out of 1800 continuations in the dataset remained. All statistical analyses in this paper report results from mixed-effect logistic regression models with the maximal random-effect structure justified by the design (Barr et al., 2013), conducted using the `lme4` R package (Bates et al., 2015; R Core Team, 2015); we report significance levels based on the Wald z statistic. In cases where we encountered convergence failure with `lme4`, we report analyses carried out using the `MCMCglmm` R package (Hadfield, 2010), which we indicate by p_{MCMC} in reporting statistical significance. Error bars in all figures are standard errors over by-participant means.

The WEAK form of the Bayesian hypothesis predicts that pronoun interpretation biases and production biases follow Bayesian principles, without fully specifying what factors affect each component. The STRONG form of the hypothesis additionally predicts that semantic factors (Verb Type) affect only next-mention biases $P(\text{referent})$ and not pronoun production biases $P(\text{pronoun}|\text{referent})$; the primary factor affecting pronoun production biases is instead predicted to be the grammatical role of the referent (Re-mentioned NP). We test these predictions in the following analyses.

Next-mention biases Figure 1 shows the proportion of continuations about NP_1 in both Free prompt and Pronoun prompt conditions. We first evaluate the next-mention biases (dark blue bars) in the Free prompt data, which serves as the prior in the Bayesian model. Analyses showed a main effect of Verb Type ($p < 0.001$): subject-biased IC-1 items prompted significantly more continuations about NP_1 (77.7%) than object-biased IC-2 items (11.7%). This indicates that verb semantics had a strong effect on next-mention biases, an expected result on both the STRONG and WEAK forms of the Bayesian hypothesis.

Production biases Figure 2 shows the rate of pronominalization conditioning on Re-mentioned NP collapsing null and overt pronouns, which serves as the likelihood in the Bayesian model. Analyses showed a large main effect of Re-mentioned NP ($\beta = 1.37$, $p < 0.001$), with NP_1 re-mentions much more likely to be pronominalized than NP_2 re-mentions. There was no significant interaction between Re-mentioned NP and Verb Type ($\beta = -0.053$, $p = 0.716$). Both results align with the predictions of both the STRONG and WEAK hypotheses, indicating a clear disassociation between next-mention biases and production biases. There was also a smaller, unanticipated main effect of Verb Type ($\beta = -0.43$, $p < 0.05$), with pronominalization rates higher in the IC-2 context than in the IC-1 context. The reasons for this effect remain unclear.

Interpretation biases Now we examine the interpretation biases (the *posterior*) and compare them to next-mention biases (the *prior*). As shown in Figure 1, there was a main effect of Verb Type ($\beta = -2.021$, $p < 0.001$), with IC-1 verbs eliciting more NP_1 mentions than IC-2 verbs. There was a main effect of Prompt Type ($\beta = 0.896$, $p < 0.001$), with the proportion of continuations about NP_1 in the Pronoun prompt condition higher than that in the Free prompt condition. Analyses also showed a interaction between Prompt Type and Verb Type ($\beta = -0.408$, $p < 0.001$), whereby the effect of Pronoun prompts was larger for IC-2 verbs than IC-1 verbs. Given next-mention and production biases, the main effects of Verb Type and Prompt Type are predicted by the Bayesian theory (the interaction's presence was neither predicted nor precluded).

Quantitative model comparisons Following Rohde & Kehler (2014), we further evaluate the Bayesian account by comparing its predictions with those of two competing models. The first competing model is the Expectancy model, according to which the interpretation bias towards a referent is equal to the probability the referent gets mentioned next (Arnold, 2001). For this model, the predicted interpretation bias is estimated to be the prior next-mention bias. We express this model in Equation (M2) below, using the $\leftarrow$ assignment operator to emphasize that this model does not follow normative probability theory.

$$P(\text{referent}|\text{pronoun}) \leftarrow P(\text{referent}) \qquad \text{(M2)}$$

The second competing model we call the Mirror model, according to which the interpretation bias towards an entity is proportional to the likelihood that a reference to that entity would have been pronominalized by the speaker. This model captures the intuition that pronoun interpretation and production are effectively mirror images of each other, as expressed by the assignment in (M3).

$$P(\text{referent}|\text{pronoun}) \leftarrow \frac{P(\text{pronoun}|\text{referent})}{\sum_{\text{referent}\in\text{referents}} P(\text{pronoun}|\text{referent})} \quad \text{(M3)}$$

To determine the quantitative predictions of each model (M1)–(M3), we used the Free prompt data to estimate [condition+item]-specific prior and likelihood probabilities $P(\text{referent})$ and $P(\text{pronoun}|\text{referent})$, using add-one smoothing to avoid zero probabilities, and then compared these predictions to [condition+item]-specific human interpretation preferences from our Pronoun prompt data. Figure 3 plots observed NP_1 interpretation rates against item-specific predictions of the three models. The $x = y$ dotted line would be a perfect model fit. The Bayesian model had the least mean-squared error (0.03), indicating the Bayesian model is a better fit than either of the competing models in predicting pronoun interpretation. In comparison, the Mirror model dramatically underpredicts cross-item/condition variability in interpretation preference, because it lacks the influence of the world-knowledge-derived prior. The Expectancy model systematically underpredicts the rate of NP_1 pronoun interpretation, because it lacks the likelihood-derived bias toward NP_1 obtaining from the speaker's choice of pronominal form.

Experiment 2

Experiment 2 further tests the generality of the Bayesian pronoun theory by pursuing a set of findings regarding the effect of voice on pronoun behavior identified for English by Rohde & Kehler (2014). Recall that Rohde and Kehler hypothesized that a difference in syntactic form — specifically active vs. passive voice — would have an effect on production biases. This expectation is based on the STRONG Bayesian hypothesis, which posits that the likelihood of a speaker pronominalizing a mention of a referent is based on the referent's degree of topicality. Hypothesizing that being the subject of a passive clause is a stronger indicator of topichood than being the subject of an active clause, Rohde & Kehler (2014) thus predicted that the syntactic subject of a passive clause is more likely to be pronominalized than that of an active clause. The results of their study confirmed this prediction. However, the results also revealed a separate effect, unexpected under the STRONG hypothesis, whereby passivization not only affected the next-mention bias as well, but did so in the opposite direction that one would expect: Passivization increased the next-mention rate of the *logical* subject, that is, the entity mentioned from within a *by-* adjunct.

Here we evaluate the predictions of the models by employing a similar voice manipulation in Mandarin Chinese. The STRONG hypothesis predicts that the difference in information structure between active and passive voice will affect production biases, and in turn interpretation biases. A finding similar to Rohde and Kehler's whereby passivization increases the next-mention bias for the logical subject is not predicted by the STRONG hypothesis, however. The WEAK hypothesis makes no commitment to the either of these predictions. Instead, the WEAK hypothesis simply predicts that any change in the pronoun production biases or the next-mention biases would affect interpretation biases, in accordance with Bayes' rule.

Methods

Participants We recruited 71 self-reported native Mandarin speakers over Witmart. Each of them was paid $4 for their participation in the experiment. None of them participated in Experiment 1.

Materials and procedures The same passage completion paradigm as Experiment 1 was employed. In addition to the Verb Type × Prompt Type design, we included a Voice (active vs. passive) manipulation. Participants were asked to write a natural continuation given the prompt (unlike Experiment 1, participants were not limited to writing explanations). The verbs were identical with the ones in Experiment 1 with one exception: We replaced "anwei" (*comfort*) with "xiaokan" (*look down upon*) since the former sounds unnatural in a passive clause. Example stimuli are shown in (5)-(8).

(5) Meihui dadong-le Jieyi. (Ta) ... [IC-1, active]
Meihui impressed Jieyi. (She) ...

(6) Jieyi bei Meihui dadong-le. (Ta) ... [IC-1, passive]
Jieyi was_by Meihui impressed. (She) ...

(7) Meihui jiegu-le Jieyi. (Ta) ... [IC-2, active]
Meihui fired Jieyi. (She) ...

(8) Jieyi bei Meihui jiegu-le. (Ta) ... [IC-2, passive]
Jieyi was_by Meihui fired. (She) ...

Each participant completed 36 target items and 36 filler items with pseudo-randomization. Prompt Type and Voice varied within participants and within items; Verb Type varied only within participants. Participants wrote passage continuations using the same web interface as in Experiment 1.

Coding The same judges coded the responses using the same criteria as those in Experiment 1.

Results and discussion

Out of 2556 continuations, 8.7% were excluded because the next-mentioned referent was coded as *unclear, both,* or *other*, or the choice of referring expressions was other than a name, an overt pronoun, or a null pronoun, leaving a total of 2334 continuations in the dataset.

Next-mention biases Next-mention results are depicted in Figure 4 in terms of the proportion of continuations about the *logical object*, which more clearly illustrates the effect of the

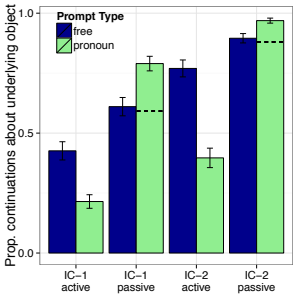


Figure 4: Prop. of continuations about the logical object, by Verb Type, Voice and Prompt Type, in Experiment 2

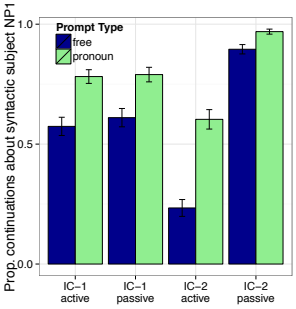


Figure 5: Prop. of continuations about the syntactic subject in Experiment 2 (representation of the data of Figure 4)

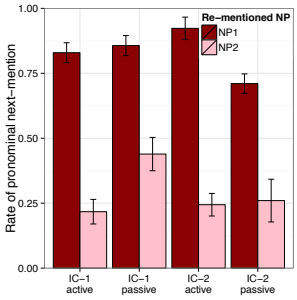


Figure 6: Rate of pronominalization, by Verb Type, Voice, and Re-mentioned NP, in Experiment 2

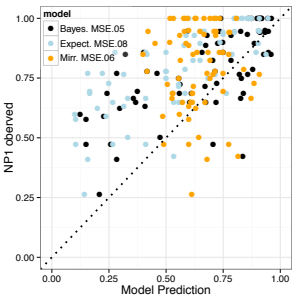


Figure 7: Quantitative model comparisons for Experiment 2

voice manipulation. The Free prompt condition (blue bars) showed a main effect of Verb Type ($\beta = -0.987,\ p < 0.001$), whereby IC-2 verbs elicited more continuations about the logical object than IC-1 verbs. Analyses also showed a main effect of Voice ($\beta = -0.477,\ p < 0.001$), whereby passivization increased continuations about the logical object. The effect of Verb Type is compatible with both the STRONG and the WEAK versions of the Bayesian hypothesis. However, the effect of Voice is only consistent with the WEAK version since the STRONG version predicts that information structure should only affect the likelihood.

Production biases Figure 6 presents the rate of pronominalization, collapsing both null and overt pronouns. Results showed a main effect of Re-mentioned NP ($p_{MCMC} < 0.001$), with NP_1 more likely to be pronominalized than NP_2 regardless of its semantic roles. There was no main effect of Voice ($p_{MCMC} = 0.38$) or Verb Type ($p_{MCMC} = 0.75$). The higher rate of pronominalization of NP_1 is compatible with both the STRONG and WEAK versions of the Bayesian hypothesis, as is the lack of effect of Verb Type. The lack of Voice effect is not compatible with the STRONG version, however, which predicts that the syntactic subject of a passive clause is more likely to be pronominalized than that of an active clause. These results instead suggest that grammatical role is the primary factor affecting pronoun production biases.

Interpretation biases To better illustrate the comparisons between next-mention biases (the prior) and interpretation biases (the posterior) in relation to NP_1 re-mentions, we graph the data from Figure 4 in terms of re-mentions of the syntactic subject in Figure 5. Analyses showed a main effect of Prompt Type ($p_{MCMC} < 0.001$), with Pronoun prompts being associated with more NP_1 re-mentions than Free prompts. Analyses also showed a main effect of Voice ($p_{MCMC} < 0.001$), with passive voice eliciting more NP_1 re-mentions than active voice, as well as a marginal effect of Verb Type ($p_{MCMC} = 0.063$), with IC-1 verbs eliciting numerically more NP_1 re-mentions collapsing Voice and Prompt Type. The Bayesian hypothesis predicts both the observed effect of Prompt Type, due to the likelihood term in Equation (M1), and the close relationship between the NP_1 re-mention rate in the Pronoun and Free prompt types across Verb Type and Voice conditions, due to the effect of the prior.

We have shown in Figure 4 (blue bars) that passivization increased next-mention rate for the logical object NP_2. The Bayesian hypothesis predicts that any effect of passivization on next-mention biases should have a corresponding effect on pronoun interpretation biases. The dashed lines in Figure 4 illustrate the NP_1 re-mention rate that would be predicted for passives under an alternative model that is identical to M1 except that the normatively correct prior term $P(\text{referent} = NP_1|\text{passive})$ is replaced with the active-voice prior term $P(\text{referent} = NP_1|\text{active})$. This alternative model underpredicts the observed rate of NP_2 re-mention, showing that the passive voice's effect of increased next-mention prior bias toward the logical object indeed manifests in pronoun interpretation preferences.

Quantitative model comparisons We again used the Free prompt data to estimate the interpretation biases of overt pronouns for each item, and then compared the predictions to the actual interpretation data measured in the Pronoun prompt condition. Figure 7 plots observed NP_1 interpretation rates against item-specific predictions of the three models. The $x = y$ dotted line would be perfect model fit. As also seen in Experiment 1, the Mirror model underpredicts cross-item variability in interpretation preferences, and the Expectancy model systematically underestimates the likelihood-derived preference toward NP_1. Once again, the Bayesian model had the least mean-squared error (0.05) of the three models.

General Discussion

The results from Experiment 1 showed that verb semantics had a strong effect on next-mention biases, where IC-1 verbs elicited more mentions of the subject than IC-2 verbs. The results also showed that the grammatical roles of potential referents had a significant impact on the likelihood of producing a pronoun given a particular referent, with re-mentions of

the subject of the previous sentence (NP_1) more likely to be pronominalized than re-mentions of the non-subject (NP_2). Importantly, this held even in the object-biased IC-2 conditions. This and the lack of interaction between Verb Type and Re-mentioned NP in Figure 2 reveal a disassociation between the factors that determine what entities get mentioned next versus those that determine whether those mentions get pronominalized. When interpreting an ambiguous pronoun, the comprehender takes into account both the prior probability of a referent being mentioned next in the context and the likelihood that the speaker would choose a pronoun to mention that referent. As illustrated in Figure 1, the presence of an overt pronoun boosted NP_1 mentions in both IC-1 and IC-2 contexts compared to the Free prompt condition. This indicates the effect of the likelihood (the probability of NP_1 being realized pronominally) on pronoun interpretation when the prior next-mention probability is kept constant.

The results from Experiment 2 show that passivization increases the rate of next-mention of the logical object in Mandarin Chinese, reflecting the affectedness of the logical object in the passive *bei* construction. Results again showed that occupying the subject position increases the likelihood of pronominalizing the re-mention of an entity. On the other hand, we found no consistent evidence showing that re-mentions of passive subjects are more often pronominalized in Mandarin than re-mentions of active subjects, unlike what Rohde & Kehler (2014) found in English.

Our results indicate that pronoun interpretation biases do not straightforwardly mirror expectations about next-mention, nor do they simply mirror pronoun production biases. Rather, pronoun interpretation biases reflect the joint influence of next-mention biases and production biases, where the comprehender reasons about the states of the world as well as the speaker's linguistic choices in conveying those states. Comparisons between the full Bayesian model and two reduced variants showed that the full Bayesian model best predicted human behavior in pronoun interpretation, lending quantitative support for the Bayesian theory. These findings are broadly consistent with previous studies in English (Kehler & Rohde, 2013; Rohde & Kehler, 2014), but also show that cross-linguistic variation in effects of grammatical voice on production preferences is mirrored in interpretation. Overall these results lend both qualitative and quantitative support to a cross-linguistically general Bayesian theory of pronoun interpretation.

Acknowledgments

We gratefully acknowledge valuable feedback from members of the UCSD Computational Psycholinguistics Laboratory, and support from NSF grant IIS-0953870 and BCS-1456081, and NIH grant 1R01HD065829 to RPL.

References

Arnold, J. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *21*(2), 137-162.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*(3), 237–273.

Crawley, R. A., Stevenson, R. J., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*(4), 245–264.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, 459–464.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*(6), 699–717.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive Science*, *3*, 67–90.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44.

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1–2), 1–37.

LaPolla, R. J. (1988). Topicalization and the question of lexical passives in Chinese. In M. Chan & T. Ernst (Eds.), *Proceedings of the Third Annual Conference on Chinese Linguistics* (pp. 170–188). Indiana University Linguistics Club, Bloomington, IN.

Li, C. N., & Thompson, S. A. (1981). *Mandarin chinese: A functional reference grammar*. Univ of California Press.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rohde, H. (2008). *Coherence-driven effects in sentence and discourse processing*. Unpublished doctoral dissertation, University of California, San Diego.

Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927.

Smyth, R. (1994). Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research*, *23*(3), 197–229.

Leveling the Field: Talking Levels in Cognitive Science

Luke Kersten (lukekersten@cmail.carleton.ca)
Institute of Cognitive Science, 1125 Colonel By Drive,
Ottawa, ON K1S 5B6 CA

Robert West (robert_west@carleton.ca)
Institute of Cognitive Science, 1125 Colonel By Drive,
Ottawa, ON K1S 5B6 CA

Andrew Brook (andrew.brook@carleton.ca)
Institute of Cognitive Science, 3A40 Paterson Hall,
Ottawa, ON K1S 5B6 CA

Abstract

Talk of levels is everywhere in cognitive science. Whether it is in terms of adjudicating longstanding debates or motivating foundational concepts, one cannot go far without hearing about the need to talk at different 'levels'. Yet in spite of its widespread application and use, the concept of levels has received little sustained attention within cognitive science. This paper provides an analysis of the various ways the notion of levels has been deployed within cognitive science. The paper begins by introducing and motivating discussion via four representative accounts of levels. It then turns to outlining and relating the four accounts using two dimensions of comparison. The result is the creation of a conceptual framework that maps the logical space of levels talk, which offers an important step toward making sense of levels talk within cognitive science

Keywords: levels; analysis; explanation; organization; cognitive science; conceptual framework

Introduction

Levels are everywhere in cognitive science. One cannot go far without hearing about the need to talk at different levels, whether it is in terms of levels of organization, explanation, description or aggregation. Levels have been called upon to do everything from adjudicate longstanding debates in cognitive modeling (Broadbent, 1985; Dawson, 1998) to motivating computational theories of mind (Pylyshyn, 1984).

Yet in spite of its widespread application and use, the concept of levels has received little sustained attention in cognitive science. Although there has been excellent wider discussion in the philosophical literature about levels (see, e.g., Wimsatt, 1976; 1994), little has been done to develop the notion specifically within cognitive science – for early attempts see Bechtel (1994) and McClamrock (1991). As Wright and Bechtel (2007) point out: "levels-talk is virtually threadbare from overuse yet [the] various conceptions of levels are rarely analyzed in any sustained, substantive detail despite there being a litany of literature on the subject" (p.55). Or, as Craver puts it: "Despite the ubiquity of levels talk in contemporary science and philosophy, very little has been done to clarify the notion" (2015, p.23).

This state of affairs calls out for improvement. In what follows, we aim to provide an analysis of the various ways levels have been deployed within cognitive science. The paper begins by introducing and motivating discussion via four representative accounts of levels. It then turns to outlining and relating the four accounts using two dimensions of comparison. The result is the creation of a conceptual framework that maps the logical space of levels talk. This conceptual framework offers one important step toward making sense of levels within cognitive science.

Talking Levels

The first account to consider is David Marr's (1977, 1982). Marr identifies three different "levels of analysis" for cognitive science.

First, there is the computational level. At the computational level, investigators look at what function a system performs, asking questions about what information-processing problem the system solves. Research at the computational level aims to translate general, everyday descriptions of cognitive phenomena into particular information-processing problems or tasks. Second, there is the algorithmic level. At the algorithmic level, researchers investigate by what steps a system solves an information-processing problem; they ask questions about the algorithms and representations used by the system. Research at this level attempts to specify in detail the set of information-processing procedures that solve a particular information-processing problem. Finally, there is the implementational level. At the implementational level, researchers attempt to determine what physical structures instantiate the algorithms used for solving the information-processing problem; what physical mechanisms realize or support the cognitive system under investigation.

For example, when investigators emphasize the spatiotemporal properties of cognitive systems using the techniques of neuroscience, such as neuroimaging or lesion studies, they operate at an implementational level. When investigators focus on how components interact so as to

produce the operations or procedures that carry out computations using the methods and techniques of cognitive psychology such as randomized block designs or error rates, they ascend to a higher, algorithmic level of analysis. Each level of analysis has a different role to play in the cognitive investigation, as each satisfies a different epistemic end.

The second account to consider is Zenon Pylyshyn's. Pylyshyn (1980, 1984) similarly identifies three "levels of description" for cognitive science, though where Marr is concerned with offering a unifying methodology for cognitive investigation Pylyshyn is concerned with articulating the foundations of cognitive science.

First, there is what Pylyshyn identifies as the "semantic" level. At this level, psychological behaviour is described and explained using the representational or semantic content of an individual's mental states; what the individual believes, desires, etc. Second, there is the "symbolic or syntactic" level. Behaviour at the symbolic level is described in terms of functional properties. For example, to explain why an individual does poorly on a memory-recall task, the symbolic level appeals to control structures or memory storage capacities. This is in contrast to explaining performance by referencing what information the individual fails to recall, for example. Third, there is the "biological level". At the biological level, individuals are described in terms of the familiar vocabulary of the physical sciences, e.g., neurology, chemistry, biology. To account for why someone is thinking of faces rather than sounds, for example, explanations at the biological level might appeal to differences in brain activity or neural-chemistry.

Similar to Marr, Pylyshyn (1980, 1984) is also impressed by the explanatory power of levels. However, whereas Marr's levels are concerned with answering different types of questions, Pylyshyn's levels are focused on capturing distinct generalizations. For Pylyshyn, different descriptive vocabularies capture distinct sets of regularities. He writes, for instance: "When we have principles of operation that cannot be stated within a certain vocabulary – but can be captured in another, more abstract (here, functional) vocabulary to which the terms of the first vocabulary stand in multiple relations – we have a prima facie case for the existence of an independent level" (1984, p.33). Some valid generalizations are only expressible at particular levels of description.

Newell (1980, 1990) offers a third account. Newell defines what he calls "systems levels". System levels are collections of components that, in virtue of their organization and interaction, produce particular functions or behaviours. System levels mark functional divisions between different sets of organized components. They define the basic "technology" by which the human cognitive architecture is constructed. Each system level is realized by components at the next system level below.

Furthermore, since complicated systems are more likely to be resistant to degradation if they are built out of assemblies of stable subcomponents, it is more likely that system levels will be successively layered. As Newell puts it: "If stable subassemblies are created, layer upon layer, then each one has a reasonable probability of being constructed out of a few parts. Thus, there exists a general argument that stability dictates the existence of levels" (1990, p.117).

In terms of identifying system levels, Newell claims that as one ascends upward from the smallest components to the largest components, system levels are revealed by their unique time signatures; as one moves up the hierarchy of system levels the size of each level will increase while speed will decrease. This follows in virtue of the aggregative size of levels. If components of one system level are of a characteristic size and those components are put together to form components at the next system level above, then it follows that the higher-level components will take longer to operate than their constitutive elements.

Methodologically, the implication is that the time required to perform different tasks exposes different system levels. Different system levels are composed of increasingly larger components that operate at successively slower speeds. Qualitative shifts in time signatures reflect substantive shifts between different system levels.

The final account to consider is William Bechtel's (1994, 2007). Bechtel's use of levels is a bit different from the preceding three, as it is formed within a larger discussion of mechanistic explanation. An explanation qualifies as mechanistic when it identifies a hierarchical system whose components, in virtue of their organization, produce some activity or behaviour. The goal of mechanistic explanation is to decompose a given mechanism, and its constitutive activities, into its underlying component parts, showing how those components conspire to produce the activity of the composite whole (see, e.g., Bechtel & Richardson, 1993; Craver, 2007). Rather than viewing levels as substantive divisions between different systems (as Newell does, for example), Bechtel prefers to view levels as local fields of analysis. Levels pick out the various explanatory strategies that can be used during mechanistic decomposition. As Bechtel writes: "[m]ultiple cycles of analysis thus give rise to a hierarchy of levels that is confined to a given mechanism" (2007, p.56).

Bechtel (2007) offers the following example. Suppose a biological mechanism involves sodium molecules crossing over a cell membrane. On the mechanistic account, the sodium molecules and cell membrane can be said to be at the same level if both are implicated in the operation of the biological mechanism in question. The status of standing at higher or lower level depends, in part, on the explanatory role occupied by the component within the larger investigation – in contrast to, for example, slotting into a global organization of size, shape or motion. If the investigator pursues another cycle of decomposition, the components might be further decomposed, but this will still be local to the analysis being offered.

On Bechtel's analysis, it is not to possible to say that one mechanism exists at a higher or lower level than another.

There is no a global ordering available for mechanisms. Levels simply define the various explanatory strategies employed at any one cycle of decomposition. The existence of levels is mechanism dependent. As Bechtel puts it: "[L]evels on the mechanistic account are real in that they deal with the particularities of actual components and their operations, but they are perspectival in that they are defined with respect to specific foci on mechanistic activities (Wright & Bechtel, 2007, p.57).

These four accounts should provide a flavor of the kind of levels talk that has prevailed in much of cognitive science. As one can see, the range is quite large. Just what a level is, how it is to be used, and why such talk is important varies notably from author to author. The question to address is how to make sense of these various uses of levels.

Analyzing Levels

There are two general points of comparison or dimensions of analysis that can be used to examine levels talk. The first deals with the types of items that figure into an account of levels; what 'relata' the view implicates. The second deals with how an account uses levels talk; to what end or purpose an account deploys levels. Consider each in turn.

First, there seem to be two general types of relata used in levels talk. The first is ontic relata. On this view, levels talk is taken to hold of activities, structures, and properties. It applies to cognitive entities as they appear *in the world*. Newell offers something like this application with his system levels account: "[A] system level is a property of nature, and not just something in the head of the observer" (1990, p.118). On the ontic application, levels are part of the configuration of the world, picking out the various features constitutive of cognition.

The second type of relata is epistemic relata. On this view, levels talk is taken to hold of 'linguistic' or 'theoretical' entities. Marr is reasonably interpreted as subscribing to something like this view. When applied to epistemic relata, levels talk functions to provide perspectives for understanding cognition. It offers a means of enhancing an investigator's epistemic situation.

When viewed as extremes, these two general applications form a continuum on which various accounts can be located. Figure 1 provides an illustration.

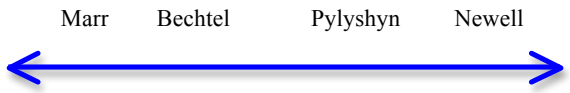


Figure 1: Levels accounts arranged along a spectrum of commitment to epistemic vs. ontic relata.

On the one hand, there are those views such as Marr's that apply levels talk to epistemic relata. On the other hand, there are also those views such as Newell's that apply levels talk to ontic relata. An account's positioning depends on the relata it implicates.

Notice that Bechtel's view provides something of an interesting middle ground, as it straddles the two poles. On Bechtel's view, levels of mechanisms are both ontic structures and observer-dependent entities. Mechanistic activities are real phenomenon in the world, but their arrangement into different levels depends on the explanatory goals of the investigation.

An additional implication of applying levels talk to epistemic versus ontic relata is that it changes how levels relate within a given account. If, on the one hand, an account applies to ontic relata, then its levels relate via a metaphysical realization relation. If, on the other hand, the account applies to explanatory or theoretical entities, then its levels relate in virtue of an explanatory realization relation.

Metaphysical realization is the familiar notion from philosophy of mind that says that two entities stand in a realization relation if and only if when there is change in the higher-level entity there is a corresponding change in the underlying entity change, but not the reverse (Kim, 1998). The standard example is the relation between mental and physical states. Since there cannot be change among mental states without a corresponding change among physical states, physical states are said to realize mental states. Metaphysical realization involves a determinative, asymmetric dependence relation between different properties, states or activities.

Explanatory realization, on the other hand, has received a little less attention than its ontological cousin. In explanatory realization, the dependence relation tracks items of particular theories, languages, or models. It addresses explanatory rather than metaphysical entities. The realization relation holds between statements rather than entities.

An illustrative example comes from theory reduction in the philosophy of science. On the standard model of theory reduction, one theory reduces to a second if and only if the first theory is derivable from the second given certain "bridge laws" (Nagel, 1961). Reduction occurs when all of the items of one theory can be translated into items of the second theory via specific law-like statements or logical connectives. For example, the theory of thermal conductivity reduces to electrical conductivity because (most) of the terms in thermal conductivity can be translated into terms of electrical conductivity via covering generalizations such as the Wiedemann-Franz Law.

Whether or not the reduction is successful is beside the point. The point to note is that the supportive relation holds between different theories or linguistic items and not ontological structures. Explanatory realization shifts attention to how descriptions of entities within theories sustain or determine the theoretical fruitfulness of entities described at other levels.

On the epistemic application, hierarchies of levels stand in similar sorts of explanatory realization relations. This is

why both Pylyshyn and Marr emphasis the collective use of several levels. On Marr's account, for example, the computational level identifies a given input-output mapping, while the algorithmic level describes the procedures by which the function is computed. The algorithmic level supplies crucial constraints on satisfying the input-output mapping. The algorithmic level makes it possible for the computational level to track real computations in the world.

Second, different accounts of levels can also be measured to the extent to which they use levels talk for epistemic versus ontic or metaphysical purposes.

On the ontic usage, levels talk attempts to provide descriptions of how cognition is organized. Levels talk is in the business of providing accurate descriptions of how the world is organized. Discussion attempts to furnish representative accounts of how cognitive systems and capacities are structured into levels.

Consider Newell's account, for example. On Newell's view, different "cognitive bands" address different sets of "system levels" – for example, the biological band addresses the neural circuit system level, while the cognitive band deals with the deliberate act system level. On Newell's view, levels talk functions to describe the organization and structure of cognition as described by the different system levels. Thus, according to the ontic usage, levels talk has a distinctly ontic role to play in discussion of cognition.

Contrast this position with a second, epistemic type of usage. On this usage, levels talk deals with how cognitive phenomena should be studied. Levels function to provide perspectives or viewpoints from which to investigate cognition. As Dawson puts it: "levels do not attempt to explain the nature of information processing devices, but instead provide an epistemology – a way to inquire about the nature of the world" (2013, p.53). Levels talk acts as a tool for inquiry rather than as an ontic description of cognition.

Consider Marr's account, for example. On Marr's view, the computational, algorithmic and implementational levels help to organize cognitive investigation into its most explanatory fruitful parts. The computational level deals with function of the cognitive system, the algorithmic level deals with the procedures used to carry out said function, and the implementational level deals with physical material in which the computation is instantiated. Each level partitions investigation into distinct units of analysis such that it structures and simplifies discussion.

When viewed as extremes, these two general usages form a second continuum on which various accounts can be located. Figure 2 provides an illustration.

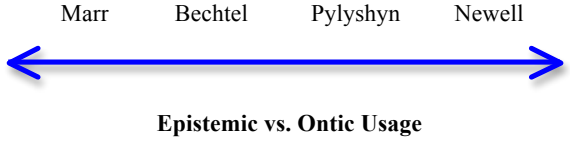


Figure 2: Levels accounts arranged along a spectrum of commitment to explanatory versus ontic usage.

On the one hand, there are those views such as Marr's that deploy levels talk for explanatory purposes – these accounts view levels talk as a tool for inquiry. On the other hand, there are those views such as Newell and Pylyshyn's that use levels talk for metaphysical or ontic purposes – these accounts view levels talk as a means to describe cognitive structures as they appear in the world. Different accounts employ levels talk to different ends.

A Conceptual Framework for Levels

The fact that levels talk can be sorted along two dimensions of analysis allows for the creation of a conceptual framework that can be used to the map the logical space of levels talk. Figure 3 provides an illustration.

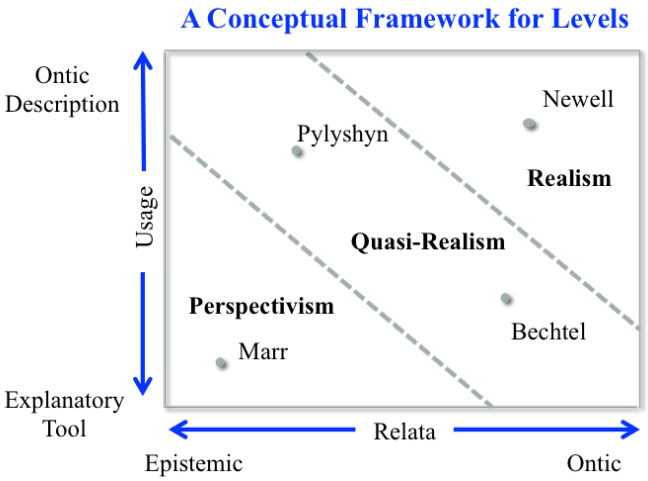


Figure 3: Levels views compared along commitment to ontic versus epistemic relata and usage.

Marr's account occupies the bottom left corner of the space. This is because it applies levels talk to exclusively theoretical or linguistic entities. Both the relata and usage are taken to be explanatory in nature. Marr's account is followed by Pylyshyn and Bechtel's accounts in the middle. These views tow a more neutral line, conceiving of levels as at once ontic and epistemic in character to some degree – Pylyshyn's view, for example, applies to epistemic relata but is used to make claims about the leveled structure of cognition. Finally, there is Newell's view in the top right hand corner of the space. Newell's position here is owed to his adherence to almost entirely ontic application of levels talk – levels talk applies to ontic relata (system levels) and is used to describe the structure of cognitive architecture in terms of the system levels.

What is interesting about this framework is that it maps many of the possible applications of levels talk. It is both descriptive and generative. Various accounts are positioned with respect to how they talk about levels, but at the same time the framework leaves it open as to how other accounts might use levels.

Now one might worry that the framework uses two dimensions of analysis where only one is necessary, particularly given that both points of comparison include an ontic versus epistemic element. The reason two dimensions are more effective than one is that two dimensions allow the current framework to distinguish positions, such as Bechtel or Pylyshyn's, that deploy level talk both epistemically and ontically. Using only one dimension would fail effectively convey the difference between positions such as these, as Pylyshyn's view implicates epistemic relata but makes ontic claims and Bechtel's view implicates ontic relata but uses levels as an epistemic tool.

One might further worry that the present analysis focuses too much on "first generation" cognitive science. That it does not pay sufficient attention to neurological or biological levels. Though this objection is well motivated, it is misplaced in present context. Nothing in the current analysis precludes discussion of neurological or biological levels. It is simply that neurological or biological levels have not featured prominently in discussion of levels – for discussion, see Craver (2015). Although it is certainly right to say that cognitive processes, states, and systems are abstract in the sense that they can be discussed without reference to material form but concrete in the sense that they are realized in a specific physical medium (Thagard, 2014), the current analysis is neutral with respect to the importance of this fact. The current analysis simply attempts to map the space of levels talk as it has been most prominently developed within cognitive science. Whether or not neurological levels are crucial to understanding cognition is of secondary concern to the present investigation.

The framework also reveals three more general, orienting positions that might be held with respective to levels. The first is what might be called the "perspectivist or perspectival" view. On the perspectival view, levels talk applies to the products or units of scientific investigation. Levels deal with theories, explanations or research programs rather than ontic structures. Different epistemic structures are sorted according to the different questions they address. Items on the perspectival view relate in virtue of standing in explanatory rather than ontic relations – recall, for example, what was said about explanatory versus metaphysical realization. Included here are views such as Marr's.

The second is what might be called the "realist" view. According to this view, levels talk applies to items *in the world*. This means that levels deal with entities, activities and properties. These items exist as part of the furniture of cognition. These structures relate via compositional relations. Each item is constituted or realized by items residing at lower levels. Often, a principle of prediction or causality is used to sort items into different levels. Newell is included under this second heading.

Finally, there is the "quasi-realist" view. On this view, levels talk applies to the components and activities of mechanisms. These are features of the world. However, in contrast to realism and closer to perspectivism, levels also have an epistemic bent. Though levels count as real parts of cognition, they only emerge during or as part of investigation. As Bechtel writes: "Constitutive strategies describe the mechanism's component parts, their operations, and their organization, showing how the mechanism's constituency is responsible for its activity" (2007, p.62). Levels are theoretical impositions, being explanatory strategies, but they are also part of the mechanisms constitutive of cognition. Bechtel falls under this third category.

Note that these three views are probably best regarded as representative positions rather than actual positions individual authors' hold. Each view represents a sort of general umbrella category or idealized position that can be used to explore different types of levels talk without getting into all the messiness of interpreting individual authors' viewpoints.

What is it about the realist, quasi-realist, and perspectivist positions that make other views cluster around them? One possibility is that the divisions reflect larger differences in overarching conceptions of cognition or science (see, e.g., van Fraassen, 1980). That the realism/anti-realism split reflects a larger divide that runs through most theoretical discussions in the sciences. Another possibility might be that the various views are just different sides of the same coin; that with additional work each can be distilled into one underlying view. For example, perhaps Marr's more perspectivist view collapses into the realist view if it is cashed out in specifically causal terms. Or, perhaps the claims of the realist can be deflated and shown to be subsumable under the quasi-realist's account.

There are three desirable features to the present framework. First, it helps to clarify confusions between different accounts of levels. For example, in their review of Marr's tripartite account of levels, Churchland, Koch, and Sejnowski write: "when we measure Marr's three levels of analysis against levels of organization in the nervous system, the fit is poor and confusing" (1990, p.38). For Churchland, Koch and Sejnowski, Marr's levels are either inadequate or patently false. They fail to offer accurate characterizations of the organization of cognition as it realized in the brain. Since there are a great many spatial levels of organization in the brain, Marr's levels makes for a poor fit.

Understood within the context of the present analysis, the issue is immediately clear: Churchland, Koch and Sejnowski have mistaken the relata of Marr's account. Marr's levels are not a poor fit for brain organization, because they are not meant to fit with brain organization. Churchland, Koch and Sejnowski have confused Marr's epistemological application of levels talk with an ontological one. In cases such as these, knowing what elements to look for can help identify key areas where potential misunderstandings might arise. The framework can play a therapeutic role in discussion.

Second, the conceptual framework furnishes interpretative benefits. This is because it allows the underlying reasoning of each account to be more easily

understood. For example, on Newell's view, evolution is thought to favor systems constructed through addition of modular parts, since it is more likely that nature will build complex systems incrementally rather than all at once. Understood in realist terms, this reasoning makes sense. Attention to the spatial and temporal features of levels weds naturally to a focus on evolutionary considerations. Once it is clear at what end each account aims, the underlying reasoning is more easily explicated.

Third, in the case of more substantive disagreements over levels, the framework helps to clarify different lines of support, and thus chart potential argumentative paths through the levels space. For example, returning to Newell's account, if one is dubious of the appeal to evolution on the grounds that such arguments are underdetermined, then one can argue that the realist position is importantly under-motivated; that application of levels talk to cognitive structures is too quick. Or, to take the contrary position, if one is concerned that the perspectival view relegates levels talk to an arbitrary status, then the current analysis provides a direction in which to mount this challenge – for example, that it fails to get at the real structures and organization of cognition. The framework helps to simply discussion by cutting to the core elements of disagreement.

For all these reasons, the present framework offers an important tool for cognitive science. It provides a means to thinking about and sorting through various types of levels talk. It also reveals key elements of the various accounts while staying neutral with respect to alternative ways of applying levels talk.

In closing, a quick qualification needs to be made. The preceding analysis is not meant to be the definitive guide to conceptualizing levels. It is entirely possible that other dimensions of analysis exist on which applications of levels talk can be analyzed. However, the present account does identify several important features that cut across a large swath of levels talk. What the present discussion has provided, we hope, is one concrete step toward making of further sense of levels in cognitive science.

References

Bechtel, W., & Richardson, R. (1993). Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. Princeton, NJ: Princeton University Press.

Bechtel, W. (1994). Levels of Description and Explanation in Cognitive Science. *Minds and Machines*, 4, 1-25.

Broadbent, D. (1985). A question of levels: Comments on McCelland and Rumelheart. *Journal of Experimental Psychology: General*, 114, 189-192.

Churchland, P.S., Koch, C., & Sejnowski, T.J. (1990). What is computational neuroscience? In Schwartz, E.L. (ed.), *Computational Neuroscience*. Cambridge, MIT Press.

Craver, C. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford, UK: Clarendon Press.

Craver, C. (2015). Levels. In T. Metzinger & J. M. Windt (Eds). Open MIND: 8(T). Frankfurt am Main: MIND Group. doi: 10.15502/9783958570498

Dawson, M. R. W. (1998). *Understanding Cognitive Science*. Blackwell Publisher.

Dawson, M. R. W. (2013). *Mind, Body, World: Foundations of Cognitive Science*. Alberta, Edmonton: Athabasca Press.

Dawson, M. R. W., Boechler, P. M., & Valsangkar-Smyth, M. (2000). Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgements. *Spatial Cognition and Computation*, 2, 181–218.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Marr, D., & Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*, 15, 470-488.

Mcclamrock, R. (1991). Marr's Three Levels: A Re-Evaluation. *Minds and Machines*, 1, 185-196.

Nagel. T. (1961). *The Structure of Science. Problems in the Logic of Explanation*, New York: Harcourt, Brace & World, Inc.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4, 66–83.

Newell, A. (1990). *Theories of Unified Cognition*. Cambridge, MA: Harvard University Press. Pinker, S. (1997). *How the Mind Works*, New York: W. W. Norton & Company.

Pylyshyn, Z. (1980). Cognition and computation: Issues in the foundation of cognitive science. *Behavioral and Brain Sciences*, 3(1), 154-169.

Pylyshyn, Z. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Wimsatt, C. (1976). Reductionism, Levels of Organization, and the Mind-Body Problem. In G. Globus, G. Maxwell & I. Savodnik, *Consciousness and the Brain* (pp. 205-267), Spring Us.

Wimsatt, C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. Canadian Journal of Philosophy, 20, 207-274.

Wilson, R., & Craver, C. (2007). Realization: Metaphysical and Scientific Perspectives. In P. Thagard (ed.), *Philosophy of Psychology and Cognitive Science*, Amsterdam: Elsvier, pp. 81-104.

Wright, C.D., and Bechtel, W. (2007). Mechanisms and Psychological Explanation. In P. Thagard (ed.), *Philosophy of Psychology and Cognitive Science* (pp. 31-77). Amsterdam: Elsvier.

Know Your Enemy: Applying Cognitive Modeling in Security Domain

Vladislav D. Veksler (vdv718@gmail.com)
DCS Corp, U.S. Army Research Laboratory
Aberdeen Proving Ground, MD, USA

Norbou Buchler (norbou.buchler.civ@mail.mil)
U.S. Army Research Laboratory
Aberdeen Proving Ground, MD, USA

Abstract

Game Theory -based decision aids have been successfully employed in real-world policing, anti-terrorism, and wildlife conservation efforts (Tambe, Jiang, An, & Jain, 2013). Cognitive modeling, in concert with *model tracing* and *dynamic parameter fitting* techniques, may be used to improve the performance of such decision aids by predicting individual attacker behavior in repeated security games. We present three simulations, showing that (1) cognitive modeling can aid in greatly improving decision-aid performance in the security domain; and (2) despite the fact that individual attackers will differ in initial preferences and in how they learn, model parameters can be adjusted dynamically to make useful predictions for each attacker.

Keywords: cognitive modeling; game theory; behavioral game theory; strategy selection; agent simulation; model tracing

Introduction

Game Theory (GT) focuses on mathematical models of rational decision-making. In recent years, GT-based decision-aiding software has received significant attention for success in real-world security domain problems, such as scheduling patrols conducted by the US Coast Guard at multiple major US ports (Shieh et al., 2012), scheduling police patrols at major airports such as LAX (Pita et al., 2008), allocating federal air marshals on flights of US Air Carriers and several other applications (Tambe, 2011; Tambe et al., 2013). Success of GT approaches can be further improved by dropping the assumption that humans are optimally rational decision-makers, and by using cognitive modeling to predict adversary strategy selection.

Humans are not perfectly rational, rather, we are boundedly rational (Simon, 1972). Our ability to make optimal decisions is limited by available information, available time, and a myriad of cognitive constraints. There is a body of literature describing biases in human decisions (e.g., Kahneman, 2011), cultural preferences (e.g., Sample, 2015), and cognitive process interactions (e.g., Anderson, 2007) that can aid in predicting attacker decisions in real-world security problems.

Behavioral game theory is a modification of rational game theory informed by, "experimental evidence and psychological intuition" (Camerer, 2003, p. 465). Ultimately, the goal of behavioral game theory is to predict behavior and inform decisions in real-world strategic situations (Gächter, 2008). Whereas the success of normative game theory in security domain comes from providing efficient randomization of security plans and processes, behavioral game theory provides a more realistic view of human strategy selection based on a large body of empirical evidence, and argues for use of behavioral/cognitive models to predict human behavior.

Cognitive modeling is a method for predicting behavior based on known cognitive processes and biases. Computational cognitive models take the form of software that simulates human decisions on a given task. Computational cognitive models have been employed to account for game play in Prisoner's dilemma (Lebiere, Wallach, & West, 2000), rock-paper-scissors (West, Lebiere, & Bothell, 2006), and a collaborative foraging Geo Game (Reitter & Lebiere, 2011).

In this paper, computational cognitive models are employed to predict human behavior in security games. The rest of this paper describes a normative game theory approach to decision-making in the security domain, and suggests an alternative approach that employs cognitive modeling for selecting the best strategy in response to individual attacker's evolving preferences. We present three simulations highlighting the advantages of using cognitive modeling over normative game theory, and examine the use of model tracing and dynamic parameter fitting for predicting individual attacker's strategy selection.

Game Theory Approach to Decision Aids

Tambe et al. (2013) describe several successful applications of game theory decision aid software in real-world security games. Airport security, coast guard, and police officers employ this software to patrol for criminal activity. Animal preservation patrols are aided with this software in their efforts to control poaching. Tambe et al. (2013) focus on Stackelberg security games where the defender must perpetually defend a set of targets with limited resources, and the attacker can choose to attack a given target after observing defender actions. The general idea of picking an optimal mix of actions (i.e. mixed strategy) for the defender so as to decrease the chances of a successful attack applies across a much wider context (e.g., sports, cyber security, anti-terrorism).

Game theory suggests that the defender's mixed strategy should be a distribution of actions that removes any incentive for the opponent to choose one action over another. For example, imagine a simple game where there are only two possible actions for the defender, $D1$ and $D2$, and two possible actions for the attacker, $A1$ and $A2$. Let us assume that attacker payoffs are probabilities of a successful attack, these probabilities/payoffs being as listed in Table 1 (when the de-

fender chooses $D1$, the probabilities of attacker success for actions $A1$ and $A2$ are .2 and .6, respectively; when the defender chooses $D2$ these probabilities are .5 and .3, respectively). In this scenario, if the defender always chose $D1$, a rational attacker would always play $A2$, winning 60% of the time. If the defender always chose $D2$, the attacker would always play $A1$, winning 50% of the time. If the defender played randomly, the attacker would have the incentive to always play $A2$, winning 45% of the time.[1] A GT-computed mixed strategy in this game would be for the defender to play $D1$ one third of the time, and $D2$ two thirds of the time. This would leave the attacker with no preference for either option: playing either $A1$ or $A2$ would only lead to a successful attack 40% of the time.

Table 1
Sample payoffs for the attacker in a security game, where there are only two possible actions for the defender, D1 and D2, and two possible actions for the attacker, A1 and A2.

	A1	A2
D1	.2	.6
D2	.5	.3

In real world security games there are many more actions, and payoffs must take into account much more than success/failure. For example, Kar, Fang, Fave, Sintov, and Tambe (2015) describe a scenario where the attacker, an animal poacher, is drawn not only by the success of a poaching effort, but also by animal density and travel time. That is, when humans play the security game as an attacker (i.e. animal poacher) they are more likely to choose an action that leads to less travel time and higher animal density, even when the risk of failure (capture) is high.

Although animal density and travel distance are characteristics of the task environment, these factors are also latent indicators of cognitive biases. Direct consideration of such cognitive biases and limitations in formal attacker models should improve the performance of decision aids in the security domain.

Cognitive Modeling Approach

There are many computational models that provide robust predictions of human behavior. For example, models of reinforcement learning provide robust accounts of human trial-and-error behavior and its neural correlates (e.g., Anderson, 2007; Fu & Anderson, 2006; Holroyd & Coles, 2002; Nason & Laird, 2005); models of declarative memory provide robust predictions of fact recall latency and probability (e.g., Anderson, 2007; Anderson & Reder, 1999; Mackintosh, 1983; Shanks, 1994); and skill acquisition models provide robust predictions of how people achieve experise (e.g., Chase & Simon, 1973; Gobet, 1998; Gobet et al., 2001). Furthermore, multiple individual process models can be integrated together to generate more complete and general predictions of behavior across many contexts (Anderson, 2007; Choi & Ohlsson, 2011; Gray, 2007; Veksler, Myers, & Gluck, 2014).

Cognitive models are typically held to account for behavior in the aggregate (average group behavior) rather than for individual differences. However, even in the absence of precise individual predicitons, general behavioral tendencies can be helpful in predicting likely behavior. This is not different from predicting large-scale events in physical sciences: fundamental principles of physics may not help us to predict exactly where a tsunami will hit, but it is useful to know that some locations are more probable than others.

Additionally, in repeated security games[2] cognitive models can be dynamically updated to provide better predictions of individual attacker's cognition and behavior by use of *model tracing* and *dynamic parameter fitting*. The model tracing technique comprises force-feeding a participant's experiences to the cognitive model. That is, if the participant and the model were to choose different strategies, model actions would be overwritten with participant actions in the model's memory. This method was employed in computerized instructional aids, "cognitive tutors", for students learning high school math in Pittsburgh (Anderson, Corbett, Koedinger, & Pelletier, 1995).Dynamic parameter fitting is used to adjust model parameters based on known data points, so as to make better individual predictions for future behavior. This method was employed to predict performance of individual F-16 pilot teams (Jastrzembski, Gluck, & Rodgers, 2009) and is employed in software that predicts optimal training schedules based on individual performance histories (Jastrzembski, Rodgers, Gluck, & Krusmark, 2014). The following simulations examine the use of these techniques for defender agent software in the security domain.

Simulation 1: Model Tracing

A cognitive model can predict general tendencies, but it is unlikely that a model will predict all decisions of a given individual, even on fairly simple tasks. Model predictions for each of the attacker's decisions contain an element of uncertainty, X. This simulation explores CM-based agent performance for varying sizes of X, and compares CM performance to a normative game theory approach.

For this simulation we employ a sample repeated security game where the attacker and defender have four strategies each, and attacker payoffs are probabilities of attacker success as represented in Table 2. Given this payoffs matrix, if the defender played a fixed strategy (e.g., always play $D2$), the attacker could find a corresponding strategy that would win 90% of the time (e.g., if defender always plays $D2$, then attacker should always play $A3$). If the defender was equally likely to choose any action, the attacker could optimize by always playing $A1$, winning 52.5% of the games. A defender agent based on GT would employ a mixed strategy, choosing actions $D1$, $D2$, $D3$, and $D4$ with probabilities .275, .240,

[1] Playing $A1$ against a random opponent would have a 35% chance of winning, and playing any mix of $A1$ and $A2$ would have a chance of winning that is between 35% and 45%.

[2] Repeated security games differ from "one shot" security games in that the attacker attempts multiple attacks in sequence.

Table 2
Payoffs for the attacker in a security game used for Simulations 1-3, where there are four possible actions for the defender {D1,D2,D3,D4}, and four possible actions for the attacker {A1,A2,A3,A4}.

		Attacker			
		A1	A2	A3	A4
Defender	D1	.15	.45	.50	.90
	D2	.55	.10	.90	.45
	D3	.50	.90	.15	.45
	D4	.90	.50	.50	.10

.275, and .210, respectively. This mixed strategy would leave the attacker without a preference, where any given action would have a 50% probability of success.

However, humans are not perfectly rational, and human attacker action preferences will change based on their experience. For example, if the attacker chooses $A1$ and happens to lose, they will be less likely to choose $A1$ in future attacks, regardless of whether $A1$ is ultimately a good choice. Conversely, if the attacker chooses $A1$ and happens to win, they will be more likely to choose $A1$ in future attacks, regardless of whether $A1$ is ultimately a poor choice.

More formally, after performing some action, A, the expected utility of this action, U_A, is incremented by the following term:

$$\Delta U_A = \alpha(R - U_A), \tag{1}$$

where α is the learning rate, and R is the value of the feedback (e.g. success/failure, reward/punishment). This type of learning (error-driven reinforcement learning) is a very robust principle of biological brains (e.g., Anderson, 2007; Bayer & Glimcher, 2005; Fu & Anderson, 2006; Kable & Glimcher, 2009).[3] The action chosen at each decision point is one with the highest value of the term $U + X$, where X represents exploratory tendencies plus all the unknown factors driving human decision-making at the given moment.

From the perspective of predicting attacker actions, decision-making can be thought of as a stochastic process, where the action with the highest U is the most likely to be selected. A defender agent based on CM would employ *model tracing* to keep track of U values for each attacker action, and then employ this knowledge to outguess the attacker. That is, once each game plays out, and actual attacker action in that game, A, and outcome of the game, R, are both known, the model of the attacker can be updated in accordance with Equation 1. For each consecutive game, CM would predict that the most likely attacker action is one with highest U, and makes a corresponding best choice (e.g., if the attacker in this simulation is likely to choose $A2$, defender will choose $D3$).

Let us assume, for now, that we know attacker learning rate (in this simulation we assume $\alpha = .2$, as is the default in the ACT-R cognitive architecture; Anderson, 2007), and their perceived rewards for success and failure (in this simulation we assume $R = 1$ for each win, and $R = -1$ for each loss), and the attacker has no prior preferences for any actions. Holding these variables constant allows us to answer what the effect of X will be – the lack of predictability in human decision-making.

For general population X is often estimated as gaussian noise with a mean of 0 and a standard deviation, σ, that varies between 5% and 25% of maximum reward values. We ran the simulation for σ values of .05, .15, and .25, averaging over 1024 simulation runs per parameter setting. The results of this simulation, displaying the number of prevented attacks over the course of 200 consecutive security games, are shown in Figure 1.

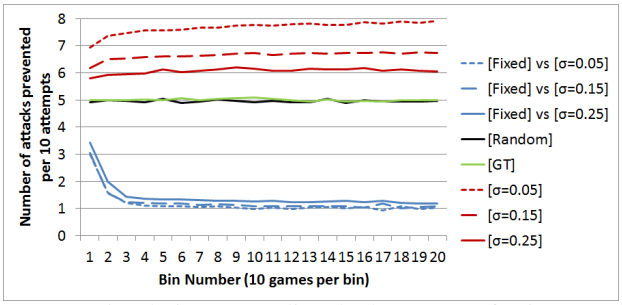


Figure 1. Simulation 1: Predicted advantages of using cognitive models in security games, given various levels of decision predictability. σ refers to uncertainty in the human attacker decisions, rather than in CM defender.

For all σ values in this simulation GT prevented 50.0% of the attacks (100.0 attacks prevented in 200 consecutive attempts). Depending on the predictability factor, CM prevented between 121.5 and 153.6 attempts on average (22%-54% more than GT). In addition to GT and CM, Figure 1 includes predicted baseline performances against random and fixed-strategy defenders. Predictably, a fixed-strategy defender does worst, losing in 75-90% of the games after about 10 initial games. Less obvious may be the fact that random-strategy agent performs almost as well as GT, winning on average 49.6% of the games, compared to 47.5% that we may have predicted against a perfectly rational agent.

The conclusion to draw from these simulation results is that humans are not normative decision makers, nor are we completely unpredictable; thus an approach that considers human cognition can perform better than normative GT. We can employ model tracing to improve defender performance in the security domain despite the fact that many attacker actions are not perfectly predictable. Finally, the less uncertainty there is in attacker behavior, the more attacks can be prevented. Thus, as we begin to account for a greater proportion of attacker cognition we can reduce the size of the X factor, and further improve defender performance.

[3]To be clear, the mechanism being described here, Reinforcement Learning, is only one of many cognitive processes that guides attacker behavior. In this paper we only focus on Reinforcement Learning in repeated security games as a clear and tractable example of how CM can aid in building better decision aids for real-world security domain problems.

Simulation 2: Preferred Strategies

Simulation 1 highlights the advantages of cognitive model predictions despite the uncertainty factor, X, in human decision-making. To isolate the effect of X we made a few assumptions, one of those being that the attacker has no initial action preferences. However, the four attacker strategies described in Simulation 1, Table 2 are not some arbitrarily named buttons $A1$-$A4$, but rather meaningful action-paths to the person(s) performing the attacks.

Let us assume, for example, that in the context of a cyber-attack, the hacker has two decisions: (1) whether to scan for vulnerabilities at a faster or a slower rate, and (2) whether to focus the attack on the main data server or on multiple perimeter machines. The hacker in this case has some risk aversion and believes that faster scans and attacking the main server present higher risks, resulting in initial perceived utilities for the *safer-perimeter*, *safer-main*, *faster-perimeter*, and *faster-main* options of +.10, +.05, 0.00, and -.10, respectively. We will refer to these initial utilities as preference set A, when they correspond to actions $A1$, $A2$, $A3$, and $A4$, respectively; and preference set B when they correspond to actions $A4$, $A3$, $A2$, and $A1$, respectively.

The question is, given that the attacker has some prior preferences, does it hurt the defender to assume that the attacker is a "blank slate"? Simulation revealed that CM would prevent about the same number of attacks against attackers with preference sets A or B as it would against blank slate attackers, see top of Figure 2. At worst, against an attacker with preference set B, σ = .05, CM prevents 2.3 attacks less in 200 attempts than against a blank-slate, σ = .05 attacker. The difference for each other attacker type is less than one attack in 200 attempts. GT agent performance remains at 50% regardless of attacker preferences.

The reason why initial preference sets A and B do not greatly disturb the model tracing approach has to do with the nature of error-driven learning (Rescorla & Wagner, 1972). This long-established principle of human learning suggests that the more surprising (i.e. unexpected) a given outcome is, the greater the change in the human (or simulated) brain. Thus, if human subjective utility for some action is 0.5 and the model assumes that utility to be 0.0, and the actual outcome value after that action was performed was 1.0, the change in action-utility in the model would be half of that in the human. After just a few experiences the model and human action-utilities would begin to converge, and model predictions would become more accurate.

A problem may occur in the instances where initial human preferences are strong enough that some actions are never even attempted. For example, let us examine preference sets C and D that are five times as strong as preference sets A and B, with initial perceived utilities for the *safer-perimeter*, *safer-main*, *faster-perimeter*, and *faster-main* options of +.50, +.25, 0.00, and -.50, respectively. Given these preference sets, an attacker (especially one with a low X factor) will be very unlikely to choose the *faster-main* option.

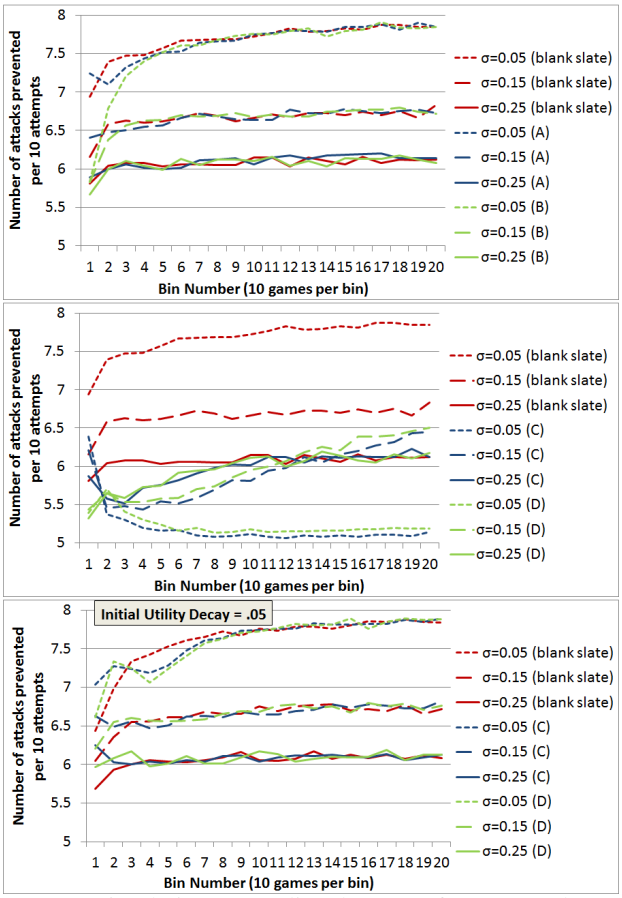


Figure 2. Simulation 2: Predicted CM performance when attacker has biases prior to the first attack. Each trend-line represents CM performance against a different attacker type (attackers differ by uncertainty factors and initial preferences). Preference sets A, B, C, and D are represented in figure legends as (A), (B), (C), and (D) respectively. Each data point represents an average over 1024 simulation runs.

This would throw off model predictions. A CM defender would still perform better than GT against attackers with such preferences, but much worse than it would against attackers with no initial preferences, see middle graph in Figure 2.

To account for potential extreme negative preferences, we can add Initial Utility Decay (IUD), dynamically adjusting initial action-utilities for any actions that are not being attempted by the attacker. The assumption is that if the attacker is not choosing a given option, then that option must have a lower subjective utility for them. For current simulation purposes we will employ linear decay, decrementing U_A by 0.05 at each decision point for every action A that the attacker does not exercise.

The results of employing a CM defender agent with initial utility decay are shown at bottom of Figure 2. As the figure suggests, initial utility decay greatly aids in accounting for strong initial preferences, resulting in almost no difference (less than 1 attack in 200 attempts) between defending against

blank-slate attackers and those with preference sets C and D. This improvement comes at a slight cost against a blank-slate attacker. CM performance with IUD is slightly worse than that without IUD against a blank-slate attacker (red lines at bottom and top of Figure 2, respectively), preventing 1.3, 0.4, and 0.4 attacks less in 200 attempts against attackers with σ values of .05, .15, and .25, respectively.

IUD is just one potential method for dynamic parameter adjustment in cognitive modeling. There are undoubtedly better alternatives to overcoming the problem of initial preferences other than linear IUD. The focus here is not in finding the optimal method, but rather in highlighting the benefits of using cognitive modeling in the security domain. This simulation suggests that even when individual attackers have unpredictable initial preferences, CM preferences can be adjusted dynamically without incurring a significant loss in performance.

Simulation 3: Learning Rate

Simulations 1 and 2 explore how CM-based defender agent performance in repeated security games is affected by decision unpredictability and initial preferences of human attackers. One other variable from Equation 1 that we have yet to discuss is the learning rate, α. This simulation focuses on CM performance given different attacker learning rates.

Table 3
Number of attacks prevented in 200 attempts by CM defender (IUD=.05). Each data point represents an average over 1024 simulations. α and σ refer to learning rate and uncertainty in the human attacker decisions. GT-based defender performance for all attackers is 100.0.

	CM learning rate				
	0.1	0.2	0.3	0.4	auto
α=0.1, σ=0.05	139.0	140.8	138.1	136.0	140.4
α=0.2, σ=0.05	139.4	152.5	152.3	151.6	151.9
α=0.3, σ=0.05	133.5	152.8	155.0	154.8	154.5
α=0.4, σ=0.05	131.6	150.1	154.4	154.7	153.9
α=0.1, σ=0.15	117.3	117.6	117.5	116.4	117.1
α=0.2, σ=0.15	130.9	133.0	133.0	132.2	132.3
α=0.3, σ=0.15	136.0	140.8	141.3	141.6	140.8
α=0.4, σ=0.15	136.1	142.6	144.4	144.4	143.9
α=0.1, σ=0.25	110.6	110.8	110.5	109.8	110.4
α=0.2, σ=0.25	120.5	121.6	121.5	121.4	121.3
α=0.3, σ=0.25	127.4	129.9	130.5	130.3	129.6
α=0.4, σ=0.25	131.3	134.6	135.9	136.0	135.1

Table 3 displays CM (IUD=.05) performance with different assumed learning rates, playing 200 consecutive security games against attackers with different actual learning rates.[4] In general, the higher the attacker learning rate (i.e., the more of a factor their experiences are in their decision-making), the easier it is to predict attacker decisions and prevent future attacks. Depending on the attacker, the range of differences between the best and worst CM performance was as high as 23 attacks in 200 attempts; though assuming a higher attacker learning rate does not hurt CM defender performance as much as assuming a lower learning rate.

The "auto" learning rate (right-most column) represents a CM defender that dynamically adjusts the learning rate parameter prior each decision. That is, the "auto" CM agent adjusts its assumption about the attacker learning rate depending on which assumed learning rate would result in a greater number of correct predictions for all known attacker decisions. As implemented in this simulation, the "auto" agent contrasts prediction history of CM agents with assumed learning rates of .1, .2, .3, and .4, and mimics the next decision of the agent with the best prediction rate.

Dynamically fitting the learning rate parameter in this way does not guarantee that the attacker learning rate will be correctly inferred at any given decision point, mostly due to the uncertainty in attacker decisions. Thus, the "auto" CM performance cannot be as good as that of an omniscient agent. However, the "auto" learning rate produces near-optimal performance, see Figure 3.

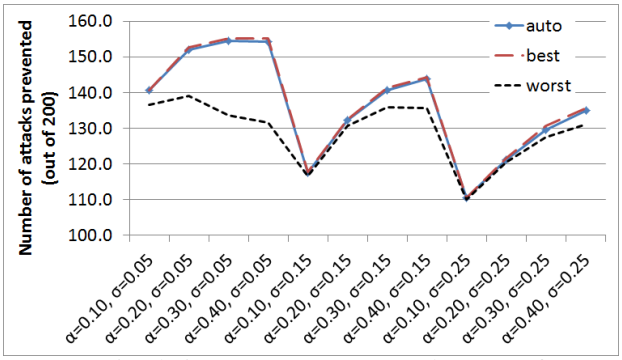


Figure 3. Simulation 3: Best, worst, and auto performance from Table 3. α and σ refer to learning rate and uncertainty in the human attacker decisions, rather than in CM defender.

To be clear, the "auto" learning rate method employed here is not the only method for adjusting model parameters. The focus here, however, is not on finding the optimal method for dynamic parameter fitting, but rather on highlighting the availability of techniques in cognitive modeling, such as model tracing and dynamic parameter fitting, for providing individual/team -tailored decision predictions that can be of great use in the security domain and beyond.

Summary & Discussion

In recent years, game theory -based decision-aid software has received significant attention for success in real-world security domain problems, such as scheduling patrols conducted by the US Coast Guard and police, allocating federal air marshals on flights, and major anti-poaching efforts. Behavioral game theory points out that normative approaches provide unrealistic predictions of human choice, and suggests the use of

[4]Results in Table 3 and Figure 3 are based on games against a blank-slate agent, but all effects hold against agents with initial preferences.

behavioral/cognitive models. In this paper we focus on the use of cognitive modeling to improve on the success of normative game-theory approaches in the security domain.

We present three simulations that highlight the potential advantages of employing cognitive models for predicting attacker decisions. The simulations suggest that (1) cognitive modeling provides performance advantages over normative game theory approaches, and (2) model parameters can be adjusted dynamically to make useful predictions about each individual human attacker despite the fact that each individual attacker may have different preferences and learning abilities.

The presented simulation results provide encouraging evidence of potential usefulness of cognitive models in the context of real-world security problems. Despite the fact that simulations in this paper are based on robust behavioral phenomena, the presented results should be taken as theoretical in nature, requiring further empirical validation. In future work we plan to gather human data and validate current simulation results.

In the current paper we only focus on a single cognitive process, as an example of how cognitive modeling may be employed in this domain. There is a wide array of established cognitive models beyond what we could explore in this paper. Integration of more models in CM-based decision aids would help in reducing the uncertainty factor, further improving the rate of prevented attacks.

Acknowledgements

This work was funded under Cooperative Agreement Number W911NF-09-2-0053.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford ; New York: Oxford University Press.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of the Learning Sciences*, *4*(2), 167–207. doi: 10.1207/s15327809jls0402_2

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology-General*, *128*(2), 186–197.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141.

Camerer, C. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Choi, D., & Ohlsson, S. (2011). Effects of multiple learning mechanisms in a cognitive architecture. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the thirty-third annual meeting of the cognitive science society* (pp. 3003–3008). Boston, MA: Cognitive Science Society.

Fu, W. T., & Anderson, J. R. (2006). From recurrent choice to skilled learning: A reinforcement learning model. *Journal of Experimental Psychology: General*, *135*(2), 184–206.

Gächter, S. (2008). Behavioral Game Theory. In *Blackwell handbook of judgment and decision making* (pp. 485–503).

Gobet, F. (1998). Expert memory: a comparison of four theories. *Cognition*, *66*(2), 115–152.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C. H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243.

Gray, W. D. (Ed.). (2007). *Integrated models of cognitive systems*. New York: Oxford University Press.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis. of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709.

Jastrzembski, T. S., Gluck, K. A., & Rodgers, S. (2009). Improving military readiness: A state-of-the-art cognitive tool to predict performance and optimize training effectiveness. In *The interservice/industry training, simulation, and education conference (i/itsec)*.

Jastrzembski, T. S., Rodgers, S. M., Gluck, K. A., & Krusmark, M. A. (2014). *Predictive performance optimizer*. Google Patents.

Kable, J. W., & Glimcher, P. W. (2009). The Neurobiology of Decision: Consensus and Controversy. *Neuron*, *63*(6), 733–745. doi: 10.1016/j.neuron.2009.09.003

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kar, D., Fang, F., Fave, F. D., Sintov, N., & Tambe, M. (2015). A Game of Thrones: When Human Behavior Models Compete in Repeated Stackelberg Security Games. In *2015 international conference on autonomous agents and multiagent systems* (pp. 1381–1390). International Foundation for Autonomous Agents and Multiagent Systems.

Lebiere, C., Wallach, D., & West, R. L. (2000). A memory-based account of the prisonerâĂŹs dilemma and other 2x2 games. In *Proceedings of international conference on cognitive modeling* (pp. 185–193).

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford, New York: Clarendon Press, Oxford University Press.

Nason, S., & Laird, J. E. (2005). Soar-RL: Integrating reinforcement learning with Soar. *Cognitive Systems Research*, *6*, 51–59.

Pita, J., Jain, M., Ordónez, F., Portway, C., Tambe, M., Western, C., & Kraus, S. (2008). ARMOR Security for Los Angeles International Airport. In *Twenty-third aaai conference on artificial intelligence* (pp. 1884–1885).

Reitter, D., & Lebiere, C. (2011). Towards cognitive models of communication and group intelligence. In *Proceedings of the 33rd annual meeting of the cognitive science society, boston* (pp. 734–739).

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In P. W. F. Black AH (Ed.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Sample, C. (2015). *Cyber + Culture Early Warning Study* (Tech. Rep.). CERT. doi: CMU/SEI-2015-SR-025

Shanks, D. (1994). Human associative learning. In N. J. Mackintosh (Ed.), *Animal learning and cognition*. San Diego, CA: Academic Press.

Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., DiRenzo, J., ... Meyer, G. (2012). Protect: A deployed game theoretic system to protect the ports of the united states. In *11th international conference on autonomous agents and multiagent systems* (pp. 13–20).

Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). msterdam: Elsevier.

Tambe, M. (2011). *Security and Game Theory* (Vol. 9781107096). doi: 10.1017/CBO9780511973031

Tambe, M., Jiang, A. X., An, B., & Jain, M. (2013). Computational game theory for security: Progress and challenges. In *Aaai spring symposium on applied computational game theory*.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2014). SAwSu: An Integrated Model of Associative and Reinforcement Learning. *Cognitive Science*, *38*(3), 580–598.

West, R. L., Lebiere, C., & Bothell, D. J. (2006). Cognitive architectures, game playing, and human evolution. *Cognition and multi-agent interaction: From cognitive modeling to social simulation*, 103–123.

Probability Prediction in Children with ASD

Zi L. Sim (zi@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Fei Xu (fei_xu@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720 USA

Abstract

Individuals with Autism Spectrum Disorder (ASD) often struggle with making inductive generalizations. Yet for typically developing children, the capacity to make such generalizations is a hallmark of human learning. This ability requires some understanding of "intuitive statistics" (i.e., the understanding that there is a relationship between samples and populations), which have been previously demonstrated to emerge early on in infancy. We hypothesized that the challenges with inductive generalization among the ASD population may have its roots in weaknesses in probabilistic reasoning. In the current study, we gave children with ASD a probability prediction task adapted from the method used with infants in Teglas et al. (2007), and our results over two experiments with two groups (one from the U.S. and one from Singapore) suggest that compared with typically developing children, children with autism may have difficulties in engaging in probabilistic reasoning.

Keywords: autism; probabilistic reasoning; prediction

Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that is characterized by impairments in social interaction and communication, as well as the presence of restricted patterns of behaviors and interests (American Psychiatric Association, 2013; Baron-Cohen, 1997). Cognitive research in the field of autism has centered on three traditions when it comes to accounting for ASD (Rajendran & Mitchell, 2007): the theory-of-mind (ToM) hypothesis, which posits that symptoms of autism manifest because of deficits in the ability to impute mental states to oneself and to others (Baron-Cohen, Leslie, & Frith, 1985; Premack & Woodruff, 1978); the executive dysfunction hypothesis, which argues that the symptoms of autism are caused by core difficulties in the planning and execution of complex actions (Hill, 2004; Ozonoff, Pennington, & Rogers, 1991); and the weak central coherence theory, which proposes that individuals with autism favor local processing (i.e., the details) over global processing (i.e., "the big picture") (Happé & Frith, 2006).

A relatively unexplored perspective is that which conceptualizes autism as a disorder of learning (following Pellicano, 2010; Solomon, Smith, Frank, Ly, & Carter, 2011). For instance, a common finding from intervention studies is that individuals with autism often fail to generalize explicitly taught skills across different contexts or to related skills (e.g., Dawson, Mottron, & Gernsbacher, 2005; Hwang & Hughes, 2000; Ivar Lovaas & Smith, 1989; Ozonoff & Miller, 1995). In more recent work, researchers have found that children and adolescents with ASD were less likely to learn from their experiences, and that their generalizations were less consistent when compared to typically developing participants (de Marchena, Eigsti, & Yerys, 2015). Clinicians have also long reported that children with autism often struggle with generalization (Rimland, 1964): for example, after being taught to brush their teeth with a green toothbrush, children with autism may appear to be at a loss when asked to brush their teeth with a red toothbrush later on.

Yet for typically developing children, the capacity to make such generalizations is a hallmark of human learning. Given small amounts of data, human learners readily make inductive inferences, formulating general principles that are extracted from the specific data. Developmental research have repeatedly demonstrated that children are extremely proficient learners, making inductive generalizations with much ease. They learn the meanings of some words with just a single labeled exemplar (Carey & Bartlett, 1978); they generalize non-obvious properties to novel objects after just a short demonstration (Baldwin, Markman, & Melartin, 1993; Gelman & Davidson, 2013; Welder & Graham, 2006), and they learn the physical rules of occlusion with just a single trial (Wang & Baillargeon, 2005).

To make such generalizations proficiently requires some understanding of "intuitive statistics", that is, understanding that a random sample enables one to make predictions about an overall population, and conversely, that a population allows one to make predictions about randomly drawn samples. This type of statistical inference can be found in almost every domain of learning, e.g., physical reasoning, social cognition, word learning, and causal reasoning (Chomsky, 1980; Gelman & Wellman, 1991; Gopnik & Sobel, 2000; Gopnik & Wellman, 2012; Griffiths & Tenenbaum, 2009; Keil, 1981; Kushnir, Xu, & Wellman, 2010; Nisbett, Krantz, Jepson, & Kunda, 1983; Xu & Tenenbaum, 2007), and the ability to make such inferences allows learners to rapidly acquire new knowledge about the world.

Previous research have demonstrated that intuitive statistics emerges very early on in development, enabling children to engage in inductive learning within the first few years of life: 6- to 12-month-old infants are sensitive to differences in probabilities (Denison, Reed, & Xu, 2011;

Téglás, Girotto, Gonzalez, & Bonatti, 2007; Xu & Garcia, 2008) For example, Téglás et al. (2007) showed that in a lottery machine-like setup that consisted of 1 yellow and 3 blue objects bouncing around, infants were more "surprised" to see a yellow object (low probability) exiting the machine, than when a blue object (high probability) did. Furthermore, this early sensitivity has been shown to guide infants and young children in making predictions and in fulfilling their goals and desires (Acredolo, O'Connor, Banks, & Horobin, 1989; Denison & Xu, 2010, 2014; Yost, Siegel, & Andrews, 1962; Zhu & Gigerenzer, 2006).

As such, we hypothesize that the challenges that autistic individuals face with generalization may have their roots in weaknesses in probabilistic reasoning. There is some preliminary evidence for this claim: in a large-scale foraging task where ASD children (8- to 12-year-olds) and matched controls had to search a room for a target among possible search locations embedded into the floor, researchers found that the autistic individuals appeared to be less sensitive to the statistical properties of the search area, taking a much longer time as compared to the matched controls to realize that one side of the room was more likely to contain the target (Pellicano et al., 2011).

In the current study, we directly examined probabilistic reasoning in children with ASD by adapting the method used with infants in Téglás et al. (2007). This method was suitable due to its relatively low task demands and was specifically about probabilistic reasoning. Over two experiments, 6- to 12-year-old children with ASD and matched controls were presented with movies displaying four objects bouncing around a lottery-like machine. The movies were identical to those shown to 12-month-old infants in Téglás et al. (2007). The two groups of children were subsequently asked to predict which object would fall out by choosing between two pictures displaying two possible outcomes.

Experiment 1

Method

Participants The sample consisted of 21 verbally fluent English-speaking 6- to 12-year-old children (18 boys) with ASD. Diagnoses were confirmed through a review of clinical diagnostic reports provided by the parents. The mean age of the sample was 104.6 months (range = 80.6 to 150.6 months). All participants were recruited from Pathlight School, which is an autism-focused school in Singapore that offers mainstream academic curriculum and life-readiness skills. Average Performance IQ was 101.3 (*SD* = 17.3). Two children were tested but excluded for failing the control task.

Materials Seven QuickTime movies simulating four three-dimensional, solid objects bouncing inside a lottery-like machine were presented on a 17-inch screen using the PsyScope software running on a MacBook Pro. These movies were provided by Luca Bonatti from the set of stimuli presented to 12-month-old infants in Téglás et al. (2007). For each movie, two picture cards (8.5 inches x 6 inches) displaying two different possible outcomes were printed in color and laminated. The back of these picture cards had a small Velcro strip with "loops." A separate A4-sized (8.27 inches x 11.69 inches) laminated card was also used, displaying the words "What happened?" above a large empty printed rectangle. At the center of the rectangle was a Velcro strip with "hooks," so a picture card could be attached to this large laminated card.

Procedure Children were tested individually in a quiet room at Pathlight School. They sat about 30 inches from the screen. The children's parents were present in the room, but were seated about 60 inches behind the child's chair. Parents filled out questionnaires throughout the session to reduce the potential for influencing their children's answers.

The procedure consisted of two phases: familiarization and test. Children were shown two familiarization movies, four experimental movies and one control movie. The order of presentation for the four experimental movies was counterbalanced across subjects. Before each movie, a visual attractor always appeared on the screen (with sounds) to orient the child's attention to the center of the screen.

Familiarization Phase In this phase, children watched two familiarization movies in which two types of objects (two of each type, e.g. two blue cubes and two yellow crosses) bounced around the lottery-like container. After approximately 10 seconds, one of the objects (e.g. a blue cube) fell out of the container accompanied by a "cuckoo" sound. 1 second later, the remaining objects in the container were occluded. This procedure was repeated for the next familiarization movie, except that the other type of object (e.g. a yellow cross) fell out of the container this time.

Test Phase In the test phase, children participated in four experimental trials and one control trial. In each experimental trial (Figure 1), three identical objects and one object of a different color and shape (e.g. three blue cubes and one yellow cross) were shown bouncing around the container. After approximately 10 seconds, the container was occluded as in the Familiarization Phase. 2 seconds later, a "cuckoo" sound was made, but the bottom half of the screen was blocked such that it was not possible to see which object had fallen out. The experimenter said, "Something happened, but the screen got blocked!" She then presented two pictures displaying the two possible outcomes (e.g., one of the three identical objects had exited the container vs. the object different in color and shape had exited the container) and said, "Now, I have these two pictures over here." Next, the "What happened?" card was presented and the experiment asked, "Can you show me what happened?" The children were requested to attach one of the picture cards to the large rectangular box. If children failed to respond, the experimenter pointed to the card on the right and said, "If you think this happened, put this card

here," and pointed to the card on the left and said, "If you think this happened, put this card here" in a neutral tone of voice. After children made a clear choice by placing one of the picture cards on to the "What happened?" card, the experimenter proceeded to the next experimental trial. The order of presentation for the four experimental movies was counterbalanced across all participants.

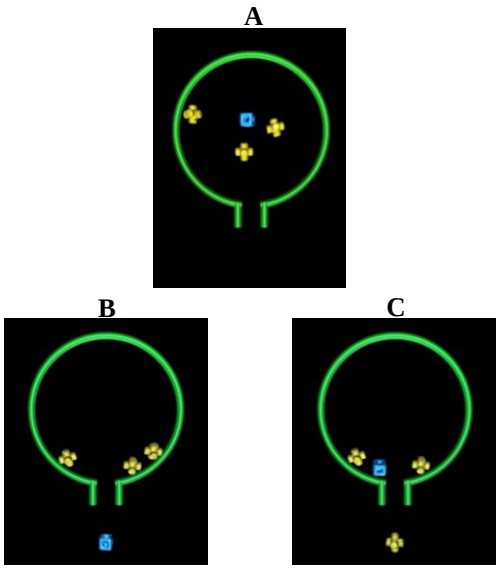


Figure 1: (A) Screenshot from an experimental movie. (B) & (C) Two pictures displaying two possible outcomes.

After the four experimental trials were completed, children were presented with a control trial (Figure 2). Like the experimental movies, the control movie presented three identical objects and one object different in color and shape bouncing around the container. However, there was now a physical barrier in the middle of the container that confined the three identical objects to the top of the container. As such, it was physically impossible for any of the three identical objects to exit the container. The procedure that followed was exactly the same as that of the experimental trials, and children were asked to choose between two picture cards displaying two different outcomes (e.g. one of the three identical objects had crossed the barrier and exited the container vs. the object different in color and shape had exited the container) as a representation of what had happened.

No feedback was given during any of the test trials. The experimenter only responded "Okay" to all of the children's choices.

Coding Children's responses in the test trials (four experimental and one control) were scored for accuracy. As children were asked to predict what had happened, choosing the probable outcome was scored as 1 point. Choosing the improbable/impossible outcome was thus scored as 0 points. A second coder recoded all of the children's responses, and the level of agreement between the coders was 100%.

Results

An alpha level of 0.05 was used in all statistical analyses. Preliminary analyses found no effects of median age-split (whether the children were younger or older than the median age of the group) and the order of presentation for the four experimental movies. Subsequent analyses were collapsed over these variables.

Overall, we found that the children in the Singapore ASD sample were able to respond correctly in the control trial. 19 out of 21 children responded correctly, and this proportion was significantly different from chance, Exact binomial p (two-tailed) < .001. However, children in the final ASD sample consisting of the 19 children who passed the control trial did not perform significantly differently from chance (.50) on the experimental trials (M = .47, SD = .38), $t(18)$ = -.301, p = .77. A conservative binomial test based on the total number of correct trials also showed that children did not perform significantly different from chance, Exact binomial p (two-tailed) = .77.

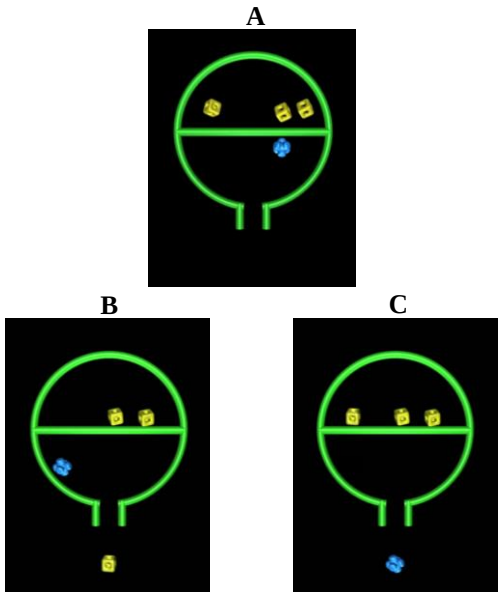


Figure 2: (A) Screenshot from the control movie. (B) & (C) Two pictures displaying two possible outcomes.

Discussion

The children in the ASD sample in Singapore did not perform reliably better than chance levels in our probability prediction task. This result is striking given that the movies used were adapted from Téglás et al. (2007), in which researchers found strong evidence that 12-month-old typically developing infants have rational expectations about future events based on single-event likelihoods. Furthermore, the low rate of success did not appear to be due to a lack of understanding with regards to the task requirements, since 19 out of 21 of the ASD children (90.5%) answered the control trial correctly. However, it remains a possibility that typically developing children

would likewise fail at such a probability prediction task. In other words, while 12-month-old infants may look longer when presented with an improbable event as compared to a probable event, older children may still find it difficult to explicitly make predictions about future events based on their likelihoods. Such an account is conceivable, as the infant procedure used in Téglás et al. (2007) was an implicit, non-verbal task, while our adapted procedure is arguably a task that is more explicit and verbal in nature. In Experiment 2, we thus collected data from a US sample consisting of both ASD and typically developing children.

Experiment 2

Method

Participants The sample consisted of nine verbally fluent English-speaking 6- to 12-year-old children (8 boys) with ASD and nine typically developing (TD) 6- to 12-year-old children (4 boys), group-matched by chronological age. All participants were recruited from Berkeley, California, and its surrounding communities through advertisements. Diagnoses of the ASD participants were confirmed through a review of clinical diagnostic reports provided by the parents. Participants in the TD group were excluded if they had first-degree relatives with an ASD diagnosis. Age was not found to be significantly different between the two groups (M_{ASD} = 97.07, SD_{ASD} = 18.77; M_{TD} = 98.44, SD_{TD} = 19.93), $t(16) = .15$, $p = .88$. Visual Spatial Index scores, which measure the ability to integrate and synthesize part-whole relationships, to evaluate visual details, and to understand visual-spatial relationships, did not differ significantly between groups (M_{ASD} = 103.29, SD_{ASD} = 11.45; M_{TD} = 111.33, SD_{TD} = 15.51) as well, $t(14) = 1.15$, $p = .27$. This index score was not obtained for two of the children in the ASD sample due to a lack of response to the component subtests.

Materials and Procedure The materials and procedure in Experiment 2 were similar to those of Experiment 1, except for the Familiarization Phase. The Familiarization Phase was modified slightly to increase children's understanding of the task, and the procedure is as follows:

Familiarization Phase In this phase, children watched each of the two familiarization movies twice. During the first familiarization movie, the experimenter commented on the occurring events in such a manner, "Objects bounce around this circle. When you hear this noise [cuckoo noise], one object falls out." The second familiarization movie was then played, and the experimenter commented on the events in the same way. Each of the two familiarization movies was then played once more, without any additional instructions.

Test Phase The test phase in Experiment 2 was identical to that of Experiment 1.

Coding Children's responses in the test trials (four experimental and one control) were scored in the same way as Experiment 1. A second coder recoded all of the children's responses, and the level of agreement between the coders was 100%.

Results

An alpha level of 0.05 was used in all statistical analyses. Preliminary analyses found no effects of median age-split (whether the children were younger or older than the median age of the group) and the order of presentation for the four experimental movies. Subsequent analyses were collapsed over these variables.

Overall, we found that children in both the ASD and TD sample were able to respond correctly in the control trial – all of the children tested made the correct prediction, and this proportion was significantly different from chance, Exact binomial p (two-tailed) < .001. Using children's responses over the four test trials, we then performed a repeated measures logistic regression with group (ASD vs. TD) as the between-subjects variable. Our results indicate that there was a significant difference between the performance of the ASD and TD children (M_{ASD} = .36, SD_{ASD} = .28; M_{TD} = .70, SD_{TD} = .21), Wald Chi-Square = 7.96, $p = .005$. A conservative binomial test based on the total number of correct trials also showed that TD children performed significantly better than chance, Exact binomial p (two-tailed) = .029, while ASD children did not, Exact binomial p (two-tailed) = .13.

Discussion

In Experiment 2, we found a significant difference between the performance of the ASD and TD children: Children in the ASD sample in Berkeley did not perform reliably better than chance levels in our probability prediction task, which replicates our findings in Experiment 1. In contrast, typically developing children were successful at this task and were able to make predictions about future events based on their likelihoods. This difference found between the two groups did not appear to be due to a difference in the children's ability to understand the task requirements, as all of the participants were able to respond to the control trial correctly.

General Discussion

The present study examined whether children with Autism Spectrum Disorder (ASD) show weaknesses in probabilistic reasoning. Using a probability prediction task adapted from a method used with 12-month-old infants (Téglás et al., 2007), we found across two different samples that high-functioning, verbally fluent children with ASD struggled with making predictions about future events based on their single-event likelihoods. In contrast, our comparison group of typically developing children were successful in making such predictions, consistent with results obtained from previous studies with infants and young children (Acredolo

et al., 1989; Denison & Xu, 2010, 2014; Yost et al., 1962; Zhu & Gigerenzer, 2006). The current findings are striking, considering that the two groups of children in Experiment 2 did not differ in their Visual Spatial Index scores, which are particularly relevant for processing the stimuli presented in our task.

May our results be better accounted for by the three dominant cognitive theories for ASD? While more research would be necessary to carefully tease these accounts apart, we have reason to believe that the current cognitive theories do not necessarily do so. According to the theory-of-mind hypothesis, the core deficit of autism is a failure/delay in taking into account others' mental states. Given that our task does not ostensibly require participants to impute the mental state of others, it is unlikely that the difference found in probabilistic reasoning between the ASD and TD groups would be related to any previously established differences in ToM. With regards to the executive dysfunction hypothesis, the high rate at which the ASD children were passing the control task suggests that they had an ability to sustain attention to the presented movies in the current study. It is possible that this group of children appeared to respond correctly on the control trial due to difficulties in inhibiting the prepotent response of selecting the dissimilar object exiting as the predicted outcome (i.e., choosing the picture with one blue object outside of the container, rather than the picture with one of the three identical yellow objects outside of the container). However, if this alternative explanation were to be true, then the ASD children should have performed significantly worse than chance on the experimental trials because of an equivalent tendency to select the low-probability outcomes. Finally, the weak central coherence theory posits that autism is characterized by a weak drive towards obtaining global coherence, such that individuals with autism are predisposed to process information in a detail-focused, piecemeal way. Again, it is unclear how such a theory would account for the different success rate that children with ASD show on the experimental trials vs. the control trials; how would attending to the details of the movies in a segmented manner lead children to pass the control but not the experimental trials?

Therefore, our results suggest that there may be early differences in probabilistic reasoning between children with autism and typically developing children. This weakness in intuitive statistics may result in impairments when making inductive generalizations. That being said, it should be noted that autism is an extremely complex phenomenon. We do not think that the demonstrated weakness in probabilistic reasoning is necessarily a defining factor for autism, or a factor that would explain the full spectrum of autism symptomatology. Instead, we speculate that this deficit may be general to all learners who have difficulties with making inductive generalizations. Given the centrality of inductive learning in almost every domain, such early difficulties may lead to a cascade of negative consequences in development, and this would be the case whether with regards to children with autism or children with other developmental disorders. Work is ongoing in our lab to examine more closely the deficit in probability prediction using other related tasks, and whether the deficits in probabilistic reasoning may be directly linked to deficits in the ability to make inductive generalizations.

Finally, we note that the current work examines autism through the lens of learning, which may allow its findings to be more amendable to the design of interventions, an aspect that is especially important to stakeholders of the ASD community. As such, we believe that the current work is a first step towards opening up new grounds in the study of autism.

Acknowledgments

We thank Pathlight School in Singapore, as well as Ariana Grande, Ashley Heywood, Sulynn Miao, Erica Shade, Grace Shen, and Sarah Stewart for their help in testing, and all the parents and children for their participation. This work was partially supported by SFARI grant 347393 from the Simons Foundation.

References

Begus, K., Gliga, T., & Southgate, V. (2014). Infants learn what they want to learn: responding to infant pointing leads to superior learning. *PloS one*, *9*(10), e108817.

Acredolo, C., O'Connor, J., Banks, L., & Horobin, K. (1989). Children's ability to make probability estimates: Skills revealed through application of Anderson's functional measurement methodology. *Child Development*, *60*(4), 933–945.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, *64*(3), 711–728.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Carey, S., & Bartlett, E. (1978). Acquiring a Single New Word. In *Proceedings of the Stanford Child Language Conference* (Vol. 15, pp. 17–29).

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(01), 1–15.

Dawson, M., Mottron, L., & Gernsbacher, M. A. (2005). Learning in Autism.

de Marchena, A. B., Eigsti, I.-M., & Yerys, B. E. (2015). Brief Report: Generalization Weaknesses in Verbally Fluent Children and Adolescents with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(10), 3370–3376. http://doi.org/10.1007/s10803-015-2478-6

Denison, S., Reed, C., & Xu, F. (2011). The emergence of probabilistic reasoning in very young infants. *… the 33rd Annual Meeting of the …*.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes.

Developmental Science, *13*(5), 798–803. http://doi.org/10.1111/j.1467-7687.2009.00943.x

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–47. http://doi.org/10.1016/j.cognition.2013.12.001

Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology*, *66*(3), 327–353. http://doi.org/10.1016/j.cogpsych.2013.02.001

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings non-obvious*. *Cognition*, *38*, 213–244.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–22.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–108. http://doi.org/10.1037/a0028044

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716. http://doi.org/10.1037/a0017201

Happé, F., & Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *36*(1), 5–25. http://doi.org/10.1007/s10803-005-0039-0

Hill, E. L. (2004). Executive dysfunction in autism. *Trends in Cognitive Sciences*, *8*(1), 26–32. http://doi.org/10.1016/j.tics.2003.11.003

Hwang, B., & Hughes, C. (2000). The effects of social interactive training on early social communicative skills of children with autism. *Journal of Autism and Developmental Disorders*, *30*(4).

Ivar Lovaas, O., & Smith, T. (1989). A comprehensive behavioral theory of autistic children: Paradigm for research and treatment. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*(1), 17–29. http://doi.org/10.1016/0005-7916(89)90004-9

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, (May).

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134–40. http://doi.org/10.1177/0956797610376652

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363. http://doi.org/10.1037//0033-295X.90.4.339

Ozonoff, S., & Miller, J. N. (1995). Teaching theory of mind: a new approach to social skills training for individuals with autism. *Journal of Autism and Developmental Disorders*, *25*(4), 415–33.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive function deficits in high-functioning autistic individuals: relationship to theory of mind. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *32*(7), 1081–105.

Pellicano, E. (2010). The development of core cognitive skills in autism: a 3-year prospective study. *Child Development*, *81*(5), 1400–16. http://doi.org/10.1111/j.1467-8624.2010.01481.x

Pellicano, E., Smith, A. D., Cristino, F., Hood, B. M., Briscoe, J., & Gilchrist, I. D. (2011). Children with autism are neither systematic nor optimal foragers. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(1), 421–6. http://doi.org/10.1073/pnas.1014076108

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, (4), 515–526.

Rajendran, G., & Mitchell, P. (2007). Cognitive theories of autism. *Developmental Review*, *27*(2), 224–260. http://doi.org/10.1016/j.dr.2007.02.001

Rimland, B. (1964). Infantile autism: The syndrome and its implications for a neural theory of behavior.

Solomon, M., Smith, A. C., Frank, M. J., Ly, S., & Carter, C. S. (2011). Probabilistic reinforcement learning in adults with autism spectrum disorders. *Autism Research : Official Journal of the International Society for Autism Research*, *4*(2), 109–20. http://doi.org/10.1002/aur.177

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19156–9. http://doi.org/10.1073/pnas.0700271104

Wang, S., & Baillargeon, R. (2005). Inducing infants to detect a physical violation in a single trial. *Psychological Science*, *16*(7), 542–9. http://doi.org/10.1111/j.0956-7976.2005.01572.x

Welder, A. N., & Graham, S. a. (2006). Infants' categorization of novel objects with more or less obvious features. *Cognitive Psychology*, *52*(1), 57–91. http://doi.org/10.1016/j.cogpsych.2005.05.003

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012–5. http://doi.org/10.1073/pnas.0704450105

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–72. http://doi.org/10.1037/0033-295X.114.2.245

Yost, P., Siegel, A., & Andrews, J. (1962). Nonverbal probability judgments by young children. *Child Development*, *33*(4), 769–780.

Zhu, L., & Gigerenzer, G. (2006). Children can solve Bayesian problems: the role of representation in mental computation. *Cognition*, *98*(3), 287–308. http://doi.org/10.1016/j.cognition.2004.12.003

Left-right mental timeline is robust to visuospatial and verbal interference

Rose K. Hendricks* (rhendricks@ucsd.edu)
Esther J. Walker* (e1walker@ucsd.edu)
Benjamin K. Bergen (bkbergen@ucsd.edu)
Lera Boroditsky (lera@ucsd.edu)
Rafael Núñez (nunez@cogsci.ucsd.edu)
Department of Cognitive Science, 9500 Gilman Dr.
La Jolla, CA 92093 USA
** Joint first authors who contributed equally to this work*

Abstract

We test the robustness of American college students' mental timeline to dual tasks that have interfered with spatial and verbal reasoning in prior work. We focus on the left-right axis for representing sequences of events. We test American college students, who read from left to right. We test for automatic space-time mappings using two established space-time association tasks. We find that their tendency to associate earlier events with the left side of space and later events with the right remains under conditions of visuospatial and verbal interference. We find this both when participants made time judgments about linguistic and non-linguistic stimuli. We discuss the relationship between these results and those obtained for mental timelines that result from learning new metaphors in language (Hendricks & Boroditsky, 2015), and the effects of the same interference tasks on number tasks (mental number-line and counting; van Dijck et al., 2009; Frank et al., 2012).

Keywords: space; time; mental timeline; metaphor; working memory; interference; implicit association

Introduction

Space and time are intricately linked in the human mind. Systematic associations between time and space show up not only in how we talk about time (e.g., *I'm looking forward to the weekend*), but also in co-speech gesture (Cooperrider & Núñez, 2009), in a variety of cultural artifacts (e.g., calendars, timelines), and in how we reason about time (e.g., Boroditsky, 2000; for review, see Núñez & Cooperrider, 2013).

In this paper we focus on representations of the order of events on the left-right axis. Prior work has shown that people who read and write from left to right (as English speakers do) tend to form a left-right representation of time with earlier events mapped to the left and later events mapped to the right (e.g., Tversky, Kugelmass & Winter, 1991; Weger & Pratt, 2008). People who read and write from right to left (as Hebrew and Arabic speakers do), accordingly tend to associate earlier events with the right side of space, and later events with the left (e.g., Fuhrman & Boroditsky, 2010).

Here we ask whether such left-right representations of time are robust to two forms of dual task interference: verbal and visuospatial. We used interference paradigms that have been shown to selectively interfere with other tasks. The current paper reports two experiments. The experiments were designed and conducted independently by two subgroups of the authors (E.W., B.B., R.N. [Exp. 1] and R.H. and L.B. [Exp. 2]), with some differences in methods and materials. Despite the differences, these independently conducted studies yielded remarkably similar results.

Both studies tested American college students on implicit space-time association tasks on the left-right axis. In both studies, participants made speeded judgments about the order of sequential events under either verbal, visuospatial, or no interference. In Experiment 1, the stimuli used were linguistic (sequences of words describing life events). In Experiment 2, the stimuli were entirely non-linguistic (picture sequences). In both studies, results revealed an implicit left-right association for time (as expected), and in both studies this left-right mapping was robust to both verbal and visuospatial interference; that is, dual tasks that are typically considered "interference" did not in fact interfere with participants' left-right mental timelines.

We discuss the relationship between these results and those obtained for mental timelines that result from learning new spatiotemporal metaphors in language (Hendricks & Boroditsky, 2015), and the effects of the same interference tasks on number tasks (mental number-line and counting; van Dijck et al., 2009; Frank et al., 2012).

Experiment 1: Linguistic Stimuli

Methods

Participant & Inclusion Criteria

72 undergraduate students at the University of California, San Diego participated for course credit.

Participants were excluded if performance was below 20% on the visuospatial or verbal interference task (five participants) or below 50% on the time judgment task (seven participants). One additional participant was removed because the program crashed halfway through the experiment. This left a total of 58 participants for analysis (22 in the verbal interference condition, 18 in the control condition, and 18 in the visuospatial interference condition).

Only trials where participants responded accurately to the time judgments were included in the analysis (93.8% of trials). Furthermore, only reaction times that were within 3

standard deviations of each participant's cell mean (97.8% of correct trials) were included in analysis.

Procedure

The design of the experiment was modeled after van Dijck, Gevers, and Fias (2009), who conducted a similar dual-task experiment on spatial representation of number. Each participant was randomly assigned to one of three conditions: control (no interference), verbal interference, or visuospatial interference. First, a baseline measure of performance was taken for the time judgment task. Then, each participant's working memory span was measured using either a visuospatial or a verbal task. This span was used to calibrate the final portion of the study, a dual-task, which is described below.

Time Judgment Task

Participants held two computer mouses, one in each hand, with each thumb placed over a single mouse button. Participants held one mouse with their left hand on their left side and the other mouse in their right hand on their right side. Before each block, participants were presented with instructions that explained the stimulus-response mappings (e.g., left response for earlier events, right response for later events) they would use for that block. The stimulus-response mappings were changed after each block. After the presentation of a fixation cross for 1000ms, participants read a reference life event written in the center of the computer screen (e.g., "high school graduation"), which remained on the screen for 2000ms. A white screen was then presented for 500ms and the text of a second life event (e.g., "college graduation") was presented and remained on the screen until the participants responded, up to 5000ms. Reaction times were measured from the onset of the second event. Participants received new instructions before each block. Participants completed four practice trials, followed by forty experimental trials during each of two blocks. Whether the participants received congruent or incongruent mappings during their first block was counterbalanced across participants.

Interference Task / Working Memory Measures

After the baseline time judgment task was completed, the verbal or visuospatial working memory spans of the participants were measured, depending on which condition they were randomly assigned to. For both the visuospatial and verbal calibration tasks, strings of items were presented in an increasing number (from three to eight items, with three strings of items per testing length). The participant's span was then defined as the highest sequence length where they recalled at least two of the three strings for that length.

The verbal working memory span was modeled after Szmalec and Vandierendonck (2007). Each trial started with a blank screen, followed by a string of single consonants, which were each separated by an empty screen. Participants were then asked to type their responses after all of the consonants were presented.

A computerized Corsi task was employed to measure visuospatial working memory. Nine white squares were presented on a black background and were positioned in the same manner as used by van Dijck et al. (2009). Each trial started with an image of the white squares, followed by a sequence of squares flashing blue, one at a time. When the sequence was over, all of the squares were left on the screen and the participant had to reproduce the sequence by clicking on the squares in the order that they saw them flash.

Dual Task

Participants in the dual task conditions were first presented with either a verbal or visuospatial sequence to remember. To ensure the dual-task was challenging, but not too difficult, the participant's working memory span minus one was used to calibrate the dual task setup. They then completed 2 time judgment trials. Finally, they were asked to recall the verbal or visuospatial sequence. This cycle was repeated 20 times for each mapping (either incongruent or congruent), for a total of 80 temporal judgment trials and 40 span trials. Each participant completed two blocks of dual-task trials, with each block using a particular set of stimulus-response mappings (congruent or incongruent). Participants in the control condition simply completed the time judgment task once more, rather than completing the dual task version.

Results

Reaction Times

A 3 (type of interference group: control, verbal, visuospatial) x 2 (congruency: congruent or incongruent) x 2 (load type: baseline or under load, which includes the Control, Verbal, and Visuospatial Interference groups) ANOVA on reaction times revealed an overall congruency effect, $F(1,55)=25.62$, $p<.001$, where participants were faster to respond to congruent (M=1360 ms) than incongruent (M=1529 ms) trials. There was also a main effect of load type, $F(1,55)=5.88$, $p=.005$, where participants were faster under load (1345 ms) than at baseline (1544 ms). However, this doesn't imply that the "under load" condition was easier. Rather, the "under load" condition was also the second time they completed the time judgment task, so this effect likely reflects a practice effect. There was, however, an interaction between interference type and load condition, $F(2,55)=5.88$, $p=.005$. Follow-up tests revealed that while participants in the Control ($F(1,21)=30.51$, $p<.001$) and Verbal ($F(1,17)=21.42$, $p<.001$) interference conditions got faster the second time they completed the task, participants in the visuospatial interference condition showed no such improvement, $p=.62$.

There was no main effect of type of interference ($p=.66$), nor was there an interaction between interference condition and congruency ($p=.16$). There was also no interaction between congruency and load type, $p=.78$, suggesting that congruency effects did not change between the baseline and when the load was introduced. Finally, no three way interaction emerged, $p=.62$. Reaction times for the load conditions are shown in Figure 1a.

Time Judgment Accuracy

A 3 (type of interference: control, verbal, visuospatial) x 2 (congruency: congruent or incongruent) x 2 (load type: baseline or under interference) ANOVA was also conducted on the accuracy of the time judgment trials. There was an overall main effect of congruency, $F(1,55)=7.87$, $p=.007$, where participants were more accurate on congruent (94.8%) than incongruent (92.7%) trials. There was also a main effect of load, $F(1,55)=7.24$, $p=.009$, again reflecting practice effects, as participants were more accurate the second time through the task (94.8%) than during baseline (92.7%). There was no main effect of interference type, $p=.26$, nor were there any interactions.

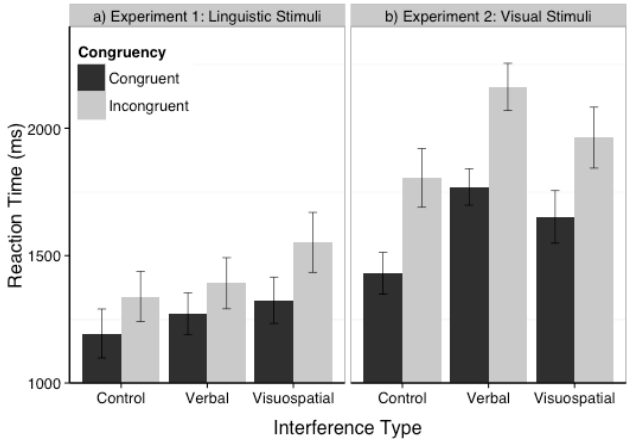


Figure 1 Reaction times to make time judgments in the main task in Experiment 1 (a) and Experiment 2 (b)

Interference Task

Based on the working memory span calibration, there was no difference in the number of interference items assigned to participants in the verbal ($M=4.0$, $SD=1.195$) and visuospatial (mean: 4.5, sd: 1.0) interference groups, $p=.15$.

Performance on the working memory interference tasks (visuospatial task or verbal task) was analyzed using a 2 (congruency: congruent/incongruent) x 2 (type of task: visuospatial or verbal) ANOVA on secondary task accuracy. Participants were more accurate in the interference task on trials that occurred with congruent time judgment trials (63.9%) than those that occurred with incongruent time judgment trials (57.6%), $F(1,38)=6.49$, $p=.015$. Participants also performed better on the verbal interference task (71.3%) than on the visuospatial interference task (47.9%), $F(1,38)=12.91$, $p<.001$. When the different interference tasks were analyzed separately, post-hoc paired t-tests revealed that participants performed better in the visuospatial interference task when the temporal judgment trial was congruent (52.6%) than when the temporal judgment was incongruent (43.4%), $t(17)=2.45$, $p=.025$. Such a congruency effect was not observed for the verbal interference task, $p=.25$ (see Figure 2a).

Discussion

In Experiment 1, we found no effect of visuospatial or verbal interference on the left-right mental timeline. Similar space-time congruency effects were observed in the control, verbal and visuospatial interference conditions, suggesting a robust left-right mental timeline. However, while an effect of interference on time judgment reaction times was not observed, evidence of interference appeared in other, more subtle, forms. First, as each participant completed the time judgment task twice (once for baseline and once under interference), one would expect them to improve the second time through the task. While this improvement was observed in the control and verbal conditions, it was not seen in the visuospatial interference condition. This suggests that participants in the visuospatial condition were more impacted by load than those in the verbal condition, reflected by slower reaction times the second time through in the visuospatial group. Second, participants performed worse on the visuospatial memory task during the incongruent blocks of the time judgment task than during the congruent blocks. This effect was not seen on the verbal memory task. While this analysis was exploratory, it suggests that the visuospatial interference task, but not the verbal interference task, requires some of the same cognitive resources as the main task. It is possible that the visuospatial interference did not impact congruency effects in the time judgment task because the stimuli were linguistic. To test for this possibility, Experiment 2 involved making similar time judgments, but the stimuli used were nonlinguistic, in the form of pictures.

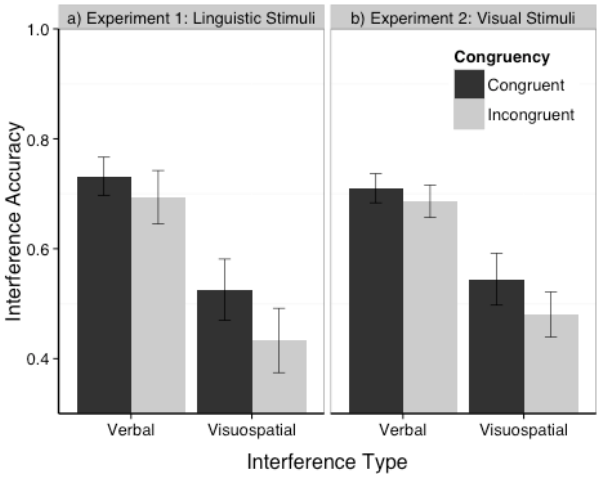


Figure 2. Interference Accuracy in both interference groups in Experiment 1 (a) and Experiment 2 (b).

Experiment 2: Visual Stimuli

Methods

Participants & Inclusion Criteria

A total of 102 UC San Diego undergraduates participated for course credit. They received a link to participate through the Psychology participant recruitment site.

We excluded 5 participants with accuracies below 25% for one time judgment task block and 7 additional participants whose accuracies on one or both blocks of the time judgment task were greater than 3 standard deviations from the sample mean accuracy. For the secondary interference task, participants who performed below 3 standard deviations from the sample mean accuracy for their interference type were eliminated. This resulted in eliminating only one participant from the verbal interference condition. After exclusions, all three conditions included 30 participants.

Only trials for which the time judgment was correct (91.0% of all trials) and reaction times were within 3 standard deviations of each participant's cell mean (97.9% of correct trials) were included.

Materials and Procedure

Each participant was randomly assigned to one of three conditions: control (no interference), verbal interference, or visuospatial interference. Participants in the interference conditions completed a calibration block at the beginning of the experiment. Everyone completed two blocks of the time judgment task with opposite key mapping instructions. The time judgment task included either visuospatial or verbal interference for participants in those conditions.

Interference Task Calibration

Participants in the verbal and visuospatial interference groups first completed a calibration task to ensure that the memory tasks were properly tuned for individual ability. The materials and calibration procedure were identical to those used in Frank et al (2012). Both conditions followed a staircase method, increasing difficulty (the number of distractors to remember) when participants got 2 consecutive trials correct, and decreasing it when they got one wrong. The number of items that a person could successfully remember was determined after 60 trials.

Time Judgment Task

The main task involved the same nonlinguistic stimuli and procedures used in prior work examining effects of writing direction and linguistic metaphors on representations of time (Fuhrman & Boroditsky, 2010; Hendricks & Boroditsky, 2015). Participants responded on a QWERTY keyboard. To begin each trial, they pressed the 'G' key. Then, the first of two images appeared on the screen (e.g., Julia Roberts in her 20s). After 2000ms, this image was replaced by a second image (e.g., either a younger or older Julia). For the congruent time judgment block, participants were instructed to press the 'D' key if the second image showed a conceptually earlier time point than the first, and the 'J' key if it showed a conceptually later point. This instruction was reversed for the incongruent block. All participants completed one congruent and one incongruent block, and whether the first block was congruent or incongruent was counterbalanced across participants.

There were 10 practice trials and 56 experimental trials in each block. Sequences of images were selected in a random order, and each sequence was shown only once in each block.

In the two interference groups, participants remembered either a string of consonants or a pattern of boxes while performing the time judgments. After completing the temporal judgment, people in the verbal interference group were prompted to type in the letter sequence they had been rehearsing. For the visuospatial interference group, the string of letters was replaced by a set of blue squares that appeared sequentially in different locations on a 4x4 grid of white blocks. After the time judgment trial, participants clicked on the appropriate boxes when given a blank grid.

Results

Reaction Times

As in Experiment 1, a 3 (interference type) x 2 (congruency) ANOVA revealed that participants were overall faster for congruent trials (M=1617ms) than incongruent trials (M=1977ms; $F(1,87) = 51.55$, $p < .00001$). This was true in all three conditions (all ps $< .0001$). The size of the congruency effect did not differ among the three conditions, confirmed by a lack of interaction between congruency and interference type ($p = .78$). Reaction times are shown in Figure 1b.

There was a main effect of interference type on overall reaction times, ($F(2,87) = 3.84$, $p = .03$). Pairwise comparisons demonstrated that those in the verbal interference group were slower overall than those in the control ($F(1,58) = 9.06$, $p = .004$), but not slower than those in the visuospatial interference group ($p = .21$). There was no difference in overall reaction times between the visuospatial interference and control groups ($p = .30$).

Time Judgment Accuracy

People were overall more accurate on congruent time judgment trials (92.4%) than on incongruent trials (89.6%; $F(1,87) = 14.93$, $p = .0002$). There was also a main effect of interference type on time judgment accuracies ($F(2,87) = 4.73$, $p = .01$). Pairwise tests revealed that participants in the visuospatial interference group were less accurate on the time judgment task (88.8%) than the participants in the verbal interference (91.5%; $F(1,58) = 4.73$, $p = .03$) and control (92.6%; $F(1,58) = 7.54$, $p = .008$) groups. There was no difference in accuracy between the verbal interference and control groups ($F(1,58) = 0.77$, $p = .38$).

Interference Task

The number of distractor items assigned to the participants for the time judgment task based on calibration was the same for both the verbal (M=4.47, SD = 1.14) and visuospatial (M=4.90, SD = 1.42) interference groups ($p = .20$).

Accuracies on the secondary memory task were higher for congruent trials (62.7%) than incongruent trials (58.3%; $F(1,58) = 6.05$, $p = .02$). They were also higher for participants who experienced verbal interference (69.8%) than those who experienced visuospatial interference (51.2%; $F(1,58) = 14.2$, $p = .0003$). As in Experiment 1, a post-hoc analysis was conducted on accuracies for each interference group separately. Participants in the visuospatial interference group were more accurate at

remembering the box pattern when the time judgment trial was congruent than when it was incongruent ($t(29) = 2.30$, $p = .03$), but there was no difference in accuracy between congruent and incongruent trials in the verbal interference group ($t(29) = 1.06$, $p = .30$). Interference task accuracies are shown in Figure 2b.

Discussion

In this experiment, the canonical left-right mental timeline was again robust to both visuospatial and verbal interference conditions. Despite the use of visual (picture) stimuli, the canonical mental timeline persisted amidst a visuospatial load, suggesting that people are able to recruit space to think about time while also maintaining visuospatial information in memory. As in Experiment 1, however, there was a subtle indication that making time judgments while experiencing a visuospatial load was more difficult than making the same judgments under a verbal load. Exploratory analyses revealed that accuracy for the visuospatial interference task was worse in incongruent trials than in congruent, but this same discrepancy was not present in the verbal interference condition.

General Discussion

Across two experiments, American undergraduates consistently demonstrated strong congruency effects (faster responses for early on the left and later on the right than for the reverse), even when faced with visuospatial or verbal dual task. These effects were also robust to differences in stimuli presentation, as they were observed in cases where the stimuli were linguistic (Experiment 1; words presented on a screen), as well as nonlinguistic (Experiment 2; images presented on a screen). However, though the interference tasks did not affect performance on the time judgment task, exploratory analyses revealed that the time judgment task appeared to affect performance on the visuospatial interference task. In both studies, participants performed worse on the visuospatial interference task during the incongruent blocks of the time judgment task than during the congruent blocks. This effect was not seen in the verbal memory task. However, due to the post-hoc nature of those analyses, future work should more carefully investigate the influence of temporal reasoning on visuospatial working memory performance.

These findings lead us to ask: what is the nature of the representation that underlies American college students' left-right mental timelines? What is it most like? Does it have important similarities or differences from the mental number line effects observed in prior work (e..g, van Dijck et al., 2009)? Or is it more like spatial representations for time that are shaped by linguistic forces?

Connection to Mental Number Lines

Left-right mental timelines are often compared to mental number lines, as they both express serial order in terms of spatial position across the lateral axis (Bonato et al., 2012). In both cases, Americans associate earlier items (e.g., 1, 2, "yesterday") with the left side of space and later items (e.g., 7, 8, "tomorrow") with the right side of space. Furthermore, writing direction influences the directionality of both mental timelines (Tversky, Kugelmass & Winter, 1991) and mental number lines (Dehaene, Bossini & Giraux, 1993). However, the persistence of a left-right mental timeline under conditions of verbal and visuospatial interference contrasts with work on the mental number line in which the same visuospatial interference task used here eliminated the bias to associate smaller numbers with left-space and larger numbers with right-space (van Dijck et al., 2009). In addition, while verbal interference did not affect the presence of a left-right mental timeline in our participants, the same verbal interference tasks have eliminated evidence of a spatial representation for number when making parity decisions (van Dijck et al., 2009), as well as eliminating people's ability to count (Frank et al., 2012).

Other work has observed differences in the mental timelines and mental number lines of blind people (Bottini et al., 2015). While the mental number line is anchored on an anatomical frame of reference for blind people and on an external frame of reference for sighted people, the mental timeline is anchored on an external frame of reference for both sighted and blind people. This is interpreted as evidence of different experiential bases for spatial representations of time and number.

Linguistic Metaphor and Mental Timeline

Our mental timelines can be shaped by the linguistic metaphors we use to talk about time (Hendricks & Boroditsky, 2015). Learning a new metaphor (that placed earlier events as either above or below later ones) in a lab setting fostered new metaphor-consistent vertical representations for time. The robustness of those new representations was tested using the same dual task conditions used in Experiment 2. Consistent with the findings we report here, neither the visuospatial nor the verbal dual task interfered with the metaphor-consistent representation for time. Together, the novel metaphor experiment and experiments reported here suggest that our spatial representations of time, though evident in a nonlinguistic implicit association task, do not rely on the same cognitive resources as those needed to complete the visuospatial and verbal dual tasks described here.

The similarities in the robustness of the space-time association between the metaphor work and current studies are notable because the origins of the associations in these experiments were very different. In the case of linguistic metaphor, participants learned a new metaphor in the lab. They practiced that metaphor for about 10 minutes before showing evidence of nonlinguistic associations consistent with the metaphors they had just acquired. Associations between space and time on the lateral axis, however, are absent from the English language. English speakers do not refer to Wednesday as to the left of Thursday, or moving a meeting to the right by 2 hours, for example. Although these associations are not evident in spoken language, they follow

the direction of reading and writing (Tversky, Kugelmass & Winter, 1991, Fuhrman & Boroditsky, 2010), an experience that accumulates throughout a person's life. Despite very different origins and developmental time scales, the robustness of the spatial representation of time remained consistent.

Spatial Representations

It is important to remember that although we use the same word, *spatial*, in different contexts, our spatial abilities are not one monolithic entity. Our conceptualization of number and time may both be grounded in an understanding of space, but as we have explored here, those spatial representations need not be the same. Similarly manifesting spatial representations can also come from different origins, developed on different time scales, though this does not necessarily mean that the representations truly share the same nature either. Because *spatial* cognition is not a single concept, a single dual task cannot be expected to interfere with all mental processes that rely on some form of spatial cognition.

In addition to exploring additional dual tasks that may interfere with congruency effects while making time judgments, future work might explore individual differences in spatial abilities. For example, Viarouge and colleagues (2014) found that stronger spatial computation and numerical semantic skills both independently predicted greater SNARC effects, suggesting that spatial conceptualizations of number are influenced by both spatial computation and number semantics. A similar individual differences approach could be useful in exploring the nature of spatial representations for time.

Conclusion

Taken together, the experiments reported here conceptually replicate findings showing that American college students think of sequences of events as unfolding from left to right (Weger & Pratt, 2008). They also demonstrate the persistence of this mental timeline to traditional interference tasks that eliminate mental number line effects involving parity judgments and counting (verbal interference; van Dijck et al., 2009; Frank et al., 2012) and magnitude judgments (visuospatial interference; van Dijck et al., 2009). These experiments highlight a difference between the way we think about time and number, encouraging us to better understand the complexity of what it means for an abstract representation to be spatial.

Acknowledgments

E.W. would like to thank Alyssa O'Neill and Tara Loveland for help with Experiment 1, which was supported by a NSERC PGS-D awarded to EW. Work on Experiment 2 was supported by a National Science Foundation award BCS 1058119 to L.B.

References

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: Evidence for a mental time line. *Neuroscience & Biobehavioral Reviews*, *36*(10), 2257–2273.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*(1), 1–28.

Bottini, R., Crepaldi, D., Casasanto, D., Crollen, V., & Collignon, O. (2015). Space and time in the sighted and blind. *Cognition*, *141*, 67–72. http://doi.org/10.1016/j.cognition.2015.04.004

Cooperrider K. & Núñez, R. (2009). Across time, across the body: Transversal temporal gestures. *Gestures, 9*(2), 181-206.

Dehaene, S., Bossini, S. & Giraux, P. (1993). The mental representation of parity and number magnitude. *J Exp Psych: General, 122*(3), 371-396.

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive Psychology*, *64*(1-2), 74–92.

Fuhrman, O., & Boroditsky, L. (2010). Cross-Cultural Differences in Mental Representations of Time: Evidence From an Implicit Nonlinguistic Task. *Cognitive Science*, *34*(8), 1430–1451.

Núñez, R.E. & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences, 17*, 220-229.

Saj, A., Fuhrman, O., Vuilleumier, P., & Boroditsky, L. (2014). Patients With Left Spatial Neglect Also Neglect the "Left Side" of Time. *Psychological Science*, *25*(1), 207–214.

Szmalec, A., & Vandierendonck, A. (2007). Estimating the executive demands of a one-back choice reaction time task by means of the selective interference paradigm. *The Quarterly Journal of Experimental Psychology*, *60*(8), 1116–1139.

Tversky, B., Kugelmass, S. and Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology, 23*, 515-557.

Van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition*, *113*(2), 248–253.

Viarouge, A., Hubbard, E. M., & McCandliss, B. D. (2014). The Cognitive Mechanisms of the SNARC Effect: An Individual Differences Approach. *PLoS ONE*, *9*(4), e95756.

Weger, U. W., & Pratt, J. (2008). Time flies like an arrow: Space-time compatibility effects suggest the use of a mental timeline. *Psychonomic Bulletin & Review*, *15*(2), 426–430.

Ulrich, R., Eikmeier, V., de la Vega, I., Fernández, S.R., Alex-Ruf, S., & Maienborn, C., (2012). With the past behind and the future ahead: Back-to-front representation of past and future sequences. *Memory and Cognition,* 40, 483-495.

A speed-accuracy trade-off in children's processing of scalar implicatures

Rose M. Schneider
rschneid@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Scalar implicatures—inferences from a weak description ("I ate some of the cookies") that a stronger alternative is true ("I didn't eat all")—are paradigm cases of pragmatic inference. Children's trouble with scalar implicatures is thus an important puzzle for theories of pragmatic development, given their communicative competence in other domains. Previous research has suggested that access to alternatives might be key. Here, we explore children's reaction times in a new paradigm for measuring scalar implicature processing. Alongside failures on scalar implicatures with "some," we replicate previous reports of failures with "none," and find evidence of a speed-accuracy trade-off for both quantifiers. Motivated by these findings, we explore the relationship between accuracy and reaction time with a Drift Diffusion Model. We find evidence consistent with the hypothesis that preschoolers lack access to the alternatives for scalar implicature computation, although this set of alternatives may be broader than previously assumed.

Keywords: Pragmatics; development; scalar implicature; diffusion models.

Introduction

Language comprehension in context is an inferential process. Listeners are not limited to interpreting the literal meaning of speakers' utterances; they can also reason about what the speaker intended, based on alternative utterances. In the case of *pragmatic implicatures* (Grice, 1975), a speaker employs a weaker literal description to imply that a stronger alternative is true. Adult listeners tend to infer from the statement "I ate *some* of the cookies" that some, but not all, of the cookies remain. This *scalar implicature* (SI) relies heavily on a knowledge of the relevant lexical alternatives in the quantifier scale <*some*, *all*>. On standard theories, a listener must be able to contrast "some" with the stronger descriptor "all" to compute the implicature (Grice, 1975; Levinson, 2000).

SIs are challenging for children until surprisingly late in development (Noveck, 2001). For example, when judging a scene in which three of three horses have jumped over a fence, five-year-olds are likely to endorse the statement "some of the horses jumped over the fence" as felicitous, despite the availability of a more informative alternative ("all"; Papafragou & Musolino, 2003). Children do seem to have some knowledge of these scalar terms, however; for example, they differentially reward speakers based on the informativeness of their scalar descriptions (Katsos & Bishop, 2011). Given this early sensitivity, why do children still struggle to compute scalar implicatures until fairly late in development?

One possible cause of children's failures is that they may not have access to relevant lexical alternatives (Barner & Bachrach, 2010). This idea, which we will refer to as the *Alternatives Hypothesis*, predicts that if children cannot quickly and reliably bring to mind the relevant alternative quantifiers (e.g., "all" in a situation where they hear "some") they will be unable to compute the implicature. The alternatives hypothesis makes a number of predictions about children's abilities in reasoning about quantifiers, some of which have been confirmed empirically. For example, consistent with the idea of inaccessible alternatives, Barner, Brooks, & Bale (2011) showed that four-year-olds could not compute the quantifier expression "only some" (which should force alternatives to be negated semantically, rather than pragmatically). In light of this hypothesis, what are the proper alternatives for SIs?

Here, empirical evidence has been changing rapidly. Although the conventional view on SI is that the primary inferential alternative is "all," a new body of evidence suggests that more alternatives may be necessary. For example, Degen & Tanenhaus (2015) found that set size can change the felicity of quantifier SIs for adults: "some" is more felicitous when participants could not say "one" or "two." In a computational reanalysis of Degen & Tanenhaus (2015) and other data, Franke (2014) showed that a high weight on the alternative "none" was critical for fitting these data. And in a recent study with children, Skordos & Papafragou (in press) found that exposing children to either "all" *or* "none" facilitated computation of subsequent SIs. Taken together, these data suggest that the availability of alternatives—particularly "none"—does affect scalar implicature processing.

This relationship to "none" is unexpected on classic Gricean theories (Grice, 1975; Horn, 1972), where the only alternatives should be those logically entailed by the original message (i.e. "all"). But it *is* in fact predicted by recent probabilistic models of implicature. Under these models, all the relevant alternatives compete with one another (Franke, 2014; Goodman & Stuhlmuller, 2013). On the other hand, all of the evidence cited above for the claim of "none" as an alternative is relatively indirect, and such a substantial revision to theory requires further evidence.

One other recent developmental study further supports the importance of "none" in SIs and provides the starting point for our current experiment. Horowitz & Frank (2015) designed a referent selection paradigm that could be used across a broad age range (3–5 years) to explore both scalar and ad-hoc (context-dependent) implicatures. In this task, children saw three book covers, each featuring four familiar objects (Figure 1). On target trials, the experimenter described a book using a semantically ambiguous description (e.g., "On the cover of my book, some of the pictures are cats" [scalar] or "On the cover of my book are cats" [ad hoc]). Children succeeded on ad-hoc trials but largely failed to make SIs, sug-

gesting they had the pragmatic competence necessary to compute the implicature, but failed to do so for scalar descriptors.

Interestingly, in Horowitz & Frank (2015), the same children who failed on SI also failed on unambiguous "none" control trials (e.g., "On the cover of my book, *none* of the pictures are cats""—and in several samples, performance was highly correlated between"none" and "some" trials. This result would be predicted if "none" were in fact an inferential alternative to "some." If children were not computing its semantics appropriately in an online fashion, they would fail in the "some" SI computation as well, leading to a correlation.

One further prediction of the alternatives hypothesis relates to processing time. Perhaps children who have a fully-established quantifier scale—and hence can make correct SIs—take additional time in using this information, due to competition between alternatives. Congruent with this prediction, our intuition in Horowitz & Frank (2015) and Horowitz, Schneider & Frank (in prep.) was that when children made correct SIs they appeared to be taking longer than when they failed. Motivated by this observation, we adapted our SI task for the iPad to collect detailed and accurate developmental reaction time data. Although reaction time measures have been commonplace in studies of adults' SI processing, they have been almost entirely absent in the developmental literature (with the exception of Huang & Snedeker, 2009, whose data showed little evidence of SI computation).

Thus, in our current study, we explore children's response latencies in an iPad adaptation of the Horowitz & Frank (2015) SI task. In our analyses, we examine overall accuracy and patterns of performance, as in Horowitz & Frank (2015), and find that children not only struggle in making SIs, but replicate the finding that they have difficulty with "none" until fairly late in development. Congruent with our predictions, in reaction time analyses we find evidence of a speed-accuracy trade-off for both quantifiers, such that children who succeed exhibit longer response latencies. Finally, we use a Drift Diffusion Model (Ratcliff, 1978; Ratcliff & Rouder, 1998) to explore the source of this increased reaction time. Overall, our findings are consistent with a version of the Alternatives Hypothesis under which "none" is an important inferential alternative in SI and its availability causes slower processing times but correct SIs.

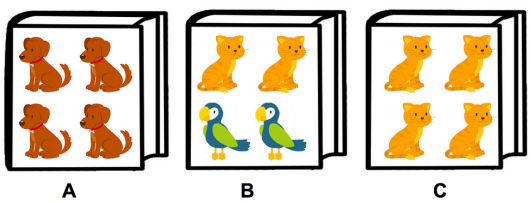


Figure 1: Example trial stimuli used in Horowitz and Frank (2015).

Age group	N	Mean	Median	SD
3–3.5 years	24	3.27	3.28	0.15
3.5–4 years	35	3.78	3.73	0.15
4–4.5 years	25	4.27	4.28	0.15
4.5–5 years	30	4.76	4.76	0.15
5–6.5 years	24	5.55	5.59	0.37

Table 1: Age information for all participants.

Method

We adapted the scalar implicature paradigm developed by Horowitz & Frank (2015) for the iPad. In addition to capturing reaction time (RT) data, this version included more trials, and standardized prosody across all trials, as well as a randomized design.[1]

Participants

Table 1 shows the breakdown of participant age information. Included are 138 children out of a planned sample of 120 participants, recruited from both a local daycare and children's museum. 39 additional children were excluded from analysis based on planned exclusion criteria of low English language exposure ($\leq 75\%$) or $< 50\%$ of trials completed. Included in our sample were 79 females and 59 males.[2]

Stimuli and design

The general format of the task was identical to Horowitz & Frank (2015), with the exception of added items for additional trials. The study was programmed in HTML, CSS, and JavaScript, and displayed to children on a full-sized iPad. Each trial displayed three book covers, each containing a set of four familiar objects (Figure 1). Each session involved 30 trials, with 10 trials per quantifier ("all", "some", and "none"). Each audio clip used the same three initial sentence frames (e.g., "On the cover of my book, *some* of the pictures…") to emphasize prosody equally across all trials. The average length of each clip (including target item phrase, e.g., "…are cats") was approximately 6s. Quantifier triad order, items (within category), target item, and quantifier were randomized for all participants. There were 270 different target items and audio clips.

Procedure

Sessions took place individually in a small testing room away from the museum floor or the classroom of the daycare. To familiarize children with the iPad, each session began with a

[1]The full experiment can be viewed online at https://rosemschneider.github.io/tablet_exp/si_tablet.html and all of our data, processing, experimental stimuli, and analysis code can be viewed in the version control repository for this paper at: https://github.com/rosemschneider/SI_tablet.

[2]Based on Horowitz & Frank (2015), we initially planned to collect data from children aged 3–5 years. After 57 participants, however, we observed significantly lower performance on implicature trials across all age groups, indicating that the iPad scalar implicature task was slightly more challenging, and included an older age group of 24 5–6.5-year-olds.

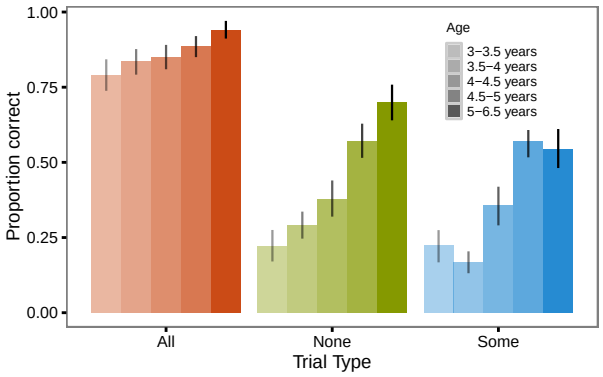


Figure 2: Children's overall accuracy for each quantifier type. Bars show mean performance for each age group. Error bars are 95-percent confidence intervals.

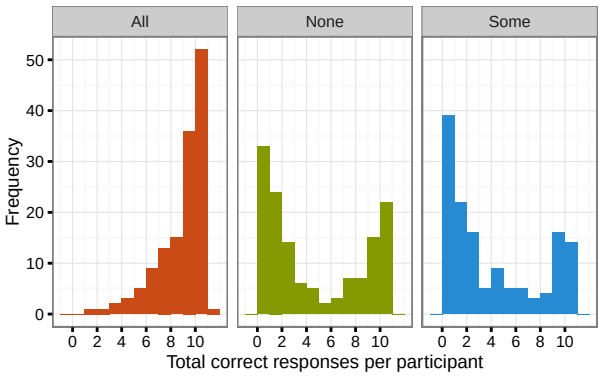


Figure 3: Frequency histogram of participant totals for each trial type, across all participants.

"dot game," which required them to press dots on the screen as fast as possible. After the dot game, the experimenter introduced them to "Hannah," a cartoon character who wanted to play a guessing game with her books. The experimenter explained that Hannah would show the child three books, and would give one hint about which book she had in mind, so they had to listen carefully. Children then saw a practice trial with an unambiguous noun referent.

Each trial allowed 2.5s for children to visually inspect the book covers before the prompt played (e.g., "On the cover of my book, *none* of the pictures are cats."). Reaction times were measured from the onset of the target word. Children could only make one selection. If a child did not hear Hannah's prompt, the experimenter repeated it, matching the original prosody. Once children made their selection, a green box appeared around the chosen book. The experiment was self-paced, and children initiated each trial by pressing a button that appeared after they made their selection.

Results

For trials where the child had missed the prompt or was not paying attention, we excluded reaction times (RTs) longer than 15s. After this initial cut, we excluded RTs outside three standard deviations of the log of mean reaction time. This cleaning resulted in RT data loss for 85 trials (2.09%). We observed a fairly wide RT distribution for all trial types. This variability in children's RTs may have been exaggerated by testing in a museum environment, or by the motoric demands associated with making a response on a tablet (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). "Success" is defined as choosing the image consistent with the scalar description (e.g., selecting the book with two cats and two birds in response to the prompt "On the cover of my book, *some* of the pictures are cats.").

Accuracy

Figure 2 shows children's mean performance for each trial type, split by age group. For each age group, we saw significantly lower accuracy for the quantifiers "some" and "none" in comparison to "all" (all ps $< .01$ in two-sample t-tests for each age group). These results generally replicate our previous findings using this paradigm (Horowitz & Frank, 2015); one difference from previous results was in implicature trials. Children aged 3–5 years performed significantly lower on "some" (implicature) trials in this task in comparison to Horowitz et al. (in prep.) ($p < .01$ for all tests). Thus, while the iPad adaptation was generally successful, implicatures were more difficult, perhaps because of the non-social nature of the iPad interaction.

We next fit a logistic mixed effects model predicting correct response as an interaction of age and trial type, with random effects of trial type and participant.[3] Performance was significantly lower on "some" ($\beta = -6.72$, $p < .0001$) and "none" trials ($\beta = -9.44$, $p < .0001$). There was also a significant interaction between age and trial type on "none" trials ($\beta = 1.52$, $p = .0005$), indicating that children's performance with this difficult quantifier increased with age. This model also showed that children's performance showed a trend towards significance for "some" trials ($\beta = 0.76$, $p = .051$).

Figure 3 shows distributions of correct responses for all trial types. Performance on "some" and "none" trials was bimodal (Hartigan's $D = 0.08$, $p < .0001$) and "none" trials ($D = 0.11$, $p < .0001$). While children's average accuracy was low for these quantifiers, there were some children who were correct on the majority of these trials ("Some": N = 30; "None": N = 37) and the others were typically incorrect on the majority of trials. Children did not appear to be responding randomly. As in previous work, we found a strong correlation between children's accuracy on "some" and "none" trials ($r = 0.49$, $p < .0001$). Children also exhibited some interesting systematicity in their errors: on incorrect "some" trials they overwhelmingly chose "all," while on "none" trials they also chose "some" at about half the rate of "all."

[3] All mixed effects models were fit in `R` using the `lme4` package. The model specification was: `correct ~ age * trial type + (trial type | subject id)`.

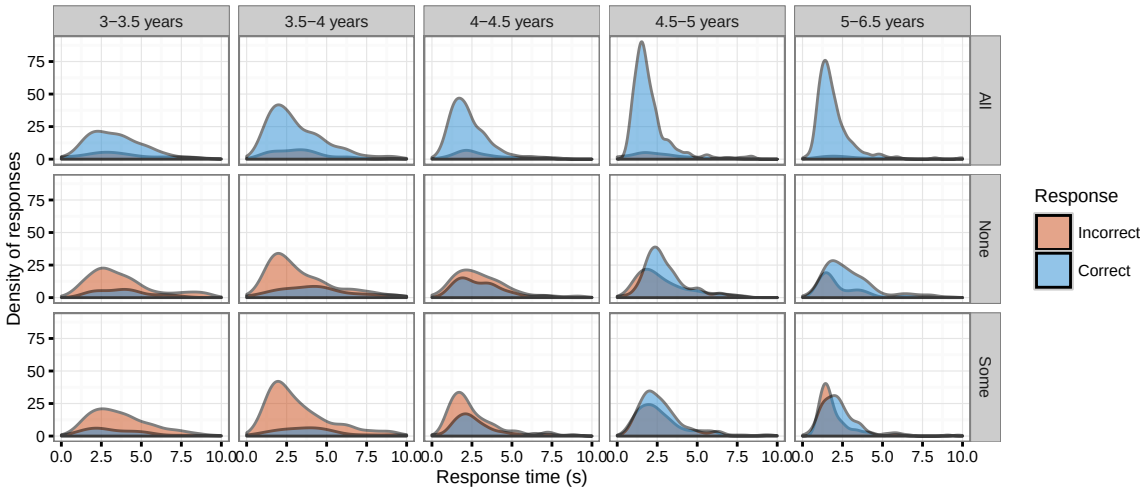


Figure 4: Density plots of reaction times for correct and incorrect responses on each trial type, split by age.

Reaction time

We fit a linear mixed effects model predicting log RT on correct trials as a function of log trial number, the interaction of age and trial type, and random effects of trial type by subject.[4] Reaction times were longer on "none" ($\beta = 0.38$, $p < .0001$) and "some" trials ($\beta = 0.22$, $p < .0001$), and reaction times decreased with age ($\beta = -0.29$, $p < .0001$). There were no significant interactions between age and trial type. The model also showed a main effect of trial number, with reaction times decreasing over the course of the study (β = -0.1, $p < .0001$).

Examination of the pattern in Figure 4 suggests that accuracy and reaction time may interact, however. In particular, while correct responses on "all" trials appear to be faster than the (few) incorrect responses, the opposite is true for "none" and "some" trials: Errors have faster RTs, potentially indicating a speed-accuracy trade-off. To test for this effect, we fit another mixed effects model, this time including accuracy and its interactions with age and trial type as predictors. This model revealed that correct trials overall had faster RTs ($\beta = -0.16$, $p = .0002$), but that this accuracy term interacted negatively with trial type such that both "none" and "some" trials had slower RTs for correct trials ($\beta = 0.34$, $p < .0001$; $\beta = 0.27$, $p < .0001$). There were no three-way interactions of trial-type and age. This model thus provides evidence of a speed-accuracy trade-off for "some" and "none" trials.

Drift diffusion models

Motivated by the evidence of a speed-accuracy trade-off, we further explored the interaction between reaction time and accuracy in more depth using drift diffusion modeling (DDM). Response latencies associated with "some" and "none" indicated that these longer RTs were associated with higher accuracy, but what components of the decision process contribute to this finding? DDM can be used in behavioral tasks to provide a more detailed view of the relationship between accuracy and reaction time (Ratcliff & Rouder, 1998). In DDM, a behavioral response (a correct or incorrect choice) is the result of noisy data accumulation through a diffusion process. Responses have *separation boundaries* that are dependent on the amount of information needed to initiate a response; *drift rate* formalizes the rate of data accumulation. *Nondecision* is the amount of time between stimuli offset and initiating the diffusion process (i.e., encoding). Finally, different responses may be *biased* in their starting point in the diffusion process. Thus, a DDM can reveal more differences in the SI decision-making process, and provide clues about the causes underlying this speed-accuracy trade-off.

Developmental analyses Although DDMs are traditionally fit to data from two-alternative forced-choice tasks, here we estimate the drift process between a correct and incorrect choice, with two options in each trial being "incorrect," and only one being consistent with the target noun and quantifier. We estimated parameters for each subject for each trial type using the `RWiener` package. We then aggregated across subjects to obtain means and confidence intervals for each age group. Figure 5 shows the parameter estimates for each age group, split by trial type. To fit the model, we excluded 85 trials with outlier RTs.

For each parameter estimate, we ran a mixed effects model, predicting parameter value as an interaction of age and trial type.[5] There was no significant effect of trial type in boundary separation, indicating that roughly the same amount of information is needed to make a decision for each quantifier. This should be expected, given our experimental design. For non-decision time, we found a significant main effect of age (β = -0.28, $p < .00001$), as well as an interaction between

[4]Model specification: `log(reaction time) ~ log(trial number) + age * trial type + (trial type | subject id)`. Age was centered for ease of interpretation of coefficients, and we calculated p values via the $t = z$ approximation.

[5]The specifications for all parameter models are as follows: `Parameter Value ~ age * trial type + (1 | subject ID)`

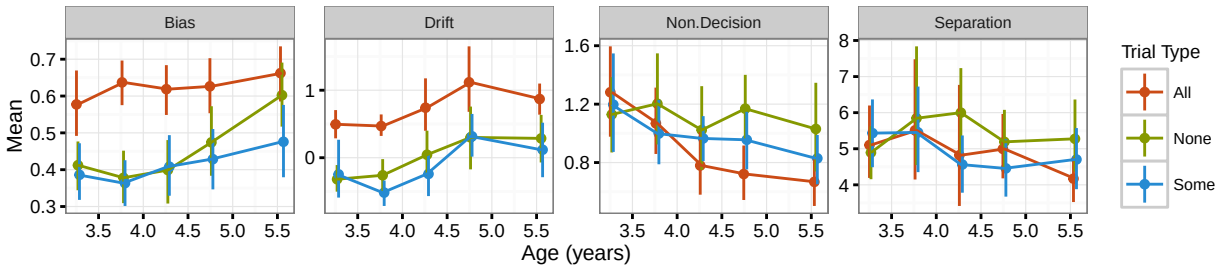


Figure 5: Parameter estimates for drift diffusion model, split by age and trial type. Error bars are 95 percent confidence intervals computed by nonparametric bootstrap.

age and "none" trials ($\beta = 0.24$, $p = .01$). As expected in drift rate, there was a negative main effect of trial type ("None": $\beta = -1.3$, $p = .0185$; "Some": $\beta = -1.18$, $p = .03$). Interestingly, for bias there was a significant negative effect of "none" trials ($\beta = -0.5$, $p = .0005$), and "some" trials trended towards significance ($\beta = -0.28$, $p = .0503$), as well as a significant interaction between age and "none" trials ($\beta = 0.07$, $p = .023$).

In sum, the parameter estimates from our DDM align with the analyses presented above: Older children are more likely to respond correctly in our scalar implicature task, while younger children's failures appear to be due to a low rate of data accumulation and a high separation boundary. One hypothesis for older children's successes in our task, and their trend towards more positive drift rates and lower biases, is rooted in a firmer grasp of the quantifier scale. In Horowitz et al. (in prep.), we observed a significant correlation between children's performance on a quantifier-knowledge task and SI computation. We found that this correlation increased with age, and that difficulties in computing an SI may be grounded in quantifier comprehension. It is possible that older children's SI success, and the observed speed-accuracy trade-off, may reflect increased knowledge of available lexical alternatives—particularly "none"—and the cost of comparing contrasting quantifiers to make SIs.

Exploratory analyses We also conducted an exploratory analysis, examining differences in the decision-making process for children who consistently made SIs compared with those who did not. This exploratory analysis was motivated by the hypothesis that successful participants in this task might display different signatures in their decision processes. Therefore, we split children into two groups by accuracy on scalar implicature trials, and then estimated parameters by accuracy group. High accuracy was defined as an average of 75% or higher performance on scalar implicature trials. Figure 6 shows parameter estimates for each accuracy group, split by trial type.

We again used mixed-effects models to predict DDM coefficients across participants. As in the developmental DDM analysis, there were no significant effects of separation or non-decision. While drift rates showed a significant effect of accuracy, because we estimated parameters for high- and low-accuracy children separately, these differences are expected.

In our bias estimates, however, we found a significant interaction between accuracy group and trial type on "some" trials ($\beta = -0.18$, $p = .0013$). This interaction suggests that bias (the starting point in the diffusion process) might be an important factor in successfully making a scalar implicature: More successful children were less biased towards incorrect response alternatives, perhaps due to greater knowledge about the quantifier scale.

General Discussion

What makes scalar implicatures using quantifiers so hard for children? The best current hypothesis posits that children do not have access to the appropriate inferential alternatives and hence fail to consider them in their pragmatic computation (Barner & Bachrach, 2010; Barner et al., 2011). But what are those alternatives? Recent work has suggested that the negative alternative "none" may compete with "some" and "all" when making SIs. Although "none" is not typically considered an alternative in Gricean theories (Grice, 1975; Horn, 1972), it nevertheless provides a relevant lexical alternative along the quantifier scale. Our findings here are consistent with this account and provide some additional support. Using a new method, we replicated the pattern found in previous studies that those children who succeed in comprehending the quantifier "none" are also able to make SIs (Horowitz & Frank, 2015; Horowitz et al., in prep.). In addition, our data revealed a speed-accuracy trade-off, such that reaction times in those trials for which children succeeded in making SIs were slower overall.

One interpretation of this speed-accuracy trade-off is that children who have more inferential alternatives accessible to them (e.g. are considering "none," "some," and "all" together) are both better at making SIs and slower to make them due to the processing cost of making the inference. This theory makes the prediction that children who have greater quantifier knowledge will be more successful in computing SIs, but will display higher RTs due to competing scalar alternatives. Our data are consistent with this account, which is also supported by an exploratory drift diffusion model analysis. We fit a DDM to our data for children who succeeded in making scalar implicatures versus children who failed. The model suggested that bias in "some" and "none" trials might be a key factor related to success—that is, children who were con-

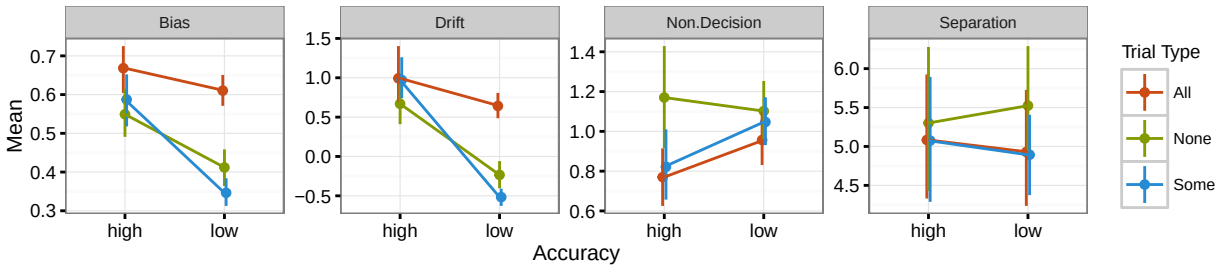


Figure 6: Parameter estimates for drift diffusion model, split by accuracy and trial type. Error bars are 95 percent confidence intervals computed by nonparametric bootstrap.

sidering "some" and "none" responses equally in their decision were more likely to make the SI. Both of these findings are again consistent with the idea that weighing alternatives appropriately in the SI computation is critical to success.

The speed-accuracy patterns we report are correlational, however, and other accounts are consistent with these findings as well. For example, some third factor (say inhibitory control) could underlie the ability to succeed in "some" and "none" trials and also explain why some children are able to inhibit their response long enough to complete the SI computation. Horowitz et al. (in prep.) did not find evidence of correlations between individuals' SI abilities and their executive function using one popular measure (the dimensional change card sort). Other versions of this account (or other accounts entirely) are still possible, however, and should be explored in future work. Nevertheless, our present work suggests that there is a meaningful relationship between children's accuracy and processing times in making scalar implicatures.

Acknowledgements

Thanks to Bing Nursery School and the San Jose Children's Discovery Museum. Thanks also to Veronica Cristiano, Rachel Walker, and Tamara Mekler for their help with data collection, and to Kara Weisman and Ann Nordmeyer for their assistance creating stimuli. This work was supported by NSF BCS #1456077.

References

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology*, *60*(1), 40–62.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in childrens pragmatic inference. *Cognition*, *118*(1), 84–93.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive Science*, *39*(4), 667–710.

Frank, M. C., Sugarman, E., Horowitz, A., Lewis, M., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1).

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 487–492).

Goodman, N. D., & Stuhlmuller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3). New York: Academic Press.

Horn, L. R. (1972). *On the semantic properties of logical operators.* (PhD thesis). University of California, Los Angeles.

Horowitz, A., & Frank, M. C. (2015). Sources of developmental change in pragmatic inferences about scalar terms. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.*

Horowitz, A., Schneider, R. M., & Frank, M. C. (in prep.). The trouble with quantifiers: Exploring children's deficits in scalar implicatures.

Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, *86*(3), 253–282.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychologial Science*, *9*(5), 347–356.

Skordos, D., & Papafragou, A. (in press). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*.

Systematic feature variation underlies adults' and children's use of *in* and *on*

Kristen Johannes (kjohann@WestEd.org)
WestEd, 300 Lakeside Dr., 25th floor, Oakland, CA 94609, USA

Colin Wilson (wilson@cogsci.jhu.edu)
Department of Cognitive Science, Johns Hopkins University, Krieger Hall
3400 N Charles St., Baltimore, MD 21218, USA

Barbara Landau (landau@cogsci.jhu.edu)
Department of Cognitive Science, Johns Hopkins University, Krieger Hall
3400 N Charles St., Baltimore, MD 21218, USA

Abstract

The spatial prepositions *in* and *on* apply to a wide range of containment and support relations, making exhaustive definitions difficult. Theories differ in whether they endorse geometric or functional properties and how these properties are related to meaning and use. This study directly examines the roles of geometric and functional information in adults' and children's use of *in* and *on* by developing a large sample of relations situated within a small gradable geometric and functional feature space. We propose that variation in features across items is systematically related to the use of *in* and *on* and demonstrate that feature-language relationships change across development: adults' expression use is sensitive to both geometric and functional features, while children's use varies only according to geometric features.

Keywords: Spatial language; spatial cognition; acquisition; language use

Background

All natural languages have a limited system of terms dedicated to the expression of spatial relationships between objects (Talmy, 1985). In the current study, we focus on the English prepositions *in* and *on*, considered the primary vehicles for expressing containment and support relations, respectively (Bowerman, 1996; Landau & Jackendoff, 1993; Talmy, 1985; Gentner & Bowerman, 2009; Herskovits, 1986, *inter alia*). Within these broad relationships, terms *in* and *on* cover a wide range of object configurations: from *apples in bowls* and *books on tables,* to *pieces in a puzzle* and *fish on a line*. The mapping between these expressions and object configurations must abstract over many fine-grained properties of individual objects, and capturing the critical properties has been challenging, especially given the early acquisition of these terms by children. Some theories endorse coarse-grained geometric properties of object representations, such as volumes, surfaces, and axes (Landau & Jackendoff, 1993), while others propose various functional properties of containment and support as relations between any objects (e.g., Garrod et al., 2009).

In this paper, we directly examine the role of geometric and functional information in English speakers' spatial uses of *in* and *on* and develop two hypotheses to address the following two questions: 1) How are geometric and functional information jointly engaged in the use of *in* and *on*? And 2) How does the relationship between functional and geometric knowledge and language develop over the course of acquisition?

Although *in* and *on* are are among the earliest words acquired by children (Johnston & Slobin, 1979), they have broad and sometimes nuanced uses in adult language, making the meanings of these terms notoriously resistant to formal definition (Herskovits, 1986; Garrod et al., 1999, *i.a.*) and their trajectory of acquisition poorly understood.

The Current Study

We selected a limited set of geometric and functional properties (hereafter 'features') of spatial relations believed to be relevant to language-independent notions of containment and support and the spatial meanings of the English terms *in* and *on*. Using a rating task, we situated a sample of 128 natural scenes within this feature space. The scenes featured object configurations that reflected a wide range of containment and support relations that varied parametrically across the geometric features and in an uncontrolled way across a functional feature. We used this feature space to test two hypotheses about the relationship between feature variation across our items and speakers' use of *in* and *on* to describe those items:

Hypothesis 1: Variation in gradable geometric and functional features is systematically related to differences in adults' probabilistic use of *in* and *on.*

Hypothesis 2: Children's probabilistic use of *in* and *on* will show early sensitivity to geometric features, evidenced pre-linguistically, but not to functional features.

Three key elements differentiate our proposal from previous accounts of *in* and *on*: 1) We developed a set of language-independent geometric features, motivated by pre-linguistic and cross-linguistic evidence, and a functional feature based on psycholinguistic findings, discussed below; 2) We selected a large and highly diverse sample of containment and support relations; and 3) We focused on mappings between our continuous feature space and probabilistic language use.

Why (not) geometry?

Geometric notions (see Landau & Jackendoff, 1993; Talmy, 1985) have served as a cornerstone for theories of the meanings of individual spatial prepositions, and the earliest accounts of prepositions are based solely on geometry. As one example, Bennett (1975) defines the preposition *in* using the highly geometric notion of an interior and a general location function, so that the expression *A is in B* for objects A and B in (1) simply means that A is located at the interior of B. Similarly, Bennett defines the preposition *on* via the location function applied to an object's surface, in (2) (Bennett, 1975, p. 71).

(1) A is *in* B: A [locative [interior of B]]
(2) A is *on* B: A [locative [surface of B]]

These strictly geometric definitions work for stereotypical cases, such as the scenes in Figure 1-A. However, relying solely on notions like interior and surface as constraints on the meanings of *in* and *on* fail to account for many of the regular (i.e., non-idiomatic) uses of the terms, such as for Figure 1-B (Herskovits, 1986; Feist, 2000, Garrod et al., 1999, Coventry et al., 1994, *i.a.*) and predicts unattested uses, such as Figure 1-C.

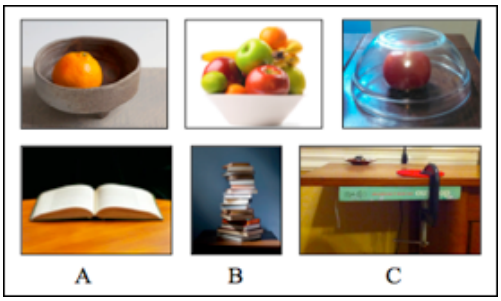


Figure 1. Purely geometric accounts predict that "The orange is *in* the bowl" and "The book is *on* the desk" felicitously describe the respective top and bottom cases for A, but not B (orange/books on top of other fruit/books), and incorrectly predicts that the expressions describe the cases in C (orange/book under bowl/desk).

These are just a few of many counterexamples that have been levied against purely geometric accounts of *in* and *on*. However, we propose that there are a number of connected reasons why an account like Bennett's might fail to capture the full range of meanings and typical uses of *in* and *on*. First, these accounts may simply choose the wrong type of geometric information by building a theory of meaning based on narrow, conventional cases of containment and support (*oranges in bowls* and *books on tables*) while ignoring less central spatial cases for which *in* and *on* can be used, such as *arrows in targets* or *clothes on a line*.

Second, previous geometric accounts may fail as a consequence of the mapping between geometric information and spatial term meaning and *use*. Many proposals, including that of Bennett (1975), state or implicitly assume a deterministic all-or-none relationship between the presence of a feature, geometric or otherwise, and the use of a spatial term: if the feature applies to the relation, then the spatial term can be used to describe it, if the feature doesn't apply, then the term cannot be used. On these accounts, features are assumed to be binary[1] — either applying perfectly to a particular object configuration or not at all — and spatial language use is assumed to be uniform — all configurations to which, for example, the term *in* applies are "equally good" instances of the features specified in the meaning of *in*.

In this paper, we argue against such a deterministic mapping and propose a probabilistic relationship between conceptual features and speaker's use of spatial language (specifically, *in* and *on*). We suggest that the geometric features relevant to containment and support relations are not binary but are instead *gradable* so that different configurations can instantiate features to varying degrees. Following this, we observe that some object configurations are intuitively "better" and more frequent instances of *in* or *on* than others (e.g., example, *on* is used more frequently for *books on a table* than *fish on a line*), and propose a probabilistic mapping between spatial relations and terms.

Support for this proposal comes from previous work by Johannes and colleagues (Johannes, Wilson, Landau, 2013, under review; Johannes et al., 2015) who have uncovered distinctions among hypothesized sub-types of containment and support relations on the basis of differences in speakers' rates of use of individual expressions. For example, across multiple languages, speakers' spatial term use reliably distinguishes cases of loose-fitting containment (e.g., *apples in bowls*) from interlocking cases (e.g., *puzzle pieces in a puzzle*), and cases of support from below (e.g., *books on a table*) from support by hanging (e.g., *fish on a line*).

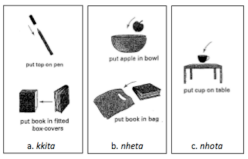
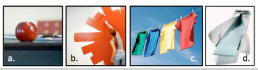
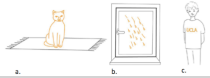
Geometric features of containment and support

We consulted a history of work on pre-linguistic and cross-linguistic spatial categorization to identify pairs of candidate features for containment and support relations. These features are likely to be language-independent in the sense of being salient pre-linguistically, suggesting that these features should not be specific to the spatial meanings of English alone (Gentner & Bowerman, 2009, Hespos & Spelke, 2004), and should be attested in the lexicons of culturally and typologically diverse languages.

Results from studies of pre-linguistic and cross-linguistic spatial concepts converge on a small set of geometric features — two for containment and two for support relations — around which, we hypothesize, the spatial meanings of *in* and *on* can be organized. These four geometric features are outlined in Table 1, along with supporting pre-linguistic and cross-linguistic evidence.

[1] On some accounts, features are not binary by design, but feature values are subject to thresholding to make binary distinctions.
[2] Children completed the rating task on an iPad alongside an

Table 1. Geometric features for containment and support items are listed with the pre-linguistic and cross-linguistic evidence used to select each feature for use in the study.

Feature	Feature Description	Pre-linguistic Evidence	Cross-linguistic Evidence
Containment Enclosure	*The extent to which one (containing) object encloses or surrounds another (contained) object*	When reasoning about containment, 3.5 month-olds are sensitive to both the solidity of a container as well as whether it has an open top (Hespos & Baillargeon, 2001).	Speakers of Jaminjung (Australia) use enclosure-sensitive coverbs *walthub* and *walyag* to encode the configurations below (Levinson & Wilkins, 2006). 
Containment Volume Match	*The amount of empty space between containing and contained objects (alternatively, the tightness of fit between a containing and contained object)*	At 5 months, infants expect objects to move separately when there is a lot of empty space between them and to move together when there is little empty space (Hespos & Spelke, 2004).	Korean speakers use contrastive verbs to encode fit: *A. kkita* for interlocking ('tight-fit') cases; B. *nheta*, used for 'loose-fit' containment; C. *nhota*, used objects put on horizontal surfaces. (Bowerman & Choi, 2001). 
Support Vertical Position	*The vertical position of one (supported) object relative to the other (supporting) object*	At 5.5 months, infants expect an object to remain supported after being placed on the top surface of an object but not after being placed against the side surface (Hespos & Baillargeon, 2008).	Dutch speakers use three prepositions to distinguish cases of support: *op* is reserved for support from below/ permanent contact; *aan* is used for hanging and attachment; *om* is used for encirclement (Gentner & Bowerman, 2007). 
Support Surface Contact	*The proportion of one (supported) object's surface that is in contact with the surface of the other (supporting) object*	6.5 month-olds reach more for an object placed with 100% of its surface in contact with a supporting object, compared to events with 15% surface contact, reasoning that objects in the latter case were permanently attached (Hespos & Baillargeon, 2008).	Speakers of *Yêlí Dnye* (Papua New Guinea), use the term *yedê* (roughly glossed as *on a surface*) to express surface contact independent of the (vertical) position of the figure and ground objects (Levinson & Wilkins, 2006). 

Functional feature: Location control for *in* and *on*

A growing number of proposals based on psycholinguistic data have considered functional information to be at the core of prepositional meanings and usage, including abstract means of *in* and *on* (Jamrozik & Gentner, 2015). Garrod and colleagues (Garrod, Ferrier, & Campbell, 1999) propose hybrid accounts of the meanings of *in* and *on* that combine geometric and functional properties. Specifically, speakers' use of *in* and *on* reflect different kinds of *locational control* between figure and ground objects, in combination with geometric properties. Successful use of *in* reflects regional enclosure with locational control, so that one object (the figure) is likely to be *in* another object (the ground) if the ground encloses the figure, and the ground (and not some other object) controls the location of the figure. Locational control differentiates cases like Figure 2-a and 2-b (taken from Garrod et al., 1999), where the pear (shaded grey object) is judged to by *in* the bowl in case (a) but not (b), despite having the same objective degree of enclosure in both cases.

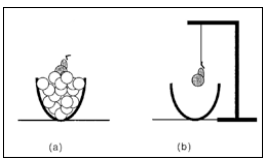


Figure 2. The pear in (a) is judged to be *in the bowl*, while the pear in (b) is not, despite the same degree of enclosure in both cases. (Figure adapted from Garrod et al., 1999)

In experiments testing speakers' confidence about *in/on* descriptions across relations that varied in the geometric properties and locational control, Garrod et al. (1999) found that for cases where regional enclosure and regional contact were high (i.e., prototypical *in* and prototypical *on* cases), judgments of locational control did not predict or influence speakers' confidence about *in* and *on*, respectively. Locational control is predicted to apply equally across cases and is only "blocked" (or reduced in importance) when geometric properties are very salient, suggesting a complex interplay between the two types of information. We examine the interplay of geometry and function in the following experiments by adding a gradable functional feature: *locational control*, operationalized as the likelihood that the figure object of a configuration will move together with the ground object.

Experiment

We developed two large item sets of internet-sourced images of objects in containment and support relations, respectively. We used a simple feature-rating task to situate containment and support items in their own 3-dimensional feature space (two geometric features and one functional feature) and then elicited simple spatial descriptions from a new group of adults to examine relationships between containment and support feature variation and rate of *in* and *on* use, respectively. Adults' patterns of use of *in* and *on* are well-captured by combinations of gradable geometric and functional features.

Methods

Design. One group of adult participants and one group of child participants completed a feature-rating task for each sets of items (containment or support). A separate group of participants from each age group completed a spatial description task with the same items.

Participants One hundred adults (mean age = 19.2 years) participated in the experiment through a self-paced online interface in return for course credit. Fifty participants provided feature ratings for a set of containment (N=25, 14 males) or support items (N=25, 12 males). The other 50 participants provided descriptions for the same containment (N=25, 13 males) or support items (N=25, 11 males). All participants were native English speakers. Twenty 6-year-olds (mean age = 6;5, 11 males) participated in a modified version of the rating task[2], and 24 6-year-olds (mean age = 6;6, 14 males) provided spatial descriptions for containment and support items.

Materials Stimuli consisted of 64 containment items (Figure 3) and 64 support items (Figure 4). The items were chosen, by hypothesis, to parametrically vary along two geometric dimensions: Enclosure and Volume Match for containment items, and Vertical Position and Surface Contact for support items. Locational Control was not manipulated in item selection but was hypothesized to vary randomly across items.

Procedure. Participants in the feature-rating task were familiarized with the set of geometric and functional features then shown either containment or support items, one at a time in random order. For each item, participants were instructed to consider the two salient labeled objects (e.g. "Object A: a sandwich", "Object B: a plate") and provide ratings on a 4-point scale for each of the features. Features, rating prompts, and the response scale endpoints provided to subjects are given below in Table 2.



Figure 3. The 64 containment items organized hypothesized Enclosure and Volume Match feature values.

[2] Children completed the rating task on an iPad alongside an experimenter, who labeled each object and read the prompts aloud.



Figure 4. The 64 support items organized by hypothesized Vertical Position and Surface Contact feature values.

Table 2. Prompts and scale endpoints for the adult and child versions of the feature-rating task.

Containment feature rating task: Prompts and scale endpoints			
	Enclosure	Volume Match	Locational Control
Adult prompts	How much of object A is enclosed by object B?	How much empty space is present between object A and object B?	If object B is moved, how likely is it that object A will move with it?
Adult endpoints	{All/hardly any} of A is enclosed by B	There is {a lot/ hardly any} empty space between A and B	A is {very likely/ unlikely} to move where B moves
Child prompts	Is [object B]...	Is there...	If I move [object B] do you think [object A] will...
Child endpoints	{All around/ not really around} [object A]	{Lots of/ no} empty space	{Definitely/ probably won't} move with it
Support feature rating task: Prompts and scale endpoints			
	Vertical position	Surface Contact	Locational Control
Adult prompts	How much of object A is higher than object B?	How much of object A's surface is in contact with object B?	(same as above)
Adult endpoints	{All/ none} of A is higher than B	{All/ hardly any} of A is in contact with B	
Child prompts	How much of [object A] is higher than [object B]?	How much of [object A] is touching [object B]?	
Child endpoints	{All, no part} is higher	{All, almost no part} of A is touching B	

Adult participants in the spatial description task typed in a brief description into the space provided in the online interface by answering the question "Where is [Object A] in relation to [Object B]?" for the labeled objects in the image. Child participants gave verbal descriptions that were later transcribed.

Results

Feature rating task Adults and children were both reliable in their geometric feature ratings (Adults: r=0.92, p<.01; Children: r=0.89, p<.01) and their mean ratings supported our hypothesized item variation across the geometric features in both containment and support item sets. Adults, but not children, were also reliable in their ratings of Locational Control. Adults' ratings of two geometric features, Enclosure for containment and Vertical Position for support, were correlated with their ratings of Locational Control (Enclosure: r=0.485, p<.01; Vertical Position: r=0.548, p<.01). Items rated highest in Enclosure or Vertical Position were also rated highest in Locational Control, suggesting that Enclosure and Vertical Position may be geometry-based mechanisms by which one object can exert control over another object.

Spatial description task Each participant provided a single description for each scene in the form: "Object A [spatial expression] Object B", where the spatial expression included one or more verbs, a preposition, and an optional modifier. Descriptions were coded for the presence of the copular verb BE combined with the prepositions *in* (for containment items), or *on* (for support items) — for example "The sandwich *is in* the bag" and "The pillow *is on* the stool". This coding was designed to exclude descriptions with lexical verbs (e.g., "The sandwich *is sealed in* the bag", "The pillow *is sitting on* the stool"), which encode additional spatial, configurational, or mechanical information (see Johannes, Wilson, & Landau, under review), as well as descriptions with prepositions other than *in* or *on*, both of which were the focus of other analyses, not reported here (54% of adult descriptions and 36% of child descriptions). This coding was used to assign a binary value to each description: a description was coded as 1 if it included *in* (for containment) or *on* (for support), and was coded as 0 otherwise.

Adults and children showed similar patterns of usage of *in* and *on* across containment and support items and adult and child uses of *in* and *on* were reliably correlated (Pearson's R_{in}= 0.67, *p<.01*; R_{on}= 0.77, *p<.01)*. Some items were described exclusively with *in* or *on*, while others were rarely described this way.

We computed four mixed-effects logistic regression models to examine the relationship between mean geometric and functional feature ratings and the distribution of adult and child speakers' use of *in* and *on* across containment and support scenes. Models included item and subject as random effects and each of the features was coded as a fixed-effect predictor[3].

Adults' use of both *in* and *on* was related to a combination of geometric and functional features (as rated by the separate group of adults). Table 3 gives examples of the containment and support items that were predicted by the models to have high and low rates of *in* and *on* use. For containment, items rated high in Enclosure and in Locational Control were most likely to be described with *in* (see Table 3 for the standardized model coefficients, β, for each feature predictor). This pattern suggests that, for containment, the geometric property of Enclosure is a means by which one object can contain another object and by which it can control the location or movement of another object. For support, items rated highly on Vertical Position and Surface Contact features, but *low* in Locational Control were most likely to be described using *on* (standardized model coefficients in Table 3). This seemingly non-intuitive pattern suggests that Locational Control is a salient means of support only when the Vertical Position and Surface Contact between objects are not. For both containment and support, the model-fitted probabilities of *in* and *on* use for each item were highly correlated with the adults' observed frequency and distribution of use.

Table 3. Results of feature-language models: items predicted to elicit high and low *in/on* use for adults and children; model standardized coefficents (β) for each predictor; and correlations between model-fitted probability of *in/on* use and observed relative frequency of *in/on* use.

Containment	**Adult Results**	**Child Results**
Item with high predicted *in* use	oranges in a bag 	hamster in a ball 
Item with low predicted *in* use	sushi in chopsticks 	(top) book in a box 
Enclosure β	**0.47, *p<.01***	**0.18, *p<.01***
Volume Match β	-0.05, *(ns)*	0.07, *(ns)*
Location Control β	**0.48, *p<.01***	-0.08 *(ns)*
Correlations: predicted and attested use	*R*= 0.67, *p<.05*	*R*= 0.61, *p<.01*
Support	**Adult Results**	**Child Results**
Item with high predicted *on* use	 sandwich on a plate	 plate on a stick
Item with low predicted *on* use	ornament on a branch 	
Vertical Position β	**0.33, *p<.01***	**0.26, *p<.01***
Surface Contact β	**0.52, *p<.01***	0.11, *(ns)*
Location Control β	**-0.29, *p<.05***	-0.04 *(ns)*
Correlations: predicted and attested use	*R*= 0.81, *p<.01*	*R*= 0.69, *p<.01*

In contrast to adults, children's use of *in* and *on* was related to single geometric features of containment (Enclosure) and support (Vertical Position) and not by children's functional

[3] Covariance between pairs of features was controlled by using the residuals of preliminary linear models predicting one feature from another.

ratings, which were noisy and unreliable, or by adults' functional ratings. Table 3 shows examples of items with predicted high and low *in* and *on* use and the standardized model coefficients for each feature predictor. We hypothesized that children's reliance on geometric over functional information stems from their robust geometric knowledge, acquired before the onset of language, along with still-developing knowledge of control relationships, which may require extremely rich knowledge about object functions and force-dynamic relationships between objects.

Discussion

In the current study, we developed a large structured sample of containment and support items to test the relationship between geometric and functional feature variation and spatial language use at two points in development. Our feature rating results confirm that a wide range of containment and support relations are organized around several geometric features, while also varying on the functional feature of location control.

Confirming our first hypothesis, geometric and functional feature variation in our sample of containment and support items was reliably related to adults' probability of using *in* and *on* to describe the items. Specifically, geometric and functional features combined to predict differences in adults' use of *in* and *on* across items.

We further examined whether the relationships between feature variation and language shown for adults also held for children's spatial language, specifically evaluating the hypothesis that children's early use of *in* and *on* is sensitive to pre-linguistically available geometric information but is not sensitive to functional feature variation among containment and support relations. Our results demonstrate that children's probabilistic use of *in* and *on* shows early adult-like sensitivity to geometric variation across relations but not to functional feature variation, suggesting that robust pre-linguistic geometric knowledge, but not functional knowledge, constrains children's early uses of *in* and *on*.

We designed our feature sets with features that were *relational*: they encode information about both the figure and the ground objects in each item. We tested our features against an alternative set of plausible, *non-relational* features: the degree of curvature of either the figure or ground object, for containment items, and horizontal orientation of either the figure or ground object, for support items. Variation in these non-relational features failed to reliably predict the use of *in* and *on* across the items in our study and ratings for these features did not correlate with ratings of the relational features in our study. Arguably, speakers' use of spatial terms like *in* and *on* is most sensitive to relational information – and not merely to properties of individual objects – in object configurations.

Conclusion

We have demonstrated that variation in combinations of geometric and functional features is related to variation in speakers' expression use across a large and diverse sample of containment and support relations. Our proposal that gradable features are related to speakers' probabilistic use of *in* and *on* reconciles the fact that, while we can use the same spatial term for a wide range of cases, some cases are better instances of *in* and *on* than others. Put another way, while relations with (any degree of) certain geometric or functional properties license a given spatial term, having those properties to a greater or lesser degree is related to how "good" a relation is as the extension of a term.

References

Bennett, D.C. (1975). *Spatial and Temporal Uses of English Prepositions*. London: Longman.

Feist, M. I., & Gentner, D. (1998). On plates, bowls, and dishes: Factors in the use of English IN and ON. *Proceedings of the 20th Annual Meeting of the Cognitive Science Society,* 345-349.

Garrod, S., Ferrier, G., & Campbell, S. (1999). In and on: investigating the functional geometry of spatial prepositions. *Cognition, 72,* 167-189.

Herskovits, A. (1986). *Language and spatial cognition: an interdisciplinary study of the prepositions in English.* Cambridge, England: Cambridge University Press.

Hespos, S.J., & Baillargeon, R. (2001). Reasoning about containment events in very young infants. *Cognition,* 78, 207-245.

Hespos, S.J., & Baillargeon, R. (2008). Young infants' action reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition, 107(1),* 304-316.

Hespos, S.J. & Spelke, E.S. (2004). Conceptual precursors to spatial language. Nature, 430, 453-456.

Johannes, K., Wang, J., Papafragou, A., & Landau, B. (2015). Systematicity and variation in the distribution of spatial expressions in three distinct languages. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society,* 997-1002.

Jamrozik, A., & Gentner, D. (2015). Well-hidden regularities: Abstract Uses of *in* and *on* retain an aspect of their spatial meaning. *Cognitive Science*, *39,* 1881-1911.

Johannes, K., Wilson, C., & Landau, B. (under review). The importance of lexical verbs in the acquisition of spatial language: The case of *in* and *on.*

Johnston, J.R., & Slobin, D.I. (1979). The development of locative expressions in English, Italian, Serbo-Croatian and Turkish. *Journal of Child Language, 6(3),* 529-545.

Landau, B. & Jackendoff, R. (1993). 'What' and 'where' in spatial language and spatial cognition. *Behavioural and Brain Sciences, 16(2),* 217-265.

Levinson, S. C., & Wilkins, D. P. (Eds.). (2006). *Grammars of space: Explorations in cognitive diversity.* Cambridge: Cambridge University Press.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen (ed.), *Language typology and syntactic description* (pp. 57-149). Cambridge: Cambridge University Press.

Modeling Triage Decision Making

J. Isaiah Harbison (jiharb@umd.edu),
Alan Mishler (amishler@umd.edu),
Thomas Wallsten (tswallst@umd.edu)
Center for Advanced Study of Language, University of Maryland
7005 52nd Avenue, College Park, MD 27642

Abstract

With the ever increasing amount of information available, the ability to prioritize the most relevant items for full processing is increasingly necessary to maintain expertise in a domain. As a result, accurate triage decisions--initial decisions about the relevance of a given article, book, or talk in order to determine whether to pursue that information further--are very important. In the present paper, we present a model of triage decision making that includes both an information search component to determine reading strategy and a decision making component to make the final decision. We apply the model to human relevance ratings as well as binary decisions of relevance for a set of emails.

Keywords: information search; information foraging; decision making; triage decision making

Introduction

When looking through Google Scholar results or the latest issue of your favorite journal, how do you decide if an article is relevant to your research? This relevance decision or triage decision requires at least two components: an information search strategy to guide search through the abstract or article, and a decision making process to determine when sufficient evidence has been collected to make the decision of relevance.

In the present paper, we introduce a model of triage decision making that combines an information search process with a decision making mechanism. We then test the model against relevance judgment data and binary decisions.

Searching for a Topic

Duggan and Payne (2009) found evidence that when skimming, people read until the rate of information gain drops below a threshold. At that point they either move on to a new section (paragraph, chapter, etc.) or they stop skimming the document. This behavior matches the behavior prescribed by the marginal value theorem (MVT). According to the MVT in order to optimize the number of items found, search in a given location or patch should continue until the rate of return drops beneath the expected rate of return in the general search environment. For example, in an online literature search, the rate of return for the general environment would be calculated by combining the average expected rate of information gain from other articles in the patch (e.g., each article would be considered a patch in the current application), including the expected time taken to find and open those articles. This rate of information gain establishes the threshold that the rate of gain from the current article is compared against.

The calculation of expected rate of return from general search relative to the rate of return of the current document would be very complicated, making the MVT implausible as a descriptive stopping or switching rule. However, there is another relatively simple stopping rule, the incremental rule that is able to implement MVT. The incremental stopping rule compares the total amount of time spent in search (or total number of search attempts performed so far) to a stopping threshold in order to determine whether to terminate the search. The threshold is incremented each time a target item is found. For example, suppose the initial stopping threshold is two minutes and the increment is five seconds. If a search proceeds for two minutes without returning any target items, then the search will be terminated. Each time a target item is found, however, five seconds are added to the time allowed to pass before terminated. After one item is found, the stopping threshold becomes two minutes and five seconds (i.e., the total time allowed for search from beginning to end would be two minutes and five seconds if only one item was found). If ten target items were found, the stopping threshold would be two minutes and fifty seconds. This strategy makes stopping decisions a function of the rate of return from a patch, ensuring that search continues longer in rich patches and terminates earlier in poor patches.

Evidence for the use of the incremental stopping rule has been found both for external search (Hutchinson, Wilke, & Todd, 2008) and internal search (e.g., memory search; Harbison, Dougherty, Davelaar, & Fayyad, 2009). Our model applies the incremental stopping rule to triage decision made about text. Information search proceeds by reading through a document until either the document has been read, the incremental stopping rule indicates search should be terminated, or the relevance of the document can be decided upon. The mechanism for latter case will be described next.

Identifying a Topic

Triage decision making is a matter of both information search and decision making. Search can be terminated as a function of the rate of return of information, but it can also be terminated based on a decision of relevance being reached if sufficient evidence has been accumulated. In the

former case, once search is terminated a decision must still be reached regarding the item's relevance. Given these two possibilities, how is the decision ultimately made about whether a text is relevant to a given topic?

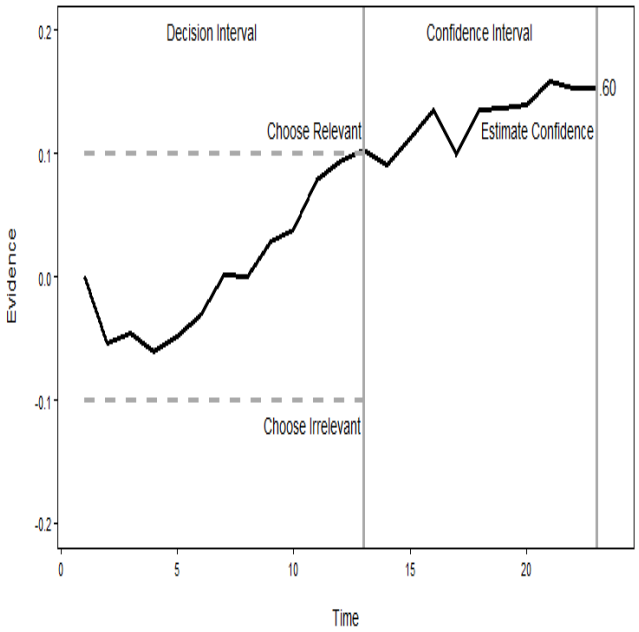


Figure 1. Schematic of the 2DSD decision making process, with a relevance judgment (first vertical line) and a confidence judgment (second vertical line).

We utilize the two-stage dynamic signal detection (2DSD) model of confidence and decision making (Pleskac & Busemeyer, 2010) to model the relevancy decision process. The 2DSD model conceptualizes the decision maker as moving between two decision thresholds. The model uses a random walk/diffusion process to account for movement between these two options, as shown in Figure 1.

In the case of triage decision making, the two decision thresholds are "relevant and "not relevant." During the decision process, the decision maker skims the material gathering evidence for and against an article's relevance. Each new piece of evidence moves the evidence state between two thresholds from its starting position. Evidence for relevance moves the evidence state toward the "Choose Relevant" threshold, and evidence against its relevance moves the evidence state towards the "Choose Irrelevant" threshold.

In previous uses of the 2DSD model, evidence was sampled from a random distribution that reflected the knowledge the individual had about the relative evidence for each of the two options. Here evidence is the information for and against each of the options. To the extent that evidence is clearly in favor one option, the probability of picking evidence in support of that option is increased, increasing the probability and speed of the dominant option being chosen (i.e., that option's threshold being passed). Once a threshold is crossed, a decision is made. If the decision maker is asked to give their confidence in their decision, this judgment is made after additional processing and additional consideration of evidence, as depicted on the right portion of Figure 1.

The difference between the present and previous applications of the 2DSD model (other than applying its application to triage decisions) is that the evidence is not sampled from a random distribution. Instead, the evidence is a function of the material being triaged. For the present modeling application, the materials are emails and the evidence is determined by the words being read. The evidential value of each word in our model consists of the point-wise mutual information (PMI, detailed below) between the word and the topic under consideration. The state of belief is determined by the running sum of the evidence values of the words read.

Triage Decision Making Model

There are three components to the triage decision making model: a computational representation of topics, a search strategy, and a model of decision making. For the first component, we represented topics as probability distributions over words. This method is common both to the natural language processing of text for guiding internet search (Blei, 2012; Xu, Yang, & Li, 2009) and to psychological models attempting to reflect human understanding of topics (Steyvers & Griffiths, 2007). The core idea is that the topic under discussion influences the chance that any given word will occur. For example, words such as "computational", "model", and "theory" are more likely to occur in conversations about cognitive science than in a conversation about Philadelphia. Therefore, it would be reasonable for a decision maker who encounters these words while reading to judge the likelihood that the text is about cognitive science more highly than the likelihood that the text is about Philadelphia.

The second component is the skimming or search strategy. We tested the following three strategies:

1. The Complete Strategy. The searcher moves from the beginning to the end of the document, making a decision about the relevance of the document only after reading the entire text.
2. The Skim Beginning Strategy. The searcher moves from beginning to end, but search is terminated once the threshold of evidence is crossed or after the stopping rule indicates that search should be terminated without a decision being made.
3. The Skim Paragraph Strategy. The searcher makes decisions about staying within or leaving individual paragraphs of the document instead of the entire document. In this strategy, the decision is made either after the threshold is crossed or after the decision maker has searched and left each of the individual paragraphs.

The third component of the model is the decision process. We used a process similar to the random walk process from the 2DSD model, considering each word as evidence for or

against a conclusion. As noted above, however, the decision process in our model is not truly a random walk. Since the evidence is represented as the PMI of each word to the topic, evidence is accrued through skimming the document through one of the search strategies. a systematic walk through the text instead of a random sampling of possible evidence.

Model Details

Information Ecology

To evaluate the model, we tested it against human judgments of topic relevance. Specifically, we applied the model to a set of emails drawn from the 2001 Topic Annotated Enron Email Data Set and compared the model's judgments to those of participants from a new experiment described below. This dataset consists of emails that have been hand-labeled with topics by a trained annotator, with one topic label per email (Berry & Browne, 2007). The topics used in our experiment were: California Analysis and Daily Business. California Analysis emails included executive summaries and analyses about the company's affairs in California. The Daily Business emails are about buying and trading shares on the stock market, setting up meetings, and confirming meetings, as well as general announcements from human resources.

Topic Representation

We create the representations of each topic by calculating the frequency distributions of the words within each set of emails for that topic and converting these into probability distributions. The evidence in the model consists of the pointwise mutual information (PMI) between the topics and the term, calculated as follows:

$$PMI(term, topic) = log \frac{p(term,topic)}{p(term)p(topic))}.$$

Participants in the experiment were asked to read the emails one at a time and indicate whether each email was relevant to the target topic (California Analysis). Half the emails came from the target topic, and the other half came from the distractor topic (Daily Business). The distribution of PMIs for words in these two topics are shown in Figure 2. These distributions are approximately normal, with the target mean just above zero (.427) and a distractor mean just below zero (-.025).

Examining the two distributions of PMIs in more detail, we calculated the mean PMI of words as a function of the sentence and paragraph in which they occurred. Many of the emails had at least five paragraphs of five sentences each, so we looked at the mean PMI for the first through fifth sentence of the first through fifth paragraph. These means are shown in Figure 3. Across paragraphs there is little variation in mean sentence PMIs. Likewise, the variation in mean PMI of sentence within paragraph does not show a clear pattern.

Given the overlap in target and distractor email PMI distributions, it is not clear how well the stimuli can be used to identify topic emails. To get an impression of how the accumulation of evidence would work for emails from both topics, we plotted the cumulative evidence (PMI) of both types of emails. The running sum of the PMI between the words in the email and the target topic is shown in Figure 4. As the figure shows, overall the PMI for the distractor emails is lower than for the target emails, suggesting that even though the average PMI for target emails and the target topic is just above 0, the summation of the evidence provided by the emails can be used to distinguish between target and distractor emails.

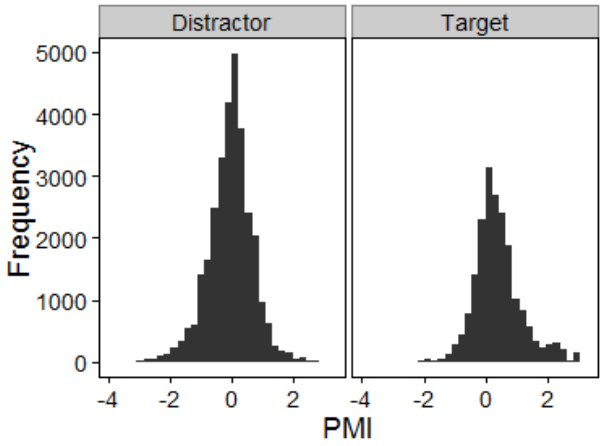


Figure 2. Distribution of pointwise mututal information values (PMI) for words within each topic.

Skimming Process

We tested three ways of moving through the text to make a triage decision: Complete, Skim Beginning, and Skim Paragraph. All three methods start with the first sentence of the first paragraph of the email. The Complete (or full read) Strategy continues reading to the end of the document, accumulating all of the evidence provided by the words along the way but only making a decision at the end. This strategy is unique in that it does not allow for early search termination for any reason.

The Skim Beginning Strategy continues reading until the evidence collected reaches a stopping threshold (θ_S) or when the incremental failure rule prescribes and end to the search (i.e., when a threshold, θ_F, number of words fail to impact the current state of the evidence). Likewise, for the Skim Paragraph Strategy, search is terminated after a decision threshold is reached. The difference between the two skimming strategies is that the Skim Paragraph Strategy will not terminate search if the information gain in the current paragraph decreases below threshold but instead moves to the beginning of the next paragraph in the email. The goal of movement between paragraphs is to allow for more effective search, with the assumption that the most informative words are likely to be in the beginning of paragraphs. Note that the Skim Paragraph Strategy will terminate search once a decision of relevance is reached or once it leaves the final paragraph.

Decision Process

Decisions are made in the model by a running sum of the evidence for and against relevance. The original application of the 2DSD followed a random walk/diffusion process (Pleskac & Busemeyer, 2010), shown in Figure 1, in which evidence is accumulated over time that moves the model closer or further away from the thresholds for determining between options (θ_D and $-\theta_D$, respectively). The random walk is instantiated by randomly sampling from a distribution of evidence values that move the current state of evidence either up or down. The distributions are different for target items than for distractor items. Drawing from the target distribution, on average, provides movement towards the "relevant" threshold and drawing form the distractor's distribution, on average, provides movement towards the "irrelevant" threshold. Once a threshold is crossed, the decision is made about the item's relevance. If participants are required to give a confidence or probability estimate, they do so after some amount of additional processing time in which it is assumed that more evidence is considered (either through additional reading or by reflecting on what has been processed so far).

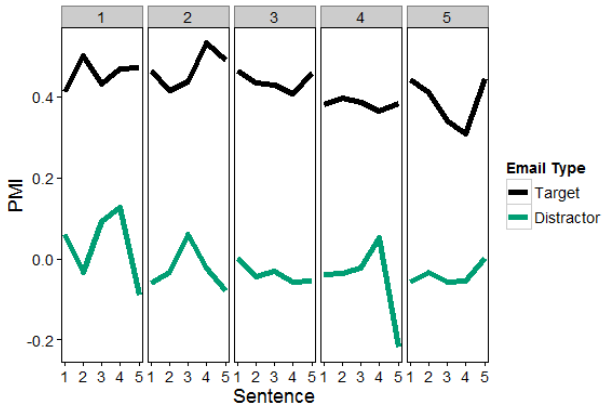


Figure 3. Mean pointwise mututal information (PMI) of each word as a function of position in the document, colored based on the topic.

This decision model has previously been applied to this data set (Harbison et al., 2015). However, the present application differs in that instead of using the random walk process by drawing randomly from target and distractor distributions, we used the skimming strategies specified above to systematically move through the documents and use the evidence collected as the basis for making the decision. That is, instead of a random walk through possible evidence values, we modeled the decision process as a systematic walk through the evidence that resulted from the skimming process. Therefore, the skimming and decision process are guided by the information ecology of the environment.

Participant Data

The data used here is drawn from a larger experiment on triage decision making that examined response type for two different topics (Harbison et al., 2015). Here we used the results of only one of the two topic conditions, California Analysis, for all response types. The response types include a continuous rating condition (Rating) in which participants responded to an item's relevancy along a sliding scale, and two binary conditions: one instructed to minimize false alarms (FA) and one instructed to maximize the number of targets found (Hit). Further details are below.

The results included 120 trials from twenty-four students in the Rating condition, twenty-six in the Hit condition, and twenty-seven in the FA condition. The trials were a mix of 60 emails about the target topic and 60 emails that were not. Each email was presented one at a time in a random order unique to each participant. Participants were instructed to imagine that they were an investigator looking for information about the Enron scandal, and they were given an excerpt from a newspaper article to read about the scandal (Berenson, 2002). A brief description of the target topic was visible on screen throughout the experiment. Responses were required to respond before the subsequent email was displayed.

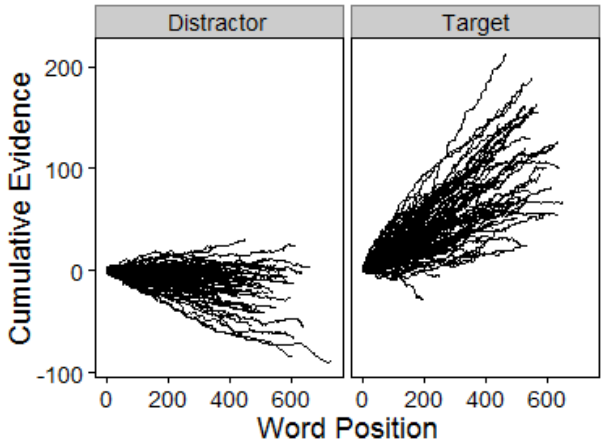


Figure 4. Running sum of the PMI for each of the distractor and target emails.

The email response portion of the study was limited to 70 minutes, with enforced one-minute breaks at 17.5 minutes, 35 minutes, and 52.5 minutes. This gave participants only 35 seconds on average to respond to each email, which forced them to spend some amount of time skimming rather than thoroughly reading every document. Participants were informed of the time constraints, which were designed to force them to make rapid judgments, as in a real-life triage decision making process.

Participants were paid a base rate of $10 in addition to a bonus of $0-15 based on task performance. The calculation of the bonus varied by condition. Prior to the main experiment, participants saw two practice trials, one relevant and one irrelevant. They received feedback on their responses to these trials and were also shown (1) what they would have earned on those trials had they been real, and (2) what other possible responses on those trials would have

earned. During the main task, participants were not given feedback on their performance.

Participants in both of the binary conditions made their responses by clicking on either the "Not Relevant" or the "Relevant" button. In both conditions, participants made no additional money for incorrect classifications (i.e., responses of "Relevant" to a distractor email or "Irrelevant" to a target email). However, the amount earned per correct response varied as a function of condition. In the Hit condition, participants were paid \$0.14 for correctly identifying target email and \$0.11 for correctly rejecting an irrelevant item. Participants in the FA condition received the inverse reward, \$0.11 for hits and \$0.14 for correct rejections.

In the Rating condition, participants gave ratings using a slider, with endpoints labeled "Completely confident it is not relevant" and "Completely confident it is relevant." The slider response was translated into a rating between 0 and 100, with 0 corresponding to the left endpoint and 100 to the right endpoint. Participants earned between \$0 and \$0.125 per email rated. Earnings were determined by the spherical scoring formula (Merkle & Steyvers, 2013):

$$earning = \frac{0.125 * r_i}{\sqrt{r_i^2 + (1 - r_i)^2)}},$$

where $r_i = d_i * f_i + (100 - d_i)(100 - f_i)$, and d_i is the relevance of the current trial (100 if relevant, 0 if irrelevant) and f_i is the participant's reported confidence that the trial is relevant.

This formula awards \$0.125 for responding "Completely confident it is not relevant" to distractor emails and for responding "Completely confident it is relevant" to target emails. It rewards \$0 for the inverse responses, and it awards amounts between \$0 and \$0.125 for responses in the middle of the scale, with the amount depending on the relevance of the email and the location of the response. Importantly, this formula incentivizes participants to accurately report their confidence that the email is relevant. That is, a participant maximizes their expected reward by responding on the point on the slider that corresponds to their actual subjective belief that the email is relevant (Merkle & Steyvers, 2013).

We are primarily interested in the relationship between the model's predictions and the participant data. However, there is one relationship within the data that is important to note. There was a negative relationship between the magnitude of participant responses and the time to make the response or reaction time (rho = -.288, p<.01). The longer participants took to respond, the closer that response was to the middle of the scale. This relationship between mean response magnitude and mean response time was also found in the two binary conditions (Hit: rho = -.443; FA: rho = -.374, p's<.001).

Simulation

We tested the model on the same emails that participants read and compared the ratings of relevance supplied by participants with the amount of evidence collected by the model. Table 1 shows three correlations for each skimming strategy. The first is the correlation between the response of the model and the reaction time or time taken for the model to make the response. The second and third correlations are between the responses made by the model and by the participant and the time taken for the model and the time taken by the participant to make a response. Note that the simulation results were not from an exhaustive exploration of the parameter space, but instead from a limited search. The decision threshold for both the skimming strategies was 20 (+20 for responding relevant and -20 for responding irrelevant). The skimming parameters were a stopping threshold (δ_S) of 20, the size of the increment was 3, and size of the change necessary to not be considered an information search failure (δ_F) was .5 for the Skim Beginning Strategy and δ_S of 1, increment of 1, and δ_F. of 5 for the Skim Paragraph Strategy.

For all three strategies there was a positive correlation between the model's evidence value (or the model's version of a response) and participant responses. The correlation was slightly smaller for the Skim Paragraph Strategy relative to the other strategies. There was also a positive correlation between participant response time and the response times predicted by the three strategies, with a slightly larger correlation between the data and the Skim Beginning Strategy's RTs. Lastly, we looked at the correlation between response time and response magnitude. We found a negative correlation between response magnitude and response time in participant data. Only the Skim Beginning Strategy predicted this negative correlation. The other two strategies incorrectly predicted a positive relationship between response magnitude and response time.

Table 1: Correlations between data and model.

		Cor. With Data	
	Resp. and RT	Response	RT
Complete	.29*	.58***	.21*
Skim-Beg	**-.46*** **	**.58*** **	**.29** **
Skim-Para	.42**	.49***	.21**

* p<.05, **p<.01, ***p<.001

We applied the most successful strategy, Skim-Beginning, to the two binary conditions, the results are shown in Table 2. The correlation between the mean participant response per email to the model's predictions were both greater than .50 and the correlation of the reaction times was correlated at greater than .30 (p's < .001).

Overall, the Skim Beginning Strategy most clearly approximates participants' actual responses and response times. Although all three strategies showed the same pattern of correlations with participant responses and reaction time. Only the Skim-Beginning Strategy yielded the negative correlation between response magnitude and reaction time that was found in all three conditions of the data. The model, with this strategy, also correlates with participant

responses and reaction times in the two binary response conditions.

Discussion

Triage decision making is increasingly relevant given the growing amount of information in any given field and the limited time and cognitive resources available to process this information. How are these decisions made? In the present paper, we presented new experimental data on the triage process as well as an initial model. This model combines what we have learned about information search termination decisions (the incremental stopping rule), a skimming strategy (skim from the beginning), a method of representing the information value of words (pointwise mutual information), and a model of decision making and confidence judgment (2DSD) to account for human decisions about the relevance of emails to a target topic.

The general performance of the model matches participant performance. It shows a moderate correlation between predicted and observed ratings of relevancy, as well as between the predicted ratings and the average triage decisions. We also found a correlation between the model's predicted reaction times and participant reaction times. Finally, the model, when using the skim-from-the-beginning skimming strategy, demonstrated the same negative correlation between response magnitude and reaction time as found in the participant data.

Table 2. Correlation between the model and the binary conditions

	Hit	FA
Response	.524***	.555***
Reaction Time	.302***	.352***

*** p <.001

One potential problem with the current representation of information distribution, particularly the relationships shown in Figure 3, is that it ignores the potential of redundancy between words. The information value of each word in each sentence was treated independently, without regard to the words that preceded it. It is doubtful that the evidence provided by each word would be treated completely independently. Just as words are more or less common given a specific topic, words are expected to show a similar relationship to each other. Words are likely related to each other such that processing one would provide evidence about the likelihood of another. This type of redundancy would cause diminishing returns from the continued skimming of a document. This could be accounted with the PMI calculation that included not only the relationship between a given word and the topic, but also the words with each other. We plan to explore this possibility in future work.

Acknowledgments

This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

References

Berenson, A. (2002, May 9). Mystery of Enron and California's Power Crisis, New York Times.

Berry, M. W., & Browne, M. (2007). 2001 Annotated (by Topic) Enron Email Data Set Documentation. Linguistics Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/docs/LDC2007T22/

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*, 77-84.

Duggan, G. B., & Payne, S. J. (2009). Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of Experimental Psychology: Applied, 15*, 228-242.

Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition, 111*, 397-402.

Harbison, J. I., Mishler, A., & Wallsten, T. (2015). Triage decision making task: An analyst-relevant outcome measure. College Park, MD: CASL.

Hutchinson, J. M., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75*, 1331-1349.

Merkle, E. C., & Steyvers, M. (2013). Choosing a Strictly Proper Scoring Rule. Decision Analysis, 10(4), 292-304. doi: doi:10.1287/deca.2013.0280.

Pleskac, T. J. & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*, 864-901.

Xu, G., Yang, S. H., & Li, H. (2009). Named entity mining from click-through data using weakly supervised latent dirichlet allocation in *KDD*, 2009.

The mismeasurement of mind:
How neuropsychological testing creates a false picture of cognitive aging

Michael Ramscar (michael.ramscar@uni-tuebingen.de)

Ching Chu Sun (ching-chu.sun@uni-tuebingen.de)

Peter Hendrix (petrus.hendrix@uni-tuebingen.de)

Harald Baayen (harald.baayen@uni-tuebingen.de)

Department of Linguistics, Wilhelmstrasse 19,
72074, Tübingen, Germany

Abstract

Age-related declines in scores on neuropsychological tests are widely believed to reveal that human cognitive capacities decline across the lifespan. In a computational simulation, we show how the behavioral patterns observed in Paired Associate Learning (PAL), a particularly sensitive measure of age-related performance change (Rabbitt & Lowe, 2000), are predicted by the models used to formalize associative learning processes in other areas of behavioral and neuroscientific research. The simulation further predicts that manipulating language exposure will reproduce the experience-related performance differences erroneously attributed to age-related decline in age-matched adults. Consistent with this, older bilinguals outperformed native speakers in a German PAL test, an advantage that increased with age. These analyses and results show that age-related PAL performance changes reflect the predictable effects of learning on the associability of test items, and indicate that failing to control for these effects is distorting our understanding of cognitive and brain development in adulthood.

Keywords: aging, learning, cognitive modeling, bilingualism

The measurement of mind across the lifespan

When adults memorize arbitrary word pairs – e.g., *jury–eagle* – during Paired Associate Learning (PAL) tests, their ability to later recall *eagle* given *jury* declines systematically with age. Together with convergent patterns of change on other neuropsychological tests, this is taken to show that cognitive ability declines across adulthood, providing a functional characterization of the age-related structural changes that occur in healthy brains (Deary et al, 2009; Salthouse, 2011; Singh-Manoux et al., 2012; Lindenberger, 2014).

PAL tests are particularly sensitive to the detrimental effects of age on cognition (Rabbitt & Lowe, 2000), which as the data from two groups of adults completing the PAL subtest of Wechsler's Memory Scale, (desRosiers & Ivison, 1988) plotted in Figure 1 shows, are evident surprisingly early in life (Singh-Manoux et al., 2012): Mean performance of 78% in 20-29 year-olds falls to 70% in 30-39 year-olds ($t(19)=5.286$, $p<0.01$; Ramscar & Port, 2015), the largest by-decade decline observed on this test across the lifespan (desRosiers & Ivison, 1988).

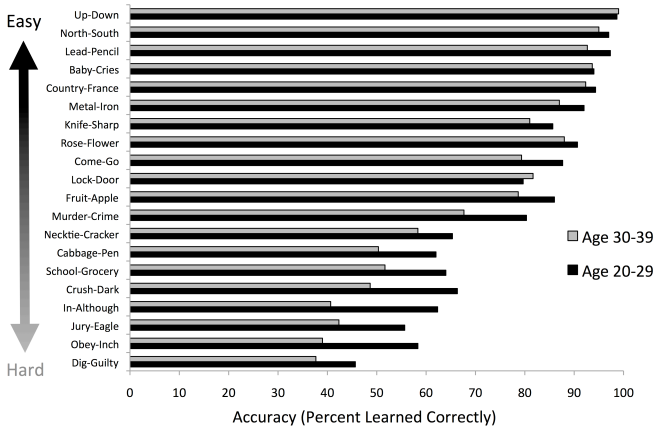


Figure 1: Average by-item performance for 400 adults aged 20-29 and 30-39 (50% females / group) on forms 1 & 2 of the WMS-PAL subtest (desRosiers & Ivison, 1988).

In what follows, we show that these changes are not evidence of cognitive decline, because raw PAL scores cannot be used to compare performance between groups whose experience varies. The reasons why begin with the fact that PAL tests present items (typically word-pairs) at a fixed rate: every word-pair is heard N times prior to testing. This means that to use raw PAL scores to compare performance in different age-groups, one must assume that learning during the test is unaffected by differences in prior experience of PAL items.

An enormous body of research has shown this assumption to be false. Psychometricians have long-known that the empirical properties of word-pairs make some easier to learn than others (desRosiers & Ivison, 1988); And research into associative learning has repeatedly shown that learning to associate a behavioral cue (the word *jury*) with a behavioral outcome (*eagle*) cannot be reliably predicted from the **association rate** of a cue and a response alone (Rescorla & Wagner, 1972; Ramscar, Dye & McCauley, 2013a).

Two other important factors have shown to influence associative learning: Cue **background rates** (Rescorla, 1968; Ramscar, Dye & Klein, 2013b), which, in PAL tests

are the frequencies at which cue words have been encountered *absent* response words, and **blocking** (Kamin, 1969; Arnon & Ramscar, 2012), the predictability of the response given the cue (in context) prior to training. While association rates tend to promote learning, blocking and background rates inhibit it, and critically, the way these factors interact to influence the learning of a specific association is entirely a function of a given learner's experience (Ramscar et al., 2013a).

This last point is critical to understanding the patterns of PAL performance across the lifespan. As can be seen in Figure 1, aging makes "hard" PAL items proportionally harder to learn than "easy" items. This non-linear pattern of change is not predicted by simple declines in plasticity or fluid-intelligence – the changes that theories of cognitive aging posit to explain declining test scores – yet as the following simulation shows, it falls naturally out of the interaction between experience and the three factors shown to influence associative learning that we just described.

Simulation Experiment

The development of word associations in a small lexicon of PAL items was simulated using the Rescorla-Wagner (1972) learning rule (computationally, this rule describes a discrimination learning network, Ramscar et al, 2013a). The lexicon comprised four "easy," or meaningful PAL pairs (*baby-cries, baby-eagle, jury-duty* and *jury-summons*) and two meaningless, or "hard" pairs (*baby-summons* and *jury-eagle*; see Figure 1).

To reflect the fact that the distributional properties of the words that occur in natural languages differ, and to show how the variations in word co-occurrences in a typical English-speaker's experience will affect learning over time, the meaningless items were pre-trained with low association rates (each was presented 10 times). Then to simulate the effect that experience of the more frequent meaningful pairs will have on the learnability of the meaningless items, we presented the model with *jury-duty* 40 times, *baby-cries* and *baby-eagle* 60 times, and *jury-summons* 80 times. The order in which individual item exemplars were presented was determined randomly, subject to the probability of their occurrence in training.

Results

The development of the word associations in the model's lexicon (Figure 2) illustrates how increased experience of a world containing *jury-duty* and *baby-eagle* serves to discriminate against the learnability of *jury-eagle*. Increased experience with the meaningful word-pairs increases the background rate of *jury* in relation to *eagle*, while simultaneously forcing *jury* into competition for associative value with the more frequent cue *baby*. This ultimately results in the model learning a negative association between *jury* and *eagle*; this would have to be unlearned in order for the model to positively associate *jury* with *eagle*.

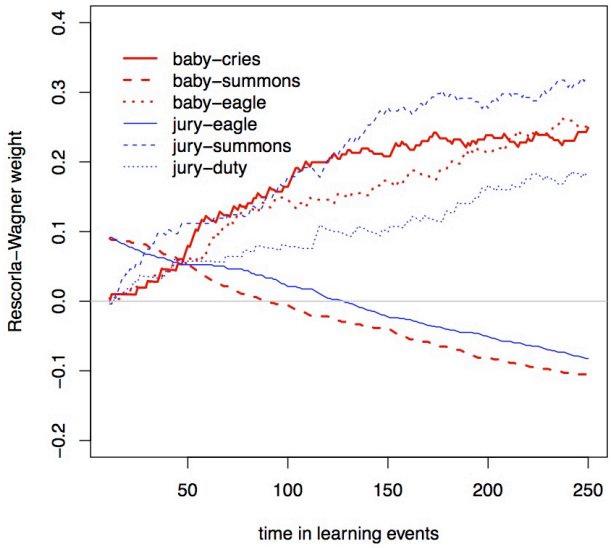


Figure 2: The development of word pair associations in the model's lexicon after each training epoch. As the model's experience of *jury* and *eagle* in other contexts increases, the strength of the *jury-eagle* association declines, such that a single new exposure will exert an ever-weaker influence of the learned strength of this pair. Eventually, the model develops a negative expectation for *eagle* given *jury*.

What do declining PAL scores reveal?

Simulating PAL using a standard associative learning rule predicts that hard items will get harder over time. Since empirical studies reveal the same pattern of change (Figure 1), this raises a question: Does lifespan PAL performance reveal cognitive declines, or the predictable outcome of learning? To examine this, Ramscar & Port (2015) used large text corpora to empirically derive parameters for the **background rates** (*w1* frequencies), **blocking** (frequency *w2* / frequency *w1*), and **association rates** (*w1-w2* co-occurrence rates) for the PAL pairs plotted in Figure 1. These parameters accounted for over 85% of the by-item variance in the observed performance of *both* the 20-29 and the 30-39 year-old age groups. As predicted by learning theory, the parameters for background rates and blocking were associated with lower scores, and the association-rate parameter was associated with higher scores, and sensitivity to each of the predictors was greater in the 30-39 than the 20-29 group (Ramscar & Port, 2015; see Ramscar, 2014 for a replication using different corpora).

An analysis of all the WMS-PAL normative data showed this pattern to be consistent across the lifespan (Ramscar et al., 2013c): oldest adults' (ages 60-69) performance showed the greatest sensitivity to the three lexical factors that learning theory predicts *should* influence the learnability of word-pairs, whereas the factors that caused negative associations to develop in the simulation did not significantly influence the youngest participants' performance (ages 20-29). In terms of learning the – obviously far more complex – ser of word associations in the English lexicon, 20-29 year-olds' performance is thus

consistent with learning at around epoch 50 in the simulation, and 60-69 year-olds' performance more akin to epoch 250.

It is thus clear that long-established principles of learning explain (1) why some PAL pairs are hard or easy (a fact psychometricians acknowledge yet fail to explain), and (2) why PAL performance ought to be expected to decline as adults age: In both cases, because of how the processes of "associative" learning have empirically been shown to work (i.e., discriminatively), experience not only teaches English-speakers which words go together, it *also* teaches them which words do not go together; and while this process increasingly differentiates meaningful and meaningless word-pairs (Figure 1), it simultaneously makes meaningless pairs ever harder to learn (Figure 2).

Behavioral Experiment

Not only do well-established models of the associative learning process predict that experience will increasingly inhibit the learning of word-pairs like *North-Dog*, they also provide a means for empirically deconfounding age and experience in PAL performance. Because of the way people are exposed to language throughout their lives, native (L1) speakers of similar ages and educational backgrounds inevitably have levels of first language experience that significantly exceed that of age-matched adult second-language (L2) speakers. Both the analysis and the simulation presented above make two clear and somewhat counterintuitive predictions about how these experiential differences will affect PAL performance:

1. Older native speakers (OAL1) ought to perform worse on lexical PAL tests than age-matched non-native speakers of a language (OAL2).
2. The difference between native and bilingual PAL performance can be expected to increase with growing experience (see Figure 2).

If, on the other hand, PAL tests actually do measure cognition independently of experience, as current practices assume, then OAL2 speakers should not out-perform OAL1s. Indeed, given that experience makes some PAL pairs easier than others, a naïve account ought to predict that OAL1's greater experience of the German items should lead them to outperform OAL2's.

Method

We tested these hypotheses on 20 young (18-28 year old) and 20 older (38-53 year old) monolingual speakers of German (a non-tonal language deriving most of its lexicon from the Germanic branch of the Indo-European language family) and two age-matched groups of 20 native speakers of Mandarin (a tonal member of the Sino-Tibetan language family), for whom German is a second language (Table 1; given that PAL is a reliable measure that is particularly sensitive to the effects of aging, Rabbitt & Lowe, 2000, and Ns < 20 are typical in neuropsychological studies that employ PAL tests, this sample was judged to be sufficient to test these hypotheses). The monolinguals completed a PAL test in German only, whereas the bilinguals completed Chinese and German PAL tests).

Table 1: By-group age and vocabulary scores for each of the groups of participants (standard deviations in brackets).

		Age	German Vocabulary	Chinese Vocabulary
Chinese-German Bilinguals	young	24.55 (2.27)	31.75 (5.35)	67.65 (6.46)
	old	43.60 (4.66)	40.25 (7.86)	64.65 (7.09)
German Monolinguals	young	23.45 (3.06)	81.95 (6.25)	- (-)
	old	44.90 (4.36)	84.10 (4.38)	- (-)

The harder items in the PAL task contained relatively infrequent words. To ensure participants knew all the words used in the test, vocabulary tests were conducted prior to the PAL task in both languages. Performance on the vocabulary test in German was highly variable, particularly for the young Chinese-German bilinguals. Participants in this group were typically graduate students at the University of Tübingen. While some of these students were highly competent in L2, others had considerably less experience. A number of these less experienced L2 learners explicitly stated that they did not know some of the words in the German PAL test.

To minimize the risk of young Chinese-German bilingual participants not knowing words in the PAL test, the 20 participants with the best performance on the vocabulary pretest in German were selected from a larger group of 34 young Chinese-German bilinguals. Vocabulary test performance was calculated as a weighted sum of the number of correct answers, with item weights being the proportion of correct answers for the items across the young and old bilingual participants.

The items for the vocabulary and paired associate learning tests in both languages were recorded from native female speakers of Mandarin Chinese and German in a sound booth using professional recording equipment.

In the vocabulary pre-test, participants were auditorally presented with a word and the 4 possible answers and were asked to select the answer that was most similar in meaning to the test word by clicking one of four buttons labeled 1 through 4 on the screen. Participants were asked to guess if they did not know the correct answer to a question.

In the paired associate learning task each block comprised a training phase and a test phase of 10 pairs. Participants were asked to memorize the pairs of words in the training phase. For the test phase, participants were asked to produce the word that formed a pair with the auditorily presented test word. The test phase was self-paced: participants were

asked to press the next button on the screen to move on to the next test word after verbally responding to a test word.

The average time required to complete the vocabulary pre-test for each language was about 30 minutes. The 3 blocks of paired associate learning took about 25 minutes per participant in each language, including a short break between each block.

Materials

The vocabulary tests for both German and Chinese consisted of 100 multiple choice questions with 4 possible answers. The 3 incorrect answers were chosen from the same parts-of-speech category as the correct answer.

The word frequency distributions for the German and Chinese vocabulary tests were matched, and items were designed to ensure that Cue Frequency, Response Frequency and Co-Occurrence Frequency (in Google documents) for the pairs in each language were approximately normal.

The PAL test in both languages consisted of three groups of 10 pairs. Pairs ranged in diffculty from easy (e.g.; “学校” - “读书” (school-study) or “Nord”-“Süd” (north-south)) to hard (e.g.; “Schiff” - “Puppe” (ship-doll) or “洋葱”-“手指” (onion-finger)). Words occurred no more than once in each of the paired associate learning tests.

Results

Analysis of our participants’ PAL performance of using generalized additive models (Wood, 2006) revealed a significant age by co-occurrence frequency interaction (χ^2 = 38.687; $p < 0.01$; co-occurrence frequency, which provides a simple, objective estimate of the easiness of PAL pairs – see Figure 2 – was measured here as the number of times *w1-w2* appear together in *Google* documents). This interaction differed depending on whether the PAL test was administered in participants’ first (L1) or second language (L2; $\chi^2 = 9.122$; $p < 0.05$). For young adults (YAs), L1 and L2 performance was similar, and the interaction between age and frequency ($\chi^2 = 19.658$; $p < 0.01$) did not differ between L1 and L2 ($\chi^2 = 1.357$; $p > 0.9$). This is consistent with previous analyses which showed that YA PAL performance is largely determined by association rates, and insensitive to background rates and blocking (Ramscar, Hendrix, Love & Baayen, 2013). By contrast, older participants performed better when the PAL test was administered in L2 (revealed by an age by co-occurrence frequency interaction ($\chi^2 = 36.335$; $p < 0.01$) that differed significantly between L1 and L2 ($\chi^2 = 14.959$; $p < 0.01$), with a main effect of L1/L2 indicating that OA performance was better in L2 than L1 ($z = 2.113$, $p < 0.05$).

The presence of a significant three-way interaction between age, association rate and first/second language reveals that the advantage for the older participants in their second language was not uniformly distributed. For the majority of the co-occurrence frequency range (19/25 of the easiest PAL pairs), the older participants performed better in L2 (Figure 6). However, for the very “hardest” PAL items – i.e., those with the lowest association rates – this pattern reversed, such that the older adults performed worse in L2.

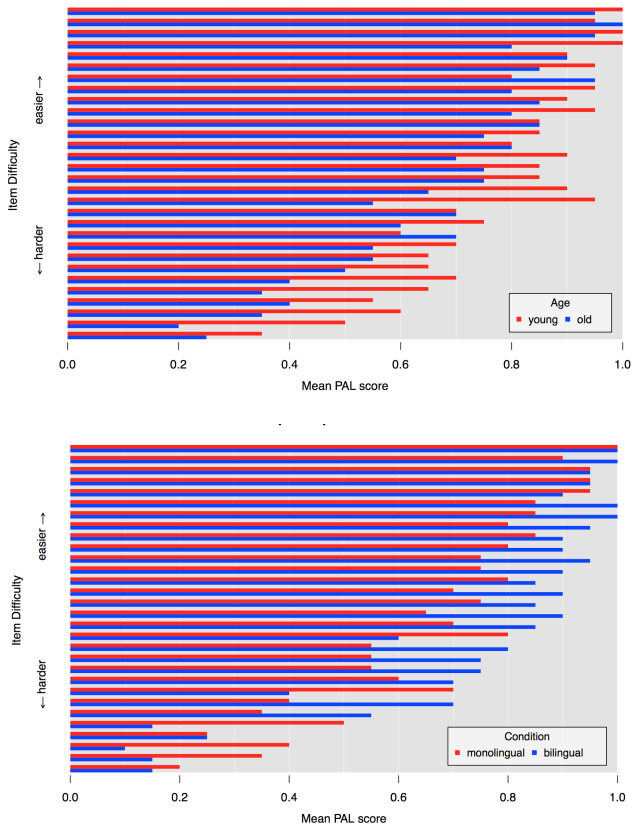


Figure 3: *Top Panel*: Average by-item German PAL performance for older and younger native German speakers (OAL1 and YAL2). *Bottom Panel*: Performance of age-matched older native German speakers (OAL1) and older Chinese-German bilinguals (OAL2) tested in German.

This highlights an important difference between older bilinguals and younger monolinguals: The former are not restricted to thinking about tasks in one language. Since there is little reason to suspect that items like *banana-lake* are any more related in Chinese than they are in German, it seems likely that when faced with learning the nonsensical links between the hardest items, OAL2s may have sought support from their native language. However, since well-learned native *dis*sociations will hinder learning on this test (and increase task complexity), this strategy will have had the opposite effect to that desired (see also Ramscar, Dye & Klein, 2013). Finally, the counterintuitive idea tested here – that increased language experience actually impairs PAL performance – gains further support from another finding from this study: OAs with doctoral degrees (the attainment of which likely involves a larger than normal amount of reading) performed significantly *worse* than OAs without them ($z = -2.073$; $p < 0.05$).

Discussion

In other work, we have shown how age-related performance changes on other neuropsychological tests are likely to reflect the effects of cumulative learning rather than cognitive decline (Ramscar et al, 2013c, Ramscar, Hendrix, Shaoul, Milin & Baayen, 2014; Blanco, Love, Ramscar, Otto, Smayda, & Maddox, 2016; Baayen, Tomaschek, Gahl, and Ramscar, *in press*). The present study extends this, showing how discriminative learning theories (Rescorla, 1988; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Ramscar et al., 2013a) predict that PAL performance will decline even when learning capacities are constant, simply because cumulative linguistic experience will make meaningless word-pairings ever harder to learn.

These predictions are supported by the results of our empirical study – which show that when age is controlled for, less linguistic experience predicts higher PAL scores – and by other studies of aging and associative learning, e.g.,: Naveh-Benjamin (2000) showed that older adults are worse at learning associations between unrelated "units of information" than when they are meaningfully related; Castel (2005) showed that older adults are better at associating realistic prices with grocery items than unrealistic prices; and Old & Naveh-Benjamin (2008) showed that adults encode less information about background context in memory tests as they age.

While these findings are usually taken reveal age-related "associative deficits" that are (somehow) lessened when associative information is consistent with the environment, it is notable that when the same pattern of learning the informative and neglecting the uninformative is seen as infants lose their sensitivity to non-native phonetic distinctions in the course of learning a language (Werker & Tees, 1984) it is not usually seen as decline. Our analyses have shown how all these patterns of data are consistent with what is known about the actual processes of associative learning (see also Ramscar, Suh & Dye, 2011), and that far from declining, associative processes appear to be stable and consistent across the lifespan.

It is often claimed that surprising results are more likely to be false (Lindsay, 2015), and since this last claim – and indeed, our main finding – may surprise many readers, it seems worth considering why. As we have sought to show, the widely held assumptions about learning that underpin current interpretations of lifespan PAL performance are fundamentally incompatible with the detailed understanding that has emerged out animal, behavioral and neuroscience research into associative learning processes (see Ramscar et al., 2013a, for a review). Some 30 years ago, Rescorla (1988) critically described the outdated and misleading way in which learning theory was taught in most areas of cognitive science, and what current interpretations of lifespan PAL data appear to reveal is that little has changed since. Accordingly, it is worth emphasizing that our main result is so well predicted by learning models that had we obtained opposite results, this would have suggested our entire understanding of learning (including Pavlovian conditioning) is wrong. This really would be surprising.

By contrast, despite widespread beliefs about its inevitability, the current behavioral evidence for "healthy cognitive decline" amounts to little more than a series of (mainly) negative correlations between neuropsychological test scores and increased adult age, while the biological evidence correlates the same declining test scores with predictable physical changes in neural morphology (such change in cortical grey matter, Sowell et al, 2003, Alemán-Gómez, 2013), response times (Deary & Ritchie, 2016), etc. It thus follows that determining whether biological changes in healthy adult brains are evidence of decline – or something else, such as a progressive adaptation to the metabolic cost of cognition (Attwell & Laughlin, 2001; Wen & Chklovskii, 2008) – depends on establishing that declining neuropsychological test scores actually do reflect functional declines. We have shown how PAL scores, thought to be particularly sensitive measures of age-related decline (Rabbitt & Lowe, 2000), reflect no such thing.

Accordingly, it is important to note that all current neuropsychological tests make use of learned information, while assuming that cognitive performance can be measured independently of experience (Deary et al, 2009; Salthouse, 2011; Singh-Manoux et al., 2012; Lindenberger, 2014). This assumption is incompatible with behavioral, animal, and neurobiological models of how information is actually learned (McDannald, Jones, Takahashi, & Schoenbaum, 2014; Daw, Courville, & Dayan, 2008; Schultz, 2006; Waelti, Dickinson, & Schultz, 2001; Sutton & Barto, 1981; Rescorla & Wagner, 1972). Learning theory differs from theories of cognitive decline in that its processes are clearly specified, and formalized in models. These models have proven adept at predicting behavior (such as that reported here), as well as the responses of the neural structures that appear to implement them in the brain, with surprising accuracy (Ramscar et al, 2013a; Schultz, 2006). Since learning theory contradicts many beliefs about cognitive decline, one or the other must be substantively wrong. We suggest it is the latter; and that it matters.

Acknowledgments

This work was supported by an Alexander von Humboldt award to R. Harald Baayen.

References

Alemán-Gómez, Y., Janssen, J., Schnack, H., Balaban, E., Pina-Camacho, L., Alfaro-Almagro, F., ... & Desco, M. (2013). The Human Cerebral Cortex Flattens during Adolescence. *Journal of Neuroscience*, 33(38), 15004-15010

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender. *Cognition*, 122(3), 292-305.

Attwell D & Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21:1133–1145

Baayen, R. H., Tomaschek, F., Gahl, S., and Ramscar, M. (in press). The Ecclesiastes principle in language change. In Hundt, Mollin & Pfenninger (Eds.), *The Changing English Language.* Cambridge, UK: Cambridge University Press.

Blanco, N. J., Love, B. C., Ramscar, M., Otto, A. R., Smayda, K., & Maddox, W. T. (2016). Exploratory Decision-Making as a Function of Lifelong Experience, Not Cognitive Decline. *Journal of Experimental Psychology: General.*

Castel, A. D. (2005). Memory for grocery prices in younger and older adults. *Psychology and Aging*, 20, 718–721

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning. *The Probabilistic Mind*, 431-452.

Deary, I. J., Corley, J., Gow, A J., Harris, S. E., Houlihan, L. M., Marioni, R. E., Penke, L., Rafnsson, S. B. & Starr, J. M. (2009). Age-associated cognitive decline. *British Medical Bulletin,* 92, 135–152.

Deary, I. J., & Ritchie, S. J. (2016). Processing speed differences between 70-and 83-year-olds matched on childhood IQ. *Intelligence*, 55, 28-33.

des Rosiers G & Ivison D. (1986). Paired-associate learning: normative data for differences between high and low associate word-pairs. *Journal of Clinical Experimental Neuropsychology*; 8: 637–42

Kamin, L. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. (New York: Crofts.

Lindenberger, U. (2014). Human cognitive aging. *Science*, 346(6209), 572-578.

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, 0956797615616374.

McDannald, M. A., Jones, J. L., Takahashi, Y. K., & Schoenbaum, G. (2014). Learning theory: a driving force in understanding orbitofrontal function. *Neurobiology of Learning and Memory*, 108, 22-27.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5), 1170.

Old, S. & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging*, 23, 104-118.

Rabbitt, P., & Lowe, C. (2000). Patterns of cognitive ageing. *Psychological Research*, 63(3-4), 308-316.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6), 909-957.

Ramscar, M., Suh E. & Dye, M. (2011) How pitch category learning comes at a cost to absolute frequency representations. *Proceedings of the 33rd Meeting of the Cognitive Science Society*, Boston, MA.

Ramscar, M., Dye, M., & McCauley, S. M. (2013a). Error and expectation in language learning. *Language*, 89(4), 760-793.

Ramscar, M., Dye, M., & Klein, J. (2013b). Children value informativity over logic in word learning. *Psychological Science*, 24(6), 1017-1023.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013c). Learning is not decline. *The Mental Lexicon*, 8(3), 450-481.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline. *Topics in Cognitive Science, 6(1), 5-42.*

*Ramsca*r, M. (2014). [Web log post]. Retrieved 1/15/2016 from ramscar.wordpress.com/2014/08/07/technical-note-bells-and-smells-and-replication/

Ramscar, M., & Port, R. F. (2015). Categorization (without categories). In Dawbroska & Divjak (Eds.), *Handbook of Cognitive Linguistics*. De Gruyter Mouton.

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53, 58-74.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of comparative and physiological psychology*, 66(1), 1.

Rescorla, R.A. & Wagner, A.R. A theory of Pavlovian conditioning. (1972) In Black & Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*. (New York: Crofts)

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151.

Salthouse, T. (2012). Consequences of age-related cognitive declines. *Annual Review of Psychology*, 63, 201.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, 57, 87-115.

Singh-Manoux, A., Kivimaki, M., Glymour, M. M., Elbaz, A., Berr, C., Ebmeier, K. P., ... & Dugravot, A. (2012). Timing of onset of cognitive decline: results from Whitehall II prospective cohort study. *BMJ*, 344.

Sowell ER, Peterson BS, Thompson PM, Welcome SE, Henkenius AL, Toga AW (March 2003). Mapping cortical change across the human lifespan. *Nature Neuroscience*, 6 (3): 309–15

Sutton, R. S., & Barto, A. G. (1981) Toward a modern theory of adaptive networks. *Psychological Review*, 88, 135

Waelti, P., Dickinson, A. & Schultz, W. (2001) Dopamine responses comply with basic

assumptions of formal learning theory. Nature, 412, 43-48.

Werker, J. & R. Tees. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*. 7, 49-63

Wen, Q. & Chklovskii, D. B. (2008). A cost-benefit analysis of neuronal morphology. *Journal of Neurophysiology* 99, 2320–2328

Wood, S. N. *Generalized additive models*. (New York: Chapman & Hall, 2006).

When are representations of causal events quantum versus classical?

James M. Yearsley (james.m.yearsley@vanderbilt.edu)
Jennifer S. Trueblood (jennifer.s.trueblood@vanderbilt.edu)
Department of Psychology, Vanderbilt University, Nashville TN, USA

Emmanuel M. Pothos (emmanuel.pothos.1@city.ac.uk)
Department of Psychology, City University London, London, EC1V 0HB, UK

Abstract

Throughout our lives, we are faced with a variety of causal reasoning problems. Arguably, the most successful models of causal reasoning, Causal Graphical Models (CGMs), perform well in some situations, but there is considerable variation in how well they are able to account for data, both across scenarios and between individuals. We propose a model of causal reasoning based on quantum probability (QP) theory that accounts for behavior in situations where CGMs fail. Whether QP or classical models are appropriate depends on the representation of events constructed by the reasoner. We describe an experiment that suggests the representation of events can change with experience to become more classical, and that the representation constructed can vary between individuals, in a way that correlates with a simple measure of cognitive ability, The Cognitive Reflection Task.

Keywords: Causal reasoning, quantum probability theory, Bayes networks, order effects, Bayesian parameter estimation

Introduction

Reasoning about the causal relationships between events is an important component of everyday cognition that allows us to make sense of the world. For example, I know that when I plug in and turn on an electric kettle, water boils. Causal reasoning is also an essential component of problem solving, for example if I plug in and turn on the same kettle and water doesn't boil, I might infer that the fuse has blown.

Human reasoners are often very competent at performing causal reasoning tasks, in the sense of providing judgments aligning well with normative prescription. Probably the most successful class of models of causal reasoning are Causal Graphical Models (CGMs) based on Bayes nets. These models are normative, since they represent causal relationships using Bayes' calculus (Pearl, 1988) and have been successfully applied to a variety of human causal reasoning problems (Tenenbaum, Griffiths & Kemp, 2006; Griffiths & Tenenmaum, 2009). In spite of the apparent success of CGMs, there have been several recent studies that report violations of some of the predictions of these models including asymmetries between predictive and diagnostic reasoning, order effects, and violations of the casual Markov condition (Sloman & Fernbach, 2011; Trueblood & Busemeyer, 2012; Rehder, 2014). Equally, there seems to be considerable variation between individuals in how closely performance matches prescription (Rehder, 2014).

The challenge of modeling behavior that violates the normative prescriptions of classical (Bayesian) probability theory is not unique to causal reasoning. Recently, some researchers have been pursuing quantum probability theory (QP) as a way of modeling such behavior (Busemeyer & Bruza, 2011). QP is essentially the mathematical theory of probability associated with quantum theory, abstracted from the physics. QP contains features, such as contextuality, order effects and constructive judgments that appear to match well the way human decision makers often reason. One aim of this paper is to demonstrate that it is possible to build models of causal reasoning based on QP which provide a good description of behavior in situations where classical models fail.

However the introduction of QP models presents us with a new question, in what situations do people adopt a quantum representation of information as opposed to a classical one? We argue that classical and QP models can be seen as different cases in a hierarchy of models which differ in terms of the way events are represented. The QP model we introduce is in some sense the simplest possible, as it involves representing events in the smallest possible sample space (2D). Experience with a scenario may allow reasoners to construct a more complex representation, with a larger dimensional sample space, ultimately resulting in a fully classical model of causal relations. Equally, it is possible that simpler representations may be used to make heuristic judgments, while more complex representations are only formed when reasoning in a more deliberative way. If this is correct, the type of representation used by a decision maker may be linked to other individual differences such as the score on a Cognitive Reflection Test (CRT) (Frederick, 2005), or may be influenced by experimental instructions, such as speed vs accuracy prompts. Thus a second aim of this paper is to show that differences in performance in causal reasoning tasks can be explained in terms of the properties of the mental representation the individual constructs, and that this can be related both to experience with the task and to individual differences in cognitive ability.

We test our ideas in an experiment where participants were asked to make judgments involving conditional probabilities about various novel categories. We constructed a classical model and a QP model and used Bayesian analysis to compare them. Three main findings emerge, 1) Both the classical and the QP models provide good fits for some sets of participants, 2) The fits of the classical/QP models get better/worse as participants gain experience in the task, 3) Participants who scored highest/lowest on the CRT task tended to be better fit by the classical/QP model.

These findings suggest that the type of representation an individual constructs to reason about causal relations can vary both across individuals and with experience. We argue that

this may shed light on why classical Bayesian models of causal reasoning seem to match behavior in some situations, but not in others.

Experiment 1

Our experiment uses a paradigm developed by Rehder (2003) to study causal reasoning with novel categories. Our aim is to assess two phenomena from the judgment literature, order effects and the "inverse fallacy" (Villejoubert & Mandel, 2002), in causal reasoning and how these phenomena are related to individual differences in cognitive ability. Order effects occur when the order of information influences judgments (e.g., $p(E|X,Y)$ is not judged equal to $p(E|Y,X)$). It has also been shown that people commit what is known as the "inverse fallacy", that is, judging $p(X|Y) = p(Y|X)$. Both of these empirical findings are difficult to reconcile with an approach to causal reasoning based on Bayesian probability theory.

Methods

60 undergraduate students from Vanderbilt University participated in the experiment online at a time of their choosing for course credit. Participants were randomly assigned to one of two novel animal categories (either Lake Victoria Shrimp or Kehoe Ants). Each animal had three binary features (X, Y, and E) where two of the features (X, Y) causally influenced the third (E) to form a common effect network. For example, in the Lake Victoria Shrimp category, X = ACh neurotransmitter (high or low amount), Y = sleep cycle (accelerated or normal), and E = body weight (high or low). Participants were given information about the typicality of feature values. For example, they were told that "Most Lake Victoria Shrimp (90%) have a high amount of ACh whereas a few (10%) a low amount of ACh". In both categories, 90% of animals had feature X_1, 10% had feature Y_1, and 50% had feature E_1. After studying this information, they took a multiple-choice test with six questions that tested them on this knowledge. Participants were required to answer each question correctly before moving on to the next one.

Participants were then told the causal relationships between features. These relationships were described as one feature causing another. In both categories, X_1 and Y_1 cause E_1. Participants were also told there were no known relationships between X and Y. After reading this information participants took another multiple-choice test with eight questions testing them on this new knowledge. As before, participants were required to answer each question correctly before moving on to the next one. Finally, participants were asked to take a few minutes to review the features and relationships one more time. After they finished reviewing this information, they completed a third multiple-choice test with 10 questions. In this final test, participants were only given one opportunity to answer each question. Their score on this test was used to gauge how well they learned the features and causal relations.

After this test, participants completed six blocks of trials where they were asked to make judgments about the value of different features. There were two block types (BX and BY) that were repeated three times in an alternating fashion (e.g., BX, BY, BX, BY, BX, BY). Participants were randomly assigned to start with either the BX or BY block.

Each block contained nine judgment questions where participants were asked to select the value of a particular feature (see Table 1). At the start of each question, participants were told that a biologist caught a new animal (either shrimp or ant) and were queried about one of the features of that animal. For example, in the Lake Victoria Shrimp category, they might be asked "What type of body weight do you think this shrimp has?" Participants were given three response options: feature value 1, feature value 2, or equally likely to be feature value 1 or 2. For example, in the question about body weight, the response options were 1) a low body weight, 2) a high body weight, and 3) equally likely to be low or high.

Some questions asked participants to make a sequence of judgments about a feature value (e.g., E) as they learned new information about the other features (e.g., X, Y). For example, they might be asked about the body weight of a shrimp given lab tests that showed the shrimp had a high amount of ACh neurotransmitter (i.e., $E|X_1$). Participants might then be asked to reevaluate body weight based on additional lab tests that showed the shrimp also had a normal sleep cycle (e.g., $E|X_1,Y_2$). Note that information about the value of the first feature (e.g., X_1) remained on the computer screen when new information about the second feature (e.g., Y_2) was presented. This was to reduce the chance of memory failures. In the BX (BY) block, information about feature X (Y) was always presented before information about feature Y (X) in sequences involving both features. This helped reduce the influence of memory on future judgments about reverse orderings.

After finishing the six judgment blocks, participants completed the CRT (Frederick, 2005). This test assesses individual's ability to suppress a spontaneous and intuitive ("System 1") wrong answer in favor of a deliberative and reflective ("System 2") correct answer. The test consists of three items using a free-response format and is scored by counting the number of correct responses across the items. The CRT has been correlated with many behavioral measures including temporal discounting, mental heuristics, and risk preferences (Frederick, 2005; Toplak, West, & Stanovich, 2011).

Results

The average score on the 10 question multiple choice test was 9.6 indicating most participants correctly learned the feature values and causal relationships during the first part of the experiment. For analyses of the judgment data, we scored responses in a similar way to Rehder (2014) by assigning the following values to the three response options: feature value 1 = 1, feature value 2 = 0, and equally likely = 0.5. Note that there were no differences between judgments in the two different animal categories and so responses were collapsed for the following analyses. We also grouped individuals into three groups based on CRT scores: high = CRT score of 3, medium = CRT score of 1,2, and low = CRT score of 0. We combined individuals with CRT scores of 1 and 2 into a sin-

Table 1: Judgments in Experiment 1

Block	Judgments									
BX	E	X	Y	$E\|X_1$	$E\|X_1,Y_2$	$E\|X_2$	$E\|X_2,Y_1$	$X\|Y_2$	$Y\|X_1$
BY	E	X	Y	$E\|Y_1$	$E\|Y_1,X_2$	$E\|Y_2$	$E\|Y_2,X_1$	$Y\|X_2$	$X\|Y_1$

gle group so we had roughly an equal number of individuals per group. There were 21 participants in the CRT high group, 19 in the CRT medium group, and 20 in the CRT low group.

Order effects were assessed by comparing the judgments $E|X_1,Y_2$ and $E|Y_2,X_1$ and the judgments $E|X_2,Y_1$ and $E|Y_1,X_2$. For each individual, we calculated an "order effect score" defined as $|(E|X_1,Y_2)-(E|Y_2,X_1)|+|(E|X_2,Y_1)-(E|Y_1,X_2)|$. Higher scores indicate larger order effects and consequently a more "quantum" representation of information. By grouping the blocks into pairs (i.e., blocks 1 and 2, blocks 3 and 4, blocks 5 and 6), we can calculate three different order effect scores for each individual. These can be used to examine changes in order effects due to experience gained through repetition. The top panel of Figure 1 shows the order effect scores for the three CRT groups across the block pairs. A repeated measures ANOVA showed a main effect of order effect score ($F(2,118) = 9.86$, $p < .001$) and CRT group ($F(2, 57) = 3.45$, $p = 0.04$). A Bayesian analysis revealed a similar conclusion with a Bayes Factor of 405.34 for a model including order effect score and CRT group over a null model.

The inverse fallacy was assessed by comparing the judgments $X|Y_1$ and $Y|X_1$ and the judgments $X|Y_2$ and $Y|X_2$. Similar to the order effect score, we can also calculate an "inverse fallacy score" defined as $|(X|Y_1)-(Y|X_1)|+|(X|Y_2)-(Y|X_2)|$. Lower scores (closer to zero) indicate larger degrees of the fallacy and consequently a more "quantum" representation of information. Like the order effect score, we grouped blocks into pairs and calculated three different order effect scores for each individual. The bottom panel of Figure 1 shows the inverse fallacy scores for the three CRT groups across the block pairs. A repeated measures ANOVA showed a significant interaction between inverse fallacy score and CRT group ($F(4, 118) = 3.22$, $p = 0.015$). A Bayesian analysis revealed a similar conclusion with a Bayes Factor of 3.13 for a model including the interaction of inverse fallacy score and CRT group over a null model.

Conclusions

The results show a clear separation between the order effect scores for the high CRT group versus the other two, suggesting the high CRT group is more likely to be using a classical representation. Figure 1 also shows that as participants gain experience with the scenarios the degree to which they display the (non-classical) order effect decreases. This suggests that participants' representation of the events changes with experience; becoming gradually more classical and less quantum. The data also show an interaction between inverse fallacy score and CRT group. This provides evidence that the low CRT group is more likely using a QP representation. Below we directly compare the performance of a classical and

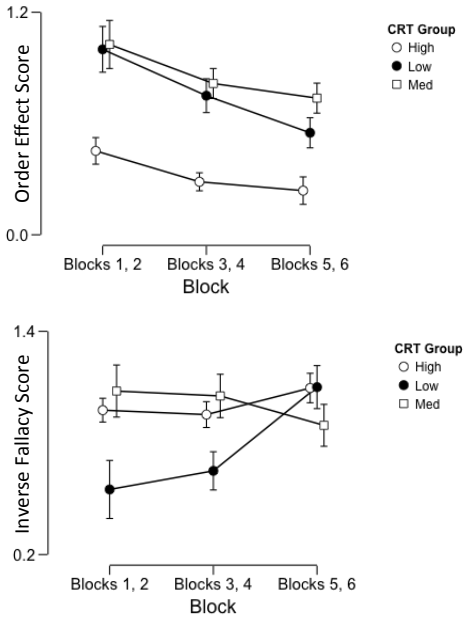


Figure 1: Top panel: Order effect scores for three CRT groups across block pairs. Bottom panel: Inverse fallacy scores for three CRT groups across block pairs. Error bars show the standard error.

QP model for this data.

Quantum Models

Before we introduce the details of the models, we provide a short introduction to QP theory. Recently a number of researchers have been investigating models of cognition based on the mathematics of QP theory (Busemeyer & Bruza, 2011). These models share the property that they are probabilistic, but the rules for assigning probabilities to events are those abstracted from quantum theory, rather than the usual Kolmogorov axioms. These models have features, such as contextuality, order effects and interference effects that seem to align well with human reasoning, at least in some cases.

One important topic in the study of quantum models of cognition is understanding the conditions under which they do and do not apply. Clearly much human decision making can be adequately described using Bayesian probability theory and in these situations, although quantum models can perfectly account for behavior also, they are clearly superfluous. The difference between classical and quantum models is often phrased in terms of the way different events are represented by a reasoner, either as *compatible* or *incompatible* (we will explain these terms shortly), and so attempts to delimit the realm of application of quantum models have often focussed on the question of whether different events may be

represented in a compatible way. There are few concrete results in this area, but it is generally believed that experience with a particular situation, either from previous familiarity or acquired through learning, may allow events to be represented in a compatible way, whereas relatively novel situations are more likely to be represented in a incompatible way. In addition, quantum models are invoked to explain a similar set of phenomena as heuristics (Busemeyer et al. 2011), and so it seems plausible that incompatible representations of events, associated with quantum models, should be preferentially used for decisions executed more quickly with little conscious deliberation.

Compatible events are ones that may be assigned a simultaneous truth value. Thus, if event X and event Y are compatible, the conjunction $X \wedge Y$ is well defined. Probabilities for compatible events obey the Kolmogorov axioms. Two immediate consequences are that for compatible events X and Y,

$$\begin{aligned} p(X \wedge Y) &= p(Y \wedge X), \\ p(X|Y) &= p(Y|X)\frac{p(X)}{p(Y)} \end{aligned} \tag{1}$$

(Note the connection to order effects and the inverse fallacy.)

Almost all events that we counter in everyday life can in principle be represented in a compatible way. However doing so requires that decision makers have access to the joint probabilities of all of these events. This may be unfeasible from the point of view of memory capacity, since the number of probabilities grows exponentially with the number of events being considered. Equally, these probabilities might be difficult to learn, since joint probabilities correspond to subsets of the sample space, and if it takes a finite number of previous experiences to learn the approximate measure of each subset, then the amount of experience required again grows exponentially with the number of events considered.

In contrast with compatible events, incompatible ones are those for which $X \wedge Y$ is undefined. Thus although the probabilities $p^Q(X)$ and $p^Q(Y)$ exist, the joint $p^Q(X \wedge Y)$ may not. Typically one can define a modified version of conjunction with an explicit ordering, eg $X \wedge Y$ is taken to mean X *and then* Y for incompatible variables. This implies that,

$$\begin{aligned} p^Q(X \wedge Y) &\neq p^Q(Y \wedge X), \\ p^Q(X|Y) &\neq p^Q(Y|X)\frac{p^Q(X)}{p^Q(Y)}. \end{aligned} \tag{2}$$

In quantum models, one can choose to model two events as either compatible or incompatible, depending on the representation one chooses. If all events are chosen to be compatible one recovers a classical model, while if no two events are compatible (except for the trivial case of an event and its negation) then one has a maximally quantum model. If there are more than two possible events then there can be intermediate representations where some subset of events are compatible. Thus quantum models encompass classical ones, and we should more accurately speak of a hierarchy of different representations, from fully quantum to fully classical.

In quantum theory events are represented by subspaces of a Hilbert space (essentially a vector space), with associated projection operators P_i, and the initial knowledge state by an operator on this space known as the density operator ρ. In a particular basis these are both simply matrices, and the computations involved in computing probabilities just linear algebra. The probability a participant assigns to event X given that her initial knowledge state is ρ is simply,

$$p(X) = \mathrm{Tr}(P_X \rho) \tag{3}$$

where Tr denotes the trace of a matrix. The probability a participant assigns to E given X is,

$$p(E|X) = \frac{\mathrm{Tr}(P_E P_X \rho P_X)}{\mathrm{Tr}(P_X \rho)} \tag{4}$$

The number of parameters it takes to specify ρ and the P_i's depends on the dimension of the Hilbert space, which in turn depends on whether the representation of different events is chosen to be compatible or incompatible. We will see in the next section how this works for the specific cases we are interested in, but generally there are some parameters that fix the initial state, and then some that determine the relationship between the projection operators representing different events (i.e., exactly how incompatible two events are).

The Classical and QP Causal Reasoning Models

In this section we briefly outline the QP and classical models of causal reasoning we propose to test. As mentioned above, the distinction between quantum and classical here is rather artificial - for our purposes a classical model is simply a QP one where all the events in question are compatible. Our models incorporate three possible events, two causes X and Y, and an effect E together with their negations. We are interested in probabilities for single events such as $p(X)$ and also conditional probabilities for the effect given one or more event, $p(E|X), p(E|X,Y)$ and for one cause given another, $p(X|Y)$.

The first choice we need to make is which events are compatible, and which incompatible. Because there are three events there are several possible options, for the present study we examine two choices; either all events are compatible, leading to a classical model, or all events are incompatible, leading to a maximally quantum model. Intermediate models exist, and may be desirable since the different models display different combinations of non-classical behaviors.

Let us consider the classical model first. Since all events are compatible it must be possible to assign truth values to propositions such as $X \wedge Y \wedge E$, and so the space needs to contain vectors which represent these states. Therefore our space needs to be 8D to encompass the set of possible states. We may use the following basis of states[1],

$$\begin{aligned} |XYE\rangle &= \begin{pmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^T, \\ |XY\bar{E}\rangle &= \begin{pmatrix}0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}^T, \\ &\vdots \\ |\bar{X}\bar{Y}\bar{E}\rangle &= \begin{pmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 1\end{pmatrix}^T \end{aligned} \tag{5}$$

[1] We use standard Bra-Ket notation (Busemeyer & Bruza, 2011).

and projection operators,

$$\begin{aligned} P_X &= |XYE\rangle\langle XYE| + |XY\bar{E}\rangle\langle XY\bar{E}| \\ &\quad + |X\bar{Y}E\rangle\langle X\bar{Y}E| + |X\bar{Y}\bar{E}\rangle\langle X\bar{Y}\bar{E}| \\ &= \text{diag}(1,1,1,1,0,0,0,0) \text{ etc.} \end{aligned} \tag{6}$$

The initial state may be a general density matrix, however it turns out that the probabilities we compute are sensitive only to the diagonal elements of ρ. Therefore we may take,

$$\rho = \text{diag}(\rho_{11}, \rho_{22}, \ldots, 1 - \rho_{11} - \rho_{22} - \ldots \rho_{77}) \tag{7}$$

It is easy to compute the various probabilities of interest in terms of the ρ_{ii}. There are therefore seven parameters in the classical model, $\{\rho_{11}, \rho_{22}, \ldots, \rho_{77}\}$. Since all events in the classical model are compatible, there are no predicted order effects, e.g. we expect,

$$p(E|X,Y) = p(E|Y,X) \tag{8}$$

Now let us turn to the maximally quantum model. Since all events are incompatible we can span the space with any single pair $\{X, \bar{X}\}$. For this reason the space we need is 2D. Any two events such as X and Y are related to each other via a unitary transformation in this space. Any unitary transformation may be parameterised in the following way,

$$R_a = \begin{pmatrix} \cos(\theta_a) & -\sin(\theta_a)e^{i\phi_a} \\ \sin(\theta_a)e^{-i\phi_a} & \cos(\theta_a) \end{pmatrix} \tag{9}$$

and we have

$$P_X = R_X \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} R_X^\dagger \tag{10}$$

We will take the initial state to be diagonal[2], $\rho = \text{diag}(\rho, 1-\rho)$ and it turns out we can set one of the ϕ parameters to be 0 without loss of generality. Thus we have six parameters in total for the quantum model, $\{\rho, \theta_E, \theta_X, \phi_X, \theta_Y, \phi_Y\}$.

One interesting feature of the 2D model is that because all the events are represented by projection operators onto one dimensional subspaces, various expressions for the probabilities simplify. One example is,

$$\begin{aligned} p(X|Y) &= \frac{\text{Tr}(P_X P_Y \rho P_Y)}{\text{Tr}(P_Y \rho)} = \frac{\langle X|Y\rangle\langle Y|\rho|Y\rangle\langle Y|X\rangle}{\langle Y|\rho|Y\rangle} \\ &= |\langle X|Y\rangle|^2 = \frac{\langle Y|X\rangle\langle X|\rho|X\rangle\langle X|Y\rangle}{\langle X|\rho|X\rangle} = p(Y|X) \end{aligned} \tag{11}$$

which means that the 2D model exhibits the "inverse fallacy".

The causal reasoning models we have developed output probabilities for various combinations of events. However the experimental set up we used involves a choice rather than a judgment similar to experiments by Rehder (2014). For this reason we run each predicted probability through a softmax function to simulate the fact that participants are forced to choose between the possible alternatives rather than outputting the exact probability. The softmax function has a standard form (eg Rehder, 2014) and involves two extra parameters, λ, τ that are the same in the classical and QP models.

[2]This is a useful trick. It is difficult to check whether a general matrix is an allowable density matrix, but writing the matrix in a basis where it is diagonal the test becomes trivial.

Model Fitting

Parameters for the classical and quantum model were estimated using a Bayesian analysis for each of the nine conditions (CRT={High, Medium, Low} $\times$ Block={1+2, 3+4, 5+6}.) For the classical model the priors for the ρ_{ii} were all taken to be uniform in the interval $[0,1]$, and then normalized to ensure $\sum_i \rho_{ii} = 1$. For the quantum model the prior for the ρ parameter was taken to be uniform in the interval $[0,1]$ and the priors for all angle parameters were taken to be $(\pi/2) \times \beta(2,2)$. Three MCMC chains were used to estimate the posterior distributions using JAGS. Chain convergence was assessed using the $\hat{R}$ statistic and all chains had good convergence behavior.

For each condition we also used JAGS to compute the DIC associated with the classical and QP models. The DIC is a generalization of the BIC, wth smaller values indicating a better model fit. A difference of three in the DIC between two models is usually taken to be indicate a significant difference in fit. The results are shown in Figure 2.

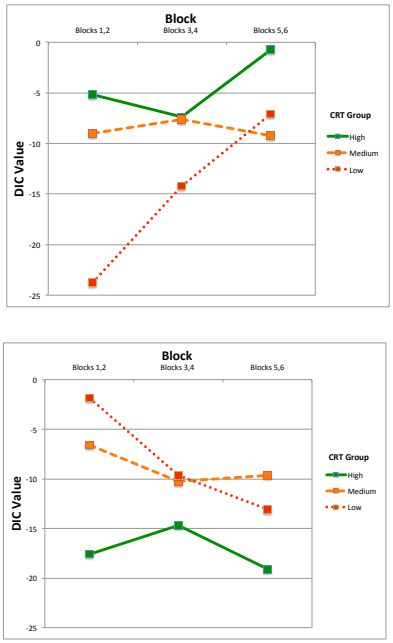


Figure 2: Top panel: DIC values for the QP model as a function of block number for each of the CRT groups. Bottom panel: DIC values for the CP model as a function of block number for each of the CRT groups.

For the QP model the DIC values show the following patterns; first the model performs better in participant groups where the CRT score is lower. Second, while the model performance for the medium and high CRT groups does not vary much across blocks, performance clearly decreases with block number for the low CRT group. For the classical model the opposite behavior is observed; first the model performs better in participant groups with a higher CRT score. Second, while performance for the high and medium CRT groups does not vary much across blocks, performance clearly increases

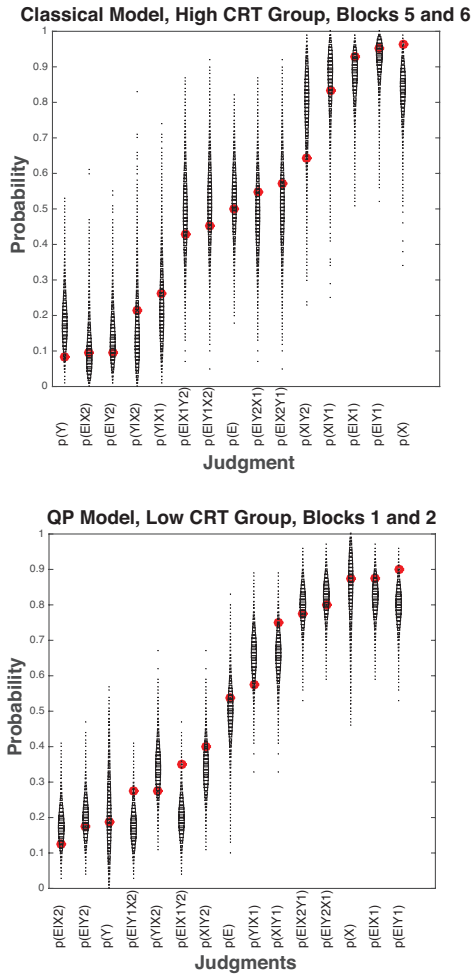


Figure 3: Top panel: Posterior probabilities (black squares) compared with empirical data (red dots) for the CP model, high CRT group and blocks 5+6. Bottom panel: Posterior probabilities compared with empirical data for the quantum model, low CRT group and blocks 1+2.

with block number for the low CRT group. Two examples of the model fits are shown in Figure 3.

Conclusion

The conclusions from the Bayesian model fitting corroborate the evidence of the diagnostic "order effect" and "inverse fallacy" scores. Overall, the representation of events that participants use to reason about causal relations appears to change as participants gain familiarity with the scenario, from an initially quantum or incompatible one to a final classical or compatible one. In addition, we found evidence that there are individual differences between participants in terms of their tendency to use classical or QP representations, which is associated with the CRT score.

Overall, our work sheds light on why causal graphical models (CGMs) have been successful in many situations, but can sometimes fail to agree with behavior. Equally it helps us understand why QP models can sometimes be successful but are superfluous in many cases. Reasoning about causal relations is neither inherently classical or quantum but rather is tied to the representation of events constructed by the reasoner. For novel scenarios or when reasoning quickly, representations of events may be incompatible and QP models are appropriate, however experience or more deliberative reasoning can lead to the formation of more complex compatible representations, which support classical causal reasoning.

Acknowledgments

JMY and JST were supported by NSF grant SES-1556415. EMP and JMY were supported by Leverhulme Trust grant RPG-2013-00. EMP was supported by Air Force Office of Scientific Research (AFOSR), Air Force Material Command, USAF, grants FA8655-13-1-3044. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright notation thereon.

References

Busemeyer, JR & Bruza, P (2011). Quantum models of cognition and decision. CUP: Cambridge, UK.

Busemeyer, JR, Pothos, EM, Franco, R, & Trueblood, J. (2011). A quantum theoretical explanation for probability judgment errors. *Psych. Rev.*, 118, 193-218.

Frederick, S (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives.* 19, 25-42.

Griffiths, TL & Tenenbaum, JB (2009). Theory-based causal induction. *Psych. Rev.* 116(4), 661-716.

Pearl, J (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.

Rehder, B (2003). A causal-model theory of conceptual representation and categorization. *JEP:LMC* 29, 1141-1159.

Rehder, B (2014). Independence and dependence in human causal reasoning. Cognitive psychology, 72, 54-107.

Sloman, SA & Fernbach, PM (2011). Human representation and reasoning about complex causal systems, *Information, knowledge, systems management* 10, 1-15.

Tenenbaum, JB, Griffiths, TL & Kemp, C (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*. 10(7), 309-318.

Toplak, ME, West, RF, Stanovich, KE (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*. 39, 1275-1289.

Trueblood, JS & Busemeyer, JR (2012). A quantum probability model of causal reasoning. *Frontiers in Cognitive Science*, 3, 1-13.

Villejoubert, G, & Mandel, DR (2002). The inverse fallacy: An account of deviations from Bayes' theorem and the additivity principle. Memory & Cognition, 30(2), 171-178.

A model of conditional probability judgment

Fintan J. Costello (fintan.costello@ucd.ie)
School of Computer Science,
University College Dublin, Dublin 4, Ireland

Paul Watts (watts@thphys.nuim.ie)
Department of Mathematical Physics,
National University of Ireland Maynooth, Maynooth, Ireland

Abstract

A standard view in cognitive psychology is that people estimate probabilities using heuristics that do not follow probability theory. We describe a model of probability estimation where people do follow probability theory in estimation, but are subject to random error or noise. This model predicts that people's conditional probability estimates will agree closely with probability theory for certain noise-cancelling expressions, but deviate from probability theory for other expressions. We describe an experiment which strongly confirms these predictions, suggesting that people estimate conditional probabilities in a way that follows standard probability theory, but is subject to the biasing effects of random noise.

Introduction

Probability theory provides a calculus of chance describing how to make optimal predictions under uncertainty. Up to the 1960s the standard view in psychology was that people's probabilistic reasoning essentially followed probability theory. However, various systematic biases in people's probability judgements (many identified in the 1970s and 1980s by Tversky, Kahneman and colleagues) led researchers to conclude that, in fact, people do not follow probability theory but instead estimate probabilities using various heuristics. While these heuristics often yield reasonable judgments, they can also produce strong biases in people's probabilistic reasoning in certain cases (Tversky and Kahneman, 1973).

In this paper we return to the view that people follow probability theory when reasoning about uncertainty. We present a simple model of conditional probability judgment (judgment of probabilities $P(A|B)$: the probability of A given that B has occurred) where people estimate probabilities according to probablity theory, but are subject to random error or noise in recall from memory. Importantly, this model predicts that bias will be 'cancelled' for certain combinations of conditional probability estimates, and so those combinations should agree closely with probability theory. In an experiment testing this prediction, we find close agreement with probability theory for noise-cancelling expressions, alongside systematic deviation from probability (in just the direction predicted by the model) for other expressions.

Estimating probabilities

In standard probability theory, the probability of some event A is estimated by drawing a random sample of events, counting the number of those events that are instances of A, and dividing by the sample size. In probability theory the expected value of these estimates is equal to $P(A)$, the probability of A; individual estimates vary with an approximately normal distribution around this value. We assume that people estimate the probability of A in exactly this way: randomly sampling episodes from memory, counting the number that are A, and dividing by the sample size. We assume a long-term memory from which a random sample of episodes or traces can be drawn. For some event A we assume that each episode i holds a flag set to 1 if i contains event A and set to 0 otherwise. An estimate for the probability of A is obtained by randomly sampling episodes from memory, counting the number where the flag for A is set to 1, and dividing by the sample size.

If this counting process was error-free, people's estimates would have an expected value of $P(A)$. Human memory is subject to various forms of random error, however. To reflect this we assume some small probability $d < 0.5$ that when some flag is read, the value obtained is not the correct value for that flag. We assume that this noise is symmetric, so that the probability of 1 being read as 0 is the same as that of 0 being read as 1. We also assume a minimal representation where every type of event, be it a simple event A, a conjunctive event $A \wedge B$, a disjunctive event $A \vee B$, or any other more complex form, is represented by such a flag, all with same probability d of being read incorrectly.

A randomly sampled event will be counted as A if the event truly is A and its flag is read correctly (this occurs with probability $(1-d)P(A)$, since $P(A)$ events are truly A and flags have a $1-d$ chance of being read correctly), or if the event is truly $\neg A$ (not A) and its flag is read incorrectly as A (this occurs with probability $(1-P(A))d$, since $1-P(A)$ events are truly $\neg A$, and flags have a d chance of being read incorrectly). The expected value or average for a noisy estimate of $P(A)$ is the sum of these two terms:

$$\langle P_E(A) \rangle = (1-2d)P(A) + d \tag{1}$$

with individual estimates varying around this value. This average is systematically biased away from the 'true' probability $P(A)$, such that estimates will tend to be greater than $P(A)$ when $P(A) < 0.5$, and less than $P(A)$ when $P(A) > 0.5$. This model explains observed patterns of bias in probability estimates such as conservatism, underconfidence, subadditivity, binary complementarity and the conjunction and disjunction fallacies (see Costello and Watts, 2014, 2016).

Table 1: 16 identities that must have a value of 0 in standard probability theory. Our model predicts if these identities are computed from people's individual probability estimates for any pair of events A, B, values obtained for identities 1 to 8 will have a mean of 0, the value required by probability theory. Our model predicts that identities 9 to 16 will have values that are positive and significantly different from 0, with identities 13 to 16 having values approximately half those of identities 9 to 12.

Label	Identity		Predicted value
1	$P(A)+P(B)-P(A\wedge B)-P(A\vee B)$	$=$	0
2	$P(A)+P(B\wedge\neg A)-P(B)-P(A\wedge\neg B)$	$=$	0
3	$P(A\|B)P(B)-P(B\|A)P(A)$	=	0
4	$P(A\|B)P(B)+P(A\|\neg B)-P(A\|\neg B)P(B)-P(A)$	=	0
5	$P(B\|A)P(A)+P(B\|\neg A)-P(B\|\neg A)P(A)-P(B)$	=	0
6	$P(B\|A)P(A)+P(A\|\neg B)-P(A\|\neg B)P(B)-P(A)$	=	0
7	$P(A\|B)P(B)+P(B\|\neg A)-P(B\|\neg A)P(A)-P(B)$	=	0
8	$P(A\|\neg B)+P(B)+P(B\|\neg A)P(A)-P(B\|\neg A)-P(A)-P(A\|\neg B)P(B)$	=	0
9	$P(A)+P(B\wedge\neg A)-P(A\vee B)$	=	d
10	$P(B)+P(A\wedge\neg B)-P(A\vee B)$	=	d
11	$P(A\wedge\neg B)+P(A\wedge B)-P(A)$	=	d
12	$P(B\wedge\neg A)+P(A\wedge B)-P(B)$	=	d
13	$P(A\wedge B)-P(A\|B)P(B)$	=	$d/2$
14	$P(A\wedge B)-P(B\|A)P(A)$	=	$d/2$
15	$P(A\wedge B)+P(A\|\neg B)P(B)-P(A)-P(A\|\neg B)$	=	$d/2$
16	$P(A\wedge B)+P(B\|\neg A)-P(B)-P(B\|\neg A)P(A)$	=	$d/2$

Predictions

Consider the identities given in Table 1. Probability theory requires that when the terms in the first identity (identity 1) are summed, the resulting value must be 0 for all events A and B. Our model also predicts that, on average, people's estimates for this identity will also sum to 0. For example suppose we ask people to estimate $P(A), P(B), P(A\wedge B)$ and $P(A\vee B)$ and combine each person's estimates in the form of identity 1. Since the expected value of a sum is equal to the sum of expected values of its terms, the expected value for this combination is, using Equation 1,

$$\langle P_E(A)\rangle+\langle P_E(B)\rangle-\langle P_E(A\wedge B)\rangle-\langle P_E(A\vee B)\rangle=$$
$$(1-2d)\,[P(A)+P(B)-P(A\wedge B)-P(A\vee B)]+2d-2d=0$$

and so we expect that the average value for this identity will be 0 just as required in probability theory. Since individual values for this sum are perturbed by random noise, we expect these individual values to be distributed symmetrically around that mean of 0. The same prediction holds for identity 2. A number of experiments have shown that these identities do in fact hold in people's probability judgments: when we ask people to estimate probabilities for the terms in these identities for a range of events, and then combine each person's estimates according to the identity, the values obtained are distributed approximately symmetrically around a mean of 0, as required by probability theory and predicted by our model (Costello and Watts, 2014, Costello and Mathison, 2014, Fisher and Wolfe, 2014).

This model also predicts that people's probability estimates will violate the requirements of probability theory for identities 9 to 12 in Table 1, with the same degree of violation for each identity. Probability theory requires that these identities must also sum to 0 for all events A and B. Substituting our model's expression for the expected value for estimates of each term gives an overall positive expected value of d, violating the requirement of probability theory. For example, the estimated value of the expression in Identity 9 is

$$\langle P_E(A)\rangle+\langle P_E(B\wedge\neg A)\rangle-\langle P_E(A\vee B)\rangle=$$
$$(1-2d)\,[P(A)+P(B\wedge\neg A)-P(A\vee B)]+2d-d=d$$

Again, a number of experiments have shown that these identities are indeed violated in people's probability estimates, in just the way predicted by the model. These results, however, apply to only to unconditional or direct probabilities. In the next section we describe our more general model of conditional probabilities, and derive a similar set of results.

Estimating conditional probabilities

Just as above, we assume that people estimate $P(A|B)$ by drawing a random sample of instances of B, counting the number that are also A, and dividing by the sample size. As before, we assume some chance of random error d in this counting process. Given this random error there are two mutually exclusive ways in an item can be read as an instance of event B: (i) when the item truly is an instance of B and is read correctly (this occurs with probability $(1-d)P(B)$); and (ii) when the item is actually $\neg B$ but is read incorrectly as B (this occurs with probability $d(1-P(B))$).

We first take case (i). Given that a randomly sampled item is read as B, the probability that the item is truly an instance of B is

$$\frac{(1-d)P(B)}{(1-d)P(B)+d(1-P(B))} = \frac{(1-d)P(B)}{(1-2d)P(B)+d}$$

with the denominator here representing the probability that an item will be read as B, and the numerator the probability that such an item was read correctly.

Given that we truly have an instance of B, there are two mutually exclusive ways in which that item can be read as A: when the item is indeed an instance of A and is read correctly, or when the item is actually $\neg A$ and is read incorrectly as A. Since $P(A|B)$ is the probability of an item being truly an instance of A given that it is truly an instance of B, the first possibility occurs with probability $(1-d)P(A|B)$; since $1-P(A|B)$ is the probability of an item being truly $\neg A$ given that it is truly an instance of B, the second possibility occurs with probability $d(1-P(A|B))$. The sum of these two probabilities is $(1-2d)P(A|B)+d$, and so the overall probability of an instance being read as A given that it was read as B (and is truly B) is

$$\frac{(1-d)P(B)\left[(1-2d)P(A|B)+d\right]}{(1-2d)P(B)+d} \tag{2}$$

Taking case (ii) and reasoning in just the same we we get that the overall probability of an instance being read as A given that it was read as B (but is truly $\neg B$) is

$$\frac{d(1-P(B))\left[(1-2d)P(A|\neg B)+d\right]}{(1-2d)P(B)+d} \tag{3}$$

Since (i) and (ii) are mutually exclusive and cover all possibilities, the sum of Equations 2 and 3 gives our predicted value for $\langle P_E(A|B)\rangle$, the average estimate for the conditional probability $P(A|B)$. Adding and using the identities

$$\begin{aligned} P(B)P(A|B) &= P(A\wedge B) \\ (1-P(B))P(A|\neg B) &= P(A\wedge\neg B) = P(A)-P(A\wedge B) \end{aligned}$$

we get

$$\langle P_E(A|B)\rangle = \frac{(1-2d)^2P(A\wedge B)+d(1-2d)\left[P(A)+P(B)\right]+d^2}{(1-2d)P(B)+d} \tag{4}$$

Just as with Equation 1, this average is systematically biased away from the 'true' probability $P(A|B)$.

A direct probability $P(A)$ is, in probability theory, equivalent to a conditional probability $P(A|B)$ where the conditioning event B has a probability of 1. Rearrangement shows that when $P(B)=1$, Equation 4 reduces to Equation 1, our expression for direct probability estimation. Equation 4 thus completely describes all probability estimates, both direct and conditional, in this model. While Equation 4 appears complicated, it follows directly from two simple assumptions: that probabilities are estimated by counting event occurrence (in accordance with probability theory) and that this counting process is subject to random noise.

Predictions

From probability theory we have a number of identities whose value must be 0 for all events A and B. One such identity is Bayes' Rule (Identity 3 in Table 1). Our model predicts that this identity should also hold in people's probability judgments, on average. To see this, suppose we ask people to estimate $P(A), P(B), P(A|B)$ and $P(B|A)$ and for each person we take the products $P(A|B)P(A)$ and $P(B|A)P(A)$. Since estimates vary independently, the expected value of the products is equal to the product of the expected values of their constituents, giving

$$\begin{aligned} \langle P_E(A|B)P_E(B)\rangle &= \langle P_E(A|B)\rangle\langle P_E(B)\rangle \\ &= (1-2d)^2P(A\wedge B)+d(1-2d)[P(A)+P(B)]+d^2 \end{aligned}$$

and similarly

$$\begin{aligned} \langle P_E(B|A)P_E(A)\rangle &= \langle P_E(B|A)\rangle\langle P_E(A)\rangle \\ &= (1-2d)^2P(A\wedge B)+d(1-2d)[P(A)+P(B)]+d^2 \end{aligned}$$

and so

$$\langle P_E(A|B)P_E(B)\rangle - \langle P_E(B|A)P_E(A)\rangle = 0$$

Thus our model predicts that the average value of this identity, computed from people's individual probability judgments, should equal 0 as required by probability theory. Since deviations from this expected average in individual estimates are due to random error, we also expect that individual values for these identities will be approximately symmetrically distributed around 0.

Similar expansion and rearrangement gives the same result for Identities 4 through 8 in Table 1. Our model therefore predicts that these identities should all have an average value of 0 in people's estimates (matching the requirements of probability theory), and that individual values for these identities will be approximately symmetrically distributed around 0.

While this model predicts agreement with probability theory for the identities given above, it also predicts that Identities 13 through 16 in Table 1 should have a positive value in people's estimates, violating probability theory. For example, using the same substitutions as above we get an expected value for Identity 13 of

$$\begin{aligned} &\langle P_E(A\wedge B)-P_E(A|B)P_E(B)\rangle \\ &\quad= (1-2d)P(A\wedge B)+d-(1-2d)^2P(A\wedge B) \\ &\qquad -d(1-2d)[P(A)+P(B)]-d^2 \\ &\quad= d(1-d)-d(1-2d)\left[P(A)+P(B)-2P(A\wedge B)\right] \end{aligned}$$

Similar substitutions gives exactly the same expected value for identities 14, 15 and 16. Probability theory requires that $0\le P(A)+P(B)-2P(A\wedge B)\le 1$ for all A and B, and since $d<0.5$ by assumption, we see that values for this expression are distributed between d^2 and $d(1-d)$ and the expected value for Identities 13 through 16 will is at the centerpoint of

Table 2: A, B weather event pairs used in the experiment.

pair	A, B pairs in Block 1	pair	A, B pairs in Block 2
1	cold, rainy	6	cloudy, rainy
2	cloudy, icy	7	cold, icy
3	cold, thundery	8	cloudy, thundery
4	cloudy, warm	9	sunny, warm
5	sunny, snowy	10	icy, snowy

this range, which is $d/2$. Our prediction, therefore, is that Identities 13 through 16 should have, on average, a value of $d/2$; half the value of Identities 9 through 12.

An Experiment

We now describe an experiment testing the predictions of our model; in particular, those concerning the identities shown in Table 1. To test these predictions we gathered 62 participants' estimates for the 10 different constituent probability terms in the identities in the Table (i.e. $P(A)$, $P(A \wedge B)$, $P(A|B)$ and so on) for five different pairs of events. We combined each participant's individual estimates for each pair according to the given identities. Participants in Block 1 saw one set of five pairs of events and those in Block 2 saw a different set: we expected the predictions to hold for both blocks.

Materials

We constructed two sets of pairs of weather events, each containing five pairs; participants in Block 1 gave estimates for one set of pairs and those in Block 2 gave estimates for the second set. The two sets are shown in Table 2; the pairs were selected so that each set contained events of high, medium and low probabilities, and with varying conditional probability relationships between events.

Method

Participants were 62 undergraduate students at the School of Computer Science and Informatics, UCD, who volunteered to take part in exchange for partial course credit. Participants were asked to estimate the ten probabilities

$$P(A), P(B), P(A \wedge B), P(A \wedge \neg B), P(\neg A \wedge B), P(A \vee B),$$
$$P(A|B), P(B|A), P(A|\neg B), P(B|\neg A)$$

for each of the five pairs of weather events, giving 50 estimation questions for each participant. For single events, conjunctions, and disjunctions participants were asked

- What is the probability that the weather will be W on a randomly-selected day in Ireland?

where the weather event W could be, for example, 'cloudy', 'cold', 'cloudy and cold', 'cloudy and not cold' and so on. For conditionals, participants were asked

- If the weather in Ireland is W on a given randomly selected day, what is the probability that the weather will also be X on that same day?

Table 3: Average value (SD) for Identities in the Experiment, computed from participants' probability estimates in Blocks 1 and 2. Values for identities 1 to 8 are close to 0, while values 9 to 16 are significantly different from 0 in one-sample t-tests, as predicted by our model. Values for identities 13 to 16 were approximately half of those for identities 9 to 12, again as predicted by our model.

	Block		
Identity	1	2	predicted
1	0.00 (0.31)	-0.03 (0.26)	0
2	-0.01 (0.26)	-0.08 (0.30)	0
3	-0.01 (0.12)	0.00 (0.16)	0
4	-0.02 (0.20)	0.02 (0.20)	0
5	0.01 (0.19)	0.02 (0.19)	0
6	-0.01 (0.21)	0.02 (0.20)	0
7	-0.01 (0.16)	0.02 (0.11)	0
8	-0.02 (0.16)	0.00 (0.19)	0
9	0.24 (0.29)*	0.17 (0.26)*	d
10	0.25 (0.31)*	0.25 (0.31)*	d
11	0.25 (0.31)*	0.28 (0.32)*	d
12	0.24 (0.29)*	0.20 (0.28)*	d
13	0.14 (0.18)*	0.10 (0.18)*	$d/2$
14	0.13 (0.20)*	0.10 (0.20)*	$d/2$
15	0.12 (0.26)*	0.12 (0.27)*	$d/2$
16	0.13 (0.22)*	0.12 (0.22)*	$d/2$

** $p < 0.0005$, with Bonferroni correction for multiple comparisons

where W and X were the two single component events of the conditional $P(X|W)$. Participants gave their estimates on a 100-point scale, with the 0 point labelled 'will never happen' and the 100 point labelled 'certain to happen'. Questions were presented in random order on a web browser. The task took around half an hour to complete. Participants' responses these were divided by 100 prior to analysis.

Results

Two participants were excluded because they gave the same response for all questions, leaving 60 participants (31 in Block 1 and 29 in Block 2). As a consistency check we split participants in each block into two random groups and calculated the average probability estimate in each group for each one of the 50 presented probability terms. If participants were responding consistently we would expect there to be a reliable correlations between these split-half averages. Both blocks showed a high correlation between split-half averages ($r = 0.96$, $p < 0.0001$ and $r = 0.97$, $p < 0.0001$), indicating consistent responses.

Deviations from probability theory As predicted by our model, average values for identities 9 to 16 were positive for every A, B pair in both blocks, representing significant deviation from probability theory's requirement that these identites have a value of 0 (see Table 3). Average values for every one of these identities were significantly different from probabil-

ity theory's value of 0 in one-sample t-tests across individual values in both blocks ($p < 0.0005$ in all cases, with Bonferroni correction for multiple comparisons), just as predicted.

Recall that our model predicts that Identities 9 through 12 should all have the same average value, equal to d (the rate of random error for a given participant), and that Identities 13 through 16 should all have the same average value, equal to $d/2$. The average values in Table 3 support this prediction: values for Identities 9 through 12 were all close to their overall mean of 0.235 and values for Identities 13 through 16 were all around half that value (close to their overall mean of 0.12).

We would expect this chance of random error, d, to vary across participants, but to be relatively constant within a given participant. To test this prediction, for each participant we calculated the average value of Identities 9 through 12 from that participant's estimates and measured the correlation between participants' values for pairs of identities. All correlations were positive (average pairwise correlation of $r = 0.57$); of the six pairs, five showed a significant correlation at the $p < 0.001$ level (with Bonferroni correction for multiple comparisons), while correlation for the remaining pair was not significant ($p = 0.17$). Similarly, for each participant we calculated the average value of Identities 13 through 16 from that participant's estimates and measured the correlation between participants' values for pairs of identities. Again, all correlations were positive (average pairwise correlation of $r = 0.71$); all pairs showed a significant correlation at the $p < 0.001$ level (with Bonferroni correction for multiple comparisons).

Agreement with probability theory

We expected reliable agreement with probability theory for the identities 1 to 8. This expectation was also confirmed. For all identities 1 to 8, participants' responses had an average value very close to probability theory's required value of 0 in both Blocks 1 and 2. Averaging across all these identities gave a grand mean of $M = -0.006 (95\%CI = [-0.002, +0.015], SD = 0.22)$. Figure 1 graphs the frequency of occurrence of values for these identities. It is clear from the graph that values for these identities are symmetrically distributed around 0, the value predicted by our model. G_1 sample skewness for values for each identity in this graph were close to zero (all fell in the range ± 0.15), and the overall sample skewness across all identities was $G_1 = -0.01$, indicating symmetric distributions (Bulmer, 2012).

This pattern of close agreement with probability theory for identities 1 to 8 also held for each individual event pair A, B. There are 80 different averages for these identities in Table 3 (10 event pairs by 8 identities); of these 74 (92.5%) fell in the range $-0.1 \ldots + 0.1$, and 48 (60%) fell in the range $-0.05 \ldots + 0.05$. We analysed the distribution of these values for individual event pairs by carrying out 80 separate one-sample t-tests. Of these 80 tests,none were significantly different from 0 at the $p < 0.01$ level; with Bonferroni correction for multiple comparisons, just one t-test was marginally significant ($p = 0.04$). JZS Bayes Factor tests on overall values for identities (across all pairs) gave evidence in favour of the null hypothesis (that the value for the identity is 0) in all but one identity (and the overall value for that identity was still relatively close to 0). These results suggest that values for all these identities are distributed around 0, just as predicted by our model. Of course, the fact that some values for these identities were different from 0 means that there may be some other factor in play that is not accounted for in our model. However, even values that were significantly different from 0 were nevertheless still close to 0; this suggests that, even if there is some other such factor, the influence it has is small.

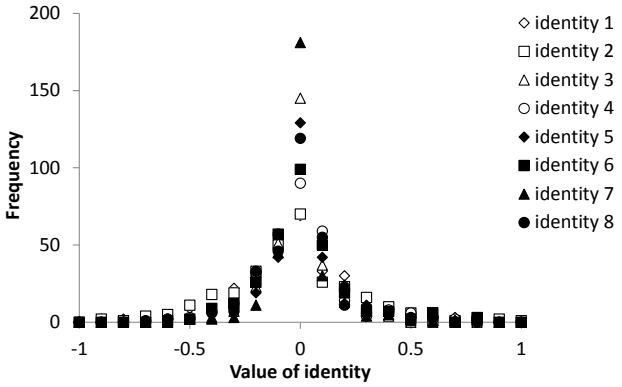


Figure 1: Frequency of occurrence of different values for Identities 1 through 8 in Experiment 1 across all A, B pairs in the experiment, grouped into 'bins' from $v - 0.05 \ldots v + 0.05$ for v from -1 to $+1$ in steps of 0.1. For example, since there were 60 participants in the experiment and each participant saw five pairs of events, the value of Identity 7 was calculated $5 \times 60 = 300$ times in total. Grouping these values into bins, we find that more than 60% of these calculations gave a value that fell in the $-0.05 \ldots + 0.05$ bin. Probability theory predicts these values will be symmetric around 0.

Discussion and Conclusions

We can summarise the main point of our work as follows: when deviations due to noise are cancelled out in people's probability judgments (as in Identities 1 through 8), those judgements are, on average, just as required by probability theory with no systematic bias. This pattern of agreement with probability theory holds for all the different event pairs A, B in our experiment. This agreement with probability theory cannot be dismissed by suggesting that our participants happened to be particularly good at probability estimation, because this agreement occurs alongside significant bias away from the requirements of probability theory for identities which do not cancel out the effects of random noise (Identities 9 through 16). For these identities the average degree of bias follows the predictions of our model (an approximately constant degree of bias d for Identities 9 through 12, and an approximately constant degree of bias $d/2$ for Identities 13

through 16). Taken together, the most natural explanation for these results seems to be that people estimate probabilities using a mechanism that is fundamentally rational (in line with frequentist probability theory), but is subject to the biasing effects of random noise.

While our results demonstrate that people's probability estimates follow probability theory (when bias due to noise is cancelled) we do not think people are consciously aware of the equations of probability theory when estimating probabilities. Indeed we doubt whether the participants in our experiment were aware of the probablity theory's requirement that thse identities should equal 0 or would be able to apply that requirement to their estimations. Instead we propose that people's probability judgments are derived from a 'black box' module of cognition that estimates the probability of an event A by retrieving (some analogue of) a count of instances of A from memory. Such a mechanism is necessarily subject to the requirements of set theory and therefore embodies the equations of probability theory.

We expect this probability module to be based on observed event frequencies, and to be unconscious, automatic, rapid, relatively undemanding of cognitive capacity and evolutionarily 'old'. Support for this view comes from that fact that people make probability judgments rapidly and typically do not have access to the reasons behind their estimations, from evidence that event frequencies are stored in memory by an automatic and unconscious encoding process (Hasher and Zacks, 1984), and from results showing that animals effectively judge probabilities (for instance, of obtaining food from a given source) and that their probabilities are typically close to optimal (Kheifets and Gallistel, 2012).

Our results have implications for current approaches to the psychology of people's probabilistic reasoning. In particular, our results are problematic for the view that people estimate probabilities via heuristics such as 'representativeness' (Tversky and Kahneman, 1983) or 'denominator neglect' (Reyna and Brainerd, 2008) that do 'do not appear to follow the calculus of chance or the statistical theory of prediction' (Kahneman and Tversky, 1973, p. 237). It seems to us that such heuristic accounts are motivated by the assumption that the observed biases and errors seen in people's probability judgments cannot be explained by probability theory. This motivation arises because probability theory is the normative model against which these biases and errors are assessed. If researchers had not taken those biases and errors as evidence that people don't reason using probability theory, they would have had no reason to propose those alternative accounts. However, our model suggests that these biases do not, in fact, count as evidence that people don't reason using probability theory. Those alternative models thus lose their fundamental motivation: there is no reason for moving from probability theory to those alternative accounts in an attempt to explain human probabilistic reasoning. There is, in contrast, an underlying motivation for the probability theory plus noise model: the probability of events in the world necessarily follow the rules of probability theory, and our reasoning processes are necessarily subject to noise.

Our results have broader implications for research on patterns of bias in aspects of people's decision-making. A common pattern in such research is to identify a systematic bias in people's responses, and to then take that bias as evidence that people are reasoning via some heuristic shortcut rather than the correct reasoning process. Our results, however, show that this inference from observed bias to inferred heuristic can be premature: random noise in reasoning can cause systematic biases in people's responses even when people are using normatively correct reasoning processes. To demonstrate conclusively that people are using heuristics, researchers must show that observed biases cannot be explained as the result of systematic effects caused by random noise.

References

Bulmer, M. G. (2012). *Principles of statistics*. Courier Corporation.

Costello, F. and Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, 121(3):463–480.

Costello, F. and Watts, P. (2016). Explaining high conjunction fallacy rates: the probability theory plus noise account. *Journal of Behavioral Decision Making*. In press, available at http://dx.doi.org/10.1002/bdm.1936.

Costello, F. J. and Mathison, T. (2014). On fallacies and normative reasoning: when people's judgements follow probability theory. In *Proceedings of the 36th annual meeting of the Cognitive Science Society*, pages 361–366.

Fisher, C. R. and Wolfe, C. R. (2014). Are people naïve probability theorists? A further examination of the probability theory + variation model. *Journal of Behavioral Decision Making*, 27(5):433–443.

Hasher, L. and Zacks, R. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *The American Psychologist*, 39(12):1372–1388.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4):237.

Kheifets, A. and Gallistel, C. R. (2012). Mice take calculated risks. *Proceedings of the National Academy of Sciences*, in press.

Reyna, V. F. and Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18(1):89–107.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315.

When does passive learning improve the effectiveness of active learning?

Kyle MacDonald
kyle.macdonald@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Much of what we learn comes from a mix of information that we select (active) and information that we receive (passive). But which type of training is better for different kinds of learning problems? Here, we explore this question by comparing different sequences of active/passive training in an abstract concept learning task. First, we replicate the active learning advantage from Markant & Gureckis (2014) (Experiments 1a and 1b). Then, we provide a test of whether experiencing active learning first or passive learning first improves the effectiveness of concept learning (Experiment 2). Across both experiments, active training led to better learning of the target concept, but "passive-first" learners were more accurate than "active-first" learners and more efficient than "active-only" learners. These findings broaden our understanding of when different sequences of active/passive learning are more effective, suggesting that for certain problems active exploration can be enhanced with prior passive experience.

Keywords: active learning, concept learning, replication

Introduction

Much of real-world learning occurs in contexts that contain *both* active and passive input. People rarely learn a new concept entirely from information generated by themselves (active learning[1]) or entirely from information received from the world (passive learning). And yet we do not have a theory about whether different sequences of active/passive input are better for different kinds of learning problems. Consider a teacher introducing a challenging math concept: should she allow students to explore first and then provide instruction, or should she teach first and then let students actively explore?

The potential benefits of active learning have been the focus of much research in education (Grabinger & Dunlap, 1995), machine learning (Settles, 2012), and cognitive science (Castro et al., 2009). In a review of this diverse literature, Gureckis & Markant (2012) suggest that active learning can be superior to passive learning because it allows people to use their prior experience and current hypotheses to select the most helpful examples (e.g., asking a question about something that is particularly confusing). But is active learning always better than passive learning?

Gureckis & Markant (2012) emphasize that the quality of active exploration is fundamentally linked to the the learner's understanding of the task: if the representation is poor, then self-directed learning will be biased and ineffective. The potential for bias in active learning suggests that receiving passive training first might be especially important for less constrained learning tasks where people are unlikely to generate examples that help them learn the target concept. For example, work by Klahr & Nigam (2004) shows that elementary school-aged children are less effective at discovering the principles of well-controlled experiments from their own self-directed learning. Moreover, Markant & Gureckis (2014) showed that active exploration provided no benefit over passive input when there was a mismatch between the target concept and learners' prior hypotheses. In both of these cases, receiving passive training first might have provided learners with a stronger task understanding that enhanced their active exploration. In fact, recent work by Thai (2015) has shown that receiving passive examples prior to performing active classifications can enhance perceptual classification learning.

In contrast, experiencing active learning first might be better for tasks with smaller problem spaces where helpful strategies can be extracted via active exploration and generalized to the passive learning context. Education research on the concept of *productive failure* shows that allowing students to struggle with a task (e.g., self-directed problem solving) can lead to better uptake of subsequent instruction (Westermann & Rummel, 2012). And work from cognitive science shows that completing a block of active learning prior to passive learning led to better word learning performance (Kachergis, Yu, & Shiffrin, 2013). Kachergis et al. (2013) suggest that active-first learners were able to develop attention and memory strategies during active training that generalized to the their passive learning experience.

In the current work, we aim to provide a direct test of how different sequences of active/passive training affect abstract category learning. We use a well-understood perceptual category learning paradigm from Markant & Gureckis (2014) since this task produces a reliable active learning advantage. We predicted that passive-first training would be better than active-first training for two reasons. First, we thought that people would generate hypotheses during passive learning, which they could then refine with highly informative examples during active exploration. Second, Markant & Gureckis (2014) reported that the quality of active learning was suboptimal early in their experiment, presumably because people were building a better understanding of the learning task. Thus, we predicted that providing passive learning first would allow people to maximize the selection of high quality evidence during active learning and improve their performance.

Experiment 1a

Experiment 1a is a direct replication of the active learning advantage for the Rule-Based (RB) category structure found in Markant & Gureckis (2014). In this task, people are asked to learn acategory boundary that is defined by a point on a single

[1]Here we focus on deliberate decisions about what to learn, as opposed to other uses of the term "active" learning (e.g., being engaged with learning materials).

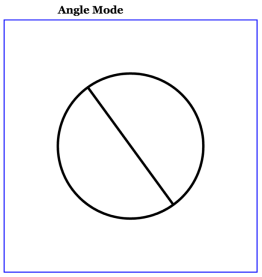


Figure 1: An example of a training trial from the active learning condition in Experiments 1 and 2. Participants could rotate the antenna or change the antenna's size before seeing which channel the antenna received.

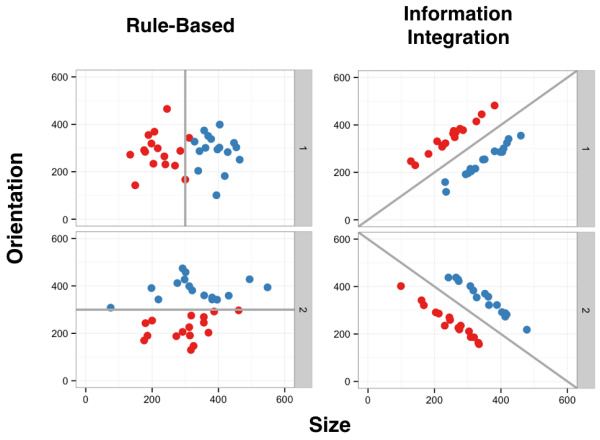


Figure 2: Example distributions of training stimuli shown to participants in the passive learning condition. Each point represents a different antenna constructed with an orientation value (vertical axis) and radius value (horizontal axis). The color of the points show the category membership of each antenna (red: channel 1; blue: channel 2). The solid line in each facet represents the optimal category boundary for both the Rule-Based category and the Information-Integration category structures.

continuous dimension (e.g., size). Active learners generate their own examples, whereas passive learners see data generated randomly from the category distributions. We used the same stimuli and followed the exact procedures as the original study with only minor differences, which we describe below. All of the stimuli, the experiments, and analysis code can be viewed and downloaded at the project page for this paper: `https://kemacdonald.github.io/Act-Learn/`.

Methods

Participants We planned a sample of 48 participants, 24 in each condition and posted a set of Human Intelligence Tasks (HITs) to Amazon Mechanical Turk. Only participants with US IP addresses and a task approval rate above 85% were allowed to participate, and each HIT paid one dollar. Data were excluded if participants completed the task more than once or if they reported to not understand the task at the end of the experiment (1 HITs). The final sample consisted of 52 participants (Passive: 27; Active: 25).

Stimuli Visual stimuli were black "antennas" on a white background (Fig 1). Each antenna could vary along two continuous dimensions – radius size or central angle – and was assigned a value between 1-600. To make sure that participants could not completely rotate the antenna, we limited the rotation to 150 degrees. The smallest radius and angle parameter values were randomized for each participant, so each participant had a unique category boundary. Finally, we used the Rule-Based category structure where the category boundary is defined by a point on a single dimension: either size or central angle.

Radius and angle values for the 96 passive training trials were generated from two Gaussian distributions with identical mean and covariance parameters as Markant & Gureckis (2014) (Fig 2). For test trials, we created a uniform grid of 192 unique test items that covered the entire parameter space. We randomly sampled 8 items from each quadrant to get 32 test trials for each block and randomized the order of training and test trials within each block and for each participant.

Design and procedure Participants saw a total of 288 trials (96 training and 192 test trials) across 6 blocks[2]. Each block consisted of 16 training and 32 test trials. Before the task, participants were told that they would see "loop antennas" for televisions and each antenna received one of two channels, and their goal was to learn the difference between the two types of antennas. We told participants that the antennas could sometimes pick up the wrong channel, and that they should learn what channel is most often received by a particular type of antenna.

After the instructions, participants were randomly assigned to one of the two between-subjects conditions (Active vs. Passive). In the Active condition, participants could design their own antennas. They modified the antenna by clicking and dragging the mouse from left to right. To change the size of the antenna, they first pressed the "Z" key. To change the angle, they first pressed the "X" key. When participants were finished with their design, they pressed the space bar to see which channel the antenna received. The channel label appeared in a text box with a green border located above the antenna.

In the Passive condition, participants saw antennas generated randomly using the size and angle values taken from the category distributions. After a two second delay they were told which channel the antenna received. The channel labels were deterministic. To help ensure that participants actually saw the channel, they had to click on the channel label in or-

[2]Six blocks is two blocks shorter than the original experiment. We reduced the length to increase our sample size.

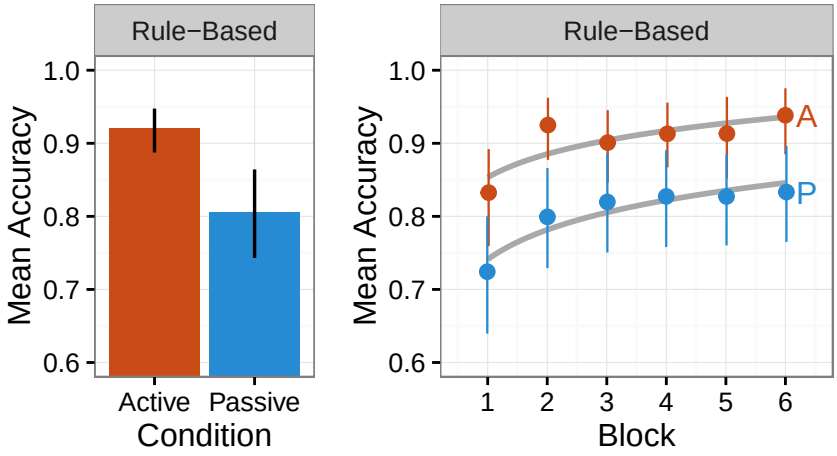


Figure 3: The left panel shows overall accuracy performance on the classification task for the Active and Passive training conditions from Experiment 1a. The right panel shows accuracy across each of the six blocks in the experiment. The grey curves are generated by a logarithmic smoother and the error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

der to advance the experiment. When they clicked, a green box appeared around the text to indicate that their response was recorded.

After training, participants proceeded to test trials. On each test trial participants saw one antenna and were asked, "Which channel does this antenna receive?" To indicate their response, participants selected one of two buttons (Ch1 or Ch2) located above the antenna. At the end of each block of test trials, participants saw a summary of their accuracy on the preceding block.

Results and Discussion

Classification accuracy First we report the planned replication analysis: a t-test to compare overall classification performance between the active and passive conditions. Active learners were more accurate, $t(49) = 2.98$, $p = 0.004$. Moreover, the effect size of the active learning advantage was large and greater than that of the original study (Original study: $d = 0.47$; Replication: $d = 0.73$).

Classification accuracy over time To quantify performance over time, we used a mixed effect logistic regression including a random effect for participants. We predicted test performance at the trial-level as a function of condition (Active/Passive) and block (1-6) and found a main effect of condition ($\beta = -0.68$, $p = 0.041$) with better performance for active learners overall, and a main effect of block ($\beta = 0.25$, $p < .001$) such that responses were more accurate later in the experiment. There was also an interaction between block and condition ($\beta = -0.1$, $p = 0.01$) with passive learners showing less of an increase in accuracy across the six test blocks.

Relations between quality of active learning and accuracy Another important result from Markant & Gureckis (2014) was the link between the quality of active learning and classification accuracy. Thus, we performed the same analysis on our replication data, quantifying the quality of evidence selection by computing the orthogonal distance between each sample and the true category boundary. Samples closer to the boundary are higher quality. We computed a mean accuracy score and a mean sample distance score for each participant, and fit a linear model using the mean sample distance score to predict accuracy. We found an effect of sample distance ($\beta = -.0004$, $p < .001$) with accuracy increasing as the sample distance decreased.

Experiment 1a provides strong evidence for a successful replication of the original results reported in Markant & Gureckis (2014). We found a comparable advantage in accuracy for active learners and we found the same relationship between quality of evidence selection and learning outcomes. Our results differ slightly from the original study in that we found an immediate advantage for active learners in the first block. We next attempt to replicate Markant & Gureckis (2014)'s findings for the more complex Information-Integration category structure and for the yoked passive learning condition.

Experiment 1b

The goals of Experiment 1b were: (a) to replicate the lack of an active learning advantage for the more difficult Information-Integration (II) category structure, where the category boundary was defined by a linear combination of an antenna's size and central angle, and (b) to replicate the finding that passive learners did not benefit from being "yoked" to active learners' data. We also performed an internal replication of the active learning advantage for the RB category that we found in Experiment 1a. We used the same stimuli and followed the exact procedures as the original study. However, we reduced the length of the experiment to two blocks in order to increase our sample size.

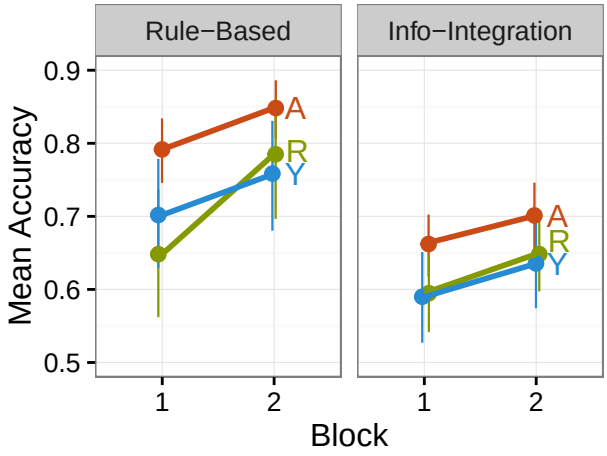


Figure 4: Accuracy across each of the two blocks in Experiment 1b for the Active, Passive, and Yoked training conditions for the Rule-Based (one dimension) and Information-Integration (two dimensions) category structures. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

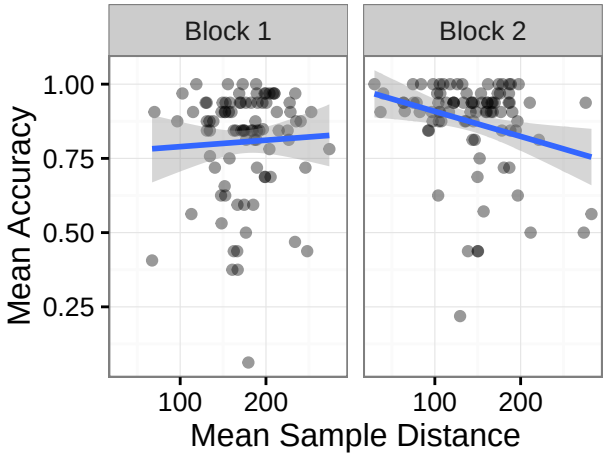


Figure 5: The relations between quality of evidence selection and accuracy on test trials across blocks for participants in Experiments 1a and 1b. Each point is an individual participant. Note that higher mean sample distance means worse evidence selection. The blue lines are linear model fits.

Methods

Stimuli Visual stimuli were identical to Experiment 1a.

Participants Participant recruitment and inclusionary/exclusionary criteria were identical to those of Experiment 1a (excluded 3 HITs). 196 HITs were posted across each of the between-subjects conditions: two category structures and three training conditions (A-RB: 66, A-II: 26, P-RB: 25, P-II: 28, Y-RB: 27, and Y-II: 24).

Design and procedure Procedures were identical to those of Experiment 1a except that we included a "yoked" learning condition where we matched each passive learning participant with training data generated by an active learner. Thus, the active and yoked participants saw the exact same data, but active learners were in control of the information flow. We also reduced the length of the experiment to two blocks.

Results and Discussion

Classification accuracy We fit the same logistic regression as specified in Experiment 1a and found a main effect of category (β = -0.82, $p < .001$) with better performance in the RB category. We also found a main effect of condition (β = -0.76, $p < .001$) such that participants in the passive and yoked conditions performed worse than participants in the active condition. We also found an interaction between block and the passive learning condition (β = 0.33, p = 0.03) with passive learners showing more learning over time, and an interaction between category type and block (β = -0.29, p = 0.04) with less learning in the II category.

Relations between evidence selections and accuracy over time Since the main goal of the current work was to test the effectiveness of different sequences of active/passive learning, we performed an exploratory analysis of the relations between sampling behavior and test accuracy over time for all active learners in Experiments 1a and 1b. We fit a linear model predicting each participant's mean accuracy based on the quality of their sampling behavior and experiment block. As expected, participants' accuracy improved in the second block (β = 0.21, p = 0.01). Interestingly, there was a reliable two-way interaction between sample distance and block (β = -0.0013, p = 0.005) such that the relationship between the quality of evidence selection and accuracy did not emerge until the second block (see Fig 5).

Experiment 1b provides additional evidence for a successful replication of the original results reported in Markant & Gureckis (2014). Importantly, yoked learners performed worse than active learners even though they had seen the same training data, suggesting that the link between people's hypotheses and the input is an important contributor to the active learning advantage. We did find that active learners performed better than both passive and yoked learners in the more complex II category structure. This advantage was not found in the original study, but this finding fits with recent work by Edmunds, Milton, & Wills (2015) suggesting that RB and II categories are not learned in qualitatively different ways. Finally, we found evidence that the link between the quality of active learning and accuracy emerges as people gain more experience with the task – an effect that we set out to explore in Experiment 2.

Experiment 2

Direct comparisons of active and passive learning are important for understanding when and why we might see advantages for one type of learning over the other. But real-world contexts are not neatly divided into active learning and passive learning. So when are different kinds of learning bet-

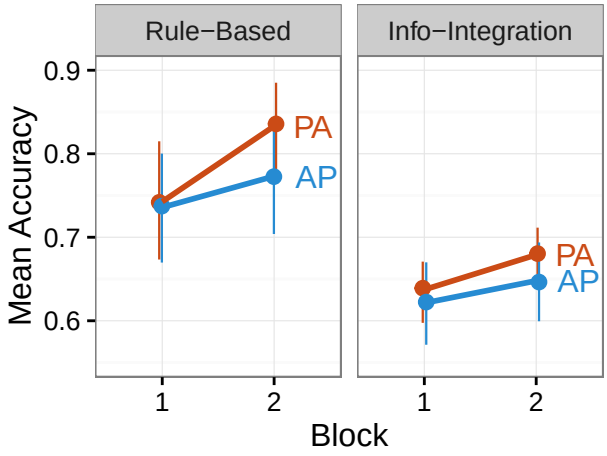


Figure 6: Accuracy across each of the two blocks in Experiment 2 for the Passive-Active (PA) and Active-Passive (AP) conditions for the Rule-Based (one dimension) and Information-Integration (two dimensions) category structures. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

ter for different kinds of learning problems? Experiment 2 provides a direct test of different predictions about the effects of different sequences of active/passive training in an abstract category learning task. Is active-first learning better because it makes people to engage more deeply with the task, enhancing their subsequent passive learning? Or is passive-first training better because it enhances later active learning?

Methods

Stimuli Stimuli were identical to Experiment 1.

Participants Participant recruitment and inclusionary/exclusionary criteria were identical to those of Experiment 1 (3 HITs). Approximately 44 HITs were posted for each training condition and category type for total of 176 paid HITs (AR-RB: 46, AR-II: 41, RA-RB: 38, RA-II: 51).

Design and procedure Procedures were identical to those of Experiment 1. Participants were randomly assigned to one the two between-subjects conditions: Active-Passive (AP) vs. Passive-Active (PA). In the AP condition, participants completed a block of active training and test trials before proceeding to a block of passive learning and test trials. In the PA condition, the order of the training/test blocks was reversed with passive training coming first.

Results and Discussion

First, we present analyses of classification accuracy and sampling behavior, focusing on just the AP and PA conditions. The key test of our hypothesis is the interaction between training condition and experiment block, since this tests the effect of training condition on improvement in classification accuracy. We then compare performance in the AP and PA conditions to performance from the Active-Active (AA) and Passive-Passive (PP) conditions from Experiments 1a and 1b, which allows for comparisons to training regimes where learners did not have to switch between types of learning.

Classification accuracy Which sequence of active and passive learning was better? We fit a logistic regression predicting test performance based on condition (PA vs. AP), block (1 vs. 2), and category type (RB vs. II). We found a main effect of block ($\beta = 0.63$, $p < .001$), with better performance in the second block across all conditions. For the key test of our hypothesis, we found a reliable two-way interaction between training condition and block ($\beta = -0.39$, $p < .001$), such that participants in the PA condition showed more learning over time compared to the AP condition. We also found a reliable two-way interaction between category type and block ($\beta = -0.44$, $p < .001$) and a marginally significant three-way interaction between condition, category type, and block ($\beta = 0.32$, $p = 0.08$). Participants showed less of an increase in accuracy for the more complex II category, and there was less of an advantage for the PA training condition in the II category.

Quality of evidence selection To test which condition produced higher quality sampling behavior, we fit a linear mixed model predicting sample quality as a function of condition and category type. We found that PA learners generated better samples than AP learners ($\beta = 22.4$, $t = 2.3$), and we found that sampling quality was lower in the II category ($\beta = 61.08$, $t = 6.46$). We also we fit a linear model predicting mean classification accuracy based on the mean distance of samples from the target category boundary, and replicate the finding from Experiments 1a and 1b: that higher quality sampling was related to higher accuracy scores ($\beta = .00006$, $p = 0.01$).

Classification accuracy for all sequences How does learning from different sequences of active/passive learning compare to learning from only active or passive data? We fit a logistic regression predicting test performance with the same specifications, but we included data from the active and passive learning conditions in Experiments 1a and 1b and coded user-defined contrasts to test classification accuracy for specific comparisons of interest. We found that receiving any active learning (AA, PA, AP) was better than receiving only passive learning (PP) ($\beta = 0.07$, $p < .01$), and that completing two blocks of active learning (AA) was marginally better than completing one block of active learning (PA or AP) ($\beta = 0.06$, $p = 0.09$).

Time costs associated with different learning sequences We were also interested in the time costs associated with different sequences of active and passive learning. We fit a linear mixed effects model with the same specifications but predicting the amount of time spent on each training trial. We found that training took longer in any active learning condition compared to passive learning ($\beta = 2.59$, $t = 11.57$). We also found that people who completed only active training (AA) took longer than people who only completed one block of active learning (PA, AP) ($\beta = -0.71$, $t = -2.22$). Surpris-

ingly, we found that people in the active-first condition (AP) took longer than people in the passive-first condition (PA) ($\beta = 2.67$, $t = 3.64$). We did not find any reliable interactions.

The main findings from Experiment 2 are that despite experiencing the exact same amount of active and passive data, passive-first learners were more accurate overall, produced higher quality sampling behavior, and spent less time on training. These results suggest that receiving passive training first improved learners' active learning by helping them to explore more efficiently.

General Discussion

Mixtures of active and passive learning are a fundamental feature of real-world learning contexts. But we do not know when different sequences of active/passive input are better for different kinds of learning problems. In the current work, we take a first step towards answering this important question. We found that that receiving passive learning first improved subsequent active exploration and led to better overall performance in an abstract concept learning task. This finding broadens our understanding of the types of learning problems where passive learning might support effective active learning.

Why was passive-first learning better in this kind of abstract concept learning problem? One possibility is passive training allowed people to generate hypotheses about the correct category boundary, which they could begin testing from the very first trial of active learning. We did find that passive-first learners produced higher quality active learning, suggesting that perhaps active-first learners used their initial active learning less effectively. In contrast, we did not see active learning strategies transfer to passive learning, as had been found previously (Kachergis et al., 2013). Perhaps we did not see transfer because this task involved incremental changes to hypotheses, which depend on generating the right example to falsify the learner's current hypothesis on a trial by trial basis. In contrast, if active learning helps learners develop better attentional/memory strategies, these strategies can often be transferred to passive contexts.

It is interesting that two blocks of active learning was slightly better than the passive-first learning. But, passive-first learning provides an additional benefit: it reduces learning costs (i.e., time and mental effort). We found that active learners took longer to complete the task and spent more time on the training blocks. Thus, even if passive-first learning does not achieve the same level of performance as active learning, it still might be preferable.

There are several limitations to our experiment. First, while the learning task we used is well-understood, like other tasks of this type it dramatically simplifies real-world concept learning. So we do not know how our sequencing finding would scale to learning more complex, higher dimensional concepts. It could be that passive learning becomes even more important as the concept becomes more complex. Second, we used a coarse sequence manipulation, at the level of training/test blocks. And sequences of active/passive learning might have differential effects on learning at finer levels of granularity (e.g., trial level).

We need a theory of what kinds of learning work best for the diverse set of problems that learners must solve. Our work here provides an important first step towards understanding when passive learning experiences could be used to support better active exploration. Overall, these results show that, for some tasks, passive learning can equip people with a better task representation, making them more effective active learners.

Acknowledgements

We are grateful to Doug Markant and Todd Gureckis for sharing the details and code from the original experiment. We thank the members of the Language and Cognition Lab for their helpful feedback on this project. This work was supported by a National Science Foundation Graduate Research Fellowship to KM.

References

Castro, R. M., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2009). Human active learning. In *Advances in neural information processing systems* (pp. 241–248).

Edmunds, C., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures. *The Quarterly Journal of Experimental Psychology*, *68*(6), 1203–1222.

Grabinger, R. S., & Dunlap, J. C. (1995). Rich environments for active learning: A definition. *Research in Learning Technology*, *3*(2).

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*(1), 200–213.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114.

Thai, K.-P. (2015). Making experts: Optimizing perceptual learning in complex, real-world learning domains.

Westermann, K., & Rummel, N. (2012). Delaying instruction: Evidence from a study in a university relearning setting. *Instructional Science*, *40*(4), 673–689.

Balancing Structural and Temporal Constraints in Multitasking Contexts

Dario D. Salvucci (salvucci@drexel.edu)
Department of Computer Science, Drexel University
3141 Chestnut St., Philadelphia, PA 19104, USA

Tuomo Kujala (tuomo.kujala@jyu.fi)
Department of Computer Science and Information Systems, University of Jyväskylä
P.O. Box 35, FI-40014 Jyväskylä, Finland

Abstract

Recent research has shown that when people multitask, both the subtask structure and the temporal constraints of the component tasks strongly influence people's task-switching behavior. In this paper, we propose an integrated theoretical account and associated computational model that aims to quantify how people balance structural and temporal constraints in everyday multitasking. We validate the theory using data from an empirical study in which drivers performed a visual-search task while navigating a driving environment. Through examination of illustrative protocols from the model and human drivers as well as the overall fit on the aggregate glance data, we explore the implications of the theory and model for time-critical multitasking domains.

Keywords: Multitasking; driving; cognitive architectures

Introduction

Multitasking often occurs in time-critical situations, such as answering a ringing phone while babysitting, cooking over a stove, or driving a vehicle. In these situations, the structure of one or both tasks may impose a sense of urgency to complete a task, perhaps due to the environment (e.g., a pot boiling over) or perhaps due to self-imposed pressures. For instance, consider Janssen, Brumby, and Garnett's (2012) example of texting the message "Running late" while driving: even as driving is clearly the task with the highest priority, a driver who is almost done typing the message ("Running lat...") will feel strongly compelled to finish typing before continuing to drive. In such situations, we continually balance our urgency to complete one task with the urgency imposed by other tasks.

A wealth of recent research has explored how we multitask, both in constrained laboratory studies and in complex real-world environments. One general finding is that task structure—how a task breaks down into smaller subtasks—strongly affects how people perform that task concurrently with other tasks (e.g., Borst, Taatgen, & van Rijn, 2015; Iqbal & Bailey, 2005). Complementary studies have shown that a task's temporal constraints can strongly affect multitasking; arguably the most studied context is that of driving, for which studies have explored the relationship between driving urgency (or uncertainty) and time looking away from the road (e.g., Kujala et al., 2015; Lee, Gibson, & Lee, 2015).

In this paper, we examine the critical relationship between structural and temporal constraints on multitask behavior and performance. Although there are multitasking scenarios in which either the structural or the temporal constraints are dominant, they are both generally present in some form: structural constraints (e.g., the chunking of a telephone number: Brumby, Howes, & Salvucci, 2009; Janssen, Brumby, & Garnett, 2012) still appear in very time-critical domains like driving, and temporal constraints (e.g., the limited time for answering a ringing phone) still appear in primarily structurally-driven domains. Although most studies have focused on only one of these constraints at a time, recent studies have examined how overall task priorities affect task switching (e.g., Brumby, Howes, & Salvucci, 2009; Janssen, Brumby, & Garnett, 2012) and glance durations between tasks (Lee, Gibson, & Lee, 2015). As yet, however, a rigorous cognitive process model that quantifies this relationship has proven elusive. Here, we propose a theoretical framework to help understand and quantify the balance between subtask structure and temporal constraints.

Balancing Structure and Time

The issues that arise in balancing structural and temporal constraints are perhaps best illustrated by behavioral protocols collected from people acting in multitasking contexts. For this purpose, we delve more deeply into a recently collected data set in which drivers (N=12) were asked to perform visual-search tasks while driving (Kujala & Salvucci, 2015). In the experiment, drivers searched through multiple screens of 6, 9, or 12 songs for a particular target song. The screens were laid out in one of two ways: a *Grid* layout, with 2, 3, or 4 rows of 3 songs each; or a *List* layout, with the songs listed vertically top to bottom. If the target song did not appear on a given screen, the driver was asked to press the down-arrow button to advance to the next screen. Here we are concerned only with screens without the target, in which drivers searched through all the songs. Drivers performed the search tasks in a driving simulator, on a display to the right of the steering wheel, and navigated a straight three-lane road at highway speed (80 km/h), occasionally performing the search task when requested.

The visual-search and driving task includes the types of structural and temporal constraints found in many

multitasking contexts. The search task comprises two subtasks that repeat for each screen, namely to ***search*** through the on-screen items and to ***press*** the button when finished. At the same time, the driving task involves increasing urgency over time as the driver looks away from the road, eventually reaching the point where the driver needs to look back. Thus, the balance between completing the search task and driving safely created the key challenge for the driver in managing both tasks concurrently.

The original treatment of these data (Kujala & Salvucci, 2015) focused on aggregate analysis and modeling of human behavior in this task. In focusing on individual behavior protocols in the data, however, we found that the aggregate analysis shrouded the interesting cognitive strategies that arose on individual trials. To this end, we now focus on examining the human protocols from the study, and in the next section develop a much more in-depth model that matches both individual and aggregate behavior.

Switching at Subtask Boundaries. As mentioned earlier, many studies have shown that people tend to switch tasks at subtask boundaries. Participants in the present study were no exception, and often switched from ***search*** to driving and from ***press*** to driving after each of these subtask boundaries. Figure 1 shows a classic example of this type of switching, taken from a person performing the search task on the Grid-6 display. Specifically, the figure shows the timeline of screen glances and button presses as the person searches through 3 consecutive screens for the target. Throughout the timeline, we see an alternation between glances for each subtask: a first glance for ***search*** and a second for ***press***. The emerging pattern is one of switching at subtask boundaries, with the total number of glances equal to the total number of subtasks completed (2 glances for each of 3 screens).

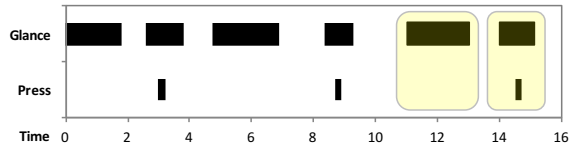


Figure 1: Human protocol showing task switching at subtask boundaries (highlighted: switch after ***search***, switch after ***press***).

Self-Interrupting during Subtasks. Although switching at subtask boundaries is commonly observed in the study protocols, there are at least two other common behaviors. One such behavior involves interrupting oneself (or "giving up": Bogunovich & Salvucci, 2011) during a subtask, when a person decides during a subtask that s/he needs to switch immediately rather than complete the subtask. Figure 2 illustrates this behavior in the timeline of glances and presses for a person doing 3 trials in the Grid-12 condition. Here, for each trial, the person divides the ***search*** subtask into three roughly equal components, and finally makes a fourth shorter glance to press the button.

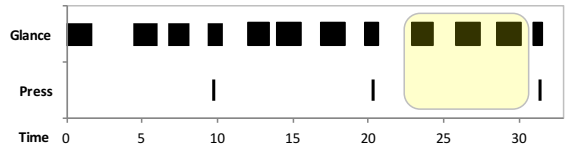


Figure 2: Human protocol showing self-interruption during subtask (highlighted: multiple glances during ***search***).

Continuing beyond Subtask Boundaries. Besides self-interrupting before subtask boundaries, people also exhibit the behavior of continuing beyond subtask boundaries—that is, reaching the subtask boundary, but then continuing to the next subtask rather than switching away from the task. Figure 3 shows the sample protocol for one example in the List-12 condition (see highlighted area): the driver presses the button and then immediately continues to search the next screen. Unfortunately, for some of these cases, continuing to the next subtask results in a long off-road glance lasting over 2 seconds, illustrating a perhaps unintuitive effect: shorter tasks and faster behavior may sometimes actually lead to longer glance times at a task, because it tempts a person into continuing with the next subtask.

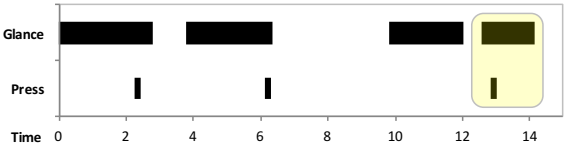


Figure 3: Human protocol showing subtask continuation (highlighted: ***press*** followed by ***search*** in same glance).

A Theory and Computational Model

The main goal of our work here is to better understand the interplay between subtask boundaries and temporal urgency illustrated by the above examples. In particular, we aim to develop a computational model that can run in simulation and thus produce behaviors directly comparable to those of human participants. In this section we describe the details of our theoretical account and computational model, to be validated with human data in the next section.

An individual task can generally be thought of as a hierarchy of higher- and lower-level tasks (or goals) and subtasks (or subgoals) (see, e.g., Schraagen, Chipman, & Shalin, 2000). To account for the interaction of structural and temporal constraints, we require that the hierarchical decomposition also specify the timing of the various components. The simplest way to achieve this goal is to assign times to the actions at the leaves of the hierarchy tree; one might assume constant times (e.g., taken from the keystroke-level model: Card, Moran, & Newell, 1980), or variable times based on aspects of the current environment (e.g., mouse movement over different distances, cognitive delays due to recalling information). Taken further, the task hierarchy could be instantiated as a running computational model that adapts continuously to its environment—for example, a model developed using a computational cognitive architecture (e.g., Anderson, 2007; Newell, 1990). For our

purposes here, a task hierarchy augmented with action times is sufficient to serve as a model of an individual task.

Given two task models, we would like to express how structural and temporal constraints are balanced to dictate how people switch tasks in a multitasking context. As a first step toward this goal, we define *urgency* as a measure of a person's desire to work on a given task; in essence, each task in a multitasking context has an associated urgency, and generally speaking, people tend to switch to (or continue with) the most urgent task at a given time. Urgency provides a convenient way to formulate the effects of structure and time into a single measure, and a way to evaluate concurrent tasks and decide whether and when to switch between them.

Urgency and Structural Constraints. Earlier we noted how task structure has a strong influence on multitasking; how can we formulate this influence in terms of urgency? Empirical studies have made clear that people value the completion of a subtask, and thus, a person's urgency should increase as s/he approaches a subtask boundary. If we assume a strong association between urgency and time, we could say that a person receives a "reward" upon completion of each subtask (e.g., Fu & Anderson, 2006), and that the reward is equal to the time spent on that subtask.

The value of receiving a reward at the completion of a subtask may also be propagated back to the actions that led to successful completion. One method allows prior actions to receive a boost to their urgency, but with a discount factor γ between 0 and 1 that reduces the reward with increasing temporal distance from the actual subtask completion. Consider a subtask S^i comprised of actions a_k^i with times (durations) t_k^i. The total reward $R(S^i)$ for this subtask is computed as the sum of its action times:

$$R(S^i) = \sum_k t_k^i$$

We then compute the structural urgency $U_s(a_k^i)$ of a particular action a_k^i as a function of the subtask reward, $R(S^i)$, and the remaining time between the end of the action and the completion of the subtask, $T(a_k^i)$:

$$T(a_k^i) = \sum_{j>k} t_j^i$$

$$U_s(a_k^i) = \gamma^{T(a_k^i)} \cdot R(S^i)$$

The final action receives the full reward ($T(a_k^i) = 0$), and each action before the final action receives the reward discounted by γ and the remaining time to completion. For example, Figure 4 graphs the structural urgency for 10 actions of 300 ms with different values of γ. This produces a hook-like function with larger γ values producing a flatter urgency function (earlier actions receiving more reward) and smaller values producing a sharper curve.

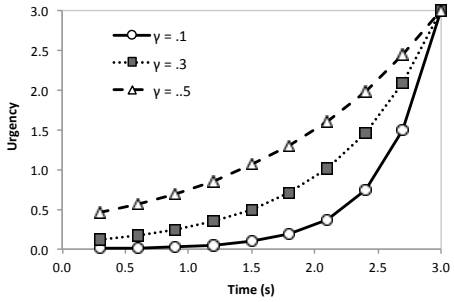


Figure 4: Sample structural urgency profiles for 10 actions each with a duration of 300 ms.

Structural urgency as defined thus far accounts for people's preferences in switching at subtask boundaries. A complementary empirical finding is that as people complete one subtask in a multitasking scenario, they are generally averse to continuing to the next subtask unless they feel they have sufficient time to complete that one as well (Bogunovich & Salvucci, 2011). This finding suggests that people have an awareness of the time needed to complete a full subtask, and that they use this information in deciding whether or not to continue. In terms of urgency, the aversion to continuing to another subtask can be represented as a negative urgency at the start of the subtask. Specifically, we define a continuation penalty when continuing from one subtask to the next, whereby we subtract the full duration of the next subtask from the structural urgency. In the context of a multitasking scenario, continuing to the next subtask will generally have a lower urgency than switching to another task; however, if no other task can proceed and/or other tasks have even lower urgency, the next subtask may then proceed.

Urgency and Temporal Constraints. In addition to the urgency contributed by its structure, a task will often have an associated temporal urgency—a feeling that compels a person to complete the task as soon as possible. Temporal urgency is influenced by the amount of time passed since switching away from the task, with urgency (typically) increasing with the passage of time. We define temporal urgency $U_t(\Delta t)$ as a function of the time since switching away Δt. The specific form of this function depends heavily on the task domain: highly time-critical domains will have a steep function with urgency rapidly increasing over time, whereas less time-critical domains will have flatter functions. In the next section we will see a concrete example of such a function for a time-critical task domain.

Deciding when to Switch Tasks. In the case of multiple concurrent tasks, we use a decision mechanism similar to the conflict resolution mechanism in ACT-R (Anderson, 2007) to determine which task will progress at a given time. First, for each task, the total urgency U is computed as the sum of its structural urgency, temporal urgency, and noise factor ϵ:

$$U(a_k^i, \Delta t) = U_s(a_k^i) + U_t(\Delta t) + \epsilon$$

As in ACT-R, the noise ϵ is sampled from a logistic distribution, with mean $\mu = 0$ and scale s, a free parameter

to be estimated (described shortly). Then, the task with the larger urgency is allowed to proceed.

Computational Simulation. We implemented the proposed framework as a Java simulation system to enable rapid testing of models and parameters settings. The system takes as input a model as described above (with subtasks and associated action times), and generates sequential behavioral protocols as output. The protocols can then be analyzed for more aggregate measures, such as the common measures of glance times to be used shortly in the forthcoming study.

Study: Visual Search and Driving

To test the proposed approach, we return to the domain discussed earlier, namely visual search and driving. The prior model in Kujala & Salvucci (2015) focused mainly on the temporal constraints of the driving task, dynamically adjusting an in-car glance duration threshold according to the stability of the vehicle after switching back to the driving task. While this prior model provides a good account of the aggregate data, it does not conform well to the individual protocols shown earlier—notably, because it does not account for the structural constraints of the visual-search task (e.g., it is equally likely to switch early or late in the visual search, whereas people show a tendency to avoid switching late in the task). Here, we use the theoretical framework above to model the two tasks of visual search and driving, and then illustrate how the theory and simulations produce behavior that better resembles that of human participants.

Models of Visual Search and Driving

As discussed earlier, the visual-search task breaks down into a repeated iteration of two basic subtasks: a ***search*** of each of the on-screen items; and (assuming the target is not found, which is always the case for the screens analyzed here) a ***press*** of the down-arrow button to advance to the next screen. The ***search*** subtask includes an encoding action for each of the on-screen items—that is, 6, 9, or 12 actions to match the items in that particular condition. The ***press*** subtask contains a single action to press the button. The time required for each action was derived from simulations of the earlier ACT-R model of this task (Kujala & Salvucci, 2015): 368 ms per item for ***search*** in the Grid layout; 291 ms per item for ***search*** in the List layout; and 564 ms for ***press*** in all cases.

The model of driving used here is derived from the ACT-R model of driver behavior (Salvucci, 2006). The core subtask is a cycle that visually encodes the near and far points of the road and updates the vehicle controls accordingly. These actions require a total of 200 ms to complete the cycle, and thus this is also the duration of the driving subtask, which is simply repeated while the model is actively driving. However, because the focus of our analysis here is on behavior in the visual-search task, we simply assume that the model drives for 1 second (5 cycles) and then switches back to the visual-search task.

Beyond the above structural details, we also require some formalization of the temporal constraints of the driving task in particular. There have been several attempts to quantify a driver's cognitive state while looking away from the road, most notably in terms of uncertainty (see Kujala et al., 2015): as time passes, the driver's uncertainty about the external environment (lane position, speed, other vehicles, etc.) gradually increases until it reaches a point at which the driver feels compelled to look back to the road. We translate these ideas into the temporal urgency of the driving task as follows. When the driver has stabilized the driving task, the urgency of further driving updates is rather low. We define a value U_{stable} to denote the low urgency of driving in this condition, a value analogous to the uncertainty threshold in prior work (e.g., Kujala et al., 2015). This value is presumed to be negative to indicate a lack of urgency—that is, it indicates that time might better be spent on some other task. In all, we define the temporal urgency of driving as:

$$U_t(\Delta t) = U_{stable} + \Delta t$$

When the driver looks away from the road ($\Delta t = 0$), the temporal urgency is equal to U_{stable}; however, as time passes and the driver continues the secondary task, the urgency of driving climbs steadily, eventually passing zero and becoming positive if the driver does not switch back to driving within U_{stable} seconds.

When we combine these models of visual search and driving, we can visualize their competing urgencies as a function of time, as illustrated in Figure 5. The urgency of the search task builds gradually because of the increasing urgency to finish the task, ramping up quickly toward the end of the subtask. Meanwhile, the urgency of driving starts low (at the assumed U_{stable} level) but increases over time due to increasing levels of uncertainty. At each point, the two urgencies are compared using the noisy conflict resolution process described earlier, forcing a switch to driving if the urgency of driving exceeds that of the search task. The graph on the right shows the probability of switching to driving at the various times: highest in the middle of the search subtask, and lowest early in the process (because driving still has a very low urgency) and late in the process (because there is high urgency to complete the search subtask). The resulting probability distribution is thus an emergent property of the theoretical mechanisms.

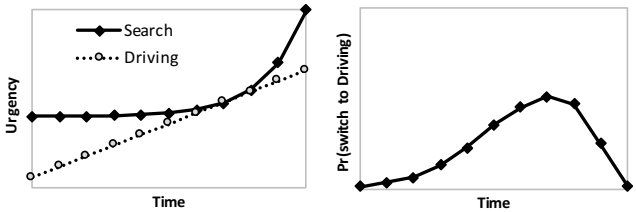


Figure 5: Sample urgency graph for visual search and driving, with associated probability of switching to driving.

For these models, we estimated the three free parameters (the urgency value U_{stable}, the scaling factor γ, and the noise scale s) by running 1000 simulations per parameter-value combination and finding the values that produced the best fit

on the aggregate data described later. The estimated values were $U_{stable} = -1.7$, $\gamma = 0.1$, and $s = 0.3$.

Model Behavior and Results

Kujala & Salvucci (2015) examined five separate measures (30 data points total) of aggregate behavior by the human participants. Because of space constraints and because our focus lies primarily in the individual protocols, we avoid a detailed comparison of these aggregate measures here. However, it should be noted that for these five measures, the overall fit of the current model was very much on par with that of the previous model. Table 1 includes the correlation (R) and normalized root-mean-squared-error ($RMSE/mean$) for both models for all measures.

Table 1: Model-human correlations and errors for prior model (Kujala & Salvucci, 2015) and current model.

	Prior Model		*Current Model*	
	R	*NRMSE*	*R*	*NRMSE*
Number of in-car glances	.99	.32	.96	.20
Total in-car glance duration	.97	.08	.99	.05
Mean in-car glance duration	.81	.13	.62	.15
Maximum in car glance duration	.94	.30	.83	.05
Percent glances over 2 seconds	.65	.31	.62	.31

While the current model matches aggregate behavior as well as the prior model, the current model importantly provides a much better account of the behavior of individual participants and trials by accounting for both temporal and structural constraints. Figure 6 shows one such behavior for the model in the Grid-6 condition, namely the classic strategy of switching at subtask boundaries. The upper portion of the graph shows a timeline of the model's glances and button presses—again, analogous to our earlier analysis of human data. The lower portion shows the competing urgency between search and driving over time. For the first three screens, including behavior for the section screen at the point labeled **A1**, the model begins the ***search*** subtask; the urgency of driving steadily grows from its starting U_{stable} value, but the urgency of completing the ***search*** subtask grows as well. When ***search*** is done, the urgency to continue with the next subtask (***press***) includes the continuation penalty defined earlier, namely subtracting the duration of the next subtask; in essence, the model is checking whether there is sufficient time to complete the next subtask, and if not, it switches back to driving. At the next opportunity, though, the model completes the pressing subtask and switches back. At the point labeled **A2**, the model switches slightly earlier but then completes the search as well as the button press on the next glance.

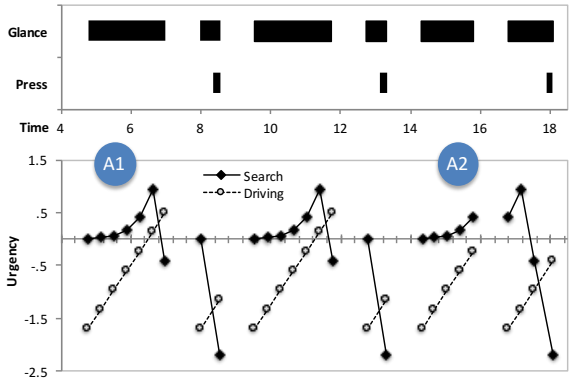


Figure 6: Model protocol showing task switching at subtask boundaries (after ***search*** and ***press***).

Not surprisingly, self-interruption during subtasks becomes more common as the number of on-screen items increases. Figure 7 shows an example in the Grid-9 condition. At point **B1**, we see a typical behavior in which the model reads several items during one glance, several more in a second glance, then finishes reading and finally makes the button press on the third glance. At point **B2**, the model splits up the item reading differently, but the end result is still a total of three glances to complete the ***search*** and ***press***, instead of two glances in the canonical behavior in Figure 6. The behaviors for the other screens show similar patterns; note that because of the noisy conflict resolution mechanism, a lower urgency can sometimes "win" over a higher one, producing a similarly large variety of protocols as for human drivers.

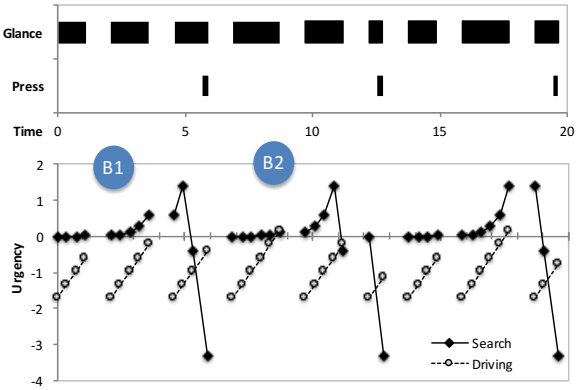


Figure 7: Model protocol showing self-interruption during subtask (multiple glances during ***search***).

As the number of on-screen items decreases, or the duration of individual actions decreases (from Grid to List), the model adapts by occasionally continuing beyond subtask boundaries, as shown in Figure 8 in the List-6 condition. At point **C1**, the model finishes searching the 6 on-screen items so quickly that the urgency for the next subtask, ***press***, is very close to that of driving; in this case (with noise), the model continues and presses the button before switching back to driving. Point **C2** illustrates a different form of continuation: after pressing the button, the urgency of driving is still very low, and again the model decides to continue and begin

searching the on-screen items. The end result for this segment of behavior is a total of 4 glances to complete 3 screens, instead of the 6 glances (2 per screen) that would result from switching at subtask boundaries. While continuation after pressing the button was observed as a relatively infrequent behavior for human and model alike, the presence of this behavior at all (again, in both human and model) emphasizes the flexible nature of the balance between structural and temporal constraints.

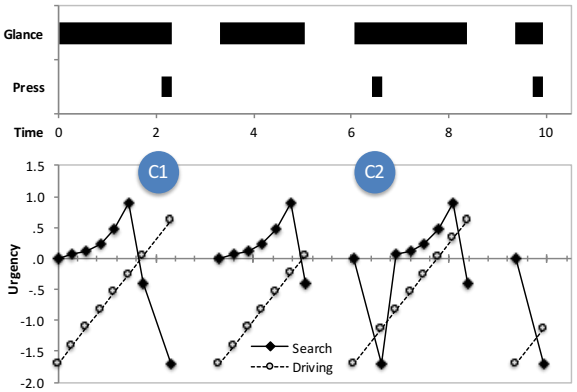


Figure 8: Model protocol showing subtask continuation (**C1**: ***press → search***; **C2**: ***search → press***).

General Discussion

The complex relationship between structural and temporal constraints presents a fascinating challenge when examining everyday multitasking behaviors, especially those in time-critical contexts. The concept of urgency developed here offers a way to unify these two important factors on multitasking, both in understanding the human behaviors that emerge, and in formalizing rigorous computational models to predict behavior in novel situations. One might consider urgency as related to task priorities that influence behavior through rational adaptation (e.g., Howes, Lewis, Vera, 2009). Empirical work along these lines have focused on manipulating the overall priority of each task (e.g., instructing participants to focus on one task or the other: Janssen, Brumby, & Garnett, 2012). Our treatment here is complementary in focusing on the rise and fall of urgency at a second-by-second level, being closely dependent on the lower-level conditions of each task. Urgency thus helps to formalize how people get "hooked on" subtasks, and how they balance structural urgency of subtasks with the temporal urgency of time-critical task domains.

As a step in this direction, the formulation of urgency has potential for incorporation into larger theories of cognition. For example, the ACT-R cognitive architecture (Anderson, 2007) posits that behavioral rules have an associated utility that can be learned and adapted using reinforcement mechanisms similar to those here (Fu & Anderson, 2006). However, whereas each rule has only one utility, a particular instantiation of a rule can have different urgency values depending on its place in the subtask structure. Urgency is more akin to threaded cognition's (Salvucci & Taatgen, 2011) "least-recently-used" heuristic in choosing the next cognitive thread to run; the heuristic might be subsumed by an improved formulation of urgency as a dynamic property of a complex task.

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: OUP.

Bogunovich, P., & Salvucci, D. D. (2011). The effects of time constraints on user behavior for deferrable interruptions. In *Proc. CHI '11*, 3123-3126.

Borst, J. P, Taatgen, N. A., & van Rijn, H. (2015). What makes interruptions disruptive? A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proc. CHI '15*, 2971-2980.

Brumby, D. P., Howes, A., & Salvucci, D. D. (2009). Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving. In *Proc. CHI '09*, 1629-1638.

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Comm. of the ACM*, *23*, 396-410.

Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*, 184-206.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review*, *116*, 717-751.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *Proc. CHI '05*, 1489-1492.

Janssen, C. P., Brumby, D. P., & Garnett, R. (2012). Natural break points: The influence of priorities and cognitive and motor cues on dual-task interleaving. *Journal of Cognitive Engineering & Decision Making*, *6*, 5-29.

Kujala, T., Mäkelä, J., Kotilainen, I., & Tokkonen, T. (2016). The attentional demand of automobile driving revisited: Occlusion distance as a function of task-relevant event density in realistic driving scenarios. *Human Factors*, *58*, 163-180.

Kujala, T., & Salvucci, D. D. (2015). Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. *International Journal of Human-Computer Studies*, *79*, 66-78.

Lee, J. Y., Gibson, M., & Lee, J. D. (2015). Secondary task boundaries influence drivers' glance durations. In *Proc. AutomotiveUI '15*, 273-280.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors*, *48*, 362-380.

Salvucci, D. D., & Taatgen, N. A. (2011). *The Multitasking Mind*. New York: OUP.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (2000). *Cognitive Task Analysis*. New York: Psychology Press.

Do additional features help or harm during category learning? An exploration of the curse of dimensionality in human learners

Wai Keen Vong (waikeen.vong@adelaide.edu.au)
Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Daniel J. Navarro (dan.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

How does the number of features impact category learning? One view suggests that additional features creates a "curse of dimensionality" - where having more features causes the size of the search space to grow so quickly that discovering good classification rules becomes increasingly challenging. The opposing view suggests that additional features provide a wealth of additional information which learners should be able to use to improve their classification performance. Previous research exploring this issue appears to have produced conflicting results: some find that learning improves with additional features (Hoffman & Murphy, 2006) while others find that it does not (Minda & Smith, 2001; Edgell et al., 1996). Here we investigate the possibility that category structure may explain this apparent discrepancy – that more features are useful in categories with family resemblance structure, but are not (and may even be harmful) in more rule-based categories. We find while the impact of having many features does indeed depend on category structure, the results can be explained by a single unified model: one that attends to a single feature on any given trial and uses information learned from that particular feature to make classification judgments.

Keywords: Category learning; supervised learning; curse of dimensionality

Introduction

Category learning can become increasingly difficult as the number of object features increases. This "curse of dimensionality" occurs because the learner must in some way search over the large number of features in order to determine how to weight the importance of each during classification (Sutton & Barto, 1998). Despite this difficulty, people – even small children – easily learn natural categories composed of objects with a very large number of features (Rosch, 1973). How do people overcome the curse of dimensionality when they learn high-dimensional categories such as these? In this paper we present simulations and empirical results that show that susceptibility to the curse depends on what is being learned: whether the categories involved follow a family-resemblance structure or are more rule-based.

Category learning experiments have traditionally avoided the curse of dimensionality by using stimuli that consist of only a few highly salient features, generally between two and four (e.g., Medin & Schaffer, 1978; Minda & Smith, 2001; Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961). Although these experiments have substantially contributed to our understanding of category learning, it remains largely an open question how learning changes (if at all) when there are a large number of features. The few studies that have investigated this empirically have yielded conflicting results. Increasing the number of features has been variously found to impair learning (Edgell et al., 1996), to facilitate learning (Hoffman & Murphy, 2006), or to not impact learning at all (Minda & Smith, 2001).

How can we resolve this apparent discrepancy? One possibility is that the studies differ in the kinds of categories being learned. After all, the curse of dimensionality stems from having so many possible stimuli configurations in a high dimensional space that it is difficult to figure out which features are the most important ones. This should lead to the greatest inefficiency when most of the possible features are not diagnostic of category membership and only one or a few matter, as in Edgell et al. (1996). By contrast, if all features are diagnostic to some degree – especially if they are not perfectly correlated with each other – then additional features should be beneficial, or at least not hurtful (Hoffman & Murphy, 2006; Minda & Smith, 2001).

This reasoning is sensible, but no studies to date have tested it by manipulating category structure and number of features while holding other factors constant. The goal of the current paper is to do this. Our results suggest that people's ability to evade the curse of dimensionality in natural categories occurs because of the family resemblance structure of natural categories – but that in rule-based categories the curse defeats us. We also show that although people's performance qualitatively varies depending on the nature of the categories to be learned, it can be accounted for by a single unified model with limited attentional abilities.

Experiment

Our experimental design involves systematically manipulating the number of features and the category structure in a simple categorization task. We were interested in how performance changed with increasing numbers of features, and how this depended on the nature of the categories being learned (family resemblance, intermediate or rule-based). As predicted, learning decreased when there were additional features when category structure was rule-based, but did not when it was more of a family resemblance structure.

Method

Participants 442 participants (238 male) were recruited via Amazon Mechanical Turk. Participants ranged in age 19 to 76 (mean 34.2). They were paid US$2.00 for completion of the experiment, which took roughly 12 minutes to complete. 14 participants failed to complete the task, and 5 participants had participated in a pilot version of this study; these data were excluded from further analyses.

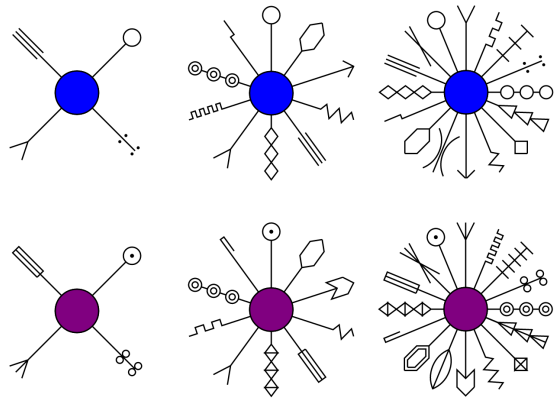


Figure 1: Six example stimuli, displaying two examples from each of the three possible *Dimensionality* conditions (left: 4, middle: 10, right: 16). Features were binary and correspond to the legs of the amoebas. Together, the two examples from the 16-FEATURE condition show all possible feature values.

Design The task was a supervised category learning experiment in which people learned to classify amoeba as either bivimias or lorifens. Each amoeba consisted of a circular base with a set of distinct binary features (legs). The full set of 16 unique pairs of features are shown on the two stimuli in the right column of Figure 1.

The nine experimental conditions were created by manipulating two factors (the *Dimensionality* of the stimuli and the category *Structure*), each with three levels, in a between-participant design. The three levels of *Dimensionality* reflect the number of binary features present on the stimuli (4-FEATURE, 10-FEATURE, or 16-FEATURE conditions). For the lower-dimensionality conditions, the set of displayed features chosen were randomly selected from a subset of the features used in the 16-FEATURE condition. The position of features on the amoeba and the mapping from feature values to category labels were randomized differently for each participant.

The three category *Structures* defined the relationship between feature values and category labels. For all three structures, one dimension was 90% predictive of the correct category label (meaning that on 90% of trials, categorizing according to that dimension would lead to a correct label). The different category structures were defined by the diagnosticity of the other features in the category. In one condition, the other features were 50% predictive of the category label;[1] we call this the RULE condition because maximum performance can be achieved by finding the one highly-diagnostic feature and ignoring all of the rest. In another condition, all of the features were 90% predictive of the category label (though all were generated independently, so none were perfectly correlated with each other). We called this the FAMILY condition because this imposes a family resemblance structure with many highly coherent and predictive features. Finally, in the INTERMEDIATE condition the other features were 70% diagnostic: thus, one feature was most diagnostic but it would be theoretically possible to achieve better performance by using all of the features in concert.

Procedure The experiment consisted of five blocks of 20 learning trials each, for a total of 100 trials. On each trial people were presented with an amoeba[2] and were asked to classify it as either a bivimia or a lorifen. They received points both for answering correctly and quickly. To make it more game-like each trial was associated with a countdown bar that decreased in size over time. After responding, feedback was given by displaying the true category label for three seconds. In addition, the circular base of the amoeba lit up with the appropriate category colour (blue for bivimias and purple for lorifens). There was a one second delay after the feedback before the presentation of the next stimulus. At the end of each block, participants were presented with a short summary of their performance across each completed block.

Results

Before addressing the main question of how learning is influenced by *Dimensionality* and *Structure*, it important to verify that learning in fact took place. As Figure 2 illustrates, participants in all nine conditions showed evidence of learning. We evaluate this quantitatively using a Bayesian mixed effects model with block as a continuous variable, and *Dimensionality* and *Structure* as discrete variables.[3] Across all conditions, accuracy increased during training: the model that included block was strongly preferred over a model that only included a random effect for each participant ($BF > 10^{44} : 1$).

How did the number of features and the structure of the categories affect learning? As is evident from Figure 2, learning was fastest in the FAMILY condition, slowest in the RULE condition, and intermediate in the INTERMEDIATE condition; the corresponding main effect of *Structure* yielded a Bayes factor of more than 700:1 in favor of a difference. It is also evident that performance was affected by the number of features, with the main effect of *Dimensionality* yielding $BF > 47 : 1$.

The main effects are sensible, but the main prediction was that we expected the effect of *Dimensionality* to be different for different *Structures*. Was such an interaction observed? Figure 2 suggests there was one, with performance decreasing with additional features in the INTERMEDIATE and RULE conditions, but not in the FAMILY conditions. Supporting this, a Bayesian mixed effects model containing block, structure, dimensionality and the interaction between structure and dimensionality was strongly preferred over a model without the interaction term ($BF > 10^3 : 1$). Overall, these results suggest that high dimensionality is only a curse as categories grow more rule-based; if they are not, the high informativeness of every feature renders the search problem less of an issue.

[1] Since there were two categories, this means they were not predictive at all.

[2] The stimuli for each person was generated randomly according to the appropriate category structure, rather than pre-generating 100 specific stimuli and showing the same ones to everybody.

[3] All mixed effects models in this paper assume a random intercept for each subject. Bayes factors were calculated using the BayesFactor package 0.9.12-2 (Morey & Rouder, 2015) in R 3.2.3. Because it is typical to obtain a range of possible factors within a confidence interval, for simplicity we report the approximate factor.

	Category Structure		
Number of Features	Family	Intermediate	Rule
4 Features	4 features all 90% predictive	1 feature 90% predictive 3 features 70% predictive	1 feature 90% predictive 3 features 50% predictive
10 Features	10 features all 90% predictive	1 feature 90% predictive 9 features 70% predictive	1 feature 90% predictive 9 features 50% predictive
16 Features	16 features all 90% predictive	1 feature 90% predictive 15 features 70% predictive	1 feature 90% predictive 15 features 50% predictive

Table 1: The nine different conditions tested in the experiment. For all of the Intermediate and Rule conditions where only 1 feature was 90% predictive, this 1 feature was chosen at random.

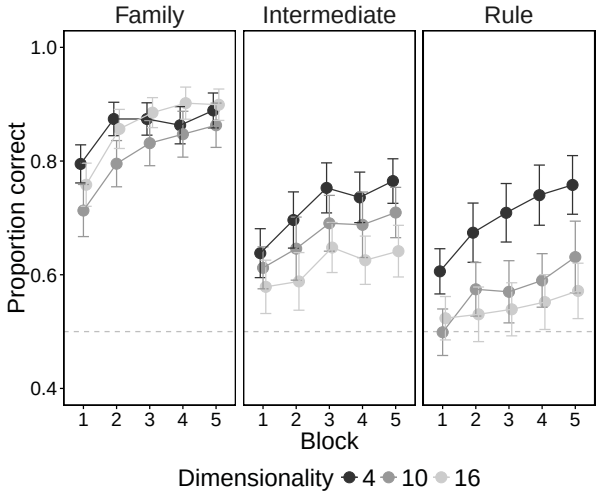


Figure 2: Mean human performance across the three *Dimensionality* and *Structure* conditions. While learning within the FAMILY-resemblance categories was unaffected by the number of features, more features meant that learning was poorer in the RULE-based and INTERMEDIATE categories. Error bars reflect 95% confidence intervals, and the dotted line reflects chance performance.

Two models of human performance

The intuitive reasoning motivating this experiment was based on the insight that if people approach category learning by searching among all possible features, then the curse of dimensionality should hurt performance only when one or few features are useful (as in rule-based categories) but not if most or all of them are (as in family-resemblance categories). The empirical results support our predictions, but another important test is whether a search-based model qualitatively reproduces human performance while a model that uses all of the information from all features does not. In this section we implement such a test by modeling people's behavior with two different learning models. Both models were simulated on nine conditions that exactly paralleled the conditions in the experiment, with three levels of dimensionality (4, 10, and 16) and the same three category structures (FAMILY, INTERMEDIATE and RULE-based).

The structure and notation of the learning environment is identical for both models. The input for each trial is a D-dimensional stimuli vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, where D is the dimensionality of the stimulus and each x_i is a binary feature ($x_i \in \{0,1\}$). The predicted category response ($\hat{y} \in \{0,1\}$) for the nth trial is defined by the feature information from the nth trial along with the representation learned by the model based on the previous $N-1$ trials. The two learning models we consider differ according to the representation they learn from experience.

Naive Bayes

The first model we consider is the Naive Bayes classifier, which uses information about every feature to determine category predictions. The model tracks the diagnosticity of each feature ($p(x_i|y)$) across all previous trials to compute an estimated probability of each category label (y) for a given stimulus ($\mathbf{x}$). The model assumes each of these features are independently diagnostic of the category label and combined as in Equation 1. The predicted category response ($\hat{y}$) of the model on each trial is the category with the highest estimated probability (Equation 2).

$$p(y|\mathbf{x}) \propto \prod_{i=1}^{D} p(x_i|y)p(y) \quad (1)$$

$$\hat{y} = \arg\max_{y} p(y|\mathbf{x}) \quad (2)$$

Hypothesis Testing

The second model is a Hypothesis Testing model which assumes that categories are defined by a single binary feature. On each trial, the model maintains a single hypothesis h_{ia} that consists of a simple rule for determining the predicted category response. All rules in the hypothesis space share the same format: `if` $x_i = a$ `then` $\hat{y} = 0$, `otherwise` $\hat{y} = 1$. The space of hypotheses is defined by a, indicating a particular feature value ($a \in 0,1$), and the feature x_i where: ($i \in \{1,\ldots,D\}$). As an example, a particular hypothesis the model might use is: "If the third feature dimension is 0 (i.e. $x_3 = 0$), then respond $\hat{y} = 0$, otherwise respond $\hat{y} = 1$."

The probability of staying with the current hypothesis after each trial is given by the utility u of the current hypothesis. The utility is proportional to prediction accuracy for previous trials on which it was the current hypothesis; it is equal to 1 for those hypotheses that have never been the current hypothesis (Equation 3). On trials in which the current hypothesis is discarded, a new hypothesis is selected from the set of all possible hypotheses. New hypotheses are selected in proportion to their utility, as in Equation 4.

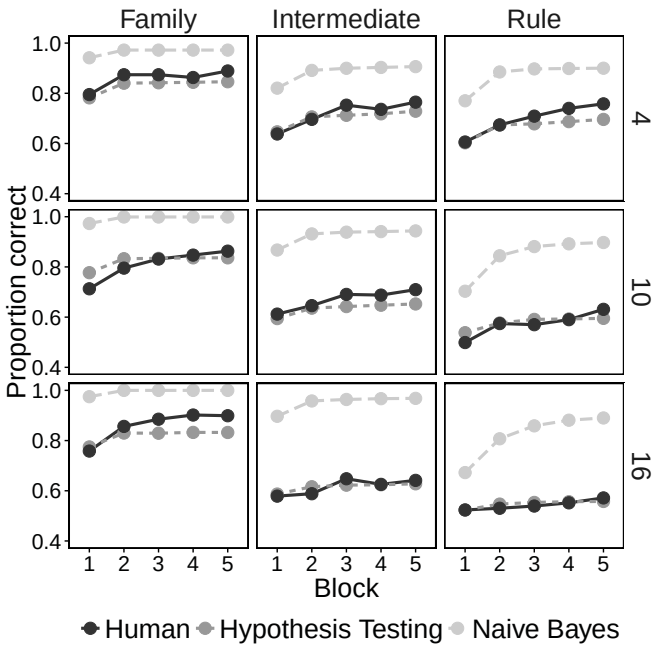


Figure 3: Predicted performance from the Hypothesis Testing and Naive Bayes models across the nine conditions. Each data point is the average of 10,000 simulations. Naive Bayes systematically overestimates performance, while the Hypothesis Testing model provides a much better fit. However, it fails to capture more subtle aspects of human performance, like the gradual learning curves.

$$u(h_{ia}) = \frac{1 + (\text{correct predictions with } h_{ia})}{1 + (\text{trials with } h_{ia})} \quad (3)$$

$$p(h_{ia}) = \frac{u(h_{ia})}{\sum_{x,y} u(h_{xy})} \quad (4)$$

Simulations

Both models were run 10,000 times in each of the experimental conditions. Each simulation mimicked one 100-trial experiment. On each trial a new stimulus was generated as in the experiment, the model made predictions, feedback was provided, and the models were updated.

Figure 3 shows the average prediction accuracy of the two models and the human data, broken down into subplots for each of the nine learning environments. The most striking finding is that Naive Bayes systematically overestimates human performance, especially for rule-based categories. This qualitative effect is mirrored in the root-mean-squared error of each model with the Naive Bayes model producing much larger overall error (0.226) than the Hypothesis Testing model (0.033). That said, both models capture many of the qualitative patterns in the human data, including the advantage for learning the FAMILY category structure relative to the RULE structure. In the FAMILY condition, both find very little effect of increasing feature dimensionality because all of the features are equally very predictive. A small positive effect of additional features does exist for Naive Bayes, because when all information is used, information from enough additional features can improve performance even above the 90% possible from any single one.

The models have similar qualitative outcomes in the RULE condition, doing worse with additional features. In the Hypothesis Testing model this occurs because of the increased difficulty in finding the most useful feature. Naive Bayes shows a very small performance decrement with more features because it does not set the feature weights of the unpredictive features to precisely zero; this effect, however, is tiny and also diminishes with time.

That said, the models make qualitatively different predictions in the INTERMEDIATE category structure: Naive Bayes predicts that additional features should improve prediction accuracy while the Hypothesis Testing model predicts the opposite. This qualitative difference emerges because of the utility of the less-predictive features in each model. The Naive Bayes model combines the information from the additional less-predictive features with the more-predictive feature to make judgments in the INTERMEDIATE condition. As a result, it makes better judgments where there are more features. By contrast, because the Hypothesis Testing model only uses one feature, it does not improve category prediction by adding additional feature information. In fact, as the number of less-useful features grows, the less likely it is for the model to switch to the hypothesis containing the most predictive feature; its performance therefore worsens.

Overall, the Hypothesis Testing model captures human performance much better than Naive Bayes, especially as the structure of the categories become more rule-based. However, the Hypothesis Testing model fails to capture some of the more subtle qualitative effects found in the human data. Most interestingly, the Hypothesis Testing model shows a sharp increase in prediction accuracy after the first block of training but does not continue to improve prediction accuracy beyond the second block. This results in a systematic underestimation of prediction accuracy in the final block across all conditions. These patterns suggest that people might be using information about more than one feature when making decisions or shifting between rules. In the following section we introduce a new model to try to account for these effects.

A hybrid model of category learning

Both the Hypothesis Testing and Naive Bayes models fail to capture all of the qualitative trends in human performance. In this section we propose a hybrid model framework that combines elements from both previous models. Like the Naive Bayes model it represents categories by assigning each feature a diagnosticity value and assumes that features are independent, but it learns in a much more limited way: on any given trial it updates the diagnosticity only for a single feature. The mechanism in the Hybrid model for determining which feature weight to update follows the same switching rule as the Hypothesis Testing model (Equation 4).

We also consider two variants of the model, corresponding to two different ways to incorporate feature information when making decisions about category membership. Both versions

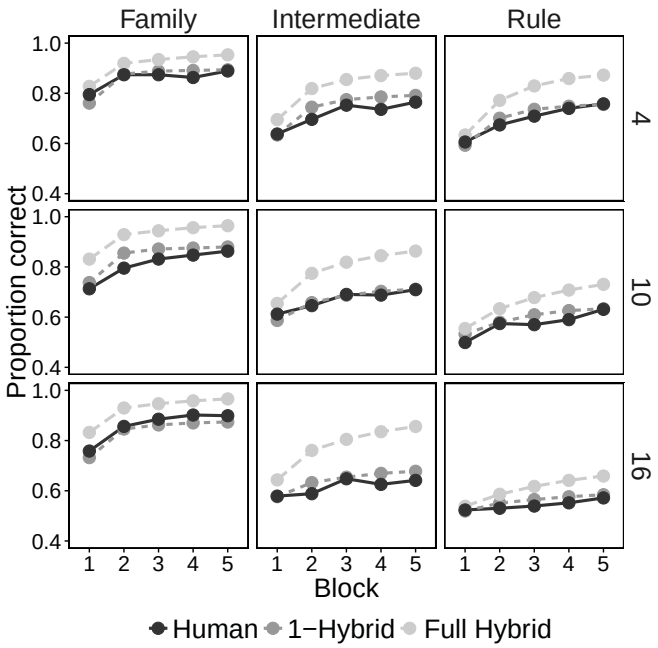


Figure 4: Predicted performance of the 1-Hybrid and Full-Hybrid models, compared against the human data across the nine conditions. Each data point is the average of 10,000 simulations. The 1-Hybrid model captures performance better than all of the other models considered.

implement the same decision rule as the Naive Bayes model (Equation 2), but differ in terms of how many features they include. The Full Hybrid model, like Naive Bayes, incorporates information from all of the features; by contrast, the 1-Hybrid model incorporates information only from the current feature when determining the category assignment.[4]

As Figure 4 shows, both Hybrid models produce learning curves whose shape closely matches that of human performance. However, the 1-Hybrid model captures the overall level and magnitude of performance much better than the Full Hybrid model; indeed, of all of the models we considered, it has the tightest quantitative fits to human data (RMS = 0.026, vs 0.033 for Hypothesis Testing and 0.107 for Full Hybrid). The poor performance of both the Full Hybrid and Naive Bayes models suggests that human learners probably do not make decisions based on combining information from all features. The improved performance of 1-Hybrid over the pure Hypothesis Testing model, however, suggests that people *do* maintain information about the diagnosticity of all features, even if only one feature is used to make category judgments at any given time.

Discussion

This paper demonstrates that the effect of additional features on category learning depends a great deal on the category structure, as reflected by the diagnosticity of the additional features. We found that more features hurt learning when the additional features were less predictive than the best ones, as in rule-based or intermediate categories. This effect was captured best by models that attend to single features for learning and prediction, rather than models that attended to or updated all features at once. These constraints are consistent with known limitations on working memory and attention in humans (e.g., Atkinson & Shiffrin, 1968).

The critical role of category structure and feature diagnosticity may explain the existing disagreements in the literature concerning the impact of feature dimensionality on category learning. Edgell et al. (1996) used a category structure in which additional features were not diagnostic of category membership and found that increasing the number of feature dimensions decreases category learning accuracy. We replicated this effect in our rule-based category structure, and both the Hypothesis Testing and 1-Hybrid model captured it. They do so because in them, increasing the number of less-useful dimensions increases the chance of switching to a less useful hypothesis and thus making a poor prediction.

In contrast, Hoffman and Murphy (2006), who used a category structure in which the new features were predictive of category membership, found that increasing the number of feature dimensions actually improved accuracy. Finally, Minda and Smith (2001) found no effect of number of features in a similar category structure. Performance in our family resemblance condition replicated the results of Minda and Smith (2001), showing no improvement in learning (but also no decrement) with additional features in a family resemblance structure.[5] This is captured by the Hypothesis Testing and 1-Hybrid models because learning about any feature is equally useful, so switching does not hurt performance.

The 1-Hybrid model accounts for the empirical data better than the other three models. The Naive Bayes and Full Hybrid models use a decision rule that produces performance that is much more accurate than human performance across all conditions, suggesting the humans do not make decisions based on information from all of the features. (Other augmentations of the models are possible as well, e.g., using a probabilistic choice rule instead of a maximizing strategy, but it is unlikely that this would change this qualitative aspect of performance). The 1-Hybrid model slightly outperforms the Hypothesis Testing model because it produces accuracy curves that continue to improve over time, while the Hypothesis Testing model underestimates improvement after the second block. This gradual improvement throughout training seems to be due to an advantage from maintaining a representations of the diagnosticity of previous feature hypotheses. In future work we will investigate this issue more precisely.

Another interesting future direction of research is to compare the 1-Hybrid model to other learning mechanisms that have been proposed to address the curse of dimensionality. These methods have focused on reducing the number of dimensions via manifold learning (Tenenbaum, 1998) or structured inference (Kemp & Tenenbaum, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Lake, Salakhutdinov,

[4] We also considered versions that included information from two to four features, but do not include these results for space reasons.

[5] Our results probably did not replicate Hoffman and Murphy (2006), because, like Minda and Smith (2001) but not it, our additional features were not perfectly correlated with existing features.

& Tenenbaum, 2015), rather than preserving the true dimensionality of the stimuli but limiting the learning mechanism by focusing on a reduced set of features on each trial. We can compare our models to such methods in the kinds of learning environments explored in this paper, as well as those that they have already been shown to successfully account for. It is, of course, possible that human learning is versatile enough to incorporate the fundamental insights from both types of models, and apply each appropriately where it is called for. This is all a matter for future work.

Overall, this research suggests that the "curse of dimensionality" negatively impacts category learning mainly in environments in which a single (or a few) features are predictive of the category, but there are many features that are not. Environments that contain many features, in which all or most of them are diagnostic of category membership, do not appear to harm performance. People's behavior can be explained by a computational model that attends to and updates a single feature at a time, shifting between features based on diagnosticity; by contrast, models that integrate information from many features or models that do not learn feature weights at all do more poorly. These results suggest that in the real world, people may be able to overcome the "curse of dimensionality" not because we are optimal learners, but rather because the structure of most natural categories is more similar to family resemblance structures in which most features are predictive of category membership.

Acknowledgments

AP was supported by grants from the Australian Research Council (DP110104949 and DP150103280, with salary support from DE12010378). The salary of AH was supported by DP110104949.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The Psychology of Learning and Motivation*, *2*, 89–195.

Edgell, S. E., Castellan Jr, N. J., Roe, R. M., Barnes, J. M., Ng, P. C., Bright, R. D., & Ford, L. A. (1996). Irrelevant information in probabilistic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1463.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 301.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=BayesFactor` (R package version 0.9.12-2)

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Tenenbaum, J. B. (1998). Mapping a manifold of perceptual observations. *Advances in Neural Information Processing Systems*, 682–688.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Desirable difficulties in the development of active inquiry skills

George Kachergis, Marjorie Rhodes, & Todd Gureckis
{george.kachergis, marjorie.rhodes, todd.gureckis}@nyu.edu
Department of Psychology, New York University
New York, NY

Abstract

This study explores developmental changes in the ability to ask informative questions. We hypothesized an intrinsic link between the ability to update beliefs in light of evidence and the ability to ask informative questions. Four- to ten-year-old children played an iPad game asking them to identify a hidden bug. Learners could either ask about individual bugs, or make a series of feature queries (e.g., "Does the hidden bug have antenna?") that could more efficiently narrow the hypothesis space. Critically the task display either helped children integrate evidence with the hypothesis space or required them to perform this operation themselves. Although we found that helping children update their beliefs improved some aspects of their active inquiry behavior, children required to update their own beliefs asked questions that were more context-sensitive and thus informative. The results show how making a task more difficult may actually improve children's active inquiry skills, thus illustrating a type of desirable difficulty.

Keywords: question asking, information search, active inquiry, hypothesis testing, scientific reasoning

Introduction

A central aim of science education is to teach students how to approach the task of understanding their environment. Rather than teaching only a catalogue of facts about the biological and physical worlds, current standards emphasize teaching the conceptual and analytic skills that underlie science: detecting patterns in environments that initially appear chaotic, abstracting the general principles that can be used to understand and predict events, and importantly, learning how to ask informative questions to reveal these patterns and principles when they are not immediately obvious (Bransford, Brown, & Cocking, 2000; Donovan & Bransford, 2005; Duschl, Schweingruber, & Shouse, 2007).

Many of the cognitive skills required for active scientific inquiry follow protracted developmental trajectories. For example, in tasks designed to assess scientific reasoning abilities, children in the older elementary school years (ages 8-10) often have difficulty adopting systematic strategies, such as testing the effects of one variable at a time or selecting interventions that will lead to determinate evidence (Chen & Klahr, 1999). Although children in the older elementary school years can be taught to engage in these strategies via direct instruction (Klahr & Nigam, 2004; Kuhn & Dean, 2005), it is notable how difficult it is for them to discover and implement them on their own.

One reason for the difficulties children show may be that active inquiry depends on the coordination of a variety of component cognitive processes (belief updating, decision making, hypothesis generation, etc.). Inefficiencies in any or all of these interrelated processes may serve as developmental limitations. For example, young learners may be able to search efficiently for information given a particular set of hypotheses but have trouble updating their beliefs correctly given new evidence. In this sense active inquiry is like a bicycle: when all the elements are properly functioning and aligned the bike moves forward. However, misalignment of any one component can be catastrophic.

The present study joins with some recent work which attempts to decompose the component processes involved in active inquiry (e.g., Bonawitz & Griffiths, 2010). In particular, we tasked four to ten-year olds to identify a hidden bug in a simple iPad variant of the classic "Guess Who?" game. Children asked questions to try to identify the hidden bug. Across conditions, we manipulated whether the computer program helped children to use the new evidence that resulted from their queries to narrow down the hypothesis space, or whether children had to use the new evidence to update the hypothesis space on their own. Our expectation was that helping children to update their beliefs accurately following the receipt of new information would free up cognitive resources and lead to more effective question-asking. Interestingly, our results opposed our main hypothesis in that elements which ostensibly made our task more difficult actually improved the quality of children's inquiry behavior.

How the ability to ask revealing questions develops

Experimental tasks based on the "Guess Who?" game have often been used to study question asking and active inquiry with both children and adults. In the game, the asker (participant) tries to determine a hidden object known only to the the answerer (experimenter) (e.g., "What animal am I thinking of?") by asking a series of yes-or-no questions. Mosher and Hornsby (1966) identified two broad question types commonly used in the game: *hypothesis- scanning* questions test a single hypothesis (e.g., "Is it a monkey?"), whereas *constraint-seeking* questions attempt to constrain the hypothesis space faster by querying features that are present or absent in multiple objects (e.g., "Is it soft?"), but that do not directly identify the answer except by virtue of elimination.

A classic finding in this literature is that younger children (e.g., aged 6) tend to ask more hypothesis-scanning questions, while older children (e.g., aged 11) use more constraint-seeking questions, and also tend to find the answer after fewer questions (Mosher & Hornsby, 1966). One explanation is that only older children have developed the ability to focus on the high-level features that group the hypotheses, whereas younger children focus on individual stimuli. Consistent with this viewpoint, manipulations that help children focus on these higher-level features (such as cuing them

with basic level category labels instead of exemplar names (Ruggeri & Feufel, 2015) increase the likelihood that young children will generate constraint-seeking questions (see also Herwig (1982). Further, although young children are often relatively less likely than older children to ask constraint-seeking questions, even younger children (ages 7-9) are more likely to do so when such questions are particularly informative, such as when the hypothesis space is large and there several equally probable solutions remaining (e.g., Ruggeri & Lombrozo, 2015; 2015).

Whereas previous work has focused on developmental changes in when children generate informative, hypothesis-scanning questions, less prior work has considered possible developmental changes in how children make use of the new evidence that their questions reveal. As described above, effective active inquiry involves the coordination of multiple cognitive processes–representing the hypothesis space, generating an informative query, updating one's representation of the hypothesis space in light of the data produced by the query, and so on. As suggested by prior work, hypothesis scanning questions might be easier for young children to generate because they do not require abstracting informative higher level features (features to query that group classes of hypotheses together and might allow them to be eliminated at once). Yet, another reason why hypothesis scanning questions might be easier for young children is that they produce evidence that is easier for them to process. As a hypothesis scanning question is answered, children are told directly whether the item they queried is correct or not. If instead children ask about a feature (as in a constraint-seeking question), additional cognitive processing is required–children have to take that new information (e.g., that a hidden animal has antennae) and consider each remaining possible exemplar in light of this information (e.g., check if each one has the antennae) and eliminate from the hypothesis space any that are ruled out by the new information. This process could be cognitively taxing, and also prone to errors. Thus, although constraint seeking questions are often more informative in theory, they might not always be so to young children, particularly if children have difficulty using the obtained information to update their representation of the hypothesis space accurately. To address these issues, in the present study we manipulated whether children received assistance in updating their hypothesis space or had to undertake this process on their own, following the receipt of new evidence obtained by their queries.

Experiment

Methods

Participants Participants in this experiment were 134 children between the ages of 5 and 10 years old who were recruited at the American Museum of Natural History's Discovery Room. Of the 134 children recruited, we analyze the data from 121 children (21 5-year-olds, 20 6-year-olds, 22 7-year-olds, 20 8-year-olds, 20 9-year-olds, and 18 10-year-olds) who completed 5 or more rounds of the game.

Stimuli On each round, sixteen bugs with the same body shape but with varying features were used as stimuli. Bugs were defined by the presence or absence of 9 features: green body, orange eyes, antennae, big spots, tiny spots, legs, leaves, water droplets, and blue "fur". Figure 1 shows an example of two of the body shapes used, each with all of the binary features present. One of the sixteen possible bugs was chosen as the "hidden bug" on each trial which children attempted to identify by asking questions. The hidden bug was randomly selected on each round, and each round had differently-shaped bug bodies, selected from a pool of 16 unique body shapes. The bug task was used to fit thematically with the content of the AMNH Discovery Room activities.

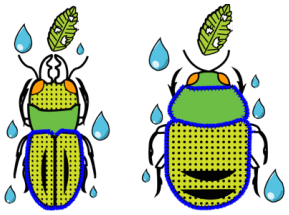


Figure 1: Examples of two bug types with all 9 of the binary features present. Each round used one of the 16 body shapes.

Design Across the sixteen items, some feature were more frequent than others (relevant to eight of the possible bugs), while some were very infrequent (relevant to only 1 bug), with an abstract structure shown in Figure 2. This introduced strong differences in the informational utility of each feature. For example, given no other information it would be informative to ask about feature F1 because is it shared with half the possible bugs. In contrast, feature F9 is less informative on the first trial because most of the bugs do no have this feature. The abstract features in Figure 2 were randomly assigned to the visual features for each participant, and then remained consistent across rounds. This gave participants the opportunity to learn the structure across rounds, for example to perhaps figure out which visual features are most relevant to ask about first.

Each of these features was represented on a button, available for participants to query. Before participants were allowed to begin, the experimenter explained at least three of these buttons, randomly selected. An additional feature button depicted a particular body shape that was not relevant to the bugs on display. Instead of choosing a feature button, participants could at any time query an exemplar to determine if it was the hidden bug or not. This paradigm thus enabled us to investigate both the qualitative strategies used by participants (constraint-seeking feature queries or hypothesis-scanning exemplar queries) and to quantify how efficiently participants searched the hypothesis space, within and across rounds as they learn a novel structured stimulus space. Moreover, we introduced a novel manipulation: after making a feature query, participants in the manual-update condition had to select the hypotheses that were consistent with the feed-

Exemplar	F1	F2	F3	F4	F5	F6	F7	F8	F9
A	**1**	0	0	**1**	0	0	0	0	0
B	**1**	**1**	0	**1**	0	0	0	0	0
C	0	**1**	0	**1**	0	0	0	0	**1**
D	0	**1**	0	**1**	0	0	0	0	0
E	**1**	0	0	0	0	0	0	0	0
F	**1**	**1**	0	0	0	0	0	0	0
G	0	**1**	0	0	0	0	**1**	**1**	0
H	0	**1**	0	0	0	0	**1**	0	0
I	**1**	0	**1**	0	0	**1**	0	0	**1**
J	**1**	0	**1**	0	0	**1**	0	0	0
K	0	0	**1**	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0
M	**1**	0	**1**	0	**1**	0	0	**1**	0
N	**1**	0	**1**	0	**1**	0	0	0	0
O	0	0	**1**	0	**1**	0	0	0	0
P	0	0	0	0	**1**	0	0	0	0

Figure 2: The abstract feature structure of the 16 exemplars used in each round. Each participant had these abstract features randomly assigned to the visual features, but had a consistent assignment used round-to-round.

back, whereas participants in the automatic-update condition had the hypothesis space automatically updated. This manipulated the ease of updating the hypothesis space: a difficult step in the cycle of active inquiry that has not been well-studied.

Procedure After being trained by an experimenter on a simpler version of the task with unrelated stimuli (a dog searching dog houses), participants played 5 or more rounds of an iPad game asking them to identify which one of 16 bugs is hidden under a rug (see Figure 3). The task alternated between the query phase and the elimination phase. In the query phase, players either queried individual bugs, or used feature queries (e.g., "Is the hidden bug green?") to find out whether the hidden bug had a particular feature. If a single exemplar was queried by tapping on it, feedback was immediate: if it happened to be the hidden bug, a smiley face appears and the round was done, whereas if the tapped exemplar was not the hidden bug, a red "X" appeared on top of the tapped bug and the bug becomes grayed out (i.e., eliminated). After a feature query, the bug gave feedback, saying "Yes!" (it has the feature; narrated by the experimenter), or "No!" (it does not have the feature). This was followed by the elimination phase, during which bugs that are inconsistent with the feedback were eliminated, thus narrowing the hypothesis space.

Participants were assigned in counterbalanced order to one of two hypothesis-updating conditions. In the automatic-update condition, after the feedback from a feature query, subjects merely pressed the "Eliminate" button and all the irrelevant bugs are eliminated (grayed out), and the game returns to the guessing phase. In the manual-update condition, after a subject made a feature query and saw feedback, they had to select each bug that was consistent with the feedback for that feature, as shown in the top right of Figure 3. Bugs were selected (denoted by a green box) by tapping, and could be deselected by tapping again. Only when participants were done selecting bugs did the experimenter press the "Eliminate" button, which eliminated any bugs that were not selected. Although manual-update participants received training for the manual elimination in the dog house training task, as well as gentle reminders in the first round of the bug game, it should be noted that it was possible for mistakes to be made during manual updating–unlike in the automatic condition.

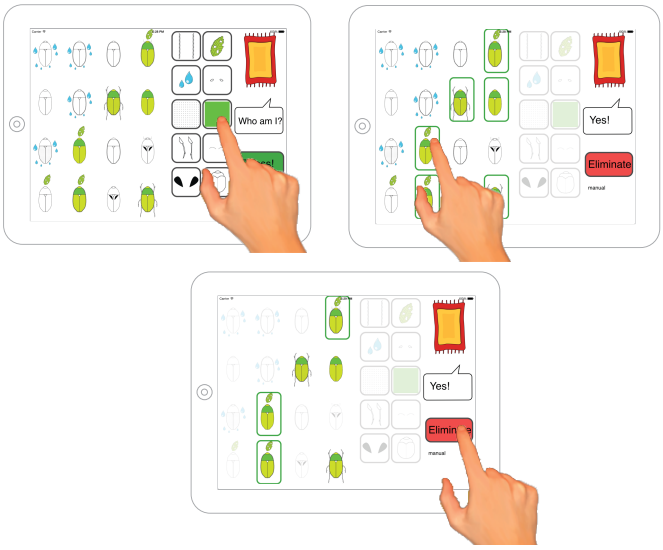


Figure 3: Task overview: in the upper left, a feature button is used, asking if the bug hidden under the rug is green. Given feedback ("Yes!"), participants in the manual update condition select the bugs that are consistent with this new information (upper right), whereas in the automatic condition the consistent bugs are selected by the game. Players in both conditions press the red button to return to the button phase, and again either choose a feature button or query a single bug.

Results

Overall We analyzed the first 10 rounds from each child (only 8 children played more than 10 rounds, including one who played 51 rounds). This covers 722 rounds from 121 children. The mean number of total queries (feature and exemplar) taken to complete a round was 6.5 in the automatic-update condition, and 7.6 in the manual-update condition. Although the median queries to complete a round in each condition was 6, the distributions were significantly different (Kolmogorov-Smirnov test, $D = 0.13$, $p < .01$). For comparison, we simulated 700 rounds of the game with an agent that clicked randomly in the task. This agent took on average 8.9 queries (median: 9) to complete a round.

Response Times Participants' median RT for each button type (feature and exemplar) was computed and these data were subjected to an ANOVA with condition (automatic, manual) and age group (5-7, 8-10) as between-subjects factors and button type as a within-subject factor. There were significant main effects of button type ($F(1,229) = 42.52$, $p < .001$) and condition ($F(1,229) = 4.14$, $p < .05$), but not a significant main effect of age group ($F(1,229) = 0.73$). On av-

erage, participants took longer to make queries in the manual condition (4800 ms) than in the automatic condition (4000 ms). Overall, participants took much longer to make feature queries (7,470 ms) than to press an exemplar button (2,680 ms), perhaps indicating more thought before making more complex queries. There was also a significant interaction effect of button type and condition (F(1,229) = 12.89, $p < .001$). Figure 4 shows the mean of subjects' median RTs for each button type, split by condition. Feature queries were slower in the manual-update condition (7900 ms vs. 5430 ms in automatic), which could indicate 1) more careful thought given to features in this condition, and/or 2) general hesitance to use feature queries, perhaps because it is time-consuming (even difficult) to manually update hypotheses. Exemplar queries were faster in the manual-update condition (1850 ms vs. automatic: 2570 ms), which could be greater readiness to use the simpler strategy.

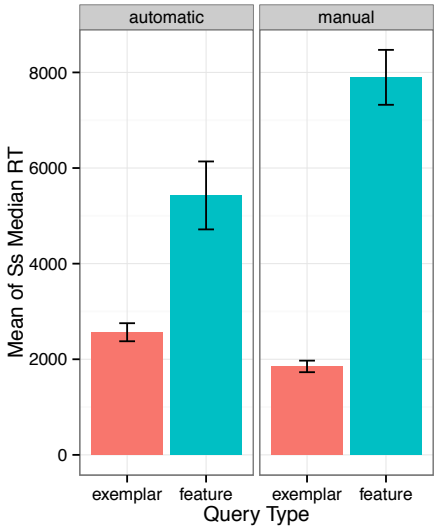


Figure 4: Mean of participants' median RT for each condition and query type. Exemplar queries were faster than feature queries, which represent a more complex strategy and thus likely required more thought. Feature queries were slower in the manual-update condition: it seems the difficulty of updating in this condition made participants think even more carefully about using feature queries. Error bars show +/-1SE.

Querying Behavior Participants' mean number of queries per round were subjected to an ANOVA with condition and age group (5-7 vs. 8-10) as between-subjects factors and query type as a within-subject factor. This analysis indicated significant main effects of condition (F(1,229) = 4.60, $p < .05$) and age group (F(1,229) = 12.20, $p < .001$), and no significant main effect of query type (F(1,229) = 0.10, $p = .75$). Overall, older children required fewer total queries to complete a round ($M_{5-7} = 4.2$, $M_{8-10} = 3.3$), also evidenced by a significant negative correlation with age ($t(119) = 3.24$, $p = .001$, $r = -.28$). There were significant interactions of condition and query type (F(1,229) = 22.18, $p < .001$), and age group and query type (F(1,229) = 12.25, $p < .001$). No other interactions were significant (all Fs < 1).

Figure 5 shows the average number of query types used per round for participants by age group. Both age groups in the manual-update condition used more exemplar queries than feature queries. In comparison to the manual condition, there were fewer exemplar queries in the automatic condition ($M_{man} = 5.0$, $M_{auto} = 3.2$, $t(103.5) = 4.1$, $p < .001$), while there were more feature queries in the automatic condition ($M_{auto} = 3.8$) ($M_{man} = 3.3$, $t(102.9) = 2.1$, $p < .05$). These query rates were all lower than the simulated random rounds' mean number of feature queries (6.5) and exemplar queries (5.3), but above the optimal.[1] Older participants used a greater proportion of feature queries than younger participants in both the automatic ($M_{5-7} = .50$ vs. $M_{8-10} = .66$, $t(57.2) = 3.12$, $p < .01$) and manual conditions ($M_{5-7} = .39$ vs. $M_{8-10} = .50$, $t(50.3) = 2.30$, $p < .05$). Thus, both conditions replicate the Mosher and Hornsby (1966) finding that older children use a greater proportion of constraint-seeking questions.

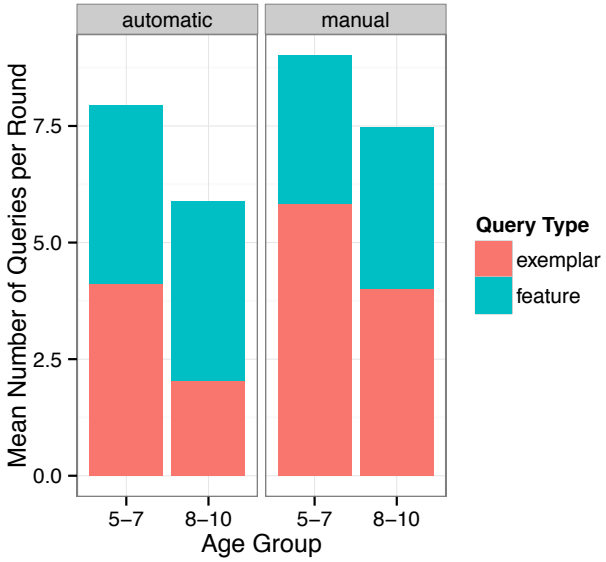


Figure 5: Mean number of queries of each type per round by age and condition. Error bars show +/-1SE.

In summary, it is clear that the manual-update condition results in fewer feature queries and more reliance on exemplar queries. Manual-update participants may be loathe to use feature queries for at least two reasons: 1) it demands more time and cognitive effort to manually update the hypothesis space after a feature query than in the automatic-update condition, and 2) the manual update process is error-prone, and any mistakes may in turn lead to more exemplar queries in order to recover.[2] Therefore we proceed to investigate errors in man-

[1] Although there were at first more exemplars (16) than feature buttons (10), after the first 1-2 clicks there would likely be few exemplars remaining, thus the expected number of exemplar queries is lower than the expected number of feature queries in the simulation.

[2] If the correct answer is mistakenly eliminated, additional clicks on the grayed-out bugs were needed to find it and finish the round.

ual updating, as well as information theoretic analyses that will indicate whether the quality of feature queries varied in the two update conditions. Although the qualitative analyses have thus far revealed interesting effects that build on the previous literature, as the game unfolds the utility of different query types (and specific queries) changes, and can be best quantified using a more sophisticated model-based approach to understanding the quality of children's question asking.

Manual Update Mistakes The manual-update condition allows participants to commit two types of error during hypothesis updating: a miss is defined as a failure to eliminate a bug, and a false alarm is a failure to keep a hypothesis that was consistent with the query. Note that a miss is an error of commission–i.e., the bug had to be tapped to be kept–whereas a false alarm is an error of omission (i.e., failing to tap a bug), and thus we expect more of the latter. Comparing the manual-update subjects' mean number of errors of each type per round, indeed there were more false alarms ($M = 6.9$, sd = 1.9) than misses ($M = 1.8$, sd = 1.3; paired $t(58) = 19.8$, $p < .001$). A MANCOVA to determine if error rates were related to age did not find a significant effect for either misses (F(1,56) = 0.77, $p > .05$) or false alarms (F(1,56) = 0.23, $p > .05$). Given the fairly high rate of errors in manual updating, it is perhaps unsurprising that fewer feature queries and more exemplar queries were made in this condition than under automatic updating of the hypothesis space. However, RT analyses indicated that feature queries took longer under manual updating: is this just reluctance, or could it be that feature queries were more carefully considered in this condition than under the ease of automatic updating? The expected information gain of children's feature queries provides a measure of their sensitivity to the information structure in the stimuli.

Expected Information Gain Each successive query reduces the size of the remaining hypothesis space to some degree: on the first move, querying the appropriate feature (F1) can cut the space in half. When two hypotheses remain, even an exemplar query will cut the space in half. The appropriate way to analyze the contextual sensitivity (i.e., are they choosing a feature that is present for half of the remaining exemplars, thus quickly reducing the hypothesis space?) of participants' queries is to calculate the Expected Information Gain (EIG) of the query they made. We first introduce key terms used to define EIG. Entropy measures uncertainty about the outcome of a random variable X. Entropy is 0 when there is only one possible outcome, and maximal when all possible outcomes are equiprobable (i.e., a uniform distribution).

$$H(X) = -\sum_{x} p(x) \cdot log(p(x)) \tag{1}$$

Mutual information gain measures the change in entropy as we receive a new piece of information Y, i.e., how much does our uncertainty about X change given that we know Y?

$$I(X;Y) = H(X) - H(X|Y) \tag{2}$$

The Expected Information Gain (EIG) of a query Q is the weighted average of the information possible from each possible answer to the query, weighted by the current probability of receiving that answer. This will be 0 (or near-0) for queries that can be expected to eliminate none or just one or two hypotheses in a large space, and more positive for queries that are likely to eliminate a larger number of hypotheses. In this task, EIG is maximal (1) for a feature query that will eliminate half the remaining hypotheses. Such a query is always available at the beginning of any round, and due to the partially-nested feature structure used, maximal EIG queries are often available at other stages of the round.

$$EIG(Q) = -\sum_{Y} p(Y|Q) I(X;Y) \tag{3}$$

EIG has often been proposed as a model of how children might evaluate the quality of possible queries. For example, Nelson, Divjak, Gudmundsdottir, Martignon, and Meder (2014) found that 8-10 year-old children can search a familiar structured domain (people with varying gender, hair color, etc.) fairly efficiently, tending to ask about frequent real-world features that roughly bisected the search space. Likewise, Ruggeri, Lombrozo, Griffiths, and Xu (2015) found evidence that children's patterns of search decisions were well-explained in terms of EIG.

In our study, the EIG for each participants' feature queries[3] were computed, and their mean EIG was subjected to an ANOVA with condition and age group (5-7 vs. 8-10) as between-subjects factors. This ANOVA indicated significant main effects of condition (F(1,115) = 55.0, $p < .001$) and age group (F(1,115) = 12.42, $p < .001$), with no significant interaction effect (F(1,115) = 0.2, $p > .05$).[4] Figure 6 shows mean EIG per feature query by age group and condition. Mean EIG of feature queries for each subject was marginally correlated with age ($t(116) = 1.77$, $p = .08$, $r = .16$), suggesting that older children tended to use more relevant feature queries. The feature queries made by participants in the automatic condition had significantly lower EIG than those made in the manual condition ($M_{auto} = .60$, $M_{man} = .74$, $t(116) = 5.49$, $p < .001$). Thus, although manual-update participants used fewer feature queries overall, and tended to make mistakes during hypothesis updating, the greater amount of time they spent when choosing a feature query tended to pay off: manual-update participants queried features with higher expected information gain than automatic-

This occurred rarely, happening in $< 10\%$ of rounds.

[3]Exemplar query EIGs are less interesting, as they are a simple function of how many remaining hypotheses there are. Participants' choice of feature query, on the other hand, indicates how sensitive they are to the relevance of each feature–and to the context of their current situation, as it is based on the remaining bugs' features.

[4]The same significant effects and similar mean EIG values were obtained when analyzing only the first two feature queries per round, when manual- and automatic-update participants were on more equal footing (i.e., before further manual errors–which could raise or lower the EIG of the remaining feature queries).

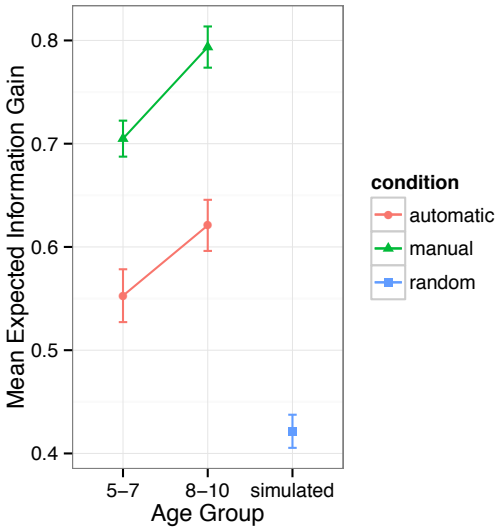


Figure 6: Mean expected information gain for feature queries by age group and condition, with simulated subjects making random feature queries for comparison. Manual- update subjects had higher EIG than automatic-update subjects, and both were better than random–but suboptimal (1). Older children had higher EIG than younger children. Bars show +/-1SE.

update participants. Indeed, there was a weak but significant correlation of participants' mean feature query RT and EIG ($r = .20$, $t(116) = 2.17$, $p < .05$), verifying that longer RTs are associated with more informative feature queries.

General Discussion

The present study asked children 5-10 years of age to learn feature distributions in an unfamiliar hypothesis space, and examined both their qualitative questioning strategies, and how efficiently they were able to search that space. Importantly, we manipulated the support children were given while updating the hypothesis space: after a feature query, participants in the automatic update condition were shown which bugs were eliminated at the press of a button, whereas manual update participants were required to select the bugs that were consistent with the feedback.

In line with previous research (Mosher & Hornsby, 1966; Ruggeri & Lombrozo, 2014), older children (ages 8-10) asked a higher proportion of constraint-seeking questions than younger children (ages 5-7), who relied more on hypothesis-scanning (i.e., exemplar queries), in both conditions. These qualitative analyses also found that children use more constraint-seeking questions (i.e., feature queries) in the automatic-update condition. On the surface then, these children were using a more efficient strategy than the manual-update children.

However, in terms of expected information gain, a context-sensitive measure of how well a chosen feature bisects the remaining hypothesis space, it turned out that children in the automatic-update condition made less informative feature queries. We suggest that the greater mental effort required by manual updating actually lead to more careful consideration of which feature query to use, and ultimately a better choice. Indeed, response times for feature queries were slower under manual updating, perhaps indicating that greater thought went into making those choices. Indeed, slower feature query RTs were correlated with higher EIG. In both conditions, older children made more informative feature queries, but even 5-7 year-olds asked far more informative questions than a simulation that chose a random sequence of queries, showing some efficiency in navigating an unfamiliar domain even after only a few minutes of experience.

In summary, this study provides evidence that hypothesis updating is a difficult, error-prone step in the active inquiry process. Moreover, children are sensitive to the difficulty of this step: if aided in hypothesis updating, they will ask more constraint-seeking questions than if they must manually update the space. However, we also uncovered evidence of a desirable difficulty in this step, for manual updating resulted in more informative, context-sensitive constraint-seeking questions than the supported update process. Future work will aim to reduce errors in hypothesis updating and discover other bottlenecks–or desirable difficulties–in active inquiry.

Acknowledgments

This work was supported by the John Templeton Foundation "Varieties of Understanding" grant to TMG and MR.

References

Bonawitz, E., & Griffiths, T. (2010). Deconfounding hypothesis generation and evaluation in bayesian models. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of CogSci 32*. Austin, TX.

Bransford, J., Brown, A., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school*. National Research Council.

Chen, Z., & Klahr, D. (1999). All other things being equal: Children's acquisition of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

Donovan, M., & Bransford, J. (Eds.). (2005). *How students learn: Science in the classroom*. Nat'l Research Council.

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades k-8*. Washington, D.C.: National Research Council.

Herwig, J. A. (1982). Effects of age, stimuli, and category recognition factors in children's inquiry behavior. *Journal of Experimental Child Psychology*, *33*, 196–206.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, *16*(11), 866–870.

Mosher, F. A., & Hornsby, J. R. (1966). Studies in cognitive growth. In (chap. On asking questions). New York, NY: Wiley.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*, 74–80.

Ruggeri, A., & Feufel, M. A. (2015). How basic-level objects facilitate asking efficient questions in a categorization task. *Frontiers in Psychology*, *6*(918), 1–13.

Ruggeri, A., & Lombrozo, T. (2014). Learning by asking: How children ask questions to achieve efficient search. In *Proceedings of CogSci 36*. Cognitive Science Society.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T., & Xu, F. (2015). Children search for information as efficiently as adults, but seek additional confirmatory evidence. In D. C. Noelle et al. (Eds.), *Proceedings of cogsci 37*.

Linguistic Signatures of Cognitive Processes during Writing

Laura K. Allen (LauraKAllen@asu.edu)
Cecile Perret (cperret@asu.edu)
Danielle S. McNamara (dsmcnama@asu.edu)
Arizona State University, Department of Psychology, P.O. Box 872111
Tempe, AZ 85281 USA

Abstract

The relationship between working memory capacity and writing ability was examined via a linguistic analysis of student essays. Undergraduate students (n = 108) wrote timed, prompt-based essays and completed a battery of cognitive assessments. The surface- and discourse-level linguistic features of students' essays were then analyzed using natural language processing tools. The results indicated that WM capacity was related to surface-level, but not discourse-level features of student essays. Additionally, the results suggest that these relationships were attenuated for students with high inferencing skills, as opposed to those with lower inferencing skills.

Keywords: writing; natural language processing; computational linguistics; strategies; working memory

Introduction

Writing is a complex cognitive and social process that involves the production of texts for the purpose of conveying meaning to others (Graham, 2006). This task involves cognitive processes such as accessing vocabulary knowledge and constructing grammatical sentences, knowledge (e.g., of language, the writing process, and the domain), and the ability to strategically use language to connect and present ideas in a meaningful way (Donovan & Smolkin, 2006; Graham, 2006; McNamara, 2013).

While a clear demarcation cannot be placed between levels of processes (McNamara, Jacovina, & Allen, 2015), one dominating question in the writing literature regards the relative roles of lower- and higher-level cognitive processes. Evidence considered for the role of lower-level processes generally comes from studies that examine relations between writing ability and working memory (WM) capacity (Kemper, Rash, Kynette, & Norman, 1990; Hoskyn & Swanson, 2003; McCutchen, Covill, Hoyne, & Mildes, 1994). Evidence considered for *higher-level* processes comes primarily from studies showing that skilled writers have more knowledge about writing norms and are more strategic, as well as from studies that show the effectiveness of writing strategy interventions (Graham & Perin, 2007).

Importantly, evidence for lower- and higher level skills tends to come from separate studies and different research camps in the writing literature. Few writing researchers have explored the *relative* influences of lower- and higher-level skills, nor have they considered the potential for higher-level skills to mitigate the impact of lower-level factors. The goal of this study is to address these gaps in the literature by examining the linguistic signatures of cognitive processes in students' essays and their relations to WM capacity.

Working Memory Capacity and Writing

Within the writing literature, WM capacity has received considerable attention and is commonly labeled as a central component of the writing process (Berninger & Swanson, 1994; Hayes, 1996; Kellogg, 2001; 2008; McCutchen, 1996). The measured capacity of an individual's WM has been theorized to relate to their writing ability because of the complex and resource-demanding nature of the task. Writing requires individuals to engage in multiple separable processes, such as accessing word knowledge, planning, activating prior knowledge about a particular domain, and making connections between ideas, all of which can place extreme demands on the cognitive system (Kellogg, 2001).

Notably, however, the link between WM capacity and writing skill has failed to be consistently supported by the literature. Although some studies have reported positive correlations between performance on WM tasks and writing quality (Babayigit & Stainthorp, 2011; Berninger & Swanson, 1994; Kellogg, 2008), others have found this relationship to be non-significant or even negative (Allen et al., 2014; Dixon, LeFevre, & Twilley, 1988).

One potential explanation for this conflicting evidence lies in the characteristics of the rubrics by which essays are assessed. The measured capacity of an individual's WM may be inconsistently related to essay quality because the definition of writing quality has not been operationalized consistently (i.e., it might focus more or less on different text properties). If this is the case, in order to understand what role (if any) WM capacity plays in the writing process, it is important to conduct analyses at multiple levels of discourse. Our supposition is that multi-dimensional discourse analyses can potentially deepen our understanding of individual differences in writing.

The Role of Higher-Level Skills

In addition to considering texts at multiple levels, researchers can benefit from a consideration of the interactions that potentially occur amongst the lower- and higher-level skills that students have developed. Despite the stronger emphasis on lower-level skills in cognitive writing research, evidence from educational research suggests that the development and use of higher-level skills (e.g., strategies) can significantly reduce the demands of the writing process and enhance writing performance.

For example, in a meta-analysis conducted on over 120 published studies of writing interventions, Graham and Perin (2007) found that strategy instruction was the most

effective form of writing instruction. Given these findings, it may be the case that the writing process is primarily constrained by WM in the absence of higher-level skills. However, once students have been trained to employ strategic processes during writing, this relationship between WM and writing ability may be significantly reduced.

Current Study

We adopt a different theoretical and methodological approach than is considered typical of studies in the writing literature. Our theoretical approach is motivated by research in cognitive science, which supposes that the development of higher-level skills, such as the ability to generate inferences, can help to make up for deficits in knowledge and cognitive capacity (e.g., McNamara & Scott, 2001).

Our methodological approach is inspired by research showing that the linguistic properties of texts reflect readers' potential levels of comprehension as measured using indices at multiple levels (e.g., Graesser & McNamara, 2011). To investigate writing, our approach is to consider the notion that there are multiple linguistic dimensions of the texts that students produce. *Surface-level* text features relate to the characteristics of the words and sentences in texts. Variations in these features can alter the style of the essay, as well as influence its readability and perceived sophistication. *Discourse-level* features, on the other hand, go beyond the individual words and sentences, and instead reflect aspects of the situation model portrayed by the text such as the degree of narrativity in the essay.

Our first hypothesis is that the surface-level features of students' essays will be related to their WM scores. Second, we hypothesize that the role of WM capacity in the writing process will be moderated by the development of strategic inferencing skills (i.e., higher-level cognitive skills). Thus, the relations between students' WM scores and the characteristics of their essays may be more or less pronounced depending on the degree to which they have developed abilities to think and write strategically.

We investigate these hypotheses through a linguistic analysis of student essays. We first examine relations between students' WM scores and the properties of their essays at both the surface- and discourse-levels of the text. Next, we examine whether these relationships differ as a function of students' inferencing skills.

Our research questions are listed below:

1) Do WM scores demonstrate significant relations to surface-level and/or discourse-level linguistic properties of students' essays?
2) Do these potential relations between WM scores and text properties vary as a function of students' inferencing skills?

Methods of Automated Text Analysis

To calculate the linguistic properties of students' essays (see Table 1 for the indices calculated), we used two natural language processing (NLP) tools: Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) and the Writing Assessment Tool (WAT; McNamara, Crossley, & Roscoe, 2013). Both tools report hundreds of linguistic indices that relate to the structure of the text, its general readability, rhetorical patterns, lexical choices, and cohesion using a combination of components that are commonly used in NLP tools (Crossley, Allen, Kyle, & McNamara, 2014).

For the purposes of the current analysis, we pre-selected 20 indices from Coh-Metrix and WAT, all of which had theoretical links to writing quality. These indices related to two primary essay levels (each containing 10 variables): surface-level and discourse-level text features. The surface-level features relate primarily to the characteristics of the individual words and sentences in the text, whereas discourse-level features relate to text cohesion and rhetorical functions. It is important to note that, although we have grouped the linguistic features into two distinct categories, they lie more realistically on a continuum. Thus, some of the discourse-level features will tend more towards the surface level than others (and vice versa). The indices selected for these two categories are described briefly below. For more thorough descriptions of these indices and their theoretical links, see McNamara et al. (2014).

Surface-level Text Indices The surface-level text indices selected describe word- and sentence-level characteristics of students' essays. Indices were broadly selected to account for a multitude of independent constructs related to texts at these levels. These indices range from simple frequency counts for certain parts-of-speech to more informative measures that describe the types of words used in a text.

Word Information. Coh-Metrix and WAT calculate multiple indices that describe the specific types of words used in texts. Word frequency measures, for instance, are used to assess how frequently certain words occur in the English language. Coh-Metrix reports indices of word frequency that are taken from the CELEX database. Coh-Metrix additionally reports the logarithm of word frequency for all words in a text and the minimum log word frequency for content words. An index of log frequency is calculated because reading times are typically linearly related to the logarithm of word frequency (rather than the raw word frequency; Haberlandt & Graesser, 1985). Coh-Metrix, therefore, provides the average minimum log frequency of words across sentences (minimum log word frequency). Average measures of word frequency are important indicators of lexical knowledge and can inform text readability because frequent words are more easily accessed and decoded than less frequent words (Perfetti, 1985).

Additionally, Coh-Metrix employs WordNet to calculate polysemy and hypernymy scores for all content words in a text. *Polysemy* scores denote the number of senses that are associated with a given word (ambiguous words have more senses). Polysemy scores provide indications of lexical proficiency (McNamara et al., 2014). *Hypernymy* is indicative of the specificity of a given word – as defined by its location within a conceptual hierarchy (e.g., *dog* would

have a lower hypernymy value than *poodle*). Hypernymy scores have been linked to lexical knowledge and production (Crossley & McNamara, 2009).

Part-of-speech and Sentence Information. Coh-Metrix and WAT additionally contain multiple indices that describe the features of the sentences in texts, such as the parts of speech they contain and the complexity of their constructions. Coh-Metrix reports incidence scores for all of the part-of-speech tags in the Penn Tree Bank Tag Set. This set includes tags at the word and phrase levels for content items, as well as for function items. These tags have been used in previous studies to classify high and low quality essays (Crossley & McNamara, 2011).

In the current study, we selected part-of-speech tags that were related to function items (to assess the surface-level grammatical structures of the text, rather than specific content) and that target independent grammatical constructs. We examine the incidence of modals (might, could), the incidence of prepositions (on, in, around, between) and subordinating conjunctions (after, because, whereas, unless), the incidence of first person pronouns (I, me), the incidence of third person pronouns (he, she), the incidence of causative subordinators (because, as), and the incidence of phrasal coordinators (and, but).

Sentence complexity is measured in Coh-Metrix with multiple indices. Higher quality essays typically contain more complex syntactic constructs (McNamara et al., 2014), which can increase WM load on readers (Graesser et al., 2006). However, it is unclear whether these syntactic constructions are related to the WM of the *writer.* We used the index *mean number of words before the main verb* as a proxy for sentence complexity.

Discourse-level Text Indices The discourse-level indices relate to the cohesion and semantic properties of texts. As with the surface-level indices, these measures were broadly selected in order to target multiple constructs. The final set of indices ranges from basic measures of cohesion to more robust component indices related to the style of the text.

Connectives. Coh-Metrix provides an incidence score for the number of connectives that are contained in a text. Connectives increase cohesion because they explicitly link ideas and clauses (Longo, 1994). Coh-Metrix also provides indices related to specific categories of connectives. Because the essays in this study were argumentative essays that rely on logical argumentation, we chose to analyze the incidence of logical connectives to serve as a measure of cohesion.

Lexical Overlap. Cohesion is also calculated through indices of overlap for certain parts-of-speech. Relevant to the current study, argument overlap calculates how often two sentences share nouns with common stems. Lexical overlap increases the readability of a given text (Kintsch & van Dijk, 1978).

Situation Model Cohesion. Coh-Metrix calculates multiple indices that attempt to tap into text cohesion beyond the word level. For instance, WordNet is used to assess spatial cohesion using two forms of information: location information and motion information. Location information is represented through nouns such as *school* or *Indiana*, whereas motion information is represented through verbs such as *fight* or *run*. Similarly, intentional cohesion is measured by the ratio of intentional particles to intentional verbs. Both of these cohesion measures help to increase the readability of a text by promoting the successful generation of inferences by readers (McNamara et al., 2014).

Semantic Cohesion. Semantic cohesion is calculated in Coh-Metrix using Latent Semantic Analysis (LSA). Semantic overlap is calculated between the paragraphs in the essay (McNamara et al., 2014). Similarly, WAT uses LSA to calculate verb overlap. This measure of verb cohesion calculates the average LSA cosine between verbs in adjacent sentences. We used two indices of semantic cohesion: LSA overlap amongst all paragraphs (LSA paragraph-to-paragraph) and LSA overlap between verbs. These indices are indicative of the extent to which certain concepts are repeated across sections of essays.

Easability Component Scores. Coh-Metrix calculates five text Easability Components that were developed to account for the multiple dimensions of text difficulty (see Graesser, McNamara, & Kulikowich, 2011, for more information). In this study, we analyzed the components related to text *narrativity, referential cohesion, and deep cohesion,* because they relate to discourse-level text properties and essay quality. Narrativity captures the genre/style of a text by calculating the amount of story-like and familiar elements it contains. Referential cohesion measures how words and ideas are repeated between sentences. Deep cohesion refers to how ideas connect throughout the text (i.e., through causal and logical relationships).

Method

Participants This study included 108 college students from a large university campus in the Southwest United States. These students were, on average, 19.75 years of age (range: 18-37 years), with the majority of students reporting a grade level of college freshman or sophomore. Of the 108 students, 52.9% were male, 53.7% were Caucasian, 22.2% were Hispanic, 10.2% were Asian, 3.7% were African-American, and 9.3 % reported other nationalities. The data for two students were lost due to a computer failure.

Study Procedure This study consisted of a 2-hour session during which students completed the following assessments (in this order): demographics questionnaire, timed-essay, vocabulary test, comprehension test, WM task, and a component processes task.

Essays All students wrote a timed (25-minute), prompt-based, argumentative essay that resembles what they would see on the SAT (previously referred to as the Scholastic Aptitude Test; sat.collegeboard.org). Students were not allowed to proceed until the entire 25 minutes had elapsed.

Working Memory Capacity Students' WM capacity was assessed using the Automated Operation Span (Aospan; Unsworth, Heitz, Schrock, & Engle, 2005). Students'

overall Aospan score reflects the total number of letters they correctly recognized and correctly ordered.

Component Processes Students' inferencing ability was measured using Hannon and Daneman's (2001) component processes task. This test assesses individual differences in higher-level cognitive skills: text memory, text inferencing, knowledge access, and knowledge integration. Previous studies have used this task to identify the processes involved in reading comprehension (Hannon & Daneman, 2001) and writing proficiency (Allen et al., 2014).

Results

Correlation and regression analyses were conducted to examine relations between students' WM scores and the surface- and discourse-level properties of their essays. We first investigate these relations for all students, and then investigate whether and how these relationships differ for students with low and high inferencing skills.

Table 1. Pearson correlations between linguistic indices and AOSPAN scores for all students, low inference students, and high inference students

	Linguistic Index	All Ss	Low Inf	High Inf
Surface-level	Word frequency[1]	**.22***	**.40****	.04
	Word polysemy	**-.23***	**-.34***	-.07
	Word hypernymy	.03	-.07	.06
	Modals[2]	-.18	**-.35***	.05
	Subordinating conj.[3]	**.34****	**.30***	**.40****
	1st person pronouns[2]	.07	.26	-.08
	3rd person pronouns[2]	-.01	-.01	-.04
	Causative subordinators[2]	**.21***	**.32***	.06
	Phrasal coordinators[2]	-.15	**-.28***	-.07
	Word before main verb	**.21***	.24	.17
Discourse-level	Logical connectives[2]	.18	**.34***	.05
	Argument overlap	.03	-.02	.20
	Location nouns[2]	.00	.07	-.08
	Motion verbs[2]	-.18	**-.34***	.03
	Intentional ratio[4]	-.13	**-.33***	.16
	LSA Paragraph overlap	.08	.03	.15
	LSA Verb overlap	.06	.04	.14
	Narrativity[5]	-.01	.09	-.09
	Referential cohesion[5]	.11	.03	.23
	Deep cohesion[5]	.03	-.02	.10

Notes: ** $p < .001$; * $p < .05$; [1]Minimum log word frequency; [2]Incidence; [3]Incidence of prepositions and subordinating conjunctions; [4]Ratio of intentional particles to intentional; [5]Easability Percentile Score

Working Memory and Text Features

Students' scores on the WM (M = 56.44, SD = 11.79) and inferencing (M = 61.16, SD = 15.24) measures were not significantly correlated ($r = .14$, $p = .17$). Therefore, it can be inferred that the tasks measured independent skills. Students' essays contained an average of 410.44 words (SD = 152.50), ranging from a minimum of 84 words to a maximum of 984 words. Table 1 presents the Pearson correlations between scores on the WM test and the surface-level and discourse-level essay properties selected for the NLP analysis. WM scores were primarily related to the surface-level properties of students' essays, such as word frequency and syntactic complexity. However, the discourse-level indices demonstrated much weaker relations with WM, with only two marginally significant correlations demonstrated (i.e., motion verbs, logical connectives).

A stepwise regression was conducted to determine which of these text properties were most predictive of WM scores. The indices that demonstrated significant or moderately significant correlations were regressed onto the WM scores, yielding a significant model [$F(3, 97) = 9.85$, $p < .001$; $R^2 = .23$] with three significant predictors: incidence of prepositions and subordinating conjunctions [$B = .34$, $t(1, 97) = 3.75$, $p < .001$], minimum log word frequency [$B = .32$, $t(1, 97) = 3.45$, $p = .001$], and number of words before the main verb [$B = .23$, $t(1, 97) = 2.47$, $p = .015$]. The results suggest that students with higher WM scores generated texts that contained more complex sentence structures, less familiar words, and a greater incidence of transition words. Importantly, all variables retained in the analysis had been classified a priori as surface-level text properties. Thus, the results provide evidence that the relationship between WM and writing may be strongest at surface levels of the text, such as in the sophistication of the words and the complexity of the sentences.

Role of Inferencing Skills

Our second research question focused on the interplay between WM and inferencing skills during the writing process. To investigate this research question, we conducted a median split on participants' component processes task scores, which resulted in two groups: low (n = 53; M = 17.49, SD = 3.09) and high inference ability students (n = 53; M = 26.55, SD = 3.07). Separate correlation and regression analyses were then conducted on the two groups to determine whether the relations between WM and the linguistic indices were weaker for the high inference ability students than for the low inference ability students.

Low inference ability students Pearson correlations were calculated between the linguistic indices and students' WM scores (see Table 1). Similar to the previous analyses, performance on the WM test was most strongly related to the surface-level linguistic indices. The surface-level linguistic indices that most strongly correlated with WM scores were: minimum log word frequency, incidence of modals and word polysemy. Thus, for low inference ability students, higher WM scores were associated with more frequent words, along with language that was less abstract and more direct (modals, such as *might*, serve as a means through which writers can hedge their arguments).

Notably, fewer (only three) discourse-level indices related to WM scores: motion verbs, ratio of intentional particles to intentional verbs, and logical connectives. Thus, students with higher WM scores produced essays with fewer motion

verbs, less intentional cohesion, and more logical connectives.

To determine which of the indices were most predictive of WM scores, a stepwise regression was calculated with the variables that demonstrated significant or marginally significant correlations. This yielded a significant model [$F(4, 45) = 9.61$, $p < .001$; $R2 = .46$], with four predictors: minimum log word frequency [$B = .40$, $t(1, 45) = 3.63$, $p = .001$], word polysemy [$B = -.32$, $t(1, 97) = -2.84$, $p = .007$], logical connectives [$B = .32$, $t(1, 45) = 2.88$, $p = .006$], and incidence of modals [$B = -.24$, $t(1, 45) = -2.09$, $p = .043$]. These results suggest that, for the low inference ability students, text properties accounted for nearly *half* (i.e., 46%) of the variance in WM scores. Additionally, performance on the WM assessment was more strongly predicted by surface-level text properties than discourse-level text properties.

High inference ability students Similar analyses were conducted for the high inference ability students (see Table 1). As with the low inference ability students, performance on the WM test was more strongly related to surface-level linguistic indices than discourse-level linguistic indices. Unlike the low inference ability students, however, only one of the indices was significantly correlated with WM scores (incidence of prepositions and subordinating conjunctions). A follow-up regression was conducted and yielded a significant model [$F(1, 49) = 9.15$, $p = .004$; $R^2 = .16$], with incidence of prepositions and subordinating conjunctions [$B = .40$, $t(1, 49) = 3.02$, $p = .004$] as a significant predictor.

Taken together, the analyses of the two student groups provide confirmatory evidence for our hypothesis that inferencing skills helped to attenuate the effects of WM capacity on the writing process. Additionally, across the analyses, surface-level text properties provided the most predictive power for WM scores, suggesting that this individual difference predominantly manifests in the surface-level features of texts. These results potentially suggest that when students develop strong inferencing skills, the consequences of WM deficits may be reduced.

Discussion

In this study, we examined the relationship between WM capacity and writing ability through a linguistic analysis of student essays. The results first confirmed the notion that WM capacity is related to the features of texts produced by writers. Namely, students with higher WM scores produced essays that contained more sophisticated vocabulary and complex sentence constructions. The discourse-level properties of the essays, on the other hand, did not vary according to students' WM capacities. Thus, students with higher WM capacities did not necessarily produce essays that were more coherent or informative than their peers. Taken together, these results emphasize the importance of investigating writing at multiple levels of the text. The differential relations between WM and surface- and discourse-level text properties may shed light on the inconsistent findings regarding WM in previous research. In particular, depending on the nature of the rubric, WM may be more or less related to the scores assigned to essays.

Importantly, the results additionally revealed information about the interactive influence of students' lower- and higher-level skills on the writing process. When considering all of the students in the analysis, the linguistic features of the essays accounted for approximately a quarter (23%) of the variance in WM scores, which may suggest that WM potentially serves an important role in the writing process – at least with respect to the surface-level text properties. This relationship differed, however, once inferencing skills were taken into consideration. For low inference ability students, the linguistic properties accounted for nearly half (46%) of the variance in WM scores, whereas they only accounted for 16% of the variance in high inference ability students' WM scores. These results suggest that higher-level inferencing skills can potentially reduce the negative effects of WM constraints during writing.

These results are also important for writing researchers and educators, as they indicate that the link between cognitive skills and the writing process may fluctuate according to the degree to which students have developed strong inferencing skills. Accordingly, writing proficiency is not only influenced by the cognitive capacity of a given student, but is also (and arguably more importantly) closely related to the degree to which that student has developed strategic skills. This finding is particularly important in the context of education, as students can be taught to generate inferences. Although educators have little means to modify a student's WM capacity, they can help students enhance their strategic skills. Thus, higher-level strategy training may be a powerful intervention tool and potentially offset the negative effects of cognitive constraints.

An additional strength of the current study is that it employs NLP techniques to analyze the linguistic properties of the students' essays. Although previous studies have investigated the role of individual differences in the writing process, they have largely relied on human judgments of essay quality or subjective human coding of specific essay elements. Here, we leveraged NLP tools to automatically calculate the surface- and discourse-level features of students' essays. These analyses afforded us the opportunity to investigate the role of WM capacity at a much finer grain size. Thus, rather than simply concluding that WM is an important component in essay quality (according to certain essay rubrics), we can claim that WM is most strongly related to the production of essays that contain sophisticated words and sentences. Overall, these fine grain linguistic analyses can serve as powerful tools for writing researchers, as they can provide more thorough descriptions for the various components of the writing process.

Acknowledgments

This research was supported in part by the Institute for Educational Sciences, US Department of Education, through Grant R305A120707. Any opinions, findings, and

conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., & McNamara, D. S. (2014) Reading components and their relation to the writing process. *L'Année Psychologique/Topics in Cognitive Psychology, 114*, 663-691.

Babayigit, S., & Stainthorp, R. (2011). Modeling the relationships between cognitive–linguistic skills and literacy skills: New insights from a transparent orthography. *Journal of Educational Psychology*, *103*(1), 169.

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writers. In E. C. Butterfield (Ed.), *Children's writing: Toward a process theory of the development of skilled writing* (pp. 57-81). Greenwich, CT: JAL.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D.S. (2014). Analyzing discourse processing using the Simple Natural Language Processing Tool (SiNLP). *Discourse Processes, 51*, 511-534.

Crossley, S. A. & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18,* 119-135.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *IJCEELL, 21,* 170-191.

Dixon, P., LeFevre, J., & Twilley, L. C. (1988). Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology, 80*, 465-472.

Donovan, C. A., & Smolkin, L. B. (2006). Children's understanding of genre and writing development. *Handbook of Writing Research,* 131-143.

Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly*, *70*, 3–22.

Graesser, A. C. & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2*, 371-398.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223-234.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of Educational Psychology* (pp. 457-478). Mahwah, NJ: Erlbaum.

Graham, S., & Perrin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114*, 357-374.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding the individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93,* 103-128.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In Levy, C. M., and Ransdell, S. (eds.), *The science of writing: Theories, methods, individual differences, and applications*, Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, pp. 1-27.

Hoskyn, M., & Swanson, H. L. (2003). The relationship between working memory and writing in younger and older adults. *Reading and Writing*, *16*(8), 759-784.

Kellogg, R. T. (2001). Long-term working memory in text production. *Memory & Cognition*, 29, 43-52.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*, 1-26.

Kemper, S., Rash, S., Kynette, D., & Norman, S. (1990). Telling stories: The structure of adults' narratives. *European Journal of Cognitive Psychology*, *2*, 205-228.

Kintsch, W. & van Djik, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Longo, B. (1994). The role of metadiscourse in persuasion. *Technical Communication*, *41*, 348– 352.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*, 299-325.

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology*, *86*, 256.

McNamara, D. S. (2013). The epistemic stance between the author and the reader: A driving force in the cohesion of text and writing. *Discourse Studies, 15*, 575-592.

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45,* 499-515.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.

McNamara, D. S., Jacovina, M. E., & Allen, L. K. (2015). Higher order thinking in comprehension. In P. Afflerbach (Ed.), *Handbook of individual differences in reading: Text and context* (pp. 164-176). Taylor & Francis, Routledge: NY.

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition, 29*, 10-17.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498-505.

Statistical Learning of Prosodic Patterns and Reversal of Perceptual Cues for Sentence Prominence

Sofoklis Kakouros (sofoklis.kakouros@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Abstract

Recent work has proposed that prominence perception in speech could be driven by predictability of prosodic patterns, connecting prominence perception to the concept of statistical learning. In the present study, we tested the predictability hypothesis by conducting a listening test where subjects were first exposed to a 5-minute stream of sentences with a certain proportion of sentence-final words having either a falling or rising pitch trajectory. After the exposure stage, subjects were asked to grade prominence in a set of novel sentences with similar pitch patterns. The results show that the subjects were significantly more likely to perceive words with low-probability pitch trajectories as prominent independently of the direction of the pitch change. This suggests that even short exposure to prosodic patterns with a certain statistical structure can induce changes in prominence perception, supporting the connection between prominence perception and attentional orientation towards low-probability events in an otherwise predictable context.

Keywords: statistical learning; prosody; prominence perception; attention; stimulus predictability

Introduction

Recent theoretical and computational studies have suggested that there is a connection between perception of prominence and the predictability of the acoustic prosodic features in speech (Kakouros & Räsänen, 2014; 2015a; in press). The general idea is that subjective perception of prominence is a response elicited to unpredictable prosodic trajectories in a normal train of speech, thereby drawing the attention of the listener. This hypothesis extends the existing probabilistic accounts of prominence that are based, for instance, on the frequency of occurrence of linguistic units such as syllables and words (see, e.g., Aylett and Turk, 2004; 2006) or on word collocation information (see, e.g., Pan & Hirschberg, 2000). The proposal presented in Kakouros & Räsänen (in press) attempts to explain how the speaker must be also capable of manipulating the expectations of the prosodic correlates of prominence in the acoustic signal, therefore going beyond the probabilistic relations at the symbolic linguistic level.

In the present study, we conducted a listening experiment where probabilities of prosodic trajectories were explicitly manipulated, first exposing subjects to different ratios of rising and falling pitch trajectories on sentence-final words and then asking the subjects to grade prominence in a set of novel utterances. The experiments and results, as described in the next sections, show that the probability of prosodic features indeed affects subjective perception of prominence.

Background

Prominence is a prosodic phenomenon that can be generally defined as the property by which linguistic units are perceived to be standing out from their environment (Terken, 1991) and is closely connected to the concept of stress. As the terminology can be ambiguous, here we use the term sentence prominence to refer to one or more words that are perceived to be standing out in a sentence (see, e.g., Terken, 1991; Cutler, Dahan, & van Donselaar, 1997, for related definitions).

Prominence has been examined from a number of different perspectives (see, Wagner et al., 2015, and references therein). From the physical perspective, a number of studies have focused on identifying the acoustic correlates of prominence. It has been now well established that energy, fundamental frequency (F0), duration (see, e.g., Fry, 1955; 1958; Lieberman, 1960; Terken, 1991; Kochanski, Grabe, Coleman, & Rosner, 2005; see also Ortega-Llebaria & Prieto, 2010, and references therein), and spectral tilt (see, e.g., Sluijter & van Heuven, 1996; but see also Campbell, 1995; Campbell & Beckman, 1997) are the acoustic parameters in speech whose variations signal prominence and constituent boundaries (see also Shattuck-Hufnagel & Turk, 1996). Another interesting aspect of these features is the degree of acoustic similarities in prominence production and perception across different languages. The central argument is that regardless of the language, all speakers are equipped with the same production and perception apparatus, therefore, the type of information conveyed through speech should not vary greatly (Vaissière, 1983). This assumption may at least be partly true for prominence as it seems that the basic acoustic correlates of prosody are the same, although the actual realizations of prosodic patterns depend on the language (see, e.g., Rosenberg et al., 2012; Maier et al., 2009). Additionally, a speaker can manipulate the acoustic prosodic features relatively independently of the linguistic content of a sentence. On the listener's side, the perceptual outcome of prominence seems to be the same across languages, that is, a

shift of the attention to a specific part in the stream of speech.

From the functional perspective, prominence has been studied with respect to its linguistic and communicative role. The realization of prominence seems to have effects on the parsing of information and syntactic structure of utterances (see, e.g., Calhoun, 2010; Shattuck-Hufnagel & Turk, 1996). For instance, prominence may indicate the word in an utterance where the most important information lies and it has been observed that reaction times (RTs) for prominent words are shorter when compared to their non-prominent counterparts (see, e.g., Cutler & Foss, 1977). In all, prominent words seem to attract the listener's attention thus allocating extra cognitive processing resources (see, e.g., Cole, Mo, & Hasegawa-Johnson, 2010). At the level of the listener's perceptual processing this implies that there are elements in speech that attract the listener' attention. Cole et al. (2010) have suggested that prominence and attention might be associated, where a listener's attention can be drawn to a word either as a response to acoustic modulation or due its relative unpredictability. In general, an attention-capturing stimulus can be seen as something that is novel or surprising (see, e.g., Itti & Baldi, 2009). Correspondingly, surprisal can be defined in a probabilistic way as something that is unpredictable. In the case of the acoustic prosodic features, this could be manifested, for instance, as an unpredictable F0 trajectory (see, e.g., Kakouros & Räsänen, 2014).

Probabilistic processing at the level of human cognition is an idea that has held a central role in many models of language processing. Predictability and frequency effects have been widely studied, providing evidence that predictability plays a role in language comprehension, production, and learning (see, e.g., Jurafsky, Bell, Gregory, & Raymond, 2001; Jurafsky, 1996). For instance, predictability of the linguistic elements (such as syllables or words) seems to affect their acoustic realization during speech production (see, e.g., Jurafsky et al., 2001). For example, frequent words are more likely to be reduced in duration than less frequent words. Several theories have emerged in an attempt to explain these probabilistic phenomena, resulting in theories such as the Probabilistic Reduction Hypothesis (Jurafsky et al., 2001) and Uniform Information Density (Frank & Jaeger, 2008). At the level of prosodic prominence, Aylett and Turk's (2004) Smooth Signal Redundancy Hypothesis is based on a similar proposal (linguistic predictability) and suggests that acoustic differences are linguistically implemented through prosodic prominence structure. However, most of the theories focus on examining the predictability of the linguistic units (e.g., phonemes, words) in speech whereas little is known about how the predictability of the low-level or suprasegmental acoustic features affects speech production and perception. Moreover, not all acoustic variation can be explained only by differences in the predictability at lexical or grammatical level. Thus, investigating how the predictability of the acoustic prosodic features might affect different phenomena in speech production and perception is of particular interest.

Previous computational modeling studies reveal that predictability of prosodic trajectories, when measured in terms of a probabilistic prosody model learned from a corpus of speech, is highly correlated with human perception of prominence in the same set of utterances (Kakouros & Räsänen, 2014; 2015b; in press). In the present work, we investigate whether it is possible to induce different prominence perception patterns in human subjects by manipulating the probabilities of different prosodic trajectories during a habituation stage. More specifically, we ask whether the predictability of F0 trajectories affects the perception of sentence prominence. The hypothesis is that low-probability F0 patterns, i.e., the patterns that are less frequent during the habituation stage, would be regarded as more prominent independently of the actual direction of the F0 change.

Experimental setup

We conducted a listening experiment to investigate whether exposure to a certain probability distribution of rising and falling F0 trajectories will affect the listeners' perception of prominence. The aim was to examine 1) whether probabilities of prosodic trajectories have an impact on subjective prominence ratings, and 2) whether a short exposure to prosodic stimuli is sufficient to alter these probabilistic expectations from the baseline perceptual system acquired through life-long experience with spoken language.

The experiment consisted of two conditions: The first, referred to as rising standard condition (RSC), involved the presentation of spoken utterances with 90% of the tokens having rising fundamental frequency (F0) during the sentence-final word while 10% of the sentences had falling F0 during the last word. The second condition, the falling standard condition (FSC), was similar to the first but with the ratio of falling and rising tones inverted with, 90% of the utterances having a falling F0 and the remaining 10% having a rising F0. The subjects in both conditions were first habituated to a 5-minute stream of utterances having the condition-specific F0 distribution while asked to perform an overt task that was designed to ensure that the participants were paying attention to the utterances. After habituation, they were tested in their judgments of prominence on a set of new utterances that also had falling or rising F0 trajectories during the last word.

Stimuli

Speech samples from the CAREGIVER Y2 FI corpus (Altosaar et al., 2010) were used in the study. The style of speech in CAREGIVER is enacted infant-directed speech (IDS) spoken in continuous Finnish, corresponding to a situation where a caregiver is talking to a child, and recorded in high quality in a noise-free anechoic room. The corpus was selected due to the availability of multiple sentences with a simple subject-verb-object (SVO) syntactic

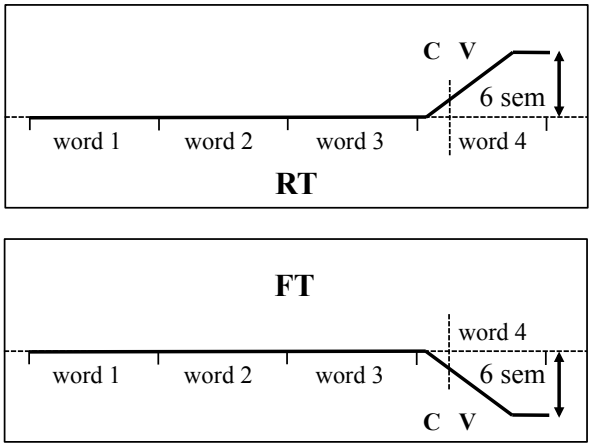


Figure 1: F0 trajectories for the two conditions. Top panel: contour for the rising trajectory (RT), Bottom panel: contour for the falling trajectory (FT). The dashed vertical line represents the vowel onset.

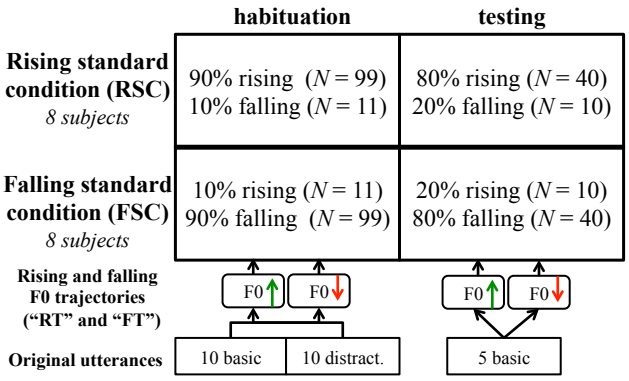


Figure 2: Overview of the experimental setup.

structure and slow speaking rate that allows easier manipulation of the F0 patterns in the sentences. After initial testing, 25 unique utterances from a female speaker (*Speaker 2*) were chosen for the experiments based on the overall naturalness of the stimuli after manipulating the F0 trajectories of the utterances (see below). Each stimulus consisted of a four-word SVO sentence with an average duration of three seconds.

Ten of the utterances were assigned as the main habituation stimuli and five utterances were used solely for the testing stage. The last ten utterances formed a so-called *distractor set* since, despite being grammatically correct, they had very unusual semantic structure (e.g., "*Vauva antaa kulmikkaan koiran*.", Eng: "*The baby gives the square dog*."). The distractors were used as targets in the overt task given to the listeners during habituation (see below).

For each of the 25 utterances, we generated two prosodic versions with either a falling F0 trajectory (FT) or rising F0 trajectory (RT) on the last word using the pitch manipulation functionality available in the Praat software (Boersma & Weenink, 2012). For each stimulus, the original pitch contour was first flattened and the F0 set to 185 Hz (approximately the average across the original stimuli), reflecting the most natural sounding pitch level for the speaker. The F0 trajectory was then modified for the fourth word ("target word") of each stimulus while keeping the rest of the contour flat (Figure 1). Since the primary stress of Finnish always falls on the first syllable of a word (e.g., Suomi & Ylitalo, 2002), the pitch excursion for the falling and rising trajectories was set to start right before the vowel onset of the first syllable of the fourth word (see, e.g., Hermes & Rump, 1994). To ensure consistency, the pitch excursion for all stimuli started 50-ms before the vowel onset and peaked 150-ms later, staying constant for the remaining part of the utterance (Figure 1). As the relation between pitch excursion size and prominence perception may vary, some studies reporting that a difference of 1.5 semitones (Rietveld & Gussenhoven, 1985) or even 4 semitones (Hart, 1981) is required for a perceptually noticeable difference, we preliminarily experimented with a number of different excursion sizes. After assessing excursion sizes between 2 and 8 semitones, we selected 6-semitone change from the flat F0 as the difference producing a clear and most natural-sounding perception of prominence across all stimuli. Thus, the modified trajectories during the target words had ±6 semitone excursion from the flat 185 Hz F0 of all previous words in the utterance (Figure 1). After resynthesizing the utterances with the modified F0 trajectories, all stimuli were amplitude normalized.

Participants

Sixteen native Finnish speakers (9 male, 7 female; average age 28 years) participated in the listening experiment. The test subjects were recruited from the personnel and students of Aalto University. All participants reported normal hearing. The subjects were randomly assigned to the two test conditions with 8 subjects per condition.

Experimental procedure

The listening experiment was conducted in a sound-isolated listening booth of the Acoustics Laboratory of the Aalto University. The habituation and testing software was run on a Mac mini with Matlab 2014b. The audio from the computer was fed through a Motu UltraLite-mk3 Hybrid into a pair of high-quality Sennheiser HD650 headphones.

Participants were given a brief description of the task by the experimenter and they were then asked to start the experiment. The experiment consisted of two parts: (i) a habituation and (ii) a testing stage (see Figure 2 for an overview). During the habituation stage, participants were asked to listen carefully to each utterance being played and press the spacebar whenever they heard a semantically incoherent sentence (the distractor). The role of the overt task was to ensure that the subjects engaged into holistic lexical and semantic processing of the stimuli during the habituation. The subjects were not given specific instructions regarding what counts as a semantically incoherent target, but were instructed to use their own judgment.

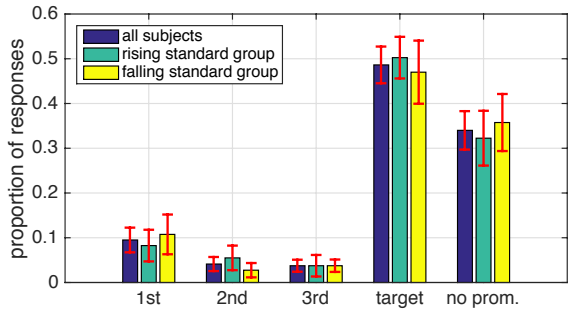


Figure 3: Proportion of responses to different words in the test sentences (relative position), including the no-prominence option. The error bars denote one standard error measured across test subjects (SE).

The only difference between the subject groups, i.e., the RSC and FSC conditions, was the distribution of the rising and falling F0 trajectories during the habituation and testing stage. In the case of RSC (FSC), 90% of the stimuli were RT (FT) ("standards") and 10% were FT (RT) ("deviants"). Ten semantically incoherent targets (distractor stimuli) were interleaved in the data and also followed the same 9:1 ratio. The ordering of the stimuli was randomized for each participant in a manner that each participant in each condition heard each lexically unique training utterance exactly the same number of times: nine times with RT (FT) and once with FT (RT). In addition, they heard each distractor once. The total duration of the habituation stage was 5 minutes, corresponding to a total of 110 utterances with a 500-ms silence interval between each utterance. To avoid repetition of the same utterance in a sequence, the stimuli were presented in blocks of 11 where each block had 10 unique training sentences and one distractor. There were no audible pauses between the blocks.

In the second part of the experiment, the subjects heard the test utterances one-by-one and, for each utterance, they were asked to grade the prominence level of only the single most prominent word on a nominal scale of 0 = no prominence, 1 = slight prominence, 2 = notable prominence. Subjects were allowed to hear each utterance only once in order to facilitate the capture of initial perceptual impressions and in order not to alter the perceived distribution of the pitch trajectories. The tasks of word selection and prominence grading were carried out using a graphical user interface (GUI) where a list of the spoken words was presented together with the prominence scale. Subjects used a mouse as the controller to make the selection. The stimuli distribution in the test stage was set to have 80% of standards (RT for RSC; FT for FSC) and 20% of deviants (FT for RSC; RT for FSC) in order to get more test samples for the deviants than what would be available from the original habituation distribution. The stimuli were presented in blocks of 5 comprising, for the case of RSC, 4 RT and 1 FT. There were 10 blocks of test stimuli, adding up to a total of 50 test tokens per subject. There were no distractor stimuli during the testing stage.

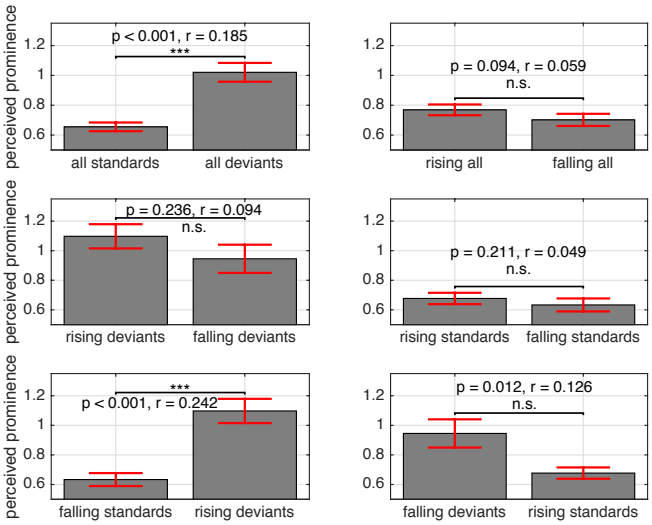


Figure 4: Means and SEs for the perceived prominence (0–2) levels for different stimulus types. Significance and effect sizes are reported using the Wilcoxon rank-sum test and using Bonferroni corrected significance level of $p < 0.0083$.

Results

We first verified how often the subjects labeled utterance-final words, the targets, as prominent instead of other words in the sentences. Figure 3 shows the proportion of responses to each of the four words (relative positions) in the test sentences and no-prominence responses across all subjects and for both sub-groups separately. As can be observed from the figure, the subjects primarily considered the manipulated target word as a prominent or not prominent with less than 18% of the responses marking one of the first three words in the utterances as prominent. In addition, the response strategies between the two subject groups are similar, suggesting that both types of pitch patterns were similarly strong attractors of prominence perception with respect to other competing words in each utterance.

Since the lexical content of the stimuli was fully independent of the pitch trajectory, each subject hearing exactly the same number of repetitions for each sentence, and each sentence occurring exactly the same number of times for both falling and rising pitch across both subject groups, we were able to pool the responses using the standard/deviant criterion, the absolute direction of the pitch pattern, or their combination. Figure 4 shows the full summary of perceived prominence levels across a number of different comparisons.

The results indicate that there is a main effect of condition with the less frequent pitch trajectories (deviants) perceived more prominent (M = 1.02, SD = 0.759) than the standard trajectory experienced during the habituation (M = 0.655, SD = 0.66) independently of the direction of the change. Although the effect is not large, it is highly significant ($p <$ 0.001, Z = 5.44, r = 0.185; Wilcoxon rank-sum test). In contrast, there are no observable differences between the perceived prominence of rising and falling pitch patterns when pooled across both groups, between rising and falling

deviants, or between rising and falling standards ($p > 0.0083$ after Bonferroni corrected for multiple comparisons; see Figure 4 for details). In other words, only the probability of the pitch trajectory during the habituation and testing has an impact on the overall prominence levels.

A closer analysis of both subject groups shows that there are indications for a reversal of preferences between the groups (Figure 4, bottom panel): Subjects habituated with frequent rising pitch consider words with a falling pitch marginally more prominent ($M = 0.945$, $SD = 0.815$ for falling and $M = 0.677$, $SD = 0.613$ for rising; $Z = 2.510$, $p = 0.012$, $r = 0.126$) whereas subjects habituated with predominant falling pitch consider words with a rising pitch significantly more prominent ($M = 1.097$, $SD = 0.695$ for rising versus $M = 0.633$, $SD = 0.705$ for falling; $Z = 4.847$, $p < 0.001$, $r = 0.242$). Although the effects are not large, the reversal of preferences between the two subject groups is clearly seen in the data.

Discussion and conclusions

The present findings suggest that the statistical distribution of prosodic cues can impact subjective perception of word prominence. This is in line with the earlier work that connects the idea of prominence to low-predictability events in the perceptual stream, i.e., "something standing out from a context" (Kakouros & Räsänen, in press; cf. Itti & Baldi, 2009) but contrasts with the idea of prosodic stress being conveyed with certain type of prosodic patterns such as rising or falling pitch contours (see, e.g., Hermes & Rump, 1994). Since there were no lexical or semantic differences between the low- and high-predictability targets, the present results also show that this type of expectation-based prosodic processing occurs in parallel to lexical processing.

From a cognitive point of view, a predictability-based system for prominence (i.e., attentional capture) would be much more flexible than one based on a fixed set of acoustic/prosodic feature detectors for prominent words. First of all, it enables "learning" of prominence perception from language experience, enabling a natural way for different languages to develop strategies for conveying prominence when constrained by the simultaneous production of phonemic contrasts of the language. Similarly, a predictability-based system allows shorter time-scale adaptation to the ongoing communicative situation where factors such as communication channel (is speech partially masked by noise?) or talker-specific idiosyncrasies (acoustic and linguistic characteristics of a specific talker) can lead to very different acoustic prosodic outcomes than what can be characterized at a language-general level. Finally, the predictability framework integrates naturally to the work on statistical learning and information theoretic models at different levels of linguistic analysis (see, e.g., Jurafsky et al., 2001; Frank & Jaeger, 2008; Aylett and Turk, 2004) and language learning (see, e.g., Saffran, Aslin, & Newport, 1996), suggesting that similar basic mechanisms for capturing statistical regularities in the sensory input may be responsible for phenomena at multiple different levels and domains of cognitive behavior.

However, the present findings are preliminary and should be investigated further in a number of additional experiments. For instance, the present setup used an arbitrarily chosen ratio of 1 deviant to 9 standards during habituation, revealing that the subjects are sensitive to such a difference. It is unclear whether the magnitude of apparent prominence is related to the probabilities of the tokens or whether the mechanism is more binary in nature. In addition, the fundamental frequency is only one of many cues to word prominence (see, e.g., Fry, 1955; 1958; Lieberman, 1960; Ortega-Llebaria & Prieto, 2010) and factors such as word position and durational cues also play a role (see, e.g., Luchkina & Cole, 2014). Since we used pre-recorded sentences from a corpora originally designed for other purposes, the speech is typical continuous speech in the sense that there is a high likelihood to have stress on sentence-final words. Despite controlling for pitch and energy, subtle cues such as fine-grained timing in syllabic structure may still be present, here seen as an inherent bias to perceive standard targets also as slightly prominent ($M = 0.655$, $SD = 0.660$, on the scale of 0–2). However, these cues were exactly the same for both of our test groups and cannot affect the group differences.

In all, the current findings provide initial behavioral support for the hypothesis that prominence and unpredictability of the acoustic prosodic features are connected. However, more work is needed in order to confirm this finding and to understand the characteristics and limits of the probabilistic framework in the perception of speech prosody. This also includes extension of the study for other languages than Finnish.

Acknowledgments

This study was funded by the Academy of Finland in the project "Computational modeling of language acquisition", Aalto ELEC Doctoral School, and Nokia Foundation.

References

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation (LREC-2010)* (pp. 1062–1068).

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*, 31–56.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, *119*, 3048–3058.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.13, retrieved from http://www.praat.org/

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*, 1–42.

Campbell, N. (1995). Loudness, spectral tilt, and perceived prominence in dialogues. *Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS-1995)* (pp. 676–679).

Campbell, N., & Beckman, M. E. (1997). Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, & G. Carayiannis (Eds.), *Intonation: Theory, Models, and Applications (Proceedings of an ESCA Workshop)* (pp. 67–70).

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*(2), 425–452.

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*, 1–10.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*, 141–201.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 933–938). Austin, TX: Cognitive Science Society.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*, 126–152.

Hermes, D. J., & Rump, H. H. (1994). Perception of prominence in speech intonation induced by rising and falling pitch movements. *The Journal of the Acoustical Society of America*, *96*, 83–92.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295–1306.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological studies in language*, *45*, 229–254.

Kakouros, S., & Räsänen, O. (2014). Statistical Unpredictability of F0 Trajectories as a Cue to Sentence Stress. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1246–1251). Austin, TX: Cognitive Science Society.

Kakouros, S., & Räsänen, O. (2015a). Analyzing the Predictability of Lexeme-specific Prosodic Features as a Cue to Sentence Prominence. *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1039–1044). Austin, TX: Cognitive Science Society.

Kakouros, S., & Räsänen, O. (2015b). Automatic Detection of Sentence Prominence in Speech Using Predictability of Word-Level Acoustic Features. *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (pp. 568–572). Dresden, Germany: International Speech Communication Association.

Kakouros, S., & Räsänen, O. (in press). Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features. *Cognitive Science,* accepted for publication.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, *118*, 1038–1054.

Luchkina, T., & Cole, J. (2014). Structural and prosodic correlates of prominence in free word order language discourse. *Proceedings of Speech Prosody (SP-2014)*.

Maier, A. K., Hönig, F., Zeißler, V., Batliner, A., Körner, E., Yamanaka, N., & Nöth, E. (2009). A language-independent feature set for the automatic evaluation of prosody. *Proceedings of Interspeech* (pp. 600–603).

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, *32*, 451–454.

Ortega-Llebaria, M., & Prieto, P. (2010). Acoustic correlates of stress in central Catalan and Castilian Spanish. *Language and Speech*, *54*, 1–25.

Pan, S., & Hirschberg, J. (2000). Modeling local context for pitch accent prediction. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 233–240).

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics*, *13*, 299–308.

Rosenberg, A., Cooper, E. L., Levitan, R., & Hirschberg, J. B. (2012). Cross-language prominence detection. *Proceedings of Speech Prosody*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, *25*, 193–247.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*, 2471–2485.

Suomi, K., & Ylitalo, R. (2002). Word stress and duration in Finnish. In C. Dunger, B. Granström, D. House & A. M. Öster (Eds.), *Proceedings of the Swedish Phonetics Conference (Fonetik-2002)* (pp. 73–76).

t Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, *69*, 811–821.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of the Acoustical Society of America*, *89*, 1768–1776.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 53–66). Springer Berlin Heidelberg.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., ... & Vainio, M. (2015). Different Parts of the Same Elephant: a Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS-2015)*. Glasgow, UK: the University of Glasgow.

Integrating identification and perception: A case study of familiar and unfamiliar face processing

Kelsey R. Allen (krallen@mit.edu), **Ilker Yildirim** (ilkery@mit.edu), **Joshua B. Tenenbaum** (jbt@mit.edu)
Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

We are very familiar with certain objects; we can quickly recognize our cars, friends and collaborators despite heavy occlusion, unusual lighting, or extreme viewing angles. We can also determine if two very different views of a stranger are indeed of the same person. How can we recognize familiar objects quickly, while performing deliberate, perceptual inference on unfamiliar objects? We describe a model combining an identity classification network for familiar faces with an analysis by synthesis approach for unfamiliar faces to make rich inferences about any observed face. We additionally develop an online non-parametric clustering algorithm for recognition of repeatedly experienced unfamiliar faces, and show how new faces can become familiar by being consolidated into the identity recognition network. Finally, we show that this model predicts human behavior in viewpoint generalization and identity clustering tasks, and predicts processing time differences between familiar and unfamiliar faces. **Keywords:** face recognition; analysis-by-synthesis; neural networks; computational

Introduction

Walking to work in the morning, we may encounter familiar faces, buildings in which we regularly have meetings, or the passing car of a colleague on their way to the office. Glancing at these objects, we can effortlessly perceive details of their shape and appearance; we can also recall associated identity-specific content (like which colleague owns that car, or the name of that office building). These two abilities can be thought of as object perception and identification respectively. Thus, perception is noticing the shape, texture and expression of a face, even if the person is a stranger. Identification is recognizing a close friend even if she has had a dramatic hair cut and is wearing a new pair of large, dark sunglasses.

Recent work in machine vision has made significant progress on both of these problems, but very different techniques have been applied to each problem. Dramatic gains in object *identification* have come from deep neural networks (Simonyan & Zisserman, 2014). These methods learn to be invariant to certain object transformations and small differences in appearance. However, they require large amounts of training data, and do not generalize to novel objects without at least re-training the top classification layers. Rich object *perception* has become possible using an alternative approach to vision, known as "analysis by synthesis" or "inverse graphics". This approach posits that the perceptual system models the generative processes that form images from scenes, and works backwards from an observed image to infer the scene most likely to have generated it (Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015). Inverse graphics methods can often recover the fine-grained geometrical and physical properties of objects in an image, but are much slower than feed-forward neural networks and have not yielded practical object identification or recognition systems.



Figure 1: How many people are depicted here?

However, object identification and object perception are not separate. For example, extensive research on face perception has studied familiar face recognition, unfamiliar face perception, and the dynamics of how new faces are recognized differently as they become increasingly familiar (O'Toole, Edelman, & Bülthoff, 1998; Burton, Bruce, & Hancock, 1999). Our goal is to build models of these two aspects of vision and their interaction in the domain of faces, and more generally to integrate object perception and identification, learning to see objects differently as they become familiar to us.

There is a wealth of experimental data, including neurophysiological, fMRI, and behavioral studies, investigating the differences between familiar and unfamiliar face processing (Eifuku, De Souza, Nakata, Ono, & Tamura, 2011; Natu & O'Toole, 2011). Many behavioral studies have found dramatic differences in processing, including differences in viewpoint generalization, reaction times for recognition tasks, and a shift from external to internal facial feature processing as faces become more familiar (Johnston & Edmonds, 2009). As a quick example: looking at Figure 1, how many identities do you see? Do you recognize any of the individuals?

One of the early conceptual models seeking to capture some of these behavioral differences was proposed by Bruce and Young (1986). They suggested that face recognition begins with a structural encoding of the face, regardless of familiarity. The structural encoding is compared to stored representations for familiar faces, and a face recognition unit is activated if a similarity threshold is reached. The associated 'person identity nodes' can interface with other identity-specific semantic modules. Quantitative implementations of just the structural encoding aspect of this model include

O'Toole et al.'s RBF model (1998) and others (Leibo, Mutch, & Poggio, 2011).

Moving beyond structural encoding, we implement a modified version of the Bruce and Young system capturing notions of familiarity and identity. Our model suggests that 'person identity nodes' are recognized in a holistic way that depends on learned individual invariances rather than by a comparison of structural encodings. Recognition is accomplished by a 'long term memory' which is represented by a neural network trained to predict identities from face images. These identities are associated with latent parameters describing the 3D structure of that person's face (our structural encoding). With this representation of a person identity node, the set of familiar identities need not be fixed, and can be expanded over time. Thus we provide a computational account of how we become familiar with a new face, which also explains how the processing of familiar faces differs from those we have only seen a few times.

The rest of the paper describes the model in more detail, including its accuracy and inference curves, as well as an online clustering algorithm for unfamiliar faces. We validate the model on three different behavioral experiments, and suggest directions for future research.

Model

Our model represents one way of combining the richness of generative models with the speed of neural networks. Inspired by the Helmholtz machine (Dayan, Hinton, Neal, & Zemel, 1995), we describe an efficient analysis-by-synthesis approach by training a recognition model to approximate the latent parameters of a generative model in a fast, feed-forward way. The approximated parameters provide initializations for top-down inference in a generative model, allowing for some fine-tuning. The generative process, inference procedure, and learned recognition models are described below.

Generative Model

We consider the 3D Morphable Face Model as described in (Blanz & Vetter, 1999). This model is obtained from a set of 200 laser scanned heads, providing a mean shape and texture vector for the eyes, nose, mouth and outline of a face, as well as a covariance matrix to generate new faces by eigendecomposition. The shape and texture are Gaussian distributed, with $N(\mu_{\text{shape}}, \Sigma_{\text{shape}})$ and $N(\mu_{\text{texture}}, \Sigma_{\text{texture}})$.

Each of the shape and texture vectors are 200 dimensional, such that a given face lives in a 400 dimensional latent space. An identity can be thought of as a cluster in this latent representation, with a corresponding mean vector μ_i and isotropic variance Σ. Here Σ has been set to 0.01 to represent perceptually indistinguishable identities. An image can be created by sampling a latent vector for a given facial identity, and rendering it at a specific pose and lighting, as seen in Figure 2b. Figure 3 shows some example faces drawn from this model.

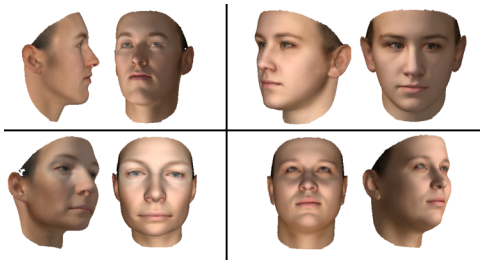


Figure 3: Pairs of images drawn from the generative model.

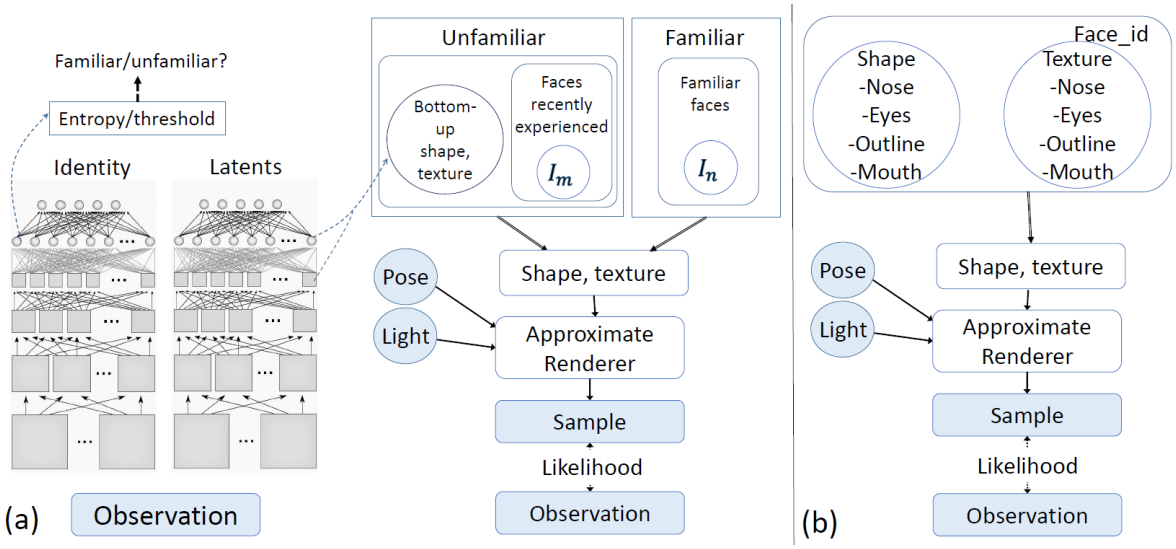


Figure 2: Pipeline used for recognizing an observed image. (a) shows our modified model, using the identity recognition network to determine familiarity, and then initializing with a draw from either the identified familiar cluster or an unfamiliar/new cluster as appropriate. (b) shows the standard generative model operating only over the base distribution in latent space.

Recognition models

Latents recognition network The first recognition network (Figure 2a, network labeled "Latents") is trained to predict the 400 dimensional latent vector that generated a given image of a face. This allows for an efficient, approximate, guess for the latent parameters of a face. The details of this network are described in (Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015). Yildirim et al. use the top convolutional and first fully connected layers of a convolutional neural network (CNN) (pre-trained on ImageNet) to train a linear model to predict the shape, texture, pose and lighting variables of a set of generated faces. Here we assume that pose and lighting are observed, so only shape and texture need to be predicted. Since the generative model created all training data for the network, it is self-supervised. This recognition model will be referred to as the 'latents recognition model'.

Familiar identity recognition network To mimic long-term memory, we include a classification network for familiar identities (Figure 2a, network labeled "Identity"). We use the first fully connected layer of the network (Simonyan & Zisserman, 2014) (also pre-trained on ImageNet) as input to a linear model, which outputs a probability for each familiar identity. Two versions of this network, an 'old' network which knows 80 identities (80 output class labels), and a 'young' network (30 familiar identities/class labels), were trained using 400 different viewing conditions for each familiar identity.

Processing Pipeline

An observed image I_D generated with the Morphable Face Model is fed to both the identity recognition network and the latents recognition network. The system first determines whether this is a familiar person by calculating the entropy across the familiar identities (Figure 2a, entropy/threshold).

Familiar faces If the entropy falls below a learned threshold, the face is classified as familiar and the identity is set to the most probable class. The latent parameters are then initialized by sampling from the stored representation associated with the determined familiar identity, rather than from the general purpose latents recognition network (Figure 2a, Familiar box).

Unfamiliar faces If the entropy falls above the threshold, the face is unfamiliar and we disregard the familiar identities. There are then two possible cases for unfamiliar faces: either the face is completely novel, or it is the same face as one which we have only seen a few times before (Figure 2a, Unfamiliar box). This can be viewed as a non-parametric clustering problem. The very first unfamiliar person we see will generate their own cluster. Each unfamiliar face we see afterwards will either be clustered with a previously encountered identity, or form its own cluster. We therefore model this process using a sequential clustering algorithm with a Chinese Restaurant Process (CRP) prior on cluster assignments for observation i, where n_k is the number of times you have seen person k before:

$$P(k) = \begin{cases} \frac{n_k}{i+\alpha} & (n_k > 0, \text{old cluster}) \\ \frac{\alpha}{i+\alpha}, & (n_k = 0, \text{new cluster}) \end{cases}$$

α is chosen to be 1 for the following experiments, but this choice has little effect on the results.

The likelihood of a specific cluster k for the current observation is computed in image space. We use the generative model to obtain an image from each cluster, rendered at the same pose and lighting as the observed image. While the likelihood for already existing clusters is trivial to compute (Gaussian in pixel space), determining the likelihood of a new cluster is more complex. We approximate it using an image rendered with the latent parameters from the latents recognition network (I_{lrn}). Thus the likelihood can be described by a Gaussian with mean I_k for old clusters, and mean I_{lrn} for the new cluster (and noise $\sigma = 0.01$).

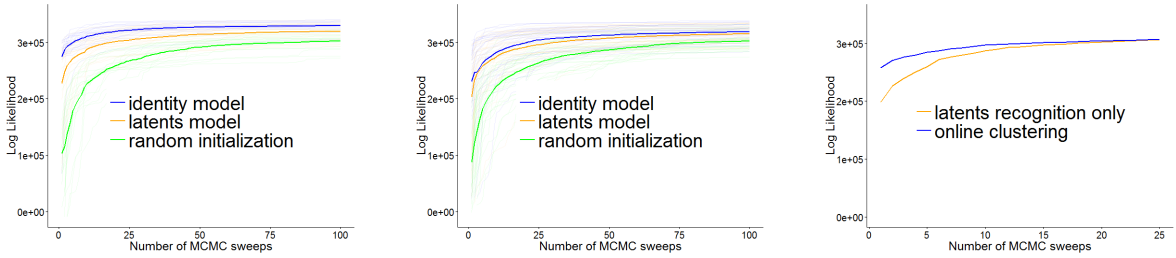
We choose as our estimate the local MAP, which gives us a good initialization for the latent parameters of the new face, even when we are unfamiliar with the observed individual. After forward inference, the cluster means in latent space are updated, reflecting the potential addition of a new cluster member. This learning procedure is the critical contribution of our approach: it presents an account of how we may become familiar with a previously unfamiliar face, even without any supervised training data.

Inference

In order to fine-tune the latent parameters for a given image, we iterate through a few sweeps of forward inference as described in (Yildirim et al., 2015) and (Kulkarni et al., 2015). After initializing the latent parameters for either a familiar face or an unfamiliar face as above, multi-site elliptical slice sampling (Murray, Adams, & MacKay, 2009), a form of MCMC, is performed on the vectors for shape and texture (Figure 2a, Approximate Renderer $\rightarrow$ Observation). At each MCMC sweep, we iterate a proposal-and-acceptance loop on the shape and texture vectors. Proposals are images that are rendered based on a set of latent parameters, a set pose, and a set lighting using a standard graphics engine. The log-likelihood with respect to the observed image is then computed (and assumed to be Gaussian in pixel space).

Simulation experiments

We analyze the performance of our model in several different scenarios. We first generated a set of 100 identities, each of which was rendered under 500 different pose and lighting conditions. We then trained an output layer on the last fully connected layer in the network from (Simonyan & Zisserman, 2014) to predict either a set of 30 identities (the young network) or 80 identities (the old network). For each identity, 400 views were used for training, leaving 100 viewpoints for testing. On the test set, the young network achieves 98.72% accuracy while the old network achieves 98.42% accuracy. Training was performed using stochastic gradient de-



(a) Familiar faces (b) Unfamiliar faces (c) Faces observed once before

Figure 4: Inference traces for (a) 20 familiar (b) 20 unfamiliar and (c) observed once before, faces. The addition of the identity recognition network improves performance for familiar identities, and doesn't hurt performance for unfamiliar faces.

scent with a learning rate of 0.001 and a maximum number of iterations of 1000.

Familiarity classification At the first stage of the pipeline, an incoming face is deemed to be familiar or unfamiliar based on the entropy over the network class labels. To determine an appropriate threshold, we maximize accuracy on a familiar/unfamiliar task using 400 views of 20 familiar faces and 400 views of 20 unfamiliar faces in the **young** network. This results in an accuracy of 91.3% for the young network. Using the same threshold for the old network yields an accuracy of 94.1%. The older network slightly outperforms the younger network, which qualitatively matches the behavioral findings of (Germine, Duchaine, & Nakayama, 2011), who showed that face recognition ability increases with age (up to a certain point).

Inference We check whether including the identity network yields a better initial estimate of the latent parameters for familiar faces compared to random initializations or initializations taken from the latents recognition network. For this experiment, we randomly sampled 20 known and 20 unknown faces rendered at 3 different viewing conditions. Each face was presented to the identity model pipeline as described earlier, but without the added "online clustering" for unfamiliar faces. The resulting log likelihood trajectories are shown in Figures 4a and 4b.

Online clustering 6 identities were chosen, each with a frontal view under random lighting and a 1/4 side view under random lighting. The model was first presented with the 6 frontal views, and correctly made 6 new clusters for these faces. The 6 side views were then presented in scrambled order, and the clustering scheme was able to successfully cluster $4/6$ of the secondary views. The average likelihood traces are shown in Figure 4c.

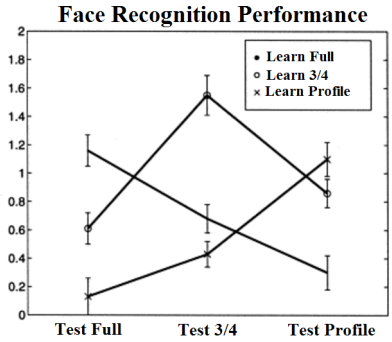
Expanding the set of familiar identities Finally, we tested how well the network consolidated new faces into long-term memory. We sampled 20 views from 3 novel identities, as well as 5 views from each of our previous 30 familiar identities as our training set (which might reflect dreaming of new faces, for example). We initialized the weights of the linear layer in the identity recognition network to those from the previous network for the familiar identities, and randomly initialized the weights for the unfamiliar identities. After training, we achieve 89.84% accuracy on the old faces, and 89.70% on the new faces, reflecting reasonable memory consolidation.

Comparisons with behavioral experiments

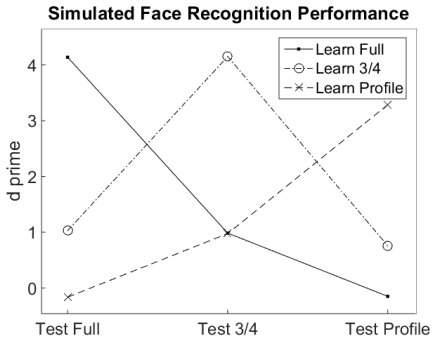
Experiment 1 In the first experiment, we show the power of the unfamiliar face processing component of the model by reproducing results from O'Toole et al. (1998). In this experiment, participants were trained on 36 unfamiliar faces from one of three views: frontal, 3/4 or profile. Participants were then shown 72 images from any of the three viewing conditions, and asked to classify each image as depicting an individual in the training set ('old'), or a new individual ('new'). D prime measures were then calculated for each individual in the task.

We simulated this task by collecting 36 random identities using the Morphable Face Model. Each individual was rendered under a profile, 3/4 and frontal view, with identical lighting conditions. We then used the latents recognition network to predict the latent parameters for each of the 36 unfamiliar faces, with all faces shown in the same viewing condition (either profile, 3/4 or frontal). This results in 36 distinct clusters for 36 individuals. In the test phase, the model observes a face at one of the three views. We then compute the likelihood for each of the 36 learned identities (in pixel space) by rendering its associated latent parameters in the same pose as the test image. These are compared to the likelihood computed with the latents predicted from the latents recognition network. If the likelihood of one of the learned clusters is higher than the likelihood from the latents recognition network, the face is classified as 'old'. Otherwise, it is classified as 'new'. The results from both the psychophysics experiment and from our simulations are shown in Figure 5.

Overall, our model is much more accurate within viewpoints than humans on this task, but follows the general trend for viewpoint-transfer generalization. The model provided by O'Toole in the paper (based on radial basis functions) also predicts this trend, although the old/new classification task would need to incorporate a learned threshold (with one free parameter), which we do not need. The most major discrep-



(a) Experimental results from (O'Toole et al., 1998)



(b) Model results

Figure 5: Results from experiment performed by O'Toole et al. showing viewpoint generalization compared with results from our model.

ancy lies in the relative inability of our model to generalize from a 3/4 view to a profile view. This may be mitigated by running a few sweeps of forward inference during the training phase (in the model) to more accurately determine latent parameters for faces viewed from the 3/4 and frontal views.

Experiment 2 We next show that our model can account for differences in processing speed for familiar and unfamiliar faces, even in very easy tasks with near ceiling performance. Balas, Cox, and Conwell (2007) performed a delayed match-to-sample task for identity, where participants were cued with a profile of either a familiar or unfamiliar person, and then asked to choose which of two individuals (shown at either a 3/4 or frontal view) matched the cue. They found that reaction times for personally familiar individuals was approximately 100 ms faster than for unfamiliar individuals, even though performance in both conditions was above 95%.

It is not obvious how models that rely on stored viewpoints for both familiar and unfamiliar faces could account for this difference in processing speed. Thus, the RBF models presented by O'Toole or those provided by Leibo et al. (2011) do not immediately explain the results of this experiment.

Our model can account for this difference regardless of whether a likelihood measure for the unfamiliar case is done in image or latent space. In latent space, the perceptual system may require a certain confidence in the latent parameters of a face before making a judgment. Therefore, in the test phase, if the images are detected as unfamiliar, they will need more MCMC inference steps to achieve the same likelihood as in the familiar case.

Alternatively, if likelihood is computed in image space, the pose and lighting of the test face need to be inferred (which requires at least one pass through the latents recognition network). The cue face must then be rendered in the appropriate viewpoint in order to compute likelihoods. This extra projection step could account for the longer reaction times.

To ensure that our model achieves comparable accuracy for this task, we trained it on 80 familiar faces, and then randomly chose 9 of these as well as 9 unfamiliar faces for the experiment. The model is first shown a cue face (in the profile view), and then shown two faces during the test phase. The test images are shown in either frontal or 3/4 view. As in the original experiment, each identity is used as a cue 4 times, giving 36 trials for the familiar cases and 36 for the unfamiliar. For 8/9 familiar identities, the model correctly identifies the individual and classifies both the cue image and one of the two test images as familiar. This requires only two passes through the feed-forward identity recognition network. In the last case, the cue was classified as unfamiliar, but the correct judgment was still made when choosing the image with the highest likelihood to the cued face.

For the unfamiliar faces, the model correctly matched the cued individual on each run. It also correctly classified every cued image as "unfamiliar" which meant that a projection back to image space was performed. In future work, we will run further controlled experiments looking at reaction times, and quantitatively compare the model's performance to human performance in recognition tasks for familiar and unfamiliar individuals.

Experiment 3 In this experiment, Jenkins et al. showed that there are massive differences between familiar and unfamiliar face recognition by asking participants to cluster images of two famous Dutch actresses (20 images for each individual) into identities (Jenkins, White, Van Montfort, & Burton, 2011). The experimenters did not specify how many identities were depicted in the collection. Strikingly, they found that participants who were unfamiliar with the actresses clustered the set of images into 6-10 identities (mode 9, range 3-16), while those who were familiar with the actresses correctly clustered the space into two individuals (mode 2, range 2-5). Interestingly, the rate of misidentification (ie. sorting the two different individuals into the same pile) was very rare for both groups.

We simulated this task in a sequential clustering experiment, with two individuals rendered under 20 different viewing conditions each. We used our 'old' network which was trained on 80 identities, but varied whether or not the two individuals were included in the training set (giving an unfamiliar condition as well as a familiar condition).

In the unfamiliar condition, the model created 5 distinct clusters for the two identities with memberships of 7, 3, 9, 7

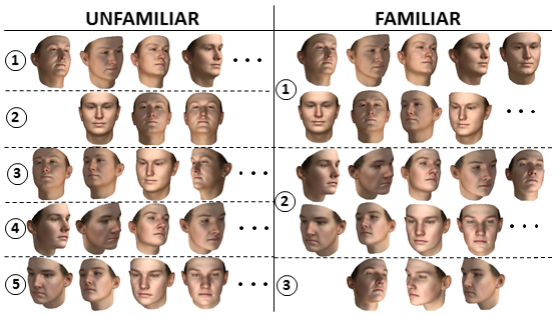


Figure 6: Clusters discovered by the model in the unfamiliar condition (left) and familiar condition (right).

and 12 images (selections from clusters are shown in Figure 6), while 2 faces were incorrectly classified as familiar (into two separate identities). In the familiar condition, the model correctly identified the first individual (with all 20 images being classified as the correct familiar person), while mostly correctly classifying the second individual (with 17/20 images). There were three minor misclassifications for the second individual, resulting in a third cluster being formed.

These results seem to reflect those found by Jenkins et al. (2011). Namely, the model also over-clusters the space when the identities are unfamiliar, but makes only three mistakes when the identities are familiar. In the latents recognition network, the inferred latents depend substantially on pose and lighting, and thus are not successfully ignored in clustering unfamiliar faces. In the familiar network, these invariances are successfully learned, allowing for accurate clustering. Additionally, matching humans, the model never forms clusters which have images from the two different identities in either the familiar or unfamiliar conditions.

Discussion

The model that we have described is both computationally powerful, and also qualitatively and quantitatively captures human behavior across a wide range of different experiments. To our knowledge, this is the first model that learns new identities in both an unsupervised and supervised way, and can account for both effects of familiar and unfamiliar face recognition.

First, we showed how the unfamiliar component of the model can predict the patterns of viewpoint generalization found by O'Toole et al. (1998), even with no explicit viewpoint dependence built in. Although the model is 3 dimensional, the reconstruction accuracy is constrained by the fact that there may be multiple sets of generative parameters that give rise to the same 2D view, which leads to this viewpoint dependent generalization. Second, we show how recognition of familiar faces can proceed significantly faster than for unfamiliar faces, predicting the experimental results from Balas et al. (2007). We discussed how this could result either from a comparison in the latent space (where a good estimate of the latents may be required, and the estimates get better faster for familiar faces) or by a projection back to image space, which is only necessary for unfamiliar faces. This cannot be immediately predicted by standard view-dependent models. Third, we show a major difference in the processing of familiar and unfamiliar identities by replicating the findings of Jenkins et al.'s clustering experiment (albeit in an online fashion). Our model over-clusters the space when faces are unfamiliar, but correctly clusters the space (with minor errors) when the faces are familiar.

We propose that this framework presents a general way of integrating identification and perception. One future line of work will investigate whether the same architecture might be applied to familiar and unfamiliar objects. We also plan to examine other methods of non-parametric clustering, run experiments using the face stimuli we generated, and do a more thorough model comparison. We will also investigate how much exposure you need with an individual in order for them to be consolidated in long-term memory, and whether or not this requires sleep.

References

Balas, B., Cox, D., & Conwell, E. (2007). The effect of real-world personal familiarity on the speed of face information processing. *PloS One*, *2*(11), e1223.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Burton, A. M., Bruce, V., & Hancock, P. J. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science*, *23*(1), 1–31.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, *7*, 889–904.

Eifuku, S., De Souza, W. C., Nakata, R., Ono, T., & Tamura, R. (2011). Neural representations of personally familiar and unfamiliar faces in the anterior inferior temporal cortex of monkeys. *PLoS One*, *6*(4), e18913.

Germine, L., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*, 201–210.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. (2011). Variability in photos of the same face. *Cognition*, *3*, 313–323.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*(5), 577–596.

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: An imperative probabilistic programming language for scene perception. *Computer Vision and Pattern Recognition*.

Leibo, J. Z., Mutch, J., & Poggio, T. (2011). Why the brain separates face recognition from object recognition. In *Advances in neural information processing systems* (pp. 711–719).

Murray, I., Adams, R. P., & MacKay, D. J. (2009). Elliptical slice sampling. *arXiv preprint arXiv:1001.0175*.

Natu, V., & O'Toole, A. J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology*, *102*(4), 726–747.

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision research*, *38*(15), 2351–2363.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Yildirim, I., Kulkarni, T., Freiwald, W., & Tenenbaum, J. (2015). Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations. In *Proceedings of the thirty-seventh annual conference of the cognitive science society.*

On the Link between Fact Learning and General Cognitive Ability

Florian Sense (f.sense@rug.nl)
Department of Experimental Psychology & Department of Psychometrics and Statistics
Groningen, The Netherlands

Rob R. Meijer (r.r.meijer@rug.nl)
Department of Psychometrics and Statistics
Groningen, The Netherlands

Hedderik van Rijn (d.h.van.rijn@rug.nl)
Department of Experimental Psychology & Department of Psychometrics and Statistics
Groningen, The Netherlands

Abstract

Adaptive fact learning systems have been developed to make optimal use of testing and spacing effects by taking into account individual differences in learning efficiency. Measures derived from these systems, capturing the individual differences, predict later performance in similar *and* different fact learning tasks. Additionally, there is a rich body of literature showing that individual differences in general cognitive ability or working memory capacity can predict scores on achievement tests. If these measures also influence fact learning, incorporating them might further enhance adaptive systems. However, here we provide evidence that performance during fact learning is neither related to working memory capacity nor general cognitive ability. This means that the individual differences captured by our adaptive learning system encapsulate characteristics of learners that are independent of their general cognitive ability. Consequently, adaptive learning methods should focus primarily on memory-related processes.

Keywords: learning; memory; working memory capacity; general cognitive ability; fluid intelligence; individual differences; computational modeling

Introduction

Research has shown that standardized measures of acquired knowledge – such as the Scholastic Aptitude Test – are among the best predictors of success in life (e.g., Kuncel & Hezlett, 2007, 2010). With an ever-growing body of knowledge, life-long learning has become a reality for many and acquiring new knowledge efficiently is paramount. Computerized learning systems are developed to streamline the acquisition of new knowledge and the most successful ones stand out because they adapt to the individual characteristics of the learner. What is not clear, however, is whether only individual differences in memory-related processes are relevant to optimize the adaptation or whether individual differences in general cognitive ability should be taken into account as well. Here, we present data that suggest that individual differences in general cognitive ability are not required to optimize the fact-learning process using the model developed in our lab.

Few findings in psychology are as reliably reproduced as the spacing effect (Donovan & Radosevich, 1999), the finding that learning yields better long-term results if repetitions are spaced over time rather than crammed together. The spacing effect holds over various time scales (e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008), indicating that it reflects a fundamental property of the human memory system. Therefore, the spacing effect should be exploited when we want to optimize fact learning (Dempster, 1988). Pavlik and Anderson (2003) extended ACT-R's declarative memory module to account for spacing effects and subsequently made first steps in using the decay-based model of human memory to enhance the learning of facts (Pavlik & Anderson, 2005, 2008). Their model uses trial-by-trial information gathered *during* learning to devise personalized learning schedules on the fly.

Van Rijn and colleagues (2009) refined the model further but the underlying principles are still the same: each item is assigned an activation value when it is first presented to the learner. The activation decays over time and the model strives to schedule a presentation of the item before the activation is too low, which would prevent a successful retrieval from memory. Throughout the learning session, the estimation of each item's activation is continuously fine-tuned based on the learner's response times to quiz-items. Specifically, on each trial, the model's predicted activation can be converted to an *expected* response time, which is compared to the *observed* response time. The discrepancy between the two is used to update the estimated decay rate of the activation. This way, one decay-related parameter is estimated for each item for each learner based on both the accuracy and speed of the response. By averaging across the item-specific parameters, we can compute a value that indicates how quickly, on average, a learner forgets the items in a particular set. We will refer to this value as the *rate of forgetting* (for detailed descriptions of the model see Sense, Behrens, Meijer, & Van Rijn, 2016; Van Rijn et al., 2009).

Nijboer (2011) has shown that the *rate of forgetting* estimated during learning of one set of items is very strongly related to subsequent test performance for the same set of items and the data presented here confirm this strong relationship. More recently, we showed that the *rate of forgetting* is stable over time within material from one topic, and relatively stable over materials from different topics, which indicates that the *rate of forgetting* can be estimated

reliably (Sense et al., 2016). That is, if the *rate of forgetting* for a learner is estimated while studying a set of Swahili words, the *rate of forgetting* for another set of Swahili words will be almost identical when estimated a week later. There is a bit more variation if material from another domain is studied but the correlation is still high. This variation might be caused by differences in difficulty of the studied materials or because learners' ability to learn different types of materials varies. The model itself is agnostic with regards to the source of the variation between individuals and materials (Sense et al., 2016).

Taken together, this suggests that someone's *rate of forgetting* is a useful and reliable measure of individual differences. What is not clear, however, is what exactly the *rate of forgetting* encapsulates. Given the parameter's roots in ACT-R's declarative memory module (Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson, 2007), one would expect that it captures memory-related processes. However, for a participant to do well in a laboratory simulation of a fact-learning session, more than the memory processes modeled by ACT-R's decay functions might be involved. Alternative individual difference measures are commonly used to study individual differences and have been shown to have high predictive power in various aspects of life (Kuncel & Hezlett, 2010). The goal of the present study is to investigate how the *rate of forgetting* relates to established measures of individual differences.

Two of the most widely used measures of individual differences are working memory capacity (WMC) and general cognitive ability or fluid intelligence (gf). There are various ways of conceptualizing and measuring both concepts. Complex span tasks are commonly used to measure WMC (Conway et al., 2005) and there are multiple standardized tests to assess general cognitive ability (e.g., the Wechsler Adult Intelligence Scale, WAIS). And while the two concepts are strongly related to each other (Ackerman, Beier, & Boyle, 2005), they are not identical (Conway, Kane, & Engle, 2003; Kane, Hambrick, & Conway, 2005).

Kane and colleagues (2007) define WMC as attentional processes that enable goal-directed behavior. Conway and colleagues (2005) recommend to administer three complex span tasks and conceptualize WMC as a composite score across those tasks. This way, task-specific components are partialled out and the derived score expresses domain general attentional processes (Kane et al., 2007). Consequently, WMC scores obtained with complex span tasks primarily reflect general executive processes (the tasks may vary but the objective is similar in other approaches, e.g., Cowan et al., 2005). Such WMC scores share variance with measures of tests of general cognitive ability because they, too, require superior executive attentional processes to obtain high scores (Engle, Tuholski, Laughlin, & Conway, 1999).

The goal of the current study is to shed light on how the *rate of forgetting* extracted from our model is related to these two measures of individual differences. A strong relationship between *rate of forgetting* and either or both WMC and general cognitive ability would suggest that the model's parameter encapsulates an executive process that is akin to what allows individuals to perform well on tests of WMC or general cognitive ability. If there were no such relationship, however, we would conclude that the *rate of forgetting* reflects a measure of individual difference that tells us something about a learner's memory retention capacities that goes beyond how well they can use their executive-attentional resources efficiently.

To this end, we report the findings from an experiment in which participants studied a set of facts so their *rate of forgetting* could be estimated. Additionally, they completed three complex span tasks (analogous to Foster et al., 2015) as well as a test of general cognitive ability.

Methods

Procedure

All participants were invited for two sessions that were spaced three days apart.

Session 1. In the first session participants spent 20 minutes learning 35 Swahili-Dutch word-pairs. Participants were randomly assigned to study with one of two methods: either they used digital flashcards or an adaptive learning method. As we will focus here on the results of the adaptive learning method, we will refrain from further discussion of the digital flashcard method. During learning, words were introduced on *study trials* which showed both the cue (Swahili word) and the correct response (Dutch word) alongside an input field. Initial *study trials* were self-paced and participants proceeded by typing in the Dutch word. All subsequent repetitions of an item were *test trials* which only displayed the cue and the input field. *Test trials* were followed by feedback: either a 600 *ms* display saying "correct" or a four second display of a *study trial* without the input field (as recommendd here: Zeelenberg, de Jonge, Tabbers, & Pecher, 2015).

Next, participants completed the three complex span tasks used by Foster and colleagues (2015). In these tasks, participants are shown items that need to be recalled in the correct order at the end of trial. Each to-be-remembered item is followed by a distractor, which requires the participant to engage executive attentional processes. This is to reduce the ability to rehearse to-be-remembered items during the distractor task. In the Operation Span task, for example, to-be-remembered items are letters and distractors are simple equations (e.g., (2 x 2) - 1 = 3), which the participant has to make true/false judgments about. The order of the tasks was identical across participants (as in Foster et al., 2015): first Operation Span, followed by Rotation Span, and then Symmetry Span.

Finally, a test of the word-pairs that were studied at the beginning of the session was administered. All 35 Swahili cues were shown on screen as a list and the participant had to provide the correct Dutch translation. The test was self-paced and because all words were visible at the same time,

participants were able to provide answers in any order they preferred. No feedback was provided.

The duration of complex span tasks varies between participants. To ensure that the retention interval between the word-learning task and the test was the same across participants, a simple lexical decision task was administered as a filler task before the test. The task was setup in a way that it would terminate as soon as the retention interval was 80 minutes, irrespective of the number of trials completed. For the task, five-letter strings were presented on screen and participants had to press one of two buttons to indicate whether the string was a Dutch word or not. By using high frequency words, the task was made relatively easy to avoid fatigue. All but two participants maintained accuracy levels above 80% and visual inspection of the response time distributions suggests the task was performed consistently. The task was chosen because it was easy to check on a trial-by-trial basis whether participants engaged in the task, and could easily be programmed to ensure that all participants started with the subsequent test at the same, relative, point in time. The data from the filler task will not be discussed further here.

Session 2. Three days later, participants came back for the second session. The second session started with a second test of the Swahili-Dutch word-pairs learned at the beginning of the first session. The test was identical to the one completed at the end of Session 1.

Subsequently, we assessed the participant's general reasoning abilities by administering the Q1000 Cognitive Capacity test on-line. Upon completion of the test, the website provided participants with feedback indicating how their performance (overall and on the sub-scales) compared to that of a norm group.

Materials

Swahili-Dutch Word-Pairs. The 35 items were randomly sampled from the list of 100 Swahili-English word-pairs provided by Nelson and Dunlosky (1994). The English responses were translated to Dutch and all participants studied the same subset of 35 word-pairs. The order in which words were introduced was randomized.

Complex Span Tasks. The code for the three complex span task was obtained from the Engle lab's website and used with their permission. It is the same code used by Foster and colleagues (2015) but all instructions were translated to Dutch. Scores reported in Table 1 are partial-credit unit scores (Conway et al., 2005).

General Cognitive Ability. As a measure of general cognitive ability, we used *Q1000 Capaciteiten Hoog* ("High Capacity"; university-educated individuals) developed by Meurs HRM. The test has been developed for a selection context to determine whether a candidate has the necessary intellectual ability to perform well in cognitively demanding jobs, but has psychometric properties akin to other standardized tests of intelligence. There are multiple sub-scales that are ordered hierarchically with the declared goal of measuring general intelligence. In contrast to more traditional tests of general cognitive ability, this test can be administered, using online tools, in a classroom setting. The Committee on Test Affairs Netherlands (COTAN) has evaluated the test and concluded that it is a valid and reliable measure of general cognitive ability (Van Bebber, Lem, & Van Zoelen, 2010). All scores reported here are z-scores relative to the highest available norm group ("WO"; people that completed university education).

Participants

A total of 42 participants were recruited from the Dutch first-year participant pool at the University of Groningen. Of those, 14 were female (33%) and the median age was 19 (SD_{age} = 1.34; $range_{age}$ = [17, 24]). No one indicated any familiarity with Swahili. All participants gave informed consent and the Ethics Committee Psychology approved the study (ID: 15006-N).

Due to technical issues, data in the Rotation Span task was lost for one participant and in the Symmetry Span task for another. The composite scores (i.e., WMC) for these two individuals are based on the z-score average of the two remaining tasks. One participant did not complete the second vocabulary test and Q1000 scores were not available for 4 participants because the university was closed due to extreme weather on the day of the second session. They did complete the second test online, though.

Results

To express a single measure of working memory capacity (WMC), the scores on the three complex span tasks are summarized into a single composite score[1]. This is done by calculating a participant's z-score for each task and then computing a z-score average (following Foster et al., 2015). Table 1 provides descriptive statistics for the scores on the individual tasks as well as their partial correlations among each other and with the resulting composite score. As expected, the complex span tasks correlate with each other and are highly correlated with the composite score. All correlation coefficients differ significantly from 0 with $p < .001$[2]. For brevity's sake, the composite score will be referred to as a participant's WMC.

There is considerable variation, both across items and participants, in the estimated parameters that are used to compute the *rate of forgetting*. This indicates that some words are more difficult to learn than others and that some participants learned the material more easily than others. As described in the Introduction, the *rate of forgetting* is obtained by computing the average across all item-specific parameters estimated by the model. Across the 42

[1] A supplement with the raw data and scripts to compute all numbers and generate the plot in this manuscript is available at: https://github.com/fsense/cogsci-2016-paper

[2] Bayesian equivalents (Wetzels & Wagenmakers, 2012) were computed and ranged from 90 to 2327 in favor of the alternative model (that $r \neq 0$) for correlation coefficients among the three tasks and were all over one billion for the three coefficients listed in row four of Table 1.

participants, the mean *rate of forgetting* is .288 with a standard deviation of 0.049 (range = [0.186; 0.404]), with higher values indicating faster forgetting. The distribution is also apparent in Figure 1 and indicates that there are considerable individual differences in the *rate of forgetting*.

Table 1. Descriptive statistics for the complex span tasks and their composite score (WMC) as well as the correlations between all measures.

	Mean	**SD**	**Range**	1.	2.	3.
1. OSpan	59.5	9.2	[39, 75]			
2. RotSpan	29.0	7.3	[10, 40]	.56		
3. Symspan	31.4	6.9	[14, 42]	.64	.54	
4. WMC	0.0	0.8	[-1.6, 1.3]	.86	.83	.85

Figure 1 provides an overview of all relevant measures and how they relate to each other. The plot depicts the distribution of each variable on the diagonal. On the off-diagonal, scatterplots with fitted linear regression lines are shown. The corresponding partial correlations are shown on the other off-diagonal, along with *p*-values expressing the probability of observing the data assuming the coefficient is 0. Also shown are the sample size and the Bayesian equivalent of the null-hypothesis significance test (Wetzels & Wagenmakers, 2012). The subscript "H0" indicates that the Bayes factors quantify the evidence the data provides for the null hypothesis (assuming no correlation, i.e., that $r = 0$) relative to the alternative hypothesis and vice versa for subscript "H1".

Participants took one test of the learned word-pairs in both sessions but only the test results from the second test are included in Figure 1. The scores from the two tests are highly correlated (r = .88) and interchangeable[3] when it comes to the conclusions that can be drawn from Figure 1.

Since a higher *rate of forgetting* indicates faster forgetting, we would expect all correlations in the left-most column of Figure 1 to be negative. This expectation is met by the performance on the second test; forgetting items more slowly is strongly related to performing well on the test. While the signs of the correlation coefficients corresponding to working memory capacity (WMC) and general cognitive ability (GCA) are also negative, both the *p*-values and the Bayes factors suggest that we can assume the correlations are zero. All other correlation coefficients have the expected positive sign but the data only provide evidence for a non-zero correlation between WMC and GCA.

Thus, the only measure that is related to test performance is the *rate of forgetting* estimated during learning, while the individual differences captured by WMC/GCA and the *rate of forgetting* do not seem to share any variance.

[3] See the Supplement for how the numbers change if scores from the first test are used. The scores were lower, on average, on the second test as one would expect. But the general trend and distribution of the data does not differ much between the two tests.

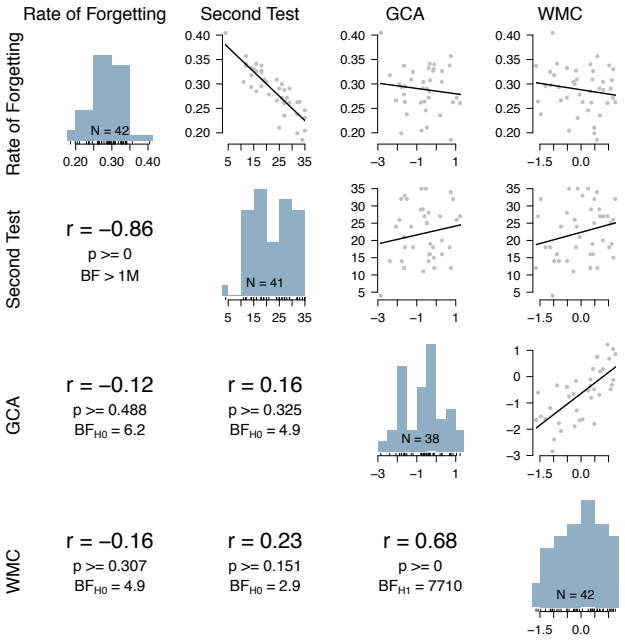


Figure 1. Depicted are the measures of interest: the estimated rate of forgetting, the score on the second test, general cognitive ability (GCA), and their working memory capacity (WMC).

Discussion

The goal of the present study is to investigate how the *rate of forgetting* estimated during learning with our adaptive fact-learning model relates to established measures of individual differences.

The very high correlation between *rate of forgetting* and test scores (see Figure 1) suggests that the parameter estimated during learning captures the learner's ability to store and retrieve information from memory. Vocabulary tests are a common way to assess word-pair learning in many school curricula, which gives face validity to the test administered here. Being able to predict test performance with 74% accuracy (the square of r = -.86; see Figure 1) is not only impressive but also denotes that the estimated parameter captures meaningful individual differences. We have shown recently that the *rate of forgetting* estimated when studying one topic is highly correlated with that estimated for another topic – even though we picked the materials to be distinct (Sense et al., 2016). This lends indirect support to the *rate of forgetting* as a robust individual difference measure but a direct test needs to be conducted to verify this idea.

We could neither establish a correlation between the *rate of forgetting* and someone's WMC nor their general cognitive ability. The lack of a correlation suggests that the *rate of forgetting* shares very little to no variance with the performance on either complex span tasks or a test of general cognitive ability. Since both these tasks make strong executive attention demands (Kane et al., 2007), we

conclude that the *rate of forgetting* encapsulates a characteristic of the learner that is not analogous to their general cognitive ability.

Earlier work provides additional evidence for the view that the *rate of forgetting* might be independent from cognitive span measures. For example, Rosen and Engle (1998) had people learn three lists of paired-associates. For half the participants, the second list re-used cues from the first list but participants had to associate them with different responses. This created interference and they were interested in whether low and high span participants were affected by the experimental manipulation differently. They found that low and high span participants learned the first list at the same rate, when just the *rate of forgetting* plays a role, but that low span participants' performance suffered a lot more from the interference than high span participants'. Their experiments suggest that high span participants perform better in the interference condition because they ward off the intrusions more successfully (also see Brewin & Beaton, 2002; however, see Oberauer, Lange, & Engle, 2004 for contradicting evidence). No such interference existed in our experiment and the finding that someone's WMC is not linked to their *rate of forgetting* is consistent with the finding that low and high span participants learn paired-associates at the same rate (Rosen & Engle, 1998). Similarly, Kane and colleagues conclude that WMC measured with complex span tasks "does not predict variability in *all* aspects of remembering" (2007, p.32) even though there is research that suggests an influence of WMC on memory-related processes (e.g., Mall & Morey, 2013).

Although the Q1000 is not a commonly used test to assess general cognitive ability, this study provides further, indirect, validation. Working memory capacity (WMC) and general cognitive ability are known to be related concepts, and in this study we find a correlation in the expected range (Ackerman et al., 2005). Nevertheless, it would be good to replicate the current findings with more commonly used tests of general cognitive ability (e.g., WAIS or Raven's advanced progressive matrices).

Based on the data presented here, we can extrapolate that the individual differences captured by someone's *rate of forgetting* are not confounded by their general cognitive abilities. We believe, however, that this relationship might emerge if a more heterogeneous sample is used. Our participants were first-year psychology students at a Dutch university, which necessarily restricts the range of general cognitive ability in an absolute sense. Figure 1 shows that there is a reasonable amount of variation in the data when it comes to general cognitive ability but this is due to the fact that the test we used is highly sensitive in the higher range of the construct it measures. Therefore, it is not unreasonable to assume that a relationship between someone's ability to learn factual knowledge and their general cognitive ability might exist in the general population. These results suggest, however, that in more restricted ranges, as observed in the population commonly used in psychological research, WMC and general cognitive ability do not influence fact learning.

Conclusion

The data presented here confirm that the *rate of forgetting* estimated by the adaptive fact-learning model developed in our lab is an excellent predictor of outcomes on tests of the same material. While the common measures of individual differences used in this study - working memory capacity and general cognitive ability - are related to each other in the way we would expect, they do not seem to be related to the estimated *rate of forgetting*. This is interesting because it suggests that the rate with which someone acquires and retains factual knowledge is not linked to their executive attentional capabilities.

Acknowledgments

We would like to thank Susan Niessen for her invaluable help in planning the experiment, collecting the data, and all aspects related to the Q1000 test. We would also like to thank Atser Damsma, Anna Leonte, and Rob Nijenkamp for their help with the translations and Lukas Preis and Ron Woytaszek for their help with data collection.

A supplement with the raw data and scripts to generate the plot and numbers in this manuscript is available at: https://github.com/fsense/cogsci-2016-paper

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: the same or different constructs? *Psychological Bulletin*, *131*(1), 30–60. doi:10.1037/0033-2909.131.1.30

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An Integrated Theory of List Memory. *Journal of Memory and Language*, *38*(4), 341–380.

Brewin, C. R., & Beaton, a. (2002). Thought suppression, intelligence, and working memory capacity. *Behaviour Research and Therapy*, *40*, 923–930. doi:10.1016/S0005-7967(01)00127-9

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–102. doi:10.1111/j.1467-9280.2008.02209.x

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. doi:10.3758/BF03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. doi:10.1016/j.tics.2003.10.005

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. doi:10.1016/j.cogpsych.2004.12.001

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. doi:10.1037//0021-9010.84.5.795

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology. General*, *128*(3), 309–331. doi:10.1037/0096-3445.128.3.309

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. doi:10.3758/s13421-014-0461-7

Kane, M. J., Conway, A. R. A., Hambrick, David Z, & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, & A. T. Miyake (Eds.), *Variation in working memory* (pp. 21–46). New York, NY: Oxford University Press. doi:10.1093/acprof

Kane, M. J., Hambrick, D. Z., & Conway, A. R. a. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71; author reply 72–75. doi:10.1037/0033-2909.131.1.66

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315*, 1080–1081. doi:10.1126/science.1136618

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and Fiction in Cognitive Ability Testing for Admissions and Hiring Decisions. *Current Directions in Psychological Science*, *19*(6), 339–345. doi:10.1177/0963721410389459

Mall, J. T., & Morey, C. C. (2013). High working memory capacity predicts less retrieval induced forgetting. *PloS One*, *8*(1), e52806. doi:10.1371/journal.pone.0052806

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory*, *2*(3), 325–335.

Nijboer, M. (2011). *Optimal fact learning: Applying presentation scheduling to realistic conditions*. University of Groningen.

Oberauer, K., Lange, E., & Engle, R. W. (2004). Working memory capacity and resistance to interference. *Journal of Memory and Language*, *51*(1), 80–96. doi:10.1016/j.jml.2004.03.003

Pavlik, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In *Proceedings of the 5th International Conference on Cognitive Modeling* (pp. 177–182). Bamberg.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–86. doi:10.1207/s15516709cog0000_14

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–17. doi:10.1037/1076-898X.14.2.101

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language*, *39*, 418–436. doi:10.1006/jmla.1998.2590

Sense, F., Behrens, F., Meijer, R. R., & Van Rijn, H. (2016). An Individual's Rate of Forgetting is Stable over Time, but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. doi:10.1111/tops.12183

Van Bebber, J., Lem, J., & Van Zoelen, L. (2010). *Q1000 Capaciteiten Hoog*. Woerden: Meurs HRM.

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115).

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057–1064. doi:10.3758/s13423-012-0295-x

Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2015). The effect of presentation rate on foreign-language vocabulary learning. *Quarterly Journal of Experimental Psychology*, *68*(6), 1101–15. doi:10.1080/17470218.2014.975730

The paradox of relational development: Could language learning be (temporarily) harmful?

Christian Hoyos (choyos@u.northwestern.edu)
Ruxue Shao (ruxueshao2018@u.northwestern.edu)
Dedre Gentner (gentner@northwestern.edu)
Northwestern University
Department of Psychology, 2029 Sheridan Rd.
Evanston, IL 60208 USA

Abstract

Recent studies report a striking decline in children's ability to notice *same-different* relations around age 3 (Walker et al., 2015). We propose that such a decline results from an object focus related to children's avid noun-learning. To test this, we examine children's performance on a classic relational task – the relational match-to-sample task (RMTS). Prior work has shown that 4-year-olds can pass this task (Christie & Gentner, 2014). However, if nominal language induces an object focus, their performance should be disrupted by a noun-labeling pretask. In two experiments, 4-year-olds either labeled objects or actions in a naming pretask. Then they completed the RMTS task. Consistent with the noun-focus explanation, the object-naming group failed the RMTS task, whereas the action-naming group and a control group both succeeded. This suggests that nominal language can lead to an object focus, and that this could explain the temporary decline in children's relational processing.

Keywords: cognitive development, relational processing, learning, language

Introduction

Relational processing is the capstone of human reasoning. Some have proposed that humans' remarkable capacity for relational processing may be the key difference between humans and non-human primates (Gentner, 2003, 2010; Penn, Holyoak, & Povinelli, 2008). The focus of the current work is to examine how language influences changes in relational processing over development.

How well do children engage in relational processing? Gentner's (1988) relational shift account posits that early in learning, children (and other novices) tend to focus on objects and surface-level attributes, and that with learning, they become able to focus on relational similarity (Gentner & Rattermann, 1991; Honomichl & Chen, 2006; Paik & Mix, 2006; Rattermann & Gentner, 1998). For instance, Gentner (1988) found that 5-6 year olds interpreted the metaphor "A tire is a shoe" based on their shared attributes ("both are black"), whereas 9-10 year olds produced more relational interpretations ("you can go places on both").

Gentner and Rattermann (1991) provided further support for the relational shift using a different paradigm. In their relational matching task, the experimenter and child each had their own set of three distinct objects. The experimenter hid a sticker under an object in their set, and the child had to find a sticker in the corresponding location in their set (i.e., if the sticker was in the center of the experimenter's set, then the child would find sticker in the center object of their set). In one condition, the objects were cross-mapped—for example, the center object in the experimenter's set was identical to the rightmost object in the child's set. Under these conditions, it was difficult for 3-year-olds to resist matching by object identity. Instead of searching in the corresponding relational location (center), they searched under the identical object. In contrast, 5-year-olds were able to resist the object matches and make relational matches.

Christie and Gentner (2014) measured children's performance on a different relational matching task used widely with other species – the relational match-to-sample task (RMTS; Premack, 1983). In this task, subjects are given a standard (AA) and must select a match between two options (BB or CD). The BB pair should be matched to the AA pair on the basis of the common *same* relation between them. Christie and Gentner (2014) found that 4-year-olds were able to make relational matches at above chance levels in the RMTS task, but 3- and 2½-year-olds were not. Together, these studies reveal that younger children tend to focus on individual objects rather than on relations. (These studies compare younger versus older children because age generally correlates with greater knowledge and thus better relational processing. However, a relational shift occurs at different times for different tasks. For example, college students show a novice-expert shift from focusing on concrete domain matches to focusing on common causal systems [Rottman et al., 2012].)

What factors improve children's relational processing? Gentner and colleagues have proposed that the acquisition of relational knowledge, via a combination of structure-mapping processes and relational language, allows learners to shift their focus towards relations (Christie & Gentner, 2014; Gentner, 2010; Gentner & Christie, 2008; Rattermann & Gentner, 1998). Other accounts emphasize the maturation of domain-general cognitive abilities. For instance, some have argued that the development of working memory (e.g.,, Halford, Wilson, & Phillips, 1998) and inhibitory control (e.g., Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2010) allow children to consider multiple relations and resist object matches. Nonetheless, there is widespread agreement that children's relational processing generally improves over time. This consensus has been challenged by recent findings.

Developmental Decline in Relational Processing

While Christie and Gentner (2014) found that 2½- and 3-year-olds failed to make relational matches in the RMTS task, the ability to generalize *same-different* relations has been found to be present much earlier in development. Ferry, Hespos and Gentner (2015) habituated 7- and 9-month-old infants to either *same* or *different* pairs of objects. At test, they were shown novel *same* and *different* pairs. Infants looked longer at the novel relation (i.e., infants in the *same* condition looked longer at the *different* pair, and vice versa). Thus, infants appear capable of analogical generalization across multiple examples (see also Marcus, Vijayan, Bandi Rao & Vishton, 1999; Saffran, Pollak, Seibel, & Shkolnik, 2007).

Toddlers have also been found to generalize *same-different* relations. Walker and Gopnik (2014; Walker, Bridgers, & Gopnik, 2015) modified the classic blicket detector paradigm and showed 18- to 30-month-olds a machine that played music only when certain pairs of objects were placed on it. During training, four unique pairs of objects were placed on the machine (twice each). In the *same* condition, the machine played music only when the experimenter placed *same* pairs on the machine, and not when *different* pairs were used. The reverse pattern held for the *different* condition. After this training, the toddlers successfully chose between a *same* and a *different* pair to make the machine play music. Thus, very young humans can generalize these relations from multiple examples.

In themselves these results may not be totally surprising. Perhaps toddlers do well on the blicket-detector task because it is an especially engaging task or because it affords opportunities for comparison. But if so, then performance should continue to improve with age and experience. Surprisingly, Walker et al. (2015) found that older children performed worse than younger children. When older children (30 to 36 months) were tested in the same task, they could select the correct pair only in the *same* condition, and not the *different* condition. An older group (36 to 47 months), failed in both the *same* or *different* conditions. These results suggest that between the ages of 18 and 47 months, there is a decline in children's ability to notice these relations *in the same task*.

The overall picture that emerges is a puzzling one. Taken together, the evidence suggests that the developmental trajectory of relational processing is a U-shaped curve. At an early age, at least some relations are available to children (with the right scaffolding). Over time, relations become temporarily less available.

What could account for this decline in relational ability? As discussed above, failure to make relational matches can result from a focus on objects. Thus if children focused strongly on individual objects in the relational blicket detector task, this could account for their failure to notice that the objects form an abstract relation. But this line of reasoning would call for an explanation of *why* the object focus should become stronger during early childhood.

Language and Relational Processing

What could explain the apparent decline in relational ability roughly between 2.5 and 3 years of age? One proposal, advanced by Walker et al. (2015), is that children's early hypothesis space gives equal priority to objects and relations, but with learning, children form "a different overhypothesis namely, that individual kinds of objects, rather than relations between them, have causal powers" (p. 2560). As children learn that individual objects are likely to be causally potent, they prioritize object-focused hypotheses over relational hypotheses. Thus, the older group fails to attend to relations in the task.

Here we consider a different hypothesis—that patterns of language learning are to blame for the object focus. Specifically, we hypothesize that young children are focused on learning names for objects and that this focus on objects temporarily impedes their ability to notice relations.

We note that this proposal runs contrary to the abundant evidence that language aids relational processing (Gentner & Christie, 2008; Gentner & Rattermann, 1991; Loewenstein & Gentner, 2005; Pruden, Levine, & Huttenlocher, 2011; Pyers & Senghas, 2009). For example, learning spatial relational terms such as *left-right* (Hermer-Vazquez, Moffet, & Munkholm, 2001) and *on* (Casasola, 2005) have been shown to help children in relational and categorization tasks. Even in *same-different* tasks like the RMTS, 3-year-olds given training in the meanings of *same* and *different* succeed in making relational matches (Christie & Gentner, 2014). However, these studies all concern the benefits of *relational* language. Our hypothesis concerns a different aspect of language. We propose that the early acquisition of nouns can have a temporarily negative influence on relational reasoning.

The Nominal Explosion There is overwhelming evidence that nouns are among the first words that children learn, and they dominate over other word classes in early vocabulary (Bates et al., 1994; Bornstein et al., 2004; Gentner, 1982; Gentner & Boroditsky, 2001; Gleitman et al., 2005; MacNamara, 1972). Bates and colleagues' (1994) used the MacArthur-Bates Communicative Development Inventory (MCDI) with a large sample of parents and showed that children's early vocabulary was dominated by nouns. The proportion of nouns relative to other words was found to peak at a vocabulary size of 100 to 200 words (55.2% of these words were found to be nouns). Peak "nouniness" occurs roughly between 1 ½ to 2 ½ years of age. As children's vocabularies grow, the proportion of verbs and other predicates increases, while the proportion of nouns decreases. This pattern of early noun dominance appears across languages, as evidenced by MCDI data collected from thousands of children across 13 languages (Braginsky, Yurovsky, Marchman, & Frank, 2015). We propose that the dominance of nouns early in language learning can make objects particularly salient.

According to the natural partitions hypothesis (Gentner, 1982), early noun dominance results from the fact that many

nouns refer to concrete, perceptually individuated entities in the environment (unlike verbs and prepositions, whose meanings vary cross-linguistically). Nouns are thus a natural entry point for referential language. Children are eager to learn language, and acquiring names for objects is a highly successful strategy for young children. This strategy can lead young children to focus strongly on objects. Of course, in the long run, a repertoire of noun meanings will aid children in learning relational terms (Gentner 1982; Gleitman, 1990). But in the short run, it may result in an object focus that renders relations temporarily less available.

We tested this idea in a pilot study. We asked whether 2-year-olds' performance on the RMTS task was related to their language development, as assessed by the MCDI. We found that 2-year-olds who performed better than chance on the RMTS (5/8 trials or more) had fewer words checked off on the MCDI (M = 441.38, SD = 153.51) than children who performed at or below chance (M = 562.23, SD = 124.82, $t(28) = 2.21$, $p = .04$, $d = 0.86$). This provides some encouragement for the idea that knowing more words could in fact impede relational processing. Our goal for the current research was to manipulate children's noun focus and examine its influence on relational processing.

Experiment 1

In the current work, we test the hypothesis that nominal language induces an object focus and disrupts children's ability to make relational matches. We used Christie & Gentner's (2014) *same*-only RMTS task, and tested children at an age at which they already pass the RMTS task – 4 years. We developed a naming task in which children are asked to provide noun labels for common objects presented on cards. By having children activate many noun labels, we hypothesized it would lead to an object focus. The goal was to bring children back into a "younger" mindset – to an earlier age at which they were captivated by objects. If nominal language induces an object focus, then children in this condition should show impaired performance on a subsequent RMTS task. We compared this group of children to a separate group who also engaged in a naming task. However, we asked this second group to produce verb labels for images of scenes. Because the verb naming group was a control, we chose transitive action verbs that could be easily depicted. We did not predict that priming transitive verbs would improve children's performance on the RMTS task relative to the baseline group. Instead, we predicted that the noun labeling group would perform poorly because of the induced object focus.

Methods

Participants Fifty-nine 4-year-olds were recruited from the Evanston/Chicago area (29 females, M_{age} = 53 months). The racial and economic composition of the sample reflected those of the local population. Children received a small gift for their participation.

Materials & Procedure The RMTS task was presented on cards. There were 8 trials. Each trial was made up of a triad of three cards: the standard card, and two possible matches. Each card depicted two shapes – either a *same* pair or *different* pair. The standard always showed a *same* pair, and children had to select between a *same* pair and a *different* pair to match the standard (see Figure 1 for a sample triad). Within each triad, each pair was made up of unique shapes and colors. At the end of the 8 target trials, there were two catch trials meant to determine whether the child understood the task. These catch trials were literal similarity matches that did not require the child to judge relational similarity. Children who failed any of the catch trials were not included in the analysis (n = 8).

For each trial of the RMTS task, the experimenter placed the standard in front of the child and asked "Do you see this one?" Then the experimenter placed the two possible matches below the standard (as in Figure 1) and asked "Which one of these two goes with this one?" No feedback was given on the child's response. The next trial was presented until the task was complete.

Children were placed in one of three between-subjects conditions: *Object Naming*, *Action Naming*, or *Baseline*. Children in the Object Naming condition first completed the object naming pretask. A set of 39 cards was created depicting everyday objects (e.g., pencil, ball, turtle). The experimenter presented each card in the object naming set one at a time and asked the child "What is this called?" If the child did not know one of the words, they were told the label and were asked to repeat it. After the child named all the objects, they completed the RMTS task.

A similar procedure was used in the Action Naming condition. Children first completed the action naming pretask. A set of 35 cards was created depicting scenes involving actions that were familiar to children (e.g., kicking, throwing, jumping). For each card, children were asked "What is (s)he doing?" They generally provided verb phrases ("he's throwing it"), or single word utterances ("jumping"). If the child could not produce a label, it was provided for them and they were asked to repeat it. After naming the action cards, children completed the RMTS task.

In the Baseline condition, children simply completed the RMTS task. This condition was created in order to test children's relational ability with no prior manipulations. We expected to replicate the performance of 4-year-olds in Christie and Gentner (2014).

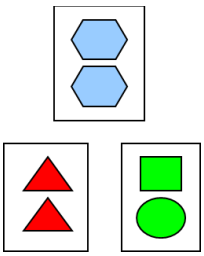


Figure 1. A sample triad from the RMTS task.

Results & Discussion

We measured the mean proportion of relational matches in the RMTS task. Two different analyses assessed performance. First, we used a one-way ANOVA to calculate differences between the conditions. We found a significant effect of condition, $F(2,48) = 4.43$, $p = .02$, $\eta^2 = 0.15$. Bonferroni post-hoc tests revealed that the Baseline group ($M = 0.76$, $SD = 0.22$) made significantly more relational matches than the Object Naming group ($M = 0.51$, $SD = 0.28$), $p = .03$. The Action Naming group ($M = 0.73$, $SD = 0.30$) was no different from the Baseline group, $p = 1.00$. The difference between the Action Naming and Object Naming groups approached significance, $p = .06$.

In the second analysis, we compared the means of each group to chance (50%). We found that the Baseline group selected relational matches significantly more than chance, $t(16) = 4.76$, $p < .001$, $d = 1.18$, replicating the findings from Christie and Gentner (2014). The Action Naming group also performed above chance, $t(16) = 3.16$, $p = .006$, $d = 0.77$. However, the Object Naming group performed no differently from chance, $t(16) = 0.11$, $p = .91$, $d = 0.04$.

As we predicted, the children who labeled objects prior to the RMTS task performed worse than those who did not (the Baseline condition). Further, those who labeled actions did not display impaired performance on the RMTS task, suggesting that the impairment was specific to noun labels and not due to fatigue. This pattern is consistent with the idea that nominal language can impair relational matching.

Experiment 2

Experiment 1 provided evidence that noun-naming can (at least temporarily) impair relational reasoning. However, before drawing conclusions we wanted to generalize the findings and address some possible concerns. In Experiment 1, we used a mix of hand-drawn and printed images, and different numbers of cards between conditions. For Experiment 2, we developed new sets of cards, using words that weren't used in Experiment 1, and we used a more uniform style across the images. The number of cards shown in each condition was also equalized (a total of 40 cards each). We also changed the instructions given in the RMTS task. In Experiment 1, the target question was "Which one of these two goes with this one?" The meaning of "goes with" may not be immediately clear to children, so in order to make sure this did not influence the results, we asked children "Which one of these two is more like this one?" (following Christie & Gentner, 2014).

Methods

Participants Twenty-nine 4-year-olds were recruited from the Evanston/Chicago area (14 females, $M_{age} = 53$ months). The demographics were the same as in Experiment 1.

Materials and Procedure Two new sets of cards were developed. Forty new objects and forty new actions were depicted on cards for the Object Naming and Action Naming conditions. The RMTS task was identical to Experiment 1, with the exception that an additional catch trial was added for a total of three literal similarity catch trials. Children that failed any of these catch trials were excluded from analysis (n = 4).

The overall procedure was the same as in Experiment 1. Children were placed in either the *Object Naming* or *Action Naming* condition. In the Object Naming condition, children were shown a card with an image of an object and were asked, "What is this called?" In the Action Naming condition, children were shown a card with an image of a scene and were asked: "What is (s)he doing?"

Following the naming pretask, children completed the same 8 trials of the RMTS task as in Experiment 1. However, the instructions were slightly different. After the two possible matches were placed underneath the standard, the target question was "Which one of these two is more like this one?" No feedback was provided.

Results & Discussion

As in Experiment 1, we measured the proportion of relational matches in the RMTS task. Although the Object Naming group ($M = 0.55$, $SD = 0.32$) appeared to perform worse than the Action Naming group ($M = 0.69$, $SD = 0.21$), the difference did not reach significance, $t(23) = -1.23$, $p = .23$, $d = 0.54$. However, whereas the Action Naming group performed above-chance on the RMTS task, $t(8) = 2.80$, $p = .02$, $d = 0.90$, the Object Naming group did not, $t(15) = 0.58$, $p = .57$, $d = 0.16$.

These results largely replicated Experiment 1. We found that with a new set of stimuli, providing noun labels for objects hindered children's ability to make relational matches in the RMTS task. This effect was not seen when children were asked to provide verb labels for actions.

General Discussion

Research on relational processing in the first few years of life has revealed a puzzling pattern. Infants and young toddlers can generalize *same-different* relations when presented with multiple examples (Ferry et al., 2015; Walker et al., 2015). In contrast, slightly older toddlers, from around 2.5 years of age, show a steadily decreasing ability to notice these relations. Here we tested the possibility that this decline results from an object focus that is a side effect of children's early focus on acquiring nouns. We hypothesize that children's interest in noun-learning can make objects more salient, and interfere with children's ability to judge relational similarity across instances.

We tested this possibility by simulating these conditions in preschoolers. We first verified that four-year-olds could easily make relational matches in the RMTS task in the absence of any manipulation via the Baseline condition. We found that this was the case, replicating Christie and Gentner (2014). We asked children in the Object Naming condition to generate noun labels for familiar objects. We predicted that this experience would lead them to focus on objects and thus perform poorly on a subsequent RMTS

task. Across two experiments, we showed that the use of noun labels had a negative influence on children's relational matching. In both studies, children in the Object Naming condition went on to perform at chance in the RMTS task—far lower than the Baseline group. In contrast, children asked to generate verbs in the Action Naming condition saw no such impairments. This finding supports our hypothesis that the relational decline is related to children's early tendency to focus on objects in the service of noun learning.

An interesting aspect of our findings is that we successfully induced a general focus on objects, rather than a focus on specific objects shown in the pretask. To illustrate this difference, let us consider one of the studies conducted by Ferry et al. (2015). Infants were exposed to specific objects in the waiting room before the experiment began. When these same objects appeared within a relational pair in one of the test trials, this impaired infants' ability to notice that the object was part of an abstract relation. They focused on the individual object they had seen earlier, and this impaired their ability to notice relational similarity. However, the current experiments differ; the objects that children labeled in the pretask never appeared in the subsequent RMTS task. Thus, activating multiple noun labels led to a general focus on objects that influenced how the children processed unrelated objects in the RMTS task. This supports our claim that language can lead to a generalized object focus.

The current work shows that focusing on nouns can direct attention to objects at the expense of relations. These findings dovetail with other results showing that noun labels lead both infants and children to focus on objects (Balaban & Waxman, 1997; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Xu & Carey, 1996). For instance, Xu and Carey (1996) found that 10-month-olds who knew the names of particular objects could keep track of the objects' identities in an object individuation task better than the infants who did not know the words.

Comparative work also provides some support for the notion that language is related to the object focus. Christie, Gentner, Call, and Haun (2016) compared the performance of nonhuman apes and 3-year-old children on the same relational mapping task. Children displayed greater relational matching ability than chimpanzees and bonobos. But when given a cross-mapping task, children were far more distracted by object matches than were the apes. Thus it does not appear that children's early noun focus results from a general phylogenetic bias. As none of the apes in this study had been taught symbol systems, language learning is at least a plausible candidate for the greater object focus seen in children.

We note that there is a slight difference between the generally assumed age of the nominal explosion and the age at which Walker et al. (2015) found a relational decline. The nominal explosion occurs between 18 and 30 months, whereas the decline in relational reasoning occurs slightly later, starting around 36 months. However, effects of language often take time to become entrenched. Noun-dominance in language learning may appear early, and affect relational attention at a somewhat later time.

The current work has focused on a temporary decline. But in the larger picture, children improve dramatically in their relational ability over the preschool and early school years (Gentner & Rattermann, 1991; Paik & Mix, 2006; Rattermann & Gentner, 1998). Although general experience (Rattermann & Gentner, 1998) and increases in executive ability (Richland et al., 2006; Thibaut et al., 2010) and processing capacity (Halford et al., 1998) may all contribute to this development, we suggest that a major factor is relational language learning (Gentner, 2010; Gentner & Rattermann, 1991). Children are acquiring relational terms—including spatial terms and mathematical terms—that support relational representation and reasoning (Casasola, 2005; Christie & Gentner, 2014; Gentner & Christie, 2008; Hermer-Vasquez et al., 2001; Loewenstein & Gentner, 2005; Pruden et al., 2011; Pyers & Senghas, 2009). Over learning, language becomes a fine-tuned instrument that can support selective attention to either objects or relations.

Our goal here was to induce an object focus via our object naming task as a microcosm of what occurs on a macro scale with the noun explosion. In ongoing studies, we aim to trace these effects at younger ages.

Acknowledgements

We thank Emma Bulzoni, Katie Cha, Tiana Hickey, and Victoria Wee for their help with data collection. This work was supported by NSF SLC Grant SBE-0541957 awarded to the Spatial Intelligence and Learning Center (SILC), and by ONR Grant N00014-92-J-1098 to Dedre Gentner.

References

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*(1), 3–26.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., … Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development, 75*(4), 1115-1139.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21,* 85 – 123.

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. C. (2015). Developmental trajectories of vocabulary composition across languages. Poster presented at the 40th Annual Boston University Conference on Language Development.

Casasola, M. (2005). Can language do the driving? The effect of linguistic input on infants' categorization of support spatial relations. *Developmental Psychology, 41*, 183-192.

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*, 383-397.

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to relational similarity and object similarity in apes and children. *Current Biology, 26*(4), 531-535.

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development, 86*, 1386-1405.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.) *Language development, volume 2: Language, thought, and culture* (pp. 301-334). Hillsdale, NJ: Lawrence Erlbaum.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*, 47-59.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195-235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*(5), 752-775.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 215-256). Cambridge, UK: Cambridge University Press.

Gentner, D., & Christie, S. (2008). Relational language supports relational cognition in humans and apes. *Behavioral and Brain Sciences, 31*, 136-137.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225-277). London: Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3-55.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, 1(1), 23–64.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral & Brain Sciences, 21*, 803-864.

Hermer-Vazquez, L., Moffet, A., & Munkholm, P. (2001). Language, space, and the development of cognitive flexibility in humans: The case of two spatial memory tasks. *Cognition, 79*(3), 263 – 299.

Honomichl, R. D., & Chen, Z. (2006). Learning to align relations: The effects of feedback and self-explanation. *Journal of Cognition and Development, 7*(4), 527–550.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology, 50*, 315-353.

MacNamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79*(1), 1 – 13.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P.M. (1999). Rule-learning in seven-month-old infants. *Science, 283*, 77-80.

Paik, J. H., & Mix, K. S. (2006). Preschoolers' use of surface similarity in object comparisons: Taking context into account. *Journal of Experimental Child Psychology, 95*(3), 194-214.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*(02), 109–130.

Premack, D. (1983). The codes of man and beasts. *Behavioral & Brain Sciences, 6*, 125–167.

Pruden, S. M., Levine, S., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science, 14*, 1417-1430.

Pyers, J. E., & Senghas, A. (2009). Language promotes false-belief understanding: Evidence from learners of a new sign language. *Psychological Science, 20*, 805–812.

Rattermann, M. J., & Gentner, D. (1998b). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development, 13*, 453-478.

Richland, L.E., Morrison, R.G., & Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*, 249–273.

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science, 36*, 919 – 932.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007) Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition, 105*(3), 669-680.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*(1), 13 – 19.

Thibaut, J.-P., French, R. M., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology, 106*(1), 1-19.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science, 25*, 161–169.

Walker, C., Bridgers, S. B., & Gopnik, A. (2015). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring "same" and "different." In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Xu, F., & Carey, S. (1996). Infants' metaphysics: The case of numerical identity. *Cognitive Psychology, 30*, 111-153.

Linguistic Priming and Learning Adjacent and Non-Adjacent Dependencies in Serial Reaction Time Tasks

Felix Hao Wang (wang970@usc.edu)[a], Elsi Kaiser (emkaiser@usc.edu)[b]

[a]Department of Psychology, University of Southern California, 3620 McClintock Ave, Los Angeles, CA, 90089
[a]Department of Linguistics, University of Southern California, Los Angeles, CA, 90089

Abstract

Although syntactic priming is well studied and commonly assumed to involve implicit learning, the mechanisms behind this phenomenon are still under debate. We tested whether implicit learning of adjacent and non-adjacent sequences occurs in a non-linguistic, finger sequence task (Serial Reaction Time task), and if so, whether these implicitly-learned dependencies can cause syntactic priming in the linguistic domain. We followed the logic that exposure to statistical patterns in the SRT task may influence language users' relative clause (RC) attachment biases, and trained participants on SRT sequences with adjacent or non-adjacent dependencies. Participants then wrote completions to relative clause fragments in a situation where they could opt for adjacent or non-adjacent linguistic structures. Participants successfully learned the adjacent and non-adjacent dependency implicitly during the SRT task, but, strikingly, their RC continuations did not exhibit priming effects. Implications for theories of syntactic priming and its relations to implicit learning are discussed.

Keywords: implicit learning; syntactic priming; relative clause attachment bias; non-adjacent dependencies

Introduction

Although the phenomenon of syntactic priming has been well studied in the literature (e.g., Bock, 1986; Pickering & Branigan 1998), the exact processes behind priming are still unclear. A burgeoning literature on syntactic priming has taken implicit learning as the one of the mechanisms for syntactic priming (e.g., Bock & Griffin, 2000; Chang, Dell & Bock 2006). In this paper, we build on insights from research on implicit learning research to look at adult sentence processing regarding the representation of abstract dependencies in language and other cognitive domains (Scheepers & Sturt, 2014; Menon & Kaiser, 2014; Van de Cavey & Hartsuiker 2016). We explore whether an abstract relation represented through statistical regularities from a finger sequence task (Serial Reaction Time Task, or SRT) can prime the attachment biases of relative clauses. Specifically, do adjacent and non-adjacent prediction relations derived from statistics prime the low versus high attachment preference during the production of English relative clauses?

The notions of adjacency and non-adjacency are relevant to the domain of syntax with regard to the representation of relative clauses (RCs). In English, sentences with the structure "NP1 of NP2 who" (e.g., *Jessica visited the doctor of the supermodels who…*) are structurally ambiguous as to which NP the relative clause attaches to. A relative clause can potentially attach to either one of the NPs. When the intended meaning is "the doctor lived in Los Angeles" (e.g. *Jessica visited the doctor of the supermodels who lives in Los Angeles)*, the relative clause attaches to the first NP (NP1), a structure called high attachment. When the relative clause attaches back to the lower NP (NP2) (e.g. *Jessica visited the doctor of the supermodels who live in Los Angeles*), this is a low attachment structure. Thus, relative clauses may modify the adjacent (local) noun (low attachment) or the non-adjacent noun (high attachment). Although such constructions may be ambiguous as to the intended attachment site, it is often the case that linguistic cues (e.g. *lives* vs. *live*) signal whether a speaker intended a low or a high attachment.

According to Scheepers 2003, the syntactic distinction between high and low in relative clause attachment bias is only a matter of syntactic sequencing: The syntactic rules used to generate high and low attachment relative clauses are the same. The only distinction between the two is that, in low attachment, the relative clause modifies the noun immediately preceding it, and in high attachment, it modifies the noun that non-adjacently precedes it. In our opinion, this provides a good opportunity for priming to be probed on an abstract level. Our previous work investigated the relationship between artificial language learning and structural priming, and suggested that relative clause attachments can be primed from statistical learning of artificial languages (Wang, Menon & Kaiser, under review). We found that participants who learned an artificial language with non-adjacent dependencies are more likely to complete relative clauses with high attachment, compared to a baseline group as well as compared to participants who were exposed to an artificial language with adjacent dependencies. Van de Cavey & Hartsuiker (2016), among others, have also provided evidence that an array of different experimental tasks (e.g., sentence comprehension, music listening, and math solving) can also prime relative sentence completion.

However, the question regarding the underlying mechanism for syntactic priming is not fully answered by existing work. Given the nature of relative clause attachment, it is not clear whether existing proposals regarding the mechanisms of syntactic priming can explain results from recent syntactic priming work, especially data from studies that successfully used non-linguistic stimuli to

induce syntactic priming (e.g. musical sequences or math equations). These kinds of data suggest that learning and processing sequences of numbers, musical notes and other information involves *domain-general implicit learning*, and that this kind of domain-general implicit learning is involved in syntactic priming. This predicts that other kinds of implicit learning should also trigger syntactic priming. However, are there limits to what can trigger syntactic priming? If other kinds of domain-general implicit learning do *not* result in syntactic priming, we may need to reexamine current models of syntactic priming.

To investigate these issues, we used a Serial Reaction Time Task (SRT task, Howard & Howard, 1997) to induce implicit learning. Our goal is to induce domain-general implicit learning of sequences. In an SRT task, participants generally find it very difficult to articulate *explicit* rules for the probabilistic finger sequences they are implicitly learning, whereas other tasks (such as math tasks and music tasks) involve a mixture of explicit and implicit learning that is difficult to untangle. This makes SRT tasks one of the best tasks for our goal of investigating *implicit learning*, given the long history of using SRT to explore implicit learning.

Our study tests whether structural representations arising from distributional information – in this case, finger sequences – can prime relative clause completions. If relative clause attachment biases come from representations that are completely separate from domain-general sequence processing representations, representations constructed based on finger sequences will not result in any changes in the completion of relative clauses. On the other hand, if relative clause attachment biases come from representations that are shared with domain-general sequential representations, the relative clause bias is predicted to change as a result of learning sequential statistics and sequence processing in general.

Experiment

In this experiment, we test whether implicit learning of finger sequences that involve adjacent or non-adjacent dependencies influences whether participants produce completions for incomplete relative clauses fragments that involve adjacent or non-adjacent syntactic dependencies.

Methods

Participants. Seventy-two undergraduates participated.

Stimuli. There are two tasks in the experiment: an SRT task and a sentence completion task. In the SRT task, we used a three circles and a dog image to indicate a position (Figure 1). The dog image changed location depending on which position was being represented. Figure 1 is an example of position 2 (since the dog is in the second position from the left). There are four possible positions. Participants' task was to press a key indicating what the position of the dog is. Participants saw multiple such screens in succession, with varying locations of the dog image (e.g. in position 2 on one screen, in position 1 on the next screen, and so on).

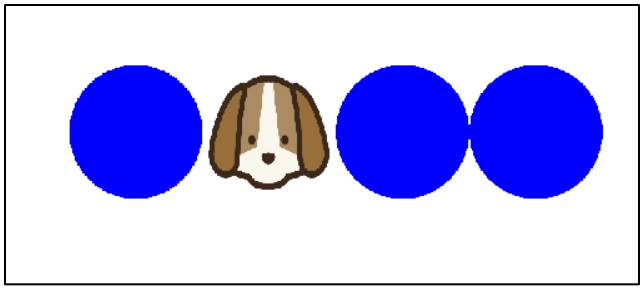


Figure 1. Interface during the SRT task. This example image depicts position 2 (since the dog is in the second position from the left).

In the sentence completion task, we created sentence fragments for participants to complete in the testing phase. There are two kinds of sentence fragments: targets and fillers. There were 30 target sentence fragments. All target sentence fragments have the structure shown in Example (1): In order to complete the fragments, participants write a relative clause modifying a complex noun phrase made of two noun phrases. As shown in Example 1, NP1 (the doctor) and NP2 (the supermodels) are connected by the preposition 'of' and are followed by the relative pronoun 'who', providing the possibility that the continuation can modify either the NP1 or NP2.

(1) John meets [the doctor]$_{NP1}$ of [the supermodels]$_{NP2}$ [who …]$_{RC}$.

In target items, the subject of sentence was always a proper name (e.g. *John*, equal numbers of male and female names), and the two NPs that make up the object (e.g. *the doctor of the supermodels)* were definite, animate nouns preceded by the definite article. NPs were controlled for number: all sentence fragments had NP1 as singular and NP2 as plural (e.g. *the doctor of the supermodels*)[1]. The differences in number makes coding of attachment easier because number marking on the verb can potentially disambiguate the attachment (e.g. *…was happy* vs. *…were happy*). For the same reason, the verbs in the main clause were in the present tense (in 3rd singular form). All verbs in the target fragments (e.g. *counted*) were non-implicit causality verbs, chosen in order to avoid verb semantics bias. Fillers were non-ambiguous English sentence fragments of similar length.

Design. The experiment consisted of sixteen blocks. Blocks 1-10, 12 and 14 consisted of Serial Reaction Time Tasks. Block 11, 13 and 15 were sentence completion tasks. Block 16 was an explicit assessment of the knowledge regarding the patterns presented in SRT task.

[1] We used singular-plural NPs only because our earlier work on this topic indicates that variance from priming came mostly with the singular/plural items (Wang, Menon & Kaiser, under review).

Serial Reaction Time Task Design. In the Serial Reaction Time task, we used a between-subjects task where half of the participants (n=36) were trained on adjacent dependencies and the other half (n=36) were trained non non-adjacent dependencies. A total of 2304 trials (i.e., screens like Figure 2, but with the dog image in varying locations) were created for both adjacent and non-adjacent dependency conditions, with 192 trials in each of the 12 blocks (described below). Between each block, participants were given information about whether the next block is an SRT block or a sentence completion block, as well as a chance to take a self-paced break.

On each trial within an SRT block, participants responded to positional information on the screen (represented by the position of the dog image), and pressed one of four keys on the keyboard (z, x, n or m) that corresponds with the position of the dog on that particular trial. After responding to the position correctly, there is a 120ms blank screen, followed by the next trial with the dog image at the position for the next trial. If the response is incorrect (if the participant pressed any key on the keyboard other than the intended key), the dog display (e.g. Figure 1) remains on the screen until the correct key is pressed. The sequence of the positions is determined by the condition that a participant was in (adjacent vs. non-adjacent dependency group).

In the *adjacent dependency condition*, 4 triplets of sequences were concatenated: R41, R13, R21, R34, where the letter R represents a random position in a display, and 1, 2, 3, and 4 represent the four positions from left to right. Note that there could be repetitions of the same position, and participants would have pressed the same key twice in that instance. The sequence was constructed such that each of the 4 triplets occurred an equal number of times, and within all occurrences of each triplet, all 4 positions occurred equal number of times. This sequence generation procedure ensures that the only dependencies are the intended *adjacent dependencies* (the dog head in position 4 on one display predicts that the dog head will be in position 1 on the next display, 1 predicts 3 on the next display, 2 predicts 1 and 3 predicts 4). For example, a dog head in position 1 can be preceded by a dog head in any of the four positions, but the probability that it was in position 4 is 3 times as the probability that it was in position 1, 2 or 3, which makes the transitional probability of 1 given 4 50% while the others are at 16.67%. Adjacency matrices were calculated (predictions of the current item given the immediate previous item, the item previous to that, and so forth). Up to the 5th order adjacency matrices were examined to make sure that no higher order prediction is possible from this sequence. An example sequence can be viewed in Example (2). In essence, participants in this group were exposed to *adjacent dependencies* across trials, since position 4 on one display predicts position 1 on the next display, 1 predicts 3 on the next display, and so on.

In the *non-adjacent dependency condition*, 4 triplets of sequences were concatenated: 4R1, 1R3, 2R1, 3R4, where the letter R represents a random position. Similar measures were taken as in the adjacent dependency condition to ensure that the only predictive relationship exists in the second order. Thus, participants in this group were exposed to non-adjacent dependencies across trials (e.g. a dog head in position 4 predicts a dog head in position 1 not on the immediately following display but on the display after that; 4R1).

The trials in the SRT were determined by a predetermined sequence, each containing relevant statistics. Two sequences, one each for the two conditions were generated (adjacent and non-adjacent). All subjects were tested on the same sequence by the condition they were in. These two sequences only differed in terms of whether the R position is in the first position of the triplet or the second position of the triplet, demonstrated in example (2) and (3):

(2) [Adjacent] 3 4 2 1 1 3 4 4 2 4 4 2 2 2 1 …
(3) [Non-adjacent] 4 3 2 1 1 3 4 4 2 4 4 2 2 2 1 …

Sentence Completion Task Design. Each participant completed 30 target sentence fragments and 48 filler sentence fragments split into three blocks (10 target and 16 filler in blocks 11, 13 & 15), in a randomized order. The targets and fillers were held constant across participants. The randomized order made sure that no two consecutive sentence fragments are target sentence fragments.

Procedure. Participants were informed that they will be doing two kinds of tasks. Then the study proceeded first with the instructions for the SRT. The instruction emphasized to use the index and middle fingers to press the four keys on the keyboard, and a demonstration phase was included to make sure participants understand the process.

After the demonstration of the SRT task, participants saw the instructions for the sentence completion task. The participants' task was to write a completion for sentence fragments (e.g. *John meets the doctor of the supermodels who*). They could write whatever first came to mind. We analyzed the target completions for whether the relative clause modifies NP1 (e.g. *who supports vaccination*) or NP2 (e.g. *who like to travel*). RCs that modify NP1 are high attachments and RCs that modify NP2 are low attachments.

In the 16th block, participants were tested on their explicit knowledge regarding the positions of the dog image. Along with the interface in Figure 2, participants were told: *"Earlier on, you saw one dog per screen. Now, you will see three rows of circles. You should imagine that these are three different screens. The question is, where will the dog on the last screen, i.e., in the third row, appear?"*

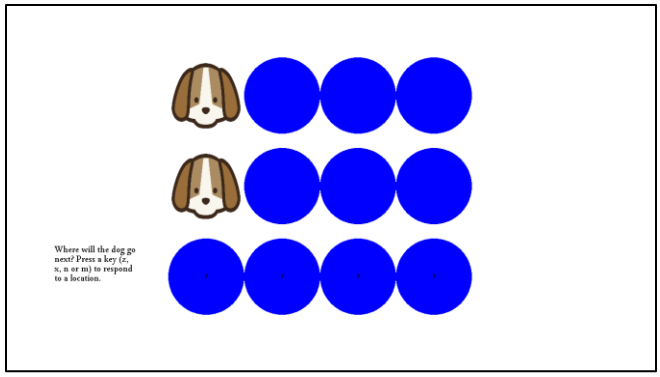


Figure 2. Interface for explicit knowledge test.

In other words, participants were asked to make judgments on where the dog head will appear given two preceding positions. This block tests whether participants have *explicit* knowledge regarding the location patterns. In this block, a total of 48 trials were given to each participant, with 3 repetitions of 16 different questions (full combinations of positions in the first and second row).

Results

We obtained three kinds of data: (i) data from the SRT task, including blocks of 1-10, 12 and 14; (ii) data from the sentence completion task, including blocks 11, 13 and 15; and (iii) the explicit knowledge task for block 16.

Serial Reaction Time Task Results For each trial in the SRT section, there is a reaction time associated with that trial. Each trial can be classified as a predictable trial or an unpredictable trial. Trials are predictable when the previous trial is predictive of the current trial. For example, in the adjacent condition, all trials with the dog head position 1 are predictable if the trial immediate before it had the dog head in position 4. As another example, in the non-adjacent dependency condition, all trials with the dog head in position 1 are predictable if two trials before it the position of the dog head was 4.

For the *adjacent dependency condition*, predictable trials were responded to faster than unpredictable trials. The mean RT for predictable trials is 415ms and the mean RT for unpredictable trials is 443ms. We used a mixed-effect linear regression to examine this difference statistically. Using subjects as a random effect and predictability of trials and block as fixed effects, we found that RT is faster for predictable trials (β= -27.86, z=-16.09, p<0.001). Block also significantly predicts RTs. For ease of modelling, we coded block as a continuous variable with block number 12 and 14 replaced as 11 & 12. There is a significant effect of block (β=-5.459, z=-21.75, p<0.001). The more the participants practice/the more blocks they go through, the fast the RTs become. In sum, participants showed learning of the adjacent statistics in general (Figure 3).

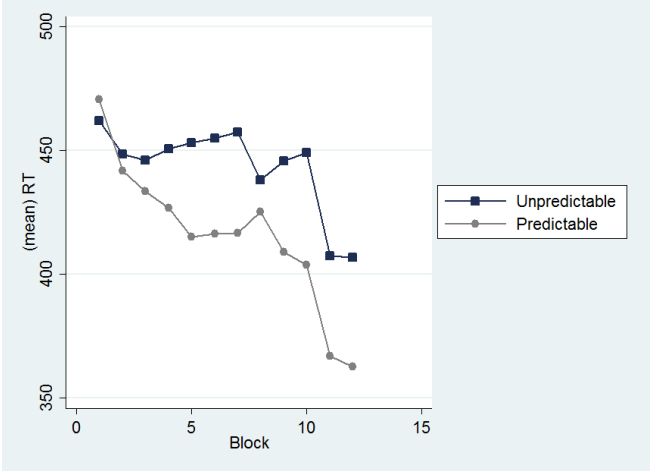


Figure 3. Mean of RT data from adjacent dependency condition. Blocks 11 and 12 in the figure represent data from experiment block 12 and 14, respectively.

For the *non-adjacent dependency condition*, predictable trials were also responded to faster than unpredictable trials. The mean RT for predictable trials is 413ms and the mean RT for unpredictable trials is 419ms. We used a mixed-effect linear regression to examine this difference statistically. Using subjects as a random effect and predictability of trials and block as fixed effects, we found that RT is faster for predictable trials (β= -5.13, z= -3.70, p<0.001). Block also significantly predicts RTs. For ease of modelling, we coded block as a continuous variable with block number 12 and 14 replaced as 11 & 12. The effect of block is significant (β=-5.61, z= -27.96, p<0.001). Similar to what we found for the adjacent group, participants become faster with practice. In sum, participants showed learning of the non-adjacent statistics in general as well, though the effect size may not be as large (Figure 4).

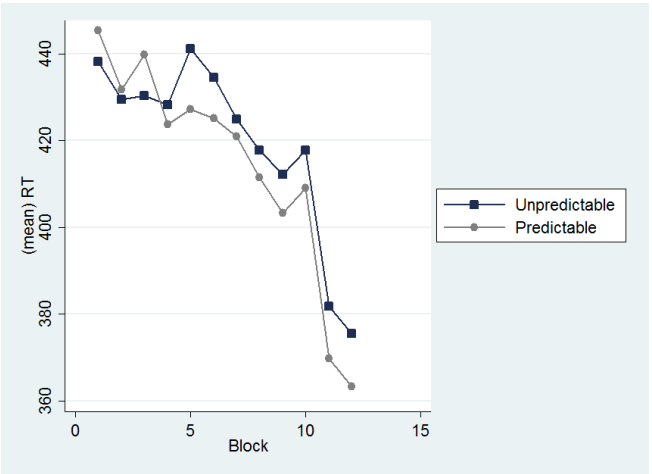


Figure 4. Mean of RT data from non-adjacent dependency condition. Blocks 11 and 12 in the figure represent data from experiment block 12 and 14, respectively.

Turning to the *sentence completion data*, we coded all the target sentence completion trials to see how participants completed the relative clause fragments. The coding of the sentences resulted in three types: high attachment (HA), low attachment (LA), and ambiguous. Coding was done with mostly syntactic considerations, given that the two NPs in our sentences differ in number. The main clue for coding comes from the verb in the relative clause that modifies the "who". More specifically, the verb in the relative clause in most sentences showed overt morphological agreement with the relevant NP. If verb number did not disambiguate (e.g. *went, asked*), semantic cues were used (e.g. Emily worked with the mother of the children *who just got tenure* => high attachment, vs. Chris counted the fans of the singer *who just finished the encore* => low attachment). If both verb marking and semantic cues were unclear, the sentence was coded as ambiguous. Alternatively, if the sentence completion was done without making the "who" the subject of the relative clause, or the sentence failed to make sense for coding purposes, it was coded as ambiguous as well.

To prepare for the mixed-effect logistic regression, high attachment sentences were coded as 1 and low attachment sentences as 0, and ambiguous was treated as missing. In the logistic regression, condition (adjacent/non-adjacent) was entered as the fixed effect, and subjects were entered as the random effect. We found no evidence of priming (β=0.031, z=0.38, p=0.704, ns). A graph of the proportions of high and low attachment completions as function of SRT group (adjacent vs. non-adjacent dependencies in SRT task) is in Figure 5. We find no signs of non-adjacent dependencies (NAD) in the SRT task priming (boosting the rate of) high attachments, compared to adjacent dependencies (ADJ).

Lastly, we examined the responses to the *explicit judgment questions* in block 16. We found that participants were 22.3% correct in their answers in the adjacent dependency condition, and 19.3% correct in the non-adjacent dependency condition. Both of these levels are statistically below chance (25% in a 4-alternative forced choice), suggesting that participants have no *explicit* knowledge of predictive relationships in the SRT task. Thus, participants have not formed explicit knowledge regarding how previous trials in the sequences are predictive of upcoming element in both the adjacent and non-adjacent condition – although their RT patterns provide evidence of *implicit* learning.

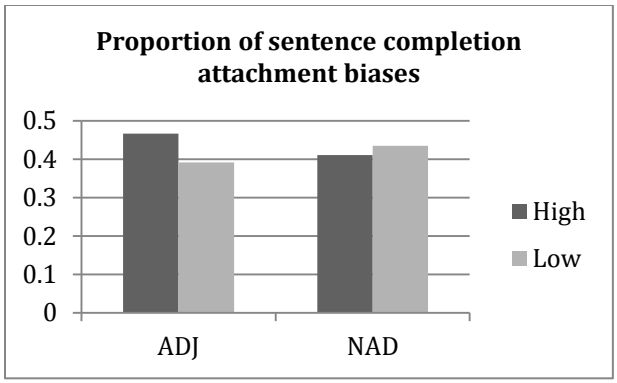


Figure 3. Proportions of high and low attachment completions as function of SRT group (adjacent vs. non-adjacent dependencies)

Discussion

We found evidence for implicit learning but no evidence of explicit learning in the SRT task. We also found no evidence of syntactic priming even when implicit learning of adjacent/non-adjacent statistics was successfully induced.

It is worth taking a closer look at the SRT task used in this experiment. The sequences in this experiment contained only one predictive relationship for each trial. That is, the only predictive relation in the adjacent condition is the current trial given the immediately previous one, and the current trial given the trial before the last in the non-adjacent dependency condition. The most common SRT task that induces representation of non-adjacency is the ASRT task (Alternating Serial Reaction Time task, Howard & Howard, 1997). In the ASRT task, participants are exposed to patterns such as 1R2R3R4R. Predictive relations exist between elements of distance 2, 4, 6, and so on. Crucially, given the configuration of ASRT sequences, it is not clear which previous trial was used for prediction for predictable trials. All we know is that non-adjacent dependencies from previous trials showed facilitated processing. This feature makes ASRT sequences unsuitable for the current relative clause priming work, because the contrast between adjacent to non-adjacent dependencies needs to be controlled. The difference between an adjacent dependency sequence and ASRT is not only between the distance between the current item and the previous item predicting it, but also the number of items that can predict the current trial. Our design avoided this difference by controlling the number of non-adjacent dependencies in the non-adjacent dependency condition.

The purpose of our SRT task was to induce implicit learning in a domain-general sense, which was successful. Indeed, we chose to use SRT to engage implicit learning and to minimize learning that exhibits explicit rule-seeking behavior. In other words, using SRT allows us to target domain general implicit learning specifically. In light of this, and given existing views regarding implicit learning in syntactic priming, there are a number of implications for syntactic priming that can be drawn from our study.

To begin with, an important question in this field is how implicit learning may induce syntactic priming. Previous work (Chang, Dell & Bock, 2006) used a connectionist model to specify how the process of implicit learning happens. In that model, it was assumed that reading sentences of a particular structure changed the weights over that structure such that the bias for the structure increased. This was in turn used to explain, in production, why syntactic priming occurs. This model provided a computational account of how implicit learning influenced syntactic behaviors of humans, but it may require further specification. For example, does the implicit learning process need to happen in the language domain in order for syntactic priming to happen? One possibility is that, in order for syntactic priming to occur, the implicit learning that takes place must be in the linguistic domain. Indeed, existing studies with linguistic materials that engaged implicit learning also successfully induced syntactic priming (Fine & Jaeger 2013; Hartsuiker, Bernolet, Schoonbaert, Speybroeck & Vanderelst, 2008). We recently provided evidence that learning an artificial language with different dependency structures successfully induced syntactic priming (Wang, Menon & Kaiser, under review). Whether the domain of implicit learning matters for syntactic priming needs to be examined more closely.

This also bring up the question of whether implicit learning in a domain-general sense (that is, other than the domain of language) can induce syntactic priming. We found implicit learning of adjacent and non-adjacent SRT finger sequences but no syntactic priming of adjacent and non-adjacent RC attachments. How do these results fit with existing data suggesting that non-linguistic representations in math and music can cause syntactic priming?

Existing evidence with non-linguistic priming (Van de Cavey & Hartsuiker, 2016; Scheepers et al., 2011; Scheepers & Sturt, 2014) may have involved both explicit rule representations in the non-linguistic domains. A subset of the experimental tasks used in these studies (arithmetic operations processing, for example), explicitly involves the use of rule-like positional operations. Thus, it may be that these findings cannot be explained by implicit learning and instead involve some other domain-general mechanisms. For example, if some kind of explicit learning drives syntactic priming, future work needs to be done to examine how this takes place. One possibility is that priming happens on a metacognitive level: Is syntactic priming in these studies the unintended result of participants trying to figure out the aims/goals of experiments, such that priming is present only when participants feel like the experiment is intending them to produce sentences with certain kinds of dependencies? Or is syntactic priming in these explicit manipulations a matter of unconscious, linguistic phenomenon, which, arguably, priming is supposed to be?

Although we do not offer conclusive answers to these questions here, our findings highlight the importance of investigating the effects implicit and explicit learning on syntactic priming (or lack thereof). We found that a finger sequence task with SRT which successfully induced implicit learning was unable to prime relative clause attachment biases and also failed to result in explicit learning of SRT sequences. This work provides preliminary evidence that implicit knowledge from a domain-general sequencing task alone does not induce syntactic priming. Future work is needed to examine whether this observation applies to other non-linguistic domains in terms of syntactic priming.

References

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, 18(2), 355-387.

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General, 129*, 177-192.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, 113, 234–272.

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578-591.

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2), 214-238.

Howard Jr, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and aging*, 12(4), 634.

Menon, M. & Kaiser, E. (2014). *Priming from Music to Language: Effects of Number Marking on the Priming of Relative Clause Attachment in Spanish*. Poster, AMLaP 2014 at University of Edinburgh, Scotland.

Rohde, H., Levy, R., & Kehler, A. (2011). Anticipating explanations in relative clause processing. *Cognition*, 118(3), 339-358.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4), 633-651.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, 89(2), 179-205.

Scheepers, C, & Sturt, P. (2014). Bidirectional syntactic priming across cognitive domains: From arithmetic to language and back. *Quarterly Journal of Experimental Psychology*. 67(8):1643-1654.

Van de Cavey, J., & Hartsuiker, R. J. (2016). Is there a domain-general cognitive structuring system? Evidence from structural priming across music, math, action descriptions, and language. *Cognition*, 146, 172-184.

Wang, F. H., Menon, M. & Kaiser, E. (under review). Statistical Structures in Artificial Language Primes Relative Clause Attachment.

Wang, F. H. & Mintz, T. H. (in revision). Characterizing the Difference Between Learning about Adjacent and Non-adjacent Dependencies.

Modeling the Visual Word Form Area Using a Deep Convolutional Neural Network

Sandy Wiraatmadja (csandyw@gmail.com)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

Garrison W. Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering, University of California San Diego
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

Abstract

The visual word form area (VWFA) is a region of the cortex located in the left fusiform gyrus, that appears to be a waystation in the reading pathway. The discovery of the VWFA occurred in the late twentieth century with the advancement of functional magnetic resonance imaging (fMRI). Since then, there has been an increasing number of neuroimaging studies to understand the VWFA, and there are disagreements as to its properties. One such disagreement is regarding whether or not the VWFA is more selective for real words over pseudowords[1]. A recent study using fMRI adaptation (Glezer, et al., 2009) provided evidence that neurons in the VWFA are selectively tuned to real words. This contradicts the hypothesis that the VWFA is tuned to the sublexical structure of visual words, and therefore has no preference for real words over pseudowords. In this paper, we develop a realistic model of the VWFA by training a deep convolutional neural network to map printed words to their labels. The network is able to achieve an accuracy of 98.5% on the test set. We then analyze this network to see if it can account for the data Glezer et al. found for words and pseudowords, and find that it does.

Introduction

The VWFA is a region of the visual cortex that is activated during visual alphabetical word reading, similar to how the fusiform face area (FFA) is responsive to faces. The idea of the existence of a specific region in the brain specialized for the reading process has been around since the nineteenth century, when a French neurologist, Joseph Jules Dejerine, who in 1892 reported a case of a patient with pure alexia due to a brain lesion. However, it was not until the late twentieth century, with advances in functional magnetic resonance imaging (fMRI), that the physical existence of the VWFA was discovered. Several brain imaging studies have been able to pinpoint this region to the same location within the left lateral occipitotemporal sulcus near the fusiform gyrus (Cohen et al., 2000; McCandliss, Cohen, & Dehaene, 2003; Vigneau, Jobard, Mazoyer, & Tzourio-Mazoyer, 2005; Dehaene & Cohen, 2011), shown in Figure 1. This area is found to be more responsive to visual words than any other similar stimuli, as corroborated by several lesion and interference studies. Lesions in the VWFA can cause pure alexia, where subjects experience severe visual reading impairment without any changes in ability to identify faces, objects, or even Arabic numerals (McCandliss et al., 2003; Dehaene & Cohen, 2011). The conclusion of this research is that the response of VWFA is strictly visual and prelexical, such that the words are recognized by VWFA visually, without giving them any meaning.

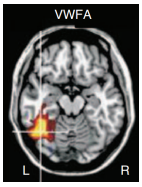


Figure 1: Activation of the VWFA in the left occipitotemporal sulcus near the fusiform gyrus (Dehaene & Cohen, 2011).

In this paper we develop a deep CNN model of the Visual Word Form Area (VWFA). We trained the network to map printed words to their labels. We then analyzed it to look for properties similar to the VWFA. We are especially interested to see if the model is consistent with the neuroimaging evidence reported by Glezer, Jiang, and Riesenhuber (2009), that suggests that the VWFA represents words discretely, and pseudowords in a more distributed manner. This is in contrast to previous studies which concluded that the VWFA is tuned to sublexical orthographic structure, and therefore has no preference for real words over pseudowords.

An influential descriptive model of the VWFA was proposed by McCandliss et al. (2003) (Figure 2). The word is first processed in ventral occipital regions V1 to V4, where the neurons are tuned to features that are increasingly complex and abstract, running posterior to anterior along the visual pathway. The features progress from horizontal and vertical bars, to individual letters, to bigrams, and so on, until the sequence of the letters is identified. This model is consistent with some neuroimaging studies (Dehaene et al., 2004; Vinckier et al., 2007). However, experiments have failed to find any further evidence of selectivity for whole words in the VWFA, leading to the hypothesis that the VWFA is tuned to the sublexical structure of a word.

We do not have space here to review the many studies of the VWFA. The reader is referred to (Dehaene & Cohen, 2011) for a relatively recent review. Here we focus on a study by Glezer et al. (2009), who used fMRI rapid adaptation (fMRI-RA) (Grill-Spector, Henson, & Martin, 2006) to investigate the nature of representations in the VWFA. fMRI-RA makes use of the fact that repeated stimuli lead to a reduction in the overall BOLD response. If a new stimulus leads to a resur-

[1] Pseudowords are usually constructed from real words by changing a consonant, and hence follow the orthographic rules of English.

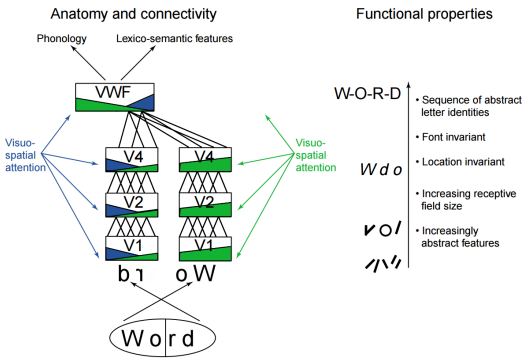


Figure 2: The anatomical and functional pathway model of visual word processing proposed by McCandliss et al. (2003).

Table 1: Some examples of prime-target word pairs that were used in the experiments done by Glezer et al. (2009).

Prime	Target Word			
	same	1L	different	
car	car	bar	lie	real words
plane	plane	place	cross	real words
health	health	wealth	ground	real words
dut	dut	dun	pim	pseudowords
nount	nount	naunt	spibe	pseudowords
shatch	shatch	chatch	jounge	pseudowords

gence of activation, that is taken as evidence that different neurons represent that stimulus, i.e., the brain sees it as different from the prime stimulus.

The subjects in the experiments were presented with prime/target pairs of real words and pseudowords, each with three different conditions: (1) "same", where the prime and target words are the same, (2) "1L", where the prime and target words differ by one letter, and (3) "different", where the prime and target words share no common letters. Examples of pairs that were used in the experiments are shown in Table 1. Of greatest interest here is their third experiment, which presented these stimuli in a within-subject design. The results (consistent with the first two experiments) showed that single letter changes gave rise to recovery of the BOLD signal that was equal to the recovery for completely different words, suggesting that even words with one letter different did not share representations. On the other hand, responses to pseudoword changes were graded, suggesting that their representations were overlapping.

Experimental Methods

As a first step in this research, we needed to develop a realistic model of the VWFA that we could analyze to look for properties that are similar to the VWFA. In order to do so, we designed a convolutional neural network for this task, created a large dataset of word images, and trained the network to recognize visual words by mapping them to their labels. Here we describe the dataset, the network, our training procedures,

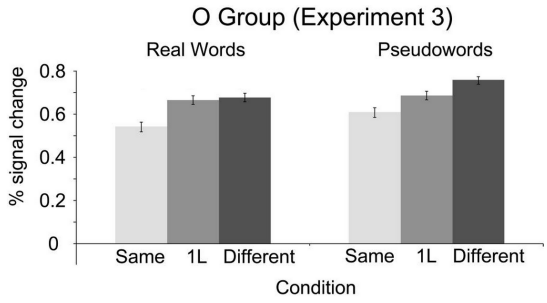


Figure 3: Mean percent signal change in the VWFA of participants in Glezer et al.'s third experiment.

and finally our experimental design to test the network on the stimuli used in Glezer et al. (2009).

Visual Words Dataset

To minimize the scope of the project, we used the 850 words of Basic English, that was created by Charles Kay Ogden. He claimed that these words are sufficient for ordinary communication in idiomatic English. The words are split into 3 different categories: (1) Operations, consisting of 100 words; (2) Things, consisting of 400 general words and 200 pictured words; and (3) Qualities, consisting of 100 general words and 50 opposite words. In order to model Glazer et al., we added 47 sets of three words each that they used for their stimuli. Since some of these words were already in the Basic English set, this resulted in 899 unique words.

Using MATLAB, the words were printed in black onto a 227x227 blank white background, and saved in PNG format. To generate an adequate variety of word images, we used 75 different font types, with three different sizes: 12, 15, and 18 pt. We also rotated the words slightly, with a rotation angle ranging from -15° to 15°, and translated the center of the text to be at least 75 and 100 pixels away from the top/bottom border and left/right border, respectively, maintaining enough space for longer words. In this manner, we generated 1,296 images per word, totaling over 1.1 million images. Some sample images in the dataset are shown in Figure 4. The images were then divided randomly into 3 sets: a training set of 899,000 images (1,000 images per word), a test set consisting of 170,000 images (200 images per word, plus four images each of the Glazer et al. words for simulating their experiments), and a validation set consisting of 86,304 images (96 images per word).

VWFA network model

Consistent with recent work in computer vision, we used a convolutional neural network (CNN) for the VWFA model. A CNN is a feed-forward neural network that uses the convolution of repeated features with small receptive fields across the image to create several feature maps. After a nonlinearity is applied, these maps are then downsampled by computing the max of a small patch of them, and the process is repeated. In this way, the receptive fields of the units get larger

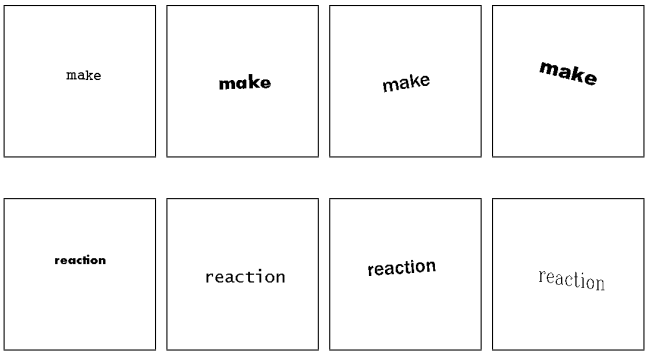


Figure 4: Example of some visual words images in the dataset

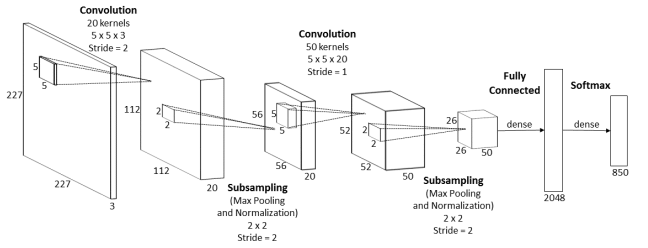


Figure 5: Architecture of the VWFANet, a modification of the LeNet-5 architecture, trained to classify visual words.

in deeper layers of the network (Krizhevsky, Sutskever, & Hinton, 2012), just as they do in visual cortex. CNNs are therefore biologically-inspired variants of multilayer perceptrons (Bengio, Goodfellow, & Courville, 2015).

A common approach in computer vision is to start with a pre-trained network, such as "Alexnet," the network that won the 2012 Imagenet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). However, for the task of recognizing images of words, Alexnet did not work well at all, presumably because it is tuned to objects, not words, which require fine-grained discrimination.

Hence we designed a modified version of LeCun's zip code reading network, LeNet-5 (LeCun, Bottou, Bengio, & Haffner, 1998). We used several refinements following ideas by Krizhevsky et al. (2012). We chose LeNet-5 as the basis for our VWFA network because both tasks involve visual character recognition. An illustration of the VWFANet architecture is shown in Figure 5.

The input to the VWFANet is a 227x227 pixel image, scaled so each pixel is in the range [0,1]. The first convolutional layer filters the input image with 20 kernels of size 5x5, and uses a stride of 2. The stride is the offset of each kernel with respect to its neighbor, so here the kernels are overlapping. This results in a feature map of dimension 112x112x20. We used Rectified Linear Units (ReLU) as the nonlinear activation function for this layer. The ReLU function is defined as $f(x) = \max(0,x)$. It has been shown empirically to learn faster than saturating nonlinearities, such as the hyperbolic tangent function $f(x) = \tanh(x)$ or the sigmoid function $f(x) = (1+e^{-x})^{-1}$. Using ReLU for this layer, which has a high dimension, produces sparse activations. These sparse features have been shown to improve the network's discriminative ability (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009).

The next step is pooling the responses of the first convolutional layer. We used max pooling on each 2x2 patch (i.e., the output of this operation is the maximum response of the four units), reducing the dimensionality to 1/4 of its previous size. Using a stride of 2, adjacent pooling units do not overlap. This produces a feature map of dimension 56x56x20.

We then applied Local Response Normalization to this output, which normalizes the activation over local regions. This scheme has heuristically been shown to aid generalization and make training faster. Each unit is divided by $(1+(\alpha/n)\sum_i x_i^2)^{\beta}$, where x is the activation of the units, n is the size of each local region, and the sum is taken over the region that is centered at that unit (Jia et al., 2014). All constants, n, α, β, are hyper-parameters. We used the values $n = 5, \alpha = 10^{-4}$, and $\beta = 0.75$, which are the same values used by Krizhevsky et al. (2012).

The third layer is another convolutional layer with 50 kernels, each of size 5x5x20, with a stride of 1. The output of this layer is a 52x52x50 feature map, to which we apply the ReLU activation function. We then use max pooling on this layer with the same parameters as before, and again use response normalization. This produces a feature map of dimension 26x26x50.

At this point, we use a fully-connected (non-convolutional) hidden layer of 2048 units, each with 26x26x50 inputs from the previous layer. Again we used ReLU as the activation function. Finally, the output of this layer is connected to an 899-way softmax. This produces a probability distribution over the 899 classes (850 Basic English words plus 49 extra words used in Glezer et al.), $P(word|input)$.

Training the Network

In order to train the network, we used Caffe, a framework designed for deep neural networks (Jia et al., 2014). We trained the weights of the VWFANet from scratch, on a single NVIDIA GeForce GTX TITAN GPU which contains 2688 cores with 6GB of memory.

The network was trained using cross-entropy loss, the minibatch stochastic gradient descent method, with momentum and weight decay. The minibatch method computes the error over a small "batch" of training examples, and then changes the weights. In our case, we used a minibatch size of 32. We used an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. The learning rate was adjusted using the inverse policy, where $\lambda_n = \lambda_0 \times (1+0.0001 \times n)^{-3/4}$, where λ_0 is the initial learning rate and n is the current iteration number. The weights initialized to have 0 mean and standard deviation equal to $1/\sqrt{m}$, where m is he fan-in to the unit. The bias in each layer was initialized to 0. We trained the network for a maximum of 100,000 iterations, while checking the validation set accuracy after every 5,000

iterations so we could stop training once the accuracy started to go down. However, even after 100,000 iterations, the error was still dropping on the validation set.

Selectivity to Real Words Experiments

We analyzed the VWFAnet to see if it has a similar selectivity for real words as the VWFA, as suggested by Glezer et al. (2009). We compared the analysis of the highest accuracy model to the results of the within-subject experimental results shown in Figure 3.

The main question now is how to model the release from adaptation effect. We assume that this effect is proportional to the difference in activation between stimuli. Hence we measure the Euclidean distance between stimuli as a proxy for signal change - a small distance means most neural activities are shared, and so there should be no or a small rise in the BOLD response, whereas a large distance means that the two stimuli do not share neural activations, so there should be a large change in the BOLD response. We measure this distance in the softmax layer. This is because each unit in this layer can be thought of as representing either a single neuron responding to a word, or a group of neurons (a Hebbian cell assembly) that work in unison.

The softmax layer imposes extreme values on the probabilities of outputs, that are likely to be more differentiated than actual neural activities. Hence we "soften" the output activity with a temperature parameter T on the softmax, as follows:

$$P(word_i|input) = \frac{\exp(x_i/T)}{\sum_{j=1}^{n}\exp(x_j/T)} \quad (1)$$

where x is the input to the softmax layer. A high temperature parameter will distribute the probability evenly, such that for $T \to \infty$, all words will have probability $1/n$. On the other hand, a low temperature will distribute the probability to only the highest value. We chose a temperature of $T = 4$, which creates a smoother probability distribution, such that there are a few labels that have non-zero probabilities, without losing the actual label information. We chose this number arbitrarily, but it is a parameter that could be fit to the data.

To model the Glezer et al. (2009) experiments, we ran each word in a pair through the network and measured the distance between their output activations. For each pair of words, we took the average of all possible input image pairs that we created out of the four generated images per word.

Results and Discussion

Network Performance

The VWFAnet achieved 98.5% accuracy on the test set. We deemed this sufficient to proceed with testing the network in a paradigm similar to that used by Glezer et al. (2009).

Some samples of correct and incorrect predictions by the VWFANet can be seen in Figure 6. We can see that even though the VWFANet overall accuracy in visual words classification task is high, there are a few outliers with less than 60% accuracy: "tall" with an accuracy of 38%, and "hour" with an accuracy of 52.5%, respectively. However, as can be seen in Figure 6, the network makes reasonable mistakes. Out of the 200 images in the test set for the word "tall", 124 of them are predicted incorrectly; 113 of those 124 images are categorized as "tail" which is very similar to "tall". Similarly, for "hour", 89 out of the 95 incorrectly predicted images are predicted to be "how" which has high resemblance to "hour".

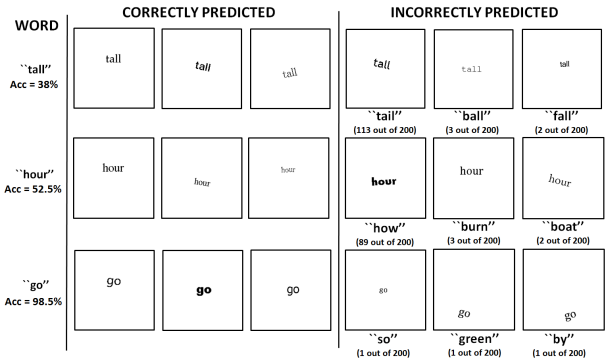


Figure 6: Some examples of correct and incorrect predictions. On the right, the incorrect predictions, and their frequency, are listed below the panels. Embiggen the pdf to see these better.

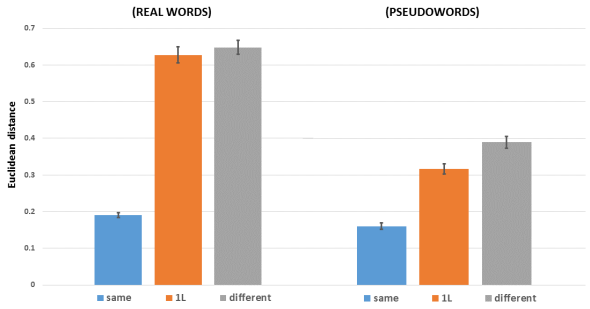


Figure 7: Euclidean distance between the activations in the output layer, with the softmax temperature set to 4. The blue bar represents the distance in activation of two instances of the same word (or pseudoword, on the right), while the orange bar represents the distance between the activations of two words or pseudowords differing by one letter. The gray bar represents the distance between two different words or pseudowords. Error bars represent the standard error of the mean. In each case, these are averages over the stimuli used in the human subject experiments.

Selectivity Analysis of VWFANet

Figure 7 shows the result of calculating the euclidean distance between the activation in the last layer of the two input images. Even though we cannot compare directly to the results of the Glezer et al. (2009) study, since the measurements are not the same, we can still observe a trend similar to the findings shown in Figure 3.

It should be noted that there is a small distance between the activations in the last layer in the "same" condition. This

is because we use different images of the same word to compute the response. For the "1L" and "different" conditions, we observe a bigger Euclidean distance compared to the one in "same", but the difference between the two conditions are not significant. This indicates that in each condition, the prime and target words activate disjoint groups of neurons, or different units in our model's output layer. This shows that even when the two stimuli differ only by one letter, the model is still able to discriminate them and so they activate different neurons, similar to when the two inputs share no common letters. A sample of the activations in the output layer for a given pair of real words is shown in Figure 8.

For the pseudowords experiment, we again notice that the smallest Euclidean distance happens when the two stimuli are images of the same pseudoword, even though the model has not been trained on these. However, what is different from the real words experiment is that we observe a gradual increase in the Euclidean distance from the "same" condition to "1L" to "different". This is also sensible because the network is trained for real words. Therefore, when the model is given a pseudoword, it activates partial representations of many different words. Figure 9 shows that the neurons from the softmax layer that get activated on pseudowords are more distributed (compare to Figure 8), but there is more overlap when the stimuli are similar, compared to when the stimuli are completely different.

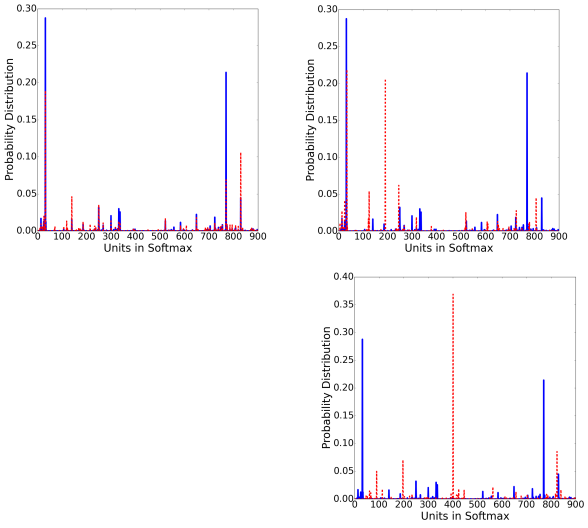


Figure 8: The activation patterns in the output layer for each pair of real words: two instances of **arm**, one of **arm** and **art**, and (bottom right) **arm** and **key**. **arm**'s activation is shown in blue. The words are sorted alphabetically, so in the upper right, the highest activation for **art** is right next to **arm**.

Conclusions

We developed a model (VWFANet) of the Visual Word Form Area using a deep convolutional neural network. Our model was trained on over a million images of almost 900 words, and achieved a very high accuracy of 98.5% on the test set.

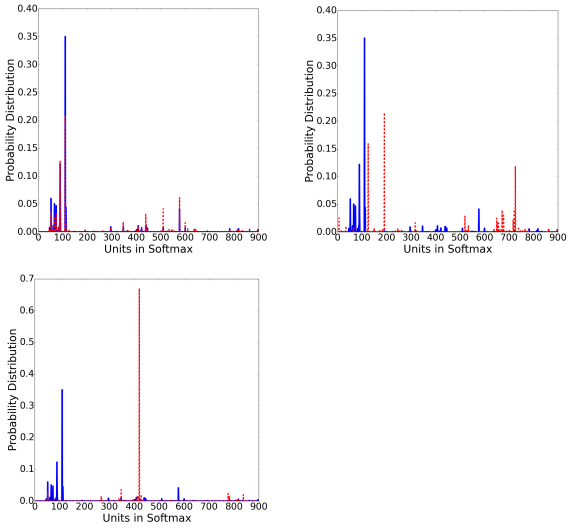


Figure 9: The activations patterns in the output layer for each pair of pseudowords: two instances of **bot**, **bot** and **sot**, and **bot** and **lan**. **bot**'s activation is shown in blue. Note that while **bot, sot** and **lan** are all real words to computer scientists, the network does not know that!

Most of the mistakes made by the network, while few, were very reasonable, such as confusing "tall" with "tail."

Our model of the fMRI-RA results was predicated on the idea that the amount of release from adaptation upon the presentation of a new stimulus should be proportional to the difference in their representations. We measured this as the Euclidean distance between the representations of words at the output layer of the network. This is consistent with previous modeling work that measures the distance between stimulus representations as the distance between their output activations, and fits human judgments of similarity quite well (Dailey, Cottrell, Padgett, & Adolphs, 2002).

Using this interpretation of release from adaptation, VWFANet was able to qualitatively reproduce the data that Glezer et al. (2009) observed in their human neuroimaging study. That is, real words, no matter how similar, were equally distant from one another at the output layer of the network. The representations developed at the output layer were also relatively sparse (Figure 8). On the other hand, pseudowords displayed more graded distances, and less sparsity at the output layer (Figure 9). These results suggest that Glezer et al.'s interpretation of their fMRI-RA results are quite reasonable - real words are represented by different populations of neurons, while pseudowords are represented by distributed patterns of activation over the representations of real words.

It is of considerable note that the fMRI-RA results support the Interactive Activation Model's representation of words versus pseudowords(McClelland & Rumelhart, 1981). In that model, words were represented by separate nodes in the network at a "word level." The word superiority effect in letter perception was explained in the model as feedback to the letter level from the word level. On the other hand, the pseu-

doword superiority effect, for example in the pseudoword "mave," was explained by the sum of feedback activation from many partially activated words at the word level (e.g., "save," "have," "wave," etc.). The current results are consistent with that account.

The current model is just the first step in modeling the VWFA. There are many other experiments that could be modeled by this same architecture. For example, Dehaene et al. (2004) compared activation to the "same" words in different cases and positions, but they also include "circular anagrams," where pairs of words can transform into one another simply by moving a single letter from the front to the back, and vice versa. An example of circular anagrams are the French words "reflet" and "trefle." If we can find some pairs of English words that are circular anagrams, we can add this to our experiments. This goes a step further than the "1L" condition because the target word is made up of exactly the same letters as the prime word, but with one location shift, instead of one letter change. If we still observe the same high percent signal change in the fMRI data, or high Euclidean distance in the VWFANet, then we provide an even stronger support for the hypothesis that the neurons in the VWFA are highly selective to whole words, instead of a broader tuning to sublexical orthographic structure of the words.

In conclusion, we have successfully trained a convolutional neural network to model the VWFA by trying to map the words to their labels. This network might not be as deep as the latest state of the art for object recognition, but it seems to perform well for the task of modeling the VWFA. The VWFANet model is able to support the recent fMRI-RA evidence that there is a preference for real words over pseudowords in the VWFA. This VWFANet model may open up more opportunities to study the VWFA and its properties further, especially with the recent increase in the neuroimaging studies for the VWFA.

Acknowledgments

This work was supported by NSF grants IIS-1219252 and SMA 1041755 to GWC.

References

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). *Deep learning.* (Book in preparation for MIT Press)

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area. *Brain*, *123*(2), 291–307.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*(8), 1158-1173.

Dehaene, S., & Cohen, L. (2011, Jun). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254262.

Dehaene, S., Jobert, A., Naccache, L., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words: Behavioral and neuroimaging evidence. *Psychological Science*, *15*(5), 307–313.

Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for highly selective neuronal tuning to whole words in the "visual word form area". *Neuron*, *62*(2), 199–204.

Grill-Spector, K., Henson, R., & Martin, A. (2006, Jan). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 1423.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009, Sept). What is the best multi-stage architecture for object recognition? In *ICCV 2009* (p. 2146-2153).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems 25* (pp. 1097–1105). Curran Associates, Inc.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the ieee* (pp. 2278–2324).

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*, 293-299.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, *88*(5), 375-407.

Vigneau, M., Jobard, G., Mazoyer, B., & Tzourio-Mazoyer, N. (2005, September). Word and non-word reading: what role for the visual work form area. *NeuroImage*, *27*(3), 694-705.

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143 - 156.

Think Fast! Mental-state Language is Related to the Speed of False-belief Reasoning in Adulthood

Erin Roby (eroby2@ucmerced.edu)
Psychological Sciences, 5200 Lake Rd.
Merced, CA 95343 USA

Rose M. Scott (rscott@ucmerced.edu)
Psychological Sciences, 5200 Lake Rd.
Merced, CA 95343 USA

Abstract

When tested appropriately, infants appear to demonstrate false-belief understanding in the first year of life. Some have argued that this is inconsistent with the well-established relationship between social experience and preschoolers' false-belief performance. We argue that these two sets of findings are not inconsistent because the ability to attribute false beliefs to others is necessary but not sufficient for false-belief performance, and we propose several ways that one social factor, hearing and using mental-state language, might relate to false-belief performance throughout the lifespan. We tested this account by examining the relationship between adults' use of mental-state language and their false-belief understanding. Participants' use of mental-state language was related to how quickly they could accurately predict the behavior of agents on the basis of desires and beliefs. These findings provide the first evidence that mental-state talk and false-belief performance are related into adulthood.

Keywords: false-belief understanding; theory of mind; mental-state reasoning; social cognition

Introduction

As members of a social species, much of our everyday life involves predicting and interpreting the behavior of other individuals. Adults often do this by attributing to others unobservable mental states such as goals and beliefs. Researchers have long been interested in the nature and development of this psychological reasoning ability.

In particular, considerable research has focused on the ability to understand that others can be mistaken, or hold false beliefs, about the world. False-belief understanding provides evidence of the ability to recognize that mental states are internal representations of the world and thus they can be false. For several decades, false-belief understanding was primarily investigated using elicited-response tasks, which require children to answer direct questions about the behavior of a mistaken agent (for a review, see Wellman, Cross, & Watson, 2001). In one such task (Baron-Cohen, Leslie & Frith, 1985), children hear a story enacted with props: Sally puts a marble in a basket and then leaves; in her absence, Anne moves the marble to a nearby box. Children are then asked where Sally will look for her marble when she returns. Beginning around age 4, children typically indicate that Sally will look in the basket, where she falsely believes the marble to be. In contrast, younger children incorrectly respond that Sally will look in the marble's actual location. This widely replicated finding suggested that the capacity to attribute false beliefs to others did not emerge until at least age 4 (e.g., Wellman et al., 2001).

However, a growing body of evidence suggests that this ability might be present much earlier than previously thought: when tested via other means, infants appear to demonstrate false-belief understanding as early as 6 months of age (for a review, see Baillargeon, Scott, & Bian, 2016). For instance, infants visually anticipate where a mistaken agent will search for an object (e.g., Southgate, Senju, & Csibra, 2007; Surian & Geraci, 2012), look reliably longer when an agent's actions are inconsistent with her false belief (e.g., Onishi & Baillargeon, 2005; Träuble, Marinović, & Pauen, 2010), and use an agent's false belief to guide their own responses to that agent (e.g., Buttelmann, Carpenter, & Tomasello, 2009; Southgate, Chevallier, & Csibra, 2010). These positive findings have led many researchers to conclude that the capacity to attribute false beliefs to others emerges in the first year of life (e.g., Baillargeon et al., 2016; Buttelmann et al., 2009; Carruthers, 2013; Kovács, Téglás, & Endress, 2010, Southgate et al., 2007).

Some researchers have challenged accounts that ascribe false-belief understanding to infants on the grounds that they fail to address the role of social factors in the development of this ability (e.g., Ruffman, 2014; San Juan & Astington, 2012). Considerable research has shown that preschoolers' performance on elicited-response false-belief tasks is correlated with individual differences in their social experience (e.g., McAlister & Peterson, 2007; Meins et al., 2003; Ruffman, Slade & Crow, 2002). In particular, there is a robust relationship between preschooler's performance on elicited-response false-belief tasks and their exposure to and personal use of mental-state language – utterances that refer to psychological states such as *think, know,* and *believe* (e.g. Ensor & Hughes, 2008; Ruffman et al., 2002). It has been suggested that if the capacity to represent false beliefs is present in the first year of life, there is "little room for social factors to play a direct role" in children's false-belief performance (p. 110, San Juan & Astington, 2012), and thus the reason for these associations is unclear (e.g. Ruffman, 2014; San Juan & Astington, 2012).

However, many researchers have argued that the capacity to represent false beliefs does not guarantee successful

performance in false-belief tasks (e.g., Baillargeon, Scott, & He, 2010; Bloom & German, 2000; Carruthers, 2013; Roth & Leslie, 1998). Studies have identified several factors that impact children's performance in both elicited-response and non-elicited-response false-belief tasks (e.g., Lewis & Osborne, 1990; Rubio-Fernández & Geurts, 2013; Scott & Roby, 2015; Yazdi et al., 2006) If, as these findings suggest, the capacity to represent beliefs is not sufficient for successful false-belief reasoning, then there is opportunity for social factors to impact false-belief performance.

Specifically, we propose that hearing and using mental-state language might be related to false-belief performance in several ways. First, individuals who frequently engage in conversations about mental states may more readily attend to and consider the mental states of others. Second, such conversations provide practice inhibiting one's own beliefs and desires in order to focus on those of another. Third, individuals who often discuss others' mental states may be more practiced at inferring others' mental states based on the (sometimes limited) information available. Finally, individuals who frequently hear and use mental-state language may more rapidly retrieve information about others' mental states from memory when necessary.

This account predicts that the relationship between mental-state talk and false-belief understanding should not be confined to performance on elicited-response tasks in the preschool years: it should be evident across the lifespan. Support for this prediction comes from recent evidence that parents' use of mental-state language is associated with 2.5-year-olds' performance on an anticipatory-looking false-belief task (Roby & Scott, 2015). In addition, deaf infants raised by hearing parents are exposed to significantly fewer mental-state references than their hearing counterparts (Morgan et al., 2014) and fail anticipatory-looking false-belief tasks (Meristo et al., 2012). These findings suggest that exposure to mental-state language is related to false-belief understanding prior to the preschool years.

Here we sought complementary evidence for this account by examining the relationship between personal use of mental-state language and false-belief performance in adults. To our knowledge, no prior studies have examined this relationship in adulthood. Although adults are certainly able to represent mental states, their mental-state reasoning exhibits considerable within- and between-individual variability (e.g., Brown-Schmidt, 2009; Harkness et al., 2010; McKinnon & Moscovitch, 2007). For instance, adults are faster at predicting the behavior of an agent who wishes to approach rather than avoid an object (e.g., Apperly et al., 2011), and adults from collectivist cultures exhibit faster and more accurate perspective taking than adults from individualistic cultures (e.g., Wu & Keysar, 2007). If, as we argue, personal use of mental-state language is related to greater attention to and consideration of others' mental states, then adults' use of such language should be related to their performance on false-belief tasks.

The present study tested this prediction. Adult participants described images of people or physical objects and then completed speeded and unspeeded mental-state reasoning tasks. We predicted that the extent to which participants used more mental-state language when discussing people than when describing objects (i.e. situations in which mental states were not relevant) would be related to their false-belief performance, and that this relationship might be most apparent when participants were required to quickly and accurately predict another individual's behavior under time pressure.

Method

Participants

71 adults (62 females) participated for course credit. An additional 4 adults were tested but excluded, 1 because she withdrew from the study, 2 because their vocabulary scores were below the 10th percentile, and 1 because he did not describe the images aloud in the picture-description task.

Materials and Procedure

All participants first completed a picture-description task. Stimuli consisted of 27 images of people or objects. Images were presented one at a time on a computer screen. Participants were instructed to describe aloud what they thought was happening in each picture. Once participants finished describing an image, they pressed a button on the keyboard to advance to the next image. Participants were told to take as much time as they needed to complete the task. During the task, audio and video were recorded to disk using the computer's built-in microphone and camera.

Participants next completed a speeded belief-desire task, a Strange Stories task, a stroop task, and a digit-span task in a randomized order.

The speeded belief-desire task was administered using PsychoPy (Peirce, 2007). On each trial, a short story about an agent was presented in the center of the computer screen. Once participants read the story, they pressed any key. The story was then replaced by a question about what the agent would do next. After 1s, two response options appeared, one in each of the bottom corners of the screen. Participants pressed either the left or right arrow key to select a response. Participants were told to respond as quickly and accurately as possible.

Participants first viewed two practice trials in which they read a story about an agent (e.g., Dennis loves pizza and hates vegetables. He is hungry and goes to the store.) and then answered a question about what the agent would do next (e.g., What will he buy?). Participants then completed 12 test trials; the stories in these trials were adapted from prior research (e.g., Bennett & Galpert, 1992; Leslie & Polizzi, 1998; Roby & Scott, 2015; Scott & Roby, 2015; Setoh, Scott, & Baillargeon, 2011; Yazdi et al., 2006). In each story, an agent held either a true or a false belief about the location of an object. In half the stories, the agent wanted to approach the object and in the other half the agent wanted to avoid the object. We varied the agent's desire because, as mentioned previously, adults find situations in

which an agent wants to avoid an object more challenging to reason about than those in which an agent wants to approach an object (Apperly et al., 2011). We therefore expected that we would be maximally likely to see individual differences in performance in avoidance situations. Belief (true, false) and desire (approach, avoid) were crossed to create four story types; participants read three stories of each type.

The Strange Stories Task (White, Hill, Happé, 2009) consisted of 14 stories: 8 described social situations (mental-state stories), 3 described events with physical causes (e.g. an old woman slips on her icy doorstep, causing her to need an X-ray; physical stories), and 3 described natural phenomena (e.g. Sun melting snow causes the formation of puddles; nature stories). Stories were presented in one of four orders; each order was randomized with the constraint that no more than 2 stories of a given type occurred in a row. Once participants finished reading the story, it was removed and they were asked to provide a written response to a question about what happened in the story. For example, in one mental-state story, a nervous woman mistakes an innocent man for a robber. She tells him to take her purse and asks him not to hurt her. The question asked, "Why did she say that?" Participants were told to take as much time as they needed to complete the task.

The stroop task (Stroop, 1935) was administered using Inquisit 4.0.3 (2004). The task consisted of 56 experimental and 28 control trials. In experimental trials, participants saw one of 4 color words (red, blue, green, black); in half the trials the color of the text matched the word (congruent trials) while in the remaining trials it did not (incongruent trials). In control trials, participants saw a box displayed in one of the four colors. Participants were asked to respond to the color of the stimulus as quickly and accurately as possible by pressing a corresponding arrow key.

The digit-span task was also administered using Inquisit (e.g., Woods et al., 2011). On each trial, participants viewed a series of digits one at a time on the screen. The digits then disappeared, and after 1s a box appeared. Participants were instructed to type the digits they had seen in either forwards or backwards order using the keyboard. They completed 14 forward trials followed by 14 backwards trials.

The stroop and digit-span tasks were included to control for differences in inhibitory control and working memory, which correlate with adults' mental-state reasoning (e.g. German & Hehman, 2006, McKinnen & Moskovitch).

Finally, participants completed the Peabody Picture Vocabulary Task (PPVT-4; Dunn & Dunn, 2012). On each trial, participants indicated which of four images corresponded to a spoken word. This measure was included to control for differences in English fluency, which could impact participants' picture descriptions or comprehension of the stories in the mental-state reasoning tasks.

Coding and Analysis

Picture-description task We transcribed participants' descriptions of each image and then used the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) program to identify utterances that included mental-state terms. These terms were divided into 5 categories: cognition (e.g., *think*, *know*), desire (e.g., *want*, *need*), emotion (e.g., *happy*, *angry*), modulations of assertion (e.g., *maybe*, *might*), and other mental-state words (e.g., *remember*, *wonder*). The mental-state words included words from previous work investigating parental mental-state talk during parent-child interactions (e.g., Ruffman et al., 2002) as well as synonyms for those words that are not generally used in child-directed speech (e.g., *devastated*, *pondering*). We omitted the word "like" because adults use it in several ways that are unrelated to mental states (e.g. "It's, like, two people at the beach").

For each participant, we calculated the percentage of utterances in each description that contained each category of word. For each category, we then averaged across all the pictures of a given type (i.e. people, objects). Next, we calculated difference scores for each category of mental-state talk by subtracting the percentage of mental-state utterances made in reference to object scenes from the percentage of mental-state utterances made in reference to images of people (see Table 1). This difference score thus reflected the extent to which participants selectively used mental-state language when discussing other individuals.

Table 1: Mean (SD) percentage of utterances in the picture-description task containing mental-state terms by image type, as well as difference scores (people – object)

	People	Object	Difference Score
Cognitive	5.6 (4.4)	4.8 (5.9)	.9 (6.0)
Desire	1.3 (1.7)	.2 (.6)	1.2 (1.8)
Emotion	11.2 (5.7)	.8 (2.2)	10.4 (5.7)
Modulations of assertion	22.3 (12.2)	18.0 (15.0)	4.3 (11.9)
Other	2.9 (3.0)	1.3 (2.9)	1.5 (3.2)
Total	36.2 (13.0)	23.8 (15.4)	12.5 (14.2)

Belief-desire Task One story was eliminated because more than 50% of participants responded incorrectly, suggesting a problem with the item. For the remaining 11 items, we eliminated responses with reaction times less than .15s or more than 3 standard deviations above the mean (n = 20). For each participant, we calculated the percentage of correct responses and average reaction time across all items, as well as separately for each story type (true-belief approach, true-belief avoid, false-belief approach, false-belief avoid).

Strange Stories Task Consistent with previous research (e.g. White et al., 2009), responses were scored on a 3-point scale. Participants received 0 points for responses that involved irrelevant or incorrect facts, 1 point for responses that were factually correct but failed to acknowledge key elements that were necessary for interpreting the story, and

2 points for fully correct responses that were factually correct and referenced the components of the story necessary for accurate interpretation. We calculated participant's average score for each story type (i.e. mental state, physical, and nature).

Inhibitory Control Trials in the stroop task with reaction times less than .15s or 2.5 standard deviations above the mean were excluded. We then subtracted participants' average reaction time on congruent trials from their average reaction time on incongruent trials.

Working Memory Consistent with prior work (German & Hehman, 2006), we used participants' backwards digit span (i.e. the highest number of digits participants recalled backwards correctly before producing two consecutive errors) as a measure of their working memory.

Receptive Vocabulary The PPVT was scored according to published procedures for this measure (Dunn & Dunn, 2012). Participants' raw scores were converted to standardized scores for analysis.

Results

Table 2: Mean (SD) percentage of correct responses and reaction times in seconds on the belief-desire task, separately by story type

	Percent Correct		Reaction Time	
	True Belief	False Belief	True Belief	False Belief
Approach	90 (21)	95 (30)	1.4 (.6)	1.4 (.7)
Avoid	79 (23)	78 (28)	1.7 (.9)	1.6 (.9)

Table 2 provides descriptive statistics for participants' performance in the belief-desire task. An ANOVA on the percentage of correct responses with desire (approach, avoid) and belief (true, false) as within-subject factors revealed a significant effect of desire, $F(1, 69) = 20.78$, $p < .001$. No other effects were significant, all Fs < 1. A corresponding ANOVA on participants' reaction times revealed a significant effect of desire, $F(1, 69) = 9.92$, $p = .002$; no other effects were significant, all Fs < 1. Consistent with prior findings (Apperly et al., 2011), participants were slower and less accurate when the agent wished to avoid rather than approach an object.

An ANOVA on participants' performance on the Strange Stories task with story type as a within-subject factor revealed no effect of story type, $F(2, 140) = 1.31$, $p = .27$. Participants performed equally well on mental-state stories ($M = 1.52$, $SD = .26$), physical stories ($M = 1.61$, $SD = .39$), and nature stories ($M = 1.54$, $SD = .34$).

To examine relationships between tasks, we computed partial correlations controlling for participants' inhibitory control, working memory, and receptive vocabulary. We report p-values corrected for multiple comparisons ($\alpha = .05$ one-tailed; Benjamini & Hochberg, 1995).

Participants' percentage of correct responses on the belief-desire task was significantly partially correlated with their performance on the mental-state stories, $r = .31$, $p = .03$, but not the physical stories, $r = .20$, $p = .15$, or nature stories, $r = .12$, $p = .25$. This suggests that the belief-desire task and mental-state stories assessed a common mental-state reasoning ability.

We next examined relationships between participants' mental-state language and their performance on the belief-desire task. We specifically examined correlations with the cognitive and total difference scores, as these categories are related to preschoolers' false-belief performance.

There were no significant relationships between participants' percentage of correct responses on the belief-desire task (overall or by story type) and their difference scores for either cognitive or total mental-state terms, all ps > .45. However, both difference scores were significantly negatively correlated with participants' reaction times on false-belief avoid stories (Table 3). Participants who used more cognitive and total mental-state terms in situations where mental states were relevant (to discuss people rather than objects) were significantly faster to respond correctly in situations where they needed to simultaneously consider an agent's false belief and desire to avoid an object.

Table 3: Partial correlations between mental-state difference scores and reaction times (correct responses only) on the belief-desire task

	True Belief		False Belief	
	Approach	Avoid	Approach	Avoid
Cognitive	-.09	-.17	-.07	-.28*
Total	.09	-.01	.00	-.25*

Note: In line with Apperly et al., 2011, only reaction times for correct responses were included. $df = 63$; * $p < .05$.

Finally, there were no significant relationships between participants' difference scores for cognitive or total mental-state terms and their performance on the Strange Stories task, all ps > .18.

General Discussion

Recent evidence suggests that the ability to attribute false beliefs to others may emerge in the first year of life. However, some have challenged this conclusion on the grounds that it is at odds with the well-established associations between social factors and preschoolers' false-belief performance (e.g. Ruffman, 2014; San Juan & Astington, 2012). If infants understand false beliefs, then why would social experience predict preschooler's performance on elicited-response false-belief tasks? We proposed that these two sets of findings are not inconsistent because the ability to attribute false beliefs to others is

necessary but not *sufficient* for false belief performance, and thus social experience might very well facilitate the latter throughout the lifespan. For instance, individuals who hear and use more mental-state language might more readily attend to, infer, and retrieve others' mental states.

Here we tested this claim by examining the relationship between mental-state language and false-belief performance in adults. Participants described images of people and objects and then completed two mental-state reasoning tasks. Participants who exhibited a greater tendency to use mental-state terms (and cognitive terms in particular) when discussing people rather than objects were significantly faster at correctly predicting the behavior of an agent who held both a false belief and a desire to avoid an object.

These findings constitute the first evidence that personal use of mental-state language is related to false-belief performance in adulthood. The fact that this relationship was specific to false-belief avoidance situations likely reflects the fact that these situations impose greater cognitive load than those involving true beliefs or approach desires (e.g., Apperly et al., 2011; Leslie & Polizzi, 1998). In particular, false-belief avoidance stories require 'double inhibition': participants must inhibit their own knowledge about the location of an object in order to appreciate the agent's false belief while also recognizing the agent's desire to *not* approach that object. Individuals who engage in more discussions of others' mental states may have more practice inhibiting their own beliefs and desires while focusing on those of another, allowing them to more rapidly predict an agent's behavior in this particular context.

Adults' use of mental-state language was not related to the accuracy of their responses on either mental-state reasoning task. This suggests that use of mental-state language is related to how quickly one attends to, encodes, and retrieves mental-state information rather than one's ability to generate correct predictions or explanations based on that mental-state information. This might explain why several prior studies have failed to find associations between school-aged children's use of mental-state language and their performance on the Strange Stories task (e.g. Charman & Shmueli-Goetz, 1998; Meins, et al., 2006).

This pattern of findings is also consistent with recent work examining the relationship between parental mental-state talk and toddlers' false-belief performance. Roby and Scott (2015) tested 2.5-year-olds on a false-belief task with two consecutive test trials: an anticipatory-looking trial followed by a preferential-looking trial. Parental use of mental-state language predicted children's performance on the anticipatory-looking trial but not the preferential-looking trial. This difference could reflect the fact these two trials had different requirements for successful performance (prediction vs. post hoc analysis). However, our results suggest that this difference may instead reflect trial order. If exposure to mental-state language is related to how quickly children are able to attend to, encode, and retrieve mental-state information, then this would lead to stronger associations on the first test trial than the second.

Our findings complement recent findings suggesting that parental mental-state talk is associated with infants' and toddlers' false-belief performance (e.g. Meristo et al., 2012; Morgan et al., 2014; Roby & Scott, 2015). Together, these results suggest that the relationship between mental-state language and false-belief performance is not constrained to the preschool years or to elicited-response tasks: hearing and using mental-state talk facilitates the ability to reason about mental states throughout life.

References

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental continuity in theory of mind: Speed and accuracy of belief–desire reasoning in children and adults. *Child Development*, *82*, 1691-1703.

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*, 110-118.

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, *67*, 159-186.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37-46.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Bennett, M., & Galpert, L. (1992). Complex belief-desire reasoning in children. *Social Development*, *1*, 201-210.

Bloom, P. & German, T. P. (2000). Two reasons to abandon the false-belief task as a test of theory of mind. *Cognition*, *77*, B25-B31.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*, 893-900.

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*, 337-342.

Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, *28*, 141-172.

Charman, T., & Shmueli-Goetz, Y. (1998). The relationship between theory of mind, language and narrative discourse: An experimental study. *Cahiers de Psychologie Cognitive*, *17*, 245-271.

Draine, S. C. (2004). Inquisit 2.0. 50401 [Computer software]. *Seattle, WA: Millisecond Software*.

Dunn, L. M., & Dunn, D. M. (2012). Peabody Picture Vocabulary Test, (PPVT-4). *Johannesburg: Pearson Education Inc*.

Ensor, R., & Hughes, C. (2008). Content or connectedness? Mother-child talk and early social understanding. *Child Development*, *79*, 201-216.

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in 'theory of mind':

Evidence from compromised belief-desire reasoning in old age. *Cognition*, *101*, 129-152.

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: Effect of sad versus happy mood induction. *Cognition and Emotion*, *24*, 497-513.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830-1834.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science*, *1*, 247-253.

Lewis, C., & Osborne, A. (1990). Three−year−olds' problems with false belief: Conceptual deficit or linguistic artifact? *Child Development*, *61*, 1514-1519.

McAlister, A., & Peterson, C. (2007). A longitudinal study of child siblings and theory of mind development. *Cognitive Development*, *22*, 258-270.

McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*, *102*, 179-218.

Meins, E., Fernyhough, C., Johnson, F., & Lidstone, J. (2006). Mind−mindedness in children: Individual differences in internal−state talk in middle childhood. *British Journal of Developmental Psychology*, *24*, 181-196.

Meins, E., Fernyhough, C., Wainwright, R., Clark-Carter, D., Das Gupta, M., Fradley, E. & Tuckey, M. (2003). Pathways to understanding mind: Construct validity and predictive validity of maternal mind-mindedness. *Child Development*, *74*, 1194–1211.

Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M. (2012). Belief attribution in deaf and hearing infants. *Developmental Science*, *15*, 633-640.

Morgan, G., Meristo, M., Mann, W., Hjelmquist, E., Surian, L., & Siegal, M. (2014). Mental state language and quality of conversational experience in deaf and hearing children. *Cognitive Development*, *29*, 41-49.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255-258.

Peirce, JW (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [Computer software]. *Austin, TX: liwc.net*.

Roby, E. M. & Scott, R. M. (2015, March). *How does social input influence false-belief reasoning? Parent mental-state language and toddlers' false-belief understanding.* Paper presented at the biennial meeting of the Society for Research in Child Development, Philadelphia, PA.

Roth, D., & Leslie, A. M. (1998). Solving belief problems: Toward a task analysis. *Cognition*, *66*, 1-31.

Rubio-Fernández, P. & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*, 27-33.

Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review*, *34*, 265-293.

Ruffman, T., Slade, L., & Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Development*, *73*, 734-751.

San Juan, V., & Astington, J. W. (2012). Bridging the gap between implicit and explicit understanding: How language development promotes the processing and representation of false belief. *British Journal of Developmental Psychology*, *30*, 105-122.

Scott, R. M., & Roby, E. (2015). Processing demands impact 3-year-olds' performance in a spontaneous-response task: New evidence for the processing-load account of early false-belief understanding. *PloS ONE*, *10*, e0142405.

Setoh, P. Scott, R. M., & Baillargeon, R. (March, 2011). *False-belief reasoning in 2.5-year-olds: evidence from an elicited-response low-inhibition task.* Paper presented at the biennial meeting of the Society for Research in Child Development, Montreal, Canada.

Southgate, V., Chevallier, C., & Csibra, G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*, *13*, 907-912.

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*, 587-592.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology*, *30*, 30-44.

Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do infants understand others' beliefs? *Infancy*, *15*, 434-444.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*, 655-684.

White, S., Hill, E., Happé, F., & Frith, U. (2009). Revisiting the strange stories: revealing mentalizing impairments in autism. *Child Development*, *80*, 1097-1117.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., ... & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*, 101-111.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science*, *18*, 600-606.

Yazdi, A. A., German, T. P., Defeyter, M. A., & Siegal, M. (2006). Competence and performance in belief-desire reasoning across two cultures: The truth, the whole truth and nothing but the truth about false belief? *Cognition*, *100*, 343-368.

Simple Trees in Complex Forests: Growing Take The Best by Approximate Bayesian Computation

Eric Schulz[1] (e.schulz@cs.ucl.ac.uk), **Maarten Speekenbrink**[1] (m.speekenbrink@ucl.ac.uk) &
Björn Meder[2] (meder@mpib-berlin.mpg.de)
[1]Department of Experimental Psychology, University College London, Gower Street, London WC1E 6BT, UK
[2]Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

How can heuristic strategies emerge from smaller building blocks? We propose Approximate Bayesian Computation (ABC) as a computational solution to this problem. As a first proof of concept, we demonstrate how a heuristic decision strategy such as Take The Best (TTB) can be learned from smaller, probabilistically updated building blocks. Based on a self-reinforcing sampling scheme, different building blocks are combined and, over time, tree-like non-compensatory heuristics emerge. This new algorithm, coined *Approximately Bayesian Computed Take The Best* (ABC-TTB), is able to recover data that was generated by TTB, leads to sensible inferences about cue importance and cue directions, can outperform traditional TTB, and allows to trade-off performance and computational effort explicitly.

Keywords: Heuristics, Take The Best, Approximate Bayesian Computation, Reinforcement Learning

Introduction

How can heuristic strategies emerge from smaller building blocks? Consider the heuristic *Take The Best* (TTB; Gigerenzer & Goldstein, 1996) as an example: TTB is a lexicographic, non-compensatory strategy to decide which of two objects scores higher on an unobserved criterion variable, based on accessible pieces of information (cues). The heuristic looks up cues sequentially (lexicographically), with the order of cues determined by their importance (predictive value). TTB is non-compensatory because once a cue discriminates between the two options, a decision is made irrespectively of the value of the other cues. Although relatively simple, TTB has been shown to outperform more sophisticated models such as regression models, nearest neighbour classifiers, and classification and regression trees, across many data sets (Gigerenzer & Brighton, 2009).

A key question is how such a successful but domain-specific strategy is learned (Gigerenzer & Gaissmaier, 2011). According to the *adaptive toolbox* metaphor, there are three building blocks that can be used to implement decision heuristics. *Search rules* dictate how information is acquired, for example looking up cues ordered by their predictive value. *Stopping rules* determine when to stop searching, for example as soon as a cue discriminates between the two options. Lastly, *decision rules* determine the actual decision after search has stopped, for example deciding which of the two objects scores higher on the target criterion, depending on the cue direction (i.e. whether a cue is positively or negatively related to the criterion). Considering TTB, we see that all of the building blocks are present.

A viable theory of heuristic decision making should also explain how heuristics strategies like TTB are learned, that is, how they emerge during learning from their constituent building blocks. One idea is that a decision maker starts off with some prior intuitions about which cues might be helpful, and then learns from experience which cue orders and directions are successful. We formalize this idea using *Approximately Bayesian Computed Take The Best* (ABC-TTB), a computational approach towards learning TTB adaptively by approximate Bayesian computation. When heuristic building blocks are fed into it, successful decision strategies are learned on the fly, paralleling a Bayesian reinforcement learning algorithm. Note that our approach critically differs from approaches that model *strategy selection* through reinforcement learning mechanisms (e.g., Rieskamp & Otto, 2006) as we do not use strategies (e.g., TTB vs. weighting-and-adding) but building blocks from within a model class (e.g., different TTB variants) as the unit of reinforcement. Our approach is based on the acceptance and rejection of simple simulations that explore the usefulness of proposal models, emerging from probabilistically updated building blocks.

We show that ABC-TTB:

1. recovers TTB when TTB provides the task structure.
2. generates sound inference based on learned cue order and directions.
3. can outperform traditional TTB and other models.
4. can balance performance and computational effort.

This is a first proof of concept of how Take The Best can emerge from smaller building blocks.

Approximate Bayesian Computation

Bayesian inference concerns updating prior beliefs in light of observed data. Given a prior distribution $\pi(\theta)$ reflecting our initial beliefs about an unknown parameter θ, the data $\mathcal{D}$ affect the posterior belief $p(\theta|\mathcal{D})$ only via the likelihood $p(\mathcal{D}|\theta)$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\pi(\theta)}{\int_\theta p(\mathcal{D}|\theta)\pi(\theta)\mathrm{d}\theta} \propto p(\mathcal{D}|\theta)\pi(\theta) \quad (1)$$

Bayesian inference thus models subjective beliefs as a mixture of prior assumptions and incoming data, providing a coherent framework to model inference under uncertainty. Accordingly, the approach has been used to model cognition in many domains (Oaksford & Chater, 2007). Since full Bayesian inference is frequently intractable (e.g., due to the curse of dimensionality, see Bellman, 1961), powerful sampling mechanisms have been developed that approximate the posterior distribution by means of Markov chain Monte Carlo methods (Gilks, 2005). Here, we focus on an alternative approach to approximate posterior inference: *Approximate Bayesian Computation* (ABC; Turner & Van Zandt, 2012). ABC is an approximative method that, instead of computing the required likelihood directly, substitutes it with surrogate simulations of a model (Csilléry et al., 2010), and then checks whether the output of these simulations comes close enough to the actual data. This simple mechanism of model simulations and reality checks provides a useful tool to approximate posteriors and has been applied to many scenarios in which the true underlying likelihood is hard to assess.

A formal description of ABC is shown in Algorithm 1. The algorithm samples a proposed parameter from a prior $\pi(\theta)$ and plugs the proposal into a model $\mathcal{M}$ in order to simulate an output $\mathbf{y}^*$ from the given data $\mathcal{D}$. It then calculates a summary statistic $\mathcal{S}$ of both the simulated data and the real data and accepts proposals that have produced a summary that is close enough (measured by a distance ρ) to the summary of the real data, allowing for some error ϵ. The proximity of the two summary statistics is normally estimated by the difference of the statistics δ, for example the Euclidean distance between two means. A more intuitive explanation of this algorithm is that an agent has some subjective beliefs about how the world works and repeatedly checks whether or not these beliefs can, on average, produce similar patterns to the ones observed. If they can, the model proposals are accepted and (possibly) reinforced. If they cannot, the proposals are rejected. Even though this algorithm is based on a computationally simple reinforcement scheme, it has been shown to provide reasonable approximate posteriors in many scenarios, in particular when inferring tree-like structures (Beaumont, Zhang, & Balding, 2002).

Algorithm 1 Approximate Bayesian Computation

Require: $\pi(\theta)$; model $\mathcal{M}$; data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$; tolerance ϵ, required sample size η
  while $j < \eta$ **do**
    Sample $\theta^* \sim \pi(\theta)$
    Simulate $\mathbf{y}^* = \mathcal{M}_{\theta^*}(\mathbf{X})$
    Calculate $\delta = \rho\big(\mathcal{S}(\mathbf{y}), \mathcal{S}(\mathbf{y}^*)\big)$
    if $\delta \leq \epsilon$ **then**
      set $\theta_j = \theta^*$ and $j = j + 1$
    else
      reject θ^*
  end while

ABC-TTB: Growing heuristic strategies

We next show how ABC can be applied to the problem of learning heuristics from smaller building blocks. More specifically, our ABC-TTB algorithm can yield a similar tree-like structure as TTB, but learns the structure of the tree (cue order and cue directions) on the fly while it makes decisions and receives feedback regarding the usefulness of the model proposals.[1] Thus, ABC-TTB randomly generates proposal trees and checks how well these trees perform in a subset of the data. Successful cues are reinforced through a kind of Bayesian reinforcement learning algorithm (Poupart, 2010). This way, ABC-TTB starts out with small building blocks, combines and tests them, and –over time– heuristic structures emerge that reflect the learning experience.

Figure 1 uses a Polya urn sampling scheme to illustrate the reinforcement-based learning mechanism underlying ABC-TTB. The model learns through updating distributions over cue importance and cue directions, from which model proposals are generated and tested. Figure 1a shows the urn representing three available cues, $c1$ to $c3$. A cue is sampled from the current distribution; in this case $c2$ is sampled from a uniform distribution over the cues. Next a cue direction is sampled from an independent urn associated with this particular cue (Figure 1b). From these two building blocks a proposal model is generated (Figure 1c) and tested against the data (Figure 1d). If the cue correctly predicts which object scores higher on the target criterion, the sampled cue and direction are reinforced by putting a $c2$ ball into the cue urn and a P ball into the urn associated with $c2$'s direction. If the prediction is wrong, the proposal is rejected and no reinforcement takes place. This process is repeated several times, and over time the distributions resulting from reinforcing successful cue combinations make it more likely that successful model proposals are generated.

Now consider a probabilistic implementation of this learning process. Associated with each cue is a Beta prior ($B(1, 1)$) that induces an implicit likelihood for each cue to generate a successful prediction. From these priors, a probability is independently sampled for each cue, which are used to generate a proposal tree by normalizing these probabilities (so they sum up to 1) and then sampling a cue according to these probabilities (similar to Figure 1a). The drawn cue is the top node of

[1] Even though TTB is not commonly labelled as a tree, the model can be represented as a tree structure, as shown in Figure 2.

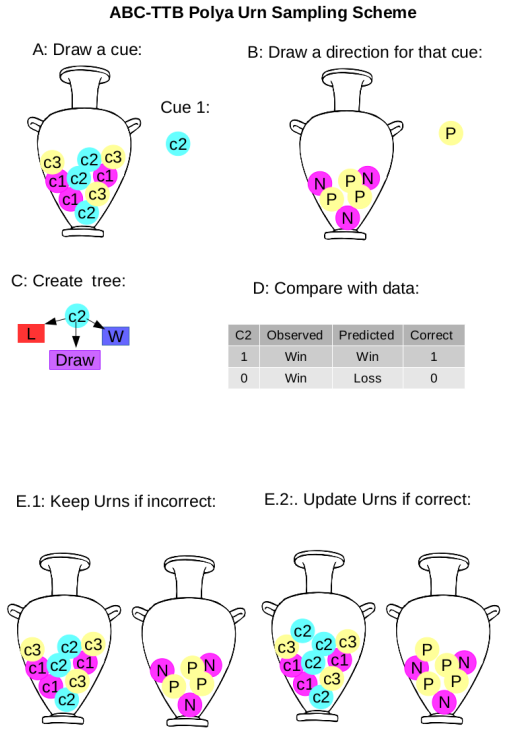


Figure 1: Sketch of ABC-TTB's sampling scheme. c1-c3 are three example cues; N and P are negative or positive weights.

the tree (i.e. the most important cue). The second cue is drawn from the remaining cues by first renormalizing the associated probabilities and then sampling accordingly. Thus, cues are drawn in order without replacement from the set of all cues to construct a lexicographic decision tree. Consider a situation with three cues, where $c1$ has a probability of 0.8 to make a correct prediction, $c2$ a probability of 0.6, and $c3$ a probability of 0.2 (all sampled from the corresponding Beta priors). To draw the top node of the tree, these probabilities are normalized to sum to 1. For example, the probability that $c1$ is sampled as the top node is $.8/(.8 + .6 + .2) = .5$, while the second and third cue have probabilities of .375 and .125 respectively. Imagine that $c1$ was drawn as the top node. The remaining probabilities for $c2$ and $c3$ are then re-normalized to .75 and .25, respectively, and the second cue is sampled. This process continues until all cues have been sampled. Associated to each cue is a second Beta prior which models the probability of the cue direction (i.e., whether it is positively or negatively related to the outcome). The resulting beta-binomial distribution model assigns to each cue the probability to have a positive direction. The cue direction is independently sampled for each cue and attached to the nodes to generate the proposal tree.

The proposal tree is then used to make predictions for a randomly sampled subset of the available data (e.g., the training sample in cross-validation). The size of the sampled subset can be varied by a parameter $0 < \phi \leq 1$, which means that the subset has to be more than 0% and can contain up to 100% of the data. If more than (1-ϵ) $\times$ 100% of the predictions are correct, the proposal is accepted and the cues involved within that tree, as well as their directions, get updated by adding a success to their posterior beta-binomial-distribution in proportion to how often they have actually made a difference (how often they successfully discriminated between the cues in the sub-sample). As the Beta distributions are updated based on a cue's success, more successful cues are sampled more often as the top node in proposal trees than less successful cues. If less than (1-ϵ) $\times$ 100% of the predictions are correct, the proposal tree is rejected and no update happens for the cues or their directions. This whole process repeats until η successful proposal trees have been generated. At this point, the updated Beta distributions are approximations to the posterior distributions of cues and directions.

ABC-TTB makes predictions for a new paired comparison based on the posterior distributions of cues and directions resulting from the learning process. The posteriors are used to generate ω proposal trees and the final decision is given by the modal prediction from all sampled proposal trees' predictions. In summary, the ABC-TTB algorithm constitutes a computational solution to how cue importance and cue directions are learned simultaneously through a simple reinforcement process. The algorithm shows how a heuristic can emerge from combining and reinforcing its building blocks, resulting in an algorithm that is able to find and utilize simple trees within the complex forest of all possible trees.

Recovering TTB

A first sanity check is to see whether ABC-TTB can recover the true underlying model from a data set that has been generated by the TTB heuristic. As TTB is a subset of all possible models (cue orders and cue directions) captured by ABC-TTB, it should be able to recover it from simulated data. We simulated a paired-comparison task with four cues, according to the TTB model shown in Figure 2. For each cue, the two objects in a paired comparison could either be identical (a draw, with probability .5), or the cue could be present in one object and absent in the other (a win, or loss, each with probability .25). In this environment the outcome of a decision is determined by the first discriminating (winning or losing) cue, just like in the original TTB. The cues were ordered by importance as $c1$, $c2$, $c3$, and $c4$. If the third cue $c3$ did not discriminate (a draw), the outcome was determined randomly. As a result, the fourth cue, $c4$, was uninformative in the simulation. The decision strat-

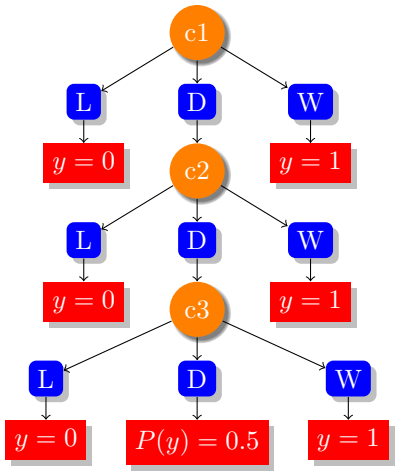


Figure 2: Underlying TTB-model generating the data set. c4 is not shown as it is not predictive.

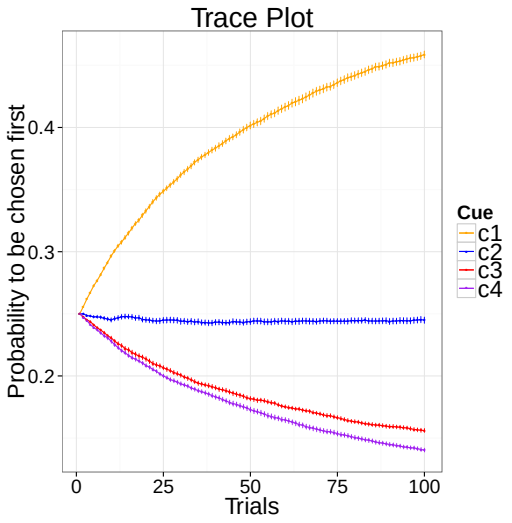


Figure 3: Trace plot of probability to be chosen first over 100 accepted proposals, averaged over 100 simulations. Legend is ordered by final cue importance. Bars represent standard errors.

egy generating this data set is identical to TTB, i.e. it can be seen as a comparison between two different items A and B, where a win means that item A has something that item B has not and an outcome of $y = 1$ means that item A is better than item B.

A data set of size $n = 1000$ was generated and the ABC-TTB algorithm was run with $\epsilon = 0.1$ and $\phi = 0.1$ until $\eta = 100$ proposed trees were accepted. Figure 3 shows the cue importance over time. The traces were calculated by dividing each Beta's posterior mean by the total sum of posterior means of all cues. This then equals the probability of a cue to be chosen as the top node in the tree and can act as a proxy variable for a cue's overall importance. Notice that the actual magnitude of that probability is not as important as the recovered order. The simulation was repeated 100 times and the probabilities were averaged per step. It can be seen that –even after a few accepted proposals– the mode tree (the most likely tree) is the same as the one generating the data. Shortly after the start of the simulation, the cue order grows to the correct order and less important cues are chosen less frequently, exactly as expected given the underlying structure. Additionally, the uninformative cue $c4$ approached a probability of 0.

Inference and performance for real data

Next, we checked whether or not ABC-TTB yields sensible and interpretable inferences in a real-world data set. For this, we applied our algorithm to the classic city size data, which has been frequently used in previous research (e.g., Gigerenzer & Brighton, 2009) and contains the population size of different German cities and whether or not they have an intercity train line, an exposition site, a soccer team, a university, are the national capital, have their own license plate, are located in the former east Germany, are a state capital, and have their own industrial belt . This data set contains 81 objects (German cities) with 9 different binary variables that can be used to decide which of two cities has more citizens (the unobserved target criterion). Figure 4 shows the trace plots for the city size simulation ($\epsilon = 0.35$, $\phi = 0.01$)[2] over 200 accepted proposal trees. Again, the results of ABC-TTB look sensible, finding that having a intercity train line, an exposition site, a major league soccer team, and a university are the four most important cues. This cue order correspond more to a frequency-adjusted validity taking into account how often a given cue can be utilized (cf. Newell et al., 2004).

Next, we compared the performance of ABC-TTB to the performance of classic TTB within the city size data set. Additionally, we also tested classification trees (CART), another tree-based classification algorithm. For the comparison, we split up the data into learning sets of $s = [5, 10, \ldots, 90]\%$, fitted TTB (using the classic cue validity), CART, and ABC-TTB ($\epsilon = 0.5$, $\rho = 0.1$) to this data and then assessed their predictive accuracy in the remaining test set. This procedure was repeated 50 times for every learning set size and the results were averaged over all trials. Results can be seen in Figure 5. ABC-TTB consistently outperformed classic TTB, which itself performed better than CART (c.f. Gigerenzer & Brighton, 2009). This is due to the fact that ABC-TTB learns the tree structure by sub-sampling items from the learning set and makes predictions based on aggregations over possible outcomes.

[2]We used different parameters as the data set was bigger and contained more variables. See next section on a more detailed treatment of hyper-parameters.

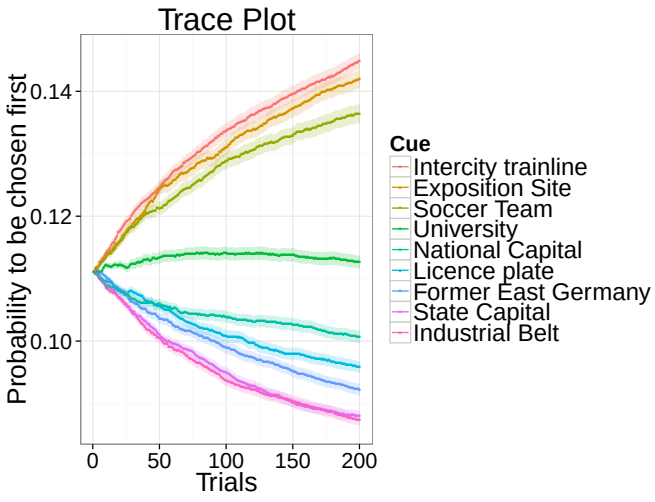


Figure 4: Trace plot of probability to be chosen first over 200 accepted proposals, averaged over 100 simulations. Legend is ordered by final cue importance.

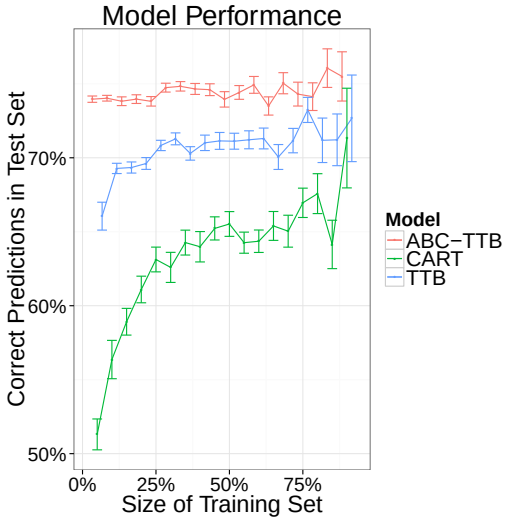


Figure 5: Performance of TTB, ABC-TTB, and CART in the city size data set. Performance increases with training set. ABC-TTB performs best overall.

As it can be seen as an ensemble of different trees, ABC-TTB will naturally generate less prediction errors than classic TTB. The posterior of ABC-TTB defines a distribution over possible trees, some of which will generate divergent predictions. Therefore, predictive variances can be calculated, which is important for future research on exploration-exploitation problems (Schulz, Konstantinidis & Speekenbrink, 2015).

Performance-effort trade-off

Another benefit of ABC-TTB is that performance and computational effort can be traded-off against each other explicitly. Assessing this trade-off lies at the heart of approaches that frame a learning agent as exhibiting a meta-reasoning mechanisms sensitive to the costs of cognition (Gershman, Horvitz & Tenenbaum, 2015), a hypothesis that can only be evaluated by making the accuracy-effort-off explicit.

As $(1-\epsilon)$ defines the proportion of predictions that have to be correct in order for a proposal tree to be accepted, decreasing ϵ will lead to more rejections of proposed trees and to acceptance of only higher quality proposals. The same holds true for the proportion of sampled data points ϕ. The higher the proportion of the sampled data set, the better the proposed tree has to be, and the longer it takes to find good trees. We can adjust these parameters explicitly to gauge the interaction between performance and computational effort. Therefore, we created all 81 possible combinations of $\epsilon = \{0.1, 0.2, \ldots, 0.9\}$ and $\rho = \{0.1, 0.2, \ldots, 0.9\}$ and applied the resulting models to the scenario in which TTB had generated the data as in the "Recovering TTB" analysis above. We tracked the number of proposed samples as a proxy variable of computational effort, averaged over 100 trials. Additionally, we tracked how well ABC-TTB described the generated data overall, measured by the mean proportion of correct predictions it would generate within that sample. Lastly, we calculated the ratio of the mean correct predictions (MCP) per proposal generated. Results are shown in Figure 6.

The leftmost plot in Figure 6 shows that the computational effort (measured by the number of generated proposal trees, plotted on a logarithmic scale) seems to increase more than exponentially (exponentially on a log-scale) when the proportion of samples and the quality of the to-be-accepted proposals increases. For the performance of the resulting models (shown in the middle), we can see that tuning the ϵ and ϕ-parameters can increase performance by up to 20%. Most importantly, there seems to be a diminishing returns property as shown in the rightmost plot, in the sense that one would have to spend more and more proposal samples for an ever smaller increase in prediction accuracy. Taking these results together, we see that there seems to be a clear trade-off between effort and performance as number of samples increases super-exponentially, but performance increases not as steeply. Using ABC-TTB, we can change the parameters directly and test this accuracy-effort trade-off explicitly thereby gaining deeper insights into when good predictions might only need few samples (Vul et al., 2014).

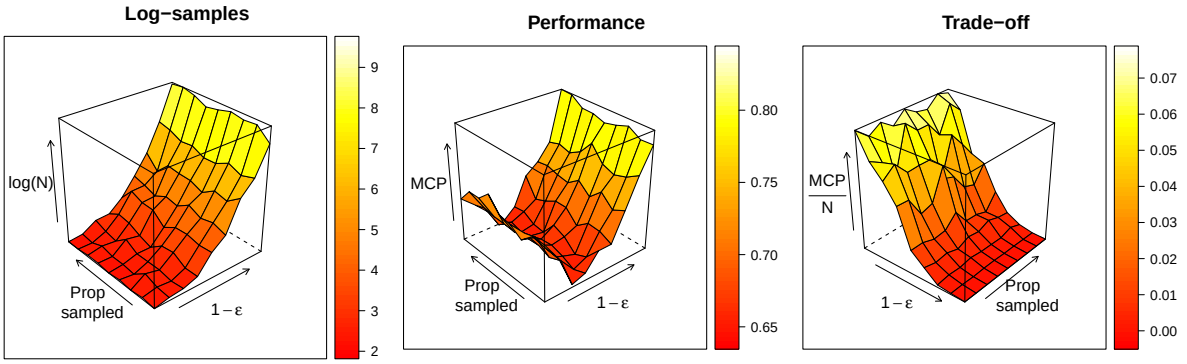


Figure 6: Assessed computational effort (log-number of proposal trees, N), performance (mean correct predictions, MCP), and trade-off (proportion correct predictions per proposal generated) for different ϵ-ϕ-combinations. Number of samples go up exponentially when bigger proportions are sampled and only high quality proposals are accepted. Performance also goes up, but there is a diminishing returns effect whereby increasingly less additional correct predictions are generated per generated proposal.

Conclusion

We have introduced Approximate Bayesian Computation as a method to let heuristic strategies emerge from smaller building blocks. ABC assesses the posterior of a model by sanity checking simulated proposals. We therefore think it could be a plausible approach to model how diverse strategies emerge over time. As a first proof of concept, we have shown how simple TTB-like trees can emerge from smaller building blocks (distributions over cues and their weights). Our new model, coined ABC-TTB, can recover TTB when data was generated by it, produces sensible inference and predictions in a real world data set, and allows to trade off computational effort and performance explicitly, consistently generating simple trees within the complex forest of all possible trees.

This is a first step towards unpacking the heuristic toolbox and future work will focus on extending our approach to the problem of letting more diverse building blocks or even different strategies emerge. Another venue for future research concerns the psychological plausibility of the proposed learning mechanism. We think that dynamic scenarios such as active (Parpart et al., 2015) or causal learning (Morais et al., 2014) are a plausible way to test sequential predictions of heuristic models, especially given that ABC-TTB generates decisions from the beginning of the learning process and provides measures of uncertainty for every prediction. Given the large number of cognitive models (Schulz, Speekenbrink & Shanks, 2014), we believe that probing how strategies emerge from compositional building blocks, and how simple structures are derived from approximative computation, is a promising endeavour.

Acknowledgements

ES is supported by the UK Centre for Doctoral Training in Financial Computing. He would like to thank the MPIB-ABC research group for hosting him multiple times. BM was supported by grant ME 3717/2-2. Code available at: github.com/ericschulz/TTBABC

References

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*, 2025–2035.

Bellman, R. E. (1961). *Adaptive control processes*. Princeton University Press, Princeton, NJ.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–418.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gilks, W. R. (2005). *Markov chain Monte Carlo*. Wiley Online Library.

Morais, A. S., Schooler, L. J., Olsson, H., & Meder, B. (2014). From causal models to sound heuristic inference. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*, (pp. 1036–1041). Cognitive Science Society.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision making: The success of success. *Journal of Behavioral Decision Making*, *17*, 117–137.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2015). Active learning as a means to distinguish among prominent decision strategies. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*, (pp. 1829–1834). Cognitive Science Society.

Poupart, P. (2010). Bayesian reinforcement learning. In *Encyclopedia of Machine Learning*, (pp. 90–93). Springer.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society*, (pp. 2122–2127). Cognitive Science Society.

Schulz, E., Speekenbrink, M., & Shanks, D. R. (2014). Predict choice: A comparison of 21 mathematical models. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*, (pp. 2889–2894). Cognitive Science Society.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*, 69–85.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Extraction of Event Roles From Visual Scenes is Rapid, Automatic, and Interacts with Higher-Level Visual Processing

Alon Hafri (ahafri@sas.upenn.edu)
Department of Psychology, University of Pennsylvania, 3720 Walnut Street
Philadelphia, PA 19104 USA

John C. Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, University of Pennsylvania, 3720 Walnut Street
Philadelphia, PA 19104 USA

Brent Strickland (stricklandbrent@gmail.com)
Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University
Institut Jean Nicod, (ENS, EHESS, CNRS)
75005 Paris, France

Abstract

A crucial component of event recognition is understanding the roles that people and objects take: did the boy hit the girl, or did the girl hit the boy? We often make these categorizations from visual input, but even when our attention is otherwise occupied, do we automatically analyze the world in terms of event structure? In two experiments, participants made speeded gender judgments for a continuous sequence of male-female interaction scenes. Even though gender was orthogonal to event roles (whether the Agent was male or Female, or vice-versa), a switching cost was observed when the target character's role reversed from trial to trial, regardless of whether the actors, events, or side of the target character differed. Crucially, this effect held even when nothing in the task required attention to the relationship between actors. Our results suggest that extraction of event structure in visual scenes is a rapid and automatic process.

Keywords: event roles; thematic roles; event perception; visual perception; switching costs

Introduction

A fundamental way we interpret the world is not just in terms of objects, but also events. Indeed, categorizing the causal relationships between people and objects is important for guiding our social behavior: Was it the boy that hit the girl, or the girl that hit the boy? The semantic relationships that exist between participants in events are called their roles (Fillmore, 1968; Gruber, 1965). For example, in *John broke the door*, John is the Agent (actor) and the door is the Patient (undergoer). The semantic properties of these roles and others (e.g., source and goal) appear to be consistent across a wide range of events, with some theorists arguing that the Agent and Patient roles subsume all others (Dowty, 1991). Relatedly, Strickland (2016) has argued that the Agent/Patient distinction is one form of "core knowledge," that is reflected across languages, pre-verbal infant cognition, and the adult visual system. While there is ongoing debate about the precise nature of the Agent/Patient dichotomy, it is clear that these representations are a crucial component of recognizing what is happening in the world.

Given the importance of event role information, then, it would be beneficial for the process of identifying roles to be automatic, continuously working in the background, since at any given moment we may be attending to other perceptual information, e.g., identifying objects or spatial properties of the scene. Recent evidence suggests that people are able to discriminate Agents from Patients from input lasting less than 75 ms, and that elements of body posture are important heuristics for these categorizations (Hafri, Papafragou, & Trueswell, 2013; see also Dobel, Gumnior, Bölte, & Zwitserlood, 2007; Wilson, Papafragou, Bunger, & Trueswell, 2011). However, it is not yet known whether this discrimination is only active when the task requires it or is a fully automatic process.

To investigate this issue, we employ a paradigm that has been used profitably in the past to investigate questions of automaticity: the switching cost paradigm (e.g., Pecher, Zeelenberg, Barsalou, 2003; Spence, Nicholls, & Driver, 2001). The logic of this paradigm is as follows: if a certain process (here role recognition) is automatically engaged, then when the role of an actor switches before participants judge an orthogonal actor property (here gender), it should result in a switching cost; i.e., a lag in reaction time.[1] If such a pattern is observed, it would provide strong evidence that analysis of event structure from the visual world is a rapid, automatic process that helps organize our understanding of the social and causal world, even when we are not explicitly focused on this information.

Experiment 1

In the first experiment, participants were asked to identify the

[1] Though previous work has referred to these reaction time effects as costs, here we cannot differentiate between switching costs vs. repetition benefits, because there is no meaningful baseline for comparison (i.e., in our stimuli, the roles always either repeated or they did not). In any case, whether the effects are a benefit or cost does not qualitatively change our conclusions.

side of the male or female actor in a sequence of scenes, and we looked for evidence of a switching cost when event roles changed from one scene to the next. To maximize our chances of finding such an effect if it exists, we included a secondary catch task on a subset of trials, in which participants were probed about what event they just observed (asking whether they just saw, e.g., *kicking* or *punching*). This secondary task made no mention of event roles, so participants were not required to extract role information to perform either task.

Methods

Participants

Twenty-four members of the University of Pennsylvania community participated and received either class credit or $10 for their participation.

Materials and apparatus

Forty color photographic images depicting 10 two-participant Agent-Patient events were used in the current experiment, taken from a previous study (Hafri et al., 2013) that investigated extraction of event categories and roles from briefly displayed and masked images (< 75 ms). These 10 showed the highest agreement for role assignment among participants when they were probed after brief displays. The events used were: *brushing, chasing, feeding, filming, kicking, looking, punching, pushing, scratching, tapping*.

Each event was depicted from a side view, and involved a male and a female actor. Six different actor pairs appeared in the 10 events, with each actor pair appearing in front of a different indoor or outdoor background. Each event was associated with only one of the actor pairs. For each event, there were four versions: the gender of the Agent was male or female, and the side of the Agent was left or right.[2] All events were normed for name agreement in the previous study. Each image was 640 × 480 pixels and subtended 19° × 15° at approximately 54 cm. Example images appear in Figure 1.

Stimuli were displayed on a 19" Dell 1908FP LCD monitor at a refresh rate of 60 Hz. Responses were collected using a PST E-Prime button box (mean latency 17.2 ms, *SD* 0.92 ms). The experiment was run in Matlab using the Psychophysics Toolbox (Brainard, 1997).

List design

Given that detecting switching costs depends on measuring the influence of one stimulus on another, we chose to implement a list design that controls for first-order carry-over effects (continuous carryover sequences; Nonyane & Theobald, 2007). These sequences are similar to randomized block and Latin square designs, with the added benefit of the following attributes: each item precedes and follows every other, including itself, exactly once; effects of the current and preceding stimuli are orthogonal to position effects in the list; and each item occurs exactly once every block, where a block represents a permutation of all items. Thus here, we use sequences of n = 40, resulting in 1601 (n^2 + 1) trials split among 40 blocks. Unique lists were generated for every participant.

Among these standard image trials, we randomly dispersed catch (Event Test) trials, in which participants were given a 2AFC test about what action just appeared in the previous trial (e.g., *tickling* vs. *scratching*). One label was correct, and the other was a foil randomly selected from the set of nine other possible actions. The purpose of these trials was to focus participants' attention on the event without explicitly testing them on event roles. There were 58 of these trials, with 1 to 3 per 40-trial block.



Figure 1: Example images (clockwise from top left) for *punching*, *brushing*, *feeding*, and *kicking*.

Procedure

Participants were told that they would view photographs of people performing actions in continuous sequences. They were instructed to press the button corresponding to the side of the screen that the male character was on as quickly and accurately as possible. They were also told that after a small set of the trials, they would be tested on what action just appeared in the previous trial. Task (male or female search) was between-subject.

Twelve practice trials preceded the main experiment (plus two catch trials), all involving the actor pairs performing joint or symmetrical actions (e.g., *writing, shaking hands, crying*). Image trials consisted of the following sequence: A "Ready?" screen for 350 ms, a central fixation crosshair for 250 ms, a blank screen for 150 ms, and the test image, at which point the participant searched for and pressed the side of the male (or female) actor as quickly as possible. Catch trials involved a similar sequence, but with text "What action did you just see?" and two event probes below (e.g., *tickling* or *scratching*). Image trials timed out if no response was given within 2000 ms, and catch trials within 3500 ms. Average

[2] We found no differences between male-Agent vs. female-Agent images in terms of the effects reported below.

duration of the experiment was 41 min and included a break every 40 image trials.

Results

Coding

Since we were investigating effects of one stimulus on the next, exclusion criteria were the following (decided in advance of analysis): response errors and those following error trials (8.8% in total), RTs faster than 200 ms (44 trials in total), timeouts (17 trials in total), trials after breaks (40 per participant), and trials after catch trials (58 per participant). An additional 63 trials were excluded due to errors in list creation. Finally, for the remaining data, trials were excluded when the RT was 2.5 standard deviations above or below each participant's mean (3.2% of trials in total), following accepted data trimming procedures. With the above criteria, a mean of 17% (*SD* 4.0%) of trials in total were excluded per subject. Average RT for the included data was 383 ms (*SD* 34 ms).

Analysis

Accuracy on catch trials was significantly above chance across subjects (mean = .85, *SD* = .10, $t(23) = 40.0$, $p < .001$, $d = 3.37$).

Individual trial reaction times from the primary task (i.e., judging gender side) were analyzed using linear mixed effects modeling with the maximal subject and item random effects structure that converged (Barr, Levy, Scheepers, & Tily, 2013). Models with and without the following factors were compared: repeated Actor Pair (repActors); repeated Side (repSide), i.e., whether the male and female actors were on the same side as in the previous trial; and repeated Role (repRole). This last factor reflects the effect of interest: how a change in the role of the person being asked to respond about affects RT (e.g., if the male remains the Agent, or switches to being the Patient). Significance of factors was tested by comparing likelihood-ratio values for models that included factors to models without them.[3]

Findings

First, and most importantly, an event role switching cost was observed. In particular, as shown in Table 1, participants were on average 6 ms slower when the role of the target character changed from one trial to the next.

This effect, though quite small, was significant: The best-fitting mixed effects model included main effects and interactions of repActors and repSide, and a main effect of repRole (the role switching cost), over a model that did not include repRole, $\chi^2(1) = 29.3$, $p < .001$. Models with additional interaction terms were not a significantly better fit, either for repActors × repRole ($\chi^2(1) = 2.54$, $p = .11$), or repSide × repRole ($\chi^2(1) = 0.57$, $p = .45$).[4]

Besides the effect of primary interest (role switching cost), post-hoc analyses re that participants were faster when the actor pair repeated, and on trials where the actor pair did not repeat, participants were *slower* when the target side repeated. Though speculative, these effects may be accounted for by two mechanisms: visual priming due to similarity of actor pair appearance (faster RTs for repeated actors), and an incorrect expectation that the response should always switch if there is significant visual change (slower RTs for different side and actor pair). See Table 1 for a summary of all effects.

Table 1: Mean RTs by condition for Experiment 1, across Subjects. Standard errors in parentheses.

	Reaction time (ms)		
Condition	Repeated	Different	Switching cost
Role	380 (6.86)	386 (6.86)	6 (0.97)**
Actors	371 (6.18)	385 (7.19)	14 (1.78)**
Side	390 (7.73)	377 (6.60)	-13 (3.03)*
Side, Repeated Actors	371 (6.22)	371 (6.52)	0 (3.09)
Side, Different Actors	393 (7.99)	378 (6.72)	-15 (3.17)**

** = Significant effect ($p < .05$) in F_1 and F_2 ANOVAs, and multilevel modeling.
* = Significant effect ($p < .05$) in either F_1 or F_2 ANOVAs, and multilevel modeling. See text for detailed statistics.

Does Agent saliency drive the effect? The switching cost could conceivably be driven by the cost of switching from Agent to Patient or vice-versa. Indeed, an Agent primacy or saliency effect has been observed both in the linguistics and vision literature (Cohn & Paczynski, 2013; Dobel et al., 2007; Dryer, 2013; Wilson et al., 2011). If so, we should observe two additional effects in our data: (1) faster RTs on Agent trials as compared to Patient trials; and (2) an asymmetry between an Agent→Patient switch and Patient→Agent switch. However, neither of these was borne out in model comparisons (all p's > .38), and parameter estimates, though not significant, indicated *greater* RTs for Agent judgments. These analyses suggest that Agent saliency cannot account for the role switching cost effect.

Summary of results. As predicted, there was a reliable role switching cost, i.e. slower RTs when switching from a judgment on the Agent on one trial to the Patient on the next, or vice-versa. Importantly, the switching cost did not interact with other factors, such as side or actor pair, and did not appear to be driven by Agent saliency.

Magnitude of the role switching cost. Although the magnitude of this effect was small (about 6 ms), it is

[3] RepActors (repeated actor pair) always entailed a repeated event since each event was depicted by only one actor pair, so all analyses reported include only repActors. Additionally, reaction times were transformed into inverse RTs by using -1000/RT as the response variable. Mean raw RTs are displayed in Tables 1 and 2 to illustrate basic effects. All models included nuisance regressors for trial number and preceding trial inverse RT to account for general temporal dependencies.

[4] Though the nature of the design results in unbalanced numbers of observations in each condition, very similar findings emerged in separate Subject (F_1) and Item (F_2) ANOVAs on mean inverse RTs for all analyses. This is true for both Exp 1 and Exp 2.

comparable to previously observed switching costs, relative to mean RTs for task (e.g., Pecher et al., 2003, obtained a cost of 29 ms relative to mean RTs of 1139 ms, compared with our 6 ms vs. 383 ms mean RTs). And while it may be surprising that such a small effect would be statistically significant, it is important to keep in mind that unlike a typical cognitive experiment, each observer provided on average 1329 data points (more than many cognitive studies), resulting in very stable performance estimates per subject and per item (e.g., note the low Standard Errors across Subjects in Table 1). Furthermore, as an indication of its robustness, 21/24 participants and 10/10 items showed a numerical difference in line with the role switching cost.

Experiment 2

Our results thus far provide compelling evidence that when observing scenes, people extract event structure, even when it is not a specific component of the task. However, it may still be the case that our secondary catch (Event Test) task, despite not being explicitly about role identification, inadvertently focused participants' attention on the event structure itself. That is, in order to perform the task of event category extraction, they may have defaulted to a strategy of analyzing event roles. Experiment 2 addresses this issue. We conducted the same experiment on a new set of participants, but removed the catch trials, and made absolutely no mention of events, actions, or roles in our instructions. If this effect is really a fact about the visual perception of scenes, then we expect to observe it even under these conditions.

Methods

Participants

An additional 24 members from the University of Pennsylvania community participated and received class credit for their participation.

Materials and procedure

All materials, apparatus, and procedure were identical to Experiment 1, except for the following changes: First, no catch (Event Test) trials were included. Second, instructions were modified to omit mention of the catch trial task, and importantly, the beginning of the instructions were rewritten to omit mention of actions ("You will be shown photographs of people in different scenes…"). Task (male or female search) was again between-subject. Average duration of the experiment was 38 min.

Results

Coding and Analysis

Data coding procedures were the same as in Experiment 1. We excluded trials with response errors and those following error trials (8.0% in total), RTs faster than 200 ms (62 trials in total), and timeouts (7 trials in total). Trials after breaks (40 per participant) and trials from the list creation error (216 in total) were also excluded. Outliers 2.5 standard deviations above or below each participant's mean were removed (mean 3.1%). With these criteria, a mean of 13% (*SD* 4.9%) of trials per subject were excluded. Average RT for the included data was 387 ms (*SD* 48 ms). Individual trial reaction times from the primary task (i.e., judging gender side) were analyzed using linear mixed effects modeling, as in Exp 1.

Findings

As in Experiment 1, a role switching cost was observed. In Table 2, we see that participants were on average 3 ms slower when the role of the target character changed from one trial to the next. Furthermore, the role switching cost interacted with repeated actor pairs, such that the role switching cost when the actor pair repeated was marginally greater than when the actor pair did not, $t_1(23) = 1.77$, $p = .09$, $d = .36$.

These effects, although small, were significant: The best-fitting mixed effects model included main effects and interactions of repActors and repSide, and a main effect of repRole and interaction of repRole × repActors. The fit of the model was significantly better than the same model without the additional interaction of repRole × repActors, $\chi^2(1) = 4.89$, $p = .03$; and significantly better than a model that did not include repRole at all, $\chi^2(2) = 15.5$, $p < .001$. Additionally, a model that also included an interaction of repRole and repSide was not a significantly better fit, $\chi^2(1) = .004$, $p = .95$.

As in Exp 1, post-hoc analyses indicated that participants were faster when the actor pair repeated, and on trials where the actor pair did not repeat, participants were *slower* when the Side repeated. See Table 2 for details.

Table 2: Mean RTs by condition for Experiment 2, across Subjects. Standard errors in parentheses.

	Reaction time (ms)		
Condition	Repeated	Different	Switching cost
Role	385 (9.62)	388 (9.79)	3 (0.77)*
Actors	371 (8.11)	390 (10.0)	19 (2.68)**
Side	394 (9.43)	380 (10.3)	-14 (3.65)*
Side, Repeated Actors	368 (6.75)	374 (9.74)	6 (5.62)
Side, Different Actors	398 (9.93)	382 (10.4)	-16 (3.51)**
Role, Repeated Actors	368 (8.12)	374 (8.29)	6 (2.73)*
Role, Different Actors	388 (9.91)	391 (10.2)	3 (0.89)*

** = Significant effect ($p < .05$) in F_1 and F_2 ANOVAs, and multilevel modeling.

* = Significant effect ($p < .05$) in either F_1 or F_2 ANOVAs, and multilevel modeling. See text for detailed statistics.

Effect of Agent saliency. In contrast to Experiment 1, there was evidence for an asymmetry in switching cost for Agent→Patient trials vs. Patient→Agent trials in model comparisons, $\chi^2(1) = 3.96$, $p = .05$. However, this was in the *opposite* direction of that predicted by Agent saliency: Switching from judging the Agent to the Patient was slightly *faster* than vice-versa. We also found some evidence for a difference between Agent and Patient judgments: Specifically, reaction times were faster for Agent judgments than for Patient judgments, but only when the actor pair repeated (368 ms vs. 372 ms; $\chi^2(2) = 6.66$, $p = .04$). Crucially,

whether people were making an Agent or Patient judgment did not interact with our effect of interest, the role switching cost, $\chi^2(1) = 2.04$, $p = .15$.

Summary of results. As predicted, we again observed a reliable role switching cost, i.e. slower RTs when the role of the target character switched, even when participants were not probed about the event. The effect again did not appear to be driven by Agent saliency.

Comparison of Experiments 1 and 2. From Tables 1 and 2, it appears that the magnitude of the role switching cost in Exp 1 is greater than in Exp 2. Formal comparison of the role switching cost across experiments revealed that the difference is significant, $t_1(45.3) = 2.14$, $p = .038$, $t_2(9) = 3.35$, $p = .009$. Indeed, more participants and items showed the numerical difference in Exp 1 than in Exp 2 (21/24 vs. 17/24 participants, and 10/10 vs. 7/10 items). Nevertheless, items drove role switching cost consistently across experiment: the effect for individual stimuli was correlated across experiment, $r = 0.37$, $t(38) = 2.43$, $p = .02$. All of these findings further attest to the stability of the measures of central tendency (i.e., subject and item means) – likely due to the large number of observations per cell.

Discussion

The two experiments presented here demonstrate that the structure of an event, i.e. who did what to whom, is continuously involved in visual processing, even when attention is directed toward other features (here, gender) that do not require extracting any information about event roles. In particular, we observed a small but reliable switching cost associated with verifying the side of the male (or female) actor in photographic images when the event roles switch from one image to the next in sequence. This effect held even when nothing in the instructions or task made reference to events or human interactions, or required making judgments about the event (Exp 2). Indeed, in debriefing, no subject explicitly guessed our hypothesis that there could be a switching cost if the person being judged had a different role than in the previous trial.

The switching cost effect appears to be robust across a range of factors: it held (1) across events, (2) across actors, and (3) across the side of the male and female characters. Furthermore, we also showed that this effect is not driven by the saliency or priority of Agents: there was no interaction between the role switching cost effect and whether participants were making Agent or Patient judgments, and there was no difference when switching from Agent to Patient judgment, compared to vice-versa (if anything, switching Agent to Patient was slightly *faster* in Exp 2).

We do not believe that the effect is due to the degree of mismatch between Agent- and Patient-like body poses from trial to trial, which could in principle drive the effect. Indeed, the fact that the role switching cost is largely invariant to the particular side (and orientation) of the target actor argues against a location-specific body-pose-matching explanation. Nevertheless, we cannot rule out location-general pose-matching. Certainly body pose is a strong and reliable heuristic to event role (Hafri et al., 2013), so higher-level body pose recognition may be the first route for identifying event roles in initial processing. We are not opposed to such an explanation, and indeed there seems to be a correspondence between Dowty's (1991) proto-role entailments and visually salient aspects of body posture, including orientation of the head and body (Dowty's "volitional involvement"), and outstretched extremities (Dowty's "ability to causally effect change").

Given that this effect is not explainable by a simple location-specific pose-matching hypothesis, what is its origin? One possibility is the following: In the gender identification task, participants search for the target character to plan their response, attending to the male (or female) character as necessary (e.g., in *girl-kicks-boy*, attention is placed on the boy). If the role of the target character changes (e.g., the next image is *boy-taps-girl*), then there is a mismatch between the previous role of the male actor (Patient) and the current role (Agent). This explanation assumes that when people identify a particular person-based attribute (e.g., gender), others "come along for the ride," and that mismatches in these attributes from trial to trial manifest in increased reaction time at the decision making stage. Our assertion here is that event role identification is automatic, and is thus consistent with the claim that Agent/Patient event role information is part of human "core knowledge" (Strickland, 2016).

Although it was not a primary focus of these studies, might our results contribute to the ongoing debate about the degree to which roles are event-general or event-specific? In particular, some have argued for an innate and limited set of semantic relations (e.g., Pinker, 1989), while others have argued that event-general properties of roles develop from observation of commonalities among event-specific roles (e.g., Tomasello, 2000). In Exp 2, we did find evidence that the role switching cost was greater *within* event (repeated actors), but we note that in both experiments, the effect still held *between* events. Thus whatever its ontogenetic origins, Agent and Patient roles are at least partly event-general.

One interesting question that arises is the extent to which this is a general property of event scene analysis, or is specific to human interactions. That is, in event scenes that involve interactions with or among inanimate objects (e.g., *A woman opens a door* or *A ball hits a rock*), are roles assigned using similar processes? If we view language as a reflection of the semantic structures available to the human mind, it is of course possible for objects to fill the same argument slots as humans do in an utterance. Yet even then, there are expectations of what can fill which roles: animacy of sentential constituents constrains on-line sentence interpretation (contrast *The defendant examined by the lawyer…* with *The evidence examined…*; Saffran, Schwartz, & Linebarger, 1998; Trueswell, Tanenhaus, & Garnsey, 1994). It may be that on the path from vision to complex, structured thought, animacy constrains role assignment in a

similar fashion. If we apply the evidence from the sentence processing literature to the visual realm, then the prediction is that more extreme switching costs would emerge when switching to an inanimate Agent/animate Patient event (I/A) from a two-animate-entity interaction (A/A). An alternate possibility, and one predicted by the animacy hierarchy within neural systems (Connolly et al., 2012; Scholl & Gao, 2013), is that switching costs may be observed regardless of animacy, but only between like-animacy pairs (so for A/A→A/A and I/I→I/I, but not for A/I→I/A or I/A→A/I).

A most exciting possibility is that our paradigm may be used to test the relationship among hypothesized event roles, or the degree to which role representations are distinct across different kinds of events, such as those involving implicit causality (e.g., *frighten*).

Conclusions

To summarize, a significant switching cost for event roles was observed and replicated across experiments. To put this effect in context, we believe that a separation of its practical and theoretical consequences is warranted. In terms of its absolute magnitude (on the order of 5 milliseconds), the role switching cost probably has few *practical* consequences for scene perception. But its magnitude is not relevant for our theoretical point: that an automatic mechanism for perceiving event structure must be operating, with a pervasive enough influence on visual perception and decision making that its presence is detectable in reaction times of participants engaged in orthogonal tasks. What our results suggest is that the human visual system is continuously engaged in extracting meaningful "high level" representations of the world, not just for the objects or space around us, but also for the causal relationships that are taking place between people.

Acknowledgements

Thanks to the actors, to Russell Epstein for lab space, and to Estee Ellis and Stamati Liapis for assistance in data collection. This research was supported by ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC, UPenn graduate research funds, and NSF Graduate Research Fellowship and NSF IGERT Traineeship (to A.H.).

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: the processing of semantic roles in visual narrative. *Cognitive Psychology*, *67*(3), 73–97.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … & Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience*, *32*(8), 2608–18.

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica*, *125*(2), 129–43.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619.

Dryer, M.S. (2013). Order of Subject, Object and Verb. In M.S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available at http://wals.info/chapter/81, Accessed 2016-01-25.)

Fillmore, C. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory*. Holt, Rinehart, & Winston.

Gruber, J.S. (1965) Studies in lexical relations. PhD Dissertation, MIT, Mass.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology. General*, *142*(3), 880–905.

Nonyane, B. A. S., & Theobald, C. M. (2007). Design sequences for sensory studies: achieving balance for carry-over and position effects. *British Journal of Mathematical and Statistical Psychology*, *60*(Pt 2), 339–49.

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different modality properties for concepts produces swtiching costs. *Psychological Science*, *14*(2), 119–124.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure.* Cambridge, MA: MIT Press.

Saffran, E. M., Schwartz, M. F., & Linebarger, M. C. (1998). Semantic influences on thematic role assignment: Evidence from normals and aphasics. *Brain and Language*, 62(2), 255–97.

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment? In M.D. Rutherford & V.A. Kuhlmeier (Eds.), *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*. MIT Press.

Spence, C., Nicholls, M. E., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, 63(2), 330–336.

Strickland, B. (2016). Language reflects "core" cognition: A new hypothesis about the origins of cross-linguistic regularities. *Cognitive Science*, 1-32.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3), 209–253.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*.

Wilson, F., Papafragou, A., Bunger, A., & Trueswell, J. (2011). Rapid Extraction of Event Participants in Caused Motion Events. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

A 3D shape inference model matches human visual object similarity judgments better than deep convolutional neural networks

Goker Erdogan (gerdogan@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY USA

Abstract

In the past few years, deep convolutional neural networks (CNNs) trained on large image data sets have shown impressive visual object recognition performances. Consequently, these models have attracted the attention of the cognitive science community. Recent studies comparing CNNs with neural data from cortical area IT suggest that CNNs may—in addition to providing good engineering solutions—provide good models of biological visual systems. Here, we report evidence that CNNs are, in fact, not good models of human visual perception. We show that a 3D shape inference model explains human performance on an object shape similarity task better than CNNs. We argue that deep neural networks trained on large amounts of image data to maximize object recognition performance do not provide adequate models of human vision.

Keywords: shape perception; object recognition; neural networks; 3D shape; deep learning;

Introduction

Despite decades of research, we know little about the neural representations underlying visual perception (Peissig & Tarr, 2007; Kourtzi & Connor, 2011). This is especially true of high-level representations involved in visual object identification and recognition. Although we understand little about how our brains accomplish visual perception, we are able to build engineering solutions that approach, and in some cases match, human performance on some visual tasks. Recently, multi-layered artificial neural networks known as convolutional neural networks (CNNs) have shown impressive object recognition performances when trained on large image data sets. Importantly for the cognitive science community, there seems to be evidence suggesting that these computer vision models may also be good models of biological visual systems (Kriegeskorte, 2015). Several studies have shown that CNNs provide good accounts of neural data from both monkey and human inferotemporal (IT) cortex, explaining almost all of the variance in some cases (Baldassi et al., 2013; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Here, we present evidence suggesting that CNNs are, in fact, not good models of human visual perception. We show that CNNs fail to capture people's responses on an object shape similarity task. Moreover, we show that a 3D shape inference model outperforms CNNs, suggesting that 3D structure is an important feature of people's visual object representations that CNNs fail to capture.

CNNs implement a sequence of convolution and subsampling operations to extract useful visual representations when trained in a supervised manner on large image data sets. Due to their huge impact on computer vision research, these models have now started to attract attention in cognitive science and neuroscience where they are actively investigated as models of biological visual perception.

A recent study by Khaligh-Razavi and Kriegeskorte (2014) compared a large set of models from computer vision and neuroscience to human fMRI and monkey neural data from IT. Similarity matrices calculated from each model were correlated with similarity matrices from human and monkey neural data. They found that AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a deep CNN trained on 1.2 million images, had the highest correlation with IT data. An ensemble model combining the outputs of each layer of AlexNet with scores from multiple categorization models trained on the features learned by AlexNet was able to capture the entire variance in IT data. The scores from animate/inanimate, face/nonface, and body/nonbody categorization models were needed to emphasize the differences between these categories, since it seems that IT gives more weight to these categorical distinctions than AlexNet did. These authors also showed that these results are not purely driven by the category structure in IT. AlexNet on its own did, in fact, capture some of the within-category structure. It is remarkable that a model trained to maximize object recognition accuracy is able to provide a good model of biological visual systems. This raises the interesting possibility that biological visual systems might be optimized primarily for object recognition. If so, a high-performing model of visual object categorization may also be a good model of biological visual systems.

Yamins et al. (2014) recently offered evidence for this claim. They showed that models that are better at categorization explain neural responses better. Instead of using a fixed set of models, they defined a model space using parameters that control various features of CNNs such as number of layers, filter sizes, and activation thresholds. Examining a large number of models in this space, their results showed that categorization performance was highly correlated with IT response predictivity. However, they also showed that an "ideal" categorization model was not highly correlated with IT responses. This suggests that solely aiming for good categorization does not, by itself, result in models predictive of IT responses. The authors claimed that it is the combination of a hierarchical architecture and high categorization performance that accounts for why CNNs provide good models of IT responses.

In spite of these impressive results, there is reason to question whether CNNs provide good models of human visual systems. CNNs are trained only to maximize object recognition performance. However, human visual systems solve not only object recognition but a myriad of visual tasks from segmentation to extraction of 3D shape. Indeed, because 3D shape and semantic category labels are highly correlated, it is unclear whether IT representations are best thought of as shape-based or semantic (Kourtzi & Connor, 2011). Even though CNNs often account for more variance in IT responses than other models, it is possible that the driving factor behind these results is shape similarity rather than semantic features. Baldassi et al. (2013) provided evidence that shape similarity, rather than semantic information, accounts for the structure of IT representations. They demonstrated that most of the semantic category structure in IT is explained by visual shape similarities within semantic categories. This result raises the question of whether the representations learned by CNNs—which receive supervised training based solely on semantic category labels—adequately characterize representations used by human visual systems. A striking demonstration suggesting these representations are, in fact, inadequate was provided by Szegedy et al. (2013). They showed that it is possible to create pairs of images that are indistinguishable to the human eye, but nonetheless are classified by CNNs into different classes. For example, it is possible to imperceptibly perturb an image that a CNN classifies as a bus such that the CNN classifies the perturbed image as an ostrich. This finding suggests that CNNs might be solving the problem of object recognition in a way that is rather different from that of human visual systems.

Here, we report behavioral evidence from an object shape similarity task suggesting that CNNs are not good models of human vision. Moreover, we show that a 3D shape inference model provides a better account for human behavior. We argue that models trained solely to maximize object recognition performance cannot capture the nature of human visual representations. A crucial feature of these representations not captured by these models is the 3D structure of objects.

Experiment

We created a set of 10 base objects using a "shape grammar" (Figures 1 and 2) where each object consisted of multiple rectangular blocks (referred to as "parts" and denoted by P in the grammar). A base object was generated as follows. To start, a root part was assigned 0-3 neighboring parts, also referred to as child parts, using the production rules of the shape grammar. A child part connected to the root part at one of its six faces. This face was chosen at random. Similarly, the width, height, and depth of a child part were randomly chosen from the range $[0,1]$. A child part could also be assigned neighboring (or grandchild) parts using the same production rules and random selections. Note that, in this framework, an object can be characterized using a "parse tree" due to our use of a shape grammar. We constrained the parse trees for our base objects to have a depth of four, which produces objects with three levels of parts (see Figures 2a and 2b for an example base object and its parse tree).

$$P \quad \rightarrow \quad P \mid PP \mid PPP \mid \varepsilon$$

Figure 1: Production rules of the shape grammar used in generating the experimental stimuli and representing shape in our 3D shape inference model. P is the only non-terminal symbol, and ε is the Null symbol.

Each base object was then used to create 8 additional objects, called variations, by applying 1 of 4 possible manipulations, referred to as *change part size*, *add part*, *remove part*, and *change connecting face of part*. Each of these four manipulations was applied at two different levels (second and third levels) of the parse trees (see Figure 2 for examples of each manipulation). When using the *change part size* manipulation, we picked one of the parts at the desired level in the parse tree and resampled its size. When using the *add part* manipulation, a new part was added to the desired level, picking its size and connecting face (i.e., the face of its parent to which it is connected) randomly. For the *remove part* manipulation, we again randomly picked one part at the desired level and removed it and all of its children parts. Lastly, for the *change connecting face* manipulation, we randomly picked a part and chose a new connecting face for it from the empty faces of its parent. This manipulation moved the part and all of its children.

Experimental stimuli consisted of images of the 10 base objects and 80 variations (90 images in total).[1] We used Blender (`http://www.blender.org`), a 3D computer graphics and animation software package, to render each object from a random viewpoint by rotating the camera around the vertical axis keeping its distance to the origin fixed. Consequently, there was significant pose variation in our experimental stimuli.

The goal of the experiment was to collect people's object shape similarity judgments. On each trial, a subject was presented with one target and two comparison objects, and was asked to pick the comparison object that he or she thought was more similar in shape to the target object. The target object was always one of the ten base objects, and the two comparisons were two randomly picked variations of the target object. (For instance, a trial may show Figure 2a as the target object, and Figures 2c and 2d as the comparison objects.) On "catch" trials, one of the comparison objects was the same as the target object. Each subject participated in 100 trials, 16 of which were catch trials. The experiment was performed on the world wide web by 41 subjects via Amazon Mechanical Turk. Five subjects were discarded because they failed to reach 85% correct performance on catch trials.

[1] The entire set of experimental stimuli can be seen online at `http://gokererdogan.github.io/CogSci16SupplementaryMaterials/`

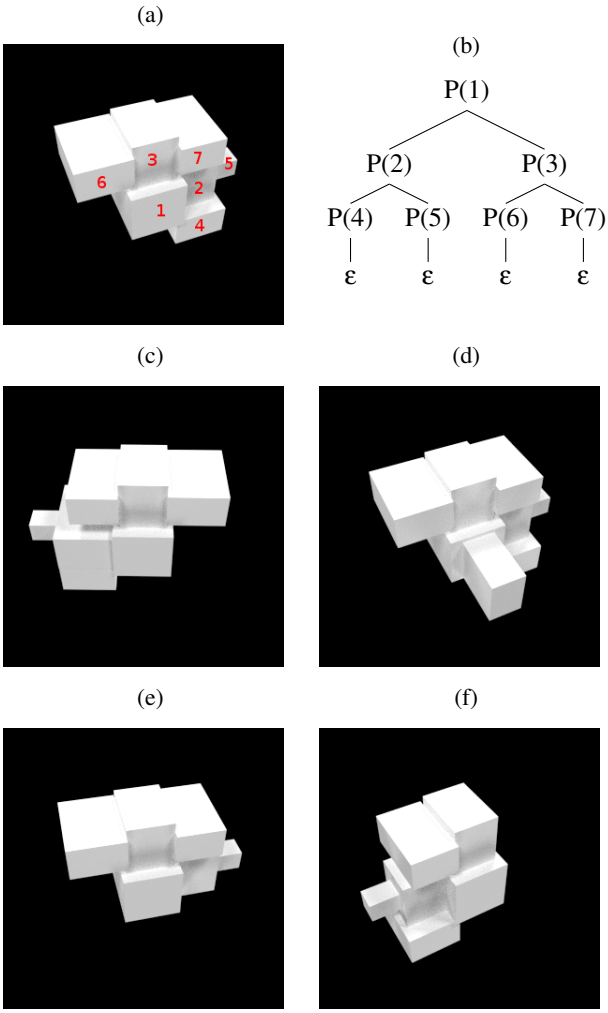


Figure 2: (a) An example base object. The numbers on parts refer to the part numbers in its parse tree. (b) Parse tree representing the object in (a). (c)-(f) Examples of *change part size* (c), *add part* (d), *change connecting face of part* (e), *remove part* (f) manipulations, respectively. The parts affected by each manipulation are Part 2 in (c), Part 4 in (e), and Part 6 in (f).

Computational Models

We compare five models on how well they account for our experimental data.

Pixel-Based model: The first one, called the Pixel-Based model, works directly on image pixel values. The dissimilarity between two objects is calculated as the Euclidean distance between their images in pixel space. The predictions of the Pixel-Based model are determined by calculating the distances between each comparison object and the target, and choosing the comparison that is closest to the target.

CNN models: Our main aim is to compare the performances of deep CNNs and a 3D shape inference model. For this purpose, we use two CNNs.[2] The first one is the eight-layer (five convolutional, three fully connected layers) CNN by Krizhevsky et al. (2012), referred to as AlexNet, trained on 1.2 million images in the ImageNet dataset. AlexNet achieved the best performance on the 2012 ImageNet Large Scale Visual Recognition Challenge. We treat each of its layers as a separate mini-model. There are, in total, 14 layers (making the three max-pooling and two normalization layers explicit). Using the standard terminology in the deep neural network literature, these layers are: *conv1, pool1, norm1, conv2, pool2, norm2, conv3, conv4, conv5, pool5, fc6, fc7, fc8*, and *prob*. The last layer, *prob*, is a 1000-dimensional vector encoding the probability of belonging to each of 1000 object categories in ImageNet. The second deep CNN that we test is by Szegedy et al. (2014), named GoogLeNet, which set the state-of-the-art performance on the 2014 ImageNet Large Scale Visual Recognition Challenge. GoogLeNet has 22 layers (with an additional five pooling layers). Our simulations used 16 layers: *pool1, conv2, inception3a-b, pool3, inception4a-e, pool5, inception5a-b, pool5, loss3* and *prob*. To make predictions from AlexNet and GoogLeNet, we input each image to a CNN and perform a bottom-up pass to calculate each layer's responses. The dissimilarity between two objects is computed as the Euclidean distance between these responses. When presented with a trial from our experiment, a mini-model chooses the comparison object that is closest in its response space to the target object.

Our 3D shape inference model: We developed a shape perception model that aims to infer 3D shape from 2D input images. Similar to our previously published 3D shape inference models (Yildirim & Jacobs, 2013; Erdogan, Yildirim, & Jacobs, 2015), this model combines a representational language characterizing 3D shape with forward models mapping from shape representations to 2D images. Using Bayesian inference, we invert this forward 3D-to-2D mapping and extract 3D shape from 2D images. Formally, a shape representation H consists of a string T from our shape grammar (Figure 1) and a spatial model S that associates a size vector ($s \in \mathbb{R}^3$) and a connecting face ($f \in \{1,2,3,4,5,6\}$) with each P node in T. The probability of H is

$$p(H) = p(S|T)p(T) \tag{1}$$

where $p(T)$ is the probability of producing parse tree T from the shape grammar. We assume production probabilities to be uniform[3] which gives the following expression for $p(T)$

$$p(T) = \frac{1}{4^{|\mathcal{P}|}}. \tag{2}$$

The probability for spatial model S consists of the probabilities of picking part sizes and connecting faces. Since we assumed part sizes to be uniform over the interval [0, 1], we only need to focus on the probabilities for connecting faces.

[2] We use the pretrained models provided by the Caffe framework (Jia et al., 2014).

[3] Production probabilities can also be integrated out, which leads to a slightly different prior distribution. Note that our results here are significantly robust to choice of prior distribution.

For a part with k available faces and c children, there are $\binom{k}{c}$ possible combinations of face assignments to its children. Since we have six empty faces for the root P node and five for the remaining P nodes (because one face is occupied by the parent), the probability of spatial model S is

$$p(S|T) = \frac{1}{\binom{6}{|O_{\text{root}}|}\prod_{n\in\{\mathcal{P}\setminus\text{root}\}}\binom{5}{(|O_n|-1)}} \tag{3}$$

where O_i refers to the set of occupied faces of node i. To map these shape representations to 2D images, we use a forward model that takes in the 3D representation and renders it as a 2D image. Because the shape representation H does not specify the viewpoint, forward model F takes in viewpoint θ along with the shape representation H and produces a 2D image I (i.e., $F:\{H,\theta\}\rightarrow I$). We used the Visualization Toolkit (VTK; `http://www.vtk.org`), a software package for 3D computer graphics, image processing, and visualization, to implement the forward model. To define the likelihood function $\mathcal{L}(H,\theta;I)$, we assume Gaussian noise on I:

$$\mathcal{L}(H,\theta;I) = p(I|H,\theta) \propto \frac{1}{\sigma^2}||I-F(H,\theta)||_F^2. \tag{4}$$

Here σ^2 denotes the variance of the noise (this is the only free parameter of the model—it was set to a value that achieves acceptance rates around 20%) on I, and $||\cdot||_F$ is the Frobenius norm. Combining the prior on shape representations and the likelihood function, we use Bayes' rule to infer likely 3D shape representations given a 2D image:

$$p(H,\theta|I) \propto p(I|H,\theta)p(H)p(\theta). \tag{5}$$

We assume $p(\theta)$ is a uniform distribution. Object similarity is computed by calculating how likely the model is to observe the image for one object given the image of the other. Denoting the images by I_1 and I_2, we calculate three similarity measures: $p(I_2|I_1)$, $p(I_1|I_2)$, and their average. We calculate $p(I_2|I_1)$ as follows (and similarly for $p(I_1|I_2)$):

$$p(I_2|I_1) = \int p(I_2|H,\theta)p(H|I_1)p(\theta)dHd\theta. \tag{6}$$

The expression inside the integral in Eqn. 6 is equivalent to inferring the 3D shape representation for I_1, picking a random viewpoint, and calculating the sum of squared error between the observed I_2 and the rendered image on the basis of inferred H and chosen viewpoint.

To sample from the posterior distribution $p(H,\theta|I)$[4] we use an MCMC procedure. We devised multiple proposal strategies to move in the hypothesis space, and used a Metropolis-Hastings (MH) algorithm to sample from the posterior.[5] These proposal strategies are: *add/remove part*, *change part size*, *change connecting face of part*, and *change viewpoint*. The *add/remove part* either adds a new P node to a random location in the tree, or removes randomly one of the P nodes with no child parts. Note that this move jumps between spaces of different dimensions; hence, we need to use a reversible-jump MCMC method. For the *change part size* move, we resample the size of a randomly picked P node. Similarly, the *change connecting face of part* move picks one P node randomly and assigns it a new random connecting face from the available faces of its parent P node. Finally, the *change viewpoint* move rotates the viewpoint around the vertical axis a random amount, which is drawn from a Gaussian distribution. Due to space limitations, we cannot go into the implementation details here.[6]

In our simulations, we ran one chain for each image used in the experiment. To speed convergence, we constrained the depth of parse trees to be at most six. Each chain was run for 200,000 iterations, and sample collection started after the first 50,000 iterations. Hence, we had 15 samples per image (see Figure 3 for two typical samples [i.e., two illustrations of the model's inferred 3D shape given an image]). To calculate the similarity between two images, we used Eqn.6, approximating the integral by a sum over samples from the posterior $p(H|I)$.

Ideal 3D observer model: As our last model, we use an "ideal" 3D observer that can perfectly extract the true 3D shape of an object from its image. Although not realistic, this model provides a useful benchmark because it defines optimal performance for our model. Two objects are compared by an alignment mechanism that rotates one object and finds the viewpoint that matches the image of the other object best (as in Eqn. 6). If we assume that shape matching is done on the basis of only the MAP sample, this ideal observer model sets the performance upper bound for our 3D shape inference model.

Results and Discussion

We calculated the predictions of each model as described in the previous section. For the experimental data, we gathered the data from all subjects and, for each trial, chose the majority response. We measured the accuracy of each model by calculating the percentage of correctly predicted trials (where a trial is correctly predicted if a model's response matches the subjects' majority response).

The results are shown in Figure 4. The Pixel-Based model has the lowest accuracy with 58%. CNNs achieve accuracies of 62% (AlexNet, using responses of the output layer) and 64% (GoogLeNet, using responses of the layer *inception5a*). Our 3D shape inference model achieves 72% accuracy using the similarities calculated by averaging $p(\text{Comparison}|\text{Target})$ and $p(\text{Target}|\text{Comparison})$. This performance is significantly better than both AlexNet's (binomial test, $p < 0.001$) and GoogLeNet's performance ($p =$

[4]To calculate Eqn. 6, we need samples from $p(H|I)$. However, $p(H|I) \approx p(H,\theta_{\text{MAP}})$ because there is only a single viewpoint from which an object H looks close to its image I. The results reported here do not change if we integrate out θ instead of using the MAP sample.

[5]The code for our shape inference model is available at `https://github.com/gokererdogan/Infer3DShape`

[6]Readers interested in these details can contact the first author.

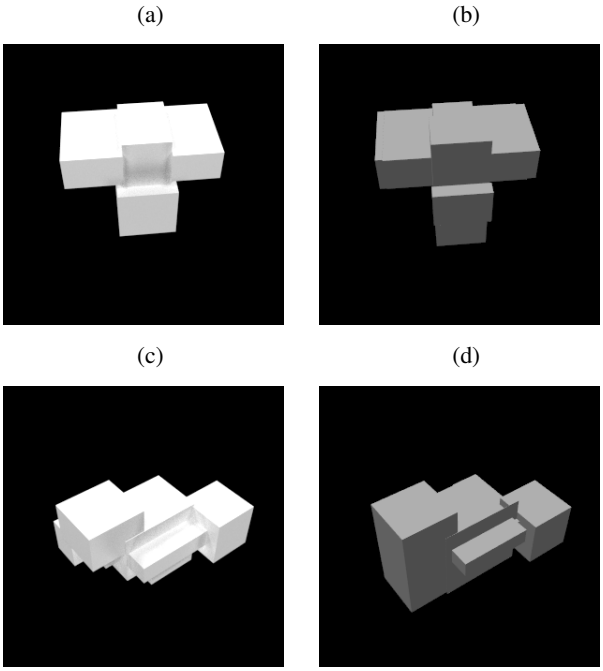


Figure 3: Sample runs of our 3D shape inference model. (a) An example input image. (b) One sample from our model for the image in (a). (c)-(d) Another example input image and a sample from our model.

0.004). The 3D ideal observer reaches an accuracy of 76%. However, the performances of the 3D ideal observer model and of our model are not significantly different ($p = 0.21$).

Because subjects did not show a strong preference for either of the comparisons in some trials, we also measured performance on only "high confidence" trials in which at least 80% of the subjects picked the same response. There were in total 120 (out of 280) high-confidence trials. Performances were as follows (due to space constraints, we omit the graph of these results). The Pixel-Based model's accuracy is 62%. Using the outputs of layer *prob*, AlexNet performs at 73%. GoogLeNet achieves an accuracy of 68% with the outputs of layer *inception5b*. Our 3D shape inference model performs the best, matching the accuracy of the 3D ideal observer model with 87% accuracy using the average similarity measure. This performance is significantly better than both GoogLeNet's performance (binomial test, $p < 0.001$) and AlexNet's performance ($p < 0.001$).

Our comparison here might seem unfair because our model knows that stimuli are built out of blocks while we used pre-trained CNNs that have never seen similar objects. However, subjects in our experiment have also never seen objects like our stimuli. In addition, previous studies presenting CNNs as good models of our visual systems used pre-trained networks. However, in order to alleviate further concerns, we have fitted the representations learned by CNNs to subjects' data using a metric-learning (Kulis, 2013) approach.[7] Accuracies have improved slightly (2%-4% increase) but not significantly, and our model still significantly outperforms both CNNs.

Taken together, these results show that a 3D shape inference model captures human performance better than deep CNNs on an object shape similarity task. This suggests that CNNs are, in fact, not good models of human vision. Although CNNs perform significantly better than chance, we believe this is due largely to the correlations between semantic object categories and shape features (Baldassi et al., 2013). Our study casts doubt on the claim that biological visual systems are optimized chiefly for object categorization, and that a system trained solely for object categorization will learn representations that are similar to ours. To the contrary, the low performances of the intermediate layers of the CNNs in our study suggests the opposite. Why does our 3D shape inference model perform better than CNNs? We believe this is due to the 3D nature of our model's shape representations. In contrast, a CNN trained to maximize object categorization performance learns to extract 3D features only to the extent that 3D information helps discriminate object categories. Therefore, it is unclear whether shape representations learned by CNNs carry 3D shape information. Since there is substantial evidence showing that human and monkey IT are selective for 3D shape (Orban, 2011), it becomes doubtful that CNNs offer good models of biological visual systems.

Lastly, we believe that our model is better suited than CNNs to understand visual perception in its totality because it is not intended simply as a model of object categorization. Biological visual systems solve a myriad of tasks from segmentation to scene perception, and our model can be readily extended to handle these diverse set of tasks. Moreover, vision is just one aspect of perception. We believe that our model—with its combination of a rich, modality-independent representational language, a forward model, and Bayesian inference—provides a promising theoretical framework for understanding not only visual, but also multisensory perception (Yildirim & Jacobs, 2013; Erdogan et al., 2015).

[7] See Supplementary Materials for further information.

Acknowledgments

This work was supported by AFOSR (FA9550-12-1-0303) and NSF (BCS-1400784) research grants.

References

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., & Zoccolan, D. (2013). Shape Similarity, Better than Semantic Membership, Accounts for the Structure of Visual Object Representations in a Population of Monkey Inferotemporal Neurons. *PLoS Comput Biol*, *9*(8), e1003167.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput Biol*, *10*(12), e1003963.

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From Sensory Signals to Modality-Independent Conceptual Rep-

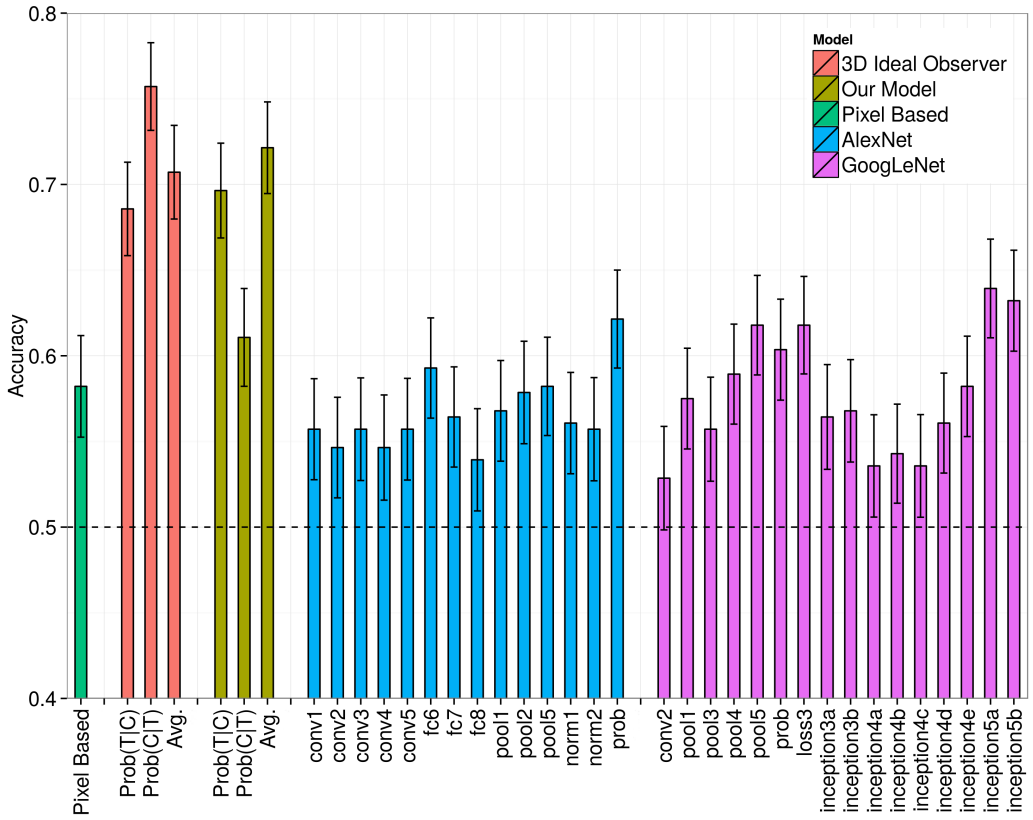


Figure 4: Prediction accuracies of each model using all experimental trials. The error bars denote SEMs and are calculated by bootstrapping with 1000 replications. P(C|T): similarity calculated from p(Comparison|Target). P(T|C): similarity calculated from p(Target|Comparison). Avg: similarity calculated from the average of these two probabilities. See the text for the layer labels for AlexNet and GoogLeNet. Note that the y-axis starts from 0.4.

resentations: A Probabilistic Language of Thought Approach. *PLoS Comput Biol*, *11*(11), e1004610.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv:1408.5093*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Comput Biol*, *10*(11), e1003915.

Kourtzi, Z., & Connor, C. E. (2011). Neural Representations for Object Perception: Structure, Category, and Adaptive Coding. *Annual Review of Neuroscience*, *34*(1), 45–67.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modelling Biological Vision and Brain Information Processing. *Annual Reviews of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Kulis, B. (2013). Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, *5*(4), 287–364.

Orban, G. A. (2011). The Extraction of 3d Shape in the Visual System of Human and Nonhuman Primates. *Annual Review of Neuroscience*, *34*(1), 361–388.

Peissig, J. J., & Tarr, M. J. (2007). Visual Object Recognition: Do We Know More Now Than We Did 20 Years Ago? *Annual Review of Psychology*, *58*, 75–96.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv:1409.4842*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing Properties of Neural Networks. *arXiv: 1312.6199*.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yildirim, I., & Jacobs, R. (2013). Transfer of Object Category Knowledge Across Visual and Haptic Modalities: Experimental and Computational Studies. *Cognition*, *126*, 135–148.

Cognitive Strategies in HCI and Their Implications on User Error

Marc Halbrügge (marc.halbruegge@tu-berlin.de)
Quality and Usability Lab, Telekom Innovation Laboratories
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin

Michael Quade (michael.quade@dai-labor.de)
DAI-Labor, Technische Universität Berlin
Ernst-Reuter-Platz 7, 10587 Berlin

Klaus-Peter Engelbrecht (klaus-peter.engelbrecht@alumni.tu-berlin.de)
Quality and Usability Lab, Telekom Innovation Laboratories
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin

Abstract

Human error while performing well-learned tasks on a computer is an infrequent, but pervasive problem. Such errors are often attributed to memory deficits, such as loss of activation or interference with other tasks (Altmann & Trafton, 2002). We are arguing that this view neglects the role of the environment. As embodied beings, humans make extensive use of external cues during the planning and execution of tasks. In this paper, we study how the visual interaction with a computer interface is linked to user errors. Gaze recordings confirm our hypothesis that the use of the environment increases when memory becomes weak. An existing cognitive model of sequential action and procedural error (Halbrügge, Quade, & Engelbrecht, 2015) is extended to account for the observed gaze behavior.

Keywords: Human Error; Memory for Goals; Eye-Tracking; ACT-R; Cognitive Modeling

Introduction and Related Work

Our daily life is dominated by routine sequential behavior like making coffee, washing clothes, or commuting to work. While we have practiced these activities hundreds of times, they are still subject to sporadic errors (Reason, 1990), a well-known example being postcompletion errors (e.g., forgetting a bank card in a vending machine after having completed a purchase, Byrne & Davis, 2006).

This paper presents recent advances of a cognitive modeling project on sequential behavior and error (Halbrügge & Engelbrecht, 2014; Halbrügge et al., 2015). The model is based on the memory for goals (MFG; Altmann & Trafton, 2002) framework which formulates how sequential behavior is controlled using declarative memory. We call this the *knowledge-in-the-head* strategy (Norman, 2002). By adding a second cognitive strategy to the model that relies more on external cues than on internal memory (the *knowledge-in-the-world* strategy; Norman, 2002), we can reproduce omission rates for different sub-tasks, comprising but not limited to postcompletion errors. Furthermore, the model can account for intrusions, an error type that has not been captured by MFG-based cognitive models before.

While the original model is targeting only error rates, the addition of the in-the-world strategy predicts behavior changes in other areas as well that allow empirical validation. In the visual domain, using the bottom-up in-the-world strategy means continously searching the UI for any suitable element, while the top-down in-the-head strategy just looks for the specific element that is needed to perform the current task. We therefore designed an eye-tracking experiment that should reproduce our previous findings and should confirm the predictions of the knowledge-in-the-world assumption in the visual domain.

The rest of this paper is organized as follows. We first give an overview on the types of errors we are aiming to address alongside the theoretical underpinnings of our model. We will then introduce an experiment that shows the connection of visual behavior to procedural error and compare the findings to the predictions of our cognitive model. We conclude with a discussion of the strengths and limitations of the model and a summary of our contributions.

Procedural Error

Human error in its most general sense is commonly refered to as "those occasions in which a planned sequence of mental or physical activities fail to achieve its intended outcome, [and] when these failures cannot be attributed to the intervention of some chance agency" (Reason, 1990). It can be further decomposed based on the level of action control on which an error occurs (Rasmussen, 1983). Knowledge-based behavior on the highest level of control is characterized by explicit planning. Skill-based behavior on the opposite, lowest level consists mainly of sensory-motor actions without conscious control. Interaction with computer systems is mainly located on the intermediate rule-based level of action control. On this level, behavior is generated using stored rules and procedures that have been formed during training or earlier encounters. Errors on the rule-based level are not very frequent (below 5%), but pervasive and cannot be eliminated through training (Reason, 1990). While Norman (2002) subsumes these errors within the 'slips' category, Reason (1990) refers to them as 'lapses'. We escape this ambiguity by using the term *procedural error*. Procedural error is defined as the violation of the (optimal) path to the current goal by a non-optimal action. This can either be the addition of an unnecessary or even hindering action, which is called an *intrusion*. Or a necessary step can be left out, constituting an *omission*.

Memory for Goals Theory

An explanation of procedural *error* must incorporate the generation of *correct* behavior as well. A very promising theoretical model of sequential action is the Memory for Goals theory (MFG; Altmann & Trafton, 2002). The MFG proposes that subgoals, i.e., atomic steps towards a goal, are represented in human memory, thereby underlying memory effects like time-dependent and noisy *activation*, *interference*, and associative *priming*.

Within the MFG theory, errors arise when the activation of a goal is not high enough to surpass interfering goals or even falls below a general retrieval threshold. Cognitive models based on the MFG assumptions have been shown to explain procedural errors in the HCI domain, namely omissions, very well (e.g., Trafton, Altmann, & Ratwani, 2011; Hiatt & Trafton, 2015; Halbrügge et al., 2015; Li, Blandford, Cairns, & Young, 2008).

The Role of the Environment

The MFG theory is clearly focused on managing task sequences in memory, i.e., within their head. This has been critized for neglecting the role of the environment (e.g., Salvucci, 2010). As embodied beings, humans strive to reduce cognitive complexity by exploiting the content and structure of the external world.

Recent research has shown that better predictions can be achieved and new error domains can be covered by extending the MFG with an activation process that relies on external cues (Halbrügge et al., 2015; Hiatt & Trafton, 2015). We propose that when a user cannot retrieve the next goal, they revert to an externalization strategy and randomly search the visual scene (i.e., the UI) for interactive elements. Whenever an element is found, the user tries to retrieve a goal that relates to this UI element. Because visually attending the element increases the activation of related goals through priming, a goal that had previously been forgotten may now surpass the retrieval threshold. As a consequence, the planned sequence of actions may be resumed. But because visual priming is independent of the currently planned sequence, it may also help retrieve an outdated goal from a previous trial, or the planned sequence may be resumed at an incorrect position.

Task- and Device-Orientation

The best known examples of procedural error are postcompletion errors (e.g., forgetting the originals in the copy machine; Byrne & Davis, 2006) and initialization errors (e.g., forgetting to reset Caps Lock before typing a password; Gray, 2000). Common to both of them is that these errors happen during procedural steps that do not directly contribute to the users' actual goals (i.e., making copies; logging into a system). This common property of goal-irrelevance of a sub-task has been coined *device-orientation* (Ament, Cox, Blandford, & Brumby, 2013; Gray, 2000), its opposite is analogously called *task-orientation*. The concept of device-orientation is based on the MFG by assuming that device-oriented tasks are "more weakly represented in memory". While Ament et al. (2013) discuss different encoding or lack of rehearsal as possible reasons for lower activation of device-oriented tasks, we are assuming lack of priming in this case (Halbrügge & Engelbrecht, 2014). We will call this the *task priming assumption* in the following.

The Role of the Application Logic

The elimination of device-oriented tasks is a reasonable design strategy for error reduction, but does not succeed in all cases. When a device-oriented step can not be avoided (e.g., the removal of the bank card from a teller machine), it is often made obligatory to make sure that users can not omit it. In the ATM example, this would mean not handing out the money before the user has taken back their card.

Previous research has shown that the impact of whether a subtask is obligatory or not is much higher than whether it is device-oriented or not, and the interaction of both needs to be taken into account when researching procedural error in real-world settings (Halbrügge et al., 2015). It is also worth noting that memory-based processing as put forth by the MFG can hardly explain why obligatory tasks are less prone to omissions. The *knowledge-in-the-world* assumption fills this gap.

Experiment

As errors are relatively infrequent, but proper statistical analysis needs sufficiently big case numbers, researchers have developed several strategies to artificially increase error rates in the laboratory. Among these are interruptions by the experimenter (Li et al., 2008; Trafton et al., 2011), secondary tasks (Byrne & Davis, 2006; Ruh, Cooper, & Mareschal, 2010), or special UIs that make users feel lost easily (Hiatt & Trafton, 2015). Because we want to maximize external validity, we rejected all of these options and chose to observe human error during repeated interaction with a real world application instead. We selected a kitchen assistance system from a "smart home" environment for the experiment. The assistant aims at helping with the preparation of meals by suggesting recipes, calculating ingredient amounts and maintaining shopping lists. A screenshot of the recipe search screen of the kitchen assistant is displayed in Figure 1.

Methods

Participants 24 members of the Technische Universität Berlin paid participant pool, 15 women and 9 men, aged between 18 and 54 (M=29.4, SD=9.4), took part in the experiment conducted in January 2015. As the instructions were given in German, only fluent German speakers were allowed.

Materials A personal computer with 23" (58.4 cm) touch screen and a 10" (25.7 cm) tablet were used to display the interface of the kitchen assistant. While the large screen operated in landscape mode, portrait mode was used for the tablet. We created two variations of the UI of the kitchen assistant that were targeted at the large screen and the tablet, respectively. All user actions were recorded by the computer system.

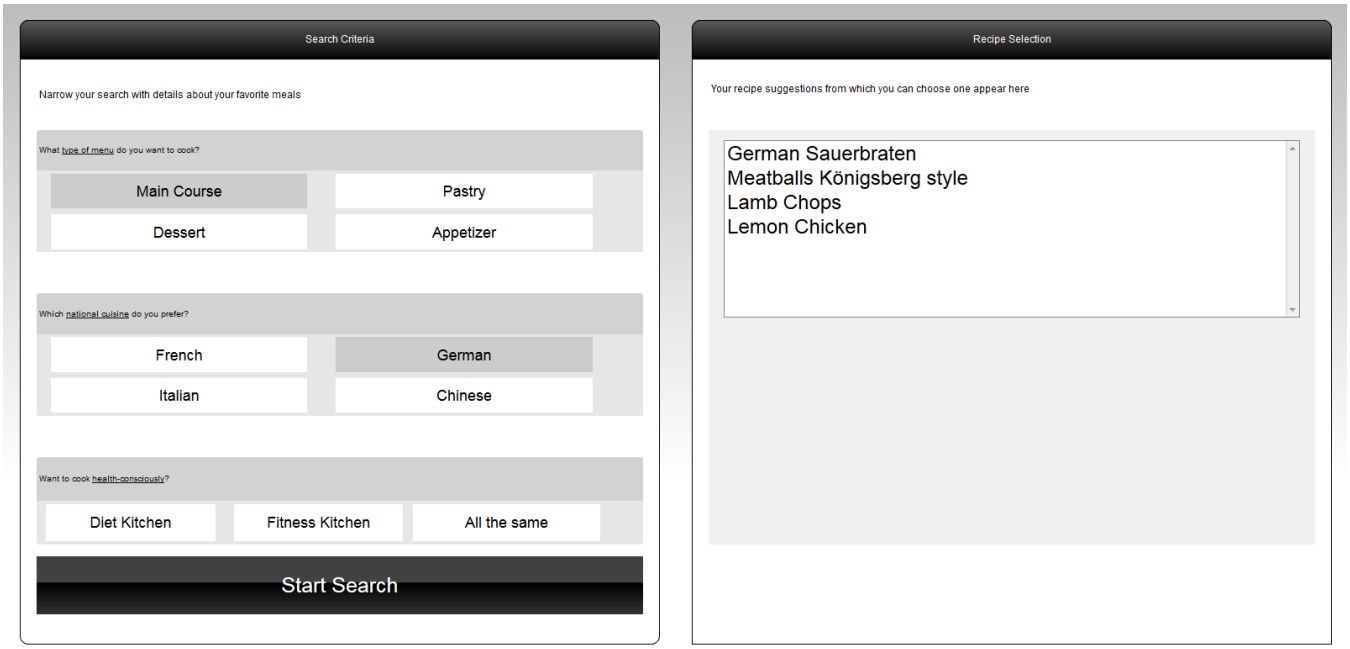


Figure 1: Screenshot of the English version kitchen assistant used for the experiment.

The participants' gaze was recorded using a SR Research Ltd EyeLink II head-mounted eye-tracker. Only the dominant eye was tracked; the sampling frequency was 250 Hz. Gaze recording was only applied during task blocks that used the large monitor because the visual angles between UI elements on the tablet were too small to obtain an unambigous mapping from gaze to element. Nevertheless, the eye-tracker was worn by the participants during the complete procedure as its scene camera view was recorded for subsequent error identification. The experiment was conducted in a laboratory with fixed lighting conditions.

Design We used a four-factor within-subject design, the factors being the physical device used, the UI variant, whether a sub-task was obligatory, and whether it was device-oriented as opposed to task-oriented. We collected user errors, task completion times, and gaze position. The participants completed a total of 46 tasks grouped into 4 blocks. Physical device, UI variant, and block sequence were varied randomly, but counterbalanced across the experiment.

Procedure After having played a simple game on each of the two devices to get accustomed with the respective touch technology, the participants received training on the kitchen assistant. The training covered all parts of the application that were used during the actual experiment. Each block of tasks began with relativly simple tasks like "Search for German main dishes and select lamb chops". Afterwards, the ingredients to a recipe were collected and some of them were added to a shopping list that is part of the kitchen assistant (e.g., "Create a shopping list for six servings and check-off garlic"). The instructions followed the experiment described in Halbrügge et al. (2015) closely, with only one new type of device-oriented non-obligatory tasks added in order to gain more insights about this special combination. When a participant made an error during a trial, they were not interrupted but informed after the trial and were given the chance to repeat this trial a single time. The complete procedure lasted approximately 45 minutes.

Results

Errors A total of 6921 clicks were recorded. We observed 85 (1.2%) omissions and 53 (.8%) intrusions. Because of the unequal number of observations per trial, mixed models were used to analyze the data (Bates, Maechler, Bolker, & Walker, 2013). Neither physical device nor UI variant showed relations to the error rate (mixed logit model with subject and task block as random factors, all $p > .3$). Obligatory task steps were less prone to errors ($z = -2.1, p = .034$) and device-oriented steps showed higher error rates than their task-oriented counterparts ($z = 7.8, p < .001$). Together with the highly significant interaction ($z = 4.4, p < .001$), this is mainly due to the relatively high omission rate for device-oriented non-obligatory task steps (e.g., checking off previously selected ingredients; see Figure 2).

Eye-Tracking The gaze of three participants could not be recorded because of calibration failure. For the remaining 21 participants, the raw screen coordinates were mapped to dynamic areas of interest (AOI) around the UI elements using a computational bridge between the eye-tracker and the web browser that rendered the UI (Halbrügge, 2015). Individual gaze positions were collapsed into fixations using a hidden Markov model (HMM) approach. Based on common rule-of-

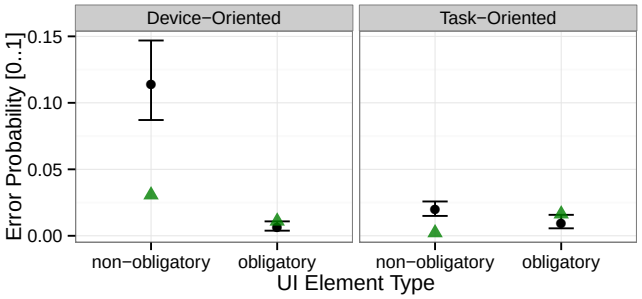


Figure 2: Error probabilities for the experiment (logarithmic scale). Error bars are 95% confidence intervals using the Agresti-Coull method. ▲ denote predictions of the revised model (400 runs).

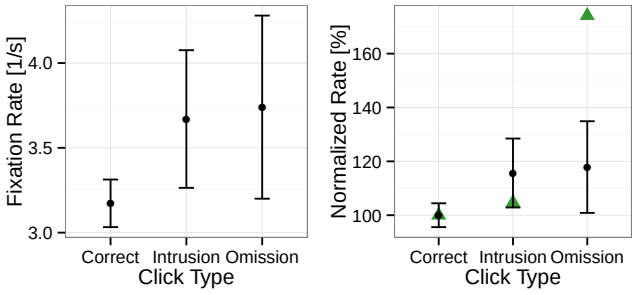


Figure 3: Fixation rate before performing a correct vs. an erroneous click. Error bars are 95% confidence intervals. Right: Rates have been normalized to the baseline of correct clicks to allow a comparison between data and cognitive model. ▲ denote predictions of the revised model (400 runs).

thumb values of maximum 100 deg/s for fixations and minimum 300 deg/s for saccades (Salvucci & Goldberg, 2000), the parameters for the states of the HMM and the respective transition probabilities were estimated from the recorded data.

In order to examine the knowledge-in-the-world assumption, the gaze recordings were split into segments based on the clicks within a trial (e.g., a trial with five clicks yielded four segments). Because of the varying length of these segments, we use the fixation rate (number of fixations divided by segment length) as dependent variable. A mixed model with subject and sub-task as random factors yielded significantly higher rates for segments before erroneous clicks compared to correct clicks ($F_{2,1952.5} = 5.3$, $p = .005$, see Figure 3). This could still mean that the users were recurrently fixating the ultimately clicked UI element instead of searching the screen. We addressed this by counting the fixations on the AOI corresponding to the element that concluded its segment. Because the resulting counts are extremely right-skewed (median = 1), we divided them into two groups for statistical analysis. A logit mixed model with subject and sub-task as random factors revealed that the probability to fixate the ultimately clicked element at least once during a segment was actually lower before errors ($z = -2.5, p = .012$).

Discussion

The results partially reproduce the findings of our previous studies. Compared to Halbrügge et al. (2015), the combination of non-obligatory device-oriented tasks yielded a much higher error rate. This is probably due to the changed set of user tasks that seems to have included device-oriented task steps that were much harder to remember than the ones used before.

The eye-tracking data confirms the knowledge-in-the-world assumption nicely. The higher fixation rate before erroneous clicks matches the proposed process of (random) search for 'inviting' elements on the surface of the UI. In principle, this could also be caused by memory-based processing, e.g., a re-fixation strategy to strengthen the activation of the current subgoal through visual priming. But the AOI-based analysis shows that the frequent fixations before errors are on other elements than the one that is eventually clicked.

Cognitive Model

To gain more insight in the implications of our theoretical assumptions, the *task priming* and the *knowledge-in-the-world* assumptions have been implemented in a cognitive model based on the ACT-R framework (Anderson et al., 2004). A simplified flow-chart of the model is displayed in Figure 4. The model had been fit to the error rates of the previous experiment and reached good fits in this domain (Halbrügge et al., 2015).[1] The new data presented here allows constraining the existing model. The model fit to the error rates of the current data is still good with R^2=.76, but somewhat increased RMSE=.044. The eye-tracking results confirm the theoretical assumptions at least qualitatively. But what about the actual quantitative predictions of the model in the visual domain?

As ACT-R's visual module operates on an abstract attention layer (as opposed to raw eye movements), the model's visual behavior cannot be compared directly to the human sample. Nevertheless, effects that were significant in the human data should be present in the model results as well. The initial model produces fixation rates that are actually lower for erroneous trials compared to correct trials, in contrary to both the theory and the human data. This happens because the memory test that is part of the knowledge-in-the-world strategy (try-retrieve-goal-for-element in Figure 4) is not restricted in time. If it fails to retrieve a matching goal chunk, this is only signaled after approximately one second. As a result, the visual search process is slowed down considerably.

Revised Model

In order to speed up the knowledge-in-the-world strategy, we make use of ACT-R's temporal buffer (Taatgen, Van Rijn, & Anderson, 2007). A timer is started with the initiation of the memory test and a single production was added that aban-

[1]The source code of the model is available for download at http://www.tu-berlin.de/?id=135088

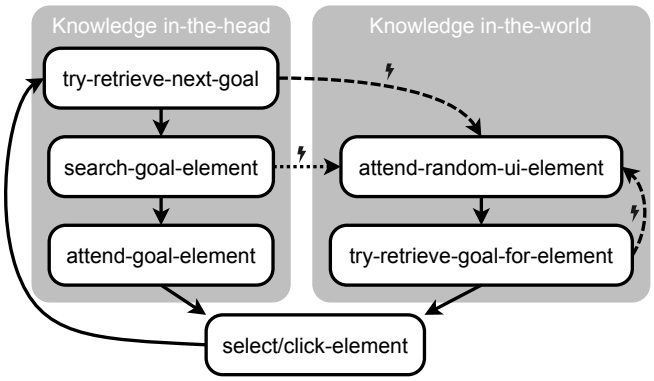


Figure 4: Simplified flow chart of the cognitive model. Dashed arrows denote retrieval errors, the dotted arrow denotes visual search failure.

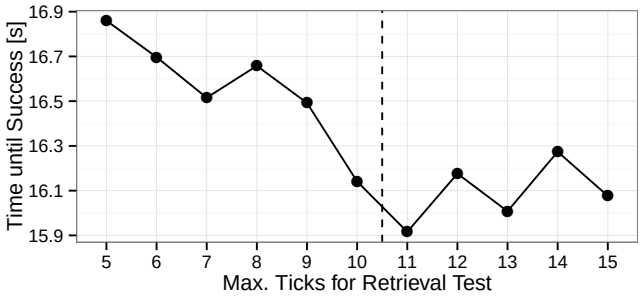


Figure 5: Average time until successful task completion when using the knowledge-in-the-world strategy as function of the maximum time ticks spent on the retrieval test. The vertical dashed line demarks the average wait tick value chosen for the model. (3000 model runs)

dons the retrieval after a fixed amount of time ticks. Based on the observed fixation rate between 3 and 4 per second before erroneous clicks, we set the tick threshold to approximately 250 ms (9 to 12 ticks using ACT-R standard parameters).

Comparing the model predictions based on the fixation rates is tricky because ACT-R only models shifts of attention and assumes completely stable gaze otherwise, which results in unnaturally low fixation rates. An exploratory approach to this based on the relative change compared to the 'correct click' baseline is shown in Figure 3. The corresponding fit is unimpressive with R^2=.41, but should be interpreted with care. Of higher importance is the qualitative result that error trials show increased fixation rates.

Errors After having optimized the model for visual behavior, the fit in the error domain somewhat degrades with R^2=.70 and RMSE=.044 (see Figure 2).

General Discussion

The current paper presents an empirical study and a cognitive user model of procedural error. The study has been designed to foster external validity, featuring a real-world application in a household scenario and drawing participants mainly from non-student populations. The model extends the activation-based Memory for Goals theory (Altmann & Trafton, 2002) by highlighting the importance of external cues during sequential action. Internal cues are divided into task-oriented, i.e., steps that directly contribute to the users' goals, and device-oriented ones (Ament et al., 2013). Together with the assumption that only task-oriented steps receive additional priming from the user's overall goal, this allows not only to explain our data, but also provides an acceptable explanation of how device-orientation actually affects sequential behavior. The role of external cues as formulated in the knowledge-in-the-world assumption is fostered by the eye-tracking results.

Eye-tracking has been used before to predict special subtypes of procedural error (Ratwani & Trafton, 2011) with good success. Ratwani and Trafton's system uses a cognitive model based on the MFG theory in combination with a statistical classifier that uses eye-tracking data as input to predict and prevent postcompletion error. The approach presented in this paper differs therefrom in two important ways. Firstly, we are targeting not only omissions of the final step within a sequence, but other types of omissions and intrusions as well. And secondly, we are trying not only to predict *when* a user makes an error, but also *why* this happens and have therefore incorporated the necessary visual processes into our cognitive model.

How do the strategies that we propose relate to the soft constraints hypothesis (SCH; Gray & Boehm-Davis, 2000; Gray, Sims, Fu, & Schoelles, 2006)? According to this hypothesis, users select (micro-) strategies based on a temporal cost-benefit tradeoff. The reduction of the waiting time for the retrieval test that is part of the knowledge-in-the-world strategy can already be viewed as the application of such a tradeoff. After a few hundred milliseconds, it may be more beneficial to try a new visual target than to wait for the successful retrieval of a matching goal chunk. We explored this hypothesis by varying the maximum time ticks the model waits until abandoning the retrieval test and comparing the success rates of the knowledge-in-the-world strategy depending on the ticks. Keeping all other model parameters fixed, the results show that the wait time chosen based on the fixation rates is also close to a local minimum of the time spent per successful click (see Figure 5). This criterion represents a time/benefit tradeoff as proposed by the SCH.

While providing good fits to the data, the model also has several limitations. Firstly, the model only covers expert behavior. The initial formation of the task sequence by novice human users is beyond its capabilities. The model also does not account for errors caused by the UI design violating general expectations of its users towards computer systems.

Secondly, the empirical basis of the model is limited. So far, we have only collected data using a single paradigm. Changing the experimental tasks resulted in changed error rates, but the overall pattern remained stable and the model fit

is still satisfactory. While the current addition of eye-tracking data represents a significant extension of its empirical foundation, the generalizability of the model still needs further investigation.

Conclusions

We have presented a model of procedural error grounded in the Memory for Goals theory. The model consists of two interdependent cognitive strategies, one relying only on memory and another that is mainly using cues from the environment. The underlying assumptions, namely the *task priming assumption* and the *knowledge-in-the-world assumption*, are confirmed by an error analysis of newly gathered data without further parameter fitting. Eye-tracking data recorded at the same time qualitatively backs our theoretical assumptions. A revised cognitive model provides reasonable fit to both the error and the gaze data.

Acknowledgments

This work was supported by DFG grant MO 1038/18-1 ("Automatische Usability-Evaluierung modellbasierter Interaktionssysteme für Ambient Assisted Living").

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive science*, *26*(1), 39–83.

Ament, M. G., Cox, A. L., Blandford, A., & Brumby, D. P. (2013). Making a task difficult: Evidence that device-oriented steps are effortful and error-prone. *Journal of experimental psychology: applied*, *19*(3), 195.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036–1060.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using eigen and s4 [Computer software manual]. (R package version 1.0-5)

Byrne, M. D., & Davis, E. M. (2006). Task structure and postcompletion error in the execution of a routine procedure. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(4), 627–638.

Gray, W. D. (2000). The nature and processing of errors in interactive behavior. *Cognitive Science*, *24*(2), 205-248.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, *6*(4), 322.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological review*, *113*(3), 461.

Halbrügge, M. (2015). Automatic online analysis of eye-tracking data for dynamic HTML-based user interfaces. In K. Gramann & T. O. Zander (Eds.), *Berliner werkstatt mensch-maschine-systeme.* Berlin: Technische Universität Berlin.

Halbrügge, M., & Engelbrecht, K.-P. (2014). An activation-based model of execution delays of specific task steps. *Cognitive Processing*, *15*, S107-S110.

Halbrügge, M., Quade, M., & Engelbrecht, K.-P. (2015). A predictive model of human error based on user interface development models and a cognitive architecture. In N. A. Taatgen, M. K. van Vugt, J. P. Borst, & K. Mehlhorn (Eds.), *Proceedings of the 13th international conference on cognitive modeling* (p. 238-243). Groningen, the Netherlands: University of Groningen.

Hiatt, L. M., & Trafton, J. G. (2015). An activation-based model of routine sequence errors. In N. A. Taatgen, M. K. van Vugt, J. P. Borst, & K. Mehlhorn (Eds.), *Proceedings of the 13th international conference on cognitive modeling* (p. 244-249). Groningen, the Netherlands: University of Groningen.

Li, S. Y., Blandford, A., Cairns, P., & Young, R. M. (2008). The effect of interruptions on postcompletion and other procedural errors: an account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied*, *14*(4), 314.

Norman, D. A. (2002). *The design of everyday things*. Basic books.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*, *13*, 257–266.

Ratwani, R. M., & Trafton, J. G. (2011). A real-time eye tracking system for predicting and preventing postcompletion errors. *Human–Computer Interaction*, *26*(3), 205–245.

Reason, J. (1990). *Human error*. New York, NY: Cambridge University Press.

Ruh, N., Cooper, R. P., & Mareschal, D. (2010). Action selection in complex routinized sequential behaviors. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 955.

Salvucci, D. D. (2010). On reconstruction of task context after interruption. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 89–92).

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78).

Taatgen, N. A., Van Rijn, H., & Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: the role of cognition, attention, and learning. *Psychological Review*, *114*(3), 577.

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*, 134–143. doi: 10.1016/j.cogsys.2010.07.010

Connectionist Semantic Systematicity in Language Production

Jesús Calvillo (jesusc@coli.uni-saarland.de)

Harm Brouwer (brouwer@coli.uni-saarland.de)

Matthew W. Crocker (crocker@coli.uni-saarland.de)
Saarland University, Germany

Abstract

A novel connectionist model of sentence production is presented, which employs rich situation model representations originally proposed for modeling systematicity in comprehension (Frank, Haselager, & van Rooij, 2009). The high overall performance of our model demonstrates that such representations are not only suitable for comprehension, but also for modeling language production. Further, the model is able to produce novel encodings (active vs. passive) for a particular semantics, as well as generate such encodings for previously unseen situations, thus demonstrating both syntactic and semantic systematicity. Our results provide yet further evidence that such connectionist approaches can achieve systematicity, in production as well as comprehension.

Keywords: systematicity; sentence production; connectionist; semantics; syntax; neural networks

Introduction

A defining characteristic of human language is systematicity: "the ability to produce/understand some sentences is intrinsically connected to the ability to produce/understand certain others" (Fodor & Pylyshyn, 1988, p. 37). Further, Fodor and Pylyshyn (1988) argue that connectionist models are not able to display systematicity without implementing a classical symbol system.

The connectionist comprehension model developed by Frank et al. (2009), however, challenges this highly debated assertion, by developing a connectionist model of comprehension which is argued to achieve relevant levels of systematicity. Their model constructs a a *situation model* (see Zwaan and Radvansky (1998)) of the state-of-affairs described by a sentence that also incorporates world knowledge-driven inferences. When the model processes a sentence like 'a boy plays soccer', for instance, it not only recovers the explicit, literal propositional content, but also constructs a more complete situation model in which a boy is likely playing outside on a field, with a ball, with others, and so forth. In this way it differs from other connectionist models of language comprehension and production, that typically employ simpler meaning representations, such as case-roles (Chang, Dell, & Bock, 2006; Mayberry, Crocker, & Knoeferle, 2009; Brouwer, 2014, among others). Crucially, Frank et al. (2009)'s model generalizes to both sentences and situations that it has not seen during training, exhibiting different levels of semantic systematicity and is argued to provide an important step in the direction of psychologically plausible models of language comprehension.

In the present paper, we examine whether the approach developed by Frank et al. (2009), is equally well suited to language production, and present a connectionist production model that generates sentences from these rich situation models. We show that our model successfully learns to produce sentences from these rich meaning representations. Crucially, we demonstrate that this model is able to describe unseen situations, demonstrating semantic systematicity similar to Frank et al. (2009), as well as produce alternative encodings (e.g. active/passive) for a given situation, that were not seen during training and thus demonstrate syntactic systematicity.

Method

We employ an extended Simple Recurrent Neural Network architecture (SRN) (Elman, 1990) to generate sentences from rich semantic representations, called Distributed Situation Space (DSS) vectors. The DSS model (Frank, Koppen, Noordman, & Vonk, 2003; Frank et al., 2009) is a distributed scheme for meaning, in which the meaning of a situation—a state of affairs—is represented as a *situation vector* in a high-dimensional "situation-state space".

This section is organized as follows: first, we introduce the Distributed Situation Space as described by Frank et al. (2003, 2009); then, we explain the vectors that we use as well as the architecture of the model; afterwards the training and evaluation schema is presented; and finally we present the results obtained from the evaluation.

Distributed Situation Space

The DSS model defines a microworld in terms of a finite set of *basic events* (e.g., $play(charlie, chess)$ and $place(heidi, bedroom)$)—the smallest meaning-discerning units of propositional meaning in that world. While these basic events can be defined in several ways, we adopt the microworld presented by Frank et al. (2009). Situations in a microworld are represented in terms of these basic events, which can be conjoined to form *complex events* (e.g., $play(charlie, chess) \wedge win(charlie)$). Not all conjunctions of basic events are, however, possible or equally likely. That is, world knowledge can pose both hard (e.g., physical) and probabilistic (e.g., preferential) constraints on event co-occurrence. A situation-state space is a large set of m situations defined in terms of n basic events, effectively yielding an $m \times n$ matrix (see Table 1). Each of the m situations in this matrix is encoded by setting basic events that *are the case* in a given situation to 1 (True) and those that are not to 0 (False). A situation-state space matrix is constructed by sampling m situations (using a non-deterministic sampling proce-

Table 1: Situation-state space.

	basic event$_1$	basic event$_2$	basic event$_3$	...	basic event$_n$
situation$_1$	1	0	0	...	1
situation$_2$	0	1	1	...	1
situation$_3$	1	1	0	...	0
...	.	.	.	...	.
situation$_m$	0	1	0	...	0

dure) such that no situation violates any hard world knowledge constraints, and such that the m situations approximate the probabilistic nature of the microworld in terms of the (co-)occurrence probability of the n basic events. The resulting situation-state space matrix is then effectively one big truth table, in which each column represents the *situation vector* for an individual basic event; that is, each column of the matrix encodes the meaning of a basic event in terms of its co-occurrence with all other basic events. The situation vectors of complex events (combinations of basic events) can be found through propositional logic, allowing to capture phenomena such as negation, conjunction and disjunction; conversely, complex events also allow us to capture aspects of modality and quantification.

This situation-state space encodes *all* knowledge about the microworld, and situation vectors capture dependencies between situations in this space, thereby allowing for 'world knowledge'-driven inference.

DSS representations have been successfully used in a connectionist comprehension model (Frank et al., 2009). That is, Frank et al. (2009) defined a small *microworld* (see section two of their paper), consisting of 44 basic events, centred around three people, and a few games, toys, and places. They constructed a situation-state space by sampling $25,000$ situations in this microworld, and reduced the dimensionality of the resulting $25k$-dimensional situation vectors to 150-dimensions using a competitive layer algorithm. Using these reduced vectors, they show that their model is not only able to comprehend sentences that it has seen during training, but that is also able to comprehend sentences and situations that it has never seen before (i.e., it shows semantic systematicity).

The vectors that we use for production are slightly different. Namely, the dimensionality of the situation-space was not reduced, since its reduction involves some loss of information. Rather, for each $25k$-dimensional situation vector associated to a particular sentence, which in turn is associated to a (complex) event; we compute a *belief situation vector*, whose dimensionality is equal to the number of basic events in the microworld (44 in this case) and the value of each dimension is equal to the conditional probability of the corresponding basic event, given the (complex) event associated to the sentence.[1] In other words, each dimension represents how likely each basic event is to be true, given the situation that is expressed by the sentence.

The Model

Our model architecture, as we can see in Figure 1, is broadly similar to the one used by Frank et al. (2009), with the main difference being that the inputs and outputs are reversed; it maps DSS representations onto sequences of localist word representations. Similar to Frank et al. (2009), we stress that it is not intended to model human language development.

The model consists of a 45-unit input layer, a 120-unit recurrent hidden (tanh) layer, and a 43 unit (softmax) output layer. The dimensionality of the input layer corresponds to the 44 basic events in the microworld, plus one extra binary unit that indicates whether the model must output an active sentence (1), or a passive one (0). The dimensionality of the output layer matches the number of available words in the grammar (42), plus the end-of-sentence marker.

Time in the model is discrete. At each time-step t, activation propagates forward following the trajectory: input $\rightarrow$ hidden $\rightarrow$ output. The activation of the output layer yields a probability distribution over the available words. We define the word produced at time-step t as the one with highest probability (highest activation). The model stops producing words when an end-of-sentence marker has been produced.

The hidden layer also receives a copy of its own activation pattern at the previous time-step $t-1$, through a 120-unit context layer, which is set to zero at the beginning of each sentence. These units help to preserve some memory of what has been produced expanding several time steps in the past, and allow the model to generate sentences of variable length.

In addition, the hidden layer receives the identity of the word that was produced at time-step $t-1$ (zeros at $t=0$) through *monitoring units* that connect the output layer to the hidden layer, where only the output unit corresponding to the produced word at time-step $t-1$ is activated (set to 1), while all other units are set to 0.

Finally, the hidden and output layers also receive input from a bias unit (with a constant activation value of 1).

Examples Set

The examples set consists of a set of pairs $\{(DSS_1, \varphi_1), \ldots, (DSS_n, \varphi_n))\}$ where each $DSS_i \in [0,1]^{45}$ corresponds to a DSS representation plus an extra bit indicating whether the system must produce an active sentence (1) or a passive one (0); and $\varphi_i = \{sent_1, \ldots, sent_k\}$ where $sent_j$ is a sentence, a sequence of words $word_1, \ldots, word_n$, expressing the information contained in DSS_i. Each set φ_i represents all the possible sentences that express the

[1] This vector is computed by calculating the dot product between the situation-state matrix and the original $25k$-dimensional situation vector, and then normalizing each dimension of the resulting vector by the sum over the dimensions of the original $25k$-dimensional situation vector.

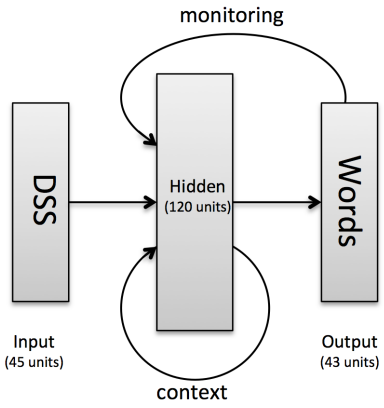


Figure 1: Model architecture.

information contained in the corresponding DSS_i and in the expected voice.

The sentences that we use are those generated by the microlanguage defined by Frank et al. (2009) (see their Tables 5–8). This microlanguage consists of 40 words that can be combined into 13556 sentences according to its grammar. We minimally modified this grammar. First, we introduced the determiners that were missing (a, the); and second, we added an end-of-sentence marker (a period) to the sentences. Leaving a total of 43 words that were encoded at the output layer as localist vectors.

Sentences that expressed unlawful situations according to the microworld rules, and therefore whose DSS belief vectors were empty, were discarded; leaving a total of 8201 lawful sentences. From these, 6782 sentences were in active voice and 1419 in passive. Note that this set contains sentences with simple semantics (e.g., "charlie plays chess ." $\rightarrow$ $play(charlie, chess)$), as well as sentences with complex semantics (e.g., "a girl plays chess ." $\rightarrow$ $play(heidi, chess) \vee play(sophia, chess)$).

There were a total of 782 unique DSS representations, from which 424 were related to both passive and active sentences. The rest (358) corresponded to situations that could only be expressed by active sentences according to the grammar. More concretely, the grammar presented in Frank et al. (2009) does not define passive sentences for situations where the object of the action is either a person (e.g. "Heidi beats Charlie.") or undefined (e.g. "Charlie plays.").

Training and Evaluation

In order to asses the performance of the model in terms of accuracy and generalization, we employed a 10-fold cross-validation schema. First, we divided the DSS representations into two sets: the first one (setAP) corresponding to those associated to both active and passive sentences and the second one (setA) corresponding to DSS representations that were related only to active sentences.

The first set (setAP) allowed for three different testing conditions:

- **Condition 1:** Situations for which the model has seen only active sentences, and a passive is queried.
- **Condition 2:** Situations for which the model has seen only passive sentences, and an active is queried.
- **Condition 3:** Completely new situations (not seen during training), passive and active sentences are queried.

The second set (setA) allowed for two different testing conditions:

- **Condition 4:** Situations for which the model has seen only active sentences, and a passive is queried.
- **Condition 5:** Completely new situations (not seen during training), passive and active sentences are queried.

These conditions represent different levels of generalization or systematicity. In all cases, the queried sentence type has never been seen by the model. For conditions 1, 2 and 4 the model has seen the situation but not in the queried voice. Importantly, for conditions 3 and 5, the model has never seen the situation itself. Finally, for conditions 4 and 5, where passives are queried, not only the system has not seen the passive sentences, but also they are not defined by the grammar.

SetAP was randomly shuffled and split into 10 folds of 90% training and 10% testing DSS representations, meaning per fold 382 training and 42 testing situations. For each fold, the testing DSS representations were further split into the 3 conditions, rendering 14 different testing DSS representations per condition, per fold.

SetA was also shuffled and split into 10 folds, but in order to preserve uniformity, for each fold and for testing, 14 DSS representations were drawn per condition; meaning that each fold contained 28 testing and 330 training DSS representations.

Finally, for condition 1 the DSS representations were coupled with their corresponding active sentences and incorporated into the training set (while the passive sentences remained in the testing set); vice-versa for condition 2. Similarly, for condition 4 the active sentences were incorporated into the training set, while during testing the system will be queried for a passive construction, even though there is none according to the grammar.

Training Procedure We trained the model using cross-entropy backpropagation (Rumelhart, Hinton, & Williams, 1986) with weight updates after each word in the sentence of each training item. Prior to training, all weights on the projections between layers were initialized with random values drawn from a normal distribution $\mathcal{N}(0, 0.1)$. The weights on the bias projections were initially set to zero.

During training, the monitoring units were set at time t to what the model was supposed to produce at time $t-1$ (zeros for $t = 0$). This reflects the notion that during training the word contained in the training sentence at time-step $t-1$ should be the one informing the next time step, regardless of

Table 2: Similarity scores for each test condition.

Cond.	Query	Similarity (%)	PerfectMatch (%)
1	pas	97.66	87.86
2	act	97.58	92.86
3	act	98.35	93.57
3	pas	96.79	83.57
5	act	95.08	85.0

the previously produced (and possibly different) word. During testing, the monitoring units are set to 1.0 for the word that was actually produced and 0.0 everywhere else.

The model was trained for a maximum of 200 epochs, each epoch consisting of a full presentation of the training set, which was randomized before each epoch. Note that each item of this set consisted of a DSS_i paired with one of the possible sentence realizations describing the state of affairs represented in DSS_i. Hence, during each epoch, the model saw all the possible realizations of DSS_i contained in the training set for a given fold. We employed an initial learning rate of 0.124 which was halved each time there was no improvement of performance on the training set during 15 epochs. No momentum was used. Training halted if the maximum number of epochs was reached or if there was no performance improvement on the training set over a 40-epoch interval.

Sentence Level Evaluation For a given DSS representation DSS_i, the model produces a sequence of words $\hat{s}_i$ constituting a sentence describing the state-of-affairs represented in DSS_i. Because a DSS_i can be described by one or more sentence(s), we assume that the output of the model is perfect if the sentence produced $\hat{s}_i$ is part of the set φ_i of all possible realizations of DSS_i in the queried voice.

However, sometimes the output of the model $\hat{s}_i$ for a DSS_i does not perfectly match any of the sentences in φ_i. As such, we also compute the similarity between the output of the model $\hat{s}_i$, and each sentence in φ_i. This similarity is derived from their Levenshtein distance (Levenshtein, 1966); which is the number of insertions, deletions or substitutions that are needed in order to transform one string into another. More formally, Levenshtein similarity $sim(s_1, s_2)$ between two sentences s_1 and s_2 is defined as:

$$sim(s_1, s_2) = 1 - \frac{distance(s_1, s_2)}{max(length(s_1), length(s_2))} \quad (1)$$

where *distance* is the Levenshtein distance. This similarity measure is 0 when the sentences are completely different and 1 when they are the same. Thus, for each item i in the training and test set, we obtain a similarity value:

$$sim(\hat{s}_i) = \max_{s \in \varphi_i} sim(\hat{s}_i, s) \quad (2)$$

Results

We trained 10 instances of the model, corresponding to each fold as described above. Each instance was initialized with a different set of weight matrices. The scores reported below are averaged over these instances.

On the training set, the model achieved an average similarity score of 99.43% (and 98,23% perfect matches). This shows that the model is able to learn to transform a DSS situation vector into a sequence of words describing the state-of-affairs that the vector represents.

Regarding the test conditions, Table 2 shows the average similarity scores for each of them. For conditions 4 and 5, where passives are queried but there are no example sentences given by the grammar, no similarity scores can be computed and in exchange a qualitative analysis will be presented.

We can notice a slight drop of similarity scores for condition 5. This could be explained because setA in general contained fewer sentences per DSS, and thus fewer training items.

The average similarity score across all conditions is of 97.1%, with 88.57% of perfect matches. This translates into roughly one to three mistakes per condition and per fold. The nature of these mistakes is addressed in the next section, however we can observe that the performance in terms of similarity is very high and almost perfect in several cases.

Qualitative Analysis Although the performance is quite high, the model elicits a number of systematic mistakes that provides us with some insight into the internal mechanism that is implemented by the network. Examples of these are shown in Table 3.

Taking a qualitative look into the produced sentences, it is evident that, with literally a couple of exceptions, all the sentences produced are syntactically correct and semantically felicitous. The vast majority of the elicited mistakes occur when the model produces a sentence that is semantically highly similar to the one expected. This pattern can be seen already during training, where the mistakes correspond to situations that are closely related, so the model is unable to distinguish them, even though it has already seen the situations/sentences (examples 1-3 in Table 3).

For the conditions shown in Table 2, the errors elicited during 5 folds were manually inspected in order to see their regularities. The errors observed (38 in total) can be classified into 2 main categories: the first one (63.2%) being errors concerning over and underspecification, and the second one (31.6%) corresponding to situations that because of the design of the microworld are remarkably similar, differing only in one aspect.

Concerning the first category, the errors can be further split into **location** under- (21.05%) and over- (7.9%) specification (examples 4-5 in Table 3), **subject** under- (15.8%) and over- (10.5%) specification (examples 6-7 in Table 3), and **object** under- (2.6%) and over- (5.2%) specification (examples 8-9 in Table 3).

The errors contained in the second category (examples 10-12 in Table 3) correspond to sentences that at first glance seem correct, it is only after taking a deep look into the microworld that one can see the mistake. First, according to this

Table 3: Examples of representative output errors.

	Output	Expected
1	someone plays with a ball outside .	a girl plays with a ball outside .
2	someone loses in the bedroom .	someone wins in the bedroom .
3	a girl loses to someone in the bedroom .	someone beats a girl at a game in the bedroom .
4	Sophia beats Heidi with ease at hide_and_seek .	Sophia beats Heidi with ease at hide_and_seek in the bedroom .
5	Sophia wins with ease at a game in the street .	Sophia wins with ease at a game outside .
6	a girl plays with a doll inside .	Heidi plays with a doll inside .
7	a game is won with ease by a girl in the bathroom .	a game is won with ease by someone in the bathroom .
8	someone plays .	someone plays with a toy .
9	Charlie plays a game in the street .	Charlie plays in the street .
10	someone wins in the bedroom at hide_and_seek .	someone loses in the bedroom at hide_and_seek .
11	Heidi loses to someone in the bedroom at hide_and_seek .	someone beats Heidi in the bedroom at hide_and_seek .
12	Sophia beats someone at hide_and_seek in the bedroom .	someone loses to Sophia at hide_and_seek in the bedroom .

microworld, whenever there is a winner, there is also a loser, which means that sentences that are apparently contradictory ("someone loses." vs "someone wins.") actually have the same implications within the microworld and therefore are semantically identical. Second, in general whenever there is a winner/loser, the loser is usually situated in the same location as the winner. However, only for the game hide_and_seek and when the participants are inside, the loser can be in the bedroom, while the winner could be in the bathroom, or vice-versa. Finally, whenever there is a prepositional phrase ("in the bedroom"), it is attached to the subject of the sentence according to the grammar, which means that in "Heidi beats Sophie in the bedroom", Heidi is in the bedroom while Sophie could be in either the bedroom or the bathroom, while in "Sophie loses to Heidi in the bedroom", it is Sophie who stays in the bedroom while Heidi could also be in the bathroom. Apart from this detail, the situations are almost identical.

According to the errors so far analyzed, we can conclude that the nature of these is primarily related to situations that are highly similar.

With regard to the test conditions 4 and 5 where a passive sentence is queried but the grammar does not properly define its characterization, Table 4 presents examples of output sentences and the situations that they were supposed to portray. As mentioned before, these situations can be of two types: the first one involving a winning/losing situation where both actors are explicitly mentioned, and the second type being situations where the object of the action is not defined. We took also a closer look into the output of the model for these conditions by manually making an analysis of the output of 3 folds (84 situations), whose results we will now present.

Even though in condition 4 the model has not seen the type sentences that are queried, and that in condition 5 the model has no experience with the queried situations, the sentences produced by the model are mostly correct and coherent with the semantic information that is given to it. One can see that some information is omitted, however, this is expected since the grammar itself does not contain rules that allow to fully encode these situations.

In general, the model learns during training that passive sentences always start by mentioning the object of the action, and that this object is never a person. Therefore, for each situation it tries first to find this object and then it tries to describe the rest of the situation.

Concerning winning/losing situations (examples 1-2 in Table 4), which conform 92.9% of the analyzed situations, the object is always a game because in the microworld winning/losing can only happen while playing a game. Thus, the model produces the specific name of the game when it is known (e.g. "soccer is..."), otherwise the sentence starts with "a game is...". Then one of the players is mentioned (one omitted) and the rest of the situation is portrayed.

For the case of situations with an underspecified object (7.1%, examples 3-4 in Table 4), it is unknown whether the subject is playing a game or with a toy, so the model is forced to choose one. In all cases the model chooses a toy, which seems reasonable because mostly the DSS representations of the underspecified sentences are more similar to situations where a toy is played with. For example, the DSS representation of "someone plays." is more similar to the one of "someone plays with a toy." (99.36% cosine similarity) than to the representation of "someone plays a game." (97.73% cosine similarity).

Similar to the other conditions, errors regarding over and underspecification occur in conditions 4 and 5, but are rare (example 5 in Table 4). And finally one type of error that appears only for these situations corresponds to the interchange of the winner/loser in game situations (example 6 in Table 4).

In sum, one can see from the output that the model is able to take the linguistic elements learned during training in order to characterize situations for which it has no experience, while being as informative as possible. The only difficulty appears to be the distinction of situations that are highly similar. However, the performance of the model is very high in general and even for the sentences that do present a mistake, the output is largely correct.

Discussion

The model learns to generate sentences from rich situation representations, and furthermore, it generalizes to novel sentences and situations.

We can see that the model is able to learn the syntactic

Table 4: Examples of passive output sentences for situations with no passive examples.

	Output	Active Sentence
1	hide_and_seek is won with ease by Heidi in the playground .	Heidi beats Sophia with ease in the playground at hide_and_seek .
2	a game is won with ease by Sophia .	Sophia beats Charlie with ease .
3	a toy is played with .	someone plays .
4	a toy is played with in the playground by Sophia .	Sophia plays in the playground .
5	a game is lost with difficulty by Charlie .	a girl beats Charlie with difficulty in the street .
6	chess is lost by Heidi in the bedroom .	the boy loses to Heidi at chess in the bedroom .

patterns of the microworld and does not just memorize sentences, thus showing *syntactic* generalization. This was observed in all test conditions, where the model was capable of producing sentences that it had never seen before. This means that the model is able to generate new combinations of words in such a way that the new combinations are in order with the syntactic rules of the grammar associated to the microworld, while at the same time being coherent with the semantic structures to which they are related.

Crucially, the model also generalizes *semantically*, as demonstrated in test conditions 3 and 5, since the semantic representations given to the model are completely novel to it, so any correct output can be regarded as arising from the regularities within the microworld from which the DSS representations are derived—cf. the comprehension results by Frank et al. (2009).

We hypothesize that the fact that the model has difficulties with highly similar situations means that the model is able to roughly reconstruct the topography of the microworld semantic space, putting together situations that are semantically related. At the same time, the model assigns linguistic structures to each area in this semantic space such that semantically similar situations are assigned linguistically similar realizations. Given that in practice the semantic space is a continuous 44-dimensional space, in theory the model should be able to generate sentences for unseen areas as long as it is given enough information during training in order to reconstruct the semantic space and the mapping between semantics and linguistic realizations, as proposed by Frank et al. (2009).

The results of the test conditions show that it is indeed the case. Conditions 3 and 5 demonstrated that the model is able to generate sentences for unseen areas in the semantic space, therefore showing semantic systematicity. Conditions 1 and 2 demonstrated that the model is able to generate sentences for semantically known situations but with a different voice (active/passive), showing syntactic systematicity. Conditions 4 and 5, where passive sentences are queried, demonstrated that it is able to produce coherent sentences even if the grammar that was used to construct the training/testing sets does not associate passive constructions with these situations.

As it is now, the semantic space related to the microworld that we use is finite, however we were able to show different levels of systematicity for unknown areas of this space, which was our main objective. In principle, the microworld can be extended by adding elements to the set of basic events, and words to its vocabulary.

Conclusion

We presented a novel connectionist model of sentence production that uses the rich semantic representations described in Frank et al. (2009). This model is able to take such representations and produce sentences that accurately describe the associated situations. The model is also able to produce alternative encodings (e.g. active vs. passive) for a particular semantics, showing an overall high performance and demonstrating that these semantic representations are not only suitable for comprehension, but also for modeling language production. And furthermore, the model is able to generate novel sentences for previously unseen situations, thus demonstrating syntactic and semantic systematicity.

References

Brouwer, H. (2014). *The electrophysiology of language comprehension: A neurocomputational model*. Unpublished doctoral dissertation, University of Groningen.

Chang, F., Dell, G., & Bock, K. (2006). Becoming syntactic. *Psychological review*, *113*(2), 234.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Frank, S. L., Haselager, W. F. G., & van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, *110*(3), 358–379.

Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, *27*(6), 875–910.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, *33*(3), 449–496.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.

Grammatical Bracketing Determines Learning of Non-adjacent Dependencies

Felix Hao Wang (wang970@usc.edu)[a], Jason Zevin (zevin@usc.edu)[a,b,c], Toby Mintz (tmintz@usc.edu)[a,b,c]

[a]Department of Psychology, University of Southern California, 3620 McClintock Ave, Los Angeles, CA, 90089
[b]Department of Linguistics, University of Southern California, Los Angeles, CA, 90089
[c]Program in Neuroscience, University of Southern California, Los Angeles, CA, 90089

Abstract

Grammatical dependencies often involve elements that are not adjacent. However, most experiments in which non-adjacent dependencies are learned bracketed the dependent material with pauses, which is not how dependencies appear in natural language. Here we report successful learning of embedded NAD without pause bracketing. Instead, we induce learners to compute structure in an artificial language by entraining them through processing English sentences. We also found that learning becomes difficult when grammatical entrainment causes learners to compute boundaries that are misaligned with NAD structures. In sum, we demonstrated that grammatical entrainment can induce boundaries that can carry over to reveal structures in novel language materials, and this effect can be used to induce learning of non-adjacent dependencies.

Keywords: non-adjacent dependency learning; grammatical entrainment

Introduction

Due to the hierarchical organization of the syntax of natural languages, lexical items (and morphemes) that are syntactically related are not always linearly adjacent. Thus, to acquire the specifics of the hierarchical grammar, learners must be able to track dependencies of linguistic items that are both linearly adjacent and non-adjacent. For example, given the dependency between the singular subject *child* and the agreeing inflected verb *runs*, the subject and the verb are adjacent in *the child runs*, and non-adjacent in *the child always runs*. Therefore, an important question in language acquisition is how learners acquire adjacent and non-adjacent grammatical dependencies, as they could help learners understand how their language is structured. There has been considerable interest in investigating learning mechanisms that could detect these dependencies in linear sequences within spoken utterances. These mechanisms could be useful for discovering syntactic structures when children start to acquire their first language, and facilitate building syntactic parses later in life. For example, many studies employing artificial and natural languages have investigated how language learners acquire non-adjacent dependencies (e.g., Gómez, 2002; Peña, Bonatti, Nespor & Mehler, 2002; Newport & Aslin, 2004; Romberg & Saffran, 2013), and how early in the acquisition process such dependencies are detected (Santelmann & Jusczyk, 1998; Gómez, 2002).

Whereas most studies on adjacent dependencies have found success in a variety of learning paradigms, the studies on non-adjacent dependency learning to date have only found evidence in limited situations, with some reporting success in learning and others reporting failure. One consistent characteristic of experiments that showed successful learning is that the minimal sequences that contained a dependency were presented as discrete chunks. In other words, the chunks were surrounded by silences, and the edges of such a chunk consisted of the (non-adjacent) dependent elements. For example, studies that have probed non-adjacent dependency learning between words in artificial languages typically have used trigrams in which the dependent words were at the trigram edges, and subjects were presented the trigrams one at a time, with silence intervening between presentations (Gómez, 2002; Romberg & Saffran, 2013). Similarly, in experiments investigating non-adjacent dependencies between syllables in syllable sequences, learning occurred only when brief pauses were introduced before (and after) each syllable trigram (Peña et al.,2002). When syllables were concatenated continuously, participants showed no learning (see also Newport & Aslin, 2004). In the studies just discussed, the fact that subjects' success in learning non-adjacent dependencies was correlated with whether the trigrams containing the dependency were pre-segmented suggests that the chunked presentation might have played an important role in learning. One reason in which pre-segmenting the material in this way could be helpful is that it places one or both dependent elements in an edge position. Indeed, Endress, Nespor & Mehler (2009) argued that edges are privileged in the kind of position-related computations they afford, and placement at edges could be an important constraint for learning non-adjacent dependencies.

But in natural languages, non-adjacent dependencies are not restricted to positions of silences (e.g., utterance boundaries), and are often embedded in longer sequences. Learning the dependency relations of a natural language generally require learning non-adjacent dependencies of items that may not always occur at silence boundaries. This may pose a problem for the learning theories mentioned above. One possibility is that the non-adjacent dependency learning with spoken language critically requires the presence of pauses as a prosodic cue. According to this interpretation (see Peña et al. 2002, for a discussion), successful learning requires an interaction between prosody and syntactic analysis. However, it is also possible that the edges that the learning mechanism requires are broader in scope, including boundaries in representations rather than directly in the signal (pauses being the latter). For example,

edges or boundaries that are represented as the result of a syntactic parse might play a similar role to pauses. Under this interpretation, prosody per se might not be critical for learning non-adjacent dependencies. But rather, the availability of any kind of boundary, in the signal or computed, could facilitate the detection of non-adjacent dependencies when the boundaries are in close alignment with them.

In order to distinguish these hypotheses, we sought a way to induce edges and bracketing via syntactic analysis without resorting to pauses. To this end, we made use of a recently published phenomenon described as grammatical entrainment. When sentences with the same syntactic structures are presented repeatedly in a cyclical pattern, MEG recordings identified cortical regions which track the syntactic structures and compute syntactic boundaries when structures repeat (Ding, Melloni, Zhang, Tian & Poeppel, 2015). One of the proposed mechanisms is that the repeated presentation of the same syntactic structure makes it possible to predict the syntax of the sentence that comes next.

We reasoned that such entrainment of structural processing might carry over to an artificial language that is presented immediately following cyclic presentation of structurally similar English sentences presented as in Ding et al. (2015). If this process induces syntactic edges in a predictable, cyclic way, it could potentially allow the same analysis to be applied and transferred to new linguistic materials that are presented also in the same frequency (words-per-second) and phase as the English material. That is, perhaps repeated structure building of familiar material (i.e., English) can be used to guide upcoming parsing of novel material (i.e., and artificial language). To this end, we placed novel artificial language after repetitions of the same English syntactic structure across a variety of English sentences, matched in word length (4 words). At the very least, we expected that the English sentences would entrain listeners to segment into 4-word sequences, effectively inducing virtual boundaries in the novel artificial language. Specifically, we predicted that when English grammatical dependencies are repeated in phase with the artificial language dependencies, they would facilitate detection of the artificial non-adjacent dependencies (Experiment 1). We further predicted that when the phase relationship between English and artificial language was mismatched, the facilitation effect would disappear (Experiment 2).

Experiment 1

Methods

Participants. Twenty-four undergraduate students at University of Southern California were recruited from Psychology Department subject pool. Half of them participated in each counterbalancing condition. The sample size was based on previous studies (Newport & Aslin 2004).

Stimuli. We recorded speech from a native English speaker and digitized the recording at a rate of 44.1 kHz. We recorded 2 types of words: English words and novel words. For English words, we recorded 5 names (Brian, John, Kate, Nate, Clair), 5 monosyllabic verbs in 3rd person singular form (turns, keeps, puts, lets, has), 5 pronouns (these, those, this, that, it) and 5 adverbs (down, on, up, off, in). For the non-adjacent dependency, we used 12 novel words: 3 at position 1 (pel, tink, blit), 3 at position 2 (swech, voy, rud), 3 at position 3 (dap, tood, wesh) and 3 at position 4 (ghire, jub, tiv).

After all the words were recorded in list intonation in a random order, we spliced the words from the recording. Each word by itself from the recording lasted between 300ms to 737ms, and we used the lengthen function in Praat (Boersma, 2001) to shorten all the words into approximately 250ms. An additional 83ms of silence was added to the end of each word to increase intelligibility and such that when words are concatenated in a continuous stream, they would occur at 3Hz. This was not intended to be a manipulation, and is certainly not the same as the design from Peña, et al., 2002 because the pauses do not pre-segment phrases into four-syllable chunks that line up with the dependency. The pauses are after every word, making it uninformative of the dependency structure.

Design and procedure. There were three blocks of training phase and testing phase, with each testing phase following a training phase. The training materials and testing materials were divided into 3 equal proportions for the 3 blocks.

Training phase. To create the training stream, we mixed English sentences and artificial sentences together. A total of 432 novel word sentences and 858 English sentences were randomly concatenated together in the following fashion. Each English sentence was created by randomly picking a name, a verb, a pronoun, and an adverb, in that order. As such, each sentence consisted of 4 syllables (with the exception of sentences containing the word Brian), and lasted 1.33 seconds. Since words were randomly selected and constrained only by position, there was no statistical dependency on the level of words.

Each novel sentence was a concatenation of 4 novel words, 1 each from choices of 3 for each position, as specified in the Stimuli section. We represent this pattern as YAXB, with one letter for each class of novel word. The second position word predicted the fourth position word (YA_iXB_i), so the fourth word is predictable from the second word. All the other words (at positions 1, 2 and 3) cannot be predicted. Given that there are 3 different words for each non-dependent position (1, 2, & 3), there were 27 possible different quadruplet artificial sentences.

The training stream was made by concatenating alternating English and artificial language sentences in the following way. We concatenated 5, 6 or 7 English sentences together to create the entrainment effect. After the English sentences, three artificial language sentences were presented, followed by more English sentences (see Figure 1 for a demonstration). There are no additional pauses between English sentences, between novel words of artificial

language, or the boundary between English words or novel words.

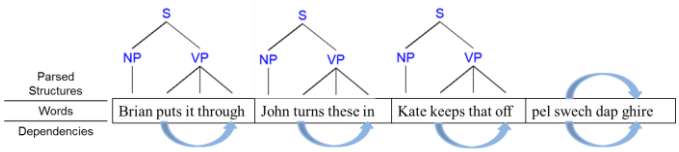


Figure 1. The design of language materials in the training phase of Experiment 1. The English sentence and the artificial dependencies are in phase. If the third sentence is [pel *swech* dap *ghire*] and the following sentence is [blit rud dap tiv], learners may learn the dependency between *swech* and *ghire* (as well as rud & tiv).

A counterbalancing condition was created such that the ungrammatical strings that occurred in the test are grammatical in the training sequence in the counterbalancing condition, similar to Gómez (2002). There were three $A_i_B_i$ frames (i.e., $A_i_B_i$, where i=1-3). The two counterbalancing languages were created by taking three pairs of A_B for the three frames in one training language ($A_1_B_1$, $A_2_B_2$, $A_3_B_3$), and three different pairs for the other training language ($A_1_B_2$, $A_2_B_3$, $A_3_B_1$). This design allowed the use of the same set of test items for the two counterbalancing training languages, where the set included both frames from the two training languages.

Participants listened to the sound stream passively through the headphones while the screen was blank. The listening part lasted about 9 minutes per block.

Test phase. Immediately after each training block, we showed instructions for the test phase on the screen. The instruction made it clear that participants would hear sound sequences and make judgment about the sequences. There were a total of 18 test trials per block, half of which were from the correct dependency, and the other half from incorrect ones. The sequence of presenting the test trials was randomized for each participant.

Participants initiated each test trial. Per trial, we played an artificial language sentence, and asked the participant to indicate whether some sequences are from the previous section that they have heard. A scale showed up after playing the sentence and participants were asked to answer the question "Do you think that you heard this sequence in the previous section?" There were five possible items to choose from, "Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not". Participants could click on any of the choices, and the trial ended and next trial began.

Results

We coded the scale of "Definitely", "Maybe", "Not Sure", "Maybe Not" and "Definitely Not" into numeric values of 1 through 5. To compare ratings statistically, we ran mixed effect linear regressions with the data. In the regression, ratings were the dependent variable, where the interaction and main effects of item type (correct vs. incorrect) and block number (1 through 3) was entered in the fixed effect with subject as the random effect. The interaction was found to be significant (chi2(2) = 8.37, p=0.0152). For this interaction, the regression revealed that the difference score between correct items and incorrect items in block 2 was not significantly different from block 1 (β=0.019, z=0.13, p=0.895) but block 3 is (β=-0.343, z=-2.44, p= 0.015). We conclude that the difference of ratings for correct vs. incorrect items was found to be significant. A plot of the data and comparisons can be found in Figure 2.

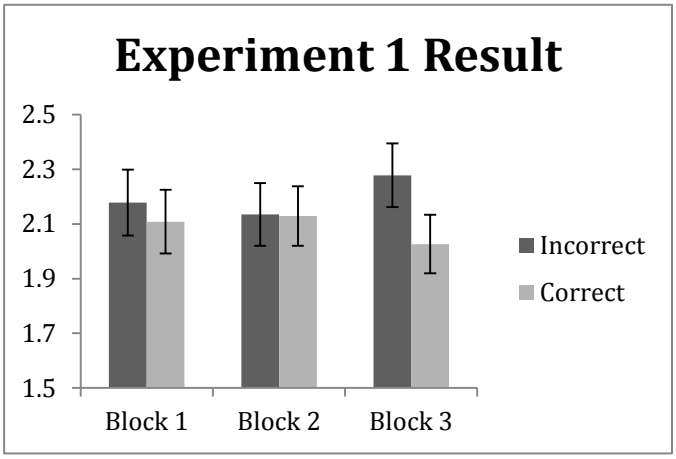


Figure 2. Ratings of test items by block for correct and incorrect items. Smaller rating indicate acceptance of the item. Error bars represent 95% confidence intervals around the mean.

Discussion

In Experiment 1, we introduced a window of analysis from English that participants can use to break into the artificial language. Given the syntactic structures in English, the bracketing of artificial language can be induced such that structural analysis can be applied to artificial language sequences. Applying the boundary every 4 words into the artificial language, we get the chunks [Kate keeps that off] [YA_iXB_i] [YA_jXB_j], which facilitates dependency learning.

It's worth characterizing the way we constructed the training material that produced grammatical entrainment. The English sentences we used did not contain any acoustic information regarding its structure unless the knowledge of English is applied. The construction was such that all the English sentences had constant transitional probability between each word, and the words between each sentences, given that the English sentences were made from a random Markov process. That is, each English word just predicts a word from a set of the next words with equal probability. If no knowledge of English is present, words come randomly without informative landmarks. Without English grammar, listening to the English words would not result in any kind of boundaries that can be used for processing of later

artificial language. The fact that our participants were native speakers of English made grammatical entrainment possible, which facilitated the detection of non-adjacent dependencies.

There are some alternative accounts to our current proposal that listeners transferred syntactic boundaries from English to the artificial language. One possibility is that the length of the English sentences (4 words) encouraged listeners to process the artificial language in smaller chunks, and that was sufficient to ease processing and facilitate detection of non-adjacent dependencies. In other words, the alignment of the chunk boundaries with the non-adjacent patterns may have been irrelevant, and all that was necessary was processing shorter sequences. A related possibility is that presenting the same syntactic structures induced syntactic priming (Bock 1986), whereby the repeated presentation of the same syntactic structure makes it easier to reuse/reactivate structures of the same type. In our case, the fact that we present one syntactic structure (verb + prep. + verb particle) over and over again may have sensitized participants to dependency lengths of three if syntactic priming is operating. Furthermore, given that the boundaries induced from English align with the switch from English to artificial words, there might be boundaries arising from this shift. These boundaries may simply serve as a starting counter for the novel word sequences, from which edge-based computations can be performed. We rule out these alternatives in Experiment 2.

Experiment 2

Methods

Participants. Twenty-four undergraduate students at University of Southern California were recruited from Psychology Department subject pool. Half of them participated in each counterbalancing condition. None of the participants participated in Experiment 1.

Stimuli. We used the same stimuli from Experiment 1.

Design and procedure. Similar to Experiment 1, there were three blocks of training phase and testing phase, with each testing phase following a training phase. The training materials and testing materials were divided into 3 equal proportions for the 3 blocks, and the only difference between Experiment 2 and Experiment 1 is the training materials (described below).

Training phase. The training phase is similar to Experiment 1 except for one key difference in how the English words were ordered. We mixed a total of 432 novel word sentences and 858 English sentences were randomly concatenated together in the same fashion as in Experiment 1, except for the order of the English words. Each English 'sentence' was created by randomly picking a verb, a pronoun, an adverb, and a name, in that order. Given such ordering of English words, parsing of these sentences described above will not start with the beginning to the end of each quadruplets (1-2-3-4); rather, it would restart from the last word of the quadruplet, making use of the name in the current quadruplet, and continuing to the next quadruplet, making a sentence out of the current 4th word and the first 3 from the next (4-1-2-3; see Figure 3 for a demonstration). The purpose of this deliberate phase shift is to change where the syntactic boundaries come in the parsing process, which will become important for the artificial language parsing later. Similar as in Experiment 1, there is no statistical dependency on the level of words, given that any one of the 4 words can appear at its position. Dependency can only be defined grammatically, given listener's knowledge of English and the parsing of the sentences.

The novel sentences were exactly the same as in Experiment 1, and the pattern can be similarly represented as YA_iXB_i. Since the phase of the English sentences was such that the first position and the third position contain the dependency (verb and verb particle), the dependency in the artificial language is out of phase with respect to the English sentences, and instead is aligned with the Y_X structures, which are independent.

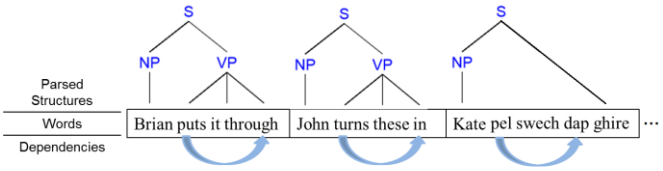


Figure 3. The design of language materials in the training phase of Experiment 2. If the third sentence is [Kate pel *swech* dap] and the following sentence is [*ghire* blit rud dap], learners fail to learn the dependency between *swech* and *ghire*.

Similarly to Experiment 1, a counterbalancing condition was created such that the ungrammatical strings that occurred in the test are grammatical in the training sequence in the counterbalancing condition.

The language stream started with a name, making the first English sentence grammatical. Participants listened to the alternating English and artificial language sentence stream passively through the headphones while the screen was blank.

Test phase. The Test phase is exactly the same as Experiment 1, with 3 blocks, each block containing the same number of correct and incorrect items to be rated.

Results

We coded the scale of "Definitely", "Maybe", "Not Sure", "Maybe Not" and "Definitely Not" into numeric values of 1 through 5. To compare ratings statistically, we ran mixed effect linear regressions with the data. In the regression, ratings were the dependent variable, where the interaction and main effects of item type (correct vs. incorrect) and block number (1 through 3) was entered in the fixed effect with subject as the random intercept. The interaction was

found to be not significant (chi2(2) = 2.50, p= 0.287). For this interaction, the regression revealed that the difference score between correct items and incorrect items in block 2 was not significantly different from block 1 (β=-0.157, z= -1.16, p=0.244), nor is in block 3 (β=0.046, z= 0.34, p= 0.732). Dropping the fixed effect of block, I rerun the regression with only item type as the fixed effect with subjects as random effects. Again, there is no evidence for learning, as correct items were not rated differently from the incorrect items (β=-0.078, z=-1.42, p=0.154). We conclude that there is no significant difference in ratings for correct vs. incorrect items. We plot the data in Figure 4.

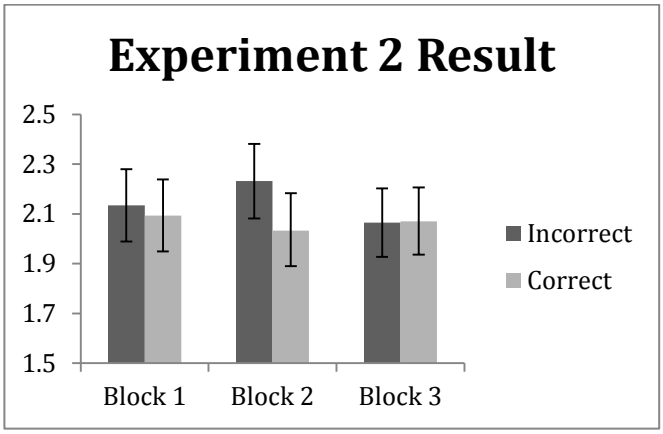


Figure 4. Ratings of test items by block for correct and incorrect items. Smaller rating indicate acceptance of the item. Error bars represent 95% confidence intervals around the mean.

Discussion

In Experiment 2, the phase of the English sentences was mismatched to the phase of the artificial language dependency whereby the non-adjacent dependency was not within each chunk of the segmented material. Participants in Experiment 2 failed to learn the artificial language dependency.

We can eliminate some of the possibilities mentioned in the discussion of Experiment 1. First, similar to Experiment 1, Experiment 2 also used English sentences that were 4 words long. If this helps the artificial language parse into 4-word chunks (as we argue it does, see below), this factor alone is not sufficient to support successful detection. Similarly, a syntactic priming account would not entirely account for the result, given that the same verb to verb particle dependency exists in Experiment 2 as well. We cannot eliminate the potential influence of syntactic priming, given that we presented the same syntactic structure (verb + prep. + verb particle) over and over again, but again, it is not enough to facilitate detection. Lastly, if there were boundaries that were due to changes of English words to artificial words, these boundaries are still present in Experiment 2. Experiment 2 thus suggests that the success of learning in Experiment 1 is not a result merely of computing boundaries in sequences of novel words.

Rather our interpretation is that the English sentences provided edges in specific locations that were sufficiently aligned with the dependencies in the artificial language. If English parsing carried over to parsing the artificial language, even at the coarsest 'sentence' level, the dependencies in Experiment 2 would have been chunked as [Kate YA_iX] [B_iYA_jX]…, and the A_iXB_i dependency would never have been considered, and never be learned as a result. We argue that the reason for failure to learn the dependency is a result of bracketing elements across boundaries. Experiment 2 contrasts with Experiment 1 in that the lack of facilitation effect given the mismatch with phase support the presence of grammatical entrainment, given all other factors (acoustic, transitional probability) are controlled for.

General Discussion

In this paper, we report our first attempt at inducing non-adjacent dependency learning with grammatical bracketing. As we mentioned, non-adjacent dependencies are generally hard to learn for a variety of reasons. For the most part, past literature suggested (Newport & Aslin, 2004; Peña et al., 2002) that pauses are critical to the learning of syllable-level non-adjacent dependencies. Our design does not contain pauses, which makes our study the first we know of that addresses the issue of handling the window of analysis in non-adjacent dependency with language learning without resorting to pauses. We show that this hard problem of learning of non-adjacent dependency can be solved when the non-adjacent dependency is entrained to an English rhythm that provided syntactic edges to the novel artificial language in such a way that the edges line up with the dependencies to be learned. Furthermore, we show that when the syntactic edges slice the novel language into chunks that do not contain the intended dependency, it leads to failure of learning.

Existing theories posits that perceptual or memory primitives guide aspects of statistical based learning, and more specifically, that edge-based computations are critically required for computing non-adjacent dependencies (Peña et al., 2002). However, this theory left the description of edges rather vague. In the studies cited here on non-adjacent dependencies from Endress and colleagues, the implementation of an edge has been a period of silence. As we mentioned, this points to a possibility that the non-adjacent dependency learning mechanisms critically require the presence of pauses as a prosodic cue, whereby an interaction between prosody and syntactic analysis. However, our data suggest that this may not be the case. One of the advantages of an edge-based computation is that it shortens the sentence length that the computations are performed on. This has great implication for how long-distance dependencies can be detected from a computational perspective. Linguistic dependency can exist between any element within a sentence to any other element the number of dependencies grows factorially with the number of elements within a sentence (between-sentence dependencies also exist in language, which is not considered here). If

boundaries are only at utterance boundaries, detection of long-distance dependencies may quickly become intractable as the length of sentence increases (see Wang & Mintz, under review, for a demonstration).

It follows that the learning mechanism for non-adjacent dependencies may require edges. Yet non-adjacent dependencies are not limited to start or end of sentences in natural languages, and as such, the learning mechanism in natural language for effective detection of non-adjacent dependencies may resort to edges beyond utterance boundaries or silences. To solve this problem, we demonstrate that non-adjacent dependencies are learnable when the boundaries in a syntactic sense brackets the continuous speech stream into chunks that contain the dependency. Our study shows that prosody by itself is not critical for learning non-adjacent dependencies. But rather, computing structural boundaries at the beginning and end of chunks that contain the dependency also facilitate their detection. In this sense, the pauses are serving the same function as the syntactic boundaries here, perhaps restricting the window of analysis the detection mechanism operate within. However, characterizing that the learning mechanism requires some kind of prosodic processes such as pauses would be an under-specification.

We can speculate about the role of the English sentences in Experiment 1 further. One possibility is that the English sentence simply sets a pace at every 4 word, bracketing artificial language to 4-word chunks. Potentially, having a chunk close to the size of dependency is enough for learning to happen, though given the low variability in the middle element, learning would occur slowly. Existing theories (Gomez, 2002, among others) suggest that the dependency is hard to detect without highly variable middle elements. In our design, the variability of the middle elements (n=3) is very low according to Gomez 2002, making the dependency hard to learn. Alternatively, processing the English sentences result in a hierarchical parse, such that the verb and the verb particle have a syntactic non-adjacent dependency (e.g., *puts … on*). If this parse transfers to new linguistic material, it would imply that the detection of non-adjacent dependency is facilitated by this narrow window of in-phase pattern matching. Learning is a result of many factors, and future research will separate these possibilities in detail. Although this is an intriguing possibility, from our data we cannot determine if the internal dependency is playing a role, as its alignment to the artificial dependency is confounded with the alignment of the of the sentence boundary and the artificial dependency.

We recognize that various aspects of our design are artificial, especially in terms of the temporal nature of the English and artificial language material. The goal was to use the rhythmic properties as a tool to stimulate syntactic transfer from processing of known structures to novel ones. At one level, this could be viewed as a kind of syntactic bootstrapping, either at a coarse grain level (the sentence level), or fine grain level (phrase level), where prior structures organize the interpretation of novel material.

There are many future directions to this preliminary work. As we just discussed, it is possible that entrainment and transfer involved the internal syntactic dependency. Moreover, it is conceivable the syntactic categories have carried over to the artificial words. There could be multiple reasons this can happen. Given the cyclical nature of the syntactic assignment of English words, the syntactic assignment can potentially carry over. That is to say, different artificial words may become more verb like or pronoun like, depending on its position and the non-adjacent dependency may facilitate syntactic categorization (Mintz, Wang & Li, 2014). Another direction to go is to examine how syntactic bracketing happens in artificial language in general. Does the presence of adjacent and non-adjacent dependencies alone induce bracketing? What kind of mechanism is involved in deducing chunks from statistical information?

In sum, we have argued that correct bracketing is crucial to learn about elements within a chunk. We propose that thinking about bracketing is a useful way of puzzle solving around learning linguistic dependencies, and that having a correct window of analysis is crucial for such purposes.

References

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, 18(3), 355-387.

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* (2001): 341-345.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in cognitive sciences*, 13(8), 348-353.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431-436.

Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. *Cognitive psychology*, 75, 1-27.

Newport, E.L., & Aslin, R.N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-162.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604-607.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive Science*, 37(7), 1290-1320.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, 69(2), 105-134.

Wang, F. H, & Mintz, T. M. (under review). Learning Non-Adjacent Dependencies Embedded in Sentences of an Artificial Language: When Learning Breaks Down.

Natural science: Active learning in dynamic physical microworlds

Neil R. Bramley[1] (neil.bramley@ucl.ac.uk), **Tobias Gerstenberg**[2] (tger@mit.edu)
Joshua B. Tenenbaum[2] (jtb@mit.edu)
[1]Experimental Psychology, UCL, London, WC1H 0DS, UK
[2]Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA

Abstract

In this paper, we bring together research on active learning and intuitive physics to explore how people learn about "microworlds" with continuous spatiotemporal dynamics. Participants interacted with objects in simple two-dimensional worlds governed by a physics simulator, with the goal of identifying latent physical properties such as mass, and forces of attraction or repulsion. We find an advantage for active learners over passive and yoked controls. Active participants spontaneously performed several kinds of "natural experiments" which reveal the objects' properties with varying success. While yoked participants' judgments were affected by the quality of the active participant they observed, they did not share the learning advantage, performing no better than passive controls overall. We discuss possible explanations for the divergence between active and yoked learners, and outline further steps to categorize and explore active learning in the wild.

Keywords: active learning; intuitive physics; causality; probabilistic inference; mental simulation.

The great majority of research on human and machine learning has focused on *passive* situations, where evidence is fixed or preselected. Participants are typically invited to make judgments based on carefully pre-chosen evidence; and machine learning algorithms compete for predictive accuracy on pre-existing datasets. In contrast, Nature's successful learners are necessarily embedded in the world they must learn about and exploit. Thus, it is the norm for human learners to exert some degree of *active* control over the evidence they see. To understand human learning then, one must also understand the myriad decisions about where to attend, and what action to take, that control and manage the flow of incoming evidence. An effective active learner will be able to bootstrap their learning and improve the utility of the information they receive by tailoring it to resolving their subjective uncertainty. On this view, we can think of the little actions in everyday life as small experiments, ranging from the automatic (e.g. cocking one's head to better locate the origin of a sound), to the deliberate (lifting a suitcase to judge its weight; shaking a present to try and guess its contents; holding a pool cue to one eye, or spinning it, to gauge its straightness). A common element in these examples is that they create situations that bring into sharper relief the physical properties of interest.

In this paper we explore this naturalistic type of learning by looking at how people learn about physical laws and properties, such as magnetism and object mass. The structure of the paper is as follows. We first survey the literatures on active learning and intuitive physics, then describe experiments that contrast passive learners with active and yoked learners. Finally, we look closely at the types of actions that active participants performed to reveal the microworlds' hidden physical properties.

Active learning

Human active learning has largely been studied in simple situations where the space of possible actions is limited and the hypothesis space is well-defined. Examples include category rule learning (Gureckis & Markant, 2009) and games like "Guess Who" (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014) and "Battleships" (Markant & Gureckis, 2010). A related line of research has explored active causal learning, where participants can intervene on causal systems (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Lagnado & Sloman, 2004). Since many causal structures cannot be distinguished by co-variational data alone (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), the concept of intervention captures a key aspect of real world active learning that goes beyond simply asking the right questions. The learner's actions can effectively create idealized situations that would rarely happen under normal circumstances, and thus uncover the true causal relationships. However, the "causal systems" explored in these studies are invariably causal Bayesian networks (Pearl, 2000) where time and space are abstracted away, and actions are limited to idealized interventions.

In general, these studies found that people select actions that are more informative than random selection, but that also tend to be more stereotyped and repetitive than those prescribed by models of optimal active selection. Bramley et al. (2015) propose that learners tailor their actions to their own limited learning capacities, testing only a subset of the possible hypotheses at any given time. If bounded learners fail to consider the true hypothesis, they can fail to generate the necessary evidence to support it, and thus perform worse than passive learners (Markant & Gureckis, 2010). This is a common problem for active learning algorithms that do not consider the whole hypothesis space (MacKay, 1992). People are typically found to be bad at balancing the costs of active learning against its benefits, typically oversampling, e.g. taking too many actions when they carry cost (Markant & Gureckis, 2012). In real-time active learning, this might lead participants to allocate too much of their limited attentional resources to controlling rather than learning (Sweller, 1994).

If learners' actions are tailored to their idiosyncratic learning trajectories, the evidence they generate may be less useful for other learners, who are considering different hypotheses while observing the active learners' choices (Markant & Gureckis, 2014). This view is broadly (Lagnado & Sloman, 2004; Sobel & Kushnir, 2006), but not always (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016), supported by experiments that include yoked conditions, where one participant observes the tests performed by another. Intuitively, the

divergence between information that is in principle available, and what participants can actually learn will be much larger in more complex and naturalistic situations, where only a fraction of the total evidence can plausibly be attended to.

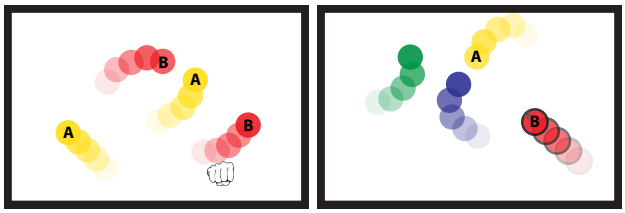
Intuitive physics

Early research into intuitive physics focused on documenting how people's understanding of some aspects of physics, such as ballistic and curvilinear motion, is sometimes systematically biased (e.g. McCloskey, Caramazza, & Green, 1980). More recent research has demonstrated how some of these biases may be explained if we assume that 1) our physical understanding is only approximately Newtonian, and 2) we are often fundamentally uncertain about some important aspects of the physical scene (e.g., the masses of the objects involved in a collision, Sanborn, Mansinghka, & Griffiths, 2013).

Battaglia, Hamrick, and Tenenbaum (2013) have argued that people's understanding of physics is best understood in analogy to a physics engine used to model physically realistic scenes. Accordingly, people have a physics simulator in their mind that they can use to approximately predict what will happen in the future (Smith & Vul, 2013), reason about what happened in the past (Smith & Vul, 2014), or simulate what would have happened if some aspect of the situation had been different (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015). The results of these experiments are consistent with the view that people have a rich intuitive theory of physics that supports approximately accurate mental simulations of key aspects of physical scenes. However, these experiments do not address the question of how we get there – how do people acquire their intuitive physical theories?

Intuitive theories can be expressed as probabilistic programs (Gerstenberg & Tenenbaum, to appear). Program induction is a thorny problem, but one where human-like performance has been demonstrated by sophisticated Bayesian machinery embodying principles of causality and compositionality (Lake, Salakhutdinov, & Tenenbaum, 2015). Ullman, Stuhlmüller, Goodman, and Tenenbaum (2014) explored intuitive physics learning by looking at how people learn about different latent physical properties of 2D "microworlds" similar to the ones shown in Figure 1. The worlds were bounded by solid walls and contained a number of colored pucks with differing weights, surfaces with differing levels of friction, as well as local (magnet-like) forces between pucks and a global (gravity-like) force pulling all the pucks in one direction. In each clip, the pucks bounced around, attracting and repelling each other, being slowed down by the friction, and being pulled by the global force. Participants identified the correct global force around 70% of the time and were much better at detecting local attraction (82%) than repulsion (53%). Ullman et al. argued that repulsion is more difficult to identify because pucks that repelled one another were rarely close enough to exhibit strong repulsion, while attracting pucks rapidly approached one another and stuck together, thus offering stronger evidence of the latent force.

Ullman et al. modeled participants' judgments by assuming a mixture of an Ideal Observer Model (IOM) and a Simulation Based Approximation Model (SBAM). The IOM compares the observed objects' trajectories to simulations of expected trajectories under the different possible worlds. The SBAM compared statistics about each clip such as the pucks' average positions, velocities and pairwise distances, to the summary statistics of repeated simulations under the different possible worlds. For instance, objects in worlds with a global force towards south generally tend to be closer to the southern wall of the world. A mixture model that combined both IOM and SBAM explained participants' judgments well.



(a) "Punch" condition in pilot study. (b) "Grab" condition in experiment.

Figure 1: Schematic displays of two "microworlds". Note: In the pilot there were two pucks of each "type" i.e. two yellow "A"s and two red "B"s while in the experiment there were two target pucks and two distractor pucks and all were randomly colored.

In the current work we build on these results, exploring how people interact with physical microworlds and how this impacts on their learning of the different physical properties.

Pilot study: From Passive to Active

For our pilot study we adapted the setup from Ullman et al. (2014). However, rather than showing participants prechosen replays, we generated the simulations on the fly to allow for active conditions in which participants could exert control over the scene. We chose two setups that differed in the extent to which participants had fine-grained control over the scene. In the *"active punch"* condition, participants controlled a fist that allowed them to roughly knock other objects around, mimicking the clumsy actions of a baby yet to develop fine motor skills. In the *"active grab"* condition, we allowed learners to use the mouse to grab the pucks with the mouse and drag them around, staging more precisely orchestrated interventions.

We were interested in whether *active* participants would be able to use these forms of control to better identify the forces than the *passive* participants; or conversely if the costs of controlling while learning would lead to worse performance. We expected the active learning advantages to be greater in the fine-grained *grab* condition, and the costs to be higher for the *punch* condition where effective control was harder.

Methods

Participants Sixty participants were recruited through Amazon Mechanical Turk (34 male, age 33.5 ± 9.7). They were paid at a rate of $6 per hour.

Materials The experiment was programmed in Javascript using a port of the Box2D physics game engine. The microworlds were displayed in a 600×400 pixel frame, with 1 m in the world corresponding to 100 pixels on the screen.

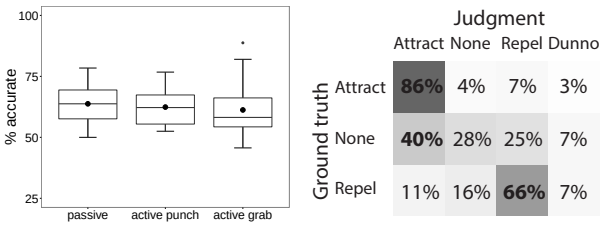
Each world was bounded by solid walls with high elasticity (90% of energy retained per collision) – and contained four pucks (2 yellow, 2 red, each with radius .25 meters, mass 1 kg and elasticity 75%). Each world either had a global force of $1\,\mathrm{m/s^2}$ in one of the four compass directions, or no global force. Each world also had up to three distinct local forces, one between the yellow pucks, one between the red pucks, and one between pucks of differing colors. Each of these could either be attractive ($3\,\mathrm{m/s^2}$), repulsive ($-3\,\mathrm{m/s^2}$), or no force.[1] The pucks' initial positions were random but non-overlapping, with initial velocities in the x and y direction drawn from $\mathrm{Unif}(-10,10)\,\mathrm{m/s}$. Whenever all pucks' velocities fell below $.15\,\mathrm{m/s}$, the simulation froze and the window went black for 500 ms before the positions and velocities of the pucks were redrawn. Each world was simulated for 30 seconds at 60 frames per second.

Conditions Participants were randomly assigned to one of three learning conditions, *passive* ($N = 21$), *active punch* ($N = 20$), *active grab* ($N = 19$, see Figure 1):

1. **Passive** Participants observed the microworlds unfold without being able to interact.
2. **"Active punch"** In addition to the four pucks, this condition featured a "fist" (see Figure 1a). The fist was the same size as the pucks but was heavier (10 kg) and less elastic (50%). The fist was initially located in the middle of the screen but strongly attracted to the location of the participant's mouse.[2]
3. **"Active grab"** In this condition, participants could grab pucks and drag them around with the mouse. Grabbed pucks retained their properties (i.e. mass and local forces and location and momentum) but became strongly attracted to the position of the mouse. When released they would continue on their current trajectory but no longer be attracted to the mouse.

Worlds Each participant either passively observed or actively interacted with 18 microworlds. The set of worlds comprised all combinations of the six possible within-color local force patterns [None-None, Attract-Attract, Repel-Repel, None-Attract, None-Repel, Attract-Repel] and the three possible between-color local forces [None, Attract, Repel]. Half of the microworlds also had a global force in one of the four compass directions. Object colors and direction of the global force were counterbalanced.

Procedure Participants were instructed about the setup of the microworlds, what judgments they had to make, and – if they were in an active condition – how they could interact with the pucks. Participants first saw two practice trials, and then 18 test trials in randomized order.



(a) Boxplots of accuracy by condition. (b) Confusion matrix for local force judgments.

Figure 2: Pilot study performance plots.

On each trial, participants answered 4 questions: One question about the direction of the global force (response options: "North", "East","South","West","None" and "Don't know"), and one question each about the local forces between pucks of the same color (red and yellow), and pucks of different color (response options: "Attract", "None","Repel" and "Don't know"). Participants took on average 22.4 ± 9.3 minutes to complete the experiment.[3]

Results

Participants in the *passive*, *active punch*, and *active grab* condition answered on average 64%, 62% and 61% of the questions correctly. Chance performance was approximately 30%.[4] Thus, judgments were well above chance in all three conditions. However, there was no main effect of condition on performance $F(2,57) = .38, p = .67$. As Figure 2a shows, both the highest and the lowest performing participants were in the *active grab* condition, suggesting that an active learning advantage for this scenario was at least possible although not generally achieved.

On the global force question people were worse at identifying when there was no force, with accuracy of only 42% when the right answer was "none" compared to an average of 85% when the right answer was one of the compass directions. The accuracy difference for identifying "none" vs. one of the other forces interacts with condition $F(4,294) = 2.6, p = .03$, with only 40% in the *active punch* and 30% in *active grab* condition identifying when there was no global force compared to 57% in the *passive* condition. For the local force questions, accuracy differed considerably depending on the ground truth (Figure 2b). Participants in the *active grab* condition were better than *passive* and *active punch* participants at identifying repel forces with an overall accuracy of 77% compared to 68% and 70%.

Due to the simulation restarting whenever all the pucks fell below a certain velocity ($2.0 \pm .8$ times per trial on average), participants in the *passive* condition actually experienced significantly more puck motion than the active participants. We can see this in terms of the total distance traveled by the four

[1] Local forces scaled with the inverse squared distance between the objects in line with Newton's universal law of gravitation. Thus the current local force L exerted on object o_1 by object o_2 (and the reverse) was given by $\frac{3}{d^2}$.

[2] We opted for strong attraction rather than simply copying the position of the mouse because this allowed the controlled object to interact reciprocally with the other objects in collisions rather than behaving as if it was infinitely heavier than the other objects.

[3] Complete specification of the settings of the Box2D simulator and demos of both experiments are available at `ucl.ac.uk/lagnado-lab/el/apl`

[4] Any "don't know" responses were treated as judgments spread evenly across the remaining 3 or 5 options. Random responding would be correct with probability $\approx \frac{1}{4} \times \frac{1}{5} + \frac{3}{4} \times \frac{1}{3} = .3$

colored pucks over the trials of 168 ± 47, 98 ± 55, 85 ± 73 meters for *passive*, *active punch* and *active grab* conditions.

For the participants in the *active grab* condition, more time spent manipulating the pucks was positively related to performance $F(1,17) = 7.2, p = .015$.

Interim Discussion

While participants' overall accuracy was not affected by learning condition, performance in the *active grab* condition was more variable, depending on how much learners controlled the pucks. *Active* learners' lower accuracy on the global force question indicates that controlling individual pucks may have led them to neglect the global properties of the scene (i.e. that the pucks congregated on one side of the world). *Active* learners were more accurate at detecting repulsion, perhaps because they were able to force repellent pucks closer together and so experience stronger evidence for repulsion.

There are several possible explanations for the lack of condition differences in overall accuracy. Restarts of the simulation in the passive condition meant that passive participants naturally experienced more balanced clips with more time during which pucks moved and interacted. Active participants had to put in work to create these experiences (evidenced by the higher variance but lower average puck motion) that were seen "for free" in the passive condition. Another possibility is that the setup of the pilot was not well-suited for active exploration. 30 seconds may have been too little time to allow for sequential, controlled testing, especially of four distinct physical properties. Third, end-of-task feedback suggested that errors were often not due to difficulties in detecting the forces but rather because of having to hold the answers in working memory until the end of the clip; or failure to segment the different worlds in memory, mixing up properties experienced in the current versus previously experienced worlds.

Main Experiment

For our main task we used the same setup as in the pilot but made a number of changes to address the issues identified above. Firstly, we improved the match between passive and active conditions by tweaking the settings of the microworlds so that objects rarely came to rest within the the length of a trial. We increased the elasticity of the pucks from 75% to 98%, leading to restarts occurring only in exceptional situations. Additionally, we replaced the *active punch* condition with a *yoked* condition (cf. Lagnado & Sloman, 2004), in which participants were matched with one of the *active grab* participants and observed their mouse movements and controlling actions. To increase the scope for active hypothesis testing, we increased the length of the trials and asked more difficult test questions (see below).

Because active testing is particularly valuable when competing causal explanations cannot be resolved by observational evidence only, we generated confounded evidence by including two distractor pucks along with two target pucks and drew local forces randomly out of attract/none/repel for all pairs of target and distractor objects. This means that it was more important to isolate the target pucks from the distractor pucks to get clear information about the target pairwise force. Instead of including a global force, which was easily identified by passive learners, we varied the relative mass of the two target objects, a property which is more difficult to infer without experiencing curated comparisons and interactions between them. To reduce memory load, we asked two rather than four questions per trial. To make it clear that each world contained new objects we drew random colors for each object and used new labels, cycling through the alphabet, for the target objects. To ensure that participants were motivated to answer the questions as well as they could, we paid a bonus for each correct response. Finally, to get a more fine-grained measure of participants' judgments, we added confidence sliders for each test question and removed the "don't know" option.

Table 1: Experiment design. *Note*: A = attract, N = none, R = repel; masses are in kg.

World	1	2	3	4	5	6	7	8	9
Target force	A	A	A	N	N	N	R	R	R
Target 1 mass	1	2	1	1	2	1	1	2	1
Target 2 mass	1	1	2	1	1	2	1	1	2

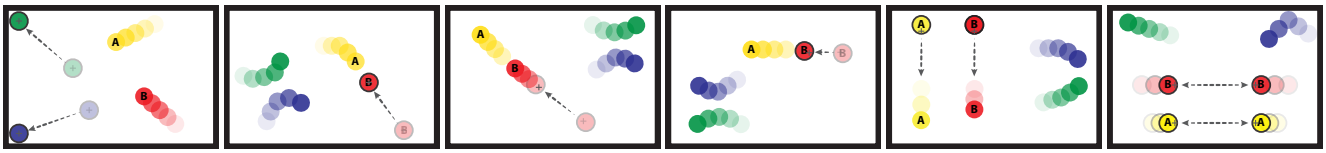
We hypothesized that in these worlds *active* participants would outperform *passive* participants, and that *yoked* participants would inherit some, but not all of this advantage.

Methods

Participants Sixty-four participants were recruited from Amazon Mechanical Turk (39 male, age 33.6 ± 10.2). Participants were paid at a rate of \$6 per hour, plus performance-related bonuses ($\$0.61 \pm .17$) .

Design The first 44 participants were randomly assigned to either the passive ($N = 24$) or the active ($N = 20$) learning condition, and the final 20 participants were yoked 1-to-1 with the active participants. Each participant watched or interacted with 9 microworlds, consisting of all combinations of target force in *attract*, *repel* and *none* and target masses in $[1,1]$kg, $[1,2]$kg and $[2,1]$kg (see Table 1). The five other pairwise forces were drawn uniformly from the three possibilities for each participant on each trial. There were no global forces.

Materials and Procedure We used the same basic set up as in the pilot, but ran the simulations for 45 rather than 30 seconds and increased the elasticity of the pucks from .75 to .98. Rather than two yellow and two red pucks, we drew four random colors for each new world. The two target pucks were labeled with new letters on each trial (e.g. "A" and "B" on trial one, "C" and "D" on trial two, cf. Figure 1b). The distractor pucks were all 1 kg as before but now one of the target pucks could weigh 2 kg. For yoked participants, the cursor of the participant to whom they were yoked (hereafter the *yoker*) was shown with a large "+" symbol whenever it was within the world, and any objects grabbed by the yoker were indicated as in the active condition with a thick black border.



(a) Deconfounding (b) Encroaching (c) Launching (d) Knocking (e) Throwing (f) Shaking

Figure 3: Examples of different interventions participants performed in the *active grab* condition.

Participants first completed instructions relevant to their condition, answered comprehension check questions, and then faced two practice trials followed by the nine test trials. Practice trials were always worlds 1 and 5 (see Table 1). The randomly drawn distractor forces, puck colors and labels differed between the practice and test instances. The two test questions appeared below the world when the time was up. Question order was counterbalanced between participants. At the end of the experiment, participants received feedback about how many of the test questions they got right, and were paid a 5c bonus for each correct answer. The experiment took 19.0 ± 7.3 minutes on average.

Results

Overall accuracy Participants answered 53%, 66% and 54% of questions correctly in the *passive*, *active* and *yoked* conditions respectively (see Figure 4, note chance was $\approx 33\%$ because there were three response options for both questions). Average performance differed significantly by condition $F(2,61) = 3.8, \eta^2 = .12, p = .03$. Post-hoc tests revealed that *active* participants answered significantly more questions correctly than *passive* participants $t(42) = 2.5, p = 0.02$, and (paired) *yoked* participants $t(19) = 2.9, p = 0.02$, with negligible difference between *passive* and *yoked* participants $t(42) = .2, p = 0.83$. Only 4 yoked participants outperformed their active counterparts, with a further 3 answering the same number of questions correctly. Yoked participants' performance was correlated with their active counterparts' $r = .49, p = .03$.

Masses vs. relationships Across conditions, participants were worse at inferring masses than forces $t(63) = -4.8, p < .0001$ and reported lower confidence in mass judgments $66 \pm 25\%$ compared to force judgments $74 \pm 25\%$ $t(63) = -4.2, p < .0001$. Again, participants were less accurate in correctly identifying when there was *no force* between the target pucks (56%) than *repulsion* (70%) or *attraction* (78%), with a main effect of question type $F(2,189) = 7.7, p < .0001$ and significant improvements going from *no force* to *attraction* $t(126) = 3.9, p = .0001$ and *repulsion* $t(126) = 2.4, p = .017$. Force type additionally interacted with condition $F(6,183) = 3.0, p < .0001$. Dummy contrasts with *no force* and *passive* as controls revealed *active* participants were significantly better at identifying *repel* than *passive* participants $t(42) = 3.2, p < .0001$ and a marginal improvement for *yoked* participants as well $t(42) = 1.9, p < .058$. There was no significant relationship between accuracy on the local force question and the number of distractor forces.

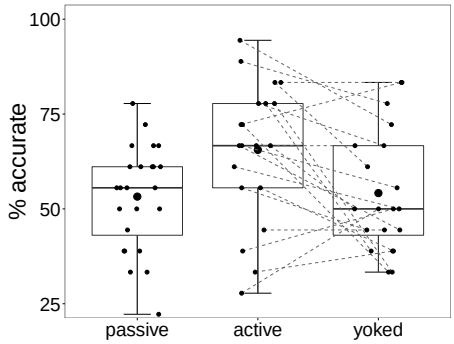


Figure 4: Performance by condition in the main task. *Note*: Large dots indicate condition means. Small dots indicating individual participants in are jittered for visibility. Dotted lines connect active participants with matched yoked participants.

Confidence judgments differed by condition $F(2,61) = 5.3, \eta^2 = .15, p = .007$, with *active* participants significantly more confident on average than *passive* $t(42) = 2.8p = .006$ or *yoked* participants $t(38) = 2.9, p = .006$. Confidence was positively correlated with accuracy $F(1,62) = 10.6, \eta^2 = .15, p = .002$ but did not interact with condition.

Natural experiments Active participants experienced slightly fewer between-puck collisions than passive participants, 59 ± 14 compared to $65 \pm 9, t(42) = 2.0, p = 0.056$. However they experienced significantly more collisions between the two target pucks 15.0 ± 8.1, compared to $9.8 \pm 4.4, t(42) = 2.7, p = 0.01$. 13.2 ± 7.8 of collisions in the active condition took place while one of the two target objects was being controlled by the participant.

Time spent controlling objects was positively related to final performance for active and yoked participants $F(1,38) = 4.8, \eta^2 = 11, p = 0.04$. Therefore, a key question is what kinds of experiments active participants used to find answers to the test questions. Space constraints prohibit a full analysis in the present paper, but we want to share some of the strategies that participants discovered (see Figure 3 and `ucl.ac.uk/lagnado-lab/aplc`):

(a) **Deconfounding** Even though participants mainly manipulated the target pucks, they also sometimes manipulated the distractor pucks. Many of these manipulations involved moving the distractor pucks out of the way and leaving them at rest in a far corner.

(b) **Encroaching** Participants grabbed one target puck and brought it toward the other target puck. This simple strategy allowed participants to infer whether and how the two pucks affected one another. In some cases, participants towed one attracting puck with the other, or pushed a repulsive puck around with the other providing a strong and extended demonstration of the force between the pucks.

(c) **Launching** Participants grabbed one of the target pucks and flicked it against the other target puck. This intervention

helps to figure out whether one of the targets is heavier than the other.

(d) **Knocking** Similar to *launching*, participants grabbed one of the target pucks and knocked it against the other (without letting it go). This intervention also reveals information about the mass of each object.

(e) **Throwing** Participants grabbed a target puck and then threw it, explicitly avoiding collision with any of the other pucks. By exerting an identical force when throwing each target puck, the results of the intervention help to figure out the mass of each object.

(f) **Shaking** Some participants discovered an effective strategy for comparing the mass of the two target objects. By rapidly shaking each in turn (moving the mouse from side to side) it was possible to see that the heavier object reacted more slugglishly. Its greater momentum takes longer to be counteracted by its attraction to the mouse location.

In line with *encroaching* (Figure 3b), we see evidence that participants in the active condition identified the local forces by bringing the two target pucks close to each other. The lower the average distance between two target objects for an *active* participant, the better they did on the force question $\beta = -.3, F(1,18) = 8.0, \eta^2 = .3, p = .001$ but this had no relationship with accuracy on the mass question $p = .87$. Conversely, in line with the *shaking* strategy (Figure 3f), participants who moved the controlled object around faster did better on the mass question $\beta = 25, F(1,18) = 15, \eta^2 = .45, p < 0.001$, but controlled object speed had no relationship with accuracy on the force question $p = .67$. Yoked participants did not inherit these differences, with no significant relationships between performance on either question and average distance between targets or controlled-object speed.

Discussion

We found a clear benefit for active over passive learning in this experiment. In particular, active participants gathered more evidence about repulsive forces by bringing target objects closer together. The quality of the control exerted by the active participants was an important determinant of the quality of the final evidence available to the yoked participants. However, the substantial drop-off from active to yoked accuracy was consistent with the idea that first-hand knowledge of *what* was being tested (e.g. relationship or mass), *when* and *how*, was likely to be crucial for learning successfully. Since there are too many objects and properties in play to track at once, it helps to align the evidence with the hypotheses currently considered. Another factor might have been that active participants were able to look ahead at the crucial locations in the scenes where diagnostic interactions were expected to occur. Yoked participants lacked the ability to foresee what will happen. Finally, active participants had an additional advantage over yoked participants by receiving direct motor feedback about their interventions. They experienced how quickly they moved the mouse or their finger on the trackpad and thus had an immediate sense for how much force they exerted.

Encroaching and *shaking* permitted simple indirect measures, and accordingly, we found shakers doing better on mass questions and encroachers doing better on relationship questions.

While the current study provides a valuable first step, there is much more to explore here. In future work we plan to extend the IOM and SBAM models to active data and use them to evaluate the informativeness of different strategies. We also plan to explore the possibility that learners have a generative grammar for constructing natural experiments; and to unpack yoking differences by looking at yoked participants' ability to infer the learning intentions and action plans of active learners.

Acknowledgments Thanks to Hongyi Zhang for initial code, and Tomer Ullman and David Lagnado for helpful comments. TG and JT were supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Forgetful conservative scholars - how people learn causal structure through interventions. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *41*(3), 708-731.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787).

Gerstenberg, T., & Tenenbaum, J. B. (to appear). Intuitive theories. In M. Waldman (Ed.), *Oxford handbook of causal reasoning*. Oxford University Press.

Gureckis, T. M., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3145–3150).

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*, 856–876.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

MacKay, D. J. (1992). Information-based objective functions for active data selection. *Neural computation*, *4*(4), 590–604.

Markant, D. B., & Gureckis, T. M. (2010). Category learning through active sampling. In *Proceedings of the of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 248–253).

Markant, D. B., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In *Proceedings of the 34th annual conference of the cognitive science society* (pp. 719 – 724).

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1-22.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children?s sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Pearl, J. (2000). *Causality*. New York: Cambridge University Press (2nd edition).

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472).

Sobel, D. M., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, *34*(2), 411–419.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295–312.

Ullman, T., Stuhlmüller, A., Goodman, N., & Tenenbaum, J. (2014). Learning physics from dynamical scenes. In *Proceedings of the 36th Annual Conference of the Cognitive Science society* (pp. 1640–1645).

Inattentional Blindness in a Coupled Perceptual–Cognitive System

Will Bridewell (will.bridewell@nrl.navy.mil)
Paul F. Bello (paul.bello@nrl.navy.mil)
U.S. Naval Research Laboratory
Washington, DC 20375 USA

Abstract

Attention is thought to be a part of a larger cluster of mechanisms that serve to orient a cognitive system, to filter contents with respect to their task relevance, and to devote more computation to certain options than to others. All these activities proceed under the plausible assumption that not all information can be or ought to be processed for a system to satisfice in an ever changing world. In this paper, we describe an attention-centric cognitive system called ARCADIA that demonstrates the orienting, filtering, and resource-skewing functions mentioned above. The demonstration involves maintaining focus on cognitive tasks in a dynamic environment. While ARCADIA carries out a task, limits on its attentional capacity result in "inattentional blindness" under circumstances analogous to those where people fail to perceive otherwise salient stimuli.

Keywords: attention; perception; vision; cognitive model; inattentional blindness

Introduction

Imagine walking across a familiar courtyard on a clear day, your cellphone rings, and you answer. Your friend invites you to a café, and as you set a time to meet, a stranger with a clipboard asks whether you noticed anything out of place. You say, "No," and look around briefly. The stranger asks, more pointedly, if you saw the unicycling clown. You did not, despite the facts that the sun is shining, that you were aware enough to avoid stumbling into anyone, that the clown was wearing bright clothes, and that he unicycled in your general direction. According to Hyman and colleagues (2010), this scenario is common with only 25% of people talking on cell phones noticing the clown compared to over 50% of others strolling through the courtyard witnessing him clearly. Their research was another poignant example of *inattentional blindness* (Mack & Rock, 1998), joining the ranks of "invisible gorilla" studies (Drew, Vo, & Wolfe, 2013; Simons & Chabris, 1999) that demonstrate the human inability to notice otherwise salient stimuli when cognitively preoccupied.

Research on inattentional blindness suggests that people consciously experience only those objects and events that receive attention. Consequently, failing to attend to an object entails failing to explicitly perceive a stimulus and subsequently failing to verbally report on it. Keeping in mind that attention is widely considered to be extremely limited, strategic focus on one aspect of the environment (e.g., a phone conversation) leaves the perceiver open to missing much of what is going on around them (e.g., a unicycling clown). Perhaps surprisingly, inattentional blindness occurs even when preattentively processed, salient, visual features (e.g., contrast in color, orientation, luminance, motion) would pop out during visual search (Yantis & Egeth, 1999). This finding indicates that saliency-based models of visual attention (Borji & Itti, 2013; Itti, Koch, & Niebur, 1998) are incomplete.

In addition to this incompleteness, what does inattentional blindness say about the character of attention and the models that researchers build? Specifically, the phenomenon indicates that for an object representation to enter awareness, to enter working memory, to be reportable, the representation must receive attention. Unsurprisingly, there are disagreements on this point. Wolfe (1999) attributes inattentional blindness to a failure of memory and says that we process and see everything in our visual field but then forget some of it, possibly due to lack of attention. That is, we are aware of everything, but we may not remember or report on the information. In contrast, Mack (2008) says that although we process and encode everything in our visual field, we only explicitly perceive those objects that receive attention. Her story is supported by recent experiments that appear to rule out memory failures (Ward & Scholl, 2015). In either case, this distinction depends on a notion of seeing that invokes conscious awareness, a topic that continues to elude computational modeling.

With consciousness out of the picture, what characteristics of attention should a cognitive model address? Studies on inattentional blindness tell us that

- cognitive tasks can interfere with the perception of visually presented objects,
- explicit object representations that do not receive attention do not persist,
- the properties (shape, color, etc.) of representations that are not persistent cannot be reported,
- visual objects that do not receive attention are still processed and represented.

Most importantly, an accurate model of attention must account for cases where otherwise salient objects go unseen.

Answering these questions requires a modeling approach that can integrate cognition and perception, where attention plays a role in memory, and where there is a distinction between preattentive and attentive information processing. One cognitive system that meets these requirements is ARCADIA.[1] After briefly describing the system, we discuss a visual processing model and highlight how it combines top-down (i.e., cognitive) and bottom-up (i.e., perceptual) factors to construct and remember object representations. We then provide an overview of early inattentional blindness experiments reported by Mack and Rock (1998). Finally, we present

[1] **A**daptive, **R**eflective **C**ognition in an **A**ttention-**D**riven **I**ntegrated **A**rchitecture

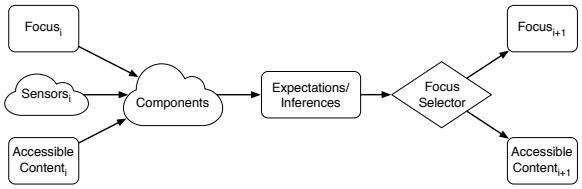


Figure 1: ARCADIA's cognitive cycle.

a model in ARCADIA that carries out a cognitive task and show that it accounts for instances of inattentional blindness by virtue of sharing the four characteristics mentioned above.

ARCADIA's Architectural Features

ARCADIA embodies a theoretical commitment to attention as a central, integrating mechanism for perception, cognition, and action. The system operates in distinct cognitive cycles that are guided by a *focus of attention*. ARCADIA's informational structure is tripartite, consisting of *components*, *accessible content*, and the focus of attention. Note that the system's design makes a principled distinction between aspects of cognition that are and are not able to be introspected or verbally reported. Briefly, *uninspectable* processing occurs in the system's components whereas potentially *inspectable*, reportable representations are housed in its accessible content. Here, we briefly summarize the organizational details of ARCADIA reported by Bridewell and Bello (2015).

The representation of data and information in ARCADIA is multifarious and differs between its uninspectable and inspectable layers. Whereas components may freely use any representations and algorithms that facilitate their operations, effective communication among components necessitates an organizational system so that their results may be located and operated upon by others. Components share their results through interlingua elements that they place in accessible content. The current focus of attention is a privileged element in accessible content.

Information within ARCADIA is processed in the cognitive cycle illustrated by figure 1. On each cycle, the system broadcasts the focus of attention to all the components. To carry out their processing, those components may poll sensors or query accessible content. When finished, the components report interlingua elements to a temporary location where ARCADIA examines them, applies a focus selection strategy to identify the next focus of attention, and populates accessible content. Note that accessible content is flushed and replenished on each cycle, so unless some components act as system memory, information about past states is lost.

ARCADIA's design reflects knowledge about attention in humans, most importantly that it directs perceptual and cognitive resources. Much of the knowledge about attention is derived from errors associated with failures to attend. Consequently, we expect models in ARCADIA to fall prey to similar errors, including inattentional blindness. Further, inasmuch as those models are motivated by theories of attention, perceptual processing, and cognition, we claim that they provide much needed explanatory accounts for those errors. To that end, after providing an overview of ARCADIA's vision model, we introduce experiments on inattentional blindness taken from the literature and describe an extended model that accounts for key phenomena in those experiments.

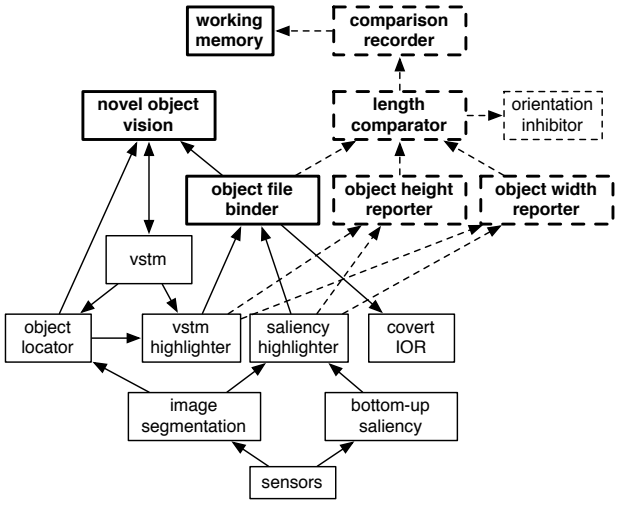


Figure 2: Component interactions in an ARCADIA model of visual processing. The boxes (except for "sensors") indicate the individual components. The arrows illustrate components that use the information output by others. All components operate over their internal state and accessible content, but those in bold font respond to the focus of attention. Dashed lines indicate parts of the model that are specific to the inattentional blindness task.

Vision in ARCADIA

Bridewell and Bello (2015) previously motivated and described a vision pipeline in ARCADIA, here we review that pipeline with a greater emphasis on the details of the components and how they work together to incrementally construct visual objects. Although components in ARCADIA are "always on" and operate in parallel, some components depend on the output of others. Referring to figure 2, the distinction between bottom-up and top-down interactions is roughly shown through the layers of components and the direction of the arrows. For instance, object locator, which tracks recently attended objects, receives bottom-up information from image segmentation and top-down guidance from visual short-term memory (vSTM).

ARCADIA's visual model is designed for object localization, identification, and tracking. To this end, the attentional strategy reflects priorities for constructing objects from visual data and updating their positions over time. ARCADIA differs from most cognitive architectures in that it operates over pixel-level, video data. Further, the system takes multiple cycles to construct object representations from low-level features (e.g., closed contours, color profiles, "retinal" location). Once constructed, the representations must be tagged as referring either to new object instances or to previously

seen objects. Because objects take multiple cycles to encode, interruptions in the normal object-construction process may result in errors that are characteristic of human perception.

As shown in figure 2, vision in ARCADIA involves a variety of components. Space precludes discussing the computational details of each one, so in this paper, we highlight each component's role and the information it communicates to others. Beginning at the base of the figure, ARCADIA incorporates sensors. For the purposes of this paper, there is one simulated camera that operates over a video file. This camera reports rectangular video frames and roughly corresponds to the system's view of a computer screen. Since the modeled experiments assume that gaze rests at a central fixation point, only covert visual attention (i.e., without saccades) is used.

The sensor feeds into visual components that process their input without any top-down influence. The first of these implements the *bottom-up saliency* calculations described by Itti, Koch, and Niebur (1998) to determine where visual attention will shift in the absence of any other information. The second of the components carries out *image segmentation* and collects information about the color and location of closed-contour regions that might correspond to objects. The resulting regions are treated as proto-objects (Rensink, 2000) that may become full-fledged objects with attention.

The next set of components operates over covert visual attention. The *saliency highlighter* reads the interlingua elements created by bottom-up saliency and image segmentation to pick out a few (here, two) proto-objects that appear at points of high saliency (Xu & Chun, 2009). If such a proto-object exists, the component produces an interlingua element representing a candidate location for visual attention. Similarly, the *vSTM highlighter* requests visual attention at the location of any object currently stored in vSTM. Working in tandem with these components, *covert inhibition of return* (McDonald, Hickey, Green, & Whitman, 2009) discourages ARCADIA from attending to only one location so that it can assemble information about multiple objects within the scene.

To this point, the components have read from sensors or accessible content on each cycle, ignoring the information in the focus of attention. Moving upwards in figure 2, if the output of one of the highlighters becomes the focus of attention, the *object file binder* responds. This component takes the proto-object at the attended location and constructs an object-file representation for storage in vSTM.

Once an object-file exists, ARCADIA can determine whether it corresponds to a new object or one that it has already seen. Currently, this determination is made only for representations in vSTM. If the object file receives focus, *novel object vision* checks whether it is identical to elements tracked by vSTM. The establishment of object identity is a complex topic and depends on the task at hand and the properties of the objects themselves. ARCADIA can examine a variety of features to distinguish objects including size and color, but in this model, only location is used. Specifically, an object is treated as an update to a memory representation if its location overlaps the most recent position of the remembered object. Novel object vision generates either a statement of equality between the focal object and one in vSTM or an assertion of a new object.

The *vSTM* component searches accessible content for these statements of equality or newness and updates its memory stores accordingly. This component holds up to four object files at a time, updating them when possible, and replacing an arbitrarily selected old object file with any new one that receives focus. Because ARCADIA components encapsulate information, vSTM reports interlingua representations of its contents on each cycle. Consumers include novel object vision, which uses the elements for comparison purposes, vSTM highlighter, which requests covert orientation, and *object locator*. The design of object locator is based on *fingers of instantiation* (see, Scholl & Pylyshyn, 1999) and is not discussed here because the objects are static. Details are provided by Bello and colleagues (Bello, Bridewell, & Wasylyshyn, 2016).

This walkthrough of ARCADIA's model of visual processing describes the components and their informational exchange but has not discussed attentional strategies, which are task specific. With that caveat, we outline a limited attentional strategy for visual perception.

1. If an object file was just created, focus on it to enable comparison to and/or encoding in vSTM.
2. Else, if a highlighter directs covert visual attention to an uninhibited location, attend to the specified proto-object.
3. Else, attend to an arbitrarily selected interlingua element.

For static scenes, this strategy is insufficient for viewing more than two or three objects because saliency is the principle means for capturing visual attention. Introducing overt visual attention, a memory for previously viewed locations, or applying the strategy to dynamic scenes can reduce the tendency to stare at one or two objects. But, under the task constraints for the experiments we are modeling, even this basic strategy (with a couple task-related augmentations) works surprisingly well. The next section describes experiments on inattentional blindness in humans, after which we introduce a handful of new components and a corresponding attentional strategy to guide focus selection.

Experiments on Inattentional Blindness

Perhaps the most systematic examination of inattentional blindness was reported by Mack and Rock (1998) in their book-length treatment of the phenomenon. Over the course of several years, the authors collected thousands of data that supported the statement, "There is no explicit perception without attention." The studies were carried out across two laboratories, and although the methodology differed according to the hypothesis, the experimental protocol for stimulus presentation was generally consistent.

Most experiments consisted of nine trials per subject. Each trial followed roughly the same structure, although several variations were explored. For brevity, we describe the ex-

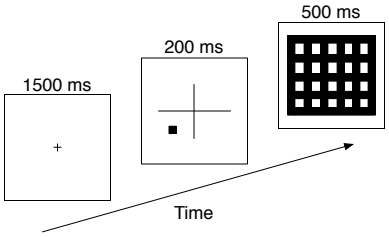


Figure 3: The time course of a single trial in the Mack and Rock (1998) studies.

perimental conditions for the specific study of interest. In this study, subjects were given the task of determining which of the two lines of a cross presented at fixation were the longest. On a standard trial the time course followed that shown in figure 3, a fixation point appeared for 1500ms, followed immediately by a stimulus presented at 200ms, and then a mask meant to overwrite iconic memory was shown for 500ms. After this presentation, the subjects were asked to report whether the horizontal or vertical line was longer.

During the first two trials, only the cross appeared. On the third trial, the cross was joined by a *critical stimulus* unrelated to the task. In this case, the critical stimulus was a black square that appeared in one of the four quadrants defined by the cross and within the rectangle implicitly defined by the length and the width of the cross. After this trial, the subjects were asked if they saw anything out of the ordinary. Three similar trials followed where the subject was asked to perform the original task while also reporting seeing any other objects. In the last set of three trials, the subjects were instructed only to report objects apart from the cross.

In this design, the critical stimulus was presented on the third, sixth, and ninth trial. During the first three trials, the subjects were considered to be naive to the presence of the critical stimulus. During the second three trials, the subjects were considered to have *divided attention*, and on the last three trials they were assumed to dedicate *full attention* to detecting the critical stimulus. Mack and Rock (1998) found that under various conditions roughly 25% of the people failed to see the critical stimulus on the third trial although almost 100% saw it during the ninth (full attention) trial.

The ability to account for perceptual differences across the various trials would be a notable success for ARCADIA's visual processing model. In particular, when the critical stimulus is present, the system should be able to create object files for the cross and the square, record those representations in vSTM, carry out the length comparison, and store the results in working memory. Because the task stimulus appears for only 200 ms, this gives the system few cycles in which to carry out this activity. Although we make no strong claims about the relationship between cycles within ARCADIA and physiological measures, we note that visual attention is linked to gamma band synchronization in the 40 to 80 Hz range (Fries, 2009). Taking the slow end of that range as the point where attended representations may shift, this would give the system 8 cycles to view and process the stimulus.

Modeling Inattentional Blindness

In the introduction, we pointed out four characteristics of attention that studies on inattentional blindness reveal. In this section, we report a model of attention that exhibits all four.

1. Cognitive tasks interfere with perception.
2. Attention is required for object representations to persist.
3. Object properties that do not persist cannot be reported.
4. Visual perceptions are processed without attention.

Furthermore, because the model receives video input that matches a rate of one frame per 25 ms, it accounts for these characteristics within a strict time limit. The task-relevant stimulus appears for only eight frames, giving ARCADIA eight cycles to orient to the visual objects, construct object files, carry out the comparison task, and memorize the result.

For this model, we emphasize those trials from the Mack and Rock experiments where the critical stimulus appears. Specifically, we will offer an explanation for why people (a) are not inattentionally blind in the full-attention condition, (b) are generally not inattentionally blind in the divided-attention condition, and (c) may or may not be inattentionally blind during the critical trial. We do not model the trials without the critical stimulus because we are not yet accounting for the process that lulls subjects into inattention.

Modeling the Mack and Rock experiments requires implementing the components with dashed borders in figure 2 and extending the attentional strategy accordingly. The changes include components for comparing object dimensions and a means for introducing a task-related inhibition of attention. In the rest of this section, we describe the new components, discuss the attentional strategy, and report the resulting behavior. We then address how the model exhibits the characteristics of inattentional blindness.

Additional Components

We developed five new components to carry out the task of comparing line lengths and remembering the result. Referring to figure 2, the *object height reporter* and *object width reporter* respond to the output of the highlighting components. In parallel with the construction of the object file representation, these components represent height and width in a comparable format. If the object file receives focus, then the *length comparator* carries out the straightforward comparison between width and height and reports its result. To ensure that the lengths and widths of other objects are not overtly compared (since that was not part of the task instructions), length comparator only operates on objects larger than the fixation mark. If a comparison was made and receives focus, then the *comparison recorder* requests that the result be memorized.

Working memory responds to attended requests for memorization, storing the associated representation for later use. Note that elements in working memory are not processed in this model, and the current implementation of working memory is little more than a storehouse. Elements in working memory are reported to accessible content on each cycle, as

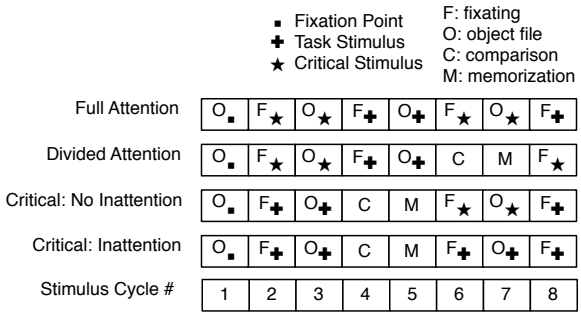


Figure 4: Focus traces for the four attentional conditions.

with vSTM, but in the current model, no other components process them.

The final task-specific component is *orientation inhibitor*. This component plays a central role because it provides task-based limits to the parts of the visual field that receive attention. On each cycle, orientation inhibitor reports a request to inhibit non-central regions of the visual field. This reflects the expectations that the cross will appear at fixation and that nothing else is important to the comparison task. There are three settings to this component.

1. Off: no inhibition signal is sent. This is the assumed operative mode for divided attention and full attention trials.
2. Slow: an inhibition signal is sent on each cycle and the component takes several cycles to revoke inhibition after task completion. This is the assumed operative mode for the critical trial when there is inattentional blindness.
3. Fast: an inhibition signal is sent on each cycle and the component revokes inhibition immediately after task completion (i.e., after a comparison element receives focus). This is the assumed operative mode for the critical trial when there is no inattentional blindness.

Note that if inhibition is revoked quickly enough, the model cannot distinguish between the Fast and Off conditions for the critical trial when there is no inattentional blindness. However, this condition shows that even if task-based inhibition is active, there is time to perceive the critical stimulus.

Attentional Strategy

The attentional strategy for this task builds on the one provided earlier in the paper. Two new kinds of elements take priority over the other preferences in that strategy. They are (1) if there is an element that encodes a comparison result, make it the focus so that it can be memorized; and (2) if there is a request to memorize some information, focus on it in order to initate memorization. To review the rest of the strategy, if there are no comparisons or memorization requests, the strategy selects, in this order of preference, (3) a newly available object file, (4) a location where it should shift its covert attention, (5) an arbitrarily selected interlingua element. These changes allow the system to read in a video that represents a trial and output a comparison result and the objects that were perceived.

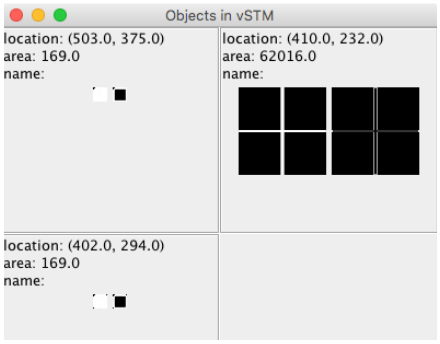


Figure 5: The resulting objects in vSTM for the divided attention trial.

Results and General Discussion

There are four conditions of interest for a model of inattentional blindness: (1) the full attention trial without a task, (2) the divided attention trial with a comparison task and object detection, (3) the critical trial where the subject is not inattentionally blind, and (4) the critical trial where the subject is inattentionally blind. Figure 4 shows the results from running ARCADIA under these conditions. Each row in the figure depicts a separate condition, starting with the cycle where the task stimulus first appears. In all the trials, the system is in the middle of constructing an object for the fixation point, although it could just as well have been shifting its covert attention to a new location (the timing in this case would not alter the qualitative results). Simultaneously during cycle 1, the image segmentation component is producing elements that will be used by the highlighters.

In cycles 2 and 3, attention in the trials without task-based inhibition is drawn to the critical stimulus, whereas attention in the other trials selects the cross. In cycles 4 and 5 of the full-attention and divided-attention trials, ARCADIA processes the cross. The system's attention in the full-attention trial then oscillates between the critical stimulus and the cross due to covert inhibition of return. Attention in the divided-attention trial moves on in cycles 6 and 7 to compare the cross dimensions and memorize the result. This condition shows that without inhibition it is possible to carry out the task and perceive the entire visual scene within 200 ms. Figure 5 shows that after the divided attention trial, vSTM contains the fixation point, the cross, and the critical stimulus.

Turning to the critical trial conditions, task-based inhibition directs ARCADIA's attention to the central cross when it appears and completes the task. If inhibition of the peripheral regions is lifted in time, the system can then process the critical stimulus. If not, then the cross continues to receive attention. These results show that even if the system is directing its attention away from the critical stimulus, there is time to complete the task and encode all the visual objects. However, inhibition may take hundreds of milliseconds to revoke, blocking the encoding of the critical stimulus and, therefore, its explicit perception.

From the outset, we stated that a model of attention must exhibit four characteristics that arise in studies of inattentional blindness. First, the different focus traces for each condition show that cognitive tasks alter the time course of perception in ARCADIA. Second, without attention, objects fail to persist as attention is required for the construction of object files and for recording task results. Third, object properties that do not persist cannot be reported. In this case, the relationship between the cross's height and width is not remembered in the full attention trial and is subsequently not reportable. Fourth, although not all visual objects are explicitly perceived, they all undergo preattentive processing. Furthermore, this model provides an account for why otherwise salient objects may go unseen.

Concluding Remarks

To be sure, researchers working with cognitive architectures have addressed certain aspects of attention, perception, and cognition, but their intersection is rarely explored. On one side of a coin, the majority of computational work on vision and especially on object recognition is divorced from the psychological literature and interactions with cognition. On the other side of the coin, psychologically plausible approaches to perception in some of the better-known cognitive architectures such as ACT-R (Nyamsuren & Taatgen, 2013) and EPIC (Kieras, 2010) rely on symbolic encodings of objects, relations, and events. As a result, there seems to be an empty space between these two families of research approaches that an architecture like ARCADIA can fill: a space for systems that operate over real sensor data, but do so in psychologically plausible ways.

In this paper, we demonstrated that emphasizing task completion produces a sort of tunnel vision that blocks out otherwise salient stimuli. We suggest that this finding applies to any system that can control its focus of attention. Part of what it means to attend to one thing is to ignore everything else.

Acknowledgments

The authors would like to acknowledge generous support from the Office of Naval Research under grants N0001414WX20179 and N0001415WX01339. The views expressed in this paper are solely the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Bello, P. F., Bridewell, W., & Wasylyshyn, C. (2016). Attentive and pre-attentive processes in multiple object tracking: a computational investigation. In *Proceedings of the thirty-eighth annual meeting of the cognitive science society.* Philadelphia, PA.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207.

Bridewell, W., & Bello, P. F. (2015). Incremental object perception in an attention-driven cognitive architecture. In *Proceedings of the thirty-seventh annual meeting of the cognitive science society* (p. 279-284). Pasadena, CA.

Drew, T., Vo, M. L.-H., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, *24*, 1848–1853.

Fries, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annual Review of Neuroscience*, *32*, 209–224.

Hyman, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, *24*, 597–607.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.

Kieras, D. (2010). Modeling visual search of displays of many objects: the role of differential acuity and fixation memory. In *Proceedings of the tenth international conference on cognitive modeling* (pp. 127–132). Philadelphia, PA.

Mack, A. (2008). Inattentional blindness: Reply to commentaries. *Psyche*, *7*, 1–6.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

McDonald, J. J., Hickey, C., Green, J. J., & Whitman, J. C. (2009). Inhibition of return in the covert deployment of attention: evidence from human electrophysiology. *Journal of Cognitive Neuroscience*, *21*, 725–733.

Nyamsuren, E., & Taatgen, N. A. (2013). Pre-attentive and attentive vision module. *Cognitive Systems Research*, *24*, 62–71.

Rensink, R. A. (2000). Seeing, sensing, and scrutinizing. *Vision Research*, *40*, 1469–1487.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38*, 259–290.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059–1074.

Ward, E. J., & Scholl, B. J. (2015). Inattentional blindness reflects limitations on perception, not memory: Evidence from repeated failures of awareness. *Psychonomic Bulletin & Review*, *22*, 722–727.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories: cognition of brief visual stimuli.* (pp. 71–94).

Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences*, *13*, 167–174.

Yantis, S., & Egeth, H. (1999). On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 661–676.

Recursive belief manipulation and second-order false-beliefs

Torben Braüner (torben@ruc.dk), Patrick Blackburn (patrickb@ruc.dk), and Irina Polyanskaya (irinap@ruc.dk)
Roskilde University, Denmark

Abstract

The literature on first-order false-belief is extensive, but less is known about the second-order case. The attainment of second-order false-belief mastery seems to mark a cognitively significant stage, but what is its status? Is it an example of *complexity only* development, or does it indicate that a more fundamental *conceptual change* has taken place? In this paper we extend Braüner's hybrid-logical analysis of first-order false-belief tasks (Braüner, 2014, 2015) to the second-order case, and argue that our analysis supports a version of the conceptual change position.

Keywords: Second-order false-belief tasks; hybrid logic; natural deduction; complexity only; conceptual change; belief formation; belief manipulation; recursion

Introduction

First-order false-belief tasks are a widely studied family of reasoning tasks in cognitive and developmental psychology. A well known example is the *Sally-Anne task*:

> *A child is shown a scene with two doll protagonists, Sally and Anne, having respectively a basket and a box. Sally first places a marble into her basket. Then Sally leaves the scene, and in her absence, Anne moves the marble and puts it in her box. Then Sally returns, and the child is asked: "Where will Sally look for her marble?"*

Children above the age of four typically handle this task correctly: they say that Sally will look in the basket, which is where Sally (falsely) believes the marble to be. Younger children, on the other hand, say that Sally will look in the box: this is indeed where the marble is, but this information is not available to Sally and hence the response is incorrect. For children with *Autism Spectrum Disorder (ASD)*, the shift to correct responses usually occurs at a later age.

The attainment of first-order false-belief mastery is a milestone in the acquisition of *Theory of Mind (ToM)*, the capacity to ascribe mental states such as beliefs to oneself and others, and some researchers account for ASD using some form of a *ToM deficit hypothesis*; see (Baron-Cohen, 1995). Many first-order false-belief tasks have been devised, and over the past 35 years both correlational and training studies (involving both typically developing and children with ASD) have yielded robust results (concerning, for example, the link between ToM and linguistic abilities; see the meta-study (Milligan, Astington, & Dack, 2007)).

Second-order false-belief tasks are less well studied. Consider the following version of the *second-order Sally-Anne task* (the bold font highlights the new text added to the first-order version just given; the bracketed [Sally will] marks the shift in word order from 'will Sally'):

> *A child is shown a scene with two doll protagonists, Sally and Anne, having respectively a basket and a box. Sally first places a marble into her basket. Then Sally leaves the scene, and in her absence, Anne moves the marble and puts it in her box.* ***However, although Anne does not realise this, Sally is peeking through the window and sees what Anne is doing****. Then Sally returns, and the child is asked: "Where* ***does Anne think that*** *[Sally will] look for her marble?"*

Again there is a transition age. Children above the age of six typically handle this task correctly: they respond that Anne thinks that Sally will look in the basket, which is where Anne (falsely) believes that Sally believes the marble to be. Younger children respond that Anne thinks that Sally will look in the box: this is where Sally knows the marble to be, but Anne does not know that Sally knows this and hence the response is incorrect. For children with ASD, the shift tends to occur at a later age, if indeed it happens at all.

The mastery of second-order false-beliefs is another key step in the acquisition of ToM, but less is known about it and many conclusions are tentative (Miller, 2009, 2012). The direction of the causal relation between second-order false-belief task competence and linguistic competence remains unclear, and though there have been second-order correlation studies on children with ASD, we know of no linguistic training studies on this population investigating the link between language and second-order false beliefs. Finally — the issue we address — there is no consensus on the status of the shift from first-order to second-order mastery. Starting with (Sullivan, Zaitchik, & Tager-Flusberg, 1994), some researchers have viewed it as a reasonably straightforward addition to first-order mastery: the acquisition of second-order mastery occurs when the child has sufficiently strengthened his or her information processing capacities, such as working memory and sequencing; following (Miller, 2009, 2012) we call this the *complexity only* position. Other researchers, starting with (Perner & Wimmer, 1985), have argued that the transition marks a more fundamental cognitive shift; again we follow Miller and call this the *conceptual change* position.

The bulk of our paper is theoretical. We extend Braüner's (Braüner, 2014, 2015) hybrid-logical analysis of first-order false-belief tasks to the second-order case, and argue that our analysis lends weight to a version of the conceptual change position. But the backdrop to our theoretical work is an ongoing training study on Danish speaking children with ASD which investigates whether training in linguistic recursion can lead to improvement in second-order false-belief mastery. We briefly link the theoretical analysis with our experimental work in the concluding discussion.

Logical analyses of false-belief tasks

There have been few previous applications of logical methods to false-belief tasks. The pioneering work is due to

Stenning and Van Lambalgen, who analyse the Sally-Anne and other first-order false-belief tasks using non-monotonic closed world reasoning (Stenning & van Lambalgen, 2008). The first-order Sally-Anne task has also been formalized using an interactive theorem prover for a many-sorted first-order modal logic (Arkoudas & Bringsjord, 2008). Applications of logical models to second-order false-belief tasks are even rarer: the clearest example is the use of Dynamic Epistemic Logic in (Bolander, 2014), though the use of game theory in (Szymanik, Meijering, & Verbrugge, 2013) to investigate performance in higher-order social reasoning is also relevant.

Our approach differs from these papers in a key respect: it takes the notion of *perspective shift* as fundamental. The work just cited formalizes false-belief reasoning from a global perspective; the hybrid logical approach of (Braüner, 2014, 2015) formalizes the local shifts of perspective required by the experimental subject when reasoning about the agents in the scenario (that is, Sally and Anne). The intuition is this. Correctly handling the first-order Sally-Anne task seems to involve taking the perspective of another agent, namely Sally, and reasoning about what she believes. So to speak, you have to put yourself in Sally's shoes.[1] In the following two sections we explain why *natural deduction* in *hybrid modal logic* enables us to formalize the reasoning underlying the first-order Sally-Anne task perspectivally.

Hybrid modal logic

Modal logics are a family of logics in which the truth-value of logical formulas is evaluated relative to models consisting of propositional information distributed over a collection of abstract points; in Braüner's approach to false-belief tasks, these abstract points are taken to be people, and the information distributed over them is their beliefs, what they see, and so on. Modal formulas are evaluated locally: truth-value evaluation begins from the perspective of one particular person and then goes on to take into account other people's perspectives.

Modal logic makes use of various (application dependent) modal operators; here we will use such operators as B (to express that someone believes something) and S (to express that someone sees something). For example, $B\phi$ expresses that someone believes the information ϕ (whatever that might be). Modal operators can be applied recursively: $BB\phi$ expresses that someone believes the information that $B\phi$, that is, someone believes that someone believes ϕ.

Hybrid modal logics are extended modal logics in which it is possible to refer to individual points. Hybrid logic is crucial to this paper as it gives us the tools needed to refer to Sally and Anne, and to model their reasoning. Technically, hybrid logic is built by extending ordinary modal logic with *nominals*, which are special propositional symbols interpreted to be true at exactly one point (so here, a person). Nominals should be thought of as naming the unique person they are true at. For example, we shall use s as a nominal true at Sally; in effect it is a 'name' or 'constant' that picks out Sally.

One other piece of hybrid logic machinery will be important: *satisfaction operators*. If ϕ is an arbitrary formula and s is the nominal that names Sally, then a new formula $@_s\phi$ can be built. The $@_s$ prefix is called a satisfaction operator, and the formula $@_s\phi$ is called a *satisfaction statement*. The satisfaction statement $@_s\phi$ says that the formula ϕ is true at a particular person, namely Sally, as she is the person s refers to. We call the nominal (here s) used in satisfaction statements the *point of view* nominal.

Let's make this just a little more concrete. Prefixing the belief formula $B\phi$ with the satisfaction operator $@_s$ yields the formula $@_sB\phi$ which says that Sally believes the information ϕ. Satisfaction statements of this form will play an important role in our analysis.

These informal examples should help guide the reader in what follows. For more on hybrid logic see (Blackburn, 2000) and (Braüner, 2011).

Figure 1: Hybrid-logical rules

$$\frac{c \qquad \phi}{@_c\phi}\,(@I) \qquad\qquad \frac{c \qquad @_c\phi}{\phi}\,(@E)$$

$$\frac{\phi_1 \;\ldots\; \phi_n \qquad\qquad \begin{array}{c}[\phi_1]\ldots[\phi_n][c]\\ \vdots\\ \psi\end{array}}{\psi}\,(Term)^{*} \qquad\qquad \frac{\begin{array}{c}[c]\\ \vdots\\ \psi\end{array}}{\psi}\,(Name)^{\dagger}$$

$*$ $\phi_1, \ldots, \phi_n$, and ψ are all satisfaction statements and there are no undischarged assumptions in the derivation of ψ besides the specified occurrences of $\phi_1, \ldots, \phi_n$, and c.
† The nominal c does not occur in ψ or in any undischarged assumptions other than the specified occurrences of c.

Seligman's natural deduction system

Logicians have developed many (very different) methods of formal proof; in this paper we make us of the one called *natural deduction*. Natural deduction was originally developed to model the structure of mathematical argumentation, but there is now some experimental backing for the claim that natural deduction is a mechanism underlying human deductive reasoning more generally; see (Rips, 2008). Be that as it may, we shall argue that the change in proof structure required as we move from modelling the first-order Sally-Anne task to modelling the second-order version supports the view that a conceptual change takes place in the developmental shift from first- to second-order false-belief mastery.[2]

[1] This might suggest that we are adopting the simulation-theory view of ToM, cf. (Gordon, 2009). But we are agnostic here: 'perspective taking' may seem reminiscent of 'simulation', but their relationship (if any) is unclear. Our talk of putting ourselves in someone else's shoes should be read in a pre-theoretic sense: it expresses an intuition that we wish to formally model in hybrid logic.

[2] So, broadly speaking, our work falls into the "mental logic" approach in the psychology of reasoning. The major alternative is the "mental models" approach, which views the construction of models as the mechanism underlying human reasoning. We earlier

Two main ideas drive natural deduction. The first is that there are two different kinds of rule for each logical connective: one to introduce it, the other to eliminate it. The second, crucial to this paper, is that *hypothetical reasoning* (or *conditional reasoning*) is hardwired into the very core of natural deduction: we can make an assumption, work out its consequences, and then discharge it (get rid of it).[3] We use the natural deduction system for hybrid logic obtained by extending the standard natural deduction system for classical propositional logic with the rules in Figure 1; the symbol c in these rules is an arbitrary nominal (that is, the name of a person).[4]

The rules $@I$ and $@E$ in Figure 1 are the introduction and elimination rules for satisfaction operators. The $@I$ rule says that if we have the information c (so we are reasoning about the person called c) and we also have the information ϕ, then we can introduce the satisfaction operator $@_c$ and conclude $@_c\phi$, which says that ϕ holds from c's perspective. The $@E$ rule says: suppose that when reasoning about the person named c, we also have the information that $@_c\phi$. Then we can eliminate the satisfaction operator $@_c$ and conclude ϕ.

But it is the *Term* rule that drives our analysis. This rule lets us switch to another person's perspective using hypothetical reasoning: the bracketed expressions $[\phi_1]\dots[\phi_n][c]$ in the statement of the rule are (discharged) assumptions. The key assumption is c, which can be glossed as: let's switch perspective and temporarily adopt c's point of view. Incidentally, when using the *Term* rule we must make at least one assumption c, but we can make several, and this is often necessary to drive the proof through. The remaining (discharged) assumptions $[\phi_1]\dots[\phi_n]$ in the rule's statement are additional assumptions we may wish to make about the information available from c's perspective.[5]

The rule works as follows. Suppose that on the basis of assumptions $\phi_1\dots\phi_n, c$ we deduce ψ from c's perspective. Then the *Term* rule tells us that if $\phi_1\dots\phi_n$ are available in the original perspective,[6] then we can discharge the assumption (which we do by bracketing them, thus obtaining the $[\phi_1]\dots[\phi_n][c]$ displayed in the statement of the rule) and conclude ψ unconditionally in the original perspective.

The *Term* rule is subtle and powerful.[7] Indeed, as we shall now see, the hybrid logical analysis of the first-order Sally-Anne task boils down to a single application of *Term*, namely one modelling the shift from the experimental subject's perspective to Sally's.

Formalizing the first-order Sally-Anne task

In this section we describe how (Braüner, 2014) formalizes correct reasoning in the Sally-Anne task. Let us call the child performing the task (the experimental subject) Peter. Let t_0, t_1 and t_2 be three successive times where t_0 is the time at which Sally leaves the scene, t_1 is the time at which the marble is moved to the box, and t_2 is the time after Sally has returned. Here's the intuition we wish to model. To answer the question, Peter views matters from Sally's perspective and reasons as follows. At the time t_0, when Sally leaves, she believes that the marble is in the basket since she sees that it is. As she sees no action to move it, when she is away at t_1 she also believes the marble is in the basket. At t_2, after she has returned, she still believes that the marble is in the basket since she was out of the room when Anne moved it at the time t_1. So Peter concludes that Sally believes that the marble is in the basket.

In our formalization we make use of the predicates $l(i,t)$ and $m(t)$ and the modal operators B and S mentioned above. The argument i in the predicate $l(i,t)$ denotes a location, and the argument t in $l(i,t)$ and $m(t)$ denotes a timepoint. We take time to be discrete, and use $t+1$ to denote the successor of t.

$l(i,t)$	The marble is at location i at time t
$m(t)$	The marble is moved at time t
S	Sees that ...
B	Believes that ...
s	Nominal naming Sally

We also make use of four *belief formation* principles:[8]

(D)	$B\neg\phi \rightarrow \neg B\phi$
$(P1)$	$S\phi \rightarrow B\phi$
$(P2)$	$Bl(i,t) \wedge \neg Bm(t) \rightarrow Bl(i,t+1)$
$(P3)$	$\neg Sm(t) \rightarrow \neg Bm(t)$

Principle (D) is a common modal principle which says if we believe something to be false, then we don't believe it.

Principle $(P1)$ states that beliefs may be formed as a result of seeing: that is, seeing leads to knowing. This is principle (9.2) in (Stenning & van Lambalgen, 2008), page 251.

Principle $(P2)$ is reminiscent of both principle (9.11) in (Stenning & van Lambalgen, 2008), page 253, and axiom $[A_5]$ in (Arkoudas & Bringsjord, 2008), page 20. Principle $(P2)$ formalizes a "principle of inertia" saying that a belief that the predicate l is true is preserved over time, unless it is believed that an action has taken place causing the predicate

mentioned (Bolander, 2014) which analyses first- and second-order false-belief tasks in Dynamic Epistemic Logic. As this is a formalism designed for a (sophisticated form of) model manipulation, this work is probably best classified as a mental models approach. There are also more practically-motivated intermediate approaches, for example the POLYSCHEME agent architecture (Kurup, Bignoli, Scally, & Cassimatis, 2011), which integrates multiple representations and has both rule-based and model-based features.

[3]For a full formulation of the discharge mechanism, we refer the reader to (Braüner, 2011).

[4]This is a modified version of Seligman's original natural deduction system for hybrid logic (Seligman, 1997) and is taken from Chapter 4 of (Braüner, 2011). We omit the rules for propositional connectives such as $\wedge$, $\rightarrow$ and $\neg$ as (a) they are standard and (b) we prefer the more perspicuous proof tree obtained by 'hiding' the simple propositional reasoning involved in the Sally-Anne task.

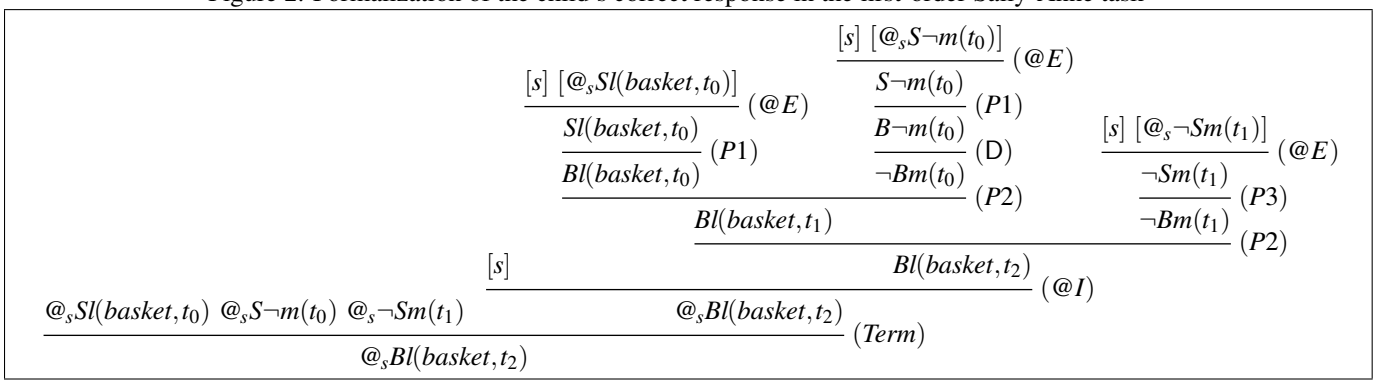
[5]The *Name* rule tells us that if we can prove the information ϕ by adopting some *arbitrary* perspective c, then ϕ also holds from the original perspective. As we won't use this rule in our analysis, we refer once more to (Braüner, 2011) for further discssion.

[6]Indicated by the premisses $\phi_1\dots\phi_n$ listed just above the horizontal line in the statement of *Term* given in Figure 1.

[7]A subtlety worth emphazing is that (as is stated in Figure 1) the assumptions $[\phi_1]\dots[\phi_n]$ must all be satisfaction statements, otherwise the rule is not sound. We refer the reader to (Seligman, 1997) and Chapter 4 of (Braüner, 2011) for further discussion.

[8]This terminology is from (Stenning & van Lambalgen, 2008). The distinction they draw between belief formation and belief manipulation is central to our paper, and we shall return to it shortly.

Figure 2: Formalization of the child's correct response in the first-order Sally-Anne task



to be false.[9]

Principle $(P3)$ encodes the constraint that if someone didn't see the marble being moved, then they can't have come to believe that it moved. Obviously this is not generally true, but the point of the formalization is to capture how Peter reasons about the Sally-Anne scenario (*Well, she can't have seen it, so I guess she won't believe it).*

The perspectival reasoning involved in the Sally-Anne task can be formalized as the derivation in Figure 2. We have already given Peter's informal perspectival reasoning; the formal proof mirrors it in full detail using a single application of *Term* in which the assumptions about s model the shift to Sally's perspective. The first two premises $@_sSl(basket,t_0)$ and $@_sS\neg m(t_0)$ taken together say that Sally at the earlier time t_0 saw that the marble was in the basket and that no action was taken to move it. The third premise, $@_s\neg Sm(t_1)$, says that Sally did not see the marble being moved at the time t_1 (since she was absent). Note that when applying the belief formation principles, we simply use them as rules. For example, when applying $P1$ (that is $S\phi \rightarrow B\phi$, or seeing leads to knowing) we simply use it to license the move from $S\phi$ on one line to $B\phi$ on the next.[10]

From belief formation to belief manipulation

The proof tree just discussed may seem complex, but it has a simple structure. The bulk of the reasoning on the (rather messy) right-hand-side of the proof tree consists of correctly sequencing applications of belief formation principles until the crucial formula $@_sBl(basket,t_2)$ — *Sally believes the ball is in the basket* — is deduced. What turns this into a formalisation of correct reasoning in the Sally-Anne task is the way the sequencing of belief formation principles is perspectivized. The right-hand-side sequencing occurs between the initial assumptions of s (which perspectives it as *Sally's* reasoning) and the final application of *Term* which lets us conclude that the crucial formula is also true from *Peter's* point of view. In short, the analysis consists of *Belief Formation + Perspectival Reasoning* correctly combined.

Analagous remarks are made by Stenning and Van Lambalgen about their own analysis of first-order false-beliefs tasks; see (Stenning & van Lambalgen, 2008), page 257. They note that the bulk of the reasoning involves belief formation principles and their analysis succeeds because it is carried out using closed world reasoning; we might summarise their approach as *Belief formation + Closed World Reasoning* correctly combined. However they then go on to remark that what they call *Belief Manipulation* rules (which codify how to reason from one belief state to another) are unnecessary.

As far as first-order false-belief reasoning is concerned, we agree completely. Indeed, until now we have provided no proof rules for manipulating the belief operator B beyond the belief formation principles. And that is because, to model the first-order Sally-Anne task, we had no need of anything else. But belief manipulation rules will be needed if we are to extend our perspectival analysis to the second-order Sally-Anne task.[11] We turn to this task now.

Formalizing the second-order Sally-Anne task

First an observation. Peter (the experimental subject) has the same beliefs about Sally in the *first-order* Sally-Anne task as Anne has about Sally in the *second-order* task. For example, in the first-order task Peter believes that Sally does not know the ball has moved (and he's right) whereas in the second-order task Anne analogously believes that Sally does not know the ball has moved. Anne is wrong about this (Sally peeked through the window) but the fact remains that her beliefs about Sally match those of Peter in the simpler task.

This suggests that we should extend the first-order analysis by (a) using our first-order analysis as an analysis of *Anne's reasoning about Sally* in the second-order task and (b) embedding this pre-existing proof into a larger proof which captures

[9] For a discussion of how the inertia principle is applied — and just as importantly, how it is *not* applied — in the second-order Sally Anne task, see (Braüner, Blackburn, & Polyanskaya, 2016).

[10] As we remarked earlier, we do this to 'hide' the simple propositional reasoning involved. Strictly speaking, deducing $B\phi$ from $S\phi$ requires us to apply the propositional rule of *modus ponens* to $S\phi \rightarrow B\phi$. Using the belief formation principles as additional natural deduction rules enables us to omit such steps and reduce the size of the proof tree.

[11] Stenning and Van Lambalgen do not analyse second-order false-belief tasks.

Figure 3: Belief manipulation rule for the B operator

$$\frac{B\phi_1 \quad \ldots \quad B\phi_n \qquad \begin{array}{c}[\phi_1]\ldots[\phi_n]\\ \vdots \\ \psi\end{array}}{B\psi}\ (\mathsf{BM})^*$$

$*$ There are no undischarged assumptions in the derivation of ψ besides the specified occurrences of $\phi_1, \ldots, \phi_n$.

Peter's reasoning about Anne in the second-order task. That is, we should add another level of nesting to the perspectival analysis. Making this work requires us to introduce a recursive belief manipulation rule for B.

We have chosen the rule given in Figure 3. We call it BM. It is a version of a rule from (Fitting, 2007) that fits naturally with our tree-style natural deduction proofs.[12] We will also make use of the following additional notation:

D Deduces that ...

a Nominal naming Anne

and of a natural deduction formulation of the belief formation principle

$(P0)$ $\quad D\phi \rightarrow B\phi$

which says that if we can deduce the information ϕ then we believe ϕ. This is principle (9.4) in (Stenning & van Lambalgen, 2008), page 251.

With this machinery in place, the reasoning in the second-order Sally-Anne task can be formalized by the proof tree in Figure 4 (where to save space, we have omitted names of the introduction and elimination rules for the @ operator). Note how this derivation is built around our analysis of the first-order case: the dots in the upper-right corner of Figure 4 mark where this earlier proof slots in.

This is indeed a formalization of the second-order Sally-Anne task. The conclusion, $@_aB@_sBl(basket,t_2)$, says that Anne believes that Sally believes that the marble is in the basket at the time t_2, and this indeed the correct response the the second-order task.

Moreover, Peter (who is now working away at the second-order task) can establish this. The first two premises used in the application of *Term* with which the proof concludes, $@_aS@_sSl(basket,t_0)$ and $@_aS@_sS\neg m(t_0)$, say that at time t_0, Anne saw that Sally saw that the marble was in the basket and that no action was taken to move it — this is the case since both Anne and Sally were present.

The third premise used in the concluding application of the *Term* rule, $@_aD@_s\neg Sm(t_1)$, says that Anne deduced that Sally did not see the marble being moved at the time t_1 — this is the case since Anne was present but Sally was absent at that time (and Anne did not see sneaky Sally peeking).

Finally, note the crucial role the belief manipulation rule BM plays in glueing the two levels of perspectival reasoning together. The embedded proof (which reasons from Sally's perspective) yields the conclusion $@_sBl(basket,t_2)$, the correct response to the *first-order* task. But this is not enough: Peter must be able to deduce that holds from Anne's perspective. The application of BM prefixes the belief operator to form $B@_sBl(basket,t_2)$, and the very next step of the proof shows that this belief holds from Anne's point of view too.

[12] In essence we are adding natural deduction machinery for the modal logic called K, the weakest normal modal logic.

Concluding discussion

We do not deny that second-order reasoning is more complex than first-order — the previous section with its embedded proof and use of the BM rule showed this clearly. Nonetheless, our analysis also suggests that the transition to second-order competence marks a more significant development than is suggested by the complexity only position: the full *reification* of beliefs. Attainment of first-order false-belief competence marks the stage at which the child becomes aware of the fact that beliefs held by other agents can be false; second-order competence marks the stage where beliefs become objects in their own right that can be recursively manipulated. This shift is mirrored in our analysis: we jumped from a logic of *Belief formation + Perspectival Reasoning* to one which allows unrestricted *Belief Manipulation* as well.

This is a significant advance. Beliefs are special objects. The child must learn that they can be embedded one inside another, and acquire the competence to carry out novel logical manipulations — and something like the BM rule, essentially a tool for handling beliefs recursively, seems to be required for this. It is tempting to speculate that at this stage of development some sort of "recursion module" is adapted to handle these strange new objects — but be that as it may, in typically developing children the reasoning architecture seems to be enriched at around the age of six in ways that suggest that a genuine conceptual change has taken place.

Recursively stacked beliefs lie at the heart of this transition, and this brings us back to our empirical work (Polyanskaya, Braüner, & Blackburn, 2016). Our logical investigations were carried out as part of an ongoing training study involving Danish speaking children with ASD. Our study is driven by the hypothesis that the delay (or even failure) experienced by children with ASD to attain second-order false-belief competence is linked to difficulties in belief manipulation. We are investigating whether children with ASD use linguistic recursion as a "scaffolding" to develop belief manipulation; this might explain why some children with ASD can pass second-order false-belief tasks, an explanation which was advanced in the first-order case by (Hale & Tager-Flusberg, 2003).

At the time of writing, our study was still work in progress. Nonetheless, we hope that these remarks show that there is a link between the abstract work of this paper and the concrete reality of psychological experimentation.

Acknowledgements

The authors acknowledge the funding received from the VELUX FOUNDATION for the project *Hybrid-Logical Proofs at Work in Cognitive Psychology* (VELUX 33305).

Figure 4: Formalization of the child's correct response in the second-order Sally-Anne task

$$\dfrac{\dfrac{\dfrac{[a][@_aS@_sSl(basket,t_0)]}{S@_sSl(basket,t_0)}(P1) \quad \dfrac{[a][@_aS@_sS\neg m(t_0)]}{S@_sS\neg m(t_0)}(P1) \quad \dfrac{[a][@_aD@_s\neg Sm(t_1)]}{D@_s\neg Sm(t_1)}(P0) \quad \begin{array}{c}[@_sSl(basket,t_0)][@_sS\neg m(t_0)][@_s\neg Sm(t_1)]\\ \vdots \\ @_sBl(basket,t_2)\end{array}}{B@_sBl(basket,t_2)}(\mathsf{BM}) \quad \dfrac{}{[a]}}{@_aB@_sBl(basket,t_2)}(Term)$$

The vertical dots in the upper-right corner represent the derivation in Figure 2. So this proof contains two applications of *Term*: the concluding application, which is shown, and the one inside the earlier proof, which is not.

References

Arkoudas, K., & Bringsjord, S. (2008). Toward formalizing common-sense psychology: An analysis of the false-belief task. In *PRICAI 2008: Trends in artificial intelligence* (Vol. 5351, pp. 17–29). Springer-Verlag.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

Blackburn, P. (2000). Representation, reasoning, and relational structures: a hybrid logic manifesto. *Logic Journal of the IGPL*, *8*, 339–365.

Bolander, T. (2014). Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In A. Herzig & E. Lorini (Eds.), *Proceedings of the European conference on social intelligence (ECSI-2014)* (pp. 87–107). IRIT-CNRS, Toulouse University, France.

Braüner, T. (2011). *Hybrid logic and its proof-theory* (Vol. 37). Springer.

Braüner, T. (2014). Hybrid-logical reasoning in the Smarties and Sally-Anne tasks. *Journal of Logic, Language and Information*, *23*, 415–439.

Braüner, T. (2015). Hybrid-logical reasoning in the Smarties and Sally-Anne tasks: What goes wrong when incorrect responses are given? In *Proceedings of the 37th annual meeting of the cognitive science society* (pp. 273–278). Pasadena, California: Cognitive Science Society.

Braüner, T., Blackburn, P., & Polyanskaya, I. (2016). Second-order false-belief tasks: Analysis and formalization. In *Proceedings of Workshop on Logic, Language, Information and Computation (WoLLIC 2016)*. Springer-Verlag.

Fitting, M. (2007). Modal proof theory. In P. Blackburn, J. van Benthem, & F. Wolter (Eds.), *Handbook of modal logic* (pp. 85–138). Elsevier.

Gordon, R. (2009). Folk psychology as mental simulation. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford University. (On-line encyclopedia article available at http://plato.stanford.edu/entries/folkpsych-simulation.)

Hale, C., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: a training study. *Developmental Science*, *6*, 346–359.

Kurup, U., Bignoli, P., Scally, J., & Cassimatis, N. (2011). An architectural framework for complex cognition. *Cognitive Systems Research*, *12*, 281–292.

Miller, S. (2009). Children's understanding of second-order mental states. *Psychological Bulletin*, *135*, 749–773.

Miller, S. (2012). *Theory of mind: Beyond the preschool years*. Psychology Press.

Milligan, K., Astington, J., & Dack, L. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*, 622–646.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." attribution of second-order beliefs by 5-to 10-year-old children. *Journal of Experimental Child Psychology*, *39*, 437–471.

Polyanskaya, I., Braüner, T., & Blackburn, P. (2016). *Linguistic recursion and Autism Spectrum Disorder.* Manuscript.

Rips, L. (2008). Logical approaches to human deductive reasoning. In J. Adler & L. Rips (Eds.), *Reasoning: Studies of human inference and its foundations* (pp. 187–205). Cambridge University Press.

Seligman, J. (1997). The logic of correct description. In M. de Rijke (Ed.), *Advances in intensional logic* (Vol. 7, pp. 107 – 135). Kluwer.

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. MIT Press.

Sullivan, K., Zaitchik, D., & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order beliefs. *Developmental Psychology*, *30*, 395–402.

Szymanik, J., Meijering, B., & Verbrugge, R. (2013). Using intrinsic complexity of turn-taking games to predict participants' reaction times. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1426–1432). Cognitive Science Society.

Syntactic Flexibility in the Noun: Evidence from Picture Naming

Nicholas A. Lester (nlester@umail.ucsb.edu)
Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)
Department of Linguistics
University of California, Santa Barbara, South Hall 3432
Santa Barbara, CA 93106 USA

Abstract

Does syntactic information affect the production of bare nouns? Research into this issue has explored word-specific features (e.g., gender). However, word-independent syntactic distributions may also play a role. For example, studies of word recognition have uncovered strong effects of the diversity of a word's syntactic distribution – its syntactic flexibility – on response times in the lexical decision paradigm. By contrast, studies of sentence production have produced strong but conflicted effects of syntactic flexibility. We propose that syntactic flexibility also affects production of individual words. We reanalyze a database of previously collected timed picture naming data using two novel measures of syntactic flexibility, one based on the relations stemming from the noun, and one based on the relations extending to the noun. Our results show that nouns that project a diverse array of structures are produced faster, and those that are integrated into a diverse array of structures are produced slower.

Keywords: syntactic flexibility; word production; picture naming; entropy

Introduction

Language production is fast – very fast – but the speed of the system is not without variability. One source of variability is the diversity of options available for encoding a given message. Most mainstream models of speech production rely on the notion of interactive competition, whereby the presence of alternatives delays encoding (so-called *interference effects;* e.g., Levelt, Roelofs, & Meyer, 1999). Such effects have been observed at the level of word production with respect to shared syntactic features of competitors: It takes longer to produce a word if other syntactic competitors are simultaneously active. Some studies have investigated similar issues at the sentence level. In one study, the number of sentence structures allowed by a given arrangement of words – their *syntactic flexibility* – was shown to correlate negatively with speech onset latencies (V. S. Ferreira, 1996). However, more recent research has reported the opposite findings: greater syntactic flexibility produced traditional interference effects, with slower production onsets for clauses with greater syntactic flexibility (Myachykov et al., 2013).

The clause-level studies have focused on the arrangement of arguments into abstract syntactic frames. One question that has not been addressed is how the syntactic flexibility of nouns affects their accessibility in production. Much work has been devoted to outlining the representation of syntactically relevant features of nouns, such as gender, as well as the time course of their activation across different sorts of tasks. However, to our knowledge, no study has specifically addressed whether a noun's retrieval depends on the diversity of its syntactic distributions. Are nouns that exhibit greater syntactic flexibility easier or harder to produce?

Syntactic Aspects of Lexical Access

All models of word production posit associations between individual lexical items and syntactic information (Caramazza, 1997; Dell, 1986; Levelt, Roelofs, & Meyer, 1999). However, they differ with respect to the time-course and directionality of activation between word-level and syntax-level nodes. For the most part, this research has focused on whether words share access to abstract specifications of gender and whether these gender-specifying nodes feed back to the lexical nodes during lexical retrieval. These studies typically rely on the picture-word interference task. In this task, participants are asked to name an image while ignoring a distractor word. The word either matches or conflicts with the target name with respect to its gender. The results have been split. If participants are asked to produce the picture name in a syntactic frame that requires access to the noun's gender (e.g., *die Katze*), then targets with matched-gender distractors are produced slower than mis-matched pairs (e.g., Schriefers & Teruel, 2000). Effects of this kind are typically attributed to competition between the distractor and target for access to a shared gender node. By contrast, when participants are asked to name a picture with a bare noun in the same task (e.g., *Katze*), then no interference effect is observed in the matched condition (La Heij, Mark, Sander, & Willeboorsde, 1998). This has led some to conclude that syntactic features are accessed preferentially during syntactic (i.e. phrasal), as opposed to purely lexical, production. However, a small number of studies eliciting bare-noun names in Romance languages have uncovered interference effects in the gender-matched condition (e.g., Cubelli et al., 2005). Thus, there is some support for activation of grammatical features in the picture-naming task, regardless of whether the form of the response requires that information.

Further support for universal access of syntactic features in bare-noun picture naming comes from a similar phenomenon – the mass/count distinction in English nouns (Gregory, Varley, & Herbert, 2012). In English, the mass/count noun distinction controls determiner type. For example, in their singular form, count nouns such as *cat* can take the determiner *each* (*each cat*), while mass nouns such

as *flour* cannot (?**each flour*). Gregory et al. (2012) found that bare-noun picture naming of count and mass nouns was facilitated by a grammatically congruent determiner prime. They attribute this finding to pre-activation of the syntactic feature shared between the determiner prime and the target word.

So far, we have seen that grammatical classes are relevant for word processing in production. Less clear is the extent to which syntactic relations – the scaffolds into which words are fit – impact lexical retrieval. To make this contrast more clear, consider that grammatical classes such as gender have consequences for the formal properties of words and their dependents (e.g., different patterns of agreement). However, they do not constrain the syntactic distributions of words. For example, the German words *Katze* and *Hund* belong to two different gender classes which take different forms of the definite determiner (*die* vs. *der*), along with different patterns of agreement elsewhere in the grammar. Nevertheless, both can occur with definite determiners, whatever their form. Going further, both may occur with adjectives, as subjects of sentences, as objects of prepositions, and so on. Crucially, gender does not dictate the availability of these relations, only the forms of the relata[1]. If information about grammatical class is activated when we access nouns for production, even when it doesn't impact the form of the utterance, then it is possible that other abstract syntactic information is likewise activated. Some recent research from comprehension supports this possibility.

Baayen et al. (2011) showed that the shape of a noun's (partially lexicalized) syntactic distribution impacts the time it takes to recognize that noun in visual lexical decision. They collected all trigrams of the form *preposition + determiner + noun* (for example, *on the table*). They then measured the diversity of the prepositions occurring with each noun and correlated this measure with lexical decision latencies. Nouns that occurred more evenly across a wider array of prepositions were recognized faster than words that skewed toward a limited array of prepositions. While interesting, these results speak only indirectly to the question raised above. This is because the trigrams used in the study do not contrast syntactic types, but rather lexical types (prepositions) within a single syntactic relation (prepositional object). Moreover, the proximity of the preposition and the noun opens the possibility that the facilitatory effect might actually be semantic in nature. Variability in small-scale co-occurrence windows is often assumed to reflect or indeed to constitute the meaning of words (Bullinaria & Levy, 2007).

In a follow-up study, Lester and Moscoso del Prado Martín (2015) compared the distributions of nouns across prepositional phrases within a parsed corpus of English writing. Crucially, the distance between noun and preposition could be of any size. First, they replicated the negative correlation between diversity and response time. To investigate the source of the effect in more detail, they compared the syntactically defined diversity space with a semantically-defined space (based on Latent Semantic Analysis vectors). They found almost zero correlation between the two. Second, they found that words with similar distributions in the syntactic space did not prime one another in an overt semantic priming task, while words with similar distributions in the semantic space did. Third, they showed that increasing average distance between the nouns and their prepositions reduced the strength of the facilitation, but did not eliminate or overturn it. That is, small-scale co-occurrence does seem to contribute something to our measure: when nouns tend to occur closer the prepositions, the facilitation is increased. However, it alone cannot explain the continued facilitation at longer distances. These findings suggest that something other than meaning must be contributing to the facilitation. A natural candidate is syntax.

These studies show that words encode abstract distributional information about their use and that this information becomes active in comprehension, even for isolated words. However, no study has yet compared syntactic types; the two studies above only treat the relation between a preposition and its object. Consequently, no study has measured diversity at a purely syntactic level, that is, without regard for the particular lexical types that anchor that relation opposite the target noun. Recall that the operationalizations of diversity outlined above included explicit reference to particular prepositions (e.g., how many times *table* occurs with *on* vs. *near* vs. *at* vs. ...). Finally, no study has yet examined how such effects play out in word production.

In the present study, we address these shortcomings by considering how flexibility across the entire set of grammatical relations afforded by a noun affects onset latencies in picture naming. We refer to this measure as the *total flexibility*. We further use the functional contrast between syntactic heads and dependents to decompose the total syntactic flexibility. On the one hand, we take the diversity of syntactic relations through which a noun serves as a dependent to some other head. Essentially, this measure captures how easily a noun can be integrated into a syntactic structure. We refer to this measure as *dependent flexibility*. On the other hand, we take the diversity of syntactic relations for which a noun can serve as head. This measure captures the structure-building potential of the noun. We refer to this measure as *head flexibility.* Importantly, all three of our measures are taken over abstract syntactic relations; they make no reference to any words besides the targets.

[1]Other grammatical classes may constrain the syntactic possibilities of nouns. For example, English mass nouns like *water* may function as their own phrase in non-metalinguistic contexts (*Water is wet*) while count nouns typically cannot (**Cat is furry*). However, these constraints are likely to be relatively minor. As the examples above demonstrate, syntactic distributions are at least partially independent of grammatical class.

Hypotheses

We contrast two general predictions regarding the behavior of our measures. We base these predictions on findings from sentence production as no comparable data exist for lexical access.

— The opportunistic hypothesis: Nouns with greater syntactic flexibility will be produced faster.

The first possibility derives from what Myachykov et al. (2013) termed the *opportunistic hypothesis*. The opportunistic hypothesis states that syntactic flexibility is generally facilitative. The assumption is that production operates most smoothly when the speaker has the most options for completing the current syntactic scaffold. Therefore, syntactically diverse nouns should be produced faster than syntactically limited nouns. Indirect support for this hypothesis comes from research on sentence production, which has shown that people will start speaking sooner when they have more syntactic options available for translating a scrambled word list into a grammatical sentence (V. S. Ferreira, 1996).

— The strategist hypothesis: Nouns with greater syntactic flexibility will be produced slower.

The second possibility derives from what we refer to as the *strategist hypothesis*. The strategist hypothesis states that more options require more deliberation before a choice can be made. Therefore, syntactically diverse nouns should be produced slower than syntactically limited nouns. Support for this possibility again comes from the sentence production literature. Myachykov et al. (2013) had speakers of Russian and English speakers describe scenes depicting transitive events. They found that Russian speakers took longer to initiate their descriptions than the English speakers. The authors attribute the finding to the much larger number of possible syntactic constructions available to Russian speakers (12 in total, as opposed to 2 for English). Hwang and Kaiser (2014), using a task similar to Ferreira (1996), likewise found slower onset latencies for syntactically flexible stimuli in Korean.

Neither hypothesis treated so far contrasts the effects of head and dependent flexibility. This is because no study to our knowledge has investigated whether different syntactic functions may respond differently to changes in diversity. This, coupled with the conflicted nature of the findings from sentence production, leads us to eschew any specific predictions with respect to our head and dependent measures.

Methods

To evaluate these hypotheses, we reanalyze the mean speech onset latencies from a large-scale picture-naming study (Bates et al., 2003; since published as part of the *International Picture Naming Project*). These data are ideal for testing the role of syntactic distributions in lexical access because they were collected using the bare-noun naming paradigm. This means that any effect of syntax that we observe cannot be attribute to features of the task (e.g., the fact that a speaker was required to activate a particular syntactic scaffold when responding).

Data

Bates and colleagues showed speakers of seven different languages a series of black-and-white line drawings and asked them to provide a bare-noun label (e.g., *cat*) for each image. They were instructed to say the name as quickly and fluently as possible. This instruction is critical: Ferreira and Swets (2002) show that speeded production tasks can induce more opportunistic processing strategies. Therefore, the data may be biased towards confirming the opportunistic hypothesis. Reaction times and errors were collected.

From the full data set, we extracted the mean onset RTs for the dominant, or majority names offered for each image. Of the 520 dominant responses, we took the subset of non-phrasal forms by omitting names that standard English orthography would divide into multiple words (such as *fishing pole*). We estimate the syntactic diversity of the remaining words on the basis of their distributions within the filtered Charniak-parsed treebank of the Open American National Corpus (OANC). This treebank contains phrase-structure parses of approximately 363,000 sentences of written American English from several genres, totaling around 11M words. We retrieved all parse trees containing the target words so long as the latter were tagged as common singular nouns We observed 416 of the original 520 word types in this sample (frequency in OANC: mean = 137.2; median = 42).

To tease apart the effects of flexibility for nouns as heads and nouns as dependents, we converted the phrase-structure trees that we culled from the OANC into *dependency graphs* using the Stanford dependency parser (de Marneffe, MacCartney, & Manning, 2006). In the dependency graph formalism, grammatical relationships are represented as ternary units consisting of a *head*, a *dependent*, and a typed *relation* (Tesnière, 1959). For example, *the fat cat* in the sentence *The fat cat ate the tuna* contains two dependencies: (1) a determiner relation *det* headed by *cat* and projected onto *the* and (2) an adjectival-modification relation *amod* headed by cat and projected onto *fat.* In the Stanford notation, these relationships surface as det(cat-3, the-1) and amod(cat-3, fat-2), respectively, where the numbers indicate sequential positioning in the clause. This formalism allows us to implement our measures in a simple, straightforward way. Dependent flexibility is operationalized as the diversity of relations for which the target noun is a dependent. Examples include the *nsubj* relation, which links a noun to a verb as its subject, or the *pobj* relation, which links a noun to a preposition as its object. Head flexibility is operationalized as the diversity of relations for which the target noun serves as head. Examples include the *amod* relation, which links an adjective to a noun, or *det*, which links a determiner to a noun. Importantly, we only care

about the target nouns and the relations; we ignore the words that are linked to the targets via the relation.

Syntactic Measures

To measure the diversity of the relation types, we generated three sparse matrices. Each matrix captures the frequency distributions of the target nouns (rows) across the set of relation types (columns). We constructed one such matrix for dependent relations, one for head relations, and one for all relations, regardless of whether the target was head or dependent.

For each word vector in each matrix, we computed the *Shannon entropy* (Shannon, 1948) of the distribution. This measure is particularly good at capturing diversity as it considers not only the number of types of relations, but also the relative frequencies of those types. The measure thus increases both with the number of types and with a more even spread of the instances across those types. Shannon entropy, denoted *H*, is defined as the mean of the weighted negative log probabilities of event *x* taken across outcomes *i* (Eq. 1).

$$H = -\sum_{i=1} p(x_i) \log\ p(x_i) \qquad (1)$$

Eq. 1 reflects the *maximum-likelihood* estimate of diversity within our vectors. As is, this measure will necessarily underestimate the true diversity of the sample. This is because we cannot be sure that the observed set of relations actually exhausts the set of possible relations for our targets. To counter this downward bias, we smooth our entropy estimates using the Chao, Wang, & Jost (2013) method (see Moscoso del Prado Martín, in press for a demonstration of the reliability of this message for word frequency distributions). We abbreviate our smoothed measure of head flexibility H_h and our smoothed measure of dependent flexibility as H_d. We abbreviate our baseline total flexibility measure as H_t.

Additional Controls

We also gathered information about the sentential distributions and semantic properties of each noun attested in our sample to avoid potential non-syntactic confounds. A word's typical positioning within a sentence may reflect aspects of its accessibility. Most models of sentence production hold that word order is partially determined by the relative speed with which the words are accessed (within the syntactic constraints afforded by the grammar of the language; Gletiman et al., 2007). Therefore, a word's general likelihood of showing up earlier in clauses might translate into faster bare-noun production in picture naming. In particular, words that frequently occur prior to the verb (e.g., as sentential subject or left-dislocated topic) might simply be more discourse-salient for speakers (better *starting points*; Konopka & Brown-Schmidt, 2014). Thus, we calculate the average relative positioning (position over sentence length) of each word, as well as the likelihood of the word preceding the sentential *root* dependency (instantiated lexically by the main verb).

Semantic effects in the timed picture naming tasks are split. Bormann (2011) found no difference in onset RTs for words from large or small semantic neighborhoods. However, Bates et al. (2003) report strong interference effects for onset RTs when the picture being described elicited more different names across participants. We likewise include the number of other names provided as a control. We also include a predictor of referent animacy, as animates are well known to be more accessible than inanimates (e.g., Ariel, 1991).

Finally, we include a number of other controls known to affect picture naming latencies: (log) lemma word frequency (CELEX), phonological complexity, visual complexity of the stimulus image, and subjective ratings of conceptual complexity.

Results

We computed several generalized additive models (GAMs) predicting mean (non-transformed) picture-naming RTs as dependent variable. As a first step, we added all control predictors (with spline-based smooths for all numeric predictors). After these, we added our total-flexibility measure H_t (with and without smoothing). As expected, we found a facilitatory linear effect of word frequency (β = -33.81, p<.0001) and a facilitatory effect of Bates and colleagues' measure of name agreement (p<.0001). Our baseline flexibility measure H_t had no effect (with or without smoothing; p>0.2).

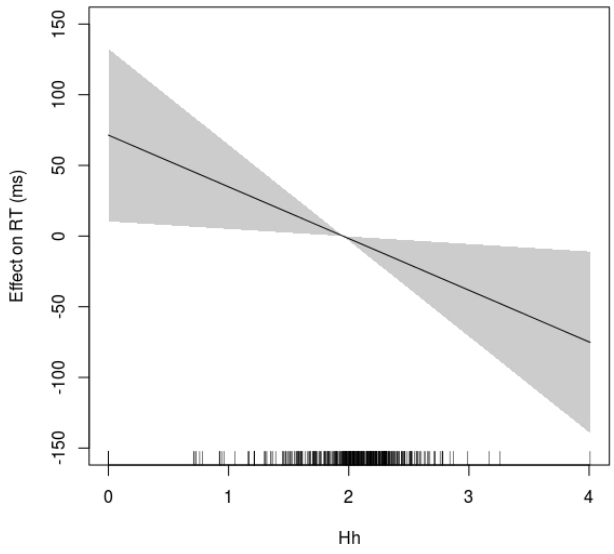


Figure 1: Facilitatory effect of H_h

In the next round, we repeated the model selection process, but substituted our measures of head flexibility H_h and dependent flexibility H_d for the total measure H_t. Collinearity between the head and dependent measures was

within the acceptable range (c=10.37)[2] as was the collinearity between each and word frequency (c<10 for both). The model returned significant effects of the same two control variables at essentially identical strengths. It also uncovered a facilitatory effect of H_h (p<.05; Figure 1) and an inhibitory effect of H_d (p<.05; Figure 2). Both effects are of roughly equal size, with a difference in extremes of ~100 ms.

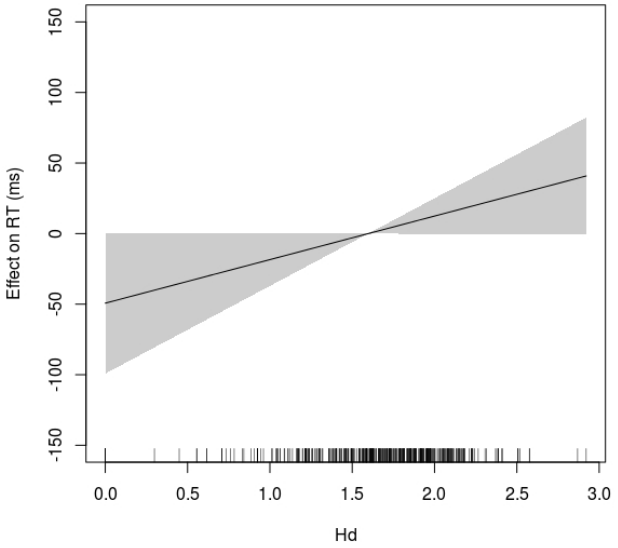


Figure 2: Inhibitory effect of H_d

Discussion

Our results provide first evidence that a noun's syntactic distribution affects its accessibility in word production. These findings support and expand prior work which has demonstrated that speakers access syntactic features when producing words in isolation (Cubelli et al., 2005; Gregory et al., 2012). Moreover, these findings suggest that words encode quite fine-grained information about their syntactic potential. Finally, we show that the function of the noun within syntactic relations plays a critical role in determining its accessibility. We find facilitation for nouns that are heads of a diverse array of relations and inhibition for those that are dependents of a diverse array of relations.

The facilitatory effect of head diversity is compatible with the opportunistic hypothesis, which states that production proceeds best when we have more options for structural encoding (V. S. Ferreira, 1996). Following this logic, nouns that provide more possibilities for expanding their phrasal nodes could be accessed more quickly because their selection does not immediately commit the speaker to any particular phrasal structure. Of course, the utterances analyzed here consist of single words. However, the opportunistic hypothesis makes reference only to *possible* structures. In that sense, the speed of lexical access need not depend on the intention to build a broader syntactic unit around the word. Rather, it may depend on the array of structures suggested by its activation. But this is only part of the story. Recall that several studies have shown that other forms of syntactic information, such as gender, can affect production of single words (e.g., Cubelli et al., 2005). Taken together, these facts suggest that the bare-noun naming paradigm may not be effective at eliminating syntactic processing, at least not entirely. This makes sense if we consider that people rarely produce words apart from a syntactic context. If when the system activates a word, it expects strongly to produce that word in some syntactic arrangement, then it may persist in activating syntax-relevant information despite the constraints imposed by the task.

Turning to dependent flexibility, we uncovered an inhibition effect: nouns that showed a greater diversity of embedding contexts were produced slower. This finding is compatible with the strategist hypothesis (Myachykov et al., 2013), which predicts that more options make it more difficult to settle on the final outcome. This result may be directly related to the task. The participants were required to produce only a single word. The single-word utterance is one of many syntactic templates in which a word could have been experienced (e.g., in exchanges like A. *What did you say?* B. *Cat.*). In dependency grammar, this would be represented as the ROOT relation (a special 'headless' dependency type that only takes a dependent). Thus, selecting the ROOT relation from among the other possible structures may have resulted in a traditional interference effect. This point distinguishes head and dependent diversity within the context of bare-noun picture naming. When producing a noun in the ROOT relation, one has actually committed to a syntactic structure – one of many available to call the noun as dependent. However, when there is no intention to elaborate the noun into a noun phrase, the diversity of its possible head relations can facilitate access without interference. The structures that would otherwise be competing for selection now support access by feeding activation back into the target lexeme. The more diverse the feedback network, the faster the access.

This pattern of results underscores two critical points regarding language processing: (1) words are finely articulated syntactic entities whose history of use partially determines how efficiently they are processed – even in isolation! – and (2) flexibility is not a simple component of syntactic entities, but one that interacts with different functional domains to help or hinder processing. Perhaps words and syntactic structures are much more tightly linked than is typically acknowledged. If so, then the effects of flexibility observed here might be attributed to the flow of

[2] Collinearity between our three flexibility measures exceeded the acceptable range (c=24.5). To clean this collinearity, we performed a principal components analysis (PCA) on the three measures. The PCA yielded a component that distinguished H_h (.64) from H_m (-.76) with near 0 loading for H_t. When we entered this component into the GAM frame described above, we found a noisier but significant effect ultimately equivalent to the individual effects of H_h and H_m (p < .05). However, to maximize the clarity of results attributable to our decomposed measures, we report the analysis that simply omits H_t.

activation within the network – top-down from the syntactic tier or bottom-up from individual words. This type of processing model would be more in line with current linguistic models of representation, for which grammar and lexicon form a continuum (e.g., Goldberg, 2006). Architecturally, it wouldn't be a matter of representing syntax *in* the noun so much as *with* the noun. At any rate, continued exploration and refinement of these diversity effects will no doubt sharpen our perspective on the relationship between words and structure.

References

Ariel , M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*, 443-463.

Baayen, R. H., Milin, P., Filipović-Đurđević, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychological Review, 118, 438-482.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, L. Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., & Tzeng, O. (2003). Timed picture naming in seven languages. *Pscyhonomic Bulletin and Review*, *10*, 344-380.

Bock, J. K. & Ferreira, V. S. (2014). Syntactically speaking. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds), The Oxford Handbook of Language Production (pp. 21-46). New York: Oxford University Press.

Bormann, T. (2011). The role of lexical-semantic neighborhood in object naming: Implications for models of lexical access. *Frontiers in Psychology*, *2*. 1-11.

Bullinaria, J. A. & Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, *39*, 510-526.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*, 177–208.

Chao, A., Wang, Y. T., & Jost, L. (2013). Entropy and the species accumulation curve: a novel estimator of entropy via discovery rates of new species. *Methods in Ecology and Evolution*, *4*, 1091–1110

Cubelli, R., Lotto, L., Paolieri, D., Girelli, M., & Job, R. (2005). Grammatical gender is selected in bare noun production: Evidence from the picture-word interference paradigm. *Journal of Memory and Language*, *53*, 42–59.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283-321.

de Marneffe, M-C., MacCartney, B., & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57–84.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, *35*, 724–755.

Gleitman, L., January, D., Nappa, R., & Trueswell, J.C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*, 544-569.

Goldberg, A. E. (2006). *Constructions at work*: *The nature of generalization in language.* Oxford: Oxford University Press.

Gregory, E., Varley, R., Herbert, R. (2012). Determiner primes as facilitators of lexical retrieval. *Journal of Psycholinguistic Research*, *41*, 439-453.

Hwang, H., & Kaiser, E. (2014). Having a syntactic choice is not always better: The effects of syntactic flexibility on Korean production. *Language, Cognition, and Neuroscience*, *29*, 1115-1131.

Konopka, A. E., & Brown-Schmidt, S. (2014). Message encoding. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds), The Oxford Handbook of Language Production (pp. 3-20). New York: Oxford University Press.

La Heij, W., Mark, P., Sander, J., & Willeboorsde, E. (1998). The gender congruency effect in picture word task. *Psychological Research*, *61*, 209–219.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical prerequisites.* Stanford: CSLI.

Lester, N. A., & Moscoso del Prado Martín, F. (2015). Constructional paradigms affect visual lexical decision latencies in English. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1320-1325). Austin, TX: Cognitive Science Society.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Moscoso del Prado Martín, F. (in press). Vocabulary, grammar, sex, and aging. *Cognitive Science.*

Myachykov, A., Scheepers, C., Garrod, S., Thompson, D., & Fedorova, O. (2013). Syntactic flexibility and competition in sentence production: The case of English and Russian. *The Quarterly Journal of Experimental Psychology*, *66*, 1601-1619.

Schriefers, H., & Teruel, E. (2000). Grammatical gender in noun phrase production: The gender interference effect in German. *Journal of Experimental Psychology: Learning memory, and Cognition*, *26*, 1368-1377.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 28(3), 379-423.

Tesnière, L. (1959). Éléments de syntaxe structurale. Paris: Klincksieck.

Alien species and alienable traits: An artificial language game investigating the spread of cultural variants between antagonistic groups

Betsy Sneller (esnell@sas.upenn.edu)
Department of Linguistics, University of Pennsylvania
255 South 36th Street, Philadelphia, PA 19104 USA

Gareth Roberts (gareth.roberts@ling.upenn.edu)
Department of Linguistics, University of Pennsylvania
255 South 36th Street, Philadelphia, PA 19104 USA

Abstract

The spread of cultural variants, such as dress or speech patterns, may be promoted or inhibited by different types of bias. In model-based bias, variants are differentially adopted according to characteristics of individuals exhibiting them. A surprising case of cross-group adoption comes from sociolinguistic fieldwork in which White speakers were observed exhibiting a feature of African-American Vernacular English, in spite of expressing aggressively negative attitudes towards their African-American neighbors. A likely explanation for this is that the feature in question had become dissociated for these speakers from the inalienable trait Blackness, but had retained associations with the more alienable trait of being "street" or tough. We tested this by conducting an artificial-language experiment in which groups of four participants played a computer game that involved typing instant messages to each other, trading resources, and fighting. Participants were assigned to one of two mutually antagonistic "alien species" (weaker Wiwos and tougher Burls) and learned an alien language with two species-specific dialects. In one condition, the Wiwos were told that that Burl dialect was mainly used by Burls; in the other condition they were told it was mainly used by "tougher aliens". Burl variants were significantly more likely to be used by Wiwos in the latter condition than in the former, even though they were associated with tougher aliens in both conditions. This suggests that cultural variants linked to more alienable traits are more likely to be adopted than those linked to inalienable ones, even if the practical implications of the two traits are very similar.

Keywords: language variation and change; dialect contact; cultural evolution; artificial language

Introduction

A core process of cultural evolution is the propagation of cultural variants – such as styles of art, technology, or dress – between individuals. Once a cultural variant has been innovated, it must spread to other individuals before cultural evolution can be said to have taken place. Crucially, this propagation occurs not only between the innovator and their immediate contacts, but also between the first contacts and their own circles of influence. In this way, the borrowing or spread of cultural variants is a crucial aspect of cultural evolution. Linguistic examples of cultural evolution abound; indeed, a conventional communication system like language relies for its success on such propagation, and dialects or dialect features may spread to huge populations of speakers covering vast geographic areas, as is the case for the Inland North dialect of North America, found in speakers across the Great Lakes from Chicago, IL to Rochester, NY (Labov, Ash, & Boberg, 2006). Like other cultural variables, speech patterns also serve as social markers, however, and this can lead to small-scale patterns of variation serving to distinguish social groups, which may be defined on the basis of personality traits such as "jocks" and "burnouts" (Eckert, 2000) or inherent attributes such as gender (Eisikovits, 1981).

For cultural variables to serve as social group markers in this way, there has to be a mechanism that limits their spread outside group boundaries. One mechanism, compatible with a neutral evolutionary model, is variation in the frequency with which individuals interact. Alternatively, individuals may be biased in their adoption of variants. Richerson and Boyd (2006, p. 69) distinguished between three kinds of bias, which may operate in isolation or in combination with one another: *content-based bias*, based on the nature of a variant, *frequency-based bias*, based on its commonness or rarity,[1] and *model-based bias*, based on characteristics of individuals bearing the variant.

In this paper we will focus on model-based bias in the transmission of linguistic variants, and will draw a distinction between two kinds of characteristics that the model might have. The first, which we term *inalienable traits*, are characteristics of the model that are inherent to the individual, and effectively do not change, such as height, race, or sex. The second kind, which we term *alienable traits*, are characteristics that are acquired, and may be learned or abandoned over time, such as "honest" or "mean." This distinction matters, assuming that model-based adoption of a cultural variant is driven by a desire to be associated with characteristics of a cultural variant's bearers. We may expect variants linked to inalienable traits to be mainly adopted by other bearers of those traits, and variants linked to alienable traits to be adopted by a wider range of individuals. High school girls, for example, may be more likely to adopt their male classmates' speech patterns if those speech patterns are associated predominately with some desirable alienable quality, like coolness, than if they're strongly associated with being a boy.

Similarly, White speakers and African American speakers have been observed to diverge from each other in both phonology and morphology within local dialects (Wolfram, 2004; Fisher & Labov, 2015; Van Herk, 2008). Where con-

[1] This should be distinguished from the role of frequency in neutral evolution, where variants may come to dominate as a result of sampling error; frequency-based bias, by contrast, can involve the *selection* of a variant on the basis of perceived rarity.

vergence occurs across racial lines, it is typically toward a prestigious mainstream dialect, associated more with such alienable traits as education, wealth and power than with race. Cross-racial shifts towards local White or African American dialect occur rarely, and usually in cases where the speaker in question holds strong social ties with speakers of the relevant dialect (Fix, 2010; Sweetland, 2002). Other cases involve more temporary shifts in which features are appropriated, allowing the claiming of social capital that comes through association with urban Blackness, such as "coolness" (Ibrahim, 1999; Cutler, 1997), by speakers relatively well disposed toward (their perception of) urban Black culture.

A striking and unusual counterexample to this was reported by Sneller (2014), who conducted fieldwork in which White speakers in a South Philadelphia neighborhood were found to be exhibiting (TH)-fronting, a feature of African-American Vernacular English (AAVE) that is well attested among African-American speakers in South Philadelphia (and elsewhere), but otherwise entirely unattested among White speakers in Philadelphia. The White speakers in question not only lacked strong social ties with African Americans, but they also espoused violently negative attitudes toward their African American neighbors. Sneller hypothesized that this paradoxical situation would be explicable if (TH)-fronting had become dissociated from being African American, an inalienable trait, and had instead become an index of street culture (Anderson, 1999) or toughness. This hypothesis is supported by the fact that all the speakers who exhibited the feature were males who were involved in street activities, such as drug deals and the stolen bike trade. Furthermore, one of the White speakers with the highest rates of fronting and one of the strongest negative attitudes toward his African American neighbors explicitly identified the fronted variants as "street," rather than as a feature associated with AAVE.

Experimental paradigm

While ethnographic studies like that of Sneller (2014) play an important role in informing models of linguistic (and other cultural) change, they do not allow hypotheses to be tested directly. Further ethnographic investigation in cases like this is also hampered by ethical concerns surrounding the involvement of serious racial tensions (including violently aggressive attitudes) and criminal activity. To test Sneller's hypothesis – and, more broadly, to investigate the role of alienable and inalienable traits in cultural transmission – we therefore conducted an experimental simulation, following an artificial-language paradigm developed by Roberts (2008, 2010), in which participants play a game with each other as aliens, and communicate by typing messages in a small artificial "alien language". Especially when employed in conjunction with studies of real-world data, this approach has a number of advantages.

First, it avoids the methodological and ethical hurdles noted above. Second, by employing a miniature artificial language, researchers are able to apply laboratory control without compromising on the cognitive plausibility of the agents, thus providing a middle-ground between ethnographic studies and computational simulations (see Galantucci & Roberts, 2012, for further discussion of the advantages of this approach). Third, the method allows language change to be observed much more rapidly than would be the case outside the laboratory. Finally, the approach has some advantages over more traditional artificial-language learning experiments (Folia, Uddén, De Vries, Forkstam, & Petersson, 2010), particularly for investigating questions with a sociolinguistic dimension. Rather than being explicitly quizzed on their knowledge of the artificial language and the social correlates of variants within the language, participants in this paradigm use the language to accomplish a nonlinguistic primary task. This helps to reduce the effect of the Observer's Paradox (Labov, 1972), wherein participants may alter their behavior as a result of being observed.

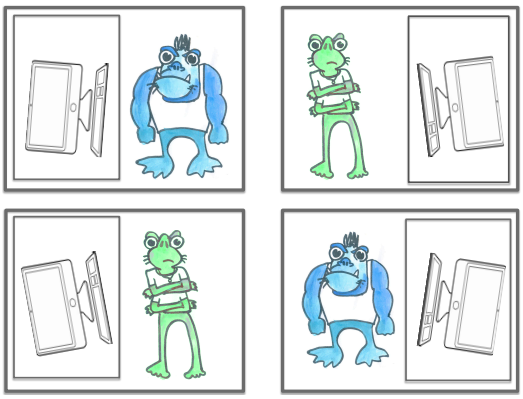


Figure 1: Players randomly assigned to species. There were two Burls and two Wiwos in each trial.

Method

Participants

Eighty students from the University of Pennsylvania participated, in groups of four, for course credit.

Procedure

Overview The experiment involved groups of four participants playing a computer game with each other. At the start of the trial, each participant was led to a separate cubicle containing a computer and was asked to log in by entering their name. Having does this, each participant was automatically assigned to one of two "alien species": the tough Burls and the weak Wiwos (Figure 1). They were then presented with the game instructions (Figure 2). As well as explaining how the game worked, these identified the species assigned to the player in question and provided information about both Wiwos and Burls, emphasizing that Burls were tougher than Wiwos and that the two species did not always get along. The difference in toughness was reinforced by images and by the names "Wiwo" and "Burl". Both the Wiwos and the Burls were explicitly told that the Burls were tougher than the Wi-

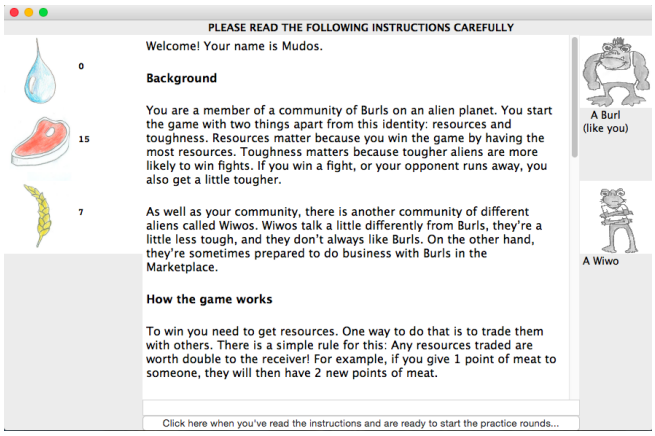


Figure 2: Instruction Screen, as presented to a Burl player.

wos, that the two species didn't always get along, and that the two species were nevertheless sometimes prepared to do business with each other in the Marketplace. In this way, we replicate the cultural setting found in Sneller (2014), where two groups may feel antagonistic toward each other but still interact. Once all participants had confirmed that they had read the instructions and had no questions, they were given an alien language to learn, presented in the form of a wordlist (see Table 1 for an example and the section on Alien Language for more details). Then they played four practice rounds before beginning the game proper, which consisted of twelve rounds. The object of the game was to win points by gaining resources from other aliens, by fighting or trading with them. The winner was announced at the end of the game. See below for more details of how the game worked.

Alien language The alien language consisted of twelve words, each consisting of two or three CV syllables and a minimal phonemic inventory of five vowels and sixteen consonants. There were two "dialects" of the language, which differed from each other with respect to both consonants (*f* vs. *b*) and vowels (*e* vs. *i* and *o* vs. *u*). Each dialect was assigned to one of the two species at random before the experiment began, so that Burls would sometimes learn words with *f* as their native variants, and sometimes with *b*. (See Table 1 for an example wordlist given to Wiwos.) We chose to vary letter representations to approximate phonological variation between the two dialects. In one dialect, *b* corresponds to *f* in the other, and central vowels *e* and *o* are raised to *i* and *u*. There were several advantages of focusing on quasi-phonological variation rather than syntactic or morphological variation. First, Sneller (2014) focused on phonological variants. Second, this simplified the task for participants, who had only to acquire words, and not syntax. Finally, borrowing across dialects is in general more likely to occur in lexical items or phonological features than in structural elements such as phonemic inventory or syntactic structure (Thomason & Kaufman, 1988). Thus by focusing on quasi-phonological features, we more closely approximate real-world cross-dialectal borrowing.

Table 1: Sample Wiwo Wordlist in "Tough Talk" Condition

hello/goodbye	buzuki (or fuzuki*)
yes	boti (or foti*)
no/not	kibo (or kifo*)
I/me	repa
you	neba (or nefa*)
have	teme
want/need	lovite
fight	bolu (or folu*)
give	viluha
water	tiluge
meat	ginuda
grain	jubu (or jufu*)
(*Tougher aliens tend to use *f* instead of *b* in words)	

Participants were exposed to the wordlists immediately after reading the game instructions, and were given approximately 14 minutes to learn the language, broken up as follows. First they had two minutes to study their wordlists. Then they played four practice rounds (each lasting approximately two and a half minutes), where they were able to practice the mechanics of the game (including chatting with each other by typing messages) with their wordlists on screen. After this they had two more minutes to study the language before the game proper began, in which participants had no access to wordlists. We note that in real languages, linguistic innovations are more frequent in informal speech registers than in formal registers (Labov, 2001). In this experiment, participants were not explicitly told anything about the register they were speaking in. However, as online chatting is an informal register (Tagliamonte & Denis, 2008), and as participants were able to engage in fighting and insulting one another, it is expected that the experiment most closely aligned with informal speech.

Game structure Each player began the game with 22 points, distributed unevenly between three resources (water, meat, and grain[2]) and shown on the left of the screen (Figure 2), and a toughness score (hidden from the player). Toughness scores varied such that, while a Burl could in principle be beaten in a fight by a Wiwo, Burls were the tougher species (a fact of which players were made aware).

The goal of the game was to obtain more points by trading resources, winning fights, or scaring off opponents. A

[2]The distribution of points over three resources had no practical significance in the game, but was designed to provide more to talk about in the Chat Stage.

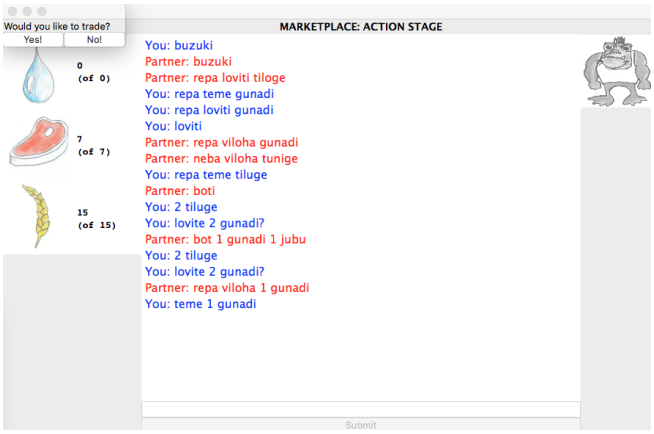


Figure 3: Beginning of the Action Stage. Chat displayed in the main box, resources displayed to the left, action choices displayed in the upper left corner, and a picture of the interlocutor displayed on the right.

game consisted of 12 rounds, each of which consisted of four stages. The first was the Setup stage, during which participants did not yet know who they would meet that round and could choose which resources they wanted to take with them to "the marketplace." After 20 seconds this stage ended and the Chat stage began, in which each participant was paired with another participant in the marketplace. The species of that alien was shown via a picture that appeared on the top right of the screen. During this stage, which lasted approximately two minutes, participants could chat with each other by typing messages using instant messaging software. At the end of the Chat Stage, the Action Stage began, in which players could decide whether they wanted to challenge their interlocutor to a fight or offer them resources in the hope of getting resources in return (see 3). They could also choose to do nothing. A participant who was challenged to a fight could choose to stand their ground or run away. If they stood their ground, the result of the fight was determined probabilistically based on a normal distribution with the player's toughness as the mean. As stated above, Burls were more likely to win fights against Wiwos, but were not guaranteed to do so.

Any resources given away were worth double to the receiver. A player who lost a fight would lose half of the resources they were carrying; a player who ran away would lose only a third. The winner of a fight would get the resources lost by the loser, multiplied by three. A player whose opponent ran away would get the resources that individual left behind multiplied by five. These values were chosen to take into account the value of leaving a potential battle unscathed. In other words, a player was better off scaring an opponent away than beating them in a fight and better off running away than losing a fight.

Following the Action Stage, participants were given feedback about the results of the round. Then a new round began. In half the rounds, Burls would meet Burls in the Chat Stage and Wiwos would meet Wiwos; in the other half, Burls would meet Wiwos. The order in which these meetings occurred, however, was randomized.

Experimental Conditions There were two conditions. In the "Burl Talk" condition, Wiwos were told that Burls sometimes used different variants. In the "Tough Talk" condition, Wiwos were told that "tougher aliens" sometimes used different variants (Table 1). In both conditions, Wiwos were exposed to the same instruction screen, which associated Burls with toughness. Thus, in both conditions the variation was implicitly associated with both Burlness and toughness; the experimental variation was in the explicit association of the variation.

In both conditions, consonantal variation was explicitly marked for the Wiwos but the vocalic variation was not. This is shown in Table 1, which displays a wordlist provided to the Wiwos. This difference in feature marking allowed us, as a secondary question, to investigate the effect of the manipulation on less salient features.

Results

Practice rounds were excluded from analysis. All words used by participants in the game proper were extracted automatically from transcripts. Because participants might use forms not included in the wordlists, as a result of typing errors, memory errors, or deliberate innovation, every word was matched automatically to the most likely intended word based on Levenshtein distance.[3] Rates of native and non-native consonant use were calculated by speaker species and interlocutor species. Given our research question, we focus here only on patterns for Wiwos. All p-values were calculated using a Wilcoxon rank-sum test.

Consonantal Variation Wiwo participants demonstrated high rates of consonantal borrowing in both conditions (Figure 4). Our hypothesis was that alienable-trait-linked features would be more frequently borrowed than features linked to an inalienable trait. In other words, we expected Wiwos to borrow Burl consonants at a higher rate under the "Tough Talk" condition than the "Burl Talk" condition. Our data supported this hypothesis: Wiwos borrowed Burl consonants significantly more in the Tough Talk condition ($\mu = 0.74$, $sd = 0.33$) than in the Burl Talk condition ($\mu = 0.47$, $sd = 0.31$) , regardless of interlocutor ($T = 2.54$, $p = 0.015$).

Furthermore, in the Tough Talk condition, Wiwos did not exhibit a difference ($T = 0.81$, $p = 0.42$) in the rates of borrowed features between chatting with a Burl interlocutor ($\mu = 0.74$, $sd = 0.33$) and a Wiwo interlocutor ($\mu = 0.65$, $sd = 0.33$). This suggests that in the Tough Talk condition, Wiwo participants treated the "tough" variant as a feature of their own dialect.

However, Wiwos did exhibit a difference in the rates of

[3]The analysis was also rerun excluding any forms not found in the wordlists; the pattern of results did not change. The same is true if practice rounds are included.

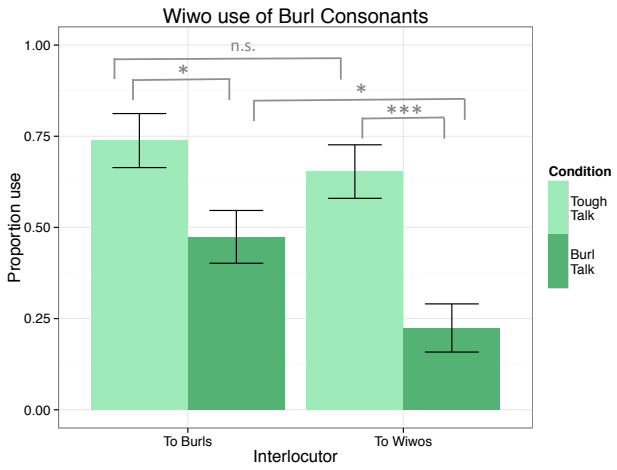


Figure 4: Wiwo use of Burl Consonants. Error bars show standard error.

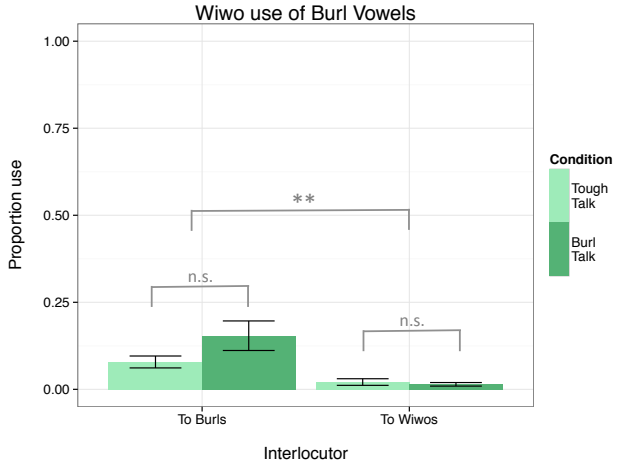


Figure 5: Wiwo use of Burl Vowels. Error bars show standard error.

borrowed features between interlocutor type in the Burl Talk condition (T = 2.55, p = 0.015). In this condition, Wiwos used significantly more Burl consonants when chatting with Burls (μ = 0.47, sd = 0.31) than when chatting with Wiwos (μ = 0.22, sd = 0.22). This behavior can be explained in terms of accommodation, whereby Wiwos in this condition were not treating the variants in question as a feature of their own dialect but rather converging with their interlocutors.

Vocalic Variation In addition to analyzing the rates of consonantal borrowing in both experimental conditions, we also analyzed the rates of vocalic borrowing. As shown in Figure 5, Wiwo participants did not engage in high rates of vocalic borrowing from their Burl interlocutors.

There are several results to be discussed in the vowel data. First, there is no statistical difference in the use of Burl vowels between the experimental conditions. To Burls in the Tough Talk condition (μ = 0.08, sd = .07) and Burl Talk condition (μ = 0.15, sd = 0.18), the difference was not significant (T = 1.6, p = 0.1). Likewise, to Wiwos in the Tough Talk condition (μ = 0.02, sd = 0.04) and Burl Talk Condition (μ = 0.01, sd = 0.02), the difference was also not significant (T = 0.6, p = 0.5). This lack of a difference between experimental conditions are consistent with our hypothesis: Since vocalic variation was associated with neither alienable nor inalienable traits, we should expect to find participants borrowing Burl vowels at the same rate in both experimental conditions. Second, Wiwos used Burl vowels more when talking to Burls (μ = 0.08, sd = 0.07) than when talking to Wiwos (μ = 0.02, sd = 0.04), regardless of experimental condition (T = 2.89, p = 0.007). This is consistent with rates of consonantal borrowing in the Burl Talk condition, as discussed above. In other words, the difference between rates of Burl vowels when chatting with Burls compared to Wiwos suggests that participants accommodated to their interlocutors even when variation was not made salient.

Discussion

We conducted an experiment in which participants learned artificial languages with quasi-phonological dialectal variation. We tested whether variants were more likely to be adopted when they were associated with an alienable traits than with an inalienable trait (species). As expected, we found higher rates of consonant adopton in the former case. These results are consistent with the patterns of real-world dialect borrowing found by Sneller (2014), who argued that White speakers who borrowed (TH)-fronting from their AAVE speaking neighbors did so because the feature had become associated with "street" identity and dissociated from urban Blackness.

An important point to be made about our study is that Wiwos can be said to be outsiders with respect to both the alienable trait (toughness) and the inalienable trait (Burls). That is, it was made clear not only that Wiwos and Burls were different species, but that the Burls were the tougher aliens. The rates of adoption of the "tough" variants by the Wiwos in the Tough Talk Condition is thus quite striking. At the same time, a limitation of the study is that participants were exposed to the variation during the learning phase of the game. In other words, the experiment can be seen as modeling a situation in which (in)alienable-trait-linked variation is already well established in the borrowing community. An important next step, currently underway, is to run the experiment without exposure to the variation during the learning phase, thus modeling an earlier stage in the spread of cultural variants.

Finally, we also note that there is an important practical difference in the benefits of Wiwos adopting the new variant between the Burl Talk and the Tough Talk condition. That is, in the Burl Talk condition, Wiwos gain no practical advantage by using the Burl features. In the Tough Talk condition, on the other hand, there is the possibility that a Wiwo using the tough features may signal actual toughness. In terms of the game, perceived toughness has a practical benefit: it may con-

vince the other player not to start a fight or even to run away if the tough-talking Wiwo starts the fight. In other words, it may be the combination of alienable trait and practical desirability that promotes the adoption of the tough features by Wiwos in the Tough Talk condition, and not the sole fact of alienability. We are currently undertaking a follow-up experiment to test whether adoption still occurs in the Tough Talk condition when there is an absence of practical benefit.

Conclusions

The aim of the study reported here was to investigate the role of two kinds of model-based bias in the adoption of cultural variants, namely quasi-phonological variants in an artificial-language game. The results of the experiment support a distinction between two types of model-based bias: *alienable* traits such as "tough" and *inalienable* traits such as race. We found support for the hypothesis that, given two hostile groups that differ with respect to both an alienable and an inalienable trait, individuals from one group are more likely to adopt cultural variants linked to the alienable than to the inalienable trait.

References

Anderson, E. (1999). *Code of the street: Decency, violence and the moral life of the inner city*. New York: W. W. Norton and Company.

Cutler, C. (1997). Yorkville crossing: A case study of the influence of hip hop culture on the speech of a white middle class adolescent in New York City. *University of Pennsylvania Working Papers in Linguistics*, *4*(1), 371–397.

Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Oxford/Cambridge MA: Wiley-Blackwell.

Eisikovits, E. (1981). *Inner-Sydney English: An investigation of grammatical variation in adolescent speech*. Unpublished doctoral dissertation, University of Sydney.

Fisher, S., & Labov, W. (2015). African American phonology in a Philadelphia community. In S. Lanehart (Ed.), *The Oxford handbook of African American language* (pp. 256–279).

Fix, S. (2010). Representations of Blackness by White women: Linguistic practice in the community versus the media. *University of Pennsylvania Working Papers in Linguistics*, *16*(2), 55–65.

Folia, V., Uddén, J., De Vries, M., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. *Language Learning*, *60*, 188–220.

Galantucci, B., & Roberts, G. (2012). Experimental Semiotics: An engine of discovery for understanding human communication. *Advances in Complex Systems*, *15*(3–4), 115–26.

Ibrahim, A. (1999). Becoming Black: Rap and hip-hop, race, gender, identity, and the politics of ESL learning. *TESOL Quarterly*, *33*, 349–369.

Labov, W. (1972). *Sociolinguistic Patterns*. Oxford/Cambridge MA: Wiley-Blackwell.

Labov, W. (2001). *Principles of linguistic change: Social factors*. Oxford: Blackwell.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, phonology and sound change*. The Hague: Mouton de Gruyter.

Richerson, P. J., & Boyd, R. (2006). *Not by genes alone: How culture transformed human evolution*. Chicago: The University of Chicago Press.

Roberts, G. (2008). Language and the freerider problem: An experimental paradigm. *Biological Theory*, *3*(2), 174–183.

Roberts, G. (2010). An Experimental Study of the Role of Social Selection and Frequency of Interaction in Linguistic Diversity. *Interaction Studies*, *11*(10), 138–159.

Sneller, B. (2014). Antagonistic contact and inverse affiliation: Appropriation of /TH/-fronting by White speakers in South Philadelphia. *University of Pennsylvania Working Papers in Linguistics*, *20*(2).

Sweetland, J. (2002). Unexpected but authentic use of an ethnically-marked dialect. *Journal of Sociolinguistics*, *6:4*, 514–538.

Tagliamonte, S., & Denis, D. (2008). Linguistic ruin? LOL! instant messaging and teen language. *American Speech*(83.1), 3–34.

Thomason, S., & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Van Herk, G. (2008). Fear of a Black phonology: The Northern Cities Shift as linguistic White flight. *University of Pennsylvania Working Papers in Linguistics*, *14*(2).

Wolfram, W. (2004). The grammar of urban African American Vernacular English. In B. Kortmann & E. Schneider (Eds.), *Handbook of varieties of English* (pp. 111–132).

Inferring priors in compositional cognitive models

Eric J. Bigelow
Dept. of Brain & Cognitive Sciences,
University of Rochester
Rochester, NY 14627
ebigelow@u.rochester.edu

Steven T. Piantadosi
Dept. of Brain & Cognitive Sciences,
University of Rochester
Rochester, NY 14627
spiantadosi@bcs.rochester.edu

Abstract

We apply Bayesian data analysis to a structured cognitive model in order to determine the priors that support human generalizations in a simple concept learning task. We modeled 250,000 ratings in a "number game" experiment where subjects took examples of a numbers produced by a program (e.g. 4, 16, 32) and rated how likely other numbers (e.g. 8 vs. 9) would be to be generated. This paper develops a data analysis technique for a family of compositional "Language of Thought" (LOT) models which permits discovery of subjects' prior probability of mental operations (e.g. addition, multiplication, etc.) in this domain. Our results reveal high correlations between model mean predictions and subject generalizations, but with some qualitative mismatch for a strongly compositional prior.

Keywords: Concepts and categories; learning; Bayesian modeling; machine learning

Introduction

Structured "Language of Thought" (LOT) models have recently become popular cognitive theories across a wide variety of domains (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp, Goodman, & Tenenbaum, 2008; Kemp, 2012; Ullman, Goodman, & Tenenbaum, 2012; S. Piantadosi, Tenenbaum, & Goodman, 2012; S. Piantadosi, Goodman, & Tenenbaum, under revision). In most of these accounts, learners are assumed to generate compositionally structured hypotheses in order to explain observed data. For instance, in Goodman et al. (2008), learners infer compositions of boolean operations and featural primitives (e.g. $RED \vee (DOTTED \wedge SMALL)$) to explain observed data, a model that explains several key phenomena in rule learning as the consequence of Bayesian rule induction. Other work models number word learning as the discovery of a counting algorithm, capturing children's developmental progression as a consequence of inferring the correct composition of operations to perform on sets (S. Piantadosi et al., 2012).

These types of LOT models typically assume fixed priors on hypotheses, which in turn provide an inductive bias for learners to prefer "simpler" compositions of primitives, consistent with behavioral tendencies (Feldman, 2000, 2003).

Of course, these priors amount to substantial assumptions about people's expectations at the start of rule learning. To test these assumptions, different particular LOTs have been tested to compare, for instance, LOT theories with distinct types of quantification or varying sets of boolean operations (Kemp, 2009, 2012; S. Piantadosi, 2011; S. T. Piantadosi, Tenenbaum, & Goodman, under review). Here, we develop a method for directly inferring the parameters of an LOT prior from behavioral data, much in the spirit of work recovering priors from behavioral data in psychophysics (Stocker & Simoncelli, 2006; Paninski, 2005). We provide a freely modifiable implementation in Python (S. T. Piantadosi, 2014) for further use and extension.

We assume that the prior parameters specify a generative model, namely a Probabilistic Context Free Grammar (PCFG). For instance, in the context of logic, we might have separate PCFG parameters corresponding to the production of a rule with disjunction ($\vee$) versus conjunction ($\wedge$). These parameters determine the relative likelihood of each operation; by inferring their values, we are able to determine how strongly subjects believe that each will be used in a novel, unobserved concept.

Our analysis technique relies on Bayesian tools, allowing us to infer both the likely parameters and the likely ranges of parameters from subjects' data. This allows us to determine exactly how much behavioral data tells us about the prior; we might discover that the behavioral data is not informative about subjects' priors, resulting in high variance in the posterior on PCFG parameters. Alternatively, we might discover that the prior probability of some but not all operations can be recovered from the data. This type of statistical inference permits inferences that are "just right" from subjects' data, indicating what a scientist should believe about otherwise unobservable cognitive operations.

The structure of this paper is as follows. First, we introduce the **Number Game**, an induction task providing a simple domain of concept-learning. Then, we discuss the structure and expressive potential of probabilistic context-free grammars as a representation for concept hypotheses. After this we present our method for Bayesian data analysis of grammar parameters and apply it to three complementary LOT formalizations.

The Number Game

We consider concept learning in the Number Game (Tenenbaum, 1999, 2000), a simple domain of cognitively-interesting induction. In the number game, concepts correspond to subsets of integers from the domain $\{1, ..., 100\}$. Subjects observe some numbers D (data) in an unobserved concept C and are asked which other numbers are in C. For instance, given observed data $D = \{16, 2, 64, 8\}$, subjects might induce that the concept used to generate these was "powers of two". This an interesting domain because this problem is under-determined, meaning that there are many solutions ("all numbers" and "even numbers" are both consistent with the data, for instance). Despite this, subjects often have strong

intuitions that some concepts are more likely.

Tenenbaum (Tenenbaum, 1999) captures these intuitions using a Bayesian model which computes $P(C \mid D)$ according to Bayes rule, or $P(C \mid D) \propto P(C)P(D \mid C)$. Here, $P(C)$ is a prior on concepts representing subjects' beliefs about likely concepts before D is observed. $P(D \mid C)$ is a likelihood model of how likely D would be if C were the true concept. This work makes a *strong sampling assumption* that the elements of D are chosen by sampling uniformly from the set specified by C. Thus

$$P(D \mid C) = \left(\frac{1}{|C|}\right)^{|D|} \quad (1)$$

This explains why, for instance, $D = \{16, 2, 64, 8\}$ suggests "powers of two" rather than "even numbers". If C is "powers of two", this D would be chosen out of the set $C = \{2, 4, 8, 16, 32, 64\}$, giving a likelihood of $(1/6)^4$. If C were $\{2, 4, 6, \ldots, 100\}$, then the likelihood would be much less, $(1/50)^4$. This critical assumption that the likelihood of data depends on the cardinality of C is known as the **size principle** and is a natural consequence of the strong sampling assumption (Tenenbaum, 1999).

Given D, inferences about whether a new number x is in the concept are made by integrating across all possible concepts:

$$\sum_C P(x \mid C)P(C \mid D) = \sum_C P(x \mid C)\frac{P(D \mid C)P(C)}{P(D)} \quad (2)$$

Here, we estimate $P(D)$ by summing over a large number of concepts, representing the vast majority of posterior probability mass (see below). In our implementation, both $P(x \mid C)$ and $P(D \mid C)$ also include a noise parameter $\alpha = 0.90$ that generates elements of C 90% of the time, and elements from $\{1, \ldots, 100\}$ uniformly 10% of the time.

The Language of Thought in the Number Game

In a formalization of learning as inductive inference over a LOT representation language, a probabilistic context-free grammar (PCFG) can be used to model concepts as compositions of simple primitives. For our purposes, each tree generated by a LOT PCFG represents a concept hypothesis C. Generation probability for a tree gives the concept's prior probability $p(C)$, calculated as $p(C) = \prod_k \lambda_k$, where λ_k is the probability of the k'th rule used to generate the tree. As in all PCFGs, this probability is conditioned on the parent of the generated node. This prior allows us to implicitly specify an infinite concept space and assign higher probability to more concise LOT expressions.

The assumed PCFG represents a hypothesis about what mental representations might be like, as well as what types of concepts people intuitively find probable. By inferring the $\{\lambda_k\}$ from data, we are assuming part of the representation (the structure of the PCFG) and discovering part (the specific probabilities) from human data. Note that we may find that some λ_k are close to zero, meaning that we have assumed rules which are not psychologically justifiable. Additionally, we may write down different PCFGs and compare their performance in explaining human behavior. This allows us to deduce some further lessons about the structure of the PCFG over and above fitting probabilities. In particular, we consider three grammars capable of modeling number game concepts (see Table 1). Next, we describe these models in more detail. We will consider as an example the prior associated with the "odd numbers" concept, for each of the three grammars.

Independent Model Grammar	**Compositional Model Grammar**
$Start \xrightarrow{\lambda_0} Math$	$Start \longrightarrow Set$
$Start \xrightarrow{1-\lambda_0} Interval$	$Set \xrightarrow{\lambda_0} Set \cup Set$
	$Set \xrightarrow{\lambda_1} Set \cap Set$
	$Set \xrightarrow{\lambda_2} Set \setminus Set$
	$Set \xrightarrow{\lambda_3} Math$
	$Set \xrightarrow{\lambda_4} Interval$
$Math \xrightarrow{\lambda_{1-9}}$ Powers of $n \quad 2 \le n \le 10$	$Math \longrightarrow Map(\lambda x.Expr, Interval)$ [1]
$Math \xrightarrow{\lambda_{10-19}}$ Multiples of $n \quad 3 \le n \le 12$	$Expr \xrightarrow{\lambda_5} Expr \cdot Expr$
$Math \xrightarrow{\lambda_{20-29}}$ Ends with $n \quad 0 \le n \le 9$	$Expr \xrightarrow{\lambda_6}$ Ends with$(Expr, Expr)$
$Math \xrightarrow{\lambda_{30-39}}$ Contains Digit $n \quad 0 \le n \le 9$	$Expr \xrightarrow{\lambda_7}$ Contains Digit $(Expr, Expr)$
$Math \xrightarrow{\lambda_{40}}$ Prime numbers	$Expr \xrightarrow{\lambda_8}$ Prime$(Expr)$
$Math \xrightarrow{\lambda_{41}}$ Even Numbers	$Expr \xrightarrow{\lambda_9} Expr^{Expr}$
$Math \xrightarrow{\lambda_{42}}$ Odd Numbers	$Expr \xrightarrow{\lambda_{10}} Expr + Expr$
$Math \xrightarrow{\lambda_{43}}$ Squares	$Expr \xrightarrow{\lambda_{11}} x$
$Math \xrightarrow{\lambda_{44}}$ Cubes	$Expr \xrightarrow{\lambda_{12}} Const_E$
	$Const_E \xrightarrow{\lambda_{13-27}} n \quad 1 \le n \le 15$
$Interval \xrightarrow{1.0}$ Range$[m, n]: \ 1 \le m \le n \le 100$	$Interval \longrightarrow Range[Const_I, Const_I]$
	$Const_I \xrightarrow{1.0} n \quad 1 \le n \le 100$ [2]

Table 1: Three PCFG constructions: a Mixture Model (not shown), a simple model with independent probabilities for each rule ("Independent Model"), and a recursive compositional rule model ("Compositional Model"). The number above each arrow gives the *unnormalized* probability of each rule expansion. The Mixture Model Grammar is identical to the Independent with all parameters λ_k are fixed to 1.0 except λ_0.

Mixture Model

The **Mixture Model** implements a very simple PCFG which combines two types of grammatical productions: those generating **mathematical rules** and those generating ranges of numbers, called **interval-based** concepts. This setup follows Tenenbaum (Tenenbaum, 1999, 2000), who constructed a concept hypothesis space for the range of numbers 1 through 100 according to representative concepts generated by additive clustering within the domain of numbers 1 through 10 (Tenenbaum, 1996). The only distinctions between this concept hypothesis space and that of our model, are that our space includes two additional types of rule-based concepts: "Ends in n" and "Contains Digit n", and that where Tenenbaum assigns an Erlang prior across interval concepts, we assign a uniform prior for simplicity.

The Mixture Model grammar has a single parameter λ_0 which determines the relative probability of using a rule-based or a interval-based grammatical production. Once the type of the production is chosen, the specific production is

[1] While other λ here refer to parameters, this refers to the abstraction operator λ.

[2] In order to narrow the extremely broad hypothesis space of the compositional model, priors of 5.0 are assigned to constants $1, 10, 20, 30, \ldots, 90, 100$.

chosen uniformly as in (Tenenbaum, 1999, 2000). For example, to generate an "odd numbers" concept, we begin at *Start* and traverse to *Math* with probability λ_0. Following the allowed rules in Table 1, we expand *Math* to "Odd Numbers" with probability $1/44$, yielding an overall prior of $p(C) = \frac{\lambda_0}{44}$.

Independent Probabilities Model

The **Independent Probabilities grammar** generates the same set of concepts as the Mixture Model grammar, but with the key difference that each rule-based concept is produced by a rule with an independent probability parameter λ_k. This grammar also has a mixture parameter λ_0 that biases the grammar towards interval- or rule-based concepts. As in the Mixture Model, interval concepts are assigned a uniform prior. The parameter space for the Independent Model allows a more intricate representation of individuals' priors, where relative bias associated a priori with particular concepts can be inferred. For example, in the Independent Model we might infer that an individual has a stronger bias associated with the concept "multiples of 4" over the concept "multiples of 7".

To generate the "odd numbers" concept, the same rules are used to generate the concept as in the Mixture Model but with different associated priors: $p(C) = \frac{\lambda_{42}}{\sum_{k=1}^{44} \lambda_k}$.

Compositional Model

The **Compositional grammar** is a recursive LOT model that expresses concepts by freely composing primitives, based on the general approach of (Goodman et al., 2008). The Compositional Model has a significantly wider concept hypothesis space, and its parameter space reflects the probability of using each primitive operation. These primitives include values and operations that can compose to create a range of numerical concepts, including all those of the other two models, as well as set operations between mathematical expressions and intervals. For example, one might infer that humans have a particularly high bias associated with the "times" operator, and a very low prior with the "powers of n" operator. Note that by the assumed statistical structure of a PCFG, every place that "times" is used, it will have the same probability. Thus, despite having a richer concept space, the Compositional grammar cannot model some of the concept-specific priors of the Independent Model — e.g. the context free assumptions of the PCFG mean that 2 in $(2 \cdot x)$ and 2 in $(2 + x)$ both are equally likely (i.e. irrespective of their use in the context of addition or multiplication).

This grammar requires a considerably deeper tree to represent the "odd numbers" concept. From *Start* we go to *Set*, then *Math* with a probability λ_3, which maps an *Expr* across the range of integers 1 to 100. This *Expr* goes to $Expr + Expr$ with probability λ_{10}; one of the two *Expr* terminates at 1 with probability λ_{13}, and the other goes to $Expr \cdot Expr$. This final expression terminates with x and 2, with respective probabilities λ_{11} & λ_{14}. In plain English, "multiples of 2, plus 1, for integers one to one-hundred." While this may seem complicated, an advantage of this model is its capacity to capture a wide range of complexities with fewer free parameters. For example, the following three concepts are ordered by increasing complexity when generated by the Compositional grammar: "powers of 2", "powers of 2 plus 1", "only primes from the set of powers of 2 plus 1"

Bayesian Data Analysis

The primary contribution of this paper is to apply data analysis techniques to an interesting domain of inductive concept learning in cognitive science, assessing parameters of the prior, $\{\lambda_k\}$, given subjects' behavioral data. The prior is the most psychologically laden aspect of LOT models because it specifies people's assumptions without any data, as well as—in compositional models—their subjective expectations about how likely each operation is to be used. Similar data analysis has been used previously to infer the parameters of the prior's PCFG (S. Piantadosi, 2011; S. T. Piantadosi et al., under review), although largely with the goal of comparing different languages. Here, we focus on the complementary question of inferring in detail the relative expectations among different primitives, or different classes of primitives.

To do this, we present a method of Bayesian data analysis (Kruschke, 2010) to infer a posterior distribution on each rule probability, given the human responses. We'll use y_i to represent the number of "yes" counts for whether a query x_i (e.g. $x_i = 16$) is in the concept, given data D_i (e.g. $D_i = \{2,4,8\}$). We'll use N_i to denote the total number of human responses for query x_i, fixed by experimental design. Let $G = \{\lambda_k\}$ be the set of rule production probabilities. For a given G, we implicitly define a space of concept hypotheses that may be generated by this grammar. We wish to find

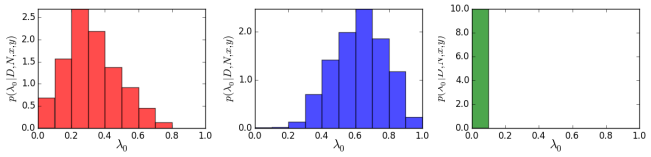
$$\begin{aligned} P(G \mid y, D, N, x) &\propto P(G) P(y, \mid G, D, N, x) \\ &= P(G) \prod_i P(y_i, \mid G, D_i, N_i, x_i). \end{aligned} \tag{3}$$

Where the product comes from assumed independence of responses, conditioned on D_i and x_i. In this equation, $P(G)$ is a prior on parameters. We assign a $Gamma(2,1)$ prior to each λ_k, representing its *unnormalized* probability. Human responses are then assumed to come from a binomial likelihood $P(y_i, \mid G, D_i, N_i, x_i)$,

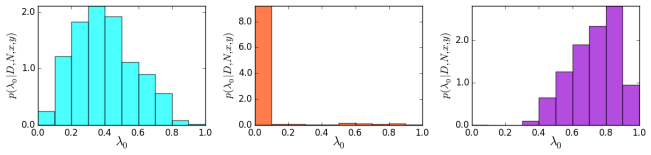
$$\binom{N_i}{y_i} [P(y_i, \mid G, D_i, N_i, x_i)]^{y_i} [1 - P(y_i, \mid G, D_i, N_i, x_i)]^{N_i - y_i} \tag{4}$$

Together, these equations specify a data analysis model over grammar probabilities that can take human inferences (e.g. 9 out of 10 people believe 16 is in the concept $\{4,8\}$ came from), and work backwards to discover the relative probability of compositional PCFG components behind these.

For the Compositional Model, we use a Markov Chain Monte Carlo sampling technique, the Metropolis-Hastings (MH) algorithm, to generate a representative subset of the infinite concept hypothesis space when computing the marginal (integral) (2). In computing the marginal, we ran 255 independent MH chains at 300,000 steps each, with each chain conditioned according to a unique concept example D from our data. All chains were initialized with the same uniform



(a) 15 in concept? $\langle 3\ yes, 3\ no \rangle$ (b) 15 in concept? $\langle 3\ yes, 30\ no \rangle$ (c) 15 in concept? $\langle 30\ yes, 3\ no \rangle$



(d) 64 in concept? $\langle 3\ yes, 3\ no \rangle$ (e) 64 in concept? $\langle 3\ yes, 30\ no \rangle$ (f) 64 in concept? $\langle 30\ yes, 3\ no \rangle$

Figure 1: Posterior distribution of parameter space for the Mixture Model grammar $p(\lambda_0 \mid D, x, y, N)$. $D = \{16\}$ for each example here, and x, y, N are each single-item sets, for either $x = \{15\}$ (top row) or $x = \{64\}$ (bottom row).

prior G_0 (approximately uniform, with higher probability assigned to terminal rules). The 1000 hypotheses with highest posterior scores for each chain (thus containing nearly all posterior probability mass) were all joined by union, yielding a total set size of 9,946. This formed a finite hypothesis space that well-approximated each posterior ($P(C \mid D)$).

MH was also used to infer the grammar parameters G in 3 for all models. Ten chains each were run for the Compositional and Independent Models, each for 50,000 steps, with 50,000 steps of burn-in; convergence was usually observed for these after less than 20,000 steps. Convergence in the Mixture Model (which has only one parameter) was usually reached after less than 500 samples were drawn, so four chains for this model were each run for 10,000 steps.

An example statistical inference

To demonstrate our approach, we first apply this data analysis to a simple toy example. Imagine you see a single example from a concept: $\{16\}$ and are asked whether 15 or 64 are part of the same concept. 15 is close to 16 in magnitude, the metric of "similarity" considered by Tenenbaum. On the other hand, 64 can be grouped with 16 according to numerical rules—for example, "even numbers", "powers of two", "powers of 4", or "perfect squares" are all candidate concepts.

We ran this simulated data set on the Mixture Model, shown in Table 1. We then used our data analysis to infer this model's single λ_0 parameter from a few data sets with intuitive answers, serving as a proof of principle for our data analysis methods. For instance, if subjects assume 15 is in the concept (given only that they know 16 is), we should infer that λ_0 is low, meaning that similarity (distance) based concepts are the most psychologically salient ones. On the other hand, if our simulated data has that subjects believe 64 is likely instead, we should see a *high* λ_0, indicating that the psychologically highest prior concepts are rule-like.

Figure 1 shows the posterior distribution on the λ_0 for six artificial datasets. Figure 1a shows the distribution on λ_0 if we find 3 subjects saying "yes" and 3 saying "no" to 15, conditioned on $\{16\}$ being in the concept. In this case, λ_0 is biased towards 0, representing a preference for interval-based concepts. A bias emerges here because mathematical concepts have trouble explaining this data, so little is gained by assigning them a high prior. However, if most subjects say "no" to 15 (Figure 1b), the model correctly implies that λ_0 must be higher: rejecting 15 means that people likely down-weight the prior probability of ranges, many of which include 15.

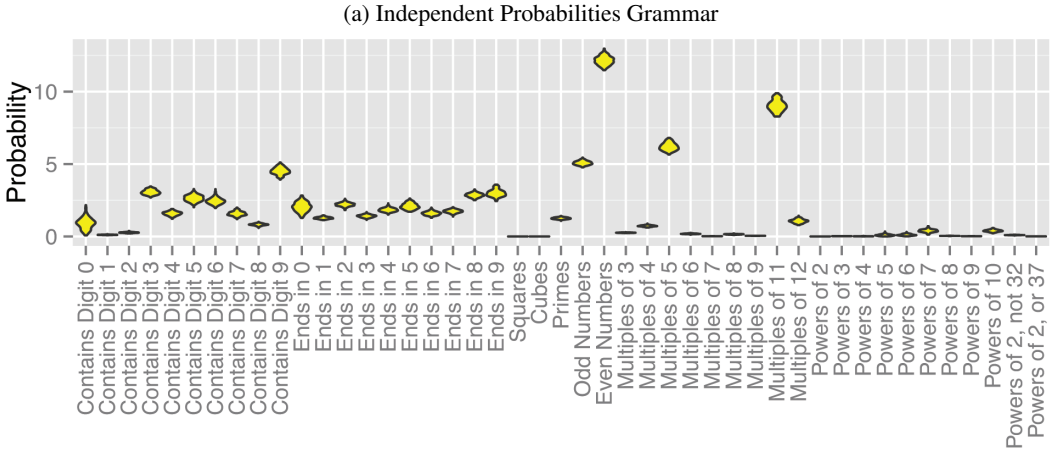
Therefore, more "yes"es than "no"s will indicate that interval-based concepts probably have a high prior bias. We see the posterior distribution in Figure 1c reflects this expectation, since the distribution over mixture ratios is greatly skewed towards interval-based concepts (low λ_0). We also see this skew towards interval-based concepts in Figure 1a, but with a more uniform distribution. With very few data points, it is unclear whether these responses are truly reflective of the prior, or simply result from noise. With a more proportional mixture ratio, we should see fewer "yes"es than "no"s due to an initial preference towards rule-based concepts, as reflected in the distribution of Figure 1b.

Imagine you are instead asked whether 64 is in the target concept given $D = \{16\}$. Maybe the pattern is "square numbers" or "powers of 2" - each of these has a small set size in the domain of 1 to 100, so it is likely that 16 would be output if one of these was the concept. Since "rule-based" concepts have such small set sizes, these are normally preferred over interval concepts given very little data (Tenenbaum, 1999). We therefore expect that with a proportional bias between interval and rule concepts, a majority of observers should assume 64 is in the target concept given $D = \{16\}$, as seen in Figure 1f. This will not be the case only when interval concepts are given a strong bias over rule concepts (Figure 1e).

Analysis of a large-scale Number Game experiment

We analyzed a large dataset of human responses for our experiment, comprised of 606 participants with demographics typical of Amazon Mechanical Turk workers (Bigelow & Piantadosi, 2016). Subjects were tested on input sets D that were generated by using a small set of "primordial" sets (all integers, evens, odds, squares, cubes, and primes) and then mapping a variety of functions (e.g. $f(n) = n+1$, $f(n) = n+2$, $f(n) = n-1$, $f(n) = 2 \cdot n$, etc.) across each. Data sets D were generated by randomly sampling sets of length 2, 3, and 4 from each resulting concept. On each trial, participants were told a set D was generated by a specific program, then indicated whether it was likely the program would generate each target from a random sequence of 30 numbers in the range $\{1, \ldots, 100\}$. No feedback was given as to the correctness of responses. Responses were trimmed for the 31 subjects with the lowest total typicality values - measured as $log(p)$ for "yes" responses and $log(1-p)$ for "no", where p is the fraction of positive responses - yielding a total 258,750 generalizations across 255 unique concepts.

Figure 2 shows an overall model summary of our results, giving the relationship between the model's predictive distri-



bution (x-axis) and humans' probability of responding "yes" (y-axis), to each of the generalizations in the data. The left column shows the initial grammar parameters — i.e., before $\{\lambda_k\}$ has been fit to human data; the right column shows the MAP grammars. This plot makes clear several aspects of our data. First, the correlations are high in the left column, meaning that the default model fits the human data relatively well. However, the model fit does improve in the right column, indicating that there is some variance to be gained by adjusting these parameters. Note that in all subplots the reported R^2 valued are on the box plot means, meaning that they do *not* represent the full variance in the data. The box plot also shows that there is considerable variance in responses for each bin, meaning that there is a substantial variability over and above the means that the model does not capture.

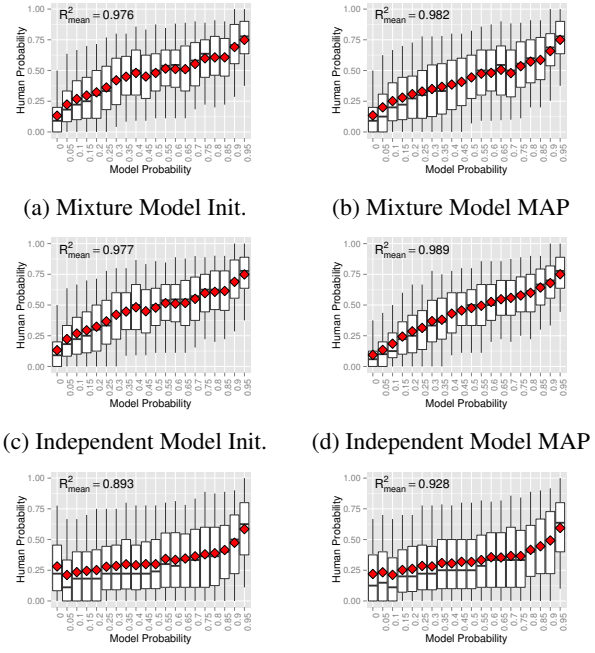


Figure 2: Box plots showing the distribution of human responses (y) binned by model predictions (x). Model predictions are shown for initial (a,c,e) and MAP (b,d,f) grammar priors. Red diamonds show mean values; box plots show 25th, 50th, & 75th percentiles, and whiskers show 5th and 95th. Correlations were computed on mean values; box plots show high variance over and above these.

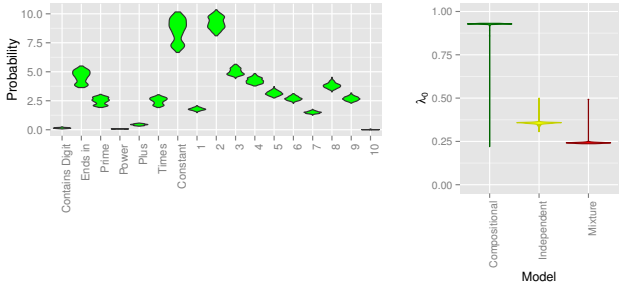


Figure 3: Posterior distribution of unnormalized grammar production probabilities λ_k. Bar width in this graph represents sample density relative to a specific parameter λ_k. Note that in the Independent Model, all rules shown are normalized relative to each other. In the Compositional Model, constant primitives (1-10) are weighted relative to other constants, and operators ("Ends in", "Times", "Constant") are weighted relative to other operators (see Table 1).

The overall shape of these plots tells us whether the model makes qualitatively correct predictions. The Mixture Model does well in this respect, as does the Independent Model, although the Independent Model must be interpreted with care since it freely fits the prior on a large number of concepts. The initially good fits of the Mixture and Independent Models indicate that an approximately uniform prior across these models' hypothesis spaces fits the data well; clearly, inferring MAP G will improve fit for some models more than others.

The Compositional Model seems to miss the qualitative match, which means falling far from the line $y = x$ even though the general increasing trend yields a respectable correlation. The qualitative mismatch is particularly salient for model predictions < 0.5, meaning the Compositional Model seems to systematically mischaracterize people's likely "no" responses. This may be because the form of the prior is highly constrained in the Compositional Model to follow the PCFG. For example, assigning a high prior to the concept "odd numbers" $(2n+1)$ requires high priors on its four components "times", "plus", "2", and "1". But high priors on these would also assign high prior to $1n+2$ ("numbers greater than two") due to the independence assumptions of a PCFG. It is possible that our Compositional Model performs poorly because we have included the wrong primitives or wrong structure in our particular PCFG. Fits such as these may in principle be used to quantitatively compare hypothesized LOTs.

These results highlight the impact of grammar design on determining the hypothesis space for models, and we expect this to be an area of future work. Another area of future work may be to explore the limitations of PCFGs. It may be that using a context-sensitive grammar will be necessary to fully capture human data, or models of another family entirely.

The most interesting aspect of our analysis is the posterior distribution on grammar productions for each model, giving what we should believe about people's relative probability of each kind of operation. Figure 3 shows these posterior distributions. The independent probabilities model shows that human data is best fit when the highest priors are associated with "Even Numbers", "Multiples of 11" (numbers with repeated digits), "Multiples of 5", "Odd numbers." Interestingly, nonzero weight is also assigned to "Primes", "Multiples of 12", and "Multiples of 4".

In the Compositional Model, human data is best fit by high priors associated with the "Ends In", "Prime", and "Multiples" ($a \cdot b$) operations. The highest prior among operations is associated with "Constant", which is to be expected as this allows for smaller hypothesis trees. Among the constants, the highest prior is associated with 2, followed by a general decreasing trend for large numbers that is reminiscent of the frequency distribution of numbers in language (Dehaene & Mehler, 1992; S. T. Piantadosi, in press).

Figure 3c shows that λ_0 is much higher for the Compositional Model than the others, as to be expected since many of its hypotheses are redundant with interval concepts (e.g. $n+2$ corresponds to "numbers greater than 2", equivalent to the interval $[3,100]$). The Independent Model's λ_0 is higher than the Mixture Model's, which is also expected as λ_k for each rule-based concept was conditioned on the same data.

Conclusion

Our work has shown how a structured inductive LOT model may be combined with a Bayesian data analysis to infer likely parameters of human subject's priors. Our analysis has revealed both an ability to freely infer priors on concepts (the Independent Model) as well as those that use the PCFG more productively, decomposing a concept's prior into a product of the priors of its parts (the Compositional Model). In these cases, Figure 3 represents our inference of these values from subject data. While the mean predictions of the Compositional Model are highly correlated with human means, the qualitative fit is worse. However, the method developed here of analyzing large-scale experiments with Bayesian data analysis provides a way forward for fitting, refining, and comparing LOT models of human cognition.

References

Bigelow, E., & Piantadosi, S. (2016). A large dataset of generalization patterns in the number game. *Journal of Open Psychology Data*, *4*(1).

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, *12*(6), 227.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

Katz, Y., Goodman, N., Kersting, K., Kemp, C., & Tenenbaum, J. (2008). Modeling semantic cognition as logical dimensionality reduction. In *Proceedings of Thirtieth Annual Meeting of the Cognitive Science Society.*

Kemp, C. (2009). Quantification and the language of thought. *Advances in neural information processing systems*, *22*.

Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, *119*, 685–722.

Kemp, C., Goodman, N., & Tenenbaum, J. (2008). Learning and using relational theories. *Advances in neural information processing systems*, *20*, 753–760.

Kruschke, J. (2010). *Doing bayesian data analysis: A tutorial introduction with r*. Academic Press.

Paninski, L. (2005). Nonparametric inference of prior probabilities from bayes-optimal behavior. In *Advances in neural information processing systems* (pp. 1067–1074).

Piantadosi, S. (2011). *Learning and the language of thought.* Unpublished doctoral dissertation, MIT.

Piantadosi, S., Goodman, N., & Tenenbaum, J. (under revision). Modeling the acquisition of quantifier semantics: a case study in function word learnability.

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, *123*, 199–217.

Piantadosi, S. T. (2014). *LOTlib: Learning and inference in the language of thought.* available from https://github.com/piantado/LOTlib.

Piantadosi, S. T. (in press). A rational analysis of the approximate number system. *Psychonomic Bulletin and Review*.

Piantadosi, S. T., Tenenbaum, J., & Goodman, N. (under review). The logical primitives of thought: Empirical foundations for compositional cognitive models.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, *9*(4), 578–585.

Tenenbaum, J. B. (1996). Learning the structure of similarity. *Advances in neural information processing systems*, 3–9.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning*. Doctoral dissertation, Massachusetts Institute of Technology.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, *12*, 59–65.

Ullman, T., Goodman, N., & Tenenbaum, J. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*.

Which is in front of Chinese people: Past or Future?
A study on Chinese people's space-time mapping

Yan Gu[1] (yan.gu@tilburguniversity.edu)

Yeqiu Zheng[2] (yeqiu.zheng@tilburguniversity.edu)

Marc Swerts[1] (m.g.j.swerts@tilburguniversity.edu)

[1]Tilburg center for Cognition and Communication (TiCC), Tilburg University, the Netherlands
[2]Department of Econometrics and Operations Research, Tilburg University, the Netherlands

Abstract

Recent research shows that Chinese, when they gesture about time, tend to put the past "ahead" and the future "behind". Do they think of time in the way as suggested by their gestures? In this study we investigate whether Chinese people explicitly have such past-in-front mappings. In experiment 1 we show that when time conceptions are constructed with neutral wording (without spatial metaphors), Chinese people are more likely to have a past-in-front-mapping than Spaniards. This could be due to cultural differences in temporal focus of attention, in that Chinese people are more past-oriented than Europeans. However, additional experiments (2 & 3) show that, independent of culture, Chinese people's past-in-front mapping is sensitive to the wording of sagittal spatial metaphors. In comparison to a neutral condition, they have more past-in-front mappings when time conceptions are constructed with past-in-front spatial metaphors ("front day", means the day before yesterday), whereas fewer past-in-front mappings are constructed with future-in-front metaphors. There thus appear to be both long-term effects of cultural attitudes on the spatialization of time, and also immediate effects of the space-time metaphors used to probe people's mental representations.

Keywords: cross-cultural differences; space and time; conceptual metaphor; Chinese; Temporal Focus Hypothesis

Introduction

Across cultures people use space to represent time. The conceptions of future and past are often linguistically expressed by the use of spatial metaphors. For instance, in English, we look *forward* to the bright future lying *ahead*, or look *back* to the hard times *behind* (e.g., Clark, 1973; Lakoff & Johnson, 1980). Interestingly, studies have shown that people not only talk about time using a front-back axis, but also tend to think about time this way, i.e., the future is mentally "ahead" of the speaker, and the past is "behind" (Boroditsky, 2000; Miles, Nind & Macrae, 2010; Ulrich et al., 2012). This seems to be consistent with the bodily experience of walking in a certain direction, so that the path that we have passed by is the past and the place that we are heading towards is the future (e.g., Clark, 1973).

However, speakers of some languages exhibit the opposite space-time mapping in the sagittal axis. For example, in Aymara, the spatial words for front and back literally mean past and future (e.g., "front year" means last year, "back year" means next year). This past-in-front mapping is also apparent from Aymara speakers' spontaneous temporal gestures (Núñez & Sweetser, 2006).

Additionally, cultures may vary about the degree to which they associate space to time. A recent study showed that different spatial-temporal mappings between Moroccans and Spaniards could be related to cross-culture differences in temporal focus (*Temporal Focus Hypothesis*). The Temporal Focus Hypothesis demonstrates that the space-time mappings in people's minds are conditioned by their cultural attitudes towards time. It is claimed to be dependent on attentional focus and can be independent from the space-time mappings enriched in language. For instance, despite the fact that front-back time metaphors in Arabic are similar to Spanish and English (future-in-front mappings), Moroccans have a strong past-in-front mapping when asked in a temporal diagram labelling experiment, whereas the majority of Spaniards have a future-in-front mapping. It has been argued that Moroccans are found to focus more on past times and old generation, and place more value on tradition in comparisons to Americans, Spaniards and other Europeans. Interestingly, that study also reveals that the focus of attention on past or future may play a role in determining the spatializing of time in people's minds. For example, after performing a short writing exercise that induces participants' focus of attention on the past, half of the Spaniards perform a past-in-front mapping, the proportion of which is higher than those without having the writing exercise (de la Fuente et al., 2014).

Although linguistic, cultural and bodily experiences have been found to have separate influences on people's spatial representation of time (e.g., Boroditsky, 2001; Fuhrman & Boroditsky, 2010; Núñez & Sweetser, 2006; Saj et al., 2014), we still have limited knowledge on why some communities adopt a future-in-front mapping whereas others a past-in-front mapping for time. For instance, Chinese people show a strong tendency to gesture according to the past-in-front mapping (Gu et al., in preparation), whereas English and Dutch people have an opposite tendency in the sagittal axis (Casasanto & Jasmin, 2012). Based on the gesture data, it would appear that Chinese speakers can think of time as the Aymara do. However, gestures about time are not only shaped by temporal thinking, but also by lexical choices (Gu et al., 2014). Given that Mandarin speakers sometimes also verbally produce "前 / qián" (front) (e.g., qián-tiān, front day, the day before yesterday) when they gesture about a past event, it could be that they just perform a forward

gesture to make it congruent with the word "qián" (front, before). Therefore, a more explicit approach will be helpful to bring a clearer picture.

Furthermore, as for the temporal focus in Chinese culture, most studies find that Chinese people are primarily past-oriented. For example, Chinese attend to a greater range of temporal information in the past than do European Canadians (Ji et al, 2009), and they perceive objects in the past as being much more valuable than their American counterparts do (Guo et al, 2012). If the cross-cultural differences in temporal focus indeed predict a different space-time mapping (*Temporal Focus Hypothesis*, de la Fuente et al., 2014), we will expect that Chinese people are more likely to have a past-in-front mapping than European people (e.g., Spaniards).

The purpose of this paper first is to find out whether Chinese people indeed have a past-in-front mapping. If they do, then we further explore possible accounts for that mapping.

Experiment 1: Do Chinese people place the past events in front?

Rather than the gesture approach that tested implicit knowledge of sagittal timeline (Gu et al., in preparation), in experiment 1 a more explicit paradigm was used to examine whether Chinese people have a past-in-front mapping. This paradigm has been used to test Moroccans and Spaniards to study the cross-cultural differences in space-time mapping at the sagittal axis (de la Fuente et al., 2014), and we use their Spanish data as comparison materials.

Method

Participants

38 Mandarin speakers participated in the experiment. They were tested in Rizhao, China, and all materials were in Mandarin.

Materials and Procedure

Participants performed a temporal diagram task adapted from de la Fuente et al., (2014, Experiment 1). Participants sat at a table and they saw a toy doll (named Xiaoming) with one box in front of the toy and one box behind it. Participants and the character faced the same direction in the sagittal plane (Fig. 1). Participants read that yesterday (昨天, zuó-tiān) Xiaoming went to visit a friend who liked eating apples, and tomorrow (明天, míng-tiān) he would be going to visit a friend who likes eating pears (or vice versa, depending on the version of the task the participant received). Participants were given an apple and a pear and were instructed to put the "apple" in the box that corresponded to *what happened at an earlier time* and the "pear" to the box that corresponded to *what would happen at a later time*. The order of mentioning of the apples and pears was counterbalanced, as were their pairs with "yesterday" and "tomorrow". Note that the temporal expressions (i.e., yesterday, tomorrow, earlier, later) in the instructions consisted of neutral wording in a sense that they had no spatial metaphors.

Instead of doing the task on paper (de la Fuente et al., 2014), we asked participants to do the task with real entities. This can not only record how participants fulfil the task, but also minimize the potential projection of vertical timeline into the sagittal axis (as in Chinese there are vertical spatial metaphors of "up" and "down" representing the time conceptions of "early" and "late"). Each participant individually did the task with the experimenter in a quiet room. After the task, s/he was given a questionnaire to fill in some background information such as gender and age. Participants were paid a small fee and signed a consent form.

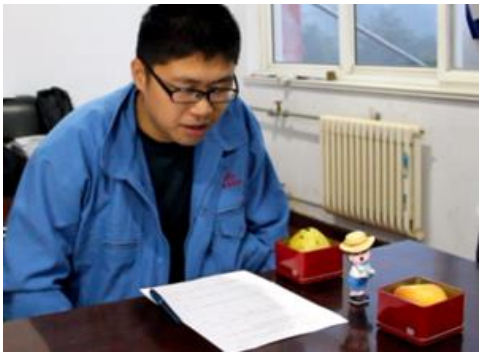


Figure 1: Setting up of Experiments 1-3.

Results and Discussion

36.8% of participants responded according to the past-in-front mapping, placing the past event in the box in front of the character and the future event in the box behind it. This rate was not significantly different from chance, $p = .14$ (a sign test, $N = 38$), which suggests that Chinese people may have no bias for the past/future-in-front mapping. In comparison to the Spaniards (12%) in de la Fuente et al's (2014) study, Chinese people were significantly more inclined to place the past in front of the character, as revealed by a binary logistic regression, Wald χ^2 (1, $N = 88$) = 6.98, $p = .008$, odds ratio = 4.28, 95% confidence interval (CI) = [1.46, 12.57].

The space-time mappings shown by this diagram task confirmed the mappings that were previously observed in native Mandarin speakers' spontaneous hand gestures (Gu et al, in preparation) and forced pointing gestures (Lai & Boroditsky, 2013): some Chinese appear to conceptualise time according to a past-in-front mapping. Furthermore, the cross-cultural differences between Chinese and Spaniards seem to indicate a long-term effect of cultural attitudes on the spatialization of time, as predicted by the Temporal Focus Hypothesis (de la Fuente et al, 2014).

In experiment 1, the temporal expressions in the instructions were constructed with neutral words. However, in Chinese very often the wording of the temporal conceptions of "the past" and "the future" contained the lexicons of "前 / qián" (front, before) and "后 / hòu" (back, after), which share the same lexicons with the spatial location of front and back (like Aymara speakers). That means, Chinese people can use past-in-

front/future-in-back spatial metaphors to express time (e.g., 后天 / hòu-tiān, back day, the day after tomorrow; 今后 / jīn-hòu, from today back, from now on). If spatial metaphors for time can have an immediate effect on people's mental representations (e.g., Boroditsky, 2000; Lai & Boroditsky, 2013), we expect that Chinese people will have more past-in-front mapping when the temporal relations are expressed with such explicit spatial markers (e.g., use "front day" and "back day" rather than "yesterday" and "tomorrow"; use "to front" and "from now back" rather than "an earlier time" and "a later time"), with a comparison to the result in Experiment 1.

Experiment 2: Does the spatial lexicon matter: past-in-front language

Method

Participants

A new group of 37 Mandarin speakers participated in the past-in-front metaphor condition. They were tested in Rizhao, China, and all materials were in Mandarin.

Materials and Procedure

Participants followed the same procedure to fulfil the temporal diagram task as described in experiment 1. However, the instructions about the temporal task were different from those in experiment 1.

First, in the statement they now read that the day before yesterday (前天, qián-tiān, front day) Xiaoming went to visit a friend who liked eating apples, and the day after tomorrow (后天, hòu-tiān, back day) he would be going to visit a friend who likes eating pears. This new pair of temporal constructs have a similar period of time unit as the pair of "yesterday" and "tomorrow", both being one or two days in reference to now. Furthermore, both pairs convey a clear contrast between the past and future time conceptions while the new pair has past-in-front / future-in-back spatial metaphors.

Second, the neutral words of "future" and "past" events in the task instruction were replaced with time conceptions consisting of spatial words. Specifically, participants were instructed to put the "apple" in the box that corresponded to the past (以前, yǐ-qián, to front, before) events and the "pear" to the box that corresponded to the future (今后， jīn-hòu, now back, from now on) events (or vice versa) (Table 1).

Results and Discussion

Interestingly, for the past-in-front metaphor condition, there were 57% of participants who responded according to the past-in-front responses. This rate was not significantly different from chance (a sign test, $p > 0.05$, $N = 37$), which may indicate that Chinese people probably do not have a bias for the past or future in-front mapping when primed by the past-in-front metaphors. However, the proportion was significantly higher than that of 36.8% in the neutral condition (Experiment 1), Wald χ^2 (1, N = 75) = 2.95, p =.086 (two tailed), odds ratio = 2.25, CI = [.89, 5.68].

The results showed that the space-time mapping was sensitive to the spatial lexical choices. When temporal conceptions were constructed with past-in-front spatial metaphors (spatial words "front" and "back" for the past and future conceptions), participants were more likely to perform a past-in-front mapping than temporal conceptions that were constructed with neutral wording. The result is consistent with previous finding on spontaneous gestures that Chinese tend to produce past-in-front temporal gestures when they are using past-in-front spatial metaphors (Gu et al, in preparation).

Table 1: Instructions for Experiments 1, 2 & 3.

Yesterday (Exp 1 & 3) / *The day before yesterday* (Exp 2) Xiaoming went to visit a friend who liked eating apples, and *tomorrow* (Exp 1 & 3) / *the day after tomorrow* (Exp 2) he would be going to visit a friend who likes eating pears. There are two boxes near Xiaoming. Please put the "apple" in the box that corresponds to [past: *what happened at an earlier time* (Exp 1) / *to front* (Exp 2) / *pass go* (Exp 3)] and the "pear" to the box that corresponds to [future: *what would happen at a later time* (Exp 1) / *now back* (Exp2) / *hasn't yet come* (Exp 3)].

Table 2: Examples of Mandarin Chinese phrases showing a Future-in-Front and a Past-in-Back Mapping.

(1)	展	望	未	来
	zhǎn	wàng	wèi	lái
	unfold	gaze-into-distance	hasn't	come
	Looking into the future			
(2)	回	首	过	去
	huí	shǒu	guò	qù
	turn-around	head	pass	go
	Looking back to the past			

Nevertheless, Chinese do not exclusively use lexical cues to associate past with front, but also have the option to use words that suggest future is in front in that sense being similar to speakers of familiar future-in-front languages (e.g., English, Dutch and Spanish). For example, apart from "以前 / yǐ-qián" (to front, before) and "今后 / jīn-hòu" (now back, from now on) (Experiment 2), "过去 / guò qù" (pass go, past) and "未/将来 / wèi/jiāng-lái" (hasn't come yet / will come, future) are common translations of past and future. Metaphorically, the word "来 / lái" (come) refers to the future as coming to us and the words "过去 / guò qù" (pass go) refers to that time as moving away from us to the past. For instance, as shown in Table 2, "未来 / wèi-lái" (hasn't come yet) is suggested to be in front and "过去 / guò qù" (pass go) is at the back. Time in these metaphors is taken as an ego-reference point, with the future in front and past in back of the speaker (Yu, 2012). In other words, the linguistic metaphors suggest a future-in-front/past-in-back mapping. If there is an immediate effect of temporal wording on

mental representation that can be independent from the culture (cf. Experiment 2), Chinese people are expected to perform fewer past-in-front mappings when the instructions of temporal concepts are in future-in-front metaphors, in comparison to that of when instructions are in past-in-front metaphors and neutral words.

Experiment 3: Does the spatial lexicon matter: future-in-front language

Method

Participants

A new group of 39 Mandarin speakers participated in the future-in-front metaphor condition. They were tested in Rizhao, China, and all materials were in Mandarin.

Materials and Procedure

Participants followed the same procedure to fulfil the temporal diagram task as described in experiment 1, except that the temporal words used in the instruction were different.

The neutral wording of "what happened at an earlier time" and "what would happen at a later time" in the task instruction of experiment 1 were replaced with "past events" and "future events", conveying future-in-front metaphors. Specifically, participants were instructed to put the "apple" in the box that corresponded to "过去 / guò-qù" (pass go, past) events, and the "pear" to the box that corresponded to "未来 / wèi-lái" (will/not yet come, future) events (Table 2).

Results and Discussion

In this future-in-front metaphor condition, only a small proportion of Chinese people performed a past-in-front mapping, which was significantly different from that in the past-in-front metaphor condition (8% vs. 57%), Wald χ^2 (1, N = 76) = 16.13, p = .0001, odds ratio = 15.75, CI = [4.10, 60.48]. The rate was also significantly different from that of the neutral wording condition (8% vs. 36.8%) Wald χ^2 (1, N = 77) = 7.99, p = .0047, odds ratio = 7.00, CI = [1.82, 26.99]. Additionally, a sign test showed that 8% was significantly lower than chance, (p < .0001, N = 39), which indicates that Chinese participants in the future-in-front metaphor condition have a bias towards future-in-front mappings.

When we merged the data from Experiments 1, 2 and 3, and recoded the three temporal wording conditions according to the extent to which they hinted past-in-front mappings: that is, the least for future-in-front metaphors, than the neutral wording, and the most for past-in-front metaphors. The result showed that wording was indeed a significant factor in predicting Chinese participants' past-in-front mappings. Wald χ^2 (1, N = 114) = 17.99, p < .0001, odds ratio = 3.51, CI = [1.96, 6.26]. In other words, the more a temporal expression is conveying a past-in-front mapping, the more likely a Chinese will conceptualise the past in the front. This again demonstrates an effect of spatial metaphors on people's mental representation of time within the Chinese culture.

To further confirm the assumption of this lexical effect, we did a random check on some participants who performed future-in-front mappings. They were shortly asked to perform the task again after receiving an oral instruction, in which the temporal expressions were changed to the past-in-front spatial metaphors (thus using the same temporal wording as in Experiment 2, i.e., "以前 / yǐ-qián" (to front, before) and "今后 / jīn-hòu" (now back, from now on). Interestingly, some Chinese people (the same participants) shifted from a future-in-front mapping to a past-in-front mapping. We immediately asked them the reason why they had two completely different placements. Their response then usually was a variant of: "Because you used the words of "以前 / yǐ-qián" (to front, past), and my feeling for what happened in "yǐ-qián" should be in front.…" Therefore, lexical spatial metaphors of time indeed seem to have an immediate influence on people's mental representation of time.

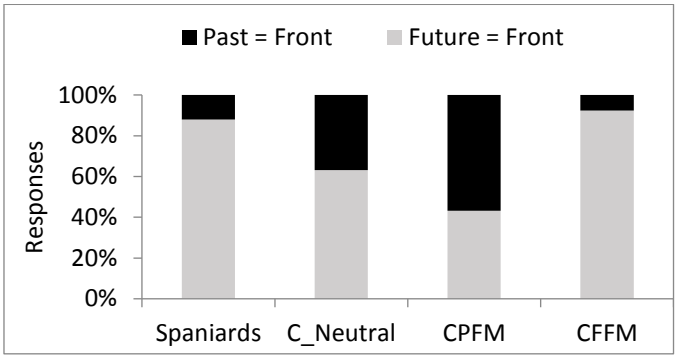


Figure 2: Results of Exps 1-3: percentage of past-in-front and future-in-front responses, separately for Spaniards (de la Fuente et al, 2014), Chinese neutral group (Exp 1) Chinese past-in-front metaphor group (CPFM Exp 2), and Chinese future-in-front metaphor group (CFFM, Exp 3).

Furthermore, in comparison to Spaniards, we see an interaction between lexical effect and culture (Fig. 2). For instance, Chinese people in the past-in-front metaphor condition were significantly different from Spaniards (57% vs 12%), Wald χ^2 (1, N = 87) = 17.12, p < .0001, odds ratio = 9.62, CI = [3.29, 28.13]. Nevertheless, Chinese people in the future-in-front metaphor condition did not exhibit significant differences from the Spaniards (8% vs. 12%), Wald χ^2 (1, N = 89) = 0.44, p = .51, odds ratio = 0.61, CI = [.14, 2.62]. When combining data from Experiments 1, 2, and 3, Chinese people were still more likely to have past-in-front mappings than the Spaniards (12% vs. 33%), Wald χ^2 (1, N = 164) = 7.38, p = .0066, odds ratio = 3.67, CI = [1.44, 9.36]. This result confirms that there were significant differences between Chinese and Spanish cultures, as predicted by the Temporal Focus Hypothesis, according to which Chinese are more past-oriented than Europeans. These differences may not be explained by the wording of the task, as approximately equal numbers of participants had past-in-front, future-in-front and neutral wording tasks.

General Discussion and Conclusion

Previous studies observed that Chinese people can perform forward temporal gestures for past events and backward gestures for future events (Gu et al, in preparation; Lai & Boroditsky, 2013), and in addition display the universally more common pattern that past is gestured towards the back and future towards the front (e.g., English, Dutch and French people). The present study used a temporal diagram task to explicitly test whether Chinese people have a past-in-front mapping. In three experiments, the lexicons of temporal expressions were manipulated as neutral, past-in-front and past-in-back metaphors. According to the results of Experiments 1 and 2, in which temporal expressions were constructed with neutral or past-in-front mappings, Chinese people did not have a bias towards past-in-front or future-in-front mappings. This pattern of space-time mapping was different from Spaniards, who predominately had a future-in-front mapping. Interestingly, when the wording of the temporal expressions consisted of future-in-front metaphors, Chinese appeared to have similar future-in-front mappings as Spaniards (Experiment 3). On average, around one third of Chinese participants (Experiments 1, 2 & 3) had past-in-front mappings, and this proportion was much larger than that of Spaniards (Fig. 2).

We further explored several aspects that can account for Chinese people's past-in-front mapping. First, the differences among three experiments show that lexical spatial metaphors have an online effect on the space-time mapping. Chinese people are more likely to have past-in-front mappings when past and future time conceptions are expressed with lexical "qián" (front, before) and "hòu" (back, after) than when they are expressed with neutral wording. By contrast, Chinese are less likely to have past-in-front mappings when past and future are expressed with lexicons of "guò-qù" (past go, past) and "wèi-lái" (hasn't come, future) than when they are expressed with neutral wording. The lexical spatial metaphor is a significant predictor of Chinese people's space-time mappings.

This raises the question as to what causes some Chinese people to use a past-in-front mapping even in the neutral condition (Experiment 1). Partly, this pattern could be related to a long term use of the past-in-front spatial metaphors, such that participants form a habitual space-time mapping even in the neutral condition.

This is in line with the proposal that speaking and learning different spatial metaphors can lead to different conceptualisations of time (e.g., Boroditsky, 2001). For instance, in a top/down plane, Chinese speakers can use vertical spatial metaphors to talk and gesture about time (e.g., "up week" means last week). Due to the habitual vertical conceptualisation of time, they also perform vertical gestures for temporal conceptions with no spatial metaphors (e.g., yesterday, tomorrow), though to a lesser extent (Gu et al., 2014).

However, we observed cross-cultural differences between Chinese and Spaniards in space-time mappings. Chinese people are significantly more often spatializing the past in front than the Spaniards, both in the lexical neutral condition and all conditions combined. One can ascribe this discrepancy to the differences in cultural values towards the past and future. If Chinese people perceive past more valuable and are more past-focused than the Europeans (Ji et al, 2009), it is plausible that Chinese people will more often have past-in-front mappings than the Spaniards. Given the fact that people usually put in front what they consider to be important, if the past is important, it is of a high priority to be placed in the front. Therefore, the differences in temporal focuses between Chinese and Spanish cultures may be part of the explanation for why Chinese people have a larger proportion of past-in-front mappings than the Spaniards. This provides new evidence supporting the Temporal Focus Hypothesis (de la Fuente, 2014).

To further explore the extent to which temporal focus plays a role in shaping Chinese people's space-time mappings, we need to do a qualitative survey on Chinese people's cultural focus of attention. Moreover, to better understand the interplay between language and cultural focus of attention, future study can research whether western learners of Chinese can form a habitual past-in-front mapping in the neutral condition after learning Chinese sagittal spatial metaphors, controlling for cultural focus of attention. Alternatively, we can also compare Mandarin speakers with Chinese signed language speakers, who have different sagittal spatial temporal metaphors within the Chinese culture (in Chinese signed language, the spatial metaphors of "front" is only used for the expression of the future temporal concepts) (Gu & Swerts, in preparation).

Furthermore, according to posthoc interviews, the various results may be due to competing time conceptions in Chinese participants. Some participants explained that the past refers to known events so one can see it in front of eyes, whereas the future is unknown and one cannot see it (so it is at the back). This explanation is in line with Aymara speakers, who also have a past-in-front mapping (Núñez & Sweetser, 2006).

Alternatively, some participants explained that they put what has happened first in the front and what has happened afterwards in the back. It is possible that those who put the past in front take a Time-Reference-Point metaphor, where earlier events in time are "in front of" later events (Núñez, Motz, & Teuscher, 2006; Yu, 2012). Specifically, they consider the series of events as a sequence from the front to the back as if they are waiting in a queue. For instance, no matter which direction you look at in the line, there is a front and back to that line according to convention. Those who are or near first position will be served earlier than those who are behind them (later), irrespective of the Ego's point of view (Núñez, Motz, & Teuscher, 2006; Walker, Bergen & Núñez, 2015). In other words, if one would be positioned in such a queue, then the people that are way back in the line will be served later (so in a more distant future).

If Chinese people think of time events as a sequence, then the anteriority refers to one time as being earlier in a sequence than another whereas posteriority refers to one time that is later in a sequence than another. This way of thinking is different from the category of "past" and "future", as a time conception of past can still be earlier in a sequence than a time conception of future, i.e., past is in

front of future. In other words, the spatial-temporal mapping depends on the sequence of the time references regardless of the time conceptions per se. For instance, even if we instruct participants by an anterior event with a past time conception and posterior event with a future time conception, participants are expected to have a past-in-front mapping if they think about time in a sequence.

Note that in our temporal diagram task, there was a character standing between the two boxes. Such a design may require participants to displace the deictic centre from their body to an external location and thus may cause them to avoid using internal deictic time. It is likely that participants mapped earlier or later events on to the inherent "frontness" and "backness" of the character, with earlier events lying ahead of the character and later events lying behind. The finding is consistent with the "earlier events lie ahead of later events" structure found in the study of psychological reality of sequential time (Gentner, Imai & Boroditsky, 2002; Núñez, Motz, & Teuscher, 2006; Walker, Bergen, & Núñez, 2015).

In sum, the experiments demonstrate a cross-cultural difference in spatial conceptions of time and explore the accounts for Chinese people's past-in-front mappings. The findings of the study support de la Fuente (2014)'s Temporal Focus Hypothesis, provide further evidence to the claim that uttering a different spatial metaphor may influence that speakers' conceptualisation of time (Boroditsky, 2001; Lai & Boroditsky, 2013), and are also consistent with previous studies on the psychological reality of sequential time (e.g., Núñez, Motz, & Teuscher, 2006). The study contributes to a growing body of evidence that spatial-temporal thinking can be rapidly affected by context (Boroditsky, 2000; Casasanto & Bottini, 2014). Moreover, there appear to be both long-term effects of cultural attitudes on the spatialization of time, and also immediate effects of the space-time metaphors used to probe people's mental representations.

Acknowledgements

The first author received financial support from The Netherlands Organization for Scientific Research, via NWO Promoties in de geesteswetenschappen (322-89-007), which is greatly acknowledged.

References

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology, 43*, 1-22.

Casasanto, D., & Bottini, R. (2014). Mirror reading can reverse the flow of time. *Journal of Experimental Psychology: General, 143*, 473-479.

Casasanto, D. & Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers. *Cognitive Linguistics, 23*, 643-674.

Clark, H. (1973). Space, time semantics, and the child. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language* (pp.27-63). New York: Academic Press.

De la Fuente, J., Santiago, J., Román, A., Dumitrache, C., & Casasanto, D. (2014). When you think about it, your past is in front of you: How culture shapes spatial conceptions of time. *Psychological Science, 25*, 1682-1690.

Fuhrman O., Boroditsky L. (2010). Cross-cultural differences in mental representations of time: evidence from an implicit nonlinguistic task. *Cognitive Science, 34*, 1430-1451.

Gentner, D., Imai, M., & Boroditsky, L. (2002). As time goes by: Evidence for two systems in processing space time metaphors. *Language and Cognitive Processes, 17*, 537-565.

Gu, Y., Mol, L., Hoetjes, M.W., & Swerts, M.G.J. (2014). Does language shape the production and perception of gestures? In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of CogSci2014* (pp. 547-552).: Austin, TX: Cognitive Science Society.

Guo, T., Ji., L., Spina, R., & Zhang, Z. (2012). Culture, temporal focus, and values of the past and future. *Personality and Social Psychology Bulletin,38*, 1030-40.

Ji, L., Guo, T., Zhang, Z., & Messervey, D. (2009). Looking into the past: Cultural differences in perception and representation of past information. *Journal of Personality and Social Psychology, 96*, 761-769.

Lai, V. T., & Boroditsky, L. (2013). The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in English, Mandarin, and Mandarin-English speakers. *Frontiers in Psychology*, 4, 1-10.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Miles, L. K., Nind, L. K., & Macrae, C. N. (2010). Moving through time. *Psychological Science, 21*, 222-223.

Núñez, R., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the cross-linguistic comparison of spatial construals of time. *Cognitive Science, 30*(3), 401-450.

Núñez, R., Motz, B., & Teuscher, U. (2006). Time after time: The psychological reality of the ego- and time-reference-point distinction in metaphorical construals of time. *Metaphor and Symbol, 21*, 133–146.

Saj, A., Fuhrman, O., Vuilleumier, P., & Boroditsky, L. (2014). Patients with left spatial neglect also neglect the "Left Side" of time. *Psychological Science, 25*, 207-214.

Ulrich, R., Eikmeier, V., de la Vega, I., Fernandez, S.R., Alex-Ruf, S., & Maienborn, C. (2012). With the past behind and the future ahead: Back-to-front representation of past and future sentences. *Memory and Cognition, 40*, 483–495.

Yu, N. (2012). The metaphorical orientation of time in Chinese. *Journal of Pragmatics, 44*, 1335-1354.

Walker, E. J., B., K. Bergen & Núñez, R. (2014). Disentangling spatial metaphors for time using non-spatial responses and auditory stimuli. *Metaphor and Symbol, 29*, 316-327.

A performance model for early word learning

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Molly L. Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Kyle MacDonald
kyle.macdonald@stanford.edu
Department of Psychology
Stanford University

Abstract

The emergence of language around a child's first birthday is one of the greatest transformations in human development. Does this transition require a fundamental shift in the child's knowledge or beliefs, or could it instead be attributable to more gradual changes in processing abilities? We present a simple model of cognitive performance that supports the second conclusion. The premise of this model is that any cognitive operation requires multiple steps, each of which require some time to complete and have some probability of failure. We use meta-analysis to estimate these parameters for two components of simple ostensive word learning: social cue use and word recognition. When combined in our model, these estimates suggest that learning should be very difficult for children younger than around a year, especially with gaze alone. This model takes a first step towards quantifying performance limitations for cognitive development and may be broadly applicable to other developmental changes.

Keywords: Speed of processing; development; word learning; meta-analysis

Introduction

Human beings begin their lives as helpless infants and yet rapidly become children who are able to perceive, act, and communicate. Infants who cannot communicate become toddlers who use words to share attention and indicate their desires. Toddlers who cannot follow the trajectory of a ball become preschoolers who can. A fundamental question of developmental psychology is how these external behavioral differences come about via internal processes of developmental change.

One possibility is that these external transitions are a product of radical internal shifts, such as the discovery of the communicative function of language, or the emergence of a theory of others' minds. Such shifts have been a centerpiece of constructivist theories of development from Piaget (1969) onward. These theories have obvious appeal, at least in part because the outward changes in children's cognitive abilities are so dramatic. Yet several decades of work with infants has revealed a surprising amount of detectable knowledge about cognitive domains, often months or even years prior to these external manifestations (Carey, 2009). How could these two sets of observations co-exist?

Perhaps children's intense performance limitations—basically, difficulties *using* knowledge or representations that they nevertheless possess—limit our abilities to observe or even to measure their competence (Chomsky, 1965). Perhaps planning to reach for an object is difficult and time-consuming enough that toddlers lose track of what they were looking for (Keen, 2003). And perhaps infants are trying to learn the meanings of words, they are just too slow and error-prone to make much progress in this task. In some sense, this hypothesis constitutes a strong null model of development: speed and accuracy *must* change, even if no internal representations do.[1]

In the current paper, we take up the challenge of building such a null model, using early word learning as a case study for exploring the role of performance limitations. The emergence of language is one area where theoretical views have differed widely. Must children master a particular insight about the role of language in communication to begin learning words in earnest (Hollich, Hirsh-Pasek, & Golinkoff, 2000), or are they pursuing the same activity throughout early childhood, but with more success later on (McMurray, 2007)?

Some empirical data support the possibility of early communicative competence. At their first birthday, infants have some expectations about the function of words in communication and show longer looking times when those expectations are violated (Vouloumanos, Onishi, & Pogue, 2012). And 6- to 9-month-olds perform above chance in word-object mapping tasks (Bergelson & Swingley, 2012). But the level of performance they show compared with older children is so limited that the data also seem to provide *prima facie* evidence for some kind of shift in representation.

We suggest instead that continuous developmental processes might be responsible, specifically increases in the speed and reliability of internal cognitive processes. We pursue this suggestion by creating a performance model for early word-object mapping. Our starting point is the idea that even the simplest word learning input for object referents involves following some kind of attentional cue (e.g., gaze or pointing) to a distal target and then processing some kind of link between a word and the target referent. Each of these abili-

[1] Note that this viewpoint does not entail any sort of nativism at all. It merely suggests that the relevant competence emerges significantly earlier than is typically supposed.

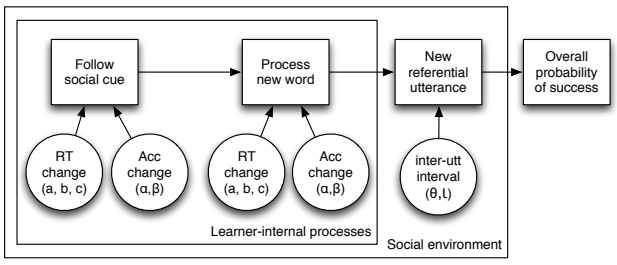


Figure 1: A schematic visualization of our model of early word learning, with the stages of processing (squares) along with the relevant parameters for each stage (circles).

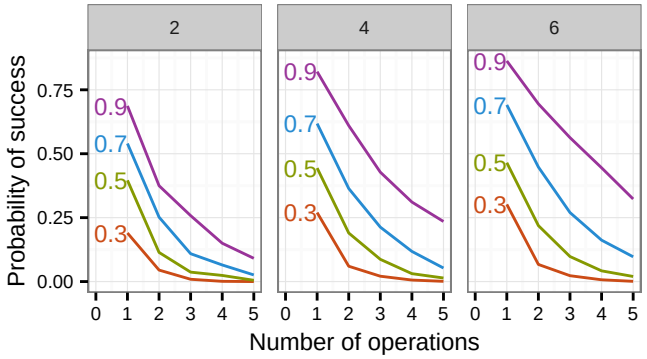


Figure 2: Probability of successfully executing chains of cognitive operations—each with their own speed and reliability—with different numbers of steps (shown by different colors). Facets show different temporal thresholds.

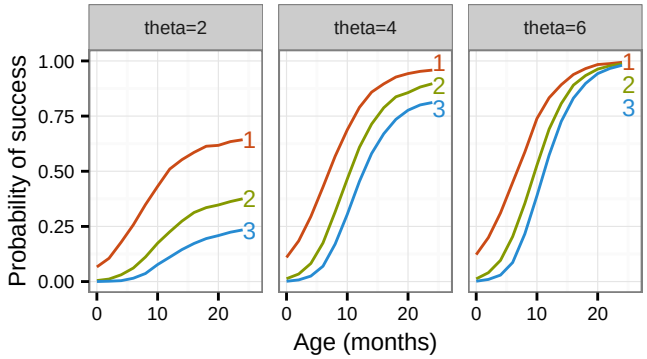


Figure 3: Probability of success for one set of developmental parameters. Different colored lines indicate chains of differing lengths, panels show different temporal thresholds.

ties has been shown to develop dramatically over the first two years and beyond. So it stands to reason that any achievement that depends on both should develop even *more* dramatically in the same period.

Our goal is to create a quantitative model that allows us to formalize this intuition. Inspired by recent meta-analyses of developmental phenomena (e.g., Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014), we conduct systematic literature reviews of the literature on social cueing (e.g., gaze following; Scaife & Bruner, 1975) and word recognition (Fernald, Pinto, Swingley, Weinbergy, & McRoberts, 1998). These meta-analytic surveys, in combination with parametric models of development (e.g., Kail, 1991) allow us to estimate the speed and accuracy for a two-component model of word learning.

The outline of the paper is as follows. We begin by describing the basic model and how it captures developmental changes. We next estimate the development of speed and accuracy independently for social cueing and word recognition. We then estimate the pace of referential utterances from a corpus, and compute children's predicted learning rate based on these parameter estimates. The conclusion of our analysis is that even if young infants were trying to learn in precisely the same way as older toddlers, they would be too slow and too fallible to extract much signal from their input data.

Model

The basic assumptions of our performance model are familiar from cognitive architectures that attempt to capture specifics of cognitive processes (e.g., ACT-R; Anderson, 1996), namely, every cognitive operation has a processing time and a probability of failure. Each complex cognitive operation is decomposable into a chain of simpler operations, any one of which can fail. And if a single link in the chain fails, then the overall operation fails as well. Thus, the probability of failure is the product of the individual failures. Similarly for timing, the total processing time for a chain of operations is the sum of the processing times for the parts.

Complex actions are describable at many different granularities. For example, word learning from an ostensive cue (e.g., parent says "doggie!" while pointing at a dog) can be decomposed into 1) social cue following and 2) word recognition/mapping. But social cue following can be further decomposed into 1a) attending to the cue, 1b) processing the directionality of the cue, and 1c) executing an eye-movement to the cue's target. Each of these could easily be broken down further. There is no one decomposition of a task, but we view this feature as a strength rather than a weaknesses of the basic framework, which can be applied to units at any grain size for which speed and reliability of processes can be measured. The overall model architecture is shown in Figure 1.

Chains of mental processes

Consider a sequence of interacting mental processes. We assume that each of these has a Bernoulli success probability s_p. Thus, the probability of a sequence of failures is exponential such that $p_{success} = \prod_p s_p$. And each operation also takes some time to complete t_p, which we assume is distributed log-normally. Thus, the total time of the chain is $rt = \sum_p t_p$.[2] Finally, consider that this operation is time-sensitive, and must be completed within a temporal threshold, also sampled from a log-normal distribution with mean θ and SD ι.

We can now approximate the probability that a chain is successful within a particular threshold. A representative set of simulations are shown in Figure 2. For these and many other parameter settings, long chains of operations are unlikely to succeed unless individual operations are very fast and very accurate.[3]

Development within the model

The two posited capacities in our model are speed and accuracy (probability of a successful operation). Both of these

[2]Note that there is not a known parametric form for the sum of multiple log normals. A variety of analytic approximations for these sums exist (Fenton, 1960), but they have some limitations, so instead we use numerical simulations here.

[3]All code and data for the simulations reported here is available at `http://github.com/mcfrank/sop`.

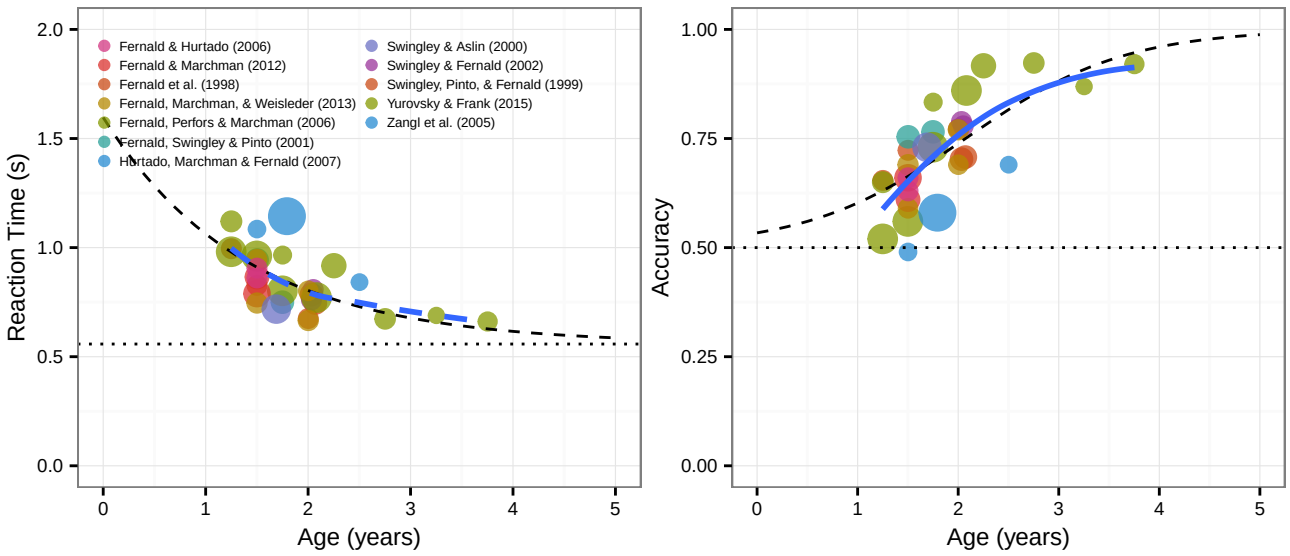


Figure 4: Reaction times (left) and accuracies (right) for two-alternative forced-choice word recognition paradigms. Each circle shows an experiment, with areas scaled by the number of participants. A generic loess smoothing function is shown in blue. For reaction times, dotted line shows adult reaction time and dashed line shows best-fitting Kail function. For accuracy, dotted line shows chance, and dashed lines show best-fitting half-logit function.

should change across development for any constituent cognitive operation, leading to dramatic changes in the cumulative speed and accuracy of chains of operations across development. To estimate these changes, we use parametric models of developmental change.

Pioneering work by Kail (1991) describes the developmental trajectory of reaction times for complex tasks, via aggregation across the published literature. Empirically, the slope of these reaction times follows an exponential, such that $Y(i) = a + be^{-ci}$, where Y is the predicted variable, a is the eventual (adult) asymptote, b is the multiplier for the (infant) intercept, c is the rate of development, and i is age. The Kail (1991) model is a model of RT multipliers. Since operations are additive, these multipliers should apply to individual operations or to chains of operations equivalently: if the multiplier is constant then it can be factored out.

Next we turn to accuracy. For simplicity, we consider the probability of success on a single operation changing across time as a simple logistic function where $Y(i) = \frac{1}{1+e^{\alpha+\beta i}}$. α sets the intercept and β marks the developmental multiplier, as in a standard logistic regression.

Preliminary simulations

We can combine these functions with the basic operation-chain simulations defined above and examine the probability of a successful chain of operations across ages. Results for one parameter set are shown in Figure 3. These simulations show that sharp developmental transitions from failure to success can be the product of relatively broad underlying functions. But the difficulty is constraining the model's predictions requires substantial information about speed and accuracy. In the next section we turn to the estimation of these parameters via meta-analysis.

Case Study: Early Word Learning

Why do children begin to show evidence of word learning around their first birthday? Although many accounts have been proposed (e.g. Tomasello, 1995's "nine-month revolution"), our null-model framework provides a simple explanation. Children may be trying to learn words from very early in development, but the basic cognitive components may be too slow and too challenging to allow for consistent learning (and consistent measurement of that learning by psychologists). The recent literature on early word learning gives some support for this contention, as careful measurement has revealed some aspects of receptive language prior to the first birthday (Bergelson & Swingley, 2012).

We focus here on learning a word that is presented ostensively via a social cue like gaze or pointing. For simplicity, we decompose the task of social word learning into two abilities: 1) social cue following, and 2) word recognition. This task analysis is an approximation: pointing is not the same as gaze following (and neither is it always necessary), and recognition is not the same as learning and retention. But it nevertheless captures some aspects of the task, namely following a social cue to a distal target and processing some language associated with that target. And it has the major benefit for our purposes of providing data on development, since each of these tasks is well-studied.

Word recognition

We first estimated developmental changes in the speed of processing for word recognition. Fernald et al. (1998) introduced the method of using eye-movements to measure children's ac-

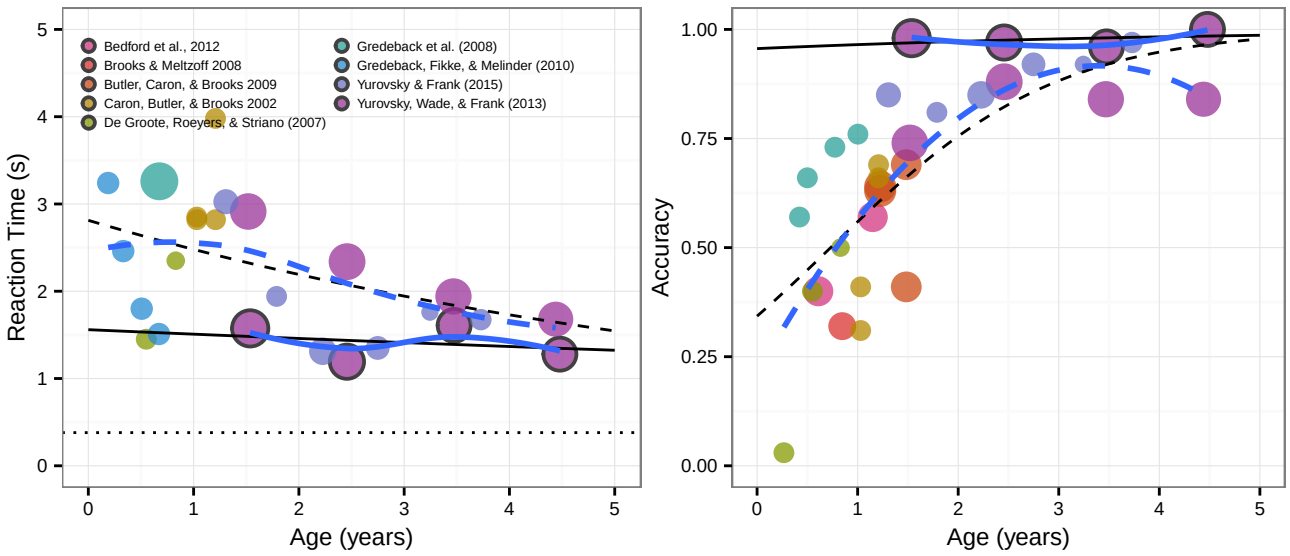


Figure 5: Reaction times (left) and accuracies (right) for social cue following. Plotting conventions are as above. Results for gaze following alone are shown with no border; results for gaze plus pointing are shown with a black border. Dashed and solid lines show fits for gaze and gaze plus pointing, respectively.

curacy and reaction time across development. Subsequent investigations have identified many factors involved in speed of word recognition (e.g., the frequency and familiarity of the target word). Nevertheless, such research generally attempts to select simple, easy words that should be accessible to most children, so we can use this literature to derive rough estimates of accuracy and reaction time across development.

We conducted a systematic literature review by using Google Scholar to identify peer-reviewed papers citing Fernald et al. (1998). We screened this sample manually to find the subsample of 12 papers that reported both accuracy and reaction time with sufficient detail to permit coding. Figure 4 shows reaction times from this sample of papers, plotted by the mean age of the children in the reported studies. The dashed line shows an exponential function fit to these data (with $a = 0.56$, $b = 1.04$, and $c = 0.72$). The intercept a is estimated as the mean reaction time for adults in control experiments.[4]

Accuracies can be estimated similarly. We fit a logistic regression to accuracy data, using a half-logit linking function ($.5 + .5 \times \frac{1}{1+e^{-x}}$) to bound accuracy between .5 and 1. Figure 4 shows the results of this analysis, with estimated parameters $\alpha = -2.62$ and $\beta = 1.27$. In both of these cases we see a qualitatively strong fit between the developmental model we assumed and the particulars of the experimental data (e.g., as estimated by a naïve smoothing model, shown in blue).

[4]Note that Kail (1991)'s original analysis was of the slope of mental rotation speeds, so the exponential curve described a multiplier on the adult slope. This analysis is simpler, fitting a curve to the mean RTs directly.

Social cue following

For our analysis of social cue following, we were interested in both pure gaze following and gaze following supplemented by pointing. We identified papers using a Google Scholar search for "gaze following" and included those studies that A) included data from typically-developing children, B) used a standard face-to-face gaze-following task, and C) reported percentage accuracy (rather than a score or other composite measure). Although we coded all papers that fit these criteria, we focused on papers with a simple two-alternative forced choice (9 papers); integrating across different numbers of alternatives added additional complexity to our model.

In our first iteration of this analysis, we found that very few studies reported reaction times for gaze following, and those that did had no data from children older than 15 months and no data from gaze plus pointing. Estimating developmental curves for these data was difficult; to remedy this issue we include new analyses of data from Yurovsky, Wade, & Frank (2013) and Yurovsky & Frank (2015). Both of these paradigms were word learning experiments in which a social cue (either brief or extended) was used to indicate a referent. Data from these studies can thus be used to estimate social cue following time; conveniently, both studies included large numbers of participants at older age ranges, constraining our analysis. We also added 440ms as a floor adult reaction time, on the basis of measurements by Driver et al. (1999).

Figure 5 shows reaction time and accuracies for gaze following, both with and without pointing. Again we see relatively good qualitative fit by the developmental models. Details of these fits are identical to our analysis of word recognition, except that we use a standard logit function. We experimented with using a half-logit link, but found that many

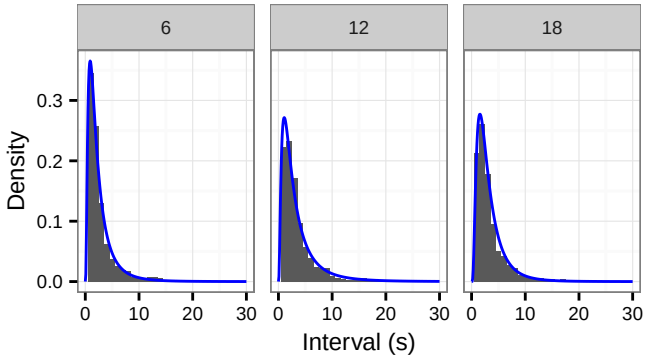


Figure 6: Histograms of inter-utterance intervals in child-directed speech. Panels show data from 6-, 12-, and 18-month-olds. Blue curves show the best-fitting log normal distribution for the full dataset.

studies reported accuracies below .5 due to children's failure to disengage from the experimenter's face.

Estimating a temporal threshold

We next estimate plausible values for θ and ι, which control the distribution of temporal thresholds at which referential utterances must be processed. To estimate these values, we turned to the Fernald and Morikawa corpus (Frank, Tenenbaum, & Fernald, 2013), which contains a set of transcribed interactions between caregivers and children as they play with pairs of objects. Critically, this corpus contains approximate timing information (Rohde & Frank, 2014) as well as annotations for social cues used by the caregivers (e.g., gaze, pointing, etc.) to indicate which toy is being talked about.

The primary variable of interest for our analysis was the distribution of time intervals between utterances using social cues to refer to objects. Preliminary analyses indicated that there were no differences in timing between referential utterances (those containing a concrete reference to an object in the current context) and non-referential utterances, so we examined the full distribution of inter-utterance intervals. Figure 6 shows the empirical distribution across ages, along with the best-fitting log-normal distribution. The mean time between utterances was 3.61 and the median was 2.5. Perhaps surprisingly, there were no major differences in the distribution between age groups (for example, $\text{median}_6 = 2$, $\text{median}_{12} = 2.5$, and $\text{median}_{18} = 2.5$), suggesting that parents were not substantially adjusting the pace of conversation to children, at least in this corpus. For our simulations below, we use the parameters of the best-fitting distribution across ages ($\theta = 0.78$, $\iota = 0.86$).

Simulations

We now run the same developmental simulations described above, but using the estimated numerical parameters for word processing, social cue following, and the timing of utterances in child-directed speech. Figure 7 shows the probability of successful ostensive learning by age, split by social cue.

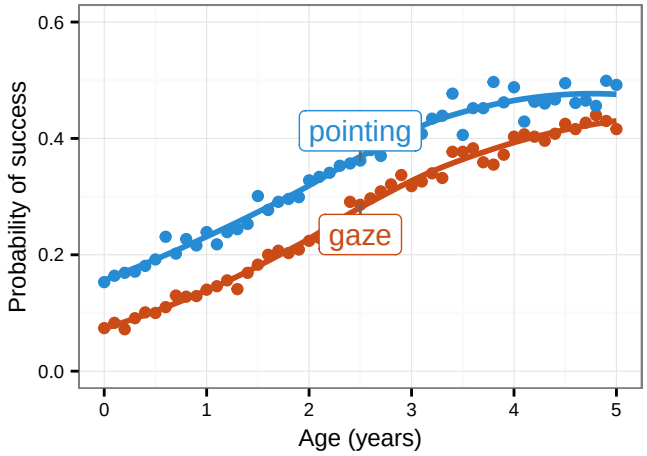


Figure 7: Simulation results using meta-analytic parameters estimated for speed and accuracy in social cue following and word recognition.

Pointing is substantially more effective than gaze following, but under this naïve model, neither cue ever leads to success more than half of the time. While low accuracy constrains performance early in development, accuracy is above 80% by age 3 and reaction time is the bounding factor.[5]

A second way of viewing these same simulations is shown in Figure 8, which shows the average number of learning instances a learner would need to achieve a single successful mapping trial under this model. This number declines from more than 10 in early infancy to an asymptote of approximately two at age five. Here the difference between gaze and pointing is more apparent, especially earlier in development.

Is it reasonable to assume that not all word mapping opportunities succeed? After all, 3–4-year-old children have been show to learn words from a single exposure (Carey, 1978). We would argue that a success rate of approximately one half at age three is actually very congruent with accuracies in the 60-80% rate shown by Markson & Bloom (1997) and others (given a handful of exposures during training). And rates of learning for younger children also show some numerical congruence (e.g., Woodward, Markman, & Fitzsimmons (1994) showed some evidence of learning from nine exposures in 13- and 18-month-olds). In sum, even though average results suggest that some children may learn from a small set of exposures, they do not imply that *all* children have learned.

Discussion

We presented a model of children's developing processing capabilities in which the sole age-related changes are in the speed and accuracy of mental operations. This strong null

[5]One consequential decision for the model is the assumed response function for social cues. Here we have used a half-logit link function for accuracy, assuming a two-alternative forced choice between referents. Without this assumption, learning probabilities for early infancy go down substantially, but the asymptote remains unchanged.

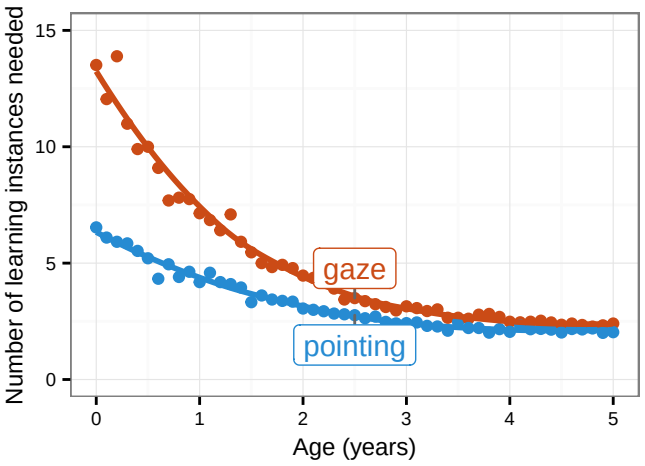


Figure 8: Average number of learning instances needed per effective learning instance, plotted by age.

model is a first step towards a baseline model of cognitive development, attempting to answer the question of what changes we would see even if there were *no* substantive differences in children's internal representations across ages. Using measurements extracted from the literature, we used the model to make predictions about early word learning, a domain in which there has been uncertainty about the presence of early representational shifts. Without including any representational changes, the model still produced large developmental changes and showed *prima facie* congruence with some previous experimental work.

The simple model we described here has many limitations. First, we estimated parameters from the data that were available rather than the data we would have liked to have (e.g., reaction times from familiar word recognition rather than from novel word learning). Second, we assumed a purely serial model of responding in which accuracy and reaction time were independent from one another; more sophisticated models of decision-making might link processes in a race model or yoke accuracy and reaction-time in a speed-accuracy tradeoff. Third, although we did not find evidence of large developmental changes in the pace of utterances, parents likely still adapt to their children's speed of processing in some instances, leading to better outcomes for those instances.

Despite the many assumptions and limitations of our model, the results should still constrain and inform our theories. If we discard the parametric form of the model and simply examine the meta-analytic reaction times we estimated, we see that they are generally longer than the interval between new utterances, suggesting that word learning through gaze is likely to be difficult in the first year under any model. This qualitative observation suggests the basic intuitions derived from our model may be useful for analyzing other domains. More generally, the null model we articulated here should reinforce the point that—even with the most sensitive measurements available—we should not infer a lack of competence from a failure in performance.

Acknowledgements

This work supported by NSF BCS #1528526. Thanks to Ellen Markman, Rebecca Saxe, and members of the Language and Cognition Lab at Stanford for helpful discussion.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Carey, S. (1978). The child as word learner. In *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carey, S. (2009). *The origin of concepts*. Oxford, UK: OUP.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT press.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, *6*(5), 509–540.

Fenton, L. F. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, *8*(1), 57–67.

Fernald, A., Pinto, J. P., Swingley, D., Weinbergy, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 1–135.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490.

Keen, R. (2003). Representation of objects and events why do infants look so smart and toddlers look so dumb? *Current Directions in Psychological Science*, *12*(3), 79–83.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*(6619), 813–815.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. Basic Books.

Rohde, H., & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science*, *38*(8), 1634–1661.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*.

Tomasello, M. (1995). Joint attention as social cognition. *Joint Attention: Its Origins and Role in Development*, 103–130.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*(32), 12933–12937.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13-and 18-month-olds. *Developmental Psychology*, *30*(4), 553.

Yurovsky, D., & Frank, M. C. (2015). Beyond nave cue combination: Salience and social cues in early word learning. *Developmental Science*.

Yurovsky, D., Wade, A., & Frank, M. C. (2013). Online processing of speech and social information in early word learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Cognitive biases and social coordination in the emergence of temporal language

Tessa Verhoef (tverhoef@ucsd.edu)
University of California, SD, Center for Research in Language

Esther Walker (e1walker@ucsd.edu)
University of California, SD, Cognitive Science

Tyler Marghetis (tmarghet@indiana.edu)
Indiana University, Bloomington, Psychological and Brain Sciences

Abstract

Humans spatialize time. This occurs within individual minds and also in larger, shared cultural systems like language. Understanding the origins of space-time mappings requires analyses at multiple levels, from initial individual biases to cultural evolution. Here we present a laboratory experiment that simulates the cultural emergence of space-time mappings. Dyads had to communicate about temporal concepts using only a novel, spatial signaling device. Over the course of their interactions, participants rapidly established semiotic systems that mapped systematically between time and space. These semiotic systems exhibited a number of similarities, but also striking idiosyncrasies. By foregrounding the interaction of mechanisms that operate on disparate timescales, laboratory experiments can shed light on the commonalities and variety found in space-time mappings in languages around the world.

Keywords: language evolution; space and time; abstract concepts; social coordination; cultural evolution

Introduction

Human language stands out for its capacity to refer to things that are elsewhere in space or time: "How was the movie?". Beyond this capacity for "displaced" reference (Hockett, 1960) to concrete objects, we can also talk about entities that are entirely abstract: the future, the number five, and so on. Indeed, natural languages deploy nuanced, systematic strategies to refer to concepts for which there are no stable perceptual referents. What are the origins of the linguistic structure that allows us to communicate about the abstract? Here, we present a novel experiment to help answer this question, focused on the language of *time*.

A test domain: Time

The conceptual domain of Time is a critical test case for theories of language evolution because, while temporal concepts are entirely divorced from perception, they are also highly structured. The concepts *next week* and *yesterday* cannot be distinguished perceptually; the events of next week will, soon enough, be those of yesterday. Temporal concepts lack stable perceptual referents. And yet our conceptualization of time is highly structured, consisting of at least three distinct facets: duration, order, and deixis (for a review: Núñez & Cooperrider, 2013). Temporal duration refers to *amount* of time. In English, we have a rich lexicon of duration terms (second, hour, year, etc.). Temporal order refers to the *sequence* in which events occurred. Pregnancy occurs *before* birth; celebrations come *after* victories. Temporal deixis refers to the placement of events in relation to *now:* past, present, or future. A celebration can be occurring now, have occurred in the past, or occur in the future. All three facets can be combined during temporal reference, so that we can speak of a *day-long* party happening *before* your next birthday (i.e., in the *future*).

Across a number of languages, a recurring strategy is to use *spatial* words to communicate about time (e.g., Haspelmath, 1997; Lakoff & Johnson, 1980). For instance, in English you can say *"**back** in the '50s"* or *"look **forward** to the weekend*," mapping past to back and future to front. How do such systems of conventionalized space-time mappings emerge? Some argue that we have an innate bias to associate space, time, and number (e.g., Walsh, 2003). On this view, cross-domain associations in language arise from neural resources that are shared across space, time, and number. Others argue that a variety of abstract domains (including time) share structural and relational features with space and thus can easily be aligned (e.g., Gentner, 1983). Once aligned, structure from the domain of space—and its associated language—is projected to the domain of time. A third view is that regularities in experience create cross-domain associations in the minds of individuals, so-called "conceptual metaphors" that, in turn, influence how we think and talk (Lakoff & Johnson, 1980). While each of these mechanisms may play a role, none on their own can account for the combination of universality and variability that we find in the languages of the world. To understand the emergence of conventionalized space-time mappings in communication, we need to consider the interplay of mechanisms that operated at various timescales (c.f., Núñez & Cooperrider, 2013).

***Table 1**: The full set of meanings used in the experiment. Meanings ranged in duration from brief moments ('now') to limitless periods ('future'), and included durations ('day'), sequential relations ('day before'), and deictic periods defined relative to the present ('yesterday').*

	Duration	***Sequence***	***Deictic***
Moments	*'second'*		*'now'*
Days	*'day'*	*'day before'* *'day after'*	*'yesterday'* *'today'* *'tomorrow'*
Year	*'year'*	*'year before'* *'year after'*	*'last year'* *'this year'* *'next year'*
Periods		*'before' 'after'*	*'past' 'future'*

Sources of structure

Our approach adopts a view of language as a complex adaptive dynamical system (Steels, 1997; Beckner et al., 2009), which foregrounds the complex links between the level of the individual (language users in a community and their individual cognitive biases) and the level of the population (languages as dynamic systems). The field of language evolution has developed experimental methods that help to isolate some of the relevant mechanisms that play a role in these links (Scott-Phillips & Kirby, 2010). These methods help explain how cognitive biases and cultural evolution together may account for some of the structures we see in language (Christiansen & Chater, 2008; Kirby, Griffiths & Smith, 2014). For instance, past work has shown that linguistic structure can emerge from pressures induced by transmission and social coordination (Kirby, Cornish & Smith, 2008; Caldwell & Smith, 2012; Verhoef, 2012; Kirby et al., 2014). Transmission of a language causes signals to be filtered through the cognitive constraints of new learners; only signals and structures that are easily learned and remembered will be reproduced and passed on. In addition, as people interact repeatedly, they have to align their signals to communicate expressively (Garrod, Fay, Lee, Oberlander & MacLeod, 2007; Fay, Garrod & Roberts, 2008; Fusaroli & Tylén, 2012; Caldwell & Smith, 2012). Signals in this case originate as conventions on the basis of social coordination and shared communication history.

Here, using methods that simulate language evolution in the lab, we study the interaction between initial biases and cultural evolution in the emergence of space-time mappings in communication.

Methods

Inspired by work in which novel semiotic systems emerged through interaction, we used a *dyadic communication game paradigm* in which pairs of participants communicated about temporal concepts using a novel signaling device, a vertical touch bar that recorded finger movement sequences. Since communication was restricted to these novel signals, successful communication required the negotiation of novel conventions for representing temporal concepts in vertical space. Undergraduate students (n = 16) at the University of California, San Diego participated for partial course credit.

Time concepts

The time concepts to be communicated fell into three broad categories: Duration (e.g. 'day'), Sequence (e.g. 'before') and Deictic (e.g. 'tomorrow'), as shown in Table 1. Meanings in one category often had closely related meanings in others. Some meanings were distinguished only by whether they involved a tenseless sequential relation rather than being placed specifically in the past, present, or future (e.g., 'year before' *vs.* 'last year'). Others involved specific durations (e.g., 'day') that are located relative to now (e.g., 'today'). While natural languages have evolved ways to encode these subtle distinctions, it is not trivial to develop a novel system of signals that is simple enough to learn but complex enough to communicate successfully about all concepts.

Signals

Pairs of participants communicated using a vertical bar on a touch screen (Fig. 1), which recorded and replayed brief signals. Signals consisted of exactly 5 seconds of vertical movements of a bubble within the bar. The bubble moved continuously, following the location of the participant's index finger. If the finger was lifted during recording, the bubble would stay in the same location until the bar was touched again or the 5 second limit was reached. A vertical signalling system was chosen because English does not make conventional use of verticality to describe time[1].



***Figure 1**: Signalling with a vertical bar on a touch screen*

Procedure

Participants were seated in separate testing rooms. They sat in front of touch screen computers that were connected

[1] Cultural artifacts sometimes associate time with vertical location, but experience with such artifacts is unlikely to generate strong, reliable associations, since artifacts vary considerably in how they map time to space. Calendars place earlier events above later ones. Many internet blogs place earlier posts below the most recent ones. Facebook combines both conventions: earlier Timeline posts are *below* more recent ones, but earlier comments are placed *above*.

through a local network. Communication was thus limited to spatial signals displayed on the vertical touch bar.

The design of the interaction game followed Verhoef, Roberts & Dingemanse (2015). On each trial, a participant was either the 'signaller' or the 'recipient', switching roles after each trial. Trials began when a meaning was displayed on the signaller's screen. The signaller then had to record a signal for this meaning using the vertical bar. Once they were happy with their signal, it was sent to the recipient, who could replay the signal on their own screen. On the basis of this signal, the recipient had to guess the intended meaning; they indicated their selection by touching the meaning on a panel that showed all meanings. Players received feedback after each trial, including both the intended target meaning given to the signaller, and the guessed meaning selected by the recipient.

Each dyad completed four rounds of eighteen trials. Each meaning appeared once per round in a random order, but such that, over the course of the experiment, each participant took the role of signaller for each meaning twice.

Results

Communicative success

We first asked whether participants could successfully communicate about these temporal concepts using the vertical signaling device. From the very first round, accuracy was above chance and improved throughout the experiment, increasing monotonically with each round: M_1: 20.1%, M_2: 21.5%, M_3: 45.1%, M_4: 47.9%. We analyzed this improvement using a mixed-logit model of accuracy on each trial, with a fixed effect of round, random effects of Dyad and Meaning, and by-dyad and by-meaning random intercepts and slopes for Round. The fixed effect of Round was centered to reduce colinearity. This model confirmed that accuracy increased significantly over the four rounds, $\beta = .699$, SE= .167, $p < .001$. Thus, despite the novelty of the task, participants quickly negotiated a set of signals that allowed them to communicate successfully.

Duration and Spatial Extent

Participants could produce a wide range of possible signals, and there was considerable between-dyad signal variability. However, there were also consistent patterns across dyads. Signals made consistent use of *spatial length* to communicate relative temporal duration (Fig. 2). Signals for 'second', a relatively "short" duration, typically involved very little motion and were restricted to the centre of the bar. By contrast, 'day' was signalled using repeated vertical movements that took up half the bar, while signals for 'year' typically covered the entire bar. Thus, *longer durations* appeared to elicit signals that occupied *more space* along the vertical axis. No dyads used the opposite strategy, associating longer *durations* with shorter *spatial extents.* To quantify this pattern, meanings with a definite duration were grouped into three categories by temporal duration: Short, which included 'second' and 'now'; Day, which included 'today', 'day after', etc.; and Year (we excluded 'before,' 'after,' 'past,' and 'future' from this analysis). For each signal, we calculated the maximal space used (MSU) as the distance between the highest and lowest points of the signal. Was there a mapping between duration and spatial extent that was consistent across dyads? We analyzed the MSU of each signal using a linear mixed-effects model, with a fixed effect of Temporal Duration, random effects of Dyad and Meaning, and by-dyad random slopes for temporal extent. As predicted, there was a systematic relation between temporal duration and MSU, with long temporal durations (e.g. 'year') associated with more space (full model compared to reduced model without temporal extent, $\chi^2 = 23.76$, $p < 0.0001$).

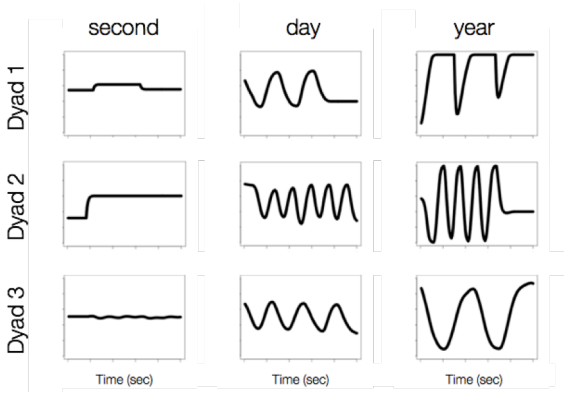


Figure 2*: Duration signals used by three dyads. Each signal is shown as a timeseries (y-axis = vertical location).*

More complex signal-meaning structure

If social interaction and coordination resulted in more complex, systematic conventions, then we should be able to observe patterns in the relations between signals and meanings. Dyads typically settled on a consistent location or movement direction to distinguish meanings that were in the *past* from those in the *future,* but differed in how they mapped past/future to location. About half the pairs used downward movements for *future* and upward movements for *past*, while other pairs reversed this mapping, using *upward* movements for *future*. There was also evidence of sequential patterning, in which dyads combined signals that had been used previously (e.g., 'before,' 'after,' 'year') to create new composite signals (e.g., 'year before,' 'year after'; Fig. 3). Note, however, that these new conventions governing compositionality did not merely reproduce the conventions of English, but sometimes reversed the sequential order (e.g., 'after' followed by 'year').

These internal systematicities can be quantified by relating meaning similarity to signal similarity. Following Kirby, Cornish & Smith (2008), we calculated the Pearson's correlation between pairwise distances in the meaning space and corresponding distances in the signal space. For each dyad, we thus computed two measures for each pair of signals: *Signal Distance,* which captured the similarity of

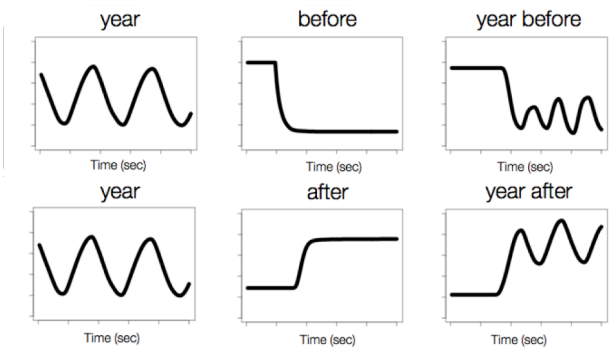


***Figure 3**: Compositional structure in the signals of a dyad.*

the signals, and *Semantic Distance,* which captured the similarity of the associated target meanings.

First, the ***Signal Distance*** (Δ_{signal}) combines the total amount of space used (i.e., MSU) with the focal location of the signal, or Location of Maximal Density (LMD). The former, MSU, is critical because, as described above, this was used consistently to communicate temporal duration. The second, focal location, appeared to be used to contrast past/future or before/after. This was calculated as the point of maximal touch density, between 0 (bottom) and 1 (top). Each signal was thus associated with two values, MSU (total space) and LMD (focal location), and the Signal Distance between two signals (*s1, s2*) was the Euclidian distance between these two points:

$$\Delta_{signal} = \sqrt{(LMD_{s1} - LMD_{s2})^2 + (MSU_{s1} - MSU_{s2})^2}$$

Second, the measure of ***Semantic Distance*** ($\Delta_{semantic}$) between two meanings combined three aspects of semantic difference: Duration (Δ_{dur}), Order (Δ_{ord}), and Class (Δ_{class}). We calculated Duration Difference (Δ_{dur}) by ordering each meaning by its temporal extent, from short or instantaneous durations ('now,' 'second'), to days ('today'), to years ('year after'), to unlimited ranges (e.g., 'past,' 'future'). For instance, for 'second' and 'next year,' the pairwise Duration Difference would be:

$$\Delta_{dur} = |\text{instant} - \text{year}| = |1 - 3| = 2.$$

Order Difference (Δ_{ord}) captured differences in relative temporal order. Sequence and deictic meanings were categorized by relative order: past/before < now < future/after. We collapsed 'past' with 'before', and 'future' with 'after,' based on the finding that English speakers associate these concepts with the same spatial locations (Casasanto & Jasmin, 2012). If both meanings were sequence or deictic, Order Difference was defined as along this ordinal scale; otherwise, if either of the meanings was a duration, Order Difference was defined as 0. For instance, for 'today' and 'next year,' Order Difference would be:

$$\Delta_{ord} = |\text{after/future} - \text{now}| = |3 - 2| = 1.$$

Lastly, we defined pairwise Class Difference (Δ_{class}) as 1 if the meanings belonged to different semantic classes (i.e., duration, sequence, deictic), and 0 if they belonged to the same class. Thus, for 'year after' and 'next year', $\Delta_{class} = 1$, since 'year after' is sequence but 'next year' is deictic. Semantic Distance summed these three components, and is thus similar to an edit distance between two meanings, accounting for duration, order, and semantic class.

Compositional structure is indicated by a positive correlation between pairwise Signal Distance (Δ_{signal}) between all possible pairs of signals in a system and pairwise Semantic Distance ($\Delta_{semantic}$) between the associated target meanings (Kirby et al., 2008). This was calculated for each dyad and each round. Structure increased over the course of dyads' interactions (Fig. 4). A linear mixed-effects model of this measure of compositionality, with Round as a fixed effect and Dyad as a random effect, showed an increase of 0.37 (S.E. +/-0.13) each round. Comparison to a reduced model without a fixed effect of Round confirmed that structure increased significantly ($\chi^2 = 6.36$, $p < 0.05$). To compare this observed structure to the structure we might observe by chance, the actual systems produced by dyads in our experiment were compared to 1000 simulated systems, in which the signal values were created randomly. The dotted line in Figure 4 indicates the upper limit of the 95% confidence interval for random data; values above this show a significant amount of internal structure. While most systems were not significantly structured in the first round, by the final round almost all systems had significantly more structure than chance.

Sensitivity to regularities in the semiotic system

Participants quickly became adept (Accuracy in Round 4, $M = 48\%$; chance < 6%). This could be because they were simply memorizing each signal-meaning pair, without having internalized any of the structure documented above. However, if participants *were* sensitive to the systematic use of space to represent different aspects of time, then we should be able to find evidence of this in their errors: incorrect guesses should be semantically related to target meanings.

To quantify the systematicities in participants' errors, we

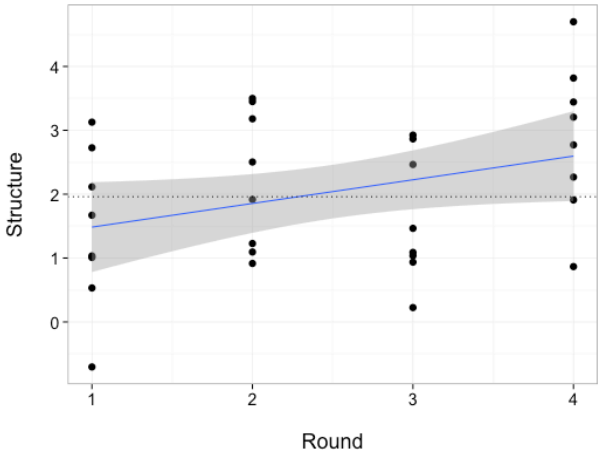


***Figure 4:** System internal structure (y-axis) increased with social interaction. Each point represents a system deployed in a particular round by one dyad.*

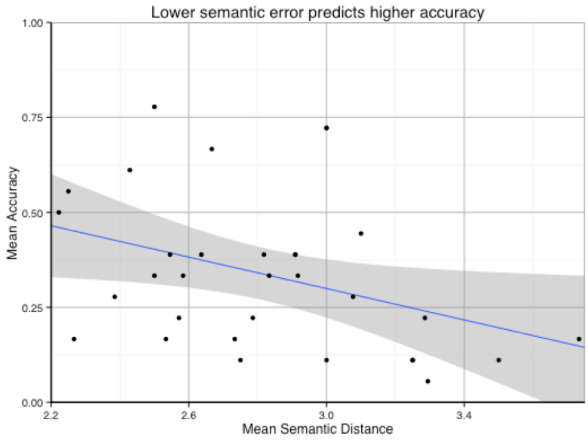


Figure 5*: Dyads were more successful overall if their incorrect guesses were semantically related to the correct target meaning. Each dot represents one round.*

used our measure of Semantic Distance ($\Delta_{semantic}$), described above, to compare incorrect guesses to correct target meanings. If dyads gradually learned the conventions of their emerging semiotic system (e.g., compositional structure), then errors should become gradually more systematic, with incorrect guesses becoming systematically closer to the correct target over time.

Indeed, Semantic Distance between target meanings and incorrect guesses decreased monotonically over the course of the interaction: M_1 = 3.13, M_2 = 2.79, M_3 = 2.75, M_4 = 2.61. To analyze this effect, we used a mixed effects model with a cumulative logit link function, since Semantic Distance is an ordered, categorical variable. The full model had a fixed effect of Round, random effects of Dyad and Target Meaning, and random intercepts and slopes. Semantic Distance shrank significantly as the game progressed, b = -0.55, z = -4.43, p < .001. This full model was better than a model without a fixed effect of Round, likelihood ratio = 10.5, p = .001.

A reduction in Semantic Distance, moreover, was accompanied by an increase in communicative success: Dyads were significantly more likely to guess correctly overall if their *incorrect* guesses were semantically closer to the target meanings. According to a linear mixed-effects model of mean accuracy in each round, with a random effect of Dyad, communicative success increased as the mean semantic distance shrank between incorrect guesses and correct targets, b = -0.21, t = - 2.24, p = .033. In other words, if a dyad's *incorrect* guesses were generally closer to the correct target meaning, then they also made more correct guesses overall (see Fig. 5).

Discussion

We investigated the emergence of structured semiotic systems for communicating about an entirely abstract domain, *time*. Despite the novelty of the signalling device, and the restricted range of possible signals, dyads established successful systems. Signal-meaning structure increased over time, with the emergence of systematic mappings between space and time. These mappings were often combined productively, suggesting compositionality. And with this increased structure came increased communicative success. Thus, structure emerged, participants used this structure when guessing, and this improved communication.

The presence of early, shared regularities and later, idiosyncratic structure sheds light on the role of individual cognitive biases, social coordination, and interaction history on the emergence of systematic mappings between space and time. The mapping between temporal duration and spatial extent, for instance, was established early in the first round and followed the same direction in all dyads (short=small). These results suggest that participants had shared, strong initial biases about the relation between spatial extent and temporal duration, perhaps due to innate intuitions, systematic polysemy in language, or systematic conflation of length and duration in our lived experience (Lakoff & Johnson, 1980; Walsh, 2003; Winter, Marghetis, & Matlock, 2015). In contrast, although almost every dyad eventually began to use location to distinguish past from future concepts, the specific strategy was more variable across dyads. This suggests that participants lacked strong, shared biases, and shared *interaction history* was thus more critical here. This finding also confirmed that our American participants did not have a strong initial bias for associating location or direction with past/future. The emerging structures of these semiotic systems reflect an interplay between individual biases and social coordination. We suspect similar dynamics were likely at work in the emergence of space-time mappings in natural language.

This combination of variability and universality mirrors regularities in human cognition and natural language. On the one hand, evidence from non-human animals and child development suggests that spatial extent and temporal duration are tightly linked, perhaps innately (for a review: Winter, Marghetis, & Matlock, 2015). On the other hand, for concepts other than magnitudes, there is considerable cross-cultural variability in how time and space are intertwined. For instance, while English associates the future with the space ahead of the body, the Aymara of Chile reverse this mapping, associating past with the space *behind*, and the Yupno of Papua New Guinea associate the future with *uphill* (Núñez & Cooperrider, 2013). Thus, the interaction between cognitive biases and social coordination during our communication game is a microcosm of a larger network of biases and pressures. These mechanisms must be considered together to account for the mix of universality and variability that we find around the world.

While the semiotic systems in our study began to capture the structure of the meaning space, no fully systematized language emerged. As we saw, some of the more similar meanings in the set (e.g. "tomorrow" *vs.* "day after") were still confused in the later rounds. What would be needed for these systems to develop ways to distinguish even these very similar sequence and deictic concepts? Our hypothesis

is that these early-emerged patterns will evolve into more regular systems once they are transmitted across multiple generations of interacting users. Iterated learning studies have found increasing regularization when languages are transmitted over generations (e.g. Kirby et al., 2008). When a system is transmitted from generation to generation it adapts to become more learnable and more structured. We are currently conducting a follow-up study using *iterated* communication (Tamariz et al, 2012; Verhoef et al., 2015).

How specific are these results to the domains of time and space? On the one hand, space and time may have a privileged relation, tightly linked in thought, language, and gesture (Núñez & Cooperrider, 2013; Winter et al, 2015). On the other hand, the domain of time is not alone in relying on space for its conceptual and linguistic structure. Abstract domains as varied as number and morality make systematic use of spatial language (Lakoff and Johnson, 1980). For number, at least, there are also strong early biases to associate numerical magnitude with spatial extent. The interaction between cognitive biases and cultural processes has been identified as an important drive in the emergence of structured semiotic systems before. What the domain of space and time contributes is a clear identifiable distinction between parts of structure that appear to be more rooted in strong shared biases and parts that result from cultural evolution, all within one emerging system. These findings are likely generalizable to other such domains.

This study is limited by the extensive linguistic experience of the participants, who already had words in their native language (English) for the meanings to be communicated. This may have influenced their ability to create a new semiotic system. Indeed, the association between length and duration is reflected in English polysemy ("long time", "long stick"). However, as we saw, much of the new structure (e.g., associations between up/down and past/future) consisted of mappings that are absent entirely from English. Future research may explore the negotiation of novel semiotic systems for concepts that lack associated lexical items.

Conclusion

To explore the link between individual-level behaviors and population-level communication systems, consideration of different mechanisms is needed (Beckner et al, 2009). Here, we focused on the interaction between initial individual biases and cultural processes and how these may contribute to the emergence of space-time mappings. By foregrounding the interplay between such mechanisms that operate on disparate timescales, laboratory experiments can shed light on the commonalities and variety found in space-time mappings in languages around the world.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., et al. (2009) Language is a complex adaptive system. *Language learning*, 59, 1-26.

Caldwell, C. A., & Smith, K. (2012). Cultural evolution and the perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE*, 7, e43807.

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *BBS*, 31(5), 489–509.

Fay, N., Garrod., S., & Roberts, L. (2008). The Fitness and Functionality of Culturally Evolved Communication Systems. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 363, 3553-3561.

Fusaroli, R., & Tylén, K. (2012). Carving language for social coordination: A dynamic approach. *Interaction Studies*, 13, 103–123.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of Representation: Where Might Graphical Symbol Systems Come From? *Cognitive Science*, 31(6), 961-987.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Haspelmath, M. (1997). *From space to time.* Munich & Newcastle: Lincom Europa.

Hockett, C. (1960). The origin of speech. *Scientific American*, 203, 88–96.

Kirby, S., Cornish, H. & Smith, K. (2008) Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*. 105(31), 10681-10686

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28C, 108-114.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago, IL: University of Chicago Press.

Núñez, R. E., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *TiCS*, 17, 220–229.

Scott-Phillips, T. & Kirby, S. (2010). Language evolution in the laboratory. *TiCS*, 14(9), 411–417.

Steels, L. (1997). The synthetic modeling of language origins. *Evolution of communication*, 1(1), 1–34.

Tamariz, M., Cornish, H., Roberts, S. & Kirby, S. (2012) The effect of generation turnover and interlocutor negotiation on linguistic structure. In *Proceedings of EVOLANG9*. World Scientific, (pp. 555-556).

Theisen, C., Oberlander, J. & Kirby, S. (2010) Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, 11(1): 14-32.

Verhoef, T. (2012) The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4(4), 357-380.

Verhoef, T., Roberts, S. G., Dingemanse, M. (2015) Emergence of systematic iconicity: Transmission, interaction and analogy. *37th Annual CogSci Conference* (pp. 2481-2486) Austin, TX: Cognitive Science Society.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *TiCS*, 7, 483–488.

Winter, B., Marghetis, T., & Matlock, T. (2015). Of magnitudes and metaphors. *Cortex*, 64, 209-224.

Transfer of Cognitive Skills in Developmental Tasks

Sarah Rupp (srupp@uos.de)
Niels A. Taatgen (n.a.taatgen@rug.nl)
Institute of Artificial Intelligence, University of Groningen, Nijenborgh 9,
9747 AG Groningen, Netherlands

Abstract

The main question we try to answer in this paper is whether stage-like progression in cognitive development can be explained by transfer of cognitive skill among tasks. We focus on the following question: To what extent does training on one task improve the performance on another task? The tasks are Piaget's (1959) Balance Scale Task and Number Conservation Task, and a task that we will call the Une-Sentence Task, which is taken from Karmiloff-Smith's (1979) experiment on the acquisition of determiners in French. We re-implemented already existing models within the framework of the PRIMs cognitive architecture (Taatgen, 2013). Each task was subdivided in certain stages related to the complexity of the problem-solving strategies. We show that mastery of a certain stage of a problem becomes easier if a higher stage of another task is mastered first.

Keywords: Transfer, PRIMs, Cognitive Architecture, Developmental Tasks, Cognitive Modeling

Introduction

A central topic of debate in developmental psychology is whether children develop in stages, or if development is a gradual monotonic process. The idea of stages was originally introduced by Piaget in his *theory of cognitive development* (1959). The claim about stages of development from early childhood up to adulthood is still influential, even though the idea of across-the-board stage transitions is unlikely to be true. Nevertheless, it is quite plausible that progress in a certain domain can also support or enable progress in another domain. In this paper we will explore this idea in the context of concrete cognitive models that learn developmental tasks that need similar strategies.

The key idea in the modeling effort is that progress on a certain task leads to new cognitive skills that might provide the key missing piece that is necessary for progress on another task. If the new skill is applicable in multiple domains, this may give the impression of a stage-like transition, even though specific learning has to be done within each of the individual tasks.

To identify phases or stages in performance, Siegler introduced the *rule-assessment approach* (Siegler, Strauss, & Levin, 1981; Siegler, 1976). The assumption of that approach is that a certain set of rules produces a particular strategy that can later be identified in behavior. For example, in the balance scale task that we will discuss in more detail later on, both weight and distance have to be taken into account to determine the correct answer. However, early in development children only take weight into account. This produces a characteristic pattern of errors that uniquely identifies the strategy. Siegler's assumption is that progress from one stage or phase to the next was produced by general problem solving strategies. In other words, progress is independent of progress on other tasks, but only dependent on knowledge that doesn't change.

Differently but also similarly, Alison Gopnik advocates the *theory-theory* (Gopnik, 2003). This theory assumes children use reasoning methods similar to what scientists use to generate and refine their knowledge about the world. The theory-theory integrates the process of scientific research and children's ability to construct new knowledge.

An alternative mechanism to explain developmental progress is transfer from other tasks. The idea of transfer was first introduced by Thorndike and Woodworth (1901). Their conclusion, however, was that there is in fact very little evidence for transfer, and that people have to discover a strategy for a new problem all over again even if it is very similar to a problem they just solved. Only when the knowledge needed for the task is identical to knowledge from another task will there be transfer: the *identical elements* theory, which is still assumed to be the dominant explanation for transfer (or, lack thereof).

The PRIMs Theory

The PRIMs theory (Taatgen, 2013) offers an alternative to both the identical elements theory, and approaches that assume that new skills are produced by general immutable strategies. The assumption of PRIMs is that general cognitive skills are learned as a byproduct of learning specific tasks. PRIMs builds on the successful ACT-R (Adaptive Control of Thought - Rational) architecture (Anderson, 2007), and inherits most of its principles. However, the key difference is the basic building block of skill. In PRIMs these are Primitive Information Processing Elements (hence PRIMs). PRIMs are clustered together into cognitive operators that carry out the skill.

This method allows you to construct nearly any task from a fixed set of building blocks. PRIMs compare information between different cognitive modules (vision, memory, etc.) or move that information from one module to another. If a particular combination of PRIMs is used more often, they will be compiled into a single unit that can be used in any situation that that combination can be used. This means that if a particular combination of PRIMs is useful and is trained in a particular task, but that combination is also useful for another task, that other task can be learned faster because it already has some of the necessary building blocks. This mechanism can explain effects of transfer, even though the process does not actually transfer knowledge from one task

to another. Instead general knowledge (the useful combinations of PRIMs) is generated by learning one task, and reused by another task. For example, a task in which a letter is displayed on the screen, and a judgment has to be made whether the letter is a vowel requires transforming a perceptual input into a memory query with a task-specific component (vowelness). This step takes four PRIMs, which can be reused for a task in which a number is displayed and the subsequent number has to be named.

Goal of this paper

This paper builds on Gittelson & Taatgen (2014), who modeled stages in the balance scale task and a decision making task. The limitation of that study was that the decision making task was not a typical developmental task. In this paper we will examine three developmental tasks in which multiple stages have been identified. The three tasks have a common strategic element: handling multiple aspects or dimensions of a particular problem. Typical stage 1 behavior is to only take one dimension into account. In a later stage, two dimensions are recognized, but are only handled one at a time. In the final stages information from multiple dimensions is integrated to reach a decision.

The approach is to build separate PRIMs models of all three tasks for each of their stages, so in total 11 models. The main question we want to answer is how easy it is to learn a new stage of a task given different types of prior knowledge. We will do this by determining the amount of transfer between combinations of models. First, we will examine how a lower stage can support a higher (typically more complex) stage. For example, to what extent is mastery of stage 1 helpful in learning stage 2. We will examine a more interesting question next: to what extent is knowledge of stage 2 of another task helpful in learning stage 2. So, is mastery of stage 2 in task A helpful in getting to stage 2 of task B. Finally, we will look at the "stage-transition" question: if you have mastered stage 1 of a task A, and stage 2 of task B, is this combination a better support to learn stage 2 of task A than just mastery of stage 1? If this is the case, progress in stage 2 in any task can trigger progress in many other tasks that are still in stage 1, producing a more general stage-like transition.

Developmental Task and Model

In the following section we will explain the three tasks we used: the Balance Scale Task (further on also Balance Task called), the Number Task and the Une-Sentence Task. We will also explain the structure of the models for each of the tasks. The first two tasks are mainly based on Siegler's (Siegler et al., 1981; Siegler, 1976) work and is re-implemented for our purposes in PRIMs. The third task was already implemented by Zondervan and Taatgen (2003), which is the basis for the PRIMs implementation.

Balance Scale Task

The Balance Task was originally described by Inhelder and Piaget (1959) and used by several other researchers. As shown in Figure 1, the children had to predict to which side the scale would tilt, or if it would stay balanced, taking into account the distance to the fulcrum and the weight of the stacks.

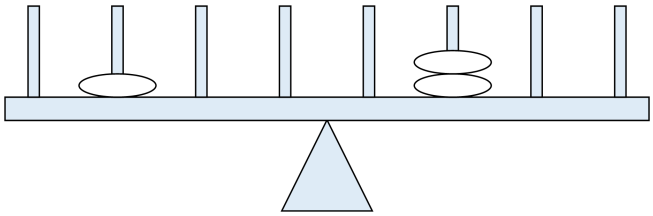


Figure 1: Balance Scale Task

Siegler labeled weight as the dominant dimension and distance as the subordinate (Siegler et al., 1981). He implemented each of the stages by a set of decision rules. For our model we re-implemented these rules directly into PRIMs. In correlation to the real world experience and according to the Piagetian stages, children need to learn to integrate multiple dimensions of the problem. In the beginning they typically only consider a single dimension. In the second stage they will consider more dimensions – but only one at a time. In the third stage they take multiple dimensions into account, and learn to weigh the dimensions properly. But there are still cases where there is a need of the proper combination of both dimensions and will always lead to the correct solution. This is what we call stage four. There is a debate going whether children even ever reach that stage without being taught explicitly. As it is proposed by Inhelder and Piaget (1959), children will reach that stage, but it has been even rejected by the findings from Siegler (1976). As we will see this is only an issue in the Balance Task, so it will not play a major role in our discussion.

In the model the first stage of the Balance Scale Task will decide just on the basis of the dominant dimension. In the second stage, the model will work just as the first stage, but in cases of equal values on the dominant dimension the model will take the subordinate dimension into account and decide on that one. In the third stage the model will take the subordinate dimension into account either way. So even in the case of different dominant dimension, it will now also take the subordination into account. If the subordinate dimension is equal, the decision is based on the dominant dimension. If both the dominant and the subordinate dimension differ, the model will guess (Siegler called this case 'Muddle through'). The fourth stage will solve the 'Muddle through' problem, with taking the dominant and the subordinate dimension comparable into account, which arises with an algebraically combination of the distance and the weight on each side (the torque law).

Number Conservation Task

Another developmental task we used is the Number Conversation Task. Like the Balance Task, it has its roots in the Piagetian account, and was also discussed by Siegler. In the number conservation task children have to decide which

of two rows has more coins. The task itself concerns two phases. The first phase consists of the initial position of the coins, which are two rows of equal length and quantity. In phase two the experimenter performs a transformation to the lower row and the participant has to determine which row contains more coins. A transformation can consist of lengthening or shortening the row, and/or adding or subtraction coins from the row. This makes the task very similar to the Balance Task: both the length and numerosity are dimensions to be considered, and length is initially dominant. The difference is, of course, that the solution only involves numerosity, the subordinate dimension.

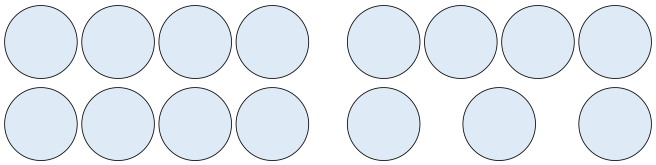


Figure 2: Number Conservation Task (Phase one is the left cluster and phase two the right cluster.)

Because of the similarity, the models for the task are very similar to the Balance Task models. Children start to learn to make the decision based on a single dimension, length. After that they will take a second dimension into account, but again only in case where the first dimension leads to a non-sufficient solution, which is the case if the values for the dominant dimension are equal, therefore considering the second dimension will help to decide. Next step is to consider both stages at any time. But in cases of two ambiguous outcomes for each dimension, children will guess. The final stage 4 rule is of course different, and similar again to stage 1, except with a different dimension.

Une - Sentence Task

The third task we modeled is based on Karmiloff-Smith's experiment on the acquisition of the function of determiners in French (Karmiloff-Smith, 1979). The French feminine word for "a", "une", has different meanings. It can be used to introduce a new item ("a balloon"), but also to indicate a specific number ("one balloon"). To assess the reasoning process that children to determine the meaning of "une", Karmiloff-Smith showed them pictures of two playgrounds, one with a girl and the other with a boy (Figure 3).

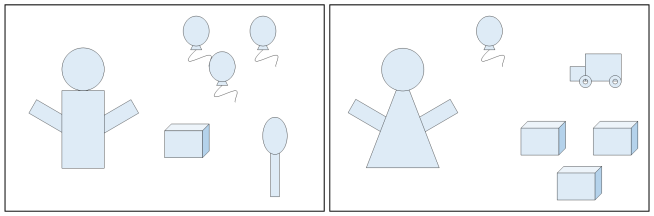


Figure 3: Playgrounds

Those playgrounds were constructed precisely; e.g. if the boy had three balloons the girl would have one. The experimenter always 'talked' to either the boy or the girl and the child had to decide to which one the experimenter was talking. Questions were of the form: 'Please lend me the X.' or 'Please lend me a X'. The word 'la' is the French word for 'the'. The French word 'une' is translated in English 'one' (number) and 'a' (indefinite article). Therefore by hearing 'Please lend me a box' (consider a playground situation as in Figure 3), the correct response of the child would be: 'You are talking to the girl'. However, if the child would answer 'boy', it would indicate they interpret 'une' as 'one'.

The task was constructed to test whether children are able to distinguish between 'une' and 'la' sentences. Results revealed that there did almost perfectly fine on 'la' sentences, but there appeared a U-shaped performance curve for 'une' sentences. At younger ages (3-4 year olds) they do well on 'une' sentences. Then there a low point at the age of 5. And then at rising ages (the testing was up to the age of 10) the performance on 'une' sentences was nearly perfect again.

Taatgen and Zondervan presented a model of the performance curve of the indefinite French article 'une' (2003) based on the representational redescription (RR) theory (Karmiloff-Smith, 1992). Generally the stages describe, as well as the stages of the other two tasks, the cognitive development. The main idea of the performance drop and the rise of performance is, that children first consider only the one dimension they know, which would be 'une' as an indefinite article – which indeed leads to the correct answer if they have to make a choice between two playgrounds: they pick the playground with multiple items. The subsequent drop is explained by a switch to the other dimension. The theory is that they now consider 'une' as a number, which leads to an incorrect response. In the final stage the children learn how to differentiate between the two interpretations by taking into account another dimension of the problem, in this case whether or not the object in question is already the focus of the discussion. If boxes are already to topic of conversation, the 'one' interpretation is correct, but if they are not, the 'a' interpretation is correct. This means that in stage 3 the decision process consists of two steps: first settle the topic question, and based on that make the right choice between the two playgrounds.

Similar to the other task and their stages, for the first stage of the Une-Sentence Task, the decision is based on one dimension (you might call it the dominant dimension). The first stage decides correctly with the information of the dominant dimension. The second stage also only take dominant dimension into account, but decides strictly incorrectly (in the context of this test, normally interpreting 'a' as 'one' does not lead to many problems). In the third stage in the task two dimensions have to be taken into account in a way that is similar, but not identical to the stages 3 in the other two tasks.

Method

To assess the amount of transfer between (stages of) tasks, we used the methodology developed by Katona (1940). This involves a comparison of two training situations. The first step is to train the model just on the task in question (let us call it task T1), and see how much progress is made after a certain amount of training (200 trials in our case). In the second step (after cleaning the environment and staring of with no prior knowledge) training of a different task will be performed. There are two cases: In the first case the model is trained on only one different task T2, which I called *one task training*. In the other case it will be trained on two different tasks (in the figure noted as T2 and T3), this will be mentioned as (two task) *shared training*. Afterwards the model is always tested on T1. We can express the benefit of T2 (or the combination T2 and T3) on T1 by comparing performance after training (Figure 4).

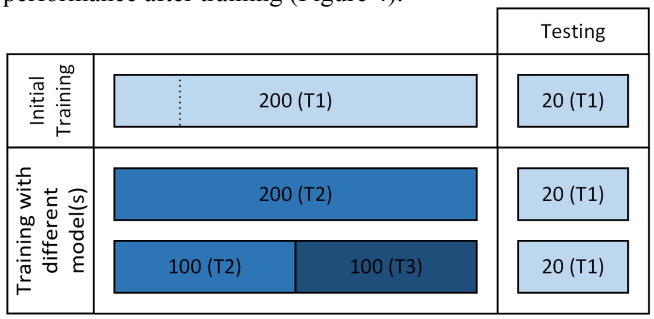


Figure 4: Illustrates the different training procedures necessary to assess transfer from task T2 to task T1, or from the combination of T2 and T3 to T1.

The transfer calculation is as follows:

$$\text{transfer} = \frac{T2_{\text{training}}}{T1_{\text{training}}} * 100\%$$

$T2_{\text{training}}$ Difference of average time of the first model (without any training) and average time after the training session on a second model (in case of shared training: the training session includes two different models)
$(T1_{\text{without_training}} - T1_{\text{with_training_T2}})$

$T1_{\text{training}}$ Difference of average time of the first model (without any training) and average time after the training session on itself.
$(T1_{\text{without_training}} - T1_{\text{with_training_T1}})$

Results

The results are separated into three parts, as indicated in the introduction.

Within-task transfer

We will first examine the transfer of a lower stage to a higher stage within a task. Figure 5 shows percentages for each of the three tasks. The arrow indicates the testing direction. For example: Balance Task, Stage 1 to Stage 2 has the value 46 – which indicates the transfer from stage 1 to stage 2. What we can see is that lower stages typically support the next stage fairly well. The most difficult transition seems to be from stage 1 to 2 (or 2 to 3 in the Sentence Task). Noticeable for the Number Task is that the transfer of the third stage to the fourth is not as high as the other values. The reason is that stage 4 is in fact a lot simpler than stage 3, whereas is most of the other cases a higher stage is an extension of the lower stage. Another point of notice for the Sentence Task is the transfer value for the first to the second stage, which is around 100% (noise in the simulation explains the deviance from 100%). This indicates that the structure of the solution is identical between the two stages. This problem arises again in the next part. The 100% does not mean that mastery of stage 1 automatically leads to mastery of stage 2 in the sentence task. The model still has to discover that it has to use a different dimension in stage 2 than in stage 1, and aspect we have not modeled here (but see Zondervan & Taatgen, 2003).

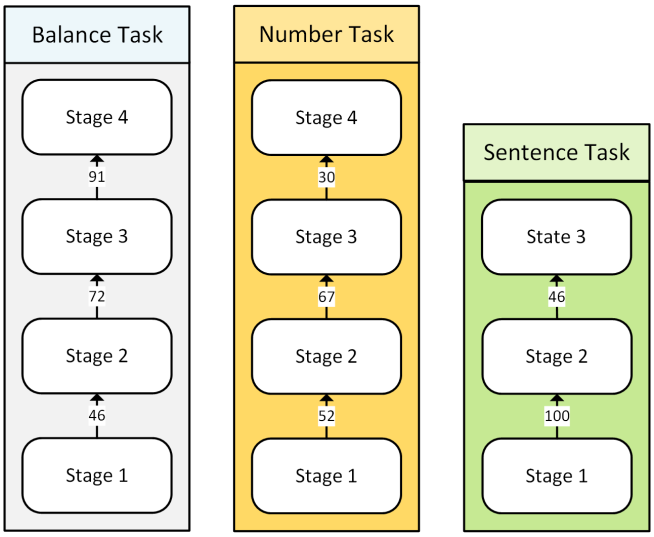


Figure 5: Within-task transfer

Transfer between almost identical tasks

The Balance Task and the Number Conservation Task are almost identical in their decision structure for each of the different stages. We therefore would expect that if a particular stage is mastered in one of the tasks, this would transfer perfectly or almost perfectly to the other task in the same stage.

Figure 6 shows that this indeed the case: transfer between the first three stages is in all cases very high. This means

that after a stage is mastered in one of the stage, the cognitive skills are available to also learn that stage in a different task.

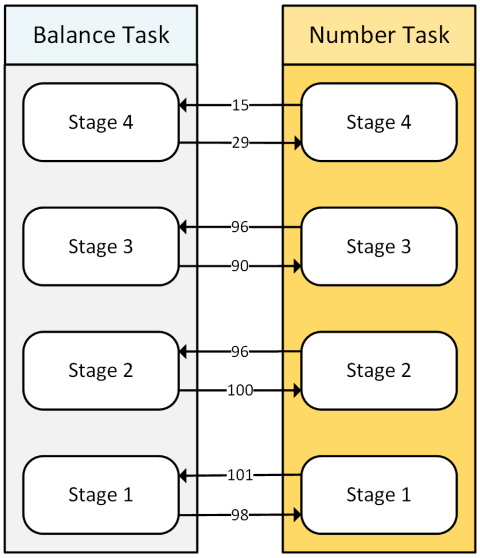


Figure 6: Transfer between the Balance and Number Task

The only exception is stage 4, but there the two tasks differ: whereas the Balance Task requires a complex multistep decision, the solution for the Number Task is quite simple, and only requires the skills learned for stage 1.

Shared Training

The most interesting case of transfer is when the model has mastered a certain stage in task A, but has already progressed to a higher stage at task B. Will skill in task B help progress on task A? To examine this we tested progress on the Balance Task given different levels of training on the Sentence task, two tasks that have similarities but are not identical in strategy.

Figure 7 shows the results. They reveal that there can be more transfer with shared training. As shown in the middle panel of the figure, training on both sentence stage 1 and balance stage 1 leads to the same transfer to balance stage 2 as just training on balance stage 1 (even though the total number of training trials is the same). Even better is the performance after shared training with the stage 3 of the sentence task: this leads to much better (62% vs. 46%) transfer than training on just stage 1 of the Balance Task. The shared training is extra effective because stage 3 of the sentence task includes skills that can handle multiple dimensions in a problem, something that is required for stage 2 of the Balance Task but that is not part of stage 1 of the Balance Task.

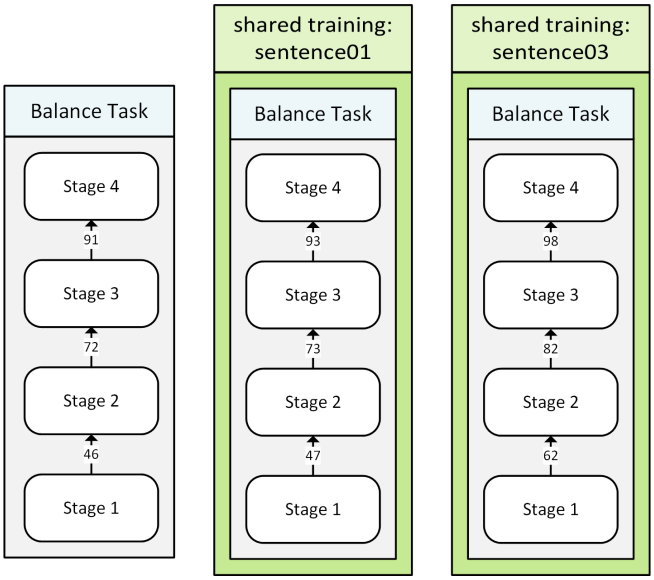


Figure 7: Shared training session. Left: repeats the within-task task transfer chart; Middle: indicates shared training with the first model of the sentence task and each model of the Balance task respectively; Right: similar to the middle column, but this time a shared training with the third model of the sentence task and each balance model.

The additional support for the stage 1 to 2 transition is particularly encouraging, because that is the hardest step to make (Figure 5). Training on stage 3 of the Sentence task is also helpful for the later transitions in de Balance task, although the effect is smaller.

Discussion

The main question of this paper is whether stage-like progress in development can be explained by the discovery of general cognitive skills. In the example we discussed here, the general skill is how different dimensions of a decision tasks are integrated. All three tasks we have looked at start with a strategy that only takes a single dimension into account, and then progress to more advanced strategies that integrate multiple dimensions. As we have seen in the simulations, different tasks can support each other. In the case of the Balance Task and the Number Conservation Task the underlying structure was almost identical. This means that progressing on one of the two tasks should make it really easy to progress to the next stage at the other task. Even when two tasks only overlap in structure, like is the case in the Balance Task and the Sentence Task, can progress on the other task help the other task. For example: We are testing the model of the second stage of the Balance Task, and train on the second stage of the Number Task, the transfer achievement is around 96 %. So we could conclude that mastery of a certain stage of task A, basically makes you are able to master another task at the same stage. If we call the mastery of a stage of a task, a skill that we achieved, we could talk about an emergent property. This holds at

least for tasks that share many similar structural elements. Therefore new strategies allow you to be better at all tasks (so all tasks within that 'strategy-level'), which would be then something similar to a so-called 'stage'. These findings do not support the idea of a sudden jump from stage to stage. It rather supports the idea of a relatively fast transition from one stage to another because of the replicated use of elements, structures or strategies of a task. Also interestingly is the fact that a shared training setting does support the performance of the tested task sometimes even better. For example: We are testing the model of the Balance Scale Task of the second stage. And we train on the model of the Balance Task stage 1 and the model of the Sentence Task stage 3. We do see, that this combination gives a better result (62 % transfer), than training only on the lower stage of Balance Task stage 2, which would be Balance Task stage 1 (which is 46 %). We can conclude from this observation that it could be the case, that training on different tasks result in better performances, than only training on one specific task.

A limitation of this study is that the actual transitions between stages are not modeled, although ACT-R models of such progress are available for both the Balance task (van Rijn, van Someren & van der Maas, 2003) and the Sentence task (Zondervan & Taatgen, 2003). What this study shows, however, is a limitation of those models, and many other models that simulate developmental tasks, in that they describe progress on a task in isolation. What we shown here is the possibility that skills in general are inter-connected, and that progress on them should be studied in a broader context than just a single task.

In the end this is a theoretical account. The simulations were performed in an isolated environment and the tasks were modeled arbitrarily, of course with regard to earlier work that has been done on this field. Therefore the models were still plausible constructed based on already existing models of that kind of task. Furthermore the data we showed are theoretical predictions, but it is indeed interesting if this would be tested and be compared to empirical data. And if our prediction turn out to be true, or partially true, it would be interesting to look at possible implications for education, in a sense of teaching procedures or methods that are commonly used. At the very least this study suggests that diversity in training has benefits over just singular training.

Acknowledgments

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen.

References

Anderson, J.R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press..

Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. In L. Antony & N. Hornstein (Eds)., Chomsky and His Critics (pp. 238-254). New York: Basil Blackwell

Gittelson, L. & Taatgen, N.A. (2014). Transferrring primitive elements of skill within and between tasks. *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 529-534). Austin, TX: Cognitive Science Society.

Piaget, J., & Inhelder, B. (1959). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures* (Vol. 84). Routledge.

Karmiloff-Smith, A. (1979). A functional approach to child language. A study of determiners and reference. Cambridge, MA: Cambridge University Press.

Karmiloff-Smith, A. (1992). Beyond Modularity: A Developmental Perspective on Cognitive Science. *Behavioral and Brain Sciences, 17*, 693–745.

Katona, G. (1940). *Organizing and memorizing*. New York: Columbia University Press.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*(4), 481–520.

Siegler, R. S., Strauss, S., & Levin, I. (1981). Developmental Sequences within and between Concepts. *Monographs of the Society for Research in Child Development, 46*(2), 1–84.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review, 120*(3), 439–471.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions: III. Functions involving attention, observation and discrimination. *Psychological Review, 8*(6), 553–564.

Van Rijn, H., Van Someren, M., & Van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science, 27*(2), 227-257.

Zondervan, K., & Taatgen, N. A. (2003). The determiners model: a cognitive model of macro development and U-shaped learning in a micro domain. *Proceedings of the Fifth International Conference on Cognitive Modeling*, (pp. 225–230).

The Description-Experience Gap in Risky Choice Framing

Gaëlle Vallée-Tourangeau (G.Vallee-Tourangeau@kingston.ac.uk)
Department of Management, Kingston University London
Kingston Hill, Kingston upon Thames, KT2 7LB, UK

Frédéric Vallée-Tourangeau (F.Vallee-Tourangeau@kingston.ac.uk)
Department of Psychology, Kingston University London
Penrhyn road, Kingston upon Thames, KT1 2EE, UK

Madhuri Ramasubramanian (madhuri.aiesec@gmail.com)
Department of Psychology, University of Madras
Chepauk, Chennai-600 005, India

Abstract

We examined whether the classical framing effect observed with the Asian Disease problem could be reversed when people make decisions from experience. Ninety-five university students were randomly allocated to one of three conditions: Description, Sampling (where the participants were allowed to sample through the outcomes presented as a pack of cards) and Interactive (where the participants were invited to spread out all possible outcomes in a sample) and made three gain-framed choices and three loss-framed choices, with two filler tasks after the first three choices. The results revealed a significant interaction effect between framing and choice condition. In the Description choice condition, participants were more risk-seeking with loss-framed problems. This pattern was reversed in the Sampling choice condition where participants were more risk-seeking with gain frames. Finally, the Interactive choice condition resulted in a classic pattern of framing effect, whereby people were more risk averse in the domain of gains.

Keywords: description-experience gap; risk-taking; framing effect; Asian disease problem; interactivity; distributed cognition.

Introduction

Choices are ubiquitous. They can be inconsequential (deciding what to wear) or involve sizeable risks (investing in a volatile stock). Decision-making often comes with a strong feeling of agency: People's preferences for choice alternatives as well as their perceptions of the risks associated may feel personal and unique yet, those preferences are highly contextual (Lichtenstein & Slovic, 2006).

An illustrious example of the impact of context on people's choices under risk is the so-called framing effect, most often illustrated with the Asian Disease choice task (Tversky & Kahneman, 1981). In this task, where people are asked to choose between two programmes for combatting "an unusual Asian Disease" expected to kill 600 people, based on the exact scientific estimates of the consequences of those programmes. Those estimates are then either presented in terms of lives saved (positive frame) or lives lost (negative frame). Thus the gin frame reads:

> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

By contrast, the loss frame reads:

> If Program A is adopted, 400 people will die.
>
> If Program B is adopted, there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

In both frames, people are asked which of the two programmes they would favor. Programme A results in a sure outcome and is considered to be a risk-averse choice whereas Programme B results in an uncertain outcome, a risk-seeking choice. Typically, the gain-framed option elicits risk averse preferences as most people select the sure option. Conversely, the loss-framed option elicits risk-seeking preferences as most people presented with this frame opt for the uncertain alternative. This seminal finding has been replicated in numerous studies and applied in various domains (e.g., see Kühberger, 1998; Maule & Villejoubert, 2007 for reviews). Importantly, it illustrates how choice preferences can arise from the design of information in a decision-maker's environment.

Recently, research has shown that people's risk preferences can, not only be influenced by information design, but also by how this information is experienced. Hertwig, Barron, Weber and Ever (2004) compared people's preferences in a series of lotteries offering certain or uncertain monetary outcomes whose distributions were either described on a computer screen (e.g., "Get $3 for sure") or experienced through sampling by pressing a computer button (e.g., $3, $3, $3, ..., $3). When lotteries were described, people preferred the sure option in gain-framed lotteries and the uncertain option in the loss-framed lotteries. The key finding of their study, however, was that this pattern of preference reversal was itself flipped when people experienced the outcome distribution through sampling instead of reading a summary description: under sampling, people preferred the risky option in gain-framed lotteries and the sure option in the loss-framed lotteries. Although the preference reversal between gain-framed lotteries and loss-framed lotteries under description is similar to the framing effects observed with the Asian disease problem, the lotteries used by Hertwig et al.

(2004) were not designed to test for framing effects and thus differed from the Asian problem on at least two noteworthy features. First, whereas the Asian disease problem specifies a reference point (an expected loss of 600 people), the reference point was not specified in Hertwig et al.'s gambling tasks. Second, expected values and framing were confounded in the lotteries used by Hertwig et al. (2004) such that the uncertain option had a higher expected value in the gain frame whereas the sure option had a higher expected value in the loss frame. By contrast, the alternatives in the Asian disease task were designed to have equivalent expected values. A first objective of the present research was therefore to examine whether experiencing the outcomes in a classic framing task rather than reading a description of these outcome would also lead to a reversal of the classic framing effect.

In both the instance where outcome information is summarised in a short description and in the instance where it is experienced through sampling, people play a relative passive role in the acquisition stage. This, we have argued elsewhere (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015; Villejoubert & Vallée-Tourangeau, 2011), seems to be a consequence of how decision-making in particular and cognitive processing in general has been conceived by scholars: as a series of mental operations on a mental representation of the outside world. An alternative, systemic perspective, recasts cognition as involving the co-occurrence of the cognitive processing of mental representations and the physical processing of material presentations. From this perspective, to understand how people reason or make decisions, one needs to understand how knowledge and thoughts emerge from the co-ordination of internal and external resources. For example, we previously showed that allowing adults to interact with the statistical information in Bayesian reasoning tasks through the manipulation of playing cards enabled most to succeed at solving these tasks without training (Vallée-Tourangeau et al., 2015).

A second objective of the present research was to examine whether (and how) increasing the affordances (i.e., the action possibilities) offered by the information display would impact framing effects. To achieve this, we allowed some of our participants to spread and rearrange the possible outcomes of the distribution payoff in any way they saw fit. We anticipated this would transform the nature of the experience. In experience through sampling, experienced outcomes typically result from minimal actions (e.g., the click of a button) and are best characterised a sequential observation of outcomes. By contrast, allowing people to interact with the payoff distribution by physically manipulating tangible representations of the different possible outcomes should augment people's ability to represent those payoffs and their probabilities. Whether this will strengthen or weaken framing effects remains to be established. On the one hand, a clearer representation may enhance people's computational abilities by reducing the cognitive load of switching focus between a gain-framed and a loss-framed representation, resulting in stable choices across frames. On the other hand, clarifying the representation of the distribution of outcomes in a given frame may reinforce the framing effect by making it harder to switch to the alternative frame.

A final objective for this experiment was to explore whether individual differences could predict framing effects. Previous research found some evidence that risk-averse individuals are more likely to shift their preference towards the sure option in gain frames (Huangfu, 2014; Mahoney, Buboltz, Levin, Doverspike, & Svyantek, 2011) although there was no evidence that risk propensity impacted choices in loss frames. We sought to expand on these findings by examining whether risk-taking and framing effects would be better predicted by individual differences in domain-specific risk taking. Moreover, interactive or experiential environments sometimes involve qualitatively different processes. For example, Weller, Villejoubert, and Vallée-Tourangeau (2011) found that insight performance in matchstick algebra problems was predicted by numeracy skills when the task was presented with a descriptive (static) statement, whereas it was predicted by matrix reasoning, a sub-component of the Beta III test (Kellog & Morton, 1999) for measuring visuo-spatial reasoning abilities, when the task was made interactive by through the use of physical and manipulable artefacts. In the present study, we therefore also sought to examine whether varying the level of interactivity afforded by the task would also moderate the impact of individual differences on risk-taking and framing effects.

Method

Participants

Ninety-five university students (54% Post-graduate, M_{age} = 23.8 years, SD = 4.5) took part in the experiment.

Materials, Design, and Procedure

Participants were randomly allocated to one of three experimental conditions: a Description choice condition, a Sampling choice condition and an Interactive choice condition and were asked to complete three problems presented in a gain frame (number of people who will be saved) and three problems presented in a loss frame (number of people who will die). Thus the experiment employed a mixed 2-within (Frame: Gain vs. Loss) x 2-between (Choice condition: Description, Sampling, Interactive) design. The order of presentation of the gain and loss frames was counterbalanced (half of the participants completed the three gain-framed problems first; the remaining half started by completing the loss-framed problems). The order of the three problems was randomized within each frame condition. For each problem solved, we recorded the time spent, the alternative chosen (sure vs. risky), and the strength of their preference (1= Extremely weak to 6 = Extremely strong).



Description condition Sampling condition Interactive condition

Figure 1: Birdseye screenshots of the experimental setting in each experimental condition.

Participants completed two self-report scales after they had completed the first set of three problems: The 30-item Domain Specific Risk-Taking Scale (DOSPERT, Blais & Weber, 2006) and the Barratt Impulsiveness Scale Version 11 (BIS-11, Patton, Stanford, & Barratt, 1995)[1]. The DOSPERT scale is composed of five subscales for measuring domain-specific risk-taking preferences: Ethical (e.g., "Revealing a friend's secret to someone else"), Financial (e.g., "Betting a day's income at the horse races"), Health & Safety (e.g., "Sunbathing without sunscreen"), Recreational ("Taking a skydiving class"), and Social (e.g., "Starting a new career in your mid- thirties"). Participants were asked to rate the likelihood that they would engage in the described activity or behavior on a 7-point labelled likert scale ranging from 1 (Extremely unlikely) to 7 (Extremely likely).

The task scenarios were adapted from Mahoney, Buboltz, Levin, Doverspike, and Svyantek (2011) and all presented a life-threatening disease. Each task was presented on a laminated A4 sheet of paper. In the Description choice condition, following the presentation of the possible consequences of each alternative, participants were asked to decide which program they would favor (A or B). In the sampling choice condition, participants were presented with two packs of 25 cards and asked to assume that the possible consequences of the programs were represented by the two packs of cards in front of them. They were informed that, upon turning a card over, they would see a possible consequence for the selected program in terms of the number of people saved [killed]. They were also told that they were not allowed to pick the packs of cards and may only look at one card at a time, from either deck. Participants in the interactive choice condition were also presented with two packs of 25 cards but they were invited to pick up the decks of cards and turn over and spread out all the cards in front of them. Within each condition, some participants were filmed, with their consent, from a camera attached to the ceiling and offering a birdseye view of their actions on the cards. Figure 1 illustrates the setting for the three experimental conditions. Figure 2 illustrates the cards used.

[1]The BIS-11 scale was used as a filler task and results will not be reported or discussed further in this report.



Figure 2: Schematic representation of the cards used in the Sampling and Interactive conditions.

Results

Choice latencies

The average choice latencies for the three problems in the gain and loss frames in the three experimental conditions are reported in the left panel of Figure 3. Participants were slowest to formulate a choice preference in the interactive condition, and fastest in the descriptive condition, while the sampling condition produced marginally slower latencies relative to the descriptive condition; choice latencies appeared unaffected by framing.

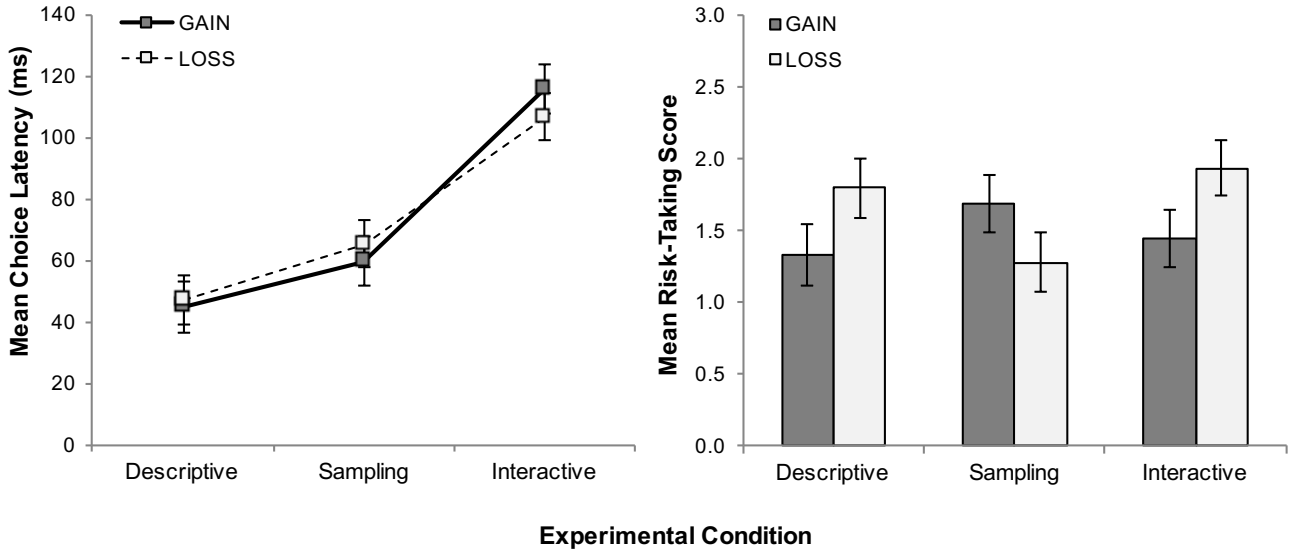


Figure 3: Mean choice latencies (left panel) and mean choice score (right panel) in the gain and loss frame in the three experimental conditions. Error bars are standard errors. .

A 2-within (Frame: Loss vs. Gain) x 3-between (Choice condition: Descriptive vs. Sampling vs. Interactive) mixed analysis of variance (ANOVA) confirmed these impressions: The main effect of condition was significant, $F(2, 92) = 26.1$, $p < .001$, $\eta^2_p = .362$, but not the main effect of frame, $F < 1$ nor the interaction, $F < 1$. Post hoc tests revealed that latencies in the interactive condition were significantly different than those in the descriptive ($p < .001$) and in the sampling ($p < .001$) condition, but that latencies in the descriptive and sampling conditions did not differ significantly ($p = .26$).

Risky choices

We computed the number of risky choices across gain-framed and loss-framed problems, respectively (from 0 to 3, in each frame). The mean composite score was higher for problems in the loss frame than in the gain frame in the descriptive and in the interactive condition (see the right panel of Fig. 3); however, in the sampling condition, the pattern was reversed such that risky choices were more frequent for problems in the gain frame than in the loss frame. The 2-within (Frame: Loss vs. Gain) x 3-between (Choice condition: Description vs. Sampling vs. Interactive) mixed ANOVA revealed that neither the main effect of condition, $F < 1$, or frame, $F(1, 92) = 2.45$, $p = .121$, $\eta^2_p = .026$, were significant, but that the interaction between frame and condition was significant, $F(2, 92) = 6.48$, $p = .002$, , $\eta^2_p = .123$. To unpack this interaction effect, we compared risk-taking scores in loss and gain frames within each experimental condition. People were significantly more likely to choose the risky option in the loss frame in the Descriptive condition, $t(29) = 2.25$, two-tailed $p = .032$, as well as in the Interactive condition, $t(32) = 2.42$, two-tailed $p = .021$. However, they were significantly more likely to choose the risky option in the *gain* frame in the Sampling condition, $t(31) = -2.08$, two-tailed $p = .046$.

Individual Differences, Risk-Taking and Framing Effects

To examine whether framing effects depended on individual differences in domain-specific risk-taking and impulsivity, we computed a framing effect index by taking the difference between the risk-taking score in loss frames and the risk taking score in gain frames for each participant. A positive score indicated a classic framing effect (higher risk-taking in losses) whereas a negative score indicated a reversed classic framing effect (higher risk-taking in gains). We then examined the correlation matrix involving scores on the Ethical, Financial, Health & Safety, Recreational, and Social DOSPERT subscales and the framing effect index, within each condition. We found no evidence that risk-taking in framing tasks was related to individual differences in risk-taking across domains. The largest correlation was with financial and gambling risk-taking in the descriptive condition, $r(28) = -.30$, $p = .11$, with health and safety risk-taking in the sampling condition, $r(30) = -.30$, $p = .10$, and with ethical risk-taking in the interactive condition, $r(31) = -.27$, $p = .13$. By and large, risk-taking in framing tasks was unrelated to individual differences in risk-taking across domains in any of the three choice conditions. The only exception was the sampling condition where risk-taking in framing tasks was negatively correlated with risk-taking in the recreational domain in both loss frames, $r(30) = -.57$, $p = .001$ and gain frames, $r(30) = -.51$, $p = .003$. In other words, the *more* likely people were to engage in recreational risk-taking, the *less* likely they were to opt for the risky alternative in the lotteries.

Discussion

This research was informed by the distributed cognition framework, which seeks to understand cognition as it unfolds through time and space, taking into account the action possibilities or affordances offered by the ecology within which cognitive processes are taking place (see also Villejoubert & Vallée-Tourangeau, 2011; Vallée-Tourangeau et al., 2015). We found little evidence to support the hypothesis that the relation between people's likelihood to engage in risky behavior and their risk-taking preferences in framing tasks could depend on the ecology within which they made their decision (i.e., whether or not they could search for information and the extent to which they could interact with it). In the interactive condition, people could spread all sample outcomes in front of them rather than sample them one at a time. This did not reduce framing effects, possibly because the artefacts used to support cognition did not afford a representational switch between a gain-framed and a loss-framed representation of the problem. This suggests that there are boundary conditions on the efficacy of an interactive ecology to support people's decision-making in a way that successfully reduces framing effects. We conceived interactivity as a feature of the ecology allowing for "sense-saturated coordination that contributes to human action" (Steffensen, 2013, p. 196). There were many possible actions in the interactive condition (picking up cards, sampling through packs, spreading cards on the table, sorting outcomes, and so on) and each action changed the manner in which the distributions of outcomes could be perceived. Stromer- Galley (2004) distinguished between interactivity as a feature of a medium and interactivity as feature of information processing in human communication. In this study, interactivity resulted from the availability of playing cards coupled with the absence of restrictions on what people could do with them; it was a product of the experimental design. Future research may examine how other forms of interactivity could impact framing effects, by using different artefacts to represent the distribution of outcomes, and by increasing interactive processing through groups (e.g., Milch, Weber, Appelt, Handgraaf, & Krantz, 2009) or game theoretic settings (e.g., Chapman et al., 2012).

Building upon the description-experience gap, we examined whether the classical framing effect observed with the Asian Disease problem could be reversed when people make decisions from experience. We also contrasted two types of experience: an experience by sampling, where participants sample single outcomes from two alternatives before making a choice, and an experience by interacting, where participants are allowed to spread all possible outcomes in a sample in front of them before making a choice.

To begin with, we replicated findings from Mahoney et al. (2001) showing that people succumb to framing effects even when they are presented with both a loss framed task and a gain framed task in a within-subject design. More importantly, we extended previous research documenting a "decision-experience gap" in monetary gambles (e.g., Hertwig et al., 2004), as we demonstrated that the framing effect could also be reversed when people sample outcomes before making a choice between a risky alternative and a sure outcome instead of reading a thumbnail description of the choice situation. Not all experiences were created equal, however, as when people were allowed to manipulate the sample of outcomes and spread them in front of them, the pattern of choices exhibited the standard framing effect, akin to what is usually observed with thumbnail descriptions. An inspection of response latencies showed that decision times increased in decision by sampling and were longest in decisions by manipulating in the interactive condition. The fact that decision times were shortest in the description condition but longest in the interactive condition suggests that framing effects do not necessarily arise from fast, non-deliberative judgements.

The reason for the framing effect reversal under sampling conditions remains to be accounted for. In the framing literature, framing reversals were created by manipulating the description of the outcomes. For example, Kühberger (1995) found that using negative descriptions (e.g., "400 people will not be saved" in the gain condition and "200 people will not die" in the loss condition) led to a reversal of the framing effects. This type of reversal is accounted for by the Explicated Valence Account (Tombu & Mandel, 2015). EVA stipulates that negative descriptions carry a negative explicated valence because they describe negative events (e.g., people who are *not* saved). Moreover, EVA proposes that people will prefer the option which maximizes the positive explicated valence, so when asked to choose between the sure outcome with a negative explicated valence ("400 people will not be saved") and a risky outcome including an event with a positive explicated valence (a chance that "600 will be saved"), people opt for the risky outcome even though the situation is framed in terms of gains (i.e., lives saved). The EVA, however, cannot be applied to explain the reversal observed in the sampling condition of the present study, since the explicated valence was always positive in the gain frame: outcomes were always described in terms of the number of lives saved.

Camilleri and Newell (2013) proposed that patterns of choices under sampling conditions result from the need to search information within the environment, and the need to integrate new information sequentially into a representation. The description-experience gap could thus result from a defective mental processing of probabilities informing the final choice in decisions from experience. There is a debate in recent literature on the description-experience gap as to whether these representations are distorted due to the limited sampling of rare events, overweighting of late observations, or both (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009). Our results provide some indirect evidence that limited sampling of rare events may hold for risk-framing tasks: in the interactive condition, participants did not need to search for information, they could display it directly in front of them. Consequently, they did not need to update their representation, they simply needed to build it from the layout

they had created. In other words, the interactive condition eliminated the sequential and potentially biased nature of information acquisition and representation that occurs in the sampling paradigm. This has been found to reduce the description – experience gap in research using monetory gambles (Camilleri & Newell, 2013) and would explain why risk-taking patterns in the interactive condition were similar to those observed in the descriptive condition.

Finally, there was some indication that recreational risk-taking could be negatively related to risk-taking in framing tasks, albeit only in the sampling condition. This suggests, somewhat counterintuitively, that risk-taking in the sampling condition was a serious and considered decision as people who took risks were those who were more likely to avoid taking risks "for fun" when making choices outside the laboratory. Future research could seek to unpack this finding to better understand the link between motivational styles (e.g., Apter, 1997) and behaviours in the sampling condition.

Acknowledgments

We thank Angelika Makri for her assistance in collecting part of the data reported in this experiment and Egle Butt for helpful comments on an earlier version of this paper. Portions of this research were included in M. Ramasubra-manian's Master's thesis. Financial support from the Kingston University Faculty of Arts and Social Sciences Research Capability Fund is also gratefully acknowledged.

References

Apter, M. J. (1997). Reversal Theory: What is it? *The Psychologist, 10,* 217-220.

Blais, A. R., & Weber, E. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*(1), 33–47.

Camilleri, A. R., & Newell, B. R. (2013). Mind the gap? Description, experience and the continuum of uncertainty in risky choice. *Progress in Brain Research, 202*, 55-71. DOI: 10.1016/B978-0-444-62604-2.00004-6.

Chapman, G. B., Li, M., Vietri, J., Ibuka, Y., Thomas, D., Yoon, H., & Galvani, A. P. (2012). Using Game Theory to Examine Incentives in Influenza Vaccination Behavior. *Psychological Science*, *23*, 1008–1015. DOI: 10.1177/0956797612437606

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-539. DOI: 10.1111/j.0956-7976.2004.00715.x.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523. DOI: 10.1016/j.tics.2009.09.004

Huangfu, G. (2014). The influence of social cues on framing effect. *Social Behavoior and Personality, 42,* 371-378. DOI: 10.2224/sbp.2014.42.3.371

Kellogg, C. E., & Morton, N. W. (1999). *Beta III manual*. New York: The Psychological Corporation, A Harcourt Assessment Company.

Kühberger, A. (1995). The Framing of Decisions: A New Look at Old Problems. *Organizational Behavior and Human Decision Processes*, *62*, 230–240. http://doi.org/10.1006/obhd.1995.1046

Kühberger, A. (1998). The influence of framing on risky decisions: A meta analysis. *Organizational Behaviour and Human Decision Processes, 78*, 23-55. DOI: 10.1006/obhd.1998.2781

Lichtenstein, S., & Slovic, P. (2006). *The Construction of Preferences*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511618031

Mahoney, K. T., Buboltz, W., Levin, I. P., Doverspike, D., & Svyantek, D. J. (2011). Individual differences in a within-subjects risky-choice framing study. *Personality and Individual Differences, 51*, 248-257. DOI: 10.1016/j.paid.2010.03.035

Maule, J., &Villejoubert, G. (2007). What lies beneath: Reframing framing effects. *Thinking and Reasoning, 13,* 25-44. DOI: 10.1080/13546780600872585

Milch, K. F., Weber, E. U., Appelt, K. C., Handgraaf, M. J. J., & Krantz, D. H. (2009). From individual preference construction to group decisions: Framing effects and group processes. *Organizational Behavior and Human Decision Processes*, *108*(2), 242–255. DOI: 10.1016/j.obhdp.2008.11.003

Patton, J. H., Stanford, M. S., & others. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768–774.

Stromer-Galley, J. (2004). Interactivity-as-Product and Interactivity-as-Process. *The Information Society*, *20*, 391–394. DOI: 10.1080/01972240490508081

Tombu, M., & Mandel, D. R. (2015). When Does Framing Influence Preferences, Risk Perceptions, and Risk Attitudes? The Explicated Valence Account: Framing and the Explicated Valence Account. *Journal of Behavioral Decision Making*. Advance online publication. DOI: 10.1002/bdm.1863

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458. DOI: 10.1126/science.7455683

Villejoubert, G., & Vallée-Tourangeau, F. (2011). Constructing preferences in the physical world: a distritubed cognition perspective on preferences and risk choices. *Frontiers in Psychology, 2*, 1-2. DOI: 10.3389/fpsyg.2011.00302

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General, 144,* 581-603. DOI: 10.1037/a0039161

Weller, A., Villejoubert, G., & Vallée-Tourangeau, F. (2011). Interactive insight problem solving. *Thinking and Reasoning, 17,* 424-439. DOI:10.1080/13546783.2011.629081.

First things first? Top-down influences on event apprehension

Johannes Gerwien (jo.gerwien@uni-heidelberg.de)
Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University
Plöck 55, 69117 Heidelberg, Germany

Monique Flecken (monique.flecken@mpi.nl)
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, the Netherlands

Abstract

Not much is known about event apprehension, the earliest stage of information processing in elicited language production studies, using pictorial stimuli. A reason for our lack of knowledge on this process is that apprehension happens very rapidly (<350 ms after stimulus onset, Griffin & Bock 2000), making it difficult to measure the process directly. To broaden our understanding of apprehension, we analyzed landing positions and onset latencies of *first fixations* on visual stimuli (pictures of real-world events) given short stimulus presentation times, presupposing that the first fixation directly results from information processing during apprehension.

Keywords: apprehension, visual attention, event construal, language production, cross-linguistic analyses

Introduction

What is event apprehension? In most sentence production research, it is assumed that during this phase the 'gist' of a scene is extracted from a visual stimulus (Henderson & Ferreira 2004). It is, however, not clear what features constitute the 'gist' of a scene, nor is there a clear differentiation between event apprehension and the next phase of the production process, namely, the generation of a pre-verbal representation ('message') in which the information that a speaker intends to verbalize, and in what way, is specified. Understanding the interrelation between these two processes though, is an important step towards better insights into the early phases of the language production process, and potential factors affecting it.

To date, studies on event apprehension show that it is a rapid process, during which scene category information and associated semantic knowledge becomes available (Henderson & Ferreira 2004), scene coherence can be evaluated (Dobel et al. 2007), and an event representation can be constructed in which visual entities are linked to event roles (agent/patient) (Griffin & Bock 2000). However, it has also been argued that only those features of objects or scenes needed for the task at hand can be perceived (Williams & Simmons 2000, Dobel et al. 2007), showing that top-down factors, i.e., task-demands, play a role already during early visual processing, and that apprehension is thus a *flexible* process - a process which can be adapted to specific circumstances. Furthermore, studies employing change blindness paradigms suggest that another top-down factor, i.e., the cultural and language background of the viewer, may also influence basic perceptual processes. Masuda & Nisbett (2006), for example, showed that East Asians are more sensitive to information in the periphery (changes in context) compared to Americans who are more sensitive to changes in focal objects. Here, we aim to explore the dimensions of the flexibility of the scene apprehension process, using eye tracking methodology.

Background

How has apprehension been studied previously? Using eye tracking, apprehension has been linked to the first, or the first few overt fixations in a trial (Griffin & Bock 2000, Bock et al. 2003). However, those early fixations have been reported to be unspecific, in the sense that no clear pattern of fixation locations was detected. First fixations tend to be placed in the middle of the stimulus, a location from where relevant information at different locations can be visually processed more or less equally well (Holmqvist et al. 2011). Moreover, Griffin & Bock (2000) interpret an increase in fixations toward one specific region of the stimulus, for example the region displaying the agent of an action, as indicating the completion of apprehension and the beginning of further linguistic processing, i.e., syntactic and phonological encoding.

Another approach to studying apprehension is to investigate what information can be extracted from a visual stimulus in what amount of time (Dobel et al. 2007). Even without overt fixations speakers are surprisingly good at identifying agents, objects, and actions, under very short presentation duration conditions (100-250ms). However, in this approach it is difficult to precisely understand what is going on during apprehension: When participants are asked to describe what they have seen, a representation of what the informant intends to say (the message) is obligatorily constructed. It is thus more than difficult to distinguish between apprehension and message generation, based on using the linguistic product (the descriptions produced) as the only measure.

In order to investigate questions about apprehension independent of message generation or other aspects of the production process, an empirical measure is needed which mainly reflects apprehension and which does not directly relate to linguistic processing. Here, we record and analyze *first fixation locations and latencies* given a manipulation of task requirements, i.e., by constraining the amount of time participants have for visual information uptake (presentation duration manipulation: 300ms, 500ms, 700ms) and by using

two different participant groups (speakers of German and Spanish). We argue that, under brief exposure times, first fixations are informative about apprehension, and not reflecting (other, later) stages in the production process.

As mentioned above, a manipulation of stimulus presentation duration has been applied previously; here, we are specifically interested in how the quality of event descriptions may change given different exposure times. By encouraging participants to produce full sentences, we can shed light on how *specific* the information extracted concerning different event elements under these circumstances can be. Furthermore, by analyzing linguistic content as well as first fixation locations/latencies we are able to compare the type of information overtly attended first to the type of information encoded in the linguistic product.

The rationale behind choosing two different languages is that, though message generation processes should be the same in speakers of any language (and they should be dependent on the same set of task- or stimulus-related factors), there is evidence of cross-linguistic variation in viewing behavior during scene description (Brown-Schmidt & Konopka 2008; Sauppe et al. 2013; Flecken et al. 2015; v. Stutterheim et al. 2016): Cross-linguistic differences concerned what type of information is considered most relevant, when, and effects were observed during message planning and/or formulation. Here, we ask whether language already exerts influence on apprehension in a description task, and thus whether the process may show flexibility in adapting to language-specific demands. Our test case is Spanish vs. German: The two languages have been shown to differ in the prominence accorded to two core event elements, i.e., the agent and the action; Spanish speakers tend to represent accidental events *action*- rather than agent-oriented (Fausey & Boroditsky 2011), whereas German speakers, conceptualize events with a strong focus on *agents* (Flecken et al. 2015). Differences are driven by specific linguistic means typically used to describe events in the two languages. We will use this variation in global agent- vs. action-prominence in event encoding to shed light on the flexibility of the apprehension process.

Aims of the present study

We address the contents and features of initial visual processing, i.e., scene apprehension, in sentence production. Our window on this process is the location of the very first fixation on stimuli showing photographed events (agent performing action on an object). Given that during apprehension the location for this first fixation must be calculated, analyzing in which region of the depicted scene the first fixation is registered should allow us to infer, at least to some extent, what information was derived from the stimulus up to this moment and what information the cognitive system demands next in order to fulfill the task, i.e., constructing a representation that can be verbalized.

A manipulation of stimulus exposure time (300ms-500ms-700ms) allows us to address the *flexibility* of the process under different degrees of 'pressure' on the system, given the complex task to produce full-fledged event descriptions. Furthermore, we investigate the flexibility of the process by contrasting speakers with native languages which differ in agent- vs action-prominence in event encoding.

If the presentation duration manipulation has no impact on the locations and/or latencies of first fixations, the apprehension process must be considered rigid, in the sense that it cannot adapt to time constraints on visual information uptake. If language shows no effect it must be concluded that apprehension is unaffected by linguistic or cultural variation.

Experiment 1 - Pretest

In Experiment 1 native speakers of German and Spanish described and rated a collection of photographs depicting everyday events, each involving an agent performing an action on an object. This experiment was done to ensure homogeneity in event recognition and labeling (e.g., choice of action verbs), and to control for a potential visual bias towards specific elements in the scenes.

Method

Participants Native speakers of German and Spanish (N = 10 in each group; university students 20-30 years of age) took part in the experiment. Data were collected at Heidelberg University, which has a large population of Spanish speaking exchange students (Erasmus programme).

Materials 73 different actions, each performed by a male and female agent, involving one specific object, were photographed. All scenes were staged identically against a white background, controlling for distance between agent and object, head, body and object position and the amount of space covered by agent and action in the photograph. There was also a mirrored version of each item.

Procedure Photographs were printed in black and white and presented to participants in a paper catalogue, in a randomized order. Each participant saw each action once, half performed by a male, half by a female actor. Participants were first asked to write down the type of event depicted (e.g. "to squeeze a lemon"). Then, they were asked to rate each photograph for a number of dimensions: - naturalness of the event, on a scale from 1 to 5 (1: unnatural; 5: natural); - ease of recognition of the action (1: action not recognizable; 5: action perfectly recognizable); - a potential visual bias of one over the other element (is either the agent or the action more easily recognizable, identifiable or prominent? or are both elements equal? (open question)). Participants were given as much time as needed, but they were instructed not to look back at previous items and to rate the pictures on the basis of their first impression.

Results

In total, 13 items had to be discarded due to a high level of heterogeneity in event description, most of them given by

cross-linguistic differences. For the selection of stimuli for Experiment 2, we made sure that the events were described in Spanish and German with a similar verbal structure and level of syntactic complexity, and that they could be described with a compact action verb. For example, the action 'ein Ei koepfen' ('to top off the lid of an egg') cannot be described by a single verb in Spanish (neither in English, actually). On the basis of the descriptions and the ratings, we selected two sets of 60 homogeneous scenes with a male and a female actor performing the same actions for Experiment 2. Selected pictures were rated 4-5 for the dimensions reported above; no visual biases were reported.

Experiment 2 - Online event description

Method

Participants In total, 40 participants were tested. Eight participants from each language were assigned to the 300ms presentation duration condition, eight from each language to the 500ms presentation duration condition and four from each language to the 700 ms condition.

Materials Figure 1 shows an example stimulus. Picture size was 600x400px.



Figure 1: Example stimulus with three Areas of Interests (AoIs); the AoIs were not visible to subjects

Procedure Eye movements were recorded with a remote Eye link 1000 system, running on Experiment Builder (SR research), with a 500 Hz sampling rate. The monocular recording mode was used. Participants were seated in front of a 19 inch computer screen, at a distance of appr. 65 cm.

Participants were given written instructions in their native language. A native speaker experimenter (Spanish/German) was present to answer questions. The instructions explicitly aimed at eliciting complete sentences, referring to dynamic events, rather than descriptions of individual elements / static properties of the scenes. A trial started with a centred fixation cross (1000ms), after which a picture appeared randomly in one of the four quadrants of the screen. This was done (1) to force participants to execute a fixation, and (2) to prevent strategies related to the predictability of a stimulus' location. The number of pictures showing agent left-action right, and agent right-action left was counterbalanced across subjects and order of presentation was randomized within subjects. Stimulus presentation duration was manipulated between subjects. After picture offset, a blank screen was shown for 9000 ms, providing sufficient time for producing utterances. Speech was audiorecorded with an external microphone.

Data treatment and Results

We analyzed first fixations[1] on the pictures for three locations (AoIs) separately: Agent (head and upper part of agent's body, white oval), Action (hands and object, white oval) and 'In-between' (part of agent's body and closely surrounding it, dark-grey area, Figure 1). We included the in-between area because there are no theoretical grounds to assume that first fixations that are placed neither in the agent nor in the action AoI are arbitrary or inaccurate. In fact, we consider the location of the first fixation as the outcome of specific calculations made by the cognitive system while apprehending.

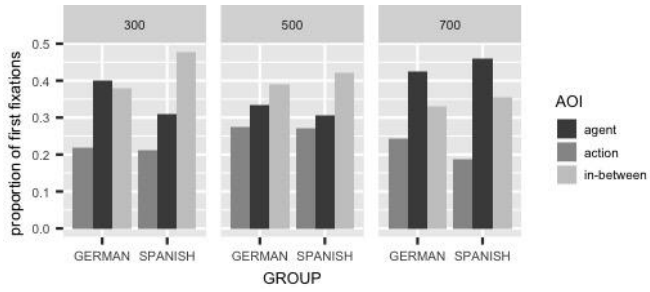


Figure 2: Overall proportions of first fixations in the AoIs, by presentation duration and language

Figure 2 shows the overall proportions of first fixations in each AoI within each group, per condition (300ms, 500ms, 700ms). First fixation locations (binary coded for each AOI separately; 1=hit, 0=no hit) were analysed with linear mixed effects regression models, including the fixed factors Language (sum coded) and Condition (treatment coded) and random intercepts for subjects and items (pictures; both versions were collapsed).

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)[2]	-0.633	0.201	-3.156	0.002	
2	Ger vs Spa	-0.212	0.164	-1.288	0.198	
3	300 vs 500	-0.171	0.185	-0.924	0.356	
4	300 vs 700	0.449	0.223	2.014	0.044	*
6	(Intercept)[3]	-0.803	0.201	-3.997	0.000	
7	500 vs 700	0.619	0.223	2.780	0.005	**
8	Ger vs SPA 300	-0.490	0.2539	-1.925	0.054	.
9	Ger vs SPA 500	-0.141	0.2531	-0.558	0.577	
10	Ger vs SPA 700	0.180	0.3491	0.516	0.606	

Table 1: 'Agent' first fixations; reference level is indicated by the first term in each line

With respect to the agent AoI, our analyses revealed an effect of condition (presentation duration): There were more first looks in the agent AoI in the 700ms condition compared to both other conditions, but there was no difference between 300ms and 500ms (Table 1, lines 3, 4, and 7). There was no effect of Language. Comparing the 3

[1] A first fixation was defined as the event that followed the first saccade after stimulus onset, as registered by the eye tracker. On each trial, the initial position of the participants' gaze was in the center of the screen due to a centred fixation cross being presented before each stimulus.

[2] Model includes both main predictors.

[3] Same model specifications as in lines 1-4 but with reference level changed to 500ms.

presentation duration conditions between languages, the model detected a small effect: In the 300ms condition, Spanish participants displayed fewer first looks to the agent than German participants (Table 1, line 8).

With respect to the action AoI, our analyses showed no effects (models are not reported).

Regarding first fixations in the 'in-between' area, there were significantly fewer of those in the 700ms compared to the 300ms condition (Table 2, line 5). A main effect of Language was detected: Spanish speakers placed significantly more first fixations in the in-between area than German speakers (Table 2, line 2). Comparing the 3 conditions between languages, the model showed a significant effect in the 300ms condition, with Spanish participants fixating the 'in-between' AoI more than German participants (Table 2, line 8).

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)	-0.317	0.130	-2.427	0.015	
2	Ger vs Spa	-0.132	0.067	-1.987	0.047	*
4	300 vs 500	-0.112	0.149	-0.753	0.452	
5	300 vs 700	-0.396	0.183	-2.166	0.030	*
6	(Intercept)	-0.428	0.130	-3.288	0.001	
7	500 vs 700	-0.284	0.183	-1.555	0.120	
8	Ger vs SPA 300	-0.226	0.104	-2.179	0.029	*
9	Ger vs SPA 500	-0.079	0.103	-0.762	0.446	
10	Ger vs SPA 700	-0.052	0.146	-0.355	0.722	

Table 2: 'In-between' first fixations; reference level is indicated by the first term in each line

First Fixation latencies

To test whether participants modulate the time for executing their first fixation depending on stimulus presentation duration, log-transformed first fixation latencies were analyzed using mixed effects linear regression models with Language and Condition as fixed factors and random intercepts for subjects and items. Results show that first fixation latencies are significantly smaller in the 300ms condition compared to the 500ms condition (Est.=0.050, SE=0.025, t=1.96, p<.05) and marginally smaller compared to the 700ms condition (Est.=0.060, SE=0.031, t=*1*.95, p=.05). There was no significant effect of Language, nor any interaction effects.

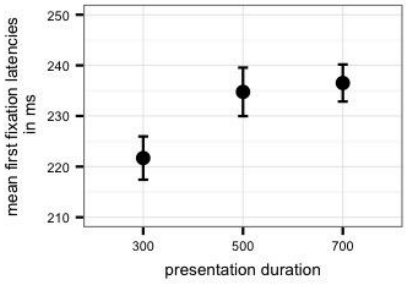


Figure 3: Mean first fixation latencies

Language data: specificity of event descriptions

The transcribed utterances were coded for specificity with respect to agent, action, and object features. Agent reference was coded as 'specific' when the subject contained a noun (phrase) with information on the agent's gender (*[there is / I see] a/the woman*), and as 'unspecific' when the sentence contained a gender-neutral noun (*a person/someone*), a pronoun or null subject. Action-specific sentences contained a specific action verb (*to draw*); the action was coded as unspecific in case of general action verbs (*to do/hold something*) or state-verbs (*to sit at a table*). Objects were coded as specific when referring to concrete items (*a sandwich/a bottle*); other object references were coded as unspecific (*something [big]*) (all specificity factors were binary coded). Mixed effect models were used to estimate the impact of the factors Language and Condition.

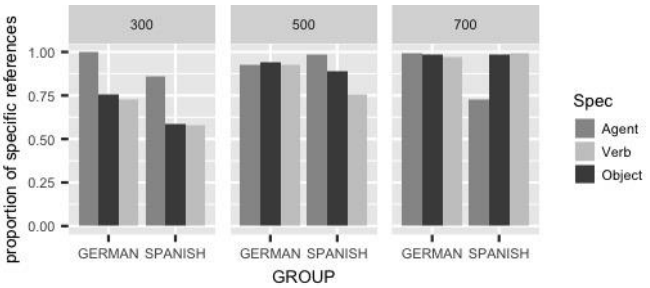


Figure 4: Overall proportions of specific encodings of agent, verb, object, by presentation duration and language

Agent specificity

There were no language or condition effects (nor interactions) for agent specificity (model is not reported). Note that in some cases Spanish participants omitted overt reference to the agent, which is licensed in specific contexts (pro-drop). It occurred mainly in the 700ms condition and these instances are coded as unspecific agent references.

		Estimate	Std. Error	z value	Pr(>|z|)	
1	(Intercept)	0.855	0.378	2.263	0.024	
2	Ger vs Spa	0.418	0.241	1.732	0.083	.
3	300 vs 500	1.562	0.522	2.992	0.003	**
4	300 vs 700	3.963	0.745	5.320	0.000	***
5	(Intercept)	2.417	0.393	6.147	0.000	
6	500 vs 700	2.403	0.746	3.219	0.001	**
7	Ger vs SPA 300	0.385	0.350	1.099	0.272	
8	Ger vs SPA 500	0.790	0.364	2.167	0.030	*
9	Ger vs SPA 700	-0.627	0.662	-0.948	0.343	

Table 3: Verb specificity; reference level is indicated by the first term in each line

Verb and object specificity

Analyses showed an effect of Condition: The longer the stimulus was displayed on the screen, the more specific action verbs were produced (Table 3, lines 4 and 5). There was also a significant Condition by Language interaction: In the 500ms condition, Spanish speakers produced fewer specific action verbs than German speakers (Table 3, line 8).

With respect to object encoding there were significant effects of Condition (Table 4, lines 3, 4, and 6), with object specificity increasing with longer exposure time. There was also an effect of language: Overall, Spanish speakers produced fewer specific object references. Most pronounced is the effect when comparing Spanish and German speakers within the 300ms condition (Table 4, line 7).

		Estimate	Std. Error	z value	Pr(>|z|)
1	(Intercept)	0.830	0.257	3.232	0.000
2	Ger vs Spa	0.406	0.164	2.475	0.013*
3	300 vs 500	2.063	0.354	5.834	0.000***
4	300 vs 700	3.862	0.554	6.908	0.000***
5	(Intercept)	2.893	0.288	10.056	0.000
6	500 vs 700	1.763	0.557	3.167	0.002**
7	Ger vs SPA 300	0.515	0.232	2.215	0.027*
8	Ger vs SPA 500	0.391	0.258	1.513	0.130
9	Ger vs SPA 700	-0.027	0.488	-0.056	0.955

Table 4: Object specificity; reference level is indicated by the first term in each line

General Discussion

The main findings of Experiment 2 are the following: (1) Presentation duration modulates the specificity of the event descriptions: Reference to actions and objects is more precise given longer viewing times; (2) presentation duration affects the timing of parafoveal and foveal information uptake, i.e., given a short viewing time participants move the eyes faster towards a location from where acute vision is possible; (3) presentation duration has an effect on the location of first looks: There were generally fewer first looks in the in-between region and more looks to the agent region in the longest presentation duration condition; (4) the language of the participants exerted an influence on first fixation locations.

The finding that verbal responses were more specific regarding actions and objects given longer presentation durations, but at the same time there was no increase with respect to first fixations registered in the corresponding Area of Interest (Action region), suggests that the location of the first fixation does not directly relate to the process of message construction, i.e., linguistic processing. Furthermore, the location of first fixations does not correlate with the sentence format produced (only approximately 40% of all first looks were registered in the agent region, but sentences were exclusively SV(O)). We thus take both findings to show that our measure indeed reflects processing *prior* to linguistic planning, i.e., scene apprehension.

To explain the effect of presentation duration on the location of the first fixation, three things must be taken into account. First, information extraction is better the nearer the eye rests on the area which contains critical information, because viewing acuity asymptotically decreases from the foveal region (Holmqvist et al. 2011). Second, the time available for information extraction *after* the first fixation is placed on the stimulus (acute vision), differs between all three conditions. In the 300ms condition participants have approximately 80ms for further visual processing after the eye has arrived at the landing site and before the stimulus disappears, whereas this time is already more than three times longer in the 500ms condition, and almost 6 times longer in the 700ms condition. Third, the time before the first fixation is launched also differs between the presentation durations, as is evident from the differences in first fixation latencies. In the 300ms condition participants remained significantly shorter at the fixation cross before directing focal attention onto the stimulus, compared to the 500ms or 700ms condition. The location of the first fixation thus reflects what type of visual information the processing system demands with the highest acuity possible, given the time available. Since first fixation patterns differ most clearly between the 300ms and the 700ms condition (generally more first looks to agents, and fewer to the in-between region in the longer presentation duration), the strongest predictor for the location of the first fixation is the probability of a following fixation: In the longest presentation duration condition participants nearly in all cases had enough time to place a total of two fixations on the stimulus before it disappeared. Calculating the location and timing the execution of a first fixation thus draws on the ability of the cognitive system to predict the approximate future time course of the processes relevant for task completion, in our study the task of constructing a message on the basis of the information available. Given these results, we can conclude that apprehension is a flexible process.

Interestingly, we did obtain a difference with respect to first fixation *latencies* but not (generally) first fixation *location* when comparing the 300ms and the 500ms condition. Given that verbal responses were more precise in the 500ms condition we can conclude that the landing site of the first fixation in the 300ms condition was already "good enough" to fulfill the task, even with this short exposure time (i.e., participants were already able to provide information on both the agent and the action in the 300ms condition). In future research it thus seems reasonable to explicitly analyze the interrelation between first fixation locations and first fixation durations.

The analysis of event descriptions revealed that the representation which is ultimately encoded by verbal means, i.e., the preverbal message, is not as "rich" in the shortest condition as it is in both longer conditions – participants produced fewer specific action verbs and object references in the 300ms condition. This implies that less specific information was passed on from the representation constructed during apprehension to the message generator. Put the other way around, a message was constructed on the basis of whatever information was available. Since messages are thought to be composed of word meanings (Levelt et al. 1999) the quality of information available from event apprehension directly influences what word meanings can be activated. This is especially evident from the main effect of presentation duration on verb specificity.

Given the role of the factor language in our experiment we may infer that during scene encoding speakers of different languages focus on different things at different times. Under time pressure, mainly Spanish speakers resort to fixating the area in between the agent and the action first, and this seems to be sufficient for retrieving information concerning both event elements, as there is no difference in encoding specificity with German speakers. German speakers, in turn, fixate agents first more frequently in the 300ms condition. In sum, language seems to matter but not in a straightforward way: Given limited exposure time, we

have identified that German and Spanish speakers have different *starting points* for their linguistic encoding processes. Nonetheless the *outcome* seems to be the same, in the present task setting. Previous research has put forward the hypothesis that in general, the Spanish language may be less agent-oriented than German or English. Our data can be interpreted in this way, but the effect mainly appeared under increased task demands: Spanish speakers showed fewer agent-first fixations than Germans in the 300ms condition. Note that this lack of a prominent agent-first strategy in Spanish is not directly linked to the fact that the language allows subject-drop; there were only few such instances in the data, limited to the longest presentation duration condition. We can conclude that the current experimental design does not provide an appropriate pro-drop licensing context in Spanish. At this point, further research is required, introducing, e.g., pro-drop biasing context (e.g., a discourse context), or more variation regarding agents and event-types (i.e., ditransitive events; controlling for animacy, another relevant factor) in the experimental stimuli. The use of different language contrasts would also shed more light on whether language background of the speaker structurally affects scene apprehension.

On a similarly critical note, given the between-subjects manipulation of stimulus presentation duration, we cannot fully exclude that participants developed some strategy for solving the task in the specific condition put to them. Importantly, however, for our participants it was impossible to predict the locations of the agent and the action on the screen, given left-right counterbalancing of agents and actions within the pictures, and the random placement of pictures on the screen. Regardless, even if explicit strategy-formation played a role of some sort, the data still show effects of task demands and language background, evidencing top-down influences on scene apprehension.

Conclusions

We measured the location of first fixations and their latencies to gain insight into the flexibility of the event apprehension process. We find top-down effects: For the first time it was shown that the time available for visual processing (presentation duration manipulation: 300ms-500ms-700ms) directly affects *when* and *where* people move their eyes first. However, first fixation locations and latencies did not directly relate to the contents (i.e., the level of specificity) of our participant's verbal responses. We thus argue that first fixations are driven by apprehension (initial gist extraction) and not message generation processes, which underlines that both processes are distinct in nature. Furthermore, we obtained small, but measurable effects of the language spoken by our participants, but mainly when time pressure on information uptake was strongest. These differences in first fixations, however, were not directly reflected in the information provided in the utterances produced. Further research is required to address language background as a factor affecting scene apprehension.

Acknowledgements

This study was made possible by a grant from the Ibero-Amerika-Zentrum (IAZ), Heidelberg University, to both authors. We thank Lina Zambrano for help with data collection and Kay Bock for discussing preliminary results with us.

References

Bock, K., Irwin, D., Davidson, D., & Levelt, W. (2003). Minding the clock. *Journal of Memory and Language, 48*(4), 653-685.

Brown-Schmidt, S., & Konopka, A. (2008). Little houses and casas pequeñas: message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition, 109(2),* 274-280.

Dobel, C., Gumnior, H., Bölte, J. & Zwitserlood, P. (2007), Describing scenes hardly seen. *Acta Psychologica, 125(2),* 129-143.

Fausey, C., &. Boroditsky. L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review, 18(1),* 150-157.

Flecken, M., Gerwien, J., Carroll, M., & v. Stutterheim, C. (2015). Analyzing gaze allocation during language planning: a cross-linguistic study on dynamic events. *Language and Cognition, 7(1),* 138-166.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11(4),* 274-279.

Henderson, J., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. Henderson & F. Ferreira (Eds), *The Interface of Language, Vision, and Action*, pp. 1-58. New York: Psychology Press.

Holmqvist, K., Nystrцm, M., Andersson, R. Dewhurst, R. Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences, 22(1),* 1-38.

Masuda, T., & R. Nisbett (2006). Culture and change blindness. *Cognitive Science, 30(2),* 381-399.

Sauppe, S., Norcliffe, E., Konopka, A., Van Valin, R., & Levinson, S. (2013). Dependencies first: eye tracking evidence from sentence production in Tagalog. In M. Knauff et al. (Eds.), *Proceedings of the 35th Annual Meeting of CogSci* (pp. 1265−1270). Austin, TX: Cognitive Science Society .

v. Stutterheim, C., Bouhous, A. & Gerwien, J. (2016). The impact of aspectual categories on the construal of motion events. The case of Tunisian and Modern Standard Arabic. Poster presented at EXAL+, NYU Abu Dhabi.

Williams, P., & Simons, D. J. (2000). Detecting changes in novel, complex three-dimensional objects. *Visual Cognition, 7,* 297–322.

Definitely maybe and possibly even probably: efficient communication of higher-order uncertainty

Michele Herbstritt (michele.herbstritt@gmail.com)
Michael Franke (mchfranke@gmail.com)
Department of Linguistics, University of Tübingen
Wilhelmstrasse 19, 72070 Tübingen, Germany

Abstract

Possibility and probability expressions, like *possibly* or *probably*, are frequently assumed to communicate that the probability of a proposition is above a certain threshold. Most previous empirical research on these expressions has focused on cases of known objective chance: if the true objective probability is given, would a speaker use *possibly*, *probably* or one of their kin? Here, we investigate the use of probability expressions when speakers have subjective uncertainty about objective chance, i.e., higher-order uncertainty. Experimental data suggest that speakers' choices of a probability expression is a complex function of their state of higher-order uncertainty. We formulate a computational probabilistic model of pragmatic speaker behavior that explains the experimental data.

Keywords: uncertainty; probability; experimental pragmatics; computational modeling

Introduction

Real-world reasoning, decision-making and communication tasks take place almost invariably under uncertainty. From everyday future planning and small talk about the weather to scientific practices and financial or legal reports, we reason, decide and speak about things that we do not know for sure. It is not surprising that possibility and probability expressions are ubiquitous in communication. To mention only a recent example, a quick search within 'The Litvinenko Inquiry', delivered by Sir Robert Owen to the Home Secretary of the United Kingdom on January 19, 2016 (329 pages),[1] resulted in 84 instances of *probable/probably/(un)likely* and 103 instances of *might/possibly/it is possible that*.

Higher-order uncertainty This paper reports experimental and modeling work on the meaning and use of the English expressions *probably* and *possibly*. These expressions (and variations thereof) have been extensively investigated in linguistics (Kratzer, 1991; Egan & Weatherson, 2011; Yalcin, 2010; Lassiter, 2011) and psychology (Beyth-Marom, 1982; Brun & Teigen, 1988; Teigen, 1988; Windschitl & Wells, 1996, 1998). What sets our work apart from most of the literature is that we investigate the use of *probably* and *possibly* in situations of what we call "higher-order uncertainty".

To illustrate, imagine an urn containing 10 balls of two different colors (e.g. red and blue). The proportion of red balls out of 10 expresses the objective chance that a random draw will result in a red ball. Knowing the objective chance means knowing the exact content of the urn, and yet it is not enough to be sure about what will happen, unless the objective chance is equal to 0 or 1. This represents the first layer of uncertainty, where having perfect information about how the world is like is not enough to perfectly predict how the world will be.

A second, higher-order, level of uncertainty comes into the picture when the true objective chance is uncertain. Imagine an agent who has only imperfect information about the urn: the agent knows that it contains 10 balls of two different colors, but the agent is only allowed to draw (and look at) a certain number of balls. The agent might observe that none, some or all of the drawn balls are red and form an uncertain belief about the content of the urn. Clearly, the lower the number of drawn balls, the less precise the agent's belief.

This work is about uncertainty of this higher-order kind. In the remainder of this section we summarize some relevant ideas from the linguistic literature on possibility and probability expressions. The following section reports on an experiment that investigates choices of probability expressions under higher-order uncertainty. Standard regression modeling suggests that higher-order uncertainty may affect lexical choices in complex ways. We therefore turn to a computational model of a pragmatic speaker that aims to predict lexical choice in a natural way. We compare model variants and scrutinize the predictions of the most credible variant.

Previous work The amount of previous linguistic research about possibility expressions such as *might*, *possible*, *possibly* (best known as "epistemic modals") is gigantic.[2] Here we refer to the milestone work by Kratzer (1991), which puts forward a uniform analysis of possibility and probability expressions as quantifiers over possible worlds. It is a purely qualitative analysis, with no reference to probability measures.

Much less research has appeared in linguistics specifically about *probably*. However, Kratzer's picture has been challenged on many grounds by several authors in recent years. Oversimplifying, there seems to be a consensus about the need of a semantics which incorporates probability measures (Yalcin, 2010; Lassiter, 2011; Moss, 2015). The simplest implementation of this semantics postulates that the meaning of a sentence of the form *Probably, p* is that the probability of *p* exceeds a certain threshold θ.

Recent work considers complex nested cases as well (Moss, 2015; Lassiter & Goodman, 2015):

(1) It might be probable that Alice is wearing green.

[1] Available online at `https://www.litvinenkoinquiry.org`.

[2] See (Egan & Weatherson, 2011) for an introduction to the contemporary debate.

(2) Alice is definitely likely to be wearing green.

Suppose that Bob is Alice's friend and has witnessed her wearing green 5 times on 8 consecutive days. On the other hand, Carol is Alice's mum and has observed that Alice was wearing green 500 times on 800 consecutive days. Despite the fact that the objective proportion of green observations is the same for both, only Carol is in the position to utter (2), while Bob should limit himself to (1). These examples raise the question —most often left unanswered— of what kind of uncertainty is in play when speakers use possibility and probability expressions. If the objective proportion is not enough to distinguish between (1) and (2), then what is needed?

This is relevant also for simple, non-nested, uses of possibility and probability expressions, which are way more frequent than nested cases. Consider a more extreme version of the urn scenario. There are 100 balls in the urn, of two different colors. Imagine drawing 8 balls and observe that 5 of them are red. Then (3) is a very appropriate thing to say and (4) is not. But imagine drawing 80 balls and observe that 50 of them are red: the proportion of observed red balls is the same, but, intuitively, (4) is more appropriate than before.

(3) A randomly drawn ball might be red.

(4) A randomly drawn ball will probably be red.

It seems that in the case of uncertainty about objective chance, the maximum likelihood guess about the latter is not all that matters. If 50 out of 80 observed balls are red, we can be much more certain about the objective chance level, than when 5 out of 8 balls are observed. This difference may give rise to complex nested uses of probability expressions, but does it have to? It seems that higher-order uncertainty is frequent in life. If there is an objective chance of rain, we do not know it, but still we do not hear the neighbor say that it will definitely maybe and possibly even probably rain.

There are two options to reconcile higher-order uncertainty with the use of simple probability expressions. One is to assume that an agent who is uncertain about the objective chance of a proposition p, flattens his higher-order belief into a single probabilistic belief about p. Simple probability expressions do not communicate genuine higher-order uncertainty, but simple one-dimensional uncertainty about p. Another possibility is that higher-order uncertainty plays a role in our choice of probability expressions. The question, to be addressed by a computational model below, would then be: how exactly would different layers of uncertainty affect the choice of probability expressions? In any case, since the mapping from higher-order certainty to flattened one-dimensional beliefs is many-to-one, it should be possible to find experimental evidence to decide between these rival options.

Experiment

Participants 50 (self reported) English native speakers with US IP-addresses were recruited via Amazon's Mechanical Turk.

Material The setup of the experiment is the urn scenario introduced above. The urn contains 10 balls of two different colors. Observations of the content of the urn are made by drawing a certain number of balls (referred to as "access") and counting how many of them are red ("observation"). Each trial showed a picture representing the urn scenario with various access and observation. A sample stimulus is shown in Figure 1. A short description that came with every picture reminded participants that they had to imagine drawing a certain number of balls, looking at them, and putting them back in the urn; then, they would draw another ball and make a prediction about its color.

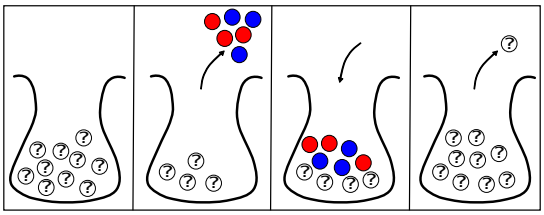


Figure 1: Sample stimulus.

We selected seven proportions of observed red balls, namely $0, 0.25, 0.33, 0.5, 0.67, 0.75, 1$. Each proportion is obtained in two different ways: one encoding what we call "low" uncertainty (access>5) and one what we call "high" uncertainty (access<5). Crossing proportions and uncertainty levels resulted in 14 experimental conditions (Table 1).

Table 1: Experimental conditions.

	0	0.25	0.33	0.5	0.67	0.75	1
high	$0/2$	$1/4$	$1/3$	$2/4$	$2/3$	$3/4$	$2/2$
low	$0/10$	$2/8$	$3/9$	$4/8$	$6/9$	$6/8$	$10/10$

The experiment consisted of *expression* and *likelihood trials*. Expression trials probe into the lexical choice of probability expressions. We asked the participants to complete a sentence of the form *The next ball will* [...] *be red* by selecting the most appropriate expression from a drop-down menu containing *certainly not*, *probably not*, *possibly*, *probably*, *certainly*. In likelihod trials participants had to answer the question *How likely do you think it is that a randomly drawn ball will be red?* by adjusting a slider ranging from 0 to 100 with a step of 5. Answers from likelihood trials hint at participants' beliefs about the probability of the crucial proposition *The next draw will be red* in each condition.

Procedure Each participant completed 12 trials, one for each of 12 conditions randomly picked from the 14 total conditions. Half of the trials were expression trials, the other half likelihood trials. Prior to the main experimental phase, participants completed a training phase that contained the following introduction:

"This experiment is an interactive two player game of chance. The players cooperate to guess the content of an urn. Both players know that the urn always contains 10 balls of different colors (for example, red and blue). But only one player (the *sender*) is allowed to draw a certain number of balls from the urn and look at them. The sender puts the balls back into the urn and gives it a nice shake, then the sender draws a new ball from it. Before looking at it, the sender sends a message to the other player (the *receiver*). The receiver reads the message and tries to guess the exact content of the urn."

The motivation for this was to avoid potential confounds, as much as possible, about the purpose of conversation when choosing between probability expressions. Previous research has established that contextual *questions under discussion*, i.e., ways of classifying what counts as important to a conversation, may affect the use and interpretation of probability expressions (Teigen, 1988; Windschitl & Wells, 1998; Lassiter, 2011; Herbstritt, 2015). Moreover, by introducing the participants to a fictive interactive game context we hoped to suggest that they should reason about the effect of their choices on another agent.

Results & analysis Figure 2 shows results from the expression trials, i.e., the percentages of choices for each message in each condition, grouped by uncertainty level. Figure 3 shows mean answers from the likelihood trials, alongside the posterior beliefs that an ideal Bayesian agent should hold.

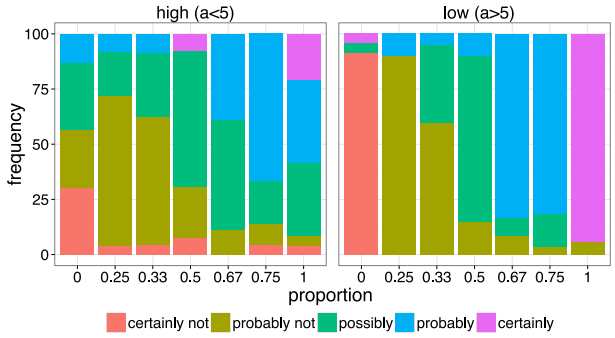


Figure 2: % of message choices in each condition.

We want to test whether choice of probability expression is governed (i) only by participants' probabilistic beliefs about the prejacent *The next draw will be red*, or (ii) by the more complex higher-order uncertainty state, expressed as a function of accessed and observed balls. To do so, we formulate and compare two kinds of multinomial logistic regression models. The first and `simple` model seeks to predict the five-level categorical factor `answer` (*certainly not*, *probably not*, *possibly*, *probably*, *certainly*) with a single metric factor `belief` which is the participants' mean answers in the relevant likelihood trials. The second and `complex` model considers metric factors `observation` and `access` as predictors.

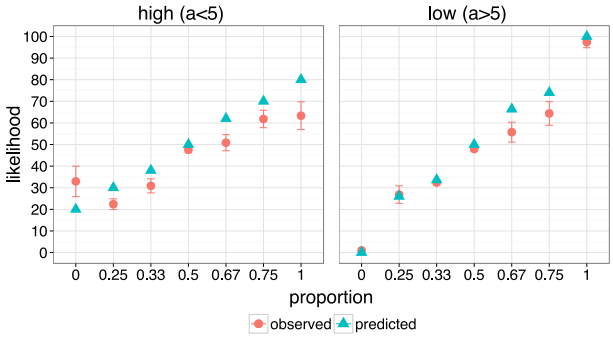


Figure 3: Observed and predicted expected likelihood of the prejacent in each condition.

We also consider an `interaction` model that contains the latter factors' interaction as well.

Table 2: AIC scores.

	`simple`	`complex`	`ineraction`
AIC	662.78	646.69	635.93

AIC scores of best-fits of these regression models are given in Table 2. Despite added complexity, `complex` seems a better model than `simple`, and `interaction` seems to be the best. This suggests that there might be more to the use of probability expressions than just the level of belief in the prejacent; the result suggests that higher-order uncertainty may affect choice of expressions in more subtle ways.[3]

How do speakers choose expressions under higher-order uncertainty? What exactly is the role of observation and access? Can we predict the speakers' choices by assuming that possibility and probability expressions have a simple semantics but the speakers use them pragmatically? To answer these questions we developed a probabilistic model based on a version of the *Rational Speech Acts* model (henceforth, RSA) proposed by Goodman and Stuhlmüller (2013) (G&S).

Computational model

Probabilistic modeling is arguably better suited for capturing the subtleties of (pragmatic) language use than the standard formal semantics framework. RSA has proven to be a successful modeling tool when it comes to complex usage patterns (or pragmatic effects) on the basis of simple semantics and largely shared insights from rational choice theory and information theory.[4] The version of RSA developed by G&S

[3]This result holds also if the factor `observation` is uniformly replaced in the models with a factor containing the proportions of observed red balls over accessed balls and, more importantly, if we restrict the data to cases of genuine high-order uncertainty, i.e. removing condition $a = 10$. Moreover, AIC score of the `simple` model using ideal beliefs instead of measured ones as predictor becomes better than the model using measured beliefs, but is still worse than the `complex` model.

[4]See recent works by Degen, Tessler, and Goodman (2015), Kao and Goodman (2015) and Hawkins, Stuhlmüller, Degen, and Goodman (2015) among others.

deals with exactly the kind of uncertain beliefs that we are interested in, although for a different purpose.

Basic model The basic setup of the model mirrors the setup of the experiment. The urn contains 10 balls, any number of which can be red. The set of natural numbers $S = \{0, \ldots, 10\}$ represents the state space: each $s \in S$ is a possible quantity of red balls in the urn. The set $A = \{0, \ldots, 10\}$ contains the possible access values. Given a state s and access value a, the number of red balls observed by the agent is denoted o, and upper-bounded by $\min(s, a)$.

The first level of uncertainty is determined by the state: $s/10$ is the objective chance that a randomly drawn ball will be red. Crucially, the agent does not know this value. Instead, the agent observes o red balls out of a. Based on this observation the agent forms an uncertain belief over the state space, modeled as the Bayes-inverted conditional probability of observing o red balls out of a drawn balls in state s:

$$\text{speaker.bel}(s|o,a) \propto P(o|a,s) * \text{prior}(s) \tag{1}$$

In turn, $P(o|a,s)$ is given by the hypergeometric model of the urn:

$$P(o|a,s) = \text{hypergeometric}(o;a,s,10) \tag{2}$$

Equation 1 defines the beliefs of a *speaker*, as we aim to model the behavior of message-sending agents. As such, a crucial component of the model is the set of available messages together with their meaning. In the spirit of RSA, we assume a simple literal semantics, expressed as follows in the standard notation of formal semantics:

$$[\![\text{certainly}(p)]\!]_s = 1 \text{ iff } P(p) = 1 \text{ in } s$$
$$[\![\text{probably}(p)]\!]_s = 1 \text{ iff } P(p) > \theta \text{ in } s$$
$$[\![\text{possibly}(p)]\!]_s = 1 \text{ iff } P(p) > 0 \text{ in } s$$

The threshold θ is a free parameter in the model (more about this below). The variable p is to be instantiated with (some sentence equivalent to) *The next ball will be red*.

An important assumption in RSA modeling —loosely based on Paul Grice's Maxim of Quantity, is that the communicative goal of the speaker is to maximize the information transferred to the listener. There are several ways of formalizing the maximization of information. We think of it as bringing the listener's beliefs as close as possible to the speaker's beliefs, i.e. minimizing the (Hellinger) distance between the probability distributions representing those beliefs.[5]

RSA models the behavior of (imperfectly) rational agents. Adopting the terminology of rational choice theory, the speaker's behavior is to soft-maximize the expected utility of the message in the current situation:

$$\text{speaker.prob}(m|o,a) \propto exp(\lambda * \text{EU}(m;o,a)) \tag{3}$$

EU is multiplied by a rationality parameter λ (free in the model) that modulates "how rational" the choice is.[6] EU is defined as negative Hellinger distance (HD) between the speaker's beliefs and the beliefs of a "literal listener":

$$\text{EU}(m;o,a) = -\text{HD}[\text{speaker.bel}(s|o,a), \text{literal.bel}(s|m)] \tag{4}$$

The literal listener is a theoretical construct needed to ground the otherwise infinite process implied in a recursive model of pragmatic reasoning. It is a dummy agent who does not perform any kind of pragmatic reasoning and simply interprets every message m according to the literal semantics:

$$\text{literal.bel}(s|m) = P(s|m \text{ is true}) * \text{prior}(s) \tag{5}$$

Parameters estimation The basic version of the model, M_0 has two free parameters: the threshold θ and the rationality λ. It assumes flat prior beliefs over the state space. We also considered three more complex versions of the model: a model where the prior beliefs are expressed by a symmetric beta-binomial distribution with free shape parameters $\alpha = \beta$ (M_1); a model with two (possibly) different free rationality parameters, λ_{low} and λ_{high}, one for each level of uncertainty (M_2); finally, a model combining these two variations (M_3).

Each model was implemented in JAGS (Plummer, 2003) and the posterior likelihoods of the free parameters given the experimental data were estimated by Bayesian inference via MCMC sampling. We remained uncommitted on the prior distributions over the parameters, except for fixing sensible intervals:

$$\theta \sim \mathcal{U}(0,1) \quad \lambda, \lambda_{\text{low}}, \lambda_{\text{high}} \sim \mathcal{U}(0,20) \quad \alpha \sim \mathcal{U}(0,20)$$

We gathered two chains of 5000 samples for each model after an initial burn-in phase of 5000. Convergence was confirmed via $\hat{R}$ (Gelman & Rubin, 1992). The results were interesting for at least three reasons.

First, the mean value inferred for θ was always equal to 0.55, regardless which model we simulated. This is an important result: our data provide evidence that an objective chance bigger than 55% is enough to consider *probably* as an appropriate expression. This is in line with intuition and previous experimental results. It speaks in favor of the model that data-driven inference recovers this intuitive value for θ without stipulating it from the start.

Second, the estimation of α in M_1 resulted in a mean value of 2.81 (HDI: 1.26-4.78).[7] Notice that $\alpha = 1$ would imply flat prior beliefs over the state space, which can thus reasonably be excluded given our data (more about this below). This

[5] We depart from the informativity criterion adopted by G&S because it does not allow the speaker to send messages that are literally false; however, we want to allow some pragmatic slack to the speaker: it is plausible to think that, for example, chances around 97% or bigger are certain enough for us to say *certainly*, even if this expression is literally true only when the odds are 100%.

[6] As $\lambda \to \infty$, the choice approaches perfect rationality.

[7] All HDIs (*highest density intervals*) reported here comprise 95% of the posterior density, such that no point outside the interval has higher density than any point within.

can be perhaps explained by looking again at the introductory text of the experiment: it is written that "[...] the urn always contains 10 balls of different colors" and it is reasonable to assume that this might have caused a number of the participants to neglect the possibility that the urn contained 0 or 10 red balls (which results in non uniform prior beliefs).

Third, the estimation of λ_{low} and λ_{high} in M_2 resulted respectively in mean values of 7.36 (HDI: 6.03-8.79) and 3.49 (HDI: 2.75-4.24). The fact that the difference between λ_{low} and λ_{high} is different from zero with completely non-overlapping HDIs suggests that our data provide evidence for assuming different rationality parameters for different uncertainty level (more about this below).[8]

Finally, the mean values for the free parameters of M_3 and their HDIs are reported in Table 3.

Table 3: Estimated parameters of M_3.

	θ	α	λ_{low}	λ_{high}
mean	0.55	2.81	6.11	3.16
HDI	0.50-0.59	1.04-5.42	4.75-7.45	2.39-3.95

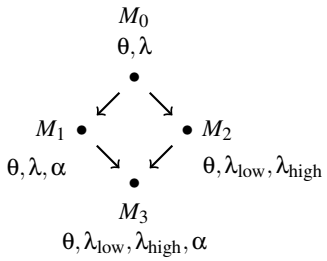


Figure 4: Nesting relations between models.

Model comparison Our four models are nested. M_0 is a special case of both M_1 and M_2, which are special cases of M_3, as depicted in Figure 4. This allows for model comparisons based on Bayes factors (BF) with the Savage-Dickey method (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). As expected, both M_1 and M_2 are more credible models given our data. The Bayes factor in favor of M_1 against M_0 is substantial (approximately $BF = 8.68$). More strikingly, the Bayes factor in favor of M_0 against M_2 is so low that it cannot be computed with normal float precision levels, which in turn means that M_2 is definitely more credible than M_0. Moving to M_3, we get a similarly striking result when we compare it to M_1: the Bayes factor in favor of M_1 approaches zero. On the other hand, the Bayes factor in favor of M_3 against M_2 is only equal to 1.40 which is barely enough to distinguish the two models on the basis of the data. Summing up, allowing our model to use two different rationality parameters for the two different uncertainty levels invariably results in a definitely more credible model.

[8] An intuitive explanation of this result may be that the more precise the participants' belief, the easier it was for them to behave more rationally. In other words, the participants' choice behavior becomes more noisy as their uncertainty increases. As pointed out by a reviewer, in light of this explanation it would seem more principled to allow λ to vary smoothly with the access value. We leave the investigation of such a model for another occasion.

Correlation and model criticism Model comparison based on Bayes factors tells us if/how much a model is better than another, but does not give us an absolute measure of how good a model is. Having picked M_3 as the best model at our disposal, we correlated the experimental observations with the predictions made by the model fitted with the maximum-likelihood estimated values of its free parameters. The results of Pearson's product-moment correlation provide us with a measure of goodness of fit for the model as its best:

$$df = 68;\ r = 0.927;\ 95\%\ \text{ci} : 0.885\text{-}0.954;\ p < 0.001$$

By that (frequently used) measure, predictions of the model look quite good. A more stringent criterion for goodness of fit are posterior predictive checks (PPC) (Kruschke, 2014).

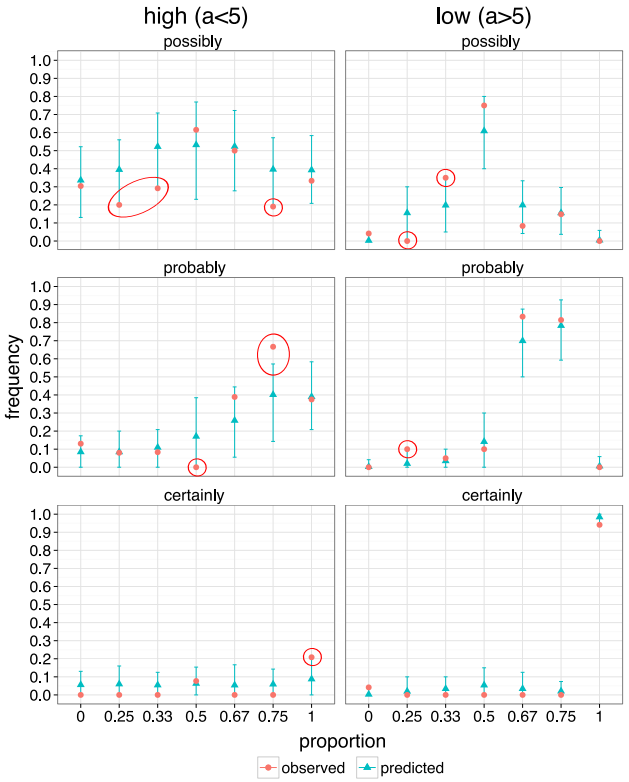


Figure 5: Posterior predictive checks. Red circles highlight discrepancies between observed data and credible values predicted by the model.

PPCs look at hypothetical data generated from a model when parameters are randomly drawn from their posterior distribution (i.e., conditional on the data). By visual means, we then check whether any actually observed data point is "surprising" in the sense that it lies outside of the 95% HDI of the hypothetically generated data. Figure 5 shows the mean

frequencies of posterior predictive message choices of M_3 (in light blue), together with their 95% HDIs. The pink dots represent observed data. Ideally, all observations should lie within the posterior HDIs. Visual inspection of the plots allows us to recognize several critical points where the posterior predictions of the model diverge from the observed data (circled in red). We observe that most critical points are found in the high uncertainty condition, plausibly where most noise occurred in the data. It is also mostly at these points where participants' mean answers in the likelihood trials diverged from idealized Bayesian belief update (see Figure 3). This suggests that failures in PPCs might not be the fault of the pragmatic part of the model, but the belief update part of the model. In other words, it might be that the pragmatic reasoning which lead to participants' message choices was based on uncertain beliefs different from those of an idealized Bayesian reasoner.

Conclusion

We presented experimental evidence suggesting that the speakers' use of possibility and probability expressions depends not only on the objective chance of the event in question but also on the speakers' state of higher-order uncertainty. We formulated a computational model based on RSA that explains the experimental data in terms of simple semantics and standard pragmatic reasoning. Despite its simplicity, the predictions of the model are quite good. Comparing AIC scores for the best-fits of the `interaction` regression model and our theory-driven computational model reveals a striking preference for the latter (AIC: 635.93 vs 270.90).

Future work will attempt to refine the model by measuring participants' full posterior beliefs about the urn and using the measured distributions in the model instead of the Bayesian beliefs. Another line of empirical research will be to investigate speakers' use of nested possibility and probability expressions under higher-order uncertainty: if given the choice, would speakers prefer a nested construction over a simple one to communicate their uncertain beliefs? Moreover, a natural continuation of this work will be to investigate listeners' interpretation of (simple and nested) possibility and probability expressions.

Acknowledgments

Financial support by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) is gratefully acknowledged.

References

Beyth-Marom, R. (1982). How probable is probable? a numerical translation of verbal probability expressions. *Journal of forecasting*, *1*(3), 257–269.

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, *41*(3), 390–404.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In D. C. Noelle et al. (Eds.), *Proceedings of CogSci37*.

Egan, A., & Weatherson, B. (2011). *Epistemic modality*. Oxford University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. In D. C. Noelle et al. (Eds.), *Proceedings of CogSci37*.

Herbstritt, M. (2015). Experimental investigations of probability expressions: a first step in the (probably) right direction. In M. Kaeshammer & P. Schulz (Eds.), *Proceedings of ESSLLI 2015 student session.*

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In D. C. Noelle et al. (Eds.), *Proceedings of CogSci37*.

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 639–650). de Gruyter.

Kruschke, J. (2014). *Doing bayesian data analysis, 2nd edition: A tutorial with r, jags, and stan*. Academic Press.

Lassiter, D. (2011). *Measurement and modality: the scalar basis of modal semantics*. Unpublished doctoral dissertation, NYU Linguistics.

Lassiter, D., & Goodman, N. D. (2015). *Nested and informative epistemics in a graphical models framework.* Talk given at Bridging Logical and Probabilistic Approaches to Language and Cognition, ESSLLI 2015, Barcelona.

Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*(5), 1–81.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of DSC3* (Vol. 124, p. 125).

Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica*, *6*(2), 157–174.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive psychology*, *60*(3), 158–189.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411–1423.

Yalcin, S. (2010). Probability operators. *Philosophy Compass*, *5*(11), 916–37.

Dynamical systems modeling of the child–mother dyad: Causality between child-directed language complexity and language development

Jeremy Irvin (jirvin@umail.ucsb.edu) **Daniel Spokoyny (dspoka@gmail.com)**
College of Creative Studies
University of California, Santa Barbara, CA 93106, USA

Fermín Moscoso del Prado Martín (fmoscoso@linguistics.ucsb.edu)
Department of Linguistics
University of California, Santa Barbara, CA 93106, USA

Abstract

We model the *causal* links between child language (CL) and child-directed language (CDL). We take pairs of sequences of linguistic measurements from a longitudinal study. Each child-mother pair of sequences is considered as an instance of the trajectory of a high-dimensional dynamical system. We then use Multispatial Convergent Cross Mapping to ascertain the directions of causality between the pairs of sequences, that is, whether the complexity of CL drives that of CDL, the complexity of CDL drives that of CL, both, or neither. We find that children are responsive to the amount of speech and the diversity of words produced by their mothers, but not vice-versa. However, the syntactic diversities of the children's utterances drive the syntactic diversity of the mothers' utterances. This is evidence for fine-grained fine-tuning of CDL in response only to the syntax of CL.

Keywords: Causality; Child-Directed Language; Convergent Cross Mapping; Dynamical Systems; Fine-Tuning; Information Theory; Language Acquisition

Introduction

Child-directed language (CDL) –sometimes referred to as "motherese"– is the pattern of language used by parents when talking to young children. It is known to exhibit distinctive characteristics with respect to regular adult language (cf., Snow & Ferguson, 1977): It typically uses shorter utterances, prosody is often exaggerated, redundancy and repetition are higher than normal, and referential context tends to be linked to very immediate contexts. It has been observed that the lexical and syntactic complexity of CDL gradually increases along a child's development (Cross, 1977; Snow, 1989), eventually converging to regular adult language. Whether the simplicity of CDL relative to regular adult language plays a functional role in facilitating language acquisition is a contentious issue in the literature. Some researchers argue that "starting small" (Elman, 1993) is a fundamental aspect that facilitates language acquisition (e.g., Dominey & Dodane, 2004). Others, however, claim that there is no facilitating role played by such simplicity (e.g., Pinker, 1994). This latter group would claim that, to some degree, children might exploit universal aspects of language structure, and their syntactic performance would be unrelated to the input to which they have been exposed. Those researchers advocating for a functional role of the simplified input refer to *fine-tuning* (Snow, 1989, 1995; Sokolov, 1993) as the process by which caregivers adjust the complexity of CDL as a function of the level of complexity of the language produced or understood by the child.

A weak and a strong version of the fine-tuning hypothesis compete in the literature. In the weak interpretation, although parents gradually increase the complexity of their CDL, they do not do so as a direct response to the specific properties of the utterances produced by their children, but rather they adjust to the children's overall level of cognitive development, irrespective of the specificities of the language they produce and understand. In this line, several studies have failed to find a direct link between the complexity of the parent's language and that of the child's (Newport, Gleitman, & Gleitman, 1977; Scarborough & Wycoff, 1986; Valian, 1999). These findings suggest that –if anything– the complexity of CDL might increase as a function of the child's age or overall level of development, but not so much as a direct response to the detailed properties of CL. In contrast, other researchers have found evidence supporting a strong version of the fine-tuning hypothesis: That parents adapt the complexity of CDL in direct response to the specific properties of CL (Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002; Kunert, Fernández, & Zuidema, 2011; Murray, Johnson, & Peters, 1990; Roy, Frank, & Roy, 2009; Snow, 1995; Sokolov, 1993). The strong version of the fine-tuning hypothesis has become the dominant view in the field, considered a well-established fact by influential researchers (e.g., MacWhinney, 2014).

Dynamical systems offer powerful tools for modeling human development (e.g., Smith & Thelen, 2003; van Geert, 1991). These models provide a mathematical framework for implementing the principle that development involves the mutual and continuous interaction of multiple levels of the developing system, which simultaneously unfold over many time-scales. Typically, a dynamical system is described by a system of coupled differential equations governing the temporal evolution of multiple parts of the system and their interrelations. In recent years, it has been noticed that, in human development, such systems extend beyond the individual. In particular, it has been found that the linguistic and behavioral interaction between parent-child dyads can be jointly considered as part of a single dynamical sytem encompassing both the child and the parent (Dale & Spivey, 2006; Steenbeek & van Geert, 2007). In this direction van Geert (1991)

shows that the interaction of components within the child-parent dyad can also be modeled on the time-scale of development itself. One difficulty that arises when trying to model a dynamical system as complex as the joint development of language in a parent–child dyad is that many factors that are important for the evolution of the system might not be available or might not be easily measurable or –even worse– there are additional variables relevant for the system of which the modeler is not even aware. In this respect, a crucial development was the discovery that, in a deterministic coupled dynamical system –even in the presence of noise– the dynamics of the whole system can be satisfactorily recovered using measurements of a single of the system's variables (Takens' Embedding Theorem; Takens, 1981).

The finding above opens an interesting avenue for understanding the processes involved in language acquisition (perhaps more suitably termed language *growth*, following van Geert, 1991). In the same way that systems of differential equations can be used to model the evolution of ecosystems (e.g., predator-prey systems), one could take measurements of the detailed properties of CL and CDL, and build a detailed system of equations capturing the macroscopic dynamics of the process. However, in order to achieve this, it is necessary to ascertain the ways in which different measured variables in the system affect each other. This problem goes beyond estimating correlations (as could be obtained, for instance, using regression models), as one needs to detect asymmetrical *causal* relations between the variables of interest, so that these causal influences can be incorporated into the models.

In this study, we investigate the causal relations between measures describing CL (i.e., number of words produced, lexical diversity, and mean length of utterances) and the equivalent measures in CDL, using the longitudinal data provided in the Manchester Corpus (Theakston, Lieven, Pine, & Rowland, 2001). In order to detect causal relations between the different measures, we make use of state space reconstruction relying on Takens (1981)'s Embedding Theorem, and recently developed techniques for assessing the strength of causal relations in dynamical systems (Multispatial Convergent Cross Mapping; Clark et al., 2015). Our results provide a detailed picture on the presence and effects of fine tuning across different linguistic strata, and provide an important inroad into building a detailed dynamical system jointly considering the co-development of CDL and CL.

Causality Detection in Dynamical Systems

The commonly-held maxim that "correlation does not imply causation" is often misinterpreted to mean that one might have correlated variables that are not involved in any causal relations. Pearl (2000) clarifies that, whenever two variables are correlated, there *must* exist some causal link between them. Namely, if variables A and B are found to be correlated, then one of four possibilities must be true: (a) A causes B, (b) B causes A, (c) A and B form a feedback loop, each causing the other, or (d) there is a third variable C causing both A and B. In order to ascertain a *direct* causal relation between any two variables, one needs to discard possibility (d) above, for which one requires knowledge of all other variables that might be of importance for the system in question. Without knowledge of these possible intervening variables, causal inference using these methods must remain strongly suspect. This problem is, of course, particularly acute when one is dealing with systems as complex as is human language.

The most common technique for assessing the presence of causal relations between time series is Granger-causality (Granger, 1981). It relies on the notion of *separability*, this is, that the information contained by a causal source is unique to it, so that just eliminating that variable from consideration suffices for eliminating the information that it contributes. Purely stochastic systems often exhibit separability. Unfortunately, however, separability is not a property that is exhibited by deterministic non-linear dynamical systems. For studying the interactions of species within ecosystems, Sugihara et al. (2012) introduced *Convergent Cross Mapping* (CCM), a causality-detection technique that is valid for non-separable systems, is capable of identifying weakly coupled variables even in the presence of noise, and –crucially– can distinguish direct causal relations between variables from effects of shared driving variables (i.e., in possibility (d) from the previous paragraph, CCM would *not* find causality).

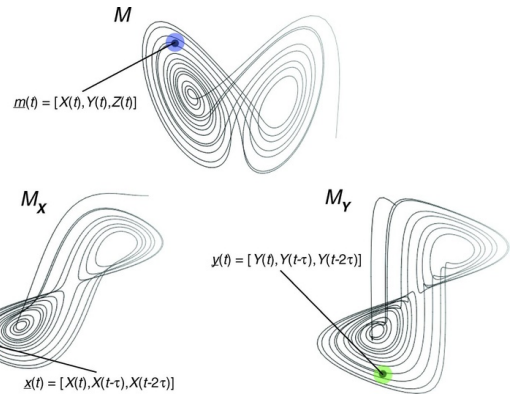


Figure 1: Reconstructed manifold for Lorenz's system (M; top), as well as the shadow manifolds reconstructed considering only X (M_X; bottom-left) and Y (M_Y; bottom-right) (reprinted with permission from Sugihara et al., 2012).

For instance, consider E. Lorenz's often studied dynamical system including three coupled variables $X(t)$, $Y(t)$, and $Z(t)$ whose co-evolution is described by the system of differential equations

$$\begin{cases} \dfrac{\mathrm{d}X}{\mathrm{d}t} = \sigma(Y - X) \\ \dfrac{\mathrm{d}Y}{\mathrm{d}t} = X(\rho - Z) - Y \\ \dfrac{\mathrm{d}Z}{\mathrm{d}t} = XY - \beta Z \end{cases} . \tag{1}$$

The first equation in this system indicates that there is a re-

lation by which Y causes X, as the change in X (i.e., its future value) depends on the value of Y (i.e., the future of X depends on the past of Y even after the past of X itself has been considered), a causal relation whose strength is indexed by parameter σ. The manifold defined by these three variables (Lorenz's famous strange attractor), which we can denote by M, is plotted in the top of Fig. 1. In many circumstances, however, not all variables of the system are available (some might be difficult to measure, or we might not even be aware of their relevance). It is at this point that Takens (1981)'s Embedding Theorem comes into play. Informally speaking, the theorem states that the properties of a coupled dynamical system's attractor can be recovered using only measurements from a single one of its variables. This is achieved by considering multiple versions of the same variable lagged in time, that is, instead of plotting $(X[t], Y[t], Z[t])$, when only measurements of X are available, we can plot $(X[t], X[t+\tau], \ldots, X[t+(E-1)\tau])$. These reconstructed manifolds are termed "shadow" manifolds. M_X denotes the shadow manifold of M reconstructed on the basis of X alone. There are well-studied techniques for finding the appropriate values for the parameters for the lag τ and the number of dimensions E (c.f., Abarbanel, Brown, Sidorowich, & Tsimring, 1993) so that the properties of the original manifold M are recovered by the shadow manifold M_X. Fig. 1 illustrates this point by plotting the shadow manifolds M_X (bottom-left) and M_Y (bottom-right) for the Lorenz system. Notice how both shadow manifolds recover much of the original's structure, using only knowledge of one of its three variables.

Each point in the original manifold M maps onto points in its shadow manifolds, as is illustrated by the points labelled $m(t)$, $x(t)$, and $y(t)$ in Fig. 1. The preservation of the topological properties of the original manifold in its shadow manifolds entails that points that are close-by in the original manifold will also be close-by in its shadow versions. This implies that, for causally linked variables within the same dynamical system, the state of one variable can identify the states of the others. Sugihara et al. (2012) noticed that, when one variable X stochastically drives another variable Y, information about the states of X can be recovered from Y, but not vice-versa. This is the basic insight of the CCM method. To test for causality from X to Y, CCM looks for the signature of X in Y's time series by seeing whether the time indices of nearby points on M_Y can be used to identify nearby points on M_X. Crucially, in order to distinguish causation from mere correlation, CCM requires *convergence*, that is, that cross-mapped estimates improve in estimation accuracy with the sample size (i.e., "library size") used for reconstructing the manifolds. As the library size increases, the trajectories defining the manifolds fill in, resulting in closer nearest neighbors and declining estimation error, which is reflected in a higher correlation coefficient between the points in the neighborhoods of the shadow manifolds. Convergence then becomes the necessary condition for inferring causation. Using both artificial systems and ecological time-series with known dynamics, Sugihara and his colleagues demonstrated that this technique successfully recovers true directional causal relations when these are present, and –crucially– is able to discard spurious causation in the case when both variables are causally driven by a third, unknown, variable, but there is no true direct causation between them.

An inconvenience of CCM, and in general of techniques that rely on manifold reconstruction, is that they generally require that relatively long time-series of the behavior of the system are available. Such long series are, however, very difficult, if not impossible, to obtain in many fields, including of course language acquisition. One can however obtain multiple short time series from different instances of a similar dynamical system. In ecology, for instance, one can obtain short sequences of measurements of the population densities of a group of species measured at different places and times. In language acquisition, we might have multiple, relatively short longitudinal sequences of measurements from different children. With this in mind Clark et al. (2015) developed Multispatial CCM (mCCM), an extension of CCM able to infer causal relations from multiple short time-series measured at different sites, making use of dewdrop regression (Hsieh, Anderson, & Sugihara, 2008) to take the additional heterogeneity into account.

Materials and Methods

We obtained from the CHILDES database (MacWhinney, 2000) the transcriptions contained in the Manchester Corpus (Theakston et al., 2001). This corpus contains annotated transcripts of audio recordings from a longitudinal study of 12 British English-speaking children (6 girls and 6 boys) between the ages of approximately two and three years. The children were recorded at their homes for an hour while they engaged in normal play activities with their mothers. Each child was recorded on two separate occasions in every three-week period for one year. Each recording session is split into two half-hour periods. The annotations include the lemmatized form of the words produced by both the children and their mothers (incomplete words and small word-internal errors were manually corrected in the lemmatization).

In order to increase the sample size in each period, we used a sliding window technique (akin to that used in Moscoso del Prado Martín, 2014), by computing measures for the samples contained in overlapping windows of three consecutive corpus files. In this way, at each point we obtained samples originating from two files from the same recording session, and a file from either the previous or the next recording session. For each child and mother, we recorded the total number of words they produced, the lexical diversity measured as the entropy of the lemmas produced (following the estimation method of Moscoso del Prado Martín, in press, which is demonstrated to be accurate and unbiased for these sample sizes), and the mean length of the utterances (MLU) they produced. Instead of measuring MLU in morphemes (Brown,

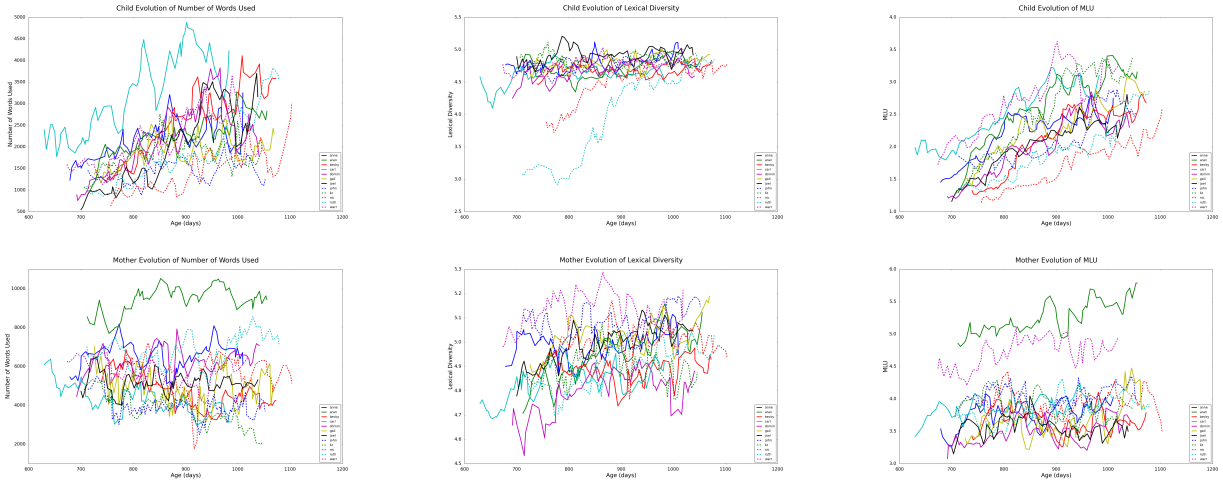


Figure 2: Evolution of the measures under consideration as a function of the children's ages, for the children (top row) and their mothers (bottom row). The total number of words produced are plotted in the left column, the lexical diversities in the middle column, and the MLUs in the right column.

1973), we used the simpler, but equally accurate measure in number of words (Parker & Brorson, 2005). In these ages, MLUs are well known to provide an accurate measure of the syntactic richness of the utterances produced (Brown, 1973), and in fact correlate almost perfectly with explicit measurements of grammatical diversity (Moscoso del Prado Martín, in press).[1] Fig. 2 plots the temporal evolution of the three measures for the children and their mothers.

Table 1: Values of the parameters used for shadow attractor reconstruction.

	Child			Mother		
	Number of Words	Lexical. Diversity	MLU	Number of Words	Lexical Diversity	MLU
τ	3	2	3	3	3	3
E	3	3	4	5	4	4

The optimal time-lags (τ) for constructing the shadow manifolds were estimated as the first local minimum of the lagged self-information in each of the time series (c.f., Abarbanel et al., 1993). The optimal embedding dimensionalities (E) were estimated by optimizing next-step prediction accuracy. The estimates were not found to differ significantly across children or mothers, and therefore for each measure, we used a single estimate of (τ, E) for all children (see Table 1), and a single estimate for all mothers. The time series were checked to ensure that they contained non-linear signal not dominated by noise using a prediction test, and the presence of directional causality between children's and mothers' measures was tested for each of the three variables using mCCM, with 1,000 bootstrapping iterations used to assess the p-values.[2] Finally, to account for our lack of *a priori* predictions on the causal directions to be tested, the p-values were adjusted for multiple comparisons using the false discovery rate for correlated data (FDR; Benjamini & Yekutieli, 2001).

Results and Discussion

Fig. 3 plots the mCCM results for each pair of shadow manifolds. The curves plot how the correlations between nearest neighbours across shadow attractors evolve as one considers increasingly larger library sizes. The p-values report whether these correlation values are significantly increasing (the p-values are obtained by a Monte Carlo method with 1,000 resamplings, and further FDR-corrected for multiple comparisons). Panel (a) shows that convergence indicates a significant causal relation between the number of words produced by the mother, and the number of words produced by her child ($p < .001$) which is not significantly present in the opposite direction ($p = .067$). A similar picture arises in the lexical diversities in panel (b): The richness of the vocabulary produced by the mothers influences the richness of the vocabulary produced by their children ($p = .014$), but the richness of the vocabulary used by children does not significantly affect that of their mothers ($p = .481$). In contrast, panel (c) shows that, in terms of MLU, the language produced by children and their mothers form a feedback loop, with significant causal relations in both directions (child to mother: $p < .001$; mother to child: $p = .005$).

In terms of the amount of speech, or the richness of the vocabulary used, these results indicate that the mothers are not increasing the complexity of CDL in response to the details

[1] We also tested a measure of inflectional diversity (Moscoso del Prado Martín, 2014, in press), which was not found to produce any reliable causal effects, and is therefore not discussed further. However, the presence of these additional tests was nevertheless taken into account when correcting for multiple comparisons.

[2] All computations, except for τ selection, we done using R package `multispatialCCM` (Clark et al., 2015).

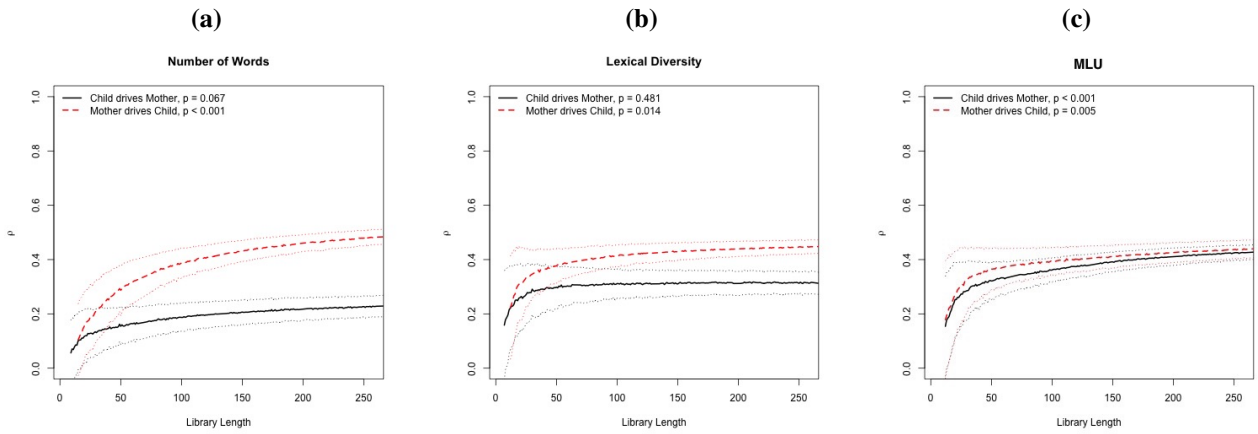


Figure 3: For each of the measures considered, the panels plot the evolution of Pearson's correlation coefficient (ρ) between the predicted and predicting shadow manifolds. The dashed lines denotes the standard deviation of the estimates. The *p*-values indicated in the legend were obtained by the bootstrapping procedure described by Clark et al. (2015). A value of ρ significantly increasing with library length is mCCM's indication of causality between two variables.

of CL, but the children's performance still benefits from the increased quantity and diversity of words. This is evidence that weak lexical fine-tuning serves a functional role. In contrast, when it comes to MLUs, the bidirectional causality provides clear evidence for a strongly coupled system with feedback. As is shown in Moscoso del Prado Martín (in press), MLUs are in fact almost perfectly correlated (i.e., Pearson's $r \approx .96$) with an explicit measure of the diversity of the syntactic structures used in a sample (i.e., the *syntactic diversity* of the sample). Mothers adjust the complexity of their syntactic structures as a direct response to the syntactic complexity of the utterances produced by their children, as is advocated by the strong version of the fine-tuning hypothesis.

Our results provide direct evidence for the fine-tuning hypothesis. For the first time, we have *explicitly* demonstrated that, in all measures studied, the children benefit from the gradual increase in complexity of CDL, as is indicated by the directional causalities found between the measures in CDL and those in CL. In addition, only for the case of syntax, we find direct evidence for the strong version of the fine-tuning hypothesis: The complexity of the syntactic structures produced by mothers are directly caused by those of the syntactic structures produced by their children.

These findings are the first step in building a macroscopic level dynamical system model of language acquisition explicitly considering children jointly with their environment. In order to build such a model, one also needs to test for the explicit causal components of complexity within an individual (e.g., what are the causal connections between increased vocabulary and increased syntactic knowledge?), and those present across individuals and linguistic strata; for instance, it has been reported that the amount of speech produced by parents influences the growth of both the vocabulary (e.g., Hurtado, Marchman, & Fernald, 2008; Weisleder & Fernald, 2013), and the MLUs in the children (Barnes, Gutfreund, Satterly, & Wells, 1983). We anticipate further research in these directions.

References

Abarbanel, H. D. I., Brown, R., Sidorowich, J. J., & Tsimring, L. Sh. (1993). The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, *65*, 1331–1392.

Barnes, S., Gutfreund, M., Satterly, D., & Wells, G. (1983). Characteristics of adult speech which predict children's language development. *Journal of Child Language*, *10*, 65–84.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA.: Harvard University Press.

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., & Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, *96*, 1174–1181.

Cross, T. G. (1977). Mothers' speech adjustments: The contribution of selected child listener variables. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children* (pp. 151–188). Cambridge, England: Cambridge University Press.

Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 391–430.

Dominey, P. F., & Dodane, C. (2004). Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics*, *17*, 121–145.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Granger, C. W. J. (1981). Testing for causality. *Journal of Economic Dynamics and Control*, *2*, 329–352.

Hsieh, C., Anderson, C., & Sugihara, G. (2008). Extending nonlinear analysis to short ecological time series. *The American Naturalist*, *171*, 71–80.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, *11*, F31–F39.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*, 337–374.

Kunert, R., Fernández, R., & Zuidema, W. (2011). Adaptation in child-directed speech: Evidence from corpora. In *SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 112–119). Los Angeles, CA.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed., Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2014). What we have learned. *Journal of Child Language*, *41*, 124–131.

Moscoso del Prado Martín, F. (2014). Grammatical change begins within the word: Causal modeling of the coevolution of Icelandic morphology and syntax. In P. Bello, M. Guarini, M. McShane, & B. Scasselatti (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2657–2662). Austin, TX: Cognitive Science Society.

Moscoso del Prado Martín, F. (in press). Vocabulary, grammar, sex, and aging. *Cognitive Science*.

Murray, A., Johnson, J., & Peters, J. (1990). Fine-tuning of utterance length to preverbal infants: effects on later language development. *Journal of Child Language*, *17*, 511–525.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children* (pp. 109–149). Cambridge, England: Cambridge University Press.

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, *25*, 365–376.

Pearl, J. (2000). *Causality: models, Reasoning and Inference*. Cambridge, England: Cambridge University Press.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper-Collins.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In N. Taatgen, H. van Rijn, K. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2106–2111). Austin, TX: Cognitive Science Society.

Scarborough, H., & Wycoff, J. (1986). Mother, I'd still do it myself: Some further non-effects of 'motherese'. *Journal of Child Language*, *13*, 431–437.

Smith, L., & Thelen, E. (2003). Development as a dynamic system. *TRENDS in Cognitive Sciences*, *7*, 343–348.

Snow, C. E. (1989). Understanding social interaction and language acquisition; sentences are not enough. In M. H. Bornstein & J. S. Bruner (Eds.), *Interaction in human development* (pp. 83–103). Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, C. E. (1995). Issues in the study of input: Fine-tuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher & B. MacWhinney (Eds.), *The Handbook of Child Language* (pp. 180–193). Oxford, England: Blackwell.

Snow, C. E., & Ferguson, C. A. (Eds.). (1977). *Talking to Children*. Cambridge, England: Cambridge University Press.

Sokolov, J. L. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, *29*, 1008–1023.

Steenbeek, H. W., & van Geert, P. L. C. (2007). A theory and dynamic model of dyadic interaction: Concerns, appraisals, and contagiousness in a developmental context. *Develpmental Review*, *27*, 1–40.

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E. R., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*, 496–500.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand & L.-S. Young (Eds.), *Dynamical Systems and Turbulence* (pp. 366–381). Berlin, Germany: Springer Verlag.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, *28*, 127–152.

Valian, V. (1999). Input and language acquisition. In W. C. Ritchie & T. K. Bhattia (Eds.), *Handbook of child language acquisition* (pp. 497–530). San Diego, CA: Academic Press.

van Geert, P. L. C. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, *98*, 3–53.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152.

The Relational SNARC: Spatial Representation of Nonsymbolic Ratios?

Percival G. Matthews (pmatthews@wisc.edu)
Department of Educational Psychology, 1025 W. Johnson Street
Madison, WI 53706 USA

Rui Meng (rmeng5@wisc.edu)
Department of Educational Psychology, 1025 W. Johnson Street
Madison, WI 53706 USA

Elizabeth Y. Toomarian (toomarian@wisc.edu)
Department of Educational Psychology, 1025 W. Johnson Street
Madison, WI 53706 USA

Edward M. Hubbard (emhubbard@wisc.edu)
Department of Educational Psychology, 1025 W. Johnson Street
Madison, WI 53706 USA

Abstract

Recent research has highlighted the operation of a ratio processing system that represents the analog magnitudes of nonsymbolic ratios. This study investigated whether such representations would demonstrate spatial associations parallel to the SNARC (spatial numeric association of response codes) effect previously demonstrated with whole number magnitudes. Participants judged whether nonsymbolic ratio test stimuli were larger or smaller than reference stimuli using response keys located alternately either on the left or on the right side of space. Larger ratio magnitudes were associated with the right side of space and smaller magnitudes with the left. These results demonstrate that nonsymbolic ratio magnitudes – defined *relationally* by pairs of components – are characterized by a left-to-right spatial mapping. The current focus on ratio magnitudes expands our understanding of the basic human perceptual apparatus and how it might provide tools that grant intuitive access to more advanced numerical concepts beyond whole numbers.

Keywords: fractions; nonsymbolic ratios; ratio processing system; SNARC effect; mental number line; magnitude representation

Introduction

Recent years have brought increased attention to human abilities to process the magnitudes of numerical fractions (Bailey, Hoard, Nugent, & Geary, 2012; Kallai & Tzelgov, 2009; Siegler et al., 2012). Although understanding this ability has practical importance and may inform the design principles for promoting school learning, there are compelling reasons to investigate the basic science behind processing fraction magnitudes as well. In particular, learning more about the representation of fraction magnitudes has implications for our understanding of how humans gain access to numerical concepts. Moreover, this line of research may provide insight into the ways humans perceive magnitudes more generally.

Much work on the psychophysics of numerical perception focuses on the ways we process whole numbers and their nonsymbolic analogs, such as the number of dots in an array or the magnitudes of whole number symbols (e.g., Kaufman, Lord, Reese, & Volkmann, 1949; Moyer & Landauer, 1967; Stevens, 1957). Indeed, seminal works have demonstrated deep parallels between the ways humans perceive the magnitudes of whole number values and their analogs in nonsymbolic numerosities. Perhaps most notable among these findings was Moyer and Landauer's (1967) discovery that, when participants were asked to compare numerals, they first converted numerals to analog magnitudes and compared them "in much the same way that comparisons are made between physical stimuli such as loudness or length of line." (p. 1520). Many have taken these parallels to suggest that the ability to process nonsymbolic numerosities may be a key foundational ability that helps ground whole number concepts (e.g., Feigenson, Dehaene, & Spelke, 2004; Nieder & Dehaene, 2009; Piazza, 2010).

Several recent studies began to suggest that humans are also capable of processing the nonsymbolic analogs of fractions values (Chesney & Matthews, 2013; Jacob, Vallentin, & Nieder, 2012; Matthews, Lewis, & Hubbard, 2015). These studies extend the psychophysics of perception to the study of nonsymbolically instantiated ratio magnitudes. What is interesting about these magnitudes is that they do not correspond to individual stimuli like a single dot array or a single line segment; instead, they are compound stimuli whose magnitudes are determined relationally. For instance, in Figure 1, panel (a) represents the same ratio as panel (b), despite the fact that each of its components is larger than those of ratio (b). Similarly, it is a smaller ratio than that in panel (c), despite the fact that its components are larger when considered individually. Focusing on the perception of nonsymbolic ratio magnitudes stands to expand our understanding of our basic perceptual apparatus and how it might provide tools that

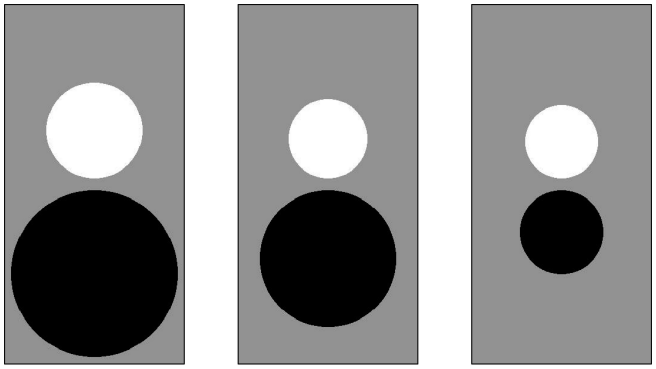


Figure 1. Sample circle stimuli; panels (a) and (b) are two possible instantiations of the magnitude 1/3; panel (c) represents the magnitude 7/9.

grant intuitive access to more advanced numerical concepts beyond whole numbers.

To date, studies have suggested that the nonsymbolic ratio processing is operative among nonhuman primates (Jacob et al., 2012), human infants (McCrink & Wynn, 2007), elementary school age children (Boyer, Levine, & Huttenlocher, 2008), typically developing adults (Chesney & Matthews, 2013), and individuals with limited number vocabularies and formal arithmetic skills (McCrink, Spelke, Dehaene, & Pica, 2013). Moreover, the ability correlates with symbolic math abilities in ways that are distinct from the approximate number system or ANS (Matthews et al., 2015, see also Hansen et al., 2015). Despite the above evidence demonstrating that humans can process nonsymbolic ratio magnitudes, much remains unknown about the nature these abilities.

In order to better understand whether non-symbolic ratios – these visuospatial fraction analogs – are processed similarly to whole number magnitudes, we tested whether they elicited spatial mappings. Many have argued for the association of number and space (e.g., Hubbard, Piazza, Pinel, & Dehaene, 2005; Newcombe, Levine, & Mix, 2015), with research on the SNARC (spatial numeric association of response codes) effect as the most well-known example (Dehaene, Bossini, & Giraux, 1993). The SNARC effect – an association between left responses and small numerical quantities and between right responses and large numerical quantities – is considered to be an indication that numbers are represented on a mental number line, spatially oriented from left to right with increasing magnitude (Dehaene et al., 1993). Indeed, in an oft-cited piece, Walsh (2003) predicted that the SNARC effect is only a special case of a SQUARC (spatial quantity of response codes) whereby all quantities represented in analog form are mapped to space.

Whether the analog perception of nonsymbolic ratios is cast in terms of number specifically or in terms of magnitude more generally, we hypothesize that such magnitudes should be associated with space, resulting in a SNARC or SQUARC effect. For convenience sake, we will simply refer to any such findings with the current stimuli as a SNARC effect. With the present study, we sought to elicit the SNARC using novel ratio stimuli.

Method

Participants

37 undergraduate students at a major Midwestern university participated for course credit (13 males; 30 right handed; ages 18 – 22).

Materials and Design

All stimuli were presented on 22" monitors using the MATLAB Psychophysics Toolbox Version 3 (PTB-3) (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). We constructed three different types of nonsymbolic ratio comparison tasks: nonsymbolic circle ratios, nonsymbolic line ratios, and nonsymbolic dot ratios (see Figure 2 for sample ratio formats). In each, participants were asked to decide whether a test ratio was larger or smaller than a reference ratio. For some symbolic SNARC tasks (e.g., Dehaene, Dupoux & Mehler, 1990; Gevers, Verguts, Reynvoet, Caessens & Fias, 2006; Santens & Gevers, 2008), stimuli are compared to a memorized standard located halfway between the extremes of the stimulus set. However, there is no nonsymbolic standard that could be referenced in memory without a label to serve as a perfect analog for this task. Thus, we used a nonsymbolic standard presented before each test stimulus in place of a memorized standard in the current paradigm. The set of test stimuli were used to instantiate magnitudes corresponding to the 26 irreducible, single-digit proper fraction values except ½, which was at the midpoint of the stimulus range and was reserved as the reference value for comparison (drawn from the symbolic fractions list used in Toomarian and Hubbard, under review). Because the stimuli were nonsymbolic, the "irreducible" and "single-digit" qualifiers do not strictly apply in the current case. The sizes of the ratio components used to instantiate those values varied from trial to trial, and participants were asked to focus on the overall ratio each time. The reference ratio was fixed to ½, and the sizes of the components used to instantiate it also varied from trial to trial. We randomly varied components as a control to reduce reliance on the magnitudes of individual components and possible pattern-recognition strategies. These tasks assessed the ability to discriminate between nonsymbolic ratio values composed of circle areas, line segments, or dot arrays.

Circle Ratios Each stimulus was composed of a pair of circles, with one serving as a figurative 'numerator' and another as the 'denominator'. Numerator circles were white and denominator circles were black, both presented on a gray background, similar to the stimuli used by Matthews et al. (2015). Ratio size was defined as the area of the white circle up top to the area of the black circle on bottom. The summed areas of the numerator circle and the denominator circle were constrained such that it occupied between 20-40% of an invisible rectangle of 600 pixels high and 300 pixels wide.

Line Ratios Each stimulus was composed of a pair of lines. Shorter white lines served as numerators, and longer black lines served as denominators for all stimuli. They were both presented on a gray background. The distance between two lines was 50 pixels, and the width of both lines was 20 pixels. The summed lengths of the numerator and denominator lines were constrained such that they occupied 30 – 60% of the height of a 600 x 300 pixel frame. Additionally, the vertical position of the shorter line varied randomly.

Dot Ratios Each stimulus was composed of a pair of dot arrays arranged to form a nonsymbolic ratio. The numerators of the nonsymbolic numerosity arrays were composed of black dots on a white background, and the denominators were composed of white dots on a black background. Both were presented in a 600 x 300 pixel frame on a gray background. Individual dot size varied between 8 and 12 pixels in diameter. To discourage participants from counting or using computational procedures to estimate the ratios, the summed numerosity of numerator dots and denominator dots fluctuated from 150 to 200. The smallest numerosity displayed in any given array was 15. Dots were randomly and evenly distributed in each array.

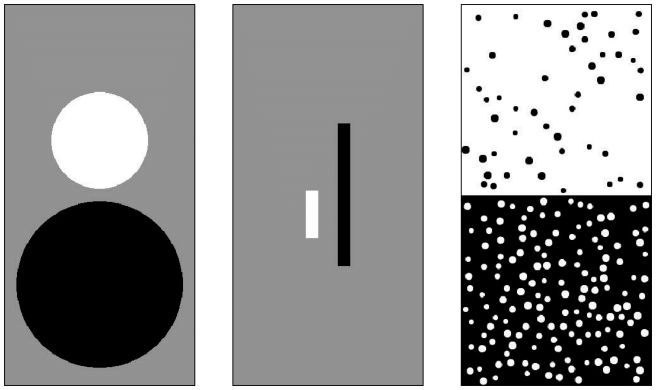


Figure 2: Sample circle, line, and dot ratio stimuli; all represent fraction magnitude of 1/3.

Procedure

Participants were presented with three blocks of comparison trials. Each block was composed wholly of trials of a particular stimulus type. Presentation order of circle and line blocks were counterbalanced across participants, and dot block was always presented last. Participants first saw instructions, then received eight practice trials, and then performed the formal experimental trials. Accuracy feedback was provided for practice trials but not for test trials. Each participant completed all tasks in one hour-long session.

Participants were instructed to press the space bar to initiate each trial. Each trial began with a fixation cross presented in the center of the screen for 500 ms, immediately followed by an 800 ms presentation of the reference ratio stimulus (various instantiations of ½). The reference stimulus was followed by another fixation cross for 500 ms, followed by the test ratio stimulus which remained on screen until participants submitted a response or timed out at 3000 ms (see Figure 3).

After viewing both the reference and the test ratios, participants were asked to decide if the second ratio was larger or smaller than the first, which was always equal to ½. Participants made decisions via button press on a standard keyboard (19 mm horizontal center-to-center distance), with either the left ("D" key) or right ("K" key) buttons indicating that the second ratio was larger. Response side was counterbalanced both within blocks of corresponding stimuli and across participants, meaning that half of the participants pressed "K" for larger with right hand and "D" for smaller with left hand first, while the other half participants pressed "D" for larger with left hand and "K" for smaller with right hand first. Each ratio value appeared eight times for circle and line stimuli, but appeared only four times for dot stimuli. There were 208 total trials for each circle and line stimuli type, and 104 trials for dot stimuli type, resulting a total of 520 trials.

Note that our primary interest was in finding the SNARC for ratios instantiated by continuous quantities (circles and lines) because – unlike dot arrays – individual circles and lines do not map to any particular number. We therefore felt that ratios of continuous nonsymbolic quantities provided a more stringent test for ratio processing than dots. Given time constraints for each experimental session, we only conducted a relatively impoverished test of the dot comparison condition: dot comparisons always came last, and stimuli were only seen half as often as the line and circle stimuli.

For line stimuli, participants were told to estimate "the ratios between line lengths." For circle stimuli, participants were specifically told to estimate "the ratios between circle *areas*, or how much room each circle takes up on the screen." They were also told stimuli would flash too briefly to measure or to use calculations, so they should "just try to feel out the ratio instead of applying a formula." These instructions parallel protocols we have used elsewhere (Chesney & Matthews, 2013; Matthews et al., 2015), and participants quickly understood the tasks.

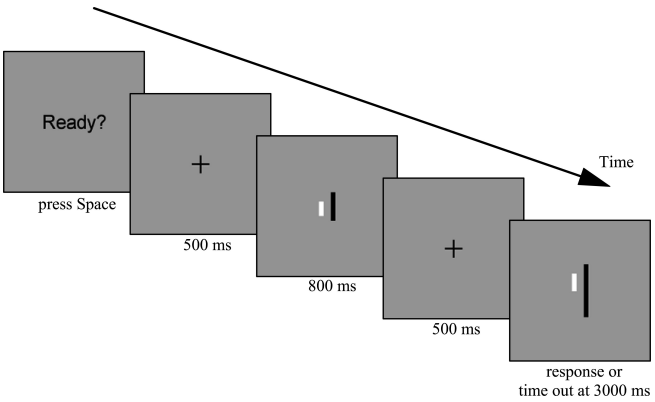


Figure 3. A sample line ratio comparison trial.

Results

Prior to analysis, we first excluded trials with reaction times below 250 ms or that were more than 3 standard deviations faster or slower than a participant's mean reaction time for that task. Nine participants were excluded from the analyses because of very low accuracy (below 60%) in the experimental task. For remaining participants (n=28), average accuracy across three tasks was: 87.31% (circle ratios), 88.70% (line ratios), and 85.58% (dot ratios). Only RTs for accurate responses were used for SNARC analyses,

To test for the presence of the SNARC effect, mean reaction times for correct responses made with the left hand were subtracted from right hand mean reaction times for each fraction magnitude across all participants. This measure, dRT, is positive when a participant responds faster with the left hand for a particular stimulus and negative if the participant responds faster with the right hand. A classical SNARC effect is noted when dRTs are more positive for small numbers and more negative for large numbers, yielding a negative slope when dRT is regressed against stimulus magnitude. However, it is often the case that dRTs for small numbers are not actually positively valued (e.g., Dehaene et al., 1993; Fias, Brysbaert, Geypens, & d'Ydewalle, 1996) because intercept is confounded with handedness effects (i.e., right hand responses are typically faster in right handed participants). Thus, slope is generally taken as the key measure of the SNARC effect.

We ran an omnibus regression on mean dRTs for against ratio magnitudes, including stimulus type and stimulus type X magnitude interactions in the model (line stimuli served as the baseline, with dummy codes for circle and dot stimuli). There was a significant regression slope based on magnitude b = -66.57, p = .03. However, there was no significant main effect for stimulus type, nor was there a significant type X magnitude interaction. This indicates that the SNARC effect for all three types of stimuli were statistically indistinguishable, suggesting that the effects on dRT really did correspond to the ratio magnitudes presented independently of format.

For ease of presentation, we depict the results of separate regressions for each format in Figures 4, 5, & 6. The slopes for each was significant ($b_{circles}$ = -59.84, p = .02; b_{lines} = -66.57, p = 0.04; b_{dots} = -68.33, p = 0.04). These figures illustrate that all three types of nonsymbolic ratio stimuli elicited robust spatial associations – classical SNARC effects.

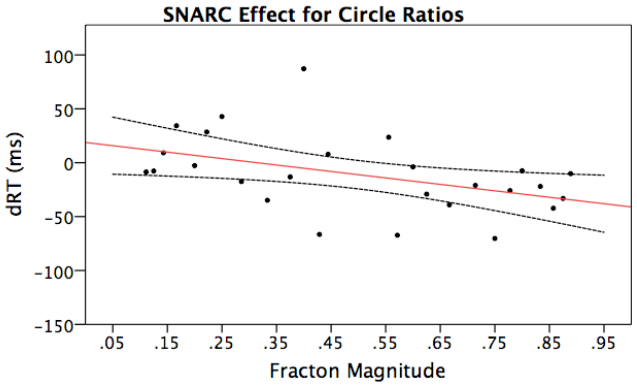


Figure 4: dRT as a function of fraction magnitudes for circle ratio stimuli.

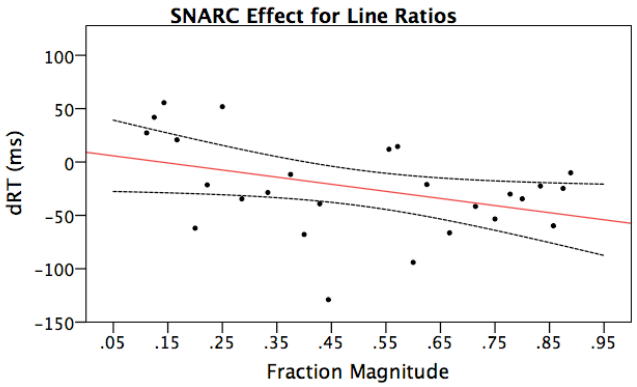


Figure 5: dRT as a function of fraction magnitudes for line ratio stimuli.

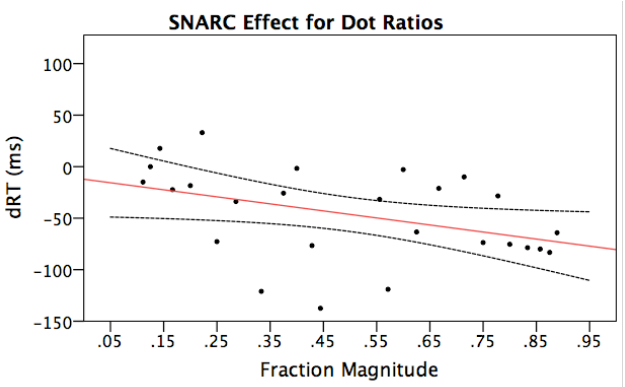


Figure 6: dRT as a function of fraction magnitudes for dot ratio stimuli.

Discussion

The current experiments demonstrated the existence of spatial associations with magnitudes instantiated using three different types of nonsymbolic ratios (i.e., dot arrays, line pairs, and circle pairs). These effects paralleled the classical SNARC, but differed in two important ways: First, the stimuli were nonsymbolic, instead of symbolic. Second, the stimuli were ratios, so the magnitudes were defined relationally as opposed to corresponding to the absolute magnitude of any individual component. Several aspects of the current experiment make important contributions to our understanding of human magnitude processing.

First, the current study provides further evidence for the existence of a ratio processing system and extends our understanding of its operation. Participants performed accurately on a perceptually based task comparing ratio

magnitudes, replicating recent work demonstrating human sensitivity to nonsymbolic ratio magnitudes (e.g., Jacob et al., 2012; Matthews & Chesney, 2015; Matthews et al., 2015; McCrink et al., 2013; Meert, Grégoire, Seron, & Noël, 2012).

This in itself represents a contribution to the basic science of human magnitude representation. The nonsymbolic aspects of the current experiment stand alongside other research demonstrating that the basic operation of the SNARC effect applies to nonsymbolic magnitudes. For instance, Holmes & Lourenco (2011) found similar spatial associations using the valence of facial expressions of emotion, and Rusconi, Kwan, Giordano, Umiltà, and Butterworth (2006) demonstrated spatial associations with musical pitch. All together these works seem consistent with speculation that the SNARC effect may be a more general association between quantities and space – a SQUARC.

Here we underscore the fact that the use of ratio magnitudes in this study represents a major departure from previous studies involving automatic associations of space and magnitude (be they symbolic or nonsymbolic). Previous SNARC/SQUARC studies have involved quantities with stand-alone magnitudes: the size of a number (Dehaene et al., 1993), the angriness of an expression (Holmes & Lourenco, 2011), or the "height" of a musical pitch (Rusconi et al., 2006). In the present case, the magnitudes involved emerged from the relations between two components, each of which had its own magnitude. That is, even though each circle in a ratio stimulus had its own magnitude, the effects found here reflected a magnitude *determined by the relation between the two* – one that can be expressed by multiple components of different sizes so long as the pairs maintain the same ratio value. The current findings suggest that this emergent relational magnitude is automatically mapped to space as well.

These results may also have implications for how we think about the human number sense from an applied perspective. Unlike research involving faces and musical pitch, the ratio stimuli used in the current experiments are fraction analogs. Each is a concrete instantiation of a fraction value, and prior work has shown that human adults can make the mapping between these nonsymbolic stimuli and symbolic numbers (Matthews & Chesney, 2015). Moreover, Matthews et al. (2015) have shown that college student's ability to compare nonsymbolic ratios predicts symbolic fraction and algebra skills. Thus, Lewis et al. (2015) have essentially argued that nonsymbolic ratios are protonumerical and that ratio processing abilities may be an early-emerging tool that provides intuitive access to fractions concepts. The existence of a SNARC for these fraction analogs adds to the potential of this speculative account.

In summary, the current study contributes another advancement in a recent string of studies shedding light on the impressive extent of human ratio processing abilities (Boyer & Levine, 2012; Jacob et al., 2012; Matthews & Chesney, 2015; Matthews et al., 2015; McCrink & Wynn, 2007; Möhring, Newcombe, Levine, & Frick, 2015). Each step foregrounding the perception of ratio magnitudes expands our understanding of the basic human perceptual apparatus. This in turn informs theory about the foundational inputs upon which concepts can be built.

Acknowledgments

This research was supported in part by NSF Grant DRL-1420211 to Percival Matthews and Edward Hubbard.

References

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology*, *113*(3), 447–455. http://doi.org/10.1016/j.jecp.2012.06.004

Boyer, T. W., & Levine, S. C. (2012). Child proportional scaling: Is 1/3 = 2/6 = 3/9 = 4/12? *Journal of Experimental Child Psychology*, *111*(3), 516–533. http://doi.org/10.1016/j.jecp.2011.11.001

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, *44*(5), 1478–1490. http://doi.org/10.1037/a0013110

Chesney, D. L., & Matthews, P. G. (2013). Knowledge on the line: Manipulating beliefs about the magnitudes of symbolic numbers affects the linearity of line estimation tasks. *Psychonomic Bulletin & Review*, *20*(6), 1146–1153. http://doi.org/10.3758/s13423-013-0446-8

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396. http://doi.org/10.1037/0096-3445.122.3.371

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 626.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. http://doi.org/10.1016/j.tics.2004.05.002

Fias, W., Brysbaert, M., Geypens, F., & d'Ydewalle, G. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical cognition*, *2*(1), 95-110.Hansen, N., Jordan, N. C., Siegler, R. S., Fernandez, E., Gersten, R., Fuchs, L., & Micklos, D. (2015). General and math-specific predictors of sixth-graders' knowledge of fractions. *Cognitive Development, 35*, 34-49. doi:10.1016/j.cogdev.2015.02.011

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: a computational model of the SNARC effect. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(1), 32.

Holmes, K. J., & Lourenco, S. F. (2011). Common spatial organization of number and emotional expression: A

mental magnitude line. *Brain and Cognition*, *77*(2), 315–323. http://doi.org/10.1016/j.bandc.2011.07.002

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, *6*(6), 435–448. http://doi.org/10.1038/nrn1684

Jacob, S. N., Vallentin, D., & Nieder, A. (2012). Relating magnitudes: the brain's code for proportions. *Trends in Cognitive Sciences*, *16*(3), 157–166. http://doi.org/10.1016/j.tics.2012.02.002

Kallai, A. Y., & Tzelgov, J. (2009). A generalized fraction: An entity smaller than one on the mental number line. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1845–1864. http://doi.org/10.1037/a0016892

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498–525.

Lewis, M.R., Matthews, P. G., & Hubbard, E. M. (2015). The non-symbolic foundations of fraction understanding. In D. B. Berch, D. C. Geary, & K. Mann Koepke (Eds.), *Development of mathematical cognition: Neural substrates and genetic influences*. San Diego, CA: Academic Press.

Matthews, P. G., & Chesney, D. L. (2015). Fractions as percepts? Exploring cross-format distance effects for fractional magnitudes. *Cognitive Psychology*, *78*, 28–56. http://doi.org/10.1016/j.cogpsych.2015.01.006

Matthews, P. G., Lewis, M. R., & Hubbard, E. M. (2015). Individual Differences in Nonsymbolic Ratio Processing Predict Symbolic Math Performance. *Psychological Science*, 0956797615617799. http://doi.org/10.1177/0956797615617799

McCrink, K., Spelke, E. S., Dehaene, S., & Pica, P. (2013). Non-symbolic halving in an Amazonian indigene group. *Developmental Science*, *16*(3), 451–462. http://doi.org/10.1111/desc.12037

McCrink, K., & Wynn, K. (2007). Ratio Abstraction by 6-Month-Old Infants. *Psychological Science*, *18*(8), 740–745. http://doi.org/10.1111/j.1467-9280.2007.01969.x

Meert, G., Grégoire, J., Seron, X., & Noël, M.-P. (2012). The mental representation of the magnitude of symbolic and nonsymbolic ratios in adults. *The Quarterly Journal of Experimental Psychology*, *65*(4), 702–724. http://doi.org/10.1080/17470218.2011.632485

Möhring, W., Newcombe, N. S., Levine, S. C., & Frick, A. (2015). Spatial Proportional Reasoning is Associated with Formal Knowledge about Fractions. *Journal of Cognition and Development*. http://doi.org/10.1080/15248372.2014.996289

Moyer, R. S., & Landauer, T. K. (1967). Time required for Judgements of Numerical Inequality. *, Published Online: 30 September 1967; | doi:10.1038/2151519a0*, *215*(5109), 1519–1520. http://doi.org/10.1038/2151519a0

Newcombe, N. S., Levine, S. C., & Mix, K. S. (2015). Thinking about quantity: the intertwined development of spatial and numerical cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(6), 491–505. http://doi.org/10.1002/wcs.1369

Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *Annual Review of Neuroscience*, *32*, 185–208. http://doi.org/10.1146/annurev.neuro.051508.135550

Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Sciences*, *14*(12), 542–551. http://doi.org/10.1016/j.tics.2010.09.008

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition*, *99*(2), 113–129. http://doi.org/10.1016/j.cognition.2005.01.004

Santens, S., & Gevers, W. (2008). The SNARC effect does not imply a mental number line. *Cognition*, *108*(1), 263-270.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., … Chen, M. (2012). Early Predictors of High School Mathematics Achievement. *Psychological Science*, *23*(7), 691–697. http://doi.org/10.1177/0956797612440101

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181. http://doi.org/10.1037/h0046162

Toomarian, E.Y. & Hubbard, E.M. Stimulus Characteristics and Strategic Variability Modulate Spatial Representations of Fractions. Manuscript submitted for publication.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. http://doi.org/10.1016/j.tics.2003.09.002

Emotional influences on time perception

Anna K. Trapp (anna.k.trapp@tu-berlin.de)
Chair of Cognitive Psychology and Cognitive Ergonomics, MAR 3-2, Marchstraße 23
10587 Berlin, Germany

Manfred Thüring (manfred.thuering@ tu-berlin.de)
Chair of Cognitive Psychology and Cognitive Ergonomics, MAR 3-2, Marchstraße 23
10587 Berlin, Germany

Abstract

In studies on prospective time perception, a prolonging effect of arousal on time estimates is commonly reported for durations under 2s while the effect vanishes for longer intervals. In this study, we investigated how arousal and pleasure induced by aural stimuli varying in volume and valence influenced reproductions in the range from 1.1s to 5s. As expected, higher arousal was associated with higher estimates for 1.1s durations. However, this effect was also found for 3.8s durations. An additional analysis with linear mixed models revealed an interaction between volume manipulation and subjective ratings regarding arousal and pleasure. Based on these results we propose that subjective experience of the emotional quality of stimuli might be interesting for further research on prospective time perception. Moreover, the results showed that not only within subject variation should be statistically controlled when analyzing such data. Instead, statistical models should also include parameters controlling for stimulus material.

Keywords: prospective time perception; reproduction; emotion; arousal; valence; linear mixed models

Introduction

The prolonging effect of emotions on prospective temporal duration judgments has been subject to a great number of research projects and publications in the last decade. Research on this phenomenon differs regarding methodology and the considered duration. Examples for the variety of methods to induce an emotional state are emotional pictures (Gil & Droit-Volet, 2012), music (Droit-Volet, Ramos, Bueno & Bigand, 2013), emotional faces (Gil & Droit-Volet, 2011), emotional sounds (Mella, Conty & Pouthas, 2011) and bodily expressions (Droit-Volet & Gil, 2015). Another variation in methodology relates to differing timing tasks, such as the bisection task (Droit-Volet, Brunot & Niedenthal, 2004), verbal estimates (Gil & Droit-Volet, 2012), and production (Gil & Droit-Volet, 2011) as well as reproduction tasks (Angrilli, Cherubini, Pavese & Manfredini, 1997). With respect to duration ranges, most of these studies focused on intervals between 400ms and 1600ms, while only a few experiments investigated longer durations up to 6000ms.

Gil and Droit-Volet (2012) asked participants to verbally estimate for how long different emotional pictures had been presented. The durations ranged between 50ms and 1600ms. The presented pictures systematically varied with respect to the arousal level they caused (high vs. low) as well as with respect to the discrete emotion they evoked (disgust, fear, sadness, or none for neutral pictures). Results showed higher estimates for emotional pictures compared to neutral ones and indicated that this effect gained in magnitude with increasing arousal (Gil & Droit-Volet, 2012). The same effect was found in all other studies reported above for durations smaller than two seconds.

These results accord with the clock speed hypothesis (c.f. attentional-gate-model, Block & Zakay, 1996) which assumes that prospective timing relies on accumulated pulses generated by an internal clock. The clock accelerates when arousal increases. This leads to a higher count of pulses compared to unchanged or decreased arousal. More pulses cause a prolonged time perception and consequently lead to higher estimates. Thus, the finding that a raising of arousal leads to *longer* estimates can be explained by an increased number of pulses due to an accelerated clock.

However, this effect seems to change for longer intervals. Noulhiane, Mella, Samson, Ragot and Pouthas (2007) studied the impact of different arousal levels (low vs. high) on time estimates for 2s-durations and found that stimuli evoking high arousal led to *shorter* estimates than those evoking low arousal. This finding was consistent over emotionally negative and positive stimuli and was found for verbal timing tasks as well as for reproductions. It clearly contradicts the clock speed hypothesis and deviates from all other studies reported above.

Noulhiane et al.'s surprising result (2007) raises the question why the arousal effect turns into the opposite direction. One explanation could be that subjects start to cognitively process emotions when confronted with an emotion induction long enough, as for two seconds (Noulhiane et al., 2007). Such a processing would bind attentional resources by reducing the attention on the timing task. Less attention on the timing task means that more pulses of the clock are missed, leading to a lower count of pulses and a shorter estimate. This explanation is in line with literature on the shortening effect of distraction on prospective timing (Brown, 2008). Another explanation could be that the reversed effect is modality specific. Noulhiane et al. (2007) used aural stimuli in their experiment. Angrilli et al. (1997) induced emotions via visual stimuli and report results consistent with the clock speed hypothesis. More precisely, they report longer estimates for high arousing stimuli than for low arousing stimuli at two seconds durations.

To sum up, evidence for the prolonging effect of arousal on time estimates of durations smaller than two seconds is very strong and supports the clock speed hypothesis. Looking at durations lasting for two seconds, however, the influence of arousal is unclear. Starting from this summary, it is obvious that the arousal effect on time perception during intervals greater than or equal to two seconds requires further inspection. Both, Angrilli et al. (1997) as well as Noulhiane et al. (2007), reported no such an effect for four seconds and six seconds durations. In their experiments, time estimates did not significantly differ between arousal levels for these intervals. Nather, Bueno, Bigand and Droit-Volet (2011) induced arousal via pictures of body postures. Even though they found the prolonging arousal effect at durations ranging from 400ms to 1600ms, they did not detect such an effect at durations ranging from two to eight seconds. These studies give support to the assumption that the lengthening effect of arousal vanishes at durations longer than two seconds.

Up to this point, only the effect of arousal on time estimates has been discussed. Another important dimension of emotions is valence and many of the studies described above did not only vary arousal but valence as well. Noulhiane et al. (2007), for example, compared low-arousing emotional with low-arousing neutral sounds and reported longer estimates for emotional compared to neutral stimuli for the 2s-duration. Likewise, Gil and Droit-Volet (2012) compared emotional stimuli to neutral ones and found the same pattern. Similar results were reported by Droit-Volet et al. (2004) as well as by Gil and Droit-Volet (2011).

Summing up the results, it is obvious that a comprehensive investigation of emotional influences on time perception ought to address both, arousal as well as valence for a larger area of durations. This leads to the following research question: How do arousal and valence affect time perception of durations ranging from under to over two seconds? To answer this question, we conducted an experiment in which valence and arousal were varied for durations between 1.1s and 5.0s. This range includes durations for which an arousal effect is commonly found, for which results are ambiguous and for which the arousal effect is expected to vanish. With respect to stimuli, we decided to focus on the same modality as Noulhiane et al. (2007). The use of aural stimuli allows a comparison with their rather uncommon findings and might help to decide between the two explanations discussed earlier.

Methodological notion on Linear Mixed Models

Another reason for the ambiguous state of affairs concerning the relationship between emotions and time perception may be methodological in nature. In this field of research, mostly ANOVAs are used to analyze data. To this account, most studies cited above used stimuli that were rated for arousal and valence beforehand and averaged the dependent variable over trials for certain groups of stimuli. However, even though all stimuli from one group are similar to each other on the predefined dimension, slight differences can cause systematic variation in the data. If systematic by-item variation is not statistically accounted for, error variance increases and thus makes type II errors more likely.

Linear mixed models (LMM) offer a solution to this problem. They are called 'mixed models' because they can include fixed effects, like factors or covariates, as well as random effects for subjects and items (Winter, 2013). Random effects account for variance between subjects and between items in three ways: by including random intercepts for both subjects and items, by including random slopes considering variance between subjects respectively items for all main effects and interactions and by considering correlations between intercepts and slopes (Bates, Kliegl, Vasishth & Baayen, 2015). All this can be done in one statistical model which gives an advantage over a single within ANOVA because the ANOVA only allows to consider either variance of subjects or of items. Thus, LMMs can help to reduce error variance (Winter, 2013) and therefore increase statistical power (Kliegl, Wei, Dambacher, Yan & Zhou, 2011). Moreover, LMMs make it possible to analyze the complete data set, instead of averaging over certain dimensions. This helps to improve the understanding of complex structures within the data as for example by identifying important covariates (Kliegl et al., 2011). Furthermore, they allow the inclusion of covariates or predictors that vary within subjects thus offering another advantage over an ANOVA.

This advantage is especially important when looking at time perception because its analysis may be strongly influenced by inter-individual differences. For example, time perception changes with age (Block, Zakay & Hancock, 1999) and studies suggest that it differs between gender (Block, Hancock & Zakay, 2000). Therefore, including statistical parameters for variation between participants may improve the understanding of important participant characteristics (Kliegl et al., 2011) and help to detect effects that were otherwise overlooked.

Method

Participants

A study was conducted with N=20 participants (11 male, 9 female). All participants were young adults (M_{age}= 25.1, SD_{age}=3.3) comparable to the studies cited above.

Design, material and procedure

The study was based on a 2x3x4 within-subjects design consisting of the factors valence (negative, neutral), volume (low, medium, loud) and duration (1.1s, 2.4s, 3.8s, 5.0s). To implement the factor valence, twelve stimuli were taken from the International Affective Digital Sounds System (Lang, Bradley & Cuthbert, 2008). Six of them were categorized as neutral and six as emotionally negative according to the norm by Lang et al. (2008). All sounds showed a continuous pattern of noise. For the factor volume,

the levels were low, medium and loud. The difference between these levels was established by amplifying the original sound from the IADS by the factors 0.04, 0.4 and 4 respectively. The mean volumes were the following: M_{low}=41dBA (SD=2.9), M_{medium}=61dBA (SD=2.5), M_{loud}=70dBA (SD=2.3). Arousal induction via volume was chosen because noise is known to stimulate the central nervous system and to increase arousal level (Hockey, 1972). Moreover, this variation allows the use of the same sounds to elicit different levels of arousal. Under each combination of valence, volume and duration, six sounds were presented, resulting in 144 stimuli. In addition to these, 36 trials with random durations between 0.9s and 5.3s were introduced to increase the variability of the durations. Altogether, this adds up to a total of 180 trials.

The experiment consisted of two successive phases, a rating and a test phase. In the rating phase, participants heard each stimulus for 6s and filled in the self-assessment manikin SAM (Bradley & Lang, 1994) to judge its emotional effect in terms of pleasure and arousal. At the beginning of the test phase, the forthcoming trials and the reproduction procedure were explained. Participants were instructed not to count while perceiving or reproducing the durations. After the instruction, they completed three practice trials with a neutral sound that was not used otherwise in the experiment. Each trial was started by the participants themselves by pressing the space key. After a 2s delay, the aural stimulus was presented, followed by a request to reproduce the duration. The reproduction was accomplished by a continuous key press without any sound. This response type was chosen because continuous key presses show less variability than responses that indicate only the end or both start and end of an interval (Mioni, Stablum, McClintock, & Grondin, 2014). After each trial, participants were asked to report whether they had counted (by pressing key 1) or not (key 3). They started the next trial by pressing space.

The presentation of stimuli and instructions as well as the recording of data was implemented in Matlab 2014b and the Psychophysics Toolbox extensions version 3 (Brainard, 1997; Pelli, 1997). All visual information (e.g. instructions, requests) were given on a 22-inch desktop display. Sounds were presented over headphones, while a fixations cross was shown on the display. Pupil dilation was measured but its data is not reported in this paper due to limited space.

Data analysis

Prior to the analysis, all trials with random duration as well as all trials in which participants had counted were eliminated from the data set. For the remaining 2839 trials, perceived time ratios (PRTs) were computed by dividing the duration of the reproduction by the duration of the presented interval (Block, Hancock & Zakay, 2010). PTRs were preferred over the original data of the reproductions because they represent a measure for goodness of estimation. A PTR of one represents a perfect estimation independently from the presented duration, while a PTR smaller respectively bigger than one indicates underestimation respectively overestimation. PTRs allow for comparing duration estimates regarding the strength of their over- or underestimation which can vary between different arousal levels or even different durations.

The data was analyzed with a repeated measurements ANOVA and with a Linear Mixed Model (LMM). We choose to use both methods to find out if the LMM provides more insights than the ANOVA. Data analysis was conducted with R (R Core Team, 2015) using the ez (Lawrence, 2013), lme4 (Bates, Maechler, Bolker & Walker, 2015) and ggplot (Wickham, 2009) packages. In accordance with common practice for ANOVAs, all data was averaged over sounds with negative valence and over sounds with neutral valence for each duration and volume level resulting in 24 values per subject. Additionally, a manipulation check for the segmentation of the 12 sounds into neutral and negative valence groups was conducted. To this account, t-tests were conducted with the subjective ratings from the rating phase regarding pleasure and arousal averaged over the neutral and over the negative sounds. As expected, the negative group of sounds was rated as inducing less pleasure than the neutral sounds, $t(19)=-10.8$, $p<.001$, $r=.93$. Regarding arousal, the negative group of sounds was rated as inducing more arousal than the neutral sounds, $t(19)=4.7$, $p<.001$, $r=.73$. Hence, both pleasure and arousal were different for the two valence conditions. Moreover, mean correlation coefficients computed via z transformation showed high correlations between arousal and valence for negative sounds, $r= -.67$, and medium correlation for neutral sounds, $r= -.35$.

Results

Using the averaged data, an ANOVA Type III was conducted with the factors duration, volume and valence (4x3x2). The ANOVA revealed significant main effects for volume, $F(1, 38)=20.8$, $p<.001$, $ges=.025$, duration, $F(1.16, 21.99)=54.2$, $p<.001$, $ges=.286$, and valence, $F(1, 19)=15.3$, $p<.001$, $ges=.008$.

Regarding the main effect of volume, low volume sounds had lower PTRs (M_{low}=0.83) compared to sounds with medium volume (M_{medium}=0.88), t[1](19)=-4.1, $p<.01$, $r=.68$. Moreover, PTRs of the latter were lower compared to those of loud sounds (M_{loud}=0.90), $t(19)=-2.2$, $p<.05$, $r=.44$. A closer inspection of the main effect of duration revealed that sounds played for 1.1s showed higher PTRs values ($M_{1.1s}$=1.04) compared to sounds played for 2.4s ($M_{2.4s}$=0.92), $t(19)=5.3$, $p<.001$, $r=.77$. Moreover, PTRs of the latter were higher than PTRs of sounds that lasted 3.8s ($M_{3.8s}$=0.80), $t(19)=7.5$, $p<.001$, $r=.83$. PTRs for the duration of 3.8s compared to those with a duration of 5.0s were also significantly higher ($M_{5.0s}$=0.73), $t(19)=7.3$, $p<.001$, $r=.86$. As the PTRs show, all durations were underestimated. Increasing duration led to stronger

[1] p-values for all post-hoc t-tests were adjusted using Bonferroni-Holm correction

underestimations, while increasing volume led to weaker underestimations. Negative sounds ($M_{negative}$=0.89) led to weaker underestimations compared to neutral sounds ($M_{neutral}$=0.86).

The ANOVA also showed a significant interaction between volume and interval, $F(2.64, 50.21)=4.9$, $p<.01$, $ges=.012$, which is visualized in Figure 1. Post-hoc t-test revealed significant differences between low and medium sounds for 3.8s, $t(19)=-4.0$, $p<.01$, $r=.68$ but not for other durations, $ts(19)>-2.4$, *n.s.*. Furthermore, they showed significant differences between low and loud sounds at 1.1s, $t(19)=-5.9$, $p<.001$, $r=.80$ and at 3.8s, $t(19)=-5.0$, $p<.001$, $r=.75$. For none of the durations, a significant difference between medium and loud sounds was found, $ts(19)>-2.6$, *n.s.*. No other interaction effect of the ANOVA proved to be significant, Fs<1, ps>.05.

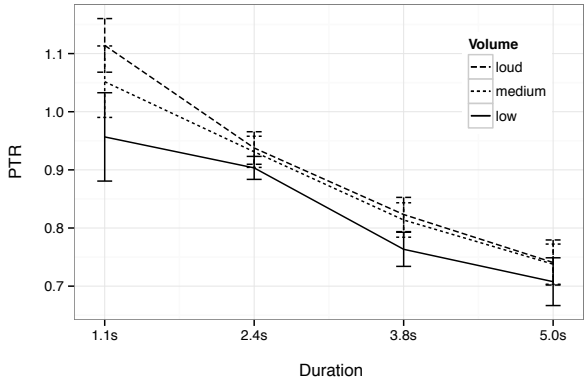


Figure 1: Interaction of the factors volume and duration. Error bars represent confidence interval computed according to Cousineau (2005).

In addition to the ANOVA, a LMM with the fixed effects duration and volume (4x3, centered variables) was fitted by restricted maximum likelihood (REML). The factor valence was replaced by two continuous predictors. These were the subjective ratings on arousal and pleasure for each sound given by each participant during the rating phase. All possible interaction terms between the four fixed effects were included. As random effect, both random intercepts for subjects and sounds as well as all random slopes for the two factors over subjects and sounds were comprised. The analysis started with the following model:

(1) PTR ~ (Duration+Volume+Pleasure+Arousal)^4 + (1+Duration+Volume+Duration:Volume|Subjects) + (1+Duration+Volume+Duration:Volume|Sounds)

To avoid over-parameterization unnecessary random effects were eliminated following the procedure suggested by Bates et al. (2015). Random effects in the final model are presented in formula (2). Random effects by subjects revealed that subjects not only differed regarding intercepts, but also with respect to the effect that duration had on PTR values. In other words, the strength of the duration effect differed between subjects. The random intercept of sound showed that the intercept changed between the 12 sounds. Furthermore, the effect of duration and volume changed in magnitude between the different sounds.

(2) (1+Duration|ID) + (1+Volume+Duration|Sound)

Regarding fixed effects, a stepwise backwards selection was conducted. During each step, a new model was fitted, in which one of the model terms was eliminated. Each new model was tested against the former model by a likelihood ratio test with an alpha criterion of .05. The remaining fixed effects and their output statistics are listed in Table 1. P-values were calculated using the lmerTest package (Kuznetsova, Brockhoff & Christensen, 2015).

Table 1: Fixed effects of model (2).

Fixed Effects	estimate	*SD*	*t*	*p*
Intercept	0.880	0.051	17.27	<.001
Duration	-0.081	0.020	-4.03	<.001
Volume	0.003	0.036	0.07	n.s.
Pleasure	0.001	0.005	0.13	n.s.
Arousal	-0.002	0.005	-0.48	n.s.
Duration:Vol.	-0.018	0.004	-4.10	<.001
Duration:Ple.	-0.005	0.002	-2.34	<.05
Volume:Ple.	0.006	0.005	1.16	n.s.
Volume:Aro.	0.008	0.005	0.16	n.s.
Pleasure:Aro.	-0.000	0.001	-0.01	n.s.
Vol.:Ple.:Aro.	-0.002	0.001	-2.01	<.05

Compared to the ANOVA the LMM showed the previously described effects of duration and of the volume-duration interaction. However, the LMM indicated that the effect of the arousal induction via volume on PTRs was dependent on the emotional qualities of the stimuli, namely arousal and pleasure. Moreover, the duration x pleasure interaction indicated that PTRs were higher for low pleasure stimuli compared to high pleasure stimuli, while this effect was stronger for longer intervals.

Discussion

In this study, we investigated the effect of arousal and valence on time perception regarding durations ranging from 1.1s to 5.0s. For this purpose, 12 sounds classified as either unpleasant or neutral were played at different volume levels and participants reproduced the duration of the presented sound. Data analysis was carried out on standardized estimates (PTRs) using both an ANOVA and a linear mixed model (LMM). Results revealed that underestimations grew stronger with increasing durations. While the ANOVA showed both an effect of volume as well as of valence, the LMM indicated that the relation between emotions and duration estimations are more complicated than a simple main effect of arousal or of valence.

The effect of volume as shown in the ANOVA and the interaction between volume and duration in both analyses

can be regarded as further support for the prolonging effect of arousal on time perception. Here, estimates increased with the volume of the presented sound. Literature on noise and its effect on arousal level indicates that high volume noise increases arousal (e.g. Hockey, 1972). These findings support our interpretation that the louder sounds elicited higher arousal which in turn caused an increase of time estimates. This arousal effect lends further support to the clock speed hypothesis as predicted by the attentional-gate-model (Block & Zakay, 1996).

However, the prolonging effect of volume differed within the duration range. At 1.1s low and loud volumes lead to significantly distinct estimates. This is not surprising because it is in line with the commonly found arousal effect for durations smaller than two seconds. In contrast to other studies (Angrilli et al., 1997; Noulhiane et al., 2007; Nather et al., 2011), however, the effect did not vanish for duration longer than two seconds. Even though the found prolonging effect did not prove as significant for 2.4s durations, a significant difference in estimates was found at 3.8s concerning the low and medium as well as the the low and loud volume. Before declaring this finding as further support for the clock speed hypothesis though, one point has to be made. The clock speed hypothesis not only predicts higher estimates when arousal is increased, it also predicts that this effect should gain in magnitude when durations grow longer. This prediction is based on the mode of action belonging to the clock. If durations are longer, the clock has more time to produce pulses. Hence, increased clock speed will not lead to an additive increase of pulses but to a multiplicative increase (Gil & Droit-Volet, 2012; Nather et al., 2011). However, when comparing effect sizes of the difference between low and loud sounds between 1.1s and 3.8s durations, the results of this study provide no evidence for an increased arousal effect. Thus, even though the arousal effect did not vanish as described by other authors, it does not show a multiplicative increase either.

Comparing our results to those by Noulhiane et al. (2007), we could not replicate the reversal of the arousal effect for two second durations. Moreover, our results do not support the assumption of a modality specific effect because they are consistent with results gained by applying visual instead of aural stimuli (Angrilli et al., 1997). Instead, the arousal effect reported in this study as well as the vanishing arousal effect described by other authors for durations longer than two seconds, might be connected to a change in cognitive processing for durations exceeding two seconds. More precisely, attentional processes might come into play for durations greater than 2 seconds. Thus, the interplay of arousal and attention as described in the attentional-gate-model (Block & Zakay, 1996) might become particularly important for longer intervals.

Regarding the effect of valence shown in the ANOVA, estimates for negative stimuli were higher than those for neutral stimuli. Even though it was only a small effect, it is in line with earlier findings (Droit-Volet et al., 2004; Gil & Droit-Volet, 2011; Gil & Droit-Volet, 2012; Noulhiane et al., 2007). Nonetheless, as shown in the manipulation check for the segmentation of stimuli into negative and neutral, both groups did not only differ regarding pleasure but also regarding arousal. Thus, the difference in PTRs between groups cannot be distinctly associated with either either dimension. In the analysis using LMM, the factor valence was therefore replaced by arousal and pleasure ratings given by each subject for each sound. The analysis revealed that PTRs were affected by an interaction of the subjective ratings. Furthermore, the interaction effect was dependent on volume, and thus on the level of arousal induction. This three-way interaction cannot be easily interpreted for two reason: First, the two rating dimensions covary and second, both ratings were measured on a nine-point scale making post-hoc comparisons of means or slopes unusable.

However, we derive two hypotheses from the three-way interaction that should be tested further in future studies. First, people differ in their subjective experience of the emotional quality of stimuli and their experience has a distinct impact on time perception. Whether this experience can be measured by ratings as done here, or if other measures (e.g. physiological data) are needed, remains open. Second, there is a close connection between arousal and pleasure and their impact on time perception. Previous studies have already looked beyond the scope of the clock speed hypothesis by including valence as an impact factor on time perception. However, we want to emphasize the importance to look at the effect of subjectively experienced emotional reactions to the stimuli rather than at the effect of predefined categories of arousal levels and valence. Such an approach seems more appropriate for the interplay of arousal and valence. An alternative operationalization might be a measure for emotional intensity, a concept that has been discussed by Reisenzein (1994) and Scherer (2005).

Summing up, the reported results of the ANOVA are in line with a prolonging arousal effect on time. Moreover, the arousal induction method by volume changes seems to be comparable to other induction methods. However, the arousal effect did not vanish for durations longer than 2s, but it did not gain in magnitude as predicted by the clock speed hypothesis either. Thus, in line with other studies (e.g. Noulhiane et al., 2007), our results suggest a change in cognitive processing for durations exceeding two seconds. When including subjective predictors with help of a LMM it becomes clear that the effect of arousal and the effect of valence on time perception are interconnected. Focusing research on the effect of their interplay might enhance the understanding of the cognitive processes involved in time perception. However, such research should include also positive stimuli and not only neutral and negative ones as done here. Reflecting the data analysis strategy, LMMs are a handy tool to include continuous predictors with within-subject variation in the statistical model. Moreover, the random effect structure of the LMM showed that it is reasonable to not only consider by-subject variations as it is done in the ANOVA, but also to consider by-item variations.

Acknowledgments

We would like to thank Tomislav Vidos and Nikolaus Rötting for their work during implementing the experiment as well as collecting data. We would also like to thank Hans A. Trukenbrod and Daniel Schad for statistical support.

References

Angrilli, A., Cherubini, P., Pavese, A., & Manfredini, S. (1997). The influence of affective factors on time perception. *Perception & Psychophysics*, *59*(6), 972–982.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *arXiv:1506.04967 [stat]*. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using eigen and S4*. R package version 1.1-8. http://CRAN.R-project.org/package=lme4.

Block, R. A., Hancock, P. A., & Zakay, D. (2000). Sex differences in duration judgments: a meta-analytic review. *Memory & Cognition*, *28*(8), 1333–1346.

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: a meta-analytic review. *Acta Psychologica*, *134*(3), 330–343.

Block, R. A., & Zakay, D. (1996). Models of psychological time revisited. In H. Helfrich (Ed.), *Time and mind* Hogrefe & Huber, Kirkland.

Block, R. A., Zakay, D., & Hancock, P. A. (1999). Developmental changes in human duration judgments: a meta-analytic review. *Developmental Review*, 183–211.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8*. University of Florida, Gainesville, FL.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 443-446.

Brown, S. (2008). Time and attention: review of the literature. In S. Grondin (Ed.), *Psychology of time*. Bingley, UK: Emerald Group Publishing.

Cousineau, D. (2005). Confidence intervals in within-subjects designs: a simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42-45.

Droit-Volet, S., Brunot, S., & Niedenthal, P. M. (2004). Perception of the duration of emotional events. *Cognition & Emotion*, *18*(6), 849–858.

Droit-Volet, S., & Gil, S. (2009). The time-emotion paradox. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1943–1953.

Droit-Volet, S., & Gil, S. (2015). The emotional body and time perception. *Cognition & Emotion*, 1–13.

Droit-Volet, S., Ramos, D., Bueno, L. J., & Bigand, E. (2013). Music, emotion, and time perception: the influence of subjective emotional valence and arousal? *Emotion Science*, *4*, 417.

Gil, S., & Droit-Volet, S. (2011). "Time flies in the presence of angry faces"… depending on the temporal task used! *Acta Psychologica*, *136*(3), 354–362.

Gil, S., & Droit-Volet, S. (2012). Emotional time distortions: the fundamental role of arousal. *Cognition & Emotion*, *26*(5), 847–862.

Hockey, G. R. J. (1972). Effects of noise on human efficiency and some individual differences. *Journal of Sound and Vibration*, *20*(3), 299–304.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. *Quantitative Psychology and Measurement*, *1*, 238.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B., (2015). *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-29. http://CRAN.R-project.org/package=lmerTest

Lawrence, M. A. (2013). ez: Easy analysis and visualization of factorial experiments.. R package version 4.2-2. http://CRAN.R-project.org/package=ez

Mella, N., Conty, L., & Pouthas, V. (2011). The role of physiological arousal in time perception: psychophysiological evidence from an emotion regulation paradigm. *Brain and Cognition*, *75*(2), 182–187.

Mioni, G., Stablum, F., McClintock, S. M., & Grondin, S. (2014). Different methods for reproducing time, different results. *Attention, Perception & Psychophysics*, *76*(3), 675–681.

Nather, F. C., Bueno, J. L. O., Bigand, E., & Droit-Volet, S. (2011). Time changes with the embodiment of another's body posture. *PLoS ONE*, *6*(5), e19818.

Noulhiane, M., Mella, N., Samson, S., Ragot, R., & Pouthas, V. (2007). How emotional auditory stimuli modulate time perception. *Emotion*, *7*(4), 697–704.

Pelli, D.G. (1997) The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437-442.

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and ocial Psychology, 67*(3)*,* 525-539.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*(4), 693-727.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York, NY: Springer.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv:1308.5499*. Retrieved from http://arxiv.org/pdf/1308.5499.

Stable Causal Relationships are Better Causal Relationships

Nadya Vasilyeva (Vasilyeva@Berkeley.Edu)*,
Thomas Blanchard (Tblanchard@Berkeley.Edu)*,
Tania Lombrozo (Lombrozo@Berkeley.Edu)
Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA

*These two authors contributed equally to this work

Abstract

We report two experiments investigating whether people's judgments about causal relationships are sensitive to the robustness or *stability* of such relationships across a wide range of background circumstances. We demonstrate that people prefer stable causal relationships even when overall causal strength is held constant, and we show that this effect is unlikely to be driven by a causal generalization's actual scope of application. This documents a previously unacknowledged factor that shapes people's causal reasoning.

Keywords: stability, causality, explanation, background conditions, moderating variables

Consider the relationship between being sexually active and developing deep vein thrombosis. Is the former a *cause* of the latter?

In fact, sexual activity increases the probability that a woman will become pregnant, which in turn increases the probability that she will develop deep vein thrombosis. Thus, there is a sense in which being sexually active is causally relevant to whether one gets thrombosis. Yet this causal relationship seems different from that between, say, sexual activity and contracting herpes. When explaining why someone contracted herpes (but not why a woman has deep vein thrombosis), we're likely to cite sexual activity. Similarly, medical websites list herpes as a sexually transmitted disease, but do not list deep vein thrombosis as a risk associated with sexual activity. Why might this be?

The asymmetry between sex and thrombosis versus sex and herpes doesn't seem due to a difference in the strength of association between the relevant factors. Few people suffer from thrombosis as a result of sexual activity; but likewise, herpes is only contracted after sexual activity in some cases (namely, when both the partner is infected and transmission occurs). And while thrombosis can be caused by many other factors besides sexual activity, herpes can also be transmitted by non-sexual forms of physical contact.

Instead, these causal associations could differ in their robustness or *stability*. While sex can elevate the risk of thrombosis under very specific conditions (most notably, when the person is a woman who becomes pregnant), there are also plenty of circumstances under which there is simply *no* causal relationship between being sexually active and deep vein thrombosis, e.g., if the person is male, or sterile, or on blood thinners, and so on. By contrast, the causal relationship between sex and herpes is more robust insofar as under most circumstances – whatever your gender, your diet, your age, etc. – sexual activity puts you at greater risk of contracting herpes.

The extent to which a causal relationship is stable is not adequately captured by measures of causal strength that have dominated research on causal inference, such as ΔP (Allan, 1980) and power-PC (Cheng, 1997). The reason is that these measures track the *average* strength of a causal relationship in a population, whereas stability has to do with the extent to which the relationship holds *across* diverse segments of the population (or across various circumstances). True, very unstable causal relationships, such as the relationship between sexual activity and thrombosis, also tend to have low average strength. But two causal relationships can be equally strong on average and yet not be equally stable. By treating both relationships on a par, standard measures of causal strength ignore an important difference between them. For instance, these measures do not allow us to capture the fact that more stable relationships provide more far-reaching and reliable means for controlling their effects.

Stability is a well-known notion in the philosophy of science, where it has been introduced and discussed most extensively by Woodward (2006, 2010). In Woodward's framework, one starts with a very undemanding notion of causal relevance, on which X is causally relevant to Y just in case X causally influences Y in at least *some* circumstance. Stability is then defined as the extent to which the causal relationship $X \rightarrow Y$ holds in a variety of background circumstances. (One can think of background circumstances as circumstances not included in X and Y.) If $X \rightarrow Y$ holds in a wide variety of background circumstances – in particular, circumstances that we regard as 'normal' or 'important' – then it is relatively stable. Woodward argues convincingly that stability considerations play an important role in scientific practice, especially in selecting appropriate levels of causal representation and explanation.

There is also some indirect evidence for the role of stability in people's intuitive causal and explanatory judgments. Lombrozo (2010) found that people are more willing to consider relationships causal when an association involves a direct physical connection rather than double prevention, and – when double prevention is involved – are more inclined to regard an agent as a cause of an outcome when the action was intentional (vs. accidental). She argues that both effects could be due to a difference in the stability of the relevant relationships. Likewise, there is evidence that people are less inclined to regard an agent as a cause of a bad outcome when a third-party intentionally controlled the agent (Phillips & Shaw, 2015; Murray & Lombrozo, 2016). A possible explanation suggested by Murray and Lombrozo is that the dependence of the outcome on the agent is very

sensitive to the third-party's intentions, and in that respect fairly unstable. However, so far few direct investigations of the role of stability in lay causal judgments have been attempted (but see Gerstenberg et al., 2012), and none that appropriately control for such relevant features as number of intermediate causes and causal strength. In particular, to show that stability has an effect over and above causal strength, it is essential to consider cases where stability varies while causal strength (e.g., ΔP) is held fixed.

We conducted two experiments to investigate whether people are sensitive to stability considerations. Participants were presented with evidence suggesting either that a causal relationship holds in only one out of two kinds of circumstances, or that it holds in both kinds of circumstances. The causal strength of the relationship (for the full set of cases) was held fixed across the two conditions. In philosophy, the notion of stability has been applied to causal relations both between *types* (Woodward, 2010) and between *token* events (Woodward, 2006), and is held to be important both for *causal* and *explanatory* judgments (Woodward, 2010). To test for these different aspects and roles of stability, we asked participants to rate either causal or explanatory judgments at either the type or token level. (This also allowed us to ensure that any observed effect wasn't merely due to idiosyncrasies in the formulation of particular questions.) If people's causal and explanatory judgments are sensitive to stability considerations, this should be reflected in a lower willingness to say that *C* causes or explains *E* when the relationship holds only in one possible circumstance.

Experiment 1

The main goal of Experiment 1 was to examine the effect of stability on judgments of causal relationships when the causal strength of the relationships is held constant. To do so, we presented participants with evidence suggesting that a factor *C* has a causal influence on an effect *E* in a certain population. We further specified that some members of the population had a certain property *D* (e.g., a behavioral or environmental characteristic) that other members of the population lacked. Participants were assigned to one of two conditions. In the *non-moderated* condition, participants were presented with further evidence suggesting that *C* has a causal influence on *E* both when *D* is present and when it is absent. In the *moderated* condition, by contrast, the evidence suggested that *C* causes *E only* when *D* is present (i.e., in the presence of the *enabling circumstance*). The causal strength (ΔP or power PC) of C → E in the overall population was the same in both conditions, but its stability varied. The relationship was stable with respect to the *moderator variable* (presence or absence of *D*) in the non-moderated condition, but unstable with respect to this variable in the moderated condition.

Participants were asked to rate statements about the relationship between *C* and *E* in the overall population. As the causal strength of the relation was the same in both conditions, an effect of stability on causal and explanatory judgments should manifest itself in higher ratings in the stable (*non-moderated*) than in the unstable (*moderated*) condition.

Method

Participants One-hundred-eighty-two participants were recruited on Amazon Mechanical Turk in exchange for $1.50. In all experiments, participation was restricted to users with an IP address within the United States and an approval rating of at least 95% based on at least 50 previous tasks. An additional 49 participants were excluded for failing a memory check.

Materials, Design, and Procedure Participants first completed a short training to ensure that they could interpret covariation tables, and were then placed in the role of a scientist studying several natural kinds on a fictional planet. Table 1 shows the four kinds – zelmos, drols, grimonds, and yuyus - each associated with a triad of variables (putative cause, effect, and moderator). We illustrate the procedure with zelmos, but the structure was matched across cases.

The scientist was described as investigating the hypothesis that eating yona plants is causally related to developing sore antennas. Participants were told that to test the hypothesis, the scientist performed an experiment, selecting a random sample of 200 zelmos and randomly assigning them to two equal groups that ate a diet either containing or not containing yonas. Participants saw the results of the experiment in the form of a 2 x 2 covariation table cross-classifying zelmos based on whether they ate yonas or not, and whether they developed sore antennas or not. The numbers in the table were selected to support a causal strength with a ΔP of about .4 (range .39-.42).

The scientist then decided to conduct a second experiment with a new, larger sample of 400 zelmos, again randomly assigning zelmos to one of the two diets. But this time the scientist discovered after the experiment that due to a miscommunication between research assistants, half of the zelmos were given salty water, and the other half were given fresh water. The two values of this potentially moderating

Table 1: Materials used in Experiments 1 (all four items) and 2 (zelmo and drol items only).

Item	 Zelmo (lizard-like species)	 Drol (mushroom)	 Grimond (mineral)	 Yuyu (bird)
Cause variable	eating yona plants	saline soil	exposure to sulfuric acid	eating marine snails
Effect variable	sore antennas	bumpy stems	surface cracks	brownish feather tint
Moderator variable	drinking water (salty vs. fresh)	exposure to forest fire smoke (occurred vs. not occurred)	temperature (hot vs. cold)	inhaling volcanic ash (occurred vs. not occurred)

Table 2: Sample causal and explanation judgments in Experiment 1, as a function of judgment type and target (type vs. token).

	Causal judgment	Explanation judgment
Type	How much do you agree with the following statement about what causes zelmos' antennas to become sore?: *For zelmos, eating yonas causes their antennas to become sore.*	How much do you agree with the following explanation of why zelmos' antennas become sore?: *For zelmos, antennas become sore because of eating yonas.*
	Your assistants select one of the zelmos with sore antennas from your second experiment. They call him Timmy. During the experiment Timmy has eaten yonas. You do not know whether Timmy drank fresh water or salty water during the experiment.	
Token	How much do you agree with the following statement about what caused Timmy's sore antennas?: *Eating yonas caused Timmy's antennas to become sore.*	How much do you agree with the following explanation of why Timmy has sore antennas?: *Timmy's antennas became sore because he ate yonas.*

variable were always said to occur normally on the planet (e.g., in the wild, zelmos drink either fresh or salty water, depending on what's available). Luckily for the scientist, the moderator and cause variables varied orthogonally. Participants were told that "to see whether drinking salty water made a difference to the effects of yonas on sore antennas, you decide to look at the results of the experiment within each of these two groups." This time participants were presented with the data split into two tables, one for the salty water subgroup, and one for the fresh water subgroup, each table cross-classifying zelmos in terms of diet and antenna soreness (see Figure 1).

We varied whether the split tables indicated a relationship that was *moderated* or *not moderated*. In the moderated cases (illustrated in Figure 1a), in one subgroup (salty water) the relationship between eating yonas and sore antennas was very strong (ΔP=.82-.85), while in the other subgroup (fresh water), the relationship nearly disappeared (ΔP= -.06-.01). In the non-moderated cases (Figure 1b), each of the split tables corresponded to relationships with a ΔP comparable to the ~.40 from the original, unsplit table. Importantly, the average strength of the relationship across the two split tables was the same in the moderated and non-moderated conditions, and equaled the strength of relationship in the first table that participants saw for each item (within .02 ΔP units).[1] The split tables were accompanied by a note for moderated [non-moderated] conditions: "The tables reveal that the data pattern looks very *different* [*similar*] for zelmos who drank salty water during the experiment and for zelmos who drank fresh water during the experiment. Please compare the two tables to see how different [similar] the patterns are."

Once all three covariation tables had been presented, participants evaluated either claims about *causal relationships* or *explanations* (Table 2). Each claim was presented either at the *type* or *token level*. All claims were general, i.e., they stated a relationship between eating yonas and sore antennas without mentioning the kind of water the zelmo(s) in question drank.[2] Across items, each participant saw two moderated cases and two non-moderated cases, presented in random order. Thus, Experiment 1 had a 2 moderator (moderated vs. non-moderated relationship) x 2 judgment (causal vs. explanatory) x 2 target (type vs. token) mixed design, with moderator manipulated within-subjects. The dependent variable was agreement with causal or explanatory claims, measured on a 1 (strongly disagree) to 7 (strongly agree) scale.

Results and Discussion

Our main question was whether relationships with known moderators support general causal and explanatory claims to the same extent as relationships without known moderators. A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) mixed ANOVA on ratings revealed a main effect of moderator: as shown in Figure 2, participants were significantly less likely to agree with claims about causal and explanatory relationships when a relationship was moderated than non-moderated, $F(1,178)=163.22$, $p<.001$, $\eta_p^2=.478$, even though moderated and non-moderated relationships were equated for overall strength (defined as the degree of covariation between putative causes and effects). There were no other significant main effects or interactions (all p's≥ .211), suggesting that the effect of moderator was not itself moderated by the nature of the judgment (causal or explanatory, type or token).

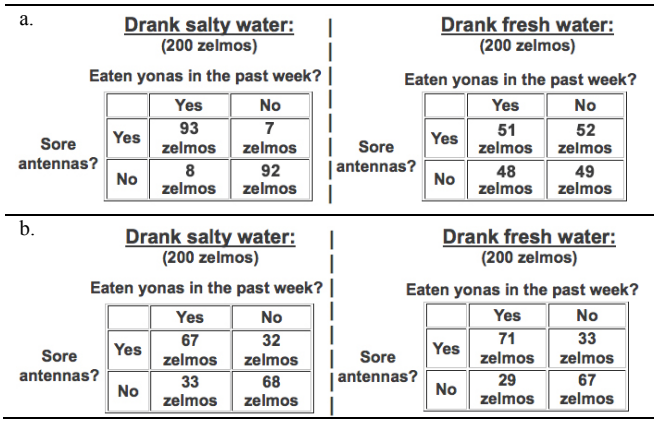


Figure 1: Sample covariation matrices from Experiment 1: (a) split tables in the moderated condition, ΔP's=.85 and .00 (*M*=.42); (b) split tables in the non-moderated condition, ΔP's=.35 and .38 (*M*=.37).

[1] For each item, average ΔP's in the moderated and non-moderated conditions could differ slightly (by no more than .05 ΔP units). Importantly, the non-moderated condition strength never exceeded the moderated condition strength, which worked against our hypothesis. (This also holds for other metrics of causal strength computed over covariation tables, e.g. causal power, Cheng, 1997.)

[2] In both Experiments 1 and 2, an additional group of participants evaluated qualified causal and explanatory claims that specified the subgroup defined by the moderator variable, e.g., *For zelmos who drank salty water, eating yonas causes their antennas to become sore*. Participants in both experiments also evaluated counterfactual claims. Due to space limitations, here we focus on unqualified claims only, and we omit counterfactual ratings.

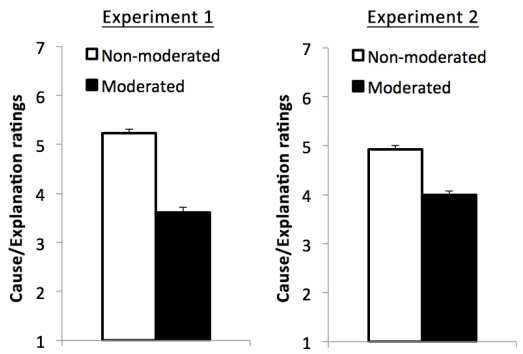


Figure 2: The effect of moderator on ratings of causal and explanatory relationships in Experiments 1 (left) and 2 (right). Error bars correspond to 1 SEM.

Thus, the results of Experiment 1 indicate an effect of stability over and above causal strength: holding causal strength fixed, causal relationships are penalized for instability. The effect was consistent across tasks, holding both for causal and explanatory judgments at both the type and token levels.

Experiment 2

The results of Experiment 1 provide initial evidence for an effect of stability (i.e., invariance across a range of circumstances) on causal and explanatory ratings. Yet these results are also amenable to an alternative interpretation. In Experiment 1, the moderated relationship had two characteristics. First, it was relatively unstable, in that it held in only one kind of circumstance described in the fictional world. The non-moderated relationship, by contrast, held in both kinds of circumstances. Second, the moderated relationship had a narrower *actual scope*, i.e., the actual proportion of the population for which it held was relatively small: as the moderating variable took the value favoring the presence of the causal relationship in half of the actual members of the population, the moderated relationship held for only 50% of the actual population. By contrast, the non-moderated relationship held in the entire actual population. But stability and actual scope are distinct: an unstable relationship can have wide actual scope if the circumstance in which it holds happens to be frequent. This suggests an alternative explanation for the results of Experiment 1: moderated relationships could be penalized merely for their narrow actual scope. This alternative explanation also suggests that the penalty for moderated relationships could be a superficial pragmatic phenomenon: a generalization about a population could be infelicitous when it applies only to a small actual portion of the population. For instance, "having sex can cause deep vein thrombosis" is potentially *misleading* if the generalization only applies to women who become pregnant.

To address the possibility that our results are driven by a concern for actual scope rather than stability, we conducted a further experiment where in addition to varying the number of circumstances in which a causal relationship holds (thus investigating effects of stability), we orthogonally varied the relative size of the two subsets of the population broken down by the moderator variable (thus varying actual scope). In Experiment 1, the proportion of the population for which the enabling circumstance (e.g., drinking salty water) held was always 50%, and therefore fixed actual scope to 50% of the population. In Experiment 2, we introduced two additional conditions: a *high-frequency* condition in which the enabling circumstance was present in 70% of the population, and a *low-frequency* condition in which the enabling circumstance was present in 30% of the population. The actual scope of the moderated relationship thus varied across frequency conditions, but its (in)stability remained the same: in all frequency conditions, there was one possible circumstance (e.g., drinking fresh water) in which the causal relationship did not hold.

Experiment 2 also included a set of ratings concerning the structure and strength of causal relationships. Participants were asked whether in their view *a causal relationship* between the cause (e.g., eating yonas) and effect (e.g., sore antennas) is *likely to exist*, and if so how *strong* it is. By using such a formulation, which taps more directly into participants' beliefs about causal relationships as opposed to communication and language use, we hoped to address the possibility that the results of Experiment 1 were due to some pragmatic infelicity of our general (unqualified) claims.

Method

Three-hundred-and-ninety-three participants (excluding an additional 83 participants who failed a memory check) were recruited on Amazon Mechanical Turk in exchange for $1.30.

Materials, Design, and Procedure The materials, design and procedure were the same as in Experiment 1, with the following exceptions. First, we presented split data tables in the context of an additional experiment designed to determine whether the moderator makes a difference (rather than a consequence of the research assistants' mistakes), and we increased the sample sizes in the hypothetical experiment to accommodate the changes in our design.

Second, we varied the *moderator frequency*: the base rate of the enabling circumstance (i.e., the moderator value for which the causal relationship held) in the natural population and in the sample. This circumstance occurred in either 30% (*low frequency*), 50% (*medium frequency*), or 70% of cases (*high frequency*). Participants were told that the sizes of the groups were intentionally matched to the frequency of the enabling circumstance in the natural population.

To keep the mean strength of causal relationships (averaged across split tables) the same (ΔP=.31) in the moderated and non-moderated condition despite variation in the base rate



Figure 4: Sample diagram provided to participants to illustrate the design of a hypothetical study (high-frequency moderator condition).

approximately equal to zero in one subgroup, the strength of the relationship in the subgroup with a causal association inevitably had to vary (ΔP=.97, ΔP=.61, and ΔP=.44 for low, medium, and high, respectively).

Thus, Experiment 2 had a 2 moderator (moderated vs. non-moderated relationship) x 2 judgment (causal vs. explanatory) x 2 target (type vs. token) x 3 moderator frequency (low 30%, medium 50%, high 70%) mixed design, with moderator manipulated within-subjects. The main dependent variables were agreement with causal or explanatory claims, measured on a 1 (strongly disagree) to 7 (strongly agree) scale.

Participants also answered questions about the structure and strength of causal relationships between pairs of variables (see Griffiths & Tenenbaum, 2005). For instance, a structure judgment might ask: "In your opinion, how likely is it that there is some causal relationship between eating yonas and having sore antennas?", rated on a scale from not at all likely (1) to very likely (7). A strength judgment might ask: "If there is a causal relationship between eating yonas and having sore antennas, how strong do you think it is?", rated on a scale from very weak relationship (1) to very strong relationships (7). Only participants who gave a rating higher than 1 in response to *structure* were asked to rate *strength.* Here we report the findings for judgments concerning the candidate cause and effect (e.g., eating yonas → sore antennas). To prevent participant fatigue given additional ratings, the number of items was reduced to two (see Table 1).

Results and Discussion

Main ratings. A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) x 3 moderator frequency (low, medium, high) mixed ANOVA on main ratings revealed two main effects. As predicted, moderated relationships were rated lower than non-moderated relationships $F(1,381)=79.19$, $p<.001$, $\eta_p^2=.172$, replicating the moderator effect from Experiment 1 (see Figure 2b). In addition, type ratings ($M=4.63$) were higher than token ratings ($M=4.26$, $F(1,381)=8.97$, $p=.003$, $\eta_p^2=.023$. There were no other main effects or interactions (all p's≥.154). Most notably, there was no effect of moderator frequency, $F(2,381)=1.60$, $p=.203$.

Ratings of causal structure and strength A 2 moderator (moderated relationship, non-moderated relationship) x 2 judgment (causal, explanatory) x 2 target (type, token) x 3 moderator frequency (low, medium, high) mixed ANOVA on causal structure ratings revealed a significant main effect of moderator, $F(1,381)=56.49$, $p<.001$, $\eta_p^2=.129$. As shown in Figure 5, ratings were on average higher for the non-moderated relationship than for the moderated relationship. There was also a significant interaction between frequency and target, $F(2,381)=3.48$, $p=.032$, $\eta_p^2=.028$; there was a trend such that token ratings were lower than type ratings in the high frequency condition but not in others; however none of the simple effects were significant (p's≥.275).

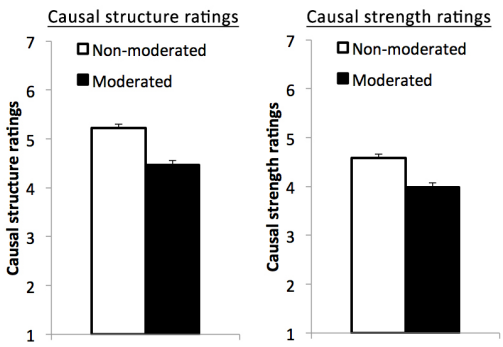


Figure 5: The effect of moderator on ratings of causal structure and strength in Experiment 2. Error bars correspond to 1 SEM.

An equivalent analysis of strength judgments also revealed a significant main effect of moderator, $F(1,380)=38.76$, $p<.001$, $\eta_p^2=.093$, with higher ratings for unmoderated than moderated relationships (Figure 5). No other effects reached significance.

In sum, Experiment 2 shows that stability affects causal judgments even controlling for causal strength, and that this influence is unlikely to reflect the frequency of the moderating circumstance or pragmatic considerations.

General Discussion

In two experiments, we document an important factor shaping people's assessments of causal relationships over and above causal strength: the *stability* of the causal relationship – that is, the extent to which it holds across various possible circumstances. While philosophers of science have stressed the importance of stability in scientific modeling and explanation, the role of stability in causal and explanatory judgments has so far been largely unacknowledged in psychology (but see Gerstenberg et al., 2012, 2015). Experiments 1 and 2 show that stability considerations play a consistent role in various contexts: people are more willing to endorse causal and explanatory claims involving stable causal relationships for statements at both type and token levels. The results of Experiment 2 also indicate that the effect of stability is not reducible to actual scope (that is, unstable causal relationships are not penalized merely because they hold in a smaller actual proportion of the population).

These findings point to an important limitation of the metrics of causal influence that have dominated the psychological research on causal reasoning and induction, such as ΔP or power PC. These measures track one aspect of causal relationships (their average strength in a population), but do not capture another important aspect that matters for causal assessment, namely the extent to which the relationship holds in a range of plausibly occurring background circumstances.

Our findings are related to the work of Liljeholm and Cheng (2007), who show that people can infer the presence of background factors interacting with a causal relationship based on differences in covariation across contexts. These findings are important in demonstrating that people are able to track the kind of evidence relevant to assessments of stability, which our experiments show will in turn affect the endorsement of general causal claims as well as ratings for causal structure and strength.

Our results also provide support for the *exportability*

theory of explanation (Lombrozo & Carey, 2006) and causal ascriptions (Lombrozo, 2010). According to this theory, a central function of explanations and causal ascriptions is to pick out patterns of dependence that are exportable in the sense that they support future predictions and interventions. If this is correct, we should expect explanatory and causal ratings to favor more stable relationships: by being insensitive to variations in background circumstances, a stable causal relationship provides more reliable opportunities for future prediction and intervention.

Our findings also suggest directions for future research. First, how does stability connect with issues of simplicity in causal representation? As Woodward (2016) notes, causal structures involving stable relationships can be represented with sparse causal graphs, whereas unstable relationships complicate the task of causal representation.

Second, what are the boundaries of the observed effects of stability? In a very different paradigm involving collisions among physical objects, Gerstenberg et al. (2012) found that the robustness of an outcome (whether a ball clearly or barely went through a gate) did not affect "cause" versus "prevent" judgments. However, it did predict choices between descriptors of causal relationships ("caused" "prevented," "almost caused/prevented," "helped (to prevent)"), and it had some effect on the responsibility assigned to potentially competing causes in complex causal structures, including causal chains (Gerstenberg et al., 2015).

Third, is it possible (and necessary) to draw a line between the stability of a causal relationship with respect to *background circumstances* (defined loosely as circumstances not included in *cause* and *effect*) versus the stability of an outcome with respect to the *manner* in which the target cause occurs (Lombrozo, 2010), and/or to the *status of intermediate causes* in a causal chain (e.g., as in Gerstenberg et al., 2012, 2015)?

Fourth, can stability account for intransitivity in causal chains? For instance, it's reasonable to say that sex causes pregnancy, which causes nausea, but it seems less reasonable to say that sex causes nausea. Johnson and Ahn (2015) show that causal chains with equally strong intermediate links may nevertheless differ in transitivity, and argue that some causal relations must be represented as "causal islands" rather than coherent networks. Could stability help explain what makes some causal relations behave as causal islands (regardless of the nature of the representation)? For example, intransitivity could arise if the component links are evaluated with respect to different sets of moderators, and/or there is little overlap between subsets of background circumstances for which the component relationships hold.

Fifth, how does the stability of a relationship across a range of circumstances relate to the degree of *guidance* it provides? Consider again the causal relationship between eating yonas and getting sore antennas in the case where it holds only in one background circumstance. One way to alleviate this instability is to explicitly build this background circumstance into the relationship: "*For zelmos who drink salty water*, eating yonas causes sore antennas." This qualified claim seems better than the bare claim that eating yonas causes sore antennas – not because it applies to a wider range of possible circumstances *per se*, but because it is more "guiding": by flagging the circumstance under which the relationship holds, it provides a better sense of *when* the relevant causal relationship can be used for prediction and control, and is therefore exportable in the sense that it contains conditions for application, whether or not those conditions hold widely. Thus one question is whether people are sensitive to considerations of *guidance* when assessing unstable relationships, and how guidance (achieved by building in background circumstances) differs from offering an explanation or causal claim with the enabling conditions instead identified as additional, interacting causes. The roles of stability and guidance in causal ascription and explanation are ripe for further investigation.

Acknowledgments

This work was supported by the Varieties of Understanding Project, funded by the John Templeton Foundation.

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*, 147-149.

Cheng, P. (1997). From covariation to causation: A theory of causal power. *Psychological Review, 104,* 367-405.

Gerstenberg, T., Goodman, N., Lagnado, D.A., & Tenenbaum, J.B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In Miyake, N., Peebles, D., and Cooper, R. P., (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society, pp. 378–383. Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N.D., Lagnado, D.A, & Tenenbaum, J.B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock,T., and Maglio, P. P., (Eds.), Proceedings of the 37th Annual Conference of the Cognitive Science Society, pp.782–787, Austin, TX. Cognitive Science Society.

Griffiths, T.L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Johnson, S. G., & Ahn, W. K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive science*, *39*(7), 1468-1503.

Liljeholm, M., & Cheng, P. (2007). Coherent generalization across contexts. *Psychological Science, 18,* 1014-1021.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions,,functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303-332.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167-204.

Murray, D., & Lombrozo, T. (2016). Effects of manipulation on attribution of causation, free will, and moral responsibility. *Cognitive Science,*.doi: 10.1111/cogs.12338.

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science, 39(6),* 1320-1347.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review, 115*, 1-50.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy, 25*, 287-318.

Woodward, J. (2016). The problem of variable choice. *Synthese, 193*, 1047-1072.

What does the crowd believe? A hierarchical approach to estimating subjective beliefs from empirical data

Michael Franke, Fabian Dablander, Anthea Schöller
{michael.franke, fabian.dablander, anthea.schoeller}@uni-tuebingen.de
Department of Linguistics, Wilhelmstraße 19
72074 Tübingen, Germany

Erin Bennett, Judith Degen, Michael Henry Tessler, Justine Kao, Noah D. Goodman
{ebennett, jdegen, mtessler, justinek, ngoodman}@stanford.edu
Department of Psychology, Stanford University
450 Serra Mall, Stanford, CA 94305 USA

Abstract

People's beliefs about everyday events are both of theoretical interest in their own right and an important ingredient in model building—especially in Bayesian cognitive models of phenomena such as logical reasoning, future predictions, and language use. Here, we explore several recently used methods for measuring subjective beliefs about unidimensional contiguous properties, such as the likely price of a new watch. As a first step towards a way of assessing and comparing belief elicitation methods, we use hierarchical Bayesian modeling for inferring likely population-level beliefs as the central tendency of participants' individual-level beliefs. Three different dependent measures are considered: (i) slider ratings of (relative) likelihood of intervals of values, (ii) a give-a-number task, and (iii) choice of the more likely of two intervals of values. Our results suggest that using averaged normalized slider ratings for binned quantities is a practical and fairly good approximator of inferred population-level beliefs.

Keywords: subjective beliefs, hierarchical modeling, Bayesian data analysis, Bayesian cognitive models

Motivation

When trying to understand observed behavior we readily ascribe beliefs and desires to fellow agents. This happens intuitively, in folk psychology, but also in science. Scientific ascription of latent mental states plays an important role in many explanations of higher-order cognition: decision making, reasoning, language use, etc. It is therefore vital to have methods for validating explanatory mental state ascriptions.

A family of models where this is particularly pressing are Bayesian models of cognition which seek to explain task behavior in a variety of domains as partially informed by what participants believe about mundane events—their *prior beliefs*. Take interpretation of language. "That watch cost a million dollars," tends to be interpreted as hyperbole: conveying affect rather than literal truth. Empirical data on whether similar statements are understood as hyperbole can be explained well by a Bayesian model of utterance interpretation (Kao, Wu, Bergen, & Goodman, 2014). This model assigns a crucial role to an empirical measure of participants' expectations about the likely or normal price of a watch—only when the uttered price is sufficiently unlikely *a priori* will hyperbolic interpretation be possible. Other examples of domains in which empirically successful models have included a measure of participants' prior beliefs include making predictions about everyday events (Griffiths & Tenenbaum, 2006), referential reasoning (Frank & Goodman, 2012), strength of pragmatic enrichments (Degen, Tessler, & Goodman, 2015), or quantifier interpretation (Schöller & Franke, 2015).

Many methods have been used to build prior distributions used in Bayesian cognitive models. One method of getting at subjective beliefs is to take actual frequencies as an approximation to subjective beliefs (e.g. Griffiths & Tenenbaum, 2006). Unfortunately, frequency data may be unavailable (e.g., one-shot events) or deviate from participants' subjective beliefs in crucial respects.

Another common approach is to empirically measure subjective beliefs by *give-a-number* tasks. The simplest version would be this: being told that John just bought a new watch, participants are asked for a single numerical estimate of its price. This task is easy to comprehend and implement, but a single number does not provide much information about the subjective belief that it is a manifestation of. One solution is to infer which parameterized distribution best explains the observed number choices, either at the individual level (Manski, 2004) or at the population level (Tauber & Steyvers, 2013). More sophisticated *give-a-number* tasks give more information, but can be difficult to implement and analyze.

More complex elicitation methods include *scoring rules* (Savage, 1971; Andersen, Fountain, Harrison, & Rutström, 2014; Schlag, Tremewan, & van der Weele, online first), prominent in economics, and the *iterated learning paradigm* (Lewandowsky, Griffiths, & Kalish, 2009). These methods are powerful but difficult to implement, and often assume very specific behavior from participants—such as optimal decision making under perfect knowledge of the payoff scheme.

In sum, there is a tradeoff between simplicity of paradigms and their information content. Ideally, we would like to have an experimental method for measuring subjective beliefs that (i) provides sufficient and reliable-enough information about subjective beliefs to derive testable predictions from cognitive models that rely on such information, (ii) is easy to understand by participants, (iii) is easy to implement, and that (iv) does not require sophisticated means of data analysis. Moreover, the ideal method would be (v) flexible enough to allow inferred subjective belief distributions beyond standard pa-

Item	Bins ({min, max} step; units)	Context sentence	GAN question	BH frame	PC frame
coffee	{<44, >200} 2; degrees	X has just fetched himself a cup of coffee from the office vending machine.	What do you think the temperature of his coffee is?	His coffee was the following temperatures	The temperature of his coffee is N degrees.
commute	{0, >98} 7; minutes	X commuted to work yesterday.	How many minutes do you think she spent commuting yesterday?	She commuted for the following numbers of minutes yesterday	She spent N minutes commuting.
joke	{0, 14} 1; children	X told a joke to 14 kids.	How many of the kids do you think laughed?	The following number of kids laughed	N of the children laughed.
laptop	{0, >7500} 500; dollars	X bought a laptop.	How much do you think it cost?	The laptop cost the following numbers of dollars	The laptop cost $N.
marbles	{0, 14} 1; marbles	X threw 14 marbles into a pool.	How many of the marbles do you think sank?	The following number of marbles sank	N of the marbles sank.
movies	{0, > 210} 16; minutes	X just went to the movies to see a blockbuster.	How many minutes long do you think the movie was?	The movie was the following numbers of minutes long	The movie was N minutes long.
TV	{0, >43} 3; hours	X watched TV last week.	How many hours do you think he spent watching TV last week?	He watched TV for the following numbers of hours last week	He spent N hours watching TV.
watch	{0, >750} 50; dollars	X bought a watch.	How much do you think it cost?	It cost the following numbers of dollars	The watch cost $N.

Table 1: Experimental items. *X* was a randomly generated name (different on each trial). *N* was one of the bins.

rameterized distributions and (vi) would provide a good approximation of the central tendency or average of subjective beliefs in a given population. The latter would allow using one set of participants for measuring prior beliefs and another for whatever other task is of interest, so as to avoid potential cross-over effects.

A simple technique that seems to fit this bill is the *binned histogram* task that has recently been used with apparent success (e.g. Kao, Wu, et al., 2014; Kao, Bergen, & Goodman, 2014; Tessler, 2015; Schöller & Franke, 2015). Participants adjust sliders to express (relative) subjective beliefs about how likely it is that the value of an uncertain contiguous quantity lies in some interval of possible values (a bin). E.g., to report beliefs about a watch's price, participants adjust sliders whose endpoints are labelled "extremely likely" and "impossible." Each slider corresponds to one of 15 bins that partition the range of plausible values (established by a pre-test), such as "$0-$50", "$50-$100", etc. up to "$700-$750" and "more than $750." Each participant's slider ratings are normalized and the results averaged across participants. The resulting *mean slider ratings* look like plausible population-level beliefs and give good results when fed into cognitive models that predict task behavior based on prior expectations.

Explanatory success aside, the question remains whether mean slider ratings measure what we would like them to. Are these good approximations of a central tendency of participants' individual beliefs? Are these measures consistent with participants' behavior in other tasks, such as *give-a-number*? To scrutinize the binned histogram task we address these issues by collecting data in a within-subject paradigm from three task types: (i) *binned histograms*, (ii) *give-a-number* and (iii) *paired comparisons*. The latter asks participants for a direct comparison of two bins from the *binned histograms* task, and was included as a further consistency check. We used a Bayesian hierarchical model to analyze this data jointly. The model infers latent subjective beliefs, where each subjective belief is, intuitively, a noisy perturbation of a population-level belief. We find that mean *a posteriori* population-level beliefs are approximated well by mean slider ratings from the binned-histograms task, suggesting that this may indeed be a sound and easy measure of what the crowd believes. Moreover, the model is able to capture participants' behavior in all three task types reasonably well, suggesting that what mean slider ratings measure is what we think it is: a population-level central tendency of latent subjective beliefs.

Experimental elicitation of prior beliefs

Participants, procedure and materials We recruited 20 self-reported English native speakers over Mechanical Turk and collected responses for eight different items (listed in Table 1) using the three different dependent measures mentioned above. Each participant rated each item using each dependent measure. Trial order was randomized.

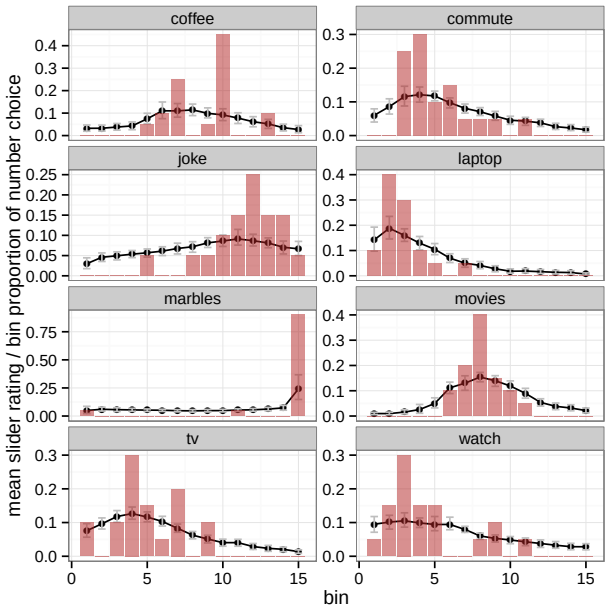
On *give-a-number* (GAN) trials, participants saw the context sentence and GAN question for each item (see Table 1) and provided one number by adjusting a slider with endpoints labeled as the lowest and highest number for that item. Min and max numbers are shown in Table 1 and were taken from previous studies that had used these items (Degen et al., 2015;

Schöller & Franke, 2015; Kao, Wu, et al., 2014). The selected number appeared above the slider so participants knew exactly what the value of the slider would be.

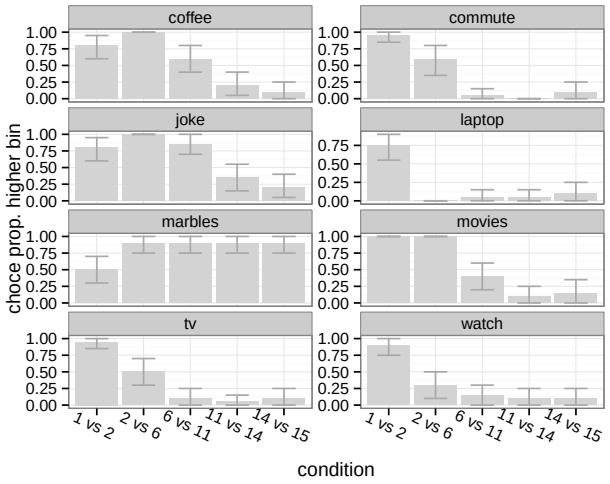
On *binned histogram* (BH) trials, participants saw an item's context sentence and were asked to *Please rate how likely it is that Y* (where *Y* came from the corresponding BH frame in Table 1). They adjusted 15 continuous sliders, one per bin, with five points labeled *impossible*, *not very likely*, *neutral*, *very likely*, and *extremely likely*. Bins were determined by dividing the interval spanned by the item's minimum and maximum value into equally sized bins (Table 1).

On *paired comparison* (PC) trials, participants saw an item's context sentence and had to click on one of two options shown side by side – whichever one they thought was more likely. Each option used the appropriate PC frame in Table 1 and numbers were filled in by comparing the following bins: 1 vs 2, 2 vs 6, 6 vs 11, 11 vs 14, and 14 vs 15. Paired comparison trials always occurred as a block of five trials, with order of comparisons randomized within block.

Results Fig. 1 shows mean slider ratings alongside the frequencies of number choices in the *give-a-number* task. It looks like the latter could be samples from the former, with a tendency to modal choices. But some *give-a-number* results seem influenced by a tendency towards round or salient numbers. E.g., in item *coffee* more than 40% of participants guessed that a coffee from the vending machine was ca. 150°F (bin 10). The results from the *paired comparison* seem consistent with the mean slider ratings as well, albeit not as straightforwardly as one might think. This is clearest from the *marbles* case. While the mean slider ratings suggest that bins 2, 6 and 11 were considered roughly equiprobable on average, almost every participant chose the higher bin in direct bin comparisons 2 vs 6 and 6 vs 11. A potential explanation for this, which we will implement formally in the data-generating model, is that the *paired comparison* condition invited participants to think about what they thought was mostly likely the case (usually: bin 15 in the *marbles* case) and to choose the bin that is closest to that.



(a) Red: proportion of bins corresponding to number choices on *give-a-number* trials. Black: mean slider ratings on *binned histogram* trials.



(b) Proportion of higher bin choices on *paired comparison* trials.

Figure 1: Average data. Bars are bootstrapped 95% CIs.

Model

The data we would like to explain are: (i) the normalized slider ratings $s_{ijk} \in [0;1]$ of participant $i \in \{1,\ldots,20\}$ for item $j \in \{1,\ldots,8\}$ and bin $k \in \{1,\ldots,15\}$ from the binned histogram task; (ii) the bins $n_{ij} \in \{1,\ldots,15\}$ in which participant i's number choice for item j was in the give-a-number task; and (iii) the binary choices $c_{ijl} \in \{0,1\}$ of whether participant i selected the higher bin for item j in the paired comparison task for each comparison l. There are two simplifications in need of commenting. In (i), we focus on slider ratings after normalizing for each participant, because we assume that slider adjustments reflect relative, not absolute estimates of subjective beliefs. In (ii), we focus on bin choices, not actual number choices, in order to avoid, as much as possible, considerations of salience of particular numbers, and also because otherwise data from items with smaller domains of plausible numbers would get more weight than data from items with a wider range of number choices. (Future work should investigate the relation to an explicit *give-a-bin* task.)

All three pieces of data are to be explained as functions of subjective beliefs P_{ij}, with P_{ijk} being participant i's belief about the relative likelihood of bin k for item j. Each P_{ij} defines a likelihood for our data, via appropriate link functions. Variance in subjective beliefs is harnessed by a population-level hyper-prior with central tendency Q_j. The structure of this model is pictured in Fig. 2.

To fill the structure in Fig. 2 with life, we need to spell out three parameterized link functions, one for each task type,

and the relation between population-level belief Q_j and individual beliefs P_{ij}. Let's start with the latter. The idea is that P_{ij} are noise-perturbed variants scattered around Q_j, with some parameter w to determine how much perturbation we should expect. To realize this, the model assumes that P_{ij} are distributed according to a Dirichlet distribution with weights given by wQ_j:

$$Q_j \sim \text{Dirichlet}(1,\ldots,1) \qquad w \sim \text{Gamma}(2,0.1)$$
$$P_{ij} \sim \text{Dirichlet}(wQ_j)$$

The higher w, the more likely it is that P_{ij} is "close" to Q_j.

The link function for the **slider rating data** uses a logit transformation to project observed slider ratings s_{ijk} and latent probabilities P_{ijk}, which are bound to lie between 0 and 1, to the reals. The likelihood of logit-transformed observation s_{ijk} is given by a Gaussian with standard deviation σ around the logit-transformed predictor P_{ijk}. On top of that, there is a parameter κ, the steepness of the logit transform of P_{ijk}, that allows response likelihoods to capture end-point affinity for $\kappa > 1$ (values of P_{ijk} close to 0 or 1 are likely mapped to 0 or 1) or end-point aversion for $\kappa < 0$ (values of P_{ijk} are likely to be realized as more median), with a prior that expects $\kappa = 1$.

$$\text{logit}(s_{ijk}) \sim \text{Norm}(\text{logit}(P_{ijk},\kappa),\sigma)$$
$$\sigma \sim \text{Gamma}(0.0001,0.0001) \quad \kappa \sim \text{Gamma}(5,5)$$

The link function for **number choice data** treats each bin n_{ij} as a draw from a categorical distribution where the probability of bin k is proportional to $\exp(aP_{ij})$, i.e., a soft-max choice from P_{ij}. The higher parameter a, the more likely n_{ij} is the mode of P_{ij}. For $a \to 0$, all bins become equiprobable.

$$n_{ij} \sim \text{Categorical}(\exp(aP_{ij})) \qquad a \sim \text{Gamma}(2,1)$$

Finally, consider the link function for **bin comparisons**. We are interested in the likelihood with which participant i selects the higher bin for item j in bin comparison condition l. Suppose l is about comparing the lower bin b_l to the higher b_h. Perhaps the most natural approach would be to link the likelihood of choosing b_h over b_l to the difference between P_{ijb_h} and P_{ijb_l}. However, as discussed above, this does not appear to be what participants were doing. Indeed, a model that implements this idea blatantly fails to capture the relevant regularities in the data. Another plausible link function is to assume that what matters is the distance to the mode of P_{ij}: soft-max prefer the bin that is closer to the mode of P_{ij}; select randomly if both bins are equally far from the prototype.[1]

$$c_{ijl} \sim \text{Bern}((1+\exp(2b(1-p^{\text{high}}_{ijl})))^{-1}) \quad b \sim \text{Gamma}(2,1)$$

$$p^{\text{high}}_{ijl} = \begin{cases} 2 & \text{if } mode(P_{ij}) \text{ is closer to higher} \\ & \text{bin of } l \text{ than to lower bin} \\ 1 & \text{if equal distance} \\ 0 & \text{otherwise} \end{cases}$$

[1] $(1+\exp(2b(1-x)))^{-1} = \frac{\exp(bx)}{\exp(bx)+\exp(by)}$ if $x = 2-y$.

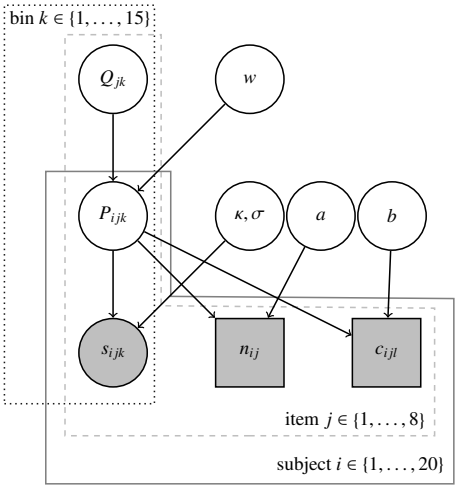


Figure 2: The data-generating model as a probabilistic graphical model, following conventions of Lee and Wagenmakers (2015). Shaded nodes are observed, white nodes are latent variables. Square nodes represent categorical, round nodes continuous variables. Boxes indicate scope of indices.

Inference

The model was implemented in JAGS (Plummer, 2003). 50,000 samples were obtained from two chains with a thinning rate of 2 after a burn-in of 100,000 that ensured convergence according to $\hat{R}$ (Gelman & Rubin, 1992).

Means and bounds of 95% high-density intervals (HDIs) of the posteriors for model parameters are in Table 2. Posterior credible levels of w allow for a limited amount of slack around Q_j. Values for κ indicate that, on average, participants had no preference for or against extreme slider ratings. Relatively high values of a indicate that participants, on average, had a strong tendency to choose modal values in the *give-a-number* task.

	w	κ	σ	a	b
lower	14.65	0.98	0.26	22.04	1.11
mean	15.55	0.99	0.28	27.43	1.27
upper	16.48	0.99	0.31	32.95	1.42

Table 2: Summary statistics for posteriors on parameters

Posterior estimates of Q_j are the most relevant. Fig. 3 shows their means with their 95% HDIs, alongside the mean slider ratings. The latter provide a very good approximation of the inferred population-level beliefs. Inspection of posteriors of individual P_{ij} shows that there is ample variation between participants. Still, the way the model suggests we should think about harnessing the individual P_{ij}s under a population-level central tendency is closely approximated by mean slider ratings. Although the match is not perfect, it is good enough to say that the latter are a practical way of approximating what the crowd believes despite individual differences.

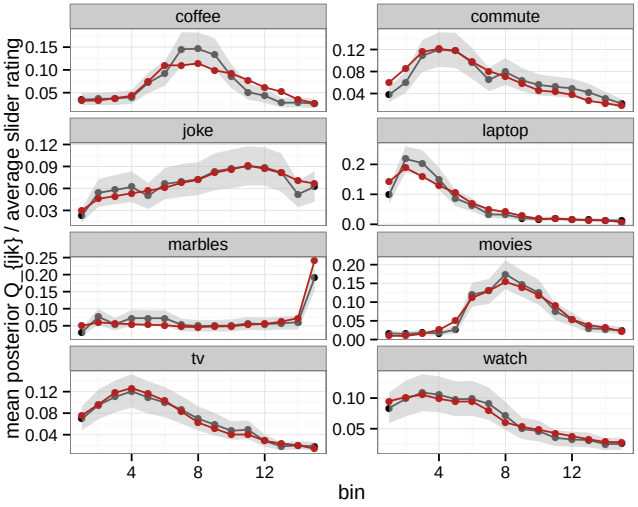
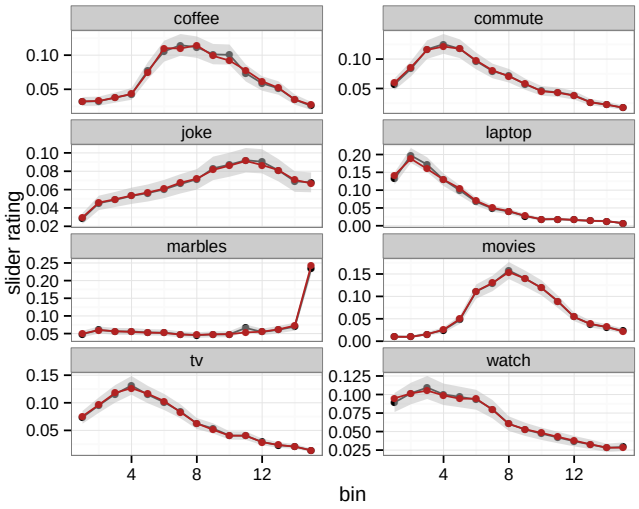


Figure 3: Means of posteriors over Q_j in black with gray area indicating 95% HDIs. Red: mean slider ratings.

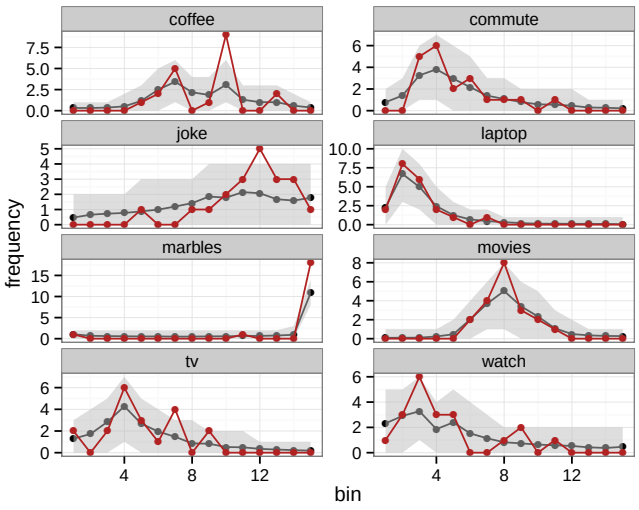
Model criticism

Inferences based on the model are only as reliable as the model itself is plausible. Model criticism is therefore important. Fig. 4 shows posterior predictive checks at the population-aggregate level for all of our three task types. For the binned histogram task, posterior predictions are spot-on. For the give-a-number task, some of the data is surprising despite the model being trained on this very data. This could have various reasons: (i) the give-a-number data does not have a huge influence on the posterior likelihood, (ii) number choices may be influenced by saliency and/or roundness of numbers after all (and thus, not accurately reflect the true beliefs Q_j). Finally, there is one condition in the paired comparison task that the model definitely got wrong. This is the choice of what is more likely: that one or that none of 14 marbles thrown into a pool would sink. The model predicts that almost everybody should answer that it is more likely that one marble sank. But that is not what we observe. It may be that participants revise beliefs about "normality" of the marbles, while holding on to an assumption that all marbles behave in the same way (Degen et al., 2015).

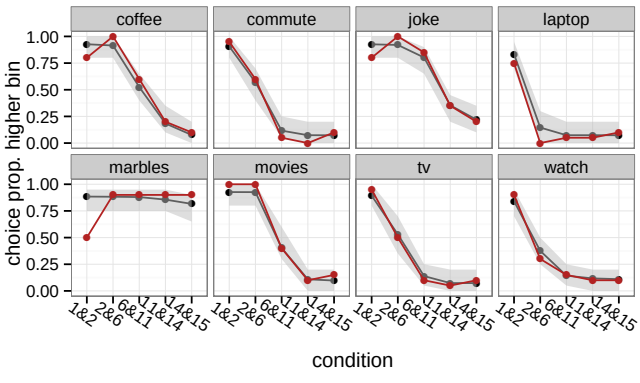
Posterior predictive checks indicate that the trained model captures patterns of answers at the aggregate population level well. To have a more fine-grained measure of model fit, we also looked at posterior predictive p-values (Gelman, Carlin, Stern, & Rubin, 2014) at the level of participants and items. We look at the binned histograms task, because this is our main focus here and population-level posterior predictive checks for BH revealed no systematic deviance. Fixing a participant and an item, observations and replicates are probability vectors of length 15. In a first analysis, we used the mean of these probability vectors as a test statistic. The minimum posterior predictive p-value over all 20 (participants) times 8 (items) cases was 0.13, suggesting that the means of observed s_{ij} are non-surprising to the trained model. In a second analysis, we used entropy as a test statistic. Two cases gave poste-



(a) Binned histogram task.



(b) Give-a-number task.



(c) Paired comparison task.

Figure 4: Posterior predictive checks for aggregate data. Red lines give empirical observations. Black lines are means of posterior predictive samples, gray areas are 95% HDIs.

rior predictive p-values lower than 0.05. These were from the two participants who gave a very extreme slider rating for the "marbles" item, basically assigning all "mass" to the last bin. What this suggests is that the model can cope reasonably well also with individual-level data, but, somewhat unsurprisingly, has problems accounting for "extreme" choices, given that the population-level hyper-prior on P_{ij} will lead to shrinkage.

Conclusion

The data and model presented here suggest that mean slider ratings are consistent with other measures of subjective belief, namely from give-a-number and paired comparison tasks. Future research should evaluate whether this holds for other possible measures of subjective beliefs as well, such as iterated-learning or scoring-rule tasks.

There are aspects of the data that the model does not capture well, but there are also natural explanations for these discrepancies. It therefore does not seem implausible that participants' latent beliefs could have generated the data from all three task types roughly in the way assumed by the model's link functions, with each subjective belief being an expression of a population-level central tendency. If that is so, then mean slider ratings from the binned histogram task are a practical and reliable approximation of what the crowd believes.

Future research should investigate whether and how our results can be extended to other types of uncertain variables. Here, guided by the needs of many previous Bayesian cognitive models, we focused on uncertainty about unidimensional contiguous variables. It remains to be seen whether the binned histogram task can be applied to higher-dimensional variables, like joint prior beliefs over, say, *height* and *weight* of an individual or, more abstractly, different dimensions of a multi-dimensional property like *intelligence*. Another challenge lies in devising means of eliciting prior expectations about properties that do not have clear and familiar measure terms that can be used to label bins—again like *intelligence*.

Finally, another aspect that our model has ignored so far are potential individual differences in the way subjective beliefs P_{ij} generate responses. It is not unlikely that there is individual variation in at least some link function parameters, such as κ, which expresses end-point attraction or end-point aversion in the binned histograms task. A detailed investigation of such individual-level differences must be left to future work. Still, we believe that the model presented here is an important first step towards finding reliable and practical means of measuring what the crowd believes.

Acknowledgments

MF's work was supported by the Institutional Strategy of the University of Tübingen (DFG, ZUK 63). MF, FD, AS and JD's work was supported by the Priority Program XPrag.de (DFG Schwerpunktprogramm 1727). This work was further supported by ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG and an SNF Early Postdoc.Mobility Award to JD. Thanks to three anonymous reviewers for helpful suggestions and criticism.

References

Andersen, S., Fountain, J., Harrison, G. W., & Rutström, E. E. (2014). Estimating subjective probabilities. *Journal of Risk and Uncertainty*, *48*, 207–229.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of CogSci 37*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd edition ed.). Boca Raton: Chapman and Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of CogSci 36*. Austin, TX: Cognitive Science Society.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *PNAS*, *111*(33), 12002–12007.

Lee, M. D., & Wagenmakers, E.-J. (2015). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, MA: Cambridge University Press.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of Distributed Statistical Computing 3*.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Schlag, K. H., Tremewan, J., & van der Weele, J. J. (online first). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*.

Schöller, A., & Franke, M. (2015). Semantic values as latent parameters: Surprising *few* & *many*. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of SALT*.

Tauber, S., & Steyvers, M. (2013). Inferring subjective prior knowledge: An integrative Bayesian approach. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the CogSci 35* (pp. 3510–3515).

Tessler, M. H. (2015). Understanding *Belief bias* by measuring prior beliefs for a Bayesian model of syllogistic reasoning. In *Proceedings of ESSLLI student session.*

Active control of study leads to improved recognition memory in children

Azzurra Ruggeri (a.ruggeri@berkeley.edu)
Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany and
Department of Psychology, University of California, Berkeley, USA

Douglas B. Markant (markant@mpib-berlin.mpg.de)
Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Todd M. Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, New York University, New York, USA

Fei Xu (fei.xu@berkeley.edu)
Department of Psychology, University of California, Berkeley, USA

Abstract

This paper reports an experiment testing whether volitional control over the presentation of stimuli leads to enhanced recognition memory in 6- to 8-year-old children. Children were presented with a simple memory game on an iPad. During the study phase, for half of the materials children could decide the order and pacing of stimuli presentation (active condition). For the other half of the materials, children observed the study choices of another child (yoked condition). We found that recognition performance was better for the objects studied in the active condition as compared to the yoked condition. Furthermore, we found that the memory advantages of active learning persisted over a one-week delay between study and test. Our results support pedagogical approaches that emphasize self-guided learning and show that even young children benefit from being able to control how they learn.

Keywords: active learning, recognition memory, exploration, metacognition, inquiry learning, cognitive development.

Introduction

Research in both psychology and education has argued that the opportunity to exert active control over what is experienced during learning can lead to improved outcomes as compared to more passive forms of instruction (see Bruner, Jolly, & Sylva, 1976; Gureckis & Markant, 2012; Montessori, 1912; Piaget, 1930). In particular, past research has highlighted the important role of active control in cognitive development (Held & Hein, 1963). Self-guided learning is particularly interesting to consider from a developmental perspective because it requires the coordination of a range of cognitive processes including decision making, metacognition, attention, memory, and learning. However, it is currently unknown whether active control is associated with improved episodic memory during development, specifically early childhood. In this paper we explore the effects of active learning on episodic memory in 6- to 8-year-old children.

Recent experimental evidence with adults shows that active control of the learning experience can lead to improvements in episodic memory for objects (Harman, Humphrey, & Goodale, 1999; Voss, Galvan & Gonsalves 2011; Voss, Gonsalves, Federmeier, Tranel, & Cohen 2011; Voss, Warren, Gonsalves, Federmeier, Tranel, & Cohen 2011), faces (Liu, Ward, & Markall, 2007), and in spatial learning tasks (Meijer & Van der Lubbe, 2011; Plancher, Barra, Orriols, & Piolino, 2013; for a review see Markant, Ruggeri, Gureckis, & Xu, 2016), as compared to conditions lacking this control. Most of the studies investigating the benefits of active control for episodic memory adopt "yoked" experimental designs involving pairs of learners. In each pair, one person is the *active* participant, who controls the flow of information during learning (e.g., selecting what to study and for how long) and the other is the *yoked* participant, who observes the learning experience generated by the active participant. By matching the content experienced during study across conditions, yoked designs isolate the effects of active decision making on learning and memory. For example, Voss et al. (2011a; 2011b; 2011c) presented adult participants with a memory task involving a set of objects arranged in six 5×5 grids, with only one object visible at a time through a moving window. Participants alternated between active study blocks and yoked blocks. In active study blocks, participants controlled the study sequence and timing by deciding how to move the window, whereas in the yoked blocks they observed the study sequence a previous participant had generated in an active study block. Participants were then tested on two different components contributing to accurate memory encoding: Their recognition memory of the studied objects (whether they had been studied before or not) and their ability to recall the locations on the grid where the objects were presented during study. The results showed an advantage for the active study condition for both object recognition and spatial recall that persisted a week after the initial study session (Voss et al., 2011a). The authors also showed that the benefits of active encoding were influenced by participants' study patterns. Objects studied for longer durations (Voss et al., 2011a) and revisited more than once within a short period of time (Voss et al., 2011c) were recognized more accurately, but only in the active learning condition (i.e., recognition of objects studied in the yoked condition did not correlate with study time or with revisits).

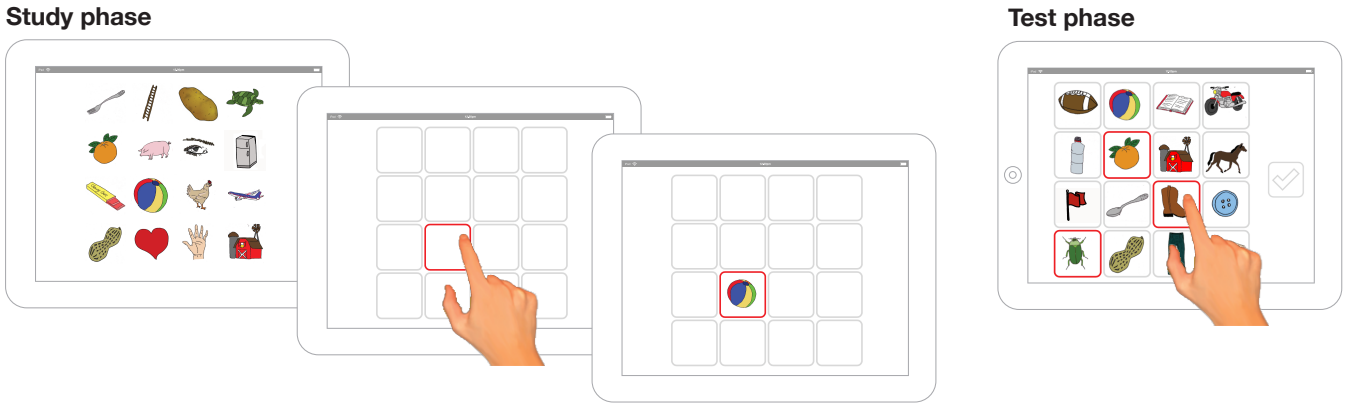


Figure 1. Each study round began with the objects displayed for two seconds. After the objects disappeared, the participant either selected a location to study (*active* condition), causing a red frame to appear, followed by the object, or clicked on the location where the object appeared (*yoked* condition), anticipated by a red frame. During each test block, participants clicked on objects that were recognized from the study phase.

Active learners in the above studies might have benefited from multiple levels of control over the study experience. For example, active study entails control over the content (i.e., choosing which object to study next), as well as the timing of the study sequence (i.e., when to move the window and for how long to study an object). In a series of experiments that varied the amount of control experienced during active study, Markant et al. (2014) found that active exploration (i.e., decisions about how to navigate the grid) was not necessary for the advantage from active study to emerge. Simpler forms of control (for example, merely controlling the timing of presentation for new objects) also led to a benefit in recognition as compared with yoked observation of the same study sequences, suggesting that multiple levels of active control contribute to improvements in episodic memory. Moreover, the authors showed that the active learning advantage for spatial recall was relatively inconsistent across different versions of the memory task, and their results did not support the correlation between recognition memory performance and objects visited more often or revisited after a short period.

The present study compares the effects of active and yoked study on episodic memory in 6- to 8-year-old children, using a variant of the task from Markant et al. (2014). Based on some of the previous literature, we expected that active control of study would generally improve learning outcomes in children. Indeed, Sim et al. (2015) showed that 7-year-olds learn more effectively when they are allowed to make decisions about what information they wish to gather, as compared to yoked observations. Previous findings also suggest that episodic memory, in particular, may benefit from the opportunity to actively control the learning process even at an early age. For example, active navigation has been shown to lead to memory improvements by age five (Feldman & Acredolo, 1979; McComas, Dulberg, & Latter, 1997; Poag, Cohen, & Weatherford, 1983), and self-directed learning has been shown to enhance short-term memory retention for novel object-word pairings in 3- to 5-year-old children (Partridge, McGovern, Yung, & Kidd, 2015).

However, there are also contradictory views in the literature. In particular, the benefits of active learning might crucially depend on children's use of metacognitive process to control study, as well as on their ability to implement successful studying strategies. Previous work suggests that the ability to allocate study time based on the difficulty or familiarity of the material develops over the course of childhood (Dufresne & Kobasigawa, 1989; Metcalfe, 2002; Metcalfe & Finn, 2013). For instance, although 6-graders demonstrate sensitivity to the strength of their own memories (Metcalfe, 2002), they are inefficient in controlling how long to study particular items in order to achieve the best level of recall. Given these conflicting results it is an open question whether active control would lead to benefits in episodic memory of 6- to 8-year-old children.

Method

Participants

Participants were 29 6- to 8-year-old children (15 female, M_{age} = 89.92 months; *SD* = 8.31 months), recruited in Berkeley, California, from a participants database. Due to technical difficulties, the data of three additional children were not recorded and therefore could not be included in the analyses.

Materials

Because some children might not have been familiar with some of the objects included in the original set of stimuli used by Voss et al. (2011a; e.g., accordion, chisel), we developed a new set of stimuli. Our set consisted of 192 line drawings of the most frequent objects mentioned by children younger than 5-year-olds in their everyday

conversations, as recorded by the CHILDES database (MacWhinney & Snow, 1985), which includes transcripts of children's natural speech collected over many years.

Design and procedure

The experimental materials were presented as a simple game where children were tasked with remembering as many of the presented objects as possible. The design and procedure were modeled after Experiment 2 in Markant et al. (2014). However, we made several modifications to the previous design to make it suitable for children. First, all the stimuli were presented on an iPad touchscreen instead of a computer screen. To select objects, children did not have to use a mouse, but could touch the objects on the screen directly. We added two familiarization blocks at the beginning of the first experimental session, aimed at introducing the goal of the game and the study procedures, and at making children comfortable playing with the tablet. To reduce the total testing time and the general cognitive load experienced by children, the main experimental session consisted of two *active* and two *yoked* study blocks (four total blocks, instead of six as in the original study), presented in alternating order (i.e., active, yoked, active, yoked). The active block was always presented first, so that children's initial active study pattern would not be influenced by the study pattern observed in the yoked condition. Each study block included 16 pictures representing different objects (see Materials), arranged on a 4×4 grid (instead of 25 pictures arranged in a 5x5 grid as in the original study), so that, in total, children were asked to memorize 64 pictures.

In contrast to the original design, all 16 objects were shown on the screen for 2 seconds at the beginning of each study block, before disappearing under occluders (see Figure 1, left). In the active blocks, children had 90 seconds (instead of a minute, as in the original design) to study the 16 objects in order to memorize them. To study an object, the child touched the corresponding occluder button once. A red frame appeared for 500ms, followed by the removal of the occluder that would reveal the hidden object. Before studying another object, the child had to touch the current object once more to make it disappear behind the occluder. In each of the yoked blocks, children were presented with the 90-second video showing the same objects and study pattern of one of the previous children's active learning blocks, and had the same task (to memorize the objects). To keep their engagement and attention level comparable to the active blocks, during yoked blocks children clicked on the objects as soon as they appeared during the video, although this click had no effect on the display. As in the active blocks, a red frame preceded each object for 500ms so that children had time to allocate their attention to the new study location before the object appeared. At the end of each block there was a twenty second break in which children were briefly reminded of the study procedure for the next block.

The study phase was immediately followed by a test phase, comprised of 8 blocks. In each test block, 16 objects were presented in a 4×4 grid, as in the study blocks (see Figure 1, right). Across the 8 test blocks, 64 objects were *old* objects the children had studied, and 64 were *new* objects that were not presented during study. The proportion of old objects in each block was randomly determined and all objects were arranged in random locations in the grid. For each block, participants indicated the objects they had studied earlier by touching them on the screen. Selected objects were framed in red (see Figure 1, right) to help participants keep track of the objects selected as recognized. The children could deselect any of the previously selected objects by clicking them again and making the red frame disappear. When finished selecting objects, participants were prompted to click on a button to proceed to the next test block, until the last test block was completed. Children were not given any feedback about their performance during or after the test phase. Note that, to shorten the testing time, the test phase was radically different from that of Markant et al. (2014), where participants were presented with individual objects and asked whether they were "Definitely OLD," "Probably OLD," "Probably NEW," and "Definitely NEW." Moreover, participants in our study were not tested on their spatial recognition memory.

After about one week (range 6 to 15 days; M = 8.45 days; SD = 1.93 days), children came to the lab for a second experimental session in which they were asked to complete 8 new test blocks. The 64 objects studied in the first session were randomly mixed with 64 *new* objects (i.e., objects that were not presented during the previous session, neither as study nor as test objects), again placed in random locations in the grid. The testing procedure was identical to the test phase from the first session.

Results

Results were analyzed with respect to: (1) the number of objects recognized among the ones studied; (2) the *false alarms*, that is, the number of objects recognized in the test blocks that had not been presented in the study blocks; and (3) the correlations between study experience and performance, to test whether certain participants' exploration strategies and patterns lead to better recognition accuracy. In particular, we examined the correlation between the recognition accuracy for a certain object and the time spent studying it, as well as the number of times it had been visited during study. We also examined the correlation between participants' average recognition accuracy and the distance between subsequent study locations (that is, the average distance on the grid between the object currently visible and the one selected next), a basic measure of how systematically a child explored the grid. Finally, although we did not test participants' spatial recognition directly, we analyzed the correlation between the location on the grid of the objects in the study phase and in the test phase, to investigate whether participants were

more likely to recognize an object when it was presented in the same location as in the study phase.

Because six children did not come to the lab for the retest session, we analyzed the data using mixed model ANOVAs with study condition (2 levels: active versus yoked) and session (2 levels: test versus one-week-later retest) as within-subject variables. We also analyzed the data using a univariate between-subjects ANOVA, considering the yoked pairs (i.e., comparing children's yoked study conditions with the active study conditions of the participants they were yoked to). Because we found no differences between these two sets of analyses, we only report the results of the mixed model ANOVAS.

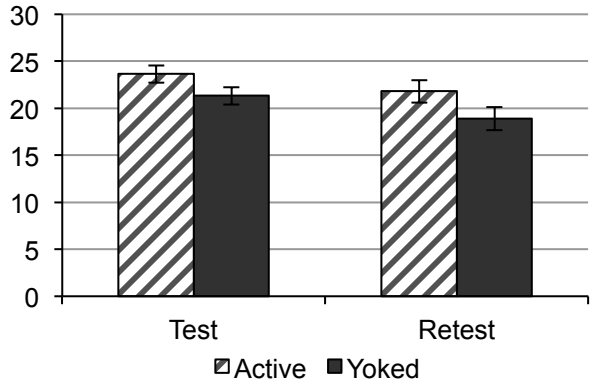


Figure 2. Number of objects correctly recognized in the test trials, displayed by study procedure (active vs. yoked) and session (test vs. retest).

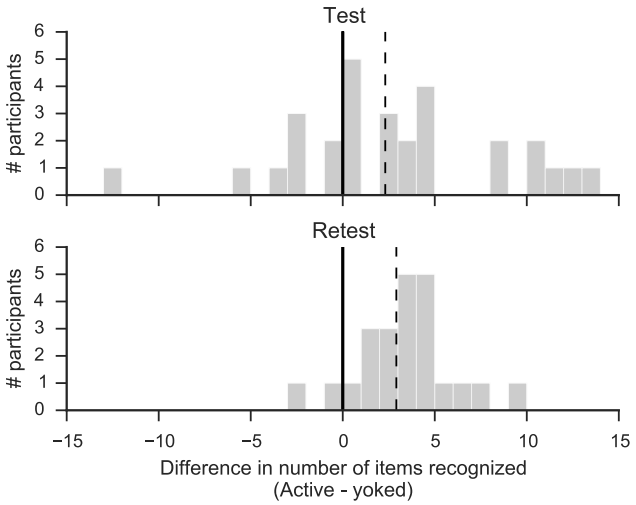


Figure 3. Distribution of within-subject differences in hit rate (active versus yoked) for studied items in the immediate test (top) and retest following a week delay (bottom). Dashed lines indicate the average difference.

Recognition of studied objects. On average, participants in the active study condition studied 30.45 (*SD* = .66) of the 32 objects presented (i.e., 95%).

The key analysis reveals a main effect of study condition, $F(1, 100) = 6.15$, $p = .015$. Children recognized more of the objects studied in the active study condition ($M_{active} = 22.81$; $SD = 5.31$) as compared to the objects studied in the yoked study condition ($M_{yoked} = 20.23$; $SD = 5.47$; see Figure 2), a 10% difference. The distributions of within-subject differences are shown in Figure 3. We also found a main effect of session, $F(1, 100) = 4.13$, $p = .045$. Children recognized more of the objects studied in the first test session ($M_{test} = 22.47$; $SD = 5.05$) as compared to the one-week-later retest session ($M_{retest} = 20.33$; $SD = 5.90$; see Figure 2). There was no reliable interaction effect between study condition and session ($p = .775$).

False alarms. We did not find a main effect of session on the number of objects recognized in the test blocks that had not been presented in the study blocks, $p = .122$. However, in general participants made more false alarms in the retest ($M_{retest} = 5.91$; $SD = 7.25$) as compared to the first test session ($M_{test} = 3.41$; $SD = 4.07$).

Correlations between study experience and performance. We found that object recognition accuracy was positively correlated with the time spent studying an object, as well as with the number of times the object had been visited, for both test and retest and for both the active and the yoked study conditions (see Table 1).

Active study condition				
Test	Correlations between tests			
	1	2	3	4
1. Accuracy in test				
2. Accuracy in retest	.459**			
3. Study duration	.212**	.174**		
4. Number of visits	.216**	.092*	.511**	
5. Distance from study position	.036	.013	.011	0
Yoked study condition				
Test	Correlations between tests			
	1	2	3	4
1. Accuracy in test				
2. Accuracy in retest	.448**			
3. Study duration	.139**	.142**		
4. Number of visits	.163**	.058	.412**	
5. Distance from study position	-.02	-.015	-.023	-.035

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

Table 1. Correlations between the recognition accuracy in test and retest, time the objects were studied, number or times they were visited and distance between the position in which the object was presented on the study grid and its position on the test, for objects presented in the active (top) or yoked (bottom) study condition.

However, we found no correlation between recognition memory and the average distance between subsequent study locations (that is, the distance on the grid between the object currently visible and the one selected next), neither in the active nor in the yoked conditions, for either test or retest (ps > .1). Finally, we found that recognition accuracy was not correlated with the distance between the location in which the objects were presented on the study grid and their location on the test grid.

Discussion

The present study examined whether active control of what to study (specifically, of when, for how long and how many times to process an object to be memorized) leads to advantages in memory encoding for 6- to 8-year-olds. Using a memory task modeled after Markant et al. (2014), we replicated most of the results previously found with adults.

First, children's episodic memory is more accurate for objects studied in the active control condition as compared to the yoked condition where children could merely observe the active study pattern of a previous participant. Note that the yoked experimental procedure allows controlling for study content and timing, which are identical across the two conditions. The magnitude of the advantage of active control for memory encoding (10% increase over the yoked-study condition) is comparable to the effect found with adults across several versions of the same recognition task (6% to 10%; see Markant et al., 2014), suggesting that such benefits are stable across development, in addition to being robust across different versions of the same tasks (see Voss, 2011a; Markant & Gureckis, 2014).

Second, we found that the benefits of active control for episodic memory persist a week after the study session (see also Voss et al., 2011a). Future work will be needed to assess if the advantage of active learning increases over longer delays between study and test, as seems to be suggested by our current results (M_{test} = .91; SD = 5.49; M_{retest} = 2.82; SD = 2.59).

Third, we found that episodic memory is influenced by children's study patterns. Objects studied for a longer time and objects visited more often were recognized more accurately, even after a week. However, different from Voss et al. (2011a; 2011c) and consistent with the results of Markant et al. (2014), we found that the correlations are present for the objects studied in the yoked condition as well: Studying objects more than once and for longer time generally led to a better memory encoding, without such study patterns being more beneficial in the active control rather than in the yoked study condition. The presence of correlations in both conditions suggests that, as in Markant et al. (2014), attentional cueing (that is, the appearance of a red frame anticipating the presentation of the next object) might be enough to extend the benefits of longer study to the yoked condition.

Finally, we found that recognition accuracy is not correlated with the distance between the location in which the object was presented on the study grid and its location on the test grid. Having the objects presented in the same location on the grid across the study and test blocks does not help recognizing them more accurately, neither in the active nor in the yoked study condition. These results might speak, though indirectly, against a robust active learning advantage for spatial recall (see Markant & Gureckis, 2014). However, only a direct test of spatial memory would allow confirmation of this hypothesis.

Although these results are largely consistent with adult behavior, it could be that an advantage from active control emerges only later on in child development, possibly as a result of formal education. To explore this hypothesis, we are currently replicating this study with 5-year-olds. The data we collected so far (N = 12) suggest that the active control of study does not lead to advantages in episodic memory for preschoolers (Test: M_{active} = 21.66; SD = 5.03; M_{yoked} = 22.75; SD = 4.41; Retest: M_{active} = 17.00; SD = 6.63; M_{yoked} = 18.80; SD = 8.69). Although preschoolers seem to perform as well as older children in the test session, their episodic memory declines faster than older children.

To explore these developmental differences further, we plan to develop new versions of the memory task used in this study in order to identify the underlying cognitive processes responsible for the memory improvement across different age groups. A number of mechanisms may mediate the effects of active control on episodic memory (Markant, Ruggeri, Gureckis, & Xu, 2016). In particular, it is important to investigate whether the advantage of active learning for memory encoding depends on the efficiency of children's study strategies and metacognitive decision making (and therefore being possibly linked to formal education and schooling), or whether it persists when such processes do not play a prominent role. For instance, Partridge et al. (2015) compared active and passive performance in a word-learning task, in which active control entailed selecting items from a grid to learn their labels, whereas passive learning involved observing items in a predetermined order. Active choice was associated with improved accuracy in an immediate test, even though the number of study events and study time was constant across conditions. Moreover, it is crucial to examine more thoroughly the role of attention and motivation on the active learning benefit for memory encoding.

In conclusion, in this paper we demonstrated that active control of study leads to advantages in memory encoding for 6- to 8-year-old children. These results have general implications for informing educational practice, which is increasingly incorporating the model of inquiry-based learning, by helping develop more generalizable insights into the effective implementation of active learning in educational settings.

Acknowledgments

We thank Celina Vicuna, Sana Alimohamed, Lesley Blair Winchester, Mandana Mostofi, Sarah De La Vega and Minh-Thy Nguyen for assistance in data collection and coding, as well as Susana Herrera and Kritika Shrestha for

drawing and coloring the stimuli. This research was supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme to AR and by grant number BCS- 1255538 from the National Science Foundation, the John Templeton Foundation "Varieties of Understanding" project, and a John S. McDonnell Foundation Scholar Award to TMG.

References

Bruner, J., Jolly, A., & Sylva, K. (1976). *Play—Its role in development and evolution.* New York: Basic Books.

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology, 47*(2)*,* 274–296.

Feldman, A., & Acredolo, L. (1979). The effect of active versus passive exploration on memory for spatial location in children. *Child Development*, *50*(3), 698–704. doi:http://dx.doi.org/10.2307/1128935

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science, 7*, 464–481. doi:10.1177/1745691612454304

Harman, K. L., Humphrey, G. K., & Goodale, M. A. (1999). Active manual control of object views facilitates visual recognition. *Current Biology, 9,* 1315–1318. doi:10.1016/S0960-9822(00) 80053-6

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Perception: An Adaptive Process, 56*:182.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 609–622. doi: 10.1037/0278-7393.32.3.609

Liu, C. H., Ward, J., & Markall, H. (2007). The role of active exploration of 3d face stimuli on recognition memory of facial information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4):895.

MacWhinney, B., & Snow, C. E. (1985). The child language data exchange system (CHILDES). *Journal of Child Language, 12,* 271-294.

Markant, D., DuBrow, S., Davachi, L., & Gureckis, T. M. (2014). Deconstructing the effect of self-directed study on episodic memory. *Memory & Cognition, 42*(8):1211–1224.

Markant, D., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Paper in revision.*

McComas, J., Dulberg, C., & Latter, J. (1997). Children's memory for locations visited: Importance of movement and choice. *Journal of Motor Behavior*, *29*(3), 223–229.

Meijer, F., & Van der Lubbe, R. H. (2011). Active exploration improves perceptual sensitivity for virtual 3D objects in visual memory. *Vision Research, 51,* 2431–2439. doi:10.1016/j.visres.2011.09.013

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131,* 349–363. doi:10.1037/0096-3445.131.3.349

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*(1): 19–46.

Montessori, M. (1912/1964). *The Montessori Method*. Schocken, New York.

Partridge, E., McGovern, M., Yung, A., & Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In Dale, R., Jennings, C., Maglio, P., Matlock, T., Noelle, D., Warlaumont, A., and Yoshimi, J., editors, *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Piaget, J. (1930). *The child's conception of physical causality.* New York: Harcourt, Brace

Plancher, G., Barra, J., Orriols, E., & Piolino, P. (2013). The influence of action on episodic memory: A virtual reality study. *Quarterly Journal of Experimental Psychology, 66,* 895–909. doi:10.1080/ 17470218.2012.722657

Poag, C. K., Cohen, R., & Weatherford, D. L. (1983). Spatial representations of young children: the role of self- versus adult-directed movement and viewing. *Journal of Experimental Child Psychology*, *35*(1), 172–9.

Sim, Z. L., Tanner, M., Alpert, N. Y., & Xu, F. (2015). Children learn better when they select their own data. In Dale, R., Jennings, C., Maglio, P., Matlock, T., Noelle, D., Warlaumont, A., and Yoshimi, J., editors, *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Voss, J., Galvan, A., & Gonsalves, B. (2011a). Cortical regions recruited for complex active-learning strategies and action planning exhibit rapid reactivation during memory retrieval. *Neuropsychologia, 49,* 3956–3966. doi:10.1016/j. neuropsychologia.2011.10.012

Voss, J., Gonsalves, B., Federmeier, K., Tranel, D., & Cohen, N. (2011b). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nature Neuroscience, 14*, 115– 120. doi:10.1038/nn.2693

Voss, J., Warren, D., Gonsalves, B., Federmeier, K., Tranel, D., & Cohen, N. (2011c). Spontaneous revisitation during visual exploration as a link among strategic behavior, learning, and the hippocampus. *Proceedings of the National Academy of Sciences, 108*, E402– E409. doi:10.1073/pnas.1100225108

Cohesive Features of Deep Text Comprehension Processes

Laura K. Allen (LauraKAllen@asu.edu)
Matthew E. Jacovina (Matthew.Jacovina@asu.edu)
Danielle S. McNamara (DSMcnama@asu.edu)
Department of Psychology, P.O. Box 872111
Tempe, AZ 85281 USA

Abstract

This study investigates how cohesion manifests in readers' thought processes while reading texts when they are instructed to engage in self-explanation, a strategy associated with deeper, more successful comprehension. In Study 1, college students (n = 21) were instructed to either paraphrase or self-explain science texts. Paraphrasing was characterized by greater cohesion in terms of lexical overlap whereas self-explanation included greater lexical diversity and more connectives to specify relations between ideas. In Study 2, adolescent students (n = 84) were provided with instruction and practice in self-explanation and reading strategies across 8 sessions. Self-explanations increased in lexical diversity but became more causally and semantically cohesive over time. Together, these results suggest that cohesive features expressed in think alouds are indicative of the depth of students' comprehension processes.

Keywords: text comprehension, self-explanation, cohesion, think-aloud

Introduction

Comprehension is a complex cognitive activity that involves the processing of information for the purpose of extracting meaning (McNamara & Magliano, 2009). Within the context of text comprehension, this process relies on the interplay between both lower-level and higher-level processes. Lower-level processes, such as word decoding, relate to the ability to understand the surface-level attributes of a text. However, comprehension does not simply occur once these processes have taken place. Rather, deep comprehension relies on a reader's ability to understand and make connections among the multiple concepts that are activated as a result of these lower-level processes.

The outcome of these comprehension processes is referred to as the mental representation. This representation contains information that was explicitly provided in the text, outside information related to the text, and inferences generated during the comprehension process. Readers achieve deep comprehension when they make connections among these information sources and develop a coherent mental representation of the text (see McNamara & Magliano, 2009, for a review). Importantly, the *coherence* of this mental representation is established through the activation of prior knowledge (from within and outside the text), the incorporation of this knowledge into the mental representation, and the development of connections among the propositions within the mental representation.

Many or most of these coherence-building comprehension processes occur *online* – in other words, they are enacted at the same time that individuals are reading the text (Kintsch, 1988). Consequently, to better understand how readers establish coherence in their mental representations, these online processes must be identified and examined. Unfortunately, most text comprehension assessments only occur after reading is complete. While these assessments can measure recall and recognition of key concepts, they often fail to detect cognitive processes associated with comprehension (Magliano, 2007).

To develop a better understanding of these online reading processes, researchers commonly use think-aloud methodologies (Pressley & Afflerbach, 1995). Generally, think-aloud instructions are designed to prompt individuals to generate the thoughts that are currently in working memory and thus easy to express (Ericsson & Simon, 1994). These methodologies have allowed researchers to identify a number of strategies that readers use to comprehend texts, such as paraphrasing and bridging (McNamara, 2004; Millis, Magliano, & Todaro, 2006).

An important goal for comprehension researchers has been to identify factors that differentiate skilled and less skilled readers. Research has identified a number of individual differences, such as prior knowledge (McNamara et al., 1996), writing ability (Allen et al., 2014), and motivation (Hidi & Harackiewicz, 2000). Relevant to the current study, research suggests that skilled and less skilled readers differ in their *strategic processing* of texts. Skilled readers generate more inferences while reading, which allows them to connect text information to prior knowledge (Oakhill & Yuill, 1996). Further, skilled readers typically establish connections at a more global level, whereas less-skilled readers tend to generate connections in more local contexts (one or two sentences; Millis et al., 2006).

Researchers have additionally *manipulated* the instructional prompts in think-aloud studies to encourage students to engage in different cognitive processes while reading (e.g., Allen, McNamara, & McCrudden, 2015). For example, self-explanation is a commonly employed instructional strategy that encourages the generation of inferences during reading. Broadly, self-explanation is the process of generating explanations to oneself about a particular topic; this strategy has been shown to improve students' deep understanding of complex concepts (Chi et al., 1989; McNamara, 2004). When students produce quality self-explanations (either spontaneously or following an instructional prompt), they make inferences that link text

content together and tie text ideas to their prior knowledge. Thus, self-explanation instructions can prompt students to behave as skilled readers. By contrast, instructions to generate surface level responses to a text (e.g., paraphrasing) do not necessarily prompt students to make connections that are characteristic of skilled readers.

We consider these connections in the reader's mental representation to be representative of *coherence* because research suggests that successful readers generate more inferences (connections) and those inferences are associated with enhanced comprehension. Hence, coherence is a construct that is indirectly inferred about the mental representation of the reader, generally from comprehension tests, but also based on computational models that simulate readers' performance (e.g., Kintsch, 1988).

Cohesion, on the other hand, refers to the incidence of explicit text cues that facilitate readers in making connections among ideas in texts (Gernsbacher, 1990; McNamara et al., 2014). For instance, words that overlap between sentences signify to the reader that the sentences are related. It has been assumed that cohesion cannot be *observed* in the mind of the reader, and that the elements of cohesion that might be related to coherence would vary with individual differences and the task requirements, among other factors (O'Reilly & McNamara, 2007).

Our goal here is to challenge that assumption.

Current Study

In the current study, we examine the extent to which elements of cohesion detected within the (expressed) thoughts of readers are indicative of the type of coherence-building processes in which readers are engaging. We adopt a multi-step methodological approach that relies on natural language processing (NLP) techniques to investigate the cohesion of students' think-alouds and self-explanations. In Study 1, we manipulate the think-aloud instructions students are given while reading texts, instructing them to either paraphrase or self-explain. We then extract the linguistic indices related to the cohesion of these different types of think-alouds that we hypothesize will be indicative of particular strategies. In Study 2, we examine how the different cohesive properties of students' think-alouds change over the course of self-explanation strategy training. The purpose of this second analysis is to examine the extent to which strategy training and practice that prompts students to engage in deeper text processing is evident in the forms of cohesion established in their self-explanations.

Cohesion in text can be established in a number of ways. For the sake of simplicity, we refer to three forms of cohesion that differ in the way in which they express relationships among ideas in texts. *Lexical cohesion* is cohesion that is established through overlapping words in a text. For instance, a text that repeats similar words will have higher lexical cohesion. *Causal cohesion* is signaled by overlapping actions (verbs) and connectives that serve to describe explicit connections among people, objects, events, and actions. Finally, *semantic cohesion* emerges from relationships among concepts, without relying on specific word overlap. For instance, two paragraphs that describe doctors and nurses, respectively, would still be *semantically* cohesive, even if they had no overlapping words.

In this study, we examine the extent to which these three forms of cohesion manifest in students' think alouds when they are engaged in tasks that lead to more coherent mental representations. To this end, we utilize an NLP tool, Coh-Metrix, to investigate the forms of cohesion present in students' think-alouds. Coh-Metrix (McNamara et al., 2014) calculates a number of linguistic indices, ranging from lower level word information to higher level information about cohesion. We selected Coh-Metrix indices that were representative of lexical cohesion (lexical overlap, lexical diversity), causal cohesion (connectives, causal ratio), and semantic cohesion (givenness, global cohesion).

Lexical overlap measures the degree to which words and phrases overlap across sentences. We selected the argument overlap index, which specifically measures overlap between nouns, noun phrases, and pronouns in adjacent sentences.

Lexical diversity is based on type-token ratios, which increase when the words in a text are less repetitive. In Coh-Metrix, lexical diversity is calculated using multiple algorithms that control for text length. Here, we used the *D* measure (Malvern, Richards, Chipere, & Durán, 2004). Lower lexical diversity is indicative of greater cohesion. In this context, higher lexical diversity can be associated with bringing in more information from outside of the text, or prior knowledge, when talking about the text.

Connectives (because, therefore) specify relationships between ideas and provide information about the properties of those relationships. Coh-Metrix provides the incidence of connectives per 1000 words in a text.

Causal ratio is calculated as the ratio of causal verbs to causal particles (McNamara et al., 2014). The causal verb count is calculated with WordNet and the causal particle count is based on a pre-defined set of causal verbs. A higher causal ratio is associated with greater causal cohesion.

Givenness measures the amount of semantic information that can be recovered from earlier in a text. In Coh-Metrix, givenness is calculated using LSA, which is a statistical method that uses large corpora to represent semantic knowledge (Landauer et al., 2007).

Global semantic cohesion is also analyzed using LSA. We utilized a paragraph-to-paragraph LSA cosine, which reflects the semantic similarity of paragraphs to each other.

Study 1

Study 1 investigates whether the form of cohesion established in readers' think-aloud responses differ when they are prompted to self-explain or paraphrase a text. Paraphrases of text focus primarily on the targeted sentence, and thus place a less emphasis on thinking about the relationships to other ideas or sentences in the text. By contrast, self-explanations are intended to create links to prior knowledge and between ideas and sentences in the text. As such, we expected paraphrases to be characterized

by higher lexical cohesion and a lower diversity of ideas. We also expected students to emphasize the causal connections between ideas more so when self-explaining than when paraphrasing.

Method

Participants Data from this study was collected as part of a larger study that examined neural correlates of strategic reading comprehension (Moss et al., 2011). Participants were undergraduate students (n = 21) recruited from large universities in the Northeastern United States. The majority of the participants were female (n = 14), with a mean reported age of 20.7 (*SD* = 2.4; range = 18-28). All participants were native speakers of English.

Procedure Participants completed two sessions, which occurred 2-5 days apart. The first session consisted of a pretest and training with a self-explanation tutoring system, iSTART (described below). During the second session, participants spent 30 minutes practicing the self-explanation strategies in iSTART. The purpose of the self-explanation tutoring during these two sessions was to ensure that all of the participants were familiar with the strategies and had practiced using them while reading. Participants then read three separate Biology texts and, for each text, they were provided instructions to engage in rereading, paraphrasing, or self-explaining. Thus, each participant performed all three of the strategies. The order in which participants performed the strategies was randomized and the assignment of reading strategies to texts was counterbalanced across the participants.

iSTART

iSTART is an automated version of the SERT (McNamara, 2004) intervention (McNamara, Levinstein, & Boonthum, 2004). It is an intelligent tutoring system (ITS) that provides high school and college students with training and practice on reading comprehension strategies to improve self-explanation and comprehension of complex texts.

iSTART training is separated into three modules, which map onto the three principles of modeling, scaffolding, and fading instruction. After training, students interact with the practice module, where direct instruction is faded and students are required to engage more deeply with the self-explanation strategies. Here, students are asked to self-explain target sentences in science texts and a teacher agent provides feedback and prompts the use of other strategies.

Text Reading Procedure Each text was divided into three text sections that consisted of four paragraphs each. These sections were presented to students one at a time. Because each text was assigned a specific reading strategy, participants were never asked to switch strategies within a particular section. Each of the four-paragraph text sections was presented before a section of another text. For instance, a student might self-explain the first text section of the text on DNA, then reread the first text section of a text on Heat, and finally paraphrase the first text section on the text about Cells. In the next section, the student would self-explain the second text section on DNA, and proceed accordingly. The texts were divided in this manner so that no reading strategy was performed more than once in each of the trials in order to control for confounding effects, such as fatigue.

Corpus To prepare students' think-aloud statements for text analysis, we first aggregated the self-explanations and paraphrases into two files (this method is discussed in greater detail in Varner et al., 2013). Paragraph breaks were added to each of the aggregated files for each of the blocks in the trial (i.e., to preserve the three separate text sections). Ultimately, this yielded two aggregated files per student: one paraphrase file and one self-explanation file.

Computational Analysis of Text Cohesion Students' paraphrase and self-explanation files were analyzed using Coh-Metrix. For the purposes of the current study, we selected Coh-Metrix indices that were representative of text cohesion: lexical overlap, lexical diversity, incidence of connectives, causal ratio, givenness, and global cohesion.

Results

Our first research question regarded whether students' self-explanations and paraphrases differed in their explicit markers of cohesion. A repeated-measures MANOVA was conducted to examine the differences in the cohesion indices across the aggregated paraphrase and self-explanation files (see Table 1 for descriptive statistics). This analysis revealed that there was a main effect of task instructions (i.e., paraphrase vs. self-explain) on the cohesion of think-alouds, F (6, 15) = 9.24, $p < .001$. Paraphrases were characterized by higher lexical cohesion, both in terms of higher lexical overlap, F (6, 15) = 11. 26, $p = .003$, and lower lexical diversity, F (6, 15) = 18.94, $p < .001$. Paraphrases also included somewhat greater semantic cohesion as measured by givenness, F (6, 15) = 3.20, $p = .089$. By contrast, more connectives were included in the self-explanations, F (6, 15) = 4.41, $p = .049$, indicative of greater specification of relationships between ideas. There were no differences between conditions in terms of the causal ratio or global semantic cohesion.

Table 1: Descriptive Statistics [Means and (SD)] for Paraphrase and Self-Explanation Conditions

Index	Paraphrase	Self-Explain
Lexical Overlap	**0.77** (0.17)	0.63 (0.16)
Lexical Diversity	50.84 (13.92)	**60.82** (9.75)
Connectives	88.76 (24.20)	**101.26** (15.48)
Causal Ratio	1.75 (1.85)	1.19 (0.85)
Givenness	**0.49** (0.05)	0.46 (0.04)
Global Cohesion	0.65 (0.09)	0.64 (0.13)

Discussion

The results from Study 1 indicate that cohesion manifested in think-aloud statements differently as a function of

whether students were instructed to paraphrase versus self-explain text. In particular, when readers were prompted to self-explain segments of the text, they used more diverse language and more connectives. However, given these same instructions, they also exhibited lower overlap among ideas, and less frequently referred to ideas that they had previously discussed. Self-explanation aims to promote the generation of inferences and, ideally, the establishment of connections between ideas in the text and prior knowledge of the world. When asked to self-explain, readers used more diverse words and less redundant information, suggesting that they may have been activating and thus referring to information from outside the text. Additionally, the self-explanation instructions prompted readers to utilize a greater number of connectives. This may suggest that when readers activated their prior knowledge, they used connectives to ensure that they explicitly established the nature of the relationships between the concepts they were discussing.

Study 2

Study 1 provides preliminarly evidence that the form of cohesion established in readers' think-aloud statements can serve as a proxy for the processes involved in deep comprehension. One possibility is that these findings are simply a by-product of readers being explicitly told when to use each of the strategies. Therefore, an important question relates to the extent to which these different cohesion indices are indicative of coherence-building processes over the course of extended practice.

The participants in Study 1 were relatively skilled, adult readers who were provided with brief self-explanation training with iSTART. Many readers, however, struggle to self-explain well (McNamara, 2004), and require extended self-explanation practice (Jackson & McNamara, 2013). Without thorough training and practice, less skilled readers are more likely to engage in shallow cognitive processes, which do not stimulate new connections among text concepts. Hence, our goal in Study 2 was to examine how cohesion manifests in the think-alouds of less-skilled (i.e., adolescent) readers and whether these aspects of cohesion change over the course of extensive training and practice.

The data in Study 2 were collected as a part of a larger study that investigated the impact of self-explanation training on readers' ability to generate high-quality self-explanations (Jackson et al., 2013). Results from this study confirmed the benefit of this training, revealing higher self-explanation scores over the course of the extended practice sessions. Our goal in the current study is to conduct a linguistic analysis of the self-explanations that students produced during these training sessions. The purpose of this analysis is to examine the cohesive devices that are related to students' comprehension processes and to determine whether these cohesion indices can provide critical information about the deep comprehension processes developed by readers during this self-explanation training.

Method

Participants Participants were 84 high school students from a mid-south urban environment (51% male; 81% African American, 13% White, 6% other; average grade completed = 10th grade; average age 15.8 years). All participants were monetarily compensated for their time.

Procedure The current study took place over 11 sessions, with a pretest, 8 training sessions, a posttest, and a delayed retention test. Students completed a pretest during the first session, which included measures of their reading and self-explanation ability. During the following eight sessions, students received training and practice in the iSTART system. This study focuses on the self-explanations generated by the students during practice.

Computational Analysis of Text Cohesion. Students' self-explanations in the iSTART systems were analyzed in a similar manner as in Study 1 with two notable exceptions. First, the aggregated self-explanations preserved the paragraph structure of the texts in iSTART. For a target text with p paragraphs and n target sentences, the resulting aggregated self-explanation file would contain p paragraphs and n self-explanations corresponding to the relative position of the target sentence. This is because iSTART prompted students to self-explain at specific target sentences, whereas in Study 1, students self-explained or paraphrased entire text segments. The second difference relates to the calculation of the cohesion indices. Students in this study self-explained multiple texts per day. Therefore, for each student, we calculated an average score for each cohesion index on each day of training.

Results

Separate repeated-measures ANOVAs for each cohesion index across the 8 training days were conducted to investigate the effect of self-explanation training on the cohesion of students' self-explanations. Table 2 presents the descriptive statistics for the first and last session of training. The results revealed that there was a significant effect of training session for all six cohesion indices. Over the course of iSTART training, students produced self-explanations that had lower lexical cohesion, both in terms of lower lexical overlap, F (1, 83) = 3.47, p = .005, and greater lexical diversity, F (1, 83) = 4.23, p = .001. Although the students decreased in their use of connectives, F (1, 83) = 9.90, p < .001, the causal ratio increased, F (1, 83) = 3.27, p = .026, indicating that the students were using connectives that were linked to causal verbs. While the self-explanations included less lexical cohesion (as found in Study 1), they were more semantically cohesive, including more given information, F (1, 83) = 6.99, p < .001, and an increase in global semantic cohesion across training, F (1, 83) = 11.99, p < .001. This suggests that students improved in establishing connections among the self-explanations produced across texts (rather than simply paraphrasing individual sentences).

Table 2: Descriptive Statistics [Means and (SD)] for Sessions 1 and 8 of Self-Explanation Practice

Index	Session 1	Session 8
Lexical Overlap	**0.76** (0.20)	0.61 (0.23)
Lexical Diversity	50.94 (13.73)	**54.29** (12.72)
Connectives	**101.31 (16.42)**	85.03 (17.86)
Causal Ratio	0.83 (0.74)	**0.96 (1.40)**
Givenness	0.35 (0.68)	**0.37 (0.09)**
Global Cohesion	0.46 (0.18)	**0.62 (0.11)**

Note: Analyses included all 8 sessions as a repeated measure

Discussion

The results revealed important information about the role of cohesion in students' self-explanations. The self-explanation training provided by iSTART promoted changes in students' use of cohesion across their self-explanations. In particular, over time, students' self-explanations became less cohesive lexically, but more cohesive at the global, semantic level. The explanations also included fewer connectives, but increased in terms of the causal ratio. The causal ratio is indicative of the use of connectives that are tied to causal verbs, and thus higher causal cohesion. These results indicate that self-explanation training prompted students to generate more inferences and establish more connections across the texts they were reading. This finding is important as it provides further confirmation that self-explanation training and practice can promote changes in students' on-line comprehension processes. Further, and most relevant to the current study, the results indicate that these coherence-building comprehension processes can be identified (at least in part) through automated analyses of the forms of cohesion in students' think-alouds.

Importantly, however, the changes observed in the cohesion of students' self-explanations across training do not directly map onto the findings from Study 1. In particular, the indices that were positively associated with self-explanation instructions in Study 1 did not necessarily increase over the course of iSTART training in Study 2. In Study 1, lexical overlap and givenness were higher for paraphrases than self-explanations, whereas connectives and lexical diversity were greater for self-explanations compared to paraphrases. In Study 2, lexical overlap and lexical diversity changed in a manner that was "consistent" with the results of Study 1, in so far as the cohesive devices become more indicative of deeper processing over time.

The incidence of connectives and givenness, however, behaved inconsistently with Study 1. These results potentially point toward important differences between the contexts of the two studies. In Study 1, skilled readers were explicitly prompted to engage in different reading strategies before generating any text. In Study 2, however, less-skilled readers were practicing a host of self-explanation strategies over extended practice. Therefore, these students were likely increasing in their use of certain deep comprehension strategies, but in some contexts, may have also needed to engage in "shallow" text processing, such as paraphrasing.

General Discussion

The current study investigated whether the cognitive processes associated with deep comprehension manifested in the cohesive properties of students' think-aloud statements. The results confirmed this prediction. In particular, cohesive indices of students' think-aloud statements differed according to task instruction and changed across time as students received self-explanation training. These results suggest that deep comprehension processes can be detected through analyses of readers' typed, verbal responses while reading texts.

The results from Study 1 indicated that prompting students to engage in shallow text processing (i.e., paraphrasing) or deep processing (i.e., self-explaining) led them to establish different levels of cohesion in their think-aloud statements. In particular, when students self-explained texts (as compared to paraphrasing), they used more diverse information and established more explicit connections among the ideas. This finding is important for a number of reasons. First, it provides further confirmatory evidence for the fact that instructions to either self-explain or paraphrase a text can dramatically alter students' on-line reading processes. Second, the results suggest that these instructional differences can be detected through analyses of the cohesion found in students' think-aloud statements. Thus, the coherence-building processes important for text comprehension may manifest in the overt cohesive cues students use when reading through the text.

Study 2 investigated whether the benefits of iSTART training could be detected through analysis of the cohesion of readers' self-explanations. The results suggested that all of the cohesion indices changed across training days and that the majority became more *consistent* with deeper levels of processing (as evidenced by the results of Study 1). In particular, after training, students produced self-explanations that were less lexically cohesive, but more causally and semantically cohesive. In particular, their self-explanations contained less explicit lexical overlap, with greater semantic connections established across the statements. These results suggest that changes in coherence-building comprehension processes can be identified by investigating the forms of cohesion in self-explanations.

Of course, this study is only an initial step in answering our questions. First, additional studies will be necessary to examine the relationship between these cohesive cues and comprehension more directly by examining students' comprehension on specific texts that they have self-explained. In the current study, we were interested in the specific influence of instructional prompts on these cohesive cues; however, future studies investigating how these cues relate to comprehension on specific question types will be necessary. Second, further research is needed to examine the generality of these effects across different types of texts and different types of comprehension goals.

Overall, the current study takes an important step towards understanding the role of cohesion in students' think-alouds during text comprehension. These findings can strengthen

our theoretical understanding of text comprehension processes, as well as for comprehension more broadly. Additionally, the results may be used to inform educational reading interventions and tutoring systems. If specific cohesion indices can be identified that systematically relate to certain comprehension processes and outcomes, educators may be able to use this information to provide more adaptive instruction and feedback to their students.

Acknowledgments

This research was supported in part by the Institute for Educational Sciences (IES R305A130124) and National Science Foundation (NSF REC0241144; IIS-0735682). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES or NSF.

References

Allen, L. K., McNamara, D. S., & McCrudden, M. T. (2015). Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. Maglio, *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Pasadena, CA.

Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., & McNamara, D. S. (2014). Reading comprehension components and their relation to the writing process. *L'année Psychologique/Topics in Cognitive Psychology, 114*, 663-691.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, R., & Glaser, R. (1989). Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge: MIT.

Gernsbacher, M.A. (1990). Language comprehension as structure building. Hillsdale, NJ: Erlbaum.

Hidi, S., & Harackiewicz, J.M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 2,* 151-179.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036-1049.

Jackson, G. T., Varner, L. K., Boonthum-Denecke, C., & McNamara, D. S. (2013). The impact of individual differences on learning with an educational game and a traditional ITS. *International Journal of Learning Technology, 8*, 315-336.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95,* 163–182.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah: Erlbaum.

Malvern, D. D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* Basingstoke, UK: Palgrave Macmillan.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourses Processes, 38,* 1–30.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36,* 222-233.

McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In J. D. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60-81). Mahwah, NJ: Erlbaum.

Millis, K. K., Magliano, J. P., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10*, 251–283

Moss, J., Schunn, C. D., Schneider, W., McNamara, D. S., & VanLehn, K. (2011). The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension. *NeuroImage, 58*, 675-686.

Oakhill, J., & Yuill, N. (1996). Higher order factors in comprehension disability: Processes and remediation. In C. Cornaldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 69–72). Mahwah, NJ: Erlbaum.

O'Reilly, T., & McNamara, D.S. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121-152.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale: Erlbaum.

Varner, L. K., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2013). Does size matter? Investigating user input at a larger bandwidth. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th Annual Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 546-549). Menlo Park, CA: AAAI Press.

Specificity at the basic level in event taxonomies: The case of Maniq verbs of ingestion

Ewelina Wnuk (e.wnuk@let.ru.nl)
Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands

Abstract

Previous research on basic-level object categories shows there is cross-cultural variation in basic-level concepts, arguing against the idea that the basic level reflects an objective reality. In this paper, I extend the investigation to the domain of events. More specifically, I present a case study of verbs of ingestion in Maniq illustrating a highly specific categorization of ingestion events at the basic level. A detailed analysis of these verbs reveals they tap into culturally salient notions. Yet, cultural salience alone cannot explain specificity of basic-level verbs, since ingestion is a domain of universal human experience. Further analysis reveals, however, that another key factor is the language itself. Maniq's preference for encoding specific meaning in basic-level verbs is not a peculiarity of one domain, but a recurrent characteristic of its verb lexicon, pointing to the significant role of the language system in the structure of event concepts.

Keywords: basic level; categorization; events; verbs; Maniq; Aslian.

Introduction

How shall an event be called? Paralleling the research question famously posed by Roger Brown (1958) for objects, an analogous problem can be raised for events. Is there a default most salient level of abstraction commonly applied to refer to events? For objects, such a default level has been termed the *basic level*, i.e. the level that "anticipates the equivalences and differences" relevant in most "dealings" with particular objects (Brown, 1958, p. 16). A label at the basic level is thus one that brings forward "an immediately important property of the referent" (Brown, 1958, p. 17). For example, the default way of referring to spoons is by the word *spoon*, rather than a more general label such as *eating implement,* or a more specific one such as *dessert spoon*. Basic-level categories are said to be optimal in the sense that they contain maximal information without being too specific, as this would lead to excessively fine-grained categorizations. In addition, basic labels are associated with cognitive advantages and are typically preferred in communication (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

Particularly revealing for the basic-level theory is the observation regarding variation as to what concepts are actually found at the basic level. Contrary to the idea that basic categories reflect some kind of objective reality and are thus invariable, there is cross-cultural variation in basic-level concepts (Dougherty, 1978). For example, some cultures have expertise in biological categories of plants and animals, and therefore treat the genus level (e.g. "pine") as basic, while other communities of speakers, e.g. some Western societies, show less interest in those categories and use superordinate labels (e.g. "tree") at the basic level (cf. also Tanaka & Taylor, 1991 for variation *within* a culture). Such variation is important as it constitutes evidence for culture-specific constraints on categorization (cf. Malt, 1995; Malt & Majid, 2013).

While there is ample evidence for a basic level for object concepts, there has been relatively little work investigating other ontological categories. For instance, we have comparatively limited knowledge on the basic level for events. The existing studies suggest that event hierarchies, just like object hierarchies, can have a basic-level structure, with one level of abstraction being more salient than others (Morris & Murphy, 1990). However, it has been pointed out that events have a more complex structure when compared to objects. Crucially, events have less clear-cut boundaries than objects and can therefore be conceptualized in a larger number of different ways (cf. Gentner, 1982). This finds reflection in a greater variability of linguistic labels encoding events, i.e. verbs are said to be among the most cross-linguistically variable part of the lexicon in terms of denotation (Gentner, 1982; Talmy, 1985; Evans, 2011). Given the freedom in event conceptualization, we may ask: To what extent is the basic level of events – as reflected in common verbs – similar across cultures and how much does it vary?

It is legitimate to expect that the basic level in verbs – similar to nouns – will be a reflection of local preoccupations and expertise of particular communities. Thus, we are likely to find more specific basic-level event labels for events which are culturally salient. For instance, it is not surprising that there is an elaborate lexicon of basic-level harvesting verbs in Dogon (Heath & McPherson, 2009) or climbing verbs in Jahai (Burenhult, 2013; Schebesta, 1929, pp. 151–152), because these events are highly salient in these cultures. However, as will be shown here, in some cultures highly specific verbs at the basic level occur not only in narrow areas of specialization, but also domains of basic human experience. This article focuses on one such area – ingestion – in Maniq, an Austroasiatic language spoken in southern Thailand. The aim is to demonstrate that specificity at the basic level in verbs is not purely a reflection of cultural concerns, but also of the

language's *typological profile*, i.e. its consistent preference for making fine-grained distinctions in basic-level verbs.

In what follows, I first briefly introduce the Maniq society and give some basic facts about their language. I then provide a detailed case study of verbs of ingestion, demonstrating that ordinary references to ingestion events in Maniq involve the use of verbs with highly specific meanings. An examination of the broader lexicalization patterns reveals further that specificity of basic-level verbs is pervasive in Maniq and is part of the logic of the language, i.e. it follows a systematic lexicalization principle applying across verbal domains (cf. Wnuk, 2016; Wnuk & Majid, 2014).

The Maniq and their Language

The Maniq language belongs to the Northern Aslian branch of Aslian, a division within the Austroasiatic language family.

The Maniq are a population of about 300 people inhabiting the Banthad mountain range of southern Thailand. They live in small groups scattered across the provinces of Trang, Satun, Phatthalung and Songkhla. Maniq speakers belong to the larger ethnographic cluster of Semang nomadic populations. Their subsistence relies on hunting and gathering as well as small-scale trade of forest products. Nomadism is still practiced by a large proportion of the population, but today there are also Maniq groups that have settled and embraced agriculture and waged labor.

Ingestion verbs in Maniq

Ingestion is a domain of importance across human communities and ingestion verbs are high-frequency words in many languages. However, only some languages distinguish specific types of ingestion events with separate basic-level verbs (cf. Bowerman, 2005; P. Brown, 1998; Burenhult & Kruspe, 2016; Heath & McPherson, 2009; Rice, 2009). For instance, in English, the verb *eat* is a single default descriptor of eating actions. And although there exist a number of more specific 'eat' verbs encoding manner, e.g. *devour, gorge, gnaw, gobble* (Levin, 1993, pp. 213–216), these are not employed in neutral contexts, but are used only when the manner is somehow salient.

In contrast, in Maniq there are multiple specific ingestion verbs applicable at the basic level of contrast. These verbs are neutral, default ways of referring to ingestion events, frequent in everyday discourse and not restricted to special registers or groups of speakers (cf. Brown, 2008, p. 169). They are the preferred labels applied spontaneously in the free naming of ordinary scenes involving ingestion events, general statements about ingestion (e.g. *I've just eaten*), as well as translations of simple sentences from Thai involving the semantically general ingestion verb *gin* 'to eat or drink, to consume' (e.g. *The boy ate the fruit*).

In fact, Maniq does not seem to have separate general 'eat', 'drink' or 'consume' verbs. All mentions of eating/drinking actions involve one of several ingestion verbs with more specific meanings, selected based on manner of ingestion and/or type of ingested object (see further below). In cases, where these parameters are unknown, speakers typically employ the verb *hãw* (cf. Table 1), which is the most frequent ingestion verb with the broadest range of application. Note, though, that this verb is not a true generic term similar to *eat* and cannot be used in contexts where other verbs apply (cf. the verb *gey* in Jahai; Burenhult & Kruspe, 2016, p. 180). Lack of superordinate monolexemic labels is common in the Maniq verb lexicon and accords well with the idea of highly specific basic-level labels (though it is not always the case that basic-level verbs lack labeled superordinates).

The discussion below explores the distinctions in Maniq ingestion verbs. Of central focus is the question of what semantic principles of categorization apply in this set. It covers ingestion defined broadly as the absorption of various kinds of substances through the mouth. The discussion includes human ingestion verbs, which place selection restrictions on the consumed objects, and animal ingestion verbs, which place selection restrictions on possible actors (cf. Bowerman, 2005).

The data were collected during fieldwork carried out with a nomadic Maniq group of the Manang district, Satun province, during five field trips in the period 2009-2014. The group numbers approximately 15-20 members, but its size and composition are not fixed and changed somewhat over this period depending on external conditions such as food availability. The generalizations made here are based on observations of spontaneous uses of the verbs in everyday situations (e.g. with reference to actual ingestion events), informal interaction with speakers, as well as interviews in which speakers provided basic ingestion verbs for various ingested substances named by the interviewer. These included a range of typical objects foraged by the group (e.g. wild yams, fruit, game), as well as those obtained via exchange with outsiders (e.g. cultivated fruit, coffee, cakes). The most extensive interviews were carried out with three speakers (1 female) in the approximate age range of 25-45.

Human ingestion

Table 1 below contains a summary of human ingestion verbs in Maniq, including 'eat', 'drink' and 'inhale'-type predicates.

Table 1: Basic-level human ingestion verbs in Maniq.

Verb	Gloss	Example objects
hãw	'to chew, to eat mainly by chewing'	rice, non-fibrous yams, pineapple, cucumber, garlic, chili, papaya, sweet potato, nut of *Canarium* sp., durian, banana, jackfruit, petai beans, leafy plants, mushrooms, cempedak, baked goods (cookies, cakes, bread)

kap	'to bite, to eat harder objects that require biting'	meat, animal fat, fish, bones, hard fruit
lɨk	'to swallow, to eat with little biting or chewing'	orange, lemon, mango, grapes, wild gandaria fruit (*Bouea* sp.), various unidentified wild fruit species (e.g. *kabiʔ yəbac*)
paŋ	'to eat with spitting out hard fibers'	fibrous yams – especially *ciyak* (*D. cf. piscatorum*) and *ciyak lapɔn* (*D. laurifolia*)[1], sugar cane
hop	'to consume nutritious and/or savory liquids'	soup, honey
buʔ	'to drink non-nutritious liquids'	water, coffee, medicinal infusions, coconut water, honey
hɔp	'to inhale medicinal smoke (does not involve blowing out)'	smoke from burning dried roots of the medicinal plant *Dianella ensifolia*
yɔt	'to inhale and blow out smoke'	tobacco

Before discussing the semantic distinctions underlying this system, a few general comments on the structural properties of ingestion verbs are in order. The items listed in Table 1 are monomorphemic verbs with no formal relationship to one another. Rather than being derived from the same root, each verb stands on its own. The only case of formal similarity is the pair *hop-hɔp*, where the verbs share the underlying template *hVp*. However, the verbs are synchronically monomorphemic and cannot be decomposed into morphological units (cf. Wnuk, 2016). Semantic richness in Maniq is thus accompanied by formal non-transparency, as the multiple meaning components are packaged into non-analyzable forms.

Maniq ingestion verbs can occur in syntactically intransitive (1) and transitive (2) constructions.

(1) ʔiŋ bah lamah, ʔiŋ hãw
1S NEG be.hungry 1S chew
'I'm not hungry, I ate.'

(2) ʔiŋ hãw ʔanciʔ
1S chew yams
'I'm eating yams.'

Note, however, that even when the object is not overtly expressed, it is implied since a particular verb is associated with a restricted range of objects. Thus, for instance, although the speaker in (1) does not mention what he ate, the listener can restrict the possible objects to those fitting with the requirements of the verb *hãw* (cf. Table 1).

The categorization of ingestion events in Maniq is influenced by several factors. For most verbs in Table 1, the primary factor is the manner of ingestion. By indicating manner, ingestion verbs covertly classify objects, as they restrict the possible range of referents to those with the consistency or texture fitting with that manner (cf. Aikhenvald, 2009, p. 106; Heath & McPherson, 2009, p. 42). For instance, *lik* is a way of eating in which the most prominent part is swallowing since the food does not require much pre-mastication. The prototypical foodstuffs described with *lik* are thus softer types of fruit such as oranges, grapes, as well as a number of wild fruit species, e.g. *kabiʔ yəbac*. Note that in its basic sense *lɨk* denotes the action of swallowing and may be employed when no eating is involved, as in e.g. swallowing one's saliva or a pill. A similar pattern is attested with the verb *kap*, referring to biting, and *hãw*, referring to chewing, which are used to describe eating events in which biting and chewing, respectively, are the most prominent parts. Compare example (3), where *kap* describes a biting event, with (4), where it is used to describe an eating event.

(3) naʔ hay miʔ kc-kac
FOC like Maniq IMFV-scratch
ʔagɛs ʔɛʔ kap
mosquito 3 bite
'It's like when Maniq scratch after they've been bitten by mosquitoes.'

(4) ʔiŋ kap tawɔh paliek
1S bite gibbon be.white
'I eat white gibbons.' (uttered in a contrastive context, 'white gibbons' are here juxtaposed with 'black gibbons')

The fact that the basic set of ingestion distinctions for solid matter is primarily manner-based constitutes a departure from the basic pattern among some of Maniq's close relatives, i.e. other Aslian languages spoken by hunter-gatherer groups (e.g. Jahai, Batek Deq, Semaq Beri). In these languages, the main ingestion verbs are linked not to manner, but to the categorial identity of a food item. For instance, in Jahai *muc* 'to eat animal' maps onto all foods classified as *ʔay* 'edible animal', *gey* 'to eat starchy food' onto foods known as *bap* 'starchy food', *but* 'to eat ripe fruit' onto *bɔh* 'ripe fruit', and *hɛ̃w* 'to eat leafy greens' onto *tʔaʔ* 'leafy greens' (Burenhult & Kruspe, 2016, p. 180). In Maniq, the categorial identity of food items is of lesser importance and in fact no similarly elaborate system of generic food classes seems to be in place. The only large classes similar to those in other Aslian languages include *ʔay* 'game' and *kabiʔ* 'fruit and some vegetables'. There is no generic label for starchy foods, leafy vegetables, or a specific label for ripe fruit, and Maniq ingestion verbs do not indicate such a classification might be operating at a covert level. Except for food of animal origin associated consistently with the verb *kap* (and mostly mapping onto the *ʔay* category), most classes of foods are connected to several verbs, depending on which specific food item from that class is involved. The verb *kap* in fact represents a mixed

[1] All identified species of wild yams discussed here are from the genus *Dioscorea*. For convenience, the genus name is abbreviated to "*D.*". The identifications are based on Maniq vernacular names provided in Maneenoon, Sirirugsa, and Sridith (2008).

pattern, since its use is sometimes triggered by the categorial identity of ingested matter (e.g. in the case of meat), and sometimes by manner of ingestion (e.g. in the case of hard fruit).

Fruit and vegetables (*kabiʔ*) can be linked to different ingestion verbs, depending on which specific food item from that class is involved. Eating soft and juicy fruit like orange and gandaria, for instance, is typically described with the "swallow" verb *lik*. Eating fruit with a somewhat harder or mushy texture such as banana or pineapple is associated with the "chew" verb *hãw*. Finally, eating hard fruit, including not fully ripe fruit, and some vegetables such as cucumbers is associated with the "bite" verb *kap* (cf. Heath & McPherson, 2009, p. 43, for a similar set of distinctions in Dogon). In addition, there is some variation as to what verbs are preferred with what fruit. In particular, fruit with texture not clearly linked to one specific manner of ingestion (e.g. mango) tends to be used with several verbs, depending on the speaker or specific context.

Perhaps the most nuanced culturally-salient contrast among the Maniq ingestion verbs is the one between the verbs *paŋ* and *hãw*. Both verbs are used with several kinds of foods, but are associated most prominently with wild yams (*Dioscorea* spp.), the traditional staple food of the Maniq (Maneenoon et al. 2008). Underlying the distinction between them is the classification into fibrous and non-fibrous yam species. *Hãw* is used with the majority of the consumed types, which usually do not contain hard fibers, while *paŋ* is employed especially with *ciyak* '(*D. cf. piscatorum*)' and *ciyak lapɔn* '(*D. laurifolia*)'[2], which have hard woody fibers that are never swallowed. Apart from these two species, a number of others (e.g. *ləntak* '(*D. glabra*)') contain fibers of a softer type, which are varyingly described with either of the two verbs. The distinction between *paŋ* and *hãw* also roughly maps onto two main labeled types/sections of tubers – the cylindrically-shaped, slender and usually fibrous part called *tənat*, and the thicker and wider part called *bahiʔ*. The consumption of the *tənat* sections of tubers is usually described with the verb *paŋ*, while *bahiʔ* sections with the verb *hãw*. The covert tuber classification presupposed by these two verbs thus reveals fine details about the indigenous botanical knowledge structure.

While for solid objects, the verb choice is determined by manner of ingestion, for liquids it is based on the category of the ingested substance. Thus, the verb *hop* maps onto liquid substances collectively classified as *leŋ* 'nutritious and/or savory liquid substance' such as soup, meat juices, and honey. The verb *buʔ*, on the other hand, is reserved for non-nutritious/non-savory items, many of which fall under the generic label *batew* 'water, liquid' (e.g. water, medicinal infusions, coconut water), but also coffee. Honey has been noted to occur both with *hop* and *buʔ*. By being linked to specific classes of liquid substances, these verbs resemble the classificatory 'eat' verbs in other Aslian languages such as Jahai and Semaq Beri (Burenhult & Kruspe, 2016).

Finally, for volatile substances the choice of verb is determined strictly by manner. Smoking of a cigarette described as *yɔt* involves a quick inhale and a voluntary action of blowing out the smoke. In contrast, the verb *hɔp* describes the inhaling of smoke from the burning of dried roots of the medicinal plant called *kasay* '*Dianella ensifolia*'. Crucially, in the case of *hɔp* the inhalation is not followed by deliberate blowing out of smoke – the person performing the practice usually attempts to inhale deeply and keep the smoke in the lungs for as long as possible. One may also apply medicinal smoke locally on a particular body part, e.g. a leg. In this case, it is common to first ingest the smoke and then blow it on the ailing body part. This activity – like cigarette smoking – is described with the verb *yɔt*, which illustrates that manner, and not the type of smoke, is the primary factor determining verb choice.

To summarize, human ingestion verbs in Maniq are sensitive to a combination of parameters, including the manner of ingestion as well as categorial identity of ingested items, which in turn depend on various physical properties of those items, e.g. their texture, nutritiousness, and taste.

Animal ingestion

Although most discussions of ingestion verb lexicons focus on the elaboration linked to the ingested object, elaboration can also be linked to the agent performing the action of ingestion. In Maniq, such verbs are associated with some characteristic ways of ingestion typical of specific animals. In such cases, the identity of the agent is linked to the verb indirectly via the manner of ingestion. For instance, the verb *coh* 'to strike with a long object' is a conventional way of describing pecking and eating-by-pecking, and thus in the ingestion context, it conveys the action of eating associated with birds.

(5)	jawãŋ	ʔɛʔ	coh	kabiʔ
	great.hornbill	3	strike	fruit
	'Great hornbills eat fruit.'			

Note that example (5) is best translated with *eat* rather than *peck* since in English *peck* in generic statements of this kind would be marked. In Maniq on the other hand, *coh* is a basic-level predicate, the most unmarked and frequent item in such contexts. Table 2 lists ingestion verbs associated with particular animals together with glosses and their covert agents.

[2] *Ciyak lapɔn* is a medicinal yam species that is too toxic for consumption (Maneenoon et al. 2008). Note, though, that the Maniq do classify it as being *paŋ*-eaten, which suggests that either this is a hypothetical response (one that would apply if it was edible), or that there is a way of detoxifying it.

Table 2: Basic-level animal ingestion verbs in Maniq.

Verb	Gloss	Agent
coh	'to eat by pecking'	birds
ɲam	'to eat fruit, involves spitting out seeds, skin and pulp'	bats
dut	'to eat by sucking out liquid substance'	insects
lɔs	'to eat fruit/nuts (of squirrels and rats)'	squirrels, rats

The verb *ɲam* is similar to the human ingestion verb *paŋ* applied with fibrous tubers in that it involves discarding some of the masticated matter. In this case, however, the discarded elements include seeds, skin and pulp of fruit consumed by fruit bats (e.g. large flying fox (*Pteropus vampyrus*)). The verb *dut* (a loanword from Thai *dùut* 'to suck') describes sucking or eating-by-sucking, as in, for instance, bees sucking nectar. Finally, *lɔs* is associated with the particular manner in which some rodents such as squirrels eat fruit and nuts, and is therefore characteristically applied in descriptions of eating actions performed by these animals.

Since some animals bite, chew, drink, etc. in ways similar to people, many human ingestion verbs have been attested with animals, too. In particular, *hãw* 'to chew-eat', *kap* 'to bite-eat' and *buʔ* 'to drink non-nutritious liquid' are used frequently with animal agents, cf. (6) and (7).

(6) yəkɔp ʔɛʔ kap basiŋ
snake 3 bite-eat dusky.leaf.monkey
'Snakes eat dusky leaf monkeys.'

(7) kaɲcɛh ʔɛʔ hãw kabiʔ
bearcat 3 chew-eat fruit
'Bearcats eat fruit.'

Summary and conclusions

The system of basic-level ingestion verbs described here reveals a more fine-grained and nuanced categorization of ingestion events in Maniq than in many other languages, where only a general division between *eat* and *drink* is made (cf. Newman, 2009; Wierzbicka, 2009). Such specificity could be argued to be related to cultural factors. This would then constitute a case parallel to objects, where it was found that more specific labels at the basic level are a reflection of community's expertise in a domain (Dougherty, 1978). Such interpretation appears to be at least partially true since, aside from drawing on basic bodily mechanics, a number of ingestion verbs in Maniq reflect ethnobiological expertise of Maniq speakers. Their semantics involve culturally relevant ethnobiological knowledge (e.g. relating to the ingested objects, behavior of animals, etc.), and presuppose familiarity with this knowledge. While this provides an account for how ingestion events are categorized in Maniq, by itself it does not explain why specific rather than general labels are preferred. Further insights into this issue can be gained by taking into account the distinct meal habits of the community and the typical composition of meals. This has been suggested by Burenhult and Kruspe (2016) for related Aslian hunter-gatherer groups. According to this account, the existence of specific verbs is linked to the fact that meals in hunter-gatherer communities typically do not involve elaborate combinations of different foodstuffs, but are often instances of "opportunistic ingestion of a single resource" (Burenhult & Kruspe, 2016, p. 194). This would suggest further that "there is no culturally salient type of ingestion event for which a general concept or label "eat" seems necessary" (Burenhult & Kruspe, 2016, p. 194). In other Aslian languages, this has been connected specifically to the existence of one particular subtype of semantically specific ingestion verbs (i.e. food-category-encoding verbs), but it is conceivable that such distinct meal habits could facilitate a fine-grained categorization of ingestion events more generally.

Although culture-specific factors like salience and distinct meal habits are likely important, by themselves they might not constitute sufficient pressure for such lexical elaboration. Ingestion is a domain of basic human experience, and specific subtypes of ingestion events are likely salient in many cultures. Yet, semantically detailed ingestion verbs are found only in a subset of the world's languages. Similarly, although there is no extensive survey of ingestion verb paradigms across hunter-gatherer communities, not all such communities seem to have equally elaborate systems of ingestion verbs (e.g. Wierzbicka, 2009). Further analysis of the Maniq verb lexicon reveals, however, that another key factor is the language itself. Verb meanings do not exist in a vacuum, but form part of a system and often pattern in systematic ways (Gentner, 1982; Talmy, 1985). Hence, the characteristic ways in which languages encode verb meaning in one domain tend to recur in other domains too, revealing general principles underlying lexicalization. In the case of Maniq, and Aslian languages more generally, specificity is such a general principle. Aslian languages have often been noted to have a penchant for encoding specific meanings in monolexemic verbs occurring across multiple semantic domains (Kruspe, Burenhult, & Wnuk, 2015; Matisoff, 2003). For instance, in Maniq they are attested not only in ingestion, but in a number of other domains, e.g. perception (*balay* 'to look up'), location (*cibɛl* 'to be located upside down'), motion (*tik* 'to move upstream'), and transportation (*galɛs* 'to carry on back') (cf. Wnuk, 2016). This typological characteristic makes Maniq and Aslian stand out from many other languages of their linguistic area, placing them together with other languages and language groups characterized by a similar marked preference for verb specificity, e.g. Mayan (P. Brown, 2008). Thus, although the basic level for events is partially a reflection of cultural factors, it is also influenced by the semantic-typological profile of the language, suggesting a significant role for language in the structure of event concepts.

Acknowledgments

I would like to thank the Maniq community of the Manang district (Satun, Thailand) and Jaroon Thotsagool. This paper has benefitted from comments and discussions with Niclas Burenhult, Asifa Majid, Nicole Kruspe, Rebecca Defina, Ilja Croijmans, and Laura Speed. This work was supported by Max Planck Gesellschaft and The Netherlands Organization for Scientific Research.

References

Aikhenvald, A. Y. (2009). "Eating", "drinking" and "smoking": A generic verb and its semantics in Manambu. In J. Newman (Ed.), *The linguistics of eating and drinking* (pp. 91–108). Amsterdam ; Philadelphia: John Benjamins.

Bowerman, M. (2005). Why can't you "open" a nut or "break" a cooked noodle? Learning covert object categories in action word meanings. In L. Gershkoff-Stowe & D. H. Rakison (Eds.), *Building object categories in developmental time* (pp. 209–243). Mahwah, N.J: L. Erlbaum Associates.

Brown, P. (1998). Children's first verbs in Tzeltal: Evidence for an early verb category. *Linguistics*, *36*(4), 713–754.

Brown, P. (2008). Verb specificity and argument realization in Tzeltal child language. In M. Bowerman & P. Brown (Eds.), *Crosslinguistic perspectives on argument structure: Implications for learnability* (pp. 167–189). Mahwah, NJ: Erlbaum.

Brown, R. (1958). How shall a thing be called? *Psychological Review*, *65*(1), 14–21. http://doi.org/10.1037/h0041727

Burenhult, N. (2013). Categories of arboreal locomotion in Jahai (Austroasiatic, Malay Peninsula). Unpublished report.

Burenhult, N., & Kruspe, N. (2016). The language of eating and drinking: A window on Orang Asli meaning-making. In K. Endicott (Ed.), *Malaysia's original people : Past, present and future of the Orang Asli* (pp. 175–199). National University of Singapore Press.

Dougherty, J. W. D. (1978). Salience and relativity in classification. *American Ethnologist*, *5*(1), 66–80. http://doi.org/10.1525/ae.1978.5.1.02a00060

Evans, N. (2011). Anything can happen: The verb lexicon and interdisciplinary fieldwork. In N. Thieberger (Ed.), *The Oxford handbook of linguistic fieldwork* (pp. 183–208). Oxford; New York: Oxford University Press.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought, and culture* (pp. 301–334). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heath, J., & McPherson, L. (2009). Cognitive set and lexicalization strategy in Dogon action verbs. *Anthropological Linguistics*, *51*(1), 38–63. http://doi.org/10.1353/anl.0.0001

Kruspe, N., Burenhult, N., & Wnuk, E. (2015). Northern Aslian. In P. Sidwell & M. Jenny (Eds.), *Handbook of the Austroasiatic languages* (pp. 419–474). Leiden; Boston: Brill.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press.

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85–148. http://doi.org/06/cogp.1995.1013

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 583–597. http://doi.org/10.1002/wcs.1251

Maneenoon, K., Sirirugsa, P., & Sridith, K. (2008). Ethnobotany of Dioscorea L.(Dioscoreaceae), a major food plant of the Sakai tribe at Banthad Range, Peninsular Thailand. *Ethnobotany Research & Applications*, *6*, 385–394.

Matisoff, J. A. (2003). Aslian: Mon-Khmer of the Malay Peninsula. *Mon-Khmer Studies*, *33*, 1–58.

Morris, M. W., & Murphy, G. L. (1990). Converging operations on a basic level in event taxonomies. *Memory & Cognition*, *18*(4), 407–418. http://doi.org/10.3758/BF03197129

Newman, J. (2009). A cross-linguistic overview of "eat" and "drink." In J. Newman (Ed.), *The linguistics of eating and drinking* (pp. 1–26). Amsterdam; Philadelphia: John Benjamins.

Rice, S. (2009). Athapaskan eating and drinking verbs and constructions. In J. Newman (Ed.), *The linguistics of eating and drinking* (pp. 109–152). Amsterdam; Philadelphia: John Benjamins.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. http://doi.org/10.1016/0010-0285(76)90013-X

Schebesta, P. (1929). *Among the forest dwarfs of Malaya*. London: Hutchinson & Co., ltd.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description: Vol. III, Grammatical categories and the lexicon* (pp. 55–149). Cambridge: Cambridge University Press.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. http://doi.org/10.1016/0010-0285(91)90016-H

Wierzbicka, A. (2009). All people eat and drink. Does this mean that "eat" and "drink" are universal human concepts? In J. Newman (Ed.), *The linguistics of eating and drinking* (pp. 65–89). Amsterdam; Philadelphia: John Benjamins.

Wnuk, E. (2016). *Semantic specificity of perception verbs in Maniq* (PhD dissertation). Radboud University, Nijmegen. Manuscript in preparation.

Wnuk, E., & Majid, A. (2014). Revisiting the limits of language: The odor lexicon of Maniq. *Cognition*, *131*(1), 125–138. http://doi.org/10.1016/j.cognition.2013.12.008

Switch it up: Learning Categories via Feature Switching

Garrett Honke (ghonke1@binghamton.edu)
Nolan Conaway (nconawa1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University (SUNY)
Binghamton, NY 13905 USA

Abstract

This research introduces the switch task, a novel learning mode that fits with calls for a broader explanatory account of human category learning (Kurtz, 2015; Markman & Ross, 2003; Murphy, 2002). Learning with the switch task is a process of turning each presented exemplar into a member of another designated category. This paper presents the switch task to further explore the contingencies between learning goals, learning modes, outcomes, and category representations. The process of successfully transforming exemplars into members of a target category requires generative knowledge such as within-category feature correspondences – similar to inference learning. Given that the ability to switch items between categories nicely encapsulates category knowledge, how does this relate to more familiar tasks like inferring features and classifying exemplars? To address this question we present an empirical investigation of this new task, side-by-side with the well-established alternative of classification learning. The results show that the category knowledge acquired through switch learning shares similarities with inference learning and provides insight into the processes at work. The implications of this research, particularly the distinctions between this learning mode and well-known alternatives, are discussed.

Keywords: concepts; learning; categorization; category use

Introduction

In light of recent work establishing the effects of category use on conceptual representation, it has been argued that studying human category learning solely through the traditional artificial classification learning (TACL) paradigm provides a limited view of the processes involved (Chin-Parker & Ross, 2002; Levering & Kurtz, 2015; Love, 2002; Yamauchi & Markman, 1998; see Kurtz, 2015 for review). Specifically, category representations acquired through TACL are in many ways specially tailored for performing the classification task itself—a procedure that requires discriminating between a set of different categories based on the features of individual examples. Considering that real-world concepts commonly serve functions beyond class discrimination, there is clear importance for developing theories of human category learning that generalize across different types of learning goals and opportunities (Murphy, 2002).

In the spirit of evaluating how distinct category representations can arise under alternative learning conditions, we introduce the *switch task* – a new category learning technique based on transforming exemplars into members of another category – inspired by the DIVA account of human category learning (Kurtz, 2007, 2015). The chief goals of this paper are to introduce the switch task and present empirical data exploring the differences in category representation that arise as a result of this learning mode.

Research Motivation: The DIVA Model

DIVA (Kurtz, 2007) is an artificial neural network (ANN) model that uses a DIVergent Autoencoder architecture trained via backpropagation (Rumelhart, Hinton, & Williams, 1986) to learn and represent categories. The model is a concrete instantiation of a theory for how humans learn and represent categories – namely, that psychological categories are task-constrained, generative models of the regularities that exist among a category's members (Kurtz, 2015). One novel property of the model is that, within the context of a classification problem, category representations are not built independently (see Figure 1).

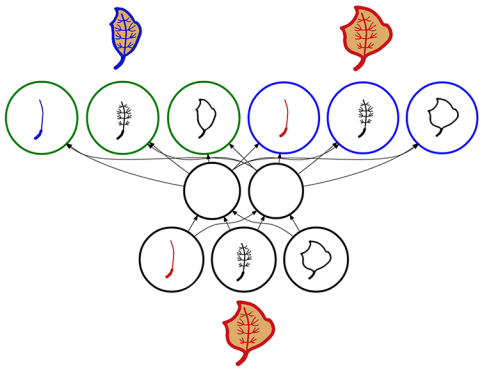


Figure 1: The DIVA model architecture with two possible reconstructions of a category exemplar. Stimuli are drawn from the present study.

The switch task is motivated by this unique design feature where each stimulus is reconstructed as a possible member of each category in the task. At the core of switch learning is an invitation to construe each presented example as a member of an alternate category, and then transform the example accordingly by changing its feature values as required. Although DIVA does not actually produce switch responses or learn categories by switching features, the key motivation of this work is to determine whether and with what properties the process of generating alternate categorical construals promotes category acquisition. To properly characterize how learning categories with the switch paradigm might create distinct category representations, the next section addresses the common properties of category knowledge after learning via classification.

The TACL paradigm

In the dominant TACL approach to studying human category learning, observers are presented with exemplars belonging to a set of experimenter-defined categories. For each observation, learners are asked to make a classification decision and are then provided corrective feedback. This guess-and-correct process is continued for a set number of iterations (or until a criterion is met) and then participants are tested on what they have learned.

An observed result of learning categories under this approach is that the task biases the resultant structure of conceptual representation (Solomon, Medin, & Lynch, 1999). In particular, individuals completing the classification task learn to select a category label based on an example with a particular set of features from among a set of alternatives. This has the effect of focusing the learner on features that have distinct values across the present categories (Chin-Parker & Ross, 2002). As such, learners need only to determine the features that are diagnostic for distinguishing between categories (when available).

Even small changes to the TACL paradigm can affect what is learned in the task. Levering and Kurtz (2015) have shown that removing the guess-and-correct component of the TACL paradigm increases the learning of within-category correlations — even when they are not useful towards predicting class membership. Likewise, asking learners to provide missing feature values (as opposed to class labels) promotes knowledge of within-category central tendencies (Markman & Ross, 2003; Yamauchi & Markman, 1998).

While the TACL approach has been favored for many valid reasons (notably, to reduce the scope of the problem and develop clear research questions), it also produces a fragmented view of the processes that underlie human category learning (Murphy, 2002, 2005; see Kurtz, 2015 for an in-depth exploration of these issues). As such, this project aims to widen current understanding of category learning task effects by presenting a comparison of two techniques: classification learning and switch learning.

The Current Study

In the present work, we primarily seek to evaluate the differences in conceptual representation fostered by the novel switch paradigm (fully described below) and traditional artificial classification learning. Specifically, questions of interest include whether either mode is better suited for different learning problems and different applications of category knowledge (i.e., inferring unobserved features, classifying partial items with features omitted).

We address these questions with an experiment consisting of a learning phase (switch or classification) and three test phases – switch, classification and inference tests. Participants learned about two categories of plant life, named *Lape* and *Tannet*, each consisting of four examples varying in three binary dimensions. The test phases included the eight examples from the learning phase plus partial exemplars. The classifications were three intermediate-difficulty category structures (Types II, III and IV; see Figure 2) from the elemental six-types problems (Shepard, Hovland, & Jenkins, 1961).

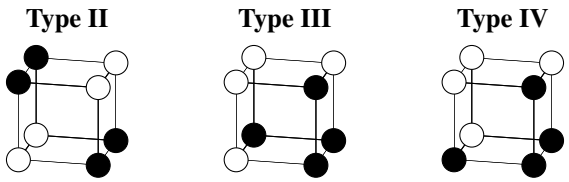


Figure 2: An instantiation of the experiment's category structures. The vertices represent individual exemplars with category marked by label and color; the spatial dimensions correspond to features of the stimuli.

The key difference between the switch task and other commonly studied learning modes is that participants are tasked with: 1) deciding which features of the example do not accord with the targeted category and 2) switching the binary values of these features so as to change the category membership. Participants use feature buttons to transform an initial example into a member of the target category. Each button shows an image of a feature value that is not currently present in the example. Examples are immediately updated on button press and the selected button is removed from the interface (the feature switch cannot be reversed). The only constraint on the switching task is that at least one feature must be switched to complete a trial. Accuracy feedback is provided at the end of each trial. The switch is considered accurate if the newly constructed example is a member of the target category.

Switch and classification training clearly differ in many respects – the central goal of this study is to evaluate the differences in conceptual representation conferred by two learning modes, as well as how these representations serve learners in putting their knowledge to work. If the process of viewing non-member exemplars as possible members of a target category helps people learn about underlying category structure, then the switch learning mode has the potential to be in some ways more effective than classification. Higher accuracy on test phases after learning would be clear evidence consistent with this hypothesis.

Alternatively, it is plausible that participants will be best at the test phase that mirrors their assigned learning condition. In accord with the construct of transfer-appropriate processing (Morris, Bransford, & Franks, 1977), this would suggest that classification learners will perform best on the classification test phase. This result would further validate a key idea guiding this research – that task conditions during category learning have a strong effect on category knowledge. Likewise, under this view, performance on the inference test phase has the highest importance: unlike the switch and classification tests, learners in both conditions have no experience completing inference trials in the context of our experiment.

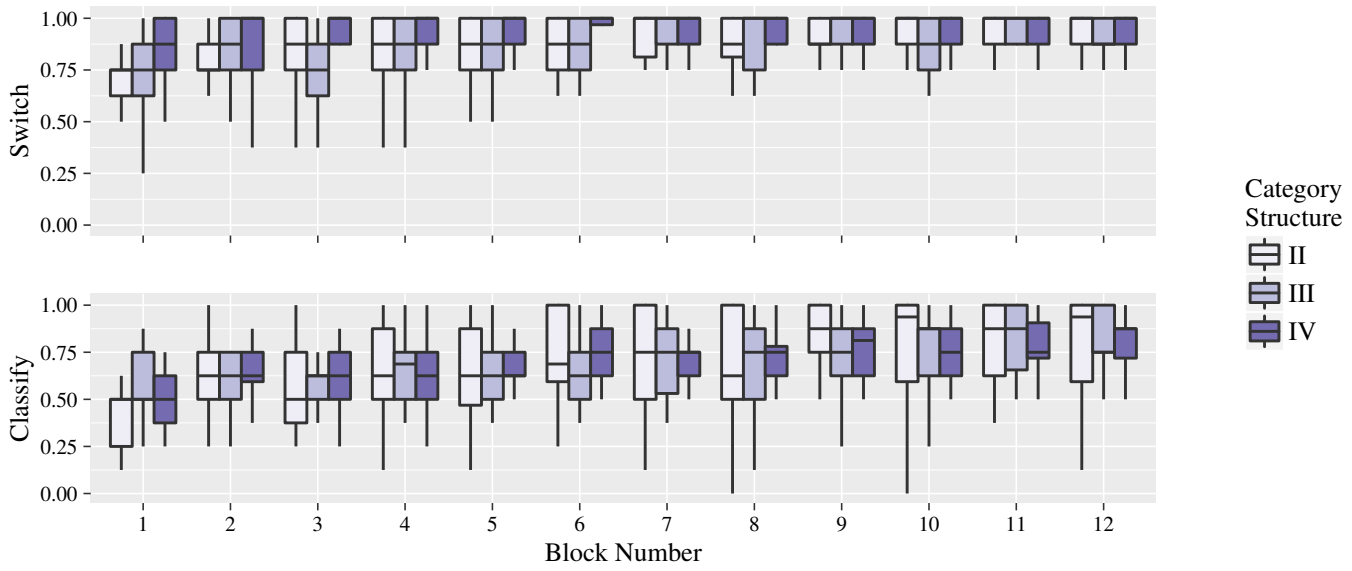


Figure 3: Mean percent correct for the training phase of the experiment. Switch learners were reliably better with Types III and IV as compared to Type II. No reliable differences were found within the classify group.

Method

Participants. All participants ($N = 170$) were recruited from the Binghamton University Psychology Department pool and randomly assigned to condition. Each participant gave consent to participate in writing and received credit towards the completion of a course requirement. Three participants were excluded from the analysis; two for failure to complete the experiment and one due to experimenter error.

Design and Procedure. There were two independent variables in the present study: learning condition and category structure. Three category types drawn from the Shepard et al. (1961) elemental category structures were included as a between-subjects variable. Learning condition was also a between-subjects variable – leading to a 2 (learning condition: classify, switch) x 3 (category structure: Type II, Type III, Type IV) design. The stimuli were leaf-like images (e.g., Figure 1) varying on three binary dimensions (color, veining, and shape). The assignment between conceptual and perceptual dimensions was counterbalanced across participants.

The experiment was conducted in private testing rooms on PCs with the use of a mouse and keyboard. The PsychoPy package was used for the development of the task interface (Peirce, 2007). Each participant was presented with instructions on screen. In the learning phase, participants completed 12 blocks of either classification or switch trials (8 trials per block) with feedback provided after every trial. Incorrect trials were repeated until a correct switch or classification was produced.

The goal for participants in the switch learning condition ($n = 84$) was to switch the features of provided examples until they matched a target category. The location of each feature button randomly varied by trial. After completing the transformation, participants used a Done button and then received feedback (see supplementary information[1] for a depiction of the switch interface).

Participants in the classification learning condition ($n = 86$) performed a task similar to that of the TACL paradigm except for the noted difference that incorrect trials were repeated until the correct category label was selected. On each trial, an example was presented and participants were asked to select the correct category label. After the classification decision was made, participants were given corrective feedback.

Three distinct test phases were included in the experiment: inference, classification and switch tests. All participants completed the inference test phase first and then the classification and switch phases (order randomly determined). On inference trials, learners were presented an incomplete example (one or two features omitted), a category label, and images of the two possible instantiations of one of the missing features as clickable buttons. Switch and classification test trials were identical to the training task. No feedback was provided during the test phases. The number of test trials in each phase depended on the category structure condition and test phase (see the supplementary materials for a depiction of the complete set of training and test examples, the task interface and all instructions[1]).

Results

The data was analyzed using generalized linear mixed effects regression (GLMER; Bates, Mächler, Bolker, & Walker,

[1] gist.github.com/ghonk/7e24c78a05280f61e866

2015) fit by maximum likelihood in the R analysis environment (R Core Team, 2016). The general analysis approach was to build regression models that predict trial success with learning condition (switch, classify), example type (complete training examples, incomplete novel examples) and their interaction. Study participants were included in the models as a random intercept.

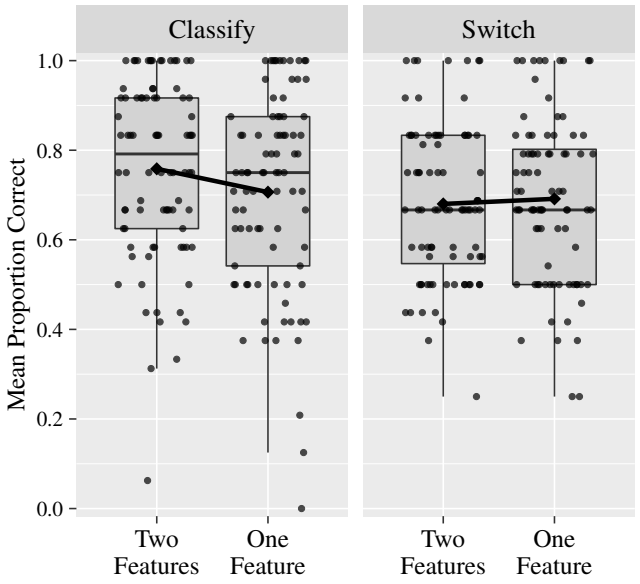


Figure 4: Mean proportion correct on the inference test phase. Tukey's boxplots present the overall accuracy pattern with black lines and diamonds indication mean differences and points representing individual means.

Training Phase Analysis. It is unclear whether the switch and classify tasks are equivalent in difficulty, prohibiting a direct comparison of training block accuracy. However, the relative difficulty of the different category structures can be examined within condition. Training accuracy was analyzed with GLMER where trial accuracy was predicted with the fixed effect of category structure and the random intercept of participant. The classify groups were not reliably different between category structures (replicating the core finding of Kurtz, Levering, Stanton, Romero, & Morris, 2013). In contrast, the switch group exhibited higher accuracy for the Type IV category structure as compared to Type II (Beta Estimate $= 0.684$, $SE = 0.24$, Wald $Z = 2.857$, $p = .004$) and Type III (Beta Estimate $= 0.687$, $SE = 0.24$, Wald $Z = 2.916$, $p = .004$) (see Figure 3).

Test Phase Analysis. To preview the test phase results, we first note that there was no reliable difference in accuracy between the learning conditions when collapsing across test phases. Breaking this down by test phase, the classify group was more accurate on inference and classification test trials and the switch group was more accurate on switch test trials. Perhaps most interestingly, learning condition and example type (partial versus full) interacted in the inference and classification tests: the switch group had higher accuracy on incomplete exemplars than on complete training examples, but the classify learners exhibited the opposite pattern. This pattern of results remains consistent when participants performing below chance are removed.

Inference Test. Inference test accuracy was analyzed as the dependent variable in a GLMER model with learning condition and example type included as interacting fixed effects and participant included as a random intercept. Overall, the switch group was less accurate than the classify group on the inference test trials (Beta Estimate $= -0.502$, $SE = 0.18$, Wald $Z = -2.736$, $p = .006$). Accuracy was worse on one-feature trials when collapsing across group (Beta Estimate $= -0.304$, $SE = 0.10$, Wald $Z = -3.08$, $p = .006$), but the interaction between learning condition and example type shows that this decrease in accuracy was not present in the switch group (Beta Estimate $= 0.38$, $SE = 0.14$, Wald $Z = 2.814$, $p = .005$) (see Figure 4).

Turning to the effect of category structure on accuracy, no reliable differences were found between category types for the classify group. The switch group, however, was more likely to be accurate if they were assigned Type IV as compared to Type II (Beta Estimate $= 0.551$, $SE = 0.24$, Wald $Z = 2.303$, $p = .02$).

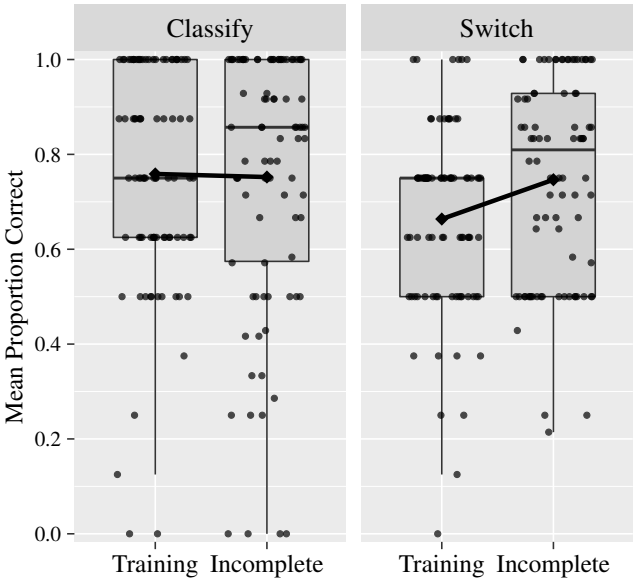


Figure 5: Mean proportion correct on the classification test phase split by condition and exemplar type.

Classification Test. The analysis approach for the classification test was similar to that of the inference test (save the dependent variable). The model uncovered reliable effects of learning condition and the interaction between learning condition and example type (See Figure 5). The switch group was less accurate overall (Beta Estimate $= -0.646$, $SE = 0.22$,

Wald $Z = -2.905$, $p = .004$). However, a significant interaction shows that the switch group was more accurate on incomplete trials than on training trials (Beta Estimate $= 0.503$, $SE = 0.18$, Wald $Z = 2.722$, $p = .007$) (see Figure 5).

Category type effects were also found for classification test. The switch group was less accurate on Type II compared to Type III (Beta Estimate $= -0.725$, $SE = 0.26$, Wald $Z = 2.805$, $p = .005$) and Type IV (Beta Estimate $= -0.762$, $SE = 0.26$, Wald $Z = 2.9$, $p = .004$). No category differences were found for the classify group.

Switch Test. Accuracy on switch test phase trials was analyzed with a GLMER model with learning condition included as a fixed effect and participant included as a random intercept. The results show that the switch group was more accurate on the switch test than the classify group (Beta Estimate $= 1.683$, $SE = 0.37$, Wald $Z = 4.576$, $p < .001$). No differences between category structures were found.

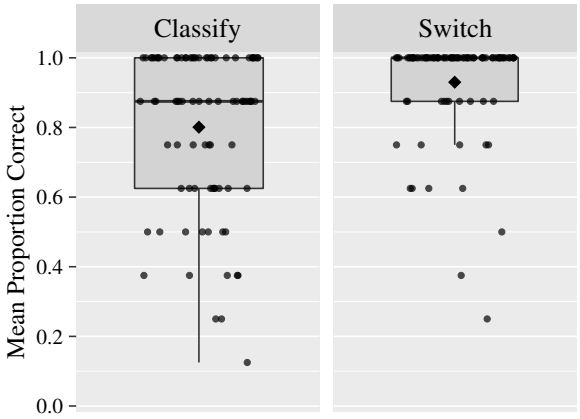


Figure 6: Mean proportion correct on the switch test phase.

Switch Test Characteristics. Given that the results have shown effects of category structure and exemplar type that are unique to the switch condition, it is of interest to characterize the switch patterns that participants used to learn the categories. These data can be compactly visualized in the style of a confusion matrix, containing the conditional probability of each possible switch, given a starting exemplar (i.e., 010 → 011). Although space does not permit a comprehensive display of this data (see the supplementary materials[1]), the switching patterns exhibited by the Type IV learners are especially illustrative (see Figure 7). These learners demonstrated a tendency to switch exemplars to the prototype of the target category. This pattern suggests that switch learners had acquired the knowledge that each class in this category structure is described by a family resemblance structure.

Discussion

We reported a novel categorization experiment exploring the differences between learning under a traditional classification learning mode and a novel switch task. Broadly speaking, the

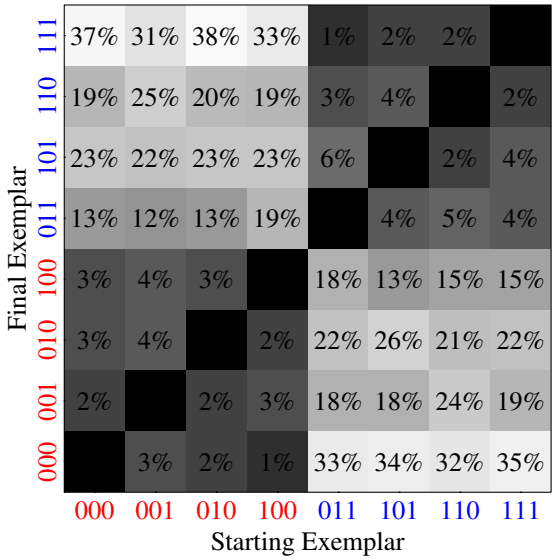


Figure 7: Confusion matrix depicting switch behavior during the Type IV training phase. Values represent the conditional probability of each switch, given a starting exemplar. Columns sum to 100%. Learners most commonly switched to the prototypes (000, 111).

results presented in this report are consistent with known effects of category use on category learning (Markman & Ross, 2003) – learners in the switch and classify conditions not only showed advantages on their respective tasks, but also differences in their inference responses to novel items of varying completeness.

In accounting for these effects, it is useful to consider the knowledge required to complete the switch task. Specifically, towards the goal of executing a successful class switch, learners need to understand how the present feature values relate to the target category, as well as how the current and target classes differ across the feature space. On its face, the advantage switch learners show on partial items (relative to complete items) may indicate that these individuals are better equipped to process items analytically – as a collection of parts rather than wholes. This interpretation is sensible given the nature of the switch task, where learners presumably become familiar with how elements of the exemplars relate to the class label. However, this would not explain the decline in performance on full items.

Past research on inference learning has uncovered higher (e.g., Sakamoto & Love, 2010) and lower (e.g., Sweller & Hayes, 2010) accuracy at test in relation to traditional classification. Considering the similarities between switch and inference, it is puzzling that the classification group was reliably more accurate on the inference test.

Still, there are many commonalities between inference learning outcomes and the results presented here. Switch learners had higher accuracy on family resemblance based categories as evidenced by the Type IV group's higher ac-

curacy on the training phase and the classification and inference test phases as compared to the Type II (non-linearly separable category structure) group. Similar advantages on family resemblance categories are found for inference as well (Yamauchi, Love, & Markman, 2002).

Switch learning produced higher accuracy on exemplars with features omitted as compared to complete exemplars while the opposite was observed in the classification learning group in our study. This result is mirrored in investigations of inference learning (Anderson, Ross, & Chin-Parker, 2002).

Another result of switch learning is that the family resemblance category structure was learned quite quickly. By evaluating the switch behavior during training (Figure 7) it can be seen that the prototypes are favored during this process – another commonality that can be tied back to a distinction made between inference and classification (Johansen & Kruschke, 2005; Yamauchi & Markman, 1998). Furthermore, the relative ease of learning across the category structures under the switch learning mode provides new evidence on the variable nature of the ordering of acquisition of the SHJ categories (Kurtz et al., 2013).

Given that real-world categories often conform to a family resemblance structure (Rosch & Mervis, 1975), future work will explore the sensitivity to within-category regularities and rapid learning of Type IV seen with switch-based learning.

Acknowledgments

We thank Kimery Levering for developing the stimuli used in this study, Sarah Laszlo and the members of the Binghamton University Modeling Meeting, and the members of the LaRC (Learning and Representation in Cognition) lab.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, *30*(1), 119–128.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, *30*(3), 353–362.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1433–1458.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. *Psychology of Learning and Motivation*, *63*, 77–114.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552–572.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & cognition*, *43*(2), 266–282.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological bulletin*, *129*(4), 592.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: MIT press.

Murphy, G. L. (2005). The study of concepts inside and outside the laboratory: Medin versus medin. In W. Ahn, R. Goldstone, B. Love, A. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab: Essays in honor of Douglas L. Medin* (pp. 179–195). Washington, DC: American Psychological Association.

Peirce, J. W. (2007). Psychopy – psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1), 8–13.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *Journal of Experimental Psychology: Applied*, *16*, 361–377.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in cognitive sciences*, *3*(3), 99–105.

Sweller, N., & Hayes, B. K. (2010). More than one kind of inference: Re-examining what's learned in feature inference and classification. *The Quarterly Journal of Experimental Psychology*, *63*(8), 1568–1589.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 585.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and language*, *39*(1), 124–148.

Episodic memory as a prerequisite for online updates of model structure

David G. Nagy[1,2], Gergo Orban [1]
{nagy.g.david, orban.gergo}@wigner.mta.hu
[1]Computational Systems Neuroscience Lab, Wigner Research Centre for Physics, Budapest, Hungary
[2]Institute of Physics, Eotvos Lorand University, Budapest, Hungary

Abstract

Human learning in complex environments critically depends on the ability to perform model selection, that is to assess competing hypotheses about the structure of the environment. Importantly, information is accumulated continuously, which necessitates an online process for model selection. While model selection in human learning has been explored extensively, it is unclear how memory systems support learning in an online setting. We formulate a semantic learner and demonstrate that online learning on open model spaces results in a delicate choice between either tracking a possibly infinite number of competing models or retaining experiences in an intact form. Since none of these choices is feasible for a bounded-resource memory system, we propose an episodic learner that retains an optimised subset of experiences in addition to semantic memory. On a simple model system we demonstrate that this normative theory of episodic memory can effectively circumvent the challenge of online model selection.

Keywords: episodic memory; semantic memory; online model selection; Bayesian modeling; bounded-resource-rationality

Introduction

In a complex, structured environment that is capable of providing a practically infinite variety of possible experiences, storing them in all their detail would take a prohibitive amount of memory and would be useless in responding to novel situations. It is more beneficial for an learning agent to extract the structure of the world into a concise model, which enables both compression and generalisation, and store this model instead of the observations. But then what is the benefit of devoting precious mental resources to encoding inconsequential contingencies by storing rich snapshots of actual experience, that is, what use is episodic memory?

We argue that online learning in open-ended hypothesis spaces under realistic resource constraints — similar to what the human brain faces — presents a computational challenge that makes such a memory system necessary. In an online learning scenario, observations arrive sequentially and predictions have to be continuously updated. Iterative updates of a particular model's parameters do not require storing the data, since it is sufficient to retain only the information relevant to the specification of the parameters. However, if the structural form of the model is a priori unknown (Kemp & Tenenbaum, 2008), then only a subset of candidate models can be tracked at any given time, since the memory cost of retaining even such compressed statistics becomes prohibitive for an infinite set of models. The inevitable information loss resulting from this restriction presents the brain with a delicate problem: relevance judgements, that is, decisions about what to forget and what to remember can only be based on the currently tracked models, but the initial guess for which models these should be is likely to be wrong because the initial data will only warrant an overly simple model and because it might be misleading about the correct structure and form. Introducing such a bias in the interpretation of new experiences towards the wrong models means that statistical power required for model updating cannot accumulate, since the evidence for alternative models and the information needed for fitting those models will often be deemed irrelevant and discarded, preventing the discovery of the correct representation.

We propose that an episodic memory can alleviate the fundamental problem of online learning described above, by retaining a selected subset of samples. This mini-batch allows evidence for a novel model to accumulate by retaining the contingent details of observations irrespective of how relevant they appear under the current model. We also argue that to take full advantage of episodic memory, its contents should be chosen selectively, so that the combination of episodic and semantic memories provide an efficient representation of the observations.

We are aware of two prior attempts to provide a normative explanation for an episodic memory based on computational principles. The complementary learning systems account of (McClelland, McNaughton, & O'Reilly, 1995) suggests that a hippocampal learning system is required in order to avoid interference with knowledge stored in a neocortical system where learning occurs via slow changes of synaptic connectivity in a network of neurons. Catastrophic interference can be seen as a special case of the detrimental consequences of an inability to maintain a lossless representation of observations during learning, but in contrast to our treatment, the complementary learning systems approach lacks a normative framework and only concerns parameter estimation within a single model. Lengyel & Dayan (2009) argue that using the data samples directly for control is advantageous at the early stages of learning in a new environment. A different but related question about how the combination of semantic and episodic memories can be used to optimize reconstruction is explored by (Hemmer & Steyvers, 2009).

While this paper is intended primarily as a normative argument for the existence of a cognitive system, the problem explored here is intimately related to the efforts in machine learning to handle the problem of online Bayesian model selection in arbitrarily complex model spaces. There are numerous proposals for methods that deal with online model selection or model selection in infinite model-spaces (Grosse, Salakhutdinov, Freeman, & Tenenbaum, 2012; Hjort, Holmes, Müller, & Walker, 2010) separately.

Recently, there have been attempts to tackle both challenges at once in a similar setting, but these are concerned with a restricted hypothesis space over possible model forms, such as mixture models (Sato, 2001; Fearnhead, 2004; Gomes, Welling, & Perona, 2008). Methods that are specific to a given model form have the potential to be vastly more efficient within their domain, but we are striving to find the principles for a general purpose computational architecture that is flexible enough to accommodate uncertainty in the structural form of the model (Kemp & Tenenbaum, 2008). To the best of our knowledge, such a scenario has not yet been explored.

Learning paradigm

In this paper we aim to study how the computational problem of learning shapes the architecture and dynamics of long-term memory. We assume that the main goal of human learning is the acquisition of a suitable representation of the world and propose that this learning process is characterised by the following fundamental properties: i) it is incremental; ii) it requires an open-ended hypothesis space which incorporates not only an arbitrary amount of complexity but also enables the discovery of the appropriate model form; and iii) it is subject to computational constraints, most notably a limited amount of memory.

Our main argument is agnostic to the choice of learning method, but we are adopting the Bayesian inference framework. This framework provides us with a consistent, general and arguably elegant solution for dealing with uncertainty during learning and is central to many state-of-the-art advances in machine learning (Ghahramani, 2015) while simultaneously being able to capture a large body of knowledge concerning the acquisition of abstract knowledge in humans (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Orbán, Fiser, Aslin, & Lengyel, 2008). In this framework the problem of learning can be formalised as the continual refinement and updating of a probabilistic generative model, where information about unobservable or currently not observed variables, parameters and candidate world structures can all be expressed as probability distributions over latent variables.

In our treatment the memory constraints are formalised such that after the model has been updated, the observation is discarded and only the sufficient statistics for the best performing model is kept. The two main challenges introduced by these constraints are that the learner needs to both: i) assess the plausibility and ii) approximate the right parameter settings of alternative models based solely on the sufficient statistics of the tracked model, without having access to the data.

We set out with an example learning problem that can demonstrate both the challenges and the power of the proposed approach: a mixture of Gaussians model (MoG) has the benefit of showing non-trivial model-learning dynamics while also providing an opportunity for analytical treatment. Mixture models are also frequently used as cognitive models of human category learning (Sanborn, Griffiths, & Navarro, 2006). We use a version where model selection corresponds to determining the correct number of mixture components based solely on the data; parameter learning consists of finding the means for the components; while mixture weights and variance of mixture components are assumed to be fixed and known. Although a more flexible model would provide richer dynamics, the main challenges stated earlier can be clearly demonstrated on this simplified model.

The rest of the paper is structured as follows: first, we show how incremental Bayesian inference works in a setting without resource constraints; next, we introduce a learning agent that only has access to a semantic memory and demonstrate that it has a propensity to discard the information that would enable model change; finally, we show that the introduction of an episodic memory substantially mitigates this problem.

Learning in an unconstrained setting

Bayesian inference provides a consistent framework for learning the form, the structure and the parameters of the model estimating the probability distribution of data. Learning entails the estimation of the posterior probability of parameters (θ) in a given model and/or that of the model (m) itself:

$$\mathrm{P}(\theta \,|\, \mathcal{D}, m) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, \theta, m)\,\mathrm{P}(\theta, m) \tag{1}$$

$$\mathrm{P}(m \,|\, \mathcal{D}) \;\propto\; \mathrm{P}(\mathcal{D} \,|\, m)\,\mathrm{P}(m) \tag{2}$$

Posterior probabilities for alternative model structures, and/or forms need to be assessed individually and the marginal likelihood (mLLH),

$$P(\mathcal{D} \,|\, m) = \int d\theta\, P(\mathcal{D} \,|\, \theta, m)\,\mathrm{P}\,(\theta \,|\, m)\,, \tag{3}$$

plays a critical role in comparing these models: even with a uniform prior probability distribution over alternative models, the mLLH function implements the automatic Occam's razor principle, which ensures that the simplest model that can account for the observed variance in the data has the highest posterior probability. Even when the model prior is flat, the evaluation of mLLH is sufficient to compare the models.

In the analytic treatment of MoG, the posterior over the means μ is a MoG again, in which the number of mixture components grows exponentially with the number of observations T. Whether learning is performed on the whole batch of data at once or is done in an online manner, Bayesian inference yields the posterior distribution of parameters for any particular model structure at any particular time (Fig. 1a-d). This posterior distribution can be used to make predictions on upcoming data and learning helps to disentangle the predictions of different models. While early in the training a complex model that reflects the actual statistics of the data adequately might be discounted because of lack of sufficient evidence, after extended experience the marginal likelihood of the simpler model will be overcome by the model of right complexity (Fig. 1a-d). Switching time in model selection is determined by the actual data samples and is defined by the evolution of the mLLH (Fig. 1e,f). The Automatic Occam's

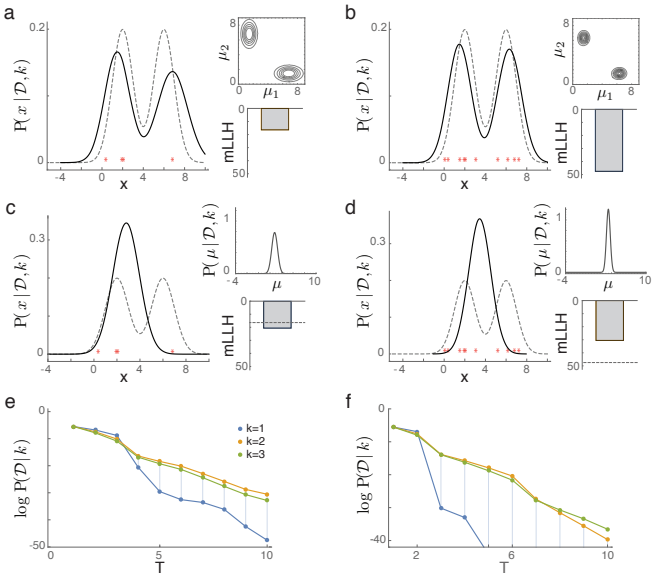


Figure 1: Illustration of model learning on a MoG model. **a**, The goal of learning is the estimation of the probability distribution of the data (*left panel, dashed grey line*) from a limited sample (*asterisks*, $n = 4$). Inference in a given model yields a posterior probability distribution over model parameters (*upper right panel*). The model assumes two mixture components ($k = 2$). Based on the posterior, the predictive posterior distribution (*solid black line*) provides our estimate on how data points are distributed. Marginal likelihood assesses the statistical power of the model (*lower right panel*). **b**, Same as **a** but using a larger data set ($n = 10$). Tighter posterior results in a tighter and more accurate predictive probability distribution and higher average marginal likelihood. **c**, **d**, Same as **a** and **b** but for a $k = 1$ model. **e**, Evolution of mLLH as more data is accumulated from a $k = 2$ model. Colours show models with different number of mixture components. Equality of mLLH at $T = 1$ is a consequence of learning limited to the means. **f**, Same as **e** but for a data set from a $k = 3$ mixture.

razor that is implemented by the mLLH function ensures that no overfitting happens: the learner discovers more complex structures if data statistics justifies such a model but keeps the model as simple as possible.

Semantic-only learner under constraints

While Eq. 1 provides a general recipe for adjusting the model parameters to data, learning can be formulated in two markedly different ways. i), In order to obtain a posterior at a particular time T, the whole data set $\mathcal{D}^T$ is evaluated according to Eq. 1. ii), Online learning relies on a parameter posterior obtained at an earlier time point $T-1$ to provide a prior for the evaluation of novel data:

$$\mathrm{P}\left(\theta \,|\, \mathcal{D}^T, m\right) \propto \mathrm{P}\left(x^T \,|\, \theta, m\right) \mathrm{P}\left(\theta \,|\, \mathcal{D}^{T-1}, m\right) \qquad (4)$$

While online learning has the same power as batch learning, it has the benefit that it is explicitly formulated such that the effect of the earlier data points is summarized in the posterior calculated for $\mathcal{D}^{T-1}$. As a consequence, online learning liberates us from the need to retain the whole data set: once the posterior has been updated the data can be discarded. As long as both parameters and models are updated, this procedure provides a consistent method to update and compare alternative hypotheses on how the model was generated without needing to keep a growing data set in memory. In contrast, if we track only a limited number of models (one model being an extreme but valid approach), discarding data prevents the consistent assessment of alternative models.

The unavailability of the original data leads to an uncertainty as to the possible past data sets that could lead to the same available statistics. An ideal learner represents this uncertainty by means of a probability distribution over possible past data sets. The learner needs a method for constructing such a distribution based solely on the posterior of the current model, since this contains all the information that it has retained. Given such a distribution, a method is required to compare alternative models (i.e. estimate the mLLHs, Eq. 3) and to assess what the parameters of the alternative models would have been had those been tracked from the beginning (i.e. estimate parameter posteriors of novel models Eq. 1). We propose that a natural approximation of the current model's estimate of the distribution of possible past data sets can be obtained by the assessment of the posterior predictive distribution, $\mathrm{P}(x \,|\, \mathcal{D}, m) = \int d\theta \, \mathrm{P}(x \,|\, \theta, m) \, \mathrm{P}(\theta \,|\, \mathcal{D}, m)$, of the tracked model. This choice is conceptually related to using "pseudopatterns" to transfer knowledge between different models (French, 1999). It has the benefit that while the parameter posteriors of different models in general span very different spaces and are thus not comparable, all models give predictions over the same data space (Fig. 1a-d). Another benefit is that the predictive distribution is presumably available for the learner in any case, since it is a fundamental component of numerous other cognitive computations as well.

Inferring the posterior of a novel model

In a given model, the posterior distribution of parameters summarises the model's knowledge about the statistics of the data. Since the predictive distribution of the tracked model carries information about the uncertainty of the parameters this can be used to approximate the posterior of the parameters in a novel model by minimising the dissimilarity of the predictive posterior distributions. Minimising the KL divergence solves exactly this problem:

$$\mathrm{P}\left(\theta \,|\, \mathcal{D}, m'\right) \approx \underset{\mathrm{P}(\theta|\mathcal{D},m')}{\operatorname{argmin}} \; \mathrm{KL}\left[\mathrm{P}(x \,|\, \mathcal{D}, m) \,||\, \mathrm{P}\left(x \,|\, \mathcal{D}, m'\right)\right]. \qquad (5)$$

Calculating the KL divergence analytically is in most cases unfeasible, therefore two approximations have been made. First, inspired by Snelson and Ghahramani (2005) we were looking for a compact representation of the predictive posterior, but instead of achieving this by simply taking a likely set of parameter settings, we've assumed that the posterior comes

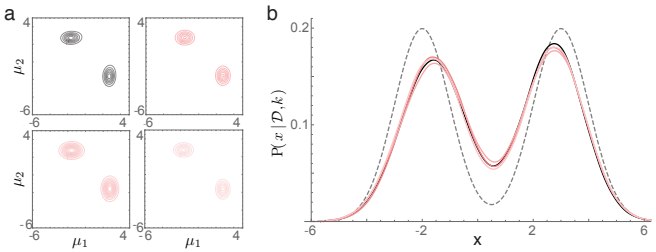


Figure 2: Reconstruction of the parameter posterior from the predictive posterior distribution. **a**, Posterior distribution of the component means in a $k = 2$ model after observing $n = 10$ data points. *Black contour plot*: posterior obtained by analytical calculation; *coloured contour plots*: posterior reconstructions. **b**, Comparison of the true predictive posterior distribution (*black line*) and its approximations. Colours are matched across panels, *dashed line*: data distribution.

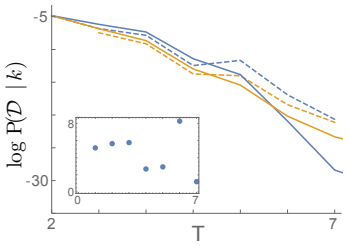


Figure 3: Inability of the memory-constrained learner to increase model complexity. Evolution of the true mLLH (analytic batch learner) of different models (*continuous lines*) and the mLLH of a constrained semantic learner (*dashed lines*). *Inset*:the data set used.

from a simple parametric distribution family:

$$\mathrm{P}\big(\theta \,|\, \mathcal{D}, m'\big) \to \mathrm{P}\big(\theta|\eta, m'\big)\,, \tag{6}$$

where η provides a parametrisation of the approximate posterior. As a result, the former functional optimization problem in (Eq. 5) reduces to

$$\hat{\eta} = \underset{\eta}{\operatorname{argmin}}\ \mathrm{KL}\left[\mathrm{P}(x|\mathcal{D}, m)\ ||\,\mathrm{P}\big(x|\eta, m'\big)\right], \tag{7}$$

where $\mathrm{P}(x|\eta, m') = \int d\theta\, \mathrm{P}(x|\theta, m')\,\mathrm{P}(\theta|\eta, m')$ is the approximate predictive posterior distribution. Eq. 7 is equivalent to minimising the cross entropy, which can be approximated using a Monte Carlo integral. After sampling $\hat{x}_i \sim \mathrm{P}(x \,|\, \mathcal{D}, m)$ we have to choose the η for which the expected value of $\log\left(\mathrm{P}(x|\eta, m')\right)$ is maximal, concluding to a maximum likelihood estimation over the generated 'fake data'

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \sum_i \log\left(\mathrm{P}\big(\hat{x}_i|\eta, m'\big)\right) \tag{8}$$

The resulting $\mathrm{P}(\theta|\hat{\eta}, m')$ is our estimate of the parameter posterior on the original data. The parameter posterior in our implementation of MoG is a MoG again, hence a convenient and effective parametrisation of the posterior uses a single mixture component. This approximate posterior effectively reproduces both the true posterior and the true predictive posterior distribution of the model (Fig. 2).

Model comparison in constrained learners

Model comparison requires the assessment of the mLLH function for alternative models (Eq. 3). However, even if we have access to the marginal likelihood of the tracked model, discarding the original data points renders the construction of the mLLH for the novel model impossible. Again, the posterior of the tracked model summarises our knowledge of the data and and therefore we rely on the predictions that can be drawn from the model posterior in order to assess the possible data sets. This can be achieved by calculating the expected value of the marginal likelihood over the predictive posterior:

$$\left\langle \mathrm{P}\big(\mathcal{D}^* \,|\, m'\big) \right\rangle_{\mathcal{D}^* \sim \mathrm{P}(\mathcal{D}^* \,|\, \mathcal{D}, m)}, \tag{9}$$

where $\mathcal{D}^*$ denotes fake data sets obtained from the predictive posterior distribution. This expected value can be evaluated by Monte Carlo sampling. Upon the arrival of a novel data point x^T, fake data sets are sampled from the predictive distribution. The novel data point is then appended to the fake dataset and the marginal likelihoods are calculated and averaged. In general, a single experience does not constitute adequate evidence for switching to an alternative model, since it lacks sufficient statistical power (Fig. 3). Note, that this claim is not true in extreme cases: there always exist outliers such that the marginal likelihood's automatic Occam's Razor effect will be overpowered by the unlikeliness of the new data (data not shown). If the present model estimate is correct, and the observed data corroborates this model then it can be integrated without information loss. We argue however, that models of differing form and complexity have different kinds of regularities that they can capture, and it is exactly the recurring appearance of features of the data that the current model is unable to represent that necessitates model change. Consequently, when a novel data point arrives which pushes the learner toward a change of model form but is insufficient in itself to force a switch, then the information loss prevents any subsequent model change (Fig. 3). This results in an inability to switch models for the memory-constrained learner even after observing arbitrary amount of evidence that supports a different one.

Episodic learner

The episodic learner differs from the semantic learner only in an additional limited capacity storage for observations. Since the semantic learner's inability to change models is a result of loss of information about past data, it is reasonable to expect that providing a buffer for data points is bound to help. However, we also require that the capacity of episodic memory necessary to enable model change should be small relative to the memory demands of a batch learner. Simply using this

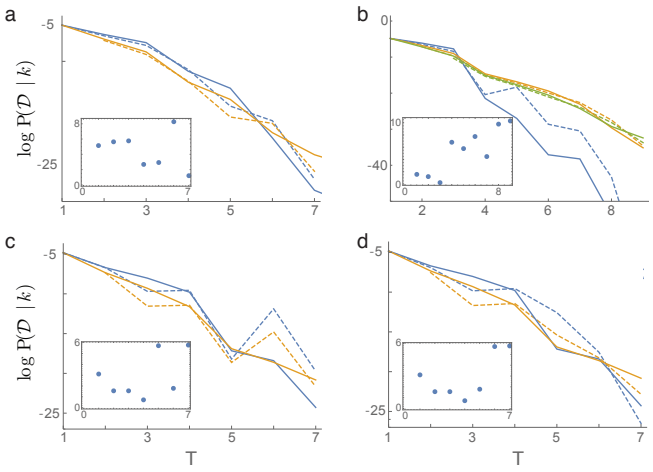


Figure 4: Effect of introducing episodic memory. mLLHs of the analytic batch learners (*solid lines)* approximate learners (*dashed lines*). **a, b**, Using ordered data and retaining the last two data points, the more complex model can obtain sufficient statistical power to overcome the Occam's razor effect at transitions $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$, respectively. **c**, For sampled data (unordered) a sliding window for two data points is insufficient to induce model switch. **d**, Episodic memory effectively rearranges data points (compare with panel **c**) such that the arrival of a subsequent data point(s) incompatible with the simple model induces model switch.

storage indiscriminately as a sliding window is inefficient for enabling model change (Fig. 4) since the experiences that taken together would provide the necessary statistical power for model change might not arrive consecutively. Taking full advantage of episodic storage requires the learner to optimise its contents and use it selectively. Thus, given a bounded capacity, the selection criterion for determining which data points to store in episodic and which in semantic memory is expected to be optimised to support the learner in dealing with online model selection.

In order to retain statistics necessary for model transitions, an episodic learner needs to identify points that have a large information content with respect to fitting the models. The Shannon definition of surprise $-log(P(\mathcal{D}|m))$ has been criticised as being unfit for this purpose because low predictive probably does not guarantee that the observation is informative with respect to the appropriateness of the model. Therefore we adopt the Bayesian definition of surprise (Itti & Baldi, 2005), which characterises the extent to which the posterior is different from the prior expectations

$$S(\mathcal{D},m) = KL(P(m|\mathcal{D})||P(m)). \tag{10}$$

Ideally, episodes that are maximally informative regarding the model form would be sought but that would require evaluating the model posterior, $P(m|\mathcal{D}_{T-1})$, which is not accessible, since the learner doesn't necessarily evaluate the same set of models at different steps. Instead, we use the surprise in the model parameters as a proxy: this selects observations that change the learner's beliefs about the parameters the most. A large change in the parameter posterior signifies a difficulty in explaining the new observations and previously seen data under the current model which suggests that a change of models might be appropriate. Another insight can shed further light on the motivation behind our choice of selection criterion. Adopting the perspective that the memory trace is a lossily compressed form of the data, it should be optimised so that the distribution over past data – used in approximating the mLLH and alternative posteriors – is going to be as accurate as possible. We can view the combination of episodic and semantic memories as jointly providing a representation of the agent's past experiences $P_{SM}(x|\eta) + \sum_{x_m \in EM} \delta(x - x_m)$. In order to achieve the best compression the learner needs to use each kind of memory system to store the information it is most suited to reconstruct. Performing such an optimisation would be relatively straightforward by comparing the combined representation with the data, but the data was previously discarded. The learner can, however, select the data points that would change the reconstruction to a large extent, by seeing how much the posterior would change if the given experience was stored in semantic memory. Taken together, we formulated the criterion for selecting a data point for storing in episodic memory by assessing whether the dissimilarity of posteriors with the novel data exceeds a fixed threshold:

$$KL(P(\mu|x_T,\eta,k)||P(\mu|\eta,k)) > \tau. \tag{11}$$

Threshold τ is measured in units of surprise and its value was determined empirically, but performance is relatively robust to its choice. At low threshold values the learner becomes non-selective, which results in accumulating sequential mini-batches. On the other hand, at high threshold levels the learner will be reluctant to store anything in episodic memory and is thus asymptotically equivalent to the semantic learner. When episodic memory is saturated the learner "consolidates" the episodes by performing batch learning on its content. Upon triggering a model change the episodes also serve to find the parameter posterior of the novel model. For demonstration we have set the maximal size of episodic memory to one and used it to show that the problems of a constrained semantic learner can be effectively alleviated (Fig.4).

We have directly contrasted the performance of learning models in the model selection task on random data sets of length $T = 12$ where the generating distribution had $k = 2$ or $k = 3$ components (Fig. 5). Besides the unconstrained learner and the semantic learner, we set up models for an episodic learner with a memory capacity of one and two items, and also a pseudo-episodic learner that does not perform optimisation on the items to be stored in episodic memory. The episodic learner can demonstrate a remarkable increase in performance even with an extremely limited capacity. In order to make a fair comparison between $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$ switches we balanced the difficulty of model switch. Our analysis on $k = 2 \rightarrow 3$ switch revealed an even more pronounced advantage of the episodic learner over the semantic learner, doubling the probability of a correct switch.

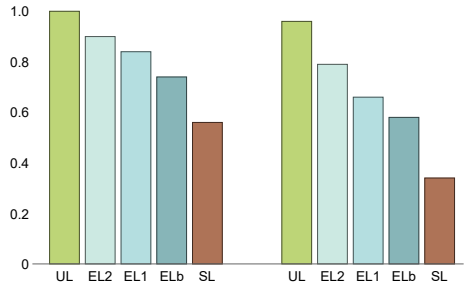


Figure 5: Comparison of model learning in different learners. UL:unconstrained; EL2: episodic with capacity 2; EL1: episodic with capacity 1; ELb: pseudo episodic with no selectivity; SL: semantic. Probability of $k = 1 \rightarrow 2$ and $k = 2 \rightarrow 3$ model switch when data comes from a MoG with $k = 2$ and $k = 3$ (*left* and *right panels*, respectively) estimated from a thousand model runs each.

Discussion

We have offered a normative argument for the existence of episodic memory by analysing a computational problem that the brain has to solve, namely online model selection in an open-ended model space. We used a simple minimal model to demonstrate that the introduction of memory constraints has dire consequences for a semantic-only learner and showed that these problems are substantially mitigated by an episodic memory, the contents of which are selected based on the Bayesian formalisation of surprise.

Our choice of model was motivated by analytical tractability which helped us to set a benchmark to model learning. While this choice constrained the form of the model and the size of the data set, the demonstrated problem is fundamental. These restrictions can be lifted by allowing the iterative posterior updates to be approximate, for example by using particle filters. Importantly, we strove to only use principles and approximations that are agnostic to the model class, so that the episodic learner can straightforwardly be extended to richer hypothesis spaces.

The overall goal of our normative account is to shed light on the dynamics underlying the organisation of long-term memory: from a continuous stream of experience, how does the human brain determine what parts to remember and what to forget? It is extensively documented that humans are prone to systematic biases in these decisions. We share the widely-held belief that these systematic memory errors reflect rational adaptations to computational resource constraints. In our assessment, a comprehensive explanation and detailed predictions on how these processes work requires an understanding of both the computational function and the constraints that shape the dynamics of long-term memory. In this paper, we aimed to provide the computational backbone for such a normative understanding: although some aspects of the current treatment are reminiscent of the characteristics of the dynamics of human memory (e.g. storing detailed representations of surprising events), a more direct comparison between model predictions and human performance will require the analysis of model classes that can be related to available human data.

Acknowledgements.

We thank Zoubin Ghahramani and Szabolcs Káli for discussions and Máté Lengyel and Balázs Ujfalussy for comments on earlier versions of the manuscript. We thank the anonymous reviewers for useful suggestions. This work was supported by an MTA Lendület Fellowship.

References

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Stat Comput*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

Gomes, R., Welling, M., & Perona, P. (2008). Incremental learning of nonparametric Bayesian mixture models. *2008 IEEE Conf on Comput Vis and Pattern Recogn*.

Grosse, R. B., Salakhutdinov, R., Freeman, W. T., & Tenenbaum, J. B. (2012). Exploiting compositionality to explore a large space of model structures. *Conference on Uncertainty in Artificial Intelligence*.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic and semantic information in memory for natural scenes. *Proc 31st Ann Conf of the Cogn Sci Soc*, 1557–1562.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge Univ. Press.

Itti, L., & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Adv neur inf proc sys* (pp. 547–554).

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proc Natl Acad Sci*, *105*, 10687–92.

Lengyel, M., & Dayan, P. (2009). Hippocampal contributions to control: The third way. *Adv Neur Inf Proc Sys*.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psych Rev*, *102*(3), 419–57.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proc Natl Acad Sci*, *105*, 2745–50.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A More Rational Model of Categorization. *Proc 28th Ann Conf of the Cogn Sci Soc*, 1–6.

Sato, M.-a. (2001). Online Model Selection Based on the Variational Bayes. *Neural Comput*, *13*, 1649–81.

Snelson, E., & Ghahramani, Z. (2005). Compact approximations to Bayesian predictive distributions. *Proc 22nd Int Conf on Machine Learning*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*, 1279–85.

Stop paying attention: the need for explicit stopping in inhibitory control

Ning Ma (nima@eng.ucsd.edu)
Department of Electrical and Computer Engineering, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92037 USA

Angela J. Yu (ajyu@ucsd.edu)
Department of Cognitive Science, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92037 USA

Abstract

Inhibitory control, the ability to stop inappropriate actions, is an important cognitive function often investigated via the stop-signal task, in which an infrequent stop signal instructs the subject to stop a default go response. Previously, we proposed a rational decision-making model for stopping, suggesting the observer makes a repeated Go versus Wait choice at each instant, so that a Stop response is realized by repeatedly choosing to Wait. We propose an alternative model here that incorporates a third choice, Stop. Critically, unlike the Wait action, choosing the Stop action not only blocks a Go response at the current moment but also for the remainder of the trial – the disadvantage of losing this flexibility is balanced by the benefit of not having to pay attention anymore. We show that this new model both reproduces known behavioral effects and has internal dynamics resembling presumed Go neural activations in the brain.

Keywords: Bayesian model, decision-making, stochastic control theory, inhibitory control, stop-signal task, neural data

Introduction

Humans and animals are often faced with the need to choose, under time pressure, an action among options with uncertain consequence. The ability to dynamically withhold or modify planned actions according changing task conditions is known as *inhibitory control*. In psychology and neuroscience, inhibitory control has been studied extensively using the stop-signal task (Logan & Cowan, 1984). In this task, subject performs a default go task on each trial, usually consisting of two-alternative forced choice (2AFC) discrimination between two stimuli (e.g. press "L" for square, press "R" for circle). On a small fraction of trials, an additional *stop* signal occurs at some time (known as SSD, or stop-signal delay) after the go stimulus onset, and the subject is instructed to withhold or stop the *go* response. When the subject succeeds to stop, the trial is considered a stop success (SS) trial; otherwise, it is considered a stop error (SE) trial. Typically the SSD is chosen such that subjects on average only achieve 50% accuracy on the stop trials.

The classical model for the stop signal task is the race model (Logan & Cowan, 1984; Boucher, Palmeri, Logan, & Schall, 2007), which is a mechanistic account that posits a race to threshold between independent *go* and *stop* processes (See Fig. 1). A stop trial results in SE if the *go* response is processed before the stop process. The race model also defines a subject-specific *stop signal reaction time* (SSRT), which is a measure of the average amount of time the observer requires to process the stop signal and cancel the go response (in practice, it is often calculated as the difference between mean Go RT and the SSD specific to each subject for achieving 50% accuracy on stop trials). Although the race model provides a simple and elegant description of the basic behavioral phenomena, it is not a normative account of how the brain *ought* to treat the stop signal task according to task demands or behavioral goals. As such, it does not have a principled basis for predicting how behavior might change according to changes in task conditions or motivational factors. It also does not possess the representational richness to identify the computational functions of all the different brain areas implicated in the stop-signal task, or to explain the distinct causes of the myriad inhibitory deficits observed in various psychiatric conditions (e.g. ADHD, depression, OCD, drug addiction).

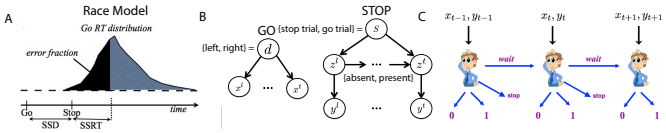


Figure 1: Models for inhibitory control in the stop-signal task. (A) The classical race mode posits that the behavioral outcome (go or stop) is determined by a race between two independent go and stop processes. (B) Bayesian graphical model for noisy sensory data generation in the rational decision-making model. (C) The decision of whether/when to Go, which Go response to select, and whether/when to Stop, are modeled as sequential decision-making, where the subject chooses at each moment whether to select a Go or Stop response, or to wait at least one more time point.

In part to overcome some of these challenges, we previously proposed a normative Bayesian Markov decision process (MDP) model for stopping (Shenoy & Yu, 2011), which assumed that subjects maintains a continually evolving, Bayes-optimal belief state about stimulus properties, and that they make an moment-by-moment optimal decision between *go* and *wait* by mapping the current belief state to optimal action. We showed that this model accounts for a range of classical and more subtle behavioral effects in the stop-signal task, includes the possibility of predicting how experimental manipulations of different cognitive factors should affect stopping behavior (Shenoy & Yu, 2011; J. S. Ide, Shenoy,

Yu*, & Li*, 2013; Ma & Yu, 2015, 2016). The model was also used to successfully identify brain regions involved in representing and predicting the probability of encountering a stop signal in healthy human subjects (J. S. Ide et al., 2013), as well as how that neural representation becomes altered in users of stimulants such as cocaine and metaphetamine (J. Ide, Hu, Zhang, Yu, & Li, 2015; Harlé et al., 2014, 2015).

One critical assumption made by the previous model (Shenoy & Yu, 2011) is that only two possible actions are entertained at each moment within the trial, Go or Wait. A Stop response is only realized in the model by choosing the Wait action repeatedly until the response deadline. Essentially, this model implies that SSRT is an emergent property, that there is no underlying stop process that terminates in a stop action. However, empirical evidence, including neuroimaging data in humans (Aron et al., 2007) and neurophysiology data in monkeys (Hanes, Schall, & Patterson, 1998), suggests that the brain may instead execute an explicit *stop* action on successful stop trials. However, adding a Stop action to the action set is non-trivial, because if both Wait and Stop block the Go action in the current moment in time, but Stop in addition blocks the possibility of choosing Go in the future while Wait allows that possibility, then it is always more rewarding to choose Wait over Stop in order to keep that possibility open. There must be some additional benefit to the Stop action that would make it worthwhile to execute. We hypothesize here that the extra benefit is a certain savings in attentional cost, such that choosing Stop alleviates the observer from the cost associated with attending to the sensory input and engaging with the task for the remainder of the trial.

Specifically, we formulate a novel Bayesian MDP model for inhibitory control, in which there are three explicit actions available to the decision-maker: Go, Wait, and Stop. In addition to the four kinds of behavioral costs included in the original model – the cost of making the wrong go response, the cost of not responding on a go trial, the cost of not stopping on a stop trial, and the cost of time incurred proportional to the total length of the trial – we incorporate an additional term, the cost of attending to the sensory input, which is terminated by either a Go action, a Stop action, or the expiration of the response period. Analogous to the stop-signal reaction time (SSRT) assumed by the Race Model, we define the Stop RT to be the temporal delay between the onset of stop signal and the time when the *Stop* action is chosen.

In the following, we first describe the model, then show how the model captures a variety of behavioral phenomena observed in the stop signal task, as well as neural data implicating neurons in the frontal eye field in participating in the initiation or execution of the Stop action. We conclude with some discussion of related work and thoughts on future directions.

The Model

As in the earlier MDP model (Shenoy & Yu, 2011), this MDP model consists of two key components, a monitoring component that formalizes sensory processing as iterative Bayesian posterior inference based on conditionally iid data, and a decision process that applies an optimal stochastic control policy. We describe the two components below, and show how the model behaves on different trial types.

Monitoring process as Bayesian statistical inference

We use the same Bayesian inference model proposed in (Shenoy & Yu, 2011) to implement the sensory processing component, and thus provide only a short overview here. Fig. 1B shows the graphical model, whereby the two hidden variables correspond respectively to the identity of the go stimulus, $d \in \{0,1\}$, and whether this trial is stop trial, $s \in \{0,1\}$. The priors of d and s, in our model, are $P(d = 1) = 0.5$ and $r = P(s = 1) = 0.25$, consistent with general experimental settings. Conditioned on the go response identity d, a sequence of iid sensory inputs are generated on each trial, x^1, ... ,x^t, ... ,where t indexes time step within a trial. The likelihoods of the sensory inputs given d are $f_0(x^t) = p(x^t|d = 0)$ and $f_1(x^t) = p(x^t|d = 1)$, which are assumed to be Bernoulli distribution with distinct rate parameters q_d and $1 - q_d$, respectively. The dynamic variable z^t denotes the presence/absence of the stop signal. $z^1 = ... = z^{\theta-1} = 0$ and $z^\theta = z^{\theta+1} = ... = 1$ if a stop signal appears at time θ. For simplicity, we assume that the onset of the stop signal θ follows a geometric distribution: $P(\theta = t|s = 1) = q(1-q)^{t-1}$. Conditioned on z^t, a stream of iid observations are generated on each trial. The likelihoods of the the sensory inputs, associated with the stop signal, are $p(y^t|z^t = 0) = g_0(y^t)$ and $p(y^t|z^t = 1) = g_1(y^t)$. We still assume that the likelihood functions, g_0 and g_1, are Bernoulli distributions with distinct parameters q_s and $1 - q_s$.

In the recognition model, Bayes' Rule is applied in the usual iterative manner way to compute the the sequential posterior probability associated with go stimulus identity, $p_d^t = P(d = 1|\mathbf{x}^t)$, where $\mathbf{x}^t = \{x^1, x^2, ..., x^t\}$ denotes all the data observed so far. Similarly, computing the posterior probability that the stop signal is already been present, $p_z^t = P(\theta < t|\mathbf{y}^t)$, involves inverting the generative model, which is a simple version of a hidden Markov model. p_z^t then can be used to compute the posterior probability that the current trial is a stop trial, $p_s^t = P(s = 1|\mathbf{y}^t) = p_z^t + (1 - p_z^t)P(s = 1|\theta > t, \mathbf{y}^t)$, where $P(s = 1|\theta > t, \mathbf{y}^t)$ represents the probability that the stop signal will occur in the future. The *belief state* at time t is defined to be the vector $\mathbf{b}^t = (p_d^t, p_s^t)$.

Decision process as optimal stochastic control

In each trial, the subject is required to make response to a go stimulus by a response deadline D, or else the trial terminates and the response is recorded as *stop*. We define a loss function that accounts for the cost and penalty structure of the stop-signal task, and assume that the observer minimizes the expected value of this loss function in choosing whether to *go*, *wait*, or *stop* at each moment in time, based on the current belief state.

Like the earlier MDP model (Shenoy & Yu, 2011), we assume that there is a basic cost c_r per unit time on each trial if the current trial is not terminated. In addition, the subject has to pay an extra attentional cost c_a per unit time if it decides to continually collect and process the new coming sensory input. The subject can explicitly choose the *stop* action to stop processing the sensory information and take the benefit of only paying the basic cost c_r in the rest of trial. Once the stop action is instantiated, the subject will no longer choose the go action again, thus will incur a unit penalty cost for missing a go option on a go trial. Making response to the *go* stimulus will terminate the current trial and save the subject the basic and attentional costs in the rest trial, but in the price of paying a unit penalty cost for making a discrimination error on a go trial or a penalty cost c_s for responding on a stop trial. The subject can also take the *wait* action to process new sensory information and update the belief state in the next time step.

Let τ_r denote the trial termination time and τ_a the time the subject is involved in the task, so that $\tau_a = \tau_r < D$ if a go response is made before the deadline D, $\{\tau_r = D, \tau_a < D\}$ if an explicit stop action is chosen, and $\tau_a = D$ if the subject waits until the deadline. $\delta \in \{0,1\}$ represents the possible binary discriminations produced by making a go response. We assume the loss function:

$$l(\tau_r,\tau_a,\delta;d,s,\theta,D) = (c_r+c_a)\tau_a \mathbf{1}_{\{\tau_r=\tau_a<D\}} + c_s \mathbf{1}_{\{\tau_r<D,s=1\}} + \mathbf{1}_{\{\tau_r<D,\delta\neq d,s=0\}} + (c_a\tau_a + c_r D)\mathbf{1}_{\{\tau_r=D\}} + \mathbf{1}_{\{\tau_r=D,s=0\}}$$

where the first three terms correspond to the cost for taking the go action and the last two denote the cost for taking stop action or waiting until the deadline. The optimal decision policy will minimize the expected loss, $L_\pi = \mathbb{E}[l(\tau_r,\tau_a,\delta;d,s,\theta,D)]$,

$$\begin{aligned} L_\pi = {} & (c_r+c_a)\mathbb{E}[\tau_a]P(\tau_a=\tau_r<D) + c_s r P(\tau_r<D|s=1) \\ & + (1-r)P(\tau_r<D,\sigma\neq d|s=0) \\ & + (c_a\mathbb{E}[\tau_a] + c_r D)P(\tau_r=D) + (1-r)P(\tau_r=D|s=0) \end{aligned}$$

It is computationally intractable to directly minimize L_π over the policy space. Fortunately, Bellman's dynamic programming principle provides an iterative relationship between the optimal state-value function and optimal action-value function. The Bellman optimality equation for optimal state-value function, $V^t(\mathbf{b}^t)$, is

$$V^t(\mathbf{b}^t) = c_a + c_r + \min_a [\int P(\mathbf{b}^{t+1}|\mathbf{b}^t;a)V^{t+1}(\mathbf{b}^{t+1})\mathrm{d}\mathbf{b}^{t+1}]$$

where a ranges over all possible actions. In our alternative model, the action space is $\{go,stop,wait\}$ associated with three optimal action-value functions (also called Q-factors), $Q_g^t(\mathbf{b}^t)$, $Q_s^t(\mathbf{b}^t)$, and $Q_w^t(\mathbf{b}^t)$, respectively. Using the Bellman optimality equation for optimal action-value function-$Q(\mathbf{b},a) = \mathbb{E}\left[c_a + c_r + V(\mathbf{b}^{t+1})|\mathbf{b}^t = b, a_t = a\right]$, we can obtained the three Q-factors

$$\begin{aligned} Q_g^t(b^t) &= (1-p_s^t)\min(p_d^t, 1-p_d^t) + c_s p_s^t \\ Q_a^t(b^t) &= c_r(D-t) + (1-p_s^t) \\ Q_w^t(b^t) &= c_a + c_r + \mathbf{1}_{\{D>t+1\}}\mathbb{E}\left[V^{t+1}(\mathbf{b}^{t+1})|b^t\right]_{\mathbf{b}^{t+1}} \\ &\quad + \mathbf{1}_{\{D=t+1\}}(1-p_s^t) \\ V^t(\mathbf{b}^t) &= \min(Q_g^t, Q_s^t, Q_w^t) \end{aligned}$$

Note that, in our model, the optimal state-value function and action-value functions only account for the future cost after the current time step, regardless of how much cost has been paid before, since only the expected futures costs matter in adjudicating among the action options. The optimal state-value function is the smallest of three optimal action-value functions. The optimal policy chooses the action corresponding to the smallest Q-factor at each time step. The value of discrimination response,δ, is 1 if $p_d^t > 0.5$ and 0 otherwise. Since the subject can no longer update the belief state nor take any action at the deadline, the optimal state-value function can be initially computed at D as $V^t(\mathrm{b}^D) = 1 - p_s^D$. The recursive relationship between the optimal action-value and state-value functions in Bellman optimality equation allows us to the compute the optimal state-value functions and Q factors backwards in time from $t = D-1$ to $t = 1$.

In the last section, we showed that the belief state $\mathbf{b}^{t+1}$ is a deterministic function of $\mathbf{b}^t$ and the observations. Thus, given V^{t+1}, we can compute $\mathbb{E}\left[V^{t+1}(\mathbf{b}^{t+1})|\mathbf{b}^t\right]$ by averaging over all possible next observations x^{t+1},y^{t+1}.

$$\mathbb{E}\left[V^{t+1}(\mathbf{b}^{t+1})|\mathbf{b}^t\right] = \sum_{x^{t+1},y^{t+1}} p(x^{t+1},y^{t+1}|\mathbf{b}^t)V^{t+1}(\mathbf{b}^{t+1}(\mathbf{b}^t,x^{t+1},y^{t+1}))$$

$$\begin{aligned} p(x^{t+1},y^{t+1}|\mathbf{b}^t) &= p(x^{t+1}|p_d^t)p(y^{t+1}|p_s^t) \\ p(x^{t+1}|p_d^t) &= p_d^t f_1(x^{t+1}) + (1-p_d^t)f_0(x^{t+1}) \\ p(y^{t+1}|p_s^t) &= (p_z^t + (1-p_z^t)h(t+1))g_1(y^{t+1}) \\ &\quad + (1-p_z^t)(1-h(t+1))g_0(y^{t+1}) \end{aligned}$$

In the simulations, we discretize the space of p_d^t and p_z^t each into 200 bins.

Results: Model Simulations

Fig. 2A shows the simulated evolution of belief state in the model for different trial types: (1) go trial (Go), where no stop signal appears, (2) stop success trial (SSS), where a stop signal is successfully processed by taking an explicit stop action, (3) stop error trial (SE), where a go response is made on a stop trial. Similar to (Shenoy & Yu, 2011), in SSS trials, the go stimulus happens to be processed slowly while the stop signal is being processed quickly, thus leading to successful stopping; conversely, on SE trials, the go stimulus happens to be processed quickly while the stop signal is being processed slowly. Note that the difference in these belief state trajectories across SSS and SE trials is solely due to sensory noise in the observation generation process.

Fig. 2B shows the simulated evolution of different Q-factors, or the expected cost of taking the three actions (Go,

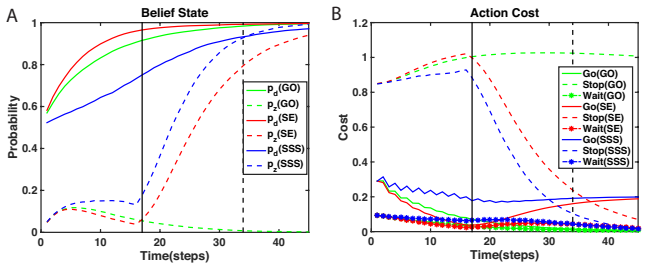


Figure 2: Mean Belief state and Q-factors. (A) Evolution of the average belief states p_d (solid line) and p_z (dashed line) for different trials-Go(green):go trials, SSS(blue): stop trials successfully stopped by choosing explicit stop action, SE (red): stop error trials. In SE trials, the go stimulus is processed faster than the stop signal, but the converse in SSS trials. We also assume that $d = 1$ for all trials in the figure for the purpose of simplicity. The onset of stop signal is θ_s = 17 time steps(solid vertical line). The dashed vertical line represents the SSRT with current parameters. (B) Go(solid line), Stop(dashed line) and Wait(dotted line) cost for the same classification of trials. In SSS trials, the go costs significantly overpass the wait cost until the stop cost drops below the wait cost. In contrast to SSS, SE trials shows a rapidly decreasing go cost which dips below the wait cost before the stop cost decreases sufficiently, leading to stop errors. Although the average go cost never falls below the average wait cost, each individual trajectory will cross over at different time due to the stochasticity of observations. We adopt most of the paramters used in (Shenoy & Yu, 2011): q_d = 0.68, q_s = 0.72, q = 0.2, r = 0.25, D = 50 steps, c_s = 0.2, c_r = 0.002, c_a = 0.002. Unless otherwise stated, there parameters are used in all the subsequent simulations.

Wait, Stop), over time on different trial types. In SSS trial, the *go* cost decreases slowly and never drops below the *wait* cost, while the *stop* cost drops rapidly after the onset of the stop signal and eventually below both the *wait* and *go* costs. The *go* cost in the SE trial shows the converse, dipping below the *wait* cost before the *stop* cost has decreased sufficiently (Q_g does not look like it dips below Q_w in the average trajectory, but it does do so on every individual trial, but at different moments, such that the average looks like it does not do so). In Go trials, the *stop* cost is large and continuously increasing, while the *go* cost is small and continuously decreases until it dips below Q_w.

Results: Model Comparison to Behavioral Data

Here, we show that the model reproduces behavioral phenomena observed in relation to the stop signal task, including all those demonstrated by the earlier MDP model (Shenoy & Yu, 2011).

A classical behavioral phenomenon is that SE frequency increases in an approximately logistic fashion as a function of SSD (inhibition function), which is captured by both the race model and the earlier MDP model (Shenoy & Yu, 2011); Fig. 3A shows that our model also captures this effect. Our model also capture the effect that the stop error decreases with the stop error penalty c_s, with the effect present at almost the whole range of SSD.

Additionally, it is known that subjects have slower Go reaction time (RT), lower SE rate, and faster SSRT when the relative of a stop error is increased via experimental design (Leotti & Wager, 2009), a phenomenon shown to naturally arise when c_s is increased in the earlier MDP model (Shenoy & Yu, 2011); Fig. 3B-D shows the new model also captures this. As c_s, the parameter specifying the stop error cost in the model, varies from low to high, our model simulation shows that (Fig. 3B) subjects can be expected to respond faster, (Fig. 3C) make fewer errors , and (Fig. 3D) have longer stopping latency (Stop RT). Note that Stop RT is analogous to SSRT in the race model, but instead of being estimated jointly from the Go RT distribution and the stop error rate as a function of SSD, Stop RT is computed directly by taking the difference between the Stop action time and the Go stimulus onset. The effects associated with Go RT and Stop error rate are generally robust for different setting of c_a and c_r.

Neural Representation of action value

In this section, we show how internal computational components of the MDP model compares to neural responses observed in the frontal eye field (FEF) region of the monkey cortex during an oculomotor version of the stop-signal task (Hanes et al., 1998). FEF is known to be important for the planning and execution of eye movements and is under strong top-down cognitive control. It has two known subpopulations of neurons, "movement" neurons and "fixation" neurons, which are respectively more active on go and stop trials, and which have been postulated to be instantiating the go and stop processes in the race model (Hanes et al., 1998).

Fig. 4 A and B show the spike density function of fixation and movement neurons, respectively. The go (no-stop-signal) trials (thin solid lines) are latency-matched to canceled trials (thick solid lines) with saccade latencies that are long enough, e.g greater than the SSD + SSRT, such that they would have been canceled if a stop signal had been presented. During canceled stop (SS) trials Fig. 4A, the activity of fixation neuron is significantly enhanced after the onset of the stop signal and peaks around the SSRT, diverging from its weak response in go (no-stop-signal) trials. Fig. 4 B shows that the activity of movement neuron also diverge on go (no-stop-signal) trial as compared to SS (canceled) trial around SSRT. These neuron data imply that these neurons may encode computations leading to the cancellation and execution of the go response.

We hypothesize that fixation neurons may encode the explicit Stop action in our model, while movement neural activities may reflect the formation of the decision to Go. Fig. 4C shows the simulated distribution of Stop RT on successfully stopped trials in which the models takes an explicit stop action. Stop RT shows a peak right near the SSRT, which

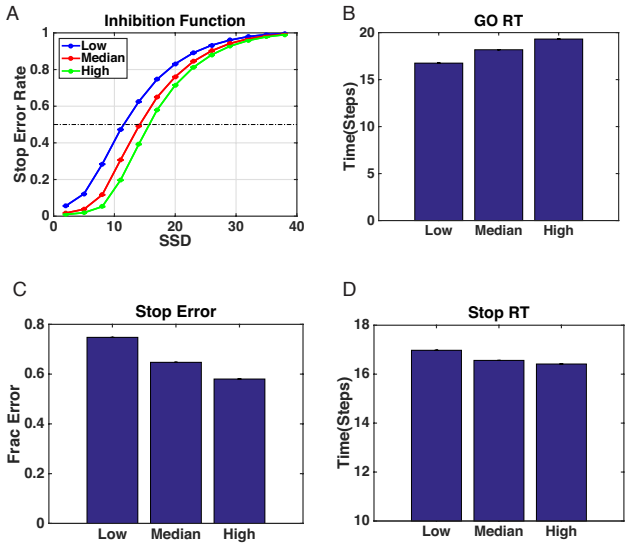


Figure 3: Influence of reward/motivation on stopping behavior. Results are averaged over 50 simulated subjects, each performing 10000 go and stop trials. Error bars (s.e.m.) are too small to be observed. (A) Model simulation produces an inhibition function (frequency of SE as a function of SSD) similar as that observed in behavioral data. Stop error (SE) decreases as the stop error penalty c_s increase (low = 0.2, median = 0.4, high = 0.6), with the effect present at almost the whole range of SSD. (B) When c_s is increased, the model responds slower in go trials, (C) make fewer stop errors (SSD = 17), and (D) exhibit shorter stopping latency (Stop RT). Stop RT is analogous to SSRT in the race model, here computed explicitly by taking the difference between the onset of stop signal and the time the decision maker chooses the *stop* action.

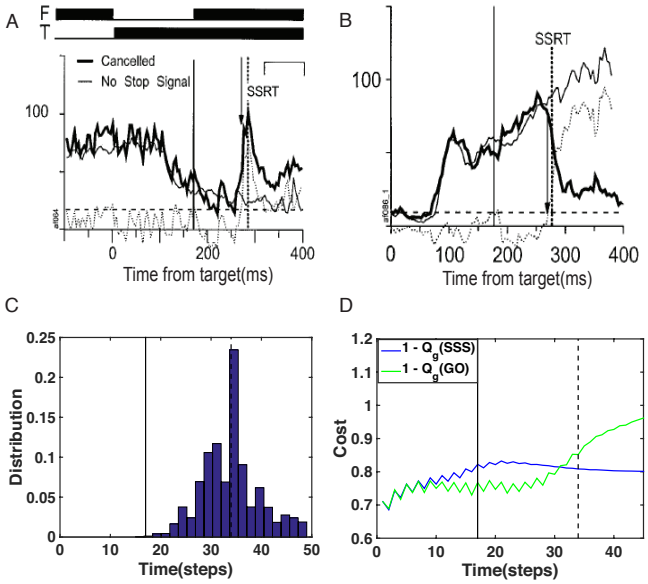


Figure 4: Neural representation of action values. (A) Average firing rate of a fixation neuron in the frontal eye field (FEF) of a monkey performing a saccade version of stop-signal task peaks at the time of the behaviorally measured SSRT (Hanes et al, 1998), suggesting fixation neurons encode a Stop action signal. (B) Movement neurons in the FEF diverge, between stop-success trials (Canceled) and go trials (No Stop Signal), at a time close to SSRT, suggesting participation in the decision process for the Go response. (C) In our model, the distribution of Stop RT in SS trials peaks near the SSRT as fixation neurons do, while (D) the expected reward (1 - expected cost) of taking the Go action diverge around SSRT on SS/Go trials, much as movement neurons do.

closely resembles the fixation neuron activity in canceled trial, implying that the fixation neurons may activate when an explicit stop action is chosen in SS trials. Fig. 4D shows the trajectories of $1 - Q_g$: the negative expected cost, or the expected reward, associated with the Go action in our model. The qualitative similarity between the expected Go reward and the activity of movement neurons suggests that the movement neuron may encode moment-by-moment estimate of Go action values.

Discussion

In this work, we presented a novel Bayesian Markov Decision Process model of inhibitory control in the stop-signal task. The key difference between this model and our earlier Bayesian MDP model (Shenoy & Yu, 2011) is that the earlier model only allowed two actions, go and wait, with the stop response only implicitly realized when the observer repeatedly chooses the Wait action until the response period expires. Here, we formally introduce an additional Stop action, the existence of which has long been postulated based on neuroimaging data in humans and neurophysiology data in monkeys. We also posited an extra attention cost associated with being engaged in the task, and which is spared when the stop action is taken. We showed that the new model can reproduce all the behavioral effects captured by the previous model. In addition, our model simulations indicate that previously observed activities of "movement" neurons in the monkey frontal eye field are consistent with their encoding the moment-by-moment valuation of the Go action, while the "fixation" neural activities may encode an explicit Stop action.

We can relate the present model to the classical race model, as the go action is typically chosen in our model when the expected go cost dips below the cost of waiting, and the stop action is typically chosen when the stop cost dips below the cost of waiting. Notably, the cost of waiting is fairly constant over the time course of the trial and stable across trial types (go, SS, SE), so that it can be thought of as the common

threshold that the go cost and stop costs race to reach first in order to determine the response outcome. However, this is not an independent race between go and stop processes, as originally envisioned in the race model (Logan & Cowan, 1984). Rather, it is closer to an interactive version of the race model that posits a late mutually inhibitory interaction between the stop and go processes (Boucher et al., 2007), except here, there is no direct antagonism between the two processes, but rather a common input (increasing sensory evidence that a stop signal is present) that drives the two processes in two opposite directions (suppressing the go process and accelerating the stop process). This leads to another notable difference between this model and the race model: the race model assumes the go process to be identical between go and stop trials, and uses that assumption to estimate the SSRT. In the present model, the go process is suppressed by the (late) onset of a stop signal, and thus the process splits into a bimodal distribution of termination times, such that there is an early mode that escapes the stop signal's suppressive influence and which ends in slightly faster average SE RT than Go RT, and a late mode that gets suppressed by the stop signal. Because of this, the true stopping latency can afford to be, and is indeed found to be, much later than the estimated SSRT. This may explain why the FEF neural response diverges between correct go and stop trials apparently too late to participate in executing the stop action (Hanes et al., 1998). This is because the race model's assumption of an unchanging go process (between go and stop trials) may be leading it to systematically under-estimate the true stopping latency. The current model would interpret FEF "fixation" neurons are signaling or relaying the decision to stop, while the FEF "movement" neurons are encoding the expected value of executing the go response.

One implication of the current work is that contextual changes in the attentional state of the observer, or the costs associated with paying attention to the task, should have systematic consequences on the observer's readiness and timing in executing a stop action. In particular, the model predicts that if the attention cost is raised, for example due to the presence of a dual task siphoning away cognitive resources, then the stop action should be chosen more readily, which would both have a behavioral consequence and be reflected in the neural dynamics. A productive line of future experimental work would be to test these predictions empirically by manipulating attentional costs. While the proposed model of inhibitory control, and the earlier MDP model that preceded it (Shenoy & Yu, 2011), may not be fully correct in describing the cognitive and neural processes underlying inhibitory control, they exemplify a powerful modeling framework for hypothesizing neural computations in the context of behaviorally defined goals and computations, which can then be tested experimentally by changing experimental conditions or behavioral objectives.

Acknowledgments

We are grateful to Pradeep Shenoy and Joseph Schilz for helping to collect data in the stop-signal task, and to Pradeep Shenoy for preliminary discussions on how to model an explicit Stop action. Funding was provided in part by an NSF CRCNS grant (BCS- 1309346) to AJY.

References

Aron, A. R., Durston, S., Eagle, D. M., Logan, G. D., Stinear, C. M., & Stuphorn, V. (2007). Converging evidence for a fronto-basal-ganglia network for inhibitory control of action and cognition. *Journal of Neuroscience*, *27*(44), 11860-4.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: an interactive race model of countermanding saccades. *Psychological Review*, *114*(2), 376-97.

Hanes, D. P., Schall, J., & Patterson, W. F. (1998). The role of frontal eye field in countermanding saccades: visual, movement and fixation activity. *J. Neurophysiol.*, *79*, 817-34.

Harlé, K. M., Shenoy, P., Steward, J. L., Tapert, S., Yu*, A. J., & Paulus*, M. P. (2014). Altered neural processing of the need to stop in young adults at risk for stimulant dependence. *Journal of Neuroscience*, *34*, 4567-4580. (*Yu and Paulus are co-senior authors)

Harlé, K. M., Steward, J. L., Zhang, S., Tapert, S., Yu*, A. J., & Paulus*, M. P. (2015). Bayesian neural adjustment of inhibitory control predicts emergence of problem stimulant use. *Brain*. (*Yu and Paulus are co-senior authors)

Ide, J., Hu, S., Zhang, S., Yu, A. J., & Li, C.-S. R. (2015). Impaired bayesian learning for cognitive control in cocaine dependence. *Drug and Alcohol Dependence*, *151*, 220–227.

Ide, J. S., Shenoy, P., Yu*, A. J., & Li*, C.-S. R. (2013). Bayesian prediction and evaluation in the anterior cingulate cortex. *Journal of Neuroscience*, *33*, 2039-2047. (*Co-senior authors)

Leotti, L. A., & Wager, T. D. (2009). Motivational influences on response inhibition measures. *J. Exp. Psychol. Hum. Percept. Perform.*, *36*(2), 430-447.

Logan, G., & Cowan, W. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psych. Rev.*, *91*, 295-327.

Ma, N., & Yu, A. J. (2015). Statistical learning and adaptive decision-making underlie human response time variability in inhibitory control. *Frontiers in Psychology*, *6*(1046). doi: 0.3389/fpsyg.2015.01046

Ma, N., & Yu, A. J. (2016). Inseparability of go and stop in inhibitory control: Go stimulus discriminability affects stopping behavior. *Frontiers in Neuroscience*, *10*(54). doi: 10.3389/fnins.2016.00054

Shenoy, P., & Yu, A. J. (2011). Rational decision-making in inhibitory control. *Frontiers in Human Neuroscience*. (doi: 10.3389/fnhum.2011.00048)

A cross-linguistic investigation on the acquisition of complex numerals

Pierina Cheung (mcheung@wesleyan.edu)
Department of Psychology, Wesleyan University, 207 High Street, Middletown, CT, 06457, USA

Meghan Dale (meghan.dale@queensu.ca)
Faculty of Education, Queen's University, 511 Union Street, Kingston, ON, K7M 5R7, Canada

Mathieu Le Corre (mathieu@uaem.mx)
Centro de Investigación Transdisciplinar en Psicologia, Universidad Autónoma del Estado de Morelos, Mexico

Abstract

Complex numerals (e.g., four hundred) have a multiplicative structure (four hundred = 4 x 100). This paper investigates whether children are sensitive to the meaning of the multiplicative structure. We designed a novel word learning paradigm and taught 4- to 6-year-old children the meaning of a novel numeral phrase (e.g., 'one gobi houses' to mean a group of three houses). We then asked whether they could generalize it to a novel context (e.g., 'two gobi butterflies' to mean two groups of three). Experiment 1 showed that only English-speaking children who received multiplier syntax training were able to generalize. Experiment 2 extended findings from Experiment 1 to Cantonese-speaking children and found that they could also generalize a novel multiplier to novel contexts. These results suggest that children as young as 4 can create a mapping between the structure of complex numerals and a multiplicative meaning.

Keywords: complex numerals, digits, multipliers, syntax, semantics, preschoolers, cross-linguistic investigation

Introduction

Numerals are built using a compositional system, in which a set of individual numerals can be combined to form many different numerals. For example, to count to one hundred in English, we only need to remember 28 words: one through nineteen, twenty, thirty, forty, fifty, sixty, seventy, eighty, ninety, and hundred. Counting to one trillion, a number that is 10,000,000,000 times larger than 100 requires only four additional words: thousand, million, billion, and trillion. How numerals are combined to form complex numerals such as 'twenty-three' and 'two hundred' are governed by compositional rules (e.g., Hurford, 1975; Ionin & Matushansky, 2006). The current study investigates the developmental origins of the compositional rules of numerals, by examining the linguistic and conceptual building blocks of the system.

Linguists have long observed that there are two types of numerals based on which compositional rules apply: digits (two, five) and multipliers (hundred, thousand; Hurford, 1975; Ionin & Matushansky, 2006). Digits and multipliers have different syntactic properties. First, multipliers are similar to singular count nouns in English. They must be preceded by a numeral or a determiner – e.g., "A/one million people watched the game." In contrast, digits can be used in its bare form – e.g., "Three people watched the game." Second, multipliers in English such as hundred and million can be pluralized – e.g., "Vaccination could save millions" – but digits cannot – e.g., "Vaccination could save threes" is ungrammatical.

Across all natural languages with a numeral system, digits and multipliers can be combined in two ways to form complex numerals: conjunction and multiplication. Numerals can be combined via conjunction explicitly with the use of 'and' as in "one hundred *and* one" or implicitly as in "twenty-three." Moreover, similar to a determiner phrase in which a determiner is combined with a noun (e.g., "a" + "dog" → "a dog"), a digit and a multiplier can be combined to form a numeral phrase (e.g., "one" + "hundred" → "one hundred"). Each of these two types of combinations maps onto a unique arithmetic operation. Conjunctions map onto addition (e.g., twenty-three means 20 + 3), and numeral phrases map onto multiplication (e.g., two hundred means 2 x 100).

In this paper, we focus on the mapping between the structure of complex numeral phrases such as 'one hundred' and its multiplicative meaning as a first step to investigate the acquisition of the compositional nature of numerals.

Acquisition of compositional rules of numerals

How do children discover the compositional rules that govern the combination of digits and multipliers? On one view, children discover the rules of the numeral system on their own (e.g., Hurford, 1975; Siegler & Robinson, 1982). For example, after encountering numeral phrases in their language, children may discover a rule that maps the structure of complex numerals onto multiplication. They are then able to apply this rule to generate new numbers in the form of 'digit multiplier' and assign them a multiplicative meaning.

Alternatively, acquisition of complex numerals may be item-based. For example, after learning 'one hundred', children still have to separately learn numerals that are of a similar form – e.g., 'two hundred', 'three hundred', and 'four hundred'. In other words, children's learning is not initially rule-based, and only later do they learn the underlying compositional rules that govern numeral structure, perhaps via explicit instruction. Thus, children may learn complex numerals by rote memorization prior to decomposing them (Fuson, 1990).

Although no previous studies have examined the acquisition of multipliers, there is related, though indirect, evidence from studies on the acquisition of the count

sequence and the place-value system that provide insights into how compositional rules for numerals may be acquired.

First, previous studies have demonstrated that children's counting experience is related to discovering the underlying structure of the count sequence. For example, Siegler and Robinson (1982) asked a group of 3- to 5-year-olds to repeatedly count from one over multiple sessions, and to count on from a particular number that was beyond their counting range. They found that children who counted between 20 and 99 always ended with 'nine' (twenty-nine, thirty-nine, etc), suggesting that they understood the within-decade structure but were limited by the knowledge of the next decade word. They also found that those who counted up to 100 understood the within-decade structure *and* showed some knowledge of the between-decade structure. Thus, a majority of children did not simply memorize up to an arbitrary number. Rather, children's counting reveals the different rules that they discover about the count sequence (e.g., the order of one to nine, the order of decade terms).

In addition, cross-linguistic studies have shown that the numeral structure of some languages may facilitate the rule discovery process (Miller, et al., 1995; Miller & Stigler, 1987). Numerals in Korean and Chinese follow a highly regular structure. For example, in Chinese, the numbers following ten (*shi*) are ten-one (*shi-yi*) and ten-two (*shi-er*), while the words for twenty and thirty are two-ten (*er-shi*), and three-ten (*san-shi*), respectively. This contrasts sharply with the irregularities in English (e.g., eleven, fourteen, twenty). Previous studies have found that 4-year-old children learning Chinese and Korean are able to count higher than their English counterparts (Miller et al., 1995), suggesting that the regularity of a language's numeral system may help children discover the count sequence structure.

Another piece of evidence comes from studies on children's understanding of the place-value system. Previous studies have found that when asked to represent 2-digit numerals (e.g., 11, 42) with blocks, children starting at around the age of 6 are able to use a combination of unit blocks and tens blocks. For example, children represent 42 using four blocks of 10 and two single blocks rather than 42 single blocks. Some have also documented cross-linguistic differences with children's place value understanding, showing that Japanese-speaking children are more likely to use a combination of tens and units blocks than English-speaking children (Miura, 1987; Miura, Kim, Chang, & Okamoto, 1988; but see Saxton & Towse, 1998; Vasilyeva et al., 2015). These results suggest that children are able to decompose complex numerals into its constituents. Nevertheless, children's understanding of the place-value system only reveals knowledge of *written* numerals and leaves open the question of what young children understand about the compositional nature of numerals *prior to* acquiring the place-value system.

The present experiments

To investigate whether and when children acquire the mapping between the compositional structure of complex numerals and their meaning, we designed a novel word learning paradigm. Specifically, in two experiments, we asked when children between the ages of 4 and 6½ recognize that a complex numeral in the form of 'digit multiplier' maps onto multiplication.

The general logic of our experiments was to teach children a novel noun phrase that described a set of three objects (e.g., the experimenter described a group of three houses as 'one gobi houses'). Then, we tested whether children have learned the meaning of the phrase (e.g., 'one gobi Xs' to refer to a group of three objects). Critically, we asked whether children can generalize the novel noun phrase to a novel context involving 'two' (e.g., Who has two gobi books?).

In Experiment 1, we provided one group of English-speaking children with informative multiplier syntax – i.e., the numeral 'one' followed by a novel multiplier 'gobi' (one gobi Xs), and another group with uninformative syntax, with only a novel word modifying the noun (gobi Xs). The two groups of children saw the same pictures but they heard verbal descriptions that differed only in the structure of the noun phrase (one gobi Xs vs. gobi Xs). We hypothesized that if children are sensitive to the mapping between complex numeral of the form 'digit multiplier' and its multiplicative meaning, then they should be more likely to generalize the learned novel numeral phrase to new contexts when they are presented with informative multiplier syntax than uninformative syntax.

In Experiment 2, we tested children learning Cantonese Chinese, which has a regular numeral structure, using the same paradigm. Cantonese Chinese, similar to Mandarin Chinese, has no irregularities in the naming of the numerals, and multiplier syntax occurs as early as 20 (*ji-sap*). We asked whether they would also demonstrate sensitivity to the multiplier syntax, and if so, whether they would show earlier knowledge of the multiplier structure than English-speaking children. To the extent that they do, we asked whether this was due to linguistic differences or other educational or cultural factors. Tests of general receptive vocabulary and mathematical competence were included as control measures. We also included a highest count measure to investigate if counting experience affects children's sensitivity to multiplier structure.

Experiment 1 – English-speaking children

We taught English-speaking children a novel numeral phrase with multiplier syntax – i.e., the numeral 'one' followed by a novel multiplier 'gobi' (one gobi Xs; Multiplier condition), and asked if they could generalize it to a novel context involving 'two'. For example, children heard 'one gobi houses' when shown a group of three houses, and they were then asked to choose which one of two sets contained 'two gobi books'. To investigate if children's understanding of multipliers is specific to multiplier syntax, we adopted a between-subjects design and presented another group of children with the same visual

stimuli but with uninformative multiplier syntax. Specifically, the novel word 'gobi' was presented without the numeral 'one,' (gobi Xs; the No Digit condition). Thus, the only difference between the two conditions was in the syntactic structure of the noun phrase.

Method

Participants A total of 98 children between the ages of 4;2 and 6;6 participated. Sixty-eight of them were assigned to the Multiplier condition (M = 5;4), and 30 in the No Digit Condition (M = 5;4). They were recruited at daycare centres and schools in southwestern Ontario. All children spoke English as their primary language.

Design and Procedure

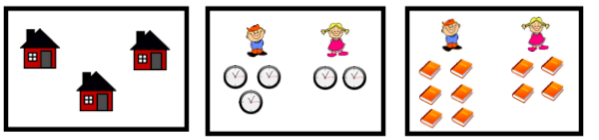
Highest count Children were asked to count as high as they can, and were stopped if they could count up to 100.

Novel Word Learning The novel word learning paradigm had two conditions: Multiplier and No Digit conditions, and each condition proceeded in three phases, including modelling, training, and generalization.

Multiplier Condition During modeling, children were told that they were going to learn a new word – gobi. They were shown sets of three objects that were labelled with a novel numeral phrase. Children in the Multiplier condition were provided with multiplier syntax, e.g., "This is one gobi houses" (see Figure 1a). There were a total of six trials. On the last modelling trial, children were shown groups of two, three, and four objects (e.g., two phones, three phones, four phones). They were told that the collection of three phones was 'one gobi phones.' Importantly, they were also told that the collection of two phones and that of four phones were *not* 'one gobi phones.'

After the modelling phase, children proceeded to the training phase. During training, children were given a forced-choiced task. They were shown a boy and a girl, one of whom had three objects and the other had either four or two. Children were asked, "Who has one gobi clocks?" (Figure 1b). Children were corrected if they chose the wrong character and praised if they chose the right one. Children passed training if they correctly answered 4 trials in a row, with a maximum of 16 trials in the training phase.

After training, children were asked to generalize to a novel numeral phrase involving the numeral 'two'. Children were shown the boy and the girl characters, and were asked, "Who has two gobi books?" (Figure 1c). Children who passed the training phase completed all generalization trials. Unlike the training phase, feedback was not provided.



a) Modelling b) Training c) Generalization

Figure 1a-c: A schematic illustration of the experimental set-up.

The forced-choice task in the generalization phase consisted of a character who had two groups of three or a group of six, and another character who had one of four different types of competitor sets: two individuals, a group of three, two groups of two, and four groups of three (see Figure 2). These competitor sets were designed to test the scope and the specificity of the acquired meaning of 'digit multiplier'. We outlined our reasoning for including the four different competitor sets below:

(1) To test whether children interpreted the novel numeral phrase as one that refers to groups or to individuals (two individuals).
(2) To test whether children interpreted the novel numeral phrase as one that applies to two groups of any number (two groups of two).
(3) To test whether children interpreted 'gobi' to mean 'three' (a group of three).
(4) To ensure that children did not simply pick the more numerous set (four groups of three).

There were a total of 24 trials, four of each type. Children completed two blocks of 12 trials. The presentation of the competitor sets was pseudo-randomized in each block. The boy and the girl had an equal number of correct answers.

The task was presented on a laptop computer. A different kind of object was used for each trial. The stimuli set contained a total of 46 objects that were familiar to preschoolers (e.g., butterflies, houses, bags, cups).

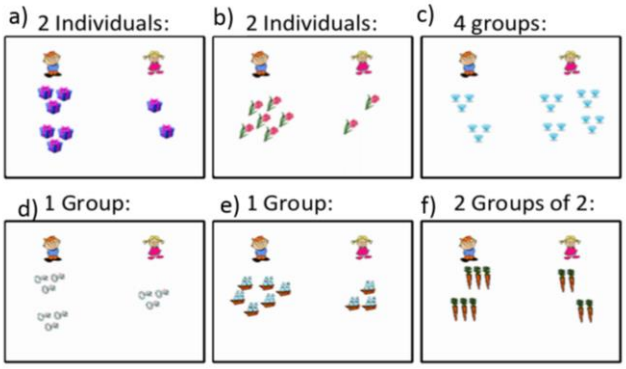


Figure 2: Six types of competitor sets. *a)* 2 groups of 3 vs. 2 individuals; *b)* a group of 6 vs. 2 individuals; *c)* 2 groups of 3 vs. 4 groups of 3; *d)* 2 groups of 3 vs. 1 group of 3; *e)* a group of 6 vs. a group of 3; *f)* 2 groups of 3 vs. 2 groups of 2

No Digit Condition Another group of children were presented with uninformative multiplier syntax in which the digit was removed from the numeral noun phrase. Specifically, children were modelled on 'gobi Xs' (e.g., This is gobi houses), trained on 'gobi Xs' (e.g., Who has gobi clocks?), and were asked to generalize to 'two gobi Xs' (Who has two gobi books?).

Control measures[1]

Receptive vocabulary We used the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2012) to measure children's receptive vocabulary. Children were asked to point to one of four pictures in response to a target word. They received one point for each correct answer, and a total score was computed for each child.

General mathematical competence We chose three subtests of the Test of Early Mathematics Ability (TEMA; Ginsburg & Baroody, 2003) as a measure of children's math skills: mental number line, verbal numerical comparison, and mental addition. Percent correct averaged across the three subtests was computed for each child.

Task Order Each testing sessions always began with a highest count task, followed by the novel word learning task. Children in the Multiplier condition also completed PPVT and TEMA.

Results and Discussion

Training phase A majority of children from the Multiplier condition (78%) passed training and were included in the analysis. Those who were included (M = 5;5) were on average five months older than those who did not pass training, $t(66) = 2.07$, $p = .042$. All children from the No Digit condition passed training and were included. Of those who passed the training phase, 85.9% of them did so within 4 to 6 trials.

Generalization phase We first examined how likely children in the Multiplier condition generalized to 'two gobi.' Children's scores in the generalization phase followed a bimodal distribution (see Figure 3). A Shapiro-Wilk test confirmed that the distribution of scores violated the normality assumption, W = .85, p < .001.

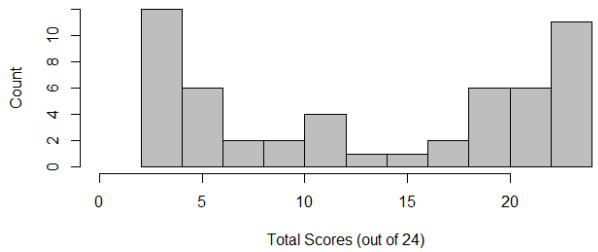


Figure 3: A histogram of children's scores in the generalization phase (Multiplier Condition).

As indicated in Figure 3, the distribution of scores formed two peaks, one at the highest end and another at the lowest end of the distribution, suggesting that some children were able to generalize from 'one gobi' to 'two gobi' consistently, and some were consistently unable to do so. To capture the dichotomous nature of children's responses and to investigate their performance in the generalization phase, we defined above chance performance as answering more than 16 out of 24 trials correct (binomial test, $p = .06$). Results showed that 25/53 (47.2%) of children scored at least 17 on the generalization task (M = 21.6, SD = 2.0); we labelled these children 'generalizers'. The remaining children scored on average 6.6 trials correct (SD = 3.7), and we labelled them 'non-generalizers'.

These results suggest that some children between the ages of 4 and 6 were able to generalize a novel numeral phrase 'one gobi Xs' to 'two gobi Xs'. This provides evidence that children are able to create a mapping between a numeral phrase in the form of 'digit multiplier' (e.g., one gobi) and a multiplicative meaning. Nevertheless, these results do not demonstrate that this mapping is unique. It is possible that the context of the training and generalization tasks allow some children to interpret any novel numeral phrase with a multiplicative structure, regardless of the actual form of the numeral phrase itself. To address this question, we analyzed how likely children in the No Digit condition generalized. We found that only 3 of them were 'generalizers', receiving a total score of 17 or higher, and a majority of them (n = 27) were 'non-generalizers'. Compared to those in the Multiplier condition, a significantly lower proportion of children in the No Digit condition generalized, $\chi^2(1) = 10.24$, $p = .001$. These results provide evidence that children who were provided with informative multiplier syntax were able to map the form 'digit multiplier' to multiplication.

Effect of age and highest count Next, we analyzed whether children's age or counting experience predicted the likelihood that they generalized in the Multiplier condition. We conducted a logistic regression with age in months and highest count as predictor variables. The dependent measure was children's status as a 'generalizer'. There was no effect of age, $\beta = .060$, $SE = .048$, $p = .21$, or counting, $\beta = .015$, $SE = .011$. $p = .17$. Nevertheless, the final model was significantly better than the constant-only model, $X^2(2) = 9.91$, $p < .001$.

Discussion

Results from Experiment 1 showed that 4 to 6½-year-old children who were trained to pair the novel numeral phrase 'one gobi' with sets of three were more likely than those in the No Digit condition to generalize to 'two gobi', suggesting that children in the Multiplier condition can create a mapping between a novel complex numeral in the form of 'digit multiplier' and a multiplicative meaning. This provides the first piece of evidence that children are sensitive to the multiplier syntax. In the following experiment, we extended this finding to another language group, and asked whether language structure affects children's sensitivity to the multiplier structure.

Experiment 2 – Cantonese-speaking children

Experiment 1 showed that English-speaking children between the ages of 4 and 6½ successfully generalized a novel numeral phrase 'one gobi' to novel contexts. We also

[1] Analyses on these control measures are reported in Experiment 2.

found that this knowledge was specific to complex numerals in the form of 'digit multiplier.' In Experiment 2, we asked whether children learning Cantonese, a language with transparent numeral system, would demonstrate earlier knowledge of the meaning of the multiplier structure. Cantonese belongs to the family of Chinese languages, and has a regular numeral structure. For example, numerals in the teens in Chinese are highly regular, unlike English – e.g., *eleven* is literally translated to 'ten one' (*sap jat*) in Cantonese Chinese, and *twelve* is 'ten two' (*sap ji*). Starting at the numeral *twenty* begins the multiplier structure – 'two ten' (*ji sap*), whereas in English, the multiplier structure appears at *one hundred*. Given these differences in numeral structure, we predicted that Cantonese-speaking children may demonstrate earlier sensitivity to the meaning of the multiplier structure.

To test this, we recruited a sample of Cantonese-speaking children and tested them using the same novel word learning paradigm as Experiment 1.

Method

Participants A total of 122 children between the ages of 4 and 6½ participated. Sixty-three children participated in the Multiplier condition (*M* = 5;3; range = 4;0 to 6;6), and 59 children participated in the No Digit condition (*M* = 5;3; range = 4;0 to 6;6). They were recruited at daycare centres in Hong Kong. All children spoke Cantonese as their primary language.

Design and Procedure The design of this experiment was identical to Experiment 1. Children were first modelled on a novel numeral phrase, followed by a training phase with feedback and a generalization phase without feedback.

Children were tested by a native Cantonese speaker. In the Multiplier condition, children were taught 'one gobi classifier Xs'. In the No Digit condition, children were trained on 'gobi classifier Xs'. All noun phrases were used with a general classifier (CL) – *goh*.[2]

Similar to Experiment 1, children also completed a highest count task, a receptive vocabulary test (PPVT) and a general mathematical competence test (TEMA).

Results

Training phase In the Multiplier condition, 90.5% of children passed the training phase (57/63 children), and among them, 94.7% of them did so within 4 to 6 trials. Children required an average of 4.72 trials (SD = 2.37 trials) to pass training. In the No Digit condition, a majority of children passed the training phase (89.8%; 53/59 children), and among them, 80.7% of them did so within 4 to 6 trials. Children required an average of 5.02 trials (SD = 2.30 trials) to pass training. Children who did not pass training were excluded (n = 6 in the Multiplier condition; n = 6 in the No-digit condition).

Generalization phase Similar to English-speaking children in Experiment 1, Cantonese-speaking children's responses in the generalization task followed a bimodal distribution (see Figure 4). A Shapiro-Wilk test confirmed that the normality assumption was violated, W = .80, p < .001.

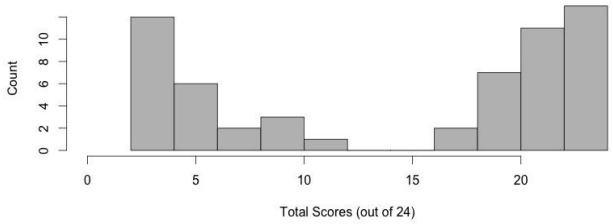


Figure 4: A histogram of children's scores in the generalization phase (Multiplier Condition).

Similar to Experiment 1, we defined above chance as answering more than 16 out of 24 trials correct (binomial test, p = .06). In the Multiplier condition, 33/57 children performed significantly above chance, $t(32) = 16.97$, $p <$ .001 (M=21.7 trials, SD=1.9 trials), and were thus termed 'generalizers'. In contrast, the 'non-generalizers' scored on average 5.3 trials correct (SD = 2.4 trials).

In the No Digit condition, surprisingly, 26/53 children performed significantly above chance, $t(25) = 12.11$, $p <$ .001 (M=20.7 trials, SD=12.0 trials), and the remaining children – 'non-generalizers' - scored on average 7.8 trials correct (SD=4.2 trials). This contrasts with the findings from Experiment 1. We discuss reasons for this diverging result in the Discussion section.

Effect of age and highest count We analyzed whether children's age or counting experience predicted the likelihood that they generalized in the Multiplier condition. We conducted a logistic regression with age in months and highest count as predictor variables. The dependent measure was children's status as a 'generalizer'. Results showed that there was no effect of age, β = .011, SE = .042, p = .78, but a marginal effect of highest count, β = .024, SE = .013, p = .064. The model with both predictors was significantly better than the constant-only model, $X^2(2) = 6.50$, p = .038.

Cross-linguistic comparison Next, to investigate whether there is a cross-linguistic difference in how likely children generalize the novel multiplier, we conducted a logistic regression with language (English, Chinese) as a predictor. We also included age in months, highest count and the two control measures – PPVT and TEMA – in the model. The dependent variable was children's status as a 'generalizer'. This analysis was performed for children in the Multiplier condition. We found no effect of language, β = -.26, SE =

[2] Cantonese is a classifier language in which nouns cannot co-occur directly with numerals, but require classifiers (CL). There is also no obligatory plural morphology. For example, 'three balls' is translated to 'three classifier ball'. In this experiment, we used the general classifier – *goh* – the most frequent classifier in Cantonese (Matthews & Yip, 1994).

.51. p = .61, suggesting that although proportionally more Chinese-speaking children were generalizers compared to English-speaking children, this difference was not reliable. None of the control measures were significant, all ps > .09.

Discussion

Experiment 2 extended the findings from Experiment 1 to children learning Cantonese, a language with regular numeral structure. We found that Cantonese-speaking children between the ages of 4 and 6½ could map complex numerals in the form 'digit multiplier' onto multiplication. We also found that some children in the No Digit condition demonstrated knowledge of the multiplier structure, which contrasts sharply with children learning English (Experiment 1). We speculate that this is due to the fact that in Cantonese, the numeral 'one' can often be dropped in conversations (Erbaugh, 2002; Matthews & Yip, 1994). This raises the possibility that even in absence of the word 'one' during modelling and training, children in the No Digit condition interpreted 'gobi CL Xs' as if it means 'one gobi CL Xs', and assigned it a multiplicative meaning. Another possibility is that the classifier may provide cues to Chinese speakers that 'gobi CL Xs' is a numeral phrase. We think that this explanation cannot fully explain the pattern of results because English-speaking children in the No Digit condition could identify a set of three objects as 'gobi Xs'. Their difficulty lies in generalizing it to novel contexts.

General Discussion

Two experiments explored whether and when children are able to map the multiplier structure, 'digit multiplier' (e.g., one hundred), onto multiplication. Using a novel word learning paradigm, we found that both English- and Cantonese-speaking children between the ages of 4 and 6½ who were taught the meaning of 'one gobi Xs' to refer to a group of three objects were able to generalize to 'two gobi Xs' as meaning two groups of three objects. We also found that knowledge of this mapping in complex numerals is specific to the form of 'digit multiplier' - when provided with uninformative syntax in which the digit was not presented, English-speaking children failed to generalize the novel noun phrase to a novel context. The current set of experiments are the first to investigate the acquisition of the multiplier syntax in children, and our results demonstrate that children as young as 4 can create a mapping between the multiplier structure – 'digit multiplier' – and a multiplicative meaning with minimal training.

Our findings provide some support for the rule-based learning account of the acquisition of complex numerals. With only six modelling trials, a majority of children learned that 'one gobi' refers to a group of three objects, and approximately half of those generalized to a novel context. Although not all children could generalize, our results indicate that even with very little input, children are able to map the syntax of multipliers to its semantics.

Nevertheless, the current study leaves open two questions. First, we found that Cantonese-speaking children in the No Digit condition were also able to generalize. Further studies are required to investigate whether Cantonese-speaking children would generalize the novel numeral phrase to novel contexts when presented with other uninformative syntactic structures. Second, we found that age, general vocabulary and mathematical skills, and the regularity of a language's numeral system did not predict how likely children generalized in the Multiplier condition. Counting experience was only marginally significant in the Chinese sample. It remains to be tested what may predict children's sensitivity to multiplier syntax.

In summary, the present results suggest that numeral syntax may provide a rich foundation for numeral learning. They also highlight the importance of investigating the linguistic aspects of the numeral system, and provide a fruitful avenue for examining the linguistic and conceptual building blocks of the acquisition of number.

References

Erbaugh, M. (2002). Classifiers are for specification: complementary functions for sortal and general classifiers in Cantonese and Mandarin, *Cahiers de linguistique-Asie orientale, 31*, 33-69.

Fuson, K. (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching. *Cognition and Instruction, 7*(4), 343-403.

Ginsburg, H., Baroody, A. (2003). *Test of Early Mathematics Ability. 3rd edition*. Autsin, TX: Pro-Ed.

Hurford, J. (1975). *The linguistic theory of numerals.* Cambridge University Press.

Ionin, T., & Matushansky, O. The composition of complex cardinals. *Journal of Semantics, 23*, 315-360.

Matthews, S., & Yip, V. (1994). *Cantonese: A comprehensive grammar*. New York: Routledge.

Miller, K. F., & Stigler, J. W. (1987). Counting in Chinese: Cultural variation in a basic cognitive skill. *Cognitive Development, 2*, 279-305.

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science, 6*(1), 56-60.

Miura, I.T. (1987). Mathematics achievement as a function of language. *Journal of Educational Psychology, 79*, 79-82.

Miura, I.T., Okamoto, Y., Kim, C., Steere, M., & Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons. *Journal of Educational Psychology, 85*, 24-30.

Siegler, R.S., & Robinson, M. (1982). The development of numerical understandings. In H. W. Reese & L. P. Lipsitt (Eds.), *Advances in child development and behavior* (pp. 241-312). New York: Academic Press.

Saxton, M., & Towse, J. N. (1998). Linguistic relativity: The case of place value in multi-digit numbers. *Journal of Experimental Child Psychology, 69*, 66–79.

Vasilyeva, M., Laski, E., Ermakova, A., Lai, W-F., Jeong, Y., & Hachigian, A. (2015). Reexamining the language account of cross-national differences in base-10 number representations. *Journal of Experimental Child Psychology, 129*, 12-25.

tDCS to premotor cortex changes action verb understanding: Complementary effects of inhibitory and excitatory stimulation

Tom Gijssels[1,2] **(tomgijssels@uchicago.edu)**

Daniel Casasanto[1,3] **(casasanto@uchicago.edu)**

[1]Department of Psychology, University of Chicago, IL
[2]Department of Linguistics, Vrije Universiteit Brussel, Belgium
[3]Grossman Institute for Neuroscience, University of Chicago, IL

Abstract

Do neural systems for planning motor actions play a functional role in understanding action language? Across multiple neuroimaging studies, processing action verbs correlates with somatotopic activity in premotor cortex (PMC). Yet, only one neurostimulation study supports a functional role for PMC in action verb understanding: paradoxically, inhibiting PMC made people respond *faster* to action verbs. Here we investigated effects of PMC excitation and inhibition on action verb understanding using tDCS. Right-handers received excitatory or inhibitory stimulation to left PMC hand areas, then made lexical decisions on unimanual action verbs and abstract verbs. tDCS polarity selectively affected how accurately participants responded to unimanual action verbs. Inhibitory stimulation to left PMC caused a relative improvement in performance for right-hand responses, whereas excitatory left PMC stimulation caused a relative impairment. tDCS polarity did not differentially affect responses to abstract verbs. Premotor areas that subserve planning actions also support understanding language about these actions.

Keywords: action; language; embodiment; premotor cortex; tDCS

Introduction

How do people understand the meaning of words? According to theories of embodied cognition, word meaning relies, in part, on neural systems for perceiving and acting on the world (Barsalou, 1999). For instance, to understand the sentence *I am petting a cat*, modality-specific brain areas may simulate the visual experience of seeing a cat, the tactile experience of feeling its fur, and the motor programs for petting it. So far, most of the empirical support for embodied semantics comes from neuroimaging studies of action language understanding. When people read action verbs like *kick*, *pick*, and *lick*, frontal motor areas tend to show somatotopic activation (i.e. *kick*, *pick* and *lick* preferentially activate leg-, hand-, and mouth-areas; Hauk, Johnsrude, & Pulvermüller, 2004, Pulvermüller, 2005, for review).

What could the motor system be contributing to the meanings of action verbs? On one view, motor simulations recapitulate previous action experiences by re-activating some of the neural circuits that were used to perform those actions (Barsalou, 1999). On an alternative view, motor simulations may not recapitulate past actions, but rather partially prepare the motor system for future actions – thus simulations are "pre-enactments" rather than reenactments (Willems, Toni, Hagoort, & Casasanto, 2010; Zwaan & Kaschak, 2008). One advantage of the "simulation as pre-enactment" view is it clarifies the computational-level motivation for simulation (Marr, 1982): partially preparing for actions should allow people to perform these actions more efficiently, if overt actions are needed (Willems, Toni, et al., 2010). If this view is correct, motor simulations should be implemented primarily in neural systems that support action planning. The available fMRI data support this proposal: processing action verbs correlates mainly with activity in motor planning areas (e.g., premotor cortex; PMC), rather than activity in motor execution areas (e.g., primary motor cortex; M1) (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Tettamanti et al., 2005; Willems, Hagoort, & Casasanto, 2010; Willems, Toni, et al., 2010, but see Pulvermüller, 2005).

To date, only one neurostimulation study has tested for a functional relationship between PMC activity and action verb understanding. Willems and colleagues (2011) used continuous theta-burst stimulation (cTBS) to change neural activity in either left or right PMC areas involved in planning right and left hand actions, respectively (Willems, Labruna, D'Esposito, Ivry, & Casasanto, 2011). The right-handed participants then performed a lexical decision task on unimanual and nonmanual action verbs. Right-handers responded faster to unimanual action verbs after cTBS to dominant hand areas (in left PMC) than after cTBS to non-dominant hand areas (in right PMC) (Willems et al., 2011). By contrast, responses to nonmanual action verbs were not differentially affected by cTBS to left vs. right PMC. The predicted interaction suggests that left PMC circuits that plan dominant hand actions play a causal role in understanding language about those actions.

Yet, the direction of Willems et al.'s (2011) reaction time (RT) pattern was unexpected. CTBS has been shown to suppress neural excitability (Huang, Edwards, Rounis, Bhatia, & Rothwell, 2005). Therefore, on the simplest prediction, left PMC stimulation might be expected to cause a *impairment* in performance instead of an improvement. Willems et al's data, therefore, appear to be an instance of "paradoxical functional facilitation," (Kapur, 1996) which arises because normal brain functioning relies on a complex interaction of excitation and inhibition. Patients with brain lesions sometimes show enhanced behavioral performance, relative to controls. Presumably, their lesions selectively affect inhibitory circuits, thus the behavior supported by these circuits becomes facilitated (Kapur, 1996; Papeo, Pascual-Leone, & Caramazza, 2013). Similarly, cTBS could have facilitated

action verb processing either by reducing interference from background noise, or by modulating the activity of inhibitory PMC circuits: inhibition may be necessary to prevent people from overtly performing the actions named by verbs, rather than covertly simulating them (Willems et al., 2011). This post-hoc explanation for Willems et al.'s findings generates testable predictions regarding the effects of excitatory vs. inhibitory stimulation of left PMC.

The current study tested for the complementary effects of excitatory and inhibitory stimulation of left PMC hand area (and surrounding tissues) on behavioral responses to manual action verbs. Right-handers received transcranial Direct Current Stimulation (tDCS) over hand areas in left and right PMC. tDCS passes a weak direct current between two scalp electrodes, causing neural excitation under the anodal electrode and, when applied to motor areas, neural inhibition under the cathodal electrode (Nitsche et al., 2008; Jacobson, Koslowsky, & Lavidor, 2012). After receiving tDCS, participants performed a lexical decision task on unimanual and abstract verbs, making responses with their left and right hand. Since the polarity of the tDCS montage determines whether its neural effects are inhibitory or excitatory, we predicted the following patterns. First, inhibitory stimulation to left PMC hand areas should cause an *improvement* in processing unimanual action verbs, as was found by Willems et al. By contrast, excitatory stimulation to the same areas should cause a behavioral *impairment* in processing unimanual action verbs; this finding would complement the pattern found by Willems et al. Second, if the meaning of unimanual action verbs relies on the same premotor circuits for preparing hand actions, then the effects of tDCS polarity should depend critically on which hand is used to make a response.

Method

Participants

73 participants from the University of Chicago community took part in the experiment. Data from 1 participant were replaced for not following task instructions and data from 1 other participant were lost due to a script error. The remaining 71 participants were monolingual native English speakers and were right-handed as established by the Edinburgh Handedness Inventory (EHI: M = 78; range = 47-100; Oldfield, 1971). Participants were healthy adults who did not report being pregnant, having sustained a stroke or brain injury, being on psychoactive medication, or having any electronic implants. All participants provided informed consent and received course credit or $30 for their participation.

Materials

198 verbs were used in this experiment: 66 unimanual verbs (e.g. *to write*), 66 abstract verbs (e.g. *to tempt*), and 66 phonotactically legal nonce words (e.g. *to frinckle*). The unimanual verbs were selected from a set of manual action verbs that elicited dominant hand responses in a pantomime elicitation study (Gijssels & Casasanto, unpublished data). Although we used a within-item design, all three verb types were matched in word length (unimanual vs. abstract: $t(130) = 1.48$, $p = .14$; unimanual vs. nonce: $t(130) = -.74$, $p = .46$; abstract vs. nonce: $t(130) = -.74$, $p = .46$). Unimanual and abstract verbs were matched in word frequency ($t(96) = .20$; $p = .85$; Coltheart, 1981).

Procedure

Transcranial Direct Current Stimulation tDCS was performed using a battery-powered Soterix Medical 1x1 (Soterix Medical, New York) with two 5x7cm saline-soaked sponges covering the electrodes. Each participants received 20 minutes of stimulation at 2 mA, which was slowly ramped up from 0 mA at stimulation onset, and ramped down to 0 mA at stimulation offset. The electrodes were placed over premotor hand areas, at FC3 and FC4 in the 10-20 electrode system (Koessler et al., 2009; Nitsche et al., 2010). In the left PMC inhibition condition (N=35), the cathode was placed at FC3 and the anode at FC4, inhibiting left PMC and simultaneously exciting right PMC. In the left PMC excitation condition (N=36) this placement was reversed, with the anode placed at FC3 and the cathode at FC4, exciting left PMC and inhibiting right PMC.

Behavioral Procedure After receiving tDCS participants performed a lexical decision task. Verbs appeared one at a time in the center of a computer screen. Participants indicated whether each stimulus was an existing English word by pressing a button corresponding to "yes" or "no" with their left or right index finger. The response mappings for each button were presented below the verb, on the left or right side of the screen. For each verb type, the "yes" response was mapped to the right button for half of the stimuli and to the left button for the other half (mapping counterbalanced across participants). The stimuli appeared in a random order, and the placement of the response labels varied unpredictably from one trial to the next.

Every trial had the following structure. First, participants saw a "Ready?" sign prompting them to push and hold down the two white "home" buttons with their left and right index finger (mapped to the 'd' and 'k' keys). Once the buttons were held down, a fixation cross appeared for a duration randomly selected between 750 and 1250 ms. Then, the stimulus and response prompts appeared. As soon as the participant had decided the correct response, they released the home button held down by the response hand and used the same hand to push the correct pink response button, after which a new trial started. Response buttons were mapped to the 'z' and 'period' keys. If participants released either of the home buttons before the stimulus was presented, the trial was restarted.

Accuracy and RTs were collected. All trials in which participants released or pressed the wrong button were classified as incorrect. Participants performed 16 practice trials and received feedback about errors during both the practice and the experimental trials.

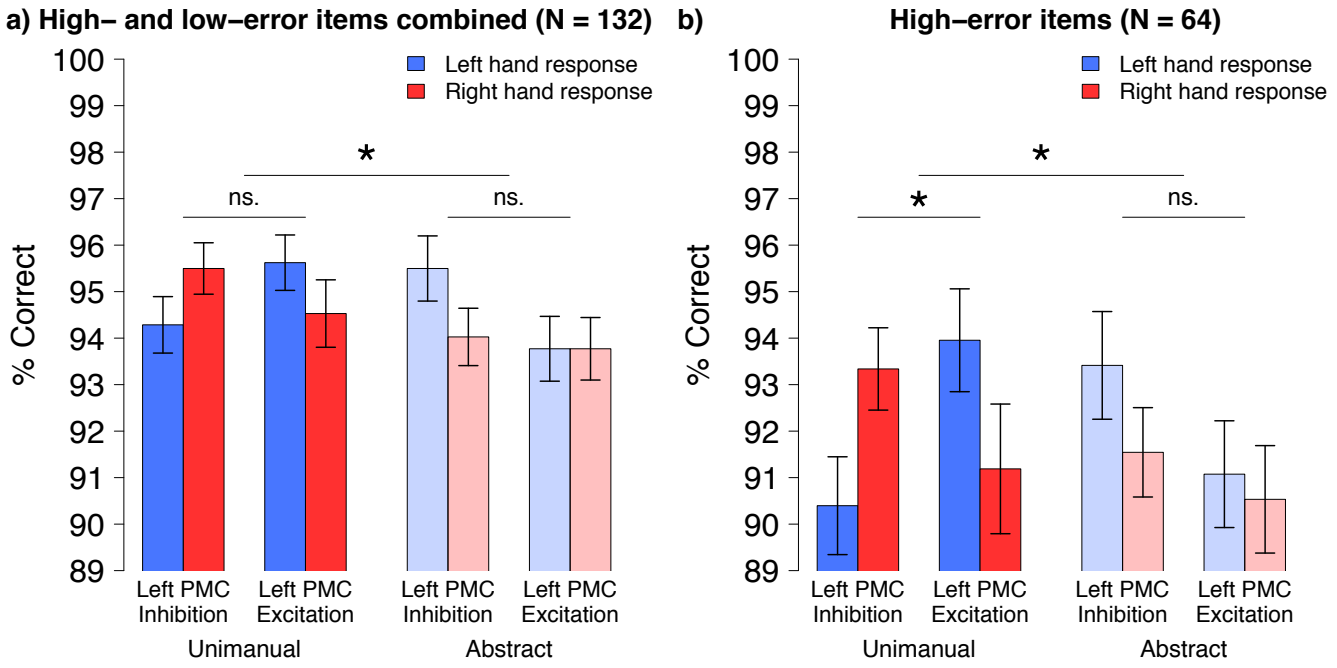


Figure 1: Mean accuracy on the lexical decision task for **a)** all items and **b)** items with a high error rate. Left PMC Inhibition = cathode left PMC, anode right PMC; Left PMC Excitation = anode left PMC, cathode right PMC. Unimanual and Abstract refer to the Verb type of the stimuli. Error bars reflect within- subject *SEM*. * indicates $p < .05$

Results

The accuracy data and RTs for target trials were analyzed with maximal mixed effects models. Both models included three two-level independent variables: tDCS polarity (left inhibitory vs. left excitatory); Verb type (unimanual vs. abstract), and Response hand for the required response (left vs. right). We used the maximal random effects structure. Nonce trials were excluded before the analyses. Accuracy data were analyzed using a general linear model with a binomial linking function. The dependent variable for this model was whether the response for each trial was correct or incorrect. For the RT model, we discarded all the incorrect trials and log-transformed RTs to reduce skew in the residuals.

Accuracy The accuracy results showed the predicted 3-way interaction of tDCS polarity x Verb type x Response hand (β = -.82, *SE* = .40, z = -2.07, p = .04; Fig. 1a). The polarity of tDCS to left PMC differentially affected the accuracy of participants' responses to unimanual and abstract verbs. In addition to the significant 3-way interaction, we also found the predicted *qualitative* pattern of results for the constituent 2-way interactions, suggesting that the predicted relationship between tDCS polarity and response hand was present, selectively for unimanual verbs. Yet, neither of the 2-way interactions was statistically significant (unimanual verbs: β = -.32, *SE* = .32, z = -.98, p = .33; abstract verbs: β = .33, *SE* = .31 , z = 1.06, p = .29).

One possible reason why the predicted 2-way interaction in the unimanual verb condition did not reach significance is that accuracy approached 100%, in all conditions: a ceiling effect. To determine whether the predicted effects were masked by this ceiling effect, we performed a second analysis on a subset of items that were not near ceiling. We calculated the number of errors for each verb (Range = 0 - 16 errors, *Median* = 3) and then performed a median split to identify items that led to more errors (High-error items: n = 31 unimanual verbs, n = 33 abstract verbs; Low-error items: n = 35 unimanual verbs, n = 33 abstract verbs). Unimanual and abstract high-error items did not differ in word frequency or word length (p's >.38).

The accuracy analysis of high-error items again showed the predicted 3-way interaction of tDCS polarity x Verb type x Response hand (β = -1.07 , *SE* = .47, z = -2.29 , p = .02; Fig. 1b). As expected, this 3-way interaction was driven selectively by responses to unimanual verbs, as shown by a significant 2-way interaction of tDCS polarity x Response hand (β = .80, *SE* = .35, z = -2.32, p = .02; Fig. 1b). After inhibitory stimulation to left PMC, participants tended to respond more accurately to unimanual verbs with their right hand than with their left hand (β = -.41, *SE* = .26, z = -1.58, p = .11; Fig. 1b). After excitatory stimulation to left PMC, we found a trend in the opposite direction: participants tended to respond less accurately with their right hand than with their left hand (β = .62, *SE* = .37, z = 1.67, p = .10; Fig. 1b). As expected, there was no statistically significant 2-way interaction of tDCS polarity and Response hand for abstract verbs (β = .20, *SE* = .36, z = .56, p = .58; Fig. 1b).

Although not of interest, for completeness we report that analysis of the low-error (near-ceiling) items showed no evidence for a 3-way interaction of tDCS polarity x Verb type x

Response hand (β = -.18, *SE* = .74, z = -.25, p = .80) nor for either of the 2-way interactions of tDCS polarity x Response hand (unimanual verbs: β =.08, *SE* = 0.68, z = .11, p = 0.91; abstract verbs: β =.43, *SE* = 0.65, z = .66, p = 0.51).

Reaction Times RTs were defined as the latency from stimulus onset to release of the "home" button. There were no statistically significant effects for the 3-way interaction of tDCS polarity x Verb type x Response hand (Wald $\chi^2(1)$ = .68, p = .41), nor for the constituent 2-way interactions of tDCS polarity x Response hand for either verb type (unimanual: Wald $\chi^2(1)$ = .10, p = .75; abstract: Wald $\chi^2(1)$ = .31, p = .58). As with the accuracy data, we analyzed the release RTs separately for the high- and low-error items. There were no statistically significant 3-way or 2-way interactions in either model (all Wald χ^2s < 1.94; all ps $> .15$).

General Discussion

This study tested whether motor circuits involved in action preparation play a causal role in action verb understanding. In right-handers, inhibitory and excitatory tDCS to hand areas in left PMC differentially affected the accuracy of responses to unimanual action verbs, but not to abstract verbs. After left-PMC inhibition, right-handers tended to make *fewer* errors to unimanual verbs with their right hand than with their left hand. After left-PMC excitation, right-handers tended to make *more* errors to unimanual verbs with their right hand than with their left hand. By contrast, the polarity of tDCS did not differentially influence how accurately participants responded to abstract verbs. These results are the first to demonstrate complementary effects of exciting and inhibiting left PMC activity on action verb processing, thus clarifying the functional contribution of PMC to understanding action language.

Some researchers have been hesitant to accept previous neurostimulation data as evidence for a functional link between motor system activity and action language understanding because there was "no evidence yet that sensory-motor cortex stimulation disrupts semantic processing," (Hauk & Tschentscher, 2013: p. 6, see also Mahon & Caramazza, 2005, 2008). Here we find both a relative improvement and a relative impairment in accuracy for lexical decisions (a task known to show semantic priming effects; Neely, Keefe, & Ross, 1989), depending on the polarity of tDCS to the motor system and on the response hand. Furthermore, the finding that response hand interacts with tDCS polarity to predict accuracy, selectively for unimanual verbs, suggests that processing these verbs not only relies on "motor circuits," but also relies on the *same* motor circuits that plan and execute unimanual hand actions (see Bedny & Caramazza, 2011).

Complementary effects of excitatory and inhibitory tDCS

Why do inhibition and excitation of left PMC lead to behavioral improvements and impairments, respectively? Across two studies, using two different neurostimulation methods, we now find a complementary pattern of results. Inhibitory tDCS to left PMC produced a conceptual replication of Willems et al.'s (2011) study that showed "paradoxical facilitation" of lexical decision after (presumably inhibitory) cTBS to left PMC hand area. Excitatory tDCS produced the complementary finding: what we could call "paradoxical impairment" of lexical decision. One potential explanation for these complementary effects of excitatory and inhibitory neurostimulation is that the mental simulation processes that contribute to action language understanding do not depend on an all-or-nothing activation of premotor circuits, but rather rely on a complex balance of neural excitation and inhibition. Left-inhibitory stimulation could facilitate performance by reducing activation of competing motor plans (see Willems et al., 2011). On another, non-mutually exclusive account, PMC inhibition may ensure that simulations do not result in the actual execution of the action named by a verb. For instance, reading the verb *throw* should not always lead to actual throwing, just like reading the word *blue* does not automatically cause the percept of blue. Left-inhibitory stimulation may have improved behavioral performance by increasing inhibition of contextually inappropriate motor plans (since our participants were asked to recognize the verbs, not to perform the actions the refer to). Left-excitatory stimulation may have impaired performance by boosting activation of all potentially relevant motor plans, thus increasing competition among them.

PMC activity affects quality of understanding

The current data also provide some of the first evidence that motor system activity not only affects how fast people understand action verbs, but also how well. Previous neurostimulation results showed that changing motor system activity affects the speed with which people process action verbs (Willems et al., 2011; Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005, though see below). Yet, as Willems and Casasanto (2011) point out, these data do not address how motor activity changes the quality with which people understand these verbs. Here we show that stimulation of hand areas affects how accurately participants process manual action verbs. In combination with previous findings, these results suggest that motor simulations contribute to both *how fast* and *how well* people construct the meaning of action language.

Why did we observe the predicted pattern in the accuracy data but not in the RT data? Since we did not speed participants (e.g., with a trial timeout), it may not be surprising that we observed the predicted interactions in the accuracy data, alone. Furthermore, the absence of an RT effect allows us to rule out the presence of a speed-accuracy trade-off. In general, it is often unclear a priori whether studies investigating motor-system contributions to action language understanding will show the predicted effect in RTs, accuracy, or both (see Mahon & Caramazza, 2008). Across our cTBS and tDCS studies, we find the predicted effects in *both* reaction times and accuracy. Since finding the predicted effect in either accuracy or RT could be interpreted as support for our

experimental hypothesis, a reasonable (though non-standard) precaution to take would be to double our *p*-values (i.e. Bonferroni correction). Even if we do so, the critical 3-way and 2-way interactions remain significant in our high-error data set.

We found the predicted 3-way interaction of tDCS, verb type, and response hand in the full data set, but the predicted lower-order interaction was not significant due to a ceiling effect. Overall, these results should be interpreted with some caution given that we had to perform an unplanned median split in order to observe the predicted 2-way interaction of tDCS by response hand. Our confidence in the data is increased, however, by the fact that (a) inhibitory left PMC stimulation produced a close conceptual replication of our previous cTBS study, and (b) excitatory left PMC stimulation produced a mirror image of these results, providing convergent evidence that inhibitory left PMC activity plays a functional role in processing unimanual action verbs.

Stimulation to M1 and action language

In contrast to PMC stimulation, stimulation of primary motor cortex (M1) has not shown consistent effects on action verb understanding. Several studies applied transcranial magnetic stimulation (TMS) to M1 but did not show the predicted behavioral effects on an action verb task (Papeo, Vallesi, Isaja, & Rumiati, 2009; Tomasino, Fink, Sparing, Dafotakis, & Weiss, 2008; Lo Gerfo et al., 2008[1]). One other study applied single pulse TMS to M1 hand and leg areas, and reported that TMS modulated lexical decision times to hand and leg action verbs differentially. However, the data do not show the predicted somatotopic specificity; the predicted effect was only found for leg verbs with no clear effect of TMS on hand verb processing(Pulvermüller et al., 2005).

The inconsistent results of M1 stimulation are problematic for the view that language-driven simulations are reenactments of previous actions, but may be compatible with the view that simulations are pre-enactments. If verb-driven motor simulations are partial preparations to perform the action named by the verb, then they should mainly rely on neural systems for action preparation (PMC), and not necessarily on systems for action execution (M1). In an fMRI test of this proposal, both motor simulations and motor imagery activated PMC, but only motor imagery activated M1 (Willems, Toni, et al., 2010). Consistent with M1 being involved selectively in imagery, the only study that showed the predicted somatotopic effects of TMS to M1 asked participants to judge the concreteness of action verbs, a task that is likely to induce imagery (Repetto, Colombo, Cipresso, & Riva, 2013, see also Tomasino et al., 2008).

[1]Parenthetically, we note that Lo Gerfo et al. (2008) reported that RTs for action-related verbs and nouns were slower after rTMS to left M1 compared to sham, according to a one-tailed test. However, this effect is not well supported by the data: not only was the use of a one-tailed test unlicensed, the comparison of interest was not licensed by the necessary higher-order interaction, nor was there any correction for multiple comparisons in their 2x2x2x2 design.

Why tDCS?

Across hundreds of studies of embodied cognition, only a few have used true experimental methods to test the functional relationship between the motor system and action verb representation. Two of these studies used tDCS to test the role of M1 in action language processing, but did not find the predicted differences among the anodal, cathodal, and sham conditions (Liuzzi et al., 2010; Vicario & Rumiati, 2012). Our results, however, show that tDCS can be a useful alternative to TMS for testing theories of language-driven motor simulation. tDCS does not lead to any noticeable muscle twitches, which might inadvertently draw participants attention to the stimulated body part (Papeo et al., 2013). tDCS is also safer, less unpleasant, and less expensive than TMS. One potential shortcoming is that excitatory tDCS to one area is often paired with inhibitory stimulation of another area. Yet, this complementary montage can also be an advantage: in the current study, we were able to simultaneously induce complementary stimulation to left and right PMC. Presumably, this montage enhanced the *relative* excitation or inhibition of the target left-hemisphere regions compared to their right-hemisphere homologues.

Conclusions

These results suggest a functional relationship between the neural systems for preparing hand actions and understanding language about those actions. tDCS to left PMC affected how accurately people processed action language, in a predictable way: inhibitory stimulation caused a relative improvement in performance (consistent with our previous cTBS results), and excitatory stimulation caused a relative impairment. These complementary effects of excitatory and inhibitory tDCS were specific to unimanual action verbs, and depended critically on the hand that participants used to respond. Previous neurostimulation results have shown that modulating PMC activity can influence how fast people can respond to action verbs. The present results show that modulating PMC activity in the hemisphere that controls the dominant hand can also affect *how well* people process verbs that name dominant-hand actions, strengthening the evidence that motor simulations contribute to the meanings of action words.

Acknowledgments

Many thanks to Nikolai Maximay for help with data collection. Research supported by a James S. McDonnell Foundation Scholar Award (#220020236) and an NSF grant (BCS-1257101) to DC, and a PhD grant from the Scientific Research Fund of Flanders to TG.

References

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology*, *16*(18), 1818–1823.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577–660.

Bedny, M., & Caramazza, A. (2011). Perception, action, and word meanings in the human brain: The case from action verbs. *Annals of the New York Academy of Sciences*, *1224*(1), 81–95.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307.

Hauk, O., & Tschentscher, N. (2013). The body of evidence: what can neuroscience tell us about embodied semantics? *Frontiers in Psychology*, *4*(50), 1-14.

Huang, Y.-Z., Edwards, M. J., Rounis, E., Bhatia, K. P., & Rothwell, J. C. (2005). Theta burst stimulation of the human motor cortex. *Neuron*, *45*(2), 201–206.

Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). tDCS polarity effects in motor and cognitive domains: a meta-analytical review. *Experimental Brain Research*, *216*(1), 1–10.

Kapur, N. (1996). Paradoxical functional facilitation in brain-behaviour research. a critical review. *Brain*, *119*(5), 1775–1790.

Koessler, L., Maillard, L., Benhadid, A., Vignal, J. P., Felblinger, J., Vespignani, H., & Braun, M. (2009). Automated cortical projection of EEG sensors: anatomical correlation via the international 10–10 system. *Neuroimage*, *46*(1), 64–72.

Liuzzi, G., Freundlieb, N., Ridder, V., Hoppe, J., Heise, K., Zimerman, M., ... others (2010). The involvement of the left motor cortex in learning of a novel action word lexicon. *Current Biology*, *20*(19), 1745–1751.

Lo Gerfo, E., Oliveri, M., Torriero, S., Salerno, S., Koch, G., & Caltagirone, C. (2008). The influence of rTMS over prefrontal and motor areas in a morphological task: grammatical vs. semantic effects. *Neuropsychologia*, *46*(2), 764–770.

Mahon, B. Z., & Caramazza, A. (2005). The orchestration of the sensory-motor systems: Clues from neuropsychology. *Cognitive Neuropsychology*, *22*(3-4), 480–494.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102*(1), 59–70.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *Henry Holt and Co Inc., New York, NY*.

Neely, J. H., Keefe, D. E., & Ross, K. L. (1989). Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. *JEP: LMC*, *15*(6), 1003–1019.

Nitsche, M. A., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., ... others (2008). Transcranial direct current stimulation: state of the art 2008. *Brain Stimulation*, *1*(3), 206–223.

Nitsche, M. A., Jakoubkova, M., Thirugnanasambandam, N., Schmalfuss, L., Hullemann, S., Sonka, K., ... Happe, S. (2010). Contribution of the premotor cortex to consolidation of motor sequence learning in humans during sleep. *Journal of neurophysiology*, *104*(5), 2603–2614.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Papeo, L., Pascual-Leone, A., & Caramazza, A. (2013). Disrupting the brain to validate hypotheses on the neurobiology of language. *Frontiers in Human Neuroscience*, *7*(148), 1-8.

Papeo, L., Vallesi, A., Isaja, A., & Rumiati, R. I. (2009). Effects of TMS on different stages of motor and non-motor verb processing in the primary motor cortex. *PLoS One*, *4*(2), e4508.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(7), 576–582.

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793–797.

Repetto, C., Colombo, B., Cipresso, P., & Riva, G. (2013). The effects of rTMS over the primary motor cortex: The link between action and language. *Neuropsychologia*, *51*(1), 8–13.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., ... Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281.

Tomasino, B., Fink, G. R., Sparing, R., Dafotakis, M., & Weiss, P. H. (2008). Action verbs and the primary motor cortex: a comparative TMS study of silent reading, frequency judgments, and motor imagery. *Neuropsychologia*, *46*(7), 1915–1926.

Vicario, C. M., & Rumiati, R. I. (2012). tDCS of the primary motor cortex improves the detection of semantic dissonance. *Neuroscience letters*, *518*(2), 133–137.

Willems, R. M., & Casasanto, D. (2011). Flexibility in embodied language understanding. *Frontiers in Psychology*, *2*(116), 1–11.

Willems, R. M., Hagoort, P., & Casasanto, D. (2010). Body-specific representations of action verbs neural evidence from right-and left-handers. *Psychological Science*, *21*(1), 67–74.

Willems, R. M., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding evidence from theta-burst transcranial magnetic stimulation. *Psychological Science*, *22*(7), 849–854.

Willems, R. M., Toni, I., Hagoort, P., & Casasanto, D. (2010). Neural dissociations between action verb understanding and motor imagery. *Journal of Cognitive Neuroscience*, *22*(10), 2387–2400.

Zwaan, R. A., & Kaschak, M. P. (2008). Language in the brain, body, and world. In P. Robbins & M. Aydede (Eds.), *The cambridge handbook of situated cognition.* New York: Cambridge University Press.

From low to high cognition: A multi-level model of behavioral control in the primate brain

Jerald D. Kralik[1,2] (jerald.kralik@gmail.com), Dongqing Shi[1] (dongqshi@gmail.com), Omar A. El-Shroa[1] (o.a.elshroa@gmail.com), Laura E. Ray[3] (laura.e.ray@dartmouth.edu)

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755 USA
[2]Department of Bio and Brain Engineering, Program of Brain and Cognitive Engineering, College of Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, South Korea
[3]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755 USA

Abstract

The basic cognitive architecture of the human brain remains unknown. However, there is evidence for the existence of distinct behavioral control systems shared by humans and nonhumans; and there is further evidence pointing to distinct higher-level problem solving systems shared by humans and other primates. To clarify the nature of these proposed systems and examine how they may interact in the brain, we present a four-level model of the primate brain and compare its performance to three other brain models in the face of a challenging foraging problem (i.e., with transparent, and thus, invisible barriers). In all manipulations (e.g., size of problem space, number of obstacles), our model never performed the best outright; however, it was always among the best, appearing to be a jack-of-all-trades. Thus, the virtues of our primate brain lie not only in the heights of thinking it can reach, but also in its range and versatility.

Keywords: cognitive control; cognitive architecture; reinforcement learning; creativity; agency; concept formation

Introduction

There is considerable evidence for the existence of two neural systems controlling mammalian behavior: (1) a goal-directed decision-making system based in the ventromedial prefrontal cortex (vmPFC), and (2) a habit system based in the striatum (Daw, Niv, & Dayan, 2005; Rangel, Camerer, & Montague, 2008). It has been argued that the goal-directed system is model-based, meaning essentially that an individual knows what will happen upon taking an action; while the habit system is not, meaning that an individual simply knows what action to take in a given state of the world (Daw et al., 2005). A model-based system potentially provides an infrastructure for mental simulation, planning, and reasoning, which in turn lead to faster learning and greater generalization across environmental conditions. Such potent abilities reduce errors when facing novel situations. However, research in engineering and computer science shows that models are notoriously brittle and therefore often break under real-world conditions. One approach to dealing with this brittleness is to have a fallback process that does not rely on the model, and this appears to be one of the main advantages of the model-free habit system. Another approach would be to have an additional system that can fix the models when they break, enabling individuals to solve these harder problems. We classify these two types of problems as *apparent* — i.e., the problem features are properly modeled by the standard model-based problem-solving system — and *nonapparent* — i.e., those that break the system.

In primates, there is neuroanatomical evidence for a distinct region of prefrontal cortex (PFC), called granular PFC (which includes lateral PFC and frontal pole), and some theorists have suggested that this region may enable primates to perform unconventional behaviors, such as looking away from a salient visual stimulus when required to do so (Passingham & Wise, 2012; Striedter, 2005). Related to this view, we hypothesize that granular PFC mediates the cognitive ability to solve *nonapparent* problems. Moreover, we believe that a detailed analysis of this ability in primates will shed light on the mechanisms that underpin creative problem solving in people.

Here, we present a computational framework and model to begin this analysis of clarifying apparent versus nonapparent problem solving, as well as to examine how these processes interrelate with the other main behavioral control systems of the primate brain. First, we focus on the classic *detour problem* in the comparative literature, in which subjects must circumvent a barrier (either via reaching or navigating) to obtain a reward item (Wynne & Udell, 2013). Most comparative research on the detour problem has focused on cases with opaque barriers as obstacles, and have generally shown that many species can solve the problem — thus, they have the basic capacity to take paths away from a goal item to reach it, at least when the obstacle is clearly defined.

However, as illustrated in Figure 1a, many nonhuman animals and human infants find the problem challenging when the barrier is transparent, as they repeatedly attempt to reach directly for the reward item, even in the face of strong, negative feedback (Diamond, 1990; Santos, Ericson, & Hauser, 1999; Wallis et al., 2001). This insensitivity to feedback is typically explained as an inability to inhibit a lower-level behavioral control system (such as a Pavlovian system). However, other experimental conditions suggest not, at least for nonhuman primates (Santos et al., 1999). For example, when first given experience with an opaque barrier, subjects tend to solve the transparent barrier problem. The ease with which they refrained from reaching directly for the food item once they had an alternative response available suggests that the major difficulty did not stem from a lack of self-control.

We propose that the difficulty results from confusion with the transparent barrier: the subjects do not readily see that

there is a barrier. Although their response is blocked, there is no apparent reason for it, and so they continue to attempt the most efficient solution of reaching directly for the goal item. We suggest that this is an example in which a problem-solving system sees a clear solution and is therefore overriding feedback to the contrary. However, it is using a broken model. Put differently, it is an example of a cognitive illusion that provides insight into how cognitive systems are constructed (Kahneman, 2011).

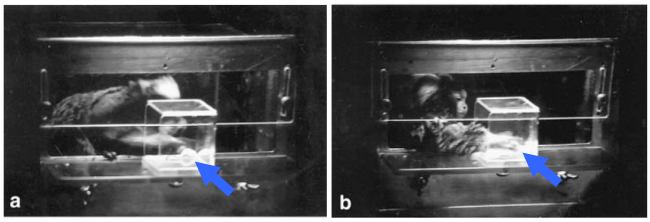


Figure 1: Illustration of a nonapparent problem. (**a**) The marmoset monkey attempts to reach directly for the marshmallow but is blocked by the transparent barrier. (**b**) The monkey learns to reach around the barrier.[1]

We further suggest that the transparent barrier problem requires a nonapparent solution. To solve it, the problem-solving system must include an obstacle in the problem formulation, thus fixing the model that did not include it. However, if an individual cannot see that an obstacle exists, it must be inferred from the effect of being blocked. Thus, taken together, we use the detour problem with a transparent barrier as an example of a nonapparent problem, and more specifically, as a case in which an individual confronts an unknown event or consequence, must infer a hidden cause, and create a (nonperceptual) concept to model it (Goswami, 2008; Holyoak & Morrison, 2012).

In addition, there is evidence that the ability to solve this nonapparent problem is subserved by a separate problem-solving system. Chiefly, it has been shown that rhesus monkeys who normally solve the transparent barrier problem lose this ability with lateral PFC lesions (Diamond, 1990). Thus, such findings implicate lateral PFC in mediating nonapparent problem solving, and suggest that it is separable from what we are calling apparent problem solving.

From these results and others, we model the primate brain with four basic levels of behavioral control. The first is based on the first main system in the vertebrate brain that controls complete goal-oriented behaviors: the hypothalamus (Swanson, 2000). That is, it is the first behavioral control system involving complete behavioral sequences that attain goals, such as goal-directed approach and ingestion behaviors when food is perceived. However, here we assume it is normally inhibited until the goal state is reached, and then is used to complete the process of actually obtaining and ingesting the food item. The second level represents the striatal-based habit system, which uses model-free reinforcement learning. The third level represents the first model-based problem solving system, i.e., that which solves *apparent* problems. This system would solve the detour problem with well-defined, opaque barriers, but would produce direct reaching with the ill-defined transparent barrier. The fourth level, then, performs *nonapparent* problem solving, and evidence suggests it is mediated by granular PFC. Evidence that granular PFC mediates nonapparent problem solving further suggests that mammalian vmPFC likely underlies the solving of apparent problems.

Because these systems of the primate mind/brain evolved under specific selection pressures, we also attempted to model these pressures to best understand the utility of each system and their interactions. More specifically, we used a foraging problem, and tested parameters to mimic basic selection pressures, such as size of the foraging environment, number of obstacles, and changing conditions, e.g., changes in the goal state position. In addition, because the primate brain evolved along a specific phylogenetic trajectory, rather than studying each system in isolation, we examine them from this phylogenetic perspective, in which each system appears to be added to a previous combination. Thus, very roughly, we compared the four-level primate brain model to an ancestral vertebrate brain, consisting of the first two levels, and an ancestral mammalian brain, consisting of the first three levels. We also compared the model to an alternative primate brain consisting of Levels 1, 2, and 4 (i.e., without the simpler *apparent* problem-solving system).

In this paper, we focus on the main difference between the levels and therefore only use transparent barriers (i.e., no opaque ones). Thus, when no obstacles are in the direct path, the apparent system can solve the problem; otherwise, the nonapparent system must be utilized.

In sum, the aims of the current project were (1) to begin specifying more clearly how nonapparent problems can be solved when the simpler model-based problem solving system fails; (2) to examine the potential advantages and disadvantages of a four-level system; and (3) to determine the foraging conditions that would provide a selective advantage for this brain architecture.

Methods

In what follows, we describe the modeling environment, and then layout the details of our model; we next present the four competing models, and then describe how the models were assessed.

Modeling environment

To focus on the main features of our model, we sought to keep the testing environment as simple and straightforward as possible. Therefore, we tested the model with a *foraging problem* using a *2D grid world*, in which the agent must learn a path from its current location to the goal location that avoids all obstacles (Figure 2).

[1]With permission from Wallis, Dias, Robbins, & Roberts (2001) © 2001 Federation of European Neuroscience Societies.

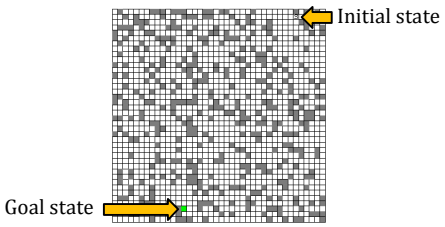


Figure 2: Example grid world, with the 'Initial state' denoted as "S" (Start) and the 'Goal state' in green. Obstacles are denoted in black, however, they are actually transparent (and thus hidden to the problem solver visually).

Model description

Our model consists of the four levels and a cognitive control mechanism that passes control between them.

Level 1 enables actual *goal attainment* once in view. For foraging, it completes the act of food consumption. This level is assumed in the four-level model, and is not explicitly modeled or tested here. With respect to the brain, Level 1 represents hypothalamus, which is considered the first behavioral control system involving complete behavioral sequences that attain goals (Swanson, 2000).

Level 2 represents the striatal-based habit system. It uses a *Markov Decision Process (MDP)* and *Reinforcement-Learning (RL) framework* (Sutton & Barto, 1998). Thus, there is a set of states, i.e., the (x, y) positions in the grid world; a set of possible actions, i.e., all eight reachable positions from a given grid world position; and a reward function that assigns values (called *action* or Q values) to the actions based on environmental feedback, with the Q values representing expected future reward. The agent then learns to choose the best action (i.e., policy) in a given state that leads to the highest future reward. Thus, rather than having a model of the world, i.e., an understanding of how the states relate to each other in the larger grid world, Level 2 sees the states independently, simply using a Q-Table to determine action values in each state (Daw et al., 2005). Level 2 is composed of *learning* and *acting* components. The *learner* updates the action values using the following Q-learning algorithm (a form of RL learning):

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\,[r_{t+1} + \gamma \max_a(s_{t+1}, a) - Q_t(s_t, a_t)]$$

where α is the learning rate $\alpha \in [0, 1]$; r_{t+1} is the actual reward received at episode $t+1$; and γ is the discount factor $\gamma \in (0, 1)$ (Sutton & Barto, 1998). The *actor* selects an action according to the Boltzmann distribution of Q values:

$$\pi(a|s_t) = \frac{e^{Q_t(s_t,a)/\tau}}{\sum_b e^{Q_t(s_t,b)/\tau}}$$

, where τ is the temperature that controls the degree of action exploration.

Cognitive control, i.e., control between levels, in the current model is generally hierarchical and modulatory (Kahneman, 2011; Miller & Cohen, 2001). Control begins at Level 2; however, when Level 2 fails (here, when the maximum Q value for an action is not unique), control passes to Level 3. Levels 3 & 4, representing *apparent* and *nonapparent* problem solving respectively, derive their behavioral control by modulating the actor in Level 2 (by modifying the action values and then passing control back to the Level 2 actor when a solution is found). If Level 3 fails to find a solution, it passes control to Level 4; and if Level 4 fails, Level 2 selects an action randomly. Future work will examine a more sophisticated controller based on, e.g., cost/benefit analysis (Daw et al., 2005; Kowaguchi et al., *accepted*).

Level 3 represents explicit problem solving when problem components are apparent, passing control to Level 4 when they are ill-defined and a solution cannot be found. Both Levels 3 and 4 look for solutions based on a cognitive model—i.e., a cognitive map—of the problem environment that is built in the background as the problem solver moves through it and experiences the state transitions: i.e., the subsequent state reached when an action is taken in a given state (with all transition probabilities = 1), e.g., $T(s_i, a_j) \rightarrow s'$, where s_i = (1, 1), a_j = *move right*; and s' = (2, 1), and thus, placing s' to the right of s_i in the map/model (Sutton & Barto, 1998; Daw et al., 2005). This model thus has an understanding of the relationships among the states in the grid world that Level 2 cannot see. Currently, we have one cognitive model that is built; however, Levels 3 and 4 contribute different elements to it and utilize what they have access to.

More specifically, for both Levels 3 and 4, all problems are considered *multi-agent problems*, with three classes of agents: self, others, and goals (Holyoak & Morrison, 2012; Shi et al., 2010; Wooldridge, 2009). All are considered agents because they could theoretically invoke a state change by virtue of their actions and functional relationships with other agents. (This would occur for the goal item if, for example, it were moving prey; however, in the current case, the goal item is stationary.) Thus, the cognitive model used by Levels 3 and 4 consists of four main components: (1) an x, y coordinate frame that defines each location in the grid world; (2) the identification and location of every agent in the problem; (3) the set of available actions each agent could take; and (4) an understanding of functional interactions among the agents (Goswami, 2008): i.e., f_i(*agent*$_j$, *target agent*$_k$), where f_i could be *acquiring* the goal item by the problem solver or *blocking* of the problem solver by an obstacle (also considered an agent by virtue of this blocking effect). For the current study, there is only one type of obstacle with only one available action: blocking.

For any novel problem in the grid world, the problem solver cannot see the entire problem immediately — the world is too large — and so a cognitive model must be developed via initial experience with each state. Model building entails developing the cognitive map of (x, y) coordinates for each state as well as whether an agent resides in each state. Again, because Level 3 can only see apparent obstacles, not invisible ones, it cannot see any of the transparent obstacles, and thus assumes there aren't any.

For problem solving, Level 3 uses the cognitive model to find a path to the goal. Since Level 3's view of the grid-

world problem appears clear of obstacles, it always finds a direct path to the goal. The potential advantage of this is that when there are clear paths to the goal, a brain that contains this system would provide fast and efficient solutions. Since Level 3 always sees a clear path, it will continue producing this solution (via hill climbing), leaving the problem solver in a potential phase of perseveration (i.e., continuing to attempt the same direct path solution). This impasse, then, is what causes cognitive control to pass to Level 4.

Level 4 assumes control when Level 3's model breaks. Level 4 therefore represents the system that attempts to find these nonapparent solutions. For the current project, this occurs every time a hidden cause, i.e., an 'invisible' obstacle, blocks the direct path to the goal. In this case, the obstacle is literally nonapparent to Level 3.

Level 4's main contributions occur with both the building of the overall cognitive model of the grid-world problem and utilizing it to solve the nonapparent problems. First, as the problem solver is moving in the problem space (the grid world) via the actor module, when it is blocked, Level 4 uses this information conceptually, inferring that there is an agent doing the blocking (Goswami, 2008; Holyoak & Morrison, 2012; Tenenbaum et al., 2011; Wynne & Udell, 2013). That is, it infers f_i(*obstacle*, *self*), where f_i is *blocking*. From this inference, Level 4 places a *blocking* agent at the grid location in the cognitive model. Unfortunately, Level 3 cannot see this nonvisual conceptual obstacle; it is beyond Level 3's comprehension.

Second, when cognitive control is passed to Level 4, it uses the complete cognitive model (including the transparent barriers) to find a path to the goal. To achieve this, Level 4 currently uses the planning algorithm A* (called 'A star') to find an efficient path around the obstacles to the goal (Russell & Norvig, 2010). Once a path is found, Level 4 then modifies the Level 2 action values so that the actor module will use the path. The main advantage of Level 4 over Level 2 (simple model-free RL) is that it can perform inductive inference, and use the internal cognitive map for mental simulation and planning, leading to rapid, one-trial learning via problem solving (as well as greater generalization to novel problems in future model development) (Passingham & Wise, 2012).

Figure 3 summarizes the key characteristics of each level.

Levels

1. See goal → obtain it
(e.g., Approach food → ingest)

2. **(a) Actor (action selection via Boltzmann distribution)**
(b) Learner (Q-learning algorithm)

3. (a) Builds internal cognitive model
(b) Hill-climbing search
(c) Change Level 2 Q-values for Actor

4. (a) Adds transparent obstacles to the internal model
(b) A* search
(c) Change Level 2 Q-values for Actor

Figure 3: Description of model levels.

Brain models

We compared our model to three others, thus testing four different multi-level models:

(1) Model 1: consisting of levels 1 & 2
(2) Model 2: levels 1, 2, & 3
(3) Model 3: levels 1, 2, & 4
(4) Our model, Model 4: levels 1, 2, 3, & 4

All four models assume the existence of Level 1. Model 1 simply uses model-free reinforcement learning, and roughly speaking, perhaps represents an ancestral vertebrate. Model 2 combines model-free RL with the ability to solve more straightforward, *apparent* problems, perhaps representing the ancestral mammalian brain. Model 3 represents a brain that contains both lower level model-free RL and the higher level *nonapparent* problem-solving system that one might argue should replace the simpler *apparent* problem-solving system altogether. Model 4 is our multi-level model of the primate brain.

Model assessment

We examined the effects of (1) grid world size, (2) number of obstacles, (3) changing initial states, (4) changing goal states, and (5) changing obstacles. The models were assessed via two measures (average from ~50 iterations): (1) *Cumulative number of steps* to reach the goal after 200 learning episodes (i.e., 200 times in which the goal item was reached); and (2) *Cumulative computational cost* needed per action, measured as the amount of processing time per action. The two measures were combined to obtain an overall *performance score*.

Simulation results

Because Level 1 was not explicitly modeled or examined here, results are from Levels 2-4. We maintain the color-coding of the model names to help keep them straight.

Grid world size

To examine the effects of grid world size, no obstacles were included. As seen in Figure 4a, performance was best for Model 1 with the smaller world and worse with the largest grid size. As the world size increased, Model 2 and Model 4 performed the best. Figure 4b shows the learning rates, and Figure 4c the cumulative computational costs for all four brain models for the largest grid world size (40x40). Model 1 was slower to learn a path to the goal (and progressively more so as the world size increased); in contrast, the other models all continued to learn very quickly. For all grid world sizes, the computational cost was greatest for Model 3 and lowest for Model 1.

Number of transparent obstacles

All remaining analyses used the large grid size. As the number of obstacles increased, Model 1 and Model 2 were slower to learn a path to the goal, while Model 3 and Model 4 continued to learn quickly. As shown in Figure 5, Model 4 performed relatively well across all numbers of obstacles,

Model 2 performed relatively well until the largest number, Model 3 was relatively better with the largest number of obstacles, while Model 1 performed the poorest at every number of obstacles.

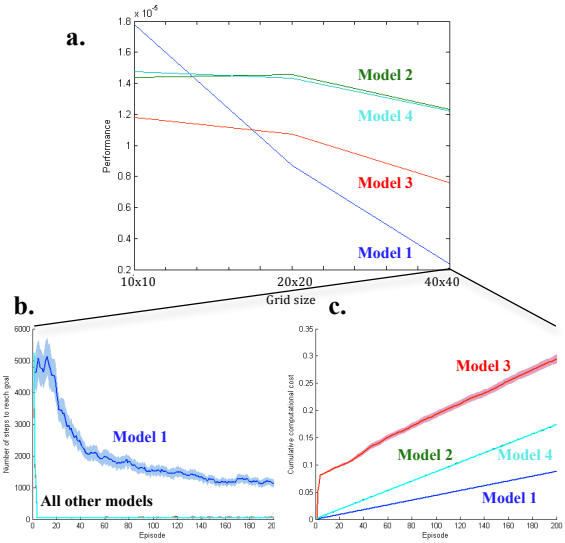


Figure 4: (**a**) Performance of the four brain models as a function of grid world size, (**b**) Learning rates, and (**c**) Cumulative computational costs for all four brain models for the largest grid world size (40x40). Bands around the curves represent standard error of the mean (SEM).

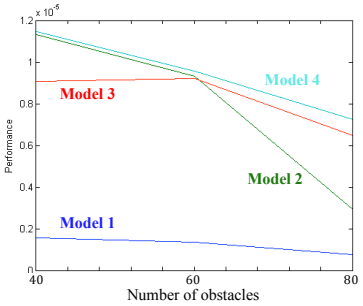


Figure 5: Model performance as a function of the number of transparent obstacles in the grid world.

Changing world

To examine the effects of a changing initial state and changing goal state, no obstacles were used.

Changing initial state The initial state of the problem solver was changed once every 50 episodes or once every 10 episodes. The models were generally robust with the changing initial states, although computational costs increased, especially for Model 3. As shown in Figure 6, overall performance was best for Model 2 and Model 4.

Changing goal state As seen in Figure 7a, rises in path length occurred when the goal state changed, especially for Model 1. With a changing goal state (and no obstacles), the computational costs for Model 3 were relatively high (Fig. 7b).

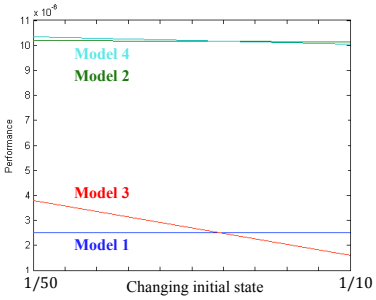


Figure 6: Model performance as a function of the rate of change of the initial state: once every 50 episodes or once every 10 episodes.

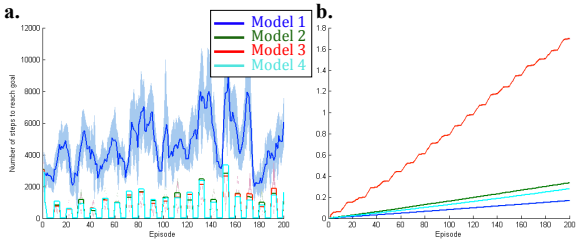


Figure 7: (**a**) Learning rates and (**b**) Cumulative computational costs for the highest rate of goal state location change (once every 10 episodes). Bands around the curves are SEMs.

As seen in Figure 8, Model 2 and Model 4 outperformed the others.

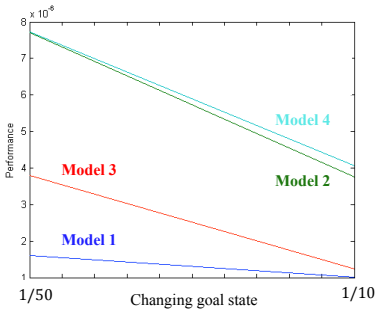


Figure 8: Performance as a function of changing goal state.

Changing (transparent) obstacles As shown in Figure 9, when the number of obstacles (600) and frequency of change (1/10) were high, Model 3 and Model 4 found a path to the goal most quickly (Fig. 9a), however, the computational demands were relatively steep (Fig. 9b).

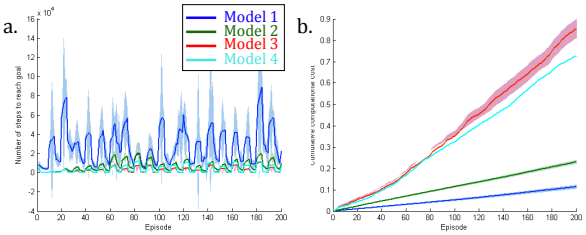


Figure 9: (**a**) Learning rates and (**b**) Cumulative computational costs for changing obstacles (600 total, once every 10 episodes). Bands around the curves are SEMs.

Discussion

Model-based problem solving potentially provides great advantages, such as lowering the number of errors during learning and generalizing to novel problems (Daw et al., 2005; Holyoak & Morrison, 2012). However, models of the real world are notoriously brittle, and thus require other approaches to problem solving when they break. It has been hypothesized that primates have evolved granular prefrontal cortex to cope with these more challenging, nonapparent problems. Here, the aims of our study were threefold: (1) to begin clarifying the mechanisms of nonapparent problem solving, (2) to examine the potential advantages and disadvantages of having four types of behavioral control systems, and (3) to determine the foraging conditions that would provide a selective advantage for this brain architecture. We used the classic detour problem in the comparative literature, and in particular, focused on the use of transparent barriers, which prove challenging for nonhuman animals. We suggest that an *apparent* problem-solving system, likely mediated by vmPFC, can solve detour problems with well-defined obstacles; while a *nonapparent* system, mediated by granular PFC, can solve the problem with 'invisible', transparent barriers. It does so by inferring the existence of a barrier from its effects on the problem solver. Thus, this system may underlie the powerful ability of humans to infer hidden causes from given events and consequences (Holyoak & Morrison, 2012; Tenenbaum et al., 2011). Other behavioral research we have conducted with monkeys suggests further possible mechanisms for the nonapparent system that help solve insight problems by both nonhuman primates and people (Murray, Kralik, Wise, 2005; Kowaguchi et al., *accepted*; Kralik, 2005, 2011). We plan to incorporate these findings in the future.

With respect to our second and third aims, the advantages of greater cognitive abilities must outweigh the disadvantages, and the advantages of a multi-level brain architecture appears to lie in its versatility. In all manipulations conducted here, our model (Levels 1-4) never performed the best outright; however, it was always among the best, appearing to be a jack-of-all-trades. Thus, rather than fitting a high-level cognitive niche best, our brain model appears to best fit a niche with problems of varying levels of complexity: a low-to-high cognitive niche. Thus, it may be useful to have multiple behavioral control systems at different levels of sophistication, which allow computational savings when facing simpler problems, and more elaborate capabilities when faced with more challenging ones (Kahneman, 2011; Rangel et al., 2008).

More theoretical development, however, is required to better understand the characteristics of such multi-level systems. For example, we plan further developments that include using a dynamic environment, different classes of agents that can both hinder or aid the problem solver in goal attainment, more sophisticated inductive reasoning and cognitive-control mechanisms, and further levels of abstraction (Botvinick, Niv, & Barto, 2009; Daw et al., 2005; Shi et al., 2010; Tenenbaum et al., 2011).

The ability to solve problems creatively across a wide range of domains embedded in complex, physical environments remains out of reach for current artificial systems; but we are extending their reach (Hélie & Sun, 2010). A detailed analysis of how it evolved in the human lineage should help to further demystify the creative process. Such an analysis can also help to clarify how primates in general, and humans in particular, have come to fill the low-to-high cognitive niche. Creative thinking represents the pinnacle of high-level cognition and underlies many of our greatest achievements. This success not only derives from the heights of thinking we can attain, but also the diversity of challenges we can master.

Acknowledgments

ONR Grant N00014-08-1-0693 supported the project.

References

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113, 262-280.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Diamond, A. (1990). Developmental time course in human infants and infant monkeys, and the neural bases of, inhibitory control in reaching. *Ann New York Ac Sci*, 608, 637-676.

Goswami, U. (2008). *Cognitive Development*. London, UK: Routledge.

Hélie, S. & Sun, R. (2010). Incubation, Insight, and Creative Problem Solving: A Unified Theory and a Connectionist Model. *Psychological Review*, 117, 994-1024.

Holyoak, K. J. and Morrison, R. G. (2012). *The Oxford Handbook of Thinking and Reasoning*. Oxford: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Kowaguchi, M., Patel, N. P., Bunnell, M. E., & Kralik, J. D. (*Accepted*). Competitive control of cognition in rhesus monkeys. *Cognition*.

Kralik, J.D. (2005). Inhibitory control and response selection in problem solving: How cotton-top tamarins (*Saguinus oedipus*) overcome a bias for selecting the larger quantity of food. *Journal of Comparative Psychology,* 119(1), 78-89.

Kralik, J.D. (2011). Rhesus Macaques (*Macaca mulatta*) spontaneously generalize to novel quantities in a reverse-reward contingency task. *Journal of Comparative Psychology*, 255-262.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167-202.

Murray, E. A., Kralik, J. D., & Wise, S. P. (2005). Learning to inhibit prepotent responses: successful performance by rhesus macaques, Macaca mulatta, on the reversed-contingency task. *Animal Behavior,* 69, 991-998.

Passingham, R. E., & Wise, S. P. (2012). *The neurobiology of the prefrontal cortex: Anatomy, evolution, and the origin of insight*. Oxford, UK: Oxford University Press.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall

Santos, L. R., Ericson, B. N., & Hauser, M. D. (1999). Constraints on problem solving and inhibition: Object retrieval in cotton-top tamarins. *J. Comparative Psychology*, 113, 186-193.

Shi, D., Sauter, M. Z., Sun, X., Ray, L. E., & Kralik, J. D. (2010). An extension of Bayesian game approximation to partially observable stochastic games with competition and cooperation. *Proceedings of the International Conference on Artificial Intelligence (ICAI)*.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press.

Striedter, G. F. (2005). *Principles of Brain Evolution*. Sunderland: Sinaur Assoc.

Swanson, L. W. (2000). Cerebral hemisphere regulation of motivated behavior. *Brain Research*, *886*(1-2), 113–164.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, Structure, and Abstraction. *Science*, 331, 1279-1285.

Wallis, J. D., Dias, R., Robbins, T. W., & Roberts, A. C. (2001). Dissociable contributions of the orbitofrontal and lateral prefrontal cortex of the marmoset to performance on a detour reaching task. *European Journal of Neuroscience*, *13*(9), 1797–1808.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. Chichester, UK: John Wiley & Sons.

Wynne, C., & Udell, M. A. R. (2013). *Animal Cognition*. London, UK: Palgrave Macmillan.

Learning in the wild - how labels influence what we learn

Samuel Rivera (rivera.162@osu.edu)
Department of Psychology, The Ohio State University
1835 Neil Avenue, Columbus, OH 43210 USA

Christopher W. Robinson (robinson.777@osu.edu)
Department of Psychology, The Ohio State University Newark
1179 University Dr., Newark, OH 43055, USA

Abstract

Learning concepts and categories in the real world is often accompanied by verbal labels. The existing theoretical accounts of how labels influence what we learn range from facilitation to overshadowing, with changes occurring over development. Studies investigating how labels influence what people learn have typically been confined to a category learning framework, where participants were tasked to learn how to discriminate categories or infer missing category properties. Here, we investigate how the absence or presence of labels, both common and unique, alter how people attend and what they remember in a more general setting. Our results suggest that unique labels may promote visual exploration of objects; whereas, there was no evidence to support the claim that hearing the same label associated with different members of a to-be-learned category directed attention to common features.

Keywords: categorization; cross-modal processing; attention; learning

Introduction

Every day we learn about objects in the world through experiences with our environment and communication with others. Language, and labeling in particular, is known to have important consequences for how we interpret and perceive categories in our environment (Lupyan, 2012; Perry & Lupyan, 2014), and effects of labeling on visual attention may illuminate a mechanism underlying how language may affect thought more generally (Whorf, 1956). The goal of this research is to understand how linguistic labels, both common and unique, influence what we learn from our environment.

Previous research has shown that adults expertly recognizing the link between labels and concepts (Lupyan, Rakison, & McClelland, 2007; Perry & Lupyan, 2014), and adults appear to appreciate the symbolic nature of linguistic input (Yamauchi & Markman, 2000). Referring to an object with the phrase "it's a jambo" immediately conveys what the object is, and that it belongs to a particular category of things. The powerful effect of labels exists even in infancy, with studies showing that linguistic labels, unlike non-linguistic sounds, facilitate category formation in very young infants (Althaus & Mareschal, 2014; Balaban & Waxman, 1997; Ferry, Hespos, & Waxman, 2010; but see Robinson & Sloutsky, 2007). The mechanism underlying this ability, especially early in life, remains an active and contentious research topic.

Improved learning has been theorized as stemming from labels indicating category membership in a natural kind class (Gelman & Markman, 1986), and may invite infants to form categories by denoting category membership and highlighting object commonalities (Waxman & Booth, 2003; Waxman & Markow, 1995). It is also possible that hearing different words associated with different objects directs attention to unique features. Support for this claim comes from research examining effects of labels on individuation (Xu, 2002, but see Robinson & Sloutsky, 2008). For example, young infants are more likely to assume that two different objects are hidden behind an occluder if the objects were associated with two unique labels. Infants do not make this assumption if they hear one label associated with both objects or if both objects are associated with two non-linguistic sounds.

The view that labels draw attention to common features and possibly unique features has been supported by behavioral and eye tracking experiments with young infants (Althaus & Plunkett, 2015a; Plunkett, Hu, & Cohen, 2008). However, it is important to note that other studies examining visual attention while learning categories have failed to find evidence that labels direct attention to category relevant features (Best, Robinson, & Sloutsky, 2011). Thus, it is possible that, early in life, labels can affect category learning without being perceived as symbols that denote category membership. They may instead be an additional perceptual cue that adds to the perceptual similarity of objects that are labeled with the same name (Sloutsky & Fisher, 2004; Sloutsky, 2003), but become more like category markers over development (Sloutsky, Lo, & Fisher, 2001).

A third proposal is that labels may actual hinder category learning early in life, with evidence of both words and non-linguistic sounds attenuating both categorization and individuation early in development (Robinson & Sloutsky, 2007, 2008). More specifically, while both of these studies provide some evidence that labels can have facilitative effects compared to non-linguistic sounds, this difference stemmed from non-linguistic sounds interfering with visual processing more than words, as neither the sound nor label condition exceeded a silent condition. This contradictory finding may be reconciled by recent research arguing that the precise synchrony of labels with visual object presentation has important consequences on whether or not labels will facilitate learning (Althaus & Plunket, 2015b; see also Roberts, 1995).

In summary, it has been proposed that labels have a facilitative effect on categorization and individuation in infants and young children by directing attention to common features and unique features, respectively (Waxman & Booth, 2003). However, most of the research supporting the facilitative effects of labels examines the outcome of learning (e.g.,

better learning in the label condition than a non-linguistics sound condition), and does not directly test the hypothesis that labels affect visual attention. Moreover, the research directly examining infants' and children's fixations while learning categories are mixed (Althaus & Plunkett, 2015b; Best et al., 2011). Therefore, the current study examined the effects of common and unique labels on adults' category learning.

Previous research suggests that for adults, labels are not just perceptual features (Yamauchi & Markman, 2000). Furthermore, adults optimize their attention to focus on category relevant features (Blair, Watson, & Meier, 2009; Hoffman & Rehder, 2010). Will adults who are experts in language learning and attention optimization direct their attention to common features when the to-be-learned individuals are associated with the same labels? Will adults fixate more on unique features when every member of the to-be-learned category is associated with a unique label? To address these questions, we taught adults three novel categories. One category was presented in silence (baseline), members of a second category were always paired with the same word (common label condition) and members of the third category were associated with unique labels (unique label condition). If common labels direct attention to common features, then fixating on common features in the common label condition should exceed the silent baseline. If unique labels direct attention to unique features, then adults in the unique label condition should look less to the common feature and more to unique features compared to the silent baseline.



Figure 1: We show example stimuli used in the experiment. Each category is defined by the presence of a particular feature: prongs on the shoulders for category A, a dorsal fin for category B, and suction cup feet for category C.



(a) Category A (b) Category B (c) Category C

Figure 2: We show examples of stimuli used in our study, with red circles denoting the areas of interest.

Method

Participants

Nineteen adults participated in the study for course credit after giving written consent. The study was evaluated and approved by an institutional review board at The Ohio State University.

Stimuli

Verbal labels were pre-recorded sound clips of count nouns in a child directed carrier phrase like "Look, here's a dax" or "this is called a dax" that were spoken by a female. Sound clips were recorded using a Yeti Pro microphone and edited using Audacity software so that each linguistic phrase was approximately one second in duration. Auditory stimuli were presented via Kensington KMW33137 headphones at approximately 65 dB. Visual stimuli were artificial creatures generated with the Spore creature creation software (Electronic Arts Inc., 2009). Creatures were from three different categories defined by the presence of a single deterministic feature. All other features varied independently of the category, and gave no information about category membership. Category A was defined by the presence of three small prongs on the front shoulders. Category B members has a dorsal fin, and category C creatures had suction-cup feet. The new items were perceptually impoverished, had fewer limbs, and were used as catcher items on test trials to make sure participants were paying attention during training.

We made an effort to normalize for the overall size and complexity of creatures within and across categories and exemplars. Examples are shown in Figure 1. On average, stimuli spanned approximately 18 by 13 degrees of visual angle and were presented on a 1920 x 1080 Benq XL2420-B monitor. The area of interest (AOI) for each creature was fixed size circle (or circles) enclosing the deterministic feature(s). Some example AOIs are shown in in Figure 2.

Procedure

We conducted a within subject study with eye-tracking to identify differential effects of labeling on attention allocation, memory for encountered objects, and attention to novel exemplars of studied categories. The label conditions encountered during training were Common (single label for category members), Silent (no labels or carrier phrase), or Unique (each category exemplar had a different label). Following training, participants performed an old vs. new recognition

task.

To begin, participants were told that they would see a series of creatures from another planet, and that they would learn some of the creatures' names. They were asked to study the creatures at their own pace since they would be asked questions in the second part of the study (testing phase). They were not told about the specific recognition task, nor were they told what they would be tested on (e.g., words, creatures, word-creature pairings), but they were informed about a test in general. Their gaze was recorded during training and testing at 500Hz by an Eyelink 1000 Plus Tower system (SR Research, Ontario, Canada). Fixation information was identified on-line during the experiment by the Eyelink system, then recorded for offline processing and analysis with custom MATLAB and Python software.

Each trial began with a central drift correction target that corrected for slight eye-tracker drift during the experiment. A trained experimenter sitting at the right side of the participant initiated each trial once the participant fixated the central cue. Stimuli were presented in a self-paced manner, in three blocks of 15 trials consisting of 5 items from each category. Participants pressed the 1 key on the number pad when they were ready to move to the next stimulus. Trials within blocks were randomized, and the same 15 items were presented in each block. Each category was consistently presented in either silence (category A), with a unique label phrase (category B), or with a common label phrase (category C). The auditory and visual stimuli shared the same stimulus onset. The carrier phrase and label terminated after about 1 s and the visual stimulus terminated when the participant pressed the 1 key.

After three training blocks, participants were presented with an old versus new recognition task. All of these test trials were presented in silence. Participants saw 5 exemplars from each of the three categories that had been previously studied (15 old items), 5 novel exemplars from the studied categories (15 lure items that had the category defining feature), and 6 completely new trials that looked completely different from any of the studied items. Each trial began with a drift correction target as before. Participants responded old or new to the stimulus by pressing the 1 key for old, and the 3 key for new. Participants were instructed to respond as fast and as accurately as possible.

If common labels facilitate categorization by directing attention to common features, then participants in the common label condition should focus more on the category defining features compared to the silent condition, and thus would be more likely to mistake critical lures as old or have worse recall in general in comparison to the unique label or silent conditions. If unique labels facilitate individuation by directing attention away from common features and to unique features, then participants in the unique label condition should focus more on unique features compared to the silent condition.

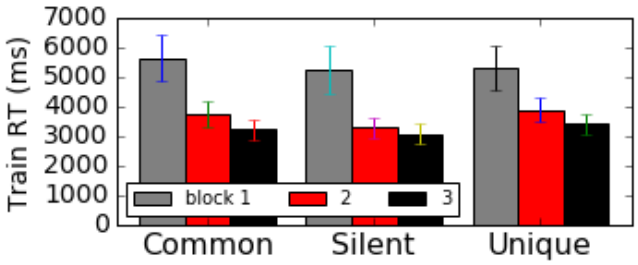


Figure 3: Here we show participant mean study time across conditions and blocks. Error bars denote standard error. Trends show that participants spent less time studying over the training blocks.

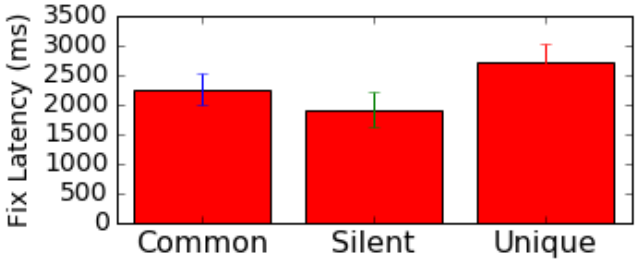


Figure 4: Here we show participant mean latency to fixate the deterministic feature. Error bars denote standard error.

Results

Training Results

Training study time We plot the mean study time in Figure 3. We ran a 3 (condition: common vs. silent vs. unique) x 3 (block) repeated measures ANOVA on log(studyTime). We found a significant effect of condition, $F(2,36) = 9.56, p < .001, \eta_p^2 = 0.35$. Pairwise comparisons showed that participants viewed the images for less time in the Silent condition than the Common ($p = 0.015$, Bonferroni corrected) and the Unique label conditions ($p < .001$). Thus, participants were faster when only the visual information was presented. There was also a main effect of block, $F(2,36) = 31.05, p < 0.001, \eta_p^2 = 0.63$. Pairwise comparison also showed that block 1 was significantly longer than blocks 2 and 3 ($ps < 0.001$). Thus, participants spent less time studying the creatures after the first block.

Training latency Deterministic Looks For several participants, there were blocks where the deterministic features were never fixated. Thus, we excluded these participants from this analysis. We considered the averages across all blocks such that only one participant was excluded. Averages are shown in Figure 4. We ran a one-way repeated measures ANOVA on latency to fixate the relevant AOI with main effect of condition (common vs. silent vs. unique). We found a significant main effect of condition, $F(2,34) = 10.78, p < .001, \eta_p^2 = .39$. Pairwise comparisons revealed that only the

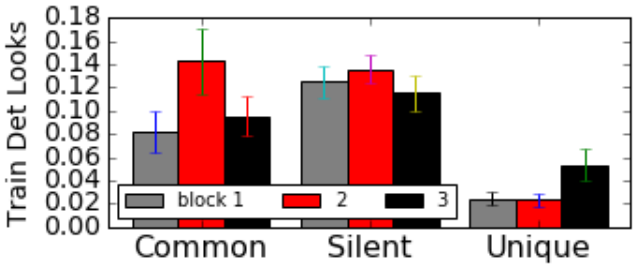


Figure 5: Here we show participant mean proportion of looking at the deterministic or category defining AOI. Error bars denote standard error.

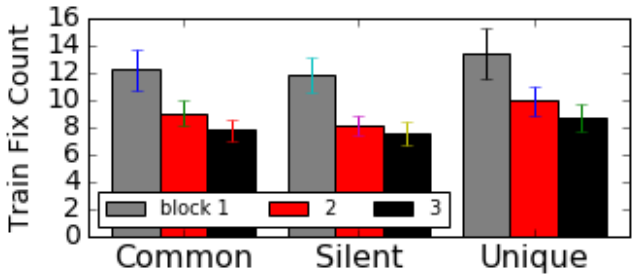


Figure 6: Here we show participant mean number of fixations during training across conditions and blocks. Error bars denote standard error.

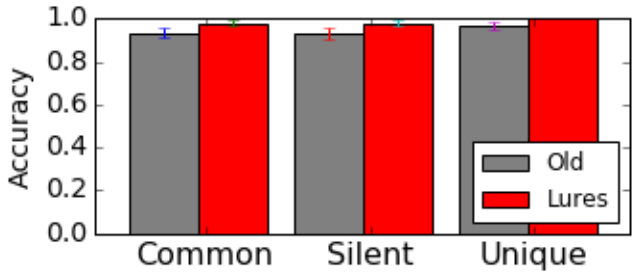


Figure 7: Here we show participant mean accuracy across conditions for both old and lure trials. Error bars denote standard error. Trends suggest that lure items were more accurately identified as new across all conditions. However the differences were not significant.

difference between the silent and unique conditions was statistically significant ($p < 0.001$, Bonferroni corrected), with latency in the silent condition being shorter. The difference between the silent and common was approaching significance ($p = 0.077$), with latency in the silent condition being shorter than the common label condition. Thus, participants in the silent condition fixated the deterministic feature earlier on trials during training. Is it possible that participants were in general fixating earlier on Silent trials? We ran a 3 (condition: common vs. silent vs. unique) x 3 (block) repeated measures ANOVA on latency of log(first fixation) to test this possibility. There were no significant main effects or interactions, ruling out the possibility that participants in the silent condition may have been faster overall.

Training - deterministic looking The latency analysis above is problematic because there were many trials where the deterministic feature was never fixated so that latency information was missing. Such cases bias the result because if the deterministic feature was never fixated, that trial was excluded from the analysis instead of providing a large latency value. A better measure, if we are interested in how label conditions contribute to deterministic looking is to consider the proportion of looking on a trial. Thus, no fixations at the deterministic feature would give a value of 0. We show the averages across conditions and blocks in Figure 5. We ran a 3 (condition: common vs. silent vs. unique) x 3 (block) repeated measures ANOVA on proportion of fixation time at the deterministic feature. We found a significant main effect of condition, $F(1.65, 29.78) = 21.39, p < 0.001, \eta_p^2 = 0.54$, Greenhouse-Geisser corrections. Pairwise comparisons revealed that the unique condition had a significantly lower proportion of fixation time at the deterministic feature than the other conditions ($ps < 0.003$, Bonferroni corrected). No other main effects or interactions were significant.

Training Fixation counts Given this difference in proportion of fixations at the deterministic feature, we wondered if this could be explained by an overall increase in the exploration of the object. Such behavior would transpire in an increase in the number of overall fixations. We plot means in Figure 6. A 3 (condition: common vs. silent vs. unique) x 3 (block) repeated measures ANOVA on number of fixations revealed a statistically significant main effect of condition, $F(2, 36) = 5.66, p = 0.007, \eta_p^2 = 0.24$. Pairwise comparisons showed that the effect was driven by a significant difference between the silent and unique condition (p=0.02, Bonferroni corrected), with more fixations in the unique condition as predicted. The number of fixations in the unique condition was trending larger than the common label condition, but the difference was not statistically significant.

There was also a significant main effect of block, $F(1.31, 23.54) = 13.63, p = 0.001, \eta_p^2 = 0.43$, Greenhouse-Geisser corrected for sphericity. Pairwise comparisons revealed that there were significantly more fixations in the first block than later blocks ($ps < 0.03$, Bonferroni corrected). Interaction between condition and block was not significant.

Test Results

Accuracy One participant reversed the response mapping, as indicated by 0% accuracy. We reversed their responses in calculating the accuracy scores below. We plot the mean participant accuracy in Figure 7. Overall, participant accuracy was near or at ceiling in the old vs. new judgement task, with accuracy above 95% across the three conditions. A 3 (condition: common vs. silent vs. unique) x 2 (test type: Lure

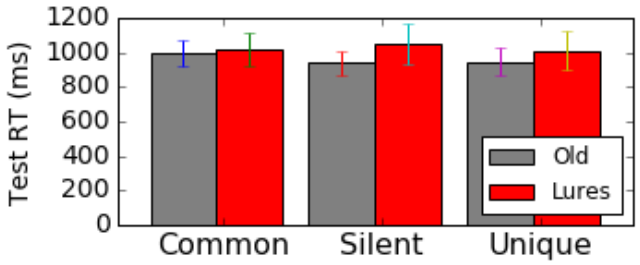


Figure 8: Here we show mean participant reaction times during test trials. Error bars denote standard error.

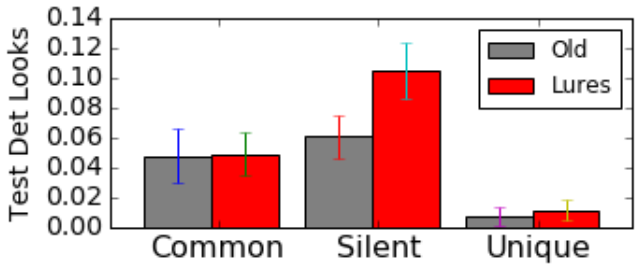


Figure 9: Here we show mean proportion of fixation time at the deterministic or category defining feature. Error bars denote standard error.

vs. Old) repeated measures ANOVA on old/new recall accuracy revealed a statistically significant main effect of test type, $F(1,17) = 5.64, p = 0.03, \eta_p^2 = 0.25$, with accuracy on lure items being higher. No other effects or interactions were significant.

Testing RT Mean reaction times during test trials are shown in Figure 8. We applied a 3 (condition: common vs. silent vs. unique) x 2 (test type: Lure vs. Old) repeated measures ANOVA to log(RT) on accurate test trials. There were no significant main effects or interactions.

Testing proportion of deterministic looking We plot the proportion of fixation time at the deterministic AOI in Figure 9. A 3 (condition: common vs. silent vs. unique) x 2 (test type: Lure vs. Old) repeated measures ANOVA on proportion of deterministic looking on test trials revealed a significant main effect of condition, $F(2,34) = 13.16, p < 0.001, \eta_p^2 = 0.44$. Pairwise comparisons showed that the effect was driven by the unique condition having a significantly lower proportion than the common label condition ($p = 0.031$, Bonferroni corrected) and the silent condition ($p < 0.001$). This result is consistent with the gaze pattern during training.

Testing Fixation counts As with the training data, we were interested in whether the relatively smaller proportion of fixations at the deterministic feature could be explained by an overall increase in the number of fixations in the unique condition. We show the average number of fixations in Figure 10.

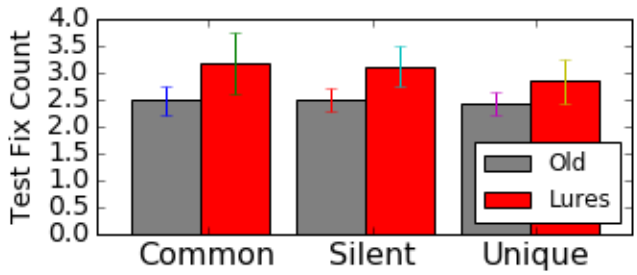


Figure 10: Here we show mean number of fixations during testing trials. Error bars denote standard error.

A 3 (condition: common vs. silent vs. unique) x 2 (test type: Lure vs. Old) repeated measures ANOVA on number of fixations revealed no significant main effects or interactions. Thus, across conditions participants fixated the same amount when making their old/new judgements. However, the unique label condition was associated with less focus on the deterministic feature(s).

Discussion

In the current study we investigated how the presence or absence of labels affect adults' visual attention while viewing novel visual stimuli. Previous research with infants and young children shows that infants are more likely to learn categories when the to-be-categorized images are paired with the same label (Balaban & Waxman, 1997) and more likely to individuate exemplars or learn two categories when the different objects are associated with different labels (Plunkett et al., 2008; Xu, 2002). One potential mechanism that may account for these effects is that infants understand the symbolic nature of language, with common and unique words (but not sounds) directing infants attention to common and unique features, respectively (Waxman & Booth, 2003). However, recent eye tracking studies directly testing this hypothesis have yielded mixed results (Althaus & Plunkett, 2015a; Best et al., 2011); therefore, this study examined visual fixations in adults, who clearly understand the symbolic nature of words and can optimize their attention (Blair et al., 2009; Lupyan et al., 2007; Rehder & Hoffman, 2005).

Participants in the current study required less exposure to images and were faster to make fixations to category defining features when the images were presented in silence. This is consistent with some of the research examining auditory dominance in a variety of cognitive tasks (see Robinson et al. 2012 for a review). Across many different behavioral and eye tracking variables, there was no support for the hypothesis that common labels directed attention to category relevant features; however, there was some support showing that hearing unique labels associated with category members decreased attention to commonalities or increased attention to unique features.

A goal of future research is to adapt this initial study to

examine labeling effects in younger populations. One advantage of the study is that it mirrors situations encounter outside the laboratory, where the way knowledge will be applied is unknown. However, this article is limited in that stimuli are defined by only a small subset of defining features, and the defining features did not vary across labeling phrases. Thus, future work should consider categories comprised of multiple defining features with proper counter-balancing of these categories with the labeling conditions.

In summary, studies have demonstrated important influences of labeling on category learning. While many of these studies show that labels and sounds can have different effects on categorization and individuation, underlying mechanism are poorly understood. The current study took a novel approach by examining fixations in real time as adults learned about category objects. While there was some support showing that unique labels push attention to unique features or away from category defining features, there was no support showing that participants looked more at defining features or made faster fixations to these defining features. These findings have implications for proposed mechanisms underlying the effects of labels on category learning and should be further examined in younger populations.

References

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS ONE*, *9*(7).

Althaus, N., & Plunkett, K. (2015a). Categorization in infancy: Labeling induces a persisting focus on commonalities. *Developmental Science*.

Althaus, N., & Plunkett, K. (2015b). Timing matters: The impact of label synchrony on infant categorisation. *Cognition*, *139*, 1–9.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month old infants? *Journal of Experimental Child Psychology*, *64*, 3–26.

Best, C. A., Robinson, C. W., & Sloutsky, V. M. (2011). The effect of labels on categorization: Is attention to relevant features a good index of infants category learning? In L. Carlson, C. Hlscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2751–2755).

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning.

Ferry, A., Hespos, S., & Waxman, S. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. *Child Development*(2), 472–479.

Gelman, S. a., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of experimental psychology. General*, *139*(2), 319–340.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label feedback hypothesis. *Frontiers in Psychology*, *3*, 54.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.

Perry, L. K., & Lupyan, G. (2014). The role of language in multi-dimensional categorization: Evidence from transcranial direct current stimulation and exposure to verbal labels. *Brain and Language*, *135*, 66 - 72.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665–681.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41.

Roberts, K. (1995). Categorical responding in 15-month-olds: Influence of the noun-category bias and the covariation between visual fixation and auditory input. *Cognitive Development*, *10*, 21–41.

Robinson, C. W., Best, C. A., Deng, S., & Sloutsky, V. M. (2012). The role of words in cognitive tasks: What, when, and how? *Frontiers in Psychology*, *3*, 1–8.

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, *11*(3), 233–253.

Robinson, C. W., & Sloutsky, V. M. (2008). Effects of auditory input in individuation tasks. *Developmental Science*, *11*, 869–881.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*(6), 246 –251.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: a similarity-based model. *Journal of experimental psychology. General*, *133*(2), 166–188.

Sloutsky, V. M., Lo, Y. F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development*, *72*(6), 1695–1709.

Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, *6*, 130–137.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive psychology*, *29*(3), 257–302.

Whorf, B. (1956). *Language, thought, and reality* (J. B. Carroll, Ed.). Cambridge, Massachusetts: MIT Press.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*, 223–250.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 776–795.

Cross-Linguistic Similarities Aid Third Language Learning in Bilinguals

James Bartolotti (j-bartolotti@u.northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Viorica Marian (v-marian@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Abstract

Learning a new language involves significant vocabulary acquisition. Learners can accelerate this process by relying on words with native-language overlap, such as cognates. For bilingual third language learners, it is necessary to determine how their two existing languages interact during novel language learning. A *scaffolding* account predicts transfer from either language for individual words, whereas an *accumulation* account predicts cumulative transfer from both languages. To compare these accounts, twenty English-German bilingual adults were taught an artificial language containing 48 novel written words that varied orthogonally in English and German wordlikeness (neighborhood size and orthotactic probability). Wordlikeness in each language improved word production accuracy, and similarity to one language provided the same benefit as dual-language overlap. In addition, participants' memory for novel words was affected by the statistical distributions of letters in the novel language. Results indicate that bilinguals utilize both languages during third language acquisition, supporting a scaffolding learning model.

Keywords: Bilingualism; Language learning

Introduction

Knowledge of multiple languages is a desirable skill. Eighty percent of Americans (Rivers, Robinson, Harwood, & Brecht, 2013) and 84% of Europeans (European Commission Special Barometer, 2006) believe that adults should be fluent in a second language, but multilingual rates are far lower than desired levels. Bilingualism in the United States is estimated at 20-26% (United States Census Bureau, 2007), and even in the European Union, where primary and secondary school instruction in two foreign languages is widespread, only 56% of adults report fluency in a second language (European Commission Special Barometer, 2006).

One factor that makes language learning difficult for adults is prior language knowledge. In particular, first language (L1) experience tends to sharpen the mind to features and regularities of future L1 input, such as word forms (Schmitt, 1997). A consequence of this linguistic sharpening is that adults become particularly attuned to learning new vocabulary in their native language (Long & Shaw, 2000). Throughout life, people rapidly learn new vocabulary; a professional psycholinguist quickly learns words like morpheme, electroencephalogram, or aphasia. Because of their similarity to existing words and patterns, acquisition comes relatively seamlessly, even to people who may have struggled to learn a foreign language.

While linguistic sharpening can facilitate acquisition of native-like words, it can also interfere with learning of words that do not match native language patterns. This has important implications for novel language learning, because lexical similarity drops across languages. Written English words, for example, have on average 5-7 times fewer neighbors (i.e., words that differ in a single letter) in related languages like Dutch, French, German, or Spanish than they do English neighbors (Marian, Bartolotti, Chabal, & Shook, 2012). Thus, when attempting to learn a new language, the L1 is less likely to confer a benefit.

When overlap between languages does exist, it can be a powerful tool. For example, cognates, which overlap in both form and meaning across languages, are easier to acquire (De Groot & Keijzer, 2000). Partial similarity is also beneficial, with learners better able to retrieve words with L1-familiar patterns during immediate recall (Storkel, Armbrüster, & Hogan, 2006; Thorn & Frankish, 2005). Novel words' L1 overlap can be characterized in several different ways; two useful metrics are neighborhood size and orthotactic probability. A broad definition of a word's neighborhood size is the number of known words that differ from it in the substitution, addition, or deletion of a single letter, and orthotactic probabilities calculate how often individual letters or sequences of letters are used in a language.

Complete word learning is a complex, protracted process, and only after a period of memory consolidation is a new word fully integrated into the lexicon (Gaskell & Dumay, 2003). Acquisition of a word's form, though, occurs relatively quickly, making form a useful target for investigating the initiation of word learning. The mechanism by which form similarity benefits language learning is the focus of the current investigation. In this study, we compare two possible models for how long-term knowledge affects novel vocabulary learning: the Scaffolding and the Accumulation models.

The Scaffolding model predicts that the ability to create a direct association between a newly-encountered word and an existing word or concept drives memory strength. Novice learners rely heavily on L1 translations during L2 vocabulary learning (Liao, 2006; Schmitt, 1997), which anchors the relatively weak novel word to a strong existing memory. In the Revised Hierarchical Model of bilingual language processing, these word-to-word associations are strongest at the onset of L2 learning (Kroll & Stewart, 1994). Beyond translation equivalents, other lexical associations can be used to remember discrete aspects of a word. For example, the keyword learning method (Shapiro & Waters, 2005) is a pedagogical

approach that emphasizes using a known word as a form intermediary between a novel word and its meaning (e.g., the word steel can be used to remember the phonological form of the French word *stylo*, meaning pen). A novel word with a more native-like form is more likely to have closely related L1 keywords available that the learner can select from to create a personally meaningful association. The scaffolding account thus emphasizes the learner's ability to directly utilize their existing linguistic framework during language acquisition.

In contrast, the Accumulation model proposes that a novel word's consistency with lexicon-wide patterns affects the fidelity with which it is represented in short-term memory and retrieved from long-term memory. When a new word is first encountered, it is vulnerable to disruption, but rehearsal processes maintain the trace in the phonological loop until it can be stored in memory. How well information is represented and maintained in the phonological loop is affected by interactions with prior knowledge, as detailed in Baddeley's Working Memory model (Baddeley, 1986). As applied to word learning, this long term knowledge may be used to enhance the strength of the initial temporary storage during encoding (Gathercole, 2006). Novel words with more native-like features are easier to repeat and maintain in working memory, because the sequencing of their letters or sounds is more predictable. In addition, newly-learned words that are composed of high probability patterns can benefit more from redintegration, the process of reconstructing a partially decayed short-term memory using long-term knowledge. Thus, a word that was incompletely encoded may nevertheless be accurately produced, reinforcing the target representation. To again use the French word *stylo* as an example, it contains common English bigrams *st* and *lo*, and the frequency of these features can facilitate accurate encoding and retrieval. These sub-lexical effects are distinct from the one-to-one whole word associations that drive learning in the scaffolding model.

In order to compare the Scaffolding and Accumulation models of word learning, we make use of a population with unique language experience and investigate acquisition of a third language in bilingual adults. We taught English-German bilinguals 48 words in an artificial language that varied orthogonnaly in their similarity to English and German. The two models make different predictions for how the three languages interact during word learning and retrieval. The scaffolding account predicts comparable learning benefits for novel words that resemble one or both known languages, because a single lexical-level link to existing knowledge is sufficient to advance learning. For each new word, quality links in either language may be available, but only one is utilized. In contrast, the accumulation account predicts tiered learning, where novel words that resemble both known languages are learned better than those that resemble a single language. This pattern is driven by cumulative sublexical frequency effects from each known language.

Methods

Participants

Twenty English-German bilinguals participated for monetary compensation. Informed consent was obtained in accordance with Northwestern University's IRB. After the experiment, participants completed the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian, Blumenfeld, & Kaushanskaya, 2007) to assess proficiency, age of acquisition, and percentage of current usage for each language. Participants also completed the LexTALE in English and German, which assesses vocabulary knowledge based on lexical decision accuracy (*LexTALE*, Lemhöfer & Broersma, 2012) (Table 1).

Table 1: Bilingual language backgrounds, mean and SD.

Measure	English	German
Proficiency (1-10)	9.50 (0.72)	8.27 (1.52)
Age of Acquisition	3.84 (5.04)	10.74 (7.69)
Current Usage (%)	75.47 (19.30)	16.18 (13.51)
Vocabulary Size (0-100)	95.22 (4.99)	77.35 (14.58)

Materials

Forty-eight orthographic CVCC words were created in an artificial language. The novel words' English and German wordlikeness were calculated as composite scores of English and German orthographic neighborhood size and orthotactic probability (sum of grams and sum of bigrams), calculations from CLEARPOND (Marian et al., 2012), and English and German word similarity judgments obtained from English-German bilinguals (N = 10, ratings on scales of 1-5). Words were divided into four groups based on median splits of the English and German wordlike composites. Fourteen words had both high English and high German wordlikeness (E+G+, e.g., *nist* or *baft*), ten had high English but low German wordlikeness (E+G- e.g., *sumb* or *gonk*), eleven had low English but high German wordlikeness (E-G+ e.g., *gach* or *kenf*), and the remaining thirteen had low English and German wordlikeness (E-G- e.g., *gofp* or *kowm*). Each novel word was paired with a color line drawing from the revised Snodgrass and Vanderwart picture set (Rossion & Pourtois, 2004). Pictures were chosen to be highly recognizable (naming reliability: M = 99.1%, SD = 2.0%, Bates et al., 2003), and did not overlap orthographically or phonologically in English or German with their paired novel-language words. The names of pictures used in each of the four conditions did not differ on lexical frequency, orthographic or phonological neighborhood size, or gram, bigram, phoneme, or biphone probabilities in English or German (calculations from CLEARPOND, Marian et al., 2012).

Procedure

Participants began training with a single exposure block of 48 randomized trials to familiarize them with the novel language. In each exposure trial, a picture was presented in the

center of the computer screen, and the written novel-language target appeared below the picture. Trials advanced automatically after two seconds. Following the exposure block, participants performed five blocks of word recognition with feedback, and five blocks of word production with feedback, alternating between the two tasks.

Word learning: Recognition. In 48 recognition trials, a random target picture and three randomly selected foil pictures were displayed in the four corners of the screen, and the written target word appeared in the center of the screen. The participant clicked on the picture that matched the word; accuracy and response time were recorded. After making a response, the three foils disappeared, and the target picture and written word remained onscreen for 1000 ms to facilitate continued learning, followed by a 1000 ms inter-trial interval.

Word learning: Production. In 48 production trials, a random target picture was presented in the center of the screen. The participant typed the name of the picture in the new language and accuracy was recorded (RTs were not analyzed due to the continuous response nature of typing in the task). After making a response, the picture and the participant's answer remained on the screen, and the correct name of the target was printed below the participant's response for 1000 ms as corrective feedback, followed by a 1000 ms inter-trial interval. After completing all 48 trials, a new testing block of recognition and production began. After the fifth series of recognition and production blocks the experiment concluded.

Data Analysis

Accuracy in each task was automatically scored by the computer. Response times in the recognition task were measured from the onset of the display until the participant clicked on a picture. RT analyses were performed on correct responses only, in order to control for accuracy differences between blocks and conditions. Outlier RTs within each combination of Block and Condition were identified and replaced with the threshold value (Mean + 2SD).

Change across blocks in recognition accuracy and RT, and in production accuracy was analyzed using growth curve analysis (Mirman, Dixon, & Magnuson, 2008), a form of multilevel regression that simultaneously estimates the effects of individuals and of experimental manipulations on time-course data. Accuracy and RT were fit with two model levels. The Level-1 submodels captured the effect of time on changes in the dependent measure over the course of training using second-order orthogonal polynomials. The intercept term describes the overall height of the curve, the linear term reflects the slope, and the quadratic term reflects the curvature. The Level-2 submodels capture the effects of experimental manipulations and individuals on each of the time terms present in the Level-1 model through a combination of population means, fixed effects, and random effects. In the current study, the fixed effects corresponded to English- and German-wordlikeness. The random effects captured individual deviances from the global mean and condition means. The effects of individual differences in English and German background on learning were assessed by correlating each measure with individuals' random effect estimates in each model.

The base Level-2 model included all time terms, fixed effects of English and German proficiency, and random effects of participant and participant-by-condition on all time terms. Additional Level-2 models were built that added three fixed effects of English wordlikeness (E+/E-), German wordlikeness (G+/G-), and their interaction to each time variable in turn. A significant improvement in model fit (a Chi-squared test on the change in model fit using -2LogLikelihood) indicates an effect of condition on independent properties of the curve (i.e., height, slope, or curvature). Parameter-specific *p*-values were estimated by using a normal approximation, treating the *t*-value from the model as a *z*-value.

Results

Word Production

Word production accuracy improved from 10.3% (SD = 14.8) to 66.3% (SD = 26.1) over blocks one to five.

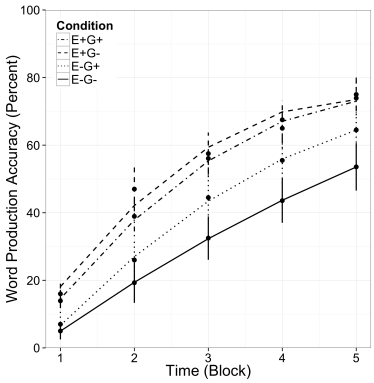


Figure 1: Novel word production accuracy (dots indicate observed values, and vertical lines standard error) and best fit quadratic growth curve models. Both English and German wordlikeness increased overall accuracy (intercept height), but the two factors did not have an additive effect.

The production accuracy model (Figure 1) was improved by adding English and German wordlikeness to the intercept (ΔLL = 25.82, $\chi^2(3) = 51.64$, $p < .001$) and to the quadratic term (ΔLL = 4.89, $\chi^2(3) = 9.78$, $p < .05$). The overall height of the curve was increased by English (Estimate = 0.218, SE = 0.023, $p < .001$) and by German wordlikeness (Estimate = 0.087, SE = 0.023, $p < .001$), and there was a significant interaction between the two terms (Estimate = -0.117, SE = 0.033, $p < .001$). The combination of the two terms revealed that whereas Englishlikeness improved accuracy by 21.8% and Germanlikeness improved accuracy by 8.7% relative to the unwordlike baseline, they combined non-additively, as the E+G+ double-wordlike condition was only 18.8% above baseline. Additionally, the benefits of English and German

wordlikeness were not equivalent, as the height of the learning curve for E-G+ words was significantly lower than both the E+G- words (Estimate = 0.131, SE = 0.023, $p < .001$) and the E+G+ words (Estimate = 0.101, SE = 0.023, $p < .001$).

The curvature of the learning gains over time (i.e., the quadratic term in the model) was significantly affected by Englishlikeness (Estimate = -0.098, SE = 0.033, $p < 0.01$), but not by Germanlikeness (Estimate = -0.045, SE = 0.033, n.s.). The baseline quadratic term was also not significant (Estimate = -0.029, SE = 0.031, n.s.), and together, these results indicate that whereas accuracy gains between blocks were nearly linear for baseline words, Englishlikeness had a non-linear effect on change in accuracy over time, with the largest accuracy gains between blocks occurring earlier during training.

Word Recognition

Participants' recognition accuracy improved from 61.6% (SD = 19.0) in the first block to 98.4% (SD = 3.7) in the fifth block. RTs became faster over time, from 3325 ms (SD = 593) in the first block to 2208 ms (SD = 493) in the fifth block.

The word recognition accuracy model (Figure 2) was improved by adding the wordlike predictors to the intercept ($\Delta LL = 4.52$, $\chi^2(3) = 8.10$, $p < .05$). Englishlikeness raised the overall height of the curve (Estimate = .0.036, SE = 0.016, $p < .05$), reflecting consistently higher accuracy for the Englishlike words compared to un-Englishlike words of 3.6% over the course of training. For recognition RT (Figure 3), there was a significant improvement to the base model by adding English and German wordlikeness to the intercept ($\Delta LL = 22.16$, $\chi^2(3) = 44.32$, $p < .001$). English wordlikeness reduced RT relative to baseline (Estimate = -0.406, SE = 0.061, $p < .001$), and there was a marginal decrease in RT by German wordlikeness (Estimate = -0.109, SE = 0.061, $p < .1$). Novel words that resembled English were thus correctly identified 406 ms faster than baseline words, whereas words that resembled German were identified 109 ms faster than baseline, with no interaction between the two factors. These wordlike increases were stable across training, even as RTs globally decreased by 1117 ms from blocks one to five.

Language Proficiency

The random effect terms in the accuracy and RT models quantify how much individual participants' performances deviated from the group mean. We correlated the random effects with measures of language aptitude and relative language balance.

Relative differences in proficiency were associated with differences between Englishlike and Germanlike learning rates. Specifically, for production accuracy, higher proficiency in English relative to German was associated with larger slopes (faster learning rate) for the Englishlike relative to the Germanlike words, and conversely, higher relative German proficiency was associated with a faster learning rate for the Germanlike words ($r = .479$, $R2 = .217$, $p < .05$).

Overall proficiency was associated with recognition accuracy. English vocabulary size (LexTALE score) was correlated with a higher intercept ($r = .647$, $R2 = .419$, $p < .01$) and a shallower slope ($r = -.624$, $R2 = .389$, $p < .01$). English and German self-rated proficiencies were each marginally correlated with higher intercepts (English proficiency, $r = .423$, $R2 = .179$, $p < .1$; German proficiency, $r = .407$, $R2 = .221$, $p < .1$) and shallower slopes (English proficiency, $r = -.470$, $R2 = .166$, $p < .05$; German proficiency, $r = -.433$, $R2 = .187$, $p = .05$). In each case, greater language knowledge increased accuracy; the shallower slopes reflect reaching ceiling performance.

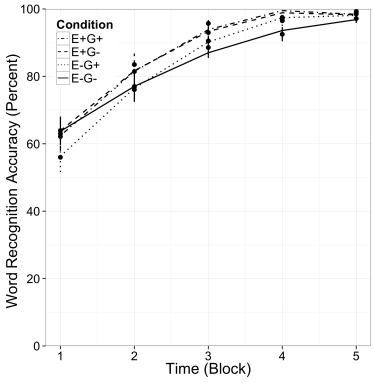


Figure 2: Novel word recognition accuracy (dots indicate observed values, and vertical lines standard error) and best fit quadratic growth curve models. English wordlikeness increased overall accuracy (intercept height).

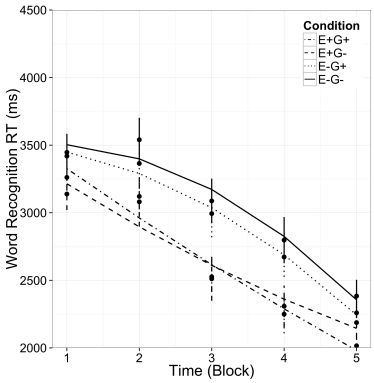


Figure 3: Novel word recognition RT (dots indicate observed values, and vertical lines standard error) and best fit quadratic growth curve models. English wordlikeness decreased overall response time across blocks and there was a marginal effect of German wordlikeness (intercept height).

Discussion

In the current study, we found that lexical similarity to a single known language improved bilinguals' learning of novel written words as much as simultaneous overlap with two known languages. These results indicate that vocabulary learning in a third language benefits from each language, and that bilinguals can flexibly transfer L1 and L2 knowledge to the L3 as appropriate at early stages of instruction. The lack

of an additive learning benefit for words with close lexical neighbors and familiar patterns in both languages suggests that early vocabulary transfer may occur through a process of linking novel words to anchors in a single language. This process most closely resembles the scaffolding model of word learning, with limited evidence for the accumulation model.

Two tasks, word recognition and production, were used to asses complementary aspects of novel word learning. The recognition task probed the formation of word-meaning links, whereas the production task assessed recollection of an exact L3 written word form. Recognition accuracy was high even after a single training exposure. In the first block, participants recalled roughly 30 of the 48 pairs correctly, and quickly approached ceiling performance. English, but not German, wordlikeness increased overall accuracy across training relative to baseline words. In addition, participants with larger English vocabularies performed better on the task, with higher accuracies across training. Marginal correlations between individuals' proficiency in either language and accuracy suggest the possibility of a generalized vocabulary size benefit on recognition learning. For word recognition, there is more support for an effect of English similarity than of German; this difference between languages reflects the overall higher English proficiency in our sample. Because the novel words were presented in their entirety during the recognition task, accurate performance depended not on memory for wordforms, but on the link between form and meaning.

The production task, in contrast, was designed specifically to probe participants' memories for the actual written forms in the artificial language. With only a single picture prompt, the participant's task was to type the matching word from memory. In this task, we saw evidence for strong effects of wordlikeness both in English and in German. All three wordlike conditions (E+G+, E+G-, and E-G+) had higher curve heights than the baseline (E-G-) words, indicating higher accuracy throughout the experiment. However, the combination of English and German wordlikeness was not additive, as the E+G+ words were no different from the better of the two single-wordlike conditions, E+G-. This pattern of results provides support for the scaffolding account, by which the novel words received a benefit to learning if they overlapped with at least one of bilinguals' known languages, but received no additional benefit for overlapping both languages.

Bilingual third language learners appear to be especially sensitive to perceived overlap, and will transfer vocabulary knowledge from their two languages preferentially based on typological similarity, regardless of other factors like age of acquisition (Cenoz, 2003). As a result, we would expect participants in the current study to transfer knowledge from the language of overlap for E+G- or E-G+ words, and either language for E+G+ words. In post-experiment debriefings, 95% of participants reported using a keyword learning method (Shapiro & Waters, 2005) to learn words' meanings by creating a mental image linking the meaning and word via a similarly spelled existing word. Participants' use of English and German anchors tended to overlap with the target word's proscribed category. For example, to learn that the novel E+G- word *sumb* meant *fork*, one participant imagined counting the *sum* total of forks in a drawer. Another participant learned that that the E-G+ word *kenf* meant *goat* by thinking of a goat eating *senf*, the German word for mustard. Words that resembled both English and German, like the word *duch* meaning *eyeglasses*, were sometimes learned through an English linking word (a *duck* wearing glasses), but other times a German link (using glasses to read a *buch*, the German word for book). By not limiting themselves to a single language of transfer, learners were able to maximize the benefits they gained from their existing knowledge. This pattern would also predict language-specific benefits for English or German monolinguals, and comparable performance between monolingual and bilingual groups on dual-overlap E+G+ words.

While both languages provided benefits to memory for novel word forms, the sizes of the effects were not equivalent. English wordlikeness, compared to German, had a larger effect on overall production accuracy, and led to a slightly different curvature over the course of training. These patterns are consistent with participants' proficiency asymmetry. Although participants were highly proficient in both languages, all participants were currently living and working in the United States, and had slightly higher English proficiency. Second language proficiency can affect the degree to which it influences third language learning (Hammarberg, 2001), and accordingly, the largest accuracy gains were seen for English wordlikeness. At the individual level, however, we found that learning patterns were influenced by relative proficiencies in English and German. Bilinguals with higher proficiency in English learned to produce the Englishlike words at a faster rate than the Germanlike words, while the opposite was true for those with higher proficiency in German. This difference may reflect either the relative ease of acquisition of individual words, or an attention allocation strategy that prioritized words resembling the learner's dominant language.

To conclude, we found evidence for effects of wordlikeness in each of a bilinguals' two languages on third language orthographic word learning. Memory for word forms was often improved by linking novel vocabulary to existing lexical anchors in either language, depending on the similarity of the novel word to lexical patterns in English and German. Importantly, a novel word's similarity to both of a bilingual's known languages does not provide an additional learning benefit beyond similarity to a single language, suggesting that orthographic knowledge does not necessarily combine additively during third language learning. These results provide support for the scaffolding account of word learning, in which existing language knowledge provides a framework upon which novel words in another language can be built, accelerating early stages of language acquisition. The persistance of these similarity effects on lexical integration and long-term retrieval will be a critical test in applying these

findings to language instruction. Foreign language instruction has long placed a premium on total immersion, but there is growing evidence that comparisons to the native language can benefit learning (Lin, 2015). We demonstrate here the learning benefits to be gained by utilizing areas of overlap between a known and a novel language.

Acknowledgments

This research was supported in part by grants NICHD R01 HD059858 to VM and T32 NS 47987-8 to JB. The authors would like to thank the members of the Northwestern University *Bilingualism and Psycholinguistics Research Group* for comments on this work.

References

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford Univeristy Press.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., ... Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, *10*(2), 344–380. doi:10.3758/BF03196494

Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: A review. *International Journal of Bilingualism*, *7*(1), 71–87. doi:10 . 1177 / F13670069030070010501

De Groot, A. M. B. & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning*, *50*(1), 1–56. doi:10.1111/0023-8333.00110

European Commission Special Barometer. (2006). *Europeans and their languages* (tech. rep. No. December 2005). Retrieved from http://ec.europa.eu/public%7B%5C_%7Dopinion/archives/ebs/ebs%7B%5C_%7D243%7B%5C_%7Den.pdf

Gaskell, M. & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. doi:10.1016/S0010-0277(03)00070-2

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543. doi:10 . 1017 / S0142716406060383

Hammarberg, B. (2001). Roles of L1 and L2 in L3 production and acquisition. In J. Cenoz, B. Hufeisen, & U. Jessner (Eds.), *Cross-linguistic influence in third language acquisition: psycholinguistic perspectives.* (Vol. 31, pp. 21–41). Multilingual Matters.

Kroll, J. F. & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. doi:10.1006/jmla.1994.1008

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. doi:10.3758/s13428-011-0146-0

Liao, P. (2006). EFL learners' beliefs about and strategy use of translation in English learning. *RELC Journal*, *37*(2), 191–215. doi:10.1177/0033688206067428

Lin, A. M. (2015). Conceptualising the potential role of L1 in CLIL. *Language, Culture and Curriculum*, *28*(1), 74–89. doi:10.1080/07908318.2014.1000926

Long, L. L. & Shaw, R. J. (2000). Adult age differences in vocabulary acquisition. *Educational Gerontology*, *26*, 651–664. doi:10.1080/03601270050200644

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PloS One*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*, 940–967. doi:10.1044/1092-4388(2007/067)

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. doi:10.1016/j.jml.2007.11.006

Rivers, W. P., Robinson, J. P., Harwood, P. G., & Brecht, R. D. (2013). Language votes: Attitudes toward foreign language policies. *Foreign Language Annals*, *46*(3), 329–338. doi:10.1111/flan.12048

Rossion, B. & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. doi:10.1068/p5117

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.

Shapiro, A. M. & Waters, D. L. (2005). An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. *Language Teaching Research*, *9*(2), 129–146. doi:10.1191/1362168805lr151oa

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. doi:10.1044/1092-4388(2006/085)

Thorn, A. S. C. & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 729–735. doi:10.1037/0278-7393.31.4.729

United States Census Bureau. (2007). *American Community Survey*. Retrieved from http://www.census.gov/acs/www/

Varieties of experience: A new look at folk philosophy of mind

Kara Weisman
(kweisman@stanford.edu)
Department of Psychology
Stanford University

Carol S. Dweck
(dweck@stanford.edu)
Department of Psychology
Stanford University

Ellen M. Markman
(markman@stanford.edu)
Department of Psychology
Stanford University

Abstract

Philosophers, psychologists, and neuroscientists have often divided the mind into fundamental component parts. Does this intuition carry over into folk philosophy of mind? In a series of large-scale studies, we explore intuitive distinctions among different kinds of mental phenomena and consider how these distinctions might organize the conceptual space of the diverse "intelligent" and "social" entities in the modern world. Across studies, independent exploratory factor analyses reveal a common latent structure underlying mental capacity attributions, centered on three types of phenomenal experiences: *physiological* experiences of biological needs (e.g., hunger, pain); *social-emotional* experiences of self- and other-relevant emotions (e.g., guilt, pride); and *perceptual-cognitive* abilities to detect and use information about the environment (e.g., hearing, memory). We argue for an expanded model of folk philosophy of mind that goes beyond agency and experience (H. M. Gray, Gray, & Wegner, 2007) to make basic and important distinctions among different varieties of experience.

Keywords: mind perception; folk theories; sentience.

Introduction

The ontology of the mind or soul has been a topic of great interest to humankind, from ancient philosophers to modern neuroscientists. In Plato's *Republic*, Socrates argued for a tripartite division of the soul into reason, spirit, and appetite, while Aristotle, in *De Anima*, posited four faculties of the soul: nutrition, perception, mind, and desires. Half a world away, the Buddha described sentient beings as aggregations of five components: material form, feelings, perceptions, impulses, and consciousness. Over two millennia later, the tradition continues, from Freud's model of the human psyche as composed of id, ego, and superego, to ongoing attempts to derive data-driven "cognitive ontologies" from neural activity (Hastings, et al., 2014). Does the intuition that the mind is composed of distinct parts carry over into *folk* philosophy of mind? Here, we set aside questions about the true organization of the mind to explore how lay people conceptualize mental capacities and mental life.

Many converging traditions in psychology and philosophy suggest that folk philosophy of mind might include a broad distinction between internal experiences and behavioral outputs (e.g., Gelman & Spelke, 1981; Knobe & Prinz, 2008). In an influential investigation of folk philosophy of mind, Gray, Gray, and Wegner (2007) conducted a large survey in which participants compared the relative mental capacities of various characters (e.g., a frog vs. a robot; a man vs. God). Participants' judgments of capacities for hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy covaried, forming a dimension that Gray et al. termed *experience*. Judgments of self-control, morality, memory, emotion recognition, planning, communication, and thought also hung together; they termed this second dimension *agency*. Gray et al. proposed that these two dimensions—experience and agency—organize people's understanding of different kinds of minds, playing a particularly important role in the identification of moral patients (capable of experience and therefore vulnerable to harm) and moral agents (capable of intentional behavior and therefore responsible for their actions).

However, we suspect that there is more to the lay ontology of mind than broad categories of experience and agency—particularly in light of Gray et al.'s (2007) untraditional analytical approach. In preparing their data for dimension reduction, the authors collapsed across many paired comparisons to estimate (non-independent) scores for 13 target characters on 18 mental capacities, and then performed a principal components analysis on this 13x18 dataset. In contrast, guidelines for dimension reduction generally recommend a much higher ratio of observations to variables (see Fabrigar, Wegener, MacCallum, & Strahan, 1999). Given the constraints of their approach, Gray et al.'s analysis would not be expected to yield much more than one dimension of mind perception; indeed, we note that the *experience* dimension actually accounted for nearly all of the variance in their data (88%, compared to 8% for *agency*). These findings thus leave open the possibility that there may be more than two dimensions of mind perception. In particular, lay people may make distinctions between different kinds of "experience," with potentially important consequences for social and moral reasoning.

In line with this, in our previous work we have proposed a three-part model of the lay concept of sentience, including two distinct forms of experience: *affect*, the ability to experience positively or negatively valenced states; and *perception*, the ability to detect information about the environment. (The third component in our model is *autonomy*, similar to Gray et al.'s (2007) "agency.") Our studies have demonstrated that when adults or young children learn that an unknown entity has one of these kinds of experience they do *not* strongly infer that it has the other; instead, affect and perception appear to be conceptually distinct (Weisman, Markman, & Dweck, 2015).

If, indeed, people consider affect and perception to be distinct capacities that do not mutually imply each other, they might also distinguish between entities that have one vs. both of these experiential capacities. For example, perceptual abilities might allow some autonomous being to

have a sensory impression of its environment, but suffering—a hallmark of higher-order mental and moral life—requires the ability to evaluate which sensations are pleasant or unpleasant. As Gray and others have argued (e.g., K. Gray, Young, & Waytz, 2012), attributions of intentional action and suffering are fundamental to moral reasoning. If these mental phenomena are thought to require combinations of distinct capacities for affect and perception, then differentiating between these varieties of experience might be as important in sociomoral reasoning as the broader distinction between experience and agency.

With these considerations in mind, the current studies examine US adults' attributions of a variety of mental phenomena, including various affective, perceptual, agentic, physiological, cognitive, social, and other capacities. Building on Gray et al.'s (2007) data-driven approach, we probe ontological distinctions among mental capacities in three large-scale studies. Converging dimension reduction analyses lead us to propose an expanded model of folk philosophy of mind, focused on intuitive distinctions among different varieties of experience.

Study 1

We begin by exploring people's attributions of mental capacities to two "edge cases" in social reasoning: a beetle and a robot. We selected targets whose existence is beyond question, but whose mental capacities were predicted to be controversial. This ensured that not all participants would endorse all mental capacity attributions (as they might if the target were a human), providing the variance necessary for the planned dimension reduction analyses. In addition, this provided a glimpse into how lay people currently think about robots, as social technologies begin to play increasingly larger roles in our everyday lives.

Methods

Participants. 405 adults participated via Mechanical Turk. All participants had gained approval for ≥95% of previous work (≥50 assignments); had verified US MTurk accounts; and indicated that they were ≥18 years old. Participants were paid $0.30 for about 3-4 minutes of their time. Repeat participation was prevented. An additional 48 respondents were excluded for not completing the survey (*n*=14), failing an attention check (19), or not providing a year of birth (15).

Materials and procedure. Participants were randomly assigned to evaluate either *a beetle*, accompanied by a photograph of a black beetle on a leaf (*n*=200); or *a robot*, accompanied by a photograph of a humanoid robot (Sony's Qrio; *n*=205). The picture and label ("a beetle" or "a robot") were present throughout the survey.

Participants read the following instructions: "On the following page, you will see a list of mental capacities. For each mental capacity, please indicate the extent to which you believe a [beetle/robot] has this capacity. Please note: We care only about your opinion or best guess—please do not do any external research about these questions." Participants then rated 40 mental capacities presented in a random order, responding to the following question: "On a scale of 0 (Not at all capable) to 6 (Highly capable), how capable is a [beetle/robot] of...?" An attention check ("Please select 4 for this question") was embedded randomly among the ratings, and respondents who failed excluded from analyses (see *Participants*).

The 40 mental capacities were generated from an a priori analysis of candidate ontological categories of mind: physiological experiences of biological needs (e.g., *getting hungry*); emotional experiences (*feeling happy*); perceptual experiences (*detecting sounds*); cognitive abilities (*remembering things*); capacities related to autonomy or agency (*having intentions*); social abilities (*experiencing guilt*); and several additional items that could have fallen into either none or more than one of these categories (*being conscious*). Each category included at least five items of varying valence, complexity, and phrasing. All 18 mental capacities from Gray et al.'s (2007) study, or close variants thereof, were included. See Table 1 for the full set of items.

Exploratory factor analysis (EFA).[1] For all EFAs reported in this paper, we used Pearson correlations to find minimum residual solutions. We examined maximal (39-factor) unrotated solutions to determine how many factors to extract. We report factor loadings from varimax-rotated solutions that included only factors that had eigenvalues >1.0 and that individually accounted for >5% of the total variance in the maximal model.

Results and Discussion

Collapsing across conditions, the first three factors of an unrotated EFA accounted for 68% of the variance in the data, with eigenvalues of 15.37 (explaining 46% of total variance), 4.36 (13%), and 3.09 (9%); all other factors individually explained ≤5% of total variance.

After rotation, the first factor captures a continuum from embodied physiological experiences of biological needs to non-bodily computational abilities, with factor loadings >0.60 for the following items, in descending order: *getting hungry*, *experiencing pain*, *feeling tired*, *experiencing fear*, *experiencing pleasure*, *being conscious*, *having free will*, *feeling safe*, *having desires*, *feeling calm*, and *feeling nauseated*. One item had a strong negative loading: *doing computations* (-0.74); there were no other loadings <-0.29.

The second factor corresponds to social-emotional experiences, with factor loadings >0.60 for the following items: *feeling embarrassed*, *experiencing pride*, *feeling love*, *experiencing guilt*, *feeling depressed*, *feeling disrespected*, *holding beliefs*, *understanding how others are feeling*, *experiencing joy*, *having a personality*, *feeling happy*, and *telling right from wrong*. No items had loadings <-0.08.

The third factor includes a mix of perceptual experiences and cognitive abilities, with factor loadings >0.60 for the

[1] Factor analyses using polychoric correlations and/or oblimin rotation; principal components analyses; correspondence analyses; and item response analyses all yielded similar structures.

Table 1: Factor loadings from exploratory factor analyses for all studies (S1-S3)

A priori category	Item *How capable is a [target] of...?*	Factor 1: Physiological S1	S2	S3	Factor 2: Social-emotional S1	S2	S3	Factor 3: Perceptual-cognitive S1	S2	S3
PHY	*getting hungry**	**0.93**	**0.93**	**0.84**	0.01	0.10	0.11	-0.08	-0.05	0.34
PHY	*experiencing pain**	**0.93**	**0.90**	**0.86**	0.10	0.16	0.14	0.01	0.01	0.33
PHY	*feeling tired*	**0.83**	**0.82**	**0.85**	0.23	0.31	0.21	0.10	0.04	0.35
EMO	*experiencing fear**	**0.82**	**0.78**	**0.83**	0.28	0.37	0.20	0.06	0.10	0.37
COG	*doing computations*	**-0.74**	**-0.80**	-0.41	0.19	0.08	0.51	0.44	0.27	0.40
	*experiencing pleasure**	**0.74**	**0.70**	**0.79**	0.43	0.51	0.36	0.11	0.11	0.35
	*being conscious**	**0.70**	**0.69**	**0.65**	0.36	0.41	0.36	0.12	0.20	0.38
AGE	*having free will*	**0.70**	**0.69**	0.59	0.37	0.40	0.42	0.09	0.18	0.39
PHY	*feeling safe*	**0.70**	**0.69**	**0.73**	0.36	0.35	0.33	0.13	0.16	0.36
	*having desires**	**0.69**	**0.73**	**0.68**	0.40	0.44	0.44	0.11	0.18	0.34
EMO	*feeling calm*	**0.65**	0.59	**0.75**	0.41	0.49	0.35	0.17	0.25	0.32
PHY	*feeling nauseated*	**0.65**	**0.64**	**0.70**	0.50	0.50	0.40	0.09	0.06	0.26
EMO	*getting angry**	0.58	0.54	**0.67**	0.57	**0.62**	0.47	0.08	0.11	0.31
AGE	*having intentions*	0.54	0.54	0.48	0.35	0.33	0.48	0.27	0.34	0.44
	being self-aware	0.52	0.49	0.38	0.48	0.48	0.59	0.22	0.26	0.35
PER	*detecting odors*	0.45	0.54	0.58	-0.01	0.04	0.22	0.43	0.41	0.58
SOC	*feeling embarrassed**	0.19	0.18	0.28	**0.85**	**0.75**	**0.82**	-0.01	0.02	0.09
	*experiencing pride**	0.28	0.24	0.43	**0.85**	**0.77**	**0.74**	0.05	0.08	0.20
SOC	*feeling love*	0.37	0.42	**0.65**	**0.81**	**0.70**	**0.58**	0.06	0.15	0.18
SOC	*experiencing guilt*	0.26	0.19	0.31	**0.80**	**0.76**	**0.82**	0.02	0.07	0.13
SOC	*feeling disrespected*	0.37	0.35	0.51	**0.78**	**0.75**	**0.63**	0.03	0.06	0.23
EMO	*feeling depressed*	0.25	0.21	0.29	**0.78**	**0.75**	**0.78**	0.04	0.07	0.18
COG	*holding beliefs*	0.11	0.10	0.19	**0.76**	**0.64**	**0.82**	0.12	0.15	0.14
SOC	*understanding how others are feeling*†	0.06	0.09	0.32	**0.70**	**0.62**	**0.72**	0.29	0.35	0.30
EMO	*experiencing joy**	0.51	0.50	0.71	**0.70**	**0.70**	0.53	0.10	0.14	0.25
	*having a personality**	0.23	0.21	**0.63**	**0.66**	**0.62**	0.54	0.31	0.37	0.33
EMO	*feeling happy*	0.55	0.52	**0.74**	**0.65**	**0.68**	0.50	0.10	0.17	0.23
	telling right from wrong†	-0.04	-0.10	0.17	**0.60**	0.51	**0.80**	0.32	0.37	0.25
COG	*having thoughts*†	0.50	0.50	**0.60**	0.55	0.55	0.50	0.22	0.33	0.33
AGE	*exercising self-restraint*†	0.24	0.19	0.24	0.55	0.56	**0.70**	0.31	0.27	0.38
COG	*remembering things*†	-0.20	-0.15	0.25	0.17	0.19	0.41	**0.72**	**0.65**	**0.71**
SOC	*recognizing someone*	-0.29	-0.27	0.43	0.29	0.29	0.39	**0.71**	**0.66**	0.59
PER	*sensing temperatures*	0.21	0.40	0.41	-0.06	-0.06	0.06	**0.66**	0.58	**0.72**
SOC	*communicating with others*†	-0.02	-0.11	0.34	0.20	0.16	0.31	**0.65**	**0.60**	**0.71**
AGE	*working toward a goal*†	0.09	0.16	0.23	0.17	0.21	0.41	**0.62**	0.56	0.57
PER	*perceiving depth*	0.11	0.20	0.28	0.11	0.11	0.32	**0.62**	0.58	**0.68**
PER	*detecting sounds*	0.06	0.20	0.45	-0.05	-0.03	0.08	**0.61**	**0.64**	**0.75**
PER	*seeing things*	0.36	0.37	0.56	-0.07	-0.04	0.13	**0.61**	0.55	**0.67**
AGE	*making choices*	0.25	0.27	0.37	0.18	0.21	0.37	**0.60**	0.59	**0.67**
COG	*reasoning about things*	-0.06	-0.13	0.16	0.47	0.41	**0.66**	0.57	0.59	0.48
	Percentage of total variance explained:	46%	47%	63%	13%	11%	10%	9%	8%	7%

Note: Factor loadings >0.60 or <-0.60 are in bold. The full set of items, used for all studies reported here, is listed in the second column. Each item is listed with its a priori category membership (first column): physiological (PHY); emotional (EMO); perceptual (PER); cognitive (COG); agentic (AGE); social (SOC); and other/multiple (unmarked). Items marked with an asterisk (*) or a dagger (†) constituted Gray et al.'s (2007) "experience" and "agency" dimensions, respectively.

following items: *remembering things*, *recognizing someone*, *sensing temperatures*, *communicating with others*, *working toward a goal*, *perceiving depth*, *detecting sounds*, *seeing things*, and *making choices*. No items had loadings <-0.09.

See Table 1 for the full set of factor loadings.

These results suggest that three latent constructs guided participants' assessment of the target characters included in this study: *physiological experiences*, characterized by embodied sensations related to biological needs; *social-emotional experiences*, characterized by positive or negative valence and relevance to the self and/or social partners; and *perceptual-cognitive abilities*, characterized by the detection and use of information about the environment. Interestingly, this analysis did not reveal any factor corresponding to agency or autonomy, as Gray, et al. (2007) and Weisman, et al. (2015) would predict. Instead, distinctions among varieties of experience dominated the correlation structure of participants' judgments when they were asked to evaluate the mental capacities of a beetle or a robot in isolation.[2]

Study 2

In Study 1, each participant evaluated a single entity in isolation. Study 2 was a within-subjects replication of Study 1, providing an opportunity to evaluate the reliability of this framework and to examine whether this way of thinking about minds is altered when people are presented with a salient contrast between an animate and an inanimate entity.

Methods

Participants. 400 adults participated via MTurk and were paid $0.50. An additional 24 respondents were excluded for not completing the survey (n=13), failing the attention check (7), or not providing a year of birth (4).

Materials and procedure were identical to Study 1, except that all participants rated both entities. Half of participants saw the beetle on the left side of the screen and half saw the beetle on the right. Although ratings for the two entities were made simultaneously, they were independent (e.g., a participant's rating of a beetle's capacity for joy did not constrain her rating of a robot's capacity for joy).

Results and Discussion

The first three factors of an unrotated EFA accounted for 67% of the variance in the data, with eigenvalues of 15.55 (explaining 47% of total variance), 3.77 (11%), and 2.63 (8%); all other factors individually explained ≤5% of total variance. After rotation, all three factors were very similar to those revealed in Study 1, corresponding to physiological experiences, social-emotional experiences, and perceptual-cognitive abilities; see Table 1.

A within-subjects design, which encouraged participants to compare an animate being with a "social" technology, revealed a very similar three-factor structure, distinguishing among physiological, social-emotional, and perceptual-cognitive experiences. This framework for mind perception appears to be quite robust, at least in participants' reasoning about "edge cases" like beetles and robots.

Study 3

In Studies 1 and 2, participants evaluated entities that we considered to be controversial in terms of their mental capacities. Were the distinctions uncovered in these studies specific to reasoning about edge cases, or would they apply more to reasoning about a wider range of entities? More broadly, how does the lay ontology of mind uncovered in Studies 1 and 2 organize the range of potentially "mental" entities people encounter in the world? In Study 3, we presented a variety of entities ranging from an inert object (a stapler) to a canonical social partner (a human adult).

Methods

Participants. 431 adults participated via MTurk and were paid $0.30. An additional 40 respondents were excluded for not completing the survey (n=15), failing the attention check (24), or not providing a year of birth (1).

Materials and procedure were identical to Study 1, except that participants were randomly assigned to evaluate one of the following entities (labeled as follows and accompanied by a photograph): *an adult*, *a child*, *an infant*, *a person in a persistent vegetative state*, *a fetus*, *a chimpanzee*, *an elephant*, *a dolphin*, *a bear*, *a dog*, *a goat*, *a mouse*, *a frog*, *a blue jay*, *a fish*, *a beetle*, *a microbe*, *a robot*, *a computer*, *a car*, or *a stapler*. The number of participants per condition ranged from 17 (*stapler*) to 24 (*dog*).

Results and Discussion

Once again, three factors emerged from the correlation structure of participants' mental capacity attributions, distinguishing social-emotional, physiological, and perceptual-cognitive abilities. Notably, this framework was revealed even when canonical minds, such as humans and familiar mammals, were evaluated. In fact, the three (unrotated) factors accounted for the vast majority of the variance in Study 3 (80%), with eigenvalues of 22.77 (63% of total variance), 3.72 (10%), and 2.42 (7%); all other factors explained ≤2% of total variance. Rotated factor loadings were very similar to Studies 1-2; see Table 1.

Target characters varied widely in their judged mental capacities (see Figure 1): While a human adult was seen to be highly capable of all mental capacities and a stapler was seen to be incapable of any, judgments of other targets revealed a diverse range of attribution patterns between these extremes. Non-human mammals were judged to be highly capable of most physiological experiences and many perceptual-cognitive abilities and to have middling social-emotional capacities. Non-mammalian animals were judged to have weaker capacities across the board, particularly in the social-emotional domain. See the General Discussion for comments on the particularly interesting case of the robot.

[2] A direct replication of Study 1 yielded very similar results, although the third factor accounted for only 5% of total variance.

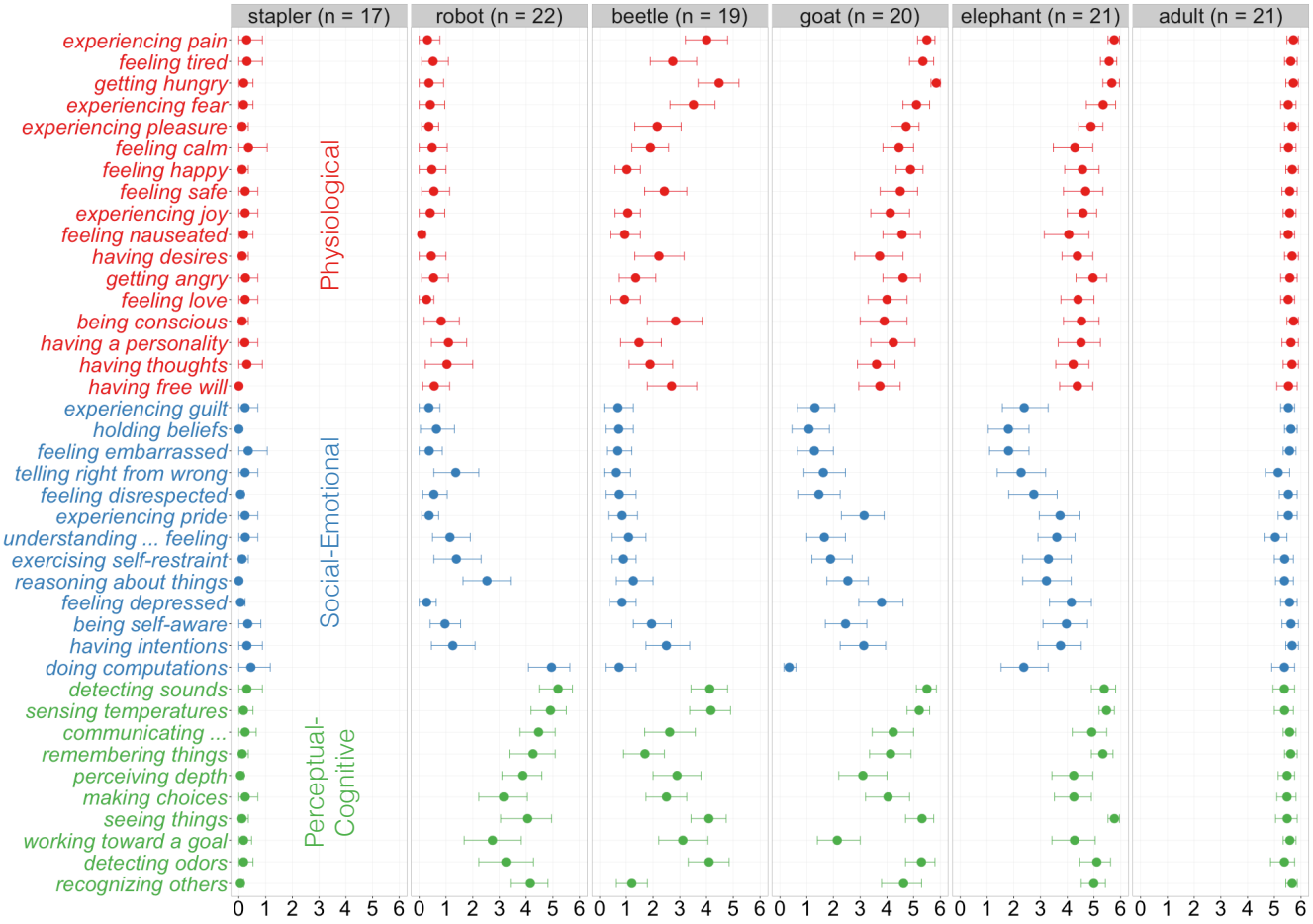


Figure 1: Mean ratings by mental capacity for a subset of target characters (Study 3)

Note. Target characters were rated on a scale from 0 ("Not at all capable") to 6 ("Highly capable"). Error bars are bootstrap 95% confidence intervals. Mental capacities are grouped according to their dominant factor loading in Study 3; see Table 1.

General Discussion

In three large-scale studies, we set out to explore what, if any, distinctions people make between categories of mental phenomena—to examine, in other words, a folk ontology of mind. Our results revealed reliable intuitive distinctions between physiological experiences of biological needs (e.g., hunger, pain); social-emotional experiences of self- and other-relevant emotions (e.g., guilt, pride); and perceptual-cognitive abilities to detect and use information about the environment (e.g., hearing, memory).

This three-factor structure seems to be quite robust. First, we note that participants each rated a wide variety of mental capacities, from multiple a priori domains, that varied in valence and complexity. Given this experimental design, additional or alternative latent factors—e.g., complex cognitive abilities, negatively valenced experiences, experiences of the self, etc.—could have emerged, but they did not. Furthermore, we observed very similar factor structures across independent analyses, both when participants judged a single "mental edge-case" in isolation (Study 1) and when participants were encouraged to compare two edge-cases that contrasted in animacy (Study 2). Finally, when a wider range of entities was included—from humans and other mammals down to microorganisms, technologies, and an inert object—this three-factor framework accounted for fully 80% of the variance in participants' judgments (Study 3). Given these observations, we conclude that distinctions among varieties of experience loom large in people's intuitive ontology of mind.

Interestingly, the agency/autonomy construct predicted by both Gray et al. (2007) and Weisman et al. (2015) did not emerge as a separate factor in any of the current studies. Instead, items that we predicted to be related to agency or autonomy were evenly distributed across the physiological (*having free will*, *having intentions*), social-emotional (*exercising self-restraint*), and perceptual-cognitive (*working toward a goal*, *making choices*) factors. This prompts us to speculate that people might also make intuitive distinctions between different aspects of agency (e.g., the experience of having intentions vs. abilities to act or not act on these intentions); a modified version of the current paradigm including a wider range of "agentic"

abilities and actions could help substantiate this suggestion. This null finding by no means rules out the possibility that lay people consider agency to be an important, distinct component of the mind. Our studies do, however, suggest that distinctions among varieties of experience are at least as prominent in people's intuitive philosophy of mind as the broad distinction between experience and agency.

We view these results as consistent with—but an important expansion upon—Gray et al.'s (2007) dimensions of mind perception. Rather than making a general distinction between agency and experience, participants in our studies focused on the extent to which mental capacities and phenomenal states are embodied, socially valenced, or perceptual in nature. These latter two kinds of experience—social-emotional experiences and perceptual-cognitive abilities—are closely aligned with Weisman et al.'s (2015) model of the lay concept of sentience, which distinguishes between affect and perception. In some sense, the current results might be seen as a combination of Weisman et al.'s theory with the classic animate–inanimate distinction, which from early in development encompasses physiological experiences of hunger and pain (Carey, 1985). Indeed, the size of the physiological factor, which accounted for 46-63% of the total variance across our studies, indicates that reasoning about biological animacy might have played an especially large role in people's judgments in this task.

Differentiating among capacities for physiological, social-emotional, and perceptual-cognitive experience could have important ramifications in social reasoning, particularly in the identification of moral patients, beings that should be protected from harm and suffering. Building on Gray et al.'s (2012) argument that "mind perception is the essence of morality," we speculate that different varieties of experience might play *different* roles in social and moral reasoning. For example, our ongoing work examines whether attributions of social-emotional experiences might be more strongly predictive of judgments of moral patiency than attributions of perceptual-cognitive abilities, at least among US adults.

Explorations of folk philosophy of mind have acquired new urgency in recent years, as people have begun interacting more frequently with increasingly sophisticated "intelligent" and "social" technologies. Interactions with robots and other social technologies are likely to be guided by intuitive understandings of the mind; in turn, these encounters might reshape lay intuitions about how minds work, and what qualifies an entity to be considered an object of sociomoral concern. The current studies provide a snapshot into how US adults are currently thinking about the "minds" of robots: The robot was judged to have virtually no capacity for physiological experiences (confirming that participants considered it inanimate), but it received notably higher ratings for many perceptual-cognitive abilities, and even for some capacities in the social-emotional domain (contra K. Gray & Wegner, 2012). In fact, in several cases, judgments of a robot's capacities for social-emotional and perceptual-cognitive abilities *exceeded* judgments of the capacities of other "edge cases" (e.g., a beetle; see Figure 1). In the agency–experience framework, this result would have been obscured by the stark discrepancy between these entities' relative capacities for physiological experiences. However, the attribution of even low-level perceptual-cognitive and social-emotional abilities to technological devices could have profound implications for how people reason about artificial intelligences as they become more enmeshed in our everyday lives—particularly if these capacities are conceptually linked to morally relevant abilities in their respective ontological categories.

There is a growing body of evidence that lay people share the ancient philosophical intuition that "the mind" is composed of distinct parts. In particular, the current studies shed light on an intuitive ontological distinction between physiological sensations, social-emotional feelings, and perceptual-cognitive abilities—three varieties of experience that may play different roles in guiding people's sense of who or what "counts" as an object of moral concern.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-114747, and by a William R. & Sara Hart Kimball Stanford Graduate Fellowship.

References

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299.

Gelman, R., & Spelke, E. S. (1981). The development of thoughts about animate and inanimate objects: Implications for research on social cognition. In J. H. Flavell & L. Ross (Eds.), *Social cognitive development: Frontiers and possible futures* (pp. 43–66). Cambridge, England: Cambridge University Press.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: mind perception and the uncanny valley. *Cognition*, *125*(1), 125–30.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Knobe, J., & Prinz, J. J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, *7*(1), 67–83.

Weisman, K., Markman, E. M., & Dweck, C. S. (2015). Reasoning about sentience and animacy: Children's and adults' inferences about the properties of unseen entities. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2625–2630). Pasadena, CA.

Benefiting from Being Alike: Interindividual Skill Differences Predict Collective Benefit in Joint Object Control

Basil Wahn (bwahn@uos.de)
Institute of Cognitive Science – Neurobiopsychology, University of Osnabrück, Albrechtstr. 28
49069 Osnabrück, Germany

Laura Schmitz (Schmitz_Laura@phd.ceu.edu)
Department of Cognitive Science, Central European University, Október 6 utca 7
1051 Budapest, Hungary

Peter König* (pkoenig@uos.de)
Institute of Cognitive Science – Neurobiopsychology, University of Osnabrück, Albrechtstr. 28
49069 Osnabrück, Germany
Institut für Neurophysiologie und Pathophysiologie, Universitätsklinikum Hamburg-Eppendorf
Hamburg, Germany

Günther Knoblich* (KnoblichG@ceu.edu)
Department of Cognitive Science, Central European University, Október 6 utca 7
1051 Budapest, Hungary

* shared senior authorship

Abstract

When two individuals perform a task together, they combine their individual skills to achieve a joint goal. Previous research has shown that interindividual skill differences predict a group's collective benefit in joint perceptual decision-making. In the present study, we tested whether this relationship also holds for other task domains, using a dynamic object control task in which two participants each controlled either the vertical or horizontal movement direction of an object. Our findings demonstrate that the difference in individuals' skill levels was highly predictive of the dyad's collective benefit. Differences in individuals' subjective ratings of task difficulty reflected skill differences and thus also turned out to be a predictor of collective benefit. Generally, collective benefit was modulated by spatial task demands. Overall, the present study shows that previous findings in joint decision-making can be extended to dynamic motor tasks such as joint object control.

Keywords: collective benefit; joint action; coordination; collaboration; task distribution; social cognition.

Introduction

In our modern world, controlling man-made objects has become an essential part of human everyday life. Human-object interactions range from simple tasks such as carrying a table (Sebanz, Bekkering, & Knoblich, 2006) to complex ones such as flying an airplane (Hutchins, 1995). Humans often do not perform these tasks alone but collaboratively in teams by distributing the task among team members to facilitate performance. For instance, when carrying a table up the stairs, the stronger person may hold the lower, heavier end of the table, or when flying an airplane, pilot and co-pilot efficiently distribute responsibilities.

While collaboration has obvious benefits, coordinating individual actions (in relation to external objects) also gives rise to costs such as the need to predict and integrate another person's action into one's own action plan. Whether benefits or costs prevail depends on a variety of aspects such as the type of control task, the information exchange between individuals, as well as individual skill levels. A study by Knoblich and Jordan (2003) tested whether groups could learn an anticipatory control strategy to jointly track a moving object. Results showed that groups managed to reach the level of individual performance *only if* given extensive training and external feedback about each group member's actions. In another task, individuals outperformed groups in a virtual object lifting task (Bosga & Meulenbroek, 2007). Yet, Masumoto and Inui (2012) provide evidence that in a joint force production task, a dyad's joint action is more successful than individual performance, provided the dyad receives external action feedback. Joint forces were also produced in a different study (van der Wel, Knoblich, & Sebanz, 2011) where participants had to control a pendulum-like object. But here, group and individual performance levels closely resembled each other. When two participants jointly controlled a ball in a virtual labyrinth game, they did not exceed the individual performance level either (Rigoli et al., 2015).

These differing findings across studies may be partly attributed to the use of different types of control tasks. Despite these task differences, it is notable that a collective benefit was only reached when coordinating individuals received some form of feedback about each other's actions. Another factor that has been shown to affect whether a group will outperform an individual or vice versa is the

magnitude of the difference in individual skill levels. Previous research has demonstrated that such differences can reliably predict whether a dyad outperforms its more skilled member in a collaborative perceptual decision-making task (Bahrami et al., 2010). In particular, a dyad's collective benefit was higher the more similar the perceptual sensitivities of the two dyad members were. Importantly, the members' opportunity to verbally negotiate the joint decision was crucial for joint success.

Considering that the specific type of task as well as the flow of information between co-actors seem to substantially affect whether joint performance is beneficial, we aimed to test whether Bahrami et al.'s finding that interindividual skill differences predict collective benefit generalizes to a *non-communicative motor* task. Specifically, in contrast to discrete decision-making where two participants first make individual decisions and subsequently communicate to reach a joint decision, we used a dynamic object control task in which participants' interaction was continuous and verbal communication was prohibited.

A further objective was to test two predictions about joint performance: We predicted that joint success should depend on the specific type of spatial control required because action distribution may be more suitable for certain types of control than for others. For instance, flying an airplane in undisturbed midair demands a more coarse control than carefully landing the plane on a small airfield. We further predicted that assigning task contributions in accordance with co-actors' individual skills should facilitate joint performance.

To summarize, it is yet unclear whether and how interindividual skill differences, spatial task demands, and unequal task contributions influence collective benefit in a dynamic task in which communicative exchange is prohibited. Hence in the present study, we used a joint object control task to examine 1) whether the correlation between interindividual skill differences and collective benefit holds not only for perceptual decision-making (Bahrami et al., 2010) but also for dynamic object manipulation, 2) whether spatial task demands modulate the group's collective benefit, and 3) whether assigning task contributions in accordance with individuals' skills facilitates joint performance. We hypothesized that:

H1 Interindividual skill differences between dyad members will be negatively correlated with the dyad's collective benefit.

H2 A dyad's collective benefit will be additionally modulated by the type of spatial task demand (i.e., coarse vs. precise control).

H3 Assigning the higher task contribution to the more skillful dyad member will facilitate joint performance while assigning the lower task contribution to the more skillful member has a detrimental effect (relative to an equal assignment of task contributions).

Method

Participants

Twelve pairs of individuals (16 female, M = 24.83 years, SD = 3.05 years) participated in the study. Data collection was partly conducted at the Central European University in Budapest, Hungary (eight pairs) and partly at the University of Osnabrück in Germany (four pairs). All participants were right-handed and had normal or corrected-to-normal vision. They signed prior informed consent and received monetary compensation. The study was performed in accordance with the Declaration of Helsinki.

Apparatus and stimuli

The experimental setup consisted of two 24" Asus computer screens (resolution 1920 x 1080 pixels, 40.25° x 22.64° visual field, refresh rate 60 Hz) which were placed next to each other. A black cardboard partition (70 x 100 cm) was set up between them (Figure 1). On each screen, the following stimuli were presented (Figure 1): A circle (outlined in black, 0.9 cm = 0.71° diameter) was centrally displayed on a white background. A second, smaller circle (filled in black, 0.5 cm = 0.38° diameter), surrounded by another larger circle (outlined in grey, 4.2 cm = 3.21° diameter) was displayed in one of 16 possible positions. The central circle represented the cursor that participants controlled. The second circle was the target to which the cursor should be moved. The target's periphery was defined as homing-in zone. The target was 12.8 cm away from the start location of the cursor.

The experiment was programmed using the Python library Pygame and the experimental procedure and data collection were controlled by Python 2.7.3. The experiment was run on two Dell Precision computers.

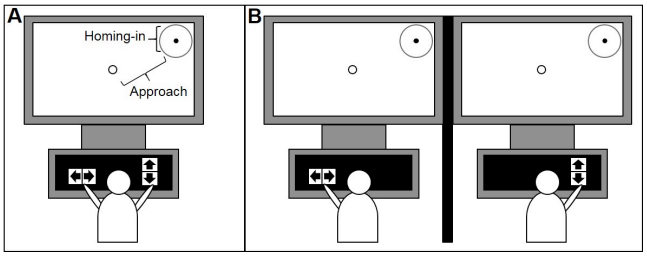


Figure 1: Via key presses, participants controlled the cursor's horizontal movement with the left and its vertical movement with the right hand. In the Individual Condition, participant controlled both dimensions (A); in the Joint Condition (B), control was distributed between participants.

Procedure

Participants were seated next to each other in front of two computer screens (at a distance of 85 cm), separated by a partition (see Figure 1B) such that they could neither see each other nor each other's screens. They were not allowed to talk and wore ear muffs throughout the experiment to shield external noise (e.g. the sound of key presses).

Participants performed the task both individually (Individual Condition) and together as members of a pair (Joint Condition). In the Individual Condition, each participant controlled the cursor displayed on their own computer screen whereas in the Joint Condition, participants jointly controlled one cursor and both screens showed identical displays. Participants first performed five practice trials in both conditions. After the practice, six experimental blocks each consisting of 40 trials followed. Individual and joint blocks alternated, i.e. participants first completed one block in the Individual or Joint Condition and then switched to the other condition in the next block. This way, three blocks in each condition were completed in a counterbalanced order across pairs.

At the start of each trial, the cursor appeared in the center of the screen. After one second, the target appeared in one of 16 possible locations (order of locations was randomized). Relative to the cursor's start location, the target was located in one of five angles (0°, 22.5°, 45°, 67.5°, 90°; see Figure 2). Each angle appeared eight times per block. Participants were instructed to move the cursor to the target as fast as possible and on the most direct path, and to stop the cursor exactly on the target. The target turned green when the cursor came to a halt on it. After the cursor had remained motionless on the target for one second, the trial was completed successfully and the next trial started. Two types of behavior were classified as errors: If the cursor moved into the homing-in zone but then exited it again, and if the cursor moved outside of the screen borders.

Participants controlled the cursor by pressing different keys on the keyboard which incremented the cursor's velocity either to the left, to the right, upwards, or downwards (compare Knoblich & Jordan, 2003). The 'left-key' (LK) and 'right-key' (RK) were controlled by the left hand and the 'up-key' and 'down-key' were controlled by the right hand (see Figure 1). If LK was pressed, the cursor started moving to the left. Another LK press accelerated the movement to the left, whereas a RK press decelerated it (by increasing its velocity into the opposite direction). Each key press caused a speed increment/decrement of 0.24°/s. Thus, if a participant had produced leftward movement by pressing LK e.g. four times, she needed to press RK also four times to decrease the cursor's velocity back to zero. If she pressed RK a fifth time, the cursor started moving to the right. In the Individual Condition, participants used both their left and right hand to control all four movement directions whereas in the Joint Condition, task control was distributed between the two members of a dyad such that the participant sitting on the left side controlled only the horizontal dimension (with her left hand) and the participant on the right side controlled only the vertical dimension (with her right hand). Participants were assigned randomly to either the left or the right side. The five different angles determined participants' relative contributions towards task completion in the Joint Condition: For 0° and 90°, only one dyad member controlled the cursor individually, as only horizontal or vertical movement was required. These angles served as baseline measure for individual performance within a joint setting ('baseline angles'). For the other three angles, both members had to contribute to achieve the task goal ('joint contribution angles'). For 45°, both members contributed equally ('50 % contribution'). For 22.5° and 67.5°, one of the members contributed more than the other ('30 % vs. 70 % contribution'); see Figure 2.

Task demands not only differed in terms of the relative contributions across trials but also regarding the two distinct movement phases within a trial, i.e., the Approach and the Homing-in phase (see Figure 1). The Approach was the interval from trial start to the moment the cursor entered the target periphery, where the Homing-in began. The Approach primarily required high movement speed whereas controlled braking and high spatial accuracy was required during the final Homing-in. As we hypothesized that the two movement phases ('Phase') as well as the three contribution angles ('Contribution') could differentially influence joint performance (cf. H2 & H3, respectively), they were included as factors in the analysis.

After task completion, participants were asked to rate (in written form) subjective task difficulty of joint and individual performance on a 7 point scale (1 = very easy, 7 = very difficult). The experiment lasted about 80 minutes.

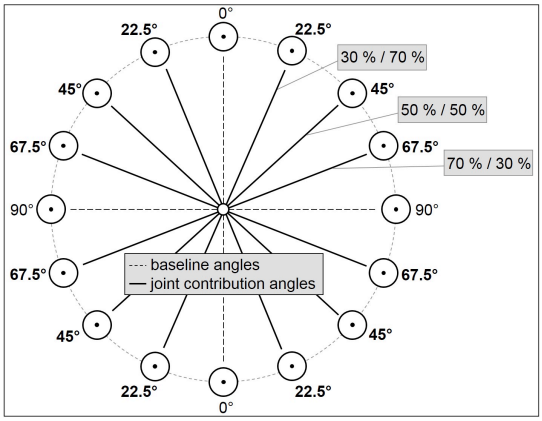


Figure 2: Contribution angles: For 0° only vertical and for 90° only horizontal control by one participant was required. For 45°, both participants contributed equally, and for 22.5° and 67.5°, one participant contributed 40 % more than the other (i.e., contributions were split 30 % / 70 %).

Data preparation and analysis

Prior to analysis, all error trials were removed from the data set (6.7 %). Statistical analyses were done in R (2014).

As performance measures, we derived reaction time (RT) and distance travelled (DIST). RT was the time between cursor appearance and arrival at target. DIST was the absolute path length the cursor traveled from start to target location. Both RT and DIST were also calculated separately for each movement phase. The values for Approach were the time/distance from cursor appearance/start location to crossing the target periphery. The remaining time/distance until arrival at target were the Homing-in values. RT and

DIST served as central measures since participants' goal was to move to the target as fast and as directly as possible.

To evaluate whether performing as a dyad was more efficient than performing alone, we compared each dyad's performance to the performance of the better member of that dyad. Following Bahrami's terminology, a *collective benefit* is achieved if the dyad outperforms the better individual. As lower values of the dependent variables indicate better performance, we calculated the collective benefit by dividing the better member's measurement by the dyad's value. For instance for RT, the faster member's RT (RT min) was divided by the dyad's RT (RT dyad). Values above 1 indicate a collective benefit because dyad RT is lower than the faster member's RT. This measure was calculated separately for the Approach and the Homing-in phase – averaged across all contribution angles as well as separately for each of the three contribution angles (30 %, 50 %, 70 %; defined for the better member's contribution, see Figure 3). We predicted (cf. H3) that the 70 % contribution should result in a higher collective benefit because the better member's contribution dominates joint performance. Conversely, a lower collective benefit should be achieved for the 30% contribution because the weaker member's contribution dominates joint performance.

Besides testing whether joint performance would result into an overall collective benefit, we investigated in what way two individuals' performances determine the dyad's joint performance. Based on previous findings (Bahrami et al., 2010), we predicted that the quality of a dyad's performance is determined by the magnitude of the difference between the two individual performances, such that the smaller the individuals' performance difference, the larger their collective benefit during joint performance (cf. H1). To test whether the difference between two individuals' performances is indeed a predictor of their collective benefit, we divided the better individual's performance value by the worse individual's performance value. For instance, the RT ratio was composed of the faster member's RT (RT min) divided by the slower member's RT (RT max). For the resulting skill ratio, a value of 1 indicates that both members perform equally well individually whereas an increasingly lower ratio indicates a larger interindividual performance difference. In our subsequent analysis, we correlated these interindividual skill ratios with the collective benefit for each of the contribution conditions, separately for Approach and Homing-in. We predicted that the collective benefit systematically increases with larger skill ratios, suggesting that dyad members who have increasingly similar individual performance levels benefit more from performing jointly – particularly when both members contribute equally to the joint performance. If however the better individual's dominance compensates for a large interindividual difference (as in the 70% contribution condition), we predicted a reduced correlation relative to the other contribution conditions (cf. H3).

Results

Social facilitation

To exclude the theoretical possibility that differences between individual and joint performance are due to social facilitation (individual performance being affected purely by someone else's nearby presence), we compared RT in the Individual and Joint Condition for the baseline angles in which only one participant controlled the cursor. This comparison was done separately for the Approach and Homing-in phase. We found no significant RT difference for these comparisons (Approach: $t(11) = -0.03$, $p = .979$, Cohen's $d = -0.04$; Homing-in: $t(11) = -0.14$, $p = .894$, Cohen's $d = -0.06$), indicating that participants' performance was not affected by the joint setting. This finding was confirmed by the results for DIST (Approach: $t(11) = -0.92$, $p = .376$, Cohen's $d = -0.27$; Homing-in: $t(11) = -0.28$, $p = .782$, Cohen's $d = -0.08$). For all further analyses, we focused on the joint contribution angles.

Collective benefit

To determine whether joint control was beneficial, we tested whether the collective RT benefit was larger than 1, using one sample t-tests. Results showed that for Approach, dyads did not outperform the better member ($t(11) = -0.88$, $p = .396$, Cohen's $d = -0.26$) whereas a collective benefit was achieved for Homing-in ($t(11) = 2.96$, $p = .013$, Cohen's $d = 0.86$) (Figure 3A). In line with our hypothesis H2, this result suggests that distributing task dimensions is advantageous for precise spatial control over short distances (as during Homing-in) but not for longer distances when speed is prioritized over spatial accuracy (as during Approach).

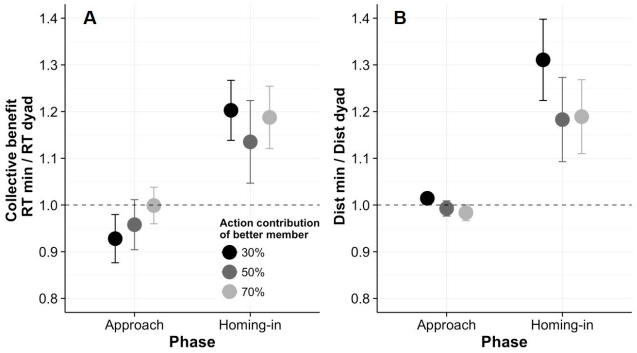


Figure 3: Mean collective benefit for A: RT and B: DIST. Dyads obtained a collective benefit > 1 (see dashed line) for Homing-in only. Error bars reflect Standard Errors.

To test our hypothesis H3 that collective benefit might differ between the three contribution angles, we performed a 2 x 3 repeated measures ANOVA (Phase x Contribution). We found a significant main effect of Phase ($F(1,11) = 20.57$, $p < .001$), indicating that the collective benefit was higher for Homing-in than Approach, as shown by the previous analysis. Neither a significant main effect of Contribution ($F(2,22) = 1.00$, $p = .384$) nor a significant interaction effect between the two factors ($F(2,22) = 1.31$, p

= .288) was found, demonstrating that collective benefit did not differ between contribution angles.

These findings were corroborated by the analysis of our second performance measure DIST (Figure 3B). As for RT, a collective benefit was present for Homing-in ($t(11) = 3.20$, $p = .008$, Cohen's $d = 0.92$) but not for Approach ($t(11) = -0.24$, $p = .822$, Cohen's $d = -0.07$) and the ANOVA results showed a significant main effect of Phase ($F(1,11) = 12.18$, $p = .005$), but neither a significant main effect of Contribution ($F(2,22) = 2.09$, $p = .147$) nor a significant interaction effect ($F(2,22) = 0.99$, $p = .389$).

Interindividual differences as predictor

Regarding the predicted correlation between interindividual skill ratios and collective benefit measures, results for RT showed that the skill ratio significantly predicts collective benefit for the 30 % and the 50 % contribution angles for Approach (30 % contribution: $r = .90$, $t(10) = 6.46$, $p < .001$; 50 % contribution: $r = .91$, $t(10) = 7.02$, $p < .001$) (Figure 4A). As hypothesized, there was no significant correlation for the 70 % contribution ($r = .48$, $t(10) = 1.75$, $p = .111$). Pairwise comparisons between the correlation value of the 70 % contribution and the values of the two other contribution angles demonstrated a significant difference for both comparisons (70 % vs. 30 %: $z = 2.87$, $p = .004$; 70 % vs. 50 %: $z = 3.01$, $p = .003$) (cf. Steiger, 1980).

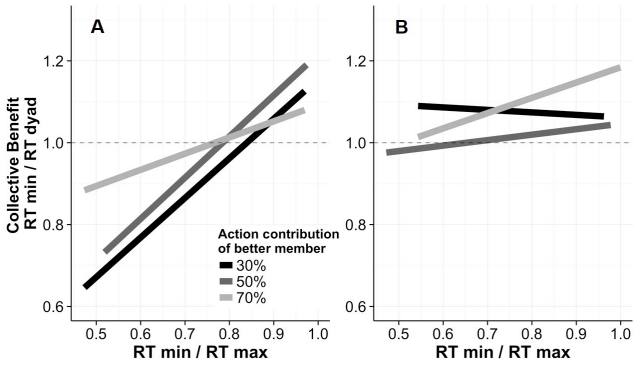


Figure 4: The dyad member's RT ratio predicts the dyad's collective benefit for A: Approach but not B: Homing-in.

In sum, these results indicate that the interindividual RT ratio highly predicts a dyad's collective benefit (cf. H1). Moreover, the predictive power of the RT ratios is significantly reduced when large interindividual differences are compensated for by the better individual's dominant contribution in comparison to the other two contribution angles, suggesting that the collective benefit depends on the individuals' relative action contribution (cf. H3).

For Homing-in (Figure 4B), analyses yielded a significant correlation for the 70 % contribution angle ($r = .58$, $t(10) = 2.27$, $p = .047$) but not for the other two angles (30 % contribution: $r = .38$, $t(10) = 1.32$, $p = .217$; 50 % contribution: $r = .41$, $t(10) = 1.43$, $p = .183$). However, pairwise comparisons showed no significant differences between contribution angles (70 % vs. 30 %: $z = 0.74$, $p = .459$; 70 % vs. 50 %: $z = 0.51$, $p = .614$). Overall, these findings indicate that the interindividual RT ratios predict the collective benefit for Approach but not for Homing-in.

In contrast to the RT results, we did not find significant correlations between interindividual DIST ratios and collective benefit for Approach (30 % contribution: $r = -.23$, $t(10) = -0.73$, $p = .48$; 50 % contribution: $r = -.28$, $t(10) = -0.91$, $p = .39$; 70 % contribution: $r = .09$, $t(10) = 0.29$, $p = .778$). For Homing-in however, interindividual DIST ratios predicted the collective benefit significantly for the 70 % ($r = .62$, $t(10) = 2.51$, $p = .030$) and close-to-significantly for the 30 % contribution angle ($r = .57$, $t(10) = 2.18$, $p = .054$), but not for the 50 % contribution angle ($r = .31$, $t(10) = 1.04$, $p = .183$). The lack of significant correlations for Approach can be explained by participants' overall low variability for DIST in Approach (Approach SD = 0.04 vs. Homing-in SD = 0.15, see Figure 3B).

Subjective difficulty as predictor

Finally, we explored whether differences in individuals' subjective experience might be related to the quality of their joint performance. We hypothesized that participants' introspective access may allow for accurate ratings of perceived task difficulty, thereby effectively providing an indirect measure of one's own performance.

We used participants' ratings of individual task difficulty and subtracted the rating of the faster dyad member from the rating of the slower member. These difference values were then correlated with the collective RT benefit measures (for each of the contribution conditions, separately for Approach and Homing-in). Results showed that for Approach, the difference in difficulty ratings indeed significantly predicted the collective benefit for all contribution angles (30 %: $r = -.73$, $t(10) = -3.33$, $p = .008$; 50 %: $r = -.76$, $t(10) = -3.67$, $p = .004$; 70 %: $r = -.58$, $t(10) = -2.23$, $p = .050$). For Homing-in, we found significant correlations for the 30 % ($r = -.76$, $t(10) = -3.74$, $p = .003$) and 70 % ($r = -.59$, $t(10) = -2.33$, $p = .042$) contribution angles, as well as a trend towards significance for the 50 % contribution angle ($r = -.55$, $t(10) = -2.09$, $p = .063$). These results indicate that the differences in participants' subjective difficulty ratings can be used to predict the collective temporal performance benefit.

The same tests were conducted with the collective DIST benefit measure. Based on the finding that only Homing-in performance is predicted by interindividual DIST ratios, we computed the correlations with the rating differences only for Homing-in. Yet no significant correlations were found (30 %: $r = -0.46$, $t(10) = -1.66$, $p = .128$; 50 %: $r = -0.33$, $t(10) = -1.10$, $p = .296$; 70 %: $r = -0.43$, $t(10) = -1.49$, $p = .167$), suggesting that subjective difficulty ratings cannot be used to predict the collective spatial performance benefit.

Discussion

The aim of the present study was to target a gap in the existing literature by examining group performance in a dynamic object control task. We investigated whether the relationship between interindividual skill differences and collective benefit previously found for perceptual decision-

making (Bahrami et al., 2010) also holds in a dynamic motor task (cf. H1). Whereas Bahrami et al. used a discrete perceptual decision-making task in which two participants first took individual decisions and subsequently communicated to reach a joint decision, we used a dynamic task in which participants' interaction was continuous. Two members of a dyad jointly controlled a cursor movement with the shared goal of moving it to a target location, while one member controlled the vertical and the other member the horizontal movement dimension. In contrast to Bahrami et al. (2010), group members were not allowed to verbally communicate. Additionally, the present task tested whether spatial control demands (coarse vs. precise control) affect collective benefit (cf. H2). Finally, we manipulated the dyad members' relative action contributions to test whether assigning the higher task contribution to the more skillful dyad member facilitates joint performance (cf. H3).

Despite using a dynamic instead of a discrete task and not allowing verbal exchange between participants, the present data are in line with Bahrami and colleagues' (2010) finding that a dyad's collective benefit varies according to the interindividual skill differences of its members (cf. H1). Thus our results generalize our current knowledge regarding individuals' skills as predictors of collective benefit to the domain of dynamic object control, demonstrating a systematic relationship between interindividual skill ratios and the quality of joint performance. Adding to Knoblich and Jordan (2003) who highlighted the influence of external action feedback on collective benefit in object tracking, we identify interindividual skill differences as a further predictor for collective benefit in a similar motor task.

When comparing the dyad's performance to the performance of the more skillful dyad member, we found that dyads outperformed the individual in the Homing-in phase of the task while there was no such collective benefit in the Approach phase. This suggests that collective benefit is modulated by spatial task demands such that joint control is advantageous only for precise but not for coarse spatial control (cf. H2). The two phases may have also differed in overall task difficulty – future studies are needed to tease apart the effects of spatial control type and of overall control difficulty. With regard to spatially coarse control, we showed that the influence of individual skill differences on collective benefit can be effectively reduced by assigning contribution percentages in accordance with the individual skill level (cf. H3).

Finally, the present study shows that not only individuals' performance differences but also differences in individuals' subjective experience predict the quality of their joint performance. Relatedly, Bahrami and colleagues' (2010) found that collective benefit was highly dependent on whether participants were allowed to share their confidence ratings when agreeing on a joint decision. Both confidence ratings as well as subjective difficulty ratings require assessing one's own performance level. In Bahrami et al.'s (2010) task though, participants *shared* their confidence before taking a joint decision. In the present study, subjective task difficulty was *privately* assessed after task completion. When assessing task difficulty, participants could rely on introspective access and/or regard their own skill in relation to their task partner's. In any case, participants' ratings provided an indirect performance measure such that interindividual rating differences predict the dyad's collective benefit. Possibly, sharing these ratings online might have further improved a dyad's performance.

Taken together, the present study shows that the collective benefit in a joint object control task depends on interindividual skill differences, as well as on the type of spatial task demand. Future research could investigate whether interindividual skill differences are viable predictors for other types of joint action.

Acknowledgments

This research was supported by H2020—H2020-FETPROACT-2014641321—socSMCs (for BW & PK) and by the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n° [609819], SOMICS, and by ERC grant agreement n°616072, JAXPERTISE (for LS & GK).

References

Bahrami, B., Olsen, K., Latham. P. E., Roepstorff, A., Rees, G., Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*, 1081-1085.

Bosga, J., & Meulenbroek, R. G. J. (2007). Joint-action coordination of redundant force contributions in a virtual lifting task. *Motor Control*, *11*(3), 234-257.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*, 265-288.

Knoblich, G. & Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 1006-1016.

Masumoto, J., & Inui, N. (2013). Two heads are better than one: both complementary and synchronous strategies facilitate joint action. *Journal of Neurophysiology*, *109*, 1307-1314.

Rigoli, L., Romero, V., Shockley, K., Funke, G. J., Strang, A. J., & Richardson, M. J. (2015). Effects of complementary control on the coordination dynamics of joint-action. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 1997-2002. Austin, TX: Cognitive Science Society.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences, 10*, 70-76.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

van der Wel, R. P., Knoblich, G., & Sebanz, N. (2011). Let the force be with us: Dyads exploit haptic coupling for coordination. *Journal of Experimental Psychology: Human Perception and Performance, 37*(5), 1420-1431.

Reading experience shapes the mental timeline but not the mental number line

Benjamin Pitt[1] (bpitt@uchicago.edu)
Daniel Casasanto[1,2] (casasanto@uchicago.edu)

[1]Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL 60637 USA
[2]Grossman Institute for Neuroscience, University of Chicago, 5812 South Ellis Avenue, Chicago, IL 60637 USA

Abstract

People conceptualize both time and numbers as unfolding along a horizontal line, either from left to right or from right to left. The direction of both the mental timeline (MTL) and the mental number line (MNL) are widely assumed to depend on the direction of reading and writing within a culture. Although experimental evidence supports this assumption regarding the MTL, there is no clear evidence that reading direction determines the direction of the MNL. Here we tested effects of reading experience on the direction of both the MTL and MNL. Participants read English text either normally (from left to right) or mirror-reversed (from right to left). After normal reading, participants showed the space-time associations and space-number associations typical of Westerners. After mirror reading, participants' space-time associations were significantly reduced but their space-number associations were unchanged. These results suggest that the MTL and MNL have different experiential bases. Whereas the MTL can be shaped by reading experience, the MNL is shaped by other culture-specific practices through which people experience numbers arrayed in left-right space.

Keywords: SNARC; Mental number line; Mental timeline; Space; Time; Reading direction; Numerical cognition

Introduction

Across many cultures, people use left-right space to think about both time and number. In English-speaking cultures, people associate earlier events with the left side of space and later events with the right, forming an implicit mental timeline (MTL) that progresses from left to right. Likewise, English speakers associate smaller numbers with the left and larger numbers with the right, forming an implicit mental number line (MNL) that increases from left to right. These spatial mappings of time and number have been demonstrated in hundreds of experiments, most often using reaction time (RT) tasks: People tend to respond faster to earlier events and smaller numbers using their left hand and to later events and larger numbers using their right hand (Dehaene, Bossini, & Giraux, 1993; Wood, Willems, Nuerk & Fischer, 2008; Bonato, Zorzi & Umiltà, 2012). These spatial mappings of time and number are also evident in people's spontaneous gestures (Casasanto & Jasmin, 2012) and eye movements (Fischer, Castel, Dodd & Pratt, 2003; Loestscher, Bockisch & Brugger, 2008) across lateral space.

What determines the directions of the MTL and MNL? On the basis of cross-cultural variation, many scholars have assumed that the directions of both the MTL and MNL are determined by the direction in which people read and write text. Yet, upon examination, the evidence for this assumption is much stronger for the MTL than the MNL.

People from Western cultures show MTLs that progress from left to right (Spaniards: Santiago, Lupáñez, Pérez, & Funes, 2007; Canadians: Weger & Pratt, 2008), whereas people from cultures where text is written from right to left show the corresponding reversal in the MTL (i.e. earlier events on the right, later events on the left; Arabic: Tversky, Kugelmass & Winter, 1991; Hebrew: Fuhrman & Boroditsky, 2010; Ouellet, Santiago, Israeli & Gabay, 2010; cf., Tversky et al., 1991). Despite this clear correlation, it is not known to what extent the direction of reading and writing is a cause or an effect of cross-cultural variation in implicit space–time mappings, in part because cultural practices tend to covary. Groups who write from left to right also tend to spatialize time on calendars and graphs from left to right (Tversky et al., 1991), and to gesture according to a left-to-right mental timeline (Casasanto & Jasmin, 2012; Cooperrider & Nuñez, 2009). This covariation leaves open many possible scenarios according to which orthography could play a primary causal role, a mediating role, or no causal role at all in determining the direction of the MTL (see Casasanto & Bottini, 2014).

Testing whether reading experience can play a causal role in determining the direction of the MTL requires experimental intervention. Casasanto and Bottini (2014) randomly assigned Dutch speakers to read text in either normal orthography (from left to right) or mirror-reversed orthography (from right to left) while classifying events as either earlier or later in time. Participants who read normally were faster to classify earlier events with their left hand and later events with their right hand, reflecting the left-to-right MTL typical of Westerners. By contrast, those who read mirror-reversed text showed the opposite pattern of RTs, showing a right-to-left MTL like that of Arabic speakers. Together, the correlational and experimental data provide strong support for the claim that reading experience can determine the direction of the MTL.

In the case of the MNL, however, the evidence is much less clear. In general, the direction of the MNL covaries with the direction of written text in a culture: Westerners tend to show MNLs that increase from left to right (e.g. French: Dehaene et al., 1993; Scots: Fischer, 2008; Canadians: Shaki, Fischer, & Petrusic, 2009), whereas people from some Arabic cultures show MNLs in the opposite direction (i.e. smaller numbers on the right, larger

numbers on the left), consistent with the right-to-left reading direction in their cultures (Palestinians: Shaki et al., 2009; Lebanese: Zebian, 2005). This cross-cultural variation has lead to a general consensus that, like the MTL, the MNL's direction is determined by the direction of orthography.

Yet, overall, the claim that reading or writing experience determines the direction of the MNL is neither well-supported by empirical evidence nor clearly motivated. First, the direction of people's MNLs appears to be only loosely correlated with the direction in which they read and write text. In their seminal study establishing the Spatial-Numerical Association of Response Codes (SNARC) effect, Dehaene et al. (1993) found that French participants responded faster to small numbers with the left hand and large numbers with the right. However, this same study found "no evidence" of a reversed SNARC effect in Iranians who had extensive exposure to right-to-left orthography. Another study found a reversed SNARC effect in Arabic-speaking Palestinians but no SNARC effect in Hebrew-speaking Israelis, who also read text from right to left (Shaki et al., 2009). One other study has demonstrated a reversed SNARC effect among Arabic-speakers (Zebian, 2005), but contrary to predictions, Arabic-English bilinguals showed reversed SNARC effects that were numerically *stronger* than those of Arabic monolinguals (and English monolinguals showed no significant SNARC effect). Although these three studies are often cited as evidence for the proposal that reading direction shapes the MNL, none of them clearly support this proposal (see also Fischer, Shaki, & Cruise, 2009).

The only direct experimental test of the effect of reading experience on the direction of the MNL produced a null result[1]. French participants responded to number words presented in either standard or mirror-reversed orthography. Participants showed normal SNARC effects in both conditions; Orthography had no effect on the strength or direction of the SNARC (Dehaene et al., 1993: Experiment 8). In spite of this result, the researchers concluded that "[t]he particular direction of the spatial-numerical association seems to be determined by the direction of writing," (Dehaene et al., 1993, pg. 394) – a conclusion that has been widely accepted for more than two decades. Yet, there is little evidence to support this conclusion, and some clear reasons to doubt it, on the basis of both correlational and experimental data.

Why might reading experience determine the direction of the MTL but not the MNL? The answer may be found in the experience of reading. When reading English text, people's eyes start on the left side of the page at an earlier time and end on the right side at a later time. In this experience, progress through time correlates with progress (rightward) through space. This experiental correlation between space and time in orthography is sufficient to determine the direction of the MTL (Casasanto & Bottini, 2014).

There is no analogous correlation between space and number in the act of reading ordinary text (as opposed to reading numbers, per se). Moving rightward across the page corresponds to moving *later in time*, but it does not correspond to moving *greater in number* (unless people count words as they read, which is unlikely). Reading or writing text creates an experiential link between space and time, but not between space and number. As such, the proposal that reading experience plays a functional role in determining the direction of the MTL is well motivated, but the proposal that reading experience plays a functional role in determining the direction of the MNL is not.

Here we tested the effects of reading experience on the direction of the MTL and MNL, by randomly assigning US participants to read either normal or mirror-reversed English text. After reading training, we assessed the strength and direction of participants' MTLs and MNLs as indexed by their RTs on matched space-time and space-number congruity tasks. We reasoned that if reading direction can play a causal role in determining the direction of both the MTL and the MNL, then participants should show normal space-time and space-number congruity effects after reading normal text, and reduced (or reversed) effects after reading mirror-reversed text, for both time and number. Alternatively, if the directions of the MTL and MNL are determined by different kinds of experience, then mirror-reversed reading should reduce (or reverse) the space-time congruity effect but not the space-number congruity effect.

Method

Participants

Sixty-four right-handed native English speakers from the University of Chicago community participated for payment or course credit. Half were randomly assigned to the standard reading condition (n=32) and the other half to the reversed reading condition (n=32).

Materials and Procedure

Participants performed a two-part experiment in which a training phase was followed by a test phase.

Training Phase. In the training phase, participants read a passage silently in either standard or mirror-reversed orthography. They were seated in front of a 24-inch Apple iMac computer (with the keyboard and mouse removed) and were told that they would be asked some comprehension questions after reading. Text appeared in black on a white background and spanned the width of the screen. The text, which was excerpted from *Zen and the Art of Motorcycle Maintenance* (Persig, 1974), consisted of 2,964 words and

[1] In a study by Shaki & Fischer (2008), Hebrew-Russian bilinguals showed weaker SNARC effects after reading in Hebrew than after reading in Russian. However, in these experiments, reading direction was confounded with language. Therefore, any difference between conditions may be due to other cultural factors that differ across these language groups.

spanned 25 pages. After reading each page, participants pressed the central key on a button box to advance to the next page. On average, reading training lasted about 12 minutes in the standard condition and 36 minutes in the reversed condition and was limited to 45 minutes by the experimenter. After reading, participants responded to five comprehension questions by selecting one of two answers.

Test Phase. The test phase immediately followed the training phase and consisted of three tasks in which participants were instructed to respond "as quickly and accurately as possible." In the months task, 3-letter abbreviations for the months of the year (February through October except June) appeared on the screen one at a time. Participants classified each month as either "earlier" or "later" than June in the calendar year by pressing one of two response keys. In one block of trials, participants used the left-hand key for months that were earlier and the right-hand key for months that were later. This response mapping was reversed in the other block of trials and block order was counterbalanced across participants. In two number tasks (digits and number words), participants classified numbers (1 - 9 except 5) as either "greater" or "less" than five. For one block, they used the left-hand key for small numbers and the right-hand key for large numbers. In the other block, this response-mapping was reversed. In the digits task, numbers were presented as Arabic numerals; in the number words task, they were presented as English number words. In training and test, all instructions and stimuli appeared in capital letters and were presented in normal orthography in the standard reading condition and in mirror-reversed orthography in the reversed reading condition.

In each block, the eight unique stimuli appeared in random order eight times, composing 128 trials per task. At the beginning of each block, the experimenter asked the participant to raise the hand corresponding to each of the responses to ensure clarity of the response mapping. Each trial began with 500 ms of a black screen followed by a fixation cross whose duration varied uniformly between 500 and 1000 ms. Throughout testing, all numbers and words were presented in white on a black background at the center of the screen. The order of months and number tasks was counterbalanced across participants such that the months task was first for half of participants and the number tasks were first for the other half of participants. Within the number tasks, the order of the digits and number words tasks was counterbalanced across participants.

After testing, participants were debriefed to determine whether they were aware of the experimental hypotheses, and they then completed a language history questionnaire and the Edinburgh Handedness Inventory (Oldfield, 1971).

Results

Three subjects who failed to follow instructions and one who guessed the purpose of the training were replaced. The error rate was significantly higher in the reversed reading condition (4.16%) than in the standard reading condition (3.20%; $\chi2(1, N=64)=16.03$, $p=.0006$) and significantly higher in the months task (4.11%) than in the number words task (3.28%; $\chi2(1, N=64)=7.94$, $p=.005$; all other p's > .10). Although statistically significant, these differences in error rates were very small (less than 1%). We excluded inaccurate trials (3.68%) and accurate trials with RTs less than 200ms or greater than 2500ms (2.53%).

To evaluate space-time congruity effects, months were coded for ordinal position in the calendar year. We conducted an analysis of variance (ANOVA) on RTs with response hand and ordinal position of months as predictors. We used the same analysis to evaluate space-number congruity effects: RTs were entered into an ANOVA with response hand and ordinal position of numbers as predictors (a measure of the SNARC effect; see Dehaene et al., 1993; Gevers et al., 2010; van Dijck & Fias, 2011)[2]. In all tests, RTs were reciprocal-transformed to approximate a normal distribution of residuals and subjects were included as a random effect.

Space-Time Associations

RTs greater than 2.5 standard deviations from subject means were removed (4.47%). In the standard reading condition, the interaction between response hand and position was highly significant ($F(1, 3622)=46.15$, $p<.0001$), indicating a reliable standard space-time congruity effect in which earlier months were associated with the left and later months were associated with the right (mean slope=-18.61ms/position). A significant space-time congruity effect was also found in the reversed reading condition ($F(1, 3570)=6.98$, $p=.008$; mean slope=-8.34ms/position); of primary interest, this effect was significantly weaker than in the standard reading condition ($F(1, 7192)=8.09$, $p=.004$; Fig. 1, left). Reading direction reliably changed the MTL, in the predicted direction.

Space-Number Associations

Digits Task RTs greater than 2.5 standard deviations from subject means were removed (4.95%). In the standard reading condition, the interaction between response hand and position was highly significant ($F(1, 3681)=44.01$, $p<.0001$), indicating a reliable standard SNARC effect in which small numbers were associated with the left and large numbers were associated with the right (mean slope=-7.68ms/digit). A nearly identical SNARC effect was found

[2]For comparison with other findings, we report and plot the SNARC effect in each task as a regression slope, following Fias, Brysbaert, Geypens, & Géry (1996), regressing dRT values (dRT=right-hand – left-hand RT) for each number or month over ordinal position. Although these slopes can also be used for inferential statistics, using them here would be inappropriate for several reasons. For example, statistical tests across these data points cannot include random effects of subjects, which can increase Type I error rate. Furthermore, because Fias et al.'s method collapses over large amounts of data (here, a 128:1 compression) it is unfit for testing the higher-order (3-way and 4-way) interactions on which our experimental questions depend.

in the reversed reading condition (F(1, 3604)=42.73, p<.0001; mean slope=-7.94ms/digit). The difference in the SNARC effects across conditions did not approach significance (F(1, 7285)=.05, p=.82).

Number Words Task RTs greater than 2.5 standard deviations from subject means were removed (5.02%). In the standard reading condition, the interaction between response hand and position was significant (F(1, 3650)=11.20, p=.0008), indicating a reliable standard SNARC effect (mean slope=-5.77ms/digit). A significant standard SNARC effect was also found in the reversed reading condition (F(1, 3625)=9.69, p=.002; mean slope=-2.75ms/digit). The difference in the SNARC effects across conditions did not approach significance (F(1, 7275)=.01, p=.92).

Comparison of number tasks To compare the effect of reading condition between the digit task and the number words task, we conducted an ANOVA on reciprocal-transformed RTs with position, response hand, reading condition, and task as predictors. The effect of reading condition on the SNARC effect did not differ between the two number tasks (F(1, 14622)=.02, p=.89). We therefore combined the RT data from the digits task and the number words task, doubling our item-wise power to detect an effect of reading direction on the MNL.

Number Tasks Combined In the standard reading condition, the interaction between response hand and position was significant (F(1, 7366)=47.69, p<.0001), indicating a reliable standard SNARC effect (mean slope=-6.69ms/digit). A reliable standard SNARC effect was also found in the reversed reading condition (F(1, 7264)=35.16, p<.0001; mean slope=-5.10ms/digit). Of primary interest, the SNARC effects did not differ across reading conditions (F(1, 14630)=.12, p=.73; Fig. 1, right). Reading direction had no effect on the direction of the MNL.

Comparison of time and number tasks To compare the effect of reading direction on the space-time and space-number congruity effects, we used a linear mixed-effects model in R. Reciprocal-transformed RTs were predicted by response hand, ordinal position, reading condition, and task, with random slopes and intercepts for subjects. The effect of reading condition on the space-time congruity effect was reliably stronger than the effect of reading condition on the space-number congruity effect ($\chi^2(1)=7.99$, p=.005). The significant effect of reading experience on the MTL was greater than its non-significant effect on the MNL.

General Discussion

The directions of the mental timeline and mental number line both vary across cultures, and both are often attributed to culture-specific habits of reading and writing. Here we tested the effect of reading direction on the direction of the MTL and MNL in the same group of participants. After reading normal English text, participants showed the space-time and space-number associations typical of Westerners. After reading mirror-reversed text (from right to left), participants' space-time associations were significantly weakened but their space-number associations were unchanged. These results provide evidence that reading direction can influence the direction of the MTL, but challenge the claim that reading direction shapes the MNL.

These findings address two shortcomings of the only other experimental test of the effect of reading direction on the MNL. Dehaene and colleagues (1993; Experiment 8) found no effect of reading direction on the SNARC effect. In principle, this null effect could result from an insufficient experimental manipulation. First, there was no training phase in Dehaene et al.'s experiment. Second, there was no manipulation check. Therefore, there is no evidence that the amount of exposure to mirror-reveresd text that participants received was sufficient to influence spatial mappings in their minds. In the current study, (a) we included a training phase to greatly increase participants' exposure to mirror-reversed text, and (b) we included a manipulation check: although reading training had no effect on the participants' MNL it had a highly significant effect on their MTL. As

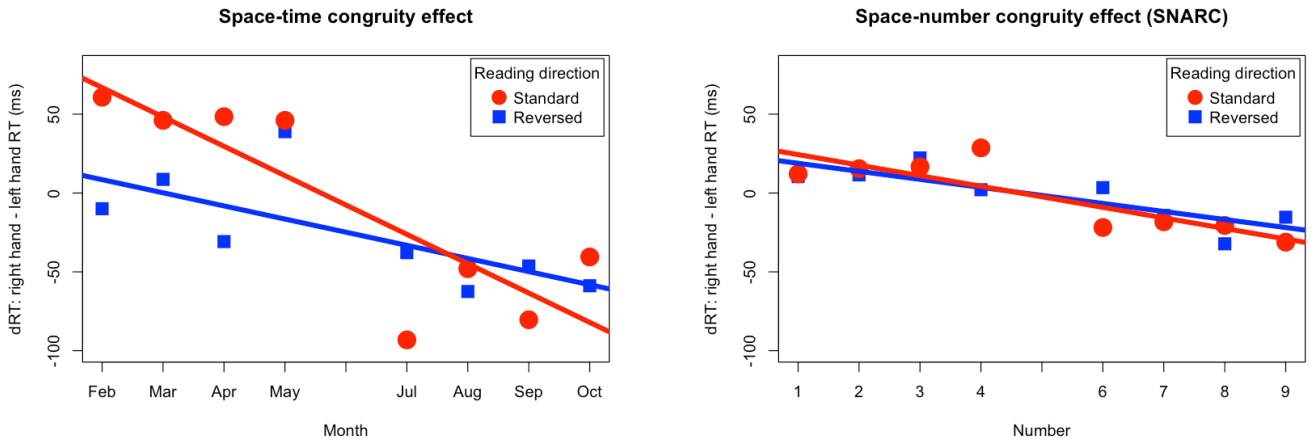


Figure 1. Left: Significant space-time congruity effects differed across reading conditions. Right: Significant space-number congruity effects did not differ across conditions (standard SNARC effect).

such, the lack of such an effect on the MNL in the present study cannot easily be attributed to a paucity of reading training. Nor can it be attributed to a lack of power: By combining data from the two number tasks, we had twice as much item-wise power to detect differences in space-number congruity effects as space-time congruity effects.

How experience shapes mental metaphors

Why does reading experience shape the MTL but not the MNL? Both space-time associations and space-number associations can be considered *mental metaphors*: point-to-point mappings between analog continuums in two different conceptual domains, in which the *source domain* (e.g. space) serves as a scaffold for representations in the *target domain* (e.g. time, number), which is typically more abstract (Casasanto, 2010; Lakoff & Johnson, 1980). The specifics of these mental metaphors are established through correlations in particular kinds of experience (Casasanto, 2013). Manipulating the kind of experience in which source and target domains are correlated should affect the mapping between them; manipulating other kinds of experience should not. This principle predicts the pattern of results we find here: space and time are correlated in the experience of reading text, but space and number are not. Therefore, manipulating reading experience affected the MTL but not the MNL.

Experiential bases of left-right time mappings What kinds of experience provide a correlation between space and time? In the act of reading, as people move their attention through space in either one direction or the other they also "move" through time. When reading each line of an English text, the reader's eyes begin on the left side (at an earlier time) and end on the right side (at a later time). This correlation between progress through time and progress *rightward* through space results in an MTL in which earlier events are associated with the left and later events are associated with the right. In other cultures, reading from right to left produces an MTL that progresses in the opposite direction. Although experience with orthography may be sufficient to determine the direction of the MTL in the laboratory (Casasanto & Bottini, 2014), it is likely that beyond the lab reading experience combines with other culture-specific experiences to shape the MTL: experiences like using calendars, graphs, or written timelines, which also provide a correlation between space and time.

Experiential bases of left-right number mappings What kinds of experience provide a correlation between space and number? There appears to be more than one answer. Here we consider 3 possible experiential bases of space-number mappings: reading and writing text; reading and writing numbers; and finger counting. All three of these practices vary across groups of people, and all three correlate roughly with the direction of the MNL across cultures. Only the last two practices, however, appear to be *causally related* to the direction of the MNL.

Reading and writing text Whereas the act of reading provides a correlation between space and time (no matter what the content of the text may be), reading does not provide a correlation between space and number. Without providing such a correlation, how could reading experience influence the MNL?

In principle, the direction of written text could have an *indirect* influence on the direction of the MNL, via the MTL. Starting in childhood, people experience numbers in a consistent temporal order in both speech and writing. When people count aloud, the word "one" is spoken before "two," etc. Given an MTL that progresses from left to right, the temporal sequence of number words in the count list could cause people to associate numbers that occur earlier with the left and numbers that occur later with the right, producing a culture-specific MNL that reflects reading direction.

However, the present findings do not support this account. To the degree that the direction of the MNL depends on the direction of the MTL, changes in the MTL should cause corresponding changes in the MNL. Contrary to these predictions, changes to the MTL did not produce changes to the MNL in our participants, who showed standard space-number congruity effects regardless of differences in their space-time congruity effects. These findings, therefore, provide no evidence for either a direct or an indirect effect of reading direction on the MNL.

Reading and writing numbers Although numbers are not systematically spatialized in text, they are systematically spatialized on written number lines, which appear on calendars, graphs, rulers, computer keyboards, and other cultural artifacts. Changing the relative left-right positions of smaller and larger numbers has been shown to modulate the SNARC effect. In a training experiment, reading recipes in which large numbers appeared on the left of the page and small numbers appeared on the right caused a positive shift in the slope of participants' SNARC effects (compared to the opposite spatialization of numbers), *even though reading direction was held constant across conditions* (Fischer, Mills, & Shaki, 2010). Thus, the spatialization of written numbers on a page can influence the MNL even when it is in direct conflict with the direction of written words.

Finger counting Like written number lines, finger counting provides a correlation between space and number. In the act of finger counting, each number is assigned to a distinct location in lateral space. Individual differences in finger-counting habits correspond to differences in the SNARC effect: People who start counting on their left hand (left-starters) were found to be more likely to show a standard SNARC effect than right-starters (Fischer, 2008). Likewise, when responding to single digits using all 10 fingers, people are fastest to respond to single digits when the response mapping between numbers and fingers matches their own finger-counting routine (DiLuca et al., 2006). In an experimental test of the role of finger counting experience in determining the direction of the MNL, participants were

trained to count on their fingers either from left to right or from right to left. Participants who counted left-to-right showed a standard SNARC effect whereas those who counted right-to-left showed no SNARC effect. Less than 15 minutes of finger counting from right to left caused a reliable change in the MNL (Pitt & Casasanto, 2014). By contrast, in the present study, more than twice as much experience reading text from right to left caused no such change in the MNL.

Conclusions

Here we challenge the assumption that the directions of both the MTL and the MNL are determined by the direction of reading and writing. Less than 40 minutes of reading mirror-reversed English text reliably changed participants' space-time associations, but left their space-number associations unchanged. These results suggest that the directions of the MTL and MNL are determined by different kinds of experience. Whereas the MTL is shaped by experiences that, like reading, provide a correlation between space and time, the MNL is shaped by other culture-specific experiences that, like finger counting, provide a correlation between space and number.

References

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: evidence for a mental time line. *Neuroscience & Biobehavioral* Reviews, *36*(10), 2257-2273.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition*, *118*(1), 123-129.

Casasanto, D. (2010). Space for thinking. In V. Evans & P. Chilton (Eds.), *Language, cognition, and space: State of the art and new directions* (pp. 453-478). London: Equinox Publishing.

Casasanto, D. (2013). Experiential origins of mental metaphors: Language, culture, and the body. In M. Landau, M.D. Robinson, & B. Meier (Eds.), *The power of metaphor: Examining its influence on social life* (pp. 249-268). Washington, DC: American Psychological Association Books.

Casasanto, D., & Bottini, R. (2014). Mirror Reading Can Reverse the Flow of Time. *Journal of Experimental Psychology: General*. 143(2), 473-9.

Casasanto, D., & Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers. *Cogit. Linguistics, 23*(4), 643–674.

Cooperrider, K., & Núñez, R. (2009). Across time, across the body Transversal temporal gestures. *Gesture*, *9*(2), 181-206.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371–396.

Di Luca, S., Granà, A., Semenza, C., Seron, X., & Pesenti, M. (2006). Finger–digit compatibility in Arabic numeral processing. *The Quarterly J. Exp. Psych.*, *59*(9), 1648-1663.

Fias, W. (1996). The importance of magnitude information in numerical processing: Evidence from the SNARC effect. *Mathematical Cognition*, *2*(1), 95-110.

Fischer, M. H. (2008). Finger counting habits modulate spatial-numerical associations. *Cortex*, *44*(4), 386–92.

Fischer, M. H., Castel, A. D., Dodd, M. D., & Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. *Nature Neuroscience*, *6*(6), 555-556.

Fischer, M. H., Mills, R. A., & Shaki, S. (2010). How to cook a SNARC: Number placement in text rapidly changes spatial–numerical associations. *Brain and Cognition*, *72*(3), 333-336.

Gevers, W., Santens, S., Dhooge, E., Chen, Q., Van den Bossche, L., Fias, W., & Verguts, T. (2010). Verbal-spatial and visuospatial coding of number-space interactions. *Journal of Experimental Psychology. General*, *139*(1), 180–90.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Loetscher, T., Bockisch, C., & Brugger, P. (2008). Looking for the answer: The mind's eye in number space. *Neuroscience*, *151*(3), 725-729.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Ouellet, M., Santiago, J., Israeli, Z., & Gabay, S. (2010). Is the future the right time? *Experimental Psychology*, *57*(4), 308–14.

Pirsig, R. M. (1974). The art of motorcycle maintenance. *William Morrow, New York*.

Pitt, B., & Casasanto, D. (2014). Experiential origins of the mental number line. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1174-1179). Austin, TX: Cognitive Science Society.

Santiago, J., Lupáñez, J., Pérez, E., & Funes, M. J. (2007). Time (also) flies from left to right. *Psych. Bulletin & Review*, *14*(3), 512-516.

Shaki, S., & Fischer, M. H. (2008). Reading space into numbers: a cross-linguistic comparison of the SNARC effect. *Cognition*, *108*(2), 590–9.

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*(2), 328–31.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, *23*(4), 515-557.

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition*, *119*(1), 114–9.

Weger, U. W., & Pratt, J. (2008). Time flies like an arrow: Space-time compatibility effects suggest the use of a mental timeline. *Psychonomic Bulletin & Review*, *15*(2), 426-430.

Wood, G., Willems, K., Nuerk, H., & Fischer, M. H. (2008). On the cognitive link between space and number : a meta-analysis of the SNARC effect. *Psychology Science Quarterly*, *50*(4), 489–525.

Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the reverse SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. *Journal of Cognition and Culture*, *5*(1-2), 1-2.

Cultural Evolution Across Domains: Language, Technology and Art

Monica Tamariz (monicatamariz@gmail.com)
Psychology, Heriot-Watt University, Edinburgh EH14 4AS, UK

Simon Kirby (simon@ling.ed.ac.uk)
Jon Carr (j.w.carr@ed.ac.uk)
Centre for Language Evolution,
The University of Edinburgh, 3 Charles Street,
Edinburgh EH9 1AD, UK

Abstract

The social and cognitive mechanisms of cultural evolution have been studied in detail for different domains: language, technology, the economy, art, etc. However, a model that incorporates the function of a cultural tradition and that is able to compare evolutionary dynamics across cultural domains has not been formulated. By exploring the dynamics of comparable linguistic, technological and artistic experimental tasks, we test the effect of domain-specific function on evolutionary mechanisms such as inheritance, innovation and selection. We find evidence that cultural domain shapes both the structure of the traditions and the way the cultural-evolutionary mechanisms operate. The simplifying effects of cultural transmission are noticeable in language and technology, but not in art; innovation is highest in art and lowest in language; and functional pressures lead to different morphological adaptations across domains. This speaks of a crucial role of function and domain in the evolution of culture.

Keywords: cultural evolution; language; technology; art; Lego; iterated learning; transmission chain; experiment

Introduction

Humans are distinct from all other species because of the scale and diversity of our culture. Our behaviour is strongly influenced by knowledge, beliefs, skills and values that are socially learned, as opposed to genetically inherited (Boyd & Richerson 1985, 2005). Cultural traditions such as languages, social habits, technologies, religion or art change over time, and this change has been characterized as an evolutionary process of descent with modification (Campbell, 1965; Durham, 1991; Whiten et al., 2011; Boyd & Richerson, 1985, 2005; Mesoudi, 2011). Traditions are often transmitted unchanged over the generations thanks to faithful imitation (Tomasello, 1996) and teaching (Csibra & Gergely, 2011). Mutated, or innovative, traits also appear in a cultural tradition due to copying error, intentional innovation or borrowing from other traditions. Cultural traits spread and wane in populations as a result of random, neutral evolutionary processes (Bentley et al., 2007) and of selection pressures (Rogers & Shoemaker, 1971; Boyd & Richerson, 1985, 2005; Rogers, 2003).

Cultural evolutionary processes have been described in detail for language (Croft, 2000; Ritt, 2004; Binder & Smith, 2013; Kirby, Dowman & Griffiths; Kirby, Griffiths & Smith, 2014), technology (Rogers, 2003; Fleming & Sorenson, 2001; Mesoudi et al., 2013), the economy (Nelson & Winters, 1982; Hodgson, 2004), social and political systems (Fukuyama, 2011; Gintis & van Shaik, 2013) and art (Haddon, 1914; Morris-Kay, 2012), among others. These and other studies indicate that culture is not a monolithic, homogeneous phenomenon and that the relevant cognitive mechanism --innovation to generate diversity and imitation to preserve the developments of past generations (Tomasello, 1996; Lewis & Laland 2012)-- interact in different ways with the functions prevalent in distinct cultural domains. A model comparing how cognitive biases interact the functions of culture has not been formulated. The present study represents the first attempt at understanding the diversity and complexity of human culture by comparing directly how cultural evolutionary processes unfold across domains. We investigate whether evolutionary mechanisms like inheritance, innovation and adaptation vary across three domains with very different cultural functions: language, technology and art.

Cultural evolutionary pressures derived from transmission

Culture evolves under two main pressures: one related to transmission, and one to usage. The first pressure requires cultural patterns to be learnable and reproducible so they can survive over generations. Under this pressure, patterns that are simple and compressible are selected for (Kirby, Cornish & Smith, 2008; Tamariz & Kirby, 2015). The second pressure requires cultural traditions to be useful for their relevant function. Under this pressure, traditions that are functional and useful --which may imply diversity or complexity or both-- are selected for (Regier, Kemp & Kay, 2015; Kirby et al., 2015).

An example of the kind of solution that evolve in the face of these two opposed forces is the systematic, regular structure of languages. Cross-linguistic studies have looked at the distribution of structure in lexical domains such as colour or kinship terms across world languages, and found that world languages simultaneously optimize simplicity and informativeness (Regier et al. 2015).

Iterated learning experiments and computer models help understand the social and cognitive mechanisms that may

have led to this distribution. Using miniature artificial languages made up of words associated to objects, Kirby et al. (2008, 2015) modeled the effects of transmission and of usage. In transmission experiments, a participant is trained on a language, and then she is tested on it. Her output is used to train the next participant, forming a transmission chain. After several generations, the languages have very few distinct words left: they have become highly compressible, and therefore easy to learn. But they cannot name all the meanings (Kirby et al., 2008). Kirby et al. (2015) ran a communication condition in which two participants are trained on a language and then use it to play communication games with each other; after many interactions, the languages stay expressive: they maintain a distinct word for each object. They also ran a condition that included both transmission and communication: a pair learns a language and then play communicative games with it; the words they produce at the end of the game are used to train the next pair in a transmission chain. Here, under pressure for transmission and for utility, the languages developed regular systematicity over generations to become simultaneously easy to learn, and also fully expressive, with a distinct word for each object.

The conclusions of the studies reviewed above lead us to predict an increase in compressibility, or simplicity, in response to cultural transmission in the three domains. In the language domain, we should also expect a pressure for expressivity arising from communicative function, which puts a limit to how simple linguistic patterns can get. This raises the question whether other domains have similar breaks on simplicity. The following section explores this question and justifies our predictions for the effects of function in each of our cultural domains.

Domain-related predictions

Art is produced for the purposes of self-expression and evaluated by criteria like beauty or artistic value, making it a highly subjective domain of culture. Each individual may have his or her own expressive biases and ideas about artistic value. The functional pressures guiding the evolution of art may thus vary across individuals, and the artistic tradition, therefore, might not show the signature of directional selection for specific features over several generations. Additionally, we might expect innovation to have some functional benefit in art. We therefore predict low transmission accuracy and also a reduction in the evolution of simplicity in this domain.

The goals of technology are usually objective and clearly defined, which leads us to predict high transmission accuracy of adaptive traits which contribute to achieving the goal at hand, combined with the predicted pressure for simplicity.

Unlike technological solutions, which are strongly constrained by the structure of the problem they have to solve, linguistic signals are not constrained by the structure of their referents. Most signals in human languages are arbitrary, that is, they do not resemble or are caused by their referent meaning (de Saussure, 1983 [1916]). In order to be communicatively functional, however, signals must be shared conventionally by the interlocutors. An innovation in the signal has the potential to disrupt a convention, as the listener might not be able to recognize it. For these reasons, and also based on the studies reviewed above, we predict low innovation of signals, in addition to the predicted selection for simplicity caused by transmission.

Experimental Testing of the Effects of Cultural Domain on Evolutionary Dynamics

In order to test these predictions we compare the evolution of cultural traditions pertaining to the three domains in a transmission chain experiment where participants use similar materials in a similar task (making Lego constructions), but with different functions, corresponding to each of the domains.

Methods

Participants. Ninety-nine participants (66 were female; ages ranged from 18 to 35, mean = 21.2, SD=3.54), took part in this study. The results from three of them were excluded because of non-adherence to the instructions. They were recruited through the University of Edinburgh employment service and also in the Psychology building. The experiment lasted 6 to 7 minutes in total, for which each participant could choose to be rewarded with between £1 or a snack (a can drink and a chocolate bar).

Materials. Participants come into the lab to find on the table 45 identical 2x4 orange loose Lego bricks, a number of objects (see next section for details) and a Lego construction standing next to each object (see e.g. Fig. 1). The objects were a tennis ball, a flower made of fabric and plastic, a £1 coin, a wooden spoon and a mesh bag containing pebbles.

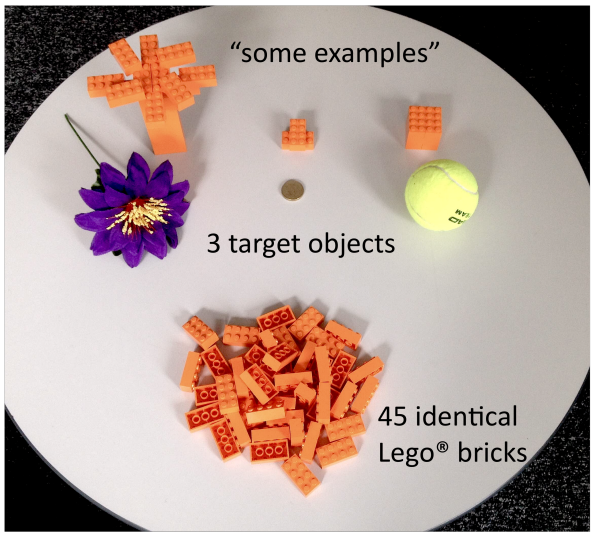


Figure 1: Experimental setup in one of the trials in the Art condition.

Procedure. After signing a consent form, participants were instructed to make one construction for each of the objects using the Lego bricks. In the Art condition three objects were present (the ball, the flower and the coin) and the participants were told to use the loose bricks to produce 'original' works of art that would be 'judged for their artistic value'. In the Technology condition, the same three objects were present and the participants were asked to make a stand out of Lego to support each of them so that the sum of heights measured from the table to the top of each objects was maximized. In the Language condition, all five objects were present, as was an interlocutor (another participant), sitting across the table. Participants had to make signals with the Lego bricks to communicate the objects to the interlocutor during a series of naming games: the experimenter showed a photograph of one of the objects to the participant, he made a Lego signal in view of the interlocutor, who had to guess which of the five objects the signal referred to. They were both told they scored a point for each correct answer, and their goal was to score as many points as possible in the time available. The experimenter showed pictures of the objects in random order, and told the interlocutor and noted down whether each response was correct. In all domain conditions, the maximum time allocated for the task was 5 minutes. Each participant was tested on only one condition.

Transmission chains. Participants were organized in transmission chains, so that the 'example' Lego constructions seen by one participant (or pair of participants in the Language condition) were the constructions produced by the participant in the previous generation in the chain. The first participant in each chain, therefore, saw no examples. They were not told what to do with the examples; they could copy them, draw inspiration from them or ignore them.

Data analysis. The 216 Lego constructions related to the three objects common to all conditions (the ball, the flower and the coin) were manually coded: we recorded the number of pieces, overall height, width and depth, volume (of the smallest cuboid box that could contain the construction), and area of the faces that were external or visible (long sides, short sides, top and bottom). Using these data, we estimated complexity using a 3D version of perimetric complexity, equal to the squared total external and any internal areas in mm^2 divided by the volume contained within those areas in mm^3 (yielding units mm^{-1}). This measure is an extension of standard 2D perimetric complexity, equal to the squared length of the inside plus outside perimeters of a drawing divided by the ink area contained within those perimeters, and known to be a good estimate of processing effort for a graphic (Pelli et al. 2006).

In order to estimate the level of innovation during transmission, each pair of constructions for the same object in consecutive generations (e.g. the construction inspired by the ball in the art condition, chain 1, produced at generations 1 and 2) was rated for difference (a proxy of innovation) on a scale of 0 to 9 points (0 if they were identical, 9 if they were completely different) by three independent coders. The inter-coder agreement was very high (Pearson's r values ranged between 0.821 and 0.831, N=180, p<0.001).

We measured functional success in two of the domains. In technology, the height in cm from the table to the top of the object. In Language, the number of successful communicative interactions in five minutes.

We used R (R Core Team, 2012) and *lme4* (Bates, Maechler & Bolker, 2012) to perform linear mixed effects analyses of the effects of Lego figure innovation and morphological features (complexity, width, length and heigth) on cultural domain. As fixed effects, we entered Domain (Art, Tech and Language), Object (Ball, Flower, Coin) and Generation (1–6 for Complexity, 1-5 for Innovation) (without interaction term) into the model. As random effects, we had an intercept for Chain (1-12), as well as by-Chain random slope for the effect of Domain, Object and Generation. *Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality*. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question (Winter, 2013).

Results

Cultural domain had a significant effect on complexity ($\chi^2(2)$=13.39, p=.0012). The least complex constructions were produced in the language domain (average 2.50mm^{-1} ± 0.68 S.E.); in the technology domain, complexity increased by 3.37mm^{-1} ± 0.74 S.E. and in the art domain, by 2.94mm^{-1} ± 0.74 S.E. (Fig. 2).

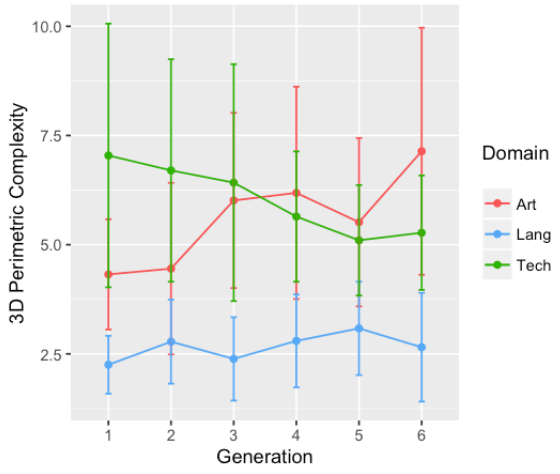


Figure 2: Complexity in the three domains over generations (95% C.I. shown throughout).

There was an interaction between domain and generation ($\chi^2(10)$=18.573, p=0.046) due to the complexity of in the art domain increasing and in the technology domain condition decreasing over generations (Fig. 2).

Object also affected complexity ($\chi^2(2)$=52.83, p=3.4x10^{-12}). Here, the least complex constructions were associated

with the coin (average 0.61mm^{-1} ± 0.68 S.E.); for the flower, complexity increased by 3.16mm^{-1}± 0.40 S.E. and for the ball, by 1.68mm^{-1} ± 0.40 S.E.

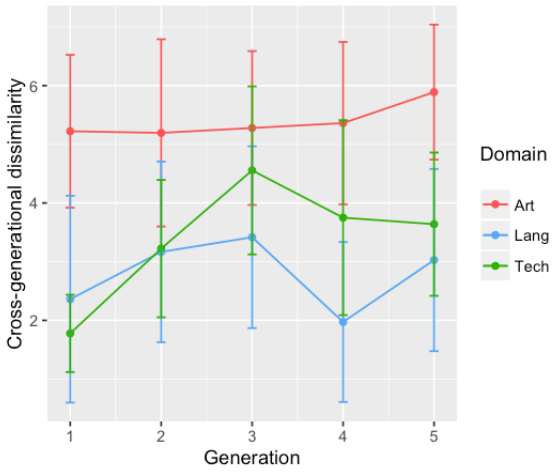


Figure 3: Innovation in the three domains over generations.

Regarding innovation, there was a significant effect of Domain ($\chi^2(2)=11.65$, $p=0.003$). Innovation was greatest in the art domain (average 5.11 points ± 0.57 S.E.), with a drop of 2.60 points ± 0.61 S.E. in the language domain and a drop of 2.00 points ± 0.61 S.E. in the technology domain (Fig. 3).

There was also a significant effect of object on innovation ($\chi^2(2)=19.80$, $p=5\text{x}10^{-5}$). Innovation was greatest for the flower (average 5.42 points ± 0.57 S.E.),, with a drop of 0.32 points ± 0.36 S.E. for the ball and a drop of 1.56 points ± 0.36 S.E. for the coin.

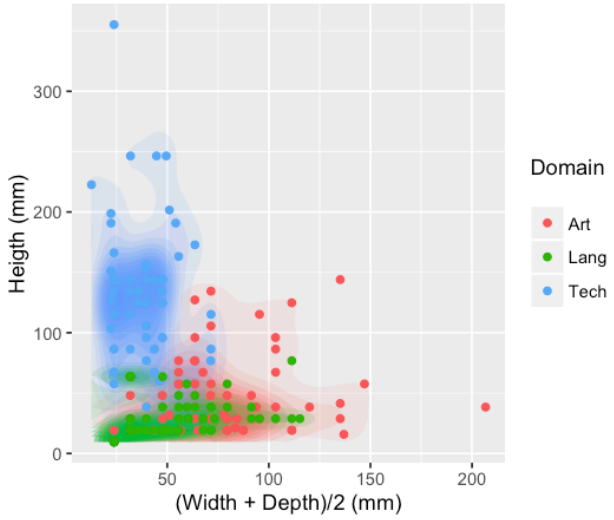


Figure 4: Height of the Lego constructions in the different domains against horizontal dimensions

As far as the morphological features are concerned there was an effect domain on Lego construction height ($\chi^2(2)=30.161$, $p=2.8\text{x}10^{-7}$), with the towers in the Technology condition were significantly taller (on average 123.06mm ± 9.14 S.E.) than the artworks (80.72mm ± 8.78 S.E. shorter) and the linguistic signals (95.3mm ± 8.78 S.E. shorter). Fig. 4 shows height versus horizontal dimensions width and depth in the three domains.

The number of Lego pieces and the volume also showed an effect of domain. For number of pieces, $\chi^2(2)=34.46$, $p=3.3\text{x}10^{-8}$; technology had the most pieces, with 15.35 ± 1.44 S.E. pieces, with a drop of 2.66 ± 0.88 S.E. pieces in art and a drop of 9.59 ± 0.88 S.E. in language. For volume, $\chi^2(2)=17.84$, $p=1.3\text{x}10^{-4}$; here, art had the largest volume, 226cm^3, with a drop of 109 cm^3 ± 40 S.E. pieces in technology and a drop of 195 cm^3 ± 40 S.E. in language.

Functional success in the language and technology domains did not change significantly over generations: there was no effect of generation on height of tower plus object in the technology domain constructions ($\chi^2(1)=0.012$, p=0.91). The effect of generation on communicative success in the language domain was not significant either ($\chi^2(1)=3.021$, p=0.082), although success in the last four generations visibly increases with respect to the first two (Fig. 5).

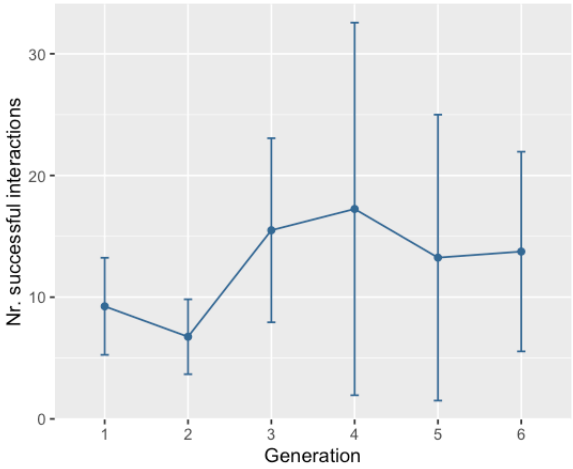


Figure 5: Success over generations in the language domain.

Discussion

Summary of results. First, for language, we predicted simplifications (or reduction in complexity) over generations and low levels of innovation. We found no evidence for the first prediction: Lego figures in the linguistic task are significantly simpler than in the other two domains (Fig. 2), but complexity does not change over generations. The second prediction is reflected in the results: out of the three domains, language shows the lowest innovation levels (Fig. 3). Second, for art, our predicted absence of simplification is clear from the results: we actually saw an increase in complexity over generations for art (Fig. 2); we also predicted high levels of innovation, which, again, are found in the results (Fig. 3). Thirdly, for technology, we found evidence for the predicted simplification of constructions (Fig. 2) and for selection of traits related to the function: Lego constructions in this

domain are clearly taller and narrower than those in art or language (Fig. 4).

Simplification. Our results clearly show that cultural domain has an effect on simplification. The absence of the expected simplification in language could be explained by the fact that the initial items were already at floor in terms of complexity. In the other domains, we see an interaction between simplicity and domain over generations. A bias in favour of innovation may have been sufficient to drive an increase in complexity in the art domain, whereas in the technology domain, we observe some evidence for the decrease in complexity predicted by previous work.

The floor effect in the language domain may be explained by the details of experimental design. In previous iterated learning experiments, the first participant in each chain was trained on random items. In our chains, in contrast, the first participant of each chain improvised signals in the absence of examples. These chain-initial signals already had very low levels of complexity.

Innovation. Cultural domain also modulates the rate of innovation (mutation) that takes place during transmission (Fig. 3). The high innovation level in the art condition may be explained by the perception that originality increases artistic value; selection for novelty thus affects the transmission accuracy. Low innovation in language may respond to the fact that conventional signals, which we share with our interlocutors, are more likely to be effective in the short term than innovative signals, even if these have other desirable properties.

The effects of selective pressures on fundamental evolutionary mechanisms like inheritance of or mutation are specific characteristics of cultural evolution that have no equivalent in the paradigmatic case of evolution, biology, where selection affects traits, not the mechanisms of inheritance or mutation. Our results show that the mechanisms of inheritance and mutation in culture can *adapt* to the functional domain of a tradition.

Adaptation. Domain also seems to pose selective pressures on the morphological traits of the Lego constructions. Significant increases in functional success (number of successful communicative interactions, tower height and artistic value) would have constituted evidence for adaptation: traits that increase functionality are more likely to be copied by the next generation, resulting in increased performance. We have not found such cumulative effect over time, but nevertheless there are clear differences in size and shape between the constructions across domains to indicate adaptation of the constructions to the domain-specific task. In the technology condition, the vertical dimension of the Lego stands is significantly larger in the works of art or linguistic signals (Fig. 4), a clear fit to the task of maximizing height. The art constructions are the largest in volume (even with fewer pieces than the technology ones), which we may interpret as an adaptation to creativity. As for the linguistic signals, they have the least pieces, volume, as well as horizontal and vertical dimensions. Communicative success showed a marginally significant upward trend over generations ($p=0.08$, Fig. 5), indicating selection for traits that adapted to the communicative task.

Object. The reported effect of object (flower, ball, coin) is also relevant to our question, because each object poses functional constraints on the Lego constructions in addition to those posed by cultural domain. Across conditions, the constructions associated with the coin had lowest complexity and innovation, those with the flower, highest complexity and innovation, with those for the ball in between in both measures. This can be explained by iconicity and size: the constructions mirrored the object size: fewest Lego bricks for coin and most for flower; the number of bricks may constrain possible innovations, with more bricks affording more changes. The analysis returned no significant interactions between object and domain for either complexity or innovation, indicating that the objects behaved similarly across domains.

Conclusion

The study reported here represents the first direct comparison of cultural evolution across cultural domains. We have found that the affordances and constraints of a linguistic, a technological and an artistic task have significant effects on the evolution of the cultural traditions, and also on key evolutionary mechanisms such as transmission fidelity and innovation. The mechanisms and pressures that operate in cultural evolution unfold differently across domain: the simplification driven by transmission is visible in the linguistic, but not in the artistic domain; and the rate of innovation is greater in art than in technology or language. This is an important step towards understanding the huge diversity of traditions found in culture. Furthermore, our study pioneers exciting possibilities of using lab techniques to test hypotheses regarding the mechanisms of cultural evolution across domains.

Acknowledgments

This research benefitted from funding from a University of Edinburgh PPLS Pilot grant.

References

Bates D, Maechler M, Bolker BM and Walker S (2015). "Fitting Linear Mixed-Effects Models using lme4." ArXiv e-print; in press, *Journal of Statistical Software*, URL: http://arxiv.org/abs/1406.5823.

Bentley, R.A., Lipo, C.P., Herzog, H.A. & Hahn, M. (2007). Regular rates of popular culture change reflect random copying. *Evolution and Human Behavior* 28: 151–158.

Binder, P.M. & Smith, K. (eds.) (2013). *The language phenomenon: Human communication from milliseconds to millennia*. Springer Press.

Boyd, R. & Richerson, P.J. (1985). Culture and the Evolutionary Process. Chicago: University of Chicago Press.

Boyd, R. & Richerson, P.J. (2005). Not by genes alone. Chicago: University of Chicago Press.

Campbell, D.T. (1965). Variation and selective retention in socio-cultural evolution. In: Barringer, H.R., Blanksten, G.I. & Mack, W. (eds.) *Social change in developing areas: A reinterpretation of evolutionary theory*, 19-49. Cambridge: Schenkman.

Christiansen, M.H. & Kirby, S. (eds.) (2003). Language evolution. Oxford: Oxford University Press.

Croft, W. (2000). Exploring language change: An evolutionary approach. Harlow: Longman.

Csibra, G. & Gergely, G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B,* 366: 1149-1157.

de Saussure, F. (1983 [1916]). *Course in general linguistics*. London: Duckworth.

Durham, W.H. (1991). *Co-evolution: Genes, culture and human diversity*. Stanford: Stanford University Press.

Fleming, L. & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy*, 30: 1019-1039.

Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French Revolution*. New York: Farrar, Straus and Giroux.

Gintis, Herbert; van Schaik, Carel (2013). Zoon Politicon: The evolutionary roots of human sociopolitical systems. In: Richerson, P. J. & Christiansen, M.H. (eds.) *Cultural evolution: Society, technology, language, and religion*, 25-44. Cambridge, MA: MIT Press.

Haddon, A.C. (1914). *Evolution in art*. London: Walter Scott Ltd.

Hodgson, G. (2004). *The evolution of institutional economics: Agency, structure and Darwinism in American institutionalism*. London: Routledge

Kirby, S., Cornish, H. & Smith, K. (2008) Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language. *PNAS,* 105: 10681-10686.

Kirby, S., Dowman & M., Griffiths, T.L. (2007). Innateness and culture in the evolution of language. *PNAS,* 104: 5241-5245.

Kirby, S., Griffiths, T. L. & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28*, 108-114.

Kirby, S., Griffiths, T. L. & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28*, 108-114.

Kirby, S., Tamariz, M., Cornish, H. & Smith, K. (2015) Compression and Communication Drive the Evolution of Language. *Cognit*ion, 141: 87-102.

Lewis, H. & Laland, K.N. (2012). Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society B*, 367: 2171-2180.

Mesoudi, A., Laland, K.N., Boye, R., Buchanan, B., Flynn, E., McCauley, R., Renn, J., Reyes-Garcia, V., Shennah, S. Stout, D & Tennie, C. (2013). The cultural evolution of technology and science. In Richerson, P.J. & Christiansen, M.H. (eds.), *Cultural evolution: Society, technology, language, and religion,* 193-218. Cambridge: MIT Press.

Mesoudi, A. (2011). Cultural evolution: How Darwinian theory can explain human culture and synthesize the social sciences. Chicago: University of Chicago Press.

Morris-Kay, G.M. (2010). The evolution of human artistic creativity. *Jounal of Anatomy*, 216: 158-176.

Nelson, R.R. & Winters, S.G. (1982). An evolutionary theory of economic change. Cambridge, MA: The Belknap Press.

Pelli, D. G., Burns, C.W., Farrell, B., & Moore, D. C. (2006). Identifying letters. *Vision Research*, 46: 4646-4674.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Regier, T., Kemp, C. & Kay, R. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (eds.), *The handbook of language emergence*, 237-263. Hoboken, NJ: Wiley-Blackwell.

Ritt. N. (2004). Selfish sounds. A Darwinian approach to language change. Cambridge: University Press.

Rogers, E.M. (2003). *Diffusion of innovations. 5th edition*. New York: Free Press.

Rogers, E.M. & Shoemaker, F.F. (1971). Communication of innovations: a cross-cultural approach. New York: Free Press.

Tamariz, M. & Kirby, S. (2015). Culture: Copying, compression and conventionality. *Cognitive Science*, 39: 171-183.

Tomasello M. (1996). Do Apes Ape? In *Social learning in animals: the roots of culture*, Heyes, C.M. & Galef, B.G. (eds.), 319-346. London, UK: Academic Press.

Whiten, A., Hinde, R.A., Laland, K.N. & Stringer, C.B. (2011). Culture evolves. *Philosophical Transactions of the Royal Society B*. 366, 1567: 938-948.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv: 1308.5499. URL: http://arxiv.org/pdf/1308.5499.pdf

A Neural Field Model of Word Repetition Effects in Early Time-Course ERPs in Spoken Word Perception

Andrew P. Valenti (andrew.valenti@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Michael Brady (michael.brady@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Matthias J. Scheutz (matthias.scheutz@tufts.edu)
Human-Robot Interaction Laboratory, Tufts University, 200 Boston Ave.
Medford, MA 02155 USA

Phillip J. Holcomb (pholcomb@tufts.edu)
Department of Psychology, Tufts University, 490 Boston Ave.
Medford, MA 02155 USA

He Pu (he.pu@tufts.edu)
Department of Psychology, Tufts University, 490 Boston Ave.
Medford, MA 02155 USA

Abstract

Previous attempts at modeling the neuro-cognitive mechanisms underlying word processing have used connectionist approaches, but none has modeled spoken word architectures as the input is presented in real-time. Hence, such models rely on the ingenuity of the modeler to establish a mapping of real-time stimulus to the model's input which may not preserve processing that happens during each time step. We present a neural field model which successfully replicates the effect of immediate auditory repetition of monosyllabic words and fits it to a component of a well-studied mechanism for analyzing language processing, the event-related potential (ERP). This represents a new modeling approach to studying the neuro-cognitive processes, one that is based on the bottom-up interaction of real-time sensory information with higher-level categories of cognitive processing.

Keywords: dynamic neural fields; event-related potential (ERP); spoken word perception; mental workload; computational modeling; word repetition

Introduction

By "spoken word perception", we mean the cognitive processes that entail the sensory intake of an acoustic waveform until the words contained in it are identified. Some early connectionist models of speech perception processes were driven by research in generalized automatic speech recognition and have shown, for example, that a good deal of phonemic information is present in the auditory signal and can be extracted from the statistical generalization of the model. Among the best-known models of speech perception is TRACE (McClelland & Elman, 1986) which has modeled several lexical effects (e.g., phonemic restoration in a noisy environment) and the time-course of word recognition. TRACE has been criticized for its biologically unrealistic handling of time and the lack of a learning mechanism (Protopapas, 1999). As a result, models were developed (Elman, 1990; Norris, 1995) which represent time through cyclical, recurring connections from one state to an earlier state in the network. One popular method by which learning is incorporated in these networks is through a gradient decent regression using backpropagation.

While these models can account for many aspects of how humans comprehend spoken and written words, none of these architectures model speech perception using real-time, human input. We present a neural field model with an efficient learning mechanism which dynamically responds to the spoken word process as it unfolds over time. A neural field sits in an equilibrium state waiting for a pattern it has tuned itself to detect, and this detection takes the form of a perturbation. Learning associates the equilibrium state of a field with its environment. Primary fields tune themselves to fall into systematic equilibrium states in response to combinations of sensory input. Deeper-processing, secondary neural fields are then enabled to tune themselves in response to their environments once primary fields have settled into predictable behaviors. With experience, the network forms representations as each neural field systematically responds to its environment through time.

Word Repetition Effects and ERPs

An event-related potential (ERP) is an electrical voltage associated with an event such as a stimulus or response. ERPs are believed to reflect the summation of post-synaptic potentials occurring in many thousands of neurons. The time course of ERPs in auditory processing can be traced starting from stimulus onset and continuing for approximately 800 ms. Our study focused on a particular ERP known as the P200 (P2) which occurs in the interval from 145 ms to 225 ms after

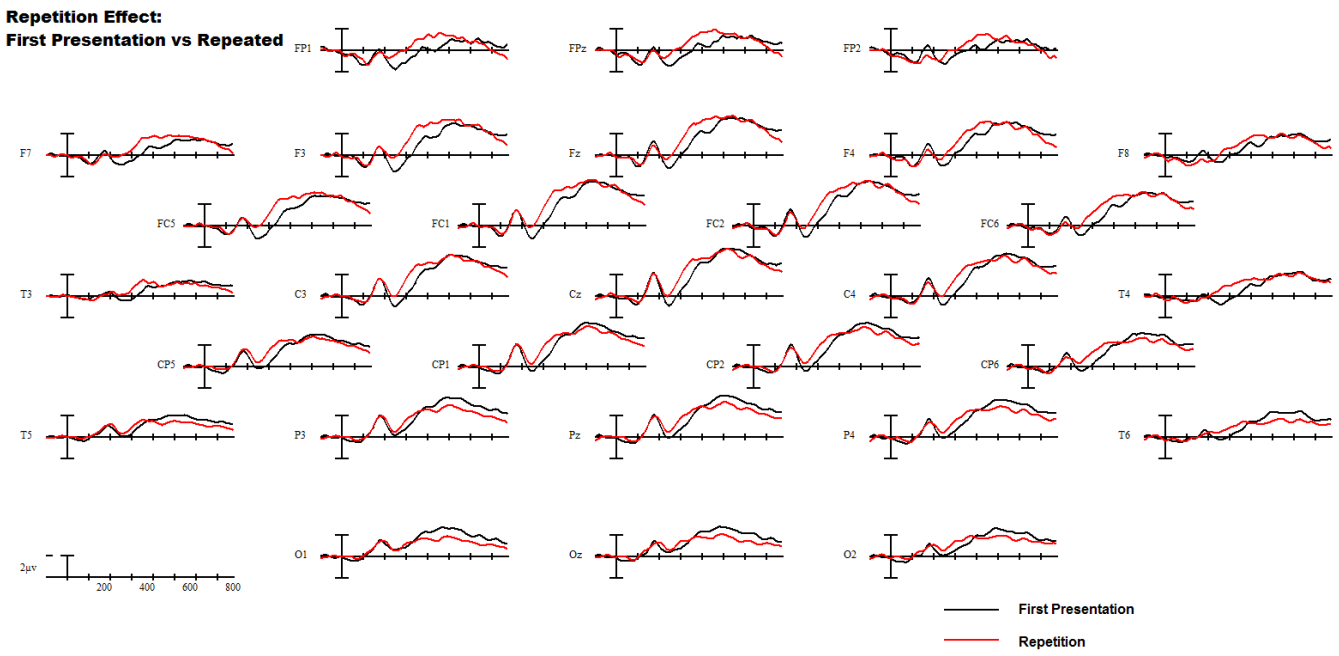


Figure 1: ERP repetition effects, seen in the difference between the first presentation (black line) or a word and the immediate repetition (red line) of that word

stimulus onset and is classically associated with top-down attention processes on early sensory processing (Hillyard & Anllo-Vento, 1998). Of particular interest, the P2 has also been associated with a word repetition effect (Luck, 2014; Molfese, Key, Maguire, Dove, & Molfese, 2005) where the P2 showed a reduced positivity (i.e., a larger negativity) to primed versus unprimed targets. Word repetition is frequently used as an investigative tool in psycholinguistic and memory research. It is a simple empirical procedure which demonstrates that subjects are usually faster in their response to the second presentation of words than the first; such responses may be captured via reaction time (RT) measures across a variety of experimental paradigms such as lexical decision or semantic categorization.

Prior research in which participants read short texts containing repeated words has found three distinct ERP components to be sensitive to repetition: a positive component peaking around 200 ms post-stimulus, a negative component at 400 ms (N400) and a later positivity (van Petten, Kutas, Kluender, Mitchnier, & McIsaac, 1991). However, van Petten et al. (1991) note that the early P2 repetition effect has not been consistently found in other studies, at times appearing with an opposite polarity. Due to the paucity of research using real-time speech signals and the conflicting early results cited, it appears that the processes which control this early component are not well-understood. Among the research questions that remain open are to what extent does deeper lexical processing and explicit memory influence the word repetition effect and what particular cognitive processes elicit this effect? While this paper did not set out to explore these questions in depth, we address some of them in the context of our results.

Human Experiments and ERP Data

Empirical ERP Data

We collected ERP data from 12 Native English speakers from Tufts University (mean age 19.6, 7 male), of which 2 were excluded due to excessive ocular artifacts. All participants self-reported as monolingual and right-handed (Oldfield, 1971), with normal or corrected-to-normal vision/hearing and normal neurological profile. Participants provided written informed consent and were monetarily compensated, as approved by the Tufts University Institutional Review Board.

Materials and Design

During ERP recording, participants completed a dual-task paradigm with a primary task of playing a video game (i.e., "Breakout": breaking pre-arranged blocks by bouncing a ball from a controllable paddle) and a secondary task of listening to words through a set of headphones. The dual-task paradigm was important for our ERP modeling task because we attempted to reduce any explicit episodic memory effect so that we could focus on more implicit repetition primary effects by introducing the primary task of playing a video game. For the primary task, we utilized a JavaScript variant of Breakout. Three game levels were chosen based on pilot results, indicating them to be similar in difficulty. For the secondary task, a female experimenter recorded 300

monosyllabic English words to be used in stimuli generation. These 300 words were split into two lists (of 150 words each) matched for psycholinguistic properties (e.g., bigram frequency, length, phonological and orthographic frequency, familiarity, and concreteness). An additional list was created from the two split lists (half from each) so that a total of three lists of 150 words were created. From each of the 3 lists, 50 of the 150 words were randomly selected to be repeated so that each list contained a total of 200 words. None of the repeated words were redundant across lists.

EEG Recording

Participants engaged in the dual-task paradigm in a dark, sound-attenuated room while their EEG was recorded using a 29-channel electrode cap. Loose electrodes recorded from 1) below the left eye (LE) to monitor for blinks and vertical eye movements, 2) at the right temple (HE) to monitor for horizontal eye movements, and 3) behind each mastoid (left: A1, right: A2) for referencing (A1) and monitoring differential mastoid activity (A2). Electrode impedances were kept under 5 k for all scalp electrodes, 10 k for both eye electrodes, and 2 k for both mastoid electrodes. We sampled the EEG at 200Hz while an SA Bioamplifier (SA Instruments, San Diego, CA) amplified the signal with bandpass of 0.01 and 40 Hz.

Experimental Results

Averaged ERPs were formed for each spoken word (using -100 and 0 ms baseline) after artifact rejection (15.67% of the trials were rejected due to ocular artifacts) and collapsed into conditions (first presentation or repeated) for comparison. The ERPs were then low-pass filtered at 15 Hz. Individual participant ERPs were then averaged into a grandmean of 10 participants, allowing for the analysis of overall auditory language processing effects. Of particular interest is the repetition effect on particular ERP components such as the P2 (van Petten & Kutas, 1991; Rugg, 1987) with an anterior scalp distribution, sensitive to lexical processing and implicated in word recognition processes (Dambacher, Kliegl, Hofmann, & Jacobs, 2006). Such repetition effects manifest in the form of attenuated amplitudes to repeated items compared to their first presentation, reflecting the ease of processing for the former relative to the latter. Results indicate the presence of a P2 repetition effect, seen clearly in anterior electrodes between 200 and 400 ms (Figure 1).

Model Description

We modeled a single layer of the hierarchical process generally regarded to represent the architecture of speech perception (Grossberg, 2005; McClelland & Elman, 1986; Norris, 1995). In Figure 2, the model architecture consists of (1) a vector of *auditory input nodes*, (2) a vector of *category nodes*, (3) a grid of processing units called a *neural field*, and (4) three fully connected sets of weights to be trained called *adaptive filters*. The field processing units are reciprocally connected to each other through non-adjustable

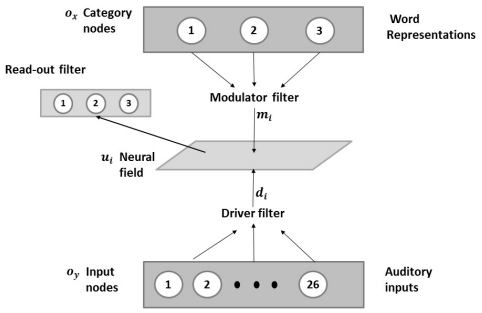


Figure 2: Neural field training. The training vector at the word representation layer develops an input signal $s = m_i$ through the modulator filter to each processing unit u_i in the neural field as a random sound exemplar of the same training vector category is played to the input nodes.

weighted connections using an on-center, off-surround "Mexican hat" distance function (Brady, 2014). The input nodes carry sensory information which is refreshed with new data at each time step. This input is passed through a "driver" filter to develop a bottom-up input signal to the field. The category nodes carry persistent labeling information which is passed through a "modulator" filter to provide a top-down input signal to the field. The labeling information is also used as the training target for a "read-out" filter.

A neural field in our model is a "sheet" of processing units. If given no input and random initial conditions, all units of the field are guaranteed to quickly fall into a stable equilibrium state with respect to each other such that the entire field may be considered to fall into an equilibrium. Different equilibrium states of the field are associated with different input patterns. The field is updated once every 10 ms (i.e., a time step) using Equation 1 which computes the change in its activation. This general equation and its variations are widely used in dynamical systems models, (e.g., Amari, 1977; Beer, 2000; Brady, 2014; Grossberg, 2005; Hopfield, 1982; Schöner & Spencer, 2015).

$$\dot{u_i} = -u_i + s_i + h + n + \sum_j \lambda(i,j) \cdot \sigma(u_j) \qquad (1)$$

The change in activation of a unit, u_i at a given time step is computed as the sum of influence to the unit at that time step minus the activation of the unit from the previous time step. Influence to a unit at a time step comes from an input signal, s_i, the field's slightly negative bias, h, a noise term, n, and from other units within the field. Influence from other units within the field is computed to be the sum of the squashed activations of neighboring units multiplied through corresponding within-field connection weights w. A stepwise squashing function, σ, is used such that only units with non-negative activations can influence their neighborhoods. Within-field con-

nection weights are specified as on-center off-surround by a Mexican hat weighting function, $\lambda(D)$. Input to the function D is the Euclidean distance between two units, u_i and u_j; the output of the function specifies their connection strength.

Neural Field Learning

We implemented a learning mechanism in which the driver and modulator filters are trained together that works as follows. The filter weights are initialized with random values which are then updated across training cycles. A training cycle consists of iterations in which the neural field is initialized with random unit activations simulating the passage of time between learning patterns. Then, a training vector is used to generate an input signal s_i through the filters to each unit of the neural field using Equation 2, and a random sound exemplar of the same category as the training vector is played to the input nodes as time unfolds. In our experiment, the training vector represents a monosyllabic word. Here, o_y is the activation of a category node, o_x is the activation of an input node, and g_1, g_2, g_3 are gain terms; $\dot{d}_i$ is the change in activation of the driver signal, $\overline{u}_i$ is the running average of the unit being updated, and $\overline{m}_i$ is the running average of the modulator signal to a unit.

$$s_i = g_1|\dot{d}_i| \cdot (g_2\overline{m}_i - g_3\overline{u}_i) \tag{2}$$

$$\overline{m}_i = \sum_y w_{iy} \cdot o_y$$

$$\dot{d}_i = \sum_x w_{ix} \cdot o_x$$

The weights of the modulator and driver filters are adjusted following Equation 3, a variant of the delta training rule.

$$\Delta w_{ix} = \eta \cdot \bar{o}_x \cdot (\overline{u}_i - \dot{d}_i) \cdot |\dot{u}_i| \tag{3}$$

$$\Delta w_{iy} = \eta \cdot \bar{o}_y \cdot (\overline{u}_i - \overline{m}_i) \cdot |\dot{u}_i|$$

Learning proceeds as the training vector persists for the duration of the input sound as the neural field adjusts itself in response to its input, updating the modulator and driver filters at each time step. Subsequently, a new iteration begins by initializing the field to a new random state and associating the transformation of that state through time with the next input training vector (i.e., new word), and so on. A cycle is completed when all training vectors have been exposed to the model in random order, at which point a new training cycle begins.

In Equation 3, η is the learning rate, $(\overline{u}_i - \dot{d}_i)$ and $(\overline{u}_i - \overline{m}_i)$ are the errors to be minimized; cyclic training continues until the learning error is reduced to asymptote. The last term of the equation, $|\dot{u}_i|$, is an innovation which allows learning to occur only if there is a change in the target neural field and therefore important associations are maintained even as learning proceeds over time.

Neural Field Model Simulations and Results

The model's read-out filter is trained in order to evaluate how well the neural field categorizes its input. The weights of this read-out filter are updated using the "delta rule" as in Equation 4. Training vectors o_y are converted to target vectors T_y by setting the negative values of the training vectors all to zero. The generated output is notated as $\hat{o}$.

$$\Delta w_{yi} = \eta \cdot u_i \cdot (T_y - \sigma(\hat{o}_y)) \tag{4}$$

Where:

$$\hat{o}_y = \sum_i w_{yi} \cdot m_i$$

We selected a subset of five monosyllabic words from the stimuli used in the empirical experiment: "beach", "dog", "soup", "bog", and "tend". Four exemplars of each word were recorded separately by a male speaker as male voices span a lower frequency range making for easier speech processing by the model. The recordings were transformed into the 26 coefficients shown in Figure 2 and were provided as input to the model in 10 ms time steps. The model was trained on three target words from this set, "beach", "dog", and "soup". How well the model learned was measured by computing the error as the sum of the differences between "readout" vector generated as output by the model and the corresponding target word.

Modeling the ERP Measure

We chose to model the ERP as the difference between the modulator signal and the field activation. This can be thought of as analogous to error values or implicit prediction error. Implicit prediction error at multiple levels of language processing is thought to play a critical role in language comprehension (Kuperberg & Jaeger, 2015). Within probabilistic frameworks, implicit prediction error has been linked to other language-related components such as the N400 ERP (Kuperberg, 2013; Xiang & Kuperberg, 2014; Kuperberg, 2016), as well as non-linguistic ERP components (e.g., Friston, 2005; Wagongne, Changeux, & Dehane, 2005). Moreover, the N400 ERP component has recently been simulated as cross-entropy error at a semantic level within a connectionist model (Rabovsky & McRae, 2014).

The ERP at time t is computed as shown in Equation 5; m_i and u_i are each unit's modulator and field activation respectively:

$$ERP_t = \sum_i |m_i - u_i| \tag{5}$$

Modeling Results

The words from the test input were presented in the following order: "soup", "dog", "dog", "dog", "beach", "dog", "bog", "tend". The neural field was trained on "soup", "dog", and "beach"; "bog" and "tend" were novel stimuli the field was not trained on. Figure 3 shows that the model replicates the repetition effects, i.e., the maximum ERP values at a t after the first exposure of the word "dog" are all smaller than

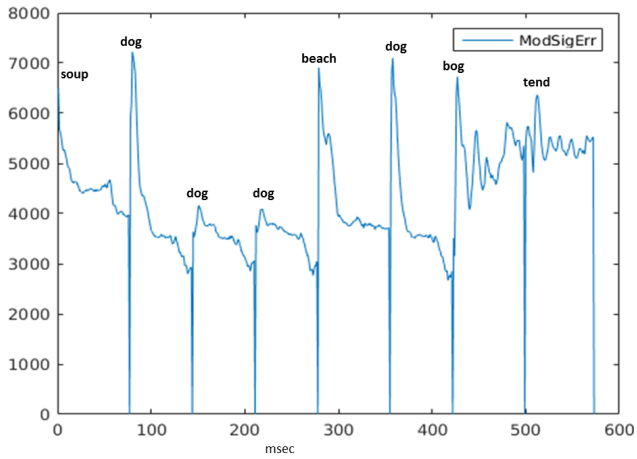


Figure 3: Modulator-Field difference for repetitions of the word "dog"

the first peak, until a different word is presented. At this time, the neural field is perturbed into a different state, releasing it from the effect. A subsequent presentation of "dog" no longer elicits a repetition effect, producing a larger peak as the field resettles into the equilibrium state for "dog". In Equation 5, the modulator signal, m_i, can be thought to "predict" the next equilibrium state the neural field u_i is likely to settle to. This suggests that a smaller amount of perturbation is required to "nudge" the settled field into a new equilibrium state upon presentation of a repeated word. The presentation of untrained, novel stimulus, i.e., "bog" and "tend", does not show the repetition effect as these words are not predicted by the modulator signal.

Table 1: Model Fitting

Interval Width	Model Proportion	ERP Data Proportion
100 ms	1.41	1.60
112 ms	1.53	1.53
120 ms	1.66	1.58
Best fit (112 ms) 144 ms - 256 ms		

We note that model does not aim to fit the polarity of the P2 ERP as what gives rise to the polarity is not well-understood and as there have been inconsistent reports on the word repetition effect as mentioned earlier (van Petten et al., 1991). Furthermore, it is the nature of ERP measurement that the interval within which a given effect is manifested varies somewhat between experimental paradigms. However, the model should fit the magnitude and the duration of the human ERP data. Thus, to compute the model fit, we looked at the ERP data intervals centered around 200 ms as this interval contains the P2 effect and computed the proportion as follows. We took the area under the ERP curve within an interval for the first presentation of the word "dog" and divided it by the identical interval contained under the repeated presentation to calculate its proportion. Referring to Figure 3, we also took the area under the ERP curve generated by the model and performed the same calculation. As shown in Table 1 we found that the 112 ms interval around 200 ms (i.e., from 145 ms to 255 ms) showed both proportions to be identical i.e., 1.53, thus demonstrating it is possible to find a good model fit to the experimental data.

Discussion

We designed our model to be a single neural field reflecting processing in the auditory cortex and hypothesized that this forms a "layer" of phonological processing. In order to provide a modulator signal, we simulated the existence of a deeper word-form layer by "clamping" the modulator signal to the three words the model was trained on (i.e., "beach", "dog", "soup") and this was fed "down" to the neural field as its modulator signal. We did not presuppose which ERP correlates would occur using only one neural field layer and did not set as a goal to identify all possible auditory effects; we were not concerned with capturing non-speech auditory processing at all.

The model succeeded in capturing the repetition effect noted in the experimental results as can be seen in Figure 1, most notably in the central scalp ERPs e.g., Cz. Figure 3 shows a diminished response to the initial presentation of the word "dog" at 75 ms with the repetition effect occurring at 150 ms and 225 ms. Note that the typical convention is to plot the ERP, with the area above the x-axis as negative and the area below as positive. Thus the model and ERP waveforms covary in amplitude and polarity with the repetition (i.e., in the model the repetition effect is "more negative" than the initial presentation).

The model demonstrated the immediate word repetition effect using a single neural field sheet, without modulator input from deeper lexical and semantic processing layers. This suggests that the ability of a single neural field layer to learn sound patterns (i.e., phonemes, monosyllabic words) alone appears to be sufficient to account for the immediate word repetition effect and the release from repetition. We believe this to be among the first computational models to match the time course of ERP events on real-world, real-time data, and the first model to do so using spoken word perception i.e., we used the same data that was presented to the experiment's participants and validated the model fit. These results suggest that our neural field approach can now be used to build additional layers and thus model later ERPs.

Conclusion

We have developed a dynamic neural field model of phonological processing of monosyllabic spoken words and compared it with a separately designed experiment which measured ERP responses of participants to spoken words. We

found a good fit between the model and the human ERP data. The model succeeded at replicating the word repetition effect showing a positive correlation with the experiment's P2 measurements. This suggests that a minimal neural field model can perform some components of auditory processing (e.g., detect immediate word repetition) and generate a correlated ERP effect. Future work will explore modeling deeper lexical and semantic processing and related mid-to-late ERP effects by connecting additional neural field layers in a hierarchy which will allow feedback from the deeper processes to affect computations at earlier layers.

Acknowledgments

Research was sponsored by the U.S. Army Natick Soldier Research, Development, and Engineering Center and was accomplished under Cooperative Agreement Number W911QY-15-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Natick Soldier Research, Development, and Engineering Center, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics*, *27*(2), 77–87.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Science*, *4*(3), 91–99.

Brady, M. (2014). A bi-directional graphical model for babble-feedback learning in speech. *Procedia Computer Science*, *41*, 220–225.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society London, Series B, Biological Sciences*, *360*(1456), 815–836.

Grossberg, S. (2005). Adaptive resonance theory. In L. Nadel (Ed.), *The encyclopedia of cognitive science* (1st ed.). Wiley.

Hillyard, S., & Anllo-Vento, L. (1998). Event-related brain potentials in the study of visual selective attention. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(3), 781–787.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the national academy of sciences* (Vol. 79, pp. 2554–2558).

Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprension. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling the behavioral, neurobiological, and genetic components of reading comprehension* (pp. 176 – 192). Paul Brookes Publishing.

Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1130233

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1102299

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: The MIT Press.

McClelland, J., & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

Molfese, D., Key, A. P. F., Maguire, M. J., Dove, G., & Molfese, V. (2005). Event-related potentials (ERPs) in speech perception. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception.* Blackwell Publishing.

Norris, D. (1995). A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive models of speech processing.* Cambridge, MA: MIT Press.

Oldfield, R. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin*, *125*(4), 410–436.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 erp component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*, 68–89.

Rugg, M. (1987). Dissociation of semantic priming, word and non-word repetition effects by event-related potentials. *The Quarterly Journal of Experimental Psychology*, *39*(1), 123–148.

Schöner, G., & Spencer, J. (2015). *Dynamic thinking: a primer on dynamic field theory*. New York, NY: Oxford University Press.

van Petten, C., & Kutas, M. (1991). Electrophysiological evidence for the flexibility of lexical processing. In G. Simpson (Ed.), *Understanding word and sentence.* Amsterdam: North-Holland Press.

van Petten, C., Kutas, M., Kluender, R., Mitchnier, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience*, *3*(2), 131–150.

Wagongne, C., Changeux, J.-P., & Dehane, S. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815 – 836.

Xiang, M., & Kuperberg, G. R. (2014). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2014.995679

Talking with tact: Polite language as a balance between kindness and informativity

Erica J. Yoon*, **Michael Henry Tessler***, **Noah D. Goodman**, and **Michael C. Frank**
{ejyoon, mtessler, ngoodman, mcfrank} @stanford.edu
Department of Psychology, Stanford University
*Authors contributed equally to this work.

Abstract

Conveying information in a false or indirect manner in consideration of listeners' wants (i.e. being polite) seemingly contradicts an important goal of a cooperative speaker: information transfer. We propose that a cooperative speaker considers both *epistemic utility*, or utility of providing the listener new and accurate information, and *social utility*, or utility of maintaining or boosting the listener's self-image (being polite). We formalize this tradeoff within a probabilistic model of language understanding and test it with empirical data on people's inferences about the relation between a speaker's goals, utterances and the true states of the world.

Keywords: Politeness; computational modeling; communicative goals; pragmatics

Introduction

Your friend gives a terrible presentation and asks for your opinion. Must you admit, "Your talk was terrible," or is it acceptable to say: "Your talk was fine"? The latter is potentially misleading but gives the listener what she might want to hear—-in other words, it would be polite.

Politeness violates a critical principle of cooperative communication: exchanging information efficiently and accurately (Grice, 1975). If information transfer was the only currency in communication, a cooperative speaker would find polite utterances undesirable because they are potentially misleading. People are polite, however, and speakers do produce polite utterances. Adults and even young children spontaneously produce requests in polite forms (Clark & Schunk, 1980; Axia & Baroni, 1985). Speakers exhibit politeness strategies even while arguing, preventing unnecessary offense to their interactants (Holtgraves, 1997). Listeners even attribute ambiguous speech to a polite desire to hide a truth that could hurt another's self-image (e.g. Bonnefon, Feeney, & Villejoubert, 2009). In fact, it is difficult to imagine human speech that efficiently conveys only the truth. Intuitively, politeness is one prominent characteristic that differentiates human speech from stereotyped robotic communication, which may try to follow rules to say "please" or "thank you" yet still lack genuine *politeness*.

Does this mean people are not cooperative communicators? Brown and Levinson (1987) recast the notion of a *cooperative speaker* as one who has both an epistemic goal to improve the listener's knowledge state as well as a social goal to minimize any potential damage to the hearer's (and the speaker's own) self-image, which they called *face*. In their analysis, if the speaker's intended meaning contains no threat to the speaker or listener's face, then the speaker will choose to convey the meaning in an efficient manner, putting it *on the record*. As the degree of face-threat becomes more severe, however, a speaker will choose to be polite by producing more indirect utterances.

In the current paper, we formalize a version of Brown and Levinson (1987)'s theory, exploring the idea that cooperative speakers attempt to balance two goals, epistemic and social. The Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) describes language understanding as recursive probabilistic inference between a pragmatic listener and an informative speaker. This framework has been successful at capturing the quantitative details of a number of language understanding tasks, but it neglects the social goals a speaker may pursue. Here we extend RSA to take into account a speaker with both the usual epistemic goal and a competing social goal: be kind. We test this model by gathering data about utterance interpretations and goal attributions in settings where the true state of the world carries affective consequences for the hearer.

Computational Model

Politeness poses a challenge for formal models of pragmatic language understanding, which assume that speakers' goals are to communicate informatively about some aspect of the world (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). RSA models a listener as reasoning about a speaker, who chooses utterances approximately optimally given a utility function. Goodman and Stuhlmüller (2013) define speaker utility by the amount of information a *literal listener* would still not know about world state s after hearing a speaker's utterance w (i.e. *surprisal*), what we will call *epistemic utility*: $U_{epistemic}(w;s) = \ln(P_{L_0}(s \mid w))$, where the literal listener is a simple Bayesian agent that takes the utterance to be true:

$$P_{L_0}(s \mid w) \propto [\![w]\!](s) \cdot P(s). \quad (1)$$

Here, $[\![w]\!](s)$ is the truth-functional denotation of the utterance w (i.e. the utterance's literal meaning): It is a function that maps world-states s to Boolean truth values. The literal meaning is used to update the literal listener's prior beliefs over world states $P(s)$.

We propose there is a second component to the speaker's utility related to the intrinsic value of the state in the eyes of the listener[1], what we will call *social utility*. We define the social utility of an utterance to be the expected utility of the state the listener would infer given the utterance w:

$$U_{social}(w;s) = \mathbb{E}_{P_{L_0}(s|w)}[V(s)],$$

[1] At this point, we do not differentiate state value to the listener from state value to the speaker, though in many situations these could in principle be different.

where V is a value function that maps states to subjective utility values—this captures the affective consequences for the listener of being in state s. We take the overall speaker utility to be a weighted combination of epistemic and social utilities:

$$U(w;s;\hat{\beta}) = \beta_{epistemic} \cdot U_{epistemic} + \beta_{social} \cdot U_{social}.$$

The speaker chooses utterances w softmax-optimally (determined by *speaker rationality* parameter λ) given the state s and his goal weights $\hat{\beta}$:

$$P_{S_1}(w \mid s, \hat{\beta}) \propto \exp(\lambda \cdot \mathbb{E}[U(w;s;\hat{\beta})]) \quad (2)$$

The pragmatic listener, denoted L_1, infers the world state based on this speaker model. We will assume the listener does not know exactly how the speaker weights his competing goals, however. Following the treatment of RSA using lifted variables (Goodman & Lassiter, 2015; Bergen, Levy, & Goodman, in press; Kao, Wu, Bergen, & Goodman, 2014), we assume the pragmatic listener jointly infers the state s and the utility weights of the speaker, $\beta_{epistemic}$ and β_{social}:

$$P_{L_1}(s, \hat{\beta} \mid w) \propto P_{S_1}(w \mid s, \hat{\beta}) \cdot P(s) \cdot P(\hat{\beta}) \quad (3)$$

Within our experimental domain, shown in Figure 2 and described in more detail below, we assume there are five possible states of the world corresponding to the value placed on a particular referent (e.g. the presentation the speaker is commenting on): $S = \{s_1, ..., s_5\}$. We further assume a uniform prior distribution over possible states of the world. The states have subjective numerical values $V(s_i) = \alpha \cdot i$, where α is a scaling parameter (later inferred from data). The set of utterances is {*terrible*, *bad*, *okay*, *good*, and *amazing*}. We implemented this model using the probabilisitic programming language WebPPL (Goodman & Stuhlmüller, 2014) and a complete implementation can be found at `http://forestdb.org/models/politeness-cogsci2016.html`.

In what follows, we measure the literal semantics in Experiment 1, then use these to predict performance in two experiments. In Experiment 2, we explore listeners' inferences about the world s given an utterance and a speaker's goal. In Experiment 3, we investigate inferences about speakers' goals given an utterance and a state. Then we compare the behavioral results to our model predictions, that people's inferences are based on a speaker model with two utilities, epistemic and social.

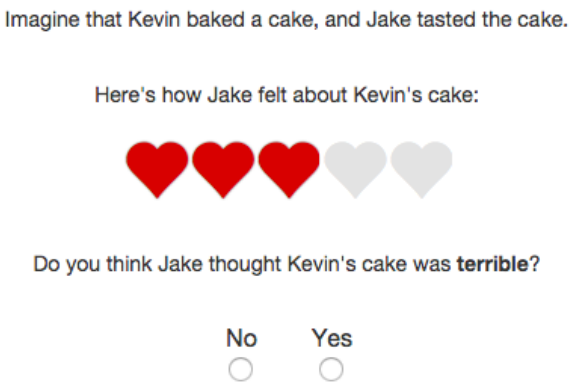


Figure 2: Example of a trial in Experiment 1.

Experiment 1: Literal semantics

Experiment 1 measured judgments of literal meanings of our target words. Responses in this experiment will be used to set expected literal meanings of words in our formal model.

Method

Participants 30 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Stimuli and Design We created 13 different context items, in which someone (e.g. Ann) gave a performance of some kind (e.g. gave a talk, performed a cello solo, baked a cookie, etc.), and another person (e.g. Bob) evaluated it. For example, in one of the contexts, Ann baked a cake, and Bob's true feelings toward Ann's cake ("*true state*") were shown on a scale out of five hearts. The question of interest was "Do you think Bob thought Ann's cake was [*terrible*, *bad*, *okay*, *good*, or *amazing*]?" Each participant read 25 scenarios (5 true states and 5 words). The order of context items was randomized.

Procedure Participants read scenarios and indicated their answer to each question by answering 'No' or 'Yes' (see Figure 2 for a screenshot of an example trial). The experiment

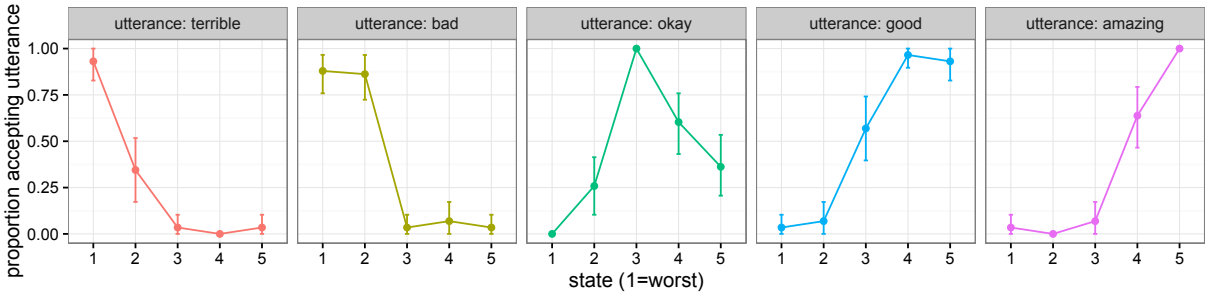


Figure 1: Results from Experiment 1. Proportion of acceptances of words (shown in different colors) given the true state represented on a scale of hearts. Error bars represent 95% confidence intervals.

can be viewed at: `http://langcog.stanford.edu/expts/EJY/polgrice/L2_J/polgrice_L2_J.html`.

Results In this and all subsequent experiments, we analyze the data by collapsing across contexts. Meanings of the words as judged by participants were as one would expect (see Figure 1). Proportion of acceptances for a word given the true state peaked where the degree of positivity, neutrality and negativity of the state matched that of the word. The fraction of participants that endorsed utterance w for state s will be used as the literal meaning $[\![w]\!](s)$ in Eq. 1.

Experiment 2: True state inference

In Experiment 2, we examined listeners' inferences about the likely state of the world s given a speaker's utterance (e.g. *"It was good"*) and a description of the speaker's intentions (e.g. the speaker wanted to be *nice*).

Method

Participants 35 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Stimuli and Design We designed scenarios in which a person (e.g. Ann) asked for another person (e.g. Bob)'s opinion on her performance. The same context items and true states as Experiment 1 were used. Additionally, we provided information on Bob's goal (to be *honest*, *nice*, or *mean*) and what Bob actually said to Ann (e.g. "It [your cake] was okay"), where Bob used one of the five possible words: *terrible*, *bad*, *okay*, *good*, or *amazing*. Then we asked participants to infer the true state of the world (e.g. how Bob actually felt about Ann's cake). Each participant read 15 scenarios (3 goals and 5 words). The order of context items was randomized.

Procedure Participants read each story (e.g. Ann baked a cake and asked Bob about it) followed by a prompt that said, e.g., "Bob wanted to be nice: "It was okay," he said. How do you think Bob actually felt about Ann's cake?" Participants indicated their answer on a scale of five hearts. The experiment can be viewed at: `http://langcog.stanford.edu/expts/EJY/polgrice/L2_S/polgrice_L2_S.html`.

Behavioral results

Inferences of the actual rating for hearer's performance, or the true state, varied depending on speaker's goal and utterance (Figure 3). Consistent with intuition, when the speaker was trying to be honest, utterances accurately mapped onto inferred states (Figure 3, left, red line). For utterances consistent with the speaker's goals (e.g. saying positive utterances when the speaker was trying to be nice; negative utterances when trying to be mean), the results are also consistent with intuition. Knowing that the speaker was trying to be nice, participants inferred a true state appreciably lower than the state inferred given honesty (green line). The reverse was true when the speaker was trying to be mean (blue).

For utterances inconsistent with the speaker's goals, we observe an interesting asymmetry. When the speaker was trying to be nice and said a negative utterance (e.g. "it was bad"), participants inferred that the true state really was bad (no difference between an *honest* "bad" and a *nice* "bad"), perhaps because of a floor effect or owing to the fact that one can be nice by being honest. When the speaker was trying to be mean, however, and said a positive utterance (e.g. "it was amazing"), participants inferred a state that is *worse* compared to the states based on honesty and niceness goal. This is likely due to participants attributing sarcasm or irony to the speaker who was trying to be mean.

Model predictions

Model fitting In this experiment, participants were told the speaker said w and described what the speakers' intentions were (e.g. *Bob wanted to be nice*). We want to explore inferences both when speaker wanted to be nice and when he

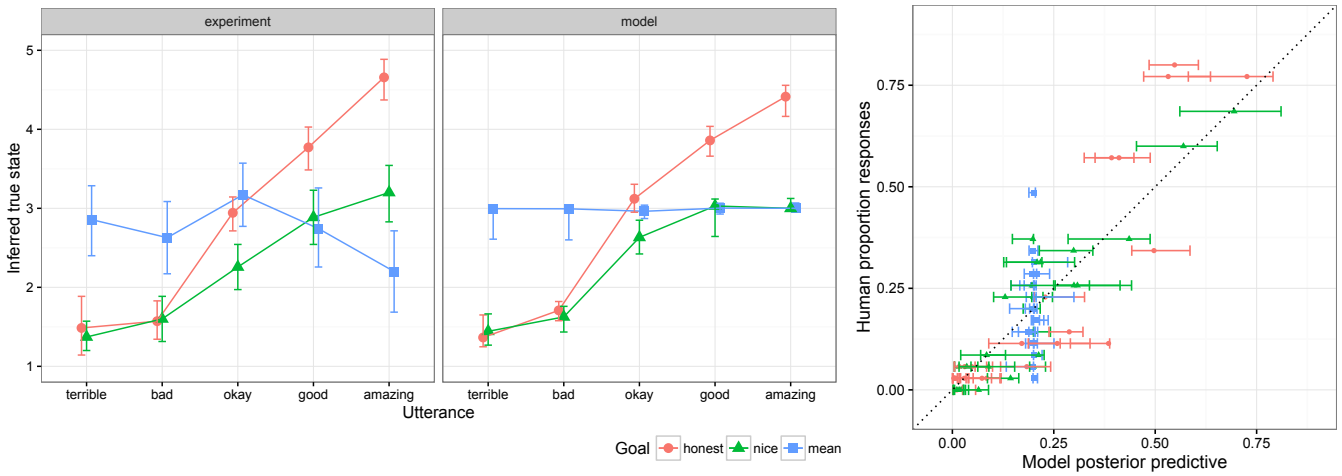


Figure 3: Results from Expt. 2 (left) and model predictions (center) for average states inferred based on a speaker's goal and utterance. Right: Full distribution of human responses vs. model predictions. Error bars represent 95% confidence intervals for the data and 95% highest density intervals for the model.

wanted to be mean: For ease of comparison to the experimental data, we assume each $\beta \in [0,1]$ and separate β_{social} into a positive component (β_{nice}) and a negative component (β_{mean}), writing the utility function now as

$$U(w;s;\beta) = \beta_{epistemic} \cdot U_{epistemic} + (\beta_{nice} - \beta_{mean}) \cdot U_{social}$$

We assume that the intentions (e.g. *wanted to be nice*) notified the listener of a particular set of goal-weights $\{\beta_{nice}, \beta_{honest}, \beta_{mean}\}$ that the speaker was using. We put uninformative priors on these weights ($\beta \sim \text{Uniform}(0,1)$) and infer their credible values separately for each goal condition ("trying to be X") using Bayesian data analytic techniques (Lee & Wagenmakers, 2014).

There are 2 additional parameters of the cognitive model: the speaker optimality parameter λ in Eq. 2 and the value scale parameter α in the utility function. We put uninformative priors on these ($\lambda \sim \text{Uniform}(0,20)$ and $\alpha \sim \text{Uniform}(0,5)$) and infer their posterior credible values from the data. We ran 2 MCMC chains for 100,000 iterations, discarding the first 50,000 for burnin. The Maximum A-Posteriori (MAP) estimate and 95% Highest Probability Density Interval (HDI) for λ is 1.4 [0.9, 8.5]; for α is 6.4 [1.9, 9.9]. To generate predictions, given our cognitive model and the inferred parameters, we evaluated the posterior predictive distribution, marginalizing out all parameters.

Results The inferred weights for each goal condition were largely as expected (Figure 4). For the "trying to be honest" condition, the model infers the speaker was using a non-zero weight on honesty, while the overall social weight is zero (since niceness and meanness are inferred to be the same values and they inverses of each other). For "trying to be nice", the model puts a high weight on niceness but also some appreciable weight on honesty. The model fits are worse for the "trying to be mean" case; all that the model infers is that the speaker was not honest.

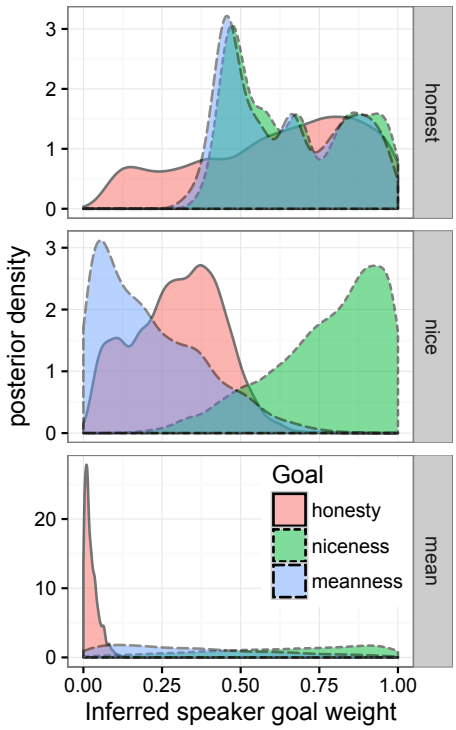


Figure 4: Inferred goal weights β for Expt. 2. Facets are different experimental conditions (trying to be X). Density plots show likely weights used in the speaker's utility function.

The predictions of the expectations of the listener model for the *true state* (Eq. 3) are shown in Figure 3 (center). The model's expected posterior over states when the speaker is trying to be *honest* increases as a function of the positivity implied by the utterances (e.g. "amazing" means a very high state). When it knows the speaker was trying to be *nice*, the pragmatic listener is more conservative in how it interprets positive utterances, with the difference between an honest utterance and a nice utterance increasing as the utterance becomes more positive. For example, the difference between a *nice* "amazing" and an *honest* "amazing" is greater than the difference between a *nice* "okay" and an *honest* "okay". Inferences when the speaker is trying to be *mean* display the associated opposite behavior when the utterance is indeed negative (e.g. "terrible"). Overall, the expected values of the model explain a lot of the variance in the average data $r^2(15) = 0.91$. The main discrepancies are with the goal "trying to be mean", as noted above. Among the other 2 goal conditions, the model explains almost all of the variance $r^2(10) = 0.97$.

The model makes predictions not only for the average data but also for the full distribution of responses (Figure 3, right). The model predicts the full distribution with relatively high accuracy $r^2(75) = 0.74$.

In sum, the model captured key aspects of our empirical findings. The largest discrepancies appear in the experimental condition of the goal to be mean. Participants thought that a mean speaker saying "[your cake] was amazing" meant the true state was below average, which the model was unable to accommodate. This deviation is likely due to the effect of irony: making an extremely positive remark about an extremely bad performance is perceived to be sarcastic and ill-intentioned (Colston, 1997). Our model does not include sarcastic interpretation though other models in the RSA family do (Kao & Goodman, 2015), and future work should address the delicate interplay of politeness and sarcasm.

Experiment 3: Goal inference

Experiment 3 probed listeners' inferences of the speaker's goals, given an utterance (e.g. *"It was good"*) and a true state (e.g. 2 out of 5 hearts).

Method

Participants 45 participants with IP addresses in the United States were recruited on Amazon's Mechanical Turk.

Stimuli and Design We presented the same context items and utterances as Experiment 2. But instead of goals, we provided information on the true states (i.e. how Bob actually

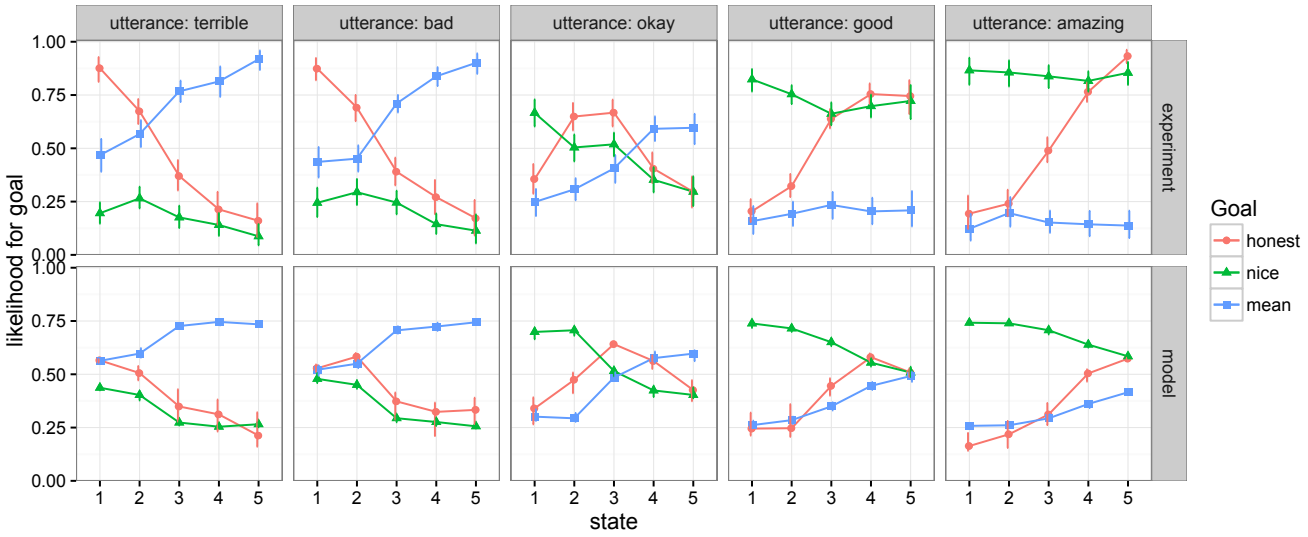


Figure 5: Results from Expt. 3 (top) and model predictions (bottom). Attribution of speaker's goals: honest, nice and mean (colors) based on the true state and utterance. Error bars represent 95% CIs for the data and 95% HDIs for the model.

felt towards Ann's performance). Then we asked participants to infer the likelihood of Bob's goals to be *honest*, *nice*, and *mean*. Each participant read 25 scenarios (5 true states and 5 words). The order of context items was randomized.

Procedure Participants read each scenario followed by a question that read, "Based on what Bob said, how likely do you think that Bob's goal was to be: honest; nice; mean," with the three goals placed in a random order below three slider bars, on which the participant could indicate each goal's likelihood. The experiment can be viewed at: `http://langcog.stanford.edu/expts/EJY/polgrice/L2_G/polgrice_L2_G.html`.

Behavioral results

Participants rated speaker's goals differentially depending on the true state and utterance (see Figure 5, top). Honesty was rated highest when the true state was most consistent with the literal semantics. As utterances became more positive, participants rated the speaker's niceness higher and meanness displayed the reverse pattern. Ratings for niceness and meanness goals were strongly anti-correlated $r = -.78$.

Interestingly, we observe an asymmetry in how positive and negative utterances map onto the goals of niceness and meanness, respectively. Participants reported that *truthfully* saying "amazing" is both honest and nice, while *truthfully* saying something is "terrible" is honest and not that mean. Here, meanness can decrease in likelihood (i.e. be *explained away*) because of honesty. And yet honesty does not explain away niceness when the speaker says a truthful "amazing".

Model predictions

Model fitting The model in Eq. 3 specifies a joint-belief distribution over the speaker's goals β and possible states of the world s. To compare to the empirical data, we condition on the true state of the world given in the experimental condition, and compare the marginal distribution of the speaker's goals β to the empirical ratings. We separate β_{social} into β_{nice} and β_{mean}, as in Expt. 2. With no prior knowledge about the speaker's intentions, we assume a uniform prior over the goal-weights in the speaker's utility function: $\beta \sim \text{Uniform}(0,1)$.

We put the same priors over the speaker optimality parameter λ and the value scale parameter α. We ran 2 MCMC chains for 40,000 iterations, discarding the first 20,000 for burnin. The MAP estimate and 95% HDI for λ is 3.8 [3.1, 4.9]; for α is 1.4 [1.23, 1.53].

Results The predictions of the model are shown in Figure 5 (bottom). Like our participants, the model believes the speaker to be more honest when the utterance matches the true state of the world, as given by the literal semantics data (Figure 5, red lines). Further, the model increases its ratings of niceness as the utterance better matches states with higher values (green lines, main effect of panel). The goal to be mean displays the opposite behavior, increasing as the utterance matches states with lower values (blue lines). Overall, the model displays a strong quantitative fit to the goal inference data $r^2(75) = 0.82$.

The model successfully captured the key patterns in the data: changing goal likelihoods based on the degree of match, and positivity/negativity bias of mismatch, between the utterance and true state. The biggest mismatch seems to be that the model under-predicts the goal to be honest for every utterance except *okay*. This may be because explicitly positive/negative utterances can also be explained by a social goal in the model. The model also does not capture the interaction that a truthful positive utterance is both honest and nice, while a truthful negative utterance is honest and not that mean. The model treats niceness and meanness as perfectly opposite to

each other, while there is an interesting asymmetry in participants' behavior. This asymmetry may be because listeners expect honesty and niceness to be correlated *a priori*, and anti-correlated with meanness, rather than independent as assumed by the model.

Discussion

Why would a speaker ever say something that is not maximally truthful and informative? Communication is often examined from the perspective of successful information transfer from speaker to listener. In the social realm, however, communication also can serve the social function of making the listener feel good and saving her face.

We proposed here that intuitively "polite" utterances arise from the desire to be kind (i.e. save face). A cooperative speaker then tries to balance the goals to be kind and to be informative, and produces utterances of varying degrees of politeness that reflect this balance. To test this proposal, we examined inferential judgments on a speaker's utterance, which was a potentially face-threatening evaluation of the listener's performance. As we predicted, participants' inferences about the true state of the world differed based on what the speaker said and whether the speaker's intended goal was to be honest, nice or mean (Expt. 2). We were also able to predict participants' attributions of different social goals to speakers depending on how well the literal utterance meaning matched the actual rating the performance deserved (Expt. 3).

The model presented here relates to other work done in game-theoretic pragmatics. Van Rooy (2003) uses a game-theoretic analysis of polite requests ("Could you possibly take me home?") to argue the purpose of polite language is to align the preferences of interlocutors. Our notion of social utility U_{social} is similar in that it motivates speakers to signal worlds that make the listener feel good. Van Rooy's analysis, however, relies on the notion that polite language is *costly* (in a social way e.g., by reducing one's social status or incurring social debt to one's conversational partner) but it's not clear how the polite behaviors explored in our experiments (not polite requests) would incur any cost to speaker or listener. Our model derives its predictions by construing the speaker utility as a collection of possible goals (here, epistemic and social goals). The speech-acts themselves are not costly.

Will machines ever be polite? Politeness requires more than merely saying conventionalized words (*please*, *thank you*) at the right moments; it requires a balance of informativity and kindness. Politeness is not an exception to rational communication; it is one important element of rational communication, serving a key social function of maintaining relationships. We extended the Rational Speech Acts framework to include social utility as a motive for utterance production. This work takes a concrete step toward quantitative models of the nuances of polite speech. And it moves us closer to courteous computation—to computers that communicate with tact.

Acknowledgments

This work was supported by NSF grant BCS #1456077 to MCF, ONR grant N00014-13-1-0788 and a James S. McDonnell Foundation Scholar Award to NDG, NSF Graduate Research Fellowship DGE-114747 to MHT, and NSERC postgraduate doctoral scholarship PGSD3-454094-2014 to EJY.

References

Axia, G., & Baroni, M. R. (1985). Linguistic politeness at different age levels. *Child Development*, 918–927.

Bergen, L., Levy, R., & Goodman, N. D. (in press). Pragmatic reasoning through semantic inference. *Semantics & Pragmatics*.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge Univ. Press.

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, *8*(2), 111–143.

Colston, H. L. (1997). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, *23*(1), 25–45.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory, 2nd edition.* Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`.

Grice, H. P. (1975). Logic and conversation. In *Readings in language and mind.* Blackwell.

Holtgraves, T. (1997). Yes, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology*, *16*(2), 222–239.

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 37th annual conference of the Cognitive Science Society.*

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge Univ. Press.

Van Rooy, R. (2003). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In *Proceedings of the 9th conference on theoretical aspects of rationality and knowledge* (pp. 45–58).

Understanding "almost": Empirical and computational studies of near misses

Tobias Gerstenberg (tger@mit.edu) & Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

When did something almost happen? In this paper, we investigate what brings counterfactual worlds close. In Experiments 1 and 2, we find that participants' judgments about whether something almost happened are determined by the causal proximity of the alternative outcome. Something almost happened, when a small perturbation to the relevant causal event would have been sufficient to bring it about. In contrast to previous work that has argued that prior expectations are neglected when judging the closeness of counterfactual worlds (Kahneman & Varey, 1990), we show in Experiment 3 that participants are more likely to say something almost happened when they did not expect it. Both prior expectations and causal distance influence judgments of "almost". In Experiment 4, we show how both causal proximity and beliefs about what would have happened in the absence of the cause jointly explain judgments of "almost caused" and "almost prevented".

Keywords: causality; counterfactuals; almost; mental simulation; intuitive physics.

> "*We have come to think of the actual as one among many possible worlds. We need to repaint that picture. All possible worlds lie within the actual one.*" – Goodman (1983, p. 57)

Contemplating what could or should have been is a pervasive feature of human cognition (Kahneman & Miller, 1986; Reid, 1986; Roese, 1997; Teigen, 1996). As the quote above suggests, our actual world is infused with different possible worlds. Despite the appeal of dozens of self-help books to live in the "now", we often cannot help but dwell on how the "now" could have turned out differently (Johnson, 1986). The word "almost" provides a window into counterfactual thoughts (cf. Kahneman & Varey, 1990). Actual and counterfactual worlds can come dangerously close, such as when a researcher almost missed her plane, or a dentist almost drilled the wrong tooth. The goal of this paper is to better understand when it is that something almost happened. By investigating what factors influence whether people agree with the statement that something almost happened, we learn more about people's concept of "almost" specifically, as well as about counterfactual thinking more generally.

A linguistic analysis of the word "almost" yields two key results (cf. Nouwen, 2006; Penka, 2006): first, *almost q* implies that *q* did not actually happen and, second, that the actual outcome *p* was close to *q* on some dimension, such as its probability of occurrence. For *q* to have almost happened, its probability must now be zero, but its (subjective) probability at some earlier stage must have been high. In their investigation of close counterfactuals, Kahneman and Varey (1990) distinguish two different kinds of probabilities: *dispositions* and *propensities*. A disposition is the prior probability of the outcome before the relevant episode started to unfold. A reliable penalty taker has a strong disposition to score a goal. Propensities express dynamic updates to the outcome probability as the episode unfolds. When the penalty taker's ball is headed toward the post, it now has the propensity to miss the goal.

When judging propensity we need to assess the causal proximity of different possible outcomes. One way of doing so is through the process of mental simulation (Kahneman & Tversky, 1982). In the penalty example above, we mentally extrapolate the trajectory of the ball to estimate whether it is going to hit or miss (cf. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014). The outcome was close if a small perturbation to the relevant causal event would have been sufficient to bring about the alternative outcome (cf. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015). If the penalty taker had kicked the ball slightly differently, then the ball would have missed the goal. We investigate the influence of causal as well as spatial proximity on people's judgments about whether something almost happened in Experiment 1.

Kahneman and Varey (1990) argued that whether something almost happened is more strongly influenced by propensity than by disposition. In fact, dispositions are thought to be neglected entirely when they are not supported by the propensities of the situation. Accordingly, whether the penalty taker is normally good or bad should not make much of a difference for whether we think that the player almost missed. What matters is how close the ball came to the post.

We believe that there is more to "almost" than propensity, and that prior expectations matter. Imagine the following scenario: The first author plays a one-on-one basketball game against Michael Jordan. Whoever scores 11 points wins the game. Let's say that the outcome of Game 1 is 11 (Jordan) to 10 (Gerstenberg). Here, it would sound fine to say that "Gerstenberg almost beat Jordan". The outcome of Game 2 is 11 (Gerstenberg) to 10 (Jordan). Now, it would sound strange to say that "Jordan almost beat Gerstenberg" even though he came very close to doing so. We would be more inclined to say that (wow!) "Jordan lost to Gerstenberg".

The asymmetry between these two situations is difficult to accommodate by considering causal proximity only. If anything, the counterfactual world in which Gerstenberg would have won in Game 1 seems further away from what actually happened than the counterfactual world in which Jordan would have won in Game 2. Intuitively, what seems to be driving the difference between these two situations is our prior expectation: we expected that Jordan would beat Gerstenberg. We look at the influence of prior expectations on people's judgments of "almost" in Experiments 2 and 3.

In Experiment 4, we examine in what situations people say that something "almost caused" an outcome to happen or "almost prevented" it from happening. In previous work (Gerstenberg et al., 2012, 2014, 2015), we have developed the *counterfactual simulation model* of causal judgment which

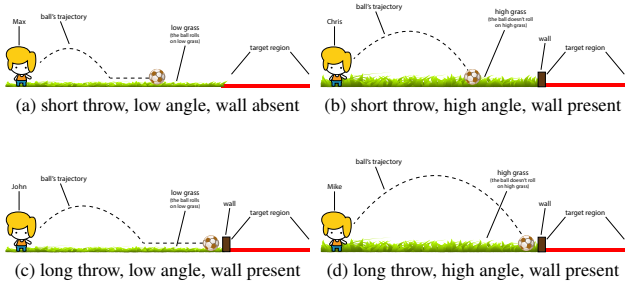
predicts that people make causal judgments by comparing what actually happened with what would have happened if the candidate cause had been absent or different. Here, we apply and extend this model to capture people's judgments about "almost caused".

Experiment 1: Causal distance

For an outcome to have almost happened it has to be close to what actually happened. But close in what sense? Consider the scenarios depicted in Figure 1. In each scenario, a boy tries to toss a ball far enough to reach a target region. In some of the situations, the spatial distance to the target region is large, and sometimes it is small. If spatial distance matters, then participants should be more likely to say that the character's ball almost reached the target region in situations c) and d) than in situations a) and b).

The situations also differ in whether the character throws the ball at a low or high angle, and whether there is a wall in front of the target region. If participants only cared about the spatial distance between the ball's final location and the target region then these aspects should not matter. However, if they care about causal distance, these aspects should matter. Recall that to estimate causal distance, we imagine what perturbation to the relevant causal event would have been sufficient to generate the counterfactual outcome (cf. Gerstenberg et al., 2015). More precisely, we can quantify the distance between the actual and counterfactual outcome by looking at how much more force the character's throw would have required to be to get the ball into the target region (Stocker, 2014; Wolff, 2007).

Consider situations c) and d). In both situations the ball comes to a stop at the exact same location. Thus, spatial distance to the target is identical. However, the two situations differ in terms of the causal distance to the counterfactual outcome. In situation d) the boy's ball could have reached the target region if he had thrown the ball just a little bit harder. In contrast, in situation c) the boy's throw would have needed to be substantially harder for the ball to reach the target region. Since the causal distance to the alternative outcome is closer in d) than it is in c), we predict that participants are more likely to say that the boy's ball almost reached the target region in d) than in c). Note that the difference between the two situations disappears if we remove the wall in front of the target region. Now both situations are similar in their causal distance to the counterfactual outcome – the target region would have been reached if each character had just tossed the ball a little harder. Thus, we expect participants' judgments to be similarly high for these two situations.



(a) short throw, low angle, wall absent (b) short throw, high angle, wall present

(c) long throw, low angle, wall present (d) long throw, high angle, wall present

Figure 1: Example of stimuli presented in Experiments 1 and 2. Note that the character throws the ball at a high angle when then grass is high and the ball does not roll (b and d), and at a low angle when the grass is low and the ball rolls (a and c).

Methods

Participants Participants in each experiment of this paper were recruited via Amazon's Mechanical Turk and compensated at a rate of $6 per hour. No participant took part in more than one experiment. 40 participants ($M_{age} = 33.8$, $SD_{age} = 11.6$, $N_{female} = 14$) completed Experiment 1.

Design We manipulated three factors within participants: (1) How far the character threw the ball (*throw*: short vs. long), (2) at what angle the ball was thrown (*angle*: high vs. low), and (3) whether there was a wall in front of the target region (*wall*: present vs. absent). Figure 1 shows four examples of the stimuli shown to participants.

Procedure After having received instructions, participants saw eight the different situations presented in randomized order. Participants were asked to what extent they agreed with the sentence that "Max's ball almost reached the target region." They indicated their response on a sliding scale with the endpoints labeled "I completely disagree" (-50) and "I completely agree" (50). A different character name was used on each trial. It took participants 2.4 minutes ($SD = 1.8$) on average to complete the experiment.

Results and Discussion

Figure 2 shows participants' mean agreement ratings for the eight different situations. There was a main effect of *throw*; agreement ratings were higher for longer throws than for shorter throws ($F(1,39) = 132.24, p < .0001, \eta_p^2 = .77$). There was also a main effect of *wall*; ratings were higher when the wall was absent ($F(1,39) = 49.37, p < .0001, \eta_p^2 = .56$). There was no main effect of the *angle* in which the ball was thrown. However, there was an interaction between *angle* and *wall* ($F(1,39) = 11.37, p < .001, \eta_p^2 = .23$). The presence of the wall reduced participants' agreement ratings more strongly for the low angle throws (e.g. situation 7 vs. 8 in Figure 2) than for the high angle throws (e.g. situation 5 vs. situation 6). There was also a significant interaction between *throw* and *wall*. The presence of the wall reduced participants' agreement ratings more for long throws rather

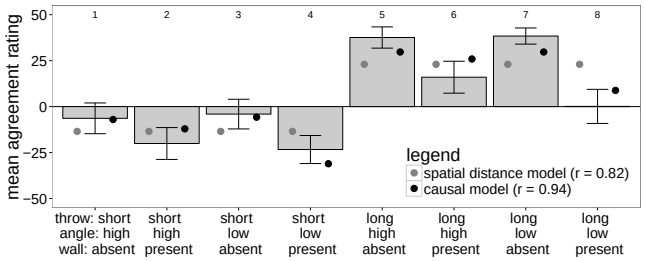


Figure 2: **Experiment 1**: Mean agreement ratings for each of the 8 different stimuli together with the predictions of the spatial model (gray points) and the causal model (black points). *Note*: Error bars in all figures indicate 95% confidence intervals.

than short throws ($F(1,39) = 12.66, p < .001, \eta_p^2 = .25$).

We implemented a causal model by recreating the diagrammatic stimuli with the physics engine Box2D.[1] For each situation, we determined the minimum additional impulse to the throw that would have been required to get the ball into the target region. We contrast the causal model with a spatial distance model that predicts participants' ratings based on the geometrical distance between the final position of the ball and the target region. As Figure 2 shows, participants' mean agreement ratings are better accounted for by the causal model ($r = .94$) than the spatial distance model ($r = .82$). The causal model correctly captures that the presence of the wall makes a bigger difference when the character tossed the ball at a low angle, as the throw would have needed to be substantially harder in these situations to reach the target region. Participants' judgments of whether the boy's ball almost reached the target region are determined by the causal rather than the spatial distance between the actual and the counterfactual outcome.

Experiment 2: Expectations

Experiment 1 showed that people's concept of "almost" implies that the counterfactual outcome was causally close to what actually happened. In Experiment 2, we investigate how prior expectations influence participants' judgments. As argued in the introduction, we expect that in addition to considering causal distance, participants are more likely to say that a counterfactual outcome *almost happened* when they had low prior expectations for that outcome to occur, and that it *didn't happen* when they had high prior expectations.

Methods

Participants 40 participants ($M_{age} = 31.3, SD_{age} = 10.8, N_{female} = 14$) completed this experiment.

Design and Procedure The design was identical to Experiment 1 except for two changes: First, we introduced information about the character's usual performance as an additional factor. For each of the eight different situations used in Experiment 1, participants were either told that the character normally gets his ball in the target region, or does not. The experiment follows a 2 (*throw*: short/long) $\times$ 2 (*angle*: high/low) $\times$ 2 (*wall*: absent/present) $\times$ 2 (*norm*: positive/negative) within-subjects design. Participants saw the 16 trials in randomized order. The information about the character's normal performance ("Normally, Max does / doesn't get his ball into the target region.") was provided in large font underneath the same stimuli images that were used in Experiment 1.

Second, participants were asked to indicate their agreement with two different statements on independent sliding scales. The statements were "This time, Max's ball *almost reached* the target region." and "This time, Max's ball *didn't reach* the target region." The order of the statements was counterbalanced between participants. It took participants 4.6 minutes ($SD = 1.9$) on average to complete the experiment.

[1] Here is a link to the model:
http://web.mit.edu/tger/www/demos/almost/

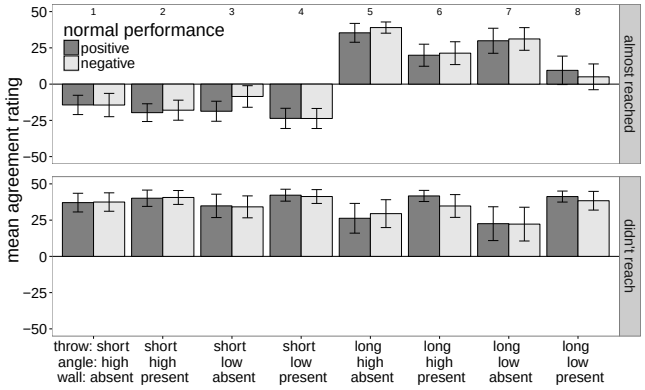


Figure 3: **Experiment 2:** Mean agreement ratings with the "almost reached" sentence (top panel) and the "didn't reach" sentence (bottom panel) as a function of each character's normal performance (bar color) for each of the 8 different types of stimuli (x-axis).

Results and Discussion

Participants' responses to the "almost reached" question closely mirrored the pattern of responses in Experiment 1 (Figure 3). Again, there was a main effect of *throw* ($F(1,39) = 12.17, p < .01, \eta_p^2 = .24$), a main effect of *wall* ($F(1,39) = 10.27, p < .01, \eta_p^2 = .21$), and interaction effects between *throw* and *wall* ($F(1,39) = 6.56, p < .05, \eta_p^2 = .14$), and between *angle* and *wall* ($F(1,39) = 4.69, p < .05, \eta_p^2 = .11$).

In general, participants tended to disagree with the "almost reached" statements in situations in which the throw was short, and tended to agree with the statement when the throw was long. In contrast, participants never disagreed with the "didn't reach" statements. There was a strong negative correlation between participants' agreement ratings with the "almost reached" and "didn't reach " statements ($r = -.72$). The more participants agreed that the character's ball almost reached the target region, the less they agreed that the character's ball didn't reach the target region. In contrast to what we hypothesized, manipulating the character's normal performance did not affect participants' agreement ratings. There was neither a main effect of norm nor any interaction effects involving norm.

Why did the norm manipulation fail to influence participants' judgments? The Jordan vs. Gerstenberg scenario mentioned above strongly suggests that prior expectations matter. We believe that the manipulation may have failed for the following reasons: First, in none of the situations does the character actually succeed in getting his ball into the target region. Participants may have simply ignored the information about normal performance since it was not indicative for how the character actually performed. Second, the norm manipulation was rather weak – we merely presented text underneath the figure stating what the character's normal performance was. Participants may have focused on the image and ignored the text underneath. However, it is also possible that our intuition was wrong and prior expectations do not influence people's belief that something almost happened.

Experiment 3: Expectations revisited

Experiment 2 found no influence of prior expectations. In this experiment, we made the norm manipulation stronger. We used a different scenario in which participants were presented with outcomes of a darts game (see Figure 4, cf. Gerstenberg, Ejova, & Lagnado, 2011). The darts scenario gives us a nice way of visually manipulating prior expectations. In the experiment, participants were shown an image that summarized the character's practice performance. Some characters had performed poorly in the practice (Figure 4a) while others had done well (Figure 4b). On the same screen that showed the practice performance, we also presented the character's shot in the crucial test trial. Some characters' test shots just missed the center region (Figure 4c) while other characters' shots were further off (Figure 4d). We hypothesized that both practice performance and closeness of the test shot to the center region would influence participants' judgments that the character almost hit the center region. We expected that participants' agreement with the "almost" statement would be greater for test shots that were close and for characters who had performed poorly in the practice.

Methods

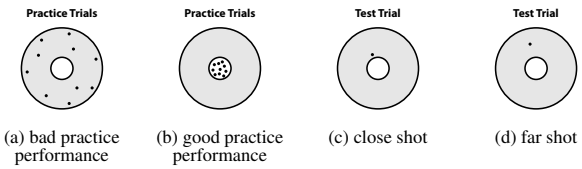
Participants 40 participants ($M_{age} = 31.4$, $SD_{age} = 8.7$, $N_{female} = 16$) completed this experiment.

Design The experiment follows as 2 (*practice performance*: good/bad) $\times$ 2 (*test shot*: close/far) within-subjects design.

Procedure After having received instructions, participants saw four trials in randomized order. On each trial, participants viewed the practice performance on the left and the test shot on the right. We generated different practice trial and test shot stimuli by simply rotating the dots around the dart board center. Participants were asked to indicate on a sliding scale which statement better describes what happened on the test trial. The left endpoint of the scale was labeled "This time, Max missed the center region." (-50), the right endpoint was labeled "This time, Max almost hit the center region." (50), and the center was labeled "unsure". It took participants 2.2 minutes ($SD = 1.6$) on average to complete the experiment.

Results and Discussion

Figure 5 shows participants' agreement ratings as a function of practice performance and the distance of the test shot to the center region. There was a main effect of distance ($F(1,39) = 78.25, p < .0001, \eta_p^2 = .67$). When the test shot was close, participants tended to say that the character almost hit the center region ($M = 19.61$, $SD = 34.24$). When the test shot was far, participants tended to say that he missed the center region ($M = -23.74$, $SD = 28.67$).



(a) bad practice performance (b) good practice performance (c) close shot (d) far shot

Figure 4: Stimuli manipulating the practice trial performance, and the distance of the test trial shot.

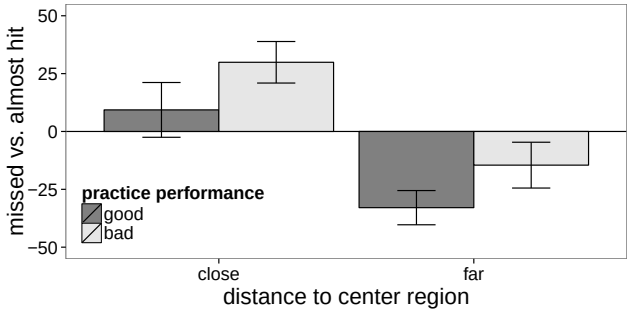


Figure 5: **Experiment 3**: Mean agreement ratings as a function of how close the test shot was to the center region (x-axis), and of how well the character had performed in the practice (bar color).

Importantly, there was also a main effect of practice performance ($F(1,39) = 12.76, p < .0001, \eta_p^2 = .25$). When the character had performed poorly in the practice, participants tended to say that he almost hit the center region ($M = 7.67$, $SD = 36.90$), whereas when the character had performed well, they tended to say that he missed the center region ($M = -11.80$, $SD = 37.31$). There was no significant interaction between practice performance and distance of the test shot.

The results of this experiment show that in line with the intuition motivated by the Jordan vs. Gerstenberg scenario, judgments about whether something almost happened are affected by prior expectations. Participants agreed more with the statement that the character almost hit the center region when he had failed to do so in the practice. The results nicely demonstrate how both causal proximity to the counterfactual outcome, as manipulated by the test shot's distance to the center region, as well as prior expectations influence participants' judgments of "almost". Interestingly, 15 out of 40 participants tended to say that the character almost hit the center region (i.e. their rating was above 0) even when the distance to the center was far, provided that the character's performance in the practice had been poor (only 5 participants' ratings were above 0 for a far shot when the practice performance was good). If our prior expectation is very low, then we may say that a counterfactual outcome almost happened even if the distance between actual and counterfactual outcome was still relatively big. If Michael Jordan threw an air ball, we would not say that he almost scored. However, if a small child who never even came close to the basket so far, finally made a shot that was close to the rim, we would be happy to say that she almost scored this time.

Experiment 4: Almost caused

The results of Experiments 1–3 show that people's judgments of whether something almost happened are sensitive to the causal distance between the actual and counterfactual world (Experiments 1 and 2), as well as prior expectations about what will happen (Experiment 3). In this experiment, we investigate what it means for something to have "almost caused" an outcome to happen, or "almost prevented" it from happening.

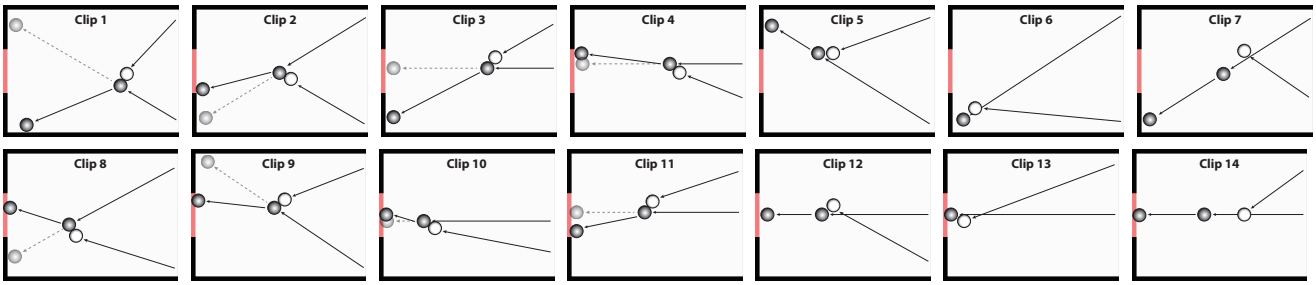


Figure 6: Stimuli used in Experiment 4. The top row shows clips in which the black ball (B) misses the gate. The bottom row shows clips in which the black ball goes through the gate. *Note*: Solid lines show the balls' actual paths, dotted lines show the black ball's trajectory if the white ball (W) had been absent.

We showed participants video clips of colliding billiard balls (see Figure 6). In half of the clips, the black ball (B) did not go through a red gate (top row), while in the other half, B went through the gate. The clips also differed in a) how close the outcome actually was, b) whether the outcome would have been different if the white ball (W) had not been present in the scene, and c) whether the outcome would have been different if W's trajectory had been a little different from what it actually was.

For example, in Clip 1, B clearly missed the gate. It would also have missed the gate if W had not been present in the scene, or if W's trajectory had been slightly different from what it actually was. In Clip 9, B almost missed the gate. It would not have gone through the gate if W had been absent, and it could well have missed the gate if W's trajectory had been somewhat different. In some of the clips, the two balls did not collide at all. For example, in Clip 5, W just misses B. If W's trajectory had been a little different, it could have knocked B into the gate.

Based on what we found in Experiments 1–3, we expected that participants agreement ratings with the "almost caused" or "almost prevented" statements, would be influenced by the perceived closeness of the counterfactual outcome, as well as their prior expectation about what would happen. Specifically, we hypothesized that participants would say that W almost caused B to go through the gate when 1) the actual outcome was close, and 2) it was clear that B would have missed the gate if W had been absent (cf. Gerstenberg et al., 2012, 2015). Conversely, we predicted that participants would agree with the "almost prevented" statement, if the outcome was close, and it was clear that B would have gone through the gate if W had not been there.

Methods

Participants 80 participants ($M_{age} = 31.6$, $SD_{age} = 10.2$, $N_{female} = 33$) completed this experiment.

Design and Procedure We manipulated what questions participants were asked to answer between conditions. Participants in the *cause condition* ($N = 40$) indicated on a sliding scale to what extent they agreed with the statement that "The white ball *almost caused* the black ball to go through the gate." (when the black ball missed), or that "The white ball *almost prevented* the black ball from going through the gate." (when the black ball went through the gate). Participants in the *counterfactual condition* ($N = 40$) indicated to what extent they agreed with three different statements about the clip: 1) "The black ball almost went through the gate.", 2) "The black ball would have gone through the gate if the white ball had not been present in the scene.", and 3) "The black ball would have gone through the gate if the white ball's path had been slightly different from what it actually was." The statements were adapted based on whether the black ball went through or missed the gate. The endpoints of the sliders for all of the statements were labeled "not at all" (-50) and "very much" (50) and the order of the statements was randomized between participants. The clips and statements were shown on the same page. Each clip was played in a loop so that participants were able to watch a clip as many times as they wanted while making their judgments. Two practice clips were presented at the beginning of the experiment. It took participants 7.2 minutes ($SD = 3.5$) on average to complete the experiment.

Results and Discussion

We will discuss the results from the *counterfactual* and *causal* condition in turn.

Counterfactual condition Figure 7 shows participants' mean agreement ratings for the three different questions they were asked for each clip. Participants judged that B almost went through the gate in Clips 2 and 4, and that it almost missed the gate in Clips 9 and 11. Participants had no trouble simulating what would have happened if W had been removed from the scene. They correctly believed that the outcome would have been different from what it actually was if W had been removed for Clips 3, 4, 8 and 9. Finally, they believed that the outcome would have been different if W's

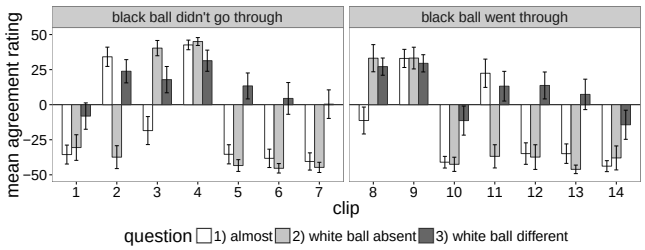


Figure 7: **Experiment 4**: Mean agreement ratings for the different counterfactual questions. *Note*: The question labels 1), 2), and 3) correspond to the questions as written in text.

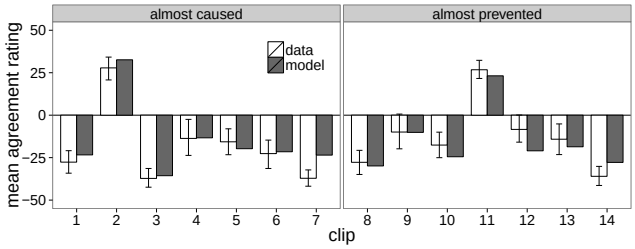


Figure 8: **Experiment 4**: Mean agreement ratings (white bars) and model predictions (gray bars, $r = .95$) for the "almost" statements.

trajectory had been somewhat different for Clips 2, 3, 4, 5, 8, 9, 11, and 12. While the answers to questions 1) and 3) were highly correlated ($r = .77$), there were situations in which they came apart. For example, for clips in which B missed the gate and the two balls did not collide (Clips 5–7), participants disagreed that B almost went through the gate. However, they tended to agree that B would have gone through the gate if W's trajectory had been somewhat different.

Causal condition Figure 8 shows participants' agreement ratings with the "almost caused" and "almost prevented" statements. Participants agreed that W almost caused B to go through the gate for Clip 2, and that it almost prevented B from going through for Clip 11. In all other situations, participants tended to disagree with the statement.

To explain people's agreement ratings in the causal condition, we use participants' responses from the counterfactual condition. The model that best accounts for participants' judgments in the causal condition combines participants' answers to questions 1) and 2) in the counterfactual condition [Model R^2 =.90; question 1 ($\beta = .96, p < .001$); question 2 ($\beta = -0.79, p < .001$); interaction ($\beta = -0.26, p = .046$)]. Question 3) was not a significant predictor of participants' causal judgments.

These results confirm our hypotheses: Participants say that W almost caused B to go through the gate when i) B actually missed the gate, ii) it was clear that B would have missed if W had not been present, and iii) B was actually close to going in. Interestingly, in situations in which W failed to collide with B, participants tended to disagree with the statement – even when a small perturbation to W's path would have been sufficient to bring about the alternative outcome (e.g. Clips 5 and 12). Thus, W must have actually made a difference for participants to say that it almost caused (or prevented) the alternative outcome. W only "almost caused" B to go through when the counterfactual outcome was close.

General discussion

When do people say that something almost happened? In this paper, we have expanded on previous work that has investigated when counterfactual worlds come close (Kahneman & Miller, 1986; Kahneman & Varey, 1990; Teigen, 1996). In line with Kahneman and Varey (1990), Experiments 1 and 2 showed that the perceived causal distance between actual and counterfactual world affects participants' judgments that something almost happened. If a small perturbation to the relevant causal event would have been sufficient to generate the alternative outcome, then that outcome almost happened. In contrast to Kahneman and Varey (1990), Experiment 3 showed that participants' "almost" judgments were, in addition to causal distance, significantly influenced by manipulating prior expectations. Participants were more likely to say that an event almost happened, when their prior expectation was low. Finally, Experiment 4 extended these findings to modeling people's judgments about causation. In line with the *counterfactual simulation model* (Gerstenberg et al., 2015), participants judged that a candidate cause "almost caused" an event to happen when the cause actually made a difference to the outcome, it was clear that the outcome would *not* have happened had the cause been absent, and when the counterfactual outcome was close.

Goodman's (1983) quote at the beginning of this paper states that all possible worlds lie within the actual one. In this paper, we have shown what makes some worlds lie a little closer to the actual world than others (cf. Phillips, Luguri, & Knobe, 2015). One of the key questions for future research is to study in more detail how prior expectations and causal proximity to the counterfactual outcome help us to bring order into possible worlds.

Acknowledgments This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

References

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Goodman, N. (1983). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Johnson, J. T. (1986). The knowledge of what might have been: Affective and attributional consequences of near outcomes. *Personality and Social Psychology Bulletin*, *12*(1), 51–62.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology*, *59*(6), 1101–1110.

Nouwen, R. (2006). Remarks on the polar orientation of almost. *Linguistics in the Netherlands*, *23*(1), 162–173.

Penka, D. (2006). Almost there: The meaning of almost. In *Proceedings of Sinn und Bedeutung* (Vol. 10, pp. 275–286).

Phillips, J., Luguri, J., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Reid, R. L. (1986). The psychology of the near miss. *Journal of Gambling Studies*, *2*(1), 32–39.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

Stocker, K. (2014). The elements of cause and effect. *International Journal of Cognitive Linguistics*, *5*(2), 121–145.

Teigen, K. H. (1996). Luck: The art of a near miss. *Scandinavian Journal of Psychology*, *37*(2), 156–171.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Tit-for-Tat: Effects of Feedback and Speaker Reliability on Listener Comprehension Effort

Nicole Craycraft* (ncraycraft@gmail.com)
Department of Psychology, University of Illinois, Urbana-Champaign

Zoe Kriegel* (zkriegel@kent.edu)
School of Health Sciences, Kent State University

Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Brain & Cognitive Sciences, University of Rochester

Jennifer Roche (jroche3@kent.edu)
School of Health Sciences, Kent State University

Abstract

Miscommunication is often seen as a detrimental aspect of human communication. However, miscommunication can differ in cause as well as severity. What distinguishes a miscommunication where conversation partners continue to put forth the effort from miscommunication where conversation partners simply give up? In this eye-tracking study, participants heard globally ambiguous statements that were either a result of an experimental error or speaker underspecification; participants either received positive or negative feedback on these ambiguous trials. We found that negative feedback, paired with the reliability of the message, will impact the amount of processing effort a comprehender puts forth—specifically, listeners were less forgiving of errors when they were penalized and when speakers' instructions lacked effort. This suggests that language users weigh conversational contexts and outcomes as well as linguistic content during communication.

Keywords: ambiguity; intentions; communication; comprehension; context

Introduction

Communication ideally ensures a successful exchange of information between speakers and listeners. Unfortunately, communication rarely functions ideally: a successful exchange will be riddled with unsuccessful attempts. But are all miscommunications equally disruptive? The answer to this question has implications for the mechanisms used in language processing.

In an egocentric account, speakers and listeners automatically use their own perspective, often failing to take the perspective of others in the conversation. Evidence for this view includes the finding that speakers are *more* likely to inadvertently reveal the identity of a hidden object when explicitly instructed to keep it secret, presumably due to the increased attention paid to it (Lane & Ferreira, 2008) and that speakers in a matching task are better able to tailor utterances to the past experience of a particular listener when item category labels are distinct for each listener rather than overlapping, reducing the memory load of pairing a specific item to a specific listener (Horton & Gerrig, 2005). In this view, miscommunication is a necessary consequence of egocentric thinking, and would often go undetected. Indeed, when task partners have different goals and were either uninformed or misinformed that their goals were the same, statements by both parties were routinely misinterpreted and only 1 pair out of 31 discovered their goals were different (Russell & Schober, 1999)

Contrasting this is an audience design account, where the perspective of and previous experiences with a conversational partner constrain interpretation or production. An assumption of this account is all parties in a conversation possess cooperative intent. Grice (1975) presented a theory of general maxims that communication must follow to be successful, including "make your contribution as informative as is required (for the purposes of the exchange)" as a maxim of Quantity and "do not say that for which you lack adequate evidence" as a maxim of Quality. From these maxims as well as others, we can derive that listeners assume speakers will provide true information with as much detail as a context necessitates. There is a body of literature suggesting that listeners can flexibly incorporate contextual information to intuit the intentions of speakers. Listeners use an ambiguous utterance as a signal of mismatching perspectives and readily adapt to the speaker's perspective to quickly identify a referent (Hanna, Tanenhaus, & Trueswell, 2003). Shared experiences may form a *referential pact* by establishing a potentially ambiguous phrase as a term for a referent; when a speaker exerts more effort and uses a new term for a previously mentioned referent,

* Indicates shared first authorship

it delays comprehension of the listener (Metzing & Brennan, 2003).

In this audience design framework, miscommunication would be a mismatch between perceived context and actual context. Roche, Paxton, Ibarra, and Tanenhaus (*under review*) posit that miscommunication can promote deeper processing and more care to resolve misaligned representations of the world. Broadly, most miscommunications can be explained as either an accident by a still cooperative party or as an indication of a partner who is an unreliable communicator.

Listeners suspend their expectations of correct usage when a demonstrably unreliable speaker mislabels common objects and misuses modifiers (Grodner & Sedivy, 2011). Even young children are sensitive to speaker reliability, as they usually rely upon adults over their peers for accuracy of new information. In fact, when preschoolers detected inaccuracies from an adult, they quickly overrode this tendency (Jaswal & Neely, 2006). In addition, young children also show capabilities to learn selectively from more reliable sources without specifically discussing prior reliability of relevant speakers (Birch, Vauthier, & Bloom, 2008). Speakers that continually provide false or vague information may cause listeners to disregard subsequent utterances by the same speaker. Speakers also have assumptions about listeners: namely, that they will pay attention. In Kuhlen and Brennan (2010), speakers told jokes they had previously learned. When speakers expected attentive listeners and received distracted listeners, they used fewer details.

In the following task, we tracked listeners' eye movements while they heard unambiguous or globally ambiguous (i.e., unresolvable) descriptions from a speaker during a matching task. Sometimes, the globally ambiguous description is due to a perspective mismatch; in other cases, the speaker simply did not provide enough information. During the globally ambiguous trials, half of the listeners received negative feedback and half received positive feedback. If miscommunication is processed purely egocentrically, trials with negative feedback should cause the participant to disengage from the task regardless of the source of the error. However, if listeners are sensitive not just to the *presence* of a miscommunication but the *cause* of it, the accidental errors should be more easily "forgiven" than lazy speaker errors, especially in the negative feedback cases. We expect listeners to process ambiguity differentially depending on the reason for the ambiguity and whether the ambiguity prevents communicative success.

We hypothesize that listeners understand communication as a shared experience and should process errors differently depending on the intention and outcome. We predict that listeners should exert more effort in processing the speaker's perspective if the error was the fault of the experimenter compared to errors due to speaker laziness. The consequence of the error should also affect the listener. If the listener is penalized for the speaker's laziness, then the listener should stop *working hard* too (tit-for-tat).

Methods

In the following study, we used a pseudo-confederate design (description below), which involved the ruse that the listener was interacting with a live person (similar to Roche, Dale, & Kreuz, 2010). During the task, the listener heard an instruction (pre-recorded statement) describing which object to click. The listener then saw feedback about the correctness of the choice she made. On some of the trials, the listener had difficulty making the correct decision because the speaker produced a globally ambiguous statement. The listener then learned the source of the ambiguity, which we predict will differentially affect processing effort on decoding the error and future effort.

Participants

Sixteen undergraduates from Kent State University (15 females; mean age = 21.5 years) participated for extra credit in a Speech Pathology & Audiology course, resulting in a total of 4,480 data points across the experimental trials. All participants were native speakers of American English with normal to normal-corrected vision. None reported speech or hearing impairments.

Materials & Stimuli

A 21inch iMac (experiment computer), Eyelink 1000 (eye-tracking computer), noise cancelling headphones, wireless mouse, and usb microphone (for pseudo-confederate recordings) were used. All materials were set up in a sound attenuated booth, with two chairs – one for the participant and one for the experimenter to control the eye tracking computer.

Visual stimuli. Participants were presented 4 shapes (square, circle, triangle, and star) x 4 colors (purple, blue, green, and red) x 2 sizes (big and small), resulting in a total of 32 possible shapes. On a given trial, two shapes were paired with each other, resulting in visual stimulus pairs that overlapped on zero, one, or two features.

Auditory stimuli. Participants were presented with pre-recorded statements from a Caucasian female talker that referenced one (e.g., big shape), two (e.g., big red shape), or three (e.g., big red triangle) possible features of the visual stimulus. The recordings were equated for RMS

amplitude to adjust stimulus sound level for more comfortable listening. This process "turned up" the volume on the sound files that were at lower amplitudes to match the highest amplitude and did not affect recording quality.

Design & Procedure

An individual listener interacted with a confederate during the informed consent portion of the study, but the listener and confederate were separated during the experimental task. After consent was obtained, the listener was seated in the sound booth with the door open, and the confederate sat at a computer and microphone around the corner. The listener and confederate received separate instructions, but the listener could hear the experimenter providing the confederate's instructions. The listener and confederate were told they would interact with each other via a computer network – much like a cell phone conversation. In fact, the listener never interacted with the confederate during trials but, instead, heard pre-recorded statements by a pseudo-confederate.

Once the pseudo-confederate ruse was established, the listener was calibrated, validated, and drift corrected using the prescribed procedure for the Eyelink 1000. During the task, the listener viewed two objects on the computer screen in addition to a bullseye (see Fig. 1, *left panel*). By clicking the bullseye, the listener initiated the trial and received a pseudo-confederate instruction.

On any given trial, the instruction could have been ambiguous or unambiguous. Sometimes, the global ambiguity forced the listener to guess which object the speaker intended (e.g., "Click on the red shape," when paired with a visual display containing two red shapes; see Fig. 1, *left panel*). On the next screen, the listener received feedback indicating whether the choice was correct or incorrect. The content of the speaker's screen was also on this screen.

A 2 **Feedback** (*Consequence* vs. *No Consequence*; between-subjects) by 3 **Error Type** (*No Error*, *Experimenter Error*, or *Lazy Speaker Error*; within-subjects) design was used. The **Feedback** conditions determined the type of feedback listeners received on globally ambiguous trials. Listeners in the *Consequence* condition always received negative feedback on the globally ambiguous trials – indicating they chose the incorrect object (see Fig. 1, *right panel*). Listeners in the *No Consequence* condition always received positive feedback on the globally ambiguous trials – indicating they chose the correct object (green check mark, instead of a red X, as indicated in Fig. 1, *right panel*).

Feedback was crossed with **Error Type**. In the *Experimenter Error* condition, the speaker did not see the same objects as the listener – the source of the error was due to mismatching visual referents (see Figure 1, *right panel*). In the *Lazy Speaker Error* condition, the speaker accidentally described the overlapping feature instead of the disambiguating feature – the speaker was being lazy and not paying attention to detail. Trials without errors comprised the No Error condition.

Over the course of the experiment, each listener was presented with a total of 280 experimental trials: ~20% of trials included a global ambiguity (60 trials total). The global ambiguity trials were pseudo-randomly distributed across the experiment with the first error occurring at trial 57 (i.e., 912 critical data trials across all participants). As a reminder, these types of errors required the listener to make a best guess. The other 80% of trials included a *resolvable ambiguity* or *unambiguous* statement that did not affect the listener's ability to choose correctly.

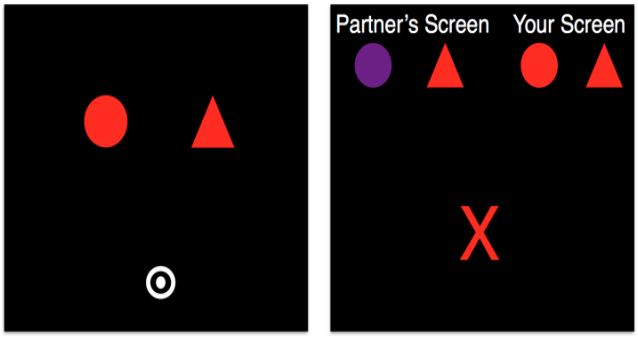


Figure 1. Sample experiment screen for an Experimenter Error trial, with negative feedback, when the listener heard: "*Click on the red shape.*" The *left panel* represents the Response/Instruction screen; the *right panel* represents the Feedback Screen.

Measures

Over the course of the experimental trial, dwell times were evaluated (i.e., how long the listener fixated within an interest area). Dwell time has been suggested to be a good measure of cognitive processing, and we use it to determine processing effort (Ehrlich & Rayner, 1981; Rayner & Duffy, 1986). It is expected that under moments of ambiguity that the listener should experience increased cognitive load while processing the ambiguity. However, interest here laid in the amount of time *and* effort the listener was willing to put forth understand the locus of ambiguity. Therefore, we chose to use dwell time measure, instead of pupil dilation – which is also an established measure of cognitive load (c.f., Kahneman, 1973).

The three areas of interest included the two objects on the Instruction Screen (Fig. 1, *left panel*) and the location of the "Partner Screen" information on the Feedback Screen (Fig. 1, *right panel*). Dwell times were calculated using the Eyelink Dataviewer software based on a fixation that landed in the interest area for a predetermined algorithm that calculates the amount of time in milliseconds.

Results

Experimenter and Lazy Speaker errors are both miscommunications but imply very different intentions. An *Experimenter Error* indicates the speaker did not realize that the ambiguous statement could have been harmful, unless she received feedback that what she said was wrong or confusing. However, the *Lazy Speaker Error* indicates that the speaker failed to put forth the necessary effort to disambiguate and thus violated the principle of collaborative effort.

The current study aims to answer three questions: 1) How does feedback affect the processing of ambiguous statements over time?; 2) How much effort will a listener put forth to understand a miscommunication?; and 3) Does the amount of effort to understand a previous misunderstanding affect the amount of effort put forth on future language comprehension? We predict that negative feedback should make the ambiguity more salient—thus recruiting more cognitive resources initially (questions 1 & 2). However, as the listener learns that the speaker's ambiguity is often unreliable (causing communication breakdown or failure) and negatively impacts the listener, we should see cognitive effort decline (question 3).

The data was analyzed using growth curve models, multivariate methods for analyzing time series data that simultaneously allows for the measurement of group and individual level effects (Mirman, Dixon, & Magnuson, 2008). The calculation of orthogonal polynomials resolves the issues of dependence associated with the time series. The first orthogonal indexes linear slope, and the second orthogonal indexes line curvature.

Understanding the Ambiguity

A growth curve model evaluated the effects of **Feedback** (*Consequence* vs *No Consequence*) and **Error Type** (*No Error*, *Experimenter Error*, vs. *Lazy Speaker Error*) on dwell time to the "Partner Screen" interest area of the Feedback screen as an indirect measure of how hard the listener tried to understand why the miscommunication occurred (see Fig. 1, *right panel* to reference the "Partner Screen" region of the Feedback Screen). Based on visual inspection of the group data, we decided to calculate up to the second orthogonal polynomial to interact Feedback and Error Type (see Table 1 for statistics and p-values).

The significant main effect of the first orthogonal indicated that as the listeners progressed through the experiment, dwell time significantly decreased—essentially, task adaptation. The main effect of **Error Type** indicated that *Experimenter* and *Lazy Speaker Errors* had longer dwell times than *No Error* trials (p < .001) – demonstrating that global ambiguity recruited more cognitive effort.

More interesting is the effect of the global ambiguity over time. The significant **first orthogonal (linear slope) polynomial** x **Error Type** shows a steeper linear decline for the *Experimenter* and *Lazy Speaker Errors* relative to *No Error*. It appears that globally ambiguous statements were seriously considered initially, but over time, the listeners gave up on exerting effort in processing. Potentially, this effect is a result of disengaging from the task. The significant **second orthogonal (line curvature) polynomial** x **Error type** interaction suggests that at least one of the Error Types was marked by a more curvilinear line, which is possibly reflective of adapting to the global ambiguity. Finally, the **Feedback x Error Type** interaction suggests that listeners' dwell times during global ambiguity trials decrease as a function of negative feedback: listeners that were penalized stopped considering the speaker's perspective faster on the global ambiguity trials than listeners that were not penalized (at least p < .05; see Fig. 2).

Table 1: Estimates, standard errors, t and p-values for the growth curve model evaluating dwell time to the "Partner Screen" region of the Feedback screen.

Effect	b	se	t
1st orth.	-250.39	45.79	-5.468***
Error	2133.48	223.54	9.54***
1st orth. x Error	-1304.67	206.68	-6.313***
2nd orth. x Error	568.33	172.12	3.30***
Feedback x Error	87.23	18.53	4.47***

Note: ***p < .001

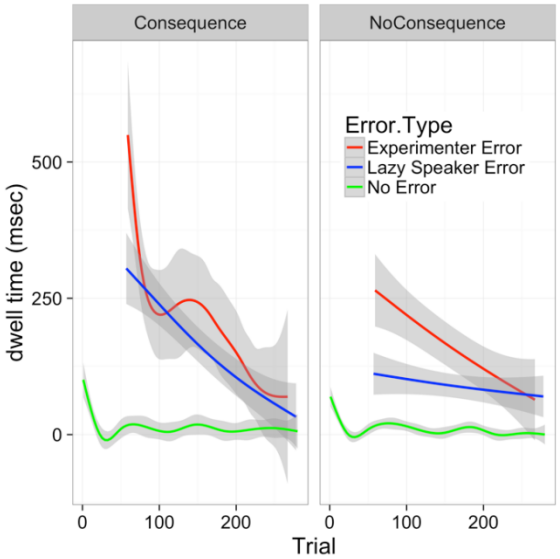


Figure 2: Average dwell time (msec) on the Partner Screen location of the Feedback screen as a function of Feedback Condition by Error Type.

Effect of Global Ambiguity on Future Success

Interestingly, the *No Error* trials had a near-sinusoidal pattern of dwell times (see Fig. 2). The purpose of this growth curve analysis was to determine if the global ambiguity trials had an effect on *No Error* trials. Specifically, we used a growth curve model to explore changes in dwell time during language comprehension (i.e., object locations on the Response Screen, where the listener has to select the target shape). We are especially interested in how hard the listener tried to understand the speaker's perspective after the listener just experienced global ambiguity on the previous trial.

A growth curve model evaluated the effects of **Feedback Condition** (*Consequence* vs *No Consequence*) and the **Error type** (*No Error*, *Experimenter Error*, vs. *Lazy Speaker Error*) on the No Error trials, up to the second orthogonal polynomial. Experimenter Error and *Lazy Speaker Error* trials were excluded from this analysis to reveal the effect of the speaker's intent on listener comprehension of the speaker's future instructions (see Table 2 for statistics and *p*-values).

Table 2: Estimates, standard errors, t and p-values for the growth curve model evaluating dwell time to the object regions of the Instruction screen.

Effect	b	se	t
1st orth.	-2095.03	646.01	-3.24**
Error	207.37	77.28	-2.68**
2nd orth. x Error	-4648.72	1539.35	-3.02**

Note: **p < .01

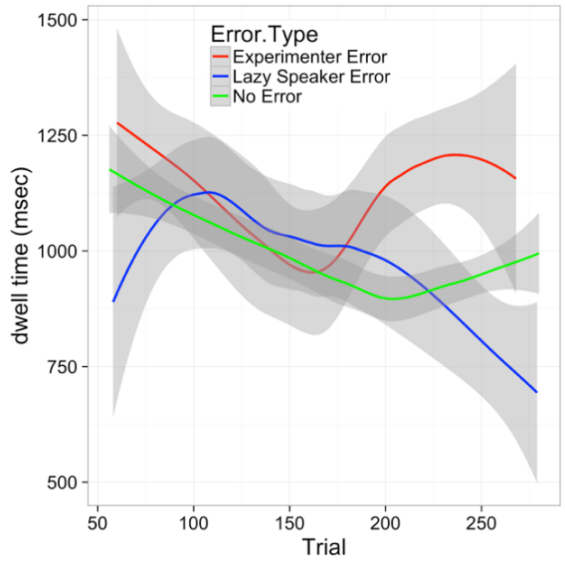


Figure 3: Average dwell time (msec) on the objects on the response screen as a function of Error Type experienced on the previous trial, for No Error Trials.

The main effect of the first orthogonal (linear slope) polynomial demonstrates that dwell time to the speaker's objects on the response screen decreases over time regardless of experimental conditions—i.e., task adaptation. Simple effects analysis of the Error main effect found if a listener had just experienced a *Lazy Speaker Error*, dwell times to the speaker's objects were significantly shorter on the next trial – indicating the *Lazy Speaker Error* negatively affected future language comprehension ($p < .05$). If the listener perceives the speaker lacking in effort, the listener will correspondingly not put in effort to comprehend the speaker's message. However, if the listener perceives the error to not be the fault of the speaker, processing effort increases on the next *No Error* trial ($p < .05$; see Fig. 3). Finally, the interaction between the second (line curvature) orthogonal polynomial and Error type showed that *Lazy Speaker Errors* produced a more curvilinear line than the *Experimenter Error* and *No Error* types ($p < .05$), indicating that a *Lazy Speaker Error* disrupted processing more than the other trial types.

Discussion

In summary, our sample of listeners in typical communication expended different amounts of effort based on the available information. We first evaluated how much effort listeners put forth in considering the speaker's perspective (i.e., "Partner Screen" interest area). Overall, it would seem that when a listener is penalized (i.e., receives negative feedback), the listener more quickly stops considering or reduces consideration for the speaker's perspective. Past errors also modulate future comprehension of the speaker's statements. When a listener experiences a **Lazy Speaker Error** (relative to **Experimenter Error** trials), the listener spends significantly less time considering the objects on subsequent **No Error** trials. It would seem that if a listener understands errors were not the speaker's *fault*, the listener would exhibit willingness to exert more cognitive effort. However, if the speaker is perceived to be *lazy*, listeners will reciprocate – by being *lazy* themselves.

In addition, the very first error presented to participants was an *Experimenter Error*. There were marginally longer processing times on the "Partner Screen" interest area in the *Consequence* condition relative to the *No Consequence* condition ($t = 1.696$, $p = .09$), showing a trend in a direction of longer dwell times when negative feedback is provided. The negative feedback seems to make the listener consider the speaker's perspective, probably because the listener was surprised and wanted to understand what happened. Even more interesting was the drastic decrease in processing time at the last instance of a **Lazy Speaker Error** for the **Consequence** condition ($t = 3.636$, $p < .001$). This suggests that listeners might put forth

significantly less effort in trying to understand their partner's perspective if they perceive their partner to not care and were penalized for the error. Therefore, it would seem that listeners and speakers may develop a *tit-for-tat* relationship when negative consequences affect the communicators over the course of an interaction. This is consistent with previous work on listeners handling ambiguous language in a rational manner. For example, Degen, Franke, and Jäger (2013) suggest that listeners interpret ambiguous referents in a game theoretical manner - assuming the ambiguous word refers to only one target because if the speaker meant to refer to the other target, they could have used an unambiguous word instead. When messages had different costs, as unambiguous costs increase, listeners make more inferences based on the ambiguous messages (Rohde et al., 2012). In this case, listeners rationally respond to a speaker that is uncooperative by disengaging from the task, and respond to a speaker that was misinformed with more attention.

Conclusion

Ambiguity need not be problematic for conversations because it can be quite helpful in reducing some of the cognitive effort exerted by both listener and speaker in typical communication. However, these benefits happen only if the ambiguity is properly situated in context. Ambiguity may also promote deeper processing by requiring repair and attention to detail.

It would seem that, in the current study, the interpretation of ambiguity can be ignored when a visual referent helps disambiguate. However, when an ambiguity becomes problematic, the listener weighs the consequence and reason for the ambiguity – subsequently affecting future processing effort. This work provides evidence that listeners are able to interpret miscommunications within the communication context and respond differentially based on both intent and direct consequences.

Acknowledgements

A special thanks goes to the research assistants in the MADI lab at Kent State University.

References

Birch, S. A. J., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018–1034.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. *Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci'13)*, 376–281.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Grice, P. (1975). Logic and conversation. In Cole, P.; Morgan, J. (Eds). *Syntax and semantics: Speech acts*. New York: Academic Press.

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. *The Processing and Acquisition of Reference*, *2327*(8957), 239–272.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Jaswal, V. K., & Neely, L. a. (2006). Adults don't always know best. *Psychological Science*, *17*(9), 757–758.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Kuhlen, A. K., & Brennan, S. E. (2010). Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. *Discourse Processes*, *47*(7), 567–587.

Lane, L. W., & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1466–1481.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*(3), 191–201.

Roche, J. M., Dale, R., & Kreuz, R. J. (2010). The resolution of ambiguity during conversation: More than mere mimicry? *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 206–211.

Roche, J. M., Paxton, A., Ibarra, A., & Tanenhaus, M. K. (*under review*). Communication breakdown prompts readjustment of informativeness.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., & Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*, 107 – 116.

Russell, A. W., & Schober, M. F. (1999). How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes*, *27*, 1–33.

Distinguishing processing difficulties in inhibition, implicature, and negation

Ann E. Nordmeyer
anordmey@stanford.edu
Department of Psychology
Stanford University

Erica J. Yoon
ejyoon@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Despite their considerable communicative abilities, young children often have difficulty interpreting complex linguistic structures in context. Two examples of this phenomenon are negation and pragmatic implicature, both of which pose sometimes surprising difficulties for preschoolers. Both of these structures require children to resist a more salient alternative interpretation; since executive function abilities develop extensively during childhood, perhaps failures are due to problems in inhibition. To test this hypothesis, we designed tasks to measure inhibitory control, negation, and implicature comprehension in children and adults. Using standard analyses as well as drift diffusion models, we found different patterns of processing on all three tasks, and no support for the hypothesis that inhibitory control per se is playing a role in either adults' or children's negation or implicature processing. Instead, our analyses reveal qualitatively different developmental trajectories, suggesting task-specific factors driving these changes.

Keywords: Inhibitory control; negation; implicature; drift diffusion model; cognitive development; pragmatics

Introduction

By ages 3–4, children are often quite proficient at general communication, using language to tell stories, coordinate actions, and navigate the social world (E. V. Clark, 2009). Despite this proficiency, the same children often fail laboratory language processing tasks that test important but complex linguistic constructions. For example, in tests of negation comprehension, children will often struggle to identify a character who has "no apples," instead choosing the character who *does* have apples (Nordmeyer & Frank, 2014b). And while adults easily make the ad-hoc pragmatic implicature that "My plate has a carrot" refers to a plate with *just* a carrot (rather than a plate with a carrot and a banana), children often fail (see example stimuli in Figure 1; Yoon, Wu, & Frank, 2015).

Children's difficulties with these tasks are surprising in light of their other abilities in each domain. Children under two years spontaneously produce negative utterances (Bloom, 1970; Pea, 1980, 1982), yet children as old as four years consistently fail tests of negation comprehension (e.g. Kim, 1985; Nordmeyer & Frank, 2014b). And children demonstrate an understanding of pragmatic principles by age three (E. V. Clark, 2009), adjusting the informativeness of their own expressions depending on the difficulty of referent disambiguation (Matthews, Lieven, Theakston, & Tomasello, 2006) and the extent of listeners' knowledge (Matthews, Butcher, Lieven, & Tomasello, 2012). Yet they stuggle to make ad-hoc implicatures before age four (Yoon et al., 2015; Stiller, Goodman, & Frank, 2015) and scalar implicatures with quantifiers beyond age five (Huang & Snedeker, 2009).

Why do children struggle on negation and implicature when production data and past experiments suggest that they have the requisite linguistic abilities? In both of these examples, children's failures might occur not because they lack linguistic understanding, but because the incorrect choice is more salient or perceptually interesting. When children are asked to "find the boy with no apples," they must look *away* from the character who is holding the labeled objects, and look at some other, unlabeled object instead. Similarly, to demonstrate comprehension of "my plate has a carrot," children must look away from the more visually salient plate (the one with two objects instead of one). In both cases, poor executive function might explain children's difficulty selecting the correct response on these comprehension tasks.

Executive function is a construct that encompasses a wide variety of functions including response inhibition, task-shifting, and working memory (Miyake et al., 2000). These abilities change dramatically during early childhood (Diamond & Taylor, 1996; Davidson, Amso, Anderson, & Diamond, 2006), especially during the period of interest (ages 3–5), making them a possible candidate factor underlying changes in language processing. The specific executive function construct most relevant to negation and implicature is response inhibition, because classic tests of this construct require participants to override a prepotent response. Participants who react quickly often make the wrong choice, and correct choices are typically slower. The presence of these sorts of speed-accuracy tradeoffs in language comprehension data might be a signal of inhibitory issues, but speed-accuracy tradeoffs can be difficult to understand with traditional analyses, which examine reaction time and accuracy separately.

Drift diffusion models (DDMs) are an important tool for addressing the speed-accuracy tradeoff that often arises in response inhibition tasks. DDMs (e.g. Ratcliff, 1978) use both accuracy and reaction time as measures of the decision making process in simple two-alternative forced choice paradigms. The model can be imagined as a noisy accumulation of evidence over time, resulting in an eventual correct or incorrect choice when the accumulated evidence reaches a predetermined decision "boundary." The model produces four key parameters: boundary separation (the amount of evidence needed to reach a positive decision), bias (to what extent is the decision process biased towards one decision or another), non-decision time (the time needed to encode basic information about the stimuli, before embarking on the decision process), and drift rate (the rate at which evidence is accumulated). Past work has shown that children typically have longer non-decision times, higher separation boundaries, and slower drift rates, suggesting that children require more time to process stimuli, more information to make a de-

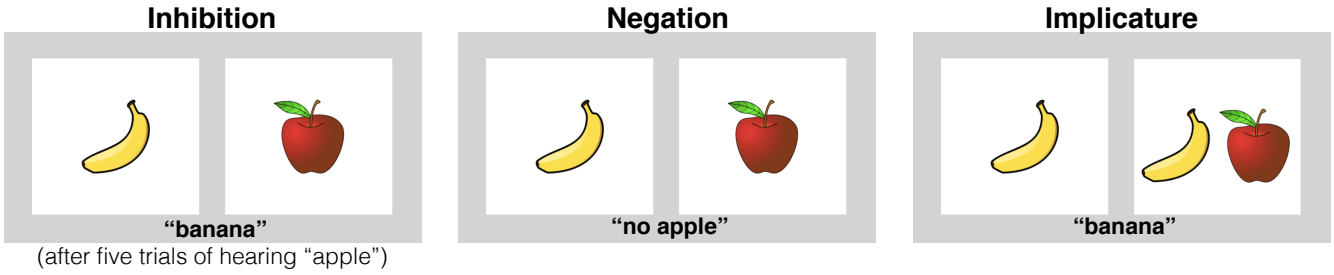


Figure 1: Example stimuli from each condition. Words identifying the target picture (left in each panel) are given at the bottom.

cision, and take longer to accumulate evidence compared to adults (Ratcliff, Love, Thompson, & Opfer, 2012).

In the current work, we test the hypothesis that changes in response inhibition underlie developmental changes in children's processing of negation and implicature. We use a simple design with three separate subtasks to test adults' and children's inhibitory control, negation comprehension, and implicature comprehension. In order to explore how children's language processing changes across development, we collected data from 4–6-year-olds, as well as adults. Despite the fact that the three tasks contained almost identical visual and auditory stimuli, results of conventional and DDM analyses suggest distinct patterns of information processing and developmental change across the three tasks. This suggests that, contrary to our initial hypothesis, children's undeveloped inhibitory control cannot entirely explain their difficulty with negation and pragmatic implicatures. Instead, our data suggest that negation and implicature appear to follow different developmental trajectories throughout early childhood.

Method

We created three simple "find the picture" tasks (stimuli pictured in Figure 1) to test inhibitory control, negation, and implicature in adults and children. Our initial version of this task was implemented on an iPad tablet (and results were similar), but in this version we used a computer to facilitate shorter reaction times and increase comparability between the children and an online sample of adults (who used a computer).

Participants

We invited parents and children at the Children's Discovery Museum in San Jose, CA to play a computer game. Ten children were excluded from analysis because their parents indicated that they heard English at home 50% of the time or less. An additional 24 children were excluded from analysis for failing to complete at least half of the trials in each task. These exclusions resulted in a final sample of 22 4-year-olds (mean age 4;7), 19 5-year-olds (mean age 5;5), and 25 6-year-olds (mean age 6;5). We also recruited adult participants on Amazon Mechanical Turk to play the computer version of the task. Two adults were excluded for failing to complete at least half of the trials, resulting in a final sample of 48.

Stimuli and design

The experiment consisted of three tasks: inhibition, negation, and implicature. In each trial, two images were presented side by side on the screen. Images were presented in yoked pairs, cycling through the same twelve possible item pairs.

A pre-recorded voice said one or two words, and participants' task was to select the correct referent as soon as they could. We used words (instead of full sentences) both to keep linguistic stimuli as similar as possible across the three tasks and to keep trials very short, allowing us to collect many trials per participant (a requirement of the planned DDM analyses). On each task, participants were instructed to select the corresponding picture as quickly and accurately as possible.

For the inhibition task, in a set of 6–8 trials, the same two pictures appeared side by side (e.g., a picture of a banana and a picture of a apple). The target picture appeared on either the left or the right side, with target side randomized across all trials. For the first 5–7 trials (control trials), one of the two objects was named (e.g., "apple"), then on the last trial (target trials), the other object was named ("banana"). Children saw a total of 12 sets of trials, and adults saw 24 sets.

For the negation task, the referents were named with or without negation. For example, given two pictures of banana and apple respectively, to refer to the banana the recorded voice said "banana" (positive trials, the control for this task) or "no apple" (negative trials, the target for this task). Children saw 60 trials, and adults saw 120 trials.

For the implicature task, in each trial there was a picture with one object (e.g., banana) and another picture with the same object as well as another (e.g., banana and apple). In control trials, the unique object was named ("apple") and in target trials, the common object was named ("banana"), implying "banana *but not apple*" (an ad-hoc implicature). Children saw 60 trials, and adults saw 120 trials.

Procedure

Parents and children were recruited on the floor of the CDM and were led to a small nearby research room. An experimenter first explained the task to children, who then went through two practice trials, where they were asked to select an obvious, unambiguous referent (e.g., "cow" as opposed to "rabbit"). Children selected the picture on the left side of the screen by pressing the 'z' key and selected by picture on the

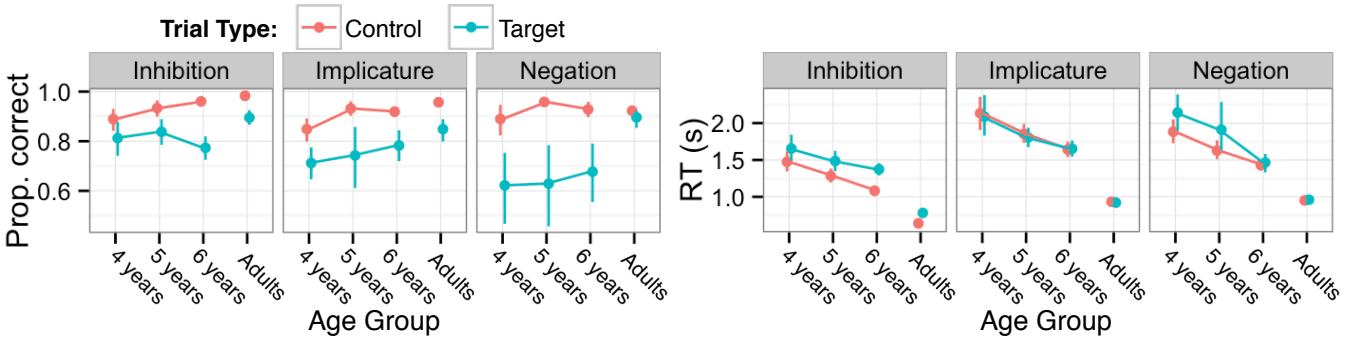


Figure 2: Adults' and children's mean proportion correct responses (left) and reaction time (right) for target and control trials across the three tasks. Control trials are shown in red, target trials are shown in turquoise. Error bars show 95% confidence intervals.

right side of the screen by pressing the '/' key; both keys were covered by blue stickers to help children find them. Then they went through the three tasks in a randomized order. Afterwards, the experimenter gave them a sticker as a gift. Adults played the same game without practice trials.

Analysis and preprocessing

Because some children responded immediately without listening to the entire word and some children took an exceedingly long time to respond to some trials, we made a decision to remove any trials with reaction times less than 200 milliseconds from the word onset, and longer than 15 seconds from the word onset.[1] After these initial exclusions, we followed our initial planned analysis by removing RTs outside of three standard deviations from the log-transformed mean.

We fit diffusion models separately to each individual participant's data using the RWiener package,[2] using the Nelder-Mead method to estimate optimal parameter values. We estimated parameters separately for each trial type (target vs. control) within each task for every participant, and then calculated the mean and 95% C.I. across all participants within each age group (4-year-olds, 5-year-olds, 6-year-olds, and adults). We removed participants with any parameter value outside of 3 standard deviations from the mean for that parameter within each game.

All regression results reported are based on linear (RT and DDM parameters) or logistic (accuracy) mixed effect models, fit using `lme4` with the maximal convergent random effect structure. Models of RT were fit using log-transformed RTs.

Results and Discussion

This experiment yielded a large data set that affords many analyses. Participants' accuracy and reaction times are shown in Figure 2; DDM fits are shown in Figure 3. We organize the results reported here around those findings that speak most directly to our question of interest. First we use traditional accuracy and reaction time measures to examine whether individual performance on the inhibition task is correlated with individual performance on the implicature and negation tasks. We then use the drift diffusion model to explore differences in processing across the three tasks, and examine the development of children's processing in each task. Raw data are available at `https://github.com/anordmey/cogsci16`, and a fuller set of analyses can be viewed at `http://anordmey.github.io/cogsci16/analysis/neginhib_cogsci16.html`.

Inhibition task performance was not correlated with negation or implicature performance

Our key hypothesis was that successful performance on the negation and implicature tasks requires inhibitory control, and hence performance on negation and implicature should be correlated with performance on inhibition. To test this hypothesis, we calculated standardized difference scores between target trials and control trials for each participant on each task, using both accuracy and RT. For accuracy, we found no significant correlations between individual accuracy on the inhibition task and the implicature task for adults ($r = .21$, $p = .16$, 95% CI [-0.08, 0.46]) or kids ($r = -.09$, $p = .49$, 95% CI [-0.32, 0.16]), nor between individual accuracy on the inhibition task and the negation task for adults ($r = -.11$, $p = .46$, 95% CI [-0.38, 0.18]) or kids ($r = -.08$, $p = .51$, 95% CI [-0.16, 0.32]). For RT, there was similarly no relationship between individual RT on the inhibition task and the implicature task for adults ($r = -.004$, $p = .98$, 95% CI [-0.29, 0.28) or kids ($r = .20$, $p = .11$, 95% CI [-0.04, 0.42]), nor between individual RT on the inhibition task and the negation task for adults ($r = .17$, $p = .25$, 95% CI [-0.12, 0.43]) or kids ($r = -.01$, $p = .92$, 95% CI [-0.24, 0.26]).

Consistent with past findings (e.g. Nordmeyer & Frank, 2014b; Yoon et al., 2015), we found a great deal of individual variability in children's performance on these tasks, with 4- to 6-year-olds struggling especially on the implicature and nega-

[1] Although this decision was made post-hoc based on some unexpectedly short and long reaction times, the decision did not affect the key findings we present in this paper.

[2] All analyses described in this paper were conducted using R version 3.2.1

tion tasks. Contrary to our hypothesis, however, performance on these two tasks was not correlated with performance on the inhibition task. One drawback to these traditional analyses is that they separate reaction time and accuracy, obscuring any potential interaction between the speed at which participants make a decision and the accuracy of their decision. In the next sections, we use the parameters of drift diffusion models to compare participants' decision-making processes.

Adults show different decision processes for inhibition, implicature, and negation

Despite the surface similarities of the tasks, adults' decision processes for control vs. target trials were different across the three tasks (Figure 3, right panels). In the inhibition task, the most striking difference between control and target (inhibition) trials was the bias towards incorrect responses on target trials. The bias parameter was significantly lower for target trials compared to control ($t(43) = 4.92$, $p < .001$); target trials also had significantly longer non-decision times compared to control trials ($t(43) = -9.36$, $p < .001$). The bias effect is easily interpretable in the context of the task, because the repeated references to the control object were designed to create such a bias.

In the implicatures task, the most noticeable difference between control and target trials was the higher boundary separation but faster drift rate for control trials compared to target trials (boundary separation: $t(44) = 4.21$, $p < .001$; drift: $t(44) = 7.15$, $p < .001$). The slower drift rate for target trials in the implicatures task might have arisen due to their ambiguity (either picture is technically correct): Participants might have been slower to accumulate information to resolve this ambiguity. The fact that the boundary separation is higher for control trials in the implicatures task, combined with this slower drift rate, explains why reaction times do not differ between trial types on the implicatures task, despite lower accuracy for target trials.

Finally, in the negation task, there was no difference in parameters between positive and negative trials. For example, there was no significant difference between drift rates ($t(43) = 0.80$, $p = .43$). This was surprising in light of past work suggesting that adults take longer to respond to negative sentences compared to positive ones (H. Clark & Chase, 1972), especially in context-free tasks such as this one (Nordmeyer & Frank, 2014a). One possibility is that the large number of repetitive trials, or the simplistic and child-friendly stimuli, made this task easier for adults.

Information processing improves with age

Across all three tasks, children at all age groups had different parameter values than adults: higher boundary separation (4-year-olds: $\beta = 1.09$, $p < .001$; 5-year-olds: $\beta = 1.15$, $p < .001$; 6-year-olds: $\beta = .74$, $p < .001$) and longer non-decision times (4-year-olds, $\beta = .18$, $p = .09$; 5-year-olds: $\beta = .18$, $p < .001$; 6-year-olds: $\beta = .10$, $p < .001$), suggesting that children take longer to encode information and need more information to make a decision compared to adults. Children also had significantly slower drift rates compared to adults (4-year-olds: $\beta = -1.63$, $p < .001$; 5-year-olds $\beta = -1.22$, $p < .001$; 6-year-olds: $\beta = -1.13$, $p < .001$), suggesting that children acquire evidence more slowly than adults.

To explore whether these parameters changed significantly throughout these early years, we next focused just on children's data and analyzed age group as a continuous variable. This analysis revealed a significant increase in drift rate across these three years ($\beta = .25$, $p < .001$), as well as significant decreases in separation bias ($\beta = -.18$, $p < .01$) and non-decision time ($\beta = -.04$, $p < .05$). These findings indicate that the speed at which children accumulate evidence and the amount of information that children need to make a decision changes rapidly across early childhood.

Children show different developmental trajectories for inhibition, implicature, and negation

In addition to the general developmental changes in drift rate, non decision time, and boundary separation described in the previous section, children's decision process for each task showed a distinct pattern of development from four to six years of age.

In the inhibition task, 4-year-olds were *less* likely to show a bias towards the incorrect trial compared to adults; this bias appears to get stronger by age 6. Linear models fit to individual parameter values for each task revealed a significant positive interaction between age group and trial type for four-year-olds' bias parameter ($\beta = .14$, $p < .05$), indicating that for four-year-olds the bias parameter for target trials in the inhibition task is higher compared to adult participants. There was no interaction between age group and trial type for five-year-olds ($\beta = .06$, $p = 37$) or six-year-olds ($\beta = -.06$, $p = .31$). These results collectively suggest that while four year olds are *less* biased towards the incorrect response on target trials compared to adults, by age five children are becoming more biased towards the incorrect response on target trials.

In the implicature task, children's drift rates increased developmentally, but the increase in drift rate occurred primarily in control trials. This pattern led to an increasing difference in drift rates between control and target trials across development, driven primarily by an increase in control trial drift rate. Drift rates for target trials remain low even in adulthood, presumably due to the ambiguous nature of these trials. Confirming these findings, linear models fit to individual parameter values for each task resulted in significant positive interactions between age group and trial type at all ages (four-year-olds: $\beta = 0.63$, $p < .01$; five-year-olds: $\beta = 0.80$, $p < .01$; six-year-olds: $\beta = 0.49$, $p < .05$).

The negation task yielded the most striking difference between adults and children. Children had much lower drift rates for target trials compared to control trials. Drift rates on negation trials for four-year-olds were close to zero, suggesting a pattern of general uncertainty for young children on this task. The primary developmental change for this task was an increase in drift rate for target trials, with significant neg-

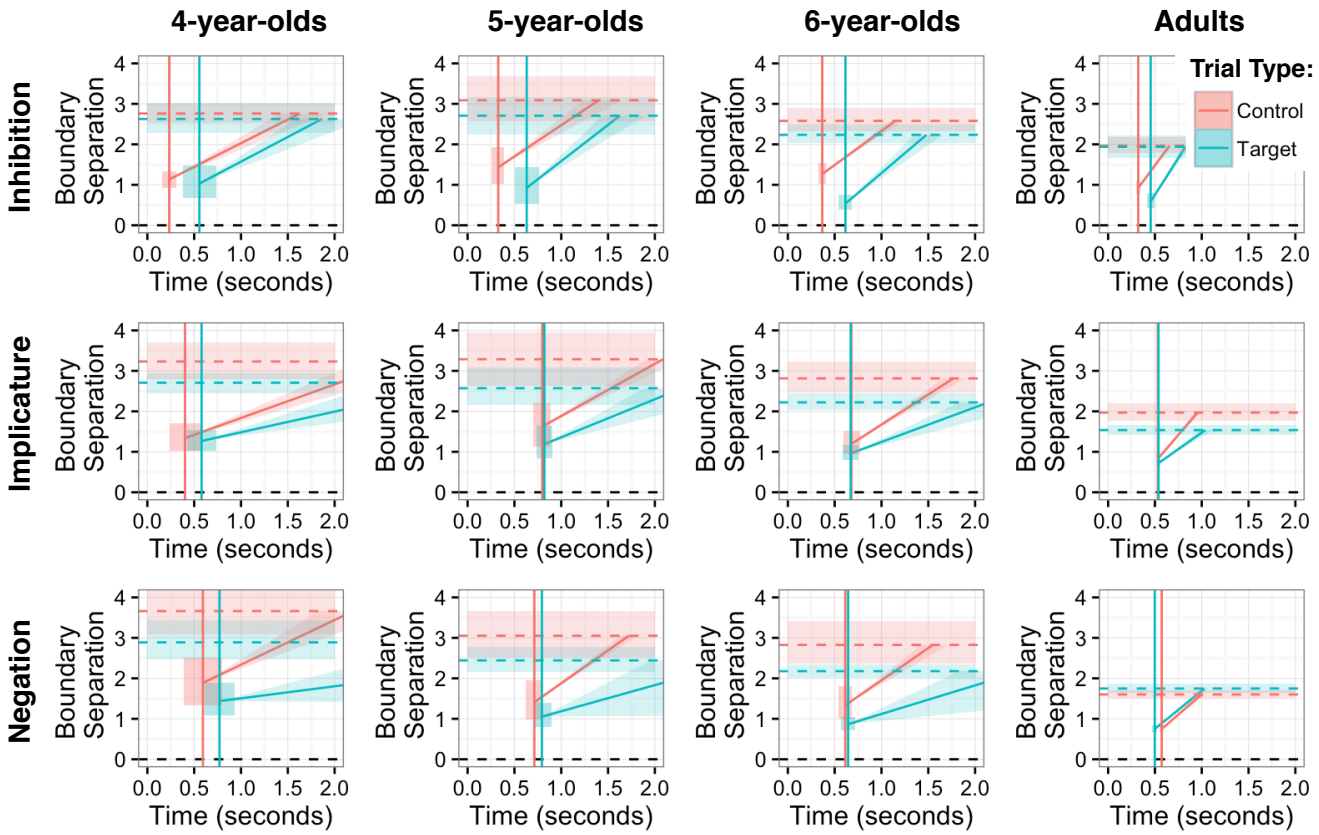


Figure 3: Visualization of the drift diffusion process across the three tasks. The process for control trials is shown in pink, and the process for target trials is shown in green. The dotted black line at zero represents the threshold for making an incorrect decision, and horizontal colored lines represent the boundary separation parameter (i.e., the threshold for making a correct decision). Vertical colored lines represent the non-decision time parameter, and the slope of the decision process (angled line) represents the drift rate parameter. The point where the decision process crosses the non-decision line represents bias $\times$ boundary separation. Ribbons around all lines represent 95% confidence intervals around each parameter.

ative interactions between age group and trial type for four, five, and six-year-olds (4-year-olds, marginally significant: $\beta = -0.53$, $p = .07$; 5-year-olds: $\beta = -1.06$, $p < .001$); 6-year-olds ($\beta = -0.83$, $p < .01$).

Although children struggled on both negation and implicature tasks, children's difficulty on the implicature task appeared to be due to general changes in information processing across development, as indicated by improvement primarily in the control trials of this task. In contrast, children's difficulty on the negation task appeared to be due to unexpectedly low drift rates for negation trials compared to control trials, a difference that disappeared in adulthood. *Why* children appear to struggle so much on the comprehension of negative sentences, despite producing similar sentences at a much younger age, remains an open question.

General Discussion

We created a set of three tasks to explore children's and adults' inhibitory control, implicature processing, and negation processing. Our experiment was designed to test the hypothesis that negation and implicature processing require inhibitory control, and that children's poor performance on these tasks is due to poor inhibitory control. Contrary to our hypothesis, we did not find any evidence of a relationship between performance on the inhibition task and performance on the negation or implicature tasks. Drift diffusion models revealed different patterns of processing on all three tasks, with no evidence that adults or children are biased towards the incorrect answer (as in the inhibition task) on implicature or negation tasks. Our analysis sheds light on the developmental trajectory of children's inhibitory control, negation processing, and pragmatic inference, suggesting that children's low performance on negation and implicature comprehension tasks may be caused by separate processing difficulties.

The current work used a unified procedure which allowed us to directly compare different linguistic and non-linguistic processes. Traditional analyses of reaction time and accuracy can tell us *that* children struggle on these tasks, but not *why* they struggle. The drift diffusion analysis, however, revealed

important processing differences between the three tasks that can help us understand the source of these difficulties. Although our three tasks were nearly identical in appearance—they used the same few trial images and labels and had the same basic instructions to select the correct picture—the decision process to choose the correct picture was very different for each task. For example, the inhibition task was characterized by a bias towards the incorrect answer on target trials, while the implicatures task was characterized by slower drift rates for target trials.

Children's difficulty on the negation and implicature games appears to be due to slow accumulation of information on these tasks. Past work suggests that children typically have slower drift rates compared to adults (Ratcliff et al., 2012), and our current work replicates that finding. In our tasks older children had faster drift rates compared to younger children on all tasks, but this developmental increase in drift rate occurred particularly for control trials on the implicature task and target trials on the negation task. The fact that children's drift rates for negation trials were significantly slower compared to adults suggests that children find it particularly difficult to process the relevant information about negation above and beyond a generally slower processing speed for children compared to adults. Whether the source of this difficulty is an undeveloped semantic understanding of truth-functional negation, or poor phonological processing—for example, perhaps some children simply miss the word "no" entirely—is a topic for future research.

A final novel aspect of this work is the use of the drift diffusion model with preschool children. Although some past work has explored children's information processing using the drift diffusion model (Ratcliff et al., 2012), that work tested second graders at the youngest. Here we expand on that work by replicating their findings with a group of much younger children. Our success using the drift diffusion model with data from such young children opens doors for future work exploring the development of children's information processing.

Acknowledgments

Special thanks to Stephen Powell, Veronica Cristiano, and Rachel Walker for their help with data collection, and to the parents, children, and staff at the San Jose Children's Discovery Museum. This work was supported by NSF grant BCS #1456077.

References

Bloom, L. (1970). *Language development: Form and function in emerging grammars.* New York, NY: MIT Research Monograph, No 59.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Clark, H., & Chase, W. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*, 2037–2078.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "do as I say, not as I do". *Developmental Psychobiology*, *29*, 315–334.

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: evidence from real-time spoken language comprehension. *Developmental psychology*, *45*(6), 1723.

Kim, K. (1985). Development of the concept of truth-functional negation. *Developmental Psychology*, *21*, 462.

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two-and four-year-olds learn to adapt referring expressions to context: Effects of distracters and feedback on referential communication. *Topics in cognitive science*, *4*(2), 184–210.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*(03), 403–422.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Nordmeyer, A. E., & Frank, M. C. (2014a). A pragmatic account of the processing of negative sentences. In *Proceedings of the 36th annual meeting of the cognitive science society.*

Nordmeyer, A. E., & Frank, M. C. (2014b). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25–39.

Pea, R. (1980). The development of negation in early child language. In D. Olson (Ed.), *The social foundations of language and thought* (pp. 156–186). New York, NY: Norton New York.

Pea, R. (1982). Origins of verbal logic: Spontaneous denials by two-and three-year olds. *Journal of Child Language*, *9*, 597-626.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child development*, *83*(1), 367–381.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176–190.

Yoon, E. J., Wu, Y. C., & Frank, M. C. (2015). Children's online processing of ad-hoc implicatures. In *Proceedings of the 37th annual meeting of the cognitive science society.*

Answering Causal Queries about Singular Cases

Simon Stephan (simon.stephan@psych.uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Queries about singular causation face two problems: It needs to be decided whether the two observed events are instantiations of a generic cause-effect relation. Second, causation needs to be distinguished from coincidence. We propose a computational model that addresses both questions. It accesses generic causal knowledge either on the individual or the group level. Moreover, the model considers the possibility of a coincidence by adopting Cheng and Novick's (2005) power PC measure of causal responsibility. This measure delivers the conditional probability that a cause is causally responsible for an effect given that both events have occurred. To take uncertainty about both the causal structure and the parameters into account we embedded the causal responsibility measure within the structure induction (SI) model developed by Meder et al. (2014). We report the results of three experiments that show that the SI model better captures the data than the power PC model.

Keywords: causal inference, generic causation, singular causation, actual causation, causal responsibility, causal attribution, Bayesian modeling

Imagine that you wake up one morning and recognize that you are haunted by a mean twinge in your head. You also know that you drank too many glasses of wine last night. Now the question arises whether your behavior last evening is causally responsible for your headache this morning. This causal query targets a *singular* instance in which one event at a specific spatio-temporal location may have caused another event that followed. The general problem with singular causal queries is that a co-occurrence of the particular events by itself does not guarantee causation. It may just be a *coincidence*. How confident can you be that this singular case of having drunk wine is the cause of your headache?

One important source that should influence our confidence is past knowledge about the contingency between drinking wine and headache. This is *generic* causal knowledge. It could either refer to cases of presence and absence of drinking and headache in an observed sample of people, or, even better with respect to the example given above, to a sample of these events in your life. The contingency in your life provides the best estimate for a generic causal relation between drinking and headache in your body.

However, knowing that there is a generic causal relation does not necessarily imply that a singular co-occurrence of the target events is causal. Unless the generic relation is deterministic, the co-occurrence may still be a coincidence. Thus, our causal judgment about singular causation must take this possibility into account.

Psychological research on causal inference adheres to two different theoretical frameworks to characterize the reasoning processes (see Waldmann, in press; Waldmann & Hagmayer, 2013; Waldmann & Mayrhofer, in press). One approach assumes that causal knowledge is grounded in knowledge about causal dependencies gleaned from observed contingencies. According to this approach, causes are *difference makers* that raise or lower the probability of an effect (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005; Meder, Mayrhofer, & Waldmann, 2014).

A different approach assumes that causal knowledge is based on a search for mechanisms and processes linking causes and effects (e.g., Ahn, Kalish, Medin, & Gelman, 1995). The two approaches need not be incompatible. Often mechanism knowledge is based on generic information about causal chains. Specific mechanisms are then simply instantiations of generic chain knowledge.

Mechanism knowledge is often not available. But even in cases in which information about intervening variables of a causal chain is in fact accessible, the question arises again how we should distinguish causation from coincidence. The joint occurrence of all elements of a chain certainly makes the possibility of a coincidence extremely unlikely, but it is in principle still a possibility. Thus, the general problem of how we should apply generic knowledge to singular cases still needs to be addressed, regardless of whether we solve this problem for a direct causal link or a causal chain.

In the present paper, we investigated how people exploit generic causal information when estimating singular causation. We here focus on cases for which it is known that both the cause event and the effect event are present, but the question needs to be answered how much confidence we can have that the co-occurrence is due to causation and not a coincidence.

Cheng and Novick (2005) have proposed a measure of *causal responsibility* that helps to answer this question. The measure is based on the assumptions of power PC theory (Cheng, 1997). It delivers the conditional probability that a cause event c is causally responsible for an effect event e given that a reasoner knows that both have occurred, $P(c \rightarrow e|c,e)$. We think that this quantity underlies judgments of singular causation when only generic information is available. One shortcoming of this measure is that it does not incorporate the reasoners' *uncertainty* concerning the underlying causal structure and the corresponding parameters (see Griffiths & Tenenbaum, 2005). Therefore, Holyoak, Lee, and Lu (2010) have proposed a causal responsibility measure that also takes parameter uncertainty into account. We go one step further here, and test a model of causal responsibility that is sensitive to both parameter uncertainty and uncertainty about the existence of the causal link between factors C and E. This

model is based on the structure induction (SI) model of diagnostic reasoning developed by Meder et al. (2014). We present three experiments in which we empirically tested the power PC theory of causal responsibility against our SI model of singular causation.

The Power PC Model of Causal Responsibility

Causal power in Cheng's (1997) theory can be understood as the probability that the target cause brings about the effect in the absence of all alternative causes of the effect. Fig. 1 (right) shows the basic causal model assumed by the power PC theory. The default assumption is that the target effect E can be independently produced by either the observed cause C with causal power (or strength) w_c or by alternative unobserved causes A with the power (or strength) w_a. It is further assumed that C and A occur independently and do not interact. Thus, C and A combine through a *noisy-OR* gate (see Griffiths & Tenenbaum, 2005; Pearl, 1988).

Based on the power PC framework, Cheng and Novick (2005) developed a measure of causal responsibility. Formally, causal responsibility is the proportion of occurrences of the effect due to the target cause C. Cheng and Novick (2005) presented formalizations of different kinds of causal responsibility. Here we are interested in how to answer the question whether an observed event c was causally responsible for an observed event e in cases in which both c and e were observed. Under the default assumptions of power PC theory, Cheng and Novick (2005) showed that this quantity is given by

$$P(c \rightarrow e|c,e) = \frac{P(c) \cdot w_c}{P(c,e)} = \frac{P(c) \cdot w_c}{P(c) \cdot P(e|c)} = \frac{w_c}{P(e|c)}, \quad (1)$$

where $P(c)$ equals b_c in Fig. 1, which denotes the base rate of C, and $P(c,e)$ denotes the joint probability of cause and effect. Since $P(c,e)$ can be rewritten as the product of $P(c)$ and the predictive probability $P(e|c)$, causal responsibility given the joint occurrence of c and e equals the causal power of c divided by the predictive probability of e given c.

The power PC theory only considers the model S_1 depicted in Fig. 1 (right) and uses maximum likelihood estimates for the parameters. It does not take into account uncertainty about the size of the parameters, nor about the causal structure (see Griffiths & Tenenbaum, 2005). Thus, it does not consider the possibility, depicted as S_0 in Fig. 1, that there is no causal arrow from C to E, meaning that all co-occurrences are due to coincidence.

A consequence of maximum likelihood estimates is that the theory predicts maximal values of causal responsibility for any contingency where the target cause is necessary (i.e., a table with an empty C cell) no matter how rarely co-occurrences of cause and effect are. Consider the two examples of such cases depicted in Fig. 2. In the left table, the effect has only occurred four times when the cause event was present. In contrast, in the right table the effect occurred sixteen times when the cause event was present. By applying the equations, one can see that the prediction of maximal responsibility is a consequence of w_c being equal to $P(e|c)$ for any table with an empty C cell.

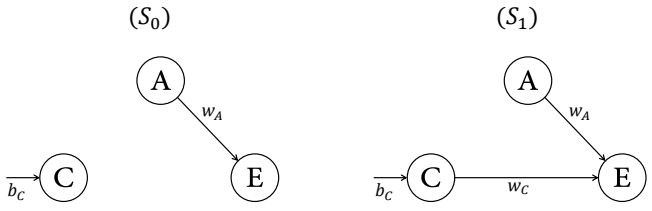


Figure 1: The two causal structures in the structure induction model. In S_0, no causal relationship exists between C and E. In S_1, C and E are causally connected. Node A represents unobservable background causes of E. The parameter b_c denotes the base rate of the cause, and w_c and w_a the causal powers of C and A, respectively.

The Structure Induction Model of Singular Causation

Contrary to the power PC theory, the structure induction (SI) model assumes that reasoners might be uncertain about both the underlying causal model and the parameters (Meder et al., 2014). Originally, the model was developed to model diagnostic inferences. Here, instead, we use the framework to model judgments about singular causation. Our key claim is that subjects assess singular causation by estimating causal responsibility within the SI model.

Causal queries about simple causal models with a single cause and a single effect are modelled by the SI theory as a Bayesian inference problem over the two mutually exclusive causal structures shown in Fig. 1 (see also Griffiths & Tenenbaum, 2005). The model formalizes the assumption that reasoners are uncertain about which of the two models underlies the data; they are also uncertain about the size of the parameters. We will briefly summarize the different computational steps of the model.

Parameter Estimation

In light of the possibility that either S_0 or S_1 might underlie the data, the model estimates the base rate and power parameters b_c, w_c, and w_a, separately for each causal structure (see Fig. 1). To express uncertainty, distributions of the parameters, rather than maximum likelihood point estimates, are inferred. According to Bayes' rule, the posterior probability distributions for the parameters of each model given the data, $P(w|D)$, is proportional to the likelihood of the data given the set of parameters w, weighted by the prior probability of the structure:

$$P(w|D) \propto P(D|w) \cdot P(w) \quad (2)$$

$P(D|w)$ denotes the likelihood of the data given the parameter values for b_c, w_c, and w_a. $P(w)$ represents the prior joint

probability of the parameters which we set to flat, uninformative *Beta*(1, 1) distributions.

Structure Estimation

The SI model separately derives the posterior probabilities for each causal structure. Applying Bayes' rule, the posterior probability for a causal structure is proportional to the likelihood of the data given the causal structure, weighted by the structure's prior probability:

$$P(S_i|D) \propto P(D|S_i) \cdot P(S_i) \tag{3}$$

$P(D|S_i)$ denotes the likelihood of the data given a particular structure, which is the integral over the likelihood function of the parameter values under the particular structure. $P(S_i)$ represents the prior probability of the structures. The model initially assumes that both structures have equal priors, i.e., $P(S_i) = 1/2$. When data are available, the posterior for a causal structure varies systematically with the observed contingency: the higher the contingency, the more likely S_1 becomes.

Causal Responsibility for Each Structure

Having estimated the parameters and the posteriors of the structures, the model computes causal responsibility separately for each parametrized structure using Equation 1. Under a noisy OR-parametrization Equation 1 can be rewritten as

$$P(c \rightarrow e|c,e) = \frac{w_c}{P(e|c)} = \frac{w_c}{w_c + w_a - w_c \cdot w_a}. \tag{4}$$

According to S_0, there is no causal connection between C and E, and any co-occurrences of c and e are coincidences. Hence, $P(c \rightarrow e|c,e) = 0$ for S_0. For S_1, Equation 4 is applied. The estimation of causal responsibility is then derived by integrating over the parameter values.

Deriving a Single Value

The final output of the model is a single estimate of causal responsibility through integrating out the causal structures by summing over the derived values of $P(c \rightarrow e|c,e;S_i)$ for each structure weighted by each structure's posterior probability:

$$P(c \rightarrow e|c,e;D) = \sum_i P(c \rightarrow e|c,e;S_i) \cdot P(S_i|D). \tag{5}$$

The incorporation of structure and parameter uncertainty by the SI model leads to systematic deviations from the predictions of the power PC model (see Fig. 2). For an illustration, consider the left contingency table depicted in (a). Although the effect never occurred in the absence of the cause, only four co-occurrences are observed. This relatively low frequency of co-occurrence is the reason for the relatively high probability of S_0. Hence, based on this information, it seems reasonable to be cautious about the causal relation of a singular co-occurrence. By contrast, consider the right table in (a). Here again, the effect never occurred in the absence of C but it occurred frequently in its presence. In this case, the data suggest a strong generic causal relationship between the factors. Now, when presented with a singular case, relatively high confidence in a case of actual causation is to be expected.

Experiment 1

The experiment tests the SI model against the power PC theory as an account of how subjects answer singular causal queries. Fig. 2 shows the predictions of the models for the two data sets that we used. The power PC model predicts invariance whereas our SI model of singular causation predicts that judgments should vary. This is because the likelihood that the observed data pattern was produced by structure S_1 varied between the two different contingencies. We also wanted to address a second question: Do subjects differentiate between generic and singular queries? To test this, we compared a generic with a singular query in the experiment.

Design, Materials, and Procedure

Sixty-one subjects completed this online experiment for monetary compensation. Fifty-six subjects (mean age = 29.16 years, *SD* = 8.4 years, 28 were female) passed our embedded attention check and were included in the analyses. The contingencies (see Fig. 2) were varied between subjects.

We used a fictitious story of a single sea devil living in an aquarium. Our subjects were asked to assume they were biologists being interested in whether noise causes the fish's antenna to flash. Subjects read they should imagine having run a single-case study with one fish to test this hypothesis. We then instructed our participants that they will see the results of this study. They read that the fish was placed in front of a loudspeaker twenty times and that the loudspeaker was off at first and activated subsequently to assess the fish's reaction.

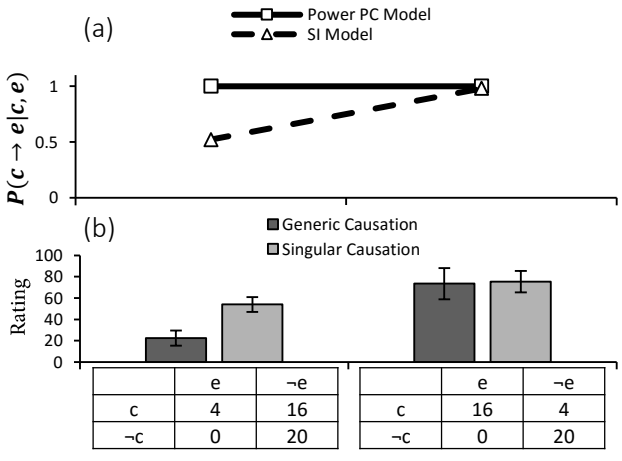


Figure 2: Predictions of the power PC and the SI model of singular causation are depicted in (a), the results in (b). Dark and light bars show means (95% CI) of the generic and singular causation ratings, respectively.

Having read the instruction, subjects saw the results of the observations arranged in a table with four columns and five rows. Below each depiction of the fish, located in front of a tiny loudspeaker, a little yellow scrip was placed that indicated the respective trial number. The initial screen showed the state of the fish as the loudspeaker was off ($\neg$ c). Next, we presented the results as the loudspeaker was activated (c), symbolized by tiny sound waves. A yellow color of the fish's flash bulb indicated the flashing of the antenna.

To test intuitions about generic causation we asked subjects how appropriate it is to say (on a rating scale from 0 to 100) that noise is a cause of the flashing of the sea devil's antenna. The corresponding question targeting singular causation asked subjects to focus on the first trial of the observations in which the fish's antenna had flashed upon noise exposure. Subjects were requested to indicate (again using a scale ranging from 0 to 100) how appropriate it is to say that the noise had caused the flashing of the fish's antenna in this particular trial. Additionally, we also asked subjects about a trial in which the antenna did not flash upon noise exposure. Furthermore, we asked them to make a predictive judgment concerning a hypothetical new trial. Here, we asked how likely it is that the fish would flash its antenna again.

Results and Discussion

The results can be seen in Table 1 and Fig. 2. Fig. 2 shows that participants judged the generic-level causal relationship between noise and antenna flashing differently for the two contingencies. On average, ratings were higher in the high contingency condition compared to the low contingency condition. As predicted by the SI model, the singular cause ratings were also different in the two conditions.

A 2 (contingency) $\times$ 4 (type of rating) mixed ANOVA with the second factor being varied within subject yielded a significant main effect of contingency, $F(1,54) = 82.00$, $p < .001$, a significant main effect of rating, $F(3,162) = 53.13$, $p < .001$, and also a significant interaction between contingency $\times$ rating, $F(3,162) = 15.74$, $p < .001$. A planned contrast comparing the singular causation ratings was also significant, $t(54) = 2.55$, $p = .01$, confirming that the ratings were higher in the high contingency condition. This difference is not predicted by the power PC model but it is predicted by the SI model.

An interaction contrast comparing the difference between the generic and the singular ratings was significant, $t(54) = 3.60$, $p < .001$. As Fig. 2 shows, this finding is due to the fact that the mean rating obtained for the singular causation question was higher than the generic causation rating in the low contingency condition. While this indicates that participants indeed conceptualized the two causal queries differently it seems that subjects answered the generic question with a causal strength rather than a structure estimate. This may also explain why there was no difference between generic and the singular ratings in the high contingency condition. As can be seen in Table 1, the additional ratings for the predictive probability were in line with the empirical predictive probability obtained in the contingency tables. Interestingly, singular ratings differed from zero in both groups for the trial in which the fish did not show a flashed antenna after noise exposure. These ratings were also higher in the high contingency condition than in the low contingency condition. This seems to reflect an unpredicted influence of generic knowledge, but it should be noted that the ratings were the lowest in the set.

In sum, the results favor our SI model of singular causation. We obtained the predicted slope of the ratings that contradicts the power PC model predictions. Furthermore, the results obtained in the low contingency condition show that subjects differentiated between a generic and a singular causal query.

Table 1. Mean ratings (*SE* of the mean) obtained for the different questions in the experiment.

	condition	
	contingency: low	contingency: high
generic causation	22.40 (3.43)	73.55 (3.43)
singular causation (e, c)	54.00 (7.12)	75.48 (4.95)
singular causation (e, $\neg$c)	10.40 (3.58)	19.03 (5.36)
predictive probability	20.40 (3.13)	76.13 (2.48)

Experiment 2

In Experiment 1 we used data sets in which the cause appeared necessary for the effect. To test our model for data sets in which all event types exist, we used in the present experiment the three contingency tables shown in Fig. 3. Again, the power PC model predicts invariance. This time it predicts values of 0.83, whereas the SI model predicts a slope. For this experiment, we also wanted to compare the SI model with a Bayesian variant of the power PC model (Holyoak et al., 2010) that takes into account parameter uncertainty but not structure uncertainty. For this model, we also used uninformative priors over the parameters. As can be seen, the predictions of this model are very similar to the predictions of the power PC model but slightly lower on average.

A second change was that the generic information did not refer to a time series of trials involving a single individual, but a sample of different individuals. Thus, we were interested in testing how subjects use generic information about a sample of different fish to respond to a causal query about a singular case.

Design, Materials, and Procedure

Thirty-eight participants (mean age = 31.23 years, *SD* = 8.90 years, 14 were female) completed this online experiment. The data sets were presented to each subject.

In this experiment, all forty observations were shown in random order on the screen in a table with five columns and eight rows. Subjects should assume they were biologists who had studied the influence of chemicals on gene mutations in populations of sea devils. They were instructed that they will see the results of these studies. Subjects also read that the

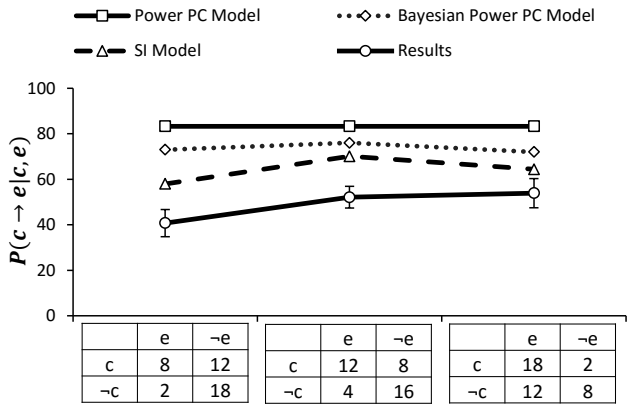


Figure 3: Model predictions and results (95% CI) for the different data sets used in Experiment 2.

chemical had been injected in half of the sample in each study and that the whole sample had later been tested for mutation. We used differently colored circles (red vs. blue) to indicate the presence vs. absence of gene mutations; a black margin around circles indicated chemical treatment. Finally, we instructed our participants that their task will be to provide a judgment concerning a single case.

For each data set, participants were asked to imagine a new individual fish which had ingested the chemical and also had the mutation. This time, we asked for the probability (on a scale from 0 to 100) that the chemical caused the gene mutation in this single case.

Results and Discussion

As can be seen in Fig. 3, ratings differed across the three data sets. A repeated-measure ANOVA with "data set" as within-subject factor was significant, $F(2,74) = 4.19$, $p = .02$. This main effect was due to the ratings for the first data set being different from those for the second and the third data set, $t(37) = 2.77$, $p < .01$. The ratings for the second and the third data set did not differ, $t(37) = 0.39$. The dampening of the upward trend across the three data sets is consistent with our SI model, although the model predicts a slight downward trend between the second and third data set. In contrast to what we observed in the first experiment, the singular causation ratings for all three data sets were lower than predicted by the SI model. One explanation for this finding may be that the type of data presentation that we used made it hard for participants to grasp the contingencies precisely, because all observations were presented on the screen in random order and also in combination with abstract symbols.

In sum, the results again favor our SI model of singular causation over the power PC model and also over a Bayesian variant of the power PC model as an account of how subjects respond to causal queries about singular causation. The present experiment further demonstrates that not only time series data but also data about samples of different individuals are used to derive a prediction for a singular case.

Experiment 3

In the first two experiments we dissociated the SI model of singular causation from the power PC model by using data sets for which the power PC model predicts invariance. To obtain additional evidence for the validity of the SI model we conducted an experiment with data sets for which the SI model predicts invariance but the power PC model does not. We obtained invariant predictions of the SI model by counteracting a slight decrease in contingency across the two contrasted data sets (see Fig. 4) with an increase of sample size.

Apart from singular causal inferences, we also tested whether participants would infer a higher probability for the existence of a generic causal relationship (i.e., for S_1) for the larger sample (Fig. 4, right), which is predicted by the SI model for generic queries. Our model predicts an interaction effect between the type of question and data set.

Design, Materials, and Procedure

All subjects saw both data sets. The type of question (confidence in generic vs. singular causation) was manipulated between subjects. 101 subjects participated in the online study (mean age = 31.40 years, *SD* = 11.10 years, 43 were female) and provided valid data.

Participants read that they should imagine to be biologists who tested the influence of the chemical "Acrinazyl" on the expression of the gene ASPM in mice in an experiment using a sample of twenty test animals. They were instructed that one half of the sample served as a control group. We presented the two halves of the sample separately on the same screen. Different colors indicated the status of ASPM. In the generic condition, we asked subjects to indicate (on a slider ranging from 0 to 100) how confident they are that the chemical does indeed raise the probability of ASPM expression. We were hoping that this test question would less ambiguously refer to generic causal relations than the one we used in Experiment 1. In the singular condition, subjects were instructed that they should imagine having picked out a single mouse from the Acrinazyl group with ASPM being expressed. They were asked how much confidence they had that it was indeed

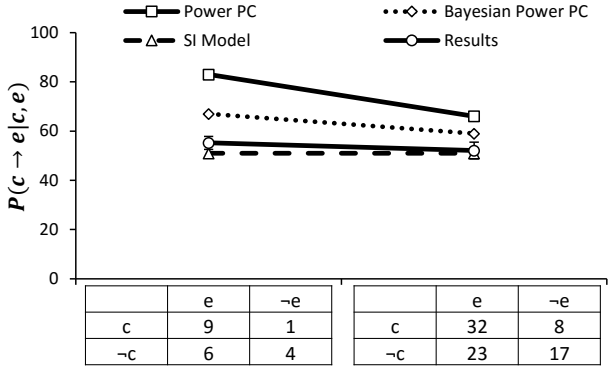


Figure 4: Model predictions and results (95% CI) for the different data sets used in Experiment 3.

the chemical that caused the gene expression in this particular mouse. Subjects indicated their confidence using a slider ranging from 0 to 100.

After participants had provided these ratings they read that they had conducted a second experiment with a larger sample size. We then showed them the results of the second study and asked them to re-assess their ratings in light of the results of this second study.

Results and Discussion

Fig. 4 shows that the singular ratings are well captured by the SI model. Moreover, as predicted by the SI model, the mean ratings for the generic causal question increased between data sets (M = 63.54, SE = 1.00; M = 68.63, SE = 1.00). A mixed ANOVA confirmed our predictions with a significant main effect for causal question, $F(1,99) = 7.42$, $p < .01$, and a significant interaction between question type and data set, $F(1,99) = 6.60$, $p = .02$. Furthermore, there was no difference between the singular causal ratings, $t(46) = 1.34$.

Overall, Experiment 3 was in line with the predictions of our SI model of singular causation. We also demonstrated again that subjects treat singular and generic causation queries differently.

General Discussion

We tested a new model of how people respond to queries about singular causation. Simply observing two consecutive events at a specific space-time location does not suffice. The question needs to be answered whether this co-occurrence manifests a causal relation or merely a coincidence. We argued that one relevant source of knowledge are generic causal relations. However, knowing, for example, that smoking generally increases the risk of lung cancer does not imply that a specific cancer patient who has smoked throughout her life has actually contracted the disease because of this risk factor. A coincidence is still possible.

One strategy that has been suggested in the literature is that judgments about singular causation should rely on unveiling causal mechanisms: observations of tar in the lung and genetic alterations in a cancer patient would strengthen the claim of singular causation. However, observing chains of singular events only helps with the question of singular causation if there is reason to assume that they are causally related. Thus, generic knowledge is necessary here as well (see also Danks, in press).

Since studying chain-like mechanisms does not fully solve the problem of how generic and singular causation are related, we here began with the simplest case with one cause and one effect event. In three experiments we showed that subjects used contingency information from both time series data of an individual and from group data when making judgments about singular causation.

We also showed that subjects differentiated between generic and singular causation. The responses to queries about singular causation were best explained by a variant of the SI model from Meder et al. (2014) that computes an estimate of causal responsibility (Cheng & Novick, 2005). Unlike its main competitors, the SI model is sensitive to the uncertainty of both the causal structures and their parameters.

Our research is just a first step. We have compared the SI model with two power PC models. However, we kept the priors uninformative and did not model different types of priors (e.g., Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008).

Another question that we want to address in future research is how subjects assess generic causal relations in time series versus group samples. Finally, it would be interesting to broaden the scope of the SI model by applying it to more complex causal models, such as causal chains, to gain insights in the important role that mechanism knowledge plays in singular causal inferences.

Acknowledgments

We thank Jonas Nagel, York Hagmayer, and Ralf Mayrhofer for helpful comments.

References

Ahn, W. K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, *112*, 694–706.

Danks, D. (in press). Singular causation. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning.* New York: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*, 702–727.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–982.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277–301.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan-Kaufmann.

Waldmann, M. R. (in press). *Oxford handbook of causal reasoning*. New York: Oxford Universtiy Press.

Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 733–752). New York: Oxford University Press.

Waldmann, M. R., & Mayrhofer, R. (in press). Hybrid causal representations. In B. Ross (Ed.), *The psychology of learning and motivation.* New York: Academic Press.

The Impact of Granularity on the Effectiveness of Students' Pedagogical Decision

Guojing Zhou, Collin F. Lynch, Thomas W. Price, Tiffany Barnes, Min Chi
Department of Computer Science
North Carolina State University
{gzhou3,cflynch,twprice, tmbarnes,mchi}@ncsu.edu

Abstract

In this study we explored the impact of student versus tutor pedagogical decision-making on learning. More specifically, we examined what would happen if we let students decide how to handle the next task: to view it as a worked example or to solve it as a problem solving. We examined this impact at two levels of task granularity: problem vs. step. This 2×2 study was conducted on an existing Intelligent Tutoring System (ITS) called Pyrenees. 279 students were randomly assigned to four conditions and the domain content and required steps were strictly controlled to be equivalent across four conditions: all students used the same system, followed the same general procedure, studied the same training materials, and worked through the same training problems. The only substantive differences among the four conditions were decision agency {Student vs. Tutor} and granularity {Problem vs. Step}. That is: who decided to present an example or to solve a problem; and was the decision made problem-by-problem or step-by-step? Our results showed that there was a significant interaction effect between decision agency and granularity on student learning and a significant main effect of granularity on time on training. That is, step level decisions can be more effective than problem level decisions but the students were more likely to make effective pedagogical decisions at problem level than step level. In general, on both problem and step levels, the students were significantly more likely to decide to do problem solving rather than study it as a worked example.

Keywords: pedagogical policy, student-centered learning, problem solving, faded worked example, granularity

Introduction

Human one-on-one tutoring is one of the most effective way to improve student learning (Bloom, 1984). Intelligent Tutoring Systems (ITSs) are computer systems that mimic aspects of human tutors and have also shown to be successful as well (Koedinger, Anderson, Hadley, & Mark, 1997; Vanlehn, 2006). Most ITSs are tutor-centered. The tutor is responsible for selecting the next action to take at any given time. Each of these decisions affects student's successive actions and performance. In the learning literature, the skills used to make such decisions are generally referred to as *pedagogical skills*. More formally, Chi et al. defined pedagogical skills are those "involve skillful execution of tactics, such as giving explanations and feedback, or selecting the appropriate problems or questions to ask the students" (M. T. H. Chi, Siler, & Jeong, 2004). Most ITSs generally employ fixed pedagogical policies that do not adapt to users' needs, or they rely on hand-coded rules that seek to implement existing cognitive or instructional theories that may not have been well-evaluated. For example, in most ITSs students are asked to *solve* a series of training problems while research showed that studying worked examples can be more effective than solving problems and the former generally takes much less time (Sweller & Cooper, 1985; McLaren & Isotani, 2011).

On the other hand, much previous research showed that it is desirable for student to experience a sense of control over their own learning (Harackiewicz, Sansone, Blair, Epstein, & Manderlink, 1987). People are likely to persist at doing constructive things, like learning, exercising, quitting smoking, or fighting cavities, when they are given the choice and when they can make decisions. Letting students make decisions during the tutorial process should make them feel that they are actively directing their own learning process and not just passively following it. Therefore, in this paper we provided the students with two different yet both reasonable choices and let them decide how they want to solve the problem next. So the question is: can students make effective pedagogical decisions that would promote their learning?

Moreover, we investigated the impact of students' decisions across two levels of granularity: problem versus step. Tutoring in domains such as math and science can be viewed as a two-loop procedure (Vanlehn, 2006). In the outer loop, the tutor makes task or problem-level decisions such as deciding what problem to solve next, while the inner loop controls step level decisions such as whether or not to give a hint. In educational literature, 'steps' often refer to the application of a major domain principle such as Newton's Third Law of Thermodynamics. Solving a complete problem generally involves applying many individual principles in a logical order.

In theory, problem-level decisions are at a larger grain size and thus once students make one 'big' decision, they can focus on comprehending an example or solving a problem. However, such "big" decision might not be very sensitive to students' specific moment-by-moment needs. For example, a student faces difficulty with a single principle then a complete worked example may rob them of the chance to exercise other skills. When making step-level decisions, by contrast, students may be better able to tailor their decisions to their immediate needs and current knowledge level. However, making many fine grain decisions can be more frustrating than beneficial over time.

In order to investigate the effectiveness of students' pedagogical decision-making at both levels of granularity, it is necessary to separate the pedagogical decisions from the instructional content, strictly controlling the content so that it is equivalent for all participants. To strictly control the content to be equivalent, 1) we used an ITS which provides equal support for all learners; and 2) we focused on tutorial decisions that cover the same domain content at both problem and step levels, in this case Worked Examples (WE) versus Problem-Solving (PS). In WE, student were given a detailed example showing the expert solution for the problem or were shown

the best step to take given their current solution state. In PS, by contrast,the students were tasked with solving the same problem using the ITS or completing an individual problem-solving step. While engaging students in decision-making within an ITS is not novel, prior researchers have generally focused on letting students dictate content by letting them decide what problem they wish to solve but not how they wished to solve it (Koedinger et al., 1997). So as far as we know, no prior research has investigated pedagogical decision-making independently of content selection.

In short, our primary research question is: will the granularity of the pedagogical decisions have an impact on the effectiveness of students' pedagogical decision-making? To investigate this question we will compare students' pedagogical decisions against tutor's decisions.

Background

WE/PS, vs. FWE

A number of researchers have examined the impacts of problem-level PS, problem-level WE, vs. *Faded Worked Example* (FWEs) (Renkl, Atkinson, Maier, & Staley, 2002; Schwonke et al., 2009; Najar, Mitrovic, & McLaren, 2014; Salden, Aleven, Schwonke, & Renkl, 2010). FWEs interleave problem-solving steps with worked example steps within a single problem. Renkl et al. compared WE-PS pairs with FWE using a fixed fading policy (Renkl et al., 2002). In that study the number of example steps and problem-solving actions were strictly equal between the conditions. They found that FWEs with the fixed fading policy significantly outperformed the WE-PS pairs. They found no significant time-on-task differences between the two groups. Schwonke et al. compared FWE with a fixed fading policy to tutored PS (Schwonke et al., 2009). Over the course of two studies, they found no significant differences between the two conditions in terms of their learning outcomes. However the FWE group spent significantly less time on task than the tutored PS group. Najar and colleagues compared FWE with an adaptive fading policy to WE-PS pairs. They found that the FWE condition significantly outperformed the WE-PS condition in their learning outcomes and spent significantly less time on task (Najar et al., 2014). Finally, Salden et al. compared three conditions: FWE with a fixed fading policy, FWE with an adaptive fading policy, and PS-only (Salden et al., 2010). They found that the adaptive FWE group outperformed the fixed FWE who, in turn, outperformed PS-only. They found no significant time-on-task differences among three groups.

Thus prior researchers have shown that FWE with *effective* pedagogical polices can outperform fixed WE-PS pairs. It has also been shown that the former may require significantly less time on task than the latter. However all of these studies relied on hand-coded tutor pedagogical polices whereas in this study, we investigated whether students can make effective pedagogical decisions on whether to do PS or study a WE at either problem level or step level.

Students Pedagogical Decision on ITS

Prior research on problem-level decision-making has primarily focused on the impact of letting the students dictate content, e.g which problem to solve but not let students to decide how, e.g. WE vs. PS. The results for student step-level pedagogical decision-making are inclusive. Aleven & Koedinger studied students' help-seeking behaviors in the Cognitive Tutor (Aleven & Koedinger, 2000) where tutor permits students to request help when they do not know what step to take next. Help is provided via a sequence of hints that progress from general top-level hints that prompt the student to consider a principle or variable, to bottom-out hints that tell them exactly what action to take. They found that students do not always have the necessary metacognitive skills to know when they need help. They tend to wait too long before requesting information, and then focus only on applying the bottom-out action rather than processing the top-level conceptual guidance. Roll et al. by contrast examined the relationship between students' help-seeking patterns and their learning (Roll, Baker, Aleven, & Koedinger, 2014). They found that asking for help on challenging steps was generally productive while help abusing behaviors were correlated with poor learning.

Therefore prior research on students' help-seeking suggests that the students can benefit substantially from effective pedagogical decision-making. Yet they often lack the metacognitive skills that are required to do so. On the other hand, help in ITSs is generally provided on demand, and better-performing students are less likely to ask for it. Thus some students may simply never need to do so. In this study we controlled for this possible conflict by focusing on WE/PS decisions, and by examining both problem and step-level decision-making. This allows us to evaluate *all* students' decision-making, not just the lower-performers and help-abusers. It also allows us to investigate the impact of granularity on student learning outcomes.

Our Approach

Previous studies on problem-level decision-making, PS vs. WE, mainly employed some fixed pedagogical polices (either WE-PS or PS-WE) and prior studies on step-level decision-making, FWE, either used fixed fading polices or relied on hand-coded adaptive polices. With adaptive policies, the system decides whether the next step is WE or PS based on a realtime assessment of the student's concept mastery. For example, students may be asked to solve steps involving the same concepts repeatedly until they demonstrate mastery and then such steps would be faded away by presented as WEs only. However there is no clear consensus on how or when students should be given a WE, nor how the faded policy should change on each level.

Therefore in this study we will investigate the impact of students' pedagogical decisions on learning by comparing students' decisions to tutors' **random** decisions at either problem or step level in order to avoid the impact of possibly misguided pedagogical policies. This study is 2 {Student,

Tutor} $\times$ 2 {Problem, Step} design with four conditions:

1. $Stud_{Prob}$: problem-level student decisions.
2. $Stud_{Step}$: step-level student decisions.
3. Tut_{Prob}: problem-level random tutor decisions.
4. Tut_{Step}: step-level random tutor decisions.

All students in this study were given the same problems in the same order. We compared the four groups using pre- and post-tests as well as their time on task.

Methods

Participants

This study was conducted in the undergraduate Discrete Mathematics course at the Department of Computer Science at North Carolina State University in the Fall of 2015. 279 students were enrolled in the course and this study was their *final* homework assignment. The students had two weeks to complete it and they were graded based upon their effort in completing the assignment, not their post-test scores.

Conditions

The students were assigned to the four conditions via balanced random assignment based upon their course section and performance on the class mid-term exam. Since the two tutor-random decision groups were already compared in our prior study (Zhou, Price, Lynch, Barnes, & Chi, 2015) and the primary goal of this work is to examine the nature and effectiveness of students' pedagogical decision-making, we assigned twice more students to the two student-decision groups, $Stud_{Prob}$ & $Stud_{Step}$, than the two tutor-random groups, Tut_{Prob} & Tut_{Step}. The final group sizes are as follows: $N = 92$ for $Stud_{Prob}$, $N = 93$ for $Stud_{Step}$, $N = 47$ for Tut_{Prob}, and $N = 47$ for Tut_{Step}.

Due to the holiday break, preparations for final exams, and length of the experiment, 212 students completed the experiment. 11 students were excluded from our subsequent analysis because they performed perfectly on the probability pretest. The remaining 201 students were distributed as follows: $N = 70$ for $Stud_{Prob}$; $N = 59$ for $Stud_{Step}$; $N = 38$ for Tut_{Prob}; $N = 34$ for Tut_{Step}. We performed a χ^2 test of the relationship between students' condition and their rate of completion and found no significant difference among the groups: $\chi^2(3) = 1.159, p = 0.763$.

Probability Tutor

Pyrenees is a web-based ITS for probability. It covers 10 major principles of probability, such as the Complement Theorem and Bayes' Rule. In prior studies Pyrenees was compared against Andes, another well-evaluated ITS (Vanlehn et al., 2005). Results showed that Pyrenees significantly outperformed Andes in both physics (VanLehn et al., 2004) and probability (M. Chi & VanLehn, 2007). This improvement was observed in part because Pyrenees teaches students domain-general problem-solving strategies, which draw students' attention to the conditions under which each domain principle is applicable. The differences were apparent on all types of test problems: simple/complex problems and isomorphic/non-isomorphic problems, and the effects were large, with Cohen's d=1.17 for overall post-test scores.

Figure 1 shows the interface of Pyrenees, which is divided into multiple windows. Through the dialogue window, Pyrenees provides messages to the students such as explaining a worked example step, or prompting them to complete the next step. Students can enter their inputs, such as writing an equation or selecting the answer to a multiple-choice question, through the response text box below. Any variables or equations that are defined through this process are displayed on left side of the screen for reference. Any time an answer is submitted, Pyrenees provides immediate feedback on whether or not it is correct.

In addition to providing immediate feedback, Pyrenees can also provide on-demand hints prompting the student with what they should do next. As with other systems, help in Pyrenees is provided via a sequence of increasingly specific hints. The last hint in the sequence, the bottom-out hint, tells the student exactly what to do. For the purposes of this study we incorporated four distinct pedagogical decision modes into Pyrenees to match the four conditions.

Procedure

In this experiment, students were required to complete 4 phases: 1) pre-training, 2) pre-test, 3) training on Pyrenees, and 4) post-test.

During the pre-training phase, all students studied the domain principles through a probability textbook. They read a general description of each principle, reviewed some examples of its application, and solved some single- and multiple-principle practice problems. After solving each problem, the student's answer was marked in green if it was correct and red if incorrect. They were also shown an expert solution at the

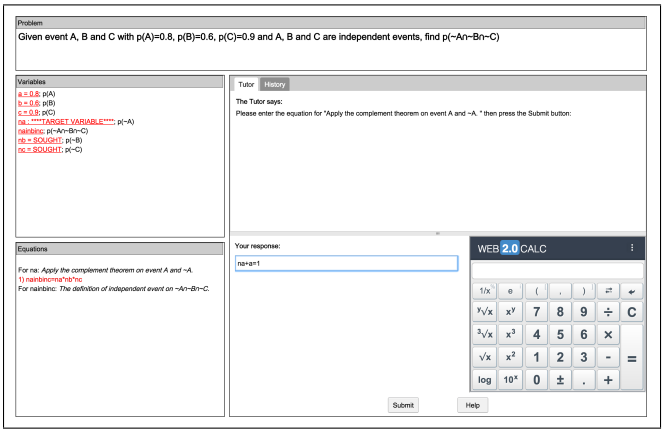


Figure 1: The interface of the Pyrenees probability tutor used in this study.

same time. If the students failed to solve a single-principle problem then they were asked to solve an isomorphic one; this process was repeated until they either failed three times or succeeded once. The students had only one chance to solve each multiple-principle problem and were not asked to solve an isomorphic problem if their answer was incorrect.

The students then took a pre-test which contained 10 problems. The textbook was not available. They were not given feedback on their answers, nor were they allowed to go back to earlier questions. This was also true of the post-test.

During phase 3, students in all four conditions received the same 12 problems in the same order on Pyrenees. Each primary domain principle was applied at least twice. The minimum number of steps needed to solve each training problem ranged from 20 to 50. The steps included variable definitions, principle applications and equation solving. The number of domain principles required to solve each problem ranged from 3 to 11. For the FWE problems, the $Stud_{Step}$ students were asked to make decision only on two types of steps: **principle selection** and **principle application**. To apply each principle, students need to first do principle selection: to choose the principle that they will use and then do principle application: to write the appropriate equation to apply it. We evaluated the students' decisions on both types of steps in our analysis below. The only procedural differences among the four conditions were the decision agency: Student vs. Tutor and the granularity of the decision: Problem vs. Step. Apart from this, the system was identical.

Finally, all of the students took a post-test with 16 problems. Ten of the problems were isomorphic to the pre-test problems given in phase 2. The remainder were non-isomorphic complicated multiple-principle problems.

Grading Criteria

The test problems required students to derive an answer by writing and solving one or more equations. We used three scoring rubrics: binary, partial credit, and one-point-per-principle. Under the binary rubric, a solution was worth 1 point if it was completely correct or 0 if not. Under the partial credit rubric, each problem score was defined by the proportion of correct principle applications evident in the solution. A student who correctly applied 4 of 5 possible principles would get a score of 0.8. The One-point-per-principle rubric in turn gave a point for each correct principle application. All of the tests were graded in a double-blind manner by a single experienced grader. The results presented below were based upon the partial-credit rubric but the same results hold for the other two. For comparison purposes, all test scores were normalized to the range of [0,1].

Results

A one-way ANOVA test on students' pre-test score show that there is no significant difference among the four groups. $F(3,197) = 1.969$, $p = 0.12$. The second column in Table 1 showed students' pretest scores and as we can see, the two Tutor decision groups, Tut_{Prob} and Tut_{Step}, had higher pretest scores than the two Student decision groups: $Stud_{Prob}$ and $Stud_{Step}$ but the difference is not significant. Next we will compare students' learning performance in the post-test and training time across the four conditions. We discuss each comparison in turn.

Learning Performance

A repeated measures analysis using test type (pre-test and isomorphic post-test) as factors and test score as the dependent measure showed a main effect for test type $F(3,197) = 163.160$, $p < 0.0001$. On the isomorphic questions, all four groups of students scored significantly higher on the post-test than on the pre-test, $F(1,69) = 68.04$, $p < 0.0001$ for $Stud_{Prob}$; $F(1,58) = 65.35$, $p < 0.0001$ for $Stud_{Step}$; $F(1,37) = 8.349$, $p = 0.004$ for Tut_{Prob}; and $F(1,33) = 32.04$, $p < 0.0001$ for Tut_{Step}. Therefore all four conditions made significant gains from pre- to post-test by training on Pyrenees. This suggests that the basic practice and problems, domain exposure, and interactivity of Pyrenees might help students to learn even when the problem- and step-level decisions are made randomly.

Table 1 shows a comparison of the pre-test, isomorphic post-test (10 isomorphic questions), and overall post-test scores among the four conditions, showing the mean (and SD) for each score. We calculated a two-way ANCOVA analysis on decision agency (Student vs. Tutor) $\times$ granularity (Problem vs. Step) using pretest scores as a covariate. We found a significant interaction effect on the isomorphic post-test scores: $F(1,196) = 5.664$, $p = 0.018$. However, there was no significant main effect on either decision agency or the granularity alone. Pairwise t-tests showed a significant difference between $Stud_{Prob}$ and Tut_{Prob} groups: $t(106) = 2.514$, $p = 0.013$, $d = 0.477$, that is, the $Stud_{Prob}$ scored significantly higher than the Tut_{Prob} on isomorphic post-test scores. Additionally, there is a trend that Tut_{Step} group out-performed Tut_{Prob} group: $t(70) = -1.853$, $p = 0.068$, $d = 0.444$. Therefore, this result showed that students were able to make effective problem-level decisions in that $Stud_{Prob}$ group learned significantly more than random decision Tut_{Prob} group but not step level decisions in that $Stud_{Step}$ is not significantly better than those trained with the random decisions Tut_{Step}.

Similarly, a two-way ANCOVA on the factors of granularity and decision using pretest scores as a covariate also showed significant interaction effect on the overall post-test score: $F(1,196) = 4.375$, $p = 0.038$. Again there was no significant main effect on either the granularity or decision

Table 1: Learning Performance

Cond	pre	Iso Post	Overall Post
$Stud_{Prob}(70)$	.684(.186)	.890(.119)	0.788(.137)
$Stud_{Step}(59)$	.671(.212)	.861(.129)	0.778(.152)
$Tut_{Prob}(38)$	.737(.189)	.818(.177)	0.726(.198)
$Tut_{Step}(34)$	.754(.167)	.882(.101)	0.811(.133)

Table 2: Time on task (in minutes)

Cond	# Stud	Total Time
$Stud_{Prob}$	70	130.39(28.26)
$Stud_{Step}$	59	148.54(42.31)
Tut_{Prob}	38	121.53(47.15)
Tut_{Step}	34	136.44(30.27)

agency alone. Post-hoc pairwise t-tests showed the Tut_{Step} group had significantly higher scores than the Tut_{Prob} group: $t(70) = -2.107$, $p = 0.039$, $d = 0.503$ and a trend that the $Stud_{Prob}$ group out-performed the Tut_{Prob} group: $t(106) = 1.933$, $p = 0.056$, $d = 0.368$. Therefore, it seems that tutor's step-level decision is more effective than tutor's problem-level decision. But no significant difference was found between the two student decision making groups.

To summarize, our results showed that: 1) the granularity can make a significant difference on student learning in that tutor's step-level decisions can be more effective than tutor's problem-level decisions; and 2) students can make better problem-level decision than random, but not better step level decisions. Therefore, one potential explanation for the lack of the difference between the two student decision groups is that: while the step level decisions can indeed be more effective than problem-level decisions, the students cannot make effective step level decisions to fully take advantage of the learning power that step level decisions can provide. Further research is needed to investigate this hypothesis.

Training Time

Table 2 shows the average amount of total training time (in minutes) students spent on Pyrenees for each condition. A two-way ANOVA analysis on granularity and decision agency revealed there is no significant interaction effect. However, there is a significant main effect of granularity: $F(1,197) = 10.283$, $p = 0.0015$ and a marginal main effect of decision agency: $F(1,197) = 3.609$, $p = 0.059$. Subsequent pairwise t-tests showed that the $Stud_{Step}$ condition spent significantly more time than the $Stud_{Prob}$ and Tut_{Prob} conditions: $t(127) = -2.902$, $p = 0.004$, $d = 0.504$ ($Stud_{Prob}$); $t(95) = -2.937$, $p = 0.004$, $d = 0.603$ (Tut_{Prob}).

Overall, we found that decision granularity can make a difference on the time on task: 1) students spent more time with step-level decisions than problem-level decisions in that the two step-level groups spent significantly more time than the two problem-level groups; 2) the two student decision groups seemingly spent more time on task than the two random tutor groups, but the difference was only marginally-significant.

Student Decisions

Our preliminary analysis on students' decision-making preference suggested that students are far more likely to choose problem solving than worked examples.

Problem Level Decisions: Table 3 shows the number of different types of problem level decisions made by the $Stud_{Prob}$ and the Tut_{Prob} groups. Columns 2 and 3 show the average number of worked examples and problem-solving problems that each condition experienced. We required each student to solve two problems in order to familiarize them with Pyrenees. Therefore each student made 10 problem-level decisions. For the $Stud_{Prob}$ group, the students chose less than two WEs on average; while the Tut_{Prob} group, the students received an almost equal number of WEs and PSs (5.45 vs. 4.55) since the tutor makes random decisions. That is, the $Stud_{Prob}$ group only received 15.8% of worked examples; while the Tut_{Prob} group received 54.5% worked examples . This difference was statistically significant: $t(106) = -13.203$, $p < 0.0001$, $d = 2.614$

Table 3: Number of problem-level decisions

Cond	WE	PS	Total
$Stud_{Prob}$	1.58(1.40)	8.44(1.40)	10
Tut_{Prob}	5.45(1.57)	4.55(1.57)	10

Table 4: Number of step-level decisions

ST	Cond	WE	PS	Total
Principle Selection	$Stud_{Step}$	9(10)	58(11)	67
	Tut_{Step}	34(4)	33(4)	67
Principle Application	$Stud_{Step}$	11(11)	56(11)	67
	Tut_{Step}	34(5)	33(5)	67

Step Level Decisions: Table 4 shows the number of different types of step level decisions made by the $Stud_{Step}$ and Tut_{Step} groups on the principle selection and principle application steps. For both groups the number of WE and PS decisions sum to 67. The $Stud_{Step}$ group selected an average of only 9 WE steps and decided to do PS on the remaining 58 steps; while in the Tut_{Step} group, since the tutor makes random decisions, the students received an almost equal number of WE and PS steps (34 vs. 33). That is, the $Stud_{Step}$ group received 14.81% WE steps while the Tut_{Step} group received 50.92% WE steps. This difference was also statistically-significant: $t(91) = 13.67$, $p < 0.0001$. We found the similar patterns on the principle application steps as well.

To summarize, the two tutor decision groups received about equal number of WE vs. PS at either problem or step levels while the two student decision groups, either problem level or step level, are significantly more likely to decide to do Problem Solving than Worked Examples.

Discussion

In this study, we investigated the impact of students' pedagogical decision-making on learning. We focused on the decisions whether to give students a WE or to engage them in PS at two levels of granularity: problem versus step. We were able to strictly control the domain content and thus to isolate the impact of *pedagogy* from *content*. And we compared the

students' pedagogical decision making performance to a random baseline with a goal of factoring out the impact of hand-coded strategies on student learning.

Interestingly, our results showed that students can make effective problem-level decisions that enabled them to significantly outperform students with random decisions. However, the students were no better than a random tutor when making step-level decisions. When comparing across the four conditions we found that the Tutor random step-level decision group outperformed the Tutor random problem-level decision group ($Tut_{Step} > Tut_{Prob}$) but no significant difference was found between the two student decision groups.

Our results suggests that different granularity of pedagogical policies can significantly impact students' performance in that the step-level decisions can potentially be more beneficial than the step level ones; however, the students are more capable of making effective problem-level pedagogical decisions than making step-level ones. This may be due to the fact that students may lack the necessary metacognitive skills to make such fine-grain decisions or because they get overwhelmed by the number of decisions to make.

Surprisingly, students selected more problem solving than worked example on both problem and step levels. The feeling of engagement may partly explain their decisions. Prior work has shown that students are more likely to be engaged in the learning process when they experience a sense of control over it(Harackiewicz et al., 1987). Therefore, the students might decide to do problem solving simply because they feel more involved in problem solving than in worked example. However, much further research is needed to fully understand why.

Currently we are applying Reinforcement Learning (RL) to induce effective pedagogical policies directly from our data. We will investigate whether the RL-induced policies can be more effective than student decision making at both levels of granularity. We will also investigate whether it is possible to combine RL-induced policies with student decision making and thus give students both beneficial guidance and an all-important sense of agency.

Acknowledgements
This research was supported by the NSF Grant #1432156: "Educational Data Mining for Individualized Instruction in STEM Learning Environments".

References

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In *Intelligent tutoring systems* (pp. 292–303).

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*, 4-16.

Chi, M., & VanLehn, K. (2007). The impact of explicit strategy instruction on problem-solving behaviors across intelligent tutoring systems. In *Proceedings of the 29th annual conference of the cognitive science society, nashville, tennessee* (pp. 167–172).

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363-387.

Harackiewicz, J. M., Sansone, C., Blair, L. W., Epstein, J. A., & Manderlink, G. (1987). Attributional processes in behavior change and maintenance: smoking cessation and continued abstinence. *Journal of Consulting and Clinical Psychology*, *55*(3), 372.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30-43.

McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In *Artificial intelligence in education* (pp. 222–229).

Najar, A. S., Mitrovic, A., & McLaren, B. M. (2014). Adaptive support versus alternating worked examples and tutored problems: Which leads to better learning? In *User modeling, adaptation, and personalization* (pp. 171–182).

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education*, *70*(4), 293–315.

Roll, I., Baker, R. S. d., Aleven, V., & Koedinger, K. R. (2014). On the benefits of seeking (and avoiding) help in online problem-solving environments. *Journal of the Learning Sciences*, *23*(4), 537–560.

Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89.

Vanlehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227–265.

VanLehn, K., Bhembe, D., Chi, M., Lynch, C., Schulze, K., Shelby, R., ... Wintersgill, M. (2004). Implicit versus explicit learning of strategies in a non-procedural cognitive skill. In *Intelligent tutoring systems* (pp. 521–530).

Vanlehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... Wintersgill, M. (2005). The andes physics tutoring system: Lessons learned. *JAIED*, *15*(3), 147–204.

Zhou, G., Price, T. W., Lynch, C., Barnes, T., & Chi, M. (2015). The impact of granularity on worked examples and problem solving. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2817–2822).

Can Musical Engagement Alleviate Age-Related Decline in Inhibitory Control?

Ruben Vromans (r.d.vromans@tilburguniversity.edu)
Tilburg center for Cognition and Communication, Tilburg University
5000 LE Tilburg, The Netherlands

Marie Postma-Nilsenová (Marie.Postma@tilburguniversity.edu)
Tilburg center for Cognition and Communication, Tilburg University
5000 LE Tilburg, The Netherlands

Abstract

The purpose of our study was to determine whether active musical engagement alleviates decline in inhibitory control due to cognitive aging. Given that musical training in young adults has been shown to improve attentional performance, we can expect this benefit to persist for older adults as well. With the help of the stop-signal procedure, we measured response inhibition of young and older adults who provided a self-reported assessment of their musical engagement, using the recently validated Goldsmiths Musical Sophistication Index. The Gold-MSI addresses a variety of musical activities and thus offers a more comprehensive measure than ability to play a musical instrument used in the past. Results of the experiment showed that older participants had longer stop-signal reaction times, independently of their musical training and engagement, but musical training and ensemble practice were negatively related to the proportion of missed responses suggesting a weak effect of certain types of musical activities on inhibitory control.

Keywords: inhibitory control; musical sophistication; attention; cognitive aging; stop-signal task.

Introduction

Inhibiting an action that has just been initiated, like crossing the street when the light suddenly turned red, requires high levels of cognitive control (Verbruggen & Logan, 2008a). Cognitive control refers to "purpose control mechanisms that modulate the operation of various cognitive subprocesses and thereby regulate the dynamics of human cognition" (Miyake et al., 2000:50); it is used as a cover term for high-level executive functions that coordinate the cognitive ability of selective attention and a range of other complex cognitive processes (e.g., Diamond, 2013; Miyake et al., 2000; Moreno & Bidelman, 2014). Metaphorically speaking, cognitive processes in the human mind are regulated by executive function in a way similar to how the conductor of a symphony orchestra coordinates the performance of individual instrumentalists (Brown, 2006).

Response inhibition is considered to be a key component of higher order executive functions (Miyake et al., 2000). This type of inhibition is often defined as the ability to stop actions or responses that are no longer required or appropriate, in order to support flexible and goal-directed behavior in dynamic environments (Verbruggen & Logan, 2009a). It is the ability of having control over one's attention, thoughts, emotions and behavior (Diamond, 2013). According to a recent study by Moreno & Farzan (2015), inhibitory control is also an important mediator in transferring skills between cognitive functions. Without response inhibition, human behavior would be dominated by impulsive reactions, habits and distracting factors in the environment.

Response inhibition also plays a central role in theorizing about human cognitive aging (e.g., Bedard et al., 2002; Hasher & Zacks, 1988; Williams et al., 1999). It has often been argued that, similar to other cognitive functions (i.e., working memory, auditory perception, or motor control), performance on response inhibition declines as we get older and considerable resources are being spent on training and improving it. Repeated practice throughout life appears to be key (e.g., Ericsson, Nandagopal, & Roring, 2009), as suggested by the *mental-exercise hypothesis* (Salthouse, 2006). However, existing programs mostly yield mixed results, particularly with respect to far-transfer to untrained tasks (Barnett & Ceci, 2002; Owen et al., 2010).

Contrary to the limited impact of specialized cognitive training tasks and (video-)games, musical training appears to result in robust domain-independent transfer effects (Schellenberg, 2004). Playing a musical instrument not only requires complex motor movements, but also demands high levels of executive control including selective attention, switching, inhibition, updating and monitoring (for a review, see Moreno & Bidelman, 2014). In fact, musicians' cognitive advantages have been found in a variety of domains including mathematics (Vaughn, 2000), verbal and non-verbal skills (Forgeard et al., 2008), and working memory in both the auditory (Chan, Ho, & Cheung, 1998; Pallesen et al., 2010) and visual domain (George & Coch, 2011; Weiss et al., 2014).

Existing studies focus on the comparison of musically trained and untrained individuals, thus disregarding the potentially advantageous effects of other types of musical engagement, such as DJ-ing, music engineering and even frequent passive exposure to music, particularly if accompanied by emotional response. In addition, the research of Corrigall, Schellenberg and Misura (2013) showed that the length of musical training during childhood and adolescence is strongly related to variables such as openness-to-experience, IQ and education of parents, suggesting that simple comparative studies of the relation between musical training and cognitive abilities in a single age group are likely to be dealing with several confounding socio-demographic and personality variables. Finally, from

a cognitive psychological point of view, it is important to gain a better understanding of the mechanisms that underlie the relationship between musical performance and different levels of inhibitory control. There is a wealth of experimental evidence suggesting that musical experience enhances inhibitory control at the level of interference control (e.g., selective attention) in both the auditory (e.g., Bialystok & DePape, 2009; Parbery-Clark, Skoe, & Kraus, 2009; Strait & Kraus, 2011) and visual domain (Rodrigues, Loureiro, & Caramelli, 2013). However, little research has focused on the effect of musical experience on inhibitory control at the level of response inhibition (e.g., Moreno et al., 2014).

Present Study

The present study addresses the following research question: 'Can musical engagement alleviate the negative effect of aging on inhibitory control?' Based on earlier research regarding the developmental link between aging and inhibitory control, we expect that young adults would be faster in stopping prepotent responses than older adults. Second, in accordance with research on musical training and enhanced inhibitory control performance, we expect that musical engagement would improve inhibitory control in all ages. In order to test these hypotheses experimentally, we made use of the stop-signal procedure (e.g., Logan, Cowan, Davis, 1984; Verbruggen & Logan, 2008a, 2009a). In this task, the primary performance variable is the amount of time that is needed to withhold your response when a stop-signal is given. This variable is also known as the stop-signal reaction time (SSRT) and provides the latency of the stop process.

Method

Participants

In total, 120 participants, consisting of two music groups (musicians and non-musicians) and two age groups (young and older adults) voluntarily took part in the study. All participants were native speakers of Dutch. Fifty-eight of the participants were young adults (28 female; M_{age} = 22.9 (*SD* = 3.1), range: 18-29 years), and sixty-two were older adults (29 female, M_{age} = 65.5 (*SD* = 5.5), range: 57-83 years). In order to ensure that skilled musicians were included in the sample, as well as to address the potential effect of ensemble musical practice, about a half of the participants (*N*=62), both young and older, was recruited from three Dutch symphony orchestras (the Philips Symphony Orchestra of Eindhoven, the Wildacker Orchestra of Goirle and the South Netherlands Philharmonic), and from the conservatory of Tilburg.

Instrumentation and Material

Goldsmiths Musical Sophistication Index. In line with recent research on musical abilities (e.g., Carey et al., 2015), we used the Goldsmiths Musical Sophistication Index (Müllensiefen, Gingras, Stewart, & Musil, 2014) to assess participants' musical experience. The Gold-MSI is a comprehensive instrument that was developed to account for individual differences in active musical engagement beyond the traditional focus on the ability to play an instrument. It is based on the assumption that other forms of engagement such as DJ-ing, music production and music engineering can have an effect on the cognitive system that is comparable to the impact of instrumental practice. The Gold-MSI consists of five sub-scales: *active engagement* (9 items addressing resources spent on music, e.g., 'I am intrigued by musical styles I am not familiar with and want to find out more'), *perceptual abilities* (9 items on accuracy of musical listening skills, e.g. 'When I sing, I have no idea whether I am in tune or not'), *musical training* (7 items regarding one's life history of formal musical training, e.g., 'I would not consider myself a musician'), *singing abilities* (7 items on singing performance, e.g., 'I am able to hit the right notes when I sing along with a recording'), and *emotions* (6 items on the ability to discuss emotional musical expressions, e.g. 'I am able to talk about the emotions that a piece of music evokes for me'). Next to the five scales, a *general* measure of musical *sophistication* is derived from a subset of items associated with the five subscales. Table 1 provides an overview of the dimensions measured by the Gold-MSI including Cronbach's alpha.

Table 1: Overview of the Gold-MSI subscales

Subscale	Cronbach's α	Nr. of Items	Score Range
Active Engagement	.838	9	9-63
Perceptual Abilities	.869	9	9-63
Musical Training	.950	7	7-49
Singing Abilities	.879	7	7-49
Emotion	.750	6	6-36
General Sophistication	.951	18	18-126

Stop-Signal Task. The inhibition task was based on the stop-signal paradigm developed by Logan et al. (1984), which is considered to be a sophisticated measure of response inhibition (Diamond, 2013; Verbruggen & Logan, 2008a). In the STOP-IT task (Verbruggen, Logan, & Stevens, 2008), the presentation of the stop-signal is adaptive in that it is presented with a variable delay. Participants are assigned a primary shape judgment task, discriminating between a square and a circle. A stop-signal (a 750 Hz tone, 75 ms long) is occasionally presented to inhibit the judgment reaction (on 25% of the trials). If inhibition is successful, the stop-signal delay (the interval between the presentation of the stimulus and the stop-signal, SSD) will increase in intervals of 50 ms from the default delay of 250 ms. In case of unsuccessful inhibition, SSD will decrease with the same interval length (the lowest SSD being 50 ms).

The task consisted of two phases: a practice block of 32 trials, designed to accustom the participants to the task, and an experimental phase of three blocks of 64 trials, each followed by performance feedback presented for 10 seconds. Subsequently, two variables were used as performance measures on the task: the stop-signal reaction

time (SSRT) and the proportion of missed responses on no-signal trials (MISS). SSRT is considered to serve as a reliable indication of overall performance, while MISS provides a useful indication as to which response strategy had been used during the stop-signal task (Verbruggen et al., 2008; Verbruggen & Logan, 2009b).

The stimuli for the STOP-IT task were displayed on two Dell Laptops running on Windows 7 Professional with a 15.6" display, or on an ASUS N56VZ-RH71 running on Windows 8 with a 15.6" display. For all three laptops, the response keys "Z" and "/" were covered with a sticker displaying a square and a circle, respectively. The auditory stop-signal was presented via a Sennheiser PC 320 headset.

Results

In order to obtain a comprehensive view of the relation between musical engagement and inhibitory performance, we made use of all the five sub-scales obtained with the Gold-MSI (*active engagement, perceptual abilities, musical training, singing abilities, and emotions)*. In an additional analysis, we also tested separately the so-called *general sophistication* (calculated with a subset of the items included in the sub-scales). Finally, we tested the relation with participants' experience with ensemble musical performance encoded as a binary variable. Table 2 presents a correlation matrix for the outcome variables of the stop-signal task (SSRT and MISS), the five Gold-MSI sub-scales and the *general sophistication* score of the Gold-MSI, as well as participants' age. As can be seen from the table, age related positively with SSRT and MISS. Furthermore, a negative correlation was found between the proportion of missed responses on no-signal trials (MISS) and musical training. Finally, all Gold-MSI measures were strongly related to each other.

Table 2: Correlations between the sub-scales of the Gold-MSI, outcome variables of the stop-signal task and age.

Variable	1	2	3	4	5	6	7	8	9
1. Age	-								
2. SSRT	.24*	-							
3. MISS	.32*	.06	-						
4. AE	-.15	.06	-.11	-					
5. PA	-.06	-.01	-.11	.74*	-				
6. MT	-.05	-.01	-.20*	.69*	.74*	-			
7. SA	.01	.08	-.01	.66*	.77*	.79*	-		
8. EMO	-.03	.04	-.05	.69*	.75*	.59*	.62*	-	
9. GS	-.02	-.09	-.13	.82*	.86*	.93*	.91*	.72*	-

Note. SSRT = stop-signal reaction time; MISS = missed responses on no-signal trials; AE = active engagement; PA = perception abilities; MT = musical training; SA = singing abilities; EMO = emotions; GS = general sophistication; *$p < .01$.

Musical Ensemble Practice. Since experience in performing with a musical ensemble could lead to a significant improvement in inhibitory control, we first examined the effect of the binary variable musical ensemble practice in combination with a participant's age as possible predictors of SSRT and MISS. A regression analysis revealed a significant effect of age for both stop-signal performance outcomes, with an additional effect of musical ensemble practice on the proportion of missed responses on no-signal trials, see Table 3.

Table 3: Summary of regression analyses for Age and Musical Ensemble Practice (MEP) as predictors of stop-signal reaction time (SSRT) and the proportion of missed responses on no-signal trials (MISS).

	SSRT			MISS		
Variable	*B*	*SE B*	*t*	*B*	*SE B*	*t*
Age	0.844	0.313	2.698**	0.129	0.035	3.644**
MEP	3.539	13.625	0.260	-3.341	1.536	-2.174*
R^2		0.06			0.13	
F		3.67*			9.1**	

Note. MEP = Musical Ensemble Practice; SSRT = stop-signal reaction time; MISS = missed responses on no-signal trials; *$p < .05$, **$p < .01$.

As can be seen from the box plots in Figure 1, there was no effect of musical ensemble experience on the stop-signal reaction time. However, the proportion of missed responses was effected by participants' musical ensemble experience; non-musicians with no experience had a significant higher percentages than orchestra musicians.

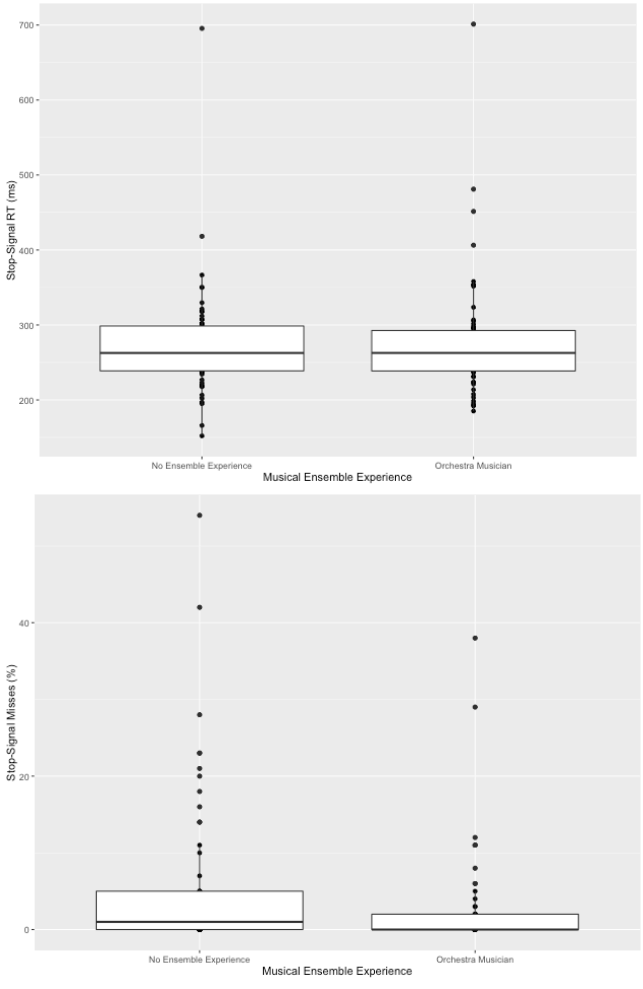


Figure 1: Box plots illustrating the stop-signal reaction times (top) and proportion of missed responses (bottom) for both participants with and without musical ensemble experience.

Five Gold-MSI Sub-scales. To capture the relation between inhibitory control, age and other forms of musical experience, we subsequently performed a regression analysis with the five Gold-MSI Sub-scales (active engagement, perceptual abilities, musical training, singing abilities, emotions) and age as predictors of SSRT and MISS. This regression showed a significant effect of age for both stop-signal performance outcomes, with an additional effect of the scale 'musical training' on the proportion of missed trials (see Table 4). The same effects were found in additional two regression analyses in which four SSRT-outliers (with SSRT around 700ms) were discarded (results not reported due to space limit).

Table 4: Summary of regression analyses for Age and the five Gold-MSI Subscales as predictors of stop-signal reaction time (SSRT) and the proportion of missed responses on no-signal trials (MISS).

		SSRT			MISS	
Variable	*B*	*SE B*	*t*	*B*	*SE B*	*t*
Age	0.862	0.317	2.717**	0.124	0.036	3.431**
AE	1.171	0.994	1.178	0.059	0.113	0.516
PA	-2.157	1.307	-1.650	-0.087	0.149	-0.582
MT	-0.860	0.817	-1.052	-0.231	0.093	-2.475*
SA	1.689	1.208	1.398	0.204	0.138	1.478
EMO	1.774	1.667	1.064	0.093	0.190	0.490
R^2		0.11			0.16	
F		2.24*			3.53**	

Note. AE = active engagement; PA = perception abilities; MT = musical training; SA = singing abilities; EMO = emotions; SSRT = stop-signal reaction time; MISS = missed responses on no-signal trials; $^{*}p < .05$, $^{**}p < .01$.

General Musical Sophistication. Finally, to assess the relation between inhibitory control, age, and general musical sophistication we carried out a regression analysis with general musical sophistication and age as predictors of SSRT and MISS. This regression analysis only revealed a significant effect of age for both stop-signal performance outcomes, as listed in Table 5.

Table 5: Summary of regression analyses for Age and General Sophistication (GS) as predictors of stop-signal reaction time (SSRT) and the proportion of missed responses on no-signal trials (MISS).

		SSRT			MISS	
Variable	*B*	*SE B*	*t*	*B*	*SE B*	*t*
Age	0.847	0.313	2.709**	0.128	0.036	3.588**
GS	0.120	0.251	0.479	-0.038	0.029	-1.337
R^2		0.06			0.11	
F		3.75*			7.47**	

Note. GS = general sophistication; SSRT = stop-signal reaction time; MISS = missed responses on no-signal trials; $^{*}p < .05$, $^{**}p < .01$.

Figure 2 illustrates the association between the six sub-scales of the Gold-MSI and the stop-signal reaction times.

Furthermore, the relation between the six sub-scales of the Gold-MSI and the proportion of missed responses on no-signal trials is shown in Figure 3.

Conclusion

Taken together, these results of the stop-signal task suggest that there is an association between inhibitory control and aging; as we grow older, our ability to stop responses that have already been initiated decreases, regardless of musical engagement and ensemble experience. However, musical training and ensemble practice were negatively related with the proportion of missed responses. Therefore, on the question whether musical engagement alleviates age-related decline in inhibitory control, we suggest a weak effect of certain types of musical activities on inhibitory control.

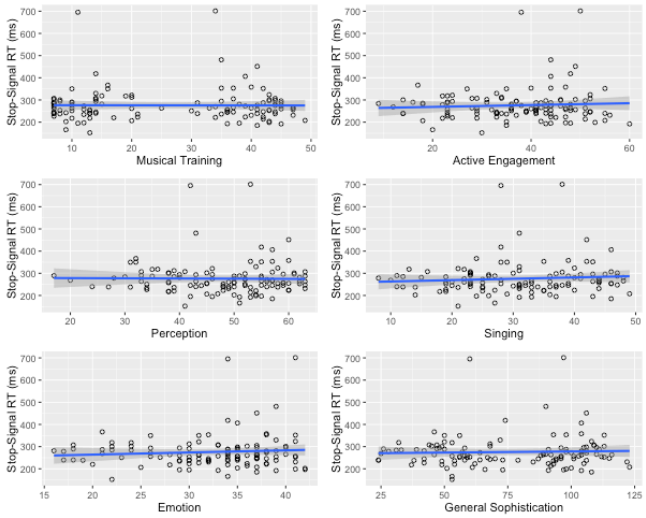


Figure 2: Plots of Gold-MSI scale in relation to stop-signal reaction times.

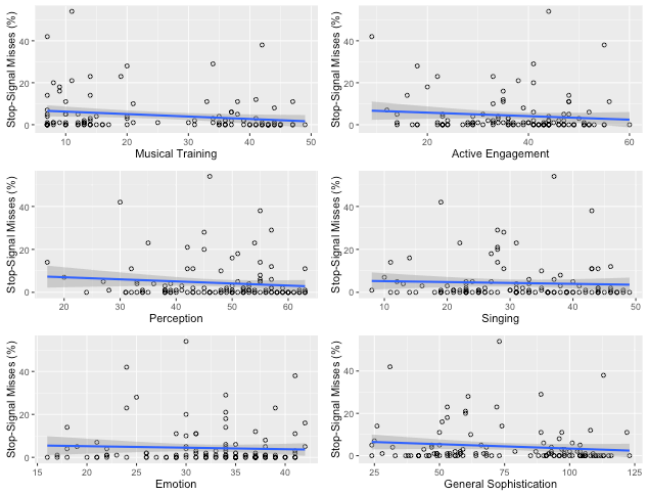


Figure 3: Plots of Gold-MSI scale in relation to stop-signal misses.

Discussion

The purpose of the current study was to determine whether different forms of musical engagement decrease age-related

decline in inhibitory control. Musical engagement was assessed with the Goldsmiths Musical Sophistication Index (Gold-MSI, Müllensiefen et al., 2014). In addition, we tested the effect of musical ensemble practice on the behavioral aspect of inhibitory control, that is, response inhibition, using the stop-signal procedure (Logan et al., 1984; Verbruggen & Logan, 2008a).

The SSRT, which is the primary performance variable and indicates the speed of process of inhibitory control (Verbruggen et al., 2008), increased with age. This is in line with previous findings (Bedard et al., 2002; Kramer et al., 1994), though see Williams et al. (1999) for an opposing view. A possible explanation is that during stop-signal trials, participants are confronted with constant cognitive demand on task goals (i.e., stopping and going), and on set shifting modalities (i.e., auditory and visual stimuli). Therefore, the observed age-related losses in inhibitory control might be caused by increased cognitive demand (Bedard et al., 2002; Verbruggen & Logan, 2009b for similar reasoning).

Next, we found that various kinds of musical engagement did not improve inhibitory control performance, contrary to the *mental exercise hypothesis* (Salthouse, 2006), which assumes that continued practice helps slow down age-related losses in cognition. Possibly, the ability to inhibit a planned action is not only one of the earliest emerging executive functions but is also maintained the longest (Barkley, 1997), given the importance of inhibitory control for survival (e.g., Bedard et al., 2002).

The results obtained in this study are inconsistent with research of Moreno and colleagues (2014). Arguably, this might also be due to the use of different paradigms (i.e., the Go/No-Go paradigm and the stop-signal paradigm). Both paradigms test different kinds of response inhibition and require different demands of cognitive control (Verbruggen & Logan, 2008b).

Finally, we noted that the proportion of missed trials (MISS) increased with age; however, older adults with usical ensemble experience and/or musical training had fewer missed responses than their peers. These outcomes suggest that older adults use a different response strategy during the stop-signal task that is aimed at waiting for a stop-signal to occur. By doing so, they reduce the risk of pressing on stop-signal trials. This observed waiting strategy could be attributed to the proactive-adjustment hypothesis (Verbruggen & Logan, 2009b), which means that during the stop-signal task, participants proactively change their response threshold when they expect a stop-signal to occur on the next few trials. Older adults with musical training apply the waiting strategy and are better in timing their response. Perhaps musicians benefit from their extensive rhythmic experiences and enhanced timing skills, and auditory-motor expectancies (Brown, Zatorre, & Penhune, 2015), which may help them anticipate thoughts or actions for unexpected events related to music, such as the stop-signal in the present study (Furuya & Soechting, 2010).

A key contribution of the present study was that we treated musical expertise as a multifaceted concept. Classically, the distinction between musicians and non-musicians was primarily based on participant's musical training background. However, the degree of musical engagement in our study was established by the recently validated Gold-MSI (Müllensiefen et al., 2014), which provided a more comprehensive measure than a simple focus on the ability to play a musical instrument. By gathering data from young and older adults, we ensured that the association between personality and early musical training played a limited role as confounding variable (Corrigall et al., 2013).

Notwithstanding, some important issues with regard to musical expertise should be mentioned. Future studies should examine the exact contribution of ensemble practice compared to solo practice. In ensemble settings, musicians are required to listen carefully to different aspects of sounds produced by other musicians (e.g., intonation, timing, and dynamics), as well as to observe the visual input provided by the conductor in order to time their reaction. This, howeverm is not need in private musical training (Habibi et al., 2014). Also, different instruments might impact inhibitory control in various ways. Past research that compared two specific instrumental groups (violinists versus pianists) (Carey et al., 2015) found differences in terms of auditory skills (e.g., violinists showed a greater ability to detect small changes in pitch than both pianists and non-musicians), suggesting possible effects on general cognitive skills as well.

To conclude, the present findings strengthen the idea that response inhibition decreases during normal cognitive aging. Although we found no evidence that both musical training and general musical engagement alleviate the negative effect of aging on inhibitory control, other life-long activities, such as occupational complexity, can be explored in the future. However, musical training and ensemble practice relates negatively to the proportion of missed responses suggesting a weak effect of certain types of musical activities on inhibitory control. Arguably, response inhibition is an example of a complex executive function and more research should be undertaken in the future, in order to enhance our understanding of its components and how they can be trained effectively.

References

Barkley, R. A. (1997). Behavioral inhibitory control, sustained attention, and executive functions. *Psych Bulletin, 121*, 65-94.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?. *Psych bulletin*, *128*, 612.

Bedard, A. C., Nichols, S., Barbosa, J. A., Schachar, R., Logan, G. D., & Tannock, R. (2002). The development of selective inhibitory control across the life span. *Developmental Neuropsychology, 21*, 93-111.

Bialystok, E., & DePape, A. M. (2009). Musical expertise, bilingualism, and executive functioning. *JEP: Human Perception and Performance, 35*, 565-574.

Brown, T. E. (2006). Executive functions and attention deficit hyperactivity disorder. *International Journal of Disability, Development and Education*, *53*, 35-46.

Brown, R. M., Zatorre, R. J., & Penhune, V. B. (2015). Expert music performance: cognitive, neural, and developmental bases. *Progr Brain Research*, *217*, 57-86.

Carey, D., Rosen, D., Krishnan, S., Pearce, M. T., Stepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effect of musical expertise on perception and cognition. *Cognition, 137*, 81-105.

Chan, A. S., Ho, Y. C., & Cheung, M. C. (1998). Music training improves verbal memory. *Nature, 396*, 128.

Corrigall, K. A., Schellenberg, E. G., & Misura, N. M. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, *4*, 10-3389.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135-68.

Ericsson, K. A., Nandagopal, K., & Roring, R. W. (2009). Toward a science of exceptional achievement: attaining superior performance through deliberate practice. *Annals of the New York Academy of Sciences, 1172*, 199-217.

Forgeard, M., Winner, E., Norton, A., & Schlaug, G. (2008). Practicing a musical instrument in childhood is associated with enhanced verbal ability and nonverbal reasoning. *Plos one*, *3*, e3566.

Furuya, S., & Soechting, J. F. (2010). Role of auditory feedback in the control of successive keystrokes during piano playing. *Exp Brain Res*, *204*, 223-37.

George, E. M., & Coch, D. (2011). Music training and working memory. *Neuropsych, 49*, 1083-94.

Habibi, A., Ilari, B., Crimi, K., Metke, M., Kaplan, J. T., Joshi, A. A., ... & Damasio, H. (2014). An equal start: absence of group differences in cognitive, social, and neural measures prior to music or sports training in children. *Frontiers in Human Neuroscience, 8*, 690.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 22, pp. 193-225). CA: Academic Press.

Kramer, A. F., Humphrey, D. G., Larish, J. F., Logan, G. D., & Strayer, D. L. (1994). Aging and inhibition: Beyond a unitary view of inhibitory processing in attention. *Psychology and Aging, 9*, 491-512.

Logan, G. D., Cowan, W. B., & Davis, K. A. (1984). On the ability to inhibit responses in simple choice reaction time tasks: A model and a method. *JEP:HPP, 10*, 276-291.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49-100.

Moreno, S., & Bidelman, G. M. (2014). Examining neural plasticity and cognitive benefit trough the unique lens of musical training. *Hearing Research, 308*, 84-97.

Moreno, S., & Farzan, F. (2015). Music training and inhibitory control: a multidimensional model. *Annals of the New York Academy of Sciences, 1337*, 147-152.

Moreno, S., Wodniecka, Z., Tays, W., Alain, C., & Bialystok, E. (2014). Inhibitory control in bilinguals and musicians. *PloS one*, *9*, e94169.

Moritz, C., Yampolsky, S., Papadelis, G., Thomson, J., & Wolf, M. (2013). Links between early rhythm skills, musical training, and phonological awareness. *Reading and Writing*, *26*, 739-769.

Müllensiefen, D., Gingras, B., Stewart, L., & Musil, J. (2014). *The Goldsmiths musical sophistication index (Gold-MSI): Technical report and documentation v1.0.* London: Goldsmiths, University of London.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*, 775-778.

Pallesen, K. J., Brattico, E., Bailey, C. J., Korvenoja, A., Koivisto, J., Gjedde, A., & Carlson, S. (2010). Cognitive control in auditory working memory is enhanced in musicians. *PloS one*, *5*, e11120.

Parbery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical experience limits the degradative effects of background noise on the neural processing of sounds. *J Neuro, 29*, 14100-7.

Rodrigues, A. C., Loureiro, M. A., & Caramelli, P. (2013). Long-term musical training may improve different forms of visual attention ability. *Brain & Cognition, 82*, 229-35.

Salthouse, T. A. (2006). Mental exercise and mental aging evaluating the validity of the "use it or lose it" hypothesis. *Perspectives on Psych Science, 1*, 68-87.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, *15*, 511-514.

Strait, D. L., & Kraus, N. (2011). Can you hear me now? *Frontiers in Psychology, 2*, 113.

Vaughn, K. (2000). Music and mathematics: Modest support for the oft-claimed relationship. *Journal of Aesthetic Education, 34*, 149-166.

Verbruggen, F., & Logan, G. D. (2008a). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences, 12*, 418-424.

Verbruggen, F., & Logan, G. D. (2008b). Automatic and controlled response inhibition. *JEP: Gen, 137*, 649-72.

Verbruggen, F., & Logan, G. D. (2009a). Models of response inhibition in the stop-signal and stop-change paradigm. *Neurosci & Biobehav Reviews, 33*, 647-661.

Verbruggen, F., & Logan, G. D. (2009b). Proactive adjustments of response strategies in the stop-signal paradigm. *JEP: HPP, 35*, 835-854.

Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). STOP-IT: Windows executable software for the stop-signal paradigm. *Behav Res Meth, 40*, 479-483.

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of inhibitory control across the life span. *Dev Psych, 35*, 205-213.

Weiss, A. H., Biron, T., Lieder, I., Granot, R. Y., & Ahissar, M. (2014). Spatial vision is superior in musicians when memory plays a role. *Journal of Vision*, *14*, 18.

A 6-month longitudinal study on numerical estimation in preschoolers

Pierina Cheung
mcheung@wesleyan.edu
Department of Psychology
Wesleyan University

Emily Slusser
emily.slusser@sjsu.edu
Child & Adolescent Development
San Jose State University

Anna Shusterman
ashusterman@wesleyan.edu
Department of Psychology
Wesleyan University

Abstract

The current study investigated the development of numerical estimation in 3- to 5-year-old children sampled monthly for six months. At each session, children completed a task that assesses verbal number knowledge (Give-N task) and a numerical estimation task that assesses approximate number knowledge (Fast Cards). Results showed that children who acquired the cardinal principle (CP) during the course of the study showed marked improvement on the estimation task. Following CP acquisition, estimation became more accurate overall but also fluctuated widely. We discuss the implications of our findings for number word learning, particularly the mapping between verbal number and the approximate number system (ANS).

Keywords: numerical estimation; approximate number; subset-knowers; cardinal principle knowers; longitudinal

Introduction

How do children acquire number word meanings? On one view, the approximate number system (ANS) – a non-verbal representation of number – provides meanings for number words. That is, the initial meanings of number words *are* ANS representations (e.g., Dehaene, 2011; Gallistel & Gelman, 2000). The current paper investigates the mapping between ANS and verbal number in a longitudinal study, by tracking the development of number word learning and the ANS-to-word mapping in a group of preschoolers.

Starting at around the age of 2, children begin to recite the count list, but they lack meanings for number words. Slowly and gradually, over the next 1 ½ to 2 years, children learn the meanings of the first few number words in order. Using the Give-N task, Wynn (1990, 1992) showed that children first acquire the meaning of 'one' – these children give one object when asked for 'one' (1-knower). Approximately 6 months later, children become 2-knowers and have exact meanings for 'one' and 'two' but not higher numbers. Later, they acquire the meaning of 'three' (3-knower) and 'four' (4-knower). Collectively, they are called 'subset-knowers' because they have acquired a subset of number word meanings. Shortly thereafter, children understand that counting can be used to generate any set size within their count list. At this stage, children are called cardinal principle knowers (CP-knowers) because they understand that the last word of a counted set denotes the cardinality of the set and have exact number word meanings for all numbers in their count list (e.g., Gelman & Gallistel, 1978; Le Corre, Van de Walle, Brannon, & Carey, 2006).

The question of how children acquire number word meanings has sparked interest on the interface between non-verbal and verbal representations of number. In particular, some have highlighted the importance of parallel individuation – a non-verbal system for representing up to 3 or 4 individual objects in parallel – as the primary source of meanings for small number words (i.e., 'one' through 'four'; Carey, 2009; Le Corre & Carey, 2007). Nevertheless, parallel individuation has a set size limit and cannot represent exact number concepts for numbers beyond four.

Unlike parallel individuation, the ANS – a system that represents number as continuous magnitudes – has no set size limit (Dehaene, 2011; Gallistel & Gelman, 2000). The ANS follows Weber's law and exhibits scalar variability (i.e., the variability of estimates increases linearly with the mean of estimates, see Feigenson, Dehaene, & Spelke, 2004, for review). The ANS is a dedicated system for representing number: it supports numerical computations such as comparison, addition, subtraction, and multiplication (e.g., Barth, La Mont, Lipton, Dehaene, Kanwisher, & Spelke, 2006; McCrink & Wynn, 2004).

A growing body of research has examined the development of approximate number knowledge and verbal number knowledge and found conflicting results with regard to when children map number words onto representations in the ANS. In a widely cited study, Le Corre and Carey (2007) tested children's number word knowledge using Give-N and developed a task – Fast Cards – to assess the mapping of number words on ANS representations. In Fast Cards, an array of dots is presented quickly (~1s) and children are asked to estimate the number of dots without counting. Le Corre and Carey (2007) analyzed the slope of average estimates produced as a function of set size. The main interest was children's estimates beyond the capacity limit of parallel individuation – sets larger than 4. They found that many subset-knowers and some CP-knowers produced slopes that were not significantly different from zero in the large number range (> 4), suggesting that these children provided similar estimates when shown sets of 6, 8, and 10. In other words, they did not provide larger estimates for larger set sizes. Further, Le Corre and Carey (2007) found that only some CP-knowers were able to provide an estimate that approximately matched target sets; these children showed positive slopes that were significantly different from zero. Based on these findings, Le Corre and Carey (2007) hypothesized further development after the CP induction, and they set a criterion for classifying children who have mapped number words onto nonverbal numerical magnitudes as having slopes greater than or equal to 0.3. They called these children "CP-mappers". On the other hand, children were identified as "CP non-mappers" if their slopes were less than 0.3. CP non-

mappers were on average 6 months younger than CP-mappers. From this cross-sectional data, they inferred that children first become CP non-mappers and eventually become CP mappers. Moreover, given that CP-knowers only mapped large sets onto the ANS *after* acquiring the cardinal principle, they argued that the ANS cannot be foundational to the acquisition of early number word meanings.

Others have, however, reported different results. Using a modified Give-N task, Wagner and Johnson (2011) found that children between the ages of 3 and 5 often *do* produce larger sets for larger requests. For example, they found that although subset-knowers gave an incorrect number when asked to give 'eight fish,' they were nevertheless more likely to generate a larger set than when asked to give 'four fish'. They took this to suggest that even subset-knowers had approximate number word knowledge, and thus, the ANS plays a crucial role in children's acquisition of number words, even before children acquire the cardinal principle. However, they did not assess 'knower-level' in the same way as other researchers in the literature, making it difficult to directly compare their results to Le Corre and Carey (2007).

More recent studies have offered some insights into these discrepant findings. First, Gunderson and colleagues (2015) showed that the range of set sizes included in the analysis for both Fast Cards and Give-N could affect the results. Specifically, the slopes of subset-knowers were found to be positive and different from zero only when the analysis included numbers that were immediately beyond the range of a child's knower-level (N+1 for an N-knower); for most children except for 4-knowers, this analysis included numbers smaller than or equal to 4. On the contrary, they did not find positive slopes when the analysis included numbers *beyond* the small number range (> 4; e.g., 'five', 'six'). These findings suggest that subset-knowers understand that numbers between N + 1 and 4 are smaller than numbers larger than 4. Given that children's representation of small numbers can be supported by parallel individuation, slopes that include numerosities in the small number range cannot provide definitive evidence for the role of ANS in number word acquisition.

Odic and colleagues (2015) offer a different view, arguing that the mapping between verbal number and the ANS is not bidirectional. In Fast Cards, children are shown an array of dots and asked to generate a numeral, and thus the mapping is from the ANS to number word (ANS-to-Word mapping); in tasks where children are asked to generate sets of objects upon a verbal request, such as Give-N, the mapping is from number word to ANS (Word-to-ANS mapping). They found that Word-to-ANS mapping develops prior to ANS-to-Word mapping, and that subset-knowers show only Word-to-ANS mappings. Given their results, it remains a possibility that ANS could play some role in children's acquisition of exact number concepts.

Taking a different approach, a recent longitudinal study tested the relation between the precision of ANS representation (Panamath task) and children's verbal number knowledge (Give-N). Shusterman and colleagues (2016) found that improvements in ANS acuity were related to the acquisition of the cardinal principle, but not to other stages in number word acquisition. They also found that the largest changes in ANS acuity were observed *at* or *after* children became CP-knowers, but not before, suggesting that while the ANS could be related to number word learning, especially cardinality, it is unlikely to be the driving force. Additionally, while the longitudinal data confirmed the knower-levels trajectory proposed by Wynn (1992), the data did not support a transition from CP non-mapper to CP mapper as proposed by Le Corre and Carey (2007).

Studies on the mapping between ANS and number words thus far suggest that the ANS and number word acquisition may be related (Odic et al., 2015; Shusterman et al., 2016; Wagner & Johnson, 2011, but see Le Corre & Carey, 2007; Gunderson et al., 2015), but none have clarified how *changes* surrounding children's number word acquisition are related to mapping to the ANS. To address this, we analyzed the numerical estimation data from the Fast Cards task in a 6-month longitudinal study, with two goals: First, we aimed to investigate children's mapping between the ANS and number words, and specifically, the relation between changes in the ANS-to-Word mapping and changes in children's number word knowledge. If mappings to ANS representations do in fact support the transition from subset to CP knower, then improvement in the quality of ANS-to-Word mappings should be evident prior to this transition. The second goal was to understand the dynamics of CP knowers' mapping between number words and ANS representations. If CP-knowers are indeed a heterogeneous group with respect to the quality of ANS-to-Word mappings (i.e., non-mappers and mappers), then individual children should exhibit some stability in their performance and clear improvement over time. Further, if ANS-to-Word mappings develop most substantially *after* the CP transition, as proposed by Le Corre and Carey (2007), such development should be apparent in this longitudinal data set.

Longitudinal Study

Methods

Participants A group of 33 children (15 male) between the ages of 36 and 55 months (M = 46.6 months) participated. Children were recruited in Central Connecticut and spoke English as their primary language. They were tested once a month for a total of six months. All children completed at least four sessions, with most participating in all six sessions.

Procedure Participants were tested individually at local preschools. Written consent was obtained from parents. In each testing session, children completed an ANS acuity task (Panamath; Halberda, Mazzocco & Feigenson, 2008), elicited counting task, Give-N (Wynn, 1990), and the numerical estimation task (Fast Cards, adapted from Le Corre

& Carey, 2007), always in that order.[1]

Elicited Counting. Children were asked to count a set of 10 objects.

Give N. Children were shown a set of 20 fish and a large bowl. The experimenter started by asking the child to put one fish in the bowl ("Can you make one fish go swimming?") If the child gave 1 fish, the experimenter asked for N + 1 fish ("Can you make two fish go swimming?") After the child placed some fish in the bowl, the experimenter asked to confirm, "Is that N?" If the child responded "no", they were given the opportunity to fix it. If the child succeeded, the experimenter moved on to the next number, and if the child failed, the experimenter asked for N – 1. This process continued until the child could give 8 fish correctly. A child was classified as an N-knower if s/he was correct 2 out of 3 times when N was asked, and failed 2 out of 3 times on requests of N + 1. Children who had only acquired a subset of the numeral meanings were collectively termed "subset-knowers" (1-knowers, 2-knowers, 3-knowers, 4-knowers). Children who could give all sets correctly up to 8 were classified as cardinal principle knowers (CP-knowers).

Fast Cards. At the beginning of the task, children were told that they would see some objects on the screen. They were encouraged to guess how many objects there were, as quickly as possible, without counting. The study began with a training session in which a set of objects, ranging from 1 to 15 objects, was presented in numerical order.

There were four blocks in the Fast Cards task. In each block, there were seven trials. Sets of 1, 2, 3, 4, 6, 10, 14 objects were used. Because it is questionable whether children could discriminate sets of 8 and 10 items, set sizes used in the original Fast Cards task, we changed the stimuli to clearly discriminable sets of 6, 10, and 14. In each trial, identical objects of the same kind were presented. Each trial lasted for 1s. Numbers were randomized within each block. On two of the blocks, total surface area co-varied with numerosity, and on the other two, total surface area was held constant.

Results

Elicited Counting. All children could count up to 10 with no more than one error.

Give-N [2] At the beginning of the study, we found 14 subset-knowers (3 1-knowers, 3 2-knowers, 4 3-knowers, 4 4-knowers), and 19 CP-knowers. At the end of the study, 8 of the subset-knowers became CP-knowers, 6 children remained subset-knowers, and 19 children remained CP-knowers throughout the study.

Fast Cards In the following analysis, we focused on three main questions. First, we aimed to document developmental changes in children's estimation ability. Second, we investigated whether the ANS is related to the cardinal principle acquisition. Third, we sought to replicate previous findings that there are two groups of CP-knowers – mappers and non-mappers. To address these questions, we analyzed the development of children's estimation ability by calculating average slopes and error rates.

Relationship between changes in children's estimation ability and verbal number knowledge To examine whether the mapping between ANS and number words differed depending on children's knower-level status, we categorized children into three knower-level groups based on developmental changes observed on the Give-N task, and included this as a variable in our analysis. The "subset-only group" included children who were subset-knowers throughout the study; the "subset-to-CP group" included children who acquired the cardinal principle during the 6-month period, and "CP-only group" included children who were CP-knowers throughout the study.

Linear slopes. To examine whether children have mapped large number words onto non-verbal number representations, we assessed the slope of children's estimates as a function of target set size. Linear slopes were computed separately in the small (1-4) and large number range (6, 10, 14). Given that our main question concerns the role of the ANS, we focused our analysis on the large number range because sets less than 4 can also be represented by parallel individuation.[3] If ANS representations support cardinality development, or vice versa, we should observe significant improvement in ANS-to-Word mappings in children who acquired the CP during the course of the study (subset-to-CP).

To test this, we asked whether children's slopes differed as a function of their Knower-Level Group (i.e., subset-only, subset-to-CP, and CP-only). We constructed mixed effects model, with random intercepts for subjects, and Age (in months), Session and Knower-Level Group as fixed effects. We found no effect of Age, $X(1) = .039$, $p = .84$, or Session, $X(5) = 4.61$, $p = .46$, suggesting that children's ANS-to-Word mapping did not significantly improve over time during the 6-month period. As predicted, we found a significant main effect of Knower-Level Group, $X(2) = 9.80$ $p = .0074$, and a significant interaction between Knower-Level Group and Session, $X(10) = 18.88$, $p = .042$.

The significant interaction indicated that the developmental trajectory in estimation differs depending on children's knower-level status. We conducted analyses separately on each of the three knower-level groups. Results showed that there was no effect of Session for the subset-only group, $X(5) = 1.45$, $p = .92$, and CP-only group, $X(5) = 2.87$, $p = .72$, but there was a significant effect of Session for subset-to-CP knowers, $X(5) = 20.25$, $p < .001$ (see Figure 1). As indicated in Figure 1, the subset-to-CP group

[1] The ANS acuity task was administered as part of a larger project and the data were reported elsewhere (Shusterman, et al., 2016).

[2] Longitudinal analyses of the Give-N task are reported in Shusterman, et al. (2016).

[3] Averaging across all sessions, children from all knower-level groups showed significantly positive slopes in the 1-4 range, ps < .008 (mean slopes: subset-only = 1.18, subset-to-CP = 1.10, CP-only = 1.07), indicating that children were able to estimate small sets and that they were engaged in the task.

demonstrated the largest improvement in slopes in the large number range.

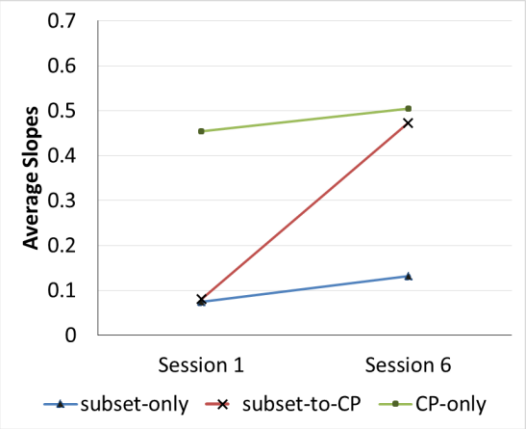


Figure 1: Average slopes from Session 1 and Session 6 in the large number range (6, 10, 14).

We further explored the changes in slopes in the large number range for the subset-to-CP group by comparing their average slopes across sessions *prior to* and *after* the CP acquisition. Consistent with the group-level analysis, we found significant improvement in slopes, $t(13.72) = -2.31$, $p = .037$. Children's slopes were close to 0 before they acquired the cardinal principle ($M_{slope} = .079$) and improved after they became a CP-knower ($M_{slope} = .33$).

While it is possible that this improvement in slopes was an effect of repeated testing, note that there was *no* effect of Session among the subset-only and CP-only groups. If practice effects are what drive children's improvements on Fast Cards, then we should observe similar levels of improvement in the subset-only and CP-only group.

Error rates. Similar analyses were conducted using error rates as a measure of accuracy. Error rates were computed as the difference between children's response and the target value. For example, when shown a set of 14 objects, the error rate was -4 if the response was 10, or +1 if the response was 15. A value of 0 indicates accurate mapping. Negative error rates indicate that children were underestimating and positive error rates indicate that they were overestimating.

Mean error rates were computed for each session. Using the same linear mixed effects model as the previous analysis on slopes, we found a main effect of Session, $X(5) = 27.22$, $p < .001$, and an effect of Knower-level Group, $X(2) = 9.54$, $p = .0085$, but there was no interaction between Knower-level Group and Session, $X(10) = 12.62$, $p = .25$. There was also no effect of Age, $X(1) = .019$, $p = .89$. Figure 2 displays average error rates for each session. Our results also showed that, consistent with previous studies, children tend to underestimate in Fast Cards (Odic et al., 2015).

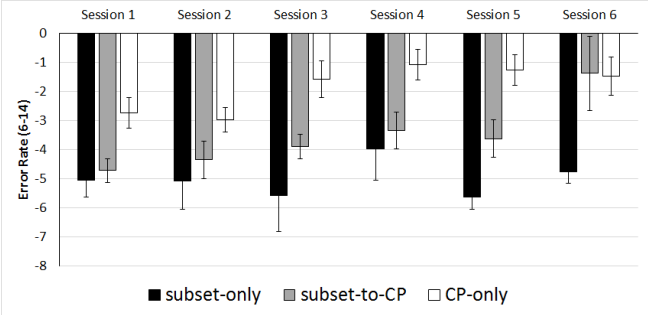


Figure 2: Average error rates across sessions in the large number range. Error bars represent SEM.

We compared overall error rates across the three groups. CP-only group was significantly more accurate than the subset-to-CP group, $t(14.69) = 2.35$, $p = .033$, and the subset-only group, $t(13.18) = 5.28$, $p < .001$. The subset-to-CP group had significantly lower error rates ($M_{error} = -3.49$) than those who remained subset-knowers throughout the study ($M_{error} = -5.17$; $t(12) = 2.26$, $p = .043$), suggesting that the ANS-to-Word mapping between subset-to-CP group and subset-only group is qualitatively different.

Results from the slope and error rate analysis both suggested that children's performance on Fast Cards improved over the course of a 6-month period. Importantly, subset-knowers who acquired the cardinal principle during the study improved to a greater extent than those who continued to be subset-knowers at the end of the study. These findings point to a previously unexplored relationship between cardinal principle acquisition and ANS-to-Word mappings—namely, that these mappings develop neither before nor after, but primarily in concert with, the CP transition. Although these results cannot definitively address the causal relation between ANS mappings and the CP acquisition, they suggest that acquiring cardinality allows children to fine-tune ANS mappings to generate more precise ANS mappings (we will return to this point in the Discussion). In our final analysis, we investigated the developmental trajectory of children's mapping ability by examining the stability of their mapper vs. non-mapper status.

Mappers vs. Non-Mappers Previous research identified separate groups of CP-knowers who had mapped large number words onto ANS as 'mappers' and those who had not as 'non-mappers' (Le Corre & Carey, 2007). The criterion for establishing mappers and non-mappers is often set at slopes ≥ 0.3, with some reporting that a criterion between 0.3 and 0.5 reveals similar results (e.g., Odic et al., 2015). To investigate whether there are two groups of CP-knowers in our sample, we conducted a Shapiro-Wilk test, which showed that CP-knowers' slopes violated the normality assumption in the first two sessions of the study (Session 1: $W = .88$, $p = .029$; Session 2: $W = .87$, $p = .001$), but the distribution was more evenly distributed from Session 3 through Session 6 ($Ws > .95$, $ps > .14$). As shown in Figure 3a, children's slopes

in Session 1 had two peaks, one at around 0 and another at 0.5, but towards the end of the study in Session 6, children had a mean slope of .43 (see Figure 3b).

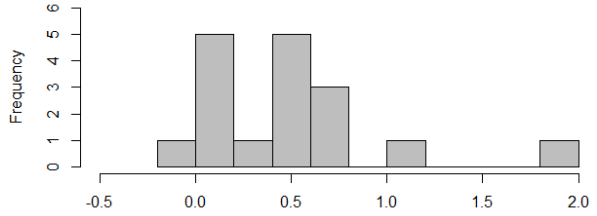


Figure 3a: Histogram of slopes in the large number range for CP-knowers at Session 1

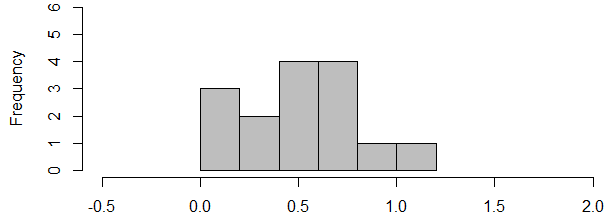


Figure 3b: Histogram of slopes in the large number range for CP-knowers at Session 6

Next, we adopted the same criterion as previous studies and identified children as 'mappers' if their slopes were ≥ 0.3 (Le Corre & Carey, 2007; Gunderson et al., 2015; Davidson et al, 2012; Odic et al., 2015). In a cross-sectional study, Le Corre and Carey (2007) found a 6-month difference in age between CP-mappers and CP-non-mappers, and Gunderson and colleagues (2015) found a 3-month difference in age. To investigate the developmental trajectory of mappers and non-mappers, we identified CP-knowers as mappers if their slopes were ≥ 0.3 at Session 1, and non-mappers if their slopes were less than 0.3. We asked (1) how likely CP-non-mappers were to become mappers at the end of the study, and (2) whether CP-mappers continued to be mappers throughout the course of the study.

We first asked how likely children who were non-mappers at Session 1 were to become mappers by Session 6. Among the 7 non-mappers, only 1 child became a mapper, 4 remained non-mappers, and 2 children did not complete Session 6. While this finding may suggest that most CP non-mappers remained non-mappers throughout the six-month period, an inspection of the individual development of slopes shows that children's slopes fluctuated over time (Figure 4). We found that the 7 non-mappers remained non-mappers for the remaining five sessions approximately 60% of the time (3 out of 5 sessions). In contrast, children who were mappers at Session 1 remained mappers 80% of the time (approximately 4 out of 5 sessions).

*Scalar variability***.** If numerals are mapped onto the ANS representations, children's estimates in Fast Cards should exhibit scalar variability (e.g., Cordes et al., 2001). That is, the mean estimates and variability should increase linearly as the set size increases. This is typically calculated as the ratio of the standard deviation over the mean estimates (i.e., the coefficient of variation, COV). We computed COVs for each target set size in each session. We hypothesized that if CP mappers used mappings to the ANS in Fast Cards, their COVs for each of the large sets should remain similar to each other.

A linear mixed effects model with Set Size (6, 10, 14), Session (1-6), children's mapper status in Session 1, and Age (in months) revealed that there was a significant effect of Set Size, $X(1) = 9.24$, $p = .002$. Pairwise comparisons revealed that COV-14 (M=.21, SE = .071) was significantly different from COV-6 (M = .32, SE = .071), p = .001, and COV-10 (M=.33, SE = .071), p = .058, but COV-6 and COV-10 did not differ from each other, p = .70. No other effects were found. Our results on COVs for 6 and 10 replicate previous studies on ANS-to-Word mapping, which did not include sets larger than 10 (Le Corre & Carey, 2007; Odic et al., 2015). Although our prediction that COVs would be constant was not validated, since the COVs for 14 were different than those for 6 and 10, we speculate that a potential explanation is that not all children could stably count to 14. Further work is needed to understand how COVs reflect children's estimation performance. Nevertheless, these results support the claim that children's estimates for sets smaller than 10 were based on mappings to ANS representations.

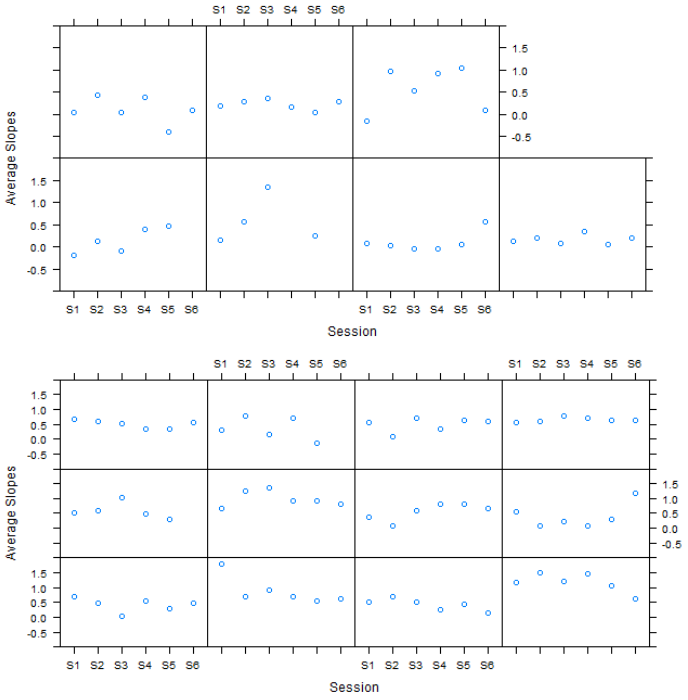


Figure 4: Scatterplots of mean slopes for each session for children who started as a) CP non-mappers (top) and b) CP mappers (bottom).

Discussion

The current study revealed two main findings. First, we found that subset-knowers who acquired the cardinal principle (subset-to-CP) showed the greatest improvement in response slopes in an estimation task – they became better at generating larger estimates for larger sets. Results with error rates also showed that these children had smaller error rates

than children who remained subset-knowers throughout the study (subset-only). These results are consistent with previous studies that suggest that the ANS may be related to the acquisition of the cardinal principle (e.g., Wagner & Johnson, 2011; Odic et al., 2015; Shusterman et al., 2016).

Our second main finding is that CP knowers' estimation slopes fluctuate over time. While it is often assumed that CP non-mappers would eventually become a mapper, our results showed that this is not necessarily the case. Children's status as a mapper or a non-mapper varied from session to session, suggesting that researchers should be cautious about using 0.3 as a 'standard' cut-off point for determining mapper status. We do not suggest that the mapper and non-mapper distinction is a false dichotomy. Indeed, similar to previous studies (Le Corre & Carey, 2007; Odic et al., 2015), we also found a non-normal distribution of CP-knowers' slopes at the beginning of the study, suggesting that CP-knowers are a heterogeneous group with regard to their ANS-to-Word mapping; furthermore, children who started as mappers were relatively more likely to retain that status across sessions. Rather, our results highlight the fragile nature of children's estimation in Fast Cards, and suggest that the period following the CP transition is characterized by large fluctuation in estimation quality, rather than consistently poor estimation.

The current paper provides an important piece of data to the debate on whether the ANS drives cardinal principle acquisition. Using a longitudinal design, we found that ANS-to-Word mapping significantly improved right at the moment when children became CP-knowers, but not prior to the CP acquisition. A previous analysis from the same longitudinal study showed that the largest improvement in Weber's fraction is observed *at* or *after* children acquire the cardinal principle (Shusterman, et al., 2016). Given that the ANS-to-Word mappings fluctuate to a much greater extent than previously assumed, especially around the subset-CP transition, it seems unlikely that the ANS underlies number word acquisition. Cumulatively, the evidence suggests that while the ANS and verbal number is closely related, acquiring the CP helps children fine-tune the ANS-to-Word mappings.

If the ANS does not drive the CP acquisition, what is the precise relationship between the non-verbal and verbal number acquisition? We speculate that in acquiring the notion of cardinality, children begin to understand the exact relation between number words and sets of objects – i.e., a set has a cardinality N if and only if you follow a stable count list and are assigning each object with a number word. Through counting an exact number of objects and noticing the difference between different cardinalities, children may thus begin to more accurately estimate a set of dots in Fast Cards.

A question that this current work raises is why the slopes of children's estimates fluctuate for such a long period of time. We speculate that factors such as children's attention and processing ability may affect their verbal estimation performance. Another possibility is the length and stability of children's count list. Future studies should examine the potential of these additional factors to predict verbal estimation.

Acknowledgments

We thank all participating families and preschools. We also thank Angela Lo, Alexander Hoyle, Anthony Gutierrez, Jessica Taggart, Sarah Edelman, and Barry Finder for their help with data collection and analysis. This work was supported by an NSF CAREER grant (DRL-0845966) awarded to A.S.

References

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E.S. (2006). Non-symbolic arithmetic in adults and young children. *Cognition, 98*, 199-222.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal and nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*(4), 698-707.

Dehaene, S. (2011). *The Number Sense: How the mind creates mathematics*. New York: Oxford University Press.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object files: Evidence from infants' manual search. *Developmental Science, 6*(5), 568-584.

Feigenson, L., Dehaene, S. & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307-314.

Gelman, R. & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Oxford: Harvard University Press.

Gallistel, C.R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. Trends in Cognitive Sciences, 4(2), 59-65.

Gunderson, E.A., Spaepen, E., & Levine, S.C. (2015). Approximate number word knowledge before the cardinal principle. *Journal of Experimental Child Psychology, 130*, 35-55.

Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differences in nonverbal number acuity predict maths achievement. *Nature*, 455, 665-668.

Le Corre, M. & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*, 395- 438.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, 15, 776-781.

Odic, D., Le Corre, M., & Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition, 138*, 102-121.

Shusterman, A., Slusser, E., Odic, D., & Halberda, J. (2016). Acquisition of the cardinal principle coincides with improvement in approximate number system acuity in preschoolers. PLoS ONE, 11(4), 1-22.

Wagner, J.B., & Johnson, S.C. (2011). An association between understanding cardinality and analog magnitude representations in preschoolers. *Cognition, 119*, 10-22.

Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology, 24*(2), 220-251.

Xu, F., & Spelke, E.S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, B1-B11.

The distorting effect of deciding to stop sampling

Anna Coenen and Todd M. Gureckis
Department of Psychology, NYU, 6 Washington Place, New York, NY 10003 USA
{anna.coenen, todd.gureckis}@nyu.edu

Abstract

People usually collect information to serve specific goals and often end up with samples that are unrepresentative of the underlying population. This can introduce biases on later judgments that generalize from these samples. Here we show that goals influence not only *what* information we collect, but also when we decide to terminate search. Using an optimal stopping analysis, we demonstrate that even when learners have no control over the *content* of a sample (i.e., natural sampling), the simple decision of when to stop sampling can yield sample distributions that are non-representative and could potentially bias future decision making. We test the prediction of these theoretical analyses with two behavioral experiments.

Keywords: information search; stopping rules; sampling; decision-making

Introduction

Information search typically improves judgment and decision making by reducing uncertainty about the the world. Imagine trying to research what qualities make a good business manager prior to making a hiring decision. To achieve this goal, it might make sense to look at the characteristics of CEOs of successful companies and find out what they have in common. In this case, your research seems to add information, improving your ability to decide. However, no matter how much data you collect, it would be a fallacy (and one that people often commit) to make the inverse inference that managers with those characteristics will necessarily lead companies to success (Denrell, 2003). In this case, the way in which information is collected can strongly bias judgment and decision making.

This scenario points to a common problem we encounter when sampling the world. In most cases, the information we obtain is not random, but rather tailored to our specific goals. Biases can arise when the information we collect with one goal or objective in mind is used to make later decisions pertaining to different goals. To give another example, imagine that we had sampled more broadly CEOs from both successful and unsuccessful companies to help answer the question what makes companies successful. Even though that broader sample is now more suitable to answer that question, it would still be unhelpful for answering other questions, like whether college dropouts do better than graduates in their subsequent careers. In this paper, we examine the relationship between search goals, information sampling decisions, and prediction.

Selective versus natural sampling

Examples of biased sampling such as those reviewed above typically involve *selective sampling*, when learners have control over *what* they sample (e.g., only successful managers). In addition to the example of conditional reasoning, selective sampling biases have also been shown when people have to trade-off reward and information (e.g. Denrell & March, 2001; Rich & Gureckis, 2014) or when feedback is asymmetric across different choice options (Le Mens & Denrell, 2011).

Unlike selective sampling, *natural sampling* refers to the process of drawing samples directly from the generating distribution without conditioning queries on any particular aspects of the sample. For example, a natural sample of businesses could be obtained by randomly selecting a number of companies that were founded in a specific year, irrespective of their subsequent success. As noted above, natural sampling is often considered a remedy for biases introduced by selective sampling because it enables learners to "conserve the properties of the universe" (Fiedler, 2008), such that the distribution of the sample will mirror the distribution of the source in an unbiased fashion.

However, even during natural sampling learners can often make the decision when to *terminate search* which may in turn be influenced by their goals (see e.g., Juni et al., in press; Lee & Paradowski, 2007; Vul et al., 2014). Here, we will argue that optimal goal-induced stopping decisions can lead to potentially non-representative samples even when they are generated by natural sampling. In particular, using an optimal stopping analysis we show that goals can have a powerful impact not only on the size of samples that learners collect, but also on the content and distributional characteristics of those samples. We then show how such samples may not reflect the statistical properties of the original distributions and how they could later on produce biased decisions in a naive learner. We finally examine these model predictions with two behavioral experiments.

An optimal stopping analysis for natural sampling of binary outcomes

To demonstrate the impact of stopping on sample composition, consider a simple information search task in which learners repeatedly observe binary outcomes from a distribution of interest. For example, learners could be picking either good or bad apples in order to learn something about the quality of the tree. The tree will have some probability of yielding good and bad apples, but learners have to rely on a finite sample of apples to estimate this probability.

Such a binary task can be modeled as a Bernoulli process (a coin flip essentially) with two possible outcomes (*h*eads or *t*ails) with the outcome probability θ (probability of heads). Assume that the learner incurs a small cost c for every draw of the distribution (every coin flip). Let's also assume that the learner will subsequently have to answer one of the following two questions (sampling goals).

1. *Binary*: Find out if θ is greater than 0.5 ("Is the coin biased towards heads or tails?").

2. *Estimation*: Find out the value of θ ("What is the bias of the coin?").

Each goal is associated with a different reward that is a function of the true value of θ and the participant's estimate. The learner's task is to decide when to stop sampling and provide their estimate given this cost function, the current sample (heads and tails), and the sampling cost. Assuming that there exists a maximum number of samples that learners are allowed to draw this can be framed as an finite-horizon *optimal stopping problem*. At every possible state (defined by the size of the current sample, n, and number of heads in that sample, h_n), an optimal decision maker should compare the expected value of stopping and of continuing and choose whichever is higher. Thus the expected value of a state, given a horizon of a maximum of T flips is

$$V_n^{(T)}(n,h_n) = max\{V_{stop}, V_{cont}\} \quad (1)$$

where

$$V_{stop} = u_{stop}(n,h_n) - nc \quad (2)$$

and

$$V_{cont} = E[V_{n+1}^{(T)}(n+1,h_{n+1})] \quad (3)$$

and where $u_{stop}(.)$ is the expected utility of the post-sampling task, which depends on a learner's sampling goal (see below). The value of continuing to obtain another sample is a learner's expectation over possible future states given their current knowledge. The probabilities in the expectation are based on a beta distribution parameterized by the outcomes observed so far. The value of the final state (when $n = T$) is just the expected value of stopping, which, along with the Markov property, means that this problem is solvable by backwards induction (Ferguson, 2012).

To compute the utility of stopping under the binary goal, note that the probability of a coin being biased towards heads is $P(\theta > .5) = 1 - I_{0.5}(\alpha,\beta)$, where $I_x(\alpha,\beta)$ is the cumulative distribution function of the beta distribution with parameters $\alpha = h+1$ and $\beta = (n-h)+1$. Assuming that the learner chooses heads when $P(\theta > .5) > 0.5$ and tails otherwise, their expected utility for stopping is

$$u_{stop}^{bin} = max\{1 - I_{0.5}(\alpha,\beta), I_{0.5}(\alpha,\beta)\}\, r \quad (4)$$

where r is the reward for making a correct choice.

The expected utility from stopping under the estimation goal requires specifying a cost function over the distance between the participant's estimate and θ. For simplicity's sake, assume that the learner's answer counts as correct whenever their response lies within a .2 interval surrounding the true value, and incorrect otherwise. The expected utility from stopping under this goal is

$$u_{stop}^{est} = \max_{\theta}\left\{\int_0^1 Beta(x;\alpha,\beta) w_{0.2}(\theta - x)dx\right\} r \quad (5)$$

where $w_{0.2}$ is a boxcar function with a .2 wide interval. By convolving it with the posterior over θ it can be used to find the interval with the largest posterior density.

Predictions

To predict behavior using this model, we first need to choose values for sampling cost, c, potential reward of the secondary task, r, and the maximum number of trials (the length of the horizon), T. To generate more realistic predictions, we additionally assume some stochasticity in people's choice behavior by using a probabilistic choice rule instead of the $max()$ operator in Equation (1). For example, using a softmax choice rule yields the following probability of stopping at each state, $P_{stop} = \frac{exp(V_{stop}/\tau)}{exp((V_{stop}+V_{cont})/\tau)}$. It requires specifying an additional temperature parameter, τ which governs the degree of probabilistic responding (when $\tau = 0$ it chooses the maximum value, as $\tau \to \infty$ it chooses randomly). In this paper we will use the following parameter settings to derive model predictions: $T = 24, c = \$0.03, r = \$2, \tau = .02$. These values were chosen for illustrative purposes and many qualitative results hold across a broader range of values (as long as the trade-off between c and r leads to stopping after a number of samples but before the horizon is reached). The model was used to derive predictions for the expected sample size, stopping probabilities, and the composition of samples after stopping.

Stopping Figure 1A shows the model's predictions for the learner's decisions to stop (white) or continue (green) given the current state (outcomes observed so far). Under the binary goal, learners' should continue sampling if the current sample is balanced (similar proportion of heads and tails) and be more likely to stop when a sample is more extreme. In contrast, the estimation goal leads to a greater probability to continue sampling in a much broader range of states. This is unsurprising because estimation requires a representative sample and overall more data than binary choice. A perhaps more surprising and subtle prediction is that the probability is not uniform for each sample size, but shows a slight pattern of earlier stopping for both very extreme and very balanced samples. Both patterns are caused by a learner's expected success at the estimation task. When samples are extreme, expected accuracy is high and learners can stop earlier with a reasonable chance at success. When samples are very mixed (close to 0.5 average), a high chance of good performance would require too many costly samples, at which point stopping early and making a best guess might have higher expected value. Basically mixed samples tend to take too many costs samples to resolve accurately.

From the stopping matrices in Figure 1A one can now derive expected probabilities for stopping points, that is learners' expected final state after stopping, for specific values of θ. Figure 1B shows the likelihood of different stopping points assuming $\theta = 0.5$. As expected, binary learners are predicted to end up in early extreme states (all heads or all tails) or later mixed states, whereas estimation learners show a wider dis-

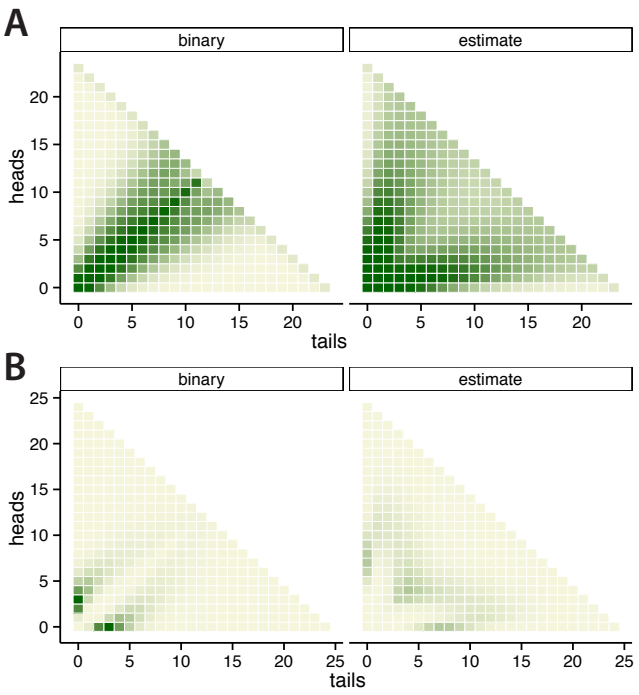


Figure 1: Model Predictions. A: Probability of continuing (green) or stopping (white) given observed data. B: Probability of having stopped in any given state. Modeled using uniform θ.

tribution of stopping states.

Sample Size Figure 2A shows the expected distribution of number of total samples taken, assuming learners are sampling from a process with underlying $\theta = 0.5$. Estimation learners are expected to collect more samples than binary learners, who face an easier secondary task and therefore terminate earlier on average.

Sample Composition The fact that stopping rules affect expected *size* of samples is not surprising given the different loss functions of the two goals. A more intriguing question is whether sampling goals also preserve the *properties of the distribution* that generates each sample. To investigate this, we will consider the proportion of heads and tails within each sample, after repeatedly sampling and stopping under the two goals. We expect that binary sampling would on average lead to more unequal samples (all heads, or all tails) since its stopping rule terminates early when outcomes are extreme (after two or three heads one can be pretty confident that heads is more common, for example) and continues when early outcomes are mixed (after one tails and two heads a decider might want to flip the coin at least one more time to be sure). On the other hand the estimation condition predicts a wider array of stopping points because the goal is to get an accurate picture of the average outcome probability.

Figures 2B and C confirm this effect of stopping rule on sample composition. It shows the expected frequency distribution of sample averages under each sampling goal, when the true $\theta = 0.5$(A) or $\theta = 0.7$(B). In the estimation condition, sample means fall around the expected value of the distribution, whereas the binary condition leads to many samples with only one type of outcome (all heads or all tails). Note that these more extreme samples tend to be small (that's when learners terminate early), but they still mostly contain more than one observation, which can be seen from Figure 1A. Thus, the difference in sample mean distributions between the two goals is not just due to the fact that the binary condition leads to samples of size 1 which necessarily have an average of 0 or 1.

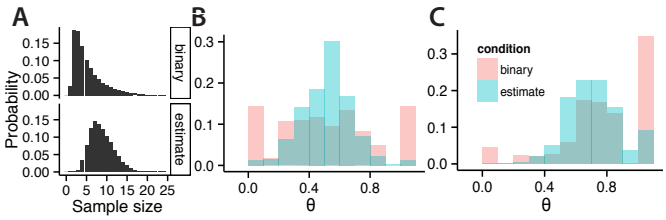


Figure 2: A: Distribution of sample size by condition, using uniform θ. B and C: Predicted distribution of sample averages, using $\theta = 0.5$, and $\theta = 0.7$.

Summary This analysis shows how sampling goals can influence on learners' expected stopping strategies even in a very simple binary sampling task. Furthermore, even though the model gives learners no influence on *what* to sample and it assumes they are choosing optimally, the resulting distribution of samples is *not* necessarily representative of the underlying generating probabilities. Instead, sampling with the binary goal leads to far more samples with extreme averages than the estimation goal, which generates more representative samples.

This illustrates that even in natural sampling tasks, in which samples generally "conserve the properties of the universe" Fiedler (2008) optimal stopping rules can distort the experienced data depending on the learner's goal during the sampling. In other words, even the minimal decision of when to stop can introduce systematic discrepancies between a sample and the population distribution.

Experiments

These theoretical results yield a number of interesting empirical questions, which will be addressed with two experiments. First, the model uses a relatively sophisticated forward-looking, optimal decision-making process to govern people's stopping decisions. The behavioral patterns in Figure 1 require not only that learners can assess the cost-accuracy trade-off and adjust the number of samples accordingly. They need to also be able to assess the expected performance *given the current sample content*. In the following two experiments we will therefore examine to what degree participants are actually engaging in such forward looking behavior when deciding to terminate or continue sampling.

Both experiments will also examine how goals affect sample composition, compared to the predictions of the model. A

key intention of the modeling effort above was to show that distributions of sample means differ depending on the goal, which suggests that even natural sampling does not preserve population characteristics in a straightforward manner.

Another question, albeit one that is more difficult to answer, is whether people's judgments about the distribution that generates the samples are affected by the different sampling strategies. For example, since binary sampling, according to the model, should produce more extreme samples, does this lead people to expect more extreme outcomes? Experiment 2 starts to address this question and yields some preliminary results.

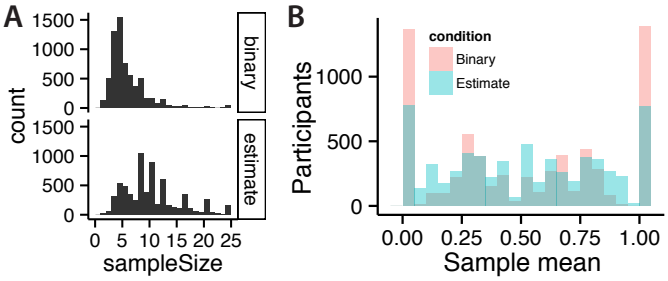


Figure 3: A: Histogram of sample sizes in Experiment 1. B:Distribution of sample averages in Experiment 1.

Experiment 1

Experiment 1 manipulated sampling goals in a simple repeated Bernoulli sampling task that shared all the characteristics of the task described in the modeling section. The sampling goal was manipulated between participants who either had to estimate the overall value of θ (*estimation condition*) or decide whether θ was greater or lower than 0.5 (*binary condition*).

We predicted that sampling goals would affect both the number of samples collected on average (higher for the estimation vs. binary goal), and the relationship between stopping points and current sample composition. As outlined in the previous section, we expected participants engaged in binary sampling to be more likely to stop early when strong evidence is encountered (average closer to 0 or 1) and more likely to continue when evidence is mixed (average closer to 0.5).

Participants 276 participants were recruited via Amazon Mechanical Turk. They were paid \$2 for participating with an option to win a bonus of up to another \$2 (explained below).

Stimuli Participants were told that they were repeatedly drawing cards from 50 different card decks consisting of 200 cards each. Cards could be either red or blue, and the distribution of red and blue cards in each deck was determined semi-uniformly (using an evenly-spaced distribution of Bernoulli probabilities θ that were then used to randomly draw the cards for each deck). Participants were explicitly told to assume a uniform distribution of numbers of red and blue cards in each deck, as well as being told that the order of cards in each deck was completely random ("well shuffled").

Procedure For each card deck, participants could repeatedly (up to 24 times) turn over cards using a button press to reveal their color. A counter on the screen would tell how much of potential bonus remained after each sample (which cost \$0.05). For each card deck, the potential bonus started at \$2. At any point participants could also decide to move on to the secondary task (binary choice or estimation) via a different button. In the binary task, participants would then make a two alternative forced choice decision of whether they thought there were more red or more blue cards in the deck. In the estimation task they gave an exact estimate using a bar slider whose handle had two lines that indicated the .2 interval within which their response would be counted as correct. After giving their estimate, all 200 cards of the card deck were revealed and participants received feedback on their choice or estimate. If their estimate was correct, their potential bonus was recorded as the bonus that remained after the sampling phase, otherwise the it was recorded as \$0. At the end of the task, one card deck was chosen at random and participants were actually paid the bonus earned on that deck.

Results & Discussion To recap, our two main predictions motivated by the modeling were that participants in the estimation conditions would on average collect larger samples, and that participants in the binary conditions would be strongly influenced by the composition of the current sample when deciding whether to stop or to continue sampling. Figure 3A shows the distribution of sample sizes by condition. As expected (compared to model predictions in Figure 2A), participants in the estimation group took larger samples than participants in the binary group ($t(211) = -11.18, p < 0.01$), suggesting that people were aware of the trade-off between sampling cost and accuracy and were willing to incur higher cost (more samples) in the more difficult task. There were also visible spikes for certain sample sizes (8, 10, and 12) in the estimation condition, suggesting that perhaps these were preferred, salient stopping points for a range of participants (cards on the screen were aligned in columns of four, so spikes at 8 and 12 may be the result of some aesthetically inspired fixed-sample-size stopping rule).

Figure 4 shows the proportion of times participants ended up in each state after terminating search. There was a clear difference between the stopping patterns of the two groups. Participants in the binary group were more likely to stop in early extreme states, but continued sampling if early evidence was mixed. Participants in the estimation group, on the other hand, showed no discernable stopping pattern based on current sample composition. For a given sample size, there appeared to be fairly equal stopping probability across different proportions of red/blue cards. Also, the previously mentioned average preference for certain sample sizes is reflected in diagonal "ridges" in the estimation goal data. What we did not observe were the earlier stopping patterns for very mixed or very extreme samples, which were predicted by the model.

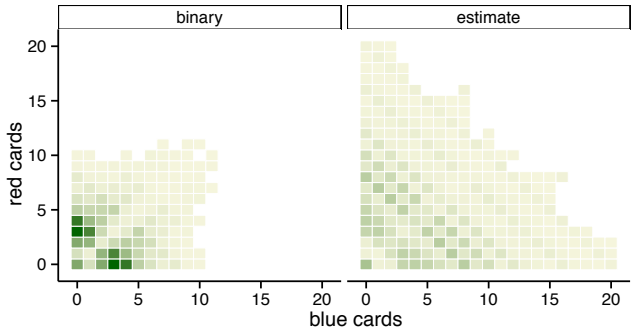


Figure 4: Stopping probabilities (dark means high) by sampling goal in Experiment 1

Possibly participants were unaware of the differences in uncertainty about the true average that emerge as a consequence of the sample composition.

Finally, we also examined the sample composition resulting from these stopping rule differences. Figure 3B shows the distribution of sample averages by condition. The binary sampling group showed a much larger proportion of extreme samples (all red/blue) compared to the estimation condition. In summary, people's stopping decisions lead to differences in the resulting sample content in a manner that is qualitatively consistent with the model predictions.

Experiment 2

Experiment 1 established that people adapt their stopping rule to different sampling goals in a way that is broadly consistent with the predictions of our optimal stopping analysis. However collecting a sample is usually not the end goal of learning. Instead learners use a sample to enable predictions about more general properties of the world. For example, people might listen to a few songs by a band then attempt to generalize from those song to the entire catalog. Experiment 2 was designed to explore how goal-induced stopping rule differences might affect people's subsequent judgments of more general population parameters.

A critical challenge was ensuring that participants were motivated to learn more general properties from the samples they collected. To that end, we altered the Experiment 1 design by grouping samples under two different generative processes (specifically two different schoolteachers who had the ability to influence the test scores of students in their class). The intention was to encourage participants to form summary representations of these parameters based on the samples collected for each. By varying the average outcome probabilities of these two groups/parameters (one high, one low), we would then be able to compare the potential impact of sampling goals on the estimates of each group's average. We expected binary samplers to be exposed to a greater difference between outcome probabilities due to a stopping rule that leads to more extreme outcomes and wanted to know if this would bias their population estimates to also be more more extreme (even higher, and even lower) as a consequence.

Participants 58 participants were recruited via Amazon Mechanical Turk and paid $2 for participation (plus bonus).

Stimuli To encourage participants to form summary representations of two separate distributions a cover story asked them to repeatedly sample binary results (correct or incorrect test scores) from a set of high school students, who were each taught by one of two teachers (teacher A or teacher B). That is, rather than randomly and uniformly drawing values for θ for each student (as in Experiment 1), outcome probabilities were now drawn from one of two hierarchical distributions. The mean accuracy of students from each teacher differed, such that one teacher yielded a higher average ($\mu_{high} = 0.7$) of correct test scores than the other ($\mu_{low} = 0.3$). Individual students' outcome probabilities were distributed as $\theta \sim Beta(3\mu_t, 3(1-\mu_t))$, depending on a student's teacher type, $t \in \{high, low\}$. Participants were each given different random draws from this distribution.

Procedure The main part of the experiment was similar to that of the model and previous experiment. Participants repeatedly sampled 100 students' test answers (correct or incorrect), while losing $0.05 per question from a potential $2 bonus per student. Each student was presented along with their respective teacher, which alternated between teacher A and B. Participants were again assigned to a binary choice or estimation condition, but unlike in the previous experiment they did not receive feedback on the true distribution for every student. After sampling 50 students per teacher, participants also rated the average quality of each teacher's students (using a slider between 0% and 100% correct). Again, they were paid a bonus based on one randomly chosen student from the sampling task.

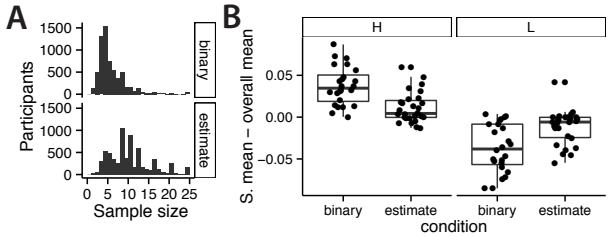


Figure 5: A: Number of samples collected in Experiment 2. B: Difference between mean of the sampling means, and mean of all samples by condition and teacher type (*H*igh or *L*ow).

Results & Discussion Figure 5A shows the distribution of sample size by condition and Figure 6 depicts participants' stopping probabilities by state. Despite substantial changes compared to Experiment 1 (addition of teachers, two different outcome distributions, omission of feedback) the results are qualitatively similar. Again, binary learners took smaller samples on average and were more likely to stop given highly

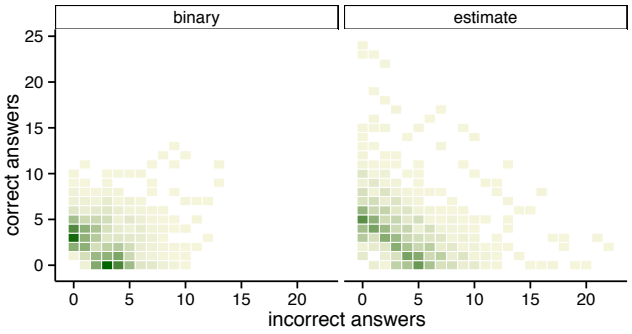


Figure 6: Stopping probabilities (dark means high) by sampling goal in Experiment 2.

unbalanced samples (all correct or all incorrect).

Next, we considered the impact of stopping on the evidence collected for each of the teachers. Because each participant saw a different sequence of outcomes and the resulting distribution of samples varied as a consequence, we developed the following measure of a participants' *potential* for being biased given the data they saw: For each participant, we computed the difference between the average sample mean for each teacher and the overall teacher average irrespective of the sample boundaries (i.e. treating all students as one sample). This difference is 0 when a participant always takes samples of the same size, but may diverge when sample sizes are uneven, particularly when sample size and sample mean are non-independent, as we expect to happen under binary sampling. Figure 5B shows this difference by teacher and condition. As expected, in the binary condition, the average sample mean was more extreme (higher for the good teacher, lower for the bad teacher) than the overall average, compared to the estimation condition, (significant condition and teacher type interaction, $t(98) = 5.62, p < 0.01$), indicating again that stopping rules affected the sample distribution.

Next, we tried to investigate the impact of this difference on people's subsequent judgments of the teachers. To do so, we regressed participants' posterior estimates of the teacher mean (one estimates per participant per teacher) on their overall teacher mean and the difference between overall mean and average sample mean (from Figure 5B). If the latter had any additional positive effect of peoples estimates beyond the overall mean, this would indicate that participants were not correcting for the unequal sample sizes that introduced this difference. However, we failed to find a significant positive effect of the difference, $t(99) = 1.627, p = 0.11$. Due to considerable variation in the actual outcomes observed by each participant (all received different sequences) and the high variability of theses posterior estimates, it may be that we lacked sufficient statistical power to detect this effect. We are currently working towards an improved design of this study that reduces variance in people's posterior estimates of teacher quality and their sample composition.

Discussion

In this paper, we presented a theoretical analysis showing how sampling goals can have a profound impact on people's stopping strategies even under natural sampling. Crucially, our results go beyond showing a difference in the *amount* of evidence collected, but demonstrate that goals also affect also the composition and statistical properties of samples. This demonstrates that natural sampling does not necessarily produce samples that mirror the statistical properties of the environment. Instead, just the simple decision of when to stop sampling can lead people to collect samples with vastly different distributional characteristics depending on their goal. While Experiment 2 did not show a robust effect of this non-representative sampling on more general predictions, it is an interesting question if people can take into account the process by which samples were gathered and perform the necessary correction to remove possible biases. Prior work on sampling and estimating binary outcomes suggests that this might be difficult for people (Fiedler, 2008; Griffin & Tversky, 1992). If this turns out to be more generally true, it suggest that simply deciding when to stop sampling information from a natural process can strongly bias judgement.

Acknowledgments This work was supported by grant number BCS-1255538 from the National Science Foundation and a John S. McDonnell Foundation Scholar Award to TMG.

References

Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, *14*(3), 227–243.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

Ferguson, T. S. (2012). *Optimal stopping and applications.* Electronic Text. Retrieved from `https://www.math.ucla.edu/ tom/Stopping/Contents.html`

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 186.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive psychology*, *24*(3), 411–435.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (in press). Information sampling behavior with explicit sampling costs. *Decision*.

Lee, M. D., & Paradowski, M. J. (2007). Group decision-making on an optimal stopping problem. *The Journal of Problem Solving*, *1*(2), 06.

Le Mens, G., & Denrell, J. (2011). Rational learning and information sampling: On the naivety assumption in sampling explanations of judgment biases. *Psychological review*, *118*(2), 379.

Rich, A. S., & Gureckis, T. M. (2014). The value of approaching bad things. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36rd annual conference of the cognitive science society.* Austin, TX.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Inferring Individual Differences Between and Within Exemplar and Decision-Bound Models of Categorization

Irina Danileiko (idanilei@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA 92617 USA

Abstract

Different models of categorization are often treated as competing accounts, but specific models are often used to understand individual differences, by estimating individual-level parameters. We develop an approach to understanding categorization that allows for individual differences both between and within models, using two prominent categorization models that make different theoretical assumptions: the Generalized Context Model (GCM) and General Recognition Theory (GRT). We develop a latent-mixture model for inferring whether an individual uses the GCM or GRT, while simultaneously allowing for the use of special-case simpler strategies. The GCM simple strategies involve attending to a single stimulus dimension, while the GRT simple strategies involve using unidimensional decision bounds. Our model also allows for simple contaminant strategies. We apply the model to four previously published categorization experiments, finding large and interpretable individual differences in the use of both models and specific strategies, depending on the nature of the stimuli and category structures.

Keywords: category learning; exemplar models; decision bound models; General Recognition Theory; Generalized Context Model; Bayesian inference, latent-mixture model

Introduction

The ability to categorize is widely regarded as a cornerstone of cognition (Murphy, 2002). It allows people to use innate or learned concepts to give meaning to stimuli, and provides a basis for the key cognitive capabilities of generalization and prediction. Existing theories of categorization vary in their assumptions about how people represent categories, and make decisions when categorizing stimuli. Accordingly, there are many different formal models of categorization, including prototype, exemplar, decision-bound, and rule-based models, as well as various hybrid models (Kruschke, 2008). Hybrid models involve representational assumptions that combine two or more approaches such as prototype with exemplar, or exemplar with rule-based representations.

A prominent exemplar model is the Generalized Context Model (GCM: Nosofsky, 1986), which assumes people store exemplars of each category in memory, attend to the relevant dimensions of the stimuli, and categorize a stimulus using similarity-based generalization from these exemplars. A prominent decision-bound model is General Recognition Theory (GRT: Ashby & Townsend, 1986), which assumes that people use a decision bound to partition the stimuli into discrete categories. Categorization seems likely to depend heavily on psychological components and processes such as memory capacity, attentional control, decision-making biases, and so on, all of which may vary across people. Accordingly, it seems reasonable to expect meaningful individual differences in categorization, and this expectation is supported by model-based and empirical evidence (e.g., Bartlema, Lee, Wetzels, & Vanpaemel, 2014; Soto, Vucovich, Musgrave, & Ashby, 2015). Previous work has studied different types of strategies within models like the GCM and GRT. For example, previous GRT modeling has emphasized the possibility that different people might use different decision bounds, including special cases like unidimensional horizontal or vertical bounds (Ashby & Gott, 1988; Maddox & Ashby, 1993).

In this paper, we present an approach to inferring the general models and specific strategies people use in categorization tasks. We do this by allowing for individual differences between the GCM and GRT models, and for individual differences in specific strategies, like unidimensional bounds, possible within each model. We develop a latent-mixture modeling approach that infers the model and strategy each person is using. Building on previous work in which both the GCM and GRT have been implemented as Bayesian graphical models (Lee & Wagenmakers, 2013; Danileiko, Lee, & Kalish, 2015), we implement our latent-mixture approach also as a graphical model, allowing for fully Bayesian inference. We apply our model to four existing categorization experiments–all involving stimuli that can be represented in terms of two underlying psychological dimensions–but with various types of stimuli and category structures. We find evidence for large individual differences both between and within models. We finish by discussing the implication of our model and results for future research in understanding how people represent categories.

Latent-Mixture Model

Our latent-mixture approach assumes that each subject uses one categorization model or specific strategy within that model, and that the overall data set is therefore a mixture of these specific components. We also allow for the possibility of contaminant subjects, who are guessing or repetitively assigning stimuli to the same category. Instead of filtering these people out, we model the contaminant behavior as another mixture component (Zeigenfuse & Lee, 2010).

Figure 1 presents a schematic graphical model that summarizes our approach. Each subject's categorization data $\boldsymbol{y}$ is generated by either the GCM, GRT, or a contaminant process. Within each of these general models, there are specific possibilities. For the GCM, either the original model with general

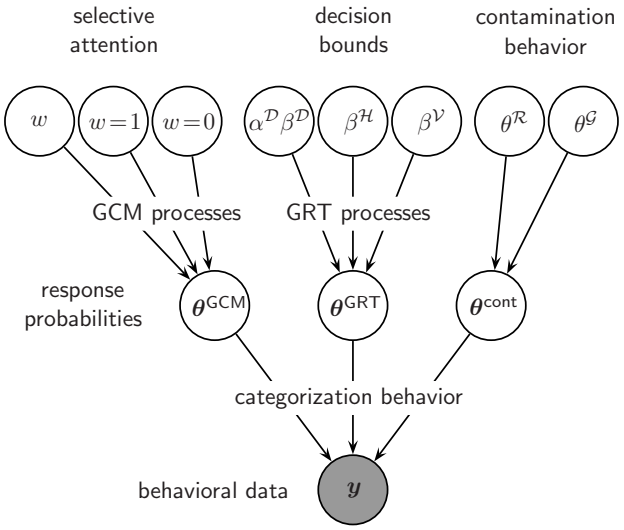


Figure 1: Schematic graphical model representation of the latent-mixture approach. A GCM or ,GRT, or contaminant categorization process generates the observed behavior of each individual. Within each model, special-case strategies involving the nature of selective attention, the decision bound, or the contaminant response probabilities are considered.

selective attention w is used, or attention is focused on only one of the stimulus dimensions, $w = 0$ or $w = 1$. For the GRT, either a general diagonal bound is used $\alpha^{\mathcal{D}}, \beta^{\mathcal{D}}$, or unidimensional horizontal $\beta^{\mathcal{H}}$ or vertical $\beta^{\mathcal{V}}$ bounds are used. For the contaminant processes, either a category is repeatedly chosen $\theta^{\mathcal{R}}$ or a guess is made on each trial $\theta^{\mathcal{G}}$.

GCM and Strategies

The GCM is an exemplar model that assumes people store all stimuli in memory and categorize a new stimulus by comparing it the stored stimuli. It is based on a similarity comparison between the presented stimulus and every other stimulus, using the concept of psychological distance (Shepard, 1957). If the stimuli are points in a two-dimensional coordinate space, psychological distance is defined as $d_{ij} = [\sum_{k=1}^{N} w_k |x_{ik} - x_{jk}|^r]^{\frac{1}{r}}$, where x_{ik} is the value of coordinate point x_i on dimension k, N is the number of dimensions, and r is either equal to 1 or 2 for separable-dimension or integral-dimension stimuli, respectively. The selective attention parameter w controls the level of attention given to one stimulus dimension. The distance is used to calculate the similarity $\eta_{ij} = (e^{-cd_{ij}})^{\gamma}$ so as the distance between points gets larger, their perceived similarity decreases exponentially. The generalization parameter c controls the steepness of the generalization gradient. The response determinism parameter γ_k controls probabilistic or deterministic responding (Ashby & Maddox, 1993). The final probability based on these processes can be affected by the bias toward each category. In our implementation, we assume there is no bias, and so set $b = \frac{1}{2}$. This means that the probability of responding J to stimulus i is equal to $\theta^{GCM} = \sum_{j \in C_j} \eta_{ij} / \sum_{K=1}^{m} (\sum_{k \in C_K} \eta_{ik})$.

In the full GCM, selective attention can range between 0 and 1. The special cases of $w = 0$ and $w = 1$ correspond to attending to just one of the two stimulus dimensions, and constitute theoretically interesting strategies. For example, if one stimulus dimension is shape and the other is color, one person might attend only to the shape dimension and place circles in one category and squares in the other. However, another person might attend only to the color dimension and place red shapes in one category and blue shapes in the other. A third person might attend to both dimensions and categorize red circles separately from blue squares. These possibilities correspond to the three GCM components included in our model, represented by the $w = 0$, $w = 1$, and w nodes respectively.

GRT and Strategies

The GRT model assumes that instead of storing each stimulus in memory, people partition the stimulus space into response regions divided by boundaries. Response probabilities are determined by these decision bounds, based on which region a noisy perception of the presented stimulus, $x_{pi} = x_i + \varepsilon_p$, belongs. Our model considers only linear decision bounds, although quadratic bounds have also been considered in the literature (Ashby & Maddox, 1992). A linear bound is defined as a discriminant function of the two dimensions satisfying the implicit line equation $h(x_1, x_2) = b_1 x_1 + b_2 x_2 + c$.

GRT assumes that there is variability in people's memory of the location of the bound. To account for this, the function is adjusted to include criterial noise ε_c. The function is compared to a bias parameter δ which captures bias toward a category. If $h(x_{pi}) + \varepsilon_c$ is smaller than δ, the response is category A. If it is larger than δ, the response is category B. If it is equal to δ, the response will be a guess between A and B. We again assumed no bias so that $\delta = 0$. Thus, the probability of a category A decision for stimulus i is $\theta^{GRT} = P(h(x_{pi}) + \varepsilon_c < 0)$.

Special cases of the general GRT model that have previously been emphasized involve unidimensional boundaries corresponding to vertical or horizontal lines. A vertical strategy is defined by an intercept value β^V, a horizontal strategy is defined by an intercept value, β^H, and a general diagonal boundary is defined by a slope and intercept α^D and $\beta^{\mathcal{D}}$. These possibilities correspond to the three GRT components included in our model. Although the vertical and horizontal strategies can be viewed as special cases of the diagonal strategy, one way to think about this in the latent-mixture approach is as a single model with a theoretically-rich prior. Including the vertical and horizontal boundaries as special cases corresponds to considering only a diagonal boundary with a prior that places significant density on boundaries with infinite and zero slope.

Contaminant Strategies

The three GCM strategies and the three GRT strategies make up six mixture components in our model. The remaining components capture contaminant subjects. In these cases, the probability values θ^{cont} do not follow from a theoretical

model, but are set directly. For guessing, the response probability is $\theta^{\mathcal{G}} = 0.5$, so that each category response is equally likely on every trial. For a repetitive contaminant behavior, the probability of a category A response is either $\theta_A^{\mathcal{G}} = 0.99$ or $\theta_B^{\mathcal{G}} = 0.01$, depending on which category choice is repeated. Adding these three contaminant possibilities leads to a total of nine components of our latent mixture model, with Figure 1 combining the two repeated contaminant possibilities.

Modeling Results

We implemented the graphical model in JAGS (Plummer, 2003), and used fully Bayesian methods based on MCMC sampling to make inferences.[1] Advantage of this methodological approach include accounting coherently for uncertainty about inferences, both in terms of model use and model-specifc parameters, and automatically controlling for the different complexity of the models and strategies considered (Lee & Wagenmakers, 2013).

We applied the model to four previously published categorization experiments. These experiments all involved a series of trials in which subjects viewed a stimulus and placed it into one of two categories, with corrective feedback after each trial. The stimuli used varied across experiments and include rectangles varying in size and interior line segment position (Kruschke, 1993), Shepard circles varying in size and radial lines (Bartlema et al., 2014), Gabor patches varying in frequency and orientation (Zeithamova & Maddox, 2006), and faces (Navarro, Lee, & Nikkerud, 2005). For the first three, there is a natural two-dimensional stimulus representation. For the faces, we assumed a two-dimensional representation based on multidimensional scaling modeling (Okada & Lee, 2016). Details of the experiments, including the number of subjects, blocks, nature of the experiment, and the various conditions, are presented in Table 2. The code for the graphical model, categorization data, detailed analysis of every subject in every experiment and condition, and other supplementary material is available on the Open Science Framework project page for this paper at `https://osf.io/ckwsn/`.

Overall Results

Table 1 summarizes our results by listing how many people are inferred most likely be using each of the possible models and strategies. The individual model-use inferences come are seen in the indicator variable in the JAGS code that assumes a uniform prior over all nine potential models, meaning that in the prior, each person is equally likely to use any of the nine. The "most likely" model for each person is taken from the posterior distribution of the indicator variable. There are four conditions in the Kruschke (1993) experiment: the first two are filtration category structures, in which the stimuli can be categorized correctly by using information from only one dimension, and the second two are condensation category structures, in which the stimuli can only be categorized correctly by using information from both dimensions. The majority of the 160 participants are inferred to use the GCM exemplar approach, but the specific selective attention strategy varies by condition. The Bartlema et al. (2014) experiment has two conditions, named after the category structures, both of which require information from both stimulus dimensions for correct categorization. The majority of the 65 participants use a decision bound approach. The Zeithamova and Maddox (2006) experiment has four conditions. The unidimensional condition is similar to the filtration condition in the Kruschke experiment and the information-integration condition is similar to the condensation condition. The "+ load" label in Table 1 indicates that that condition also involved a simultaneous working memory load task. The majority of the 170 participants use the decision-bound approach, with the vertical strategy being most common. This experiment involves the most contaminant subjects, who are inferred primarily to be guessing. The Navarro et al. (2005) experiment has four conditions. These involved categorizing faces based on gender, hair color, perceived level of trust, and a random condition with no logical structure. The majority of the 40 subjects use an exemplar approach, with selective attention that considers both available dimensions, but there is large individual variation over both models and strategies across the conditions.

Kruschke (1993) Results The results from the Kruschke (1993) experiment are shown at an individual level, for selected subjects, in Figure 2. The circles show the eight stimuli. The dark-colored circles show a response of category A while the light-colored circles show a response of category B. The size of the circle shows the number out of the total number of trials that each stimulus was placed in either category. The smallest circle means that stimulus was placed into that category exactly half of the time while the largest circle means that stimulus was placed into that category all the time. The bar graphs on the top of each panel show the uncertainty in the inference about which model and strategy the subject used. Each bar shows the posterior probability for a model or strategy. A tall bar showing one strategy means that we can be more certain of that person's inferred strategy than when there are shorter bars showing many strategies. The text at the bottom right corner of each panel indicates the inferred most likely strategy. For the general GCM w strategy, the 95% credible intervals and posterior mean for w are listed. For the GRT possibilities, the bound corresponding to the posterior mean is shown as a thick line, and the upper and lower bounds f to the 95% credible intervals are shown as thin lines.

These subject in Figure 2 are chosen to include at least one subject from each condition. The top-left came subject from the Filtration 1 condition, the top-middle subject came from the Filtration 2 condition, the top-right and bottom-left subjects came from the Condensation 1 condition, and the bottom-middle and bottom-right subjects came from the Condensation 2 condition. The first two subjects from the filtration conditions are inferred to be most likely using an exem-

[1] Our results are based on 3 independent chains with 100 samples each, collected after discarding the first 500 burn-in ones from each chain, and testing for convergence using the standard $\hat{R}$ statistic.

		Exemplar			Bound			Contam.	
		w	0	1	V	H	D	G	R
Kruschke	Filtration 1	10	-	30	-	-	-	-	-
	Filtration 2	6	30	1	3	-	-	-	-
	Condensation 1	15	4	5	-	4	8	4	-
	Condensation 2	19	6	10	2	-	1	2	-
	Total	**50**	**40**	**46**	**5**	**4**	**9**	**6**	**-**
Bartlema et al.	Diagonal	-	3	-	5	15	7	1	-
	Criss-Cross	-	3	4	8	8	4	7	-
	Total	**-**	**6**	**4**	**13**	**23**	**11**	**8**	**-**
Zeithamova & Maddox	Unidimensional	-	2	-	31	1	1	5	1
	Unidimensional + load	3	5	7	22	5	2	5	1
	Information-Integration	-	-	-	20	-	11	2	1
	Information-Integration + load	1	9	3	19	0	10	3	-
	Total	**4**	**16**	**10**	**92**	**6**	**24**	**15**	**3**
Navarro et al.	Gender	9	-	1	-	-	-	-	-
	Hair	3	3	1	-	3	-	-	-
	Trust	4	1	1	2	-	1	1	-
	Random	4	-	2	-	1	-	3	-
	Total	**20**	**4**	**5**	**2**	**4**	**1**	**4**	**-**

Table 1: Number of participants inferred to use an exemplar, decision bound, or contaminant strategy in each data set. (w: uniform w strategy; 0: w=0 strategy; 1: w=1 strategy; V: vertical; H: horizontal; D: diagonal; G: guess; R: repeat (either)

plar strategy with $w = 1$ and $w = 0$, with some possibility of the general GCM w strategy. The subjects from the condensation conditions are inferred to be more likely to use either a diagonal boundary or a general GCM w strategy with a mean value close to $w = 0.5$ in one case, and $w = 0.83$ in the other. The last subject is inferred to be a guessing contaminant, with a larger degree of uncertainty.

Bartlema et al. (2014) Results The results from the Bartlema et al. (2014) experiment are shown at an individual level in Figure 3. The top panels come from the diagonal condition and the bottom panels come from the criss-cross condition. The top-left subject is inferred to be using a horizontal boundary, but with some uncertainty about the possible use of a more general diagonal boundary. The bottom-left subject is also inferred to be using a horizontal boundary, but there is a possibility of a diagonal boundary, or a contaminant guessing strategy. The top-right subject is inferred to be using a diagonal boundary, with greater uncertainty. The bottom-right subject is inferred to be using a vertical boundary, also with a high level of certainty.

Zeithamova and Maddox (2006) Results The results from the Zeithamova and Maddox (2006) experiment are shown at an individual level in Figure 4. The top-left subject comes from the unidimensional condition. The top-middle and top-right subjects come from the unidimensional + load condition. The bottom-left subject come from the information-integration condition. The bottom-middle and bottom right subjects come from the information-integration + load condition. In this experiment, very few of the subjects were inferred to be using an exemplar strategy, perhaps as a result of the large number of stimuli required to keep in memory. Even though most subjects were inferred to be using a decision bound, there is still great variation in the specific shape of the boundaries, with varying slopes and intercepts. Two of the subjects selected for Figure 4 are inferred to be using a vertical boundary, even though they come from conditions with different category structures. Similarly, two of the subjects are inferred to be using a diagonal boundary, but one with more uncertainty than the other about the location of the boundary. This experiment also involved a large number of subjects inferred to be contaminants, one of whom is shown in Figure 4. The top-right panel shows one subject who was inferred to be using an exemplar approach with a w=1 strategy, although there is large uncertainty about this inference, consistent with the poor categorization performance shown.

Navarro et al. (2005) Results The results from the Navarro et al. (2005) experiment are shown at an individual level in Figure 5. The top-left subject comes from the gender con-

Experiment	# Su	# B	# St	# C	Type of St
Kruschke	160	8	8	4	Rectangles
Bartlema et al.	65	40	8	2	Shepard circles
Zeithamova & Maddox	170	5	80	4	Gabor patches
Navarro et al.	40	8	25	4	Faces

Table 2: Properties of the categorization experiments (Su: subjects; B: blocks; St: stimuli; C: conditions).

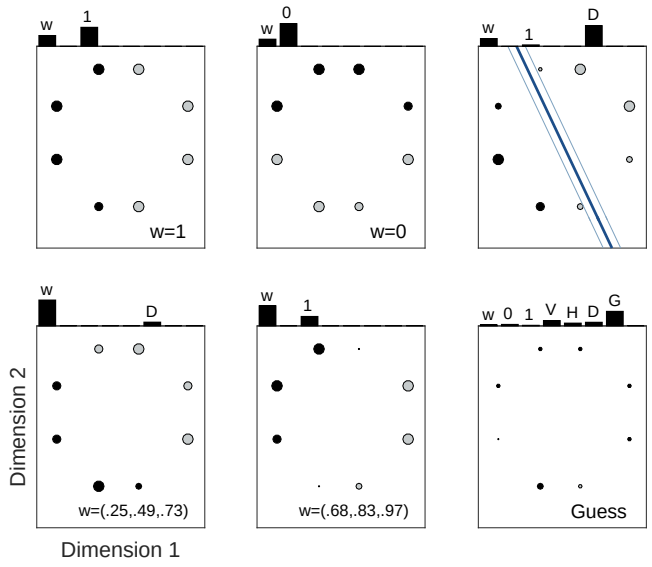


Figure 2: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Kruschke (1993) experiment.

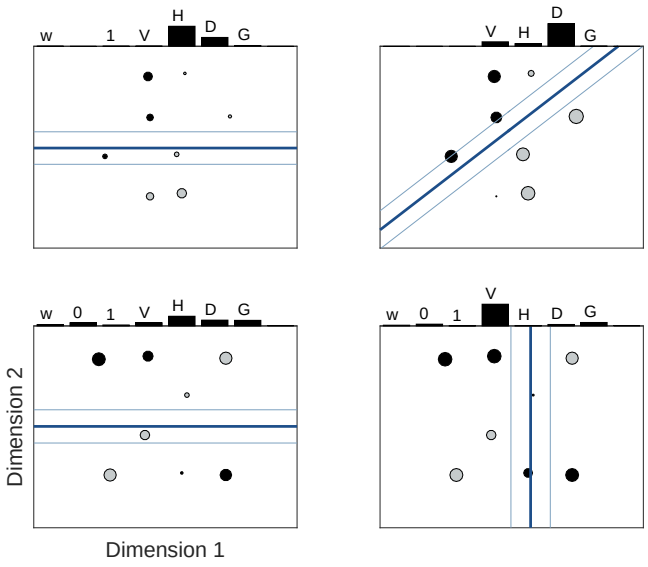


Figure 3: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Bartlema et al. (2014) experiment.

dition, the top-middle and top-right subjects come from the hair-color condition, the bottom-left subject comes from the trust condition, and the bottom-middle and bottom-right subjects come from the random condition. Most of the subjects are inferred to be using the GCM, perhaps as a result of the stimuli being faces and not easily separable into psychologically interpretable dimensions. Two of the selected subjects are inferred to be using the general GCM *w* strategy, with varying mean values depending on the condition. A few subjects are inferred, with less certainty, to be using a decision-bound approach. The random condition has the most contaminants, as for the subject in the bottom-right panel, typically with large uncertainty about model use.

Discussion

We have presented a latent-mixture model that infers which of the two models—the GCM or the GRT—each person is using, and whether they are using a specific strategy within that model. Our individual differences analysis showed that different people's categorization behavior can best be explained by different model strategies, depending on the types and number of stimuli involved, and the nature of the category structures. Instead of continuing a debate of a "one model fits all" answer where all behavioral data must be in accordance with one type of model, applications of our modeling approach to individual subject data can potentially reveal multiple models and strategies being used by different people.

Future work could apply our general method, and the specific model we have implemented and demonstrated, to other categorization experiments, exploring how individual differences change with the type of stimuli and category structures involves. It would be interesting to understand individual differences for more complicated real-world stimuli, such as faces, with the goal of understanding how people categorize in everyday life. It is straightforward to extend our model to include other theoretical accounts of categorization behavior, and different specific strategies within them. These could incorporate other categorization models, such as hybrid models that combine prototype with exemplar or rule-based representations. It would also be possible to extend the model to allow for shifts in categorization within an individual, allowing for possibilities like rapid shifts in attention, or the adaptation of an overly simple unidimensional bound to a more general diagonal bound on the basis of feedback. Examination of the strategy shifts that occur can be useful for further predictive modeling of when we can expect participants to switch strategies. Collectively, these extensions allow for broader and deeper investigation of the individual differences in the way people represent and use categories.

Acknowledgments

This work was supported by NSF Award 1431635.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensionalstimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 33–53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 50-71.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of cat-

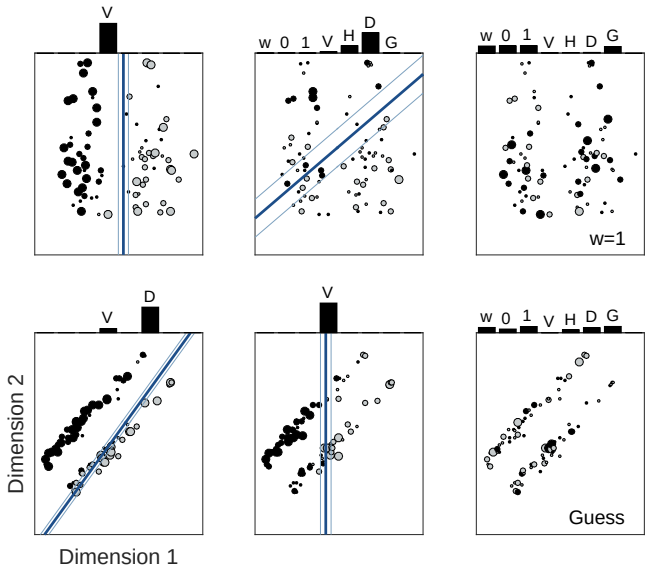


Figure 4: Inferred model or strategy use, and attention values or decision bounds, for selected subjects from the Zeithamova and Maddox (2006) experiment.

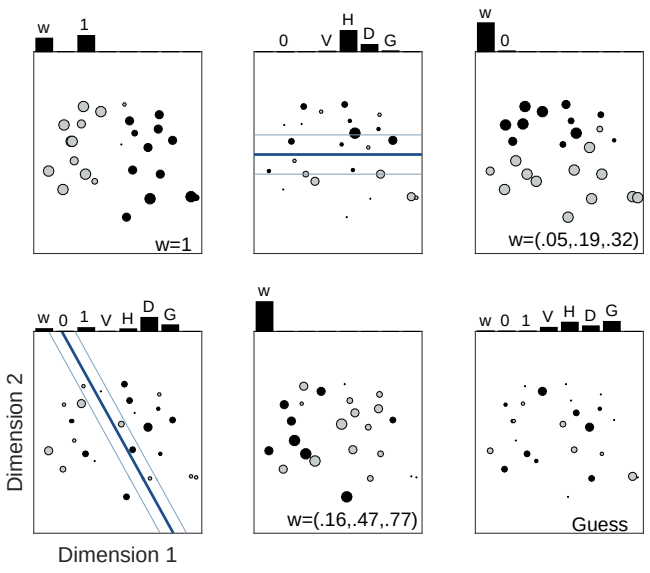


Figure 5: Inferred model or strategy use, and attention values or decision bounds for selected subjects from the Navarro et al. (2005) experiment.

egorization. *Journal of Mathematical Psychology*, *37*, 372-400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–79.

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150.

Danileiko, I., Lee, M. D., & Kalish, M. L. (2015). A Bayesian latent mixture approach to modeling individual differences in categorization using General Recognition Theory. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (p. 501-506). Cognitive Science Society.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 267–301). New York: Cambridge University Press.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*, 49–70.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Navarro, D. J., Lee, M. D., & Nikkerud, H. (2005). Learned categorical perception for natural faces. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1600–1605). Mahwah, NJ: Erlbaum.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental psychology: General*, *115*, 39–57.

Okada, K., & Lee, M. D. (2016). A Bayesian approach to modeling group and individual differences in multidimensional scaling. *Journal of Mathematical Psychology*, *70*, 35–44.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalizationto distance in psychological space. *Psychometrika*, *22*, 325–345.

Soto, F. A., Vucovich, L., Musgrave, R., & Ashby, F. G. (2015). General recognition theory with individual differences: a new method for examining perceptual and decisional interactions with an application to face perception. *Psychonomic Bulletin & Review*, 1–24.

Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, *133*, 283–295.

Zeithamova, D., & Maddox, W. T. (2006). Dual task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.

Adults' guesses on probabilistic tasks reveal incremental representativeness biases

Habiba Azab[1] **(hazab@mail.bcs.rochester.edu)**
David Ruskin[1] **(druskin@mail.bcs.rochester.edu)**
Celeste Kidd[1,2] **(celestekidd@gmail.com)**
[1]Department of Brain and Cognitive Sciences and [2]Center for Visual Sciences
University of Rochester, Meliora Hall, Rochester, NY, 14627

Abstract

Participants in most binary-choice tasks with multiple trials tend to probability-match (Vulkan, 2000) — i.e., provide responses that match the probability distribution of the presented population. Given a single trial, however, participants usually choose the majority option (James & Koehler, 2011). By using a method that visually presents the probabilities of the two competing options, we examine responses when participants are given only a single trial, and initial responses when participants are given multiple trials. While we still observe aggregate probability-matching in the multiple-trial condition, we find robust sequence effects in participants' initial responses, including robust maximizing behavior on the first response. This suggests that both maximizing in single-trial experiments and aggregate probability-matching in multiple-trial ones can be explained by a single, underlying mechanism; one that seeks to provide a representative sample at each point during sequence generation.

Keywords: Decision making; statistics; psychology; human experimentation; probability-matching; maximizing.

Participants Adopt Sub-Optimal Strategy in Simple Binary-Choice Tasks

Economics, as well as daily dealings and discourse, rest upon a fundamental assumption of human rationality: that individuals will make choices that maximize their expected benefit. However, in one of the simplest experimental setups—the binary-choice task—adult participants fail to conform to this assumption. In this task, participants are repeatedly asked to choose between two options, each exclusively offering a reward with a certain probability (for example, an orange button is rewarding on 70% of trials and a green button on the remaining 30%). These probabilities are fixed, and the distribution of rewards across trials is independent of reward history or participants' performance. The optimal strategy—the one that maximizes total payout—is to select the option with the highest probability (the majority or dominant response) on every trial (e.g. given 10 trials, press the orange button exclusively). Yet participants tend to probability-match, matching the distribution of their responses to the reward distribution of the options (e.g. pressing the orange button 7 times and the green button 3 times). This result is well established (Vulkan, 2000 for a comprehensive review; Grant, Hake & Hornseth, 1951; Hake & Hyman, 1953; Gardner, 1957; Rubinstein, 1959; West & Stanovich, 2003). Previous research has demonstrated that adults may be coaxed more towards maximizing by using higher payoffs (Toda, 2009), risk of losing gains (Siegel & Goldstein, 1959; Siegel, 1961), detailed feedback (Friedman & Massaro, 1998; Birnbaum & Wakcher, 2002), an extensive number of trials (Edwards, 1965), emphasizing the random nature of the task (Goodnow, 1955), or some combination of these (Shanks, Tunney & McCarthy, 2002). Even then, however, not all adults maximize.

Shanks, Tunney and McCarthy (2002) argue that the typical experimental setup, with a large number of rapid, successive binary choices—typically 100 to 400 (Vulkan, 2000) and as many as 1,000 (Edwards, 1961)—might be ecologically questionable. Humans are not typically faced with the same choice hundreds of times on a single day, especially with static reward probabilities. They predict that participants might maximize if given fewer, infrequent choices. To test this, James and Koehler (2011) presented participants with a sequence of ten statistically identical choices in the guise of ten different game setups—emphasizing the unique nature of each gamble. They predicted that participants would be more likely to maximize in the single-shot games. True to this intuition, participants were more likely to maximize in the unique-gambles condition and probability-match in the repeated-gambles condition.

James and Koehler (2011) suggest that the mechanism underlying adults' guessing behavior depends on their expectations about how many guessing opportunities they will have. In essence, adults could employ two distinct strategies depending on how many guesses they will have to make—maximizing when they have only one, and probability-matching when they have many. We propose an alternative hypothesis: the *incremental-representativeness hypothesis*, inspired by Kahneman and Tversky's representativeness bias (1972). This hypothesis proposes a trial-by-trial mechanism where participants provide a sample that is *incrementally* representative of the population—i.e. at each point of the experiment. We believe this hypothesis is capable of explaining maximizing behavior given a single trial, probability-matching behavior over several trials, and order effects observed in the latter setting, particularly over participants' first few responses.

The Incremental-Representativeness Hypothesis

More specifically, the incremental-representativeness hypothesis suggests that participants observe the distribution of the population and, at each guess, aim to make their previous responses representative of that population. This hypothesis makes particularly strong predictions for responses in the first few trials of an experiment. At the first guess in a series, participants are expected to maximize—i.e. choose the more probable option—regardless of how many guesses are

to come afterwards. This is because, given only one trial, the majority choice is more representative of a skewed distribution than the minority one. Over several participants, order effects would be most strongly observed during initial trials—more specifically, the minimum number of samples necessary to fully describe the distribution (e.g. 3 in a 2:1 distribution, 5 in a 2:3 distribution, 100 in a 99:1 distribution, etc).

In contrast, trial-by-trial probability-matching predicts that the likelihood of a particular response remains constant throughout the experiment, and matches the probability of that response in the population. For example, in a population of 70% orange and 30% green gumballs, the trial-by-trial probability-matching hypothesis predicts that participants are 70% likely to guess orange and 30% likely to guess green at each trial during the experiment, regardless of the order of that trial in the overall sequence of responses. The underlying mechanism behind this hypothesis is best visualized by picturing the participant flipping a weighted-coin: the coin is always (say) 70% likely to turn up heads and 30% likely to turn up tails, regardless of how many times it is flipped, or which flip in the sequence we are at.

Note that both the incremental-representativeness hypothesis and the trial-by-trial probability-matching hypothesis would give rise to the same aggregate sample statistics, where the proportions of the two available alternatives match their proportions in the given population. These hypotheses differ, however, in the predictions they make at the level of a single-trial. While the incremental-representativeness hypothesis can explain both aggregate probability-matching (given several trials) and maximizing (given only one trial), probability-matching at the single-trial level fails to explain maximizing given only one trial, and predicts no significant order effects.

We tested our hypothesis with a simple binary-choice task (Experiment 1). We randomly assigned adult participants to either a single-trial or a multiple-trial condition. We predict that adults will maximize on the first trial, regardless of whether the first guess was their only guess. We also predict that participants will probability-match over the entire sequence of given trials, although not on each individual trial, in the 10-trial condition. Alternatively, participants could probability-match on each individual trial of the experiment (including the first), supporting the trial-by-trial probability-matching hypothesis, and suggesting that participants utilize different guessing strategies depending on the number of trials they expect to be given.

Participants in our single-trial condition, as a group, maximized—replicating James and Koehler's (2011) findings and providing an experimental estimate of maximizing behavior. In line with our predictions, participants in the multiple-trial condition also maximized on the first trial, rather than probability-matched. This result suggests that maximizing behavior is not elicited by unique gambles, but by the first gamble in general. Responses in the first three trials of the experiment suggest that, rather than probability-matching on individual trials, participants seemed to provide incremental samples that were representative of the population not only after all trials were completed, but after each successive trial.

We further tested this hypothesis by asking an independent pool of participants to evaluate the likelihood of different sequences being drawn at random from the same population (Experiment 2). We find that the sequences participants tended to provide in Experiment 1 were rated highly by participants in Experiment 2. These results further support the incremental-representativeness hypothesis.

Experiments

We chose to present our binary-choice population using a static image of a gumball machine, filled with 70% orange gumballs and 30% green gumballs. We chose this particular task design for three main reasons. First, unlike the lights task (Anderson, 1960; Derks & Paclisanu, 1967)—in which the probability distribution is only apparent after some number of trials,—the gumball machine provides an immediate, non-sequence-dependent display of the distribution of the outcomes. This offers three distinct benefits: 1) it places fewer demands on working memory, 2) it avoids the risk of introducing unintentional and entirely random sequential patterns in the initial presentation of the population's distribution, and 3) unlike most previous experimental setups, it requires no training to learn the distribution of the population, allowing us to look at participants' responses from the very first trial onwards, where we are most likely to see the order effects predicted by our hypothesis[1].

Second, the gumball machine makes the primary assumptions of the task intuitive and reasonable: that outcomes are random and independent, and not subject to human control. Similar experimental setups have been used in recent studies with children, in which probabilities were represented using ping pong balls in clear containers (Xu & Garcia, 2008; Xu & Denison, 2009), and more recently using a gumball machine (Denison & Xu, 2014). Finally, we chose to provide no feedback until the end of the experiment to eliminate both pattern-seeking and 'win-stay, lose-shift' behaviors.

However, this framework is not without certain drawbacks. Drawing gumballs from a gumball machine is a process of sampling without replacement. We thus did not give participants any more than 10 trials, ensuring that sampling from the gumballs would not skew the population's distribution significantly. Also, like many modern experiments, ours relies on a computer-simulated task, not an actual gumball machine. This could interfere with participants' perceptions of randomness. Ideally, this experiment would be replicated with an actual gumball machine in a natural setting with replacement, to dispel any notions the participants might have about our interfering with the outcome of trials.

It is possible that the order effects we observe in these

[1] While past experiments have reported participants' responses in the initial trials, these are essentially random as participants are in the process of learning the underlying population distribution.

data are somehow due to the specifics of the task—such as the sequential nature of responding, by which participants give their guesses one at a time and are not shown what they guessed previously, nor are they given any feedback (to match the 1-trial and 10-trial conditions of this experiment as closely as possible). To eliminate this hypothesis, we ran a follow-up experiment. In Experiment 2, we ask a different group of participants to evaluate the likelihood of various sequences of gumballs, while emphasizing the order in which they come out of the machine through a simple visual animation. This relieved participants of the need to remember previous guesses, and allowed us to determine whether participants rate highly the same sequences that they are likely to respond with in the first task. This manipulation also gives us an experimental measure of what sequences participants actually consider to be representative.

Method

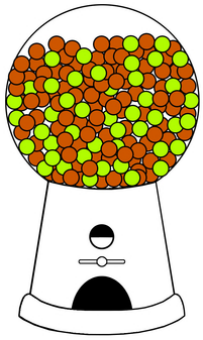


Figure 1: Gumball machine image used in both experiments

Experiment 1

173 participants[2] (mean age = 32 years, range = 18-66 years old), recruited through Amazon Mechanical Turk, were presented with the image of a gumball machine (Figure 1), in which 70% of gumballs are orange and 30% are green[3]. They were asked to look at the machine carefully and guess what color gumball would come out of it next. They were also told that they would be given a bonus based on their good performance and feedback. Participants were assigned to conditions at random; 86 participants were allowed only 1 guess, while 87 participants were allowed 10 guesses. The number of guesses remaining was shown throughout the experiment. Two gumball-shaped (i.e., round) buttons—one orange and one green—were displayed at the bottom of the screen; participants were required to choose one, then click the "Guess" button to make their guess and, in the 10-trial condition, proceed to the next trial. The locations of the buttons indicating the two choices were counterbalanced; in the 10-trial group, the locations were also switched at random between trials. At the end of the experiment, participants were given "feedback" about their performance; this was, in fact, unrelated to the guesses they made and standardized for all participants ("You guessed correctly!" for 1-trial participants, and "You guessed correctly on 8 trials!" for 10-trial participants) [4]. Participants were then asked to estimate the percentage of orange and green gumballs in the machine, and to provide a brief description of the strategy they used to make their choices.

Experiment 2

In a follow-up experiment, 154 participants[5] (mean age = 34 years, range = 19-68 years old), also recruited through Amazon Mechanical Turk, were presented with the same gumball-machine image shown in Figure 1. They were then shown a sequence of 1 to 4 gumballs, asked to look at the machine carefully, and then rate the likelihood of the shown sample on a 7-point scale (1: least likely, 7: most likely). They were presented with all 30 possible color permutations of 1, 2, 3 or 4 gumballs, as well as 4 catch trials with purple (distractor) gumballs to assess attentiveness. (Since the image contained no visible purple gumballs, attentive participants would be expected to assign these sequences low probabilities.) After all trials were completed, participants were asked to estimate the percentage of each color in the machine.

Results

Experiment 1

Participants estimated the percentage of orange gumballs at 66.07% [6] (median = 65%, standard deviation: 7.58%). This was slightly lower than the actual percentage of orange gumballs (70%). This could be due to differences in the brightness of the colors presented, or indicating that participants might be estimating the closest ratio (2:1), rather than a percentage. For all subsequent analyses, we will use the mean proportion of orange gumballs that participants predicted (0.66).

[2] 33 additional participants were excluded from all analyses due to failure to understand the task or inattentiveness: either by estimating both options to be equally distributed, or by estimating the percentage of green gumballs to be higher than that of orange ones.

[3] We never explicitly refer to the colors of the gumballs as "orange" or "green", since different monitors might display these shades differently. For all responses, we ask participants to click on a button of the same color as the gumball.

[4] Feedback at this stage of the experiment could not affect participants' performance, since they had already completed all trials.

[5] 57 additional participants were excluded from all analyses due to failure to understand the task or inattentiveness. Given that this was a longer task, we define more stringent criteria for subject inclusion. 13 participants were excluded for rating combinations with a purple (distractor) gumball at 4 or higher (where 7 = 'very likely'), 6 more participants were excluded for providing estimates of the percentage of orange and green gumballs that did not add up to 100 +/- 5%, an additional 3 participants were excluded for estimating the percentage of green gumballs to be higher than that of orange gumballs, and the rest (35 participants) were excluded for estimating the percentage of green gumballs to be equal to that of orange gumballs. While we realize this is a high percentage of participants to exclude, all reported effects remain even when no participants are excluded.

[6] 15 participants provided estimates of orange and green gumballs that did not add up to 100%, but seemed to indicate a ratio rather than a percentage. These responses were converted to percentages.

We note, however, that the pattern of results remains the same when the actual proportion (0.7) is used instead.

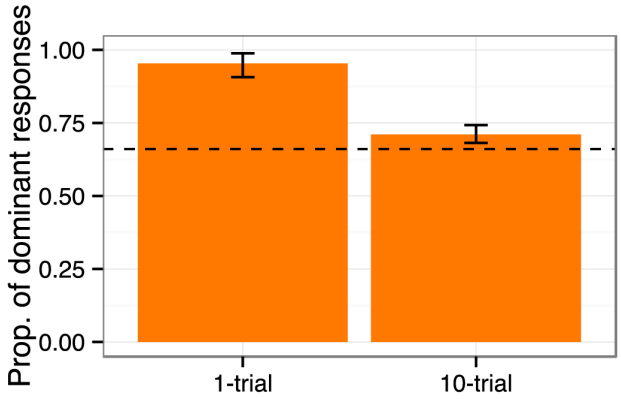


Figure 2: Comparing the mean of majority (orange) responses in the 1-trial and 10-trial conditions. Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

Results show that, overall, participants maximized in the 1-trial condition, and probability-matched in the 10-trial condition. Figure 2 shows the mean proportion of majority responses from participants in the 1-trial and 10-trial conditions; these means were significantly different (Wilcoxon rank-sum test: W=6601, $p < 10^{-15}$). The mean proportion of majority responses in the 1-trial condition was significantly larger than 0.66 (Wilcoxon signed rank test: V=3403, $p < 10^{-12}$), while the same proportion from the 10-trial condition was not significantly different from 0.66 (Wilcoxon signed rank test: V=2176, $p = 0.265$). In the 10-trial condition, we were also able to compare the proportion of majority responses per subject against their own estimated fraction of orange gumballs. These do not differ significantly, either (Wilcoxon signed rank test: V=1279, $p = 0.110$). This indicates that while, overall, participants in the 10-trial condition are probability-matching, participants in the 1-trial condition are maximizing. This replicates previous findings (James & Koehler, 2011).

Figure 3 shows the proportion of participants who gave each possible number of majority responses. This plot shows that probability-matching is not a behavior observable only at the population level; the majority of our participants probability-match. It is worth noting, however, that a number of participants (13) maximized—choosing the majority response exclusively. We re-ran all analyses while excluding this population of maximizers, and all results remain qualitatively the same.

We then examined the proportion of majority responses in each individual trial of the 10-trial condition (Figure 4). If participants were strictly probability-matching at the trial level, we would expect to see all of these proportions at or around 0.66. However, the fraction of majority responses in

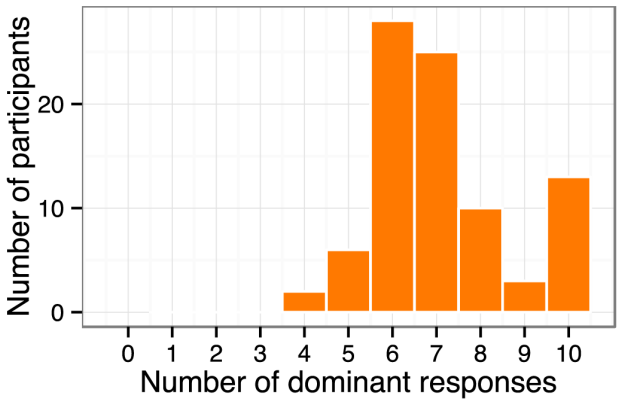


Figure 3: Proportion of 10-trial participants who gave 'x' majority-option (orange) responses over entire experiment (n=87)

the first and second trials were significantly higher than 0.66 (Wilcoxon signed rank test, 1st trial: V=3240, $p < 10^{-9}$; 2nd trial: V=2850, $p < 10^{-4}$). This fraction then decreased below 0.66 on the third trial (Wilcoxon signed rank test, V=1275, $p = 0.0052$), after which sequence effects seem to fade. Note that 3 trials are sufficient to represent a roughly 2:1 distribution.

Finally, we used our estimate of maximizing behavior from the 1-trial condition to analyze behavior on the first trial in the 10-trial condition. We found no significant difference between the proportion of majority responses chosen on this trial, and on the only trial in the 1-trial condition (Figure 5; Wilcoxon rank-sum test: W=3868, $p = 0.364$). This suggests that participants utilize the same strategy on the first trial of an experiment, regardless of whether they have more trials left.

Experiment 2

The survey data show a general correspondence between the initial sequences (length 1-4) participants are most likely to respond with in Experiment 1, and ones they tend to rate as most likely in Experiment 2. Figure 6 shows a plot of the average normalized rating per sequence[7], and the likelihood of that sequence in the dataset previously discussed. Note particularly the high average ratings for '1'('orange'), '11'('orange, orange'), and '110'('orange, orange, green'). Compare these to the relatively low rating for '011'('green, orange, orange'), which is objectively as likely as '110', but rated much lower. This suggests a strong order bias in both participants' responses and evaluations.

Discussion

We presented two groups of participants with a binary-choice task in which they had to predict which color gumball would

[7]Results are qualitatively similar when raw, un-centered ratings are used.

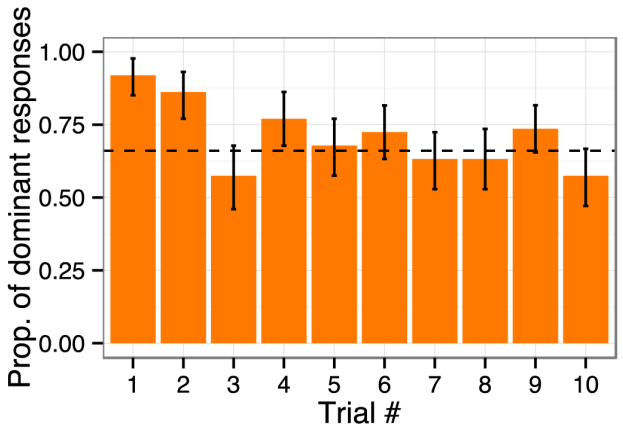


Figure 4: Proportion of majority responses at each successive trial in the 10-trial condition (total: 87 participants). Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

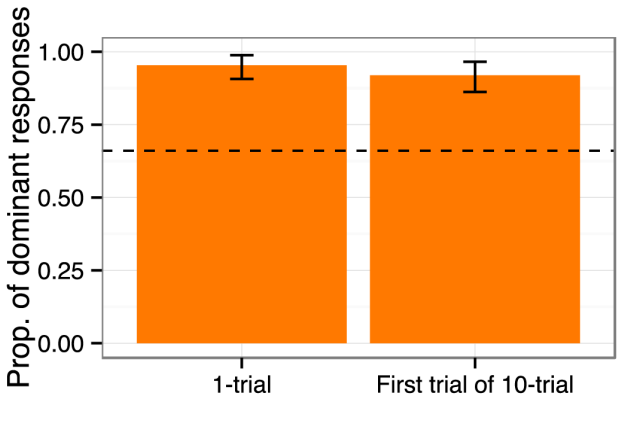


Figure 5: Comparing the mean proportion of majority (orange) responses in the only trial of the 1-trial condition and the first trial of the 10-trial condition. Dashed line indicates participants' estimate of the proportion of orange gumballs (66%). Error bars are bootstrapped 95% confidence intervals.

come out of a gumball machine next. Participants in the multiple-trial group were asked to guess the colors of the next 10 gumballs, while participants in the single-trial group were only asked to guess once. As a population, participants the single-trial condition collectively maximized, while participants in the multiple-trial condition were more likely to probability-match. Interestingly, though, participants in the multiple-trial condition show strong sequence effects: usually favoring the majority option on the first two trials, and the minority option on the third. This creates a representative sample of the population after only three trials. This finding supports our hypothesis—the incremental-representativeness hypothesis—positing the existence of a single mechanism that seeks to create incrementally representative samples at each trial of the experiment; explaining the aggregate patterns of maximizing and probability-matching we observe (in the single-trial and multiple-trial conditions, respectively), as well as the order effects we see in initial responses.

Participants in Experiment 2 also preferred sequences that started with the majority gumball. Especially notable are sequences where the non-ordered probabilities should be identical, such as '110'and '011'. As highlighted in Figure 6, though, both free-responses and ratings over sequences show a significant preference for the order that begins in a more representative manner. Sequences that do not begin with a majority gumball are regarded as less likely. This suggests that Kahneman and Tversky's representativeness heuristic is employed from the very first trial. Participants aim to make every sample—even a sample of one—a faithful representation of the distribution of the population. Further experiments with different population distributions and trial numbers are planned in order to verify this assumption.

Much like many prior studies, our results also show that not every participant uses the same strategy. For example, as in Figure 3, a few participants in the multiple-trial condition chose to maximize, and often justified this strategy. One subject explicitly stated that they thought if they chose the majority choice in every trial, they would "get more correct guesses". Other participants in this condition, while probability-matching and explicitly referring to the probability distribution in their comments, did not display any sequential effects. They chose the majority response consistently for the first fraction of trials, then chose the minority response exclusively for the rest. In spite of the above, we have chosen to report aggregate analyses whenever possible. While an aggregate analysis across participants may gloss over individual idiosyncrasies, it does reveal important patterns of behavior, and point to the causal mechanisms behind them.

Conclusion

While numerous studies have been conducted using binary-choice tasks, with participants' responses varying with a myriad of experimental variables, the mechanisms underlying participants' behaviors on these tasks remain poorly understood. In this study, we propose a unifying mechanism that explains seemingly different behavior when participants are given one vs. more trials, as well as initial order effects we observed when subjects were asked to give multiple responses. We explored these phenomena particularly in the absence of training or feedback. We found that participants given a single trial tend to maximize, while participants given multiple trials probability-match, replicating previous results (Vulkan, 2000; James & Koehler, 2011). However, we found no significant difference between participants' responses in the single-trial condition and those in the first trial of the multiple-trial condition, indicating that maximizing behavior is not exclusively a response to unique gambles. This sug-

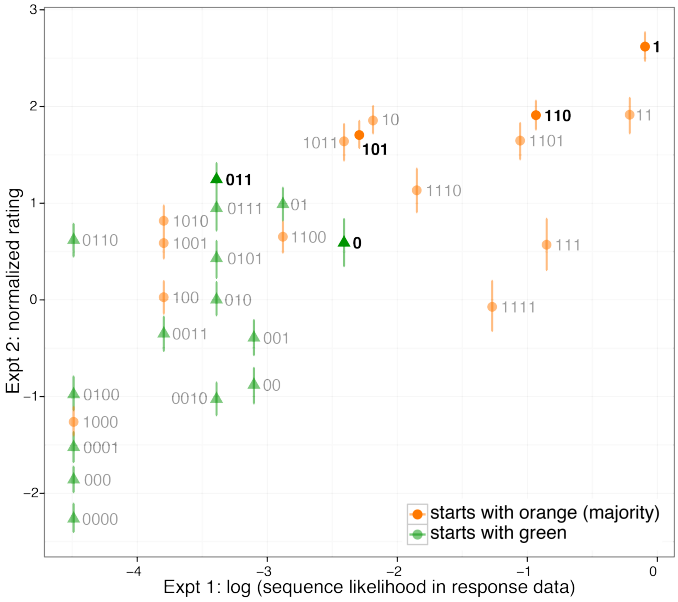


Figure 6: Average normalized rating of each possible sequence (obtained from Experiment 2), versus the likelihood of participants generating this sequence initially (obtained from Experiment 1), in log scale. These two variables are significantly correlated (Pearson correlation coefficient = 0.536 $\pm$ 0.021, $p < 10^{-15}$)

gests a single, underlying mechanism responsible for generating behavior in both of these conditions. While participants seem to probability-match over the course of the entire experiment, they do not do so exclusively for each trial. We also observe sequence effects—particularly in the initial three trials—suggesting that participants attempt to provide samples that are not only representative of the population distribution over the course of the entire experiment, but are also incrementally representative. We call this effect the incremental-representativeness hypothesis—participants seek to provide a sequence of responses that, truncated at any point during the experiment, still produces a sample representative of the population. This is in contrast to a trial-by-trial probability-matching mechanism, which fails to explain maximizing behavior observed in single-trial experiments.

Author Contributions

D.R. developed the study concept. All authors contributed to the study design. Testing and data collection were performed by H.A. H.A. performed the data analysis and interpretation under the supervision of C.K. H.A. drafted the manuscript, and D.R. and C.K. provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank the University of Rochester for funding this research. Special thanks to Frank Mollica and Amanda Yung for their assistance in setting up and running the task, and Steve Piantadosi and members of the Kidd Lab for helpful feedback and comments regarding this work.

References

Anderson, Norman H (1960). "Effect of first-order conditional probability in a two-choice learning situation." In: *Journal of Experimental Psychology* 59.2, pp. 73–93.

Birnbaum, Michael H & Sandra V Wakcher (2002). "Web-based experiments controlled by JavaScript: An example from probability learning". In: *Behavior Research Methods, Instruments, & Computers* 34.2, pp. 189–199.

Denison, Stephanie & Fei Xu (2014). "The origins of probabilistic inference in human infants". In: *Cognition* 130.3, pp. 335–347.

Derks, P L & M I Paclisanu (1967). "Simple Strategies in Binary Prediction by Children and Adults." In: *Journal of Experimental Psychology*.

Edwards, Ward (1961). "Probability learning in 1000 trials." In: *Journal of Experimental Psychology* 62.4, pp. 385–394.

— (1965). "Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing". In: *Journal of Mathematical Psychology* 2.2, pp. 312–329.

Friedman, Daniel & Dominic W Massaro (1998). "Understanding variability in binary and continuous choice". In: *Psychonomic bulletin & review* 5.3, pp. 370–389.

Gardner, R Allen (1957). "Probability-Learning with Two and Three Choices". In: *The American Journal of Psychology* 70.2, p. 174.

Goodnow, Jacqueline Jarrett (1955). "Determinants of Choice-Distribution in Two-Choice Situations". In: *The American Journal of Psychology* 68.1, p. 106.

Grant, David A, H W Hake & John P Hornseth (1951). "Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement." In: *Journal of Experimental Psychology* 42.1, pp. 1–5.

Green, C S et al. (2010). "Alterations in choice behavior by manipulations of world model." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.37, pp. 16401–16406.

Hake, H W & R Hyman (1953). "Perception of the statistical structure of a random series of binary symbols." In: *Journal of Experimental Psychology*.

James, Greta & Derek J Koehler (2011). "Banking on a bad bet. Probability matching in risky choice is linked to expectation generation." In: *Psychological Science* 22.6, pp. 707–711.

Kahneman, Daniel & Amos Tversky (1972). "Subjective Probability: A Judgment of Representativeness". In: *The Concept of Probability in Psychological Experiments*. Dordrecht: Springer Netherlands, pp. 25–48.

Rubinstein, Irvin (1959). "Some factors in probability matching." In: *Journal of Experimental Psychology* 57.6, pp. 413–416.

Shanks, David R, Richard J Tunney & John D McCarthy (2002). "A re-examination of probability matching and rational choice". In: *Journal of Behavioral Decision Making* 15.3, pp. 233–250.

Siegel, Sidney (1961). "Decision Making and Learning under Varying Conditions of Reinforcement". In: *Annals of the New York Academy of Sciences* 89.5, pp. 766–783.

Siegel, Sidney & Donald Aaron Goldstein (1959). "Decision-making behavior in a two-choice uncertain outcome situation." In: *Journal of Experimental Psychology* 57.1, pp. 37–42.

Toda, M (2009). "Guessing Sequence under Various Conditions of Payoff". In: *Japanese Psychological Research*, p. 11.

Vulkan, N (2000). "An economist's perspective on probability matching". In: *Journal of economic surveys*.

West, R F & K E Stanovich (2003). "Is probability matching smart? Associations between probabilistic choices and cognitive ability". In: *Memory & Cognition*.

Xu, Fei & Stephanie Denison (2009). "Statistical inference and sensitivity to sampling in 11-month-old infants". In: *Cognition* 112.1, pp. 97–104.

Xu, Fei & Vashti Garcia (2008). "Intuitive statistics by 8-month-old infants." In: *Proceedings of the National Academy of Sciences of the United States of America* 105.13, pp. 5012–5015.

A Bayesian Metric for Network Similarity

Richard M. Shiffrin (shiffrin@indiana.edu)
Psychological and Brain Sciences / Cognitive Science Program
1101 E 10th Street, Bloomington IN 47405 USA

Keywords: Network Similarity, Network Inference, Network Comparison, Bayesian Methods

Abstract

Networks of every kind and in numerous fields are omnipresent in today's society (e.g. brain networks, social networks) and are the intense subject of research. It would be of great utility to have a computationally efficient and generally applicable method for assessing similarity of networks. The field (going back to the 1950s) has not come up with such a method (albeit a few moves in this direction exist, such as Jaccard coefficients, QAP--quadratic assignment procedure, and more recently Menezes & Roth, 2013, and Asta & Shalizi, 2014). I present a Bayesian-based metric for assessing similarity of two networks, possibly of different size, that include nodes and links between nodes. I assume the nodes are labeled so that both the nodes and links between two nodes that are shared between the two networks can be identified.

The method calculates similarity as (a monotonic transformation of) the odds that the two observed networks, termed V and W, were produced by random sampling from a single master network, termed G, as opposed to generation by two different but similar networks, termed Gv and Gw. The simplest form of the method ignores strengths that could be assigned to nodes and links, and considers only nodes and links that are, or are not, shared by the networks. Suppose there are n_V nodes and N_V links only in V, n_W nodes and N_W links only in W and n_c nodes and N_c links shared between the networks. Thus the number of nodes in V is $n_c + n_V$ and the number in W is $n_c + n_W$. The number of unique nodes in both V and W is $n_c + n_V + n_W = n$. The number of links in V is $N_c + N_V$ and the number in W is $N_c + N_W$. The number of unique links in both V and W is $N_c + N_V + N_W = N$.

The single master network, G, is assumed to consist of the union of the nodes and links in the two networks, and has n nodes and N links. The probability a given shared node will be randomly and independently sampled twice is $[(n_V+n_c)/n][(n_W+n_c)/n]$. The probability a given shared link will be randomly and independently sampled twice is $[(N_V+N_c)/N][(N_W+N_c)/N]$.

If there are two generating networks I assume they each have n nodes and N links. I also assume they are similar, because we would not be comparing dissimilar networks. The degree of similarity is controlled by 'tuning' parameters[1]: Gv and Gw are assumed to share αn nodes and βN links. The probability a given shared node will be sampled twice is then $\alpha[(n_V+n_c)/n][(n_W+n_c)/n]$, and the probability a given shared link will be sampled twice is $\beta[(N_V+N_c)/N][(N_W+N_c)/N]$. The likelihood ratio λ_{js} for G vs (GV, GW) as generator of a given shared node is then $1/\alpha$ and the likelihood ratio π_{js} of a given shared link is then $1/\beta$.

For a non-shared node, say in V, similar reasoning gives a likelihood ratio λ_{kV} of

$[1-(n_W+n_c)/n)] /[1- \alpha(n_W+n_c)/n]$

and for a non-shared link a likelihood ratio π_{kV} of

$[1-(N_W+N_c)/n)] /[1- \alpha(N_W+N_c)/N]$

For a non-shared node or link in W substitute a W subscript for the V subscript in these likelihood ratios.

Computational efficiency is a necessity if the similarity metric is to be applied to large networks. For this reason I do not calculate the exact probabilities for the numbers of shared and non-shared nodes and links that are observed (the combinatoric complexity of such calculations is enormous). Instead I make the simplifying assumption that each node and link contribute the likelihood ratios given above and that the total odds is obtained by multiplying all the likelihood ratios together. This simplification can perhaps be justified if similar distortion is produced by this simplifying assumption for both the cases of G and (G_V,G_W) as generators. Under this simplifying assumption the overall odds becomes:

$$\varphi(1/2) = (\lambda_{js})^{nc}(\lambda_{kV})^{nV}(\lambda_{jW})^{nW}(\pi_{js})^{Nc}(\pi_{kV})^{NV}(\pi_{jW})^{NW}$$

Taking the log of this product converts the calculation to sums and makes calculation highly efficient.

This abstract is too short to permit giving the different and more complex results that hold for the several cases when the nodes and/or links have associated strengths. I give a summary of some of the results here. The results for links and nodes are similar so consider the results for nodes. Let there be just one set of strength values, Si for the i-th node. Norm these to sum to 1.0. For either generation by G or (Gv,Gw) assume sampling is made without replacement and proportional to strength. Let Ziv and Ziw be the probabilities that node i will be sampled by n_v+n_c samples, or n_w+n_c samples respectively. The Z's would be difficult to

obtain analytically but could be estimated by Monte Carlo sampling. Consider two possibilities for the way that Gv and Gw overlap. In Case A the probability a node will be shared is simply α, independent of strength. In Case B, the probability a node will be shared is an increasing function of strength, Y_i.

For Case A the likelihood ratio for a shared node i is: $1/\alpha$. For a node k only in V the likelihood ratio is: $\lambda_{kV} = (1\text{-}Zkw)/\{1 - \alpha\ (1\text{-}Zkw)\}$. For a node only in W exchange the subscripts v and w. Then we have for the odds due to nodes: $\varphi_D = (1/\alpha)^{nc}\Pi_k(\lambda_{kV})\Pi_j(\lambda_{jW})$.

For Case B the likelihood ratio for a shared node i is $1/Y_i$. For a node k only in V the likelihood ratio is: $\lambda_{kV} = (1\text{-}Zkw)/\{1-Y_k(1\text{-}Zkw)\}$. Again switch v and w subscripts for a node only in W. Then we have for the odds due to nodes: $\varphi_D = \Pi_i(1/Y_i)\Pi_k(\lambda_{kV})\Pi_j(\lambda_{jW})$.

These expressions would have analogous forms for links, with different Ns, Z's and Y's, and the overall odds would, as before, be a product of the odds for nodes and the odds for links.

The critical difference between Cases A and B is the degree to which evidence based on an observed shared node or link is strength dependent: For Case B this evidence rises as strength decreases. This should raise concerns: However strengths are obtained there is likely to be measurement noise that reduces the reliability of low strength values. This might argue in favor of using Case A, or if one preferred Case B to restrict the Yi values to lie above a lower bound. The idea would be to let evidence depend most on the nodes (or links with high strength values.

It should be observed that the existence of a computationally efficient and generally applicable metric for network similarity would allow alignment of non-labeled networks. One would search for the alignment of nodes that would maximize the metric.

I have many relevant publications demonstrating some degree of expertise in Bayesian modeling (e.g.: Shiffrin & Chandramouli, in press; Shiffrin, Chandramouli, & Grünwald, 2015; Chandramouli & Shiffrin, 2015; Nelson & Shiffrin, 2013; Cox & Shiffrin, 2012; Shiffrin, Lee, Kim, & Wagenmakers, 2008; Cohen, Shiffrin, Gold, Ross, & Ross, 2007; Denton & Shiffrin; Huber, Shiffrin, Lyle, & Ruys, 2001; Shiffrin & Steyvers, 1997). I note that the present results are in a vague sense an extension of the metric proposed for matching memory probes to memory traces that are given in Cox and Shiffrin (2012) and in the appendix of Nelsonb and Shiffrin (2013).

References

Asta, D., & Shalizi, C. R. (2014). Geometric Network Comparison. arXiv preprint arXiv:1411.1350.

Busemeyer, J. R., Zheng, W., & Shiffrin, R. M. (2015). Bayesian model comparison favors quantum over standard decision theory account of dynamic inconsistency. *Decision*, 2, 1-12.

Chandramouli, S. H., & Shiffrin, R. M. (2015). Extending Bayesian induction. *Journal of Mathematical Psychology*. http://dx.doi.org/10.1016/j.jmp.2015.10.002

Cohen, A. L., Shiffrin, R. M., Gold, J. M., Ross, D. A., & Ross, M. G. (2007). Inducing features from visual noise. *Journal of Vision*, 7(8):15, 1-14.

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science*, 4(1), 135-150.

Denton, S. E., & Shiffrin, R. M. (2012). Primes and flankers: Source confusions and discounting. *Attention, Perception, & Psychophysics*, 74 (5), 852-856.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, 108, 149-182.

Menezes, T., & Roth, C. (2014). Symbolic regression of generative network models. *Scientific reports*, 4.

Nelson, A. B. and Shiffrin, R.M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, 120(2); 356-394.

Shiffrin, R. M., & Chandramouli, S. H. (in press). Model selection, data distributions, and reproducibility. In H. Atmanspacher and S. Maasen (Eds.) *Reproducibility: Principles, Problems, and Practices*. John Wiley, New York.

Shiffrin, R. M., Chandramouli, S. H., & Grünwald, P. D. (2015). Bayes Factors, relations to Minimum Description Length, and overlapping model classes. *Journal of Mathematical Psychology*. http://dx.doi.org/10.1016/j.jmp.2015.11.002

Shiffrin, R. M., Lee, M. D., Kim, W. & Wagenmakers,, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonomic Bulletin and Review*, 4(2), 145-166.

1. α and β would have to be quite large to produce reasonable similarity values. Research will be needed to determine whether the values can be fixed for all types of networks being compared, or adjusted for different network types.

Cognition without Behaviour: Cognitive Functions in Behaviourally Non-Responsive Individuals

Boris Kotchoubey (boris.kotchoubey@uni-tuebingen.de)
Department of Medical Psychology, University of Tübingen, Silcherstr. 5, 72076 Tübingen, Germany

Keywords: Consciousness; disorders of consciousness; cognitive neuroscience; language comprehension

General

Patients, which are completely non-responsive albeit awake (i.e., not in coma), were first described by the famous German neurologist Ernst Kretschmer (1940) who designate this condition "apallic syndrome" meaning that there is no cortical activity altogether. In his description of the syndrome, Kretschmer used such terms as panapraxia (i.e., the complete lack of all functions of action control) and panagnosia (i.e. the complete lack of all cognitive functions). In the 1970ies, the not less famous Scottish neurosurgeon Bryan Jennett proposed a new name "vegetative state" (VS) for the patients of the kind as depicted by Kretschmer. While the term "apallic" (Greek "without cortex") referred to anatomical and physiological entities, "vegetative" suggests that the patients lack a "subjective world" in form of feelings, thoughts, sensations, etc., while their vital functions such as respiration and circulation remains intact. The latter fact, again, strongly distinguishes this condition from coma.

For the last twenty years, a considerable body of data has been accumulated indicating that many VS patients do possess the abilities to cognitive processing at various levels of complexity. Because of the absence of any goal-directed behaviour, these data were obtained using direct measurement of patients' brain responses by means of the electroencephalography (EEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). The data indicate that VS patients are definitely not "apallic" and that at least some of them are not "vegetative" either.

The proposed presentation critically reviews the empirical evidence of cognitive processing in VS and in two other conditions of behavioural non-responsivity: minimally conscious state (MCS) and locked-in syndrome (LiS). While VS is, by definition, a state of wakefulness without any subjective experience, MCS is characterized by minimal and highly instable signs of subjective awareness, and the LiS is defined as a condition in which consciousness and cognition are largely preserved but cannot be manifested in behaviour due to an extremely severe paralysis of the entire musculature including speech muscles. However, these statements describe the definitions of the respective states; since the rate of misdiagnoses is known to be very high, it is unclear how far these definitions characterize single patients..

Structure

The proposed review is broken down as follows:

1. Evidence of the preservation of main cortico-thalamic circuits necessary for cognitive operations.
2. Evidence of simple sensory processing.
3. Evidence of higher-level cognitive processing.
4. Evidence of language processing.
5. Evidence of subjective awareness.

Principal Conclusions

1 A significant portion of VS and MCS patients demonstrates clear markers of cognitive processing at different levels of complexity, including processing of semantic relations between linguistic objects.

2 Both VS and MCS populations are highly heterogeneous in respect of the number and nature of the obtained indicators of their cognitive functions. The overlap between these populations is much larger than might be expected on the basis of the different clinical diagnoses. At least one large subgroup of VS patients can be described that does not differ from MCS in terms of their cognitive abilities.

3 Semantic processing, sometimes at a high level of complexity, can be found in both VS and MCS patients. Notably, this processing is not necessarily related to conscious language apprehension. The two statements "the patient's brain adequately processes sematic relationships between verbal stimuli" and "the patient understands semantic relationships between words" are not equivalent.

4 Speaking about consciousness, we should distinguish between high-level, language-based cognitive awareness, on the one hand, and low-level, language-independent subjective experience (e.g., experience of pain and pleasure), on the other hand. The former can be found in about 5 to 10% of VS and MCS patients; if found in VS, it unequivocally indicates a misdiagnosis. The latter can characterize a much larger patient group. However, criteria of low-level consciousness are disputable.

5 As expected on the basis of the clinical diagnosis, indicators of higher-level cognitive functions (including high-level consciousness) are obtained in LiS much more frequently than in VS and MCS. However, their manifestation in LiS is inferior as compared with normal population. Thus the idea that LiS is a state, in which cognitive abilities are fully intact and only their overt

expression is impossible due to the severe motor impairment, is not supported by the data.

Acknowledgments

The studies were supported by the German Research Society (Deutsche Forschungsgemeinschaft), the European Commission, and the Alexander von Humboldt Foundation. The author thanks Niels Birbaumer, Andrea Kübler, Simone Lang, Friedemann Müller, Tao Yu, as well as the relatives of the examined patients.

Author's Main Publications to the Topic

Kübler, A., Kotchoubey, B., Hinterberger, T., et al. (1999). The thought translation device: A neurophysiological approach to communication in total motor paralysis. *Experimental Brain Research, 124*, 223-232.

Kotchoubey, B., Lang, S., Baales, R., et al. (2001). Brain potentials in human patients with severe diffuse brain damage. *Neuroscience Letters, 301*, 37-40.

Kotchoubey, B., Lang, S., Bostanov, V., & Birbaumer, N. (2002). Is there a mind? Psychophysiology of unconscious patients. *News in Physiological Sciences, 17*, 38-42.

Kotchoubey, B., Lang, S., Winter, S., & Birbaumer, N. (2003). Cognitive processing in completely paralyzed patients with amyotrophic lateral sclerosis. *European Journal of Neurology, 10*, 551-558.

Neumann, N., & Kotchoubey, B. (2004). Assessment of cognitive functions in severely paralysed and severely brain-damaged patients: Neuropsychological and electrophysiological techniques. *Brain Research Protocols, 14*, 25-36.

Kotchoubey, B., Lang, S., Mezger, G., et al. (2005). Information processing in severe disorders of consciousness: Vegetative state and minimally conscious state. *Clinical Neurophysiology, 116*, 2441-2453.

Hinterberger, T., Wilhelm, B., Mellinger, J., Kotchoubey, B., & Birbaumer, N. (2005). A tool for detection of cognitive brain functions in severely brain injured patients integrated in the thought-translation-device. *IEEE Transactions of Biomedical Engineering, 52*(2), 211-220.

Kotchoubey, B., Jetter, U., Lang, S., et al. (2006). Evidence of cortical learning in vegetative state. *Journal of Neurology, 253*(10), 1374-1376.

Kübler, A., & Kotchoubey, B. (2007). Brain-computer interfaces in the continuum of consciousness. *Current Opinion in Neurology, 20*, 643-649.

Kotchoubey, B. (2007). Editorial: Event-related potentials predict the outcome of the vegetative state. *Clinical Neurophysiology, 118*, 477-479.

Daltrozzo, J., Wioland, N., Mutschler, V., & Kotchoubey, B. (2007). Predicting outcome of coma using event-related brain potentials: A meta-analytic approach. *Clinical Neurophysiology, 118*, 606-614.

Kotchoubey, B. (2009). Vegetative state. In L. Squire (Ed.), *Encyclopedia of Neuroscience* (Vol. 10, pp. 61-66). Amsterdam: Elsevier.

Kotchoubey, B., & Lang, S. (2011). Intuitive versus theory-based assessment of consciousness: The problem of low-level consciousness. *Clinical Neurophysiology, 122*, 430-432.

Kotchoubey, B., Yu, T., Markl, A., et al. (2011a). On the way to the deep layers of consciousness. *Advances in Clinical Neuroscience and Rehabilitation, 11*(4), 10-13.

Kotchoubey, B., Yu, T., Markl, A., et al. (2011b). Differentiation between disorders of consciousness and disorders of movement using functional MRI. *Grand Rounds, 11*, 60-65.

Kotchoubey, B., & Lotze, M. (2013). Instrumental methods in the diagnostics of locked-in syndrome. *Restorative Neurology and Neuroscience, 31*, 25-40.

Kotchoubey, B., Merz, S., Lang, S., et al. (2013). Global functional connectivity reveals highly significant differences between the vegetative and the minimally conscious state. *Journal of Neurology, 260*, 975-983.

Markl, A., Yu, T., Vogel, D., Müller, F., Kotchoubey, B., & Lang, S. (2013). Brain processing of pain in patients with unresponsive wakefulness syndrome. *Brain and Behavior, 2*, doi: 10.1002/brb1003.1110.

Yu, T., Lang, S., Vogel, D., Markl, A., Müller, F., & Kotchoubey, B. (2013). Patients with unresponsive wakefulness syndrome respond to the pain cries of other people. *Neurology, 80*, 345-352.

Kotchoubey, B., Veser, S., Real, R., et al. (2013). Towards a more precise neurophysiological assessment of cognitive functions in patients with disorders of consciousness. *Restorative Neurology and Neuroscience, 31*, 473–485.

Kotchoubey, B., Yu, T., Mueller, F., Vogel, D., Veser, S., & Lang, S. (2014). True or false? Activations of language-related areas in patients with disorders of consciousness. *Current Pharmaceutical Design, 20*, 4239-4249.

Kotchoubey, B., Vogel, D., Lang, S., & Müller, F. (2014). What kind of consciousness is minimal? *Brain Injury, 28*(9), 1156–1163.

Kotchoubey, B., Pavlov, Y. G., & Kleber, B. (2015). Music in research and rehabilitation of disorders of consciousness: Psychological and neurophysiological foundations. *Frontiers in Psychology, 6*, Article 1783.

Kotchoubey, B. (2015). Event-related potentials in disorders of consciousness. In A. O. Rossetti & S. Laureys (Eds.), *Clinical Neurophysiology of Disorders of Consciousness* (pp. 107-124). Vienna - Heidelberg - New York - Dordrecht - London: Springer.

Visualizing Events: Simulating Meaning in Language

James Pustejovsky (jamesp@cs.brandeis.edu)
Department of Computer Science, Brandeis University 415 South Street
Waltham, MA 02453 USA

Nikhil Krishnaswamy (nkrishna@brandeis.edu)
Department of Computer Science, Brandeis University 415 South Street
Waltham, MA 02453 USA

Keywords: event semantics; simulation semantics; habitats; affordances; qualia structure; dynamic event models

Background

In this note, we present results from a new approach to modeling the semantics of natural language, *Multimodal Semantic Simulations (MSS)*, being pursued by the first author and his students. This approach assumes both a richer formal model of events and their participants, as well as a modeling language for constructing 3D visualizations of objects and events denoted by natural language expressions. The Dynamic Event Model (DEM) encodes events as programs in a dynamic logic with an operational semantics, while the language VoxML, Visual Object Concept Modeling Language, is being used as the platform for *multimodal semantic simulations* in the context of human-computer communication.

For over two decades, Pustejovsky has been investigating the nature of events in language: the Aktionsarten and event typing associated with lexemes and phrases; how they are mapped to syntax; and how nominal semantics interacts with type coercion behaviors involving events, as modeled in Generative Lexicon. Generative Lexicon has developed a fairly sophisticated model (both data structures and mechanisms) for how the function and use of an object can be encoded as part of its lexical semantics. In addition, he has worked to develop open standards for encoding the temporal and spatial information associated with events. The resulting languages, ISO-TimeML and ISOspace, are ISO standards for their areas, and ISO-TimeML has become the most widely adopted language within the field for representing temporal information in open domains.

Multimodal Semantic Simulation Theory

According to Goldman (2006), simulation provides a process-driven theory of mind and mental attribution, differing from the theory-driven models. From the cognitive linguistics tradition, simulation semantics has come to denote the mental instantiation of an interpretation of any linguistic utterance (Feldman, 2006; Bergen, 2012). See (Markman, Klein, & Suhr, 2012) for a general review of simulation theory in psychology. While these communities do not often reference each other, it is clear from our perspective, that they are pursuing similar programs, where distinct linguistic utterances correspond to generated models that have differentiated structures and behaviors (Narayanan, 1999; Siskind, 2001; Goldman, 2006).

Prior work in visualization from natural language has largely focused on object placement and orientation in static scenes (Coyne & Sproat, 2001; Siskind, 2001; Chang, Monroe, Savva, Potts, & Manning, 2015). In previous work (Pustejovsky & Krishnaswamy, 2014; Pustejovsky, 2013a), we introduced a method for modeling natural language expressions within a 3D simulation environment, Unity. The goal of that work was to evaluate, through explicit visualizations of linguistic input, the semantic presuppositions inherent in the different lexical choices of an utterance. This work led to two additional lines of research: an explicit encoding for how an object is itself situated relative to its environment; and an operational characterization of how an object changes its location or how an agent acts on an object over time. The former has developed into a semantic notion of situational context, called a *habitat* (Pustejovsky, 2013a; McDonald & Pustejovsky, 2014), while the latter is addressed by dynamic interpretations of event structure (Pustejovsky & Moszkowicz, 2011; Pustejovsky, 2013b; Mani & Pustejovsky, 2012; Pustejovsky, 2013a). The requirements on a "multimodal simulation semantics" include, but are not limited to, the following components:

(1) a. A minimal embedding space (MES) for the simulation must be determined. This is the 3D region within which the state is configured or the event unfolds;
b. Object-based attributes for participants in a situation or event need to be specified; e.g., orientation, relative size, default position or pose, etc.;
c. An epistemic condition on the object and event rendering, imposing an implicit point of view (POV);
d. Agent-dependent embodiment; this determines the relative scaling of an agent and its event participants and their surroundings, as it engages in the environment.

In order to construct a robust simulation from linguistic input, an event and its participants must be embedded within an appropriate minimal embedding space. This must sufficiently enclose the event localization, while optionally including room enough for a frame of reference visualization of the event (the viewer's perspective).

Dynamic Event Models

Models of processes and events using updating typically make reference to the notion of a state transition (Harel, Kozen, & Tiuyn, 2000). This is done by distinguishing between formulae, ϕ, and programs, π. A formula is interpreted as a classical propositional expression, with assignment of a truth value in a specific model. We interpret models by reference to specific states, which is a set of propositions with assignments to variables at a specific index. Atomic programs are input/output relations ($[\![\pi]\!] \subseteq S \times S$), and compound programs are constructed from atomic ones following rules of dynamic logic. Events are interpreted as state sequences (sentences), created by the programs (transitions) and tests between them, which can be seen as *dependencies*. An event can be modeled as a dependency graph, where:

(2) a. Events are word sequences, $\langle e_1, e_2, \ldots, e_n \rangle$.
b. We "label" the dependencies (programs) between words as in syntactic dependency structures.
c. We distinguish between the *Object Model* and *Action Model*.
d. An *Event Model* is the composition of these two.

Dependency labels are of three kinds:

(3) a. **Program** (object and action): the program moving between states.
b. **Test**: conditions that must hold during a transition
c. **Content**: the propositional content of a state

VoxML: Visual Object Modeling Language

We have developed a specification for a modeling language for constructing 3D visualizations of natural language expressions, used as the platform for creating multimodal semantic simulations of spatial and event semantics (Pustejovsky & Krishnaswamy, 2016). This modeling language allows for an encoding of how an object is situated relative to its environment, and an operational characterization of how an object changes its location (the object model) or how an agent acts on an object over time (the action model), thus allowing both design and representation of data structures that generate simulations of compositions of these two models into the aforementioned event model. The VoxML specification allows for dependency-parsed linguistic data to be annotated and transformed into a Dynamic Event Model, expressed within dynamic interval temporal logic (DITL), and from there into a procedurally generated visual simulation of the described event (for example, see Figures 1 and 2) .

put(obj,y)
loc(obj) := x, target(obj) := y; b := x;   *(x := w; x ≠ w; d(b,x) < d(b,w), d(b,y) > d(y,w)*$^+$

Figure 1: DEM/DITL expressions for *put(obj,y)*.

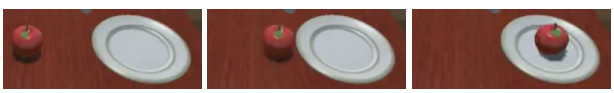


Figure 2: Program is executed

Future Directions

We are currently creating simulations of caused motion, where causation is not needed as a primitive in the logic. Rather, we employ a technique we call *rigging*, where the affected object and the actor are rigged together as one moving object. This is computationally very efficient and has interesting consequences for how to encode event causation.

Relevant Publications

Pustejovsky (2013a), Pustejovsky (2013b), Pustejovsky and Moszkowicz (2011), Pustejovsky and Krishnaswamy (2014), Pustejovsky and Krishnaswamy (2016), Mani and Pustejovsky (2012), McDonald and Pustejovsky (2014).

References

Bergen, B. K. (2012). *Louder than words: The new science of how the mind makes meaning*. Basic Books.

Chang, A., Monroe, W., Savva, M., Potts, C., & Manning, C. D. (2015). Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

Coyne, B., & Sproat, R. (2001). Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 487–496).

Feldman, J. (2006). *From molecule to metaphor: A neural theory of language*. MIT press.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. OUP.

Harel, D., Kozen, D., & Tiuyn, J. (2000). *Dynamic logic* (1st ed.). The MIT Press.

Mani, I., & Pustejovsky, J. (2012). *Interpreting motion: Grounded representations for spatial language*. OUP.

Markman, K. D., Klein, W. M., & Suhr, J. A. (2012). *Handbook of imagination and mental simulation*. Psychology.

McDonald, D., & Pustejovsky, J. (2014). On the representation of inferences and their lexicalization. In *Advances in cognitive systems* (Vol. 3).

Narayanan, S. (1999). Reasoning about actions in narrative understanding. *IJCAI*, *99*, 350–357.

Pustejovsky, J. (2013a). Dynamic event structure and habitat theory. In *Proceedings of the 6th international conference on generative approaches to the lexicon* (pp. 1–10).

Pustejovsky, J. (2013b). Where things happen: On the semantics of event localization. In *Proceedings of isa-9: International workshop on semantic annotation.*

Pustejovsky, J., & Krishnaswamy, N. (2014). Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)*, 99.

Pustejovsky, J., & Krishnaswamy, N. (2016). Voxml: A visual object modeling language. *Proceedings of LREC*.

Pustejovsky, J., & Moszkowicz, J. (2011). The qualitative spatial dynamics of motion. *The Journal of Spatial Cognition and Computation*.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.(JAIR)*, *15*, 31–90.

An Architectural Account of Variation in Problem Solving and Execution

Pat Langley (patrick.w.langley@gmail.com)
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA

Background and Motivation

Theories of the cognitive architecture (Langley et al., 2009) aim to specify the constant features of the human mind. They make assumptions about the representation and organization of memory, as well as about performance and learning mechanisms that operate over its structures. They specifically do not include content that changes, whether rapidly through reasoning and problem solving or gradually through learning.

However, we also know that humans are highly adaptive and they can work on a given task in many different ways. This creates a tension between the desire to identify universals in human cognition and to explain observed variation. Some architectures, like Soar (Laird et al., 1987) and Epic (Kieras & Meyer, 1997), make few commitments except for knowledge representation and the basic cognitive cycle. In contrast, ICARUS (Langley et al., 2009) makes more specific claims about structures and mechanisms, committing to means-ends analysis for problem solving and to reactive control for execution.

In recent research, we have sought a middle ground, devising a new cognitive architecture that incorporates stronger assumptions than Soar and Epic, but weaker ones than ICARUS. In this paper, we present the phenomena it attempts to cover, its theoretical postulates, and planned extensions to the framework. We will see that the theory builds on classic ideas but combines them in novel ways that move beyond its predecessors.

Variation in Problem Solving

Humans exhibit the ability to solve novel problems, with an important case being generation of multi-step plans that achieve goals. The standard theory of problem solving, due to Newell and Simon (1972), states this involves search through a problem space that operates over states, goals, and operators encoded as symbol structures. Heuristics or rules of thumb guide this search process and make it tractable even with limited resources.

Despite these high-level regularities, people also show considerable variety in problem-solving behavior. For instance, they use means-ends analysis on puzzles like the Tower of Hanoi (Newell & Simon, 1972), but rely on forward search in games like chess (de Groot, 1978). Domain expertise accounts for some of these differences, but others appear due to generic strategies that determine search direction, operator selection, and other matters. Our research has led to an expanded theory of human problem solving that comprises five postulates:

- Plans are represented by hierarchical decomposition (AND) trees, in which each node is a problem and every child denotes a subproblem of its parent.
- Problem solving involves the recursive decomposition of problems into subproblems, with alternative candidates organized into a search (OR) tree.
- Problem solving operates in a cognitive cycle of five stages: problem selection, intention generation, subproblem creation, and failure/success checking.
- Strategic knowledge – encoded as domain-independent control elements – governs decisions at each stage to produce different problem-solving strategies.
- Domain expertise takes the form of generalized decompositions that specify how to break a problem into subproblems and that serve as higher-level operators.

We have incorporated these tenets into FPS, a problem solver that implements the theory (Langley et al., 2014). We have tested the system on multiple domains, including puzzles like the Tower of Hanoi and planning tasks involving logistics. We have demonstrated that FPS can solve problems in these domains, that its strategic knowledge reproduces familiar behaviors like means-ends analysis and forward search, and that domain knowledge reduces search and makes problem solving tractable.

Variation in Plan Execution

Humans also exhibit the ability to carry out complex sequential activities over time to achieve their goals (Miller et al., 1960). There are different accounts of such extended behavior, but they generally agree on major features. This topic has been studied mainly in the context of skilled behavior, but similar accounts apply to the execution of plans generated during problem solving, which we argue involve two distinct but linked processes.

Again, although human execution of skills and plans follows some high-level regularities, there remains considerable variation. The field of motor behavior saw a long debate about whether people utilize closed-loop (Adams, 1971) or open-loop (Schmidt, 1982) control, yet it is now clear this differs across individuals and situations. Our theory of plan execution includes four claims:

- Plans and skills are stored as hierarchical decomposition trees like those produced during problem solving.
- Plan execution operates by traversing these decomposition trees from top to bottom and from left to right, with physical actions occurring at terminal nodes.

- This process involves a cognitive cycle with five stages: intention selection, condition checking, intention enaction, perceptual inspection, and effects checking.
- Strategic knowledge – stated as domain-independent control elements – governs decisions at each stage of the cycle to produce different execution strategies.

We have incorporated these ideas into FPE, a flexible plan executor that embodies the theory. We have tested this module on the same domains as FPS, demonstrating that it can carry out hierarchical plans in simulated environments and that it supports a variety of execution strategies, such as open-loop and closed-loop control.

Interleaving Execution with Planning

Just as people exhibit different strategies for plan generation and plan execution, so can they interleave these two processes in distinct ways. This interaction has been less studied than either problem solving or execution in isolation, but some things are clear. In certain settings, a person can generate a complete plan before executing it in an open-loop manner. In others, the problem is so complex, as in difficult puzzles, or the environment is sufficiently unpredictable, as in playing chess, that they must alternate between extending and executing a plan.

Our theory of how humans interleave planning with execution adds two postulates to our earlier accounts:

- Domain-independent strategic control elements determine whether the system proceeds to the next stage of processing or transfers control to the other module.
- The primary loci of control reside in the fourth stage of problem solving – success checking – and the second stage of execution – condition checking.

We have extended FPS and FPE to incorporate these assumptions, letting them operate jointly as parts of an integrated architecture. This uses the problem solver to generate a plan, then shifts to execution and returns to problem solving when monitoring reveals the need.

Exactly when the shifts occur depends on the control elements. We have used different settings to reproduce execution only after forming a top-level plan, alternation between looking N steps ahead and taking a single action (as in game-playing systems), and execution upon solving each subproblem (as in ICARUS). Moreover, computational studies have shown that strategy effectiveness interacts with domain characteristics (Bai et al., 1015).

Strategy Variation and Adaptation

These accounts of variation in problem solving, plan execution, and interleaving make a clear contribution to our understanding of high-level cognition. However, they remain incomplete in that they do not clarify the conditions under which humans utilize different strategies. In ongoing work, we are extending the theory and its implementation to address this challenge.

The central idea is that strategy variation is the result of mental adaptation. The human cognitive architecture has access to each strategy handled by our theory, and it can switch among them as appropriate. This functionality requires access to meta-level information about the state of problem solving, such as branching factors in the forward and backward directions, and execution, such as the reliability of actions. This means that the strategic control elements must include conditions that match against such meta-level data.

We are curently extending the FPS and FPE modules to support such adaptive behavior. We intend to demonstrate that the revised architecture decides for itself which strategies to use during problem solving, plan execution, and interleaving of these activities. The result will be a more complete account of high-level cognition that is consistent with earlier theories but combines their ideas in novel ways to explain important phenomena.

Acknowledgements

This research was supported in part by Grant N00014-15-1-2517 from the Office of Naval Research. C. Pearce, Y. Bai, C. Worsfold, and M. Barley at the University of Auckland contributed to the results reported here.

References

Adams, J. A. (1971). A closed-loop theory of motor learning. *Journal of Motor Behavior*, *3*, 111–149.

Bai, Y., Pearce, C., Langley, P., Barley, M., & Worsfold, C. (2015). An architecture for flexibly interleaving planning and execution. *Poster Collection: The Third Annual Conference on Advances in Cognitive Systems*.

de Groot, A. D. (1978). *Thought and choice in chess* (2nd Ed.). The Hague: Mouton Publishers.

Kieras, D., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, *12*, 391–438.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Langley, P., Choi, D., & Rogers, S. (2009). Acquisition of hierarchical reactive skills in a unified cognitive architecture. *Cognitive Systems Research*, *10*, 316–332.

Langley, P., Pearce, C., Barley, M., & Emery, M. (2014). Bounded rationality in problem solving: Guiding search with domain-independent heuristics. *Mind and Society*, *13*, 83–95.

Miller, G. A., Galanter, E., & Pribram, K. A. (1960). *Plans and the structure of behavior*. New York: Holt, Rhinehart, & Winston.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Schmidt, R. A. (1982). More on motor programs. In J. A. S. Kelso (Ed.), *Human motor behavior: An introduction*. Hillsdale, NJ: Lawrence Erlbaum.

Verbalizing navigation: Explicit and implicit concepts

Thora Tenbrink (t.tenbrink@bangor.ac.uk)
School of Linguistics and English Language, College Road, Bangor University (Wales)
Bangor LL57 2DG, UK

Keywords: Cognitive Discourse Analysis; spatial cognition; verbalization; cognitive processes; reference frames

Introduction

Every day, we navigate our environments with astonishing ease. Most of our paths are familiar to us and can be navigated without (much) conscious thought; in other cases, we use various strategies to find our way (Tenbrink & Wiener, 2007). Since these processes are at the heart of human spatial cognition they have been researched extensively, often based on route directions as the most common verbalizations of navigation. Our research extends this tradition across various wayfinding contexts, addressing street network scenarios (Hölscher, Tenbrink, & Wiener, 2011), complex buildings (Tenbrink, Bergmann, & Konieczny, 2011), alpine environments (Egorova, Tenbrink, & Purves, 2015), and including effects of automatic systems as producers (Tenbrink & Winter, 2009) or recipients (Moratz & Tenbrink, 2006; Tenbrink et al., 2010) of spatial directions. In all of these studies natural language data are used to address concepts of navigation, some of which are expressed explicitly, while others remain implicit and only indirectly reflected through the ways in which speakers use language in spatial navigation contexts.

Navigation and language

Whenever we use language, we reveal something about our thoughts and concepts – in part intentionally and explicitly, and in part without being aware of it, reflected in the use of language in a particular way. In the context of navigation, even a common route segment description like *Go straight on the main road until you cross the traffic lights* reflects a wealth of concepts. For instance, the recipient is expected to infer relevant movement directions from *straight on* (e.g., what is the initial orientation from which to move straight on? Is it still 'straight on' if there's a bend in the road? Etc.) and *cross* (in which direction should the road be crossed?), and to identify suitable referents for *the main road* and *the traffic lights*, drawing on world knowledge and current perception of the environment. The two definite articles reveal the speaker's consideration of these references as unambiguous, clearly without conceptualizing any competing main roads or traffic lights. The term *until* highlights the concept of an end point of the segment in question, marked by the traffic lights as a recognizable landmark in the environment.

Route givers do not necessarily consciously consider all of these concepts, or their possible alternatives; they use language as it comes to mind, spontaneously and naturally. Nevertheless, features such as these are frequent and systematic in route descriptions. Patterns can be (and have been) identified by close cognitive-linguistic data analysis (e.g., Allen, 2000; Denis, 1997; and others), enhanced by theories and findings from a wealth of research in linguistics, psychology, and more.

Building on such findings and related studies across domains, Cognitive Discourse Analysis (CODA, Tenbrink, 2015) was developed to support the systematic analysis of language reflecting various levels of thought. Using CODA, Hölscher et al. (2011) identified explicitly verbalized strategies on route choices along with linguistic indicators that highlighted how route givers accounted for the needs of the wayfinder (see also Tenbrink & Winter, 2009). Crucially, our results highlighted the strategy of heading into a particular direction while navigating, as opposed to orienting towards the street network while planning ahead and explaining the route in advance.

Other studies examining the verbalization of spatial navigation concepts include Klippel et al. (2013), who showed how speakers intuitively conceptualize intersections differently depending on how they are used for navigation, focusing on either structure or function. Here, different linguistic expressions systematically reflect the underlying relevance of an intersection concept. Mast et al. (2014) recognized patterns of categorization in speakers' directional concepts that were reflected through grouping behavior as well as verbalization. They identified two principles of categorizing directions such as *left* and *right*, one based on a prototype structure and the other on discrete boundaries. Both exist in parallel in cognition and in language, with distinct linguistic expressions associated with different types of categories. Tenbrink & Seifert (2011) identified explicit and purposefully formulated strategies for tour planning along with implicitly represented conceptual switches between maps and real environments. Together, these studies demonstrate that language is a powerful representation tool that reflects different levels of cognition, accessible through systematic analysis of linguistic patterns.

Navigation in virtual space

Our recent research in virtual space (Tenbrink & Salwiczek, in press) addressed a fundamental distinction frequently encountered in navigation contexts – orientation on the basis of an *allocentric* vs. an *egocentric* reference frame. We reasoned that if verbal data reliably convey navigation concepts at several levels of awareness (as in the studies cited above), this should also be true for something as pervasive and basic as reference frames. To address this we used a task that has, in various variations, frequently been

used to differentiate between humans' preferred reference systems (e.g., Goeke et al., 2015; Gramann et al., 2005).

Participants were presented with a simple, perceptually sparse virtual tunnel shown on a computer screen (similar to common virtual reality scenarios in research and gaming contexts), and were asked to indicate the direction of the tunnel's entrance after the journey ended. They were asked to *think aloud* while doing this task, and also produced retrospective reports after a set of tunnel journeys. Mirroring previous results, directional responses in our study fell into two major categories. One of these had in previous studies been interpreted as an allocentric conceptualization of the journey, and the other as an egocentric one. Since these two reference frames generally have clear reflections in language across task settings, our analysis addressed the extent to which speakers conceptualized and verbalized these basic reference frames in this scenario.

Surprisingly, the expected correspondence between verbalizations and behavioral response patterns could not be verified in the data. In fact, the verbalizations only corresponded to a limited extent to *any* kinds of previously reported reference systems in the literature. Instead, they indicated a multitude of conceptual strategies and concepts emerging in this simple navigation scenario, reflecting the participants' struggles to maintain orientation in this perceptually sparse scenario. These results challenge previous assumptions that the two kinds of behavioral responses to the navigation task consistently and reliably reflect basic allocentric and egocentric reference systems.

Spatial cognition, trivially, depends on *space* – which (non-immersive) virtual reality can only simulate. While visual perception may be similar to real navigation, other sensory input (most prominently proprioception) is lacking – and this can pose major complications for orientation in virtual space. Sparse orientation task settings may thus not be suitable for assessing reference frame preferences that are normally at work in far richer real world contexts.

Acknowledgments

Funding by the DFG for I5-[DiaSpace] and I6-[NavTalk] as part of the SFB/TR8 Spatial Cognition, Bremen/Freiburg, and by the Volkswagen Foundation for a Tandem Project (with J. Wiener, 2007-2009) is gratefully acknowledged.

References

Allen, G.L. (2000). Principles and practices for communicating route knowledge. *Applied Cognitive Psychology,* 14, 333-359.

Denis, M. (1997). The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de Psychologie Cognitive, 16*(4):409-458.

Egorova, E., Tenbrink, T., & Purves, R. (2015). Where snow is a landmark: Route direction elements in alpine contexts. In S.I. Fabrikant, M. Raubal, M. Bertolotto, C. Davies, S. Freundschuh, & S. Bell (Eds.), *Spatial Information Theory*. Berlin: Springer.

Goeke, C., Kornpetpanee, S., Köster, M., Fernández-Revelles, A.B., Gramann, K., & König, P. (2015). Cultural background shapes spatial reference frame proclivity. *Scientific Reports 5, 11426*.

Gramann, K., Müller, H.J., Eick, E.M., & Schönebeck, B. (2005). Evidence of separable spatial representations in a virtual navigation task. *Journal of Experimental Psychology: Human Perception and Performance* 31:6, 1199-1223.

Hölscher, C., Tenbrink, T., & Wiener, J. (2011). Would you follow your own route description? *Cognition 121*, 228-247.

Klippel, A., Tenbrink, T., & Montello, D. (2013). The role of structure and function in the conceptualization of directions. In M. Dimitrova-Vulchanova & E. van der Zee (Eds.), *Motion Encoding in Spatial Language*. Oxford: Oxford University Press.

Mast, V., Wolter, D., Klippel, A., Wallgrün, J.O., & Tenbrink, T. (2014). Boundaries and prototypes in categorizing direction. In C. Freksa, B. Nebel, M. Hegarty, & T. Barkowsky (Eds.), *Spatial Cognition 2014*. Berlin: Springer.

Moratz, R. & Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction. *Spatial Cognition and Computation 6*:1, 63-106.

Tenbrink, T. (2015). Cognitive Discourse Analysis: Accessing cognitive representations and processes through language data. *Language and Cognition 7*:1, 98-137.

Tenbrink, T., Bergmann, E., & Konieczny, L. (2011). Wayfinding and description strategies in an unfamiliar complex building. *Proceedings of the 33th annual conference of the cognitive science society* (pp. 1262-1267). Austin, TX: Cognitive Science Society.

Tenbrink, T., Ross, R.J., Thomas, K.E., Dethlefs, N., & Andonova, E. (2010). Route instructions in map-based human-human and human-computer dialogue: a comparative analysis. *Journal of Visual Languages and Computing 21*:5, 292–309.

Tenbrink, T. & Salwiczek, L. (in press). Orientation and metacognition in virtual space. *Journal of Experimental Psychology: Human Perception and Performance*. Doi:10.1037/xhp0000178

Tenbrink, T. & Seifert, I. (2011). Conceptual layers and strategies in tour planning. *Cognitive Processing 12*:1, 109–125.

Tenbrink, T. & Wiener, J. (2007). Wayfinding strategies in behavior and language: A symmetric and interdisciplinary approach to cognitive processes. In T. Barkowsky, M. Knauff, G. Ligozat, & D. Montello (Eds.), *Spatial Cognition V*. Berlin: Springer.

Tenbrink, T. & Wiener, J. (2009). The verbalization of multiple strategies in a variant of the traveling salesperson problem. *Cognitive Processing 10:*2, 143-161.

Tenbrink, T. & Winter, S. (2009). Variable granularity in route directions. *Spatial Cognition & Computation 9*, 64-93.

The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color

Emily Cibelli[1] (emily.cibelli@northwestern.edu)
Yang Xu[2,3] (yang_xu_ch@berkeley.edu)
Joseph L. Austerweil[4] (joseph_austerweil@brown.edu)
Thomas L. Griffiths[3,5] (tom_griffiths@berkeley.edu)
Terry Regier[2,3] (terry.regier@berkeley.edu)

[1] Department of Linguistics, Northwestern University, Evanston, IL 60208 USA
[2] Department of Linguistics, University of California, Berkeley, CA 94720 USA
[3] Cognitive Science Program, University of California, Berkeley, CA 94720 USA
[4] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912 USA
[5] Department of Psychology, University of California, Berkeley, CA 94720 USA

Keywords: Sapir-Whorf hypothesis; linguistic relativity; color memory; category effects; probabilistic inference.

The Sapir-Whorf hypothesis holds that our thoughts are shaped by our native language, and that speakers of different languages therefore think differently. This hypothesis is controversial in part because it appears to deny the possibility of a universal groundwork for human cognition, and in part because some findings held to support it have not reliably replicated. We argue that considering this hypothesis through the lens of probabilistic inference has the potential to resolve both issues, at least with respect to certain prominent findings concerning color cognition.

We cast color memory as inference under uncertainty, and explore this idea using a version of the category adjustment model of Huttenlocher et al. (1991) (see also Bae et al., 2015; Persaud & Hemmer, 2014). The model holds that color memory involves the probabilistic combination of evidence from two sources: a fine-grained representation of the specific color seen, and the language-specific category in which it fell (e.g. *green*). Both sources of evidence are represented in a universal perceptual color space, yet their combination yields language-specific bias patterns in memory. The model predicts that such category effects will be strongest when perceptual information is uncertain. It thus has the potential to explain the mixed pattern of replications of Whorfian effects in the literature (e.g. Brown et al., 2011; Wright et al., 2015): non-replications could result from high perceptual certainty.

We first describe the results of an experiment showing language-consistent biases in color memory in English speakers, and stronger bias when perceptual information is uncertain, consistent with this model, and replicating recent findings (Bae et al., 2015). We then show that the same model also accounts for existing data on cross-language differences in color memory in speakers of English, Berinmo, and Himba (Roberson et al., 2000; 2005). Finally, we show that this model accounts for existing cross-language data on within-category discrimination (Hanley & Roberson, 2011). We suggest that these ideas may help to clarify the debate over the Sapir-Whorf hypothesis.

Acknowledgments

We thank Paul Kay and Charles Kemp for their comments. This research was supported by NSF under grants DGE-1106400 (EC) and SBE-1041707 (YX, TR).

References

Bae, G.Y., Olkkonen, M., Allred, S.R., & Flombaum, J.I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144*, 744-763.

Brown, A.M., Lindsey, D.T., Guckes, K.M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision, 11*, 2.

Hanley, J.R. & Roberson, D. (2011). Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin and Review, 18*, 355-363.

Huttenlocher, J., Hedges, L.V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review, 98*, 352-376.

Persaud, K. & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 1162-1167.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General, 129*, 369-398.

Roberson, D., Davidoff, J., Davies, I.R.L., & Shapiro, L.R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*, 378-411.

Wright, O., Davies, I.R.L., & Franklin, A. (2015). Whorfian effects on colour memory are not reliable. *The Quarterly Journal of Experimental Psychology, 68*, 745-758.

Do gestures serve an interpersonal function?

Yiwei Chen
University of Alberta

Robyn Enns
University of Alberta

Elena Nicoladis
University of Alberta

Abstract: Some researchers argue that gestures serve an interpersonal function, such as making the intended message clear (e.g., Gallagher & Frith, 2003; cf. Kita, 2000). In this study, we tested whether gestures serve an interpersonal function, specifically predicting that the higher participants' autism spectrum quotient, the less frequently they would gesture. Participants completed the Autism Spectrum Quotient questionnaire (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001). To elicit gestures, participants did two tasks. In one, they explained spatial and social concepts. In another, they told the story of a cartoon. The dependent variable is the gesture rate (gestures per word), to account for individual differences in volubility. Participants completed a standardized vocabulary test. The initial results show no correlation between gesture rate in either task and ASQ scores. There is a negative correlation between ASQ and vocabulary scores. These results are inconsistent with the argument that gestures serve an important interpersonal function.

Emergence of Euclidean geometrical intuitions in hierarchical generative models

Arianna Yuan
Stanford University

Te-Lin Wu
Stanford University

James L. McClelland
Stanford University

Abstract: In this study, we aim to understand the origins of human intuitions about Euclidean geometry by simulating geometric concepts acquisition with unsupervised learning in hierarchical generative models. Specifically, we build a deep neural network that learns a hierarchical generative model of sensory inputs. The results show that hidden layer activities can support the categorization of different geometric objects and distinguish among various spatial relationships between geometric figures. Specifically, hidden layer activities can be decoded to compare line orientations, detect right triangles, and judge whether two triangles are similar or not. We further analyze the response profiles of hidden layers units and find some units resembling parietal neurons in the brain. Using unsupervised deep learning, the current modeling work provides a possible explanation of how Euclidean geometrical intuitions might emerge from daily visual experience, which has significant implications for cognitive psychology and computational neuroscience.

Implicit Emotional Priming of Traumatic Events: The Effects of Semantic Level and Emotional Activation

Jae-Ho Lee
Dept. of Psychology, Keimyung University

Yun-Kyeung Choi
Dept. of Psychology, Keimyung University

Abstract: Vivid representations are often made by traumatic events with intense emotions. The emotions may be activated automatically from memory on the mere exposure of an affect-loaded stimulus. The aims of this study were to investigate the implicit emotional processing of traumatic events and the moderation of priming by semantic level of the events, using primed naming task at short stimulus onset asynchrony (150ms). A 3 semantic level of traumatic primed events (general, domestic, or foreign words) by 3 target emotions (high-arousal negative, moderate positive, low-arousal negative words) repeated design was used. When the primed words were general (e.g. terror) or domestic (e.g. Sewol ferry disaster) events, response time of high-arousal negative words (e.g. fear or angry) were significantly longer than other emotion words (e.g. happy or sadness). Our findings suggest contrast effects of affective priming as a result of automatic implicit regulation.

Current Research in Cognitive Science at Educational Testing Service (ETS)

Irvin Katz
Educational Testing Service

Abstract: The Cognitive Science Research group at ETS conducts research and development at the forefront of educational assessment, using cognitive theory in the design of assessments, building cognitive models to guide interpretation of test-takers' performance, and researching cognitive issues in the context of assessment. Moving beyond traditional (e.g., multiple-choice) tests, the group explores assessments that use innovative, highly interactive digital environments such as online games, virtual labs or other simulations, and human-agent conversation-based interactions. Researchers also investigate how to draw appropriate inferences about test-takers' knowledge and skills from complex data sources such as eye-movements, interaction logs, and other sequential information. I will provide an overview of the group's research, including the use of cognitive models to interpret test-takers' actions within interactive assessment tasks and empirical studies on how test-takers externalize their knowledge when problem solving in domains such as science inquiry, inter-cultural competence, and mathematics argumentation.

Free-form response vs. yes/no-question methodologies in the study of human reasoning

Salvador Mascarenhas
University of Oxford, St Catherine's College, Oxford, United Kingdom

Philipp Koralus
University of Oxford, St Catherine's College

Abstract: There are two widespread strategies for testing experimentally whether a conclusion follows naively from a sequence of premises. The free-form response strategy (FFR) presents participants with the premises and asks them "what, if anything, follows?" In the simplest case, participants' responses are coded as to whether they made the predicted inference. On the yes/no question strategy (YNQ), after presenting the premises, the researcher puts forth a sentence C and asks whether C follows from the premises.

We compare the two methodologies with respect to six types of fallacious problems involving propositional connectives from the mental-models literature, to address the question of whether the methodologies are equally valid. We found that the two methodologies overwhelmingly yield identical results. Interestingly, the exceptions we found show that in some cases FFR fails to detect an attractive fallacious conclusion that can be reliably probed with YNQ.

Noticing causal properties of objects from sequence statistics

Anna Leshinskaya
University of Pennsylvania, Philadelphia, PA, USA

Sharon L Thompson-Schill
University of Pennsylvania, Philadelphia, PA, USA

Abstract: How do we learn non-physical properties of physical objects? We explored how the statistical structure of events can be a source of object property learning. Twenty-five participants saw sequences of visual events surrounding two distinct objects. Object identity determined 1) the direction of a high transition probability between two events, and 2) the frequency of two other events. Learning was unsupervised and unguided. However, participants spontaneously noticed these regularities. In an explicit, verbal judgment task, they discriminated between frequent vs. rare events ($t(24) = 10.7$, $p < 0.00001$) and between predictive vs. non-predictive event pairs ($t(24) = 3.04$, $p<0.01$), as appropriate to the object. These statistics gave rise to distinct conceptual interpretations: participants ascribed a causal interpretation to the predictive statistics ($t(24) = 1.91$, $p<0.05$) more than to events frequently co-occurring with the objects ($t(24) = 3.00$, $p<0.01$). Such learning may underlie concept acquisition, particularly of functional kinds like artifacts.

Filling in the gaps: Event segmentation is robust to missing information

Jessica Kosie
University of Oregon, Eugene, OR, USA

Dare Baldwin
University of Oregon, Eugene, Oregon, United States

Abstract: Fluent event processing involves segmenting streaming sensory information into discrete units. Adults and children selectively attend to these meaningful moments within event streams, which predicts later memory. In natural environments, however, uninterrupted attention is unlikely. Consequently, some information is missed, including event boundary information. To what extent does missing information alter the attentional dynamics of processing, specifically viewers' ability to target remaining boundaries with enhanced attention? Adults advanced at their own pace through slideshows of unfolding activity. Slides were systematically deleted to enable comparison of viewers' attentional dynamics when specific content was present versus absent. Average dwelling per slide increased with missing content. However, the attentional dynamics of processing were unaltered; attention to boundaries displayed comparable enhancement regardless of missing content. Attention modulation during processing of relatively familiar events appears to be highly robust to missing information. What occurs with more novel events is an interesting question for future research.

Effects of the Plausibility of the Retracting Information on Memory Updating

Kyung Soo Do
Sungkyunkwan University

Jin Ju Gu
Sungkyunkwan University

Kwanghyeon Yoo
Sungkyunkwan University

Abstract: The plausability of the retracting information, which can affect whether the retracted information be replaced by the retracting information, had not been studied much in previous studies on continued influence effect. An experiment was conducted to investigate how the plausability of a retracting information affects the understanding of an event in which some of the information had been retracted. In the experiment, the plausibility of the retracting information (low versus high) and the manner of correcting information (simple retraction versus supplying an alternative) were manipulated, and the participants were asked to answer memory questions and inference questions. Retracted information was better remembered for memory questions, but was used less frequently for inference questions when the retracting information was less plausible compared to when the retracting information was quite plausible. The results were interpreted to support the 'memory-based' theory of memory updating.

Which Statistic Matters? Effects of Category Size and Distribution on Statistical Category Learning

Chi-hsin Chen
Indiana University

Paulo Carvalho
Indiana University

Chen Yu
Indiana University

Abstract: The present study investigates whether, and if so in what way, adult learners are sensitive to the properties of the statistical input, such as frequency and skewedness, when learning and generalizing category labels. Participants were presented with novel objects belonging to four different categories and heard category labels in a cross-situational learning task. The four categories were matched for the total amount of exposure but varied in category size and shape of distribution. Participants learned object-to-label mappings better for categories with a skewed distribution of fewer objects. Moreover, object-to-label mapping performance was positively related to the ability to extend category knowledge to novel items. Co-occurrence frequency or category size alone were not good predictors of label learning and generalization. The results indicate the importance of input distribution in word and category learning processes.

Reciprocal altruism in preschool-aged children

Nadia Chernyak
Boston University

Yarrow Dunham
Yale University

Peter Blake
Boston University

Abstract: Young children are remarkably prosocial, yet the mechanisms driving their prosociality are poorly understood. In two studies, we looked at whether a need for reciprocity drives children's prosocial behavior. In Experiment 1, children were given a puzzle task to complete in which they were either missing 2 pieces (experimental group) or not (control group). All children then received 2 puzzle pieces from a confederate, resulting in either necessary instrumental help (experimental group) or unnecessary help (control group). Children were more prosocial (shared a greater proportion of their resources) with the confederate after receiving instrumental help than after receiving unnecessary help. In Experiment 2, we investigated the types of principles children use when paying back help. We found that children employed a mix of exact reciprocation and "needs-based" help when paying back individuals. Our results suggest an important role of gratitude and reciprocity in the development of early prosociality.

Heuristics in exploration: Distributional information is selectively used for active learning

Elizabeth Lapidow
Rutgers University - Newark

Elizabeth Bonawitz
Rutgers University - Newark

Abstract: Everyday decision-making is filled with choices about what to act on, with outcomes playing a critical role in learning. Information gain is oft cited as a valuable approach to maximize potential learning, but its computation is costly. It entails evaluating the probability of multiple outcomes given any possible action, and then considering the degree of belief-change over all possibilities. Given the computational complexity of this evaluation, it becomes important to ask whether learners employ cues to information gain; are there heuristics that drive choice in active learning? Our experiments ask participants to choose between two options (varying in distributional characteristics) in either a "learning-condition" or "collecting-condition". Our results suggest that adults are sensitive to cues (e.g. variance) that tend to correlate with information gain. These cues are only favored in learning-goal contexts, suggesting that certain distributional qualities are not always appealing, but rather are selectively-employed heuristics towards information gain.

On the psychological reality of linguistic event structures

Jayden Ziegler
Harvard University

Jesse Snedeker
Harvard University

Abstract: How language represents meaning remains a central topic of debate in linguistics. On some accounts, the noun phrases in a sentence are identified semantically by a list of independent atomic labels (thematic roles), ordered relative to one another depending on the position of the nouns around the verb (e.g., AGENT-THEME-GOAL). Others instead capture such interdependencies with complex, non-atomic event structures (e.g., [x CAUSE [y TO-COME-TO-BE-AT z]]). Here, we use structural priming to investigate the psychological reality of these two theories of semantic representation. On the thematic role approach, we should expect to see priming between theme-first locatives and prepositional-object datives (both VP-NP-PP syntactically) precisely because their thematic ordering is consistent across the two constructions. The event structure approach posits no such minimal semantic structural similarity, such that we should not see priming cross-constructionally. We find only within-construction priming (N=52) and not across-construction priming (total N=344), in favor of event structures.

Effects of Analogical Processing: Evidence for Re-representation

Daniel Silliman
Binghamton University, Binghamton, NY, United States

Kenneth Kurtz
Binghamton University, Binghamton, NY, United States

Abstract: Re-representation is a mechanism for aligning non-identical structurally corresponding predicates during comparison. Re-representation is therefore a shift in the encoding of a stimulus from an initial set of elements to an altered set. For example, when comparing two analogous sentences, different verbs could be re-represented to allow a match – thereby altering construal. Re-representation is theoretically and intuitively compelling, but difficult to demonstrate. We had participants compare pairs of short text passages and judge them as potential analogies. At test, participants read an altered version of a prior passage and were asked to detect changes from the original. The alterations increased the semantic match between the passage and its analog. Participants who had compared the original passage to the analog were less likely to detect changes than those who compared to a non-analogous passage. As would be expected due to re-representation, analogical comparison made participants less sensitive to the changes.

A computational theory of temporal inference

Sangeet Khemlani
Naval Research Laboratory

Anthony Harrison
Naval Research Laboratory

Greg Trafton
Naval Research Laboratory

Abstract: We describe a novel model-based theory of how individuals reason deductively about temporal relations. It posits that temporal assertions refer to mental models – iconic representations of possibilities – of events (Khemlani, Harrison, & Trafton, 2015; Schaeken, Johnson-Laird, & d'Ydewalle, 1996). In line with recent accounts of spatial reasoning (Ragni & Knauff, 2013), the theory posits that individuals tend to build a single preferred model of a temporal description. The more models necessary to yield a correct answer, the harder that problem is. The theory is implemented in a computer program, mReasoner, which draws temporal deductions by building models. It varies four separate factors in the process: the size of a model, its contents, the propensity to consider alternative models, and the propensity to revise initial conclusions. Two studies corroborated the predictions of the theory and its computational implementation. We conclude by discussing temporal and relational inference more broadly.

An interactive model accounts for both ultra-rapid superordinate classification and basic-level advantage in object recognition

Qihong Lu
University of Wisconsin-Madison, Madison, WI, United States

Timothy Rogers
University of Wisconsin-Madison, Madison, WI, United States

Abstract: While people are faster to categorize objects at an intermediate or basic level of specificity (e.g. "bird"), several recent studies have shown them to have much earlier access to more general category information (e.g. "animal"). Ultra-rapid superordinate classification has been taken as evidence that recognition processes are largely feed-forward. In simulations with a deep neural network model, we show that this conclusion does not follow: even a model that is fully recurrent and interactive shows ultra-rapid superordinate classification patterns when tested with analogs of behavioral tasks such as rapid serial visual presentation or deadline classification. Moreover, this recurrent model explains recently-observed similarities and differences in the time-course of classification as estimated by electro-encephlography (EEG) versus human electro-corticography (ECoG), and also account for the well-known basic-level advantage in non-speeded classification. These results provide evidence that ultra-rapid and unconstrained visual object recognition is supported by interactive processes in the brain.

Handedness and Mathematics: Toward a More Comprehensive Model

Giovanni Sala
University of Liverpool, UK

Martina Bolognese
University of Milan, ITA

Giulia Barsuola
University of Milan, ITA

Michela Signorelli
University of Milan, ITA

Fernand Gobet
University of Liverpool, UK

Abstract: The relationship between handedness and mathematical abilities is controversial. Whilst some researchers have claimed that left-handers are gifted in mathematics and strong right-handers perform the worst in mathematical tasks, it has been more recently proposed that mixed-handers are actually the most disadvantaged group. To disentangle these discrepancies, we conducted five experiments in several Italian schools (total participants: N = 2,308) involving students of different ages (6 to 17 years) and a range of mathematical tasks. The results showed that (a) the percentage of variance in mathematics scores explained by handedness was moderate (about 5%) but statistically significant, and (b) the shape of the relationship between handedness and mathematical ability depended on age, task, and gender. We concluded that the different outcomes reported in the literature probably reflected the dissimilarities between the studies about the above variables. Therefore, a more comprehensive model is needed, which explains how these variables interact.

Shifting meanings: The fluidity of signal-meaning mappings in a minimal communicative task

Jennifer Misyak
University of Warwick

Takao Noguchi
University College London

Nick Chater
University of Warwick

Abstract: We used a non-linguistic experimental paradigm to explore the instantaneous creation of new communicative conventions. Participants played a computer game, in which they sent and interpreted minimal signals to obtain shared rewards within a virtual scene. Trials manipulated the space of possible signals that could be sent, and the meanings to be expressed (locations and quantities of rewards); as such, optimal success in the task required participants to jointly construct signal-meaning mappings that functioned as part of a system, rather than in isolation.

We observed different signalling strategies among participants, but with some individuals using 'system-mapping' conventions that globally reorganized in light of changing task constraints. Such behaviour reflects the principle of pre-emption in pragmatics, where the inferred meaning of an utterance depends on its relationship among a set of alternatives. Our initial findings provide a basis for future research, investigating contexts that are conducive to this phenomenon.

Process of visual input does not decide the accuracy imitation performance

Takashi Mizuguchi
Shinshu University

Abstract: The associative sequence learning (ASL) model states that error patterns in observed actions during physical imitation and verbal description are identical, because of the critical role played by the process of visual input compared to the process after visual input. Action models were presented that comprised four elements: using right or left hand, using right or left stick, tapping right or left side of a box, and placing a stick on right or left side. In the condition in which identical elements of video stimuli and manipulated objects placed in front of participants had the same color, the colored element was correctly performed compared to the condition with different colored elements. However, colored element was not correctly performed in the condition in which particular elements of video stimuli were colored, whereas manipulated objects were not colored. These results suggest the important role of the process occurring after the visual input.

Reduced benefit from regularities in language among Dyslexics

Eva Kimel
PhD student at The Hebrew University of Jerusalem

Merav Ahissar
Professor at The Hebrew University of Jerusalem

Abstract: The "Anchoring Deficit" hypothesis (Ahissar et al., Nat Neurosci. 2006) proposed that Dyslexics have a difficulty in automatic extraction of simple stimulus regularities in sound sequences. JaffeDax et al. (J Neurosci. 2015) modelled these difficulties as yielding noisy priors.

The current study was aimed to assess the impact of long-term regularities in language, which listeners had life long experience with. Our assumption was that this familiarity would enhance Controls' performance more than Dyslexics' due to a noisier prior formation among Dyslexics. This question was addressed in a series of experiments - in each there was one condition for which information accumulated over the life span could be utilized.

In all three experiments Dyslexics did not benefit as much as Controls from the long term statistics associated with the input. These results suggest that Dyslexics could not compensate for the deficit despite multiple exposures to lingual input with the same statistics.

How the Physicality of Space Affects How We Think about Time

Jennifer Kolesari
University of Notre Dame, Notre Dame, IN, 46556

Laura Carlson
University of Notre Dame, Notre Dame, IN, 46556

Abstract: Time is an abstract concept that is better understood when it is mapped onto space. One mechanism to accomplish this mapping is a reference frame. Previous research has shown the orientation and direction parameters of a spatial reference frame are involved in understanding time. For example, for English speakers, time is organized horizontally and runs from left (past) to right (future). The current experiments focus on the scale parameter. Experiment 1 changes temporal scale across trials, and illustrates that the scale parameter is set, as evidenced by a cost when the parameter value changes. Experiment 2 examines the correspondence between the spatial scale and the temporal scale, requiring participants to map small or large temporal distances to small or large spatial distances. The results illustrate flexibility in this mapping. Together these experiments support the idea that all the parameters of a spatial reference frame are used when understanding time.

Embodiment Effects in Evolutionary Robotics

Nicholas Livingston
Vassar College

Anton Bernatskiy
University of Vermont

Kenneth Livingston
Vassar College

Marc Smith
Vassar College

Jodi Schwarz
Vassar College

Joshua Bongard
University of Vermont

David Wallach
Vassar College

Evan Altiero
Vassar College

John Long
Vassar College

Abstract: We evolve simple neural network controllers in swimming robots in order to test the hypothesis that, given distinct dimensions of control for the tail structure, evolution will favor the emergence of modular neural networks as most likely to enhance fitness (successful light harvesting). Evolution does lead to improved fitness, but this does not appear to result from increases in modularity. However, an unexpected result highlights the importance of embodiment for the evolution of the agent. The output of the neural network controller is high frequency with many extreme excursions, but the actual movements of the tail are damped by the physics of the body as it interacts with the aquatic environment. Subsequent simulations establish the role of these physical parameters in dampening noisy network controller output. Thus, morphology can increase evolvability by acting as a low pass filter of high-frequency controller dynamics.

Perception of math and non-math content in children's storybooks

Colleen Uscianowski
Teachers College Columbia University, New York, NY, United States

Ma. Victoria Almeda
Teachers College Columbia University, New York, NY, United States

Cassiopeia Freeman
Teachers College Columbia University, New York, NY, United States

Herbert Ginsburg
Teachers College Columbia University, New York, NY, United States

Abstract: Young children acquire informal math ideas through everyday experiences. Research demonstrates that storybooks can link their informal notions to abstract concepts (Murphy, 1999). The integration of visual and written depictions of mathematics, along with conversations arising from the story, can bolster children's capacity to communicate and think mathematically. Despite the growing literature supporting use of math storybooks, little is understood about how educators perceive and interpret the embedded math content. Our study presents storybook pages to educators and asks questions to determine whether they are more likely to identify concepts in math (e.g., number) or non-math (e.g., vocabulary) domains. We also analyze the association between domain and the degree of abstraction in the language used to describe content in that domain. Preliminary data suggest that educators ask questions about number concepts at higher levels than expected and character's actions at lower levels than expected.

Your Obstacle on My Mind: Task Co-representation in Coordination is Modulated by External Timing Cues

Laura Schmitz
Central European University, Budapest, Hungary

Cordula Vesper
Central European University, Budapest , Hungary

Natalie Sebanz
Central European University, Budapest, Hungary

Guenther Knoblich
Central European University, Budapest, Hungary

Abstract: When acting in a social context, people have an automatic tendency to represent another person's task – to the extent that another's task constraints may influence one's own movement performance. Task co-representation will also affect co-actors' performance in joint action coordination; however, how exactly movement parameters are influenced is unclear. We investigated this question in four experiments. Pairs of participants performed arm movements back and forth between two targets, instructed to synchronize their landing times while external metronome tones provided timing cues. We predicted that actors would represent their co-actors' task constraints such that when the co-actor moved over an obstacle the actor without obstacle would move higher as well. Results confirmed this prediction, suggesting that joint action partners co-represented each other's task constraints. Moreover, this obstacle effect increased significantly when timing cues were removed, indicating a stronger need for co-representing the partner when demands on interpersonal coordination are amplified.

Prediction of Single-Trial Behavior using a Layered Dynamic Systems Model with Evolutionary Algorithm Updating

Richard Prather
University of Maryland, College Park, MD, USA

Sara Heverly-Fitt
University of Maryland, College Park, MD, USA

Abstract: In this study we attempted to predict individual participants single trial behavior (response and reaction time) on a non-symbolic number comparison task. Experimental sessions included the completion of the number comparison task along with concurrent EEG measures. We then used a dynamic systems model with evolutionary algorithm updating to predict behavior for each participant independently. The computational model approximated neural coding of number by calculating tuning curves implemented through multilayered dynamic systems architecture. Typically dynamical systems models of cognition have fixed parameters tailored to the particular task being modeled and selected by the researcher. The models used were designed to adapt such that each participant's model is individually customized to their particular data. Average ERP amplitude across occipitoparietal areas were used as model input in addition to participant's prior responses and reaction time.

The role of viewpoint in event segmentation

Khena Swallow
Cornell University, Ithaca, NY, 14853

Ayse Candan
Cornell University, Ithaca, NY, 14853

Jovan Kemp
Cornell University, Ithaca, NY, 14853

Abstract: The ability to perceive and understand experience is influenced by a process that divides it into meaningful parts. This process, called event segmentation, is frequently studied by asking participants to identify meaningful units of activity in films that depict it from a third person perspective. However, because segmentation is associated with changes in the perceptual and conceptual features of film, it could differ for films that present the same activity from a first person perspective. This study therefore examined segmentation for simultaneously recorded films that depicted identical activities from different perspectives. In several experiments participants were asked to segment these movies into natural and meaningful units of activity. Measures of segmentation frequency and agreement provided little evidence that segmentation reliably differed across first and third person perspectives. These preliminary findings suggest that participants identify similar events when they are viewed from different perspectives.

Individual differences in verbalization predict change detection performance: A new perspective on the language-thought debate

Junko Kanero
Temple University, Philadelphia, PA, USA

Eileen Kitrick
Colorado College, Colorado Springs, CO, USA

Kathy Hirsh-Pasek
Temple University, Philadelphia, PA, USA

Kevin Holmes
Colorado College, Colorado Springs, CO, USA

Abstract: The question of whether language affects nonlinguistic processes remains unresolved. Whereas many studies find that effects of language on such processes are disrupted when verbalization is inhibited, others show that they persist. We explored individual differences in the tendency to verbalize as a potential resolution to this discrepancy. We hypothesized that if language is spontaneously accessed during nonlinguistic tasks, individual differences in verbalization should predict task performance. Participants completed a visual change-detection task and the Verbalizer-Visualizer Questionnaire (VVQ), a self-report measure of cognitive styles linked to modality-specific neural systems. We found that higher scores on the "verbalizer" dimension of the VVQ predicted faster but less accurate change detection. These results suggest that some individuals are more likely than others to use language when performing tasks that do not require it, and hence that effects of language on nonlinguistic processes are more likely to be observed in such individuals.

Do Subliminal Hints Facilitate Sequential Planning When Solving a Spatial Insight Problem?

Hiromitsu Miyata
The University of Tokyo

Hiroaki Suzuki
Aoyama Gakuin University, Shibuya, Tokyo, Japan

Abstract: Subliminal information has been suggested to facilitate insight problem solving. The present study examined whether subliminal hints may influence not only retention of the goals but also planning of sequential steps required to solve the nine-dot problem. Using continuous flash suppression, participants were subliminally presented with either an image depicting the entire solution of the problem, the three steps to solve the problem in a sequential order, or an image of the nine dots that does not involve solution of the problem. Participants presented with the entire solution of the problem tended to show better solution performance and greater relaxation of constraints than those in the latter two conditions, whose performance failed to significantly differ from each other. These results indicate that subliminal information may be stored as a global and static visuo-spatial representation to influence solution but may not involve planning of each step to achieve insight problem solving.

Electrophysiological markers indicate disturbance of involuntary attention, but not voluntary attention, in adult ADHD patients

Frank Nonnenmacher
University of Tübingen, Tübingen, BW, Germany

Boris Kotchoubey
University of Tübingen, Tübingen, BW, Germany

Abstract: When we cannot concentrate on reading a book, we have problems with voluntary attention. When we stand up and stumble on a chair leg, we have a problem with involuntary attention. A bimanual Stroop task (ST) and the Ericson's flanker task (EFT) were used for the analysis of voluntary and involuntary attention, respectively. Electrophysiological markers of attention were applied in adult ADHD patients and yoked control individuals. Behavioral incongruence effects were stronger in patients than in controls in the EFT. P3 latency in the incongruent condition was identical in patients and controls in ST but strongly delayed in patients compared with controls in EFT. A significant positive lateralized readiness potential indicating the activation of the false response channel was obtained in the incongruent condition of EFT, being significantly larger in patients than controls. The data indicate a disorder of automatic attentional control in ADHD adults despite nearly normal voluntary attention.

Biased Attention to Spatial Dimensions Predicts Children's Spatial Word Acquisition

Hilary Miller
University of Wisconsin Madison

Haley Vlach
University of Wisconsin, Madison, Madison, Wisconsin, United States

Abstract: Children's spatial language abilities relate to their spatial skills. We propose that this relation arises from attention to spatial dimensions influencing both spatial word and spatial skill acquisition. This study tests whether attending to spatial dimensions in a word learning task predicts spatial vocabulary. Three to 5-year-olds completed a novel word assessment testing categorization of angles, shapes, and a test of spatial vocabulary. In the novel word assessment, children were presented with an exemplar angle with a novel label and asked to select another angle sharing the label. Foils matched the exemplar in degree, orientation, color, or size. Significant age differences occurred in children's bias to select foils based on angle degree (but no age differences occurred in exemplar choices based on shape). Children showing an angle bias had significantly higher spatial vocabulary than those who did not. These findings show that attending to relevant spatial dimensions predicts spatial vocabulary.

On Measuring the Difficulty of Scrabble-like Problems

Morgan Magnus Fleming
University of California, Merced

Paul P. Maglio
University of California, Merced

Abstract: Scrabble-like tasks have increased in popularity as a means of exploring cognitive phenomena, such as embodied problem solving, mastery, and creativity. Many of these tasks make assumptions about the key factors driving relative difficulty of word-finding problems; these factors include average frequency of words produced, number of words produced, and number of readily apparent bi-grams in the initial presentation of the letters. This study measures the effects of each of these factors on cognitive load by systematically and empirically exploring such factors, comparing how these various attributes influence the number of words participants produce in different circumstances.

Metaphorical Color Representations of Emotional Concepts in English and Chinese Speakers

Junqing Chen
The Graduate Center and Brooklyn College, CUNY, New York, NY, USA

Natalie Kacinik
Brooklyn College, CUNY, Brooklyn, NY, USA

Yingjun Chen
Shanghai Normal University, Shanghai, China

Nianyang Wu
Shanghai Normal University, Shanghai, China

Abstract: This project examined whether the cultural and linguistic experiences of English and Chinese speakers can result in different metaphorical representations of emotion in those individuals. The Brief Implicit Association Test (BIAT) technique was used to measure how strongly various colors are associated with anger, sadness, happiness, fear, envy, shame and shyness. The results showed that some metaphorical associations like red-anger are common in both English and Chinese speakers, whereas other associations are culturally-specific (e.g., red is also associated with happiness in Chinese, while only English individuals associate blue with sadness). Some interesting gender differences were also obtained, such that Chinese females associate shyness with pink, but males with red. Black was associated with fear in both genders in Chinese, but only present in English males. This study thus demonstrates that the conceptual representations of different emotions are shaped by an individual's linguistic and cultural experience.

Feature distinctiveness in verbs: links between verb distinctiveness, child directed speech and age of acquisition

Tomas Engelthaler
University of Warwick

Thomas Hills
University of Warwick

Abstract: Feature distinctiveness is a measure representing the uniqueness of objects' features. Previous research found links between noun feature distinctiveness and age of acquisition (i.e. nouns referring to objects with relatively unique features are learned earlier). The present work investigates the links between feature distinctiveness and age of acquisition in verbs. Using high-dimensional vector space modelling, noun and verb feature distinctiveness was represented as Manhattan distance between word nodes. Both nouns and verbs showed negative correlations between feature distinctiveness and age of acquisition (words of more distinctive objects learned earlier), suggesting a general distinctiveness bias. This effect was stronger for nouns. An investigation of child directed speech (CDS) from the CHILDES corpus showed a correlation between child directed word frequency and feature distinctiveness for nouns (featurally distinctive nouns are more common in CDS), but not for verbs. The possible link between distinctiveness in CDS and age of acquisition effects is discussed.

The Effects of Grammatical Aspect and Visual Perspective on Imaginging Accomplishments

Deanna Hall
Wilfrid Laurier University

Todd Ferretti
Wilfrid Laurier University

Jeffrey Hong
Wilfrid Laurier University

Abstract: We examined how grammatical aspect and visual perspective taking (first- or third-person) influence the ability to imagine accomplishments. Our main prediction was that it would be easier to imagine completed (I had built the fence.) than ongoing events (I was building the fence.) because accomplishments include natural temporal end points. Slow cortical brain potentials were examined as an index of the difficulty associated with imagining. Our results showed that participants had more difficulty imagining ongoing than completed accomplishments, and that it was easier to imagine from the third- versus first-person perspective. The ability of participants to imagine from different visual perspectives was not influenced by grammatical aspect. Participants indicated that the people in their imagined events were more vivid when they imagined ongoing versus completed accomplishments, as well as when they imagined from a third- versus first-person perspective. As expected, grammatical aspect influenced which temporal components of events were imagined.

A Cognitive Model of Fraction Arithmetic

David Braithwaite
Carnegie Mellon University

Robert Siegler
Carnege Mellon University

Abstract: Learning about fractions is a critical step on the path to high school mathematics, yet many children never master basic knowledge such as fraction arithmetic procedures. To better understand these difficulties, the present study describes a computational model of fraction arithmetic problem solving. The model demonstrates that the majority of empirically observed errors over all four arithmetic operations can be explained by only two error-generating mechanisms – overgeneralization and repair. Further, by assuming probabilistic selection of solution procedures using associative strengths learned from experience, the model predicts two other empirical phenomena: (1) variation in error rates and relative frequencies of specific errors as a function of problem features, and (2) variable strategy selection within and between problems and individuals. Beyond providing a formal account of errors, the model was used to simulate the effects of variation of instructional parameters, leading to novel predictions regarding potentially effective instructional designs.

The role of regional topography in route planning

Andrew Deslauriers
Colorado College

Kevin Holmes
Colorado College

Abstract: When planning the most efficient route from one location to another, people tend to prefer southern routes over northern routes of equal length and complexity. This asymmetry has been attributed to implicit associations between cardinal direction and relative elevation (i.e., north = higher), and holds even when regional topography conflicts with these associations. No such asymmetry has been observed between eastern and western routes. Here we provide evidence for an eastern-western asymmetry in participants residing in an environment with east-west topography differences. Residents of Colorado Springs, CO, where topography is mountainous to the west, showed a reliable preference for eastern routes over equal-length western routes on a Colorado Springs map, but not an unfamiliar map. This pattern held even though the represented areas contain minimal elevation differences. Our findings suggest that regional topography can induce a novel, physically unfounded asymmetry in otherwise unbiased representations of the spatial environment.

Sign languages reveal spatial mappings of valence and magnitude

Eileen Kitrick
Colorado College

Kevin Holmes
Colorado College

Abstract: Much research indicates that concepts of magnitude and valence are represented spatially, with more/less and positive/negative relations mapped to vertical and horizontal axes. While these mappings are sometimes manifested linguistically through conventional metaphors (e.g., "prices fell"), recent evidence suggests that they may be built into the very forms of words – traditionally assumed to be arbitrarily related to their meanings. Following previous research, we examined whether the directions of hand motions constituting words in two sign languages predicted the meanings of their English translation equivalents. Upward-moving signs were more positively valenced than downward-moving signs, as found previously, but were also greater in magnitude, or intensity. Additionally, rightward-moving signs (from the signer's perspective) were more positively valenced than leftward-moving signs, consistent with the bodily experience of right-handers. Our findings demonstrate systematic encoding of multiple spatial-conceptual mappings in words, adding to the growing literature showing non-arbitrary links between linguistic form and meaning.

Informant effort expenditure impacts young children's learning, eye gaze, and trust

Molly Schlesinger
University of California Riverside

Rebekah Richert
University of California Riverside

John Franchak
University of California Riverside

Abstract: Abstract: Recent research has suggested informant trust is an important factor in preschoolers' observational learning. This poster will present data from an ongoing study examining if 3.5- to 6.5-year-old children (current n=24) relate perceptions of effort and trust. Children watched two informants solving problems using different solutions, exerting either high or low effort. Children's eye gaze, trust of each informant, and learning from informants were measured. There were no significant differences in trust of the two informants, but children were significantly more likely to learn the solution demonstrated by the high effort informant, $t(23) = 2.161$, $p = 0.041$. High effort informant trust was also significantly related to time spent looking at the high effort informant, $r = 0.675$, $p < 0.01$. These findings indicate children are more likely to watch informants who exert high effort and are more likely to use those solutions when faced with a novel problem.

Audio-Visual Task Switching in Multisensory Environments

Christina Wasylyshyn
Naval Research Laboratory, Washington, District of Columbia, United States

Abstract: The majority of task-switching research has focused on shifting attention between multiple tasks in the same perceptual modality (i.e., visual) within a single task domain. However, typical environments are not unisensory, and typical response decisions often involve multiple task domains. This study examines multisensory task-switching costs and the interactions of several variables, including perceptual modality of the cue, perceptual modality of the target task, type of task completed (i.e., spatial or identity decisions), and availability of foreknowledge. The design is marked by no redundant multisensory information and minimal memory demands. Performance costs varied as a function of whether participants had foreknowledge of upcoming task and/or modality presentation. Consistent with previous research, the current results also show that performance costs between tasks were significantly smaller (and essentially, eliminated) when the sensory modality of the task switched versus when it repeated. However, this result was contingent on manipulations of the experimental design.

Causal Representation in Foresight: Can We Improve Memory and Novel Understanding?

Edward Munnich
University of San Francisco

Emma Weinberger
University of San Francisco

Ana Maria Hoffmann
University of San Francisco

Abstract: Learning facts without first considering what they could be can lead to hindsight bias. How might this impact (a) memory and (b) understanding of novel topics? Foresight participants read about five psychology studies, including mean performance of one group; then they estimated the mean performance of another group and stated causes for the difference; finally, they received the second group's actual performance. Hindsight participants learned about both groups' performance at the beginning, then imagined what estimates and causes they would have indicated had they not seen actual means. A week later, half of each group recalled the means they had learned, and other half estimated means for a novel set of studies. We considered the extent to which: 1. foresight promotes long-term memories as opposed to providing an anchor that biases memories; 2. foresight cultivates a habit of considering alternative possible outcomes that might help one understand novel topics.

From computation to automization: How practice alters initial neural response to familiar arithmetic problems

Caitlin Tenison
Carnegie Mellon University, Pittsburgh, PA, USA

John Anderson
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract: Building and validating models of skill acquisition that explain speedup effects has been limited by difficulty distinguishing quickly executed cognitive processes (e.g. Anderson, 1982; Logan, 1988; Rickard, 1997). In this experiment, magnetoencephalography (MEG) data are collected from participants solving a repeated math problem set. We use MEG signal to test the three-phase model of skill acquisition that describes the transition from problem-solving strategies of computation, to retrieval, to an automatic stimulus-response process (Fitts & Posner, 1967). We hypothesize that the processes of familiarity and recollection are early features that distinguish the three phases of skill acquisition. Analyzing event-related fields, we test two predictions. First, early frontal activation (akin to the FN400 old-new effect of ERP studies) should diminish in strength with each successive phase transition. Second, parietal activation (corresponding to the ERP P600 old-new effect) should be present in the second phase, but not in the first or last phase.

Influence of Need for Cognition and Cognitive Closure on Magic Perceptions

Lisa Grimm
The College of New Jersey

Nicholas Spanola
The College of New Jersey

Abstract: From children's parties to acts in Las Vegas, magic is one of the world's most timeless forms of entertainment. Current psychological research on magic has started to focus on how magicians are best able to elicit the observer reactions associated with their craft, such as what methods are most successful, as well as what cognitive mechanisms are specifically driving the observers' reactions. However, while research examining the practice of magic from a psychological perspective has been expanding, few studies have looked at how cognitive individual differences influence an observer's magic perceptions and experiences. In a collaboration with award-winning magician, Joshua Jay, we examined the impacts of Need for Cognition (NFC) and Need for Cognitive Closure (NFCC) on magic perceptions. Results showed that NFC and NFCC had opposite effects on engagement (i.e., rewatching and solution generation) and that frustration levels were behavior drivers for participants with high NFC or low NFCC.

Asymmetric derivational priming in recognition of Greek nouns and verbs

Sofia Loui
University of Athens; Athens Metropolitan College

Athanassios Protopapas
University of Athens

Abstract: We examined differences between the processing of inflectional versus derivational morphology, using Greek nouns and verbs with a primed lexical decision task. Previous work suggested that both noun and verb targets were significantly primed by the same grammatical class. However, when preceded by different grammatical class, verb but not noun targets showed priming. We attributed the asymmetrical priming to the materials used: noun stimuli were derived by their verb counterparts, suggesting an important inherent asymmetry between nouns and verbs. To further investigate this suggestion, we used materials with the opposite asymmetry (verbs derived by nouns) expecting an asymmetry in the opposite direction to emerge for derivationally related words. A clear explanation of the asymmetry would allow us conclusions about the (debated) existence of differences in representation and processing between inflectional and derivational morphological relations and thus provide evidence for or against a fully decompositional view of processing morphologically complex words.

Linear separability and human category learning: Revisiting a classic study

Kimery Levering
Marist College, Poughkeepsie, NY

Nolan Conaway
Binghamton University, Binghamton, NY

Kenneth Kurtz
Binghamton University, Binghamton, NY

Abstract: The ability to acquire non-linearly separable (NLS) classifications is well documented in the study of human category learning. In particular, one experiment (Medin & Schwanenflugel, 1981; E4) is viewed as the canonical demonstration that, when within- and between- category similarities are evenly matched, NLS classifications are not more difficult to acquire than linearly separable ones. The results of this study are somewhat at issue due to non-standard methodology and small sample size. We present a replication and extension of this classic experiment. We did not find any evidence of an advantage for linearly separable classifications. In fact, the marginal NLS advantage observed in the original study was strengthened: we found a significant advantage for the NLS classification. These results are discussed with respect to accounts provided by formal models of human classification learning.

Concept of a Deity: Structure and Properties

Nicholas Shaman
University of California, Riverside, Riverside, CA, United States

Anondah Saide
University of California, Riverside, Riverside, CA, United States

Rebekah Richert
University of California, Riverside, Riverside, CA, United States

Abstract: Individuals attribute more psychological (e.g., forget) than biological (e.g., eat) or physical (e.g., be touched) properties to supernatural beings (Shtulman, 2008). It is unclear how those domains each contribute to an overall conception of a supernatural being (e.g., God). Undergraduate students (N = 341) responded to nine questions representing the three domains or factors (psychological, biological, and physical), composing an overall measure of God's anthropomorphic properties.

A confirmatory factor analysis was performed to assess the structure of undergraduates' anthropomorphic concept of God. Fit indices suggest acceptable model fit, $\chi^2(24) = 73.09$, $p < .001$, CFI $= 0.952$, SRMR $= .051$. All loadings were significant. Biological (0.99; 0.01) and physical (0.90; 0.19) factors loaded more strongly onto anthropomorphism, and had smaller variances, than the psychological (0.67; 0.56) factor. These findings suggest there are varied ways of conceptualizing the psychological (versus non-psychological) properties of God; thus, non-psychological properties are more predictive in God concepts.

Do peer interactions influence infant communication development?

Anne Warlaumont
The University of California, Merced, Merced, CA, USA

Gina Pretzer
The University of California, Merced, Merced, CA, USA

Abstract: Studying infants in daycare or school settings enables us to ask whether infants influence each others' development, and if so, whether peer influences are similar to influences from adult caregivers. Answering these questions will not only inform infant educators but can also help us understand the mechanisms underlying infant learning. We have collected audio recordings from 21 1- to 21-month-old infants in two infant rooms in our campus early childhood education center. Recordings took place nearly every school day over a continuous 8 month period, for an average of 231 hours of recording per child. Multiple infants within the same room were recorded simultaneously. We will present our approach to synchronizing, coding, and analyzing these recordings toward the goal of understanding peer influences on vocal communication development, present preliminary results, and seek input on how to further analyze this large and unique dataset.

How event endstates are conceptualized in adults and infants

Angela Xiaoxue He
Boston University, Boston, Massachusetts, United States

Sudha Arunachalam
Boston University, Boston, MA, USA

Abstract: Many event descriptions are true only when the event comes to its natural end point: e.g., a "feeding" event culminates when the feed-ee has eaten, not simply when food is provided. Do non-linguistic event conceptualizations reflect attention to natural culmination points? We tested adults and 14-month-olds to ask: provided two events with the same ACTION but different ENDPOINTs - one a naturally expected result, the other only partially achieved - do adults and infants perceive them as members of the same event category or of different categories? Adults were asked to rate the similarity between the two events; infants were habituated to one event and tested for dishabituation when it was switched to the other. Adult data suggest the difference between a complete and a partially-complete event is registered, and carries more psychological weight than a mere perceptual difference. Infant data (ongoing) will show the developmental origin of such conceptualizations.

ADHD modulates link between event processing and recall

Robbie Ross
University of Oregon, Eugene, Oregon, United States

Leah Child
University of Chicago, Chicago, IL

Dare Baldwin
University of Oregon, Eugene, Oregon, United States

Abstract: How might ADHD symptomatology influence adults' ability to process and recall the actions of others? University undergraduates observed a woman packing a suitcase by advancing through a self-paced slide show of still images extracted from a digital video, and were then asked to recall as many actions as possible. Results showed that lower self-reported retrospective ADHD symptomatology was associated with a) longer overall dwelling on images from the slideshow, $r(91) = -.315$, $p=.002$, and b) recall of more actions $r(85) = -.237$, $p=01$. Further, exploratory analyses indicated that ADHD symptomatology modulated the specific linkage between dwell time patterns and stronger recall: Attention to fine-grain details within activity improved recall for those reporting higher ADHD symptomatology; those reporting lower ADHD symptomatology displayed stronger recall when prioritizing attention at a more coarse-grain level, $F(1,83) = 4.19$, $p=.04$. These findings offer suggestive novel evidence that ADHD has implications for event processing.

The sequence of study changes what is encoded during category learning.

Paulo Carvalho
Indiana University

Robert Goldstone
Indiana University

Abstract: This work investigates how the sequence of study influences encoding and memory for different properties of the categories studied. We used a transfer task with different types of items and show that following blocked study learners are sensitive to category properties that were presented frequently in the category (but were not diagnostic of category membership). However, following interleaved study learners do not seem to be sensitive to changes on these non-diagnostic properties. Moreover, when asked to judge different properties for their relevance for category learning (cue and category validity), participants rate discriminating properties more highly than similarities following interleaved study, but not following blocked study. These results are consistent with previous evidence and are captured by an exemplar model that takes into account the sequence of exemplars during learning by changing the likelihood of attending to and encoding different object properties depending on sequential similarities.

Finding Clarity Amidst the Clutter: How Parents Name Objects

Charlene Tay
Indiana University Bloomington

Linda B. Smith
Indiana University Bloomington

Chen Yu
Indiana University Bloomington

Abstract: A core issue in the study of word learning is understanding how beginning learners cope with referential ambiguity in the clutter of natural learning environments, and how parents may help them find the referent in that clutter. Here we ask how sensitive parents are in taking advantage of optimal visual moments where a single object is visually large in view to provide linguistic labels for their infants. Using a mini-head camera, we recorded parent-child free play interactions and studied the parent naming events for 12 and 30 month old children from the infant-perspective in a context of high clutter (30 objects dumped on the floor). Despite the cluttered context, parents and infants frequently created infant-perspective scenes in which one object was visually singled out. At both age levels, parents named objects in these moments of visual clarity and almost never named objects in sub-optimal moments.

Relational discovery in category learning: interactions of learning strategy and task structure

Micah Goldwater
University of Sydney, NSW Australia

Hilary Don
University of Sydney, NSW Australia

Evan Livesey
University of Sydney, Sydney, NSW, Australia

Abstract: Often failures of problem solving on educational assessments are failures of problem categorization. That is, when reasoners do not properly classify a novel problem they do not know what solution to apply. For example, often physics students do not recognize the underlying commonalities in the relationships among the variables in different problems concerning Newton's laws of motion (Chi, Feltovich, & Glaser, 1981). Addressing this challenge there have been separate lines of research examining 1. how differences in students' learning strategies or cognitive abilities affects their propensity to discover relational commonalities (e.g., Little & MacDaniel, 2015) and 2. how variations in task structure change the likelihood of successful categorization (e.g., Roher & Pashler, 2010). However, relatively little research has examined whether the optimal task structure depends on the learner's strategy or ability. Across several experiments, we demonstrate multiple dependencies between the effectiveness of different task structures on differences in learning strategy.

The effect of loss and gain expression in the riddle on insight problem solving

Reiko Yakushijin
AOYAMA GAKUIN UNIVERSITY, TOKYO, JAPAN

Katsutochi Endo
AOYAMA GAKUIN UNIVERSITY, TOKYO, JAPAN

Abstract: Previous studies suggested that heuristics sometimes obstructed solving particular insight problems (e.g. Knoblich, 1999). Abe & Nakagawa (2008) took up the Cheater Detection Model (CDM: Cosmides, 1989) as an adaptive heuristic within social environment, and suggested that it has a negative influence on the 'missing dollar' riddle. In this study, we examined the same type of insight problem from a different point of view, i.e., the balance of loss and gain. We made two isomorphic riddles: in one riddle some amount of money was lost and was never found (the loss condition), and in another riddle the same amount of money was spent for additional service (the benefit condition). The percentage of correct answer was significantly higher for the latter. The result suggested that the balance of loss and gain influenced the cognitive set of the participants. The occurrence of loss might draw their attention on outflow of the money.

Beyond the 64 Squares: Does Chess Instruction Enhance Children's Academic and Cognitive Skills? A Meta-Analysis

Giovanni Sala
University of Liverpool, UK

Fernand Gobet
University of Liverpool, UK

Abstract: In recent years, pupils' poor achievement in mathematics has been a concern in many Western countries. Chess instruction has been proposed as one way to remedy this state of affairs, as well as improving other academic topics and general cognitive abilities. The aim of this paper is to quantitatively evaluate the available empirical evidence that skills acquired during chess instruction in schools positively transfer to mathematics, reading, and general cognitive skills. The selection criteria were met by 24 studies (40 effect sizes), with a total of 5,221 participants. The results show (a) a moderate overall effect size (g = 0.34), and (b) a significant positive effect of duration of treatment ($p < .05$). However, almost no study controlled for placebo effects by using an active control group. For this reason, there are still doubts about the real effectiveness of chess instruction – in spite of some promising results.

Memory biases in matching and recall: Evidence from initial consonant clusters

Mackenzie Young
Johns Hopkins University

Colin Wilson
Johns Hopkins University

Abstract: Perception and memory of linguistic information is biased in favor of stimuli that conform to structural regularities. At the level of word form, there is evidence that initial consonant clusters varying in grammatical status (e.g., br vs. *bn, *bd, *rb) differentially affect response times in same-different matching (e.g., slower responses to rbif - REBIF than to brif - BERIF; Berent & Lennertz, 2010). Previous results are consistent with two hypotheses: non-conforming clusters could be modified by a specific 'repair' (e.g., rbif recoded as rebif), or the encoding of such clusters could be more uncertain and their recall more variable. A series of matching and full-recall experiments support the second hypothesis: the response time effect for non-conforming clusters is observed for both *rbif - REBIF and *rbif - RBIFE, but only the former 'repairs' the cluster; furthermore, errors made in recall exhibit high variability and do not systematically improve structural well-formedness.

Assessing Science Inquiry using MDP Goal Detectors

Michelle Lamar
Educational Testing Service

Janet Koster van Groos
Educational Testing Service

Abstract: Complex cognitive tasks, such as science inquiry, often involve a sequence of goals, each of which is pursued through a sequence of actions. Effective assessment of inquiry performance requires identification of these student goals. Markov decision processes (MDPs) have been used to infer goals and beliefs over a single directed sequence of actions (Baker et al., 2009), but multi-goal complex systems are computationally prohibitive to model. This research investigates the use of targeted MDPs as goal detectors, embedded within a larger hidden Markov model (HMM) that accounts for the transition between goals. This multi-layer approach allows the MDP state spaces to remain small while modeling complex cognition. Because canonical HMM estimation is complicated by the dynamic nature of MDPs, in which action probabilities depend on context, we explore several different estimation methods. The approach is applied to log-file data of test-taker interactions with a simulation-based science inquiry assessment.

Metaphor and Memory: How Metaphors Instantiate Schemas in and Influence Memory of Narrative

Sarah Cox
Oberlin College

Paul Thibodeau
Oberlin College

Abstract: Metaphoric frames are prominently featured in public discourse. They highlight certain aspects of the target issues they are used to describe, thereby encouraging specific patterns of inference. Our goal was to test whether they would influence memory as well. Building off prior work, we contrasted two metaphors for crime: virus and beast. In a pilot study, we identified specific causes, examples, and solutions to crime that were congruent with each frame (one but not the other; e.g., people thought "drug use" better exemplified a crime virus, whereas "murder" better exemplified a crime beast). Participants (n = 469) read or listened to a short metaphorically-framed crime report, completed a filler task, and were prompted for the information they had seen/heard. Results indicated the virus metaphor facilitated memory, overall, but not the specific frame-congruent information, suggesting a more general influence of the frame than predicted.

Is human learning driven by Prediction Error?

Jiří Čevora
Cognition and Brain Sciences Unit, Medical Research Council

Máté Lengyel
Computational and Biological Learning Lab, Department of Engineering, University of Cambridge

Richard Henson
Cognition and Brain Sciences Unit, Medical Research Council

Abstract: Prediction Error [PE] is a core component of some of the most influential theories of how animals use experiences to update their knowledge (e.g, Rescorla & Wagner, 1972). The classic demonstration of PE is the single-cell recording done by Schultz and colleagues (1997). However, there is no evidence that this signal plays any role in learning.

Only two studies have related a neural correlate of PE to learning performance so far (Gläscher, Daw, Dayan, & O'Doherty, 2010; McGuire, Nassar, Gold, & Kable, 2014). We provide a formal analysis demonstrating that non-PE learning can also explain the results of these studies if the imaging signal they identify relates to the size of weight updates instead of PE.

We conclude that the case for PE driving many forms of animal learning is not yet sufficiently proven, and identify approaches which can potentially resolve this question in future.

Modeling the Influence of Knowledge on Recognition: Connecting visual recognition behavior across development to PDP computational models of semantic knowledge

Clint Jensen
University of Wisconsin - Madison

Vanessa Simmering
University of Wisconsin - Madison

Timothy Rogers
University of Wisconsin - Madison

Abstract: Recent behavioral findings in children's selection of a "real" versus "silly" animal demonstrated a developmental change wherein younger children select chimeric animals with regular forms (e.g., a seal with four legs, a camel with no hump) as real. To reduce verbal demands while maintaining the same stimuli, we developed a touch screen change-detection task in which children (4 – 7-years-old) were instructed to locate a single changing feature (e.g., a rhinoceros with and without a horn) as fast as possible. Children were faster to find changes when the feature appeared on animals with more prototypic animal forms (e.g., a donkey with and without a hump) when compared to animals with atypical forms (e.g., a camel with and without a hump). Alongside exploration via computational models, these findings suggest that children's real-world object recognition is supported by the interplay of semantic knowledge, informed by covariation among visual features, and visual recognition.

Explanation-based discourse inferences support early word learning

Jessica Sullivan
Skidmore College

Juliana Boucher
Skidmore College

Rachel Goodkind
Skidmore College

Reina Kiefer
Skidmore College

Raymond Skyer
Skidmore College

Katie Williams
Skidmore College

Abstract: Children can learn new words from non-instructive contexts (e.g., overheard speech). Recently, it has been proposed that one way that children do this is by using the surrounding discourse to constrain the interpretation of new words (Sullivan & Barner, 2015). However, little is known about what sort of discourse inferences children might compute when learning. In the present study, we adopt a discourse-coherence framework (e.g., Rohde et al., 2006) in order to explain how preschoolers (N = 96, M = 49.2 months, range: 28-65 months) learn new words from discourse. We ask whether young children compute adult-like discourse coherence relations across clauses, and provide some of the earliest evidence that they do. We then relate children's ability to compute these discourse coherence relations to their ability to learn a novel word from that discourse, demonstrating the relation between the computation of discourse coherence and early word learning.

Moral language in the Basel Accords: A quantitative analysis

Eyal Sagi
University of St. Francis, Joliet, IL

Hadar Jabotinsky
Hebrew University, Jerusalem, Israel

Abstract: Technical documents are generally perceived as objective and free of opinion. The Basel Accords, a global financial regulatory standard, fall into this category. Therefore, political texts have to appear as morally neutral.

In this paper we argue that some moral arguments and convictions can be found in most texts, including technical ones. To test this hypothesis, we employ a novel quantitative analysis, based on corpus statistics and uncover elements of moral language that are present in the Basel Accords. In particular, we investigate the differences between the language used in different parts of the Accords and how it evolved over time.

Our results show an increase in moral language that emphasizes fairness following the financial crisis of 2007-2009. In contrast, moral language relating to authority greatly decreased in the most recent version of the Accords.

Measuring Cognitive Skills through Conversation-Based Assessment

G. Tanner Jackson
Educational Testing Service, Princeton, New Jersey, United States

Katherine Castellano
Educational Testing Service, United States

Diego Zapata-Rivera
Educational Testing Service, Princeton, New Jersey, United States

Abstract: Conversation-Based Assessment (CBA) represents a relatively new method of measuring student cognitive skills (e.g., science reasoning) through adaptive dialogues with automated characters. This approach leverages the openness of natural language with the interactivity of spoken dialogue to engage students in verbal reasoning and constructive processes (i.e., cognition). These two dimensions differentiate CBA from other assessment items (e.g., multiple choice and essays) by allowing for more freedom in responses along with the ability to adapt and follow-up on particular threads of information. The conversational exchange affords a rich data stream that can provide additional explanatory evidence of students' cognition exhibited through conversational content and dialogue paths. The current work, built on the AutoTutor dialogue engine, will discuss the affordances and constraints of CBA along with how this approach may complement and enhance other methods of measuring of cognitive skills.

Single-kernel models of single-voxel visual selectivities in convolution neural networks

Daniel Leeds
Fordham University, Bronx, New York, United States

Ivan Iotzov
Fordham University, Bronx, New York, United States

Abstract: The translation of retinal images into recognizable objects and scenes is not yet well understood. Beyond edge-detection in primary visual cortex, higher stages of cortical representation are still uncertain. We use a multi-layer convolutional neural network (Krizhevsky, 2012) to provide models for visual selectivities in the ventral visual pathway. We examine individual neural units, or "kernels", in CNN layer 2, correlating kernel activity to single fMRI voxel activity for 1750 natural images (Kay, 2008). Building on Güçlü (2015), we find most significant voxel-kernel correlations in V2, with additional matches throughout the ventral pathway. Notably, only 25% of kernels correlate with voxel responses — many voxels correlate with a consistent small set of kernels. Inhibition of voxel response for kernel selectivities also was observed. Our results indicate a limited number of CNN kernels may be used to gain a finer understanding of voxel level representations in the mid-level ventral visual pathway.

Attentional Resource Allocation in Multisensory Processing is Task-dependent

Basil Wahn
University of Osnabrück, Osnabrück, Lower Saxony, Germany

Peter König
University of Osnabrück, Osnabrück, Lower Saxony, Germany

Abstract: Human information processing is constrained by limited attentional resources. A matter of ongoing debate in multisensory research is whether attentional resources are shared or distinct across sensory modalities. Previous research suggested that the type of tasks that humans perform in separate sensory modalities determines whether attentional resources are shared or distinct across sensory modalities. Here, we investigated the relation between attentional resources and the performed type of tasks in four experiments using a dual task paradigm. We found shared attentional resources for vision, haptics and audition when two purely spatial tasks were performed in separate sensory modalities (Experiment 1 & 2) while we found distinct attentional resources for the same sensory modalities when a spatial task was performed together with a discrimination task (Experiment 3 & 4). Overall, our findings suggest that the distribution of attentional resources is operating at a task-level independent of the involved sensory modalities.

Cognitive Predictors of Timed and Untimed Early Arithmetic Performance

Olivia Wassing
King's University College at Western University , London , Ontario, Canada

Marcie Penner-Wilger
King's University College at Western University, London, ON, Canada

Abstract: Do established predictors of children's arithmetic performance differentially predict performance on timed versus untimed calculation tests? We investigated phonological awareness (i.e., CTOPP), phonological working memory (i.e., digit span), and visuo-spatial short-term memory (i.e., Corsi blocks) as predictors of timed and untimed calculation, both concurrently in Grade 1 (N= 116) and longitudinally in Grade 2 (N = 79). Timed calculation was operationalized as single-digit addition fluency and untimed calculation was operationalized as performance on the Woodcock Calculation subtest and KeyMath Numeration subtest. Examined concurrently, separate multiple regressions revealed that phonological awareness predicted timed calculation and all three cognitive measures predicted untimed calculation performance. Examined longitudinally, separate multiple regressions revealed that phonological awareness again predicted timed calculation and that phonological awareness and visuo-spatial short-term memory predicted untimed calculation performance. These results suggest a difference in the predictive set between timed and untimed calculation tests; furthermore, a difference between concurrent and longitudinal predictors.

On the Evaluability of Effort: Influences of Single and Joint Evaluation on Judgments of Subjective Effort in Memorial, Motor, and Perceptual Domains

Timothy Dunn
University of Waterloo, Waterloo, ON, Canada

Evan Risko
University of Waterloo, Waterloo, ON, Canada

Abstract: Theories attempting to explain the evaluation of subjective values often stress the importance of the context in which a judgment is made. One such theory, the General Evaluability Theory (GET; Hsee & Zhang, 2010), suggests judgments are particularly sensitive to evaluation mode (i.e., a single or joint evaluation). Importantly, deviations in the patterns of judgments of a value across single and joint modes are argued to reflect the degree to which individuals can consistently evaluate that value (i.e., the extent of evaluability). We applied this framework to a novel context, specifically the evaluation of effort. Individuals made judgments of effort across memorial, motor, and perceptual domains in single and joint evaluation modes. Results demonstrated that memorial and motor effort judgments remained largely consistent across modes, whereas perceptual effort judgments did not. These results provide initial evidence that at least some types of effort may not be evaluable.

Towards a Pre-Newtonian Intuitive Physics of Object Collisions

Ralf Mayrhofer
University of Göttingen

Michael Waldmann
University of Göttingen

Abstract: Some researchers have argued that mass perception, causal ascriptions, and predictions in simple billiard ball interactions can be modeled as inductive Bayesian inference over a (noisy) Newtonian representation of the world. However, there are phenomena, such as the asymmetrical ascription of forces to colliding objects, that are conceptually incompatible with the symmetry of Newtonian physics. We propose that human inference in physical scenarios operates over a pre-Newtonian physical representation that is based on impetus intuitions. Impetus theories assume that object movements are caused by an internal force, impetus, that is transferred and reflected when objects collide with each other. Moreover, impetus interactions are inherently asymmetric. We present a mathematical model that implements impetus theory and show that the theory is well suited to model perceived causal asymmetry. Moreover, the theory can also explain phenomena that so far have been presented as unique evidence for (noisy) Newtonian representations.

Parsing Selective Attention and Executive Function in Children

Staci Weiss
Temple University, Philadelphia, PA, USA

Peter Marshall
Temple University, Philadelphia, PA, USA

Abstract: Selective attention is related to a range of cognitive abilities, including executive function (Lawson et al., 2014). Orienting attention to visual and auditory targets are component skills inherent in many cognitive assessments, making it often difficult to parse cognitive capacities from selective attention abilities.The fundamental, early-developing nature of somatosensory processing (Saby et al., 2015) make it a compelling sensory domain within which to study top-down attentional processes. This presentation describes the initial results of a study examining how the electrophysiological indicators of selective attention, specifically the ability to focus attention on tactile stimuli, relate to children's executive function. Results will parse the relations between a composite of executive function tasks and the EEG mu rhythm response of participants when anticipating tactile stimulation of the hand. The implications of individual differences in somatosensory selective attention are discussed in light of cognitive assessment design, SES-related discrepancies in attention, and bodily awareness.

The effect of disfluency on mind wandering during text comprehension

Myrthe Faber
University of Notre Dame, Notre Dame, Indiana, United States

Caitlin Mills
University of Notre Dame, Notre Dame, Indiana, United States

Kristopher Kopp
University of Notre Dame, Notre Dame, Indiana, United States

Sidney D'mello
University of Notre Dame, Notre Dame, Indiana, United States

Abstract: Perceptual disfluency of a text can operate as a desirable difficulty, presumably because it leads to better comprehension. However, little is known about what cognitive mechanisms support this benefit. Here, we investigate whether sustained attention, as measured by reports of mind wandering (i.e., lapses in attention) during reading, mediates the relationship between disfluency and text comprehension. We manipulated the typeface (fluent: Arial; disfluent: Comic Sans) of two texts on research methods. A total of 208 participants recruited online read either one of these texts, either in a fluent or disfluent typeface, followed by a series of text level and inference level comprehension questions. We found that mind wandering was less frequent when participants read disfluent text. Importantly, our results show that the relationship between disfluency and text level comprehension was indirectly mediated by mind wandering, suggesting that sustained attention is one of the cognitive mechanisms by which disfluency influences comprehension.

Choosing Poorly: Reward-Induced Strategy Shifts in Estimating the Probabilities of Conjunctions and Disjunctions

James Tripp
University of Warwick

Adam Sanborn
University of Warwick

Neil Stewart
University of Warwick

Takao Noguchi
University College London

Abstract: Human estimates of the probabilities of combinations of events show well-established violations of probability theory, most notably the conjunction and disjunction fallacies. These violations have led researchers to conclude that the rules of probability are too complex for most people to use, and that cognitively-easier approximations such as averaging are used instead. Unlike previous work that has assumed that individuals use only a single combination rule, we collected repeated estimates of conjunctions and disjunctions and investigated whether individuals consistently used a single rule or used a repertoire of rules in a trial-by-trial Bayesian analysis. When not incentivized, most participants were best described as randomly selecting a combination rule on each trial, and the correct rule was the most often used. Despite this, when incentivized to use their single-best strategy participants were more likely to use the incorrect averaging rule. People do not seem to understand their own strategies well.

The presence of meaning constrains productive language processes: A test of the language game hypothesis in type writing.

Sebastian Wallot
Max Planck Institute for Empirical Aesthetics

Joachim Grabowski
University of Hannover

Abstract: How does meaning influence cognitive processes involved in the production and reception of language? The language-game hypothesis (LGH) states that meaning acts to constrain the cognitive processes involved in language comprehension. The degree of constraint can be gauged by measures of structuredness of a process, e.g. using Recurrence Quantification Analysis (RQA). LGH was originally formulated in the area of reading. The present study investigate its generalization to productive language processes, i.e. writing. In this study participants copy-typed a comprehensible text, written in their native language, and an incomprehensible text, written in an unfamiliar language. The writing process was recorded via key-logging and the time-series of inter-stroke-intervals was subjected to RQA. Results showed that comprehensible texts significantly increased the degree of structuredness of the writing process compared to incomprehensible texts. This suggests that meaning does indeed constrain language processes, and that this is the case for receptive and productive language tasks.

Modular versus Integrated Causal Learning

Bob Rehder
New York University

Kelly Goedert
Seton Hall University

Ciara Willett
Seton Hall University

Raymond Blattner
Seton Hall University

Abstract: Many pieces of information are potentially important to causal inference. Determining whether vitamin C prevents colds may entail knowing the frequency with which colds occur without vitamin C, other cold inhibitors, and the frequency of vitamin C use. Do reasoners integrate all this information to create coherent beliefs? In contrast to models emphasizing modular causal learning (e.g., Cheng, 1997), McDonnell, Tsividis, & Rehder (2013) proposed an integrated model, positing that individuals simultaneously update their beliefs about all components of a causal network. We tested modular versus integrated learning in two experiments using a retrospective inhibition design. In both, participants learned about two causes of headaches sequentially across two phases. We manipulated the base rate of headaches in phase II to be either consistent or inconsistent with phase I learning. Across experiments, participants failed to use base rate information as predicted by the integrated model, supporting modular causal modular learning.

A normative theory of visual working memory limitations

Ronald van den Berg
University of Uppsala

Wei Ji Ma
New York University

Abstract: There are many benefits to having a highly accurate representation of the environment. Why, then, has evolution equipped us with a visual working memory (VWM) system that can represent only a handful of items with high accuracy? Here, we offer a normative explanation for this limitation by conceptualizing VWM as a system that balances between two conflicting goals: keeping memory errors small and spiking activity low. We formalize this trade-off in a loss function and show that minimization of loss dictates a strategy in which memory precision declines with the number of remembered items. Using psychophysical data from 67 human subjects in 5 delayed-estimation experiments, we show that this normative model provides an excellent account of human VWM limitations. These results suggest that human VWM implements an optimal compromise between two conflicting ecological goals.

Verbs of Explanandum seem Crucial in Evaluating Explanations

Kwanghyeon Yoo
Sungkyunkwan University

Kyung Soo Do
Sungkyunkwan University

Abstract: Many previous studies have assumed that the domain of the explanandum determines which type of explanation, mechanistic or teleological, people prefer (domain theory). In this study, I proposed that the explanandum's thematic relation, which is mostly determined by the predicate, i.e., action verb or state verb, is crucial for the explanation type preference (thematic relation theory). To compare the two theories, participants were asked to read a sentence describing the explanandum, and then judge the appropriateness of two explanations for the explanandum, mechanistic and teleological, one after the other. Order of the two explanations were counterbalanced over participants. The domain and thematic relations were manipulated by varying the subject and the predicate of the explanandum. Mechanistic explanations were preferred when the predicate was a state verb, whereas teleological explanations were preferred for an action verb. Results of the experiment gave support for the thematic relation theory.

Network Analysis of Characters' Relationship in "Chronicle of Death foretold" using Graph Theory

Jongin Kim
Seoul National University

Yohan Moon
Sung Kyun Kwan University

Hwiyeol Jo
Seoul National University

Jeong Ryu
Yonsei University

Abstract: Over the past decade, there has been an explosion of interest in network analysis research across the social sciences and computer science. As it is an idea that can be applied in many fields, this study, in particular, its influence in the literature. We present a method for extracting social networks from literature. This study focuses on the relation between novel itself, narration in fiction and was carried out experiments with 89 undergraduate students. They were instructed to write down their remembered memory of the novel after reading the novel 'Chronicle of a death foretold.' We extract features from the social networks of characters in students' recall story and examine their differentiation with one another, as well as novel's setting. This study compares graph theory–based cohesion measures characters' relationship in novel and students' story. Our results suggest an alternative explanation for difference in social networks.

The construction of function representations

Brian Montambault
Brown University, Providence, RI, USA

Christopher Lucas
University of Edinburgh, Edinburgh, UK

Joseph Austerweil
Brown University, Providence, RI, USA

Abstract: Whether learning how pressing on the gas pedal of rental car will affect its acceleration or learning how changing the volume of speakers affects the perceived loudness of the sound they produce, humans can quickly learn functions from a few examples. Recent hybrid models (Lucas et al., 2015) combine the structure of rule-based models with the flexibility of similarity-based models by exploiting the equivalence of Bayesian linear regression and Gaussian processes. We expand on these models by taking advantage of the compositional nature of Gaussian processes and imposing a generative grammar over a set of base components in order to build the structured but diverse hypothesis spaces that appear to be represented by people. Subsequent testing will compare this model's ability to reproduce people's learning difficulty rankings of different functions, extrapolation results, and representations of multiple overlapping functions to that of other hybrid models.

Infants' Developing Coordinated Visual-Manual Object Exploration and Links with Vocabulary Development

Lauren Slone
Indiana University

Chen Yu
Indiana University

Linda Smith
Indiana University

Abstract: Research has demonstrated links between visual and manual object exploration and infants' object perception (e.g., Soska, Adolph, & Johnson, 2010). However, systematic investigation of the development of visual and manual object exploration and potential cascading effects on early word learning is lacking. In a longitudinal study of infants aged 9 to 24 months, we captured dynamic visual and manual information using head-mounted eye tracking and motion tracking of infants' hands as infants and their parents played with objects. Parents completed the MCDI vocabulary assessment at every visit. We will present preliminary data investigating individual and developmental differences in visual and manual object exploration, the resulting object views that are generated, and their relation to word learning. The results will inform our understanding of the relations between motor development, visual attention, and word learning in infancy.

Measuring individual and developmental differences in children's sense of confidence

Carolyn Baer
University of British Columbia, Vancouver, British Columbia, Canada

Darko Odic
University of British Columbia

Abstract: From something as simple as judging the time to more complicated behaviours like answering trivia questions, our cognitive systems always provide us with a representation of confidence: the probability of being correct. The development of confidence has been a long-standing issue in cognitive and developmental science. However, most studies assess children's confidence through either extensively trained numerical or verbal scales ("I am sure"), or by asking children to gamble on their answer. These measures stand to confuse metacognition with the development of language and inhibitory control. Here, we validate a novel model and task that measures individual and developmental differences in confidence relatively ("Are you more confident in X or Y"). Subsequently, we apply this task to demonstrate that metacognitive abilities of children aged 5–8 show significant development in the domain of intuitive number representations. These results are discussed in a broader context of theory and measurement of metacognition.

Assessing children's reading comprehension by the component processes tasks

Chi-Shun Lien
National Chung Cheng University, Chia-Yi, Taiwan

Yuhtsuen Tzneg
National Chung Cheng University

Hsin-Yin Chien
National Taitung University, Taitung, Taiwan

Chin-Chih Chen
National Chung Cheng University

Abstract: The purposes of this study were to develop a theoretical-based, comprehension process assessment and to measure children's reading comprehension processes. This assessment was based on Hannon & Daneman's (2001) paradigm and Hannon & Frias' (2012) component processes tasks, including the memory measure, the inference measure, knowledge access and integration measure, and modified to two parts in order to assess 4th to 6th graders' reading comprehension processes. We reduced the difficulties and complexity of this comprehension measure for younger children. Four-hundred-and-fifty participants (at 4th to 6th grade level) were recruited from four elementary schools in Chia-Yi, Taiwan. The results show that the Cronbach's alpha coefficients were .75 to .87 and the citerion-reference validity was around .70 to .75 with the Chinese Reading Comprehension Test. There were good item discriminations and difficulties, analysed by the Rash model.

Linguistic alignment with artificial entities in the context of second language acquisition

Astrid Marieke Rosenthal-von der Pütten
University of Duisburg-Essen

Carolin Straßmann
University of Duisburg-Essen

Nicole Krämer
University of Duisburg-Essen

Abstract: Native-speakers often adapt to non-natives in order to foster mutual understanding and successful communication, sometimes with the negative outcome of interfering with successful second language acquisition (SLA) on a native-speaker level. In two experimental studies we explored the potential of artificial tutors to avoid inhibition effects and exploit linguistic alignment processes in HCI for SLA. Study 1 (n=130 non-native speakers) investigated the influence of system voice (text-to-speech vs. pre-recorded speech) and embodiment (virtual agent vs. robot vs. speech based interaction) on participants' perception of the system, their motivation, their lexical and syntactical alignment during interaction and their learning effect after the interaction, while in Study 2 (n=85) embodiment and the presence of expressive nonverbal behavior were varied. The variation of system characteristics had barely influence on the evaluation of the system or participants' alignment behavior. Moreover, although participants linguistically aligned this did not result in significant short-term learning effects.

Temporal event clustering in speech versus music

Butovens Médé
University of California Merced, Merced, California, United States

Ramesh Balasubramaniam
University of California Merced, Merced, California, United States

Chris Kello
University of California Merced, Merced, California, United States

Abstract: Both speech and music can be organized as hierarchical, nested groupings of units. In speech, for instance, phonemes can group to form syllables, which group to form words, which group to form sentences, and so on. In music, notes can group to form phrases, which group to form chord progressions, which group to form verses, and so on. We present a new method for extracting events (amplitude peaks in Hilbert envelopes of filter banks) from speech and music recordings, and quantifying the degree of nesting in temporal clusters of events across timescales (using Allan Factor analysis). We apply this method to monologue recordings of speech (TED talks) and also to solo musical performances of similar lengths. We found that both types of recordings exhibit nested clustering, revealing similar organizational principles, but that clustering is more pronounced on shorter timescales (milliseconds) for speech, but longer timescales (seconds+) for music.

Developmental deficit in autobiographical episodic memory: Evidence from Williams syndrome

Katrina Ferrara
Georgetown University

Alexandra Cohen
Johns Hopkins University

Barbara Landau
Johns Hopkins University

Abstract: Williams syndrome (WS) is a genetic developmental disorder characterized by severe spatial impairments and structural and functional abnormalities in the hippocampus (Meyer-Lindenburg et al., 2006). Although the spatial deficit is well-documented, we know little about other deficits that would be predicted by the hippocampal abnormalities. Here, we examine episodic memory (i.e. memory for personally experienced events in a spatio-temporal context, Tulving, 1983), asking people with WS to recount past personal events. We use an interview method developed for patients (Levine et al., 2002) and typically developing children (Willoughby et al., 2012). People with WS recounted significantly fewer episodic details than age-matched controls. Importantly however, they offered just as many semantic details (reflecting general world knowledge), indicating that global factors (e.g., verbal skill or IQ) cannot account for these results. Our work identifies a specific cognitive deficit in WS and further highlights the critical involvement of the hippocampus in episodic memory.

Building bilingual semantic representations based on a corpus-based statistical learning algorithm

Xiaowei Zhao
Emmanuel College

Abstract: In the current study, we applied a corpus-based statistical learning algorithm to derive semantic representations of words under bilingual situations (English and Chinese). The algorithm relies on the analyses of contextual information extracted from a text corpus, specifically, analyses of word co-occurrences in a large-scale electronic database of text. Particularly, we examined how the semantic structure of L2 words can be built based on and influenced by the semantic representations of L1 words in a sequential L2 learning situation. We got the semantic representations under various conditions and the results were processed and illustrated on self-organizing maps, an unsupervised neural network model that projects the statistical structure of the context onto a 2-D space. We further discussed a couple of factors that affected the validity of the representations.

Deconstructing the multi-dimensional Aha! experience

Amory H. Danek
University of Illinois at Chicago, Chicago, IL, USA

Shannon Menard
University of Illinois at Chicago, Chicago, IL, USA

Jennifer Wiley
University of Illinois at Chicago, Chicago, IL, USA

Abstract: The Aha! experience is not a unitary construct, and several different dimensions have been proposed as its constituents. However, a systematic analysis of how much each purported dimension predicts the overall Aha! experience is needed. Presented with a large set of difficult problems (magic tricks), participants were asked to rate their solving experience with regard to suddenness in the emergence of the solution, certainty about the solution, surprise, pleasantness, relief, and drive. The strongest correlations with an overall Aha! rating on correct solutions were found for the dimensions of pleasantness, relief and certainty. Suddenness and drive were correlated to a lesser extent. No significant correlation was found for surprise. These results question the wisdom of the established approach of using a multi-component operational definition for the Aha! experience that encompasses suddenness, certainty and surprise. The positive affect that comes with discovery seems better expressed as pleasantness or relief than surprise.

On Constancy in Spatial Perception

Louise Daoust
University of Pennsylvania, Philadelphia, Pennsylvania

Abstract: The perceptual constancies are at the heart of the scientific and philosophical study of perceptual experience, for they are responsible for our enjoying stable percepts despite fluctuating proximal stimulation. For some time, it has therefore seemed natural to appeal to the constancies as a way of explaining the factivity of perception - how (in veridical cases) we present or represent our environments as they are. Notably, a number of theorists now reject the suggestion that color constancy straightforwardly allows us to track mind-independent physical properties, such as surface spectral reflectances. In the spatial literatures, however, the constancies remain tasked with accounting for the perceptual presentation or representation of objective values as they are independent of perceivers. In this presentation, I outline the unacceptable normative consequences of these latter views, and sketch an alternative, more ecologically plausible understanding of veridicality in spatial perception.

Analogies and Graphics can lead to Illusions of Understanding

Jennifer Wiley
UIC

Thomas Griffin
UIC

Allison Jaeger
Temple University

David Sarmento
UIC

Marta Mielicki
UIC

Abstract: Many people experience illusions of understanding for explanations of scientific phenomena (Rozenbleit & Keil, 2002) and readers tend to be poor at gauging how well they have understood what they have read in expository science texts (Dunlosky & Lipko, 2007; Maki, 1998; Thiede, Griffin, Wiley, & Redford, 2009). The present line of research includes studies demonstrating that metacomprehension accuracy may be especially poor when students are presented with texts that include features such as diagrams, graphs, animations, and analogical examples. Although these adjuncts are meant to improve comprehension, they can often lead to illusions of understanding. An important theme of this research is articulating the kinds of instruction and skills that students may need before they can learn effectively from expository science texts including graphics or analogies.

Auditory N1 Amplitude Varies Across Multiple Acoustic and Phonological Dimensions in Speech

Olivia Pereira
Villanova University, Villanova, Pennsylvania, United States

Joe Toscano
Villanova University, Villanova, PA, United States

Abstract: Listeners are sensitive to numerous fine-grained acoustic cues in speech. However, there has been little work examining how listeners encode these cues at early stages of perception. The event-related potential (ERP) technique provides a tool to help us address this. Previous work shows that the amplitude of the auditory N1 ERP component varies with differences along VOT continua, but it is not clear which other cues show similar effects. We present data examining a large set of minimal pair stimuli spanning 18 consonants. Results reveal widespread differences in N1 amplitude for stops, fricatives, and nasals, including distinctions primarily caused by temporal cues (stop voicing; /b,d,g/ vs. /p,t,k/) and spectral cues (place of articulation; /b,p/ vs. /d,t/ vs. /g,k/). Our results suggest that early speech processing is based on fine-grained acoustic cues, rather than articulatory differences, and that the ERP technique provides a useful tool for measuring speech sound encoding.

Does Contrast or Comparison Help More? The Role of Learning Mode and Category Type

Jan Andrews
Vassar College

Ken Livingston
Vassar College

Calais Larson
Vassar College

Ken Kurtz
Binghamton University

Abstract: Recent work suggests that classification training and observational learning may differ with regard to the benefits of different types of item presentation. In particular, there is evidence that between-category contrast is most helpful for traditional classification learning of feature-based categories, while a supervised observational mode promotes learning of relational categories via within-category comparison. The purpose of this study is to begin to tease apart the role of the learning mode versus the type of category in producing this pattern of results by replicating an earlier study that used classification training and feature-based categories, and adding in observational learning conditions. If under these conditions, contrast is beneficial for both learning modes, it will suggest that the type of category being learned is the key to the previously observed difference. If, on the other hand, that same difference is observed, it will suggest that the cause is the learning mode itself.

Speech Perception Across the Lifespan: Using a Gaussian Mixture Model to Understand Changes in Cue Weighting Between Younger and Older Adults

Sarah Vrabic
Villanova University, Villanova, Pennsylvania, United States

Elke Nordeen
Villanova University, Villanova, Pennsylvania, United States

Joe Toscano
Villanova University, Villanova, Pennsylvania, United States

Abstract: In order to understand speech, listeners must weight and combine multiple acoustic cues. For example, voice onset time (VOT) is a reliable cue to stop consonant voicing, while onset F0 provides information, but is much less reliable. Consequently, we would expect listeners to weight VOT higher than F0. This is the pattern observed for most listeners. However, these cue weights also change over time, and older adults tend to rely less on VOT than young adults, even in listeners without hearing loss. One hypothesized mechanism for this change is a decreased ability to detect temporal differences in sounds, which renders temporal cues (e.g., VOT) less reliable and leads to a greater reliance on spectral information (F0). We simulate this using a weighted Gaussian mixture model and find evidence in support of this mechanism: decreased temporal cue reliability leads to the same pattern of differences observed between younger and older listeners.

Individual Differences in the Acquisition of Strategies in a Complex Task

Edward Cranford
Mississippi State University

Aaron Wong
Mississippi State University

Jaymes Durriseau
Mississippi State University

Kevin Barnes
Mississippi State University

Winston Jones
Mississippi State University

Gary Bradshaw
Mississippi State University

Jarrod Moss
Mississippi State University

Abstract: A multi-session experiment explored the relationship between individual differences and the development of strategies in a complex task environment. In the first session, participants completed measures of working memory and adaptivity. Participants then performed 4.5 hours of a multitasking activity that involved prioritizing, selecting, and sorting objects into bins under time pressure. The analyses reported here focus on how participants prioritized objects in a queue of objects and selected objects from that queue for sorting. Priority selection strategies were automatically extracted using machine learning methods. Differences in strategy use were related to measures of working memory and adaptivity. Strategy use and strategy change mediated the relationship between task performance and individual differences. A hierarchical clustering analysis revealed patterns of strategy shifts that distinguished between participants who improved and those who did not. These results provide a basis for examining strategy training geared toward individuals' cognitive abilities.

Multimodal Dynamics of Explaining the Mechanisms of Global Warming

Alexandra Paxton
University of California, Berkeley

Drew Abney
University of California, Merced

Kristina Castellanos
University of California, Merced

Gilbert Sepulveda
University of California, Merced

Abstract: Human communication is a complex multimodal behavior that is deeply embedded within our environment. From narrative monologues (Dale, 2014) to dyadic task performance (Paxton, Abney, Kello, & Dale, 2014), recent efforts have sought to identify multimodal signatures of different types of communication. We extend these efforts in the current project by investigating the multimodal signatures of learning about a pressing but publicly controversial issue: global warming. Here, we explore how personal political stances and previous scientific understanding affect patterns of multimodal behavior (i.e., language use and gaze patterns) when participants are asked to learn about and then describe the mechanisms behind global warming (Ranney et al., 2013). Quantifying understanding – and exploring how personal traits affect that understanding – is not only vital to better describing communication dynamics overall but may also shed light on emerging efforts to educate the public on important scientific concerns.

Imagining activities: The role of perspective and grammatical aspect

Jeffrey Hong
Wilfrid Laurier University

Todd Ferretti
Wilfrid Laurier University

James Siklos-Whillans
Wilfrid Laurier University

Deanna Hall
Wilfrid Laurier University

Abstract: The ability to imagine events is important to regular thought processes such as remembering and understanding the world in general. Two EEG experiments were conducted to investigate the difficulty associated with imagining activities from different visual perspectives. Experiment 1 involved participants imagining ongoing activities (e.g., I was skating) from a first and third person perspective. Experiment 2 involved completed activities (I skated) and also included a condition in which participants imagined other people from a third person perspective (Karen skated). Slow cortical brain potentials revealed that the third-person perspective was generally the most difficult to imagine and that the third-person-self perspective was more difficult than the third-person-other perspective. Imagining activities as ongoing or completed did not influence the pattern of results. This research provides novel neurocognitive and behavioural insight into how event representation is influenced by temporal information associated with verbs and the perspective from which an event is represented.

The role of higher order relational structure in relational category label extension

Sean Snoddy
Binghamton University, Binghamton, NY, United States

Kenneth J. Kurtz
Binghamton University, Binghamton, NY, United States

Abstract: The perceived soundness of an analogy is influenced by shared relational structure between the analogs with higher-order relational (HOR) structure being the primary determinant (Gentner, Rattermann, & Forbus, 1993). We conducted a replication and extension to investigate whether the same pattern holds when deciding whether to extend a relational category label from a base example to a target. Participants were assigned to judge either category extension or analogical soundness (using a more direct version of the original measure) across four targets that shared HOR structure, surface similarity, neither, or both (literal similarity) with the base passage. We found that shared HOR structure led to a higher likelihood of both extending the category label and judging the analogy to be good. No effect of surface similarity was found. These results suggest that the generalization of relational categories follows the same principles of structure-mapping theory that are seen in analogical processing.

Memory for the Meaningless: Experts' Advantage at Recalling Unstructured Material

Giovanni Sala
University of Liverpool, UK

Fernand Gobet
University of Liverpool, UK

Abstract: The ability to recall domain-specific unstructured material (e.g., random chess positions) is a litmus test for theories of expert memory. Theories emphasising high-level memory structures or holistic processing of stimuli predict no difference between experts and novices at recalling unstructured material, because no large structure or whole are present in such material. Conversely, theories assuming small memory structures (e.g., chunks) predict a skill effect, because even in scrambled material some small meaningful structures occur by chance. This meta-analysis assessed the correlation between expertise and recall of unstructured material in several domains, including board games, programming, sports, and music. We found a moderate but significant overall correlation ($r = .42$, $p < .001$), and the presence of an effect in nearly every domain. This outcome suggests that experts base their superiority on a vaster knowledge of small memory structures, in addition to high-level structures or holistic processing.

“He will try to learn it because he doesn’t know it.” Young children’s understanding of learning based on their knowledge states

Jeein Jeong
University of Pennsylvania, Philadelphia, PA, United States

Douglas Frye
University of Pennsylvania, Philadelphia, PA, United States

Abstract: If we already know how to tie our shoelaces, it should not be necessary to learn again. When somebody shows you how to tie them, if you already know how, you may not regard the person as a source of knowledge. Do preschoolers understand the role of learner’s knowledge states in learning the same way? The current study, with seventy-two 3- to 5-year-olds, tested preschoolers’ understanding of learning. Children listened to three teaching stories that a peer tries to teach a knowledgeable, neutral, or ignorant child something, and three not-teaching stories that a knowledgeable, neutral or ignorant child accidently sees the peer do that same thing. We asked if the child would try to learn from the peer, and whether s/he really learned the knowledge from the peer. Results showed an age change in understanding of learning intention and source of knowledge. Relevance to children’s theory of mind is discussed.

The Relationship Between Mental-state Language and False-belief Understanding in Adulthood

Erin Roby
University of California Merced, Merced, California, United States

Alexandra Paxton
University of California, Berkeley, Berkeley, CA, USA

Rick Dale
University of California, Merced

Rose Scott
University of California Merced, Merced, CA, United States

Abstract: Research has revealed a robust relationship between preschooler's use of mental-state language (e.g. think, know) and performance on false-belief tasks (e.g. Ruffman, Slade & Crow, 2002). However, investigations of this relationship with school-aged children have shown mixed results, making it unclear whether mental-state talk continues to play a role in false-belief understanding following the preschool years (e.g. Charman & Shmueli-Goetz, 1998; Grazzini & Ornaghi, 2012). This discrepancy may result from the fact that preschooler's talk has consistently been assessed during interpersonal interactions with peers, siblings, and parents, while school-aged children's talk has been assessed via descriptions of wordless picture books or absent friends. The present study bridges this gap by exploring whether adults' use of mental-state language during interaction correlates with their false-belief performance. In doing so, we help to shed light on an important issue in theoretical accounts of the development of false-belief understanding.

The use of dispositional cues to causality in judgements of mechanical and living interactions

Julia Schneider
Departement of Psychology, University of Berne, Switzerland

Sufi Abbaspour
Departement of Psychology, University of Berne, Switzerland

Trix Cacchione
Departement of Psychology, University of Berne, Switzerland

Abstract: White (2013) stated that dispositional causal thinking derives from experiences of acting on objects aquired early in life. He made evident that, under uncertainty, particular cues in an interaction between an agent and a patient (e.g., two entities, agent focuses on patient, contact, effect in patient) guide people's perception of causality. This study systematically examines the predictive strength of eight causal cues worked out by White (2013) and aims at comparing people's reliance on these cues in the physical and the biological domain.

Children (7-year-old) and adults judged a prototype (mechanical collision event or stinging event) and another nine prototype related events, with systematically omitted cues.

A general linear mixed models analysis revealed a significant effect for the number of cues in an event. Both age groups rely on singular causal cues when interpreting physical and biological events. Moreover, the disposition of causality appears to harden with increasing age.

Learning Hierarchical Labels through Cross-situational Learning

Chen Yu
Indiana University

Chi-hsin Chen
Indiana University Bloomington

Abstract: An increasing body of research has demonstrated that human learners are able to use co-occurrences among words and objects to form word-object associations (e.g., Yu & Smith, 2007). In this study, we further investigated learners' ability to use statistical information to learn labels at different hierarchical levels. Participants were presented with objects and words in ambiguous learning trials. In some learning trials, participants saw multiple objects and heard their individual labels presented in a random order, while in other trials, category labels were presented instead. Results from three experiments provided converging evidence that adults were able to use word-object co-occurrences across different situations to learn hierarchical labels. Moreover, participants generalized category labels to novel members at the same level but not to superordinate-level instances. There was also an interaction between the level of ambiguity in learning contexts and performance in label learning and generalization.

Intensional Probability Judgments and Inclusion Fallacies With Generics

Momme von Sydow
University of Munich, Munich, Germany

Abstract: The discussion of conjunction fallacies or, more generally, inclusion fallacies (IFs) is usually limited to dyadic relationships. Bayesian logic formalizes a rational intensional probability, predicting IFs and supplementing standard extensional probabilities (von Sydow, 2011, 2016). They treat logical patterns as explanatory patterns (explanans) given some data (the explanandum). We here address the even more basic issue of nested hypotheses in a single polytomous dimension (von Sydow, 2015) and present a corresponding variant of Bayesian logic (BL). The experiments use materials from the Linda tasks (one concerned with jobs, the other with political attitude) and they explore the polysemous character of 'AND' (Hertwig, Benz & Krauss, 2008; von Sydow, 2014). BL stresses that pattern probabilities should depend on the representation of subclasses. As predicted, the results show substantial deviations from standard probability and here corroborate a pattern approach. They are also at odds with a confirmation account.

Motor cortex excitability during processing of handwritten and typed non-action-related text

Chelsea Gordon
University of California, Merced, Merced, CA, United States

Ramesh Balasubramaniam
University of California, Merced, Merced, CA, United States

Michael Spivey
University of California, Merced, Merced, CA, United States

Abstract: Motor cortex has been found to play a crucial role in processing the semantics of spoken and written action-related language as well as in early speech perception. One possibility is that the motor system is always involved in perception and cognition, picking up any available motor information in the environment. If this is true we should see increased corticospinal excitability when subjects are looking at anything that affords motor behaviors or possible simulation of motor behaviors. We used Transcranial Magnetic Stimulation and electromyography to investigate corticospinal excitability while participants read handwritten or typed words and non-words from a computer screen. Results show that for typed words, there is an increase in excitability for words compared to non-words, while the reverse is true for handwritten words. We discuss implications for the possible role of the motor system in early language perception in different contexts.

In the nick of time: Using temporal cues to examine ongoing event representations

Victoria Scharp
University of Pittsburgh , Pittsburgh, PA, United States

Connie Tompkins
University of Pittsburgh, Pittsburgh, PA, United States

Michael Dickey
University of Pittsburgh, Pittsburgh, PA, United States

Abstract: Time cues are ubiquitous in language and the ability to interpret them is essential for understanding events during discourse comprehension. Temporal markers that signal ongoing versus completed events, like the progressive and simple past tense, prompt distinct mental event representations. However, the detailed properties of ongoing event representations remain unexplored. Drawing from both the simulation and semantic association approaches to knowledge representation, this study examines the novel prediction that ongoing events engender incremental discourse representation updating processes. Experimental sentences cued either early or late phases of an ongoing event (e.g. Alice had recently started/almost finished baking a cake). Targets in a post-sentential lexical decision task were strongly associated with either early or late event phases (e.g. EGGS/AROMA). Facilitation priming was predicted for congruent sentence-target pairs. Implications of the results for models of knowledge representation, theories of semantic priming, and discourse model updating will be discussed.

Toward a Simulation Platform for Comparing Computational Cognitive Neuroscience Models

David Noelle
University of California, Merced

Abstract: While computational cognitive models serve many purposes, perhaps their primary utility is in formalizing specific hypotheses in order to facilitate evaluation in light of empirical results. Such evaluations are inherently relative, comparing the explanatory power of proposed models to alternatives. Direct comparisons are hindered, however, when competing hypotheses are framed within different cognitive architectures, as the contributions of non-focal aspects of those architectures cannot necessarily be yoked. In order to help address this problem, a novel computational framework for model comparison is proposed, grounded in gross neuroanatomy. This framework supports the hierarchical specification of connections between brain systems, producing computational architectures based on neuroscientific data. This approach shifts from modeling particular cognitive processes, which might differ across cognitive architectures, to modeling established brain systems, for which there may be greater consensus. The framework supports the direct comparison of models of a given system by fixing the function of other systems.

Consequences of bilingualism for perceptions of categories and similarity

Kevin Holmes
Colorado College

Jacob Brodsky
Colorado College

Abstract: Speakers of two languages have access to two semantic systems that, while largely similar, may differ in subtle ways. The existence of multiple similar systems offers the potential for comparison of their structures and discovery of the differences between them. We hypothesized that if bilinguals engage in such a comparison process, they may be (a) less likely than monolinguals to view the categories of any single language as natural kinds, and (b) more likely than monolinguals to discern differences among high-similarity items more generally. Monolingual and bilingual participants indicated their level of agreement with statements equating social categories with natural kinds and judged the similarity of pairs of perceptual relations. Compared to monolinguals, bilinguals were less willing to endorse naturalness statements and showed more variability in their similarity judgments. These results suggest that bilingualism may promote sensitivity to differences among highly similar stimuli, linguistic and otherwise.

Mandarin-English Bilinguals Match Lexical-Tone Processing to the Language Context

Carolyn Quam
University of Arizona

Sarah Creel
University of California, San Diego

Abstract: Proficient bilingual listening requires differential processing of sound variation in each language context. We considered context-based processing of pitch information by Mandarin-English bilinguals, for whom pitch indicates word distinctions in one language but not the other. In an eye-tracked word-learning experiment, 58 bilinguals and 28 English monolinguals each learned English-like and Mandarin-like wordsets, words referring to images. Wordsets differed primarily in that English-like words contained final consonants. We explained that some words might differ only in their pitch patterns, and included training on minimal tone pairs. In test, two pictures appeared on the screen with referents differing in either tone or vowel. One picture was labeled. Bilinguals processed tones more efficiently ($t(78) = 3.54$, $p = .001$) and more accurately ($t(84) = 3.78$, $p < .001$) than monolinguals only in the Mandarin context. Mandarin-English bilinguals thus appear to tailor tone processing to the within-word language context.

Modifying Cognitive Load Component Survey for K-12 STEM Testing

Jingyuan Huang
Educational Testing Service, Princeton, NJ, USA

Janet Koster van Groos
Educational Testing Service, Princeton, NJ, USA

Michelle Lamar
Educational Testing Service, Princeton, NJ, USA

Abstract: Test-taker's capability to answer questions is influenced by available cognitive resources for problem solving. Due to the limited working memory capacity, excessive cognitive load for interpreting instruction would impact test-taker's construct-relevant process and test validity. Especially in STEM assessment where multimedia and interactive design are widely used, test-takers can easily get overwhelmed by a large amount of visual or audio information. Testing materials should be designed to minimize the unnecessary cognitive load in order to increase cognitive resources for problem solving in the task. The Cognitive Load Component Survey is one of the first self-report measurements distinguishing different types of cognitive load: intrinsic cognitive load, extraneous cognitive load, and germane cognitive load. We report modifications of this survey to fit into K-12 educational assessment, results of measuring cognitive loads in a simulation-rich science assessment, and implications to use this survey for future assessment development.

The Development of Intuitions about the Controllability of Thoughts, Emotions, and Behavior

Brittany Klimek
Lehigh University, Bethlehem, PA, United States

Amanda Brandone
Lehigh University, Bethlehem, PA, United States

Abstract: From early in development children show impressive knowledge about mental states such as beliefs and desires. However, less is known about the development of knowledge about more sophisticated aspects of mental activity, including the adult intuition that the mind is an independent agent over which we have some but not total control. This project explored 8- to 11-year-olds' (n = 46) and adults' (n = 48) beliefs about the extent to which thoughts, emotions, and behaviors are controllable. Results indicated that both children and adults viewed thoughts and emotions (in contrast to behavior) as relatively involuntary. Children and adults also generally rejected the notion that mental activities and behaviors are chronic. However, while adults were skeptical about whether people can stop their own thoughts, emotions, and behavior, children fully endorsed this type of control. Overall, data suggest that intuitions about the controllability of mental activities continue to mature throughout childhood.

A shape-heavy vocabulary does not a shape bias make: A comparison of the content of English-learning children's and Spanish-learning children's typical vocabularies

Emily Russell
CSUN, Los Angeles, California

Christina Schonberg
University of California, Los Angeles, Los Angeles, CA, United States

Shawntel Barreiro
CSUN, Los Angeles, California

Abstract: We asked why Spanish-monolingual children exhibit a weaker, slower-to-develop shape bias in word-learning contexts compared to English-monolingual children (Hahn & Cantrell, 2012). Ten English-monolingual adults and nine English-Spanish bilingual adults rated the perceptual similarity of items indicated by subsets of words from the English MCDI and Spanish MCDI, respectively. Consistent with previous research with similar methodology (Samuelson & Smith, 1999), words for shape-similar items predominated in the content of the English MCDI (47.72%; agreement: 70%, $p < .05$). Interestingly, words for shape-similar items also predominated in the content of the Spanish MCDI (56.67%; agreement: 70%, $p < .05$). Results suggest that the types of words that children learn play a less important role in the development of the shape bias than other proposed factors (e.g., syntactical regularities; Smith, 2000). Additional findings and implications for children with various language backgrounds will be discussed.

Motion Capture of Phase Change Transitions During Insight Problem Solving

John Hart
Arizona State University

Chelsea Johnson
Arizona State University

Nicholas Duran
Arizona State University

Abstract: Insight problem solving refers to the phenomenon of experiencing a sudden flash of insight when discovering novel problem solving strategies. This sudden transition in thinking suggests a phase change in human cognition as an emergent property of the self-organizing complex system of coupled neural activations. In our study, we developed a method of measuring this phase change within an embodied cognition paradigm. We used 3D motion capture to measure the precise body movements of 21 participants at 120 Hz resolution while they solved 3 different types of insight problems. We analyzed a sliding time series of postural sway and head displacement using recurrence quantification and spectral analyses to determine changes in entropy in the participants' movements. These measures allow us to make inferences about changes in level of self-organization as the participant's neural activation transitions from one state to another.

Exploring the Cognitive and Social Profile of Science Rejection

Emilio Lobato
Illinois State University

Corinne Zimmerman
Illinois State University

Steve Croker
Illinois State University

Abstract: We examined the degree to which cognitive style, cultural worldview, conspiracy ideation, and religious and political demographic variables correspond to agreement with scientific claims across four domains. Participants rated their level of agreement with scientific statements in four domains (evolution, GMOs, vaccinations, climate change) along with open-ended questions to investigate participants' reasons for their support or rejection, filled out individual difference measures, and completed a demographics questionnaire asking about frequency of attendance at religious services as a proxy for religiosity, and political ideology along the liberal-conservative spectrum. Lower agreement with scientific statements was found to be related to a lower analytic thinking style and a stronger conservative political ideology. Our results contribute to a better understanding of the cognitive and social profiles of individuals who reject scientific conclusions and can be useful in designing future research efforts aimed at investigating science acceptance and science denial.

Technological Shaping of Verbal Working Memory: A Difference between Chinese Phonology-Based and Orthography-Based Typing

Jenn-Yeu Chen
National Taiwan Normal University, Taipei, Taiwan, ROC

Adam Li
National Taiwan Normal University, Taipei, Taiwan, ROC

Abstract: Typing Chinese words on a computer can be carried out with a phonology-based or an orthography-based method. Phonological typing constantly engages typists' verbal working memory (VWM), while orthographic typing engages their visual-spatial working memory (VSWM). Accordingly, habitual phonological typists would develop a better VWM capacity, while habitual orthographic typists would have a better VSWM capacity. Five VWM tests and five VSWM tests were administered to 24 phonological typists and 23 orthographic typists. The results showed that the phonological typists scored higher than the orthographic typists on the VWM tests, but no significant differences on the VSWM scores were observed. The latter result is attributed to the notoriously abundance of homophones in Chinese, which forces the phonological typists to keep attending to the orthographic forms of the characters being typed. Our findings suggest that individual cognitive systems develop and adapt flexibly, subject to shaping by technology within a life's time.

Adults' drawing and recognition of familiar objects and substances: Nonsolids are hard to identify

Lynn Perry
University of Miami, Coral Gables, Florida, United States

Abstract: In English, categories of solid objects (e.g., couch) are similar in shape, but vary in color and material; categories of nonsolid substances (e.g., yogurt) are similar in material, but vary in color and shape (Samuelson & Smith, 1999). Although even infants can discriminate between how solids and nonsolids should behave (Hespos et al., 2009), increasing evidence suggests recognizing specific substances is difficult for children (Perry et al., 2014). This begs the question, what do adults even know about nonsolids? Twenty adults drew 23 familiar solids and nonsolids. 116 participants from Amazon Mechanical Turk attempted to identify each drawing. Participants more accurately identified drawings of solids (M=.70) than nonsolids (M=.25), X2(1)=13.87, p=.0002. Drawings of nonsolids leading to accurate identification often depicted prototypical containers (e.g., milk carton). These results suggest visual recognition—even of nonsolids—is aided by shape and that adults may conceptualize nonsolids as more object-like than was previously thought.

How Spatial Ability and Stress Impact Escape Path

Benjamin Nelligan
University of Notre Dame; Johns Hopkins University

Alexandra Amorati
Carnegie Mellon University

Deanne Adams
Microsoft

Christopher Galeucia
University of Notre Dame

Laura Carlson
University of Notre Dame

Abstract: Individual differences and situational factors can both affect how and how well one navigates. This study examined the effects of stress and spatial ability, measured as mental rotation ability, on navigation during an emergency situation. Participants learned a virtual mall environment and were subsequently either told to meet a friend at the far exit (control) or to use the far exit to escape a fire. In an emergency, participants made an initial movement faster, made more errors during navigation, and overestimated the amount of time they took to exit relative to controls. Relative to controls, emergency low spatial participants more often reversed a learned path to exit the mall, whereas high spatial participants more often directly used a previously learned path. The results illustrate that stress from an emergency situation negatively impacts navigation, and that the behavioral consequences of this are in part dependent upon one's spatial abilities.

Rapid emotion discrimination in the infant brain

Alexandra Marquis
Ryerson University

Bruno Rossion
University of Louvain (UCL)

Margaret Moulson
Ryerson University

Abstract: The ability to recognize facial expressions of emotion in social partners is important for successful social interactions. It is unknown how accurately and rapidly the infant brain discriminates between emotions with different valences (e.g., happy vs. fearful) and between emotions with similar valences (e.g., fearful vs. angry). The current study uses a novel approach—Fast Periodic Visual Stimulation (FPVS)—to evaluate emotion discrimination in infancy. FPVS is an electrophysiological technique that relies on rapid presentation of stimuli to create corresponding oscillations in the brain that can be measured at the scalp surface. Preliminary results (n = 6) indicate that infants are indeed sensitive to the visual stimulation: EEG power, averaged over occipital and occipitotemporal areas, was 11.55 times larger at 6Hz compared to surrounding frequencies. This study aims to shed light on a longstanding theoretical debate of whether emotion recognition is innate or learned through experience.

Mental representations and processing of radical expressions

Purav Patel
University of Minnesota - Twin Cities

Sashank Varma
University of Minnesota - Twin Cities

Abstract: Mathematical cognition researchers have studied the mental representations of natural numbers, integers, and fractions extensively. We investigated the representations of irrational and perfect square numbers in a laboratory setting. Eighty participants performed (1) a magnitude comparison task (MC) by indicating which of two numbers is greater or lesser, (2) a number line estimation task (NLE) that required subjects to estimate the positions of natural and radical numbers on a number line, and (3) a numeracy test. On the MC task, participants were slower for radical expressions than for natural numbers and showed distance and size effects for both. When comparing radical expressions, they were faster when both numbers were perfect squares. This suggests a privileged mental representation for perfect squares. On the NLE task, participants were less accurate when locating radical expressions. Performance on the numeracy test revealed broad deficits in conceptual and procedural knowledge of irrational numbers.

Data Shows Human Behavior is Not Random, period.

William G. Kennedy
George Mason University

Abstract: Simulations of many people's decisions are used in public health and safety as well as to support policymaking. These simulations rely on creditable models of individual decision-making. An obvious approach is to develop a list of plausible actions and to then evaluate the benefits of each in the current situation to make the decision. However, such evaluations can be implausible, e.g., zero-intelligence traders in economics, or impracticable because the approach is computationally intensive for large-scale simulations. As a result, a commonly used approach is to select randomly from the plausible actions. Without data on how people would actually chose, a random number from a uniform distribution over the plausible options is often used to represent the unknown cognition. However, we claim that substituting a uniform random distribution for how people make decisions is making very strong claims about the process and we will present data demonstrating it is simply wrong.

Feedback Markers in Mandarin: Tracking Cognitive Status in Conversation

Li-chiung Yang
Tunghai University, Taichung, Taiwan

Abstract: A key aspect of conversation is the interactive exchange of information and the cooperative discourse process that functions to bring about a mutually satisfactory sharing of information. The encountering of new information gives rise to emotional responses as well as on the certainty and degree of cognitive reorientation to the pre-existing knowledge state. In this study we present our results on prosody and contextual meaning of two feedback markers: "dui" and "oh", showing that the marker "dui" 'right', functions to mark the perceived status of lexical truth and degree of understanding, approval, or agreement, thus acting as an indicator of cognitive understanding to organize topic development through signals of a shared knowledge state, whereas the marker "oh" acts as an indicator of temporary cognitive difficulty or reorientation of cognitive states, thus providing the complementary function to "dui" with respect to certainty and uncertainty of information and knowledge states in discourse.

The Semantic Stroop Effect: An Ex-Gaussian Analysis

Darcy White
University of Waterloo, Waterloo, Ontario, Canada

Evan Risko
University of Waterloo

Derek Besner
University of Waterloo

Abstract: The standard Stroop effect (which typically uses color words that form part of the response set) is robust and well documented in mean RT. Ex-Gaussian analyses reveal that this effect is seen in the mean of the normal distribution (mu), in the standard deviation of the normal distribution (sigma), and (c) in the tail (tau) of the ex-Gaussian distribution. The present experiments investigate whether the semantically based Stroop effect (which contrasts incongruent color-associated words with neutral controls) is seen in the three ex-Gaussian parameters. This analysis yielded a semantic Stroop effect in the arithmetic mean and mu, but no semantic Stroop effect was observed in tau. These data are consistent with the conclusion that interference associated with response competition on incongruent trials is absent in the semantic Stroop effect (at least in the tail of the distribution).

Investigating the Effect of Experience on Concrete and Abstract Word Processing

Selgun Yuceil
Middle East Technical University, ANKARA, Turkey

Didem Gokcay
Middle East Technical University, ANKARA, Turkey

Abstract: As shown in previous studies, semantic processing of words is mainly affected by frequency, context and concreteness. Recently sensory, motor and emotional components are also examined to explain the concreteness effect within the embodiment framework. Concreteness effect, which adopts the processing advantages of concrete words over abstract ones, is supported by studies, which show better remembering, faster processing and faster recognizing performances for concrete words. In this study, concreteness effect was examined via two experiments on experts and controls in which verbal fluency and lexical decision tasks were employed. Lawyers were considered as an expert group with their intense deal with abstract concepts. A novice lawyer group and age-matched participants other than lawyers were used as control groups. Results showed that concreteness effect disappeared in the expert group as a matter of expert's verbal experience.

A cognitive model of online event segmentation

Anthony Harrison
Naval Research Laboratory

Sangeet Khemlani
Naval Research Laboratory

Greg Trafton
Naval Research Laboratory

Abstract: People automatically segment online perceptual and conceptual experiences into events (Newston, 1973). A new model-based theory explains how people construct temporal markers and prioritize those changes to build representations of events (Khemlani et al., 2015). The theory is implemented within an embodied extension of the ACT-R cognitive architecture (Anderson, 2007) called ACT-R/E (Trafton et al., 2013). Its principal parameter is the prioritization scheme by which certain detectable changes (e.g., in a perceived location) are preferred over others (e.g., in perceived states of an object). We tested the predictions of the theory and its computational model against an experiment on narrative event segmentation. Participants in the study read an excerpt of text and were asked to assess whether certain lines marked the start of a new event. The computational model readily accounted for their segmentation behavior. We conclude by discussing event segmentation and its relation to embodied cognition and cognitive robotics.

Human-object interaction understanding without objects

Anna Belardinelli
University of Tübingen

Johannes Lohmann
University of Tübingen

Martin V. Butz
University of Tübingen

Abstract: During object manipulation the actor's eye movements are directed to the target of the interaction and to the relevant sites where this takes place. Eye movements during grasping observation are influenced by low-level motor information, helping inferring the target from hand shape. In an eye-tracking experiment, we investigated which factors influence understanding when observing bimanual object interactions, if no objects are visible but only the movements reproduced by an avatar. Participants watched ten different actions (e.g., pour water from a bottle into a cup) and guessed among ten possibilities. Also perspective was varied (frontal, side, head-centered). Preliminary results show higher response accuracy in the frontal perspective. During the interaction phase participants spent more time fixating closer to the interaction point between the hands, where the objects would be, than on the single hands, suggesting this is the best vantage point to make sense of the observed action without other cues.

The Effect of Book Design on Beginning Readers' Attention Allocation

Karrie Godwin
Carnegie Mellon University

Anna Fisher
Carnegie Mellon University

Abstract: Books for beginning readers typically intermix text and pictures in close proximity. The proximity of difficult-to-decode text to pictures may induce competition between these sources of information. As a result, children may frequently shift gaze between text and pictures, which may degrade memory representations of the text and reduce comprehension. A mobile eye tracker was used to measure children's attention allocation while reading commercially available books for beginning readers. Preliminary evidence suggests that pictures capture young children's (N=12, Mage=7.14 years) attention while they are engaged in guided reading. Even when the text was short (on average 6.94 words per page), children frequently shifted their attention between text and pictures. Per page, children made on average 2.80 alterations from text to pictures (Range: 0.93 to 6.57 alterations). These findings are consistent with the hypothesis that the close proximity of text and pictures may result in competition between these sources of information.

Generating Predictions Non-consciously: Evidence from Invisible Motion with and without Obstacles

Ariel Goldstein
Hebrew university, Jerusalem, Israel

Ido Rivlin
Hebrew university, Jerusalem, Israel

Ran Hassin
Hebrew university, Jerusalem, Israel

Abstract: Previous research has established that conscious stimuli can lead to non-conscious predictions. Yet, it also suggests that conscious awareness of stimuli is a necessary condition for using them in predictions. We use subliminal movement – with and without obstacles – to examine predictions from subliminal stimuli. In four experiments, a moving object was masked with continuous flash suppression. After the object had stopped moving, a conscious probe appeared in a location that was either consistent with the movement or not. In the first three experiments the movement was linear, and non-conscious predictions were based on both direction and speed of movement. In Experiment 4, the moving object collided with an obstacle. Response times revealed predictions on the deflection route. We thus conclude that humans can use dynamic subliminal information to generate active predictions about the future

Using analogical comparison to help children learn the day-night cycle

Benjamin Jee
Worcester State University

Florencia Anggoro
College of the Holy Cross

Natalie Evans
College of the Holy Cross

Caitlin Murphy
College of the Holy Cross

Jessica Tran
College of the Holy Cross

Caroline Morano
College of the Holy Cross

Amanda McCarthy
College of the Holy Cross

Victoria Jackson
College of the Holy Cross

Abstract: Children have difficulty reconciling their observations of the sky (an Earth-based perspective) with scientific models of the solar system (space-based perspectives) (e.g., Vosniadou & Brewer, 1994). Analogical comparison could be an effective way to address this cognitive challenge. By comparing and aligning different perspectives on events, such as sunrise, children may develop a more coherent understanding of the solar system. The present experiment tested this theory by varying the presence of explicit comparisons between Earth-based and space-based perspectives during a multi-day lesson about the day-night cycle. Children (N=63, Mean age=8.57) were randomly assigned to one of four learning conditions: one that involved guided comparison of perspectives, two that involved similar tasks but without comparison, or a control (no instruction) condition. We found that children in the guided comparison condition had the greatest learning gains on a task that involved demonstrating the day-night cycle using a model Earth and Sun.

A Preliminary Model of Situation Awareness in a Cognitive Architecture.

Ehsan Khosroshahi
Drexel University, Philadelphia, Pennsylvania, United States

Dario Salvucci
Drexel University, Philadelphia, PA, USA

Abstract: Although maintaining situation awareness (SA) is a critical skill for many complex tasks, there have thus far been few rigorous computational approaches to modeling SA behavior and performance. We developed a preliminary computational model of SA focused on remembering the locations of static objects in the visual field. We built this model on the foundation of the ACT-R cognitive architecture, using its declarative memory and vision modules to specify the process of scanning the field and remembering object locations. In the current work, we demonstrate how this model accounts for human behavior and performance in two recent experiments: one, a study of object location memory with identities similar to air-traffic control call signs; and another, a study of remembering the location of shapes of varying size, color, and pattern.

Verifying who ”jumped more” or ”higher” in simple events

Haley Farkas
Northwestern University

Alexis Wellwood
Northwestern University

Abstract: Sentences with “more” can be used to compare along many different dimensions (e.g., number, height, etc.). Barner, Wagner, and Snedeker (2008) found that participants strongly preferred number as the relevant dimension for comparatives with deverbal nominals like “more jumping,” even when height was an available choice. Would this preference manifest as a choice of number over height pitting the two dimensions against one another with verbal “jumped more”? We animated two objects, A and B, and varied each’s height, duration, and number of jumps, counterbalancing how often A “won” along each dimension. In separate blocks, participants judged whether “A jumped higher/longer/more times/more than B,” and unambiguously chose height with “higher,” duration with “longer,” and number with “more times.” With bare “more,” however, participants said “yes” both by height and number. This study challenges Barner et al’s (2008) idea that the lexical root “jump” determines a comparison by number with “more”.

Language influences attention to Japanese event components in native English-speaking 21- to 24-month-olds

Natalie Brezack
University of Delaware

Haruka Konishi
Michigan State University

Roberta Golinkoff
University of Delaware

Kathryn Hirsh-Pasek
Temple University

Abstract: Japanese and American 13-15-month-old infants distinguish between crossing a bounded ground (e.g., a street) versus an unbounded ground (e.g., a field). In English, the same verb – crossing – expresses both types. While language has been hypothesized to guide infants' progression from language-general to language-specific event perception (Göksun et al., 2011), no prior studies examined this hypothesis. We presented toddlers who no longer perceive this Japanese distinction in events with novel spatial prepositions (N = 24) or nonlinguistic tones (N = 12) to label bounded versus unbounded grounds. Children presented with labels, but not tones, attended to the differences in ground categories by looking significantly longer to the novel ground type at test. This suggests that above and beyond the attention-getting function associated with non-linguistic auditory stimuli, language uniquely facilitates categorization of event components.

Comparing predictions of lexical norm data obtained using word associations and word collocation

Hendrik Vankrunkelsven
University of Leuven, Leuven, Belgium

Steven Verheyen
University of Leuven, Leuven, Belgium

Simon De Deyne
University of Adelaide, Adelaide, SA, Australia

Gert Storms
University of Leuven, Leuven, Belgium

Abstract: We compared the quality of prediction of word variables based on a Dutch word association and text corpus. We derived estimates for: valence, arousal, dominance, concreteness and age of acquisition (AoA) for 2831 words. Based on the similarity between words we: (1) used projections on a dimension identified as the variable in question in a multidimensional representation, (2) used the k-nearest neighbors values, weighted according to their proximity. Estimates prevailed when based on word associations. Differences between the predictions of the two methods were small. Based on the word association corpus it yielded correlations of .92, .85, and .85, for valence, arousal, and dominance, respectively. Its corresponding correlations based on the text corpus were .80, .74, and .67. For concreteness and AoA, both the association and the text corpus yielded correlations of .88 and .73, respectively. This suggests word associations are better at capturing human ratings of affective word variables.

Does GIScience Training Enhance Spatial Navigation Ability?

Annisa Ahmed
Temple University, Philadelphia, Pennsylvania, United States

Alina Nazareth
Temple University, Philadelphia, Pennsylvania, United States

Danielle Burnetta
Temple University, Philadelphia, Pennsylvania, United States

Magda Abozeed
Temple University, Philadelphia, Pennsylvania, United States

Nora Newcombe
Temple University, Philadelphia, Pennsylvania, United States

Abstract: Research on the reciprocal influence of spatial thinking and GIScience training is limited (Wakabayashi & Ishikawa, 2011). In the current project, we examine improvement in spatial navigation in undergraduates enrolled in GIS classes over the course of a semester. Students enrolled in strategic communications (SC), a low-spatial content class, were used as a control group. Fifty students were trained and tested at 2 time points – beginning and end of a semester – in a virtual navigation task (Silcton; Weisberg, et. al., 2014). We hypothesize that a significantly higher number of GIS students as compared to SC students will be Integrators i.e. they will find the most targets within-route and between-route in the virtual environment. Furthermore, we hypothesize that GIS students will show a significantly greater improvement at time point 2 as compared to SC students. This research has important implications for spatial training and educational pedagogy in STEM disciplines.

Plasticity of Categorization: Developmental Differences in Category Learning and Transfer between Children and Adults

Sophia Deng
The Ohio State University

Vladimir Sloutsky
The Ohio State University

Abstract: How do people learn categories and transfer learning? This study addressed this question by examining the role of attention in the development of category learning and transfer. Participants (adults and 4-year-olds) were trained with two categories including deterministic and probabilistic features and their attention was directed to either type of features. After training, participants learned two new categories and their categorization and memory for exemplars were tested. Results indicated that adults and 4-year-olds were able to be trained to use either a similarity-based or rule-based strategy. However, adults failed to transfer and went back to their default rule-based strategy in novel situations, whereas 4-year-olds transferred the learned strategy. Furthermore, in contrast to adults exhibiting better memory for features used in categorization, 4-year-olds remembered multiple features well regardless of categorization. These results have important implications for understanding the development of categorization and the role of attention in cognitive development.

Unsupervised learning of VerbNet argument structure

Jesse Mu
Boston College

Timothy O'Donnell
Massachusetts Institute of Technology

Joshua Hartshorne
Boston College

Abstract: The relationship between a verb and the syntactic frames in which it can appear has been closely studied by psychologists and linguists. Research suggests that the semantics of a verb and its arguments determine the verb's syntactic frames, but various theories (Levin & Hovav, 2005) disagree on the nature and complexity of these relationships, in part because most investigations have focused on a small subset of verbs that may not generalize. Investigating the semantic and syntactic relationships present in larger sets of verbs would provide more substantial evidence for evaluating and selecting theories of verb argument structure. We report on initial analyses of the 6000+ verbs and 280+ syntactic frames of VerbNet (Kipper et al., 2008), the largest English verb syntax resource available, using nonparametric Bayesian methods (e.g. Shafto et al., 2006) for cluster analysis and dimensionality reduction.

Moral Social Media: Heavy Facebook Users Accept Harsher Moral Criticism for Microaggressions

Boyoung Kim
Brown University

Joanna Korman
Brown University

Bertram F. Malle
Brown University

Abstract: Microaggressions are unintentional or thoughtless behaviors that convey negative messages to members of minority groups. Due to the attributional ambiguity of microaggressions, people often differ in their judgments about how morally bad acts of microaggression are. To account for this individual variation, we explored the potential influence of heavy social media use on individuals' moral judgments of microaggressive behaviors. We hypothesized that, because of the relative acceptance of strong blame expressions on social media, heavy Facebook users would endorse intense moral criticism of microaggressions. Participants read about several agents who committed microaggressions and judged the appropriateness of moral criticism (pretested to vary in degrees from "disapprove of," "chastise," "chew out," to "lash out at") that a friend directed towards the offender. Consistent with our hypothesis, we found a strong correlation ($r = .47$) between increasing degrees of Facebook use and increasing acceptance of harsher moral criticism for microaggressions.

Memory-based decision making: Examining the relative influence experimental and pre-experimental exposure.

Justin Olds
University of Lausanne

Julian Marewski
University of Lausanne

Abstract: We examine the role of memory accessibility across two different memory-related judgments: episodic recognition (e.g., "Was this person's name presented earlier in the experiment?") and probabilistic inference (e.g., "How famous do you consider this person to be?"). For both judgments (episodic recognition and probabilistic inference), we observe the influence of both pre-experimental exposure, which is approximated by web-frequencies (e.g., Google search results), and experimental exposure, which is manipulated through an incidental study phase (e.g., a vowel counting task). The results of these experiments allow for an integrative understanding of how different sources of memory accessibility (experimental vs. pre-experiential) are combined, and possibly interfere with one another, depending on the type of memory-related judgment.

The role of emotional mediation in musical and vocal sound-color correspondence

Erin Isbilen
Cornell University, Ithaca, New York, United States

Abstract: This study investigates the role of emotional mediation in sound-color cross-modal correspondence, using two complementary sets of validated stimuli: the Montreal Affective Voices (MAV; Belin et al., 2008), and Musical Emotional Bursts (MEB; Paquette et al., 2013). These stimuli were presented to participants for color associations, emotional associations, and rated for arousal and valence. The results demonstrated that the same pattern of color association applied across both vocal and musical sounds, which strongly correlated with the perceived emotional connotation of the sound. Sounds across both domains that were rated as high arousal/negative valence were associated with red (anger), sounds rated as high arousal/positive valence were associated with yellow (happiness), and sounds rated as low arousal/negative valence were associated with blue (sadness). The results thus replicate previous research indicating that arousal and valence govern sound-color correspondence, suggesting that cross-modal associations may reflect reciprocal interactions between the connotative meanings of different stimuli.

Collaborative Story Construction and Telling for Second Language Learning

Mei Si
Rensselaer Polytechnic Institute

Abstract: Though utilizing information from the internet is common in language learning, it is often a spontaneous behavior and is not supported by computer programs that are specialized for language learners. In this project, we create a collaborative learning environment for Mandarin Chinese that implements the idea of learning by teaching. The students will learn and practice their language skills by collaboratively designing interactive presentations, such as introducing a place or a type of food with a computer agent, and the scenarios can be used later on by their peers for practicing. During the design process, storytelling techniques such as making an analogy, foreshadowing and flashback will be automatically suggested by the agent for making the presentation more interesting and clear.

Partnering with Autonomous Vehicles

Umesh Krishnamurthy
University of California Merced, Merced, California, United States

Paul Maglio
University of California Merced, Merced, California, United States

Abstract: Despite the growing prominence and capabilities of self-driving cars, little is known about human interaction with such autonomous vehicles. I am studying human interaction with multiple levels of automation and navigation in a simple driving simulator to discover principles for interaction design that encourage effective partnership between people and automated vehicles. I have conducted two experiments so far. In the first, participants completed driving tasks based on two different factors: automated versus manual control and navigation in low-level versus high-level language, which were combined via a factorial design into four conditions for each participant in the form of different driving scenarios. Performance was measured by travel time, frequency of driving and navigational errors, and levels of user satisfaction. The second iteration had all automated driving, with participants merely supervising while I gave varying amounts of verbal input.

Mind Wandering during Film Comprehension: The Role of Prior Knowledge and Situational Interest

Sidney Dmello
University of Notre Dame

Kris Kopp
University of Notre Dame, Notre Dame, Indiana, United States of America

Caitlin Mills
University of Notre Dame

Abstract: We assessed mind wandering (MW) during film comprehension. We predicted that prior-knowledge would aid in the construction of a situation model of the film, which would suppress MW by directing attention towards task-related thoughts, and that interest would moderate this effect. In our experiment, 108 participants either read a short story that depicted the plot (i.e., prior-knowledge condition) or read an unrelated story of equal length (control condition) prior to viewing the 32.5 minute film The Red Balloon. Participants self-reported their interest in viewing the film immediately before the film presentation and reported self-caught instances of MW while viewing the film. The prior-knowledge condition reported less MW compared to the control condition. MW also decreased over the course of the film, but only for the prior-knowledge condition, thereby suggesting a suppression effect. Finally, prior-knowledge effects on MW were only observed when interest was average or high.

A concurrent task facilitates, not impedes, the heel-to-toe standing balance in children: The case of a dual-task benefit

Rong-Ju Cherng
Department of Physical Therapy, College of Medicine, National Cheng Kung University, Tainan. Taiwan, ROC

Hsien-Ying Chen
Department of Physical Therapy, College of Medicine, National Cheng Kung University, Tainan. Taiwan, ROC

Chiu-Yu Cho
Department of Physical Therapy, College of Medicine, National Cheng Kung University, Tainan. Taiwan, ROC

Jenn-Yeu Chen
Department of Chinese as Second Language, National Taiwan Normal University, Taipei, Taiwan, ROC

Abstract: Performance in a dual task is typically worse than performance in a single task due to the sharing of limited cognitive capacity. The present study found the opposite results when the task involved postural control in non-typical standing. Thirty-six children aged 4-9 years stood on a force plate for 10 seconds with a normal or heel-to-toe stance. In the dual-task condition, they also performed an auditory or a visuospatial task. They were instructed to achieve high accuracy on the concurrent task while maintaining balance. Standing balance, expressed in terms of the velocity and the trajectory of the center of pressure on the force plate, was significantly better in the dual-task than in the single-task condition. Performances on the concurrent tasks were also better in the dual-task condition. The overall dual-task benefits are attributed to the increased deployment of cognitive capacity specially called for by the balance challenge in non-typical standing.

The effect of language impairment on non-symbolic exact quantity representation

John Verbos
Duquesne University, Pittsburgh, PA, USA

Sarah Wallace
Duquesne University, Pittsburgh, PA, USA

Alexander Kranjec
Duquesne University, Pittsburgh, PA, USA

Abstract: Both English-speakers whose access to number language is artificially compromised by verbal interference and the Pirahã (an Amazonian tribe without exact number words) appear to rely on analog magnitude estimation for representing non-symbolic exact quantities greater than 3. Here, 10 participants with aphasia from stroke performed the same 5 counting tasks from these previous studies. Performance was poorest when targets were not visible during response (70% correct) and best when targets were presented as subitizable groups of 2 and 3 (98% correct). Western Aphasia Battery-Revised subtest scores were reliably correlated with performance across counting tasks suggesting ways that both speech and naming may contribute to errors. Coefficients of variation for particular tasks, and significant correlations between target magnitude with both error rate and size across tasks suggests use of analog magnitude estimation for verbally impaired participants. Diverse forms of language impairment may contribute to errors on nonverbal counting tasks.

Do Open Specifier Positions at Clause Edges Alleviate Working Memory Load?

Ryan Rhodes
University of Delaware

Abstract: In this study I examine the relationship between syntactic working memory and intermediate gap positions in long-distance filler-gap dependencies. An "intermediate" gap position is a structural position, distinct from verbal argument positions, which is available to the filler at a clause edge between its surface dislocated position and its original canonical position.

I will use anterior negativity (AN) as an index of working memory resources as the filler is held by the parser before its integration in the final gap position. Under the hypothesis that such an available intermediate gap position offers a temporary integration of the filler, an attenuation of the anterior negativity is expected at the intermediate gap site. However, this attenuation is not observed, suggesting either that this intermediate gap position is in fact not available to the parser as a site of temporary integration, or that such integration has no mitigating effect on working memory resources.

Attentional Enhancement at Event Boundaries

Dare Baldwin
University of Oregon, Eugene, Oregon, United States

Eric Pederson
University of Oregon, Eugene, Oregon

Abstract: A fundamental aspect of everyday processing involves identifying discrete events within continuously unfolding sensory experience. However, the processes enabling determination of event boundaries remain poorly understood. Recently, inconsistent conclusions have emerged regarding attentional processes associated with detection of event boundaries. Use of the Dwell-Time Paradigm has indicated enhanced attention at event boundaries (e.g., Hard, Recchia, & Tversky, 2011), whereas evidence from the Rapid Serial Visual Presentation paradigm (e.g., Huff, Papenmeier, & Zacks, 2012) indicates impairment. We employed a change-detection procedure similar to the RSVP, except that the change to be detected was uniform across the entire visual field, rather than varying with respect to the viewer's spatial locus of attention. Changes occurred either at event boundaries or mid-stream within event segments. With spatial locus of attention rendered irrelevant, participants displayed significantly faster reaction time to changes coinciding with event boundaries, implying that viewers selectively target event boundaries with heightened attention.

Historical Semantics of Risk

Ying Li
University of Warwick

Thomas Hills
University of Warwick

Ralph Hertwig
Max Planck Institute for Human Development

Abstract: The key insight in this work is that the events people associate with risk is under systematic change over the past 200 years. Leveraging Latent Dirichlet Allocation (Topics Modeling) and the Google Ngram Corpus, we identified historical and newly-emerging events associated with risk and tracked their relevance over time. We also computed the probability of risk co-occurring with words associated with those identified events to capture a more accurate trend. Several highlights of the findings include: attention on risk has been spreading from one general domain (about losing life and war/battle) to a set of wider, more specific events and activities such as cancer, sex, HIV, smoke, and finance; in addition, the concept of risk has recently become more differentiated, incorporating both cost and benefits, long and short-term consequences. This approach could be extended to study semantic history of a number of other concepts of interest.

An explicit theory of sortal representation and some evidence for it.

Sandeep Prasada
Hunter College, CUNY

Abstract: Does the acquisition of words like "dog", "table", and "sand" require the support of sortal concepts? In arguing for and against sortals, theorists typically contrast representations of unsorted individuals (bare particulars) and sortal representations. As such, bare particular and sortal representations are presented as alternative means of representing concepts like "dog", "table" and "tree". Arguments for and against sortals typically proceed in the absence of an explicit characterization of the form of sortal representations. I present an explicit theory of the form of sortal representations. It turns out that, for sortals to do the work they need to do, they must incorporate bare particular representations into the sortal representation. Two experiments provide evidence for the predicted by the proposed theory of sortal representations. I also show how the proposed theory of sortal representation is consistent with recent findings by Rips and colleagues that seem to provide empirical evidence against sortals.

The influence of temporal order on the recognition of causal relations

Jane Neal
Northern Illinois University, DeKalb, IL

Katja Wiemer
Northern Illinois University, DeKalb, IL

Abstract: Fenker, Waldmann, & Holyoak (2005) found that participants are faster to recognize a causal relation between words presented in predictive (cause first) than in diagnostic order (effect first). We extended these findings to a comparison of abstract and concrete word pairs. Causality may play a more prominent role in abstract concept relations. Given that causally related abstract concepts are not always observable, they may often involve diagnostic reasoning (e.g. inferring motives). Across two experiments, participants made timed judgments of whether abstract and concrete word pairs of equal bidirectional associative strength were causally related. Items were presented in blocks comprising pairs in either predictive or diagnostic order. Reaction times were significantly lower for predictive order compared to diagnostic. This was not moderated by abstractness, but there was a slightly greater effect for concrete pairs. These data indicate that causal relations are likely stored in memory for both abstract and concrete concepts.

Can Distributional Fitting of Short Semantic Fluency Results Predict ADHD?

Janelle Szary
Indiana University

Michael N. Jones
Indiana University

Abstract: Remembering information can be likened to a search through a network composed of semantically related information. An appropriate search through this network requires an adaptive balance between exploratory and exploitative search behaviors. Without exploration, a searcher will perseverate too long on a semantic area devoid of resources. Without exploitation, a searcher may jump around aimlessly, failing to find the semantic areas with plentiful resources (see Hills, Jones, & Todd, 2012). The semantic fluency task, in which subjects are asked to recall items from a given semantic category, can be used to measure these behaviors. Classically, this task has been used to predict Alzheimer's susceptibility, but other clinically relevant predictions have required involved semantic analyses. Here, we show that distribution fitting applied to the gross time series of recall events, can be used to easily predict measures of clinical relevance such as Wender Utah ADHD and Zuckerman's Sensation Seeking scores.

Representations of Entropy and of the Relations Same and Different Early in Human Development

Rebecca Zhu
Harvard University, Cambridge, MA, United States

Jean-Remy Hochmann
CNRS, Bron, France

Sophia Sanborn
University of California, Berkeley

Susan Carey
Harvard University, Cambridge, MA, United States

Abstract: Animals typically fail 2-item Relational Match to Sample (RMTS), whereas animals from pigeons through primates succeed at 16-item RMTS. Furthermore, training on the 16-item arrays does not transfer to 2-item arrays in these non-human species. Animal researchers conclude that success on 16-item RMTS reflects a perceptual property of the set, variability or entropy, rather than conceptual representations of the relations 'same' and 'different'. Four experiments explore young children's ability to pass 2-item and 16-item RMTS. Like non-human animals, three- and four-year-olds fail 2-item RMTS while passing the16-item task. As with animals, training with 16-item cards does not facilitate success on 2-item RMTS in four- and five-year-olds. These data, as well as data from within the 16-item task, suggests that young children, like non-human animals, rely on entropy in RMTS tasks. Data from 5 and 6-year-olds suggest a representational change late in the preschool years.

Predictors of lexical stability in an artificially learnt language

James Brand
Lancaster University

Padraic Monaghan
Lancaster University

Abstract: Lexical items in the vocabulary of a language undergo dramatic changes over time, explaining the mechanisms that cause this change has been an important topic for the cognitive sciences. One particular focus for researchers has been understanding the dynamics of change in word forms. The rate (or half-life) at which word forms change over time varies greatly, and corpus-based cladistic studies have shown that certain properties, such as word frequency, length and age of acquisition, can be used to predict this variation. We test through the use of an artificial language learning paradigm the extent to which these psycholinguistic factors affect accurate learning of word forms, linking processes of acquisition with processes of evolutionary change. Our findings provide an insight into the underlying mechanisms that drive diachronic change within a language's vocabulary, highlighting the important role that the learning process has on lexical change.

Predictable stimulus onsets improve memory

George Kachergis
New York University

Shannon Tubridy
New York University

Todd Gureckis
New York University

Abstract: Exploring and remembering are fundamental to many human activities. Characterizing influences on recognition memory can help clarify the workings of memory systems and facilitate design of effective learning environments. Studies of self-directed learning show that a key determinant of self-directed benefits is in choosing when to see the next stimulus, but these results do not establish whether it is the act of choosing or the knowledge of stimulus arrival times that primarily matters. We disentangle these factors by asking whether predictable stimulus timing that is not under participant control still leads to a memory benefit. Participants saw pictures of objects one at a time with either a constant or unpredictable inter-stimulus interval (ISI) and showed better memory with constant timing across a range of ISIs. These results speak to interactions between attention and memory, the efficiency of study protocols, and the factors influencing effective self-directed learning.

Can a Bayes' Net approach capture intuitive use of sequential testimonies in a legal reasoning paradigm?

Jens Koed Madsen
Birkbeck, University of London, London, United Kingdom

Saoirse Connor Desai
City University, London, United Kingdom

Adam Harris
University College London, London, United Kingdom

David Lagnado
University College London

Abstract: The studies apply a Bayesian source credibility model to a legal setting to test epistemic influence of witness testimonies. The model amalgamates perceived witness trustworthiness and access to accurate information as independent elements that describe and predict the impact of the testimony of that particular witness.

Across two studies, the model enjoys a good fit with observed posterior ratings of the likelihood of guilt (study 1: R2 = .867, study 2: R2 = .701). Study 1 (n = 101) employs different witness types and reports whilst study 2 (n = 102) employs different witness types, access to accurate information, and reports.

The studies suggest the applicability of a Bayesian source credibility model in a legal setting to account for the impact of different witness types. We show that participants are sensitive to the type of witness and that different witnesses have a predictable impact on the perception of the testimony.

Exploring Individual Differences in Preschooler's Causal Reasoning Skills in the Physical and Digital Domains

Jessie-Raye Bauer
University of Texas at Austin

Amy Booth
University of Texas at Austin

Cristine Legare
University of Texas at Austin

Abstract: Do children reason about causal events differently in physically 'live' and digital domains? To answer this question, we introduced 35 3-year-olds to the traditional live version and a newly developed digital version of the "blicket detector" task. In both formats, the "blicket detector" first produced an interesting event (e.g., lit up) when a distinctive object (e.g., cylindrical block) touched its surface, but then failed to do so when a different object (e.g., triangular block) did so. After both blocks then simultaneously touched (and activated) the "blicket detector," children were asked to identify the 'causal' block. Children's performance correlated significantly across the physical and digital trials ($r = .4$, $p = .02$). Not only does this study further our understanding of children's causal reasoning skills in the digital domain, it introduces a major methodological advance with the development of a highly efficient and reliable digital version of the "blicket" task.

A subject-object asymmetry in the online processing of 'only': evidence from eye-tracking

Pooja Paul
Harvard University

Tanya Levari
Harvard University

Dylan Hardenbergh
Harvard University

Jesse Snedeker
Harvard University

Abstract: While most formal semantic accounts of focus-sensitive particles such as 'only' acknowledge that their interpretation requires the integration of contextual information with the linguistic representation, it is less clear how this interaction plays out in real-time. Recent psycholinguistic work in this domain favors an incremental processing story, but divergent results elsewhere complicate this picture. Our findings from two Visual World eye-tracking studies (n = 33, 32) help resolve this conflict, and confirm the existence of an adult processing asymmetry: sentences in which 'only' associates with the subject ('Only John bought an apple') take longer to process than object-only sentences ('John only bought an apple'). We find that current accounts of the representation and exhaustification of propositional alternatives invoked by 'only' do not explain this effect. We suggest that differences at the event-structural level — which propositional alternatives arguably map onto — might explain the asymmetry.

Taking the Easy Way Out: Effects of Feedback on Children's Metacognitive Control

Allison O'Leary
The Ohio State University

Vladimir Sloutsky
The Ohio State University

Abstract: It is a well-known finding that adults are "cognitive misers," in that they optimize performance by reducing effort whenever possible. We examined how this tendency emerges in the course of development. We incentivized participants to acquire as many points for correct responses as possible, and allowed them to choose between two games of differential difficulty on every trial. Whereas adults systematically chose the easier of the two games, 5-year-olds did not. However, when we gave children feedback on the basis of their choice rather than accuracy, they exhibited evidence of optimization by selecting the easier game. Further, they showed the same pattern when the two games were modified to be identical in difficulty, with only feedback supporting the choice of one game over the other. These findings suggest that children may be cognitive misers, but they rely on external feedback rather than internal signals of effort to optimize their behavior.

Instance-based Learning in Multi-cue Diagnosis

Christopher Myers
Air Force Research Laboratory

Kevin Gluck
Air Force Research Laboratory

Andrew Halsey
ORISE at the Air Force Research Laboratory

Jack Harris
Air Force Research Laboratory

Abstract: Decision heuristics are often described as fast and frugal, taking little time and few computations to make a good decision (Gigerenzer & Todd, 1999). Fast & Frugal Trees (F&FTs) are a type of decision heuristic that are a special case of decision trees in which there is a possible exit out of the decision process at every node in the tree (Luan, Schooler, & Gigerenzer, 2011). We present predictions from a computational process model of learning in a multi-cue diagnosis task with and without information acquisition costs and across different penalty values for errors. The model uses Instance-Based Learning Theory (IBLT) to acquire new knowledge and makes precise performance predictions across a range of dependent variables. We will compare the a priori model predictions to F&FT-constrained machine learning results on the same stimuli and also to empirical results collected from human participants making decisions in the same environment.

Rapid acquisition of novel information: Is disjunctive syllogism necessary for fast mapping?

Hillary Abel
Villanova University, Villanova , Pennsylvania, United States

Anna Drummey
Villanova University, Villanova , Pennsylvania, United States

Stephanie Vicari
Villanova University, Villanova , Pennsylvania, United States

Irene Kan
Villanova University, Villanova , Pennsylvania, United States

Abstract: We investigated two possible mechanisms that may mediate the rapid acquisition of novel words and their corresponding referents (i.e., fast mapping, FM). In the standard paradigm that examines FM, a novel label is presented alongside a novel object and a familiar object, and subjects are asked to identify the item that corresponds to the novel label. Acquisition of name-object pairing is subsequently assessed. One possible mechanism underlying FM is disjunctive syllogism: the active rejection of the familiar item, which allows for the novel object-to-label mapping (e.g., "I know this is a cricket, so "torato" can't be referring to that; it must refer to the unfamiliar item."). Another possible mechanism involves activation of a relevant semantic network (e.g., insect) into which the novel concept can be incorporated. We found that semantic network activation alone is sufficient, and that active rejection is not necessary, for the rapid acquisition of novel object-name associations.

Information Processing during Intertemporal Choice: An Investigation using Eye Movement Data

Kelli L. Johnson
Stony Brook University

Michael T. Bixter
Georgia Institute of Technology

Christian C. Luhmann
Stony Brook University

Abstract: Intertemporal choices consist of trade-offs between reward magnitude and the delay until those rewards are received. Distaste for delay (i.e., impatience) is related to various undesirable variables including drug use, credit card debt, and low grade point average. These findings have underscored the critical need to better understand intertemporal preferences. Previous work has shown that forcing participants to wait 9 seconds before making an intertemporal choice yields greater patience than when they are forced to wait 3 seconds. Unfortunately, the mechanisms that produced this effect are currently unknown. The current study uses a similar choice paradigm but collects eye movement data in order to non-invasively investigate decision makers' information processing strategies under short and long deliberation times. Eye movements over the various pieces of choice-relevant information (i.e., delays and magnitudes of rewards) were related to the deliberation time manipulation, individual choices, and individual differences in patience. Theoretical implications are discussed.

Neural Resonance Theory: Entrainment to Missing Pulse Rhythms

Charles Wasserman
University of Connecticut, Storrs, CT, USA

Jung Nyo Kim
University of Connecticut, Storrs, CT, USA

Edward Large
University of Connecticut, Storrs, CT, USA

Erika Skoe
University of Connecticut, Storrs, CT, USA

Abstract: Many rhythm perception experiments employ simple isochronous rhythms, in which synchronous neural or behavioral responses are observed. However, responses at the stimulus frequency do not allow one to distinguish whether synchrony occurs as a response to common input, or as the result of an emergent population oscillation that entrains at a particular frequency. This study aimed to investigate whether the sensorimotor system, as measured by 32- channel cortical EEG, would entrain to a complex rhythm at the pulse frequency even when the complex rhythm contained no spectral power at that frequency. Dynamical analysis predicts neural oscillation will emerge at such a "missing" pulse frequency. We report evidence of response in the EEG to missing 2 Hz pulse rhythms. These data support the theory that rhythmic synchrony occurs as the result of an emergent population oscillation that entrains at this particular frequency. We also discuss generators of the missing pulse response.

Young children's estimation of difficulty and time

Hyowon Gweon
Stanford University

Mika Asaba
Stanford University

Abstract: Even infants have a remarkably sophisticated understanding of objects, agents, and how they interact. We investigated young children's ability to reason about the relationship between complexity of physical structures created by agents, their perceived difficulty, and the time required for creating these structures. Seventy 4-5 year-olds were shown trials consisting of pairs of agents who had the same numbers of blocks but made different structures (e.g., horizontal line vs. vertical tower, castle structure vs. two piles of blocks). Children were asked which structure was easier to make (Difficulty condition) or who was done first (Time condition). Even the youngest participants were successful in determining which structure is more difficult, but their estimates of time showed improvement with age. These results offer novel insights into how an early understanding of difficulty and time shape young children's beliefs about how agents intervene on the physical world to induce changes in their states.

Exploring the Use of Conversations with Virtual Agents in Assessment Contexts

Diego Zapata-Rivera
Educational Testing Service, Princeton, New Jersey, United States

G. Tanner Jackson
Educational Testing Service, Princeton, New Jersey, United States

Irvin R. Katz
Educational Testing Service, Princeton, New Jersey, United States

Abstract: Conversations with computer agents can be used to measure skills that may be difficult to accomplish using traditional multiple-choice assessments. In order to achieve natural conversations in this form of assessment, we are exploring issues related to how test-takers interact with computer agents, such as what dialogue moves lead to interpretable responses, the influence of "cognitive characteristics" of computer agents, how should the system adapt to test-taker responses, and how these interactions impact test-taker emotions and affect. In this presentation we will discuss our current research addressing these questions, illustrating important dimensions that are involved with designing a conversation space and how each design decision can impact multiple factors within assessment contexts.

Mothers' use of emotion words to their 15- to 18- month old children.

Mai Hamana
the Universtiy of Tokyo

Abstract: Young children can apply certain emotion words in a manner that is too wide for facial expressions or situations and use other emotion words too narrowly. To explore the extent to which children's overextension or underextension of emotion words is influenced by mothers' use of those words, the present study analyzed how mothers talk to their 15- to 18-month-old children. Mothers were asked to explain to their children the emotions of a model using particular facial expressions or those of characters in stories. The results indicated that mothers used some emotion words in a manner that was too wide and other emotion words too narrowly. Mothers' use of emotion words is partly related to children's application patterns of those words, especially regarding facial expressions. Regarding the use of emotion words for characters in stories, other factors aside from mothers' use can contribute to children's semantic domain of emotion.

Effects of categorization on item memory and forgetting

Kevin Darby
The Ohio State University, Columbus, OH, USA

Vladimir Sloutsky
The Ohio State University, Columbus, OH, USA

Abstract: Memory and categorization have different goals. The goal of memory is to keep a distinct record of each individual item. In contrast, categorization aims to treat items as equivalent in some way despite differences. When memory for items is the priority, research suggests that adults are able to bind many elements of an experience to form a complex memory structure, and that such binding may help guard against forgetting due to interference from learning in similar situations. When categorization is the priority, however, complex binding structures in memory may impede generalization. In this experiment adults demonstrated robust memory for items after learning one set of categories, but much worse memory for items that were categorized differently in a second set. Results suggest that this interference was due to failure to form complex binding structures in memory, as a result of selective attention during categorization.

Optimal Predictions in Illness Cognition

Talia Robbins
Rutgers University, New Brunswick, NJ, U.S.A

Pernille Hemmer
Rutgers University, New Brunswick, New Jersey

Abstract: People make accurate predictions for many real world events e.g. human life spans (Griffiths & Tenenbaum, 2006). Accurate predictions are particularly important in the domain of health, where illness knowledge directly influences patient outcomes. To understand how well peoples' illness expectations were aligned, we asked participants to estimate durations for 9 illnesses, and compared their responses to the real-world distributions. We found that for common acute illnesses (e.g., the cold) people make accurate predictions, whereas for rare chronic illnesses (e.g., COPD) people make comparatively poor predictions. Further, we found that participants overestimate the prevalence of every illness, especially for those that are more common (e.g., the cold). Taken together, these results suggest that people more accurately estimate the duration of common acute illnesses, but this may cause them to overestimate the prevalence of these illnesses. Results will be discussed in terms of implications for both cognition and behavioral health theory.

Language acquisition as sparse foraging: mapping path-dependence in word acquisition

Riccardo Fusaroli
Aarhus University

Malte Lau Petersen
Aarhus University

Abstract: How does the acquisition of a new word affect the successive acquired ones? We model language acquisition in terms of foraging in an unprecedentedly fine-grained 1-child corpus (0-3y, RoyEtAl2015). We assess whether successive words are learned in close semantic clusters or not (exploitation vs. exploration) and the structure of these clusters. Words are defined in terms of topic distribution in the parental input (Latent Dirichlet Allocation). Distance between successively learned words is measured as cosine distance between their topic distribution.

Word acquisition can be accurately described as foraging in a sparse resource environment with very local path dependence (power law distribution: alpha=3.9$\pm$0.2; detrended fluctuation analysis: alpha=0.6). Words are acquired in semantically close clusters of 2-4 successive words (Recurrence Quantification Analysis: L=2, LMAX=4, V=2, VMAX=4). The effects remain when controlling for shuffled baselines and temporal distance between word acquisition.

Multiple Systems for Modal Cognition

Jonathan Phillips
Harvard University

Fiery Cushman
Harvard University

Abstract: The capacity for representing and reasoning over sets of possibilities, or modal cognition, has long been understood as central to many high level judgments. To date, however, little empirical research has sought to directly investigate the connection between these high-level judgments and the underlying cognition that allows humans to represent and reason over sets of possibilities. The present studies build on previous developmental research which suggests that the early emerging system for modal cognition treats norm-violations (e.g., immoral actions) as impossible. Across two studies, we provide evidence that a similar system for representing possibilities persists in human adults, despite the development of an additional capacity for reasoning about possibilities in a way that is independent of considerations of normality. Study 1 distinguishes between these two ways of representing possibilities. Study 2 demonstrates that the early-emerging system is often recruited when adults make high-level cognitive judgments.

The Effect of Varying Problem Contexts on Learning Probability Rules

Tiantian Jin
Columbia University

Marianna Lamnina
Columbia University

Abstract: While previous research shows that varying problem contexts generally facilitates learning (Ranzijin, 1991), it is still unknown how much variability is ideal. Since it is often more economical for teachers to use consistent problem contexts, it is valuable to know how much variability is needed. We examined this in teaching probability. Students randomly assigned to one of three groups learned four rules with four worked examples each, differing in context variability: One group learned four rules with the same cover story (all examples for all rules used cards), the second group with different cover stories per rule (multiplication taught with cards, permutation, with spinners), and the third group with varying cover stories within each rule (addition taught with cards, spinners, marbles, and dice). Learning with context varying within rules led to the greatest learning gains from pretest to posttest. We discuss implications of these findings and underway follow-up research.

Analyzing experimental paradigms under modification on web-based experiment platforms

Michael Pacer
University of California, Berkeley

Abstract: Experimental paradigms are particular experiments that can be modified along a variety of dimensions to answer questions different than the original inquiry but which have similar content or structure. For example, the original looking-time study is an experimental paradigm that has shaped developmental psychology. Thomas Kuhn proposed that, under normal conditions, experimentation by modification of previous paradigms is how science progresses.

Web-based experiment platforms (e.g., psiturk and Wallace) are installed with collections of working experimental paradigms that are pre-populated with structures necessary to use the platform, but which can be modified to allow users to generate novel experiments. Because these experiments are implemented in code, we can identify exactly how the code is modified in practice, and begin to directly measure and even test Kuhn's hypothesis regarding the progress of normal science. I will describe possible methods for achieving this, ideally providing others a paradigm to modify.

Sex Differences in Mental Rotation Performance: The Self-fulfilling Prophecy of Gender Stereotypes

Alina Nazareth
Temple University

Shannon Pruden
Florida International University

Abstract: Male advantage on spatial tasks may be explained in part by gender stereotypes (Nash 1975). The current study investigated the effect of awareness of sex differences in mental rotation on mental rotation (MR) performance. We hypothesized that students with negative stereotypes would score significantly lower than students who were unaware or held positive stereotypes. Participants – 285 undergraduates — completed the Shepard & Metzler (1971) MR task followed by a short online survey. Preliminary analysis revealed a significant sex difference in mental rotation performance F (1, 256) = 9.68, p=. 002. There was no main effect of awareness on MR performance. Interestingly, there was a significant interaction between sex and awareness on MR performance, F (1,256) =6.77, p=. 010. Results on the role of awareness in cognitive strategy selection will be presented. By understanding gender stereotypes associated with spatial ability, we can reduce the gender gap found in STEM disciplines.

Relative influence of anchoring and centering biases in reconstructive memory

Sean Duffy
Rutgers University

L. Elizabeth Crawford
University of Richmond

John Smith
Rutgers University

Abstract: We report the results of an experiment probing the relative influence of centering and anchoring biases in reconstructive memory for line lengths. On each of 90 trials participants (N=120) viewed a target line, which they reproduced after a delay by adjusting a response line. We manipulated the starting size of this response line in three conditions: one provided an anchoring bias opposite the centering bias (expand condition), one in the same direction (contract condition) and one that provided no anchoring bias (control). We eliminated the centering bias in the expand condition, increased it in the contract condition, and showed an attenuated centering bias in the control condition. We discuss the implications for these results in relation to cognitive models of stimulus reproduction that employ the method of serial reproduction. We suggest that experiments of this type should carefully control for the possible influence of anchoring biases in reconstructive memory.

Making invisible ”trouble” visible: Self-repair increases abstraction of referring expressions

Gregory Mills
University of Groningen, Groningen, Netherlands

Gisela Redeker
University of Groningen, Groningen, Netherlands

Abstract: A central finding in dialogue research is that interlocutors rapidly converge on referring expressions which become progressively contracted and abstract. However, there is currently no consensus on which mechanisms underpin convergence: The interactive alignment model (Pickering and Garrod) favours priming, the grounding model (Clark, 1996) prioritizes positive feedback, while Healey (2002) demonstrates the importance of miscommunication. To investigate convergence in closer detail we report a variant of the “maze-task”. Participants communicate via a text-based chat tool which selectively transforms participants’ private turn-revisions into public self-repairs that are made visible to the other participant. Dyads who received these artificially transformed turns used more abstract referring expressions, but performed worse at the task. We argue this is due to self-repairs having a beneficial effect of amplifying naturally occurring miscommunication (Healey et al., 2013), while also having a deleterious effect of decreasing participants’ confidence in the conventions established during the task.

Inferring actions, intentions, and causal relations in a neural network

Andrew Saxe
Harvard University, Cambridge, MA, USA

Abstract: From a young age, we can select actions to achieve desired goals, infer the goals of other agents, and learn causal relations in our environment through social interactions. Crucially, these abilities are productive or generative: for instance, we can impute desires to others that we have never held ourselves. This capacity has been captured by the powerful Bayesian Theory of Mind formalism, but it remains to forge connections to the rich neural data around action selection, goal inference, and social causal learning. How can productive inference about actions and intentions arise within the neural circuitry of the brain? Using the recently-developed linearly solvable Markov decision process, we present a neural network model which permits a distributed representation of tasks. Such a representation allows the expression of infinite possibilities by combining a finite set of bases, enabling truly generative inference of actions, goals, and causal relations in a neural network framework.

Forming spatial habits in the wild: Examining stabilization in classroom seating arrangements

Mona Zhu
University of Waterloo

Evan Risko
University of Waterloo

Abstract: Previous research suggests that individuals tend to form spatial habits (i.e., placing objects in consistent locations in space) when interacting with objects in the environment (Zhu & Risko, in press). However, it remains unclear how such spatial habits develop over time. One hypothesis suggests that spatial habit formation may involve a process of stabilization whereby individuals' spatial behaviour becomes progressively fixed over time. We examined this hypothesis by tracking students' seating behaviour in a classroom over the course of 12 weeks. Although individuals' overall seating choice tended to cluster near where they initially sat, we did not find evidence that seating behaviour stabilized over time. However, a significant curvilinear relation was found between seating choice and time such that seating choice near the beginning and end of the 12-week period were more varied than those in the middle. Implications of this study for understanding spatial habit formation will be discussed.

An action dynamics study of the onset of prediction

Maryam Tabatabaeian
University of California Merced, merced, California, United States

Rick Dale
University of California, Merced

Abstract: A recent approach in cognitive science argues that prediction is a core concept underlying cognition, to the extent that brains could be referred to as "prediction machines" (Clark, 2013). Extending experimental paradigms to explore prediction facilitate tests of this claim. In a previous study we introduced a statistical learning paradigm to detect when participants are predicting during implicit/explicit learning. The results revealed that participants tend to rapidly switch into a predictive mode almost as a discrete strategy. While this intriguing possibility was not directly explored in the original study, in this project we repurposed the task to explicitly explore the onset of predictive behaviors. Through a mouse-tracking task, participants get to learn the statistical structure in a sequence of flashing dots while their mouse movements are being recorded. Findings reveal the level of statistical structure in the environment that triggers the rapid onset of predictive behavior in participants.

The ON/IN scale of semantic extensions of adpositions: testing through artificial language learning

Matthew Stave
University of Oregon, Eugene, Oregon, USA

Eric Pederson
University of Oregon

Abstract: Cross-linguistically, the semantic extensions of certain adpositions obey an implicational scale, ranging from the situation class of prototypical ON (support from below) to prototypical IN (full containment) (Bowerman & Pederson 1992, Levinson & Meira 2003, Feist 2008). However, the nature of this phenomenon remains unresolved. Two general accounts are possible: 1) the patterned output is the simply the result of diachronic factors and speakers learn which extensions go with each adposition without any particular analysis, or 2) there may be some online psychological reality to the scale.

Language learning may be a fruitful domain to test this second account. Using an artificial language learning paradigm, subjects are trained on alternative forms paired with scenes representing some situation classes on the scale. In testing, they generalize the use of these forms to novel classes. Preliminary findings suggest a generalization of forms to uses in ways consistent with a psychological representation.

Time Course of Fidelity and Contributing Factors to Long-Term Memory

Kimele Persaud
Rutgers, the State University of New Jersey, New Brunswick, New Jersey

Pernille Hemmer
Rutgers University, New Brunswick, New Jersey

Abstract: Various models have been implemented to explain long-term memory (Brady, et al., 2013; Lew, et al., 2015), with some being derived from studies of visual working memory (Bays, et al. 2009; Zhang & Luck, 2008). The implicit assumption is that processes and mechanisms of working memory also exist in long-term memory. However, the findings of fidelity and contributing factors are highly varied (e.g., Persaud & Hemmer, 2014; Schurgin & Flombaum, 2015) To address what happens to memory traces as they transition from visual working into long-term memory and what factors, such as prior knowledge and guessing, contribute to the "lifespan" of long-term memory, we implemented three models: the standard remember-guess model, a three-component remember-guess model, and a Bayesian mixture model and evaluated these models against data from a continuous recall task. The results clarify the time course of fidelity in long-term memory and pinpoints specific factors that contribute to memory.

The Neurorobotics Platform of the Human Brain Project

Florian Roehrbein
Technical University of Munich

Marc-Oliver Gewaltig
École Polytechnique Fédérale de Lausanne

Cecilia Laschi
Scuola Superiore Sant' Anna

Gudrun Klinker
Technical University of Munich

Paul Levi
Forschungszentrum Informatik

Alois Knoll
Technical University of Munich

Abstract: The aim of the neurorobotics platform of the Human Brain Project (HBP) is to offer scientists from various fields a software and hardware infrastructure that allows them to connect brain models to detailed simulations of robot bodies and environments. In the ramp-up phase of the HBP a first version of this platform has been developed, which allows researchers to design and run simple experiments in cognitive neuroscience using simulated robots and simulated environments linked to simplified versions of HBP brain models. The developed tools, i.e., various designers, simulation engines and simulation viewers, allow researchers to operate robots remotely, to repeat in-silico experiments and to visualize the behavior of the robots in real-time. Together with five other ICT platforms developed in the HBP, these technologies will also enable the development of brain-inspired computing systems. The first version of this platform has been released March 2016 and is described in this contribution.

Understanding human facial attractiveness from multiple views

Amanda Song
University of California San Diego, La Jolla, CA, United States

Linjie Li
University of California San Diego, La Jolla, CA, United States

Vicente Malave
University of California San Diego, La Jolla, CA, United States

Gary Cottrell
University of California San Diego, La Jolla, CA, United States

Angela Yu
University of California San Diego, La Jolla, CA, United States

Abstract: Facial attractiveness has long been a topic of interest for cognitive scientists. Early psychological research has found that averageness, distinctiveness and familiarity of a face can influence facial attractiveness. However, faces also convey rich social information. How various social features are related to facial attractiveness hasn't been systematically studied before. We investigate facial attractiveness in the context of social feature evaluation and find that social attributes like appearing interesting and sociable contribute to facial attractiveness whereas appearing boring, and humble are negatively correlated with attractiveness. We further compare social features of faces with the physical configuration of faces and we are able to use geometric features to predict facial attractiveness. We further study the individual differences on attractiveness perception and find out that. Our study illustrates that social attributes and pixel information can go hand in hand to facilitate attractiveness prediction.

Causal perception is constrained by principles of Newtonian mechanics

Jonathan Kominsky
Yale University

Brent Strickland
Institut Jean Nicod

Annie Wertz
Max Planck Institute for Human Development

Claudai Elsner
Max Planck Institute for Human Development

Karen Wynn
Yale University

Frank Keil
Yale University

Abstract: Humans irresistibly perceive certain events as causal. We show, for the first time, that there is not one monolithic representation of causality in perception. Rather, there are multiple categories of causal events in perception, one of which is constrained by an approximation of a Newtonian mechanical principle: in an elastic collision, a struck object cannot move at more than double the speed of the object striking it. We show that adults are sensitive to causal (but not non-causal) events that violate this principle in a visual search task (Experiment 1), that this sensitivity is due to a categorical boundary and not the salience of this event (Experiment 2), and that the threshold for detecting these events approximates this Newtonian limit (Experiment 3). Finally, we argue that categorical boundaries are a core feature of causal perception, as they are present around the age at which causal perception first emerges (Experiment 4).

Finger Gnosis Predicts Children's Numeracy, Despite Controlling for Visuo-Spatial Memory

Marcie Penner-Wilger
King's University College at Western University

Sheri-Lynn Skwarchuk
University of Winnipeg

Carla Sowinski
Carleton University

Jo-Anne Lefevre
Carleton University

Abstract: Finger gnosis, the ability to mentally represent one's fingers, predicts numeracy in children (Penner-Wilger et al., 2007, 2009) and adults (Penner-Wilger et al., 2014, 2015). It has been argued that the relation may reflect visuo-spatial memory, rather than finger gnosis ability per se. This rival hypothesis was not supported in adults (Penner-Wilger et al., 2015), but here we examined it in children, using both a novel set of Grade 1 participants (N = 119) and a separate set previously reported on (Penner-Wilger et al., 2007, 2009; N=146). In multiple regressions, for each set of participants, finger gnosis significantly predicted numeracy skills, measured using the KeyMath Numeration subtest. Moreover, the relation between finger gnosis and numeracy held for both sets of participants, despite controlling for visuo-spatial memory, measured using a Corsi-block test. These findings suggest that the relation between finger gnosis and numeracy is robust and does not reflect visuo-spatial memory.

Does Experience with Physics Concepts Improve Mental Rotation Performance?

Rosalie Odean
Florida International University, Miami, FL, USA

Alina Nazareth
Temple University , Philadelphia , Pennsylvania

Jessica Bartley
Florida International University, Miami, FL, USA

Eric Brewe
Florida International University, Miami, FL, USA

Angela Laird
Florida International University, Miami, FL, USA

Shannon M. Pruden
Florida International University, Miami, FL, USA

Abstract: STEM disciplines have been shown to positively impact an individual's visuospatial skills (Kozhevnikov, 1999). The current study examines improvement in spatial thinking in physics undergraduate students over the course of a semester. Students completed the Shepard and Metzler (1971) task at two time points– beginning and end of a semester – where they were asked to determine if two 3D figures were a match or mirror-images of each other. A Tobii X60 eye-tracker was used to record eye movement as an indirect measure of cognitive strategy selection. Preliminary analysis show a significant improvement in mental rotation performance from time point 1 (M=31.867, SD=5.027) to time point 2 (M=35.333, SD=3.885) t(14)=-3.014, p=.009. A latent profile analysis will be used to model cognitive strategies selected at time points 1 and 2 and analyzed for sex differences. The findings of this study are important for understanding the underrepresentation of women in STEM.

Pragmatic inference in definite and indefinite contexts

Kelsey Moty
Lehigh University

Mahesh Srinivasan
University of California, Berkeley

Abstract: Previous studies have shown that children have difficulty inferring intended referents of definite/indefinite determiners: e.g., that "Give me the ball" implies a specific ball, while "Give me a ball" requests any ball from a larger set. Here, we show that these findings need not indicate a fragile capacity for pragmatic reasoning because adults only make such inferences within specific contexts. Across four studies, we found that when presented with novel labels in definite contexts (the dax), adults consistently selected unique objects as the referent (though they were not at ceiling), suggesting they interpreted the definite as conveying specificity. Strikingly, however, when presented in indefinite contexts (a dax), subjects did not reliably link novel labels to objects of a larger set of kind-members, unless the context explicitly encouraged them to reason about the intended addressee. Together, these findings suggest that failures to make inferences about definiteness need not reflect pragmatic incompetence.

Using eye tracking data to compare models of numerical estimation

Bradley Morris
Kent State University, Kent, Ohio, USA

Christopher Was
Kent State University, Kent, OH, USA

Amy Masnick
Hofstra University, Hempstead, NY, US

Bushra Aldosari
Kent State University, Kent, OH, USA

Abstract: People accurately compare and estimate means without using formal calculations, however, little is known about the cognitive processes underlying these behaviors. We used objective, behavioral data (e.g., eye fixation patterns), which are compatible with multiple representations, to compare cognitive models. Specifically, we compared seven cognitive models including working memory activation (weighting values as a function of the number of and duration of fixations), working memory constraint (e.g., recency + primacy, last four), or Bayesian models (e.g., first fixation set as prior).

Our task presented sets of 5 to 10 3-digit numbers (framed as the result of a home run derby) and asked participants to predict how far the next ball would be hit. The same fixation data were loaded into each model to create a unique estimate, which was then compared to the participant's actual prediction. The difference between the model and actual was calculated to create an accuracy index.

A Cognitively Realistic Model of Decision Making in Ocean Ecology

Philipp Koralus, Jens Madsen, Ernesto Carella, and Richard Bailey

There is strong evidence that a key determinant of the ecological state of the world's oceans is the decision-making of fishers and policy-makers. There is a large empirical literature on the complex reality of human decision-making, but a comparative lack of work bringing detailed cognitive facts to models of aggregate behavior of this kind. We show how a psychologically realistic description of decision-makers can be integrated into a large-scale ocean systems model, going beyond profit maximization agent-based models. In particular, we seek to model the questions that frame decisions of different types of fishers and the motivating reasons on the basis of which those questions are resolved. The broader purpose of the model is to see how those patterns could inform regulatory policy. We present the basic architecture of the model and the data gathering tools that will allow us to adapt the model to fishers in different environments.

Neural bases of semantic-memory deficits for events

Chia-Ming Lei (chl162@pitt.edu)
Department of Communication Science and Disorders, 6035 Forbes Tower
Pittsburgh, PA 15260 USA

Haley C. Dresang (haleydresang@pitt.edu)
Department of Communication Science and Disorders, 6035 Forbes Tower
Pittsburgh, PA 15260 USA

Michelle B. Holcomb (mbh32@pitt.edu)
Department of Psychology, Learning Research & Development Center, 3939 O'Hara Street
Pittsburgh, PA 15260 USA

Tessa C. Warren (tessa@pitt.edu)
Department of Psychology, Learning Research & Development Center, 3939 O'Hara Street
Pittsburgh, PA 15260 USA

Michael Walsh Dickey (mdickey@pitt.edu)
VA Pittsburgh Healthcare System, University Dr C
Department of Communication Science and Disorders, 6035 Forbes Tower
Pittsburgh, PA 15260 USA

Abstract

This study investigated the neural bases of event-related semantic-memory deficits among people with aphasia due to left-hemisphere (LH) stroke. A novel task using naturalistic photographic stimuli and patient-friendly procedures was used to test event-related semantic knowledge. In the task, participants decided whether depicted events were normal (represented in semantic memory) or were abnormal (not represented in semantic memory). Performance on this Event task was correlated with deficits in action- and object-concept processing and on standardized language measures, especially action- and verb-processing deficits. Logistic regression analyses examined lesion correlates of patient performance on the Event task. Surprisingly, increasing LH lesion size in action ROIs was associated with *improved* performance on the event-knowledge task. These findings suggest that action processing may play a special role in event-related semantic memory representations. Furthermore, they are consistent with recent claims that the right hemisphere may be especially important for activation of event-related knowledge.

Keywords: semantic memory; event-related knowledge; left-hemisphere lesion; aphasia; lesion-deficit analysis

Introduction

Semantic memory refers to an individual's core world knowledge, ranging from general facts to specific features of objects and actions (Squire, 1987; Tulving, 1972; Yee, Chrysikou, & Thompson-Schill, 2013). Semantic memory therefore impacts a wide variety of cognitive domains, including language. Semantic memory representations have been examined clinically through standardized assessments for object-related conceptual semantics, such as the Pyramids and Palm Trees task (PPT; Howard & Patterson, 1992), and action-related conceptual semantics, such as the Kissing and Dancing Task (KDT; Bak & Hodges, 2003). However, the focus of most research has been on concrete concepts (e.g., object semantics). There has been much less investigation of abstract concepts (e.g. event-related semantics).

Event-related semantic memory consists of knowledge about common events, such as playing soccer or eating breakfast. Event-related representations bind together information about actions and their common participants (such as agents, themes, and locations: Ferretti, McRae & Hatherell, 2001; McRae & Matsuki, 2009). Event-related semantic memory representations are especially important for rapid language comprehension and online language prediction (Matsuki et al., 2011; Metusalem et al., 2012). Event-related semantic knowledge can be quickly activated by isolated words, agent-verb combinations, or morphosyntactic cues, and has been shown to guide expectations for upcoming words and concepts (Hare et al., 2009; Kamide, Scheepers & Altmann, 2003; McRae & Matsuki, 2009). These event-related expectations contribute to comprehension at both the sentence and the discourse level (Metusalem et al., 2012). Semantic-memory impairments for events can thus contribute to debilitating communication deficits.

Cognitive neuropsychology has a long history of investigating individuals with semantic deficits (e.g., Warrington & Shallice, 1984). Persons with aphasia (PWA) and persons with neurodegenerative disorders such as semantic dementia or fronto-temporal dementia often

experience semantic-memory impairments. These impairments can impact both language production and comprehension. Size, location, and relative atrophy of lesions, as well as cortical and cognitive reorganization post-stroke, yield individual differences in semantic-memory deficits (Antonucci & Reilly, 2008). For example, patients with either lesions or neurodegenerative disorders affecting left-hemisphere (LH) temporal regions exhibit deficits in object processing (e.g., Bak & Hodges, 2003; Binney et al., 2010; Kim & Thompson, 2004; see discussion in Lambon Ralph, 2014). In contrast, patients with atrophy or lesions affecting LH frontal regions exhibit deficits in action-concept processing (Bak & Hodges, 2003; Hillis, Oh & Ken, 2004; Kemmerer et al., 2012; for a review see Mätzig, et al., 2009). There is also evidence that action or verb processing activates posterior LH superior temporal gyrus and adjacent angular gyrus (Boylan, Trueswell, Thompson-Schill, 2015; den Ouden, et al., 2009).

Still, it remains unclear how event-related semantic-memory representations are organized within the brain. Events are complex, abstract concepts, and involve constellations of actions, objects, spatial, and temporal information. We know little about how these different components contribute to the neural representation of event-related conceptual semantic knowledge. The evidence described above suggests that distinct cortical networks are associated with action versus object conceptual processing. Event-related semantic representations may lean more strongly on action-related networks, given that actions are at the core of event-related knowledge (McRae & Matsuki, 2009), but this remains unknown.

One study to date has directly examined how this type of event-related knowledge is represented in the brain. Metusalem et al. (2016) used divided visual-field presentation to examine how event-related knowledge is activated in the two cerebral hemispheres. They presented short discourses from a previous ERP study (Metusalem et al., 2012) that were intended to activate event-related knowledge representations. These short discourses were followed by sentences with a critical word that was expected (*snowman*), unexpected but related to the event activated by the previous context (*jacket*), or unexpected and unrelated to the event (*towel*):

(1) A huge blizzard swept through town last night. My kids ended up getting the day off from school. They spent the whole day outside building a big [snowman/jacket/towel] in the front yard.

Metusalem et al. (2012) found that the expected word elicited a smaller N400 response than the two unexpected words. In addition, they found that the event-related word elicited an attenuated N400 response compared to the unrelated, unexpected word. This is parallel to many other ERP findings showing that words that are semantically related to a strongly expected word elicit attenuated N400 responses (Federmeier, 2007). Metusalem et al. (2016) found that the advantage for event-related words was present when the critical word was presented to the left visual field/right hemisphere (RH), but absent when it was presented to the right visual field/LH. This finding suggests that event-related knowledge may be activated or represented primarily in the RH. This finding is in contrast to the lesion-based and neuroimaging evidence noted above, which suggests a LH bias in action and object conceptual representations.

The current study examined the neural basis of event-related semantic memory via a combination of behavioral testing and lesion-deficit analyses. Participants with left-hemisphere lesions were tested on a novel, patient-friendly assessment of event knowledge, along with standardized measures of object- and action-related conceptual processing and language performance. The influence of lesions to LH action- and object-related regions of interest on performance in the Event task was then modeled. The study tested two hypotheses: (1) that participants' performance on the Event task will be more strongly correlated with their action-concept/verb processing deficits than their object-concept/noun processing deficits; and (2) that lesions to LH action-processing regions will negatively impact event-knowledge task performance, but lesions to object-processing regions will not.

On the other hand, if Metusalem et al.'s (2016) claim regarding the RH bias for event-related representations is correct, LH lesions may be unrelated or even positively related to event task performance. Increasing LH lesion may force participants to lean more heavily on RH to perform the Event task, much as larger LH lesion prompts PWA to recruit RH regions for language tasks (Saur et al., 2006). Although this RH recruitment may be maladaptive for language processing, it could result in better performance if event-related semantic memory is RH-biased.

Method

Participants

All participants (n=26, 8 female) had unilateral left-hemisphere (LH) strokes. They ranged in age from 50 to 82 (mean=73.4) and were between 17 and 276 months post-stroke (mean=73.4). They all had diagnosed aphasia, confirmed by their performance on the Comprehensive Aphasia Test (Howard, Swinburn, & Porter, 2004; CAT Modality Mean T-score mean=55.42). All 26 participants completed behavioral testing designed to measure their event-related knowledge and its connection to action and object processing. A subset of these participants (n=10, 4 female) also completed neuroimaging procedures. They ranged in age from 54 to 80 (mean=68), were between 19 and 276 months post-stroke (mean=102.7) and had a mean CAT Modality Mean T-score of 54.17.

Behavioral Tasks

A novel task utilizing naturalistic photographic stimuli and patient-friendly procedures (the Event task) was used to test event-related semantic knowledge. Participants were presented with images that depicted either typical events (n=130; left image, Fig. 1) or events that involved one or more abnormal event participants, such as a highly unlikely agent, theme, location, or instrument (n=130; right image, Fig. 1). These images were taken from an ERP study by Proverbio and Riva (2009) that found that the images of abnormal events elicited a robust centro-parietally distributed N400 response compared to the typical-event images in college-aged adults. In the Event task, participants had to decide whether the depicted events were normal (represented in semantic memory) or were abnormal (not represented in semantic memory).



Figure 1: Sample images from the Event task.

Participants also completed standardized measures of language performance and of action and object processing. Language performance was tested using the CAT (Swinburn, et al., 2004). Action processing was tested by the Kissing and Dancing Test (KDT: Bak & Hodges, 2003), whereas object processing was tested by Pyramids and Palm Trees (PPT: Howard & Patterson, 1992). KDT and PPT are picture-based tasks in which participants indicate which of two actions (KDT) or objects (PPT) best matches a reference action/object (e.g., a palm tree matches a pyramid better than a pine tree does).

Procedure

Event task stimuli were presented on a computer using E-Prime 2.0 (Schneider, Eschmann & Zuccolotto, 2012) and participants responded by pressing one of two keys on the keyboard. They were instructed to press a left-hand key marked with a '1' label if the image showed "something that might normally happen," and to press a right-hand key marked with a '5' label if it showed "something that might not normally happen." Participants had 5 seconds to respond. Each participant was given 4 practice trials with feedback. The remaining trials did not have feedback. KDT and PPT stimuli were also presented via E-Prime, with participants pressing the keys labeled with '1' and '5' to indicate whether the left-hand (1) or right-hand (5) image better matched the reference action or object. The CAT was administered and scored by a trained speech-language pathologist. The Action Naming and Object Naming subtests of the CAT were intended to measure verb and noun processing, respectively. The CAT, KDT, PPT, and the Event task were all administered in a single session.

Neuroanatomical Data

Scan Protocol

Magnetic resonance imaging procedures were performed at the Neuroscience Imaging Center (http://www.nic.pitt.edu/) using a Siemens Allegra 3T scanner. High-resolution T1-weighted structural images were collected from each participant using an MPRAGE sequence. The scan parameters for the T1 image were as follows: Field of View: 256mm x 256mm; 192 slices in sagittal plane, 1mm thick; Flip angle: 8°; TE/TR = 3.04/1540 ms; Voxel size: 1mm x 1mm x 1mm. In addition to T1 images, either high-definition 3D T2-weighted or 3D FLAIR images were collected for each participant. Scan parameters for the 3D FLAIR image were as follows: Field of View: 256mm x 220mm; 160 slices in sagittal plane, 1mm thick; Flip angle: 180° – 90° – 180°, IR sequence; TE/TR/TI = 353/6200/2000 ms; Voxel size: 1mm x 1mm x 1mm. Scan parameters for the 3D T2 image were as follows: Field of View: 256mm x 192mm; 160 slices in sagittal plane, 1mm thick; Flip angle: 90° – 180°, spin-echo sequence; TE/TR = 354/3500 ms; Voxel size: 1mm x 1mm x 1mm. All structural images were collected in a single scanning session.

Image Preprocessing

The acquired DICOM images were converted to NifTi format using MRIcron (Rorden & Brett, 2000). Brain lesions were demarcated directly on coronal slices of the T1 image using ITK-SNAP (Yushkevich et al., 2006). Lesion boundaries on coronal slices were compared to axial and sagittal slices, with T2/FLAIR images serving as reference images to help disambiguate lesion from cerebral-spinal fluid, ventricles, and white-matter atrophy. At least two tracers delineated the lesion boundaries in each brain. The tracers were blind to the individual's identity and language-impairment profile. Inter-tracer reliability for lesion tracings was computed by comparing the discrepant areas voxel by voxel in AFNI (Cox, 1996). Less than 30% discrepancy was achieved for all brains (range: 8% - 27%). All lesion tracings were then reviewed by the first author and an additional experienced tracer (Michelle Gravier, Ph.D.), who was blind to the individual's identity and profile. The lesion tracing that was judged by the first author and the additional tracer to be the best for each brain was selected as the consensus lesion tracing for that brain.

The consensus lesion tracings served as the lesion mask for each participant. All lesion masks were normalized using SPM8 (Ashburner et al., 2012) to an age-matched brain template (Rorden, et al., 2012). Then, cost-function masking was applied to demarcate the lesion boundaries so as to reduce 'lesion bleeding' during transformation (Bates et al., 2003; Brett, et al., 2001). Finally, each patient's brain scan was warped onto the older brain template using the Clinical Toolbox for SPM8. The transformation matrix obtained was

also applied to each patient's lesion mask, so that all the lesions were transformed to the template space.

Lesion Measures

Two LH ROIs were defined using the parcellations of the automated anatomical labeling (AAL: Tzourio-Mazoyer et al., 2002). These ROIs were derived from lesion-based and fMRI evidence regarding the neural correlates of object- and action-concept processing (see Table 1).

Table 1: ROIs used in analysis.

ROI	Anatomical regions in LH
Object	Middle temporal gyrus Inferior temporal gyrus Fusiform gyrus Temporal pole
Action	Precentral gyrus Postcentral gyrus Inferior frontal gyrus (pars triangularis, orbitalis, and opercularis) Paracentral lobule Superior temporal gyrus Superior parietal gyrus Angular gyrus Temporal pole

These ROIs were intended to measure the influence of object- and action-related semantic processing on performance in the Event task. In addition to these ROIs, additional ROIs were defined for: (1) the whole brain and (2) LH Heschl's gyrus. The whole-brain ROI was intended to measure the influence of overall lesion volume on performance. It served as a control variable in regression models examining the effects of specific ROIs above. The Heschl's gyrus ROI served as an additional control ROI. Because the Event task does not involve auditory stimuli, lesion to primary auditory cortex (Heschl's gyrus) should not affect Event-task performance.

The SPM toolbox MarsBaR (Brett, Anton, Valabregue, Poline, 2002) was applied to extract masks for the ROIs described above. These masks were intersected with lesion masks, to determine the overlapping area between the normalized lesions and the ROIs. The voxels of the overlapping areas were then used to calculate the lesion volume (proportion of lesioned voxels) in each ROI for each participant. The proportion of lesion in each ROI was used as predictors in lesion-deficit models.

Results

Behavioral Tasks

Data from the behavioral tasks came from the full set of 26 participants. Accuracy for the Event task ranged from 47.7% to 94.5%, with a mean of 86.2%. Accuracy was higher for normal events than abnormal events (88.3% vs. 84.1%), suggesting there may have been a 'yes' bias among our participants (e.g., Mitchum, Haendiges & Sloan Berndt, 2004). A logistic regression analysis of trial-level Event data in R using lme4 (Bates, et al., 2015) confirmed this, showing that stimulus type (normal versus abnormal) predicted the likelihood of a correct response (estimate=0.552; z=2.656, p<0.01). Given this bias, we used d-prime as our measure of Event task performance and included stimulus type as a nuisance variable in the lesion-deficit models reported below.

Correlation analyses examined the relationship between Event task performance and (1) action/object-processing deficits (KDT, PPT performance) and (2) language-processing impairments (CAT performance). Accuracy data for KDT and PPT and d-prime data for the Event task were z-score transformed. These z-scores were used for correlation analyses. Scores on the Event task were significantly correlated with both KDT and PPT, but more strongly correlated with KDT than PPT (KDT: r=.637, PPT: r=.526; both p<0.01). Event performance was compared to performance on the CAT Action Naming and Object Naming subtests. Event z-scores were significantly correlated with both Action Naming and Object Naming T-scores, but more strongly correlated with Action Naming than Object Naming (Action Naming: r=.411, Object Naming: r=.331; both p<0.05).

Neuroimaging Data

The distribution of lesions for the 10 participants who were scanned is depicted in Figure 2.

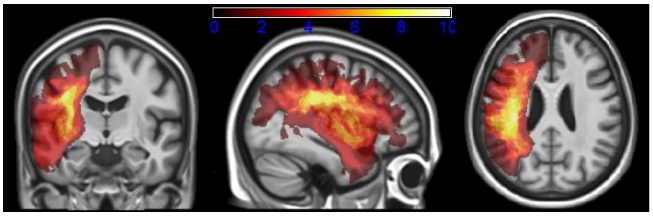


Figure 2: Lesion overlap for scanned participants

Logistic regression models examined the relationship between trial-level Event responses and lesions to control, action, and object ROIs. All models contained whole-brain lesion volume as a covariate, in order to measure the effect of lesions to a specific ROI while controlling for the effect of overall lesion volume. In addition, all models contained fixed effects of stimulus type (normal vs. abnormal), to control for the effect of response bias on Event performance, as well as stimulus-type-by-ROI interaction. The main effect of stimulus type was significant in all models.

As expected, there was no effect of lesion to the control ROI (Heschl's gyrus) on Event performance (estimate=0.878; z=1.090, p>0.2), and no interaction between control ROI lesion and stimulus type (estimate=0.556; z=-1.627, p>0.1). There was also no effect of lesion to the object ROI (estimate=-0.168; z=-0.078,

p>0.9), though there was a negative interaction of object ROI lesion and stimulus type (estimate=-1.846; z=-2.073, p<0.05). In contrast, there was a marginally significant effect of lesion to action ROI (estimate=10.42; z=1.674, p=0.094) and a significant interaction of action ROI lesion and stimulus type (estimate=-2.311; z=-1.973, p<0.05). Interestingly, the effect of action ROI lesion was positive: participants with larger lesions to the LH action ROI were *more* likely to respond accurately in Event, as shown in Figure 3. Similarly, the interactions of action and object ROI lesion and stimulus type (a measure of response bias) were negative: participants with larger action or object ROI lesions had a smaller response bias.

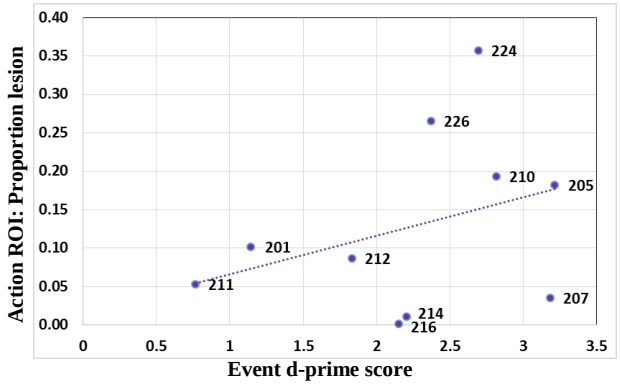


Figure 3: Relationship between Action ROI lesion and Event performance

Discussion and Conclusions

This study examined the neural bases of semantic-memory deficits for events in a sample of people with LH damage and aphasia, using a novel task (the Event task) that is picture-based and patient-friendly. Results from the behavioral tasks showed that Event-task performance is strongly correlated with performance on other picture-based measures of semantic memory, especially action-related measures (KDT). Furthermore, Event-task performance is correlated with deficits in language processing, especially verb processing (Action Naming on the CAT). These findings suggest that semantic memory representations for events may be closely aligned with action-concept representations and verb processing. This is consistent with evidence linking verb processing and event-related knowledge (e.g., Hare, et al., 2009).

Consistent with the behavioral evidence, lesion-deficit analyses showed that Event-task performance was most strongly related to lesions in LH action-related ROIs. However, this relationship was a negative one: greater LH lesion was associated with better Event-task performance, as well as a reduction in response bias. This finding suggests that damage to LH regions responsible for action processing may force participants to rely on complementary RH regions to perform the Event task. This LH damage may impair action-related language performance (such as verb processing), since language is a LH-localized function. However, it could result in better Event-task performance if event-related representations are partially or primarily RH-localized. This finding is thus consistent with evidence that event-related knowledge may be activated or represented primarily in the RH (Metusalem, et al., 2016).

Although the current findings suggest that event-related knowledge (measured by Event-task performance) is most closely related to action-concept processing, Event-task performance was also related to both object and noun processing. It was also somewhat related to damage to LH object ROIs. This pattern is consistent with the fact that events connect actions, objects, spatial, and temporal information. Further research with larger samples of participants is needed to elucidate the nature of these relationships.

Ackowledgments

This research was supported by the National Institutes of Health through grant R01DC011520 to the last two authors. It is the result of work supported with resources and the use of facilities at the VA Pittsburgh Healthcare System. Thanks to Michelle Gravier for assistance with lesion-tracing analysis, normalization, and visualization procedures, to Anish Kumar and Derrick Krieder for help with lesion-tracing procedures, and to Alice Proverbio and Federica Riva for sharing stimuli from Proverbio and Riva (2009).

References

Antonucci, S. M., & Reilly, J. (2008). Semantic memory and language processing: A primer. *Seminars in Speech and Language, 29*(1), 5-17.

Ashburner, J., Barnes, G., Chen, C. –C., Daunizeau, J., Flandin, G., Friston, K., … & Phillips, C. (2012). SPM8 manual. Functional Imaging Lab, Inst. of Neurology. http://www.fil.ion.ucl.ac.uk/spm/doc/maual.pdf.

Bak, T. H., & Hodges, J. R. (2003). Kissing and dancing- a test to distinguish the lexical and conceptual contributions to noun/verb and action/object dissociation. *Journal of Neurolinguistics, 16*(2), 169-181.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H. …Rcpp, L. (2015). Package "lme4." *Convergence, 12*, 1.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., & Dronkers, N. F. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience, 6*(5), 448-450.

Binney, R. J., Embleton, K. V., Jefferies, E., Parker, G. J. M., & Ralph, M. A. L. (2010). The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory. *Cerebral Cortex, 20*(11), 2728-2738.

Boylan, C., Trueswell, J. C., & Thompson-Schill, S. L. (2015). Compositionality and the angular gyrus: A multi-voxel similarity analysis of the semantic composition of nouns and verbs. *Neuropsychologia, 78*, 130-141.

Brett, M., Leff, A. P., Rorden, C., & Ashburner, J. (2001). Spatial normalization of brain images with focal lesions

using cost function masking. *Neuroimage, 14*(2), 486-500.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using an SPM toolbox. Presented at the 8th International Conference on Functional Mapping of the Human Brain (June, Sendai, Japan). Available on CD-ROM in *Neuroimage*, *16*(2).

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162-173.

Den Ouden, D. –B., Fix, S., Parrish, T. B., & Thompson, C. K. (2009). Argument structure effects in action verb naming in static and dynamic conditions. *Journal of Neurolinguistics, 22*(2), 196-215.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491–505.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44*(4), 516-547.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*(2), 151-167.

Hillis, A. E., Oh, S., & Ken, L. (2004). Deterioration of naming nouns versus verbs in primary progressive aphasia. *Annals of Neurology, 55*(2), 268-275.

Howard, D., & Patterson, K. E. (1992). *The Pyramids and Palm Trees Test: A test of semantic access from words and pictures.* Thames Valley Test Company.

Howard, D., Swinburn, K., & Porter, G. (2004). *Comprehensive aphasia test*. Psychology Press.

Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research, 32*(1), 37-55.

Kemmerer, D., Rudrauf, D., Manzel, K., & Tranel, D. (2012). Behavioral patterns and lesion sites associated with impaired processing of lexical and conceptual knowledge of actions. *Cortex, 48*(7), 826-848.

Kim, M., & Thompson, C. K. (2004). Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization. *Brain and Language, 88*(1), 1-20.

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 369*(1634).

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(4), 913.

Mätzig, S., Druks, J., Masterson, J., & Vigliocco, G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex, 45*(6), 738-758.

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass, 3*(6), 1417-1429.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language, 66*(4), 545-567.

Metusalem, R., Kutas, M., Urbach, T. P., & Elman, J. L. (2016). Hemispheric asymmetry in event knowledge activation during incremental language comprehension: A visual half-field ERP study. *Neuropsychologia*, 84, 252–271.

Mitchum, C., Haendiges, A., & Sloan Berndt, R. (2004). Response strategies in aphasic sentence comprehension: An analysis of two cases. *Aphasiology, 18*(8), 675–692.

Proverbio, A. M., & Riva, F. (2009). RP and N400 ERP components reflect semantic violations in visual processing of human actions. *Neuroscience Letters, 459*(3), 142-146.

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., & Karnath, H. O. (2012). Age-specific CT and MRI templates for spatial normalization. *Neuroimage, 61*(4), 957-965.

Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural neurology, 12*(4), 191-200.

Saur, D., Lange, R., Baumgaertner, A., Schraknepper, V., Willmes, K., Rijntjes, M., & Weiller, C. (2006). Dynamics of language reorganization after stroke. *Brain, 129*(6), 1371-1384.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2012). E-Prime v2.0. *Pittsburgh, PA: Psychology Software Tools.*

Squire, L. R. (1987). Memory and the Brain. In S. L. Friedman, K. A. Klivington, & R. W. Peterson (Eds.), *The Brain, Cognition, and Education* (pp. 171-202). Orlando, FL: Academic Press, Inc.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381-402). New York, NY: Academic Press, Inc.

Tzouriou-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subj. brain. *Neuroimage, 15*(1), 273-289.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*(3), 829-853.

Yee, E., Chrysikou, E. G., & Thompson-Schill, S. L. (2013). The cognitive neuroscience of semantic memory. *Oxford Handbook of Cognitive Neuroscience, Vol. 1*, 353-374.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage, 31*(3), 1116-1128.

Seeing the Bees Buzz and Hearing the Diamonds Glisten: The Effect of the Mode of Presentation of Stimuli on the Modality-Switch Effect

Elisa Scerrati
University of Bologna, Bologna, Italy

Luisa Lugli
University of Bologna

Anna M. Borghi
University of Bologna

Roberto Nicoletti
University of Bologna

Abstract: Previous studies showed that the sequential verification of different sensory modality properties for concepts (e.g., BLENDER-loud; BANANA-yellow) incurs a processing cost, known as the modality-switch effect (Pecher et al. 2003; 2004). We assessed the influence of the mode of presentation of stimuli on the modality-switch effect in a property verification priming paradigm. Participants were required to perform a property verification task on a target sentence (e.g., "butter is yellowish", "leaves rustle") presented either visually or aurally after having been presented with a prime sentence (e.g., "the light is flickering", "the sound is echoing") that could either share both, one or none of the target's mode of presentation and content modality. Results showed that the presentation and the content-driven effects were not cumulative. We conclude that the MSE is a two-fold effect which can occur at two different levels of information processing (i.e., perceptual and semantic).

Bilingual Proficiency Affects Inhibitory Control: A study of Stroop Performance in 8-year-old English-Chinese Singaporean Children

Carissa Kang
Cornell University, ITHACA, New York, United States

Abstract: Inconsistent results in the field of bilingualism and cognition may be largely influenced by variation in the nature of bilingual language proficiency. Here we explore the relationship between inhibitory control and bilingual proficiency in 43 8-year-old English-Chinese children in Singapore where bilingualism is prolific. Proficiency estimates are based on Oral and Written exam scores and caretaker estimates (including use and exposure). Children completed English and Chinese Stroop, where each task comprised 75% incongruent trials. Stroop effects were calculated for both languages. Higher English scores (written and oral) and English use predicted smaller English Stroop interference. Conversely, higher Chinese exposure and use predicted smaller Chinese Stroop interference. Thus, language proficiency, use and exposure influence inhibitory control, reiterating the need to consider bilingual proficiency when studying the relationship between bilingualism and attention. Since Stroop effects differ depending on language, bilinguals should be tested in both languages for verbal EF tasks.

Linguistic recursion and Autism Spectrum Disorder

Irina Polyanskaya
Roskilde University

Torben Braüner
Roskilde University

Patrick Blackburn
Roskilde university

Abstract: Both first-order and second-order false-belief mastery are important in acquisition of Theory of Mind. Our logical analysis of second-order false-belief tasks shows that this sort of reasoning involves recursion. Language involves recursion as well; recursive possessive and complements clauses are examples.

Second-order social reasoning depends on both individual cognitive resources and immersion in a wide range of interactive contexts. But since the 'usual' interactive contexts do not make the same sense to children with Autism Spectrum Disorder (ASD), it has been proposed that they use language as scaffolding in false-belief understanding.

We hypothesize that competency in linguistic recursion predicts second-order false-belief mastery for children with ASD. We investigate this by training children with ASD to better comprehend and produce recursive possessive and complement clauses. We have developed and validated a tool to measure the recursion competency in the Danish language, and we apply this in a randomized controlled training study.

Computational explanation of "fiction text effectivity" for vocabulary improvement: Corpus analyses using latent semantic analysis

Keisuke Inohara
The University of Electro-Communications

Akira Utsumi
The University of Electro-Communications

Abstract: Previous studies have suggested that fiction book reading has a stronger positive effect on vocabulary development than nonfiction. In this study, we examined this phenomenon in terms of word appearance information in fiction (story texts), nonfiction (explanation texts), and web text using latent semantic analysis (LSA). In a human experiment with Japanese undergraduates, we replicated fiction (story) text effectivity. Participants who often read story texts achieved the highest vocabulary test scores. Then, in a corpus experiment, we constructed a story text corpus, explanation text corpus, and web text corpus of identical size. Based on these corpora, we calculated the LSA similarities between words, and simulated answering the same vocabulary test as used in the human experiment. The corpus experiment demonstrated the nonfiction (explanation) text effectively, that is, the explanation corpus was the highest. The cause of discrepancy in the results and the educational implications of this study were also discussed.

Interaction, abstraction and complexity

Kristian Tylén
School of Communication and Culture, Aarhus University

Riccardo Fusaroli
School of Culture and Society, Aarhus University

Pernille Smith
School of Business and Social Sciences, Aarhus University

Jakob Arnoldi
School of Business and Social Sciences, Aarhus University

Abstract: In an experimental study, we test the hypothesis that constitutive properties of social interaction - such as diversity in cognitive styles, knowledge and experience - enhance cognitive processes of abstraction. Through three sessions, individuals and dyads categorized aliens based on combinations of features such as the shape and color of their body parts. We manipulated relations among the features to elicit increasingly complex categories. Furthermore, to assess the character of participants' evolving categories, after each training session they were presented to a new test set of aliens that differed in appearance, but shared relations among features with aliens from the training set. We found that dyads outcompete individuals in categorization accuracy across levels of complexity. We also found that this effect is due to the more abstract and rule-based character of dyads emerging representations as evidenced by their performance on test items.

Using Nature Language Processing to Improve Optical Character Recognition

Lining Xu
Tsinghua University

Yongxu Wu
Beijing Information Science and Technology University

Abstract: OCR (Optical Character Recognition) has developed over 100 years. However, if the document or picture is stained, it could not work well. Considering that words in text can be connected by logical relationship, with the help of the idea that reducing the size of word stock which references from license plate recognition, this paper established N-GRAM model, used the results of Google search engine to improve its text sparsity. The use of residual features of the original stained characters can improve the recognition rate and accuracy with the help of a smaller size of the word stock successfully.

Artificial Language Learning: The Context of Negation

Ariel Mathis
University of Memphis

Stephanie Huette
University of Memphis

Abstract: Words are learned in various contexts and over various timescales throughout our lives. The current study explored the role of context in the form of negation in artificial language learning. It was predicted that words trained in an entirely negated context would show lower average correctness in the testing phase than those trained entirely in the affirmative or in the combined contexts. Eighteen artificial nouns were trained using the prefixes "an-" meaning "not the" or "o-" meaning "the" to mark negation. In the testing phase, participants were tested without the prefix on word stems only. Findings indicated words learned solely in the affirmative context led to a higher average correctness while those learned solely in the presence of negation showed the least average correctness in the testing phase.

Source Expertise and Question Type Effects in Conversation-Based Assessment

Jesse R. Sparks
Educational Testing Service

Jessica J. Andrews
Educational Testing Service

Diego Zapata-Rivera
Educational Testing Service

Blair Lehman
Educational Testing Service

Kofi James
Educational Testing Service

Abstract: Conversational discourse is a cognitive and social process influenced by both discourse content and pragmatic factors, such as the participants' prior knowledge; these factors may also affect how simulated conversations with virtual agents unfold, with implications for design. This study explored effects of question content and perceived expertise of a virtual agent on students' interactions with a conversation-based assessment (CBA) measuring science inquiry skills. Twenty-four middle school students were randomly assigned to work with a High- or Low-Knowledge virtual peer to collect data and generate weather predictions. Students evaluated their own data relative to the peer's; they could either "Choose" which note to keep, or to "Agree/Disagree" with the peer's suggested choice of note. Students rated the peer as more expert in the High-Knowledge condition, but peer expertise did not affect performance. However, the Agree/Disagree condition improved students' accuracy in their note choice, and yielded marginally higher pre-post learning gains.

Implicit updating of object representation via predictive associations

Ru Qi Yu
University of British Columbia, Vancouver, BC, Canada

Jiaying Zhao
University of British Columbia, Vancouver, BC, Canada

Abstract: An adaptive function of the visual system is that it flexibly updates existing representations of objects upon changes. Such updating can also alter the representations of associated objects that are not directly observable. What mechanism supports this process? We propose that statistical learning provides a channel through which changes in one object are automatically transferred to related objects. Observers viewed a temporal sequence of paired circles. One circle in each pair then changed in size, and observers recalled the size of the other circle. When the first circle enlarged (or shrank), the second circle was judged to be larger (or smaller), suggesting that the change was automatically transferred to the predicted object (Experiment 1). The same, however, was not true if the second circle changed in size (Experiment 2). No observer was explicitly aware of the circle pairs. Thus, statistical learning enables the implicit updating of representations through predictive associations.

Roles of Metacognitive Suggestions in Hypothesis Revision

Sachiko Kiyokawa
Nagoya University

Kazuhiro Ueda
The University of Tokyo

Yoshimasa Ohmoto
Kyoto University

Abstract: We investigated whether metacognitive suggestions alone can function as a means of encouraging recipients' reflection and facilitate hypothesis revision. We also examined whether it is necessary to ground the suggestions on the recipients' thinking processes for the facilitative effects. 108 participants were assigned to one of the four conditions: metacognitive suggestion collaboration with/without grounding support; free collaboration; and sole. They were asked to engage in a rule discovery task. The task was designed so that hypothesis revision was necessary for the participants to find the correct rule. The results showed that performance both in the free collaboration and in the metacognitive suggestion collaboration with grounding support conditions was higher than that in the solo condition. We concluded that metacognitive suggestions alone can facilitate hypothesis revision as well as free collaboration and that grounding support is necessary for the facilitative effect to be obtained.

Tactile protocol analysis: Observations of novices reading data tables by touch

Peter Cheng
University of Sussex

Panayiota Polycarpou
Cyprus University of Technology

Grecia Garcia Garcia
University of Sussex

Frances Aldrich
University of Sussex

Abstract: The "Cognitive Science of Tactile Graphics Project" at the University of Sussex is studying how experienced and novice users of tactile graphics read diagrams by touch. One goal of the project is to design novel tactile formats that specifically support ready access to, and rich interpretations of, information in tactile materials by people with visual impairment. This presentation focuses upon the development of a Tactile Protocol Analysis (TPA) method for the transcription, segmentation, coding and interpretation of tactile graphic reading behaviours. TPA is challenging because both hands may be performing separate actions over different timescales and fingers of one hand may themselves be performing different actions. To initially demonstrate the utility of the method, 24 novice tactile readers' performance on a data table search task was analyzed, from which hypotheses were formulated about the impact of object recognition skill on overall patterns of search.

In the event of a turn exchange: Visual information the perception of turn-taking behavior in natural conversation

Nida Latif
Queen's University, Kingston, Ontario, Canada

Agnès Alsius
Queen's University, Kingston, Ontario, Canada

K.g. Munhall
Queen's University, Kingston, Ontario, Canada

Abstract: Everyday conversation is composed of a rapid exchange of turns between talkers as they communicate. The speed of these exchanges implies simultaneous perception and production of conversational cues relevant to turn-taking behavior. Natural face-to-face conversation involves a rich set of these social cues including visual information whose contribution to perception of turns has yet to be examined. Our studies investigated the influence of visual information in perceiving a turn exchange. We examined the time-course of the use of these visual cues during turn judgments. Results show that visual information is sufficient but not necessary for perceiving turn exchanges. Further, the temporal precision with which auditory cues influence turn perception is greater than that of visual information. We suggest that although auditory cues dominate the perception of turn exchanges, reliance on the various sources of information is flexible and may follow highly sensitive timelines.

An Eye For Figurative Meaning: The Effects of Familiarity on Metaphor Comprehension

Chris Genovesi
Carleton University , Ottawa , Ontario, Canada

Michael Vertolli
Carleton University, Ottawa, Ontario, Canada

Abstract: The career of metaphor hypothesis suggests that processing preference is a result of conventionality whereby conventional metaphors are processed through categorization, and novel ones processed through comparison. Alternatively, the categorization model predicts that apt metaphors are processed as categorizations whether or not they are conventional. However, research has largely ignored another known factor to influence metaphor processing, namely familiarity. The categorization model predicts familiarity to play no role in deciding on processing strategy. On the other hand, the career of metaphor hypothesis predicts that familiarity to play a facilitating role in metaphor comprehension. In this experiment, we used the eye tracking paradigm and controlled for aptness and conventionality, and manipulated familiarity in order to test these predictions. Our initial results support the career of metaphor hypothesis suggesting that familiarity facilitates metaphor processing. We discuss the implications these results have on the psycholinguistic models and briefly speculate on their philosophical consequences.

The Concept of Transcendent God and the Emergence of Cognitive Level D in the Four-Level-Cognitive-Development Theory

Vladimir Glebkin
RANEPA

Abstract: The author addresses the emergence of cognitive level D in the Four-Level-Cognitive-Development Theory (FLCDT). He argues that the concept of transcendent God, 'Maker of Heaven and Earth and of all things visible and invisible' was an important factor for that. The problem of God's qualities and the view of numbers in medieval Christian and Arabic philosophy are examined as some arguments for the key point of this paper.

Implicit versus explicit language learning: Differential effects of working memory and learning styles

Sarah Grey
Fordham University Pennsylvania State University

Christine Schoetensack
Lancaster University

Kimberly Bell
Lancaster University

Padraic Monaghan
Lancaster University

Patrick Rebuschat
Lancaster University

Abstract: Understanding the process of adult language learning has recently undergone advances due to the consideration of how individual differences (IDs) in cognitive processing, such as working memory (WM), affect acquisition. We know that implicit versus explicit learning conditions also influence learning, however, the potential interactions between IDs, the efficacy of implicit versus explicit learning, and different types of linguistic information are largely unknown. In this study, we tested learning of syntax and grammatical case under two conditions: incidental and explicit rule-provision ("instructed"). We also assessed individuals' WM, phonological working memory (PWM), and learning styles. Significant learning effects were found for word order and case in both learning conditions. For case, but not for word order, the instructed group outperformed the incidental group. Regarding IDs, incidental learning of case was marginally related to individuals' WM; instructed learning of case was related to PWM. For learning styles, there was a negative relationship between learning of word order in the instructed condition and a deductive learning style. These results reveal the complex relationships among cognitive processes in explicit and implicit language learning across different aspects of language structure, and in relation to cognitive IDs.

Enhancing Creativity in Children by Imparting Chess Training

Ebenezer Joseph
University of Madras, Chennai, India

Veena Easvaradoss
Women's Christian College, Chennai, India

Suneera Abraham
Emmanuel Chess Centre, Chennai, India

Michael Brazil
Emmanuel Chess Centre, Chennai, India.

David Chandran
Emmanuel Chess Centre, Chennai, India.

Abstract: Creativity is the ability to find hidden patterns, to make connections between seemingly unrelated phenomena, and to generate solutions. The present study, funded by Indian Government, analysed the effect of one year chess training program on the creativity of school-going children of both genders. A pre-test post-test with control group design was used. The sample comprised 64 children: 32 children each in experimental (Mean age=11.86, SD=1.44) and control (Mean age=12.03, SD=1.14) groups. Children in the experimental group underwent weekly chess training with Winning Moves Chess Curriculum. Creativity was assessed by Indian adaption of Wallach-Kogan Creativity Test. Pre–intervention equivalence of the two groups was established. Independent t test analysis revealed that the experimental group had statistically significant gains in total creativity and Instances and Alternate Uses subtests compared to the control group. The authors conclude that systematic chess intervention increases creativity in children. The educational implications are significant.

The role of grammatical form in generalizing principled and statistical properties

Kim Kristina Fuellenbach
Oxford University

E. Matthew Husband
Oxford University

Abstract: Asymmetric generalization patterns for definite and indefinite singular generics

In two experiments, we investigated the role of grammatical form in inferring the conceptual status of properties in generic sentences. We trained participants on novel properties in pseudoword sentences with bare plural (BP), indefinite (IS) and definite singular (DS) subjects. In the test phase, participants rated the relationship between trained properties and novel subjects: We found that, compared to BP, properties trained in the IS condition increase expectations of principled connections whereas DS-trained properties diminish expectations of statistical connections. BP subjects acted as a control since they were equally judged to be statistically or principally connected. These results support the theoretical claim that IS have quantificational force and DS directly refer to kinds. They leave open the puzzle of the nature of BP subjects which seem to be ambiguous between the two and also the only one to convey statistical connections.

Priming Spatial Reference Frames for Memory

Elena Andonova
New Bulgarian University

Abstract: We examined the role of spatial frames of reference in memory. Participants first verified verbal descriptions of visual scenes and spatial relations among objects. The intrinsic and relative frames of reference (FoR) were used in the descriptions with varying degrees of frequency (availability) and veridicality (reliability). Descriptions in the two reference frames could either be equally distributed in terms of validity or were biased towards one of the two spatial frames. Participants' performance on the memory task was sensitive to priming from the spatial FoRs and their information distribution characteristics. These findings provide evidence that spatial frames of reference can influence spatial memory and that this influence depends on the frequency of use of a given frame of reference and on the frequency with which it is associated with valid and reliable information.

Deep Learning and Attentional Bias in Human Category Learning

Leyla Roksan Caglar
Rutgers University, Newark

Stephen Jose Hanson
Rutgers University, Newark

Abstract: Human category learning is known to be a function of both the complexity of the category rule and attentional bias. A classic and critically diagnostic human category problem involves learning integral stimuli (correlated features) using a condensation rule, or separable stimuli (independent features) using a filtration rule. Human category learning shows differential learning based on category rules that either require attentional binding or ignoring features. It has been shown that neural networks learning with backpropagation cannot differentially learn or distribute attention without built in perceptual bias. In effect neural networks fail to integrate the complexity of learning with the representational bias of the stimuli. In this paper we show that Deep Learning networks, through successive re-encoding and the development of more sensitive feature detectors, learn both the category rules while modeling the attentional bias consistent with the human performance in a task of categorizing realistic 3D-modeled faces.

Re-presenting a Story by Emotional Factors using Sentiment Analysis Method

Hwiyeol Jo
Seoul National University

Yohan Moon
Sungkyunkwan University

Jongin Kim
Seoul National University

Jeong Ryu
Yonsei University

Abstract: Remembering events is affected by personal emotional status. We examined the psychological status, personal factors, and social factor of undergraduate students (N=64) and got summaries of a story, Chronicle of a Death Foretold from them. As transfer learning, we collected 38,265 movie review data to train a sentimental analysis model based on convolutional neural network, using the model to score each summary. The results of CES-D and PANAS show the relationship between emotion and memory retrieval; depressed people have shown a tendency of representing a story more negatively, and seemed less expressive. People with full of emotion have retrieved their memory more expressively, using more negative words. The contributions of this study can be summarized as follows: First, we lighten the relationship between emotion and its effects on storing or retrieving memories. Second, we suggest objective methods to evaluate the intensity of emotion in words, using a sentimental analysis model.

Reexamining the Unaccusative Hypothesis: a Visual World Paradigm study

Yujing Huang
Harvard University, Cambridge, Massachusetts, United States

Laine Stranahan
Harvard University

Jesse Snedeker
Harvard University

Abstract: The Unaccusative Hypothesis (UH) predicts that the subject of an unaccusative is mentally reactivated in the object position. Previous psycholinguistic studies have reported evidence of reactivation (Friedmann et al 2008, Koring et al 2012). However, these studies did not equate the unaccusative and unergative stimuli, resulting in confounds which jeopardize their conclusions. We reexamined UH with two Visual World Paradigm experiments carefully controlled for the potential confounds in the previous studies. On each trial, participants (n=40; n=52) saw 4 black-and-white drawings and heard a sentence. In the test condition, but not the control condition, one image was semantically related to subject of the sentence. We measured the proportion of looks to the target image at three time regions after the verb onset and found a robust match-effect (p's$<$.05) but no differences between the unaccusative and unergative conditions. We will discuss the consequences of our findings for syntactic theory.

Negation affects processing of correct and incorrect information: A visual-world paradigm for misinformation

Jeffrey Viaud
University of Memphis

Stephanie Huette
University of Memphis

Abstract: The current study investigated how lexical priming and negation affects encoding and retrieval of information. Studies have shown people encode and retrieve misinformation from memory, but the mechanisms of encoding and retrieval are not well understood. To address this, an eye-tracking paradigm was designed to examine probabilistic activation during retrieval of accurate or inaccurate information. Participants read four different kinds of texts that varied by if they were affirmative or negated and whether they contained accurate or inaccurate information. After participants read all texts, eye fixations were tracked in a visual world paradigm with four plausible answers on screen in each corner to choose from. Suppression was observed in groups that did not produce misinformation. When participants answered correctly, despite reading misinformation, we observed misinformation being inhibited instead of primed. Mechanisms of processing true and false information and the interplay between language and conceptual formation are discussed.

Lexicalization Typology of Realization Events in Mandarin Chinese

Yu Deng
School of Foreign Languages, Beihang University, Beijing, China

Fuyin Li
School of Foreign Languages, Beihang University, Beijing, China

Abstract: There has been a hot debate on the typological status of Mandarin Chinese in Talmyan framework of Verb-framed languages (V-languages) and Satellite-framed Languages(S-languages). However, most previous studies focus on motion events, while other macro-events (Talmy, 2000) receive little attention. The present study aims to investigate event of realization in Mandarin Chinese with experimental method. The analysis of elicited data shows that: (1) predicates of Mandarin realization events are mostly bipolar resultative verb compounds, which have the semantic feature of [+agent], [+instrument], and [+state change]. This proves that "result" is a semantic prime in Chinese verb semantics. (2) Lexicalization patterns of realization events in Mandarin represent more of S-language, but Mandarin also shows the characteristics of V-language. The difference between S-language patterns and V-language patterns is significant, and the general tendency is: S-language>V-language. Overall, the results indicate that the lexicalization typology of Mandarin realization events falls into a complementary typological framework.

An Analogical Model of Pretense

Irina Rabkina
Northwestern University

Ken Forbus
Northwestern University

Abstract: Pretense has been implicated as playing a role in the development of many cognitive and social skills. Despite the importance and ubiquity of this phenomena, few computational models of pretense exist. We propose a model of pretense via analogical abduction. We suggest that pretense occurs via structural alignment. Where a mismatch occurs (i.e. something in the pretend scenario is not the same as its aligned match in an equivalent real-world scenario), a re-representation must take place in order for pretense to continue. For example, a seashell may be re-represented as a cup (as in Fein, 1975) or an empty cup may be re-represented as full (as in Onishi et al., 2007). We show that this model can explain results from two empirical studies, including failures in pretense.

A Model of Language-Guided Concept Formation using a Common Framework for Unsupervised and Supervised Learning

Paul Munro
University of Pittsburgh

Abstract: A general learning rule, "BCM-δ", is proposed that subsumes both unsupervised learning as a form of the BCM rule (Bienenstock, Cooper, Munro, 1982; Munro, 1984) and the delta rule (Rosenblatt, 1958; Rumelhart, Hinton, and Williams, 1986). The "BCM-δ" unit is composed of two subunits, T and L, each integrating distinct input streams across distinct sets of synapses. The two subunits follow a common Hebb-like learning procedure that reduces to an unsupervised rule for the T subunit and a supervised rule for the L subunit in which the T response is the training signal. This model suggests a neurally plausible mechanism for the shaping of concepts by labels. More generally, stimuli from one modality can shape the response properties of a unit to another modality using a framework that is biologically plausible and gives clues to the source of a teaching signal for supervised learning.

Learning Spiking Neural Controllers for In-Silico Navigation Experiments

Mahmoud Akl
Technical University of Munich

Florian Walter
Technical University of Munich

Florian Röhrbein
Technical University of Munich

Abstract: Artificial neural networks have been employed in many areas of cognitive systems research, ranging from low-level control tasks to high-level cognition. However, there is only little work on the use of spiking neural networks in these fields. In this project, we developed a virtual environment to explore solving navigation tasks using spiking neural networks. We first used an existing experimental setup and compared the results to validate the developed environment. An evolutionary approach is used to set the parameters of a spiking neural network controlling a robot to navigate without collisions. In a second set of experiments, we trained the network via reinforcement learning which was implemented as a reward-based STDP protocol. Our results validate the correctness of the developed virtual environment and demonstrate the usefulness of using such a platform. The virtual environment guarantees the reproducibility of our experiments and can be easily adapted for future research.

Why are we (un)systematic? the (empirical) costs and benefits of learning universal constructions

Steven Phillips
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan

Yuji Takeda
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan

Fumie Sugimoto
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan

Abstract: A theoretical challenge for cognitive science is to explain both the presence and absence of systematicity. One explanation (Phillips & Wilson, 2010) says systematicity derives from universal constructions. We tested this theory with an experiment that required learning cue-target pair maps whose underlying structures were either products (universal construction), or non-products (control). Each series was learned in either ascending or descending order of size: number of unique cue/target elements constituting pairs, which varied from three to six. Only performance on the product series was affected by order: systematicity was obtained universally in the descend group, but only on large sets in the ascend group. The results suggest that learning small maps directly, without reference to the underlying product, may be perceived as more cost-effective, i.e., acquisition of a universal construction, hence systematicity, depends on an empirical cost-benefit tradeoff.

Reasoners are influenced by conversational pragmatics in abstract conditional reasoning tasks

Paul Thibodeau
Oberlin College

Daniel Grodner
Swarthmore College

Abstract: People demonstrate systematic logical failures when reasoning about conditional statements. In the Wason selection task, a test typically interpreted as a measure of abstract deductive reasoning, only about 10% of participants choose the cards prescribed by deductive logic. One possibility is that people are simply bad at hypothesis testing – biased toward confirming rather than falsifying abstract conditional rules. A second possibility, however, is that performance on the task is strongly influenced by pragmatic effects of linguistic interpretation. In three experiments, we find that manipulating the instructions to emphasize falsification and that changing the formulation of the rule to increase the pragmatic salience of the correct choices improves performance. These results arise because people do not merely decode the logical content of linguistic expressions. Rather they attempt to understand the communicative intentions of the individual who produced the expression even in abstract reasoning tasks.

A Brain-Based Feature Model of Adjective-Noun Composition

Elizabeth A. Shay
University of Rochester

Rajeev D. S. Raizada
University of Rochester

Abstract: Brain-based features of meaning (sensory-motor features: sound, color, manipulation, motion, and shape) are used to compare two popular models of adjective-noun semantic composition: element-wise vector addition and multiplication. A large literature (e.g. Fernandino et al., 2015) suggests that perceptual systems contain information that can be extracted using neural decoding (e.g. Anderson, Murphy & Poesio, 2014). Using Amazon's Mechanical Turk, participants rated how much each of the words and phrases (made of all combinations of the selected adjectives and nouns) evoked the features. Both multiplication and addition surpass chance at matching the correct phrase, but addition outperformed multiplication (addition = 7.6/60, multiplication = 13.4/60). Addition allows the adjective to weight the important sensory-motor attributes for the noun. Based on these behavioral results, we predict, and will test in upcoming work, that addition will also be successful when using brain activity (from fMRI) as the representations of the adjectives, nouns, and phrases.

Experience Representation of Artificial Cognitive System in Interaction with Real World

Liwei Zhang
School of Mechanical Engineering and Automation, Fuzhou University

Zhen Deng
TAMS, Department of Informatics,University of Hamburg

Jianwei Zhang
TAMS, Department of Informatics,University of Hamburg

Zhixian Chen
Shenzhen Institutes of Advanced of Technology, Chinese Academy of Sciences

Ying Hu
Shenzhen Institutes of Advanced of Technology, Chinese Academy of Sciences

Abstract: In this paper, we propose a novel experience representation approach for artificial cognitive system (such as a robot). The artificial cognitive system with the ability to store experiences and to adapt plans and behavior according to experiences will be beneficial for understanding the human representation of experience and be useful for developing practical service robot. Here an artificial cognitive system experience is defined as a record about the events occurred in the past. Three kinds of experiences (ontology, robot activities and environment activities) are introduced. In this work, we demonstrate the mobile cognitive system with a PR2 platform in a restaurant environment and a corresponding simulation environment. Four different scenarios (Serve-A-Coffee, Deal-with-Obstacles, Clear-Table and Well-set-Table) have been set to demonstrate the performance and collect the corresponding experiences.

Is it Living? Insights from Modeling Event-Oriented, Self-Motivated, Acting, Learning and Conversing Game Agents

Martin Butz
University of Tübingen, Tübingen, Germany

Mihael Simonic
University of Tübingen, Tübingen, Germany

Marcel Binz
University of Tübingen, Tübingen, Germany

Jonas Einig
University of Tübingen, Tübingen, Germany

Stephan Ehrenfeld
University of Tübingen, Tübingen, Germany

Fabian Schrodt
University of Tübingen, Tübingen, Germany

Abstract: A cognitive architecture is presented, which combines insights from artificial intelligence with cognitive psychology, biology, and linguistics. Using a Super Mario clone, we equipped the simulated agents with (i) motivational behavioral systems, (ii) reasoning and planning capabilities, (iii) event-based schema learning and sensorimotor exploration, and (iv) speech comprehension and generation mechanisms. The motivational system activates goal events to maintain internal homeostasis. To invoke selected events, hierarchical action planning and control unfolds both on an event-schematic and a sensorimotor level. Schema learning is based on the detection of event changes, which are not predicted by the basic sensorimotor forward model. Language is comprehended and generated using context-free grammars linked to the schema-based knowledge structure. The work offers an approach to develop and thus to ground conceptual, semantic world knowledge in sensorimotor interactions and to couple this knowledge with a language to generate and comprehend language about the agent's virtual world meaningfully.

Cognitive Sciences Strategies for Futures Studies (Foresight)

Ahmad Mahdeyan
Strategic Management School, Supreme National Defense University, Tehran, Tehran, Iran

Abstract: It seems before understanding the mind structure and laws governing the process of perception, we've gone into the application of science to build better futures. This is because the science begins with attitude not the reality.The major goal of this study is to identify qualitative relationships among key variables across of Cognitive Sciences (CS) with Futures Studies (FS).Ideas and subjects that explained here are just introduction to this research main goal. As a conclusion, by some strategies of CS for FS theFuture Oriented Artificial Intelligence Machine as a science fiction ideahas been introduced.

Extending the Talmyan Typology: A Macro-event as Event Integration and Grammaticalization

Fuyin Li
Beihang University (BUAA)

Abstract: This article addresses the macro-event, a fundamental concept for the Talmyan two-way typology. It proposes a diachronic aspect of macro-event, an aspect that seems to be under-appreciated or even neglected. It argues that a macro-event integrates two simple events through grammaticalization. This hypothesis is supported by the behaviors of directional verbs in Mandarin Chinese in that these forms alone can express almost all five types of macro-events that Talmy analyzed, and that these macro-events themselves represent an integration of two simple events, and exhibit various degrees of grammaticalization. This study relates two seemingly unrelated areas of research, i.e. event structure and grammaticalization and provides a new perspective to the Talmyan typological paradigm.

Perceptual Decision Making of Humans and Deep Learning Machines: a Behavioral Study

Ge Chen
Beihang University

Yuzhong Liu
Beihang University

Qiyang Zhao
Beihang University

Abstract: Human visual perception system is a key issue of the cognitive researches. It is also an inspiring prototype of the cutting-edge artificial intelligence researches - deep learning. It is interesting to investigate the behaviors of humans and deep learning machines on vision tasks. In this paper, we focus on the perceptual decision making and object recognition on distorted images. We found that in a wide range of distortion levels, the recognition rates of human subjects are smoothly increased along with the decreases of distortions. Although the deep learning machines perform obviously worse than human subjects, their recognition rates vary with the similar trends. It indicates that the deep learning machines make a good simulation to the human being on the perceptual decision making tasks.

Age Related Differences in Episodic Memory Recollections: Applying Latent Dirichlet Allocation to Free-Writings on Driving Incidents by Older and Young Drivers

Ritsuko Iwai
RIKEN, Wako, Saitama, Japan

Ritsuko Iwai
Kyoto University

Takatsune Kumada
Kyoto University

Takatsune Kumada
RIKEN, Wako, Saitama, Japan

Daisuke Kawahara
Kyoto University

Sadao Kurohashi
Kyoto University

Abstract: Reporting driving incidents depends on episodic memories formed at a certain time point in the past, and its retrieval with subjective feelings. This study examined aging effects on episodic memory recollections by analyzing free writing reports by older and young drivers. Unstructured hand-writing samples from 199 older (Mage = 69.2) long-experienced (Mdriving = 43.0 years), and 299 young (Mage = 21.5) novice (Mdriving = 2.2 years) drivers were avalyzed by Latent Dirichlet Allocation. This identified a 6-topic model labeled: (1) operational failures, (2) control aspects, (3) other vehicles, (4) jump-outs, (5) traffic lights, and (6) attention. Posterior distribution analysis revealed that older drivers reported less in topics concerning own driving operations. In addition, older drivers less attributed these topics to self than environment relative to young drivers. The age differences of episodic memory retrieval for free reports and applicability of natural language processing to psychology are discussed.

The Temporal Cheerleader Effect: Attractiveness Judgments Depend on Surrounding Faces Through Time

Alexia Toskos Dils
Purchase College, SUNY, Purchase, New York, United States

Scott Niedopytalski
Purchase College, SUNY, Purchase, New York, United States

Jeffrey Arroyo
Purchase College, SUNY, Purchase, New York, United States

Stephen Flusberg
Purchase College, SUNY, Purchase, New York, United States

Abstract: Previous research has found that people are seen as more attractive when they appear in a group rather than in isolation. Do faces that surround us in time also affect how attractive we seem? Participants rated the attractiveness of famous female faces presented in a sequence of three and in isolation. We found that people do integrate information about attractiveness over time, but that temporal context has the opposite effect of static context. People perceived faces as less attractive in a series than when those same faces were presented in isolation. We also varied the attractiveness of surrounding faces in order to examine how the serial position of contextual information figures into people's judgments of a face. Faces presented early in the sequence figured more heavily into people's judgments than did faces presented at the end. These findings highlight the role that temporal context plays in the perception of attractiveness.

Information Acquisition: Stopping Rules For Varying Levels of Probabilities and Consequences

Gustavo Gauer
Universidade Federal do Rio Grande do Sul

Roberto Nonohay
University of Michigan; Universidade Federal do Rio Grande do Sul

Richard Gonzalez
University of Michigan

Guilherme Lannig
Universidade Federal do Rio Grande do Sul

Abstract: We performed an exploratory experiment aiming to assess the use of stopping rules in information acquisition. Participants were requested to make a decision or procrastinate on 24 economic/financial scenarios after buying information pieces. Behavioral and EEG data were recorded for analysis. Results showed that participants decided according to Bayesian calculations to stop information acquisition and decide. Moreover, information acquisition strategies seemed consistent with prospect theory, with participants weighing information pieces differently and seeking more or less information given different manipulations in scenario probability, consequence valence and intensity. EEG data suggests a lateralization at frontal electrode sites. With probabilities stated, low negative consequence scenarios showed a positive peak at F3 around 200 ms before a decision was made. When probabilities were not stated, high positive consequences scenarios evoked a negative deflection at F4 around 400 ms before a decision.

Categorization of Probability Word Problem: Effects of Prior Statistical Training and Semantic Schema

Chenmu Xing
Teachers College, Columbia University, New York, NY, USA

James Corter
Teachers College, Columbia University

Doris Zahner
Council for Aid to Education

Abstract: A problem sorting task was used to examine how the semantic content of probability word problems affects problem understanding and categorization, for students with various levels of statistical training. In the task, undergraduate and graduate students were asked to sort probability problems into groups by similarity of solution. The problems varied by relevant probability principle, by type of semantic schema, and by cover-story surface content. Results showed that both less-trained students and more-trained students tended to sort problems by relevant probability principle, but students with more statistics training did this more consistently. Both groups of students tended to be affected in the sorting task by semantic schema, defined here as intermediate-level abstractions of the problem structure. For example, when a permutation problem described assignment of people to people, students showed a strong tendency to group it with independent-events problems with a people-to-people matching schema.

Differential Processing for Actively Ignored Pictures and Words

Scott Sinnett
University of Hawaii at Manoa, Honolulu, Hawaii, USA

Maegen Walker
University of Hawaii at Manoa, Honolulu, Hawaii, USA

Margeaux Ciraolo
University of Hawaii at Manoa, Honolulu, Hawaii, USA

Andrew Dewald
University of Hawaii at Manoa, Honolulu, Hawaii, USA

Abstract: Previous work suggests pictures may be processed more readily than words, likely because pictures appear to maintain more direct access to semantic and conceptual representations (Amit, Algom, & Trope, 2009). However, it is unclear how words and pictures may be processed differently when they are actively ignored. Our earlier work demonstrated a facilitated recognition for actively ignored words, provided they appeared frequently with an attended target in a previously presented repetition detection task (Dewald, Sinnett, & Doumas, 2012). The current study adapted this paradigm to examine the extent to which unattended pictures may be processed under analogous conditions. Overall, ignored pictures were recognized more often than ignored words. Moreover, recognition for ignored pictures did not benefit from target-alignment whereas ignored words did. These findings suggest that unattended pictures may continue to be processed more readily than words even under conditions in which attention is not directed toward them.

Kindergarteners and adults learn fraction-rules in a categorization task

Tasha Posid
The Ohio State University

Vladimir Sloutsky
The Ohio State University

Abstract: Both children and adults can learn new categories when presented with a rule about a perceptual feature. Like categorization, numerical abstraction requires the ability to ignore irrelevant (non-numeric) perceptual features when making decisions about relevant (numeric) features. The present study fuses these two lines of research by training 5-7 year olds and adults in a categorization task, in which they must form a rule about a fraction-based category. Can children form this mathematical category readily? Will they be able to do so without any formal instruction? How does this ability develop or change across the lifespan? We find that young children and adults readily form fraction-based categories, indicating that children can think about proportional information prior to formal schooling. Additionally, an ability to map between visual and symbolic representations aided both children and adults in this numeric categorization task, with children showing additional gains in traditional fraction knowledge.

Promoting prosocial behavior and emotional awareness in preschoolers

Qiuwen Dou
Department of Psychological Studies, The Hong Kong Institute of Education

Sarah-Jane Vick
Department of Psychology, University of Stirling

Michael C. W. Yip
Department of Psychological Studies, The Hong Kong Institute of Education

Abstract: Human-animal interaction (HAI) supports a theory that animals may enhance social support both directly, as a source of comfort, and indirectly, as a facilitator of human interaction. Existing research found that dogs are suitable in helping children develop healthy self-esteem and empathy for others and the classroom interaction between assistance dogs and children is also effective in developing children's sense of responsibility and empathy. The current study implemented dog-assisted interaction (DAI) and video interaction (VI) as two different experimental conditions to test their efficacies in promoting preschoolers' prosocial behaviors and emotional awareness. A total of 146 Chinese children (aged from 3 to 5.5 years old) participated in the study. The results indicated that both DAI and VI are more effective in promoting collaborative behaviors than regular teaching method on their collaborative ability and emotional awareness while only DAI significantly enhanced children's emotional awareness.

A Cognitive Approach to Modeling Sentence Level Prominence Based on Stimulus Unpredictability

Sofoklis Kakouros
Aalto University

Okko Räsänen
Aalto University

Abstract: The human sensory system is capable to rapidly respond to novel input, allowing for quick allocation of attentional resources to the stimulus. In a similar manner, prominent words in speech seem to attract the listeners' attention and facilitate or alter interpretation. Sentence prominence has been typically studied across languages by examining configurations of acoustic prosodic features during prominent words. Recent studies have provided evidence that, in addition to the predictability of the lexical units in speech, manipulating the predictability of the acoustic prosodic features can also signal prominence. In this work, we provide a high-level description of a cognitive framework that attempts to characterize sentence prominence as a phenomenon that is connected with the unpredictability of suprasegmental acoustic features, thereby capturing the attention of the listener and causing differential processing of prominent speech.

Simulating the cost of cooperation: A recipe for collaborative problem solving

Alessandro Lazzeri
University of Pisa

Andrea Guazzini
University of Florence

Daniele Vilone
CNR - Italy

Giorgio Gronchi
University of Florence

Abstract: Crowdsourcing consists in engaging a community of people in solving complex problems. Collaborative problem solving is affected by many variables (e.g., group size, difficulty of the task, tendency to cooperate) in a complex way. In this study, we extend the results of Guazzini et al. (2015) by means of a numerical simulation exploring the impact of the cost of cooperation in collaborative problem solving. We observed that the cooperation costs have damaging effect with smaller groups that face hard problems. When groups fail to solve the problem there is a long-term reduction in fitness (since the group is not able to learn) as well as a short-term loss of a payoff. So, when facing small group and hard task in concrete application, it is better to control the cooperation costs with ad hoc interventions.

Choice magnitude and decision time: investigating magnitude sensitivity in value-based decision making

Angelo Pirrone
University of Sheffield, UK

Philip Pärnamets
Lund University, Sweden

James Marshall
University of Sheffield, UK

Tom Stafford
University of Sheffield, UK

Abstract: From an evolutionary perspective, it has been proposed that decision making should be sensitive to the overall magnitude of the alternatives under consideration in order to resolve costly deadlocks and thus improve long-term reward intake. We provide initial evidence that the overall magnitude of the alternatives affects decision making, by speeding up decision time in order to maximise a speed-value trade off. Implications for current computational models of decision making, in particular for the Drift Diffusion Model, are discussed.

Multimodal Object Recognition and Categorisation by Interactive Behaviours

Haojun Guan
University of Hamburg, Hamburg, Hamburg, Germany

Jianwei Zhang
University of Hamburg, Hamburg, Germany

Abstract: Human beings have an excellent ability which can form and recognise object categories. In this paper, a novel system of multimodal object recognition and categorisation by perform- ing interactive behaviours is introduced. Video clips are filmed as the raw input of the system. A dataset of 100 objects with 18 categories and 5 different interactions is used to evaluated the performance. Convolutional neural network is used to train the classifier and learn the categories. The result shows the highest, lowest and average recognition accuracies of every specific object in every category and the receiver operating characteristic for every category. The connection between the presented system and human cognitive system is discussed in the conclusion and future works.

Preschoolers evaluate risk and reward in exploration-exploitation tasks

Elizabeth Lapidow
Rutgers University - Newark

Elizabeth Baraff Bonawitz
Rutgers University - Newark

Abstract: Children are drivers of their own discovery. To develop a complete characterization of the factors that drive exploration in early childhood, we must first understand how competing factors influence children's decision making. We investigated preschool-aged children's decision-making on explore-exploit tasks where the available information about the distribution of rewards was controlled. When probability information is unknown, children preferred to exploit known rewards over exploring unknown ones. However, performance in Experiment 2 shows that children can use probabilistic information to form accurate expectations about possible outcomes to effectively choose between exploiting and exploring. The degree to which individual children are "exploratory" is also shown to be consistent over weeks, suggesting that individual children have "trait-like" exploratory drives. On aggregate, children incorporate these individual tendencies towards exploration or exploitation with probability information; thus children readily form estimations of expected reward and use this information to guide efficient exploratory behavior.

Transfer at the Level of Human-Computer System: Problem Solving using Procedure-Automation Software

Dorrit Billman
San Jose State University

Debra Schereckenghost
TRACLabs

Zachary Caddick
San Jose State University

Abstract: Issues of transfer and of human-computer systems are central, but largely separate questions in cognitive science. We take the human-computer system as the unit of analysis and explore how well a human-computer system transfers to tasks outside the scope of the humans' training and the software's design. In two experiments, participants used the procedure automation software (PRIDE) to control simulations of International Space Station habitat systems. Both the software design and the user training addressed routine procedure execution. In the transfer problems the conditions assumed for routine procedure execution were not met, requiring novel problem solving. We report on our methods for complex behavior analysis and our results showing high though imperfect transfer, noteworthy given the widespread difficulty of transfer. Further investigation of transfer at the level of human-computer system is important for understanding what combinations of technology design and user experience enable effectively dealing with the unexpected.

What Makes Campaign Messages Popular On Twitter ?

M. Afzal Upal
Defence R & D Canada

Pritham Marupaka
Defence R & D Canada

Abstract: The rapid adoption of social media by billions of people from all over the world has unleashed unprecedented opportunities for marketers and cognitive scientists to better understand why some message become popular while other die quickly. We designed a novel technique for automatically learning to differentiate popular tweets from unpopular ones and to predict how popular a given tweet will become in a given target audience. To demonstrate the effectiveness of our approach, we applied it to real world data collected from six social media messaging campaigns run by a variety of marketing as well as non-profit organizations including Proctor and Gamble's Always Campaign. The studies showed that our approach can be highly effective (achieving accuracy scores from 92% to 99%) for automatically learning what makes a message popular in any given group as well as for automatically predicting how popular a message will be in a given target audience.

Investigating the Explore/Exploit Trade-off in Adult Causal Inferences

Erik Herbst
University of Toronto, Toronto, ON, Canada

Christopher Lucas
University of Edinburgh

Daphna Buchsbaum
University of Toronto, Toronto, ON

Abstract: We explore how adults learn counterintuitive causal relationships, and whether they interpret evidence and discover hypotheses by incrementally revising beliefs. We examined how adults learned a novel, unusual causal rule when given data that initially appeared to follow a simpler, more salient rule. Adults watched a video of blocks placed sequentially on a detector that activated when a block was a "blicket", then were asked to determine the underlying causal structure. We contrasted two causal learning problems. In both cases, one rule could be used to determine which objects were blickets; in the first problem this rule was complex, but could be found by making incremental improvements to a simple and salient initial hypothesis. The second problem's rule was simpler, but to adopt it, participants had to ignore initial beliefs. Our results provide some of the first evidence for an inference trade-off analogous to the "explore-exploit" trade-off in active learning.

Author Index

Abbasi, Yasaman 1391
Abbaspour, Sufi 2942
Abbott, Joshua 2363
Abel, Hillary 2998
Abney, Drew 1439, 2936
Aboody, Rosie 770
Abozeed, Magda 2973
Abraham, Suneera 3047

Abramyan, Meline 241
Adams, Deanne 2957
Ahissar, Merav 2866
Ahmed, Annisa 2973
Akhavan, Niloofar 716
Akin, Heather 1619
Akl, Mahmoud 3057
Aldosari, Bushra 3024
Aldrich, Frances 3042
Alexopoulou, Theodora 1050
Alishahi, Afra 486
Allen, Kelsey 2495
Allen, Laura 764, 2483, 2681
Allred, Sarah R. 2273
Almeda, Ma. Victoria 812, 2869
Alsius, Agnès 3043
Altiero, Evan 2868
Amon, Mary Jean 1553
Amorati, Alexandra 2957
Anderson, Andrew 2249
Anderson, John 2887
Anderson, Nathaniel 51
Andonova, Elena 3049
Andrews, Jan 2933
Andrews, Jessica J. 3039
Andrikopoulou, Angeliki 87
Anggoro, Florencia 2968
Anthony, Susan 468, 2153
Apfelbaum, Keith 2195
Arnhold, Anja 562
Arnoldi, Jakob 3036
Arnon, Inbal 1469, 1643
Arroyo, Jeffrey 3067
Arunachalam, Sudha 2893
Asaba, Mika 135, 1529, 3001
Asano, Yuki 1775
Asselin, Jennifer 1979
Atmanspacher, Harald 568
Aubin, Sean 2021
Austerweil, Joseph .. 313, 1158, 1679, 1907, 2847, 2921
Azab, Habiba 2831
Azar, Zeynep 1295
Azari, Banafsheh 462
Baayen, Harald 550, 2441
Bach, Dominik 1140
Backus, Ad 1295
Badger, Julia 199
Baer, Carolyn 2923
Bailey, Richard 3025
Baker, Ryan 812
Balasubramaniam, Ramesh 2926, 2945
Baldwin, Dare 1337, 2854, 2894, 2985
Baltaretu, Adriana Alexandra 978
Bambach, Sven 1631
Baratgin, Jean 1104
Barbieri, Christina 592
Barbu, Roxana-Maria 1961
Barkoczi, Daniel 918
Barner, David 259, 1595, 2381
Barnes, Kevin 2935
Barnes, Tiffany 2801
Barreiro, Shawntel 2952
Barsuola, Giulia 2863
Bartels, Daniel 1727, 2303
Bartley, Jessica 3022
Bartolotti, James 2735
Barton, Alexis 1044
Bauer, Jessie-Raye 2994
Beck, Sara 1817
Beekhuizen, Barend 2255
Begg, Steve 544
Begolli, Kreshnik 2015

Author Index

Behme, Christina 205
Belardinelli, Anna 2965
Bell, Kimberly 3046
Beller, Sieghard 43
Bello, Paul 1517, 2573
Ben-Asher, Noam 1391
Bender, Andrea 43
Bennett, Erin 2669
Benton, Deon 830
Berends, Floris 193
Bergelson, Elika 21
Bergen, Benjamin 2417
Bergmann, Christina 33, 1331
Berman, Marc 1721
Bernatskiy, Anton 2868
Berretta, Matias 1535
Bertel, Sven 462
Bertenthal, Bennett 936, 1044
Besner, Derek 2962
Besserman, Ana 877
Bhatia, Sudeep 1937, 1943
Bhowmick, Sankha 2135
Bigelow, Eric 2597
Billman, Dorrit 3078
Binz, Marcel 3062
Bixter, Michael T. 2999
Blackburn, Patrick 2579, 3034
Blake, Peter 2857
Blanchard, Thomas 2663
Blascovich, Jim 2369
Blass, Joseph 556
Blattner, Raymond 2917
Bloomfield, Amber 990
Blouw, Peter 1583
Boghrati, Reihane 1463
Bolognese, Martina 2863
Bonawitz, Elizabeth Baraff 3077
Bonawitz, Elizabeth 13, 1086, 2858
Bongard, Joshua 2868
Booth, Amy 2994
Booth, Julie 592
Borghi, Anna M. 3032
Boroditsky, Lera 1164, 2417
Borovsky, Arielle 1421
Borst, Jelmer 2189
Börütecene, Ahmet 1871
Bostan, İdil 1871
Boucher, Juliana 2905
Boyer, Ty 936
Bradley, Petra 990
Bradshaw, Gary 2935
Brady, Michael 2765
Braginsky, Mika 1691
Braithwaite, David 2881
Bramley, Neil 2567
Brand, James 2991
Brandone, Amanda 2951
Branigan, Holly 1715
Brannon, Elizabeth 1841
Braüner, Torben 2579, 3034
Braus, Niels 105
Brawner, Stephen 1158
Brazil, Michael 3047
Brennan, Nicole 420
Brewe, Eric 3022
Brezack, Natalie 2971
Bridewell, Will 1517, 2573
Bridgers, Sophie 710, 1985
Brochhagen, Thomas 2081
Brodsky, Jacob 2948
Broniatowski, David 1667
Brook, Andrew 2399
Brothers, Brock 1205
Brouwer, Harm 2555
Brueggemann, Stephan 1032
Bryson, Joanna 1170
Buchler, Norbou 2405
Buchsbaum, Daphna 3080
Bunce, John 1439
Burigo, Michele 580

Author Index

Burnetta, Danielle 2973
Busemeyer, Jerome 25
Bütepage, Judith 1014
Butz, Martin V. 2965
Butz, Martin 1787, 3062
Buwalda, Trudy 2189
Byers, Patrick 1128
Byrne, Kaileigh 2357
Cacchione, Trix 2942
Caddick, Zachary 3078
Caglar, Leyla Roksan 3050
Caldwell, Christine 43
Calvillo, Jesus 2555
Campbell, Susan 990
Campos, Joana 1403
Candan, Ayse 2872
Cao, Rui 438
Capelier-Mourguy, Arthur 336
Carbajal, M. Julia 889
Carella, Ernesto 3025
Carey, Susan 2990
Carlson, Laura 2867, 2957
Carr, Jon W. 2759
Carstensen, Alexandra 2231
Carvalho, Paulo 2327, 2856, 2895
Casasanto, Daniel 1056, 2717, 2753
Cassani, Giovanni 1152
Castellano, Katherine 2907
Castellanos, Kristina 2936
Castillo, Ramon D. 1080
Catrambone, Richard .. 47, 271, 1925, 2009
Caudek, Corrado 171
Čevora, Jiří 2903
Chadwick, Krysten 2201
Chan, Antoni B. 182, 854, 1032
Chan, Cynthia Y. H. 854
Chandran, David 3047
Chang, Ya-Ning 123
Chater, Nick 2864
Chen, Chi-hsin 2856, 2943
Chen, Chin-Chih 2924
Chen, Dar-Wei 1925
Chen, Dawn 574
Chen, Ge 3065
Chen, Hsien-Ying 2982
Chen, Jenn-Yeu 2955, 2982
Chen, Jidong 2267
Chen, Junqing 2878
Chen, Stephanie 1727
Chen, Yingjun 2878
Chen, Yiwei 2848
Chen, Zhixian 3061
Cheng, Patricia 307, 480
Cheng, Peter 3042
Cherng, Rong-Ju 2982
Chernyak, Nadia 1607, 2159, 2857
Chesney, Dana 1559
Cheung, Pierina 2711, 2813
Cheyette, Samuel 614
Chi, Min 2801
Chien, Hsin-Yin 2924
Child, Leah 2894
Chin, Jessie 538
Cho, Chiu-Yu 2982
Cho, Pyeong Whan 1487
Choi, Dongkyu 620
Choi, Yun-Kyeung 2850
Christiansen, Morten 43, 81, 420, 901
Chu, Junyi 49, 1595
Chuk, Tim 182
Chung, Susana T. L. 1499
Chwilla, Dorothee 111
Cialdella, Vincent 1913
Cibelli, Emily 2847
Ciraolo, Margeaux 848, 3070
Coderre, Emily 37
Coenen, Anna 2819
Cohen Priva, Uriel 1895
Cohen, Alexander 644
Cohen, Alexandra 2927

Author Index

Cohen, Jonathan D. 1547
Cohen, Jonathan 1751, 1793
Cohn, Neil 37, 2057
Colflesh, Gregory 289
Collins, Michael 75
Collister, Kevin 2315
Colombo, Matteo 432
Conaway, Nolan 2375, 2693, 2890
Congdon, Eliza 2141
Connor Desai, Saoirse 626, 2993
Conway, Christopher ... 51, 396, 740, 1074, 1481, 2129
Cooper, Rick 53
Cooperrider, Kensy 692
Corbett, Natalie 592
Cordes, Sara 586, 1607, 1733
Corter, James 3069
Cosgrove, Laura 1176
Costello, Fintan 408, 2453
Cottrell, Garrison 217, 1409, 2519
Cottrell, Gary 3019
Couper, Lisa 372
Cowan, William 1649
Cowell, Rosemary A. 2087
Cox, Gregory 1589
Cox, Sarah 2902
Crandall, David 1631
Cranford, Edward 2935
Crawford, L. Elizabeth 3011
Craycraft, Nicole 2783
Creel, Sarah 2949
Criss, Amy 1589
Cristia, Alejandrina 33, 1331
Crocker, Matthew 2555
Croijmans, Ilja 141
Croker, Steve 1913, 2954
Crone, Philip 842
Crossley, Scott 764
Crump, Matthew 1325
Cruz, Nicole 1104
Cummins, Chris 1241
Cushman, Fiery 3007
D'mello, Sidney 2914
Daamen, Jeroen 2147
Dablander, Fabian 2669
Daelemans, Walter 1152
Dahlbäck, Nils 734
Dale, Meghan 2711
Dale, Rick 901, 1433, 2941, 3015
Danek, Amory H. 2929
Danileiko, Irina 2825
Daoust, Louise 2930
Darby, Kevin 3004
Dascalu, Mihai 764
Davenport, Jodi 372
Davies, Jim 1415
Davis, Zachary 1002
Daw, Nathaniel 1751, 1793
Dayer, Alex 1913
De Deyne, Simon 2972
de Jong, Chiara 1116, 1901
de Kleijn, Roy 193, 1601
de Melo, Celso 2369
de Vries, Erica 450
Degen, Judith 2261, 2669
Dehghani, Morteza 638, 1463
Dell, Gary 51
Demberg, Vera 366
Deng, Sophia 2974
Deng, Yu 3054
Deng, Zhen 3061
Denison, Stephanie 2117
Deocampo, Joanne 1074, 2129
Derringer, Cory 758
Deslauriers, Andrew 2882
Destefano, Marc 668
Dewald, Andrew 848, 3070
Dey, Biswadip 1547
Dickey, Michael Walsh 3026
Dickey, Michael 2946
Dils, Alexia Toskos 3067

Author Index

Dimov, Cvetomir 492
Dingemanse, Mark 1823
Djalal, Farah M. 1068
Dmello, Sidney 2981
Do, Kyung Soo 2855, 2919
Doesburg, Inge 650
Don, Hilary 2897
Dou, Qiuwen 3072
Doumas, Leonidas 2279
Doyle, Gabriel 2093
Dresang, Haley C. 3026
Drummey, Anna 2998
Duffy, Sean 3011
Dunham, Yarrow 2857
Dunn, Timothy 2911
Dupoux, Emmanuel 662, 889, 1331
Duran, Nicholas 2953
Durgin, Frank 1769
Durriseau, Jaymes 2935
Dweck, Carol 2741
Dye, Melody 1763
Easvaradoss, Veena 3047
Edmunds, Charlotte 69, 2243
Eghbalzad, Leyla 2129
Ehrenfeld, Stephan 3062
Einig, Jonas 3062
Eisenberg, Ian W. 2039
Eisenberg, Michelle 53
El Banani, Mohamed 235
El-Shroa, Omar 2723
Eliasmith, Chris 1122, 1583, 2021, 2183
Elsner, Claudai 3020
Endo, Katsutochi 2898
Engelbrecht, Klaus-Peter 2549
Engelhardt, Paul 526
Engelthaler, Tomas 2879
Enns, Robyn 2848
Erdogan, Goker 2543
Ettinger, Allyson 1445
Evans, Natalie 2968
F. Malle, Bertram 2976
Faber, Myrthe 2914
Fairchild, Sarah 93
Farkas, Haley 2970
Favela, Luis H. 1673
Favela, Luis 948
Fazio, Lisa K. 265
Feher, Olga 41
Feldman, Naomi 1367, 1445, 1919
Fér, Radek 889
Ferrara, Katrina 2927
Ferretti, Todd 2880, 2937
Finch, Dustin 2327
Fischer, Eugen 526
Fisher, Anna 188, 199, 812, 2966
Flecken, Monique 2633
Fleming, Morgan Magnus 2877
Flombaum, Jonathan I. 2273
Flores, Shaney 53
Flusberg, Stephen ... 866, 1571, 2315, 3067
Fong, Nathan 728
Forbus, Ken 836, 1277, 3055
Forbus, Kenneth 556
Forestier, Sébastien 1859
Forsyth, Carolyn 824
Foster-Hanson, Emily 1541
Foster, Katlyn 1685
Fotiadis, Fotis A. 277
Fournier, Lisa 1361
Fourtassi, Abdellah 662
Foushee, Ruthe 1697, 2177
Fox, Amy 450
Franchak, John 2884
Franconeri, Steve 2141
Frank, Michael C. 1385, 2771
Frank, Michael C 354
Frank, Michael 2093, 2423, 2609
Frank, Mike . 33, 319, 842, 1691, 2459, 2789
Frank, Stefan L. 2207
Frank, Stefan 2105

Author Index

Franke, Michael 2081, 2639, 2669
Freeman, Cassiopeia 2869
Freeman, William T. 1781
French, Bob 1349
Freudenthal, Daniel 602, 674

Freund, Michael 1625
Frissen, Eefje 99
Frost, Rebecca 81
Frye, Douglas 2940
Fuellenbach, Kim Kristina 3048
Fum, Danilo 1134
Furutani, Ryohei 972
Fusaroli, Riccardo 901, 3006, 3036
Fuyama, Miho 283
Galantucci, Bruno 1211
Galeucia, Christopher 2957
Gao, Dingguo 1901
Gao, Tao 17
Garcia Garcia, Grecia 3042
Garten, Justin 1463

Gatt, Albert 159
Gauer, Gustavo 3068
Geana, Andra 1751, 1793
Gelman, Susan 1307
Genovesi, Chris 3044
Gentner, Dedre 692, 1038, 2507
George, Nathan 111
Geraci, Carlo 55
Gershman, Samuel 1493
Gerstenberg, Tobias 1745, 2567, 2777
Gervits, Felix 1146
Gerwien, Johannes 2633
Gewaltig, Marc-Oliver 3018
Ghafurian, Moojan 1223
Gibbert, Michael 895
Gijssels, Tom 2717
Gillis, Steven 1152
Gilman, Anne 1062
Gimbel, Sarah 638
Ginsburg, Herbert 2869
Glady, Yannick 1349
Gleason, Emily 1895
Glebkin, Vladimir 3045
Gluck, Kevin 75, 2997
Gobet, Fernand 602, 674, 1355, 1475, 1973, 2863, 2899, 2939
Godwin, Karrie 812, 2966
Goedert, Kelly 2917
Gokcay, Didem 2963
Goksun, Tilbe 716, 1871
Goldin-Meadow, Susan 692, 2141
Goldstein, Ariel 2967
Goldstone, Robert 39, 41, 2327, 2895
Goldwater, Micah 39, 2897
Golinkoff, Roberta 2971
Gomez, Rebecca 51
Gonzalez, Cleotilde 614, 1391
Gonzalez, Richard 3068
Goodkind, Rachel 2905
Goodman, Noah D. 2771
Goodman, Noah ... 35, 41, 360, 1110, 1613, 1655, 1889, 2261, 2669
Gopnik, Alison 770
Gordon, Barry 1625
Gordon, Chelsea 1439, 2945
Gorman, Chris 342
Gosmann, Jan 2183
Grabowski, Joachim 2916
Graesser, Art 253
Graf, Caroline 2261
Grant, Erin 2309
Gratch, Jonathan 2369
Gray, Wayne 45, 668
Greenwald, Amy 1158
Grey, Sarah 3046
Griffin, Thomas 2931
Griffiths, Thomas L. 27, 1505
Griffiths, Thomas 2363, 2847
Griffiths, Tom .. 41, 1008, 1697, 2063, 2075
Grigoroglou, Myrto 782
Grimm, Lisa 2888
Grimm, Robert 1152

Author Index

Grinberg, Maurice 704, 1511
Grodner, Daniel 3059
Gronchi, Giorgio 171, 331, 3074
Gros, Hippolyte 1271
Gu, Jin Ju 2855
Gu, Yan 2603
Gualtieri, Samantha 2117
Guan, Haojun 3076
Guazzini, Andrea 3074
Gumbsch, Christian 1787
Gunzelmann, Glenn 75
Gureckis, Todd M. 2675
Gureckis, Todd 2051, 2477, 2819, 2992
Gweon, Hyowon .. 135, 710, 924, 966, 1529, 1985, 3001
H. Arai, Noriko 2297
Haarmann, Hank 1283
Hafri, Alon 2537
Hagoort, Peter 1823
Halbrügge, Marc 2549
Hall, Deanna 2880, 2937
Hall, Lars 301
Halsey, Andrew 2997
Hamamouche, Karina 1733
Hamana, Mai 3003
Hamrick, Jessica B. 27
Hansen, Steven S. 19, 2045
Hanson, Stephen Jose 3050
Haque, Moinuddin M. 486
Harbison, J. Isaiah 990, 2435
Hardenbergh, Dylan 2995
Harding, Samuel 936, 1044
Harman, Jason 614
Harmon, Zara 520
Harris, Adam 390, 2993
Harris, Jack 2997
Harris, Paul 1607
Harrison, Anthony 2861, 2964
Hart, John 2953
Hartshorne, Joshua 15, 1745, 2975
Hass, Richard 996
Hassin, Ran 2967
Hawkins, Robert X.d. 1889
Hawkins, Robert 41, 2261
Hawthorne, Kara 562
Hayashi, Yugo 247, 818
He, Angela Xiaoxue 2893
Hegarty, Mary 47
Heller, Daphna 2123
Hembacher, Emily 354
Hemmer, Pernille 3005, 3017
Hendricks, Rose 1164, 2417
Hendrickson, Andrew 2471
Hendrix, Peter 2441
Henson, Richard 2903
Herbst, Erik 3080
Herbstritt, Michele 2639
Herd, Brooke 2015
Hertwig, Ralph 2986
Hespos, Susan 55
Heverly-Fitt, Sara 2871
Heyman, Gail 259
Heyman, Tom 1068
Hidaka, Shohei 283
Hilgard, Joseph 1619
Hilliard, Elizabeth 1158
Hills, Thomas ... 895, 907, 1565, 2879, 2986
Hirsh-Pasek, Kathryn 2971
Hirsh-Pasek, Kathy 2873
Hitczenko, Kasia 1367
Ho, Mark 1158, 1679
Hochmann, Jean-Remy 2990
Hoffmann, Ana Maria 2886
Holcomb, Michelle B. 3026
Holcomb, Phillip 2765
Holden, John G. 1553
Holmes, Kevin 2873, 2882, 2883, 2948
Holodny, Andrei 420
Hommel, Bernhard 193, 1601
Honda, Hidehito 223
Hong, Jeffrey 2880, 2937

Author Index

Hong, Min Kyung 265
Honke, Garrett 2693
Hoover, Joe 1463
Hopkins, Emily 1829
Hopper, William J. 1235
Horne, Zachary 39
Hostetter, Autumn 1685
Howard, Lauren 53
Hoyos, Christian 2507
Hristova, Evgeniya 704, 1511
Hsiao, Janet H. 182, 854, 1499
Hsiao, Janet 1032
Hu, Helen 1062
Hu, Ying 3061
Huang, Jingyuan 2950
Huang, Ren 1901
Huang, Yujing 3052
Hubbard, Edward 2651
Huber, David E. 1235
Huette, Stephanie 253, 3038, 3053
Huettig, Falk 2267
Huppert, Theodore 199
Hurst, Michelle 586, 1733
Husband, E. Matthew 55, 3048
Huys, Quentin 1140
Iain, Oliver 1355
Icard, Thomas 800
Igami, Tsuyoshi 414
Imai, Mutsumi 49
Inohara, Keisuke 3035
Iotzov, Ivan 2908
Irvin, Jeremy 2645
Isbilen, Erin 2978

Ishiguchi, Akira 1523
Iwai, Ritsuko 3066, 3066
Iyiewuare, Peace 1535
Jabotinsky, Hadar 2906
Jackson, G. Tanner 2907, 3002
Jackson, Victoria 2968
Jacobs, Robert A. 2543
Jacobs, Robert 1259
Jacovina, Matthew 2681
Jaeger, Allison 2931
Jaeger, T. Florian 2351
Jaeggi, Susanne 2015
Jagannatha Rao, Sanjeev 217
James, Kofi 3039
Jamieson, Randall 1325
Jansen, Aren 1919
Jara-Ettinger, Julian 710, 2033
Järvikivi, Juhani 550, 562
Jee, Benjamin 2968
Jenkins, Gavin 508
Jenkins, Sarah 390
Jensen, Clint 2904
Jeong, Jeein 2940
Jiang, Chenfanfu 17, 1805
Jimenez, Sofia 913
Jin, Tiantian 3008
Jo, Hwiyeol 2920, 3051
Johannes, Kristen 372, 2429
Johanson, Megan 1146
Johansson, Petter 301
Johns, Brendan 1325, 1763, 2291
Johnson-Laird, P.n. 2165
Johnson, Chelsea 2953
Johnson, Kate M. 1463
Johnson, Kelli L. 2999
Johnson, Samuel 63, 776, 1577, 1967
Johnston, Benjamin 514
Jones, Gary 602, 674
Jones, Michael N. 1397, 2989
Jones, Michael 1092, 1325, 1433, 1763, 1799, 2291
Jones, Winston 2935
Jordan, J. Scott 1913
Joseph, Ebenezer 3047
Jost, Ethan 420
Jun, Kwang-Sung 1907
Kachergis, George ... 193, 1601, 2477, 2992
Kacinik, Natalie 2878

Author Index

Kadreva, Veselina 704
Kaesler, Matthew 502
Kahan, Dan 1619
Kai, Shimin 812
Kaiser, Elsi 877, 2513
Kajic, Ivana 2183
Kakouros, Sofoklis 2489, 3073
Kan, Irene 2998
Kanero, Junko 111, 2873
Kang, Carissa 3033
Kao, Justine 35, 2669
Kapatsinski, Vsevolod 520
Kaplan, Jonas 638
kar, Debarun 1391
Kardan, Omid 1721
Karnick, Harish 596
Karuzis, Valerie 289
Kaschak, Michael 1421
Katz, Irvin 2851
Kawahara, Daisuke 3066
Keehner, Madeleine 824
Kehler, Andrew 2393
Keil, Frank 776, 1577, 1967, 3020
Kello, Chris 1865, 2926
Kelly, Deirdre 1415
Kemp, Jovan 2872
Kendall, L.n. 37
Kenett, Yoed 313, 1907
Kennedy, William G. 2960
Kershaw, Trina 2135
Kersten, Luke 2399
Kerster, Bryan 1865
Khemlani, Sangeet 23, 2165, 2861, 2964
Khooshabeh, Peter 2369
Khosroshahi, Ehsan 2969
Kidd, Celeste 698, 2831
Kiefer, Reina 2905
Kikuchi, Yusuke 794
Kim, Boyoung 2976
Kim, Dan 752
Kim, Eunkyung 638
Kim, Haylie 776
Kim, Jaeah 199
Kim, Jongin 2920, 3051
Kim, Jung Nyo 3000
Kim, Kristen 1577
Kimel, Eva 2866
Kirby, Simon 1241, 2759
Kirsh, David 1253
Kitrick, Eileen 2873, 2883
Kiyokawa, Sachiko 3041
Kjellström, Hedvig 1014
Kleiman-Weiner, Max 1158, 1679
Kleinschmidt, Dave 2351
Klichowicz, Anja 129
Klimek, Brittany 2951
Kline, Melissa 15
Klinker, Gudrun 3018
Kloos, Heidi 1080
Kneissler, Jan 1787
Knobe, Joshua 800
Knoblich, Guenther 2870
Knoblich, Günther 2219, 2747
Knoeferle, Pia 580
Knoll, Alois 3018
Knott, Alistair 342, 532, 2345
Kobayashi, Munehiro 2297
Kocab, Annemarie 1343
Kodama, Kentaro 794
Koedinger, Kenneth 954, 1229, 2069
Kolesari, Jennifer 2867
Komatsu, Takanori 930
Komer, Brent 1122
Kominsky, Jonathan 3020
König, Peter 2747, 2909
Konishi, Haruka 2971
Konovalova, Elizaveta 722
Konstantinidis, Emmanouil 614
Koolen, Ruud 159
Kopp, Kris 2981

Author Index

Kopp, Kristopher 2914
Koralus, Philipp 2852, 3025
Korman, Joanna 2976
Kornmeier, Jürgen 568
Kosie, Jessica 1337, 2854
Koster van Groos, Janet 2901, 2950
Kotabe, Hiroki 1721
Kotchoubey, Boris 2839, 2875
Kragic, Danica 1014
Krahmer, Emiel 99, 159, 486, 978, 1901
Kralik, Jerald 2723
Krämer, Nicole 2925
Kranjec, Alexander 2983
Krause, Andreas 1140
Kregor, Jennifer 1176
Krems, Josef F. 129
Kriegel, Zoe 2783
Krishnamurthy, Umesh 2980
Krishnaswamy, Nikhil 2841
Kristiansson, Mattias 734
Krusmark, Michael 75
Kryven, Marta 1649
Kubricht, James 1805
Kuhn, Jeremy 55
Kujala, Tuomo 2465
Kumada, Takatsune 3066, 3066
Kumcu, Alper 2387
Kunda, Maithilee 235
Kurohashi, Sadao 3066
Kurtz, Ken 2933
Kurtz, Kenneth J. 2938
Kurtz, Kenneth 2333, 2375, 2693, 2860, 2890
Kushnir, Tamar 770, 1307, 1877
Kvam, Peter 25, 1098
Lachman, Margie 907
Lagnado, David 626, 2993
Lai, Jun 1901
Laird, Angela 3022
Lake, Brenden 644, 2051
Lamar, Michelle 824, 2901, 2950
Lamnina, Marianna 3008
Landau, Barbara 2429, 2927
Landrum, Asheley 1619
Landy, David 1301, 1799
Landy, Justin 2303
Lane, Peter 1475
Langley, Pat 2843
Langstein, Benjamin 1211
Lannig, Guilherme 3068
Lapidow, Elizabeth 2858, 3077
Large, Edward 3000
Lark, Murray 390
Larson, Calais 2933
Laschi, Cecilia 3018
Lassiter, Daniel 1110
Latif, Nida 3043
Lau Petersen, Malte 3006
Lazzeri, Alessandro 3074
Le Corre, Mathieu 2711
Le Mens, Gael 722
Lee, Daye 1571
Lee, Himchan 1176
Lee, Jae-Ho 2850
Lee, Michael 211, 2825
Lee, Tatia M. C. 854
Leech, Kathryn 2159
Leeds, Daniel 2908
Lefevre, Jo-Anne 656, 3021
Legare, Cristine 2994
Lehman, Blair 3039
Lei, Chia-Ming 3026
Lengyel, Máté 2903
Leong, Yuan Chang 966
Lerch, Rachel A. 2273
Leshinskaya, Anna 2853
Leslie, Sarah-Jane 1541
Lester, Nicholas 1847, 2585
Levari, Tanya 2995
Levering, Kimery 2890
Levi, Paul 3018

Author Index

Levy, Roger P. 2393, 2393
Levy, Roger 378
Lew, Timothy 746
Lewis, Molly 33, 1385, 2609
Lhommet, Margot 788
Li, Adam 2955
Li, Fuyin 3054, 3064
Li, Linjie 1427, 3019
Li, Nan 1619
Li, Sheng 117
Li, Tze Kwan 1499
Li, Ying 2986
Lieder, Falk 2075
Lien, Chi-Shun 2924
Lin, Rebecca 2369
Link, Daniela 1883
Lins, Jonas 1931
Lippa, Katherine 496
Littman, Michael 1158, 1679
Litvinova, Aleksandra 638
Liu, Ran 2069
Liu, Xun 1457
Liu, Yuzhong 3065
Livesey, Evan 2897
Livingston, Ken 2933
Livingston, Kenneth 2868
Livingston, Nicholas 2868
Lloyd-Kelly, Martyn 1475
Lobato, Emilio 2954
Lockwood, Gwilym 1823
Loehr, Abbey 1739
Loewenstein, Jeffrey 39
Lohmann, Johannes 2965
Lombrozo, Tania 1979, 2663
Long, John 2868
Lõo, Kaidi 550
Loomes, Graham 1943
Lorenzino, Martina 171
Lou-Magnuson, Matthew 2105
Loucks, Jeff 53
Loui, Sofia 2889
Lu, Hongjing 229, 574, 1805
Lu, Qihong 2862
Lucas, Christopher 2921, 3080
Lugli, Luisa 3032
Luhmann, Christian C. 2999
Lukowski, Angela 241
Luo, Yu 984
Lynch, Collin 2801
Ma, Ning 2705
Ma, Wei Ji 2918
Ma, Zheng 2273
MacDonald, Kyle 2459, 2609
MacDonald, Maryellen 1026
MacGlashan, James 1158
Madsen, Jens Koed 165, 2993
Madsen, Jens 3025
Maehigashi, Akihiro 414
Maes, Alfons 978
Magliano, Joe 37
Maglio, Paul P. 2877
Maglio, Paul 2980
Mahdeyan, Ahmad 3063
Majid, Asifa 141, 1451, 2267
Malave, Vicente 1427, 3019
Malle, Bertram F. 924
Malt, Barbara C. 1955
Malt, Barbara 49, 2213
Mandalaywala, Tara 1319
Marchman, Virginia 1691
Marewski, Julian N. 1883
Marewski, Julian 2977
Marghetis, Tyler 2381, 2615
Margulieux, Lauren 271, 2009
Marian, Viorica 348, 2735
Markant, Doug 444
Markant, Douglas B. 2675
Markman, Ellen 2741
Marquis, Alexandra 2958
Marsella, Stacy 788

Author Index

Marsh, Jessecae 1020
Marshall, James 3075
Marshall, Peter 2913
Marti, Louis 698
Martin, Andrea 2279
Marupaka, Pritham 3079
Mascarenhas, Salvador 2852
Masnick, Amy 3024
Massey, Christine 1176
Mata, Rui 907

Mathis, Ariel 253, 3038
Matsuka, Toshihiko 223
Matsumuro, Miki 176
Matsuzaki, Takuya 2297
Matthews, Percival 2651
Mayrhofer, Ralf 2912
McCarthy, Amanda 2968
McClelland, James L. 19, 2045, 2849
McDaniel, J. Dean 860
McFate, Clifton 836, 1277
McNamara, Danielle ... 45, 764, 2483, 2681
Mech, Emily 1026

Médé, Butovens 2926
Meder, Björn 2531
Meijer, Rob R. 2501
Meltzoff, Andrew 53
Menager, David 620
Menard, Shannon 2929
Meng, Rui 2651
Mesgarani, Nima 1757
Mewhort, Douglas 1325, 2291
Meylan, Stephan C. 27
Mielicki, Marta 1811, 2931
Milburn, Evelyn 57
Miller, Hilary 1709, 2876
Mills, Caitlin 2914, 2981
Mills, Gregory 3012
Milton, Fraser 2243
Mintz, Toben 1661, 2561
Mirman, Daniel 1193
Miscevic, Filip 1313
Miser, Tracey 632
Mishler, Alan 2435
Mistry, Percy 211
Misyak, Jennifer 2864
Miwa, Kazuhisa 176, 247, 414, 686
Miyata, Hiromitsu 2874
Mizuguchi, Takashi 2865
Mollica, Francis 698
Monaghan, Padraic 51, 81, 123, 1703, 2991, 3046
Montambault, Brian 2921
Moon, Jung Aa 824
Moon, Yohan 2920, 3051
Moralez, Larry 948
Morano, Caroline 2968
Moreno, Keith 1685
Morgan, Caitlin 2099
Morgan, Thomas J. H. 27
Mori, Kensaku 414
Morris, Bradley 3024
Morrison, Don 1391
Moscoso del Prado Martin, Fermin ... 2585, 2645
Moss, Jarrod 2935
Moty, Kelsey 3023
Moulson, Margaret 2958
Mozuraitis, Mindaugas 2123
Mu, Jesse 2975
Mueller-Trede, Johannes 2285
Muldner, Kasia 656
Munhall, K.g. 3043
Munnich, Edward 2886
Munro, Paul 3056
Munroe, Abisha 474
Murakami, Akira 1050
Murphy, Caitlin 2968
Musslick, Sebastian 1547
Myers, Christopher 2997
Myers, David 866
Nagamine, Tasha 1757
Nagel, Jonas 806

Author Index

Nagy, David Gergely 2699
Nakamura, Kuninori 960
Nakamura, Yoshihiko 414
Nan, Weizhi 1457
Navarro, Daniel 544, 2471
Nazareth, Alina 2973, 3010, 3022
Neal, Jane 2988
Nelligan, Benjamin 2957
Nematzadeh, Aida 1313, 2309
Newcombe, Nora 2973
Newell, Ben 211
Newport, Elissa L. 2321
Nicoladis, Elena 2848
Nicoletti, Roberto 3032
Niedopytalski, Scott 3067
Niemi, Laura 1745
Nikitin, Evgenii 1565
Nilsenova, Marie 432
Noelle, David 2947
Noguchi, Takao 2864, 2915
Nonnenmacher, Frank 2875
Nonohay, Roberto 3068
Nordeen, Elke 2934
Nordmeyer, Ann 2789
Nosofsky, Robert 438
Novack, Miriam 2141
Novick, Laura 47
Nowak, Eric 895
Nozari, Nazbanou 716, 1193, 1625
Nunez, Rafael 2417
O'Donnell, Timothy 2975
O'Grady, Shaun 872, 1008
O'Leary, Allison 2996
O'Rourke, Polly 289
Oaksford, Mike 1104
Oda, Masahiro 414
Odean, Rosalie 3022
Odic, Darko 2923
Ohlsson, Stellan 942
Ohmoto, Yoshimasa 3041
Olds, Justin 2977
Ong, Desmond 135, 360
Onnis, Luca 2105
Opfer, John 752
Orban, Gergo 2699
Ortega, Gerardo 1182
Oudeyer, Pierre-Yves 1859
Ouimette, Brett 2369
Over, David 1104
Overlan, Matthew 1259
Overmann, Karenleigh 43
Özcan, Oğuzhan 1871
Ozcimder, Kayhan 1547
Özyürek, Asli 1182
Ozyurek, Asli 1295
Pacer, Michael D. 27, 1505
Pacer, Michael 3009
Paiva, Ana 1403
Pan, Chelsey 966
Pang, Bo 2357
Pani, John 47
Papaeliou, Christina F. 87
Papafragou, Anna 93, 782, 1146
Parker, Jeffrey 2225
Parker, Jessica 2201, 2237
Pärnamets, Philip 301, 3075
Parrini, Giacomo 331
Patel, Purav 2959
Patel, Rony 954, 1229, 2069
Patson, Nikole 1188
Patterson, John 2333
Patwary, Mostofa 1547
Paul, Iman 2225
Paul, Pooja 2995
Pavlova, Margarita 325
Paxton, Alexandra 2936, 2941
Peck, Kyung 420
Pederson, Eric 2985, 3016
Peeters, David 1637
Peloquin, Benjamin 319

Author Index

Peng, Yujia 229
Penner-Wilger, Marcie 2910, 3021
Pereira, Olivia 2932
Perez, Ray 45
Perfors, Amy 2471
Perret, Cecile 2483
Perry, Lynn 1026, 2956
Persaud, Kimele 3017
Peterson, Joshua 2363
Peterson, Mark 1062
Peterson, Rebecca 2135
Petkov, Georgi 325
Petrusz, Stephanie 474
Phillips, Colin 1445
Phillips, Jonathan 3007
Phillips, Steven 3058
Piantadosi, Steven 698, 1259, 2597
Pickering, Martin 1715
Pierguidi, Lapo 331
Pine, Julian 602, 674
Pipergias Analytis, Pantelis 918
Pirrone, Angelo 117, 3075
Pitt, Benjamin 1056, 2753
Poldrack, Russell A. 2039
Polyanskaya, Irina 2579, 3034
Polycarpou, Panayiota 3042
Poppels, Till 378
Posid, Tasha 3071
Postma-Nilsenova, Marie 2807
Postma-Nilsenová, Marie 1379
Postma, Eric 1379
Pothos, Emmanuel 2339, 2447
Powell, Derek 307
Powers, Jacklyn 372
Prasada, Sandeep 2987
Prather, Richard 2871
Pretzer, Gina 2892
Price, Thomas 2801
Protopapas, Athanassios 87, 277, 2889
Pruden, Shannon M. 3022
Pruden, Shannon 3010
Pu, He 2765
Purdy, John 1481
Pustejovsky, James 2841
Qian, Zhiying 608
Qing, Ciyang 1110
Quade, Michael 2549
Quam, Carolyn 2949
R. Katz, Irvin 3002
Rabkina, Irina 3055
Rabovsky, Milena 2045
Rączaszek-Leonardi, Joanna 901
Raizada, Rajeev D. S. 3060
Raizada, Rajeev 2249
Rakison, David 830
Ramasubramanian, Madhuri 2627
Ramscar, Michael 1763, 2441
Räsänen, Okko 1757, 2489, 3073
Rauwolf, Paul 1170
Raviv, Limor 1469, 1643
Ray, Laura 2723
Reali, Florencia 426
Rebuschat, Patrick 51, 3046
Redeker, Gisela 3012
Regier, Terry 2231, 2847
Rehder, Bob 1002, 1853, 2917
Rehg, James 235
Rehr, Victoria 1062
Reimers, Stian 626
Reinholtz, Nicholas 2225
Reitter, David 1223
Relander, Charlotta 586
Resnik, Philip 1445
Reyna, Valerie 1667
Rhodes, Marjorie 1319, 1541, 2477
Rhodes, Ryan 2984
Rhodes, Theo 474
Riano, Cesar 426
Richert, Rebekah 2884, 2891
Richland, Lindsey 2015

Author Index

Richter, Caitlin 1919
Richter, Mathis 1931
Righi, Stefania 331
Rips, Lance 55, 2111
Risko, Evan 2911, 2962, 3014
Rittle-Johnson, Bethany 265, 1739
Rivera, Samuel 1247, 2729
Rivlin, Ido 2967
Robbins, Talia 3005
Roberts, Gareth 1211, 2591
Roberts, Rachel 1841
Robinson, Chris 2201, 2237, 2729
Roby, Erin 2525, 2941
Roche, Jennifer 2783
Rodríguez, Alba 580
Roehrbein, Florian 3018
Rogers, Bradley 1301
Rogers, Timothy 2862, 2904
Röhrbein, Florian 3057
Romberg, Alexa 1283
Roney, Deborah 1062
Rosenthal-von der Pütten, Astrid Marieke .. 2925
Ross, Kimberly 740
Ross, Robbie 2894
Rossion, Bruno 2958
Rotaru, Armand S. 2207
Rothe, Anselm 2051
Rottman, Benjamin 153, 758
Rowe, Meredith 2159
Ruggeri, Azzurra 2675
Rupp, Sarah 2621
Ruskin, David 2831
Russell, Emily 2952
Ryan, Kaitlin 913
Ryu, Jeong 2920, 3051
Sadil, Patrick 2087
Sagi, Eyal 2906
Saide, Anondah 2891
Sala, Giovanni 1973, 2863, 2899, 2939
Salay, Nancy 295
Salgado, Harini 1919
Salomon, Meghan 2111
Salvucci, Dario 2465, 2969
Sanborn, Adam 2915
Sanborn, Sophia 2990
Sanches de Oliveira, Guilherme 1373
Sander, Emmanuel 1271
Sandham, Beth 1607
Saner, Lelyn 990
Sarmento, David 2931
Sarnecka, Barbara W. 241
Sass, Larry 1253
Satoh, Tetsuji 972
Saxe, Andrew 19, 3013
Saylor, Megan 913
Sayonno, Hannah 2015
Scerrati, Elisa 3032
Schaeffer, Laura 271
Scharp, Victoria 2946
Schereckenghost, Debra 3078
Scheutz, Matthias 2765
Schlenker, Philippe 55
Schlesinger, Molly 2884
Schloesser, Daniel 1913
Schmitz, Laura 2219, 2747, 2870
Schneider, Julia 2942
Schneider, Rose 2423
Schoeller, Anthea 2669
Schoener, Gregor 1931
Schoenherr, Jordan 147
Schoetensack, Christine 3046
Scholz, Agnes 129
Schonberg, Christina 2952
Schooler, Lael J. 1883
Schrater, Paul 2285
Schrodt, Fabian 3062
Schroeder, Scott R. 348
Schuler, Kathryn D. 2321
Schulz, Eric 1140, 2531
Schulz, Laura 2033

Author Index

Schurz, Gerhard 456
Schwarz, Jodi 2868
Scott, Rose 1439, 2525, 2941
Searle, Ellen 1205
Sebanz, Natalie 2870
Seidenberg, Mark 1026
Seino, Yuki 972
Seltman, Howard 812
Semmler, Carolyn 502
Senghas, Ann 1343
Sense, Florian 2501
Sepulveda, Gilbert 2936
Shafto, Patrick 1086

Shahriari, Bobak 1289
Shalin, Valerie 496
Shaman, Nicholas 2891
Shao, Ruxue 1038, 2507
Shapiro, Derek 2315
Sharma, Sugandha 1122
Shatz, Marilyn 49
Shay, Elizabeth A. 3060
Shi, Dongqing 2723
Shields, Joshua 474
Shiffrin, Richard 438, 2837
Shimojo, Shinsuke (shin) 182
Shipley, Thomas 47
Shipp, Nicholas 468, 2153
Shtulman, Andrew 2099
Shu, Tianmin 574
Shusterman, Anna 2813
Si, Mei 860, 2979
Siegel, Max 1265, 1493
Siegler, Robert 2881
Signorelli, Michela 2863
Siklos-Whillans, James 2937
Sikos, Les 1769
Silberglitt, Matt 372
Silliman, Daniel 2860
Sim, Zi L. 2411
Simmering, Vanessa 1709, 2904
Simonic, Mihael 3062
Sims, Chris R. 2273
Singh, Sonia 396
Sinnett, Scott 848, 2201, 3070
Sintov, Nicole 1391
Sirota, Miroslav 1949
Sivén, David 301
Skoe, Erika 3000
Skwarchuk, Sheri-Lynn 3021
Skyer, Raymond 2905
Sloman, Steven 1217
Slone, Lauren 2922
Sloutsky, Vladimir .. 632, 1247, 2195, 2974, 2996, 3004, 3071
Slusser, Emily 2813
Smith, Gretchen 1481
Smith, Jack 402
Smith, John 3011
Smith, Kenny 41, 43, 1241
Smith, Linda B. 2896
Smith, Linda 1631, 1835, 2922
Smith, Marc 2868
Smith, Pernille 3036
Smithwick, Daniel 1253
Smolensky, Paul 1487
Snedeker, Jesse 1343, 2859, 2995, 3052
Sneller, Betsy 2591
Snoddy, Sean 2938
Solomon, Gregg 45
Song, Amanda 1427, 3019
Soo, Kevin 153
Sovansky, Erin 942
Sowinski, Carla 3021
Spanola, Nicholas 2888
Sparks, Jesse R. 3039
Speed, Laura 1451, 2267
Speekenbrink, Maarten 1140, 2531
Spivey, Michael 1439, 2945
Spokoyny, Daniel 2645
Sprenger, Jan 432, 1901
Srinivasan, Mahesh . 1697, 2213, 2381, 3023

Author Index

Srivastava, Nisheeth 2285
Stafford, Tom 384, 3075
Starr, Ariel 1841
Stave, Matthew 3016
Steacy, Sandy 544
Steiner, Samantha 1217

Stephan, Simon 806, 2795
Stevenson, Suzanne . 1313, 2123, 2255, 2309
Stewart, Neil 2915
Stewart, Terrence 1122, 2183
Stine-Morrow, Elizabeth 538
Storms, Gert 1068, 1955, 2972
Stranahan, Laine 3052
Strandberg, Thomas 301
Straßmann, Carolin 2925
Strehlau, Sascha 129

Strickland, Brent 55, 2537, 3020
Stubblefield, Alexandra 1361
Suanda, Sumarga 1835
Suchow, Jordan W. 27, 1505, 2063
Suffill, Ellise 1715
Sugimoto, Fumie 3058
Sugimoto, Takayo 883
Sullivan, Emily 562
Sullivan, Jessica 2905
Sun, Ching Chu 2441
Sun, Felix 2033
Sun, Qizhang 895
Sun, Yanlong 402, 1457
Suzuki, Hiroaki 2874
Swallow, Khena 2872
Swerts, Marc 2603
Szary, Janelle 1397, 2989
Taatgen, Niels ... 31, 650, 2147, 2189, 2621
Tabatabaeian, Maryam 3015
Takac, Martin 532, 2345
Takano, Katsumi 2171
Takeda, Yuji 3058
Tamariz, Monica 2759
Tambe, Milind 1391
Tan, Shawn 656
Tanenhaus, Michael 2783
Tay, Charlene 2896
Taylor, Jordan 1829
Tenbrink, Thora 29, 2845
Tenenbaum, Josh .. 1158, 1265, 1493, 1649, 1679, 2033, 2495, 2777
Tenenbaum, Joshua B. 1781
Tenenbaum, Joshua 2567
Tenison, Caitlin 2887
Terai, Hitoshi 247, 686
Terzopoulos, Demetri 1805
Tessler, Michael Henry 1655, 2771
Tessler, Michael 2669
Tezuka, Taro 972
Thibaut, Jean-Pierre 1271, 1349
Thibodeau, Paul 866, 1535, 1571, 1769, 2315, 2902, 3059
Thomas, Ashley J. 241
Thomas, Sujith 596
Thompson-Schill, Sharon L 2853
Thompson, Robin L. 2387
Thomsen, Lotte 241
Thomson, Robert 147
Thorn, Paul 456
Thorstad, Robert 1991, 2003
Thüring, Manfred 2657
Thurman, Steven 229, 574
Tillman, Katharine 49, 2381
Ting, Titus 770
Tiv, Mehrgol 57
Toivonen, Ida 1961
Tokita, Midori 1523
Tomm, Brandon 1199
Tompkins, Connie 2946
Toomarian, Elizabeth 2651
Torabi Asr, Fatemeh 366, 1092
Toscano, Joe 2932, 2934
Toyama, Naohiro 686
Trafton, Greg 23, 2861, 2964
Tran, Jessica 2968
Trapp, Anna K. 2657

Author Index

Trausan-Matu, Stefan 764
Trimbach, Carl 1158
Trinchero, Roberto 1973
Tripp, Bryan 1583
Tripp, James 2915
Trueblood, Jennifer 25, 2447
Trueswell, John 2537
Tseng, Christine 2231
Tsuji, Sho 33
Tubridy, Shannon 2992
Tupper, Paul 508, 1289
Twomey, Katherine 336, 680
Tylén, Kristian 901, 3036
Tzneg, Yuhtsuen 2924
Ueda, Kazuhiro 223, 3041
Ullman, Tomer 1493, 1613, 1649
Unger, Layla 188, 199
Upal, M. Afzal 3079
Urminsky, Oleg 1727
Uscianowski, Colleen 2869
Utsumi, Akira 2171, 3035
Valdez, Gerardo 1481
Valenti, Andrew 2765
Vallee-Tourangeau, Frederic 468, 2153
Vallée-Tourangeau, Frédéric 2627
Vallee-Tourangeau, Frederic 1949
Vallée-Tourangeau, Gaëlle 2627
Vallee-Tourangeau, Gaelle 1949
Van Deemter, Kees 159
van den Berg, Martin 450
van den Berg, Ronald 2918
van der Loo, Janneke 99
Van Gompel, Roger 159
van Rij, Jacolien 550
van Rijn, Hedderik 2501
van Rooij, Marieke M. J. W. 1673
van Rooij, Marieke 568
van Rooij, Robert 2081
van Vugt, Marieke 2147, 2189
Vankrunkelsven, Hendrik 2972
Varma, Sashank 2959
Vasilyeva, Nadya 2663
Veksler, Vladislav Daniel 2405
Vélez, Natalia 966, 1985
Ventura, Salvatore 1973
Verbos, John 2983
Verheyen, Steven 1955, 2972
Verhoef, Tessa 2615
Vertolli, Michael 3044
Vesper, Cordula 2219, 2870
Viaud, Jeffrey 3053
Vicari, Stephanie 2998
Vick, Sarah-Jane 3072
Viggiano, Maria Pia 331
Vigliocco, Gabriella 2207
Vilone, Daniele 3074
Vinos, Michael 87
Vinson, David 1433
Vitale, Jonathan 514
Vlach, Haley 2876
Voelker, Aaron 2021
Vogt, Paul 486, 1116
von Sydow, Momme 105, 2944
Vong, Wai Keen 2471
Vonk, Jennifer 1205
Voorspoels, Wouter 1068
Vrabic, Sarah 2934
Vromans, Ruben 2807
Vul, Edward 746, 2285
Wagner, Katie 49, 1595
Wahn, Basil 2747, 2909
Wakefield, Elizabeth 2141
Waldmann, Michael R. 2795
Waldmann, Michael 307, 2912
Walin, Haley 872
Walk, Anne 396, 1481
Walker, Esther 2417, 2615
Walker, Jessica 480
Walker, Maegen 848, 3070
Wallace, Sarah 2983

Author Index

Wallach, David 2868
Wallot, Sebastian 2916
Wallsten, Thomas 2435
Walsh, Matthew 75
Walter, Florian 3057
Waltzer, Talia 1080
Wang, Hao 1661, 2513, 2561
Wang, Hongbin 402, 1457
Wang, Panqu 1409
Wang, Yufei 217
Wang, Zheng 25
Warlaumont, Anne 2892
Warren, Tessa C. 3026
Warren, Tessa 57
Was, Christopher 3024
Washington, Jeffery 1685
Wasserman, Charles 3000
Wassing, Olivia 2910
Wasylyshyn, Christina 1517, 2885
Watts, Paul 408, 2453
Weber, Andrea 1775, 1817
Weber, René 1847
Weinberger, Emma 2886
Weisberg, Deena 1829
Weisman, Kara 2741
Weiss, Daniel 111
Weiss, Staci 2913
Welbourne, Stephen 123
Wellwood, Alexis 55, 2970
Welsh, Matthew 502, 544
Wennekers, Thomas 2183
Wente, Adrienne 770
Wertz, Annie 3020
West, Robert 2399
Westermann, Gert 336, 680
White, Anne 1955
White, Darcy 2962
Widen, Sherri 49
Wiemer, Katja 2988
Wiese, Eliane 954, 1229
Wieth, Mareike 1685
Wiley, Jennifer 1811, 2929, 2931
Wilkenfeld, Daniel 1979
Willett, Ciara 2917
Williams, Alexander 55
Williams, Katie 2905
Williams, Mary-Anne 514
Willits, Jon 1092, 1799
Willke, Theodore L. 1547
Wills, Andy 69, 2243
Wilson, Colin 2429, 2900
Wilson, Robert 1751, 1793
Wiraatmadja, Sandy 2519
Wittenberg, Eva 15
Wnuk, Ewelina 2687
Woensdregt, Marieke 1241
Wolff, Phillip 1991, 2003
Wong, Aaron 2935
Wong, J. J. 854
Woodward, Amanda 53
Worthy, Darrell 2357
Wu, Charley 918
Wu, Jiajun 1781
Wu, Nianyang 2878
Wu, Te-Lin 2849
Wu, Yongxu 3037
Wuethrich, Charles 462
Wulff, Dirk U. 907
Wynn, Karen 3020
Xing, Chenmu 3069
Xu, Fei 13, 872, 1008, 2177, 2411, 2675
Xu, Lining 3037
Xu, Tian 1997
Xu, Yang 1613, 2213, 2231, 2847
Yakushijin, Reiko 2898
Yamada, Seiji 930
Yamagiwa, Hideo 794
Yang, Charles 2321
Yang, Li-chiung 2961
Yazbec, Angele 1421

Author Index

Yazdi, Haleh 259
Ye, Jian-Ping 480
Yearsley, James 25, 2339, 2447
Yeo, Darren J. 265
Yildirim, Ilker 1265, 2495
Yip, Michael C. W. 3072
Yoo, Kwanghyeon 2855, 2919
Yoon, Erica J. 2771
Yoon, Erica 2789
Yoon, Sangsuk 728
Young, Laura 592
Young, Liane 1745
Young, Mackenzie 2900
Yu, Angela 1427, 2705, 3019
Yu, Chen 1631, 1835, 1997, 2027, 2856, 2896, 2922, 2943
Yu, Lap-Fai (craig) 17
Yu, Ru Qi 3040
Yu, Yue 1086
Yuan, Arianna 2849
Yuceil, Selgun 2963
Yun, Hongoak 2105
Yurovsky, Daniel 1691, 2093
Zacks, Jeffrey 53
Zahner, Doris 3069
Zaki, Jamil 360, 966
Zapata-Rivera, Diego 2907, 3002, 3039
Zarl, Francesca 1134
Zemla, Jeffrey 1217, 1907
Zeveney, Andrew 1020
Zevin, Jason 1661, 2561
Zhan, Meilin 2393
Zhang, Chengkai 1781
Zhang, Jianwei 3061, 3076
Zhang, Liwei 3061
Zhang, Marianna 1967
Zhang, Qi 117
Zhang, Renqiao 1781
Zhang, Yayun 2027
Zhao, Jiaying 984, 1199, 3040
Zhao, Qiyang 3065
Zhao, Xiaowei 2928
Zhao, Xin 1877
Zhao, Xuan 924
Zhao, Yibiao 17
Zheng, Yeqiu 2603
Zhou, Guojing 2801
Zhu, Mona 3014
Zhu, Rebecca 2990
Zhu, Song-Chun 574, 1805
Zhu, Yixin 17, 1805
Ziegler, Jayden 2859
Zimmerman, Corinne 2954
Zinszer, Benjamin 2249

Reviewers

Joshua Abbott
Akinori Abe
Drew Abney
Dor Abrahamson
Darren Abramson
Cengiz Acarturk
Rakefet Ackerman
Staub Adrian
Sheeraz Ahmad
Woo-Young Ahn
Nameera Akhtar
Garnham Alan
Martha Alibali
Afra Alishahi
Carsten Allefeld
Kelsey Allen
Laura Allen
Erik Altmann
Ben Ambridge
Elena Andonova
Janet Andrews
Mark Andrews
Florencia Anggoro
Keith Apfelbaum
Joanne Arciuli
Selene Arfini
Engin Arik
Blair Armstrong
Clementking Arockiasamy
Rosa Arriaga
Sudha Arunachalam
Mika Asaba
Richard Ashley
Natsuki Atagi
Nart Bedin Atalay
Nada Attar
Luc Augier
Joseph Austerweil
Edmond Awad
Habiba Azab
Roger Azevedo
Harald Baayen
Ted Bach
Annalisa Baicchi
Todd Bailey
Jonathan Bakdash
Chris Baker
Lewis Baker
Ryan Baker
Colin Bannard
Bonny Bannerjee
Jean Baratgin
Mike Barley
David Barner
Daniel Bartels
Christopher Bates
Malcolm Bauer
Peter Baumann
Merja Bauters
Paul Baxter
Barend Beekhuizen
Randall Beer
Christina Behme
Anna Belardinelli
Daniel Belenky
Sieghard Beller
Paul Bello
Andrea Bender
Viridiana L. Benitez
Kuipers Benjamin
Jordan Benko
Elika Bergelson
Christina Bergmann
Kirsten Bergmann
Sven Bertel
Luc Berthouze
Tommaso Bertolotti
Tarek R. Besold
Brad Best
Sudeep Bhatia
Francesco Bianchini
Sarah Bibyk
Klinton Bicknell
Perrin Bignoli
Dorrit Billman
Jennifer Bittner
Michael Bixter
Leslie Blaha
Kristen Blair
Tommy Blanchard
Fernando Blanco
Julia Blau
Stephen Blessing
Peter Blouw
Reihane Boghrati
Mario Bonato
Luca Bonatti
Elizabeth Bonawitz
Rebecca Boncoddo
Jean-Francois Bonnefon
Anna Borghi
Arielle Borovsky
Heather Bortfeld
Solveig Bosse
Mona Roxana Botezatu
Lewis Bott
Oliver Bott
Bastien Boutonnet
Ty Boyer
David Braithwaite
Neil Bramley
Rebecca Brand
Amanda Brandone
Holly Branigan
Richard Breheny
Johns Brendan
Jon Brennan
Will Bridewell
Sophie Bridgers
Henry Brighton
Mirjam Broersma
Thomas Brooks
Geoffrey Brookshire
Harm Brouwer
Joshua Brown
Meredith Brown
Michael Brownstein
Riccardo Brunetti
Angela Brunstein

Reviewers

Joerg Brunstein
Tad Brunye
Monica Bucciarelli
David Buchanan
Daphna Buchsbaum
Cameron Buckner
Marc Buehner
John Bunce
Ann Bunger
Curt Burgess
Michele Burigo
Joseph Burling
Bruce Burns
Daniel Burnston
Hendrik Buschmeier
Lucas Butler
Martin Butz
Angelo Cangelosi
Reality Canty
valentina cardella
Floriana Carmeci
Peter Carruthers
Paulo Carvalho
Claudia Casadio
Daniel Casasanto
Marisa Casillas
Monica Castelhano
Gianluca Caterina
Richard Catrambone
Tom Caudell
Gabrielle Cayton-Hodges
Joel Chan
Margaret Chan
Sanjay Chandrasekharan
Franklin Chang
Emmanuel Chemla
Dar-Wei Chen
Eva E. Chen
Jidong Chen
Lang Chen
Peter Cheng
Erik Cheries
Nadia Chernyak
Dana Chesney
Fabienne Chetail
Pierina Cheung
Jane Childers
Seth Chin-Parker
Morten Christiansen
Stella Christie
Evangelia Chrysikou
Adam Chuderski
Mina Cikara
Benjamin Cipollini
William Clancey
John Clapper
Timothy Clausner
Meghan Clayards
Catherine Clement
Charles Clifton
Donna Coch
Moreno I. Coco
Ruben Coen-cagli
Anna Coenen
Trevor Cohen
Matteo Colombo
Arzu Coltekin
Eliana Colunga
Christopher Conway
Susan Cook
Rick Cooper
Kensy Cooperrider
Sara Cordes
Mario Córdoba
Kathleen Corriveau
Fintan Costello
Gary Cottrell
Caitlin Coughlin
Seana Coulson
Ben Cowan
Rosie Cowell
Christopher Cox
Gregory Cox
Scotty Craig
Patrick Cravalho
Nicole Craycraft
Christopher Crick
Amy Criss
William Croft
Steve Croker
Vincenzo Crupi
Jennifer Culbertson
Chris Cummins
Fred Cummins
Fiery Cushman
Christine Cuskley
Ilaria Cutica
Francesca D'errico
Rick Dale
Gregory Dam
Irina Danileiko
David Danks
Judith Danovitch
Isabelle Dautriche
Jules Davidoff
Jim Davies
Felipe De Brigard
Irmgard de La Vega
Josh de Leeuw
Paul De Palma
Virginia de Sa
Gedeon Deák
Christelle Declercq
Judith Degen
Morteza Dehghani
Guillermo Del Pinal
Gary Dell
Sara Dellantonio
Claire Delle Luche
Wei (sophia) Deng
Stephanie Denison
Jean-Louis Dessalles
Melissa Dewolf
Barbara Di Eugenio
Julien Diard
Brian Dillon
Alexia Toskos Dils
Jacob Dink
James Dixon

Reviewers

Tim Donovan
Alex Doumas
Gabe Doyle
Anna Drummey
Sarah Duffy
Sean Duffy
Sophie Dufour
Susan Duncan
Frank Durgin
Kelley Durkin
Melody Dye
Sarah Dziura
Baxter Eaves
Catharine Echols
Charlotte Edmunds
Suzanne Egan
Paul Egre
Brannon Elizabeth
Michelle Ellefson
Giulia Elli
Jan Engelmann
Russell Epstein
Susan Epstein
Lucy Erickson
Margaret Evans
Zoe Falomir
Mandy Faretta-Stutenberg
Mirko Farina
Igor Farkas
Thomas Farmer
Simon Farrell
Caitlin Fausey
Luis H. Favela
Mark Fedyk
Maryia Fedzechkina
Michele Feist
Naomi Feldman
Brock Ferguson
Heather Ferguson
Ferreira Fernanda
Diego Fernandez Duque
Mark Finlayson
Sara Finley
Bridgid Finn
Jozsef Fiser
Anna Fisher
Cynthia Fisher
Julia Fisher
Paula Fitzpatrick
Monique Flecken
Stephen Flusberg
Jonathan Folstein
Ken Forbus
Alice Foucart
Ruthe Foushee
Ana Franco-Watkins
Steve Franconeri
Mike Frank
Stefan Frank
Michael Franke
Diego Frassinelli
Scott Fraundorf
Alexandra Frazer
Christian Freksa
Melinda Fricke
Marcello Frixione
Caren Frosch
Wai-Tat Fu
Danilo Fum
Richard Futrell
Emily Fyfe
Rami Gabriel
Francesco Gagliardi
Sarah Gaither
Alexia Galati
Chiara Gambi
Timothy Gann
Ricardo Garcia
Bärbel Garsoffky
Justin Garten
Susan Gelman
Silvia Gennari
Dedre Gentner
Louann Gerken
Altmann Gerry
Samuel Gershman
Tobias Gerstenberg
Valeria Giardino
Nina Gierasimczuk
Tom Gijssels
Daniel Gildea
Maureen Gillespie
Michael Glass
Joseph Glavan
Vladimir Glebkin
Kevin Gluck
Matthias Gobel
Silke Göbel
Fernand Gobet
Karrie Godwin
Peter Michael Goebel
Kelly Goedert
Ashok Goel
Tilbe Goksun
Micah Goldwater
David Gomez
Rebecca Gomez
Tao Gong
Laura Gonnerman
Cleotilde Gonzalez
Noah Goodman
Martijn Goudbeek
Katharine Graf Estes
Erin Grant
Mario Graziano
Christopher Green
Nathan Greenauer
Stefan Th. Gries
Oren Griffiths
Sascha Griffiths
Tom Griffiths
Maurice Grinberg
John Grundy
Ricardo Gudwin
Amal Guha
Elizabeth Gunderson
Glenn Gunzelmann
Todd Gureckis
Robin Gustafson

Reviewers

Hyowon Gweon
Marc Halbrügge
John Hale
Matthew Hall
Tim Halverson
Kiley Hamlin
Jessica Hamrick
J. Isaiah Harbison
Mary Hare
Kevin Harris
Joshua Hartshorne
Robert Hartsuiker
Paul Haward
Kara Hawthorne
Mary Hayhoe
Angela Xiaoxue He
Karl Healey
Mary Hegarty
Marco Heimann
Steve Hekkanen
Daphna Heller
Pernille Hemmer
Marlone Henderson
Li Heng
Seth Herd
Stefan Herzog
Gail Heyman
Laura Hiatt
Shohei Hidaka
Patrick Hill
Caitlin Hilliard
Thomas Hills
Xavier Hinaut
Dale Hirsch
Guy Hochman
Marieke Hoetjes
Janina Anna Hoffmann
Annette Hohenberger
Anders Højen
John Holden
Kevin Holmes
garrett honke
Joe Hoover
Zachary Horne
Gernot Horstmann
Joseph Houpt
Michael Hout
Christine Howes
Evgenia Hristova
Penka Hristova
Janet Hsiao
Nina Hsu
Yi Ting Huang
Edward Hubbard
Stefan Huber
Eva Hudlicka
John Hummel
Yu-Chen Hung
Julie Hupp
E. Matthew Husband
Larisa Hussak
Hyesung Grace Hwang
Thomas Icard
catalina iricinschi
Alistair Isaac
Veronique Izard
Eduardo Izquierdo
Josh Jackson
Tanner Jackson
Cassandra Jacobs
T. Florian Jaeger
Georg Jahn
Ariel James
Karolina Janacsek
Julian Jara-Ettinger
Alan Jern
William Jimenez-Leal
Mark Johansen
Elizabeth johnson
Kate Johnson
Mark Johnson
Samuel Johnson
Angie Johnston
Gary Jones
Lara Jones
Jerome Scott Jordan
Pamela Jordan
David Joyner
George Kachergis
Elsi Kaiser
Michael Kalish
Rachel Kallen
Irene Kan
Frank Kanayet
Padmapriya Kandhadai
Katerina Kantartzis
Justine Kao
Stec Kashmiri
White Katherine
Irvin Katz
John Kearns
Andrew Kehler
John Kelleher
Frank Keller
Christopher Kello
Damian Kelty-Stephen
William Kennedy
Trina Kershaw
Alan Kersten
Sangeet Khemlani
Peter Khooshabeh
Celeste Kidd
Evan Kidd
Douwe Kiela
David Kieras
Dahee Kim
Eunkyung Kim
Shinwoo Kim
Ildiko Kiraly
Szabolcs Kiss
Max Kleiman-Weiner
Daniel Kleinman
Dave Kleinschmidt
Melissa Kline
Alexander Klippel
Heidi Kloos
Joshua Knobe
Pia Knoeferle
Alistair Knott

Reviewers

Adam Koblinger
Ken Koedinger
Judith Koehne
Melissa Koenig
Takanori Komatsu
Jonathan Kominsky
Piotr Konderak
Agnieszka Konopka
Emmanouil Konstantinidis
Ruud Koolen
Stefan Kopp
Nate Kornell
Boris Kotchoubey
Emiel Krahmer
Ekaterina Kravtchenko
Josef Krems
Carol Lynne Krumhansl
Angelos Krypotos
Marta Kryven
Karolina Krzyżanowska
Stefanie Kuchinsky
Sarah Kucker
Anna Kuhlen
Kai-Uwe Kühnberger
Bora Kumova
Maithilee Kunda
Lars Kunze
Kenneth Kurtz
Chigusa Kurumada
Tamar Kushnir
Peter Kvam
Tei Laine
Itziar Laka
Brenden Lake
Tuan Lam
Christine Lamarre
David Landy
Jill Lany
Gabriella Lapesa
Chris Lawson
David Leake
Gianluca Lebani
Christian Lebiere
Jessica Lee
Michael Lee
Blair Lehman
Jaroslaw Lelonkiewicz
Alessandro Lenci
Nick Leonard
Giuseppe Leonardi
Beth Levin
William Levine
Joshua Levy
Roger Levy
Simon Levy
Casey Lew-Williams
Molly Lewis
Jixing Li
Peggy Li
Peipeng Liang
Melissa Libertus
Patricia Lichtenstein
Karen Lichtman
Falk Lieder
Constantine Lignos
Jared Linck
Shane Lindsay
Marcus Lindskog
Hannah Little
Brandon Liverence
Tania Lombrozo
Yanjin Long
Deryle Lonsdale
Tamara Lorenz
Jeff Loucks
Max Louwerse
Andrew Lovett
Hongjing Lu
Luca Luca Surian
Christopher Lucas
Elliot Ludvig
George Luger
Christian Luhmann
Steven Luke
Jian Luo
Stephen Lupker
Gary Lupyan
David Luque
Dermot Lynott
Nieuwland M.
Kyle MacDonald
Maryellen MacDonald
Ross Macdonald
Luís Macedo
Edouard Machery
Michael Mack
Alex Madva
Hagit Magen
Brian Magerko
Rachel Magid
Lorenzo Magnani
James Magnuson
Kyle Mahowald
Steve Majerus
Asifa Majid
Vicente Malave
Laurence Maloney
Tara Mandalaywala
Eric Mandelbaum
Tyler Marghetis
Bedny Marina
Doug Markant
Art Markman
Lori Markson
Jessecae Marsh
Louis Marti
Andrea Martin
Salvador Mascarenhas
Amy Masnick
Ingrid Masson Carro
J. Douglas Mastin
Bryan Matlen
Michi Matsukura
Danielle Matthews
Justin Matthews
Percival Matthews
Laura Matzen
Ralf Mayrhofer
Reiko Mazuka

Reviewers

Stewart McCauley
John McCoy
Janet McLean
Conor McLennan
Marjorie McShane
Bjoern Meder
Jared Medina
Catherine Mello
David Mendonca
Hugo Mercier
Lotte Meteyard
Douglas Mewhort
Stephan Meylan
Evelyn Milburn
Lynden Miles
Petar Milin
Craig Miller
Gregory Mills
Toby Mintz
Jelena Mirkovic
Reza Moazzezi
Jeff Moher
Monika Molnar
Padraic Monaghan
Jessica Montag
Daniel Montello
Jung Aa Moon
Claire Moore-Cantwell
Adam Moore
Laura Morett
Emily Morgan
Bradley Morris
Fermin Moscoso del Prado Martin
Jarrod Moss
Johannes Mueller-Trede
Edward Munnich
Paul Munro
Dafne Muntanyola
Stefan Münzer
Kou Murayama
Gregory Murphy
Christopher Myers
Myrto Mylopoulos

Jonas Nagel
Shweta Naik
Kuninori Nakamura
Bhuvana Narasimhan
Dezso Nemeth
Ani Nenkova
Nora Newcombe
George Newman
Andreas Højlund Nielsen
Laura Niemi
Marie Nilsenova
Nicolae Nistor
Yael Niv
David Noelle
Nicholaus Noles
Jonas Nölle
Nader Noori
Ann Nordmeyer
Ara Norenzayan
Catherine Norris
Miriam Novack
Ira Noveck
Jared Novick
Nazbanou Nozari
Padraig O'Seaghdha
Mike Oaksford
Tim Oates
Unaizah Obaidellah
Darko Odic
Marta Olivetti
Ana-Maria Olteteanu
Alessandro Oltramari
Akira Omaki

Desmond Ong
Luca Onnis
John Opfer
Gary Oppenheim
Gergo Orban
Daniel Osherson
Korinn Ostrow
Iris Chuoying Ouyang
David Over
Amanda Owen Van Horne

Michael Pacer
Sebastian Padó
Bozena Pajak
Kiran pala
Zsolt Palatinus
Martha Palmer
Bo Pang
John Pani
Peter Pantelis
Carita Paradis
Woosuk Park
Dan Parker
Jeff Parker
Philip Pärnamets
Fey Parrill
Luigi Pastore
Andrea Patalano
Thanujeni Pathman
Sturt Patrick
Nikole Patson
Muentener Paul
David Pautler
Brennan Payne
Lisa Pearl
Neal Pearlmutter
Sarah Bro Pedersen
David Peebles
Madeline Pelz
Marcie Penner-Wilger
Antonino Pennisi
Alfredo F. Pereira
Amy Perfors
Marcus Perlman

Pamela Perniss
Lynn Perry
Megan Peters
Joshua Peterson
Matthew Peterson
Niki Pfeifer
Eugene Philalithis
Colin Phillips
Jonathan Phillips
Steven Phillips

Reviewers

Steve Piantadosi
Maria M. Pinango
Benjamin Pitt
Alex Pitti
Alessio Plebe
Amanda Pogue
Guy Politzer
Ferran Pons
Marie Postma-Nilsenová
Eric Onno Postma
Emmanuel Pothos
Derek Powell
Lindsey Powell
Lucia Pozzan
Ranjani Prabhakaran
Sandeep Prasada
Vitor Profeta
Athanassios Protopapas
Guillermo Puebla-Ramirez
Cynthia Putnam
David Pynadath
Jennifer Queen
Joao Queiroz
Hugh Rabagliati
Gabriel Radvansky
Anna Rafferty
Marco Ragni
Maartje Raijmakers
jason ralph
Gary Raney
Michael Ranney
Okko Johannes Räsänen
Martina Rau
Kyle Rawlins
Florencia Reali
Elizabeth Redcay
Gisela Redeker
A. David Redish
Stephen Reed
Patricia Reeder
Terry Regier
Bob Rehder
Nicholas Reinholtz
Margaret Renwick
Kate Revill
Russell Revlin
Theo Rhodes
Ariane Rhone
Alexander Rich
Mike Richardson
Rebekah Richert
Elizabeth Richey
Lindsey Richland
Chris Riesbeck
Lance Rips
Lilia Rissman
Samuel Rivera
Talia Robbins
Katherine Robinson
Jennifer Roche
Florian Roehrbein
Ardi Roelofs
Tanja Roembke
Etienne Roesch
Hannah Rohde
Katharina Rohlfing
Matthias Rolf
Alexa Romberg
Joost Rommers
Samuel Ronfard
Paul Rosenbloom
Federico Rossano
Eleonora Rossi
Constantin Rothkopf
Benjamin Rottman
Joshua Rottman
Brandon Roy
Azzurra Ruggeri
Joshua Rule
Rachel Ryskin
Mark Sabbagh
Matthew Sachs
Eyal Sagi
William Sakas
Joanna Salapska-Gelleri
Elena Salillas
Adam Sanborn
Raúl Sánchez–garcía
Julio Santiago
Brown-Schmidt Sarah
Karanam Saraschandra
Barbara Sarnecka
Cristina Savin
Megan Saylor
Brian Scassellati
Adena Schachner
Daniel Schlegel
Karen Schloss
Ute Schmid
Tom Schoenemann
Kathryn Schuler
Eric Schulz
Stephanie Schwartz
Florian Schwarz
Rose Scott
Nicholas Searcy
Colleen Seifert
Elizabeth Seiver
Andrea J. Sell
J. Ignacio Serrano
Thomas Serre
Annalisa Setti
Mike Shafto
Natela Shanidze
Kailen Shantz
Azim Shariff
Alex Shaw
Jason Shaw
Shlomi Sher
Wenlei Shi
Atsushi Shimojima
Andrew Shtulman
Mo Shukla
Noam Siegelman
Tal Siloni
Catriona Silvey
Victoria Simms
Scott Sinnett
Miroslav Sirota

Reviewers

Paul Skarratt
Vladimir Sloutsky
Kevin Smith
Linda Smith
Hannah Snyder
William Snyder
Joo-Hyun Song
Park Soojin
George Spanoudis
Jesse Sparks
Ann Speed
Laura Speed
Maarten Speekenbrink
Jennifer Spenader
Stephen Spiller
Michael Spranger
Jon Sprouse
Maria Spychalska
Mahesh Srinivasan
Nisheeth Srivastava
Tom Stafford
Maria Staudte
Mark Steedman
Nancy L Stein
Shane Steinert-Threlkeld
Keith Stenning
Carl Stern
Most Steven
Terrence Stewart
Mallory Stites
Kurt Stocker
Britta Stolterfoht
Brent Strickland
Sumarga Suanda
Jordan Suchow
Jacob Suher
Jess Sullivan
Naveen Sundar G
Abigail Sussman
Khena Swallow
Daniel Swingley
Gabriel Synnaeve
Kristen Syrett
Jakub Szymanik
Niels Taatgen
Whitney Tabor
Darren Tanner
Heike Tappe
Caglar Tas
Loukia Taxitari
Julia Taylor
Thora Tenbrink
Josh Tenenbaum
katya tentori
Kinneret Teodorescu
Susan Teubner-Rhodes
Khanh-Phuong Thai
Jean-Pierre Thibaut
Paul Thibodeau
Erik Thiessen
Serge Thill
Kathleen Thomas
Bill Thompson
Clarissa Thompson
Jim Thompson
Robert Thomson
Alex Tillas
Kevin Tobia
Ida Toivonen
Joe Toscano
Greg Trafton
Bastien Trémolière
Melissa Troyer
John Trueswell
Ricky Van Yip Tso
Shannon Tubridy
Luca Tummolini
Alba Tuninetti
Peter Turkeltaub
Barbara Tversky
Katherine Twomey
Kristian Tylen
Taiji Ueno
Tomer Ullman
Monika Undorf
Christoph Ungemach
Layla Unger
Oleg Urminsky
David Uttal
Irem Uz
Virginia Valian
Juan Valle-Lisboa
Frederic Vallee-Tourangeau
Ronald van den Berg
Sabine van der Ham
Natalie van der Wal
Julie Van Dyke
Jo Van Herwegen
Donald Alexander Varakin
Alexandra Varga
Sashank Varma
Nadya Vasilyeva
Ashish Vaswani
Argiro Vatakis
Richard Veale
Bella Veksler
Natalia Vélez
Steven Verheyen
Francesco Vespignani
Timothy Vickery
Ferreira Victor
David Vinson
Jonathan Vitale
Haley Vlach
Carl Vogel
Paul Vogt
John Voiklis
Regina Vollmeyer
Titus von der Malsburg
Bettina von Helversen
Gottfried Vosgerau
Ed Vul
Matt Wagers
Adrian Walker
Caren Walker
Erin Walker
Esther Walker
Sebastian Wallot
Matthew Walsh

Reviewers

Felix Wang
Panqu Wang
Tony Wang
Liza Ware
Anne Warlaumont
Tessa Warren
Christopher Was
Duane Watson
Kirsten Weber
Deena Weisberg
Daniel Weiss
Alexis Wellwood
Matthew Welsh
Janet Werker
Chris Westbury
Gert Westermann
Andrew Whalen
Aaron Steven White
Jeffrey White
Alex Wiegmann
Eliane Wiese
Joseph Jay Williams
Andy Wills
Colin Wilson
Bodo Winter
Eva Wittenberg
Ralph Wojtowicz
Phillip Wolff
Sharon Wood
Jiajun Wu
Fei Xu
Shaorong Yan
Charles Yang
Lee-Xieng Yang
Defu Yap
Eiling Yee
Saiwing Yeung
Ilker Yildirim
Hyungwook Yim
Michael Yip
Erica Yoon
Si On Yoon
Masaya Yoshida
Jeffrey Yoshimi
Galit Yovel
Chen Yu
Yue Yu
Dan Yurovsky
Jeff Zacks
Jeffrey Zemla
Cristina Zepeda
Arkady Zgonnikov
Shunan Zhang
Changkun Zhao
Jiaying Zhao
libo zhao
Xuan Zhao
Yibiao Zhao
Corinne Zimmerman
Iraide Zipitria
Megan Zirnstein